Discovery & Modeling of Genomic Regulatory Networks with Big Data

Hamid Bolouri
Division of Human Biology
Fred Hutchinson Cancer Research Center

labs.fhcrc.org/bolouri

I have no financial relationships with any commercial entities.

Frontiers in Laboratory Science

Overview:

(1) What do I mean by Genomic Regulatory Networks (GRNs)?

- Why go genome-wide?
- Why focus on regulatory interactions?

(2) What do I mean by big (molecular biology) data?

(3) Overview of integrative network discovery/modeling

(4) Examples of available bioinformatics resources

(5) Limitations and common pitfalls

(6) Conclusion: *Proceed With Caution!*

- What are Genomic Regulatory Networks?

  Molecular interaction networks regulating gene expression

- Why ***genomic***?

  Genome-wide for unbiased discovery of network components

  Cost-effective exploration via high-throughput sequencing

- Why focus on ***regulatory*** interactions?

  Determinants of cellular identity and state

- What are the main challenges?

  Complex, intricate (much detail), and large-scale networks

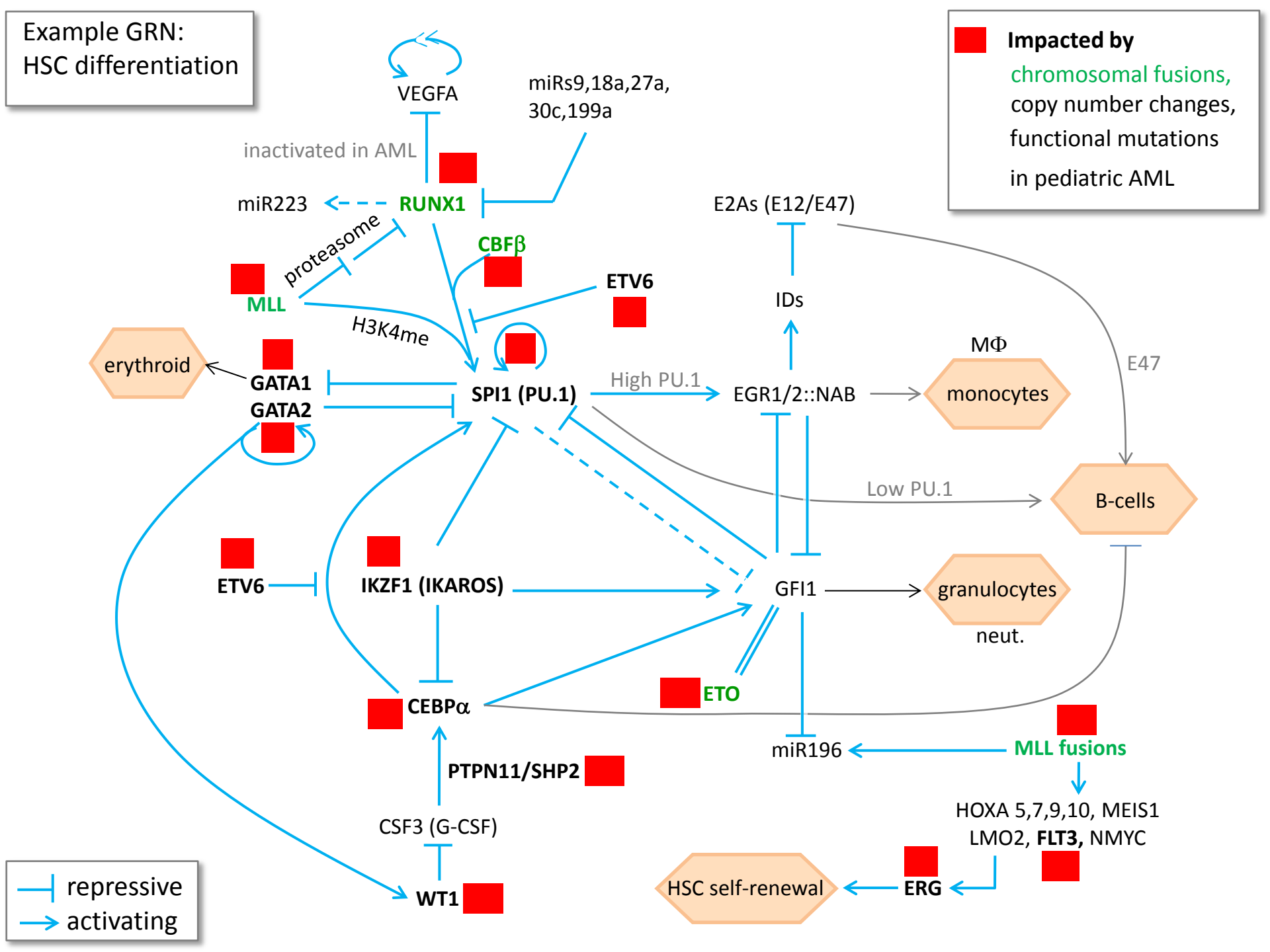
  Large-scale, noisy data

Example regulators of gene expression

- ❑ Enhancers, promoters, silencers, insulators/boundary elements
- ❑ Transcription factor state, concentration, localization, binding partners
- ❑ DNA methylation
- ❑ Histone/chromatin state regulation
- ❑ Histone variants
- ❑ Nucleosome remodeling
- ❑ 3D DNA conformation, nuclear localization
- ❑ miRNAs, other ncRNAs

Example processes regulating gene expression

- ❑ Chromatin accessibility
- ❑ TF complex occupancy
- ❑ Rate of RNA polymerase complex formation/transcription initiation
- ❑ Abortive & paused transcription initiation, elongation
- ❑ RNA processing (splicing, capping, editing, transport, etc.)
- ❑ Translation initiation and elongation
- ❑ Post-translational modifications


Example GRN:
HSC differentiation
Impacted by
chromosomal fusions,
copy number changes,
functional mutations
in pediatric AML
VEGFA
miRs9,18a,27a,
30c,199a
inactivated in AML
miR223
RUNX1
proteasome
CBFβ
MLL
ETV6
H3K4me
E2As (E12/E47)
IDs
MΦ
E47
erythroid
GATA1
GATA2
SPI1 (PU.1)
High PU.1
EGR1/2::NAB
monocytes
Low PU.1
B-cells
ETV6
IKZF1 (IKAROS)
GFI1
granulocytes
neut.
ETO
CEBPα
miR196
MLL fusions
PTPN11/SHP2
HOXA 5,7,9,10, MEIS1
LMO2, FLT3, NMYC
CSF3 (G-CSF)
HSC self-renewal
ERG
WT1
repressive
activating

Example sources of big data for GRN discovery and analysis

Public datasets:

- ✓ Gene Expression Omnibus (GEO): **1,335,996 samples** (ncbi.nlm.nih.gov/geo/). See also GTEXportal.org
- ✓ 1000 Genomes Project: DNA sequence variants in **2,504 individuals** from 26 populations (www.1000genomes.org). See also NHLBI exome seq project: (evs.gs.washington.edu, **6,500 exomes**)
- ✓ Regulatory regions & interactions in DNA: ENCODE (**human 5,040 samples, mouse, 667 samples**) www.encodeproject.org/. See also modENCODE (fly, worm, www.modencode.org/)
- ✓ Roadmap Epigenomics Consortium. DNA methylation, histone mods, chromatin accessibility & small RNA transcripts in 23 stem cells and primary ex vivo tissues (**3,135 samples**, www.roadmapepigenomics.org)
- ✓ Protein localization in tissues (www.proteinatlas.org/), & organelles (locate.imb.uq.edu.au/)
- ✓ The Cancer Cell Line Encyclopedia (CCLE): sequence, expression, drug sensitivity in **~ 1000 cell lines** www.broadinstitute.org/ccle . See also: www.cancerrxgene.org/ , discover.nci.nih.gov/cellminer , cancer.sanger.ac.uk/cancergenome/projects/cell_lines/
- ✓ NCI TCGA, TARGET projects. Comprehensive multi-omics and clinical patient data (see also ICGC).

Your (genome-wide) data:

- ✓ DNA-seq (whole genome / exome / targeted sequencing)
- ✓ RNA-seq (mRNA-seq, miRNA-seq, ribosome profiling)
- ✓ ChIP-seq (TFs, pol2, histone modifications, methylated-DNA seq). Also DNase/MNase seq, FAIRE, etc.
- ✓ Chromatin interactions, location, …

Examples of combining in-lab and 3rd party data to discover GRNs

Use data from related cell types to augment/test in-lab findings, e.g.:

- ❑ Correlate mRNA data with 3rd party
  - miRNA expression
  - miR binding motifs in UTR sequences
  - DNA-methylation
- ❑ Match up-regulated transcription factors (TFs) to
  - TF ChIP-seq peaks
  - TF binding motifs in open chromatin (e.g. DNAse1 HS / footprints) regions
- ❑ Match DNA-sequence variations to
  - Changes in mRNA isoforms
  - mRNA/miRNA expression
  - Changes in correlated gene expression

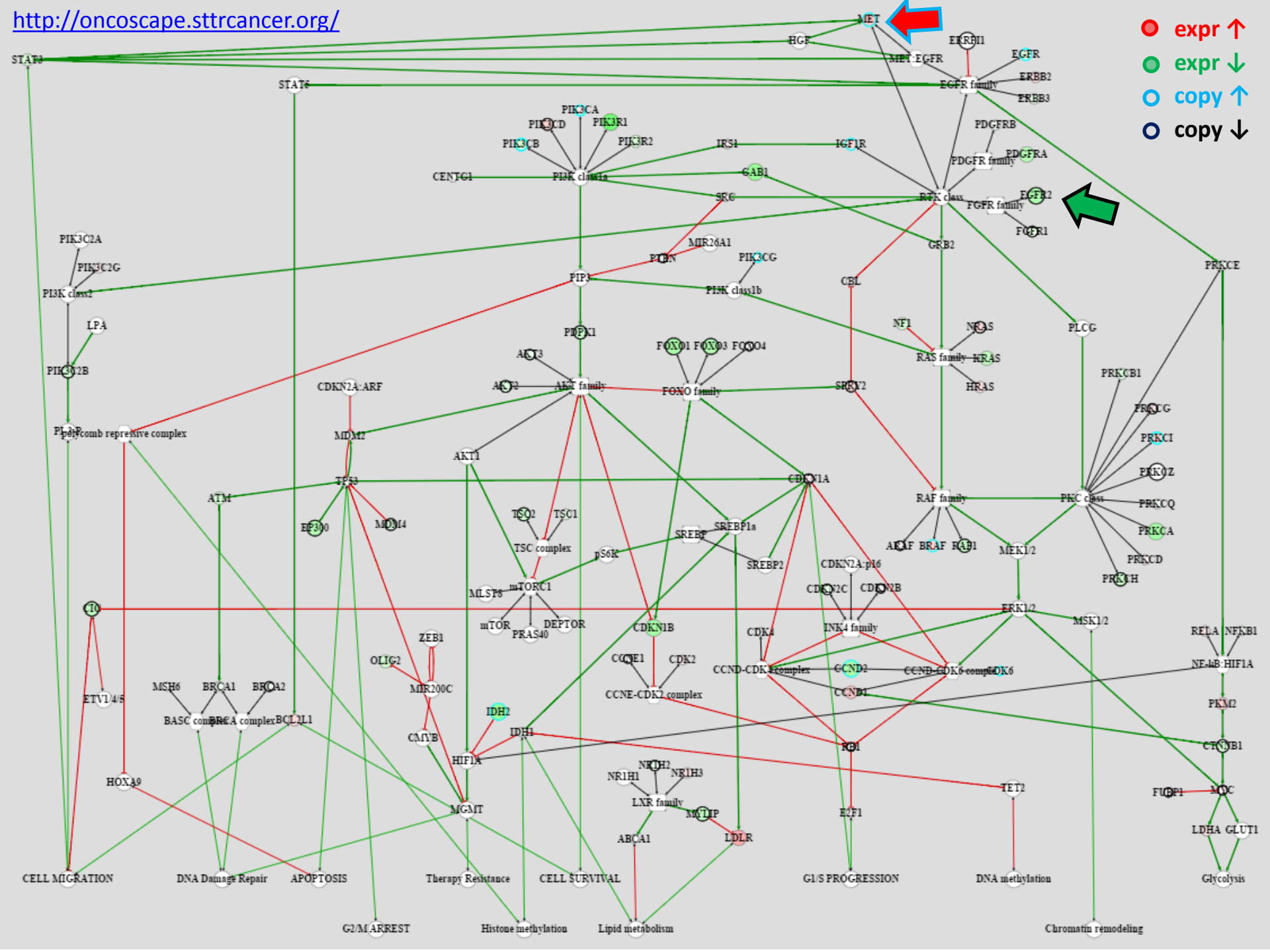
GRN discovery/modeling example approach (1):

visual exploration of ‘omics data supper-imposed on an interaction/pathway map

Example software:

Cytoscape: cytoscape.org (see also http://js.cytoscape.org)

BipTapestry: http://www.biotapestry.org/#download


http://oncoscape.sttrcancer.org/
expr ↑
expr ↓
copy ↑
copy ↓
MET
HGF
ERBB1
MET:EGFR
EGFR
ERBB2
ERBB3
EGFR family
STAT3
STAT6
PIK3CA
PIK3CD
PIK3R1
PIK3CB
PIK3R2
IRS1
IGF1R
PDGFRB
PDGFRA
PDGFR family
CENTG1
PI3K class1a
GAB1
SRC
RTK class
FGFR family
FGFR2
FGFR1
PIK3C2A
PIK3C2G
MIR26A1
PTEN
PIK3CG
GRB2
PRKCE
PIP3
PI3K class1b
CBL
PI3K class2
LPA
PIK3C2B
PDPK1
NF1
NRAS
PLCG
FOXO1
FOXO3
FOXO4
RAS family
KRAS
AKT3
HRAS
PRKCB1
CDKN2A:ARF
AKT2
AKT family
FOXO family
SRC2
PRKCG
PI3P
polycomb repressive complex
MDM2
PRKCI
AKT1
PRKCZ
TP53
CDKN1A
RAF family
PKC class
ATM
PRKCQ
TSC2
TSC1
EP300
MDM4
PRKCA
SREBP
SREBP1a
TSC complex
ARAF
BRAF
RAF1
MEK1/2
PRKCD
pS6K
SREBP2
PRKCH
CDKN2A:p16
CDKN2C
CDKN2B
MLST8
mTORC1
ERK1/2
TP
mTOR
PRAS40
DEPTOR
CDKN1B
INK4 family
MSK1/2
RELA
NFKB1
ZEB1
CDK4
OLIG2
CCNE1
CDK2
NF-kB
HIF1A
CCND-CDK4 complex
CCND2
CCND-CDK6 complex
CDK6
MIR200C
CCNE-CDK2 complex
CCNB1
PKM2
MSH6
BRCA1
BRCA2
ETV1/4/5
IDH2
BASC complex
BRCA complex
BCL2L1
CTNNB1
CMYB
IDH1
HIF1A
RB1
HOXA9
TET2
NR1H2
NR1H1
NR1H3
FUBP1
MYC
LXR family
MYLIP
E2F1
ABCA1
LDLR
LDHA
GLUT1
CELL MIGRATION
DNA Damage Repair
APOPTOSIS
Therapy Resistance
CELL SURVIVAL
G1/S PROGRESSION
DNA methylation
Glycolysis
G2/M ARREST
Histone methylation
Lipid metabolism
Chromatin remodeling

GRN discovery/modeling example approach (2):

GRN inference by integration of complementary ‘omics data sets

Example software:

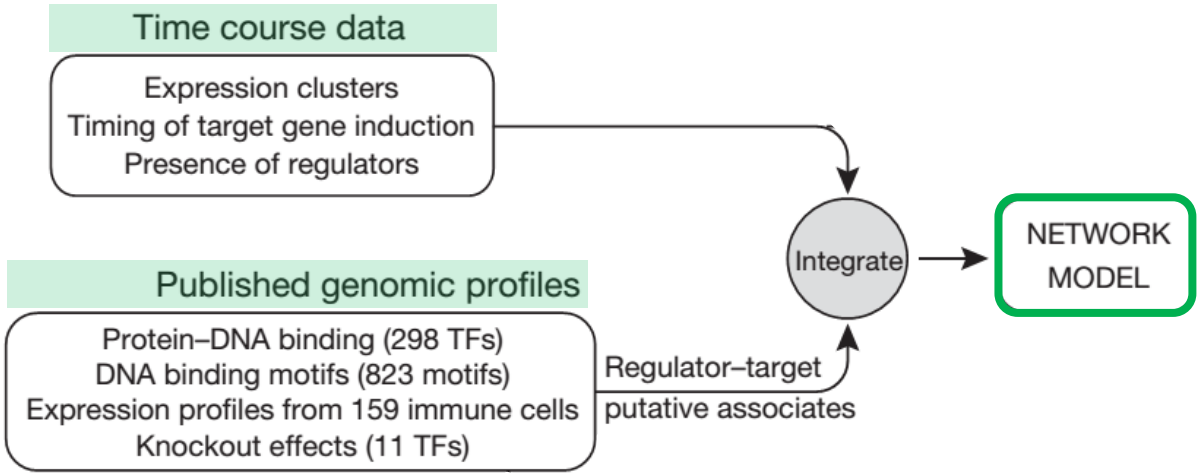
PARADIGM, Vaske, Benz et al, Bioinformatics (2010) 26 (12):i237-i245.

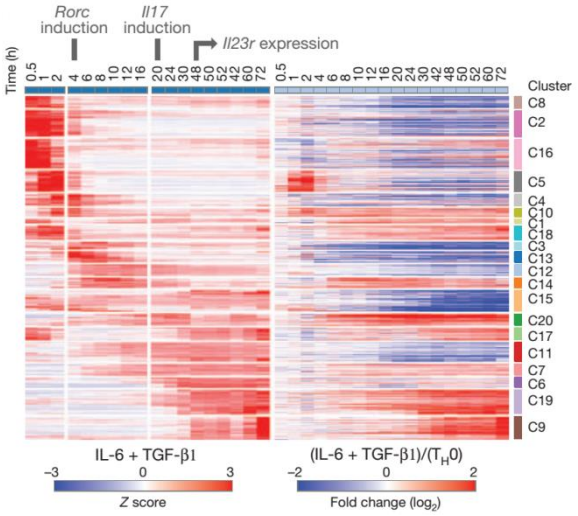
sbenz.github.com/Paradigm

Dynamic regulatory network controlling T_H17 cell differentiation

25 APRIL 2013 | VOL 496 | NATURE | 461

Nir Yosef, Aviv Regev & colleagues





mRNA profiling → Modelling → Perturbation (siRNAs, nanowire delivery) → Validated model

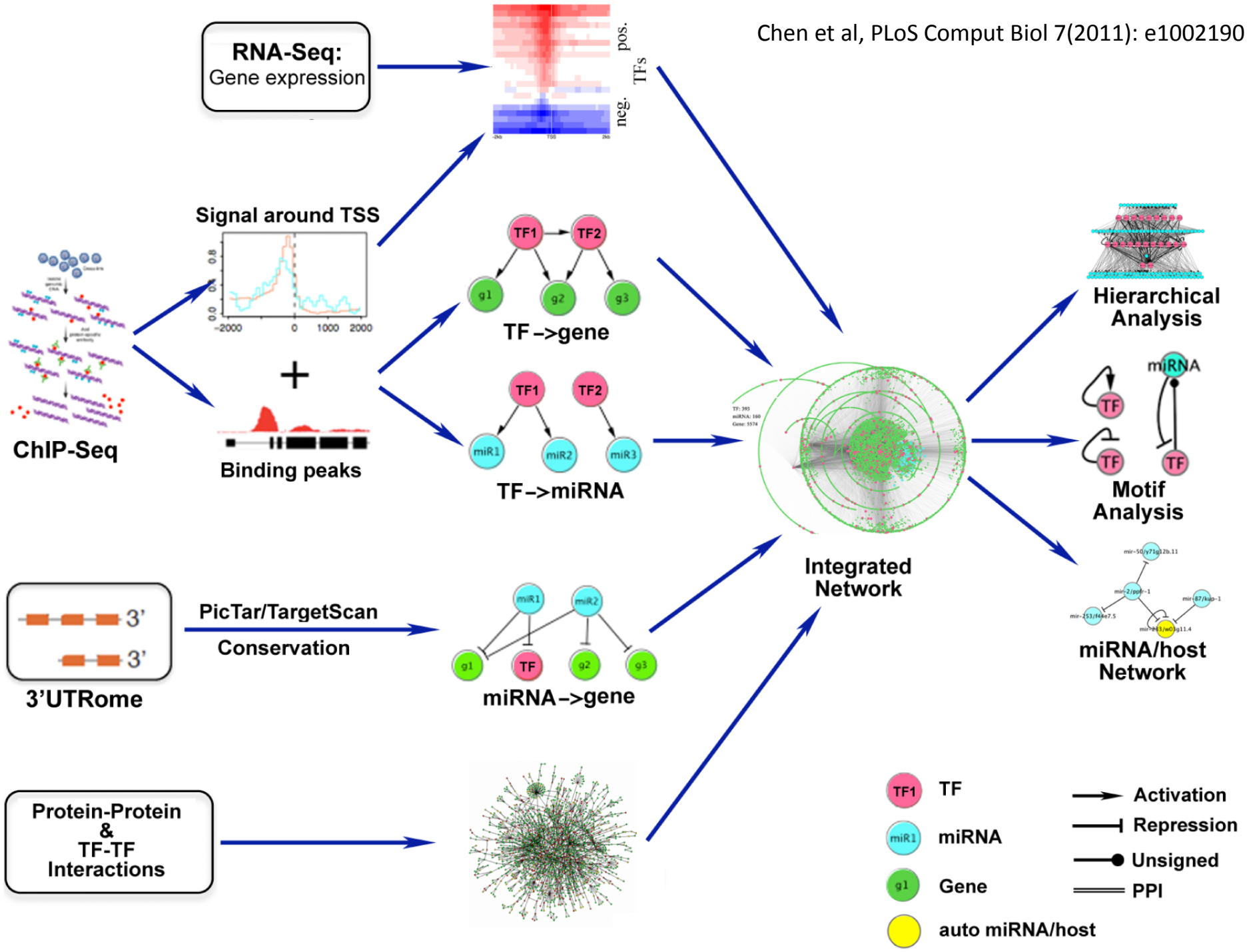
18 time points (0.5–72 h; ± IL-23: 48–72 h)

3 temporal networks; 71 regulators; ~1,200 target genes

29 transcriptional regulators; 10 receptors

22 T_H17-positive regulators antagonistically coupled to 5 T_H17-negative regulators

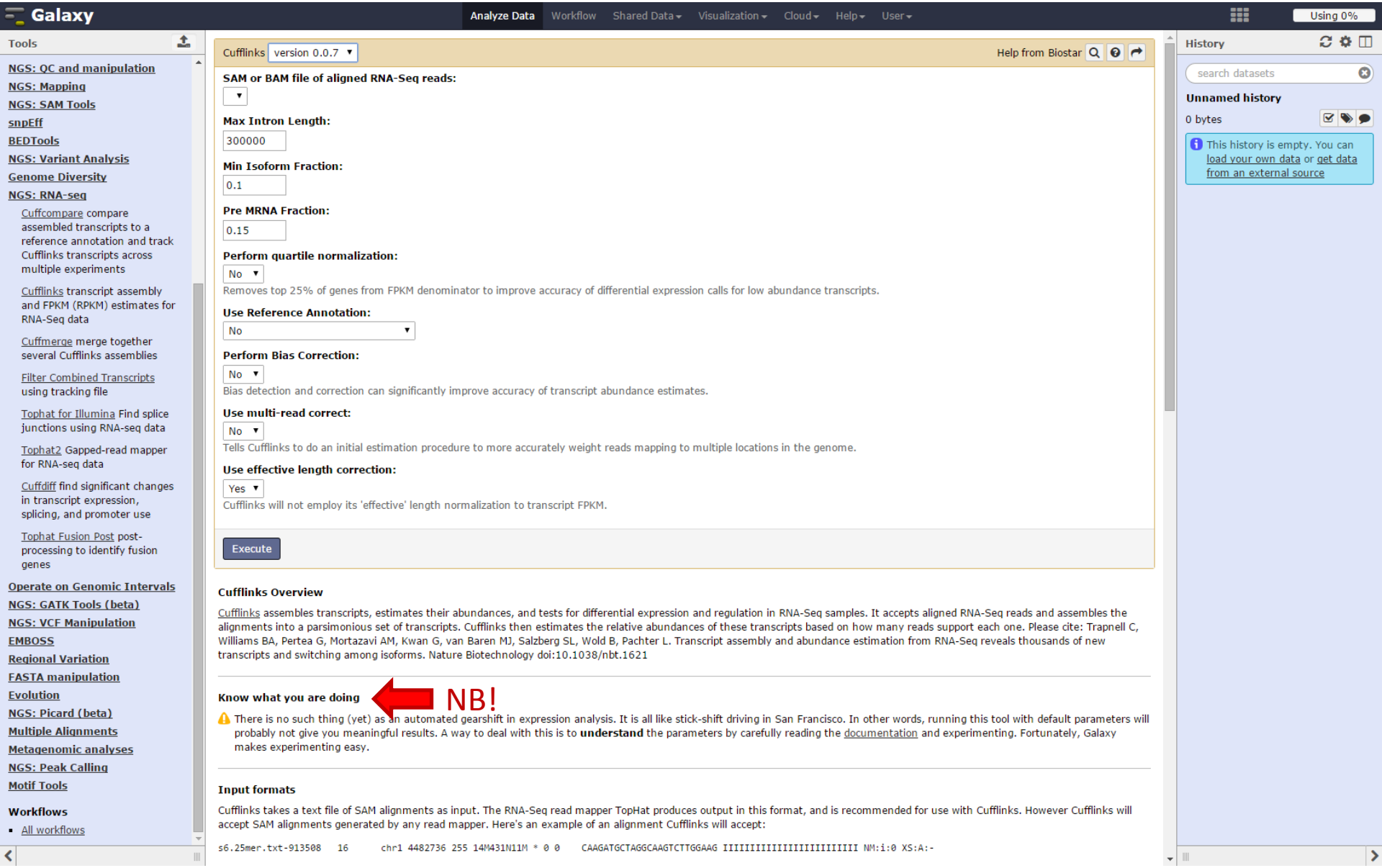
Chen et al, PLoS Comput Biol 7(2011): e1002190



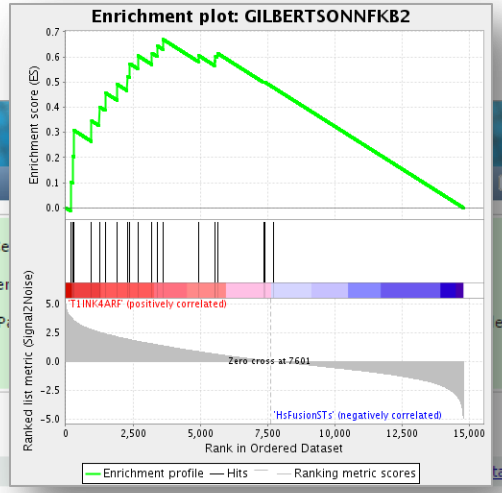
Examples of available bioinformatics resources

… for people who do not write Perl, Python, R (Bioconductor), etc.

https://usegalaxy.org



http://genepattern.broadinstitute.org/



GenePattern

hbolouri

Modules & Pipelines | Suites | Job Results | Resources | Help | GenomeSpace

8.3 GB / 30 GB

Modules | Jobs | Files

Search Modules & Pipelines

No Jobs Processing

Browse Modules ›

Favorite Modules

ssGSEAProjection

Recent Modules

GSEA

GSEALeadingEdgeViewer

-- *Feb 11* -- **FCSNormalization** v2 now available (via GParc and BC Cancer Research Ce

-- *Jan 20* -- **IGV** updated: addresses previous issues with launching IGV from GenePatter

-- *9:10am Jan 20* -- The software update to 3.9.1 is complete. The "what's new in GeneP

gp-help with your questions and comments.
Thanks!

ease contact us at

GSEA version 14

tation

Gene Set Enrichment Analysis

* required field

Reset | Run

expression dataset*

Upload File... | Add Path or URL... Drag Files Here

Batch

2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.

Dataset file - .res, .gct

gene sets database

Gene sets database from GSEA website.

gene sets database file

Upload File... | Add Path or URL... Drag Files Here

Batch

2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.

Gene sets database - .gmt, .gmx, .grp. Upload a gene set if your gene set is not listed as a choice for the gene sets database parameter.

number of permutations*

1000

Number of permutations to perform

phenotype labels*

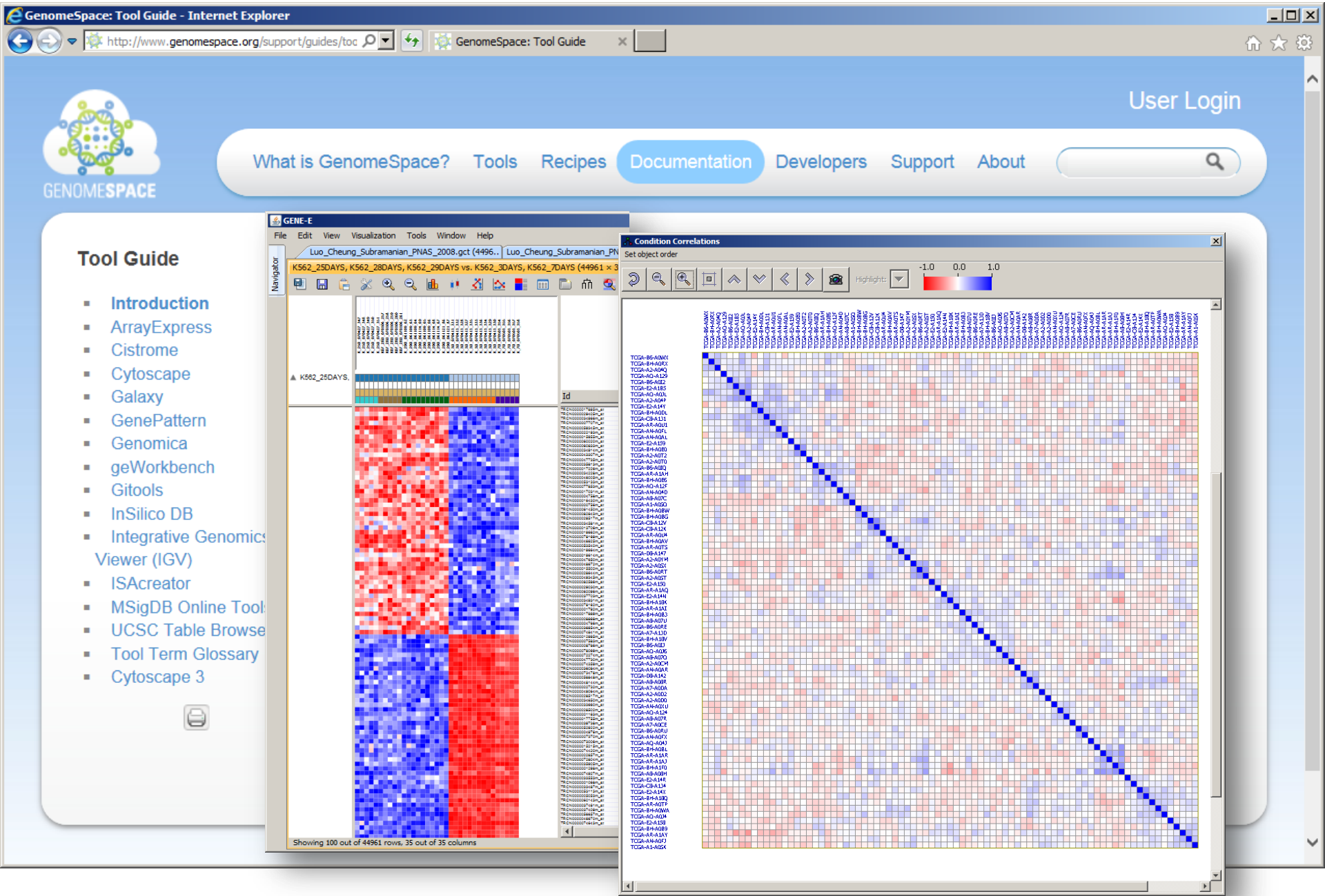
Upload File... | Add Path or URL... Drag Files Here

Batch

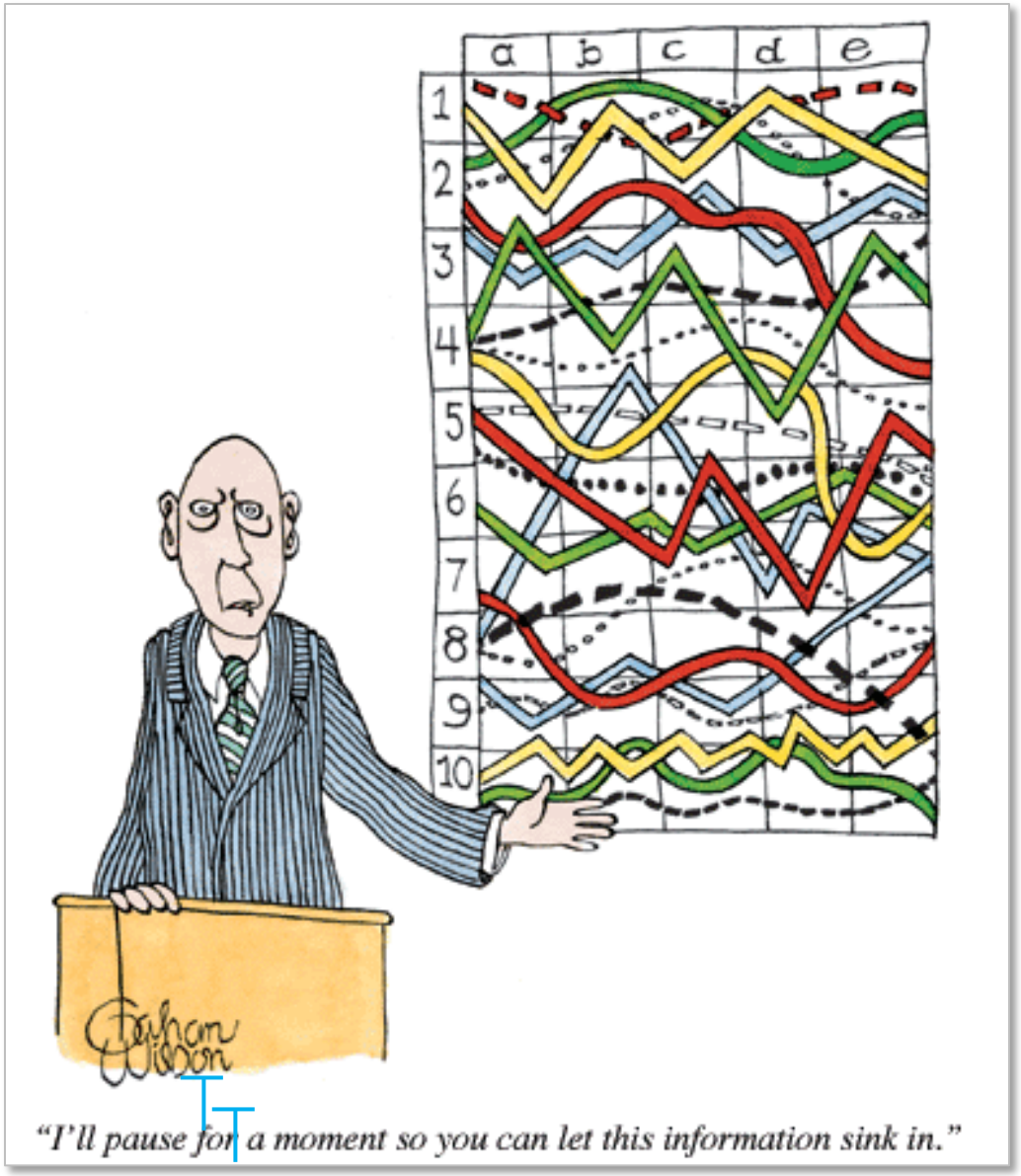
2GB file upload limit using the Upload File... button. For files > 2GB upload from the Files tab.

About GenePattern | Contact Us

http://genomespace.org/



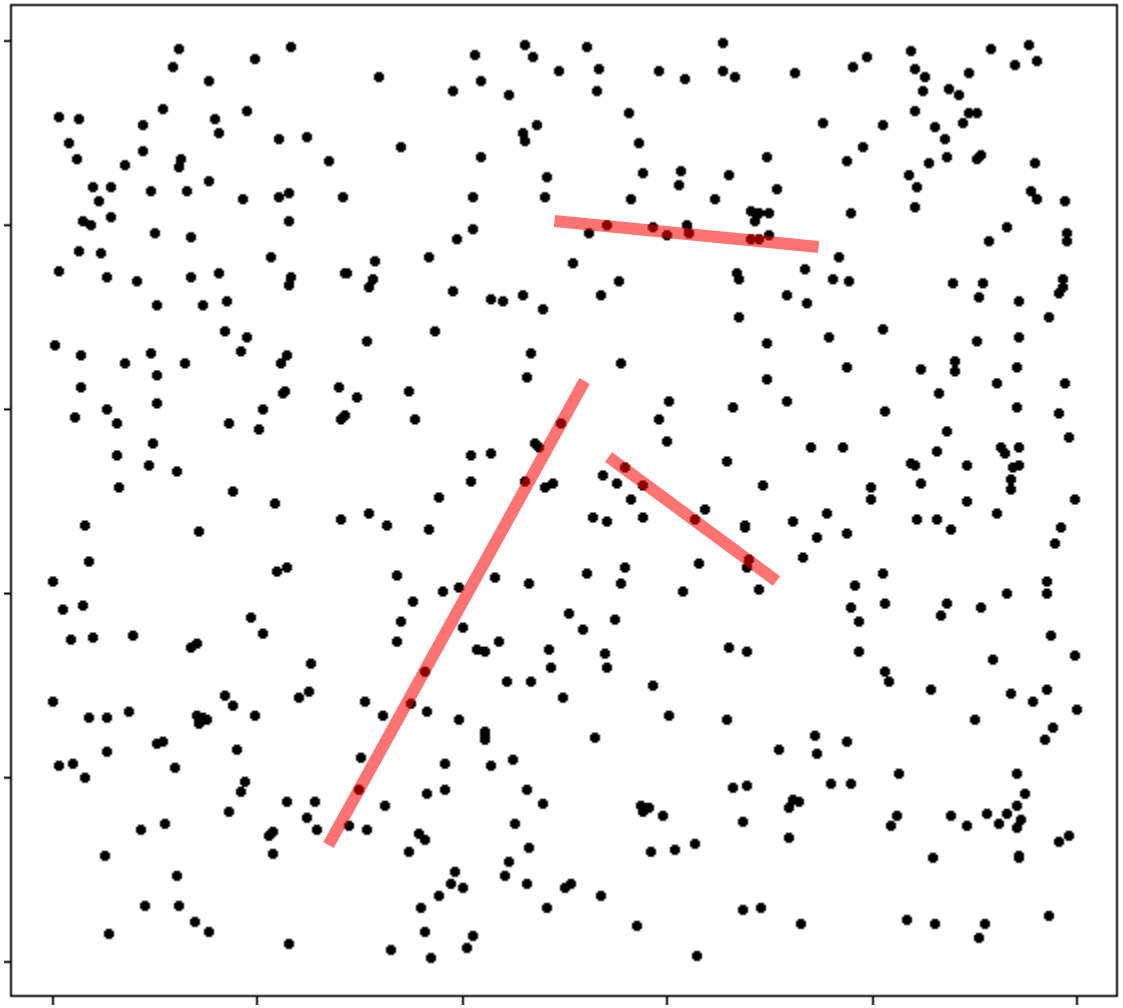
Limitations and common pitfalls



"I'll pause for a moment so you can let this information sink in."

In large-scale (e.g. genome-wide) data, patterns will occur by chance

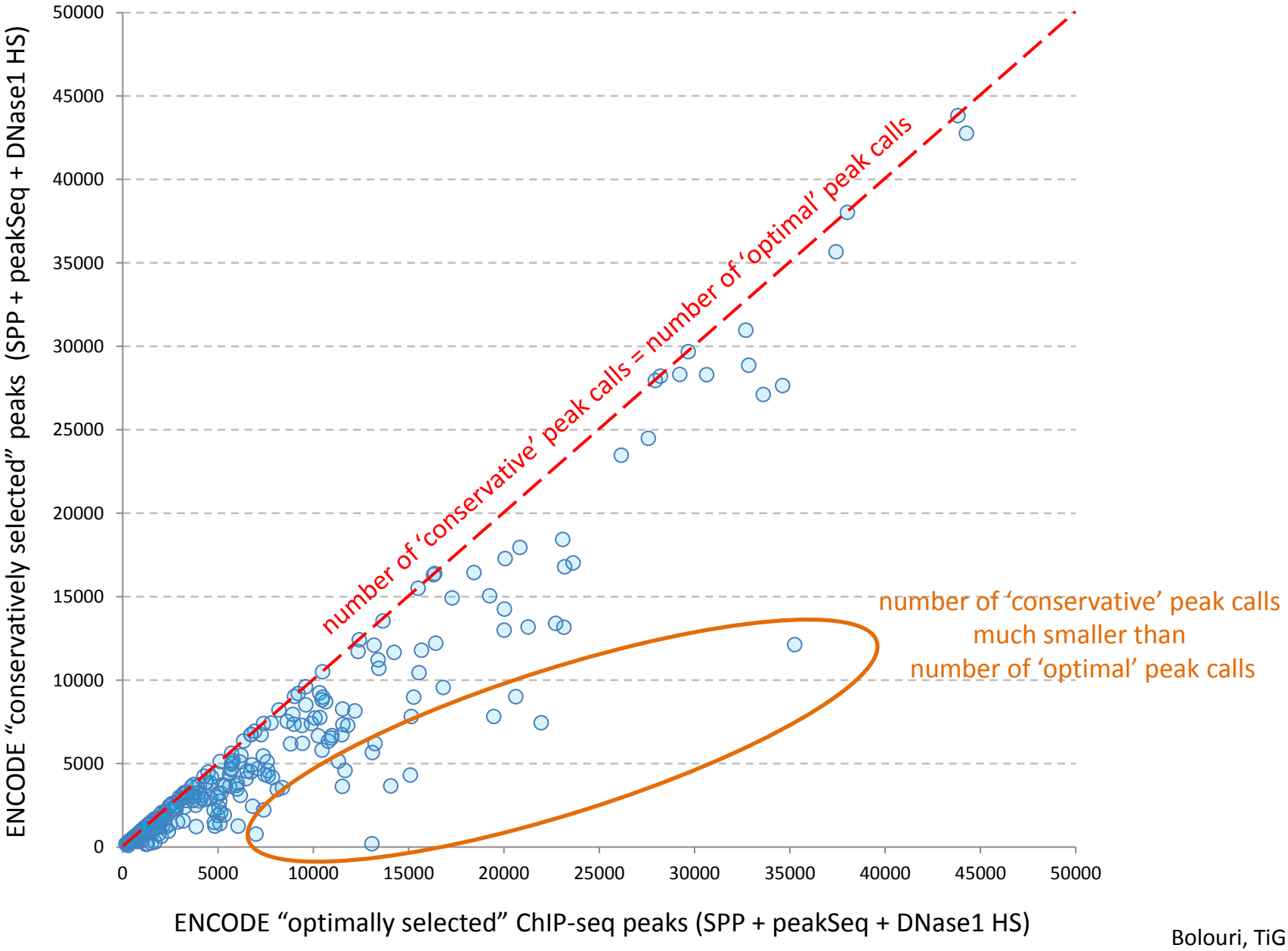
Always test the statistical significance of finding a pattern



Simpson's paradox

Results are often statistically reliable, but individually noisy

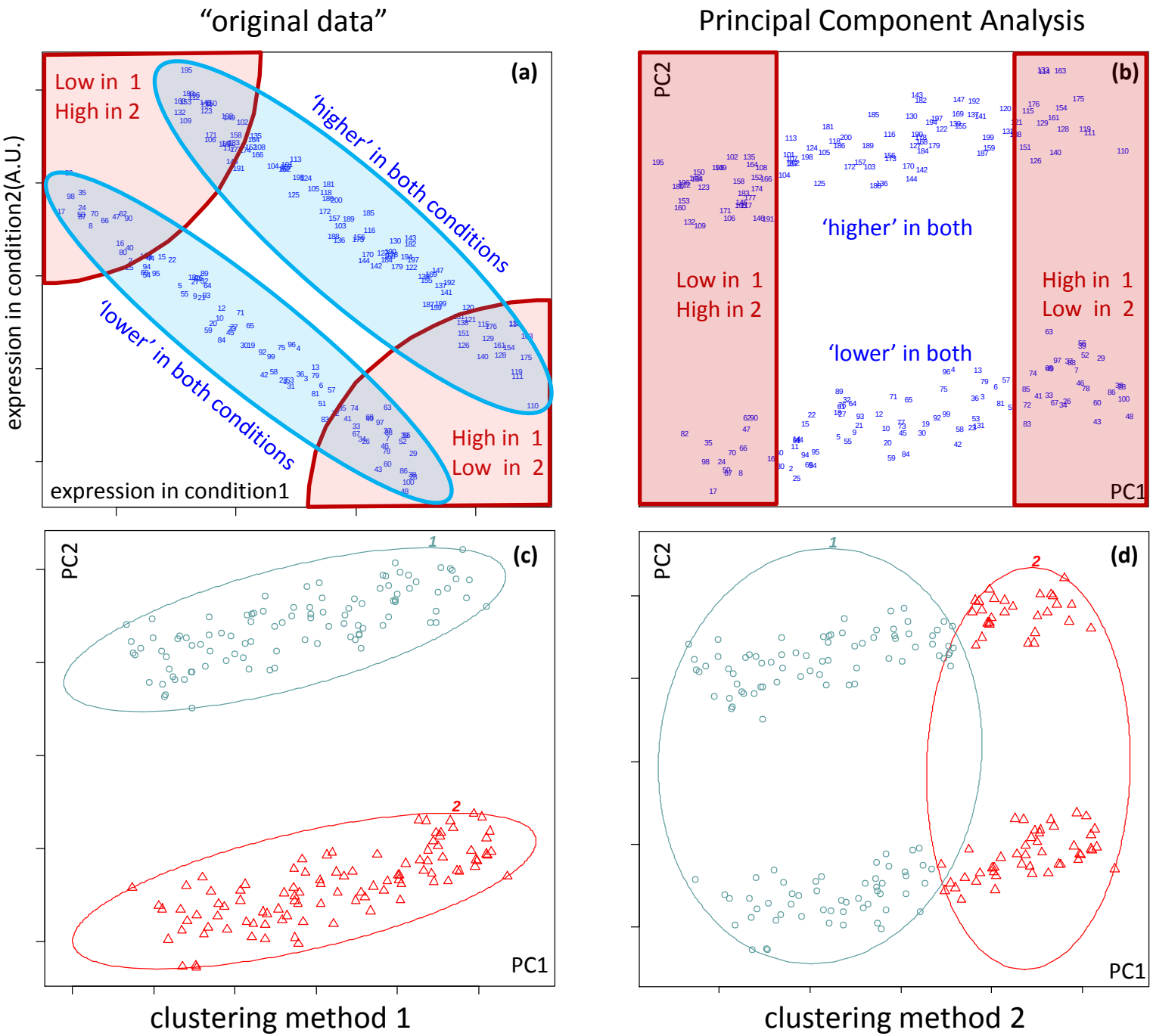
Per sample variability in ENCODE ChIP-seq peak calls (254 transcription factors)



Bolouri, TiG 2014

The importance of knowing your tools:

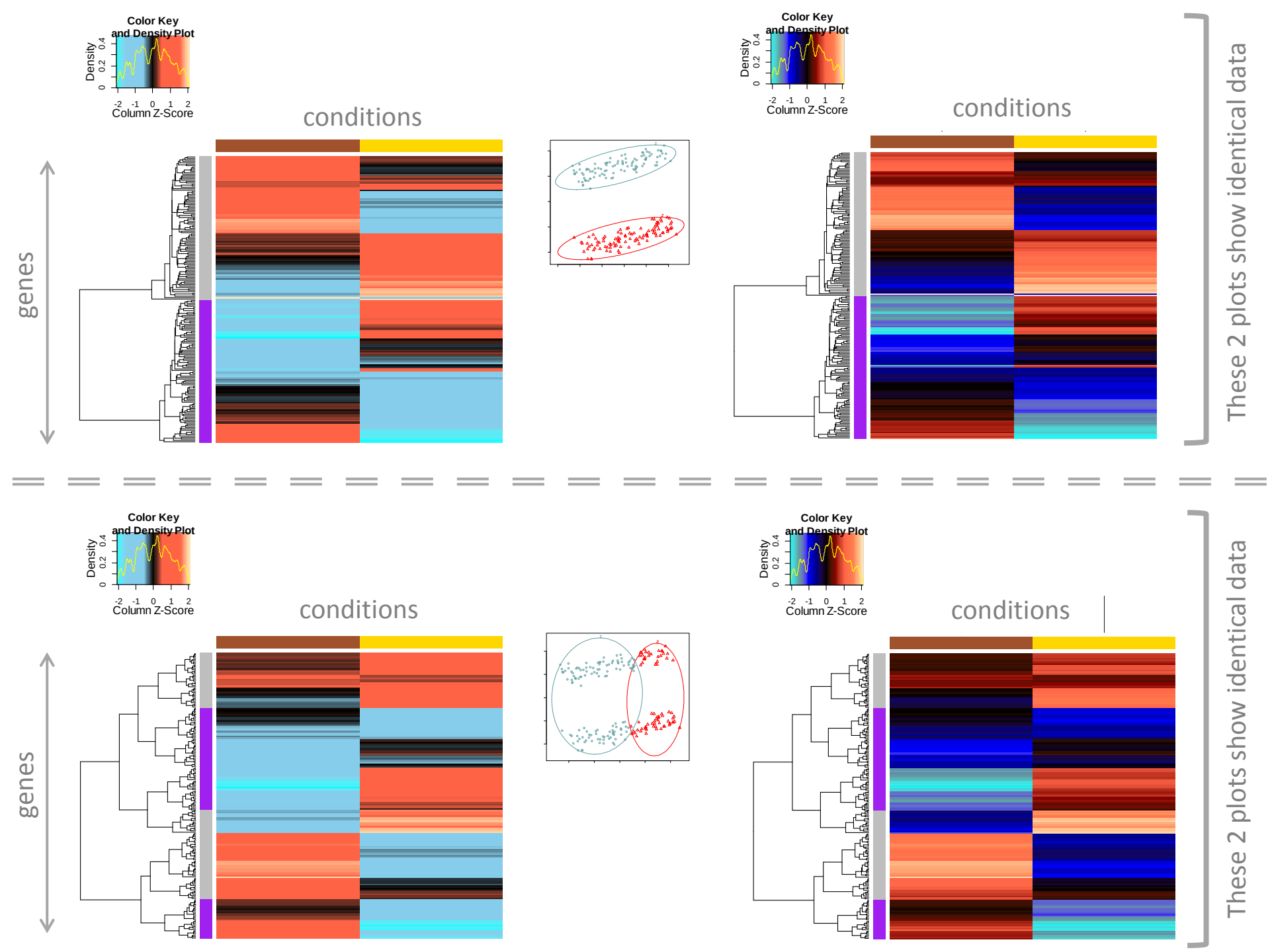
choices of statistical methods & parameters affect results



Bolouri, TiG 2014

The importance of full disclosure

Example: how a missing color-scale key can mislead

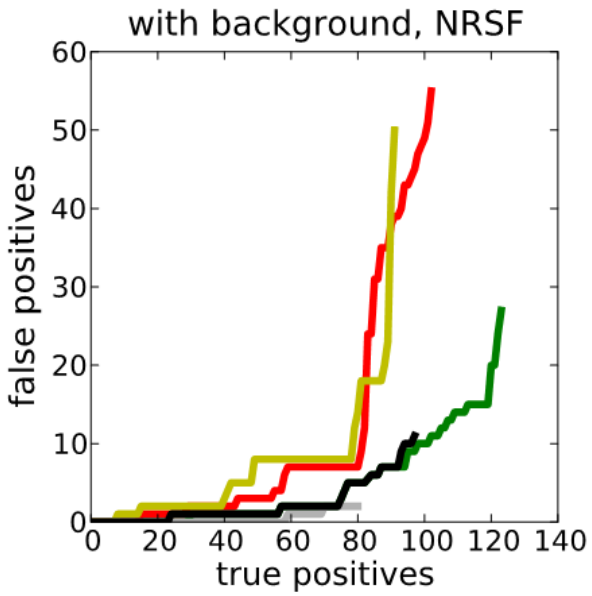

Color Key and Density Plot
Density
Column Z-Score
conditions
genes
These 2 plots show identical data
Color Key and Density Plot
Density
Column Z-Score
conditions
These 2 plots show identical data

- Process data with multiple algorithms and parameters
- When possible use matched controls

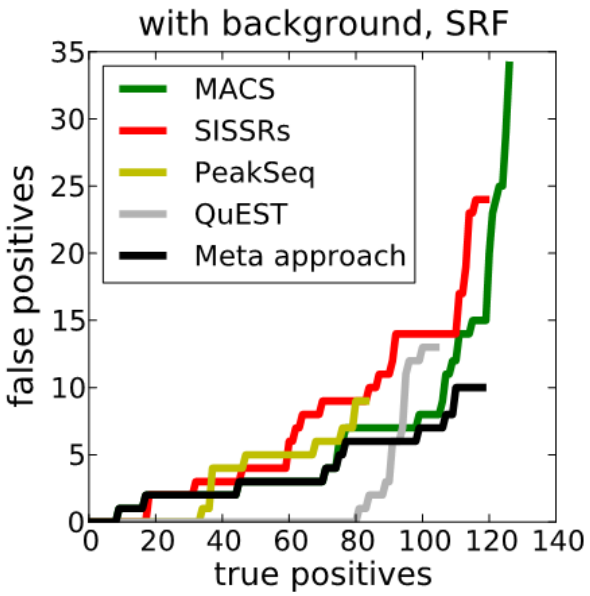
Published online 26 November 2010 *Nucleic Acids Research, 2011, Vol. 39, No. 4* ***e25***
doi:10.1093/nar/gkq1187

A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs

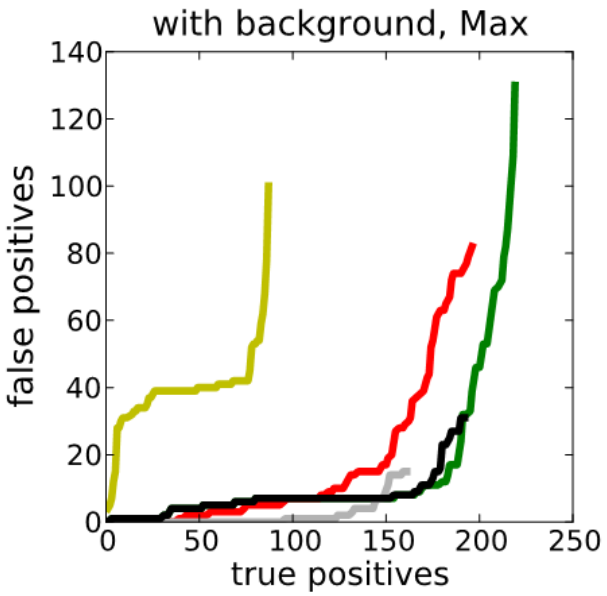
Morten Beck Rye[1,*], Pål Sætrom[1,2] and Finn Drabløs[1]



(out of 138 manually curated binding sites)



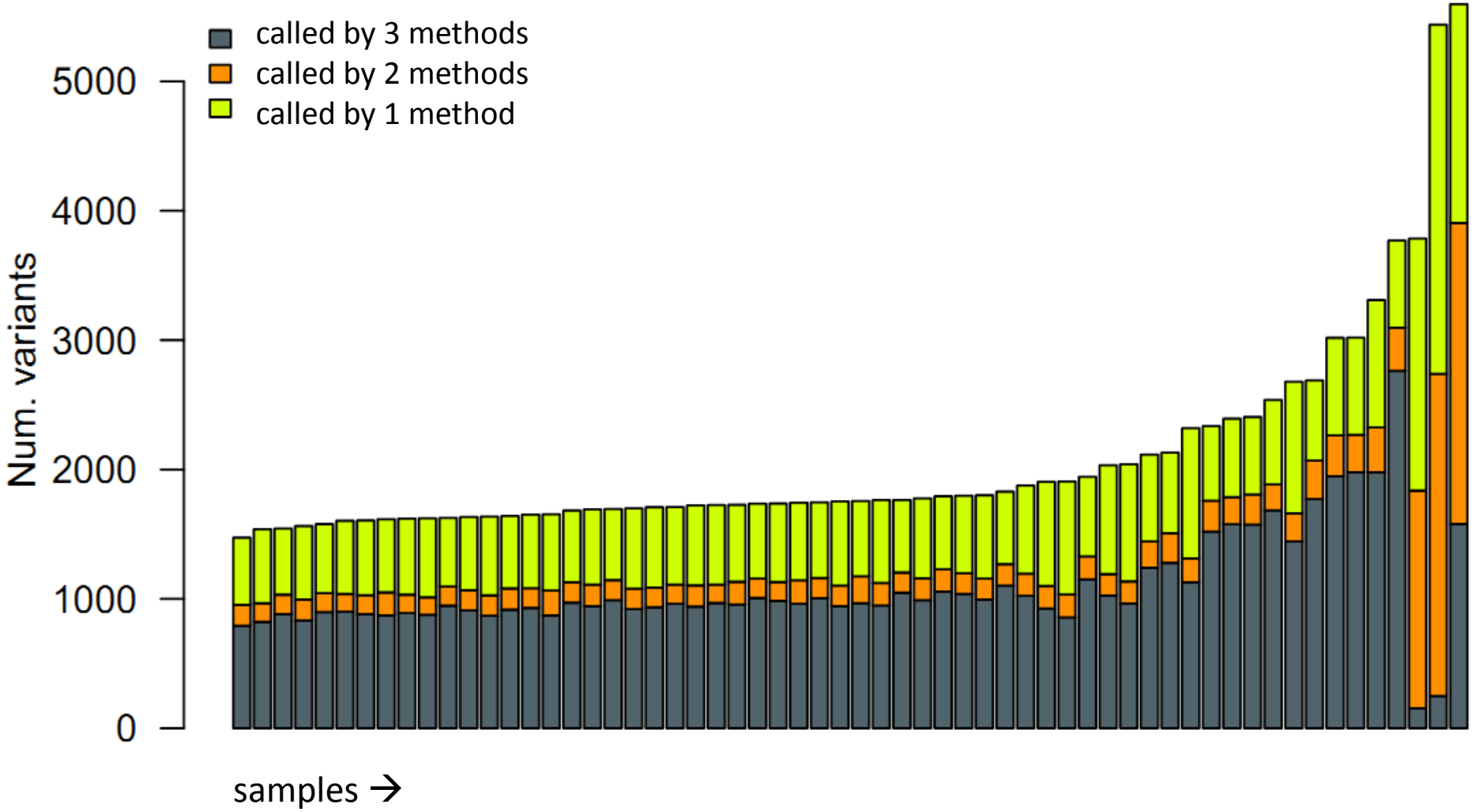
(out of 136 manually curated binding sites)



(out of 226manually curated binding sites)

SNAs detected by 3 methods applied to the same 80X tumor exomes

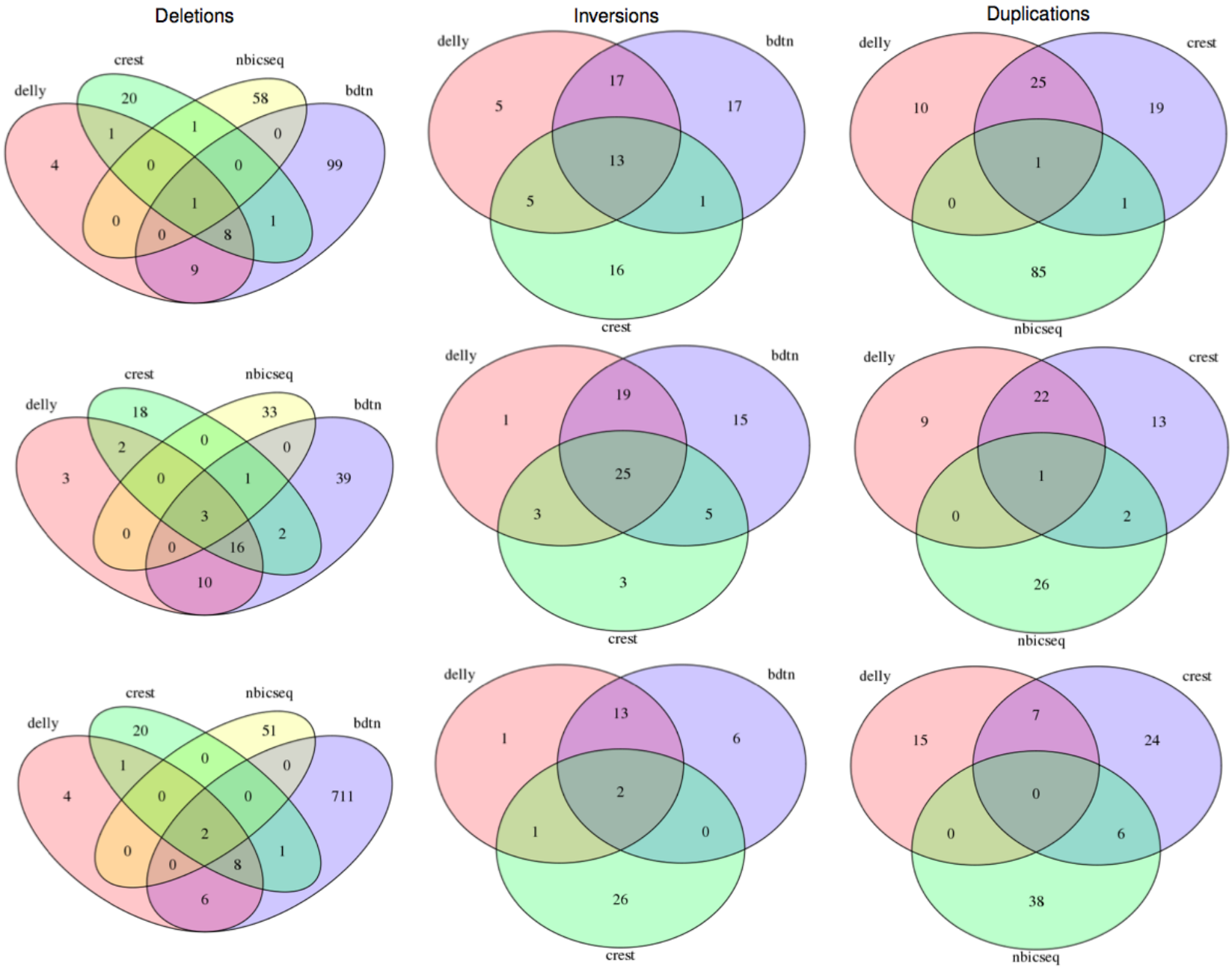
(methods: MuTect, Strelka, Virmid)



2x125bp HiSeq 2500 (high output mode),
Library Prep: Agilent SureSelectXT V4 71mb+UTR Exome

Overlap among four structural variation callers for three 30X WGS tumor samples

2x150bp HiSeq X, Library Prep: TruSeqDNA Nano 350bp



Overview:

(1) What do I mean by Genomic Regulatory Networks (GRNs)?

- Why go genome-wide?
- Why focus on regulatory interactions?

(2) What do I mean by big (molecular biology) data?

(3) Overview of integrative network discovery/modeling

(4) Examples of available bioinformatics resources

(5) Limitations and common pitfalls

(6) Conclusion: ***Proceed With Caution!***

Indels detected by 2 methods applied to the same 80X tumor exomes

(methods: SID, Strelka)