Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

Mark Gerstein, Yale

Slides freely downloadable from Lectures.GersteinLab.org
& “tweetable” (via @markgerstein). See last slide for more info.
Large-scale RNA

• Recent advent of many consortia & group producing large scale RNA-seq following on DNA sequencing
• Often this is of human subjects (eg TCGA, PCAWG, GTEx) and needs to be protected
• Useful to build tools & approaches that interact with these data
ERCC Organization

Data Management & Resource Repository (DMRR)

- Curated Gene Context, Network Modules, Pathways

- Data Coordination Center (DCC)
  - Aleks Milosavljevic, PI

- Scientific Outreach Component (SOC)
  - David Galas, PI

- Data Integration & Analysis Component (DIAC)
  - Mark Gerstein, PI

- Develop Metadata Stds
- Create & Host exRNA Atlas
- Data Analysis & Visualization

RFA RM-12-011
Reference Profiles

RFA RM-12-012
Biogenesis, Biodistr, Uptake, and Effector Function

RFA RM-12-013
Clinical Utility for Biomarkers

RFA RM-12-014
Clinical Utility for Therapy Development

- Develop Analysis Pipelines/Tools
- Integrative Analysis
- Profile/Gene/Network Modules
Unlocking the Mysteries of Extracellular RNA Communication

Once thought to exist only inside cells, RNA is known to travel outside of cells and play a role in newly discovered mechanisms of cell-to-cell communication.

www.exrna.org
Regional exRNA Mapping Centers and Data Coordination Center

Pacific Northwest Res. Inst. (Seattle)

U. Mass Med School (Worcester)

Beth Israel (Boston)

NCBI Washington, D.C.

Baylor College of Medicine (Houston): Data Coordination Center

UCSF

UC San Diego

UC Michigan
RNA-Seq Profiling of Human exRNAs – Multiple Biological Fluids

- Short & long non-coding, non-coding, circular coding
- Plasma
- Serum
- Urine
- Saliva
- Cerebrospinal fluid
- Cord blood
- Seminal fluid
- Bronchoaveolar fluid
- Placenta
- Endogenous vs exogenous (environment/diet)
### ExRNA Atlas

**Analysis: EXR-DGPLAS00-AN**

<table>
<thead>
<tr>
<th>Status:</th>
<th>Protect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis Type:</td>
<td>Reference Alignment</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sample Code</th>
<th>Total Mapped Reads</th>
<th>Other Genomic Loci</th>
<th>tRNAs</th>
<th>miRNAs</th>
<th>tRNAs</th>
<th>piRNAs</th>
<th>snoRNAs</th>
<th>Rfam RNAs</th>
<th>Plant and Virus miRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXR-DGPLAS01-BS</td>
<td>6822465</td>
<td>46.232%</td>
<td>51.062%</td>
<td>2.296%</td>
<td>0.216%</td>
<td>0.011%</td>
<td>0.002%</td>
<td>0.000%</td>
<td>0.181%</td>
</tr>
<tr>
<td>SM11_norm1, Plasma, Scientific Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS02-BS</td>
<td>6318178</td>
<td>47.003%</td>
<td>51.495%</td>
<td>1.150%</td>
<td>0.186%</td>
<td>0.010%</td>
<td>0.002%</td>
<td>0.000%</td>
<td>0.155%</td>
</tr>
<tr>
<td>SM12_norm2, Plasma, Scientific Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS03-BS</td>
<td>5943384</td>
<td>48.815%</td>
<td>49.283%</td>
<td>1.542%</td>
<td>0.210%</td>
<td>0.012%</td>
<td>0.002%</td>
<td>0.000%</td>
<td>0.135%</td>
</tr>
<tr>
<td>SM13_norm3, Plasma, Scientific Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS04-BS</td>
<td>1490041</td>
<td>51.510%</td>
<td>47.307%</td>
<td>0.924%</td>
<td>0.119%</td>
<td>0.016%</td>
<td>0.002%</td>
<td>0.000%</td>
<td>0.122%</td>
</tr>
<tr>
<td>SM1_crc1, Plasma, Colorectal Carcinoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS05-BS</td>
<td>2872815</td>
<td>45.518%</td>
<td>52.513%</td>
<td>1.643%</td>
<td>0.191%</td>
<td>0.018%</td>
<td>0.003%</td>
<td>0.000%</td>
<td>0.114%</td>
</tr>
<tr>
<td>SM2_crc2, Plasma, Colorectal Carcinoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS06-BS</td>
<td>3690661</td>
<td>46.638%</td>
<td>51.518%</td>
<td>1.498%</td>
<td>0.215%</td>
<td>0.018%</td>
<td>0.002%</td>
<td>0.000%</td>
<td>0.110%</td>
</tr>
<tr>
<td>SM3_crc3, Plasma, Colorectal Carcinoma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS07-BS</td>
<td>3304333</td>
<td>47.004%</td>
<td>51.877%</td>
<td>0.661%</td>
<td>0.345%</td>
<td>0.020%</td>
<td>0.006%</td>
<td>0.000%</td>
<td>0.088%</td>
</tr>
<tr>
<td>SM6_uc1, Plasma, Ulcerative Colitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS08-BS</td>
<td>3613655</td>
<td>45.448%</td>
<td>53.622%</td>
<td>0.590%</td>
<td>0.196%</td>
<td>0.018%</td>
<td>0.002%</td>
<td>0.000%</td>
<td>0.125%</td>
</tr>
<tr>
<td>SM7_uc2, Plasma, Ulcerative Colitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXR-DGPLAS09-BS</td>
<td>4719012</td>
<td>44.496%</td>
<td>51.860%</td>
<td>3.297%</td>
<td>0.218%</td>
<td>0.015%</td>
<td>0.003%</td>
<td>0.000%</td>
<td>0.111%</td>
</tr>
<tr>
<td>SM8_uc3, Plasma, Ulcerative Colitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals
Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org

- Long-RNA pipeline
  - RSeqTools & anonymized MRF format

- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels

- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous

- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals
**Light-weight formats**

- Some lightweight 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><strong>Alignment Blocks</strong></td>
<td><strong>ID</strong></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>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**ID** | **Sequences**
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GTCGTGTCTGTATCCA...</td>
</tr>
<tr>
<td>2</td>
<td>ATGGCTCGTTGGATT...</td>
</tr>
<tr>
<td>3</td>
<td>CCTCTGGTCTGTTACC...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Reads**

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

**Mapping coordinates without variants (MRF)**

[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)]
Various Data Types (tracks, files, ROIs, etc)

Specific information on files/samples selected in the “Data Selector”

Tells the tool to use this input data/file

Tells the tool where to deposit results
Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals
eQTL Mapping Using RNA-Seq Data

- eQTLs are genomic loci that contribute to variation in mRNA expression levels.
- eQTLs provide insights on transcription regulation, and the molecular basis of phenotypic outcomes.
- eQTL mapping can be done with RNA-Seq data.
Information Content and Predictability

\[ ICI(\text{individual has variant genotypes } g_1, g_2, \ldots, g_n \text{ for variants } V_1, V_2, \ldots, V_n) = \log\left(\frac{1}{\text{Frequency of } V_1 \text{ genotype } g_1 = 2}\right) + \log\left(\frac{1}{\text{Frequency of } V_2 \text{ genotype } g_2 = 1}\right) + \ldots + \log\left(\frac{1}{\text{Frequency of } V_n \text{ genotype } g_n = 2}\right) \]

\[ e^{\rho} \quad \text{(Expression of gene } k) \quad \text{(Genotype of variant associated with expression of gene } k) \]

\[ \pi(V|E = e) = e^{-H(V|E = e)} \]

[Harmanci et al. Nat. Meth. ('16)]
Representative Expression, Genotype, eQTL Datasets

- mRNA sequencing for 462 individuals
  - Publicly available quantification for protein coding genes
- Approximately 3,000 cis-eQTL (FDR<0.05)
- Genotypes are available from the 1000 Genomes Project
Per eQTL and ICI Cumulative Leakage versus Genotype Predictability

Colors by absolute correlation

[Harmanci et al. Nat. Meth. ('16)]
Cumulative Leakage versus Joint Predictability

[Harmanci et al. Nat. Meth. ('16)]
Linking Attack Scenario

Phenotype dataset (Public)

<table>
<thead>
<tr>
<th>Phenotype ID</th>
<th>HIV Status</th>
<th>Phenotype 1</th>
<th>Phenotype 2</th>
<th>Phenotype q</th>
</tr>
</thead>
<tbody>
<tr>
<td>PID-1</td>
<td>HIV+</td>
<td>0.1</td>
<td>-2.7</td>
<td>... 90.3</td>
</tr>
<tr>
<td>PID-2</td>
<td>HIV-</td>
<td>0.5</td>
<td>8.6</td>
<td>... 63.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>PID-n</td>
<td>HIV-</td>
<td>-0.2</td>
<td>5.4</td>
<td>... 50.3</td>
</tr>
</tbody>
</table>

Phenotype-Genotype correlation dataset

- Phenotype 1 ↔ Variant 1
- Phenotype 2 ↔ Variant 2
- Phenotype q ↔ Variant q

Genotype dataset (Stolen/Hacked/Queried)

<table>
<thead>
<tr>
<th>Genotype ID</th>
<th>Variant 1</th>
<th>Variant 2</th>
<th>Variant q</th>
</tr>
</thead>
<tbody>
<tr>
<td>GID-1</td>
<td>0</td>
<td>1</td>
<td>... 1</td>
</tr>
<tr>
<td>GID-2</td>
<td>2</td>
<td>1</td>
<td>... 0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>GID-m</td>
<td>1</td>
<td>2</td>
<td>... 1</td>
</tr>
</tbody>
</table>

1. Genotype prediction
2. Genotype comparison and matching
3. Predicted/Matched genotypes

[Harmançi et al. Nat. Meth. ('16)]
Steps in Instantiation of a (Mock) Linking Attack

**Step 1**
- Phenotype and genotype selection
  - Absolute Value of Correlation
- G-P correlation dataset

**Step 2**
- Genotype prediction
  - Maximum a Posteriori Genotype
- Prediction methodology

**Step 3**
- Linking
  - Minimum Distance between Predicted and Individual Genotypes
- Auxiliary information
  - Gender, Population, Age
- Estimate Reliability of Linking
  - How far is the linked genotype distance from second in ranked list? ($d_{1,2}$) (Higher: More accurate linking)

[Harmanci et al. Nat. Meth. (’16)]
Levels of Expression-Genotype Model Simplifications:
Extremity based linking with homozygous genotypes

Attacker can estimate the reliability of linkings

Sensitivity: Fraction of correctly linked Individuals among all individuals

PPV: Fraction of correctly linked individuals among selected individuals

[Harmanciet al. Nat. Meth. (in revision)]
Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org

- Long-RNA pipeline
  - RSeqTools & anonymized MRF format

- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels

- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous

- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals
Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

• Large-scale data from consortia #1
  – exRNA.org

• Long-RNA pipeline
  – RSeqTools & anonymized MRF format

• Privacy risk in eQTLs
  – Quantifying & removing variant info from expression levels + eQTLs
  – Linking Attack using extreme expression levels

• Short-RNA pipeline
  – exceRpt
  – Appreciable mapping beyond miRNAs & mRNAs, including exogenous

• Large-scale data from consortia #2
  – Co-authorship networks show data flow through key broker individuals
for a typical cellular sample...

- exRNA samples typically much noisier
- cascade of read-alignment steps mitigates contamination
**extra-cellular RNA processing toolkit**

- automatic pre-processing and QC of sequence reads
- explicit filtering of contaminants & rRNA
- quantification of spike-in sequences and many different smallRNA biotypes
- support for random barcodes (Bioo)
- choice of 3 end-points:
  1. endogenous only
  2. exogenous miRNA + rRNA
  3. exogenous genomes
total reads by biotype

- large contribution from miRNA and mRNA
- also some signal from exogenous sequences
• extremely simple to use (1 input, 1 output)

• can process multiple samples in parallel

• very customisable (choice of smRNA libs, calibrators, etc)
Transcriptome Analysis: Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org

- Long-RNA pipeline
  - RSeqTools & anonymized MRF format

- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels

- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous

- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals
Papers authored by ENCODE consortium members vs. those that use ENCODE data but were not funded by ENCODE

[Wang et al., TIG ('16)]
ENCODE co-authorship network

[Wang et al., TIG ('16)]
Network statistics highlight change in modularity with consortium rollouts (L) & importance of broker role (R)

Coalesced into single module due to ENCODE consortium papers in 2007

Some separation but retention of a unified modular structure

Network statistics highlight change in modularity with consortium rollouts (L) & importance of broker role (R)

Coalesced into single module due to ENCODE consortium papers in 2007

Some separation but retention of a unified modular structure

[Wang et al., TIG ('16)]
Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

- Large-scale data from consortia #1
  - exRNA.org
- Long-RNA pipeline
  - RSeqTools & anonymized MRF format
- Privacy risk in eQTLs
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels
- Short-RNA pipeline
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous
- Large-scale data from consortia #2
  - Co-authorship networks show data flow through key broker individuals
Transcriptome Analysis:
Large-scale data, high-throughput pipelines & privacy considerations

- **Large-scale data from consortia #1**
  - exRNA.org

- **Long-RNA pipeline**
  - RSeqTools & anonymized MRF format

- **Privacy risk in eQTLs**
  - Quantifying & removing variant info from expression levels + eQTLs
  - Linking Attack using extreme expression levels

- **Short-RNA pipeline**
  - exceRpt
  - Appreciable mapping beyond miRNAs & mRNAs, including exogenous

- **Large-scale data from consortia #2**
  - Co-authorship networks show data flow through key broker individuals
Acknowledgments

PrivaSeq. gersteinlab.org
A Harmanci

RSEQtools. gersteinlab.org
L Habegger, A Sboner,
TA Gianoulis, J Rozowsky,
A Agarwal, M Snyder

“Encode authors”
D Wang, KK Yan,
J Rozowsky, E Pan

exRNA.org & exceRpt
R Kitchen
J Rozowsky
A Milosavljevic
M Roth
S Subramanian

Hiring Postdocs. See gersteinlab.org/jobs
Extra

STOP
Info about content in this slide pack

• General PERMISSIONS
  – This Presentation is copyright Mark Gerstein, Yale University, 2015.
  – Please read permissions statement at www.gersteinlab.org/misc/permissions.html.
    – Feel free to use slides & images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).
    – Paper references in the talk were mostly from Papers.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 kwotptp, that can be easily queried from flickr, viz: http://www.flickr.com/photos/mbgmbg/tags/kwotptp