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.
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

- Chromosomal fusions, copy number changes, functional mutations in pediatric AML

- MiRs 9, 18a, 27a, 30c, 199a inactivated in AML

- MLL fusions

- MiRs 9, 18a, 27a, 30c, 199a

- PTPN11/SHP2

- HOXA 5, 7, 9, 10, MEIS1, LMO2, FLT3, NMYC

- ETV6

- MΦ

- B-cells

- E47
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 3\textsuperscript{rd} party data to discover GRNs

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

- Correlate mRNA data with 3\textsuperscript{rd} 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/
GRN discovery/modeling example approach (2):

GRN inference by integration of complementary ‘omics data sets

Example software:


sbenz.github.com/Paradigm
Dynamic regulatory network controlling T_H17 cell differentiation

Nir Yosef, Aviv Regev & colleagues

**Time course data**
- Expression clusters
- Timing of target gene induction
- Presence of regulators

**Published genomic profiles**
- Protein–DNA binding (298 TFs)
- DNA binding motifs (823 motifs)
- Expression profiles from 159 immune cells
- Knockout effects (11 TFs)

**Integrate**

**Network Model**

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

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

**Perturbation (siRNAs, nanowire delivery)**

**Validated model**
- 29 transcriptional regulators
- 10 receptors
- 22 T_H17-positive regulators antagonistically coupled to 5 T_H17-negative regulators
Examples of available bioinformatics resources

... for people who do not write Perl, Python, R (Bioconductor), etc.
https://usegalaxy.org
http://genomespace.org/
“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.
Results are often statistically reliable, but individually noisy
Per sample variability in ENCODE ChIP-seq peak calls (254 transcription factors)

Bolouri, TiG 2014

number of ‘conservative’ peak calls much smaller than number of ‘optimal’ peak calls
The importance of knowing your tools:

choices of statistical methods & parameters affect results
Gene expression in condition 2
Component 2

These two components explain 100% of the point variability.
The importance of full disclosure

Example: how a missing color-scale key can mislead
These 2 plots show identical data.
☐ Process data with multiple algorithms and parameters

☐ When possible use matched controls
A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs

Morten Beck Rye¹,* Pål Sætrom¹,² and Finn Drabløs¹

(out of 138 manually curated binding sites)

(out of 136 manually curated binding sites)

(out of 226 manually 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
Overview:

(1) What do I mean by Genomic Regulatory Networks (GRNs)?
   o Why go genome-wide?
   o 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)