ANALYSIS

nature genetics

Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis

Po-Ru Loh[1,2], Gaurav Bhatia[1,2], Alexander Gusev[1,2], Hilary K Finucane[3], Brendan K Bulik-Sullivan[2,4], Samuela J Pollack[1,2,5], Schizophrenia Working Group of the Psychiatric Genomics Consortium[6], Teresa R de Candia[7], Sang Hong Lee[8,9], Naomi R Wray[8], Kenneth S Kendler[10], Michael C O'Donovan[11], Benjamin M Neale[2,4], Nick Patterson[2] & Alkes L Price[1,2,5]

Heritability analyses of genome-wide association study (GWAS) cohorts have yielded important insights into complex disease architecture, and increasing sample sizes hold the promise of further discoveries. Here we analyze the genetic architectures of schizophrenia in 49,806 samples from the PGC and nine complex diseases in 54,734 samples from the GERA cohort. For schizophrenia, we infer an overwhelmingly polygenic disease architecture in which ≥71% of 1-Mb genomic regions harbor ≥1 variant influencing schizophrenia risk. We also observe significant enrichment of heritability in GC-rich regions and in higher-frequency SNPs for both schizophrenia and GERA diseases. In bivariate analyses, we observe significant genetic correlations (ranging from 0.18 to 0.85) for several pairs of GERA diseases; genetic correlations were on average 1.3 tunes stronger than the correlations of overall disease liabilities. To accomplish these analyses, we developed a fast algorithm for multicomponent, multi-trait variance-components analysis that overcomes prior computational barriers that made such analyses intractable at this scale.

Over the past 5 years, variance-components analysis has had considerable impact on research in human complex trait genetics, yielding rich insights into the heritable phenotypic variation explained by SNPs[1–3], its distribution across chromosomes, allele frequencies and functional annotations[4–6], and its correlation across traits[7,8]. These analyses have complemented GWAS: whereas GWAS have identified individual loci explaining significant portions of trait heritability, variance-components methods have aggregated signal across large SNP sets, finding information about polygenic effects invisible to association studies. The usefulness of both approaches has been particularly clear in studies of schizophrenia, for which early GWAS achieved few genome-wide significant findings yet variance-components analysis indicated a large fraction of heritable variance spread across common SNPs in numerous loci, over 100 of which have now been discovered in large-scale GWAS[5,9–12].

Despite these advances, much remains unknown about the genetic architecture of schizophrenia and other complex diseases. For schizophrenia, known GWAS-identified loci collectively explain only 3% of variation in disease liability[12]; of the remaining variation, a sizable fraction has been shown to be hidden among thousands of common SNPs[5,11], but the distribution of these SNPs across the genome and the allele frequency spectrum remain uncertain. Even for traits such as lipid levels and type 2 diabetes for which loci of somewhat larger effect have been identified, the spatial and allelic distribution of the variants responsible for the bulk of known SNP heritability remains a mystery[13,14]. Variance-components methods have potential to shed light on these questions using the increased statistical resolution offered by tens or hundreds of thousands of samples[15,16]. However, although study sizes have increased beyond 50,000 samples, existing variance-components methods[2] are becoming computationally intractable at such scales. Computational limitations have forced previous studies to split and then perform meta-analysis on data sets[6], a procedure that results in loss of precision for variance-components analysis, which relies on pairwise relationships for inference (in contrast to meta-analysis in association studies)[15,16].

Here we introduce a much faster variance-components method, BOLT-REML, and apply it to analyze ≈50,000 samples in each of two very large data sets—the Psychiatric Genomics Consortium (PGC2)[12] and Genetic Epidemiology Research on Aging (GERA; see URLs)—obtaining several new insights into the genetic architectures of schizophrenia and nine other complex diseases. We harnessed the computational efficiency and versatility of BOLT-REML variance-components analysis to estimate components of heritability,

[1]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts, USA. [5]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. [6]A full list of members is provided in the **Supplementary Note**. [7]Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, Colorado, USA. [8]Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia. [9]School of Environmental and Rural Science, University of New England, Armidale, New South Wales, Australia. [10]Department of Psychiatry and Human Genetics, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, USA. [11]Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK. Correspondence should be addressed to P.-R.L. (loh@hsph.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

Received 5 June; accepted 2 October; published online 2 November 2015; doi:10.1038/ng.3431

infer levels of polygenicity, partition SNP heritability across the common allele frequency spectrum and estimate genetic correlations among GERA diseases. We corroborated our results using an efficient implementation of PCGC regression[17] when computationally feasible.

RESULTS

Overview of the methods

The BOLT-REML algorithm employs the conjugate gradient-based iterative framework for fast mixed-model computations[18,19] that we previously harnessed for mixed-model association analysis using a single variance component[20]. In contrast to that work, BOLT-REML robustly estimates variance parameters for models involving multiple variance components and multiple traits[21,22]. BOLT-REML uses a Monte Carlo average information restricted maximum-likelihood (AI REML) algorithm[23], which is an approximate Newton-type optimization of the restricted log likelihood[24] with respect to the variance parameters being estimated. In each iteration, BOLT-REML rapidly approximates the gradient of the log likelihood using pseudorandom Monte Carlo sampling[25] and approximates the Hessian of the log likelihood using the average information matrix[26]. Full details, including simulations verifying the accuracy of BOLT-REML heritability parameter estimates and standard errors (which are nearly identical to those from standard REML), are provided in the Online Methods and the **Supplementary Note**. We have released open source software implementing BOLT-REML (see URLs).

Efficiency of BOLT-REML variance-components analysis

We assessed the computational performance of BOLT-REML, comparing it to GCTA software[2] (see URLs) in REML analyses of GERA disease phenotypes on subsets of the GERA cohort. We observed that, across three types of analyses, BOLT-REML achieved order-of-magnitude reductions in running time and memory use in comparison to GCTA, with relative improvements increasing with sample size (**Fig. 1**). The running times we observed for BOLT-REML scaled roughly as $\approx MN^{1.5}$ for M SNPs and N samples, consistent with previously reported empirical results for BOLT-LMM association analysis[20]; in contrast, standard REML analysis required $O(MN^2 + N^3)$ running time (**Fig. 1a** and **Supplementary Table 1**). BOLT-REML also only required $\approx MN/4$ bytes of memory (with the amount nearly independent of the number of variance components used), in contrast to standard REML analysis, which required $O(N^2)$ memory per variance component (**Fig. 1b** and **Supplementary Table 1**). Consequently, GCTA could only analyze at most half the cohort; indeed, computational constraints have forced previous studies to split large cohorts for analysis[6], increasing standard errors. In contrast, BOLT-REML enabled us to perform a full suite of heritability analyses on $N = 50{,}000$ samples with tight error bounds[15,16].

SNP heritability of schizophrenia and GERA diseases

We analyzed 22,177 schizophrenia cases and 27,629 controls with well-imputed genotypes at 472,178 markers of minor allele frequency (MAF) ≥2% in the PGC2 data[12] (**Supplementary Table 2**) and nine complex diseases in 54,734 randomly ascertained samples typed at 597,736 SNPs in the GERA cohort (Online Methods; quality control included filtering to unrelated European-ancestry samples and pruning markers by linkage disequilibrium (LD) to $r^2 \leq 0.9$). To remove possible effects from population stratification, all analyses included ten principal-component covariates; PGC2 analyses further included 29 study indicators. We estimated liability-scale SNP heritability (h^2_{g}; ref. 1) for schizophrenia in the PGC2 data set and all 22 disease phenotypes in the GERA data set assuming a liability-threshold model; we assumed a schizophrenia population risk of 1% (refs. 5,11,12), and we assumed that GERA disease population risks matched case fractions in the GERA cohort. For GERA diseases, we estimated h^2_{g} by applying BOLT-REML directly to observed case-control status—obtaining raw observed-scale heritability parameter $h^2_{\text{g-cc}}$ estimates—and then converting $h^2_{\text{g-cc}}$ to liability-scale h^2_{g} using the linear transformation from ref. 3 (**Table 1** and **Supplementary Table 3**). Given the very low values of $h^2_{\text{g-cc}}$ for many GERA diseases, we restricted further GERA analyses to the nine individual diseases with the highest $h^2_{\text{g-cc}}$ values (**Table 1**). For schizophrenia, we estimated h^2_{g} by developing and applying a computationally efficient implementation of PCGC regression[17] (see URLs and Online Methods) in light of the known downward bias of large-sample REML h^2_{g} estimates for ascertained case-control traits[17,27]. Indeed, upon performing REML analyses on full data sets as well as on random 2–10× subsamples, we observed significant downward bias of schizophrenia h^2_{g} estimates with increasing sample size, whereas we observed no such trend in GERA, which is a randomly ascertained cohort study (**Supplementary Table 4**). REML h^2_{g} estimates on PGC2 data down-sampled by 10× ($N \approx 5{,}000$) were consistent with the PCGC regression estimate (**Supplementary Table 4**).

These analyses help explain a previously mysterious observation of decreasing h^2_{g} estimates for schizophrenia with increasing aggregation of cohorts[5]. This phenomenon was attributed to phenotypic heterogeneity[5,11], as suggested by estimates of between-cohort genetic correlation of <1 (ref. 5). Our analyses implicate ascertainment-induced downward bias of estimated h^2_{g} (worsening with increasing sample size) as an additional explanation of this effect (**Supplementary**

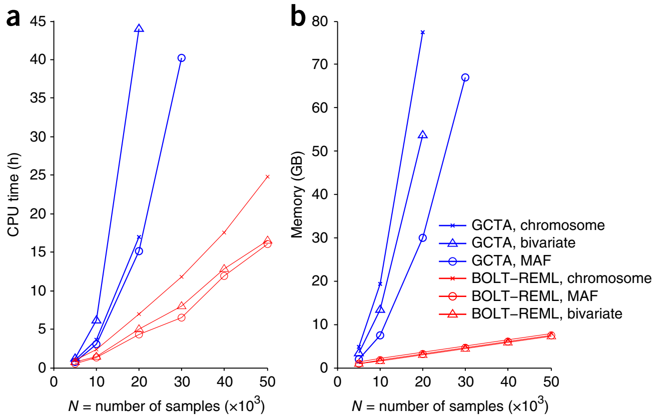


Figure 1 Computational performance of the BOLT-REML and GCTA heritability analysis algorithms. Benchmarks of the BOLT-REML and GCTA algorithms are shown for three heritability analysis scenarios, including analysis with heritability partitioned across 22 chromosomes, analysis with heritability partitioned across six MAF bins and bivariate analysis. (**a**,**b**) Run times (**a**) and memory (**b**) are plotted for runs on subsets of the GERA cohort with a fixed SNP count of $M = 597{,}736$ and increasing sample size (N), using dyslipidemia as the phenotype in the univariate analyses and hypertension as the second phenotype in the bivariate analyses. Reported run times are the medians for five identical runs using one core of a 2.27-GHz Intel Xeon L5640 processor. Reported run times for GCTA are the total time required to compute the genetic relationship matrix (GRM) and perform REML analysis; time breakdowns and numeric data are provided in **Supplementary Table 1**. Data points not plotted for GCTA represent scenarios in which GCTA required more memory than the 96 GB available. Software versions: BOLT-REML, v2.1; GCTA, v1.24.

Table 1 Estimated proportions of variance in disease liability explained by SNPs

Disease	Cases	Controls	h_g^2 (SE)
Schizophrenia	22,177	27,629	0.274 (0.007)
Allergic rhinitis	13,437	41,297	0.074 (0.015)
Asthma	8,929	45,805	0.152 (0.018)
Cardiovascular disease	14,861	39,873	0.092 (0.015)
Diabetes type 2	6,845	47,889	0.297 (0.022)
Dyslipidemia	29,511	25,223	0.263 (0.014)
Hypertension	27,921	26,813	0.255 (0.014)
Macular degeneration	3,700	51,034	0.242 (0.029)
Osteoarthritis	19,832	34,902	0.098 (0.014)
Osteoporosis	5,337	49,397	0.195 (0.024)

Schizophrenia cases and controls are from the PGC2 data set[12]; the h_g^2 estimate assumes a population risk of 1% and was computed using PCGC regression to avoid REML bias induced by overascertainment of cases[17,27]. Cases and controls for the other nine diseases are from the GERA data set; h_g^2 estimates assume random sample ascertainment and were computed using BOLT-REML.

Tables 4 and **5**). In theory, the extent of ascertainment-induced bias could be used to infer the extent of case overascertainment and hence infer population risk, but we found in simulations that larger sample sizes would be required (**Supplementary Table 6**). Finally, we note that, although our reported schizophrenia h_g^2 estimate assumes a population risk of 1% (refs. 5,11,12), this assumption does not affect estimates of the relative partitioning of SNP heritability across SNP subsets; in the partitioning analyses that follow, h_g^2 serves only as a scale factor (Online Methods). Similarly, although our use of an LD-pruned marker set to alleviate LD bias[28–30] (Online Methods) results in a higher h_g^2 estimate than when using unpruned markers (**Supplementary Table 5**), this choice does not otherwise affect the analyses that follow.

Contrasting polygenicity of schizophrenia and GERA diseases

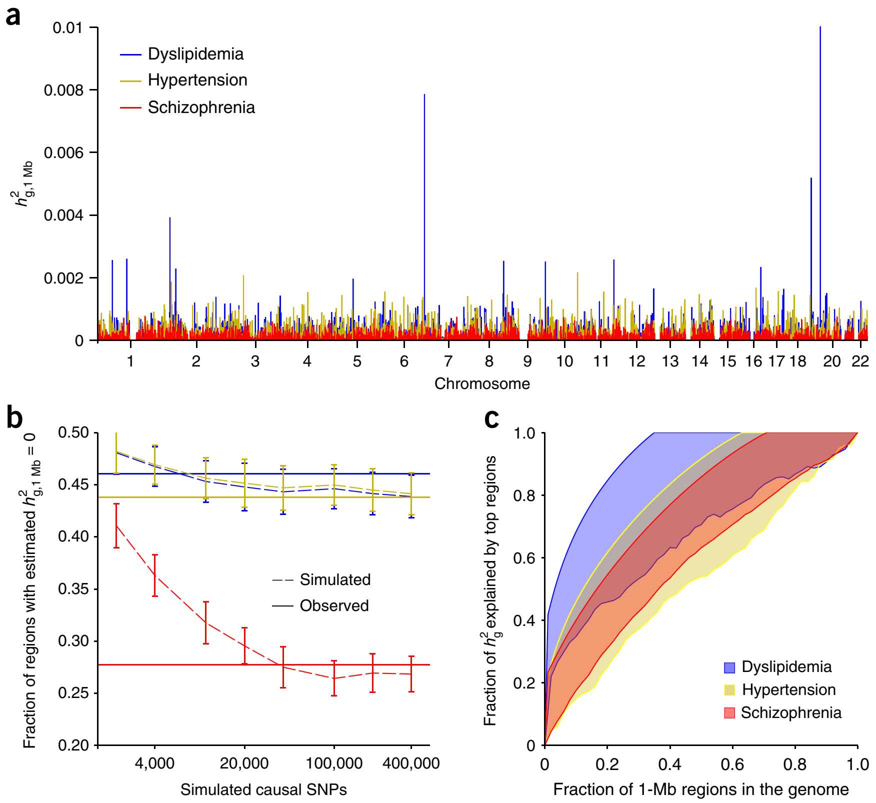
We next performed a detailed investigation of the polygenicity of schizophrenia and the GERA diseases. Specifically, we estimated the SNP heritability explained by each 1-Mb region of the genome, $h^2_{g,1\ \mathrm{Mb}}$ (defined in the Online Methods) (**Fig. 2a**); we confirmed in simulations that 1-Mb regions are sufficiently wide to ensure negligible leakage of heritability across region boundaries due to LD or incomplete tagging of variants (**Supplementary Tables 7** and **8**). We restricted our primary analyses of GERA diseases to dyslipidemia and hypertension, the diseases with the highest observed-scale SNP heritability, $h^2_{\text{g-cc}}$ (**Supplementary Table 3**); we had insufficient statistical power to analyze diseases with lower $h^2_{\text{g-cc}}$ (**Supplementary Fig. 1**). As expected, SNP heritability estimates for individual 1-Mb regions were individually noisy (mean estimated $h^2_{g,1\ \mathrm{Mb}}$ /mean SE ($h^2_{g,1\ \mathrm{Mb}}$) = 0.85 for schizophrenia and 0.51 for dyslipidemia and hypertension, where SE is the standard error), although we did see substantial SNP heritability in some 1-Mb regions (particularly for dyslipidemia, which has SNPs of relatively large effect[13]; in contrast, no 1-Mb region was estimated to explain more than 0.1% of schizophrenia liability). We therefore sought to draw inferences from the bulk distribution of per-megabase SNP heritability estimates (**Supplementary Fig. 2**). We note that a limitation of BOLT-REML is that it does not compute likelihood-ratio test statistics for testing whether individual variance components contribute nonzero variance (**Supplementary Note**).

To understand the effect of different levels of polygenicity on the distribution of per-megabase SNP heritability estimates, we simulated quantitative traits of varying polygenicity (2,000–597,736 causal SNPs) with h_g^2 matching the genome-wide observed-scale $h^2_{\text{g-cc}}$ estimates for schizophrenia, dyslipidemia and hypertension (**Supplementary Table 3**) using PGC2 and GERA genotypes. We then applied the same procedures we used for the real phenotypes to obtain per-megabase SNP heritability estimates, and we compared the simulated distributions of per-megabase heritability to the observed distributions, focusing on the fraction of 1-Mb regions with $h^2_{g,1\ \mathrm{Mb}}$ estimates of zero (**Fig. 2b**). Intuitively, traits that are more polygenic have heritability spread more uniformly across 1-Mb regions and hence have fewer $h^2_{g,1\ \mathrm{Mb}}$ estimates of zero, as our simulations confirmed. On the basis of this statistic, our analyses suggest that schizophrenia has a genetic architecture involving >20,000 causal SNPs; however, we caution that, unlike our analyses below, this estimate is contingent on our parameterization of simulated genetic architectures, as are previous estimates[11,31].

We further interrogated our real and simulated distributions of per-megabase SNP heritability estimates to obtain nonparametric bounds on the cumulative fraction of h_g^2 explained by various numbers of top 1-Mb regions, that is, the regions that harbor the most SNP heritability in the population, for schizophrenia, dyslipidemia and hypertension (**Fig. 2c**). We observed that the probability of observing an $h^2_{g,1\ \mathrm{Mb}}$ estimate of zero for a given 1-Mb region is a convex function of the true SNP heritability of that region (**Supplementary Figs. 3** and **4**), and we harnessed this observation to create an upper bound for the cumulative heritability explained by true top regions (Online Methods). To create a lower bound for this quantity, we applied a cross-validation procedure (similar to the one in ref. 32) in which we selected top regions using subsets of the data and estimated the heritability explained by these regions using left-out test samples (Online Methods). Combining the upper and lower bounds produced conservative 95% confidence intervals for the heritability explained by the top regions (**Fig. 2c**), as we verified in simulations (**Supplementary Fig. 5**). In particular, we inferred that schizophrenia has an extremely polygenic architecture, with most 1-Mb regions (conservative 95% confidence interval (CI) = 71–100%) contributing nonzero SNP heritability and very little concentration of SNP heritability in the top 1-Mb regions, in contrast to dyslipidemia (**Fig. 2c**). Notably, these bounds are not contingent on any particular parametric model of genetic architecture (**Supplementary Fig. 6**): this inference uses simulation data only to interrogate the sampling variance in $h^2_{g,1\ \mathrm{Mb}}$ estimates, which is largely independent of the distribution of heritability across the SNPs in a region (**Supplementary Fig. 4**)[28]. (We report only conservative 95% confidence intervals, without parameter estimates, because obtaining point estimates would require assuming a parameterization of genetic architecture.) We repeated all of these analyses using 0.5-Mb regions and observed no qualitative differences in the results (**Supplementary Figs. 2**, **3** and **7**, and **Supplementary Table 7**).

Having computed per-megabase SNP heritability estimates, we checked for correlations between estimated $h^2_{g,1\ \mathrm{Mb}}$ and genomic annotations that vary slowly across the genome. Specifically, we tabulated the GC content, genic content[6], replication timing[33], recombination rate[34], background selection[35] and methylation quantitative trait loci (meQTLs)[36] per megabase of the genome. Each of these annotations had autocorrelation of r^2 >0.3 across consecutive 1-Mb segments (**Supplementary Table 9**). For each disease (schizophrenia, dyslipidemia and hypertension), we observed the greatest correlation with GC content ($P < 1 \times 10^{-5}$) (**Supplementary Table 10**). We also observed significant correlations of per-megabase SNP heritability with genic content, replication timing and recombination rate; however, upon including GC content—which is correlated with each of the other annotations (**Supplementary Table 11**)—as a covariate, all other correlations became non-significant (**Supplementary Table 10**). To further investigate this finding, we stratified 1-Mb regions into

Figure 2 Extreme polygenicity of schizophrenia in comparison to other complex diseases. (**a**) Manhattan-style plots of estimated SNP heritability per 1-Mb region of the genome, $h^2_{\mathrm{g,1\,Mb}}$, for dyslipidemia, hypertension and schizophrenia. The *APOE* region of chromosome 19 is an outlier with an $h^2_{\mathrm{g,1\,Mb}}$ estimate of 0.022. (**b**) Fractions of 1-Mb regions with estimated $h^2_{\mathrm{g,1\,Mb}}$ equal to its lower-bound constraint of zero in disease phenotypes (solid lines) and simulated phenotypes with varying degrees of polygenicity and with h^2_{g} matching the $h^2_{\mathrm{g\text{-}cc}}$ of each disease (dashed lines). The simulation data plotted are the means over five simulations; error bars, 95% prediction intervals assuming Bernoulli sampling variance and taking into account standard error. (**c**) Conservative 95% confidence intervals for the cumulative fractions of SNP heritability explained by the 1-Mb regions that contain the most SNP heritability. Lower bounds are from a cross-validation procedure involving only the disease phenotypes, and upper bounds are inferred from the empirical sampling variance in $h^2_{\mathrm{g,1\,Mb}}$ estimates (Online Methods).



quintiles on the basis of GC content and partitioned SNP heritability across the strata, observing a clear enrichment of heritability with increasing GC content (**Fig. 3**), which we verified was not due to systematic differences in SNP count or MAF distribution across the GC quintiles (**Supplementary Fig. 8** and **Supplementary Table 12**) and was not explained by differences in meQTL counts (**Supplementary Fig. 9**). To quantify this enrichment, we performed finer partitioning of 1-Mb regions into 50 GC strata and regressed SNP heritability estimates against GC content (Online Methods). We found that a 1% increase in GC content (relative to the median) corresponded to a 1.0%, 4.4% and 3.2% increase in heritability explained (relative to the mean) for schizophrenia, dyslipidemia and hypertension, respectively (95% CI = 0.3–1.6%, 2.1–6.7% and 1.8–4.6%). Again, repeating these analyses using 0.5-Mb regions produced no qualitative differences in the results (**Supplementary Fig. 10** and **Supplementary Tables 10** and **11**). We also observed that including ten principal-component covariates per variance component or applying extremely stringent quality control had a negligible impact on our results (**Supplementary Table 13**). Likewise, repeating our analyses using PCGC regression instead of BOLT-REML produced consistent results with slightly larger standard errors (**Supplementary Table 13**).

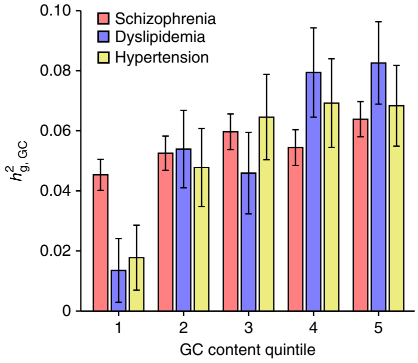


Figure 3 SNP heritability of disease liabilities partitioned by GC content. GC content was computed at 1-Mb resolution, after which 1-Mb regions were stratified into GC content quintiles for variance-components analysis. Quintiles 1–5 have median GC contents of 35.7%, 38.1%, 40.2%, 42.8% and 47.2%, respectively. Error bars, 95% confidence intervals based on REML analytic standard errors.

Finally, we performed chromosome partitioning of SNP heritability for each disease, as previously done for schizophrenia using N = 21,258 samples[5]. We confirmed a strikingly linear relationship between the per-chromosome SNP heritability of schizophrenia and chromosome length (**Supplementary Fig. 11**), consistent with a highly polygenic disease architecture. In contrast, the trend for dyslipidemia was noticeably less linear, consistent with the existence of large-effect loci (**Supplementary Fig. 11**).

Enrichment of SNP heritability in higher-frequency SNPs

Given the high observed-scale heritability of schizophrenia in the full data set with N = 49,806 samples (**Supplementary Table 3**), we reasoned that analyses partitioning SNP heritability for schizophrenia by allele frequency would yield tight partitioning estimates, providing greater resolution than previous inferences using N = 21,258 samples[5]. To calibrate this analysis (accounting for incomplete LD between tagging SNPs and true causal SNPs), we first ran MAF-partitioned heritability analyses of simulated quantitative phenotypes based on UK10K sequencing data (see the Online Methods and URLs). We simulated genetic architectures in which causal SNPs were drawn from SNPs with MAF $p \geq 0.1\%$ and were randomly assigned allele effect sizes with variances proportional to $(p(1-p))^{\alpha}$ for various values of α between −1 and 0 (refs. 28,29) (Online Methods). Under this parameterization, $\alpha = -1$ corresponds to a model in which rare SNPs have larger per-allele effects, such that all SNPs have the same expected

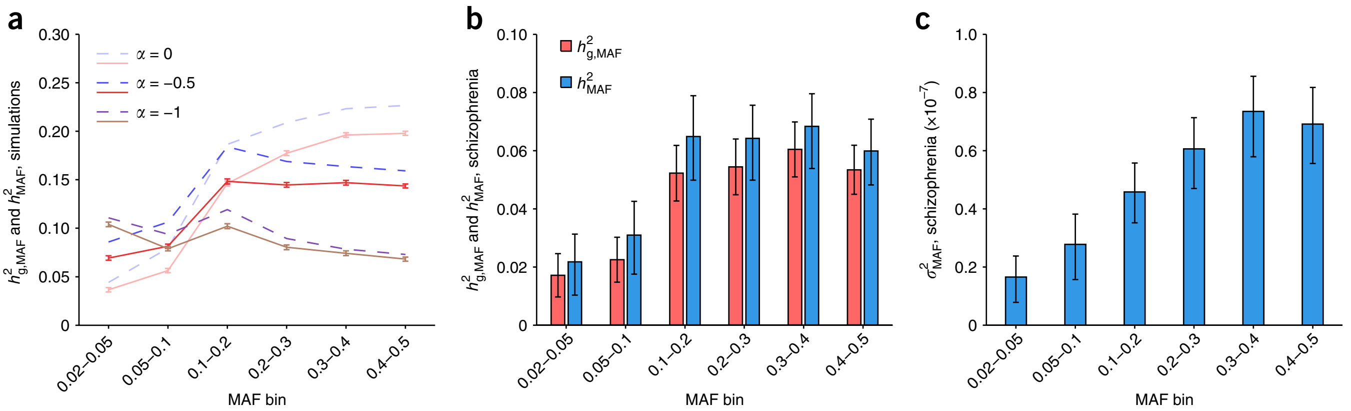


Figure 4 Inferred heritability of schizophrenia liability due to SNPs of various allele frequencies. (**a**) Simulated narrow-sense heritability per MAF bin (h^2_{MAF}; dashed blue curves) and estimated SNP heritability per MAF bin ($h^2_{g,MAF}$; solid red curves) for quantitative phenotypes with genetic architectures in which SNPs of MAF p have average per-allele effect size variance proportional to $p(1 - p)^{\alpha}$. Simulations used causal SNPs with MAF ≥0.1% in UK10K sequencing data and tagging SNPs from our PGC2 analyses; error bars, 95% confidence intervals based on 4,000 runs. (**b**) SNP heritability (red) and inferred narrow-sense heritability (blue) of schizophrenia liability partitioned across six MAF bins. Point estimates of narrow-sense heritability per bin are based on interpolated values of the $h^2_{g,MAF} / h^2_{MAF}$ ratio at $\alpha = -0.28$, which provided the best weighted least-squares fit between observed $h^2_{g,MAF}$ and interpolated $h^2_{g,MAF}$ for the simulations in **a** (**Supplementary Fig. 12**). (**c**) Inferred narrow-sense heritability of schizophrenia liability explained per SNP in each MAF bin, that is, h^2_{MAF} in **b** normalized by UK10K SNP counts (**Supplementary Table 14**). Schizophrenia $h^2_{g,MAF}$ error bars, 95% confidence intervals based on REML analytic standard errors. Schizophrenia h^2_{MAF} and σ^2_{MAF} error bars, unions of 95% confidence intervals assuming $-1 \le \alpha \le 0$.

contribution to variance[1], whereas $\alpha = 0$ corresponds to a model with no selection[37] in which all alleles have similar per-allele effects, where rarer SNPs contribute less variance. We performed MAF-partitioned analyses[29] over six MAF bins (partitioning the 2–50% MAF range) using tagging SNPs from the PGC2 data set, and we observed that the heritability captured by tagging SNPs in each bin ($h^2_{g,MAF}$; defined in the Online Methods) accounted for most but not all of the true heritability contributed by causal UK10K variants in each bin (h^2_{MAF}; defined in the Online Methods) (**Fig. 4a**).

We then performed MAF partitioning of schizophrenia h^2_g by running BOLT-REML on the full PGC2 data set with variance components corresponding to the same six MAF bins (**Fig. 4b**). We estimated the total narrow-sense heritability contributed per MAF bin, h^2_{MAF} (**Fig. 4b**), by performing an inverse variance–weighted least-squares fit of observed $h^2_{g,MAF}$ against data from our simulations, interpolated for $-1 \le \alpha \le 0$; this procedure yielded a best-fit value of $\alpha = -0.28$ (jackknife SE = 0.09) (**Supplementary Fig. 12**), from which we inferred h^2_{MAF}. To keep our inferences robust to model parameterization, we computed conservative 95% confidence intervals for h^2_{MAF} (independent of the best-fit α) by taking the union of the 95% confidence intervals assuming different values of α ($-1 \le \alpha \le 0$). Finally, we divided h^2_{MAF} by the number of UK10K SNPs per bin (**Supplementary Table 14**) to estimate the average heritability explained per SNP in each MAF bin, σ^2_{MAF} (**Fig. 4c**), observing a clear increase in per-SNP heritability with increasing allele frequency. Repeating the MAF partitioning using PCGC regression produced consistent results with slightly larger standard errors (**Supplementary Table 13**). We observed the same general trend in analyses of GERA diseases, although the results were noisier because of smaller $h^2_{g\text{-cc}}$ (**Supplementary Fig. 13**).

Genetic correlations across GERA diseases

Because GERA samples were phenotyped for multiple diseases, we also estimated genetic correlations and total correlations (r_g and r_l; defined in the Online Methods) among GERA disease liabilities (**Fig. 5** and **Supplementary Table 15**). We estimated genetic correlations using bivariate BOLT-REML on each pair of case-control traits[7] and total liability-scale correlations using Monte Carlo simulations to match total observed-scale correlations (Online Methods). We first ran the analysis using only our standard set of covariates (age, sex, ten principal components and Affymetrix kit type) (**Fig. 5a**) and then reran the analysis including body mass index (BMI) as an additional covariate (**Fig. 5b**). We verified that, of the nine survey-derived covariates provided with the GERA data set, BMI was the only one relevant to our analysis (**Supplementary Fig. 14**). Interestingly, we observed that adjusting for BMI produced (on average) a 25% (SE = 5%) relative reduction in genetic correlations and a 19% (SE = 3%) relative reduction in total correlations,

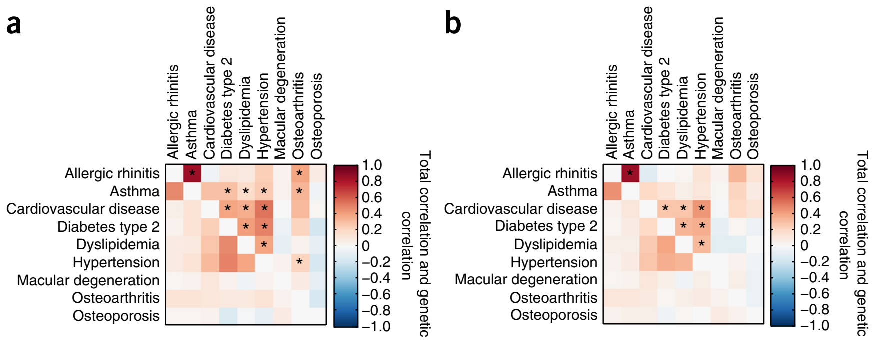


Figure 5 Genetic correlations and total correlations of GERA disease liabilities. (**a**) Correlations from bivariate analyses using only age, sex, ten principal components and Affymetrix kit type as covariates. (**b**) Correlations from bivariate analyses including BMI as an additional covariate. Genetic correlations are shown above the diagonals, and total liability correlations are shown below the diagonals. Asterisks indicate genetic correlations that are significantly positive ($z > 3$), accounting for 36 trait pairs tested. Numeric data, including standard errors, are provided in **Supplementary Table 15**.

as assessed by regressing BMI-adjusted correlations on unadjusted correlations, suggesting that some correlation signal among these diseases is mediated by BMI. Of the 13 significant genetic correlations in the unadjusted analysis, six became non-significant upon adjusting for BMI, leaving a very strong genetic correlation between asthma and allergic rhinitis ($r_g = 0.85$, SE = 0.11) and a cluster of six moderately strong genetic correlations among cardiovascular disease, type 2 diabetes, dyslipidemia and hypertension ($r_g = 0.27$–0.43) (**Supplementary Table 15**).

We further investigated the relationship between genetic correlations (r_g) and total correlations (r_l) among disease liabilities. We observed that r_g significantly exceeded r_l for asthma and allergic rhinitis ($r_g = 0.85$ versus $r_l = 0.46$; $P = 0.008$ when adjusting for 36 hypotheses); no other pair of diseases reached significance. We also observed an approximately linear relationship between genetic correlation and total liability correlation; regressing r_g on r_l yielded a proportionality constant of $r_g/r_l = 1.3$ (SE = 0.1; with the caveat that the 36 trait pairs are not independent), robust to the choice of whether to use BMI as a covariate (**Supplementary Fig. 15**).

DISCUSSION

We have introduced a new fast algorithm, BOLT-REML, for variance-components analysis involving multiple variance components and multiple traits, and we demonstrated that it enables previously intractable large-sample heritability analyses. Such analyses will be essential to attaining the statistical resolution necessary to gain deeper insights into the genetic architecture of complex traits (**Supplementary Table 16**)[15,16]. We have applied BOLT-REML to perform ≈50,000-sample analyses of the PGC2 and GERA data sets, uncovering multiple insights into complex disease architecture, including the extreme polygenicity of schizophrenia, enrichment of SNP heritability in GC-rich regions and in higher-frequency SNPs, and significant genetic correlations among several GERA diseases.

Our per-megabase analyses of SNP heritability in schizophrenia, dyslipidemia and hypertension found contrasting levels of polygenicity, with schizophrenia exhibiting an exceptionally polygenic architecture. Our inference that most 1-Mb regions of the genome (71–100%) contain schizophrenia-associated loci evokes the concern that increasingly powered complex trait GWAS will ultimately implicate the entire genome, becoming uninformative[38]. Recent very large-scale GWAS[12,32,39] have begun grappling with this problem by focusing on biological pathways or gene sets instead of individual SNPs[40]. Although previous studies have provided evidence for a highly polygenic architecture for schizophrenia[9,41], no previous study has quantified polygenicity at the extreme level we have observed here; in light of this result, methods that further interrogate associations at the pathway level will be essential to extracting further biological insights about schizophrenia[42]. This finding also raises the question of whether polygenicity would diminish in analyses with more homogeneous sample recruitment or phenotype (for example, treatment-resistant schizophrenia); future studies may be sufficiently powered to answer this question. As to our observation of enrichment of SNP heritability with increasing GC content, further study will be required to disentangle the mechanisms underlying this phenomenon; previous work has shown that GC architecture has complex effects on recombination and replication timing[33] as well as DNA methylation[43].

Our results partitioning the SNP heritability of schizophrenia and GERA diseases across the 2–50% allele frequency spectrum shed light on the extent to which rarer SNPs tend to have larger per-allele effects, as predicted by evolutionary models[44,45]. Our analysis of schizophrenia, based on well-imputed SNPs with MAF ≥2%, does not assess the contribution of rare variants (MAF <1%) because of the need for stringent quality control in heritability analyses of ascertained case-control cohorts[3]; however, the trend for SNPs with MAFs of 2–50% (**Fig. 4b**,**c**) strongly suggests that rarer SNPs have larger effect sizes per allele yet explain less variance per SNP. Although further study of more phenotypes and rarer variants is needed, this observation implies that the implicit assumption of $\alpha = -1$ made by standard analyses of heritability[1] and mixed-model association[20,27] may be suboptimal, leaving room for further improvement on both fronts.

Our correlation analyses of GERA diseases identified a very strong genetic correlation ($r_g = 0.85$, SE = 0.11) between asthma and allergic rhinitis. Although the link between asthma and allergy has long been known and recent GWAS have identified many shared associations, the extent to which these two diseases are genetically related has not previously been quantified[46–48]. Among other disease pairs, our observation of significant genetic correlations among metabolic diseases confirms and adds resolution to previous estimates[49,50], and our observation of significant broad decreases in genetic and total correlations upon including BMI as a covariate highlights the importance of carefully considering the effects of heritable covariates when conducting and interpreting genetic analyses[51]. Additionally, our empirical observation of an approximately linear relationship between correlations of total liability and genetic correlations[52], viewed in conjunction with a similar (but noisier) empirical observation among a set of seven quantitative metabolic traits[50], suggests the generality of such a trend for human complex traits.

Methodologically, although the variance-components (REML) approach[1] that we have applied and accelerated here enjoys widespread use, three alternative approaches to heritability analysis (with various tradeoffs) have recently been proposed. First, the Bayesian sparse linear mixed model[53] adapts the variance-components approach to better model traits with large-effect loci, slightly reducing standard errors at the expense of much larger computational cost; integrating this approach into BOLT-REML is a potential future direction. Second, PCGC regression[17], which generalizes Haseman-Elston regression[54], is not subject to downward bias under case-control ascertainment; we therefore recommend PCGC regression for the purpose of estimating genome-wide h_g^2 in such situations. (For partitioning SNP heritability across subsets of SNPs, PCGC estimates have slightly higher standard errors than REML.) Third, LD Score regression[49,55] is a very different approach that makes inferences using only GWAS summary statistics—not genotype data. LD Score regression has the disadvantage of somewhat higher standard errors (in comparison to REML) that further increase if inference is desired for small regions of the genome; as such, we are not currently aware of a method for assessing the degree of polygenicity using summary statistics. All of these methods have the limitation that they assume independence of genetic and environmental effects; violation of this assumption may cause bias.

In comparison to existing REML methods, the BOLT-REML algorithm we have proposed is more computationally efficient; however, our approach does have limitations. First, because BOLT-REML achieves its increased speed by avoiding direct computation of likelihoods, it is unable to compute likelihood-ratio tests to assess whether variance parameters are significantly nonzero. In fact, the assumptions underlying REML analytic standard errors break down for parameter estimates of zero (and, more generally, at the parameter space boundary; **Supplementary Note**). GCTA[2] provides an unconstrained optimization feature that allows negative variance estimates, thereby sidestepping this issue and also reducing constraint-induced bias; incorporating such a feature into BOLT-REML is a potential future direction. Second, BOLT-REML, like all REML algorithms, occasionally fails to converge when variance parameters are poorly constrained, typically for multicomponent models at small sample sizes (N <5,000). Given that sample sizes are steadily increasing, however, we expect BOLT-REML to be a robust choice for harnessing the full power of large-scale cohorts to further elucidate complex trait architectures.

URLs. BOLT-REML software and source code (implemented in the BOLT-LMM v2.1 package), http://www.hsph.harvard.edu/alkes-price/software/; GCTA software, http://www.complextraitgenomics.com/software/gcta/; PCGC regression efficient software, http://github.com/gauravbhatia1/PCGCRegression/; PLINK2 software, http://www.cog-genomics.org/plink2; KING software, http://people.virginia.edu/~wc9c/KING/; EIGENSOFT v6.0.1, including open source implementation of FastPCA, http://www.hsph.harvard.edu/alkes-price/software/; GERA data set (database of Genotypes and Phenotypes (dbGaP), phs000674.v1.p1), http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1; UK10K Project, http://www.uk10k.org/.

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 are grateful to K. Galinsky, T. Hayeck, P. Palamara, J. Listgarten, V. Anttila, S. Sunyaev, D. Howrigan, R. Walters, P. Sullivan, M. Keller, M. Goddard, P. Visscher, J. Yang, S. Ripke, D. Golan and S. Rosset for helpful discussions. This research was supported by US National Institutes of Health grants R01 HG006399 and R01 MH101244 and US National Institutes of Health fellowship F32 HG007805. H.K.F. was supported by the Fannie and John Hertz Foundation. Members of the Schizophrenia Working Group of the Psychiatric Genomics Consortium are listed in the **Supplementary Note**. Statistical analyses of PGC2 data were carried out on the Genetic Cluster Computer (http://www.geneticcluster.org/) hosted by SURFsara and financially supported by the Netherlands Scientific Organization (NWO 480-05-003, principal investigator D. Posthuma) along with a supplement from the Dutch Brain Foundation and VU University Amsterdam. Analyses of GERA data were conducted on the Orchestra High-Performance Compute Cluster at Harvard Medical School, which is partially supported by US National Center for Research Resources grant 1S10RR028832-01.

AUTHOR CONTRIBUTIONS

P.-R.L., N.P. and A.L.P. designed experiments. P.-R.L. performed experiments. P.-R.L., G.B., A.G., H.K.F., B.K.B.-S., S.J.P. and A.L.P. analyzed data. All authors wrote the manuscript.

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. 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).
3. Lee, S.H., Wray, N.R., Goddard, M.E. & Visscher, P.M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
4. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
5. Lee, S.H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
6. 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).
7. Lee, S.H., Yang, J., Goddard, M.E., Visscher, P.M. & Wray, N.R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism–derived genomic relationships and restricted maximum likelihood. *Bioinformatics* **28**, 2540–2542 (2012).
8. Lee, S.H. *et al.* Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
9. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
10. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium. Genome-wide association study identifies five new schizophrenia loci. *Nat. Genet.* **43**, 969–976 (2011).
11. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat. Genet.* **45**, 1150–1159 (2013).
12. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
13. Willer, C.J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
14. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nat. Genet.* **46**, 234–244 (2014).
15. Visscher, P.M. *et al.* Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet.* **10**, e1004269 (2014).
16. Visscher, P.M. & Goddard, M.E. A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. *Genetics* **199**, 223–232 (2015).
17. Golan, D., Lander, E.S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. USA* **111**, E5272–E5281 (2014).
18. Legarra, A. & Misztal, I. Computing strategies in genome-wide selection. *J. Dairy Sci.* **91**, 360–366 (2008).
19. VanRaden, P.M. Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).
20. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
21. Henderson, C. *Application of Linear Models in Animal Breeding* (Univ. Guelph, 1984).
22. Henderson, C. & Quaas, R. Multiple trait evaluation using relatives' records. *J. Anim. Sci.* **43**, 1188–1197 (1976).
23. Matilainen, K., Mäntysaari, E.A., Lidauer, M.H., Strandén, I. & Thompson, R. Employing a Monte Carlo algorithm in Newton-type methods for restricted maximum likelihood estimation of genetic parameters. *PLoS ONE* **8**, e80821 (2013).
24. Patterson, H.D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).
25. García-Cortés, L.A., Moreno, C., Varona, L. & Altarriba, J. Variance component estimation by resampling. *J. Anim. Breed. Genet.* **109**, 358–363 (1992).
26. Gilmour, A.R., Thompson, R. & Cullis, B.R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450 (1995).
27. Yang, J., Zaitlen, N.A., Goddard, M.E., Visscher, P.M. & Price, A.L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
28. Speed, D., Hemani, G., Johnson, M.R. & Balding, D.J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–1021 (2012).
29. Lee, S.H. *et al.* Estimation of SNP heritability from dense genotype data. *Am. J. Hum. Genet.* **93**, 1151–1155 (2013).
30. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLoS Genet.* **9**, e1003993 (2013).
31. Stahl, E.A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.* **44**, 483–489 (2012).
32. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46**, 1173–1186 (2014).
33. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
34. International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. *Nature* **449**, 851–861 (2007).
35. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet.* **5**, e1000471 (2009).
36. Banovich, N.E. *et al.* Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.* **10**, e1004663 (2014).
37. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. USA* **111**, E455–E464 (2014).
38. Goldstein, D.B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
39. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
40. Pers, T.H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
41. Gottesman, I.I. & Shields, J. A polygenic theory of schizophrenia. *Proc. Natl. Acad. Sci. USA* **58**, 199–205 (1967).

42. Sullivan, P.F. Puzzling over schizophrenia: schizophrenia as a pathway disease. *Nat. Med.* **18**, 210–211 (2012).
43. Gelfman, S., Cohen, N., Yearim, A. & Ast, G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res.* **23**, 789–799 (2013).
44. Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2011).
45. Lohmueller, K.E. The impact of population demography and selection on the genetic architecture of complex traits. *PLoS Genet.* **10**, e1004379 (2014).
46. Ferreira, M.A. *et al.* Identification of *IL6R* and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006–1014 (2011).
47. Bønnelykke, K. *et al.* Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nat. Genet.* **45**, 902–906 (2013).
48. Hinds, D.A. *et al.* A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. *Nat. Genet.* **45**, 907–911 (2013).
49. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
50. Vattikuti, S., Guo, J. & Chow, C.C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet.* **8**, e1002637 (2012).
51. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96**, 329–339 (2015).
52. Cheverud, J.M. A comparison of genetic and phenotypic correlations. *Evolution* **42**, 958–968 (1988).
53. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genet.* **9**, e1003264 (2013).
54. Haseman, J.K. & Elston, R.C. The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**, 3–19 (1972).
55. Finucane, H.K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

ONLINE METHODS

BOLT-REML algorithm. The overall framework of the BOLT-REML algorithm is Monte Carlo AI REML[23], a Newton-type iterative optimization of the (restricted) log likelihood with respect to the variance parameters sought. BOLT-REML begins a multi–variance component analysis by computing an initial estimate of each parameter using the single–variance component estimation procedure of BOLT-LMM[20] (which is the only analysis possible with BOLT-LMM). Then, in each iteration, BOLT-REML rapidly approximates the gradient of the log likelihood using pseudorandom Monte Carlo sampling[25] and the Hessian of the log likelihood using the average information matrix[26]. BOLT-REML efficiently computes both approximations using conjugate gradient iteration[18,19] with the performance optimizations applied by BOLT-LMM[20]. The approximate gradient and Hessian produce a local quadratic model of the likelihood surface, which we optimize within an adaptive trust region radius—key to achieving robust convergence—to obtain a proposed step. To evaluate the success of the proposed step (that is, to determine whether to accept the step, whether to change the trust region radius and whether the optimization has converged), we introduce a gradient-based approximation to the change in log likelihood achieved by the step. These procedures allow BOLT-REML to consistently achieve convergence in $\approx O(MN^{1.5})$ time; in contrast, existing multicomponent REML algorithms either are less robust or require $O(MN^2 + N^3)$ time (for example, GCTA[2]). Details are described in the **Supplementary Note**.

Accuracy of BOLT-REML variance-components analysis. We verified the accuracy of BOLT-REML analysis by simulating quantitative traits with infinitesimal architectures using genotypes from subsets of the GERA data set and partitioning heritability by chromosome. On a first set of 50,000 simulations using genotypes from $N = 2{,}000$ samples on chromosomes 21 and 22, BOLT-REML correctly estimated components of heritability, computing nearly identical results to GCTA[2] when run with 100 Monte Carlo trials and incurring only 1.03 times higher standard errors when run with 15 Monte Carlo trials (**Supplementary Table 17**), consistent with theory (**Supplementary Note**). On additional sets of 100 simulations using genotypes from $N = 10{,}000$ samples on chromosomes 1 and 2, BOLT-REML correctly estimated genetic correlations in bivariate analyses of simulated quantitative traits[7] (**Supplementary Table 18**) and randomly ascertained case-control traits using a liability-threshold model[3] (**Supplementary Table 19**). Finally, in simulated $N = 50{,}000$ case-control cohorts overascertained for cases (including population stratification and varying polygenicity), we observed that, whereas absolute estimates of heritability were biased downward, as previously demonstrated[17,27], the relative contributions of variance components and their standard errors were still accurately estimated when partitioning heritability by chromosome or MAF (**Supplementary Figs. 16–19**).

PGC2 data set. We analyzed the PGC2 schizophrenia data set[12], applying the following filters. Of 39 European-ancestry cohorts available to us for analysis, we first eliminated the ten cohorts (containing 12% of the available samples) with the lowest numbers of well-imputed SNPs. We further filtered out samples with <90% European ancestry as determined by SNPweights v2.0 (ref. 56). Finally, we extracted an unrelated subset of individuals (pairwise genetic similarity <0.0884) using KING v1.4–unrelated–degree 3; see URLs (refs. 57,58), comprising 22,177 cases and 27,629 controls (**Supplementary Table 2**). Of the imputed genotypes previously computed for each cohort, we restricted to well-imputed autosomal markers (genotype call confidence $P > 0.8$ with <2% missing rate in the cohort), given that stringent quality control is critical to avoid inflated estimates of components of heritability in ascertained case-control data[3]. We then merged the 29 cohorts, taking the union of remaining markers across cohorts and then restricting to markers with total missing rate <5%, leaving 4.4 million markers. We further imposed a >2% MAF threshold based on the imputation quality of typical arrays at low MAF[59], yielding 3.9 million markers in substantial LD, to which we applied two rounds of LD pruning at $r^2 = 0.9$ (PLINK2 (ref. 60); --indep-pairwise 50 5 0.9; see URLs), reducing the number of markers to 596,583 and finally 472,178. Our primary motivation for pruning was to reduce susceptibility of REML h_g^2 estimation to LD bias[28–30]; additionally, pruning reduced computational costs.

GERA data set. We analyzed GERA samples (see URLs; dbGaP study accession phs000674.v1.p1) typed on the GERA EUR chip[59] with phenotypes available for each of 22 disease conditions based on electronic medical records. (Our primary analyses did not include survey-derived phenotypes such as BMI, as the data use conditions stipulated that these phenotypes could only be used as covariates.) We applied similar filters as above, eliminating samples with <90% European ancestry and samples with missing sex and extracting an unrelated subset of 54,734 individuals using PLINK2 (--rel-cutoff 0.05). We removed SNPs deviating from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) and SNPs with missing rate >2%, leaving 597,736 autosomal SNPs.

UK10K data set. Our simulations used UK10K genotypes from sequencing data (see URLs); we merged the ALSPAC and TwinsUK cohorts, intersected marker sets and eliminated multiallelic variants (leaving 18 million variants) and extracted 3,567 unrelated individuals using PLINK2.

Definitions of heritability parameters. We define h_g^2 as the proportion of population variance in disease liability (assuming a liability threshold model[61]) explained by the best linear predictor using typed variants[6]. We call this quantity 'SNP heritability' (ref. 1) (although the set of well-imputed variants in our PGC2 data set included a small fraction of biallelic indels). We define $h_{\mathrm{g,MAF}}^2$ as the proportion of population variance in disease liability explained by the subset of variants in a particular MAF range within the same best linear predictor (jointly fit using all typed variants) and define $h_{\mathrm{g,1\ Mb}}^2$ and $h_{\mathrm{g,chr}}^2$ analogously[6]. We define h^2 as the total narrow-sense heritability—that is, the proportion of population variance explained by the best linear predictor using all variants (including untyped variants)—and we define h_{MAF}^2 as the proportion of population variance explained by all variants in the MAF range (within a predictor using all variants). Finally, we note that we abuse notation slightly by using the above symbols to refer to both true population parameter values and estimates thereof.

Estimating the SNP heritability of disease liabilities. We estimated h_g^2 for each GERA disease by running BOLT-REML on all samples and all markers in our filtered data set. In all our GERA analyses, we adjusted for age, sex, Affymetrix kit type and ten principal-component covariates by residualizing genotypes and phenotypes accordingly. We included principal-component covariates (computed using FastPCA[62]; see URLs) to eliminate phenotypic variance explained by ancestry. We transformed raw REML parameter estimates (denoted $h_{\mathrm{g\text{-}cc}}^2$) to h_g^2 using the linear transformation of ref. 3 assuming case fraction for each GERA disease matched population risk.

For the PGC2 data set, which is overascertained for schizophrenia cases, we estimated h_g^2 using PCGC regression[17] (see below) to avoid ascertainment-induced REML bias[17,27]. In all our PGC2 analyses, we included sex, 29 study indicators and ten principal components as covariates and assumed a schizophrenia population risk of 1% (refs. 5,11,12).

Computationally efficient implementation of PCGC regression. To run PCGC regression on $N = 50{,}000$ samples, we developed a new, efficient software implementation of PCGC regression (see URLs). The new software (i) eliminates in-memory storage of $N \times N$ matrices by accumulating dot products among regressors on the fly (i.e., streaming the GRM inputs); (ii) speeds up jackknife computations (by streaming the GRMs in one pass); and (iii) eliminates storage of 'cleaned' GRMs (i.e., GRMs with principal components projected out) by projecting principal components on the fly.

Partitioning SNP heritability across genomic regions. We estimated per-chromosome $h_{\mathrm{g,chr}}^2$ by running BOLT-REML on all samples and markers using one variance component per chromosome and rescaling raw REML parameter estimates and standard errors by $h_g^2 / h_{\mathrm{g\text{-}cc}}^2$ (**Supplementary Table 3**), noting that relative variance contributions are accurately estimated by REML even under case-control ascertainment (**Supplementary Figs. 16–19**). Estimating per-megabase $h_{\mathrm{g,1\ Mb}}^2$ in an analogous manner would have required fitting a >2,500–variance component model, which was computationally intractable; therefore, we instead performed the computation on contiguous chromosomal segments of up to 100 regions at a time, parallelizing computations using GNU parallel[63]. We used joint multi–variance component analyses rather than fixed-effect analyses of one region at a time to improve robustness against potential confounding (for example, subtle structure or LD between SNPs in nearby windows): any such confounding would contribute to multiple one-region-at-a-time fixed-effect analyses, whereas it is spread across a joint random-effects analysis. (Additionally, we note that fixed-effect regression run on one region at a time would incur strong upward bias: each regressor, even if uncorrelated with the phenotype, can still be used to explain ~1/N of the variance.) For schizophrenia, we used one variance component per 1-Mb region in the segment (discarding regions containing <5 markers) plus a single additional variance component containing all remaining markers. (This approach is similar to that in ref. 64 but computationally cheaper than directly applying the method in ref. 64 using BOLT-REML.) Including all markers in the model was necessary because of ascertainment-induced genome-wide LD among causal variants[27]; we observed that analyses without the all-remaining-markers variance component produced inflated estimates. For the GERA diseases, we did not observe this phenomenon, as expected for a randomly ascertained trait, so for computational efficiency we included only markers in flanking 1-Mb regions in the additional variance component. We ran BOLT-REML with 15 Monte Carlo trials for the extensive computations in this section; we used 100 Monte Carlo trials in all other analyses. We note that we were unable to perform these analyses using PCGC regression because of the disk space requirements of storing 100 different 50,000 × 50,000 GRMs. We also note that the choice of 1 Mb as the window size reflects a tradeoff between fine resolution and the need to preserve a reasonable signal-to-noise ratio of $h^2_{\mathrm{g,1\,Mb}}$ estimates (using N = 50,000 samples) for downstream analyses. A larger sample size would allow decreasing the window size. In the limit of infinite sample size, analysis using one variance component per SNP would theoretically be possible, but, in this limit, the variance-component model would also converge to standard multivariate (fixed-effect) regression.

We estimated $h^2_{\mathrm{g,GC}}$ for each GC quintile by stratifying 1-Mb regions into GC quintiles and running BOLT-REML as above with one variance component per quintile. To obtain finer resolution for regression analyses, we further stratified 1-Mb regions into 50 GC-content strata. We then performed a series of BOLT-REML analyses with one variance component containing the first n strata and a second variance component containing the last 50 − n strata, and we estimated $h^2_{\mathrm{g,GC}}$ of the nth stratum as the difference between the SNP heritability estimates for n and n − 1 strata.

Bounding the SNP heritability explained by the top 1-Mb regions. We bounded the population variance in disease liability explained by the 1-Mb regions with the largest true $h^2_{\mathrm{g,1\,Mb}}$ using the following procedure. We inferred an upper bound by analyzing the observed distribution of $h^2_{\mathrm{g,1\,Mb}}$ estimates and accounting for sampling variance. Explicitly, we analyzed the probability of obtaining a zero $h^2_{\mathrm{g,1\,Mb}}$ estimate, $P(0)$, as a function of the actual value of $h^2_{\mathrm{g,1\,Mb}}$ (relative to its mean). Because of sampling noise and the non-negativity constraint on our REML $h^2_{\mathrm{g,1\,Mb}}$ estimates, $P(0)$ is always positive. In lieu of an analytic formula for $P(0)$ as a function of actual $h^2_{\mathrm{g,1\,Mb}}$, we obtained Monte Carlo estimates of $P(0)$ by simulating quantitative traits (for the samples analyzed, using their actual genotypes) with heritability equal to the $h^2_{\mathrm{g\text{-}cc}}$ of the actual disease status (**Supplementary Table 3**). We distributed heritability across varying numbers of causal variants (13 values ranging from 2,000 random markers to all available markers) and assigned each normalized causal variant a normally distributed effect size, repeating each simulation five times. For each of the 65 simulated traits, we estimated $h^2_{\mathrm{g,1\,Mb}}$ for each 1-Mb region. Combining these data with the actual $h^2_{\mathrm{g,1\,Mb}}$ per region (that is, the sum of squared simulated effect sizes), and, aggregating the data from all simulations and all 1-Mb regions, we obtained a clean empirical estimate of $P(0)$ as a function of actual $h^2_{\mathrm{g,1\,Mb}}$, which we observed was well fit by a sum of two exponentials (**Supplementary Fig. 3**). Although the empirical curve was based on simulation data, it is robust to the genetic architecture used in simulations (for example, varying numbers of causal SNPs and normal versus Laplace effect size distributions; **Supplementary Fig. 4**), as it simply measures the sampling distribution of constrained REML estimates for our genotype data at a given actual $h^2_{\mathrm{g,1\,Mb}}$.

To interpret the observed fraction of zero $h^2_{\mathrm{g,1\,Mb}}$ estimates in light of this information, we harnessed the fact that the decay curve of $P(0)$ versus actual $h^2_{\mathrm{g,1\,Mb}}$ is convex (**Supplementary Fig. 3**). In particular, if a set of 1-Mb regions has a fixed average actual $h^2_{\mathrm{g,1\,Mb}}$, their average $P(0)$ is minimized when all the regions have equal actual $h^2_{\mathrm{g,1\,Mb}}$ (by Jensen's inequality). Conversely, an uneven distribution of actual $h^2_{\mathrm{g,1\,Mb}}$ across regions tends to increase the number of zero $h^2_{\mathrm{g,1\,Mb}}$ estimates. These observations allowed us to bound the maximum fraction of h^2_{g} that could be explained by top 1-Mb regions and still be consistent with the observed fraction of zero $h^2_{\mathrm{g,1\,Mb}}$ estimates. Explicitly, if a certain number of top regions explain SNP heritability $h^2_{\mathrm{g,top}}$, then the sum of $P(0)$ over all regions is minimized by setting $h^2_{\mathrm{g,1\,Mb}}$ of each top region to ($h^2_{\mathrm{g,top}}$ divided by the number of top regions) and $h^2_{\mathrm{g,1\,Mb}}$ of each remaining region to ($h^2_{\mathrm{g}} - h^2_{\mathrm{g,top}}$) divided by the number of non-top regions. We therefore bounded $h^2_{\mathrm{g,top}}$ by requiring this minimum expected number of zero $h^2_{\mathrm{g,1\,Mb}}$ estimates to be at most the observed number of zero $h^2_{\mathrm{g,1\,Mb}}$ estimates (plus 1.96 times its standard error for a conservative 95% confidence bound). We checked the accuracy of this procedure using simulated case-control ascertained data sets with varying numbers of causal SNPs (**Supplementary Fig. 5**).

We obtained lower bounds on the fraction of h^2_{g} explained by top 1-Mb regions by threefold cross-validation. For each fold in turn, we estimated $h^2_{\mathrm{g,1\,Mb}}$ for each region using the remaining two folds, ranked regions accordingly and then estimated the SNP heritability explained by top-ranked regions using the left-out fold. We repeated this procedure three times, obtaining nine estimates per fraction of regions, and computed the mean minus 1.96 times the standard deviation/3 as a conservative 95% confidence lower bound on the SNP heritability explained by top regions. We estimate standard error using standard deviation/3 because the variance of heritability estimates scales with the number of sample pairs (N^2) for $N << M$ (refs. 15,16). This standard error estimate is not theoretically precise because of the complexities of sample reuse in cross-validation[65], but a rough estimate (see **Supplementary Table 4** for empirical support) suffices given that the lower bound is probably a substantial underestimate (that is, very conservative): the finite sample size of the training folds prevents an accurate ranking of regions, especially those contributing small amounts of variance.

Partitioning SNP heritability across allele frequency bins. We computed per–MAF bin $h^2_{\mathrm{g,MAF}}$ estimates in a manner analogous to $h^2_{\mathrm{g,chr}}$ estimates. To infer per–MAF bin h^2_{MAF} explained by untyped as well as typed variants, we ran simulations using UK10K sequencing data to assess the tagging efficiency of our PGC2 and GERA marker sets in various MAF ranges. Specifically, we simulated fully heritable quantitative traits in which normalized SNPs with MAF p ≥0.1% (in the UK10K data) were selected as causal with probability 0.5% and assigned normally distributed effect sizes with variance $(p(1-p))^{\alpha}$. (This setup assumes that UK10K SNPs explain all narrow-sense heritability, but, given that we are only interested in tagging efficiency at MAF ≥2%, our estimation procedure is robust to violations of this assumption. We also note that our choice of a normal distribution of effect sizes is inconsequential given the robustness of REML estimates to a wide range of genetic architectures[28].) We performed 4,000 simulations for each of $\alpha = 0, -0.25, -0.5$ and -1. For each marker set, we then computed REML estimates of h^2_{MAF} for each simulated trait across six MAF bins (**Fig. 4**) using one variance component per bin[29] and restricting to SNPs in the marker set. A small subset of the PGC2 marker IDs (8%) and GERA SNP IDs (4%) were not present among the UK10K SNP IDs, so we did not include these markers in our REML analyses of simulated traits; we verified that the inclusion versus exclusion of these markers had a negligible effect on schizophrenia $h^2_{\mathrm{g,MAF}}$ estimates (**Supplementary Fig. 20**). We performed REML analyses of UK10K simulated traits using a slightly modified version of GCTA v1.21 (ref. 2) to perform robust unconstrained REML (allowing negative $h^2_{\mathrm{g,MAF}}$ estimates); at low sample sizes, constrained REML estimates are upward biased because of noise and the positivity constraint. (We modified GCTA to improve robustness in this setting by adding a trust region framework to its REML optimization.) Finally, we computed $h^2_{\mathrm{g,MAF}}$ for the simulated traits by summing squared simulated effect sizes.

Estimating genetic correlations and total correlations of disease liabilities. For each pair of GERA diseases, we estimated the genetic correlation (denoted r_{g}) directly from bivariate BOLT-REML, which models both genetic and residual covariance, using all samples and markers. Under a liability-threshold model, the estimated genetic correlation (using observed case-control phenotypes) accurately reflects the genetic correlation of underlying disease liabilities, so we did not need to transform raw BOLT-REML r_g parameter estimates[7]. However, the total correlation of observed case-control phenotypes is damped relative to the total correlation of underlying disease liabilities (which we denote by r_l): assuming two diseases have bivariate normal liabilities l_1 and l_2 with correlation r_l, the correlation of case-control phenotypes is $r_p = \mathrm{corr}(l_1 > z_1, l_2 > z_2)$, where z_1 and z_2 are appropriate liability thresholds. In general, $|r_p| \le |r_l|$ under a bivariate normal liability-threshold model; for example, two traits with the same liabilities ($r_l = 1$) but different thresholds ($z_1 \neq z_2$) have $r_p < r_l$. We recovered r_l from r_p by straightforward Monte Carlo simulation, performing a binary search to determine the value of r_l producing the observed r_p assuming values of z_1 and z_2 corresponding to GERA case fractions. Similarly, we obtained a standard error for r_l by transforming the 95% confidence interval for r_p (based on its standard error of $(1-r_p^2)/\sqrt{N}$) in the same way. Finally, we note that, for analyses in which we included BMI (coded on a scale of 1–5 in the GERA data) as a covariate, we included an additional missing indicator covariate marking samples with missing BMI (5%).

56. Chen, C.-Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* **29**, 1399–1406 (2013).
57. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
58. Manichaikul, A. *et al.* Population structure of Hispanics in the United States: the multi-ethnic study of atherosclerosis. *PLoS Genet.* **8**, e1002640 (2012).
59. Hoffmann, T.J. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79–89 (2011).
60. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 1–16 (2015).
61. Falconer, D.S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.* **29**, 51–76 (1965).
62. Galinsky, K.J. *et al.* Fast principal components analysis reveals independent evolution of *ADH1B* gene in Europe and East Asia. *bioRxiv* doi:10.1101/018143 (24 August 2015).
63. Tange, O. GNU Parallel—the command-line power tool. *USENIX* **36**, 42–47 (2011).
64. Kostem, E. & Eskin, E. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *Am. J. Hum. Genet.* **92**, 558–564 (2013).
65. Bengio, Y. & Grandvalet, Y. No unbiased estimator of the variance of *k*-fold cross-validation. *J. Mach. Learn. Res.* **5**, 1089–1105 (2004).