Program Project Grant on

"Statistical Methods for Integrative Genomics in Cancer"

DIVISIONS OF BIOSTATISTICS AND BIOINFORMATICS

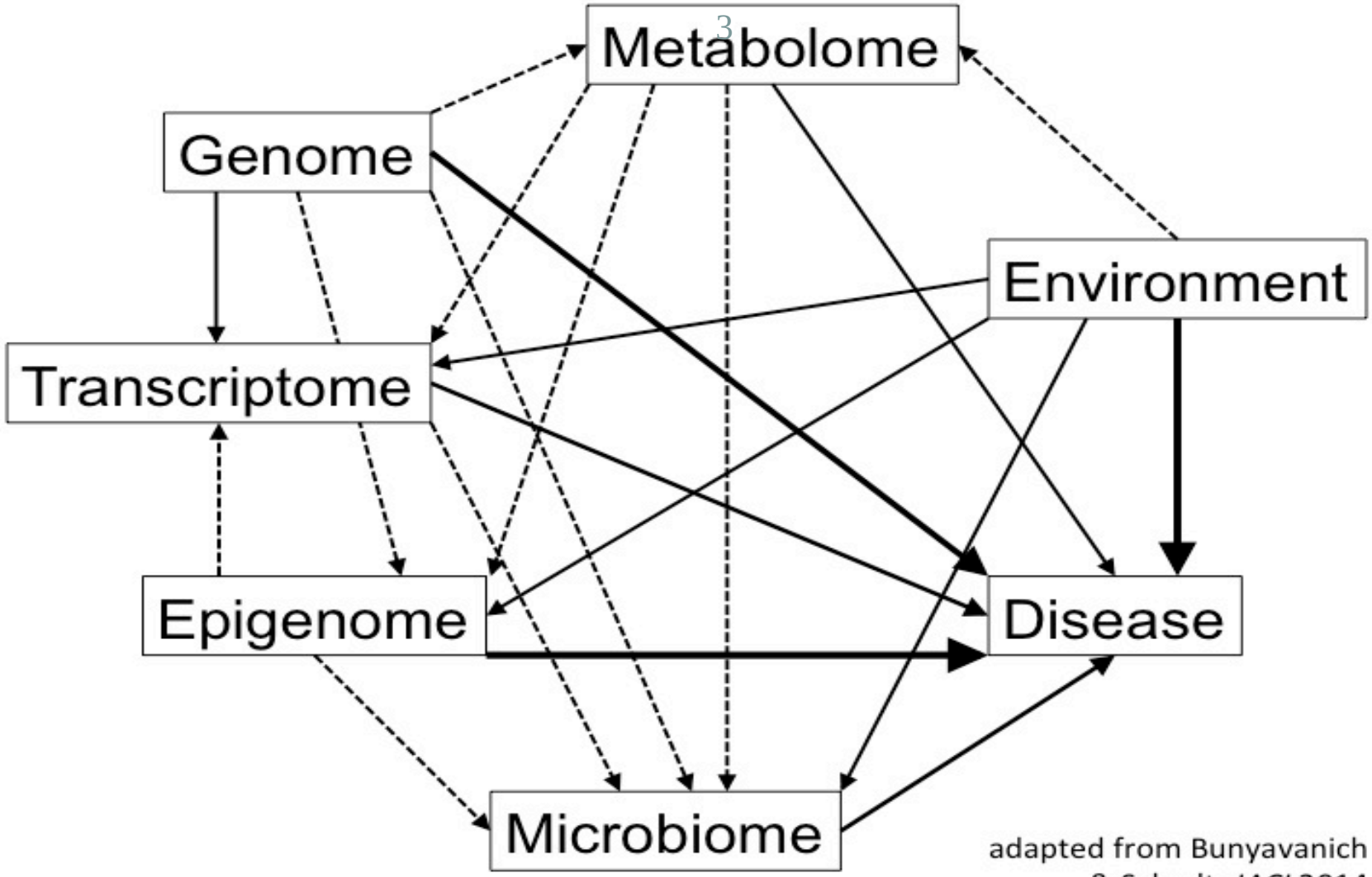
UNIVERSITY OF SOUTHERN CALIFORNIA

Keck School of Medicine of USC

Context: Integrative Genomics

- In the "post-GWAS" world, emphasis has shifted to
  - Mining the "lower Manhattan" to discover biological pathways by exploiting external information
  - Further discovery, e.g., sequencing for rare variants
  - Characterization of biological function
  - Somatic variation and heterogeneity
  - Building and validating comprehensive risk models incorporating multiple genes and environmental factors
- To accomplish this, many groups are exploring techniques of "systems biology" and "integrative genomics" to combine information across many different data types

A Knowledge Network for Asthma and Allergic Diseases



adapted from Bunyavanich & Schadt, *JACI* 2014

Theme

We aim to develop novel statistical methods to address some of the major problems facing cancer genetic epidemiologists in the "post-GWAS" era.

Our overall focus is on integrating information across different data types:

- Across –omics platforms
- Across phenotypes
- Across species

and for especially incorporating the wealth of prior biological knowledge available in various "ontologies" for genomic annotation, pathways, etc.

Examples include tumor heterogeneity, the microbiome, the exposome.

Our methods will be translational, providing tools for

- Development and evaluation of robust risk prediction models from high-dim data
- Assessing genetically-targeted risk factor intervention or screening programs.

Projects

1. Hierarchical Modeling (Lewinger & Conti)
2. Inference of gene function (P Thomas)
3. Exposome & Microbiome (D Thomas & Millstein)
4. Tumor evolution (Siegmund)

Cores

A. Administration (Gauderman & Thomas)

includes a distinguished External Advisory Committee

(Schaid, Schadt, Freedman, Peters, Miller, Lange, Coetzee, Suchard)

B. Genome annotation (Mi)

C. High-performance computation & simulation (Chen)

D. Data management & software development (Gauderman)

Project 1: Integrated Analysis for Observational Genomic Data

PI: DAVID CONTI

CO-INVESTIGATORS:

DUNCAN THOMAS

GARY CHEN

Keck School of Medicine of USC

Integration of Information

- Public databases that provide genome annotation.
  - Gene Ontology
- Sub-studies that obtain additional data in same samples from larger observational studies.
  - Colon Cancer Family Registry Folate Biomarker Study
- Experimental studies that obtain –omic data on independent samples.
  - The Cancer Genome Atlas

Hierarchical Models

- Incorporates external knowledge about pathways as "prior covariates" for coefficients of a data model
- Level I: Epidemiologic data model:

$$\text{logit}\ \Pr(Y_i = 1 | X_i) = \beta_0 + \sum_p \beta_p X_{ip}$$

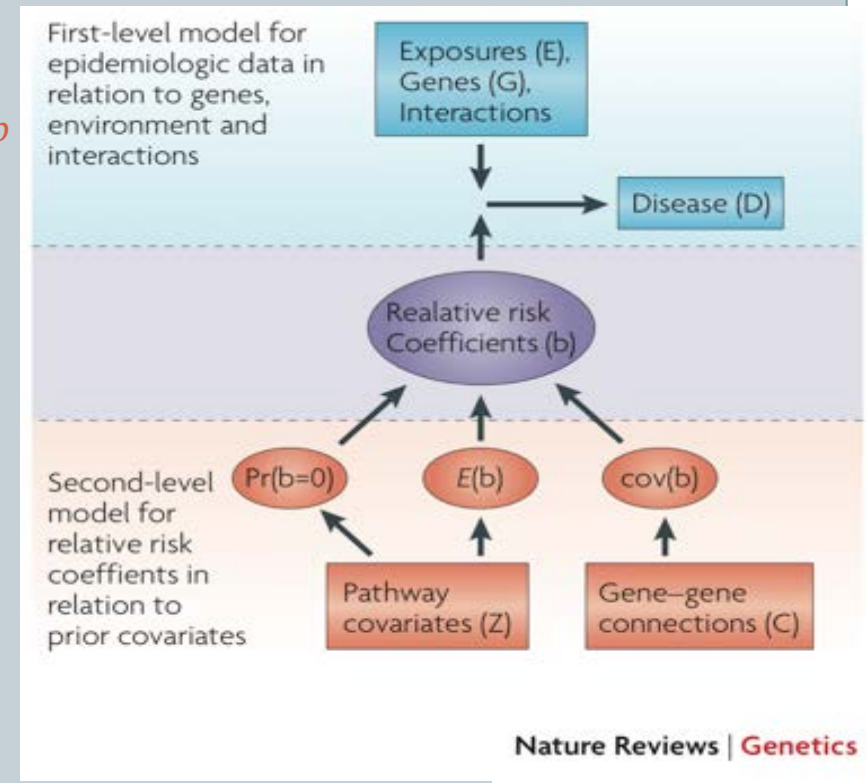
$$X = (G, E, GxE, GxG, GxGxE, \ldots)$$

- Level II: Pathway model:

$$\beta_p \sim N(\Sigma_v\, \alpha_v Z_{pv}, \sigma^2)$$

Z_{pv} = prior covariates:

  - other –omics data
  - experimental assays
  - ontologies

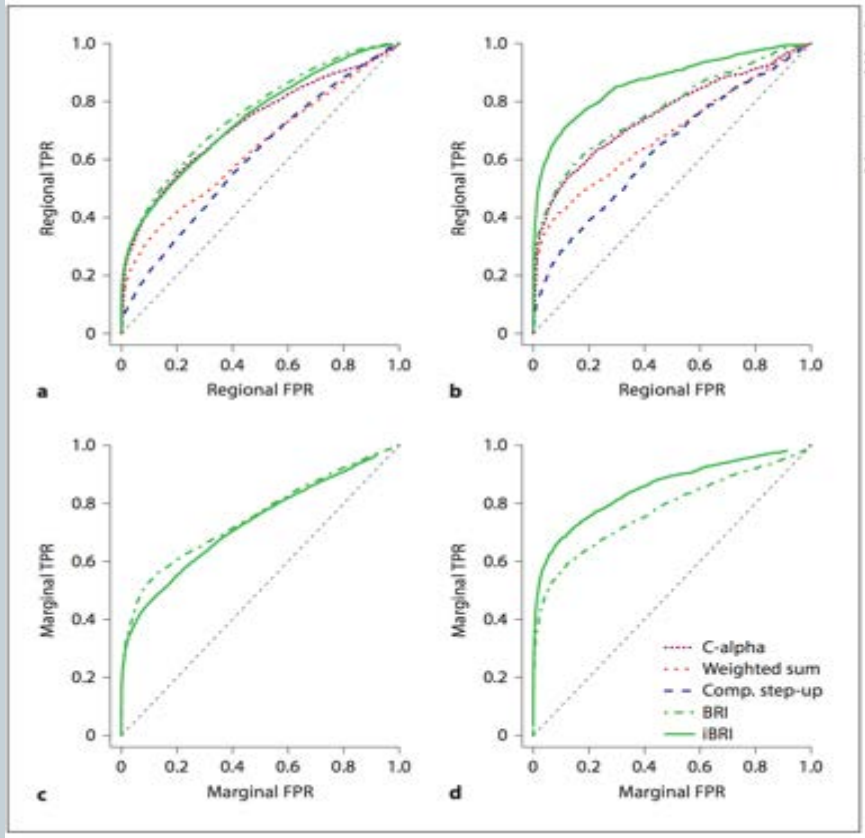


Nature Reviews | Genetics

Keck School of Medicine of USC

Filtering vs. Hierarchical Modeling

- Filtering improves power by reducing multiple comparisons burden
  - but only if external information is good
  - otherwise can reduce power by excluding true positives from consideration
- Hierarchical modeling has little loss of power from including some irrelevant prior covariates
  - they get little or no weight in second level model
  - requires no arbitrary weighting of multiple sources of external information



Quintana et al, *Hum Hered* 2012: 74: 184-95

Specific Aims

Aim 1: To integrate multiple types of **genomic data from tumors** into prediction models via structured regularized regression.

- Structured penalties based on biological knowledge
- Efficient and scalable fitting algorithms (convex optimization: MM, coordinate descent)
- Model selection methods that incorporates the biological structure
- Performance metrics that better reflect prediction goal

Specific Aims

Aim 2: To incorporate gene function information **from the gene ontology database (GO)** to enhance prediction modeling using structured regularized regression

- Structured penalties based on GO
- Incorporate uncertainty due to incomplete annotation
- Efficient fitting algorithms (convex optimization: MM, coordinate descent)
- Model selection methods that incorporates the GO structure

Specific Aims

- Aim 3: Causal inference methods to evaluate the impact of genomic-based prediction tools for clinical use.
- Adapt counterfactual framework and G-computation to estimate the causal effects.
- Can also be used to evaluate impact of other types of interventions: personalize screening protocols based on genetic risk score (e.g. colonoscopy frequency) or modifiable risk factors (e.g. obesity, diet smoking, alcohol)

REMOVED!

Project 2:
Inference of gene function

PI: PAUL THOMAS

CO-INVESTIGATORS:
HUAIYU MI
GARY CHEN
DUNCAN THOMAS?

Keck School of Medicine of USC

Rationale

- Human-curated evolutionary inference has been used by the GO Consortium to solve this challenge
  - Approach published (Gaudet et al, Briefings Bioinform. 2011)
  - has been used to dramatically increase the GO annotations for ~800 human genes
- Scaling up to all 20,000 genes will take decades, during which time much more experimental information will accumulate
- A computational approach is much faster, and can be easily updated to include inferences from new experimental information

BRIEFINGS IN BIOINFORMATICS. VOL 12. NO 5. 449–462
Advance Access published on 27 August 2011

Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium

Pascale Gaudet, Michael S. Livstone, Suzanna E. Lewis and Paul D. Thomas

Submitted: 8th March 2011; Received (in revised form): 8th June 2011

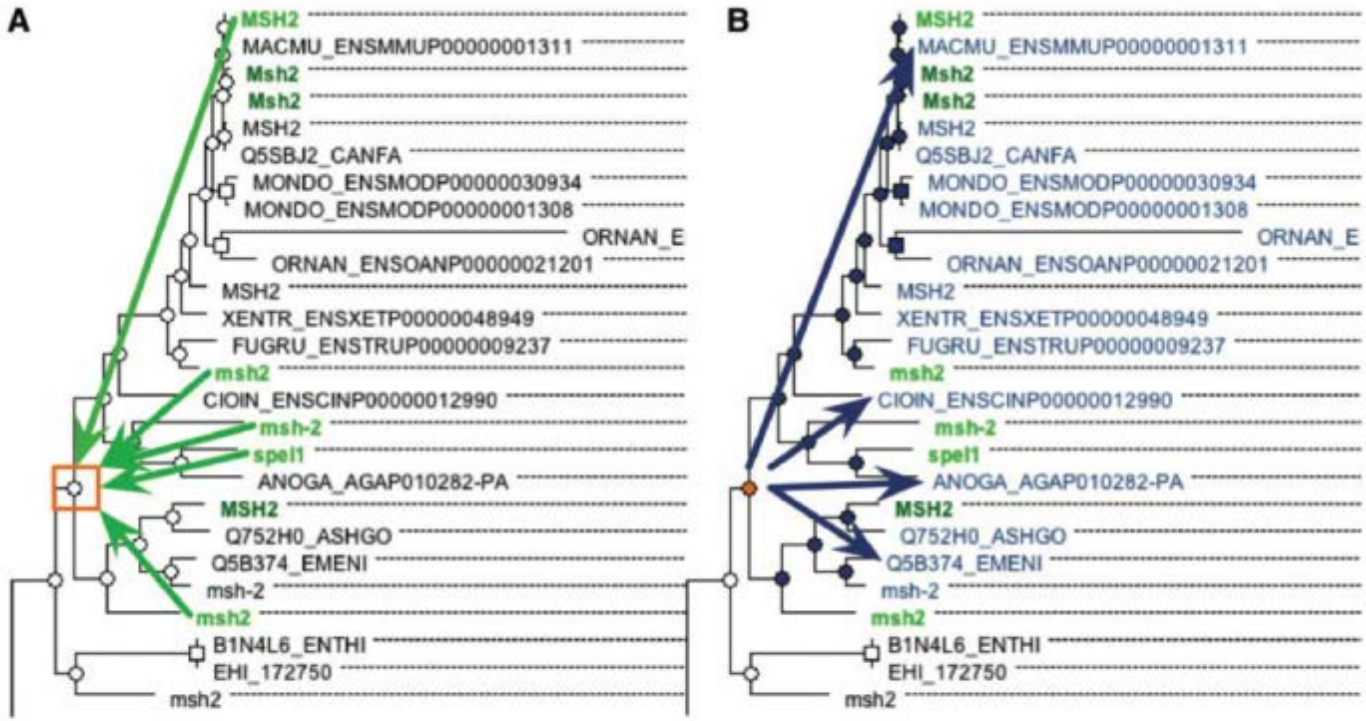
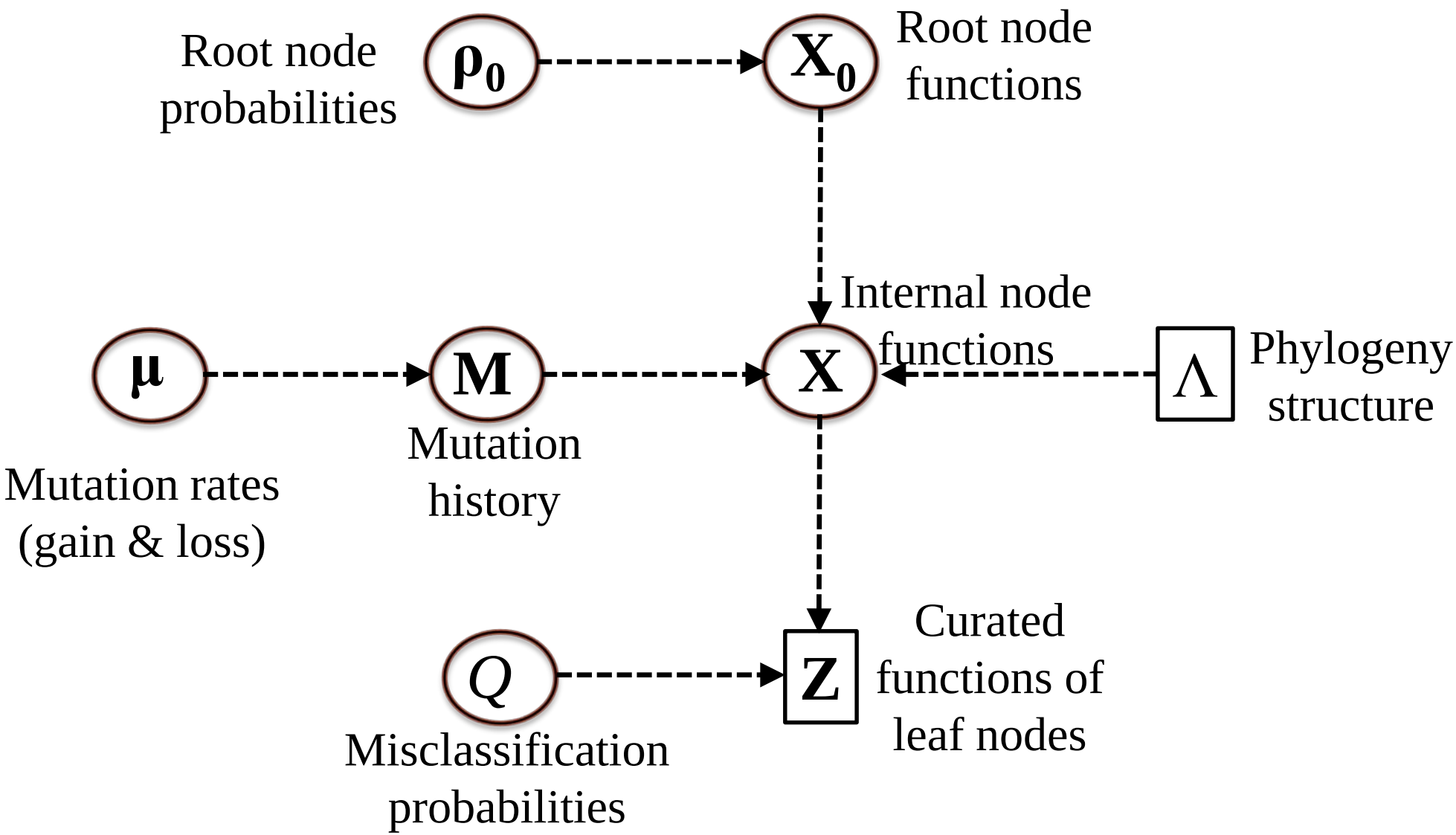
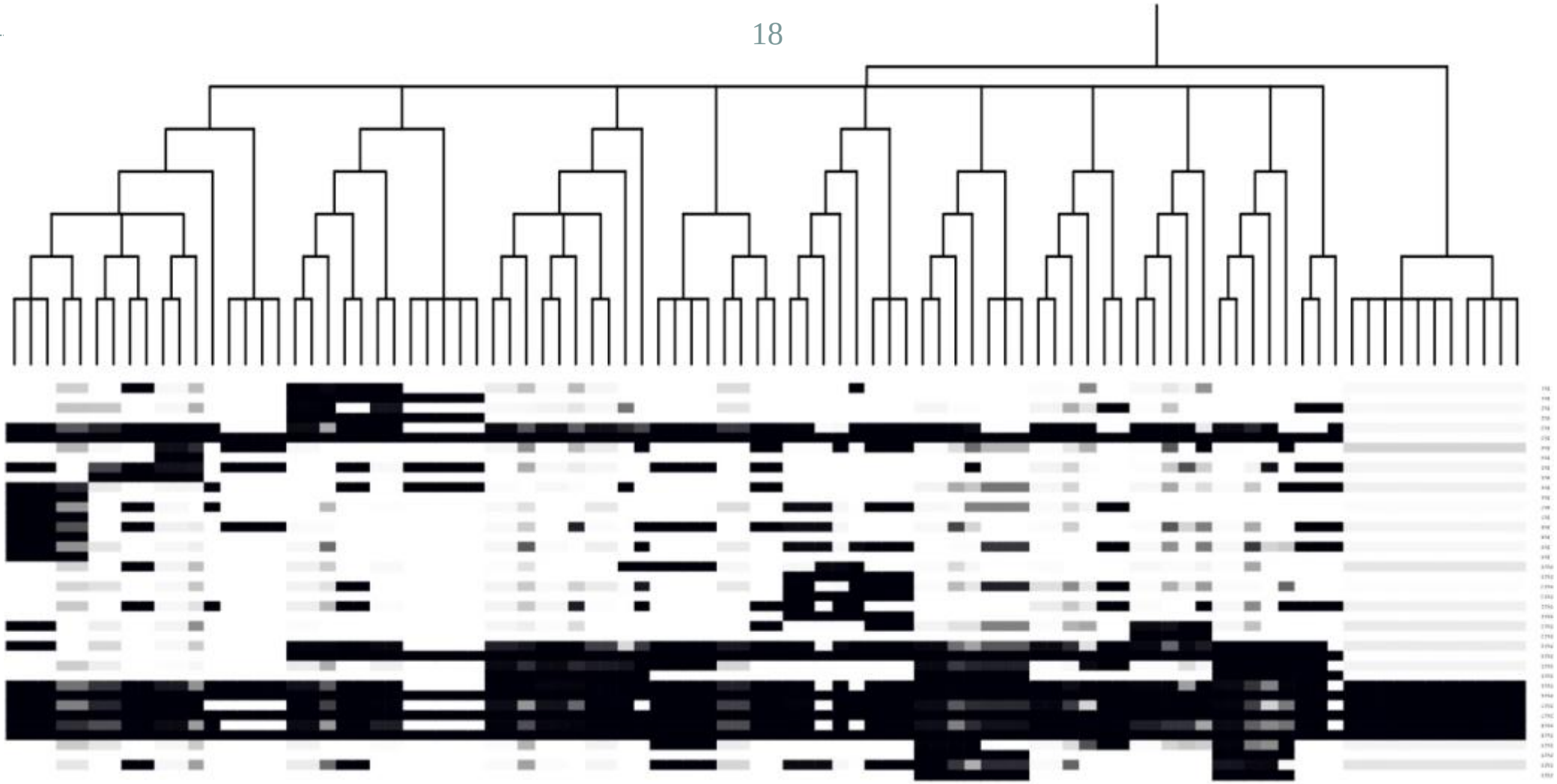


Figure 1: The concept of PAINT. This example presents a MutS homolog family showing experimental evidence for 'GO term'. (**A**) Primary experimentally based annotations to one term or any of its ancestors (light green labels) are used to infer that the most recent common ancestor (CA) of the all those proteins also had that function. The curator notes this by dragging the term onto the node of the MCRA (orange box). (**B**) Subsequently, PAINT propagated this annotation forward to other descendant leaves (blue labels).


Root node probabilities
ρ_0
X_0
Root node functions
Internal node functions
μ
M
X
Λ
Phylogeny structure
Mutation rates (gain & loss)
Mutation history
Q
Z
Curated functions of leaf nodes
Misclassification probabilities

20 Curated vs Predicted Functions for Caspase Genes Across 92 Species



Project 3:
The Internal Environment: Exposome and Microbiome

PI: DUNCAN THOMAS

COINVESTIGATORS:

JOSH MILLSTEIN

DAN STRAM

WENDY COZEN

Exposome: Motivation

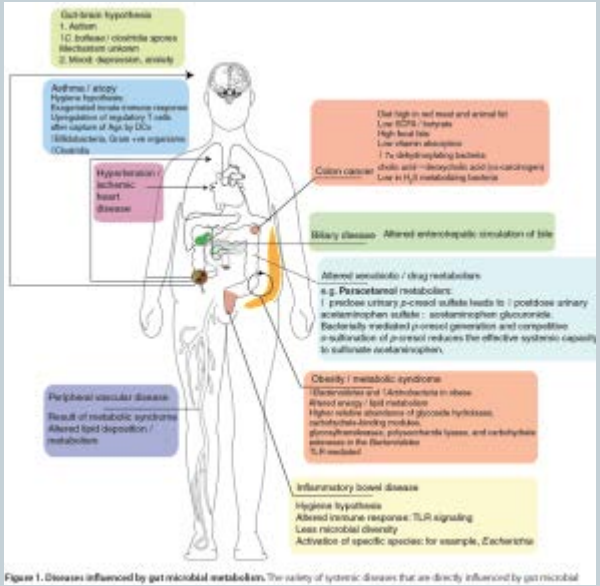
- Opportunity to examine environmental exposures in a comprehensive manner using mass spectrometry of biospecimens (Wild, CEBP 2005)
- Opportunity to conduct disease association (EWAS) and GxE studies (GEWIS) in an agnostic manner comparable to GWAS (Patel, PLoS One 2010)
- Study design and statistical analysis issues largely unstudied, but pose several novel challenges (Chadeau-Hyam et al, Env Mol Mut 2013; Biomark 2011)
- Unlike the genome, the exposome is modifiable

Unique Challenges in Exposomics

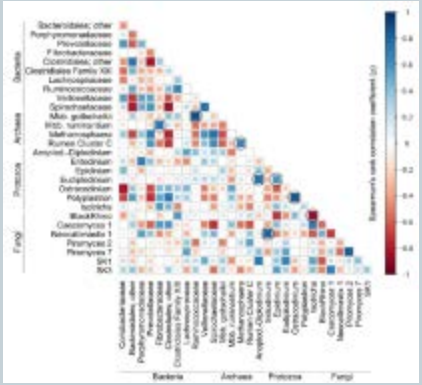
- Time-dependent exposures
- Reverse causation
- Confounding by population substructure (culture, SEP, diet, etc.)
- Measurement error
- Multiple comparisons (correlation structure quite different from LD)
- Relevant sources of biospecimen

The Microbiome: Motivation

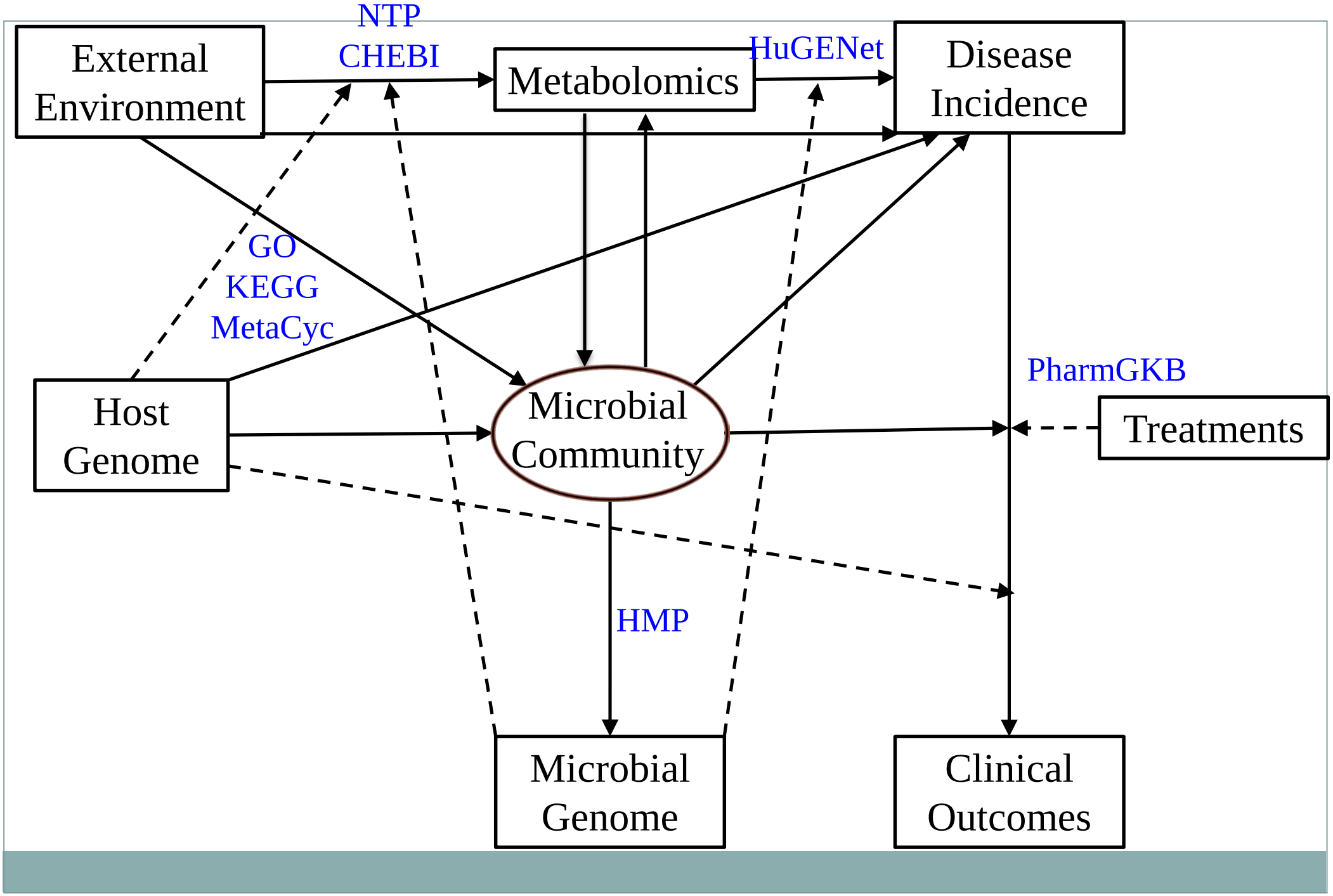
- Gut microbiome has been implicated in diverse disease states including cancer and obesity, as well as drug metabolism and toxicity.
- Gut microbiome provides a potential point of therapeutic intervention.
- Current methods describe association but do not identify directed causal relationships.



Kinross 2011



Kittelmann 2013


External Environment
NTP
CHEBI
Metabolomics
HuGENet
Disease Incidence
GO
KEGG
MetaCyc
Host Genome
Microbial Community
PharmGKB
Treatments
HMP
Microbial Genome
Clinical Outcomes

Integrated Aims

1. Modeling the influence of exogenous exposures and host genotype on the internal environment (metabolome and microbiome) and their inter-dependence
2. Modeling the effect of the internal environment on disease risk and its mediation of the effects of exposures and genes
3. Build a comprehensive Bayesian network for the entire system

... details to follow (after P01 overview)

Real Data Applications

- Multi-ethnic cohort obesity P01
  - project 1: diet, hormones, and obesity phenotypes
  - project 2: GWAS
  - project 3: metabolome
  - project 4: microbiome
    - 2000 MEC participants with exquisite obesity phenotyping
    - nested case-control study of breast and colorectal cancer (1000 cases & 1000 controls each)
- Microbiome of twins discordant polyps
  - 60 MZ pairs with diet (incl. recent changes), 16s rRNA microbiome (3 months apart), 25 immune markers
- ColoCare consortium
  - 1140 CRC patients with clinical outcomes at 3, 6, 12, 24 & 36 months, metabolomics (urine, plasma, fat), microbiome, GWAS, expression, methylation

Project 4:
Tumor Evolution

PI: KIM SIEGMUND

CO-INVESTIGATORS:

PAUL MARJORAM
CHRISTINA CURTIS
DARYL SHIBATA

Keck School of Medicine of USC

A Big Bang model of human colorectal tumor growth

Andrea Sottoriva[1,6], Haeyoun Kang[2,3], Zhicheng Ma[1,6], Trevor A Graham[4,5], Matthew P Salomon[1], Junsong Zhao[1], Paul Marjoram[1], Kimberly Siegmund[1], Michael F Press[2], Darryl Shibata[2] & Christina Curtis[1,6]

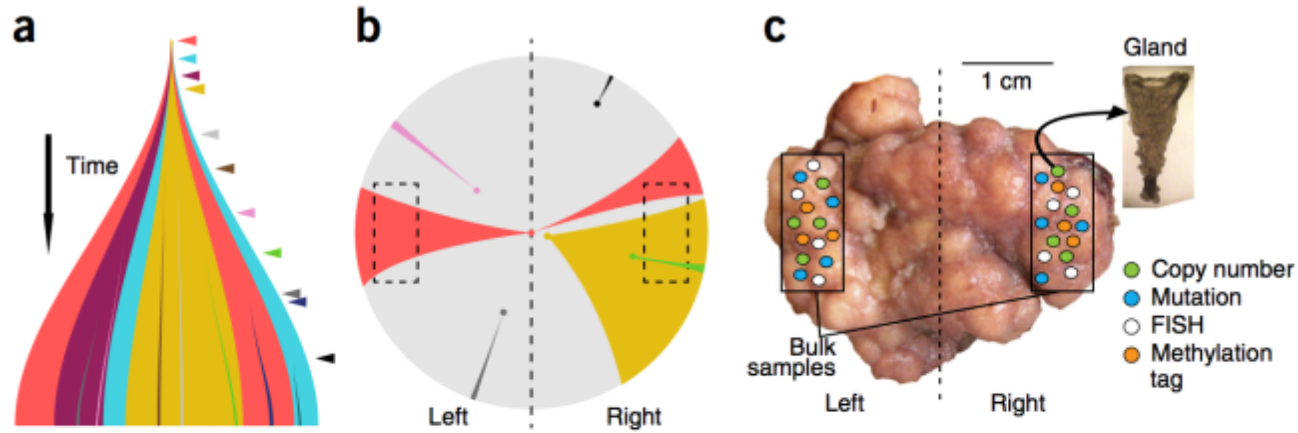


Figure 1 The Big Bang model of tumor growth. (**a**) After initiation, a tumor grows predominantly as a single expansion populated by numerous heterogeneous subclones. ITH results from private alterations (colored arrowheads) that continuously accumulate owing to replication errors. In addition to public alterations present in the first transformed cell, private alterations acquired early persist and become pervasive in the final tumor although remaining non-dominant (colored segments). Late-arising alterations are only present in small regions of the tumor. (**b**) In the Big Bang model, the pervasiveness of private alterations depends on when the alteration occurs during growth, rather than on selection for that alteration. The schematic illustrates how early private alterations, despite remaining non-dominant, are pervasive within the tumor (for example, red and yellow) and can be found in distant regions, thus appearing variegated (for example, red). This is owing to aberrant subclone mixing in the primordial tumor, followed by scattering during expansion. Late alterations are restricted to small regions (for example, black, pink, gray) and are essentially undetectable by conventional bulk genomic profiling. Distance from the dashed vertical axis corresponds to increasingly late onset for alterations. Dashed boxes represent sampled regions. (**c**) We sampled an average of 23 individual tumor glands (<10,000 cells) from distant regions (~0.5 cm^3 in size) and bulk (left and right) samples from the remaining tissue. Samples were profiled using several genomic techniques, including copy number analysis, whole-exome and targeted sequencing, neutral methylation tag sequencing and FISH, providing a panoramic view of genomic alterations throughout the tumor on multiple spatial scales.

Nature Genetics **47**, 209–216 (2015)

Keck School of Medicine of USC

Approach

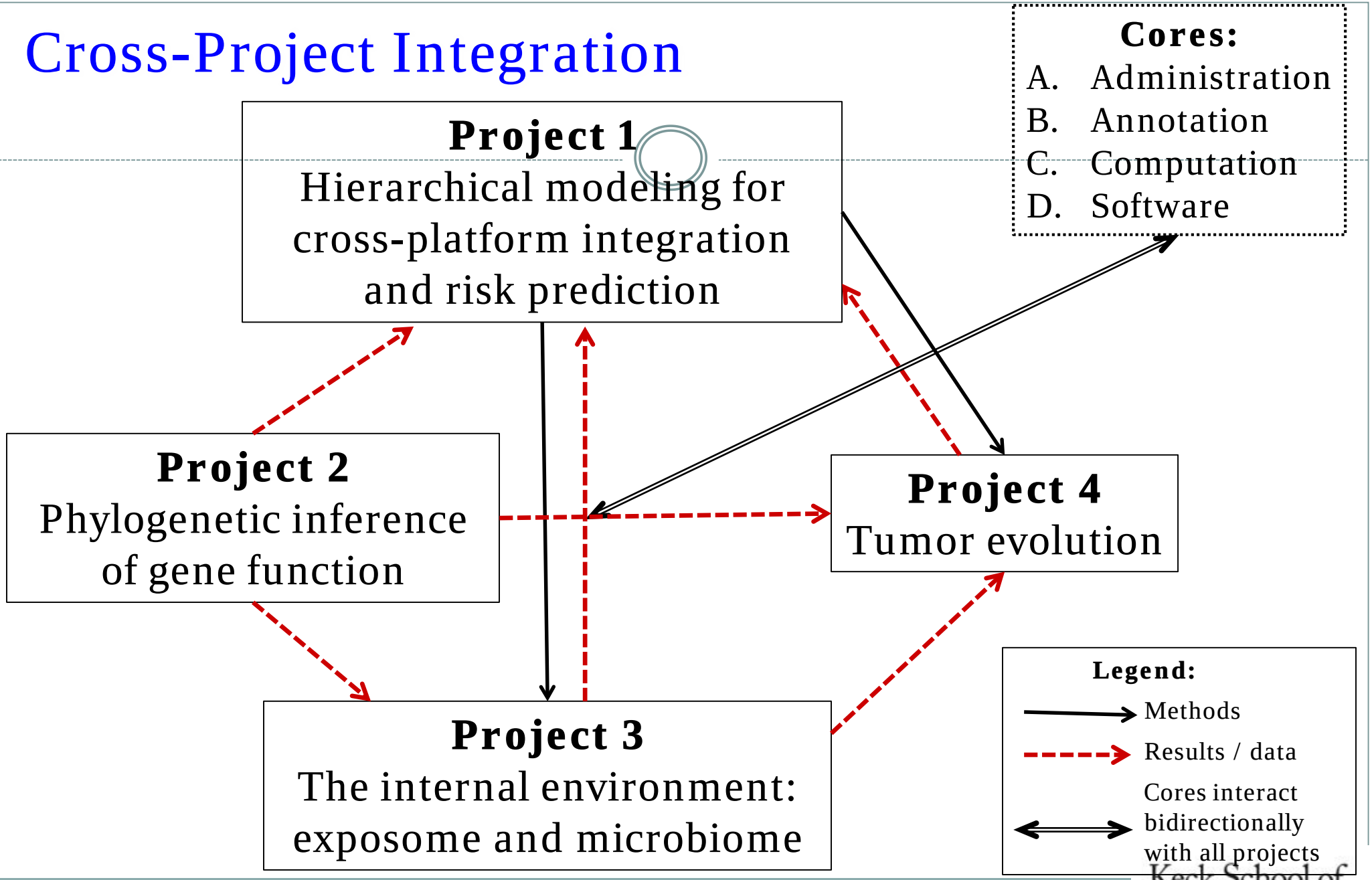
Challenge: Cannot observe tumor growth prospectively

Approach: Model the growth retrospectively

Develop 'Molecular Phylogeny' to model tumor growth

- Cancer cells share common ancestor
- Cancer cells show many 'passenger changes'
  - DNA methylation 'clocks', somatic mutations (WES), DNA copy number alt.
- Sample multiple subpopulations of cells/tumor
  - Isolated glands (~10,000 cells)

Cross-Project Integration



Back to Project 3: Exposome-specific Aims

1. To explore statistical issues that distinguish exposome studies from analogous genome studies and to develop procedures that will be robust to these problems
2. To develop study designs and methods of analysis for EWAS and GEWIS
3. To extend our approaches to examining epigenetic mediation of exposure-response relationships in a similarly agnostic manner

Microbiome-specific Aims

1. Develop network based approach to integrate data such as diet and genomic variation in the host with gut microbiome data and disease phenotypes in order to identify disease pathways.
2. Develop approach to quantify uncertainty in the estimated network/disease pathway and validate in simulation studies.
3. Develop software in the form of a Bioconductor package to freely distribute to the scientific community.

Mediation of Exposure and Genetic Effects on Disease Through the Internal Environment

- Overall hierarchical model structure:

$$\text{logit Pr}(Y_i=1) = \beta_0 + \mathbf{M}_i'\beta^{(1)} + \mathbf{C}_i'\beta^{(2)} + \mathbf{M}_i \otimes \mathbf{C}_i'\beta^{(3)}$$

$$\beta^{(k)} \sim N_{p^{(k)}}(\mathbf{Z}_j'\pi_j, \mathbf{A}_j)$$

- Aim 1: Exposure (E) and genetic (G) effects on metabolome (M) and microbial community (C)

$$\mathbf{M}_{ik} = \boldsymbol{a}_i + \mathbf{E}_{ik}'\alpha_1 + \mathbf{G}_i'\alpha_2 + \mathbf{E}_{ik} \otimes \mathbf{G}_i'\alpha_3 + \mathbf{M}_{i,k-1}'\alpha_4 + \mathbf{C}_{i,k-1}'\alpha_5 + e_{ik}^{M}$$

$$\mathbf{C}_{ik} = \boldsymbol{b}_i + \mathbf{E}_{ik}'\beta_1 + \mathbf{G}_i'\beta_2 + \mathbf{E}_{ik} \otimes \mathbf{G}_i'\beta_3 + \mathbf{C}_{i,k-1}'\beta_4 + \mathbf{M}_{i,k-1}'\beta_5 + e_{ik}^{C}$$

- Aim 2: Joint effects of external and internal factors on disease

$$\lambda_i(t) = \lambda_0(t) \exp[\beta^{(1)}E_i + \beta^{(2)}G_i + \beta^{(3)}E_i \times G_i + \gamma M_i(t)]$$

growth curve model for $M(t)$:

$$M_{ik}^* = M_i(t_k) + e_{ik}$$

where $M_i(t) = a_{0i} + \alpha_{01}E_i + \alpha_{02}G_i + \alpha_{03}E_i \times G_i + (a_{1i} + \alpha_{11}E_i + \alpha_{12}G_i + \alpha_{13}E_i \times G_i)\, t$

"Meeting in the Middle": A Hierarchical Bayes Approach

- Basic idea: to identify a subset of metabolomic markers that are associated with both exposure and disease

Vineis & Chadeau-Hyam, *Curr Opin Oncol* 2011;**23**(1): 100-5.
Chadeau-Hyam et al., *Biomarkers* 2011;**16**(1): 83-8
Assi et al., *Mutagenesis* 2015

- Multi-level model:

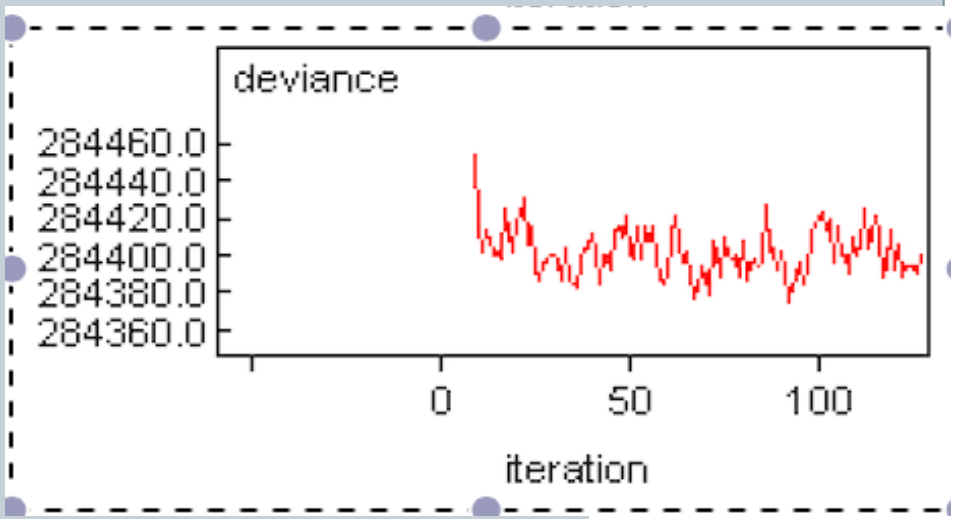
$$f(E(B_j) = \alpha_0 + \alpha_{1j} X V_j$$
$$f(E(Y)) = \beta_0 + \beta_{1j} B_j W_j$$

$$\Pr(V_j = v, W_j = w) = \pi_{vw}$$
$$(\alpha_{1j}, \beta_{1j}) \sim N_2(\mathbf{Z}_j'\boldsymbol{\eta}, \boldsymbol{\Sigma})$$

- Inference is based on posterior probability that $V_j = W_j = 1$.
- Could be extended to multivariate $\mathbf{X}$
- Prior models can be added to α_{1j} and β_{1j}

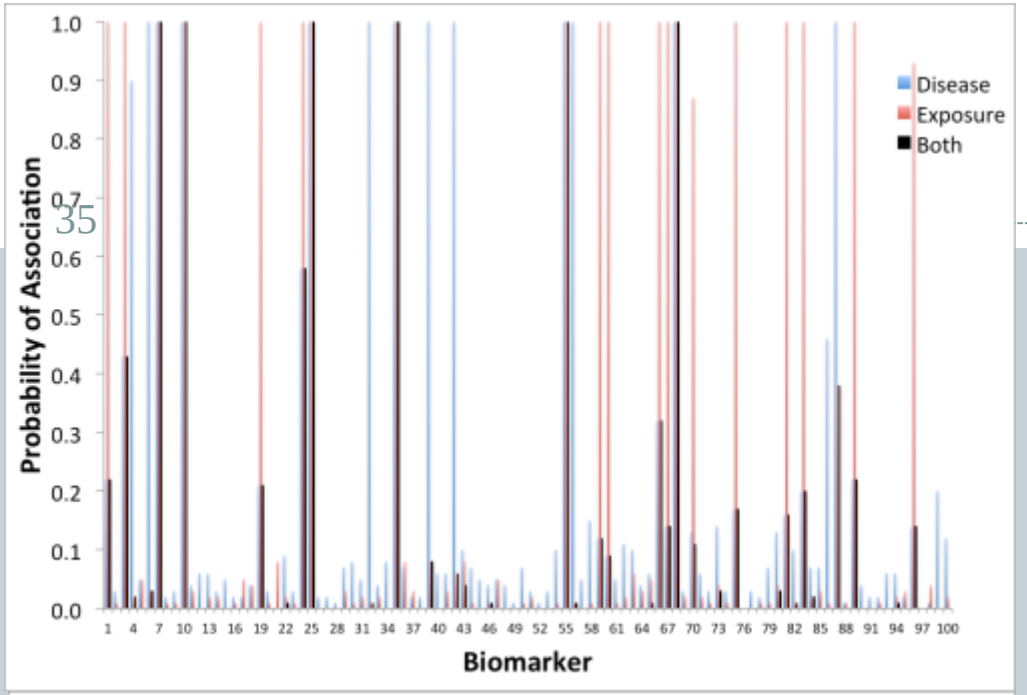
"Meeting in the Middle": A Hierarchical Bayes Approach

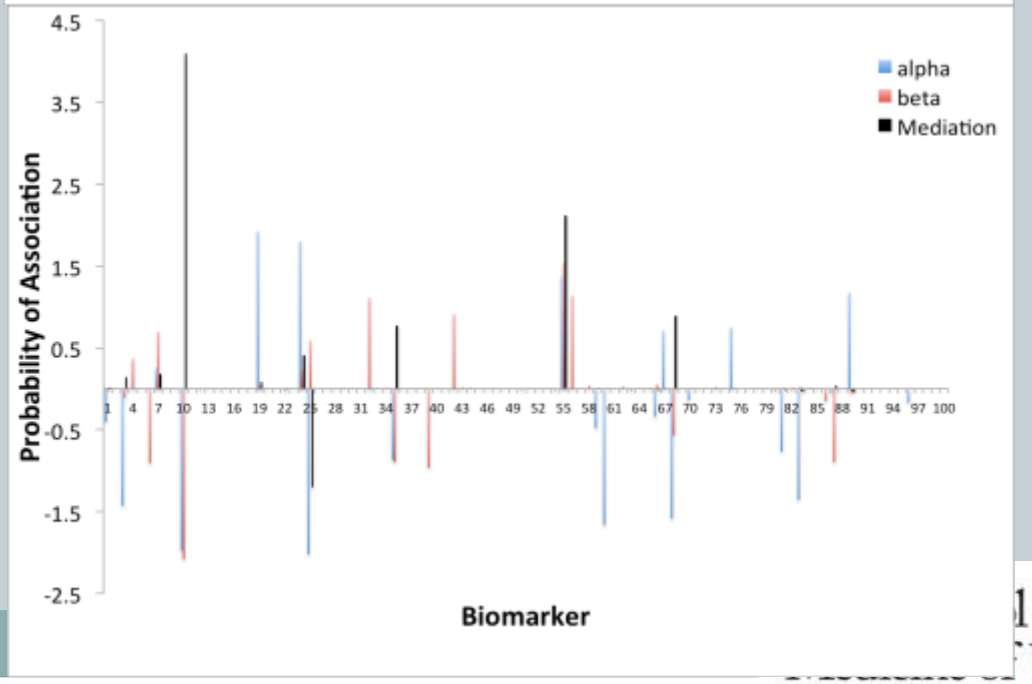
- Simulated data
  - one exposure variable X
  - one disease outcome Z
  - 100 biomarkers: some related to X, some to Y, some to both
  - 1000 subjects
- Fitting by MCMC (WinBUGS)
- Mediation estimates: $\alpha_{1j}\ \beta_{1j}\ \pi_{11j}$



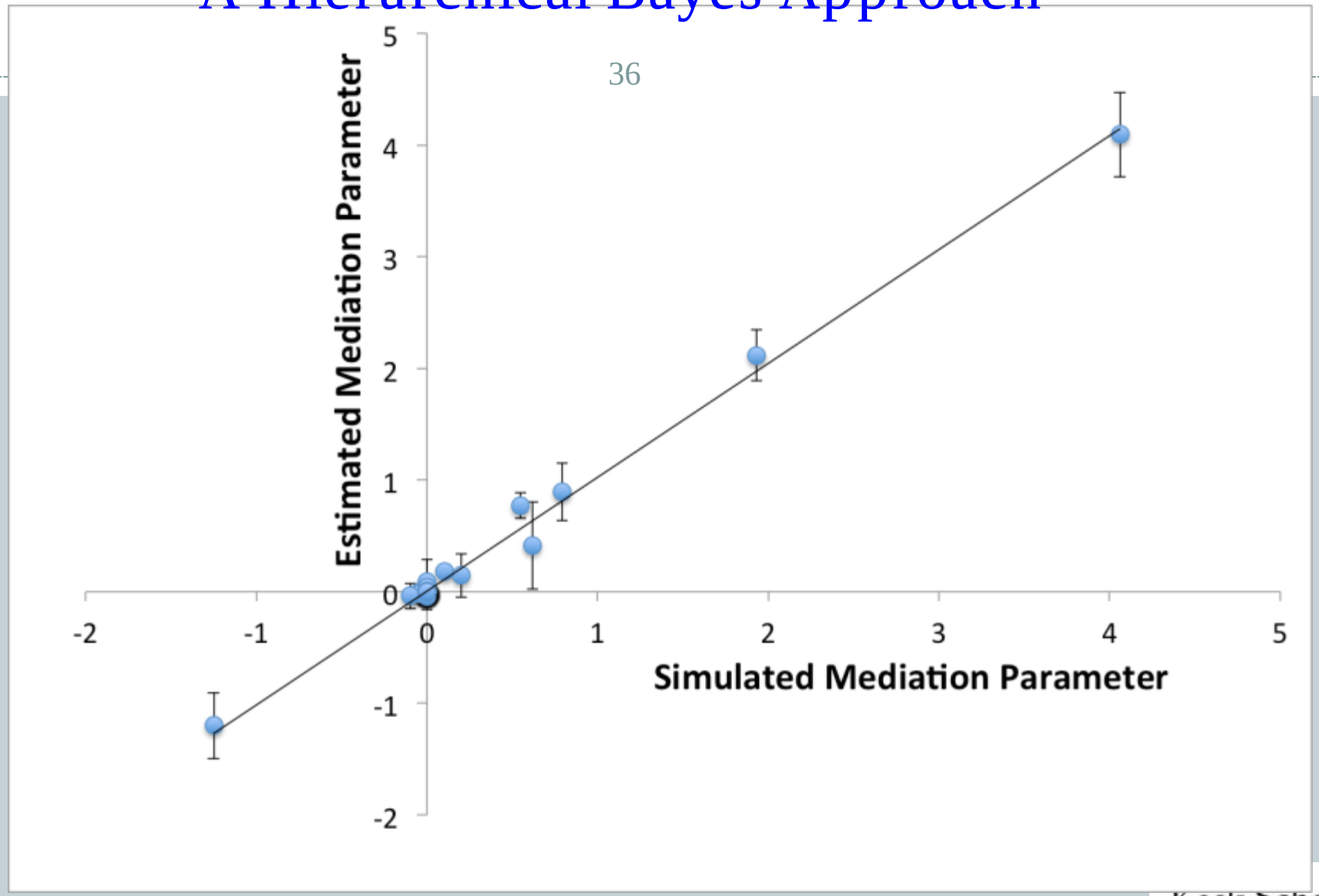
"Meeting in the Middle": A Hierarchical Bayes Approach

- Probabilities of association and MITM
- Separate estimates of association and mediation

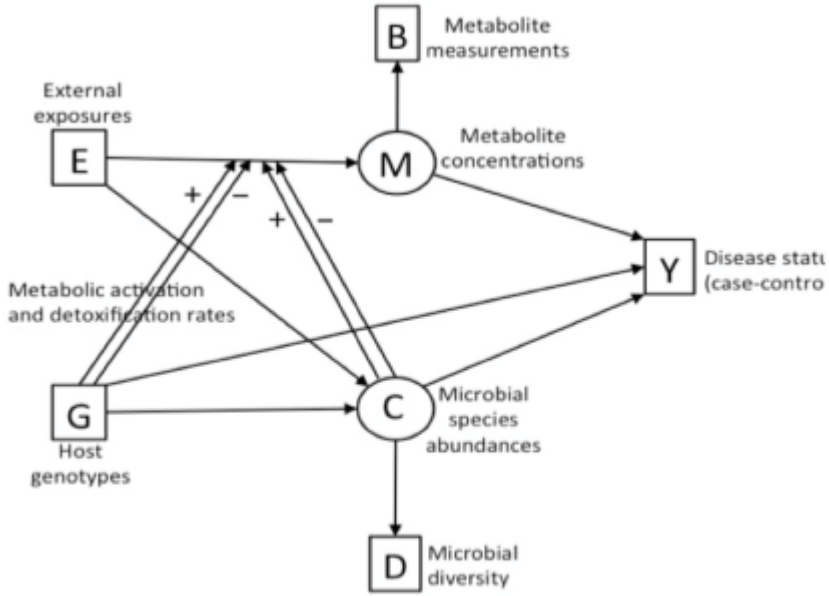




"Meeting in the Middle": A Hierarchical Bayes Approach



Network Analysis Simulation



$\mathbf{E}_i \sim MLN(\mu\mathbf{1},\Sigma)$ where Σ is block diagonal wit

$G_{ig} \sim$ Bernouilli(q) iid, where $q = 0.2$

logit pr($C_{is} > 0$) = E(C_{is} | $C_{is} > 0$) = $\alpha_0 + \alpha_E \Sigma_e E_i$

$D_i = \{ \Sigma_c I(C_{ic} > 0, \Sigma_c C_{ic}, \mathrm{var}(C_{ic} | C_{ic} > 0), \mathrm{Shar}$

$$M_{ie} = \left(\frac{\lambda_e(G_i,C_i)}{\lambda_e(G_i,C_i) + \mu_e(G_i,C_i)} \right) E_{ie} \quad \text{where} \quad \begin{cases} \lambda_e(G,C) = \exp(\\ \mu_e(G,C) = \exp(\end{cases}$$

$B_{ie} \sim LN(M_{ie},\omega^2) I(M_{ie} > \tau)$

logit pr($Y_i = 1$) = $\beta_0 + \beta_E \Sigma_e E_{ie} I_{eY} + \beta_G \Sigma_g G_{ie} I_{gY} +$

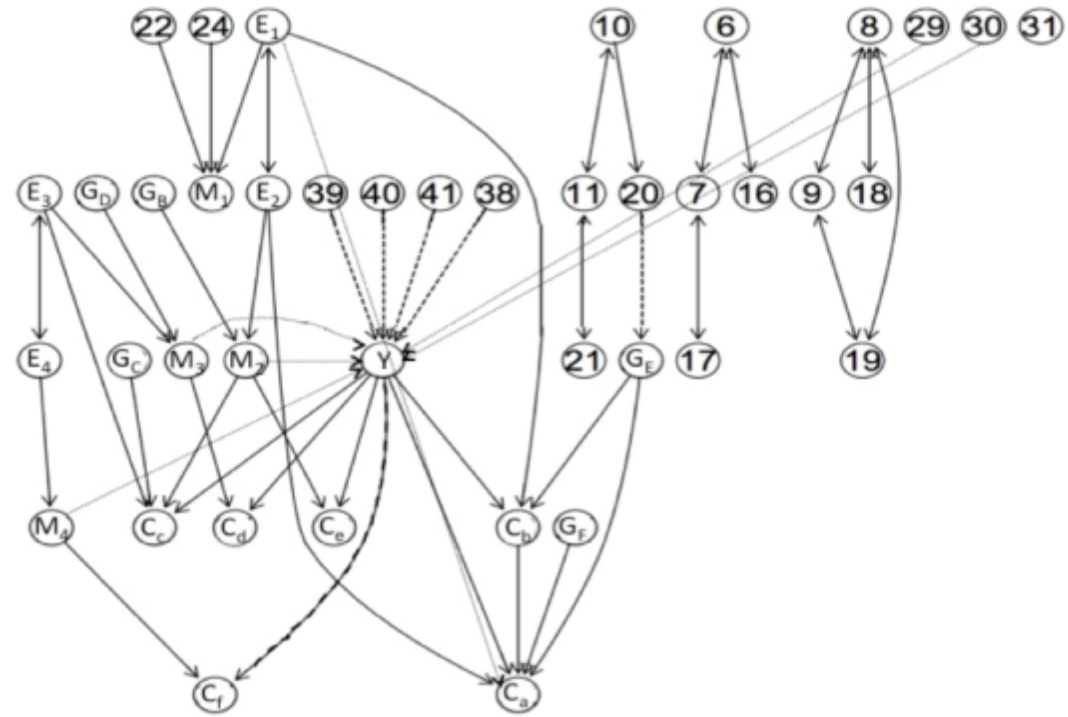


Figure 3: Example of a Bayes network reconstruction based on 42 of the nodes, 18 being causal (E=exposures, M=metabolites, G= genes, C=microbial species, Y=disease; numbered nodes are noncausal variables in the simulated dataset; dashed lines are edges identified by pcalg not in the simulated model, dotted lines are causal relations not identified by pcalg)

Environmental Epigenetics

- Not part of P01 application, but a related interest

Hum Genet
DOI 10.1007/s00439-012-1189-8

REVIEW PAPER

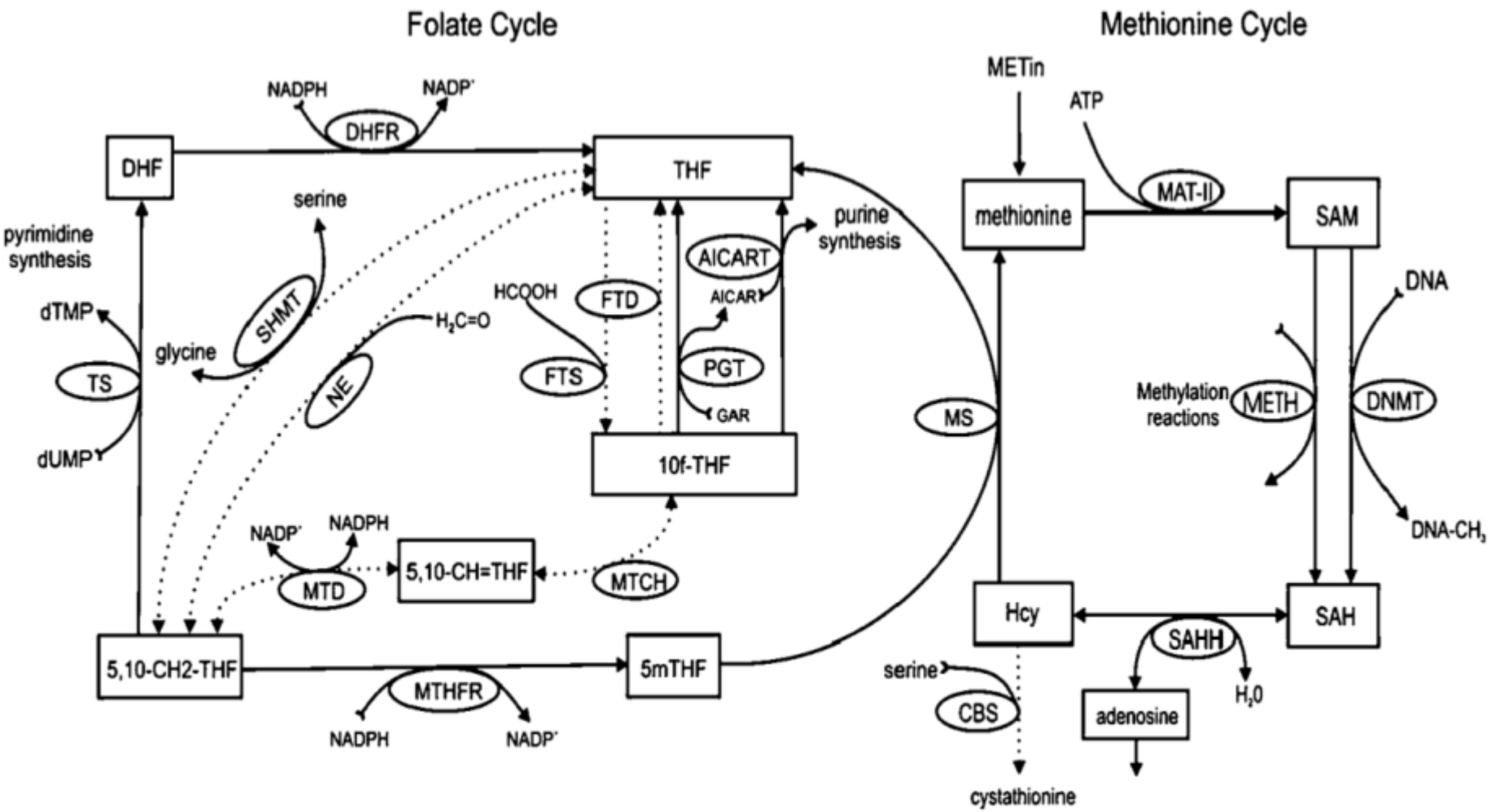
Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships

Victoria K. Cortessis · Duncan C. Thomas · A. Joan Levine · Carrie V. Breton · Thomas M. Mack · Kimberly D. Siegmund · Robert W. Haile · Peter W. Laird

Hum Genet 2012;**131**(10): 1565-89

- Statistical methods in:

Thomas DC (2015). A statistical framework for mediation in environmental epigenetics. In: *Complex Phenotypes: Methodological Issues in Gene-Environment Interactions.* M. Windle (Ed.). Boston, MA, MIT Press**:** in press.



Ulrich et al., *CEBP* 2008: 17:1822-31

Thomas et al., *Hum Genom* 2009;4:21-42

Folate metabolites in the tissue compartment:

$$\mathbf{X} = (X_1,\ldots,X_5) \sim MVN(\alpha_0 + \alpha_1 A + \alpha_2 G, \Sigma_G)$$

where A = alcohol
G = *MTHFR* genotype (A677T coded 0,1,2)

X_1 = total folate
X_2 = homocysteine
X_3 = *TS* activity
X_4 = methionine
X_5 = methylation rate $\equiv E$

Biomarkers in blood:

M_1 = 5mTHF
M_2 = homocysteine
M_3 = SAM
M_4 = SAH
M_5 = methionine

Epigenetic markers: $\mathbf{B} = (B_1,B_2,B_3) \sim MVN(\eta_0 + \eta_1 E, \Omega),\ b = 1,2,3$

B_1 = SAM/SAH (plasma)
B_2 = *DNMT* activity
B_3 = *GNMT* activity

Disease model:

$$\text{logit Pr}(Y=1) = \beta_0 + \beta_1 A + \beta_2 X_1 + \beta_3 A \times X_1 + \beta_4 X_2 + \beta_5 E + \beta_6 X_4 + \beta_7 G$$

node	mean	sd	I
SD.F	0.308	0.02633	
SD.H	0.1798	0.2058	
SD.M	0.1269	0.0192	
SD.T	0.181	0.008571	
alpha.0[1]	-3.527	0.05114	
alpha.0[2]	1.61	0.1117	
alpha.0[3]	2.467	0.02495	
alpha.0[4]	3.857	0.03901	
alpha.0[5]	0.9404	0.07265	
alpha.A[1]	0.05835	0.02137	
alpha.A[2]	0.7648	0.1673	
alpha.A[3]	-1.04E-04	0.01098	
alpha.A[4]	0.0733	0.01709	
alpha.A[5]	0.07265	0.02485	
alpha.G[1]	0.01554	0.02442	
alpha.G[2]	0.06866	0.2422	
alpha.G[3]	-0.003799	0.01564	
alpha.G[4]	-0.003796	0.01425	
alpha.G[5]	0.1012	0.03266	
beta[1]	1.798	0.4151	
beta[2]	0.7447	0.2081	
beta[3]	0.08874	0.09105	
beta[4]	0.03755	0.05627	
beta[5]	-0.6822	0.5047	
beta[6]	0.2719	0.4738	
beta[7]	0.8416	0.1666	
beta[8]	-0.09862	0.08803	
deviance	-7984	582	
eta.0[1]	1.25	0.04622	
eta.0[2]	3.448	0.009361	
eta.0[3]	3.464	0.008401	
eta.E[1]	-0.8031	0.1111	
eta.E[2]	-2.846	0.1719	
eta.E[3]	-2.794	0.14	

The Team

Statistical Investigators

- Duncan Thomas
- Jim Gauderman
- David Conti
- Juan Pablo Lewinger
- Paul Thomas
- Huaiyu Mi
- Josh Millstein
- Daniel Stram
- Kim Siegmund
- Paul Marjoram
- Gary Chen
- Kai Wang

Epidemiologic Investigators

- Graham Casey
- Wendy Cozen
- Steve Gruber
- Darryl Shibata
- John Heidelberg

Consultants & Data Providers:

- Loic Le Marchand
- Cornelia Ulrich
- Urike Peters
- Christina Curtis
- Sylvia Richardson
- Paul Newcombe

Setting & Research Environment: Genetic Epidemiology of Cancer

- Biostatistics and Bioinformatics Divisions of the Department of Preventive Medicine (90 faculty w great breadth in G & E epidemiology)
- Members of the Genetic Epidemiology Program of the USC/Norris Comprehensive Cancer Center (DCT co-directs) & other programs
- History of contributions to study design and statistical analysis methods for genetic and environmental epidemiology
- Longstanding collaborations with cancer genetic epidemiologists, molecular biologists, clinicians
- Rich data resources:
  - GAME-ON Consortium (c0lon, prostate, breast, ovary)
  - Multi-ethnic cohort study, Colon CFR, CA twin & teachers cohorts, …
  - Gene Ontology project
- Highly interactive environment and training w/in division & department
  - Weekly seminars
  - T32 training grant in "Environmental Genomics"

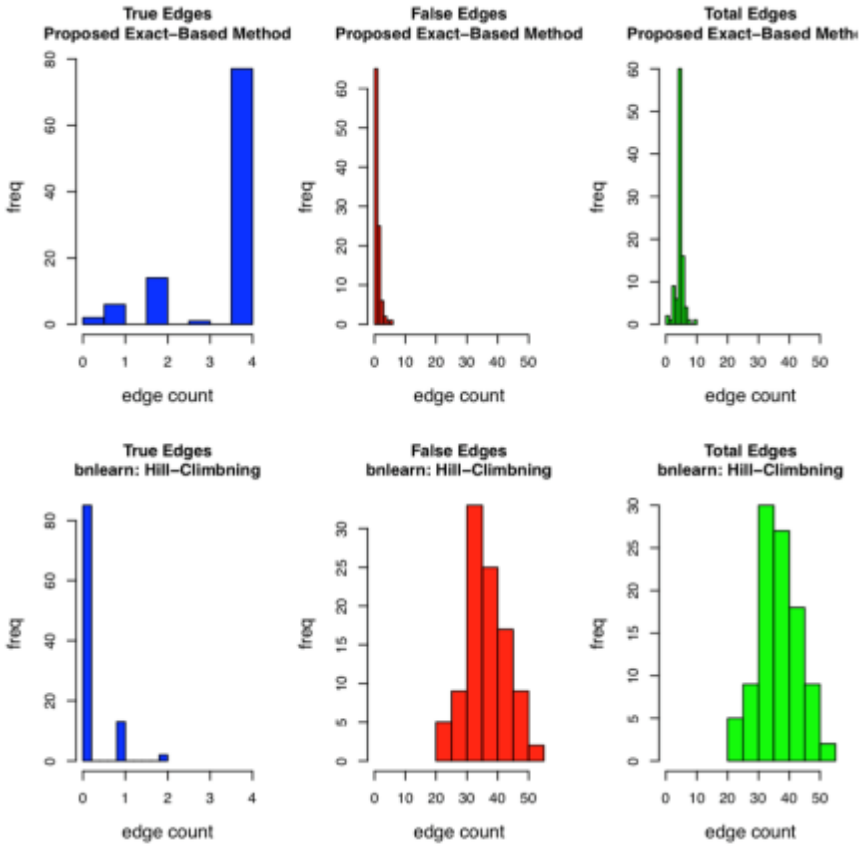


Figure 4. Comparison of the novel proposed exact-based method against the more standard Hill-Climbing approach as implemented in the R package bnlearn in 100 simulated datasets, each with 50 variables and 200 subjects. The true underlying network includes 4 edges connecting 5 nodes. Comparison of upper vs lower graphs shows that the proposed method is more accurate and more parsimonious.

Specific Aims

1. Develop a computational method for functional inference, based on automation of manual Gene Ontology evolutionary annotation process
2. Iteratively improve method on a human-curated training set (75 families), and test against an independent test set (75 families) to avoid overfitting
3. Compare (and validate where possible) hierarchical modeling results (Project 1) with these expanded annotations set to original set

Areas of Current (and Future) Application

- *Primary* focus of examples will be colorectal cancer
  - CORECT & GECCO
  - C-CFR
  - MEC
  - GWASeq
  - Epigenome Center & TCGA
  - ColoCare Cohort
- Other cancers: esp. prostate, breast, ovary ...
  - Involved in 4/5 GAME-ON projects
  - And numerous other case-control & cohort studies
- Non-cancer (asthma, diabetes, ...)

Cores

A. Administration (Gauderman, Thomas, Casey)

B. Genomic Annotation (Mi)
   - Compile protein-coding and non-coding regulatory annotations
   - Distribute in multiple formats

C. HPC and simulation (Chen)
   - Develop hardware & software solutions for HPC needs
   - Generate simulation shells

D. Data management and software development (Gauderman)
   - Merge source data and create analysis-ready files
   - Develop and distribute analysis software

External Advisory Committee

- Expertise
  - Statistical genetics
  - Cancer epidemiology
  - Molecular biology
  - Computation
  - Bioinformatics
  - Exposure assessment
- EAC roles:
  - Attend one annual meeting in Los Angeles
  - Be available to review progress, manuscripts, reports, etc.

Members:

Dan Schaid (Mayo)
Eric Schadt (Mt. Sinai)
Matt Freedman (Harvard)
Riki Peters (FHCRC)
Gary Miller (Emory)
Ken Lange (UCLA)
Gerry Coetzee (USC)
Mark Suchard (UCLA)

Expertise of the Team

KEY PERSONNEL:

- Biostatistical methods in epidemiology
- Statistical genetics
- Bioinformatics
- High-performance computing
- Software development

CLOSE COLLABORATORS:

- Cancer genetic epidemiology colleagues
- Environmental epidemiology colleagues
- Molecular biology colleagues

Integration: Examples related to CRC

1. How can we use gene ontologies and functional annotations to link GWAS hits to potential pathways? (P1, P5, C2-C4)
2. How do host factors (via the exposome and microbiome) modify the effects of mutations on CRC risk? (P1, P3, P4, C2-C4)
3. How do germline and somatic mutations jointly affect initiation and growth of CRC tumors? (P2, P5, C2-C4)
4. How can we improve current CRC risk prediction models using selected 'omic' data? (P1, P5, P6, C2-C4)
5. How can models of integrative genomic data be used to improve our understanding of cancer biology and risk? (P1-P6, C2-C4)

Why a P01?

- Strong integration of projects toward common goals
- Strong cores that enhance focus and synergy among projects and translation of methods to the outside world
- Study team with proven track record of collaboration leading to important methods and applications
- Tackling integrative genomics in a comprehensive way is essential, but is 'too big' for a single RO1

Table 1: MCMC parameter estimates for 3 gene families: the Caspase family and two versions of the Cholinesterase family. The cholinesterase family was pruned to generate the "subset", allowing us to see whether the results were robust to the number of genes in the tree and the number of annotated functions.

Gene family:	Caspase	Cholinesterase subset	Cholinesterase full
# leaves	92	190	1051
# internal nodes	64	121	519
# functions	20	3	20
Root node probability	0.22 ± 0.06	0.29 ± 0.16	0.15 ± 0.10
Function change probabilities (per 1000 evolutionary events, ± SE)			
Speciation nodes :			
Gain of function	1.1 ± 0.7	0.0 ± 0.0	3.8 ± 2.1
Loss of function	2.1 ± 0.5	3.1 ± 1.5	3.3 ± 1.1
Duplication nodes:			
Gain of function	0.3 ± 0.4	1.1 ± 0.9	4.2 ± 2.0
Loss of function	0.3 ± 0.2	2.6 ± 0.9	1.8 ± 0.7
Misclassification rates for leaf annotations (%, ± SE)			
False positives	0.4 ± 0.4	0.1 ± 0.4	0.3 ± 0.0
False negatives	0.3 ± 0.5	0.0 ± 0.1	0.6 ± 0.0

Specific Aims

1. To develop methods to partition somatic alterations into
   i. Present in most recent common ancestor (1st transformed) cell
   ii. Occur during tumor growth
2. To develop phylogenetic models that allow us to estimate parameters characterizing cancer growth
3. To develop multi-scale models that correlate host factors (e.g. age, E, M, G) with cell behavior (e.g. mutation frequency, cell motility) in human cancer

Motivation

- Functional knowledge (e.g. Gene Ontology annotations) for human genes is very incomplete
- Increase in association detection power using prior biological knowledge depends strongly on annotation completeness (Preliminary Results)
- Phylogenetic inference of annotations allows vast experimental knowledge in model systems (e.g. mouse, fruit fly, yeast) to augment human gene annotations
  - Challenge: which knowledge can be extrapolated to related human genes, and which cannot?

Critique of Initial Submission (Sept 2014; rev Feb 2015)

Recurring Themes:

- Moderate innovation
- Advantages over currently available ones
- More preliminary studies
- Scalability
- How to evaluate & validate new methods

Minor changes in budget & structure: all projects survived

Integration good, outstanding investigators, etc. …

Motivation

- Cancers are heterogeneous in etiology, and growth
- Textbook model for cancer growth:
  Little tumors grow into larger, more malignant tumors
- Recent large cancer screening study presents mixed results
  Removing small lesions does not always improve survival
  New lesions are not always Stage 1
- Is the model wrong?
  Newer hypothesis: Some cancers are “Born to be Bad!”