School of Computer Science Seminar Series

Computational Advances in High-Throughput Biological Data Analysis

Mike Langston

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

7 March 2011

Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning

Outline of Talk

Toolchains, Clustering, Thresholding, FPT

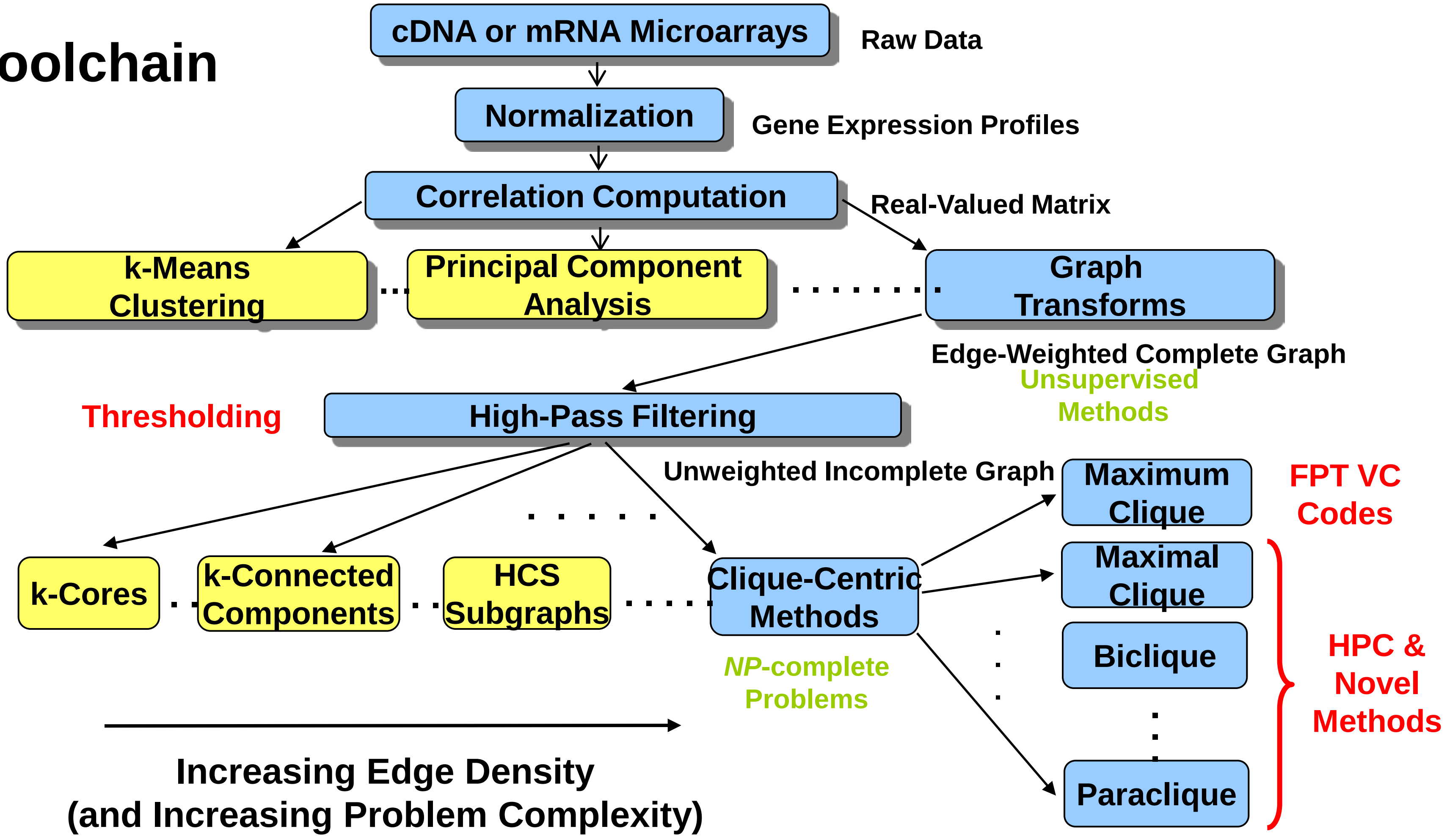
Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning

Clustering
A Classic Application

Toolchain



cDNA or mRNA Microarrays

Raw Data

Normalization

Gene Expression Profiles

Correlation Computation

Real-Valued Matrix

k-Means Clustering

Principal Component Analysis

Graph Transforms

Edge-Weighted Complete Graph

Unsupervised Methods

Thresholding

High-Pass Filtering

Unweighted Incomplete Graph

k-Cores

k-Connected Components

HCS Subgraphs

Clique-Centric Methods

NP-complete Problems

Maximum Clique

FPT VC Codes

Maximal Clique

Biclique

Paraclique

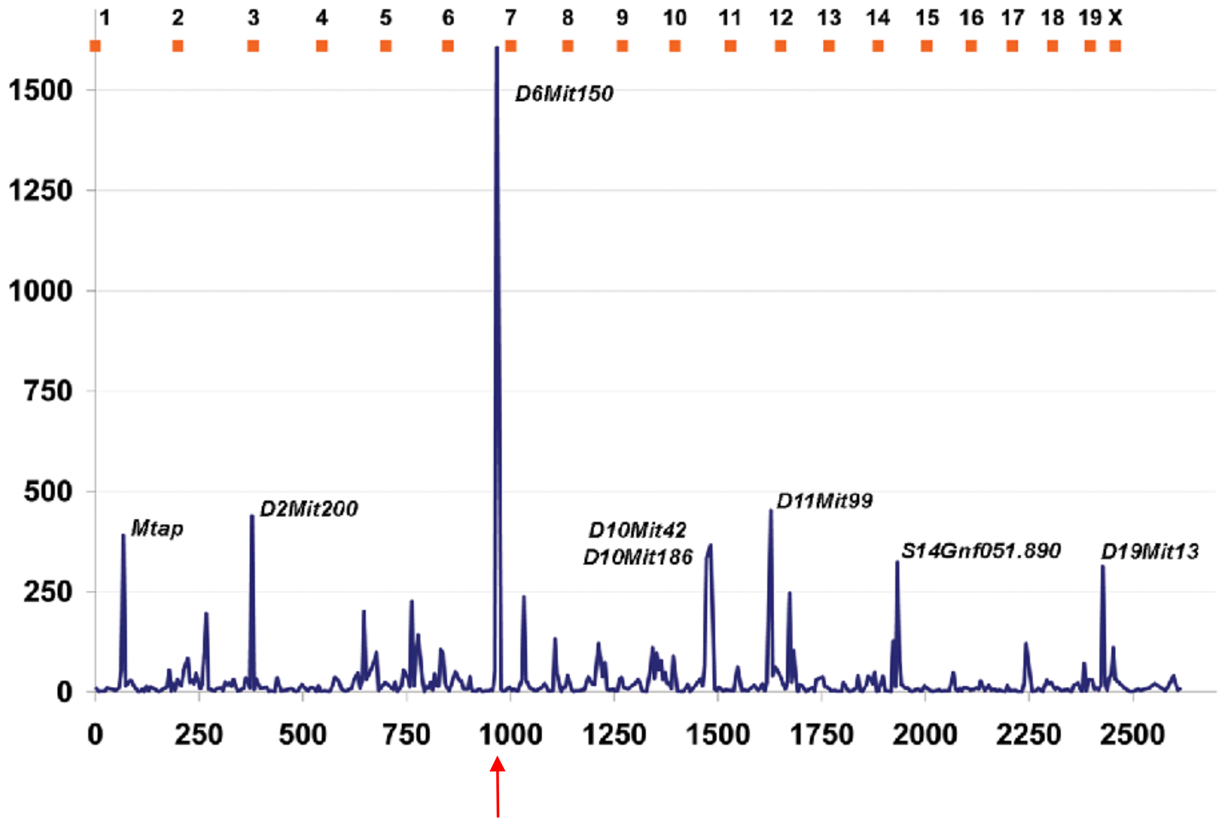
HPC & Novel Methods

Increasing Edge Density
(and Increasing Problem Complexity)

Algorithms Ranked by Quartile Comparisons

		Small (3-10 genes)		Medium (11-100 genes)		Large (101-1000 genes)	
Clustering Method	Average Quartile	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

Coexpression Analysis

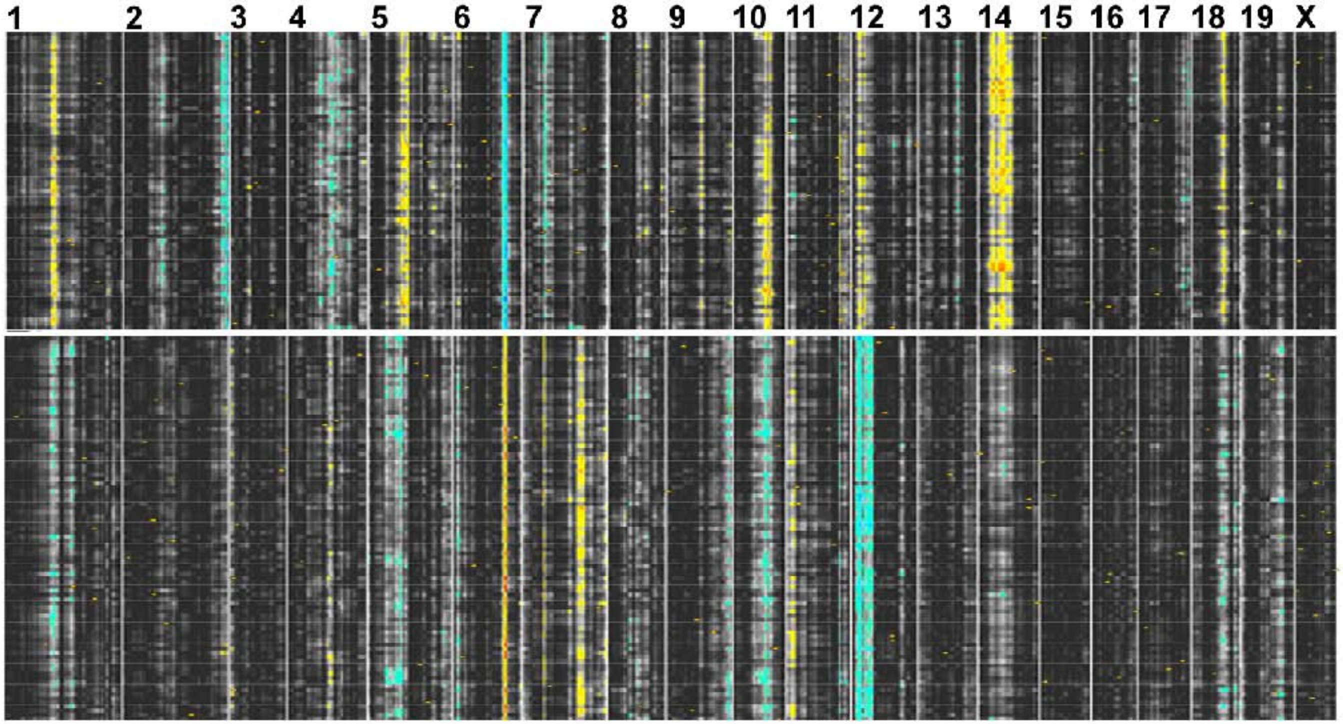


Seven Quantative Trait Loci

Transcript abundance can be the phenotype!

There's a high probability that somewhere in **here** is a polymorphism controlling this trait.

Coexpression Analysis



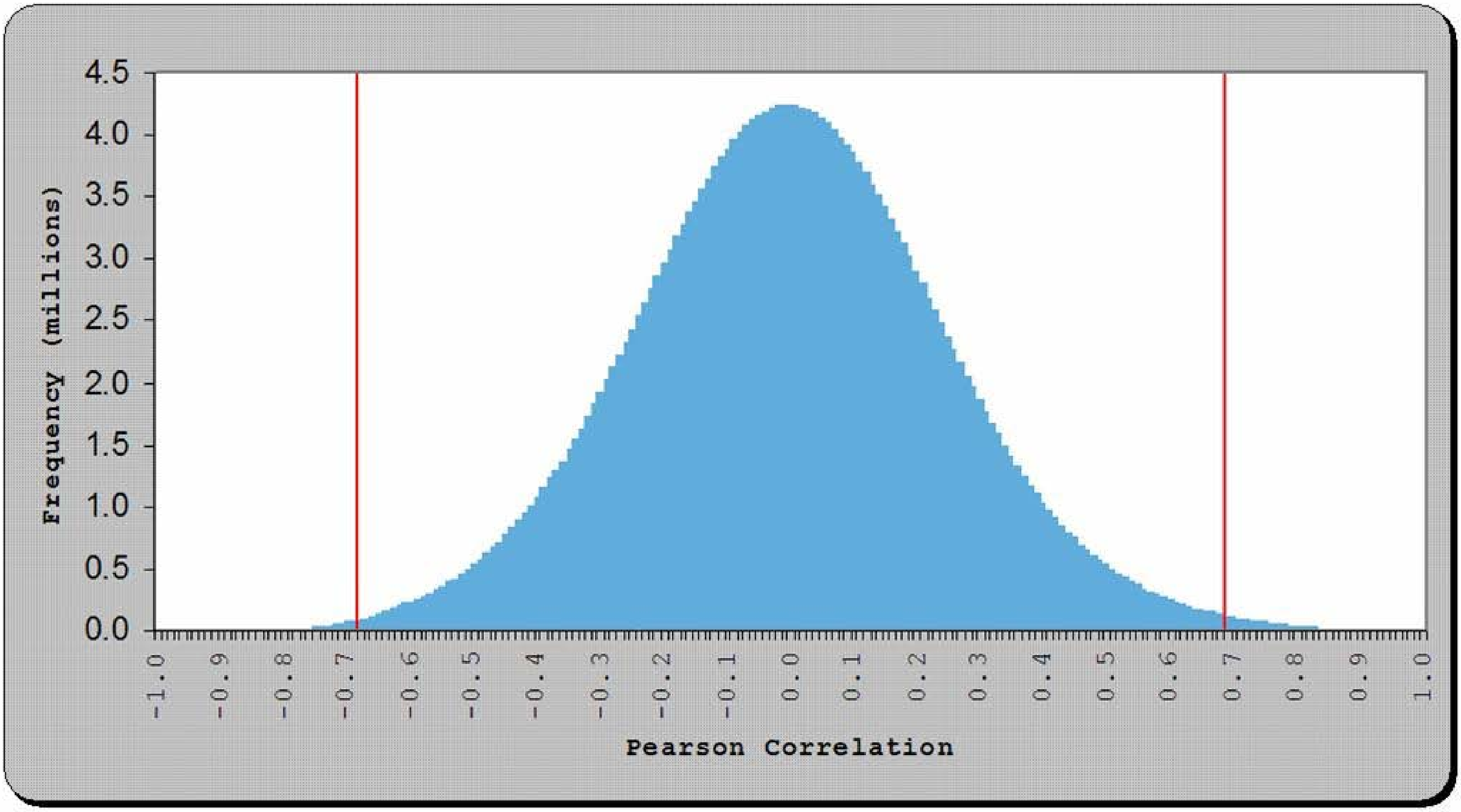
1

Two Paracliques

2

Concentrated Parental Alleles

Thresholding



Thresholding

Method	Anoxia	Reoxygen-ation	Alpha	Absolute deviations from GO threshold
GO Functional Similarity	0.97	0.92	0.85	
Spectral Clustering	**0.93**	**0.97**	**0.89**	**0.04+0.05+0.04=0.13**
Maximal Clique-2	**0.90**	**0.91**	**0.74**	**0.07+0.01+0.11=0.19**
Power	0.88	**0.94**	**0.96**	0.09+0.02+0.11=0.22
Bonferroni adjustment	0.85	**0.93**	**0.95**	0.12+0.01+0.10=0.23
Control-Spot	0.93	0.83	0.70	0.04+0.09+0.15=0.28
Maximal Clique-3	0.87	0.89	0.60	0.10+0.03+0.25=0.38
Top 1 Percent	0.81	0.81	0.72	0.16+0.11+0.13=0.40

Estimated threshold for each dataset, sorted by performance of the methods. GO functional similarity thresholds are the standard against which the methods are compared, summing absolute deviations across datasets (thresholds above GO are in bold).

Fixed-Parameter Tractability

Pioneering approach going back twenty-five years

- Well-Quasi-Order theory
- nonuniform measure of complexity

Exploit knowledge of the solution space

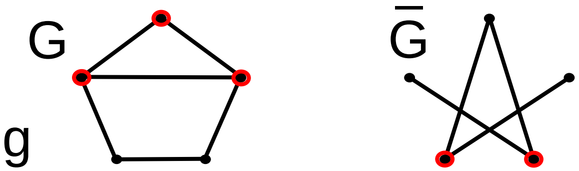
- Consider an algorithm with a time bound such as $O(2^{kn})$.
- And now one with a time bound more like $O(2^k n)$.
- Both are exponential in parameter value(s).
- But what happens when *k* is fixed?
- Fixed-Parameter Tractable (FPT) iff $O(f(k)n^c)$
- Confines superpolynomial behavior to the parameter

Duality

- We solve **vertex cover**, clique's complementary dual
- $O(1.2738^k k^{1.5}+kn)$ time

Key features

- Kernelization, branching and interleaving



Outline of Talk

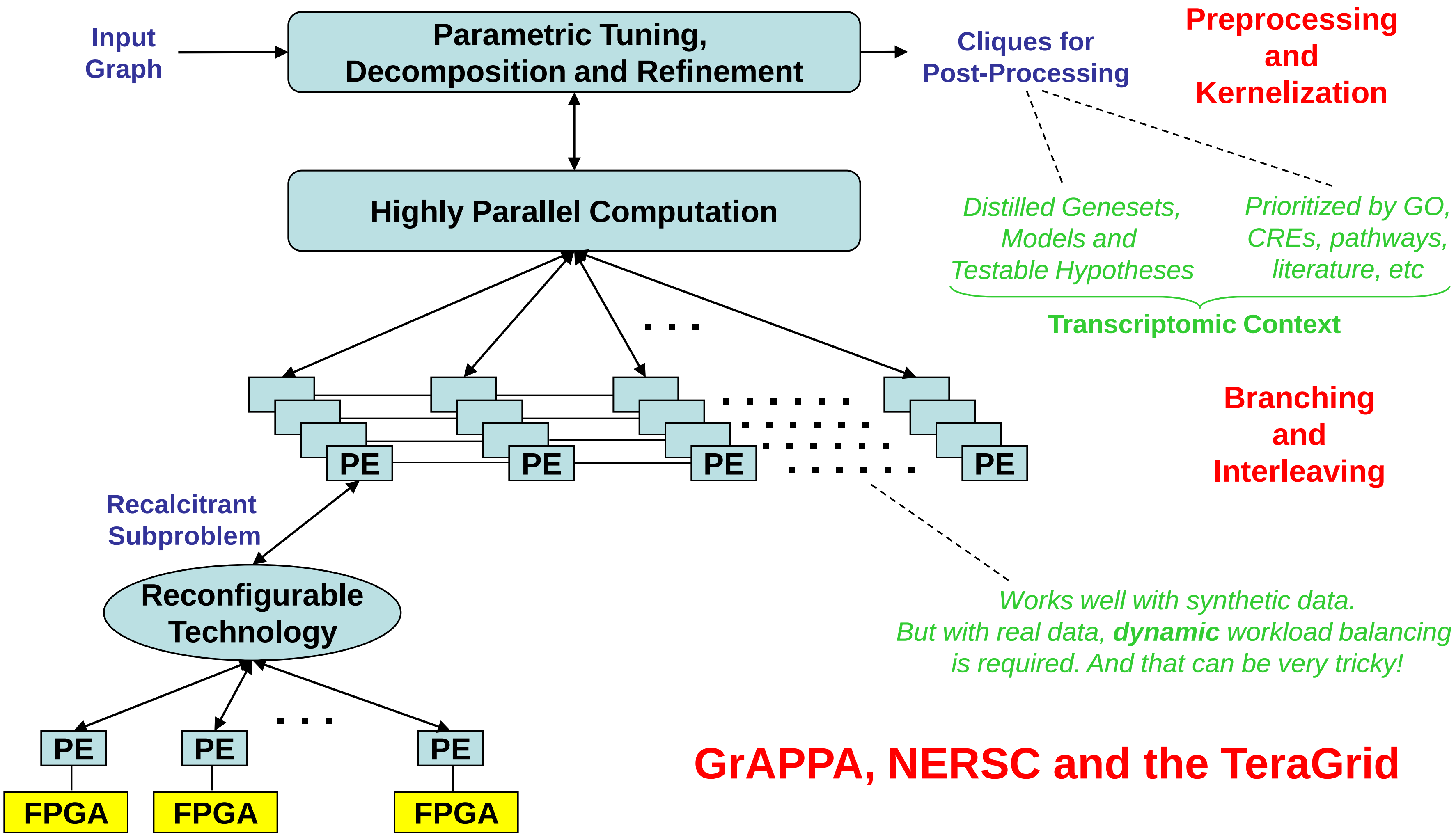
Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning

A Clique Compute Engine


Input Graph
Parametric Tuning, Decomposition and Refinement
Cliques for Post-Processing
Preprocessing and Kernelization
Highly Parallel Computation
Distilled Genesets, Models and Testable Hypotheses
Prioritized by GO, CREs, pathways, literature, etc
Transcriptomic Context
. . .
PE
PE
PE
PE
Branching and Interleaving
Recalcitrant Subproblem
Reconfigurable Technology
Works well with synthetic data.
But with real data, dynamic workload balancing is required. And that can be very tricky!
PE
PE
. . .
PE
FPGA
FPGA
FPGA

GrAPPA, NERSC and the TeraGrid

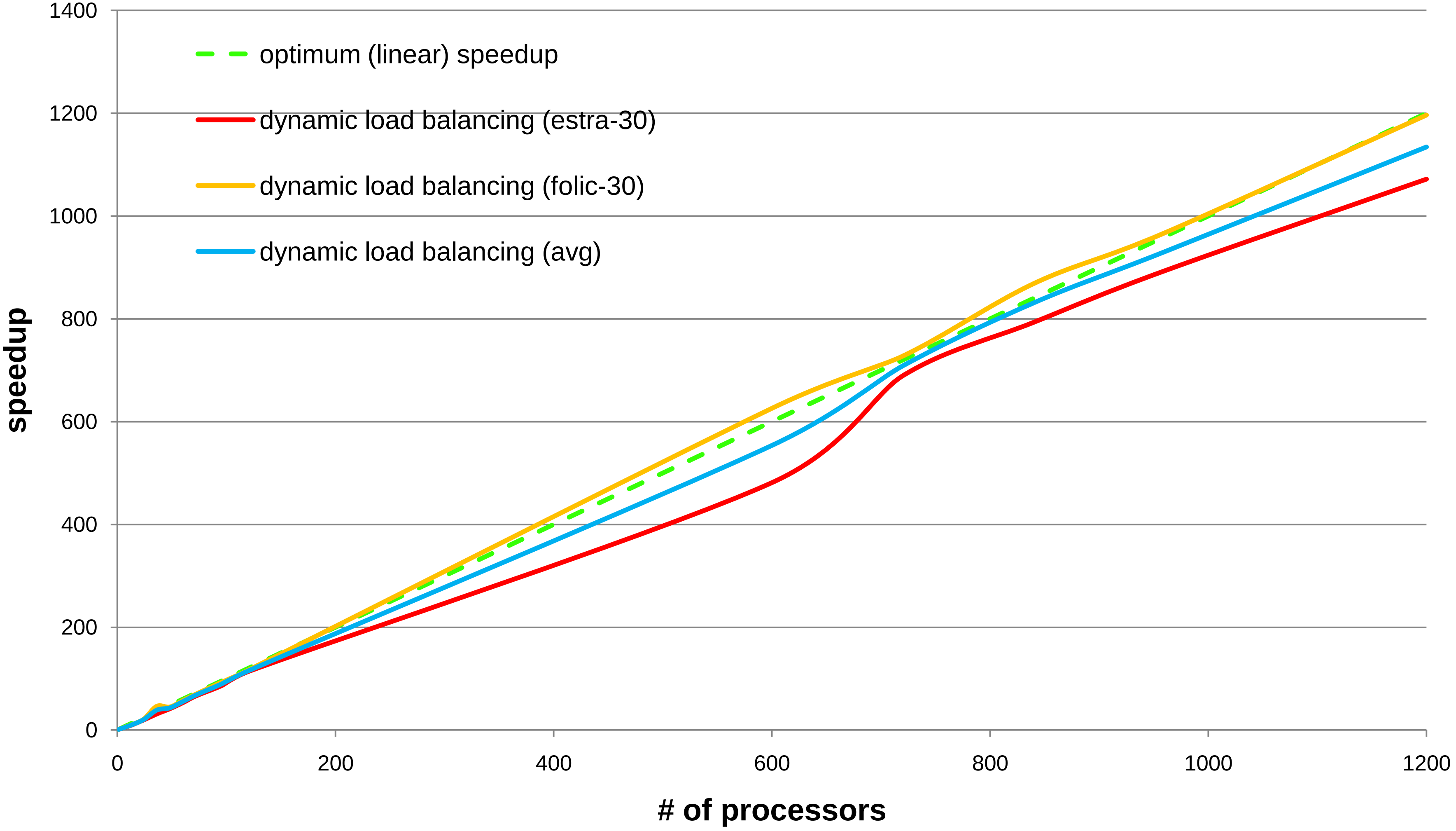
Supercomputer Implementations

Now also using new ORNL-UT Cray XT5 system, Kraken

- currently the world’s largest academic (non defense) computer
- 10^5 processor cores (and expanding)
- nearly 10^{12} calculations per second (a petaflop)
- quite a beast to harness, at least for combinatorial work

Carleton UNIVERSITY Canada's Capital University

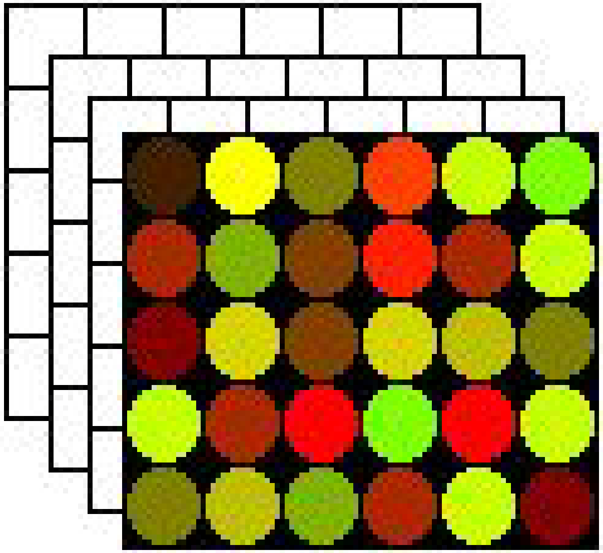
Workload Balancing and Speedup



Differential Analysis

Gene (vertex) comparisons:

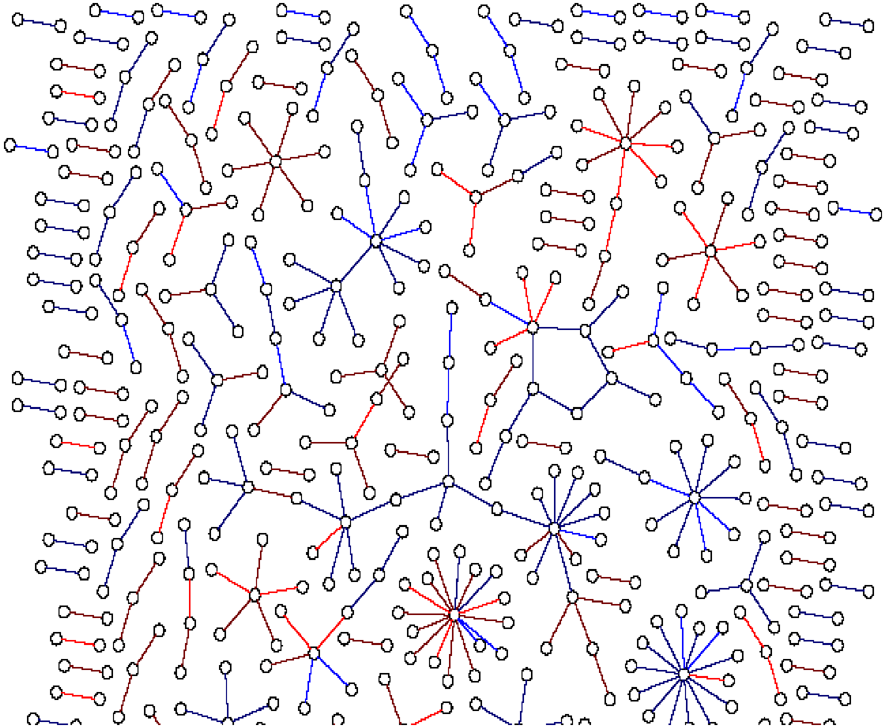
- differential expression
- does not require multiple conditions
- compare the two lists of gene expression levels



Differential Analysis

Correlate (edge) comparisons

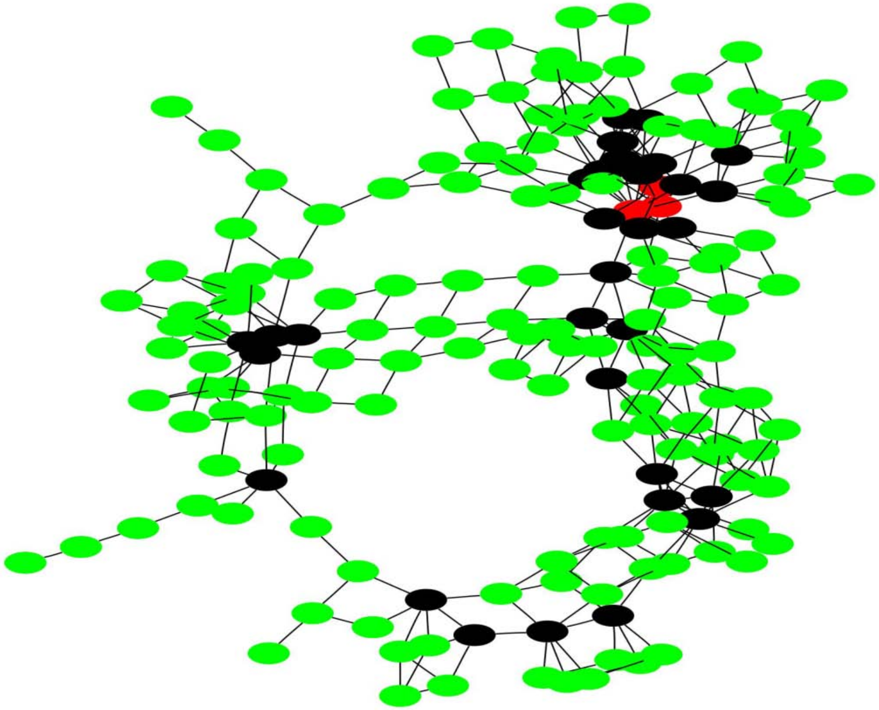
- differential correlation
- requires multiple conditions in control versus stimulus
- compare two lists of gene-gene correlations



Differential Analysis

Putative network (clique) comparisons

- differential topology
- compare dense subgraphs, sort by ontology, CREs, etc
- consider granularity, for example, with the clique intersection graph



Outline of Talk

Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

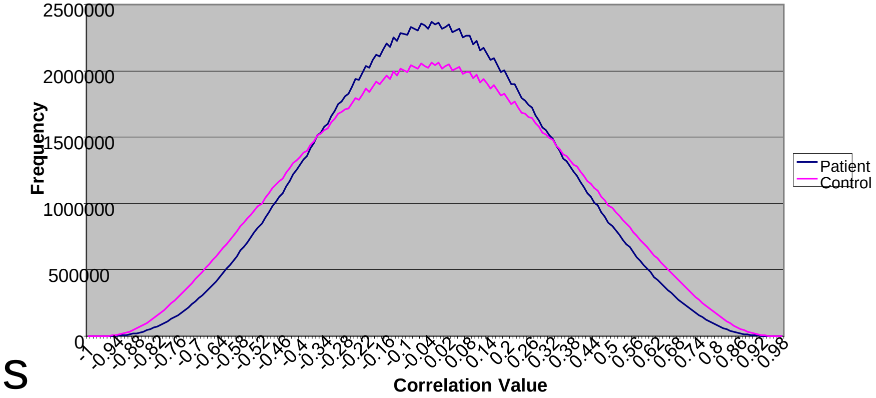
Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning

Application, Allergy

Data Description

- Mikael Benson, Göteborg, Sweden, 56 patients and 39 controls
- Affymetrix HU133 arrays
- roughly 33,000 genes
- 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

- chosen to balance graph densities
- AFFX spots retained for quality control

Application, Allergy

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 ChIP-chip technologies, etc.

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

Application, Cancer

Data Inhomogeniety

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

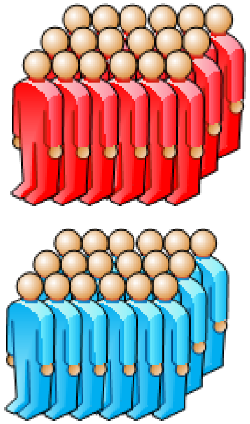
Creative Use of Graph Algorithms

- perform multiple data views
- drive correlations with both persons and genes
- exclude outliers with clique-centric tools
- perform differential analysis to distill biomarkers from genome

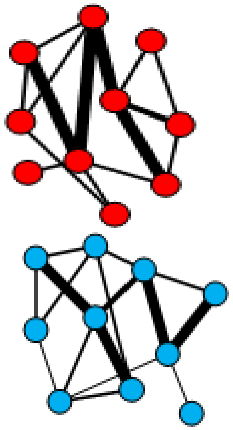
Application, Cancer

Case

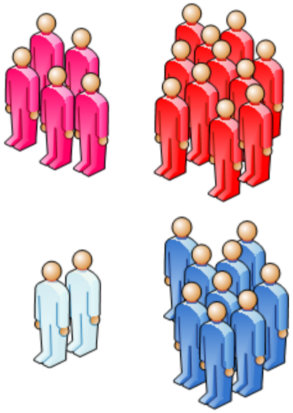
Control



Genes Drive Person-Person Correlations

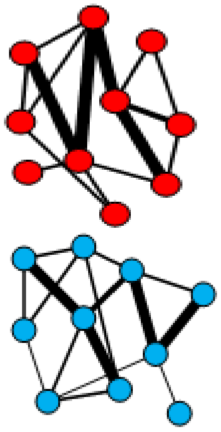


Select Thresholds, Extract Cliques

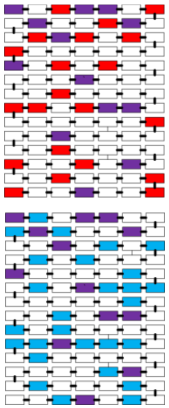


Classify Subtypes and Eliminate Outliers

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, Radiation

Low dose ionizing radiation and its impact on human health

- Sources of low dose radiation exposures
  - medical diagnostics
  - hazardous waste abatement
  - handling materials for nuclear weapons and power systems
  - even terrorist acts such as dirty bombs
- In all these the major type of exposures will be low dose IR (primarily X- and gamma-radiation) from fission products
- Are low doses safe, perhaps even therapeutic?
- Identify biological pathways that are activated or repressed by IR
- Understand the risks so that we may protect the workforce

Sample Result: Gene for Tubby-like Protein 4 (*Tulp4*)

- a nucleus of six genes are putatively coregulated in dose
- in fact they appear together in 5765 dose cliques
- yet no more than two occur together in any control clique
- this nucleus includes genes known to be involved in
  - immune function
  - stress mediation
  - and so these are consistent with IR response
- but one of these is Tulp4...why is a tubby-like protein here?
- original classification
  - based on sequence similarity to Tub, an adipose tissue protein
  - responsive to oxidative stress
- it's in 4.7% of the dose cliques and only 0.01% of control
- novel role for Tulp4 as a transcriptional regulator of immune response to IR?

Outline of Talk

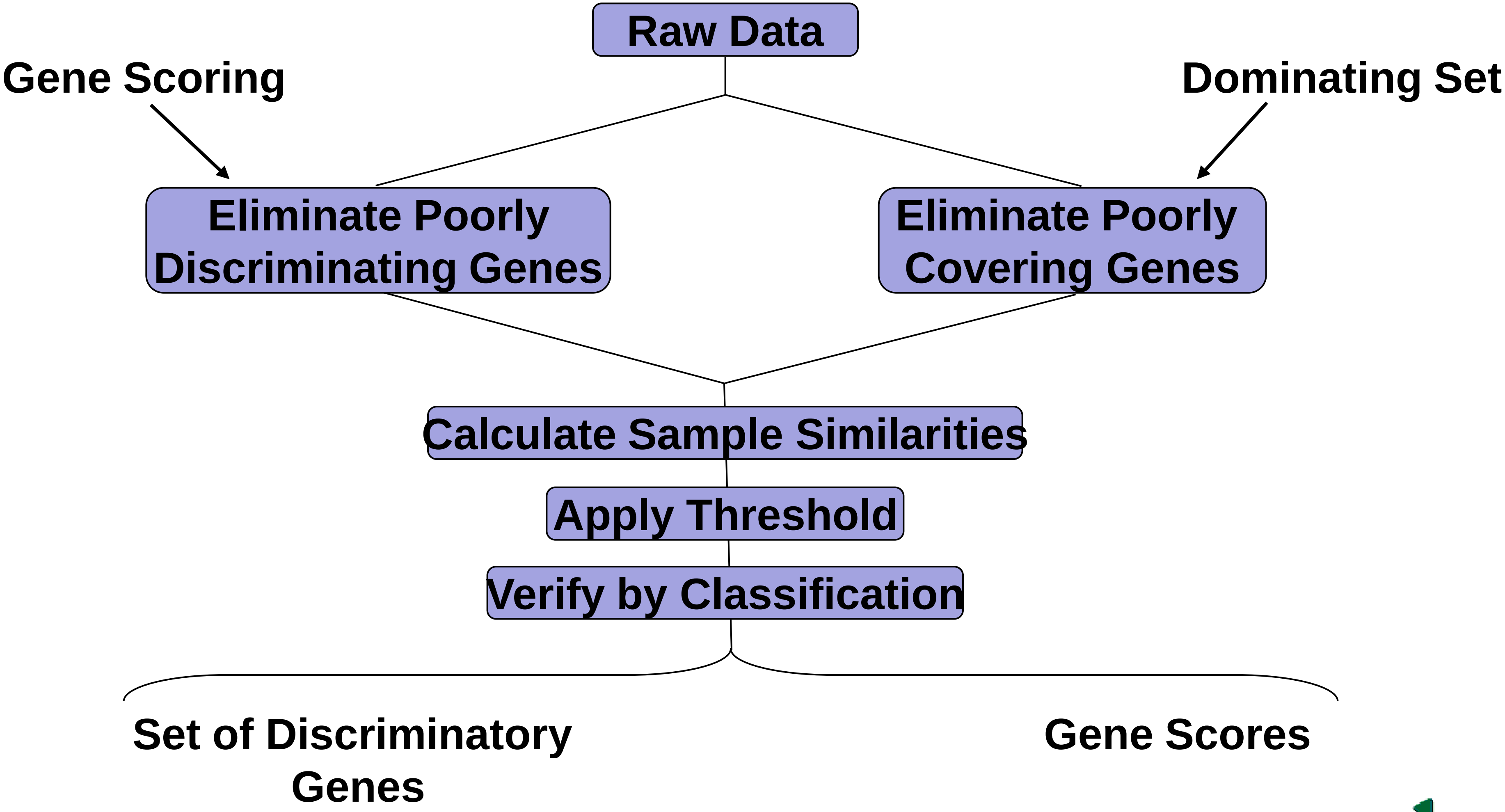
Toolchains, Clustering, Thresholding, FPT

Computation, Workload Balancing, Differential Analysis

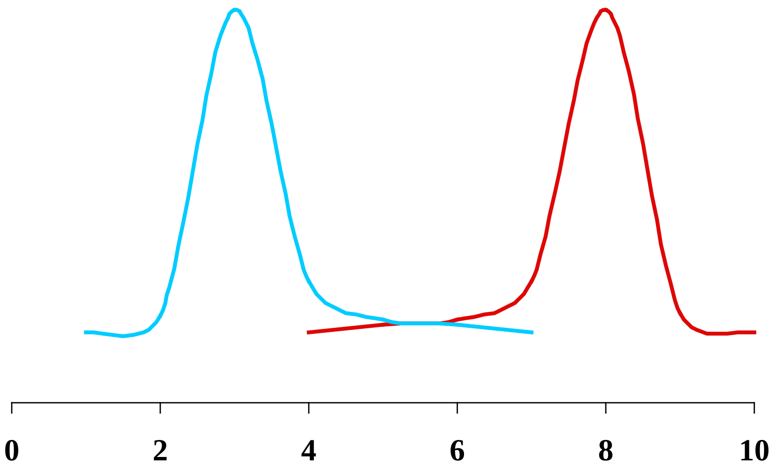
Sample Applications: Allergy, Cancer, Radiation

Biomarkers and Machine Learning

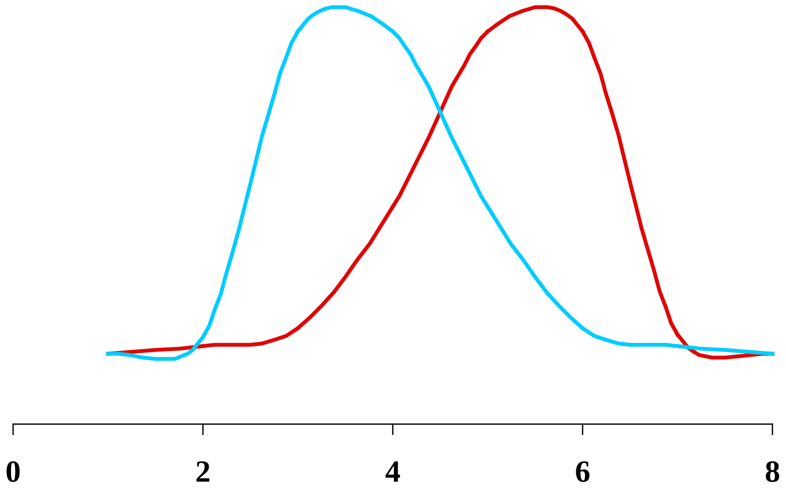
Algorithmic Training


Raw Data
Gene Scoring
Dominating Set
Eliminate Poorly Discriminating Genes
Eliminate Poorly Covering Genes
Calculate Sample Similarities
Apply Threshold
Verify by Classification
Set of Discriminatory Genes
Gene Scores

Gene Scoring



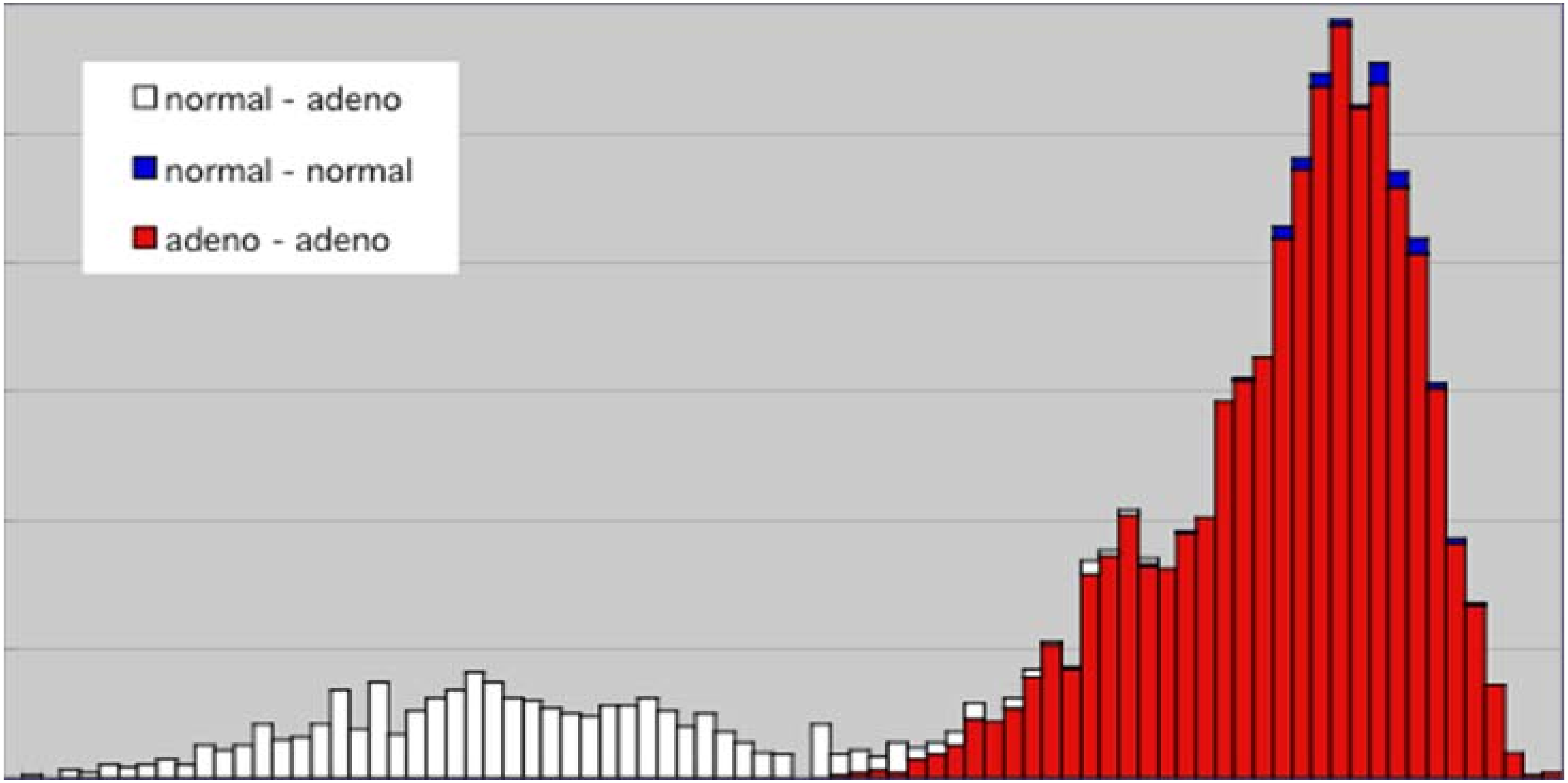
vs.



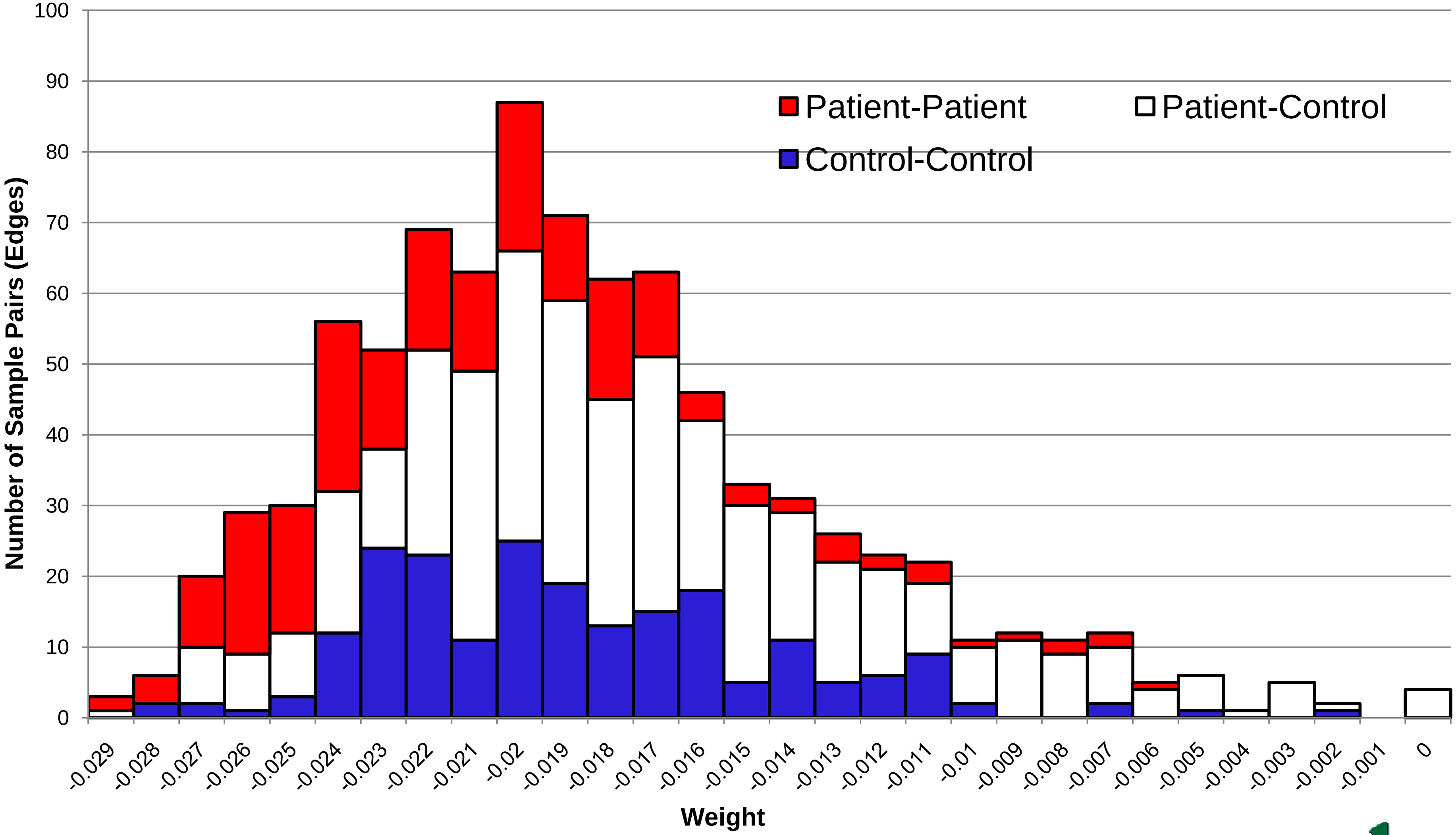
$$score(gene_i) = \left| m_{classA} - m_{classB} \right| - \left| \sigma_{classA} + \sigma_{classB} \right|$$

Followed by edge weighting.

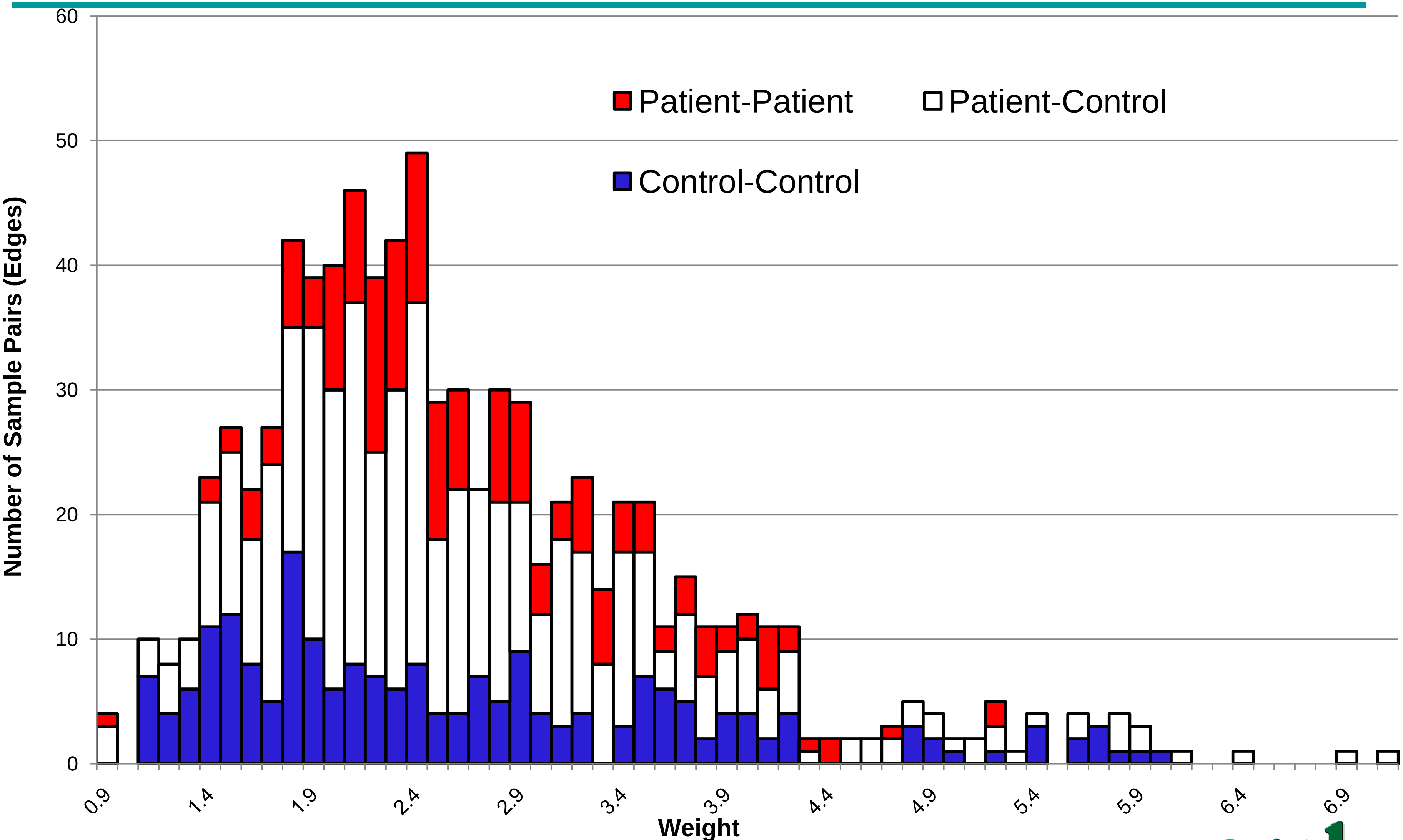
Edge Weight Spectrum



Allergic Rhinitis: Top 100 Genes

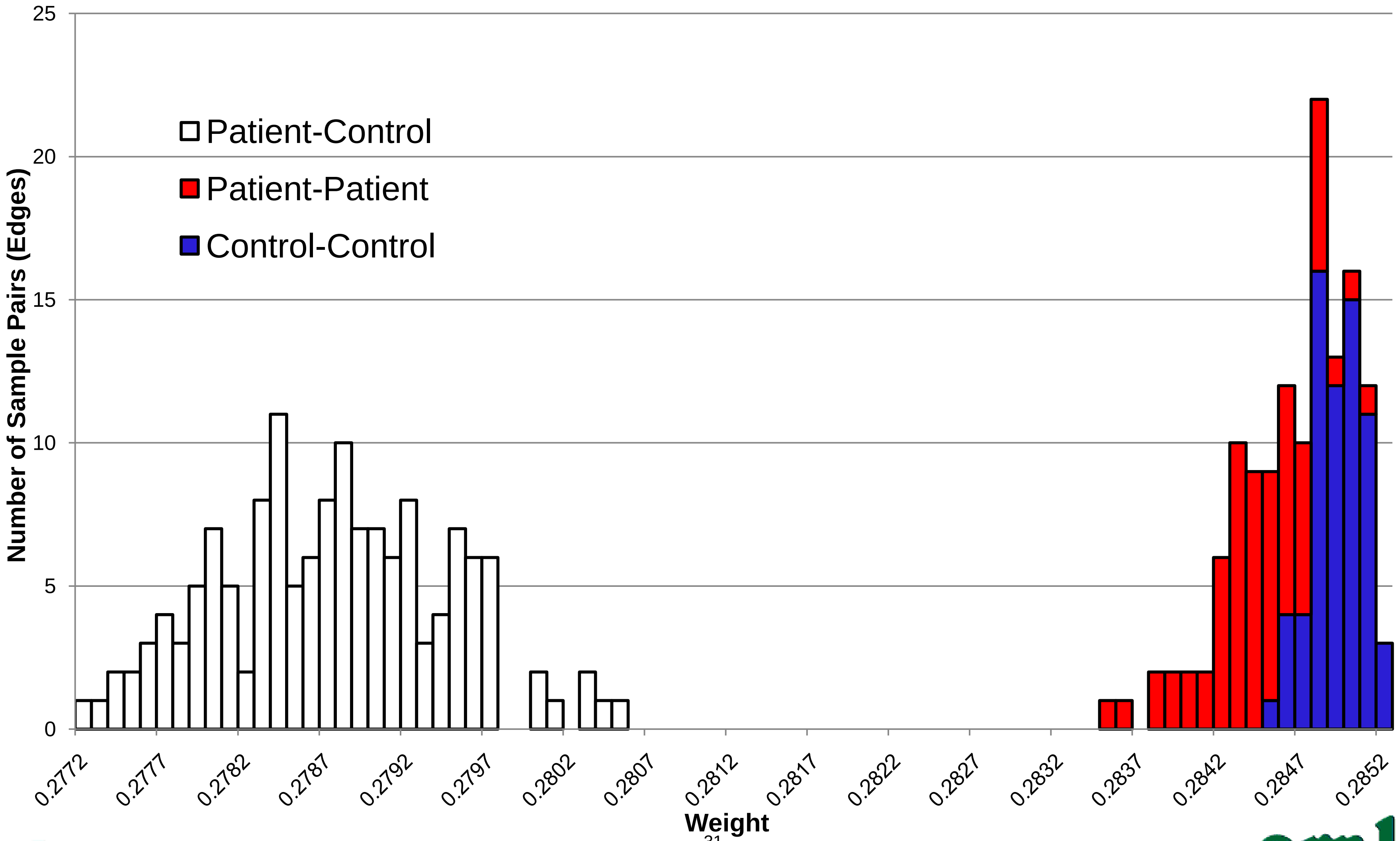


Allergic Rhinitis: Top 100 µRNAs

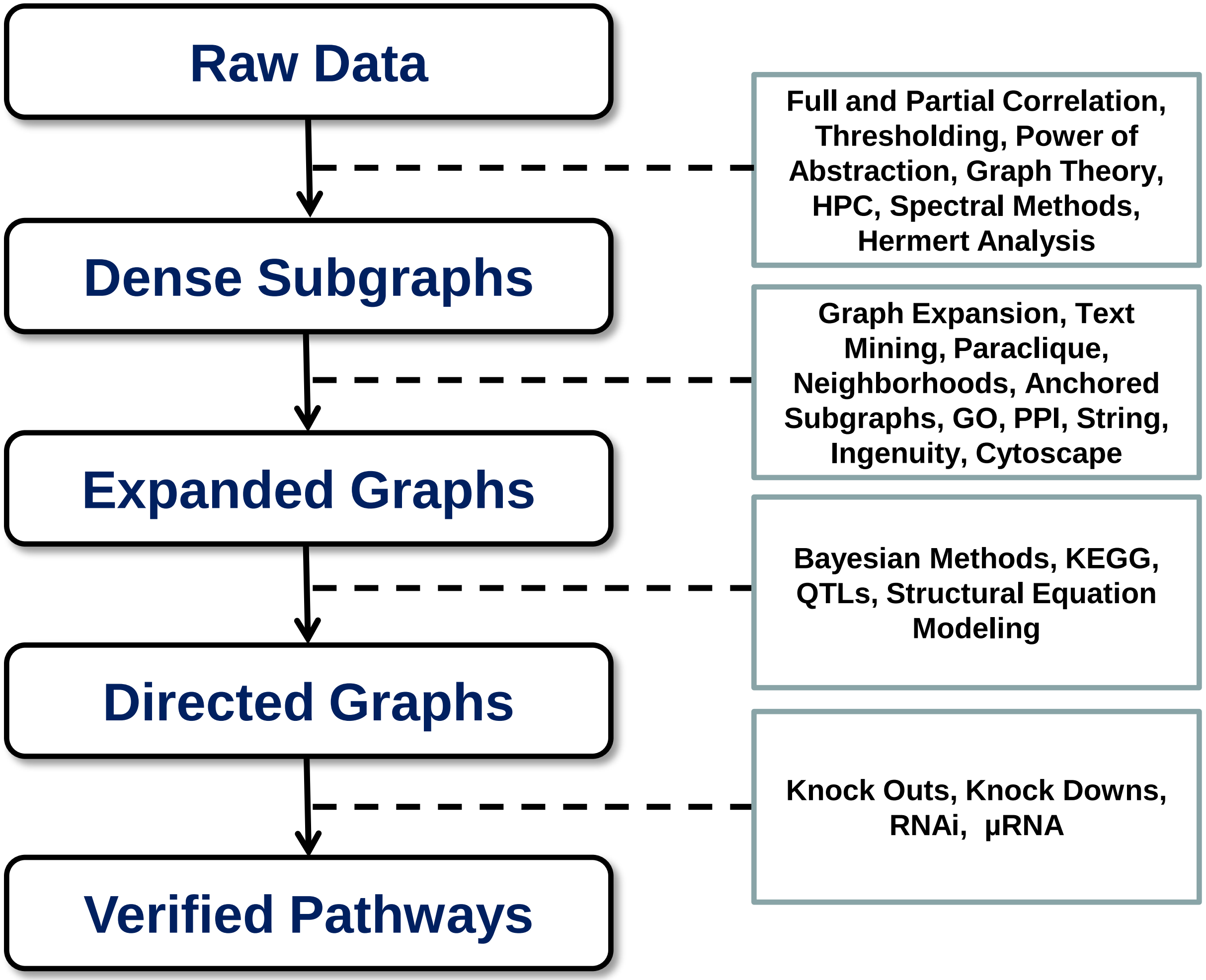


Carleton UNIVERSITY Canada's Capital University

Allergic Rhinitis: Top 100 Methylation Sites



Where We Are, Where We're Going



Raw Data

Dense Subgraphs

Expanded Graphs

Directed Graphs

Verified Pathways

Full and Partial Correlation, Thresholding, Power of Abstraction, Graph Theory, HPC, Spectral Methods, Hermert Analysis

Graph Expansion, Text Mining, Paraclique, Neighborhoods, Anchored Subgraphs, GO, PPI, String, Ingenuity, Cytoscape

Bayesian Methods, KEGG, QTLs, Structural Equation Modeling

Knock Outs, Knock Downs, RNAi, µRNA

Carleton UNIVERSITY Canada's Capital University

The Langston Lab

Computer Science, Mathematics, Molecular Biology, Statistics