OPTIMIZATION, RANDOM RESAMPLING, AND MODELING IN BIOINFORMATICS

BY

WEIHAO GE

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biophysics and Computational Biology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Eric Jakobsson, Chair
Professor Liudmila Sergeevna Mainzer
Professor Saurabh Sinha
Professor Mark Nelson
Professor Kenton McHenry

ABSTRACT

Quantitative phenotypes regulated by multiple genes are prevalent in nature and many diseases falls into this category. High-throughput sequencing and high-performance computing provides a basis to understand quantitative phenotypes. However, finding a statistical approach correctly model the phenotypes remain a challenging problem. In this work, I present a resampling-based approach to obtain biological functional categories from gene set and apply the approach to analyze lithium-sensitivity of neurological diseases and cancer. Then, the non-parametrical permutation-based approach is applied to evaluate the performance of a GWAS modeling procedure. While the procedure performs well in statistics, search space reduction is required to address the computation challenge.

ACKNOWLEDGEMENTS

I would sincerely thank my committee, supervisors, and colleagues, without whose insightful instructions and discussions this work would be impossible. Especially, I would like to express my gratitude to my advisor, Dr. Eric Jakobsson, for his support in academic, financial, and emotional aspects all these years. I am also thankful to my TA supervisor Dr. Kenneth Chapman for enabling me to regain my passion in teaching, science, and reading.

I would also like to thank my family and friends for their love and encouragement in the past 10 years of graduate school life thousands of miles away from home.

TABLE OF CONTENTS

CHAPTER 1: Introduction - Massive Genomic Data, Database, and Analysis 1

CHAPTER 2: Using Optimal F-measure and Random Resampling in Gene Ontology Enrichment Calculation 15

CHAPTER 3: Systems Biology Understanding of the Effects of Lithium on Affective and Neurodegenerative Disorders...46

CHAPTER 4: Systems Biology Understanding of the Effects of Lithium on Cancer 63

CHAPTER 5: An Assessment of True and False Positive Detection Rates of Stepwise Epistatic Model Selection as a Function of Sample Size and Number of Markers 79

CHAPTER 6: Future Work - Search Space Reduction Approaches for GWAS Epistatic Model Selection..102

CHAPTER 1: Introduction - Massive Genomic Data, Database, and Analysis

Continuous Phenotypes

A central problem for biologists is devising how to summarize complex phenomena using a concise set of descriptors, when in fact many biological characteristics are not categorical at the phenome level of description. For example, height[1], BMI[2], and crop nutrient production[3], are all "quantitative phenotypes", to name a few. While these phenotypes do have genomic basis and some finite heritability, but they are more complex than strictly Mendelian traits. Many widely-studied phenotypes, including detrimental diseases[4] such as schizophrenia[5] and autism[6], are also considered to be quantitative, with both genetic and environmental components contributing to observed phenotypic variability. Some examples of these environmental components include soil quality for crop harvest index[7], nutrition for obesity[8], and nurturing environment for intelligence[9]. A central theme of this thesis is the proper statistical methods and their implementation to understand biological phenomena in the context of the above descriptors.

One century ago, Fisher proposed an “infinitesimal model” which successfully described genetic effects on quantitative phenotypes[10]. In this model, trait values are composed of a large number of heritable factors with small contributions in an additive way. This model was later developed to incorporate other factors such as recombination, selection, migration, drifting, mutation, and epistasis[11]. The model explains why heritability in continuous trait is partially "missing". Large number of common SNPs with statistically insignificant individual contributions collectively bring a compelling effect[12]. Therefore, exploration of continuous phenotypes calls for large-scale studies that require huge amount of sequencing, expression, and interaction data. Biologically, the "infinitesimal model" is supported by expression data and regulation network analysis, which find genetic regulation network to be essential in phenotype emergence[13].

Technology Advancement in Data Acquisition, Annotation, and Processing

In the paper "Big Data, Astronomical or Genomical?"[14], Stephens, et. al. compared the amount of genomic data available to-date to that of Astronomical data, YouTube and Twitter, and reviewed the unique challenges in genomic data acquisition, storage, distribution, and analysis.

The distribution stage is more related to hardware architecture and ethical issues, while the acquisition, storage, and analysis stages motivated developments in sequencing technologies, and computational tools and databases.

Technological advancement in Next Generation Sequencing enables efficient, low-cost sequencing, so that enormous amount of data impossible using traditional Sanger method can be produced. For example, a popular sequencing technique is Illumina[15]. Illumina applied bridge PCR to amplify sequence fragments on a solid surface instead of flowing in a homogenized liquid, so that multiple fragments can be amplified at the same time[15]. Fluorescent-tagged nucleotides are attached to the end of fragments, one at a time, during the process of amplification. The fluorescent signal indicates which nucleotide is attached to the amplified fragment at each round, sequencing cooperatively with amplification. Raw sequence data deposited in the Sequence Read Archive (SRA) rapidly increases with the aid of advanced sequencing technologies.[16]

Meanwhile, algorithms and large-scale computing resources have been developed to assemble and align the fragments sequenced. Burrows-wheeler transform[17] indexes all the rotations of a long string. Then the rotated strings are arranged in a binary tree to enable fast search of substrings for alignment. De Bruijn graph organizes overlapped fragments in a directed graph and quickly finds combined sequence, enabling de novo assembly[18]. The genomic data types are largely heterogeneous. Tools addressing different problems have been developed and organized onto platforms and packages. Galaxy[19] provides an online platform mainly focused on sequencing alignment and assembly with simple text manipulation and statistics. OMICtools[20] is an online platform which categorizes online tools by evaluating literature and provides an AI-aided tool choice and workflow construction corresponding to the biological question of interest. For command-line workflows, Biopython[21] provides tools for sequence parsing, alignment, motif prediction, annotation, and statistics. Bioconductor[22] has about 1560 project-oriented tools deposited. Besides providing a convenient access to databases, Bioconductor, based on R language, is especially strong for performing statistical analysis and presenting data.

Databases have evolved to accompany the sequencing data and annotate the genomic information with their biological products and function. Oxford journal website provides a comprehensive collection of databases for convenient search[23]. Database updates are annually reported on the Nucleic Acid Research database issue[23]. Sequencing annotation databases include

SRA for raw sequencing data and alignment[16], DDBJ for nucleotide and protein sequences[24], and GenBank integrating nucleotide sequencing with gene products, protein structure, and biomedical literature[25]. ENCODE[26] project collects highly specified sequencing data in human, mouse, *C. elegans*, and fruit fly, including genome sequences, epigenetic patterns, regulatory binding sites, and chromatin contact information. Databases annotating genomic polymorphisms include dbSNP[27] and dbVar[28], which annotate mutations and structure variations in human chromosome, their frequencies, and impact on function and disease susceptibility. JASPER[29] and DBTSS[30] integrate sequence information for transcription factor binding sites from various species and cell types based on immunoprecipitation, methylation, and RNA sequencing experiments. FANTOM5[31] provides an atlas for regulatory RNA in human and mouse, as well as promoters and enhancers in human. ArrayExpress[32] and Gene Expression Omnibus[33] archive microarray experiment data and make expression data set available for further regulation network study and functional category analysis. UniProt[34] provides a hub annotating protein sequences with rich information including but not limited to: function, activity, active sites, binding regions, post-translational modification, interaction, structure, mutations, and related diseases. OMA[35], OrthoDB[36] and Pfam[37] organize protein by sequence and domain similarity to explore evolution in function. Protein-protein interaction databases such as STRING[38] and BioGRID[39] compile networks through physical contact and functional regulations. KEGG[40] and GO (Gene Ontology)[41] link genes to a systems-functional level of pathways and biological knowledge. An important task is to integrate these data sources for use in projects that tackle biological problems with multiple aspects. One effort is to provide direct mapping or even a common language for cross-talk between databases. BioMart[42] from Ensembl project[43] provides a mapping among different systems of gene and gene product symbols and their functional attributes. Gene Ontology Consortium initially begun with providing a cross-species description of genetic functions[41]. Based on similar approaches, KaBOB[44] system semantically integrates 18 biomedical databases and provides consistent representation of biomedical data and concepts.

Permutation-based Resampling Method for Assessing Many-to-Many Mapping

Associating genomic data with phenotype is a multi-dimensional problem. When the gene-function annotation is available, what functional categories and pathways are overrepresented our gene set? In this case, many ontology terms or pathways are simultaneously tested for their enrichment. When the gene/variant-function annotation is unavailable, what would be a proper model describing the system? In this case, multiple models are tested.

When multiple hypotheses are tested, one might find more false positives than suggested by an uncorrected *p*-value with a fixed threshold. For example, when we evaluated m_0 null hypotheses and found m rejected with a threshold α, then the probability to have at least one false positive would be approximately $m_0\alpha$. Bonferroni correction is a stringent threshold so that the possibility to mistakenly reject 1 true null-hypothesis is under α. While it is too stringent for gene ontology analysis, it is still widely used in genetic variant association studies due to its simplicity. Benjamini-Hochberg correction controls false discovery rate (FDR) under α assuming independent or positively-correlated hypotheses[45]. The Benjamini-Yekutieli method extends corrections to negatively correlated hypotheses[46]. A recent modification of the Benjamini-Hochberg method aims to account for acyclic tree-like structures[47].

These methods are very useful, but none provide a completely reliable estimation of false positives, due to the extremely interconnected nature of the elements in gene ontology, other biological databases, and genomic models. The assumption of independency or positive-correlation is not always valid. Therefore, non-parametric simulated statistics, is essential for accurate assessment of the significance of correlations. Chapters 1-3 discusses that when association is known, genes are permuted with background gene list to show how many annotation terms are overrepresented by chance. Chapter 4 would present the case when association is unknown, permuting phenotypes would show what associations occurs by chance and therefore sets a threshold to control False Positive Rate.

Improving Statistics of Gene Set Analysis, and overview of Thesis Chapter 1

Understanding biology in terms of categories generally reduces to a set of binary classification problems. For each potential category that might describe a system, the question is: "Does that category fit?" Next, since the data are stochastic by nature, we also ask: "How confident are we that this category fits?" and "Is it more important to find as many relevant categories as possible, or to be very sure that the identified categories are correct?" Chapter 1 of the thesis deals with these questions as they apply to gene annotation enrichment studies, in which a list of genes is analyzed for enrichment in a number of biological process-related categories, thus pinpointing which biological processes these genes enable.

The most straightforward way to quickly obtain an overview of biological information in a large data set is to evaluate whether a biological feature is overrepresented in the gene set. Huang, et.al. (2008)[48] reviewed the gene ontology enrichment tools available at the time and categorized them into singular enrichment analysis (SEA), gene set enrichment analysis (GSEA), and modular enrichment analysis (MEA). SEA uses hypergeometric test, Fisher's exact test, or Chi-squared test to evaluate whether the fraction of the genes associated with a certain biological feature in the gene set of interest is reached or exceeded by chance contrasting a background gene set[48]. GSEA applies permutations between experiment and control gene sets to find distribution of maximum enrichment score and its significance[48]. MEA incorporates network relationships between annotated terms with SEA[48]. In the present thesis, a pipeline improving SEA with a GSEA-like random-sampling like procedure is developed, so that the permutation procedures would also be applicable even if gene rank metric or case-control format are hard to obtain.

MEA tools try to overcome the disadvantage of multiple hypothesis correction of SEA by including the acyclic tree structure. TopGO options "elim"[49] and "parent-child"[50] adjust the candidate gene association with a parent term if a child term is found to be enriched. DAVID[51] provides a hierarchical enrichment option so that the terms at the same level would be evaluated at the same time. However, the hierarchical enrichment overlooks the fact that a child term can have multiple parents across levels and the levels of gene ontology terms themselves are poorly defines. Moreover, MEA tools overlook the overlap between the sibling terms, cousin terms, and distant terms due to genetic pleiotropy. For these terms, selection of an overlapped gene would bring positive correlation, while selection of a non-overlapped gene would bring negative correlation. Therefore, we choose to combine SEA tools with resampling method.

In chapter one, a resampling-based method is introduced for false positive rate control. "Null sets", the gene sets randomly selected from background gene set, are evaluated as an estimation for false positives.

Another aspect discussed in chapter one is inclusion of false negatives, which is motivated by application of gene ontology enrichment in our data set of interest, the human orthologs of honey bee alarm pheromone set. The honey bee alarm pheromone set is a set of genes differentially expressed when honey bee aggression behavior is triggered by the chemical alarm pheromone[52]. This honey bee gene set is shown to be more conserved with social placental mammals such as human, than non-social insects[53]. It is likely to provide understanding to genetic basis to social behavior. However, using both heuristic and resampling method with FDR<=0.05 brings only the most general gene ontology terms. The false negative is investigated. Metrics[54,55] , such as F-measure and Matthews Correlation Coefficient, balancing false negatives and true positives are introduced. Optimizing these metrics provides a statistical basis to rationally alter p-value thresholds. Relaxing significance threshold to balance signal and noise may bring back more specific gene ontology terms, than an arbitrary threshold.

The biological basis of complex continuous phenotypes also justifies increasing thresholds to embrace the previously thought "insignificant" terms. Complex continuous phenotypes are regulated by large quantity of genes with small contributions cooperating in a non-centralized network. The small alarm-pheromone set human ortholog is not contradicting to the feature in that a non-centralized network still maintains its general properties even if a certain portion of nodes/genes is removed. On the other hand, contributions of "insignificant" ontology terms would be large due to the large population of these terms. Therefore, it is also biologically reasonable to relax the threshold if optimization of balancing metrics requires.

Application of Simulated Enrichment Statistics on Lithium Sensitive Gene Sets, and overview of Thesis chapters 2 and 3

Chapters two and three apply statistical method from chapter one for a systems analysis of lithium-sensitive genes. Jakobsson, et.al. have provided a comprehensive review of the biochemical mechanism of lithium function and its importance in neurodegenerative disease, affective disorder, and cancer[56]. Lithium has been used as a treatment in major depressive disorder[57], bipolar disorder[58], and schizophrenia[59]. In combined treatment, lithium has enhanced effects of SSRI[60]. Epidemiology studies have revealed that lithium concentration negatively correlate with dementia[61] and rising rates of Alzheimer's disease mortality with age[62]. Retrospective studies have revealed that lithium reduced cancer risk in patients with bipolar disorder[63]. Li^+ competes with Mg^{2+} in binding with many kinases including GSK3B[64], which regulates a large number of substrates[65]. Among the substrates inhibited by GSK3B, BDNF[66] has shown neuro-protective function and has been a therapeutic target for Alzheimer's disease[67].

Lithium inhibition of inositol monophosphatase, which is another Mg^{2+} binding enzyme, have been shown to induce autophagy[68].

To systematically study the effect of lithium, it is key to understand how the Mg^{2+} binding enzyme interactive network is involved in the disease-related biochemical pathways. The protein-protein interaction database STRING[38] provides annotations based on experiment, ortholog evidence, and computational prediction. The interactome largely covering same genes are merged together. Eventually, 10 distinctive interactomes become the input of the enrichment analysis. Each interactome is then processed by the pipeline developed in chapter 1 for a resampling-based KEGG pathway enrichment study using 1000 null sets.

For each disease annotated in the KEGG database[40], either they are annotated to a single disease term or to multiple pathway terms. A collective *p*-value is computed by the geometric mean of all the *p*-value of mutual enrichment of interactome and KEGG terms associated with the disease. The lower the *p*-value, the more sensitive the disease would be to lithium. We used the analysis to identify which cancers and neurological diseases are likely to be most responsive to lithium therapy and which genes are most promising targets in a multidrug therapy involving lithium.

The calculation result has shown that except for very few outliers, cancer terms has achieved the lowest collective p-value. A majority of neurodegenerative disease and affective disorder have good enrichment in lithium interactome. Metabolic pathways are not responsive to lithium interactomes. The bipolar disorder and major depressive disorder, which are clinically proved to be sensitive to lithium treatment, have shown high enrichment.

The key genes are evaluated by how many times they appeared in the intersection of a disease-related pathway and lithium interactome. The number is normalized by the number of pathways associated with the disease. Among the highly ranked genes, many are known to be the drug target for disease treatment. For example, the APP is identified as a key protein in Alzheimer's disease and major depressive disorder. In both diseases, amyloid-beta produced by abnormal cleavage of APP are observed clinically[69,70]. MAPK3, associated with cell apoptosis[71], is both highly ranked in neurodegenerative disorder and cancer.

The study is limited with direct annotation and the contribution of individual genes are assumed to be the same, which is not the case in reality. However, it is a quick way to indicate which diseases are sensitive to lithium treatment, yet regardless of the direction. Many of the well-identified drug targets, including the key enzymes themselves, are absent from the direct annotation. The interaction network showed how these genes regulates the highly ranked genes found in the study and bridges lithium effects on the diseases.

Quantitative Model for GWAS Data and Search Space Reduction: overview of Chapter 4

Gene mutation, differential expression sensitivity to environment, and interaction are ultimately associated with phenotype variations. Especially, mutations in non-coding region is expected to result in less detrimental consequences to phenotype than those in coding regions. But these variations would alter phenotypes by regulating the expression of genes[72]. Therefore, it is necessary to include variations in both coding and non-coding regions to build a genotype-to-phenotype model of high resolution. It is also found that mutations interact (epistasis), resulting in greater effect on phenotype than would have been the sum of individual contributions, particularly in complex continuous phenotypes[73]. Therefore, to describe a continuous phenotype, a quantitative model including both additive and epistatic effects based on regression is necessary.

In chapter four, a Stepwise Procedure for constructing an Additive and Epistatic Multi-Locus model (SPAEML) is described. This model includes both first order (additive) and second order (two-way epistatic) terms and a normal random noise term[74]. The model is fitted by step-wise model selection procedure[75]. To evaluate the effectiveness of the model, it is compared to two other statistical approaches. One is Joint Linkage analysis[76], which is a multi-locus model that does not incorporate epistasis. The other is FastEpistasis[77], which only includes one pair of epistatic markers in each model.

The applications were tested on synthetic data sets and validated by linear fitting using R. False positive rate, detection rate, and specification rate are evaluated. It is confirmed that the multi-locus models (SPAEML and JL) outperforms single-locus model. The SPAEML is able to specify epistatic and additive effect when minor allele frequency is high. However, due to combinatorial growth of number of models to test, SPAEML takes much longer time to run than JL (a week vs three hours). Therefore, search space reduction is necessary.

Future work: Search Space Reduction in Model Selection

Currently, many efforts have been taken to reduce the possible search space. Pure statistical methods are quickly developed. For example, LASSO[78] captures the strongest effect by adding constraint term in regression procedure. "Screen and Clean"[79] showed a two-staged workflow finding all the additive effects with LASSO and then only fit the interactive pairs between the SNPs identified. In addition to statistical methods, it would also be helpful to take advantage of the prior biological knowledge to effectively exclude the models that are not biologically meaningful. MDR[80] categorized multi-locus data into high-risk and low-risk group so that the combination of multiple factors is reduced to two groups. Later versions of MDR also work with quantitative phenotypes.[81]

Haplotyping is one such approach. Experiments have shown that human chromosomes are organized in haplotype blocks where the groups of genomic variations[82] are likely to be inherited together. Therefore, grouping SNPs in haplotype blocks would effectively reduce the candidate markers in the model[83,84,85].

Additionally, the molecular basis of epistasis lies in gene regulation network[86]. Magnum[87] builds on about 200 tissue-specific regulatory network information from FANTOM5[31]and validated by GTEx[88]. KnowENG[89] incorporates the interaction information from databases like STRING[38], annotation databases like KEGG[40] and GO[41], and protein domain similarity information to predict interaction between exons. Juicer tool packages[90] process Hi-C data to reveal 3D structure of chromatin and bring information about contact probability of long-distance genetic elements. These software packages enable evaluation of interaction probability and provide effective ways to integrate regulatory and co-expression information in modeling.

Stochastic search and non-parametrical model building approaches have been extensively developed for effectively build a model for GWAS data. Ljungberg, et.al. (2004)[91] provided an efficient algorithm by hierarchically and stochastically partitioning the search space to find global optimal in 2 and 3 QTN models with many-way interactions. Wan, et.al. (2009)[92] developed a machine-learning tree search algorithm to identify interactive SNPs. However, there are no unified workflow that can take the genotype and phenotype table and automatically reduce search space, select SNP and SNP pairs, and fit for the model. My next goal is to explore these algorithms and incorporate statistical structure in the given genomic data set and priori biological knowledge, so that a straightforward workflow can be developed for accurately description of continuous phenotype using multi-locus, interaction-inclusive models.

Conclusion

In this work, I have demonstrated that non-parametrical statistics is now a feasible way to analyze genomic data. For gene annotation enrichment analysis, resampling simulation estimates false discovery rate in an unbiased way regardless the structure of the database. A flexible significance threshold balancing detection and correctness is better than an arbitrary threshold in exploring highly noised data set such as the honey bee alarm pheromone set. Application the resampling approach for mutual enrichment in lithium-sensitive gene set and disease-associated pathways is a fast and straightforward way to identify disease responsive to lithium and potential pharmaceutical targets for these diseases. In GWAS analysis where data set is much larger, permutation method is more computationally challenging. For a small data set, I was able to validate the step-wise procedure including both additive and epistatic effects for multi-locus model (SPAEML). The performance for the modeling procedure was evaluated using permutation method and simulated data set. For realistic-sized data set, search-space-reduction for potential model is essential.

[1] Jelenkovic, A., Sund, R., Hur, Y. M., Yokoyama, Y., Hjelmborg, J. V. B., Möller, S., ... & Aaltonen, S. (2016). Genetic and environmental influences on height from infancy to early adulthood: An individual-based pooled analysis of 45 twin cohorts. Scientific reports, 6, 28496.

[2] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., ... & Croteau-Chonka, D. C. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature, 518(7538), 197.

[3] Owens, B. F., Lipka, A. E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., ... & Buckler, E. S. (2014). A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. Genetics, 198(4), 1699-1716.

[4] McCarthy, Mark I., et al. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature reviews genetics 9.5 (2008): 356.

[5] Sullivan, Patrick F., Kenneth S. Kendler, and Michael C. Neale. "Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies." Archives of general psychiatry60.12 (2003): 1187-1192.

[6] Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., ... & Lotspeich, L. (2011). Genetic heritability and shared environmental factors among twin pairs with autism. Archives of general psychiatry, 68(11), 1095-1102.

[7] Cassman, K. G. (1999). Ecological intensification of cereal production systems: yield potential, soil quality, and precision agriculture. Proceedings of the National Academy of Sciences, 96(11), 5952-5959.

[8] Speakman, J. R. (2004). Obesity: the integrated roles of environment and genetics. The Journal of nutrition, 134(8), 2090S-2105S.

[9] Davies, G., Tenesa, A., Payton, A., Yang, J., Harris, S. E., Liewald, D., ... & McGhee, K. (2011). Genome-wide association studies establish that human intelligence is highly heritable and polygenic. Molecular psychiatry, 16(10), 996.

[10] R.A. Fisher. "The correlation between relatives on the supposition of Mendelian inheritance."

Trans. R. Soc. Edinb., 52 (1918), pp. 399-433

11 Barton, Nick H., Alison M. Etheridge, and Amandine Véber. "The infinitesimal model." bioRxiv (2016): 039768.

12 Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., ... & Goddard, M. E. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature genetics, 42(7), 565.

13 Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. "An expanded view of complex traits: from polygenic to omnigenic." Cell 169.7 (2017): 1177-1186.

14 Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., ... & Robinson, G. E. (2015). Big data: astronomical or genomical? PLoS biology, 13(7), e1002195.

15 "Illumina Sequencing Technology Highest data accuracy, simple workflow, and a broad range of applications." [pdf file] https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

16 Kodama, Y., Shumway, M., & Leinonen, R. (2011). The Sequence Read Archive: explosive growth of sequencing data. Nucleic acids research, 40(D1), D54-D56.

17 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754-1760.

18 Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research, 18(5), 821-829.

19 Afgan, E., Baker, D., Van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., ... & Grüning, B. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic acids research, 44(W1), W3-W10.

20 Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J., & Desfeux, A. (2014). OMICtools: an informative directory for multi-omic data analysis. Database, 2014.

21 Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... & De Hoon, M. J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25(11), 1422-1423.

22 Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., ... & Hornik, K. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome biology, 5(10), R80.

23 Galperin, M. Y., Fernández-Suárez, X. M., & Rigden, D. J. (2017). The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. Nucleic acids research, 45(D1), D1-D11.

24 Kodama, Y., Mashima, J., Kosuge, T., Katayama, T., Fujisawa, T., Kaminuma, E., ... & Nakamura, Y. (2014). The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data. Nucleic acids research, 43(D1), D18-D22.

25 Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. Nucleic acids research, 41(D1), D36-D42.

26 ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57.

27 Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic acids research, 29(1), 308-311.

28 Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., ... & Paschall, J. (2012). DbVar and DGVa: public archives for genomic structural variation. Nucleic acids research, 41(D1), D936-D941.

29 Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., ... & Lim, J. (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic acids research, 42(D1), D142-D147.

30 Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuchihara, K., Sugano, S., ... & Nakai, K. (2014). DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. Nucleic acids research, 43(D1), D87-D91.

31 Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., ... & Gil, L. (2014). Ensembl 2015. Nucleic acids research, 43(D1), D662-D669.

32 Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., ... & Mani, R. (2006). ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic acids research, 35(suppl_1), D747-D750.

33 Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 30(1), 207-210.

34 UniProt Consortium. (2014). UniProt: a hub for protein information. Nucleic acids research, 43(D1), D204-D212.

35 Altenhoff, A. M., Škunca, N., Glover, N., Train, C. M., Sueki, A., Piližota, I., ... & Gonnet, G. H. (2014). The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. Nucleic acids research, 43(D1), D240-D249.

36 Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, F. A., Ioannidis, P., ... & Kriventseva, E. V. (2016). OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic acids research, 45(D1), D744-D749.

37 Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., ... & Sonnhammer, E. L. (2013). Pfam: the protein families database. Nucleic acids research, 42(D1), D222-D230.

38 Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... & Kuhn, M. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic acids research, 43(D1), D447-D452.

39 Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic acids research, 34(suppl_1), D535-D539.

40 Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

41 Gene Ontology Consortium. (2014). Gene ontology consortium: going forward. Nucleic acids research, 43(D1), D1049-D1056.

42 Kasprzyk, A. (2011). BioMart: driving a paradigm change in biological data management. Database, 2011.

43 Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., ... & Durbin, R. (2002). The Ensembl genome database project. Nucleic acids research, 30(1), 38-41.

44 Livingston, K. M., Bada, M., Baumgartner, W. A., & Hunter, L. E. (2015). KaBOB: ontology-based semantic integration of biomedical databases. BMC bioinformatics, 16(1), 126.

45 Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society. Series B (Methodological), 289-300.

46 Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165-1188.

47 Bogomolov, M., Peterson, C. B., Benjamini, Y., & Sabatti, C. (2017). Testing hypotheses on a tree: new error rates and controlling strategies. arXiv preprint arXiv:1705.07529.

48 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1-13.

49 Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 22(13), 1600-1607.

50 Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics, 23(8), 980-987.

51 Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44.

52 Alaux, C., Sinha, S., Hasadsri, L., Hunt, G. J., Guzmán-Novoa, E., DeGrandi-Hoffman, G., ... & Robinson, G. E. (2009). Honey bee aggression supports a link between gene regulation and behavioral evolution. Proceedings of the National Academy of Sciences, 106(36), 15400-15405.

53 Liu, H., Robinson, G. E., & Jakobsson, E. (2016). Conservation in mammals of genes associated with aggression-related behavioral phenotypes in honey bees. PLoS computational biology, 12(6), e1004921.

54 Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

55 Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451.

56 Jakobsson, E., Argüello-Miranda, O., Chiu, S. W., Fazal, Z., Kruczek, J., Nunez-Corrales, S., ... & Pritchet, L. (2017). Towards a Unified Understanding of Lithium Action in Basic Biology and its Significance for Applied Biology. The Journal of membrane biology, 250(6), 587-604.

57 Treiser, S. L., Cascio, C. S., O'Donohue, T. L., Thoa, N. B., Jacobowitz, D. M., & Kellar, K. J. (1981). Lithium increases serotonin release and decreases serotonin receptors in the hippocampus. Science, 213(4515), 1529-1531.

58 American Psychiatric Association. (2002). Practice guideline for the treatment of patients with bipolar disorder (revision). American Psychiatric Pub.

59 Hasan, A., Falkai, P., Wobrock, T., Lieberman, J., Glenthoj, B., Gattaz, W. F., ... & WFSBP Task force on Treatment Guidelines for Schizophrenia. (2013). World Federation of Societies of Biological Psychiatry (WFSBP) guidelines for biological treatment of schizophrenia, part 2: update 2012 on the long-term treatment of schizophrenia and management of antipsychotic-induced side effects. The world journal of biological psychiatry, 14(1), 2-44.

60 Crossley, N. A., & Bauer, M. (2007). Acceleration and augmentation of antidepressants with lithium for depressive disorders: two meta-analyses of randomized, placebo-controlled trials. The Journal of clinical psychiatry.

61 Kessing, L. V., Gerds, T. A., Knudsen, N. N., Jørgensen, L. F., Kristiansen, S. M., Voutchkova, D., ... & Ersbøll, A. K. (2017). Association of lithium in drinking water with the incidence of dementia. JAMA psychiatry, 74(10), 1005-1010.

62 Fajardo, V. A., Fajardo, V. A., LeBlanc, P. J., & MacPherson, R. E. (2018). Examining the Relationship between Trace Lithium in Drinking Water and the Rising Rates of Age-Adjusted Alzheimer's Disease Mortality in Texas. Journal of Alzheimer's Disease, (Preprint), 1-10.

63 Martinsson, L., Westman, J., Hällgren, J., Ösby, U., & Backlund, L. (2016). Lithium treatment and cancer incidence in bipolar disorder. Bipolar disorders, 18(1), 33-40.

64 Ryves, W. J., & Harwood, A. J. (2001). Lithium inhibits glycogen synthase kinase-3 by competition for magnesium. Biochemical and biophysical research communications, 280(3), 720-725.

65 Beurel, E., Grieco, S. F., & Jope, R. S. (2015). Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases. Pharmacology & therapeutics, 148, 114-131.

66 Mai, L., Jope, R. S., & Li, X. (2002). BDNF‐mediated signal transduction is modulated by GSK3β and mood stabilizing agents. Journal of neurochemistry, 82(1), 75-83.

67 Nagahara, A. H., & Tuszynski, M. H. (2011). Potential therapeutic uses of BDNF in neurological and psychiatric disorders. Nature reviews Drug discovery, 10(3), 209.

68 Sarkar, S., Floto, R. A., Berger, Z., Imarisio, S., Cordenier, A., Pasco, M., ... & Rubinsztein, D. C. (2005). Lithium induces autophagy by inhibiting inositol monophosphatase. J Cell Biol, 170(7), 1101-1111.

69 Julia, T. C. W., & Goate, A. M. (2017). Genetics of β-amyloid precursor protein in Alzheimer's disease. Cold Spring Harbor perspectives in medicine, 7(6), a024539.

70 Pomara, N., & Bruno, D. (2016). Major depression may lead to elevations in potentially neurotoxic amyloid beta species independently of Alzheimer Disease. The American Journal of Geriatric Psychiatry, 24(9), 773-775.

71 Minutoli, L., Antonuccio, P., Polito, F., Bitto, A., Squadrito, F., Di Stefano, V., ... & Altavilla, D. (2009). Mitogen-activated protein kinase 3/mitogen-activated protein kinase 1 activates apoptosis during testicular ischemia–reperfusion injury in a nuclear factor-κB-independent manner. European journal of pharmacology, 604(1-3), 27-35.

72 Cubillos, F. A., Coustham, V., & Loudet, O. (2012). Lessons from eQTL mapping studies: non-coding regions and their role behind natural phenotypic variation in plants. Current opinion in plant biology, 15(2), 192-198.

73 Carlborg, Ö., & Haley, C. S. (2004). Epistasis: too often neglected in complex trait studies?. Nature Reviews Genetics, 5(8), 618.

74 Bogdan, M., Ghosh, J. K., & Doerge, R. W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics, 167(2), 989-999.

75 Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nature genetics, 44(7), 825.

76 Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics, 23(19), 2633-2635.

77 Schüpbach, T., Xenarios, I., Bergmann, S., & Kapur, K. (2010). FastEpistasis: a high performance computing solution for quantitative trait epistasis. Bioinformatics, 26(11), 1468-1469.

78 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

79 Wu, J., Devlin, B., Ringquist, S., Trucco, M., & Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. Genetic epidemiology, 34(3), 275-285.

80 Hahn, L. W., Ritchie, M. D., & Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. Bioinformatics, 19(3), 376-382.

81 Lou, X. Y., Chen, G. B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., & Li, M. D. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. The American Journal of Human Genetics, 80(6), 1125-1137.

82 Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., ... & Liu-Cordero, S. N. (2002). The structure of haplotype blocks in the human genome. Science, 296(5576), 2225-2229.

83 Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., & Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. PLoS computational biology, 12(1), e1004714.

84 Cowman, T., & Koyutürk, M. (2017). Prioritizing tests of epistasis through hierarchical representation of genomic redundancies. Nucleic acids research, 45(14), e131-e131.

85 Zhang, Y., Zhang, J., & Liu, J. S. (2011). Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. The annals of applied statistics, 5(3), 2052.

86 Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. Trends in Genetics, 27(8), 323-331.

87 Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. Nature methods, 13(4), 366.

88 GTEx Consortium. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science, 348(6235), 648-660.

89 Sinha, S., Song, J., Weinshilboum, R., Jongeneel, V., & Han, J. (2015). KnowEnG: a knowledge engine for genomics. Journal of the American Medical Informatics Association, 22(6), 1115-1119.

90 Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell systems, 3(1), 95-98.

91 Ljungberg, K., Holmgren, S., & Carlborg, Ö. (2004). Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. Bioinformatics, 20(12), 1887-1895.

92 Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N. L., & Yu, W. (2009). Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics, 26(1), 30-37.

CHAPTER 2: Using Optimal F-measure and Random Resampling in Gene Ontology Enrichment Calculations

(Submitted to *Frontiers in Applied Mathematics and Statistics*, in Review)

Weihao Ge, Zeeshan Fazal and Eric Jakobsson*

*Correspondence: jake@illinois.edu

Abstract

Background: A central question in bioinformatics is how to minimize arbitrariness and bias in analysis of patterns of enrichment in data. A prime example of such a question is enrichment of gene ontology (GO) classes in lists of genes. Our paper deals with two issues within this larger question. One is how to calculate the false discovery rate (FDR) within a set of apparently enriched ontologies, and the second how to set that FDR within the context of assessing significance for addressing biological questions, to answer these questions we compare a random resampling method with a commonly used method for assessing FDR, the Benjamini-Hochberg (BH) method. We further develop a heuristic method for evaluating Type II (false negative) errors to enable utilization of F-Measure binary classification theory for distinguishing “significant” from “non-significant” degrees of enrichment.

Results: The results show the preferability and feasibility of random resampling assessment of FDR over the analytical methods with which we compare it. They also show that the reasonableness of any arbitrary threshold depends strongly on the structure of the dataset being tested, suggesting that the less arbitrary method of F-measure optimization to determine significance threshold is preferable.

Conclusion: Therefore, we suggest using F-measure optimization instead of placing an arbitrary threshold to evaluate the significance of Gene Ontology Enrichment results, and using resampling to replace analytical methods

Keywords: Gene Ontology; F-measure; False Discovery Rate; Microarray Data Analysis

Background

Gene Ontology (GO) enrichment analysis is a powerful tool to interpret the biological implications of selected groups of genes. The gene lists from experiments such as microarrays, are gathered into clusters associated with biological attributes, and defined as GO terms[1]. The GO terms are arranged in an acyclic tree structure from more specific to more general descriptions, including biological process (BP), cellular component (CC), and molecular function (MF). GO aspires to create a formal naming system to define the biologically significant attributes of genes across all organisms. Each enriched GO term derived from a list of genes is evaluated by its significance level, i.e. the probability that the measured enrichment would be matched or exceeded by pure chance.

Enrichment tools have been developed to process large gene lists to generate significantly enriched ontologies. Huang *et.al* (2009) summarizes the tools widely used for GO enrichment[2] . Different tools emphasize different features. Gorilla[3], DAVID[4], g:profiler[5] are web interfaces that integrate functional annotations including GO annotations, disease and pathway databases etc. Blast2GO[6] extends annotation of gene list to non-model organisms by sequence similarity. GO-Miner[7], Babelomics[8], FatiGO[9], GSEA[10],[11], and ErmineJ[12] apply resampling or permutation algorithms on random sets to evaluate the number of false positives in computed gene ontologies associated with test sets. DAVID [4] and Babelomics[8] introduced level-specific enrichment analysis; that is, not including both parents and children terms. TopGO contains options, “eliminate” and “parent-child”, which eliminate or reduce the weight of genes in the enriched children terms when calculating parent term enrichment[13]. TopGO[14] and GOstats[15] provide R-scripted tools for ease of further implementation. Cytoscape plugin in BinGO [16] is associated with output tree graphs.

To calculate raw *p*-values for GO enrichment without multiple hypothesis correction, methods used include exact or asymptotic (i.e. based on the hypergeometric distribution or on Pearson's distribution), one- or two-sided tests[17]. Rivals *et. al.* discussed the relative merits of these methods[17].

Generally, inference of the statistical significance of observed enrichment of categories in gene ontology databases can’t be assumed to be parametric, because there is no *a priori* reason to postulate normal distributions within gene ontology terms. Randomization methods are powerful tools for testing nonparametric hypotheses[18]. However, heuristic methods for testing nonparametric hypotheses have long been widely used due to lack of adequate computational resources for randomization tests. In gene ontology enrichment, a widely-used heuristic method is that of Benjamini and Hochberg[19]. In their original paper, Benjamini and Hochberg tested their method against a more computationally intensive resampling procedure for selected input data and found no significant difference, Thus the more computationally efficient Benjamini-Hochberg method was justified.

Benjamini-Hochberg has been widely applied in enrichment tools such as BinGO[16], DAVID[4], GOEAST[20], Gorilla[3], and Babelomics[8], to name a few. The similar Benjamini-Yekutieli method is included in the GOEAST package which enables to control the FDR even with negatively correlated statistics[20] [21]. A recent approach published by Bogomolov, *et.al.* (2017) deals with multiple hypothesis correction and error control for enrichment of mutually dependent categories in a tree structure using a hierarchical Benjamini-Hochberg-like correction[22]. Gossip provides another heuristic estimation of false positives that compares well with resampling in the situations tested[23].

A randomized permutation method for assessing false positives is embedded in the protocol of Gene Set Enrichment Analysis (GSEA)[10]. Kim and Volsky[24] compared a parametric method (PAGE) to GSEA and found that PAGE produced significantly lower *p*-values (and therefore higher putative significance) for the same hypotheses. They suggest that PAGE might be more sensitive because GSEA uses ranks of expression values rather than measured values themselves. However, they do not demonstrate that the hypothesis of normal distributions in gene ontology databases that underlies PAGE is generally true.

Noreen[25] considered the potential of using more widely available computer power to do exact testing for the validity of hypotheses, in order to be free of any assumptions about the sampling distributions of the test statistics, for example the assumption of normality. The essence of the more exact methods is the generation of a null hypothesis by the creation and analysis of sets of randomly selected entities (null sets) that are of the same type as the test set. Then the extent to which the null hypothesis is rejected emerges from comparing the results of conducting the same analysis on the null sets and the test set. As exemplified by the over one thousand citations of this work by Noreen, these methods have been widely adopted in many areas in which complex datasets must be mined for significant patterns, as for example in financial markets.

In the present paper we utilize a straightforward random resampling method for creation of null sets and compare resultant assessments for estimating false positives with commonly used analytical methods as applied to gene ontology enrichment analysis. We also evaluate the computational cost of this method relative to analytical methods.

In applying all the cited methods and tools, it is common to apply a threshold boundary between "significant enrichment" and "insignificance". Such assignment to one of two classes is an example of a binary classification problem. Often such classifications are made utilizing an optimum F-measure[26]. Rhee, *et.al.* (2008) have suggested application of F-measure optimization to the issue of gene ontology enrichment analysis[27]. In the present work, we present an approach to gene enrichment analysis based on F-measure optimization, which considers both precision and recall and provides a flexible reasonable threshold for data sets depending on user choice as to the relative importance of precision and recall. We also compare a resampling method to the Benjamini-Hochberg method for estimation of FDR and use with F-measure optimization.

We also consider the argument made by Powers [26] that the F-measure is subject to biases, and that instead of precision and recall (the constituents of the F-measure) the constructs of markedness and informedness should be considered. Whereas precision and recall are entirely based on the ability to identify positive results, informedness and markedness give equal weight to identification of negative results. We note that the Matthews Correlation Coefficient (MCC), another well-vetted measure of significance[28], is the geometric mean of the markedness and informedness.

Our results in this paper will suggest that resampling is preferable to analytical methods to estimate FDR, since the compute costs are modest by today's standards and that even well-accepted and widely used analytical methods may have significant error. Our results also suggest that F-measure or MCC optimization is preferable to an arbitrary threshold when classifying results as "significant" or "insignificant". For the particular analyses in this paper, we found no significant difference in utilizing F-measure vs. MCC. in assessing significance of results in computing enrichment in gene ontology analysis.

Methods

Enrichment Tool

For results reported in this study (described below), the TopGO[14] package is implemented to perform GO enrichment analysis, using the "classic" option. In this option, the hypergeometric test is applied to the input gene list to calculate an uncorrected *p*-value.

FDR Calculation

The empirical resampling and Benjamini-Hochberg (BH) methods are used to estimate the FDR. The *p*-value adjustment using Benjamini-Hochberg is carried out by a function implemented in the R library. http://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html

The resampling method is based on the definition of p-value as the probability that an observed level of enrichment might arise purely by chance. To evaluate this probability, we generate several null sets, which are the same size as the test set. The genes in the null sets are randomly sampled from the background/reference list. GO enrichment analysis was carried out on both test set and null set. The average number of enriched results in the null sets would be the false positives. In all the results shown in this paper, 100 null sets were used to compute the average, unless otherwise indicated. In the pipeline, available for download in Supplementary material, the number of null sets is an adjustable parameter. The ratio of false positives to predicted positives is the FDR.

F-measure Optimization and the Matthews correlation coefficient.

To evaluate F-measure and MCC, we started with evaluating true/false positive/negatives and the metrices derived from the true/false positive/negatives. The number of "predicted positive" is the number of GO terms found at a threshold. For an analytical method such as BH, the "false positive" would be (predicted positive) multiply by FDR, which is estimated by the corrected p-value. For resampling, the "false positive" would be the average number of GO terms found by null sets. The "true positive" is calculated by:

$$True\ Positive\ =\ (Predicted\ Positive)\ -\ (False\ Positive)$$

Then, we calculate the precision:

$$Precision = \frac{True\ Positive}{Total\ Positive}$$

Recall is defined as

$$Recall = \frac{True\ Positive}{Real\ Positive}$$

“Real Positive” is defined by

$$Real\ Positive\ =\ True\ Positive\ +\ False\ Negative$$

In the absence of the ability to calculate "False Negatives" directly, we estimate the number of real positives as the maximum true positive achieved across the range of possible p-values. This procedure is shown graphically in Figure 1 for the BH method of computing false positives, using as an example a gene list to be described in detail later in the paper. In this figure we plot predicted positives, false positives (False Discovery Rate x predicted positives), and true positives (predicted positives – false positives) vs. uncorrected *p*-value for the entire range of *p*-values from 0 to 1. At very lenient *p*-values the FDR approaches 1, resulting in the true positives approaching 0. It is difficult to evaluate false negatives and thus assign a number for "real positives", since a false negative is an object that escaped observation, and thus can't be counted directly. Yet such estimation is essential to applying F-measure. In our case, if we follow the trajectory of the true positives in Figure 2.1 as the threshold is relaxed, we see that at very stringent *p*-values all positives are true positives. As the threshold is relaxed further, more false positives are generated, so the predicted positive and true positive curves start to diverge. At $p = 0.13$ (a far higher value than would ordinarily be used as a cutoff) the true positives reach a maximum, and the number of true positives starts to decline as *p* is further relaxed. We utilize this maximum value as the maximum number of GO categories that can be possibly regarded as enriched in the data set; i.e., the number of real positives.

Based on precision and recall at each raw *p*-value cut-off, we can obtain a table and curve of F-measure vs uncorrected *p*-value. The F_1-measure is an equally weighted value of precision and recall. A generalized F-measure introducing the parameter β can be chosen based on the research question, whether minimization of type I (false positive) or type II (false negative) error, or balance between the two, is preferred, according to the equation:

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{\beta^2 Precision + Recall} \qquad Equation\ 1$$

The larger the magnitude of β the more the value of F_β is weighted towards recall; the smaller the value of β the more the value of F_β is weighted towards precision. Optimizing F-measure provides us a threshold which emphasize precision (β<1) or recall (β>1), or balance of both (β=1). Note that precision and recall are extreme values of F-measure; that is, Precision=F_0 and Recall=F_∞.

To compare the different thresholds, we also calculated for each of them the Matthews correlation coefficient (MCC) [28]. Originally developed to score different methods of predicting secondary structure prediction in proteins, the MCC has become widely used for assessing a wide variety of approaches to binary classification, as exemplified by the 2704 citations (at this writing) of the original paper. Perhaps even more telling, the citation rate for the seminal MCC paper has been increasing as the method is being applied in a greater variety of contexts, reaching 280 citations in 2017 alone.

In the expression below for the MCC, the True Negative (TN) is estimated using total number of GO categories in the database minus predicted positive and false negative.

$$MCC = \frac{TP \cdot FN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad \textit{Equation}\ 2$$

The MCC can be expressed in an equivalent expression using definition of informedness and markedness, which includes precision and recall, as well as the inversed precision and recall evaluating the proportion of true negatives:

$$invPrecision = \frac{True\ Negative}{True\ Negative + False\ Negative} \qquad \textit{Equation}\ 3$$

$$invRecall = \frac{True\ Negative}{True\ Negative + False\ Positive} \qquad \textit{Equation}\ 4$$

$$informedness = recall + invRecall - 1 \qquad \textit{Equation}\ 5$$

$$markedness = precision + invPrecision - 1 \quad \textit{Equation}\ 6$$

Combining Equations 2-6 and some algebra we find:

$$MCC = \sqrt{markedness \cdot informedness}\ \textit{Equation}\ 7$$

In an analogous fashion to the manner in which the F-measure may be generalized to weight either precision or recall more strongly by a variable β, so also the MCC can be generalized to more strongly weight either markedness or informedness by the expression

$$MCC = \sqrt[1+\beta]{markedness \times informedness^{\beta}} \qquad \text{Equation 8}$$

Data Sets

- *Environmental Stress Response (ESR)*

First dataset is the Yeast Environmental Stress Response (ESR) data [29], a robust data set for a model organism. The ESR set is list of genes commonly differentially expressed in response to environmental stresses such as heat shock, nutrient depletion, chemical stress, etc. Approximately 300 genes are up-regulated, and 600 genes are down-regulated in the ESR set. We expect this set to be “well-behaved” (give reasonable results with standard methods of analysis), since the data come from a very well annotated model organism subject to a widely studied experimental intervention.

- *Alarm Pheromone (AP)*

The second data set is comprised of human orthologs to the honey bee Alarm Pheromone set[30]. The Alarm Pheromone set is a list of genes differentially expressed in honey bee brain in response to the chemical alarm pheromone, which is a component of the language by which honey bees communicate with each other. Previous studies have shown that the Alarm Pheromone set is enriched in placental mammal orthologs, compared to other metazoans including non-social insect orthologs[31]. The Alarm Pheromone set is much smaller than the ESR set, with 91 up-regulated genes and 81 down-regulated genes. We expect the AP set to be not so “well-behaved” compared to the ESR set, as we are using model organism orthologs (human) to a non-model organism (honey bee) and the organisms diverged about 600 million years ago.

- *Random Test Sets*

To generate a baseline of the analysis for each data set using different FDR calculation methods, we have applied the pipeline to analyze randomly-generated sets as “test” set inputs, where FDR should equal to 1 for all uncorrected *p*-values.

The BH FDR curves are calculated in the following way: The R program p.adjust is applied to generate a list of analytically calculated FDR (BH) corresponding to uncorrected *p*-values for each “test” sets. Then the lists of FDRs are merged and sorted by uncorrected *p*-values. The FDRs are smoothed by a “sliding window” method: at each uncorrected *p*-value point, the new FDR is the average value of 11 FDRs centered by the uncorrected *p*-value point.

The Resampling FDR curves are calculated in the following way: The output uncorrected *p*-values are binned in steps of 1E-4. The counts below the upper bound of each *p*-value bin for the “test” set enrichment categories are the “Predicted positives”, and average counts for the null set enrichment categories are the “False Positives”. The process is repeated for the multiple “test” sets, and corresponding to each test set, 100 null sets were generated for “False Positive” calculation. Then the number of total and false positives are averaged, respectively. The FDR would be the quotient of the averaged total and false positives. Then, all the FDRs are plotted against the uncorrected *p*-values.

Results

In this section, we present the results of applying our methods to the two previously published sets of data introduced in the Methods section, the ESR set and the human orthologs of the Alarm Pheromone set. For both above data sets, we show the results from analyzing the genes using the biological process (BP) category of the gene ontology. These results will show 1) areas of agreement and difference between Benjamini-Hochberg and random resampling in evaluation of FDR, 2) how the assessment of significance of enrichment varies according to the particular database that is being probed, and 3) how the assessment of significance of enrichment varies according to the weight assigned to precision vs. recall.

ESR Set (Environmental Stress Response, yeast)

- *Benjamini-Hochberg (BH)*

Figure 2.2 shows the results of F-measure optimization on the ESR data based on FDR calculated by Benjamini-Hochberg (BH) method. As expected by their definitions, precision (F_0) decreases with increasing *p*-value while recall increases with increasing *p*-value. $F_{0.5}$ (precision-emphasized), F_1 (precision and recall equally weighted) and F_2 (recall-emphasized) all show relative maxima, providing a rational basis for assigning a threshold for significance. The horizontal scale is extended far enough to visualize the determination of the number of real positives. In the case of the up-regulated gene set, maximum F_1 occurs at an uncorrected *p*-value close to 0.05. In the case of the down-regulated gene set however, it appears that a much more stringent cutoff would be appropriate.

- *Resampling*

Figure 2.3 shows the results of F-measure optimization on the ESR data using resampling to calculate FDR. The false positives are calculated by average number of GO categories enriched in random sets. For the up-regulated set, all the F-measures optimize at much lower uncorrected *p*-values than do the F-measures calculated by the BH method. For the down-regulated set, resampling-calculated $F_{0.5}$ is optimized at a lower uncorrected *p*-value than BH method while F_1 and F_2 are optimized at slightly higher uncorrected *p*-value.

Comparing the results in Figure 2.2 and Figure 2.3 show that the optimum cutoff (as measured by maximum F_1) varies widely, depending on the gene set to be tested and the method for assessing FDR. Using BH the optimum cutoff is .0476 for upregulated ESR and .012 for downregulated ESR. Using resampling, the optimum cutoff is .0096 for upregulated ESR and .0126 for downregulated ESR.

Also, as expected, the optimum cutoff is relaxed when recall is emphasized (F_2 instead of F_1) and made more stringent when precision is emphasized ($F_{0.5}$ instead of F_1).

Alarm Pheromone Set (human orthologs)

- *Benjamini-Hochberg (BH)*

Figure 2.4 shows exactly the corresponding results as Figure 2.2, this time on the human orthologs to the honey bee alarm pheromone set. F-measures are maximized at much higher thresholds than for the ESR set. The difference in optimal F-measure is largely due to the different shapes of the recall curves. For the ESR set, precision drops significantly more rapidly with increasing uncorrected *p*-value than does the AP set. Therefore, a higher uncorrected *p*-value can be used for the latter set with essentially the same degree of confidence.

- *Resampling*

Figure 2.5 shows the number of GO categories and F-measures for the alarm pheromone set human orthologs using resampling method. The resampling method have found more false positives than BH, and therefore the precision is much lower than the precision calculated from BH, and the F-measures are optimized at lower uncorrected *p*-values than the F-measures calculated from BH.

From the above Figures 2.2-2.5, we can note the stepped structure in the number of enriched GO categories. The stepped structure lies in the fact that the number of genes associated with any GO category, in the test set or reference set, must be an integer with limited number of choices. Therefore, the uncorrected *p*-values calculated would be in a discrete set instead of a continuum. Consequently, the number of positives as a function of *p*-values increases in a stepped way. As a result, the F-measures derived from the number of GO categories have spikes. But as our graphs have demonstrated, the optimal F-measures reflect the different weights on precision and recall despite the spikes.

Comparing the results in Figures 2.4 and 2.5 shows that, for the AP gene sets as for the ESR gene sets, the optimum cutoff threshold is different for the upregulated and downregulated gene sets and also is different when BH is used to determine the FDR as compared to resampling.

Comparison of F-Measure with MCC for Optimization of Threshold Choice

As indicated in the section on methods, a widely used alternative to the F-measure for optimization is the Matthews Correlation Coefficient (MCC) which, unlike the F-measure, gives equal weight to negative as well as positive identifications. Figure 2.6 shows MCC optimization for exactly the same data set (ESR) and False Discovery Rate determination (Resampling) as in Figure 2.5. The most important lesson from this Figure is that the uncorrected p-value that maximizes MCC_1 is the same as the uncorrected p-value that maximizes F_1. Inspection of the formulas reveals the reason. The divergence between MCC and F-measure occurs only when the false negatives are a significant fraction of the total negatives. Since there are tens of thousands of terms in the gene ontology database this condition does not pertain to our situation, so optimization of the F-measure is an adequate strategy. However, we agree with Powers [26] that optimization of the MCC is the more universally correct strategy.

Comparison of FDR (False Positive) Calculation by Benjamini-Hochberg (BH) and Resampling

In the previous section, we have demonstrated how to use F-measure optimization to obtain a flexible threshold based on whether precision or recall is more heavily weighted by the researcher. In that section the FDR is calculated but not shown explicitly. The present section explicitly compares the FDR as calculated by the BH method and by random resampling. In each case the random resampling FDR is computed based on the average of 50 randomly sampled null sets of the same size as the test set. Figure 2.7 shows that for the ESR set, the BH method and resampling estimate similar FDR at low *p*-value. As the threshold increases, the BH method estimates lower false discovery rate, and therefore higher precision, than the resampling method at the same uncorrected *p*-value. By contrast, for the Alarm Pheromone set, the BH method estimates lower FDR than resampling.

To further evaluate the methods, we carried out multiple runs using random (null) sets as test sets. In this case, the FDR should in principle be 1, for any uncorrected *p*-value. The results of this test are shown in Figure 8a, where for each segment of *p*-values (bin size = 0.0001) we show the mean plus/minus the standard deviation. The resampling method passes the test on the average, but the results are noisy. The BH method systematically underestimates FDR. Figure 7b shows that the noise in the resampling method results in Figure 7a are largely due to the variation in the random null sets, and that the noise level in using random resampling for real data is acceptably low.

Statistical Summary of Results from Different Threshold Criteria.

Table 2.1 shows the statistical summary of using all different criteria for the distinction between significant and non-significant enrichment. Notable features of this table include: 1) Variation of the threshold within the range explored in this study made relatively little statistical difference for the ESR set. Over the entire range of thresholds, both the precision and the recall for the ESR set are good, and the number of terms returned does not change very much. 2) Variation of the threshold within the range explored in this study makes a very large difference in the results of the AP set. For the most stringent choice of threshold, the precision is high, but the recall is quite low. Relaxing the threshold improves the recall, but at a cost to the precision, so there is a distinct tradeoff between precision and recall, and 3) We discovered that optimizing F_1 is exactly equivalent to optimizing the Matthews correlation coefficient. F.5 is optimized at a lower uncorrected p-value than F1 while F2 is optimized at a higher p-value, and the same pattern is seen for MCC.

Identity of Enriched Terms Using Different Threshold Criteria.

- *Higher order relatively general terms.*

The enriched GO terms are categorized by their parent terms, 1st order parent being direct children of the root term “Biological Process” (GO:0008150), 2nd order parent being direct children of the 1st order parent terms. Each enriched GO term is traced back to the root by the shortest route. Tables 2.2 through 2.5 below provide an outline of the complete gene ontology results by showing the high order terms that are either themselves enriched according to the described criteria or have child terms enriched, or both. In each case the results from three different thresholds are shown, BH FDR<.05, optimum $F_{.5}$, and optimum F_1. The most striking pattern is that for the ESR sets (Tables 2.2 and 2.3), modifying the threshold within the parameters of this paper did not change the identity of the putatively enriched higher order terms very much. However, for the AP sets (Tables 2.4 and 2.5), relaxing the threshold caused a substantial increase in the number of high order terms judged to be putatively significant. However, from Table 1 is it seen that the precision (confidence) of the additional terms for the AP sets is substantially lower than for the terms returned using the most stringent threshold. Thus, for the AP set we clearly see that we can’t simultaneously have high precision and high recall. We must trade one for the other.

- *Relatively Specific Terms.*

Specific, or “child” terms returned in these calculations are too numerous to delineate completely in the body of the paper. They are instead provided in the spreadsheet “AllGOTermsInTree_Final (supplementary material 2.1)” Separate tabs delineate the returns from ESR upregulated, ESR downregulated, AP upregulated, and AP downregulated. Each entry in the spread sheet is color coded with the code given in the tab labeled “color coding”. Entries that are shaded are either primary or secondary (more general) classes, which will also be shown in Table 2.1. Entries colored in black appear at "standard" threshold: BH FDR<0.05. Entries colored in blue emerge at the threshold determined by optimal $F_{0.5}$. For AP Up, the standard threshold is the most stringent while for all other sets, the optimal $F_{0.5}$ is the most stringent. Entries colored in red first emerge at the least-stringent threshold for that data set, which corresponding to optimal F_1. The format of the spreadsheet for each of the data sets is as follows: Column A is the identifying number of the GO class that is returned as significant, column B is the name of that class, and column C is the raw enrichment p-value for that class. Column D is non-zero only for the rows belonging to primary or secondary GO classes (which are shown explicitly in Tables 2.2-2.5 for the four data sets). The numerical value in column D represent the smallest uncorrected *p*-value of all the classes under the primary or secondary class shown in that row. The spread sheet is organized to be sectioned off according to primary or secondary classes. To illustrate the sectioning, under the “AP up” is the primary class “cellular process” and immediately under that the secondary class “protein folding”. This is followed by more specific classes under “protein folding” such as “chaperone-mediated protein folding” and others. The columns E and farther to the right are GO numbers representing the lineage of the particular term in that row starting with the primary class and continuing to the particular term in that row.

Because the trade-offs with varying threshold are most clear with the AP sets, we select those now for discussion. One biologically interesting feature emerging from varying the threshold consists of the more specific GO classes emerging from general classes already identified with a more stringent threshold. For example, in the “AP up” set “protein folding” was identified as a secondary class of interest by virtue of a very strong enrichment score. On relaxing the threshold more specific “child” classes emerged, such as “chaperone cofactor-dependent protein folding”, “endoplasmic protein folding”, and others. While these more specific classes are identified with less confidence than the overall “protein folding” class they are subsumed into, they do provide the most likely subclasses within protein folding to be biologically meaningful. Similarly, under the secondary class of “signal transduction” more specific subclasses such as “ER-nucleus signaling pathway”, “stress-activated

MAPK cascade" and others emerge with modest threshold relaxation. This pattern is seen throughout the spreadsheet. Relaxing the threshold provides not only improved recall, but improved specificity, which will help in biological interpretation of GO enrichment results.

- *Summary*

In general, when thresholds are varied, a tradeoff can plainly be seen between precision and recall. When looking at the specific GO classes that are returned at different choices of threshold a second tradeoff emerges, between generality and specificity. As threshold is relaxed some more general terms are revealed, but the greater effect is that more specific terms are revealed within general terms that were suggested at more stringent thresholds. These specific terms can help to provide a more focused interpretation of the biological results.

Conclusions

In this work, we have addressed two issues with the commonly used methods in the GO enrichment analysis: the relationship between resampling vs. Benjamini-Hochberg theory for estimating false discovery rate, and the arbitrariness of the threshold for significance.

To consider resampling vs. Benjamini-Hochberg we made five independent comparisons. Four consisted of upregulated and downregulated genes separately for two different animal experiments. The fifth was an array of random gene lists (null sets). For the yeast ESR sets the two methods gave almost the same results for uncorrected p-value<.04 but diverged substantially for more relaxed *p*-values, with the BH underestimating the FDR. For the honeybee AP set the BH method underestimated the FDR significantly at all uncorrected *p*-values. For the random or null sets, we know that the correct FDR is 1, because there is no significance to the results. Yet for the null sets the BH method produced FDR<1 by a large margin for the full range of uncorrected *p*-values. By contrast the resampling method, although noisy, does not systematically deviate from 1 in its prediction of FDR for the null sets.

It is of interest to consider why the BH method, while very useful and successful in some cases, sometimes fails. It is understood that the method will always work when the true inferences are independent. Strictly speaking, this will not be true of Gene Ontology data since many genes belong in multiple Gene Ontology categories. However, Benjamini and Yekutieli[21] showed that the method was still valid for dependent hypotheses provided that the related hypotheses that failed the null test showed positive regression of likelihoods. Consideration of the tree-like structure of Gene Ontology data[32] shows that this is true to a great extent. The branches of the tree-like structure clearly show positive regression within each branch; if a child category is enriched a parent is more likely to be enriched, and vice versa. Thus, as long as the enriched classes fall along a few well-delineated branches of the Gene Ontology tree structure, BH will work well. This appears to be largely the case for the yeast ESR set at relatively stringent *p*-values, in which the experimental intervention activated well-defined and annotated pathways. Thus, for relatively stringent cutoffs the BH FDR works well for this data set. However, some genes are members of categories in multiple branches, compromising the positive regression criterion. In the ESR set at relatively relaxed thresholds, and for the AP set at all thresholds, many Gene Ontology categories in different branches but with overlapping gene membership are represented in the returned categories, so that both independence and the positive regression criterion are violated. These considerations tell us why BH fails dramatically for the completely null sets. Neither independence nor positive regression are satisfied, except sometimes completely accidentally.

For the issue of the arbitrariness of the threshold, we introduced optimization of F-measures so that both type I and II errors are considered. Unlike arbitrarily applied threshold of BH FDR<0.05 or uncorrected *p*-value<0.01 for any data set, the F-measure optimization approach provides a flexible threshold appropriate to the nature of the data set and the research question. If the data set is high in noise-to-signal ratio and the penalty for letting in false positive is high, we can choose to optimize F-measures weighing more on precision. If the data set fails to show much enrichment by commonly-applied methods, we can relax the threshold and extract the best information indicated by F-measure optimization.

A concern is that, because of the nature of the problem, we were forced to use a heuristic (albeit reasonable) method to estimate the false negatives, essential for calculating recall. We judge that this concern is more than offset by the advantage of enabling the replacement of an arbitrary threshold with F-measure optimization.

We found that for the particular class of problems dealt with in this paper the F-measure is as appropriate an optimization criterion as the Matthews Correlation Coefficient.

By examination of the specific GO categories that are returned by our analysis, we find that relaxing the threshold, we see revealed the most likely specific subcategories within the general categories that are revealed at the most stringent threshold. Thus, varying the threshold not only reflects the tradeoff between precision and recall, but also between generality and specificity.

In the supplementary material we present the spreadsheet **"AllGOTermsInTree_Final",** which shows all the specific GO terms returned in the work described in this paper. Also, in the supplementary material, we present our automatic pipeline integrating TopGO with resampling and analyzing functions to carry out the whole process of resampling, enrichment analysis, F-measure calculation, and representing results in tables and figures. The pipeline also includes a GOstats[15] module for easy analysis of under-represented terms and a STRINGdb[33] module for KEGG pathway terms. As demonstrated, the pipeline can also calculate analytical FDR including, but not limited to, the BH method.

In summary, we suggest replacing a fixed *p*-value for assigning a threshold in enrichment calculations with an optimal F-measure, which incorporates the well-established and well-defined concepts of precision and recall.

Abbreviations

GO: gene ontology; FDR: false discovery rate: BH: Benjamini-Hochberg method; BP: biological process; CC: cellular component; MF: molecular function; BY: Benjamini-Yekutieli method; ESR environmental stress response genes; AP: honey bee genes in response to Alarm Pheromone, human orthologs.

Ethics approval and consent to participate

N/A

Consent for publication

N/A

Availability of data and material

Computer codes available in Supplementary Material

Competing interests

The authors declare that they have no competing interests.

Authors contributions

WG and ZF both did parts of the calculation and worked together to initially develop the automated pipeline. WG did final enhancement and debugging. EJ suggested the overall direction of the work. WG wrote the first draft of the manuscript. All three authors worked on refining the manuscript

Acknowledgements

We gratefully acknowledge useful discussions with especially Dr. Enes Kotil and Santiago Nunez-Corrales and also other members of our research group in the Beckman Institute. Professor Saurabh Sinha of our Computer Science Department provided useful discussions and wise advice.

Funding

We gratefully acknowledge support from R01GM098736 from National Institute of General Medical Science to EJ.

1 Ashburner,M. et al. (2000) Gene Ontology: tool for the unification of biology. Nat. Genet., 25, 25–29.

2 Huang,D.W. et al. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res., 37, 1–13.

3 Eden,E. et al. (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics, 10, 48.

4 Huang,D.W. et al. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc., 4, 44–57.

5 (2016) g:Profiler—a web server for functional interpretation of gene lists (2016 update). Nucleic Acids Res., 44, W83–W89.

6 Conesa,A. et al. (2005) Blast2GO: A universal annotation and visualization tool in functional genomics research. Application note. Bioinformatics, 21, 3674–3676.

7 Zeeberg,B.R. et al. (2005) High-throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). BMC Bioinformatics, 6, 168

8 Al-Shahrour,F. et al. (2006) BABELOMICS: A systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res., 34.

9 Al-Shahrour,F. et al. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 20

10 Subramanian, Aravind, et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proceedings of the National Academy of Sciences 102.43 (2005): 15545-15550

11 Subramanian,A. et al. (2007) GSEA-P: A desktop application for gene set enrichment analysis. Bioinformatics, 23, 3251–3253.

12 Ballouz,S. et al. (2016) Using predictive specificity to determine when gene set analysis is biologically meaningful. Nucleic Acids Res., gkw957

13 Alexa,A. et al. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics, 22, 1600–1607.

14 Alexa,A. and Rahnenfuhrer,J. (2010) topGO: topGO: Enrichment analysis for Gene Ontology. R package version 2.18.0. October.

15 Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. Bioinformatics, 23, 257–258.

16 Maere,S. et al. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 21, 3448–3449.

17 Rivals,I. et al. (2007) Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics, 23, 401.

18 Dwass, Meyer. "Modified randomization tests for nonparametric hypotheses." The Annals of Mathematical Statistics (1957): 181-187

19 Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B, 289–300.

20 Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res., 36.

21 Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under depencency. Ann. Stat., 29, 1165–1188.

22 Bogomolov, Marina, et al. "Testing hypotheses on a tree: new error rates and controlling strategies." arXiv preprint arXiv:1705.07529 (2017).

23 Blüthgen,N. et al. (2005) Biological profiling of gene groups utilizing Gene Ontology. Genome Informatics, 16, 106–115.

24 Kim, Seon-Young, and David J. Volsky. "PAGE: parametric analysis of gene set enrichment." BMC bioinformatics 6, no. 1 (2005): 144.

25 Noreen, Eric W. Computer-intensive methods for testing hypotheses. New York: Wiley, 1989.

26 Powers,D.M.W. (2011) Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation. J. Mach. Learn. Technol., 2, 37–63

27 Rhee,S. et al. (2008) Use and misuse of the gene ontology annotations. Nat. Rev. Genet., 9, 509–515.

28 Matthews, Brian W. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme." Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2 (1975): 442-451.

29 Gasch,A.P. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell, 11, 4241–4257

30 Alaux,C. et al. (2009) Honey bee aggression supports a link between gene regulation and behavioral evolution. Proc. Natl. Acad. Sci., 106, 15400–15405.

[31] Liu,H. et al. (2016) Conservation in Mammals of Genes Associated with Aggression-Related Behavioral Phenotypes in Honey Bees. PLoS Comput. Biol., 12.

[32] Zhang, Bing, Denise Schmoyer, Stefan Kirov, and Jay Snoddy. "GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies." BMC bioinformatics 5, no. 1 (2004):

[33] Franceschini, Andrea, et al. "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration." Nucleic acids research 41.D1 (2012): D808-D815.

Tables

Data Set	Threshold	Uncorrected p-value	# enriched categories	Precision	Recall	MCC
ESR Up	BH FDR<0.05	0.00459	118	0.936	0.798	0.864
	RS opt $F_{0.5}$	0.0029	110	0.964	0.765	0.858
	RS opt F_1	0.0096	146	0.890	0.939	0.914
	Max MCC	0.0096	146	0.890	0.939	0.914
ESR Down	BH FDR<0.05	0.00689	211	0.948	0.883	0.914
	RS opt $F_{0.5}$	0.0016	185	0.989	0.808	0.894
	RS opt F_1	0.0126	251	0.902	1	0.948
	Max MCC	0.0126	251	0.902	1	0.948
AP Up	BH FDR<0.05	0.00116	57	0.807	0.0974	0.290
	RS opt $F_{0.5}$	0.012	246	0.600	0.312	0.429
	RS opt F_1	0.0636	699	0.416	0.615	0.500
	Max MCC	0.0636	699	0.416	0.615	0.500
AP Down	BH FDR<0.05	0.00138	58	0.759	0.353	0.517
	RS opt $F_{0.5}$	4.00E-04	44	0.909	0.321	0.540
	RS opt F_1	0.0073	146	0.534	0.626	0.577
	Max MCC	0.0073	146	0.534	0.626	0.577

Table 2.1. Precision, Recall, and Matthews Correlation Coefficients (MCC) at thresholds BH FDR<0.05, Resampling optimal F0.5, and Resampling optimal F1. For the four data sets examined, we have found that optimal F_1 is the position that MCC reaches maximum. This correspondence between optimum F_1 and optimum MCC was unanticipated but emerged from independent calculation of both quantities. For the ESR set, the MCC is high for all thresholds. For AP set, MCC is relatively low, and the MCC for BH FDR<0.05 is the lowest.

GO ID	Parent Term	Minimum raw *p*-value of child terms
GO:0008152	Metabolic Process (80,85,100)	3.40E-13
GO:0050896	response to stimulus (22,23,26)	7.40E-13
GO:0065007	biological regulation (4,5,7)	9.00E-05
GO:0009987	cellular process (4,5,13)	0.00035
**GO:0032502	developmental process (0,0,1)	0.00589

Table 2.2. ESR, Up-regulated Set Each row corresponds to a 1st order Parent Terms of enriched GO categories of ESR set, Up regulated genes. The three numbers in parentheses reflect the total number of terms in the Parent family (Parent plus children). We found no difference in the high order terms between BH FDR<.05 and $F_{.5}$. However, the developmental process parent term (labeled with "**") emerges when the threshold is increased to optimal resampling F_1. The groupings as defined by the parent terms do not change very much, but the number of more specific child terms increases moderately.

GO ID	Parent Term	Minimum raw *p*-value of child terms
GO:0008152	Metabolic Process (120,139,168)	1.00E-30
GO:0009987	Cellular process (6,6,7)	1.00E-30
GO:0071840	Cellular component organization or biogenesis (31,32,36)	1.00E-30
GO:0051179	Localization (21,22,22)	5.20E-28
GO:0065007	biological regulation (7,11,15)	3.20E-12
*GO:0050896	response to stimulus (0,1,2)	0.00357

Table 2.3. ESR, Down-regulated Set 1st order Parent Terms of enriched GO categories of ESR set, down regulated genes. For this data set the optimum $F_{.5}$ was more stringent than the BH FDR <.05. The term "response to stimulus" (labeled with "*" does not meet the optimum $F_{.5}$ criterion but does for the other two criteria. The numbers in the parentheses refer to the numbers of enriched terms in each parent category, ordered from low to high. As with the up-regulated genes, relaxing the threshold did not change the parent terms much, but did increase the number of more specific child terms moderately.

GO ID	**Parent Term**	**Minimal raw *p*-value of child terms**
GO:0009987	Cellular process (13,36,96)	1.10E-10
GO:0050896	Response to stimulus (57,71,119)	1.40E-08
GO:0065007	Biological regulation (28,113,288)	4.30E-05
GO:0008152	Metabolic process (9,44,113)	5.00E-05
GO:0032502	Developmental process (1,9,33)	0.00043
GO:0071840	cellular component organization or biogenesis (1,6,12)	0.00102
*GO:0051179	Localization (0,8,37)	0.00138
*GO:0022414	reproductive process (0,2,7)	0.00192
*GO:0002376	immune system process (0,2,8)	0.00504
*GO:0032501	multicellular organismal process (0,5,19)	0.00509
*GO:0040011	Locomotion (0,1,2)	0.00932
**GO:0051704	multi-organism process (0,0,11)	0.02
**GO:0008283	cell proliferation (0,0,2)	0.02962

Table 2.4. 1st order Parent Terms of enriched GO categories of AP set, Up regulated genes. The terms with “*” appears when the threshold is increased from BH FDR<0.05 (uncorrected *p*-value<0.00116) to optimal resampling $F_{0.5}$-measure (uncorrected *p*-value<0.012). Terms with “**” emerges when the threshold is increased to that for optimal resampling F_1(uncorrected *p*-value<0.0096). The number in the brackets refers to the number of enriched terms within each parent category at each threshold, ordered from low to high. Unlike the ESR sets, for this data set relaxing the threshold caused significantly greater returns in both general terms and their children.

GO ID	**Description**	**Minimal *p*-value of child terms**
GO:0008152	Metabolic Process (40,7,25)	3.20E-08
GO:0009987	cellular process (3,4,13)	7.00E-06
GO:0071840	cellular component organization or biogenesis (1,0,5)	7.90E-06
*GO:0051179	Localization (0,3,16)	0.00052
**GO:0065007	biological regulation (0,0,15)	0.00145
**GO:0050896	response to stimulus (0,0,7)	0.00174
**GO:0022414	reproductive process (0,0,1)	0.00441
**GO:0051704	multi-organism process (0,0,1)	0.00441
**GO:0032501	multicellular organismal process (0,0,3)	0.00441
**GO:0032502	developmental process (0,0,1)	0.00534

Table 2.5. 1st order Parent Terms of enriched GO categories of AP set, Down regulated genes. The terms with "*" disappears when the threshold is decreased from BH FDR<0.05 (uncorrected *p*-value<0.00138) to optimal resampling $F_{0.5}$-measure (uncorrected *p*-value<4.00E-4). Terms with "**" emerges when the threshold is increased at optimal resampling F_1(uncorrected *p*-value<0.0073). The number in the brackets refers to the number of enriched terms at each threshold, low to high. Unlike the ESR sets, for this set relaxing the threshold caused substantial increases in the putative enriched categories at both the general level and the more specific child level.

Figures

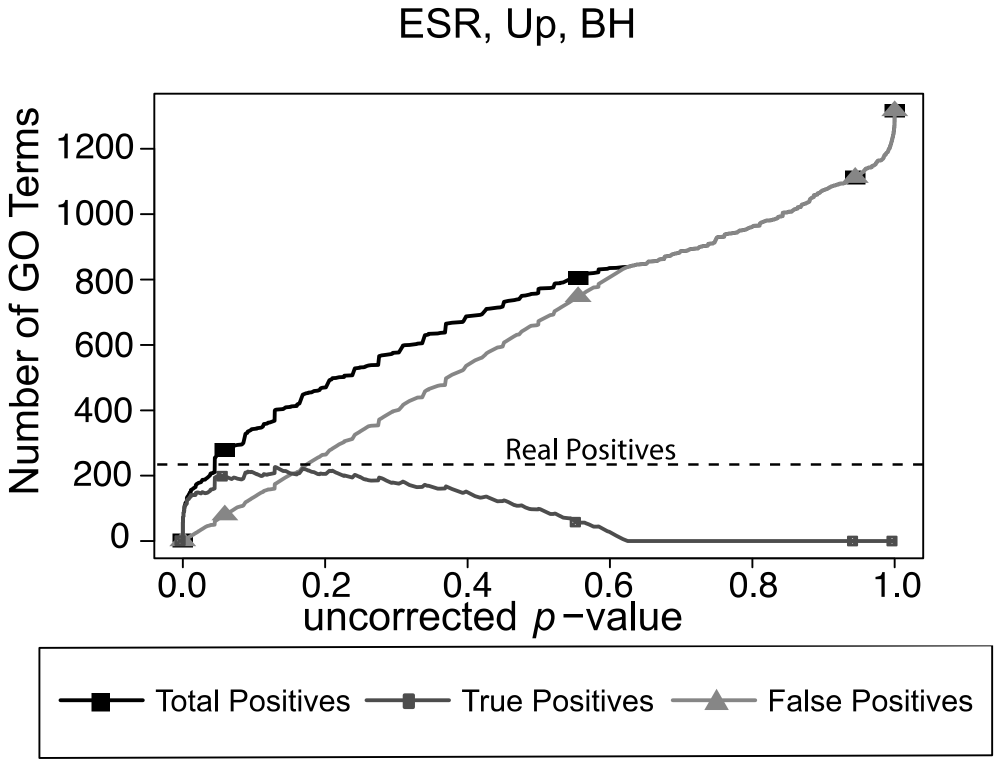


Figure 2.1. Number of positives for the yeast environmental stress response (ESR) set over the full range of uncorrected *p*-values from 0 to 1. “Predicted positives” is the number of Biological Process GO categories returned as a function of the *p*-value threshold for significance. “False Positives” is the number of predicted positives multiplied by the False Discovery Rate as calculated by the Benjamini-Hochberg formulation. “True Positives” is “Predicted Positives” minus “False Positives”. “Real Positives”, necessary to estimate number of false negatives, is estimated as the largest number of true positives computed at any uncorrected p-value.

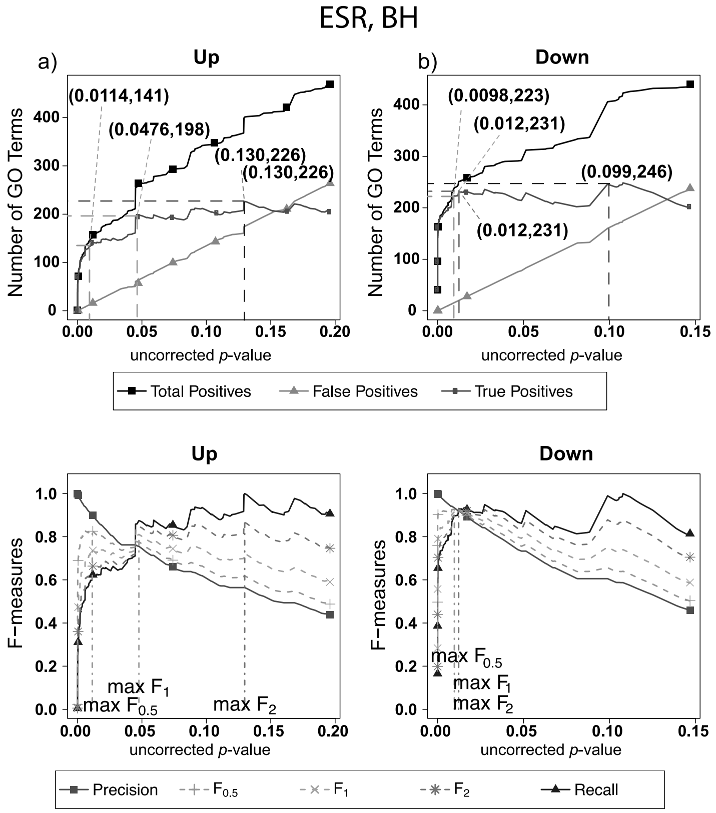


Figure 2.2. Number of positives and F-measure values for ESR set, BH-estimated FDR. a) Shows the number of enriched biological process Gene Ontology categories as a function of uncorrected *p*-value, the Benjamini-Hochberg number of false discoveries, and the projected true positives, namely the difference between the predicted positives and the false positives, for the upregulated ESR gene set. This panel is from the same data set at Figure 1. The number pairs in parenthesis are respectively (uncorrected *p*-value maximizing $F_{0.5}$, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_1, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_2, number of true positives at that *p*-value), (uncorrected *p*-value maximizing true positives, number of true positives at that *p*-value) b) is the same as a) for the downregulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of real positives, necessary to calculate recall (and therefore (F-measure)), is approximated by (predicted positives – false positives) $_{max}$. The *p*-value at which the computed true positives are a maximum is 0.13 for upregulated gene list (a) and at 0.099 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the p-value and number of returned GO terms at significant markers, specifically at maximum $F_{0.5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall), and Recall where we obtain an estimation of relevant elements by maximizing true positive).

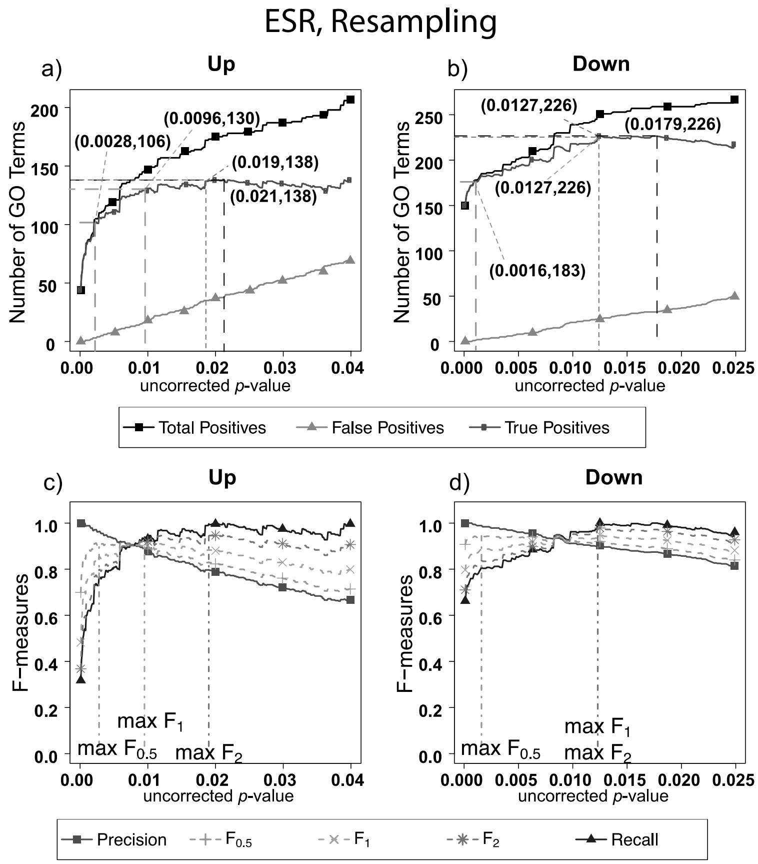


Figure 2.3. Number of positives and F-measure values for ESR set, Resampling-estimated FDR. a) Shows the number of enriched biological process Gene Ontology categories as a function of uncorrected *p*-value, the average number of enriched Gene ontology categories from the random set as the false positives, and the projected true positives, namely the difference between the predicted positives and the false positives, for the up-regulated ESR gene set. The number pairs in parenthesis are respectively (uncorrected *p*-value maximizing $F_{0.5}$, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F1, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_2, number of true positives at that *p*-value), (uncorrected *p*-value maximizing true positives, number of true positives at that *p*-value) b) is the same as a) for the down-regulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of real positives, necessary to calculate recall (and therefore (F-measure)), is approximated by (predicted positives – false positives) max. The *p*-value at which the computed true positives are a maximum is 0.021 for upregulated gene list (a) and 0.0179 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the p-value and number of returned GO terms at significant markers, specifically at maximum $F_{0.5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall), and Recall (where we obtain an estimation of relevant elements by maximizing true positive).

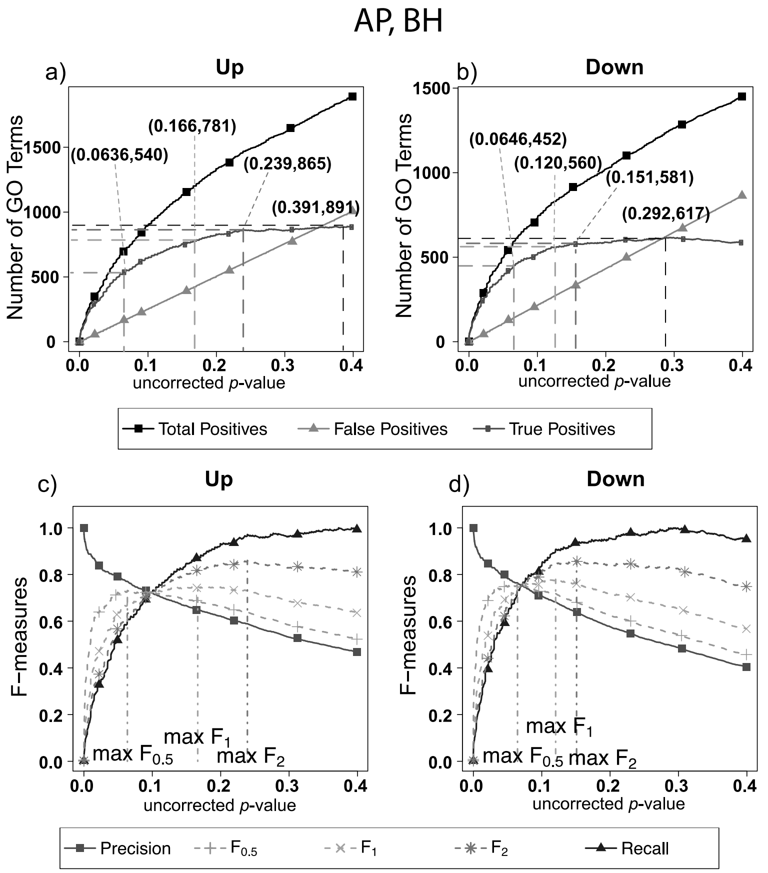


Figure 2.4. Number of positives and F-measure values for Alarm Pheromone set, BH-estimated FDR. a) shows the number of enriched biological process Gene Ontology categories as a function of uncorrected *p*-value, the Benjamini-Hochberg number of false discoveries, and the projected true positives, namely the difference between the predicted positives and the false positives, for the upregulated alarm pheromone human orthologs gene set. The number pairs in parenthesis are respectively (uncorrected *p*-value maximizing $F_{0.5}$, number of true positives at that *p*-value), (uncorrected p-value maximizing F_1, number of true positives at that *p*-value), (uncorrected *p*-value maximizing F_2, number of true positives at that *p*-value), (uncorrected *p*-value maximizing true positives, number of true positives at that *p*-value) b) is the same as a) for the downregulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of real positives, necessary to calculate recall (and therefore (F-measure)), is approximated by (predicted positives – false positives) max. The *p*-value at which the computed true positives are a maximum is 0.391 for upregulated gene list (a) and at 0.292 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the p-value and number of returned GO terms at significant markers, specifically at maximum $F_{0.5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall) and Recall (where we obtain an estimation of relevant elements by maximizing true positive).

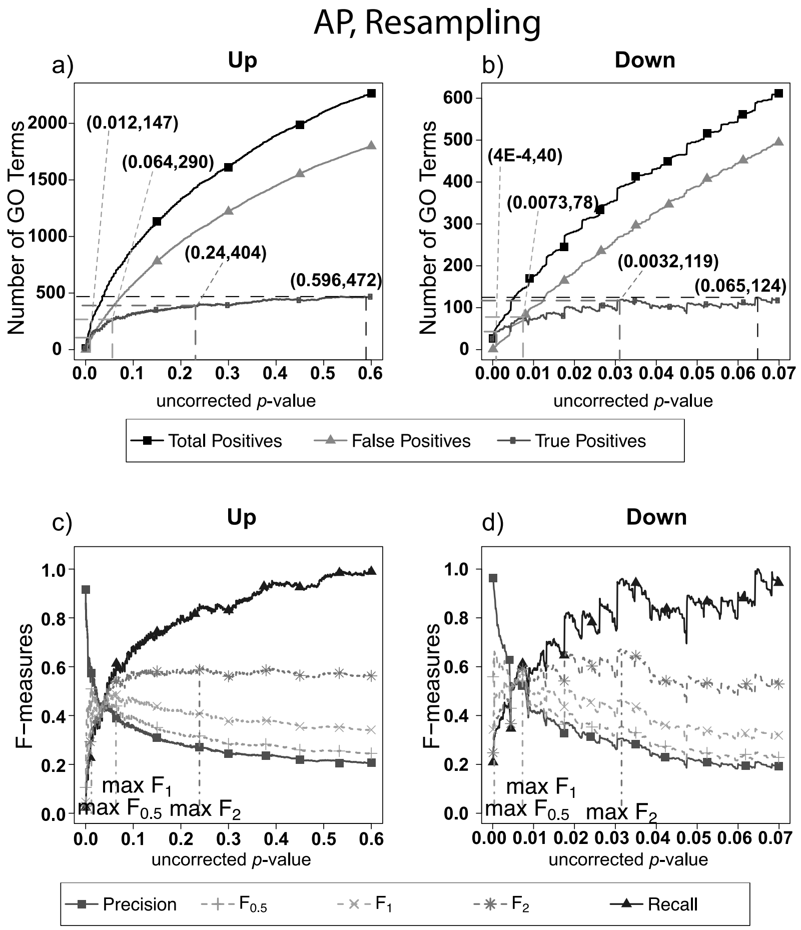


Figure 2.5. Number of Positives and F-measure values for AP set, Resampling-estimated FDR. The figure shows the number of enriched biological process Gene Ontology categories as a function of uncorrected *p*-value, the average number of enriched Gene ontology categories from the random set as the false positives, and the projected true positives, namely the difference between the predicted positives and the false positives, for the up-regulated alarm pheromone human orthologs gene set. b) is the same as a) for the down-regulated gene set. c) shows the F-measures computed from a) and d) the F-measures computed from b). Number of real positives, necessary to calculate recall (and therefore (F-measure)), is approximated by (predicted positives – false positives)$_{\max}$. The *p*-value at which the computed true positives are a maximum is 0.596 for upregulated gene list (a) and at 0.065 for downregulated gene list. (b) The pairs of numbers in parenthesis in a) and b) indicate the *p*-value and number of returned GO terms at significant markers, specifically at maximum $F_{0,5}$ (emphasizing precision), F_1 (balanced emphasis between precision and recall), F_2 (emphasizing recall), and Recall (where we obtain an estimation of relevant elements by maximizing true positive).

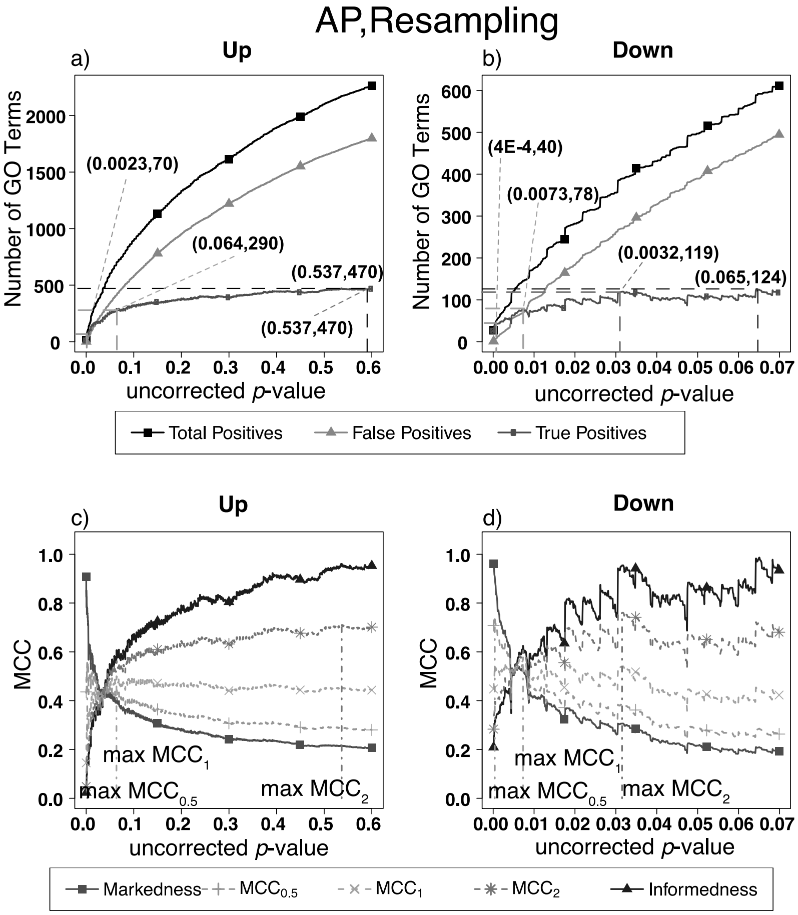


Figure 2.6. Number of Positives and MCC-measure values for AP set, Resampling-estimated FDR. This figure is the same as Figure 5 except that the optimization to determine significance-insignificance threshold is Matthews Correlation Coefficient (MCC) rather than F-measure. Note that the uncorrected p-value threshold for optimum MCC_1 is the same as for F_1. Examination of the expressions for the two quantities shows that the reason for the convergence is that in this case the number of false negatives is very small compared to the number of total and true negatives, so the fractional variation in true negatives is very small. This is true for all the data sets.

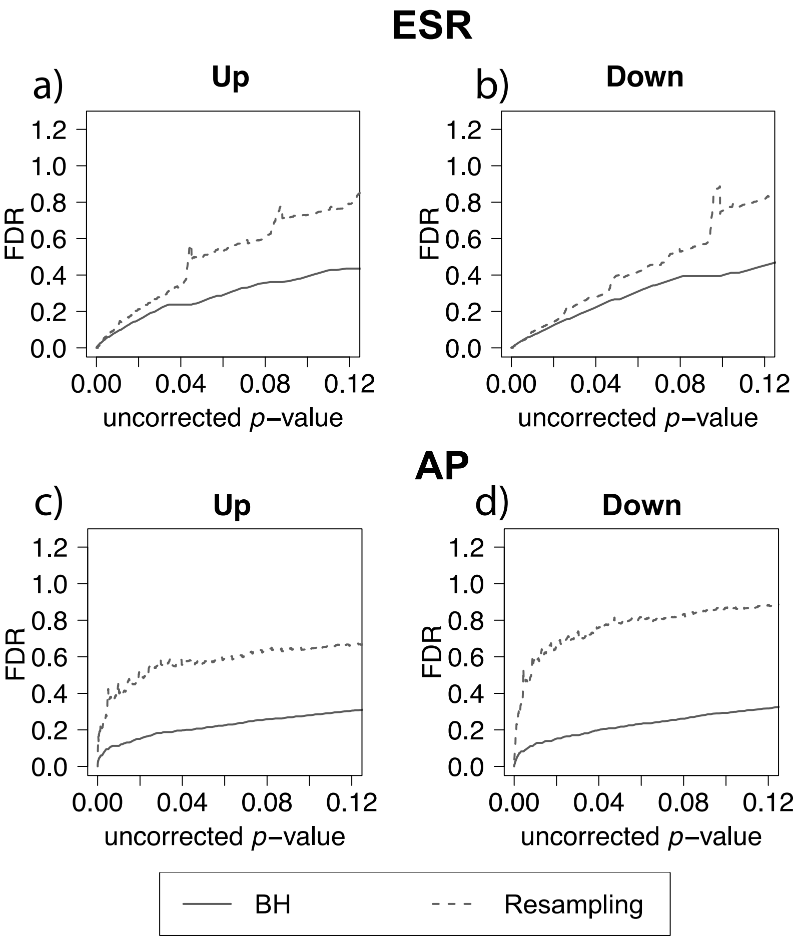


Figure 2.7. False discovery rate comparison. False discovery rate estimated by Benjamini-Hochberg (solid curve) and Resampling (dashed curve) for the ESR set and Alarm Pheromone set. Figure 7 compares the number of false discovery rate calculated by Benjamini-Hochberg (solid) and Resampling (dashed) in each set: a) up-regulated ESR, b) down-regulated ESR, c) up-regulated Alarm Pheromone set, and d) down-regulated Alarm Pheromone set. Generally, resampling has found higher false discovery rate than Benjamini-Hochberg. At low *p*-values, the BH and resampling methods get similar estimation of false discovery rate for the ESR set.

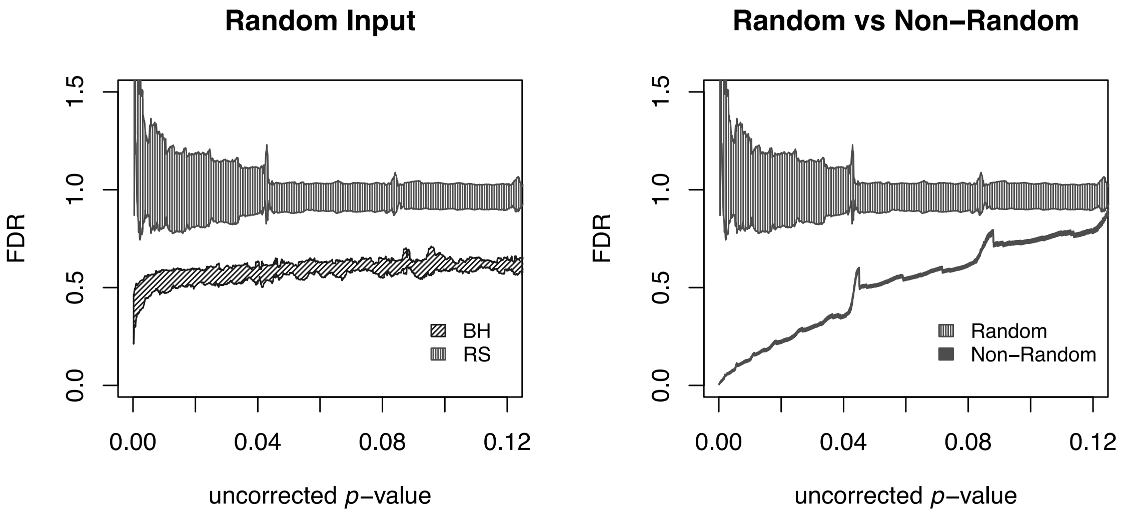


Figure 2.8. Comparison of different FDR calculation method on accuracy and convergence. a) Comparison of BH and Resampling on random "test" sets. At each *p*-value (*p*-values binned at intervals of .0001), the mean and standard deviation are calculated and plotted as shown. The random test sets consist of 281 yeast genes, against the background of the entire yeast genome. For each of the methods 50 test sets were used and the mean plus/minus standard deviation plotted as shown. Resampling hits the mark on the average but with substantial noise, while BH systematically underestimates FDR. b) Evaluation of resampling convergence on a real data set, ESR upregulated considered in this paper. This set is run against five different ensembles of null sets, each ensemble containing 100 null sets. The mean and standard deviation are plotted and compared to the results from the random test sets. It is seen that the noise of the resampling method on a real data set is acceptable.

Additional Files

Additional file 2.1--- AllGOTermsInTree_Final.xlsx

This is the spreadsheet showing all enriched terms at thresholds: BH FDR<0.05, optimal $F_{0.5}$, and optimal F_1. The terms are arranged by the primary and second-order parent terms.

Additional file 2.2 --- pipelinemanual .docx

"A TopGO- and GOstats-based automated pipeline for GO enrichment analysis using F-measure optimization based on resampling and traditional calculation"
This is a word document giving detailed description of how to run the pipeline for resampling or analytical FDR calculation and obtain thresholds maximizing F-measures

Additional file 2.3 --- pipeline.gz

This file contains source codes of the pipeline and the ESR and AP data sets for demo runs.

CHAPTER 3: Systems Biology Understanding of the Effects of Lithium on Affective and Neurodegenerative Disorders

(Submitted to *Frontiers of Neuroscience,* in review)

Weihao Ge and Eric Jakobsson*

Correspondence *Eric Jakobsson jake@illinois.edu

Abstract

Lithium has many widely varying biochemical and phenomenological effects, suggesting that a systems biology approach is required to understand its action. Multiple lines of evidence point to lithium intake and consequent blood levels as important determinants of incidence of neurodegenerative disease, showing that understanding lithium action is of high importance. In this paper we undertake first steps towards a systems approach by analyzing mutual enrichment between the interactomes of lithium-sensitive enzymes and the pathways associated with affective and neurodegenerative disorders. This work integrates information from two important databases, STRING and KEGG pathways. We find that for the majority of neurodegenerative disorders the mutual enrichment is many times greater than chance, reinforcing previous lines of evidence that lithium is an important influence on incidence of neurodegeneration. Our work suggests rational prioritization for which disorders are likely to be most sensitive to lithium and identifies genes that are likely to be useful targets for therapy adjunct to lithium.

Keywords: Lithium, systems biology, affective disorders, neurodegenerative disorders, biochemical pathways, biochemical networks

Introduction

Lithium is typically the first line therapy for bipolar disorder, including associated depression as well as mania.[1] A comprehensive review of the literature confirms that lithium is also effective against unipolar depression with unique anti-suicidal effectiveness, and may also be useful against cancer and neurodegenerative disease.[2]

Significant insights have been gained into the biochemical bases of lithium's action. The lithium-sensitive enzyme glycogen synthase kinase 3-beta (GSK3B)[3] inhibits signaling induced by Brain-Derived Neurotrophic Factor (BDNF).[4] Thus lithium would be expected to enhance activity of BDNF. BDNF may be a key bridge between affective and neurodegenerative disorders, since levels of this enzyme have been implicated in depression[5], bipolar disorder [6] [7], and dementia[8]. Indeed, in animal experiments, lithium was shown to induce brain-derived BDNF. [9] In addition, BDNF has been shown to play an important role in survival of adult and developing central neurons both in culture and in vivo.[10] [11] [12] [13] [14] The role of lithium in increasing activity of BDNF plus the role of BDNF in survival of neurons support the hypothesis that lithium might have a role to play in the treatment of neurodegenerative disease.[15]

Other reported research results have supported the potential of lithium for treatment of neurodegenerative disease.[16] However relevant clinical trials remain to be done. In the absence of clinical trial results, insights may be obtained from comparative studies on bipolar patients who have received long-term lithium treatment, and those who have not. In one such study, in an otherwise well-matched cohort of elderly (approximately 70 years old), 5% of those on long-term lithium therapy (continuous for the previous five years) were diagnosed with Alzheimer's disease (AD), while 33% of those not receiving consistent lithium therapy were diagnosed with AD.[17]

Epidemiological studies on the general population are suggestive. A recent nationwide study in Denmark showed that lithium level in the drinking water was significantly correlated with incidence of dementia, with higher lithium levels showing lower levels of dementia.[18] A more recent epidemiological study in Texas showed a similar specific effect for Alzheimer's disease.[19] An important feature of the epidemiological studies is that they involve levels of lithium ingestion that are many times smaller than those used for bipolar therapy, and are therefore almost certainly without significant side effects.

One neurodegenerative disorder, frontotemporal dementia (FTLD), initially presents with behavioral symptoms resembling mania,[20] posing a challenge for diagnosis. A definitive diagnosis in the early stage of the disease requires neuroimaging.[21] The consensus is that FTLD is invariably fatal, with a more rapid progression than Alzheimer's Disease.[22] However, there may be one documented apparent exception to the incurability of FTLD, in a case history presented by Monji et al.[23] In this study a middle-aged man presented manic symptoms that had no apparent origin in early life. Because imaging revealed abnormalities typical of FTLD, a diagnosis of FTLD was made. However, because the psychiatric symptoms had a pattern like bipolar disease, lithium therapy was begun. In a little under two years the psychiatric symptoms had been completely mitigated and new brain images appeared normal. The authors concluded that the initial diagnosis of FTLD was in error. However, the data presented in the paper were also consistent with the hypothesis that the FTLD diagnosis was correct and that the lithium therapy reversed the course of the disease. Dr. Monji, first author on the study, confirmed in an email to us that this hypothesis was consistent with their data.

A case history suggests efficacy of lithium for alleviating agitation and psychosis in both FTLD and Alzheimer's disease.[24] The efficacy of lithium for FTLD patients is to be tested in a recently announced clinical trial,[25] although only with respect to relief of the behavioral symptoms cited in the above reference over the course of a 12-week trial. The limited scope of this study is a continuation of a line of thought that considers affective and neurodegenerative aspects of FTLD as relatively separate[26], a line of thought that we question because of the evidence discussed above.

Dysfunction of autophagy is strongly implicated in neurodegenerative disease.[27] [28] [29] [30] Lithium has been shown to induce autophagy, due to its inhibition of inositol monophosphatase.[31] This is the basis of a pathway for autophagy enhancement, independent of the well-studied effects of mTOR on autophagy.[32] This additional pathway for autophagy enhancement has led to the suggestion of a combined lithium-rapamycin treatment for Huntington's Disease, with lithium inhibiting inositol monophosphatase and rapamycin inhibiting mTOR.[33]

The full range of lithium effects on autophagy is complicated,[34] as might be expected because of lithium's lack of specificity.

Because lithium affects many different biological molecules and processes[2], it is essential to utilize the tools of systems biology[35] if a comprehensive understanding of lithium action and its prospects for therapy are to be obtained. Important concepts for organizing biological information in a systems context are pathways and networks. A very useful tool for obtaining data about known pathways is the KEGG database.[36] An equally useful and complementary tool is the STRING database of interacting proteins.[37]

In the present paper we investigate further the possible linkages among 1) lithium, 2) affective disorders, and 3) neurodegenerative disorders by analyzing the mutual enrichment between the interactomes of lithium-sensitive enzymes, and the KEGG pathways associated with affective and neurodegenerative disorders.

Methods

Analysis was performed on the interactomes of lithium-sensitive genes, as identified by prior literature search[2]. This search suggested BDNF, BPNT1, DISC1, DIXDC1, FBP1, FBP2, GSK3A, GSK3B, inositol monophosphatases (IMPA1, IMPA2, and IMPAD1), INPP1, and PGM1 as key to understanding the broad biological actions of lithium. The interactomes of these genes were extracted from the STRING database (https://string-db.org). For each key gene, we adjust confidence level and order of neighbors (nearest only or next nearest included), so that each set contains a few hundred genes. This size is large enough for statistically reliable enrichment analysis. Table 1 shows the minimum confidence level and the maximum order of interaction (direct, removed by one, etc.) for each set. Very similar sets were merged; in particular FBP1 and FBP2 were merged into one set, and the inositol monophosphatases were merged into one set. On the other hand, GSK3A and GSK3B showed sufficient differences to be considered separately. Overall, we consider 10 distinct lithium-sensitive entities.

Gene	Confidence level	Order of Neighbor	Interactome Size
BDNF	0.4	1	335
BPNT1	0.6	2	388
DISC1	0.8	2	113
DIXDC1	0.6	2	378
FBP1	0.9	2	175
GSK3A	0.4	1	307
GSK3B	0.4	1	225
IMPAD	0.9	2	504
INPP1	0.7	2	228
PGM1	0.4	1	176

Table 3.1 Interactome parameters and sizes for lithium-sensitive genes.

Disease Association

We used the R-package KEGGgraph[38 39] to identify the genes associated with the pathways of interest. For one condition, bipolar disorder, there was no annotated pathway in the KEGG database. In lieu of an annotated pathway, we used the list of bipolar-related genes compiled by Nurnberger et al[40].

Empirical *p*-value calculation

The fundamental question we address is whether there is significant overlap or mutual enrichment between the interactomes of lithium-sensitive genes and the pathways or gene sets implicated in affective and/or neurodegenerative disorders.

For each of the 10 lithium sets, an ensemble of 1000 null sets are generated by random selection from the human genome. Each null set is the same size as the corresponding lithium set. Then we used the R-package STRINGdb[41] to perform KEGG pathway enrichment analysis. This operation is a particular example of the powerful technique of gene-annotation enrichment analysis.[42] In gene-annotation enrichment analysis a test list of genes (often derived from gene expression experiments) is compared to an organized database of gene annotations, often referred to as a gene ontology[43], an array of gene lists corresponding to different biological functions, molecular functions, or locations in the cell. The output of the gene-annotation enrichment analysis is expressed as the likelihood that the list overlaps could have occurred by chance (p-value). A very low p-value implies that the degree of overlap is highly significant statistically and very likely is significant biologically. In our study the gene lists we are comparing are the interactomes of lithium sensitive enzymes on the one hand, and KEGG pathways or otherwise derived lists associated with neural disease on the other hand. For each KEGG term retrieved, a null distribution of uncorrected *p*-value is generated by the 1000 null sets. This gives us a measure of the false discovery rate, since any overlap between the null sets and the KEGG pathways is purely accidentally. Then the fraction of null set uncorrected *p*-values smaller than or equal to the lithium-sensitive set uncorrected *p*-value would be the empirical *p*-value. For a detailed discussion of empirical p-value determination see Ge et al[44].

Key Gene Prediction

We predict key genes by counting how many times a gene appears in the cross section of interactomes and pathways associated with a particular disease. Then the counts are normalized by number of pathways associated with each disease. In this way, we predict which genes might be robust in disease-related pathways. Then, the genes are scored by the sum of mean counts over all diseases. A higher ranking indicates a gene would be associated with an important factor in many diseases.

Results

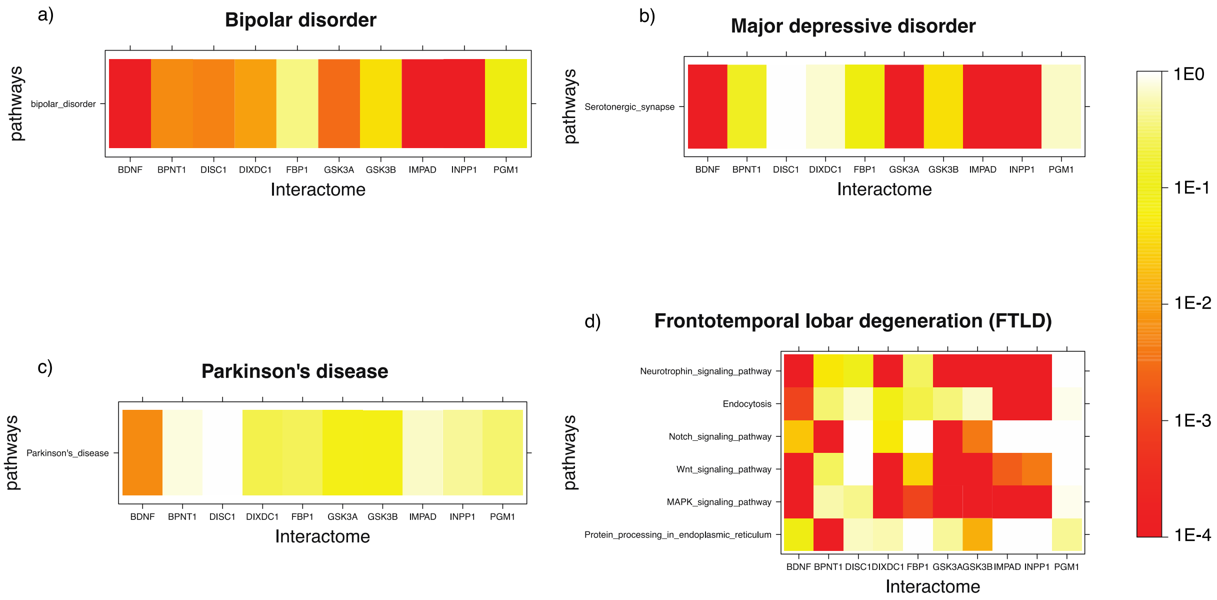


Figure 3.1. Heatmap for Lithium-sensitive enzyme interactome enrichment in disease-related pathways. The empirical enrichment p-value was calculated for each set of disease-associated genes. a) and b) are disorders where lithium treatment has proved to be effective and both show high enrichment. c) and d) are diseases where the effect of lithium treatment is unknown. c) shows relatively low enrichment while d) shows high enrichment.

Figure 3.1 shows a few examples of mutual lithium interactome enrichment with specific disease pathways, represented by heatmaps. Each area on the heatmap is a color-coded representation of the degree of mutual enrichment between the genes in the interactome of the indicated lithium sensitive enzyme and the genes in the indicated pathway. The darker the shade, the more significant the mutual enrichment of the interactome-pathway combination is. Fig.3.1 a) and b) shows two diseases where lithium treatments have been effective, and both show very strong enrichment. Fig.3.1 c) and d) shows two diseases where the effect of lithium treatment has not been explored. Parkinson’s disease, Fig.1 c), shows low enrichment while FTLD, Fig.3.1 d), shows high enrichment. We infer that FTLD is a more likely disease target for lithium treatment than Parkinson’s Disease. A spreadsheet providing p-values for the mutual enrichment of the lithium sensitive interactomes and the relevant pathways for all 112 diseases studied are provided in supplementary material.

For each of the 112 diseases considered we computed the geometric mean of the inverse of the p-values for each interactome/pathway enrichment and propose this as a “lithium sensitivity index” for the disease. The lithium sensitivity index is just the reciprocal of the mean p-value for each of the mutual enrichments between lithium-sensitive interactomes and the relevant pathways, where the mean p-value is:

$$p_{mean}=1/((1/p_1)x(1/p_2)x(1/p_3)x\cdots\cdots x(1/p_n))^{1/n} \qquad \text{Equation (3.1)}$$

We note that the individual p-values vary by several orders of magnitude. The method of averaging in Equation (3.1) ensures that both strong and weak enrichments contribute significant weight to the mean. Note also that our method is bounded at the low end of p-values by the number of null samples it is reasonable to compute, given compute time constraints. For ten thousand null sets as used in this paper, the lowest p-values are not numbers but the expression <1E-4, which means that the degree of mutual enrichment was greater for the test set than for all ten thousand of the null sets. For computing the inverse of the lowest p-values, we set the inverse at 1E+4.

Table 3.2 shows the top 34 ranked diseases out of the 112. Note that two diseases for which lithium is known to be effective therapy, bipolar disorder and major depression disorder, rank high, 9 and 29 respectively. Other notable diseases shown in Table 3.2 include Alzheimer's (20 out of 112), for which there is epidemiological evidence[17 above], FTLD (30 out of 112) for which there is evidence via case history[23], and schizophrenia (22 out of 112) for which there is some evidence of efficacy as an adjunct to antipsychotics.[45] Scores for all 112 diseases are provided in supplementary material.

Disease	**Sensitivity index**	**Mean p-value**
Dravet syndrome	1718.943899	.0006
HTLV1-Associated Myelopathy (HAM)	626.8531541	.0016
Congenital pain insensitivity with anhidrosis	466.9837537	.0021
Hemorrhagic destruction of the brain, subependymal calcification, and cataracts	418.143026	.0024
Rasmussen encephalitis	293.1481892	.0034
Lattice corneal dystrophies (LCD)	263.6451883	.0038
Subependymal giant cell astrocytoma	246.0470815	.0041
Bipolar Disorder	239.2876	.0042
Familial episodic pain syndrome (FEPS)	231.5937968	.0043
Familial exudative vitreoretinopathy (FEVR)	205.3474156	.0049
Focal dermal hypoplasia	205.3474156	.0049
Choroid plexus papilloma	198.0197413	.0051
Juvenile-onset dystonia	183.6715435	.0054
Prion diseases	175.0031999	.0057
Axenfeld-Rieger syndrome (ARS)	169.1174332	.0059
Congenital stromal corneal dystrophy (CSCD)	169.1174332	.0059
Ring dermoid of cornea	169.1174332	.0059
Stapes ankylosis with broad thumb and toes	169.1174332	.0059
Benign familial neonatal and infantile epilepsies	153.4488999	.0065
Alzheimer's disease	148.6362283	.0067
Neurosis	132.0111986	.0076
Schizophrenia	132.0111986	.0076
Pituitary adenomas	123.9488444	.0081
Febrile seizures	108.3195689	.0093
Episodic ataxias	104.5457633	.0095
Familial or sporadic hemiplegic migraine	104.5457633	.0095
Cerebral amyloid angiopathy (CAA)	89.36992044	.0112
Major depressive disorder	89.36992044	.0112
Epileptic encephalopathy with continuous spike-waves during slow-wave sleep	87.85098473	.0114
Frontotemporal lobar degeneration (FTLD)	75.82609324	.0132
Cerebral palsy	72.92882566	.0137
Generalized epilepsy and paroxysmal dyskinesia (GEPD)	69.3705138	.0144
Amyotrophic lateral sclerosis (ALS)	67.25422275	.0149
Fleck corneal dystrophy (FCD)	63.08690002	.0158

Table 3.2. Top 34 neuron-related disease by lithium sensitivity.

As a control on our methods, we compared the statistical distribution of scores for neural disease with corresponding scores for metabolic pathways (also from KEGG), and with random gene sets (null sets). This comparison is shown in box plots in Figure 3.2. As expected the scores for the null sets are quite low, collapsing into a range between 1 and 2.05. The scores for the metabolic pathways are also low, reflecting fact that lithium has not been found to be major modulator of metabolism. Just two metabolic pathways account for the height of the upward extension of the metabolic box plot, carbohydrate metabolism and nucleotide metabolism. On the other hand, the scores for neural diseases are quite high. These scores, together with large numbers of cell, animal, and epidemiological studies suggesting lithium may play a role in ameliorating this class of disease, suggest moving forward into clinical trials for selected affective and neurodegenerative disorders. Even in studies in which lithium is not the primary variable, environmental lithium should be measured and correlated with outcomes or used as an experimental variable, because of the possibility that lithium and another drug may be synergistic. For example, lithium and rapamycin stimulate autophagy by independent pathways, leading to a suggestion that they might be a promising combination therapy for Huntington's disease.[46]

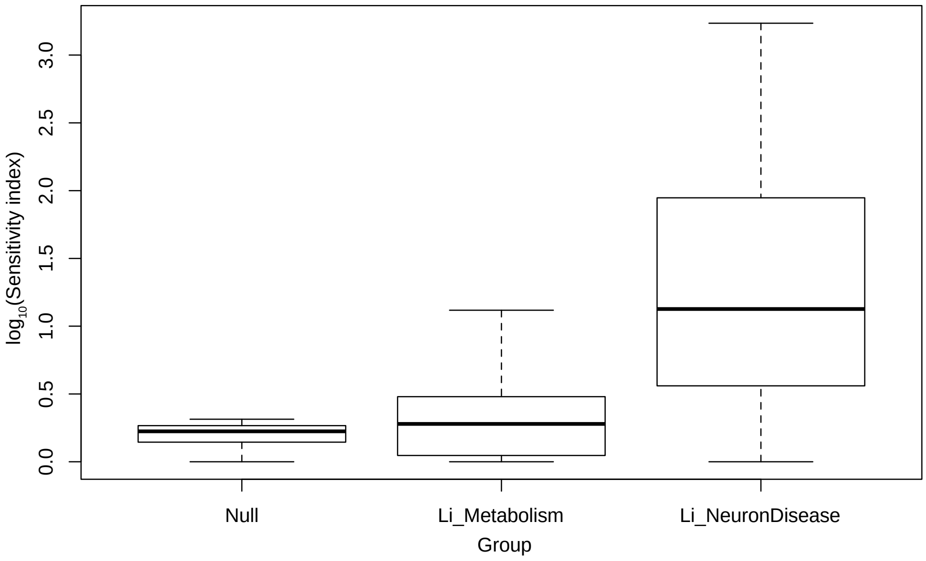


Figure 3.2 Log$_{10}$ of sensitivity index of lithium-sensitive interactome for null sets, metabolic pathways, and pathways associated with neural disease.

	Schizophrenia	Bipolar	AD	ALS	FTLD	Prion	MDD	Sum
MAPK3	0	0	**4**	0	**1.33**	**4**	**4**	13.33
APP	0	0	**6**	0	0	0	**6**	12
TP53	0	0	0	**5**	**2.5**	0	0	7.5
RAC1	0	0	0	**4**	**2**	0	0	6
PSEN1	0	0	**4**	0	**2**	0	0	6
PLCB3	0	0	**3**	0	0	0	**3**	6
PLCB2	0	0	**3**	0	0	0	**3**	6
PLCB1	0	0	**3**	0	0	0	**3**	6
PPP3CC	0	0	**3**	**3**	0	0	0	6
PRKACB	0	0	0	0	0	**3**	**3**	6
ITPR1	0	0	**3**	0	0	0	**3**	6
PPP3CA	0	0	**3**	**3**	0	0	0	6
PRKACG	0	0	0	0	0	**3**	**3**	6
NOS1	0	0	**3**	**3**	0	0	0	6
PLCB4	0	0	**3**	0	0	0	**3**	6
PRKACA	0	0	0	0	0	**3**	**3**	6
CYCS	0	0	**3**	**3**	0	0	0	6
HTR2A	**1**	**2**	0	0	0	0	2	5
NOTCH1	0	0	0	0	0	**5**	0	5
GAPDH	0	0	**5**	0	0	0	0	5
MAP2K1	0	0	0	0	**0.67**	**2**	**2**	4.67
BAX	0	0	0	**2**	**0.67**	**2**	0	4.67
GRM1	**1.5**	**3**	0	0	0	0	0	4.5
TNF	0	0	**2**	**2**	0	0	0	4
GNAQ	0	0	**2**	0	0	0	**2**	4
GNG2	0	**2**	0	0	0	0	**2**	4
ITPR3	0	0	**2**	0	0	0	**2**	4
CDK5	0	0	**4**	0	0	0	0	4
FYN	0	0	0	0	0	**4**	0	4
IL1B	0	0	**2**	0	0	**2**	0	4
ITPR2	0	0	**2**	0	0	0	**2**	4
PRKCA	0	0	0	0	0	0	**4**	4
PPP3CB	0	0	**2**	**2**	0	0	0	4

Table 3.3. Gene counts normalized by pathway number for genes appearing at intersection of interactomes and pathways. Table truncated for ease of display. Full table in supplementary material.

In addition to pathways we examined our results to identify specific genes within the lithium sensitive interactomes that may be important in modulating lithium effect on disease are useful to identify. Table 3.3 indicates the genes that occur with the greatest frequency at the intersection of the lithium-sensitive interactomes and pathways associated with selected neural diseases. The complete tally for all 112 diseases considered in this study is provided in supplementary material. We suggest that genes that appear prominently at the intersection of lithium sensitivity and neural disease pathways, and their promoter regions, should receive attention as possible sites of important mutations affecting lithium response to neural disease, and possibly as targets for drugs more specific than lithium. This is in addition to the ten lithium-sensitive genes that were used as a starting point for this study, based on their previous mentions in the literature.

The genes in Table 3.3 are ranked by total number of appearance across the diseases. The high rank indicates that a gene might be 1) found associated with multiple diseases or 2) associated with multiple interactomes for a particular disease-associated pathway. For the former case, the gene might indicate similar mechanisms for the multiple diseases. For the latter case, the gene would be a promising target in treatment of that particular disease.

For example, Table 3.3 have shown that MAPK3 is a shared gene by Alzheimer's Disease (AD), Prion, Major Depressive Disorders (MDD), and Frontotemporal lobar degeneration (FTLD), indicating that these diseases might have some shared mechanism. MAPK3 appeared in 4 interactome-pathway cross-sections in AD, Prion and MDD, and on average associated with 1.33 interactome-pathway cross-section in FTLD. MAPK3 is an essential component of the MAP signal transduction pathway that carries signals from cell surface to the nucleus. In analysis of normal as compared to AD brain tissue, MAPK3 is one of a small number of genes found to have alternative promoter usage and splicing.[47]

Another prominent gene in Table 3.3 is APP (amyloid precursor protein), which gives a strong signal in both AD (Alzheimer's Disease) and MDD (Major Depressive Disorder). Published studies implicate specific mutations in APP in incidence of AD [48], and implicate amyloid beta, the cleavage product of APP, in incidence of MDD.[49] Lithium has well established efficacy in the treatment of MDD,[50] and regulates the production of amyloid beta.[51] Taken together these findings suggest that influence of lithium on APP may be a common mode of action of lithium effect on both Alzheimer's disease and major depressive disorder.

Summary and Discussion

We have conducted a pathway and network analysis of the role of lithium in 122 neurodegenerative and affective disorders. We have found that for the large majority of such disorders, there is high mutual enrichment between the interactomes of lithium-sensitive enzymes and the pathways associated with those diseases, indicating that lithium is very likely to affect the course of the disease. We have also identified specific genes that exist frequently at the intersection of lithium-sensitive interactomes and neural disease pathways, suggesting these genes as possible targets for more specific drugs than lithium.

We hope that the results described in this paper and more detailed supplementary material will contribute to prioritizing and designing clinical trials of lithium for neural disease. To provide context for such prioritization and design, it is essential to take into account the ways in which lithium is unique, both as a pharmaceutical and as an ion that is ubiquitous in the environment, and therefore ubiquitous in the water and food we ingest[2]:

1. Unlike other ions, lithium is not regulated by selective membrane transport processes. Therefore, lithium concentration in both extracellular and intracellular compartments, rather than being roughly constant, is roughly proportional to lithium ingestion.
2. Unlike other pharmaceuticals, lithium is wildly nonselective in its biochemical effects. The major underlying mechanism for the lack of selectivity is lithium's general propensity to inhibit the many enzymes that have magnesium as a cofactor.
3. Unlike other pharmaceuticals, lithium is an essential nutrient. The question with lithium is not whether it should be ingested or not, but rather how much. Extreme lithium deprivation results in failure to thrive, while too much lithium is toxic.

In the light of all these factors, we suggest that the correct question to ask with respect to lithium and a particular disease is not, "Should lithium be administered for this particular disease?" but rather, "What is the optimum blood level of lithium for this individual, given his or her disease history, status, and genetic propensities?" Unlike other pharmaceuticals that are far more specific and inhibit or activate one or a small number of genes, the model for lithium action is that it alters the balance between a large number of interacting processes and pathways. Thus, a dose-response curve for lithium is likely to be highly nonlinear and not always monotonic.

There are just a few well-established markers for optimum concentrations. For a patient with a reliable diagnosis of bipolar disorder a common target would be 0.8-1 mM. Significantly higher concentrations will result in acute toxicity, while significantly lower will result in loss of effectiveness. Epidemiological studies on bipolar patients who are, and are not, on lithium therapy suggest that this level also protects against Alzheimer's disease. However, this level has some side effects when sustained for years or decades, namely an increased risk of kidney damage and lowered thyroid activity. Thus, for other conditions one would like to find lower effective concentrations; indeed one would like to do that for bipolar disorder as well, perhaps by combining lithium with other mood stabilizers that act in a synergistic fashion, enabling the lithium dose to be reduced.

At the other end of the dosage scale, epidemiological evidence is compelling that geographical variations in concentration of lithium in the drinking water are correlated with incidence of Alzheimer's; the lower the lithium the higher the incidence of mania. It thus seems that for Alzheimer's, an optimum level of blood lithium would be higher than the naturally occurring range, but perhaps lower than the therapeutic dose for bipolar disorder in order to minimize possible side effects of the bipolar therapeutic dose.

Another important marker is provided by a study showing that over a four-year period a lithium level of .25-.4 mM of lithium (1/3 to 1/2 of the bipolar therapeutic dose) did not incur any renal damage[52]. This study suggests that clinical studies exploring low to medium-dose lithium could be undertaken with relatively minimal concerns for side effects.

One of the authors (EJ) is an elderly person (79) and has found the evidence cited above sufficiently compelling that he self-administers lithium calibrated to a blood level of .3-.4 mM, in order to reduce the pace of age-related neurodegeneration. His outcome, however important it may be to him personally, has no statistical significance. We need a clinical study involving many subjects addressing the same question.

In general, it seems clear that whatever other studies are undertaken with respect to affective and neurodegenerative disorders, lithium blood levels should be monitored, since even geographical variations may have significant effects. The cost of adding lithium level to the routine blood tests is minimal, especially compared to the potential benefits. Beyond that, multiple studies should be undertaken in which low- to moderate-level lithium supplements are administered, since these are likely to be safe (although of course side effects should be monitored).

Perhaps our results, especially as scored in Table 1 and combined with other considerations, might help to focus on which neurodegenerative diseases might be most useful to consider for lithium therapy. Other considerations might be: 1) whether the disease impacts a large number of people, so that alleviating the condition would relieve much suffering, 2) the age at which the condition strikes, considering that the impact on individual, family, and others may be more if the disease strikes at a younger age, 3) the mortality rate, and 4) how rapidly the disease progresses, since the more rapidly progressing the disease the more rapidly meaningful statistics may be gathered from an intervention trial.

Many conditions that score highly in Table 1 might be usefully considered. One condition that looms large to us, because of the loss of a person close to us at the age of 46, is FTLD. The mean p-value for FTLD pathway mutual enrichment with lithium-sensitive interactomes is .0132, which is highly significant. While not as common as Alzheimer's, FTLD is not rare. Estimated lifetime risk is 1/742, so many millions of people each year die of FTLD.[53] The ratio of official incidence to mortality is 0.97; it is generally accepted to be 100% lethal. Life expectancy after diagnosis depends on the variant, but ranges from 3 to 9 years, so progression is much more rapid than Alzheimer's, and permits meaningful statistical analysis of any clinical trial in a relatively short time. Age of onset is most typically middle- to late middle-age when the individual is still employed and a crucial part of nuclear and extended family, in contrast to typically later onset of Alzheimer's. We have noted earlier in this paper that the initial symptoms of FTLD are sufficiently similar to mania (which is treated successfully with lithium) to sometimes lead to confusing diagnoses, which may suggest a common underlying biochemistry.

We will be happy to collaborate on further specific pathway or network analysis relevant to any of the neural diseases for which lithium may be.

[1] Post, Robert M. "Treatment of bipolar depression: evolving recommendations." *Psychiatric Clinics* 39, no. 1 (2016): 11-33.

[2] Jakobsson, Eric, Orlando Argüello-Miranda, See-Wing Chiu, Zeeshan Fazal, James Kruczek, Santiago Nunez-Corrales, Sagar Pandit, and Laura Pritchet. "Towards a Unified Understanding of Lithium Action in Basic Biology and its Significance for Applied Biology." *The Journal of membrane biology* 250, no. 6 (2017): 587-604.

[3] Freland, Laure, and Jean-Martin Beaulieu. "Inhibition of GSK3 by lithium, from single molecules to signaling networks." *Front Mol Neurosci* 5 (2012).

[4] Mai, Lian, Richard S. Jope, and Xiaohua Li. "BDNF-mediated signal transduction is modulated by GSK3β and mood stabilizing agents." *Journal of neurochemistry* 82, no. 1 (2002): 75-83.

[5] Karege, Félicien, Guillaume Perret, Guido Bondolfi, Michèle Schwald, Gilles Bertschy, and Jean-Michel Aubry. "Decreased serum brain-derived neurotrophic factor levels in major depressed patients." *Psychiatry research* 109, no. 2 (2002): 143-148.

[6] Cunha, Angelo BM, Benicio N. Frey, Ana C. Andreazza, Júlia D. Goi, Adriane R. Rosa, Carlos A. Gonçalves, Aida Santin, and Flavio Kapczinski. "Serum brain-derived neurotrophic factor is decreased in bipolar disorder during depressive and manic episodes." *Neuroscience letters* 398, no. 3 (2006): 215-219.

[7] Post, Robert M. "Role of BDNF in bipolar and unipolar disorder: clinical and theoretical implications." *Journal of psychiatric research* 41, no. 12 (2007): 979-990.

[8] Weinstein, Galit, Alexa S. Beiser, Seung Hoan Choi, Sarah R. Preis, Tai C. Chen, Demetrios Vorgas, Rhoda Au et al. "Serum brain-derived neurotrophic factor and the risk for dementia: the Framingham Heart Study." *JAMA neurology* 71, no. 1 (2014): 55-61.

[9] Hashimoto, Ryota, et al. "Lithium induces brain-derived neurotrophic factor and activates TrkB in rodent cortical neurons: an essential step for neuroprotection against glutamate excitotoxicity." *Neuropharmacology* 43.7 (2002): 1173-1179.

[10] Berton, Olivier, Colleen A. McClung, Ralph J. DiLeone, Vaishnav Krishnan, William Renthal, Scott J. Russo, Danielle Graham et al. "Essential role of BDNF in the mesolimbic dopamine pathway in social defeat stress." *Science* 311, no. 5762 (2006): 864-868.

[11] Ghosh, Anirvan, Josette Carnahan, and Michael E. Greenberg. "Requirement for BDNF in activity-dependent survival of cortical neurons." *Science* 263.5153 (1994): 1618-1623.

[12] Acheson, Ann, Joanne C. Conover, James P. Fandl, Thomas M. DeChiara, Michelle Russell, Anu Thadani, Stephen P. Squinto, George D. Yancopoulos, and Ronald M. Lindsay. "A BDNF autocrine loop in adult sensory neurons prevents cell death." (1995): 450-453.

[13] Conover, J. C., J. T. Erickson, D. M. Katz, L. M. Bianchi, W. T. Poueymirou, J. McClain, L. Pan et al. "Neuronal deficits, not involving motor neurons, in mice lacking BDNF and/or NT4." (1995): 235-238.

[14] Jones, Kevin R., Isabel Fariñas, Carey Backus, and Louis F. Reichardt. "Targeted disruption of the BDNF gene perturbs brain and sensory neuron development but not motor neuron development." *Cell* 76, no. 6 (1994): 989-999.

[15] Chuang, De-Maw. "Neuroprotective and neurotrophic actions of the mood stabilizer lithium: can it be used to treat neurodegenerative diseases?." *Critical Reviews™ in Neurobiology* 16.1&2 (2004).

[16] Forlenza, Orestes Vicente, Vanessa de Jesus Rodrigues De Paula, and Breno S. Diniz. "Neuroprotective effects of lithium: implications for the treatment of Alzheimer's disease and related neurodegenerative disorders." *ACS chemical neuroscience* (2014).

[17] Nunes, Paula V., Orestes V. Forlenza, and Wagner F. Gattaz. "Lithium and risk for Alzheimer's disease in elderly patients with bipolar disorder." *The British Journal of Psychiatry* 190.4 (2007): 359-360.

[18] Kessing, Lars Vedel, Thomas Alexander Gerds, Nikoline Nygård Knudsen, Lisbeth Flindt Jørgensen, Søren Munch Kristiansen, Denitza Voutchkova, Vibeke Ernstsen et al. "Association of lithium in drinking water with the incidence of dementia." *JAMA psychiatry* 74, no. 10 (2017): 1005-1010.

[19] Fajardo, Val Andrew, Val Andrei Fajardo, Paul J. LeBlanc, and Rebecca EK MacPherson. "Examining the Relationship between Trace Lithium in Drinking Water and the Rising Rates of Age-Adjusted Alzheimer's Disease Mortality in Texas." *Journal of Alzheimer's Disease* Preprint (2018): 1-10.

[20] Woolley JD, Khan BK, Murthy NK, Miller BL, Rankin KP. The diagnostic challenge of psychiatric symptoms in neurodegenerative disease: rates of and risk factors for prior psychiatric diagnosis in patients with early neurodegenerative disease. The Journal of clinical psychiatry. 2011 Feb 15;72(2):1-478

[21] McMillan, Corey T., Brian B. Avants, Philip Cook, Lyle Ungar, John Q. Trojanowski, and Murray Grossman. "The power of neuroimaging biomarkers for screening frontotemporal dementia." *Human brain mapping* 35, no. 9 (2014): 4827-4840

[22] Roberson, E. D., J. H. Hesse, K. D. Rose, H. Slama, J. K. Johnson, K. Yaffe, M. S. Forman et al. "Frontotemporal dementia progresses to death faster than Alzheimer disease." *Neurology* 65, no. 5 (2005): 719-725.

[23] Monji, Akira, Keisuke Motomura, Yoshito Mizoguchi, Tomoyuki Ohara, Shingo Baba, Takashi Yoshiura, and Shigenobu Kanba. "A case of late-onset bipolar disorder with severely abnormal behavior and neuroimaging observations very similar to those of frontotemporal dementia." *The Journal of neuropsychiatry and clinical neurosciences* 26, no. 1 (2014): E35-E35.

[24] Devanand, Davangere P., Gregory H. Pelton, Kristina D'Antonio, Jesse G. Strickler, William C. Kreisl, James Noble, Karen Marder, Anne Skomorowsky, and Edward D. Huey. "Low-dose Lithium Treatment for Agitation and Psychosis in Alzheimer Disease and Frontotemporal Dementia: A Case Series." *Alzheimer Disease & Associated Disorders* 31, no. 1 (2017): 73-75.

[25] https://clinicaltrials.gov/ct2/show/NCT02862210

[26] Huey, Edward D., Karen T. Putnam, and Jordan Grafman. "A systematic review of neurotransmitter deficits and treatments in frontotemporal dementia." *Neurology* 66, no. 1 (2006): 17-22.

[27] Hara, Taichi, Kenji Nakamura, Makoto Matsui, Akitsugu Yamamoto, Yohko Nakahara, Rika Suzuki-Migishima, Minesuke Yokoyama et al. "Suppression of basal autophagy in neural cells causes neurodegenerative disease in mice." *Nature* 441, no. 7095 (2006): 885.

[28] Komatsu, Masaaki, Satoshi Waguri, Tomoki Chiba, Shigeo Murata, Jun-ichi Iwata, Isei Tanida, Takashi Ueno et al. "Loss of autophagy in the central nervous system causes neurodegeneration in mice." *Nature* 441, no. 7095 (2006): 880.

[29] Nixon, Ralph A. "The role of autophagy in neurodegenerative disease." *Nature medicine* 19, no. 8 (2013): 983.

[30] Menzies, Fiona M., Angeleen Fleming, Andrea Caricasole, Carla F. Bento, Stephen P. Andrews, Avraham Ashkenazi, Jens Füllgrabe et al. "Autophagy and neurodegeneration: pathogenic mechanisms and therapeutic opportunities." *Neuron* 93, no. 5 (2017): 1015-1034.

[31] Sarkar, Sovan, R. Andres Floto, Zdenek Berger, Sara Imarisio, Axelle Cordenier, Matthieu Pasco, Lynnette J. Cook, and David C. Rubinsztein. "Lithium induces autophagy by inhibiting inositol monophosphatase." *J Cell Biol* 170, no. 7 (2005): 1101-1111.

[32] Kim, Young Chul, and Kun-Liang Guan. "mTOR: a pharmacologic target for autophagy regulation." *The Journal of clinical investigation* 125, no. 1 (2015): 25-32.

[33] Sarkar, Sovan, Gauri Krishna, Sara Imarisio, Shinji Saiki, Cahir J. O'kane, and David C. Rubinsztein. "A rational mechanism for combination treatment of Huntington's disease using lithium and rapamycin." *Human molecular genetics* 17, no. 2 (2007): 170-178.

[34] Motoi, Yumiko, Kohei Shimada, Koichi Ishiguro, and Nobutaka Hattori. "Lithium and autophagy." *ACS chemical neuroscience*5, no. 6 (2014): 434-442.

[35] Kitano, Hiroaki. "Systems biology: a brief overview." *Science*295, no. 5560 (2002): 1662-1664.

[36] Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. "KEGG: new perspectives on genomes, pathways, diseases and drugs." *Nucleic acids research* 45, no. D1 (2016): D353-D361.

[37] Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos et al. "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible." *Nucleic acids research* (2016): gkw937.

[38] Zhang JD and Wiemann S (2009). "KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor." *Bioinformatics*, pp. 1470–1471.

[39] Zhang JD (2017). *KEGGgraph: Application Examples*. R package version 1.38.0.

[40] Nurnberger, John I., et al. "Identification of pathways for bipolar disorder: a meta-analysis." *JAMA psychiatry* 71.6 (2014): 657-664.

[41] Franceschini, A (2013). "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." *Nucleic Acids Research (Database issue)*, **41**

[42] Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research* 37, no. 1 (2008): 1-13.

[43] Thomas, P. D. "Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium." *Nucleic Acids Res* 45 (2017): D331-D338.

[44] Ge, Weihao, Zeeshan Fazal, and Eric Jakobsson. "Using Optimal F-Measure and Random Resampling in Gene Ontology Enrichment Calculations." *bioRxiv* (2017): 218248.

[45] Leucht, Stefan, Werner Kissling, and John McGrath. "Lithium for schizophrenia." *Cochrane Database Syst Rev* 3, no. 3 (2007).

[46] Sarkar, Sovan, Gauri Krishna, Sara Imarisio, Shinji Saiki, Cahir J. O'kane, and David C. Rubinsztein. "A rational mechanism for combination treatment of Huntington's disease using lithium and rapamycin." *Human molecular genetics* 17, no. 2 (2007): 170-178.

[47] Twine, Natalie A., Karolina Janitz, Marc R. Wilkins, and Michal Janitz. "Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease." *PloS one* 6, no. 1 (2011): e16266.

[48] Julia, T. C. W., and Alison M. Goate. "Genetics of β-Amyloid Precursor Protein in Alzheimer's Disease." *Cold Spring Harbor perspectives in medicine* 7, no. 6 (2017): a024539.

[49] Pomara, Nunzio, and Davide Bruno. "Major depression may lead to elevations in potentially neurotoxic amyloid beta species independently of Alzheimer Disease." *The American Journal of Geriatric Psychiatry* 24, no. 9 (2016): 773-775

[50] Bschor, Tom. "Lithium in the treatment of major depressive disorder." *Drugs* 74, no. 8 (2014): 855-862.

[51] Su, Yuan, John Ryder, Baolin Li, Xin Wu, Niles Fox, Pat Solenberg, Kellie Brune et al. "Lithium, a common drug for bipolar disorder treatment, regulates amyloid-β precursor protein processing." *Biochemistry* 43, no. 22 (2004): 6899-6908.

[52] Aprahamian, Ivan, Franklin S. Santos, Bernardo dos Santos, Leda Talib, Breno S. Diniz, Márcia Radanovic, Wagner F. Gattaz, and Orestes V. Forlenza. "Long-term, low-dose lithium treatment does not impair renal function in the elderly: a 2-year randomized, placebo-controlled trial followed by single-blind extension." *The Journal of clinical psychiatry* 75, no. 7 (2014): 672-678.

[53] Coyle-Gilchrist, Ian TS, Katrina M. Dick, Karalyn Patterson, Patricia Vázquez Rodríquez, Eileen Wehmann, Alicia Wilcox, Claire J. Lansdall et al. "Prevalence, characteristics, and survival of frontotemporal lobar degeneration syndromes." *Neurology* 86, no. 18 (2016): 1736-1743.

CHAPTER 4: Systems Biology Understanding of the Effects of Lithium on Cancer

(Submitted to *Frontiers of Cancer*, in review)

Weihao Ge and Eric Jakobsson *

Correspondence *Eric Jakobsson jake@illinois.edu

Abstract

Lithium has many widely varying biochemical and phenomenological effects, suggesting that a systems biology approach is required to understand its action. Multiple lines of evidence point to lithium as a significant factor in development of cancer, showing that understanding lithium action is of high importance. In this paper we undertake first steps towards a systems approach by analyzing mutual enrichment between the interactomes of lithium-sensitive enzymes and the pathways associated with cancer. This work integrates information from two important databases, STRING and KEGG pathways. We find that for the majority of cancer pathways the mutual enrichment is many times greater than chance, reinforcing previous lines of evidence that lithium is an important influence on cancer.

Keywords: Lithium, systems biology, biochemical pathways, biochemical networks

Introduction

Clinical and Epidemiological Context for Lithium and Cancer

By far the most common medical use of lithium is as a first line therapy for bipolar disorder, including associated depression as well as mania.[1] A comprehensive review of the literature confirms that lithium is also effective against unipolar depression with unique anti-suicidal effectiveness, and may also be useful against cancer and neurodegenerative disease.[2]

One line of evidence for the possible use of lithium as an anticancer agent is epidemiological. A retrospective study showing that psychiatric patients undergoing lithium therapy for bipolar disorder had a much lower incidence of cancer than a matched group not receiving lithium therapy.[3] More recent studies of similar design, one conducted nationwide across Sweden, and another across Taiwan, achieved the same result.[4] [5] On the other hand another nationwide study, this time from Denmark, showed no correlation of lithium with colorectal adenocarcinoma.[6] On closer look, the Denmark study does not contradict the Swedish study. The Swedish study also found that for the entire population lithium was not correlated with cancer incidence, but in addition found that bipolar individuals not treated with lithium had a higher incidence of cancer than the general population. Lithium-treated bipolar patients, on the other hand, had essentially the same cancer incidence as the general population.

One piece of experimental evidence for lithium's potential as a cancer therapeutic modality is that it was observed to inhibit prostate tumor growth,[7] presumably through its ability to inhibit GSK3. A detailed study of molecular mechanisms by which lithium inhibition of GSK3-beta inhibits proliferation of prostate tumor cells in culture was presented by Sun et al.[8] The work was subsequently extended to an animal model.[9] A clinical trial for the effect of lithium coupled with prostatectomy on men has been conducted but as of this writing results have not yet been published.[10]

With respect to other cancers, lithium has been found to be lethal to neuroblastoma cells but not to normal nerve cells.[11] The experimentally determined effective dose was 12 mM, a level which would be lethal if achieved systemically in a human or model organism but perhaps could be induced locally. A similar effect was found in ovarian cancer cells,[12] although a subsequent similar study on ovarian cancer cells suggests only a more modest benefit.[13] It is not clear from our reading of the two ovarian cancer papers why the results are significantly different from each other.

With respect to colorectal cancer, one study suggests that lithium inhibits proliferation of a colorectal cancer cell line.[14] Another study on colon cancer cells showed that lithium specifically induced a reversal of the epithelial-to-mesenchymal transition characteristic of the cancer cells.[15]

Two studies with relatively small sample size suggested a possible link between lithium and tumors of the upper urinary tract.[16] [17] However a large-scale study involving all urinary tract cancers in Denmark over a multi-year period found no correlation with lithium use.[18]

Because lithium therapy is systemic rather than topical or local, it follows that lithium might inhibit metastasis. Evidence that this is the case for colon cancer comes from observation of inhibition of metastasis-inducing factors by lithium and by observation on reduced metastasis in model animals given lithium therapy.[19]

Autophagy is a key cellular process in the inhibition of cancer.[20] Lithium has been shown to induce autophagy, due to its inhibition of inositol monophosphatase.[21] The full range of lithium effects on autophagy is complicated,[22] as might be expected because of lithium's lack of specificity.[2]

Because of the promising indications as cited above, lithium has been suggested as one of a number of drugs commonly used for other reasons, to be repurposed for cancer.[23]

Biochemical Context for Lithium and Cancer

Much of lithium's biochemical action may be summarized by noting that it inhibits enzymes that have magnesium as a co-factor.[2] There are many published examples of such competition. Lithium appears to inhibit β-adrenergic and muscarinic receptor coupling to G proteins by competing with magnesium, which facilitates such coupling. [24] [25] [26] [27] [28] A particularly important example is substitution of a lithium ion for a magnesium ion acting as a cofactor in inositol monophosphatase, mentioned earlier in this paper. In this protein the binding site for lithium is not revealed in crystallography nor in solution NMR but can be identified in magic angle spinning solid state NMR, which is more suitable for systems with large internal motion.[29] Another target of lithium, also a magnesium-dependent phosphatase and with relevance to neural effects, is bisphosphate 3-prime-nucleotidase (BPNT1).[30] [31] These findings are consistent with a hypothesis that lithium inhibits at least some magnesium-dependent enzymes by displacing magnesium from its binding site thereby reducing the structural stability and lowering activity of the enzyme.

One mode of action with many consequences is lithium inhibition of glycogen synthase kinase 3 beta (GSK3B), initially shown in vitro and in intact cells,[32] and in the context of embryonic development.[33] It was later shown that lithium exerted its inhibitory effect on GSK3B by competing with magnesium for an essential binding site.[34] There are two closely related forms of GSK3, termed alpha (GSK3A) and beta (GSK3B), which are equivalently inhibited by lithium.[35] The two forms of GSK3 have substantial functional redundancy.[36] However some of their physiological properties are different, as demonstrated by the fact that GSK3B knockout mice are not viable,[37] but GSK3A knockout mice survive.[38] The very widespread nature of GSK3B effects is related to the large number of transcription factors that it regulates.[39] It functionally modulates cellular threshold for apoptosis,[40] it is central to mediating mitochondrial response to stress;41 it facilitates immune responses by enabling the nuclear export of NF-ATc;42 it regulates inflammation;43 it regulates cardiac hypertrophy and development,44 to name just a few. Based on microarray studies of brain cells in animals, lithium alters gene expression patterns significantly,45 to be expected due to the large number of transcription factors regulated by GSK3B. Mice heterozygous for GSK3B exhibit similar behavioral traits to wild type littermates treated with 1mM lithium (a concentration that inhibits about 25% of GSK3 activity, in line with 1 of the 4 alleles of GSK3 inactivated in the GSK3B heterozygous mice)46

In addition to inhibiting the activity of GSK3B, lithium also inhibits its transcription.[47] Of all kinases, GSK3 appears to have the largest number of known substrates, over 100 known[48] and about 500 predicted by theory based on scanning and interpreting relevant motif sequences in the human genome.[49] Lithium will thus to some extent modulate activity along all pathways containing the hundreds of GSK3 substrates. So far, to our knowledge there are no published counterexamples to the hypothesis that lithium will exert an inhibitory effect on all proteins with essential magnesium binding sites, of which there are estimated to be over three thousand.[50]

A second major widespread effect of lithium is as a cofactor with magnesium in interacting with phosphate groups. The primary energy source for cells and the substrate for phosphorylating enzymes is not bare ATP, but rather magnesium-associated ATP (MgATP).[51] NMR studies show that lithium associates with MgATP.[52] Based on this admittedly small amount of data, we hypothesize that lithium associates with all magnesium-phosphate complexes and will thus modulate to some extent all phosphorylation reactions and all ATP-splitting processes. This is a reasonable interpretation of early work by Willis and Fang, in which lithium was found to increase the activity of the sodium-potassium pump without itself being transported significantly.[53] We have noted earlier in this paper the inhibitory effect of lithium on GSK3 by the mechanism of competing with Mg. Here we note that lithium also inhibits the activity of GSK3 by a second method, that is, by increasing phosphorylation.[54] Depending on context of relevant protein-protein interactions, lithium's effect on phosphorylation of a particular protein may be to either increase it or decrease it. For example, lithium decreases phosphorylation of tau-protein, presumably because it inhibits GSK3B, which is implicated in the phosphorylation of the tau-protein.[55]

Because lithium affects many different biological molecules and processes[2], it is essential to utilize the tools of systems biology[56] if a comprehensive understanding of lithium action and its prospects for therapy are to be obtained. Important concepts for organizing biological information in a systems context are pathways and networks. A very useful tool for obtaining data about known pathways is the KEGG database.[57] An equally useful and complementary tool is the STRING database of interacting proteins.[58]

In the present paper we investigate further the possible linkages among 1) lithium, 2) affective disorders, and 3) neurodegenerative disorders by analyzing the mutual enrichment between STRING-derived interactomes of lithium-sensitive enzymes, and the KEGG pathways associated with cancer.

Methods

Analysis was performed on the interactomes of lithium-sensitive genes, as identified by prior literature search[2]. This search suggested BDNF, BPNT1, DISC1, DIXDC1, FBP1, FBP2, GSK3A, GSK3B, inositol monophosphatases (IMPA1, IMPA2, and IMPAD1), INPP1, and PGM1 as key to understanding the broad biological actions of lithium. The interactomes of these genes were extracted from the STRING database (https://string-db.org). For each key gene, we adjust confidence level and order of neighbors (nearest only or next nearest included), so that each set contains a few hundred genes. This size is large enough for statistically reliable enrichment analysis. Very similar sets were merged; in particular FBP1 and FBP2 were merged into one set, and the inositol monophosphatases were merged into one set. On the other hand, GSK3A and GSK3B showed sufficient differences to be considered separately. Overall, we consider 10 distinct lithium-sensitive entities.

Disease Association

We used the R-package KEGGgraph[59] [60] to identify the genes associated with the pathways of interest.

P-value calculation

The fundamental question we address is whether there is significant overlap or mutual enrichment between the interactomes of lithium-sensitive genes and the pathways or gene sets implicated in various cancers.

For each of the 10 lithium sets, an ensemble of 1000 null sets are generated by random selection from the human genome. Each null set is the same size as the corresponding lithium set. Then we used the R-package STRINGdb[61] to perform KEGG pathway enrichment analysis. This operation is a particular example of the powerful technique of gene-annotation enrichment analysis.[62] In gene-annotation enrichment analysis a test list of genes (often derived from gene expression experiments) is compared to an organized database of gene annotations, often referred to as a gene ontology[63], an array of gene lists corresponding to different biological functions, molecular functions, or locations in the cell. The output of the gene-annotation enrichment analysis is expressed as the likelihood that the list overlaps could have occurred by chance (p-value). A very low p-value implies that the degree of overlap is highly significant statistically and very likely is significant biologically. In our study the gene lists we are comparing are the interactomes of lithium sensitive enzymes on the one hand, and KEGG pathways and Kegg pathways associated with cancer on the other hand. For each KEGG term retrieved, a null distribution of uncorrected *p*-value is generated by the 1000 null sets. This gives us a measure of the false discovery rate, since any overlap between the null sets and the KEGG pathways is purely accidentally. Then the fraction of null set uncorrected *p*-values smaller than or equal to the lithium-sensitive set uncorrected *p*-value would be the empirical *p*-value. For a detailed discussion of empirical p-value determination see Ge et al[64].

Key Gene Prediction

We predict key genes by counting how many times a gene appears in the cross section of interactomes and pathways associated with a particular disease. In this way, we predict which genes might be most important in disease-related pathways. Then, the genes are scored by the sum of mean counts over all diseases. A higher ranking indicates a gene would be associated with an important factor in many diseases.

Results

Figure 4.1 shows mutual lithium interactome enrichment with specific cancer pathways, represented by heatmaps. Each area on the heatmap is a color-coded representation of the degree of mutual enrichment between the genes in the interactome of the indicated lithium sensitive enzyme and the genes in the indicated pathway. The darker the shade, the more significant the mutual enrichment of the interactome-pathway combination is. The light areas on the heatmap represent situations where a lithium-sensitive interactome has little or no mutual enrichment with a cancer pathway. The dark areas, deep orange and red, represent situations where enrichment is very strong—far greater than could be expected by chance. Three genes stand out as being not strongly connected to cancer pathways: BPNT1, DISC1, and PGM1. Of the cancer pathways, breast cancer stands out as being not likely to be strongly influenced by lithium levels. For the remainder of the genes and the remainder of the cancers, the relationship between the lithium-sensitive interactome and the cancer phenome is strong.

For each of the cancer-associated pathways we wished to compute a single number representing the relative likely sensitivity of the disease to lithium, in order to contribute to prioritizing which diseases are most likely to benefit from clinical trials with lithium. There is a significant literature on combining p-values,[65] with choices among methods depending on the detailed structure of the data. We adopt a relatively simple approach, which is to compute the geometric mean of the individual p-values for each pathway-interactome mutual enrichment value.

$$p_{mean}=(p_1 \text{ x } p_2 \text{ x } p_3 \text{ x} \ldots\ldots p_n)^{1/n} \qquad \text{Equation (1)}$$

The method of averaging in Equation (1) ensures that both strong and weak enrichments contribute significant weight to the mean. Note that all of the p-values that go into Equation (1) are corrected for false discovery rate by random resampling. Thus, no further false discovery rate correction is necessary for computing p_{mean}. Note also that our method is bounded at the low end of p-values by the number of null samples it is reasonable to compute, given compute time constraints. For one thousand null sets as used in this paper, the computed p-value will be zero when none of the thousand null sets shows the degree of enrichment of the test sets. For purposes of computing the p_{mean} in equation (1) we substitute 10^{-4} for zero for each of these cases.

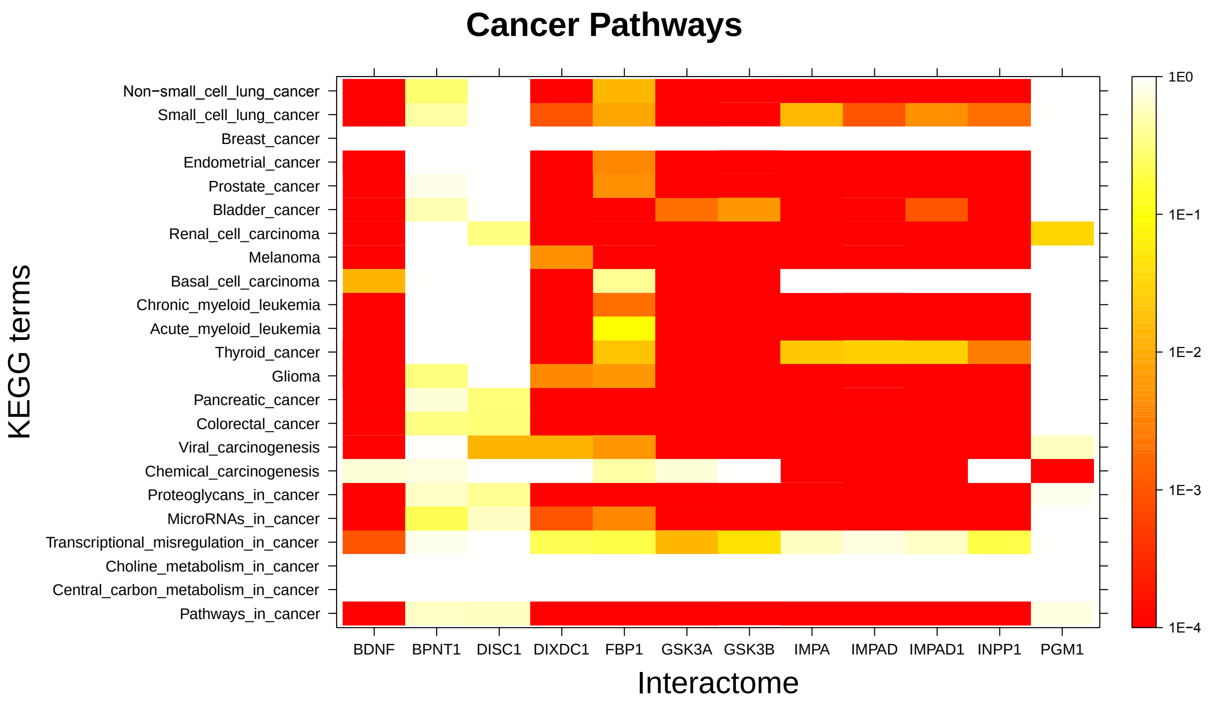


Figure 4.1. Visual representation of mutual enrichment patterns between cancer-associated pathways and the interactomes of lithium-sensitive gene products. Calibration of *p*-value vs. color is indicated by a vertical scale to the right of the heat map. Red or dark orange indicates very strong enrichment while lighter color indicates weak or, if white, no enrichment. Three genes stand out as being not strongly connected to cancer pathways: BPNT1, DISC1, and PGM1. Of the cancer pathways, breast cancer stands out as being not likely to be strongly influenced by lithium levels. For the remainder of the genes and the remainder of the cancers, the relationship between the lithium-sensitive interactome and the cancer phenome is strong.

Cancer-related KEGG pathway	p_{mean}	$1/p_{mean}$
Colorectal_cancer	0.00049903	2003.89209
Pancreatic_cancer	0.00053973	1852.76828
Proteoglycans_in_cancer	0.00054323	1840.83131
Renal_cell_carcinoma	0.00056283	1776.72794
Pathways_in_cancer	0.00056569	1767.76501
Chronic_myeloid_leukemia	0.00085134	1174.61894
Non-small_cell_lung_cancer	0.00090896	1100.15513
Endometrial_cancer	0.00091244	1095.95823
Prostate_cancer	0.00091481	1093.12212
MicroRNAs_in_cancer	0.00093106	1074.04618
Melanoma	0.00093303	1071.77346
Viral_carcinogenesis	0.00099762	1002.38273
Glioma	0.00122264	817.904951
Acute_myeloid_leukemia	0.00125638	795.934615
Bladder_cancer	0.00150566	664.158631
Small_cell_lung_cancer	0.00352993	283.291551
Thyroid_cancer	0.00567253	176.288152
Basal_cell_carcinoma	0.03711938	26.9401078
Chemical_carcinogenesis	0.05277248	18.9492716
Transcriptional_misregulation_in_cancer	0.14953148	6.68755483
Central_carbon_metabolism_in_cancer	1	1
Choline_metabolism_in_cancer	1	1
Breast_cancer	1	1

Table 4.1. Rank order of significance of enrichment between lithium-sensitive interactomes and KEGG cancer-associated pathways. It is seen that for the great majority of pathways the mutual enrichment is very significant, with *p*-values significantly below .01. Breast cancer is unusual; it appears there is no enrichment beyond chance. The table also displays a “lithium sensitivity index”, which is $1/p_{mean}$

Table 4.1 shows in rank order the significance of enrichment between lithium-sensitive interactomes and KEGG cancer-associated pathways. It is seen that for the great majority of pathways the mutual enrichment is very significant, with *p*-values significantly below .01. Breast cancer is unusual; it appears there is no enrichment beyond chance. The table also displays a “lithium sensitivity index”, which is $1/p_{mean}$

We should note that sensitivity to lithium does not necessarily imply a *beneficial* sensitivity. There are some indications for some cancers that lithium might be beneficial, as described in the Introduction section of this paper, but because of the complexity of the feedback relationships in these pathways, a complicated relationship between lithium ingestion and cancer incidence is very possible.

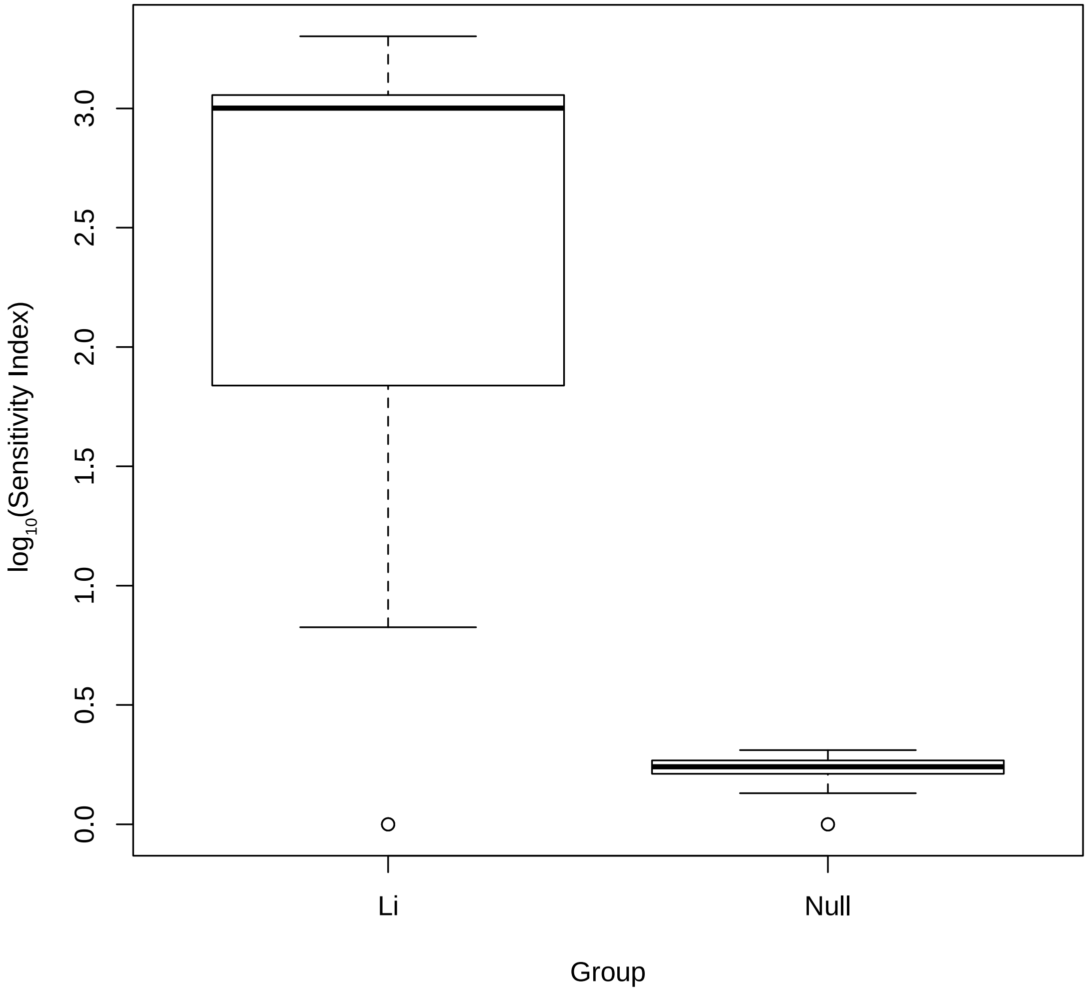


Figure 4.2. This figure visualizes the strength of the projected lithium influence on cancer pathways. The logarithms of the lithium sensitivity indices ($1/p_{mean}$) are shown in boxplot format together with the corresponding results when the lithium interactomes are replaced with random gene sets. Essentially this figure shows the signal-to-noise ratio of our results and demonstrates conclusively that lithium ingestion is overwhelmingly likely to modulate the incidence of a wide range of cancers.

Figure 4.2 visualizes the strength of the projected lithium influence on cancer pathways. In this figure the logarithms of the lithium sensitivity indices ($1/p_{mean}$) are shown in boxplot format together with the corresponding results when the lithium interactomes are replaced with random gene sets. Essentially this figure shows the signal-to-noise ratio of our results and suggests that lithium ingestion is overwhelmingly likely to influence the incidence of a wide range of cancers.

Summary and Discussion

We have conducted a pathway and network analysis exploring the role of lithium in multiple cancers. The results show that for the large majority of such cancers, there is high mutual enrichment between the interactomes of lithium-sensitive enzymes and the pathways associated with those diseases, indicating that lithium is very likely to affect the incidence and course of the disease. Our results are consistent with a variety of lines of evidence from both epidemiology and from experiment, cited in the Introduction section of this paper, suggesting possible influence of lithium on the incidence and progression of cancer.

We hope that the results described in this paper will contribute to prioritizing and designing clinical trials of lithium for cancer. To provide context for such prioritization and design, it is essential to take into account the ways in which lithium is unique, both as a pharmaceutical and as an ion that is ubiquitous in the environment, and therefore ubiquitous in the water and food we ingest[2]:

1. Unlike other ions, lithium is not regulated by selective membrane transport processes. Therefore, lithium concentration in both extracellular and intracellular compartments, rather than being roughly constant, is roughly proportional to lithium ingestion.
2. Unlike other pharmaceuticals, lithium is wildly nonselective in its biochemical effects. The major underlying mechanism for the lack of selectivity is lithium's general propensity to inhibit the many enzymes that have magnesium as a cofactor.
3. Unlike other pharmaceuticals, lithium is an essential nutrient. The question with lithium is not whether it should be ingested or not, but rather how much. Extreme lithium deprivation results in failure to thrive, while too much lithium is toxic.

In the light of all these factors, we suggest that the correct question to ask with respect to lithium and a particular disease is not, "Should lithium be administered for this particular disease?" but rather, "What is the optimum blood level of lithium for this individual, given his or her disease history, status, genetic propensities, and other medications?" Unlike other pharmaceuticals that are far more specific and inhibit or activate one gene or a small number of genes, the model for lithium action is that it alters the balance between a large number of interacting processes and pathways. Thus, a dose-response curve for lithium is likely to be highly nonlinear and not always monotonic.

There are just a few well-established markers for optimum concentrations. For a patient with a reliable diagnosis of bipolar disorder a common target for optimality would be blood concentration of 0.8-1 mM. Significantly higher concentrations will result in acute toxicity, while significantly lower will result in loss of effectiveness. However, this level has some side effects when sustained for years or decades, namely an increased risk of kidney damage and lowered thyroid activity.

At the other end of the dosage scale, epidemiological evidence is compelling that geographical variations in concentration of lithium in the drinking water are correlated with a variety of health and wellness markers.

Another important marker is provided by a study showing that over a four-year period a lithium level of .25-.4 mM of lithium (1/4 to 1/2 of the bipolar therapeutic dose) did not incur any renal damage[66]. This study suggests that clinical studies exploring low to medium-dose lithium could be undertaken with relatively minimal concerns for side effects.

One possible piece of low-hanging fruit for a clinical trial would be low- to medium-dose lithium for men undergoing active surveillance (AS) for advance of prostate cancer. From studies of AS outcomes, a large fraction of patients on AS ultimately require invasive treatment, as reviewed by Dall'Era et al[67]. When this need arises it typically comes after only a few years. Thus, a trial of lithium in this context would produce significant results in a short time and would be relatively inexpensive. One of us (EJ) conducted an informal one-person trial on himself after being diagnosed with prostate cancer in 2014, ingesting lithium supplements sufficient to bring his blood lithium to .3-.4mM while undergoing AS by Memorial Sloan Kettering Cancer Center. MSK did not prescribe the lithium but agreed to include lithium level measurement in periodic blood tests.) In October 2017 EJ was told that there was no longer a need for AS. One case, important as it is to EJ, does not have statistical significance. We need clinical trials with significant numbers of people.

We will be happy to collaborate on further specific pathway or network analysis relevant to any of the cancers for which lithium may be a promising component of therapy.

Author Contributions

The work was planned jointly in conversations between EJ and WG. WG did the computations and prepared the figures and tables. WG wrote the first draft of the Methods and Results sections. EJ wrote the first draft of the Introduction and Conclusions sections. Both authors shared in the final refinement of the manuscript.

Conflict of Interest

The authors have no personal, professional, or financial relationships that could be construed as a conflict of interest with the work described in this paper.

[1] Post, Robert M. "Treatment of bipolar depression: evolving recommendations." *Psychiatric Clinics* 39, no. 1 (2016): 11-33.

[2] Jakobsson, Eric, Orlando Argüello-Miranda, See-Wing Chiu, Zeeshan Fazal, James Kruczek, Santiago Nunez-Corrales, Sagar Pandit, and Laura Pritchet. "Towards a Unified Understanding of Lithium Action in Basic Biology and its Significance for Applied Biology." *The Journal of membrane biology* 250, no. 6 (2017): 587-604.

[3] Cohen, Y., A. Chetrit, P. Sirota, and B. Modan. "Cancer morbidity in psychiatric patients: influence of lithium carbonate treatment." *Medical Oncology* 15, no. 1 (1998): 32-36.

[4] Martinsson L, Westman J, Hällgren J, Ösby U, Backlund L. Lithium treatment and cancer incidence in bipolar disorder. Bipolar disorders. 2016 Feb 1;18(1):33-40.

[5] Huang, Ru-Yu, Kun-Pin Hsieh, Wan-Wen Huang, and Yi-Hsin Yang. "Use of lithium and cancer risk in patients with bipolar disorder: population-based cohort study." *The British Journal of Psychiatry* (2016): bjp-bp.

[6] Pottegård A, Ennis ZN, Hallas J, Jensen BL, Madsen K, Friis S. Long-term use of lithium and risk of colorectal adenocarcinoma: a nationwide case–control study. British journal of cancer. 2016 Mar 1;114(5):571-5.

[7] Mazor, Michal, Yoshiaki Kawano, Hanneng Zhu, Jonathan Waxman, and Robert M. Kypta. "Inhibition of glycogen synthase kinase-3 represses androgen receptor activity and prostate cancer cell growth." *Oncogene* 23, no. 47 (2004): 7882-7892.

[8] Sun, Aijing, Ilanchezian Shanmugam, Jiawu Song, Paul F. Terranova, J. Brantley Thrasher, and Benyi Li. "Lithium suppresses cell proliferation by interrupting E2F–DNA interaction and subsequently reducing S–phase gene expression in prostate cancer." *The Prostate* 67, no. 9 (2007): 976-988.

[9] Zhu, Qing, Jun Yang, Suxia Han, Jihong Liu, Jeffery Holzbeierlein, J. Brantley Thrasher, and Benyi Li. "Suppression of glycogen synthase kinase 3 activity reduces tumor growth of prostate cancer in vivo." *The Prostate* 71, no. 8 (2011): 835-845.

[10] https://clinicaltrials.gov/ct2/show/NCT02198859

[11] Duffy, David J., Aleksandar Krstic, Thomas Schwarzl, Desmond G. Higgins, and Walter Kolch. "GSK3 inhibitors regulate MYCN mRNA levels and reduce neuroblastoma cell viability through multiple mechanisms, including p53 and Wnt signaling." *Molecular cancer therapeutics* 13, no. 2 (2014): 454-467.

[12] Cao, Qi, Xin Lu, and You-Ji Feng. "Glycogen synthase kinase-3β positively regulates the proliferation of human ovarian cancer cells." *Cell research* 16, no. 7 (2006): 671-677.

[13] Novetsky, Akiva P., Dominic M. Thompson, Israel Zighelboim, Premal H. Thaker, Matthew A. Powell, David G. Mutch, and Paul J. Goodfellow. "Lithium and inhibition of GSK3β as a potential therapy for serous ovarian cancer."*International journal of gynecological cancer: official journal of the International Gynecological Cancer Society* 23, no. 2 (2013): 361.

[14] Li, H., Huang, K., Liu, X., Liu, J., Lu, X., Tao, K., Wang, G. and Wang, J., 2014. Lithium chloride suppresses colorectal cancer cell survival and proliferation through ROS/GSK-3β/NF-κB signaling pathway. *Oxidative medicine and cellular longevity*, *2014*.

[15] Costabile, V., Duraturo, F., Delrio, P., Rega, D., Pace, U., Liccardo, R., Rossi, G.B., Genesio, R., Nitsch, L., Izzo, P. and De Rosa, M., 2015. Lithium chloride induces mesenchymal-to-epithelial reverting transition in primary colon cancer cell cultures. *international journal of oncology*, *46*(5), pp.1913-1923.

[16] Rookmaaker, Maarten B., Heleen AJM van Gerven, Roel Goldschmeding, and Walther H. Boer. "Solid renal tumours of collecting duct origin in patients on chronic lithium therapy." *Clinical kidney journal* 5, no. 5 (2012): 412-415.

[17] Zaidan, Mohamad, Fabien Stucker, Bénédicte Stengel, Viorel Vasiliu, Aurélie Hummel, Paul Landais, Jean-Jacques Boffa, Pierre Ronco, Jean-Pierre Grünfeld, and Aude Servais. "Increased risk of solid renal tumors in lithium-treated patients." *Kidney international* 86, no. 1 (2014): 184-190.

[18] Pottegård, Anton, Jesper Hallas, Boye L. Jensen, Kirsten Madsen, and Søren Friis. "Long-term lithium use and risk of renal and upper urinary tract cancers." *Journal of the American Society of Nephrology* (2015): ASN-2015010061.

[19] Maeng, Yong-Sun, Rina Lee, Boram Lee, Seung-il Choi, and Eung Kweon Kim. "Lithium inhibits tumor lymphangiogenesis and metastasis through the inhibition of TGFBIp expression in cancer cells." *Scientific reports* 6 (2016).

[20] Zhong, Zhenyu, Elsa Sanchez-Lopez, and Michael Karin. "Autophagy, inflammation, and immunity: a troika governing cancer and its treatment." *Cell* 166, no. 2 (2016): 288-298.

[21] Sarkar, Sovan, R. Andres Floto, Zdenek Berger, Sara Imarisio, Axelle Cordenier, Matthieu Pasco, Lynnette J. Cook, and David C. Rubinsztein. "Lithium induces autophagy by inhibiting inositol monophosphatase." *J Cell Biol* 170, no. 7 (2005): 1101-1111.

[22] Motoi, Yumiko, Kohei Shimada, Koichi Ishiguro, and Nobutaka Hattori. "Lithium and autophagy." *ACS chemical neuroscience*5, no. 6 (2014): 434-442.

[23] Sleire, Linda, Hilde Elisabeth Førde-Tislevoll, Inger Anne Netland, Lina Leiss, Bente Sandvei Skeie, and Per Øyvind Enger. "Drug repurposing in cancer." *Pharmacological research* 124 (2017): 74-91.

[24] Avissar, Sofia, Dennis L. Murphy, and Gabriel Schreiber. "Magnesium reversal of lithium inhibition of β-adrenergic and muscarinic receptor coupling to G proteins." *Biochemical pharmacology* 41, no. 2 (1991): 171-175.

[25] Mota de Freitas, Duarte, M. Margarida CA Castro, and Carlos FGC Geraldes. "Is competition between Li+ and Mg2+ the underlying theme in the proposed mechanisms for the pharmacological action of lithium salts in bipolar disorder?." *Accounts of chemical research* 39, no. 4 (2006): 283-291.

[26] Yoshikawa, Tomoko, and Sato Honma. "Lithium lengthens circadian period of cultured brain slices in area specific manner." *Behavioural Brain Research*314 (2016): 30-37.

[27] Bauer, Michael, and Michael Gitlin. "What Is Lithium and How Does It Work?." In *The Essential Guide to Lithium Treatment*, pp. 33-43. Springer International Publishing, 2016.

[28] Wang, Hoau-Yan, and Eitan Friedman. "Effects of lithium on receptor-mediated activation of G proteins in rat brain cortical membranes." *Neuropharmacology* 38, no. 3 (1999): 403-414.

[29] Haimovich, A., Eliav, U. and Goldbourt, A., 2012. Determination of the lithium binding site in inositol monophosphatase, the putative target for lithium therapy, by magic-angle-spinning solid-state NMR. *Journal of the American Chemical Society*, *134*(12), pp.5647-5651.

[30] Spiegelberg, Bryan D., June dela Cruz, Tzuo-Hann Law, and John D. York. "Alteration of lithium pharmacology through manipulation of phosphoadenosine phosphate metabolism." *Journal of Biological Chemistry* 280, no. 7 (2005): 5400-5405.

[31] Meisel, Joshua D., and Dennis H. Kim. "Inhibition of lithium-sensitive phosphatase BPNT-1 causes selective neuronal dysfunction in C. elegans." *Current Biology* 26, no. 14 (2016): 1922-1928.

[32] Stambolic, Vuk, Laurent Ruel, and James R. Woodgett. "Lithium inhibits glycogen synthase kinase-3 activity and mimics wingless signalling in intact cells." *Current Biology* 6, no. 12 (1996): 1664-1669.

[33] Klein, Peter S., and Douglas A. Melton. "A molecular mechanism for the effect of lithium on development." *Proceedings of the National Academy of Sciences* 93, no. 16 (1996): 8455-8459.

[34] Ryves, W. Jonathan, and Adrian J. Harwood. "Lithium inhibits glycogen synthase kinase-3 by competition for magnesium." *Biochemical and biophysical research communications* 280, no. 3 (2001): 720-725.

[35] Freland, Laure, and Jean-Martin Beaulieu. "Inhibition of GSK3 by lithium, from single molecules to signaling networks." *Front Mol Neurosci* 5 (2012).

[36] Doble, Bradley W., Satish Patel, Geoffrey A. Wood, Lisa K. Kockeritz, and James R. Woodgett. "Functional redundancy of GSK-3α and GSK-3β in Wnt/β-catenin signaling shown by using an allelic series of embryonic stem cell lines." *Developmental cell* 12, no. 6 (2007): 957-971.

[37] Hoeflich, K. P., Luo, J., Rubie, E. A., Tsao, M. S., Jin, O., and Woodgett, J. R. (2000). Requirement for glycogen synthase kinase-3beta in cell survival and NF-kappaB activation. *Nature* 406, 86–90

[38] MacAulay, K., Doble, B. W., Patel, S., Hansotia, T., Sinclair, E. M., Drucker, D. J., Nagy, A., and Woodgett, J. R. (2007). Glycogen synthase kinase 3alpha-specific regulation of murine hepatic glycogen metabolism. *Cell Metab.* 6, 329–337

[39] Grimes, Carol A., and Richard S. Jope. "The multifaceted roles of glycogen synthase kinase 3β in cellular signaling." *Progress in neurobiology* 65, no. 4 (2001): 391-426.

[40] Beurel, Eléonore, and Richard S. Jope. "The paradoxical pro-and anti-apoptotic actions of GSK3 in the intrinsic and extrinsic apoptosis signaling pathways." *Progress in neurobiology* 79, no. 4 (2006): 173-189.

[41] Juhaszova, Magdalena, Dmitry B. Zorov, Suhn-Hee Kim, Salvatore Pepe, Qin Fu, Kenneth W. Fishbein, Bruce D. Ziman et al. "Glycogen synthase kinase-3β mediates convergence of protection signaling to inhibit the mitochondrial permeability transition pore." *The Journal of clinical investigation* 113, no. 11 (2004): 1535-1549.

[42] Beals, Chan R., Colleen M. Sheridan, Christoph W. Turck, Phyllis Gardner, and Gerald R. Crabtree. "Nuclear export of NF-ATc enhanced by glycogen synthase kinase-3." *Science* 275, no. 5308 (1997): 1930-1933.

[43] Beurel, Eléonore, Suzanne M. Michalek, and Richard S. Jope. "Innate and adaptive immune responses regulated by glycogen synthase kinase-3 (GSK3)." *Trends in immunology* 31, no. 1 (2010): 24-31.

[44] Hardt, Stefan E., and Junichi Sadoshima. "Glycogen synthase kinase-3β a novel regulator of cardiac hypertrophy and development." *Circulation research* 90, no. 10 (2002): 1055-1063.

[45] McQuillin, Andrew, Mie Rizig, and Hugh MD Gurling. "A microarray gene expression study of the molecular pharmacology of lithium carbonate on mouse brain mRNA to understand the neurobiology of mood stabilization and treatment of bipolar affective disorder." *Pharmacogenetics and genomics* 17, no. 8 (2007): 605-617.

[46] O'Brien, W. Timothy, Amber DeAra Harper, Fernando Jové, James R. Woodgett, Silvia Maretto, Stefano Piccolo, and Peter S. Klein. "Glycogen synthase kinase-3β haploinsufficiency mimics the behavioral and molecular effects of lithium." *The Journal of neuroscience* 24, no. 30 (2004): 6791-6798.

[47] Mendes, Camila Teixeira, Fábio Borges Mury, Eloísa de Sá Moreira, Fernando Lopes Alberto, Orestes Vicente Forlenza, Emmanuel Dias-Neto, and Wagner Farid Gattaz. "Lithium reduces Gsk3b mRNA levels: implications for Alzheimer disease." *European archives of psychiatry and clinical neuroscience* 259, no. 1 (2009): 16-22.

[48] Beurel, Eleonore, Steven F. Grieco, and Richard S. Jope. "Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases." *Pharmacology & therapeutics* 148 (2015): 114-131.

[49] Linding, Rune, Lars Juhl Jensen, Gerard J. Ostheimer, Marcel ATM van Vugt, Claus Jørgensen, Ioana M. Miron, Francesca Diella et al. "Systematic discovery of in vivo phosphorylation networks." *Cell* 129, no. 7 (2007): 1415-1426.

[50] Piovesan, Damiano, Giuseppe Profiti, Pier Luigi Martelli, and Rita Casadio. "The human" magnesome": detecting magnesium binding sites on human proteins." *BMC bioinformatics* 13, no. 14 (2012): 1.

[51] Gout, Elisabeth, Fabrice Rébeillé, Roland Douce, and Richard Bligny. "Interplay of Mg2+, ADP, and ATP in the cytosol and mitochondria: unravelling the role of Mg2+ in cell respiration." *Proceedings of the National Academy of Sciences* 111, no. 43 (2014): E4560-E4567.

[52] Briggs, Katharine T., Gary G. Giulian, Gong Li, Joseph PY Kao, and John P. Marino. "A Molecular Model for Lithium's Bioactive Form." *Biophysical Journal* 111, no. 2 (2016): 294-300.

[53] Willis, J. S., and L. S. T. Fang. "Li+ stimulation of ouabain-sensitive respiration and (Na+-K+)-ATPase of kidney cortex of ground squirrels." *Biochimica et Biophysica Acta (BBA)-Biomembranes* 219, no. 2 (1970): 486-489.

[54] Jope, Richard S. "Lithium and GSK-3: one inhibitor, two inhibitory actions, multiple outcomes." *Trends in pharmacological sciences* 24, no. 9 (2003): 441-443.

[55] Hong, Ming, Daniel CR Chen, Peter S. Klein, and Virginia M-Y. Lee. "Lithium reduces tau phosphorylation by inhibition of glycogen synthase kinase-3." *Journal of Biological Chemistry*272, no. 40 (1997): 25326-25332.

[56] Kitano, Hiroaki. "Systems biology: a brief overview." *Science*295, no. 5560 (2002): 1662-1664.

[57] Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. "KEGG: new perspectives on genomes, pathways, diseases and drugs." *Nucleic acids research* 45, no. D1 (2016): D353-D361.

[58] Szklarczyk, Damian, John H. Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos et al. "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible." *Nucleic acids research* (2016): gkw937.

[59] Zhang JD and Wiemann S (2009). “KEGGgraph: a graph approach to KEGG PATHWAY in R and Bioconductor.” *Bioinformatics*, pp. 1470–1471.

[60] Zhang JD (2017). *KEGGgraph: Application Examples*. R package version 1.38.0.

[61] Franceschini, A (2013). “STRING v9.1: protein-protein interaction networks, with increased coverage and integration.” *Nucleic Acids Research (Database issue)*, **41**

[62] Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research* 37, no. 1 (2008): 1-13.

[63] Thomas, P. D. "Expansion of the gene ontology knowledgebase and resources: the gene ontology consortium." *Nucleic Acids Res* 45 (2017): D331-D338.

[64] Ge, Weihao, Zeeshan Fazal, and Eric Jakobsson. "Using Optimal F-Measure and Random Resampling in Gene Ontology Enrichment Calculations." *bioRxiv* (2017): 218248.

[65] Loughin, Thomas M. "A systematic comparison of methods for combining p-values from independent tests." *Computational statistics & data analysis* 47, no. 3 (2004): 467-485.

[66] Aprahamian, Ivan, Franklin S. Santos, Bernardo dos Santos, Leda Talib, Breno S. Diniz, Márcia Radanovic, Wagner F. Gattaz, and Orestes V. Forlenza. "Long-term, low-dose lithium treatment does not impair renal function in the elderly: a 2-year randomized, placebo-controlled trial followed by single-blind extension." *The Journal of clinical psychiatry* 75, no. 7 (2014): 672-678.

[67] Dall’Era, Marc A., Peter C. Albertsen, Christopher Bangma, Peter R. Carroll, H. Ballentine Carter, Matthew R. Cooperberg, Stephen J. Freedland, Laurence H. Klotz, Christopher Parker, and Mark S. Soloway. "Active surveillance for prostate cancer: a systematic review of the literature." *European urology* 62, no. 6 (2012): 976-983.

CHAPTER 5: An Assessment of True and False Positive Detection Rates of Stepwise Epistatic Model Selection as a Function of Sample Size and Number of Markers

(Submitted to *Heredity*, in Review)

Angela H. Chen, Weihao Ge, William Metcalf, Eric Jakobsson, Liudmila Sergeevna Mainzer*, and Alexander E. Lipka*

***Corresponding authors:**

Alexander E. Lipka alipka@illinois.edu

Liudmila Sergeevna Mainzer lmainzer@life.illinois.edu

Abstract

Association studies have been successful at identifying genomic regions associated with important traits, but routinely employ models that only consider the additive contribution of an individual marker. Because quantitative trait variability typically arises from multiple additive and non-additive sources, utilization of statistical approaches that include main and two-way interaction marker effects of several loci in one model could lead to unprecedented characterization of these sources. Here we examine the ability of one such approach, called the Stepwise Procedure for constructing an Additive and Epistatic Multi-Locus model (SPAEML), to detect additive and epistatic signals simulated using maize and human marker data. Our results revealed that SPAEML was capable of detecting quantitative trait nucleotides (QTNs) at sample sizes as low as n = 300 and consistently specifying signals as additive and epistatic for larger sizes. Sample size and minor allele frequency had a major influence on SPAEML's ability to distinguish between additive and epistatic signals, while the number of markers tested did not. We conclude that SPAEML is a useful approach for providing further elucidation of the additive and epistatic sources contributing to trait variability when applied to a small subset of genome-wide markers located within specific genomic regions identified using *a priori* analyses.

Key words: Epistasis, stepwise model selection, genome-wide association study, quantitative genetics

Introduction

The ability to identify genomic regions containing gene(s) associated with quantitative phenotypes has great potential for elucidating the genetic architecture of traits (e.g. number of genes, their effect sizes, additive vs. non-additive sources), as well as identifying targets for marker-assisted selection in plants and animals and therapy in humans. One analysis that seeks to identify such regions is the genome-wide association study (GWAS), in which statistical analyses are conducted on a set of markers spanning a species' entire genome to determine which marker subsets exhibit the strongest associations with a trait of interest (reviewed in Lipka *et al*, 2015)[1]. In general, statistically significant marker-trait associations suggest that functional variants for the trait under study are located in the surrounding genomic region. To date, GWAS have been able to identify genes associated with many important traits, e.g. predisposition to breast cancer and diabetes in humans [2,3], and provitamin A levels in maize[4]. At present, GWAS is one of the most actively researched and applied methods for investigating the genomic underpinnings of Alzheimer's disease[5], coronary heart disease[6], Parkinson's disease[7], carotenoid biosynthesis in maize[8], and disease resistance in cattle[9], among others. Thus, the ability of GWAS to identify specific genomic regions associated with traits critical for human health and agronomic performance has been demonstrated, and continued refinement of the statistical approaches in GWAS could make this analysis even more relevant for quantitative genetics research and its applications.

The simplest and most widely used analytical approach for GWAS is to perform a separate statistical test for association between each marker and the evaluated trait. For example, a GWAS conducted to identify loci associated with the presence/absence of a disease in humans might perform either a Pearson's chi-square test or conduct logistic regression separately for every marker in a genome-wide marker set[10,11]. Similarly, a GWAS conducted for a quantitative agronomic trait in a given crop[12] might use the unified mixed linear model (MLM)[13] that includes both fixed effect covariates to account for false positives arising from population structure and random effect covariates to account for those arising from familial relatedness.

Although testing each marker individually has been effective in identifying statistically significant marker-trait associations in a wide variety of species and traits, it suffers from two major biological drawbacks. First, the consideration of only one marker at a time makes it impossible to quantify the simultaneous contributions of multiple functional variants located throughout the genome in one statistical model. Second, these single-marker statistical tests typically do not consider the contributions of certain types of non-additive sources of variation, such as epistasis. Improvements to the typical statistical models used for GWAS could lead to more effective models.

Both theoretical[14,15] and empirical[16,17,18] quantitative genetics research suggest that quantitative trait variation is under the control of multiple functional variants. Thus, statistical approaches need to complement this by including multiple markers in one model. Stepwise model selection is one of the simplest approaches for simultaneously estimating the additive effects of multiple loci. Here the additive effect of every marker throughout the genome is considered for inclusion as an explanatory variable in an optimal model. An extremely useful application of this approach is the multi-locus mixed-model (MLMM)[19]. In the MLMM, stepwise model selection is conducted on a given set of markers and false positives are controlled for by including the same fixed and random effects covariates as those used in the unified MLM[13]. An important advantage of the MLMM and similar approaches over single marker analyses is their capability to substantially lower false positive detection rates of marker/trait associations[19]. The MLMM has been shown to be useful for GWAS in crop diversity panels, especially as an extra step to further elucidate the signals already identified by an initial genome-wide scan using the unified MLM[4,20,21]

Another application of stepwise model selection in GWAS is found in the US maize nested association mapping (NAM) panel[22,23,24], where it is called joint linkage (JL) analysis. The maize NAM panel consists of 25 recombinant inbred line (RIL) families that share a common parent. To account for the family structure of the NAM panel, JL analysis starts with a baseline model containing the trait of interest as the response variable and the families as a fixed effect. Stepwise model selection is then conducted, where the nested additive effect of each marker within each family is considered for inclusion into an optimal JL analysis model. The use of JL analysis on the US maize NAM population data has proven fruitful for dissecting the genomic sources of many quantitative traits, including flowering time[22], inflorescence[25] and leaf blight[26]. Although the number of markers considered in these studies is orders of magnitude smaller than those currently available from high-throughput genotypic and/or phenotypic data, it is encouraging that this statistical approach successfully provided insight into the genetic architecture of those traits. To facilitate its broader adoption in GWAS, JL analysis has been made available in the graphical user interface (GUI) of TASSEL5[27], a publicly available Java package.

Non-additive sources of genetic variation are hypothesized to contribute to the discrepancies reported between the observed signals identified in GWAS and what is theoretically expected given the heritability of the trait under study[28]. Epistasis, generally defined as the interaction effect between alleles at two or more genomic loci[29], is one such non-additive source. The direct quantification of epistatic effects by inclusion into multi-locus statistical models could improve our understanding of the genomic architecture of traits. A number of statistical approaches have been described for this purpose (e.g. Cordell, 2002; Haley and Knott, 1992; Jannink and Jansen, 2001; Karkkainen *et al*, 2015)[30,31,32,33] and computationally efficient software has been developed. In particular, FastEpistasis[34], Glide[35], EpiGPU[36], Boost[37], multiEpistSearch[38] and EPIQ[39] explicitly search for pairwise epistasis among a set of markers provided by the user. However, none of the statistical models used in these packages can incorporate contributions from multiple pairs or higher-order combinations of interacting loci. This is a significant drawback, as a substantial proportion of non-additive variation could be attributable to multiple sets of epistatically interacting loci. In this manuscript we evaluate the Stepwise Procedure for constructing an Additive and Epistatic Multi-Locus model (SPAEML), which could potentially remedy that drawback.

We extended the TASSEL5 code for JL analysis to implement SPAEML and tested its ability to detect additive and epistatic QTNs as a function of sample size and number of markers. To achieve this, we used genomic data from 2,648 individuals from the North Central Regional Plant Introduction Station (NCRPIS) maize diversity panel[40] and from an Alzheimer's disease (AD) case-control cohort consisting of 2,099 human subjects[41] to simulate traits with different heritabilities and QTN effect sizes. Since these were not nested association mapping populations, the effect of nesting was not enabled in any of our analyses. We compared SPAEML to two other methods. The first, JL analysis, constructs a multi-locus model for additive marker effects and therefore will always misspecify any epistatic markers included in the model as additive. In contrast, FastEpistasis focuses on the interaction effect of one marker pair at a time; thus any additive signals identified by this approach will be misspecified as epistatic. Our hypothesis was that SPAEML can detect and correctly specify both types of markers.

Materials and Methods

Stepwise procedure for constructing an additive and epistatic multi-locus model (SPAEML)

The statistical approach implemented for SPAEML is similar to those previously described (e.g. Bogdan *et al*, 2004; Yu *et al*, 2008)[24,42]. Briefly, this procedure involves identifying the optimal version of the multi-locus linear model that combines additive and epistatic effects:

$$Y_i = \mu + \sum_{j \in I} \beta_j x_{ij} + \sum_{(u,v) \in U} \gamma_{uv} x_{iu} x_{iv} + \varepsilon_i \qquad \text{(i)}$$

for a data set consisting of *n* individuals and *m* markers denoted by $x_1, \dots, x_m$. In this model, Y_i is the observed trait value of the *i*th individual (e.g., human subject or plant accession); μ is the grand mean; β_j is the additive effect of the *j*th marker; γ_{uv} is the two-way epistatic term between the *u*th and the *v*th marker; I is a subset of the m markers with additive effects included in the model; U is another subset of markers with two-way epistatic effects included in the model; x_{ij}, x_{iu}, and x_{iv} denote the observed genotypes coded additively at the *j*th, *u*th, and *v*th marker loci respectively for the *i*th individual; and ε_i represents a normally distributed random error term. A stepwise model selection procedure is used to determine the optimal sets of markers belonging to I and U.

Simulation Study

- *Genotypic and phenotypic data*

To evaluate the statistical performance of SPAEML we conducted two independent simulation studies: one using genotypic data from a maize diversity panel, and one using genotypic data from a human case-control study. The maize data were from the NCRPIS maize diversity panel[40], consisting of a collection of 2,815 diverse maize inbred lines from throughout the world. We focused on a subset of 2,648 individuals genotyped for 681,257 single nucleotide polymorphisms (SNPs) using genotyping-by-sequencing (GBS)[43]. These data are publicly available at: http://cbsusrv04.tc.cornell.edu/users/panzea/filegateway.aspx?category=Genotypes. The second dataset is from the Mayo Clinic late-onset Alzheimer's disease GWAS, which consists of 844 Alzheimer's disease (AD) cases and 1,255 controls[41]. All 2,099 of these individuals were genotyped using 213,528 SNPs located within +/- 100 kb of 24,526 genes whose transcript levels were measured in Zou *et al* (2012)[41]. These data are available at:
https://www.synapse.org/#!Synapse:syn2910256.

Within each species, we constructed multiple test datasets varying in sample size and number of markers. All test datasets consisted of either the full set of individuals (n = Max; i.e. 2,648 maize or 2,099 human individuals), or the same random subset of n = 300 individuals in each species. Similarly, the test datasets included either a random subset of m = 15,000 SNPs or a random subset of m = 5,000 SNPs. For both species, all SNPs in the 5,000-marker set were also included in the 15,000-marker set.

Traits were simulated as previously described in scheme 2 of (Zhang *et al*, 2010)[44] for each of the above data subsets. First, additive and/or epistatic quantitative trait nucleotides (QTN) were randomly selected from a subset of markers that were present in both the 5,000- and 15,000-marker subsets from each species. For consistency across all simulation settings, the range of possible QTN effect sizes was bounded by 0 and 1. A total of five simulation settings were used (Table 5.1), each with differing numbers of additive and epistatic QTN, their effect sizes, and the broad-sense heritability values (H^2). To empirically evaluate the false positive detection rate of SPAEML in the absence of genomic signals, the traits simulated in the first setting had zero QTN and $H^2 = 0$. The genomic sources of variation underlying the traits in the next setting consisted of four markers that were randomly selected to be additive QTN and four additional marker pairs that were randomly selected to be epistatic QTN. The additive and epistatic QTN both followed a geometric series; that is, the QTN with the j^{th} largest effect size was 0.95^j. Since the purpose of this setting was to evaluate the ability of SPAEML to identify signals for a trait with an ideal genetic architecture, the heritability was set at $H^2 = 0.99$. To assess whether SPAEML can distinguish between additive and epistatic signals, the next setting consisted of two simulated QTN containing both nonzero additive and nonzero epistatic effects. Thus, the additive effects of these two QTN were 0.90 and 0.81, and the epistatic effect of these two QTN was 0.9. All traits simulated at this setting had a broad-sense heritability of $H^2 = 0.95$.

The next simulation setting strove to emulate the genetic architecture of a trait one might expect to find in a crop species. Thus, traits simulated in this setting were loosely based on the contrasting genetic architecture of inflorescence traits between maize and teosinte[16,45]. The genetic underpinnings of these simulated traits consisted of one two-way epistatic QTN of effect size 0.90, 26 additive QTN with the effect size of the j^{th} QTN set to 0.45^j, and a broad-sense heritability of $H^2 = 0.92$. In a similar vein, the next setting was based on the genetic architecture of Alzheimer's disease in humans[46,47,48]. For this setting, a large-effect additive QTN with effect size of 0.90 and a geometric series of 19 additive QTN with the effect size of the j^{th}QTN set to 0.40^jwere simulated. In addition, a two-way epistatic QTN with effect size 0.70 was simulated. To imitate the contributions of the *APOE* gene to Alzheimer's disease[46], one of the two loci contributing to this epistatic QTN was the same as the large-effect additive QTN with effect size of 0.90. Consistent with the literature[48], all traits simulated in this setting had a broad-sense heritability of H^2= 0.34.

A total of 100 traits were simulated for each setting, species, and sample size. For a given simulated trait and sample size, the cumulative additive and epistatic QTN effects were calculated across all individuals. The variance of these cumulative effects comprised the genetic variance component of the trait. Finally, for a given H^2, we simulated a normal random variable with mean 0 and variance σ_r^2, where σ_r^2 is determined from H^2. That is, if we let σ_g^2 denote the genetic variance component of the trait, then the variance of this normal random variable is calculated by solving the following equation for σ_r^2:

$$H^2 = \frac{\sigma_g^2}{\sigma_g^2+\sigma_r^2} \qquad \text{(ii)}$$

Thus, any simulated trait value from a particular individual equals the sum of the cumulative QTN effects and the observed value of the aforementioned normal random variable. For the first setting, with zero QTN and $H^2 = 0$, this normal random variable was simulated with a variance of $\sigma_r^2 = 1$.

- *Statistical models fitted to each trait at each setting*

For each trait that was generated in the simulation study, SPAEML, JL analysis, and FastEpistasis were conducted to identify markers exhibiting peak associations with additive and epistatic QTN. For each of the five simulation settings, sample sizes, species, and number of markers, two separate permutation procedures (described in Churchill and Doerge, 1994)[49] were conducted 100 times to empirically determine the inclusion and exclusion *P*-value thresholds that control the Type I error rate at 0.05: once for SPAEML and once for JL anlaysis. We conducted SPAEML using a Java package derived from the original TASSEL5 suite, but with the added ability to include epistasis (https://bitbucket.org/wdmetcalf/tassel-5-threaded-model-fitter). Additionally, the built-in stepwise model selection procedure from TASSEL was used to conduct the stepwise model selection procedure that only considered additive marker effects, a procedure which we refer to as joint linkage (JL) analysis. The FastEpistasis package was obtained from http://www.vital-it.ch/software/FastEpistasis, and Bonferroni correction was applied to control for multiple testing. FastEpistasis only tests one pair of markers at a time and constructs a model that includes additive effects for each marker and a two-way interaction term that models their epistatic effect.

- *Criteria used to quantify the detection of QTNs*

For a trait simulated under a given sample size, marker number, species, and setting, a QTN was said to have been detected by one of the three statistical approaches if either a marker contributing to the QTN itself or at least one marker located within a surrounding +/- 250 kb window was i.) included as a main (additive) effect in SPAEML or JL analysis, ii.) included as part of a two-way interaction (epistatic) effect by SPAEML, or iii.) included as part of a two-way interaction effect with a *P*-value less than or equal to the Bonferroni-adjusted α = 0.05 threshold when analyzed in FastEpistasis. Thus, an approach's (i.e., SPAEML, JL analysis, or FastEpistasis) detection rate of a QTN was defined to be the proportion of the corresponding 100 simulated traits in which a QTN was detected. A similar metric specific to SPAEML, the specification rate of a QTN, was defined as the proportion of 100 traits where an additive QTN was correctly identified by SPAEML as additive, and both loci contributing to an epistatic QTN were correctly identified by SPAEML as epistatic. A window size of +/- 250 kb has been previously used in maize diversity panels to designate local regions of genomic proximity in maize[50,51]. To enable a side-by-side comparison of results between the two species, the same +/- 250 kb window size was used in the human data.

A false positive (FP) detection was said to occur for i.) each main effect detected by SPAEML or JL analysis corresponding to a marker located outside of the +/-250 kb windows surrounding all QTN and ii.) each two-way interaction effect detected in SPAEML where both corresponding markers were located outside of the +/- 250 kb windows surrounding all QTN, or when iii.) a statistically significant association outside of these windows was identified by FastEpistasis. Hence the FP rate of a given approach was defined to be the proportion of 100 traits simulated at a given sample size, marker number, species, and setting with at least one FP.

Results

We conducted a simulation study to explore the impact of sample size and number of markers on the ability of SPAEML to identify additive and epistatic QTN. To enable a thorough investigation, traits with different genetic architectures ranging in complexity were simulated using genotypic data from a maize diversity panel and then again with genotypic data from a human case-control cohort (Table 5.1). Figure 5.1 shows that the distributions of minor allele frequencies (MAFs) of the 15,000 markers considered in both species are vastly different. While the majority of the 15,000 SNPs in the maize diversity panel have MAFs below 0.1, the majority of the 15,000 SNPs in the human case-control study have MAFs that are greater than 0.1. Within both data sets, the MAFs of the markers randomly selected to be QTNs span the entire range of observed MAFs. These patterns enabled us to observe the way the collective distribution of allele frequencies in a marker set influenced the performance of SPAEML.

Observed false positive rates across the five genetic architectures

The purpose of simulating traits under the "Null" setting was to evaluate the effectiveness of the premutation procedure (used for SPAEML and JL analysis) and the Bonferroni procedure (used for FastEpistasis results) to control the type I error rate at α = 0.05. The observed empirical FP rates across species, number of markers, and sample sizes suggest that these procedures are controlling for type I errors reasonably well, with SPAEML having empirical FP rates that are most consistently close to α = 0.05 (Figure 5.2). The FP rates are generally higher for simulation settings other than "Null," especially for the traits simulated under the "Ideal" genetic architecture (H^2 = 0.99, all QTNs have large effect; Table 5.1) in maize (Supplementary Figure 5.1). These results are not surprising because it is theoretically possible for all three approaches to identify markers that are in linkage disequilibrium (LD) with the simulated QTN. FastEpistasis, which tests the epistatic effect of one pair of loci at a time, tended to yield higher FP rates than the other two stepwise approaches, while SPAEML tended to have low FP rates at the maximum sample sizes in both datasets (Supplementary Figure 5.1).

Accuracy of SPAEML at a limited sample size of n = 300 individuals

The results from these simulation studies show that sample size has a substantially greater impact on QTN detection than the number of markers, underscoring the well-established importance of having sufficient sample sizes when conducting quantitative genetics analysis[52]. Nevertheless, to ascertain the limits of the ability of SPAEML to identify genomic signals, all five simulation settings were run with n = 300 individuals. One of the most detrimental impacts of small sample size on the accuracy of SPAEML appeared to be on the FP rate; substantially high FP rates from SPAEML were observed only at n = 300 (Supplementary Figure 5.1). In contrast, the FP rates for JL analysis and FastEpistasis were more consistent across sample sizes. At n = 300, SPAEML detected QTN at rates vastly superior to those of FastEpistasis, but not as high as those of JL analysis (Figure 5.3a; Supplementary Figures 5.2-5.9). Finally, we observed that at n = 300, SPAEML is more likely to misspecify additive QTN as epistatic and identify only one locus contributing to an epistatic QTN (Figure 5.3b; Supplementary Figures 5.10-5.17). In contrast at n = max, SPAEML yielded i.) minimal FP rates, ii.) QTN detection rates that were comparable to JL analysis, iii.) greater capability to identify both loci underlying epistatic QTN, and iv.) the capacity to distinguish between additive and epistatic signals in traits simulated in the human dataset. In light of this contrast and the negligible impact of the number of tested markers on the simulation results, the remaining sections present findings based on n = max individuals and m = 15,000 markers.

Distinguishing between additive and epistatic signals at the same locus

Among the three approaches that were evaluated, only the output of SPAEML provide results for both additive and epistatic terms fitted to one model. To characterize the ability of SPAEML to distinguish between additive and epistatic signals, both of the QTNs considered in the "Additive vs. Epistatic" setting harbored non-zero additive and epistatic effects. At this setting, we observed contrasting results between the two species. In maize, SPAEML classified the signals at these QTNs as epistatic 100% of the time, suggesting that SPAEML was unable to distinguish between these additive and epistatic effects (Supplementary Figures 5.12-5.13). Contrastingly, SPAEML identified the additive and epistatic signals underlying both QTN simulated in the human dataset for all simulated traits. Similar results were obtained for the Alzheimer's disease-like ("AD-like") setting, where the large-effect additive QTN also include a substantially large epistatic signal (Supplementary Figures 5.16-5.17).

Accuracy in more complex genetic architectures

We compared the accuracy of the three approaches in simulation settings 4 and 5, which approximate the polygenic underpinnings of maize inflorescence ("Inflorescence-like" in Table 5.1) and Alzheimer's disease ("AD-like"). Two important characteristics distinguish these two settings. First, "Inflorescence-like" was highly heritable (H^2 = 0.92) while "AD-like" was not (H^2 = 0.34). Secondly, the effect size of the epistatic QTN was substantially higher relative to those of the additive QTN in the "Inflorescence-like" setting, whereas the strength of the epistatic QTN in the "AD-like" setting was not.

The detection rate for additive QTN improved as a function of the effect size for both JL and SPAEML, with roughly comparable accuracy between the two approaches (Figure 5.4A). In the human dataset, SPAEML provided the added advantage of always correctly identifying both loci contributing to the epistatic QTN, and correctly specifying additive QTNs as a function of their effect size (Figure 5.4A and Supplementary Figures 5.16-5.17). This latter result intuitively makes sense: the stronger the QTN effect, the more likely it is to be distinguished from a non-additive signal. Among the corresponding traits simulated with maize data, SPAEML was at most capable of detecting one out of two loci contributing to an epistatic QTN, and all additive QTNs were misspecified as epistatic (Figure 5.4B). We hypothesize that the generally lower MAF observed in the markers from the maize dataset provided weaker statistical support for each of the simulated QTN, resulting in the observed misspecification.

Discussion

Statistical approaches that consider the additive and epistatic contributions of multiple genomic loci could enable unprecedented quantification of the genetic architecture of agronomically important and human health-related quantitative traits. Using genotypic data from a maize diversity panel and a case-control study of Alzheimer's disease in humans, we conducted a simulation study to determine the accuracy and limits of applicability of SPAEML. Specifically, we assessed the impact of sample size, number of markers, MAF, and the genetic architecture underlying a given trait on the ability of SPAEML to detect and correctly specify additive and epistatic QTN. Our results suggest that sample size has greater influence on the performance of SPAEML than the number of markers, in all considered cases. Additionally, the capability of SPAEML to distinguish between additive and epistatic QTN was much greater when traits were simulated in the human data set, possibly due to the generally higher values of marker MAFs. At the maximum evaluated sample sizes, the detection rate of SPAEML was comparable to JL analysis, and unequivocally superior to that of FastEpistasis.

Our study builds upon previous work[31,32,33,53] that explicitly assesses the ability of stepwise-based or similar approaches to identify and distinguish between additive and epistatic genomic signals. Novel state-of-the-art computational approaches[54] and inexpensive genotyping protocols[43,55] are resulting in extremely large amounts of genotypic and phenotypic data. Larger sample sizes facilitate improved accuracy of analyses. However, exhaustive searches for multiple sets of epistatically interacting loci on a genome-wide scale in large datasets faces a difficult multiple-testing problem[33]. Stepwise model selection and related approaches have been successful in circumventing this problem in the past[25,56,57] by considering a relatively small number of total markers in their analyses; this past success drove us to investigate SPAEML.

Based on our results, we expect SPAEML will be particularly useful for quantifying additive and epistatic marker-trait associations in specific genomic regions that have been identified in *a priori* biological or statistical analyses. This will result in the analysis of a smaller set of markers, thus yielding a smaller search space for optimal models and enabling researchers to capitalize on the accuracy of SPAEML that we demonstrate here. Our exploration of the factors that influence the accuracy of SPAEML is not exhaustive, but is sufficient to complement so-called "search space reduction" efforts[58,59] by providing a rough assessment of the number of markers to target within the genomic regions identified in *a priori* analyses.

Effect of sample size

Our study confirmed the common expectation that sample size positively affects the accuracy of SPAEML. We also demonstrated that SPAEML is capable of true positive detection and even correct specification of additive and epistatic QTN at the smaller sample size that we explored (n = 300). However, the accuracy improves dramatically at larger sample sizes. Although this result is unsurprising, direct quantification of SPAEML's ability to identify additive and epistatic QTN at different sample sizes is informative, as it is useful to know how a model will behave on a smaller dataset when desired sample sizes are unavailable. Our results show that even in those cases SPAEML will find many significant SNPs and epistatic pairs, although they may be misspecified in the final model.

Effect of the marker set size

In contrast to the substantial impact of sample size on the accuracy of SPAEML, we observed similar true and false positive rates at the two marker sizes that were tested. From a statistical perspective, these results suggest that for these simulated data, the conservativeness of the multiple testing problem is similar for both 5,000 and 15,000 markers. Thus, the larger marker set does not decrease the accuracy of SPAEML. This is important, as our marker sets are orders of magnitude smaller than those currently available on a genome-wide scale in heavily researched species. We hope the usefulness of SPAEML holds for larger sets, although direct extrapolation is not recommended. We believe this method is best used on a set of markers that has been whittled down by using prior biological information, such as linkage disequilibrium analysis, hypothetical relations between markers and cellular pathways, or other preliminary analyses that remove markers that are least likely to contribute to the final model function. This will bring the problem into the setting of optimal performance for SPAEML, and also reduce the computational burden from testing both additive and epistatic effects, which grows binomially with the marker set size.

Effect of the minor allele frequency

We found SPAEML to be much more capable of distinguishing between additive and epistatic signals for traits simulated in the human dataset, despite that the same number of markers, similar number of individuals, and the same simulated genetic architectures were evaluated in the maize and human datasets. We propose two distinct but not mutually exclusive hypotheses to explain these results. First, differences between the underlying characteristics of the maize and human genomes could result in LD-related properties being more favorable for SPAEML to work optimally in the human dataset. The second hypothesis is that the differences in accuracy are a downstream ramification of the difference in MAF distribution across the two datasets (Figure 5.1), potentially explained by the procedures for data collection. The maize dataset is a diversity panel, meaning that it consists of a wide variety of genetically diverse species[40]. Thus rare variants are prominent, and consequently SNPs with low MAFs are observed. Although rare variants are undoubtedly also present in the human genome, recent research suggests that the humans tend to be far less genetically diverse than plants, having gone through multiple rounds of purifying selection during inter-continental migrations in human evolution[60,61]. Combined with the fact that the human data we analyzed were from a case-control study, low MAFs are less prominent. In any case, the differences in SPAEML accuracy suggest that both the genomic characteristics of a species and the distribution of MAFs among the tested markers could exhibit a critical impact on the results.

Conclusions and next steps

To ensure that the most appropriate biological conclusions are made by breeding, medical, and quantitative genetics research communities, it is imperative that statistical models which approximate the genetic architecture of traits are accurate. By design, both JL analysis and FastEpistasis oversimplify the intricate patterns of main effects and multifaceted interactions between loci contributing to phenotypic variability. While JL is designed to only consider additive effects in a multi-locus model, FastEpistasis is designed to only test for epistasis, one pair of markers at a time. In contrast, we demonstrate that SPAEML is a sensitive and accurate approach capable of identifying and distinguishing between additive and epistatic genomic signals, at least for datasets of several thousand samples and markers. We suggest that SPAEML, which conducts model selection for all possible main effects and two-way interaction effects of a set of markers, is best used for constructing an accurate model on a limited set of markers identified through an a priori analysis, once the least-contributing markers have already been eliminated. To reduce the inherent computational burden with running SPAEML, we are currently working to migrate the Java code for SPAEML into Scala for a more scalable deployment on Spark. This will enable massive parallelization of the procedure. In the meantime, the Java program that conducts SPAEML is already available to the public free of charge.

Acknowledgements

We would like to thank Timothy M. Beissinger, Gota Morota, and three anonymous referees for their helpful suggestions and insight towards the contribution of epistatic factors to phenotypic variability. Their collective suggestions have made this work stronger and more relevant to quantitative genetics research endeavors. The Alzheimer's Disease study data were provided by the Mayo Clinic Alzheimer's Disease Genetic Studies, led by Dr. Nilufer Taner and Dr. Steven G. Younkin, Mayo Clinic, Jacksonville, FL using samples from the Mayo Clinic Study of Aging, the Mayo Clinic Alzheimer's Disease Research Center, and the Mayo Clinic Brain Bank. AD data collection was supported through funding by NIH grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. The research conducted in this manuscript was supported by the United States Department of Agriculture National Institute of Food and Agriculture project accession number 1005564, the University of Illinois CompGen Student Fellowship Program, the University of Illinois startup funds, and the UIUC Center for Computational Biotechnology and Genomic Medicine (06/01/2017-06/01/2018 NSF IUCRC 2017 at UIUC). We are grateful to the National Center for Supercomputing Applications for the generous provision of computational resources and administrative support. Special thanks to Katherine Kendig for editorial comments.

Conflict of Interest

The authors declare no conflict of interest.

Data Archiving

The Java code used to perform SEMS is available for download at https://bitbucket.org/wdmetcalf/tassel-5-threaded-model-fitter. The genotypic data, simulated trait data, and code to simulate the traits are available at: https://github.com/ncsa/EpiQuant_GWAS_Simulations .

[1] Lipka AE, Kandianis CB, Hudson ME, Yu J, Drnevich J, Bradbury PJ *et al* (2015). From association to prediction: statistical methods for the dissection and selection of complex traits in plants. *Curr Opin Plant Biol* **24:** 110-118.

[2] Billings LK, Florez JC (2010). The genetics of type 2 diabetes: what have we learned from GWAS? *Ann NY Acad Sci* **1212**(1)**:** 59-77.

[3] Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE *et al* (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**(7)**:** 870-874.

[4] Owens BF, Lipka AE, Magallanes-Lundback M, Tiede T, Diepenbrock CH, Kandianis CB *et al* (2014). A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. *Genetics* **198**(4)**:** 1699-1716.

[5] Wang HZ, Bi R, Hu QX, Xiang Q, Zhang C, Zhang DF *et al* (2016). Validating GWAS-Identified Risk Loci for Alzheimer's Disease in Han Chinese Populations. *Mol Neurobiol* **53**(1)**:** 379-390.

[6] Dehghan A, Bis JC, White CC, Smith AV, Morrison AC, Cupples LA *et al* (2016). Genome-Wide Association Study for Incident Myocardial Infarction and Coronary Heart Disease in Prospective Cohort Studies: The CHARGE Consortium. *PloS one* **11**(3)**:** e0144997.

[7] Siitonen A, Nalls MA, Hernandez D, Gibbs JR, Ding J, Ylikotila P *et al* (2017). Genetics of early-onset Parkinson's disease in Finland: exome sequencing and genome-wide association study. *Neurobiol Aging* **53:** 195 e197-195 e110.

[8] Azmach G, Menkir A, Spillane C, Gedil M (2018). Genetic Loci Controlling Carotenoid Biosynthesis in Diverse Tropical Maize Lines. *G3-Genes Genom Genet* **8**(3)**:** 1049-1065.

[9] Coussé A, Francois L, Stinckens A, Buys N, Elansary M, Abos R *et al* (2016). P6038 Tackling the itch: GWAS-based candidate genes for psoroptic mange sensitivity in Belgian Blue cattle. *J Anim Sci* **94**(supplement4)**:** 167-168.

[10] Nakamura M, Nishida N, Kawashima M, Aiba Y, Tanaka A, Yasunami M *et al* (2012). Genome-wide Association Study Identifies TNFSF15 and POU2AF1 as Susceptibility Loci for Primary Biliary Cirrhosis in the Japanese Population. *Am J Hum Genet* **91**(4)**:** 721-728.

[11] Wang S, Zhang Y, Dai W, Lauter K, Kim M, Tang Y *et al* (2016b). HEALER: homomorphic computation of ExAct Logistic rEgRession for secure rare disease variants analysis in GWAS. *Bioinformatics* **32**(2)**:** 211-218.

[12] Belcher AR, Cuesta-Marcos A, Smith KP, Mundt CC, Chen XM, Hayes PM (2018). TCAP FAC-WIN6 Elite Barley GWAS Panel QTL. I. Barley Stripe Rust Resistance QTL in Facultative and Winter Six-Rowed Malt Barley Breeding Programs Identified via GWAS. *Crop Sci* **58**(1)**:** 103-119.

[13] Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF *et al* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**(2)**:** 203-208.

[14] Fisher RA (1930). *The genetical theory of natural selection: a complete variorum edition*. Oxford University Press.

[15] Orr HA (1998). The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. *Evolution* **52**(4)**:** 935-949.

[16] Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S *et al* (2011). Distinct Genetic Architectures for Male and Female Inflorescence Traits of Maize. *PloS Genet* **7**(11).

[17] Flint J, Mackay TF (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Res* **19**(5)**:** 723-733.

[18] Valdar W, Solberg LC, Gauguier D, Burnett S, Klenerman P, Cookson WO *et al* (2006). Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat Genet* **38**(8)**:** 879-887.

[19] Segura V, Vilhjalmsson BJ, Platt A, Korte A, Seren U, Long Q *et al* (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nat Genet* **44**(7)**:** 825-830.

[20] Jaiswal V, Gahlaut V, Meher PK, Mir RR, Jaiswal JP, Rao AR *et al* (2016). Genome Wide Single Locus Single Trait, Multi-Locus and Multi-Trait Association Mapping for Some Important Agronomic Traits in Common Wheat (T-aestivum L.). *PloS One* **11**(7).

[21] Rincker K, Lipka AE, Diers BW (2016). Genome-Wide Association Study of Brown Stem Rot Resistance in Soybean across Multiple Populations. *Plant Genome-Us* **9**(2).

[22] Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C *et al* (2009). The genetic architecture of maize flowering time. *Science* **325**(5941)**:** 714-718.

[23] McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q *et al* (2009). Genetic properties of the maize nested association mapping population. *Science* **325**(5941)**:** 737-740.

[24] Yu JM, Holland JB, McMullen MD, Buckler ES (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics* **178**(1)**:** 539-551.

[25] Brown PJ, Upadyayula N, Mahone GS, Tian F, Bradbury PJ, Myles S *et al* (2011). Distinct Genetic Architectures for Male and Female Inflorescence Traits of Maize. *PloS Genet* **7**(11).

[26] Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011). Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *P Natl Acad Sci USA* **108**(17)**:** 6893-6898.

[27] Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**(19)**:** 2633-2635.

[28] Zuk O, Hechter E, Sunyaev SR, Lander ES (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A* **109**(4)**:** 1193-1198.

[29] Phillips PC (1998). The language of gene interaction. *Genetics* **149**(3)**:** 1167-1171.

[30] Cordell HJ (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* **11**(20)**:** 2463-2468.

[31] Haley CS, Knott SA (1992). A Simple Regression Method for Mapping Quantitative Trait Loci in Line Crosses Using Flanking Markers. *Heredity* **69:** 315-324.

[32] Jannink JL, Jansen R (2001). Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **157**(1)**:** 445-454.

[33] Karkkainen HP, Li Z, Sillanpaa MJ (2015). An Efficient Genome-Wide Multilocus Epistasis Search. *Genetics* **201**(3)**:** 865-870.

[34] Schupbach T, Xenarios I, Bergmann S, Kapur K (2010). FastEpistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* **26**(11)**:** 1468-1469.

[35] Kam-Thong T, Azencott CA, Cayton L, Putz B, Altmann A, Karbalai N *et al* (2012). GLIDE: GPU-based linear regression for detection of epistasis. *Hum Hered* **73**(4)**:** 220-236.

[36] Hemani G, Theocharidis A, Wei W, Haley C (2011). EpiGPU: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* **27**(11)**:** 1462-1465.

[37] Wan X, Yang C, Yang Q, Xue H, Fan X, Tang NL *et al* (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Hum Genet* **87**(3)**:** 325-340.

[38] González-Domínguez J, Kässens JC, Wienbrandt L, Schmidt B (2015). Large-scale genome-wide association studies on a GPU cluster using a CUDA-accelerated PGAS programming model. *Int J High Perform C* **29**(4)**:** 506-510.

[39] Arkin Y, Rahmani E, Kleber ME, Laaksonen R, Marz W, Halperin E (2014). EPIQ-efficient detection of SNP-SNP epistatic interactions for quantitative traits. *Bioinformatics* **30**(12)**:** i19-25.

[40] Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM *et al* (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biol* **14**(6).

[41] Zou F, Chai HS, Younkin CS, Allen M, Crook J, Pankratz VS *et al* (2012). Brain expression genome-wide association study (eGWAS) identifies human disease-associated variants. *PLoS Genet* **8**(6)**:** e1002707.

[42] Bogdan M, Ghosh JK, Doerge RW (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**(2)**:** 989-999.

[43] Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES *et al* (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**(5)**:** e19379.

[44] Zhang ZW, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA *et al* (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* **42**(4)**:** 355-360.

[45] Doebley J, Stec A, Gustus C (1995). teosinte branched1 and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* **141**(1)**:** 333-346.

[46] Combarros O, Cortina-Borja M, Smith AD, Lehmann DJ (2009). Epistasis in sporadic Alzheimer's disease. *Neurobiol Aging* **30**(9)**:** 1333-1349.

[47] Medway C, Morgan K (2014). Review: The genetics of Alzheimer's disease; putting flesh on the bones. *Neuropathol Appl Neurobiol* **40**(2)**:** 97-105.

[48] Wilson RS, Barral S, Lee JH, Leurgans SE, Foroud TM, Sweet RA *et al* (2011). Heritability of different forms of memory in the Late Onset Alzheimer's Disease Family Study. *J Alzheimers Dis* **23**(2)**:** 249-255.

[49] Churchill GA, Doerge RW (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**(3)**:** 963-971.

[50] Chen AH, Lipka AE (2016). The Use of Targeted Marker Subsets to Account for Population Structure and Relatedness in Genome-Wide Association Studies of Maize (Zea mays L.). *G3-Genes Genom Genet* **6**(8)**:** 2365-2374.

[51] Lipka AE, Gore MA, Magallanes-Lundback M, Mesberg A, Lin HN, Tiede T *et al* (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3-Genes Genom Genet* **3**(8)**:** 1287-1299.

[52] Doerge RW (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* **3**(1)**:** 43-52.

[53] Sehgal D, Autrique E, Singh R, Ellis M, Singh S, Dreisigacker S (2017). Identification of genomic regions for grain yield and yield stability and their epistatic interactions. *Sci Rep* **7**.

[54] Gittens A, Devarakonda A, Racah E, Ringenburg M, Gerhardt L, Kottalam J *et al* (2016). Matrix Factorizations at Scale: a Comparison of Scientific Data Analytics in Spark and C plus MPI Using Three Case Studies. *Proc IEEE Int Conf Big Data***:** 204-213.

[55] Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**(2)**:** e1000862.

[56] Mathew B, Leon J, Sannemann W, Sillanpaa MJ (2018). Detection of Epistasis for Flowering Time Using Bayesian Multilocus Estimation in a Barley MAGIC Population. *Genetics* **208**(2)**:** 525-536.

[57] Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S *et al* (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat Genet* **43**(2)**:** 159-U113.

[58] Ritchie MD (2011). Using Biological Knowledge to Uncover the Mystery in the Search for Epistasis in Genome-Wide Association Studies. *Ann Hum Genet* **75:** 172-182.

[59] Wei WH, Hemani G, Haley CS (2014). Detecting epistasis in human complex traits. *Nat Rev Genet* **15**(11)**:** 722-733.

[60] Reich D (2018). *Who We Are and How We Got Here: Ancient DNA and the new science of the human past*. Oxford University Press.

[61] Schlebusch CM, Jakobsson M (2018). Tales of Human Migration, Admixture, and Selection in Africa. *Annu Rev Genomics Hum Genet*.

TITLES AND LEGENDS TO FIGURES

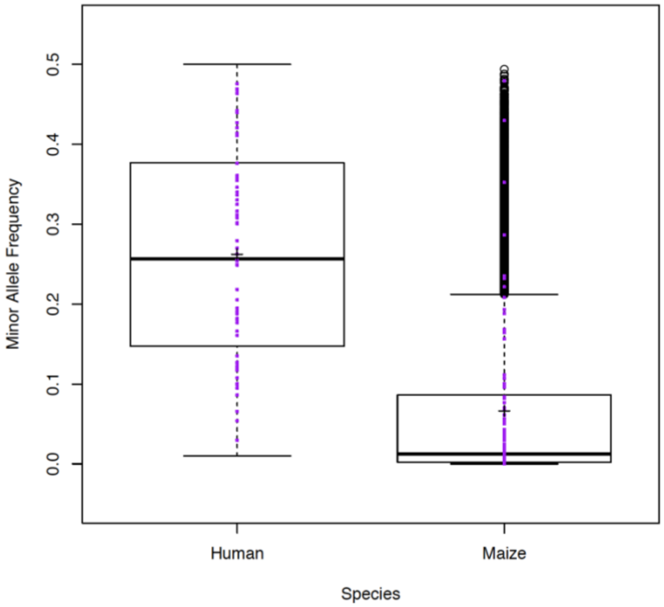


Figure 5.1. Distribution of the minor allele frequencies (MAFs) of the evaluated single nucleotide polymorphisms (SNPs). Box plots depicting the MAFs (Y-axis) of the 15,000 SNPs that were tested in the human dataset and the 15,000 SNPs that were tested in the maize dataset (X-axis). The MAFs of all SNPs that were randomly selected to be quantitative trait nucleotides (QTNs) for the simulation studies are denoted by purple dots. These box plots illustrate that the MAFs of the SNPs in the maize dataset tend to be lower than those in the human dataset.

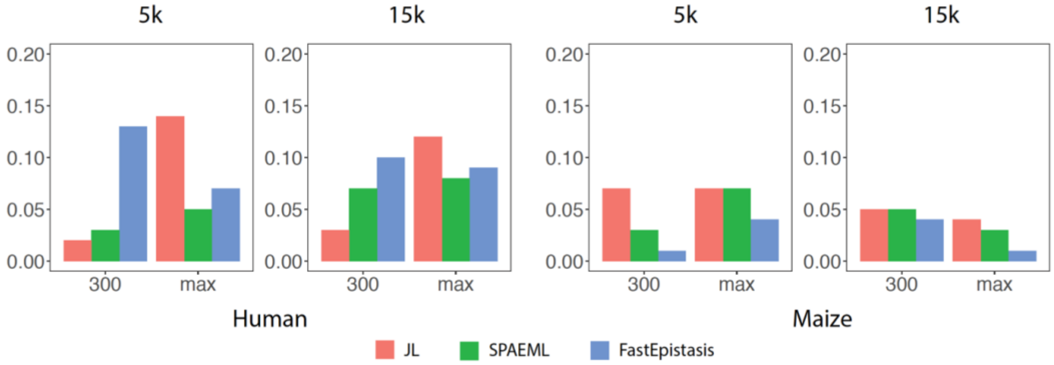


Figure 5.2. Comparison of false positive rates for the three approaches evaluated in "Null" setting where no quantitative trait nucleotides (QTNs) were simulated. The rate of false positive detection, defined as a SNP located outside of +/- 250 kb of any of the QTNs, for joint linkage (JL) analysis, the stepwise procedure for constructing an additive and epistatic multi-locus model (SPAEML), and FastEpistasis are plotted on the Y-axis of each graph. Starting from the left, the first two graphs show the results for the traits simulated in the human data, while the last two columns show the results for the maize simulated data. The graphs with the title "5k" show the results when 5,000 markers were tested, and the graphs with the title "15k" show the results when 15,000 markers were tested. The X-axis of each graph show the sample sizes that were tested, with max indicating the maximum sample size of each dataset (2,648 in the maize dataset and 2,099 in the human dataset).

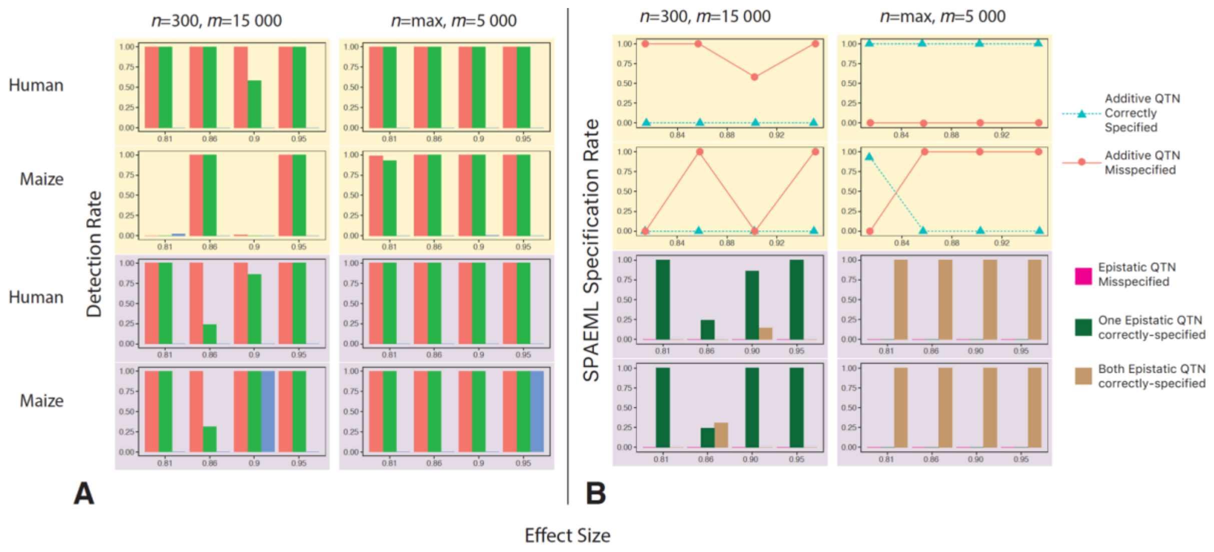


Figure 5.3. Detection (A) and specification (B) rates of simulated quantitative trait nucleotides (QTNs) for the three approaches evaluated in the "Ideal" genetic architecture with setting with four large-effect additive QTN and four large-effect epistatic QTN and heritability equal to 0.99 (A). The detection rates of the additive QTNs, defined as the proportion of SNPs located within +/- 250 kb of any of the simulated QTN detected using joint linkage (JL) analysis (red bar), the stepwise procedure for constructing an additive and epistatic multi-locus model (SPAEML; green bar), and FastEpistasis (blue bar) are plotted on the Y-axis of each graph. The first two rows (shaded pale yellow) show results for the simulated additive QTN, while the bottom two rows (shaded pale purple) show results for the simulated epistatic QTN. The first and third rows show results for the simulations conducted in the human dataset, while the second and fourth rows show results for the simulations conducted in the maize dataset. The X-axis on each graph depict the effect sizes of the QTN. The left column shows results for n = 300 individuals and m = 15,000 markers, while the right column shows results for n = max individuals (i.e., n = 2,099 in humans and n = 2,648 in maize) and m = 5,000 markers. Both JL and SPAEML are able to detect the additive and epistatic effects, while FastEpistasis failed to detect all the additive effects and most of the epistatic effects. (B) Specification rates of SPAEML, defined as the proportion of times that a detected additive QTN was correctly specified in the SPAEML model as additive, misspecified as epistatic (first two rows); or the proportion of times for a detected epistatic QTN that it was misspecified as additive, only one locus contributing to the QTN was detected, or both loci contributing to the QTN (bottom two rows). These proportions are depicted on the Y-axis of each graph. The X-axes of each graph, and how they are subdivided into rows and columns, are the same as in (A). Optimal specification is obtained at n = max; m = 5,000 marker setting and in the human data.

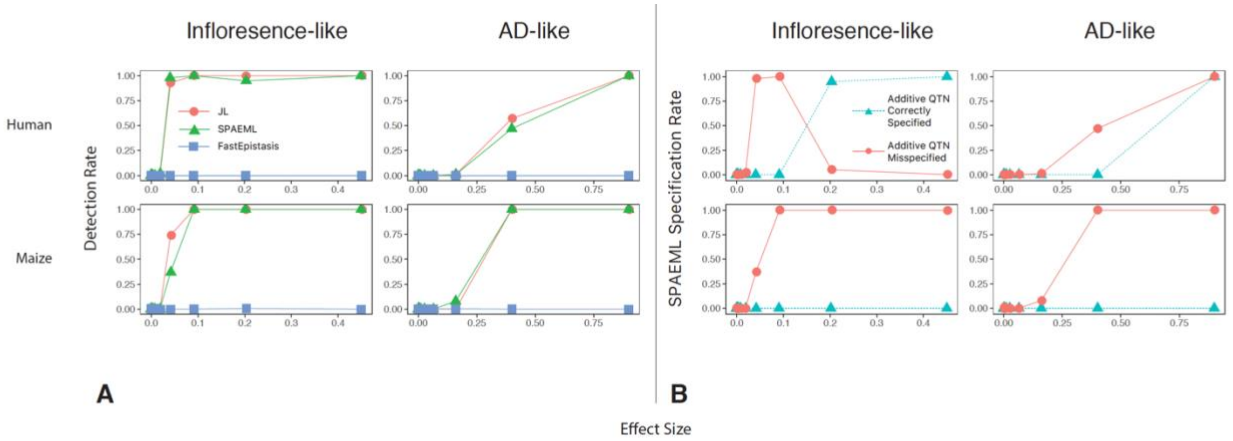


Figure 5.4. Detection (A) and specification (B) rates of simulated additive quantitative trait nucleotides (QTNs) for the three approaches evaluated in the two complex genetic architectures at a maximum number of individuals (*n* = 2,099 human subjects and *n* = 2,648 maize lines) and 15,000 markers (A) The detection rates of the additive QTNs, defined as the proportion of SNPs located within +/- 250 kb of any of the simulated QTNs detected using joint linkage (JL) analysis, the stepwise procedure for constructing an additive and epistatic multi-locus model (SPAEML), and FastEpistasis are plotted on the Y-axis of each graph. The first row shows results for the simulations conducted in the human dataset, while the second row shows results for the simulations conducted in the maize dataset. The X-axis on each graph depict the effect sizes of the additive QTN. The left column shows results for the inflorescence-like genetic architecture, while the right column shows results for the AD-like genetic architecture. Similar detection rates were observed across JL analysis and SPAEML, while FastEpistasis failed to detect all the additive effects. (B) Specification rates of SPAEML, defined as the proportion of times that a detected additive QTN was correctly specified in the SPAEML model as additive or misspecified as epistatic, are depicted on the Y-axis of each graph. The X-axes of each graph, and how they are subdivided into rows and columns, are the same as in (A). Correct specification of additive QTN occurs in the traits simulated using human data. “Inflorescence-like” = setting with 26 additive QTN, one epistatic QTN and heritability equal to 0.92; “AD-like” = setting with 20 additive QTN, one epistatic QTN and heritability = 0.34

Simulation setting	No. of individuals	No. of markers	Heritability	No. of additive QTN[a] (range of effect sizes)	No. of epistatic QTN (range of effect sizes)
1 = “Null”	300; max[b]	5,000; 15,000	0	0	0
2 = “Ideal”	300; max	5,000; 15,000	0.99	4 (0.81-0.95)	4 (0.81-0.95)
3 = “Additive vs epistatic”	300; max	5,000; 15,000	0.95	2 (0.81-0.90)	1 (0.90)
4 = “Inflorescence -like”	300; max	5,000; 15,000	0.92	26 (9.63 x 10^{-10} – 0.45)	1 (0.90)
5 = “AD-like”[c]	300; max	5,000; 15,000	0.34	20 (2.75 x 10^{-8} – 0.9)	1 (0.70)

Table 5.1. Description of the number of individuals, markers, and genetic architecture considered in the five tested simulation settings.

[a] QTN, quantitative trait nucleotide
[b] max, denotes maximum sample size, that is 2,648 maize individuals and 2,099 human individuals
[c] AD, Alzheimer’s disease

CAPTER 6: Future Work - Search Space Reduction Approaches for GWAS Epistatic Model Selection

The main focus of this thesis has been the development of methods to overcome multiple-testing problem in complex biological scenarios. In chapters 1-3, I presented the application of non-parametric resampling method in gene ontology enrichment analysis. This method allows accurate estimation of p-values despite large number of tests performed on the dataset. However, this approach is not the end-all for every biological problem. In some situations, the number of possibilities to explore is so large that even the standard permutation procedure is not sufficient. In chapter 4, the permutation method is applied to evaluate a multi-locus model including both additive and epistatic approach for GWAS data of small size. While that worked well for smaller sample sizes, it was still very computationally intensive. The entire space of all possible models to evaluate grows as an exponential of binomials of the number of genomic markers. When evaluating more than ten thousand markers, brute-force approach becomes computationally infeasible. Therefore, one must first reduce the number of possible models to consider, a process known as "search space reduction". This is best accomplished by narrowing down the number of markers to those most likely to contribute to the phenotype. A number of tool packages and algorithms have been developed to address this problem, yet a consistent workflow is yet to be developed. In this chapter I provide an overview of the approaches that have been discussed in the literature, and suggest a way to synthesize a single, meaningful solution. Generally those approaches cluster into (1) machine learning or statistics algorithms to filter out SNPs that are least likely to carry a signal; (2) removal of SNPs by applying a-priori considerations not related to the biological problem in question, and (3) using prior knowledge about the biological problem being studied to narrow down on the SNPs most likely to be involved.

Removal of SNPS least likely to contribute: LASSO, MDR, MAPIT

It makes sense that among hundreds of thousands of SNPs present in each individual, only a small fraction actually contributes to the phenotype being studied. Thus, it is desirable to have some way to very quickly eliminate SNPs that are clearly unrelated to the problem and thus should not be taken into account when building the genotype-to-phenotype model.

One such method is LASSO: The Least-Absolute Shrinkage and Selection Operator[1]. It efficiently searches for the strongest effects while solving a constrained least-square problem, while constraining the sum of coefficients' absolute value below a fixed number[5].

$$\min_{\beta_0,\boldsymbol{\beta}}\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\}\ +\lambda\sum_{j=1}^{p}|\beta_j|$$

where

$$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{ip})^T\quad for\ the\ i^{th}\ case\ out\ of\ N\ total\ cases$$

$$\lambda\text{: } the\ parameter\ defining\ the\ penalty\ strength$$

LASSO effectively shrinks coefficients of insignificant terms to zero, and could be used to quickly pre-screen a large number of markers before proceeding with actual model construction. The Screen and Clean package[2] is built on this principle that a set of SNPs was selected by LASSO and then fitted into a more precise model. Other examples of using LASSO for GWAS include Graphic-guided fused lasso (GFLASSO)[3] and Adaptive Group LASSO[4]. GFLASSO combines LASSO with genetic regulatory network. In the regulatory network, the SNPs associated with two directly interacting genes would be "fused" by having an additional penalizing term, so that the highly-correlated SNPs would be selected or filtered together[3].

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-\boldsymbol{x}_i^T\boldsymbol{\beta})^2\right\}+\lambda\sum_{j=1}^{p}|\beta_j|+\gamma\sum_{(m,l)interacting\ pair}|f(r_{ml})||\beta_{jm}-sign(r_{ml})\beta_{jl}|,$$

where

r_{ml} is the correlation derived from the interaction network.

Kim, *et.al.* (2009) applied GFLASSO method on both simulation study and a case study on asthma data set and demonstrated that GFLASSO has improved accuracy compared to regular LASSO[3]. Yang, *et.al.* (2010)[4] developed the model by updating penalizing coefficient recursively (adaptive group LASSO, AGL) and applied it on rheumatoid arthritis data set from the Wellcome Trust Case Control Consortium (WTCCC)[5]. Yang, *et.al.* (2010)[4] suggested that while the adaptive group LASSO outperforms regular LASSO in detecting and specifying interactive pairs, it is still limited by the search space and therefore requires search space reduction with biological filters.

By definition, LASSO would quickly converge to the strongest effects. While many epistatic effects do not have a strong additive effect, LASSO is a good tool to build an initial coarse model.

Multifactor Dimensionality Reduction (MDR)[6] is a method of processing SNPs such that their combinations are collectively replaced by a single list of predictors. MDR categorizes allele combinations into high/low risk groups by phenotype values followed by a data-mining procedure. First, data is partitioned into different sections and the program start with one section. For each combination of alleles, the ratio of diseased phenotype is calculated[3]. If the ratio is higher than a pre-set threshold, the combination is considered “high-risk”[3]. Otherwise, the combination would be “low-risk”[3]. Then, the grouping into high and low risk is modified and cross-validated by other data sections[3]. This works very well for case-control studies. Generalized versions of MDR[7,8] also extend its application to quantitative traits. Family Multifactor Dimensionality Reduction (FAM-MDR)[9] includes familial correlations derived from theoretical model. All these approaches derived from MDR method effectively collapse allele combination into 1D vector. SNPHarvester[10] is a package that combines MDR method and LASSO-like regression to identify interacting SNPs, which efficiently detect disease-associated SNPs. The authors for SNPHarvester points out that the package would perform better if guided by knowledge of biological knowledge to pre-select "Good" SNPs[10]. Similarly, prior knowledge of biological pathways, as implemented in Pathway Genetic Load (PGL)[11]. It generates weighed sum of important loci in the pathway associated with the phenotype and uses weighed sum instead of a vector of genotype values for regression. Instead of predicting which combination is most likely to contribute to the trait based only on statistics, PGL pre-selects alleles that are associated with pathways and combines them with a score. The scores would again replace the SNP combinations to be associated with the trait. Crawford, *et.al.* (2017) proposed PGL to evaluate SNPs affecting traumatic injury prognosis for sepsis and death by analyzing loci in the TLR4 signaling and response pathway[14].

Random Forests is a machine learning procedure especially suitable for categorical phenotypes. SNPs are partitioned so that combination of SNPs would lead to almost clear separation of phenotype categories. For example, Goldstein, *et.al.* (2010) applied random forest on multiple sclerosis (MS) case-control dataset and obtained genes bolstered by previous study as well as new candidate genes[12]. However, Winham, *et.al.* (2012) has pointed out that the performance of random forest decreases as dimensionality increases and therefore might not be suitable for complex phenotypes[13].

MAPIT is a method that identifies variants that exhibit non-zero epistatic interactions with any other variant without the need to identify the specific marker combinations that drive the epistatic association[14]. Therefore, it is also a good candidate in pre-selecting candidates for epistatic pair for more detailed fitting later. MAPIT is designed to study quantitative trait with two-way interaction, but the model is readily extensible to categorical trait and higher-order interactions. Since the marginal epistatic effect instead of explicit combination is evaluated, MAPIT can efficiently search candidate SNPs for interactions. However, MAPIT explicitly went through each SNP, which renders it less efficient than LASSO.

Overall, LASSO might be a straightforward and efficient pre-filtering method to quickly build an additive model, if the potential model would be sparse. Other methods might be included later to further improve the pipeline to accommodate different data features.

A-priori biological considerations: LD pruning

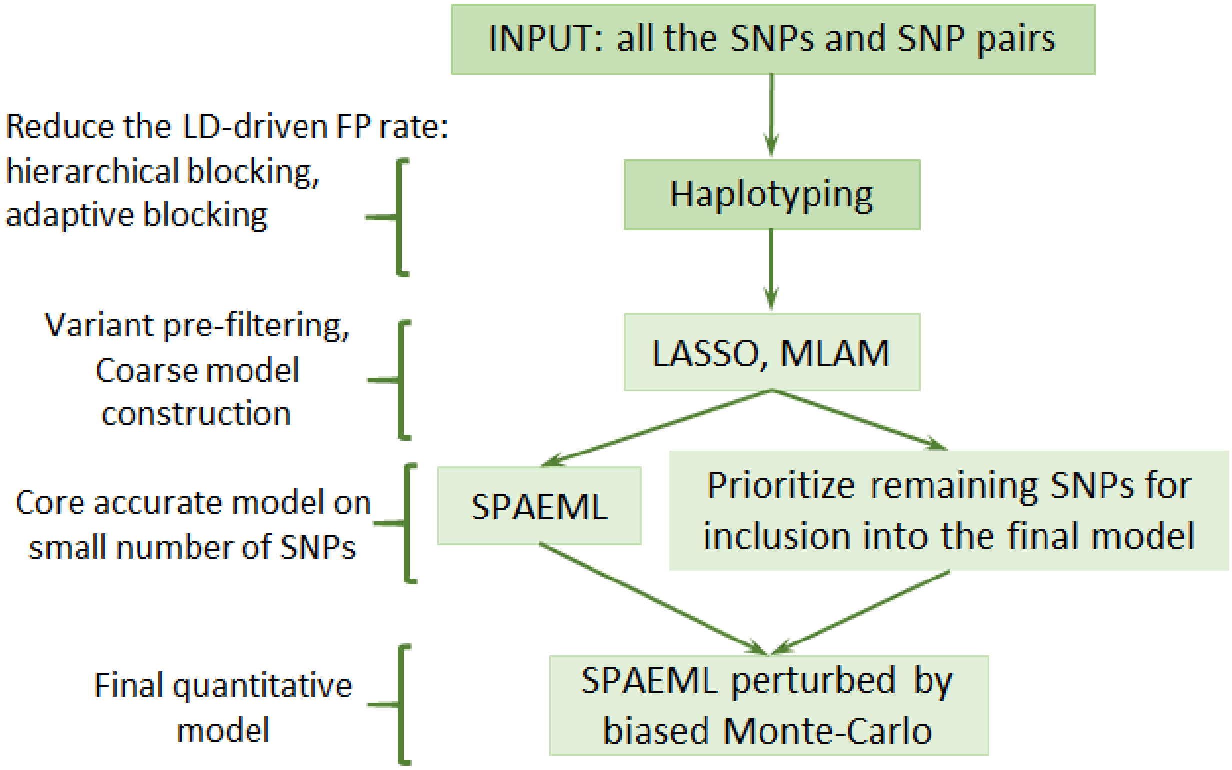
Most organisms, especially complex multicellular eukaryotes, have chromosomal regions called "haplotype blocks" wherein groups of genetic variants are likely to be inherited together[15]. Replacing single SNPs with haplotyping blocks (or randomly choosing one SNP out of a set in one block) drastically reduces the total number of SNPs that participate in the model selection process. Linkage-Disequilibrium is a statistical approach to identify haplotype regions on chromosomes[1]. Software packages LINDEN[16] and BEAM[17] facilitate this process by analyzing high-LD regions. PASCAL[18] and SKAT[19] annotates SNPs onto genes that are in high-LD with them. After haplotyping SNPs, each haplotype group is considered as one variable in the model and represented by a randomly select SNP in the group.

Using information from biological pathways and functions

In addition to haplotyping and PGL, SNP-pair search methods based on prior biological hypotheses have also been explored. For example, Reverse Pathway Genetic Approach starts with knowledge of pathways associated with a Mendelian disorder, identifies SNPs that are in epistasis with the genes of the pathway, and compares these SNPs with rare, major effect mutations involved in the pathway[20]. Ritchie (2011)[21] has reviewed the potential of combining regulatory networks and pathways for filtering out SNPs that are likely to interact with each other. Liu et.al. (2012)[22] selects SNPs in disease-associated regulatory network and performs two-locus analysis to detect epistatic pairs. In group LASSO, an additional penalize term is included so that the coefficient for SNPs sharing an edge in regulatory network would shrink to zero at the same time. These methods involve the accuracy of pre-selecting the candidate SNPs but do not utilize the regulatory network data to expand the final model. Using similar methods, one could build an interaction matrix describing how likely the epistasis pair would form. To build the interaction matrix, several data source can be utilized: (1) sequencing data such as ChIP-Seq for direct transcription factor binding information[23] and Hi-C for chromatin structure[24], (2) networks derived from the sequencing data such as GTEx project[25] and Juicebox toolkit[24], (3) co-expression data from microarray experiments, (4) protein-protein interaction data from BioGRID[26] and STRING[27]. Software like d2z[28] and Jellyfish[29] calculate the k-mer distribution in promoter region, which can be applied to evaluate similarity in the promoter regions and predict co-regulation. MatInspector also calculates promoter region similarity for co-regulation prediction but based on alignment instead of k-mer distribution.[30] A study on Merino sheep pigmentation used MatInspector to construct regulatory network and used the network to predict epistatic pairs[31]. A recent algorithm called SPADIS further give more weights to SNP groups that are distantly connected in a regulatory network to ensure coverage over biologically meaningful pathways that might otherwise blurred by nearby high LD regions[32]. Hi-C data is included in the SPADIS method to consider physical distance of SNPs in a 3D structure[24,32]. (4) Exploration into evolutionary history of chromosome rearrangement might bring some insight into inter-chromosome epistasis.

These biological toolkits and databases indicate the likelihood a SNP (group) would be interacting with another SNP (group). With these information, the candidate SNP to be added into the model would be ranked. The step-wise procedure would then include and evaluate the candidate SNPs stochastically according to their rank.

Proposed workflow to synthesize the above methods into one automated solution



Haplotyping using PASCAL or SKAT would be the first step. These packages not only group SNPs, but also associate them with genes. After this step, SNPs are grouped for reduced number of variable, and annotated to incorporate biological analysis.

Then, a LASSO procedure is conducted to build an initial model. LASSO would quickly filter for the haplotypes with strongest effect as candidate.

Meanwhile, multiple packages are applied in analyzing biological network depending on the input gene list size. If hundreds of genes are annotated, Packages like KnowEng[33] and Magnum[34] would incorporate multiple biological databases to predict candidate interactions. If SNPs of interest are collapsed near only a few genes, probably databases like co-expression[35], or promoter region similarity analysis like D2z[28] would indicate the most likely co-regulatory genes. Moreover, our script generating direct annotation to chromosome region, active CTCF binding sites, and Hi-C data using ANNOVAR[36] would model the contact probability in chromatin regions. Then, a matrix of pairwise interaction probability is generated from all the analysis.

In the SPAEML procedure, candidate SNPs are added into the model following the interaction probability stochastically. When significance is reached, search for model would stop.

In summary, the next stage of the work is to build a modeling pipeline integrating search space reduction method with SPAEML. SNPs are grouped in haplotype blocks and each block is replaced by single representative SNP. A coarse model is built by LASSO. To avoid exhaustive search, candidate SNP and pairs are prioritized with biological information and added into the model in a stochastic way.

[1] Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in medicine*, *16*(4), 385-395.

[2] Wu, J., Devlin, B., Ringquist, S., Trucco, M., & Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic epidemiology*, *34*(3), 275-285.

[3] Kim, S., Sohn, K. A., & Xing, E. P. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics*, *25*(12), i204-i212.

[4] Yang, C., Wan, X., Yang, Q., Xue, H., & Yu, W. (2010). Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC bioinformatics*, *11*(1), S18.

[5] Wtccc, T. (2007). Association scan of 14,500 nsSNPs in four common diseases identifies variants involved in autoimmunity. *Nature genetics*, *39*(11), 1329.

[6] Hahn, L. W., Ritchie, M. D., & Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions. *Bioinformatics*, *19*(3), 376-382.

[7] Agarwal, G., Tulsyan, S., Lal, P., & Mittal, B. (2016). Generalized multifactor dimensionality reduction (GMDR) analysis of drug-metabolizing enzyme-encoding gene polymorphisms may predict treatment outcomes in Indian breast cancer patients. *World journal of surgery*, *40*(7), 1600-1610.

[8] John, J. M. M., Van Lishout, F., & Van Steen, K. (2011). Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data. *European Journal of Human Genetics*, *19*(6), 696.

[9] Cattaert, T., Urrea, V., Naj, A. C., De Lobel, L., De Wit, V., Fu, M., ... & Edwards, T. L. (2010). FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals. *PLoS One*, *5*(4), e10304.

[10] Yang, C., He, Z., Wan, X., Yang, Q., Xue, H., & Yu, W. (2008). SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics*, *25*(4), 504-511.

[11] Huebinger, R. M., Garner, H. R., & Barber, R. C. (2010). Pathway genetic load allows simultaneous evaluation of multiple genetic associations. *Burns*, *36*(6), 787-792.

[12] Goldstein, B. A., Hubbard, A. E., Cutler, A., & Barcellos, L. F. (2010). An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics*, *11*(1), 49.

[13] Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., & Biernacka, J. M. (2012). SNP interaction detection with random forests in high-dimensional genetic data. *BMC bioinformatics*, *13*(1), 164.

[14] Crawford, L., Zeng, P., Mukherjee, S., & Zhou, X. (2017). Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS genetics*, *13*(7), e1006869.

[15] Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *4*(8), 587.

[16] Cowman, T., & Koyutürk, M. (2017). Prioritizing tests of epistasis through hierarchical representation of genomic redundancies. *Nucleic acids research*, *45*(14), e131-e131.

[17] Zhang, Y., Zhang, J., & Liu, J. S. (2011). Block-based Bayesian epistasis association mapping with application to WTCCC type 1 diabetes data. *The annals of applied statistics*, *5*(3), 2052.

[18] Lamparter, D., Marbach, D., Rueedi, R., Kutalik, Z., & Bergmann, S. (2016). Fast and rigorous computation of gene and pathway scores from SNP-based summary statistics. *PLoS computational biology*, *12*(1), e1004714.

[19] Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *The American Journal of Human Genetics*, *92*(6), 841-853.

[20] Mitra, I., Lavillaureix, A., Yeh, E., Traglia, M., Tsang, K., Bearden, C. E., ... & Weiss, L. A. (2017). Reverse pathway genetic approach identifies epistasis in autism spectrum disorders. *PLoS genetics*, *13*(1), e1006516.

[21] Ritchie, M. D. (2011). Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Annals of human genetics*, *75*(1), 172-182.

[22] Liu, Y., Maxwell, S., Feng, T., Zhu, X., Elston, R. C., Koyutürk, M., & Chance, M. R. (2012). Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data. *BMC systems biology*, *6*(3), S15.

[23] ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science*, *306*(5696), 636-640.

[24] Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems*, *3*(1), 99-101.

[25] Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... & Foster, B. (2013). The genotype-tissue expression (GTEx) project. *Nature genetics*, *45*(6), 580.

[26] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., & Tyers, M. (2006). BioGRID: a general repository for interaction datasets. *Nucleic acids research*, *34*(suppl_1), D535-D539.

[27] Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., ... & Kuhn, M. (2014). STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, *43*(D1), D447-D452.

[28] Kantorovitz, M. R., Robinson, G. E., & Sinha, S. (2007). A statistical method for alignment-free comparison of regulatory sequences. *Bioinformatics*, *23*(13), i249-i255.

[29] Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, *8*(8), 1494.

[30] Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., ... & Werner, T. (2005). MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, *21*(13), 2933-2942.

[31] García-Gámez, E., Reverter, A., Whan, V., McWilliam, S. M., Arranz, J. J., Kijas, J., & International Sheep Genomics Consortium. (2011). Using regulatory and epistatic networks to extend the findings of a genome scan: identifying the gene drivers of pigmentation in merino sheep. *PloS one*, *6*(6), e21158.

[32] Yilmaz, S., Tastan, O., & Cicek, A. E. (2018). SPADIS: An Algorithm for Selecting Predictive and Diverse SNPs in GWAS. *bioRxiv*, 256677.

[33] Sinha, S., Song, J., Weinshilboum, R., Jongeneel, V., & Han, J. (2015). KnowEnG: a knowledge engine for genomics. *Journal of the American Medical Informatics Association*, *22*(6), 1115-1119.

[34] Marbach, D., Lamparter, D., Quon, G., Kellis, M., Kutalik, Z., & Bergmann, S. (2016). Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nature methods*, *13*(4), 366.

[35] Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T., & Kinoshita, K. (2014). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic acids research*, *43*(D1), D82-D86.

[36] Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, *38*(16), e164-e164.