Uncovering Latent Relationships in High Dimensional Data with High Performance Parameterized Algorithms: A Smorgasbord of Applications

Mike Langston

Professor
Department of Electrical Engineering and Computer Science
University of Tennessee
USA

8 July 2012

APAC 2012

Outline of Talk

Some Thoughts on What, Really, is an "Application"

A Few Relevant Algorithms and Implementations

A Smorgasbord or Mélange of Applications

Some Ongoing Work

APAC 2012

Outline of Talk

Some Thoughts on What, Really, is an "Application"

A Few Relevant Algorithms and Implementations

A Smorgasbord or Mélange of Applications

Some Ongoing Work

APAC 2012

What, Really, is an Application?

I would propose that there are at least three levels.

1

Computational Scientists	**Complexity Theory**
Reductions & Proofs	**Algorithm Design**
Worst Case Analysis	**No or Synthetic Data**

APAC 2012

What, Really, is an Application?

I would propose that there are at least three levels.

1

Computational Scientists
Reductions & Proofs
Worst Case Analysis

Complexity Theory
Algorithm Design
No or Synthetic Data

3

Domain Scientists
Messy Work
Long Range Planning

Deep Understanding
Experimental Design
Real Data Only

APAC 2012

What, Really, is an Application?

I would propose that there are at least three levels.

1

Computational Scientists	Complexity Theory
Reductions & Proofs	Algorithm Design
Worst Case Analysis	No or Synthetic Data

2

Computational Scientists	Domain Scientists
Design & Implementations	Practical Restrictions
HPC & Scalability	Real & Synthetic Data

3

Domain Scientists	Deep Understanding
Messy Work	Experimental Design
Long Range Planning	Real Data Only

APAC 2012

Outline of Talk

Some Thoughts on What, Really, is an "Application"

A Few Relevant Algorithms and Implementations

A Smorgasbord or Mélange of Applications

Some Ongoing Work

Everyone's Favorite: Vertex Cover

Everyone's Favorite: Vertex Cover

- **Most widely studied FPT problem**
- **Dual to maximum clique**
- **Window to biclique, paraclique, etc**
- **Also to maximal clique and assorted enumerations**
- **Meld notions of FPT and graph coloring**

APAC 2012

A Few Algorithms and Implementations

Everyone's Favorite: Vertex Cover

- **Most widely studied FPT problem**
- **Dual to maximum clique**
- **Window to biclique, paraclique, etc**
- **Also to maximal clique and assorted enumerations**
- **Meld notions of FPT and graph coloring**
- **The damn thing simply will not go away**

Nonblocker aka Dominated Set

A Few Algorithms and Implementations

Nonblocker aka Dominated Set

- Second best known FPT problem?
- Dual to dominating set
- A bit misunderstood but highly useful
- We employ nonplanar red/blue dominating set
- Optimization lags well behind vertex cover

Cluster Edit

APAC 2012

A Few Algorithms and Implementations

Cluster Edit

- **Another handy FPT problem**
- **Peter Damaschke, and several here, have worked on generalizations and restrictions**
- **Frank Dehne, yours truly, others have had students work on implementations**

APAC 2012

A Few Algorithms and Implementations

Cluster Edit

- **Another handy FPT problem**
- **Peter Damaschke, and several here, have worked on generalizations and restrictions**
- **Frank Dehne, yours truly, others have had students work on implementations**
- **Not monotonic**
- **Can lack applications fidelity without tweaks**

APAC 2012

A Few Algorithms and Implementations

Supercomputers: Kraken, an ORNL-UT Cray XT5 system

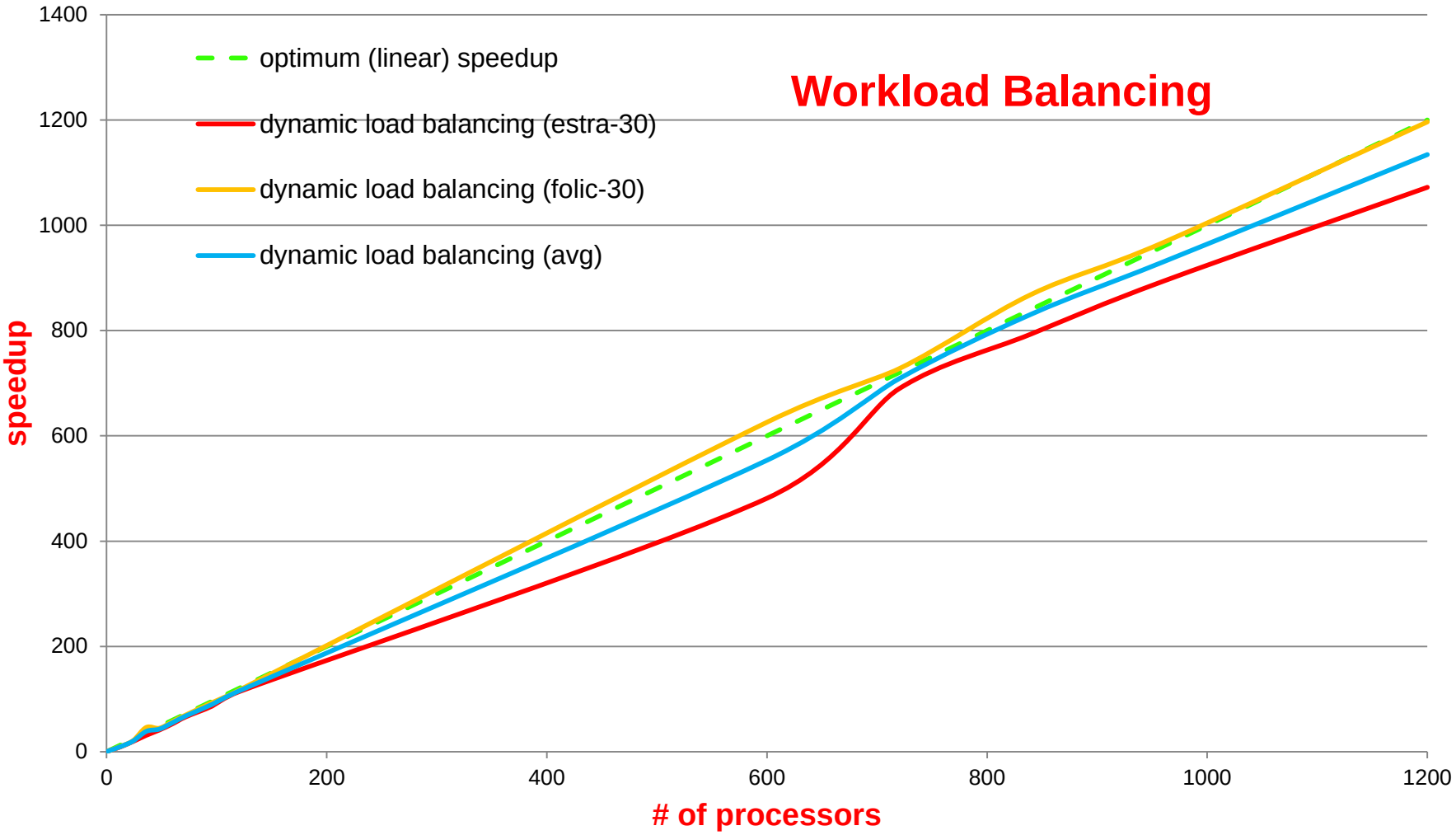
- 10^5 processor cores with distributed memory
- roughly 10^{12} calculations per second (a petaflop)
- until recently the world's most powerful computer for open science
- now moving to Jaguar then Titan

NERSC & XSEDE
(formerly the TeraGrid)

Serendipity:
Keller's Conjecture
Resolved

APAC 2012

A Few Algorithms and Implementations



APAC 2012

Outline of Talk

Some Thoughts on What, Really, is an "Application"

A Few Relevant Algorithms and Implementations

A Smorgasbord or Mélange of Applications

Some Ongoing Work

APAC 2012

Meta-Application: Co-Expression Analysis

Clustering Algorithms Ranked by Quartile Comparisons

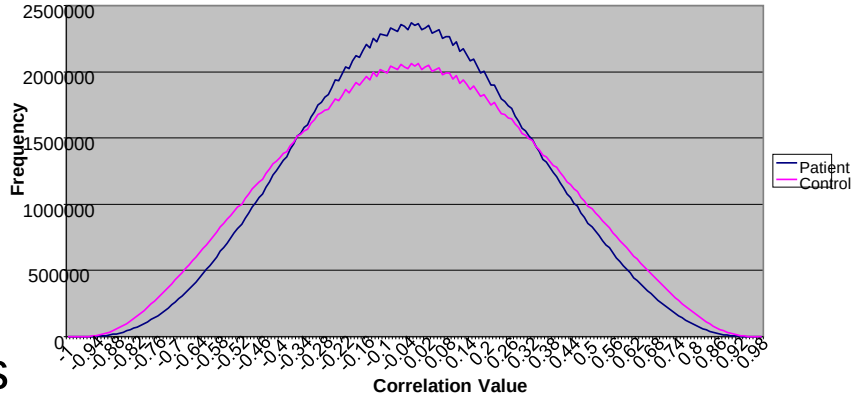
Clustering Method	Average Quartile	Small (3-10 genes)		Medium (11-100 genes)		Large (101-1000 genes)	
		Quartile	BAT5 Jaccard	Quartile	BAT5 Jaccard	Quartile	BAT5 Jaccard
K-Clique Communities	**1.00**	**1**	**0.7531**	**1**	**0.4465**	**1**	**0.4915**
Maximal Clique	**1.00**	**1**	**0.8433**	**1**	**0.4081**		**0.0000**
Paraclique	**1.00**	**1**	**0.7576**	**1**	**0.4285**	**1**	**0.4169**
Ward (H)	1.33	2	0.5782	1	0.4011	1	0.5723
CAST	**1.67**	**1**	**0.7455**	**3**	**0.3146**	**1**	**0.4994**
QT Clust	2.00	2	0.5473	2	0.3670	2	0.3944
Complete (H)	2.33	3	0.3933	2	0.3677	2	0.3419
NNN	2.67	2	0.5521	2	0.3705	4	0.2406
K-Means	3.00	4	0.2573	3	0.3015	2	0.3463
SOM	3.00	4	0.3260	2	0.3286	3	0.3282
WGCNA	3.00	3	0.4391	3	0.3106	3	0.2949
Average (H)	3.33	3	0.4087	4	0.2792	3	0.3037
McQuitty (H)	3.33	3	0.4594	3	0.3065	4	0.2868
SAMBA	3.50		0.0000	4	0.1860	3	0.3298
CLICK	4.00	4	0.0339	4	0.1453	4	0.2817

APAC 2012

Application: Allergic Rhinitis

Data Description

- Mikael Benson, Göteborg, Sweden, 56 patients and 39 controls
- Affymetrix HU133 arrays
- roughly 33,000 genes (~1B interactions)
- nasal secretions, lymphocytes, skin
- hay fever, eczema



Preprocessing

- MAS5.0
- log transformed
- replicates averaged
- centered around zero with *z* scores
- probesets with consistently low expression levels removed

Threshold Selection

- spectral graph theory used to set and balance graph densities
- AFFX spots retained for quality control

APAC 2012

Application: Allergic Rhinitis

Clique profiles using the five most highly represented genes:

Control		Patient	
Gene Symbol	Clique membership	Gene Symbol	Clique membership
UBE1C	29%	*FGFR2*	66%
RANBP6	27%	*NFIB*	65%
DKFZP564O123	26%	*PPL*	64%
SLC25A13	24%	*FGFR3*	64%
GTPBP4	21%	*CDH3*	56%

ribosomal or RNA-related **T-lymphocytes or epithelial cells**

Applied differential screens, then chromatin immunoprecipitation

Sample Result: Discovered a novel and key role for *ITK* (IL2-inducible T-cell kinase)

APAC 2012

Application: Prostate Cancer

Data Inhomogeniety

- Pablo Moscato, Uni Newcastle, Australia
- huge problem without model organisms
- no recombinant inbred human populations
- tumors and other diseases are often not uniform

Creative Use of Graph Algorithms

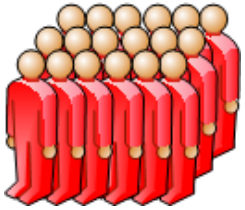
- perform multiple data views
- drive correlations with both persons and genes
- exclude outliers and separate subtypes
- perform differential analysis to distill gene differences

APAC 2012

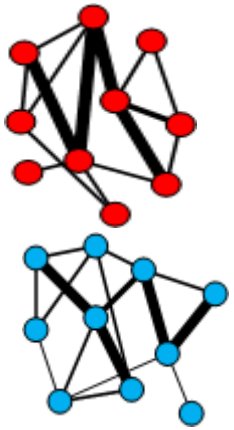
Application: Prostate Cancer

Case

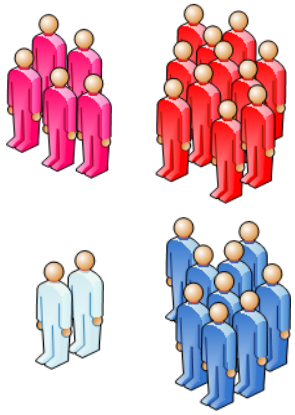
Control



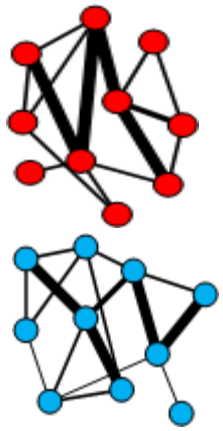
Genes Drive Person-Person Correlations



Select Thresholds, Extract Cliques

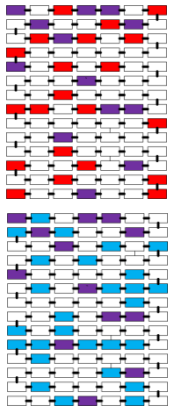


Classify Subtypes and Eliminate Outliers



Persons Drive Gene-Gene Correlations

Select Thresholds, Extract Cliques



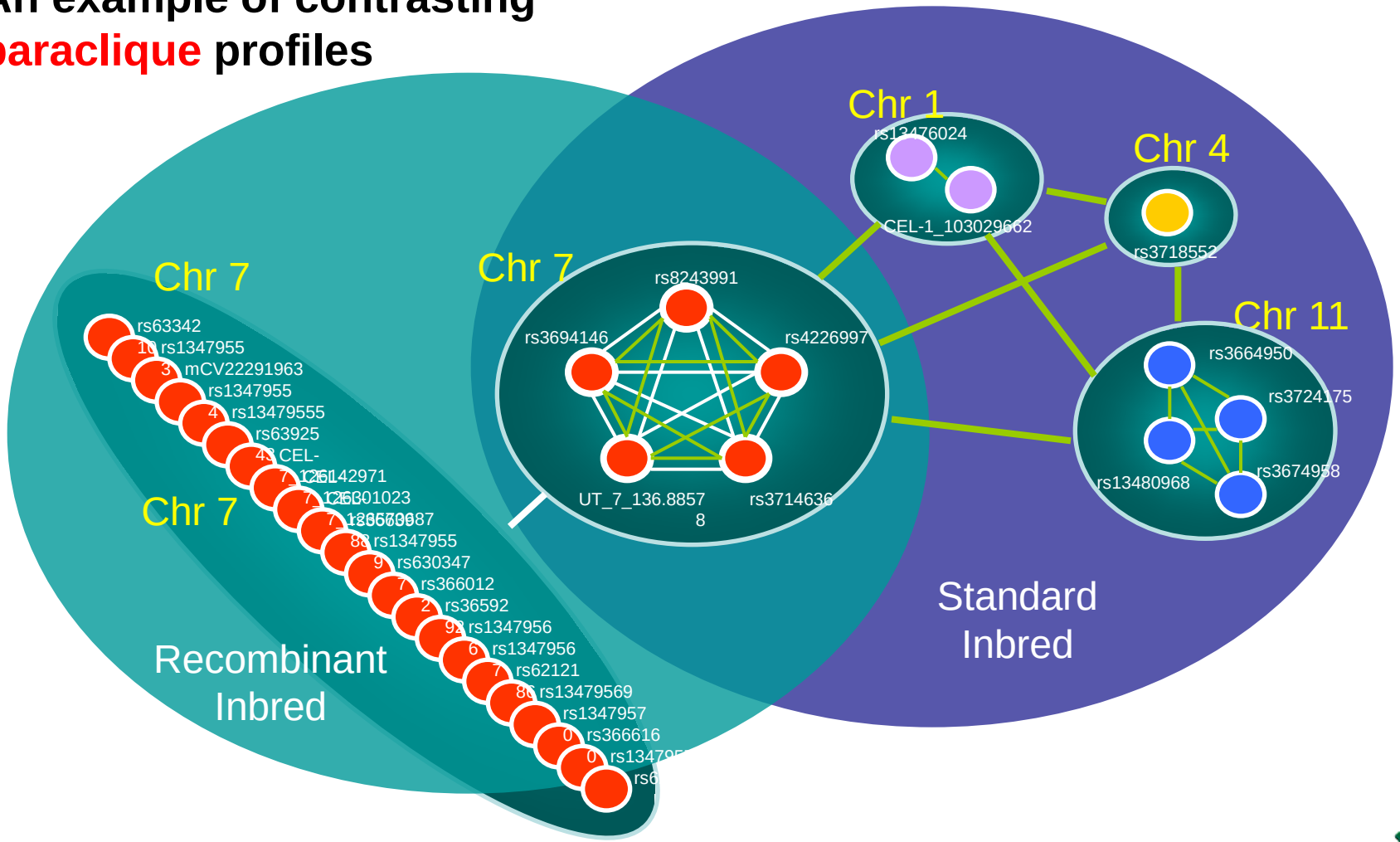
Perform Assorted Forms of Differential Analysis to Identify Network Differences

Sample Result: Putative Prostate Cancer Biomarkers

KLK3 = PSA
ETS1
MAZR
KROX
NFKB

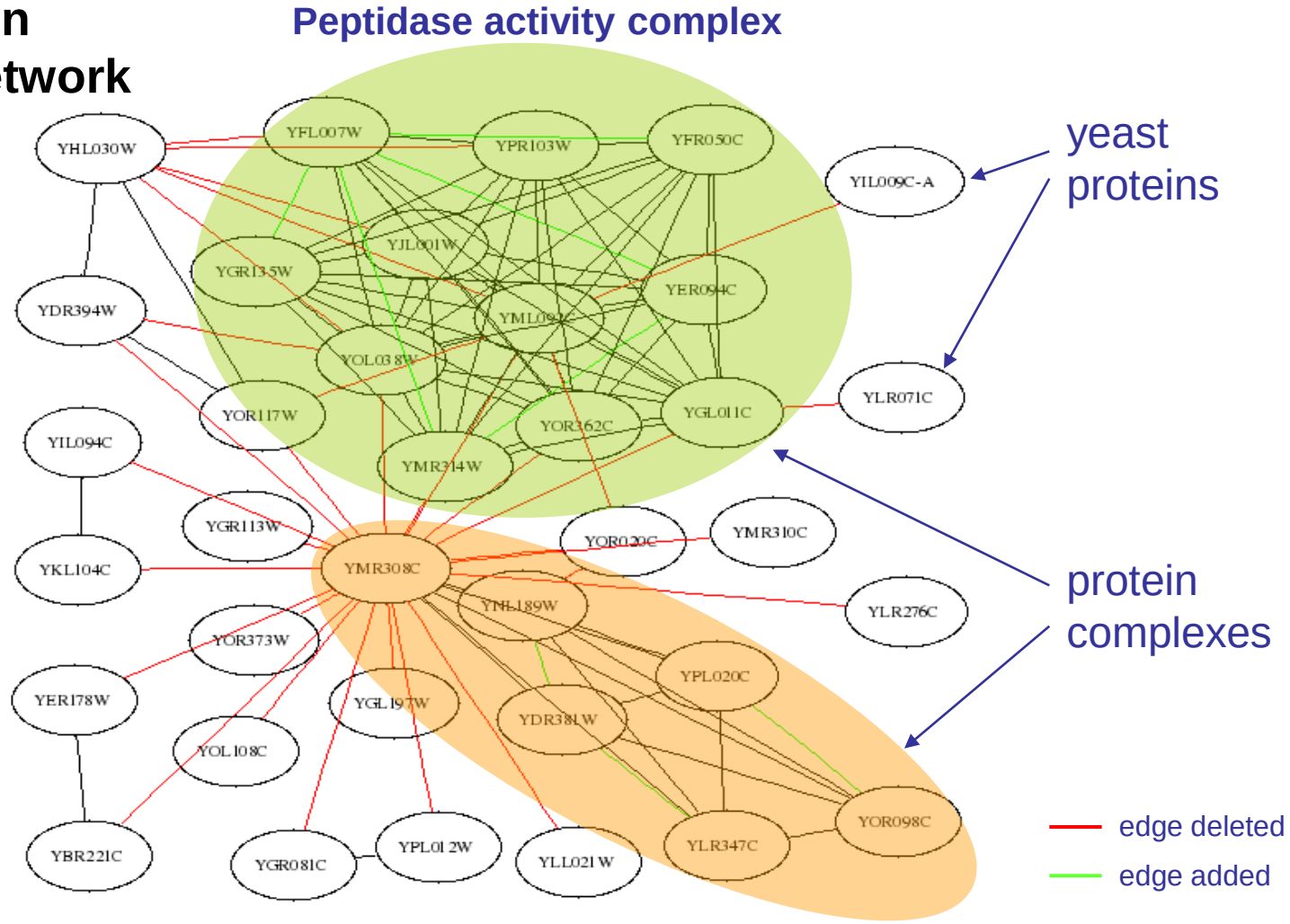
Application, Linkage Disequilibrium

An example of contrasting paraclique profiles



Application, Protein Complex Prediction

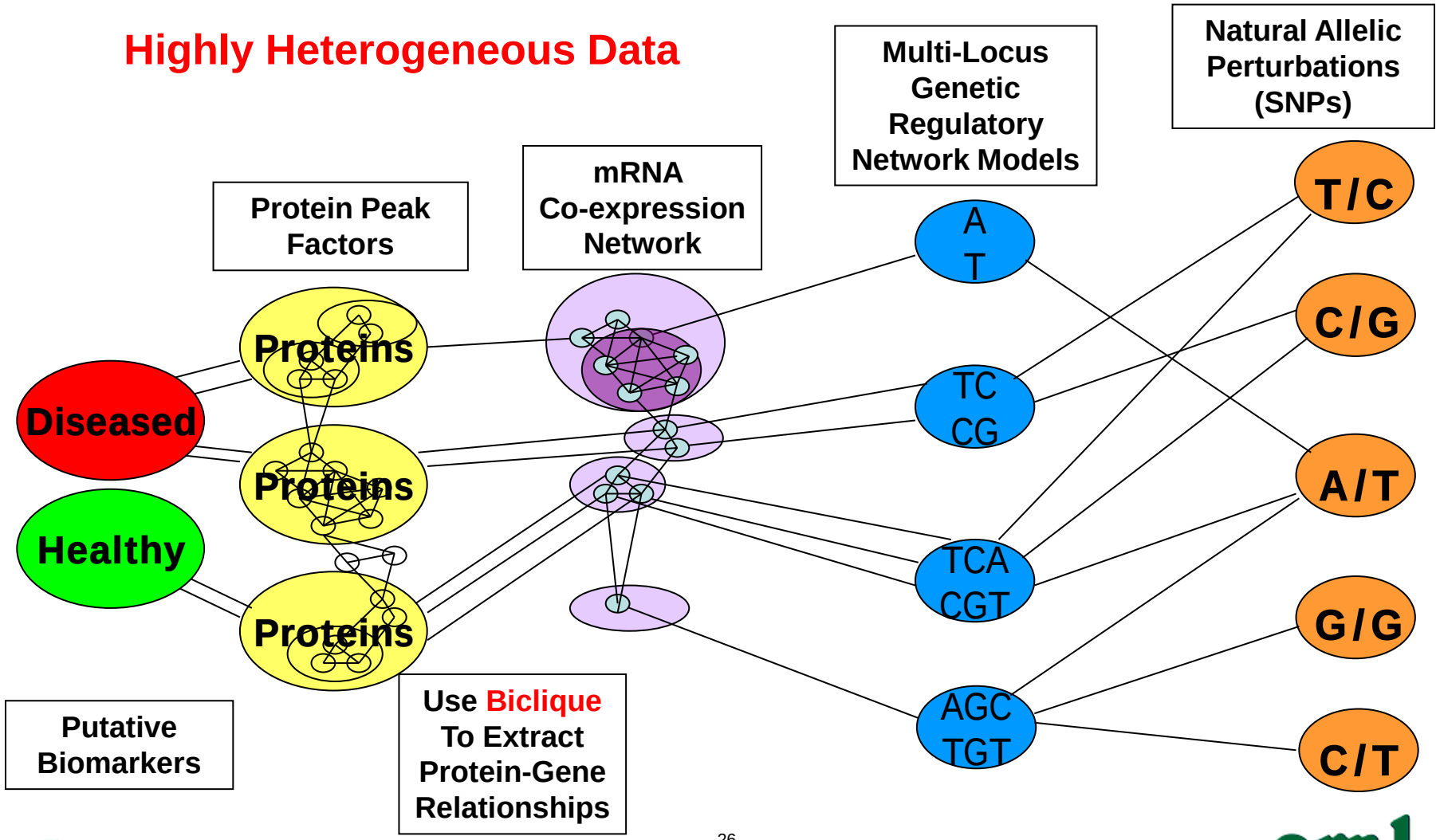
Protein-Protein Interaction Network



Recognize as Cluster Edit

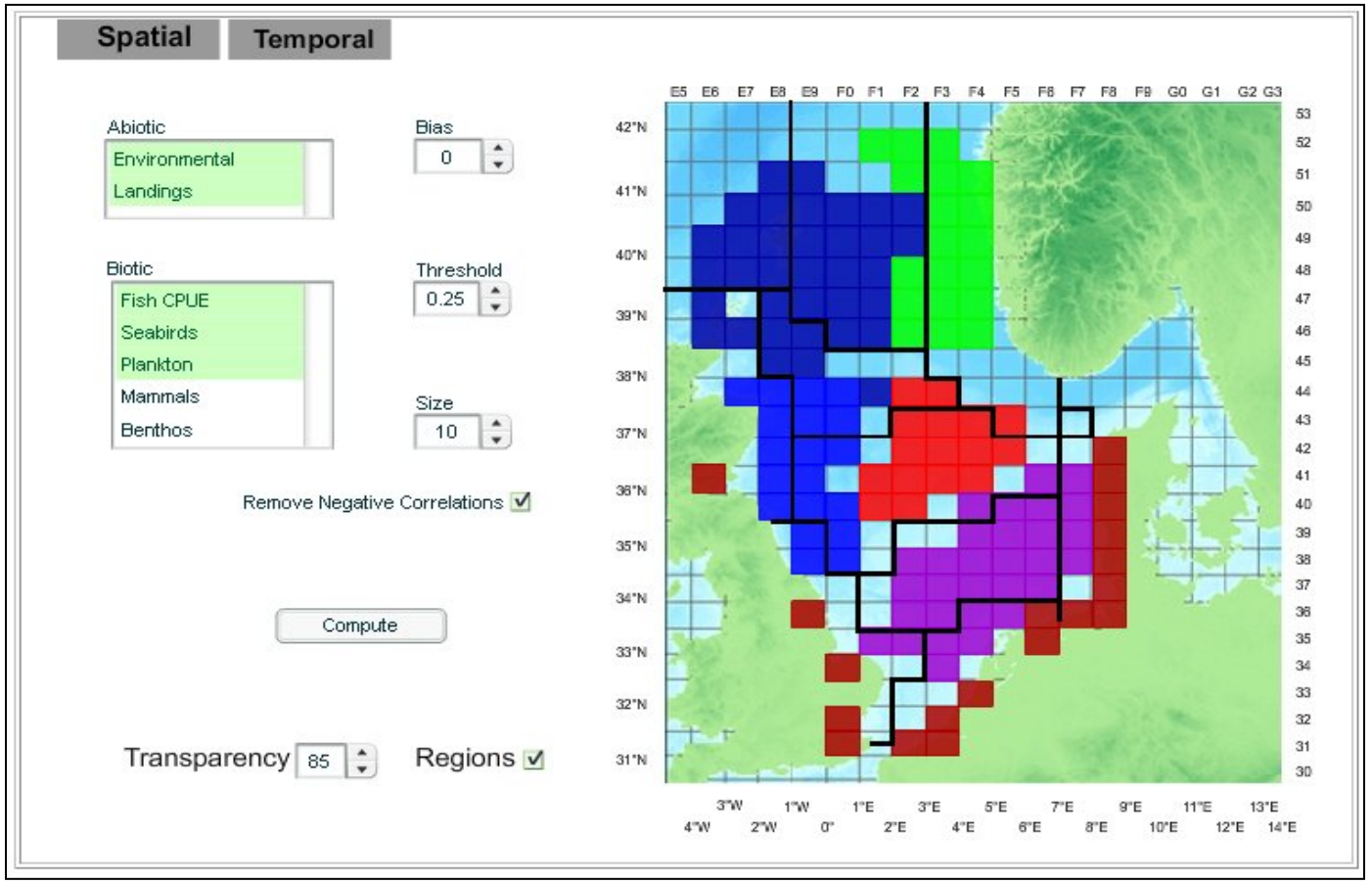
APAC 2012

Application, Data Integration

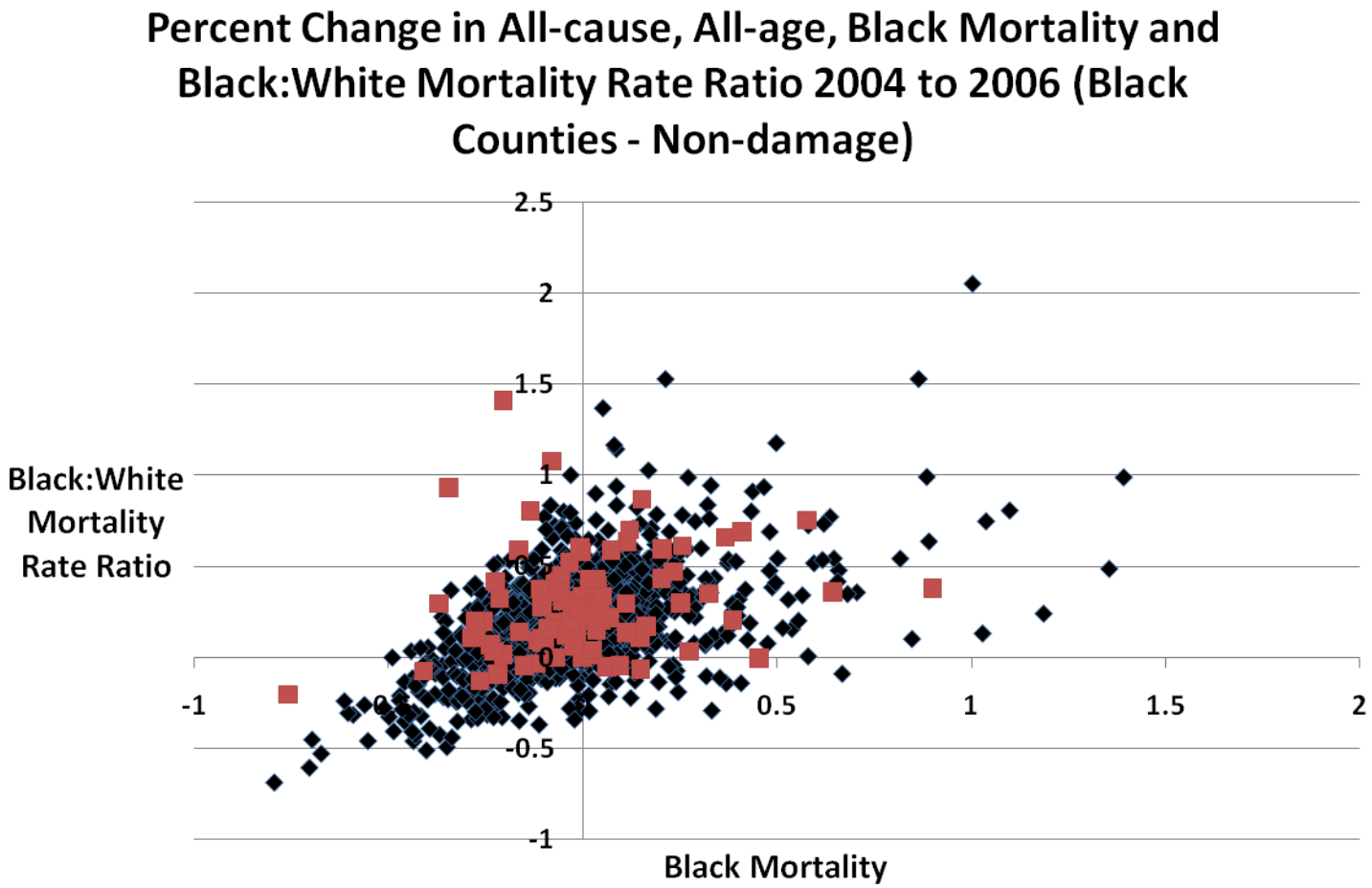


Diverse Sets of Oceanographic Data

Spatial analysis via paraclique supports the North Sea flushing rate



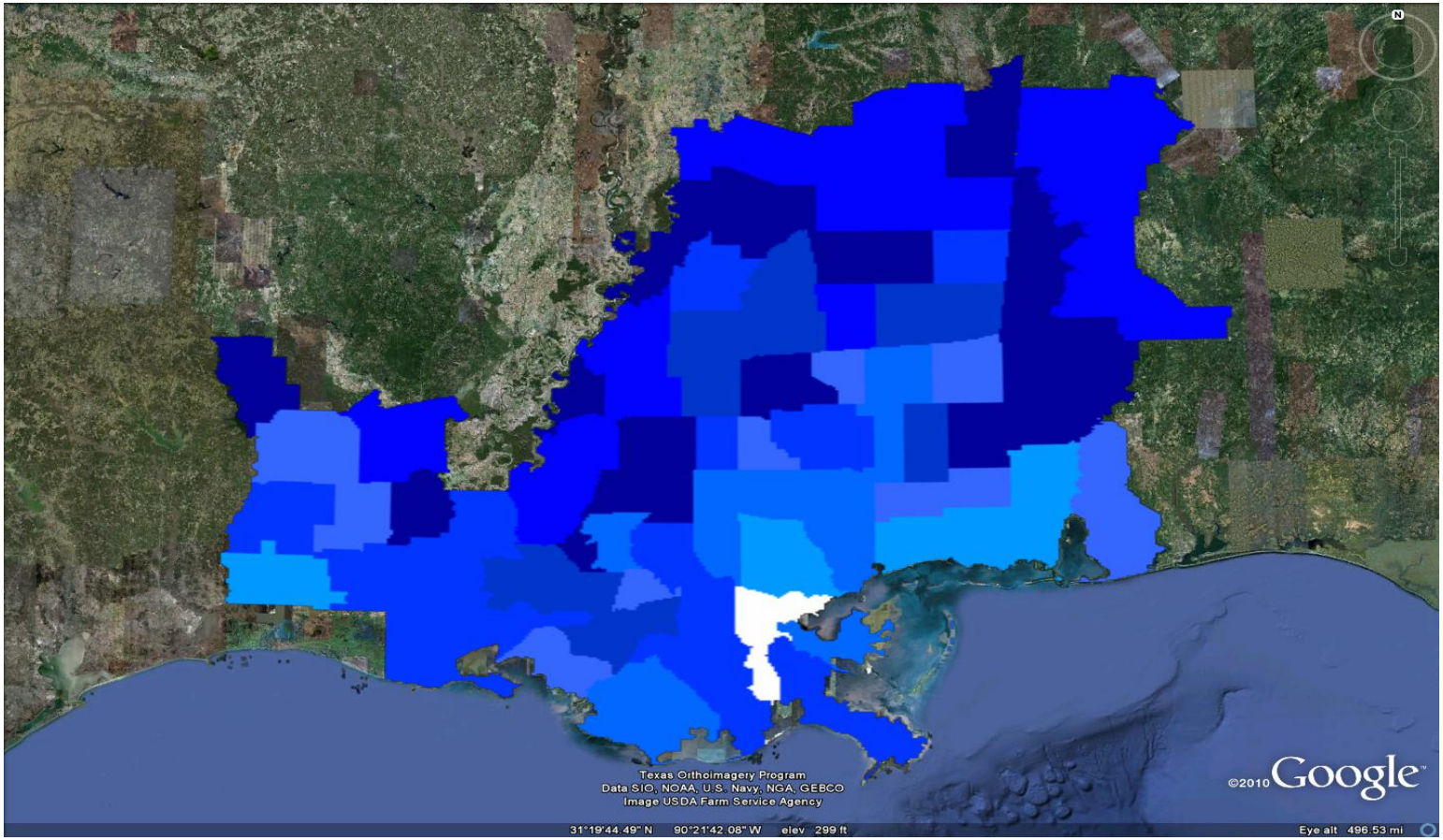
Sample Gulf Coast Analysis: Mortality Quadrants circa 2005



APAC 2012

Application, Social Disparities

Hurricane Katrina's impact on minority populations, as measured by mortality, morbidity and displacement

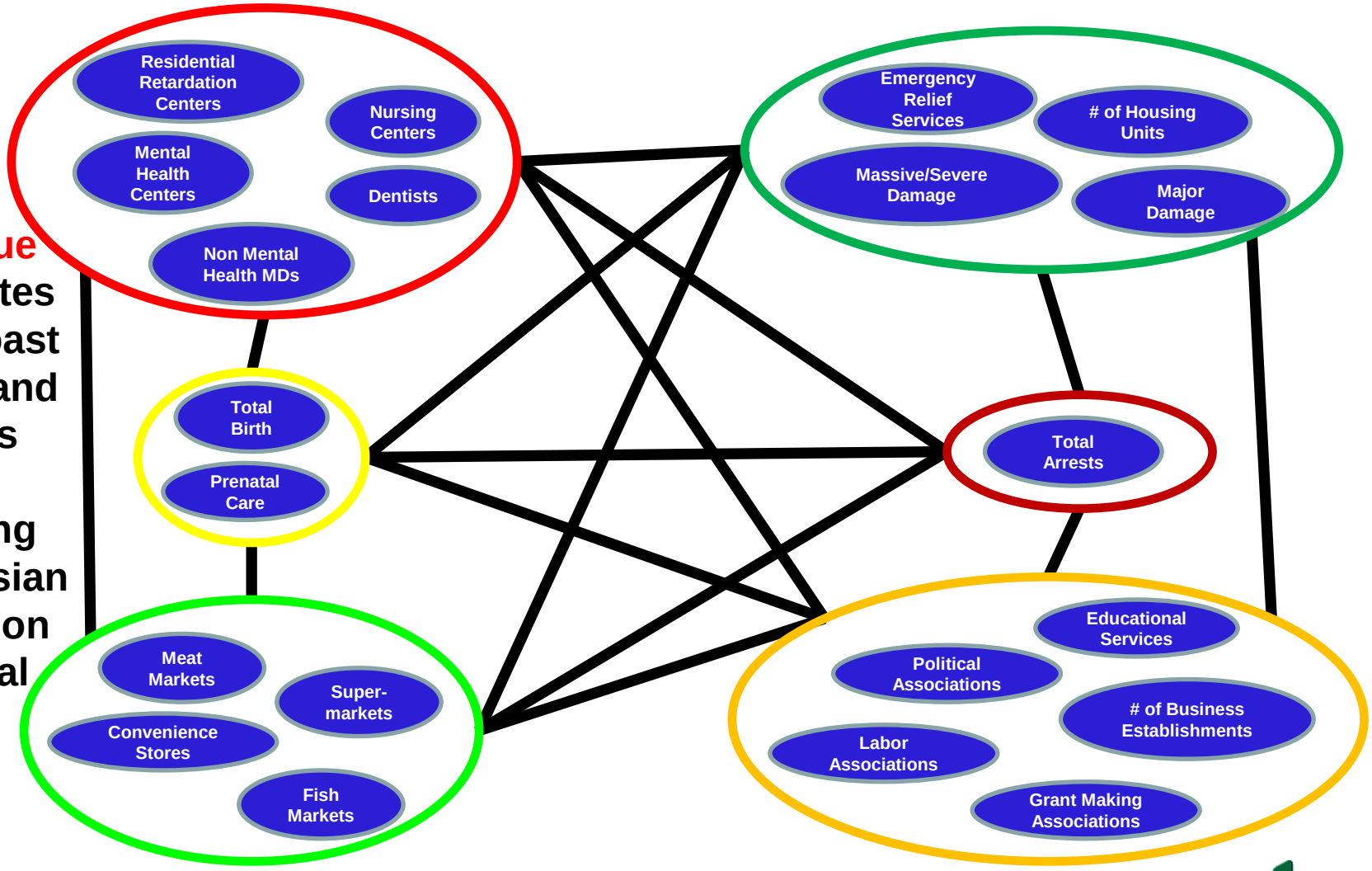


APAC 2012

Application, Social Disparities

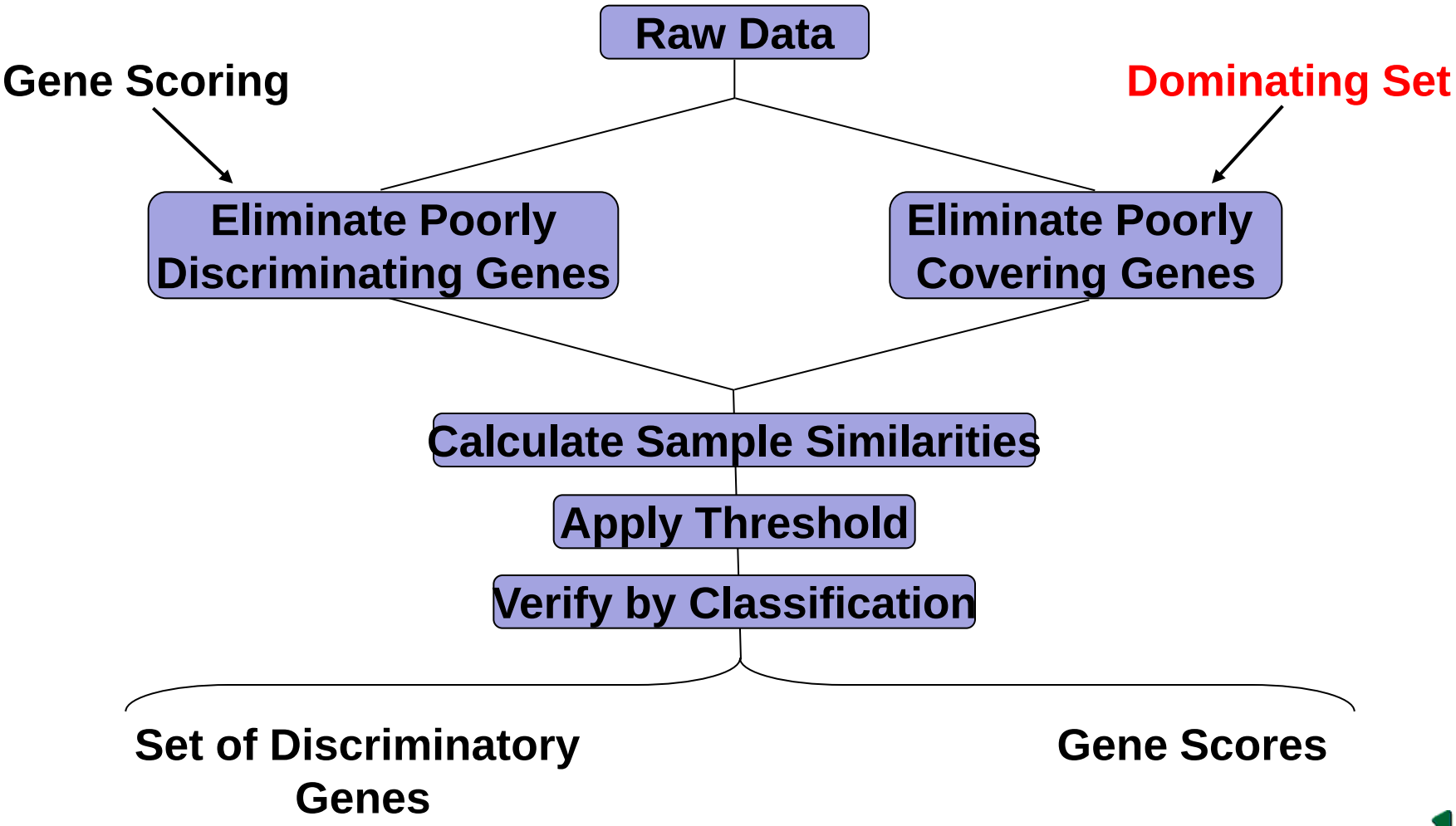
Paraclique of correlates in Gulf Coast counties and parishes

Now using with Bayesian and SEM on a national level



APAC 2012

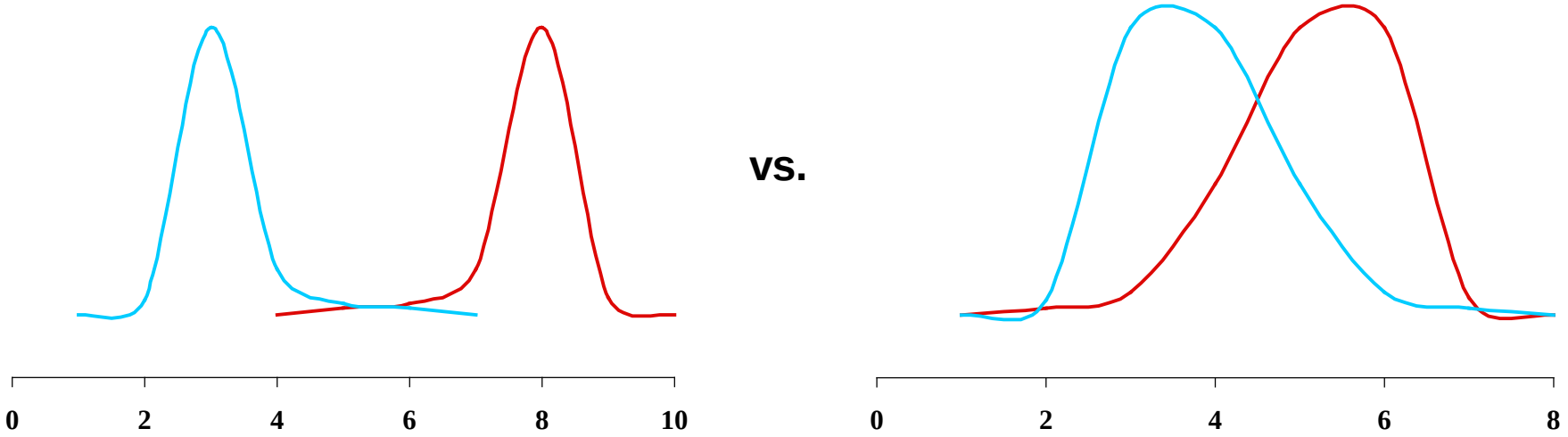
Application, Biomarker Discovery



APAC 2012

Application, Biomarker Discovery

Gene Scoring

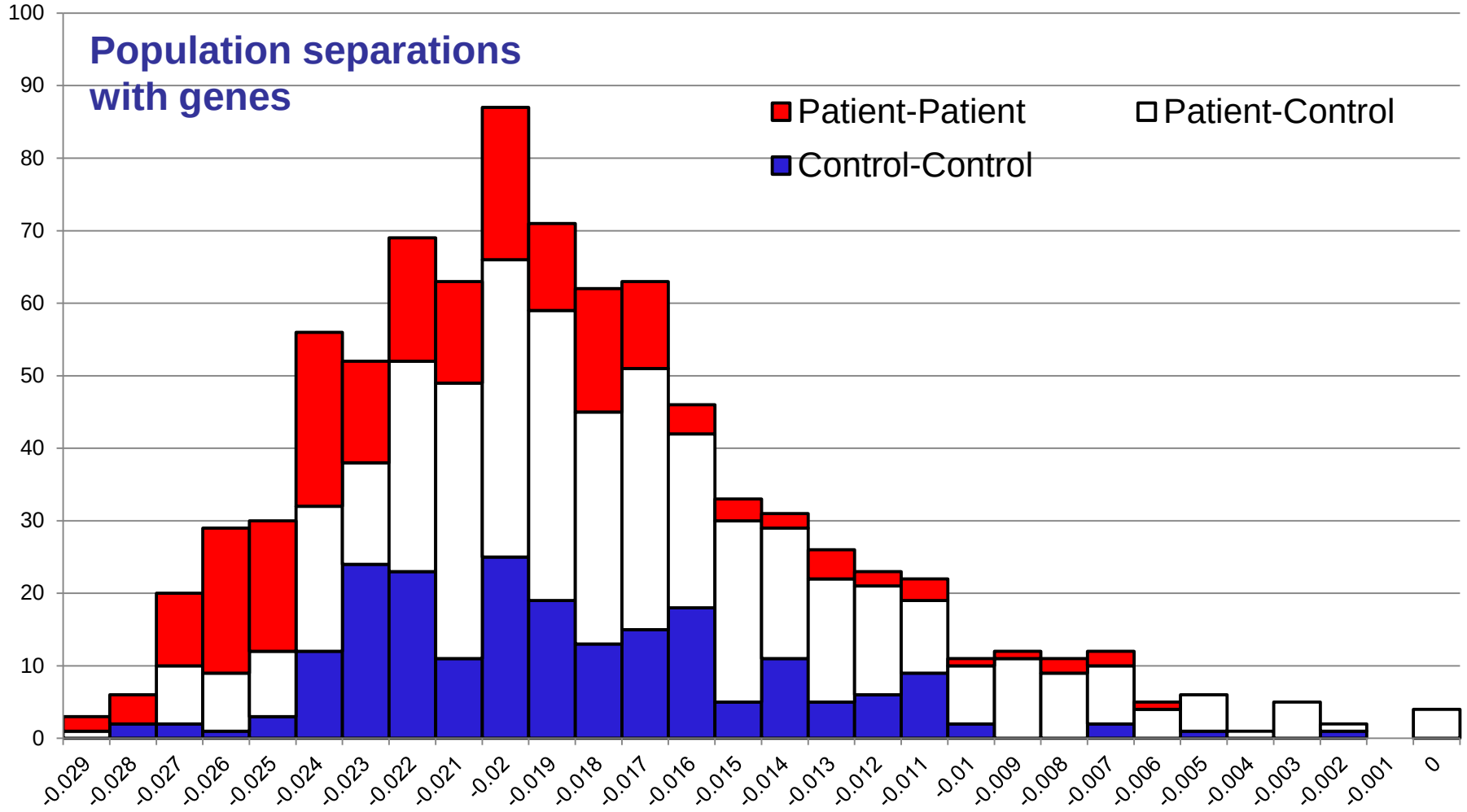


$$score(gene_i) = \left|m_{classA} - m_{classB}\right| - \left|\sigma_{classA} + \sigma_{classB}\right|$$

Followed by bipartite graph construction, edge weighting, thresholding, and application of nonplanar red/blue dominating set.

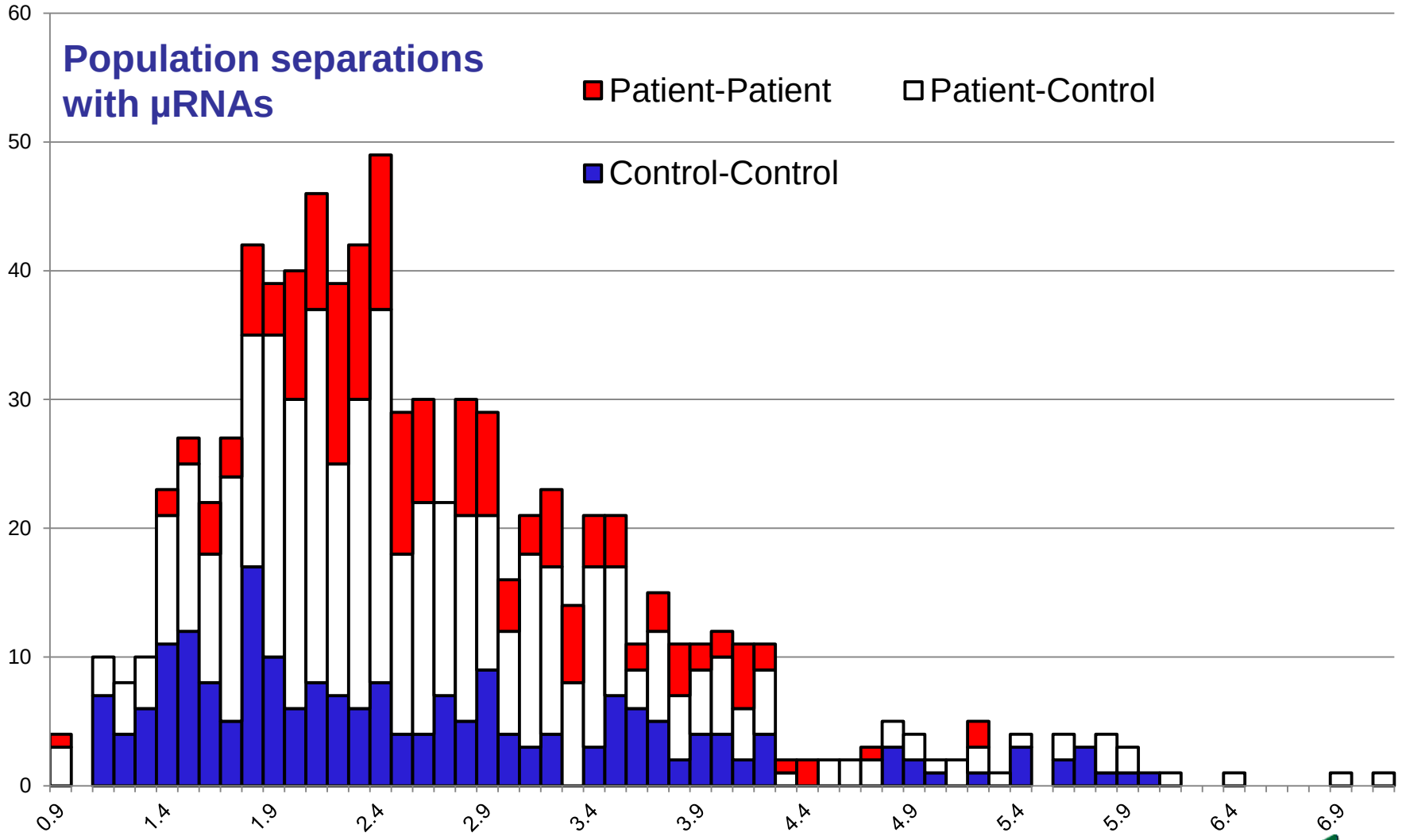
APAC 2012

Application, Biomarker Discovery



APAC 2012

Application, Biomarker Discovery



APAC 2012

Application, Biomarker Discovery

Population separations with methylation sites

APAC 2012

Outline of Talk

Some Thoughts on What, Really, is an "Application"

A Few Relevant Algorithms and Implementations

A Smorgasbord or Mélange of Applications

Some Ongoing Work

APAC 2012

Some Ongoing Work

Comparison graphs. Graphs are compared, on enormous scales, to detect anomalies and extract path, subgraph and other topological differences.

Time series or dynamic graphs. We perform differential analysis to determine how correlate relationships change over time and possibly space.

Noisy graphs. We use fuzzy clustering, soft and/or spectral thresholding, and other techniques to mitigate noise and see past its effects.

Graphs with repeated patterns. Motifs are identified in large and sometimes dynamic graphs. We employ maximum common subgraph and special cases of subgraph isomorphism.

Multivariate graphs. We have built libraries of algorithms to analyze graphs that represent heterogeneous data. Our methods frequently proceed with the use of multi-partite and other special graph classes.

Metagraphs. Also known as networks of networks, data is analyzed at multiple levels of granularity. A typical example is the use of clique intersection graphs to coarsen topological network data.

APAC 2012

The Langston Lab

Computer Science, Mathematics, Molecular Biology, Statistics