Human Molecular Genetics, 2010, Vol. 19, No. 19 **3863–3872**
doi:10.1093/hmg/ddq304
Advance Access published on July 16, 2010

Locus category based analysis of a large genome-wide association study of rheumatoid arthritis

Jan Freudenberg[1,*], Annette T. Lee[1], Katherine A. Siminovitch[2], Christopher I. Amos[3], David Ballard[1], Wentian Li[1] and Peter K. Gregersen[1]

[1]Robert S. Boas Center for Human Genetics and Genomics, The Feinstein Institute for Medical Research, Manhasset, NY, USA, [2]Mount Sinai Hospital, Toronto, ON, Canada and [3]M. D. Anderson Cancer Center, University of Texas, Houston, TX, USA

Received March 1, 2010; Revised and Accepted July 13, 2010

To pinpoint true positive single-nucleotide polymorphism (SNP) associations in a genome-wide association study (GWAS) of rheumatoid arthritis (RA), we categorize genetic loci by external knowledge. We test both the 'enrichment of associated loci' in a locus category and the 'combined association' of a locus category. The former is quantified by the odds ratio for the presence of SNP associations at the loci of a category, whereas the latter is quantified by the number of loci in a category that have SNP associations. These measures are compared with their expected values as obtained from the permutation of the affection status. To account for linkage disequilibrium (LD) among SNPs, we view each LD block as a genetic locus. Positional candidates were defined as loci implicated by earlier GWAS results, whereas functional candidates were defined by annotations regarding the molecular roles of genes, such as gene ontology categories. As expected, immune-related categories show the largest enrichment signal, although it is not very strong. The intersection of positional and functional candidate information predicts novel RA loci near the genes *TEC/TXK*, *MBL2* and *PIK3R1/CD180*. Notably, a combined association signal is not only produced by immune-related categories, but also by most other categories and even randomly defined categories. The unspecific quality of these signals limits the possible conclusions from combined association tests. It also reduces the magnitude of enrichment test results. These unspecific signals might result from common variants of small effect and hardly concentrated in candidate categories, or an inflated size of associated regions from weak LD with infrequent mutations.

INTRODUCTION

Genome-wide genetic association studies (GWAS) provide comprehensive information about the correlation between common genetic variation and phenotypic variation. In recent years, GWAS have led to the identification of genetic loci for many human disease phenotypes (1). The principle behind these studies was the search for single-nucleotide polymorphism (SNP) associations that achieve a genome-wide significance, which was combined with the replication of such findings in independent samples. This strategy requires stringent significance thresholds for reliably distinguishing true from false-positive markers. Because the effect size of most SNP associations is weak, these significance thresholds can be met only with very large samples. Nevertheless, true association signals almost certainly exist below the formal significance threshold for separate genetic loci (2,3). Therefore, the further interpretation and exploitation of the subthreshold signal in GWAS data sets is warranted.

One phenotype for which GWAS have been quite successful is rheumatoid arthritis (RA; MIM180300). RA is a common autoimmune disorder affecting ~1% of individuals in populations of European origin, with its predominant manifestation being inflammation with bone and cartilage destruction in diarthrodial joints. The genetic basis for RA is complex, with several genes accepted as associated with disease in populations of European origin, including *HLA-DRB1*, *PTPN22*, *STAT4*, *TRAF1*, *TNFAIP*, *CD40*,

*To whom correspondence should be addressed at: Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research, 350 Community Drive, Manhasset, NY 11030, USA. Tel: +1 5165621542; Email: jan.freudenberg@nslij-genetics.org

CTLA4 and *REL* (4–7). Similar to other complex diseases, the known susceptibility loci explain only a relatively small fraction of the phenotypic variation (around 20%) (6) and it is apparent that additional risk variants remain to be discovered. A convincing demonstration of individual loci will require further increases in samples size, given that their effect will be quite modest. As a complementary approach, we sought to apply computational methods that use independent knowledge to predict true-positive disease loci in a recently expanded GWAS of RA (7).

In order to integrate biological knowledge into the statistical analysis of GWAS data sets, several strategies have been proposed that use information from multiple SNPs to search for higher-order associations between gene functions and a phenotype (8–22). Most of these strategies look for categories that are overrepresented among loci with strong association signals. The alternative approach does not rely on the comparison of loci in a category to other loci in the genome. It instead only tests whether the loci from a candidate category show a stronger combined signal than would be expected by chance in the absence of any true case–control difference. The former approach has the advantage that it is more robust against unspecific effects, whereas the latter approach may be able to detect weaker signals for certain categories, regardless of whether or not other categories exhibit any associations.

With both strategies, the assignment of genetic loci to categories is required. The respective categories may be determined not only by knowledge about genes and functional sequence elements, but also by the locations of signals arising from other independent GWAS data. In the following text, we will refer to the former as 'functional' candidates and to the latter as 'positional' candidates. Regardless of how categories are defined, the number of SNPs per locus (in the following, referred to as 'SNP density' of a category) is likely to differ among categories, and this is further complicated by the fact that SNPs at a same locus are not independent of each other due to linkage disequilibrium (LD). This can influence the number of loci that is called associated for a category, which has to be considered by methods for category-based analysis of GWAS data. In the present study, we have used LD blocks to define genetic loci and permutation analysis to simulate the correct null distribution and to identify differences between categories that are related to the affection status.

We first show that there exists a considerable excess of weakly associated loci in our data. We further show that this excess signal is enriched at immune candidate loci. However, the enrichment at immune candidate loci is not particularly strong. This can be explained by the finding that a large part of the excess signal is diffusely distributed across the genome. Nevertheless, we can point out novel putative RA loci by inspecting the loci that account for the seen enrichment signals, which demonstrates the potential of the category-based analyses of GWAS data sets.

RESULTS

As described in detail in Materials and Methods, we utilized LD blocks from the HapMap database (23) to define genetic loci and to account for redundant SNP associations (Fig. 1). We refer to any SNP whose association *P*-value in our genotypes is lower than the specified threshold P_0 as 'SNP association' and any LD block that harbors at least one such SNP association as 'associated locus'. We started by comparing the observed and the expected frequency of associated loci for different SNP association parameters. This showed a maximal proportional increase of the total number of associated loci in the observed data when compared with the permuted data for SNP association thresholds near $P_0 < 0.1$ (Fig. 2, Supplementary Material, Fig. S0a), which was highly significant ($P < 10e-16$, $\chi^2 = 72$, $df = 1$). Importantly, an increased number of associated loci in the observed data still existed, when we excluded the major histocompatibility complex (MHC) region and other known RA genes from the present GWAS catalog (1) (Supplementary Material, Fig. S0b).

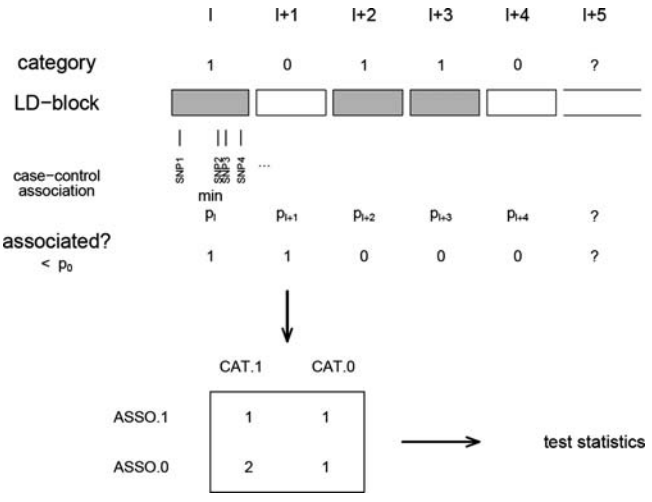

Figure 1. Graphical illustration of the applied procedure for defining the test statistic. The genome is partitioned into non-overlapping LD blocks with labels l, . . . , l + 5. Each block is represented by a certain number of genotyped SNPs. The best-scoring SNP from each block is used to decide whether the locus is significantly associated (ASSO.0 versus ASSO.1) with the phenotype based on the threshold parameter P_0. Blocks are further categorized based on the external information (CAT.0 versus CAT.1). This leads to a 2 × 2 table of count data. To test the enrichment of a category for associated loci, the odds ratio OR_C is calculated from this table. Because loci with more SNPs are more likely to belong to ASSO.1 due to the minimization procedure applied at each locus, the hypergeometric distribution cannot be used to evaluate the OR_C statistic. Our second test statistic NA_C only takes the count of loci in the fields ASSO.1 and CAT.1. Both statistics are evaluated by comparison with the case–control permuted data.

A logical question to ask is, whether this excess of associated loci is concentrated in certain locus categories. To formally answer that question, we used two different metrics. First, we quantified the enrichment of associated loci in a category: we calculated for candidate categories the odds ratio (OR_C) to contain loci with at least one SNP association [defined by Eq. (1) in Materials and Methods]. We defined as our null hypothesis that case–control differences do not contribute to any enrichment (as measured by OR_C) of SNP associations in a category. Of course, independently from any true case–control signal, an OR_C statistic larger than 1 is expected for categories with higher SNP density (i.e. more SNPs per locus) due to the minimization procedure applied to the SNPs at each locus. With our second metric, we quantified the 'combined association of loci' from a category: here,

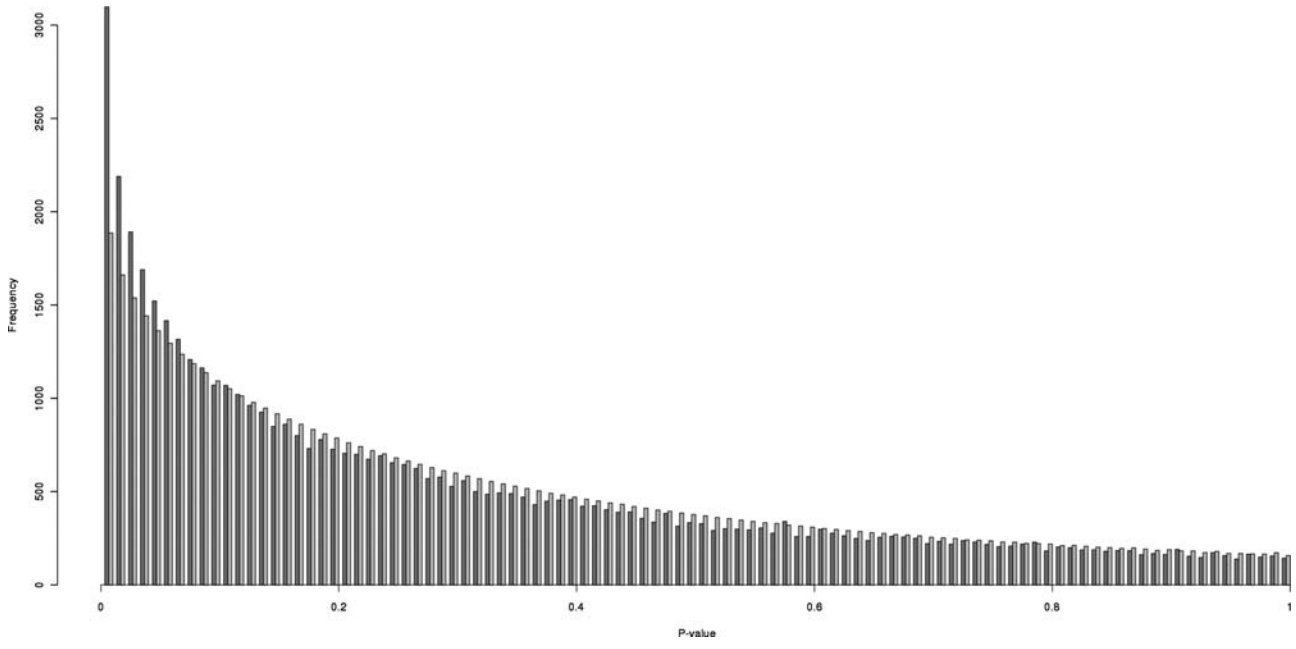

Figure 2. Histogram of distribution of the minimal SNP *P*-value within each of the 51 291 LD blocks that are represented by at least one SNP in the NARAC data. The observed distribution is shown by dark grey bars. The expected distribution (light grey bars) was estimated from the average number of loci within each *P*-value bin over 1000 permutation of the affection status. Because the best-scoring SNP association is used to represent each block, the distribution is strongly skewed toward smaller *P*-values. The skew is present both in the observed and in the expected distribution, but it is stronger in the former, indicating an excess of smaller *P*-values due to the case–control differences.

we simply counted the number NA_C of loci in a category that contain at least one SNP association [formally defined by Eq. (3) in Materials and Methods]. Both the OR_C and the NA_C statistic were compared with their expectation under the null hypothesis of no case–control differences as obtained from the permutation of the affection status. Importantly, the NA_C statistic tests the role of a category independent from any genomic background signal, whereas the OR_C statistic compares a category to the rest of the genome.

Analysis of randomly defined locus categories

We first evaluated the behavior of the enrichment (OR_C) and combined association (NA_C) statistic in the absence of any external biological knowledge, but in the presence of true case–control signal. This was primarily intended to understand the influence of factors like category size and SNP density and how to control for these factors. To this end, we generated 100 random categories for each of eight different size parameters, varying the number of loci in a random category from 12823 to 201 (1/4 to 1/256 of all LD blocks). Thus, random categories were sampled from the list of all LD blocks that contain at least one genotyped SNP, including blocks with and without any coding regions. We then applied the thresholds $P_0 < 0.1$ and $P_0 < 0.001$ for calling SNP associations and associated loci for these random categories. This showed that OR_C was on average close to 1, as expected, and independent of category size (by the term 'category size', we refer to the number of loci that belong to a category). However, the OR_C statistic showed an increased variance for smaller categories (those with fewer loci) and under more stringent threshold parameters P_0 (Supplementary Material, Fig. S1a). Because this property of OR_C hampers its comparability across parameters, we further used the permutation of the affection status to calculate a normalized score nOR_C [as defined by Eq. (2)]. The variance of this normalized enrichment score nOR_C does not depend on the category size or the SNP association threshold P_0 and is distributed around zero for random categories, as expected (Supplementary Material, Fig. S1b).

Using our second test statistic NA_C [defined by Eq. (3)], we next addressed the question whether random categories show any combined association. Because the distribution of the NA_C statistic also depends on category size and P_0 (Supplementary Material, Fig. S2a), NA_C was analogously normalized as OR_C [Eq. (4)]. Interestingly, after normalization, the nNA_C scores of random categories are mostly greater than their expected value of zero (Supplementary Material, Fig. S2b). This pattern is particularly pronounced for large categories (categories with many loci) and under loose thresholds P_0. To further confirm that this increased nNA_C score of random categories is indeed due to the case–control signal in our data, we calculated the nNA_C score for each category for each of 1000 permuted data sets (where the case–control signal is removed). The mean of these nNA_C scores of random categories (averaged from the 1000 nNA_C scores from the permuted data sets) is very close to its expected value of zero (Supplementary Material, Fig. S2c). Thus, the above nNA_C scores greater than zero are a consequence of case–control differences in the actual GWAS data. Obviously, such combined associations of random categories raise to question the specificity of possible combined association results for other candidate categories.

To further evaluate the influence of SNP density (the number of SNPs per locus in a category) on the category

enrichment test statistic OR_C, we constrained random categories by the requirement that loci harbor at least the average of five genotyped SNPs. Thus, we randomly sampled category members only from the set of LD blocks with at least five SNPs, which produced categories with a mean/median SNP density of ~10 SNPs per block. Unsurprisingly, these categories display on average >2-fold increased OR_C statistic (Supplementary Material, Fig. S3a). However, after normalization, the nOR_C scores of these categories were still greater than zero (Supplementary Material, Fig. S3b). This result is somewhat surprising, because nOR_C is designed to correct for the influence of SNP density. That nOR_C properly corrects for the influence of SNP density is demonstrated by the nORc scores that were obtained for 1000 case–control permuted data sets for each of the 100 random categories. In these data sets, under the absence of any true case–control signal, constrained random categories with high SNP density display nOR_C scores very close to zero (Supplementary Material, Fig. S3c). Thus, the increased nOR_C scores of categories with higher SNP density indicate that their loci are more likely to contain susceptibility mutations or that their loci are more likely to capture association signals that originate from neighboring blocks.

One may further ask how coding regions compare with non-coding regions. To answer that question, we constrained random categories such that all their loci overlap at least one coding exon. This showed an increased nORc score for such categories (Supplementary Material, Fig. S3d), which is consistent with the expectation of more susceptibility mutations in coding regions. Another question to ask is how SNP density influences the combined association analysis. When again constraining random categories to have at least five genotyped SNPs at each locus, we saw that the nNA_C score was increased when compared with categories of equal or smaller size without this constraint (Supplementary Material, Fig. S3e). This result would be expected, because categories with higher SNP density still tend to be enriched for associated loci after correcting for the effect of SNP density on the test statistic, as seen above.

Analysis of positionally defined locus categories

In the next step, we defined positional candidate categories based on SNP associations in the earlier Wellcome Trust Case–Control Consortium (WTCCC) GWAS (4). We defined LD blocks as positional candidate loci, if they contained at least one SNP association in the WTCCC study for the threshold parameter P_{WTCCC} under the frequentist additive model. Thus, positional candidate categories were now constructed based on SNP associations in the WTCCC GWAS with RA, type 1 diabetes (T1D), type 2 diabetes (T2D), Crohn's disease (CD), bipolar disorder (BD), coronary artery disease (CAD) and hypertension (HT). Then, we looked for associated loci in the North American Rheumatoid Arthritis Consortium (NARAC) data set that map to these positionally defined locus categories.

When looking at the positional candidates defined by the RA phenotype, we found them enriched among associated loci in the NARAC data. This is consistent with an overproportionally large overlap between the results of the two independent GWAS for RA (Table 1, Supplementary Material, Table S1). When applying loose thresholds for retrieving positional candidate loci from the WTCCC GWAS and for defining associated loci in the NARAC study ($P_{WTCCC} < 0.1$ and $P_0 < 0.1$), the enrichment signal is weak ($nOR_C = 1.39$) and only shows a non-significant trend ($P = 0.09$). However, when increasing the stringency of either P_{WTCCC} or P_0, this enrichment becomes more prominent (Supplementary Material, Fig. S4). Accordingly, the enrichment is strongest ($nOR_C = 5.81$, $P < 0.001$), for the most stringent SNP association thresholds ($P_{WTCCC} < 0.0001$ and $P_0 < 0.0001$).

We next wanted to know whether other WTCCC autoimmune loci (CD and T1D) are enriched among RA loci and whether this enrichment was absent for the WTCCC phenotypes that are not typically viewed as autoimmune disorders (BD, CAD, HT and T2D). Consistent with our expectation of a shared genetic etiology of autoimmune disease, we found an enrichment of T1D and CD hits among associated loci (Supplementary Material, Table S1), with enrichment scores nORc ranging from 0.35 to 3.77 (corresponding to *P*-values from 0.37 down to ≤0.001). As seen above for the WTCCC RA loci, the strength of enrichment scores for the other immune disease loci increases, when applying a more stringent threshold. Contrarily, more moderate enrichment scores exist for positional candidates defined by non-immune phenotypes from the WTCCC GWAS (BD, CAD, HT and T2D), where nORc scores range from −0.46 to 1.13 (corresponding to *P*-values from 0.69 to 0.13).

In addition to looking at the enrichment of positional WTCCC candidate loci by means of OR_C, we also looked at their combined association as tested by NA_C. These tests showed a highly significant increase across parameters and candidate categories (Fig. 3, Supplementary Material, Table S1). For positional candidates defined by any of the seven disease phenotypes, the nNA_C score was strongly increased under weak SNP association thresholds ($P_0 < 0.1$ and $P_{WTCCC} < 0.1$), whereas it was moderately increased for the more stringent choice of P_0 and P_{WTCCC}. However, only under more stringent SNP association thresholds, the magnitude of these combined associations (as measured by nNA_C) is visibly larger for autoimmune than non-immune categories (Fig. 3), which in turn produces an enrichment of associated loci (as measured by nOR_C).

To better understand the above enrichment of WTCCC immune candidate loci at associated loci from the NARAC study, we next retrieved those LD blocks that harbor SNPs with $P_0 < 0.001$ in both GWAS (Supplementary Material, Table S2). A notable fraction of these LD blocks is located within the MHC region, which raises the question whether non-MHC loci are sufficient to produce a significant overlap among studies. Therefore, we repeated the above analysis without any loci from this region (20–40 Mb on chromosome 6). This still showed an enrichment of WTCCC RA loci among NARAC RA loci at least for the parameter $P_0 < 0.001$, whereas the enrichment of CD or T1D loci was now rendered non-significant (Supplementary Material, Table S3). On the other hand, the combined association of candidate loci from any of the seven disease phenotypes was only marginally altered by exclusion of the MHC region and remained highly significant in particular for weak SNP association thresholds.

Table 1. Results of the category-based analysis of immune system candidate loci

Locus category	NARAC SNP association threshold (P_0)	Number of associated loci in category	Number of non-associated loci in category	Number of associated loci not in category	Number of non-associated loci not in category	Normalized score for enrichment (nOR_C)	P-value for enrichment	Normalized score for combined association (nNA_C)	P-value for combined association
WTCCC RA ($P < 10e-03$)	0.1	315	261	16 247	34 468	1.96	0.025	10.58	$\leqslant 0.0001$
	0.001	63	513	480	50 235	3.77	0.001	5.15	$\leqslant 0.0001$
Mouse Immune System Phenotype	0.1	2333	3992	14 229	30 737	1.75	0.044	9.96	$\leqslant 0.0001$
	0.001	102	6223	441	44 525	1.59	0.041	6.39	$\leqslant 0.0001$

LD blocks are categorized as candidate loci, if they harbor an SNP association with RA with $P_{WTCCC} < 0.001$ in the WTCCC study or if the mouse ortholog of their nearest gene is annotated with an immune system phenotype. These two candidate categories are analyzed for RA-associated loci in the NARAC data, where the threshold for calling SNP associations is set to $P_0 < 0.1$ and $P_0 < 0.001$, respectively. For the two candidate categories, the numbers of associated and non-associated loci under the respective threshold P_0 are shown. The enrichment of associated loci in a category is tested by the odds ratio of a category (OR_C), whereas the combined association is tested by the number of associated loci in a category (NA_C). Both the OR_C and the NA_C test statistics are normalized based on their simulated null distribution to obtain the normalized scores nOR_C and nNA_C.

Figure 3. Combined association scores of positional candidate categories. LD blocks are categorized as positional candidate loci, if they harbored SNP associations under the threshold $P_{WTCCC} < 0.1$ (dark grey bars) or $P_{WTCCC} < 0.001$ (light grey bars) with any of the seven phenotypes from the WTCCC GWAS. Associated loci were called based on SNP associations with $P_0 < 0.001$ in the NARAC GWAS. This analysis shows the presence of more associated loci than expected among immune candidate loci (as defined by the phenotypes RA, CD and T1D) and non-immune candidate loci (as defined by the phenotypes BD, CAD, HT and T2D). Immune candidates have visibly larger combined association score than non-immune candidate for $P_{WTCCC} < 0.001$ (light grey), but not for $P_{WTCCC} < 0.1$ (dark grey).

Analysis of functionally defined locus categories

We finally used functional genome annotations to define RA candidate loci in the human genome. To this end, we retrieved the list of genes with an immune system phenotype annotated to their mouse ortholog (24). In total, we found 1929 human genes for which the term 'immune system phenotype' was annotated to the respective mouse ortholog gene. These mapped to 6325 LD blocks that are covered by 40 663 SNPs in our data. As expected, this set of LD blocks is enriched for RA-associated loci, but the increase is not very strong (Table 1).

For moderate SNP association thresholds ($P_0 < 0.1$), the enrichment of associated loci at mouse immune loci is borderline significant ($nOR_C = 1.75$, $P = 0.04$). For the more stringent threshold $P_0 < 0.001$, this remained largely unchanged ($nOR_C = 1.59$, $P = 0.04$; Table 1). Nevertheless, it is now interesting to look at the intersection of loci that belong to both functional and positional candidate categories for RA. For the SNP association threshold of $P_0 < 0.001$, these loci included the already known RA genes *PTPN22*, *CTLA4*, *TNFRSF14* and *TNFAIP3* (Table 2). In addition, both NARAC and WTCCC data support an RA locus near *TEC/TXK*. Several other genes are located in the same LD block as *TEC* and its adjacent neighbor *TXK*, but only these two are associated with immune function in the mouse, and *TEC* harbors the most significant SNP association in this block. Both the NARAC and the WTCCC CD data furthermore support an autoimmune locus near the functional candidate gene *MBL2* and another locus between *PIK3R1* and *CD180*.

Table 2. Loci that belong to the intersection of functional candidates (mouse immune system loci) and positional candidates (immune disease associations in the earlier WTCCC GWAS)

NARAC SNP	$-\log_{10}$(NARAC-Pval)	Gene(s) mapped to locus	WTCCC phenotype	WTCCC SNP
rs2476601	16	*MAGI3*, *PHTF1*, *PTPN22*, *RSBN1*, *BCL2L15*, *AP4B1*, *DCLRE1B*	RA, CD, T1D	rs10858002
rs6748358	5.2	*CTLA4*, *ICOS*	T1D	rs231790
rs2327832	4.8	*OLIG3*, *TNFAIP3*	RA	rs2327832
rs1903942	3.2	*MBL2*	CD	rs1903929
rs4976139	3.8	*PIK3R1*	CD	rs12234107
rs2089510	3.4	*CORIN*, *NFXL1*, *CNGA1*, *NIPAL1*, *TXK*, *TEC*	RA	rs7679010
rs3890745	3.7	*PLCH2*, *PANK4*, *HES5*, *TNFRSF14*, *C1orf93*, *MMEL1*	RA	rs10910097

For each locus, the rs numbers are printed for the SNP with the strongest association signal at this locus in the respective data set.

Genes annotated with an immune system phenotype in the mouse are strong functional RA candidate loci, but they are functionally quite diverse. Therefore, we next categorized LD blocks based on gene ontology (GO) annotations, which provide functional annotations on multiple levels of specificity. We assigned GO annotations to LD blocks using the GO annotations of their nearest gene and dropped all those GO categories that were annotated to less than 200 or more than 20 000 blocks, which led to 1252 GO categories. When ranking GO categories by their nOR_C scores, most of the leading categories are related to immune system functions (Table 3, Supplementary Material, Table S4a). Further GO categories with an increased nOR_C score include 'cell surface receptor-linked signal transduction' and 'transcription'. Owing to their relatively high absolute numbers of associated loci, these enrichment signals should be robust against a dominating role of a subset of potentially misleading loci. Accordingly, after excluding the MHC region, similar GO categories attained the largest enrichment scores (Supplementary Material, Table S4b). However, the magnitude of these enrichment scores is again not large. If one would further perform an (albeit conservative) Bonferroni correction for the number of GO categories tested, hardly any GO category would meet formal significance thresholds.

Notably, the majority of GO categories show a combined association with RA, as indicated by nNA_C scores greater than zero (Supplementary Material, Table S4a). Thus, most GO categories have more associated loci in the observed than in the permuted data. Accordingly, the comparison of the distribution of empirical *P*-values for the NA_C statistic (Supplementary Material, Fig. S5b) is clearly different from that for the OR_C statistic (Supplementary Material, Fig. S5a) over the 1252 included GO categories. Most GO categories attain a significantly increased NA_C statistic, despite the fact that only few have a significantly increased OR_C statistic. When case–control differences are spread over many categories, the NA_C statistic of a category can be highly significant whereas OR_C is not significant, because OR_C compares locus association signals with the background of other loci, whereas NA_C tests a category independent from any background signal. This leads to the situation, where most GO categories display significantly more RA-associated loci than expected by chance, but only a few GO categories are enriched for associated loci when compared with the remaining genome.

We finally retrieved all LD blocks, which harbor SNP associations with $P_0 < 0.001$, which belong to the GO categories immune response/immune system process (GO:0006955/GO:0002376) or which are located near mouse immune genes. This produced an expanded list of 124 LD blocks, several of them located near confirmed RA genes (Supplementary Material, Table S5). More than 100 of these loci are found outside the MHC region. Given the autoimmune nature of RA and the immune function annotations of these loci, they might be considered prime targets for further replication and fine-mapping studies.

DISCUSSION

We have applied two different formal tests for the locus category-based analysis of a large GWAS of RA. With our first strategy, we tested the enrichment of a category by calculating the ratio between the odds for loci from the category and the odds for loci not from the category to harbor at least one SNP association. With our second strategy, we tested the combined association of a category by counting the number of loci in a category that harbor at least one SNP association. For both strategies, we used HapMap LD blocks to minimize redundant SNP associations and we further performed the permutation of the affection status to calculate normalized scores. Future implementations might extend these strategies toward the proposed method of randomly redrawing SNPs until they cover the same number of genes as observed (14), i.e. to redraw SNPs until they cover the same number of LD blocks as observed.

Our analysis of random categories demonstrates that SNP density, category size and the SNP association threshold P_0 exert an influence on the distribution of the test statistic. However, the normalized scores correct for these factors (and also correlate well with empirical *P*-values as shown in Supplementary Material, Fig. S6). After normalization, we notice that random categories often display a combined association signal, which is particularly pronounced for moderate SNP association thresholds and large categories. This could be explained by a multitude of common risk alleles with weak effects on the phenotype, as has been proposed for schizophrenia (3). Under this explanation, one would most likely expect that weak SNP associations are concentrated in plausible candidate categories. A second, not mutually exclusive, explanation would be that infrequent disease variants lead to weak SNP associations in the assayed common variation. Because infrequent disease variants may be evolutionarily more recent, they may often exist on longer haplotypes (25,26). This may

Table 3. Analysis of functional candidate loci defined by GO annotations

Locus category	Description	Number of associated loci in a category	Number of non-associated loci in a category	Number of associated loci not in a category	Number of non-associated loci not in a category	Normalized score for odds ratio (nOR_C)	*P*-value for odds ratio	Normalized score for number of associated loci (nNA_C)	*P*-value for number of associated loci
GO:0006955	Immune response	48	1462	495	49 286	2.44	0.001	4.45	≤0.0001
GO:0050896	Response to stimulus	136	7623	407	43 125	2.41	0.003	7.78	≤0.0001
GO:0002376	Immune system process	58	2452	485	48 296	2.35	0.002	5.28	≤0.0001
GO:0002253	Activation of immune response	11	199	532	50 549	2.30	0.009	4.21	≤0.0001
GO:0006606	Protein import into nucleus	14	274	529	50 474	2.25	0.009	3.96	≤0.0001
GO:0051170	Nuclear import	14	285	529	50 463	2.21	0.010	3.92	≤0.0001
GO:0007165	Signal transduction	172	10 787	371	39 961	2.17	0.011	8.3	≤0.0001
GO:0042742	Defense response to bacterium	10	221	533	50 527	2.11	0.012	4.12	≤0.0001
GO:0034504	Protein localization in nucleus	14	286	529	50 462	2.11	0.013	3.79	≤0.0001
GO:0004872	Receptor activity	94	5131	449	45 617	2.09	0.009	6.16	≤0.0001
GO:0006350	Transcription	83	4179	460	46 569	2.04	0.011	6	≤0.0001
GO:0007166	Cell surface receptor-linked signal transduction	95	5333	448	45 415	1.97	0.016	6.27	≤0.0001

Loci are categorized based on the GO annotations of their nearest gene. The 12 GO categories with the largest normalized nOR_C score under the SNP association threshold of $P_0 < 0.001$ are shown. The full list, including the remaining 1240 GO categories, is provided in Supplementary Material, Table S4a.

result in weak associations with alleles in larger distance to disease mutations. This could inflate the number of associated loci in the observed data when compared with the permuted data, because the correlation of infrequent mutations with common SNPs might be poorly reflected in LD patterns among common SNPs. That rare variants may generate association signals for common variants in relatively large distance was further supported by a recent simulation study (27). Finally, we have considered the, also not mutually exclusive, possibility that there is uncorrected population stratification in the data or that disease-associated alleles are themselves ancestry informative. However, we have corrected for the effects of population stratification by principal component analysis (28), making a major role of this explanation unlikely.

Not surprisingly, for categories defined positionally by the earlier WTCCC GWAS (4), we saw an enrichment of autoimmune categories (RA, T1D and CD). After excluding the MHC region, this overlap remained significant only for RA and for more stringent SNP association thresholds. When further using functional annotations to define categories, we found categories related to immune function enriched for RA-associated loci as expected. These immune candidate categories necessarily show a combined association signal. However, non-immune categories also display a rather strong combined association. These non-specific signals may overshadow weak effect signals in category enrichment analyses with less stringent SNP association thresholds. Accordingly, the considerable amount of weak SNP associations that exists in our data is hardly found concentrated even in immune candidate categories.

One motivation for carrying out this analysis was to develop evidence for additional loci involved in susceptibility to RA. To this end, we looked at the list of loci from the intersection of those positional and functional candidate categories, which display an enrichment signal. This points to three unrecognized loci near mouse immune genes that display SNP associations both in the NARAC study and with an autoimmune phenotype in the WTCCC GWAS (with a *P*-value of <0.001). These include loci closest to *TEC*, which plays a role inflammation-induced bone destruction (29), *PIK3R1*, its loss resulting in a marked reduction in *REL* expression (30), consistent with the established role of *REL* in RA (7), and *MBL2*, which cooperates with toll-like receptors in innate immune response (31). Clearly, more studies are necessary to confirm and fine map these loci in RA susceptibility, as well as to identify the causative genetic variation that underlies the extensive weak association signal for RA across the human genome.

MATERIALS AND METHODS

Data sources

Genotype data for 2418 RA cases and 4504 controls were obtained from a collaborative study with the NARAC and have been described in detail elsewhere (7). In total, our analysis was based on the comparison of the allele frequency of 270 343 autosomal SNP markers. In order to minimize the potential influence of population stratification on the results, we corrected genotypes and phenotypes along the 10 major

principal components, as implemented in the Eigenstrat program (28). SNPs from the MHC region (chromosome 6: 29–33.5 Mb) were excluded from the principal component analysis.

SNP association results for RA, CD, T1D, T2D, BD, HT and CAD were retrieved from the WTCCC database (4). To define positional candidates based on the GWAS results from these phenotypes, SNP markers from the WTCCC study were assigned to the LD blocks in which they are located.

Gene model annotations were retrieved from the Ensembl database (www.ensembl.org) (32). Gene function annotations of human genes were retrieved from the file 'gene_associations.goa_human' provided by the GO database (www.geneontology.org) (33,34). Because GO aims to annotate genes as specifically as possible, annotations were expanded to less specific terms where required. Mouse immune system annotations for human gene orthologs were retrieved based on the term 'MP:0005387' from the file 'HMD_HumanPhenotype.rpt' provided by the Mouse Genome Informatics database (www.informatics.jax.org) (24).

Statistical methods for locus category analysis

Given that the genome is partitioned into N_L separate loci, each locus l belongs to category C or it does not belong to C. If one or multiple SNPs are mapped to a locus l, the case–control association score of l is given by its best-scoring SNP as $p_l = \min(p_{l\_i})$, where i runs through all SNPs at the locus l. If only one mutated haplotype exists in a candidate region, the best-scoring SNP was shown in earlier analyses to be a powerful representation of this region (10,35). A threshold is set at P_0, such that when $p_l < P_0$, the locus l is called associated with the phenotype. This leads to a 2×2 table that may be used for enrichment testing (Fig. 1). Because the threshold parameter for which a signal is called significant is known to play a role in gene expression analysis (36), we looked at different thresholds P_0 for which we called a locus associated.

Denote the identity symbol by I (i.e. $I(x) = 1$ if x is true and $I(x) = 0$ if x is untrue), we now define the odds ratio statistic $\mathrm{OR_C}$ to measure the enrichment of phenotype-associated loci in category C as:

$$\mathrm{OR_C} = \frac{\sum_{l=1}^{N_L} I(l \in C \wedge p_l < P_0) / \sum_{l=1}^{N_L} I(l \in C \wedge p_l \geq P_0)}{\sum_{l=1}^{N_L} I(l \notin C \wedge p_l < P_0) \sum_{l=1}^{N_L} I(l \notin C \wedge p_l \geq P_0)} \quad (1)$$

Note that (i) $\mathrm{OR_C}$ implicitly depends on the threshold value P_0; (ii) if associated loci are randomly distributed over categories, one may expect $\log(\mathrm{OR_C})$ to be normally distributed and with mean zero; (iii) the distribution of $\mathrm{OR_C}$ under the null hypothesis of no enrichment signal due to true-positive SNP associations will be influenced by factors like SNP density; (iv) $\mathrm{OR_C}$ is undefined when one of the sums is equal to zero.

Categories with higher SNP density (i.e. more SNPs per block) are expected to be enriched for associated blocks independent from any true case–control differences, because $\min(p_{l\_i})$ will be smaller, when the number of SNPs within an LD block is larger. Therefore, Fisher's exact test cannot be used to evaluate the significance of $\mathrm{OR_C}$. To determine the significance of an observed $\mathrm{OR_C}$ statistic, we therefore estimated its empirical P-value as the fraction of case–control permutations, for which the $\mathrm{OR_C}$ statistic for the loci from category C is at least equally large as in the observed data. Thus, category P-values close to zero denote an observed statistic that exceeds the expectation, whereas P-values close to 1 indicate an observed statistic below the expectation.

We furthermore used permutation analysis to define a normalized score $\mathrm{nOR_C}$ based on the simulated null distribution of $\mathrm{OR_C}$ from K permutations of the affection status:

$$\mathrm{nOR_C} = \frac{\log(\mathrm{OR_{C\_obs}}) - \mathrm{mean}(\log(\mathrm{OR_{C\_permuted}}), K)}{\mathrm{SD}(\log(\mathrm{OR_{C\_permuted}}), K)} \quad (2)$$

Note (i) the purpose of this transformation is to construct a standard normal distribution and (ii) the value of $\mathrm{nOR_C}$ is typically not sensitive to the number of permutations K. Taking the difference between the observed statistic and the mean of the expected statistic corrects for the influence of the SNP density of a category on $\mathrm{OR_C}$. The division by the standard deviation additionally corrects for factors that influence the variance of $\mathrm{OR_C}$ (such as category size and the value of the parameter P_0). A similar procedure was applied elsewhere to normalize a weighted Kolmogorov–Smirnov-like running sum statistic (10). As would be expected, our normalized score is highly correlated with the empirical P-value across categories of different size or different thresholds P_0 (Supplementary Material, Fig. S6a and b).

As a second test statistic that evaluates the combined association of loci in category C, we used the number $\mathrm{NA_C}$ of associated loci in C:

$$\mathrm{NA_C} = \sum_{l=1}^{N_L} I(l \in C \wedge p_l < P_0) \quad (3)$$

The number of associated loci $\mathrm{NA_C}$ necessarily depends on the size of a category and the applied threshold parameter P_0. To calculate for each category a normalized score that corrects for the influence of SNP density, category size and P_0, we again calculated a normalized score $\mathrm{nNA_C}$ as follows:

$$\mathrm{nNA_C} = \frac{\log(\mathrm{NA_{C\_obs}}) - \mathrm{mean}(\log(\mathrm{NA_{C\_permuted}}), K)}{\mathrm{SD}(\log(\mathrm{NA_{C\_permuted}}), K)} \quad (4)$$

This approach provides a score for the combined signal that originates from a set of predefined 'loci', whereas an earlier method scored a set of predefined 'alleles' (3).

LD blocks as unit of association

To delineate associated loci, we made use of recombination hotspot predictions that were retrieved from the file 'hotspots.txt' from the human HapMap phase II database (www.hapmap.org) (23). These hotspots had been inferred from the patterns of LD in the HapMap data set across the three HapMap populations of African, Asian and European origin.

Each recombination hotspot and each interval between hotspots were defined as LD block. Because blocks are not overlapping, each SNP is contained in exactly one block. A total of 51 291 autosomal blocks were represented by at least one SNP in the NARAC data. An average of 5.1 SNPs were located in each block, but there were also many blocks represented by only one or two SNPs (Supplementary Material, Fig. S7).

If SNPs would be treated as genetic loci instead of LD blocks, association signals in genomic regions with more extensive LD would receive a disproportionately large weight. This is likely to relevant, because human gene functions are known to differ systematically in LD (23,37,38). Therefore, SNPs need to be decorrelated. One possibility would be to perform an initial LD pruning of genotyped SNPs, but this strategy is ignorant with respect to the actual locations of SNP associations and it wipes out a notable fraction of the case–control signal. For instance, LD pruning of our set of genotyped SNPs based on a pairwise threshold of $r^2 < 0.05$ (using a window size of 100 SNPs and an overlap of 25 as additional parameter with PLINK) leads to a reduced set of only 15 612 SNPs. This corresponds to a substantial reduction in statistical power to detect true SNP associations. In several earlier studies, genes were viewed as candidate loci and represented by their best-scoring SNP (10–14,17–21). However, larger genes may receive a disproportionately small weight. Moreover, certain types of candidate loci, such as those arising from independent GWAS, are not primarily defined in terms of any gene annotations.

We therefore used LD block annotations as a computationally efficient way for decorrelating SNP associations and for partitioning SNPs into separate loci. Multiple SNP associations within a same block are considered as mutually dependent, whereas SNP associations from different blocks were considered to belong to different loci. The possibility of multiple associated haplotypes within a block may cause non-redundant SNP association within a block, whereas the presence of LD among SNPs from different blocks may lead to dependency across blocks. To obtain an impression on the extent to which this might apply, we determined the fraction of pairs of SNP above various LD thresholds, distinguishing SNPs located at different or at the same locus. For that purpose, we made use of LD estimates for the European population from the HapMap database that are provided for all SNP pairs within 250 kb. This showed many SNP pairs from 'different' LD blocks which display weak LD (~33% of SNP pairs have $r^2 > 0.05$), but only few SNP pairs which display strong LD (0.06% of pairs have $r^2 > 0.8$). These fractions would be slightly increased when partitioning SNPs based on gene annotations, despite the much smaller number of resulting loci (34 and 0.26%, respectively). Vice versa we see that SNP pairs from the 'same' blocks often show at least some LD (34% of pairs have $r^2 > 0.2$), which would be less when using gene annotations to partition SNPs (12%).

Positional candidate loci were defined based on the location of an SNP association from the WTCCC GWAS (4). Functional candidate loci were defined based on the assignment of the annotations of the nearest gene for each LD block. Where multiple genes are mapped to one block, the annotations of all genes were assigned to this LD block. Where different genes in a block shared functional annotations, these annotations were assigned to the block only once. Most GO categories acquired between 250 and 1000 blocks (Supplementary Material, Fig. S8).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their detailed comments on the manuscript.

Conflict of Interest statement: None declared.

FUNDING

Part of this work was supported by NIH grants NO1-AR-2-2263 and RO1 AR44422. J.F. is supported by a NARSAD Young Investigator Award. K.A.S. is supported by a Canada Research Chair, the Sherman Family Chair in Genomic Medicine and Canadian Institutes for Health Research grants MOP79321 and IIN84042.

REFERENCES

1. Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
2. Evans, D.M., Visscher, P.M. and Wray, N.R. (2009) Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet.*, **18**, 3525–3531.
3. Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F. and Sklar, P. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**, 748–752.
4. Wellcome_Trust_Case_Control_Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
5. Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Plant, D., Gibbons, L.J., Wilson, A.G., Bax, D.E. *et al.* (2008) Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat. Genet.*, **40**, 1156–1159.
6. Raychaudhuri, S., Remmers, E.F., Lee, A.T., Hackett, R., Guiducci, C., Burtt, N.P., Gianniny, L., Korman, B.D., Padyukov, L., Kurreeman, F.A. *et al.* (2008) Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat. Genet.*, **40**, 1216–1223.
7. Gregersen, P.K., Amos, C.I., Lee, A.T., Lu, Y., Remmers, E.F., Kastner, D.L., Seldin, M.F., Criswell, L.A., Plenge, R.M., Holers, V.M. *et al.* (2009) REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.*, **41**, 820–823.
8. Hoh, J. and Ott, J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.*, **4**, 701–709.
9. Lesnick, T.G., Papapetropoulos, S., Mash, D.C., Ffrench-Mullen, J., Shehadeh, L., de Andrade, M., Henley, J.R., Rocca, W.A., Ahlskog, J.E. and Maraganore, D.M. (2007) A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.*, **3**, e98.
10. Wang, K., Li, M. and Bucan, M. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.
11. Torkamani, A., Topol, E.J. and Schork, N.J. (2008) Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, **92**, 265–272.

12. Askland, K., Read, C. and Moore, J. (2009) Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. *Hum. Genet.*, **125**, 63–79.
13. Baranzini, S.E., Galwey, N.W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B.M., Kappos, L., Polman, C.H. *et al.* (2009) Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, **18**, 2078–2090.
14. Holmans, P., Green, E.K., Pahwa, J.S., Ferreira, M.A., Purcell, S.M., Sklar, P., Owen, M.J., O'Donovan, M.C. and Craddock, N. (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am. J. Hum. Genet.*, **85**, 13–24.
15. Hong, M.G., Pawitan, Y., Magnusson, P.K. and Prince, J.A. (2009) Strategies and issues in the detection of pathway enrichment in genome-wide association studies. *Hum. Genet.*, **126**, 289–301.
16. O'Dushlaine, C., Kenny, E., Heron, E.A., Segurado, R., Gill, M., Morris, D.W. and Corvin, A. (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, **25**, 2762–2763.
17. Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D. and Daly, M.J. (2009) Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.*, **5**, e1000534.
18. Sun, J., Jia, P., Fanous, A.H., Webb, B.T., van den Oord, E.J., Chen, X., Bukszar, J., Kendler, K.S. and Zhao, Z. (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases-schizophrenia as a case. *Bioinformatics*, **25**, 2595–6602.
19. Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J.P., Russell, R.K., Sleiman, P.M., Imielinski, M., Glessner, J., Hou, C. *et al.* (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. *Am. J. Hum. Genet.*, **84**, 399–405.
20. Yu, K., Li, Q., Bergen, A.W., Pfeiffer, R.M., Rosenberg, P.S., Caporaso, N., Kraft, P. and Chatterjee, N. (2009) Pathway analysis by adaptive combination of P-values. *Genet. Epidemiol.*, **33**, 700–709.
21. Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J.D., Jin, L. *et al.* (2010) Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur. J. Hum. Genet.*, **18**, 111–117.
22. Ruano, D., Abecasis, G.R., Glaser, B., Lips, E.S., Cornelisse, L.N., de Jong, A.P., Evans, D.M., Davey Smith, G., Timpson, N.J., Smit, A.B. *et al.* (2010) Functional gene group analysis reveals a role of synaptic heterotrimeric G proteins in cognitive ability. *Am. J. Hum. Genet.*, **86**, 113–125.
23. International_HapMap_Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
24. Shaw, D.R. (2009) Searching the mouse genome informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Curr. Protoc. Bioinformatics*, **Chapter 1**, Unit1 7.
25. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
26. Gusev, A., Lowe, J.K., Stoffel, M., Daly, M.J., Altshuler, D., Breslow, J.L., Friedman, J.M. and Pe'er, I. (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res.*, **19**, 318–326.
27. Dickson, S.P., Wang, K., Krantz, I., Hakonarson, H. and Goldstein, D.B. (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol.*, **8**, e1000294.
28. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
29. Shinohara, M., Koga, T., Okamoto, K., Sakaguchi, S., Arai, K., Yasuda, H., Takai, T., Kodama, T., Morio, T., Geha, R.S. *et al.* (2008) Tyrosine kinases Btk and Tec regulate osteoclast differentiation by linking RANK and ITAM signals. *Cell*, **132**, 794–806.
30. Matsuda, S., Mikami, Y., Ohtani, M., Fujiwara, M., Hirata, Y., Minowa, A., Terauchi, Y., Kadowaki, T. and Koyasu, S. (2009) Critical role of class IA PI3K for c-Rel expression in B lymphocytes. *Blood*, **113**, 1037–1044.
31. Ip, W.K., Takahashi, K., Moore, K.J., Stuart, L.M. and Ezekowitz, R.A. (2008) Mannose-binding lectin enhances Toll-like receptors 2 and 6 signaling from the phagosome. *J. Exp. Med.*, **205**, 169–181.
32. Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
33. Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R. and Apweiler, R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res.*, **32**, D262–D266.
34. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
35. Ballard, D.H., Cho, J. and Zhao, H. (2010) Comparisons of multi-marker association methods to detect association between a candidate region and disease. *Genet. Epidemiol.*, **34**, 201–212.
36. Fury, W., Batliwalla, F., Gregersen, P.K. and Li, W. (2006) Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **1**, 5531–5534.
37. Smith, A.V., Thomas, D.J., Munro, H.M. and Abecasis, G.R. (2005) Sequence features in regions of weak and strong linkage disequilibrium. *Genome Res.*, **15**, 1519–1534.
38. Freudenberg, J., Fu, Y.H. and Ptacek, L.J. (2007) Enrichment of HapMap recombination hotspot predictions around human nervous system genes: evidence for positive selection ? *Eur. J. Hum. Genet.*, **15**, 1071–1078.