Published in final edited form as:
Hum Genet. 2009 August ; 126(2): 289–301. doi:10.1007/s00439-009-0676-z.

Strategies and issues in the detection of pathway enrichment in genome-wide association studies

Mun-Gwan Hong[1], **Yudi Pawitan**[1], **Patrik K.E. Magnusson**[1], and **Jonathan A. Prince**[1]

[1] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden

Abstract

A fundamental question in human genetics is the degree to which the polygenic character of complex traits derives from polymorphism in genes with similar or with dissimilar functions. The many genome-wide association studies now being performed offer an opportunity to investigate this, and although early attempts are emerging, new tools and modeling strategies still need to be developed and deployed. Towards this goal we implemented a new algorithm to facilitate the transition from genetic marker lists (principally those generated by PLINK) to pathway analyses of representational gene sets in either threshold or threshold-free downstream applications (e.g. DAVID, GSEA-P, and Ingenuity Pathway Analysis). This was applied to several large genome-wide association studies covering diverse human traits that included type 2 diabetes, Crohn's disease, and plasma lipid levels. Validation of this approach was obtained for plasma HDL levels, where functional categories related to lipid metabolism emerged as the most significant in two independent studies. From analyses of these samples we highlight and address numerous issues related to this strategy, including appropriate gene based correction statistics, the utility of imputed vs. non imputed marker sets, and the apparent enrichment of pathways due solely to the positional clustering of functionally related genes. The latter in particular emphasizes the importance of studies that directly tie genetic variation to functional characteristics of specific genes. The software freely provided that we have called ProxyGeneLD may resolve an important bottleneck in pathway-based analyses of genome-wide association data. This has allowed us to identify at least one replicable case of pathway enrichment but also to highlight functional gene clustering as a potentially serious problem that may lead to spurious pathway findings if not corrected for.

Keywords

pathway; genome-wide; association; gene; enrichment; ontology

Introduction

The extent to which common genetic polymorphism contributes to variance in complex human traits has been explored for decades but it has only recently become possible to perform hypothesis-free studies at fine scale on a genome-wide level (Klein et al. 2005). These studies strive to reveal the underlying genetic architecture of complex human diseases and quantitative traits and although the genetic effect sizes have typically been small, the statistical evidence

Correspondence: Dr. Jonathan A. Prince, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Nobels väg 12A, 171 77 Stockholm, Sweden, Phone: +46 (0)8 524 86008, Fax: +46 (0)8 31 49 75, Jonathan.Prince@ki.se.

Conflict of Interest Statement

The authors declare no conflict of interest.

implicating individual genes has been strong (Barrett et al. 2008; Willer et al. 2008; Zeggini et al. 2008). However, most genome-wide association studies performed to date depict only a few significant loci (http://www.genome.gov/gwastudies/). Among the many remaining questions is whether or not data from these studies can also be used to generate additional insight into the biological pathways that influence disease. If the suggestion that complex trait genetics follows an L-shaped distribution of effect sizes is true (Dixon et al. 2007; Sing and Boerwinkle 1987), then it might be possible to explore the rightmost tail of the distribution for significant genes that share common functions. This represents a natural extension of the approach used in gene expression profiling, where pathway analyses based upon controlled gene descriptions, such as from the Kyoto Encyclopedia of Genes and Genomes or the Gene Ontology projects, is common practice (Ashburner et al. 2000; Goto et al. 1997). As reflected by the recentness of genome-wide association studies themselves, attempts to consider pathways are also relatively new. The first study to explore for pathway enrichment from genome-wide marker data was based upon relatively small samples on Parkinson's disease and age-related macular degeneration (Wang et al. 2007). This was followed up by a second paper based upon several human diseases from the larger public Wellcome-Trust Case-Control Consortium data (WTCCC) (Torkamani et al. 2008). The use of pathway modeling has also been an additional analytical component to some of the primary genome-wide association studies, the first of which being a study of human height by Gudbjartsson et al. (Gudbjartsson et al. 2008). There are now several recent additional examples (Askland et al. 2009; Baranzini et al. 2009; Vink et al. 2009; Wang et al. 2009), and with the exception of one study on type 2 diabetes that reports negative findings (Perry et al. 2009), all previous studies have claimed significant evidence of pathway enrichment.

A principal bottleneck in these kinds of analyses, in contrast to gene expression profiling, is that genome-wide association studies produce genetic marker lists, not gene lists, and it is the latter for which pathway annotations exist. The reliable conversion of markers to representative genes is not trivial and shares the same inherent ambiguity as any genetic association study where a genomic region is implicated in disease or quantitative trait variation and where linkage disequilibrium (LD) remains a vital consideration. We thus set out to create an application that would allow the automatic conversion of any marker list in flat text file format to a representative gene list, also taking into account LD. Our principal target was output from PLINK (Purcell et al. 2007) which is at present the most commonly used software for performing the statistical testing of markers from genome-wide association studies. In the present study, we have applied our software to several large genome-wide association data sets and used the results as a foundation to examine a number of issues that may have relevance for the success of this strategy. We obtained relatively strong evidence validating the importance of lipid related pathways in the determination of plasma lipid levels, but we also highlight a key result that positional gene clustering, if not corrected for, can lead to spurious results.

Results

There were three core issues that we considered essential to investigate prior to attempting pathway enrichment analyses on disease or trait results from genome-wide association studies i) the extent of bias introduced by the higher likelihood of low p-values in long genes (or genes with many tested polymorphisms) ii) the potential local clustering of genes with similar function and iii) differences in marker lists, in particular as it relates to imputed vs. non-imputed data sets (Marchini et al. 2007).

For the first question, the impact of gene length was assessed using the gene list derived from an HDL data set (table 1; sample 1), in order to provide an indication of what might be expected from genome-wide marker data rather than all annotated genes in the human genome. We note that this sample consists of approximately 2.3 million directly typed and imputed markers, and thus gene representation reflects what is obtainable from HapMap data (release 21). After conversion using our algorithm, this set also resulted in the largest number of genes (16,977) from among the samples listed in table 1. These were sorted by average transcript length of known splice-forms (see methods) and the top 1%, 2%, and 3% of genes were tested by enrichment analysis using DAVID with the total (16,977) as the base set (see methods). All three divisions showed a large enrichment of specific ontology terms with progressively increasing significance from the 1% to 3% groupings (for clarity we only show the 3% grouping and the top 5 terms in table 2). Sorting this gene set by the number of SNPs extending across the longest splice-form of each gene resulted in a similar result (table 2). Finally, assessing genes according to their unadjusted p-value for the HDL trait association itself also gave rise to similar term enrichment, although significance was attenuated (table 2). Simple linear regression of gene length, number of markers, and unadjusted marker significance was as expected high ($r^2 = 0.8$ for length vs. marker number, and $r^2 = 0.15$ for both comparisons of unadjusted p-value vs. length and marker number, $p \ll 0.0001$ for all three comparisons). These results illustrate the importance of applying strict corrections for gene length (or number of markers), and for this reason we elected to base our correction upon an LD-adjusted number of markers for each gene (see methods). A related topic is the change in the rank of associated genes following the weighted adjustment by our algorithm. This is shown in figure 1, where the ranks of the top 3% of adjusted genes for the HDL trait are plotted against their ranks prior to adjustment (there we also depict a running average of the rank ratios).

The second question pertaining to the local clustering of genes with similar functions was in our view a potentially complicated problem. This has been explored to some extent previously (e.g. Lee and Sonnhammer 2003; Yi et al. 2007), but we thought it might be valuable to generate a depiction of the degree of clustering in the genome based upon gene ontology annotations. For this we again using the gene list derived from the HDL data set (table 1; sample 1) and a systematic walk was performed along each chromosome from pter to qter, taking 100 genes at a time (this resulted in an average spacing of 17 megabases and a total of 163 bins) and those lists tested for enrichment using DAVID. Gene lists were expanded to 100<N<200 at the qter tails. We recognized that the true problem resides in the capturing of multiple genes by virtue of linkage disequilibrium (LD) with a single association signal, but that LD which seldom extends beyond a few hundred kilobases (Weiss and Clark 2002) would result in gene lists that are too small to empower an enrichment analysis. An equivalent analysis was performed using 100 random bins of 100 genes each to generate a null distribution for comparison. In each case, the maximum statistic and associated GO term were recorded and we illustrate the results in figure 2 (full data ordered by chromosome and segment interval as well as specifically enriched terms are presented in supplementary table 1). This revealed that clustering is extensive, with a large proportion of bins showing strong evidence of pathway enrichment. We scrutinized the 10 most significant bins to assess the size of the chromosomal segments and the LD structure across the genes that contribute to the signals. In general, the clusters giving rise to significance spanned single contiguous regions much smaller than the total region examined (not shown). However, there were exceptions to this. For example, olfactory genes located on chromosome 11 occur in multiple, separate clusters.

To explore how imputation (Marchini et al. 2007) might impact pathway analyses, we examined the WTCCC diabetes data set (table 1; subset of sample 2), for which both directly genotyped and imputed data were available. We anticipated that an increase in gene representation would occur with imputed data, but sought to estimate its degree. After marker conversion our algorithm resulted in 15,510 genes being represented in the directly typed set (from 393,143 markers) vs. 16,347 in the imputed set (from 2,308,536 markers). We noted that the average length of the represented genes from the imputed data were much shorter than those in the non-imputed list (8,995 bases for the 837 additional genes vs. 63,591 bases for the non-imputed set, $p \ll 0.0001$). The position of the imputed genes by unadjusted p-value rank for the disease association in the full list was also assessed to determine if these were randomly distributed. Average rank for the imputed genes was 11,812 vs. 7,977 for the non-imputed set, $p \ll 0.0001$. We consider the more important result here to be the inclusion by imputation of short genes, as this last statistic related to rank simply reflects the smaller likelihood of obtaining low p-values for short genes. A second question is the relative change in the ranks of genes following imputation. To estimate this, the ranks of the top 3% of genes in the directly typed set were plotted vs. their ranks following imputation and the same was done for the top 3% of imputed genes vs. the typed set. This is shown in figure 3, where the ranks of the majority of genes remain the same, but where clear outliers exist. Of particular note, while the increase following imputation for some genes was expected (figure 3a), the loss of rank (figure 3b) derives from our algorithm using LD to predict significant un-typed markers in adjacent genes, but where imputation ultimately provides evidence against those particular genes being significant. One additional analysis was thus conducted, changing the LD threshold to 0.9 from its default of 0.8. This resulted in slightly fewer genes that decreased rank following imputation, but did not affect the number of genes that increased rank (not shown).

To initiate pathway analyses of real association results, the HDL set (table 1; sample 1) was selected with the premise that this might represent a validation of this approach and provide a benchmark for the degree of statistical evidence to be expected. Thus, the emerging genetic architecture of the HDL trait indicates that several of the most significant genes previously implicated on a genome-wide level play a role in lipid metabolism (e.g. Zeggini et al. 2008). The data set comprises a meta-analysis of 8656 individuals from 3 individual studies, with results representing the marker-by-marker test statistics for 2.3 million markers (directly typed and imputed). We performed a marker-to-gene conversion for this sample and analyzed the top 1% of genes for pathway enrichment (see methods). This threshold was decided upon primarily due to the concern that larger lists might increase the likelihood of enriching for functional gene clusters. From this analysis, we noted that there were at least two possible instances of exaggerated enrichment due to clustering. In the first, 3 genes contribute to statistical evidence of enrichment of several lipid related terms, these being *FADS1*, *FADS2*, and *FADS3*, all residing in the same genomic interval and LD block. Removing 2 of these from the input list, retaining only the gene with the highest rank, and re-evaluating the statistics for previously significant terms resulted in an anticipated decrease in the statistical evidence. In contrast, we also noted two genes on chromosome 16q, in relatively tight proximity, that both contribute the enrichment of the term “cellular lipid metabolic process”, these being *LCAT* and *LYPLA3*. The most significant markers in these genes were not in high LD, and thus both were retained. Based on these observations we concluded that while it was trivial and informative to manually curate the results of a pathway analysis by examining the specific genes that contribute to enrichment and to adjust the number downward based upon LD and retaining the gene(s) with maximum significance, that an automated process might be preferable in some cases. For this reason, we added a feature to our software to automatically generate “trimmed” lists to account for genes in high LD (see methods). This feature also allowed us to provide an estimate of the average level of excess gene representation across the entire data set. For this sample that number was 0.53; in other words, each gene had on average 0.53 additional genes that were indiscernible given a certain LD threshold (in this case $r^2 = 0.8$). The results of the manually trimmed analysis of the HDL trait are presented in table 3a. There, the genes that contribute to the enrichment of the term “cellular lipid metabolic process” are specifically highlighted since they served to illustrate the rank assignments from our algorithm of some of the most well-known genes that influence HDL levels (table 3b). We extended this analysis to both LDL and triglyceride levels. The results for LDL are shown in table 4, where lipid related terms also emerged at the top. In contrast, for TG, there were no significant terms, even at a liberal uncorrected p-value threshold of 0.05.

We next applied the above analytical scheme (first converting PLINK results to gene lists and performing enrichment tests using DAVID) to two additional genome-wide data sets, selected from the growing list of studies (http://www.genome.gov/gwastudies/) due to their size (table 1; samples 2 and 3). The most notable finding from these sets was the possible enrichment of microtubule genes in type 2 diabetes (table 5a). An interesting feature of this result was the marked difference from what was observed for lipids in that the ranks of the component genes that contribute to term significance are all much lower, but this set also had the benefit of these genes each residing in unique loci (table 5b). In contrast, the analysis of Crohn's disease was more problematic with the apparent enrichment of several terms related to immune response appearing to arise due to clustering of genes in the well-characterized HLA region on chromosome 6p. We elected for this sample to present both the pre-trimmed and trimmed results for this data set since it provides an illustration of the impact of clustering, and where the requisite trimming changes the significance substantially (table 6). Notably, even after trimming some possible clustering still remains involving 3 genes, although they were not formally in strong LD (*MICA*, *HLA-DQA2*, and *HLA-C*).

While DAVID represents an excellent starting point for pathway analyses given its flexibility and fairly comprehensive coverage of recent annotations, there are two important alternatives. For the first, in order to complement the above analyses we chose to focus again on plasma lipid traits, and a search for enrichment was attempted using gene set enrichment analysis (GSEA; see methods) which was originally designed to circumvent the arbitrary nature of threshold definitions (Subramanian et al. 2005; see methods). For HDL levels in particular, this produced only modest evidence implicating lipid metabolism (table 7). Performing GSEA based analyses on the LDL trait in this same sample failed to identify significant terms (not shown). For TG, a single term was significant at $p = 0.0002$ (heparin binding). However, a problem emerged when we simply removed the top ranked gene (CETP) from analysis of HDL, which almost completely eliminated the significance for lipid terms (table 7). We think this reiterates the importance of more closely examining the specific genes that contribute to term enrichment statistics. Perhaps more importantly, this highlights a potential problem with GSEA in that strong statistical support for a pathway can be driven by a single gene. The second software that is achieving wide-spread use now is IPA (Ingenuity Pathway Analysis). We again subjected our lipid results for sample 1 (top 1%) for analysis using IPA specifically restricting our assessment to the category of molecular and cellular functions. For this analysis we chose to focus on pathways with 3 or more contributing genes after trimming of any identified positional clusters. The best evidence of pathway enrichment was obtained for "homeostasis of cholesterol" in relation to plasma LDL levels with an uncorrected p-value of 9.3×10^{-6} (note that IPA bases its analyses on a Fisher's Exact Test which is comparable to DAVID) and included the genes *ABCA1*, *APOB*, *APOE*, *LDLR*, and *PLSCR3*. The next best evidence was for "metabolic process of lipid" in relation to HDL levels with an uncorrected p-value of 1.3×10^{-4} and included the genes *CETP, FADS1, IKBKB, IL6, LCAT, LPL, PLA2G15, PPARA, RAC1*, and *SMDPD2*. These two lists can be directly compared to tables 3b and 4b, where several genes are represented in both analyses, but where there are differences. This highlights an important aspect of which down-stream analysis software one opts to use, as there clearly are gene annotation discrepancies. There was no evidence for enrichment of any category with more than 3 genes at $p < 0.05$ for TG.

Finally, we sought evidence of replication for the above analyses by focusing on a second independent sample with plasma lipid traits from a recent genome-wide association study (table 1; sample 4). We decided to focus on IPA for this analysis, given its somewhat better apparent performance compared to the other approaches above. We were unable to validate the result for LDL levels, with no functional categories with 3 or more genes achieving $p < 0.05$. However, for HDL levels, results were highly consistent with the data from sample 1, with multiple categories replicating between the two sets and with "lipid metabolism" being the most significant high-level category in both samples. The best evidence of replication was obtained for the lower-level term “homeostasis of cholesterol” with $p = 3.3 \times 10^{-4}$ for sample 1 and $p = 5.3 \times 10^{-4}$ for sample 4 (using Fisher’s combined probability test gave $p = 2.3 \times 10^{-6}$). This represented the best study-wide significance across all tests conducted.

Discussion

We developed and implemented a new software program to facilitate the conversion of genetic marker lists to gene lists, with the primary goal of uncovering evidence of pathway enrichment from genome-wide association studies. To accomplish this, only the largest of the many available genome-wide association study data sets were targeted, with an initial focus on plasma lipid levels that we thought might have a detectable pathway basis. We obtained a reasonably high level of statistical evidence by replicating an enrichment of lipid related terms for plasma HDL levels in two independent samples. We consider however the more important aspect of this work to be the broader exploration of various parameters that may affect the validity of this strategy.

A central issue in approaches such as ours is how to best obtain representation of genes from markers. Any strategy like that presented will both increase gene diversity in a region of a single association signal, potentially causing pathway dilution, and create the possible “appearance” of enrichment due to the positional clustering of functionally related genes. For the former, we noted that across the various data sets, the inflation factor was around 0.5, or on average 50% more genes than would be expected if LD was $r^2 \geq 0.8$ between the markers showing the highest significance. The consequence of this will be to increase type II error. The latter problem however, in our view, is much more serious and a few cases are exemplified where the statistical support decreases following correction for clustering. Although we attempt to resolve clustering with both manual and/or automatic curation, the only real solution will require the identification of true functional variants that can be shown to influence specific genes. Thus, functional studies on gene regulation, including identifying markers that directly affect gene expression and can be tied to a specific gene are extremely important for pathway-based analyses. There are a number of impressive *in silico* tools now appearing for marker and gene prioritization that may also aid this endeavor (Chen et al. 2008; Gaulton et al. 2007; Ge et al. 2008; Pico et al. 2009; Tranchevent et al. 2008). The many published genome-wide data sets also lend themselves to strategies to test for marker independence (Sun et al. 2008), but there will also be merit in exploring alternative human populations, where LD is less of a problem (Cox et al. 2002). For the time being, we acknowledge that our algorithm leads to the inclusion of an excess of genes around significant signals, but we do note that the software includes options for changing LD thresholds. Nonetheless, in terms of pathway analyses it may be argued that results of enriched terms where the contributing genes all reside in independent loci may be regarded as more reliable than those where the genes are clustered, even if not in strict LD.

We think there is reason to be cautious about biological interpretations of pathway results across the data sets here and elsewhere. The space of variables that can be tuned for these kinds of analyses is quite large, ranging from the statistics and covariates chosen in the original genome-wide studies to which pathway enrichment tool/statistic is used. This is particularly important given the relatively large number of analysis programs now available, including those used here (DAVID, GSEA-P, and IPA). Against this background the possible enrichment of terms associated with lipid metabolism in two independent data sets for HDL levels represents a reasonable validation of this approach. However, this statistical evidence should be taken in context with that of the most significant individual gene (CETP; $P \approx 10^{-20}$). If this is an indication that the evidence one can expect from pathway analyses is likely to be weaker than for individual genes, then even larger samples than those currently employed may be required for detection. An interesting contrast to the HDL result is the essential lack of evidence in the Crohn's disease sample. In particular, the role of the immune system for this disease is well documented (Lettre and Rioux 2008) yet is only marginally implicated by our pathway analysis. The result however reiterates a classic problem in genetics, namely the possibility that there are multiple independent functional variants acting in the HLA region that are indiscernible due to LD. This may have implications for other immune diseases, such as rheumatoid arthritis (Raychaudhuri et al. 2008). The result for diabetes yielded yet a third contrasting scenario, where there was a suggestion of the involvement of microtubule genes but no confounding by positional clustering. An intriguing aspect of this result was the inclusion of the *KIF11* gene which resides in an LD block that includes *HHEX* and *IDE*. We have focused on *IDE* in the past (Gu et al. 2004), and others have highlighted maximum signals nearer *HHEX* (Zeggini et al. 2007). Still, though there is no strong precedent for the involvement of kinesin proteins in type 2 diabetes from the literature, the additional information included by a pathway analysis may help to prioritize specific genes for further functional assessment.

At the outset of this investigation we were interested in what the evolutionary implications might be should evidence of pathway enrichment emerge. Given that both allelic and locus heterogeneity act to influence phenotypes, a natural concern was that population differences could contribute to apparent enrichment. Thus, two genes with similar annotations and acting detectably on a phenotype but exclusively in separate sub-populations would be considered together. For genes such as *CETP* this is unlikely to be the case since it has been demonstrated to be highly significant in numerous populations, but for genes farther down in rank it may be. This ultimately represents a trade-off in terms of obtaining high statistical power for single marker analyses (via meta-analysis) and the risk of enriching a pathway due to locus heterogeneity across different populations. Nonetheless, the assumption that diseases should arise by mechanisms involving similar genes has long been one the corner-stones of genetic association studies. Complex phenotypes have evolved to be resilient to insult (Buchanan et al. 2006), and the emergence of modest to weak effects across multiple genes in a pathway might be seen to support this.

In summary, pathway analyses of genome-wide association data based upon controlled gene annotations are now emerging, with many groups claiming evidence of enrichment for a range of phenotypes (Baranzini et al. 2009; Torkamani et al. 2008; Wang et al. 2007). The data presented in this paper also support the prospect that biological pathways can be detected in genome-wide association data, but we have still only scratched the surface of the bulk of genome wide data becoming available and we think there is still reason to be cautious. In particular, the observation that persists both here and in other studies is that there is no case where the evidence for any pathway enrichment exceeds that of the previously reported single locus findings. This can be taken as a strong contrast to gene expression studies, where striking evidence of pathway enrichment often emerged in the absence of obvious single gene effects (Mootha et al. 2003). The application of pathway approaches to genetic marker data is nonetheless relatively new territory and may represent a valuable addition to ongoing genome-wide studies given the vast range of phenotypes now being investigated (http://www.genome.gov/gwastudies/). This needs to be tempered however against a number of issues that have previously not been dealt with, the most important in our view being positional gene clustering. The software provided here should nonetheless help to relieve the bottleneck of automating the transition from marker lists to gene lists, and thus expedite further analyses of genome-wide data in a pathway context.

Materials and Methods

From SNP lists to Gene lists

A Perl program, named ProxyGeneLD, was developed to automate the assignment of genotyped SNPs from genome-wide association studies to specific genes, flexibly taking into consideration linkage disequilibrium (LD). This was based upon several assumptions. First, a functional SNP is more likely to affect the gene if the variant is located within its open reading frame or in the promoter binding region. Second, the number of SNPs derived from HapMap phase II is large enough to provide coverage of most genes annotated in the various databases describing pathway terms (e.g. Gene Ontology). Third, the program also assumes that the number of common non-HapMap SNPs which are solely located in a gene and in strong LD with a directly observed SNP is negligible.

The program requires 5 different data files from public databases, which were obtained as follows. The positional data of the entire set of validated polymorphic SNPs from CEU samples of HapMap phase II (release 22) and the file containing NCBI RefSeq transcript locations were attained using the UCSC table browser tool (hg18). Information on pair-wise LD estimates for all HapMap CEU SNPs was downloaded from the official HapMap homepage (release 23a). In the program, gene definitions followed Entrez Gene database conventions and GeneIDs were used as reference identification numbers. Because GSEA-P, one of the downstream applications we utilized, accepts only gene symbols, a step to convert Entrez GeneIDs to gene symbols was also implemented in our program. The data files for GeneIDs and linked official gene symbols were downloaded from Entrez Gene ftp. All of the aforementioned 5 data files were preprocessed creating more compact files in order to speed up the program. The program initiates by reading the output data file of original genome-wide association study (GWAS) that is tab-, space-, or comma-delimited text files and which contains list of SNPs and corresponding p-values.

Two sets of SNPs are considered in the program. The first is the set of SNPs from HapMap CEU phaseII and the other comprises the SNPs tested in original GWAS, which were termed “HapMap SNPs” and “study SNPs”, respectively. The program first examines LD structure between HapMap SNPs and groups two HapMap SNPs at a user defined LD threshold to create a “proxy cluster”. In most cases the condition $r^2 \geq 0.8$ was used as a default. The proxy cluster is then expanded by iteratively adding one more HapMap SNP in high LD with any member SNP of the cluster or by merging with another proxy cluster that shares an identical member. The program retains information on which HapMap SNPs were in high LD to form proxy clusters. Each HapMap SNP regardless of membership in a proxy cluster is then assigned to the longest known transcript of each gene including a 1kbp extension to include promoter regions. Likewise, every proxy cluster is assigned to the nearest gene according to the positions of the SNPs that comprise the cluster. All genes for which any study SNP or any proxy cluster is assigned are then listed, ignoring SNPs which are not represented in HapMap. The significance level of association for each gene before adjustment is chosen as the lowest p-value among the single study SNPs that do not belong to any proxy cluster and those in proxy clusters that include one or more study SNP that have been assigned to the gene. The total number of single SNPs and proxy clusters is then multiplied by the pre-adjustment significance level. An illustration of this process is shown in supplementary figure 1.

Automated Trimming of Gene Clusters

A gene cluster was defined as a gene set for which the most significant SNP assigned to each gene is a member of the same proxy cluster. In other words, signals from all genes in a gene cluster had a high likelihood of originating from a single association between one marker and phenotype. The program automatically provides a column showing for each gene any other gene(s) that were indiscernible members of the same cluster. As an optional step, it additionally creates one more list of genes in which clusters are sorted out, whereby the gene with the highest ranked SNP is retained and the remaining genes moved to bottom of the list. A note of caution with this is that it might not be appropriate to use this function for cases in which rank is important (e.g. for GSEA-P).

GSEA

Gene set enrichment analysis (GSEA v2.0) is implemented as a JAVA application named GSEA-P and requires a text file containing gene symbols and a weight for gene ranking. For our analyses, gene ranks were calculated by; (weight) = $-\log_{10}$(adjusted gene-wide p-value). Since this produces a large number of negative values due to gene-based adjustment, to retain the relative rank positions for all genes beyond the inflection a linear scale ending at zero was created. All reported results using GSEA are based upon 5000 permutations, p = 1 weighting, maximum set size of 500 and minimum set size of 15. A false discovery rate < 25% was used as an indication of potentially interesting findings.

Threshold-based ontology analysis (DAVID)

DAVID is one of the many publicly available web-applications for searching for over-represented pathways in gene lists. It provides extensive options for the interrogation of approximately 40 alternative databases (e.g. KEGG, Interpro, Pfam), but there is a large degree of redundancy when multiple data sources are tested and GO has the greatest number of gene annotation records. Even for GO annotations, redundant terms arise with only marginal differences in the component gene lists. For this reason, for all analyses using DAVID results are presented based upon the use of functional annotation clustering and only the top five single terms from among clusters are reported. For presentation purposes, this was truncated at a maximum of five results. Analyses presented here used DAVID with a March 2008 GO annotation update.

Ingenuity pathway analysis (IPA; Ingenuity Systems)

IPA is a commercial web-delivered application implemented in JAVA of which one of functions is to calculate the probability of observing an association between certain gene sets and pathways by random chance. It applies the right-tailed Fisher's Exact test and Benjamini-Hochberg method of multiple testing correction (Benjamini and Hochberg 1995) to find over-represented pathways in the gene set comparing to a reference set, which in our analyses is the total genes monitored in each original GWAS. In IPA, annotations of genes to pathways are based on the Ingenuity Knowledge Base, that was built on the extracted findings in major life sciences literature and data in established public databases such as GO and EntrezGene.

General statistics

For various analyses, gene length was taken as the average of all known splice-forms according to RefSeq of NCBI build 36.1. Total unadjusted SNP number for each gene was established according to the genomic interval spanning the longest known splice-form. Analyses involving gene ranks were performed using the Mann-Whitney U test. Gene length, number of markers, and unadjusted genome-wide p-values were $\log_{10}$-transformed prior to simple linear regression.

Software Availability

The software ProxyGeneLD is available for free download at http://ki.se/ki/jsp/polopoly.jsp?d=26072&l=en. It consists of the main program proxyGeneLD.pl, and two additional programs for preprocessing.

URLs

DAVID 2008: http://david.abcc.ncifcrf.gov/; Entrez Gene download data: ftp://ftp.ncbi.nlm.nih.gov/gene/; Gene Ontology (GO): http://www.geneontology.org/; HapMap bulk data download, LD data: http://ftp.hapmap.org/ld_data/?N=D; UCSC genome browser: http://genome.ucsc.edu/; Ingenuity Pathway Analysis software: http://www.ingenuity.com/; GSEA-P: http://www.broad.mit.edu/GSEA

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are greatly indebted to the scientists responsible for making their genome-wide data accessible. We are grateful for financial support from The Swedish Medical Research Council (grant 2007–2722) and the National Institutes of Health (grant AG028555).

References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–9. [PubMed: 10802651]

Askland K, Read C, Moore J. Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission. Hum Genet 2009;125:63–79. [PubMed: 19052778]

Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BW, Janssens AC, Wilson JF, Spector T, Martin NG, Pedersen NL, Kyvik KO, Kaprio J, Hofman A, Freimer NB, Jarvelin MR, Gyllensten U, Campbell H, Rudan I, Johansson A, Marroni F, Hayward C, Vitart V, Jonasson I, Pattaro C, Wright A, Hastie N, Pichler I, Hicks AA, Falchi M, Willemsen G, Hottenga JJ, de Geus EJ, Montgomery GW, Whitfield J, Magnusson P, Saharinen J, Perola M, Silander K, Isaacs A, Sijbrands EJ, Uitterlinden AG, Witteman JC, Oostra BA, Elliott P, Ruokonen A, Sabatti C, Gieger C, Meitinger T, Kronenberg F, Doring A, Wichmann HE, Smit JH, McCarthy MI, van Duijn CM, Peltonen L. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat Genet 2009;41:47–55. [PubMed: 19060911]

Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, Barkhof F, Radue EW, Lindberg RL, Uitdehaag B, Johnson MR, Angelakopoulou A, Hall L, Richardson JC, Prinjha RK, Gass A, Geurts JJ, Kragt J, Sombekke M, Vrenken H, Qualley P, Lincoln RR, Gomez R, Caillier SJ, George MF, Mousavi H, Guerrero R, Okuda DT, Cree BA, Green A, Waubant E, Goodin DS, Pelletier D, Matthews PM, Hauser SL, Kappos L, Polman CH, Oksenberg JR. Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 2009;18:767–78. [PubMed: 19010793]

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, Heath S, Laukens D, Mni M, Rutgeerts P, Van Gossum A, Zelenika D, Franchimont D, Hugot JP, de Vos M, Vermeire S, Louis E, Cardon LR, Anderson CA, Drummond H, Nimmo E, Ahmad T, Prescott NJ, Onnie CM, Fisher SA, Marchini J, Ghori J, Bumpstead S, Gwilliam R, Tremelling M, Deloukas P, Mansfield J, Jewell D, Satsangi J, Mathew CG, Parkes M, Georges M, Daly MJ. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet 2008;40:955–62. [PubMed: 18587394]

Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) 1995;57:289–300.

Buchanan AV, Weiss KM, Fullerton SM. Dissecting complex disease: the quest for the Philosopher's Stone? Int J Epidemiol 2006;35:562–71. [PubMed: 16540539]

Chen YH, Liu CK, Chang SC, Lin YJ, Tsai MF, Chen YT, Yao A. GenoWatch: a disease gene mining browser for association study. Nucleic Acids Res 2008;36:W336–40. [PubMed: 18440974]

Cox R, Bouzekri N, Martin S, Southam L, Hugill A, Golamaully M, Cooper R, Adeyemo A, Soubrier F, Ward R, Lathrop GM, Matsuda F, Farrall M. Angiotensin-1-converting enzyme (ACE) plasma concentration is influenced by multiple ACE-linked quantitative trait nucleotides. Hum Mol Genet 2002;11:2969–77. [PubMed: 12393808]

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. A genome-wide association study of global gene expression. Nat Genet 2007;39:1202–7. [PubMed: 17873877]

Gaulton KJ, Mohlke KL, Vision TJ. A computational system to select candidate genes for complex human traits. Bioinformatics 2007;23:1132–40. [PubMed: 17237041]

Ge D, Zhang K, Need AC, Martin O, Fellay J, Urban TJ, Telenti A, Goldstein DB. WGAViewer: software for genomic annotation of whole genome association studies. Genome Res 2008;18:640–3. [PubMed: 18256235]

Goto S, Bono H, Ogata H, Fujibuchi W, Nishioka T, Sato K, Kanehisa M. Organizing and computing metabolic pathway data in terms of binary relations. Pac Symp Biocomput 1997:175–86. [PubMed: 9390290]

Gu HF, Efendic S, Nordman S, Ostenson CG, Brismar K, Brookes AJ, Prince JA. Quantitative trait loci near the insulin-degrading enzyme (IDE) gene contribute to variation in plasma insulin levels. Diabetes 2004;53:2137–42. [PubMed: 15277398]

Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, Sulem P, Thorlacius S, Gylfason A, Steinberg S, Helgadottir A, Ingason A, Steinthorsdottir V, Olafsdottir EJ, Olafsdottir GH, Jonsson T, Borch-Johnsen K, Hansen T, Andersen G, Jorgensen T, Pedersen O, Aben KK, Witjes JA, Swinkels DW, den Heijer M, Franke B, Verbeek AL, Becker DM, Yanek LR, Becker LC, Tryggvadottir L, Rafnar T, Gulcher J, Kiemeney LA, Kong A, Thorsteinsdottir U, Stefansson K. Many sequence variants affecting diversity of adult human height. Nat Genet 2008;40:609–15. [PubMed: 18391951]

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385–9. [PubMed: 15761122]

Lee JM, Sonnhammer EL. Genomic gene clustering analysis of pathways in eukaryotes. Genome Res 2003;13:875–82. [PubMed: 12695325]

Lettre G, Rioux JD. Autoimmune diseases: insights from genome-wide association studies. Hum Mol Genet 2008;17:R116–21. [PubMed: 18852199]

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–13. [PubMed: 17572673]

Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003;34:267–73. [PubMed: 12808457]

Perry JR, McCarthy MI, Hattersley AT, Zeggini E, Weedon MN, Frayling TM. Interrogating Type 2 Diabetes Genome-Wide Association Data Using a Biological Pathway-Based Approach. Diabetes. 2009

Pico AR, Smirnov IV, Chang JS, Yeh RF, Wiemels JL, Wiencke JK, Tihan T, Conklin BR, Wrensch M. SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. Nucleic Acids Res 2009;37:D803–809. [PubMed: 18984625]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75. [PubMed: 17701901]

Raychaudhuri S, Remmers EF, Lee AT, Hackett R, Guiducci C, Burtt NP, Gianniny L, Korman BD, Padyukov L, Kurreeman FA, Chang M, Catanese JJ, Ding B, Wong S, van der Helm-van Mil AH, Neale BM, Coblyn J, Cui J, Tak PP, Wolbink GJ, Crusius JB, van der Horst-Bruinsma IE, Criswell LA, Amos CI, Seldin MF, Kastner DL, Ardlie KG, Alfredsson L, Costenbader KH, Altshuler D, Huizinga TW, Shadick NA, Weinblatt ME, de Vries N, Worthington J, Seielstad M, Toes RE, Karlson EW, Begovich AB, Klareskog L, Gregersen PK, Daly MJ, Plenge RM. Common variants at CD40 and other loci confer risk of rheumatoid arthritis. Nat Genet 2008;40:1216–23. [PubMed: 18794853]

Sing CF, Boerwinkle EA. Genetic architecture of inter-individual variability in apolipoprotein, lipoprotein and lipid phenotypes. Ciba Found Symp 1987;130:99–127. [PubMed: 3327665]

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50. [PubMed: 16199517]

Sun J, Zheng SL, Wiklund F, Isaacs SD, Purcell LD, Gao Z, Hsu FC, Kim ST, Liu W, Zhu Y, Stattin P, Adami HO, Wiley KE, Dimitrov L, Li T, Turner AR, Adams TS, Adolfsson J, Johansson JE, Lowey J, Trock BJ, Partin AW, Walsh PC, Trent JM, Duggan D, Carpten J, Chang BL, Gronberg H, Isaacs WB, Xu J. Evidence for two independent prostate cancer risk-associated loci in the HNF1B gene at 17q12. Nat Genet 2008;40:1153–5. [PubMed: 18758462]

Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. Genomics 2008;92:265–72. [PubMed: 18722519]

Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y. ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res 2008;36:W377–84. [PubMed: 18508807]

Vink JM, Smit AB, de Geus EJ, Sullivan P, Willemsen G, Hottenga JJ, Smit JH, Hoogendijk WJ, Zitman FG, Peltonen L, Kaprio J, Pedersen NL, Magnusson PK, Spector TD, Kyvik KO, Morley KI, Heath AC, Martin NG, Westendorp RG, Slagboom PE, Tiemeier H, Hofman A, Uitterlinden AG, Aulchenko YS, Amin N, van Duijn C, Penninx BW, Boomsma DI. Genome-wide association study of smoking initiation and current smoking. Am J Hum Genet 2009;84:367–79. [PubMed: 19268276]

Wang K, Li M, Bucan M. Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet 2007;81:1278–1283.

Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, Sleiman PM, Imielinski M, Glessner J, Hou C, Wilson DC, Walters T, Kim C, Frackelton EC, Lionetti P, Barabino A, Van Limbergen J, Guthery S, Denson L, Piccoli D, Li M, Dubinsky M, Silverberg M, Griffiths A, Grant SF, Satsangi J, Baldassano R, Hakonarson H. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet 2009;84:399–405. [PubMed: 19249008]

Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. Trends Genet 2002;18:19–24. [PubMed: 11750696]

Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, Strait J, Duren WL, Maschio A, Busonero F, Mulas A, Albai G, Swift AJ, Morken MA, Narisu N, Bennett D, Parish S, Shen H, Galan P, Meneton P, Hercberg S, Zelenika D, Chen WM, Li Y, Scott LJ, Scheet PA, Sundvall J, Watanabe RM, Nagaraja R, Ebrahim S, Lawlor DA, Ben-Shlomo Y, Davey-Smith G, Shuldiner AR, Collins R, Bergman RN, Uda M, Tuomilehto J, Cao A, Collins FS, Lakatta E, Lathrop GM, Boehnke M, Schlessinger D, Mohlke KL, Abecasis GR. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 2008;40:161–9. [PubMed: 18193043]

Yi G, Sze SH, Thon MR. Identifying clusters of functionally related genes in genomes. Bioinformatics 2007;23:1053–60. [PubMed: 17237058]

Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, Ardlie K, Bostrom KB, Bergman RN, Bonnycastle LL, Borch-Johnsen K, Burtt NP, Chen H, Chines PS, Daly MJ, Deodhar P, Ding CJ, Doney AS, Duren WL, Elliott KS, Erdos MR, Frayling TM, Freathy RM, Gianniny L, Grallert H, Grarup N, Groves CJ, Guiducci C, Hansen T, Herder C, Hitman GA, Hughes TE, Isomaa B, Jackson AU, Jorgensen T, Kong A, Kubalanza K, Kuruvilla FG, Kuusisto J, Langenberg C, Lango H, Lauritzen T, Li Y, Lindgren CM, Lyssenko V, Marvelle AF, Meisinger C, Midthjell K, Mohlke KL, Morken MA, Morris AD, Narisu N, Nilsson P, Owen KR, Palmer CN, Payne F, Perry JR, Pettersen E, Platou C, Prokopenko I, Qi L, Qin L, Rayner NW, Rees M, Roix JJ, Sandbaek A, Shields B, Sjogren M, Steinthorsdottir V, Stringham HM, Swift AJ, Thorleifsson G, Thorsteinsdottir U, Timpson NJ, Tuomi T, Tuomilehto J, Walker M, Watanabe

RM, Weedon MN, Willer CJ, Illig T, Hveem K, Hu FB, Laakso M, Stefansson K, Pedersen O, Wareham NJ, Barroso I, Hattersley AT, Collins FS, Groop L, McCarthy MI, Boehnke M, Altshuler D. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet 2008;40:638–45. [PubMed: 18372903]

Zeggini E, Weedon MN, Lindgren CM, Frayling TM, Elliott KS, Lango H, Timpson NJ, Perry JR, Rayner NW, Freathy RM, Barrett JC, Shields B, Morris AP, Ellard S, Groves CJ, Harries LW, Marchini JL, Owen KR, Knight B, Cardon LR, Walker M, Hitman GA, Morris AD, Doney AS, McCarthy MI, Hattersley AT. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007;316:1336–41. [PubMed: 17463249]

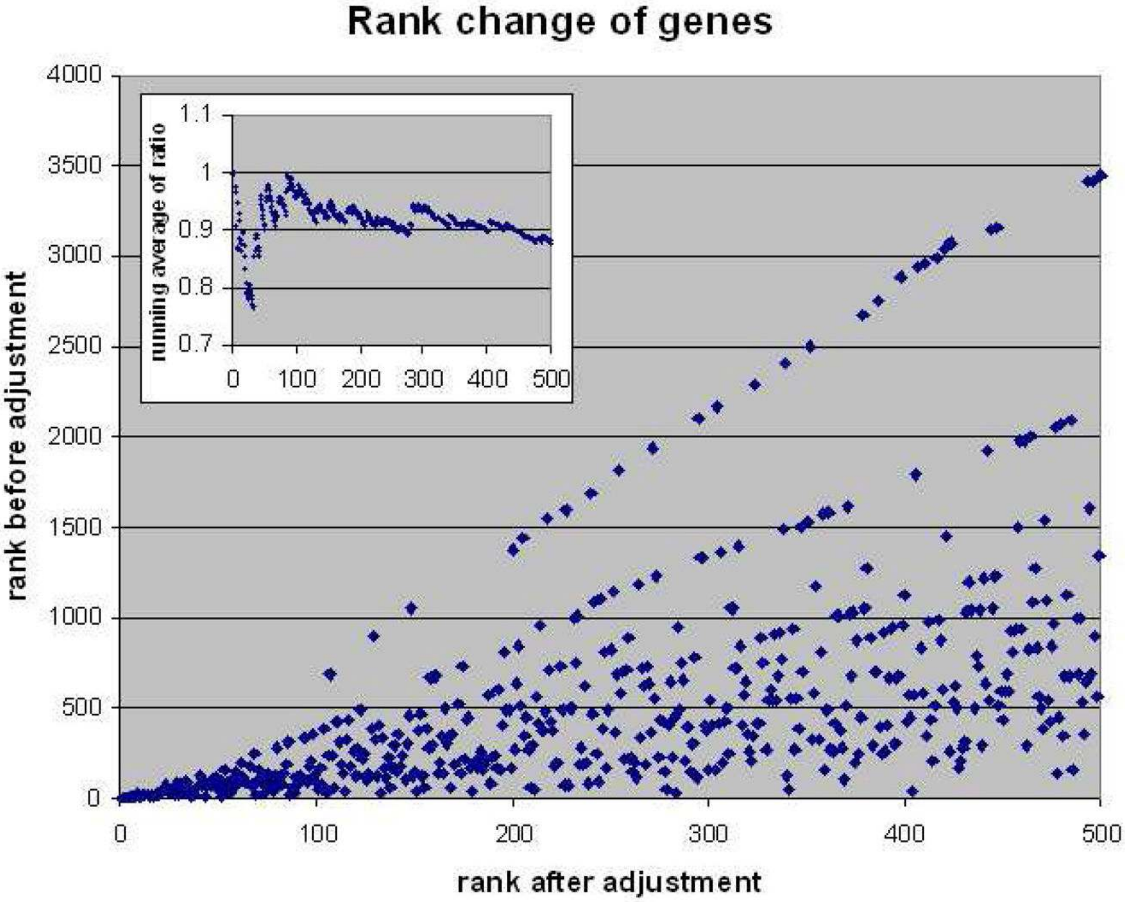

Fig. 1.
Change in the ranks of genes in sample 1 (GWAS of plasma HDL levels) following gene based adjustment for the number of markers spanning the longest known splice-form. Adjusted ranks are plotted along the x-axis and unadjusted ranks along the y-axis. The sub-panel shows the running average of the ratio between the unadjusted and adjusted ranks and illustrates that the relative change is large for the top-ranked genes. The appearance of lines in the main panel reflects groups of genes with similar numbers of markers (the steepest being genes represented by single markers, the second steepest by 2 markers, and so on).

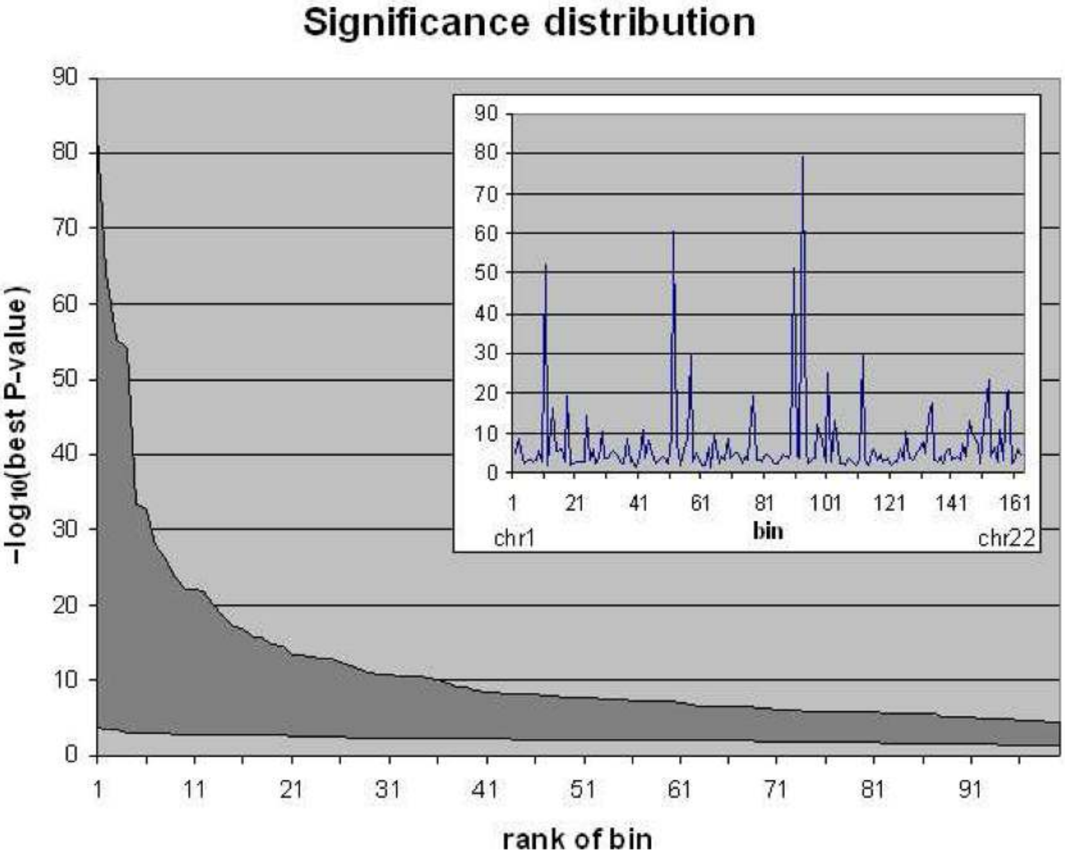

Fig. 2.

Evidence of wide-spread ontology-based clustering of genes in the human genome. The main panel illustrates the distribution of significance for the first 100 clusters (dark shaded area) against a null distribution for random genes (light shaded area at base). The sub-panel depicts significant clusters according to genome position, ordered from pter of chromosome 1 to qter of chromosome 22. The top three peaks represent “olfactory receptor activity”, “cell adhesion”, and “keratinization”, respectively. See supplementary table 1 for a full listing of significant terms.

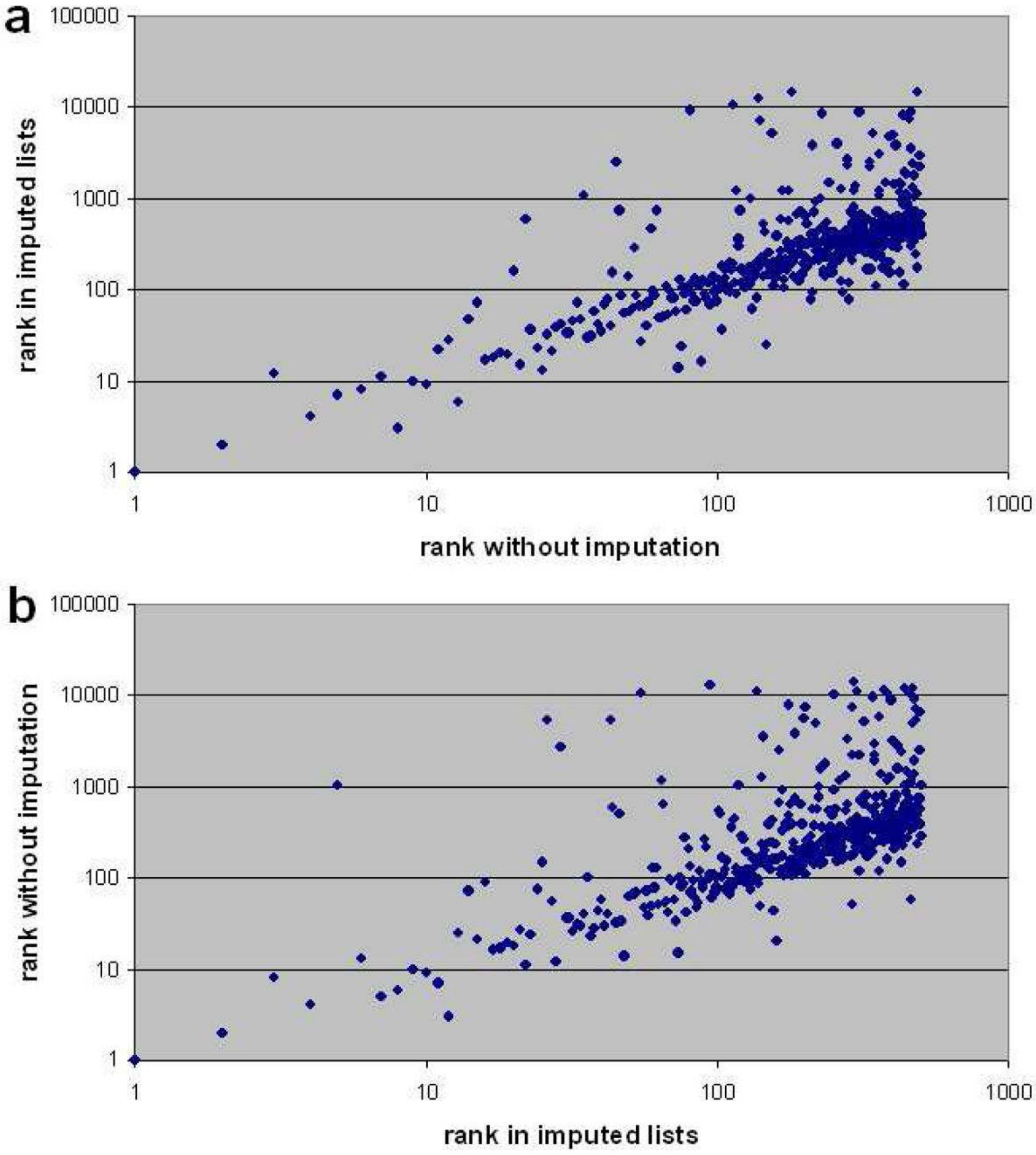

Fig. 3.
Change in gene ranks for a subset of sample 2 (WTCCC type 2 diabetes) using the top 3% of genes from directly genotyped marker lists (x-axis) compared to the same genes derived from imputed lists (y-axis) (3a). For comparison, in figure 3b the top 3% of genes from imputed lists (x-axis) are compared to directly typed markers (y-axis). In both cases, while some extreme outliers occur, most genes retain similar rank. All axes are drawn in $\log_{10}$ scale.

Table 1

Genome-wide association studies

Sample	Trait(s)	Size	Markers[a]	Genes[b]	Reference
Sample 1	HDL, LDL, TG	8816	2.56M	16977	Willer et al. 2008
Sample 2	Type II Diabetes	9562	2.20M	16128	Zeggini et al. 2008
Sample 3	Crohn’s Disease	8059	0.63M	16525	Barrett et al. 2008
Sample 4	HDL, LDL, TG	1754	0.32M	16934	Aulchenko et al. 2009

[a] Total number of typed and/or imputed markers used for gene list creation

[b] Number of genes represented after marker conversion in the meta-analysis data sets.

Table 2

GO terms enriched among long genes

Term	Genes with term	Length	SNP number	HDL trait p-value
synapse	37	4.80E-17	1.00E-16	NS
cell adhesion	61	4.00E-13	2.60E-14	1.00E-06
plasma membrane part	106	2.20E-12	1.30E-10	2.90E-07
membrane	254	3.10E-10	1.50E-11	4.30E-05
nervous system development	52	4.90E-09	5.40E-11	NS

Enriched terms among the longest 3% of genes, by number of markers, and by unadjusted association with the HDL trait. Genes in term refers only to the number in the first analysis of length.

NS = not significant at $\alpha = 0.05$.

Table 3a

GO terms enriched for plasma HDL levels

Term	Genes with term	p-value	Fold change
carboxylesterase activity	7	9.60E-04	6.1
response to external stimulus	14	3.50E-03	2.5
fatty acid metabolic process	7	4.60E-03	4.5
cellular lipid metabolic process	13	6.30E-03	2.4
sterol transport	3	8.60E-03	21

The five most significant terms from among the top functional annotation clusters are shown.

Table 3b

Specific genes contributing to enrichment of the term “cellular lipid metabolic process”

Gene	Rank in GWAS	Marker	p-value (original)
CETP	1	rs7203984	4.55E-20
LPL	2	rs328	2.29E-11
LIPC	3	rs1077834	5.18E-10
ABCA1	7	rs4149274	7.36E-08
MVK	17	rs7954144	4.76E-06
FADS1	21	rs174545	4.18E-05
LYPLA3	26	rs17688076	4.18E-05
PPARA	99	rs9615264	4.44E-05
SMPD2	120	rs6911838	1.14E-03
BMP6	136	rs2876117	5.15E-05
CPT1A	147	rs597539	1.53E-04
LCAT	151	rs4986970	1.07E-03
ADIPOR2	160	rs11061935	1.79E-04

Rank in GWAS is according to the LD-adjusted p-value of the most significant marker associated with each gene. The presented p-value is the unadjusted significance of the specified marker in the original data set.

Table 4a

GO terms enriched for plasma LDL levels

Term	Genes with term	p-value	Fold change
sterol transport	4	4.30E-04	25
lipid transporter activity	6	6.40E-04	8.6
intracellular organelle	88	1.30E-03	1.3
nucleic acid metabolic process	47	8.10E-03	1.4
macromolecular complex assembly	12	1.10E-02	2.4

The five most significant terms from among the top functional annotation clusters are shown.

Table 4b

Specific genes contributing to enrichment of the term "sterol transport"

Gene	Rank in GWAS	Marker	p-value (original)
LDLR	4	rs6511720	6.76E-10
APOB	5	rs531819	1.01E-09
APOE	93	rs405509	1.70E-03
ABCA1	105	rs2000069	2.25E-05

Rank in GWAS is according to the LD-adjusted p-value of the most significant marker associated with each gene. The presented p-value is the unadjusted significance of the specified marker in the original data set.

Table 5a

GO terms enriched for type 2 diabetes

Term	Genes with term	p-value	Fold change
microtubule-based process	10	1.30E-04	5.1
response to nutrient	5	6.10E-04	12.6
advanced glycation end-product receptor activity	3	9.90E-04	59
intracellular organelle	82	1.40E-02	1.2
cellular component assembly	12	1.60E-02	2.3

The five most significant terms from among the top functional annotation clusters are shown.

Table 5b

Specific genes contributing to enrichment of the term "microtubule-based process"

Gene	Rank in GWAS	Marker	p-value (original)
KIF2A	49	rs152189	1.03E-04
KIF21A	78	rs4768736	9.96E-05
KIF3C	100	rs2384298	1.63E-04
HCN4	121	rs1564345	1.77E-04
PCLO	123	rs1157530	4.06E-05
MAP3K11	129	rs7946115	6.47E-04
CENPE	137	rs12506065	5.46E-04
KIF11	150	rs4933734	2.76E-04
DNAL4	158	rs760482	7.89E-04
KPNA2	161	rs4638	3.25E-03

Rank in GWAS is according to the LD-adjusted p-value of the most significant marker associated with each gene. The presented p-value is the unadjusted significance of the specified marker in the original data set.

Table 6a

GO terms enriched for Crohn’s disease

Term	Genes pre-trim	p-value pre-trim	p-value post-trim (genes)	Fold change post-trim
MHC protein complex	6	1.30E-05	NS	NA
immune response	18	4.60E-05	0.011 (13)	2.3
regulation of apoptosis	13	5.30E-03	0.033 (11)	2.1
nucleocytoplasmic transport	6	1.20E-02	no change	4.3
kinase activity	16	1.30E-02	no change	1.9

The five most significant terms from among the top functional annotation clusters are shown.

NS = not signficant at α = 0.05, NA = not applicable.

Table 6b

Specific genes contributing to enrichment of the term "immune response" after trimming

Gene	Rank in GWAS	Marker	Genomic Position	p-value (unadjusted)
IL23R	1	rs11465804	chr1:67.4 Mb	1.01E-35
NOD2	2	rs2076756	chr16:49.2 Mb	3.42E-32
NKX2-3	7	rs10883371	chr10:101.2 Mb	1.82E-10
IRF1	17	rs2188962	chr5:131.8 Mb	4.58E-09
TNFSF15	39	rs6478109	chr9:116.6 Mb	2.61E-07
CCR6	50	rs7749278	chr6:167.4 Mb	3.29E-07
CCL11	77	rs991804	chr17:29.6 Mb	4.01E-06
MICA	88	rs2596503	chr6:31.5 Mb	3.94E-06
HLA-DQA2	99	rs7768538	chr6:32.8 Mb	7.92E-06
HLA-C	121	rs2905747	chr6:31.3 Mb	1.50E-05
IFI30	122	rs7125	chr19:18.1 Mb	3.61E-05
PTPN22	150	rs2476601	chr1:114.2 Mb	1.81E-05
CSF3	160	rs8078723	chr17:35.4 Mb	5.51E-05

Genomic position is according to NCBI Build 36.1 of the human genome.

Rank in GWAS is according to the LD-adjusted p-value of the most significant marker associated with each gene. The presented p-value is the unadjusted significance of the specified marker in the original data set.

Table 7a

GO term enrichment for HDL from GSEA analyses

Term	Core genes	p-value - sample 1	p-value w/o CETP
lipid transport	6	4.00E-04	0.037
lipid transporter activity	5	8.00E-03	0.67

Reported significance is based upon 5000 permutations. The analysis excluding CETP was done in sample 1.

Table 7b

Specific genes contributing to enrichment of the term "lipid transport"

Gene	Rank in GWAS	Rank Metric Score
CETP	1	18.0
ABCA1	7	5.2
PPARA	101	2.8
LCAT	153	2.5
PSAP	257	2.1
ABCD3	292	2.1
PPARD	734	1.5

Rank in GWAS is from sample 1.