Human Genome Annotation
(Large-scale Comparative & Functional Description in terms of Variable Block Deletions & Regulatory Networks)
Jim Gray’s 4th Paradigm

Science Paradigms

- Thousand years ago: science was empirical describing natural phenomena
- Last few hundred years: theoretical branch using models, generalizations
- Last few decades: a computational branch simulating complex phenomena
- Today: data exploration (eScience) unify theory, experiment, and simulation
  - Data captured by instruments
  - Or generated by simulator
  - Processed by software
  - Information/Knowledge stored in computer
  - Scientist analyzes database / files using data management and statistics

#3 - Simulation
Prediction based on physical principles (e.g., Exact Determination of Rocket Trajectory)
Emphasis on: Supercomputers

#4 - Data Mining
Classifying information & discovering unexpected relationships
Emphasis: networks, “federated” DBs

Gray died in ’07.
Book about his ideas came out in ‘09…..
Since ‘07, continued “steady” increase in computation; Even Larger Data Growth in Last Decade, Giving Rise to the Analysis of Many Big Data Sets

- Commercial World
  - Financial & Retail Data (Walmart)
  - Sensor Network
  - Cheap Imaging
- Social Media
  - Online behavior, life-logging
  - Mining documents/literature
- High-throughput scientific instruments
  - Sequencers
  - Telescopes
  - High Energy Accelerators
Sequencing Data Explosion in last 5 years

From ‘00 to ~’20, cost of DNA sequencing expt. shifts from the actual seq. to sample collection & analysis

[Sboner et al. (‘11) GenomeBiology]
Big Data: a current buzzword

Data Scientist: The Sexiest Job of the 21st Century
by Thomas H. Davenport and D.J. Patil

Artwork: Tamar Cohen, Andrew J Buboltz, 2011, silk screen on a page from a high

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business ne-
up. The company had just under 8 million accounts, and the number was growing qu-
friends and colleagues to join. But users weren’t seeking out connections with the pe-
rate executives had expected. Something was apparently missing in the social exp-

[Oct. ‘12 issue]
What do people do with Big Data?

• Fundamental goal is general understanding & answering specific Qs: modeling & making predictions

• Explicit Description of Data not Important -- Fast query, hiding underlying structure (e.g. Google Search)

• Explicit Description of Data Important – Organization highlighting underlying structure (e.g. Google Maps)

[ Nature 489: 208 ]

Higgs Boson: Searching Through Many Events for a Few Needles

"Golden" Events
One $H \rightarrow 4 \ell / \text{Billion}$
Making Intuitive Maps, Highlighting Large-scale Structure of Stars & the Earth

Revealing Statistics on Large-scale Human Activity

Large-scale twitter mining for “Mood”


[Golder & Macy, Science Vol. 333 no. 6051 pp. 1878]
Informative Models
Summarizing Activity Patterns

Avg. Article Accesses v. t

Multiplicative Model of Information Diffusion

2 Decay Regimes

PLoS ONE 6(5): e19917. doi:10.1371/journal.pone.0019917
Human Genome Annotation – a non-intuitive map

- Large-scale organization, providing an overview of the genome
- Integration of heterogeneous data
Hierarchical ENCODE Presentation

• Raw data (reads) at the bottom
• Progressive Processed Summaries
  - Signals (e.g. representing the degree to which DNA is bound by TFs)
  - Site locations
  - Regulatory networks, chromatin states & models
• At top are linked publications documenting everything, forming metadata

[ PLOS CB 4:e1000158; Nature.com/encode + Nature 489: 208 ]
How might we annotate a human text?

**The Semicolon Wars**

Brian Hayes

---

**Every programmer knows there is one true programming language. A new one every week**

---

If you want to be a thoroughgoing world traveler, you need to learn 6,912 ways to say “Where is the toilet, please?” That’s the number of languages known to be spoken by the peoples of planet Earth, according to Ethnologue.com.

If you want to be the complete polyglot, you also have quite a challenge ahead of you, learning all the ways to say:

```
printf("hello, world\n");
```

(This one is in C.) A catalog maintained by Bill Kinnersley of the University of Kansas lists about 2,500 programming languages. Another survey, compiled by Diamuid Piggott, puts the total even higher, at more than 8,500. And keep in mind that whereas human languages have had millennia to evolve and diversify, all the computer languages have sprung up in just 50 years. Even by the more-conservative standards of the Kinnersley count, that means we’ve been inventing one language a week, on average, ever since Fortran.

For ethnologists, linguistic diversity is a cultural resource to be nurtured and preserved, much like biodiversity. Or am I? A good-enough notation—for expressing an algorithm or defining a data structure.

There are programmers of my acquaintance who will dispute that last statement. I expect to hear from them. They will argue—zealously, ardent, vehement—that we have indeed found the right programming language, and for me to claim otherwise is willful ignorance. The one true language may not yet be perfect, they’ll concede, but it’s built on a sound foundation and solves the main problems, and now we should all work together to refine and improve it. The catch, of course, is that each of these friends will decide which end of a boiled egg to crack.

This famous tempest in an egg cup was replayed 250 years later by designers of computer hardware and communications protocols. When a block of data is stored or transmitted, either the least-significant bit or the most-significant bit can go first. Which way is better? It hardly matters, although life would be easier if everyone made the same choice. But that’s not what has happened, and so quite a lot of hardware and software is needed just to swap ends at boundaries between systems.

This modern echo of Swift’s Endian wars was first pointed out by Danny Cohen of the University of Southern California in a brilliant 1980 memo, “On holy wars and a plea for peace.” The memo, subsequently published in *Computer*, was widely read and admired; the plea for peace was ignored.

Another feud—largely forgotten, I think, but never settled by truce or treaty—focused on the semicolon. In Algol and Pascal, program statements have to be separated by semicolons. For example, `in x := 0; y := x+1; z := 2` the semicolons tell the compiler where one statement ends and the next begins. C
Concepts:

Raw Tracks to Networks & Relating Genomic Inputs to Outputs

- Inputs (mostly Chip-seq of TFs & Chromatin)
- Outputs (mostly RNA-seq)

1. Signal processing of raw experimental data:
   - Removing artefacts
   - Normalization
   - Window smoothing

2. Segmentation of processed data into active regions:
   - Binding sites
   - Transcriptionally active regions

3. Group active regions into larger annotation blocks

4. Further analysis: Building regulatory networks

References:

- *Science* 330:6012
Comparative Analysis, a second way to annotate non-coding elements

Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

- Annotation via Large-scale Identification of Variable **Blocks** in the Population
  - **Methods**
    - **Read-depth**: MSB+CNVnator
    - **Breakpoints** & Split Read: SRiC, AGE & BreakSeq
  - **Applications**: 1000G & Somatic Variation
- A **Networks** View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a **hierarchy** with information flow bottlenecks
  - Understanding the **impact of variation** and constraint on the network
    - Particularly with network **analogies**
Main Steps in Genome Resequencing

[Snyder et al. Genes & Dev. ('10)]

Step 0: Generate Reads

Step 1: Call SNPs
using uniquely and correctly mapped reads

Step 2: Find SVs
with aberrant paired-end reads, split-reads, read-depth analysis and CGH array data

Step 3: Assemble New Sequences
with split-, spanning- and misleading-reads

Step 4: Phasing
mostly with paired-end reads
1. Paired ends

Methods to Find SVs

2. Split read

3. Read depth (or aCGH)

4. Local Reassembly

[Snyder et al. Genes & Dev. ('10)]
Different Approaches Work Differently on Different Events

**Deletions**

- **Split-read analysis**
- **RP (fosmid)**
- **RP (454)**
- **RP (Solexa/SOLiD)**
- **hr-aCGH**
- **dbSNP**

<table>
<thead>
<tr>
<th>Method</th>
<th>Indel Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split-read</td>
<td>&gt; 1 bp</td>
</tr>
<tr>
<td>Fosmid</td>
<td>&gt; 8 kb</td>
</tr>
<tr>
<td>454</td>
<td>&gt; 3 kb</td>
</tr>
<tr>
<td>Solexa/SOLiD</td>
<td>&gt; 0.1 kb</td>
</tr>
<tr>
<td>hr-aCGH</td>
<td>&gt; 0.5 kb</td>
</tr>
<tr>
<td>dbSNP</td>
<td>1–28 bp</td>
</tr>
</tbody>
</table>

**Insertions**

- **dbSNP**
- **RP (Solexa/SOLiD)**
- **hr-aCGH**
- **RP (454)**
- **RP (fosmid)**
- **Split-read analysis**

<table>
<thead>
<tr>
<th>Method</th>
<th>Indel Size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbSNP</td>
<td>1–28 bp</td>
</tr>
<tr>
<td>Solexa/SOLiD</td>
<td>100–250 bp</td>
</tr>
<tr>
<td>hr-aCGH</td>
<td>&gt; 0.5 kb</td>
</tr>
<tr>
<td>454</td>
<td>2–3 kb</td>
</tr>
<tr>
<td>fosmid</td>
<td>8–40 kb</td>
</tr>
<tr>
<td>Split-read</td>
<td>1–250 bp</td>
</tr>
</tbody>
</table>

[Zhang et al. (‘11) *BMC Genomics*]
Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

• Annotation via Large-scale Identification of Variable **Blocks** in the Population
  • Methods
    • **Read-depth**: MSB+CNVnator
    • **Breakpoints** & Split Read: SRiC, AGE & BreakSeq
  • **Applications**: 1000G & Somatic Variation

• **Networks** View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a **hierarchy** with information flow bottlenecks
  - Understanding the **impact of variation** and constraint on the network
    • Particularly with network **analogies**
Array Signal

Read depth

Individual genome

Reads

Mapping

Reference genome

Counting mapped reads

Read depth signal

Zero level

Mean-shift-based (MSB) segmentation: no explicit model

• For each bin attraction (mean-shift) vector points in the direction of bins with most similar RD signal
• No prior assumptions about number, sizes, haplotype, frequency and density of CNV regions
• Not Model-based (e.g. like HMM) with global optimization, distr. assumption & parms. (e.g. num. of segments).
• Achieves discontinuity-preserving smoothing
• Derived from image-processing applications

[Abyzov et al. Gen. Res. ('11)]
Objective: Find the densest region
Distribution of identical billiard balls

Observed depth of coverage counts as samples from PDF

Kernel-based approach to estimate local gradient of PDF

Iteratively follow grad to determine local modes

Region of interest
Center of mass
Mean Shift vector
Example of Application of CNVnator to RD data

NA12878, Solexa 36 bp paired reads, ~30x coverage
Split-read Analysis

[ Zhang et al. BMC Genomics ('11) ]
SRiC: Split Read Pipeline

[Zhang et al. ('11) BMC Genomics]
Deletions are the Easiest to Identify

- **Simple SVs**
  - Deletion
  - Insertion, small
  - Insertion, large

- **Complex SVs**
  - Duplication
  - Translocation

[Zhang et al. ('11) BMC Genomics]
Using Simulation to Parameterize SRiC: Deletions Easier than Insertions

[Figure S3. (Zhang et al. ('11) BMC Genomics)]
Using Simulation to Parameterize SRiC: Coverage & Read Length

[Diagram A] Deletion: Percentage of called deletions vs. read length (bp) for different read lengths (50, 100, 200, 400, 800 bp) and coverage levels (2x, 5x, 10x, 15x, 20x).

[Diagram B] Insertion: Percentage of called insertions vs. read length (bp) for different read lengths (50, 100, 200, 400, 800 bp) and coverage levels (1x, 5x, 10x, 15x, 20x).

[Diagram C] Deletion: Percentage of called deletions vs. read length (bp) for different coverage levels (1x, 5x, 10x, 15x, 20x) and read lengths (50, 100, 200, 400, 800 bp).

[Diagram D] Insertion: Percentage of called insertions vs. read length (bp) for different coverage levels (1x, 5x, 10x, 15x, 20x) and read lengths (50, 100, 200, 400, 800 bp).

[Zhang et al. ('11) BMC Genomics]
Difficulties in Defining Exact Breakpoints

- **NW alignment**
  - Small gap penalty
  - Large gap penalty

- **SW alignment**
  - Small gap penalty
  - Large gap penalty

- **Optimal alignment**
  - A)
  - B)
  - C)
AGE
Alignment with
Gap Excision

Given scoring scheme (match, mismatch, gap open, gap extend) find an optimal alignment of two sequences (i.e. with highest score) where ONE gap is NOT penalized

[Abyzov & Gerstein ('11) Bioinfo.]
Using BreakSeq with a Breakpoint Library

Map reads onto

Library of SV breakpoint junctions

Junctions can be put on a chip

* Read overlaps <10 bp to one side of the breakpoint is discarded and read matches also to the reference genome is classified as non-unique match

[Lam et al., ('10) Nat. Biotech.]
SV Ancestral State Analysis

Inferring **Insertion** according to Ancestral State

Inferring **Deletion** according to Ancestral State

Region in Reference Genome inferring Deletion State

Region in Reference Genome inferring Insertion State

Junction A  Junction C  Junction B

1000 bp

Syntenic Primate Region inferring Insertion State

Syntenic Primate Region inferring Deletion State

[sv.jpeg]

[Lam et al., ('10) Nat. Biotech.]
SV Formation Mechanism

**NAHR**
(Non-allelic homologous recombination)

Flanking repeat
(e.g. Alu, LINE...)

**NHEJ**
(Non-homologous-end-joining)

No (flanking) repeats.
In some cases <4bp microhomologies

**TEI**
(Transposable element insertion)

L1, SVA, Alus

**VNTR**
(Variable Number Tandem Repeats)

Number of repeats varies between different people
SV Mechanism Classification

NAHR

Highly similar with minor offset

Single RETRO

Multiple RETRO

[Jacobs et al., ('10) Nat. Biotech.]
SV Mechanism Classification

1 kb ≤ SV ≤ 1 Mb

- Has flanking sequences
  - yes
  - Annotate SV and flanking regions by RepeatMasker
  - no
    - Unclassified
  - Has extensive coverage by VNTR regions
    - yes
      - VNTR
    - no
      - Extract a window at each breakpoint and align the two sequences
        - Two sequences share high similarity; Homologous regions have minor offsets, correct orientations and span the breakpoints
          - yes
            - NAHR
          - no
        - no

SV region covered by a single TE

- Yes
  - Potential processed pseudogene and other ambiguous cases
- No

SV region covered by multiple successive TEs

- Yes
  - Has a poly-A tail and TSD
    - yes
      - NHR
    - no
      - Annotated as fragments from a single TE
        - yes
          - MTEI
        - no
          - STEI
- No
Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

• Annotation via Large-scale Identification of Variable Blocks in the Population
  • Methods
    • Read-depth: MSB+CNVnator
    • Breakpoints & Split Read: SRiC, AGE & BreakSeq
  • Applications: 1000G & Somatic Variation

• A Networks View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  - Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
• **Many** different callers compared & used (including SRiC & CNVnator but also VariationHunter, Cortex, NovelSeq, PEMer, BreakDancer, Mosaik, Pindel, GenomeSTRiP, mrFast....)

• AGE (& Tigra_SV) for precise breakpoints on the resulting SVs

• BreakSeq Mechanism classification

• GenomeSTRiP genotyping

Deletion Size Distribution

NA12878, Solexa 36 bp paired reads, ~30x coverage

Uniformly Determined Deletion Size Distribution in a High Coverage Individual vs. Overall Distribution in Pilot 1 (~180 low cov. individuals)

[Abyzov et al. Gen. Res. ('11); Mills et al.]
Summary of Mechanism Classification of ~8900 Deletion Breakpoints in 1000G Phase I

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>&lt;500 bps</th>
<th>500-1000 bps</th>
<th>1-10 kbps</th>
<th>&gt;10 kbps</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAHR</td>
<td>9 (2.6%)</td>
<td>294 (23.3%)</td>
<td>1420 (22.6%)</td>
<td>255 (24.7%)</td>
</tr>
<tr>
<td>NHR</td>
<td>284 (82.8%)</td>
<td>889 (70.4%)</td>
<td>4642 (73.7%)</td>
<td>748 (72.4%)</td>
</tr>
<tr>
<td>MEI</td>
<td>47 (13.7%)</td>
<td>67 (5.3%)</td>
<td>124 (2.0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>VNTR</td>
<td>2 (0.6%)</td>
<td>7 (0.6%)</td>
<td>64 (1.0%)</td>
<td>23 (2.2%)</td>
</tr>
<tr>
<td>Undefined</td>
<td>1 (0.3%)</td>
<td>6 (0.5%)</td>
<td>45 (0.7%)</td>
<td>7 (0.7%)</td>
</tr>
<tr>
<td>Total</td>
<td>343 (100%)</td>
<td>1263 (100%)</td>
<td>6295 (100%)</td>
<td>1033 (100%)</td>
</tr>
</tbody>
</table>

[1000 Genomes Consortium, Nature (2012)]
Breakpoint Features Analysis

[Image of a graph showing SVs vs. Telomeres and Breakpoints]

[Lam et al., ('10) Nat. Biotech.]
Does reprogramming of human somatic cells into iPSC (hiPSC) cause de novo CNVs?

7 individuals, 21 hiPSC lines

Few previous studies reported de novo CNVs in hiPSC

Is the hiPSC genome stable during reprogramming?

What is the cause/origin of hiPSC line manifested CNVs (LM-CNVs)?

Data (one HiSeq lane per hiPSC and fibroblast):
- Poly-A RNAseq
- Whole genome DNAseq

Abyzov et al., Nature, Nov 18 (online)
Example of line manifested CNV (LM-CNV)

One lane of genomic DNA sequencing reads by Illumina HiSeq (2 x 100 nt) at ~10x coverage

Gene in duplication has increased expression
Hypothesis about somatic CNVs

Somatic CNVs present in few fibroblast cells (i.e., at low allele frequency) give rise to hiPSC, as those are derived from a single (or just a few) fibroblast cell.

Abyzov et al., Nature, Nov 18 (online)
Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

- Annotation via Large-scale Identification of Variable **Blocks** in the Population
  - Methods
    - **Read-depth**: MSB+CNVnator
    - **Breakpoints & Split Read**: SRiC, AGE & BreakSeq
  - **Applications**: 1000G & Somatic Variation
- A **Networks** View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a **hierarchy** with information flow bottlenecks
  - Understanding the **impact of variation** and constraint on the network
    - Particularly with network **analogies**
Current Annotation:
1D Browser Tracks

Will this scale to 1000+ tracks? What will next-gen annotation look like?...
Networks occupy a midway point in terms of level of understanding.

1D: Complete Partslist ("Elements" in genomic tracks)

~2D: Network Wiring Diagram of a Molecular System

3D & 4D: Detailed structural understanding of cellular machinery

[UCSC genome browser]
[Chiu et al. Trends in Cell Biol, 16:144]
Networks as a common language in many scientific contexts

- Internet
- Food Web
- Electronic Circuit
- Neural Network
- Disease Spread
- Protein Interactions
- Social Network
Exploiting Network Analogies to Gain Intuition

Guilt by association

Finding the causal regulator (the "Blame Game")

[NY Times, 2-Oct-05, 9-Dec-08]
Data Flow: Chip-seq expts. to co-associating peaks

119 TFs from 458 ChIP-Seq experiments (2 Tb tot.)

- Mostly in Tier 1 cell lines
  - K562, GM12878, H1h-ESC…
- Matching RNA-Seq data in all cell-lines
- SPP & PeakSeq
- thresholding w. IDR (replicas)

Signal Tracks

7M Peaks from Uniform Peak Calling

TF1

TF2

TF119

2785 GATA1 (focus-factor) peak locations

Data Flow: peaks to proximal & distal networks

Peak Calling

Assigning TF binding sites to targets

Filtering high confidence edges & distal regulation
Based on stat. model combining signal strength & location relative to typical binding

~500K Edges

~26K Edges

[ Cheng et al., Bioinfo. ('11); Gerstein et al. Nature (in press, '12); Yip et al., GenomeBiology (in press, '12)]
Identifying Potential Enhancer-like Elements from Discriminative Model & then Linking these to Targets (via cell-line correlations) to Create Distal Edges

[ Yip et al. Genome Biology (in press, '12) ]
Identifying potential enhancers

1. TF peaks from ChIP-seq
   - Human genome in 100bp bins
   - TF A
   - TF B
   - ...

2. Use peaks as examples to learn chromatin features of binding active regions
   - Chroatin features
   - Machine learning
   - Prediction
   - BAR scores
   - Thresholding
   - BARs
   - Filtering
   - DRMs

3. Filter close to genes to get enhancer list
   - Genencode genes
   - Predicted genes

4. Use peaks as examples to learn chromatin features of binding active regions
   - DNase I
   - FAIRE
   - H3K4me3
   - ...

5. Find TFs binding enhancer-like elt. in cell lines with strong HMs
   - Human genome in 100bp bins
   - Gene 1
   - Gene 2
   - Gene 3
   - Scale

6. Draw distal edges from TFs to targets
   - HM signals
   - Expression levels
   - Cell lines

~20K distal edges tot. from ~130K enhancer-like elements

(Related to but more “targeted” than enhancer “states” from unsupervised segmentation, M Hoffman et al. & J Ernst et al.)

Identifying Potential Enhancer-like Elements from Discriminative Model & then Linking these to Targets (via cell-line correlations) to Create Distal Edges

[ Yip et al. GenomeBiology (in press, '12) ]
Network Stats to Identify Bottlenecks & Hubs

[Yu et al., PLOS CB (2007)]
Hierarchy Height Statistic = (normalized TF Out deg. – In deg.)
Strongest Proximal Regulatory Edges Can be Arranged into a Hierarchy

Hierarchy height distribution approximated by 3 levels

Optimally arrange TFs into 3 levels by sim. annealing, maximizing downward-pointing edges

Global wiring pattern of TFs

Middle level has highest betweenness, creating info. flow bottlenecks
Comparing Proximal & Distal Networks

Integration of TF hierarchy with other ‘omic information: more influential & connected TFs on the top

Avg. correlation betw. binding signal of TF & gene expr. of its target

Integration of TF hierarchy with other ‘omic information: more influential & connected TFs on the top

Avg. # of PPI for each TF

Integration of TF hierarchy with other ‘omic information: more influential & connected TFs on the top

Sig. corr. w/ TF hubbiness (.24 & .62)

# regulating miRNAs & # regulated miRNAs

Avg. values

Cooperation of Mid-level Regulators

From looking at expr of shared vs unshared targets

Co-binding & co-regulation

Co-association  Cooperativity  Physical Interaction

Enrichment or Depletion (Observed/Expected)

T  M  B
Network Motif Analysis: Enrichment of FFLs

3-node motifs

| Freq. | 868 (0.84) | 490 (0.81) | 729 (0.72) | 26 (0.62) | 0 (0) | 0 (0) | 8 (0.70) | 64 (0.91) | 6 (1.8) | 6 (2.9) | 2 (5.6) | 16 (4.8) | 122 (1.4) |

Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

• Annotation via Large-scale Identification of Variable **Blocks** in the Population

  • Methods
    • **Read-depth**: MSB+CNVnator
    • **Breakpoints** & Split Read: SRiC, AGE & BreakSeq

  • **Applications**: 1000G & Somatic Variation

• A **Networks** View on Large-scale Organization of Genomic Elements
  - Understanding the human regulatory network as a **hierarchy** with information flow bottlenecks
  - Understanding the **impact of variation** and constraint on the network
    • Particularly with network **analogies**
Allele-Specific Behavior in the Regulatory Network

- In GM12878, determine ASB for ~50 TFs & ASE using RNA-Seq
  - AlleleSeq pipeline using personal diploid genome & annotation constructed for NA12878 (to prevent reference bias)
  - ~20% of expressed genes show ASE
  - ~10% of binding sites show ASB
- GM12878 Allele-Specific "Difference" Network
  - Just proximal edges with ASB
  - Just target nodes with ASE

RNA/ChIP-Seq Reads

[ Rozowsky et al. MSB ('11) ]
Many Technical Issues in Determining ASE/ASB:
Reference Bias
(naïve alignment against reference)

[Rozowsky et al., MSB (‘11)]
Construction of a Personal Diploid Genome & Transcriptome

Reference Genome

Paternal Haplotype

Personal Genome

Maternal Haplotype

vcf2diploid

Genotyping, Phasing, Filtering

Personal Variants

SVs  Indels  SNPs

Reference
Deletion
SNP
Insertion
Personal Haplotype

[Rozowsky et al., MSB (in press,'11)]
Maternal & Paternal Personal Regulatory Networks: combinatorial coordination of ASE & ASB

[ Rozowsky et al. MSB ('11) ]
More "allelic" components under weaker selection

(from 1000G pilot & phase I)
More generally, more connected components ("hubs") have less variation

TF target in-degree

Neg. corr. with
(SCC=-.2, P<0.5)

dN/dS
(from chimp alignments)

More "allelic" components under weaker selection

TF target in-degree
&
TF out-degree

Neg. corr. with

ns SNP density, pN/pS, avg. DAF

(from 1000G pilot & phase I)

TFs at the Top Under Stronger Negative Selection

SNP dens. $\times 10^{-3}$

1.0

3.1

3.8

Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

• Annotation via Large-scale Identification of Variable Blocks in the Population
  • Methods
    • Read-depth: MSB+CNVnator
    • Breakpoints & Split Read: SRiC, AGE & BreakSeq
  • Applications: 1000G & Somatic Variation
• A Networks View on Large-scale Organization of Genomic Elements
  – Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  – Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
More Connectivity, More Constraint: A theme borne out in many studies

Proteins that have a more central position evolve more slowly and are more likely to be essential. This phenomenon is observed in many organisms and different kind of networks: Fraser et al. (’02) Science, (’03) BMC Evo. Bio. [yeast PPI]; Butland et al. (’04) Nature [E coli. PPI]; Hahn et al. (’05), MBE [worm, fly PPI]; Cheng et al. (’09), BMC Genomics [miRNA nets]

- Sequence variation v. centrality
  - Nonsyn / synonymous SNPs v. deg. centrality
    \[ \rho = -.1, P < 4.0e-4 \]
    \[ \rho = -.3, P < 2.2e-16 \]
    (updated to 1000G phase I)

Proteins that have a more central position evolve more slowly and are more likely to be essential. This phenomenon is observed in many organisms and different kind of networks: Fraser et al. (’02) Science, (’03) BMC Evo. Bio. [yeast PPI]; Butland et al. (’04) Nature [E coli. PPI]; Hahn et al. (’05), MBE [worm, fly PPI]; Cheng et al. (’09), BMC Genomics [miRNA nets]
E. Coli Transcriptional regulatory network vs Linux call graph

<table>
<thead>
<tr>
<th>Basic properties of systems</th>
<th>E. coli transcriptional regulatory network</th>
<th>Linux call graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>Genes (TFs &amp; targets)</td>
<td>Functions (subroutines)</td>
</tr>
<tr>
<td>Edges</td>
<td>Transcriptional regulation</td>
<td>Function calls</td>
</tr>
<tr>
<td>External constraints</td>
<td>Natural environment</td>
<td>Hardware architecture, customer requirements</td>
</tr>
<tr>
<td>Origin of evolutionary changes</td>
<td>Random mutation &amp; natural selection</td>
<td>Designers’ fine tuning</td>
</tr>
</tbody>
</table>

A subnetwork in the Linux call graph from CodeViz

[Yan et al., PNAS ’10]
E. coli transcriptional regulatory network

Linux call graph

master regulator

middle manager

workhorse

[Yan et al., PNAS (2010), in press]
Comparison: hierarchical organization

<table>
<thead>
<tr>
<th></th>
<th>% in E. coli regulatory network</th>
<th>% in Linux call graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>master regulator</td>
<td>4.6</td>
<td>29.6</td>
</tr>
<tr>
<td>middle manager</td>
<td>5.1</td>
<td>58.2</td>
</tr>
<tr>
<td>workhorse</td>
<td>90.2</td>
<td>12.3</td>
</tr>
</tbody>
</table>

[Yan et al., PNAS ('10)]
We can track the evolution of a function as the “rate of evolution of a function” ~ the number of times it got revised.
Distribution of Evolutionary Rates of components in E. coli vs Linux

Persistent genes evolve slowly

Two classes of persistent functions

[Yan et al., PNAS '10]
Centrality is correlated with variation in Linux, opposite of the trend of constraint with centrality in biological networks

[Yan et al., *PNAS* '10]
Perspectives on Random Change v Intelligent Design

- Central points = hubs & bottlenecks
- If changes random, best not to put them in central pts.
- If changes made rationally, can put them into central pts.
  - Moreover, good to do this, as these more often used
  - i.e more efficient
  - Why there’s so much GWB construction
Trying to Annotate the Genome Comparatively & Functionally, in terms of Variable Blocks & Networks

• Annotation via Large-scale Identification of Variable Blocks in the Population
  • Methods
    • Read-depth: MSB+CNVnator
    • Breakpoints & Split Read: SRiC, AGE & BreakSeq
  • Applications: 1000G & Somatic Variation

• A Networks View on Large-scale Organization of Genomic Elements
  – Understanding the human regulatory network as a hierarchy with information flow bottlenecks
  – Understanding the impact of variation and constraint on the network
    • Particularly with network analogies
Summary

Comparative analysis
- Large-scale sequence similarity comparison
- Identify large blocks of repeated and deleted sequence:
  - Within the human reference genome
  - Within the human population
  - Between closely related mammalian genomes
- Identify smaller-scale repeated blocks using statistical models

Functional analysis
- Signal processing of raw experimental data:
  - Removing artefacts
  - Normalization
  - Window smoothing
- Segmentation of processed data into active regions:
  - Binding sites
  - Transcriptionally active regions
- Group active regions into larger annotation blocks
- Further analysis: Building regulatory networks

Integrate comparative and functional annotations

Comparative Approaches to Identifying Structural Variants in the Human Population

• **RD: MSB+CNVnator**
  - Mean-shift segmentation approach following grad. of PDF
  - Not model based so flexible:
    - Equally applied to aCGH & depth of coverage of various types of short reads

• **SR: AGE+BreakSeq**
  - Accurate alignments at breakpoints allowing for large jumping gap
  - Building a breakpoint library
  - Running against reads in newly seq. genome to genotype new SVs
  - Building a pipeline for characterizing breakpoints according to SV formation mechanisms

• **Applications**
  - 1000G: mostly NHR deletions, steep falloff in deletion freq. >1kb
  - Somatic copy number mosaicism in human skin revealed by IPS cells
Networks as Next-Gen Annotation: structure of the human regulatory network

- Reformulating peak lists into simplified network hierarchies, with “downward” proximal regulation
- Distal regulation has a very different connectivity pattern
- TFs on the top are more "influential" and well-connected (e.g. w/ miRNAs)
- Info.-flow bottlenecks in middle, ameliorated by co-operation of mid-level regulators
  - Cooperation evident in network motifs (e.g. prevalence of noise-buffering FFLs)
- Maternal & Paternal personal regulatory networks manifest combinatorial coordination
- More connected components are under greater constraint (understanding through analogies)
TopNet – an automated web tool

Similar tools include Cytoscape.org, Idekar, Sander et al.

Encodenets.gersteinlab.org
(tYNA: Normal website + Downloaded code (JAVA) + Web service (SOAP) with Cytoscape plugin)

[Yu et al., NAR (2004); Yip et al. Bioinfo. (2006);
Similar tools include Cytoscape.org, Idekar, Sander et al]
Acknowledgements – SVs

• H Lam, A Abyzov, Z Zhang, X Mu, J Korbel, L Wang

• 1000G genomes, particularly in the SV group

• Somatic collaboration: A Abyzov, J Mariani, D Palejev, Y Zhang, MS Haney, L Tomasini, AF Ferrandino, LA Rosenberg Belmaker, A Szekely, M Wilson, A Kocabas, NE Calixto, EL Grigorenko, A Huttner, K Chawarska, S Weissman, AE Urban, M Gerstein, FM Vaccarino
Acknowledgements

(11 Main Projects, ~50 labs, >700 substantial contributors + NHGRI)

Networks/Elements (~60 participants):


+ (for enhancers) E Birney, P Bickel, & B Brown