Functional Genomics Data Integration: Should we store & share high or low level data?

Mark Gerstein
Yale

Slides at Lectures.GersteinLab.org
(See Last Slide for References & More Info.)
Functional Genomics Data Integration: Should we store & share high or low level data?

• In relation to functional genomics, what is high & low level data? (ie signals vs reads)
• Some representative programs for manipulating high-level data: PeakSeq, RSeqTools & ACT
• Advantages of high-level data: Privacy, Compactness
• Illustration of where low-level data is necessary: Allelic effects
Low-Level Data for RNA-seq & Chip-seq

Reads (fasta)  
+ quality scores (fastq)  
+ mapping (BAM)  

Reads => Signal (Intermediate file)  

Accumulating @ >1 Pbp/yr (currently),  
~20% of tot. HiSeq output
Higher level Information from Chip-seq

TFs with Peaks
- K562 Pol2 Sig
- K562 Pol2 Pk
- K562 c-Myc Sig
- K562 c-Myc Pk
- K562 mlgG Sig
- K562 H3K4me3 S
- K562 H3K36me3 S

Control
- K562 Pol2 Sig
- K562 c-Myc Sig
- K562 H3K4me3 S
- K562 H3K36me3 S

Hist. Marks (broad)

Networks

Aggregations

[Science 330: 1775 + ENCODE Data Sources TFs & Control: Yale HMs: UW & Broad]
Higher level Information from RNA-seq:
Avg. signal at exons & TARs (RPKM)s

[PNAS 4:107: 5254 ; IJC 123:569]
Functional Genomics Data Integration: Should we store & share high or low level data?

- In relation to functional genomics, what is high & low level data? (ie signals v reads)
- Some representative programs for manipulating high-level data: PeakSeq, RSeqTools & ACT
- Advantages of high-level data: Privacy, Compactness
- Illustration of where low-level data is necessary: Allelic effects
PeakSeq: Scoring Relative to Controls

Filter for Potential Targets based on "Mappability" Simulation

Scale Input Relative to ChIP

Score Relative to Binomial Expectation

Enriched Sites

[Rozowsky et al. Nat. Biotech ('09)]
RSeqTools: An efficient RNA-seq analysis pipeline

Habegger et al., Bioinformatics ('11)
ACT: Aggregation and Correlation Toolbox for Analyses of Genomic Tracks

- Aggregation
- Correlation
- Saturation

Jee et al. Bioinformatics (‘11)
ACT: Aggregation Analyses

Saturation Plots for coding and non-coding Transcribed Regions in C. elegans

ChIP-Seq Signal Over TSSs

Aggregate of human and worm Pol II ChIP-Seq Signals
Functional Genomics Data Integration: Should we store & share high or low level data?

- In relation to functional genomics, what is high & low level data? (ie signals v reads)
- Some representative programs for manipulating high-level data: PeakSeq, RSeqTools & ACT
- Advantages of high-level data: Privacy, Compactness
- Illustration of where low-level data is necessary: Allelic effects
Privacy Considerations in Personal Genomics

• Personal genomic sequences are fundamentally private data
• Not clear whether they can be treated with “open data” ethos of circa 2000
• Great potential for misuse
  – Insurance
  – Discrimination

• Personal Genomic info.
  currently essentially meaningless but will it be in 20 yrs? 50 yrs
  – Genomic sequence very revealing about one’s children
  – Once put on the web it can’t be taken back

• Yet large-scale mining of this information is important for medical research

[Greenbaum et al. PLOS CB (‘11)]
Privacy Considerations: Different for Functional Genomics

- Human genome reseq. all about variants vs. reference
- Situation diff. for func. genomics
  - Often variant info. is determined incidentally
- On one hand: **Reads have variant information** in most functional regions (deep RNA-seq expt. essentially exome seq.)
- On other hand: **high-level summaries and signal tracks** mostly what is used (80%) and do not involve variant info. Helpful to make this freely available and easy to use

[Greenbaum et al. *PLOS CB* ('11)]
### Light-weight formats

- Some light-weight format clearly separate public & private info., aiding exchange
- Files become much smaller
- Distinction between formats to compute on and those to archive with – become sharper with big data

---

**Anonymization (Optional)**

<table>
<thead>
<tr>
<th>Public</th>
<th>Private</th>
</tr>
</thead>
<tbody>
<tr>
<td>AlignmentBlocks</td>
<td>ID</td>
</tr>
<tr>
<td>chr1:+:201:250:1:50</td>
<td>1</td>
</tr>
<tr>
<td>chr5:--:561:510:1:50</td>
<td>2</td>
</tr>
</tbody>
</table>

- Reads (linked via ID, 10X larger than mapping coord.)

---

[Bioinformatics 27: 281]
MRF Examples

10X Compression Ex.

Raw ELAND export file has uncompressed file size: ~4 GB; total number of reads: ~20 million; number of mapped reads: ~12 million.

MRF file is significantly smaller (~400 MB uncompressed, ~130 MB compressed with gzip).

BAM file has a size of ~1.2 GB.

Reference based compression (ie CRAM) is similar but it stores actual variant beyond just position of alignment block.

[Habegger et al., Bioinformatics (‘11)]
Need Reads & Quality Scores for Archival Storage

80% of time lightweight summary (mapping w/o reads) is sufficient allowing calculation of high-level info. (e.g. RPKMs, splicing, peaks…)

20% of time absolutely need reads b/c
1. Mapping & scoring depends on many parameters, with no current consensus.
2. Some calculation intrinsically depend on reads (e.g. ASB)

Manageable, given scale (~20%) & clever compression (e.g. reference-based)
Plummeting Cost of Sequencing & Exponential Amount of Data Being Stored

[Greenbaum et al., Am. J. Bioethics (08)]
AlleleSeq

Allele-Specific Binding & Expression
Inferring Allele Specific Binding/Expression using Actual Sequence Reads

**RNA/ChIP-Seq Reads**
- ACTTTGATAGC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TCAATC
- CTTTGATAGC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAATC
- CTTTGATAGC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- TTGACAGC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- TGATAGC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- ATAGCGT	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- TAGCGT	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- CGTCAAC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- GTCAATG	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- CAATG	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- AATGC	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC
- TGCACGG	\text{T\text{C{\hphantom{C}}}GA{\hphantom{T}}}TGAAC

10 x T
2 x C

Interplay of the annotation and individual sequence variants
Many Technical Issues in Determining ASE/ASB:
Reference Bias

(naïve alignment against reference)

ASE/ASB Example:
…GTCAATGCAC
…GTCAATGCACG
…GTCAATGCACGT
…GTCAATGCACGTC
…GTCAACGCAGTGC
…GTCAACGCAGGAG
CAATGCAGTCGGGA
AAA

Null Example:
ACTTTGATAGCGTCAA
CTTTGATAGGTCAAACGC
TTGACAGGTCAATGAC
ATAGCGTCAATGACGT...
TAGGTCAACGCACGT...
CGTCAACGCACGT...
CAATGCACGT...
AAATGCACGT...

[Binomial Null Distribution
(no allele-specific behavior)]

[Allele-Specific SNPs]

Reference Allele 0.0 0.2 0.4 0.6 0.8 1.0
Fraction of Reads Mapping to Alternative Allele

Frequency 0 500 1000 2000

[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: TGGAAAGAAACCGTTT...
Deletion: TGGXXAAGACCGTTT...
SNP: TGGXXAAGACCGTTT...
Insertion: TGGXXAAGACCGTTT...
Personal Haplotype: TGGAAAGGAGGTTT...

[Rozowsky et al., MSB (in press, ‘11)]
Align reads to paternal haplotype
Align reads to maternal haplotype

Compare to find best alignment

Counts over het SNPs to determine allele specificity

Filter SNPs in CNVs using read-depth

Overlap ASB SNPs with TF binding sites
Overlap ASE SNPs with gene annotation

Report ASB and ASE SNPs with significance in VCF format

[Align reads to paternal splice-junction library]

[Align reads to maternal splice-junction library]

Compare to find best alignment

Counts over het SNPs to determine allele specificity

[Overlapping ASB SNPs with TF binding sites]

[Overlapping ASE SNPs with gene annotation]

[Reads]

[Genes]

[Binding Sites]

[Filter SNPs in CNVs using read-depth]

[Counts over het SNPs to determine allele specificity]

[Report ASB and ASE SNPs with significance in VCF format]

[Rozowsky et al., MSB (in press, '11)]
Specific Data Sets Used

<table>
<thead>
<tr>
<th>Data</th>
<th>Number of reads (millions)</th>
<th>Number of mapped reads (millions)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-Seq</td>
<td>393.9</td>
<td>164.7</td>
</tr>
<tr>
<td>Pol II ChIP-Seq</td>
<td>128 (33)</td>
<td>69.5 (13.2)</td>
</tr>
<tr>
<td>Pol III ChIP-Seq</td>
<td>12</td>
<td>7.5</td>
</tr>
<tr>
<td>cMyc ChIP-Seq</td>
<td>125</td>
<td>65.5</td>
</tr>
<tr>
<td>Max ChIP-Seq</td>
<td>79</td>
<td>46.1</td>
</tr>
<tr>
<td>JunD ChIP-Seq</td>
<td>133</td>
<td>72.5</td>
</tr>
<tr>
<td>cFos ChIP-Seq</td>
<td>84</td>
<td>30.4</td>
</tr>
<tr>
<td>NFκB ChIP-Seq</td>
<td>62</td>
<td>35.5</td>
</tr>
<tr>
<td>CTCF ChIP-Seq</td>
<td>46</td>
<td>26.4</td>
</tr>
</tbody>
</table>

- GM12878 is the immortalized lymphoblastoid cell-line from NA12878, the daughter in one of the deeply sequenced 1000G trios

[Rozowsky et al., MSB (in press, '11)]
Reference Bias Revisited

Assessing Reference Bias for GM12878 RNA-Seq data using Naïve reference mapping vs NA12878 mapping

[Rozowsky et al., MSB (in press, ’11)]
Specific & General Results

~20% sites show ASE, ~10% show ASB; equal betw. Mat & Pat (except on chr X)

[Note: The text includes a table with data on Genomic elements, Number of Elements Accessible for Allele-Behavior, Number with ASE or ASB, and Fraction with Allele-Specific Behavior.]

[Rozowsky et al., MSB (in press, ‘11)]
Allele-Specific Regulatory Network: coordination of ASE & ASB

Network Motifs

Expression

Gene
Novel TAR

TF

Maternal
Paternal

Binding

Single TF

<table>
<thead>
<tr>
<th>Maternal TF</th>
<th>Maternal Expression</th>
<th>Paternal Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maternal Regulation</td>
<td>81</td>
<td>22</td>
</tr>
<tr>
<td>Paternal Regulation</td>
<td>31</td>
<td>64</td>
</tr>
</tbody>
</table>

Multiple TFs (MIM)

<table>
<thead>
<tr>
<th>Maternal Expression</th>
<th>Paternal Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Maternal Regulation</td>
<td>40</td>
</tr>
<tr>
<td>Both Paternal Regulation</td>
<td>4</td>
</tr>
<tr>
<td>Mixed Regulation</td>
<td>3</td>
</tr>
</tbody>
</table>

Single TF (SIM)

<table>
<thead>
<tr>
<th>Both Maternal Expression</th>
<th>Both Paternal Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both Maternal Regulation</td>
<td>2,840</td>
</tr>
<tr>
<td>Both Paternal Regulation</td>
<td>254</td>
</tr>
</tbody>
</table>

[Rozowsky et al., MSB (in press, '11)]
High or Low Level:
What Functional Genomics Data Should We Store & Share for Integrative Analyses?

• In relation to functional genomics, what is high & low level data? (ie signals v reads)
• Some practical programs for manipulating high-level data: PeakSeq, RSeqTools & ACT
• Advantages of high-level data:
  Privacy, Comptactness
• Illustration of where low-level data necessary:
  Allelic effects
Acknowledgements

- **Act.gersteinlab.org**:  
  J Jee, J Rozowsky, KY Yip,  
  L Lochovsky, R Bjornson, G Zhong, Z Zhang,  
  Y Fu, J Wang, Z Weng

- **Peakseq**:  
  J Rozowsky,  
  G Euskirchen, RK Auerbach, ZD Zhang,  
  T Gibson, R Bjornson, N Carriero, M Snyder

- **RSeqTools**:  
  L Habegger, A Sboner,  
  TA Gianoulis, J Rozowsky, A Agarwal, M Snyder

- **Alleleseq.gersteinlab.org**:  
  J Rozowsky, A Abyzov,  
  J Wang, P Alves, D Raha, A Harmanci, J Leng,  
  R Bjornson, Y Kong, N Kitabayashi, N Bhardwaj,  
  M Rubin, M Snyder

- **Genomic Privacy**:  
  D Greenbaum,  
  A Sboner, XJ Mu
Default Theme

- Default Outline Level 1
  - Level 2
More Information on this Talk

SUBJECT: Networks

DESCRIPTION:

NOTES:
This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2010. Please read permissions statement at http://www.gersteinlab.org/misc/permissions.html. Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see http://streams.gerstein.info. In particular, many of the images have particular EXIF tags, such as kwpotppt, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwpotppt.