Working with a whole bunch of genetic data
...and teaching data science

Shannon E. Ellis
Assistant Teaching Professor
COGS1 (Spring 2019)
A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD
SHANNON: I LIKE ALL THE THINGS?
SHANNON: I like all the things.

HIGH SCHOOL O'MANY
Upper Darby, PA

S: Science is pretty cool!
SHANNON
I LIKE ALL THE THINGS?

HIGH SCHOOL O'MANY
Upper Darby, PA

...but what did that software actually do?

GENETICS IS AWESOME!

KING'S COLLEGE
Wilkes-Barre, PA

SCIENCE IS PRETTY COOL!
SHANNON: I like all the things.

HIGH SCHOOL O'MANY: Upper Darby, PA

GENETICS IS AWESOME!

JOHNS HOPKINS: Baltimore, MD

KING'S COLLEGE: Wilkes-Barre, PA

SCIENCE IS PRETTY COOL!
SHANNON: I LIKE ALL THE THINGS!

HIGH SCHOOL O'MANY:
Upper Darby, PA

GENETICS IS AWESOME!
...but what did that software actually do?

JOHNS HOPKINS:
Baltimore, MD

KING'S COLLEGE:
Wilkes-Barre, PA

DATA ANALYSIS IS WHERE IT'S AT!

SCIENCE IS PRETTY COOL!
SHANNON: I LIKE ALL THE THINGS!

HIGH SCHOOL O'MANY
Upper Darby, PA

GENETICS IS AWESOME!

but what did that software actually do?

JOHNS HOPKINS
Baltimore, MD

KING'S COLLEGE
Wilkes-Barre, PA

DATA ANALYSIS IS WHERE IT'S AT!

JEFF LEKK: COME DO A POSTDOC WITH ME!
A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD
The quickest history of human genetics

1900 - Rediscovery of Mendel’s work

1950s - DNA is the genetic material & Structure of DNA

2001 - Human Genome Project
   - Thought this would unlock understanding of all disease!
   - Not exactly how things worked out - way more complex than initially thought

Mid-2000s - GWAS set out to sort all this out
   - 2007: 240 papers published
   - 2017: 3800+

2010s - a move toward additional approaches
Multi-omic Data Provide a More Complete Understanding of the Autistic Brain

That’s the title of my thesis dissertation. I’m not great at coming up with titles...ever.
I. Autism Background

II. Transcriptome Analyses
   A. Gene expression differences in the autistic brain
   B. Cross-disorder transcriptomic overlap

III. Epigenome of the Autistic Brain
   A. CpG methylation
   B. nonCpG methylation
RNA-Sequencing in Autism Brains

RNA-Seq

BA10
(Executive Function)

BA44/45
(Semantic Decision)

BA19
(Visual Association)

RNA

RNA

RNA-Seq

RNA-Seq

RNA-Seq

Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism

Simone Gupta1, Shannon Els1, Faran N. Asha2, Anna More3, Joel S. Bader1,3, Janan Zhan2, Andrew B. West3 & Dan E. Ariki1

ARTICLE
Received 28 Sep 2014; Accepted 3 Nov 2014; Published 10 Dec 2014

nature COMMUNICATIONS
The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

gene

exons

slide adapted from alyssa frazee

slide adapted from jeff leek
The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

transcription

RNA

AUCAGUCGAUCACCCGAU

slide adapted from alyssa frazee

slide adapted from jeff leek
The Central Dogma of Genetics

DNA

ACTGACCTAGATCAGTCGATCGATCGTATACGATTACAAAATCATCGGCAT

RNA

AUCAGUCCGAUCACCAGAU

proteins

slide adapted from jeff leek
Two copies of **DNA** -> many **transcripts** -> many **proteins**

<table>
<thead>
<tr>
<th>role in the cell</th>
<th># copies/cell</th>
<th>functional unit</th>
<th># unique functional units</th>
</tr>
</thead>
<tbody>
<tr>
<td>blueprint</td>
<td>2</td>
<td>gene</td>
<td>20,000</td>
</tr>
</tbody>
</table>

**DNA**

**RNA**

**proteins**
Two copies of **DNA** $\rightarrow$ many **transcripts** $\rightarrow$ many **proteins**

<table>
<thead>
<tr>
<th>DNA</th>
<th>RNA</th>
<th>proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>role in the cell</strong></td>
<td><strong># copies/cell</strong></td>
<td><strong>functional unit</strong></td>
</tr>
<tr>
<td>blueprint</td>
<td>2</td>
<td>gene</td>
</tr>
<tr>
<td><strong>DNA</strong></td>
<td><strong>RNA</strong></td>
<td><strong>proteins</strong></td>
</tr>
<tr>
<td><strong>carry out cellular functions</strong></td>
<td><strong>varies</strong></td>
<td><strong>proteins</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(metabolites, hormones, etc.)</td>
</tr>
</tbody>
</table>
Two copies of **DNA** -> many **transcripts** -> many **proteins**

<table>
<thead>
<tr>
<th>DNA</th>
<th>role in the cell</th>
<th># copies/cell</th>
<th>functional unit</th>
<th># unique functional units</th>
</tr>
</thead>
<tbody>
<tr>
<td>blueprint</td>
<td>2</td>
<td>gene</td>
<td>20,000</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RNA</th>
<th>messenger</th>
<th>varies</th>
<th>transcript</th>
<th>~100,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>~360,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>proteins</th>
<th>carry out cellular functions</th>
<th>varies</th>
<th>proteins (metabolites, hormones, etc.)</th>
<th>~100,000</th>
</tr>
</thead>
</table>
Variability at the level of RNA allows for a heart cell to function differently than a brain cell.
Extract mRNA

Prepare a sequencing library

Illumina Sequencing (100bp SE)

Pool of polyA+-selected mRNA

Align those reads to the reference genome

Approximate gene expression by counting the number of reads in each gene

Average read length

Sequencing Output

Green Gene > Blue Gene
I. Autism Background

II. Transcriptome Analyses
   A. Microglia playing a role in the autistic brain
   B. RNA levels show similar patterns across conditions

III. Epigenome of the Autistic Brain
   A. CpG methylation
   B. nonCpG methylation
DNA methylation is most often studied at CpG dinucleotides.

50% methylated

http://www.tokresource.org/
I. Autism Background

II. Transcriptome Analyses
   A. Microglia playing a role in the autistic brain
   B. RNA levels show similar patterns across conditions

III. Epigenome of the Autistic Brain
   A. CpG methylation does not differ
   B. Increased global nonCpG methylation
Conclusions:
Toward a More Complete Understanding of the Autistic Brain
A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD
What makes primary cancer different than metastatic cancer?
What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples
What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

Collect patient samples and information

~6 months
What makes primary cancer different than metastatic cancer?

- Find a researcher with access to patient samples
- Collect patient samples and information: ~6 months
- Extract DNA/RNA from samples: 1-2 wks
- Sequence samples: 2-4 wks
What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

- Collect patient samples and information: ~6 months
- Extract DNA/RNA from samples: 1-2 wks
- Sequence samples: 2-4 wks
- Process sequencing data: 3-6 months
- Analyze data and answer biological question: 1-3 months

Data cleaning: 1 month - 1+ years

Sequence samples: 3-6 months

Extract DNA/RNA from samples: 2-4 wks

Collect patient samples and information: 1-2 wks
What makes primary cancer different than metastatic cancer?

- Find a researcher with access to patient samples
- Collect patient samples and information
- Extract DNA/RNA from samples
- Sequence samples
- Process sequencing data
- Analyze data and answer biological question

Timeline:
- ~6 months
- 1-2 wks
- 2-4 wks
- 3-6 months
- 1-3 months
- 1 month - 1+ years

Total: 2+ years
Biologists have recently gotten pretty good at making their data available to the public.
What makes primary cancer different than metastatic cancer?

1. **Find a researcher with access to patient samples**
   - Time: 3~6 months

2. **Collect patient samples and information**
   - Time: 1-2 wks

3. **Extract DNA/RNA from samples**
   - Time: 2-4 wks

4. **Sequence samples**
   - Time: 3-6 months

5. **Process sequencing data**
   - Time: 1-3 months

6. **Data cleaning**

7. **Analyze data and answer biological question**
   - Time: 1 month - 1+ years

**Total:** 1+ years
Biologists have recently gotten pretty good at making their data available to the public.

...but they’re *not great* at making these data easily accessible and well-annotated.
What makes primary cancer different than metastatic cancer?

1. Find a researcher with access to patient samples
2. Collect patient samples and information
   - 3~6 months
3. Extract DNA/RNA from samples
   - 1-2 wks
4. Sequence samples
   - 2-4 wks
5. Process sequencing data
   - 3-6 months
6. Data cleaning
   - 1-3 months
7. Analyze data and answer biological question
   - 1 month - 1+ years

Total: 1+ years
Measuring Transcription
Next Generation Sequencing (NGS) in one slide

**Step 1:** Extract RNA to get sample of interest

**Step 2:** Chop up RNA into smaller pieces

**Step 3:** Sequence the sample

**Step 3:** Obtain short read data from the sequencer
Next Generation Sequencing (NGS) in one slide

RNA

ATCAGTCGATCACCAGAT

A short read tells you the sequence of that read
We first need to align these reads back to the genome.

The number of reads at each position lets us know abundance.

Coverage vector:

2 6 0 11 6
RNA-Seq = estimate expression across entire genome

expression $\approx$ # RNA-Seq reads
Scaling Up

slide adapted from jeff leek
Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLID System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.
<table>
<thead>
<tr>
<th>Project</th>
<th>No. of Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx</td>
<td>9,962</td>
</tr>
<tr>
<td>Genotype-Tissue Expression Project</td>
<td></td>
</tr>
<tr>
<td>TCGA</td>
<td>11,284</td>
</tr>
<tr>
<td>The Cancer Genome Atlas</td>
<td></td>
</tr>
<tr>
<td>SRA</td>
<td>49,848</td>
</tr>
<tr>
<td>Sequence Read Archive</td>
<td></td>
</tr>
</tbody>
</table>
recount2 is an online resource consisting of RNA-seq gene and exon counts as well as coverage bigWig files for 2041 different studies. It is the second generation of the ReCount project. The raw sequencing data were processed with Rail-RNA as described at bioRxiv 038224 which created the coverage bigWig files. For ease of statistical analysis, for each study we created count tables at the gene and exon levels and extracted phenotype data, which we provide in their raw formats as well as in RangedSummarizedExperiment R objects (described in the SummarizedExperiment Bioconductor package). We also computed the mean coverage per study and provide it in a bigWig file, which can be used with the derfinder Bioconductor package to perform annotation-agnostic differential expression analysis at the expressed regions-level as described at bioRxiv 015370. The count tables, RangedSummarizeExperiment objects, phenotype tables, sample bigWigs, mean bigWigs, and file information tables are ready to use and freely available here. We also created the recount Bioconductor package which allows you to search and download the data for a specific study. By taking care of several preprocessing steps and combining many datasets into one easily-accessible website, we make finding and analyzing RNA-seq data considerably more straightforward.

Related publications


The Datasets

SRP025982  1720  human

We present primary results from the Sequencing Quality Control (SEQC) project, coordinated by the United States Food and Drug Administration. Examining Illuma HiSeq, Life Technologies SOLID and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, we assess RNA sequencing (RNA-seq) performance for sequence discovery and differential expression profiling and compare it to microarray and quantitative PCR (qPCR) data using complementary metrics. At all sequencing depths, we discover unannotated exon-exon junctions, with >80% validated by qPCR. We find that
expression data for ~70,000 human samples

<table>
<thead>
<tr>
<th></th>
<th>GTEx</th>
<th>SRA</th>
<th>TCGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>9,962</td>
<td>49,848</td>
<td>11,284</td>
</tr>
</tbody>
</table>
expression data for ~70,000 human samples

<table>
<thead>
<tr>
<th></th>
<th>GTEx</th>
<th>SRA</th>
<th>TCGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>9,962</td>
<td>49,848</td>
<td>11,284</td>
</tr>
</tbody>
</table>
What makes primary cancer different than metastatic cancer?

- Find a researcher with access to patient samples
- Collect patient samples and information: 3–6 months
- Extract DNA/RNA from samples: 1–2 wks
- Sequence samples: 2–4 wks
- Process sequencing data: 3–6 months
- Determine expression differences between primary cancer and metastatic cancer: 1 month - 1+ years
- Data cleaning: 1–3 months
- Find publicly available RNA-sequencing data from primary cancer and metastasis: Total: 1+ years
What makes primary cancer different than metastatic cancer?

1. Find a researcher with access to patient samples: 3~6 months
2. Collect patient samples and information: 1-2 wks
3. Extract DNA/RNA from samples: 2-4 wks
4. Sequence samples: 3-6 months
5. Process sequencing data: 1-3 months
6. Get already processed and summarized RNA-Seq data from recount2: 1 month - 1+ years
7. Determine expression differences between primary cancer and metastatic cancer: Data cleaning

Total: months
expression data for ~70,000 human samples

Answer meaningful questions about human biology and expression

GTEx  SRA  TCGA
N=9,962 N=49,848 N=11,284
expression data for ~70,000 human samples

Answer meaningful questions about human biology and expression

<table>
<thead>
<tr>
<th>Expression estimates</th>
<th>Phenotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx N=9,962</td>
<td>TCGA N=11,284</td>
</tr>
<tr>
<td>SRA N=49,848</td>
<td></td>
</tr>
</tbody>
</table>
in-silico Phenotyping
SRA phenotype information is far from complete

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>Tissue</th>
<th>Race</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>6620</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6621</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6622</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6623</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6624</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6625</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6626</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6627</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6628</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6629</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6630</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6631</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6632</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6633</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6634</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6635</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6636</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
### SRA phenotype information is far from complete

<table>
<thead>
<tr>
<th></th>
<th>Sex</th>
<th>Tissue</th>
<th>Race</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>6620</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6621</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6622</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6623</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6624</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6625</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6626</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6627</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6628</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6629</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6630</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6631</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6632</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6633</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6634</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6635</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>6636</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
Even when information *is* provided, it’s not always clear...

**Sex across the SRA:**

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>95</td>
</tr>
<tr>
<td>female</td>
<td>2036</td>
</tr>
<tr>
<td>Female</td>
<td>51</td>
</tr>
<tr>
<td>M</td>
<td>77</td>
</tr>
<tr>
<td>male</td>
<td>1240</td>
</tr>
<tr>
<td>Male</td>
<td>141</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3640</strong></td>
</tr>
</tbody>
</table>
Even when information *is* provided, it’s not always clear...

Sex across the **SRA**:

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>95</td>
</tr>
<tr>
<td>female</td>
<td>2036</td>
</tr>
<tr>
<td>Female</td>
<td>51</td>
</tr>
<tr>
<td>M</td>
<td>77</td>
</tr>
<tr>
<td>male</td>
<td>1240</td>
</tr>
<tr>
<td>Male</td>
<td>141</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3640</strong></td>
</tr>
</tbody>
</table>

“1 Male, 2 Female”, “2 Male, 1 Female”, “3 Female”, “DK”, “male and female” “Male (note: ....)”, “missing”, “mixed”, “mixture”, “N/A”, “Not available”, “not applicable”, “not collected”, “not determined”, “pooled male and female”, “U”, “unknown”, “Unknown”
Even when information *is* provided, it’s not always clear...

Sex across the **SRA**:

<table>
<thead>
<tr>
<th>Level</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>95</td>
</tr>
<tr>
<td>female</td>
<td>2036</td>
</tr>
<tr>
<td>Female</td>
<td>51</td>
</tr>
<tr>
<td>M</td>
<td>77</td>
</tr>
<tr>
<td>male</td>
<td>1240</td>
</tr>
<tr>
<td>Male</td>
<td>141</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>3640</strong></td>
</tr>
</tbody>
</table>

“1 Male, 2 Female”, “2 Male, 1 Female”, “3 Female”, “DK”, “male and female” “Male (note: ....)”, “missing”, “mixed”, “mixture”, “N/A”, “Not available”, “not applicable”, “not collected”, “not determined”, “pooled male and female”, “U”, “unknown”, “Unknown”

<table>
<thead>
<tr>
<th># of NAs</th>
<th># w/sex assigned</th>
</tr>
</thead>
<tbody>
<tr>
<td>44,957</td>
<td>4,700</td>
</tr>
</tbody>
</table>
Goal: to accurately predict critical phenotype information for all samples in recount2
Goal:

to accurately predict critical phenotype information for all samples in recount2
Goal:

to accurately predict critical phenotype information for all samples in recount2.
## Missingness limited in GTEx phenotype data

<table>
<thead>
<tr>
<th>Sex</th>
<th>Tissue</th>
<th>Race</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>Lung</td>
<td>White</td>
<td>59</td>
</tr>
<tr>
<td>male</td>
<td>Brain</td>
<td>White</td>
<td>27</td>
</tr>
<tr>
<td>female</td>
<td>Heart</td>
<td>Black or African American</td>
<td>23</td>
</tr>
<tr>
<td>male</td>
<td>Brain</td>
<td>White</td>
<td>51</td>
</tr>
<tr>
<td>male</td>
<td>Skin</td>
<td>White</td>
<td>27</td>
</tr>
<tr>
<td>male</td>
<td>Lung</td>
<td>White</td>
<td>68</td>
</tr>
<tr>
<td>female</td>
<td>Brain</td>
<td>White</td>
<td>61</td>
</tr>
<tr>
<td>female</td>
<td>Adipose Tissue</td>
<td>White</td>
<td>42</td>
</tr>
<tr>
<td>male</td>
<td>Brain</td>
<td>White</td>
<td>40</td>
</tr>
<tr>
<td>female</td>
<td>Uterus</td>
<td>White</td>
<td>33</td>
</tr>
<tr>
<td>female</td>
<td>Nerve</td>
<td>White</td>
<td>50</td>
</tr>
<tr>
<td>male</td>
<td>Muscle</td>
<td>White</td>
<td>54</td>
</tr>
<tr>
<td>female</td>
<td>Ovary</td>
<td>White</td>
<td>31</td>
</tr>
<tr>
<td>male</td>
<td>Blood</td>
<td>White</td>
<td>53</td>
</tr>
<tr>
<td>female</td>
<td>Brain</td>
<td>White</td>
<td>56</td>
</tr>
<tr>
<td>male</td>
<td>Muscle</td>
<td>White</td>
<td>44</td>
</tr>
</tbody>
</table>
Missingness limited in **GTEx phenotype data**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Tissue</th>
<th>Race</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>male Lung</td>
<td>White</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>male Brain</td>
<td>White</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>female Heart</td>
<td>Black or African American</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>male Brain</td>
<td>White</td>
<td>51</td>
</tr>
<tr>
<td>5</td>
<td>male Skin</td>
<td>White</td>
<td>27</td>
</tr>
<tr>
<td>6</td>
<td>male Lung</td>
<td>White</td>
<td>68</td>
</tr>
<tr>
<td>7</td>
<td>female Brain</td>
<td>White</td>
<td>61</td>
</tr>
<tr>
<td>8</td>
<td>female Adipose Tissue</td>
<td>White</td>
<td>42</td>
</tr>
<tr>
<td>9</td>
<td>male Brain</td>
<td>White</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>female Uterus</td>
<td>White</td>
<td>33</td>
</tr>
<tr>
<td>11</td>
<td>female Nerve</td>
<td>White</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>male Muscle</td>
<td>White</td>
<td>54</td>
</tr>
<tr>
<td>13</td>
<td>female Ovary</td>
<td>White</td>
<td>31</td>
</tr>
<tr>
<td>14</td>
<td>male Blood</td>
<td>White</td>
<td>53</td>
</tr>
<tr>
<td>15</td>
<td>female Brain</td>
<td>White</td>
<td>56</td>
</tr>
<tr>
<td>16</td>
<td>male Muscle</td>
<td>White</td>
<td>44</td>
</tr>
</tbody>
</table>

**Sex across GTEx:**

<table>
<thead>
<tr>
<th>level</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>female</td>
<td>3,626</td>
</tr>
<tr>
<td>male</td>
<td>6,036</td>
</tr>
<tr>
<td>NA</td>
<td>0</td>
</tr>
</tbody>
</table>
Goal: to accurately predict critical phenotype information for all samples in recount2
Goal:

to accurately predict critical phenotype information for all samples in recount2
Goal: to accurately predict critical phenotype information for all samples in recount2.

GTEx
Genotype Tissue Expression Project
N=9,662

TCGA
The Cancer Genome Atlas
N=11,284

SRA
Sequence Read Archive
N=49,848

- divide samples
- build and optimize phenotype predictor
- test accuracy of predictor
- predict phenotypes across samples in TCGA

Training Data
Validation Data
Test Data
Goal: to accurately predict critical phenotype information for all samples in recount2
Identify regions with differential expression for each level

Expression

sex

male
female

expression ~ phenotype

βr

phenotype (P)

Apply coefficient estimates to the extracted data

Expression @ filtered regions

male
female

female
male

new data set samples

Build predictor

Filter regions

Extract data

Predict phenotype and assess accuracy in training set data

Prediction accuracy: 100%

Explain the relationship between expression and phenotype:

\[ E \{ \{ E^*_r \}_{r \in S} \} = \hat{\beta}_r \gamma^* \]
gene, exon, exon-exon junction and expressed region RNA-Seq data

SRA
Sequence Read Archive
N=49,848

TCGA
The Cancer Genome Atlas
N=11,284

GTEx
Genotype Tissue Expression Project
N=9,662

Test Data

phenopredict

Input Data

r regions \times \text{individuals}

functions

filter_regions()
build_predictor()
test_predictor()
extract_data()
predict_pheno()

Accuracy

100%
100%
100%
100%

Make predictions!
Let’s get predicting...
Sex prediction is accurate across data sets.

- GTEx (training): 99.9%
- GTEx (validation): 99.8%
- TCGA (test): 99.0%
- SRA: 86.3%

<table>
<thead>
<tr>
<th>Data Set Used</th>
<th>Number of Regions</th>
<th>Number of Samples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (training)</td>
<td>40</td>
<td>4,769</td>
</tr>
<tr>
<td>GTEx (validation)</td>
<td>40</td>
<td>4,769</td>
</tr>
<tr>
<td>TCGA (test)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRA</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sex prediction is accurate across data sets

<table>
<thead>
<tr>
<th>Data Set Used</th>
<th>Number of Regions</th>
<th>Number of Samples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (training)</td>
<td>40</td>
<td>4,769</td>
</tr>
<tr>
<td>GTEx (validation)</td>
<td>40</td>
<td>4,769</td>
</tr>
<tr>
<td>TCGA (test)</td>
<td>40</td>
<td>11,245</td>
</tr>
<tr>
<td>SRA</td>
<td>40</td>
<td>3,640</td>
</tr>
</tbody>
</table>

Accuracy:
- GTEx (training): 99.9%
- GTEx (validation): 99.8%
- TCGA (test): 99.0%
- SRA: 86.3%
Are a few studies driving decrease in accuracy across the SRA samples?
To assess misreporting of sex in the SRA, we can use Y-chromosome expression.
To assess misreporting of sex in the SRA, we can use Y-chromosome expression.
To assess misreporting of sex in the SRA, we can use Y-chromosome expression.
Expression from the Y chromosome suggests misreporting of sex in the SRA

SRP040547
Expression from the Y chromosome suggests misreporting of sex in the SRA
There’s a well-documented history of male sex-bias in biomedical research...

...so let’s take a closer look within recount2
Across the ~70,000 samples in *recount2*, there are more samples predicted to be *female* than *male*. 
GTEx is male-biased; TCGA is female-biased

...but that has been previously reported by the consortia
Is there a sex bias across the SRA?
No evidence for sex bias in samples across the SRA
There are many female-only, male-only, and male-female projects available in recount.
Can we use expression data to predict tissue?

http://www.rna-seqblog.com/
Tissue prediction is accurate across data sets.

- **GTEx (training):** 97.7%
- **GTEx (validation):** 96.6%
- **TCGA (test):** 76.8%
- **SRA:** 51.9%

<table>
<thead>
<tr>
<th>Data Set Used</th>
<th>Number of Regions</th>
<th>Number of Samples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (training)</td>
<td>2,281</td>
<td>4,769</td>
</tr>
<tr>
<td>GTEx (validation)</td>
<td>2,281</td>
<td>4,769</td>
</tr>
<tr>
<td>TCGA (test)</td>
<td>2,281</td>
<td>7,317</td>
</tr>
<tr>
<td>SRA</td>
<td>2,281</td>
<td>8,951</td>
</tr>
</tbody>
</table>
Prediction is more accurate in healthy tissue.

<table>
<thead>
<tr>
<th>Tissue Prediction</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (training)</td>
<td>97.7%</td>
</tr>
<tr>
<td>GTEx (validation)</td>
<td>96.6%</td>
</tr>
<tr>
<td>TCGA: healthy tissue</td>
<td>92.7%</td>
</tr>
<tr>
<td>TCGA: cancer</td>
<td>75.3%</td>
</tr>
<tr>
<td>SRA</td>
<td>51.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Regions</th>
<th>2,281</th>
<th>2,281</th>
<th>2,281</th>
<th>2,281</th>
<th>2,281</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples (N)</td>
<td>4,769</td>
<td>4,769</td>
<td>613</td>
<td>6,704</td>
<td>8,951</td>
</tr>
</tbody>
</table>
Across the samples in *recount*, brain, blood, and skin are the three most frequently predicted tissues types.
A sample reported to be *Intestine* is predicted to be *Colon*. That makes good sense.
Tissue prediction is largely accurate across *recount2*

- Tissue can be accurately predicted from expression data.
- Discordant predictions are often made to biologically similar tissues.
- Sometimes, predictions are inaccurate.
What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

Collect patient samples and information

Collect patient samples and information

Extract DNA/RNA from samples

Sequence samples

Process sequencing data

Get already processed and summarized RNA-Seq data from recount2

Determine expression differences between primary cancer and metastatic cancer

Data cleaning

Total: months

3~6 months

1-2 wks

2-4 wks

3-6 months

1-3 months

1 month - 1+ years

3~6 months
What makes primary cancer different than metastatic cancer?

Collect patient samples and information: 3~6 months

Find a researcher with access to patient samples

Extract DNA/RNA from samples: 1-2 wks

Sequence samples: 2-4 wks

Process sequencing data: 3-6 months

Determine expression differences between primary cancer and metastatic cancer: 1 month - 1+ years

Get already processed and summarized RNA-Seq data from recount2 with sample information

Data cleaning: 1-3 months

Total: months
Ok. Ok. What about actually *using* these data and predictions...?
What makes primary cancer different than metastatic cancer?
A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients

Seon-Kyu Kim\textsuperscript{a,1}, Seon-Young Kim\textsuperscript{a,1}, Jeong-Hwan Kim\textsuperscript{a}, Seon Ae Roh\textsuperscript{b,c}, Dong-Hyung Cho\textsuperscript{c,d}, Yong Sung Kim\textsuperscript{a,c,**}, Jin Cheon Kim\textsuperscript{b,c,*}

\textsuperscript{a}Medical Genomics Research Centre, Korea Research Institute of Bioscience and Biotechnology, Daejeon, Korea
\textsuperscript{b}Department of Surgery, University of Ulsan College of Medicine, Seoul, Korea
\textsuperscript{c}Department of Cancer Research, Institute of Innovative Cancer Research and Asan Institute for Life Sciences, Asan Medical Centre, Seoul, Korea
\textsuperscript{d}Graduate School of East-West Medical Science, Kyung Hee University, Gyeonggi-do, Korea

\textit{Molecular Oncology, July 2014}
Kim et al. analysis looked to identify genes that contribute to metastasis in colon cancer.

N=18

1. Healthy Colon (NC)
2. Primary Cancer (PC)
3. Liver Metastasis (MC)
*Kim et al.* analysis looked to identify genes that contribute to metastasis in colon cancer.
Predictions can be used to:

1. Identify studies of interest
2. Appropriately analyze data
Predictions can be used to:

1. Identify studies of interest
2. Appropriately analyze data
Predictions can be used to:

1. Identify studies of interest
2. Appropriately analyze data

Are the same genes found when sex is included in the analysis?

Significantly differentially expressed genes by analysis:

- MC:PC: Kim et al., 1846 vs. no covariates, 1088
- Comparison: Kim et al., 2861 vs. no covariates, 946

NC: non-cancerous
PC: primary cancer
MC: metastatic cancer
Predictions can be used to:

(1) Identify studies of interest

(2) Appropriately analyze data
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B?

**Step 1:** Order genes from each analysis by significance

**Step 2:** Starting with the most significant gene, determine the number of concordant genes at each gene.
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B?
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B?
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B?
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B?

The top results from A and B are the same.

Some differences at less significant genes.
Concordance at the Top (CAT) Plots

How similar are the results from Analysis A and Analysis B?

The results of the red condition are less similar between Analysis A and B than the green condition.
Predictions can be used to:

(1) Identify studies of interest

(2) Appropriately analyze data
Predictions can be used to:

(1) Identify studies of interest
(2) Appropriately analyze data
Predictions can be used to:

1. Identify studies of interest
2. Appropriately analyze data
Predictions can be used to:

1. Identify studies of interest
2. Appropriately analyze data
Loss of concordance suggests that differential expression is detecting tissue differences, not cancer-related changes.
We have expression data from both healthy liver and colon samples (GTEx)...
So...what if we compared the MC:PC results with differential expression between colon and liver?

**Hypothesis:** MC:PC results should be most similar to GTEx colon vs. liver.
Comparison of results with GTEx colon vs. liver suggests differential expression results detecting tissue differences.
What makes primary cancer different than metastatic cancer?

Find a researcher with access to patient samples

Collect patient samples and information

3~6 months

Extract DNA/RNA from samples

1-2 wks

Sequence samples

2-4 wks

Process sequencing data

3-6 months

Data cleaning

1-3 months

Determine expression differences between primary cancer and metastatic cancer

1 month - 1+ years

Total: days-months
The Leek group

- Jack Fu
- Aboozar Hadavand
- Leslie Myint
- Kayode Sosina
- Sara Wang
- Jeff Leek

Collaborators

- Andrew Jaffe
- Kasper Hansen
- Margaret Taub
- Leah Jager
- Sean Kross
- Ben Langmead
- Abhi Nellore
- Kai Kammers
- Leo Collado-Torres
- Ashkaun Razmara
A quick tour of a geneticist turned data scientist

Background

Projects

1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD
Chromebook Data Science (CBDS)
Find a **partner organization**

Collaboratively **develop course content**

Develop **new technology** as needed

Design in-person **tutoring program**

Launch program, **teach the stuff** & get learners **jobs**
Chromebook Data Science
Course 7: Data Visualization

The instructor has published 100% of this course.

Data Visualization

Jeffrey Leek

$0.00 / MINIMUM  $25.00 / SUGGESTED
YOU PAY

$25.00

Add Course To Cart
The Timeline

Feb 2018

- content development starts

April

- first meeting with Yo

May

- May 21st: Learning Begins!

Launch program, teach the stuff & get learners jobs
The Timeline

Feb 2018: 
- Content development starts

April: 
- First meeting with Yo

May: 
- May 21st: Learning Begins!

Sept: 
- Aug 31: Projected Course Set Completion

Launch program, **teach the stuff** & get learners **jobs**
The Timeline

Feb 2018

Feb 2018
content development starts

April

first meeting with Yo

May

May 21st: Learning Begins!

#1 #2 #3 #4 #5

0-3 4-5 6-7 8-9 10-11

June
July
Aug
Sept

Aug 31: Projected Course Set Completion

Launch program, teach the stuff & get learners jobs
The Timeline

Feb 2018
content development starts

April
first meeting with Yo

May
Learning Begins!

June
#1 0-3

July
#2 4-5

Aug
#3 6-7

Sept
#4 8-9

Oct
#5 10-11

Oct 5th Learners finish coursework!!!

Launch program, teach the stuff & get learners jobs
The Timeline

- **Feb 2018**: Content development starts
- **April**: First meeting with Yo
- **May 21st**: Learning Begins!
- **Sept**: Launch CBDS
- **Oct 5th**: Learners finish coursework!!!

Launch program, **teach the stuff** & get learners **jobs**
rstudio and Chromebook Data Science are so fun I am working on them on a Saturday night so thanks a buncha buncha @kierisi for the recommendation.
Chromebook Data Science Team

Collaboratively develop course content
A quick tour of a geneticist turned data scientist

Background

Projects
1. PhD work studying the genetic basis of autism
2. Postdoctoral work working with 70,000 samples
3. Working toward accessible data science education

What I do here at UCSD
COGS9 : Intro to Data Science

Brad has previously written about COGS9 in his super interesting and thoughtful blog post Data Science at UC San Diego, but very briefly here, COGS9 is an intro level course designed to get undergraduate students interested in data science, familiar with what data science is, and excited to learn more. It is neither math nor computationally-heavy, but is rather taught through concepts and examples. When it first ran there were 24 students. Now, each time it is offered, there are hundreds. This quarter, when numbers settled out, I had 326 students in the course.

http://www.shanellis.com/blog/cogs9-introduction-to-data-science/
Welcome to COGS18: Introduction to Python!

COGS 18 · Shannon Ellis · Spring 2019 · UCSD

Overview

Introduction to Python (COGS18) is a course offered by the Department of Cognitive Science of UC San Diego, taught by Shannon Ellis. It is a hands-on programming course, focused on teaching students in Cognitive Science and related disciplines an introduction on how to productively use Python.

Current Iteration

Introduction to Python is currently running for Spring Quarter 2019, for which you can check out the current syllabus and schedule. Course lectures are recorded and are publicly available as screencasts from here.

https://cogs18.github.io/intro/
COGS108 - Data Science in Practice

Course materials for Hands-On Data Science.

UC San Diego  COGS108@gmail.com

Repositories 23  People 19  Teams 4  Projects 0  Settings

Pinned repositories

Overview
Overview and map of the organization, which services COGS108: Hands-On Data Science, from UCSD.

Lectures-Sp19
Slides and Notebooks used in Lecture for Sp19

Section_Workbooks
Workbooks for practice during discussion section

Tutorials
Tutorial notebooks for hands-on data science, following along with the course topics.

Projects
Final Project materials and description.

Readings
A curated list of suggested reading materials.

https://github.com/COGS108