NZIMA
Napier
2008

New Zealand Institute of Mathematics & its Applications
programme in algorithmics

Analysis of High-Throughput Biological Data Part II: Computational Bottlenecks and Novel Applications

Mike Langston

Professor
Department of Electrical Engineering and Computer Science
University of Tennessee
and
Collaborating Scientist
Biological Sciences Division
Oak Ridge National Laboratory
USA

21 February 2008

NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

Application to Human Health

Protein Complex Prediction

Application to Model Organisms

NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

Application to Human Health

Protein Complex Prediction

Application to Model Organisms

Foundations

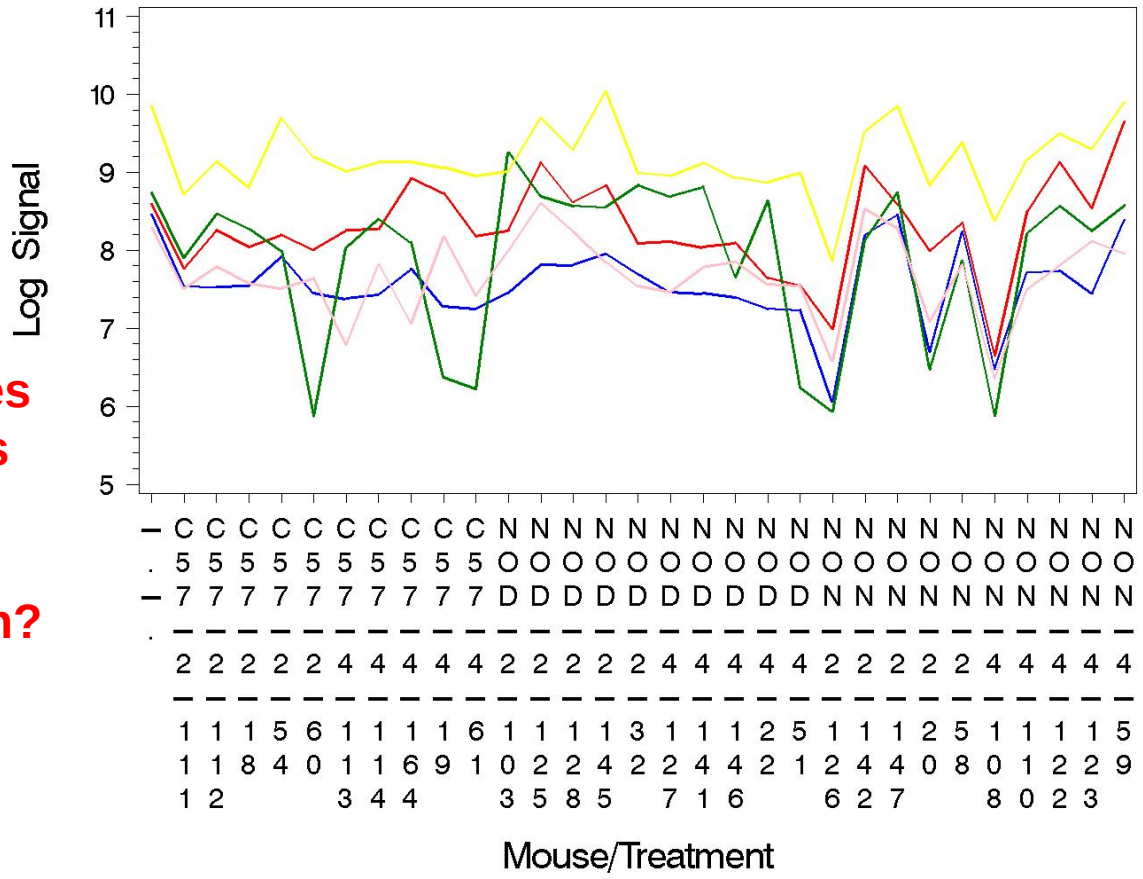
Systems Biology

- How do biological entities function in unison and at all levels of scale?
- Linkage, communication and networks (graphs!)

Foundations

Systems Biology

Correlation



Here are five mouse genes with Pearson correlations of at least 0.65. What of

- **noise?**
- **experimental design?**
- **circadian rhythms?**
- **other confounds?**
- **other metrics?**

Foundations

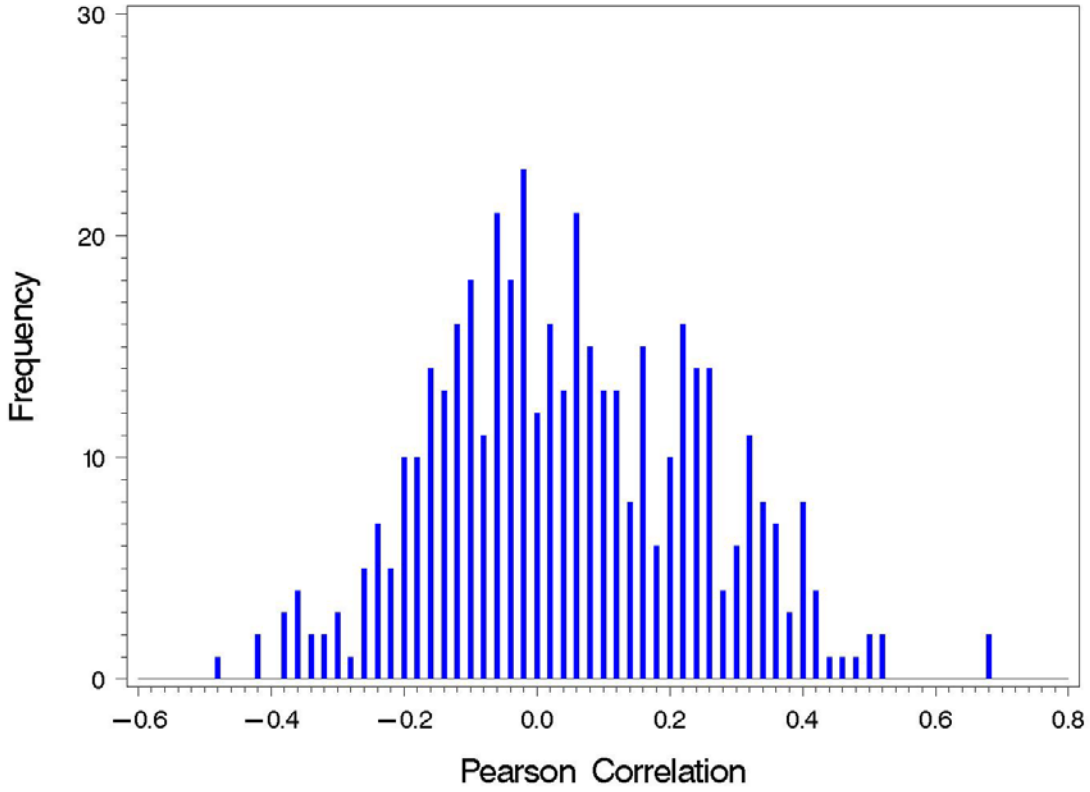
Systems Biology

Correlation

Coefficient Profiles

Sometimes via
- Pearson
- Spearman
- Mutual Information
- Etc

Other times we need
- p-values
- Bonferroni corrections
- q-values
- false discovery rates...



Foundations

Systems Biology

Correlation

Omics: key to deciphering complex systems

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics: key to deciphering complex systems

Humans: 10^{14}+ cells, 200+ cell types

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics: key to deciphering complex systems

Humans: 10^{13}+ cells, 200+ cell types

Genome (blueprint, 20K+ genes, 10M+ polymorphisms)

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics: key to deciphering complex systems

- **Humans: 10^{13}+ cells, 200+ cell types**
- **Genome (blueprint, 20K+ genes, 10M+ polymorphisms)**
- **Proteome (functional units, unknown # of proteins)**

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics: key to deciphering complex systems

- **Humans: 10^{13}+ cells, 200+ cell types**
- **Genome (blueprint, 20K+ genes, 10M+ polymorphisms)**
- **Proteome (functional units, unknown # of proteins)**
- **Transcriptome**
  - **Translation (tRNA) via transcription (mRNA)**
  - **Function and Signaling (siRNA, miRNA, etc)**

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics: key to deciphering complex systems

- **Humans: 10^{13}+ cells, 200+ cell types**
- **Genome (blueprint, 20K+ genes, 10M+ polymorphisms)**
- **Proteome (functional units, unknown # of proteins)**
- **Transcriptome**
  - **Translation (tRNA) via transcription (mRNA)**
  - **Function and Signaling (siRNA, miRNA, etc)**
- **Other: metabalome, lipidome, interactome, omeome!**

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics

Visualization

- highly dependent on scale

NZIMA
Napier
2008

Foundations

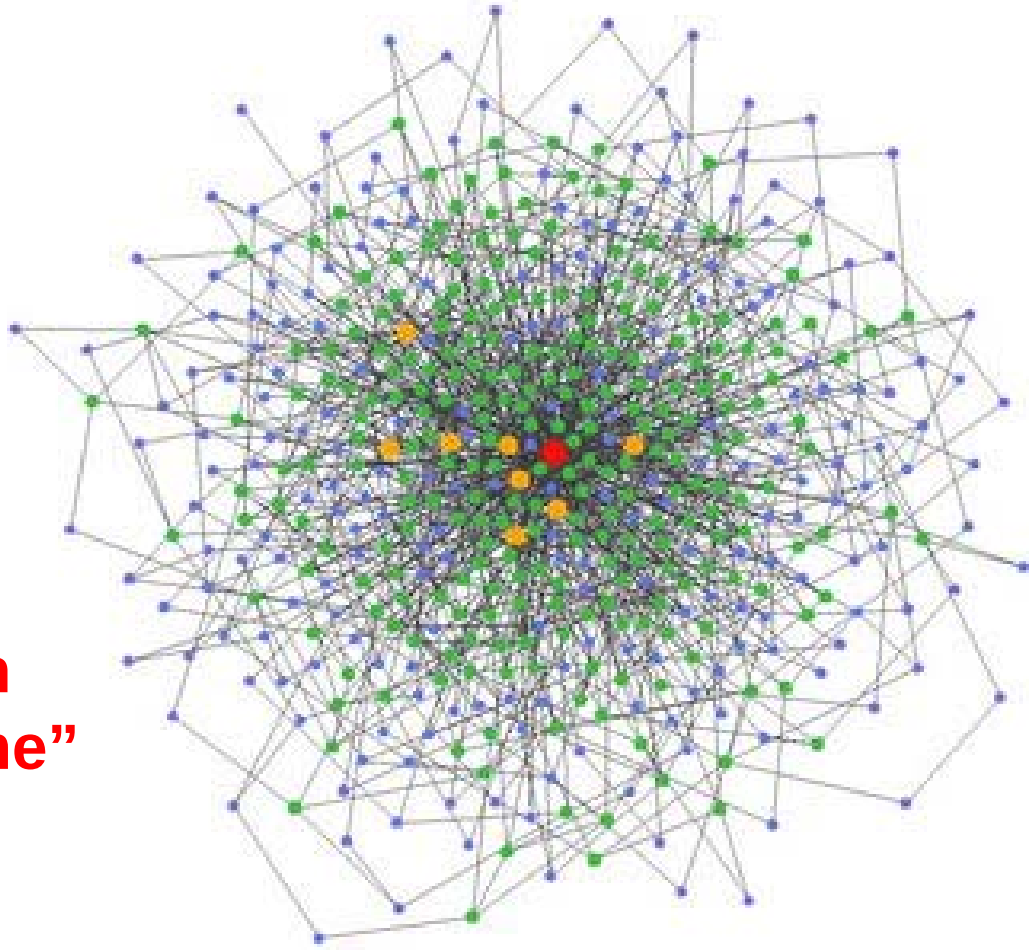
Systems Biology

Correlation

Omics

Visualization

- **highly dependent on scale**
- **the only omics often seen is a “rediculome”**



NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics

Visualization

Computational Tools - focus usually on dense subgraphs

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics

Visualization

Computational Tools

Maximum Clique

- **must run often**
- **time is a limiting factor**
- **exploit fixed-parameter tractability (FPT)**

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics

Visualization

Computational Tools

- **Maximum Clique**
- **Maximal Clique**
  - **huge outputs**
  - **various orderings**
  - **memory is often the limiting factor**

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics

Visualization

Computational Tools

- **Maximum Clique**
- **Maximal Clique**
- **Biclique**
  - **new algorithms**
  - **bipartite graphs**

NZIMA
Napier
2008

Foundations

Systems Biology

Correlation

Omics

Visualization

Computational Tools

- **Maximum Clique**
- **Maximal Clique**
- **Biclique**
- **Paraclique**
  - **noisy data**

NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

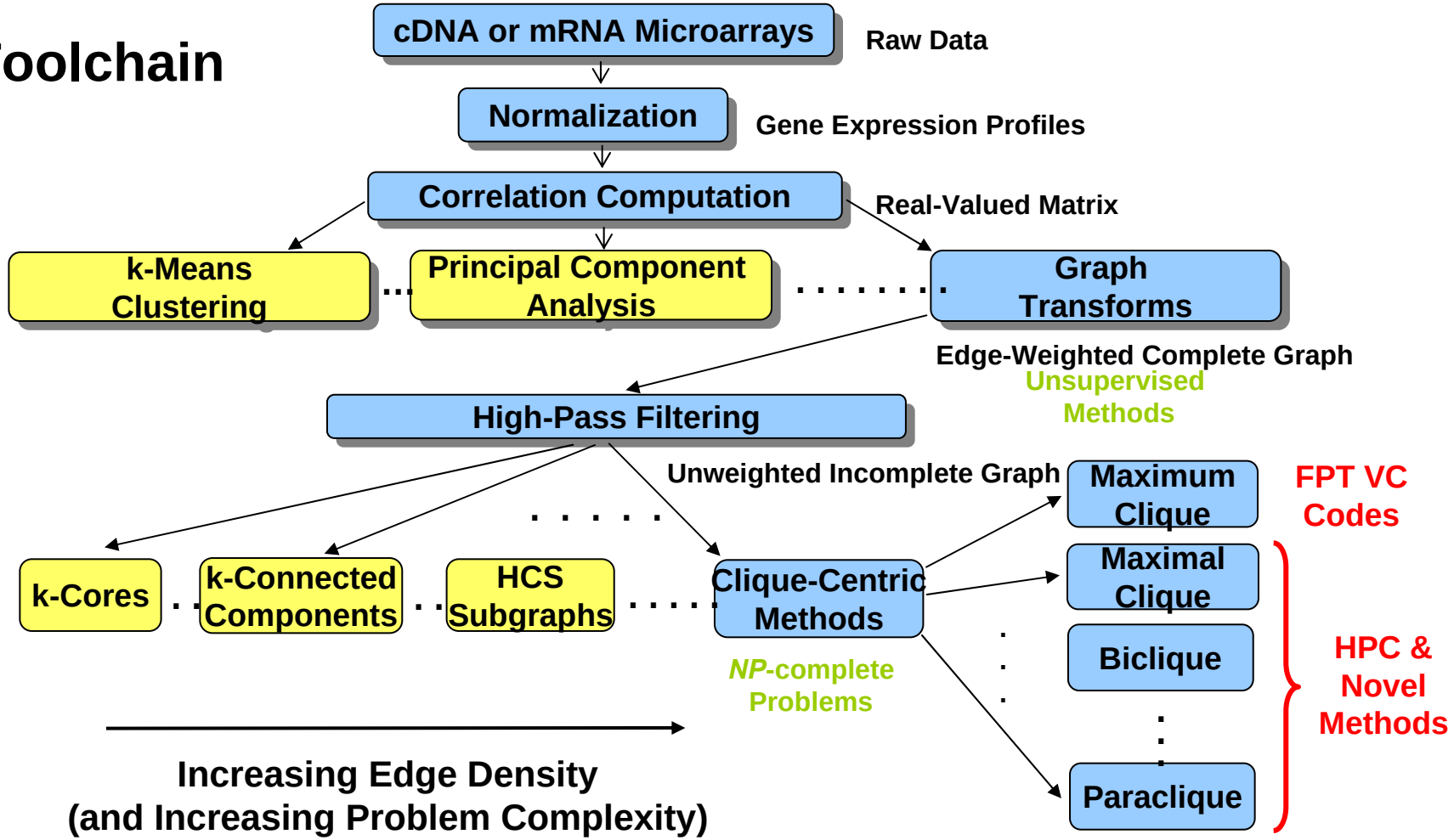
Application to Human Health

Protein Complex Prediction

Application to Model Organisms

Coexpression Analysis

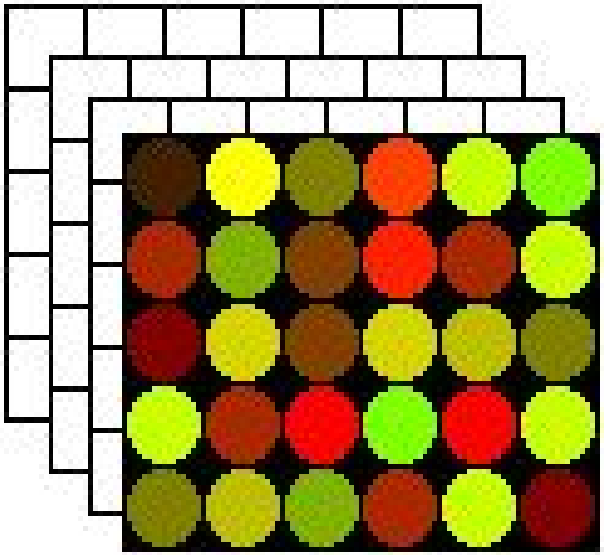
Toolchain



Coexpression Analysis

Gene (vertex) comparisons:

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

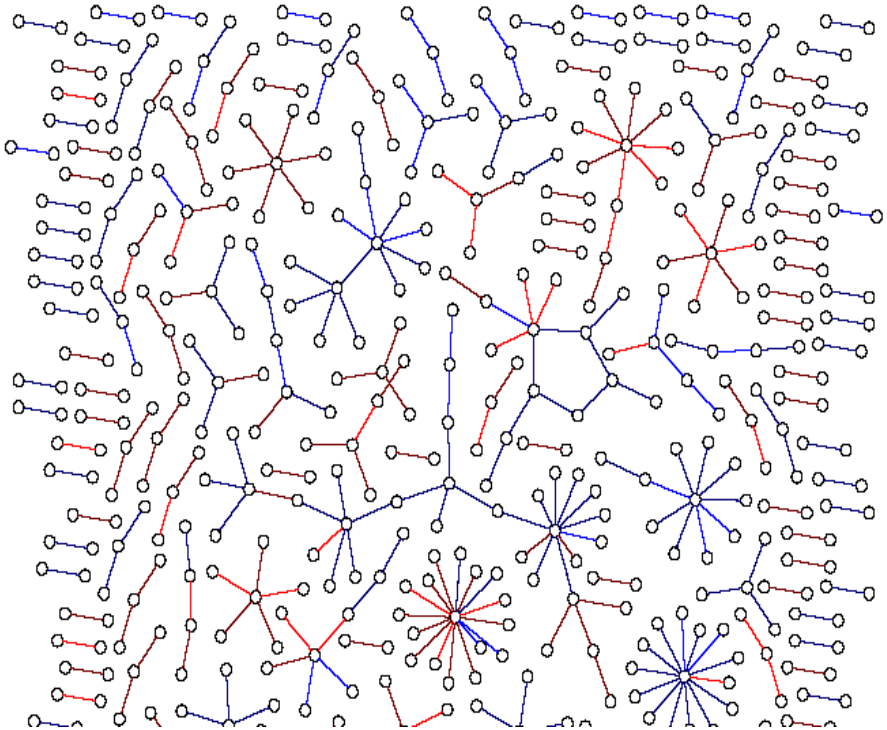


NZIMA
Napier
2008

Coexpression Analysis

Correlate (edge) comparisons

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

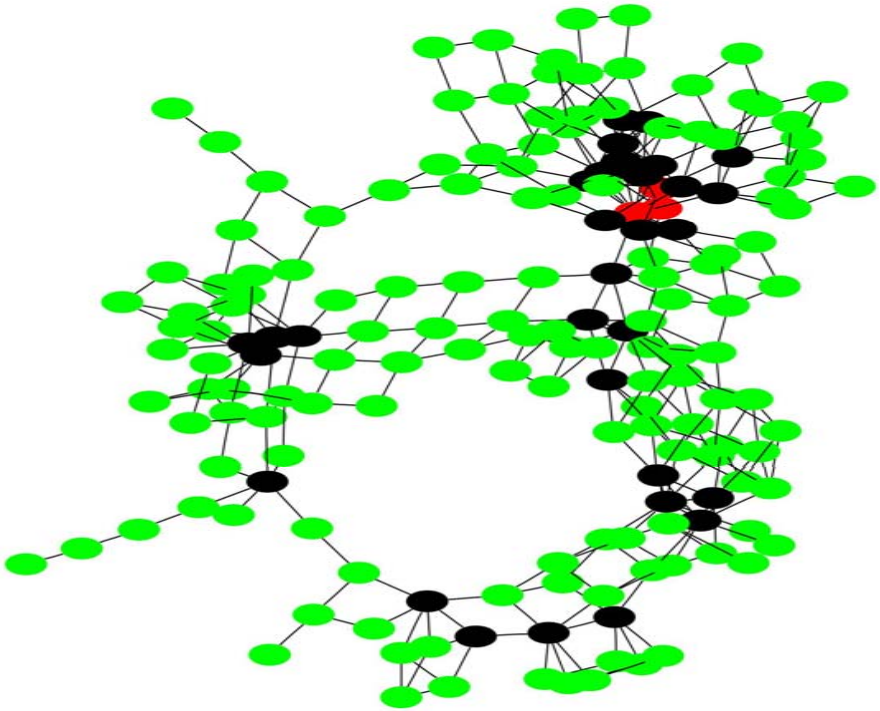


NZIMA
Napier
2008

Coexpression Analysis

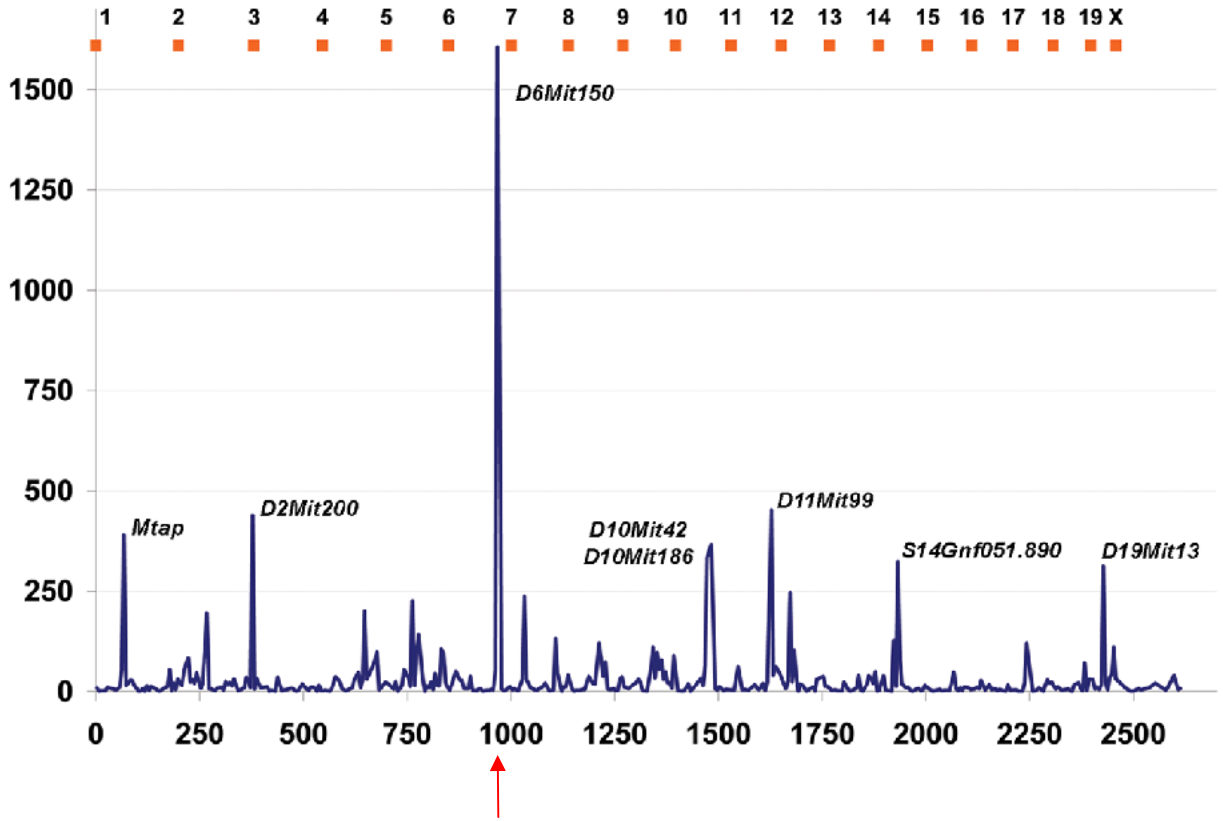
Putative network (clique) comparisons

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



NZIMA
Napier
2008

Coexpression Analysis



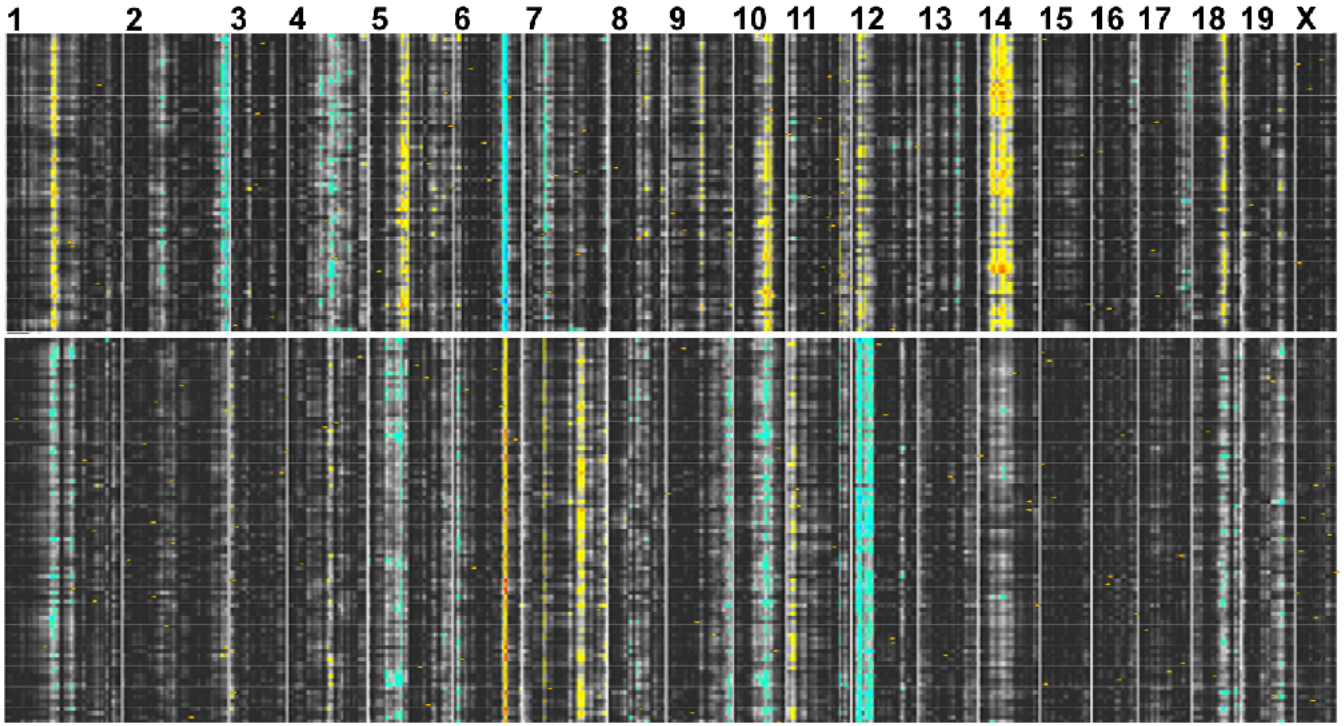
Seven Quantative Trait Loci

Transcript abundance can be the phenotype!

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

NZIMA
Napier
2008

Coexpression Analysis



NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

Application to Human Health

Protein Complex Prediction

Application to Model Organisms

NZIMA
Napier
2008

Data Integration

Phenotypic Data (e. g., diseased versus healthy patients)

NZIMA
Napier
2008

Data Integration

Phenotypic Data (e. g., diseased versus healthy patients)

Proteomic Data (e. g., amino acid peaks from mass spec)

NZIMA
Napier
2008

Data Integration

Phenotypic Data (e. g., diseased versus healthy patients)

Proteomic Data (e. g., amino acid peaks from mass spec)

Transcriptomic Data (e.g., gene expression from μarrays)

NZIMA
Napier
2008

Data Integration

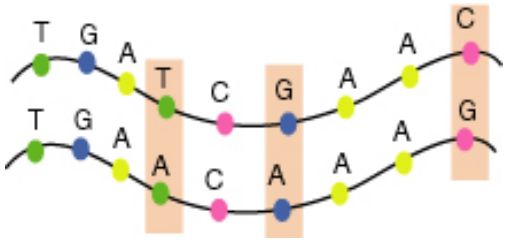
Phenotypic Data (e. g., diseased versus healthy patients)

Proteomic Data (e. g., amino acid peaks from mass spec)

Transcriptomic Data (e.g., gene expression from μarrays)

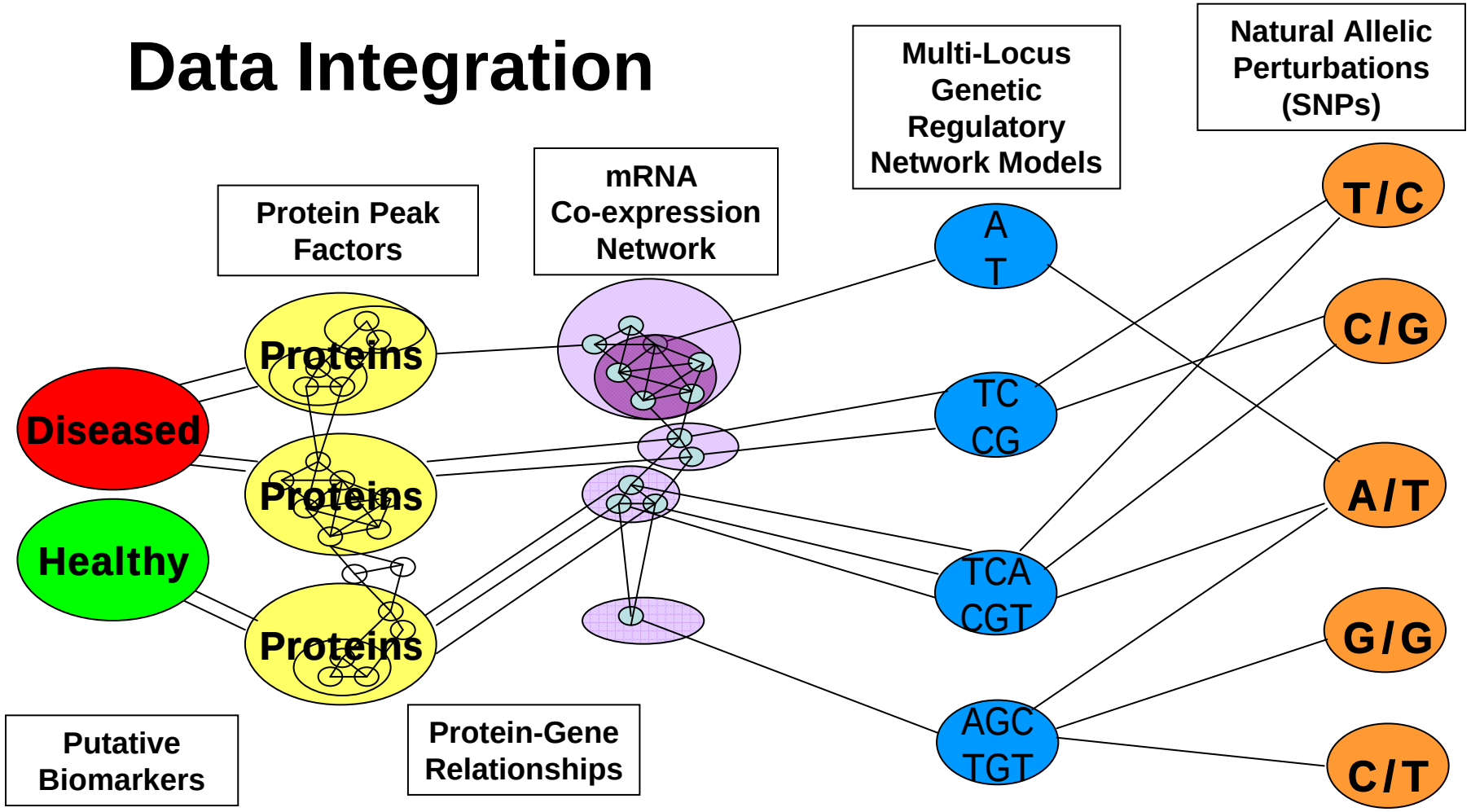
Genotypic Data: SNPs

- **DNA sequence variations, each occurring when a single nucleotide in the genome differs between members of a species**



- **highly conserved throughout evolution and within population**
- **almost always just two alleles**
- **detected with SNP arrays designed to detect polymorphisms**

Data Integration



NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

Application to Human Health

Protein Complex Prediction

Application to Model Organisms

NZIMA
Napier
2008

Application, Human Health

Data Description

- Göteborg, Sweden: 56 patients and 39 controls
- Affymetrix HU133 arrays
- roughly 33,000 genes
- hay fever, eczema
- nasal secretions, lymphocytes, skin

NZIMA
Napier
2008

Application, Human Health

Data Description

- Göteborg, Sweden: 56 patients and 39 controls
- Affymetrix HU133 arrays
- roughly 33,000 genes
- hay fever, eczema
- nasal secretions, lymphocytes, skin

Preprocessing

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

NZIMA
Napier
2008

Application, Human Health

Data Description

- Göteborg, Sweden, 56 patients and 39 controls
- Affymetrix HU133 arrays
- roughly 33,000 genes
- hay fever, eczema
- nasal secretions, lymphocytes, skin

Preprocessing

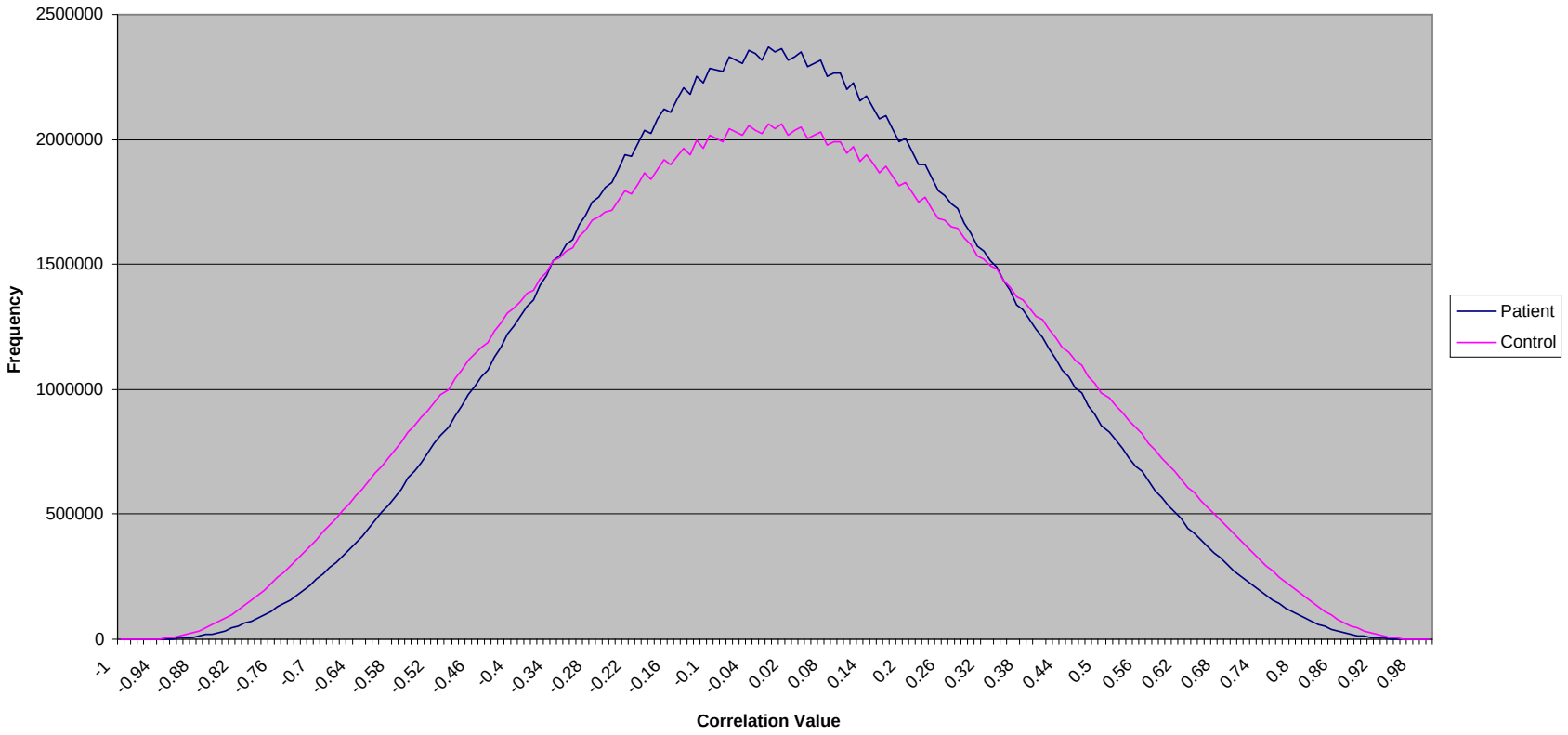
- MAS5.0
- log transformed
- centered around zero with *z* scores
- probesets with consistently low expression levels removed
- replicates averaged

Threshold Selection

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

Application, Human Health

Correlation Coefficient Distribution



Application, Human Health

Graph Properties

Control

Threshold	Vertices	Edges	Maximal Cliques	Maximum Size
0.88	8009	256346	240146378	84
0.89	7169	178144	15067064	79
0.90	6254	118900	1579041	71
0.91	5317	75541	243232	66
0.92	**4415**	**45471**	**51315**	**59**

ribosomal or RNA-related

Patient

Threshold	Vertices	Edges	Maximal Cliques	Maximum Size
0.88	5809	91152	2298595	61
0.89	4999	62271	447176	52
0.90	**4146**	**40933**	**114030**	**45**
0.91	3405	26031	41605	35
0.92	2628	11322	11322	28

T-lymphocytes or epithelial cells

Application, Human Health

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%

Application, Human Health

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%

Of course gene representation is only a small part of the story.

NZIMA
Napier
2008

Application, Human Health

We can use traditional algorithmic tools

- extract cores, cliques and other dense subgraphs
- check for scale-freeness, putative TFs, hubs, etc

NZIMA
Napier
2008

Application, Human Health

We can use traditional algorithmic tools

- extract cores, cliques and other dense subgraphs
- check for scale-freeness, putative TFs, hubs, etc

We can use commercial and other tools

- sort subgraphs by ontological enrichment, CREs, etc
- compare to literature, databases, etc
- match genes and gene products with known interactions

NZIMA
Napier
2008

Application, Human Health

We can use traditional algorithmic tools

- extract cores, cliques and other dense subgraphs
- check for scale-freeness, putative TFs, hubs, etc

We can use commercial and other tools

- sort subgraphs by ontological enrichment, CREs, etc
- compare to literature, databases, etc
- match genes and gene products with known interactions

It's tempting to scan for your favorites...

NZIMA
Napier
2008

Application, Human Health

We can use traditional algorithmic tools

- extract cores, cliques and other dense subgraphs
- check for scale-freeness, putative TFs, hubs, etc

We can use commercial and other tools

- sort subgraphs by ontological enrichment, CREs, etc
- compare to literature, databases, etc
- match genes and gene products with known interactions

It's tempting to scan for your favorites...

But our goal is to identify altered interactions

NZIMA
Napier
2008

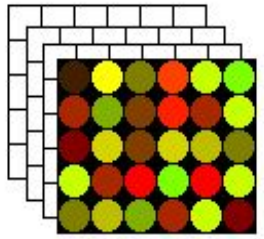
Application, Human Health

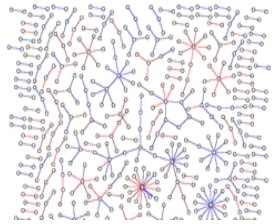
Differential Analysis

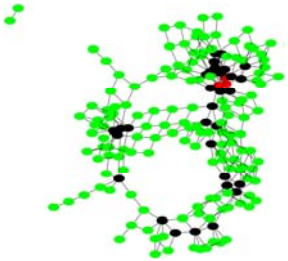
Gene (vertex) comparisons:
- differential expression
- does not require multiple conditions
- compare the two lists of gene expression levels

Correlate (edge) comparisons
- differential correlation
- requires multiple conditions in control, in dose
- compare the two lists of gene-gene correlations

Putative network (clique) comparisons
- differential topology
- focus on network aka clique differences
- consider the clique intersection graph



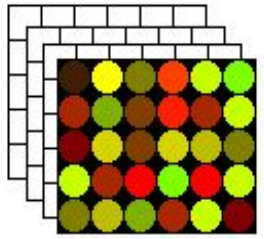


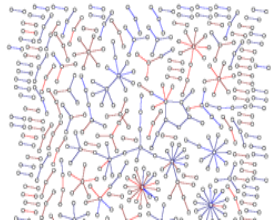


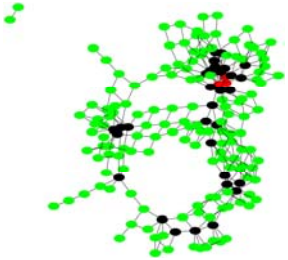
NZIMA
Napier
2008

Application, Human Health

Differential Analysis







Gene (vertex) comparisons:
- differential expression
- does not require multiple conditions
- compare the two lists of gene expression levels

Correlate (edge) comparisons
- differential correlation
- requires multiple conditions in control, in dose
- compare the two lists of gene-gene correlations

Putative network (clique) comparisons
- differential topology
- focus on network aka clique differences
- consider the clique intersection graph

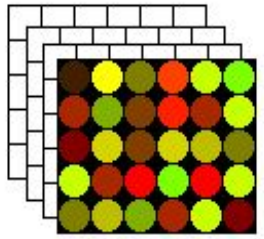
Ongoing Work

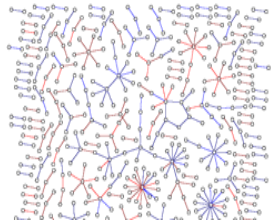
- 62 genes pass all three screens, 6 match a known pathway
- *ITK* (IL2-inducible T-cell kinase), studying in depth...moving to Illumina

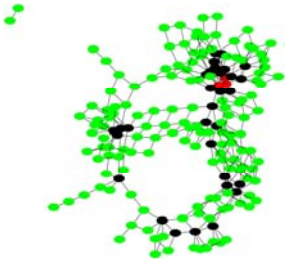
NZIMA
Napier
2008

Application, Human Health

Differential Analysis







Gene (vertex) comparisons:
- differential expression
- does not require multiple conditions
- compare the two lists of gene expression levels

Correlate (edge) comparisons
- differential correlation
- requires multiple conditions in control, in dose
- compare the two lists of gene-gene correlations

Putative network (clique) comparisons
- differential topology
- focus on network aka clique differences
- consider the clique intersection graph

Ongoing Work

- 62 genes pass all three screens, 6 match a known pathway
- *ITK* (IL2-inducible T-cell kinase), studying in depth...moving to Illumina

For Impact

- concentrate on real data, and working with bench biologists
- strategic publications (e.g., *Nature Genetics, PLoS Comp Bio*, etc)

NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

Application to Human Health

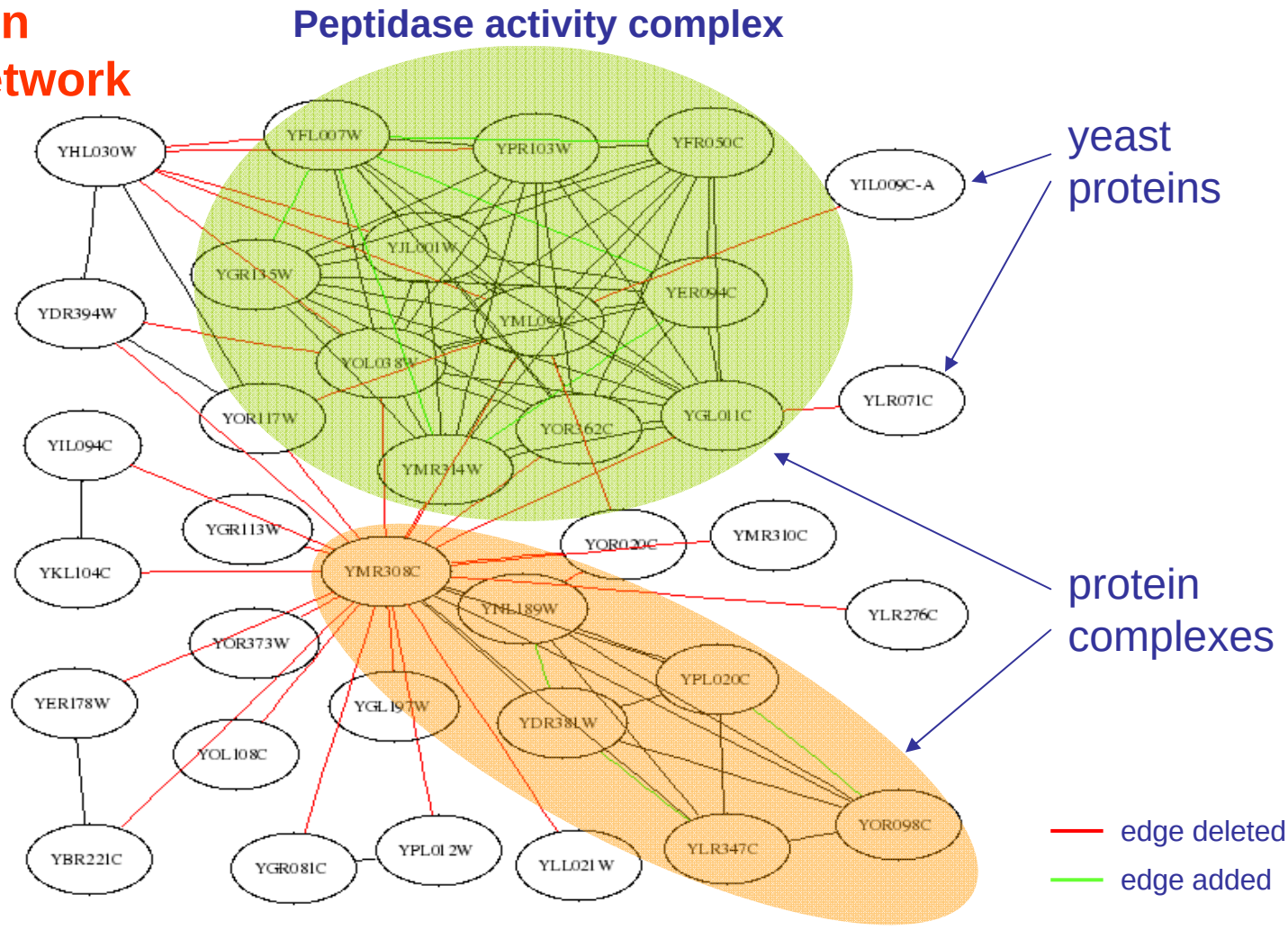
Protein Complex Prediction

Application to Model Organisms

NZIMA
Napier
2008

Protein Complex Prediction

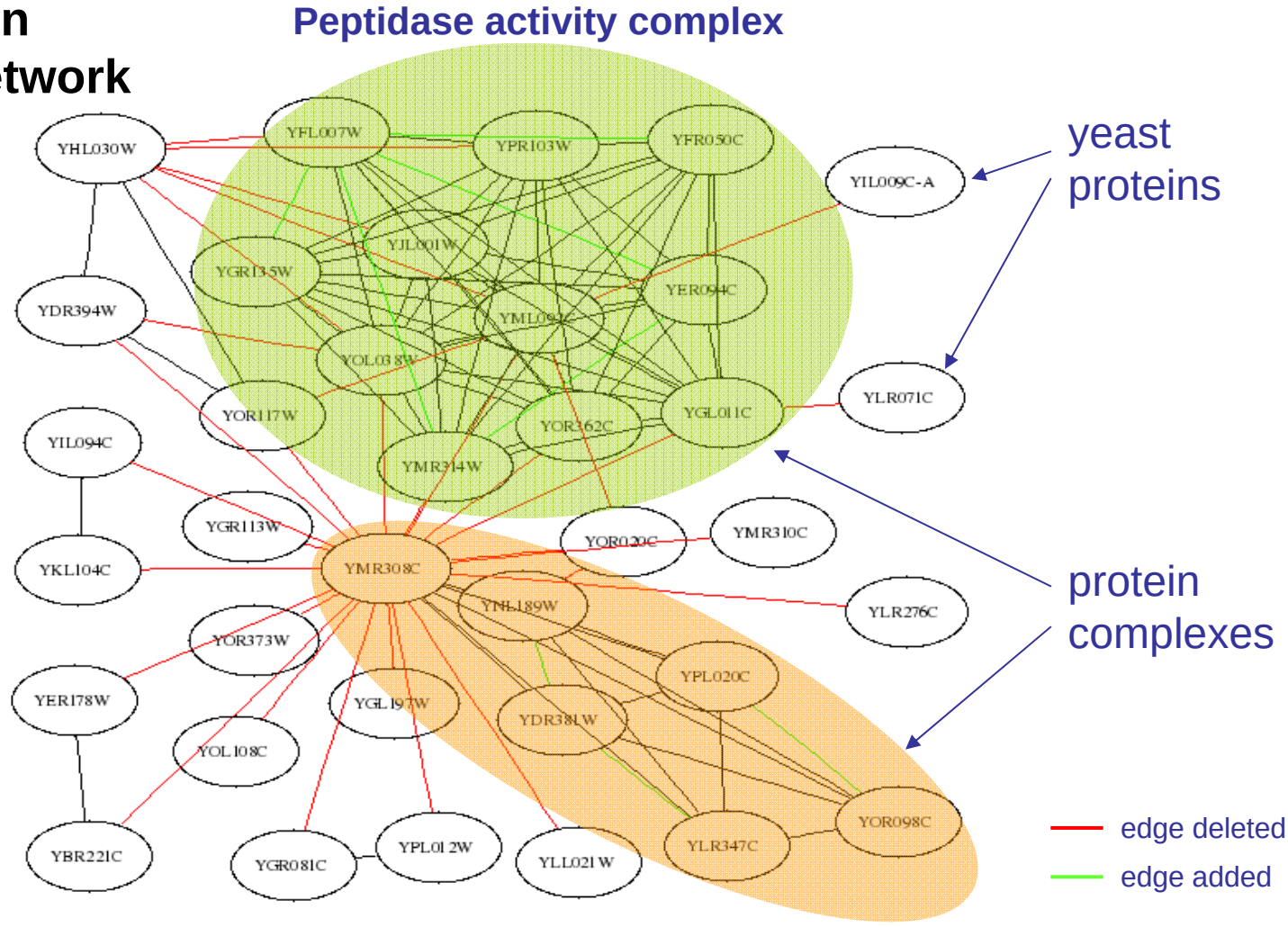
Protein-Protein Interaction Network



NZIMA
Napier
2008

Protein Complex Prediction

Protein-Protein Interaction Network



Recognize as Cluster Editing

NZIMA
Napier
2008

Protein Complex Prediction

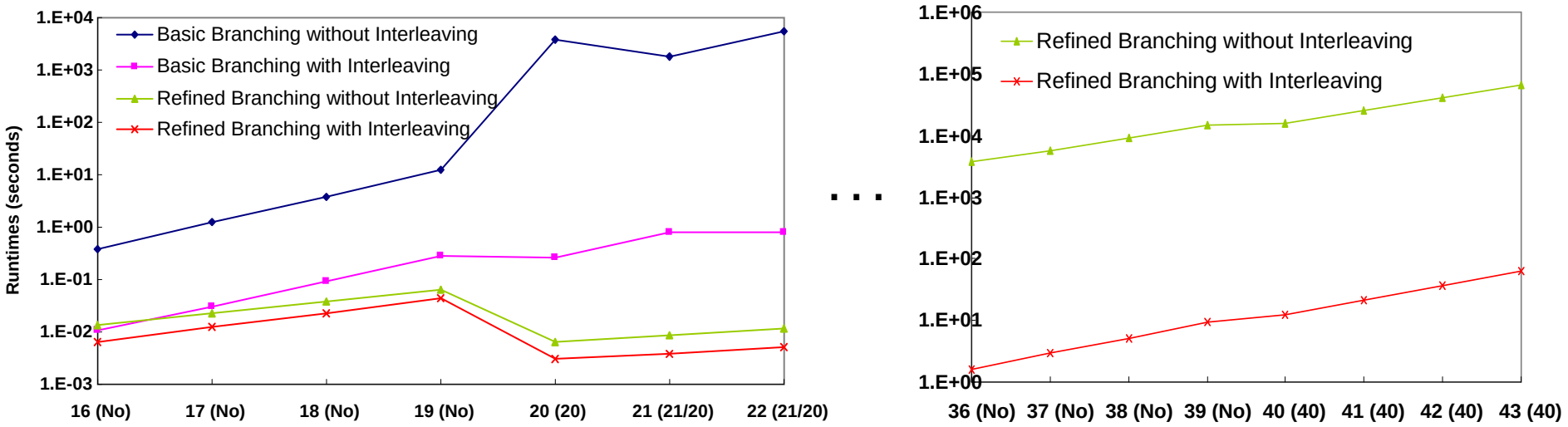
Computational Experience

- algorithms studied by Guo, Niedermeier, Damaschke, others
- synthetic graphs
  - known edit distances
  - various sizes, densities and distances

Protein Complex Prediction

Computational Experience

- non-monotonic behavior
- importance of interleaving
- benefits of refinement

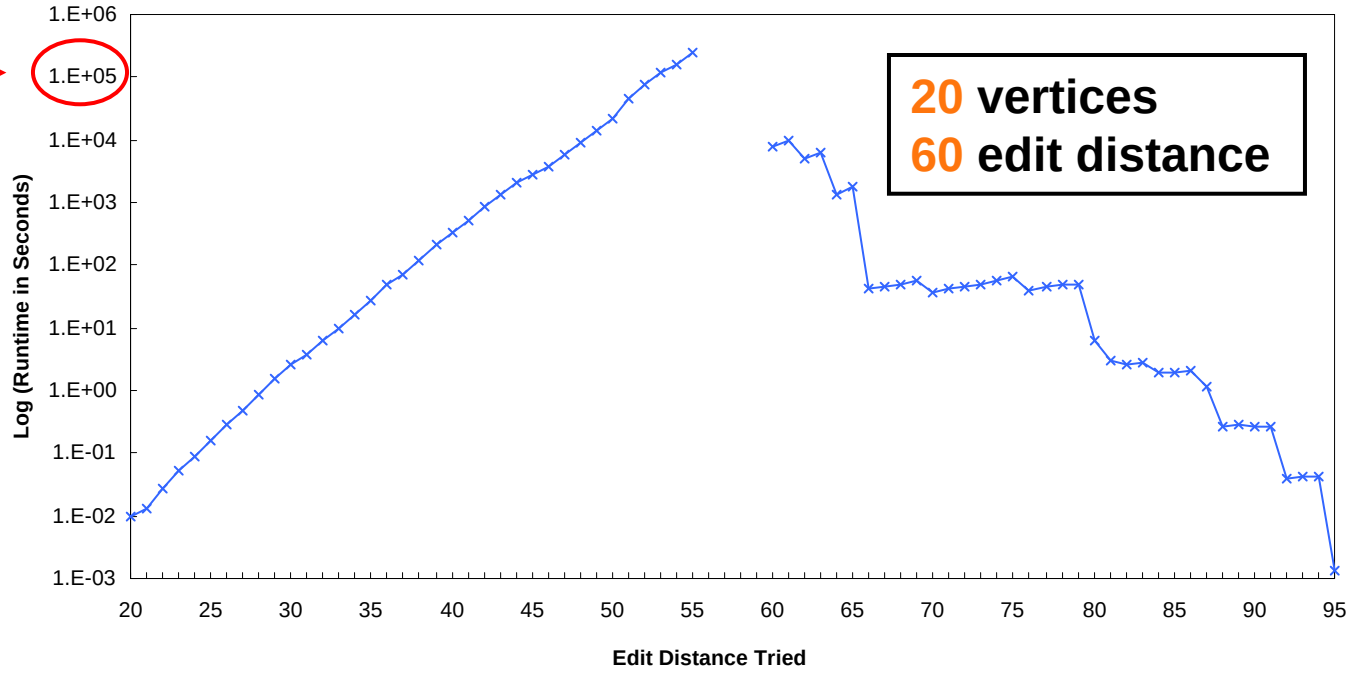


Edit distance tried (found)

Protein Complex Prediction

Nice application, but best methods still too slow

27+ hours →



No instances Yes instances

NZIMA
Napier
2008

Outline of Talk

Foundations

Gene Coexpression Analysis

Data Integration

Application to Human Health

Protein Complex Prediction

Application to Model Organisms

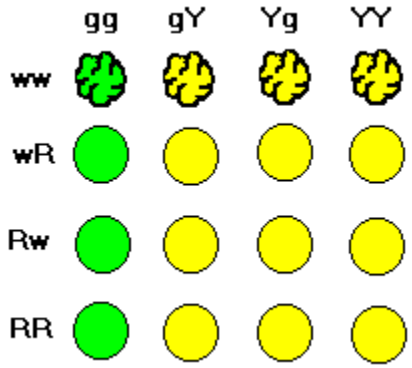
NZIMA
Napier
2008

Application, Model Organisms

Gregor Mendel, 1822-1884

pea experiments

- green vs yellow
- round vs wrinkly



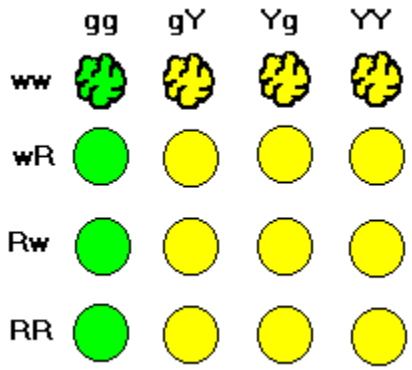
- inheritance, dominant and recessive traits (alleles)
- monogenetic phenotypes
- very “lucky”

NZIMA
Napier
2008

Application, Model Organisms

Gregor Meldel, pea experiments

- green vs yellow
- round vs wrinkly



- inheritance, dominance, monogenetic phenotypes
- but most traits appear to be “complex” (polygenetic)
- many allelic combinations convey evolutionary (dis)advantage
- simple rules of Mendelian inheritance do not apply
- need a measure of independence: Linkage Disequilibrium (LD)

NZIMA
Napier
2008

Application, Model Organisms

LD: a measure of statistical dependence between genetic markers

- non-random association of alleles at two or more loci
- the occurrence in a population of two linked alleles at a frequency higher or lower than expected on the basis of the individual frequencies
- not necessarily on the same chromosome

NZIMA
Napier
2008

Application, Model Organisms

LD: a measure of statistical dependence between genetic markers

- non-random association of alleles at two or more loci
- the occurrence in a population of two linked alleles at a frequency higher or lower than expected on the basis of the individual frequencies
- not necessarily on the same chromosome

Reflects biologically meaningful association of loci

NZIMA
Napier
2008

Application, Model Organisms

LD: a measure of statistical dependence between genetic markers

- non-random association of alleles at two or more loci
- the occurrence in a population of two linked alleles at a frequency higher or lower than expected on the basis of the individual frequencies
- not necessarily on the same chromosome

Reflects biologically meaningful association of loci

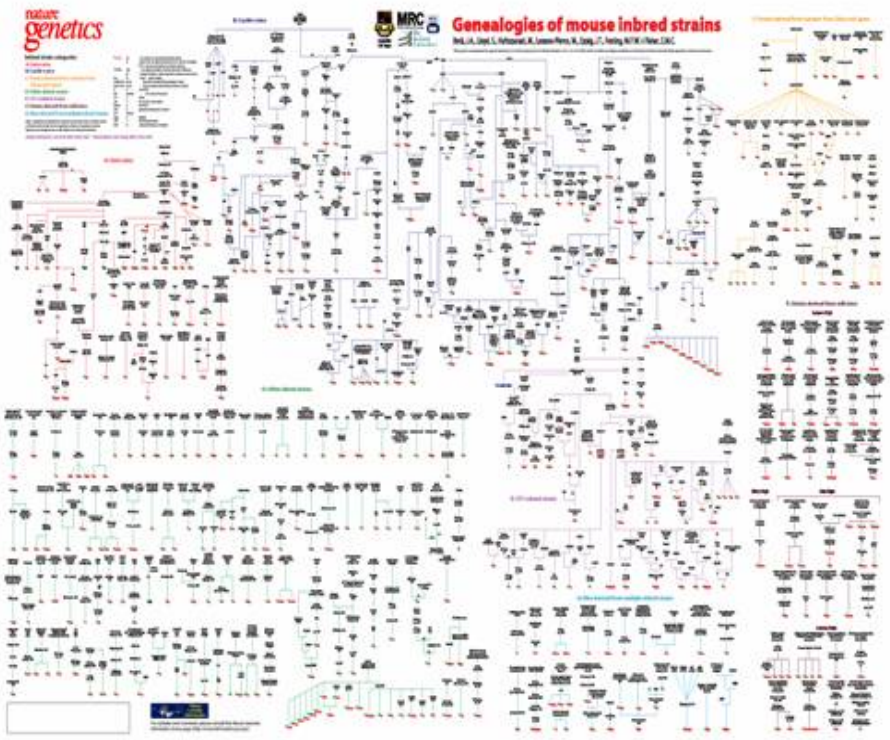
Generally a result of population history

- population genealogy
- recombination frequency
- co-adaptive allele selection
- natural selection
- other factors

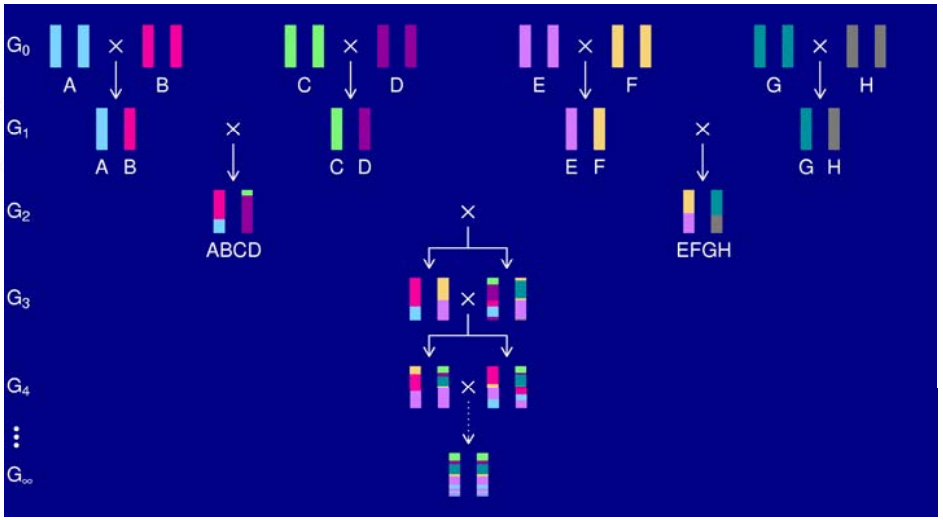
Application, Model Organisms

Evaluation of *Mus musculus* breeding strategies

Standard Inbred (SI)



Recombinant Inbred (RI)
BXD, LXS, etc
The Collaborative Cross

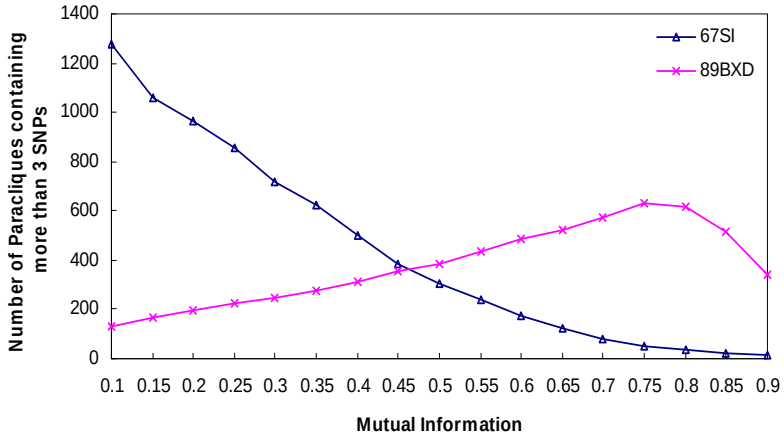


Solution: Use SNPs, correlation, paraclique and proximity

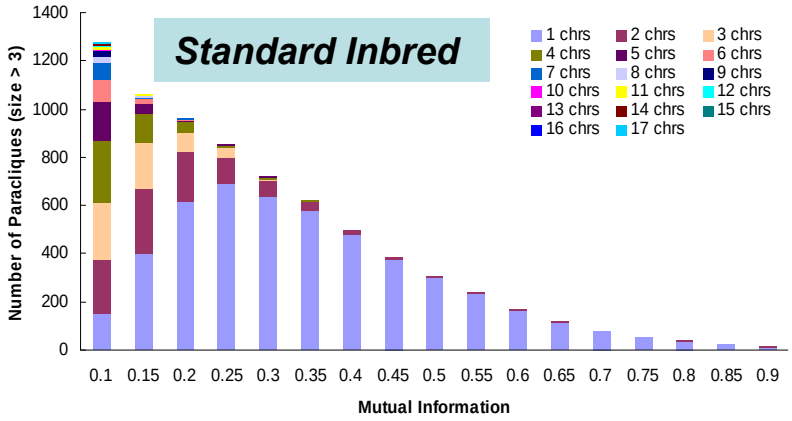
NZIMA
Napier
2008

Application, Model Organisms

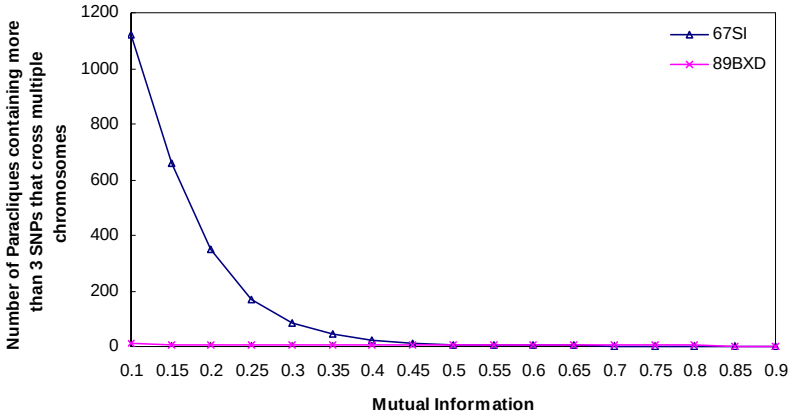
Number of LD Networks

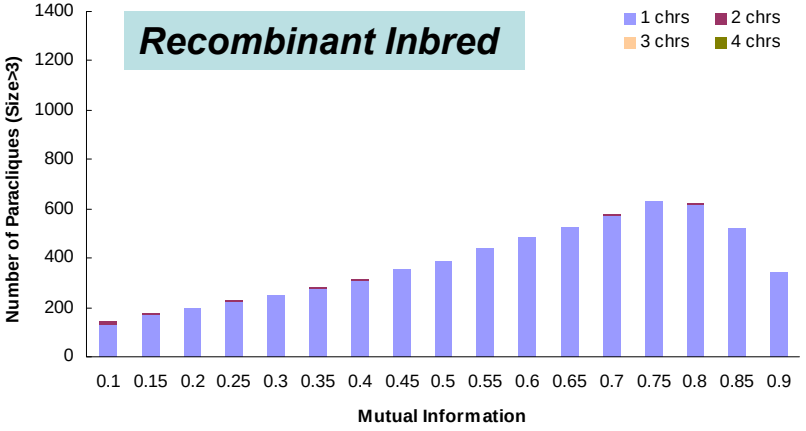


Chromosome Coverage



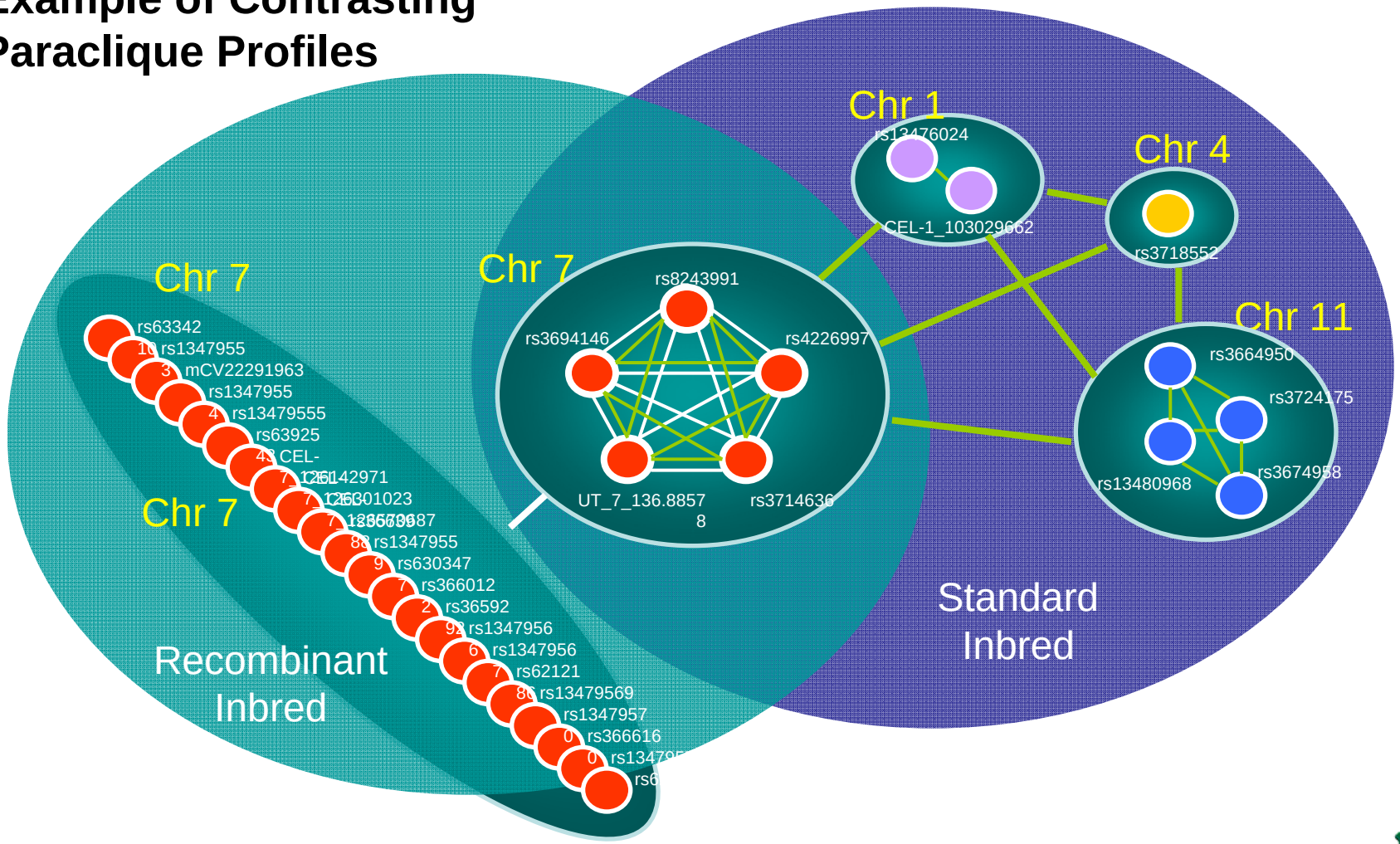
Number of Non-Syntenic LD Networks





Application, Model Organisms

Example of Contrasting Paraclique Profiles



NZIMA
Napier
2008

Collaborators

Research Scientists (Incomplete!):

Mikael Benson
Elissa Chesler
Frank Dehne
Mike Fellows
Ivan Gerling
Dan Goldowitz
Malak Kotb
Mark Ragan
Arnold Saxton
Brynn Voy
Rob Williams
Bing Zhang

Current Students:

Bhavesh Borate
Patricia Carey
John Eblen
Jeremy Jay
Zuopan Li
Sudhir Naswa
Andy Perkins
Vivek Philip
Charles Phillips
Gary Rogers
Jon Scharff
Yun Zhang

NZIMA
Napier
2008

Geeks Я Us