Detecting eQTLs from high-dimensional sequencing data using recount2

Kai Kammers

The Johns Hopkins University School of Medicine
Baltimore, MD, USA

July 7, 2017
RNA-seq

DNA

Exon 1

Exon 2

Exon 3

Exon 4

Exon 5

RNA

Exon 1

Exon 2

Exon 3

Exon 4

Exon 5

Alternative Splicing

mRNA

1

2

3

4

5

Translation

Protein A

Translation

Protein B

Translation

Protein C

RNA-seq: Alignment using Rail-RNA

[ Nellore et al. 2016, Bioinformatics, PMID: 27592709 ]
<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>
Rail-RNA: Scalable analysis of RNA-seq splicing and coverage

Abhinav Nellore\textsuperscript{1,2,3,*}, Leonardo Collado-Torres\textsuperscript{2,3,4}, Andrew E. Jaffe\textsuperscript{2,3,4,5}, José Alquicira-Hernández\textsuperscript{2,6}, Christopher Wilks\textsuperscript{1,3}, Jacob Pritt\textsuperscript{1,3}, James Morton\textsuperscript{7}, Jeffrey T. Leek\textsuperscript{2,3}, and Ben Langmead\textsuperscript{1,2,3,*}

[ Nellore et al. 2016, Bioinformatics, PMID: 27592709 ]
recount2

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>
Kai Kammers: Detecting eQTLs from high-dimensional sequencing data using recount2

recount2

expression data for ~70,000 human samples

[ Slide courtesy of Shannon Ellis ]
recount2

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

[ Slide courtesy of Shannon Ellis ]
recount2

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></td>
</tr>
<tr>
<td>SRA: N=49,848</td>
<td></td>
</tr>
<tr>
<td>TCGA: N=11,284</td>
<td></td>
</tr>
</tbody>
</table>

[ Slide courtesy of Shannon Ellis ]
recount2

- Provides expression summaries at levels of genes, junctions, exons and coverage vectors.

- 70,000 human RNA-seq samples from the SRA, GTEx, TCGA.

- Workflows illustrating differential expression analysis, meta-analysis, annotation-free base-level analysis, and replication of smaller studies using data from larger studies.

- Bioconductor package
  http://bioconductor.org/packages/recount.

- The resource is available at
  https://jhubiostatistics.shinyapps.io/recount/.
A multi-experiment resource of analysis-ready RNA-seq gene and exon count datasets

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 deriveR package to perform expression-agnostic differential expression analysis at the expressed regions-level as described at Collado-Torres et al, Genome Research, 2017.

The count tables, RangedSummarizedExperiment objects, phenotype tables, sample bigWigs, mean bigWigs, and 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

Show | 10 entries

accession | number of samples | species | abstract | gene | exon | junctions | phenotype | files | info
---|---|---|---|---|---|---|---|---|---
SRP025982 | 1720 | human | We present primary results from the Sequencing Quality Control (SEQC) project, coordinated by the United States Food and Drug Administration. Examining Illumina 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 measurements of relative expression are | RSE counts | RSE counts | RSE jx bed | jx cov counts | link | link

[ https://jhubiostatistics.shinyapps.io/recount/ ]
eQTL analysis

- **eQTL (expression Quantitative Trait Locus):** Genomic locus (genotype, SNP) that contribute to variation in expression levels of mRNAs (gene/transcript expression).

\[
E = \alpha + \beta S + \sum_{k=1}^{K} \gamma_k C_k + \epsilon.
\]

MatrixEQTL package in R.
eQTL analysis

- **eQTL (expression Quantitative Trait Locus):** Genomic locus (genotype, SNP) that contribute to variation in expression levels of mRNAs (gene/transcript expression).

- **Approach:** Fit a linear model for each gene-SNP pair to test the association between transcript expression $E$ and genotype $S$ (and additional covariates $C_k$, $k = 1, ..., K$)

$$ E = \alpha + \beta S + \sum_{k} \gamma_k C_k + \epsilon. $$
eQTL analysis

- **eQTL (expression Quantitative Trait Locus):** Genomic locus (genotype, SNP) that contribute to variation in expression levels of mRNAs (gene/transcript expression).

- **Approach:** Fit a linear model for each gene-SNP pair to test the association between transcript expression $E$ and genotype $S$ (and additional covariates $C_k, k = 1, ..., K$)

  $$ E = \alpha + \beta S + \sum_k \gamma_k C_k + \epsilon. $$

- MatrixEQTTL package in R.
eQTL example

AA: rs2523404 and HLA-F

log2 (FPKM + 1)

Genotype

0 1 2
local eQTLs [ or cis eQTLs ]:
eQTLs that map to the approximate location of their gene-of-origin.

distant eQTLs [ or trans eQTLs ]:
eQTLs that map far from the location of their gene-of-origin, often on different chromosomes.

[ details: Albert and Kruglyak, 2015, Nat Rev Genet, PMID: 25707927 ]
eQTL data

- Pheno data set (IDs, covariates).
- Gene expression data set.
- Genotype data set.
- Genomic positions of transcripts.
- Genomic positions of SNPs.
gEUVADIS: Genetic European Variation in Health and Disease

- gEUVADIS RNA-sequencing project
  [http://www.geuvadis.org/web/geuvadis/rnaseq-project]

- Transcriptome and genome sequencing data on 465 lymphoblastoid cell line (LCL) samples from 5 populations of the 1000 Genomes Project.

- Downloaded and preprocessed genotype and RNA-sequencing data.

- Performed eQTL analysis with CEU people [approx. 10 min].

- Result: 3,801 cis eQTLs and 127 trans eQTLs [p.adj < 0.05].
library('recount')

download_study('ERP001942', type='rse-gene')  #(76.8 MB)

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

rse <- scale_counts(rse_gene)
Reproducible RNA-seq analysis using *recount2*

Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper D Hansen, Andrew E Jaffe, Ben Langmead & Jeffrey T Leek

*Affiliations | Corresponding authors*

Published online 11 April 2017
GeneSTAR Research Center: Genetic Studies of Atherosclerosis Risk
http://www.genestarstudy.com/
PIs: Rasika Mathias, Diane Becker, Lewis Becker, ...

**Goal:** Understanding the biology of platelet aggregation.
Introduction

- GeneSTAR Research Center: Genetic Studies of Atherosclerosis Risk
  http://www.genestarstudy.com/
  PIs: Rasika Mathias, Diane Becker, Lewis Becker, ...

**Goal:** Understanding the biology of platelet aggregation.

- Platelets in the circulating blood mediate normal hemostasis, but may also initiate pathological arterial thrombosis that produces heart attacks and strokes.

- GWAS studies have identified common variants associated with platelet aggregation.

- The biological mechanism has remained largely undefined because most signals have occurred in introns or intergenic regions rather than in protein coding regions of known genes.
Platelets are derived from megakaryocytes (large platelet-producing bone marrow cells) in the bone marrow, but themselves are anucleate.
Platelets are derived from megakaryocytes (large platelet-producing bone marrow cells) in the bone marrow, but themselves are anucleate.
Introduction

- Platelets are derived from megakaryocytes (large platelet-producing bone marrow cells) in the bone marrow, but themselves are anucleate.

Efficient Generation of Megakaryocytes From Human Induced Pluripotent Stem Cells Using Food and Drug Administration-Approved Pharmacological Reagents

YANFENG LIU, a,* YING WANG, a,b,c,* YONGXING GAO, a,b JESSICA A. FORBES, a REHAN QAYYUM, d LEWIS BECKER, e LINZHAO CHENG, a,b ZACK Z. WANG a,b

- Induced Pluripotent Stem Cells $\rightarrow$ Megakaryocytes $\rightarrow$ Platelets

[ Liu et al. 2015, Stem Cells Transl Med, PMID: 25713465 ]
Integrity of Induced Pluripotent Stem Cell (iPSC) Derived Megakaryocytes as Assessed by Genetic and Transcriptomic Analysis

Kai Kammers¹,², Margaret A. Taub², Ingo Ruczinski², Joshua Martin³, Lisa R. Yanek³, Alyssa Frazee², Yongxing Gao⁴, Dixie Hoyle⁴, Nauder Faraday³, Diane M. Becker³, Linzhao Cheng⁴, Zack Z. Wang⁴, Jeff T. Leek², Lewis C. Becker³*, Rasika A. Mathias³

¹ Division of Biostatistics and Bioinformatics, Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, United States of America, ² Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America, ³ The GeneSTAR Research Program, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America, ⁴ Division of Hematology and Institute for Cell Engineering, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America

[ Kammers et al. 2017, PloS one, PMID: 28107356]
eQTL analysis

eQTL 2D plot (AA)

SNP position (Gb)

Gene position (Gb)

1

0 0.2 0.6 1 1.2 1.6 2 2.2 2.6 3

2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

Kai Kammers : Detecting eQTLs from high-dimensional sequencing data using recount2
eQTL analysis

eQTLs for AA, layover EA, direct matches

SNP position (Gb)  Gene position (Gb)
1 0 0.2 0.6 1 1.2 1.6 2 2.2 2.6 3
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

cis−eQTLs (EA):
p .a d j < 0.05
p .a d j ≥ 0.05
not in EA

cis−eQTLs (EA):
p .a d j < 0.05
p .a d j ≥ 0.05
not in EA