Published in final edited form as:

Genet Epidemiol. 2013 September ; 37(6): 614–621. doi:10.1002/gepi.21739.

Encore: Genetic Association Interaction Network Centrality Pipeline and Application to SLE Exome Data

Nicholas A. Davis^{1,a,b}, Caleb A. Lareau^{2,3,a}, Bill C. White¹, Ahwan Pandey¹, Graham Wiley³, Courtney G. Montgomery³, Patrick M. Gaffney³, and B.A. McKinney^{1,2,c}

¹Tandy School of Computer Science, University of Tulsa, Tulsa, OK

²Department of Mathematics, University of Tulsa, Tulsa, OK

³Arthritis and Clinical Immunology Research Program, Oklahoma Medical Research Foundation, Oklahoma City, OK

Abstract

Open source tools are needed to facilitate the construction, analysis, and visualization of gene-gene interaction networks for sequencing data. To address this need, we present Encore, an open source network analysis pipeline for GWAS and rare variant data. Encore constructs Genetic Association Interaction Networks or Epistasis Networks using two optional approaches: our previous information-theory method or a generalized linear model approach. Additionally, Encore includes multiple data filtering options, including Random Forest/Random Jungle for main effect enrichment and Evaporative Cooling and Relief-F filters for enrichment of interaction effects. Encore implements SNPrank network centrality for identifying susceptibility hubs (nodes containing a large amount of disease susceptibility information through the combination of multivariate main effects and multiple gene-gene interactions in the network), and it provides appropriate files for interactive visualization of a network using tools from our online Galaxy instance. We implemented these algorithms in C++ using OpenMP for shared-memory parallel analysis on a server or desktop. To demonstrate Encore's utility in analysis of genetic sequencing data, we present an analysis of exome resequencing data from healthy individuals and those with Systemic Lupus Erythematosus (SLE). Our results verify the importance of the previously associated SLE genes *HLA-DRB* and *NCF2*, and these two genes had the highest gene-gene interaction degrees among the susceptibility hubs. An additional 14 genes previously associated with SLE emerged in our epistasis network model of the exome data, and three novel candidate genes, *ST8SIA4*, *CMTM4*, and *C2CD4B*, were implicated in the model. In summary, we present a comprehensive tool for epistasis network analysis and the first such analysis of exome data from a genetic study of SLE.

Software Availability: <http://insilico.utulsa.edu/encore.php>.

Keywords

epistasis network; machine learning; network analysis; network centrality; Systemic Lupus Erythematosus

^cCorresponding author: Tandy School of Computer Science, University of Tulsa, 800 S. Tucker Drive, Tulsa, OK 74104, brett.mckinney@gmail.com.

^aThese authors contributed equally to this work.

^bCurrent Affiliation: Medical Informatics, University of Oklahoma-Tulsa, Tulsa, OK.

The authors declare no conflicts of interest.

Introduction

A challenge of genome-wide association studies (GWAS) is explaining the variability of disease not attributable to single common variant effects [Maher, 2008]. One proposed solution is to interrogate the genome at a finer scale by sequencing rare variants [Iyengar and Elston, 2007]. Another hypothesis emphasizes the importance of analyzing gene-gene interactions in a complex disease system [Eichler et al., 2010; McKinney and Pajewski, 2011]. The tool described herein, called Encore, focuses on “deep mining” of both common and rare variants to identify susceptibility hubs in a disease/phenotype-specific network. We define a susceptibility hub as a node containing a large amount of disease susceptibility information through the combination of multivariate main effects and multiple gene-gene interactions in the Genetic Association Interaction Network (GAIN). We calculate the degree to which a gene is a susceptibility hub based on a network centrality algorithm we developed called SNPrank [Davis et al., 2010]. In this work, we describe the implementation of Encore and demonstrate its efficiency and utility via performance on simulated GWAS data and the analysis of exome data for Systemic Lupus Erythematosus (SLE) cases and controls.

Recent studies have identified new variants and genes associated with a spectrum of diseases. We recently demonstrated the replication of pathways and increased power to detect important variants using epistasis network centrality analysis of two GWAS for bipolar disorder [Pandey et al., 2012]. Restricting ourselves to a smaller sample size, epistasis network centrality uncovered genetic effects, such as Ankyrin-3 (ANK3) for bipolar disorder, that have come to be understood as important through traditional association methods after recourse to much larger sample sizes. Additionally, in a candidate gene analysis, epistasis networks captured important features of the genetic architecture of bladder cancer [Hu et al., 2011]. These examples demonstrate the usefulness of gene interaction network analysis as applied to complex disease. However, a computationally efficient, statistically powerful, and user-friendly tool is needed to facilitate the widespread examination of interactive effects in other complex diseases. We present Encore as such a tool.

Encore is an inclusive software tool that incorporates machine-learning feature selection, construction of interaction networks, and network centrality ranking of variants based on the aggregation of gene-gene interactions and main effect contributions. Encore further facilitates visualization of synergistic relationships and community structures in a complex network. Additionally, to increase the ease of use for those accustomed to analysis of genetic data, we created a library version of PLINK [Purcell et al., 2007] to simplify the import of standard statistical analysis of genetic data into Encore (and other external software applications as well). Encore thus includes much of the functionality of PLINK while expanding to include support for network and machine-learning analysis. Thus, the machine-learning and other algorithms employed in the early steps of the analysis pipeline are designed to handle PLINK file formats, but Encore also supports a variety of other data formats.

The construction of a GAIN is a fundamental step in the Encore pipeline analysis. In addition to the original Shannon information theoretic formalism for network construction, Encore offers a novel regression-based implementation, called regression GAIN or reGAIN. We previously demonstrated the impact of information theoretic GAIN (itGAIN) in the association of the retinoid X receptor α (RXRA) gene with antibody response to smallpox vaccine [Davis et al., 2010], and a version of reGAIN without covariates was used to identify the association of ANK3 with bipolar disorder [Pandey et al., 2012]. While we have observed scenarios where itGAIN has more power to detect interactions for case-control

studies, the version of reGAIN presented and tested in the current study affords greater flexibility by allowing for both quantitative phenotypes and covariates. Encore can employ either method of network construction.

In order to characterize important nodes in complex epistasis networks, the Encore pipeline includes a fast network centrality algorithm called SNPrank. Motivated by Google's PageRank, SNPrank determines the most important variants in a GAIN matrix [Davis et al., 2010]. A primary goal of GAIN and SNPrank analysis is to discover the importance of variants that have unremarkable individual effects but whose relevance to the phenotype is derived from gene-gene interactions [Davis et al., 2010]. In addition, the multivariate models of SNPrank include main effect contributions. The edges of GAIN are derived from the gene-gene interaction scores and the weighted nodes are derived from their main effect associations. The output of SNPrank is a list of variants ordered based on importance to the phenotype; the scores are network centralities of nodes in the phenotype-specific interaction network. A final annotated network can then be viewed in Cytoscape [Shannon et al., 2003] or using our Galaxy [Blankenberg et al., 2010] instance via any web browser at <http://insilico.utulsa.edu/galaxy>.

As many genetic datasets contain hundreds of thousands, if not millions, of variants, the calculation of interactions between all pairs of variants is computationally prohibitive. To alleviate this computation and remove noise variants, Encore includes a variety of filtering methods. In addition to linkage disequilibrium (LD) pruning and association analysis, Encore includes machine-learning feature selection algorithms such as EC, Relief-F, and Random Jungle. To minimize the computational time required to run Encore, both reGAIN and EC have been parallelized using OpenMP, which takes advantage of all available processors and shared memory on a desktop or server to make the construction of genome-wide networks very fast [OpenMP, 2008]. Furthermore, both reGAIN and SNPrank are written in C++ and integrate a highly optimized linear algebra library. A tutorial of this pipeline can be found at <http://insilico.utulsa.edu/encore-tutorial.php>.

Methods

Figure 1 illustrates the typical Encore analysis pipeline steps, from PLINK data to optional pruning and filtering, gene interaction construction, network centrality, and final visualization. Here, we discuss the implementation details of the major steps. Note that step 0 is not an analysis step *per se* but a representation of the genetic data contained in the original file.

Step 1 is an optional LD pruning step using PLINK's algorithm. We recommend applying this step since many genetic datasets contain hundreds of thousands of variants. This step removes highly correlated variants and reduces network construction time. If the LD pruning threshold removes too many variants, pruning may reduce the power to detect residual interaction effects. Thus, the potential reduction in power should be weighed against the reduction in computation time required for the other methods in the pipeline, particularly the GAIN analysis (step 3). While LD pruning filters for independent variants, the remaining variants/genes will not necessarily be related to the phenotype. To maximize the efficiency of the pipeline, Encore offers three different methods of dimensionality reduction (step 2). These three methods, Evaporative Cooling, Random Jungle, and ReliefF, each have different advantages discussed below.

Encore's implementation of EC [McKinney et al., 2009] uses Random Jungle [Schwarz et al., 2010] for main effect weightings and ReliefF [Robnik-Sikonja, 2003] for sensitivity to interaction effects. Random Jungle (RJ) is available in Encore as a component of EC and for

stand-alone analysis for identifying main effects. RJ is a recent C++ implementation of Random Forests, which has been successful in classification for genetic data [Schwarz et al., 2010]. Random Forests is an ensemble classification technique, where each individual decision tree casts a “vote” for the classification [Breiman, 2001]. RJ employs the same general algorithm as Random Forests but has been customized for application to genetic association studies and is able to harness multiple CPUs and cores, employing the techniques of multi-threading and Message Passing Interface (MPI). Random Jungle is the only tool in the pipeline that uses MPI for distributed parallel computation. All other Encore tools use shared-memory parallelization using OpenMP, which was chosen to meet the large memory requirements for the simultaneous analysis of all variants in very large genome-wide data sets.

Relief-F comprises the other half of EC. While RJ is adept at uncovering main effects, we have shown that it has weak power to identify gene-gene interactions in high-dimensional data [McKinney et al., 2009]. Relief-F, on the other hand, performs well at uncovering epistatic interactions, but feature importance of relevant predictor variables may be underestimated in high-dimensional data. Traditionally, Relief-F has been used as a pre-processing step before a model is learned, but the algorithm can also serve as a good general attribute estimator [Robnik-Sikonja, 2003]. While either of these algorithms can be used individually to filter variants in Encore, EC’s incorporation of both main effect and interactive features provides the most balanced filtering option for discovery in the pipeline.

Step 3 in Fig. 1 represents the statistical interaction network, GAIN. This is the most computationally intensive step in Encore as it calculates the pair-wise interaction between each of the n variants to create a weighted complete graph with n nodes. To construct these networks, Encore provides two network construction methods: itGAIN and reGAIN. The utility of some of the components have been shown previously: EC and itGAIN in Ref. [McKinney et al., 2009], itGAIN and SNPrank in Ref. [Davis et al., 2010], and a previous version of reGAIN plus SNPrank in Ref. [Pandey et al., 2012]. The current reGAIN algorithm in Encore includes novel features that we discuss now in the implementation.

The impetus behind the development of a new version of GAIN is derived from the limitations of the previous information theoretic construction. As itGAIN is constrained to a case/control phenotype, it also has a limited ability to include covariates in an analysis. The reGAIN approach adds the ability to handle quantitative phenotypes, such as antibody titers in an immune response phenotype, and categorical or continuous covariates. The reGAIN algorithm uses a generalized linear model (GLM) for dichotomous phenotypes with a logistic model or continuous/quantitative phenotypes with a linear model. The interaction network weights are calculated using the interaction beta values (β_{n+3}) of the generalized linear model seen in Equation 1. The terms β_1 and β_2 are corrections for the main effects of variants A and B. The covariate terms, $COV_1 \dots COV_n$, and their corresponding beta weights are optional terms in the model that are inserted when the covariate data is available and relevant to the phenotype.

$$Y \sim \beta_0 + \beta_1 A + \beta_2 B + \beta_3 COV_1 + \beta_4 COV_2 + \dots + \beta_{n+2} COV_n + \beta_{n+3} AB + \epsilon \quad (1)$$

An important innovation reGAIN offers is the parallelization of the computation of the interaction terms. Rather than execute the interaction calculations sequentially, reGAIN uses the OpenMP [OpenMP, 2008] framework to divide the computational burden across the multiple cores available on most modern machines and servers. Thus, not only is reGAIN written in a computationally efficient language, C++, but the tool also utilizes multiple

cores, both of which are important features for efficient bioinformatics software [Davis et al., 2011]. The performance characteristics are discussed below.

The primary output of step 3 is the construction of a GAIN matrix, which is used as input into step 4. This $n \times n$ matrix, where n is the number of variants, is a symmetric matrix with the main effect of each variant along the diagonal. The off-diagonal terms are the interaction scores taken from the β_{n+3} calculation. A .sif file (a variation of the Cytoscape Simple Interaction File format) is also created, which lists the interaction between two variants, one per line. This .sif file can then be used to visualize the gene network through two tools: Cytoscape [Shannon et al., 2010] and our interactive web tool, Enviz, available through the previously mentioned Galaxy instance.

As both reGAIN and itGAIN compute all two-way interactions for all variants in the dataset, the visualization immediately after step 3 may lead to a “hairball” appearance (that is, more edges than are reasonably interpretable). Encore alleviates this in two ways: pruning nodes with SNPrank and edges with a false discovery rate (FDR) procedure. SNPrank ranks the importance of each genetic node in the interaction network based on its main effect and the context of the whole network [Davis et al., 2010]. The SNPrank scores are elements of a normalized eigenvector, and, thus, can be treated as a probability field. One can then filter variants based on QQ plots of the SNPrank scores [Pandey et al., 2012]. Filtering and scaling node radii based on SNPrank scores enables the user to distinguish the most important nodes in the interaction network and visualize those genetic features in the phenotype-specific interaction network.

To prune edges of a reGAIN network, Encore takes advantage of the statistical distribution of the interaction coefficient statistics. An edge threshold may be established by manually choosing a p-value and β_{n+3} threshold for interactions in reGAIN or by selecting an FDR coupled with the Encore implementation of the Benjamini Hochberg (BH) [Benjamini and Hockberg, 1995] FDR correction. When the uncorrected p-value exceeds the BH rejection threshold, the interaction score goes to 0, effectively removing the interaction from the network. In addition, Encore has a utility to create volcano plots of edge weights versus $-\log$ p-values and then to partition the plot for additional edge pruning. The volcano utility uses the $-\log$ p-value from the FDR calculation to create a horizontal cut, below which edges are pruned, and then a vertical cut is created such that the FDR percentage of the accepted points with the lowest edge weights may also be pruned from the network.

To assess the performance of Encore, we benchmark the software on artificial GWAS data with 700 subjects with quantitative traits and covariates. We then analyze real data from an exome resequencing experiment of 209 SLE cases and 195 healthy controls of European Ancestry. Exome sequence enrichment was performed using the Truseq Exome Enrichment kit as per the manufacturers protocols (Illumina, San Diego, CA). Sequencing of the enriched exome pools was performed on an Illumina HiSeq 2000 instrument with 100bp, paired-end reads. Sequences were aligned to the human reference sequence hg19 using the Burrows-Wheeler Aligner (BWA)[Li and Durbin, 2009] followed by local realignment or problematic indels and base quality score recalibration with the Genome Analysis Tool Kit (GATK)[McKenna et al., 2010]. Variants were identified using the GATK Unified Genotyper program. Variants were filtered from further analysis if they did not meet the following criteria: Alignment quality score of 30 or greater, a read depth of at least 8, a quality by depth score of at least 2.5, presence of a homopolymer run of 5 bases or less, a strand bias score of less than -10.0 , and a map quality score greater than 25.

Results

We first tested performance benchmarks of Encore on an artificial GWAS dataset with a quantitative phenotype and covariates. This synthetic data has dimensions (700 subjects, 650,000 variants) similar to a recent vaccine GWAS study that used antibody response as the phenotype and included multiple covariates [Ovsyannikova et al., 2012]. We used two common types of machines for benchmarking: a desktop (Quad Core Intel Xeon W3520 2.66GHz CPU, 6GB RAM) and a high-performance server (12-Core Intel Xeon Z5650 2.67 GHz CPU, 48GB RAM). Table Ia shows the performance results for the main options in Step 2 of Figure 1 as well as PLINK association. Each of these methods assigns either a score or p-value to each variant, allowing a user to determine the most appropriate variants for further network-analysis. While PLINK association is not computationally as intensive as the other methods, this method also does not consider the potential for interactions, likely inhibiting the inclusion for highly interactive variants. As PLINK algorithms execute serially, there was no clear advantage to using the server over the desktop for analysis. However, other algorithms like EC not only consider interactions and main effects when assigning scores but also take advantage of additional cores available. Though EC is the most computationally intensive filter method in Step 2, our previous simulation studies [McKinney et al., 2009] and the increased enrichment of known SLE associated genes indicate that EC is the strongest method in Encore for later discovery in an epistasis network.

Table Ib shows the results of the computational efficiency of the novel reGAIN and C++ SNPrank algorithms when executed in an environment of multiple cores. This step follows the feature selection step, which resulted in a smaller data set for reGAIN construction. Data set sizes of 100, 1000, and 10,000 variants were used to determine run times for reGAIN and SNPrank. The data reported in this table indicates a clear advantage in using a machine with additional cores for analysis, illustrating the importance of OpenMP in the Encore software. While reGAIN is clearly the most computationally intensive piece in the Encore pipeline, we note that its performance increases dramatically with a reduction in the number of variants analyzed. Thus, we stress the importance of careful variant selection using the methods in Step 2.

While Encore was originally designed for GWAS data, we present the results of a novel application of the software to an exome study of SLE. The goal of this exome network analysis is to refine our understanding of the genetic etiology of SLE by analyzing known genetic factors within the context of data-driven network interactions. The inferred network allows for a more detailed characterization of known associations and discovery of new associations. Using the coordinates of a manually curated list of 40 known SLE genes (Supplementary Table 1), we constructed a dataset of 719 bi-allelic and LD-pruned variants merged with 1281 additional variants selected by the Evaporative Cooling algorithm. This combined dataset was designed to visualize known SLE genes in an interactive network and to enrich for discovery of relevant variants based on gene-gene interactions. In total, these 2,000 variants were analyzed using reGAIN and subsequently SNPrank. We selected variants with a SNPrank network centrality score greater than .01, and we pruned interactions with β_{n+3} scores less than 20. These edge threshold values were chosen to create a visually clear and concise interaction network. The variants meeting the SNPrank threshold were mapped to genes, creating the 31 unique nodes shown in our centralized network. Figure 2 shows a static version of this epistasis network. An interactive web-based network visualization can be explored at http://insilico.utulsa.edu/Encore_SLE_Exome.html

Table II contains a summary of the nodes used to construct the network in Figure 2. Though Encore was originally designed for GWAS data, our novel application to Exome data

showed no evidence for bias toward selecting common variants. In fact, only 1 of the top 31 variants determined by SNPrank had a minor allele frequency (MAF) above 5% in the data. While the MAFs of 23 of the variants were between 1 and 5%, 7 variants implicated in the epistasis network were very rare (MAF < 1%). While most existing methods of associating rare variants employ some form of collapsing method, our analysis shows there is merit to incorporating individual variants into a multivariate network model.

The biological results from this analysis of SLE exome data convey Encore's utility in discovery and formulating hypotheses pertaining to a complex disease. To estimate the enrichment of known SLE effects in the exome epistasis network (Fig. 2), we manually curated a list of 40 SLE genes (Supplementary Table 1) that were found in the literature to have at least one genome-wide significant variant ($p < 5 \times 10^{-8}$) from studies of populations with multiple ethnic backgrounds. Of the 31 most significant nodes in Fig. 2 based on SNPrank network centrality for the European Ancestry exome data, 14 nodes (marked in orange) remained in the final network from the set of known 40 SLE genes. In contrast (see below), only 1 of the known SLE genes remains when using univariate filtering in the pipeline.

The remaining 17 genes in Fig. 2, labeled in blue, were nodes added through EC enrichment. Though absent from our list of 40 SLE genes, TBX21 and PHRF1 (blue in Fig. 2) have both been associated with Lupus [You et al., 2010; Salloum et al., 2010]. The gene TBX21 was not included on our list because it was associated with SLE as a risk haplotype instead of through a GWAS association [You et al., 2010]. While PHRF1 and IRF7 were both identified in the same linkage study, the biological role of IRF7 in the immune system was the primary gene identified in the study [Salloum et al., 2010]. Thus, IRF7 was present on our curated gene list while PHRF1 was not included. The results of this study, however, implicate PHRF1, suggesting both genes could play a role in SLE etiology. The selection of these nodes by EC and SNPrank verify the Encore's ability to uncover genes relevant to the disease phenotype.

The gene corresponding to our strongest SNPrank association was ST8SIA4. This gene encodes an enzyme required for the synthesis polysialic acid, which primarily functions as modulator of neural cell adhesion [Becker et al., 1996]. However, a more recent study has suggested a role for polysialic acid in immune regulation [Drake et al., 2008]. Other genes, such as C2CD4B, involved in cell adhesion and inflammation [Warton et al., 2004], and CMTM4, which has been shown to be over-expressed during cellular exposure to the Epstein-Barr Virus [Ehlin-Henriksson et al., 2009], are new, promising candidate genes that have yet to be identified with traditional association methods.

While Encore's ability to uncover new candidate genes is a valuable feature of the software, the edges displayed in the network can also lead to new hypotheses about the genes involved in complex disease pathways. The connection between JAKMIP1 and TYK2 demonstrates this usefulness, as their interaction helps mediate STAT4 activation [Watford et al., 2004; Raelson et al., 2007]. As STAT4 is another gene associated with SLE, this particular edge illustrates the power in this method. Furthermore, while NCF2 and IRF8 are involved in B-cell activation and differentiation and both previously identified as Lupus genes [Busslinger, 2004; Vene et al., 2010; Cunninghame Graham et al., 2011], no study has tested the interaction between these two genes on SLE onset. As our epistasis network shows a strong interaction between these two genes, we hypothesize the mechanism of these two genes acting on B-cells could shed new light on the SLE phenotype.

Though each method in Encore is informative as a stand-alone tool, the sequential application optimizes the information that can be gathered through a genetic network

analysis. To better demonstrate the importance of this sequential application and thus the utility of the novel pipeline we present herein, we compared our results to a model without using EC. Specifically, we replaced the 1281 variants promoted by EC with the same number of variants with the lowest p-values from a traditional association test and we repeated the other steps in the analysis. While the resulting graph contained multiple valid biological interactions, only 1 gene from the manually curated list of 40 SLE genes (Supplementary Table 1) remained in the resulting univariate-filtered exome network versus 14 out of 40 for the EC-filtered exome pipeline (orange nodes in Fig 2). Moreover, the nodes and interactions had much weaker SNPrank and interaction scores than those displayed in Fig. 2. This result demonstrates the need for EC feature selection step along with the epistasis network step in the Encore pipeline (Fig. 1) for increased power to identify disease-relevant variants.

Discussion

To address the apparent “missing” heritability for common phenotypes, attention has turned to the role of gene-gene interactions and rare variants. We recently identified novel effects among common variants by aggregating main effects and the network of gene-gene interactions. In the current study, we described this epistasis network pipeline and demonstrated its utility for rare variants with an application to exome data for SLE. Not only did Encore replicate over a dozen genes previously associated with SLE, but the pipeline also uncovered new, biologically plausible candidate genes and interactions. Furthermore, the presence of strong interactions in our enriched data-driven network generates new hypotheses for the activity of two or more genes on a complex disease phenotype, specifically SLE. Encore provides an inclusive means for generating and visualizing these statistical interactions.

Encore provides an innovative pipeline to analyze genetic data (both GWAS and rare variants) in the context of a phenotype-specific interaction networks. By incorporating PLINK as a library and providing support for its data formats, Encore provides a user-friendly interface for anyone familiar with genetic association analysis. By optimizing our algorithms using OpenMP, Encore can efficiently analyze large datasets and provide concise, annotated network results. With our Galaxy instance, we provide a simple, web-based interface to allow users an opportunity to evaluate these tools for a small test dataset prior to deciding to install Encore locally. Source code and tutorials are available for users and developers to run on local desktops and servers.

As with any network construction, two fundamental issues arise: what are the appropriate set of nodes (variants) and set of edges (gene interactions) to include in the final network construction? For nodes, we offer a variety of machine learning filters to reduce the number of variants to a more manageable number. In particular, the Evaporative Cooling algorithm weights variants on both main effects and interactions, often providing the most promising variants for analysis in an interaction network. Further, the C++ implementation of SNPrank uses a network centrality algorithm to determine the most important nodes in a network. These tools reduce the potentially hundreds of thousands of nodes down to a more tractable number for visualization. For edge selection, reGAIN offers a false discovery pruning option and a volcano plot utility to remove interactions that are consistent with the null hypothesis. As each edge is assigned a particular interaction score, a simple threshold can be applied to determine the strongest interactions.

In addition to standard genetic association measures, such as p-values, regression coefficients, and odds ratios, Encore provides information on each gene-gene interaction and a centrality score (SNPrank) for each variant that represents a global statistical network

effect. The connections, edge weights, node weights, and centralities of the epistasis network adapt to the context of the given dataset, perturbation or phenotype. Thus, the power of Encore lies in its ability to generate hypotheses and biological mechanisms for genes in a complex interaction network while complementing the metrics and discoveries of other tools.

This study has focused on Encore's optimization of network analysis with quantitative and categorical phenotypes and covariates in GWAS and exome sequencing data. In this study, we treated environmental factors as covariates. However, within the Encore tool, environmental factors may also be treated as nodes in a heterogeneous network alongside genetic nodes. In addition to GWAS and exome data, Encore may be applied to the network analysis of methylation data. Other tools in Encore support the analysis of quantitative attributes, such as mRNA-Seq, Microarray gene expression, and neuroimage data. For these utilities, we plan to refine the support of multiple data formats and data types for network analysis. Ultimately, we aim to extend Encore to include comprehensive integrative network analysis, in which multiple heterogeneous data, including environmental factors, are combined in a unified network.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We are grateful for Jennifer Kelly's work for creating and sustaining the supplementary table of SLE genes. We would also like to thank Stuart Glenn for his IT support in managing the Exome data and James Robertson for his technical assistance in producing the sequence libraries.

Funding: The development of Encore was supported by NIH Grant no. R56 AI-80932 (PI: BA McKinney). Exome resequencing was supported by NIH Grants nos. RC2 AR058959 (PI: PM Gaffney), R56 AI063274 (PI: PM Gaffney), and P20 GM103456 (PI: PM Gaffney). Dr. McKinney is also supported by the William K. Warren foundation.

References

- Becker CG, Artola A, Gerardy-Schahn R, Becker T, Welzl H, Schachner M. The polysialic acid modification of the neural cell adhesion molecule is involved in spatial learning and hippocampal long-term potentiation. *J Neurosci Res.* 1996; 45(2):143–152. [PubMed: 8843031]
- Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological.* 1995; 57(1): 289–300.
- Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol.* 2010; Chapter 19(Unit 19):101–121.
- Breiman L. Random forests. *Machine Learning.* 2001; 45(1):5–32.
- Busslinger M. Transcriptional control of early B cell development. *Annu Rev Immunol.* 2004; 22:55–79. [PubMed: 15032574]
- Cunninghame Graham DS, Morris DL, Bhangale TR, Criswell LA, Syvanen AC, Ronnblom L, Behrens TW, Graham RR, Vyse TJ. Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with systemic lupus erythematosus. *PLoS Genet.* 2011; 7(10):e1002341. [PubMed: 22046141]
- Davis NA, Crowe JE Jr, Pajewski NM, McKinney BA. Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. *Genes Immun.* 2010; 11(8):630–636. [PubMed: 20613780]

- Davis NA, Pandey A, McKinney BA. Real-world comparison of CPU and GPU implementations of SNPrank: a network analysis tool for GWAS. *Bioinformatics*. 2011; 27(2):284–285. [PubMed: 21115438]
- Drake PM, Nathan JK, Stock CM, Chang PV, Muench MO, Nakata D, Reader JR, Gip P, Golden KP, Weinhold B, et al. Polysialic acid, a glycan with highly restricted expression, is found on human and murine leukocytes and modulates immune responses. *J Immunol*. 2008; 181(10):6850–6858. [PubMed: 18981104]
- Ehlin-Henriksson B, Liang W, Cagigi A, Mowafi F, Klein G, Nilsson A. Changes in chemokines and chemokine receptor expression on tonsillar B cells upon Epstein-Barr virus infection. *Immunology*. 2009; 127(4):549–557. [PubMed: 19604305]
- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010; 11(6):446–450. [PubMed: 20479774]
- Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*. 2011; 12:364. [PubMed: 21910885]
- Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods Mol Biol*. 2007; 376:71–84. [PubMed: 17984539]
- Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25(14):1754–1760. [PubMed: 19451168]
- Maher B. Personal genomes: The case of the missing heritability. *Nature*. 2008; 456(7218):18–21. [PubMed: 18987709]
- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20(9):1297–1303. [PubMed: 20644199]
- McKinney BA, Crowe JE, Guo J, Tian D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genet*. 2009; 5(3):e1000432. [PubMed: 19300503]
- McKinney BA, Pajewski NM. Six Degrees of Epistasis: Statistical Network Models for GWAS. *Front Genet*. 2011; 2:109. [PubMed: 22303403]
- OpenMP. Application Program Interface. 2008 Available from <http://www.openmp.org/mp-documents/spec30.pdf>.
- Ovsyannikova IG, Kennedy RB, O'Byrne M, Jacobson RM, Pankratz VS, Poland GA. Genome-wide association study of antibody response to smallpox vaccine. *Vaccine*. 2012; 30(28):4182–4189. [PubMed: 22542470]
- Pandey A, Davis NA, White BC, Pajewski NM, Savitz J, Drevets WC, McKinney BA. Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder. *Transl Psychiatry*. 2012; 2:e154. [PubMed: 22892719]
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007; 81(3):559–575. [PubMed: 17701901]
- Raelson JV, Little RD, Ruether A, Fournier H, Paquin B, Van Eerdewegh P, Bradley WE, Croteau P, Nguyen-Huu Q, Segal J, et al. Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc Natl Acad Sci U S A*. 2007; 104(37):14747–14752. [PubMed: 17804789]
- Robnik-Sikonja M, Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*. 2003; 53(1–2):23–69.
- Salloum R, Franek BS, Kariuki SN, Rhee L, Mikolaitis RA, Jolly M, Utset TO, Niewold TB. Genetic variation at the IRF7/PHRF1 locus is associated with autoantibody profile and serum interferon-alpha activity in lupus patients. *Arthritis Rheum*. 2010; 62(2):553–561. [PubMed: 20112359]
- Schwarz DF, König IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*. 2010; 26(14):1752–1758. [PubMed: 20505004]

- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*. 2003; 13(11):2498–2504. [PubMed: 14597658]
- Vene R, Delfino L, Castellani P, Balza E, Bertolotti M, Sitia R, Rubartelli A. Redox remodeling allows and controls B-cell activation and differentiation. *Antioxid Redox Signal*. 2010; 13(8): 1145–1155. [PubMed: 20367281]
- Warton K, Foster NC, Gold WA, Stanley KK. A novel gene family induced by acute inflammation in endothelial cells. *Gene*. 2004; 342(1):85–95. [PubMed: 15527968]
- Watford WT, Hissong BD, Bream JH, Kanno Y, Muul L, O'Shea JJ. Signaling by IL-12 and IL-23 and the immunoregulatory roles of STAT4. *Immunol Rev*. 2004; 202:139–156. [PubMed: 15546391]
- You Y, Zhao W, Chen S, Tan W, Dan Y, Hao F, Deng G. Association of TBX21 gene haplotypes in a Chinese population with systemic lupus erythematosus. *Scand J Rheumatol*. 2010; 39(3):254–258. [PubMed: 20429676]

Figure 1. Flow chart illustrating analysis steps with Encore

The rectangles with a folded corner represent various input and output files used in the pipeline. Rectangles with rounded corners represent processes in the overall analysis. Optional paths in the workflow are labeled with dashed arrows while the recommended paths have blue shaded arrows.

Figure 2. Genetic Association Interaction Network for SLE Exome data

The nodes of the graph represent the 31 top variants (SNPrank score $> .01$) determined by the Encore pipeline (Evaporative Cooling feature selection, reGAIN, and SNPrank). The variants are mapped to their respective genes. Nodes labeled in orange represent known SLE-associated genes that SNPrank replicates in the exome analysis. Genes in blue represent discovery genes. The 61 edges shown in blue had an interaction score greater than 20 (top 1.3% of all interactions) as computed by reGAIN. The radius of each node is scaled by its degree (number of edges).

Table I**Encore Benchmarks**

(a) Runtimes for Encore filters (Evaporative Cooling, Random Jungle, and Relief-F) and PLINK Association for an artificial GWAS with quantitative phenotype and covariates (700 samples $\times$ 650,000 SNPs) performed on a 4-core desktop and a 12-core server. (b) Runtimes for reGAIN and SNPrank on a quad-core desktop (4) and a 12-core server (12) for different numbers of SNPs (10k, 1k, and 100 SNPs). In a typical analysis, one of the methods in (a) is first used to filter the features for network construction and analysis based on methods in (b).

(a)	EC	RJ	ReliefF	PLINK Assoc.
4 cores	2.48h	55m3s	1.73h	5m21s
12 cores	1.17h	23m9s	53m9s	5m12s

(b)	reGAIN (4)	reGAIN (12)	SNPrank (4)	SNPrank (12)
10k variants	5.45h	2.75h	19m11s	2m53s
1k variants	3m24s	1m19s	2.5s	1.8s
100 variants	13s	12s	< 1s	< 1s

Table II

Summary of SLE Lupus Interaction Network

Top genes found by Encore network analysis applied to SLE exome data. Rows are sorted by SNPPrank epistasis network centrality score. Columns include the gene name, hg19 coordinates of the variant, rsID when applicable, degree of the node (seen in Figure 2), SNPPrank score (network centrality), and minor allele frequency.

Gene	Chr-BP	rsID	Degree	SNPPrank Score	MAF
ST8SIA4	5-100146923	rs80313395	4	5.96E-02	0.06117
HLA-DRB	6-3723609	rs79944865	8	5.19E-02	0.02394
IRF8	16-85955663	rs305072	4	4.19E-02	0.007979
TBX21	17-45810919	rs2240017	3	4.15E-02	0.01242
ZNF643	1-40915825	-	8	3.73E-02	0.006098
TYK2	19-10488926	rs55762744	5	3.68E-02	0.02139
NCF2	1-183532364	rs35012521	7	3.58E-02	0.005263
BANK1	4-102751014	rs35978636	2	3.23E-02	0.02394
JAZF1	7-27871845	-	6	3.06E-02	0.01064
ETS1	11-128332262	rs34846069	6	3.02E-02	0.01862
USP6NL	10-11502653	-	3	2.90E-02	0.01042
AFF1	4-88036116	-	6	2.78E-02	0.01058
ARID1B	6-157100092	-	4	2.76E-02	0.04545
ITSN2	2-24583303	-	4	2.71E-02	0.01623
WDFY4	10-50029121	rs79406124	5	2.62E-02	0.007979
CMTM4	16-66730270	rs117522360	5	2.58E-02	0.03158
ZNF260	19-37003606	rs73032471	1	2.37E-02	0.0102
JAKMIP1	4-6202236	rs4689358	4	2.31E-02	0.01796
EIF4G1	3-184032388	-	3	2.30E-02	0.02695
C2CD4B	15-62456982	-	3	2.30E-02	0.01437
TNIP1	5-150413202	rs62382331	3	2.28E-02	0.009202
ICAM1	19-10396065	rs5499	2	2.19E-02	0.005348
CD44	11-35226083	rs11607491	4	2.19E-02	0.0133
RILP	17-1552519	-	3	2.18E-02	0.01974
KIAA1841	2-61293116	-	3	1.79E-02	0.01623
CRB2	9-126135490	-	3	1.67E-02	0.02551

Gene	Chr-BP	rsID	Degree	SNPrank Score	MAF
FHAD1	1-15573787	-	4	1.55E-02	0.0122
EPB49	8-21924297	rs9657287	2	1.53E-02	0.02239
PHRF1	11-582056	rs117660271	3	1.52E-02	0.01064
ITGAM	16-31343304	-	1	1.45E-02	0.005348
TNFAIP3	6-138203536	rs5029967	3	1.45E-02	0.01064