Reproducible RNA-seq analysis with recount2

Leonardo Collado-Torres
@fellgemon
#bioc2017

LIEBER INSTITUTE for
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES
GTEx

NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS

TCGA BY THE NUMBERS

TCGA produced over 2.5 PETabytes of data

TCGA data describes 33 DIFFERENT TUMOR TYPES, including 10 RARE CANCERS

To put this into perspective, 1 petabyte of data is equal to 212,000 DVDs

Based on paired tumor and normal tissue sets collected from 11,000 patients using 7 DIFFERENT DATA TYPES

TCGA RESULTS & FINDINGS

- Molecular basis of cancer
- Revolutionized how cancer is classified
- Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development
- Therapeutic targets
- TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations

What's next?

20 COLLABORATING INSTITUTIONS across the United States and Canada

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with...
Sequence Read Archive (SRA) growth

1 Pbp

Terabases

Total

Open access

2 -> 3 Pbp in ~10 months

Slide adapted from Ben Langmead
Preprocess → Align reads end-to-end to genome

Preprocess → Aggregate duplicate reads → Split into readlets → Aggregate duplicate readlets → Align readlets to genome → Correlation clustering for readlet alignments → Call splice junction → Align readlets to junction co-occurrence index → Write spliced alignment BAMs → Write junction & indel BEDs

Preprocess → Bowtie 2

Preprocess → Align reads end-to-end to genome → Bowtie

Preprocess → Align readlets to genome → Bowtie

Preprocess → Align readlets to junction co-occurrence index → Bowtie

Merge exon differentials → Compile sample coverages → Write bigWigs → Write normalization factors

http://rail.bio/
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 in the recount2 paper and at Nellore et al, Genome Biology, 2016 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 Collado-Torres et al, Genome Research, 2017. 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

Reads

Gene

Isoform 1

Isoform 2

Potential isoform 3

Expressed region 1:
potential exon 5

Exon 1
Exon 2
Exon 3
Exon 4

Covery

exon 1
exon 2
exon 3
exon 4
exon 1

exon 2

exon 3
disjoint exon 1

disjoint exon 2

disjoint exon 3
\[ 3 + 3 + 5 + 4 + 4 = 19 \]
\[ 2 + 2 + 3 = 7 \]
\[ 4 + 4 + 2 = 10 \]
\[ \text{Gene} = 19 + 7 + 10 = 36 \]
\[
\sum_{i}^{n} \frac{\text{coverage}_i}{\text{Read Length}} \times \frac{\text{target}}{\text{mapped}} = \text{scaled read counts}
\]
### Genome Coverage

| 3 | 3 | 5 | 4 | 4 | 2 | 2 | 3 | 1 | 3 | 3 | 1 | 4 | 4 | 2 | 1 |

**AUC** = area under coverage = 45
\[
\sum_{i}^{n} \frac{\text{coverage}_i}{\text{Read Length}} \times \frac{\text{target}}{\text{mapped}} = \text{scaled read counts}
\]

\[
\sum_{i}^{n} \frac{\text{coverage}_i}{\text{AUC}} \times \text{target} = \text{scaled read counts}
\]
> library('recount')

> download_study('ERP001942', type='rse-gene')

> load(file.path('ERP001942', 'rse_gene.Rdata'))

> rse <- scale_counts(rse_gene)

https://github.com/leekgroup/recount-analyses/
Recount has been very useful for me over the years in developing and testing methods.
> library('recount')

> download_study('SRP029880', type='rse-gene')
> download_study('SRP059039', type='rse-gene')
> load(file.path('SRP029880', 'rse_gene.Rdata'))
> load(file.path('SRP059039', 'rse_gene.Rdata'))
> mdat <- do.call(cbind, dat)

https://github.com/leekgroup/recount-analyses/
Average Log2 Fold Change

- Same Tissue
- Different Tissues
- Tissue vs. Batch

Collado Torres et al. Nat. Biotech 2017
Reads

Gene

Isoform 1

Isoform 2

Potential isoform 3

Expressed region 1: potential exon 5
Postmortem Human Brain Samples

Discovery data
- Fetal
- Infant
- Child
- Teen
- Adult
- 50+

6 / group, N = 36

Replication data
- Fetal
- Infant
- Child
- Teen
- Adult
- 50+

6 / group, N = 36

Jaffe et al, Nat. Neuroscience, 2015
SOX11, 619 bp from tss: overlaps 3'
expression data for ~70,000 human samples

slide adapted from Shannon Ellis
expression data for ~70,000 human samples

Answer meaningful questions about human biology and expression

<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>
expression data for ~70,000 human samples

Answer meaningful questions about human biology and expression

<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>
Even when information *is* provided, it’s not always clear…

<table>
<thead>
<tr>
<th>Category</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”

slide adapted from Shannon Ellis
SRA phenotype information is far from complete

<table>
<thead>
<tr>
<th>SubjectID</th>
<th>Sex</th>
<th>Tissue</th>
<th>Race</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>6620</td>
<td>NA</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6621</td>
<td>NA</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6622</td>
<td>NA</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6623</td>
<td>NA</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6624</td>
<td>NA</td>
<td>female</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6625</td>
<td>NA</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6626</td>
<td>NA</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6627</td>
<td>NA</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6628</td>
<td>NA</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6629</td>
<td>NA</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6630</td>
<td>NA</td>
<td>male</td>
<td>liver</td>
<td>NA</td>
</tr>
<tr>
<td>6631</td>
<td>NA</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
</tr>
<tr>
<td>6632</td>
<td>NA</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
</tr>
<tr>
<td>6633</td>
<td>NA</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
</tr>
<tr>
<td>6634</td>
<td>NA</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
</tr>
<tr>
<td>6635</td>
<td>NA</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
</tr>
<tr>
<td>6636</td>
<td>NA</td>
<td>NA</td>
<td>blood</td>
<td>NA</td>
</tr>
</tbody>
</table>
Goal:

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

to accurately predict critical phenotype information for all samples in recount

GTEx
Genotype Tissue Expression Project
N=9,662

TCGA
The Cancer Genome Atlas
N=11,284

SRA
Sequence Read Archive
N=49,848

build and optimize phenotype predictor

divide samples

build and optimize phenotype predictor

test accuracy of predictor

test set

train set

slide adapted from Shannon Ellis
Goal:

to accurately predict critical phenotype information for all samples in *recount*
Goal: to accurately predict critical phenotype information for all samples in recount.
select_regions()

Input Data

Coverage matrix (data.frame)
Region information (GRanges)

Output:

slide adapted from Shannon Ellis
Sex prediction is accurate across data sets

<table>
<thead>
<tr>
<th></th>
<th>GTEx: training</th>
<th>GTEx: test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Number of Samples (N)</td>
<td>4,769</td>
<td>4,769</td>
</tr>
</tbody>
</table>

slide adapted from Shannon Ellis
Sex prediction is accurate across data sets.

<table>
<thead>
<tr>
<th></th>
<th>GTeX: training</th>
<th>GTeX: test</th>
<th>TCGA</th>
<th>SRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Number of Samples (N)</td>
<td>4,769</td>
<td>4,769</td>
<td>11,245</td>
<td>3,640</td>
</tr>
</tbody>
</table>

Proportion Correct

- GTeX: training: 99.8%
- GTeX: test: 99.6%
- TCGA: 99.4%
- SRA: 88.5%

slide adapted from Shannon Ellis
Can we use expression data to predict tissue?
Tissue prediction is accurate across data sets.

<table>
<thead>
<tr>
<th></th>
<th>GTEx: training</th>
<th>GTEx: test</th>
<th>TCGA</th>
<th>SRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Regions</td>
<td>589</td>
<td>589</td>
<td>589</td>
<td>589</td>
</tr>
<tr>
<td>Number of Samples (N)</td>
<td>4,769</td>
<td>4,769</td>
<td>7,193</td>
<td>8,951</td>
</tr>
<tr>
<td>Proportion Correct</td>
<td>97.3%</td>
<td>96.5%</td>
<td>71.9%</td>
<td>50.6%</td>
</tr>
</tbody>
</table>

slide adapted from Shannon Ellis
Prediction is more accurate in healthy tissue.

Tissue Prediction

- GTEx: training: 97.3%
- GTEx: test: 96.5%
- TCGA: healthy tissue: 91.0%
- TCGA: cancer: 70.2%
- SRA: 50.6%

Number of Regions | 589 | 589 | 589 | 589 | 589
Number of Samples (N) | 4,769 | 4,769 | 613 | 6,579 | 8,951

slide adapted from Shannon Ellis
> library('recount')

> download_study('ERP001942', type='rse-gene')

> load(file.path('ERP001942', 'rse_gene.Rdata'))

> rse <- scale_counts(rse_gene)

> rse_with_pred <- add_predictions(rse_gene)

https://github.com/leekgroup/recount-analyses/
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>

Answer meaningful questions about human biology and expression

- GTEx: \(N=9,962\)
- SRA: \(N=49,848\)
- TCGA: \(N=11,284\)

slide adapted from Shannon Ellis
Collaborators

The Leek Group
Jeff Leek
Shannon Ellis

Hopkins
Ben Langmead
Chris Wilks
Kai Kammers
Kasper Hansen
Margaret Taub

OHSU
Abhinav Nellore

LIBD
Andrew Jaffe
Emily Burke
Stephen Semick
Carrie Wright
Amanda Price
Nina Rajpurohit

Funding

NIH R01 GM105705
NIH 1R21MH109956
CONACyT 351535
AWS in Education
Seven Bridges
IDIES SciServer
http://research.libd.org/recountWorkshop/

help(package = recountWorkshop)

file.edit(
  system.file('doc/recount-workshop.Rmd', package = 'recountWorkshop')
)

Leonardo Collado-Torres
@felligeron
#bioc2017
(Multiple) Postdoc positions available to
- develop methods to process and analyze data from recount2
- use recount2 to address specific biological questions
This project involves the Hansen, Leek, Langmead and Battle labs at JHU

Contact: Kasper D. Hansen (khansen@jhsph.edu | www.hansenlab.org)