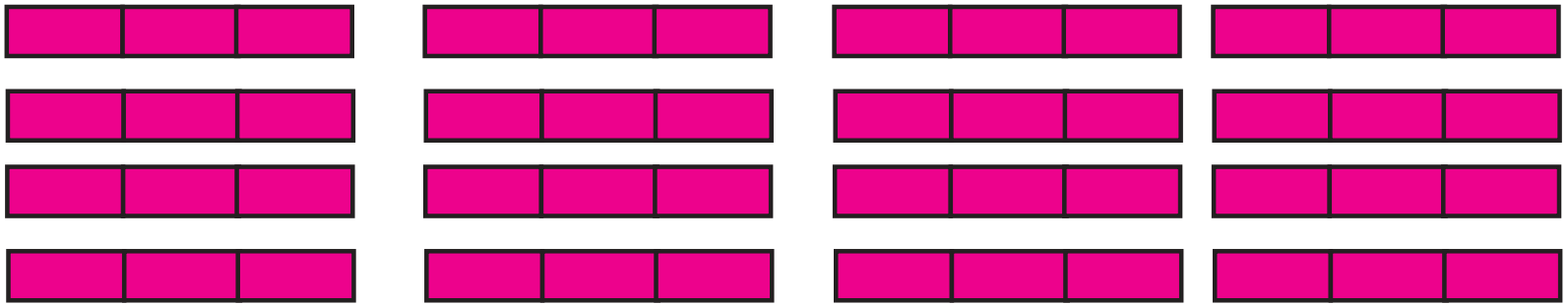
LIBD

Reproducible RNA-seq analysis with

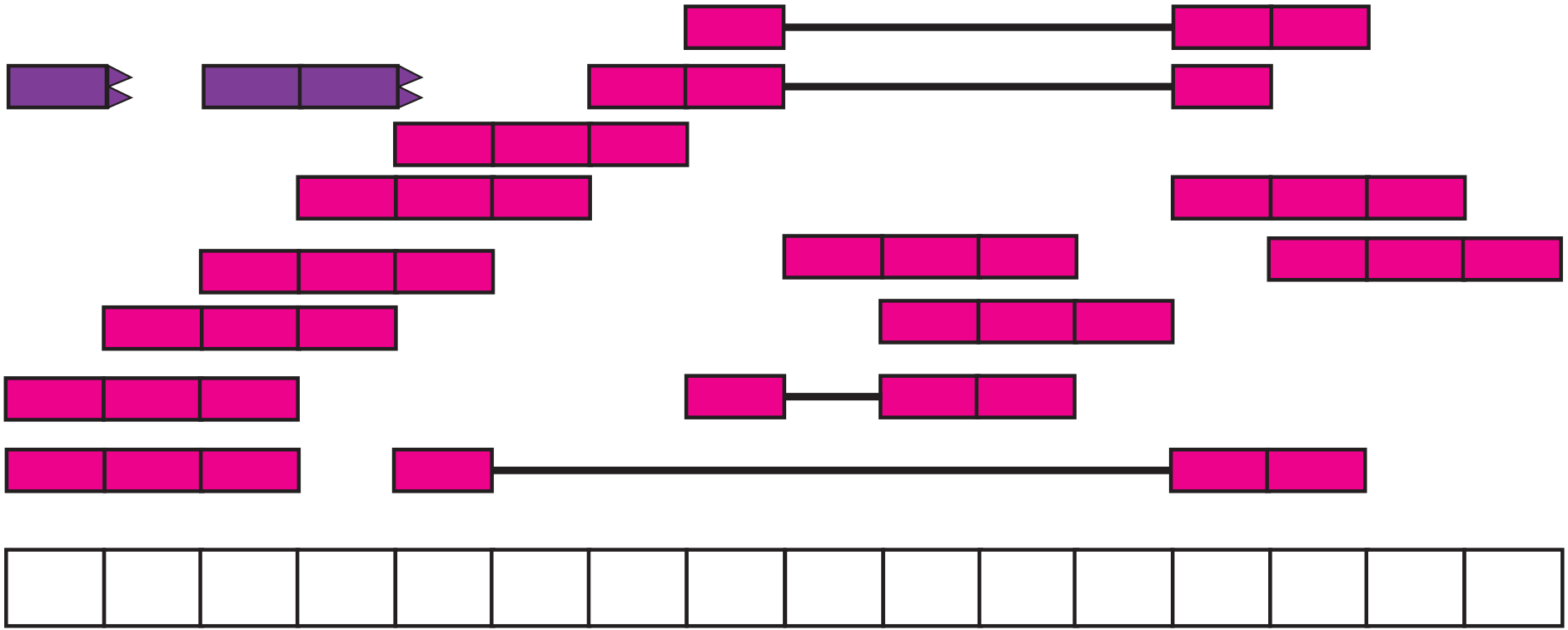
Leonardo Collado-Torres
@fellgernon
#bioc2017

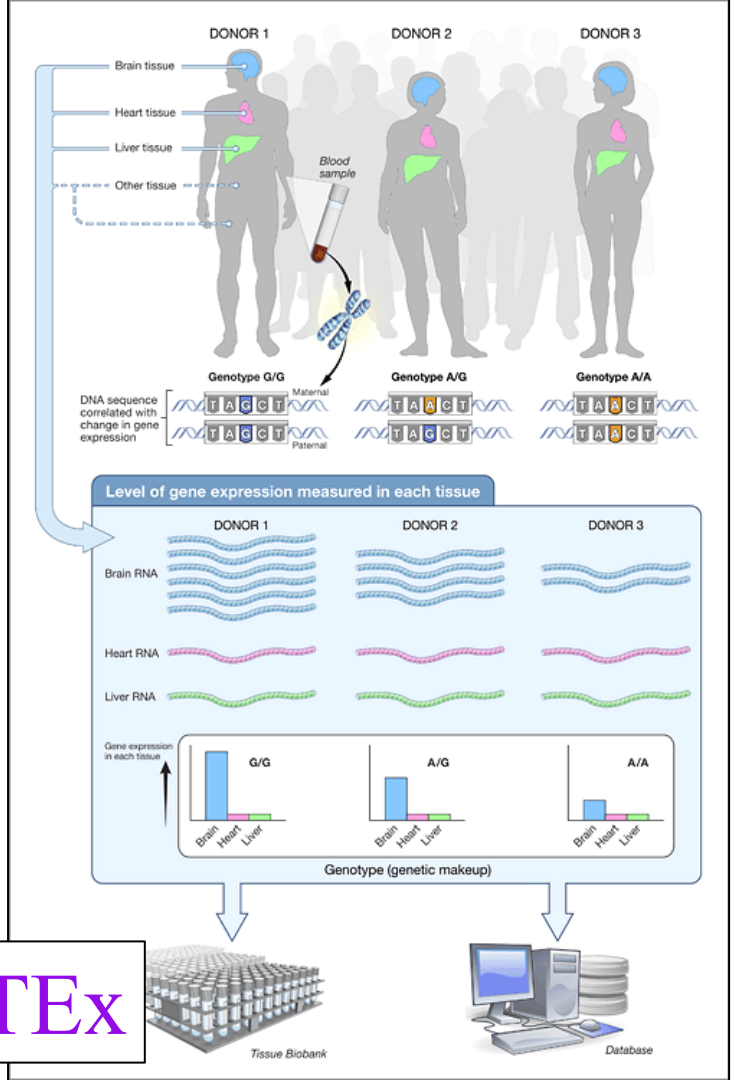
LIEBER INSTITUTE *for*
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES

Reads

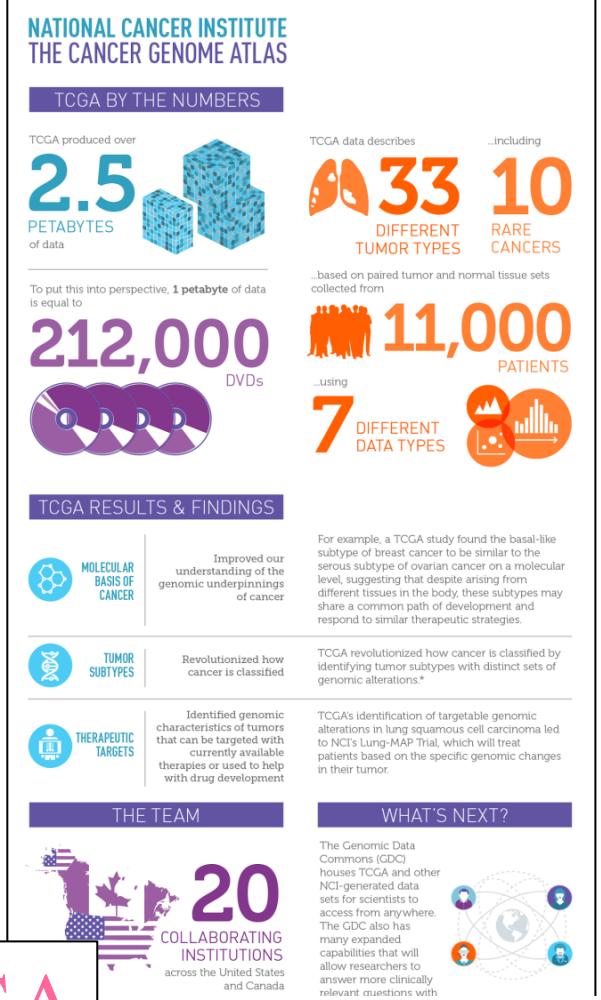


Reference genome



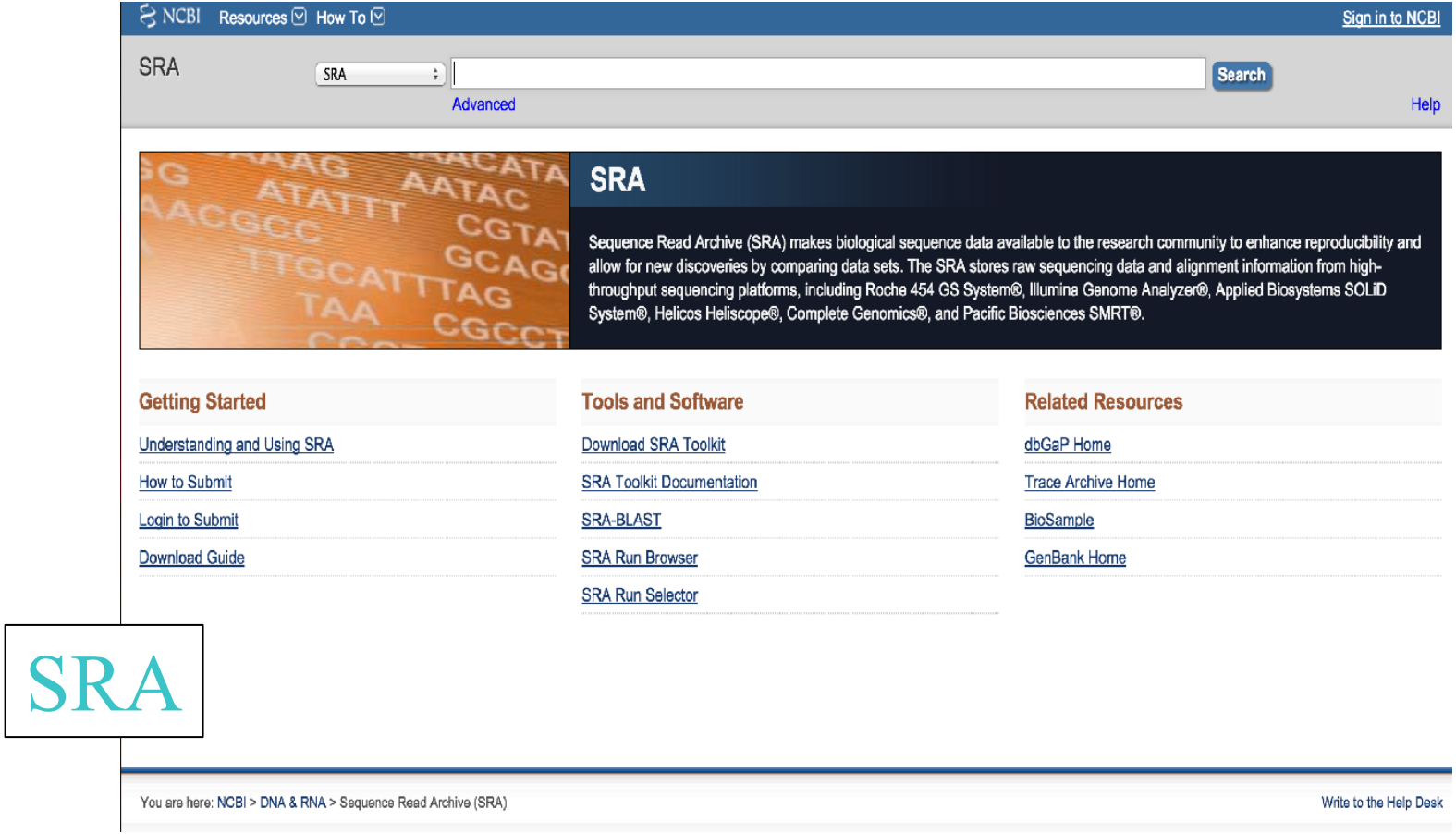


GTEx



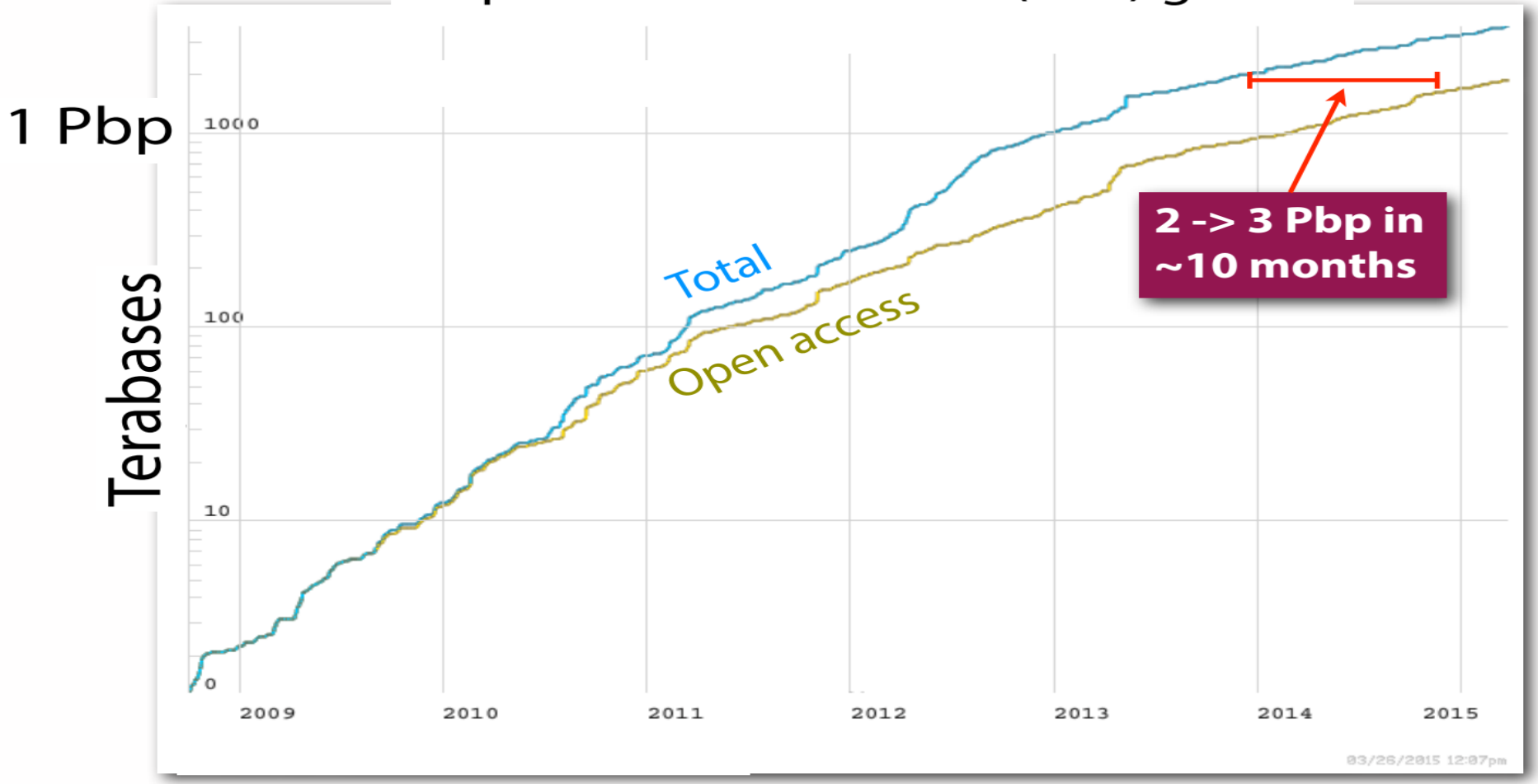
TCGA

slide adapted from Shannon Ellis

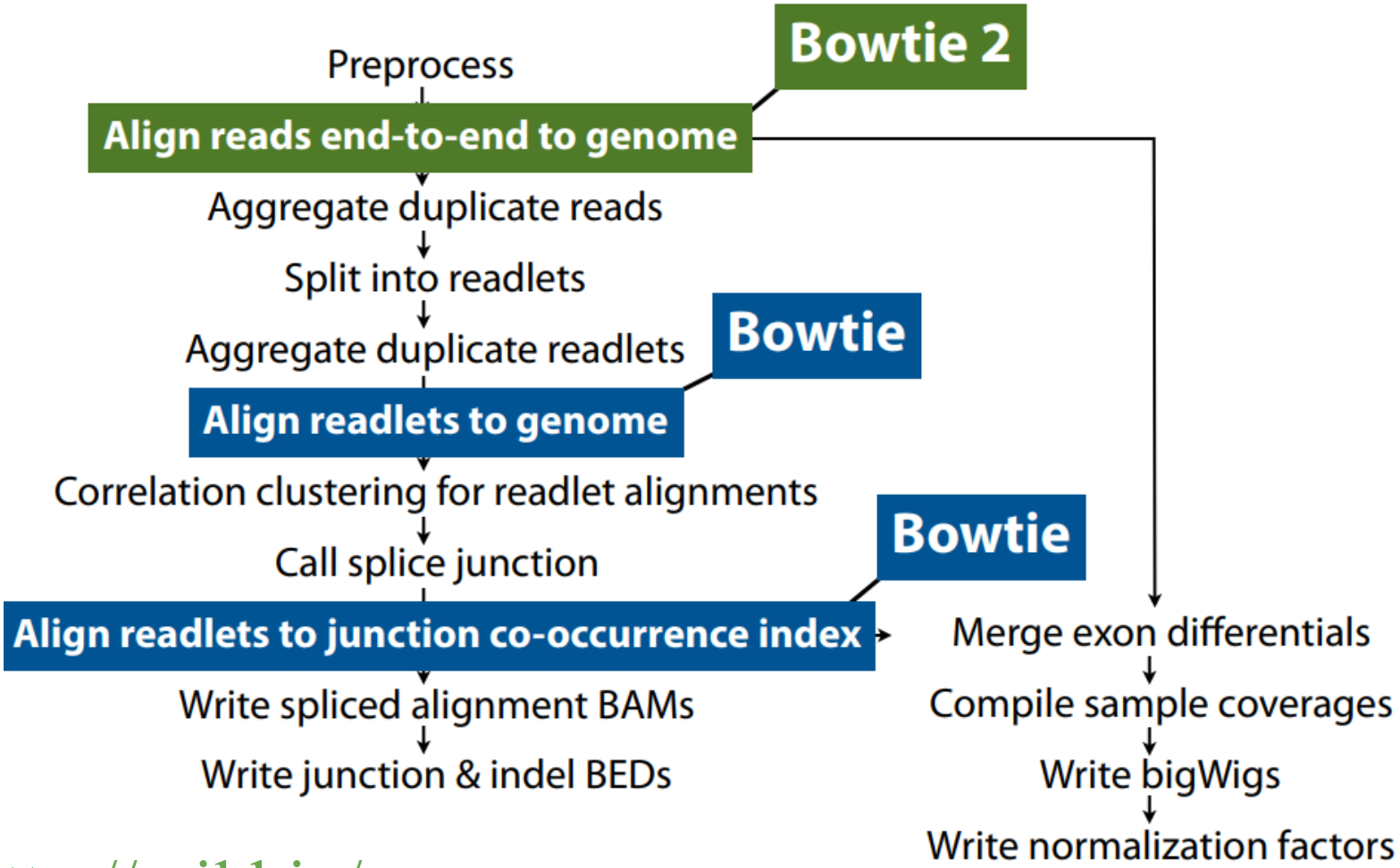

NCBI Resources How To
Sign in to NCBI
SRA
SRA
Search
Advanced
Help
SRA
Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information 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

Sequence Read Archive (SRA) growth

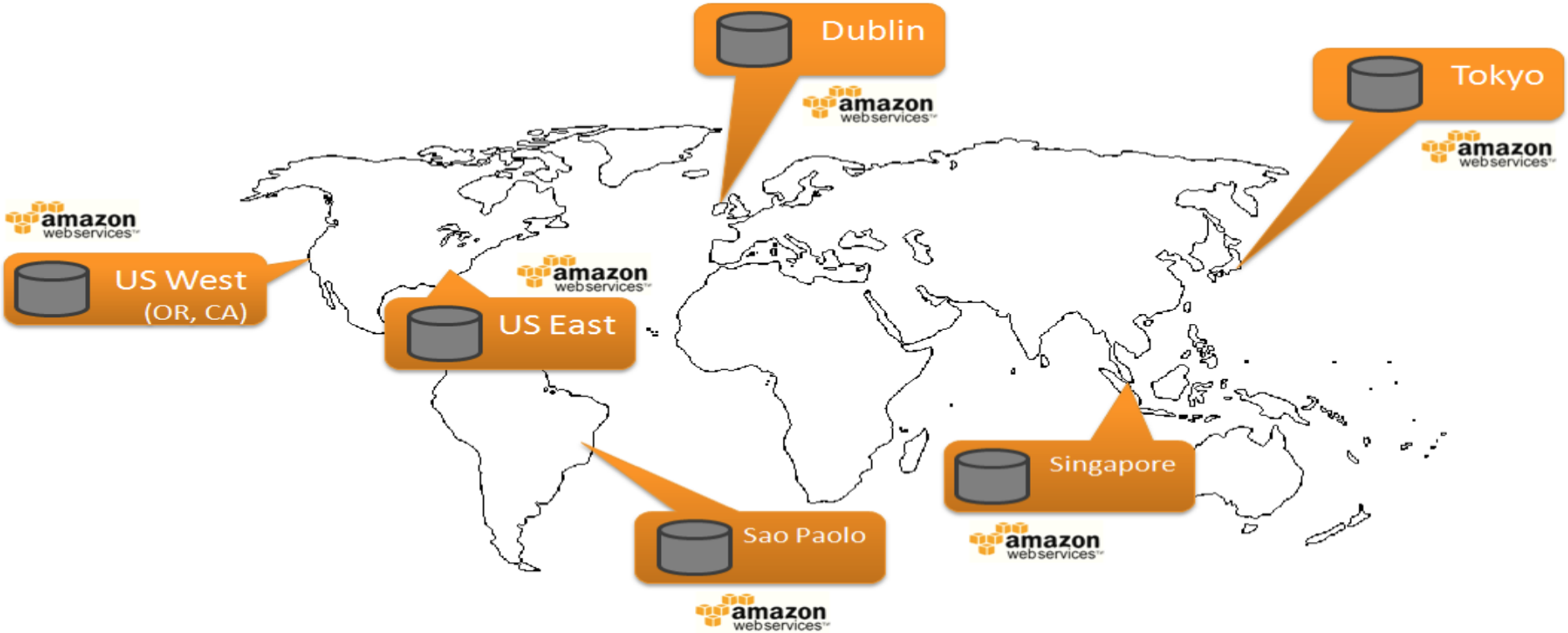


Slide adapted from Ben Langmead



http://rail.bio/

Slide adapted from Ben Langmead



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

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

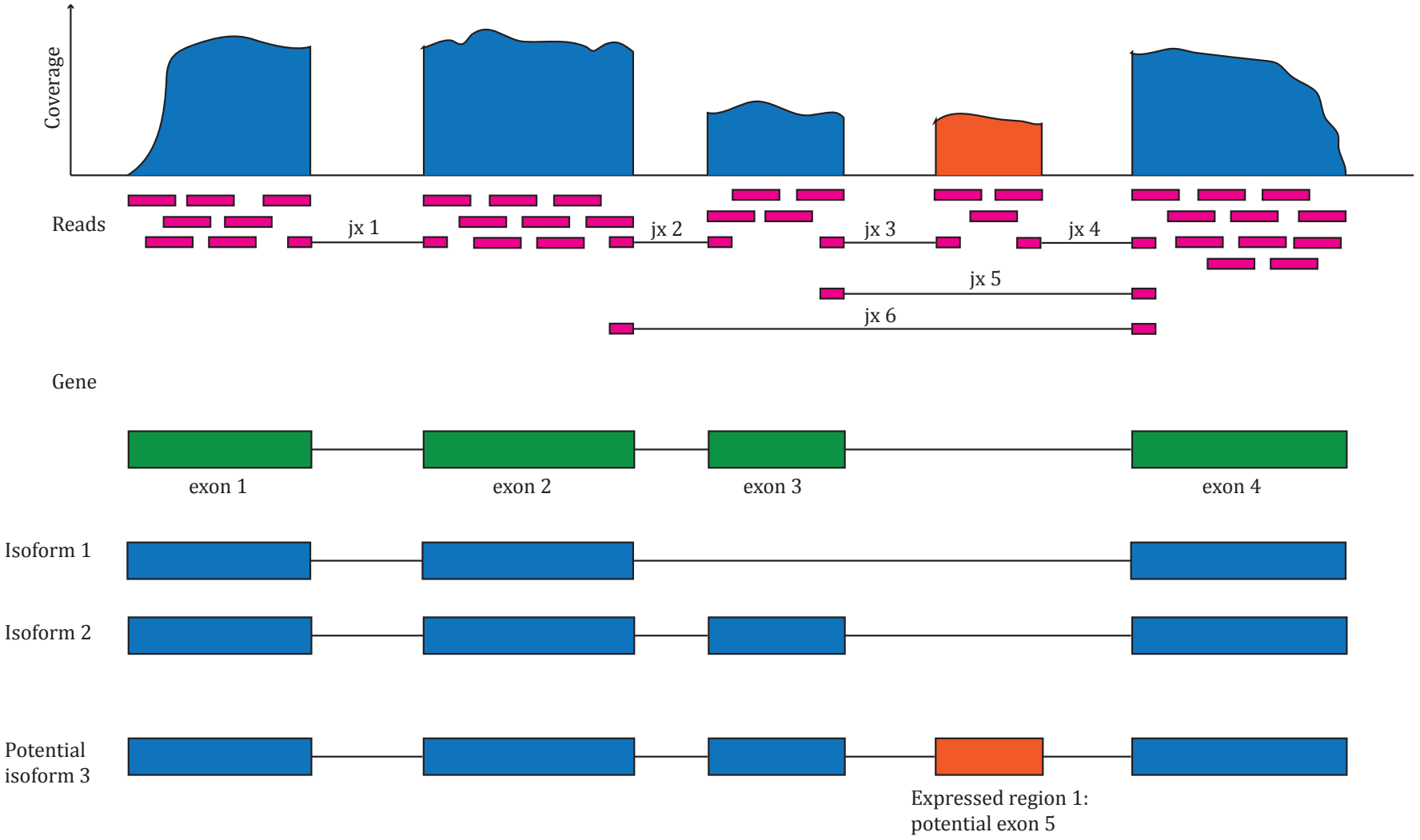
Collado-Torres L, **Nellore A**, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using *recount2*. *Nature Biotechnology*, 2017. doi: 10.1038/nbt.3838.

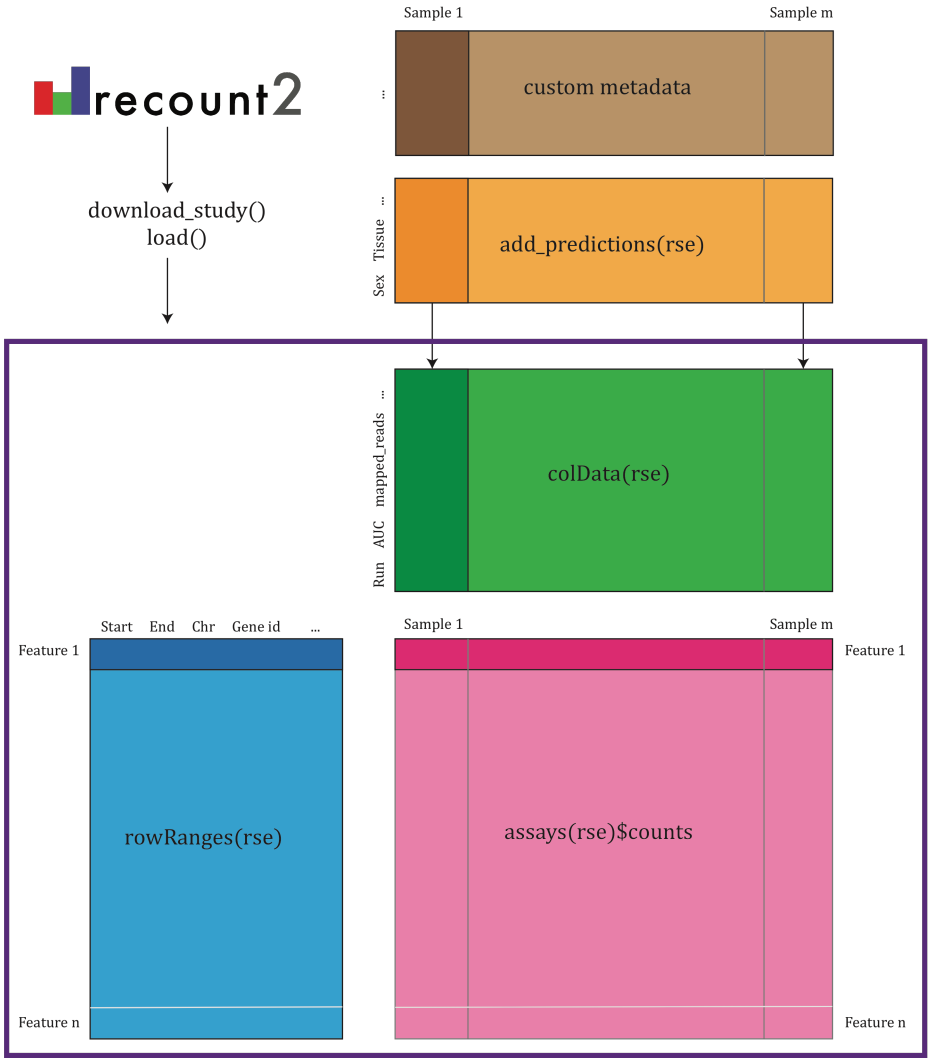
The Datasets

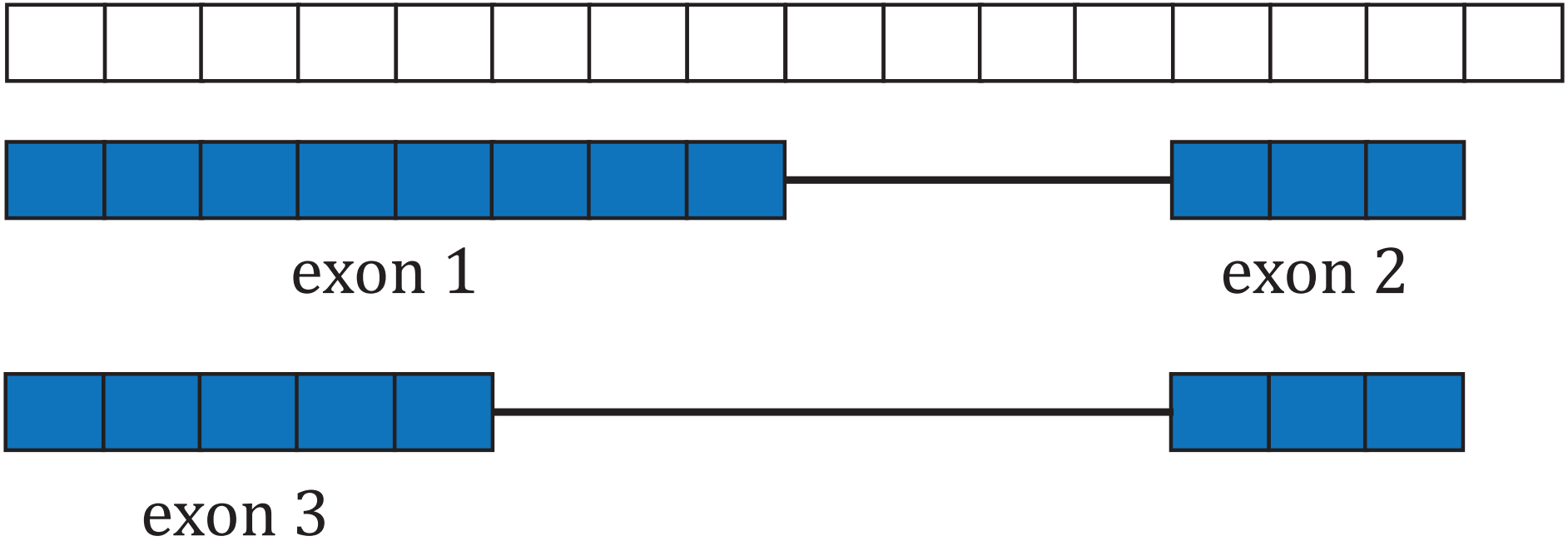
https://jhubiostatistics.shinyapps.io/recount/

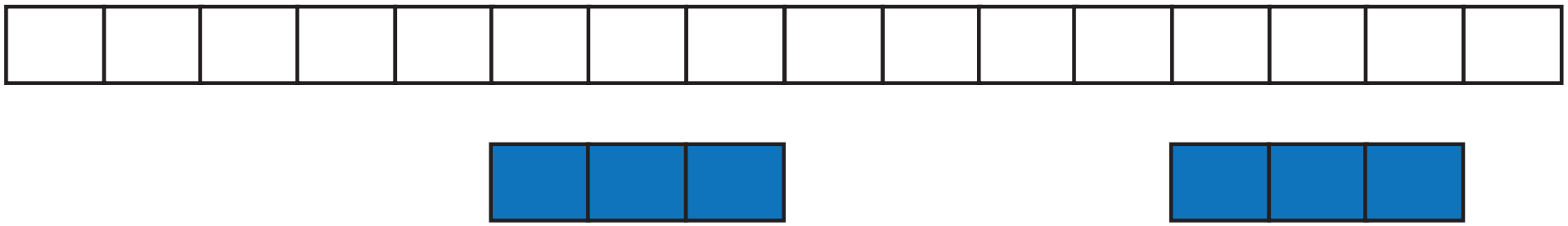
Show 10 entries

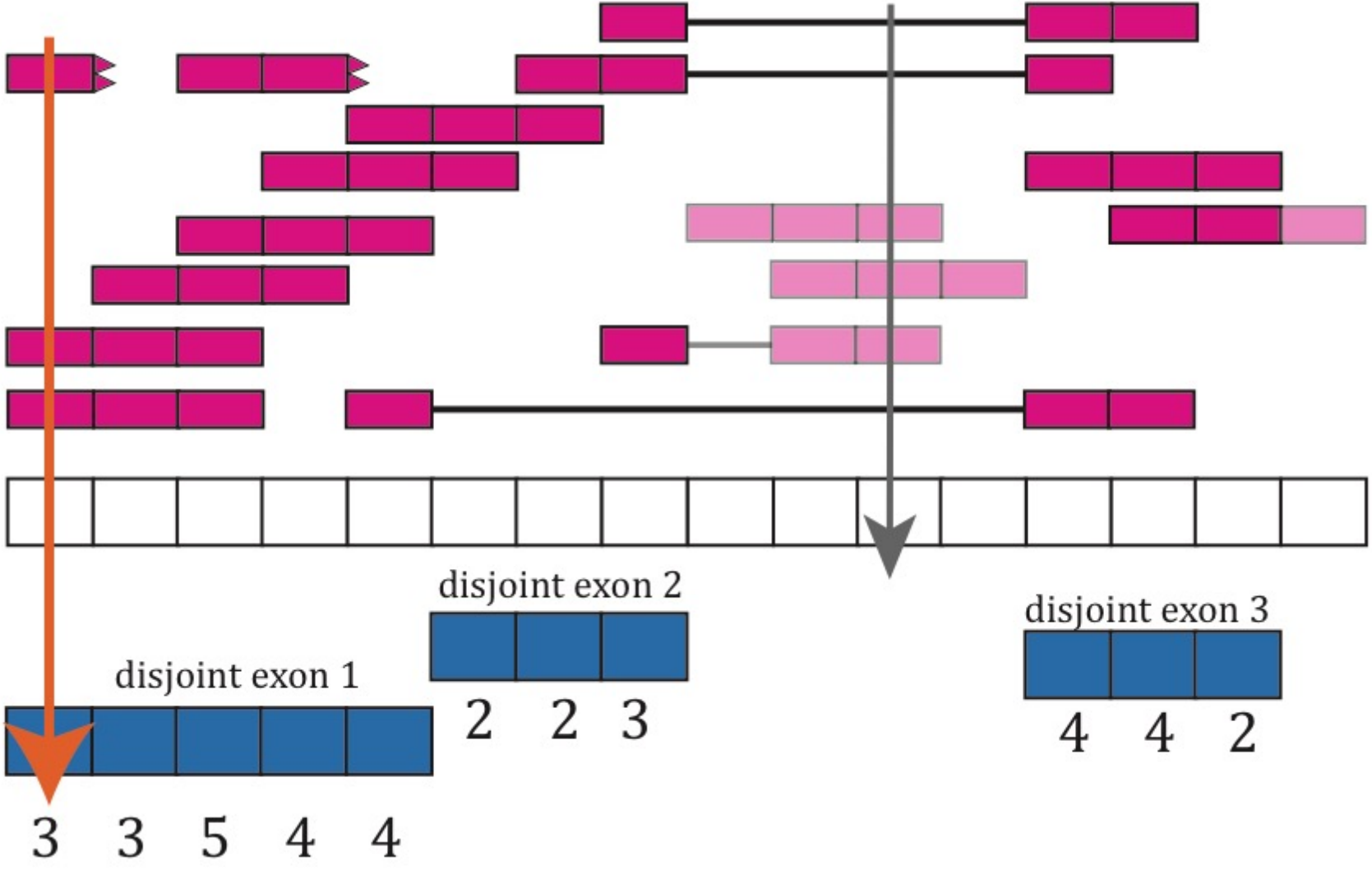
accession	number of samples	species	abstract	gene	exon	junctions	phenotype	files info
All	All	All	libd	All	All	All	All	All
SRP045638	72	human	RNAseq data of 36 samples across human brain development by age group from LIBD	RSE counts	RSE counts	RSE jx_bed jx_cov counts	link	link

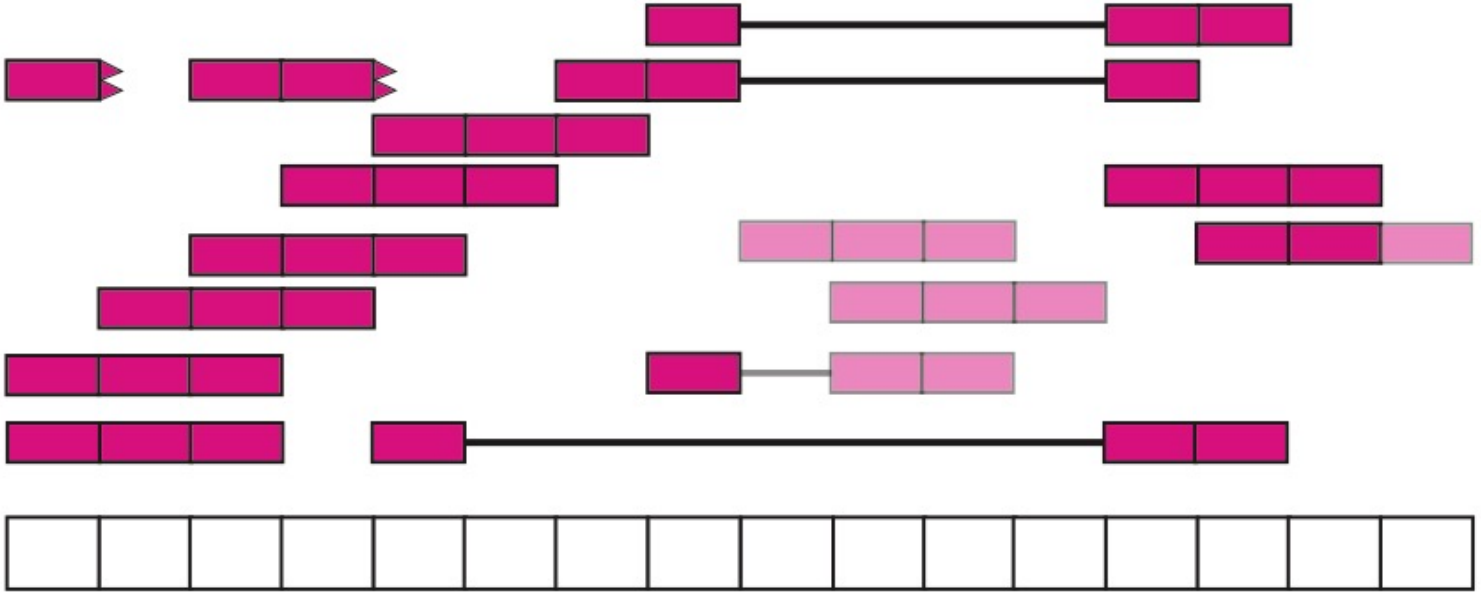

Coverage
Reads
jx 1
jx 2
jx 3
jx 4
jx 5
jx 6
Gene
exon 1
exon 2
exon 3
exon 4
Isoform 1
Isoform 2
Potential isoform 3
Expressed region 1: potential exon 5


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

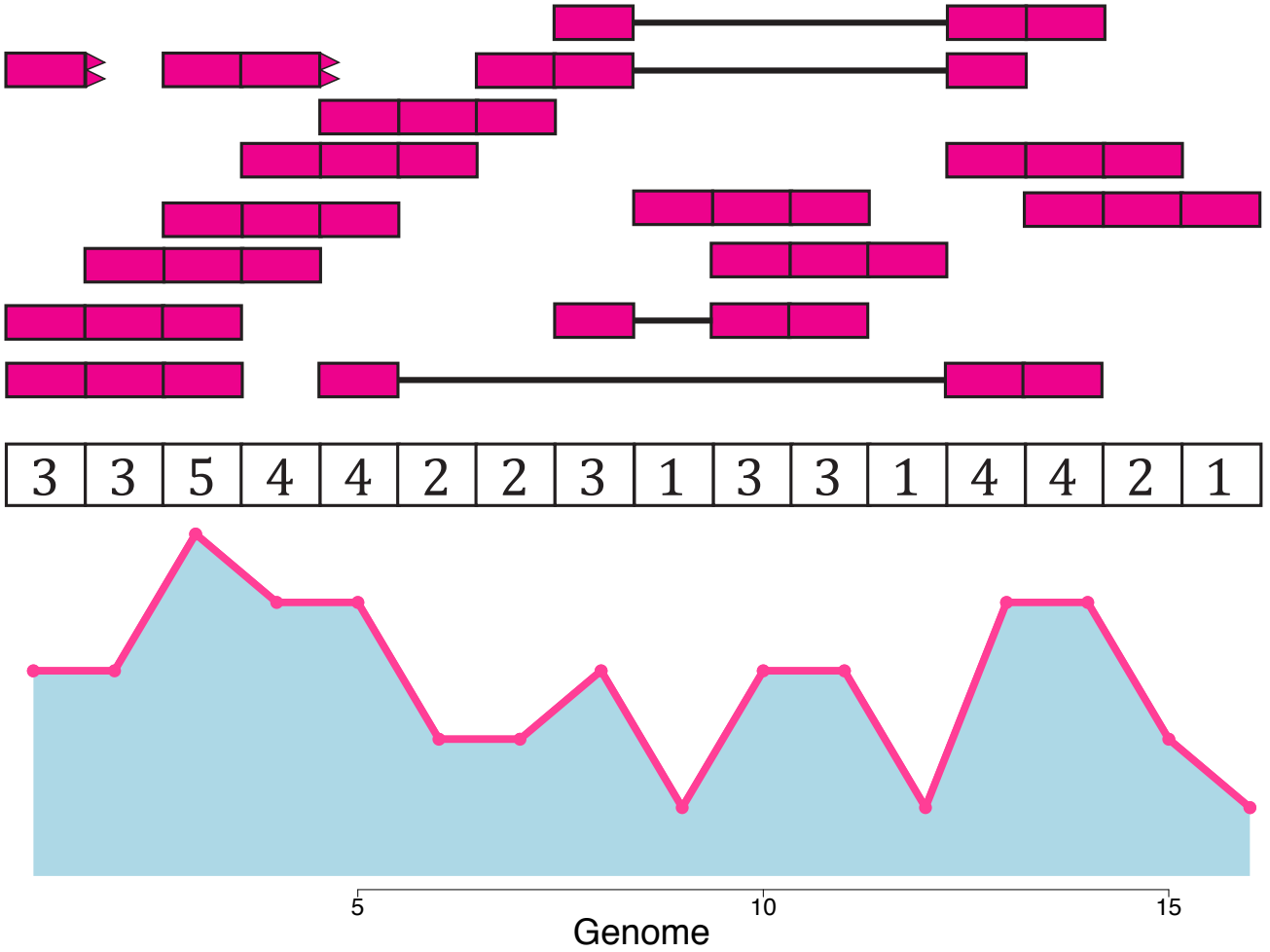
2+2+3=7

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

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

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

```
> 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

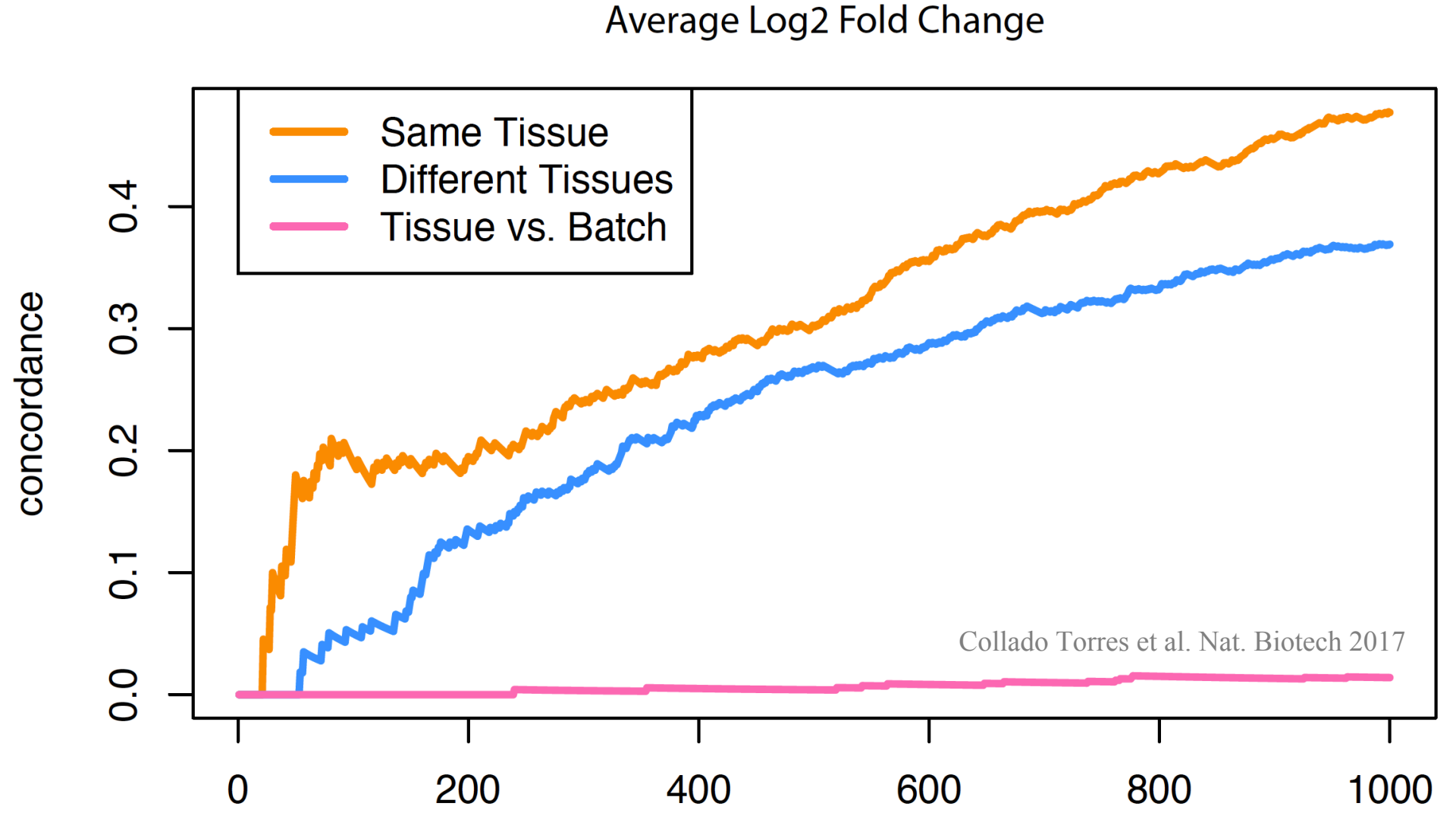
10:17 AM - 11 Apr 2017

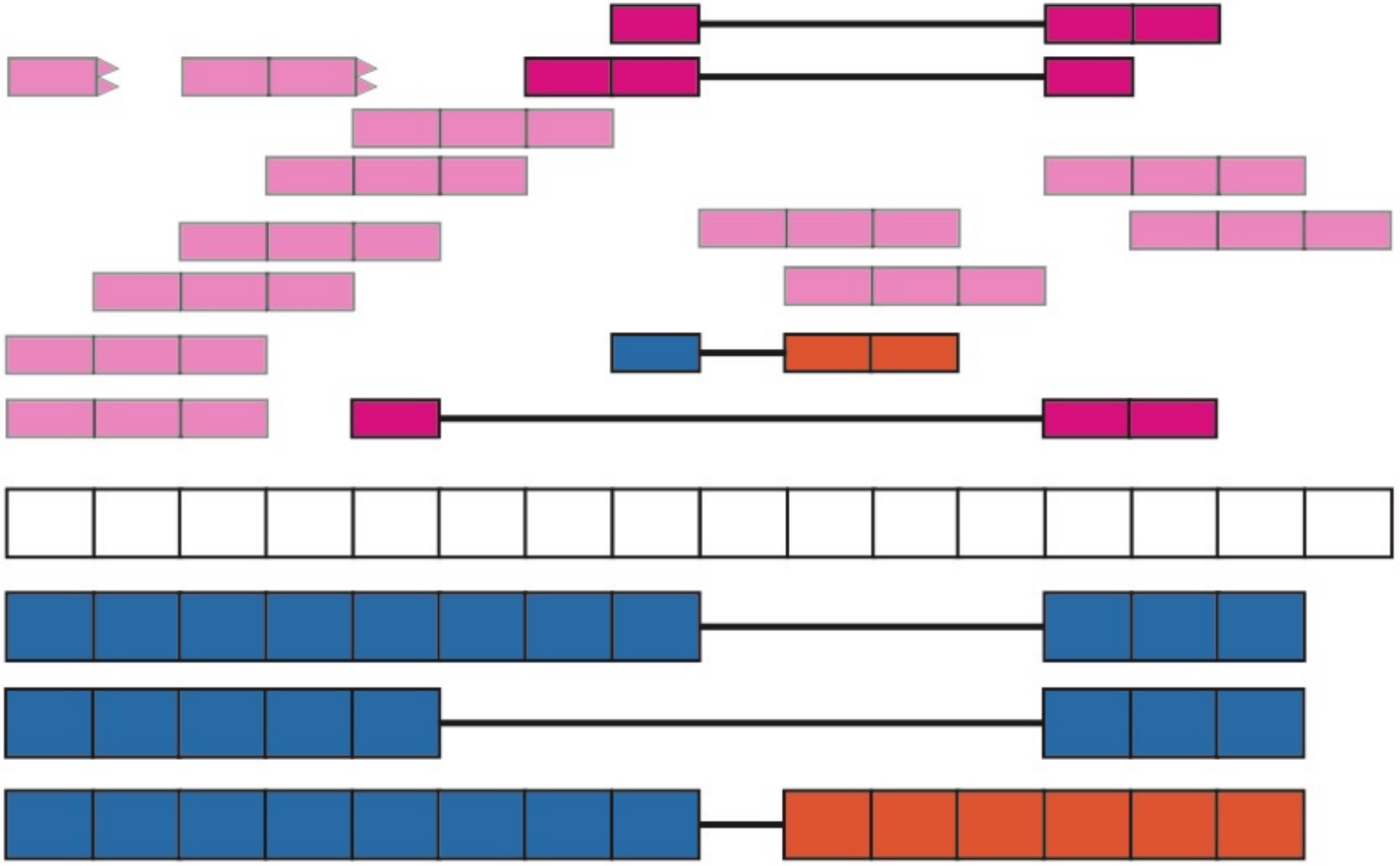
4 5

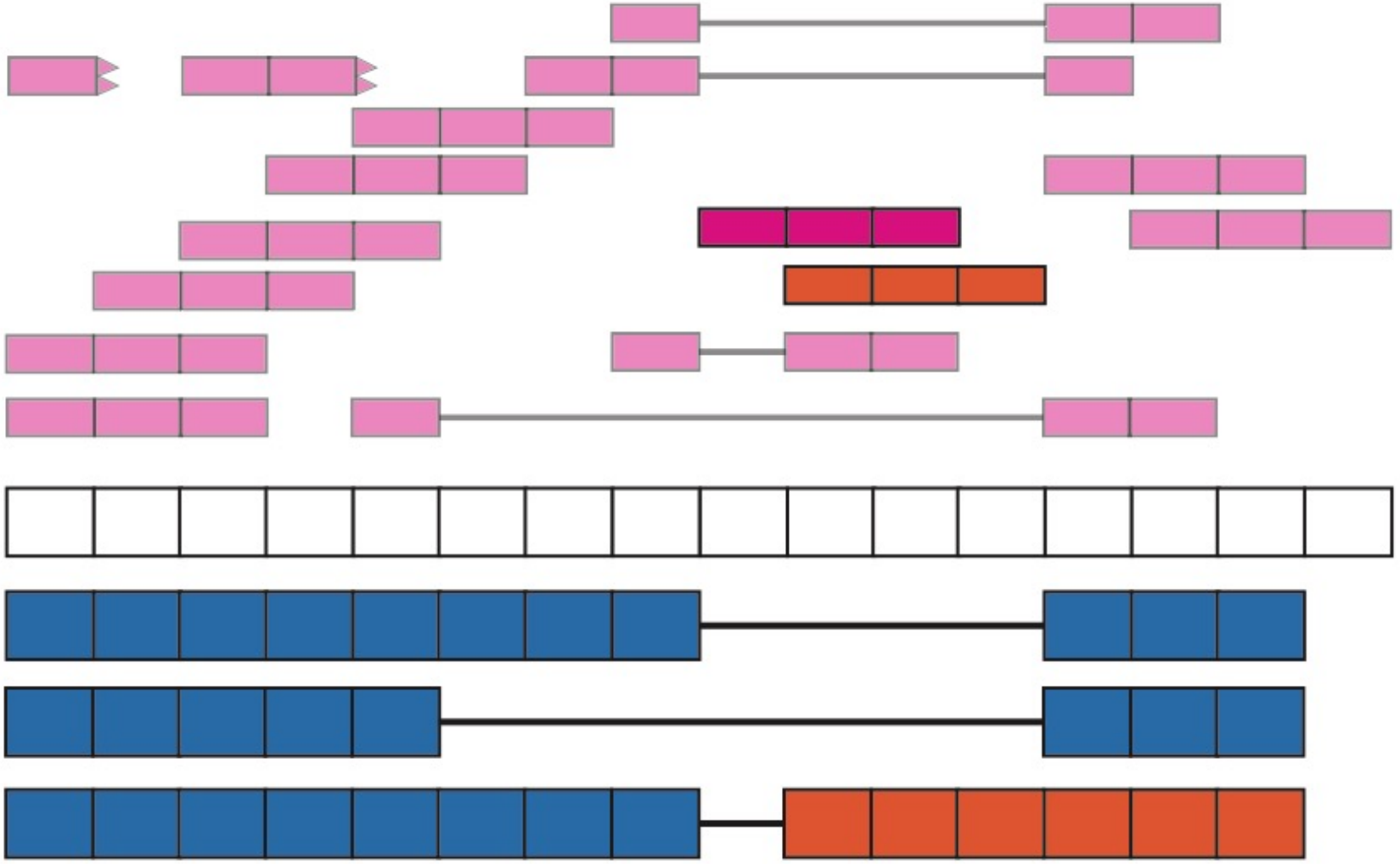
slide adapted from Jeff Leek

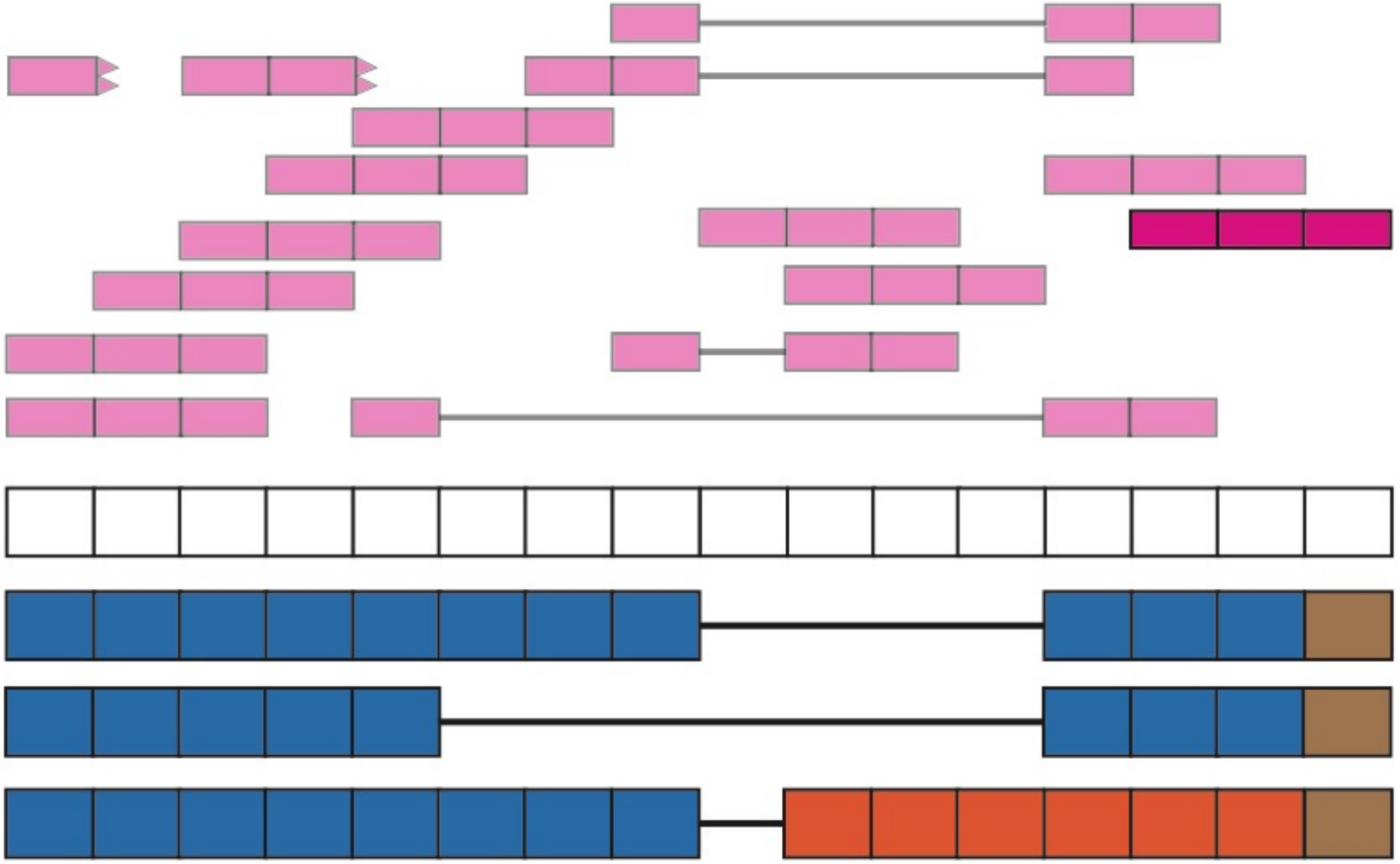
```
>library('recount')

> 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)
```

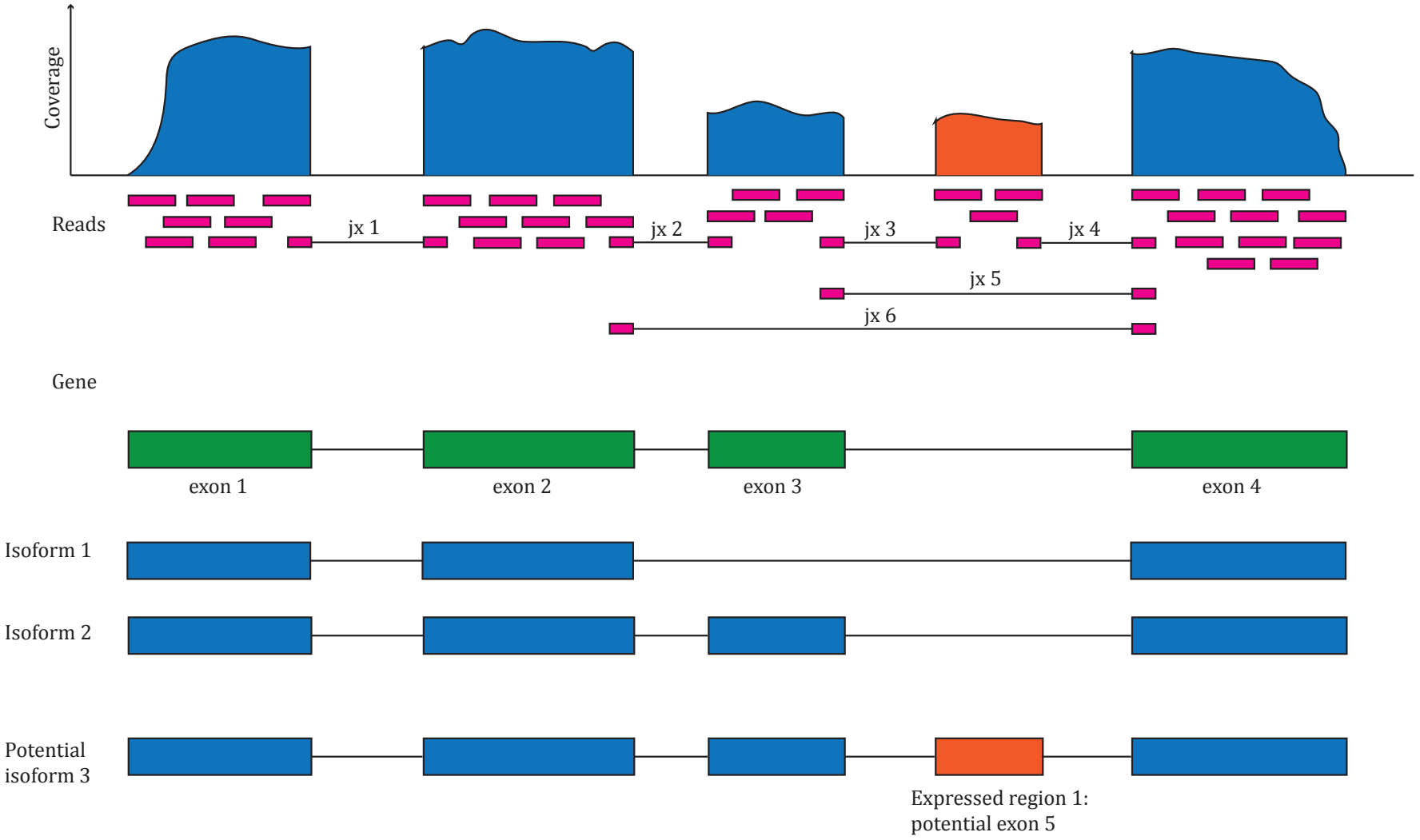
https://github.com/leekgroup/recount-analyses/

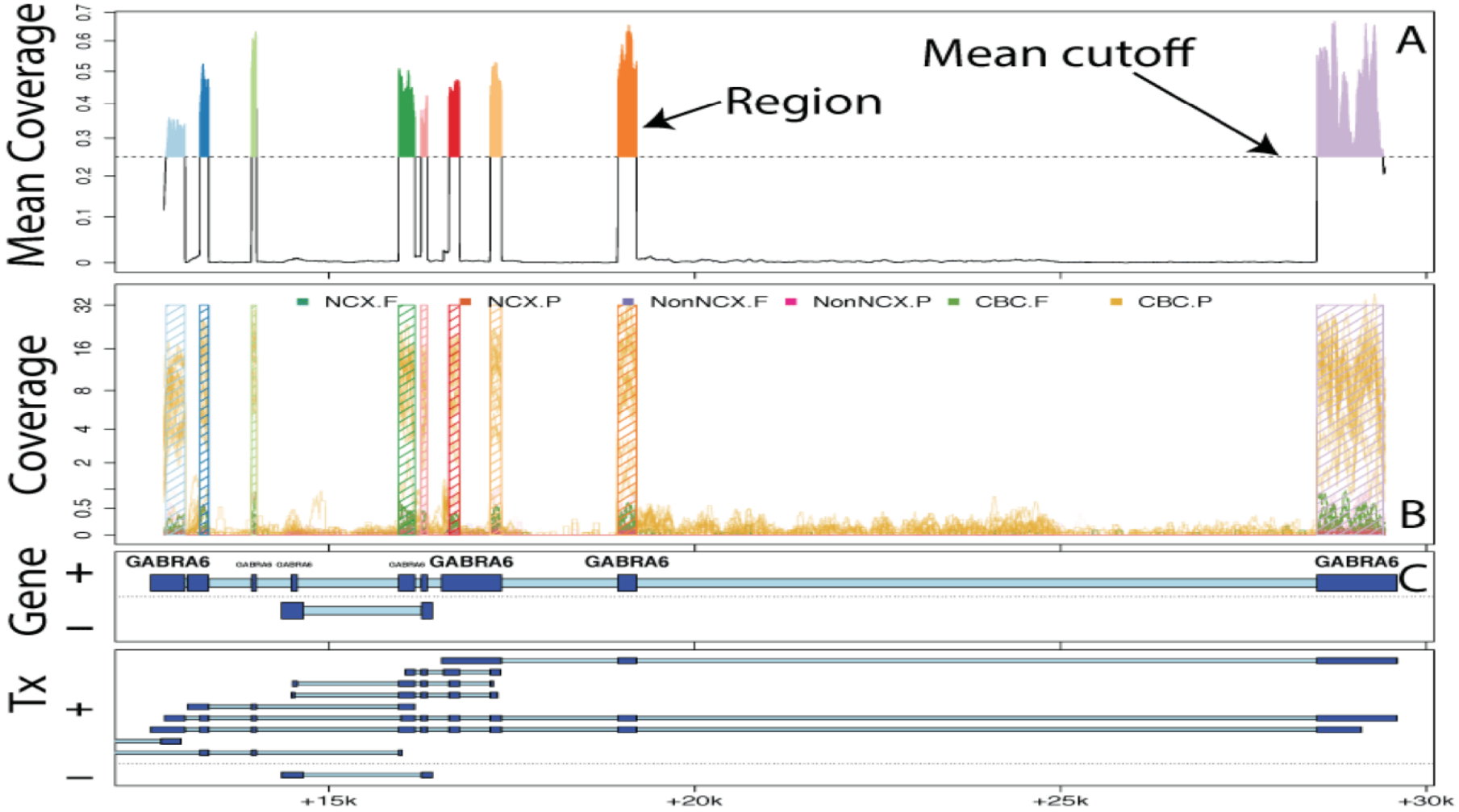

Average Log2 Fold Change
Same Tissue
Different Tissues
Tissue vs. Batch
concordance
0.0
0.1
0.2
0.3
0.4
0
200
400
600
800
1000
Collado Torres et al. Nat. Biotech 2017





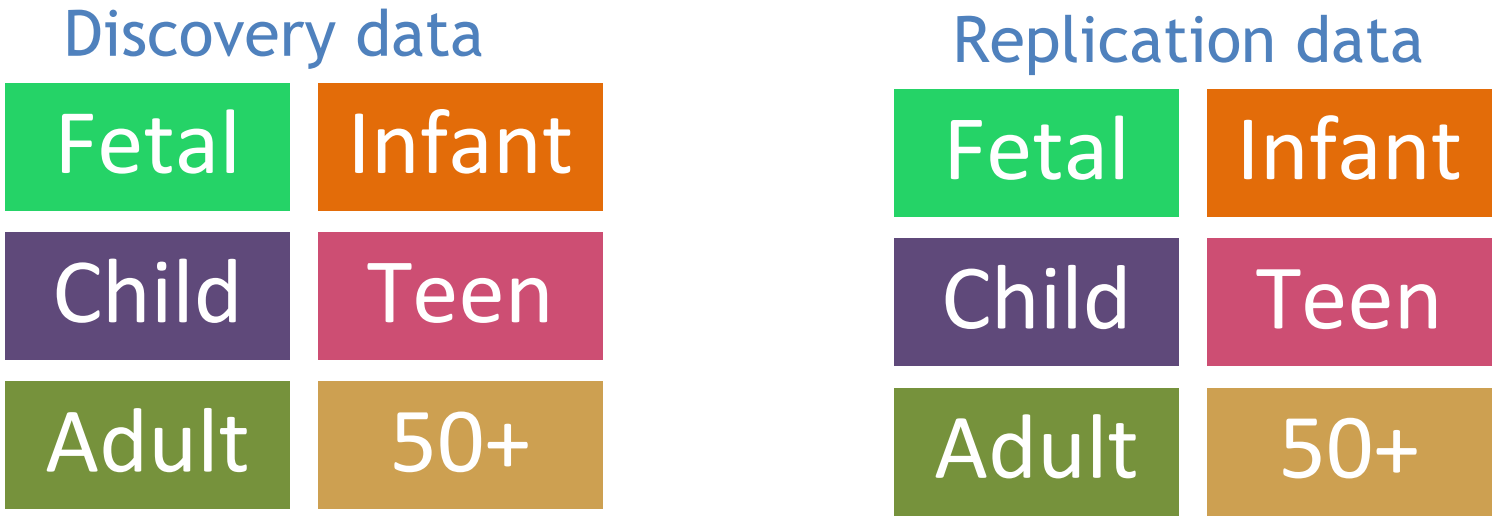



Coverage
Reads
jx 1
jx 2
jx 3
jx 4
jx 5
jx 6
Gene
exon 1
exon 2
exon 3
exon 4
Isoform 1
Isoform 2
Potential isoform 3
Expressed region 1: potential exon 5

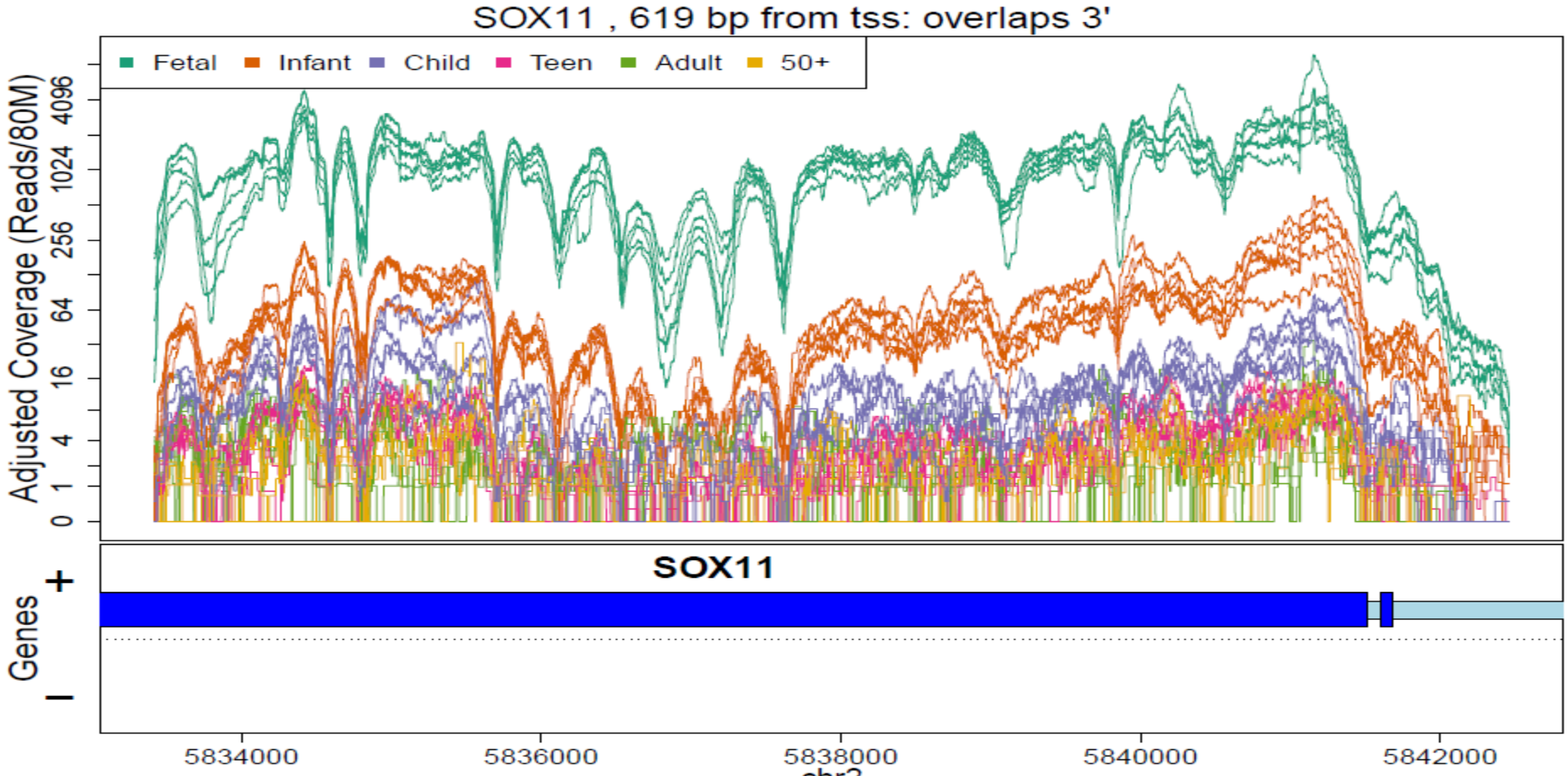


Collado-Torres et al, *NAR*, 2017

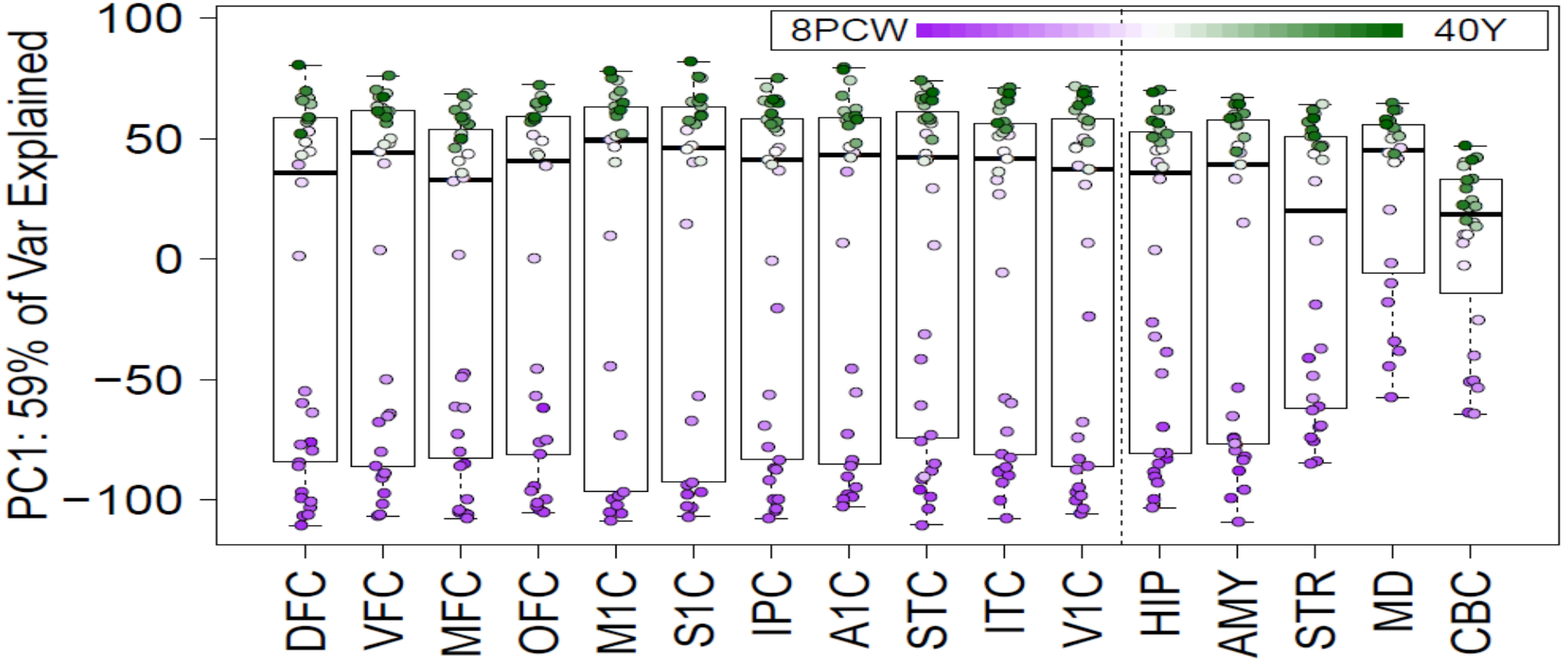
Postmortem Human Brain Samples



Jaffe et al, *Nat. Neuroscience*, 2015



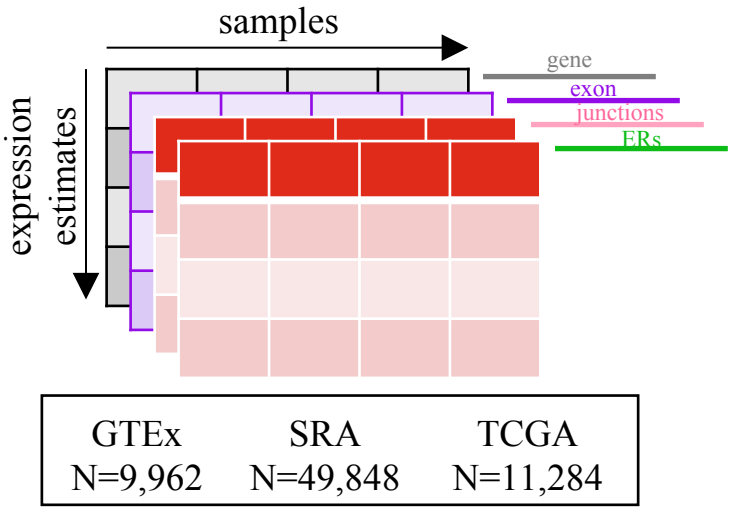
Jaffe et al, *Nat. Neuroscience*, 2015



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

recount2

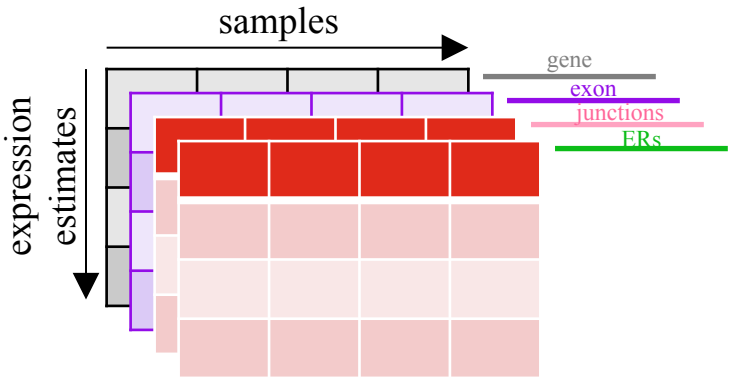
expression data for ~70,000 human samples



slide adapted from Shannon Ellis

recount2

expression data for ~70,000 human samples



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

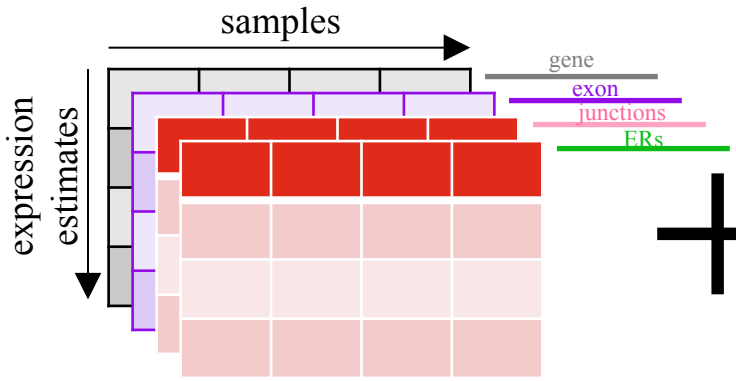
→

Answer meaningful questions about human biology and expression

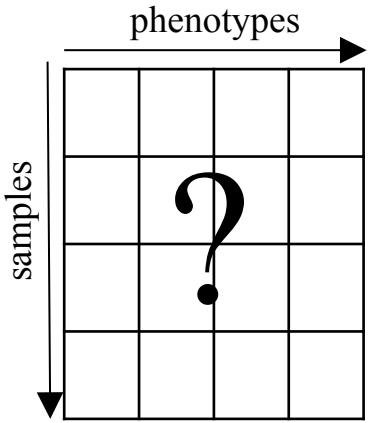
slide adapted from Shannon Ellis

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$Se
x

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

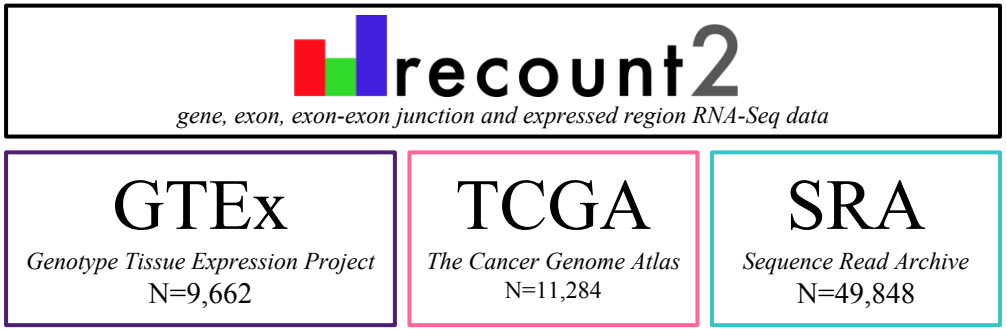
SRA phenotype information is far from complete

	SubjectID	Sex	Tissue	Race	Age
6620	NA	female	liver	NA	NA
6621	NA	female	liver	NA	NA
6622	NA	female	liver	NA	NA
6623	NA	female	liver	NA	NA
6624	NA	female	liver	NA	NA
6625	NA	male	liver	NA	NA
6626	NA	male	liver	NA	NA
6627	NA	male	liver	NA	NA
6628	NA	male	liver	NA	NA
6629	NA	male	liver	NA	NA
6630	NA	male	liver	NA	NA
6631	NA	NA	blood	NA	NA
6632	NA	NA	blood	NA	NA
6633	NA	NA	blood	NA	NA
6634	NA	NA	blood	NA	NA
6635	NA	NA	blood	NA	NA
6636	NA	NA	blood	NA	

slide adapted from Shannon Ellis

Goal :

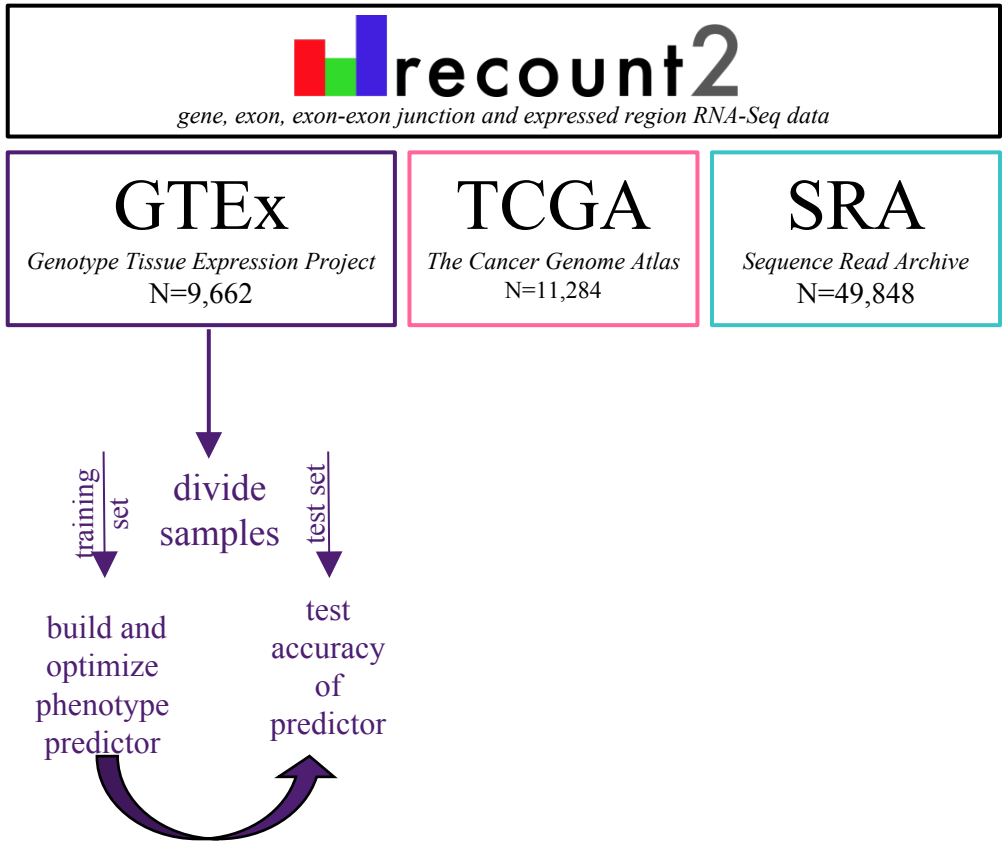
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

Goal :

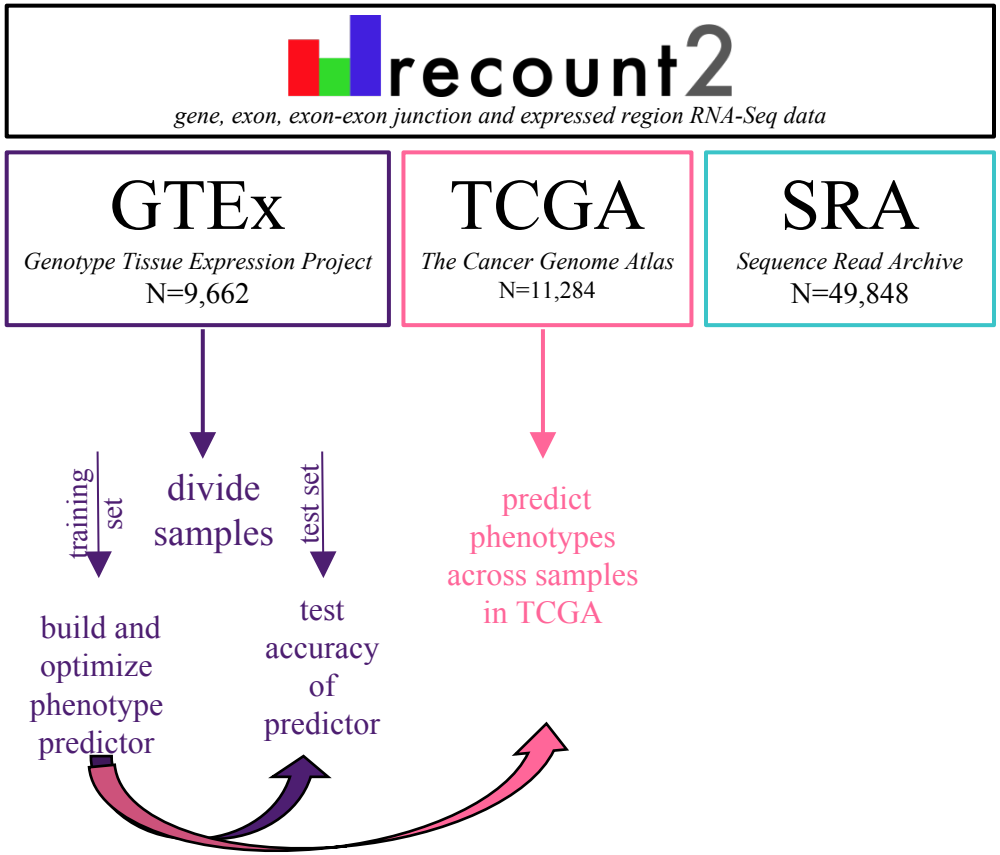
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

Goal :

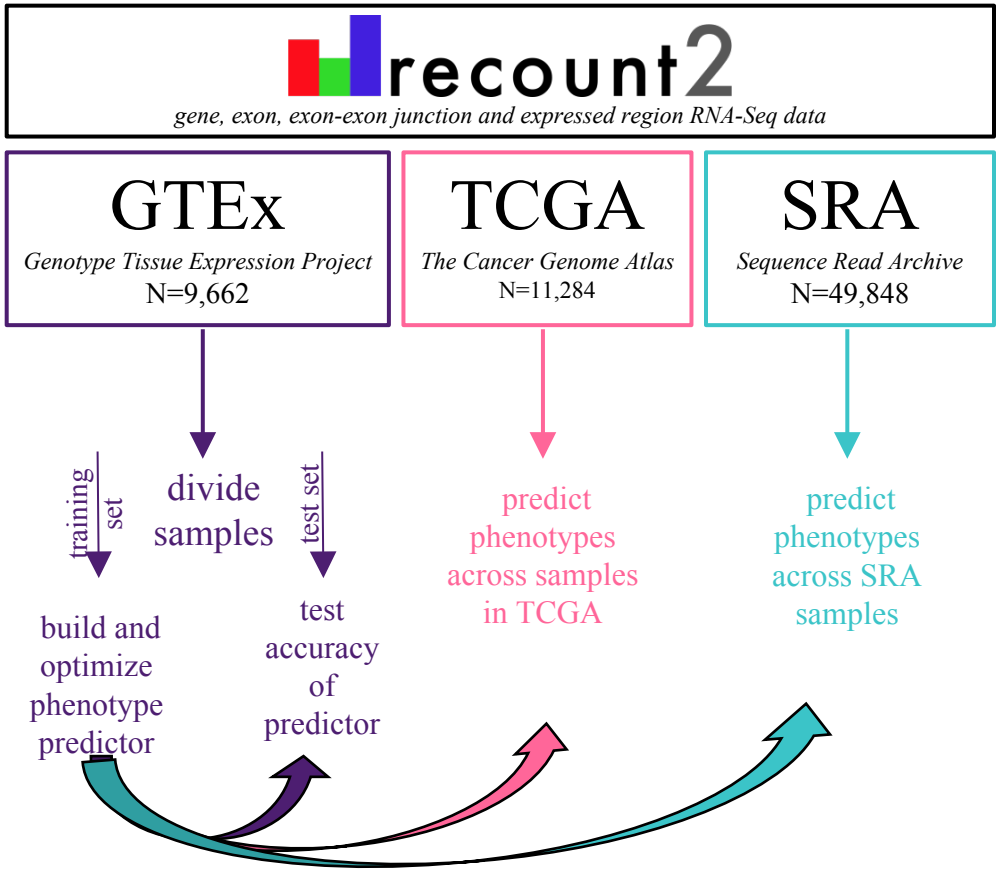
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

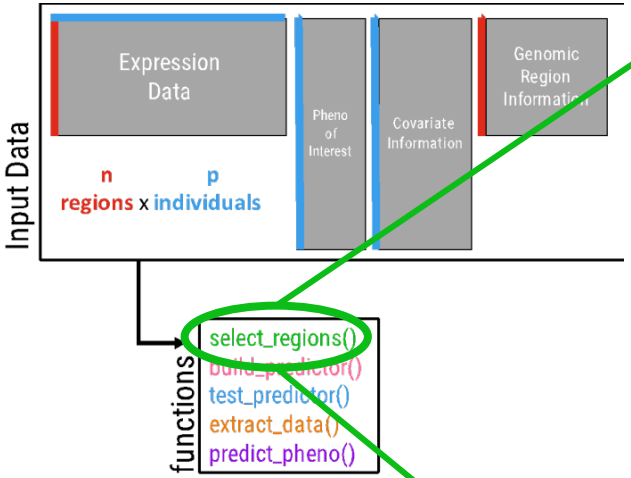
Goal :

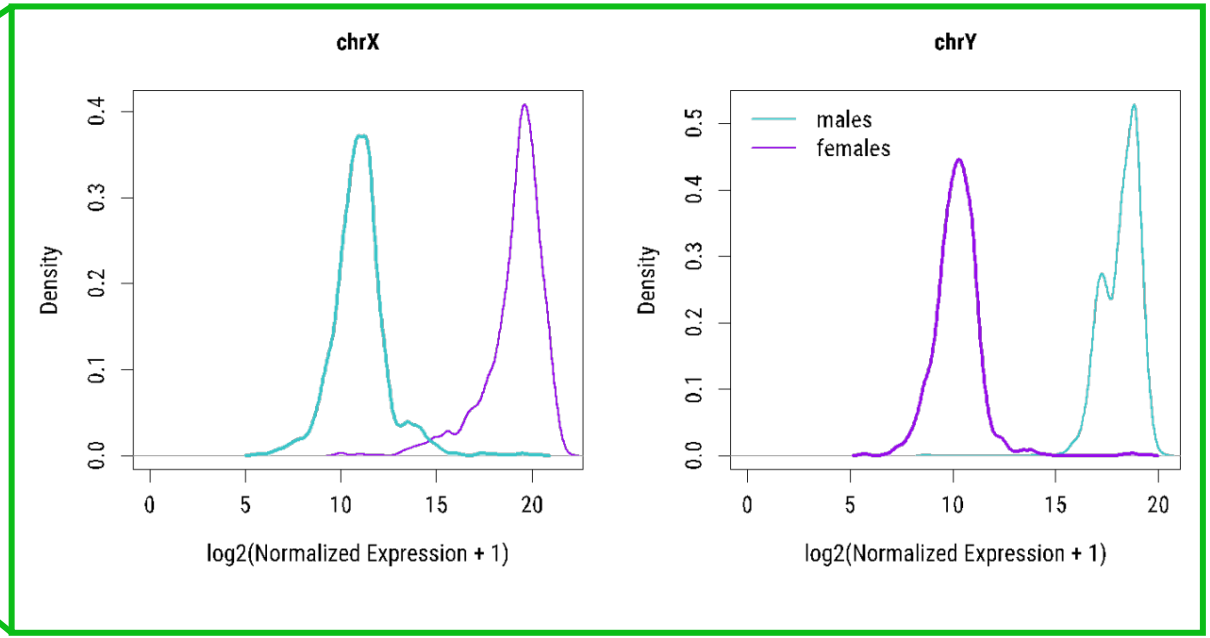
to accurately predict critical phenotype information for all samples in *recount*



slide adapted from Shannon Ellis

select_regions()

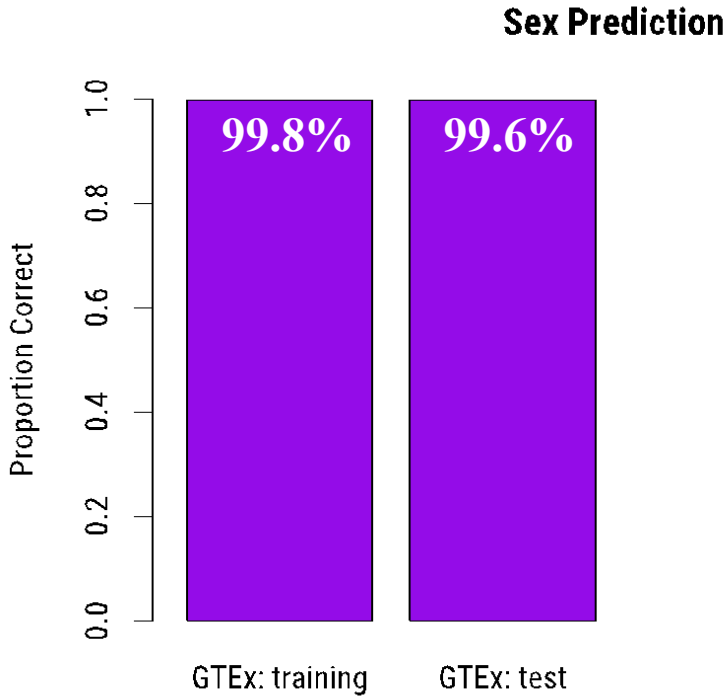




Output:
Coverage matrix (data.frame)
Region information (GRanges)

slide adapted from Shannon Ellis

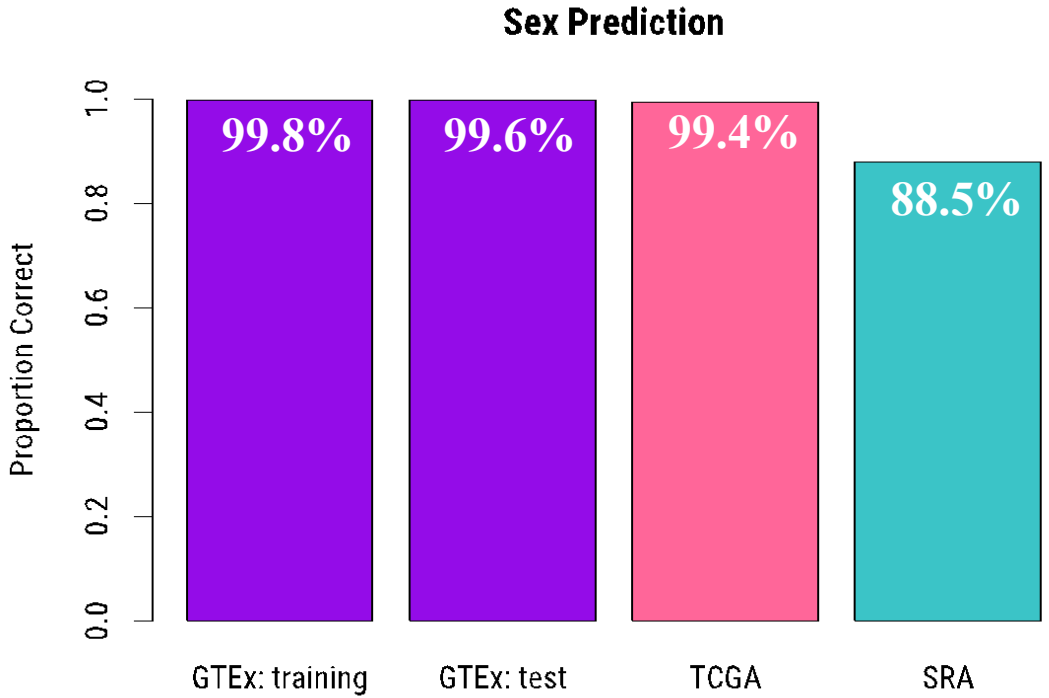
Sex prediction is accurate across data sets



Number of Regions	**20**	**20**
Number of Samples (N)	**4,769**	**4,769**

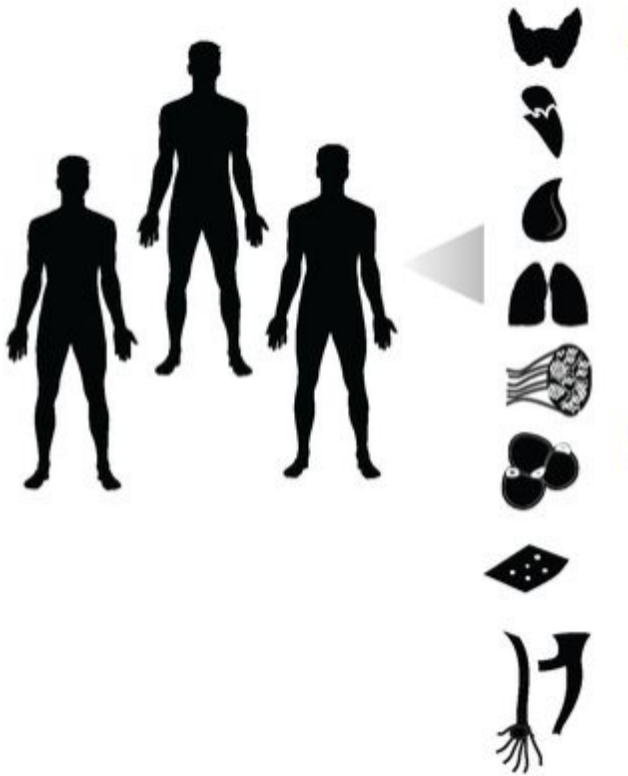
slide adapted from Shannon Ellis

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

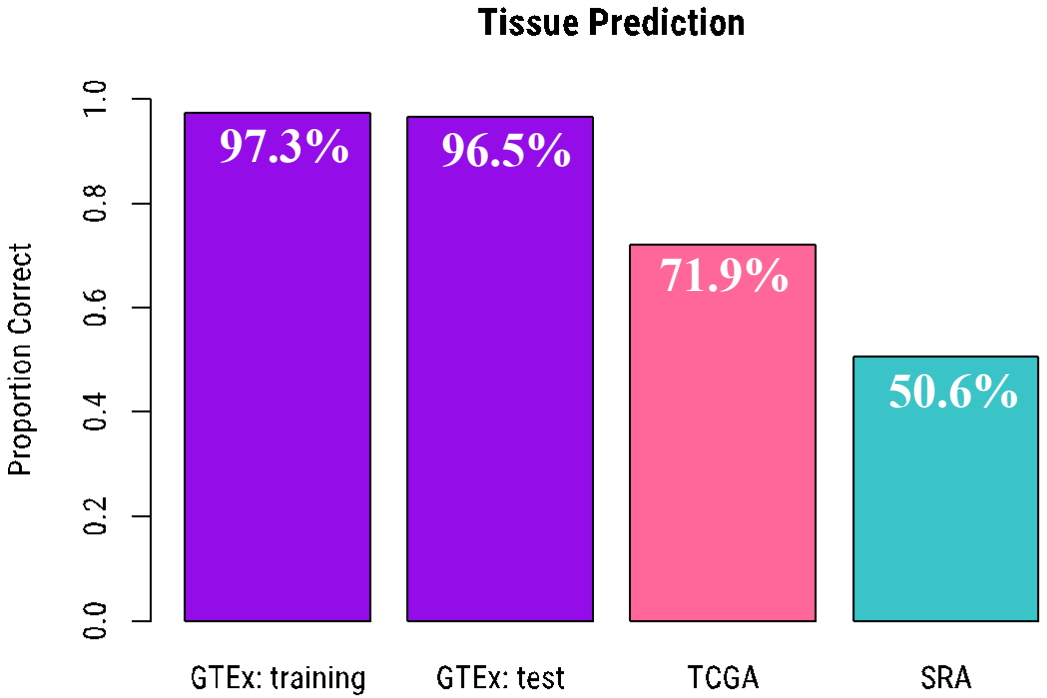


http://www.rna-seqblog.com/

Can we use expression data to predict tissue?

slide adapted from Shannon Ellis

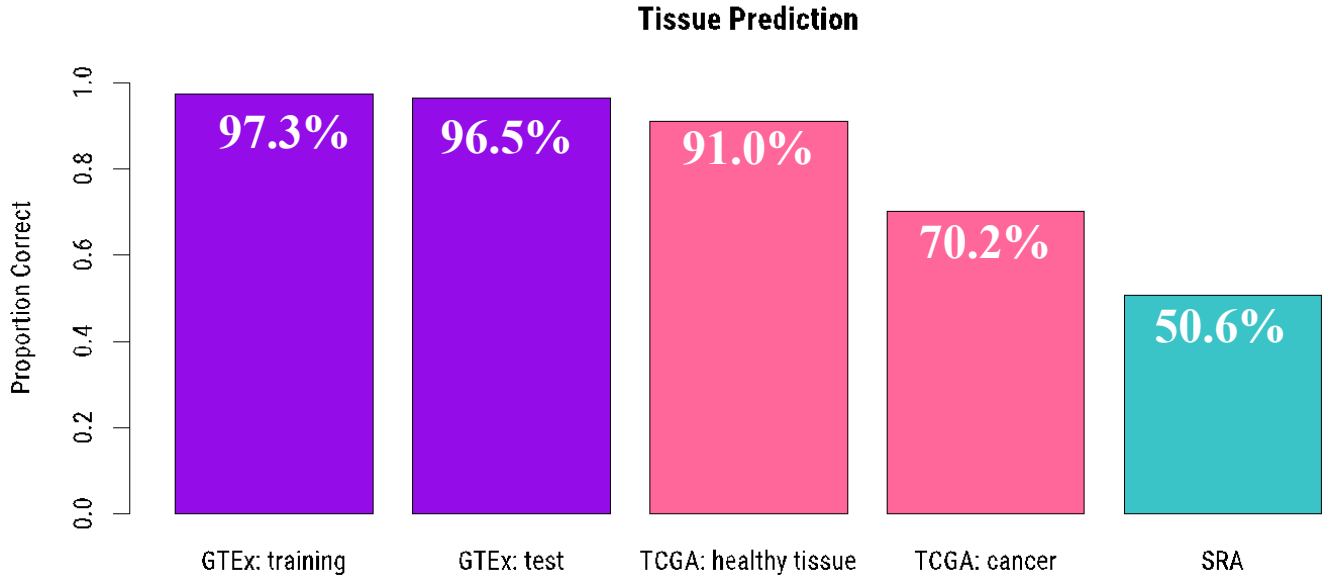
Tissue prediction is accurate across data sets



Number of Regions	589	589	589	589
Number of Samples (N)	4,769	4,769	7,193	8,951

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

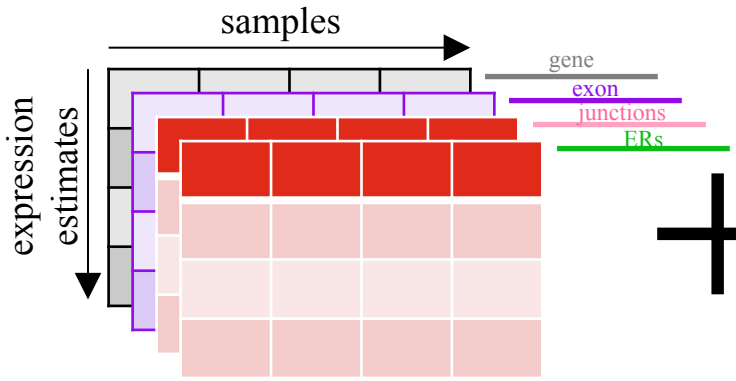
slide adapted from Shannon Ellis

```
> 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)
```

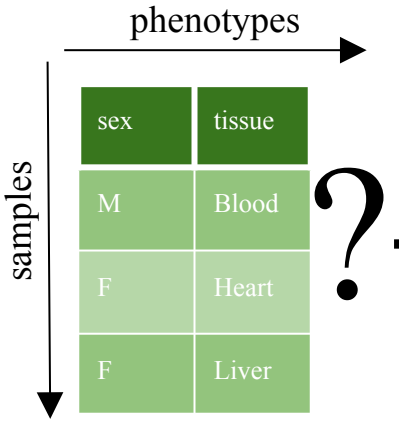
https://github.com/leekgroup/recount-analyses/

recount2

expression data for ~70,000 human samples

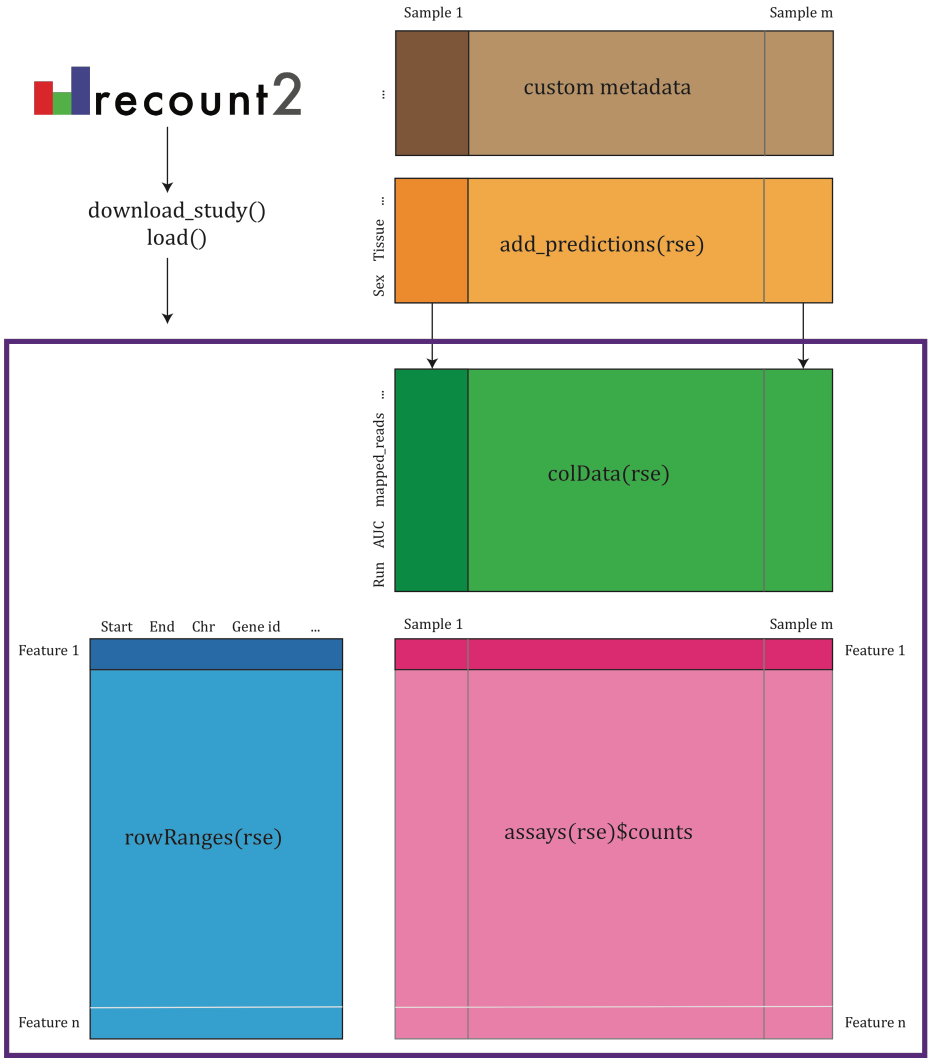


GTEx	SRA	TCGA
N=9,962	N=49,848	N=11,284



Answer meaningful questions about human biology and expression

slide adapted from Shannon Ellis


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

LIBD

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

LIEBER INSTITUTE *for*
BRAIN DEVELOPMENT
MALTZ RESEARCH LABORATORIES

LIBD

recount2

http://research.libd.org/recountWorkshop/

```
help(package = recountWorkshop)

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

Leonardo Collado-Torres
@fellgernon
#bioc2017

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)