Published in final edited form as:
Stat Appl Genet Mol Biol. ; 11(2): . doi:10.2202/1544-6115.1717.

Candidate Pathway Based Analysis for Cleft Lip with or without Cleft Palate

Tian-Xiao Zhang, **Terri H. Beaty**, and **Ingo Ruczinski**
Johns Hopkins University

1 Introduction

Clefts of the lip with/without cleft palate (CL/P) are one of the most common birth defects in humans, and create a major public health burden for the affected children and their families. Although there is strong evidence for genetic control of CL/P, it has proven difficult to identify any single gene as being causal, and it is more likely that several genes contribute to the etiology of this complex and heterogeneous malformation (Dixon et al., 2011). The challenge lies in identifying which genes are involved when several may contribute to risk (Schliekelman and Slatkin, 2002). It is clear that mutations within a single gene can control a recognized Mendelian syndrome that includes CL/P as a hallmark feature, and can also show consistent evidence of association in families with isolated, non-syndromic CL/P. In particular, mutations in the IRF6 gene on chromosome 1q32 account for most cases of Van der Woude Syndrome, an autosomal dominant syndrome (with a prevalence of 1/50,000 among livebirths) involving CL/P generally accompanied by lip pits in individuals carrying risk alleles. Polymorphic markers in IRF6 also show consistent evidence of statistical association with isolated, non-syndromic CL/P in both case-control and family based tests, suggesting common variants in this gene also influence risk, not just rare mutations. Traditional genetic approaches such as linkage analysis using multiplex families (i.e. those with two or more affected individuals) are effective in mapping causal genes controlling Mendelian syndromes (and were critical in identifying IRF6 as causal for Van der Woude syndrome). Meta-analysis of multiple linkage studies have identified several regions of the genome that likely harbor causal genes controlling risk to non-syndromic CL/P (Marazita et al., 2004, 2009). However, there is considerable locus heterogeneity among these multiplex families used in these linkage studies, where different genes appear to be acting in different families. Furthermore, only a modest fraction of isolated, non-syndromic CL/P cases have any positive family history (i.e. most cases are from simplex families where no other relatives are affected beyond the proband).

Genome wide association studies (GWAS) represent a useful study design for identifying causal genes associated with polymorphic markers that tag unobserved high risk alleles through linkage disequilibrium (LD), and this approach can be exploited to identify causal genes in an unbiased genome wide context. GWAS have proven to be useful in identifying novel genes (some of which may be causal) for complex and heterogeneous diseases (McCarthy et al., 2008; Hindorff et al., 2009; Manolio and Collins, 2009). There have been two GWAS of CL/P using population based study designs, both with cases and controls of European ancestry (Birnbaum et al., 2009; Grant et al., 2009). Both of these studies identified a novel region of 8q24 as strongly associated with risk to CL/P, but the markers showing the strongest signal were not located in any known gene and in fact appeared to be in a “gene desert”. Subsequent analysis of the German case-control data supplemented by additional case-parent trios, strengthened evidence for two additional genes (VAX1 on chromosome 10q25 and a region on chromosome 17q22 near NOG) (Mangold et al., 2010).

A key problem with case-control studies is their susceptibility to confounding due to population stratification which becomes critical when drawing cases from multiple, genetically distinct populations. As part of the project International Consortium to Identify Genes & Interactions Controlling Oral Clefts, we conducted a GWAS to identify genes influencing risk to oral clefts, either directly, or through interaction with common maternal exposures, using case-parent trios assembled from an international consortium. The case-parent trio design, where the key genetic (allelic or genotypic) contrasts are within a family, minimizes the potential for the above described confounding, and gives a more robust test for association between markers (and potentially causal genes), and the outcome of interest. Study subjects were recruited from 13 different sites in Europe, the US, China, Taiwan, Singapore, Korea and the Philippines, and maternal exposures such as smoking and alcohol consumption were recorded. In 2009, genotyping using the Illumina 610Quad array was completed for more than 2000 case-parent trios. Analysis of the entire genome wide marker panel using the allelic transmission disequilibrium test (Spielman et al., 1993; Spielman and Ewens, 1996) yielded several regions of significance from this case-parent trio study (Beaty et al., 2010). Similar to the findings of Birnbaum et al. (2009) and Grant et al. (2009), the most significant signal was in the gene desert on chromosome 8q24, where markers well removed from any known gene yielded strong signal of linkage and association. Further, two novel genes (MAFB on chromosome 20 and ABCA4 on chromosome 1) achieved genome wide significance.

The underlying rationale for GWAS is the assumption that common complex diseases are attributable in part to allelic variants reasonably common in a population (“common disease, common variant” hypothesis). And while GWAS have been very successful in identifying hundreds of genetic markers associated with many different complex diseases, any individual variant typically only represents a small increment in risk for a particular disease, and together, they can usually explain only a small proportion of the familial clustering (heritability) observed (Manolio et al., 2009; Eichler et al., 2010). The potential sources of the “missing heritability” are manyfold, and much attention has for example shifted towards assessing the effects of rare variants (with possibly larger effect sizes), which are poorly tagged by standard genotyping arrays (Manolio et al., 2009; McClellan and King, 2010; Dickson et al., 2010). Other possible explanations for the rather limited impact of individual markers identified from GWAS include the presence of DNA copy number variants, epigenetic effects, epistatis and gene-environment interactions, but also the ambiguity in the definition of heritability and the potential role of environmental variables, and in the instances of highly complex diseases, the potential difficulties in accurate phenotype delineation. We are currently mining our oral cleft case-parent trio data to assess effects of gene-environment interactions (with maternal smoking, alcohol consumption, and vitamin supplementation as environmental variables, Wu et al., 2010; Beaty et al., 2011), epistatic interactions (Wang et al., 2011), de-novo copy number events in the affected probands (Scharpf et al., 2011), and parent of origin effects (Sull et al., 2008; Shi et al., 2011). In addition, we recently obtained imputed genotypes for subjects of European and Asian ancestry in these consortium data, generated with the software package BEAGLE (Browning and Browning, 2009) using a HapMap Phase III reference panel, and we are assessing effects of genetic distance and SNP selection biases on such imputed data, using a newly developed genotypic transmission disequilibrium test that accommodates uncertain or imputed genotypes (Taub et al., 2011; Murray et al., 2011).

A criticism of the commonly employed approaches for GWAS analyses is that they do not use the full information available, and there remains the need to develop novel research strategies beyond current genome-wide association approaches (Manolio et al., 2009). The predominant strategy for analyzing GWAS data is to carry out SNP specific (marginal) tests such as the Cochran-Armitage trend test, sort the results from the smallest to the largest p-

value, and declare significance based on a Bonferroni correction to limit the family-wise error rate at a fixed level (typically, $\alpha = 5\%$). The benefit of this approach is its straightforward and reproducible implementation, but from both scientific and statistical perspectives, this approach is sub-optimal. Enforcing a tight control on the family-wise error rate, i.e. the probability of declaring at least one SNP significant that is not associated with the phenotype, comes at the expense of truly associated SNPs which do not achieve "genome wide" significance. This stringent type 1 error (false positive) control becomes particularly problematic for complex disease GWAS, since typically many SNPs with very modest effect sizes contribute to disease risk. Thus, many or most of those markers might be missed, particularly when the sample size is small and/or the respective minor allele frequencies are low. Further, controlling the family wise error rate via Bonferroni completely ignores type 2 errors (false negatives), and employs a uniform threshold for significance without considering power. Using such a constant threshold in a frequentist setting leads to an inconsistent procedure, and viewed from a decision theory perspective, leads to an inadmissible procedure (Wakefield, 2007, 2008, 2009). Possibly better suited with regard to these criticisms are Bayes ranking procedures (Louis and Ruczinski, 2010). This approach also ignores other possibly valuable information, for example findings from prior linkage studies that could boost power to detect associations when properly taken into account (Roeder et al., 2006, 2007). Other biological information available *a priori* concerns genes and their respective roles in biological pathways, delineated from gene or protein networks. Pathway-based approaches to jointly consider multiple variants and/or genes can complement GWAS results, and even reveal completely new findings that would have been missed when testing marginal associations alone, and thereby illuminate important genetic contributions to complex disease (Baranzini et al., 2009; Eleftherohorinou et al., 2009; Cantor et al., 2010). In this manuscript we employ the pathway-based approach proposed by Wang et al. (2007) to analyze case-parent trios from our International Cleft Consortium.

2 Methods

A typical pathway-based analysis requires the selection of pre-defined pathways from an existing data base such as KEGG (Kanehisa et al., 2010), GO (Ashburner et al., 2000; Gene Ontology Consortium, 2010) or PANTHER (Thomas et al., 2003). Further required are strategies to assign SNPs to genes, approaches to summarize the information available within genes and pathways, and statistical methods for inference, valid with regards to type I errors, coverage, etc. We were interested in pathways that contained genes previously implicated in facial cleft linkage and association studies. In a benchmark study, Jugessur et al. (2009) proposed such a list of 356 candidate genes that had shown some evidence of playing a rule in facial clefting, included the previously reported genes TGFA and IRF6 among other promising candidate genes for clefts. The pathway-based analysis in this manuscript was then carried out following the proposed procedure of Wang et al. (2007), a modification of the gene-set enrichment analysis (GSEA) algorithm (Subramanian et al., 2005). This method is implemented in the freely available software suite GenGen (www.openbioinformatics.org/gengen/). In short, the approach can be described as follows:

1. Based on a list of biologically relevant genes (here, from Jugessur et al., 2009), delineate pathways containing these genes (here, pathways identified by PANTHER; Thomas et al., 2003). To avoid testing overly narrow or broad functional categories, we limited ourselves to pathways containing at least 5, but not more than 200 genes.
2. Map markers (SNPs) to genes. We included markers within 500kb of annotated genes, since most enhancers and repressors fall within 500kb from recognized genes, and most LD blocks are less than 500kb in size. SNPs located in a region shared by two genes were mapped to both genes. Note that individual SNPs are

rarely shared by two genes, however, sharing genes across pathways is more common.

3. For each SNP the χ^2 statistic and the p-value were calculated using the allelic transmission disequilibrium test. The gene-wise statistic is defined as the highest χ^2 statistic among all markers within that gene.

4. For all N genes ($G_1,\ldots,G_N$) in the analysis, rank the test statistics in a descending order. Denote those as $r_{(1)},\ldots,r_{(N)}$.

5. A statistic called the "enrichment score" (ES) is calculated for each gene set (S) defined by a pathway (cardinality N_H). This statistic is defined as

$$ES(S) = \max_{1\le j\le N}\left\{\sum_{G_{j^*}\in S, j\le j^*}\frac{|r_{(j^*)}|}{N_R} - \sum_{G_{j^*}\notin S, j\le j^*}\frac{1}{N-N_H}\right\},$$

where $N_R = \sum_{G_{j^*}\in S}|r_{(j^*)}|$, and reflects the over-representation of significant genes from the gene set S among all genes in the analysis. Citing Wang et al. (2007): "The enrichment score, ES(S), measures the maximum deviation of concentration of the statistic values in gene set S from a set of randomly picked genes in the genome. Therefore, if the association signal in S is concentrated at the top of the list, then ES(S) will be high." Note that this test statistic can be understood as a weighted Kolmogorov-Smirnov-like running-sum statistic (Hollander and Wolfe, 1999) giving higher weight to genes with large test statistics (see Wang et al., 2007, for more details).

6. Since the enrichment score depends on the maximum test statistic, and thus the number of markers per gene, a permutation procedure is carried out. The transmitted/non-transmitted status for alleles at each marker are randomly shuffled, and the enrichment scores for each gene set S is re-calculated. For a certain number of permutations (here, n=1000 permutations), these enrichment scores are denoted by $ES(S, \pi)$, where π refers to the individual permutation. Then, a "normalized enrichment score" (NES) can be calculated to to make gene sets directly comparable:

$$NES(S) = \frac{ES(S) - \text{Mean}[ES(S,\pi)]}{SD[ES(S,\pi)]}$$

7. These permutations are also used to calculate nominal p-values for each gene set (pathway), simply as the proportion of $ES(S, \pi)$ exceeding the observed enrichment score $ES(S)$. In addition, the family-wise error rate (FWER) can be controlled by an adjustment procedure based on the "normalized enrichment score". The FWER p-value for a gene set can be calculated as the proportion of all permutations whose highest NES score across all gene sets is higher than the observed NES score for this particular gene set.

Since the FWER is a very stringent criterion, aiming to exclude any false positive results (typically, at the expense of false negatives), the set of significant pathways after adjusting for multiple comparisons this way can be small, or even empty. Thus, biologically relevant information might be discarded if none of the gene sets exceeds the FWER criterion, but signal is present in the data nonetheless, which would be evidenced by an enrichment of low p-values, for example. The false discovery rate (FDR) is often employed to assess the

expected number of true and false positives among a set of hypothesis tests declared significant. Another approach to assess overall departure from randomness is Fisher's inverse χ^2 test (Fisher, 1925). This approach derives a combined statistic from the p-values of k independent hypothesis tests (p_1,…, p_k). Under the global null of no signal for any of the hypothesis tests, the test statistic $-2\sum_{j=1}^{k}\log\left(p_j\right)$ asymptotically follows a χ^2 distribution with $2k$ degrees of freedom. However, in the results presented here the assumption of independence is clearly violated, since two or more pathways can contain the same gene (and their markers). A second layer of permutations was carried out to circumvent this issue (Figure 1). Similar to the above, the test statistic was based on the nominal gene set p-values, defined as $T=-2\sum_{j=1}^{|S|}\log\left(p_j+\delta\right)$. Here, δ is a small constant to avoid numeric problems in the inference, and in this application was chosen (ad-hoc) as 10^{-3}. Note that the nominal p-values of each gene set were calculated using the initial 1000 permutations as reference group (Figure 1, left). To delineate the null distribution of this test statistic T, a second layer of 1000 permutations was carried out by shuffling the transmitted/non-transmitted labels for alleles at each SNP, and deriving enrichment and normalized enrichment scores for each pathway. (Figure 1, right). For each of the new permutations, gene set p-values are derived by comparing the enrichment scores to the same initial 1000 permutations that served as reference for obtaining the nominal p-values for the observed data. That is, for each permutation $\tilde{\pi}$, we calculate $T_{\tilde{\pi}}=-2\sum_{j=1}^{|S|}\log\left(p_j^{\tilde{\pi}}+\delta\right)$, which serves as the null distribution for the observed T. The overall p-value can de derived as the proportion of $T_{\tilde{\pi}}$ that exceed T. The procedure for obtaining this exact p-value was implemented in the statistical environment R (http://cran.r-project.org/).

3 Results

Case-parent trios used here originated from an international consortium (The Gene, Environment Association Studies consortium, GENEVA) formed in 2007 by several research groups to conduct a genome-wide search for genes influencing risk to oral clefts using a case-parent trio design. A total of 5,742 individuals from 1,908 CL/P case-parents trios (1,591 complete families with one or more affected probands, and 317 incomplete trios where one parent was missing) were genotyped using the Illumina Human610-Quad v.1B BeadChip at the Center for Inherited Disease Research (CIDR), one of the two genotyping centers supported by the GENEVA consortium (Beaty et al., 2010). Genotype clusters for each SNP were determined using the BeadStudio Module (version 3.3.7), and combined intensity data from 99.2% of samples were used to define clusters and call genotypes. Both the mean SNP call rates and the mean sample call rates were over 99.8%. Genotypes were released for 589,945 SNPs (99.56% of those attempted). Duplicate samples from both HapMap and study subjects were included on each plate, and reproducibility rates in the raw data were 99.99% among 161 duplicated subjects.

Individuals were dropped if they had 1) unacceptably high rates of missing genotype calls (larger than 5%), or 2) unacceptably high rates of Mendelian errors between parents and children (larger than 5%). Four categories of quality control flags for all autosomal SNPs were set: 1) unacceptably high rates of missing genotype calls (larger than 10%), 2) low minor allele frequency (minor allele frequency less than 1%), 3) unacceptably high rates of Mendelian errors between parents and offspring (larger than 5%), and 4) deviation from Hardy-Weinberg equilibrium in founders ($p<0.00001$). As described in Beaty et al. (2010), 569,294 autosomal SNPs were available for analysis. The GWAS yielded several genome-wide significant hits, in previously implicated and novel genes, but also various intergenic regions (Table 1).

A total of 51 pathways were identified by PANTHER (Thomas et al., 2003) that included at least one of the 356 oral cleft candidate genes (Jugessur et al., 2009). Among those pathways, 42 contained between 5 and 200 genes. DNA from 5742 individuals (2589 affected oral cleft probands, and 3153 parents) of predominantly Asian and European descent was collected from 13 different populations (Beaty et al., 2010). For the pathway analysis, a total of 40,208 autosomal markers (subject to quality control constraints, see above) located in 1564 genes (or within 500kb up- or downstream from a gene), contained in at least one of the 42 pathways, were considered. The test statistics relevant for pathway analysis were derived from 1,604 complete CL/P case-parent trios (Table 2).

Analysis of the entire genome wide marker panel yielded several regions of significance from this case-parent trio study (Beaty et al., 2010). However, the most significant signal was in a gene desert on chromosome 8q24, where markers well removed from any known gene yielded the most significant p-values. Further, two novel genes (MAFB on chromosome 20 and ABCA4 on chromosome 1) achieved genome wide significance, but since these were not recognized candidate genes for CL/P listed by Jugessur et al. (2009), they were not included in this pathway analysis. The one recognized candidate gene for CL/P, IRF6 on chromosome 1q32, was not included in any of the pathways identified in the PANTHER data base, and so did not get included in this analysis. One marker included in the pathway analysis reached genome-wide significance in the original GWAS (rs11538422 located in gene PDK1, belonging to the TCA cycle, with a p-value of 6.7×10^{-10}). The lowest individual p-value after this was 1.7×10^{-6} for SNP rs704574 in the COL8A1 gene, contained in the “Integrin signaling pathway”.

We carried out the pathway-based analysis following the procedure proposed by Wang et al. (2007), as described in the Methods section, and found several pathways exhibiting nominally significant ($p < 0.05$) NES scores (Table 3). However, none of these pathways could be considered significant when controlling the family-wise error rate at significance level 5% via the Bonferroni correction ($p_{\text{threshold}} = 0.05/42 \approx 0.0012$), or the less conservative permutation based FWER control. Noticeably, three out of the five pathways with the lowest multiple comparsion corrected p-value (i.e., the Angiogenesis, Huntington disease, Cadherin signaling pathways) were among the largest pathways, containing up to 173 genes for the Angiogenesis pathway. The most “significant” pathway (Cytoskeletal regulation, $p_{\text{FWER}} = 0.212$) contained 82 genes. Among the smaller pathways, the TCA cycle had one of the lowest nominal p-values ($p = 0.03$) but was not among the top ten after multiple comparisons correction via permutation tests $p_{\text{FWER}} = 0.983$. However, the distribution of the observed p-values appears to show a clear deviation from a Uniform (0,1) which would be expected under the global null of “no signal in the data” (Figure 2). This is also reflected in the respective q-values (Storey and Tibshirani, 2003) of these pathways: for example, 13 pathways are selected at a false discovery rate of 20% (Table 3).

To further assess whether or not this deviation can be explained by chance alone, we carried out a second layer of permutation tests, as described in the methods section. The combined test statistic derived from the p-values (Figure 2) was compared to those derived from an additional 1000 permutations, letting the initial 1000 permutations serve as reference (Figure 3). Only 29 out of these 1000 permutations yielded a more extreme test statistic than the observed data, and thus, we can quote an estimated p-value of 2.9%. While this pathway-based analysis did not yield a clear significant result for any particular pathway, we can conclude that one or more of the genes and pathways considered here likely do play a role in oral clefting.

4 Discussion

In this manuscript, we presented a pathway-based approach to analyze case-parent trios from our International Cleft Consortium. The analysis is part of our ongoing efforts to find additional signal in the genome scans of the CL/P probands and their parents, beyond the hits reported in Beaty et al. (2010). Assessing the risk of oral clefts, we detected the presence of interactions between candidate genes and environmental factors such as maternal smoking and maternal alcohol consumption (Wu et al., 2010; Beaty et al., 2011), reported on parent of origin effects (Sull et al., 2008; Shi et al., 2011), and detected an epistatic interaction between markers of two candidate genes (Wang et al., 2011). Unfortunately, the pathway-based analysis did not yield a clear result, as none of the pathway enrichment tests was significant after multiple comparisons correction. Nonetheless, we believe the analysis has generated some novel clues about the genetic under-pinning of oral clefts. We observed a departure from randomness, indicating one or more pathways investigated here is involved in these biological processes. Figure 2 and in particular Figure 3 suggest that these could actually be a dozen or so, though we readily admit that this statement is not completely on solid quantitative grounds. There are very limited clues in the literature how the pathways reported here might relate to oral clefts. An exception is TP63, a gene that encodes a member of the p53 family of transcription factors, and belongs to the Huntington disease pathway (#4 in Table 3), the p53 pathway feedback loops 2 (#10), and the p53 pathway (#13). Mutations in the TP63 gene have been implicated in several Mendelian malformation syndromes that can include CL/P, and are the cause for Split-Hand/Foot Malformation (Ianakiev et al., 2000) and Ankyloblepharon-Ectodermal Dysplasia-Clefting syndrome (McGrath et al., 2001), and limb-mammary syndrome (van Bokhoven et al., 2001).

We note one shortcoming of such a gene-set based approach is the fact that for every genotyping platform there will be genes not covered by the respective probes. The distribution of these probes across the genome is mainly driven by considerations of coverage (loosely speaking, the average proportion of the variation in the genome that can be captured with these probes), and while there is some enrichment on some of the platforms for exonic regions, a fair fraction of genes will not be tagged (in our case 238 genes could not be used due to the lack of markers). Even worse, some pathways such as the folate synthesis pathway containing only six genes and thought to be related to CL/P, could not be assessed at all. These issues will likely be alleviated with the advances in whole exome and whole genome sequencing studies. However, from a statistical vantage point, the vastly different number of genes within pathways and SNPs within genes raise a flag. In the approach of Wang et al. (2007), each gene is represented by its most significant marker, and thus, larger genes have higher gene-wise test statistics on average. When permutation tests are correctly employed, the significance levels of the hypothesis tests are protected. However, while we do not have to worry about the type I error, the type II error is of concern, and it is virtually impossible to know precisely in advance what the power of the procedure will be, and how it could be improved. The same concern also applies for the vastly differing number of genes per pathway. We anticipate further development of pathway based approaches, possibly those incorporating other biological data such as transcript levels from RNAseq, will be a highly active research area.

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, the gene ontology consortium. Gene ontology: tool for the unification of biology. Nat Genet. 2000; 25:25–29. [PubMed: 10802651]

Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BMJ, Kappos L, Consortium GSA, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. Hum Mol Genet. 2009; 18:2078–2090. [PubMed: 19286671]

Beaty TH, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski JB, Liang KY, Wu T, Murray T, Fallin MD, Redett RA, Raymond G, Schwender H, Jin S-C, Cooper ME, Dunnwald M, Mansilla MA, Leslie E, Bullard S, Lidral AC, Moreno LM, Menezes R, Vieira AR, Petrin A, Wilcox AJ, Lie RT, Jabs EW, Wu-Chou YH, Chen PK, Wang H, Ye X, Huang S, Yeow V, Chong SS, Jee SH, Shi B, Christensen K, Melbye M, Doheny KF, Pugh EW, Ling H, Castilla EE, Czeizel AE, Ma L, Field LL, Brody L, Pangilinan F, Mills JL, Molloy AM, Kirke PN, Scott JM, Scott JM, Arcos-Burgos M, Scott AF. A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. Nat Genet. 2010; 42:525–529. [PubMed: 20436469]

Beaty TH, Ruczinski I, Murray JC, Marazita ML, Munger RG, Hetmanski JB, Murray T, Redett RJ, Fallin MD, Liang KY, Wu T, Patel PJ, Jin S-C, Zhang TX, Schwender H, Wu-Chou YH, Chen PK, Chong SS, Cheah F, Yeow V, Ye X, Wang H, Huang S, Jabs EW, Shi B, Wilcox AJ, Lie RT, Jee SH, Christensen K, Doheny KF, Pugh EW, Ling H, Scott AF. Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. Genet Epidemiol. 2011; 35:469–478. [PubMed: 21618603]

Birnbaum S, Ludwig KU, Reutter H, Herms S, Steffens M, Rubini M, Baluardo C, Ferrian M, de Assis NA, Alblas MA, Barth S, Freudenberg J, Lauster C, Schmidt G, Scheer M, Braumann B, Berg SJ, Reich RH, Schiefke F, Hemprich A, Ptzsch S, Steegers-Theunissen RP, Ptzsch B, Moebus S, Horsthemke B, Kramer F-J, Wienker TF, Mossey PA, Propping P, Cichon S, Hoffmann P, Knapp M, Nthen MM, Mangold E. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nat Genet. 2009; 41:473–477. [PubMed: 19270707]

Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet. 2009; 84:210–223. [PubMed: 19200528]

Cantor RM, Lange K, Sinsheimer JS. Prioritizing GWAS results: A review of statistical methods and recommendations for their application. Am J Hum Genet. 2010; 86:6–22. [PubMed: 20074509]

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 2010; 8:e1000294. [PubMed: 20126254]

Dixon MJ, Marazita ML, Beaty TH, Murray JC. Cleft lip and palate: understanding genetic and environmental influences. Nat Rev Genet. 2011; 12:167–178. [PubMed: 21331089]

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet. 2010; 11:446–450. [PubMed: 20479774]

Eleftherohorinou H, Wright V, Hoggart C, Hartikainen A-L, Jarvelin M-R, Balding D, Coin L, Levin M. Pathway analysis of GWAS provides new insights into genetic susceptibility to 3 inflammatory diseases. PLoS One. 2009; 4:e8068. [PubMed: 19956648]

Fisher, RA. Statistical Methods for Research Workers. Oliver and Boyd; Edinburg and London: 1925.

Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010; 38:D331–D335. [PubMed: 19920128]

Grant SFA, Wang K, Zhang H, Glaberson W, Annaiah K, Kim CE, Bradfield JP, Glessner JT, Thomas KA, Garris M, Frackelton EC, Otieno FG, Chiavacci RM, Nah H-D, Kirschner RE, Hakonarson H. A genome-wide association study identifies a locus for nonsyndromic cleft lip with or without cleft palate on 8q24. J Pediatr. 2009; 155:909–913. [PubMed: 19656524]

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]

Hollander, M.; Wolfe, DA. Nonparametric statistical methods. Wiley; New York: 1999.

Ianakiev P, Kilpatrick MW, Toudjarska I, Basel D, Beighton P, Tsipouras P. Split-hand/split-foot malformation is caused by mutations in the p63 gene on 3q27. Am J Hum Genet. 2000; 67:59–66. [PubMed: 10839977]

Jugessur A, Shi M, Gjessing HK, Lie RT, Wilcox AJ, Weinberg CR, Christensen K, Boyles AL, Daack-Hirsch S, Trung TN, Bille C, Lidral AC, Murray JC. Genetic determinants of facial clefting: analysis of 357 candidate genes using two national cleft studies from Scandinavia. PLoS One. 2009; 4:e5385. [PubMed: 19401770]

Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010; 38:D355–D360. [PubMed: 19880382]

Louis TA, Ruczinski I. Efficient evaluation of ranking procedures when the number of units is large, with application to SNP identification. Biom J. 2010; 52:34–49. [PubMed: 20131327]

Mangold E, Ludwig KU, Birnbaum S, Baluardo C, Ferrian M, Herms S, Reutter H, de Assis NA, Chawa TA, Mattheisen M, Steffens M, Barth S, Kluck N, Paul A, Becker J, Lauster C, Schmidt G, Braumann B, Scheer M, Reich RH, Hemprich A, Ptzsch S, Blaumeiser B, Moebus S, Krawczak M, Schreiber S, Meitinger T, Wichmann H-E, Steegers-Theunissen RP, Kramer F-J, Cichon S, Propping P, Wienker TF, Knapp M, Rubini M, Mossey PA, Hoffmann P, Nthen MM. Genome-wide association study identifies two susceptibility loci for nonsyndromic cleft lip with or without cleft palate. Nat Genet. 2010; 42:24–26. [PubMed: 20023658]

Manolio TA, Collins FS. The hapmap and genome-wide association studies in diagnosis and therapy. Annu Rev Med. 2009; 60:443–456. [PubMed: 19630580]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

Marazita ML, Lidral AC, Murray JC, Field LL, Maher BS, McHenry TG, Cooper ME, Govil M, Daack-Hirsch S, Riley B, Jugessur A, Felix T, Morene L, Mansilla MA, Vieira AR, Doheny K, Pugh E, Valencia-Ramirez C, Arcos-Burgos M. Genome scan, fine-mapping, and candidate gene analysis of non-syndromic cleft lip with or without cleft palate reveals phenotype-specific differences in linkage and association results. Hum Hered. 2009; 68:151–170. [PubMed: 19521098]

Marazita ML, Murray JC, Lidral AC, Arcos-Burgos M, Cooper ME, Goldstein T, Maher BS, Daack-Hirsch S, Schultz R, Mansilla MA, Field LL, e Liu Y, Prescott N, Malcolm S, Winter R, Ray A, Moreno L, Valencia C, Neiswanger K, Wyszynski DF, Bailey-Wilson JE, Albacha-Hejazi H, Beaty TH, McIntosh I, Hetmanski JB, Tunbilek G, Edwards M, Harkin L, Scott R, Roddick LG. Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35. Am J Hum Genet. 2004; 75:161–173. [PubMed: 15185170]

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9:356–369. [PubMed: 18398418]

McClellan J, King M-C. Genetic heterogeneity in human disease. Cell. 2010; 141:210–217. [PubMed: 20403315]

McGrath JA, Duijf PH, Doetsch V, Irvine AD, de Waal R, Vanmolkot KR, Wessagowit V, Kelly A, Atherton DJ, Griffiths WA, Orlow SJ, van Haeringen A, Ausems MG, Yang A, McKeon F, Bamshad MA, Brunner HG, Hamel BC, van Bokhoven H. Hay-Wells syndrome is caused by heterozygous missense mutations in the SAM domain of p63. Hum Mol Genet. 2001; 10:221–229. [PubMed: 11159940]

Murray T, Ruczinski I, Hetmanski JB, Scott AF, Taub M, Patel P, Zhang TX, Murray JC, Marazita ML, Munger RG, Wilcox AJ, Ye X, Wang H, Wu-Chou YH, Shi B, Chong SS, Yeow V, Lie RT, Beaty TH. The impact of genetic distance and SNP selection bias on the strength of Cleft Lip/Palate signal in chromosome 8q24 between Asians and Europeans. 2011 under review.

Roeder K, Bacanu S-A, Wasserman L, Devlin B. Using linkage genome scans to improve power of association in genome scans. Am J Hum Genet. 2006; 78:243–252. [PubMed: 16400608]

Roeder K, Devlin B, Wasserman L. Improving power in genome-wide association studies: weights tip the scale. Genet Epidemiol. 2007; 31:741–747. [PubMed: 17549760]

Scharpf RB, Beaty TH, Schwender H, Scott AF, Ruczinski I. Fast detection of de novo copy number variants from SNP arrays: a case-parent study of cleft lip and palate. 2011 under review.

Schliekelman P, Slatkin M. Multiplex relative risk and estimation of the number of loci underlying an inherited disease. Am J Hum Genet. 2002; 71:1369–1385. [PubMed: 12454800]

Shi M, Murray JC, Marazita ML, Munger RG, Ruczinski I, Hetmanski J, Wu T, Murray T, Redett RA, Wilcox AJ, Lie RT, Wu-Chou YH, Chen PK, Wang H, Ye X, Yeow V, Chong S, Shi B, Christensen K, Scott AF, Patel P, Cheah F, Beaty TH. Genome wide study of maternal and parent-of-origin effects on the etiology of orofacial clefts. 2011 under review.

Spielman RS, Ewens WJ. The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet. 1996; 59:983–989. [PubMed: 8900224]

Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52:506–516. [PubMed: 8447318]

Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003; 100:9440–9445. [PubMed: 12883005]

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; 102:15545–15550. [PubMed: 16199517]

Sull JW, Liang K-Y, Hetmanski JB, Fallin MD, Ingersoll RG, Park J, Wu-Chou Y-H, Chen PK, Chong SS, Cheah F, Yeow V, Park BY, Jee SH, Jabs EW, Redett R, Jung E, Ruczinski I, Scott AF, Beaty TH. Differential parental transmission of markers in RUNX2 among cleft case-parent trios from four populations. Genet Epidemiol. 2008; 32:505–512. [PubMed: 18357615]

Taub M, Schwender H, Beaty TH, Louis TA, Ruczinski I. Incorporating genotype uncertainties into the genotypic TDT for main effects and gene-environment interactions (in revision). 2011

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. Panther: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13:2129–2141. [PubMed: 12952881]

van Bokhoven H, Hamel BC, Bamshad M, Sangiorgi E, Gurrieri F, Duijf PH, Vanmolkot KR, van Beusekom E, van Beersum SE, Celli J, Merkx GF, Tenconi R, Fryns JP, Verloes A, Newbury-Ecob RA, Raas-Rotschild A, Majewski F, Beemer FA, Janecke A, Chitayat D, Crisponi G, Kayserili H, Yates JR, Neri G, Brunner HG. p63 gene mutations in EEC syndrome, limb-mammary syndrome, and isolated split hand-split foot malformation suggest a genotype-phenotype correlation. Am J Hum Genet. 2001; 69:481–492. [PubMed: 11462173]

Wakefield J. A bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet. 2007; 81:208–227. [PubMed: 17668372]

Wakefield J. Reporting and interpretation in genome-wide association studies. Int J Epidemiol. 2008; 37:641–653. [PubMed: 18270206]

Wakefield J. Bayes factors for genome-wide association studies: comparison with p-values. Genet Epidemiol. 2009; 33:79–86. [PubMed: 18642345]

Wang H, Wu T, Hetmanski JB, Ruczinski I, Schwender H, Liang KY, Murray T, Fallin MD, Redett RA, Raymond G, Jin SC, Wu-Chou YH, Chen PK, Yeow V, Park BY, Chong SS, Cheah F, Jee SH, Ingersoll RG, Jabs EW, Scott AF, Beaty TH. The FGF and FGFR gene family and risk of cleft lip with/without cleft palate. 2011 in press.

Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet. 2007; 81:1278–1283. [PubMed: 17966091]

Wu T, Liang KY, Hetmanski JB, Ruczinski I, Fallin MD, Ingersoll RG, Wang H, Huang S, Ye X, Wu-Chou Y-H, Chen PK, Jabs EW, Shi B, Redett R, Scott AF, Beaty TH. Evidence of gene-environment interaction for the IRF6 gene and maternal multivitamin supplementation in controlling the risk of cleft lip with/without cleft palate. Hum Genet. 2010; 128:401–410. [PubMed: 20652317]

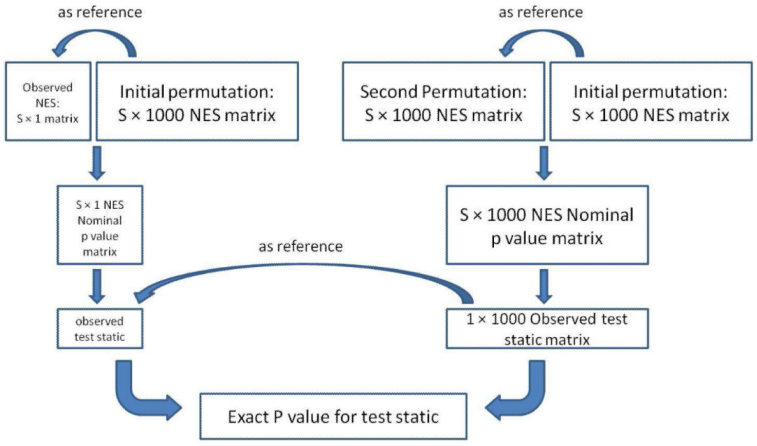

Figure 1.

The permutation procedures used in the analysis. Initially, 1000 permutations were carried out and served as reference group to calculate a nominal p-value for each of the $S = 42$ gene sets (left). A test statistic based on the distribution of these nominal p-values, similar to the one in Fisher's inverse χ^2 test, was used to assess overall signal in the data. Since gene sets can share some of the same genes, the p-values are not independent, and a second layer of permutations (1000 iterations) was carried out to delineate the null distribution of the test statistic, taking the dependency between pathways into account (right). For each of these iterations in the "second layer" of permutations, the pathway p-values (that form the basis of the respective test statistic) were derived using the same null distributions as in the observed data (indicated by the "as reference" arrow in the upper right).

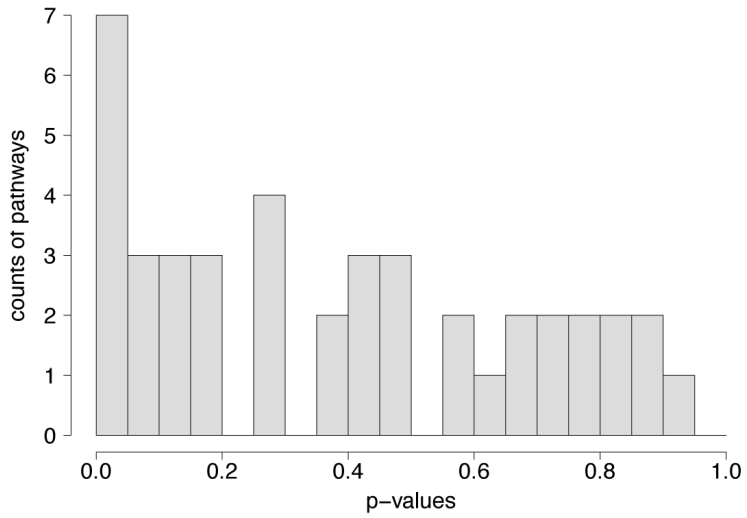

Figure 2.
Histogram of the p-values for the “pathway significance” (Table 3). While none of the p-values withstands a Bonferroni correction or family-wise error control via permutation tests, there appears to be an enrichment among pathways showing low p-values, compared to a Uniform (0,1) distribution expected under the null.

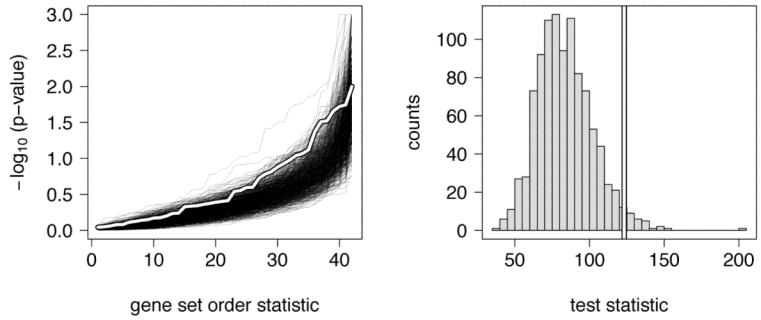

Figure 3.

[Left panel] The p-values (shown on the $-\log_{10}$ scale) from the permutation test. Each thin dark line represents one permutation. The p-values for the 42 gene sets, derived in each permutation, are ordered and shown from least significant (left) to most significant (right). The thicker white line shows the $-\log_{10}$ p-values for the observed data (Table 3). While the most significant nominal p-value can be reasonably explained by chance, there appears to an enrichment of low p-values in the distribution of the observed data, even compared to the permuted scores. This is quantified in the [Right panel]. Each of the 1000 permutation is summarized by calculating a combined test statistic from the 42 permuted gene-set p-values, and the distribution of these combined test statistics serves as empirical null distribution for the observed test statsitic, shown as a vertical bar. Only 29 of the permuted test statistics exceed the observed test statistic, yielding a p-value of 2.9%.

Table 1

The 20 most significant GWAS findings based on the allelic transmission disequilibrium test. Shown are chromosome and SNP "RS" number, gene or region name, and its respective position. Also listed are the number of times the major allele was transmitted (T) or not transmitted (U) from a heterozygous parent to the offspring. The respective odds ratios, test statistics, and p-values (based on an asymptotic χ^2_1 distribution) are also shown. The complete findings are listed in the supplementary material of Beaty et al. (2010).

Chr	SNP	Gene/region	Position	T	U	OR	χ^2	p
8	RS987525	hCG 1814486*	130015336	554	311	1.78	68.3	1.4E-16
11	RS1784394**	JAM3	133468426	84	228	0.37	66.5	3.6E-16
1	RS10863790	IRF6	208054670	312	538	0.58	60.1	9.1E-15
8	RS1519847	hCG 1814486	129984942	670	430	1.56	52.4	4.6E-13
8	RS882083	hCG 1814486	130051938	616	393	1.57	49.3	2.2E-12
8	RS12542837	hCG 1814486	129995843	668	435	1.54	49.2	2.3E-12
1	RS560426	ABCA4	94326026	885	618	1.43	47.4	5.7E-12
8	RS1519841	hCG 1814486	129988982	668	440	1.52	46.9	7.4E-12
8	RS12548036	hCG 1814486	130017064	628	408	1.54	46.7	8.2E-12
20	RS13041247	MAFB	38702488	620	881	0.70	45.4	1.6E-11
20	RS11696257	MAFB	38704230	621	881	0.71	45.0	2.0E-11
8	RS1530300	hCG 1814486	129988640	530	333	1.59	45.0	2.0E-11
1	RS2013162	IRF6	208035307	620	879	0.71	44.8	2.2E-11
20	RS17820943	MAFB	38701930	620	877	0.71	44.1	3.1E-11
1	RS2073485	IRF6	208029417	513	745	0.69	42.8	6.1E-11
8	RS1850889	hCG 1814486	129959587	288	465	0.62	41.6	1.1E-10
8	RS1519850	hCG 1814486	129966003	251	415	0.61	40.4	2.1E-10
8	RS7017252	hCG 1814486	130020026	560	367	1.53	40.2	2.3E-10
8	RS11787407	hCG 1814486	130054622	604	404	1.50	39.7	3.0E-10
8	RS1470206	hCG 1814486	130046646	661	452	1.46	39.3	3.7E-10

*gene desert region of 8q24;

**SNP showed poor clustering even though it was not flagged in quality control process.

Table 2

The number of complete CL/P trios used in the data analysis, split by ancestry and gender of the proband. A few families of European descent had more than one affected offspring, which resulted in more than one complete trio for those families. Since the association test assesses the departure from independent assortment, the respective allele transmission and contributions to the test statistic are independent.

Ancestry	Proband	CL/P trios	Total	Families
European	Male	431		
	Female	250	681	668
Asian	Male	582		
	Female	313	895	895
Other	Male	16		
	Female	12	28	28

Table 3

Results of the gene set analysis, showing the pathway names, number of genes per pathway (N), the enrichment score (ES), the normalized enrichment score (NES), the nominal p-value (P), the respective q-value (Q), and the FWER corrected p-value derived from the permutation test.

	Pathway	N	ES	NES	P	Q	P_{FWER}
1	Cytoskeletal regulation by Rho GTPase	82	0.39	2.51	0.010	0.096	0.212
2	Insulin/IGF-protein kinase B signaling	28	0.54	2.14	0.022	0.096	0.423
3	Angiogenesis	173	0.43	2.10	0.018	0.096	0.453
4	Huntington disease	144	0.39	2.05	0.019	0.096	0.484
5	Cadherin signaling	131	0.49	1.94	0.031	0.096	0.552
6	5-Hydroxytryptamine degredation	17	0.51	1.43	0.043	0.114	0.886
7	Endothelin signaling	83	0.48	1.42	0.076	0.169	0.887
8	VEGF signaling	67	0.45	1.42	0.086	0.169	0.887
9	PDGF signaling	134	0.41	1.37	0.091	0.169	0.915
10	p53 feedback loops 2	45	0.40	1.30	0.105	0.178	0.940
11	TCA cycle	15	0.48	1.12	0.030	0.096	0.983
12	De novo purine biosynthesis	29	0.40	1.01	0.118	0.183	0.992
13	P53	98	0.33	0.96	0.165	0.205	0.994
14	Phenylethylamine degradation	9	0.57	0.96	0.132	0.189	0.994
15	Hypoxia response via HIF activation	29	0.43	0.93	0.153	0.203	0.994
16	Hedgehog signaling	23	0.43	0.84	0.189	0.220	0.998
17	Alzheimer disease-presenilin	111	0.38	0.68	0.251	0.261	1.000
18	Nicotinic acetylcholine receptor signaling	87	0.37	0.58	0.279	0.268	1.000
19	Ras	66	0.36	0.53	0.288	0.268	1.000
20	Plasminogen activating cascade	18	0.35	0.52	0.253	0.261	1.000
21	Androgen/estrogene/progesterone biosynth.	16	0.37	0.29	0.386	0.312	1.000
22	JAK/STAT signaling	19	0.33	0.24	0.436	0.312	1.000
23	EGF receptor signaling	119	0.34	0.21	0.395	0.312	1.000
24	2-arachidonoylglycerol biosynthesis	7	0.49	0.17	0.422	0.312	1.000
25	Apoptosis signaling	113	0.28	0.16	0.407	0.312	1.000
26	Formyltetrahydroformate biosynthesis	8	0.44	0.16	0.469	0.312	1.000
27	Metabotropic glutamate receptor group I	31	0.53	0.12	0.454	0.312	1.000
28	GABA-B receptor II signaling	35	0.41	0.07	0.461	0.312	1.000

	Pathway	N	ES	NES	P	Q	P_{FWER}
29	Interferon-gamma signaling	27	0.30	−0.17	0.583	0.361	1.000
30	Integrin signalling	165	0.34	−0.23	0.576	0.361	1.000
31	Vitamin D metabolism and	14	0.35	−0.27	0.674	0.380	1.000
32	Gamma-aminobutyric acid synthesis	6	0.41	−0.28	0.710	0.385	1.000
33	Interleukin signaling	129	0.30	−0.33	0.634	0.380	1.000
34	T cell activation	91	0.36	−0.45	0.667	0.380	1.000
35	Toll receptor signaling	53	0.26	−0.67	0.724	0.385	1.000
36	Blood coagulation	45	0.27	−0.67	0.751	0.388	1.000
37	Notch signaling	42	0.28	−0.77	0.771	0.388	1.000
38	B cell activation	68	0.36	−0.93	0.830	0.397	1.000
39	Heterotrimeric G-protein signaling	145	0.35	−0.94	0.832	0.397	1.000
40	PI3 kinase	92	0.31	−1.10	0.867	0.401	1.000
41	FGF signaling	107	0.33	−1.21	0.891	0.401	1.000
42	TGF-beta signaling	124	0.29	−1.28	0.906	0.401	1.000