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

Jan Freudenberg1,*, Annette T. Lee1, Katherine A. Siminovitch2, Christopher I. Amos3, David Ballard1, Wentian Li1 and Peter K. Gregersen1

1Robert S. Boas Center for Human Genetics and Genomics, The Feinstein Institute for Medical Research, Manhasset, NY, USA, 2Mount Sinai Hospital, Toronto, ON, Canada and 3M. 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,
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 P0 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 P0 < 0.1 (Fig. 2, Supplementary Material, Fig. S0a), which was highly significant (P < 10e–16, χ² = 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).

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 (ORC) 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 ORC) of SNP associations in a category. Of course, independently from any true case–control signal, an ORC 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,
we simply counted the number \( N_{AC} \) 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 \( N_{AC} \) 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 \( N_{AC} \) 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 (\( N_{AC} \)) 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 \( N_{AC} \) [defined by Eq. (3)], we next addressed the question whether random categories show any combined association. Because the distribution of the \( N_{AC} \) statistic also depends on category size and \( P_0 \) (Supplementary Material, Fig. S2a), \( N_{AC} \) was analogously normalized as \( OR_C \) [Eq. (4)]. Interestingly, after normalization, the \( nN_{AC} \) 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 \( nN_{AC} \) score of random categories is indeed due to the case–control signal in our data, we calculated the \( nN_{AC} \) score for each category for each of 1000 permuted data sets (where the case–control signal is removed). The mean of these \( nN_{AC} \) scores of random categories (averaged from the 1000 \( nN_{AC} \) scores from the permuted data sets) is very close to its expected value of zero (Supplementary Material, Fig. S2c). Thus, the above \( nN_{AC} \) 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 ORc, 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 ORc statistic (Supplementary Material, Fig. S3a). However, after normalization, the nORc scores of these categories were still greater than zero (Supplementary Material, Fig. S3b). This result is somewhat surprising, because nORc is designed to correct for the influence of SNP density. That nORc 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 nORc scores very close to zero (Supplementary Material, Fig. S3c). Thus, the increased nORc 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 nNAC 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 PWTCCC 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 (PWTCCC < 0.1 and P0 < 0.1), the enrichment signal is weak (nORc = 1.39) and only shows a non-significant trend (P = 0.09). However, when increasing the stringency of either PWTCCC or P0, this enrichment becomes more prominent (Supplementary Material, Fig. S4). Accordingly, the enrichment is strongest (nORc = 5.81, P < 0.001), for the most stringent SNP association thresholds (PWTCCC < 0.0001 and P0 < 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 13).

In addition to looking at the enrichment of positional WTCCC candidate loci by means of ORc, we also looked at their combined association as tested by NAc. 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 nNAC score was strongly increased under weak SNP association thresholds (P0 < 0.1 and PWTCCC < 0.1), whereas it was moderately increased for the more stringent choice of P0 and PWTCCC. However, only under more stringent SNP association thresholds, the magnitude of these combined associations (as measured by nNAC) is visibly larger for autoimmune than non-immune categories (Fig. 3), which in turn produces an enrichment of associated loci (as measured by nORc).

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 P0 < 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 P0 < 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.
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 (\(P < 0.1\)), the enrichment of associated loci at mouse immune loci is borderline significant (\(nORC = 1.75, P = 0.04\)). For the more stringent threshold \(P < 0.001\), this remained largely unchanged (\(nORC = 1.59, P = 0.04; \text{Table 1}\)). 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 \(P = 0.001\), 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 1. Results of the category-based analysis of immune system candidate loci

<table>
<thead>
<tr>
<th>Locus category</th>
<th>NARAC SNP association threshold (P &lt; 10^{-03})</th>
<th>Number of associated loci in category</th>
<th>Number of non-associated loci in category</th>
<th>Number of associated loci not in category</th>
<th>Number of non-associated loci not in category</th>
<th>Normalized score for enrichment ((nORC))</th>
<th>(P)-value for enrichment</th>
<th>Normalized score for combined association ((nNAC))</th>
<th>(P)-value for combined association</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTCCC RA ((P &lt; 10^{-03}))</td>
<td>0.1</td>
<td>315</td>
<td>3992</td>
<td>14</td>
<td>229</td>
<td>1.96</td>
<td>0.025</td>
<td>10.58</td>
<td>≤ 0.0001</td>
</tr>
<tr>
<td>Mouse Immune System Phenotype</td>
<td>0.001</td>
<td>261</td>
<td>232</td>
<td>14</td>
<td>102</td>
<td>1.59</td>
<td>0.001</td>
<td>6.39</td>
<td>≤ 0.0001</td>
</tr>
</tbody>
</table>

LD blocks are categorized as positional candidate loci, if they harbored SNP associations under the threshold \(P_{\text{WTCCC}} < 0.1\) (dark grey bars) or \(P_{\text{WTCCC}} < 0.001\) (light grey bars) with any of the seven phenotypes from the WTCCC GWAS. Associated loci were called based on SNP associations with \(P < 0.001\) 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 \(P_{\text{WTCCC}} < 0.001\) (light grey), but not for \(P_{\text{WTCCC}} < 0.1\) (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 \((P < 0.1)\), the enrichment of associated loci at mouse immune loci is borderline significant (\(nORC = 1.75, P = 0.04\)). For the more stringent threshold \(P < 0.001\), this remained largely unchanged (\(nORC = 1.59, P = 0.04; \text{Table 1}\)). 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 \(P < 0.001\), 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 \(CD180\).
control differences are spread over many categories, the NAC association with RA, as indicated by nNAC scores greater for the number of GO categories tested, hardly any GO cat-
further perform an (albeit conservative) Bonferroni correction complementary Material, Table S4b). However, the magnitude of GO categories attained the largest enrichment scores (Sup-
loci. Accordingly, after excluding the MHC region, similar
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 nORC scores, most of the leading categories are related to immune system functions (Table 3, Supplementary Material, Table S4a). Further GO categories with an increased nORC score include ‘cell surface receptor-linked signal transduction’ and ‘transcrip-
tion’. 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 (Sup-
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 cat-
ogy would meet formal significance thresholds.

Notably, the majority of GO categories show a combined association with RA, as indicated by nNAC 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 NAC statistic (Supplementary Material, Fig. S5b) is clearly different from that for the ORC statistic (Supplementary Material, Fig. S5a) over the 1252 included GO categories. Most GO categories attain a signific-
ance increased NAC statistic, despite the fact that only few have a significantly increased ORC statistic. When case–
control differences are spread over many categories, the NAC statistic of a category can be highly significant whereas ORC is not significant, because ORC compares locus association signals with the background of other loci, whereas NAC tests a category independent from any background signal. This leads to the situation, where most GO categories display signific-
antly 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

<table>
<thead>
<tr>
<th>NARAC SNP</th>
<th>log10(NARAC-Pval)</th>
<th>Gene(s) mapped to locus</th>
<th>WTCCC phenotype</th>
<th>WTCCC SNP</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs2476601</td>
<td>16</td>
<td>MAGI3, PHTF1, PTPN22, RSBN1, BCL2L15, AP4B1, DCLKE1B</td>
<td>RA, CD, T1D</td>
<td>rs10858002</td>
</tr>
<tr>
<td>rs6748358</td>
<td>5.2</td>
<td>CTLA4, ICOS</td>
<td>T1D</td>
<td>rs231790</td>
</tr>
<tr>
<td>rs2327832</td>
<td>4.8</td>
<td>OLG13, TNFAIP3</td>
<td>RA</td>
<td>rs2327832</td>
</tr>
<tr>
<td>rs1903942</td>
<td>3.2</td>
<td>MBL2</td>
<td>CD</td>
<td>rs1903929</td>
</tr>
<tr>
<td>rs4976139</td>
<td>3.8</td>
<td>PIK3R1</td>
<td>CD</td>
<td>rs12234107</td>
</tr>
<tr>
<td>rs2089510</td>
<td>3.4</td>
<td>CORIN, NFNL1, CNGA1, NIPAL1, TXK, TEC</td>
<td>RA</td>
<td>rs769010</td>
</tr>
<tr>
<td>rs3890745</td>
<td>3.7</td>
<td>PLCH2, PANK4, HES3, TNFRSF14, Clorf93, MMEL1</td>
<td>RA</td>
<td>rs10910097</td>
</tr>
</tbody>
</table>

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 nORC scores, most of the leading categories are related to immune system functions (Table 3, Supplementary Material, Table S4a). Further GO categories with an increased nORC score include ‘cell surface receptor-linked signal transduction’ and ‘transcrip-
tion’. 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 (Sup-

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)

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 random-

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 schiz-
ophrenia (3). Under this explanation, one would most likely expect that weak SNP associations are concentrated in plausible candidate categories. A second, not mutually exclusive, explana-

The effect of these disease variants may be evolutionarily more recent, they may often exist on longer haplotypes (25,26). This may
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 distances 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 unrecognised 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.geneontology.org’ 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 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 \( p_l = \min(p_{ij}) \), 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 \( \times \) 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 \( OR_C \) to measure the enrichment of phenotype-associated loci in category \( C \) as:

\[
OR_C = \frac{\sum_{i=1}^{N_l} I(l \in C \land p_l < P_0) / \sum_{i=1}^{N_l} I(l \in C \land p_l \geq P_0)}{\sum_{i=1}^{N_l} I(l \not\in C \land p_l < P_0) / \sum_{i=1}^{N_l} I(l \not\in C \land p_l \geq P_0)}
\]

(1)

Note that (i) \( OR_C \) implicitly depends on the threshold value \( P_0 \); (ii) if associated loci are randomly distributed over categories, one may expect \( \log(OR_C) \) to be normally distributed and with mean zero; (iii) the distribution of \( OR_C \) under the null hypothesis of no enrichment signal due to true-positive SNP associations will be influenced by factors like SNP density; (iv) \( OR_C \) 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 \( \min(p_{ij}) \) 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 \( OR_C \). To determine the significance of an observed \( OR_C \) statistic, we therefore estimated its empirical \( P \)-value as the fraction of case–control permutations, for which the \( OR_C \) 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 \( nOR_C \) based on the simulated null distribution of \( OR_C \) from \( K \) permutations of the affection status:

\[
nOR_C = \frac{\log(OR_{C,obs}) - \text{mean}(\log(OR_{C,permuted}), K)}{\text{SD}(\log(OR_{C,permuted}), K)}
\]

(2)

Note (i) the purpose of this transformation is to construct a standard normal distribution and (ii) the value of \( nOR_C \) 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 \( OR_C \). The division by the standard deviation additionally corrects for factors that influence the variance of \( OR_C \) (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 \( NAC \) of associated loci in \( C \):

\[
NAC = \sum_{i=1}^{N_l} I(l \in C \land p_l < P_0)
\]

(3)

The number of associated loci \( NAC \) 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 \( nNAC \) as follows:

\[
nNAC = \frac{\log(NAC_{C,obs}) - \text{mean}(\log(NAC_{C,permuted}), K)}{\text{SD}(\log(NAC_{C,permuted}), K)}
\]

(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 were 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 ($\sim 33\%$ 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.

**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 N01-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 IN84042.

**REFERENCES**


