nature genetics

Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis

Eli A Stahl[1–3]*, Daniel Wegmann[4], Gosia Trynka[5], Javier Gutierrez-Achury[5], Ron Do[2,6], Benjamin F Voight[7], Peter Kraft[8], Robert Chen[1–3], Henrik J Kallberg[9], Fina A S Kurreeman[1–3], Diabetes Genetics Replication and Meta-analysis Consortium[10], Myocardial Infarction Genetics Consortium[10], Sekar Kathiresan[2,6], Cisca Wijmenga[5], Peter K Gregersen[11], Lars Alfredsson[9], Katherine A Siminovitch[12], Jane Worthington[13], Paul I W de Bakker[2,3,14,15], Soumya Raychaudhuri[1–3,16] & Robert M Plenge[1–3,16]

© 2012 Nature America, Inc. All rights reserved.

The genetic architectures of common, complex diseases are largely uncharacterized. We modeled the genetic architecture underlying genome-wide association study (GWAS) data for rheumatoid arthritis and developed a new method using polygenic risk-score analyses to infer the total liability-scale variance explained by associated GWAS SNPs. Using this method, we estimated that, together, thousands of SNPs from rheumatoid arthritis GWAS explain an additional 20% of disease risk (excluding known associated loci). We further tested this method on datasets for three additional diseases and obtained comparable estimates for celiac disease (43% excluding the major histocompatibility complex), myocardial infarction and coronary artery disease (48%) and type 2 diabetes (49%). Our results are consistent with simulated genetic models in which hundreds of associated loci harbor common causal variants and a smaller number of loci harbor multiple rare causal variants. These analyses suggest that GWAS will continue to be highly productive for the discovery of additional susceptibility loci for common diseases.

GWAS have led to the discovery of many common variants that are associated with complex traits. Given the number of SNPs tested in GWAS, an association must achieve a stringent threshold of statistical significance ($P < 5 \times 10^{-8}$) to be considered validated[1], and contemporary GWAS are underpowered to achieve this genome-wide significance for SNPs with modest effects on disease risk[2]. Assuming that disease-associated SNPs follow the distribution of effect sizes suggested by the validated associations, it is probable that many more true positive associations reside within GWAS data[3] that have only suggestive statistical evidence of association. Indeed, as sample sizes have increased, many more common variants of modest effect have been discovered for a variety of complex traits[4–7]. However, validated SNP associations explain only a portion of the liability-scale genetic variance or heritability of disease estimated from classical family studies, leading to the concept of missing heritability[8,9]. Elucidating the remaining sources of heritability will allow investigators to prioritize resources for future genetic studies, including acquisition of additional samples, technology development for variant discovery and testing (for example, next-generation genotyping arrays or sequencing) and analytical development for detecting associations of causal variants across the allele frequency spectrum.

Recently, two statistical methods were developed to assess the contributions of common SNPs that do not reach genome-wide significance: polygenic analysis[10] and mixed linear modeling[11]. Both methods test many SNPs in aggregate for a collective effect on phenotype. In the first method, an additive polygenic risk score based on SNPs that are below a *P* value threshold in a discovery GWAS is tested in an independent set of samples. Using this approach, polygenic effects have been shown in schizophrenia[10], multiple sclerosis[12], heart rate[13], height[4] and body mass index[5]. The second method estimates additive genetic variance (heritability) caused by common SNPs using linear mixed-effect modeling including a random effect that represents the polygenic component of trait variation[11,14]. Applied to height[11], endometriosis[15], Parkinson's disease[16] and other complex traits[14,17], this method has provided estimates of the heritability caused by common SNPs that are scattered throughout the genome. An additional third method[3] uses power correction based on validated SNP associations to estimate the number of additional SNPs with similar effect sizes, but this method estimates the contribution of more modest associations only by making strong assumptions about the distribution of effect sizes.

Although these methods show that additional variance can be explained by common SNPs in GWAS data, they have not offered meaningful estimates of the number and effect sizes of associated SNPs in the context of a GWAS of a common complex disease. Here, we develop a method integrating polygenic analysis[10] and the simulation of GWAS data under a polygenic disease model, using approximate Bayesian computation, to infer liability-scale additive genetic variance and the numbers, allele frequencies and effect sizes of common SNPs weakly associated with complex disease.

To understand the contribution of common SNPs to the heritability of rheumatoid arthritis, we applied our method to published GWAS data on >28,000 samples from rheumatoid arthritis case-control studies[2,18]. We compared the results of this analysis with those from

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

Received 6 May 2011; accepted 1 March 2012; published online 25 March 2012; doi:10.1038/ng.2232

Table 1 Common disease GWAS data

Disease	Discovery and test data (cohorts)	Cases	Controls	Total	SNP platform	
					N after QC	*N* after LD pruning
Rheumatoid arthritis	Discovery (5)	3,964	12,052	10,565	HapMap2	
					2,100,000	84,000
	Test (WTCCC)	1,521	10,557	5,318		
Celiac disease	Discovery (3)	2,091	3,218	4,776	Illumina 550K	
					503,000	91,000
	Test (UK2)	1,849	4,936	5,380		
Early onset MI/CAD	Discovery (MIGEN)	2,967	3,075	6,040	HapMap2	
					1,800,000	90,000
	Test (WTCCC)	1,926	2,935	4,652		
T2D mellitus	Discovery (7)	6,206	8,713	17,427	HapMap2	
					2,000,000	76,000
	Test (WTCCC)	1,924	2,938	4,651		

WTCCC, Wellcome Trust Case Control Consortium; MIGEN, Myocardial Infarction Genetics Consortium; UK2, Stage 1 Collection 2 from reference 19; QC, quality control.

family based heritability studies, a linear mixed model analysis and a simulation study of common or rare causal variant models. We then extended our analyses to published GWAS data for three additional diseases: celiac disease[19], myocardial infarction and coronary artery disease (MI/CAD)[20,21] and type 2 diabetes (T2D)[22]. Our results suggest that in all four of these common diseases, many hundreds of common SNP associations remain to be identified, with total genetic contributions accounting for the majority of the heritability of disease. Our results further suggest that common causal variants of weak effect underlie the vast majority of these genetic contributions.

RESULTS

Polygenic risk scores for rheumatoid arthritis

We used rheumatoid arthritis GWAS data from six independent case-control collections including a total of 5,485 seropositive individuals with rheumatoid arthritis (cases) and 22,609 individuals without rheumatoid arthritis of European descent (**Table 1**)[2,18]. We imputed the GWAS data genome wide using the HapMap2 European CEU reference panel for a total of over 2.5 million SNPs. We used a study design in which one dataset was used as the 'test' data and the other five datasets were used for 'discovery' so that case-control batch effects, as well as population stratification, would not be consistent across the discovery and test data.

For the polygenic analysis, we first performed a discovery GWAS using logistic regression with five eigenvectors from the principal-component analysis as covariates within each dataset and combined the results across the GWAS datasets using an inverse-variance–weighted meta-analysis. We then removed all known rheumatoid arthritis risk loci (**Supplementary Table 1**) to focus on previously unidentified SNP associations and pruned SNPs by their linkage disequilibrium (LD, $r^2 < 0.1$) (Online Methods), preferentially retaining the SNPs with lower discovery GWAS *P* values (P_{GWAS}), to obtain a set of maximally associated independent SNPs with unknown status with respect to disease risk. We selected sets of SNPs reaching nine different P_{GWAS} threshold values ($P_{GWAS} < 10^{-4}$, 10^{-3}, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4 and 0.5), and for each SNP set, we summed the log-odds–weighted risk allele counts for each individual in an independent test dataset using discovery-GWAS–estimated risk alleles and effect sizes. We tested the resulting polygenic risk scores for association with case-control status using logistic regression with gender and five principal component covariates.

Polygenic risk scores based on large numbers of SNPs were significantly associated with rheumatoid arthritis case-control status across a range of P_{GWAS} threshold values (**Fig. 1**). The most significant score was from SNPs with $P_{GWAS} < 0.05$ (12,788 SNPs had $P = 3 \times 10^{-9}$). We also analyzed scores based on SNPs with P_{GWAS} in nonoverlapping intervals (for example, $0.001 < P_{GWAS} < 0.01$) and found that significant polygenic risk score associations were caused by SNPs with $P_{GWAS} \leq 0.05$ (**Supplementary Table 2**). These results were consistent when we used alternative datasets for testing, alternative quality control and LD pruning thresholds, or alternative strategies for removing previously known associations. In addition, cases with non-autoimmune diseases in the Wellcome Trust Case Control Consortium dataset served as the negative controls and did not have significant polygenic risk score associations (**Supplementary Fig. 1**). Finally, we found the polygenic risk score effects to be scattered diffusely throughout the genome; many chromosomes contributed to a significant polygenic risk score ($P_{GWAS} < 0.05$) signal (**Supplementary Table 3**), and, consistent with the results using an independent method in other complex traits[17], the polygenic risk score effect sizes estimated here were correlated with chromosome size ($R^2 = 0.27$, $P = 0.007$). Thus, polygenic risk score associations seemed to be genuinely caused by polygenic effects that are specific to rheumatoid arthritis disease risk.

Polygenic risk scores in other common diseases

We continued testing our method using datasets for three additional diseases. We performed a polygenic analysis on GWAS data for celiac disease[19], MI/CAD[20] and T2D[22] (**Table 1**). Again, we used the samples

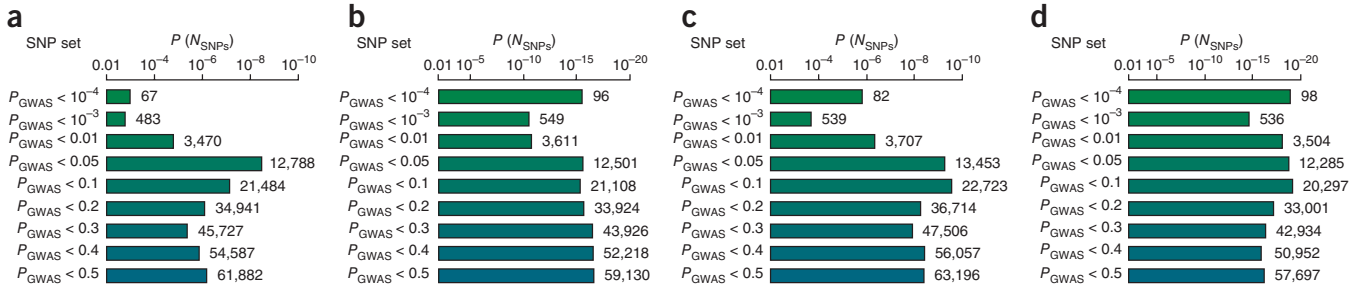

Figure 1 Association of polygenic risk scores with common disease case-control status in independent validation datasets. Association *P* values ($\log_{10}$ scale) are plotted, with the number of SNPs used for the calculation of the risk scores shown at right, for SNP sets based on P_{GWAS} thresholds ranging from 10^{-4} (top, green) to 0.5 (bottom, blue). (**a**) Rheumatoid arthritis (all known risk loci removed). (**b**) Celiac disease (with the extended MHC region removed). (**c**) Myocardial infarction (discovery data) and coronary artery disease (test data). (**d**) T2D.

© 2012 Nature America, Inc. All rights reserved.

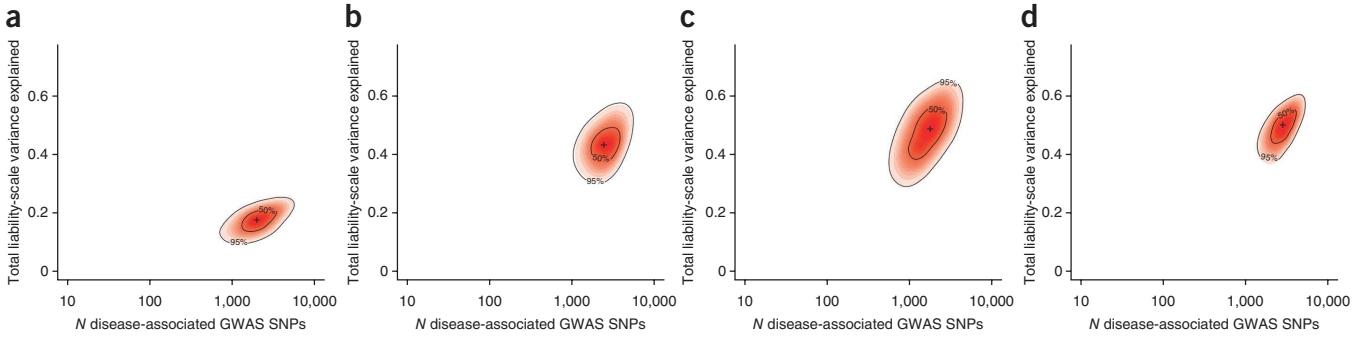

Figure 2 Posterior probability densities of the number of associated SNPs and the total liability-scale variance explained for the Bayesian analysis of the polygenic analysis results. N_{SNPs} are shown on the $\log_{10}$ scale on the *x* axis, and V_{tot} values are shown on the *y* axis. The heat map colors represent the probability density height, with darker colors indicating higher density. Contour lines show the highest posterior density and the 50%, 90% and 95% credible regions. (**a**) Rheumatoid arthritis (with all known risk loci removed). (**b**) Celiac disease (with the extended MHC region removed). (**c**) MI/CAD. (**d**) T2D.

© 2012 Nature America, Inc. All rights reserved.

from the UK as test data, as these data had restricted geographic origins relative to the discovery GWAS data and showed little stratification[19,21]. For celiac disease, we removed the major histocompatibility complex (MHC) region, which has a very strong effect on risk and on complex long-distance LD patterns; we did not remove any other known risk loci.

We used published GWAS data to show that each disease has a strong polygenic signal. As we saw in the rheumatoid arthritis datasets, the polygenic risk scores were highly significantly and specifically associated with all three of these additional common diseases (**Fig. 1** and **Supplementary Fig. 1**). Although known SNPs associated with disease risk may underlie the polygenic risk score associations for the lowest significance threshold, $P_{GWAS} < 10^{-4}$ (celiac disease, 96 SNPs, polygenic risk score $P = 2 \times 10^{-16}$; MI/CAD, 82 SNPs, $P = 1 \times 10^{-6}$; T2D, 98 SNPs, $P = 1 \times 10^{-19}$), adding thousands of independent SNPs with the marginally significant $P_{GWAS} < 0.1$ did not dilute the significance of the polygenic risk score associations (celiac disease, 21,108 SNPs, $P = 3 \times 10^{-16}$; MI/CAD, 22,723 SNPs, $P = 3 \times 10^{-10}$; T2D, 20,297 SNPs, $P = 7 \times 10^{-20}$).

Disease-associated SNPs and total variance explained

Polygenic scores are made up of an unknown number of true-positive SNPs (signal) as well as many unassociated SNPs (noise). To determine how much signal underlies our results, and, specifically, to estimate the number of associated SNPs along with their total variance explained, we conducted Bayesian inference analyses on our polygenic analysis results. Briefly, we analyzed a polygenic disease model in which independent SNPs (N_{SNPs}) additively contributed a total liability-scale variance explained (V_{tot}), with additional parameters included for the distributions of risk allele effect sizes and frequencies (**Supplementary Fig. 2**). We simulated the discovery GWAS and polygenic analysis for associated and null SNPs and used polygenic risk score logistic regression R^2 values[23] across nonoverlapping SNP sets, including scores stratified by risk-allele frequency (**Supplementary Table 2**), as summary statistics to compare the simulated and observed results. We used approximate Bayesian computation with rejection sampling and general linear model post-sampling adjustment (ABC-GLM)[24,25] to estimate the posterior densities of polygenic disease model parameters given the polygenic analysis results (**Supplementary Fig. 3**). See the Online Methods and the **Supplementary Note** for details.

Figure 2 shows the joint posterior probability densities of the two key polygenic disease model parameters, N_{SNPs} and V_{tot}. These densities are well restricted to within the range of uniform priors for these two parameters (N_{SNPs}, 10–10,000 on a $\log_{10}$ scale; V_{tot}, 0.01–0.5 for rheumatoid arthritis and 0.01–0.99 for the other diseases). For rheumatoid arthritis, excluding all known risk loci, the posterior density mode provided estimates of 18% (95% credible interval, 11–24%) of the total variance being explained by 2,231 independent disease-associated SNPs (95% credible interval 846–4,608) (**Fig. 2** and **Table 2**). Results were robust to alternative prior distributions of the N_{SNPs} and for the effect size paramater β_v, and validation analyses indicated that the parameters were inferred with reasonable bias and precision under a wide range of models (**Supplementary Fig. 4**). We also applied the previously developed linear mixed-effects modeling (LMM) method[11,14,26]. This complementary approach yielded consistent results for the variance explained by common SNPs (directly comparable to our V_{tot} results; **Table 2**). Given that rheumatoid arthritis recurrence rates for relatives of affected individuals yield estimates of a narrow-sense heritability of about 0.55 (refs. 27,28) and that previously validated risk loci contribute an estimated heritability of 0.18 (refs. 2,29) (**Supplementary Table 1**), these results show that

Table 2 Comparison of results of different polygenic methods across diseases

			Caused by common GWAS SNPs		
				Polygenic modeling and Bayesian inference	
Disease	Prevalence (%)	Family based heritability[a]	LMM-based heritability (s.e.)	Total variance explained (50% CI)	*N* SNPs (50% CI)
Rheumatoid arthritis	1	0.53–0.68 (–0.13 MHC)[b]	0.32 (0.037)	0.18 (0.15–0.20) (+0.04 known non-MHC)[b]	2,231 (1,588–2,740)
Celiac disease	1	0.5–0.87 (–0.35 MHC)[b]	0.33 (0.042)	0.44 (0.40–0.47)	2,550 (1,907–3,061)
MI/CAD	6	0.3–0.63	0.41 (0.067)	0.48 (0.43–0.54)	1,766 (1,215–2,125)
T2D mellitus	8	0.26–0.69	0.51 (0.065)	0.49 (0.46–0.53)	2,919 (2,335–3,442)

[a]Family based heritability estimates were taken from previous data for rheumatoid arthritis[27,28], celiac disease[18,30], MI/CAD[31,32] and T2D[33,34]. [b]We excluded some loci in certain analyses: although the family based heritability estimates are based on the whole genome, the extended MHC region was removed from the common GWAS SNP analyses for rheumatoid arthritis and celiac disease, and validated non-MHC loci were further removed from the polygenic modeling analysis of the rheumatoid arthritis GWAS data. 50% CI, 50% credible interval; s.e., standard error.

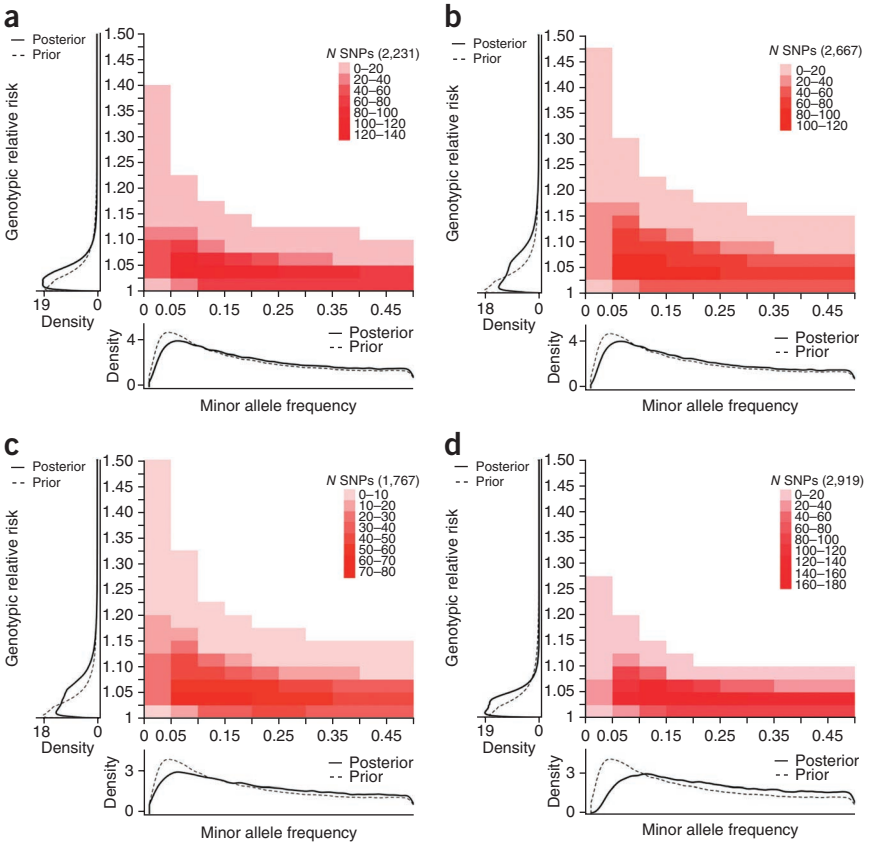

Figure 3 Posterior probability distributions of the relative risk and minor allele frequency of the inferred disease-associated SNPs. The GRR is shown on the *y* axis in the left and middle images, and the MAF is shown on the *x* axis in the middle and bottom images. Heat map colors indicate the mean posterior numbers of SNPs in risk allele frequency (RAF)-GRR bins scaled to the posterior mean number of disease-associated SNPs (indicated in the legend). The graphs on the left and at the bottom show the marginal posterior (solid line) and prior (dashed line) probability densities. (**a**) Rheumatoid arthritis (with all known risk loci removed). (**b**) Celiac disease (with the extended MHC region removed). (**c**) MI/CAD. (**d**) T2D.

roughly 65% of the heritability of rheumatoid arthritis can be accounted for by purely additive effects of common SNPs in the GWAS data that tag causal alleles.

We applied the same polygenic-model inference method to celiac disease, MI/CAD and T2D (**Fig. 2** and **Table 2**). We found substantial total liability-scale variance (V_{tot}) explained by GWAS SNPs (celiac disease, 0.43, outside of the MHC; MI/CAD, 0.48; T2D, 0.49). For a comparison, validated common SNP associations explain 5% (celiac disease, 27 non-MHC loci), 4% (25 MI/CAD loci)[30] and 10% (44 T2D loci)[31] of the total liability-scale disease variance in these three diseases. Taking into account the uncertainty in both methods, heritabilities caused by common SNPs estimated using LMM[11,14,26] were consistent with the V_{tot} values estimated using polygenic modeling and Bayesian inference, with no clear pattern of overestimation seen by using one method compared to the other. Although family based heritability estimates vary widely for these three diseases[19,32–36], the majority of their heritability is explained by common SNPs in GWAS data, without exception (**Table 2**): 83–100% of the heritability for celiac disease (heritability of 0.5–0.87, with 0.35 caused by HLA alleles in the MHC[19,37]), 80–100% of the heritability for MI/CAD and 70–100% of the heritability for T2D is explained by common SNPs.

Risk allele frequencies and effect sizes

Our Bayesian analysis generated a posterior distribution of polygenic disease model parameters, which determine the minor allele frequencies (MAFs) and genotypic relative risks (GRRs) of the inferred common SNP associations. We calculated the mean posterior distributions of the MAFs and GRRs of the associated SNPs from 1,000 samples from the joint posterior density (**Fig. 3**). We also determined the marginal prior distributions for the MAFs and GRRs that were implied by the model parameters' Bayesian priors (**Supplementary Fig. 2**). For all four common diseases, the posterior distribution of the MAFs of the associated SNPs was shifted from that of the prior distribution (all GWAS SNPs after LD pruning) toward the SNPs of more intermediate frequency. The posterior distribution of the GRRs indicates that the effect sizes of most of the disease-associated SNPs ranged from almost 1 to approximately 1.05, with larger GRRs being seen for less common MAFs (1–5%).

Notably, the number of SNPs with moderate effect sizes (measured by liability-scale variance explained; GRR > 1.05 for SNPs with MAF = 0.5 and GRR > 1.1 for SNPs with MAF = 0.05), and the total variance explained by these SNPs, varied markedly across the four diseases. Substantial numbers of SNPs with moderate effect sizes contributed the majority of the inferred total liability-scale variance explained for celiac disease (981 (95% credible interval 663–1,417) SNPs of the 2,666 total N_{SNPs} explained 0.33 (95% credible interval 0.2–0.45) of the V_{tot} of 0.43) and MI/CAD (597 (95% credible interval 319–874) SNPs of the total 1,766 N_{SNPs} explained 0.34 (95% credible interval 0.13–0.5) of the V_{tot} of 0.48). Fewer SNPs of moderate effect size explained much smaller proportions of the total disease variance in rheumatoid arthritis (212 (95% credible interval 0–492) SNPs of the total 2,231 N_{SNPs} explained 0.05 (95% credible interval 0–0.14) of the V_{tot} of 0.18) and T2D (298 (95% credible interval 0–588) SNPs of the total 2,919 N_{SNPs} explained 0.14 (95% credible interval 0–0.31) of the V_{tot} of 0.49).

Modeling causal variants

To assess what causal genetic models could explain our results, we performed simulations with causal variants and the resulting tag-SNP associations. Recent theoretical studies have posited that multiple rare causal variants may result in common SNP associations[38]. Such 'synthetic associations' probably do not account for most of the validated GWAS signals[39], but the contribution of these associations to weaker undiscovered common SNP associations has not been previously considered. We used 1000 Genomes Project[40] data and HAPGEN software[41] to simulate 10-Mb haplotypes in case-control populations under genetic models with varying numbers and effect sizes of either common (MAF > 5%) or rare (MAF < 1%) causal variants and determined the patterns of association at marker SNPs interrogated in the GWAS data. This approach allowed us to identify causal variant models in which GWAS marker SNPs were consistent

© 2012 Nature America, Inc. All rights reserved.

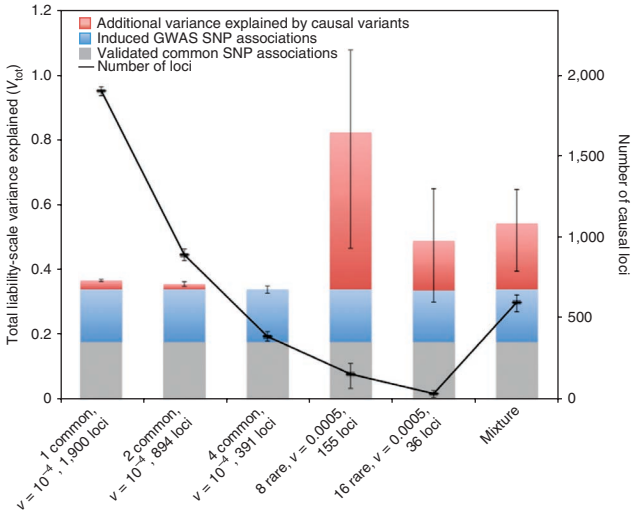

Figure 4 Causal variants underlying the rheumatoid arthritis polygenic disease architecture inferred from the GWAS data. Plotted are the liability-scale variances explained (V_{tot}, bars, left *y* axes) and the number of loci harboring causal variants (black line, right *y* axes). The colored sections in the bars partition the V_{tot} values for previously validated common SNP associations (gray), undiscovered GWAS SNP associations induced by causal variants (blue) and causal variants (V_{tot}, in addition to the values for GWAS SNPs, red). Error bars show 95% confidence intervals for causal variant numbers and V_{tot} values based on simulations achieving a GWAS SNP V_{tot} value equal to that inferred from the polygenic modeling. Six plausible causal variant models are plotted (left to right): (i) 1,900 loci each with a single common (MAF > 5%) causal variant, (ii) 894 loci each with 2 common causal variants, (iii) 391 loci each with 4 common causal variants, (iv) 155 loci each with 8 rare (MAF < 1%) causal variants, (v) 16 rare causal variants per locus with $v = 0.0005$ and (vi) a mixture (60:40 ratio of model 2 to model 4 in terms of GWAS SNPs V_{tot} values, implying 536 common causal variant loci and 62 rare causal variant loci). The per-causal–variant liability-scale variances explained (v) for models that are consistent with the polygenic modeling and inference results were $v = 0.0001$ for common causal variants and $v = 0.0005$ for rare causal variants.

with our polygenic modeling inference in terms of both their number and total variance explained (**Supplementary Table 4**), as well as their allele frequency and effect size distributions (**Supplementary Fig. 5**). Thus, we could directly address allelic heterogeneity and rare causal variant hypotheses underlying weak, polygenic effects in GWAS data.

Only models with few (1–4) common causal variants per locus and those with many (8–16) rare causal variants per locus resulted in associated GWAS SNPs that were consistent with the Bayesian inference results (**Supplementary Table 4** and **Supplementary Fig. 5**). We emphasize that to explain weak undiscovered common SNP associations, causal variants must themselves have weaker effects than have been studied previously, particularly for rare causal variants[10,38,39,42–45]. For consistent causal variant models, we simulated the number of loci genome wide that yielded our inferred total variance that was explained by the associated marker SNPs and calculated the contribution of the causal variants themselves to heritability (**Fig. 4**). Under genetic models with common causal variants, our simulations suggested that many hundreds to thousands of common causal variants spread across hundreds of loci would account for roughly the same proportion of heritability as their GWAS marker SNP tags (**Fig. 4**) but would not account for all of the disease heritability. In contrast, under models in which the causal variants are rare, only a small number of loci explain all of the common disease heritability; with larger numbers of loci, heritability owing to causal variants quickly exceeds realistic heritability estimates.

DISCUSSION

The biometrical model proposed by R.A. Fisher[46] posited that a large number of additive genetic factors inherited in a Mendelian fashion could account for the familial patterns of complex traits. In 1916, Fisher's model was criticized by Karl Pearson as being "out of the range of experiment by Mendelian methods"[47]. With the advent of GWAS that interrogate millions of common SNPs with high-throughput genotyping arrays and imputation, it is now possible to test Fisher's model of inheritance. In our study, we used polygenic analyses of GWAS data to show that a substantial proportion of SNPs reaching at best suggestive levels of statistical significance contribute to common disease risk when considered in aggregate (**Fig. 1**).

Our study extends a previously developed method[10] by performing approximate Bayesian computation (ABC-GLM) to estimate the credible region of polygenic disease model parameters (for example, number of SNPs, effect size and allele frequency) that can account for polygenic risk score associations. Bayesian inference, together with consistent results obtained using the previously developed LMM method[11,14], provide convincing evidence that substantial variance in disease liability can be explained by common SNPs captured in contemporary GWAS data (**Table 2**). For rheumatoid arthritis, the hidden heritability is on par with the variance explained by the validated risk loci, such that a total of ~36% of the overall disease liability, or ~65% of the total heritability, can be attributed to the purely additive effects of common SNPs. For celiac disease, MI/CAD and T2D, our results suggest that the true heritabilities are on the high sides of the ranges of the family based estimates and that at least ~70% of the heritability of these diseases is explained by common GWAS SNPs.

Bayesian analyses allow for computation of the posterior distribution of polygenic disease model parameters, which can then be used to address questions relating to the genetic architecture of common disease. Here, in addition to estimating the number of SNPs and their total variance explained (**Fig. 2**), we generated the posterior distribution of the allele frequencies and effect sizes of the inferred, risk-associated SNPs (**Fig. 3**) and investigated plausible causal variant models (**Fig. 4**). Other potential applications of this type of analysis include performing power calculations to predict the outcomes of future genetic studies, developing future discovery efforts such as Bayesian and pathway-based GWAS[48,49], estimating the accuracy of the risk prediction that is attainable with additional validated or unvalidated risk alleles[50,51] and developing and testing hypotheses for the polygenic adaptation[52,53] that has affected the risk of complex disease.

Although our results were qualitatively similar across the four common diseases we studied, the inferences did vary, with rheumatoid arthritis having a lower estimate (0.18) of total liability-scale variance explained (V_{tot}) by GWAS SNPs than the other three common diseases (which ranged from 0.43 to 0.49). This difference is largely a result of the exclusion of known loci for rheumatoid arthritis (~30 risk loci that together explain ~18% of the phenotypic variance). Furthermore, the inferred distributions of the effect sizes of the associated SNPs (measured on the liability scale, implying larger genotypic relative risks for lower minor allele frequencies) varied markedly across diseases: the ratios of the V_{tot} caused by SNPs with moderate compared to weak liability-scale effect sizes (corresponding to GRR > 1.05 compared to 1.01 < GRR < 1.05 for MAF = 0.5) ranged from roughly three for celiac disease and MI/CAD to roughly one-third for rheumatoid arthritis and T2D. These differences in our estimates between diseases may

© 2012 Nature America, Inc. All rights reserved.

have implications for the genetics of these diseases and will be validated and better characterized in future studies.

Our simulations incorporating causal variants and GWAS marker SNPs are consistent with results from other recent studies[10,42–45] and indicate that common causal alleles with weak effects can explain most of the polygenic signal observed in GWAS data. Unlike previous studies, we examined the impact of causal variant models on multiple weakly associated GWAS SNPs rather than considering only the single most strongly associated SNP. We found that relatively weak causal variant effect sizes (GRR ~ 1.04, 1.1, 1.5 or 3.5 for MAF = 50%, 5%, 1% or 0.1%, respectively) are required to be consistent with the polygenic analysis of GWAS data.

We show that underlying genetic models with either common (MAF > 5%) or rare (MAF < 1%) causal variants can be consistent with the data in terms of the total number of associated GWAS SNPs and the variance explained. However, under rare causal variant models for complex traits, on the order of ten causal loci are required or the variance explained by causal variants will exceed the heritability of disease[38]. This is because rare causal variants result in many weakly associated GWAS SNPs (because they are not well tagged by any single common SNP) with less total variance explained than the amount explained by the causal variants themselves and because substantial allelic heterogeneity (eight or more rare causal variants per locus) is required to induce associations throughout the common SNP frequency spectrum[38,39,45]. As our polygenic analysis suggested that the associations are diffuse throughout the genome, we conclude that the majority of the causal variants that underlie the polygenic signal of association in the GWAS data are themselves common and not rare. Common causal variants would account for a proportion of heritability only slightly greater than that of the SNPs associated within GWAS, leaving some heritability still unexplained.

We do not rule out the possibility of a contribution of rare causal variants. Indeed, a genetic model positing a mixture of loci harboring common and/or rare causal variants would fit the posterior distribution of associated GWAS SNPs better than any single model we simulated; this conclusion is based on the observation that the common causal variant models generated slightly fewer low-frequency, moderate-effect–size GWAS alleles compared to our posterior distribution, whereas rare-variant models generated slightly more (**Supplementary Fig. 5**). A genetic model that posits a mixture of common and rare causal variants could explain all of the heritability of disease but would still be dominated by common causal variants (**Fig. 3**). Finally, we note that many extremely rare causal variants that segregate privately within families would not induce SNP associations within GWAS data and, therefore, could contribute to the remaining estimated heritability under the causal variant models we studied.

Even if a complex disease is highly polygenic, it is probable that risk loci will implicate a limited number of disease-relevant biological pathways. Recent studies have shown that genes in validated rheumatoid arthritis risk loci are functionally related in terms of their descriptions in the literature[29,54], their physical interactions[55] and the tissues in which they are specifically expressed[56]. Furthermore, larger sets of suggestive loci show an over-representation of broad functional categories[57] and tissue-specific expression[56] and contribute to the disease associations of canonical molecular biological pathways[49]. By extension, many additional validated risk loci would hold great promise for bioinformatic analyses to be able to point to the mechanisms of common disease pathogenesis.

Our results have major implications for the design of future genetic association studies to identify additional common disease risk loci. Ideally, whole-genome sequencing in large case-control collections would capture all types of variants (SNPs, indels and copy number variants) across the entire range of allele frequencies (common to low frequency to private). However, such a study is prohibitively expensive at this time and comes with its own challenges, both computationally and in the interpretation of the results. The polygenic model posterior distributions for each of the four diseases examined here give expectations of hundreds of SNPs with moderate effect sizes (GRR > 1.05), especially for celiac disease and MI/CAD. Although the contributions of previously validated SNPs must be accounted for in further analyses, the difference between the V_{tot} inferred here and the variance explained by validated SNPs strongly suggests that there exist many associations that would be detectable in larger GWAS. Therefore, our results indicate that the common variant GWAS approach will continue to be a highly productive method of identifying additional risk alleles for common disease.

URLs. Full SNP results from a previous rheumatoid arthritis meta-analysis[2], http://www.broadinstitute.org/ftp/pub/rheumatoid_arthritis/Stahl_etal_2010NG/; ABCtoolbox software package, http://cmpg.iee.unibe.ch/content/softwares__services/computer_programs/abctoolbox/index_eng.html; GCTA software package, http://gump.qimr.edu.au/gcta/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

R.M.P. is supported by grants from the US National Institutes of Health (NIH) (R01-AR057108, R01-AR056768, U01-GM092691 and R01-AR059648) and holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund. S.R. is supported by an NIH Career Development Award (K08AR055688-01A1). The Brigham Rheumatoid Arthritis Sequential Study Registry is supported by a grant from Crescendo and Biogen-Idec. The North American Rheumatoid Arthritis Consortium is supported by the NIH (NO1-AR-2-2263 and RO1-AR44422). This research was also supported in part by the Intramural Research Program of the National Institute of Arthritis, Musculoskeletal and Skin Diseases of the NIH and by a Canada Research Chair and grants to K.A.S. from the Canadian Institutes for Health Research (MOP79321 and IIN-84042) and the Ontario Research Fund (RE01061). We acknowledge S. Purcell, A. Price and N. Zaitlen for help with the design and implementation of the study and analysis.

AUTHOR CONTRIBUTIONS

Study design: R.M.P., E.A.S., S.R. and P.I.W.d.B. Analysis: E.A.S. (lead), D.W., G.T., J.G.-A., R.D., B.F.V. (primary contributors), R.C., H.J.K. and F.A.S.K. Samples and data: C.W., S.K., B.F.V., the Myocardial Infarction Genetics Consortium, the Diabetes Genetics Replication and Meta-analysis Consortium, J.W., L.A., P.K.G., K.A.S. and R.M.P. Writing: R.M.P., E.A.S. (leads), D.W., P.K. (primary contributors) and all other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Wellcome Trust Case Control Consortium. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
2. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
3. Park, J.H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
4. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

© 2012 Nature America, Inc. All rights reserved.

5. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–948 (2010).
6. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
7. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
8. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
9. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
10. Purcell, S.M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
11. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
12. Bush, W.S. *et al.* Evidence for polygenic susceptibility to multiple sclerosis—the shape of things to come. *Am. J. Hum. Genet.* **86**, 621–625 (2010).
13. Eijgelsheim, M. *et al.* Genome-wide association analysis identifies multiple loci related to resting heart rate. *Hum. Mol. Genet.* **19**, 3885–3894 (2010).
14. 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).
15. Painter, J.N. *et al.* Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. *Nat. Genet.* **43**, 51–54 (2011).
16. Do, C.B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet.* **7**, e1002141 (2011).
17. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
18. Chen, R. Fine mapping the TAGAP locus in rheumatoid arthritis. *Genes Immun.* **12**, 314–318 (2011).
19. Dubois, P.C. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
20. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41**, 334–341 (2009); erratum **41**, 762 (2009).
21. Wellcome Case Control Consortium. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
22. Voight, B.F. *et al.* Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat. Genet.* **42**, 579–589 (2010).
23. Nagelkerke, N.J.D. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692 (1991).
24. Leuenberger, C. & Wegmann, D. Bayesian computation and model selection without likelihoods. *Genetics* **184**, 243–252 (2010).
25. Wegmann, D., Leuenberger, C., Neuenschwander, S. & Excoffier, L. ABCtoolbox: a versatile toolkit for approximate Bayesian computations. *BMC Bioinformatics* **11**, 116 (2010).
26. 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).
27. MacGregor, A.J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* **43**, 30–37 (2000).
28. van der Woude, D. *et al.* Quantitative heritability of anti-citrullinated protein antibody–positive and anti-citrullinated protein antibody–negative rheumatoid arthritis. *Arthritis Rheum.* **60**, 916–923 (2009).
29. Raychaudhuri, S. Recent advances in the genetics of rheumatoid arthritis. *Curr. Opin. Rheumatol.* **22**, 109–118 (2010).
30. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
31. Wheeler, E. & Barroso, I. Genome-wide association studies and type 2 diabetes. *Brief. Funct. Genomics* **10**, 52–60 (2011).
32. Nisticò, L. *et al.* Concordance, disease progression, and heritability of coeliac disease in Italian twins. *Gut* **55**, 803–808 (2006).
33. Marenberg, M.E., Risch, N., Berkman, L.F., Floderus, B. & de Faire, U. Genetic susceptibility to death from coronary heart disease in a study of twins. *N. Engl. J. Med.* **330**, 1041–1046 (1994).
34. Nora, J.J., Lortscher, R.H., Spangler, R.D., Nora, A.H. & Kimberling, W.J. Genetic-epidemiologic study of early-onset ischemic heart disease. *Circulation* **61**, 503–508 (1980).
35. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811–2819 (2011).
36. Poulsen, P., Kyvik, K.O., Vaag, A. & Beck-Nielsen, H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia* **42**, 139–145 (1999).
37. van Heel, D.A. *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nat. Genet.* **39**, 827–829 (2007).
38. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D.B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
39. Wray, N.R., Purcell, S.M. & Visscher, P.M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.* **9**, e1000579 (2011).
40. 1000 Genomes Project Consortium. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010); erratum **473**, 544 (2011).
41. Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet.* **5**, e1000477 (2009).
42. Wang, K. *et al.* Interpretation of association signals and identification of causal variants from genome-wide association studies. *Am. J. Hum. Genet.* **86**, 730–742 (2010).
43. Orozco, G., Barrett, J.C. & Zeggini, E. Synthetic associations in the context of genome-wide association scan signals. *Hum. Mol. Genet.* **19**, R137–R144 (2010).
44. Park, L. Identifying disease polymorphisms from case-control genetic association data. *Genetica* **138**, 1147–1159 (2010).
45. Spencer, C., Hechter, E., Vukcevic, D. & Donnelly, P. Quantifying the underestimation of relative risks from genome-wide association studies. *PLoS Genet.* **7**, e1001337 (2011).
46. Fisher, R. The correlation between relatives on the supposition of Mendelian inheritance. *Phil. Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
47. Norton, B. & Pearson, E.S. A note on the background to, and refereeing of, R. A. Fisher's 1918 paper 'On the correlation between relatives on the supposition of Mendelian inheritance'. *Notes Rec. R. Soc. Lond.* **31**, 151–162 (1976).
48. Stephens, M. & Balding, D.J. Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681–690 (2009).
49. Eleftherohorinou, H. *et al.* Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE* **4**, e8068 (2009).
50. Cornelis, M.C. *et al.* Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry. *Ann. Intern. Med.* **150**, 541–550 (2009).
51. Wei, Z. *et al.* From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet.* **5**, e1000678 (2009).
52. Pritchard, J.K., Pickrell, J.K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208–R215 (2010).
53. Pritchard, J.K. & Di Rienzo, A. Adaptation—not by sweeps alone. *Nat. Rev. Genet.* **11**, 665–667 (2010).
54. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
55. Rossin, E.J. Proteins encoded in genomic regions associated to immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
56. Hu, X. *et al.* Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *Am. J. Hum. Genet.* **89**, 496–506 (2011).
57. Freudenberg, J. *et al.* Locus category based analysis of a large genome-wide association study of rheumatoid arthritis. *Hum. Mol. Genet.* **19**, 3863–3872 (2010).

[1]Division of Rheumatology Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts, USA. [2]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [3]Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts, USA. [4]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California, USA. [5]Department of Genetics, University Medical Center Groningen and University of Groningen, Groningen, The Netherlands. [6]Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA. [7]Department of Pharmacology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [8]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [9]Institute of Environmental Medicine, Karolinska Institutet Hospital Solna, Stockholm, Sweden. [10]A full list of members is provided in the **Supplementary Note**. [11]The Feinstein Institute for Medical Research, North Shore–Long Island Jewish Health System, Manhasset, New York, USA. [12]Department of Medicine, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, Ontario, Canada. [13]Arthritis Research UK Epidemiology Unit, Manchester Academic Health Science Centre, University of Manchester, Manchester, UK. [14]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. [15]Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands. [16]These authors contributed equally to this work. Correspondence should be addressed to E.A.S. (estahl@rics.bwh.harvard.edu) or R.M.P. (rplenge@partners.org).

© 2012 Nature America, Inc. All rights reserved.

ONLINE METHODS

GWAS data, quality control and filtering. Quality control filtering, principal components analysis and genome-wide imputation were conducted as previously reported for rheumatoid arthritis[2], celiac disease[19], MI/CAD[20] and T2D[22]. For this study, the rheumatoid arthritis Epidemiological Investigation of Rheumatoid Arthritis (EIRA) I and II GWAS datasets[2,18] were combined, quality-control filtered and imputed. For MI/CAD, HapMap2 SNPs were extracted from data imputed into the 1000 Genomes European (CEU) reference panel (August 2009 release) (R.D. & S.K., data not shown). See the **Supplementary Note** for further details.

Known rheumatoid arthritis risk loci were removed by excluding the extended MHC region (chromosome 6, 25–35 Mb), 2 Mb across the *PTPN22* region and 1-Mb regions centered on other previously validated SNPs, extended to the furthest SNPs with LD with the known SNPs ($r^2 > 0.1$ in HapMap2 release 24) (**Supplementary Table 1**). The extended MHC region (chromosome 6, 20–40 Mb) was removed from the celiac GWAS data.

Polygenic risk score analysis. Polygenic analyses were conducted as described[10]. The discovery-set GWAS was conducted as previously described (inverse-variance–weighted meta-analysis for rheumatoid arthritis[2,18] and T2D[22], Cochran-Mantel-Haenszel tests for celiac celiac disease[19] and a combined analysis for MI/CAD[20]), excluding the samples used as test data (**Table 1**).

LD pruning by association was conducted to achieve a set of independent SNPs that retained as much association signal as possible from the discovery GWAS. Starting with the most strongly associated SNP, all SNPs in LD ($r^2 > 0.1$) were excluded until no SNPs with $r^2 > 0.1$ remained in the data. LD was calculated from HapMap2 release 24 data for all pairs of SNPs less than 1 Mb away from each other or across long-distance LD regions (*PTPN22*, chromosome 1: 113–115 Mb; MHC, chromosome 6: 25–35 Mb; chromosome 8: 6–10 Mb; chromosome 17: 40–50 Mb). A Perl program to conduct LD pruning is available on request.

Polygenic risk scores were calculated for sets of SNPs that were selected based on nine discovery-set GWAS statistical significance thresholds: $P_{GWAS} < 10^{-4}$, 10^{-3}, 0.01, 0.05, 0.10, 0.20, 0.30, 0.40 and 0.50 or $P_{GWAS} = 0.01$, 0.10 and 0.50 for the analyses stratified by RAF quintile. For each SNP set, the additive weighted polygenic risk scores PRS_i were calculated for each validation set individual i as $PRS_i = \sum_{j \in SNPs} \hat{\hat{\beta}}_j d_{ij}$, where $\hat{\hat{\beta}}_j > 0$ is the discovery GWAS log-odds ratio for the risk allele and d_{ij} is individual i's dosage (0–2) of that allele. Polygenic risk scores were tested for association with disease status by logistic regression with gender and five principal component covariates.

Polygenic disease modeling. A polygenic disease model was parameterized in which a number of independent biallelic polymorphisms (N_{SNPs}) contributed additively to a total liability-scale variance explained (V_{tot}). The relative per-SNP liability-scale variance explained was modeled by a β-distribution function, $\beta(1, \beta_v)$, allowing for a wide range of effect size heterogeneity (**Supplementary Fig. 2a**). The RAF distribution was modeled by the product of the empirical distribution (after LD pruning) and the β distribution $\beta(\alpha_{RAF}, \beta_{RAF})$, allowing for mostly rare or mostly common risk alleles (**Supplementary Fig. 2b**). Given a SNP's variance explained, v, and RAF, p, its genotypic relative risk was calculated according to the liability threshold model[10,58]: $GRR = 1 + \sqrt{vI^2 / 2p(1-p)}$, where I is the quotient of the standard normal density at the disease liability threshold and the disease prevalence K. Thus, polygenic disease was modeled by five parameters: N_{SNPs}, V_{tot}, β_v, α_{RAF} and β_{RAF}.

Polygenic analyses—from discovery GWAS to polygenic risk score association tests—were simulated for comparison with observed results for associated and null SNPs roughly equal in number to the total numbers of SNPs obtained after LD pruning of real data (80,000 SNPs for rheumatoid arthritis and MI/CAD and 70,000 SNPs for celiac disease and T2D; **Table 1**). Discovery GWAS results were directly simulated from the GRRs and RAFs of the associated SNPs and were sampled from case-control–permutated, LD-pruned GWAS replicates for null SNPs. Polygenic analysis simulations incorporated the exact study design as was used for the real data, except that (i) gender or population stratification was not modeled and covariates were not used in the simulated association tests, and (ii) independent SNPs were simulated (for comparison with real data LD pruned to $r^2 < 0.1$). See the **Supplementary Note** for additional details.

Approximate Bayesian computation. To infer the model parameters given the polygenic analysis results, we sampled parameter values from prior distributions, simulated a polygenic analysis and performed an approximate Bayesian computation with rejection sampling and general linear model post-sampling adjustment (ABC-GLM)[24]. See the **Supplementary Note** for full details.

We sampled polygenic disease model parameters from wide-ranging priors that were as 'uninformative' as possible to consider a broad range of genetic architectures for disease risk. For the primary analyses, V_{tot} values were sampled from a uniform prior on the interval 0.01–0.99 (0.01–0.5 for rheumatoid arthritis), and N_{SNPs} values were sampled from a $\log_{10}$-uniform prior on the interval 10–10,000 (that is, $\log_{10} N_{SNPs} \sim U(1, 4)$). Prior distributions for the parameters β_v ($\sim U(1, 10)$), α_{RAF} ($\sim U(0.5, 10)$) and β_{RAF} ($\sim U(0.5, 10)$) were chosen to generate a wide range of prior distributions of risk allele frequencies and effect sizes (**Supplementary Fig. 2**), and we assessed the sensitivity of our inference to alternative priors for N_{SNPs} and β_v.

We used ABC-GLM[24] to perform rejection sampling and post-sampling adjustment to estimate posterior probability densities of the model parameters given our observed polygenic analysis results. We determined rejection-sampling–based Euclidean distances between the simulated and observed transformed summary statistics based on 1,000,000 simulation replicates with a 0.2% acceptance rate for the primary analyses. The observed polygenic analysis results were not significantly less likely under the polygenic model than the accepted simulated results, and the marginal posterior densities of all the individual parameters are shown in **Supplementary Figure 3**. We validated our analysis by performing ABC-GLM with the simulated data and verified (i) that our method successfully inferred the key properties of simple, intuitive underlying disease models and (ii) that the known parameters were roughly uniformly distributed across their posterior probability density quantiles (**Supplementary Fig. 4**).

We extended the ABC-GLM to estimate joint posterior densities and to sample from joint posterior distributions using the Markov chain Monte Carlo method. Samples from the joint posterior of all five parameters were generated by Markov chain Monte Carlo with a uniform updating distribution, and convergence was assessed by comparing samples within and between independent chains and by comparing the samples with marginal densities estimated by AMC-GLM. Full joint posterior samples were used to obtain posterior distributions of the allele frequencies and variances explained (v) for the associated SNPs, which were truncated at a v corresponding to a minimum GRR of 1.01 (for MAF = 0.5; the GRRs were larger for smaller MAFs) to generate posterior distributions of V_{tot} and N_{SNPs} and of MAF and GRR.

ABC-GLM and the supporting analyses were conducted using the ABCtoolbox software package[25]. Extensions to this method are implemented in a new version of ABCtoolbox (see URLs).

Linear mixed-effects model heritability estimation. Heritability caused by common SNPs was directly estimated from the GWAS datasets using an LMM that regressed phenotype on a random-effects kinship matrix estimated from genotyped SNPs, with gender as a fixed effect and including principal component covariates, using the GCTA software[11,14,26]. Kinship matrices were estimated from genotyped SNPs after stringent quality control (missing data rate <1%, case-control differential missing data $P > 0.01$ and Hardy-Weinberg equilibrium $P > 0.001$) and adjusted for finite-SNPs estimation; individuals showing low-level relatedness were removed, and the results were converted to the population-liability scale.

Causal variant modeling. We used 1000 Genomes Project[40] data and HAPGEN software[41] to simulate a case-control population with a range of underlying causal genetic models varying by the allele frequency, number and effect size of the causal variants (**Supplementary Note**, **Supplementary Table 4** and **Supplementary Fig. 5**). We simulated 10-Mb regions with an average SNP density and genetic length, and calculated case and control haplotype frequencies based on the numbers of rare (MAF < 1%) or common (MAF > 5%) causal variants randomly selected from within 100-kb 'loci'. We then calculated

© 2012 Nature America, Inc. All rights reserved.

doi:10.1038/ng.2232

case and control allele frequencies at HapMap2 SNPs (an average of ~8,000 SNPs per region) and LD pruned the SNPs by association (average of ~240 SNPs). These single-locus simulation results were extrapolated genome wide by resampling with replacement until the expected total liability-scale variance explained of induced GWAS SNP associations reached our polygenic modeling Bayesian inference (**Supplementary Table 4**). We identified plausible causal variant models in which the simulated marker SNPs and the inferred Bayesian posterior distributions were consistent in terms of the number of associated marker SNPs and their allele frequencies and effect sizes.

58. Falconer, D. & Mackay, T. *Introduction to Quantitative Genetics* . 4th edn (Longman, 1996).

© 2012 Nature America, Inc. All rights reserved.