LIBD

Using the recount2 resource and related tools

Leonardo Collado-Torres
@fellgernon @LieberInstitute
#CONABIO2019

LIEBER INSTITUTE *for*
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES

History

Undergrad in Genomic Sciences

2005-2009

Data Science Division Leader

2009-2011

Ph.D. Biostatistics

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

2011-2016

PIs:

- Jeff Leek: 2012+
- Andrew Jaffe: 2013+

Staff Scientist I → II
Data Science Team I

LIBD

August 2016+

PI: Andrew Jaffe

Interests

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

2008+

@fellgernon
2010+

LIBD Rstats

@LIBDrstats
2018+

Comunidad de Desarrolladores de Software en Bioinformática

@CDSBMexico
2018+

- BioC 2008-2011, 2014, 2017
- useR!2013
- rOpenSci unconf 2018
- RStudio::conf 2019

RStudio

Guest
@RLadiesBmore

Blog: http://lcolladotor.github.io
2011+

R-bloggers

weekly

FB: 75k, Tw: 66k

DataCamp COMMUNITY

Defunct: BmoreBiostats, Biostats Cultural Mixers

recount2 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 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

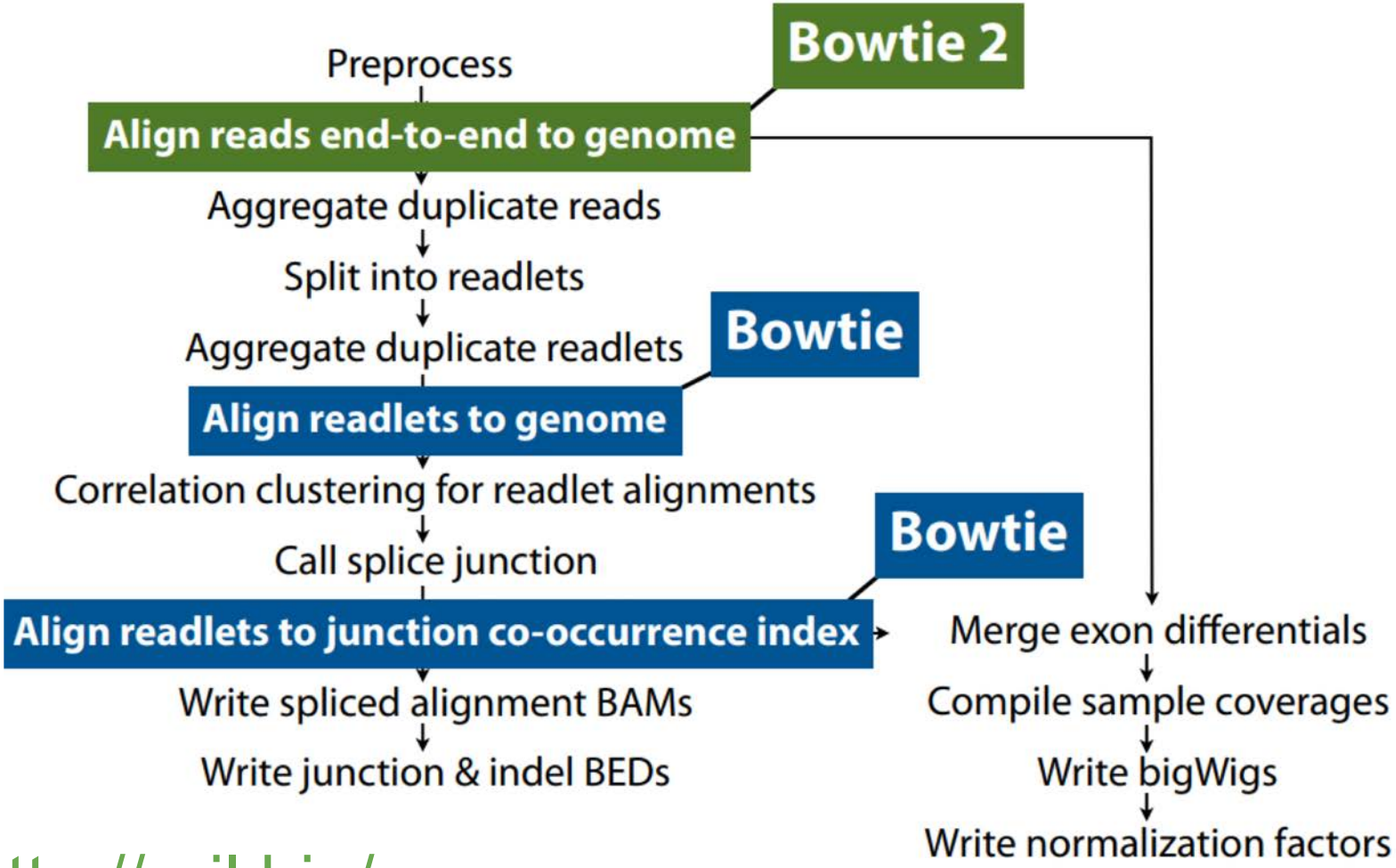
Collado-Torres L, **Nellore A**, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using *recount2*. *Nature Biotechnology*, 2017. doi: 10.1038/nbt.3838.

The Datasets

https://jhubiostatistics.shinyapps.io/recount/

Show 10 entries

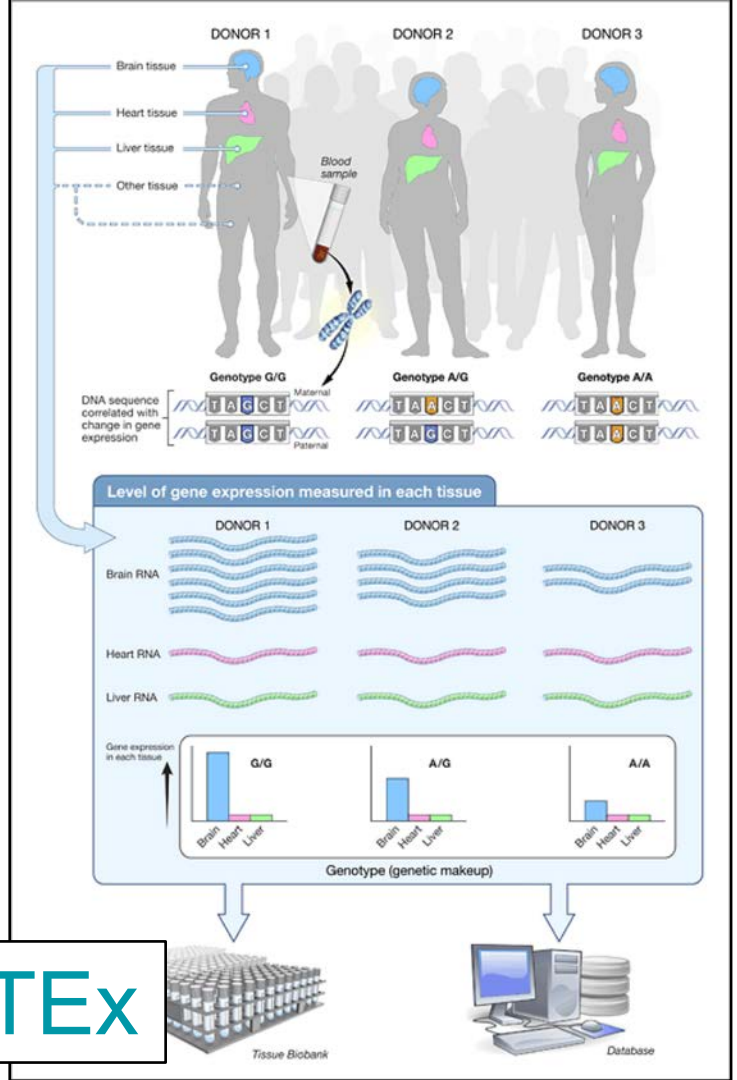
accession	number of samples	species	abstract	gene	exon	junctions	phenotype	files info
All	All	All	libd	All	All	All	All	All
SRP045638	72	human	RNAseq data of 36 samples across human brain development by age group from LIBD	RSE counts	RSE counts	RSE jx_bed jx_cov counts	link	link



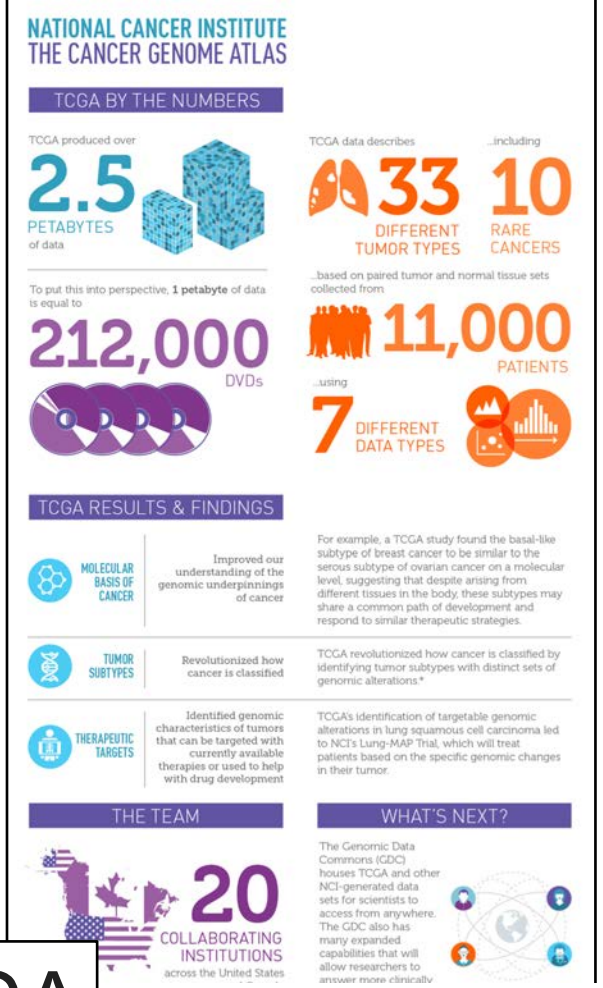
http://rail.bio/ by Abhinav Nellore

Slide adapted from Ben Langmead

http://blogs.citrix.com/2012/10/17/announcing-general-availability-of-sharefile-with-storagezones/

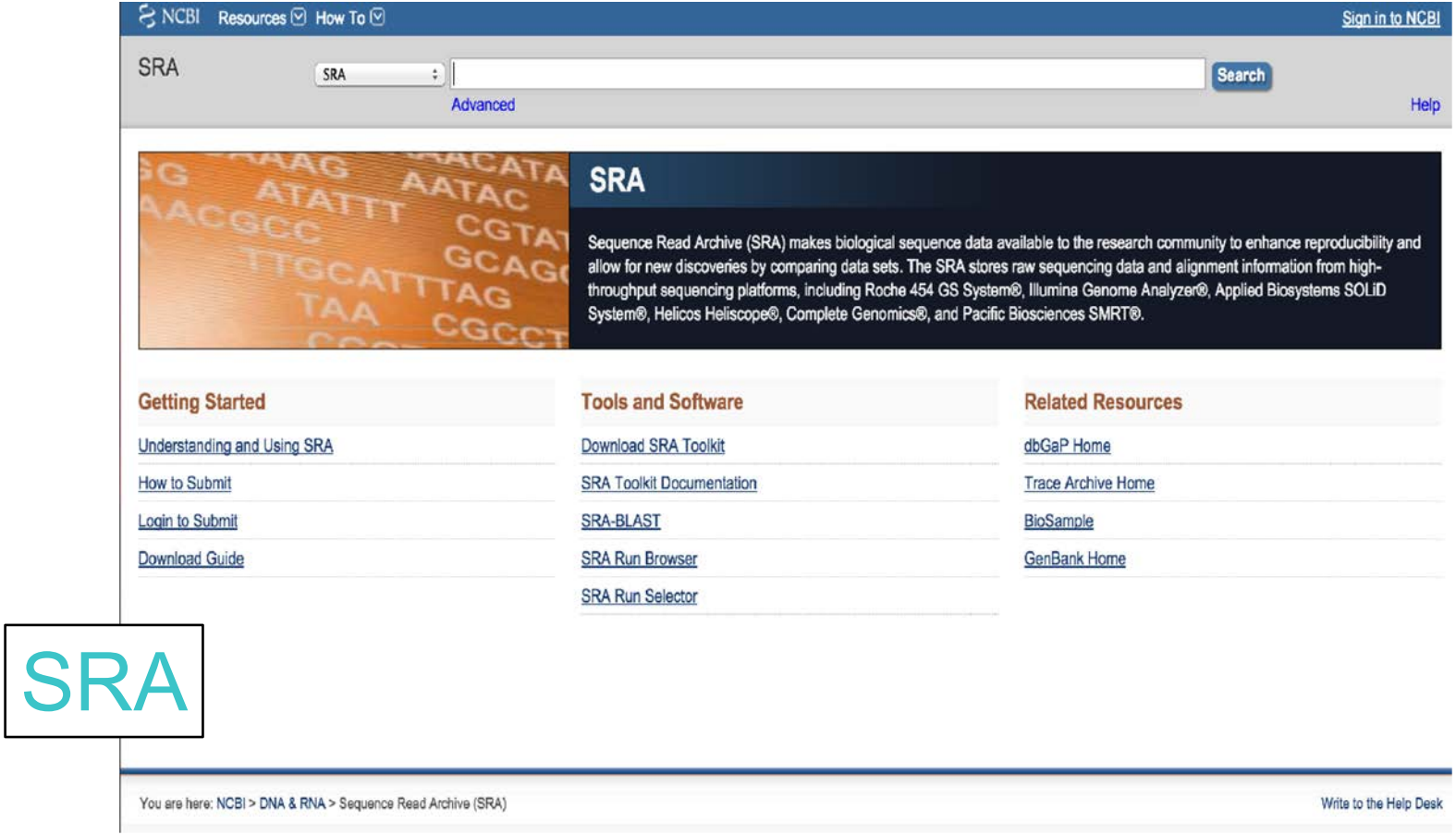


GTEx



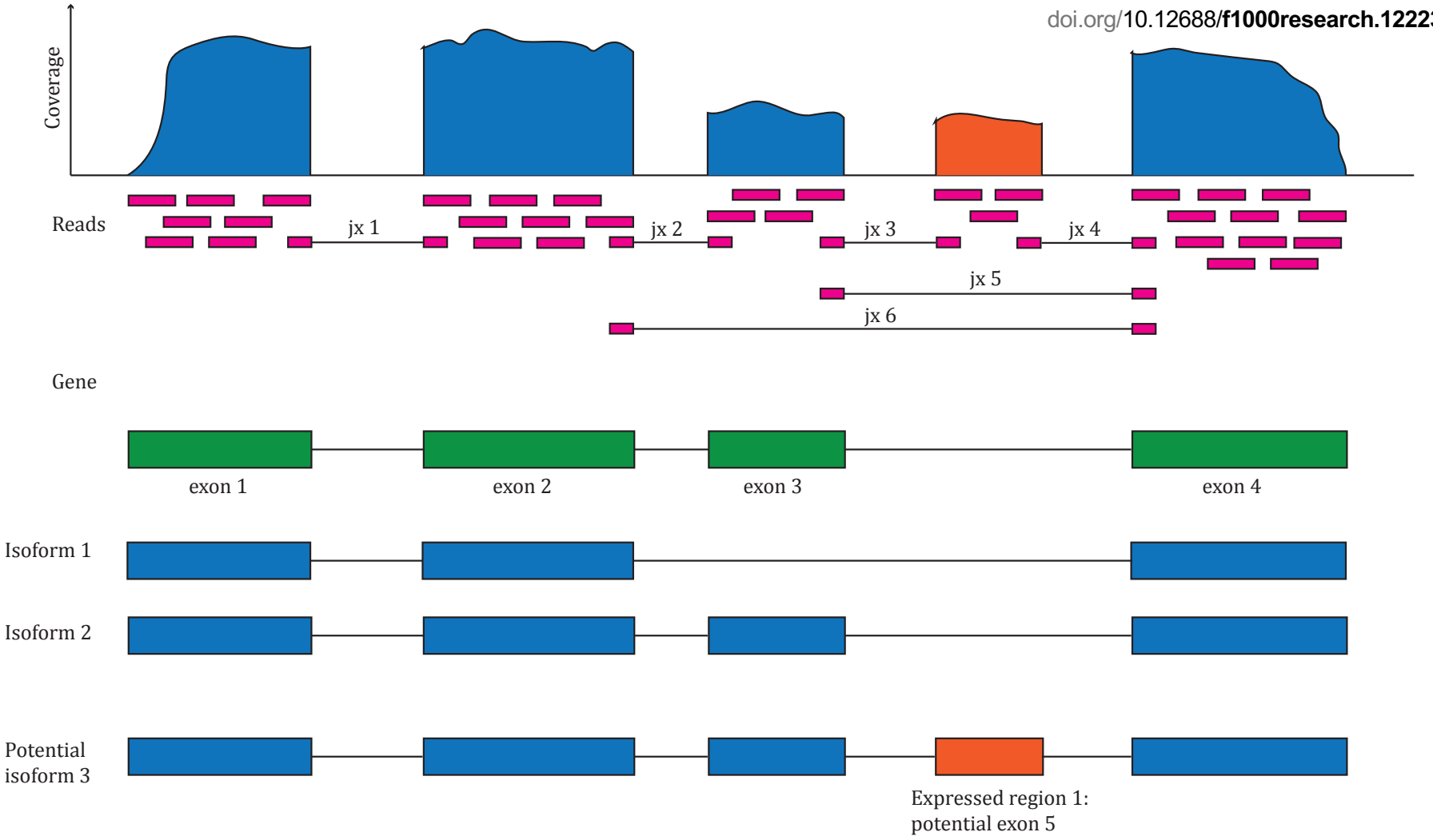
TCGA

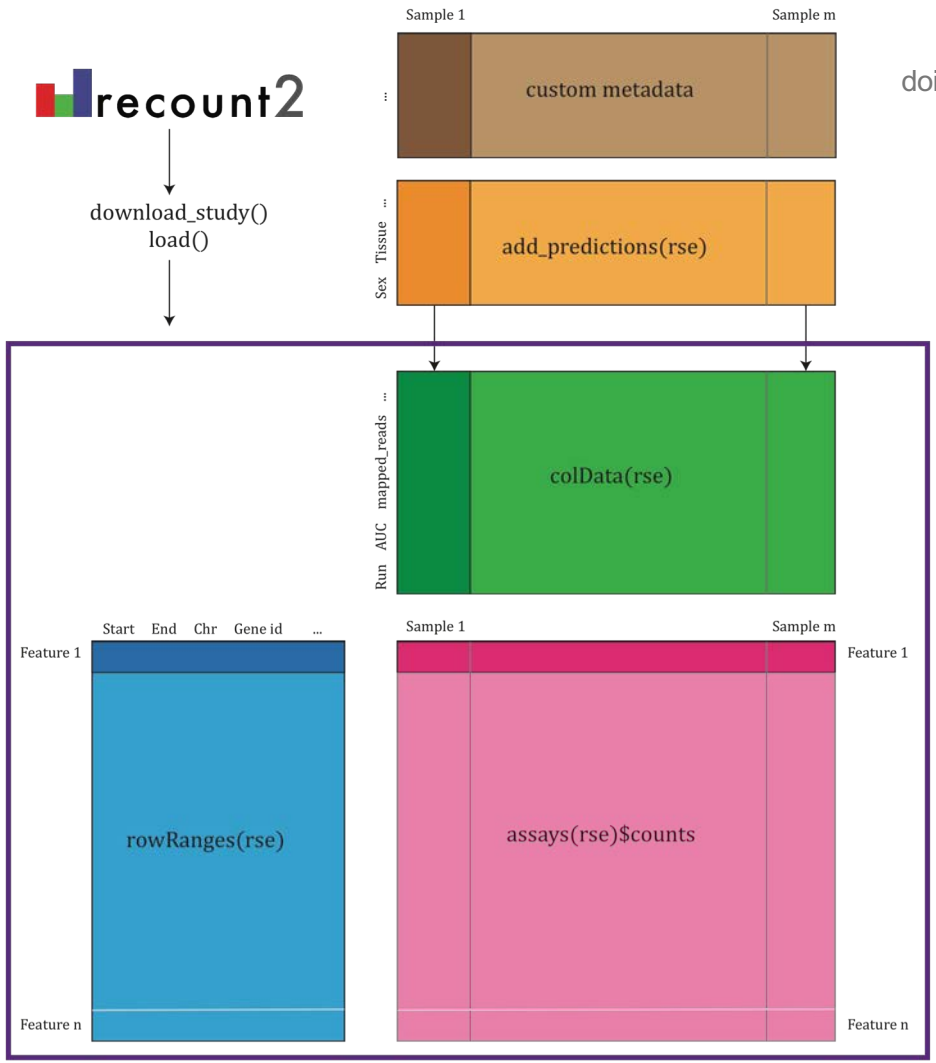
slide adapted from Shannon Ellis


NCBI Resources How To
Sign in to NCBI
SRA
SRA
Search
Advanced
Help
SRA
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®.
Getting Started
Understanding and Using SRA
How to Submit
Login to Submit
Download Guide
Tools and Software
Download SRA Toolkit
SRA Toolkit Documentation
SRA-BLAST
SRA Run Browser
SRA Run Selector
Related Resources
dbGaP Home
Trace Archive Home
BioSample
GenBank Home
You are here: NCBI > DNA & RNA > Sequence Read Archive (SRA)
Write to the Help Desk

SRA

doi.org/10.12688/**f1000research.12223.1**





doi.org/10.12688/**f1000research.12223.1**

```
> 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/

Mike Love
@mikelove

Following

Replying to @jtleek

Recount has been very useful for me over the years in developing and testing methods

RETWEETS	LIKES
4	5

10:17 AM - 11 Apr 2017

4 5

slide adapted from Jeff Leek

related projects

- Bioconductor recountWorkflow: doi.org/10.12688/**f1000research.12223.1**
- Shannon Ellis & Leek: phenotype prediction doi.org/10.1093/**nar/gky102**
- Jack Fu & Taub: transcript estimations doi.org/10.1101/**247346**

- Madugundu & Pandey (JHU):
  proteomics doi.org/10.1002/**pmic.201800315**
- Luidi-Imada & Marchionni (JHU):
  FANTOM (non-coding) and cancer doi.org/10.1101/**659490**
- Kuri-Magaña & Martínez-Barnetche (INSP Mexico):
  immune expression doi.org/10.3389/**fimmu.2018.02679**
- Ryten (UCL):
  Guelfi: validating expressed region (ER) eQTLs doi.org/10.1101/**591156**
  Zhang: improving the detection of ERs doi.org/10.1101/**499103**

Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples

Christopher Wilks ✉, Phani Gaddipati, Abhinav Nellore, Ben Langmead ✉

Bioinformatics, Volume 34, Issue 1, 01 January 2018, Pages 114–116,
https://doi.org/10.1093/bioinformatics/btx547
Published: 01 September 2017 **Article history** ▾

Christopher Wilks et al.
http://snaptron.cs.jhu.edu/snapcount_vignette.html
https://github.com/langmead-lab/snapr

Installation

```
# Install the development version from GitHub:
# install.packages("devtools")
devtools::install_github("langmead-lab/snapr")
```

Usage

snapr can be used with either a procedural interface

```
library(snapcount)

query_jx(compilation = "gtex", genes_or_intervals = "CD99")
#> class: RangedSummarizedExperiment
#> dim: 3485 9662
#> metadata(0):
#> assays(1): counts
#> rownames(3485): 28340058 28340273 ... 28352407 28352408
#> rowData names(12): DataSource:Type snaptron_id ... coverage_median
#>   source_dataset_id
#> colnames(9662): 50099 50100 ... 59759 59760
#> colData names(322): rail_id Run ... junction_coverage
#>   junction_avg_coverage
query_jx(compilation = "gtex", genes_or_intervals = "CD99", range_filters = exprs(samples_count == 10))
#> class: RangedSummarizedExperiment
#> dim: 25 9662
```

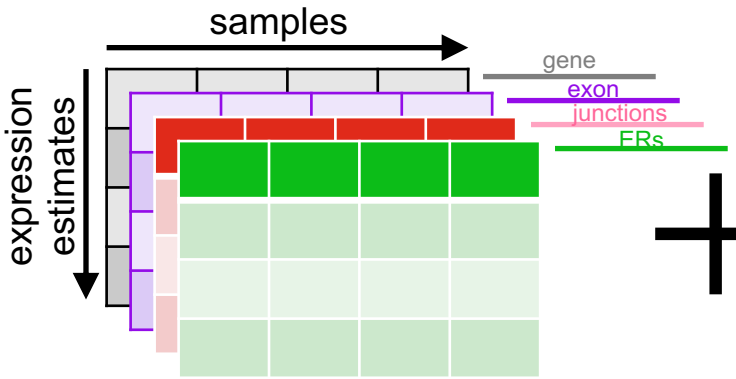
Christopher Wilks et al.

http://snaptron.cs.jhu.edu/snapcount_vignette.html

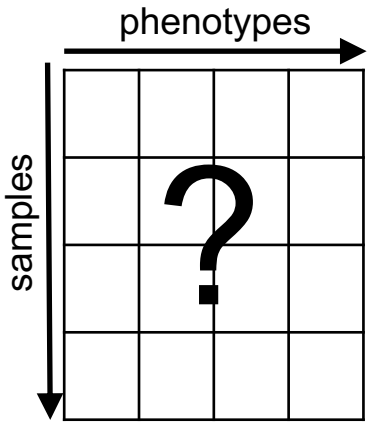
https://github.com/langmead-lab/snapr

recount2

expression data for ~70,000 human samples



GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284



Answer meaningful questions about human biology and expression

slide adapted from Shannon Ellis

Even when information *is* provided, it's not always clear…

sra_meta$Sex

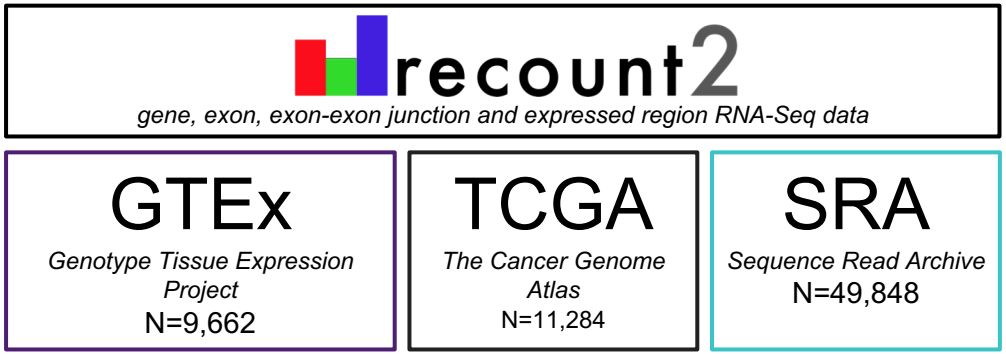
Category	Frequency
F	95
female	2036
Female	51
M	77
male	1240
Male	141
Total	**3640**

"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

Goal :

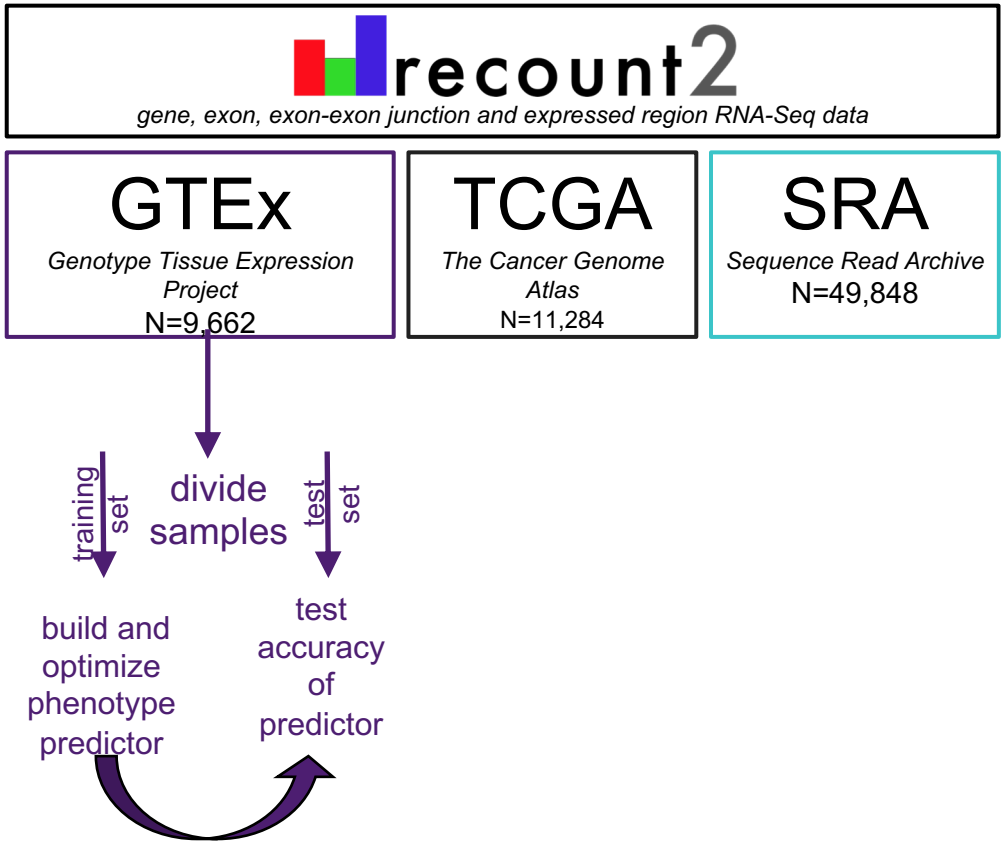
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

Goal :

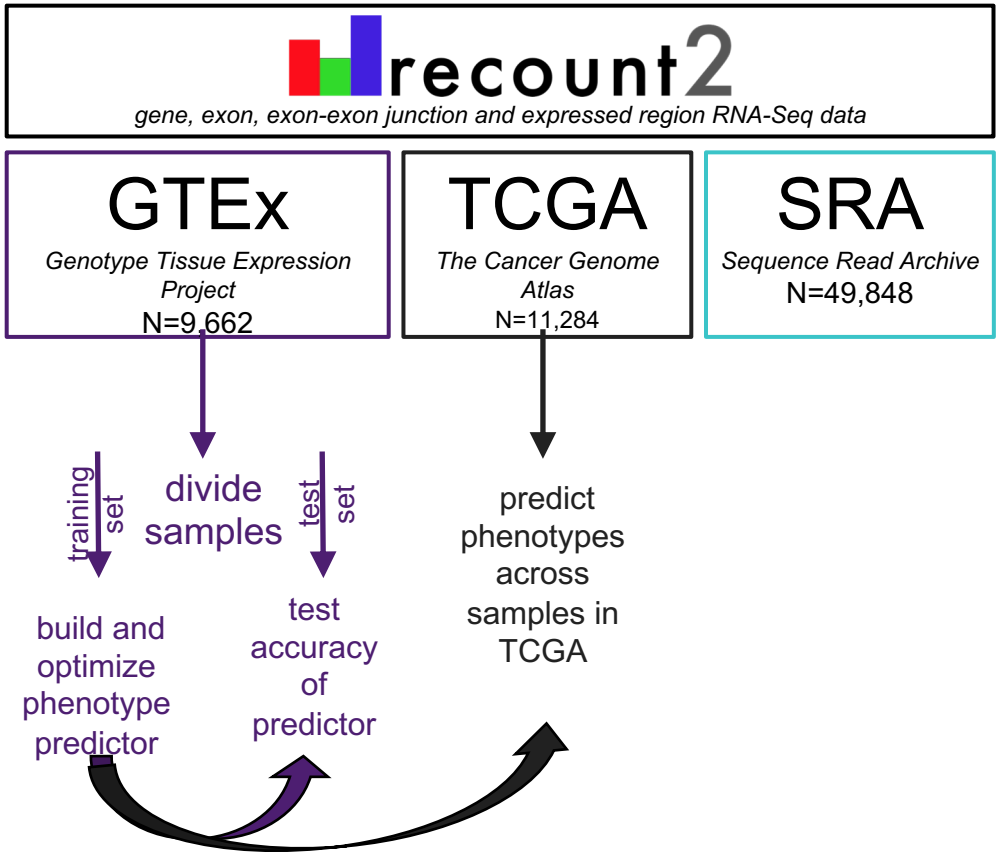
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

Goal :

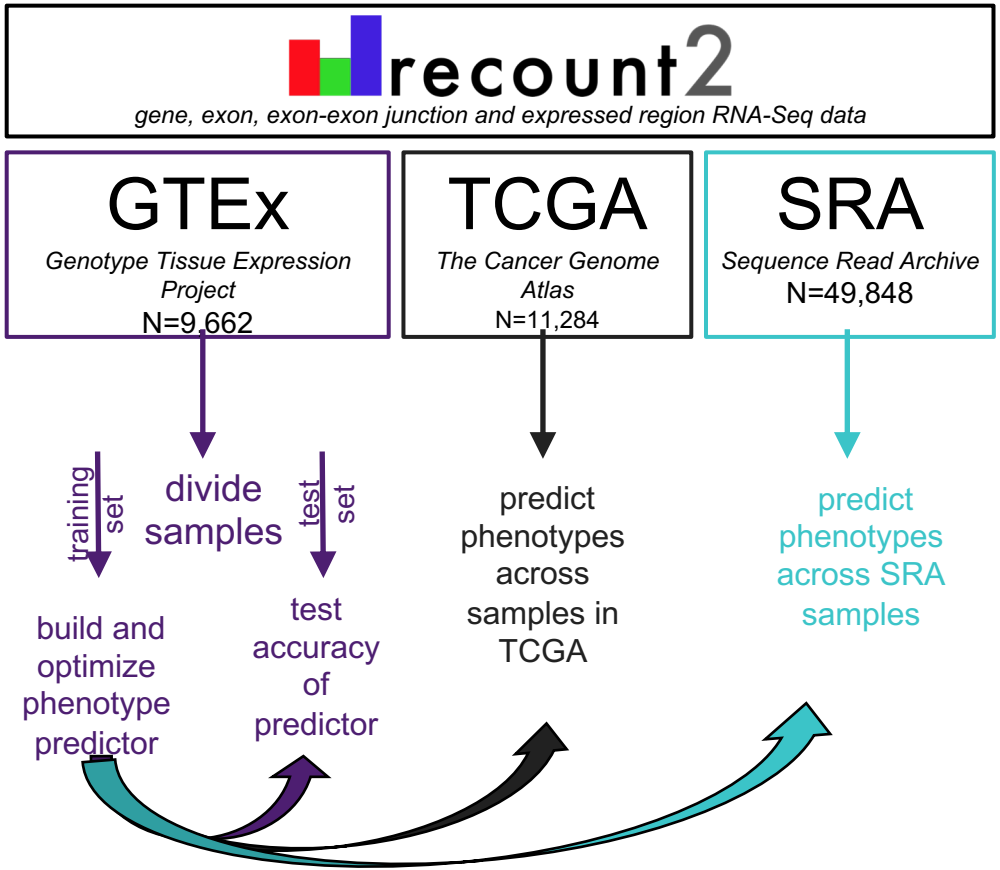
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

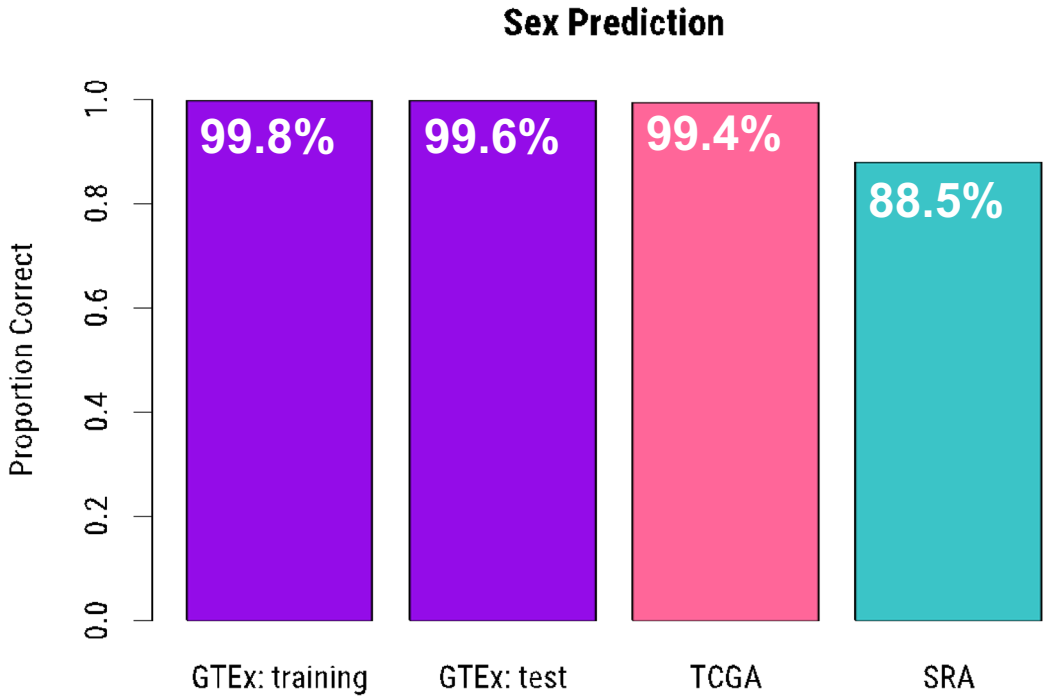
Goal :

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



slide adapted from Shannon Ellis

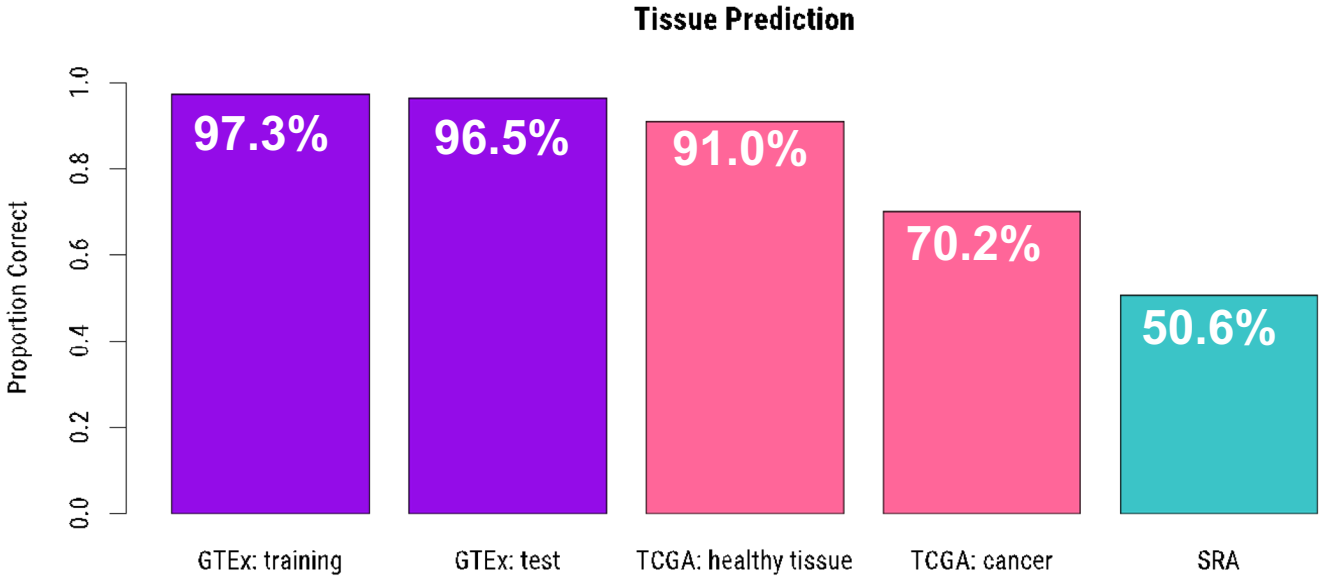
Sex prediction is accurate across data sets



Number of Regions	**20**	**20**	**20**	**20**
Number of Samples (N)	**4,769**	**4,769**	**11,245**	**3,640**

slide adapted from Shannon Ellis

Prediction is more accurate in healthy tissue

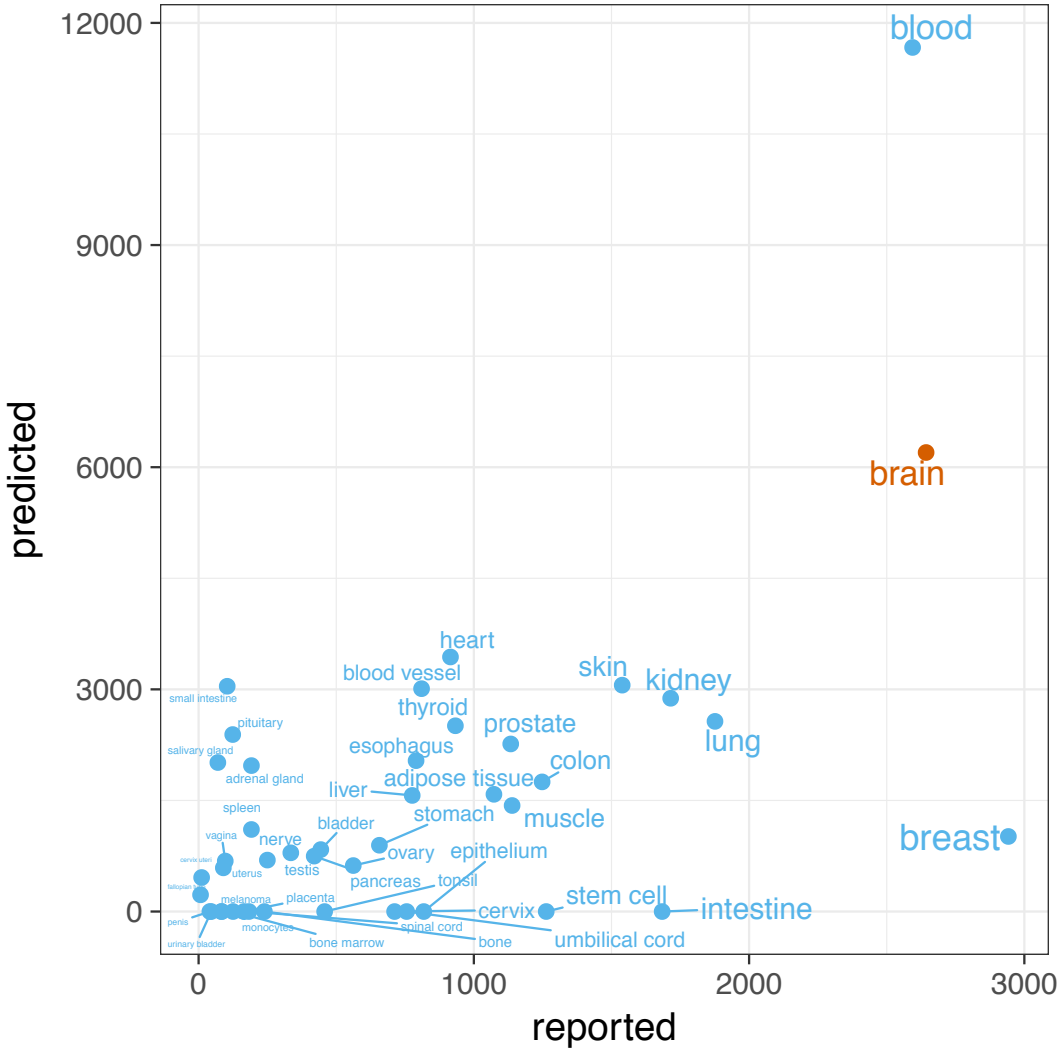


Number of Regions	589	589	589	589	589
Number of Samples (N)	4,769	4,769	613	6,579	8,951

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/


blood
brain
breast
heart
blood vessel
skin
kidney
small intestine
thyroid
prostate
lung
pituitary
esophagus
colon
salivary gland
adrenal gland
adipose tissue
liver
stomach
muscle
spleen
bladder
vagina
nerve
ovary
epithelium
testis
pancreas
tonsil
uterus
stem cell
intestine
melanoma
placenta
cervix
penis
monocytes
spinal cord
umbilical cord
urinary bladder
bone marrow
bone
12000
9000
6000
3000
0
0
1000
2000
3000
predicted
reported

- 62 SRA studies
- 4,431 rows by 48 columns

```
> with(recount_brain_v1, addmargins(table(present_in_recount, 'SRP025982' = sra_study_s == 'SRP025982')))
                   SRP025982
present_in_recount FALSE TRUE  Sum
             FALSE    39 1178 1217
             TRUE   1494 1720 3214
             Sum    1533 2898 4431
```

Ashkaun Razmara, et al doi.org/10.1101/**618025**

recount2 BRAIN

Sex	Female	Male		
Age/Development	Fetus	Child	Adolescent	Adult
Race/Ethnicity	Asian	Black	Hispanic	White
Tissue Site 1	Cerebral cortex	Hippocampus	Brainstem	Cerebellum
Tissue Site 2	Frontal lobe	Temporal lobe	Midbrain	Basal ganglia
Tissue Site 3	Dorsolateral prefrontal cortex	Superior temporal gyrus	Substantia nigra	Caudate
Hemisphere	Left	Right		
Brodmann Area	1-52			
Disease Status	Disease	Neurological control		
Disease	Brain tumor	Alzheimer's disease	Parkinson's disease	Bipolar disorder
Tumor Type	Glioblastoma	Astrocytoma	Oligodendroglioma	Ependymoma
Clinical Stage 1	Grade I	Grade II	Grade III	Grade IV
Clinical Stage 2	Primary	Secondary	Recurrent	
Viability	Postmortem	Biopsy		
Preparation	Frozen	Thawed		

Ashkaun Razmara, et al doi.org/10.1101/**618025**

LIBD

Reproducibility document

recount2 BRAIN

https://github.com/**LieberInstitute/recount-brain**/tree/master/**metadata_reproducibility**

- Overall curation steps: starts by downloading SRA Run Table info, then info from the publications
- Details for each SRA study

6, SRP019762

https://doi.org/10.1038/ncomms4584

- Methods:
  - Samples used for metabolite and RNA-seq experiments: "from frozen postmortem tissue" / "corresponding to Brodmann area 10"
    - Viability: Postmortem
    - Preparation: Frozen
    - Brodmann Area: 10
- Supplementary Information
  - Supplementary Table 11. Sample information for the prefrontal cortex samples used in metabolite and RNA-seq measurements:
    - Sex:
    - Age:
    - RIN:
    - PMI:
    - Brain Bank:

Ashkaun Razmara, et al doi.org/10.1101/**618025**

recount2 BRAIN

Download recount_brain

```
library('recount')
recount_brain = add_metadata()
```

Find project(s) of interest

```
interest = subset(recount_brain)
project = interest$sra_study_s
```

Download expression data

```
download_study(project)
load('rse_gene.Rdata')
```

recount2

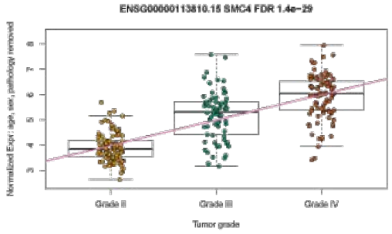
Adds sample metadata

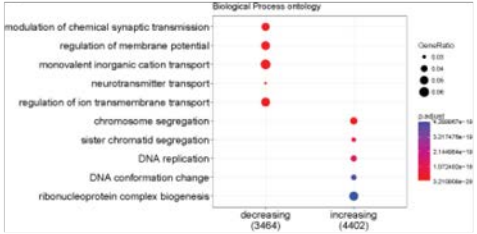
```
add_metadata(rse_gene)
```

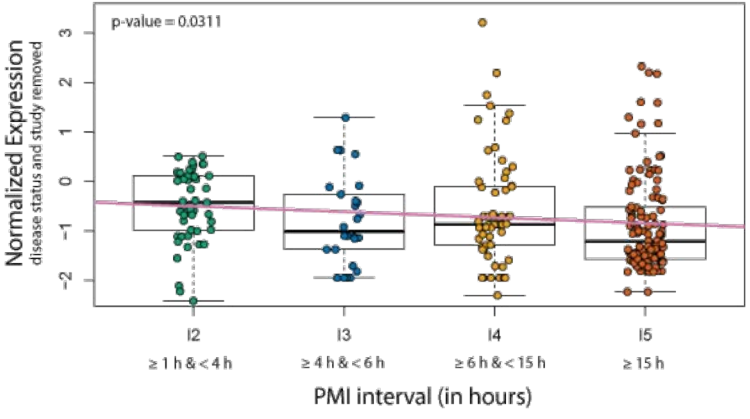
recount2 BRAIN

Perform analyses

```
....
```

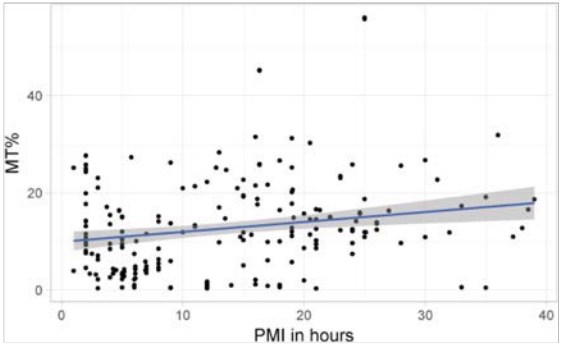




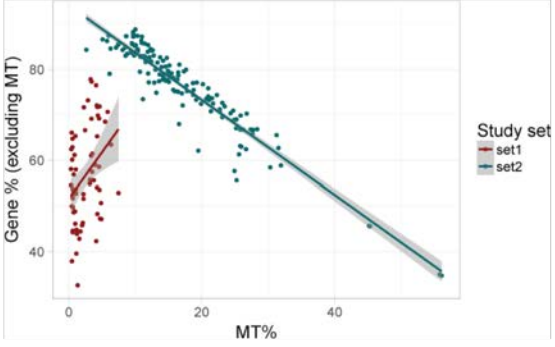


Replicates part of the GTEx PMI paper by Ferreira et al. doi.org/10.1038/s41467-017-02772-x

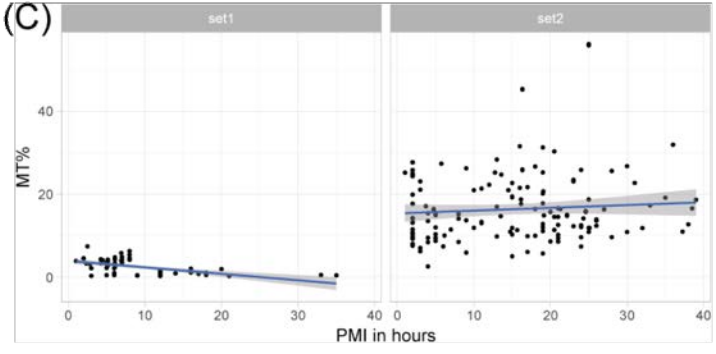
(A)



(B)



(C)



Code Example:

research.libd.org/recount-brain/example_PMI/example_PMI.html

research.libd.org/recount-brain/example_PMI/example_PMI.Rmd

Replicates part of the GTEx PMI paper by Ferreira et al.
doi.org/10.1038/s41467-017-02772-x

LIBD

recount_brain

recount2 BRAIN

* Jeff Leek presented Shannon Ellis' prediction work in Toronto (around April 2018)
https://docs.google.com/presentation/d/1FgUZZU6pW91J7zH0OqrEgxfnV1Py_ZGL3ZKHfbOZskY/edit#slide=id.g2f831fd4ae_0_306

* Dustin J. Sokolowski from Michael D. Wilson's lab is using *recount2*

* Dustin joins the project and merges *recount-brain* with GTEx and TCGA

* Met Sean Davis (NIH) at #biodata18, helped us with mapping to ontologies

The SRA samples in *recount-brain* are complemented by 1,409 GTEx (GTEx Consortium 2015) and 707 TCGA (Brennan et al. 2013; Cancer Genome Atlas Research Network et al. 2015) samples covering 13 healthy regions of the brain and 2 tumor types, respectively. In total, there are 6,547 samples with metadata in *recount-brain* with 5,330 (81.4%) present in *recount2*

Ashkaun Razmara, et al doi.org/10.1101/**618025**

LIBD

github.com/LieberInstitute/recount-brain

The *recount-brain* team

Hopkins
Ashkaun Razmara
Shannon E. Ellis
Jeff T. Leek

University of Toronto
Dustin J. Sokolowski
Michael D. Wilson

NIH
Sean Davis

LIBD
Andrew E. Jaffe

Funding

NIH R01 GM105705
NIH 1R21MH109956
NIH R01 GM121459
CIHR, NSERC
Ontario Ministry of Research

Hosting *recount2*

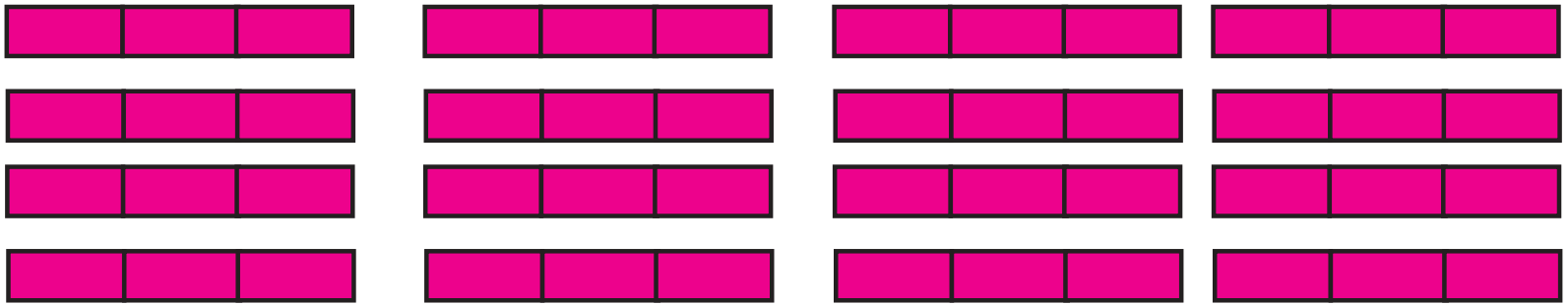
IDIES SciServer

LIEBER INSTITUTE *for* BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES

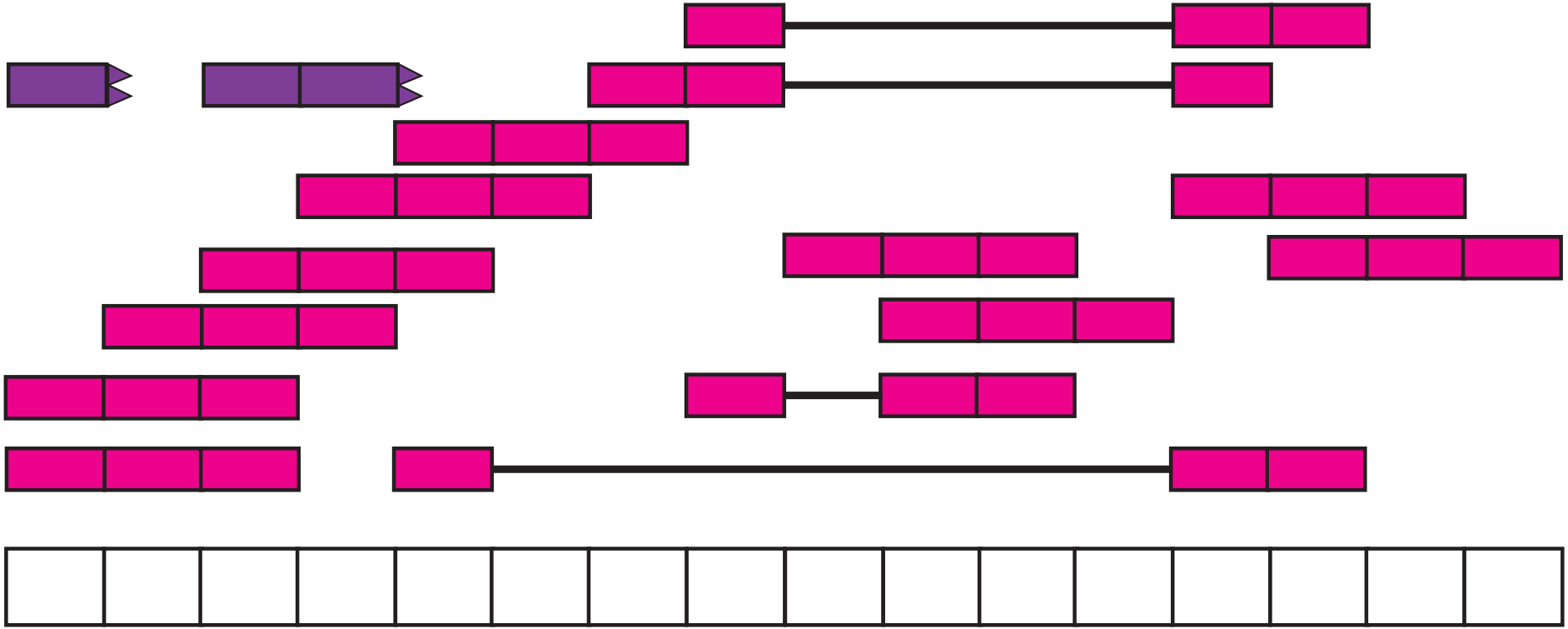
```
> library('recount')

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

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

> rse <- scale_counts(rse_gene)
```

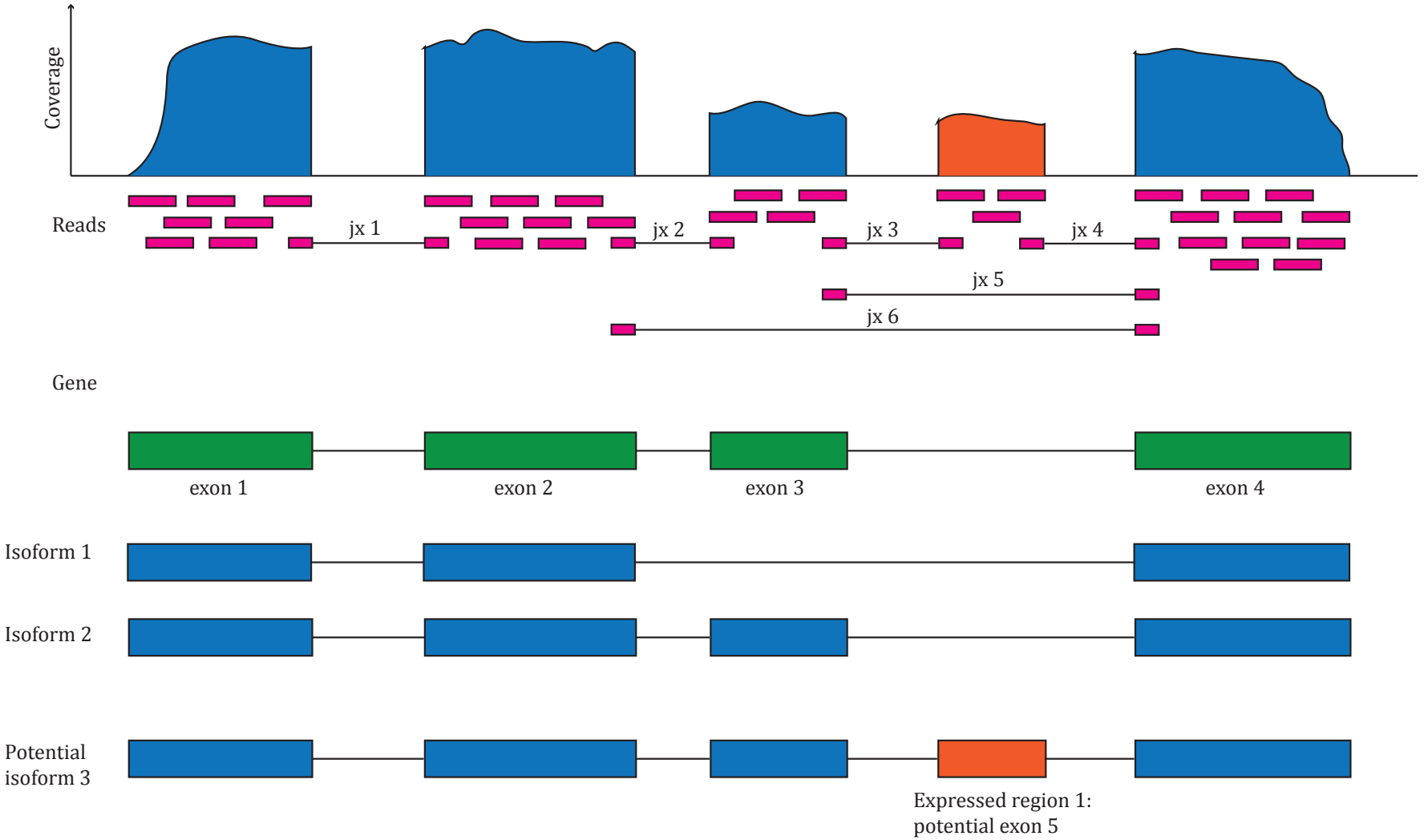
https://github.com/leekgroup/recount-analyses/

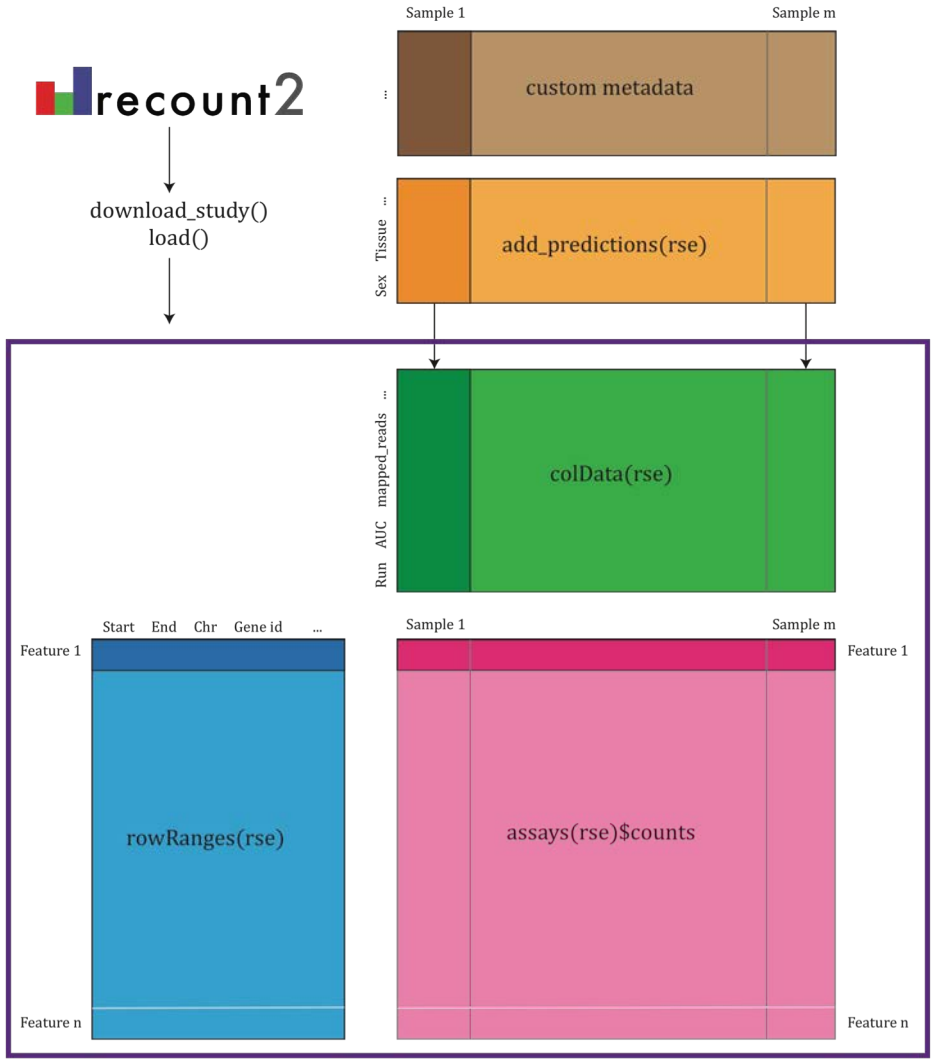
Reads

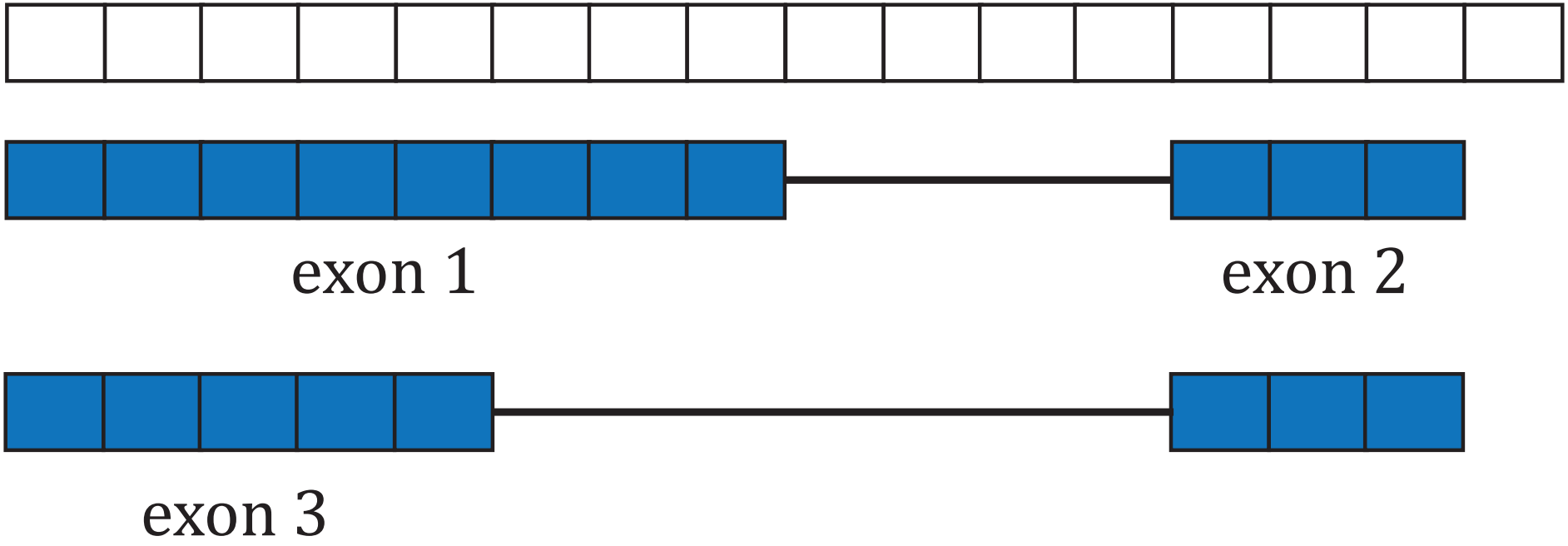


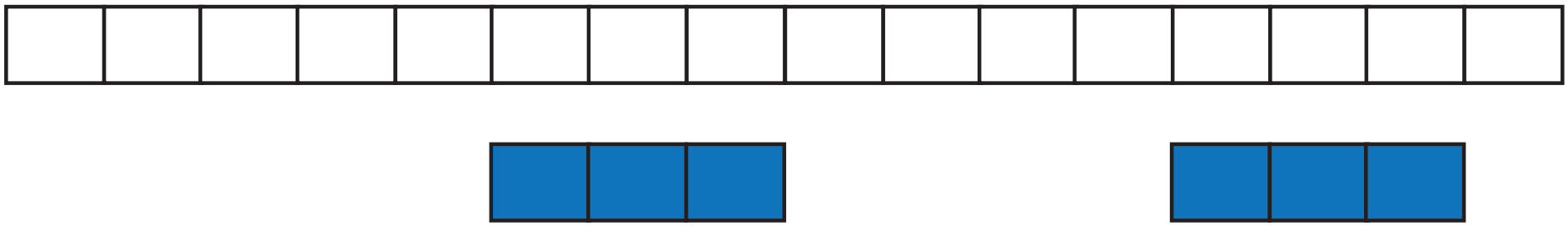
Reference genome

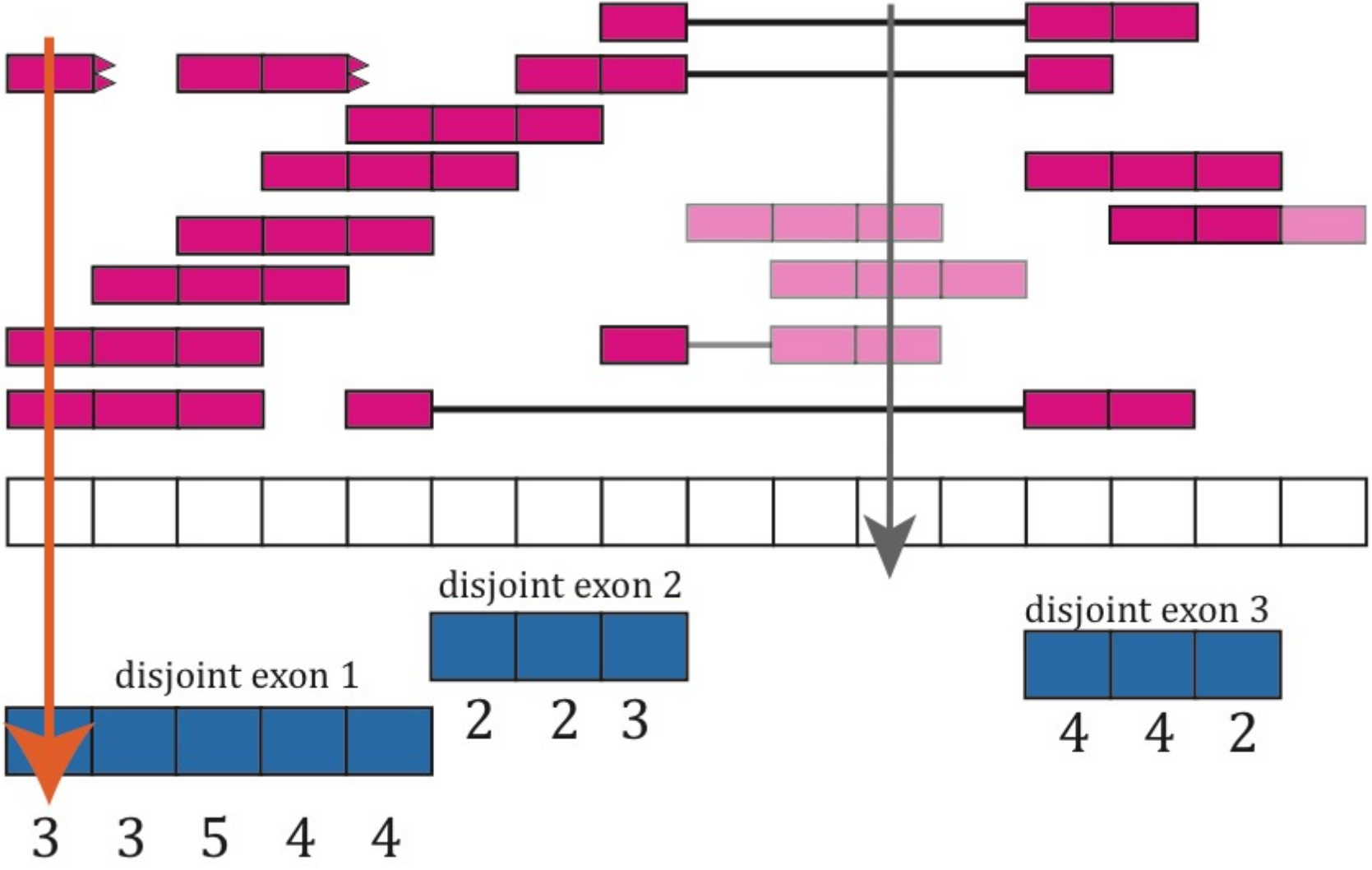


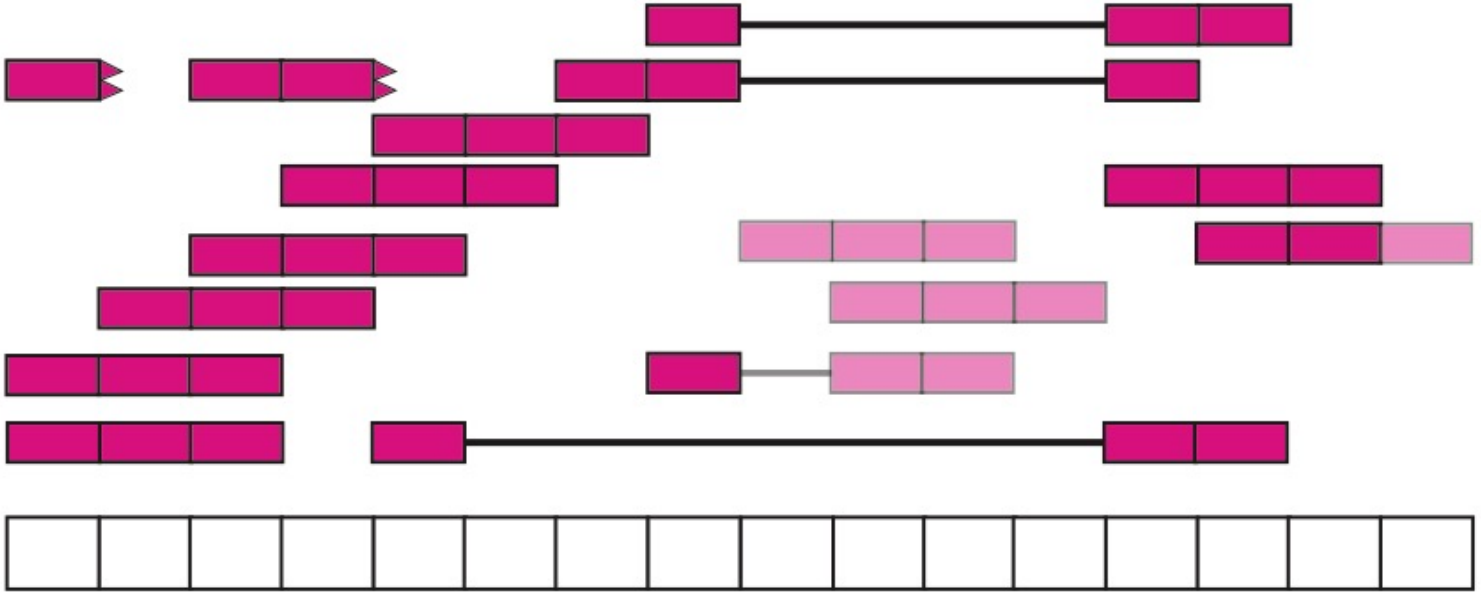

Coverage
Reads
jx 1
jx 2
jx 3
jx 4
jx 5
jx 6
Gene
exon 1
exon 2
exon 3
exon 4
Isoform 1
Isoform 2
Potential isoform 3
Expressed region 1: potential exon 5


recount2
download_study()
load()
Sample 1
Sample m
custom metadata
add_predictions(rse)
Sex
Tissue
...
colData(rse)
Run
AUC
mapped_reads
...
Start
End
Chr
Gene id
...
Feature 1
Feature n
rowRanges(rse)
Sample 1
Sample m
Feature 1
Feature n
assays(rse)$counts


exon 1
exon 2
exon 3


disjoint exon 2
disjoint exon 3
disjoint exon 1


disjoint exon 2
disjoint exon 3
disjoint exon 1
2 2 3
4 4 2
3 3 5 4 4



disjoint exon 2

disjoint exon 3

disjoint exon 1

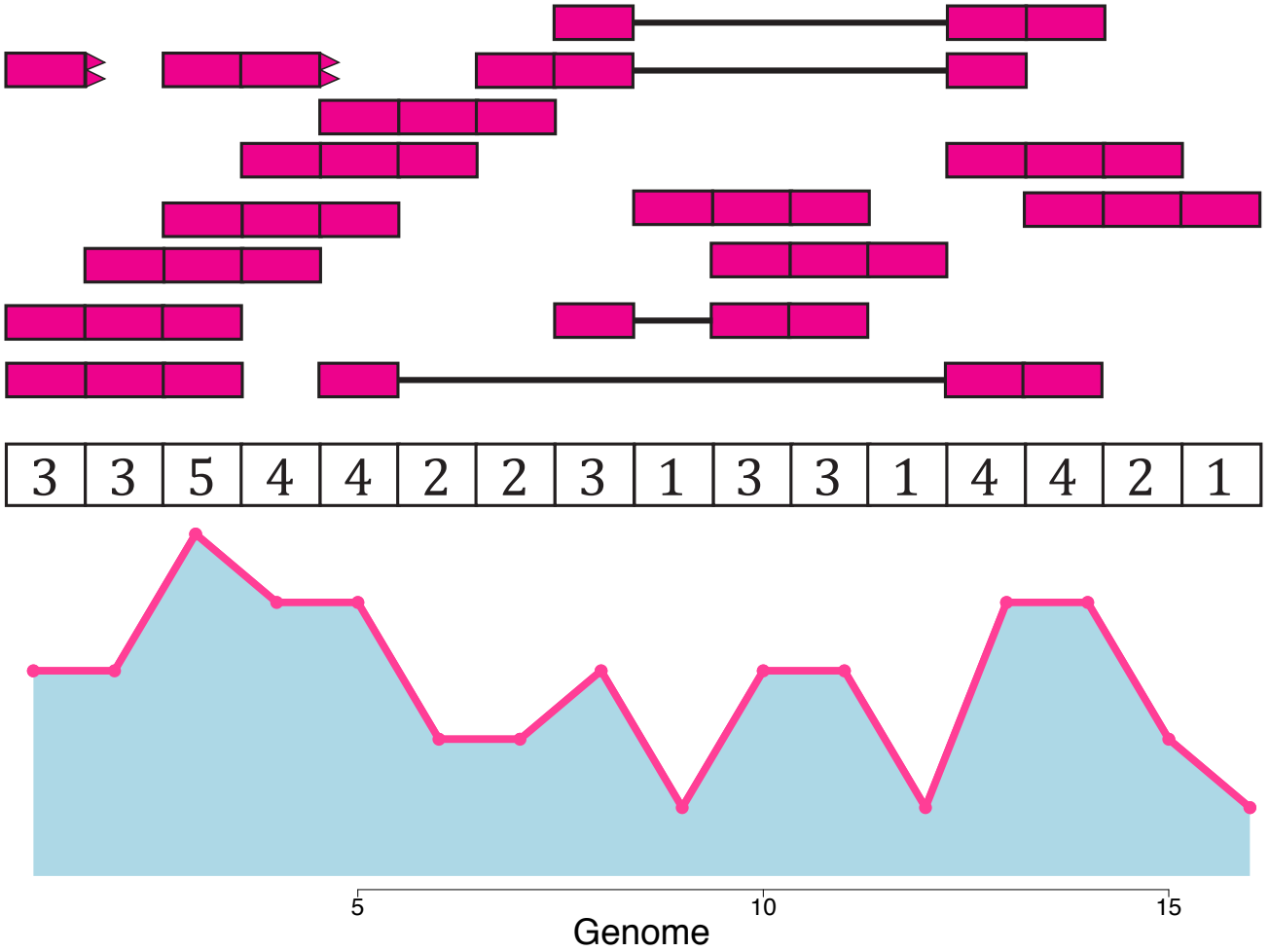
2+2+3=7

4+4+2=10

3+3+5+4+4 = 19

Gene = 19 + 7 + 10 = 36

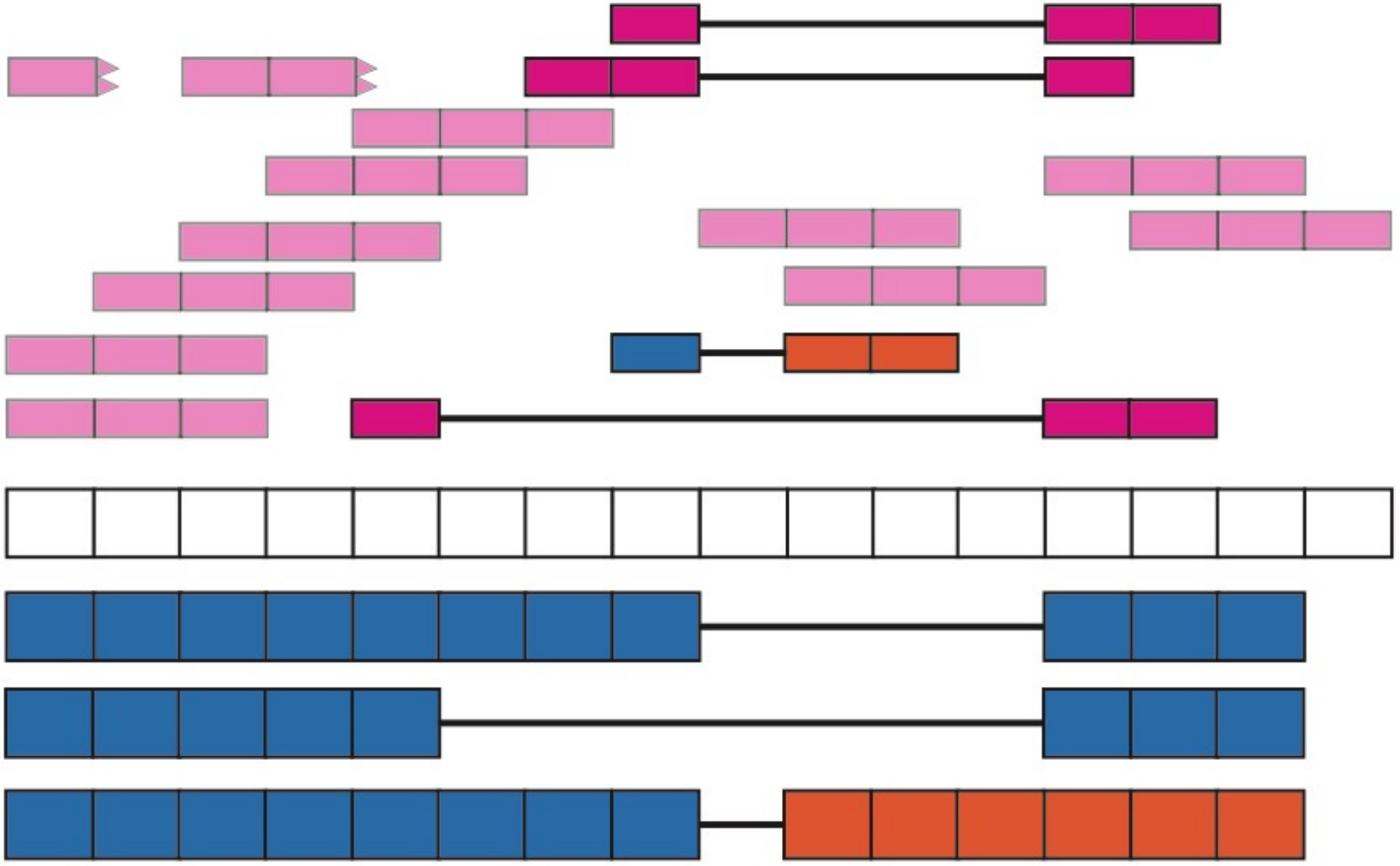
$$\frac{\sum_{i}^{n} \text{coverage}_{i}}{\text{Read Length}} * \frac{\text{target}}{\text{mapped}} = \text{scaled read counts}$$

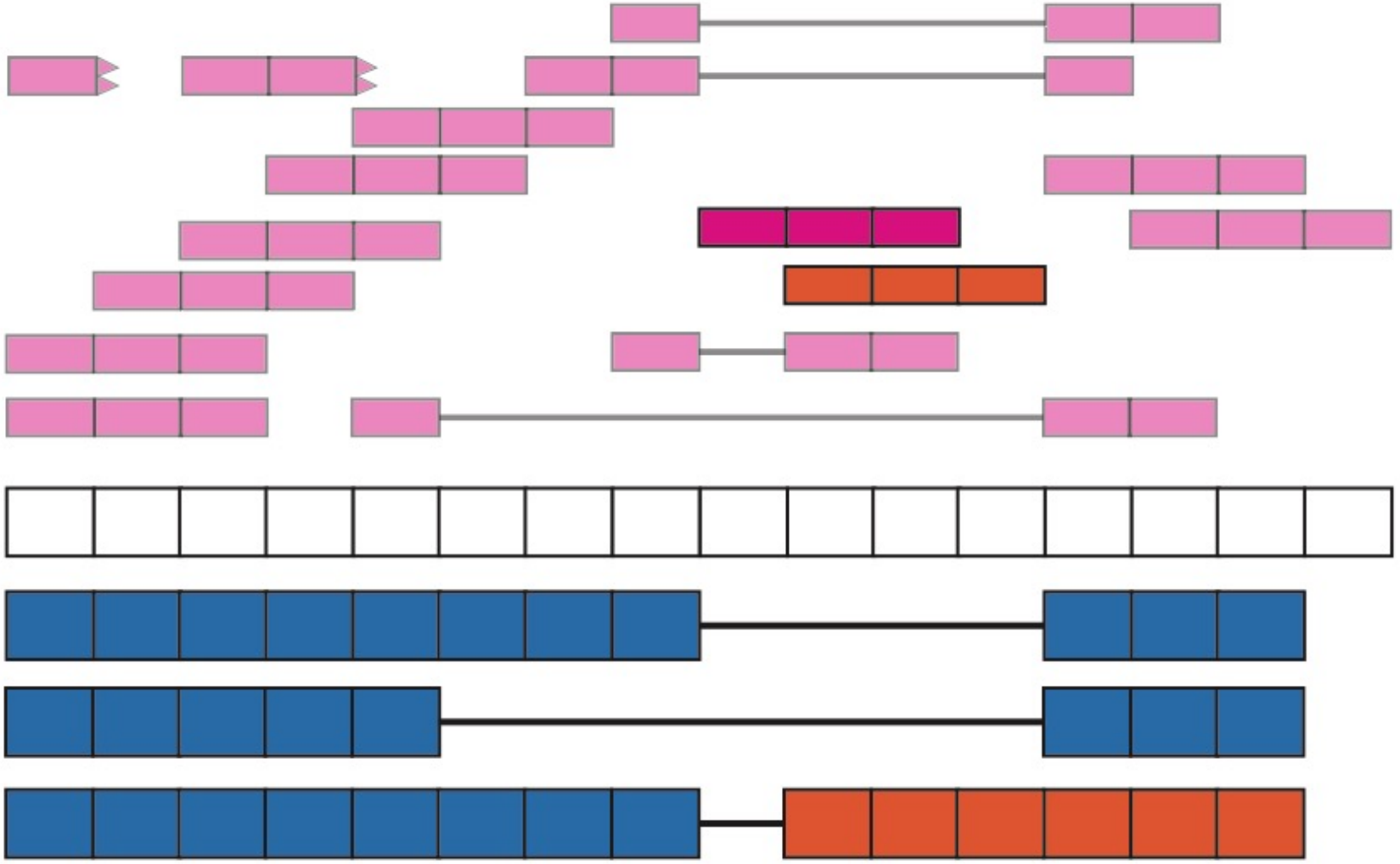


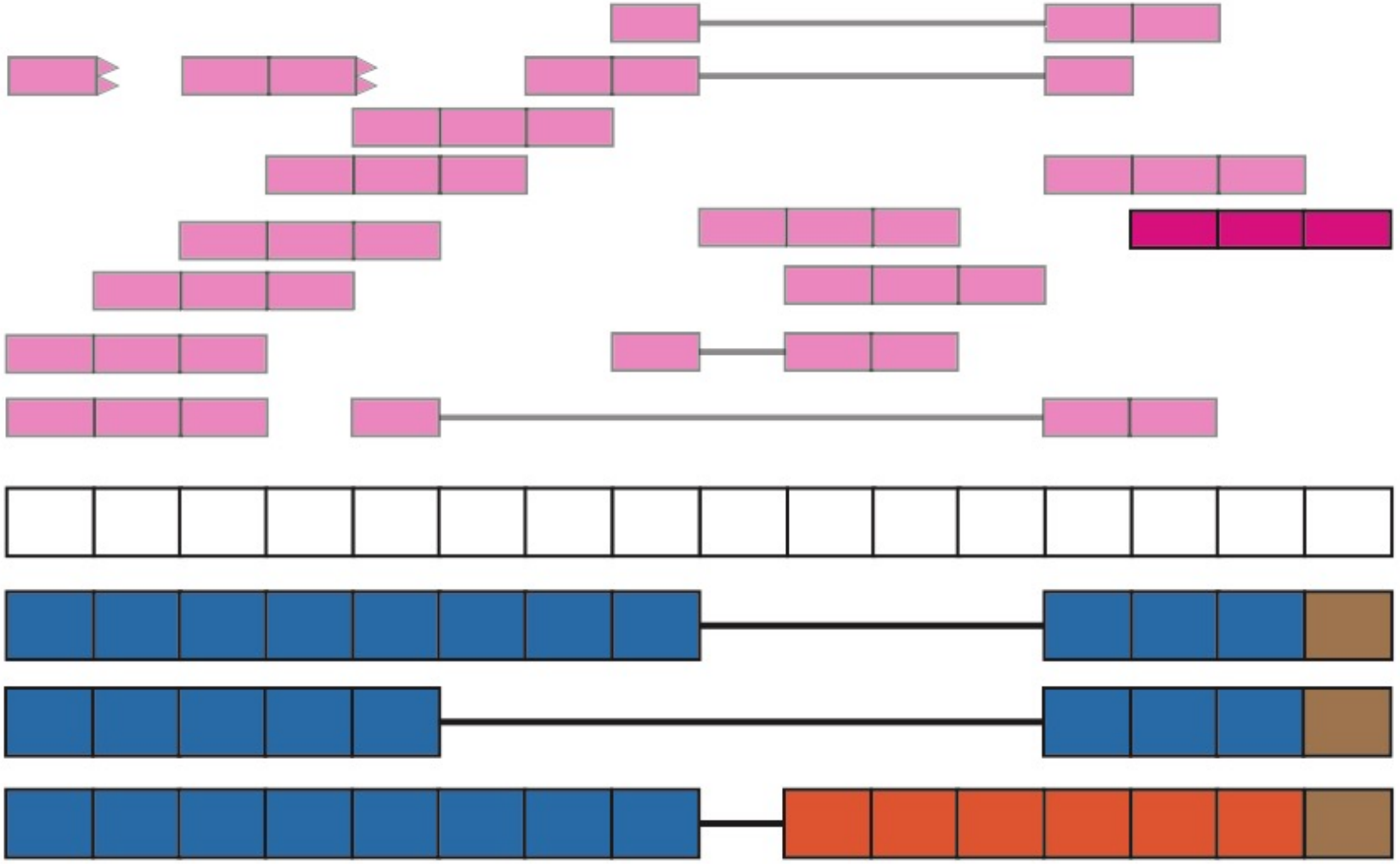
AUC = area under coverage = 45

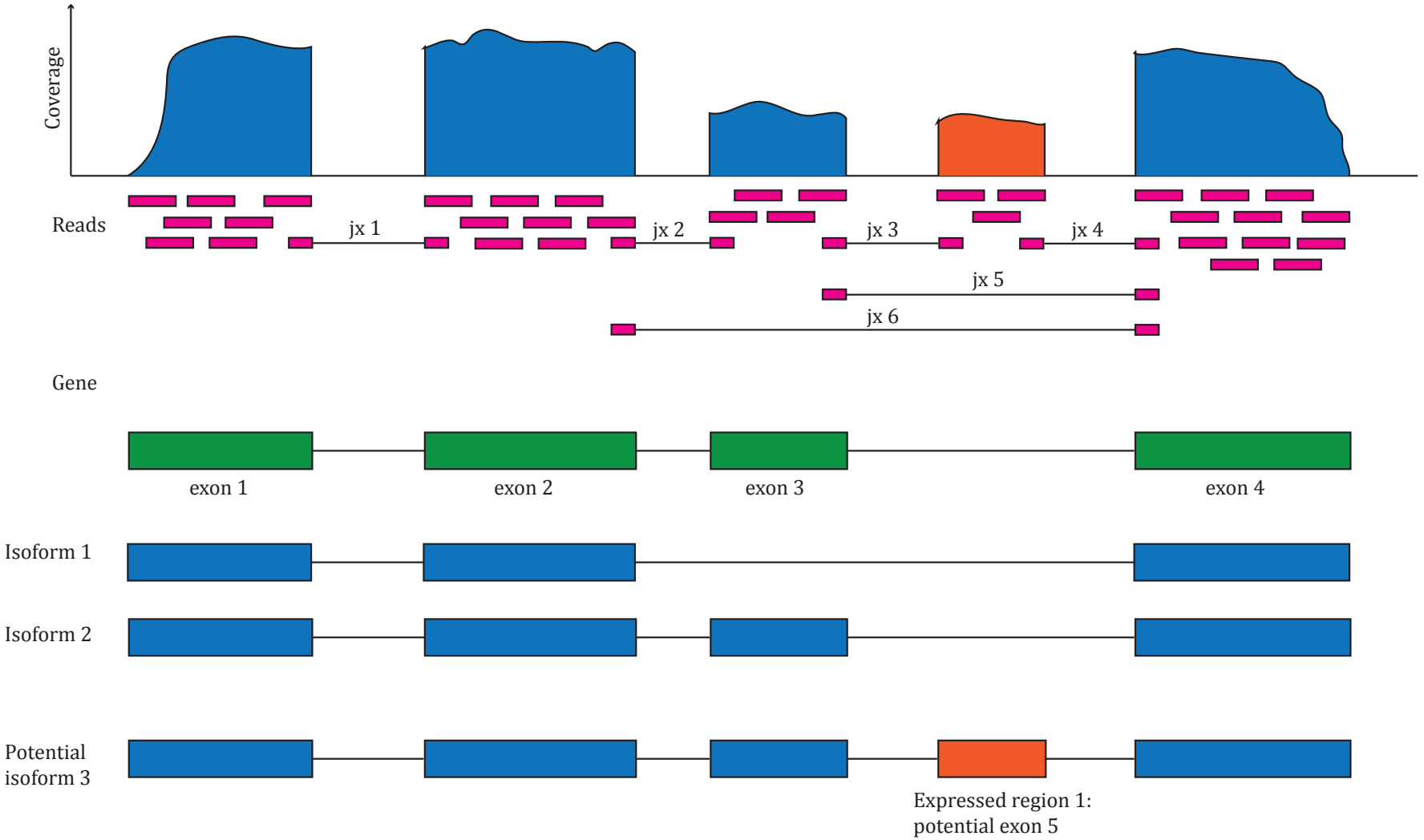
$$\frac{\sum_i^n \text{coverage}_i}{\text{Read Length}} * \frac{\text{target}}{\text{mapped}} = \text{scaled read counts}$$

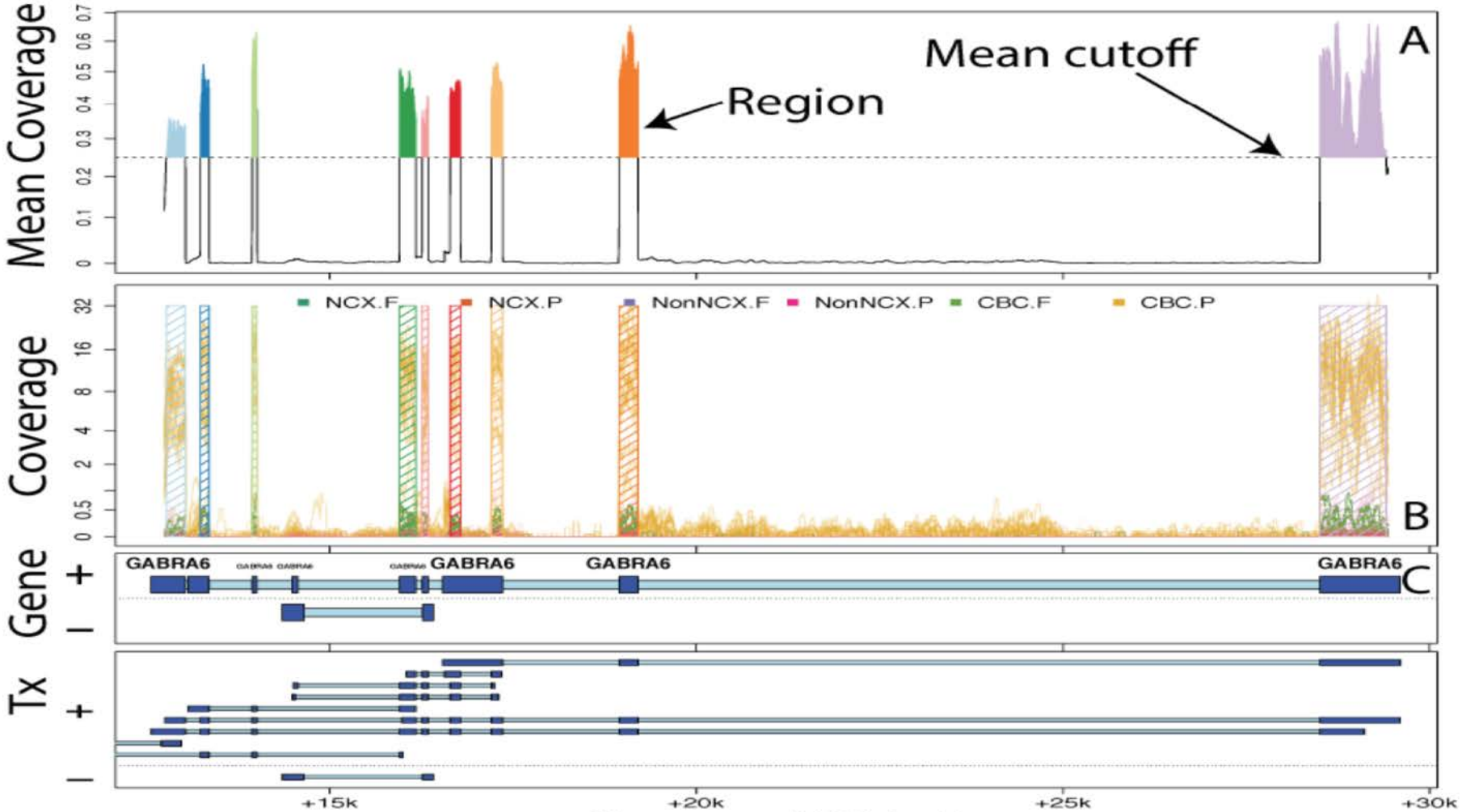
$$\frac{\sum_i^n \text{coverage}_i}{\text{AUC}} * \text{target} = \text{scaled read counts}$$





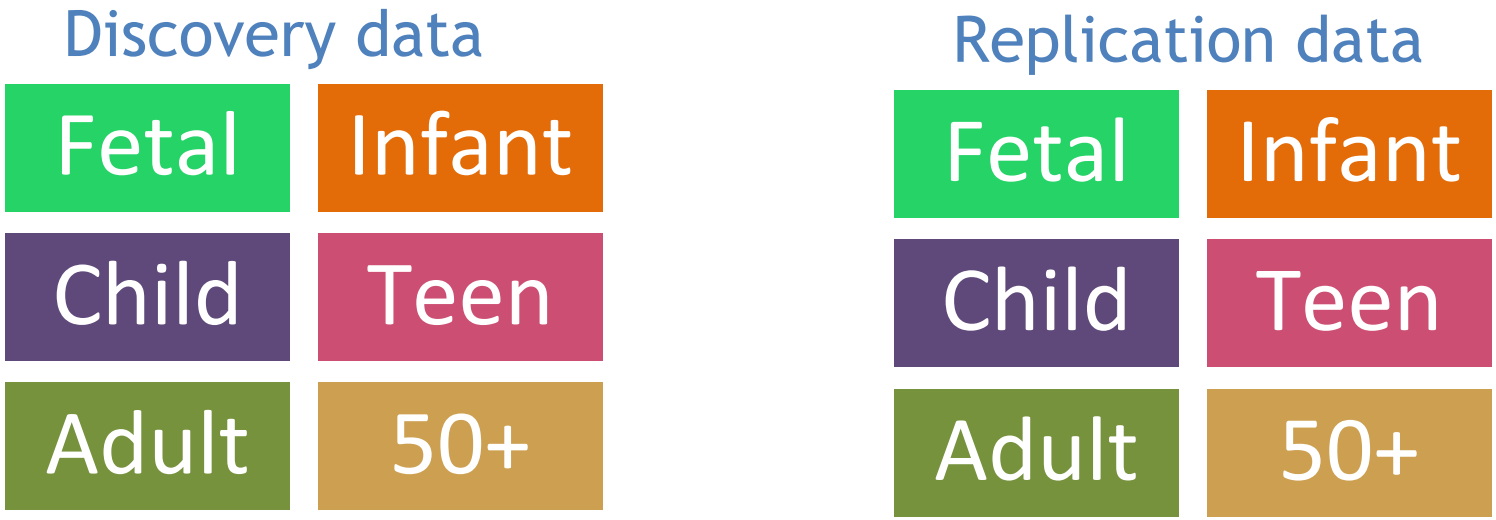



Coverage
Reads
jx 1
jx 2
jx 3
jx 4
jx 5
jx 6
Gene
exon 1
exon 2
exon 3
exon 4
Isoform 1
Isoform 2
Potential isoform 3
Expressed region 1: potential exon 5

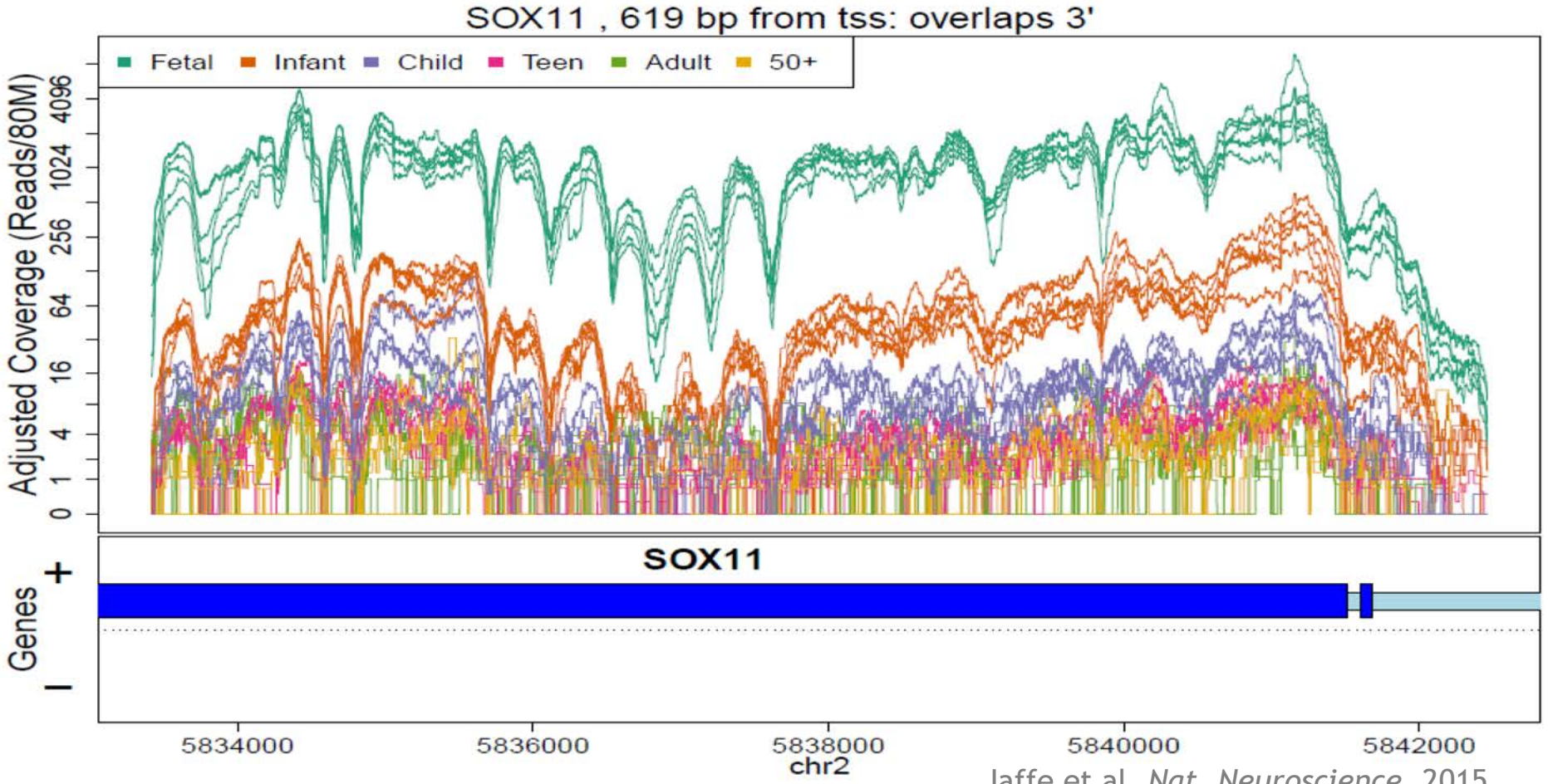


Collado-Torres et al, *NAR*, 2017

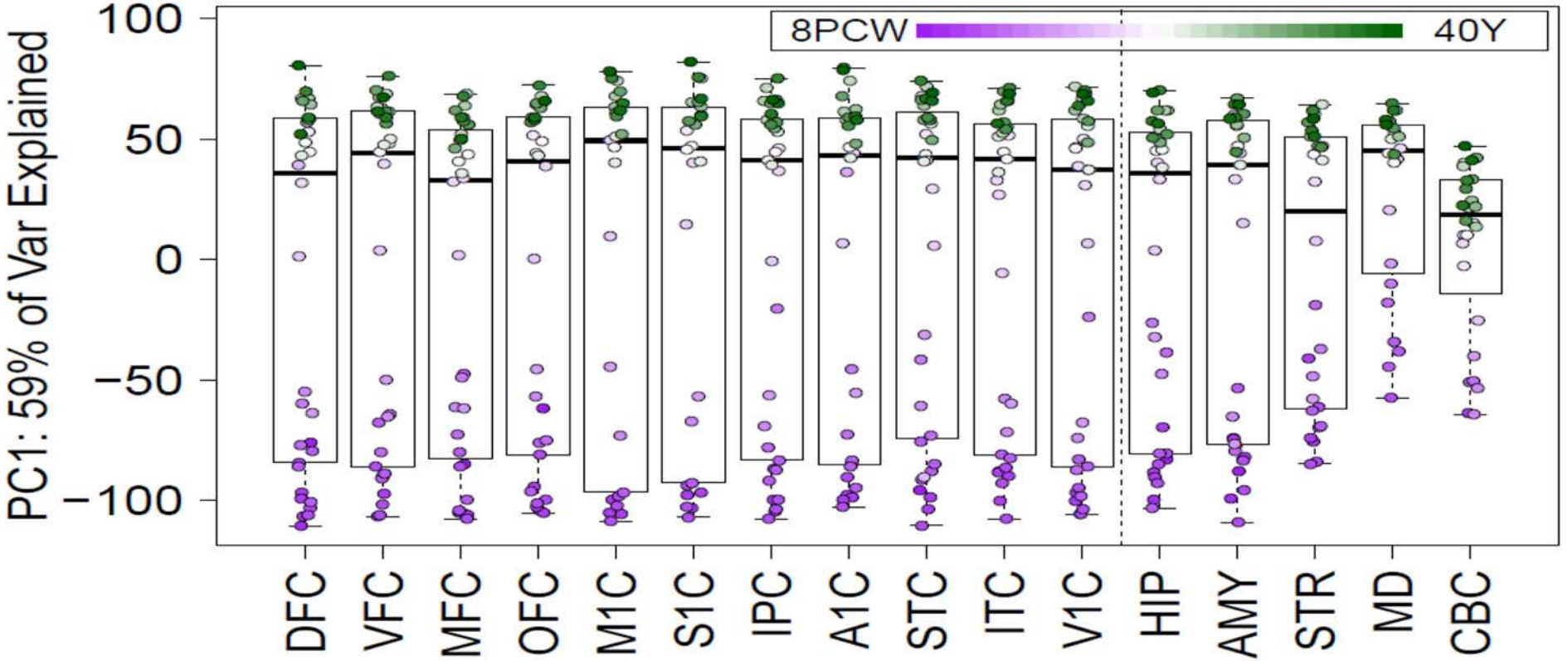
Postmortem Human Brain Samples



Jaffe et al, *Nat. Neuroscience*, 2015



Jaffe et al, *Nat. Neuroscience*, 2015



BrainSpan data
Jaffe et al, *Nat. Neuroscience*, 2015

LIBD

Collaborators

UCSD
Shannon Ellis

Hopkins
Jeff Leek
Ben Langmead
Christopher Wilks
Kai Kammers
Kasper Hansen
Margaret Taub

OHSU
Abhinav Nellore

LIBD
Andrew Jaffe

Funding

NIH R01 GM105705
NIH 1R21MH109956
CONACyT 351535
AWS in Education
Seven Bridges
IDIES SciServer

LIEBER INSTITUTE *for*
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES

expression data for ~70,000 human samples

(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)

help(package = recountWorkshop2019)

vignette('recount-workshop', 'recountWorkshop2019')

https://rebrand.ly/biocworkshops2019

Leonardo Collado-Torres
@fellgernon
#CONABIO2019

LIEBER INSTITUTE *for*
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES