The Future of Data Science and Bioinformatics-Anchored Advances in Human Health

Yves A. Lussier, M.D.
Professor of Medicine
Yves @ email.Arizona.edu

Assoc. Vice-President for Health Sciences & Chief Knowledge Officer
Director for Precision Health and Cancer Informatics, (Cancer Ctr)
Associate Director, BIO5 Institute,
Director, Center for Biomedical Informatics and Biostatistics

Lead Investigator, Beagle Supercomputing, CI, Un of Chicago & Argonne Lab

Fellow, IGSB, The University of Chicago & Argonne Laboratory

orcid.org/0000-0001-9854-1005
Scopus Author ID: 7004101223
ResearcherID: N-4891-2017
Disclosures: overlapping Interests

Intellectual property

Boards at for profit orgs (inactive)

Boards non-profit orgs

Institutions with former research trainees
• Personal dynamic transcriptomes can be interpreted affordably
  – challenged *in vitro* to unveil personal genomic response to environment

• Personal intergenic and non-coding genetic interactions are druggable
  – as they determine one’s human transcriptome
Plan

• Precision Medicine Initiative (Bipartisan / White House)
• Recent developments
  – Dynamics of personal transcriptome & interpretation (G×E)
  – Mechanisms of non coding and intergenic SNPs unveiled through genetic interactions (G×G)
• Route to application
Your Personal Nutrition Test Results

Biochemistry

Genetics

Transcriptome (mRNA)
Problem: unproductive assumptions for discovery of transcriptome biomarkers in common diseases

- 30,000 NIH “biomarker” grants in 25 yrs (> $2.5 billion/year) \(^1\)
  - unproductive: only 12 FDA-approved cancer biomarkers (2012-2017)
  - limited success in clinical practice

- Conventional transcriptome biomarker discovery designed for an average patient:
  - single biomolecule assumed concordantly altered across patients
  - patient-specific biomarker signal remains undetected

Insights from mice Genome by Environment interactions (G×E)

Problem: interpreting personal ‘omics

Only one observation in each condition?!

Statistician’s Nightmare

Reprinted from Funny Times / PO Box 18530 / Cleveland Hts. OH 44118
phone: 216.371.8600 / email: ft@funnytimes.com
Problem: drugs for non-coding genome?

• 3% of the genome is protein-coding
  – vast majority of drugs target proteins

• 97% of disease-associated polymorphisms
  – yet, “It's all druggable”

Nat Genet. 2017 Jan 31;49(2):169. doi: 10.1038/ng.3788
  • siRNA
  • CRISPR/CAS9

Nat Genet. 2017 Jan 31;49(2):169
LONG-TERM GOALS

Create a research cohort of > 1 million American volunteers who will share genetic data, biological samples, and diet/lifestyle information, all linked to their electronic health records if they choose.

Pioneer a new model for doing science that emphasizes engaged participants, responsible data sharing, and privacy protection.
White House (WH) Announces: University of Arizona's (UA) Lussier Group Involvement in National Precision Medicine Initiative

https://uanews.arizona.edu/story/white-house-announces-ua-s-involvement-in-national-precision-medicine-initiative

- UA Drs. Ojo & Lussier funded to recruit 150,000 subjects for DNA Sequencing
- UA/WH Lussier group is launching new precision medicine initiatives
  - System-wide dissemination of an on-demand "case-based reasoning" system that intelligently searches and analyzes entire databases of electronic medical records. This will give clinicians the power to develop an individualized and effective treatment plan for unusual or complex clinical conditions, grounded on practice-based evidence.
  - expand the clinical utility of its open-source, patient-centric analytic methods to aid physicians in interpreting the dynamic disease-associated gene expression changes (dynamic transcriptome) arising from patients’ own DNA blueprint
  - Development of genetic assays to predict an individual's response to therapy and prevention of adverse reactions, termed "pharmacogenomics".
Convergent pathway deregulation of diseases with distinct causes (genetic, epigenetic, environmental)
- coagulopathies can be inherited (Mendelian)
- or acquired (G×E)

In other words, convergent phenotypes attributable to distinct gene products dysregulation within the same pathway
Nonsense and missense mutations in hemophilia A*
Disruption of the Metabolic Pathways of *Inherited* Coagulopathies

genome/epigenome × environment regulation of "bleeding time" trait

Acquired (environmental) Coagulopathies: decrease expression of pathway genes in absence of mutations

Coagulopathy secondary to chronic liver disease
(e.g. cirrhosis of alcohol abuse)
Model for Network Medicine

**Gaps between networks of mechanisms at different scales**
Summative effect of diverse genetics = similar transcriptome

Pathway (KEGG)

Genetic (OMIM)

Disease (SNOMED)

Semantics (SNOMED)

Clinical system (SNOMED)

hsa04601 Complement and coagulation cascade

Factor VII deficiency (disorder) 37193007

Hereditary factor XI deficiency disease (disorder) 49762007

Afibrinogenemia (disorder) 278504009

281833003 Hematological system (body structure)
Background: single-subject transcriptome analyses of altered pathways

1. One transcriptome against a reference cohort
   • **Pathifier** (PNAS 2013;110:6388); **IndividPath** (Brief Bioinform 2016;17:78); **iPAS** (Bioinform2014;30:I422)
   • **Limitation:** requires an heterogenic cohort, may lead to false positives and false negatives due to heterogeneity and distinct environments

2. Two paired transcriptomes (e.g. before & after treatment, over time for a disease, control tissue vs affected tissue)
   • **N-of-1-pathways** methods (Lussier Group). **Wilcoxon:** J Am Med Inform Assoc 2014;21:1015; **Mahalanobis Distance:** Bioinformatics 2015;31:i293; **ClusterT:** Statistical Methods in Medical Research 2017; **MixEnrich:** BMC Medical Genomics 2017;10:27; **kMen:** J Biomed Inform 2017;66:32.
   • **Advantage:**
     o well controlled biologically
     o statistically powerful comparison to an isogenic baseline just a in studies of cell lines or isogenic animal models (e.g. mouse)

3. Multiple paired isogenic measures (>2) (e.g. time series replication, etc)
   • **Timevector** (Bioinformatics 2017)
   • **Advantage:** more powerful than two paired transcriptomes
   • **Limitation:** limited availability of clinical tissue
Transcriptome analysis

Cohort

Case

Control

Gene Expression

Common

Gene / pathway signature

Cohort

Case / Control
Paired samples

Gene Expression

Individual

Control / Case
Paired samples

Gene Expression

Common Pathway signature

Individual Pathway signature
Design the right tool

1. N-of-1 Individual
2. Paired Samples
3. Dynamic Gene Expression
4. Pathways Analysis

- Predicting or monitoring response to therapy
- Understanding personal disease mechanisms
II. N-of-1-\textit{pathways} Mahalanobis Distance (MD)


\begin{align*}
\text{Gene-level Differential Expression} = d_j &= \sqrt{\frac{S_N^2}{S_N^2 S_T^2 - (S_{NT})^2} (T_j - N_j)}
\end{align*}
III. N-of-1-\textit{pathways} k-means clustering and enrichment (kMen)


\begin{figure}
\centering
\includegraphics[width=\textwidth]{single-subject-analysis}
\caption{Single-Subject Analysis}
\end{figure}
Single-subject (SS) pathway-level studies ➔ emerging cross-subject pathway signal

- **Hypothesis:** pathway-level signal emerges from heterogeneous dysregulated genes in each patient (responsive genes), as they coordinate to alter a multi-gene function (e.g. pathway)

- **Pathway biomarker Framework:**
  - Identify responsive genes (red & blue below) and altered pathways in each subject (single-subject studies)
  - Followed by cross-subjects pathway-level statistics

**Figure:** Three SS studies. Same altered pathway in each patient, discoverable in each single subject study

Simulation parameters:
20% responsive genes
50% up-regulated genes

Legend
- Red = Up Regulated
- Blue = Down Regulated
Conventional cohort-based analyses (heterogenic)

Gene 1

<table>
<thead>
<tr>
<th>Normal</th>
<th>Tumor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Patie t 1 →</td>
<td></td>
</tr>
<tr>
<td>Patie t 2 →</td>
<td></td>
</tr>
<tr>
<td>Patie t N →</td>
<td></td>
</tr>
</tbody>
</table>

Gene 2

<table>
<thead>
<tr>
<th>Normal</th>
<th>Tumor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Gene M

<table>
<thead>
<tr>
<th>Normal</th>
<th>Tumor</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Paired t-test

Cohort Pvalues, effect size
Single-subject study (powerful isogenic conditions)

Gene 1

Gene 2

Gene M

Patient 1

Patient 2

Patient N

Single subject
Pvalues, effect size
(e.g., kMen)
Single-subject studies followed by metanalysis across studies

Gene 1  Gene 2  Gene M
\{ r \}  \{ r \}  \{ r \}

Patient 1 →

Patient 2 →

Patient N →

\[ \text{kME}_n \]

For each gene set

Take median p-value as cohort prediction
Geneset Size 200
N = 30 patients

Legend:
Black = SS anchored discovery
Red = Conventional discovery
Problem: Accurately predicting the biologic and statistical significance of the deregulation of a pathway from one normal tissue and one cancer tissue of an individual patient’s transcriptome

TCGA lung adenocarcinoma dataset
Patient individual comparison to external GS (heterogeneity index of dysregulated pathways)

*J Am Med Inform Assoc. 2014 Nov-Dec;21(6):1015-25*
A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations

A genome-by-environment interaction classifier for precision medicine: personal transcriptome response to rhinovirus identifies children prone to asthma exacerbations

Results

Virogram assay + Classifier prediction

Clinical phenotype

Exacerbated

Non-Exacerbated

Exacerbated

Non-Exacerbated
<table>
<thead>
<tr>
<th>Single Subject Analysis</th>
<th>N-of-1-pathways MD</th>
<th>N-of-1-pathways Wilcoxon</th>
<th>FAIME</th>
<th>DEG Enrich / GSEA</th>
<th>DEG / DESeq</th>
</tr>
</thead>
<tbody>
<tr>
<td>clinically relevant</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>unpaired samples</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>paired samples</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>pathway testing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>pathway visualization</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Transcript-level</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>interpretation</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
II. N-of-1-pathways Mahalanobis Distance (MD)

N-of-1-\textit{pathways} Mahalanobis Distance
Producing a Clinically Relevant Metric (CRM)

Pathway Bivariate Gene Expression

Gene-level Differential Expression

\[ d_j = \sqrt{\frac{S_N^2}{S_N^2 S_T^2 - (S_{NT})^2}} (T_j - N_j) \]

Pathway-level Dysregulation

\[ \bar{d} = \frac{1}{m} \sum_{j=1}^{m} d_j \]

Interpreted as (adjusted)
pathway log fold-change
Analysis of aggregated cell-cell statistical distances within pathways unveils therapeutic-resistance mechanisms in circulating tumor cells.

Case study: Drug resistance in CTCs

scRNA-seq Data of CTCs from prostate cancer patients

Miyamoto et al., Science, 2015

- 13 Patients
- 77 CTC RNA-seq
  - 1 to 12 per patient

Knowledge Base Pathway Interaction Database (PID)

Schaefer et al., Nucleic Acids Research, 2009

- 187 pathways w/ at least 15 gene products and high curation confidence
- Signaling pathways often implicated in cancer

EVT-Naïve (41 CTCs)  vs  EVT-Resistant (36 CTCs)
Single-cell statistics for precision medicine in circulating tumor cells (CTCs)

*Bioinformatics.* 2016 Jun 15;32(12):i80-i89

Examples: 2 distinct subjects
Comparison of single-cell RNA-seq analysis methods

<table>
<thead>
<tr>
<th></th>
<th>Aggregated cell-cell distances</th>
<th>SCDE$^2$ + Enrichment</th>
<th>sc Latent Variable Model$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small # of cells</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>2-sample DEP tests</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>2-sample DEG tests</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Cell-centric stats &amp; viz</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Identify cell subpopulations</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

$^1$Buettner et al., *Nature Biotech*, 2015

$^2$Kharchenko et al., *Nature Methods*, 2014
Isogenic single subject method MixEnrich is more accurate than heterogenic cohort-based using larger samples (FDR 5%)

Cohort-based methods were performed across 3, 6 and 12 patients (Pt). The gold standard was built using paired head and neck adenocarcinoma vs non-tumor tissue in 45 subjects.
Haiquan Li, Ikbel Achour, Lisa Bastarache, Joanne Berghout, Vincent Gardeux, Jianrong Li, Younghee Lee, Lorenzo Pesce, Xinan Yang, Kenneth S. Ramos, Ian Foster, Joshua C. Denny, Jason H. Moore, and Yves A. Lussier

Nature partner journals, Genomic Medicine
Article number: 16006 (2016)
What is the central dogma of non-coding (dark) DNA?
Intergenic SNP

Chromatin interaction
ChIA-PET

Suppressor

Enhancer
DNaseI seq

Transcription Factor
binding sites (TFBS)
ChIPseq

Enhancer

The Encyclopedia of DNA Elements
Measure + Annotation

RegulomeDB

nature
ENCODE
GUIDEBOOK TO THE HUMAN GENOME
The ENCODE project description
Intergenic SNP
Enhancer
DNaseI seq
Transcription Factor binding sites (TFBS)
ChIPseq
Suppressor
Chromatin interaction ChIA-PET

Studies
GWAS SNPs in Regulatory Elements

Systematic Localization of Common Disease-Associated Variation in Regulatory DNA
Matthew T. Maurano et al.
Science 337, 1190 (2012);
DOI: 10.1126/science.1222794

Linking disease associations with regulatory information in the human genome
Marc A. Schaub, 1 Alan P. Boyle, 2 Anshul Kundaje, 1 Serafim Batzoglou, 1,3 and Michael Snyder 2,3,4
Our Hypothesis

SNPs associated to same disease

- Intergenic (noncoding) SNPs
- Intragenic SNPs

shared biological mechanisms?
SNP$_1$ to Disease$_1$ through GWAS association.

SNP$_2$ to Disease$_2$ through GWAS association.

Are there shared biological mechanisms between disease risks loci?
Genome-wide expression Quantitative Trait Loci (eQTL)

Causal mechanisms for eQTL


- The transcription of a gene is governed by DNA binding transcription factors (TFs) that switch the gene on or off that are modulated by promoters and enhancers.

- **Cis effect eQTL**: an intragenic polymorphisms might have a clear effect on the expression of a nearby gene (e.g. mutation in a promoter region)

- **Trans effect eQTL**: An intergenic (noncoding) polymorphism may affect the transcription (expression) of distant genes (even on other chromosomes, e.g. distal enhancer)
SNP_1 \quad \text{GWAS association} \quad \text{Disease}_1

\text{eQTL association}

\text{mRNA}_x \quad (1) \text{ shared mRNAs or}

\text{mRNA}_z \quad (2) \text{ similar function}

\text{eQTL association}

SNP_2 \quad \text{GWAS association} \quad \text{Disease}_2
Intergenic SNP\textsubscript{1}

RNA\textsubscript{x}

eQTL

shared mRNAs?

RNA\textsubscript{y}

Intragenic or intergenic SNP

Intragenic

Methods: Significance of shared mRNA over-representation between two SNPs

- Conservative permutation resampling keeping node degree of eQTL-associated mRNAs to GWAS-associated SNPs constant
- Hypergeometric distribution considered anticonservative
Intergenic SNP → eQTL → RNA

Intragenic SNP → RNA → Protein → Pathway

downstream biological mechanism?
Intergenic SNP

Intragenic or intergenic SNP

RNA

Protein

Pathway $i$

Identity or similarity

Pathway $k$
Methods: from gene expression to pathway imputation using ontologies

Gene Ontology
- gene sets annotated to biological processes
- organized as a directed acyclic graph
**Methods:** Reducing Network size with Information Theoretic Semantic Similarity

**SNP1**

- mRNA
  - mRNA
  - mRNA

**SNP2**

- GO
- mRNA
- mRNA

**Share biological mechanisms?**
– Common theoretical bases

• Information content of a concept c:
  \[-\log p(c)\]  
  \(p(c)\) is the occurrence frequency of c and all its descendants in the dataset

• Shared information content between a and b:
  \[-\log p(ms(a,b))\]
  \(ms(a,b)\) is d
  \(ms\): minimal subsumer, the common ancestors with minimal descendants

Methods: 1st level - Semantic Similarity in DAG Ontologies (e.g. GO)

Lin D. An information-theoretic definition of similarity; Proc 5th Int Conf Machine Learning (ICML'98); 1998; 296–304.

\[
\text{sim}_{\text{Lin}}(a,b) = -2\log p(ms(a,b)) - \log p(a) - \log p(b)
\]
similarity score between two concepts (Lin et al.)

\[
\text{Sim}_{\text{Lin}}(a, b) = 2 \times \left( -\log \frac{2034}{82678} \right) / \left[ -\log \frac{19}{82678} - \log \frac{33}{82678} \right] = 0.457
\]

Note: the number beside each concept is the number of occurrences of the concept and all its descendants in a studied GOA file.
Methods: 2nd level - similarity between Two \textit{groups} of GO terms associated to mRNAs*  


\[ SIM(A, B) = 2 \times \sum_{(a_i, b_j) \in P, \text{sim}(a_i, b_j) \geq t} \frac{\text{sim}(a_i, b_j)}{|A| + |B|} \]

- \( P \): the set of paired-up concepts
- \( t \): the threshold value to filter out noise
- \(|A| + |B|\): the total number of concepts in group A and B

- Eliminates comparison of irrelevant concepts
- Standardized (maximum=1; using Lin’s equation)
**Approach**

**Methods:** 3rd level - nested information theoretic similarity: SNP (mRNAs[GO])

- SNPs is associated with a set of mRNAs $G(s1)$,
- $|G(s1)|$ is the cardinality of the set $G(s1)$,
- **GENEITS** is the information theoretic biological similarity of two mRNAs*
- The SNP_ITS provides a score that ranges from 0 to 1; a value of 1 indicated two SNPs with common GO–MFs or GO–BPs, and a value of 0 corresponded to two SNPs with unrelated GO–BPs or GO–MFs.

\[
\text{SNP } _\text{ITS}\left( s_1, s_2 \right) = \sum_{g_i \in G(s_1)} \max_{g_j \in G(s_2)} \left( \text{GENE}_\text{ITS}\left( g_i, g_j \right) \right) + \sum_{g_j \in G(s_2)} \max_{g_i \in G(s_1)} \left( \text{GENE}_\text{ITS}\left( g_i, g_j \right) \right)
\]

\[
= \frac{\sum_{g_i \in G(s_1)} \max_{g_j \in G(s_2)} \left( \text{GENE}_\text{ITS}\left( g_i, g_j \right) \right) + \sum_{g_j \in G(s_2)} \max_{g_i \in G(s_1)} \left( \text{GENE}_\text{ITS}\left( g_i, g_j \right) \right)}{|G(s_1)| + |G(s_2)|}
\]

* Bioinformatics. 2007 Jul 1;23(13):i529-38.
Methods: Reducing Network size with GO Information Theoretic Semantic Similarity (ITS)
Methods: Reducing Network size with mRNA Information Theoretic Semantic Similarity (ITS)

\[
\text{mRNA}_\text{ITS} = 0.7
\]
Methods: Reducing network size with 3 nested similarities: SNP_ITS{ mRNA_ITS [GO_ITS] }
Intergenic SNP\textsubscript{1} \text{ GWAS\textsubscript{1} } \text{ GWAS\textsubscript{2} } \text{ Intragenic or intergenic SNP\textsubscript{2}}

RNA \xrightarrow{\text{eQTL}} \text{ Protein} \xrightarrow{\text{Gene ontology}} \text{ Pathway}_i \text{ identity or similarity} \text{ Pathway}_z

same disease
Methods: Assessing the pvalues of SNP\_ITS\{ mRNA\_ITS [GO\_ITS] \}

Li (...) Lussier. Nature partner journals, Genomic Medicine; 16006 (2016)

- **Scale-free permutation resampling**: node degree of each GO terms, each mRNA, each SNP remains constant in each permutation, simply with different molecules connected to each other.
- Theoretical network over-representations/enrichment calculations were shown anticonservative.
- ~20,000,000 core hours of high-throughput computations were conducted
  - Beagle GLOBUS61 GRID computing
  - Cray XE6 Supercomputer of the Computation Institute at the Argonne National Laboratory (http://beagle.ci.uchicago.edu/).
  - peak performance of 151 teraflops generated by 17,424 compute cores
Methods: big data / multiscale permutation resampling were performed for imputing complex disease mechanisms in our scale-free network. Theoretical distributions are anti-conservative.
Approach: Input

- GWAS
- eQTLs

1092 intergenic SNPs
1266 intragenic SNPs
associated to
467 diseases
6301 mRNAs

~ 2 Million Lead Pairs

- Lead SNP pairs:
  - Intergenic - Intergenic
  - Intergenic - intragenic
Datasets

GWAS (NHGRI)
- SNPs → disease

ENCODE
- Regulome-DB
- ChIP-seq, ChIA-PET

eQTLs (LCLs, Liver)
- SNPs → mRNAs

Protein-Protein Interaction (PPI, STRING)

Gene Ontology
- Gene/mRNA → Molecular Function (MF)
- Gene/mRNA → Biological Process (BP)
Approach: Input, excludes SNPs in Linkage Disequilibrium (LD)

Pairwise Analysis

~ 2 Million Lead Pairs (LD r²<0.8)
Approach overview

Pairwise Analysis

- ~ 2 Million Lead Pairs
- Intergenic
- Intragenic

Prioritization

- Shared mechanisms

Biological Knowledge
(Gene Ontology)

- Prioritized inter-inter & inter-intra SNP pairs by overlap and/or similarity of:
  - mRNA
  - Biological process
  - Molecular function

Lead SNP pairs:
- Intergenic - Intergenic
- Intergenic - intragenic

High-throughput computing

Prioritized Lead SNP pairs
Non-prioritized Lead SNP pairs
Approach overview

Pairwise Analysis

Prioritization
Shared mechanisms

Validation
Shared genetics & molecular mechanisms

~ 2 Million Lead Pairs

Biological Knowledge
(Gene Ontology)

1-Genetic Interaction validation
2-Encode Validation

Prioritized inter-inter & inter-intra SNP pairs by overlap and/or similarity of:

- mRNA
- and/or
- Biological process
- and/or
- Molecular function

High-throughput computing

Prioritized Lead SNP pairs
Non-prioritized Lead SNP pairs

Enrichment analysis
Result 1.2 – Prioritized Lead SNP-pairs?

Count of distinct Lead SNPs among 5011 Lead SNP pairs sharing similar mechanisms.

Input:

- **1,092** Intergenic Lead SNPs
- **1,266** Intragenic Lead SNPs
- **6,301** Associated mRNAs (eQTL)
- **467** Diseases

Percent prioritized:

- **100%** at FDR<0.05
- **50%** at FDR<0.01
- **25%** at FDR<0.05

Prioritization threshold

Statistical mRNA overlap
Molecular function similarity
Biological process similarity
**Result 1.4** Cicos plot of SNP-pairs prioritized within the same disease showing SNP sharing mechanism across chromosomes!
Result 1.5 – SNP-SNP network
Result 1.5 – SNP-SNP network across diseases within class
Result 1.6 – Prioritized Lead SNP-pairs?

at FDR<0.05
Same mechanisms shared among prioritized SNP pairs associated with same disease
Result 2.1 – Prioritized Lead SNP-pairs for Rheumatoid Arthritis (RA)
Result 2.3 – Example of Lead SNPs sharing mechanism in RA

antigen processing and presentation of peptide or polysaccharide antigen via MHC class II

- HLA-DQB1
- HLA-DRB3
- HLA-DRB1
- HLA-DRB4
- HLA-DRB5

Antigen processing presentation

- HLA-C
- HLA-B
- MICB
- MICA

gamma-delta T cell activation

natural killer cell lectin-like receptor binding

rs6457620 chr 6

rs615672 chr 6

rs7404928 chr 16

PRKCB1

Rheumatoid arthritis
Result 2.4 – Disease specific enrichment?

Prioritized SNP pairs by Statistical mRNA overlap

Prioritized SNP pairs by Molecular Function similarity

Prioritized SNP pairs by Biological Process similarity

<table>
<thead>
<tr>
<th></th>
<th>Same disease SNP-pairs</th>
<th>Distinct disease SNP-pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prioritized SNP-pairs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non Prioritized SNP-pairs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Contingency table
Odds ratio
p-value
Result 2.5 – Disease-specific enrichment?

Prioritized SNP pairs sharing similar mechanisms are more likely to be associated to the same disease to unrelated pathologies.
Result 3 – Genetic Interaction validation in RA PheWAS

<table>
<thead>
<tr>
<th>Combinations of alleles: OR [95% CI]</th>
<th>SNP (1) rs6457617</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T/T</td>
</tr>
<tr>
<td>T/T</td>
<td>1.33 n.s.</td>
</tr>
<tr>
<td>T/G</td>
<td>1 n.s.</td>
</tr>
<tr>
<td>G/G</td>
<td>2.16 [1.95-2.40]</td>
</tr>
</tbody>
</table>

**SNP (2) rs9272219**
### Result 3 – Genetic Interaction validation in GWAS

Non-additive genetic interaction of prioritized inter–inter and inter–intra Lead SNP pairs validated in independent GWAS studies

<table>
<thead>
<tr>
<th>Disease</th>
<th>Prioritized SNP pairs</th>
<th>SNPs with synergistic effects</th>
<th>Entropy P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s</td>
<td>rs4509693–rs753129 (chr10, inter) (chr4, inter) rs7081208*–rs9331888* (chr10, FRMD4A) (chr8, CLU, MIR6843)</td>
<td>rs4509693–rs753129–rs7081208*</td>
<td>0.046</td>
</tr>
<tr>
<td>Bladder cancer</td>
<td>rs8102137–rs1014971 (chr19, inter) (chr22, inter)</td>
<td>rs8102137–rs1014971</td>
<td>0.039</td>
</tr>
</tbody>
</table>
Result 4 – ENCODE validation
Legend: **TF**=transcriptional factor  **PPI**=protein-protein interaction network (STRING)

**ChIP-seq**= ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing (seq) to identify the binding sites of DNA-associated proteins.

**ChIA-PET**= Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) is a technique that incorporates chromatin immunoprecipitation (ChIP)-based enrichment, chromatin proximity ligation, Paired-End Tags (PET), and High-throughput sequencing to determine de novo long-range chromatin interactions genome-wide.
Approach

Result 4 – ENCODE validation, any prediction of SNP-pair

Prioritized Lead SNPs are in enriched in similar regulatory elements
Conclusion

intergenic and intragenic SNPs associated to the same disease

1- most likely affect:
   > expression of the same mRNAs
   > mRNAs involved in similar biological pathways
   > mRNAs governed by similar regulatory mechanisms
      (e.g. chromatin interactions)

2- as many as 40% could display synergetic and antagonistic genetic interactions with SNPs of the same disease or those of another one

3- provide druggable targets downstream of noncoding intergenic SNPs for novel common disease risk prevention and treatment
Take home message

- Therapies for disease risks found in the intergenic genome
  - by extending the central dogma of molecular biology using Bois’ information theory
  - pathologic mechanisms identified at the convergence of eQTL signals
  - Validations of convergent signal
    - overrepresentation of drug bank targets
    - Overrepresentation of ENCODE mechanisms
    - diseases sharing intergenic mechanisms found comorbid in EHRs
    - predicts novel genetic interactions in Alzheimer’s and in bladder cancer confirmed in GWAS

- Genome-by-environment interaction (G × E) assays & interpretation:
  - enables single-subject studies
  - smaller classifier sizes
  - democratizing costs: pathway-level signal using qPCRs and representative transcript
Yves A. Lussier, MD
Haiquan Li, PhD
Ikbel Achour, PhD
Colleen Kenost, EdD
Jianrong Li, MSc
Joanne Berghout, PhD
Nima Pouladi, MD, PhD
A. Grant Schissler, MS
Qike Li, MS

Yves@email.arizona.edu
Hiring post-doc fellows & Research Assistant Professors