Next-Generation Sequence Data for Functional Genomics (RNA-seq & ChIP-seq)

Mark Gerstein
Yale
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
Control
Hist. Marks (broad)

Networks

H3K4me2

Aggregations

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

Connectivity Information

- Paired ends & split reads
- Assembling transcript structures & determining abundances of alt. transcripts
- Coming for Chip-seq (HiC & ChiA-Pet)

[PNAS 4:107: 5254 ; Science 330: 1775]
Privacy Considerations

• Human genome reseq. all about variants vs. reference
• Situation diff. for func. genomics
• 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
Light-weight formats

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

Anonymization (Optional)

Public

<table>
<thead>
<tr>
<th>AlignmentBlocks</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<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>

Private

<table>
<thead>
<tr>
<th>ID</th>
<th>Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GTGTGTGTGTATCCA...</td>
</tr>
<tr>
<td>2</td>
<td>ATGGCTCGTGGGATT...</td>
</tr>
<tr>
<td>3</td>
<td>CTCTGGTCTGTTACC...</td>
</tr>
</tbody>
</table>

Reads

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

Mapping coordinates without variants (MRF)
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)
• ENCODE Chip-seq peak-caller comparison
  – Many differences
• RGASP + ENCODE RNA analysis
  – Many differences betw. programs determining transcript abundances
  – Significant differences betw. RNA mapping approaches
Inferring Allele Specific Binding/Expression using Actual Sequence Reads

ChIP-Seq Reads
ACTTTGATAGCGTCAATG
CTTTGATAGCGTCAATGC
CTTTGATAGCGTCAACGC
TTGACAGCGTCAATGCAC
TGATAGCGTCAATGCACG
ATAGCGTCAATGCACGTCA
TAGCGTCAATGCACGTCA
CGTCAACGCACGTCGGGA
GTCAATGCACGTCGAGAG
CAATGCACGTCGGAGTTT
AATGCACGTCGGAGTTTG
TGCACGTGAGGAGTTGGC

Need to map separately to Maternal and Paternal personal genomes.
~11% of “accessible” Pol2 sites show ASB, which involves shifting >100K reads from equiv. positions in Mat. & Pat. chromosomes

Haplotypes with a Heterozygous Polymorphism

[Rozowsky et al., Mol Sys Biol (in press)]
[encodewiki.ucsc.edu]