USING NETWORKS TO UNDERSTAND THE GENOTYPE-PHENOTYPE CONNECTION

John Quackenbush
Dana-Farber Cancer Institute
Harvard TH Chan School of Public Health
Essentially, all models are wrong, but some are useful.

– George E. Box
No Free Lunch Theorem

David H. Wolpert and William G. Macready

Abstract—A framework is developed to explore the connection between effective optimization algorithms and the problems they are solving. A number of “no free lunch” (NFL) theorems are presented which establish that for any algorithm, any elevated performance over one class of problems is offset by performance over another class of problems. These theorems result in a geometric interpretation of what is suitable to an optimization algorithm. Benchmark measures of performance quantify the performance of algorithms using a priori “head-to-head” comparisons against known optimization algorithms, distilling the empirical performance of an algorithm down into a type of mathematical theorem.

Index Terms—Evolutionary optimization.

I. INTRODUCTION

The past few decades have seen an increased interest in general-purpose “black-box” optimization algorithms that exploit limited knowledge concerning the optimization problem on which they are run. In large part these algorithms have drawn inspiration from optimization processes that occur in nature, such as evolutionary algorithms [1]–[3] and others.

VIII. CONCLUSIONS

A framework has been presented in which to compare general-purpose optimization algorithms. A number of NFL theorems were derived that demonstrate the danger of comparing algorithms by their performance on a small sample of problems. These same results also indicate the importance of incorporating problem-specific knowledge into the behavior of the algorithm. A geometric interpretation was given showing what it means for an algorithm to be well suited to solving a certain class of problems. The geometric perspective also suggests a number of measures to compare the similarity of various optimization algorithms.
What is a network?

- A network is a way of representing complex association patterns between various entities.

- Networks can be abstract or represent physical/regulatory interactions.
How we do Network Analysis

Genes (Phenotype I)

Expression data

Conditions

Infer phenotype-specific network

Analyze Network Topology and Structure

Simultaneously compare differential structure and expression

Genes (Phenotype II)

Expression data

Conditions

Infer phenotype-specific network
Starting Assumptions

• There is no single “right” network

• The structure of the network matters and network structure often changes between states.

• We have to move from asking “Is the network right?” to asking “Is the network useful?”

• The real question is “Does a network model inform our understanding of biology?”
The Methodological Zoo

Preliminary Data & GRN construction

Public Data
- Gene Expression
- Genetic Profiles
- Proteomics?
- mRNA
- eQTLs

Data Integration & Network Modelling
- PPI
- TF-gene
- miRNA-gene

Public Data Resources
- StringDB
- FIMO/motif-scan
- TargetScan

Algorithm Key
- PANDA – Passing Attributes between Networks for Data Assimilation
- SPIDER – Seeding PANDA Iterations Determining Epigenetic Regulation
- PUMA – PANDA Using Micro-RNA Associations
- CONDOR – Complex Network Description Of Regulators
- LIONESS – Linear Interpolation to Obtain Network Estimates for Single Samples
- ALPACA – ALtered Partitions Across Community Architectures
- MONSTER – MOdeling Network State Transitions from Expression and Regulatory data

Input Data
- Network Algorithm
- Algorithm Output
- Hypothesis Test

Comparative Network Analysis
- edgeQTLs (NEW)
- Edge Communities (set of edges perturbed by SNPs)
- GRN
- GRN
- GRN
- GRN
- GRN

Legend
- Differential Node Communities
- Differential TFs

Node Communities & Core-scores
- prioritize SNPs
- single-sample networks

Conclusions
- Comparative network
- model covariates
- integrate variants

Comparative Network Analysis
- Differential Node Communities
- Differential TFs

Comparative Network Analysis
- edgeQTLs (NEW)
Question 1: Can we solve the "GWAS Puzzle"?
Genome Wide Association Studies (GWAS)

DNA from different individuals sequenced

Variation at a single nucleotide

Some individuals will have one version of the SNP, some the other

Sample with disease

A higher than expected incidence in a disease group suggests SNP1G is associated with a disease (or SNP1A is protective)

Normal population

In a population, a certain percentage will have one version, the rest the other

© Gibson & Muse, A Primer of Genome Science
Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

697 SNPs explain 20% of height
~2,000 SNPs explain 21% of height
~3,700 SNPs explain 24% of height
~9,500 SNPs explain 29% of height
Genetic studies of body mass index yield new insights for obesity biology

Obesity is heritable and predisposes to many diseases. To understand the genetic basis of obesity better, here we conduct a genome–wide association study and MetaChips meta–analysis of body mass index (BMI), a measure commonly used to define obesity and assess adiposity, in up to 339,224 individuals. This analysis identifies 97 BMI–associated loci (p < 5 × 10⁻⁸), 56 of which are novel. Five loci demonstrate clear evidence of several independent association signals, and many loci have significant effects on other metabolic phenotypes. The 97 loci account for ~2.7% of BMI variation, and genome–wide estimates suggest that common variation accounts for >20% of BMI variation. Pathway analyses provide strong support for a role of the central nervous system in obesity susceptibility and implicate new genes and pathways, including those related to synaptic function, glutamate signalling, insulin secretion/action, energy metabolism, lipid biology and adipogenesis.

97 SNPs explain 2.7% of BMI
All common SNPs may explain 20% of BMI

Do we give up on GWAS, fine map everything, or think differently?
The genetic architecture of type 2 diabetes

...large-scale sequencing does not support the idea that lower-frequency variants have a major role in predisposition to type 2 diabetes.
eQTL Analysis

Expression Quantitative Trait Locus Analysis (eQTL Analysis) uses genome-wide data on genetic variants (SNPs) together with gene expression data.

Treat gene expression as a quantitative trait.

Ask, “Which SNPs are correlated with the degree of gene expression?”

Most people concentrate on cis-acting SNPs.

What about trans-acting SNPs?
cis-eQTL Analysis

Population Sample

Expression

CC  CG  GG
trans-eQTL Analysis

Population Sample

Expression

CC  CG  GG
**eQTL Networks: A simple idea**

- Perform a “standard eQTL” analysis:
  \[ Y = \beta_0 + \beta_1 \text{ADD} + \varepsilon \]
  where \( Y \) is the quantitative trait and \( \text{ADD} \) is the allele dosage of a genotype.

Representing eQTLs as a network and analyzing its structure should provide insight in the complex interactions that drive disease.
Many strong eQTLs are found near the target gene. But what about multiple SNPs that are correlated with multiple genes?

Can a network of SNP-gene associations (cis and trans) inform the functional roles of these SNPs?
The Result: A Hairball

Some random hairball I grabbed. I was too lazy to make one.
Results: COPD

~30,000 SNPs and ~3,400 Genes

Degree – number of links per node
What about GWAS SNPs?

The “hubs” are a GWAS desert!
Can we use this network to identify groups of SNPs and genes that play functional roles in the cell?

Try clustering the nodes into “communities” based on the network structure.
Communities are groups of highly intra-connected nodes

- Community structure algorithms group nodes such that the number of links within a community is higher than expected by chance.
- Formally, they assign nodes to communities such that the modularity, $Q$, is optimized.

$$Q = \sum_i (e_{ii} - a_i^2)$$

- Fraction of network links in community $i$
- Fraction of links expected by chance

Newman 2006 (PNAS)
Communities in COPD eQTL networks
Communities in COPD eQTL networks

• We identified 52 communities, with Q = 0.79 (out of 1)
• Of 34 communities in the giant connected component, 11 are enriched for genes with coherent functions (GO Terms; P<5x10^{-4})
Communities in COPD eQTL networks

- Chromatin Assembly
- DNA conformation change
- Nucleosome assembly

- Immune response
- Stress response
- T cell stimulation

- Microtubule organization
- Cell cycle
- Centrosome
Identifying community cores

- Score each SNP by its contribution to the modularity of its community
- Do these “core scores” reflect known biology?

\[ Q_{ih} = \frac{Q_i}{Q_h} \]

Newman 2006 (PNAS)
What about COPD GWAS SNPs?

- Use a meta-analysis by Cho et. al. and consider 34 COPD GWAS SNPs (FDR < 0.05)

Core Scores for COPD GWAS SNPs

The median core score for the 34 FDR-significant GWAS SNPs is 20.3 times higher than the median for non-significant SNPs.
How general is this?
Bipartite Community Structure of eQTLs

John Platig, Peter J. Castaldi, Dawn DeMeo, John Quackenbush

1 Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 2 Department of Bioinformatics, Harvard Chan School of Public Health, Boston, Massachusetts, United States of America, 3 Channing Division of Network Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, 4 Division of General Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America, 5 Harvard Medical School, Boston, Massachusetts, United States of America, 6 Division of Pulmonary and Critical Care Medicine, Brigham and Women’s Hospital, Boston, Massachusetts, United States of America

*platig@hsph.harvard.edu

Abstract

Genome Wide Association Studies (GWAS) and expression quantitative trait locus (eQTL) analyses have identified genetic associations with a wide range of human phenotypes. However, many of these variants have weak effects and understanding their combined effect remains a challenge. One hypothesis is that multiple SNPs interact in complex networks to influence functional processes that ultimately lead to complex phenotypes, including disease states. Here we present CONDOR, a method that represents both cis- and trans-acting SNPs and the genes with which they are associated as a bipartite graph and then uses the modular structure of that graph to place SNPs into a functional context. In
Exploring regulation in tissues with eQTL networks

Maud Fagny\textsuperscript{a,b}, Joseph N. Paulson\textsuperscript{b,b}, Marieke L. Kuijjer\textsuperscript{a,b}, Abhijeet R. Sonawane\textsuperscript{c}, Cho-Yi Chen\textsuperscript{a,b}, Camila M. Lopes-Ramos\textsuperscript{a,b}, Kimberly Glass\textsuperscript{d}, John Quackenbush\textsuperscript{b,b,d,1}, and John Platig\textsuperscript{b,b,1}

*Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115; \textsuperscript{b}Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA 02115; \textsuperscript{c}Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School Boston, MA 02115; and \textsuperscript{d}Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115

Edited by Jasper Rine, University of California, Berkeley, CA, and approved August 4, 2017 (received for review May 3, 2017)

Characterizing the collective regulatory impact of genetic variants on complex phenotypes is a major challenge in developing a genotype to phenotype map. Using expression quantitative trait locus (eQTL) analyses, we constructed bipartite networks in which edges represent significant associations between genetic variants and gene expression levels and found that the network structure informs regulatory function. We show, in 13 tissues, that these eQTL networks are organized into dense, highly modular communities grouping genes often involved in coherent biological processes. We find communities representing shared processes across tissues, as well as communities associated with tissue-specific processes that coalesce around variants in tissue-specific active chromatin regions. Node centrality is also highly informative, with the global and community hubs differing in regulatory potential and likelihood of being disease associated.

GTEX, expression quantitative trait locus, eQTL, bipartite networks, GWAS

biological pathways. In particular, we find three aspects of the eQTL network topology that inform tissue-level regulatory biology: (i) Communities—which are composed of SNPs and genes with a high density of within-group edges—are enriched for pathways, functionally related genes, and SNPs in tissue-specific active chromatin regions (actively transcribed and open regulatory regions); (ii) community hubs (core SNPs)—which are SNPs highly connected to genes in their community—are enriched for active chromatin regions close to the transcriptional start site and for GWAS association; and (iii) global hubs—which are connected to many genes throughout the network—are enriched for distal elements such as noncoding enhancers and devoid of GWAS association. The picture that emerges from analysis of the eQTL networks is a complex web of associations that reflects the polygenic architecture across tissues and that provides a natural framework for understanding both the shared and tissuespecific effects of genetic variants. These networks, along with SNP and gene network properties across all 13 tissues, are avail-

Now in thirteen tissues
We need a big sandbox

HUMAN GENOMICS

The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans

The GTEx Consortium*†

Understanding the functional consequences of genetic variation, and how it affects complex human disease and quantitative traits, remains a critical challenge for biomedicine. We present an analysis of RNA sequencing data from 1,641 samples across 43 tissues from 175 individuals, generated as part of the pilot phase of the Genotype-Tissue Expression (GTEx) project. We describe the landscape of gene expression across tissues, catalog thousands of tissue-specific and shared regulatory expression quantitative trait loci (eQTL) variants, describe complex network relationships, and identify signals from genome-wide association studies explained by eQTLs. These findings provide a systematic understanding of the cellular and biological consequences of human genetic variation and of the heterogeneity of such effects among a diverse set of human tissues.

Over the past decade, there has been a marked increase in our understanding of the role for any given GWAS locus or disease. Hence, understanding the role of regulatory variants, statistical power, we prioritized RNA sequencing of samples from nine tissues that were most frequently collected and that routinely met minimum RNA quality criteria: adipose (subcutaneous), tibial artery, heart (left ventricle), lung, muscle (skeletal), tibial nerve, skin (Sun-exposed), thyroid, and whole blood (Table 1) (‡).

We performed 76-base pair (bp) paired-end mRNA sequencing on a total of 1,749 samples, of which 1,641 samples from 43 sites, and 175 donors, constituted our final "pilot data freeze" reported on here (‡). Median sequencing depth was 82.1 million mapped reads per sample (fig. S3A). The final data freeze included samples from 43 body sites: 29 solid-organ tissues, 11 brain subregions (with two duplicated regions), a whole-blood sample, and two cell lines derived from donor blood [EBV-transformed lymphoblastoid cell lines (LCLs)] and skin samples (cultured fibroblasts) (Table 1 and tables S1 and S2). Median sample size for the nine high-priority tissues was 105; median sample size for the other 34 sampled sites was 18.5.

Gene expression across tissues

We examined the patterns of expression of 53,934 transcribed genes across tissues [on the basis of Gencode V12 annotations] (‡, †). The number of biotypes (protein-coding genes, pseudogenes, and long noncoding RNAs [lncRNAs]) that were transcribed above a minimal threshold (reads per kilo-
GTEx eQTL Workflow

12 tissues, 450 individuals, 4904 samples
Imputed Genotypes (5,640,985 SNPs)
RNAseq data (29,242 genes)

Data Preprocessing

Genotype filtering (plink)
MAF ≥ 0.5 & SNP calling > 0.9
RNAseq data QC & normalization (qsmooth)

eQTL mapping

cis- and trans-eQTLs mapping (MatrixEQTl)

Characterization

Genomic localization of eQTLs
Enrichment in biological functions, pathways, association with diseases

Communities detection
Find community cores
Bipartite Modularity Maximization (condor)

Bipartite Networks

SNPs
Genes

Characterization

Genomic localization of eQTLs
Enrichment in biological functions, pathways, association with diseases

Communities detection
Find community cores
Bipartite Modularity Maximization (condor)

Bipartite Networks

SNPs
Genes

GTEx eQTL Workflow

12 tissues, 450 individuals, 4904 samples
Imputed Genotypes (5,640,985 SNPs)
RNAseq data (29,242 genes)

Data Preprocessing

Genotype filtering (plink)
MAF ≥ 0.5 & SNP calling > 0.9
RNAseq data QC & normalization (qsmooth)

eQTL mapping

cis- and trans-eQTLs mapping (MatrixEQTl)

Characterization

Genomic localization of eQTLs
Enrichment in biological functions, pathways, association with diseases

Communities detection
Find community cores
Bipartite Modularity Maximization (condor)

Bipartite Networks

SNPs
Genes
eQTL networks are highly modular
GWAS SNPs are cores, but not hubs
GWAS SNPs not Hubs—in every tissue
Core SNPs are more likely to be functionally annotated

<table>
<thead>
<tr>
<th>Category</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>TF binding + matched TF motif + matched DNase footprint + DNase peak</td>
</tr>
<tr>
<td>B</td>
<td>TF binding + any motif + DNase footprint + DNase peak</td>
</tr>
<tr>
<td>C</td>
<td>TF binding + matched TF motif + DNase peak</td>
</tr>
<tr>
<td>D</td>
<td>TF binding + any motif + DNase peak</td>
</tr>
<tr>
<td>E</td>
<td>TF binding + matched TF motif</td>
</tr>
<tr>
<td>F</td>
<td>TF binding + DNase peak</td>
</tr>
<tr>
<td>G</td>
<td>TF binding or DNase peak</td>
</tr>
<tr>
<td>H</td>
<td>Motif hit</td>
</tr>
<tr>
<td>I</td>
<td>No Information</td>
</tr>
</tbody>
</table>

![Brain annotation distribution graph](chart.png)
Core SNPs are different from Hubs: Roadmap Epigenomics Project

Data from 8 tissues
Core SNPs map to open chromatin

- We find tissue-specific eQTLs map to tissue-specific communities
- And those tissue-specific communities are enriched for SNPs in tissue-specific open chromatin
What does this tell us?

• The SNPs that are global hubs are not GWAS hits—meaning that they are not linked to diseases or traits.

• The SNPs and genes group into communities that share function—a family of SNPs regulate a function.

• Disease-associated (GWAS) SNPs map to communities whose genes have functions that make biological sense.

• “Core” SNPs are far more likely to be disease SNPs.

• Tissue-specific functions are in tissue-specific communities with tissue-specific genes organized around SNPs in tissue-specific open chromatin.
Question 2: Can we model gene regulatory processes?
Integrative Network Inference: PANDA

Passing Messages between Biological Networks to Refine Predicted Interactions

Kimberly Glass¹,², Curtis Huttenhower², John Quackenbush¹,², Guo-Cheng Yuan¹,²

¹ Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America. ² Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, United States of America

Abstract

Regulatory network reconstruction is a fundamental problem in computational biology. There are significant limitations to such reconstruction using individual datasets, and increasingly people attempt to construct networks using multiple, independent datasets obtained from complementary sources, but methods for this integration are lacking. We developed PANDA (Passing Attributes between Networks for Data Assimilation), a message-passing model using multiple sources of information to predict regulatory relationships, and used it to integrate protein-protein interaction, gene expression, and sequence motif data to reconstruct genome-wide, condition-specific regulatory networks in yeast as a model. The resulting networks were not only more accurate than those produced using individual data sets and other existing methods, but they also captured information regarding specific biological mechanisms and pathways that were missed using other methodologies. PANDA is scalable to higher eukaryotes, applicable to specific tissue or cell type data and conceptually generalizable to include a variety of regulatory, interaction, expression, and other genome-scale data. An implementation of the PANDA algorithm is available at www.sourceforge.net/projects/panda-net.
Regulation of Transcription

regulatory sequences

promoter

Specific transcription factors
A Simple Idea: Message Passing

Transcription Factor
The TF is Responsible for communicating with its Target

$$R_{ij}^{(t)} = \sum_k A_{ik}^{(t-1)} C_{kj} A_{ik}^{(0)} / \sum_k C_{kj} A_{ik}^{(0)}$$

Downstream Target
The Target must be Available to respond to the TF

$$A_{ij}^{(t)} = \sum_k R_{ik}^{(t-1)} C_{kj} R_{ik}^{(0)} / \sum_k R_{ik}^{(0)}$$

Kimberly Glass, GC Yuan
Code and related material available on sourceforge: http://sourceforge.net/projects/panda-net/
Subtypes of Ovarian Cancer

Angiogenic mRNA and microRNA Gene Expression Signature Predicts a Novel Subtype of Serous Ovarian Cancer

Stefan Bentink¹,6,9, Benjamin Haibe-Kains¹,6,9, Thomas Risch¹, Jian-Bing Fan³, Michelle S. Hirsch⁴,7, Kristina Holton¹, Renee Rubio¹, Craig April³, Jing Chen³, Eliza Wickham-Garcia³, Joyce Liu²,7, Aedin Culhane¹,6, Ronny Drapkin⁴,5,7, John Quackenbush¹,2,6,9*, Ursula A. Matulonis⁵,7*

1 Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 2 Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 3 Illumina, Inc., San Diego, California, United States of America, 4 Department of Pathology, Division of Women's and Perinatal Pathology, Brigham and Women's Hospital, Boston, Massachusetts, United States of America, 5 Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, United States of America, 6 Harvard School of Public Health, Boston, Massachusetts, United States of America, 7 Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Ovarian cancer is the fifth leading cause of cancer death for women in the U.S. and the seventh most fatal worldwide. Although ovarian cancer is notable for its initial sensitivity to platinum-based therapies, the vast majority of patients eventually develop recurrent cancer and succumb to increasingly platinum-resistant disease. Modern, targeted cancer drugs intervene in cell signaling, and identifying key disease mechanisms and pathways would greatly advance our treatment abilities. In order to shed light on the molecular diversity of ovarian cancer, we performed comprehensive transcriptional profiling on 129 advanced-stage, high-grade serous ovarian cancers. We implemented a re-sampling-based version of the...
PANDA: Integrative Network Models

Comparison/Identify Differences

Network for Angiogenic Subtype

Network for Non-angiogenic Subtype

Expression data (Angiogenic)

Expression data (Non-angiogenic)

Motif Data

Interaction Data

Kimberly Glass, GC Yuan
Inner ring: key TFs
Colored by Edge
Enrichment (A or N)

Outer ring: genes
Colored by Differential
Expression (A or N)

Interring Connections
Colored by
Subnetwork (A or N)

Ticks – genes
annotated to
“angiogenesis” in GO,
Complex Regulatory Patterns Emerge

<table>
<thead>
<tr>
<th>TF1</th>
<th>TF2</th>
<th>sig.</th>
<th>#</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARID3A</td>
<td>PRRX2</td>
<td>1.16E-23</td>
<td>244</td>
<td>A+</td>
</tr>
<tr>
<td>ARID3A</td>
<td>SOX5</td>
<td>1.01E-14</td>
<td>155</td>
<td>A+</td>
</tr>
<tr>
<td>PRRX2</td>
<td>SOX5</td>
<td>3.83E-12</td>
<td>157</td>
<td>A+</td>
</tr>
<tr>
<td>ARNT</td>
<td>MZF1</td>
<td>5.83E-23</td>
<td>92</td>
<td>N-</td>
</tr>
<tr>
<td>AHR</td>
<td>ARNT</td>
<td>6.13E-16</td>
<td>382</td>
<td>N-</td>
</tr>
<tr>
<td>ETS1</td>
<td>MZF1</td>
<td>9.08E-16</td>
<td>148</td>
<td>N-</td>
</tr>
</tbody>
</table>

Kimberly Glass, GC Yuan
Regulatory Patterns suggest Therapies

**ANGIOGENIC BEHAVIOR**

HIF1a → ARNT

HIF2a → ETS1 → VEGF production and angiogenesis

**TREATMENT MODEL**

1. Prevent ARNT/HIF1a and ETS1/HIF2a dimerization

2. Promote ARNT/AHR and ETS1/AHR dimerization

3. Decrease genome-wide methylation

Kimberly Glass, GC Yuan
More application papers coming….
Understanding Tissue-Specific Gene Regulation

Abhijeet Rajendra Sonawane,1,2 John Platig,3,4 Maud Fagny,3,4 Cho-Yi Chen,3,4 Joseph Nathaniel Paus disgusted by the similarity.
Camilo Miranda Lopes-Ramos,3,4 Dawn Lisa DeMeo,1,2,3,5 John Quackenbush,1,2,3,4,6 Kimberly Glass,1,2,7,8 and Marieke Lydia Kuiper9,8,*

1Channing Division of Network Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA
2Department of Medicine, Harvard Medical School, Boston, MA 02115, USA
3Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
4Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
5Division of Pulmonary and Critical Care Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA
6Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
7These authors contributed equally
8Lead Contact
*Correspondence: kimberly.glass@channing.harvard.edu (K.G.), mkuiper@jimmy.harvard.edu (M.L.K.)
https://doi.org/10.1016/j.celrep.2017.10.001

SUMMARY

Although all human tissues carry out common processes, tissues are distinguished by gene expression patterns, implying that distinct regulatory programs control tissue specificity. In this study, we investigate gene expression and regulation across 38 tissues profiled in the Genotype-Tissue Expression project. We find that network edges (transcription factor to target gene connections) have higher biological function requires the combinatorial action of multiple regulatory elements, primarily transcription factors that work together with other genetic and environmental cues to mediate the transcription of genes and their products (Vagkarizas et al., 2009).

Gene regulatory network modeling provides a framework that can summarize the complex interactions among transcription factors, genes, and gene products (Ottaviani, 2004; Gerstung et al., 2015). Despite the complexity of the regulatory process, the most widely used methods are based on pairwise gene co-expression.

Highlights

- Regulatory network connections are more tissue specific than nodes (genes and transcription factors)

More application papers coming....
Question 3: Can we move beyond THE Network?
We generally estimate “Aggregate” Networks.
Gene Expression Data

Genes

Gene Expression Data

Multiple Samples

Network Reconstruction

Pearson Correlation, Mutual Information, CLR, PANDA, etc.

Linear Interpolation to Obtain Network Estimates for Single Samples (LIONESS)

One Network

Represents information from all the input samples
Single-Sample Networks (LIONESS)

Network $e^{(\alpha)}$
Representing contributions from **all samples**

Network $e^{(\alpha-q)}$
Representing contributions from **all samples except q**

$N_s \left( e^{(\alpha)} - e^{(\alpha-q)} \right) + e^{(\alpha-q)} = e^{(q)}$

Sample q’s contribution to $e^{(\alpha)}$

Sample q’s network

Network estimated without sample q

Scale factor
A Quick Test Using Yeast Cell Cycle Data

- Data includes 48 total expression arrays taken over a time-course (every 5 min) on synchronized yeast cells (~2 cell cycles)
- Includes technical replicates (Cy3/Cy5 and Cy5/Cy3)
- Estimate single-sample networks using data for each replicate.

Replicate 1 (24 Samples → 24 Networks)

Replicate 2 (24 Samples → 24 Networks)
• Single-sample networks for TCGA glioblastoma patients (n=525)
• Defined “good” (>1.7 years; 127) and “poor” (<1.7 years; 339) prognosis groups; 59 excluded
• Used PANDA+LIONESS to get individual networks
• LIMMA analysis using network “edge weights”
• Found 148 significant TF-gene edges
• A strong edge can indicate induction/repression
Glioblastoma network signatures

- Look at the targets of differential edges.
- Do a GSEA on the target genes.
- Immune response appears to play a big role in good prognosis.
- Data suggests repressed PD1 signaling improves survival.
The future is here.
It's just not widely distributed yet.

- William Gibson
Before I came here I was confused about this subject.
After listening to your lecture, I am still confused but at a higher level.

- Enrico Fermi, (1901-1954)
Acknowledgments

Gene Expression Team
Fieda Abderazzaq
Aedin Culhane
Jessica Mar
Renee Rubio

DFCI Radiology
Hugo Aerts

University of Queensland
Christine Wells
Lizzy Mason

CDNM, Brigham and Women's Hospital
Peter Castaldi
Michael Cho
Dawn DeMeo
Kimberly Glass
Ed Silverman
Xiaobo Zhou

Center for Cancer Computational Biology
Fieda Abderazzaq
Stas Alekseev
Nicole Flanagan
Ed Harms
Lev Kuznetsov
Brian Lawney
Antony Partensky
John Quackenbush
Renee Rubio
Yaoyu E. Wang

Students and Postdocs
Joseph Barry
Joey (Cho-Yi) Chen
Marieke Kuijjer
Camila Lopes-Ramos
Zachary McCaw
Megha Padi
Joseph Paulson
John Platig
Daniel Schlauch
Daphne Tsoucas

Alumni
Martin Aryee
Stefan Bentink
Kimberly Glass
Benjamin Haibe-Kains
Kaveh Maghsoudi
Jess Mar
Melissa Merritt
Alejandro Quiroiz
J. Fah Sathirapongsasuti

http://compbio.dfci.harvard.edu

CDNM, Brigham and Women's Hospital
Peter Castaldi
Michael Cho
Dawn DeMeo
Kimberly Glass
Ed Silverman
Xiaobo Zhou

National Institute of Allergy and Infectious Diseases
National Library of Medicine
National Cancer Institute
National Heart, Lung, and Blood Institute
National Institute of Mental Health
Geoffrey Beene Foundation
Alzheimer's Initiative
NVIDIA