Published in final edited form as:
Appl Evol Comput (2014). 2014 April ; 8602: 890–901. doi:10.1007/978-3-662-45523-4_72.

Genotype Correlation Analysis Reveals Pathway-Based Functional Disequilibrium and Potential Epistasis in the Human Interactome

William S. Bush[1] and **Jonathan L. Haines**[2]

[1]Center for Human Genetics Research, Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

[2]Institute for Computational Biology, Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, USA

Abstract

Epistasis is thought to be a pervasive part of complex phenotypes due to the dynamics and complexity of biological systems, and a further understanding of epistasis in the context of biological pathways may provide insight into the etiology of complex disease. In this study, we use genotype data from the International HapMap Project to characterize the functional dependencies between alleles in the human interactome as defined by KEGG pathways. We performed chi-square tests to identify non-independence between functionally-related SNP pairs within parental Caucasian and Yoruba samples. We further refine this list by testing for skewed transmission of pseudo-haplotypes to offspring using a haplotype-based TDT test. From these analyses, we identify pathways enriched for functional disequilibrium, and a set of 863 SNP pairs (representing 453 gene pairs) showing consistent non-independence and transmission distortion. These results represent gene pairs with strong evidence of epistasis within the context of a biological function.

1 Introduction

In 1912, William Bateson first coined the term epistasis, (from the Greek for standing upon) when he observed an allele at one locus masking the effect of an allele at a second, independent locus [1]. Bateson's concept has also been described as biological epistasis, similar to a biochemist's observation that variation in the physical interaction of biomolecules affects a phenotype [2, 3]. Several years later, R.A. Fisher also used the term epistasis in a statistical context, observing multi-allelic segregation patterns that can be mathematically described as a deviation from additivity in a linear model of genotypes [4]. Given the complexities of known biological pathways that involve numerous inter-molecular interactions, epistasis is presumed to be ubiquitous both statistically and biologically [3]. This belief is driven largely by the notion that networks of gene regulation and protein-protein interaction have a functional endpoint that may be influenced by the simultaneous presence of multiple variants in those genes [3, 5]. Epistasis has been well-documented in model organisms, and was discovered early in the field of genetics. In 1918, Lancefield described a two-locus inheritance pattern for the forked bristle phenotype in Drosophila [6]. A year later, Bridges reported statistical epistasis in Drosophila eye color, where combinations of several different alleles Mendelize with various eye color phenotypes [7].

These alleles influence a biochemical pathway controlling eye pigmentation that was described many years later [8]. More recently, studies of mouse and rat chromosome substitution strains revealed substantial epistasis in over 140 quantitative trait loci [9]. But outside the exploration of these model systems, the concept of epistasis was largely ignored in the field of human genetics. Over the last fifteen years, however, the concept has resurged as the study of common complex human phenotypes has become more prominent.

Epistasis is an attractive concept for complex traits because techniques used to characterize strong single-gene effects (such as linkage analysis) typically fail to consistently identify genomic regions that explain variation in complex traits. Twin studies and family-based segregation analysis establish heritable genetic components to these traits, yet the source of genetic trait variation often remains unknown. One potential source of the unexplained heritability is that a larger proportion of trait variation is due to epistasis – combinations of genotypes at multiple loci -- rather than single independent loci [10]. Epistasis also fits well with the general notion that complex traits have complex underlying genetic etiologies.

Statistically, the concept of epistasis analysis is very similar in theory to haplotype analysis. Genetically, a haplotype occurs when loci in close physical proximity are linked by a stretch of chromosome and are thus often inherited together. When this occurs in a large population, these loci are said to be in *linkage disequilibrium*, and the alleles of these loci form haplotypes. Because these linked alleles have a high likelihood of being inherited together in the population, the genotypes of these loci are correlated, or alternatively their genotypes are non-independent.

It is also possible that there is correlation between genotypes of loci that are not physically linked on the chromosome. This phenomenon is sometimes referred to as *gametic phase disequilibrium*, as the alleles non-randomly segregate within gametes, but are not physically tethered on the chromosomes [11]. Even though alleles are not linked physically, they may still be linked on some higher biological level that causes the occurrence of the genotypes to be non-independent in the population, presumably by some function that confers a change in evolutionary fitness. We loosely define this phenomenon as functional disequilibrium, and the alleles of these functionally linked loci form a functional psuedo-haplotype.

The work of the International HapMap Project has characterized patterns of linkage disequilibrium among common SNPs in multiple human sub-populations. These patterns are useful for gene mapping studies to determine which portions chromosome (and marker loci) are typically co-inherited within a population, and thus reducing the number of genetic markers needed to effectively capture common variation in the genome. Also, the patterns of linkage disequilibrium established for a population identify haplotypes that can be tested for association with disease phenotypes or other traits. From a broader perspective, the HapMap provides an overview of the structural interdependencies of the human genome, which has given insight into various basic human genetics questions regarding recombination rates [12], segregation distortion [13], genomic regions of selection [14], and even mate choice [15].

Similarly, patterns of functional disequilibrium may exist in human populations that encapsulate common genetic variation into functional (rather than structural) units. These patterns may provide insight into previously unknown interdependencies in biochemical pathways, such as gene expression patterns that detrimentally or beneficially alter pathway kinetics or function. Characterizing functional disequilibrium also builds a better understanding of the general genetic variation in the interactome, and could lead to a new understanding of the biochemistry of these systems.

Functional disequilibrium should also have consequences for disease etiology. Biological pathways likely have distinct genetic architectures that influence overall function, and some genetic architectures may alter susceptibility to disease. Also, alterations in pathway function may influence how environmental exposures are processed, leading to increased or decreased risk of disease upon exposure, such as with nicotine metabolism and lung cancer [16].

As such, a catalog of pathway-based pseudo-haplotypes would be an excellent resource for conducting candidate epistasis studies using genome-wide association data. With these goals in mind, in this work we investigate the presence of functional disequilibrium, observed as correlated genotypes in non-linked SNPs, among a set of core biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

2 Methods

2.1 Data

For this study, we used publicly available Single Nucleotide Polymorphisms (SNPs) from the Hapmap Phase III dataset. 1,403,896 SNPs genotyped in 57 trios from Utah (Centre d'Etude du Polymorphisme Humain (CEPH) Collection) and 1,484,416 SNPs genotyped in 54 trios from the Yoruba population of Ibadan, Nigeria.

2.2 Domain Knowledge

The Kyoto Encyclopedia of Genes and Genomes [17-19] [accessed 4/27/2009] contains 203 metabolic and regulatory pathways. 183 of these pathways, containing mappings to human genes and of manageable size, were used as gene groups encompassing 4,826 unique genes. Entrez-gene IDs from the KEGG database were mapped to Ensembl gene IDs using the Ensembl database [20]. From these gene groups, 2,096,620 unique gene pairs were constructed by forming all possible pairs of genes within each gene group. Using the Ensembl Variation database, SNPs residing within the Ensembl gene physical (base-pair) start and end were mapped. SNP pairs were created by forming all possible combinations of two SNPs across the two genes. Pairs of SNPs that fall within the same gene, or within 500 KB of each other on the same chromosome were excluded from this analysis as genotypes of these SNPs may be non-independent due to linkage disequilibrium. Two-SNP models were generated using the Biofilter procedure outlined in [21].

2.3 Statistical Analysis

The non-independence of genotypes for each SNP pair was assessed within each dataset using a chi-square test of independence. The chi-square test compares the observed frequency of a genotype combination to the frequency expected if the genotypes are independent. Analysis was conducted using an internally developed C++ program incorporated into the Biofilter framework. Internal software was validated with STATA 10.1.

SNP pairs with genotypes that are non-independent were further analyzed. SNP pairs with a minor allele frequency < 0.10 were excluded from further analysis. We did not filter SNPs based on Hardy-Weinberg Equilibrium tests because unviable or lethal combinations of SNPs could appear out of Hardy-Weinberg Equilibrium if analyzed alone. For the remaining SNP pairs, r^2 correlation coefficients were computed using PLINK software [22, 23]. Using the haplotype transmission disequilibrium test implemented in PLINK, the co-transmission of SNP pairs within CEU and YRI trios was assessed. This test uses a chi-square statistic to measure multi-locus segregation distortion. In this application, the test determines if pathway-based pseudo-haplotypes observed in the parent generation are significantly over- or under-transmitted to offspring in the population, based on the parental haplotype frequencies.

3 Results

3.1 Analysis Overview

To investigate the presence of functional disequilibrium in the human genome, we used a bioinformatics approach to group genes together by functional relationships. 183 pathways from the KEGG database were used to group genes by function, and these gene groups were used to construct SNP pairs that exclude haplotype effects (the SNPs must be > 500 KB apart). Pathway-based SNP pairs were evaluated in the HapMap phase III dataset for Yoruba (YRI) and Caucasian (CEU) populations.

As an initial screen, unrelated individuals (parents) were extracted from the YRI (n=108) and CEU (n=114) datasets and a chi-square test of independence was conducted to assess the correlation between the genotypes of each pathway-based SNP pair. SNP pairs with chi-square statistics > 9.487 ($\alpha = 0.05$, df = 4) were carried forward to the next phase of analysis. To provide additional evidence of functional disequilibrium between the SNP pairs identified in the screen, we conducted a transmission disequilibrium test (TDT) to determine if there was non-independent transmission of pseudo-haplotypes (pathway-based genotype combinations) to offspring in the sample. Because we are testing transmission of the pseudo-haplotype, this test is independent of the chi-square test used in the initial screen.

Using these analyses, we present pathways potentially enriched for non-independent genotypes in both populations, pathway-based pseudo-haplotypes that show distorted transmission, and an overall collection of gene pairs showing evidence of functional disequilibrium.

3.2 Initial Screen

In the initial screen phase, we evaluated roughly 428 million CEU SNP pairs and 479 million YRI SNP pairs generated from gene combinations found in KEGG pathways. The overall significance rate for the screen was 0.0284 for CEU and 0.0303 for YRI. Both the peptidoglycan biosynthesis (CEU 0.25, YRI 0.15) and atrazine degradation (CEU 0.16, YRI 0.04) pathways had high proportions of significant results, however these two pathways contained relatively few SNP pairs (903 and 2437 respectively). Nearly all of the pathways with high proportions of significant results in the screen were metabolic rather than regulatory pathways. In fact, several large regulatory pathway groups, such as “Pathways in cancer” (CEU 0.0045, YRI 0.0053), axon guidance (CEU 0.0116, YRI 0.0148), tight junction (CEU 0.0145, YRI 0.0186), and focal adhesion (CEU 0.011, YRI 0.0063) had a very low proportion of significant results.

In this screening phase of the analysis, we used a liberal significance threshold ($\alpha = 0.05$). Corrections for multiple hypothesis testing in this setting are difficult due to the correlation between tests; we therefore rely on a two-phase design where results from the initial screen are validated using an independent approach.

3.3 Confirmation

We exploit a unique property of genetic data to conduct a confirmatory analysis; based on Mendel's law of independent assortment, the transmission of two alleles at unlinked loci should be independent. If the potential functional SNP pairs discovered in our screening analysis are transmitted together more or less often than expected by chance, this could further indicate a functional relationship between the loci. Using the full set of 57 CEU trios and 54 YRI trios, we assessed transmission distortion using the haplotype-based TDT for all significant SNP pairs identified in the screening phase. Of the 40,312,276 tests conducted, the TDT identified 1,698,521 (4.21%) significantly distorted haplotype transmissions in CEU. For the YRI samples, 50,175,211 of 2,187,530 (4.36%) tests were significant. The proportion of significant tests by pathway is shown in figure 2.

3.4 Gene-Gene Pairings with Putative Epistasis

From the results of our genotypic non-independence and pseudo-haplotype transmission tests, we compiled a list of SNP-SNP and subsequent gene-gene pairs that indicate putative epistasis. These SNP-SNP pairs had correlated genotypes and significant pseudo-haplotype TDT statistics in both CEU and YRI samples. The most compelling results are SNP pairs that were correlated in both samples, and also whose haplotypes were identical and similarly distorted in the TDT statistics. 863 of these cases were detected. Of these, 763 SNP pairs contained two intronic SNPs, 98 SNP pairs contained only one intronic SNP (others were coding, within a splice site, or within the 3′ or 5′ UTR), and only 2 SNP pairs contained two non-intronic SNPs. The two non-intronic SNP pairs are shown in table 1.

The distribution of gene pairs exhibiting putative epistasis by pathway is shown in figure 4. A database of all significant results from the confirmation phase of this study is also available upon request.

4 Discussion

In this work, we illustrate how a bioinformatics analysis of population-based genetic data can reveal allelic dependencies between genes of biochemical pathways. Just as the physical structure of the chromosome gives rise to correlations among genotypes called linkage disequilibrium, the structure of biochemical systems can likewise give rise to correlations among genotypes that presumably alter offspring viability or evolutionary fitness in some way, a phenomenon we loosely phrase functional disequilibrium. Gene pairs that contain SNPs exhibiting functional disequilibrium are potentially indicative of epistasis in relation to some phenotype.

The results of the initial screen seem to indicate that a higher degree of functional disequilibrium is present in more purely metabolic pathways. Despite this observation, the strongest and most consistent examples of functional disequilibrium occur mostly in regulatory and signaling pathways. Interestingly, pathways with high numbers of implicated gene pairs are heavily involved in nervous signal transduction, such as tight junction, chemokine signaling, and Wnt signaling and general nerve cell function, such as focal adhesion, axon guidance, and regulation of actin cytoskeleton. Several neurological phenotype pathways are well represented in this respect also, such as Alzheimer's disease, Parkinson's disease, and Huntington's disease. Genotypic dependencies among the elements of these disease related pathways should be further investigated, and may lead to new insights into population level risk for these conditions, and for general neurological development.

A specific compelling example from this study is the functional disequilibrium between rs1053454, a SNP located in the 3′ untranslated region of the 1-phosphatidylinositol-5-phosphate 4-kinase type II alpha gene (*PIP4K2A*) and rs749338, a synonymous SNP in the inositol 1,4-5-triphosphate receptor type 3 gene (*ITPR3).* These genes function in the phosphatidyinositol signaling pathway (KO:04070), a signal transduction mechanism involved in multiple physiological functions, including neurotransmitter release and other aspects of the nervous system. These two SNPs have non-independent genotypes in CEU and YRI unrelated individuals (CEU p = 0.0366, YRI p = 0.0323), and the “CT” pseudo-haplotype of these SNPs is significantly and consistently over-transmitted to offspring in both CEU and YRI samples (CEU hap-TDT = 0.026, YRI hap-TDT = 0.026).

Figure 5 illustrates the biochemical relationships between these two genes in phosphatidylinositol signaling pathway. PIP4K2A converts 1-Phosphatidyl-1D-myo-inositol 5-phosphate to 1-Phosphatidyl-D-myo-inositol 4,5-bisphosphate, which is then converted to Inositol 1,4,5-trisphosphate (IP3) by phospholipase C enzymes (PLC). IP3 then binds to the IP3 receptor (IP3R) to activate downstream calcium release. Phosphatidylinositol signaling has been implicated in neuronal function and development[24].

There are several important limitations to this work. There are numerous pathway databases that could be used for this type of analysis. We chose the KEGG database because it is a well-established and supported collection of biochemical and regulatory pathways. Other sources of functional information that relate genes could be used as well, and will be explored in future research. We elected to use the phase III Hapmap data only because this data is the most recent large scale collection of genotypes from multiple ethnicities. Using the full collection of Hapmap SNPs was logistically and computationally prohibitive for this work, but is also an area of future research.

The chi-square test of independence is not appropriate for contingency tables with fewer than 5 observations per cell -- a Fisher's exact test should be used in these cases. The computational complexity of a 3×3 Fisher's exact test calculation precluded us from conducting that calculation in these experiments, and instead we filtered the significant results from the chi-square test by minor allele frequency to limit this bias. The haplotype transmission disequilibrium test implemented in PLINK software was intended for true haplotypes of SNPs in linkage disequilibrium on the same chromosome, and performs an expectation maximization (EM) procedure to estimate the chromosomal phase of the haplotypes. When performing the EM procedure on genotypes across chromosomes, the phased haplotype distribution should very closely match the observed multi-locus genotype distribution, and when compared for randomly selected example SNPs they match well. It notable, however, that we are employing this test outside its original design, and the phasing procedure may slightly alter the distribution of transmitted and untransmitted pseudo-haplotypes. Furthermore, it is extremely difficult to assess the false positive rate for this study. Linkage disequilibrium, for example among 10 SNPs of gene 1 and 7 SNPs of gene 2, causes correlations between the tests statistics of all SNP combinations spanning gene1 and gene2.

Finally, for simplicity, we are using the Ensembl definition of a gene region (3′ to 5′ untranslated region), which does not include upstream or downstream regulatory elements. It is likely that these regulatory elements also contain variants that in combination alter pathway function. These combinations of variants would not be detected in this analysis due to our myopic gene definition.

This work is an initial first step in cataloging correlated collections of functionally related genetic variations in multiple human populations. Future directions include expanding the datasets to include all 11 populations in the Hapmap data, expanding the bioinformatics stores to include protein-protein interaction databases and protein family information, and further refining the statistical analysis of non-independence by conducting multi-locus Hardy-Weinberg Equilibrium tests. Correlated pairs of genetic variants could further be annotated to include evolutionary conservation information, potential gene-based function (such as presence in or near a regulatory sites), and local linkage disequilibrium data. Stored in a public database system, these results could provide insight into new biochemical or regulatory mechanisms, and would provide a set of potential ethnic specific differences in pathway dynamics and function.

Acknowledgments

This work was supported by National Institutes of Health grants AG19085, AG27944, NS32830, and NS49477.

References

1. Bateson, W. Mendel's Principles of Heredity. Cambridge University Press; Cambridge: 1909.
2. Moore JH, Williams SM. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays. 2005; 27:637. [PubMed: 15892116]
3. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. Hum Hered. 2003; 56:73. [PubMed: 14614241]
4. Fisher RA. Transactions of the Royal Society of Edinburgh. 1918; 52:399.
5. Moore JH, Williams SM. Bioessays. 2005; 27:637.
6. Lancefield DE. An autosomal bristle modifier affecting a sex-linked character. American Naturalist. 1918; 52:462.
7. Bridges CB. Specific modifiers of eosin eye color in Drosophila melanogaster. J Experimental Zoology. 1919; 28:337.
8. Lloyd V, Ramaswami M, Kramer H. Not just pretty eyes: Drosophila eye-colour mutations and lysosomal delivery. Trends Cell Biol. 1998; 8:257. [PubMed: 9714595]
9. Shao H, Burrage LC, Sinasac DS, Hill AE, Ernest SR, O'Brien W, Courtland HW, Jepsen KJ, Kirby A, Kulbokas EJ, Daly MJ, Broman KW, Lander ES, Nadeau JH. Genetic architechture of complex traits: large phenotypic effects and pervasive epistasis. Proc Natl Acad Sci USA. 2008; 105(50): 19910–4. [PubMed: 19066216]
10. Cordell HJ. Detecting gene-gene interactions that underlie dieseases. Nat Rev Genet. 2009
11. Wang X, Elston RC, Zhu X. The meaning of interaction. Human Heredity. 2010; 70:269. [PubMed: 21150212]
12. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851. [PubMed: 17943122]
13. Zollner S, Wen X, Hanchard NA, Herbert MA, Ober C, Pritchard JK. Evidence for extensive transmission distortion in the human genome. Am J Hum Genet. 2004; 74:62. [PubMed: 14681832]
14. Sabeti PC, Varily P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, Schaffner SF, Lander ES. the International HapMap Consortium, Genome-wide detection and characterization of positive selection in human populations. Nature. 2007; 449:913. [PubMed: 17943131]
15. Chaix R, Cao C, Donnelly P. Is mate choice in human MHC-depedent? PLoS Genet. 2008; 4:e1000184. [PubMed: 18787687]
16. Derby KS, Cuthrell K, Caberto C, Carmella SG, Franke AA, Hecht SS, Murphy SE, Le Marchand L. Nicotine metabolism in three ethnic/racial groups with different risks of lung cancer. Cancer Epidemiol Biomarkers Prev. 2008; 17:3526. [PubMed: 19029401]
17. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000; 28:27. [PubMed: 10592173]
18. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006; 34:D354. [PubMed: 16381885]
19. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. KEGG for linking genomes to life and the envrionment. Nucleic Acids Res. 2008; 36:D480. [PubMed: 18077471]
20. Flicek P, et al. Ensembl 2008. Nucleic Acids Res. 2008; 36:D707. [PubMed: 18000006]
21. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput. 2009:368. [PubMed: 19209715]
22. Purcell S. PLINK 1.01 Ref Type: Computer Program.
23. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferrieira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559. [PubMed: 17701901]

24. Kim D, Jun KS, Lee SB, Kang N, Min DS, Kim Y, Ryu SH, Suh P, Shin H. Phospholipase C isozymes selectively couple to specific neurotransmitter receptors. Nature. 1997; 389:290. [PubMed: 9305844]

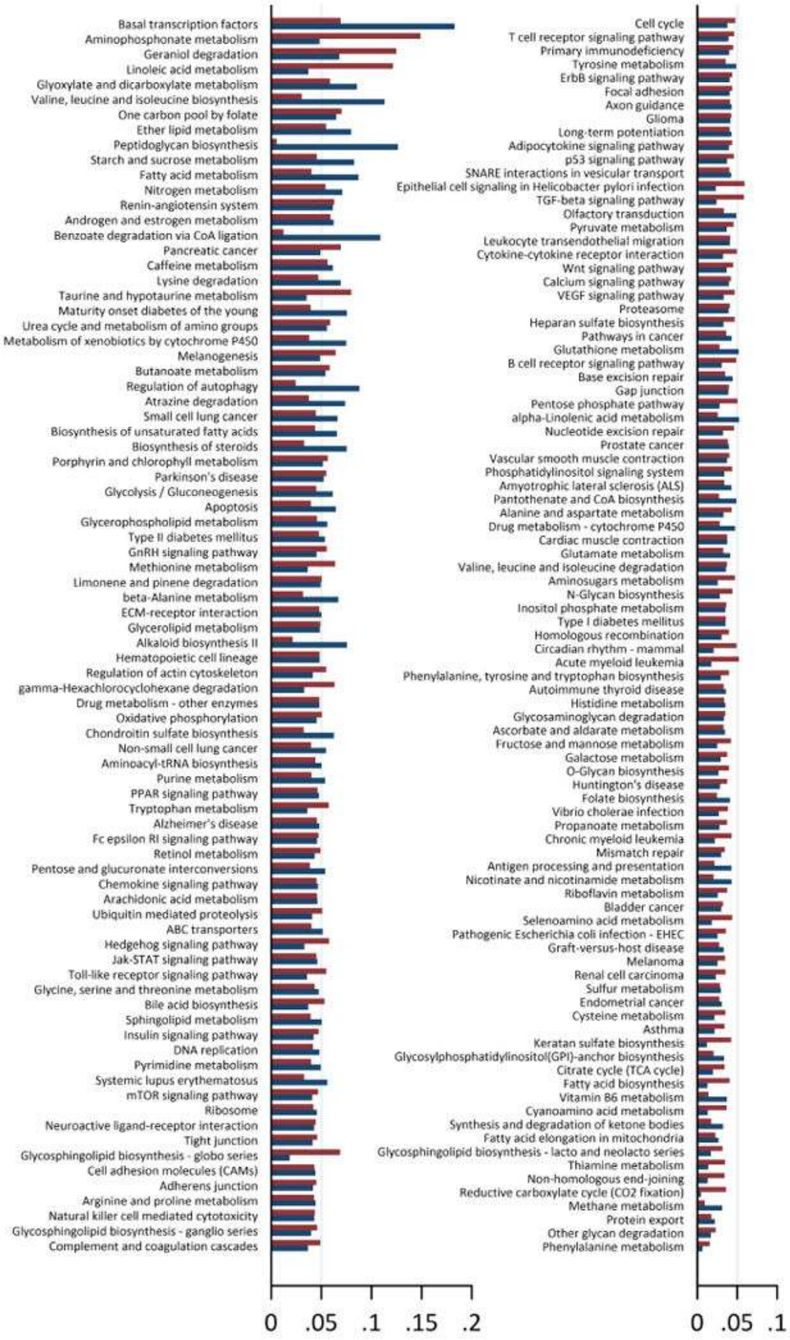

Figure 2. Distributions of Significant Haplotype TDT. YRI in red, CEU in blue

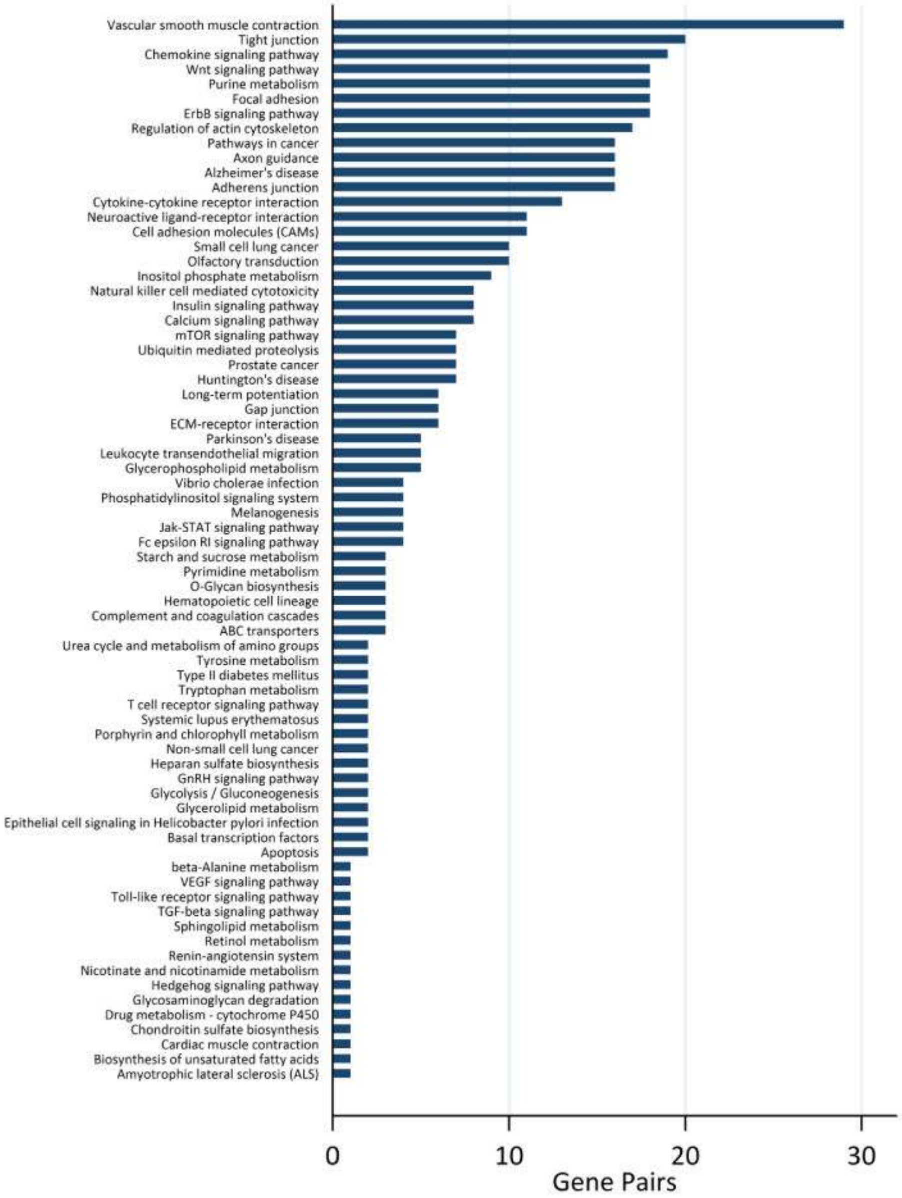

Figure 4. Distribution of gene-pairs exhibiting strong evidence of epistasis across both CEU and YRI populations, listed by biological pathway

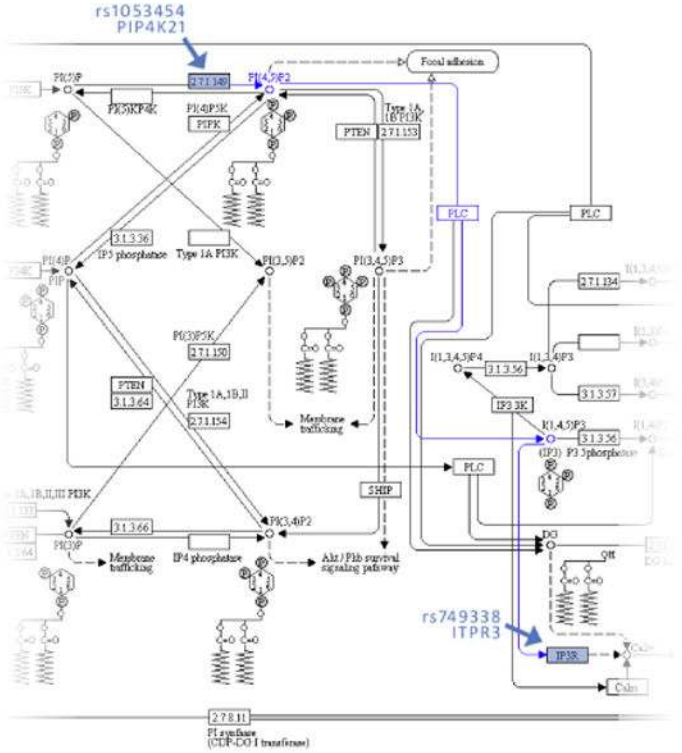

Figure 5. Putative epistasis in the Phosphatidylinositol Signaling pathway (Adapted from KEGG KO:004070)

Table 1

Two non-intronic SNP pairs showing strong evidence of probable epistasis within biological pathways (pathway 1: Phospatidylinositol signaling, Pathway 2: Olfactory transduction)

SNP Pair	Gene Pair	SNP Type	CEU Freq	YRI Freq	CEU X2	YRI X2	Haplotype	CEU TDT	YRI TDT	Path
rs1053454 rs749338	PIP4K2A ITPR3	3′ SYN	0.41 0.44	0.12 0.10	0.036	0.032	C T	0.026	0.026	1
rs2900373 rs6679056	OR13C9 OR10R2	NON NON	0.19 0.41	0.33 0.50	0.014	0.044	A A	0.029	0.029	2