Whole transcriptome framework to identify human genetic variants relevant to health and disease

Michał Seweryn[1,2*], Samuel K. Handleman[3], Katherine Hartmann[3], Grzegorz Rempala[2] and Wolfgang Sadee[3];
[1]Mathematical Biosciences Institute, The Ohio State University, [2]Division of Biostatistics, College of Public Health, The Ohio State University, [3]Department of Pharmacology, The Ohio State University Wexner Medical Center, [*]Corresponding Author

Identifying genetic variants

Genome wide association studies provide a means of identifying genetic variants that may contribute to the heritability of a trait by searching for a statistical association between variants and the trait of interest. GWA studies are free from bias introduced by the current understanding of disease pathophysiology and accordingly can identify novel genes and variants that would otherwise remain unexplored (Manolio 2009).

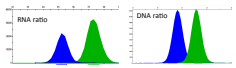
The evidence for a causal variant from GWAS results is strengthened if the variant is in a hub gene in multiple shared pathways. Therefore, we have developed methods to search for gene networks and hub-genes as a means to prioritize variants of likely significance. Our methods utilize deep sequencing transcriptome data (RNAseq) from public resources and from our own RNAseq data. These data include expression levels of coding and non-coding RNAs and their isoforms from multiple human tissues: brain, heart, liver, and lymphocytes.

We have developed a novel approach to gene network analysis that recovers properties of the broader network based on statistically significant pairwise interactions and to prioritizes hub genes. Local subnetworks are identified using topological and information-theoretic procedures. Algebraic methods are used for virtual gene knockouts in local networks using to further refine variant priorities. Network robustness is crucial in identifying hubs and and local networks useful for prioritization; epistasis and evolutionary constraints will serve to further refine high priority variants.

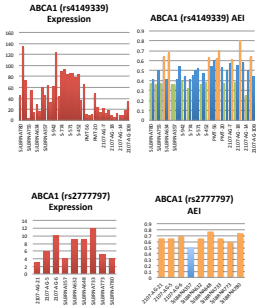
Definition

Let there be given a gene interaction network $\mathcal{N}$. We shall say that a gene g is a hub if both its adjacency index and the average strength of interaction is high.
By a module we understand a clique or a semiclique.

Identifying genetic variants using allelic expression imbalance (AEI)



Alleles can be distinguished from one another when heterozygous at a single nucleotide polymorphism (SNP). Genomic DNA exists in an equal, 1:1 ratio and in the absence of regulatory differences between alleles, RNA does as well. Differing amounts of each allele in mRNA signifies the presence of a cis-acting regulatory variant, RNA editing, loss of heterozygosity, copy number variation, or allele specific epigenetic silencing (Smith 2013).



Transcriptomes

Next generation sequencing techniques allow for sequencing of RNA transcriptomes on a large scale. RNA isolated from tissues is used to generate cDNA libraries that are sequenced and aligned to the human reference genome. RNA sequencing datasets report RNA expression values (measured in transcript counts and reported as reads per kilobase million mapped reads or RPKM to account for the effect of gene size and sample volume). These data sets also provide information about genetic variants (SNPs).

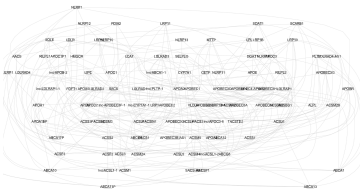
- RNA expression (RPKM)
- SNP/variant calls.

Table 1. Available RNAseq datasets.
The OSU RNAseq runs continue to accumulate.

Transcriptome Source	Number of Samples					
	Brain**	Liver	Adipose	Heart	Vascular	PBLs, LCLs
OSU – RNAseq*	150	12	---	---	4	20
PGRN RNAseq	---	24	25	25	---	80**
GTEx – RNAseq	>300	6	111	95	119	176

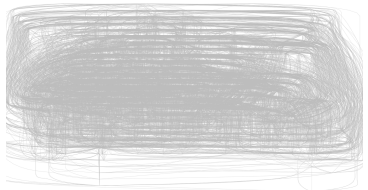
* cDNA library preparation with poly-dT and random hexamer primers; ** multiple brain regions
**40 antihypertensives highly responsive to hydrochlorothiazide and 40 non-responsive patients (collaboration with J. Johnson (PGRN)

Protein Coding RNA Hairball



Hubs	Pairs	Modules
APOC3, LDLRAD2, ABCA1	SOAT1 x APOC3, SCARB1 x AACS, PLTP x TACSTD2	SQLE, FDFT1, HMGCR

Non-coding RNA Hairball



Hubs	Pairs	Modules
U47924.28, Lnc.ADHFE1.1, LDLR	SOAT1 x LINC00843, SCARB1 x AACS, PLTP x AC013268.5	SQLE, FDFT1, lnc.KIAA1462.1, ACSL3, CTD. 3025N20.3, HMGCR, lnc.HPS3.2, lnc.ABCA8.1

Network building framework

Renyi entropy $H_\alpha(Q) := \frac{1}{1-\alpha}\log\left(\sum_i q_i^\alpha\right) \quad \alpha \in [0,\infty)$

Shannon entropy $H_1(Q) = \sum_i q_i \log\left(\frac{1}{q_i}\right)$

Renyi divergence $F_\alpha(P,Q) = \frac{1}{\alpha-1}\log\left(\sum_i \frac{p_i^\alpha}{q_i^{\alpha-1}}\right)$

Kullback-Leibler divergence $F_1(P,Q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$

Lemma

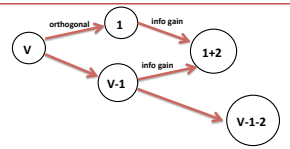
Renyi divergence satisfies the data processing inequality

$$F_\alpha(P_{|\mathcal{H}}, Q_{|\mathcal{H}}) \le F_\alpha(P,Q)$$

for any σ-subalgebra $\mathcal{H}$, where $P_{|\mathcal{H}}$ and $Q_{|\mathcal{H}}$ denote the restrictions of P and Q to $\mathcal{H}$.

Algorithm

Let $\mathcal{G} = \{g_1, \ldots, g_m\}$ be a set of genes and $\{\mathcal{X}^i = (X^i_{g_1}, \ldots, X^i_{g_m}),\ i = 1,2,\ldots n\}$ be the expression level of these genes in n individuals. We build an interaction network based on pairwise Renyi divergences using the bagging procedure to select the parameter α, which gives the most robust network (over bootstrap replicates). As a next step we use a likelihood approach (assuming Gaussian structure) to validate the significance of interactions. As a last step we use the Data Processing Inequality to prune the edges (as in ARACNE (Margolin, 2006)).



Algorithm

Let $C = [c_{ij}]$ be an array of variants (participants×variants). We aim to find linkage in the haplotype. To this aim we shall use the I-index defined as

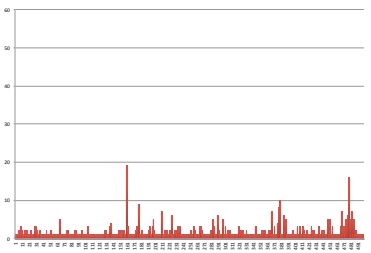
$$I_\alpha([c_{ij}]) = 1 - [F_\alpha([c_{ij}], [p_{i,\bullet} * p_{\bullet,j}]) / H_{2-\alpha}([p_{i,\bullet}])]\,.$$

And proceed as follows:

- *Take an initial set of variants*
- *remove a variant (1) 'most orthogonal' to the rest*
- *remove a variant (2) from the initial set which is BOTH: 'most orthogonal' to the initial set and 'most parallel' to (1)* – Information Gain
- *repeat step 3 until information gain is negative*
- *go to next level of hierarchy*

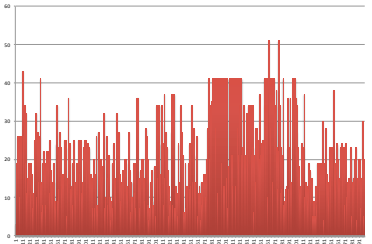
Haplotypes

CETP (56995835...57017757) rs5883, rs247616
Divisive Clustering



Haplotypes

CETP (56995835...57017757) rs5883, rs247616
Pairwise Comparisons



Acknowledgements

This research was supported by NIGMS U01 GM09265 and NCI R01CA-152158.

References

Manolio, Teri A, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." Nature 461 (7265) (October): 747–753. doi:10.1038/nature08494.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics. 2006;7 Suppl 1:S7. PMCID: 1810318.

Smith, Ryan M, Amy Webb, Audrey C Papp, Leslie C Newman, Samuel K Handelman, Adam Suhy, Roshan Mascarenhas, John Oberdick, and Wolfgang Sadee. 2013. "Whole Transcriptome RNA-Seq Allelic Expression in Human Brain." BMC Genomics 14: 571. doi:10.1186/1471-2164-14-571.