2014

Pathway Analysis Approaches for Rare and Common Variants: Insights From Genetic Analysis Workshop 18

Stella Aslibekyan

Marcio Almeida

Nathan L. Tintle
Dordt College, nathan.tintle@dordt.edu

Pathway Analysis Approaches for Rare and Common Variants: Insights From Genetic Analysis Workshop 18

Abstract

Pathway analysis, broadly defined as a group of methods incorporating a priori biological information from public databases, has emerged as a promising approach for analyzing high-dimensional genomic data. As part of Genetic Analysis Workshop 18, seven research groups applied pathway analysis techniques to whole-genome sequence data from the San Antonio Family Study. Overall, the groups found that the potential of pathway analysis to improve detection of causal variants by lowering the multiple-testing burden and incorporating biologic insight remains largely unrealized. Specifically, there is a lack of best practices at each stage of the pathway approach: annotation, analysis, interpretation, and follow-up. Annotation of genetic variants is inconsistent across databases, incomplete, and biased toward known genes. At the analysis stage insufficient statistical power remains a major challenge. Analyses combining rare and common variants may have an inflated type I error rate and may not improve detection of causal genes. Inclusion of known causal genes may not improve statistical power, although the fraction of explained phenotypic variance may be a more appropriate metric. Interpretation of findings is further complicated by evidence in support of interactions between pathways and by the lack of consensus on how to best incorporate functional information. Finally, all presented approaches warranted follow-up studies, both to reduce the likelihood of false-positive findings and to identify specific causal variants within a given pathway. Despite the initial promise of pathway analysis for modeling biological complexity of disease phenotypes, many methodological challenges currently remain to be addressed.

Keywords

way analysis, whole-genome sequence, hypertension, family studies, Genetic Analysis Workshop 18

Disciplines

Genetics and Genomics | Medicine and Health Sciences | Statistics and Probability

Comments

This is a pre-publication author manuscript of the final, published article. The definitive version is published by Wiley and available from Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21831

Title: Pathway analysis approaches for rare and common variants: Insights from GAW18

Authors: Stella Aslibekyan[1], Marcio Almeida[2] and Nathan Tintle[3]

[1]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, Alabama

[2]Department of Genetics-Texas Biomedical Research Institute, San Antonio, TX

[3]Department of Mathematics, Statistics and Computer Science, Dordt College, Sioux Center, IA

Running Title: Pathway analysis approaches at GAW18

Corresponding Author

Dr. Nathan Tintle

Department of Mathematics, Statistics and Computer Science

Dordt College

498 4th Ave. NE

Sioux Center, IA 51250

United States

Phone: +011 (712) 722-6264

Email: nathan.tintle@dordt.edu

Abstract

Pathway analysis, broadly defined as a group of methods incorporating *a priori* biological information from public databases, has emerged as a promising approach for analyzing high-dimensional genomic data. As part of Genetic Analysis Workshop 18 (GAW18), seven research groups applied pathway analysis techniques to whole genome sequence data from the San Antonio Family Study. Overall, the groups found that the potential of pathway analysis to improve detection of causal variants by lowering the multiple testing burden and incorporating biologic insight remains largely unrealized. Specifically, there is a lack of best practices at each stage of the pathway approach: annotation, analysis, interpretation, and follow-up. Annotation of genetic variants is inconsistent across databases, incomplete, and biased towards known genes. At the analysis stage, insufficient statistical power remains a major challenge. Analyses combining rare and common variants may have an inflated type 1 error rate and may not improve detection of causal genes. Inclusion of known causal genes may not improve statistical power, although the fraction of explained phenotypic variance may be a more appropriate metric. Interpretation of findings is further complicated by evidence in support of interactions between pathways as well as the lack of consensus on how to best incorporate functional information. Finally, all presented approaches warranted follow-up studies, both to reduce the likelihood of false positive findings and to identify specific causal variants within a given pathway. Despite the initial promise of pathway analysis for modeling biological complexity of disease phenotypes, many methodological challenges currently remain to be addressed.

Key words: pathway analysis, whole genome sequence, hypertension, family studies

Introduction

The advent of next generation sequencing, has produced a wealth of high-resolution genomic data at an unprecedented scale. These technologies are enabling novel, comprehensive investigations of disease phenotypes, but also creating new challenges for analysis and interpretation. Namely, despite the data explosion, we still have relatively low power to find genetic associations. Therefore, the central challenge lies in how to best use statistical relationships to infer biological mechanisms from detailed sequence information. Pathway analysis, broadly defined as a group of statistical methods that exploit *a priori* knowledge of pathways (broadly defined as sets of genes with a known biological relationship) stored in public databases such as KEGG (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012), Gene ontology (T. G. O. Consortium, 2000), Reactome (Croft et al., 2010), and others, offers a naturally attractive approach to modeling biological complexity and improving detection of statistical associations (Khatri, Sirota, & Butte, 2012). More specifically, pathway analysis methods use a variety of different strategies to aggregate or interpret individual marker or gene based phenotype association statistics to yield a single interpretable test statistic (or p-value) summarizing the strength of evidence of association between the pathway and the phenotype.

Initially based in the context of gene expression arrays, modern pathway analysis methods have recently been extended to next-generation sequence data, including structural variants and rare genetic polymorphisms (Hu, Xu, Cheng, Xing, & Paterson, 2011; Petersen et al., 2011; Tintle et al., 2011; Yang & Gu, 2011). The range of analytical methods that fall under the pathway analysis definition is rapidly gaining traction among biomedical researchers, evidenced by a more than tenfold rise in PubMed citations since the completion of the human genome sequence in 2003 (Ramanan, Shen, Moore, & Saykin, 2012). This rise in popularity is not surprising, because pathway analysis holds great promise both from the standpoint of interpretation (by placing findings in context of prior knowledge) as well as analysis (reducing the multiple comparisons burden inherent to agnostic genome-wide approaches by limiting the number of hypotheses tested to the number of pathways and potentially aggregating multiple weaker signals to a stronger signal). However, realizing the promise of pathway analysis is not straightforward. Most notably, as for many new methodological approaches, pathway analysis suffers from a lack of "gold standards" at every step of implementation: annotation, analysis, interpretation, and design of follow-up studies. As a result, much of the potential associated with pathway analysis remains untapped.

Applying biological knowledge-driven methods to whole genome sequence data as part of Genetic Analysis Workshop 18 (GAW18) highlighted both the promises and the limitations of the pathway approach. In this manuscript, we summarize the results of the work carried out by the members of the pathway analysis working group, leveraging the common themes to suggest several best practices for future investigations. To that end, we will sequentially move through each step of pathway analysis, emphasizing both lessons learned and questions that remain open for further research and discussion.

Methods

Genotype data and pedigree structure

GAW18 genotype data was obtained from 959 participants who are part of the San Antonio Family Sample of the T2D-Genes project [Cite when paper is available]. Detailed sample descriptions are provided elsewhere [Cite when paper is available], but we provide a brief overview below. Of the 959 participants, 483 underwent whole genome sequencing using the services of Complete Genomics Inc., while the sequence of the remaining 476 individuals was imputed based on a combination of (1) pedigree information, (2) genotypes from a ≥500K SNP microarray and (3) the completely sequenced 483 individuals, using a novel imputation pipeline. The final dataset, consisting of 8,348,674 single nucleotide variants (SNVs) spread across the odd numbered chromosomes and was made available to the workshop for analysis, with sample minor allele frequencies (MAF) ranging from 0.1% (singletons; 1/959) to 50%. The 959 participants in the sample were derived from 21 distinct multi-generational large Mexican-American pedigrees.

Real and simulated phenotypes

GAW18 participants had the option to analyze either real or simulated hypertension-related phenotypes. In particular, real systolic and diastolic blood pressure (SBP and DBP) measurements, knowledge of the use of antihypertensive medications, and tobacco smoking were provided for each of the 959 individuals in the sample at one to four time points. Sex, age and years of examinations were also provided. Blood pressure measurements (SBP and DBP measurements), as well as hypertension diagnoses, medication use and tobacco use status were also simulated at up to four time points using a set of variants known to be associated with BP and a complex genetic disease model (see data description paper (cite when available) for details). Two hundred replicates of the simulated phenotype data were generated.

Methods used by participants in the pathway group

The seven contributions to the GAW18 Pathways group used a variety of distinct and innovative ways to explore potential associations between phenotypes and multiple SNV genotypes, where the SNV genotypes are related due to common pathway annotations (Table 1). The approaches taken by the group can be broadly categorized into two distinct groupings based on how the pathway annotation information is incorporated into statistical analyses. The first approach leveraged *a priori* knowledge of candidate pathways to substantially limit the scope of the analysis, while the second approach, more agnostic in nature, explored large sets of pathways for novel relationships with the phenotypes of interest.

Three groups (Almeida et al., 2013; Aslibekyan et al., 2013; Greco et al., 2013) used knowledge of *a priori* associations in their analyses. Aslibekyan et al. focused on 54,309 SNVs within 50 kb of 31 genes in pathways known to be associated with hypertension. Aslibekyan et al. then used a variance components approach in order to evaluate the proportion of variation in real SBP and DBP measurements explained by different subsets of the SNVs (e.g., MAF, location). A similar approach was used by Almeida et al., who used simulated SBP and DBP data to evaluate the contributions of KEGG pathways containing genes associated with blood pressure, though with the addition of a an empirically estimated pathway-specific kinship matrix (PSGRM) to the model. Finally, Greco et al. created synthetic sets of genes containing between zero and five genes known to contain hypertension-related variants. They then evaluated different approaches for summarizing variant-phenotype associations at the pathway level using simulated phenotypes.

The four remaining groups attempted to find pathways significantly associated with the real or simulated phenotypes (Alsulami, Liu, & Beyene, 2013; Dufresne, Oualkacha, Forgetta, & Greenwood, 2013; Edwards et al., 2013; Hu & Paterson, 2013). Dufresne *et al.* started by calculating gene-based phenotype associations with DBP, adjusting for the complex pedigree structure in a mixed effects model (Oualkacha et al., 2013). They then used Cytoscape (Shannon et al., 2003) to determine if any pathways showed an overabundance of genes showing at least modest association with DBP (p<0.05). Finally, using a sample of unrelated individuals, Dufresne used a partial least squares approach to identify multi-dimensional linear combinations of SNVs in significant pathways or genes which maximally explained changes in DBP over time. A similar approach was taken by Edwards et al., who first determined pedigree-adjusted gene-based association statistics for BP phenotypes, followed by analyses to identify pathways showing aggregations of associated genes. In addition, Edwards et al. considered the longitudinal nature of the data by defining the phenotype of interest as an individual's average yearly change in SBP or DBP, assuming a linear relationship between age and BP. Alsulami et al. first evaluated the evidence for association between individual genes on chromosome 3 and BP related phenotypes using a variable weight test (VW-TOW) in a set of 129 unrelated individuals. Computation of gene-based p-values was followed by application of gene set enrichment analysis (GSEA) to 3638 pathways containing both at least one gene from chromosome 3 as well as 10 genes from the odd chromosomes in total. Lastly, Hu and Paterson applied an extended hierarchical generalized linear model to determine associations between genes on chromosome 3 and simulated blood pressure in a set of 142 unrelated individuals. Hu and Paterson then applied GSEA (Michaud et al., 2008) to 531 sets of genes each containing at least five genes on chromosome 3.

Results

In order to provide a clear and comprehensive overview of the similarities, differences and themes in the pathway analysis group, we have organized the results in the general order in which pathway analysis is conducted. We focused first on issues of annotation, addressing topics like how genes and pathways are identified and what databases are most commonly used. Next, we focused on decisions that must be made in the analysis of sequence data using pathway approaches including whether the analysis will adjust for covariates, whether rare and common variants should be analyzed simultaneously, the effects of non-causal SNPs and multiple testing penalties. Third, we discussed issues of interpretation including significance of overlapping/interacting pathways and developing a biological narrative.

Annotation

Gene boundaries. All pathway group members at GAW18 used bioinformatic knowledge about the SNVs to perform multiple aspects of their analyses. In particular, all groups utilized information about gene boundaries. Gene start and stop positions are dependent on the choice of the reference sequence and may vary among data sets. To address this issue and potentially incorporate regulatory sites, the presented analyses extended gene boundaries between 0 and 50kb upstream of the transcription start site and 0-50kb downstream of the stop codon. All groups restricted their analyses to variants in or near coding regions (without giving special consideration to splicing variants) rather than the whole genomic interval.

Pathway annotation. Six of the seven groups utilized external pathway databases containing sets of genes known to be functionally related with databases including KEGG (Kyoto Encyclopedia of Genes and Genomes (Kanehisa et al., 2012)), GO (Gene Ontology(T. G. O. Consortium, 2000)), Reactome (Croft et al., 2010), the GSEA (Gene Set Enrichment Analysis) molecular signature database (Subramanian et al., 2005), and Biocarta (www.biocarta.com). The wide variety of pathway databases used helps underscore the controversy surrounding the very definition of pathways. Furthermore, despite a large number of options for pathway definitions, genome coverage is still low. For example, using Cytoscape (Shannon et al., 2003), Dufresne *et al.* identified pathways for only 35% of the candidate genes (120 annotated genes of 600 in the selected list).

Analysis

Covariate adjustment. With the exception of the analysis by Greco et al., all approaches incorporated adjustment for potential confounders such as age or use of anti-hypertension medication. Simple covariate adjustment for medication use, however, may have diminished statistical power of the BP analyses (Tobin et al., 2005). However, non-genetic covariates explained a relatively modest fraction of outcome variance, estimated at 19% by Aslibekyan et al. Models presented by Almeida et al. also included ancestry principal components to control for possible population stratification and found them to account for 5% of the total variance in BP. Additionally, all papers that analyzed related individuals estimated a kinship matrix to account for correlations inherent to family data.

Rare and common variants. All analyses utilized genotypes across the entire allele frequency spectrum, either stratifying between rare and common SNVs or considering them simultaneously. Hu et al. observed severe inflation of type 1 error rate for analyses combining rare and common variants when compared with a stratified analysis using a hierarchical generalized linear model. Alsulami et al. addressed this concern by applying differential weights to SNVs based on their minor AF (MAF). Although Aslibekyan et al. established that rare and very rare variants contribute more to the overall phenotypic variation, the analysis by Almeida et al. found that including rare variants did not improve the model's ability to detect causal pathways.

Causal and non-causal variants. Analyses of simulated data highlighted the relevance of including causal genes to increase the ability to detect a true effect, i.e. statistical power. Specifically, Greco et al. found that for most statistical tests, the number of causal genes in the set or the proportion of causal variants did not improve power, but the fraction of phenotypic variance explained by causal SNVs did. While the three articles that used simulated data (Almeida *et al*. [2013], Greco *et al*. [2013], and Hu *et al*. [2013]) had the explicit knowledge of which genes were causal in the pathogenesis of hypertension, the other analyses relied on various proxy measures of causality. For example, both Dufresne *et al*. [2013] and Edwards *et al*. [2013] used only exonic variants, and all other analyses were restricted to variants that could be mapped to known genes. Aslibekyan *et al*. [2013]

explicitly compared variance contributions of SNVs within known blood pressure pathway genes with those of SNVs located within 50kb upstream of the transcription start site and within 50Kb downstream of the stop codon, and found that genic regions played a more prominent role in the genetic architecture of hypertension than the flanking regions.

Correcting for multiple testing. An important factor motivating all seven approaches was lowering the astronomical multiple testing burden of whole genome sequence analyses by incorporating *a priori* biological knowledge. However, reducing the number of tested hypotheses to the number of pre-defined pathways rarely yielded statistically significant results, even with relatively liberal correction methods such as the false discovery rate (Alsulami *et al*. [2013], Dufresne *et al*. [2013], and Hu *et al*. [2013]). Moreover, Greco *et al*. [2013] reported poor performance of various collapsing techniques in the context of very large sets of SNVs, suggesting that the reduction in analytic dimensions was either insufficient or masked causal signal. Alternative approaches, such as the one implemented by Edwards *et al*. [2013], effectively ignored the multiple comparison issue by using a two-stage design, simply ranking SNVs on the basis of their p-values to inform subsequent pathway analyses.

Interpretation

Overlap of pathways between phenotypes. Two of the analyses (Dufresne *et al*. [2013], Hu *et al*. [2013]) found evidence of biological pathway cross-talk, defined as sharing at least one functional locus or gene. Hu *et al*. [2013] observed cross-talk between four pathways consistently enriched across adjacent time periods. Similarly, Dufresne *et al*. [2013] evaluated overlapping networks between baseline DBPand the change in DBP over time, and also found four common enriched pathways. The evidence in support of cross-talk between biological pathways is further supported by the high degree of correlation between the analyzed phenotypes.

Biological narrative vs. data. Many fundamental biological processes such as cell growth or adhesion are common to a plethora of phenotypes, which makes it easy to construct a plausible biological story out of pathway findings, but can also create challenges in interpretation. For example, *UBC* emerged as a top hub in the analyses presented by Edwards *et al*. [2013] despite not even being included in the input gene lists. *UBC* encodes ubiquitin, a protein that affects a wide range of cellular processes that are not necessarily unique to the phenotype of interest. At approximately 6000 bp, UBC is not a particularly large gene (i.e. it does not contain many SNVs), nor is it proximal to any of the causal BP variants to suggest linkage disequilibrium. In a setting of limited statistical power, driven by a very large number of genetic predictors with small effects and a comparatively small number of study participants, genes like *UBC* may represent false positive findings with misleading biological plausibility.

Discussion

The seven members of the pathway analysis group applied a diverse set of novel and existing methods of pathway analysis to investigate its utility at finding evidence of plausible pathway-phenotype relationships based on next-generation sequencing data. While the groups remain generally optimistic about the potential for pathway analysis approaches to the analysis of sequence data, performance on GAW18 data was generally poor (likely related, at least in part, to the availability of only half of the genome, small sample size, and small average effect size), and raised more questions than answers. In particular, the lack of standards or best-practices for pathway analysis remains a significant hurdle when applying pathway analysis to single nucleotide variant data.

Despite its promise of potentially improving statistical power due to a reduction in multiple testing and aggregation of independent signals, significant questions about the power of pathway analysis in practice still exist. In particular, best practices for extending gene boundaries to potentially include transcriptional sites do not exist since transcription factors could regulate expression of distantly located target genes due to the chromatin tridimensional structure. However, approximately 85% of active sites are located 2 kb upstream of a transcriptional start site (Iwama & Gojobori, 2004) and inclusion of non-causal variants in aggregation tests are known to significantly diminish power for most standard approaches using either common (Petersen, Alvarez, De Claire, & Tintle, 2013) or rare (Liu, Fast, Zawistowski, & Tintle, 2013) variants. Nevertheless, some approaches are more robust to the inclusion of non-causal variants than others (Liu et al., 2013; Petersen et al., 2013). Recent results assigning function to about 80% of the genome (ENCODE Project Consortium, 2012) may substantially change the way we evaluate variants in these regions and define biologically meaningful pathways. Further methodological work exploring robust and powerful pathway analysis methods is needed particularly in order to limit the impact of non-causal variants through statistical approaches and bioinformatic prediction.

Pathway analysis is also significantly limited by the multitude of options that exist for choosing pathways. First, there is the choice between the many existing pathway databases, some of them overlapping. Second, the principle of reducing multiple testing penalties by testing pathways instead of genes, or even single markers, is predicated on, *a priori*, identification of minimally overlapping pathways which correctly partition the set of all genes (or variants) so that causal genes (variants) are in the same pathway, or, if not, the sensitivity of statistical approaches to identify a pathway as significant based on a single variant. Stated differently, while multiple testing penalties for testing all *m* genes (variants) individually are substantial when *m* is large, multiple testing penalties are substantially worse if all possible subsets of the *m* genes (2^m possibilities) are tested. Further work is needed to determine sensitive and efficient approaches for identifying the subset of the 2^m possible pathway sets most likely to show strong association with the phenotype. The hierarchical nature of pathways are also generally ignored.

Compounding the problem even further is the bias towards well-known and understood genes in pathway databases. Funding limitations for database curation and a bias towards testing well-known genes, instead of investigating higher-risk, but potentially higher reward, novel genes, means that only a modest fraction of genes are well-annotated, which significantly limits approaches like pathway analysis which require annotation information in the analysis.

Another issue for pathway analysis is the lack of general approaches which incorporate all potential sources of evidence for genotype-phenotype relationships. In particular, the GAW18 data is complex due to the inclusion of covariates, family structure, and longitudinal, correlated phenotypes. Most groups had to restrict their analyses to certain portions of the data in order to use published or novel methods, due to limitations of those models which limited their application to unrelated individuals, cross-sectional phenotypes, no covariates, etc. As already noted, lack of statistical power is an issue in pathway analysis, thus further methodological development is necessary to ensure that all sources of evidence can be incorporated in downstream statistical analyses of genetic data.

Finally, all members of the pathway group framed their results as a starting point for further investigations rather than as providing conclusive evidence of causal associations, both due to limited statistical power to detect true effects and the imposed restriction to odd-numbered chromosome variants. As with other high-throughput genomic approaches replication remains the gold standard to establish validity of the findings. In pathway analysis, replication efforts are often thwarted by differences in the choice of curated databases with only limited overlap between gene sets, as evidenced by the variety of approaches among the seven members of the pathway group. Furthermore, because of difficulties in interpretation of pathway findings discussed above, such high-level analyses may be followed up by gene-based or single variant tests to facilitate future research or clinical applications. Finally, questions remain about whether a significant pathway association is indicative of a single strong variant association, multiple correlated signals, or associations with multiple variants of the same outcome (e.g. DBP and SBP), which necessarily impacts choices about how to follow-up pathway analysis.

While numerous questions remain in pathway analysis for sequence data, we are encouraged and convinced that the field is worthy of continued methodological development for a number of reasons. First, the promise of improved statistical power via multiple testing and aggregated signals is too alluring to ignore. This allure is underscored by the potential for improved biological interpretation via pathway analysis. Second, the best-practices questions for pathways mirror many of those for gene-based tests, thus methodological effort in these areas should be synergistic. Third, it is important to note the recent and growing success of pathway analysis in the understanding of cancer biology by way of gene expression data. While it is possible that the genetic architecture of cancer itself (mutations or deregulation of several genes) makes it more conducive to pathway analysis, we believe that many of the most debilitating, complex diseases known to have a heritable component (e.g., cardiovascular outcomes; psychiatric disorders; diabetes) likely also follow a polygenic disease architecture. Finally, combining pathway approaches with transcriptome data may enable future studies to generate pathways or gene sets that cover a larger portion of the genome and enhance the biological meaning of such findings.

Conclusion

Despite a host of methodological questions about best practices, and the generally underwhelming performance of current pathway analysis approaches on GAW18 data, the promise of reducing biological complexity and improving statistical power warrants further methodological efforts.

Tables

Table 1. Comparison of the GAW18 Gene Pathway working group methods.

	Phenotype	Pedigree or unrelated?	Test for association	Software	Pathway Analysis	Database
Almeida et al.	Simulated	Pedigree	Variance Components	SOLAR	SOLAR	KEGG
Alsulami et al.	Real	Unrelated	VW-TOW	VW-TOW	GSEA	MSigDB
Aslibekyan et al.	Real	Pedigree	Variance Components	ASSOC	--	Qiagen
Dufresne et al.	Real	Both	Variance Components	ASKAT	Cytoscape	Reactome
Edwards et al.	Real	Pedigree	Linear Model	Golden Helix	IPA	Ingenuity
Greco et al.	Simulated	Pedigree	Multiple	--	--	Simuled
Hu and Paterson	Simulated	Unrelated	Linear Model	BhGLM	GSEA	Multiple

Acknowledgments

T2D-GENE is supported by NIH grants: U01 DK085524, U01 DK085501, U01 DK085526, U01 DK085584 and U01 DK085545. The Genetic Analysis Workshop 18 was supported by

NIH grant R01 GM031575. NT was supported by NIH grants: R15 HG006915 and R15 HG004543.

References

Almeida, M., Peralta, J., Farook, V., Puppala, S., Kent, J., Duggirala, R., Blangero, J. (2013). Pedigree-based random effect tests to screen gene pathways. *BMC Proceedings*, *In press*.

Alsulami, H., Liu, X., Beyene, J. (2013). Pathway-based analysis of rare and common variants to test for association with blood pressure. *BMC Proceedings*, *In press*.

Aslibekyan, S., Wiener, H., Wu, G., Zhi, D., Shrestha, S., De los Campus, G., Vazquez, A. (2013). Estimating proportions of explained variance: a comparison of whole genome subsets. *BMC Proceedings, In Press*.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock. G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*, 25–29.

Croft, D., O’Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., Stein, L. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, *39*, D691–D697.

Dufresne, L., Oualkacha, K., Forgetta, V., Greenwood, C. M. T. (2013). Pathway analysis for genetic association studies: to do, or not to do, that is the question. *BMC Proceedings*, *In press*.

Edwards, J., Atlas, S., Wilson, S., Cooper, C., Luo, L., Stidley, C. (2013). Integrated statistical and pathway approach to next-generation sequencing analysis: a family-based study of hypertension. *BMC proceedings*, *In press.*

Greco, B., Luedtke, A., Hainline, A., Alvarez, C., Beck, A., Tintle, N. L. (2013). Application of family-based tests of association for rare variants to pathways. *BMC proceedings*, *In press.*

Hu, P., Paterson, A. D. (2013). Dynamic pathway analysis of genes associated with blood pressure using whole genome sequence data. *BMC proceedings, In press*.

Hu, P., Xu, W., Cheng, L., Xing, X., Paterson, A. D. (2011). Pathway-based joint effects analysis of rare genetic variants using Genetic Analysis Workshop 17 exon sequence data. *BMC Proceedings*, *5*(Supplement 9), S45.

Iwama, H., Gojobori, T. (2004). Highly conserved upstream sequences for transcription factor genes and implications for the regulatory network. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(49), 17156–61.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular datasets. *Nucleic Acids Research*, *40*, D109–D114.

Khatri, P., Sirota, M., Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, *8*(2), e1002375. doi:10.1371/journal.pcbi.1002375

Liu, K., Fast, S., Zawistowski, M., Tintle, N. L. (2013). A geometric framework for evaluating rare variant tests of assocation. *Genetic Epidemiology*, *In press*.

Michaud, J., Simpson, K. M., Escher, R., Buchet-Poyau, K., Beissbarth, T., Carmichael, C., Ritchie, M. E., Schutz, F., Cannon, P., Liu, M., Shen, X., Ito, Y., Raskind, W. H., Horwitz, M. S., Osato, M., Turner, D. R., Speed, T. P., Kavallaris, M., Smyth, G. K., Scott, H. S. (2008). Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, *9*, 363. doi:10.1186/1471-2164-9-363

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P. E., Spector, T. D., Hammond, C. J., Richards, J. B., Ciampi, A., Greenwood, C. M. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology*, *37*(4), 366–376.

Petersen, A., Alvarez, C., De Claire, S., Tintle, N. L. (2013). Assessing methods for assigning SNPs to genes in gene-based tests of association using common variants. *PLoS One*, *In press*.

Petersen, A., Sitarik, A., Luedtke, A., Powers, S., Bekmetjev, A., Tintle, N. L. (2011). Evaluating methods for combining rare variant data in pathway-based tests of genetic association. *BMC Proceedings*, *5*(Suppl 9), S48. doi:10.1186/1753-6561-5-S9-S48

Ramanan, V., Shen, L., Moore, J., Saykin, A. (2012). Pathway analysis of genomic data: concepts, methods and prospects for future development. *Trends in Genetics*, *28*(7), 323–332.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. a, Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E.S., Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–50. doi:10.1073/pnas.0506580102

Tintle, N., Aschard, H., Hu, I., Nock, N., Wang, H., Pugh, E. (2011). Inflated type I error rates when using aggregation methods to analyze rare variants in the 1000 Genomes Project exon sequencing data in unrelated individuals: summary results from Group 7 at Genetic Analysis Workshop 17. *Genetic Epidemiology*, *35*(Suppl 1), S56–60. doi:10.1002/gepi.20650

Tobin, M. D., Sheehan, N. A., Scurrah, K. J., Burton, P. R. (2005). Adjusting for treatment effects in studies of quantitative traits: antihypertensive therapy and systolic blood pressure. *Statistics in Medicine, 24*(19), 2911-2935.

Yang, W., Gu, C. (2011). Enrichment Analysis of Genetic Association in Genes and Pathways by Aggregating Signals from both Rare and Common Variants. *BMC Proceedings, 5*(Suppl 9), S52.