nature
genetics

Partitioning heritability by functional annotation using genome-wide association summary statistics

Hilary K Finucane[1,2,19], Brendan Bulik-Sullivan[3,4,19], Alexander Gusev[2], Gosia Trynka[5–9], Yakir Reshef[10], Po-Ru Loh[2], Verneri Anttila[3,4,8], Han Xu[11], Chongzhi Zang[11], Kyle Farh[3,12], Stephan Ripke[3,4], Felix R Day[13], ReproGen Consortium[14], Schizophrenia Working Group of the Psychiatric Genomics Consortium[14], The RACI Consortium[14], Shaun Purcell[5,6,15], Eli Stahl[15], Sara Lindstrom[2], John R B Perry[13], Yukinori Okada[16,17], Soumya Raychaudhuri[5–8,18], Mark J Daly[3,4,8], Nick Patterson[8], Benjamin M Neale[3,4,8,20] & Alkes L Price[2,8,20]

Recent work has demonstrated that some functional categories of the genome contribute disproportionately to the heritability of complex diseases. Here we analyze a broad set of functional elements, including cell type–specific elements, to estimate their polygenic contributions to heritability in genome-wide association studies (GWAS) of 17 complex diseases and traits with an average sample size of 73,599. To enable this analysis, we introduce a new method, stratified LD score regression, for partitioning heritability from GWAS summary statistics while accounting for linked markers. This new method is computationally tractable at very large sample sizes and leverages genome-wide information. Our findings include a large enrichment of heritability in conserved regions across many traits, a very large immunological disease–specific enrichment of heritability in FANTOM5 enhancers and many cell type–specific enrichments, including significant enrichment of central nervous system cell types in the heritability of body mass index, age at menarche, educational attainment and smoking behavior.

In GWAS of complex traits, much of the heritability lies in SNPs with associations that do not reach genome-wide significance at current sample sizes[1,2]. However, many current approaches that leverage functional information[3,4] and GWAS data to inform disease biology use only SNPs in genome-wide significant loci[5–8], assume only one causal SNP per locus[9] or do not account for linkage disequilibrium (LD)[10]. We aim to improve power by estimating the proportion of genome-wide SNP heritability[1] attributable to various functional categories, using information from all SNPs and explicitly modeling LD.

Previous work on partitioning SNP heritability has used restricted maximum likelihood (REML) as implemented in GCTA[1,11–14]. REML requires individual genotypes, but many of the largest GWAS analyses are conducted through meta-analysis of study-specific results, and thus only summary statistics, not individual genotypes, are typically available for these studies. Even when individual genotypes are available, using REML to analyze multiple functional categories becomes computationally intractable at sample sizes in the tens of thousands. Here we introduce a method for partitioning heritability, stratified LD score regression, that requires only GWAS summary statistics and LD information from an external reference panel with ancestry matching the population studied in the GWAS.

We apply our new approach to 17 complex diseases and traits with an average sample size of 73,599. We first analyze annotations that are not cell type specific and find heritability enrichment in many of these functional annotations, including a large enrichment in conserved regions across many traits and a very large immunological disease–specific enrichment of heritability in FANTOM5 enhancers. We then analyze cell type–specific annotations and identify many cell type–specific enrichments of heritability, including enrichment of central nervous system (CNS) cell types in body mass index (BMI), age at menarche, educational attainment and smoking behavior.

RESULTS

Overview of the methods

Our method for partitioning heritability from summary statistics, called stratified LD score regression, relies on the fact that the χ^2 association statistic for a given SNP includes the effects of all SNPs tagged by this SNP[15,16]. Thus, for a polygenic trait, SNPs with a high LD score will have higher χ^2 statistics on average than SNPs with a low LD score[16]. This phenomenon might be driven either by the higher likelihood of these SNPs tagging an individual large effect or their ability to tag multiple weak effects. If we partition SNPs into functional categories with different contributions to heritability, then LD to a category that is enriched for heritability will increase the χ^2 statistic of a SNP more than LD to a category that does not contribute to heritability. Thus, our method determines that a category of SNPs is enriched for heritability if SNPs with high LD to that category have higher χ^2 statistics than SNPs with low LD to that category.

More precisely, under a polygenic model[1], the expected χ^2 statistic of SNP j is

$$E\left[\chi_j^2\right] = N\sum_C \tau_C \ell(j,C) + Na + 1 \qquad (1)$$

A full list of affiliations appears at the end of the paper.

Received 2 February; accepted 21 August; published online 28 September 2015; doi:10.1038/ng.3404

where N is sample size, C indexes categories, $\ell(j,C)$ is the LD Score of SNP j with respect to category C (defined as $\ell(j,C)=\sum_{k\in C} r_{jk}^2$), a is a term that measures the contribution of confounding biases[16] and τ_C represents the per-SNP contribution to heritability of category C. In particular, if the categories are disjoint, τ_C is the per-SNP heritability in category C and, if the categories overlap, the per-SNP heritability of SNP j is $\sum_{C:j\in C} \tau_C$. Equation (1) allows us to estimate τ_C via the (computationally simple) multiple regression of χ^2 statistics against $\ell(j,C)$, for either a quantitative or case-control study. We define the enrichment of a category to be the proportion of SNP heritability in the category divided by the proportion of SNPs in that category. We estimate standard errors with a block jackknife[16] and use these standard errors to calculate z scores, P values and false discovery rates (FDRs). We have released open source software implementing the method (see URLs; for further details, see the Online Methods and **Supplementary Note**).

To apply stratified LD score regression (or REML), we must first specify which categories are included in our model. We created a 'full baseline model' from 24 publicly available main annotations that are not specific to any cell type (**Supplementary Table 1**; see URLs and the Online Methods). Below, we show that including many categories in our model leads to more accurate estimates of enrichment. The 24 main annotations include coding, UTR, promoter and intronic regions[14,17]; the histone marks monomethylation (H3K4me1) and trimethylation (H3K4me3) of histone H3 at lysine 4 (refs. 3–5), acetylation of histone H3 at lysine 9 (H3K9ac)[3–5] and two versions of acetylation of histone H3 at lysine 27 (H3K27ac)[18,19]; open chromatin, as reflected by DNase I hypersensitivity sites (DHSs)[5,14]; combined chromHMM and Segway predictions[20], which make use of many Encyclopedia of DNA Elements (ENCODE) annotations to produce a single partition of the genome into seven underlying chromatin states; regions that are conserved in mammals[21,22]; super-enhancers, which are large clusters of highly active enhancers[19]; and enhancers with balanced bidirectional capped transcripts identified using cap analysis of gene expression in the FANTOM5 panel of samples, which we call FANTOM5 enhancers[23]. For histone marks and other annotations that differ among cell types, we combined the data from the different cell types into a single annotation for the full baseline model by taking a union (except for the repressed category, for which we took an intersection). To prevent our estimates from being biased upward by enrichment in nearby regions[14], we also included 500-bp windows around each of the 24 main annotations in the full baseline model, as well as 100-bp windows around chromatin immunoprecipitation and sequencing (ChIP-seq) peaks when appropriate (Online Methods). This yielded a total of 53 (overlapping) functional categories in the full baseline model, including a category containing all SNPs.

In addition to the analyses using the full baseline model, we performed analyses using cell type–specific annotations for the four histone marks H3K4me1, H3K4me3, H3K9ac and H3K27ac. Each cell type–specific annotation corresponds to a histone mark in a single cell type—for example, H3K27ac in liver cells—and there were 220 such annotations in total (Online Methods and **Supplementary Table 2**). When generating these 220 cell type–specific annotations, we wanted to control for overlap with the functional categories in the full baseline model but not for overlap with the 219 other cell type–specific annotations. Thus, we added these annotations individually to the baseline model, creating 220 separate models, each with 54 annotations. Then, for a given phenotype, we ran LD score regression once each on the 220 models and ranked the cell type–specific annotations by the P value of the coefficient τ_C of the annotation in the corresponding analysis. This P value tests whether the annotation contributes significantly to SNP heritability after controlling for the effects of the annotations in the full baseline model.

We also divided the 220 cell type–specific annotations into 10 groups: adrenal and pancreas, CNS, cardiovascular, connective and bone, gastrointestinal, immune and hematopoietic, kidney, liver, skeletal muscle and other. We took a union of the cell type–specific annotations within each group, resulting in ten new cell type group annotations (for example, SNPs with any of the four histone modifications in any CNS cell type). We then repeated the cell type–specific analysis described above with these ten cell type groups instead of the 220 cell type–specific annotations.

Simulation results assessing power and bias

In our first set of simulations, we assessed the power and bias of the method with a variety of settings for SNP heritability (h_g^2), sample size (N) and proportion of causal SNPs (p_{causal}) (Online

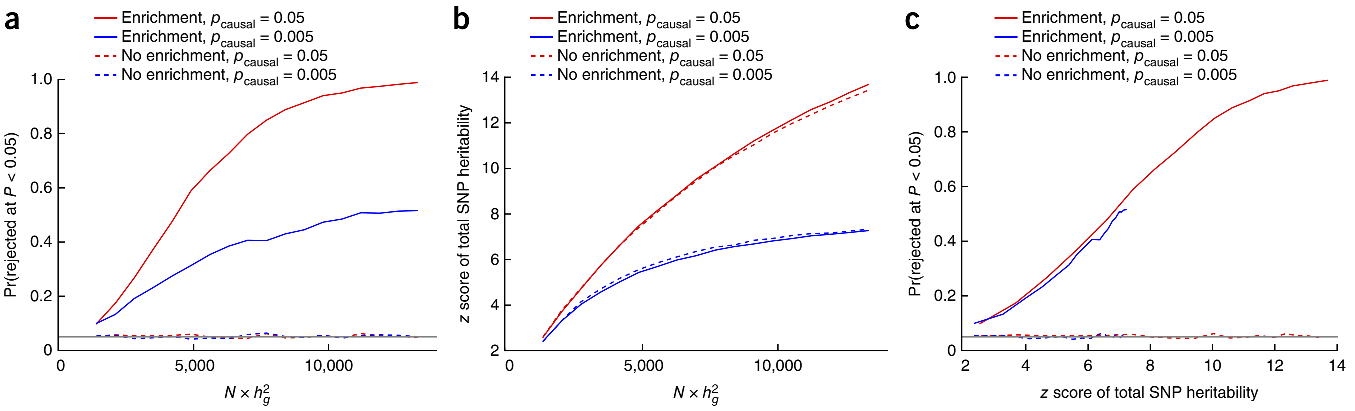

Figure 1 Simulation results for null calibration and power. We simulated genetic architectures with positive total SNP heritability, with and without functional enrichment, for two values of p_{causal} and a range of values of $N\times h_g^2$. (**a**) Proportion of simulations in which a null hypothesis of no functional enrichment is rejected, as a function of $N\times h_g^2$ and p_{causal}. (**b**) The z score of total SNP heritability depends on $N\times h_g^2$ and p_{causal} but does not depend on the presence or absence of functional enrichment. (**c**) Proportion of simulations in which a null hypothesis of no functional enrichment is rejected, as a function of the z score of total SNP heritability. Here the z score of total SNP heritability for p_{causal} = 0.005 did not exceed 7.3, even at maximum $N\times h_g^2$.

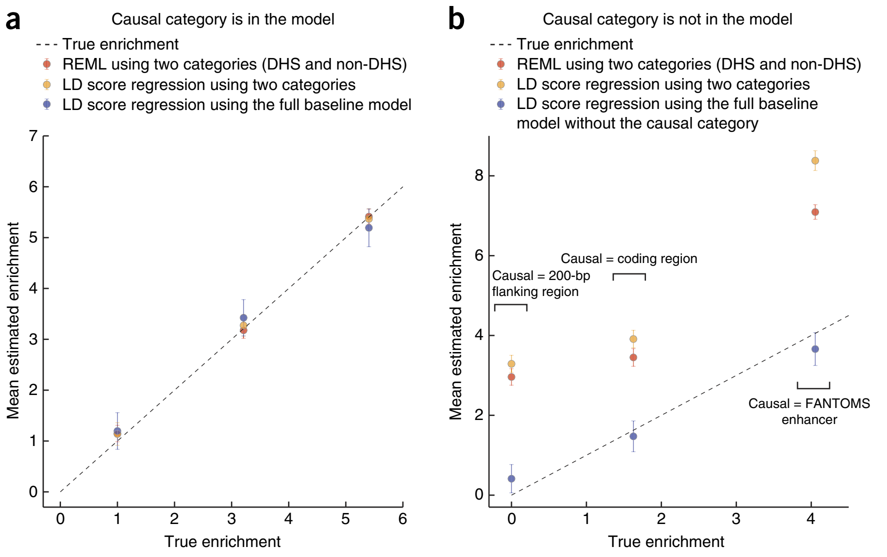

Figure 2 Simulation results for model misspecification. Enrichment is the proportion of heritability in DHS regions divided by the proportion of SNPs in DHS regions. Bars show 95% confidence intervals around the mean of 100 trials. (**a**) From left to right, the simulated genetic architectures are 1× DHS enrichment, 3× DHS enrichment and 5.5× DHS enrichment (100% of heritability from DHS SNPs). (**b**) From left to right, the simulated genetic architectures are 200-bp flanking regions causal, coding regions causal and FANTOM5 enhancer regions causal. For simulations with coding region or FANTOM5 enhancer as the causal category, we removed the causal category and the 500-bp window around that category from the full baseline model to simulate enrichment in an unknown functional category.

Methods). These simulations demonstrated well-calibrated type I error at all settings for h_g^2, N and p_{causal} tested (**Fig. 1**). At a fixed value for p_{causal}, power depended on N and h_g^2 only through $N \times h_g^2$ (**Supplementary Fig. 1**) and increased as $N \times h_g^2$ increased and as p_{causal} increased (**Fig. 1a**). We also looked at the z score for total SNP heritability in our analysis, which increased as $N \times h_g^2$ and p_{causal} increased (**Fig. 1b**). We found that the relationship of the heritability z score at power was the same for both values of p_{causal} (**Fig. 1c**), indicating that the heritability z score is a good indicator of power at a variety of sample sizes, heritabilities and values of p_{causal}. For this report, we chose to analyze only traits with a heritability z score above 7, which corresponds to $N \times h_g^2$ of roughly 4,500 for very polygenic traits and 12,500 for less polygenic traits.

In each of these simulations, stratified LD score regression gave unbiased estimates of heritability and of the heritability of the CNS cell type group (**Supplementary Figs. 2a**,**b** and **3a**,**b**). Although, in theory, the ratio of these two unbiased estimators could be a biased estimator of the proportion of heritability (and, therefore, the estimates that we report here), in practice, we saw only negligible bias in our estimates of proportion of heritability (**Supplementary Figs. 2c** and **3c**). Using LD computed from an out-of-sample reference panel caused some downward attenuation bias in estimates of total SNP heritability and category-specific heritability but also gave unbiased estimates of the proportion of heritability and properly calibrated type I error (**Supplementary Fig. 4**).

Simulation results assessing model misspecification

In our second set of simulations, we compared stratified LD score regression to REML, a method that also estimates partitioned heritability but requires genotype data, in scenarios with and without model misspecification (Online Methods). We estimated the enrichment of the DHS category, that is, (proportion of h_g^2)/(proportion of SNPs), using three methods: (i) REML with two categories (DHS and non-DHS); (ii) stratified LD score regression with two categories (DHS and non-DHS); and (iii) stratified LD score regression with the full baseline model (53 categories). Because REML with 53 categories did not converge at this sample size and would be computationally intractable at sample sizes in the tens of thousands, we did not include it in our comparison; an advantage of stratified LD score regression is that it is possible to include a large number of categories in the underlying model. We report means with s.e.m. over 100 independent simulations.

We first performed three sets of simulations without model misspecification, where the causal pattern of enrichment was well modeled by the two-category (DHS and non-DHS) model. In these simulations, enrichment of the DHS region varied from 1× (no enrichment) to 5.5× (full enrichment, where DHS SNPs explain 100% of heritability). All three methods gave unbiased estimates, although stratified LD score regression with the full baseline model had larger standard errors around the means (**Fig. 2a**).

Next, to explore the realistic scenario where the model used to estimate enrichment does not match the (unknown) causal model, we performed three sets of simulations where all causal SNPs were in a particular category but the model used to estimate heritability did not include this causal category. The three sets of simulations included (i) all causal SNPs in coding regions, yielding a true 1.6× DHS enrichment due to coding and DHS overlap; (ii) all causal SNPs in FANTOM5 enhancers, yielding a true 4.0× DHS enrichment due to FANTOM5 enhancer and DHS overlap; and (iii) all causal SNPs in 200-bp DHS-flanking regions, yielding a true 0× DHS enrichment. For the coding region and FANTOM5 enhancer causal simulations, we transformed the full baseline model into a misspecified model by removing the causal category and window around the causal category; the baseline model includes a 500-bp window around DHSs but not a 200-bp window and so is misspecified also in that case. The results from these simulations are displayed in **Figure 2b**. The two-category estimators were not robust to model misspecification and consistently overestimated DHS enrichment by a wide margin. Stratified LD score regression with the full baseline model gave more accurate mean estimates of enrichment.

In summary, although these simulations include exaggerated patterns of enrichment (for example, where 100% of heritability is present in DHS-flanking regions), the results highlight the possibility that two-category estimators of enrichment can yield incorrect conclusions. Although we cannot entirely rule out model misspecification as a source of bias for stratified LD score regression with the full baseline model, we have shown here that this method is robust to a wide variety of patterns of enrichment because including many categories gives it the flexibility to adapt to the unknown causal model.

Simulation results for cell type and cell type group analyses

We simulated realistic baseline plus enrichment in a cell type group (Online Methods), and we performed our cell type group analysis on the resulting summary statistics. First, we calibrated simulated enrichment of the causal cell type group to give us a realistic average

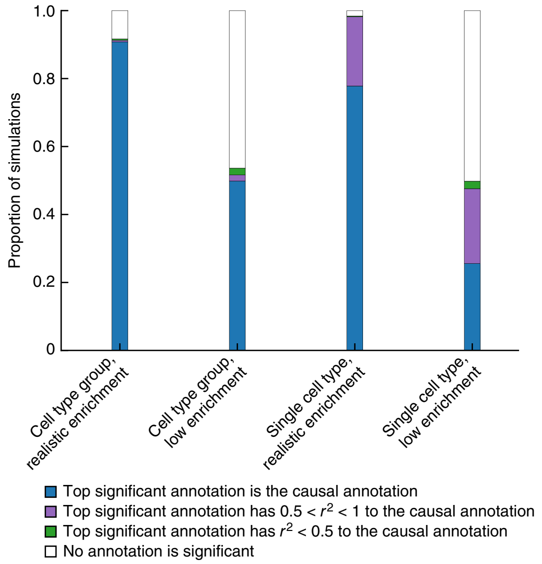

Figure 3 Simulation results for ranking cell type groups and cell types. For each cell type group, 500 simulations were performed with baseline enrichment and either realistic enrichment or low enrichment in that cell type group. Results for the left two columns are aggregated over the ten cell type groups; results for individual groups are displayed in **Supplementary Figure 5**. The right two columns represent 500 simulations each of realistic or low enrichment of a single cell type–specific annotation—H3K4me3 in fetal brain cells.

top $-\log_{10}$ (P value) based on results for the real data sets analyzed below (Online Methods). In the simulations in which at least one cell type group reached significance, we found that the top cell type group was the cell type group simulated to be causal 99% of the time (**Fig. 3**). Next, we simulated weaker enrichment, calibrated so that only 50% of replicates included a significant cell type group. In these simulations, the cell type group simulated to be causal was the top cell type group in 95% of simulations with at least one significant cell type group, and a cell type group with r^2 >0.5 with the causal group was the top cell type group in half of the remaining simulations with at least one significant cell type group (**Fig. 3**). Results separated by the ten individual cell type groups are displayed in **Supplementary Figure 5**.

We next repeated these simulations with a cell type–specific mark—H3K4me3 in fetal brain cells—instead of a cell type group as the simulated causal category. There are many more pairs of cell types that are highly correlated than there are highly correlated pairs of cell type groups, and we are testing all cell types every time (**Supplementary Fig. 6**). We found that, when the level of enrichment was calibrated to give a realistic $-\log_{10}$ (P value) (based on results for the real data sets analyzed below; Online Methods), the simulated causal cell type was the most significant cell type in 78% of simulations, a cell type with $r^2 > 0.5$ with the causal cell type was most significant in 20% of simulations and there was no significant cell type in 2% of simulations. In simulations with weak enrichment—again calibrating so that 50% of simulations had at least one significant cell type—we found that, of the simulations with at least one significant

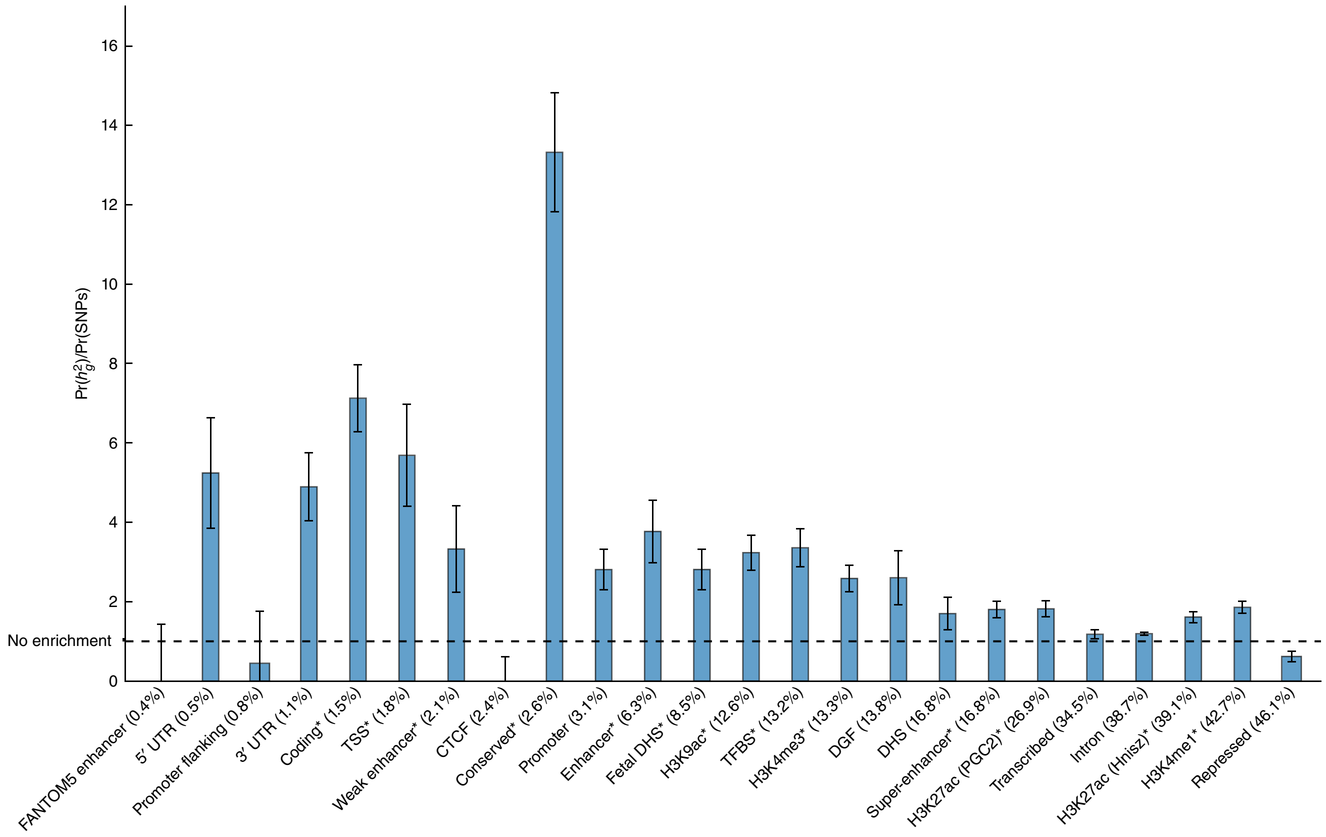

Figure 4 Enrichment estimates for the 24 main annotations, averaged over nine independent traits. Annotations are ordered by size. Error bars represent jackknife standard errors around the estimates of enrichment, and an asterisk indicates significance at $P < 0.05$ after Bonferroni correction for the 24 hypotheses tested. Negative point estimates, significance testing and the choice of nine independent traits are discussed in the Online Methods and **Supplementary Note**. TFBS, transcription factor binding site. DGF, digital genomic footprint.

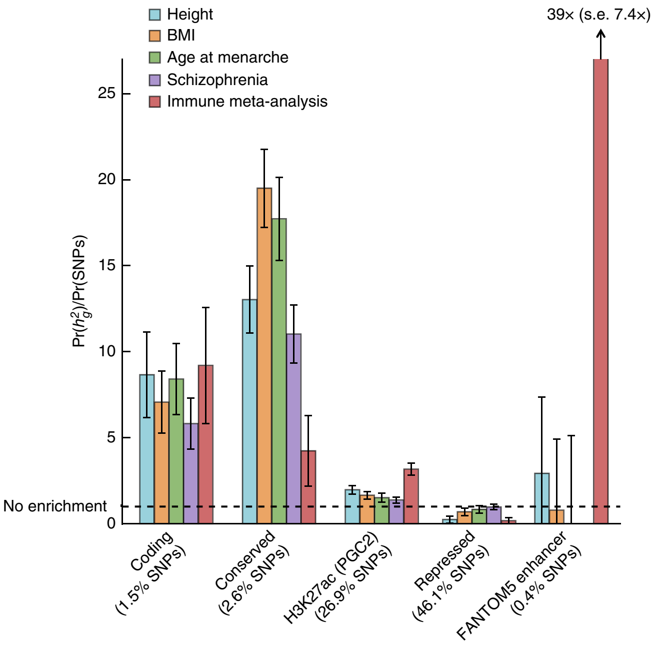

Figure 5 Enrichment estimates for selected annotations and traits. Error bars represent jackknife standard errors (s.e.) around the estimates of enrichment.

cell type, only 4% had as the top cell type a cell type with $r^2 < 0.5$ with the causal cell type.

In conclusion, the cell type group analysis reliably reports the causal annotation as the top annotation if at least one cell type group passes statistical significance. The analysis of individual cell types, because it is testing more cell types that are more correlated, often gives a highly correlated cell type as the top cell type, just as in a GWAS the top SNP in a locus is not always the causal SNP.

Analysis of 17 traits using the full baseline model

We applied stratified LD score regression to 17 diseases and quantitative traits: height, BMI, age at menarche, low-density lipoprotein (LDL) levels, high-density lipoprotein (HDL) levels, triglyceride levels, coronary artery disease, type 2 diabetes, fasting glucose levels, schizophrenia, bipolar disorder, anorexia, educational attainment, smoking behavior, rheumatoid arthritis, Crohn's disease and ulcerative colitis[18,24–36] (see the URLs and **Supplementary Table 3**). This analysis includes all traits with publicly available summary statistics with sufficient sample size, SNP heritability and polygenicity measured by the *z* score of total SNP heritability; specifically, we restricted our analysis to traits for which the *z* score of total SNP heritability was at least 7 (**Supplementary Table 4** and **Supplementary Note**). We removed the major histocompatibility complex (MHC) region from all analyses because of its unusual LD and genetic architecture.

We applied stratified LD score regression with the full baseline model to the 17 traits. Results for the 24 main functional annotations, averaged across nine independent traits, are shown in **Figure 4** (Online Methods). Trait-specific results for selected annotations and traits are shown in **Figure 5** (**Supplementary Note**). Meta-analysis and trait-specific results for all traits and all 53 categories in the full baseline model are shown in **Supplementary Tables 5** and **6**.

We observed large and statistically significant enrichments for many functional categories. A few categories stood out in particular. First, regions conserved in mammals[21] showed the largest enrichment of any category, with 2.6% of SNPs explaining an estimated 35% of SNP heritability on average across traits ($P < 1 \times 10^{-6}$ for enrichment). This is a significantly higher average enrichment than for coding regions and provides evidence for the biological importance of conserved regions, despite the fact that the biochemical function of many conserved regions remains uncharacterized[37]. Second, FANTOM5 enhancers[23] were extremely enriched in the three immunological diseases, with 0.4% of SNPs explaining an estimated 15% of SNP heritability on average across these three diseases ($P = 1 \times 10^{-4}$, 2×10^{-4} and 0.03 for Crohn's disease, ulcerative colitis and rheumatoid arthritis, respectively) but showed no evidence of enrichment for non-immunological traits (**Fig. 5**). The immune-specific enrichment could be due to immune cells having altered degradation, a higher number of enhancers and/or better sequence coverage in the FANTOM5 experiments. We did not see a large enrichment of super-enhancers in comparison to regular enhancers; the estimates for enrichment were 1.8× (standard error = 0.2) for super-enhancers versus 1.6× (standard error = 0.1) for regular enhancers using data from the same report[19] (denoted "H3K27ac (Hnisz)" in **Fig. 4**). We also did not see increased cell type specificity for super-enhancers (**Supplementary Note**). This lack of enrichment supports the hypothesis that super-enhancers may not have a much more important role in regulating transcription than regular enhancers[38]. For many annotations, there was also enrichment in the 500-bp flanking regions (**Supplementary Table 5**); this could be because the boundaries of functional regions are not well defined, because the boundaries of these regions are different in different individuals or because unknown regulatory elements often appear close to known regulatory elements. Analyses stratified by derived allele frequency produced broadly similar results (**Supplementary Table 7**; see Online Methods).

Table 1 Enrichment of individual cell types

Phenotype	Cell type	Tissue	Mark	$-\log_{10}(P)$
Height	Chondrogenic differentiationf[b]	Bone	H3K27ac	6.81
BMI	Fetal brain[a]	Fetal brain	H3K4me3	4.48
Age at menarche	Fetal brain[a]	Fetal brain	H3K4me3	12.25
LDL	Liver[a]	Liver	H3K4me1	4.76
HDL	Liver[a]	Liver	H3K4me1	4.51
Triglycerides	Liver[a]	Liver	H3K4me1	3.99
Coronary artery disease	Adipose nuclei[a]	Adipose	H3K4me1	4.21
Type 2 diabetes	Pancreatic islets	Pancreas	H3K4me3	2.87
Fasting glucose	Pancreatic islets[a]	Pancreas	H3K27ac	3.93
Schizophrenia	Fetal brain[b]	Fetal brain	H3K4me3	18.51
Bipolar disorder	Mid-frontal lobe[a]	Brain	H3K27ac	4.42
Anorexia	Angular gyrus	Brain	H3K9ac	2.61
Years of education	Angular gyrus[b]	Brain	H3K4me3	6.63
Ever smoked	Inferior temporal lobe[a]	Brain	H3K4me3	3.21
Rheumatoid arthritis	$CD4^+CD25^-IL17^+$ stimulated T_H17[b]	Immune	H3K4me1	6.76
Crohn's disease	$CD4^+CD25^-IL17^+$ stimulated T_H17[b]	Immune	H3K4me1	7.59
Ulcerative colitis	$CD4^+CD25^-IL17^+$ stimulated T_H17[b]	Immune	H3K4me1	6.37

We report the cell type with the lowest *P* value for each trait analyzed.
[a]FDR < 0.05. [b]Significant at $P < 0.05$ after Bonferroni correction for multiple hypotheses. Sample sizes are in **Supplementary Table 3**.

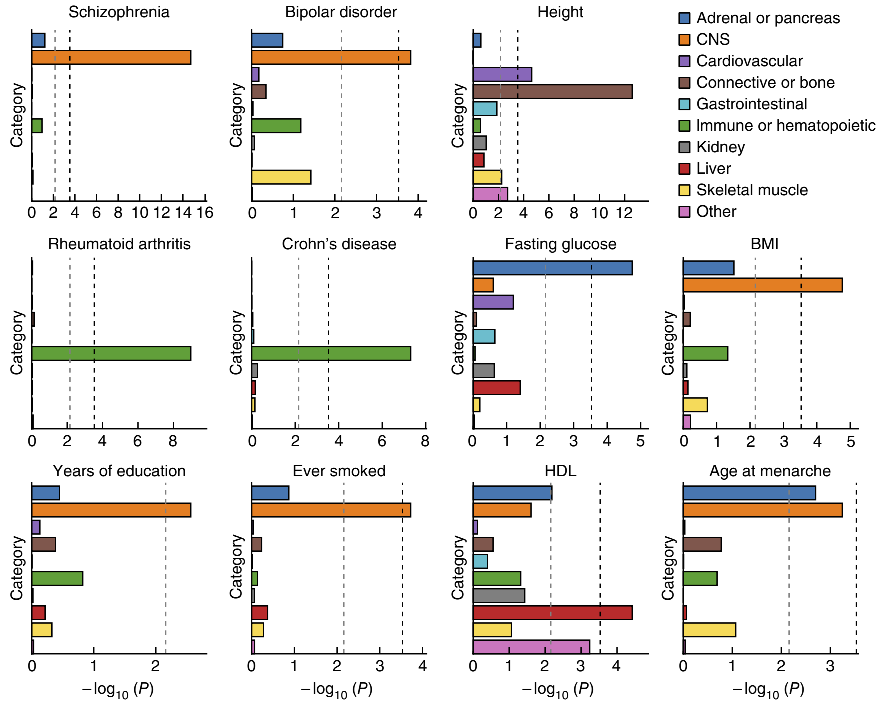

Figure 6 Enrichment of cell type groups. We report the significance of enrichment for each of ten cell type groups for each of 11 traits. The black dashed lines at $-\log_{10}(P) = 3.5$ is the cutoff for Bonferroni significance. The gray dashed lines at $-\log_{10}(P) = 2.1$ is the cutoff for FDR < 0.05. For HDL, three of the top individual cell types are adipose nuclei, which explains the enrichment of the 'other' category.

Cell type–specific analysis of 17 traits

We performed two different cell type–specific analyses: an analysis of 220 individual cell type–specific annotations and an analysis of 10 cell type groups. For the analysis of single cell types, we assessed statistical significance at $P < 0.05$ after Bonferroni correction for the 220 × 17 = 3,740 hypotheses tested, and, for the cell type group analysis, we corrected for the 10 × 17 = 170 hypotheses tested. These thresholds are conservative, as the 220 cell type–specific annotations are not independent and neither are the ten cell type group annotations. We also report results with FDR < 0.05, computed over 220 cell types for each trait in the cell type–specific analysis and over all cell type groups and traits in the cell type group analysis. For 15 of the 17 traits, the top cell type passed an FDR threshold of 0.05; for 16 of the 17 traits (all traits except anorexia), the top cell type group passed an FDR threshold of 0.05. The top cell type for each trait is displayed in **Table 1**, with additional top cell types reported in **Supplementary Table 8**. Cell type group results for the 11 traits with the most significant enrichments (after pruning closely related traits) are shown in **Figure 6**, with remaining traits shown in **Supplementary Figure 7**.

These two analyses were generally concordant and showed highly trait-specific patterns of cell type enrichment. They also recapitulated several well-known findings. For example, the top cell type for each of the three lipid traits was liver (FDR < 0.05 for all three traits). For both type 2 diabetes and fasting glucose levels, the top cell type was pancreatic islets (FDR < 0.05 for fasting glucose levels but not type 2 diabetes). For the three psychiatric traits, the top cell type was a brain cell type and the top cell type group was CNS (FDR < 0.05 for schizophrenia and bipolar disorder but not anorexia). These results are concordant with the medical literature[39,40] and with previous analysis of these GWAS data sets[9,18,27,31,41,42].

There were also several new insights among these results. For example, the three immunological diseases showed patterns of enrichment that reflect biological differences. Crohn's disease had 40 cell types with FDR < 0.05, of which 39 were immune cell types and one (colonic mucosa) was a gastrointestinal cell type. In contrast, the 39 cell types with FDR < 0.05 for ulcerative colitis included nine gastrointestinal cell types in addition to 30 immune

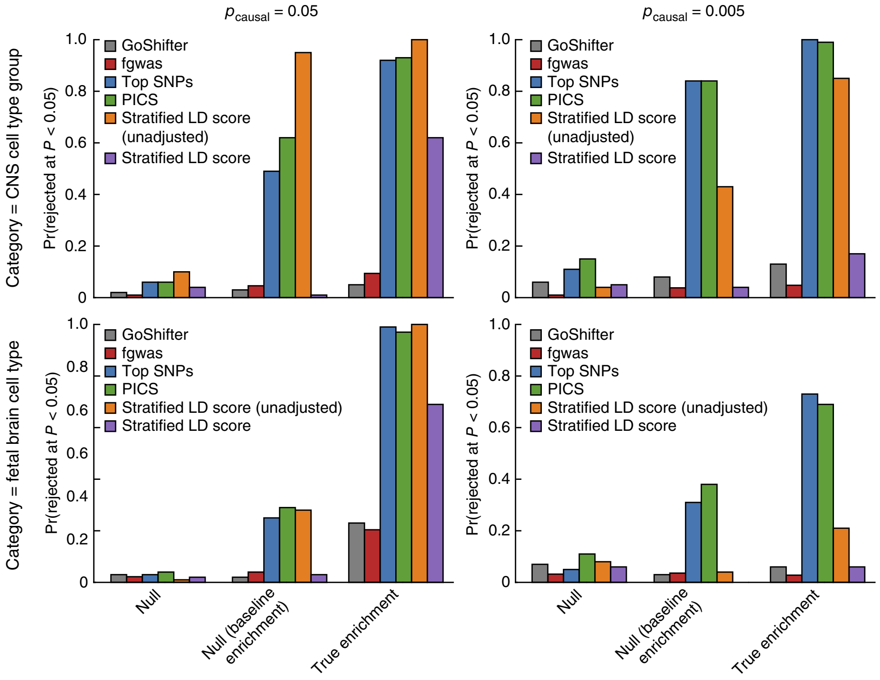

Figure 7 Comparison of stratified LD score regression to other methods for identifying enriched cell types. In null simulations, there is no enrichment. In null (baseline enrichment) simulations, there is enrichment in the baseline categories, some of which overlap the cell type or cell type group, but no additional enrichment in the cell type or cell type group. In the true enrichment simulations, there is enrichment in either the CNS cell type group (top) or the fetal brain cell type (bottom). In all simulations, $N = 14{,}000$ and $h_g^2 = 0.7$. We report the proportion of 100 simulations in which the null model is rejected for six methods: GoShifter[6], fgwas[9], Top SNPs[10], PICS[7], stratified LD score (unadjusted) and stratified LD score. Stratified LD score (unadjusted) refers to total unadjusted enrichment, that is, (proportion of h_g^2)/(proportion of SNPs); LD score refers to the coefficient τ of the category, controlling for all other categories in the model.

cell types, whereas all 39 cell types with FDR < 0.05 for rheumatoid arthritis were immune cell types. The top cell type for all three traits was CD4^{+}CD25–IL17^{+} phorbol 12-myristate 13-acetate (PMA)- and ionomycin-stimulated T helper 17 (T_H17) primary cells. T_H17 cells are thought to act in opposition to regulatory T (T_{reg}) cells, which have been shown to suppress immune activity and whose malfunction has been associated with immunological disorders[43].

We also identified several non-psychiatric phenotypes with enrichments in brain cell types. For both BMI and age at menarche, cell types in the CNS ranked highest among individual cell types, and the top cell type group was CNS, all with FDR < 0.05. These enrichments support previous human and animal studies that propose a strong neural basis for the regulation of energy homeostasis[44]. For educational attainment, the top cell type group was CNS (FDR < 0.05) and, of the ten cell types that were significant after correction for multiple testing, nine were CNS cell types. This finding is consistent with the understanding that the genetic component of educational attainment, which excludes environmental factors and population structure, is highly correlated with IQ[45]. Finally, for smoking behavior, the CNS cell type group was significant and the top cell type was again a brain cell type, likely reflecting CNS involvement in nicotine processing.

DISCUSSION

We developed a new statistical method, stratified LD score regression, for identifying functional enrichment from GWAS summary statistics that uses genome-wide information from all SNPs and explicitly models LD. We applied this method to summary statistics for 17 traits with an average sample size of 73,599. Our method identified strong enrichment for conserved regions across all traits and immunological disease–specific enrichment for FANTOM5 enhancers. Our cell type–specific enrichment results confirmed previously known enrichments, such as liver enrichment for HDL levels and pancreatic islet enrichment for fasting glucose levels. In addition, we identified enrichments that would have been challenging to detect using existing methods, such as CNS enrichment for smoking behavior and educational attainment—traits with only one and three genome-wide significant loci, respectively[33,34]. Stratified LD score regression represents a significant departure from previous methods that require raw genotype data[11], use only SNPs in genome-wide significant loci[5–8], assume only one causal SNP per locus[9] or do not account for LD[10] (**Fig. 7**; see the Online Methods for a discussion of other methods). Our method is also computationally efficient, despite the 53 overlapping functional categories analyzed.

Although our polygenic approach has enabled a powerful analysis of genome-wide summary statistics, it has several limitations. First, for the method to have reasonable power, the data set analyzed must have a very large sample size and/or large SNP heritability and the trait analyzed must be polygenic (**Fig. 1**). Second, the method requires an LD reference panel matched to the population studied to give accurate results; all results here are from European data sets and use 1000 Genomes Project Europeans as a reference panel (Online Methods and **Supplementary Fig. 4**). Third, our method is currently not applicable to studies using custom genotyping arrays (for example, the Metabochip; **Supplementary Note**). Fourth, our method is based on an additive model and does not consider the contribution of epistatic or other non-additive effects, nor does it model the causal contributions of SNPs not in the reference panel; in particular, it is possible that patterns of enrichment at extremely rare variants may be different from those inferred using this method (Online Methods). Fifth, the method is limited by available functional data: if a trait is enriched in a cell type for which we have no data, we cannot detect the enrichment. Sixth, our method currently gives large standard errors when applied to very small categories (**Supplementary Fig. 8** and **Supplementary Note**). Last, although we have shown our method to be robust in a wide range of scenarios, we cannot rule out bias due to model misspecification, caused by enrichment in an unidentified functional category, as a possible source of bias; however, our simulations show that our method gives nearly unbiased results even under very extreme scenarios of unmodeled functional categories (**Fig. 2**).

In conclusion, the polygenic approach described here is a powerful and efficient way to learn about functional enrichments from summary statistics. It will likely become increasingly useful as the amount of functional data continues to grow and the quality of these data improve, and as GWAS of larger sample size are conducted.

URLs. ldsc software, http://www.github.com/bulik/ldsc; baseline and cell type group annotations, http://data.broadinstitute.org/alkesgroup/LDSCORE/; 1000 Genomes Project, http://www.1000genomes.org/; height[24] and BMI[25] summary statistics, http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files; age at menarche summary statistics[26], http://www.reprogen.org/; LDL, HDL and triglyceride summary statistics[27], http://www.broadinstitute.org/mpg/pubs/lipids2010/; coronary artery disease summary statistics[28], http://www.cardiogramplusc4d.org/; type 2 diabetes summary statistics[29], http://www.diagram-consortium.org/; fasting glucose summary statistics[30], http://www.magicinvestigators.org/downloads/; schizophrenia[18], bipolar disorder[31], anorexia[32] and smoking behavior[33] summary statistics, http://www.med.unc.edu/pgc/downloads; educational attainment summary statistics[34], http://www.ssgac.org/; rheumatoid arthritis summary statistics[35], http://plaza.umin.ac.jp/yokada/datasource/software.htm; Crohn's disease and ulcerative colitis summary statistics[36], http://www.ibdgenetics.org/downloads.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank B. Bernstein, M. Finucane, A. Forrest, E. Hodis, D. Kotliar, X.S. Liu, M. Kellis, M. O'Donovan, B. Pasaniuc, A. Sandelin, A. Sarkar, P. Sullivan, B. Vilhjalmsson, A. Veres and the anonymous reviewers for helpful discussions and/or comments. This research was funded by US National Institutes of Health grants R01 MH101244, R01 HG006399, R03 CA173785, R21 CA182821, F32 GM106584 and U01 HG0070033. H.K.F. was also supported by the Fannie and John Hertz Foundation. G.T. is supported by the Wellcome Trust Sanger Institute (WT098051). Y.R. was supported by award T32 GM007753 from the National Institute of General Medical Sciences. S. Raychaudhuri is supported by funding from the Arthritis Foundation and by a Doris Duke Clinical Scientist Development Award. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. This study made use of data generated by the Wellcome Trust Case Control Consortium (WTCCC) and the Wellcome Trust Sanger Institute. A full list of the investigators who contributed to the generation of the WTCCC data is available at http://www.wtccc.org.uk/. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113.

AUTHOR CONTRIBUTIONS

H.K.F., B.B.-S., A.G., G.T., Y.R., P.-R.L., V.A., S. Raychaudhuri, M.J.D., N.P., B.M.N. and A.L.P. conceived and designed the experiments. H.K.F. and B.B.-S. performed the experiments, performed the statistical analysis and analyzed the data. H.X., C.Z., K.F., S. Ripke, F.R.D., S.P., E.S., S.L., J.R.B.P. and Y.O. contributed reagents. H.K.F., B.B.-S., B.M.N. and A.L.P. wrote the manuscript with feedback from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
2. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
4. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
5. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).
6. Trynka, G. *et al.* Disentangling effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex trait loci. *Am. J. Hum. Genet.* **97**, 139–152 (2015).
7. Farh, K.K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
8. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
9. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).
10. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
11. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
12. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
13. Davis, L.K. *et al.* Partitioning the heritability of Tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS Genet.* **9**, e1003864 (2013).
14. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
15. Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
16. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
17. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
18. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
19. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
20. Hoffman, M.M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).
21. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
22. Ward, L.D. & Kellis, M. Evidence of abundant purifying selection in humans for recently-acquired regulatory functions. *Science* **337**, 1675–1678 (2012).
23. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
24. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
25. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
26. Perry, J.R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).
27. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
28. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
29. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
30. Manning, A.K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
31. Psychiatric GWAS Consortium Bipolar Disorder Working Group. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near *ODZ4*. *Nat. Genet.* **43**, 977–983 (2011).
32. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol. Psychiatry* **19**, 1085–1094 (2014).
33. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467–1471 (2013).
34. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
35. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
36. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
37. Stamatoyannopoulos, J.A. What does our genome encode? *Genome Res.* **22**, 1602–1611 (2012).
38. Pott, S. & Lieb, J.D. What are super-enhancers? *Nat. Genet.* **47**, 8–12 (2015).
39. Lilly, L.S. *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (Lippincott Williams & Wilkins, 2012).
40. Kettyle, W.M. & Arky, R.A. *Endocrine Pathophysiology* (Lippincott Williams & Wilkins, 1998).
41. Parker, S.C.J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. USA* **110**, 17921–17926 (2013).
42. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).
43. Wang, W. *et al.* The Th17/Treg imbalance and cytokine environment in peripheral blood of patients with rheumatoid arthritis. *Rheumatol. Int.* **32**, 887–893 (2012).
44. Farooqi, I.S. Defining the neural basis of appetite and obesity: from genes to behaviour. *Clin. Med.* **14**, 286–289 (2014).
45. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* doi:10.1038/ng.3406 (28 September 2015).

[1]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [4]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [6]Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [7]Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, USA. [8]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [9]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [10]Department of Computer Science, Harvard University, Cambridge, Massachusetts, USA. [11]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [12]Epigenomics Program, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [13]Medical Research Council (MRC) Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [14]A full list of members appears in the **Supplementary Note**. [15]Department of Psychiatry, Mount Sinai School of Medicine, New York, New York, USA. [16]Department of Human Genetics and Disease Diversity, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Tokyo, Japan. [17]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [18]Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK. [19]These authors contributed equally to this work. [20]These authors jointly supervised this work. Correspondence should be addressed to H.K.F. (hilaryf@mit.edu), B.B.-S. (bulik@broadinstitute.org), B.M.N. (bneale@broadinstitute.org) or A.L.P. (aprice@hsph.harvard.edu).

ONLINE METHODS

Stratified LD score regression. We assume a linear model

$$y_i = \sum_j X_{ij}\beta_j + \varepsilon_i$$

where y_i is a quantitative phenotype in individual i, X_{ij} is the standardized genotype of individual i at SNP j, β_j is the effect size of SNP j and ε_i is mean-zero noise. We define heritability by

$$h^2 = \sum_j \beta_j^2$$

and the heritability of a category C to be

$$h^2(C) = \sum_{j \in C} \beta_j^2$$

We model β as a mean-zero random vector with independent entries. We have C functional categories $C_1, C_2, \ldots, C_C$, and we allow the variance of β_j, that is, the per-SNP heritability at SNP j, to depend on these functional categories via the equation

$$\mathrm{Var}\left(\beta_j\right) = \sum_{c: j \in C_c} \tau_c \qquad (2)$$

In the case that the categories are disjoint, we have $\tau_c = h^2(C_c)/M(C_c)$, where $M(C_c)$ is the number of SNPs in category C_c. Each SNP must be in at least one category; in practice, we either have a set of categories that form a disjoint partition of the genome or we include the set of all SNPs as one of the categories.

In the **Supplementary Note**, we show that, under this model,

$$E\left[\chi_j^2\right] = N \sum_c \tau_c \ell\left(j,c\right) + 1 \qquad (3)$$

where χ_j^2 is the marginal association test statistic at SNP j, N is the sample size of the study and $\ell\left(j,c\right) := \sum_{k \in C_c} r_{jk}^2$. An extension of this derivation to case-control traits appears in Bulik-Sullivan *et al.*[45].

Given a vector of χ^2 statistics and LD information, either from the sample or a reference panel, equation (3) allows us to obtain estimates $\hat{\tau}_c$ of τ_c by computing $\ell\left(j,c\right)$ and regressing χ_j^2 on $\ell\left(j,c\right)$. For some analyses, including the cell type and cell type group analyses described in this manuscript, estimating τ_c is the goal. For other analyses, including the baseline analyses described in this manuscript, the goal is to estimate $h^2\left(C_c\right) := \sum_{j \in C_c} \beta_j^2$ or $h^2(C_c)/h^2$. Because β is a mean-zero random vector, we can approximate $h^2(C_c)$ with its expectation, $\sum_{j \in C_c} \mathrm{Var}\left(\beta_j\right)$. When the categories are disjoint, $\mathrm{Var}(\beta_j) = \tau_c$ where SNP j is in category C_c, and so $\hat{h}^2\left(C_c\right) = \left|C_c\right| \cdot \hat{\tau}_c$. When the categories overlap, we apply equation (2), which gives us

$$\hat{h}^2\left(C_c\right) = \sum_{j \in C_c} \widehat{\mathrm{Var}}\left(\beta_j\right) = \sum_{j \in C_c} \sum_{c': j \in C_{c'}} \hat{\tau}_c$$

In this report, we use HapMap Project Phase 3 (HapMap 3)[46] SNPs for our regression and 1000 Genomes Project[47] SNPs for our reference panel, and we only partition the heritability of SNPs with minor allele frequency (MAF) above 5% (**Supplementary Note**). The details of the regression, including outlier removal, out-of-bounds estimates, regression weights and genomic control (GC) correction are given in the **Supplementary Note**.

Significance testing. We estimate standard errors using a block jackknife over SNPs with 200 equally sized blocks of adjacent SNPs[16]. This gives us an empirical covariance matrix of coefficient estimates. In the baseline analysis, to evaluate whether a category is enriched for heritability, we want to test whether $\frac{h^2(C)}{h^2} > \frac{|C|}{M}$. This is the same as testing whether the per-SNP heritability is greater in the category than out of the category, that is, whether $\frac{h^2(C)}{|C|} - \frac{h^2 - h^2(C)}{M - |C|} > 0$. Because our estimates of the regression coefficients are approximately normally distributed and $\frac{h^2(C)}{h^2}$ is therefore not normally distributed but $\frac{h^2(C)}{|C|} - \frac{h^2 - h^2(C)}{M - |C|}$ is, we use the latter expression to test for significance. Because this expression is linear in the coefficients, we can estimate its standard error using the covariance matrix for the coefficient estimates and then compute a z score to test for significance. This procedure is well calibrated (**Fig. 1**). We also report the jackknife standard errors of the proportion of heritability, even though this is not what we use to assess significance.

For the cell type–specific analyses, we use the z score of the coefficient directly.

Code availability. Stratified LD score regression is available as open source software at http://www.github.com/bulik/ldsc.

Full baseline model. The 53 functional categories, derived from 24 main annotations, were obtained as follows:

- Coding, 3′ UTR, 5′ UTR, promoter and intron annotations from RefSeq gene models were obtained from the UCSC Genome Browser[17] and post-processed by Gusev *et al.*[14].
- Digital genomic footprint and transcription factor binding site annotations were obtained from ENCODE[3] and post-processed by Gusev *et al.*[14].
- The combined chromHMM and Segway annotations for six cell lines were obtained from Hoffman *et al.*[20]. The CTCF, promoter-flanking, transcribed, transcription start site (TSS), strong enhancer and weak enhancer categories are each a union over the six cell lines; the repressed category is an intersection over the six cell lines.
- DHSs are a combination of ENCODE and Roadmap Epigenomics data, post-processed by Trynka *et al.*[5]. We combined the cell type–specific annotations into two annotations for inclusion in the full baseline model: a union of all cell types and a union of only fetal cell types.
- Cell type–specific H3K4me1, H3K4me3 and H3K9ac data were all obtained from Roadmap Epigenomics and post-processed by Trynka *et al.*[5]. For each mark, we took a union over cell types for the full baseline model and used the individual cell types for our cell type–specific analyses.
- Cell type–specific H3K27ac data were obtained from Roadmap Epigenomics and post-processed[18]. A second version of H3K27ac data was obtained from Hnisz *et al.*[19]. For each mark, we took a union over cell types for the full baseline model. We also used the individual cell types of the Roadmap Epigenomics H3K27ac data for our cell type–specific analyses.
- Super-enhancers were also obtained from Hnisz *et al.*[19] and comprise a subset of the H3K27ac annotation from that paper. We took a union over cell types for the full baseline model.
- Regions conserved in mammals were obtained from Lindblad-Toh *et al.*[21] and post-processed by Ward and Kellis[22].
- FANTOM5 enhancers were obtained from Andersson *et al.*[23].
- For each of these 24 categories, we added a 500-bp window around the category as an additional category to keep our heritability estimates from being inflated by heritability in flanking regions[14].
- For each DHS, H3K4me1, H3K4me3 and H3K9ac site, we added a 100-bp window around the ChIP-seq peak as an additional category.
- We added an additional category containing all SNPs.

When we report results in **Supplementary Tables 5–7**, we do not report results from the category containing all SNPs, as it has 100% of the heritability with zero standard error. (It might have a coefficient τ_c that is non-trivial, but in these tables we report proportions of heritability.)

According to our simulations (**Fig. 2**), including these 53 categories in our baseline model allows us to obtain unbiased or nearly unbiased estimates of enrichment for a wide range of potential new categories. To estimate the enrichment of a new annotation, we perform analyses using a model with these 53 annotations plus the new annotation. For example, for the cell type–specific analysis, we add each cell type–specific annotation to the baseline model, one at a time, and asses enrichment using the z score of the cell type–specific annotation.

Simulations. For the simulations shown in **Figure 1**, we used genotypes from the Wellcome Trust Case Control Consortium (WTCCC)[48]. Quality control was performed as described in Gusev *et al.*[14]: we removed any SNPs that were below a MAF of 0.01, had missingness above 0.002 or deviated from Hardy-Weinberg equilibrium at $P < 0.01$. The resulting data set had 14,526 individuals and 162,574 SNPs. We let heritability vary from 0.1 to 0.9, with the proportion of causal SNPs equal to 0.05 or 0.005 (8,129 and 813 causal SNPs on average, respectively), and we simulated quantitative phenotypes from an additive model. For each simulation, the effect sizes for causal SNPs were drawn from a normal distribution with a mean of zero and variance (average per-SNP heritability) determined by functional categories. To simulate realistic enrichment for the 53 categories in the baseline model plus the CNS cell type group, we fit the model to the schizophrenia summary statistics[18] and took the resulting coefficients, replacing negative coefficients with 0. We then scaled these coefficients as needed to give the desired heritability at the desired level of polygenicity. For each simulation, we used stratified LD score regression with the full baseline model plus the CNS cell type group to estimate total heritability, the heritability of the CNS cell type group and the proportion of heritability in the CNS cell type group.

For the simulations shown in **Figure 2**, for computational ease using REML, we decreased our sample size to the 2,680 samples in the NBS and 1966 Birth Cohort control cohorts of the WTCCC1 data set, and we correspondingly restricted ourselves to only SNPs on chromosome 1. For this set of simulations, a dense set of SNPs was particularly important, so we used genotypes imputed to integrated phase 1 v3 of the 1000 Genomes Project[47] (see URLs), giving us 360,106 SNPs after quality control. We again simulated quantitative phenotypes using an additive model, with the effect sizes of causal SNPs drawn from a normal distribution with a mean of zero and variance determined by functional categories. Heritability was set to 0.5, and all SNPs were causal unless in a category simulated to have zero variance.

For the simulations shown in **Figure 3**, we began with the simulations of realistic enrichment in the baseline categories and the CNS cell type group as in **Figure 1**. Then, for each other cell type group, we removed the CNS cell type group and added the new cell type group to the model, scaling the coefficient τ_c of the new cell type group to keep the total heritability constant. We then increased the coefficients of the cell type groups by a multiplicative constant so that the average top z score over 5,000 simulations (10 cell type groups × 500 replicates each) was close to the mean top z score found in our analysis of 17 real traits. In a second set of simulations, we decreased the coefficients so that the top cell type group was significant 50% of the time. We then repeated the process with the H3K4me3 fetal brain annotation (although with just one annotation instead of ten cell type groups). First, we fit a model with this annotation plus the baseline model to the schizophrenia summary statistics[18]. We then scaled the coefficient of the cell type–specific annotation until the mean z score over 500 replicates matched the mean z score in real data. In a second set of simulations, we decreased the coefficient so that the top cell type group was significant in 50% of 500 replicates.

Meta-analysis across traits. We chose nine phenotypes with low phenotypic correlation and sample overlap: height, BMI, age at menarche, LDL levels, coronary artery disease, schizophrenia, educational attainment, smoking behavior and rheumatoid arthritis (**Supplementary Note**). We performed a random-effects meta-analysis of the proportion of heritability over the nine phenotypes listed above for each functional category. The results are shown in **Figure 4** and **Supplementary Table 5**. Results from meta-analysis over all 17 traits are shown in **Supplementary Figure 9**; however, these results have artificially deflated standard errors because of correlated traits such as HDL, LDL and triglyceride levels being treated as independent.

Robustness to derived allele frequency. Stratified LD score regression is based on the assumption that the effect size per normalized genotype of a SNP is drawn i.i.d. (independently and identically distributed) with a mean of zero, conditioned on functional annotation. Thus, if allele frequency bins are not included as annotations in the model, then we assume that per-allele effect sizes have variance proportional to $(p(1-p))^{-1}$ for allele frequency p.

To check that our results were not affected by allele frequency–dependent genetic architecture, we repeated the meta-analysis over traits using the full baseline model with seven derived allele frequency bins as extra annotations. This allowed for effect size to depend on derived allele frequency, independently of functional annotation. The results were very similar to our results without bins of derived allele frequency (**Supplementary Table 7**).

In this report, we do not consider heritability from very rare SNPs. If stratified LD score regression were to be used to analyze a data set with rare variants, there would be several issues to consider that did not come up in our analysis. For example, in the current analysis, we could use LD estimates from a reference panel because the LD patterns in the reference panel matched the LD patterns in our samples for the allele frequency range in which we were interested; this might not hold for rare variants[49]. Also, our analysis showed that allele frequency–dependent architectures do not cause bias in our current analyses, but this robustness might not extend to potential future analyses of data sets with rare variants.

Comparison to other methods. We are not aware of any other methods designed to estimate genome-wide components of heritability from summary statistics. However, there are existing methods that identify enriched functional categories and cell types from summary statistics. We compared our method to four other methods, described below; each of these methods has provided valuable biological insights. For each of these methods, we assessed the rejection rate over 100 simulations for true cell type–specific enrichment, null baseline enrichment (baseline enrichment with no cell type–specific signal) and null simulations with no enrichment in any category. We performed this analysis for both a cell type (fetal brain for H3K4me3) and a cell type group (CNS) and for two proportions of causal SNPs, 0.05 and 0.005. All simulations had a sample size of 14,000 and h_g^2 of 0.7. Results are displayed in **Figure 7**; below, we discuss the results for each method individually.

GoShifter is a recent method of Trynka *et al.*[6] (see also their previously published work[5]). GoShifter is conservative in its identification of enrichment, comparing to a null obtained by local shifting rather than a genome-wide null, and it only uses statistically significant SNPs. It had properly calibrated type I error in all four situations we performed. In these four situations, stratified LD score regression had higher power than GoShifter in the more polygenic scenarios, and the two methods performed comparably in the less polygenic scenarios, in which there were more significant SNPs.

A method by Pickrell[9] combines GWAS data with functional data to identify enriched and depleted functional categories and leverages the resulting model to increase GWAS power. The method, called fgwas, is effective at increasing association mapping power and identifies many interesting enrichments in the published paper. In our simulations, we saw good null calibration but low power to detect enrichment. In the four simulations with true enrichment, fgwas performed best when identifying enrichment of the smaller category (fetal brain) in the more polygenic trait ($p_{\text{causal}} = 0.05$); however, stratified LD score regression had higher power than fgwas in all four situations. fgwas could have an advantage for annotations smaller than the ones tested in this manuscript, but we do not explore that possibility here.

Maurano *et al.*[10] use enrichment of SNPs passing P-value thresholds of increasing stringency to identify important cell types. Using this method, Maurano *et al.*[10] found striking patterns of cell type–specific enrichment. However, this approach implicitly assumes that the functional annotation of a GWAS SNP matches the functional annotation of the causal SNP, which could be true for functional annotations composed of very wide regions but is not likely to be true for functional annotations composed of smaller regions, such as conserved regions. Moreover, the method does not account for total LD and so could give biased results if used to compare functional annotations with different average amounts of total LD[1]. We implemented a 'top SNPs' method analogous to the method of Maurano *et al.*[10] that tests for enrichment of the functional category among SNPs that pass statistical significance. Because the method is not intended to control for any other annotations, it had a high rejection rate for the null baseline simulations, detecting cell type–specific signal where there was none. Thus, its high rejection rate for the cell type–specific simulations was not reflective of true power. It remains a powerful method for traits with many significant SNPs, if the goal does not include controlling for other categories.

Similarly, PICS, a recent method from Farh *et al.*[7], focuses on fine mapping and considers only genome-wide significant loci. On real data[7], the results from this method were compelling and consistent with biology. This method performed similarly to the top SNPs method in our simulations, with a high rejection rate in null simulations with baseline enrichment and also a high rejection rate for true enrichment.

In addition to stratified LD score regression as used in this manuscript for cell type–specific analyses, we also performed unadjusted stratified LD score regression, that is, LD score regression used to test for enrichment in the total proportion of heritability, not controlling for other methods, in a way analogous to the top SNPs and PICS methods. As expected, this unadjusted version had a high rejection rate both for null baseline enrichment and true cell type–specific signal, for the same reasons that the top SNPs and PICS methods did.

Of the three methods with properly calibrated rejection rates for the null simulations with baseline enrichment (GoShifter, fgwas and stratified LD score regression), stratified LD score regression was the most powerful for the polygenic traits. For the less polygenic traits, stratified LD score regression had power similar to GoShifter for the cell type group, and none of the three methods had any power for the single cell type with less polygenic genetic architecture.

In very recent work, Kichaev *et al.*[8] introduced a new method (PAINTOR) that leverages functional data for improved fine mapping. The method also outputs annotations associated with disease. Although the method is clearly effective in increasing fine-mapping resolution, it is unclear whether the method is effective at ranking cell types; for example, the cell types identified as contributing the most to HDL, LDL and triglyceride levels (using data from Teslovich *et al.*[27]) were muscle, kidney and fetal small intestine, respectively, whereas the top cell type was liver for all three phenotypes when using our method (also using data from Teslovich *et al.*[27]). The uncertain effectiveness of this method in ranking cell types may be due to it being primarily aimed at fine mapping and thus considering only genome-wide significant loci.

46. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
47. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
48. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **446**, 661–663 (2007).
49. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).

doi:10.1038/ng.3404