PLOS | GENETICS

RESEARCH ARTICLE

Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies

Qiongshi Lu[1☯], Ryan Lee Powles[2☯], Qian Wang[2], Beixin Julie He[3], Hongyu Zhao[1,2*]

1 Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, United States of America, **2** Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **3** Division of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, Connecticut, United States of America

☯ These authors contributed equally to this work.
* hongyu.zhao@yale.edu

OPEN ACCESS

Citation: Lu Q, Powles RL, Wang Q, He BJ, Zhao H (2016) Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS Genet 12(4): e1005947. doi:10.1371/journal.pgen.1005947

Editor: Hua Tang, Stanford University, UNITED STATES

Received: November 13, 2015

Accepted: March 1, 2016

Published: April 8, 2016

Data Availability Statement: GWAS signal prioritization software and all annotation data are available from the GenoSkyline website at http://genocanyon.med.yale.edu/GenoSkyline.

Funding: This study was supported in part by the National Institutes of Health grants R01 GM59507, the VA Cooperative Studies Program of the Department of Veterans Affairs, Office of Research and Development, and the Yale World Scholars Program sponsored by the China Scholarship Council. The funders had no role in study design,

Abstract

Extensive efforts have been made to understand genomic function through both experimental and computational approaches, yet proper annotation still remains challenging, especially in non-coding regions. In this manuscript, we introduce GenoSkyline, an unsupervised learning framework to predict tissue-specific functional regions through integrating high-throughput epigenetic annotations. GenoSkyline successfully identified a variety of non-coding regulatory machinery including enhancers, regulatory miRNA, and hypomethylated transposable elements in extensive case studies. Integrative analysis of GenoSkyline annotations and results from genome-wide association studies (GWAS) led to novel biological insights on the etiologies of a number of human complex traits. We also explored using tissue-specific functional annotations to prioritize GWAS signals and predict relevant tissue types for each risk locus. Brain and blood-specific annotations led to better prioritization performance for schizophrenia than standard GWAS p-values and non-tissue-specific annotations. As for coronary artery disease, heart-specific functional regions was highly enriched of GWAS signals, but previously identified risk loci were found to be most functional in other tissues, suggesting a substantial proportion of still undetected heart-related loci. In summary, GenoSkyline annotations can guide genetic studies at multiple resolutions and provide valuable insights in understanding complex diseases. GenoSkyline is available at http://genocanyon.med.yale.edu/GenoSkyline.

data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Author Summary

Genome-wide association study has been a productive approach to studying human complex diseases in the past decade, yet challenges still remain in both identifying and interpreting disease-associated loci. In our previous work, we showed that integrated analysis of GWAS summary statistics and genomic functional annotations could enhance the performance of signal prioritization. In this paper, we further improve our annotation through integrating a rich collection of epigenomic data from the Roadmap Epigenomics Project. We introduce GenoSkyline, a statistical framework to predict tissue-specific functional regions in the human genome, and demonstrate its ability to capture tissue-specific functionality through extensive case studies. We then illustrate a variety of ways that GenoSkyline could benefit post-GWAS analysis. The performance of signal prioritization is further improved using annotations of disease-related tissue types. Furthermore, combining GenoSkyline annotations with GWAS results allow us to partition heritability by tissue types and generate new hypotheses regarding the disease etiology behind each risk locus, thereby providing novel biological insights to many human complex diseases. GenoSkyline is powerful, robust, and customizable. We believe that GenoSkyline and its applications can guide genetics research at multiple resolutions and greatly benefit the broader scientific community.

Introduction

Functionally annotating the human genome is a major goal in human genetics research. After years of community efforts, a variety of experimental and computational approaches have been developed and applied for genomic functional annotation. Comparative genomics studies have shown that approximately 4.5% of the human genome is conserved across mammals [1]. Furthermore, the rich collection of epigenomic data generated by large consortia, e.g. ENCODE [2] and Roadmap Epigenomics Project [3], also provides great insight for understanding the functional effects of the genome, especially in terms of non-coding regulatory machinery. To best utilize these rich data, we recently developed GenoCanyon [4], a non-coding functional prediction approach based on integrative analysis of annotation data, whose performance was demonstrated through predicting well-studied regulatory DNA elements. GenoCanyon provides general predictions of non-coding functional regions in the human genome but does not fully utilize cell-type-specific information of epigenomic data. Incorporating cell-type-specific or tissue-specific information into annotation tools is essential not only for understanding the basic biology of the genome, but also for better characterizing genetic variation, as in the functional interpretation of risk loci identified from genome-wide association studies (GWAS).

GWAS has been a great success in the past decade, yet challenges still remain in both identifying additional risk variants and interpreting GWAS results. Current practice employs a significance threshold (i.e. 5×10^{-8}) that controls family-wise error rate. However, this approach is known to be underpowered when effect sizes are weak or moderate at risk loci [5]. Moreover, nearly 90% of the genome-wide significant hits in published GWAS are located in non-coding regions whose functional impact to human complex traits is largely unknown [6]. Complex linkage disequilibrium (LD) patterns also hinder our ability to identify real functional sites among correlated SNPs. Several methods have been proposed to integrate annotation data for better prioritizing GWAS signals and their effectiveness has also been well demonstrated [7–10]. Tissue-specific functional annotations have the potential to bring even more biological insights to post-GWAS analysis and help understand complex disease etiology.

In this paper, we introduce GenoSkyline, a tissue-specific functional prediction tool based on integrated analysis of epigenomic annotation data. We demonstrate its ability to identify tissue-specific functionality from its performance to rediscover a number of experimentally validated non-coding elements. Next, we show valuable biological insights GenoSkyline can provide in post-GWAS analysis through integrative analysis of 15 human complex traits. We believe that GenoSkyline will prove to be a powerful tool for human genetics research because of its abilities to assess tissue-specific enrichment of GWAS signals, better prioritize GWAS signals, and offer biological interpretations of risk loci.

Results

Predicting tissue-specific functional regions in the human genome

The posterior probability of being functional given the annotation data is used to quantify tissue-specific functional potential of each nucleotide in the human genome (Methods). It will be referred to as GenoSkyline (GS) score in following sections. We calculated GS scores for 7 human tissue types; brain, gastrointestinal tract (GI), lung, heart, blood, muscle, and epithelium (S1 Table). With a GS score cutoff of 0.5, 22.2% of the human genome is predicted to be functional in at least one of these tissue types, while 1.7% is functional in all 7 tissues (Fig 1A). Since GS score has a bimodal pattern, these results are not sensitive to the cutoff choice (S1 Text).

Across tissue types, the percentage of predicted functional genome ranges from 5.4% (Lung) to 9.7% (GI) (Fig 1B and S2 Table). The overlap between heart-specific and muscle-specific functional regions is the largest among all pairs of tissues. Interestingly, although the percentage of functional genome in blood (8.4%) is similar to other tissue types, it overlaps less with the functional regions in other tissues (Fig 1C). This is consistent with the recent discovery that blood has the lowest levels of eQTL sharing with other tissues [11].

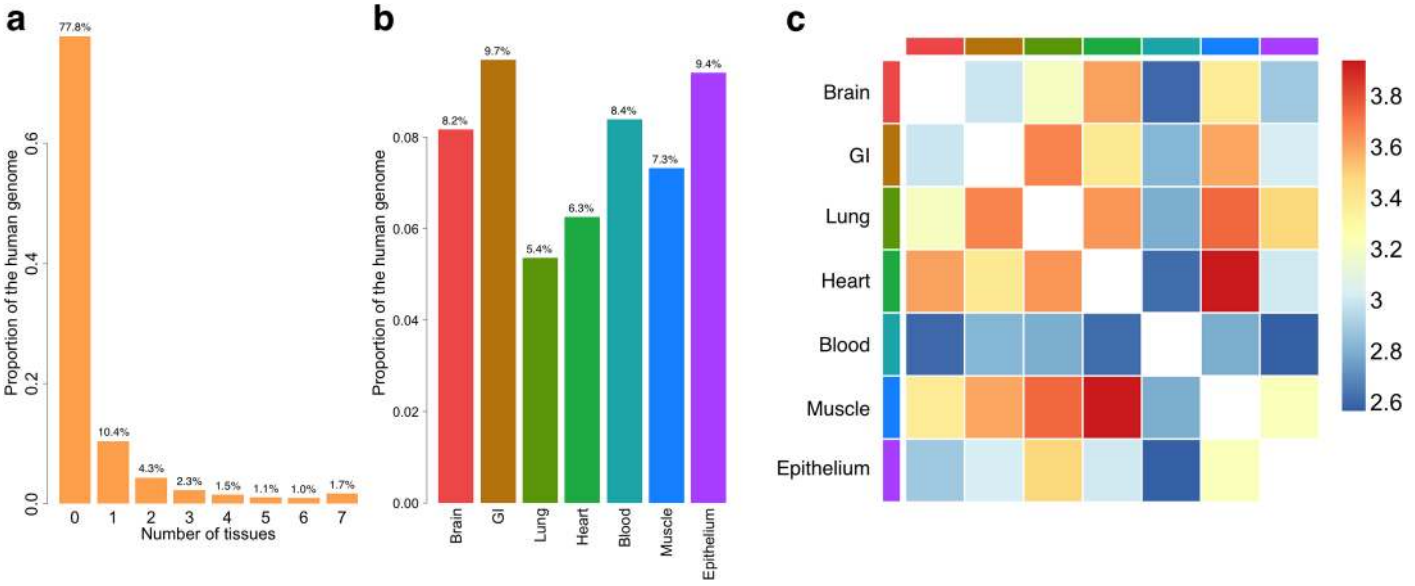

Fig 1. General characteristics of GenoSkyline annotations. (a) Number of tissues in which nucleotides are functional. (b) Proportion of functional genome for each tissue type. (c) Overlap of functional regions across seven tissue types. The scale is log odds ratio.

doi:10.1371/journal.pgen.1005947.g001

Investigating the performance of tissue-specific functional annotations

Beta-globin gene complex. We now demonstrate GenoSkyline's ability to predict tissue-specific functionality using a variety of experimentally validated functional machinery. Beta-globin (*HBB*) gene complex is an extensively studied genomic region on chromosome 11, containing 6 genes and 23 cis-regulatory modules (CRMs) that are known to control both the timing and the spatial pattern of gene expression [12,13]. We compared GS scores for different tissue types in this region. Not surprisingly, blood-specific functionality was observed (Fig 2A). Among the 6 genes in this region, adult globin genes *HBB* and *HBD*, as well as pseudogene *HBBP1* are captured well by blood-specific GS scores (S3 Table). However, embryonically expressed *HBE1*, fetally expressed *HBG1* and *HBG2*, and the CRMs that regulate these genes

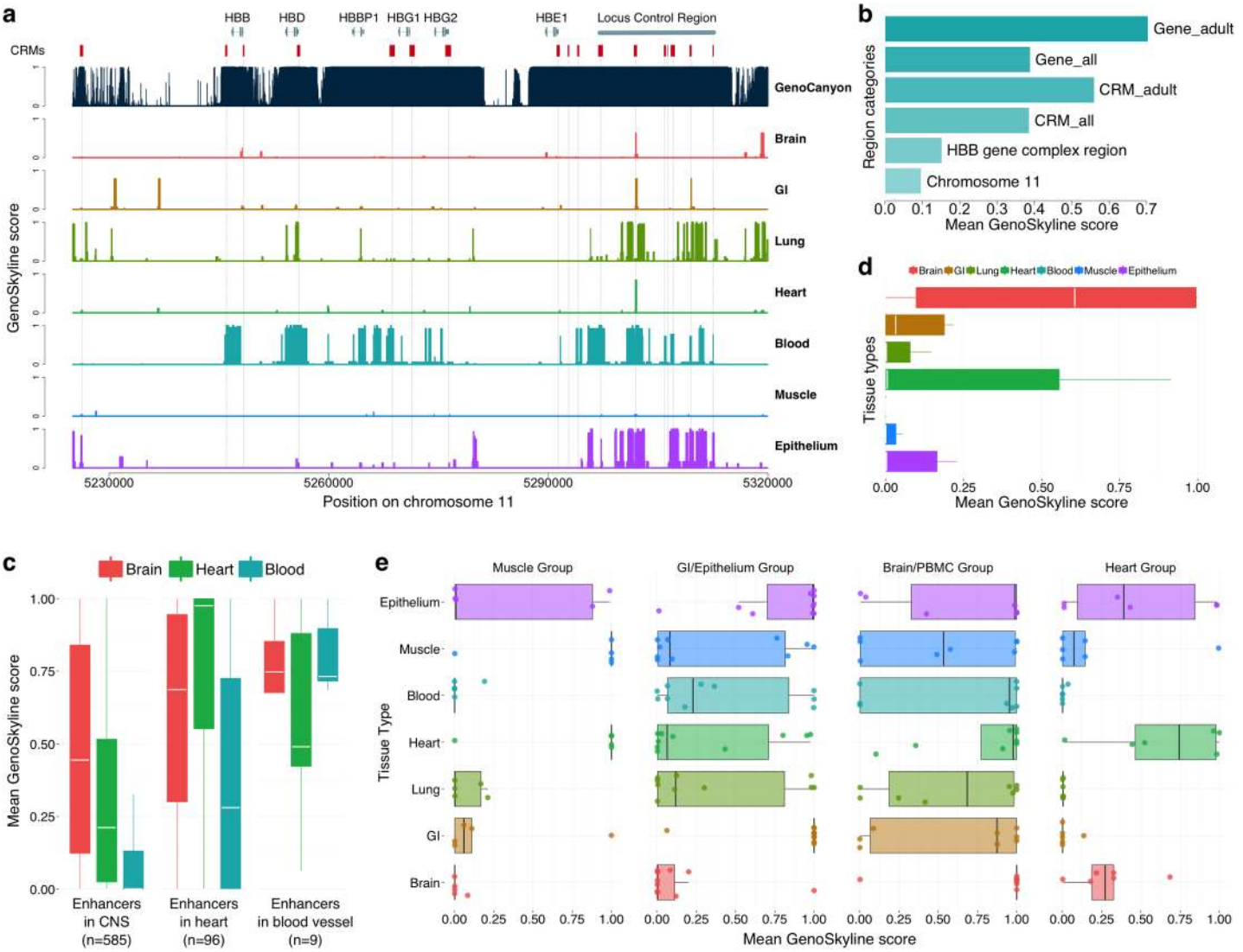

Fig 2. Case studies of *HBB* gene complex, *in vivo* enhancers, and regulatory miRNAs. (a) Comparison of GenoCanyon prediction and GenoSkyline scores for seven tissues in *HBB* gene complex region. Red boxes mark the locations of CRMs. The number of red boxes is less than 23 because some CRMs are next to each other. (b) Mean blood-specific GS score for different region categories. (c) Boxplot of mean GS scores for enhancers in CNS, heart, and blood vessel. (d) Boxplot of mean GS scores for 11 human-accelerated elements near *NPAS3*. (e) Boxplot of mean GS scores for tissue-specific regulatory miRNAs.

doi:10.1371/journal.pgen.1005947.g002

have lower GS scores. This is possibly due to the fact that 18 of the 24 cell lines used for developing blood-specific GS scores were acquired from adult samples (S1 Table). The mean blood-specific GS score in these genes increases from 0.388 to 0.704 after removing *HBG1*, *HBG2*, and *HBE1*. Similarly, a substantial boost in mean GS score is observed after removing CRMs regulating the embryonic and fetal globin genes (Figs 2B and S1). Compared with GenoCanyon, GenoSkyline provides less sensitive but highly specific functional predictions. Its ability of identifying tissue-specific functional coding and non-coding DNA elements has the potential to benefit diverse types of biological studies.

Tissue-specific enhancers. *In vivo* enhancers with tissue-specific activity in central nervous system (CNS; n = 585), heart (n = 96), and blood vessel (n = 9) were downloaded from the VISTA enhancer browser [14] (Methods). Mean GS scores for brain, heart, and blood tissues were calculated for each enhancer. Brain-specific and heart-specific GS scores were significantly higher in their respective enhancer categories compared to GS scores of non-relevant tissue types (S4 Table). Additionally, the mean blood-specific GS score also stands out for enhancers with observed activity in blood vessel despite the limited sample size (Fig 2C). In a separate study, 11 human-accelerated elements near the brain developmental transcription factor *NPAS3* have been identified to act as tissue-specific enhancers within the nervous system [15]. Brain-specific GS scores for these enhancers are substantially higher than those for other tissue types (Fig 2D and S4 Table), concurrent with previous results.

Regulatory miRNAs. Next, we test if GenoSkyline could also capture miRNAs expressed exclusively in certain tissue types. Liang et al. studied the tissue specific expression pattern of eight groups of miRNAs [16]. We extracted and annotated four groups (groups I, II, III+IVa, and V from Liang et al.) that could be represented by the currently available tissue types in GenoSkyline annotations. These four groups of miRNAs were found to be expressed preferentially in skeletal/cardiac muscle, organs lined with epithelium, brain/peripheral blood mononuclear cell (PBMC), and heart, respectively through unsupervised clustering. The most relevant tissue types suggested by GenoSkyline for these four groups are muscle/heart, GI/epithelium, brain, and heart, respectively (Fig 2E and S4 Table). Our results based on integrative analysis of epigenomic data are consistent with the tissue-specific expression pattern reported by Liang et al.

Inter-genic regulation of myosin heavy chain. We applied GenoSkyline to a validated biologic switch in cardiac development and disease. Myosin heavy chain (MHC) is the major contractile protein in human striated muscle [17]. Cardiac muscle cells primarily express two isoforms, alpha-MHC (*MYH6*) and beta-MHC (*MYH7*) [18]. The ratio of alpha-to-beta isoforms determines cardiac contractility and allows for effective response to a wide range of physiologic and pathologic stimuli [19]. Alpha-to-beta ratio decreases in cardiac diseased states [20,21], and reversal of this shift is associated with better clinical outcomes [22]. miRNAs can regulate alpha-to-beta isoform shift, and prior studies in rodents have outlined a network of crosstalk between intronically expressed miRNAs and their host muscle genes [23,24]. For instance, mir-208a, on an intron of *MYH6*, is a positive regulator of beta-MHC by targeting transcription factors that repress its expression [24]. GS scores for *MYH6* and mir-208a accurately reflect their cardiac-specific expression, whereas *MYH7* and mir-208b exhibit strong signals in both skeletal and cardiac tissue (Fig 3A and S5 Table). This corresponds to known expression pattern of *MYH7* and mir-208b in slow twitch skeletal muscle fibers [17] as well as heart. We also explored tissue-specific functionality of two known distal enhancers of mir-208b identified on VISTA Enhancer Browser, hs2330 and hs1670. GS scores for hs2330 mirror *MYH7*/mir-208b signals. Interestingly, GS scores for hs1670, a distal enhancer flanking mir-208b, are also strong in nervous and GI tissue, a finding that agrees with its observed expression pattern in other tissues (based on VISTA Enhancer Browser data). Collectively, these

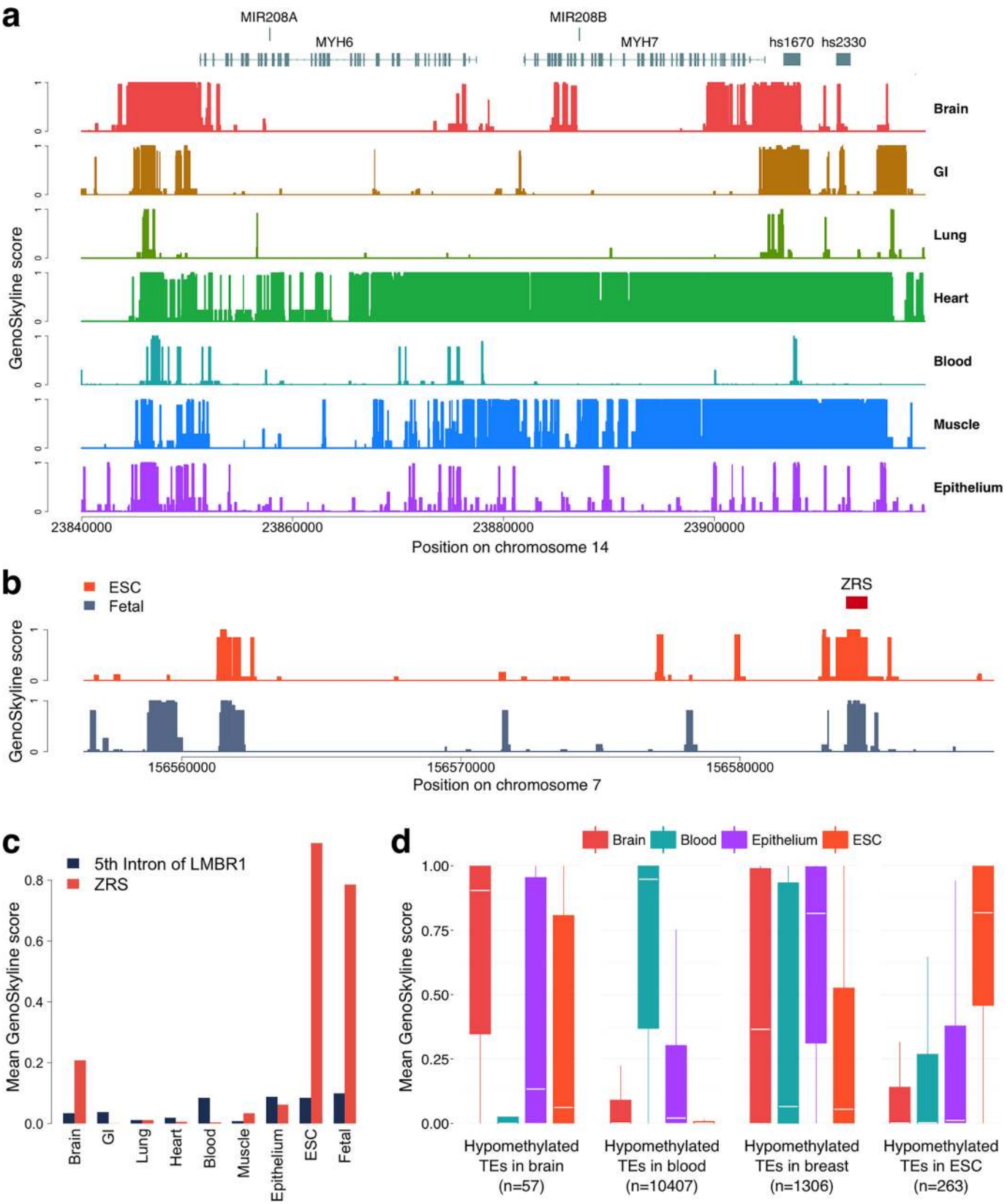
a
MIR208A
MIR208B
MYH6
MYH7
hs1670
hs2330
Brain
GI
Lung
Heart
Blood
Muscle
Epithelium
GenoSkyline score
23840000
23860000
23880000
23900000
Position on chromosome 14
b
ESC
Fetal
ZRS
GenoSkyline score
156560000
156570000
156580000
Position on chromosome 7
c
5th Intron of LMBR1
ZRS
Mean GenoSkyline score
Brain
GI
Lung
Heart
Blood
Muscle
Epithelium
ESC
Fetal
d
Brain
Blood
Epithelium
ESC
Mean GenoSkyline score
Hypomethylated TEs in brain (n=57)
Hypomethylated TEs in blood (n=10407)
Hypomethylated TEs in breast (n=1306)
Hypomethylated TEs in ESC (n=263)

Fig 3. Case studies of MHC, ZRS, and hypomethylated TEs. (a) GenoSkyline scores for seven tissues in the genomic region surrounding *MYH6* and *MYH7*. (b) ESC-specific and fetal-cell-specific GS scores for the 5th intron of *LMBR1*. The red box marks the location of ZRS. (c) Bar plot of the mean GS scores for the 5th intron of *LMBR1* and ZRS across nine tissue and cell types. (d) Boxplot of mean GS scores for four groups of hypomethylated TEs.

doi:10.1371/journal.pgen.1005947.g003

results show that GenoSkyline can replicate the tissue-specific expression pattern of a complex inter-gene regulatory network.

Zone of polarizing activity regulatory sequence. GenoSkyline can also be generalized to identify tissue specificity outside of the 7 core categories discussed here, based on available experimental data. For example, Zone of polarizing activity regulatory sequence (ZRS), a well-studied developmental enhancer, is located in the fifth intron of *LMBR1* gene. Acting as an enhancer of *SHH*, ZRS has been shown to play a crucial role in limb development [25]. However, none of the seven tissue types in GenoSkyline suggest ZRS's functionality (S2 Fig). In order to see if ZRS could be identified using epigenomic data of other cell types, we extended GenoSkyline to two new groups of cells that are potentially important for development, embryonic stem cells (ESC) and fetal cells (S6 Table). Both ESC and fetal-cell-specific GS scores successfully identified ZRS with high resolution (Fig 3B and 3C). This example shows that GenoSkyline is a flexible framework. Researchers could develop their own cell-group-specific functional annotations if ChIP-seq data are available for the cells of interest.

Hypomethylated transposable elements. A recent study of genome-wide DNA methylation status identified tissue-specific hypomethylated transposable elements (TE) exhibiting enhancer activities [26]. We downloaded four groups of TEs that are hypomethylated in ESC H1, fetal brain/primary neural progenitor cells, adult breast epithelial cells, and PBMC/adult immune cells, respectively (Methods). Although DNA methylation data were not used for developing GenoSkyline, we were still able to provide highly consistent and significant results, suggesting tissue-specific functionality of these TEs in ESC, brain, epithelium, and blood cell, respectively (Fig 3D and S4 Table).

Analyzing tissue-specific enrichment for 15 human complex traits

In the sections above, we demonstrated GenoSkyline's ability of identifying tissue-specific functional regions in the human genome. Next, we focus on how GenoSkyline could help us understand human complex traits. Finucane et al. recently proposed using LD score regression to partition heritability of complex traits by functional categories [27]. We applied LD score regression on 15 human complex diseases and traits (S7 Table), and calculated the tissue-specific enrichments using GenoSkyline annotations (Methods).

Our analysis successfully replicated some well-known findings and also provided novel insights to these complex traits (Figs 4 and S3). For schizophrenia, enrichment in brain is much stronger than in any other tissue type ($p = 3.26\times10^{-19}$), while highly significant enrichment could be observed in heart ($p = 5.02\times10^{-8}$) and blood ($p = 1.23\times10^{-6}$) as well. For three autoimmune diseases (Crohn's disease, ulcerative colitis, and rheumatoid arthritis), the strongest enrichment was in blood. However, solid enrichment in GI could also be observed for both Crohn's disease and ulcerative colitis, but not rheumatoid arthritis. Sex-stratified summary statistics were available for two anthropometric traits—body mass index (BMI) and waist-hip ratio (WHR) adjusted for BMI [28,29]. Therefore, we performed gender-specific analyses for these two traits. Consistent with recently published results [27], brain possesses the strongest enrichment for BMI. Interestingly, the enrichment in brain is stronger in female samples ($p = 1.44\times10^{-8}$) than in male samples ($p = 5.67\times10^{-7}$), while epithelial tissue may play a more important functional role in male samples ($p = 5.18\times10^{-4}$ in males and 1.17×10^{-2} in

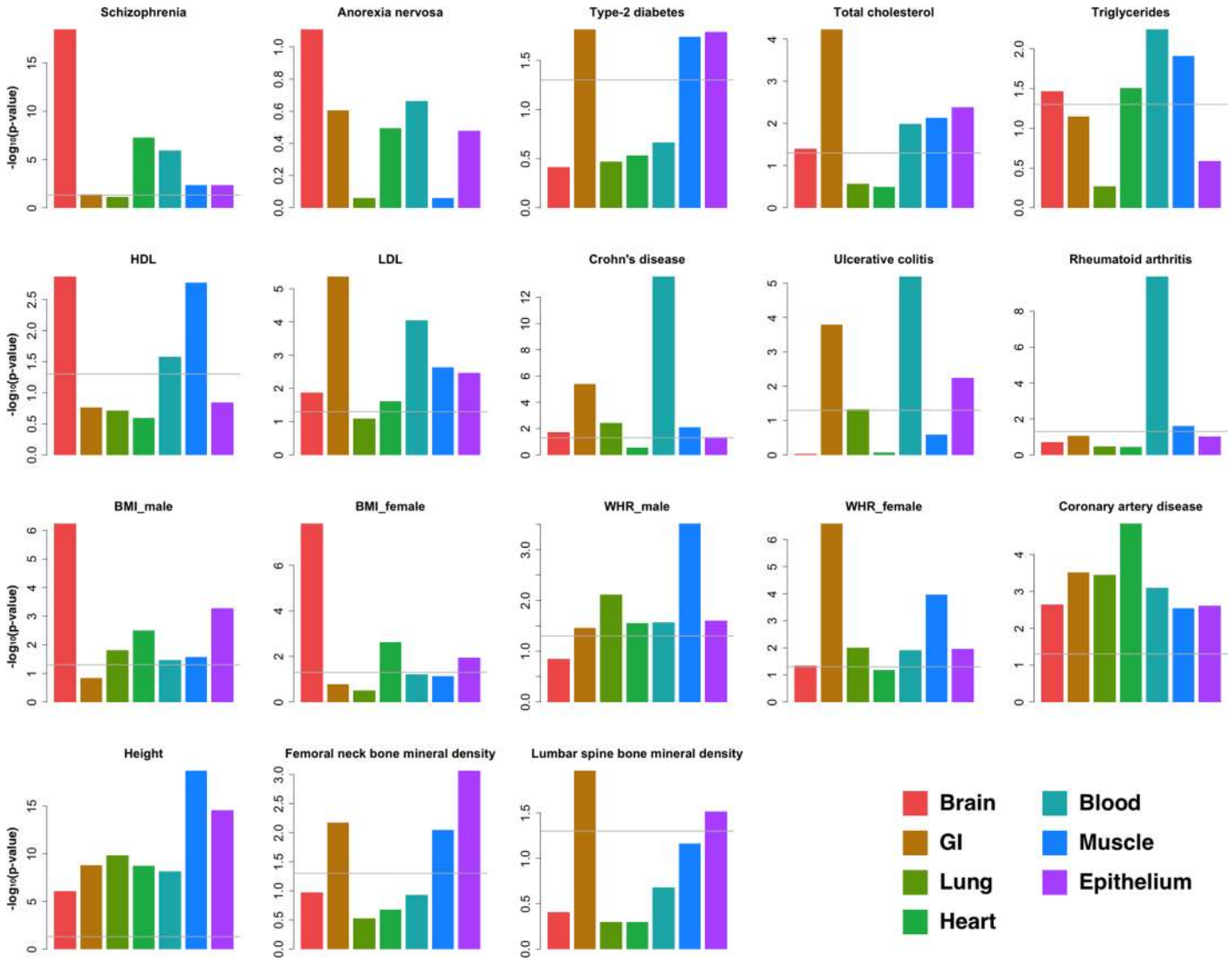

Fig 4. Tissue-specific enrichment of GWAS signals. Enrichment p-values were calculated using LD score regression. The grey line is the 0.05 cutoff for p-value.

doi:10.1371/journal.pgen.1005947.g004

females). Some patterns of gender-specific enrichment were also observed for WHR. GI is the dominant tissue for females ($p = 2.51\times10^{-7}$) but seems less important in male samples ($p = 3.44\times10^{-2}$), while enrichment in muscle is consistent between males and females.

It is worth noting that extra caution is needed when interpreting these enrichment results. For example, Finucane et al. reported connective/bone as the most enriched tissue type for human height [27], but GenoSkyline annotations for this tissue is not available at this moment due to incomplete epigenomic data (Methods). Similarly, we are not yet able to investigate the relationship between lipid traits and liver tissue because of the lack of tissue-relevant functionality data.

GWAS signal prioritization using tissue-specific functional annotations

We recently developed Genome Wide Association Prioritizer (GenoWAP), and showed that GWAS signals could be better prioritized through integrating GWAS summary statistics with GenoCanyon annotation [10]. From the results of tissue-specific enrichment analysis, it could

be seen that some complex traits are strongly related to a few tissue types. In this section, we show that the performance of GWAS signal prioritization could be further improved through integrating GenoSkyline annotations of relevant tissue types.

Using both tissue-specific GS scores and GenoCanyon scores that quantify the overall functionality, we calculate the posterior probability $P(Z_D = 1, Z_T = 1|p)$ to measure the importance of each SNP. In this calculation, Z_D is the indicator of disease/trait-specific functionality, Z_T is the indicator of tissue-specific functionality, and p is the p-value acquired from standard GWAS analysis (Methods). Psychiatric Genomics Consortium (PGC) has published two large GWAS meta-analyses for schizophrenia, a major psychiatric disorder. We applied our method to the smaller study [30] and attempted to replicate the findings of the larger study [31]. This analysis demonstrates GenoSkyline's ability to prioritize association signals that are more likely to be replicated in a larger sample. These two studies will be referred to as PGC2011 and PGC2014 studies in the following discussion.

Enrichment analysis suggests that brain is the most enriched of schizophrenia GWAS signals compared with other tissue types, and strong enrichment could also be observed in heart and blood (Fig 4). For each SNP in the PGC2011 study, mean GenoCanyon score of its surrounding region and mean GS scores of brain, blood, and heart tissues were calculated (Methods). SNPs in these tissue-specific functional regions and the SNPs in general functional regions are all enriched for associations with schizophrenia (Figs 5A and S4). Notably, tissue-specific functional regions are more enriched for associations with schizophrenia relative to general functional regions, with blood showing the strongest enrichment. It is also worth noting that non-functional regions are enriched of GWAS associations as well, most likely due to the LD between functional and non-functional SNPs [10].

Next, we define a new SNP-level metric for the tissue-specific GenoSkyline posterior (GSP) scores (i.e. $P(Z_D = 1, Z_T = 1|p)$) of brain, blood, and heart, as well as the non-specific functionality posterior (NSFP) scores (i.e. $P(Z_D = 1|p)$; see Methods) for each SNP in PGC2011 study. Enrichment analysis using GTEx whole-blood eQTLs [11] found that the top SNPs based on tissue-specific GSP scores are substantially more enriched of eQTLs than NSFP scores and p-values. As expected, blood GSP scores showed the strongest enrichment of whole-blood eQTLs (Fig 5B). When using a set of quantitative trait loci in human brain [32], tissue-specific GSP scores also showed superior performance, with the brain-specific scores dominating others as the number of top SNPs increase (Figs 5C and S5).

A total of 108 schizophrenia-associated loci were identified in the PGC2014 study. We removed three loci on chromosome X due to the absence of SNPs on sex chromosomes in the PGC2011 dataset. All the SNPs in the PGC2011 study were ranked based on their p-values, NSFP scores, and tissue-specific GSP scores, respectively (S8 Table). The maximum ranks at each of the 105 schizophrenia-associated loci based on these different criteria were then compared (S9 Table). Brain GSP score showed better performance in prioritizing these loci when compared with p-value. Sixty-seven out of 105 loci had an increased rank (p-value = 0.003, one-sided binomial test). The performance of heart GSP score was slightly worse than brain-specific score, but still better than p-value ranking. Blood GSP score showed comparable performance with p-value ranking. Notably, the performance of brain and heart GSP scores was still significantly better than NSFP score, although NSFP score outperforms ranking based on p-value.

Tissue-specific functional annotations could provide even deeper insight when prioritizing SNPs locally at risk loci. The schizophrenia-associated locus on chromosome 8q21 is located in the intergenic region upstream of *MMP16* gene (Fig 5D). The p-values in the PGC2014 study clearly suggested two signal peaks. One is located near the transcription start site of *MMP16*, while the other resides nearly 200,000 bases upstream and shows slightly stronger signal.

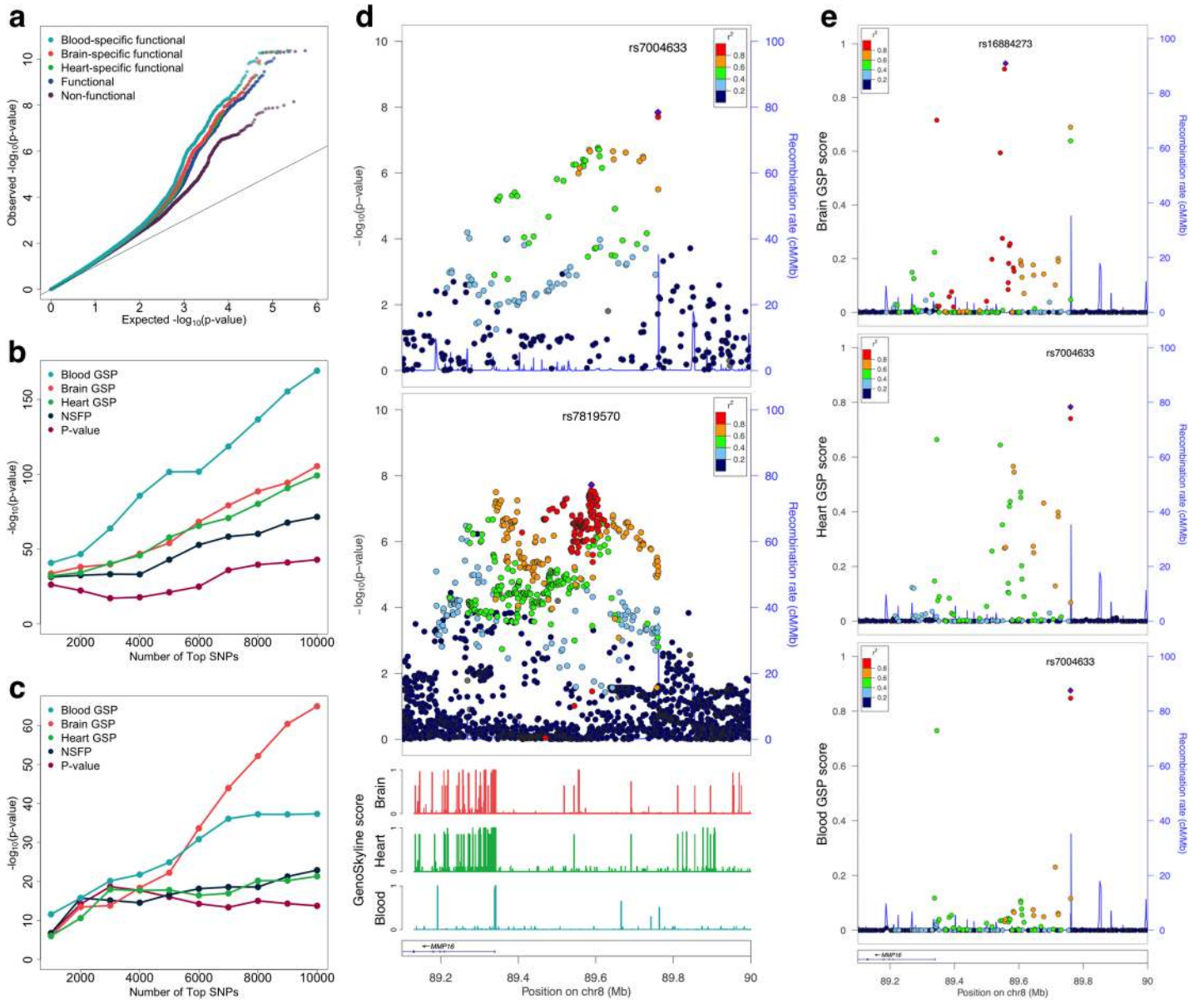

Fig 5. Prioritizing schizophrenia GWAS signals using GenoSkyline annotations. (a) Tissue-specific functional regions are more enriched of schizophrenia associations than generally functional regions and non-functional regions. (b) Enrichment of GTEx whole-blood eQTLs in top SNPs from PGC2011 study. (c) Enrichment of human brain quantitative trait loci in top SNPs from PGC2011 study. (d) Summary statistics at the schizophrenia-associated locus on chromosome 8q21 near *MMP16* gene. The top and middle panel show p-values from PGC2011 and PGC2-14 studies, respectively. The bottom panel shows GenoSkyline annotations at this locus. (e) Locus plots for tissue-specific posterior scores. From top to bottom, the three panels show posterior scores of brain, heart, and blood tissues, respectively.

doi:10.1371/journal.pgen.1005947.g005

However, the two-peak pattern was not clear in the PGC2011 study. Instead, two SNPs close to the end of the LD block near 89.8M showed the strongest signal. We compared the local predictions based on brain, heart, and blood-specific GSP scores at this locus (Fig 5E). Brain GSP scores successfully revealed the multi-peak nature at this locus, suggested the importance of the peak near 89.6Mb, and diminished the signal strength at the two SNPs near 89.8Mb, concurrent with the PGC2014 results. Although the method was applied on PGC2011 p-values, the results after prioritization matched the signal pattern in the PGC2014 study very well. Heart

GSP scores also suggested the existence of the signal peak near 89.6M. However, the posterior scores have lower values, and the overall signal pattern does not match the PGC2014 study very well. The signal peak near 89.6M was completely lost in the blood-specific results. The two SNPs near 89.8M, however, had large GSP scores. The differences across tissue types are concurrent with GS scores at this locus (Fig 5D). Upstream of *MMP16*, near 89.6M, several functional segments can be seen in brain, only one remains in heart, and none exists in blood. Through comparing the tissue-specific prioritization results with the p-values in PGC2014 study, we see that brain-specific GSP scores had the strongest signal strength, which can be quantified using the local maximum GSP score (Methods). The highly matched signal pattern also suggested that brain might be the tissue type in which this locus plays a functional role.

Further insight on risk loci associated with coronary artery disease

Next, we applied our method to another GWAS to further illustrate the biological insight that GenoSkyline can provide for understanding complex diseases. The CARDIoGRAM consortium published a large-scale GWAS meta-analysis of coronary artery disease (CAD) comprising 22,333 cases and 64,762 controls [33], in which they replicated 10 out of 12 previously reported risk loci and identified 13 new loci associated with CAD. We applied our method on the summary statistics and used the local maximum GSP score to measure the relatedness between each risk locus and different tissue types (Methods). We removed the locus on chromosome 1q41 (*MIA3*) and the locus on chromosome 6q25.3 (*LPA*) due to incomplete data in the meta-analysis stage of CARDIoGRAM study. The remaining 21 CAD-associated loci are summarized in Tables 1 and S10.

The first impression of these results is that despite the strong overall enrichment of GWAS signals (Fig 4), heart is the most relevant tissue type for only two loci. On the contrary, a substantial proportion of risk loci (9 out of 21) seem to be functional in the GI tissue. Interestingly, GI was the most enriched tissue type for several known risk factors for CAD including LDL and total cholesterol (Fig 4). These results suggest not only the larger effect sizes of CAD-associated loci in the gastrointestinal system, but also a substantial amount of undetected heart-related loci. Furthermore, brain was the least enriched tissue type for CAD GWAS signals, but the risk locus on chromosome 14q32.2 near *HHIPL1* and *CYP46A1* was predicted to be functional in brain. In fact, the *CYP46A1* gene encodes for Cholesterol 24-hydroxylase that is present mainly in brain, where it converts cholesterol from degraded neurons into 24S-hydroxychoelesterol [33,34]. This process is crucial for eliminating cholesterol from the brain since cholesterol is usually unable to pass the blood-brain barrier [35].

A larger GWAS for CAD was published during the preparation of this manuscript [36]. This large study may be used to validate the performance of our approach when its summary statistics become publicly available in the future.

Discussion

In this paper, we introduced GenoSkyline, an integrative framework for predicting tissue-specific functional regions in the human genome. Through integrating GenoSkyline annotations with GWAS summary statistics, we illustrated a variety of ways that GenoSkyline could help researchers understand human complex diseases and traits. We also showed that the GenoSkyline framework is customizable so that researchers can develop their own functional annotations for a selected group of cells. Imputed epigenetic data could also be directly used as the model input when needed. As epigenomic ChIP-seq data become available for an increasing number of cell types in the future, GenoSkyline's ability to facilitate studies of complex disease will be further enhanced.

Table 1. Risk loci for coronary artery disease.

Chr	Start[a]	Stop[a]	Genes in region	Tissue type[b]	Posterior score[c]
1p13.3	109,700,000	109,900,000	*SORT1*	GI	0.99991
2q33.1	203,600,000	204,000,000	*WDR12*	GI	0.99999
3q22.3	137,900,000	138,200,000	*MRAS*	Heart	0.99625
6p24.1	12,700,000	13,100,000	*PHACTR1*	GI	0.99997
9p21.3	21,900,000	22,200,000	*CDKN2A*, *CDKN2B*	NA[d]	NA[d]
10q11.21	44,400,000	44,900,000	*CXCL12*	GI	0.83314
19p13.2	11,000,000	11,400,000	*LDLR*	Blood	0.99967
21q22.11	35,500,000	35,700,000	*MRPS6*	Lung	0.9993
1p32.2	56,900,000	57,100,000	*PPAP2B*	GI	0.99699
6p21.31	34,600,000	35,300,000	*ANKS1A*	Heart	0.95156
6q23.2	134,000,000	134,300,000	*TCF21*	GI	0.99998
7q32.2	129,600,000	129,900,000	*ZC3HC1*	Muscle	0.9995
9q34.2	136,000,000	136,400,000	*ABO*	GI	0.99531
10q24.32	104,400,000	105,000,000	*CYP17A1*, *CNNM2*, *NT5C2*	GI	0.94966
11q23.3	116,500,000	116,700,000	*ZNF259*, *APOA5-A4-C3-A1*	Blood	0.9997
13q34	110,700,000	111,200,000	*COL4A1*, *COL4A2*	Muscle	0.99704
14q32.2	100,000,000	100,300,000	*HHIPL1*, *CYP46A1*	Brain	0.97293
15q25.1	78,900,000	79,200,000	*ADAMTS7*	Muscle	0.99934
17p13.3	2,000,000	2,300,000	*SMG6*, *SRR*	GI	0.98812
17p11.2	17,400,000	18,000,000	*RASD1*, *SMCR3*, *PEMT*	Blood	0.96864
17q21.32	46,800,000	47,200,000	*UBE2Z*, *GIP*, *ATP5G1*, *SNF8*	Blood	0.96206

[a]These coordinates are the roughly estimated boundaries of risk loci. The inference of relevant tissues is not sensitive to the boundary coordinates.

[b]The tissue type that provides the largest local maximum posterior score.

[c]Local maximum GSP score for the most relevant tissue type.

[d]Not applicable due to ties. See S9 Table.

doi:10.1371/journal.pgen.1005947.t001

Our approach is not without limitation. First, the annotation results are incomplete due to currently unavailable tissue types, and as a result, the GWAS enrichment results may not be comprehensive (e.g. liver may also be highly related to CAD, but there is no complete annotation data from liver yet). Second, some risk loci (or independent functional segments at the same locus) may play active roles in multiple tissue types. For example, in our PGC GWAS analysis, although local maximum GSP scores suggest that brain may be more relevant with the risk locus upstream of *MMP16*, two SNPs near 89.8MB are located near several functional segments in blood. Whether these SNPs can be functionally linked to schizophrenia remains to be investigated. Moreover, we emphasize that our method identifies regions of likely functionality, but does not provide conclusive proof of functionality for any individual SNP or locus. That said, our method still provides a simple and intuitive summary statistic that measures the relatedness between risk loci and sets of functionally related tissues. It has great potential to become a standard step in downstream GWAS analysis to help researchers generate new hypotheses regarding the etiology behind each risk locus.

The increasing accessibility of GWAS summary statistic datasets, coupled with the method's independence from requiring individual-level genotype and phenotype, make Genoskyline tissue-specific prioritization useful and easy to implement. Moreover, GWAS signal integration is just one way to utilize GenoSkyline annotations. Its nucleotide-level functional prediction based on unsupervised learning and the good predictive performance in non-coding regions

promise a potential role in many fields of genomics, such as next-generation sequencing studies and understanding somatic mutations. GenoSkyline scores of seven tissue types and two additional cell types have been pre-calculated for the entire human genome and can be readily downloaded. Source code is available for all major OSes and can be accessed at (http://genocanyon.med.yale.edu/GenoSkyline). As far as we are aware of, our work is the first principled method in the literature that is capable of integrating vast amounts of information from tissue-specific functional annotations to prioritize and interpret GWAS results. We believe that GenoSkyline and its applications can guide genetics research at multiple resolutions and greatly benefit the broader scientific community.

Methods

Consolidated epigenomes

Epigenetic data were selected from the Epigenomics Roadmap Project's 111 consolidated reference epigenomes database [3] (http://egg2.wustl.edu/roadmap/) based on anatomy type and mark availability. Each tissue type is a clustering of relevant samples in order to contain at least one of each of the following: H3k4me1, H3k4me3, H3k36me3, H3k27me3, H3k9me3, H3k27ac, H3k9ac, and DNase I Hypersensitivity. Samples are reduced to a per-nucleotide binary encoding of presence or absence of narrow contiguous regions of ChIP-seq signal enrichment compared to input (Poisson p-value threshold of 0.01), and a union of all tissue-specific samples for that mark is taken. The set of 8 marks was chosen due to the well-understood, localized regulatory interactions of histone marks [37] and DNase I [38]. We created seven unique tissue clusters (S1 Table) and two additional cell type clusters (S6 Table) based on these annotations to represent common, physiologically-related organ systems. To reflect actual tissue-specific epigenetic behavior, a majority of samples chosen are primary tissues and cultures, and inclusion of immortalized cell lines has been kept to a minimum.

GenoSkyline model and estimation

Lu et al. previously proposed a method that applies unsupervised-learning techniques on genomic annotations to predict the functional potential of a genomic region [4]. Given a set of annotations $\boldsymbol{A}$, we assume the joint distribution of $\boldsymbol{A}$ along the genome to be a mixture of annotations at locations with no functionality, i.e. $f(\boldsymbol{A} \mid Z = 0)$, and annotations at locations that are functional, i.e. $f(\boldsymbol{A} \mid Z = 1)$. We assume that each annotation in $\boldsymbol{A}$ is conditionally independent given Z, allowing the conditional joint density of $\boldsymbol{A}$ given Z to be factorized as

$$f(\boldsymbol{A}|Z = c) = \prod_{i=1}^{8} f_i(A_i|Z = c), \qquad c = 0, 1 \tag{1}$$

Since all annotations used are binary classifiers, the Bernoulli distribution was used to model the marginal functional likelihood given each individual annotation.

$$f_i(A_i|Z = c) = p_{ic}^{A_i}(1 - p_{ic})^{1-A_i}, \qquad i = 1, \ldots, 8; \;\; c = 0, 1 \tag{2}$$

Assuming a prior probability π of being functional ($\pi = P(Z = 1)$), we can estimate the parameter p_{ic} of each annotation with the Expectation-Maximization (EM) algorithm, and calculate the posterior probability at a given genomic coordinate, referred to as the GS score.

$$\begin{aligned} P(Z = 1|\boldsymbol{A}) &= \frac{\pi f(\boldsymbol{A}|Z = 1)}{\pi f(\boldsymbol{A}|Z = 1) + (1 - \pi) f(\boldsymbol{A}|Z = 0)} \\ &= \frac{\pi \prod_{i=1}^{8} f_i(A_i|Z = 1)}{\pi \prod_{i=1}^{8} f_i(A_i|Z = 1) + (1 - \pi) \prod_{i=1}^{8} f_i(A_i|Z = 0)} \end{aligned} \tag{3}$$

We must estimate 17 parameters for each tissue tract.

$$\boldsymbol{\Theta} = (\pi, \boldsymbol{P}_1, \boldsymbol{P}_0) \tag{4}$$

Where

$$\boldsymbol{P}_c = (p_{1c}, p_{2c}, \ \cdots \ , p_{8c}), \quad c = 0, 1 \tag{5}$$

Parameters were estimated using the GWAS Catalog [6], downloaded from the NHGRI website (http://www.genome.gov/gwastudies/), which at the time of download, contained 13,070 unique SNPs found to be significant in at least one published GWAS. These SNPs were expanded into 1k bp intervals, and formed a genome sampling covering 12,801,840 bp of the genome. While significant SNP associations are likely to tag the effects of nearby functional elements, the size and distance of these functional elements varies for each individual SNP. As a result, the total sampling serves as an effective and robust representation of functional and non-functional regions along the genome (S1 Text).

Case studies of experimentally validated functional machinery

VISTA enhancers [14] were downloaded from the VISTA Enhancer Browser (http://enhancer.lbl.gov/), where enhancers with E11.5 reporter staining experimental data were selected. Brain enhancers were selected based on staining results identifying any CNS-related tissues (neural tube, cranial nerve, hindbrain, mesenchyme derived from neural crest, trigeminal V, forebrain, and midbrain). Heart enhancers were enhancers identified for positive reporter results in the heart region of E11.5 mouse reporter assays. Blood vessels enhancers were identified by selecting for "blood vessels" expression pattern. Hypomethylated TE loci in H1ES, brain, breast, and blood were downloaded from http://epigenome.wustl.edu/TE_Methylation/. All genomic coordinates were converted to genome build hg19.

SNP prioritization using tissue-specific functional annotation

We identify three disjoint cases for a given GWAS SNP:

1. The SNP is in a genomic region that is functional for the given phenotype and tissue ($Z_D = 1, Z_T = 1$).
2. The SNP is in a genomic region that is functional for the given tissue, but that region has no functionality in the phenotype ($Z_D = 0, Z_T = 1$).
3. The SNP is in a genomic region that is not functional in the given tissue ($Z_T = 0$).

A useful metric for prioritizing SNPs is the conditional probability that the SNP is classified under case-I given its p-value in a given GWAS study, i.e. $P(Z_D = 1, Z_T = 1 | p)$. We can calculate this probability by employing Bayes formula and considering all three cases as follows:

$$P(Z_D = 1, Z_T = 1|p) = \frac{f(p|Z_D = 1, Z_T = 1) \times P(Z_D = 1, Z_T = 1)}{f(p|Z_D = 1, Z_T = 1) \times P(Z_D = 1, Z_T = 1) + f(p|Z_D = 0, Z_T = 1) \times P(Z_D = 0, Z_T = 1) + f(p|Z_T = 0) \times P(Z_T = 0)} \tag{6}$$

First, the case in which $Z_T = 0$ can be directly identified by assigning each SNP a prior probability of tissue-specific functionality (i.e. $P(Z_T = 1)$) defined as the average GS score of its surrounding 10,000 base pairs for that tissue (S1 Text). We partition all the SNPs into two subgroups based on a mean GS score threshold of 0.1. Notably, these probabilities take on a bimodal distribution and are not sensitive to changing threshold [10]. In this way, we can use these partitions to directly estimate $f(p|Z_T = 0)$ by applying density estimation techniques on

the SNP subgroup with low GS scores. More specifically, we apply a histogram approach for density estimation and use cross validation to choose the optimal number of bins.

Second, we estimate the p-value density of our second case, where $Z_D = 0$ and $Z_T = 1$. We can intuitively assume that, regardless of local LD structure, SNPs that are functional in a tissue but not relevant to the phenotype will have similar p-value behavior to all other SNPs that are not relevant to the phenotype, which in turn behave similarly to SNPs that are not functional at all (S1 Text). More formally, we can describe this relationship as follows:

$$f(p|Z_D = 0, Z_T = 1) = f(p|Z_D = 0) = f(p|Z = 0) \tag{7}$$

We can effectively estimate $f(p|Z = 0)$ by using a similar approach to estimating $f(p|Z_T = 0)$, but partitioning SNPs using the general functionality GenoCanyon score instead of tissue-specific GS score.

Next, we consider the following formulas.

$$P(Z_D = 1, Z_T = 1) = P(Z_D = 1|Z_T = 1) \times P(Z_T = 1) \tag{8}$$

$$P(Z_D = 0, Z_T = 1) = P(Z_D = 0|Z_T = 1) \times P(Z_T = 1) \tag{9}$$

The prior probability $P(Z_T = 1)$ can be calculated directly from GS scores as stated above, but the conditional probabilities of disease-specific functionality given tissue-specific functionality remains to be estimated.

Finally, we estimate all the remaining terms in formula 6 using the EM algorithm. In the first step of the estimation procedure, we acquired the subset of SNPs located in tissue-specific functional regions. The p-value distribution of these SNPs is the following mixture.

$$f(p|Z_T = 1) = P(Z_D = 1|Z_T = 1) \times f(p|Z_D = 1, Z_T = 1) + P(Z_D = 0|Z_T = 1) \times f(p|Z_D = 0, Z_T = 1) \tag{10}$$

Density $f(p|Z_D = 0, Z_T = 1)$ has been estimated in earlier steps. Applying the findings of Chung et al., we assume a beta distribution of the p-values of functional SNPs (i.e. $f(p|Z_D = 1, Z_T = 1)$) as a reasonable approximation under general assumptions of SNP effect size [9].

$$(p|Z_D = 1, Z_T = 1) \sim Beta(\alpha, 1), \qquad 0 < \alpha < 1 \tag{11}$$

The EM algorithm is then applied to the SNP subset located in tissue-specific functional regions. The beta assumption guarantees a closed-form expression in each iteration and all the remaining parameters can be subsequently estimated (S1 Text, S11 Table). We now have all the necessary terms for Eq 6, and define this as our posterior probability of tissue-specific and disease-specific functionality (GSP score). The feature of integrating tissue-specific functional annotations to prioritize GWAS signals has been added to the GenoWAP software available on our server (http://genocanyon.med.yale.edu/GenoSkyline).

SNP prioritization using GenoCanyon annotation

Non-tissue-specific GenoCanyon scores are assigned to GWAS signals using GenoWAP [10]. Briefly, GenoWAP calculates the posterior score $P(Z_D = 1|p)$ using a simpler model for functionality.

$$P(Z_D = 1|p) = \frac{f(p|Z_D = 1) \times P(Z_D = 1)}{f(p|Z_D = 1) \times P(Z_D = 1) + f(p|Z_D = 0) \times P(Z_D = 0)} \tag{12}$$

This conditional probability can be calculated similarly to GS scores, making use of Eq (7) to empirically estimate $f(p|Z_D = 0)$, a beta distribution on partitioned Genocanyon scores

(calculated with 22 non-tissue-specific functionality annotations[4]) to estimate $f(p|Z_D = 1)$, and the EM algorithm on the functional marker p-value density to calculate $P(Z_D = 1)$ as described in Lu *et al*. These are referred to in the results as the NSFP scores to which GenoSkyline SNP prioritization is compared.

Calculating tissue-specific enrichment using LD score regression

Enrichment of GWAS signals in GenoSkyline tissue-specific annotations was calculated using LD score regression [27]. First, stratified LD scores were estimated using GS scores with cutoff 0.5, 1000 Genomes data of European ancestry [39], and a 1-centiMorgan (cM) window. Then the GenoSkyline annotations were analyzed together with the 53 baseline annotations of LD score regression. For each tissue-specific annotation, partitioned heritability was estimated and enrichment was calculated as the ratio of the proportion of explained heritability and the proportion of SNPs in each annotated category [27].

Measuring relevant tissue types for GWAS risk loci

A large GSP score is obtained if the p-value for the SNP is small and the SNP is located in a highly functional region for the tissue type under investigation. Therefore, the maximal GSP score at a risk locus effectively measures how well the p-values match the pattern of GenoSkyline annotations, thereby measuring the relatedness between the GWAS locus and different tissue types. For each tissue, the maximal GSP score is acquired at the risk locus of interest. These scores are then compared across tissue types. The largest score is referred to as local maximum GSP score, and the corresponding tissue type is predicted to be the most relevant tissue.

Bioinformatics tools

Locus plots were generated using LocusZoom [40]. The "ggbio" R package [41] was used to plot genes. The "bigmemory" R package [42] was used to access and handle large datasets.

Supporting Information

S1 Text. Supplementary notes to the GenoSkyline model.
(PDF)

S1 Fig. Mean GS score for different region categories in *HBB* gene complex across seven tissue types. For each tissue, the four bars from left to right indicate all 23 CRMs, adult CRMs, all genes, and adult globins, respectively.
(TIFF)

S2 Fig. GenoCanyon score and GenoSkyline scores for seven tissues in the 5th intron of *LMBR1*. The red box marks the location of ZRS.
(TIFF)

S3 Fig. Tissue-specific fold enrichment of GWAS signals.
(TIFF)

S4 Fig. Histograms of p-values for SNPs located in non-functional, functional, and tissue-specific functional regions.
(TIFF)

S5 Fig. Fold enrichment of eQTLs in top SNPs from PGC2011 study. (a) GTEx whole-blood eQTLs. (b) Human brain quantitative trait loci.
(TIFF)

S6 Fig. Distribution of GenoSkyline scores on chromosome 22.
(TIFF)

S7 Fig. Blood-specific annotations in the *HBB* region after removing one mark from the model.
(TIFF)

S8 Fig. Blood-specific annotations in the *HBB* region based on various collections of epigenetic marks.
(TIFF)

S9 Fig. Comparison of LD score densities on chromosome 22 across different SNP categories.
(TIFF)

S1 Table. Cell types used for developing GenoSkyline annotations of seven tissue types.
(XLSX)

S2 Table. Proportion of functional genome across seven tissue types under GS score cutoff 0.5.
(XLSX)

S3 Table. Mean GS scores for functional elements in the *HBB* gene complex.
(XLSX)

S4 Table. Quantitative comparison of GS scores in several case studies.
(XLSX)

S5 Table. Mean GS scores for functional elements near *MYH6* and *MYH7*.
(XLSX)

S6 Table. Cell types used for developing GenoSkyline annotations of ESC and fetal cells.
(XLSX)

S7 Table. List of 15 complex diseases and traits.
(XLSX)

S8 Table. Ranks of top signals under different criteria at 105 schizophrenia-associated loci.
(XLSX)

S9 Table. Ranking performance comparison. The value in each cell is the p-value acquired from one-sided binomial test.
(XLSX)

S10 Table. Largest GSP scores for each tissue type at CAD-associated risk loci.
(XLSX)

S11 Table. GenoSkyline parameter estimates for brain tissue.
(XLSX)

S12 Table. Comparison of GWAS signal enrichment in annotations based on promoter- and enhancer- associated marks.
(XLSX)

S13 Table. GWAS signal enrichment in multiple annotated categories.
(XLSX)

Author Contributions

Conceived and designed the experiments: QL RLP HZ. Analyzed the data: QL RLP QW BJH. Wrote the paper: QL RLP HZ. Advised on statistical and genetic issues: HZ.

References

1. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476–482. doi: 10.1038/nature10530 PMID: 21993624
2. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489: 57–74. doi: 10.1038/nature11247 PMID: 22955616
3. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317–330. doi: 10.1038/nature14248 PMID: 25693563
4. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K- H, et al. (2015) A Statistical Framework to Predict Functional Non-Coding Regions in the Human Genome Through Integrated Analysis of Annotation Data. Sci Rep 5.
5. Efron B (2010) Large-scale inference: empirical Bayes methods for estimation, testing, and prediction: Cambridge University Press.
6. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106: 9362–9367. doi: 10.1073/pnas.0903103106 PMID: 19474294
7. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, et al. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies.
8. Pickrell JK (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. The American Journal of Human Genetics 94: 559–573. doi: 10.1016/j.ajhg.2014.03.004 PMID: 24702953
9. Chung D, Yang C, Li C, Gelernter J, Zhao H (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation.
10. Lu Q, Yao X, Hu Y, Zhao H (2015) GenoWAP: GWAS Signal Prioritization Through Integrated Analysis of Genomic Functional Annotation. Bioinformatics.
11. Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, et al. (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348: 648–660. doi: 10.1126/science.1262110 PMID: 25954001
12. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, et al. (2014) Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A.
13. King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, et al. (2005) Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res 15: 1051–1060. PMID: 16024817
14. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499–502. PMID: 17086198
15. Kamm GB, Pisciottano F, Kliger R, Franchini LF (2013) The developmental brain gene NPAS3 contains the largest number of accelerated regulatory sequences in the human genome. Mol Biol Evol 30: 1088–1102. doi: 10.1093/molbev/mst023 PMID: 23408798
16. Liang Y, Ridzon D, Wong L, Chen C (2007) Characterization of microRNA expression profiles in normal human tissues. BMC Genomics 8: 166. PMID: 17565689
17. Jandreski MA, Sole MJ, Liew CC (1987) Two different forms of beta myosin heavy chain are expressed in human striated muscle. Hum Genet 77: 127–131. PMID: 3653886
18. Gorza L, Mercadier JJ, Schwartz K, Thornell LE, Sartore S, et al. (1984) Myosin types in the human heart. An immunofluorescence study of normal and hypertrophied atrial and ventricular myocardium. Circ Res 54: 694–702. PMID: 6234108
19. Baldwin KM, Haddad F (2001) Effects of different activity and inactivity paradigms on myosin heavy chain gene expression in striated muscle. J Appl Physiol (1985) 90: 345–357.
20. Gupta MP (2007) Factors controlling cardiac myosin-isoform shift during hypertrophy and heart failure. J Mol Cell Cardiol 43: 388–403. PMID: 17720186
21. Miyata S, Minobe W, Bristow MR, Leinwand LA (2000) Myosin heavy chain isoform expression in the failing and nonfailing human heart. Circ Res 86: 386–390. PMID: 10700442
22. Lowes BD, Gilbert EM, Abraham WT, Minobe WA, Larrabee P, et al. (2002) Myocardial gene expression in dilated cardiomyopathy treated with beta-blocking agents. N Engl J Med 346: 1357–1365. PMID: 11986409

23. van Rooij E, Quiat D, Johnson BA, Sutherland LB, Qi X, et al. (2009) A family of microRNAs encoded by myosin genes governs myosin expression and muscle performance. Dev Cell 17: 662–673. doi: 10.1016/j.devcel.2009.10.013 PMID: 19922871

24. Callis TE, Pandya K, Seok HY, Tang RH, Tatsuguchi M, et al. (2009) MicroRNA-208a is a regulator of cardiac hypertrophy and conduction in mice. J Clin Invest 119: 2772–2786. doi: 10.1172/JCI36154 PMID: 19726871

25. VanderMeer JE, Ahituv N (2011) cis-regulatory mutations are a genetic cause of human limb malformations. Dev Dyn 240: 920–930. doi: 10.1002/dvdy.22535 PMID: 21509892

26. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, et al. (2013) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat Genet 45: 836–841. doi: 10.1038/ng.2649 PMID: 23708189

27. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, et al. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics.

28. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. Nature 518: 197–206. doi: 10.1038/nature14177 PMID: 25673413

29. Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, et al. (2015) New genetic loci link adipose and insulin biology to body fat distribution. Nature 518: 187–196. doi: 10.1038/nature14132 PMID: 25673412

30. Consortium SPG-WAS (2011) Genome-wide association study identifies five new schizophrenia loci. Nature genetics 43: 969–976. doi: 10.1038/ng.940 PMID: 21926974

31. Consortium SWGotPG (2014) Biological insights from 108 schizophrenia-associated genetic loci. Nature 511: 421–427. doi: 10.1038/nature13595 PMID: 25056061

32. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, et al. (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet 6: e1000952. doi: 10.1371/journal.pgen.1000952 PMID: 20485568

33. Schunkert H, König IR, Kathiresan S, Reilly MP, Assimes TL, et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nature genetics 43: 333–338. doi: 10.1038/ng.784 PMID: 21378990

34. Pikuleva IA (2006) Cytochrome P450s and cholesterol homeostasis. Pharmacol Ther 112: 761–773. PMID: 16872679

35. Russell DW (2003) The enzymes, regulation, and genetics of bile acid synthesis. Annu Rev Biochem 72: 137–174. PMID: 12543708

36. Consortium CD (2015) A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nature Genetics.

37. Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. Cell research 21: 381–395. doi: 10.1038/cr.2011.22 PMID: 21321607

38. Crawford GE, Holt IE, Mullikin JC, Tai D, Green ED, et al. (2004) Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proceedings of the National Academy of Sciences of the United States of America 101: 992–997. PMID: 14732688

39. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, et al. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491: 56–65. doi: 10.1038/nature11632 PMID: 23128226

40. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, et al. (2010) LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337. doi: 10.1093/bioinformatics/btq419 PMID: 20634204

41. Yin T, Cook D, Lawrence M (2012) ggbio: an R package for extending the grammar of graphics for genomic data. Genome Biol 13: R77. doi: 10.1186/gb-2012-13-8-r77 PMID: 22937822

42. Kane MJ, Emerson JW, Weston S (2013) Scalable Strategies for Computing with Massive Data. Journal of Statistical Software 55: 1–19.