bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Contrasting regional 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[†], Teresa R de Candia[6], Sang Hong Lee[7], Naomi R Wray[7], Kenneth S Kendler[8], Michael C O'Donovan[9], Benjamin M Neale[2,4], Nick Patterson[2], Alkes L Price[1,2,5]

[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] Department of Psychology and Neuroscience, University of Colorado Boulder, Boulder, Colorado, United States.

[7] The Queensland Brain Institute, University of Queensland, Brisbane, Queensland, Australia.

[8] Department of Psychiatry and Human Genetics, Virginia Institute of Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, Virginia, USA

[9] MRC Centre for Neuropsychiatric Genetics and Genomics, Institute of Psychological Medicine and Clinical Neurosciences, Cardiff University, Cardiff, UK

[†] A full list of members is provided in the Supplementary Note.

Correspondence should be addressed to P.-R.L. (loh@hsph.harvard.edu) or A.L.P. (aprice@hsph.harvard.edu).

Heritability analyses of 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 architecture 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 $\geq$71% of 1Mb genomic regions harbor at least one 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) among several pairs of GERA diseases; genetic correlations were on average 1.3x stronger than correlations of overall disease liabilities. To accomplish these analyses, we developed a fast algorithm for multi-component, multi-trait variance components analysis that overcomes prior computational barriers that made such analyses intractable at this scale.

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Over the past five 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 genome-wide association studies (GWAS): while GWAS have identified individual loci explaining significant portions of trait heritability, variance components methods have aggregated signal across large SNP sets, revealing information about polygenic SNP effects invisible to association studies. The utility 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 loci are collectively estimated to 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 across the allele frequency spectrum has remained 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 variants responsible for the bulk of known SNP-heritability has remained a mystery [13, 14]. Variance components methods have the potential to shed light on these questions using the increased statistical resolution offered by tens or hundreds of thousands of samples [15, 16]. However, while study sizes have increased beyond 50,000 samples, existing variance components methods [2] are becoming computationally intractable at such scales. Computational limitations have thus forced previous studies to split and then meta-analyze 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 roughly 50,000 samples in each of two very large data sets—from the Psychiatric Genomics Consortium (PGC2) [12] and the Genetic Epidemiology Research on Aging cohort (GERA; see URLs)—obtaining several new insights into the genetic architectures of schizophrenia

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and nine other complex diseases. We harnessed the computational efficiency and versatility of BOLT-REML variance components analysis to estimate components of heritability, 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 to do so.

Results

Overview of 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 contrast, our previous work [20] used a rudimentary quasi-Newton approach that sufficed only for univariate optimization.) 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 standard REML), are provided in Online Methods and the Supplementary Note. We have released open-source software implementing the method (see URLs).

Computational efficiency of BOLT-REML variance components analysis

We assessed the computational performance of BOLT-REML, comparing it to the GCTA software [2] (see URLs) for REML variance components analyses of GERA disease phenotypes on subsets of the GERA cohort of increasing size. We observed that across three types of analyses, BOLT-REML achieved order-of-magnitude reductions in running time and memory use compared

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to GCTA, with relative improvements increasing with sample size (Figure 1). The running times we observed for BOLT-REML scale roughly as $\approx MN^{1.5}$, consistent with previously reported empirical results for BOLT-LMM association analysis [20], whereas standard REML analysis requires $O(MN^2 + N^3)$ running time (Figure 1a and Supplementary Table 1). BOLT-REML also only requires $\approx MN/4$ bytes of memory (nearly independent of the number of variance components used), in contrast to standard REML analysis, which requires $O(N^2)$ memory per variance component (Figure 1b and Supplementary Table 1). Consequently, GCTA could only analyze at most half of our available samples; indeed, computational constraints have forced previous studies to split large cohorts into multiple subgroups for analysis [6], increasing standard errors and reducing statistical power. In contrast, BOLT-REML enabled us to perform a full suite of heritability analyses of N=50,000 samples with tight error bounds [15, 16].

Estimates of SNP-heritability for 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) $\geq$2% in the PGC2 data [12] (Supplementary Table 2) as well as nine complex diseases in 54,734 randomly ascertained samples typed at 597,736 SNPs in the GERA cohort (see Online Methods; QC procedures included filtering both data sets to unrelated samples of European ancestry and LD-pruning markers to $r^2 \leq 0.9$). To remove possible effects of population stratification, all analyses included 10 principal component covariates and PGC2 analyses further included 29 study indicators (see Online Methods). We computed liability-scale SNP-heritability estimates (h_g^2, 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 schizophrenia population risk of 1% (ref. [5, 11, 12]), and we assumed population risks of GERA diseases matched case fractions in the GERA cohort. For the GERA diseases, we estimated h_g^2 by applying BOLT-REML directly to observed case/control status—obtaining raw observed-scale heritability parameter estimates $h_{g\text{–}cc}^2$—and then converting $h_{g\text{–}cc}^2$ to liability-scale h_g^2 using the linear transformation of ref. [3] (Table 1 and Supplementary Table 3). Given the very low values of $h_{g\text{–}cc}^2$ for many GERA diseases, we restricted further GERA analyses to the nine individual diseases with highest $h_{g\text{–}cc}^2$ (Table 1). For schizophrenia, we estimated h_g^2 by developing and applying a com-

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

putationally efficient implementation of PCGC regression [17] (see URLs and Online Methods) in light of the known downward bias of large-sample REML ${h_g}^2$ estimates for ascertained case-control traits [17, 27]. Indeed, upon performing REML analyses on full data sets as well as on subsamples of each data set with 2x–10x fewer samples, we observed significant downward bias of schizophrenia ${h_g}^2$ estimates with increasing sample size, whereas we observed no such trend for data from GERA, which is a cohort study not subject to case-control ascertainment (Supplementary Table 4). REML ${h_g}^2$ estimates on 10x-downsampled ($N \approx 5{,}000$) PGC2 data corroborated the PCGC regression estimate (Supplementary Table 4), but we believe that PCGC regression is the most appropriate method for estimating genome-wide ${h_g}^2$ in ascertained case-control data.

These analyses help explain a previously mysterious observation of decreasing estimated schizophrenia ${h_g}^2$ with increasing aggregation of cohorts [5]. This phenomenon was attributed to phenotypic heterogeneity [5, 11], as suggested by estimates of between-cohort genetic correlation <1 (ref. [5]). Our analyses implicate ascertainment-induced downward bias of estimated ${h_g}^2$ (worsening with sample size) as an additional explanation of this effect (Supplementary Tables 4 and 5). In theory, the extent of ascertainment-induced bias could be used to infer the extent of case over-ascertainment and hence infer population risk, but we found in simulations that larger sample sizes would be required (Supplementary Table 6). Finally, we note that while our reported schizophrenia ${h_g}^2$ assumes a population risk of 1% (ref. [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, while 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 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 turned to a detailed investigation of the polygenicity of schizophrenia and the GERA diseases. Specifically, we estimated SNP-heritability explained by each 1Mb region of the genome, ${h_{g,1Mb}}^2$ (defined in Online Methods; Fig. 2a); we confirmed in simulations that 1Mb regions are sufficiently wide to ensure negligible leakage of heritability across region boundaries due to linkage

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

disequilibrium 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 $h_{g\text{–}cc}^2$ (Supplementary Table 3), because we had insufficient statistical power to make inferences for diseases with lower $h_{g\text{–}cc}^2$ (Supplementary Fig. 1). As expected, SNP-heritability estimates for individual 1Mb regions were individually noisy (mean estimated $h_{g,1Mb}^2$ / mean s.e.($h_{g,1Mb}^2$) = 0.85 for schizophrenia and 0.51 for dyslipidemia and hypertension), although we did see substantial SNP-heritability in some 1Mb regions (particularly for dyslipidemia, which has relatively large-effect SNPs [13]; in contrast, no 1Mb 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 is does not compute likelihood ratio test statistics for testing whether individual variance components contribute nonzero variance; see 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 (2K–600K causal SNPs) with h_g^2 matching the genome-wide observed-scale $h_{g\text{–}cc}^2$ estimates for schizophrenia, dyslipidemia, and hypertension (Supplementary Table 3) using PGC2 and GERA genotypes. We then applied the same procedures we applied to the real phenotypes to obtain per-megabase SNP-heritability estimates for the simulated traits (Online Methods) and compared the simulated distributions of per-megabase estimates to the observed distributions, focusing on the fraction of 1Mb regions with $h_{g,1Mb}^2$ estimates of zero (Figure 2b). Intuitively, more polygenic traits have heritability spread more uniformly across 1Mb regions and hence have fewer $h_{g,1Mb}^2$ estimates of 0, as our simulations confirmed. (Based on this statistic, our analyses suggest that schizophrenia has a genetic architectures 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 varying numbers of *true* top 1Mb regions—i.e., those that harbor the most SNP-heritability *in the population*—for schizophrenia, dyslipidemia, and hypertension (Figure 2c). We observed that the

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

probability of observing an ${h_{\mathrm{g,1Mb}}}^2$ estimate of zero for a given 1Mb region is a convex function of the true SNP-heritability of that region (Supplementary Figures 3 and 4), and we harnessed this observation to obtain upper bounds on the cumulative heritability explained by true top regions. To obtain lower bounds on this quantity, we applied a cross-validation procedure (similar to ref. [32]) in which we selected top regions using subsets of the data and estimated heritability explained using left-out test samples (see Online Methods). Combining the upper and lower bounds allowed us to obtain conservative 95% confidence intervals for heritability explained by top regions (Figure 2c), as we verified in simulations (Supplementary Fig. 5). In particular, we inferred that schizophrenia has an extremely polygenic architecture, with most 1Mb regions (conservative 95% CI: 71%-100%) containing nonzero contributions to the overall SNP-heritability and very little concentration of SNP-heritability into top 1Mb regions, in contrast to dyslipidemia (Figure 2c). Importantly, 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 of ${h_{\mathrm{g,1Mb}}}^2$ estimates, which is largely independent of the distribution of heritability across SNPs in a region (Supplementary Fig. 4) [28]. (We note that 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 0.5Mb regions and observed no qualitative differences in the results (Supplementary Figures 2, 3, and 7 and Supplementary Table 7).

Having computed per-megabase ${h_{\mathrm{g,1Mb}}}^2$ estimates, we checked for correlations between estimated ${h_{\mathrm{g,1Mb}}}^2$ and genomic annotations that vary slowly across the genome. Specifically, we tabulated GC content, genic content [6], replication timing [33], recombination rate [34], background selection [35], and methylation QTLs [36] per megabase of the genome. (Each of these annotations had an autocorrelation across consecutive 1Mb segments of at least 0.3; see Supplementary Table 9.) For each of schizophrenia, dyslipidemia, and hypertension, we observed the greatest correlation with GC content ($p < 10^{-5}$) (Supplementary Table 10). We also observed significant correlations of per-megabase ${h_{\mathrm{g,1Mb}}}^2$ 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 1Mb regions into GC

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

content quintiles and partitioned SNP-heritability across these strata, observing a clear enrichment of heritability with increasing GC content (Figure 3), which we verified was not due to systematic differences in SNP counts or MAF distributions across GC quintiles (Supplementary Table 12 and Supplementary Fig. 8) and not explained by differences in meQTL counts (Supplementary Fig. 9). To quantify this enrichment, we performed finer partitioning into 50 GC content 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 1.0%, 4.4%, and 3.2% increases in heritability explained (relative to the means) for schizophrenia, dyslipidemia, and hypertension (95% confidence intervals, 0.3–1.6%, 2.1–6.7%, and 1.8–4.6%). Once again, repeating these analyses using 0.5Mb regions produced no qualitative differences in results (Supplementary Fig. 10 and Supplementary Tables 10 and 11). We also observed that including 10 principal component covariates per variance component or applying extremely stringent QC had 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).

Finally, we performed chromosome partitioning of SNP-heritability for each disease, as previously done for schizophrenia using N=21K samples [5]. We confirmed a strikingly linear relationship between SNP-heritability of schizophrenia explained per chromosome 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 on the full N=50K data set (Supplementary Table 3), we reasoned that analyses partitioning schizophrenia SNP-heritability by allele frequency would produce results with small enough standard errors to yield high-confidence conclusions, providing greater resolution than the results of ref. [5] based on N=21K samples. We began by running minor allele frequency (MAF)-partitioned heritability analyses of simulated quantitative phenotypes based on UK10K sequencing data (see Online Methods and URLs). We

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (ref. [28, 29]) (Online Methods). Under this parameterization, $\alpha = -1$ corresponds to a model in which rare SNPs have larger per-allele effects, so that all SNPs have the same expected contribution to variance [1], while $\alpha = 0$ corresponds to a model with no selection [37] in which all alleles have similar per-allele effects, so that on average rarer SNPs contribute less variance. We performed MAF-partitioned analyses [29] over six MAF bins (partitioning the 2–50% MAF range) using tag SNPs from the PGC2 data set, and we observed that the heritability captured by tag SNPs in each bin ($h_{\mathrm{g,MAF}}{}^2$, defined in Online Methods) accounted for most but not all of the true heritability contributed by causal UK10K variants in each bin ($h_{\mathrm{MAF}}{}^2$, defined in Online Methods) (Fig. 4a).

We next performed MAF-partitioning of schizophrenia $h_{\mathrm{g}}{}^2$ by running BOLT-REML on the full PGC2 data set with variance components corresponding to the same six MAF bins (Fig. 4b). We then estimated total narrow-sense heritability contributed per MAF bin, $h_{\mathrm{MAF}}{}^2$ (Fig. 4b), by performing an inverse-variance weighted least-squares fit of observed $h_{\mathrm{g,MAF}}{}^2$ against data from our simulations, interpolated for $-1 \leq \alpha \leq 0$; this procedure yielded a best-fit value of $\alpha = -0.28$ (jackknife s.e.=0.09) (Supplementary Fig. 12), from which we inferred $h_{\mathrm{MAF}}{}^2$. To keep our inferences robust to model parameterization, we computed conservative 95% confidence intervals for $h_{\mathrm{MAF}}{}^2$ (independent of the best-fit α, which is not our focus here) by taking the union of 95% confidence intervals assuming different values of α ($-1 \leq \alpha \leq 0$). Finally, we divided $h_{\mathrm{MAF}}{}^2$ by the number of UK10K SNPs per bin (Supplementary Table 14) to estimate the average heritability explained per SNP in each MAF bin, $\sigma_{\mathrm{MAF}}{}^2$ (Fig. 4c), observing a clear increase in heritability explained per SNP 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 due to smaller $h_{\mathrm{g-cc}}{}^2$ (Supplementary Fig. 13).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genetic correlations across GERA diseases

The availability of multiple phenotypes across all GERA samples also allowed us to estimate the genetic correlations and total correlations (r_{g} and r_l, defined in Online Methods) among disease liabilities (Figure 5 and Supplementary Table 15). We estimated genetic correlations by running bivariate BOLT-REML for each pair of case-control traits [7] and total liability-scale correlations by 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, 10 principal components, and Affymetrix kit type) (Fig. 5a) and then reran the analysis including 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 lowered genetic correlations by a multiplicative factor of 0.75 (s.e.=0.05) and total correlations by a factor of 0.81 (s.e.=0.03), as assessed by regressing BMI-adjusted correlations on unadjusted correlations, suggesting that some correlation signal among these diseases may be 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, s.e.=0.11) and a cluster of six moderate 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 vs. r_l=0.46; p=0.008) after adjusting for 36 hypotheses tested; no other pair 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 (s.e.=0.1, with the caveat that the 36 trait pairs are not independent) robust to the choice of whether or not 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 demonstrated that it enables accurate

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

large-sample heritability analyses that were previously computationally intractable. Such analyses will be essential to attaining the statistical resolution necessary to reveal deeper insights into the genetic architecture of complex traits (Supplementary Table 16) [15, 16]. We have applied BOLT-REML to study human complex diseases in roughly 50K samples from each of the PGC2 and GERA data sets. At this sample size, we uncovered multiple insights into complex disease architecture, including extreme polygenicity of schizophrenia, enrichment of complex disease 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 revealed contrasting levels of polygenicity, with schizophrenia exhibiting an exceptionally polygenic architecture. Our inference that the large majority of 1Mb regions of the genome (71–100%) contain schizophrenia loci evokes the concern that complex-trait GWAS of increasing sample sizes 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]. While previous studies have provided evidence for a highly polygenic architecture for schizophrenia [9, 41], no previous study has provided a quantification of the extreme level of polygenicity we have observed here; in light of this result, methods that further interrogate association signal at the pathway level will be essential to extracting further biological insights about schizophrenia [42]. An additional question that this finding raises is whether the polygenicity would diminish in analyses with more homogeneous sample recruitment or phenotype (e.g., treatment resistant); 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 from 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$\geq$2%, does not assess the contribution of rare variants (MAF$<$1%) due to the need for stringent QC in heritability analyses of ascertained case-

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

control cohorts [3]; however, the trend for SNPs with MAF 2–50% (Fig. 4b,c) strongly suggests that rarer SNPs have larger effect sizes per allele, yet explain less variance per SNP. While 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 disease phenotypes identified a very strong genetic correlation (r_{g}=0.85, s.e.=0.11) between asthma and allergic rhinitis. While 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], while 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 conjuction 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, while the variance components (REML) approach [1] that we have applied and accelerated here enjoys widespread use, three alternative approaches to heritability analysis (with various trade-offs) 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 recommmend PCGC regression for the purpose of estimating genome-wide ${h_{\mathrm{g}}}^2$ 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 inference using only GWAS summary statistics—not genotype data. LD Score regression has the disadvantage of somewhat higher standard errors (vs. REML) that further increase if inference is desired for small regions of the genome; as such, we are not

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

currently aware of a method for assessing 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.

Compared to existing REML methods, the BOLT-REML algorithm we have proposed is much more computationally efficient; however, our approach does have limitations. First, because BOLT-REML achieves its speedup 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; see 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 multi-component models at small sample sizes ($N \ll 5{,}000$). 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, `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000674.v1.p1`.

UK10K project, `http://www.uk10k.org/`.

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 PI: Posthuma) along with a supplement from the Dutch Brain Foundation and the 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 grant NCRR 1S10RR028832-01.

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 success of the proposed step (i.e., 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 multi-component REML algorithms either are less robust or require $O(MN^2 + N^3)$ time (e.g., 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 21–22, 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 1–2, 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=50K

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

case-control cohorts over-ascertained for cases (including population stratification and varying polygenicity), we observed that while absolute estimates of heritability were downward biased, as previously demonstrated [17, 27], relative contributions of variance components and their standard errors were still accurately estimated when partitioning heritability by chromosome or minor allele frequency (Supplementary Figures 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 10 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 (ref. [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 QC 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 [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 individu-

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

als using PLINK2 (`--rel-cutoff 0.05`). We removed SNPs deviating from Hardy-Weinberg equilibrium ($p<10^{-6}$) 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 multi-allelic 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" [1] (although the set of well-imputed variants in our PGC2 data set included a small fraction of biallelic indels). We define $h_{g,MAF}^2$ 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 $h_{g,1Mb}^2$ and $h_{g,chr}^2$ analogously [6]. We define h^2 as the total narrow-sense heritability—i.e., 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 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 10 principal component (PC) covariates by residualizing genotypes and phenotypes accordingly. We included PC covariates (computed using FastPCA [62]; see URLs) to eliminate phenotypic variance explained by ancestry. We transformed raw REML parameter estimates (denoted $h_{g\text{–}cc}^2$) 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 over-ascertained for schizophrenia cases, we estimated h_g^2 using PCGC regression [17] (see below) in order to avoid ascertainment-induced REML bias [17,27]. In all our PGC2 analyses, we included sex, 29 study indicators, and 10 principal components as covariates and assumed schizophrenia population risk of 1% (ref. [5, 11, 12]).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Computationally efficient implementation of PCGC regression In order to run PCGC regression on N=50K 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 genetic relationship matrix inputs); (ii) speeds up jackknife computations (by streaming the GRMs in one pass); (iii) eliminates storage of "cleaned" GRMs (i.e., GRMs with PCs projected out) by projecting PCs 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 ${h_{\mathrm{g}}}^2/{h_{\mathrm{g-cc}}}^2$ (Supplementary Table 3), noting that relative variance contributions are accurately estimated by REML even under case-control ascertainment (Supplementary Figures 16–19). Estimating per-megabase ${h_{\mathrm{g,1Mb}}}^2$ in an analogous manner would have required fitting a >2500-variance component model, which was computationally intractable, so 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-VC analyses rather than fixed effect analyses of one region at a time to improve robustness against potential confounding (e.g., 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. For schizophrenia, we used one variance component per 1Mb region in the segment (discarding regions containing <5 markers) plus a single additional variance component containing all remaining markers. (This approach is similar to ref. [64] but computationally cheaper than directly applying ref. [64] using BOLT-REML.) Including all markers in the model was necessary because of ascertainment-induced genome-wide "linkage disequilibrium" 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 1Mb 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

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

unable to perform these analyses using PCGC regression due to the disk space requirements of storing 100 different 50K $\times$ 50K GRMs.

We estimated per-GC quintile $h_{\mathrm{g,GC}}{}^2$ by stratifying 1Mb 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 1Mb 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 $h_{\mathrm{g,GC}}{}^2$ of the nth stratum as the difference between the SNP-heritability estimates for n and $n - 1$ strata.

Bounding SNP-heritability explained by top 1Mb regions. We bounded the population variance in disease liability explained by the 1Mb regions with largest true $h_{\mathrm{g,1Mb}}{}^2$ using the following procedure. We inferred an upper bound by analyzing the observed distribution of $h_{\mathrm{g,1Mb}}{}^2$ estimates and accounting for sampling variance. Explicitly, we analyzed the probability of obtaining a zero $h_{\mathrm{g,1Mb}}{}^2$ estimate, $P(0)$, as a function of the actual value of $h_{\mathrm{g,1Mb}}{}^2$ (relative to its mean). Because of sampling noise and the nonnegativity constraint on our REML $h_{\mathrm{g,1Mb}}{}^2$ estimates, $P(0)$ is always positive. In lieu of an analytic formula for $P(0)$ as a function of actual $h_{\mathrm{g,1Mb}}{}^2$, 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 $h_{\mathrm{g\text{–}cc}}{}^2$ 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 $h_{\mathrm{g,1Mb}}{}^2$ for each 1Mb region. Combining this data with the actual $h_{\mathrm{g,1Mb}}{}^2$ per region (i.e., the sum of squared simulated effect sizes), and aggregating the data from all simulations and all 1Mb regions, we obtained a clean empirical estimate of $P(0)$ as a function of actual $h_{\mathrm{g,1Mb}}{}^2$, which we observed was well-fit by a sum of two exponentials (Supplementary Fig. 3). While the empirical curve was based on simulation data, it is robust to the genetic architecture used in simulations (e.g., varying numbers of causal SNPs and normal vs. 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 $h_{\mathrm{g,1Mb}}{}^2$.

To interpret the observed fraction of zero $h_{\mathrm{g,1Mb}}{}^2$ estimates in light of this information, we har-

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

We obtained lower bounds on the fraction of ${h_{\mathrm{g}}}^2$ explained by top 1Mb regions by 3-fold cross-validation. For each fold in turn, we estimated ${h_{\mathrm{g,1Mb}}}^2$ 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 s.d./3 as a conservative 95% confidence lower bound on SNP-heritability explained by top regions. We estimate s.e. using s.d./3 because the variance of heritability estimates scales with the number of sample pairs (N^2) for $N \ll M$ [15, 16]. This s.e. estimate is not theoretically precise due to 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 (i.e., 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 ${h_{\mathrm{g,MAF}}}^2$ estimates in a manner analogous to ${h_{\mathrm{g,chr}}}^2$ estimates. To infer per-MAF bin ${h_{\mathrm{MAF}}}^2$ 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

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

simulated fully heritable quantitative traits in which normalized SNPs with MAF $p \geq 0.1\%$ (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$\geq$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 α = 0, –0.25, –0.5, –1. For each marker set, we then computed REML estimates of $h_{\mathrm{g,MAF}}{}^2$ 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 vs. exclusion of these markers had a negligible effect on schizophrenia $h_{\mathrm{g,MAF}}{}^2$ estimates (Supplementary Fig. 20). We performed REML analyses of UK10K simulated traits using a slightly modified version of GCTA v1.21 [2] in order to perform robust unconstrained REML (i.e., allow negative $h_{\mathrm{g,MAF}}{}^2$ estimates); at low sample sizes, constrained REML estimates are upward biased due to 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 $h_{\mathrm{MAF}}{}^2$ 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 their 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}= corr($l_1 > z_1$, $l_2 > z_2$), where z_1 and z_2 are appropriate liability thresholds. In general, $|r_{\mathrm{p}}| \leq |r_l|$ under a bivariate normal liability

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

threshold model; e.g., 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 an s.e. for r_l by transforming the 95% confidence interval for r_p (based on its s.e. 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 1–5 scale in the GERA data) as a covariate, we included an additional missing indicator covariate marking samples with missing BMI (5%).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **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. *American Journal of Human Genetics* **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. *American Journal of Human Genetics* **88**, 294–305 (2011).

4. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43**, 519–525 (2011).

5. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* **44**, 247–250 (2012).

6. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics* **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. *Nature Genetics* (2013).

9. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).

10. Ripke, S. *et al.* Genome-wide association study identifies five new schizophrenia loci. *Nature Genetics* **43**, 969 (2011).

11. Ripke, S. *et al.* Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nature Genetics* **45**, 1150–1159 (2013).

12. Ripke, S. *et al.* 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. *Nature Genetics* (2013).

14. Mahajan, A. *et al.* Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics* **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 Genetics* **10**, e1004269 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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* (2015).

17. Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences* **111**, E5272–E5281 (2014).

18. Legarra, A. & Misztal, I. Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360–366 (2008).

19. VanRaden, P. Efficient methods to compute genomic predictions. *Journal of Dairy Science* **91**, 4414–4423 (2008).

20. Loh, P.-R. *et al.* Efficient Bayesian mixed model analysis increases association power in large cohorts. *Nature Genetics* (2015).

21. Henderson, C. *Application of Linear Models in Animal Breeding* (University of Guelph, 1984).

22. Henderson, C. & Quaas, R. Multiple trait evaluation using relatives' records. *Journal of Animal Science* (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. *Journal of Animal Breeding and Genetics* **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* 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. *Nature Genetics* **46**, 100–106 (2014).

28. Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics* **91**, 1011–1021 (2012).

29. Lee, S. H. *et al.* Estimation of SNP heritability from dense genotype data. *American Journal of Human Genetics* **93**, 1151–1155 (2013).

30. Gusev, A. *et al.* Quantifying missing heritability at known GWAS loci. *PLOS Genetics* **9**, e1003993 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

31. Stahl, E. A. *et al.* Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nature Genetics* **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. *Nature Genetics* **46**, 1173–1186 (2014).

33. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics* **91**, 1033–1040 (2012).

34. The International HapMap Consortium, K. A., Frazer *et al.* 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 Genetics* **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 Genetics* **10**, e1004663 (2014).

37. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proceedings of the National Academy of Sciences* **111**, E455–E464 (2014).

38. Goldstein, D. B. Common genetic variation and human traits. *New England Journal of Medicine* **360**, 1696 (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. *Nature Communications* **6** (2015).

41. Gottesman, I. I. & Shields, J. A polygenic theory of schizophrenia. *Proceedings of the National Academy of Sciences* **58**, 199–205 (1967).

42. Sullivan, P. F. Puzzling over schizophrenia: schizophrenia as a pathway disease. *Nature Medicine* **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 Research* **23**, 789–799 (2013).

44. Gibson, G. Rare and common variants: twenty arguments. *Nature Reviews Genetics* **13**, 135–145 (2012).

45. Lohmueller, K. E. The impact of population demography and selection on the genetic architecture of complex traits. *PLOS Genetics* **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).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

47. Bønnelykke, K. *et al.* Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. *Nature Genetics* **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. *Nature Genetics* **45**, 907–911 (2013).

49. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *bioRxiv* 014498 (2015).

50. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common snps for metabolic syndrome traits. *PLOS Genetics* **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. *American Journal of Human Genetics* (2015).

52. Cheverud, J. M. A comparison of genetic and phenotypic correlations. *Evolution* 958–968 (1988).

53. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genetics* **9**, e1003264 (2013).

54. Haseman, J. & Elston, R. The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics* **2**, 3–19 (1972).

55. Finucane, H. K. *et al.* Partitioning heritability by functional category using GWAS summary statistics. *bioRxiv* 014241 (2015).

56. Chen, C.-Y. *et al.* Improved ancestry inference using weights from external reference panels. *Bioinformatics* btt144 (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 multiethnic study of atherosclerosis. *PLOS Genetics* **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* (2015).

61. Falconer, D. S. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics* **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* 018143 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

63. Tange, O. GNU Parallel - The Command-Line Power Tool. *The USENIX Magazine* **36**, 42–47 (2011). URL `http://www.gnu.org/s/parallel`.

64. Kostem, E. & Eskin, E. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *American Journal of Human Genetics* **92**, 558–564 (2013).

65. Bengio, Y. & Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research* **5**, 1089–1105 (2004).

66. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* 1–38 (1977).

67. Searle, S. R., Casella, G. & McCulloch, C. E. *Variance components* (John Wiley & Sons, 2006).

68. Liu, J. S. & Wu, Y. N. Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**, 1264–1274 (1999).

69. Foulley, J.-L. & Van Dyk, D. A. The PX-EM algorithm for fast stable fitting of Henderson's mixed model. *Genetics Selection Evolution* **32**, 1–21 (2000).

70. Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature Methods* **11**, 407–409 (2014).

71. Meyer, K. *et al.* PX $\times$ AI: Algorithmics for better convergence in restricted maximum likelihood estimation. In *8th World Congress on Genetics Applied to Livestock Production* (2006).

72. Groeneveld, E. A reparameterization to improve numerical optimization in multivariate REML (co)variance component estimation. *Genetics Selection Evolution* **26**, 537–545 (1994).

73. Wei, G. C. & Tanner, M. A. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85**, 699–704 (1990).

74. Matilainen, K., Mäntysaari, E. A., Lidauer, M. H., Strandén, I. & Thompson, R. Employing a Monte Carlo algorithm in expectation maximization restricted maximum likelihood estimation of the linear mixed model. *Journal of Animal Breeding and Genetics* **129**, 457–468 (2012).

75. Kuk, A. Y. & Cheng, Y. W. The Monte Carlo Newton-Raphson algorithm. *Journal of Statistical Computation and Simulation* **59**, 233–250 (1997).

76. McCulloch, C. E. Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170 (1997).

77. Gould, N. I., Orban, D., Sartenaer, A. & Toint, P. L. Sensitivity of trust-region algorithms to their parameters. *4OR* **3**, 227–241 (2005).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

78. McCulloch, C., Searle, S. & Neuhaus, J. *Generalized, linear, and mixed models* (Wiley, 2008), 2nd edn.

79. Barry, R. P. & Kelley Pace, R. Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra and its Applications* **289**, 41–54 (1999).

80. Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nature Genetics* **44**, 1066–1071 (2012).

81. Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29**, 1526–1533 (2013).

82. Tucker, G., Price, A. L. & Berger, B. A. Improving the power of GWAS and avoiding confounding from population stratification with PC-Select. *Genetics* (2014).

83. Johnson, S. G. The NLopt nonlinear-optimization package. URL `http://ab-initio.mit.edu/nlopt`.

84. Svanberg, K. A class of globally convergent optimization methods based on conservative convex separable approximations. *SIAM Journal on Optimization* **12**, 555–573 (2002).

85. Kraft, D. Algorithm 733: TOMP–Fortran modules for optimal control calculations. *ACM Transactions on Mathematical Software (TOMS)* **20**, 262–281 (1994).

86. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).

87. Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* **178**, 1709–1723 (2008).

88. Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Research* gr–169375 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

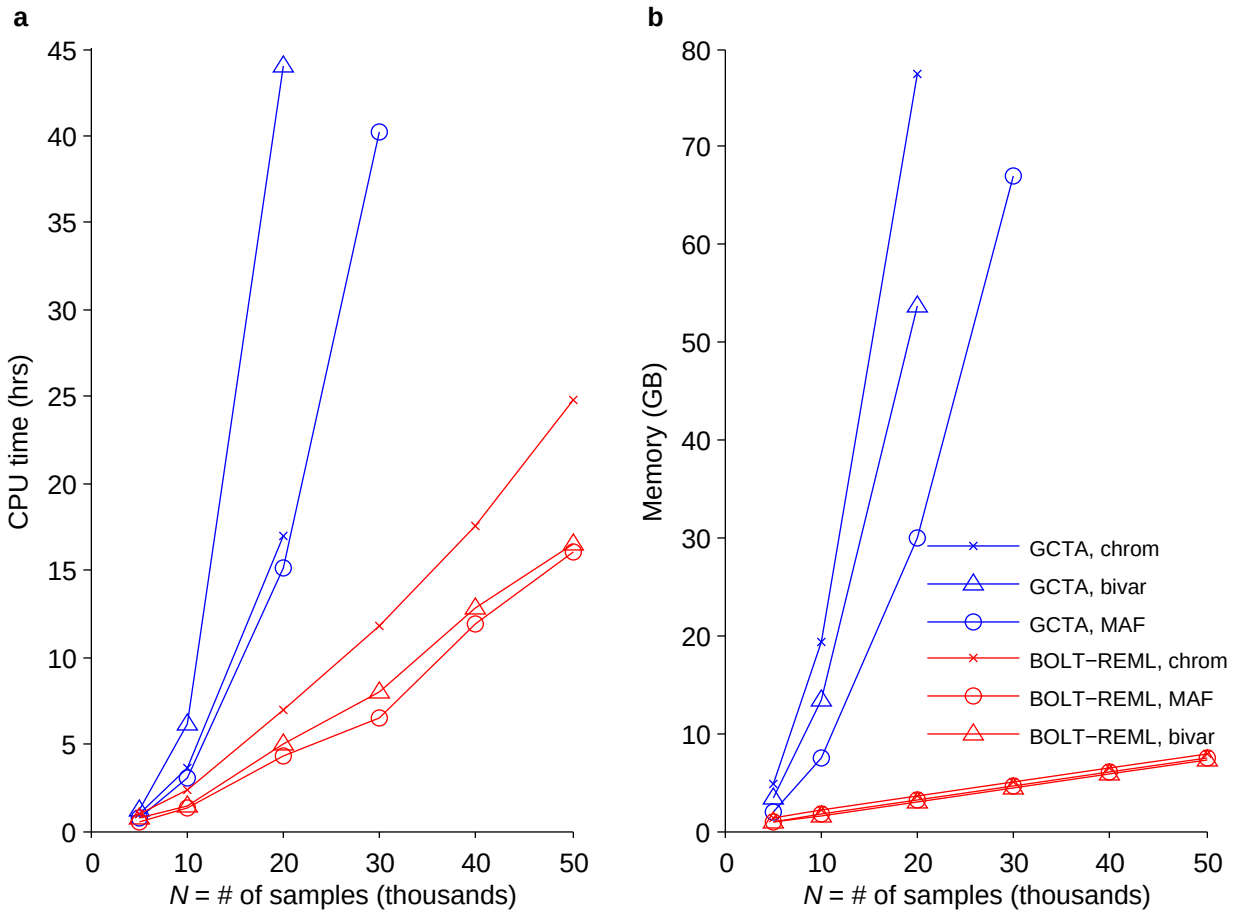

Figure 1. Computational performance of BOLT-REML and GCTA heritability analysis algorithms. Benchmarks of BOLT-REML and GCTA in three heritability analysis scenarios: partitioning across 22 chromosomes, partitioning across six MAF bins, and bivariate analysis. Run times (**a**) and memory (**b**) are plotted for runs on subsets of the GERA cohort with fixed SNP count 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 analysis. Reported run times are medians of five identical runs using one core of a 2.27 GHz Intel Xeon L5640 processor. Reported run times for GCTA are total times required for computing the GRM and performing REML analysis; time breakdowns and numeric data are provided in Supplementary Table 1. Data points not plotted for GCTA indicate scenarios in which GCTA required more memory than the 96GB available. Software versions: BOLT-REML, v2.1; GCTA, v1.24.

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

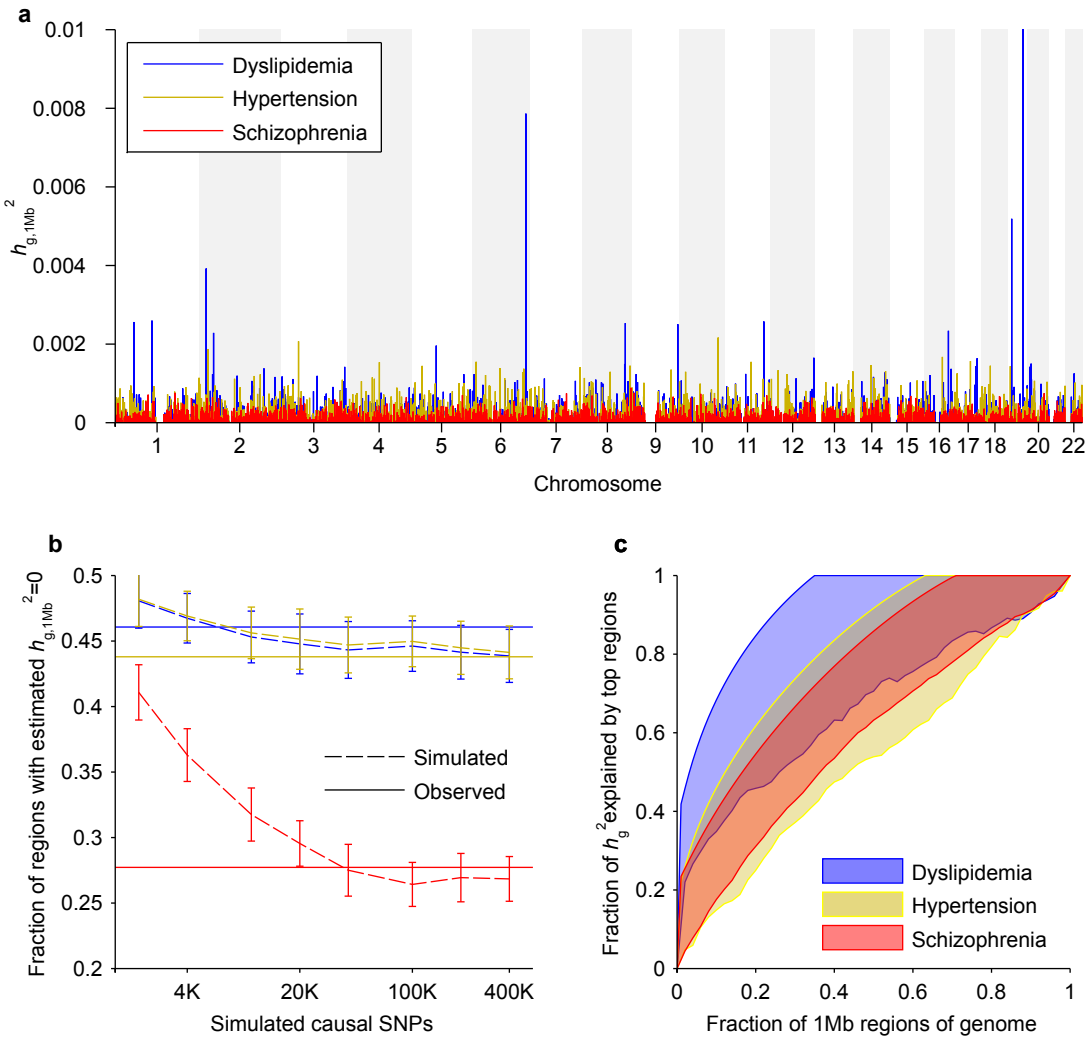

Figure 2. Extreme polygenicity of schizophrenia compared to other complex diseases. (**a**) Manhattan-style plots of estimated SNP-heritability per 1Mb region of the genome, $h_{g,1Mb}^2$, for dyslipidemia, hypertension, and schizophrenia. The *APOE* region of chromosome 19 is an outlier with an $h_{g,1Mb}^2$ estimate of 0.022. (**b**) Fractions of 1Mb regions with estimated $h_{g,1Mb}^2$ equal to its lower bound constraint of zero in disease phenotypes (solid) and simulated phenotypes with varying degrees of polygenicity and with h_g^2 matching the h_{g-cc}^2 of each disease (dashed). Simulation data plotted are means over 5 simulations; error bars, 95% prediction intervals assuming Bernoulli sampling variance and taking into account s.e.m. (**c**) Conservative 95% confidence intervals for the cumulative fraction of SNP-heritability explained by the 1Mb regions that contain the most SNP-heritability. Lower bounds are from a cross-validation procedure involving only the disease phenotypes while upper bounds are inferred from the empirical sampling variance of $h_{g,1Mb}^2$ estimates (Online Methods).

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

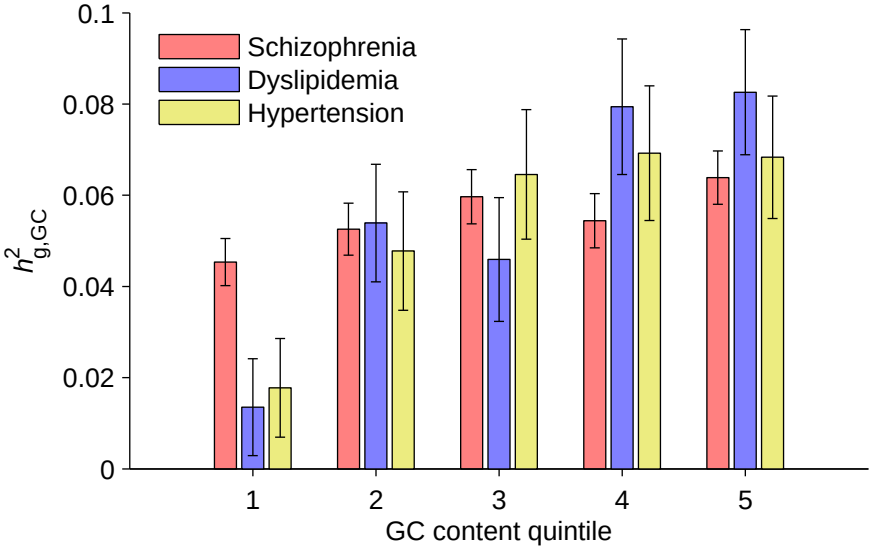

Figure 3. SNP-heritability of disease liabilities partitioned by GC content. GC content was computed at 1Mb resolution, after which 1Mb regions were stratified into GC quintiles for variance components analysis. Quintiles 1–5 have median GC contents of 35.7%, 38.1%, 40.2%, 42.8%, and 47.2%, respectively.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

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

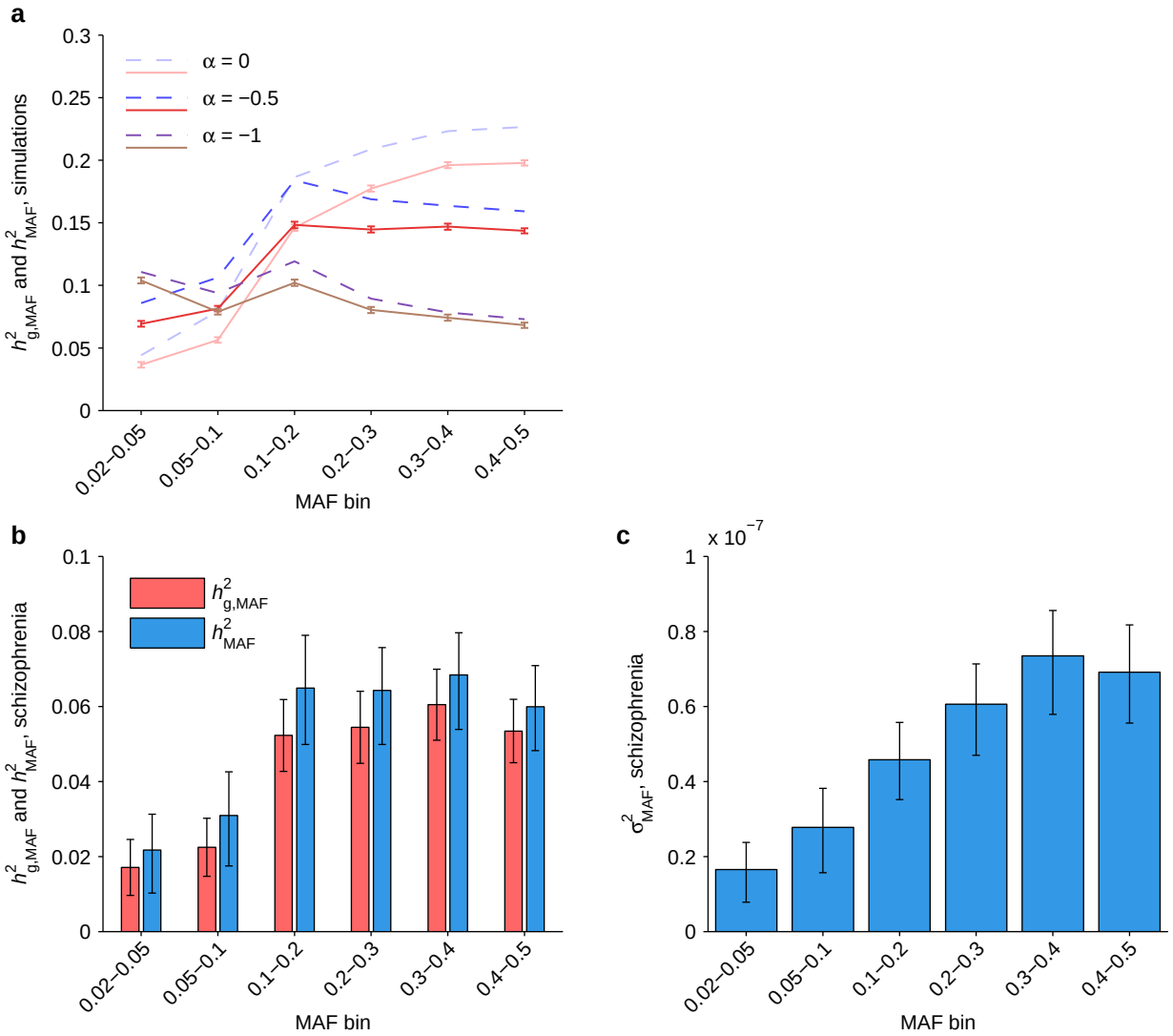

Figure 4. Inferred heritability of schizophrenia liability due to SNPs of various allele frequencies. (**a**) Simulated narrow-sense heritability per MAF bin ($h_{\mathrm{MAF}}{}^2$, dashed blue curves) and estimated SNP-heritability per MAF bin ($h_{\mathrm{g,MAF}}{}^2$, solid red curves) for quantitative phenotypes with genetic architectures in which SNPs of minor allele frequency p have average per-allele effect size variance proportional to $(p(1-p))^{\alpha}$. Simulations used causal SNPs with MAF$\geq$0.1% in UK10K sequencing data and tag 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 ratio $h_{\mathrm{g,MAF}}{}^2/h_{\mathrm{MAF}}{}^2$ at $\alpha = -0.28$, which provided the best weighted least-squares fit between observed $h_{\mathrm{g,MAF}}{}^2$ and interpolated $h_{\mathrm{g,MAF}}{}^2$ from the simulations in panel (a) (Supplementary Fig. 12). (**c**) Inferred narrow-sense heritability of schizophrenia liability explained per SNP in each MAF bin, i.e., $h_{\mathrm{MAF}}{}^2$ in panel (b) normalized by UK10K SNP counts (Supplementary Table 14). Schizophrenia $h_{\mathrm{g,MAF}}{}^2$ error bars, 95% confidence intervals based on REML analytic standard errors. Schizophrenia $h_{\mathrm{MAF}}{}^2$ and $\sigma_{\mathrm{MAF}}{}^2$ error bars, unions of 95% confidence intervals assuming $-1 \leq \alpha \leq 0$.

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

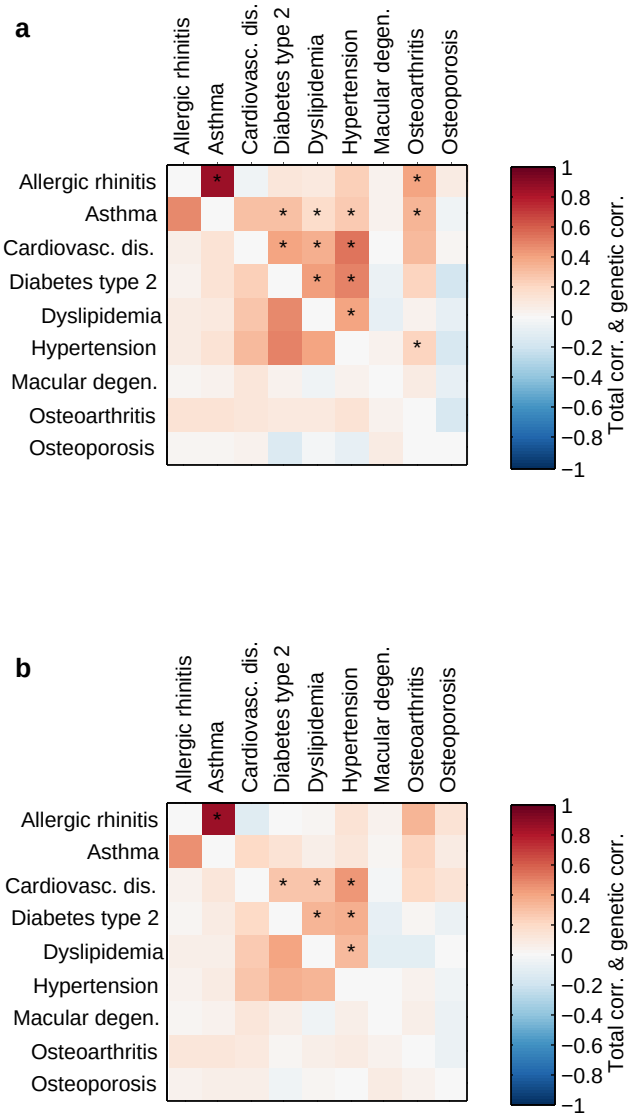

Figure 5. Genetic correlations and total correlations of GERA disease liabilities. (**a**) Correlations from bivariate analyses using only age, sex, 10 principal components, and Affymetrix kit type as covariates. (**b**) Correlations from bivariate analyses including BMI as an additional covariate. Genetic correlations are above the diagonals; total liability correlations are 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.

bioRxiv preprint doi: https://doi.org/10.1101/016527; this version posted June 5, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Disease	Cases	Controls	h_g^2 (s.e.)
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)
Cardiovasc. dis.	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 degen.	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 over-ascertainment of cases [17, 27]. Cases and controls for the other 9 diseases are from the GERA data set; h_g^2 estimates assume random sample ascertainment and were computed using BOLT-REML.