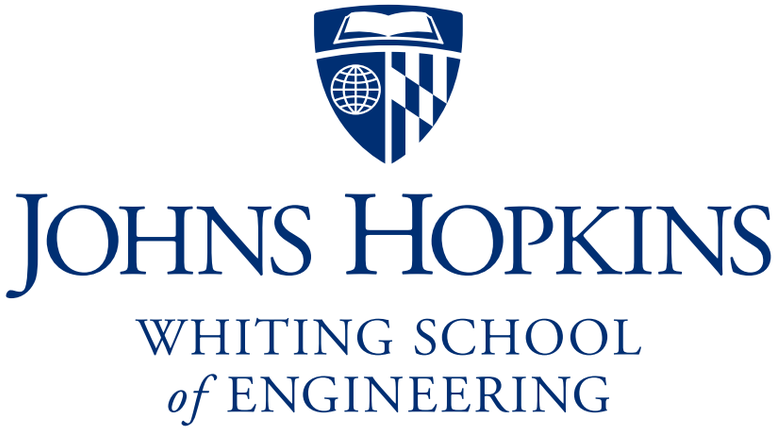
Marshaling public data for lean & powerful splicing studies

Ben Langmead
Associate Professor, JHU Computer Science
langmea@cs.jhu.edu, langmead-lab.org, @BenLangmead

National Cancer Institute
January 16, 2020



Lab goals

To make high-throughput life science data as usable as possible for scientific labs, especially small ones

Efficient

Software: Bowtie 1&2, Dashing, Kraken 2
Topics: applied algorithms, text indexing, sketching, thread scaling

Scalable

Software: Rail-RNA, recount2, Snaptron
Topics: parallel & high-performance computing, cloud computing, indexing

Interpretable

Software: Qtip, FORGe, r-index
Topics: modeling mapping quality, graph-genome variants, addressing biases

Lab goals

To make high-throughput life science data as usable as possible for scientific labs, especially small ones

Efficient

Software: Bowtie 1&2, Dashing, Kraken 2
Topics: applied algorithms, text indexing, sketching, thread scaling

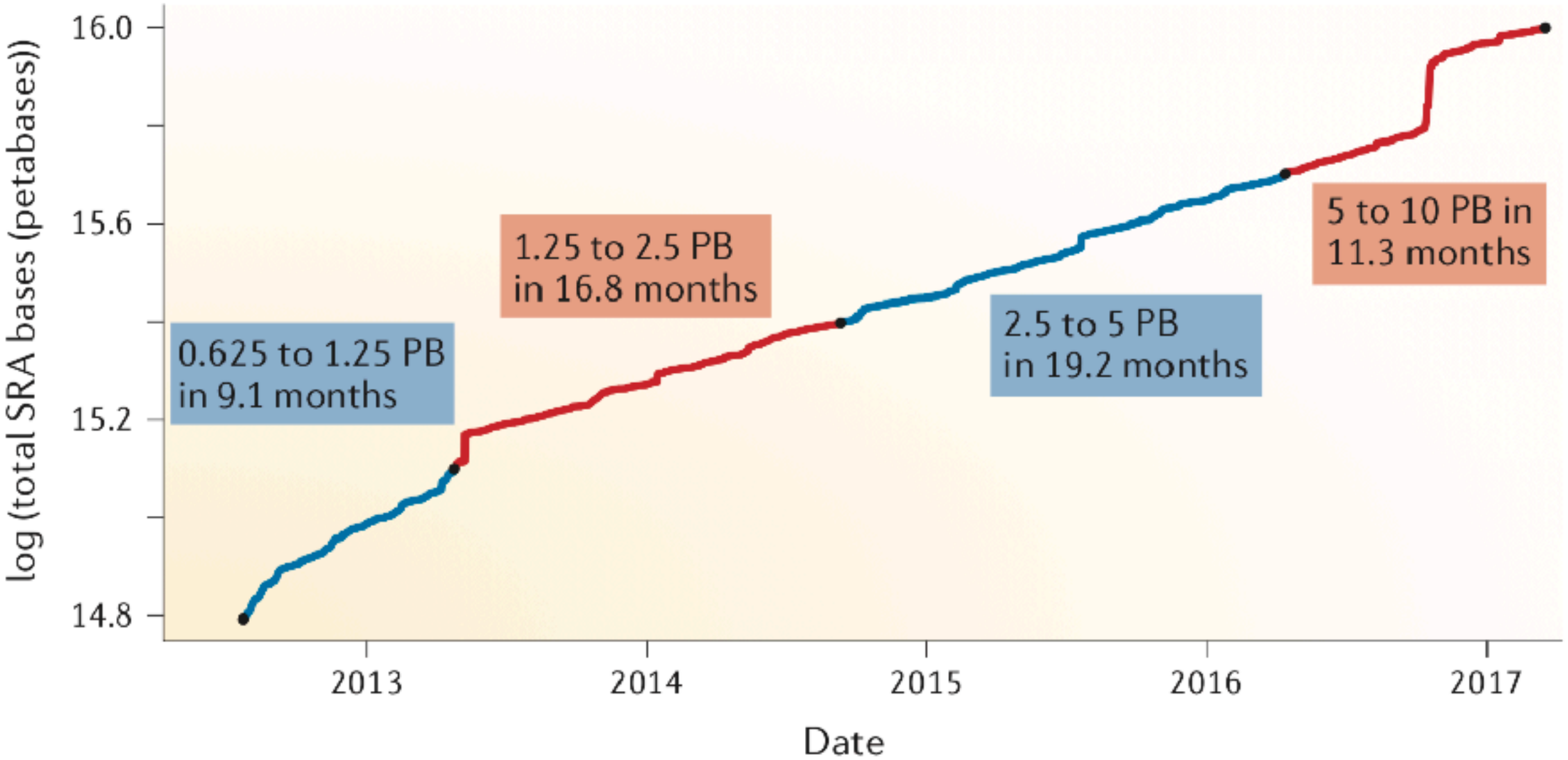
Scalable

Software: Rail-RNA, recount2, Snaptron
Topics: parallel & high-performance computing, cloud computing, indexing

Interpretable

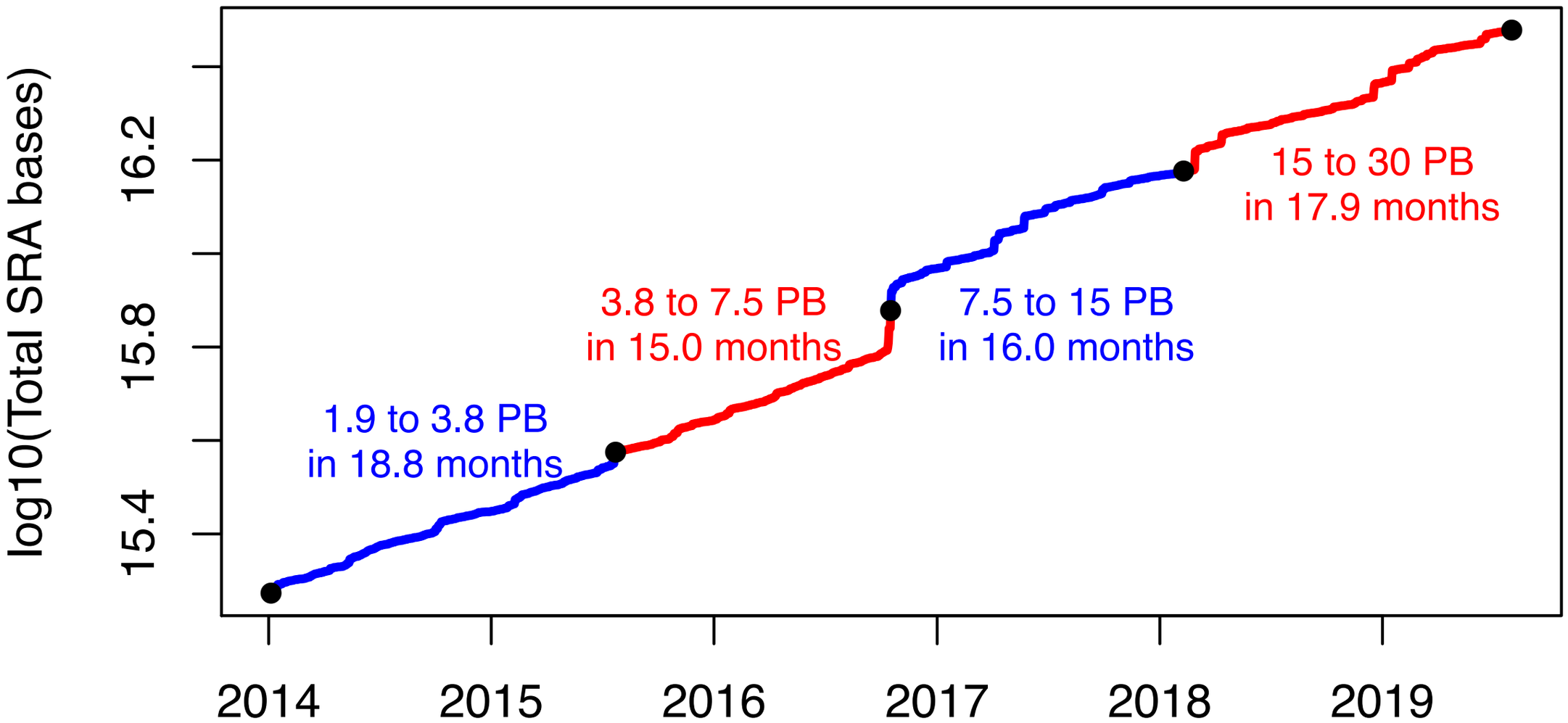
Software: Qtip, FORGe, r-index
Topics: modeling mapping quality, graph-genome variants, addressing biases

Sequence Read Archive



Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018 May;19(5):325.

Sequence Read Archive



Currently ~ 34 petabases

http://bit.ly/sra-growth

An index is a great leveler

GB Shaw

An index is a great leveler

GB Shaw

Summaries are good too

Not GB Shaw

Public summaries of sequencing data

Table 5 | **Summarized data sets, services and resources**

Name	Website	Notes
ArrayExpress[95]	www.ebi.ac.uk/arrayexpress	Archives processed data from high-throughput functional genomics experiments
Beacon	beacon-network.org	Platform for sharing genetic mutations across web services called 'beacons'
Bravo	bravo.sph.umich.edu	TOPMed data browser for accessing alleles across over 60,000 whole genomes
Expression Atlas[121]	www.ebi.ac.uk/gxa	Gene expression information across 3,000 transcriptomic experiments from ArrayExpress
PCAWG	docs.icgc.org/pcawg	Called germline and somatic variants, including structural variants, from over 5,600 tumour and normal samples across ICGC projects
recount2[61]	jhubiostatistics.shinyapps.io/recount	Web and R/Bioconductor resource for accessing genome coverage data from over 70,000 archived human RNA-seq samples, including publicly available SRA, TCGA and GTEx samples
RNASeq-er[93]	www.ebi.ac.uk/fg/rnaseq/api	Provides programmatic access to processed outputs for all archived publicly available RNA-seq samples
Snaptron	snaptron.cs.jhu.edu	Allows rapid querying of splice junctions, splicing patterns and metadata from recount2
Tatlow-Piccolo[68]	osf.io/gqrz9	Quantified transcripts across TCGA and CCLE
Toil[63]	xenabrowser.net/data-pages/?host=https://toil.xenahubs.net	Processed outputs from over 20,000 RNA-seq samples including TCGA and GTEx
Xena[94]	xena.ucsc.edu	Visualizes investigators' new functional genomics data next to publicly available data

CCLE, Cancer Cell Line Encyclopedia; GTEx, Genotype-Tissue Expression Project; ICGC, International Cancer Genome Consortium; PCAWG, Pan-Cancer Analysis of Whole Genomes; RNA-seq, RNA sequencing; SRA, Sequence Read Archive; TCGA, The Cancer Genome Atlas; TOPMed, Trans-Omics for Precision Medicine.

Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018 May;19(5):325.

Search engine for RNA-seq

Index & query engine w/ REST API
snaptron.cs.jhu.edu
doi:10.1093/bioinformatics/btx547

Clean summaries of data, metadata, packaged as R objects
jhubiostatistics.shinyapps.io/recount/
doi:10.1038/nbt.3838

Scalable, cloud-based spliced alignment of archived RNA-seq datasets
rail.bio
doi:10.1093/bioinformatics/btw575

Search engine for RNA-seq

Index & query engine w/ REST API
snaptron.cs.jhu.edu
doi:10.1093/bioinformatics/btx547

Index summaries

Clean summaries of data, metadata, packaged as R objects
jhubiostatistics.shinyapps.io/recount/
doi:10.1038/nbt.3838

Summarize

Scalable, cloud-based spliced alignment of archived RNA-seq datasets
rail.bio
doi:10.1093/bioinformatics/btw575

Reads

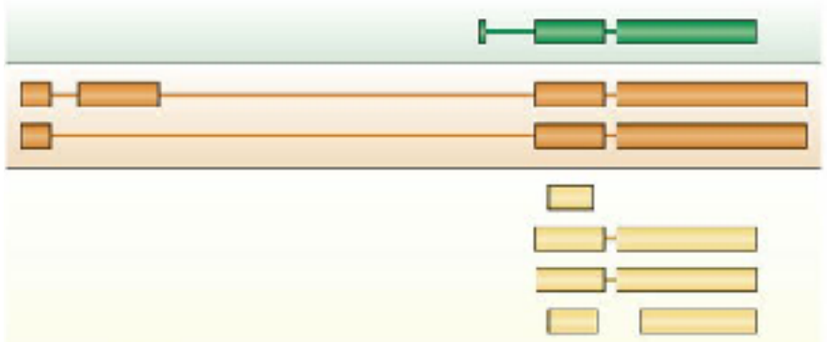
Themes

- Clouds & grids are natural fits for public data

Themes

- Clouds & grids are natural fits for public data

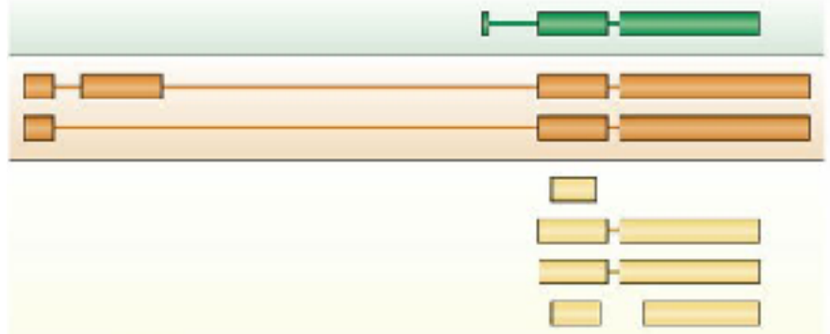
- Think outside the gene annotation



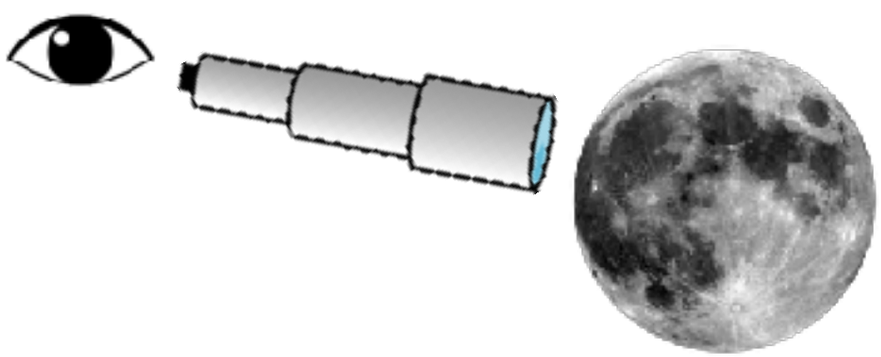
Themes

- Clouds & grids are natural fits for public data

- Think outside the gene annotation



- Much of the work is in the "last mile"



Rail-RNA

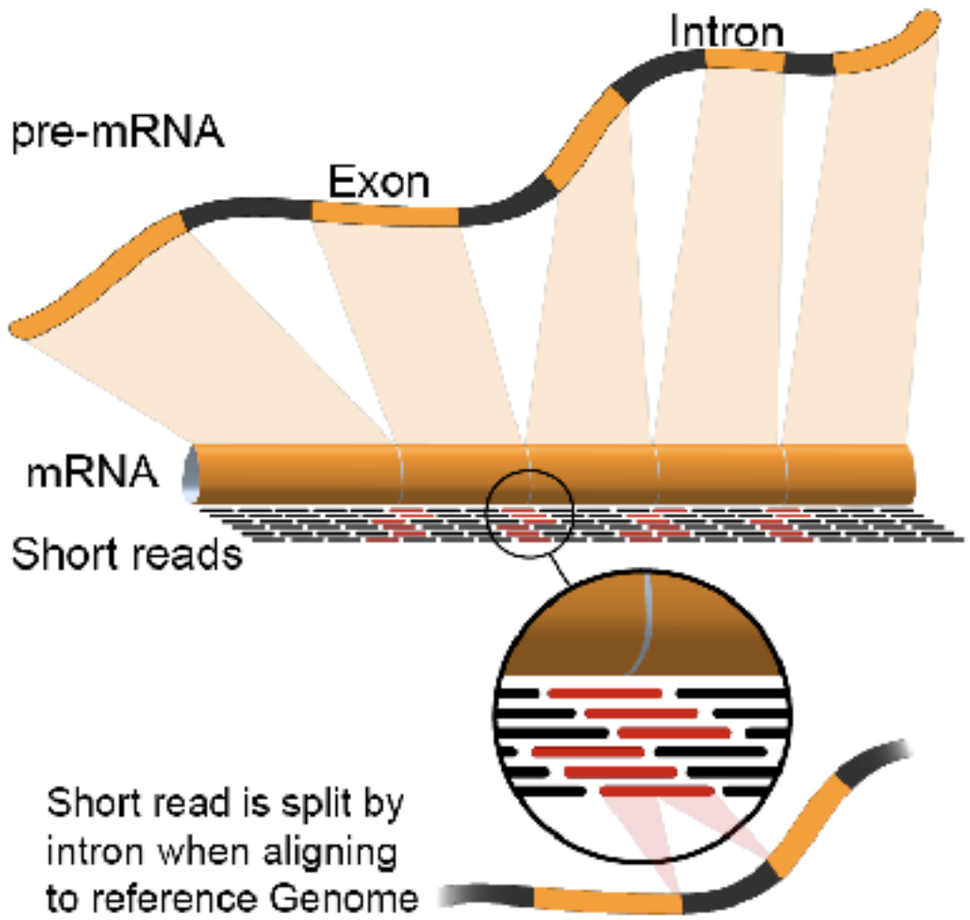


Image by Rgocs

Abhinav Nellore
OHSU

Jeff Leek,
JHU

http://rail.bio

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 33(24):4033–4040, Dec 2017

Spliced RNA-seq aligner for analyzing many samples at once

http://rail.bio

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 33(24):4033–4040, Dec 2017

Spliced RNA-seq aligner for analyzing many samples at once

- Group across samples to *borrow strength* and *eliminate redundant alignment* work

http://rail.bio

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 33(24):4033–4040, Dec 2017

Spliced RNA-seq aligner for analyzing many samples at once

- Group across samples to *borrow strength* and *eliminate redundant alignment* work
- Let data prune false junction calls, *not annotation*

http://rail.bio

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 33(24):4033–4040, Dec 2017

Rail-RNA

Spliced RNA-seq aligner for analyzing many samples at once

- Group across samples to *borrow strength* and *eliminate redundant alignment* work
- Let data prune false junction calls, *not annotation*
- *Concise outputs*: junctions & coverage vectors; *no alignments, unless asked for*

http://rail.bio

Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 33(24):4033–4040, Dec 2017

Spliced RNA-seq aligner for analyzing many samples at once

- Group across samples to *borrow strength* and *eliminate redundant alignment* work
- Let data prune false junction calls, *not annotation*
- *Concise outputs*: junctions & coverage vectors; *no alignments, unless asked for*
- Runs easily on *commercial AWS cloud*

http://rail.bio

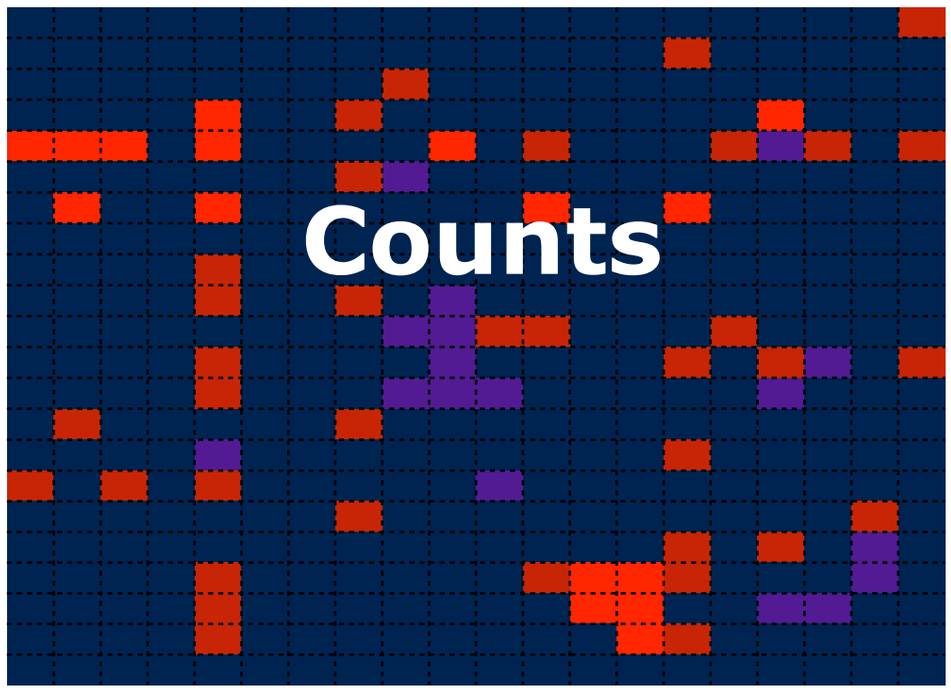
Nellore A, Collado-Torres L, Jaffe AE, Alquicira-Hernández J, Wilks C, Pritt J, Morton J, Leek JT, Langmead B. Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*. 33(24):4033–4040, Dec 2017

Intropolis

- Analyzed **~21,500** human RNA-seq samples with Rail-RNA; about **62 Tbp**

Samples (21.5K)

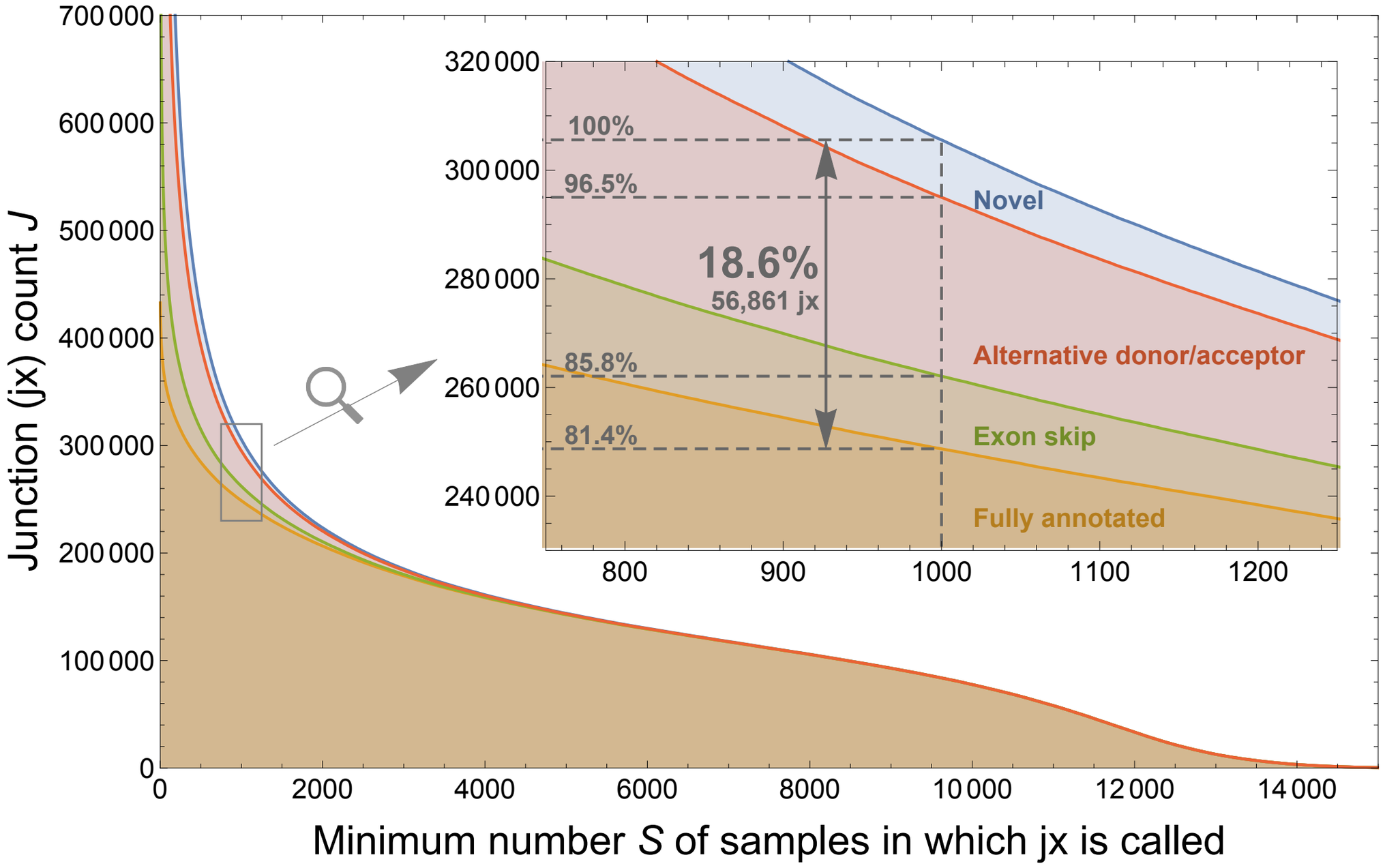
Exon-exon junctions (10s of millions)



http://intropolis.rail.bio

Nellore A, et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 2016 Dec 30;17(1):266.

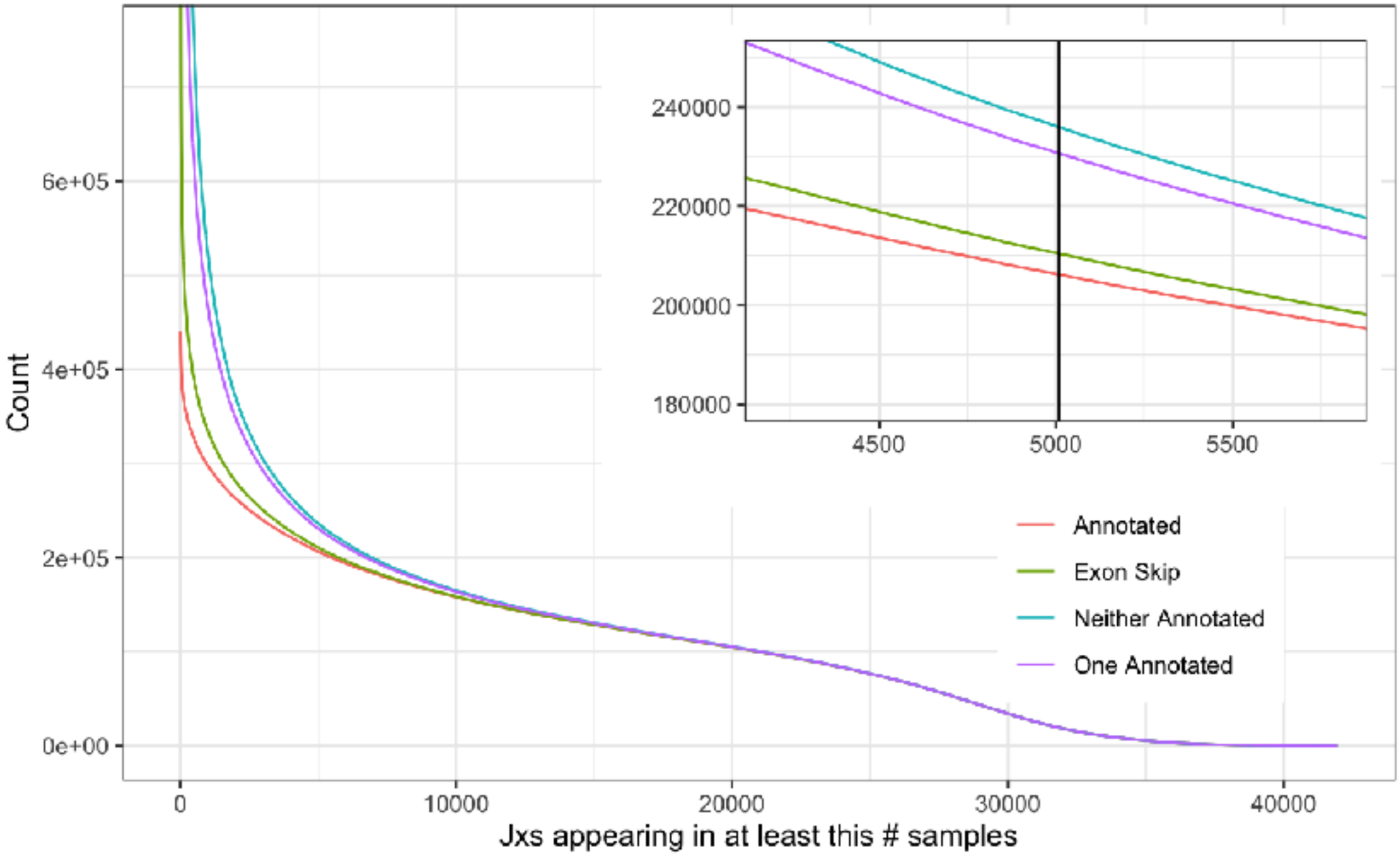
Annotations: UCSC, GENCODE v19 & v24, RefSeq, CCDS, MGC, lincRNAs, SIB genes, AceView, Vega



http://intropolis.rail.bio

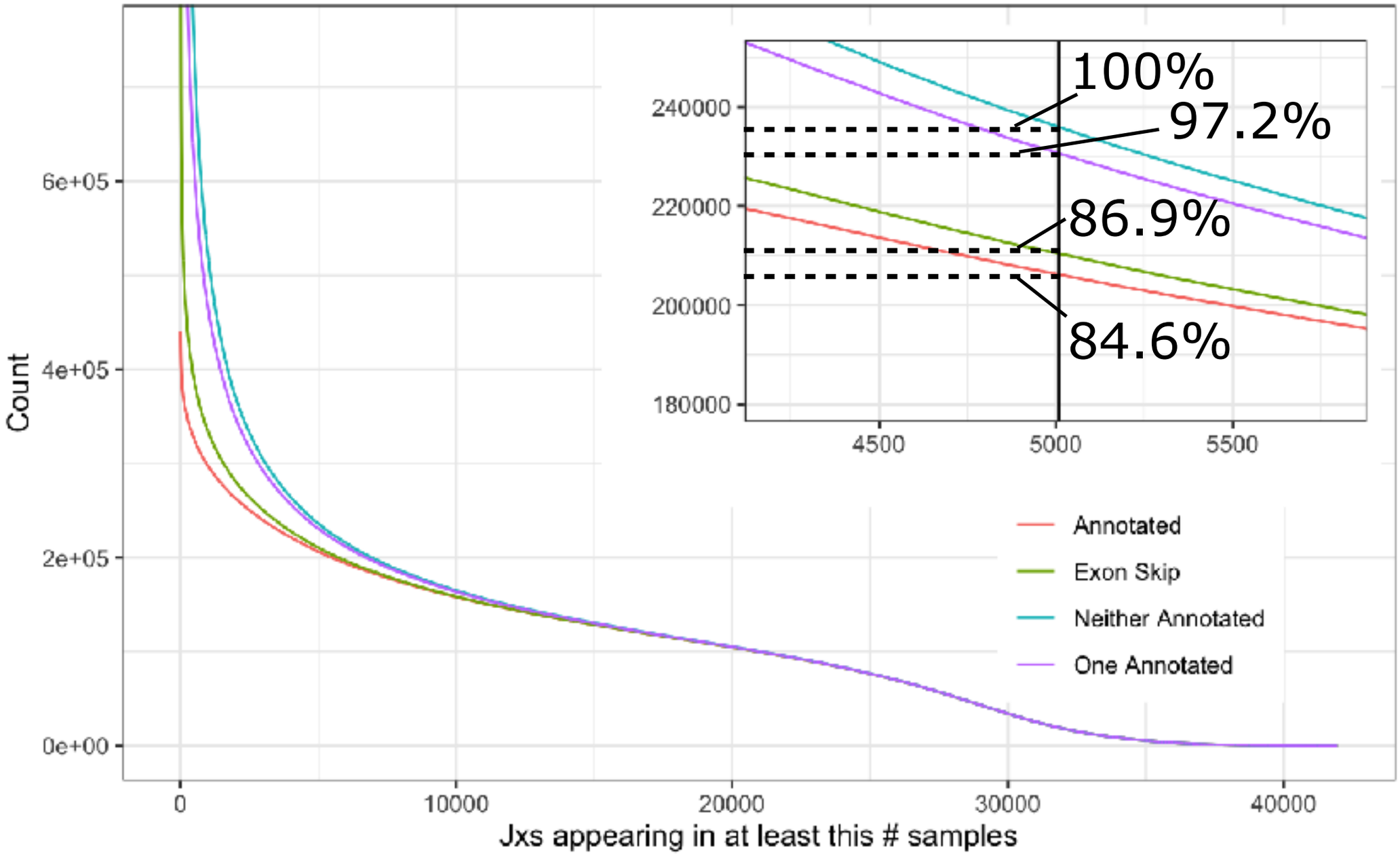
Nellore A, et al. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 2016 Dec 30;17(1):266.

With >50K samples and comparing to GENCODE v24



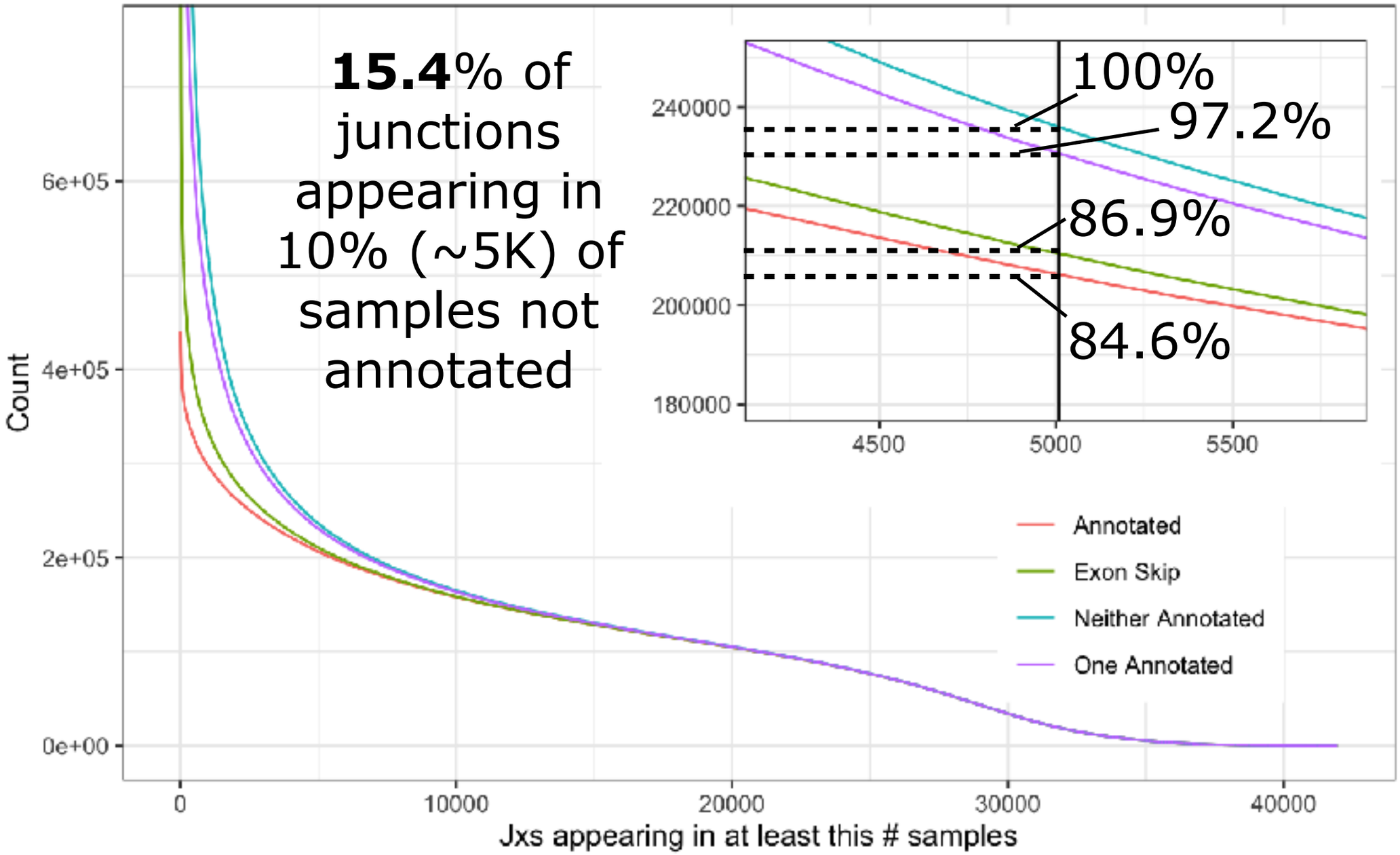
https://github.com/BenLangmead/cgsi18 (jx1.Rmd)

With >50K samples and comparing to GENCODE v24



https://github.com/BenLangmead/cgsi18 (jx1.Rmd)

With >50K samples and comparing to GENCODE v24



https://github.com/BenLangmead/cgsi18 (jx1.Rmd)

recount2

- **>50K** human RNA-seq samples from SRA (open)
- **>10K** human RNA-seq samples from TCGA (dbGaP)

Image: https://www.sevenbridges.com/welcome-to-the-cancer-genomics-cloud-2/

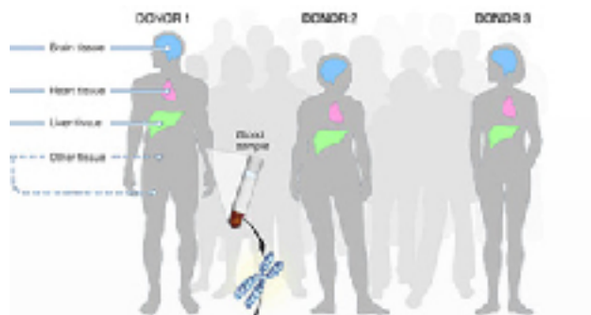


Image: doi:10.1038/ng.2653

Kasper Hansen

- **>10K** human RNA-seq from GTEx (dbGaP)
- Total: ~4.4 trillion reads, 100s of terabases

Abhinav Nellore

Leo Collado Torres

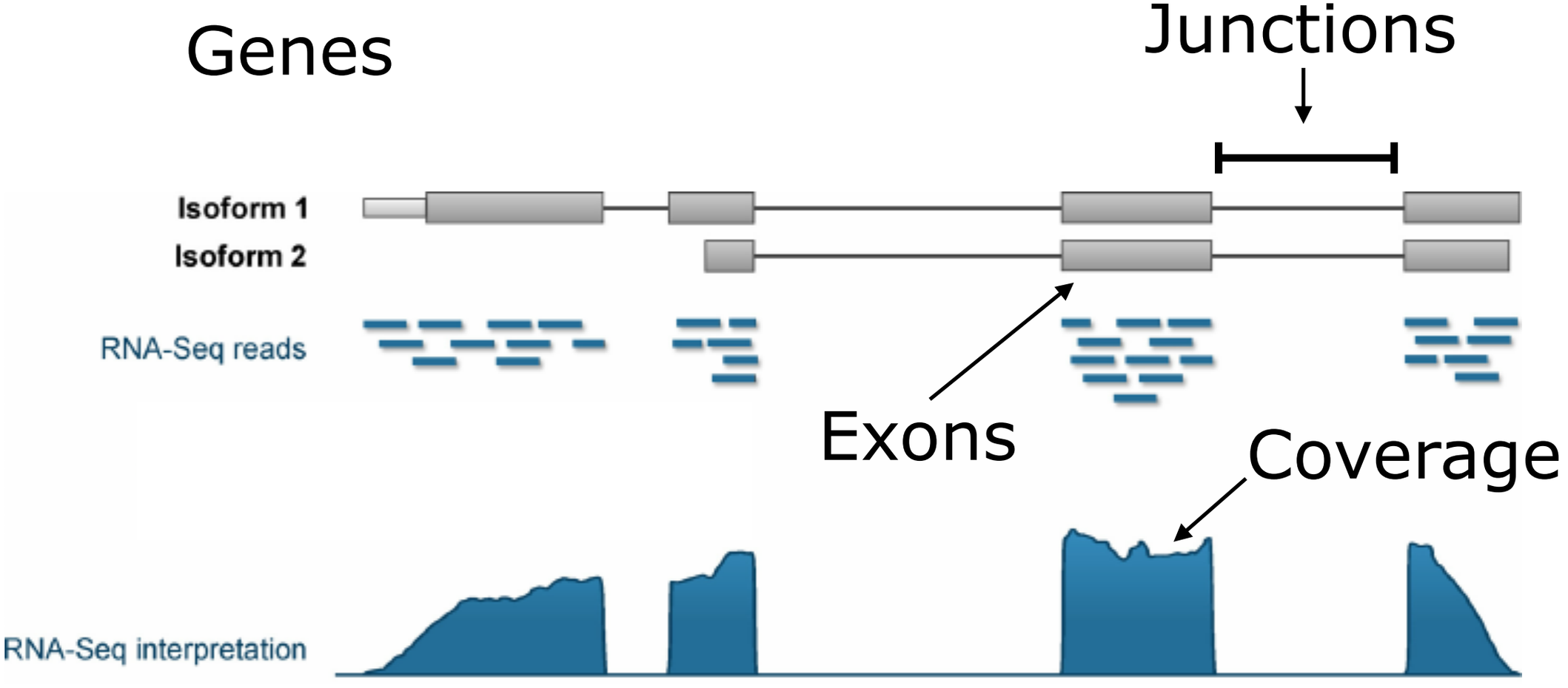
Jeff Leek

Andrew Jaffe

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 Apr 11;35(4):319-321.

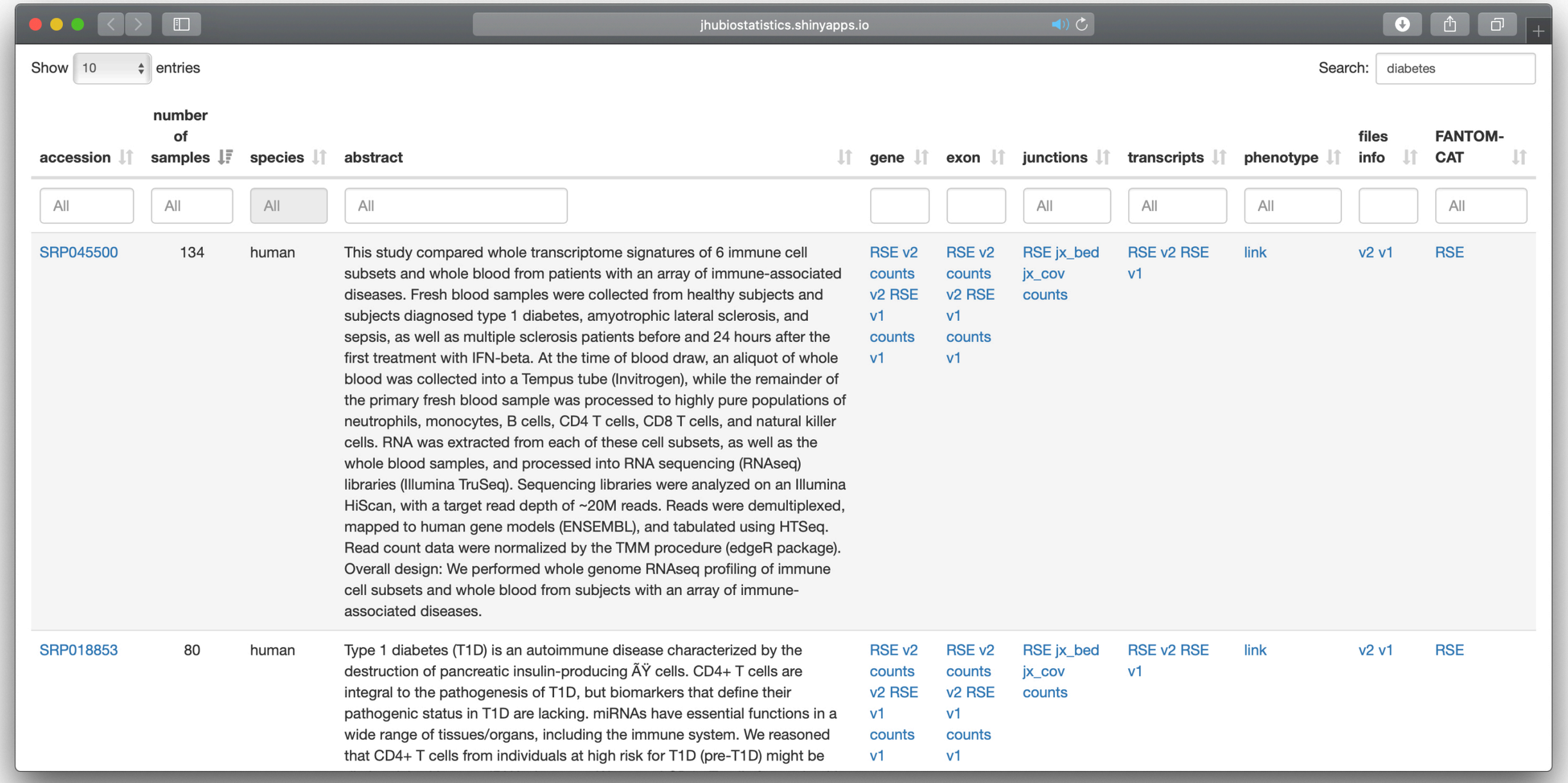
recount2

Summarized at levels of **genes**, **exons, junctions**, and **coverage vectors** + more



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 Apr 11;35(4):319-321.

recount2

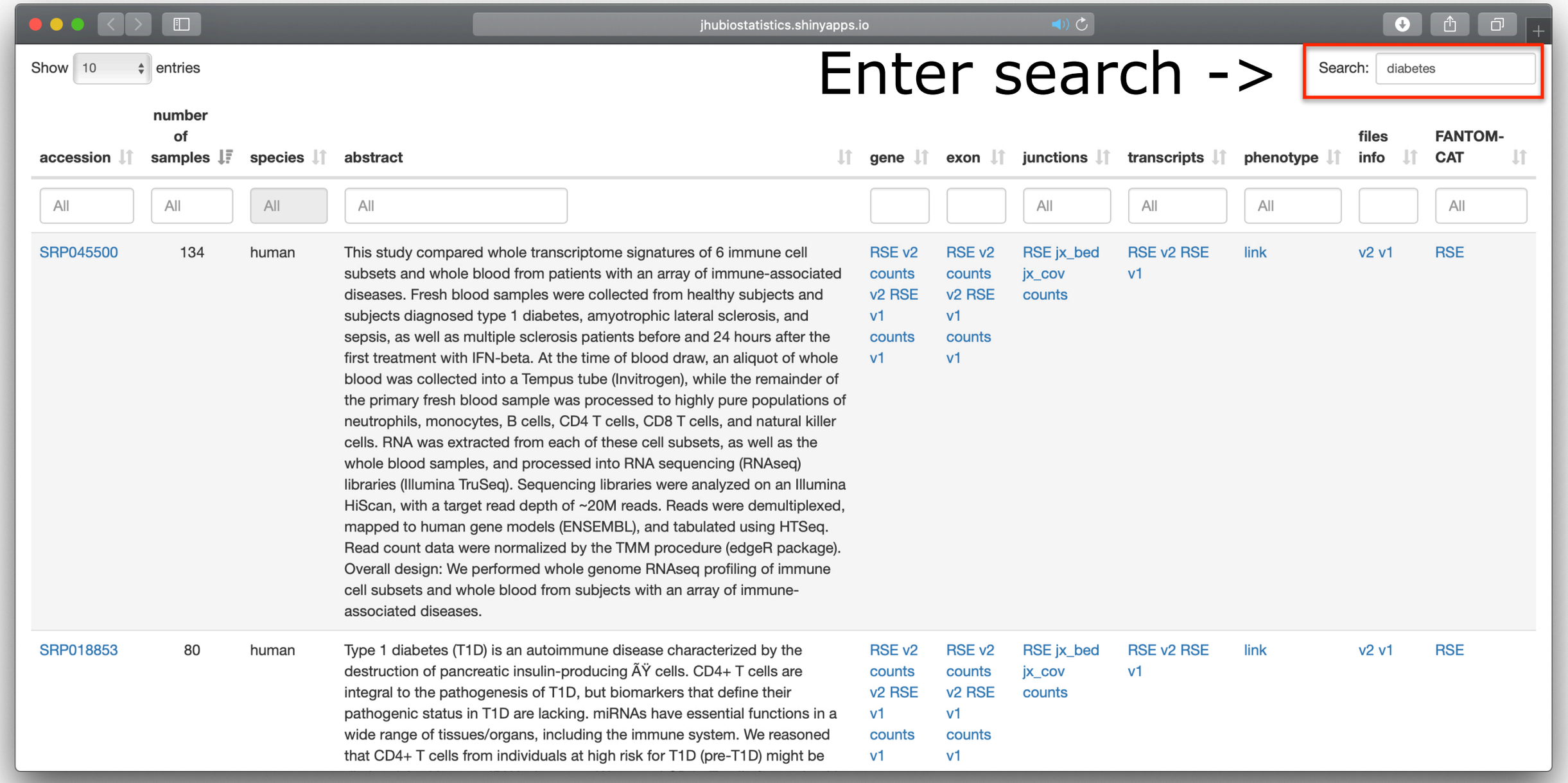


accession	number of samples	species	abstract	gene	exon	junctions	transcripts	phenotype	files info	FANTOM-CAT
All	All	All	All			All	All	All		All
SRP045500	134	human	This study compared whole transcriptome signatures of 6 immune cell subsets and whole blood from patients with an array of immune-associated diseases. Fresh blood samples were collected from healthy subjects and subjects diagnosed type 1 diabetes, amyotrophic lateral sclerosis, and sepsis, as well as multiple sclerosis patients before and 24 hours after the first treatment with IFN-beta. At the time of blood draw, an aliquot of whole blood was collected into a Tempus tube (Invitrogen), while the remainder of the primary fresh blood sample was processed to highly pure populations of neutrophils, monocytes, B cells, CD4 T cells, CD8 T cells, and natural killer cells. RNA was extracted from each of these cell subsets, as well as the whole blood samples, and processed into RNA sequencing (RNAseq) libraries (Illumina TruSeq). Sequencing libraries were analyzed on an Illumina HiScan, with a target read depth of ~20M reads. Reads were demultiplexed, mapped to human gene models (ENSEMBL), and tabulated using HTSeq. Read count data were normalized by the TMM procedure (edgeR package). Overall design: We performed whole genome RNAseq profiling of immune cell subsets and whole blood from subjects with an array of immune-associated diseases.	RSE v2 counts v2 RSE v1 counts v1	RSE v2 counts v2 RSE v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1	RSE
SRP018853	80	human	Type 1 diabetes (T1D) is an autoimmune disease characterized by the destruction of pancreatic insulin-producing ÃŸ cells. CD4+ T cells are integral to the pathogenesis of T1D, but biomarkers that define their pathogenic status in T1D are lacking. miRNAs have essential functions in a wide range of tissues/organs, including the immune system. We reasoned that CD4+ T cells from individuals at high risk for T1D (pre-T1D) might be	RSE v2 counts v2 RSE v1 counts v1	RSE v2 counts v2 RSE v1 counts v1	RSE jx_bed jx_cov counts	RSE v2 RSE v1	link	v2 v1	RSE

bit.ly/recount2 (jhubiostatistics.shinyapps.io/recount/)

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 Apr 11;35(4):319-321.

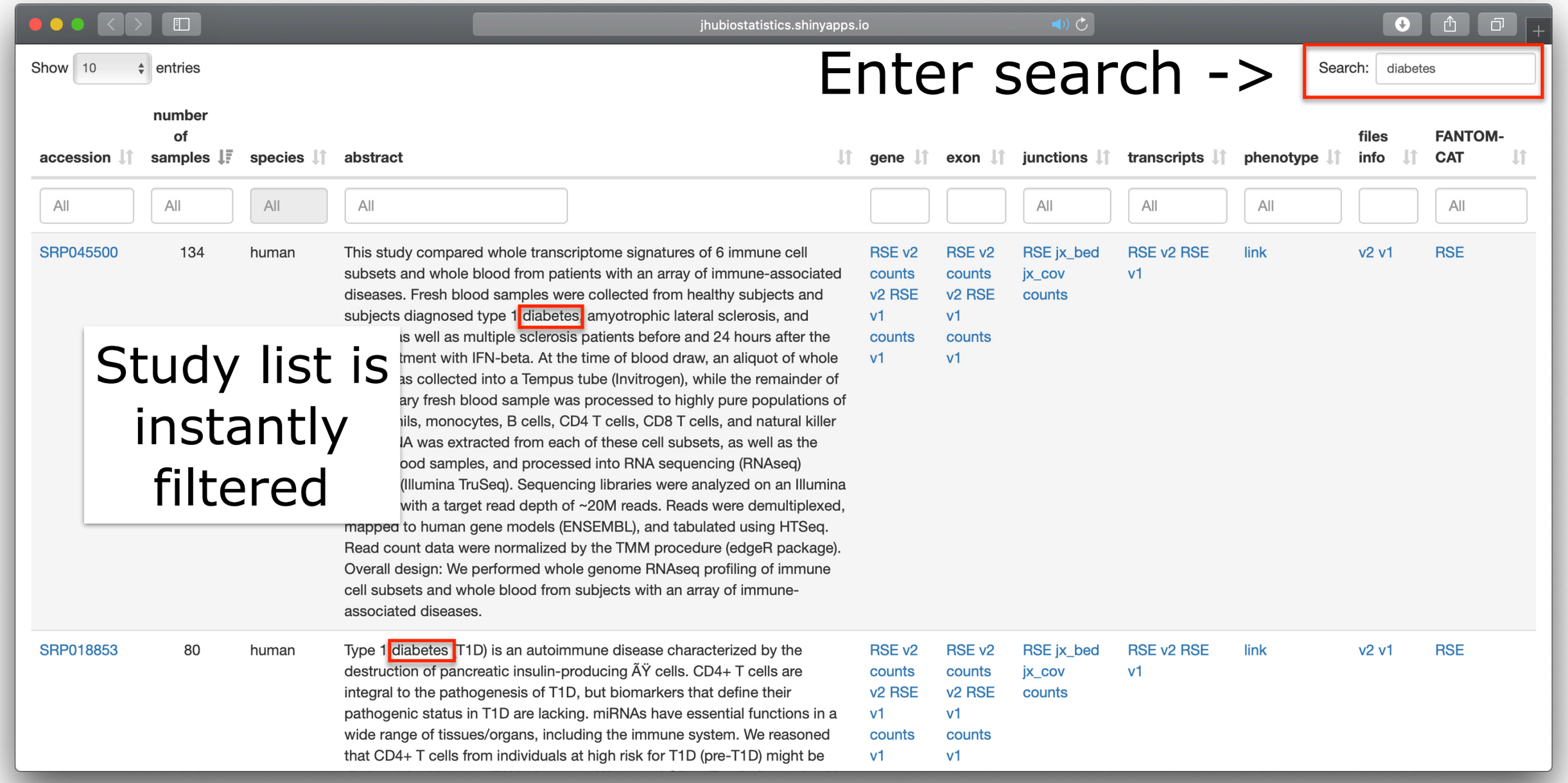
recount2



bit.ly/recount2 (jhubiostatistics.shinyapps.io/recount/)

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 Apr 11;35(4):319-321.

recount2

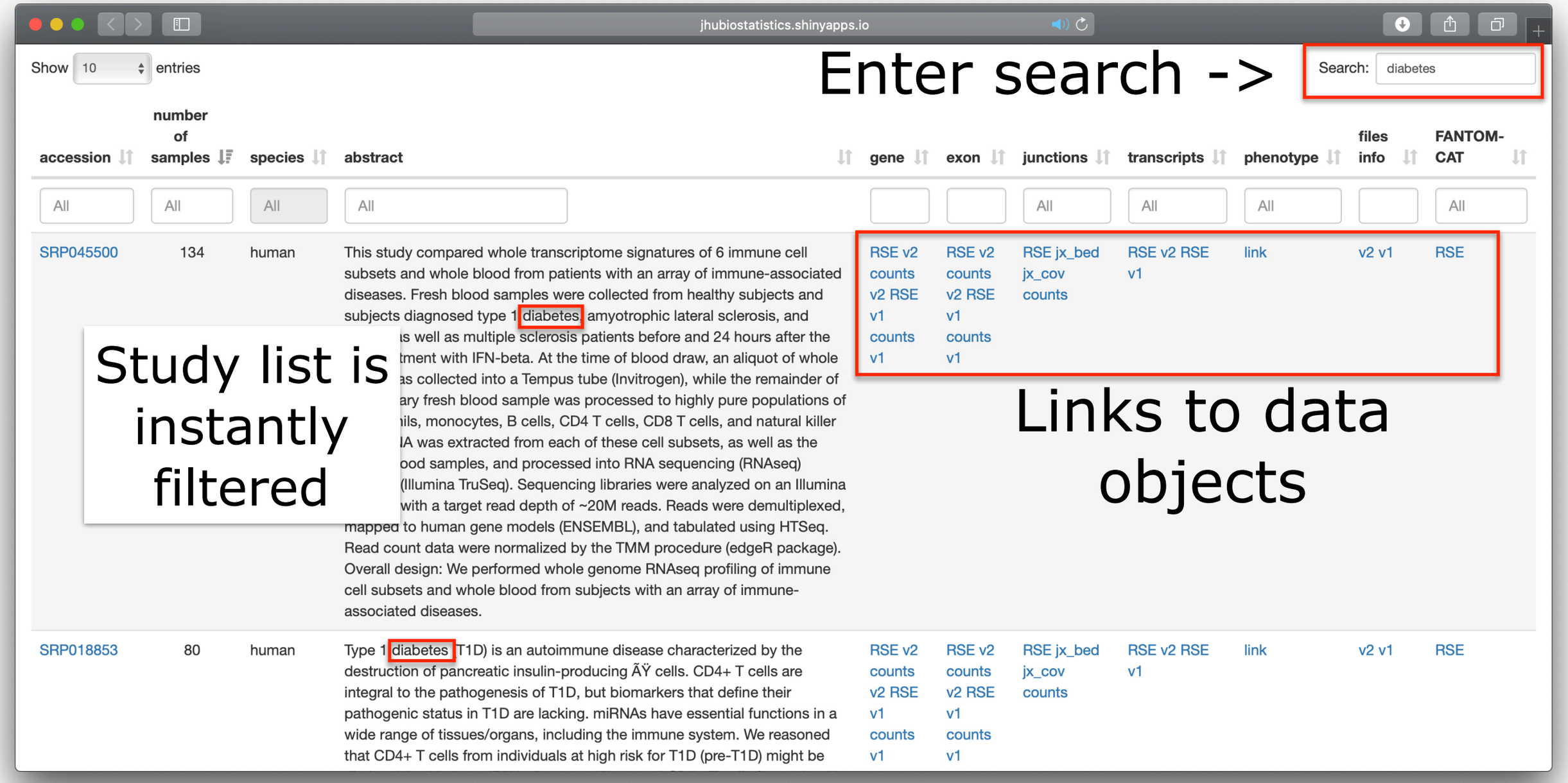


Study list is instantly filtered

bit.ly/recount2 (jhubiostatistics.shinyapps.io/recount/)

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 Apr 11;35(4):319-321.

recount2



Study list is instantly filtered

Links to data objects

bit.ly/recount2 (jhubiostatistics.shinyapps.io/recount/)

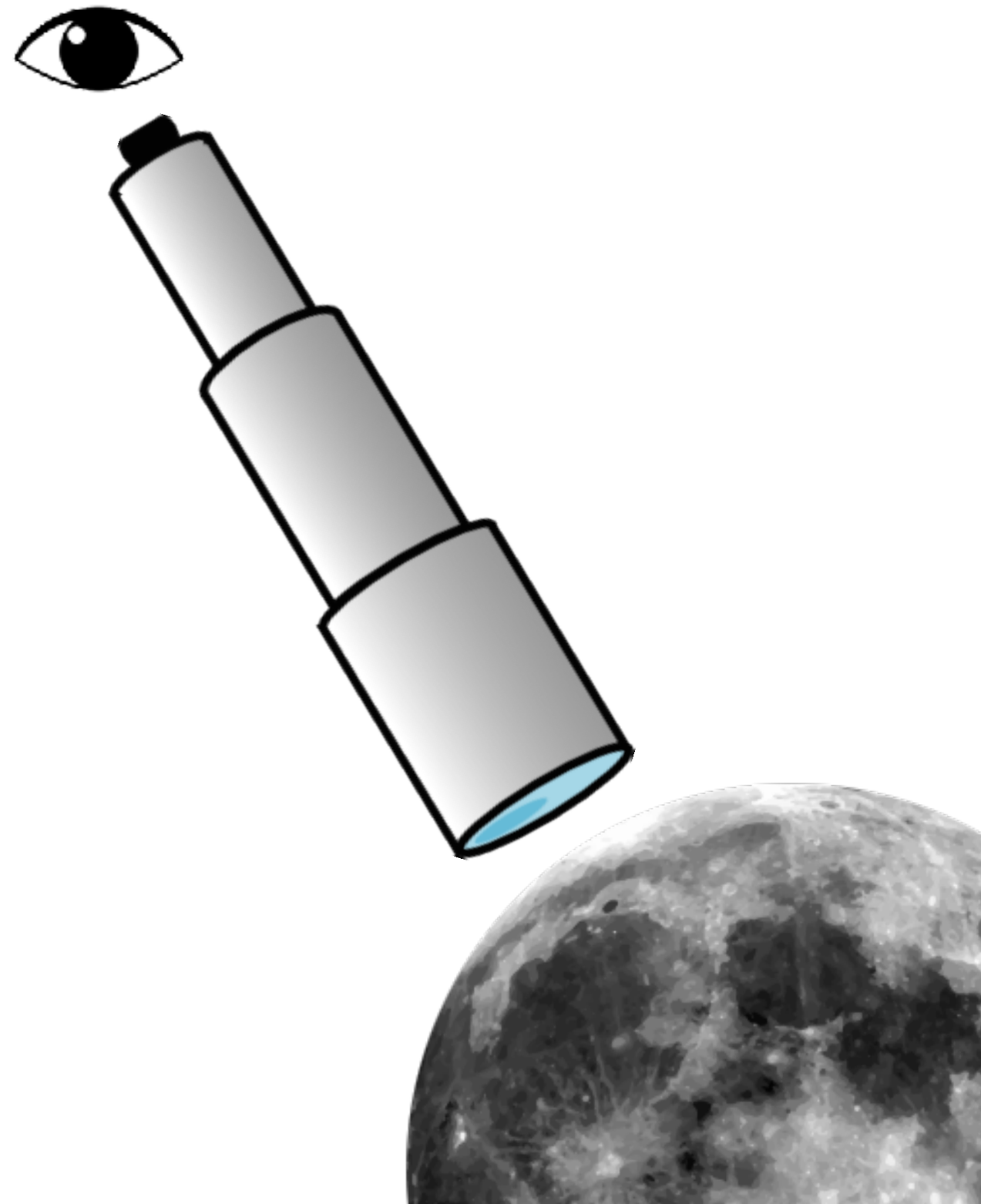
Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. Reproducible RNA-seq analysis using recount2. *Nature Biotechnology*. 2017 Apr 11;35(4):319-321.

Search engine for RNA-seq

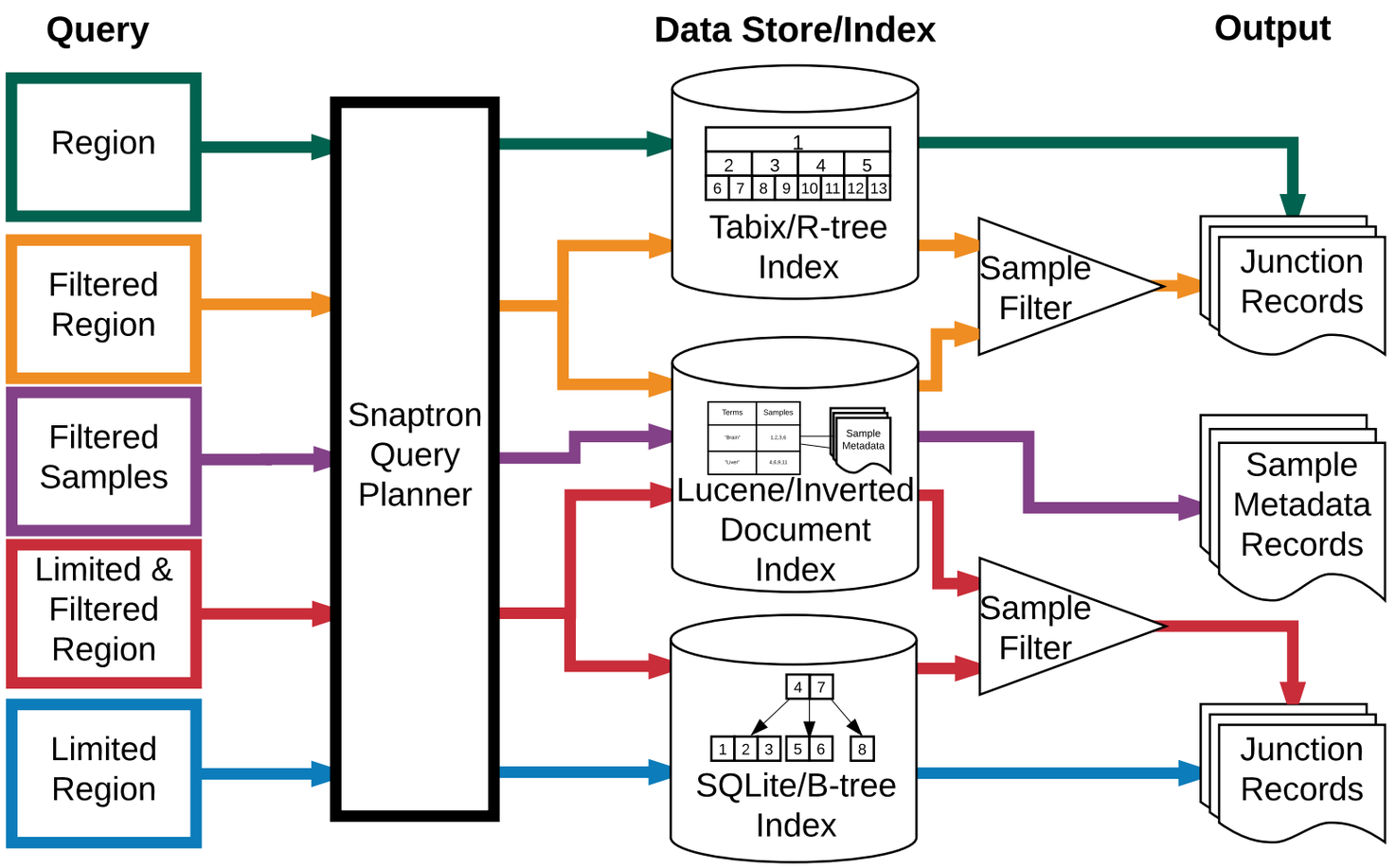
Snaptron

recount2

Rail-RNA



Snaptron

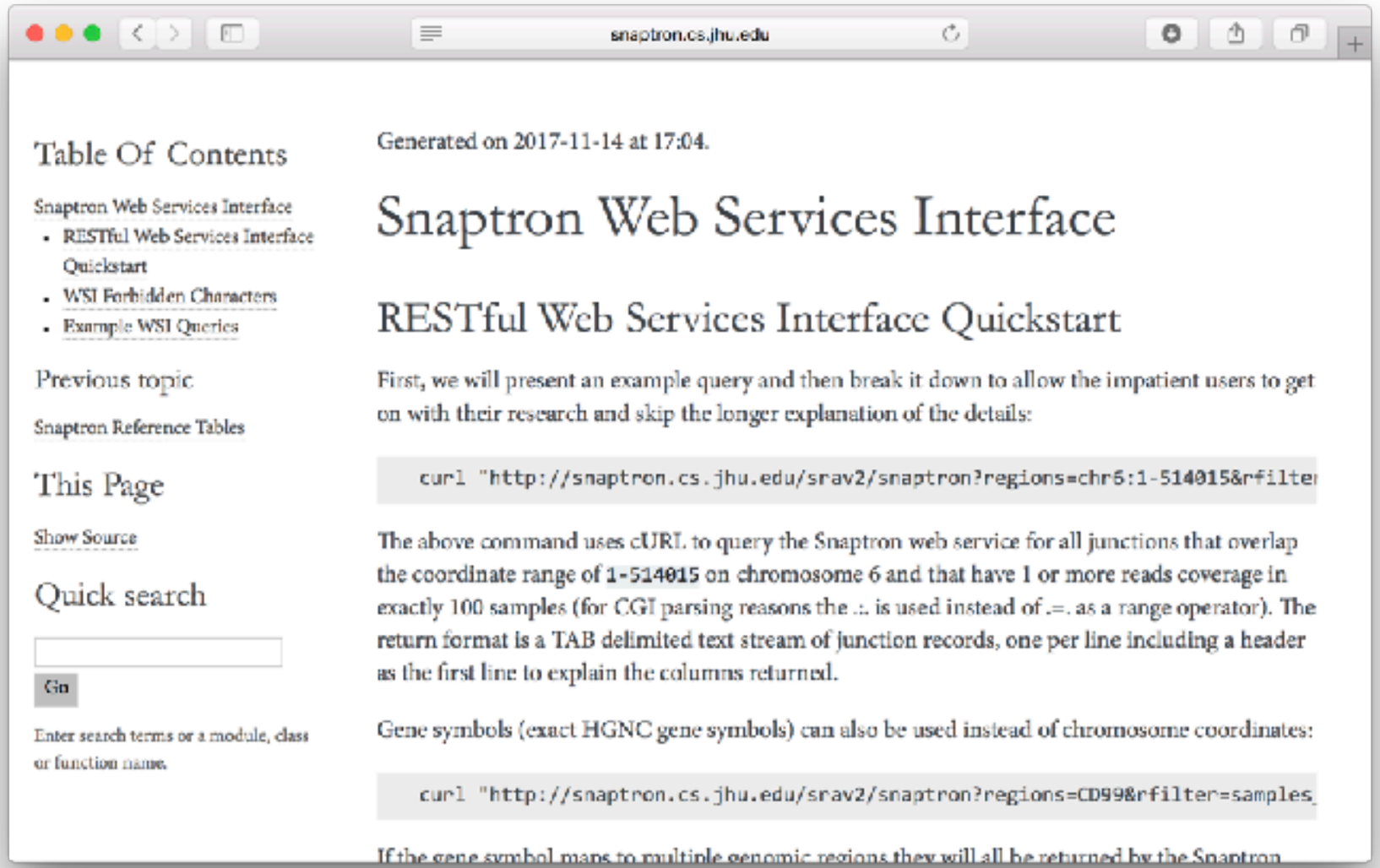


Chris Wilks

Query planner breaks down queries, delegates to appropriate systems (sqlite, tabix, Lucene) and indexes (R-tree, B-tree, inverted index)

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Snaptron



Command-line tool and REST API for querying junctions.
New: Also genes, exons and coverage vectors

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Snaptron

http://snaptron.cs.jhu.edu

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Snaptron

- For each junction in a gene, what is its read support in each of 50K SRA samples?

http://snaptron.cs.jhu.edu

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Snaptron

- For each junction in a gene, what is its read support in each of 50K SRA samples?

- What is a junction's *tissue specificity* in GTEx?

http://snaptron.cs.jhu.edu

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Snaptron

- For each junction in a gene, what is its read support in each of 50K SRA samples?
- What is a junction's *tissue specificity* in GTEx?
- In which samples is splicing pattern A overrepresented relative to pattern B?

http://snaptron.cs.jhu.edu

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Case study

- Goldstein *et al* searched for novel cassette exons in Illumina BodyMap 2.0 RNA-seq

PLOS | ONE

Prediction and Quantification of Splice Events from RNA-Seq Data

Leonard D. Goldstein[1,2]*, Yi Cao[1], Gregoire Pau[1], Michael Lawrence[1], Thomas D. Wu[1], Somasekar Seshagiri[2], Robert Gentleman[1¤]

1 Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA, United States of America, **2** Department of Molecular Biology, Genentech Inc., South San Francisco, CA, United States of America

Case study

- Goldstein *et al* searched for novel cassette exons in Illumina BodyMap 2.0 RNA-seq
- Identified 249 within known genes, not overlapping a RefSeq-annotated exon

PLOS | ONE

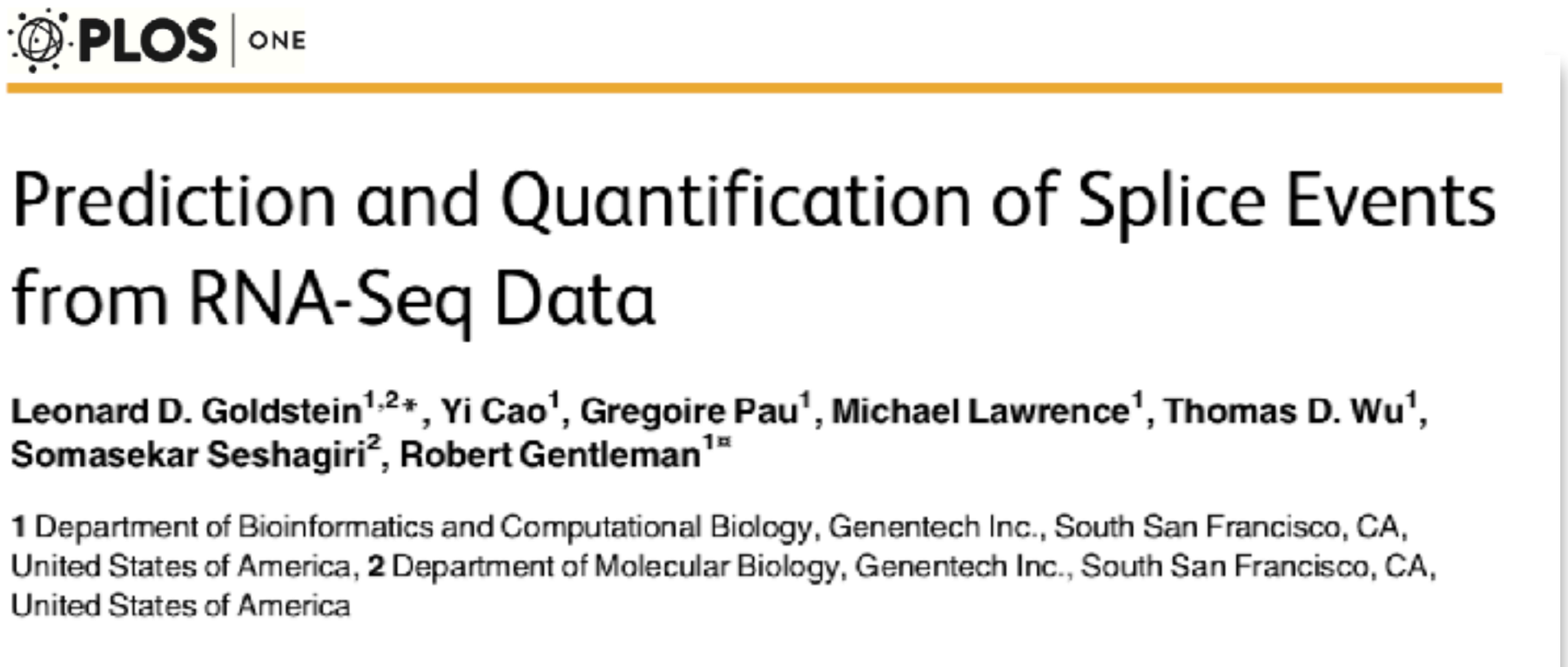
Prediction and Quantification of Splice Events from RNA-Seq Data

Leonard D. Goldstein[1,2]*, Yi Cao[1], Gregoire Pau[1], Michael Lawrence[1], Thomas D. Wu[1], Somasekar Seshagiri[2], Robert Gentleman[1¤]

1 Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA, United States of America, **2** Department of Molecular Biology, Genentech Inc., South San Francisco, CA, United States of America

Case study

- Goldstein *et al* searched for novel cassette exons in Illumina BodyMap 2.0 RNA-seq
- Identified 249 within known genes, not overlapping a RefSeq-annotated exon
- Validated 216 out of 249 in independent sample via RNA-seq



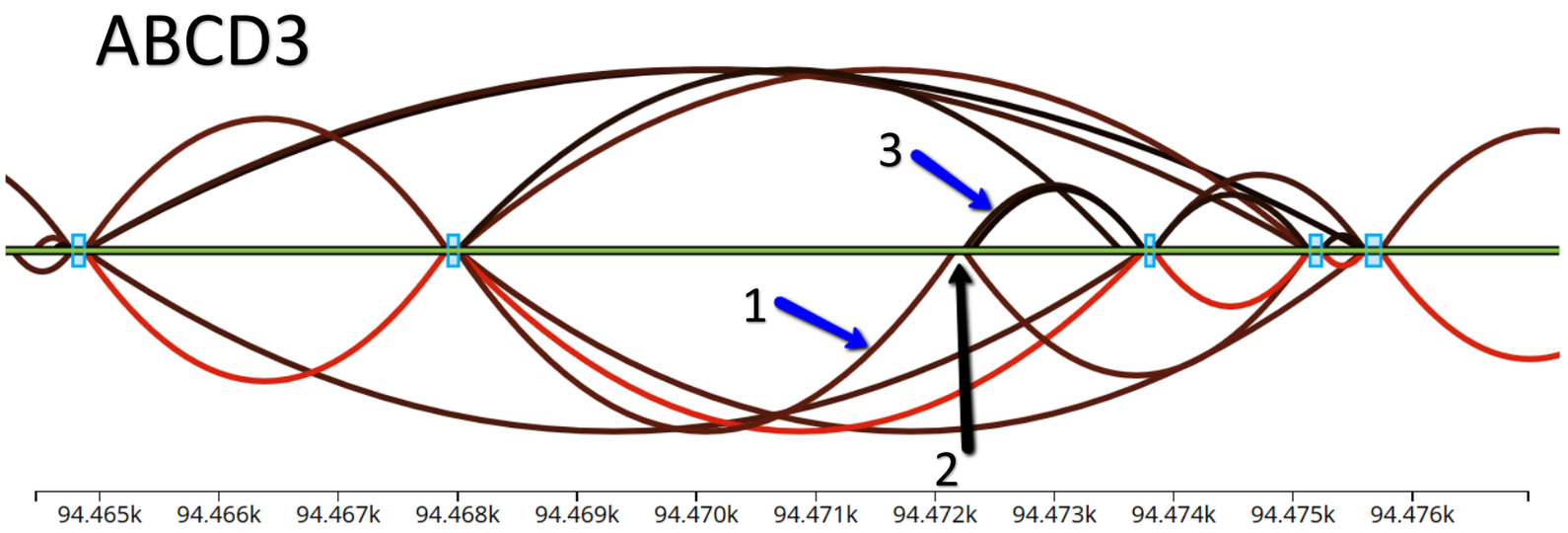
PLOS ONE

Prediction and Quantification of Splice Events from RNA-Seq Data

Leonard D. Goldstein[1,2]*, Yi Cao[1], Gregoire Pau[1], Michael Lawrence[1], Thomas D. Wu[1], Somasekar Seshagiri[2], Robert Gentleman[1¤]

1 Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA, United States of America, **2** Department of Molecular Biology, Genentech Inc., South San Francisco, CA, United States of America

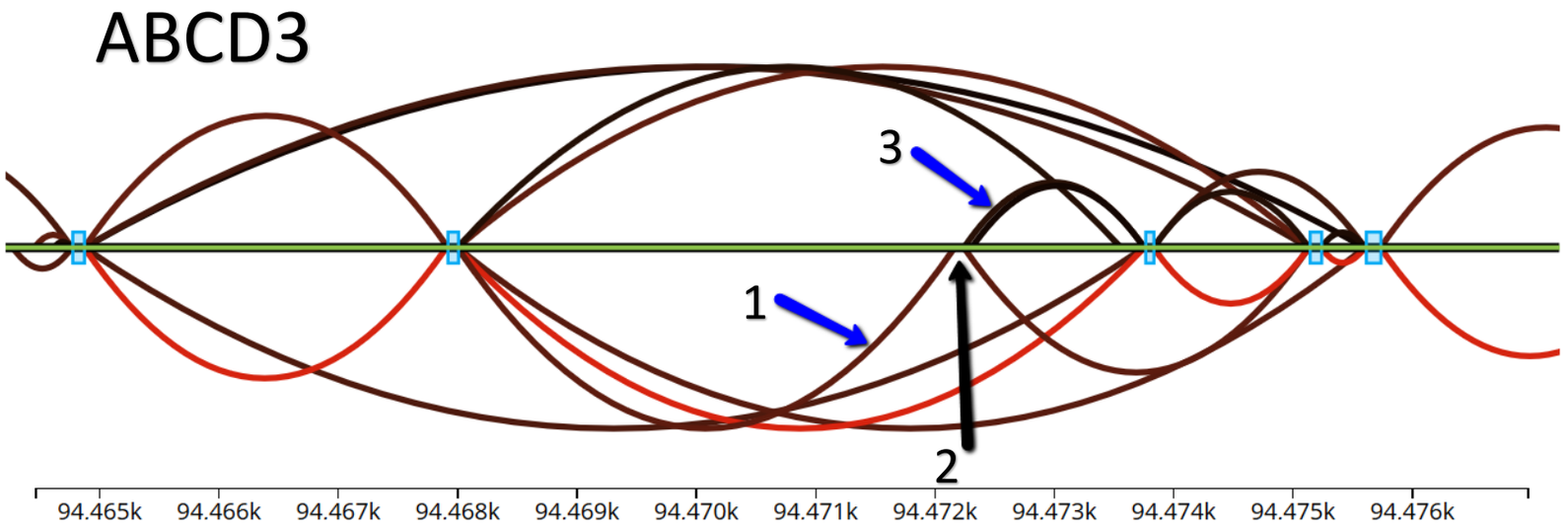
Case study



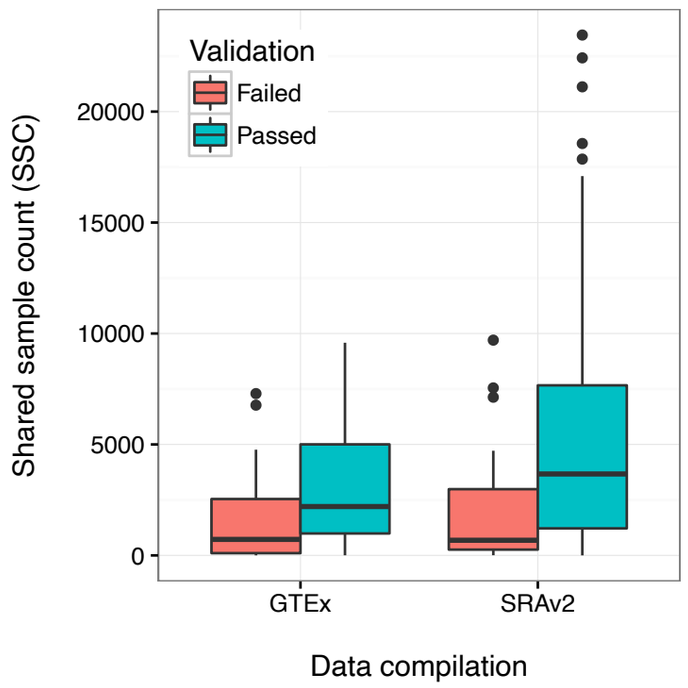
- Of the 249 novel exons, 236 (94.8%) occurred in GTEx (one shown above)

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Case study

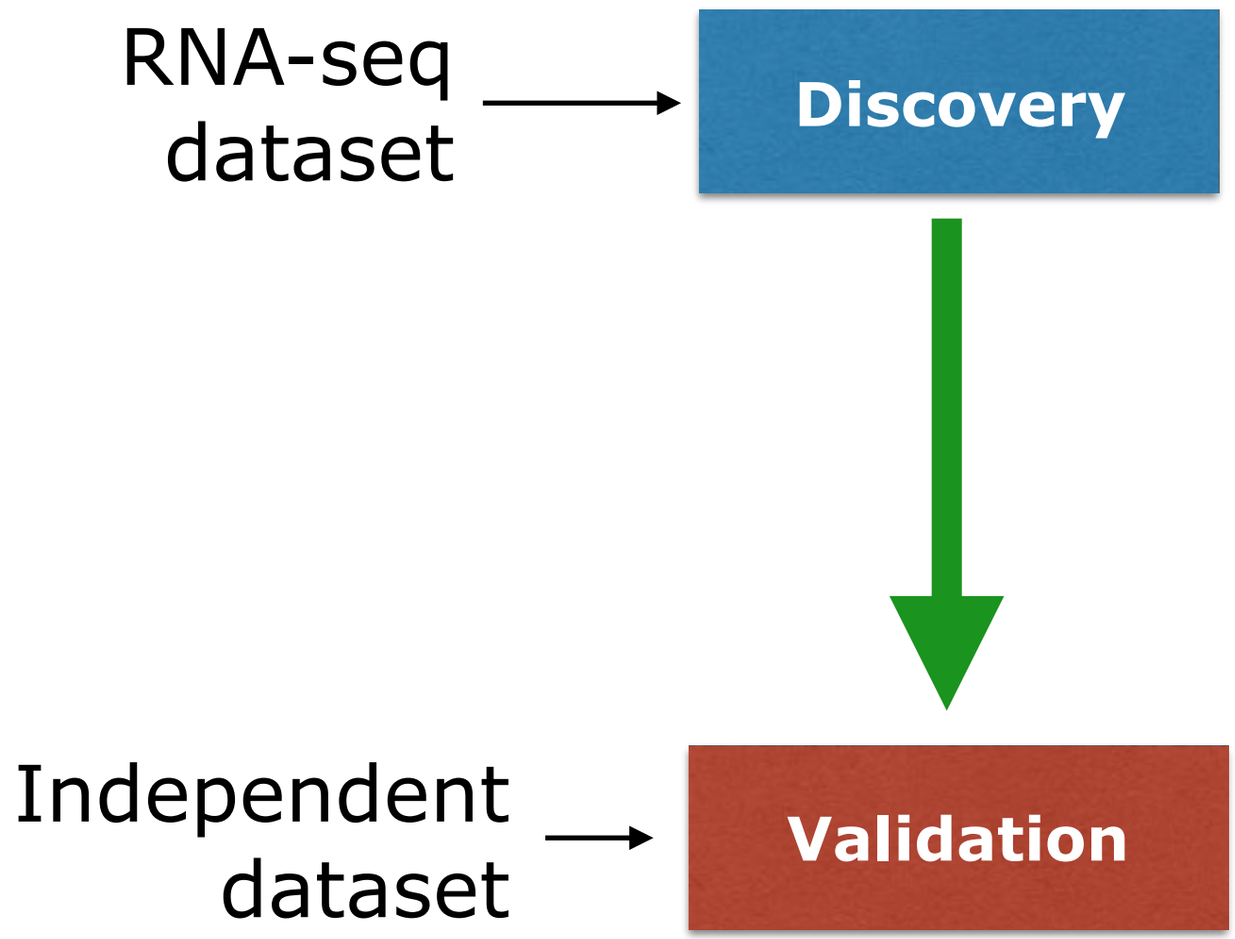


- Of the 249 novel exons, 236 (94.8%) occurred in GTEx (one shown above)
- Shared sample count predicts how likely novel exons were to validate (right)

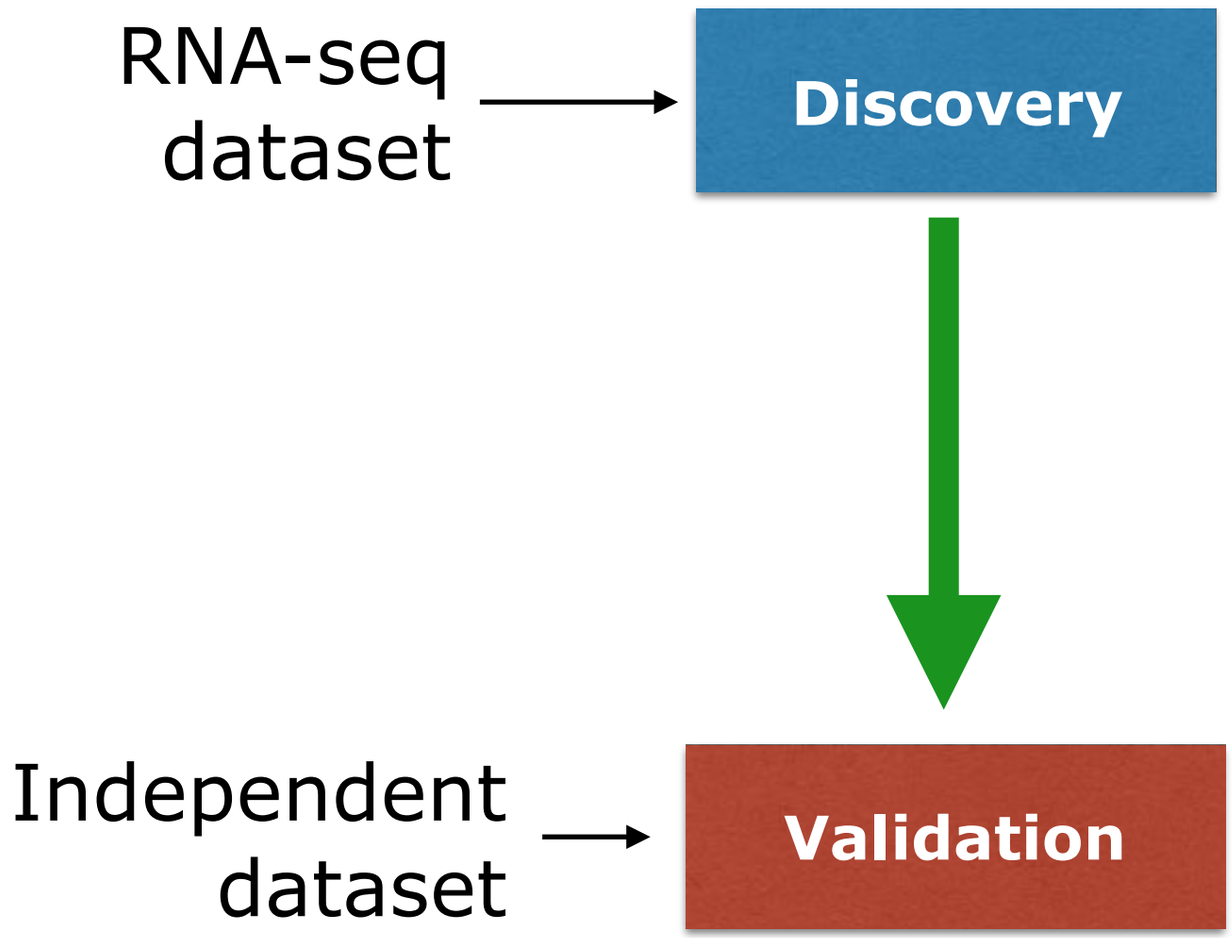


Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2018 34(1), 114–116.

Case study

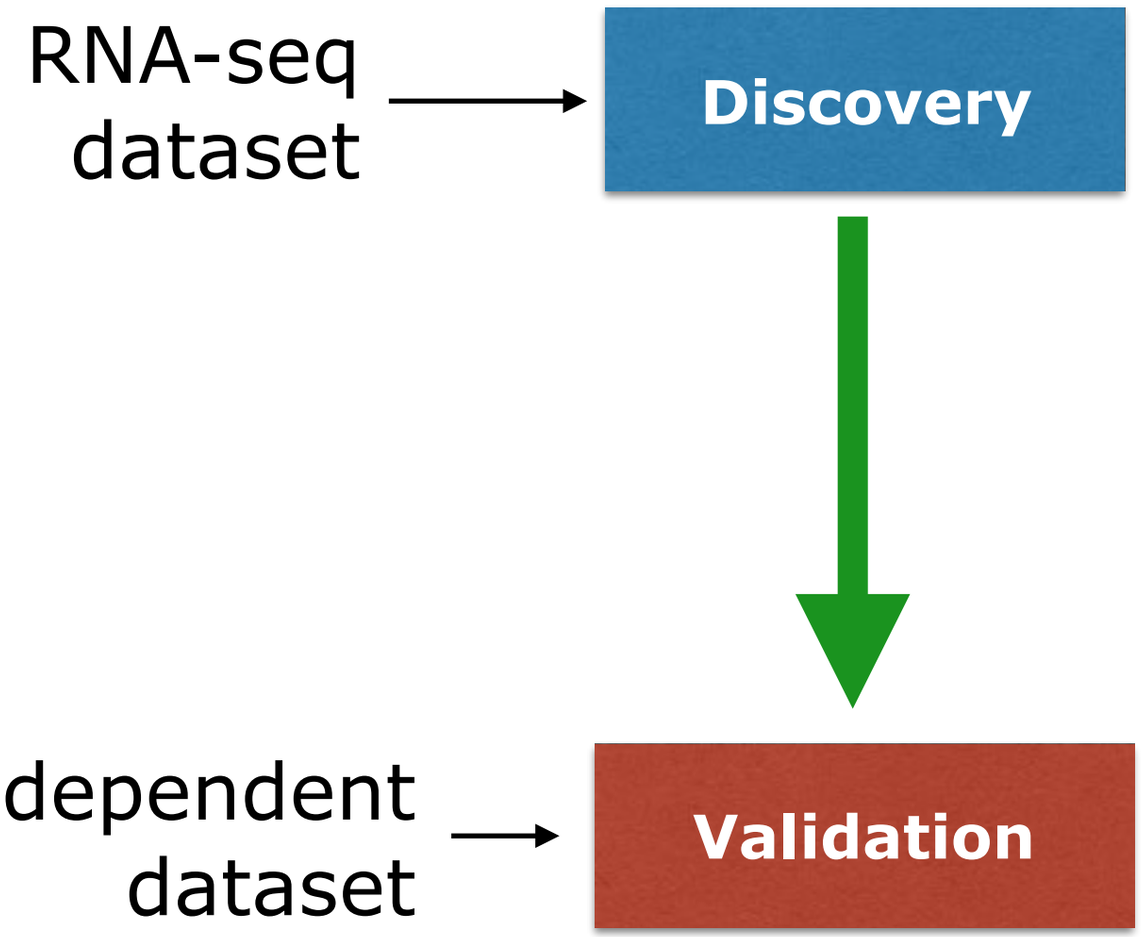

RNA-seq dataset
Discovery
Independent dataset
Validation

Case study



- Snaptron for ***discovery***: what exists? what's prevalent? what's specific?

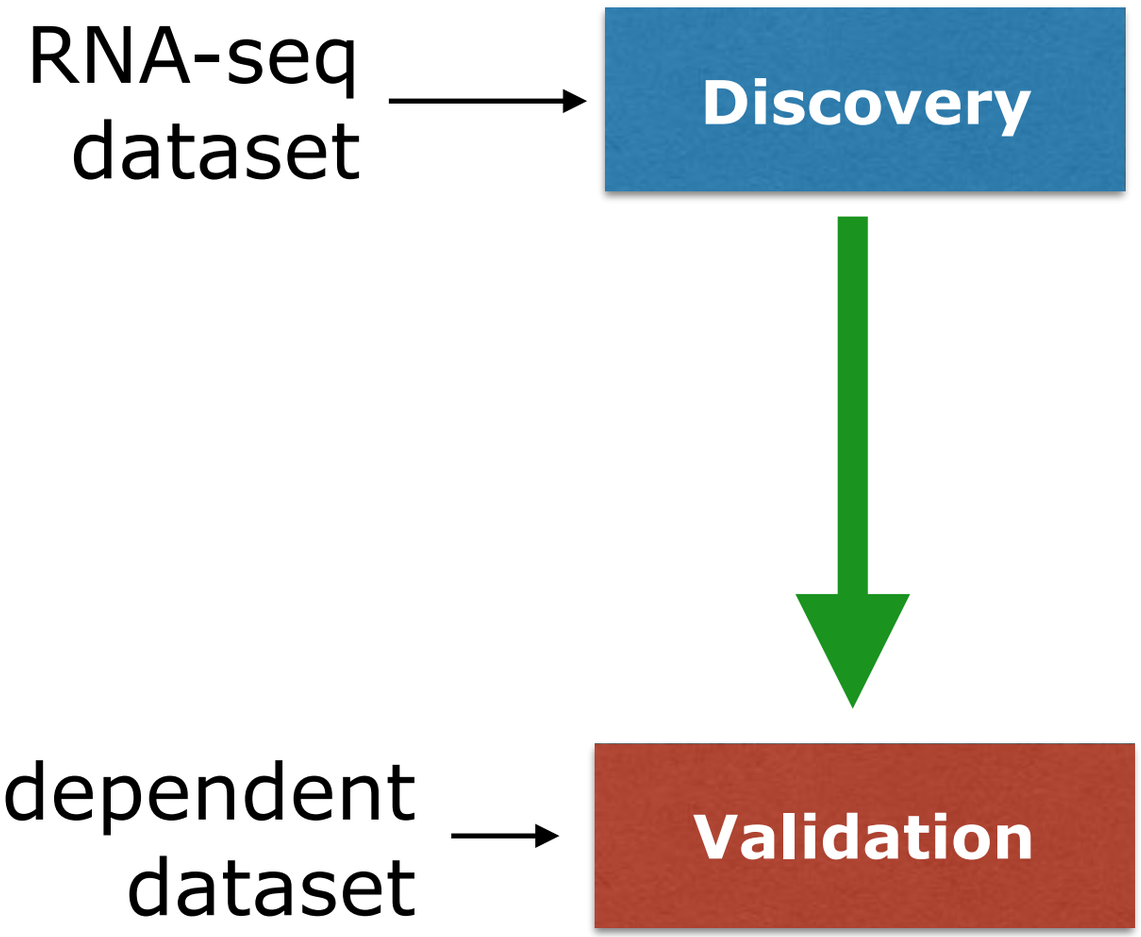
Case study



- Snaptron for ***discovery***: what exists? what's prevalent? what's specific?

- Snaptron for ***validation***: what discoveries have support in public data?

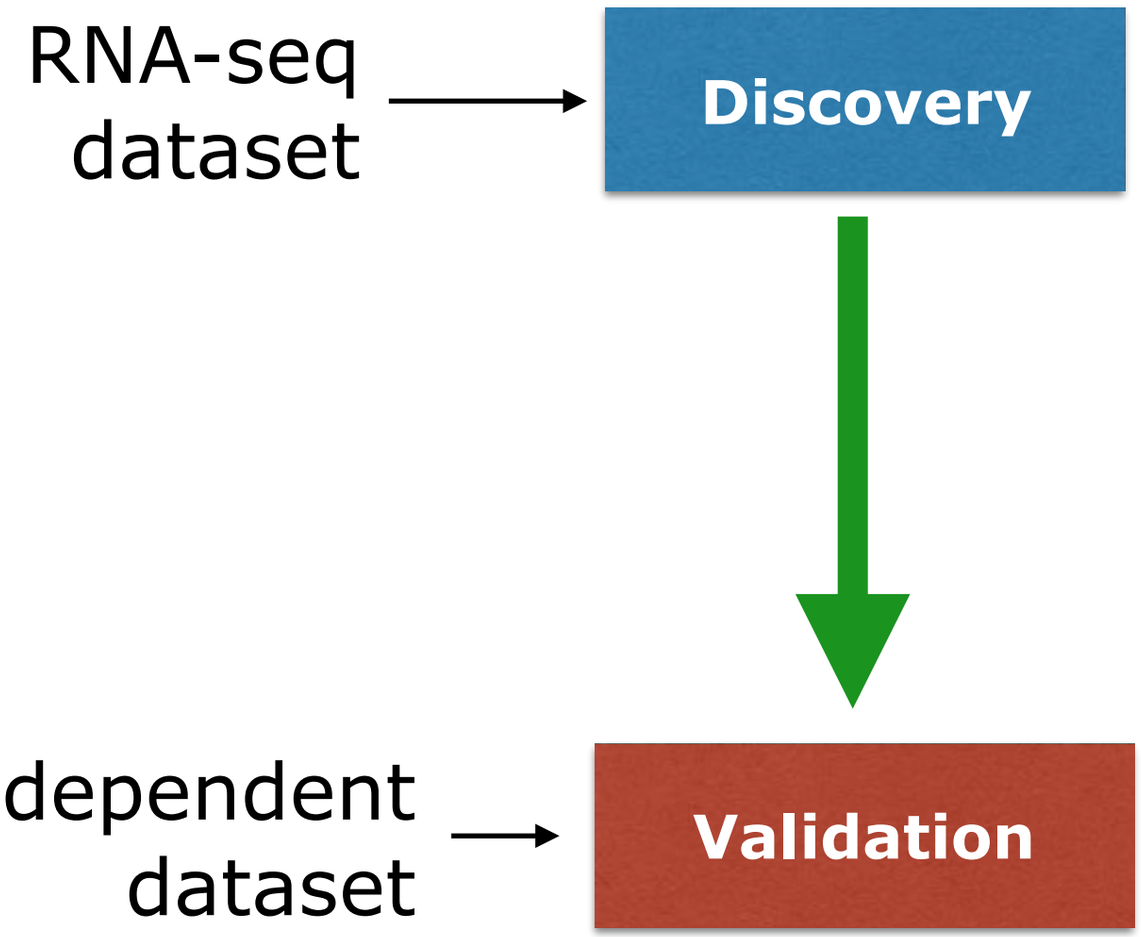
Case study



- Snaptron for ***discovery***: what exists? what's prevalent? what's specific?
- Snaptron for ***prioritization*** of potential discoveries: what discoveries are best supported?
- Snaptron for ***validation***: what discoveries have support in public data?

Case study

Snaptron for ***hypothesis generation & study design***



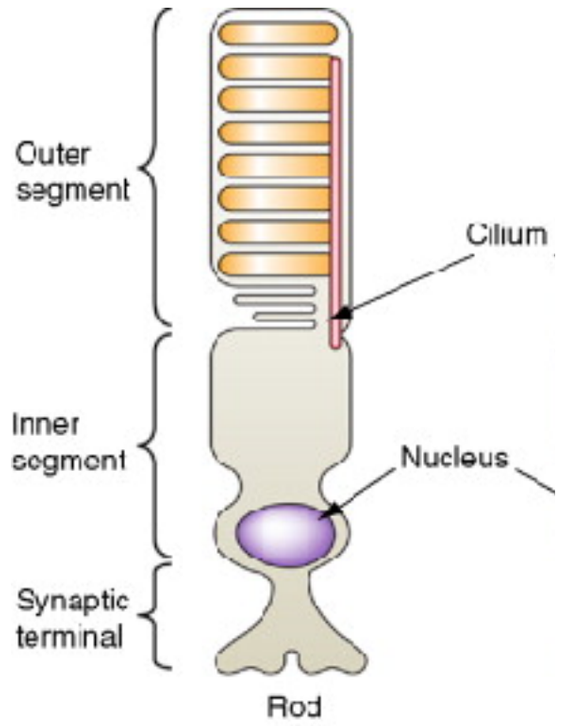
- Snaptron for ***discovery***: what exists? what's prevalent? what's specific?
- Snaptron for ***prioritization*** of potential discoveries: what discoveries are best supported?
- Snaptron for ***validation***: what discoveries have support in public data?

Rod photoreceptors

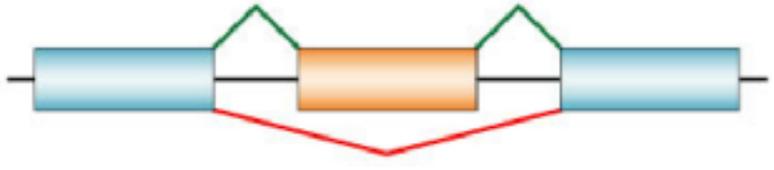
- Detect light & transduce signal to brain
- Degeneration is main cause of hereditary blindness; few treatments

Jonathan Ling

Seth Blackshaw



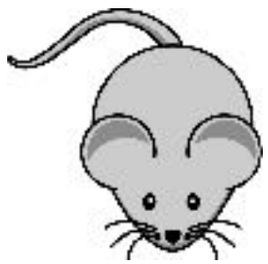
Can we find **rod-specific patterns & splicing factors**, to create a rod-like model from a human cell line?



Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

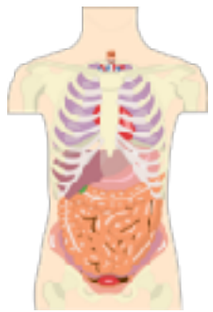
Rod photoreceptors

1. Rods and retinal cells have characteristic exon-usage patterns



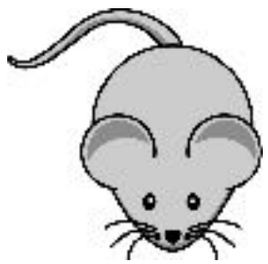
Purified tissue
(FACS/affinity)

GTEx
(+retinal)



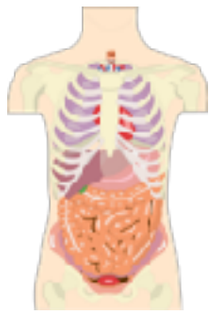
Rod photoreceptors

1. Rods and retinal cells have characteristic exon-usage patterns

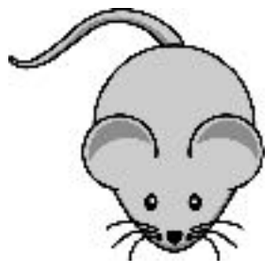


Purified tissue
(FACS/affinity)

GTEx
(+retinal)



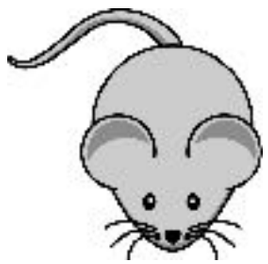
2. Certain exons are utilized only in rods



Purified
tissue

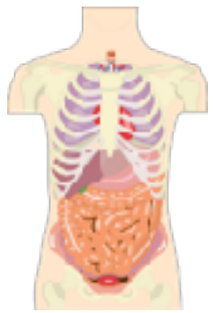
Rod photoreceptors

1. Rods and retinal cells have characteristic exon-usage patterns

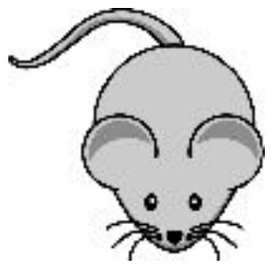


Purified tissue (FACS/affinity)

GTEx (+retinal)

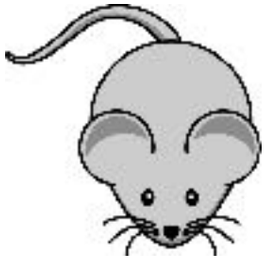


2. Certain exons are utilized only in rods



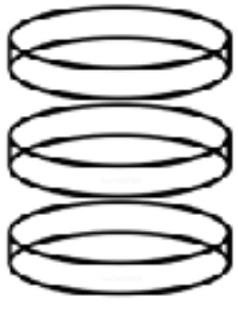
Purified tissue

3. Certain splicing factors work specifically in rods



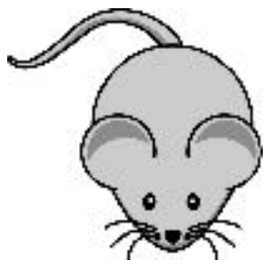
Purified tissue

ENCODE



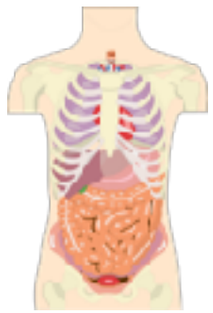
Rod photoreceptors

1. Rods and retinal cells have characteristic exon-usage patterns

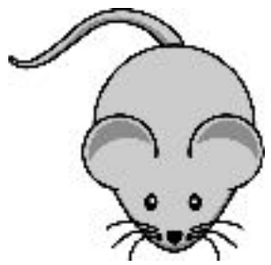


Purified tissue
(FACS/affinity)

GTEx
(+retinal)

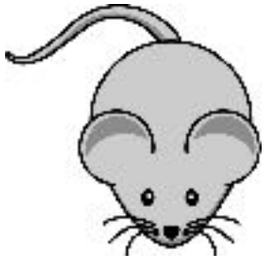


2. Certain exons are utilized only in rods



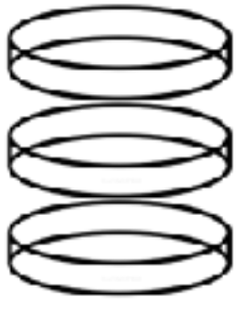
Purified
tissue

3. Certain splicing factors work specifically in rods

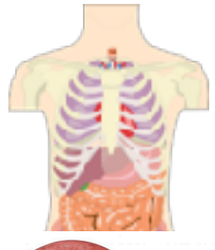


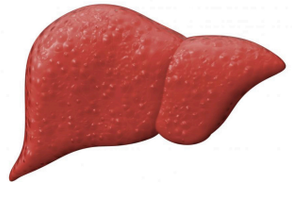
Purified tissue

ENCODE

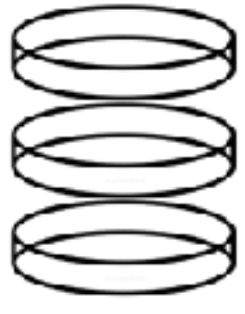


4. Up-regulating those factors induces rod-like splicing in a human cell line

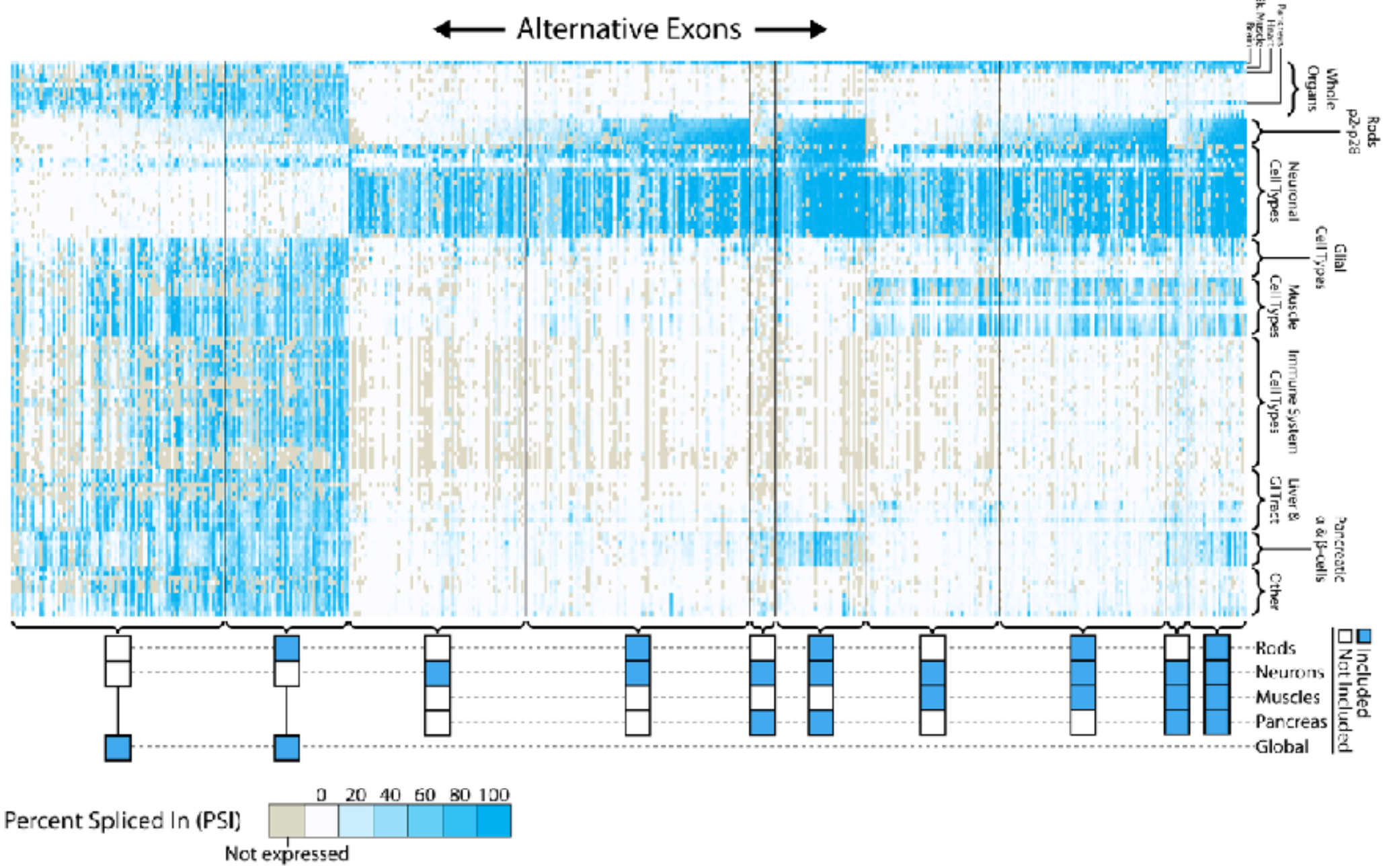




New data,
HepG2 cell line



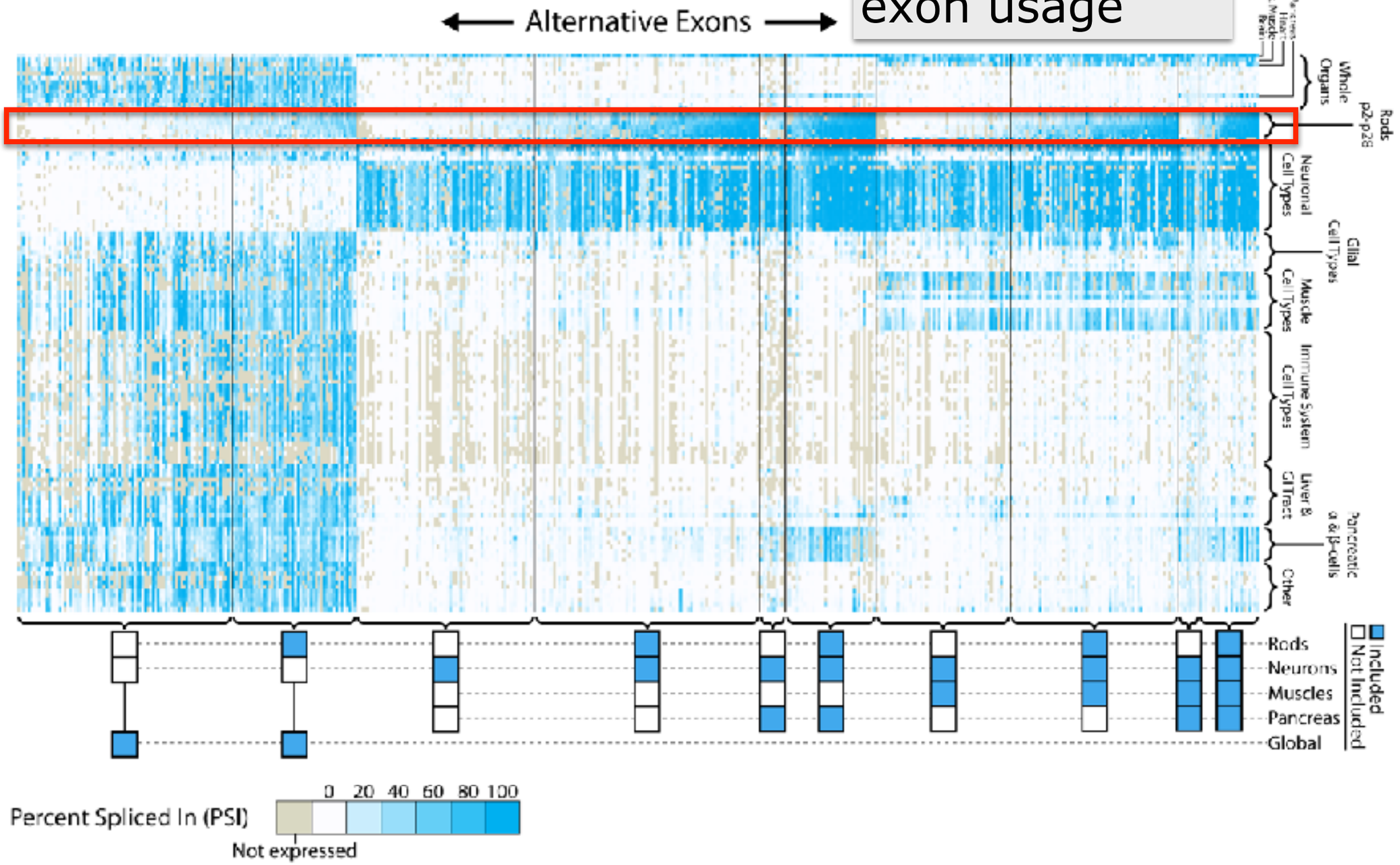
Rod photoreceptors



Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

Rod photoreceptors

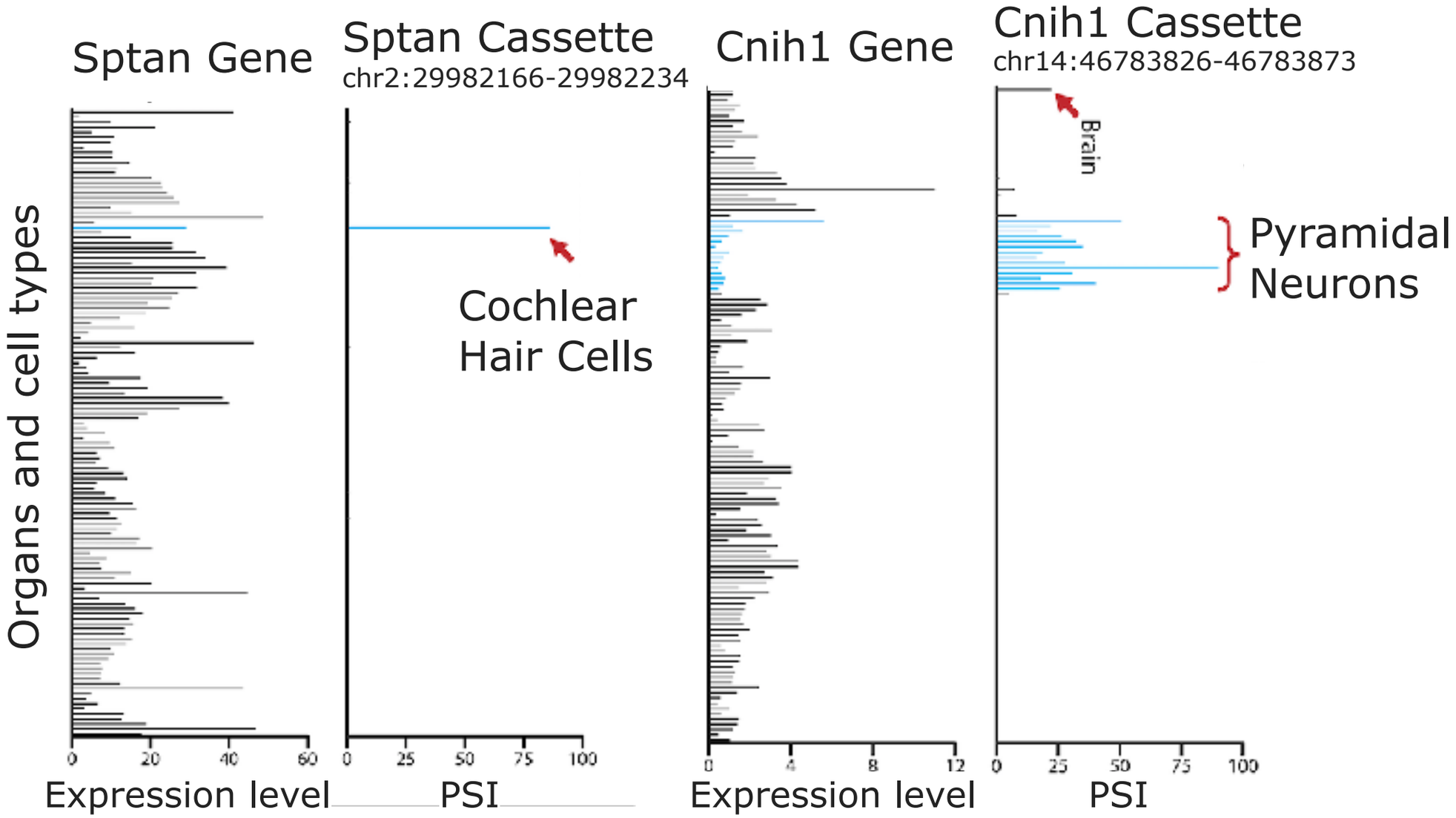
Rods have characteristic patterns of exon usage



Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

Rod photoreceptors

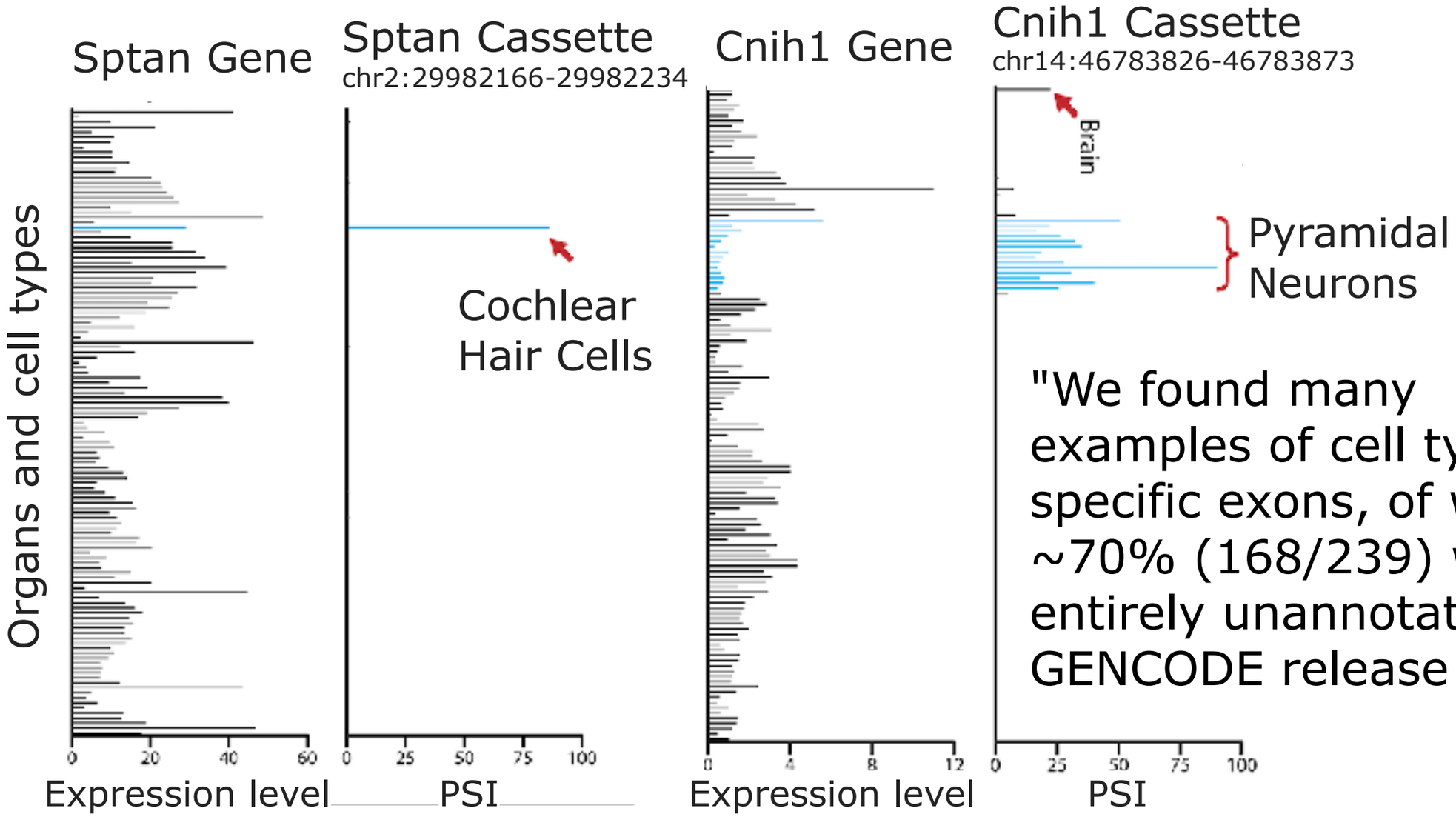
Exon usage can be a cell-type signature; sometimes invisible at gene level



Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

Rod photoreceptors

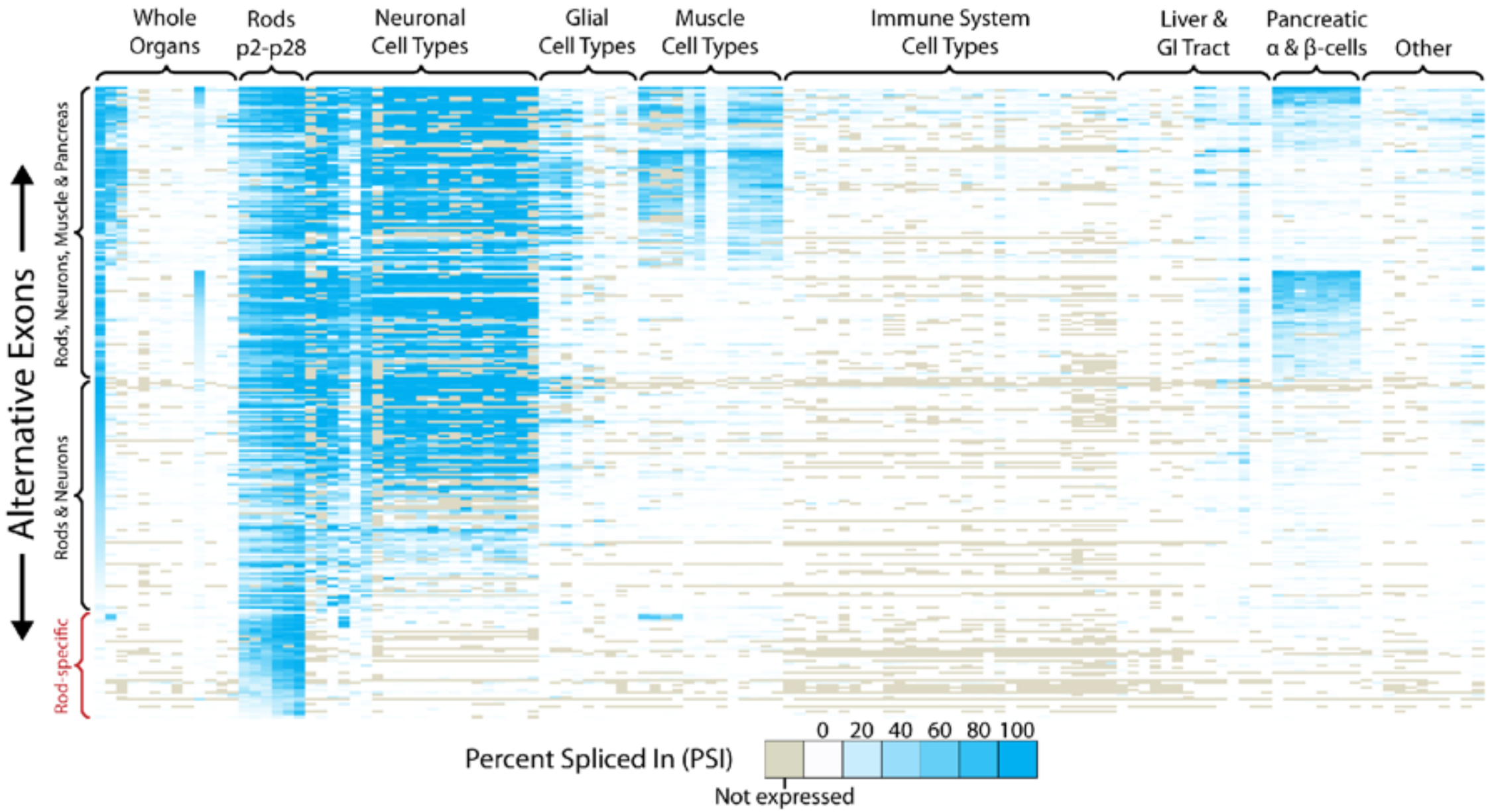
Exon usage can be a cell-type signature; sometimes invisible at gene level



"We found many examples of cell type-specific exons, of which ~70% (168/239) were entirely unannotated in GENCODE release M20"

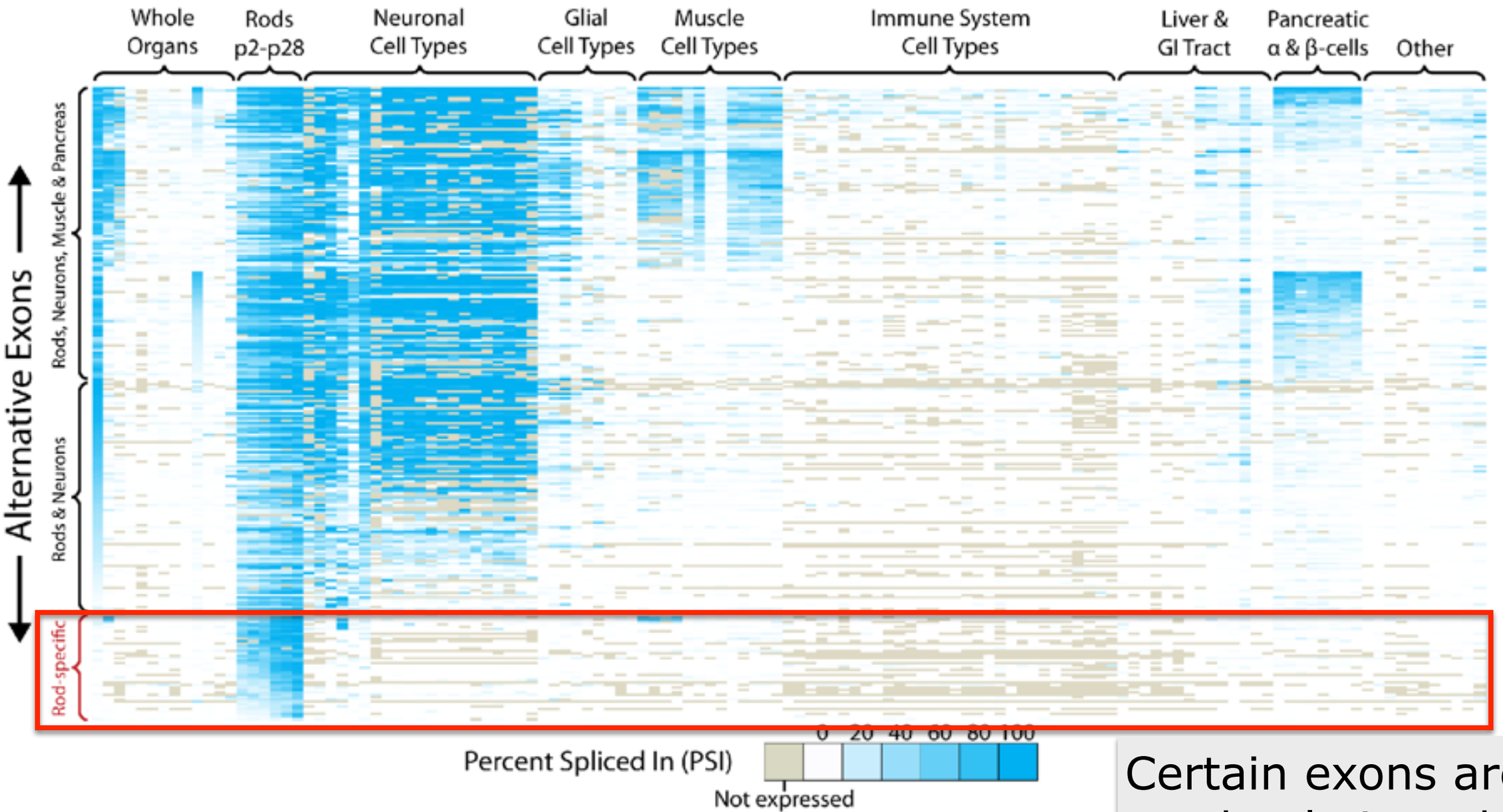
Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

Rod photoreceptors



Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

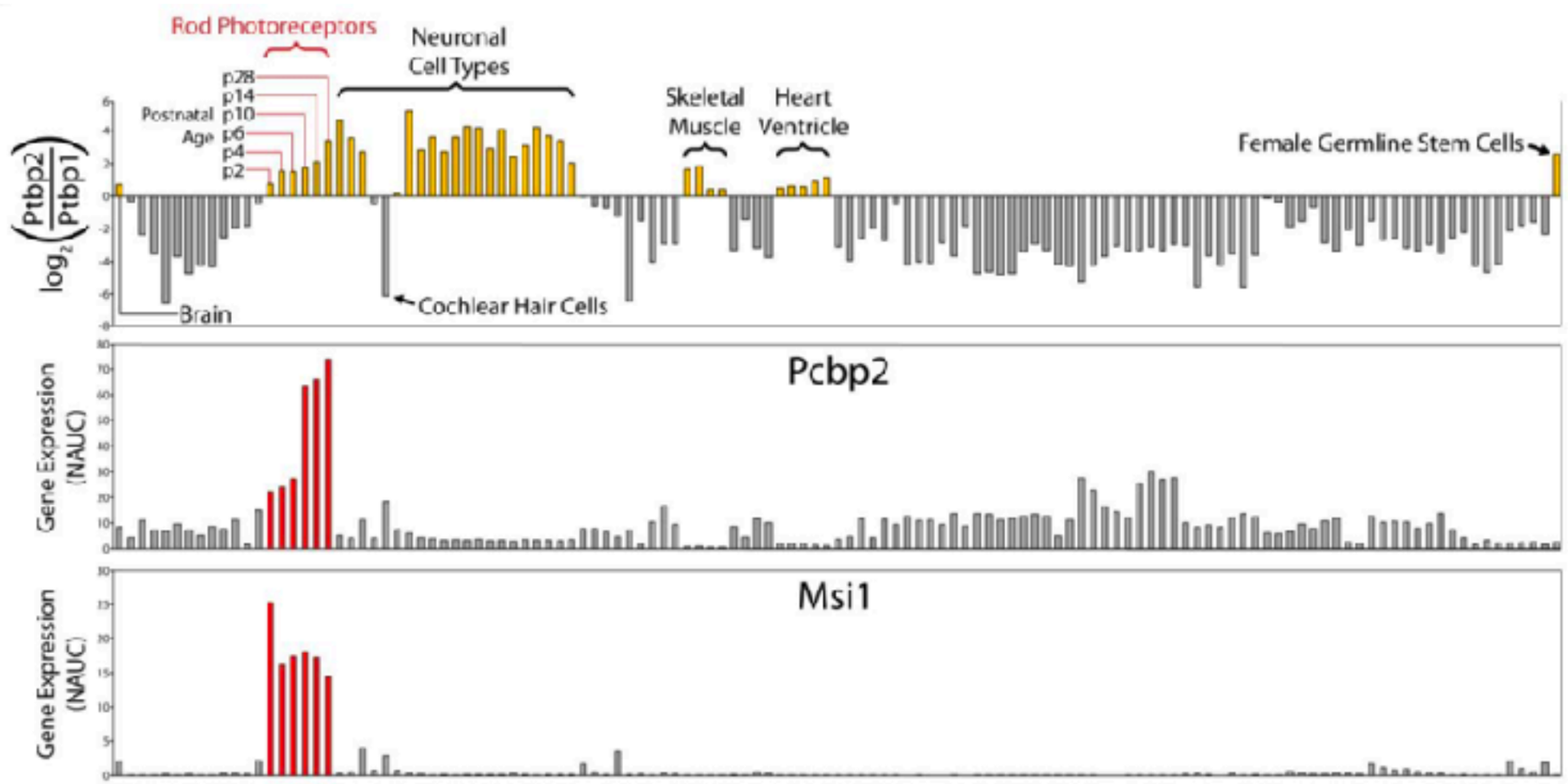
Rod photoreceptors



Certain exons are used only in rods

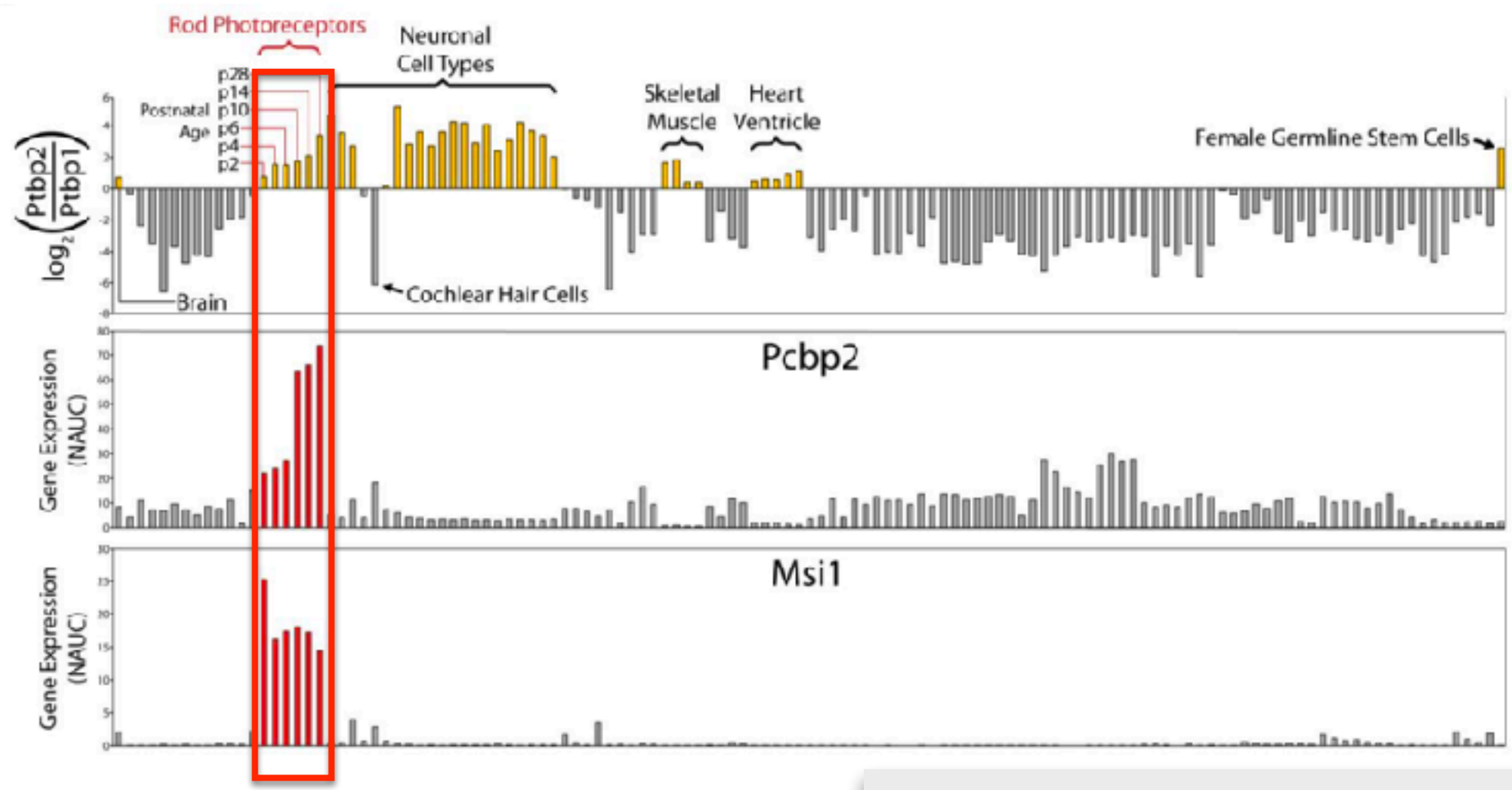
Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

Rod photoreceptors



Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

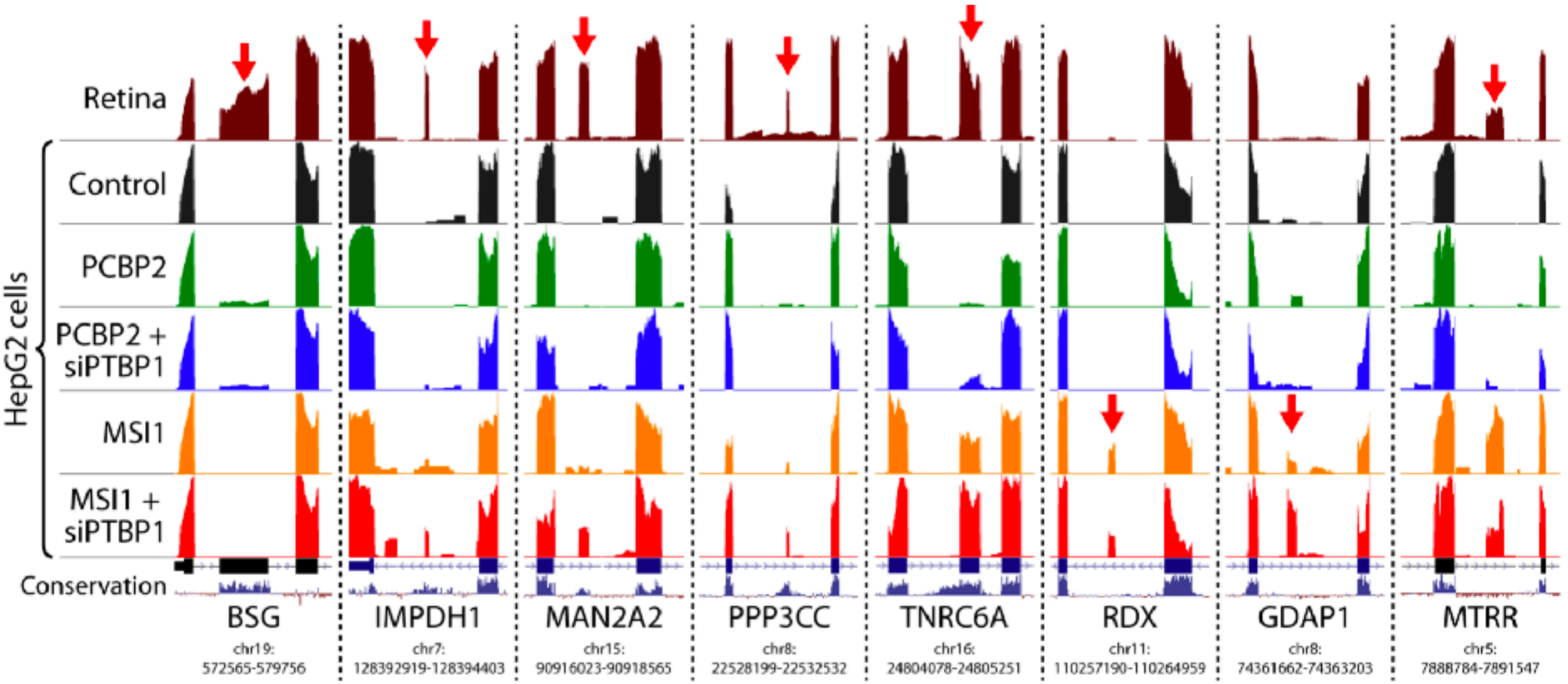
Rod photoreceptors



Certain *splicing factors* are specific. Do they drive rod-specific splicing?

Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

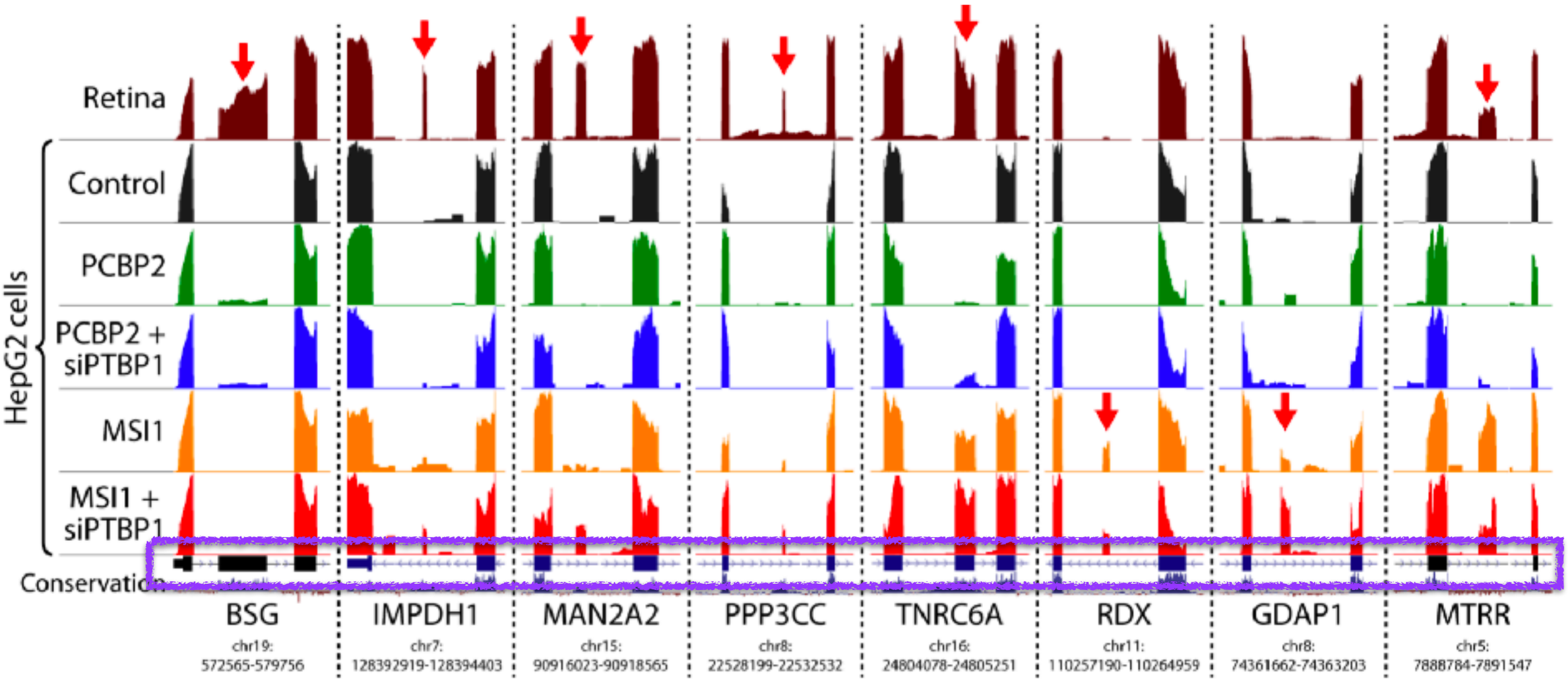
Rod photoreceptors



Up-regulating those splicing factors recovers rod-like splicing in HepG2 cells

Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

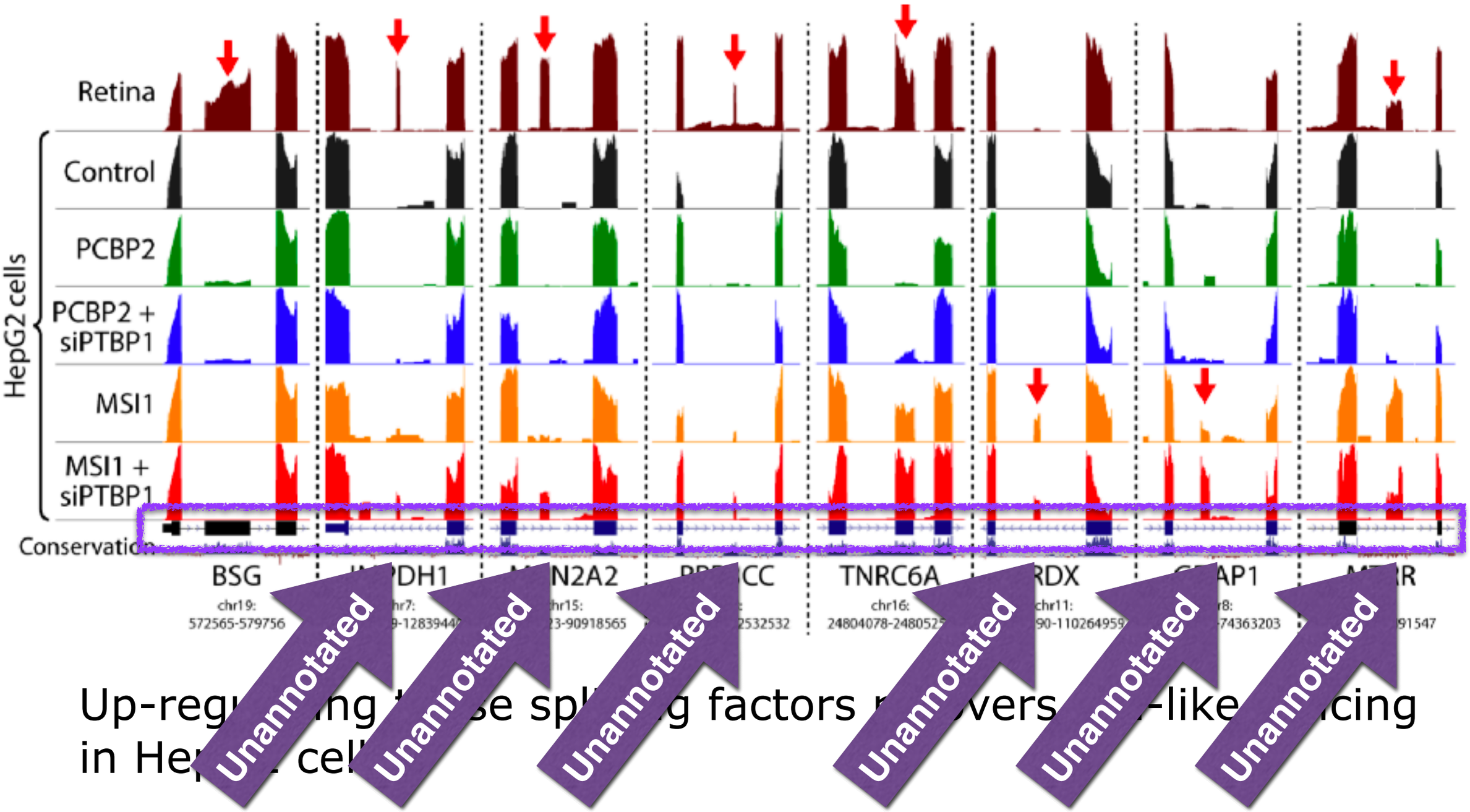
Rod photoreceptors



Up-regulating those splicing factors recovers rod-like splicing in HepG2 cells

Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

Rod photoreceptors

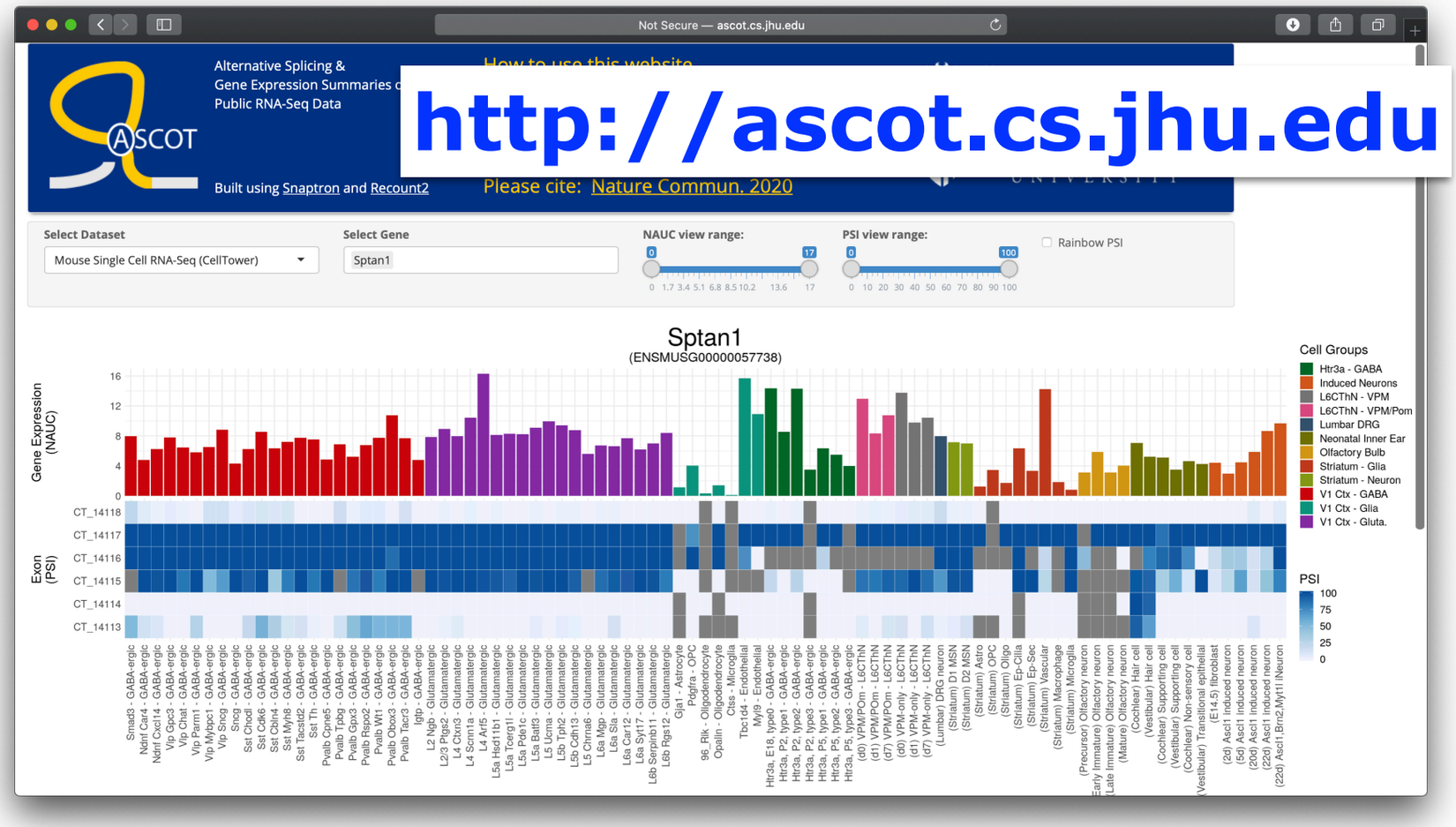


Up-regulating these splicing factors recovers neuron-like splicing in HepG2 cells

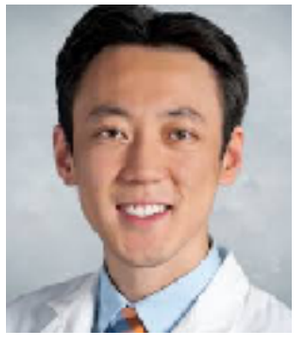
Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

ASCOT

- Explore alternative splicing events in same datasets we used
- Mouse purified tissues, Smart-seq in mouse & human, ENCODE knockdowns

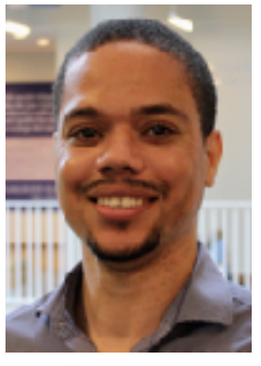


Seth Blackshaw



Jonathan Ling

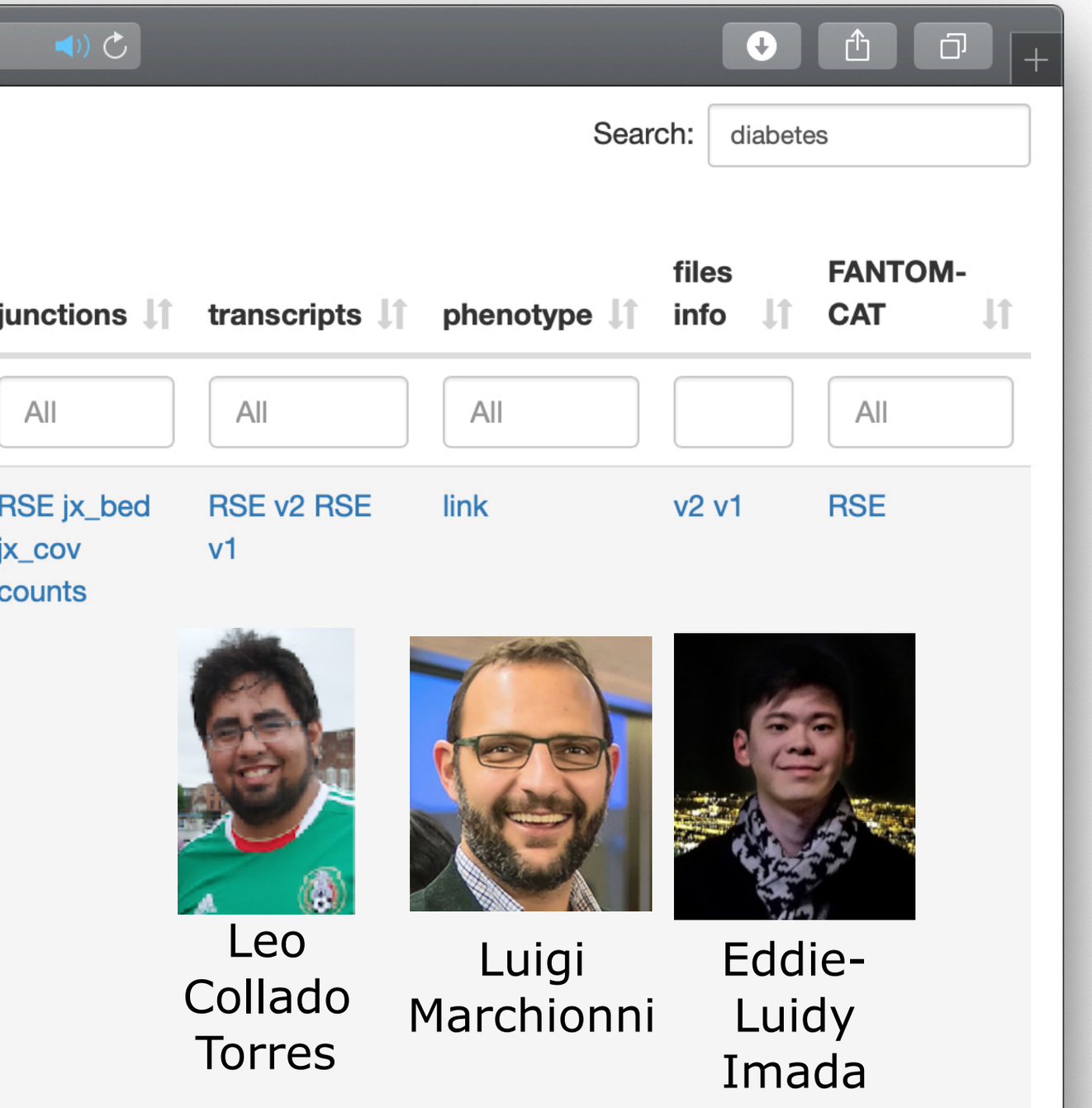
Chris Wilks



Rone Charles

Ling JP, Wilks C, Charles R, Leavey PJ, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, and Blackshaw S. ASCOT identifies key regulators of neuronal subtype-specific splicing. *Nature Communications*, 11(1):137, Jan 2020

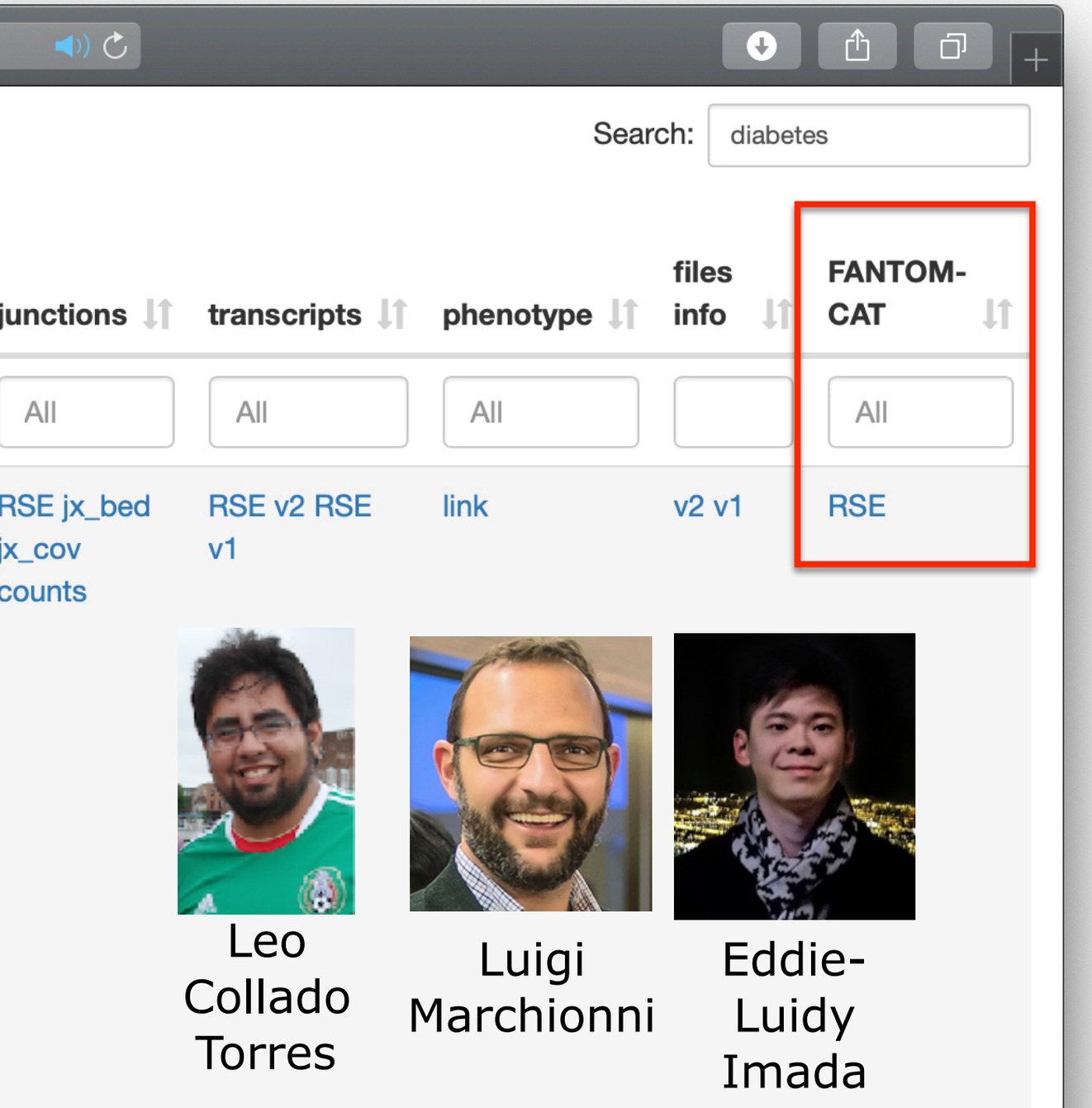
Cancer & ncRNAs



- We also quantified all of recount2 using the more ncRNA-rich FANTOM-CAT annotation

Imada, EL, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, Stupnikov A et al. "Recounting the FANTOM Cage Associated Transcriptome." *BioRxiv* (2019): doi:10.1101/659490.

Cancer & ncRNAs



Leo Collado Torres

Luigi Marchionni

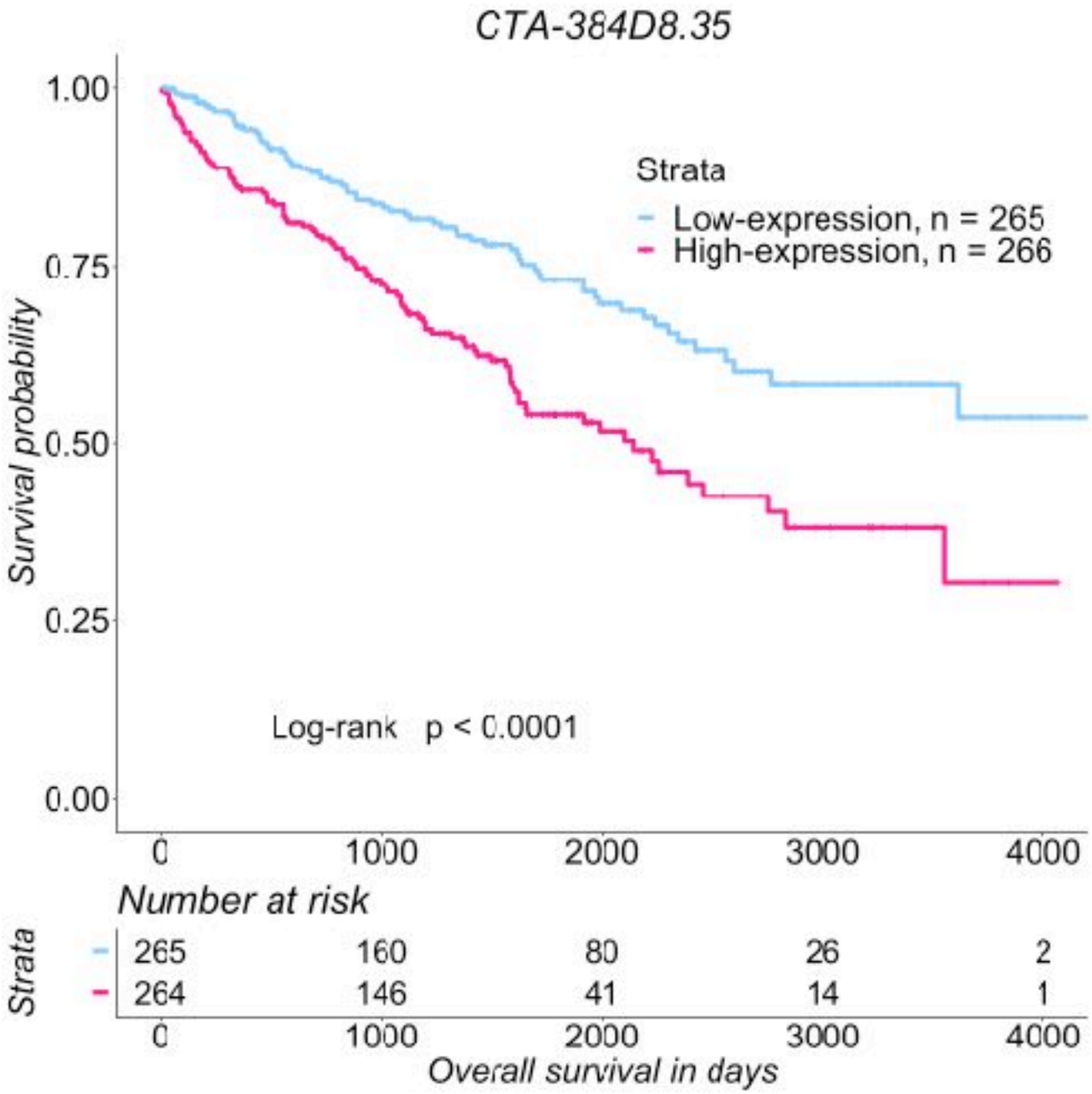
Eddie-Luidy Imada

- We also quantified all of recount2 using the more ncRNA-rich FANTOM-CAT annotation

Imada, EL, Sanchez DF, Collado-Torres L, Wilks C, Matam T, Dinalankara W, Stupnikov A et al. "Recounting the FANTOM Cage Associated Transcriptome." *BioRxiv* (2019): doi:10.1101/659490.

Cancer & ncRNAs

- Enhancer with prognostic value for kidney cancer



Chen H, Li C, Peng X, Zhou Z, Weinstein JN; Cancer Genome Atlas Research Network, Liang H. A Pan-Cancer Analysis of Enhancer Expression in Nearly 9000 Patient Samples. Cell. 2018 Apr 5;173(2):386-399.e12.

snapcount in Bioconductor

Chris Wilks

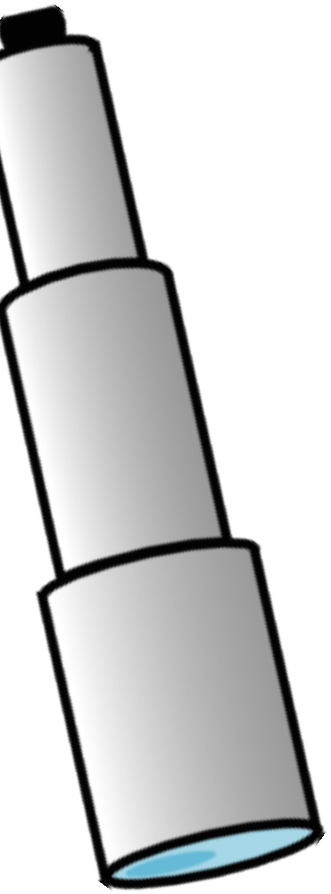
Rone Charles

Future: public data

Rod photoreceptor study involved >90K public datasets

Used public data *only,* up to HepG2 experiment

Desire: querying public data as an *everyday activity* in bio research



WORLD VIEW *A personal take on events*

MEGHNA ABRAHAM

Don’t let useful data go to waste

Researchers must seek out others’ deposited biological sequences in community databases, urges **Franziska Denk.**

One of the best ways for a neuroscientist like me to keep up to date with what colleagues are working on is to attend conferences. But on recent trips I have noticed a problem. Too few researchers are consulting and using publicly available data — my own included. What is going on?

Massive amounts of biological information are being accumulated using high-throughput sequencing techniques. Many scientists have used some of those resources, such as the Encyclopedia of DNA Elements (ENCODE) launched by the US National Human Genome Research Institute. But many more laboratories in neuroscience and other subdisciplines of cell and molecular biology generate their own data sets. These data are piling up in community databases and offer information on gene expression and regulation. Unless this discrepancy, and propose a biologically valid reason for it.

Why are so many bench biologists overlooking this wealth of cell-type-specific expression data?

My hunch is there are two reasons. First, researchers underestimate how many of these data have been published over the past few years because they are being generated across so many different fields. Second, they are wary of the data. Because you need bioinformatics knowledge to generate and analyse sequencing results, people assume that they also need such expertise to locate and interpret them.

Not so. In the past five years, improvements in technology, together with stricter deposition guidelines, mean that simple Excel files commonly accompany papers. These can be downloaded in minutes from the Supplementary Information of a relevant paper, or from the ‘GEO

Future: data science

Public data quickly challenges us with **technical confounders** & **missing/incorrect metadata**

One dataset

All of SRA

What questions can we answer *robustly?*
At what points on the spectrum?

Is metadata fixable?

Ellis SE, Collado-Torres L, Jaffe A, Leek JT. **Improving the value of public RNA-seq expression data by phenotype prediction**. *Nucleic Acids Res*. 2018 May 18;46(9):e54.

Future: data science

Public data quickly challenges us with **technical confounders** & **missing/incorrect metadata**

One dataset

All of SRA

What questions can we answer *robustly?*
At what points on the spectrum?

Is metadata fixable?

Ellis SE, Collado-Torres L, Jaffe A, Leek JT. **Improving the value of public RNA-seq expression data by phenotype prediction**. *Nucleic Acids Res*. 2018 May 18;46(9):e54.

Future: data science

Public data quickly challenges us with **technical confounders** & **missing/incorrect metadata**

One dataset

All of SRA

What questions can we answer *robustly?*
At what points on the spectrum?

Is metadata fixable?

Ellis SE, Collado-Torres L, Jaffe A, Leek JT. **Improving the value of public RNA-seq expression data by phenotype prediction**. *Nucleic Acids Res*. 2018 May 18;46(9):e54.

Abhinav Nellore

Leo Collado Torres

Chris Wilks

Rone Charles

Kasper Hansen

Jeff Leek

Seth Blackshaw

Jonathan Ling

Margaret Taub

Shannon Ellis

Kai Kammers

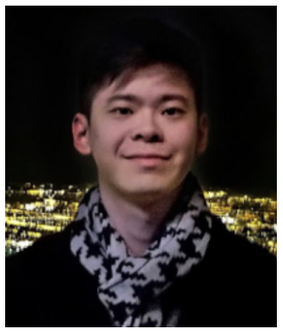
Jamie Morton

José Alquicira-Hernández

Andrew Jaffe

Jacob Pritt

Luigi Marchionni



Eddie-Luidy Imada

- NIH R01GM118568 (Langmead)
- NIH R01GM121459 (Hansen)
- NSF CAREER IIS-1349906 (Langmead)
- NIH R01GM105705 (Leek)
- NIH Cloud Credits, CCREQ-2017-03-00086 (Langmead)
- NSF XSEDE projects (TG-CIE170020, TG-DEB180021)

IDIES Seed funding
SciServer
SciServer Compute

langmead-lab.org, @BenLangmead

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Cloud computing

Cloud computing is a natural fit for reanalyzing public data and for far-flung collaborations

COMPUTATIONAL TOOLS

Cloud computing for genomic data analysis and collaboration

Ben Langmead[1] and Abhinav Nellore[2]

Abstract | Next-generation sequencing has made major strides in the past decade. Studies based on large sequencing data sets are growing in number, and public archives for raw sequencing data have been doubling in size every 18 months. Leveraging these data requires researchers to use large-scale computational resources. Cloud computing, a model whereby users rent computers and storage from large data centres, is a solution that is gaining traction in genomics research. Here, we describe how cloud computing is used in genomics for research and large-scale collaborations, and argue that its elasticity, reproducibility and privacy features make it ideally suited for the large-scale reanalysis of publicly available archived data, including privacy-protected data.

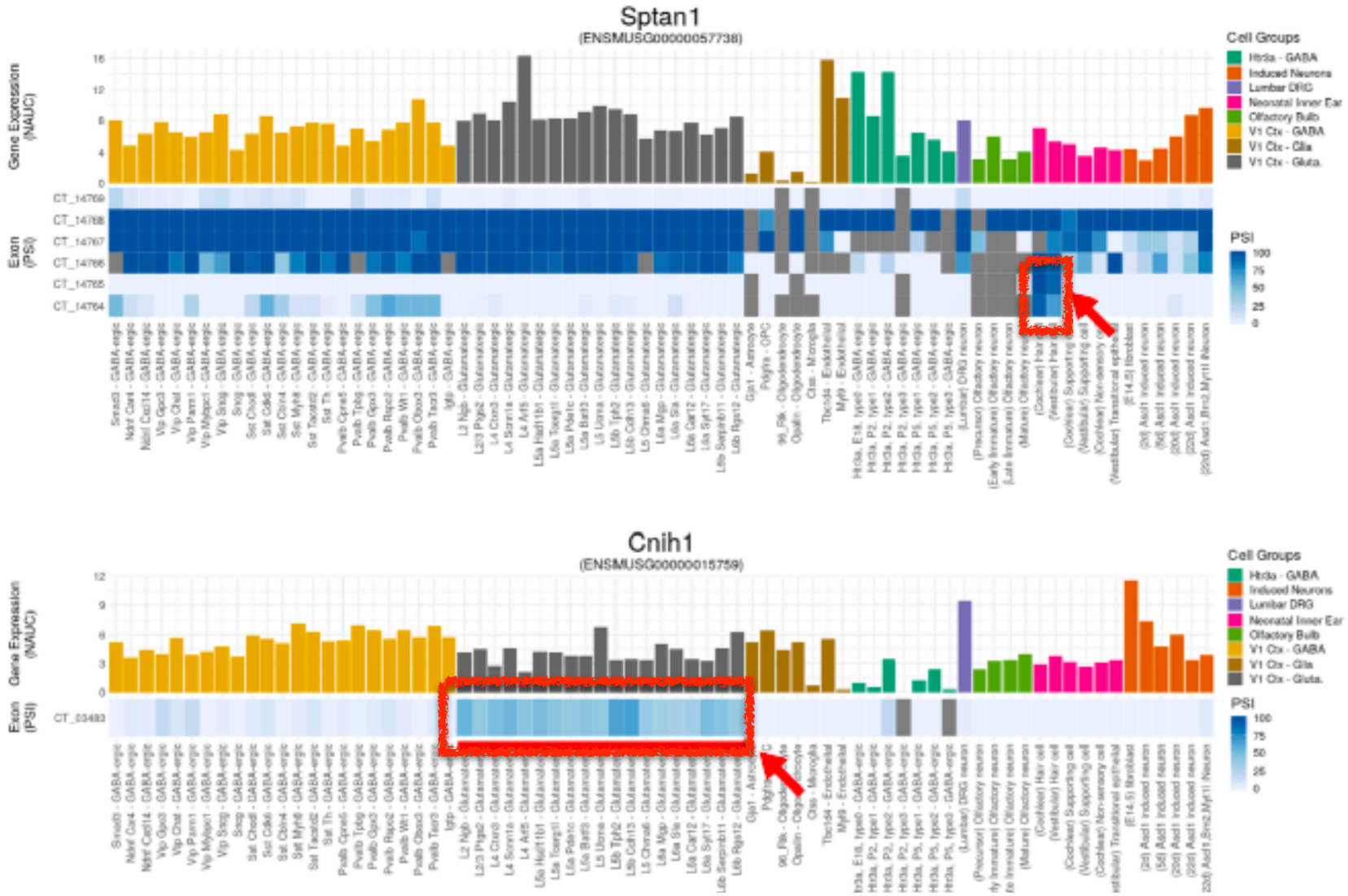
Next-generation sequencing (NGS) technologies have been improving rapidly and have become the workhorse technology for studying nucleic acids. NGS platforms work by collecting information on a large array of polymerase reactions working in parallel, up to billions at a time inside a single sequencer[1]. The speed and decreasing cost of NGS have led to the rapid accumulation of raw sequencing data (sequencing reads), used in published studies, in public archives[2] such as the Sequence Read Archive (SRA)[3,4], which is hosted by the US National Center for Biotechnology Information (NCBI), and the European Nucleotide Archive (ENA)[5], which is hosted by the European Molecular Biology Laboratory at the European Bioinformatics Institute

programme[17], among others (TABLE 1). gnomAD now spans over 120,000 exomes and over 15,000 whole genomes. ICGC encompasses over 70 subprojects targeting distinct cancer types, which are conducted in more than a dozen countries and have already collected samples from more than 20,000 donors. Aligned sequencing reads for ICGC require over 1 petabyte (PB; that is, a million GB) of storage. The TOPMed programme, which plans to sequence more than 120,000 genomes[17], has already deposited more than 18,000 human whole-genome sequencing data sets in the SRA, comprising approximately 2.3 petabases or about 16.5% of the entire archive. Large observational studies currently in progress, such as the Precision Medicine Initiative[18] and

Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nature Reviews Genetics. 2018 Apr;19(4):208-219.

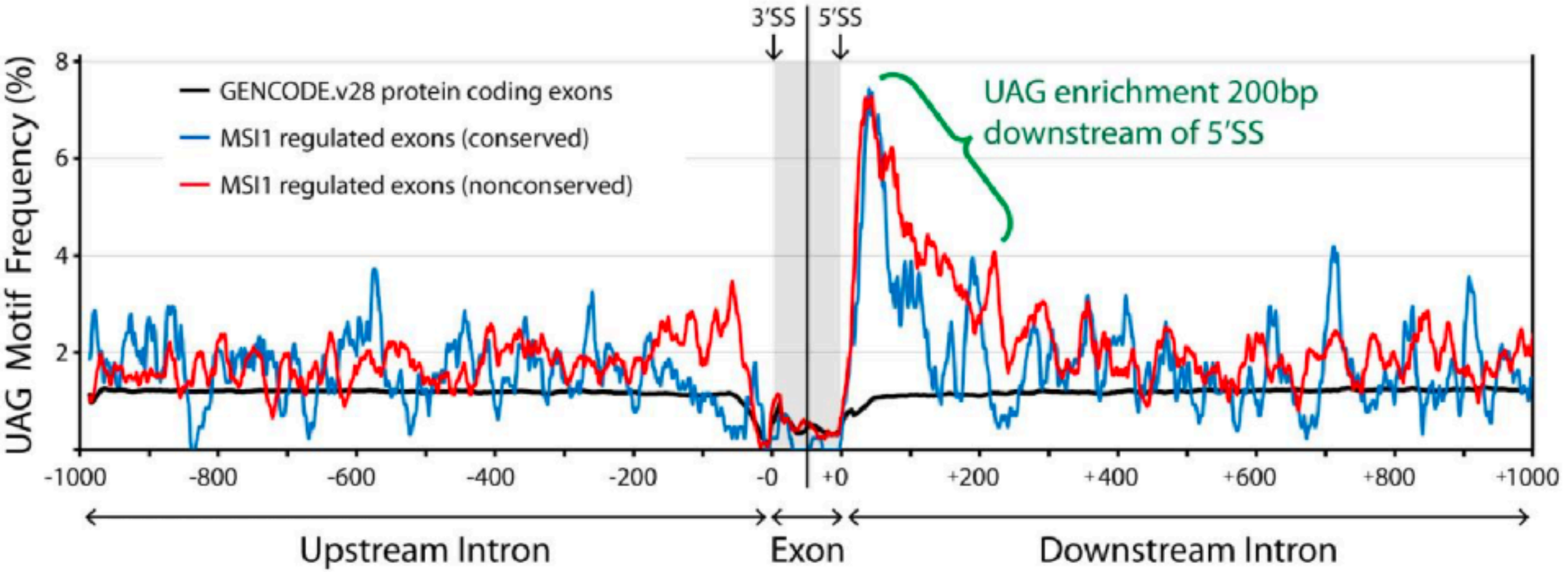
Rods: single-cell

Same exon usage patterns also seen in full-transcript single-cell RNA-seq data



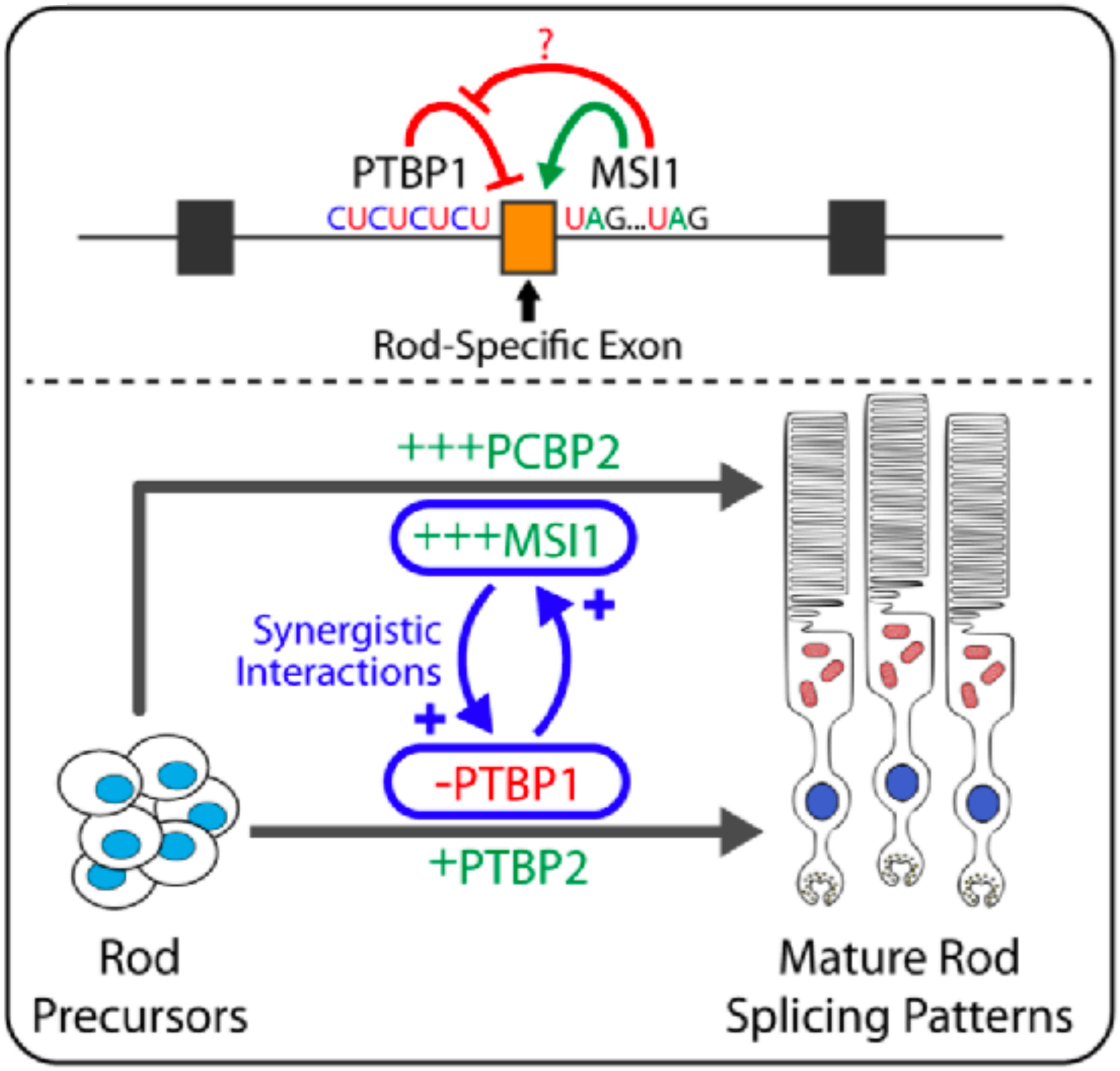
Ling JP, Wilks C, Charles R, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, Blackshaw S. ASCOT identifies key regulators of photoreceptor-specific splicing. *In preparation.*

Rods: MSI1 binding

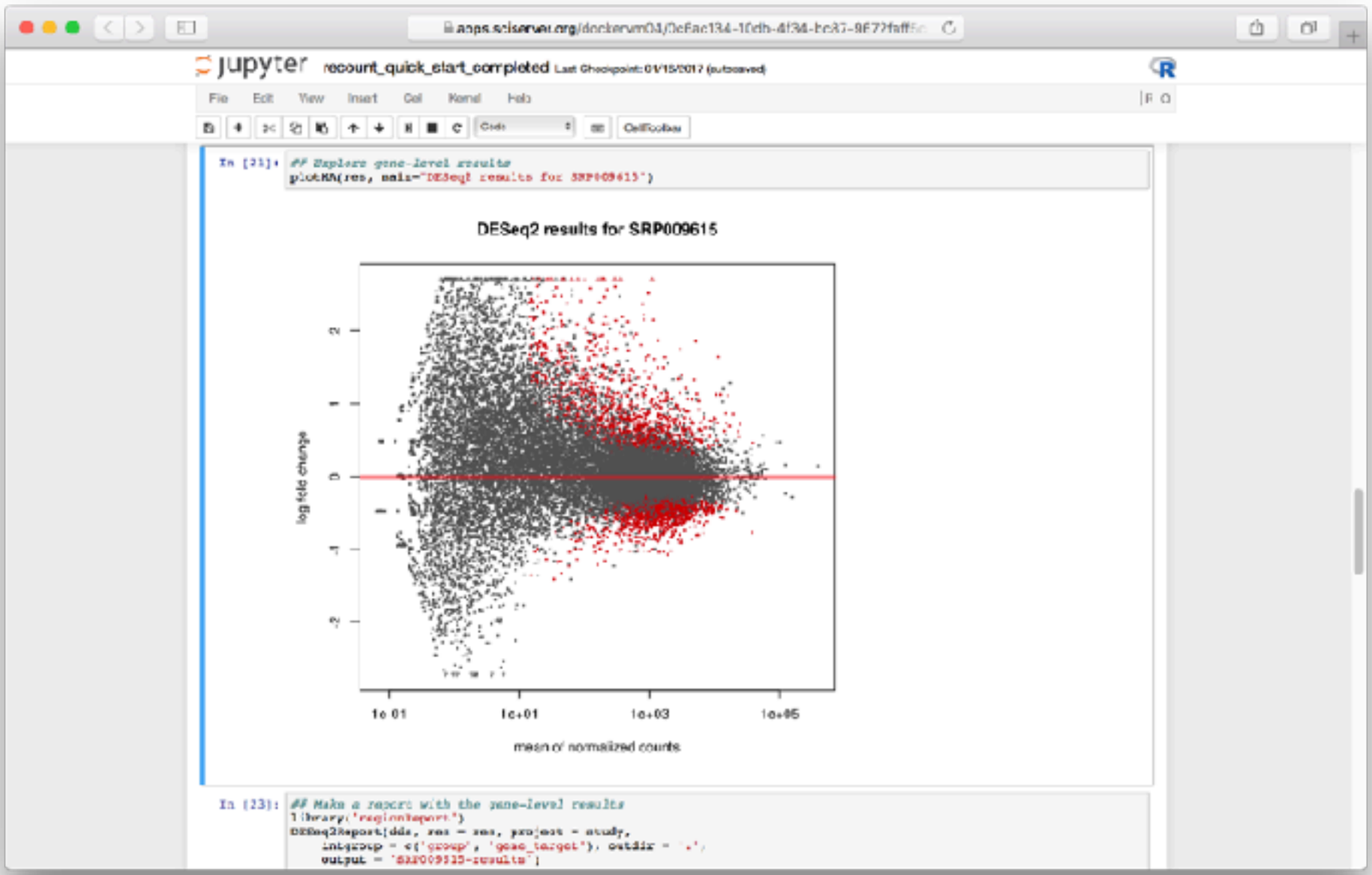


UAG -- consensus MSI1 binding site -- is enriched in downstream intron near MSI1-regulated exon

Rods: proposed mechanism



recount2



http://bit.ly/recount_sciserver

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 Apr 11;35(4):319-321.

Snaptron vignette 2

Darby MM, Leek JT, Langmead B, Yolken RH, Sabunciyan S. Widespread splicing of repetitive element loci into coding regions of gene transcripts. Hum Mol Genet. 2016 Nov 15;25(22):4962-4982.

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2017 Sep 1. btx547.

Snaptron vignette 2

- Darby *et al* studied prevalence of repeat element (RE) expression in the human orbitofrontal cortex

Darby MM, Leek JT, Langmead B, Yolken RH, Sabunciyan S. Widespread splicing of repetitive element loci into coding regions of gene transcripts. Hum Mol Genet. 2016 Nov 15;25(22):4962-4982.

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2017 Sep 1. btx547.

Snaptron vignette 2

- Darby *et al* studied prevalence of repeat element (RE) expression in the human orbitofrontal cortex
- Used RNA-seq to find junctions linking annotated exons to REs in annotated introns, indicating exonization

Darby MM, Leek JT, Langmead B, Yolken RH, Sabunciyan S. Widespread splicing of repetitive element loci into coding regions of gene transcripts. Hum Mol Genet. 2016 Nov 15;25(22):4962-4982.

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2017 Sep 1. btx547.

Snaptron vignette 2

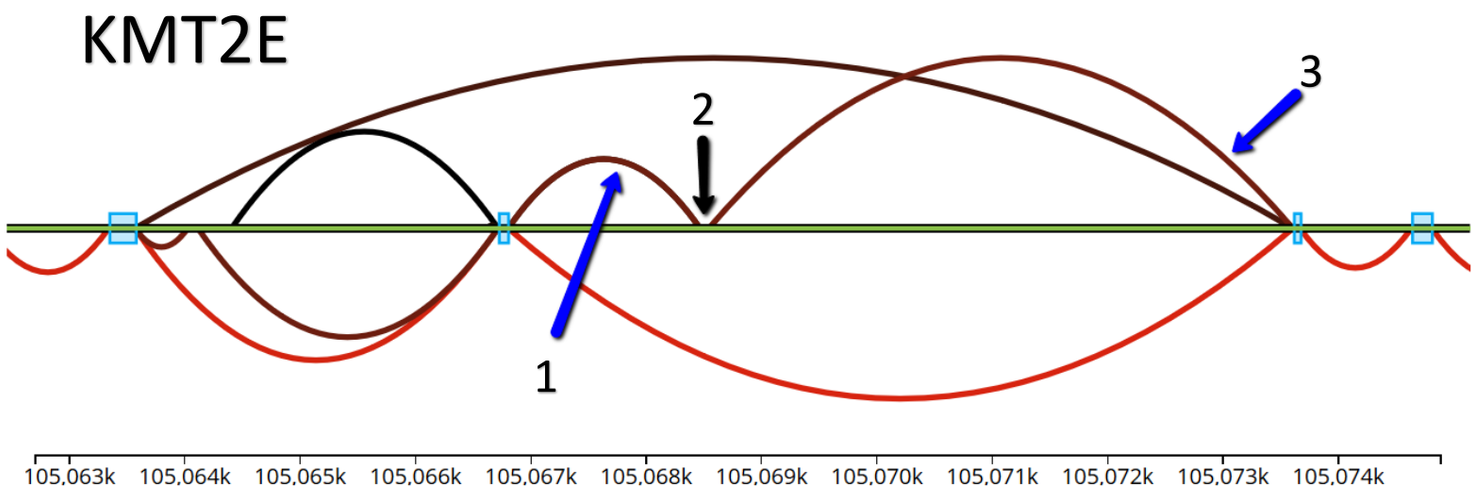
- Darby *et al* studied prevalence of repeat element (RE) expression in the human orbitofrontal cortex
- Used RNA-seq to find junctions linking annotated exons to REs in annotated introns, indicating exonization
- They supplied us 5 events where RE exon was *unannotated*; Snaptron SSC query confirmed all 5 occurred at least 35 times in SRAv2 & GTEx

Darby MM, Leek JT, Langmead B, Yolken RH, Sabunciyan S. Widespread splicing of repetitive element loci into coding regions of gene transcripts. Hum Mol Genet. 2016 Nov 15;25(22):4962-4982.

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2017 Sep 1. btx547.

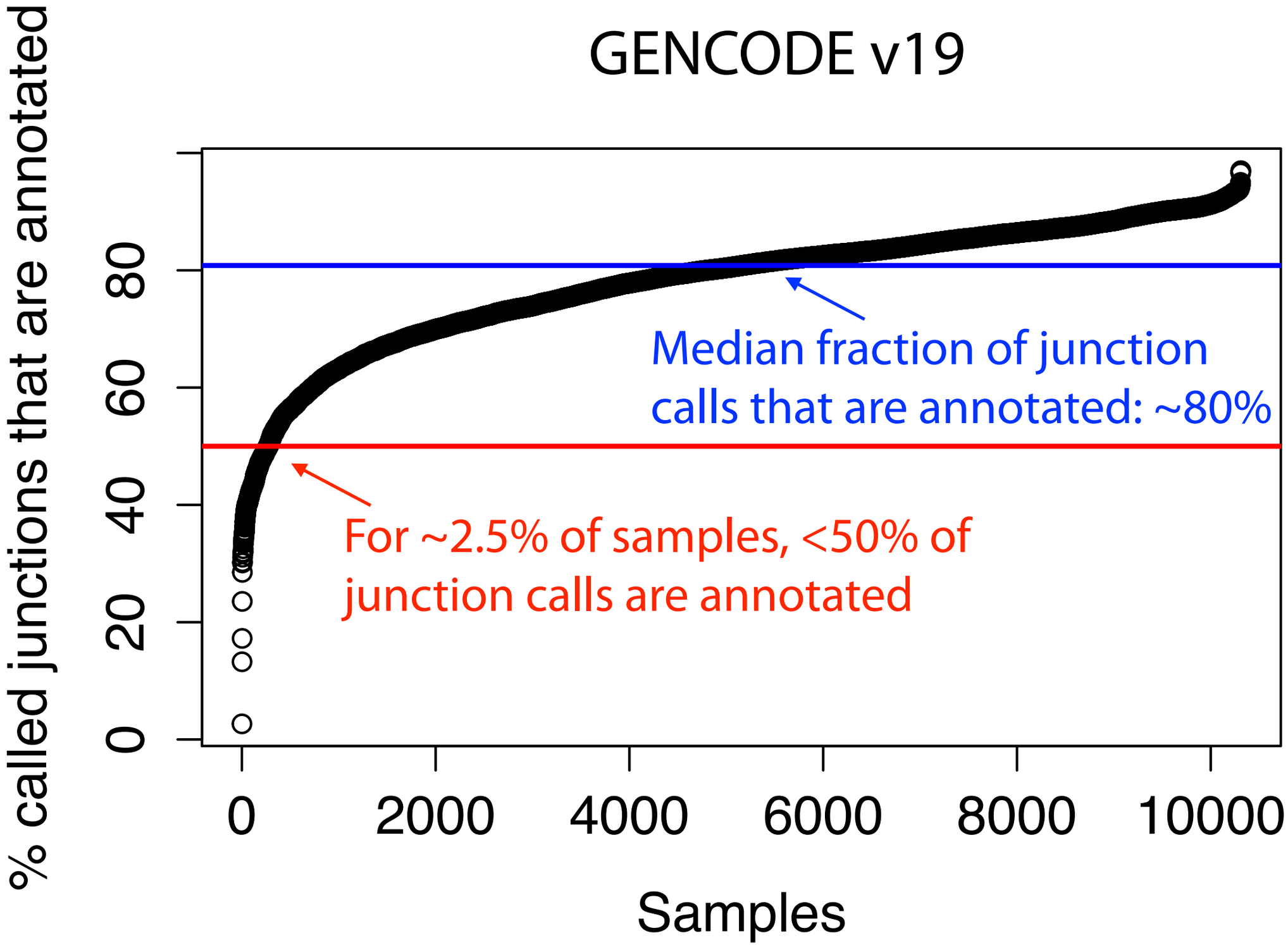
Snaptron vignette 2

- One of the 5 shown here (arrow 2)

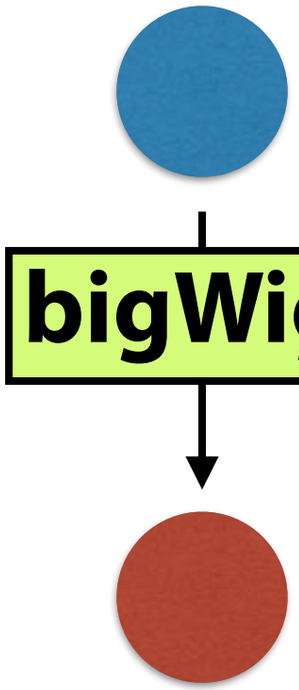


- *Tissue specificity* query showed all 5 events were expressed in a tissue-specific pattern in GTEx (Kruskal-Wallis $P < 0.01$)

Wilks C, Gaddipati P, Nellore A, Langmead B. Snaptron: querying and visualizing splicing across tens of thousands of RNA-seq samples. *Bioinformatics*. 2017 Sep 1. btx547.


GENCODE v19
% called junctions that are annotated
0
20
40
60
80
0
2000
4000
6000
8000
10000
Samples
Median fraction of junction calls that are annotated: ~80%
For ~2.5% of samples, <50% of junction calls are annotated

A third way



Spliced alignment

Differentially expressed region finder

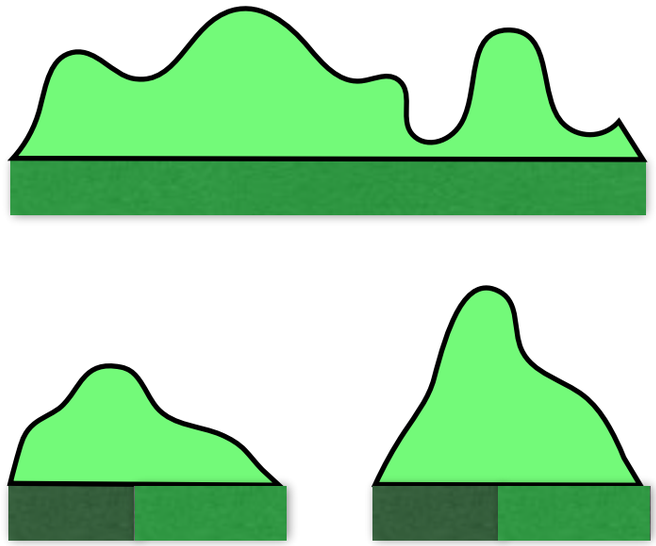
Rail-RNA: accurate, annotation-agnostic

derfinder: region-based, annotation-agnostic

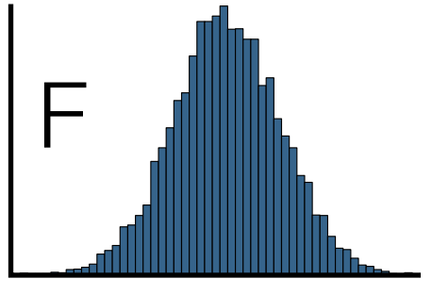
Collado-Torres L, Nellore A, **Frazee AC**, Wilks C, Love MI, Langmead B, Irizarry RA, **Leek JT**, Jaffe AE. Flexible expressed region analysis for RNA-seq with derfinder. Nucleic Acids Res. 2017 Jan 25;45(2):e9.

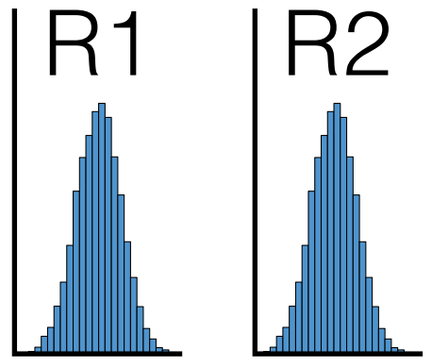
Boiler: RNA-seq alignment compression

Coverage

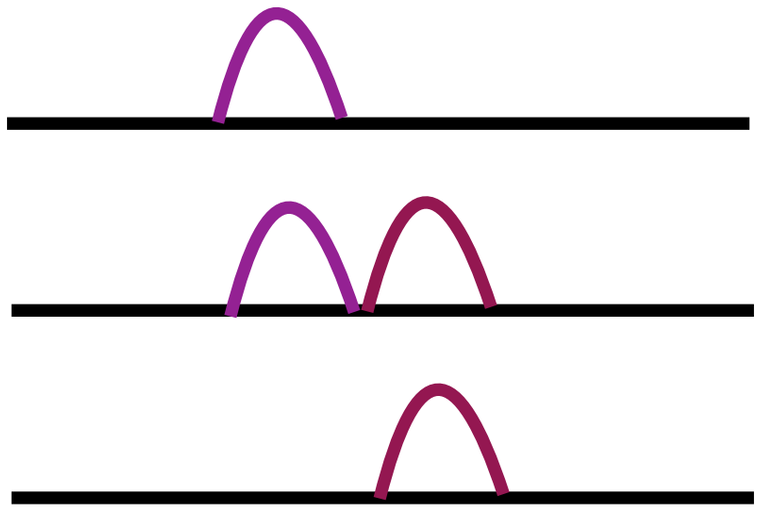


Length tallies





Co-occurrence patterns



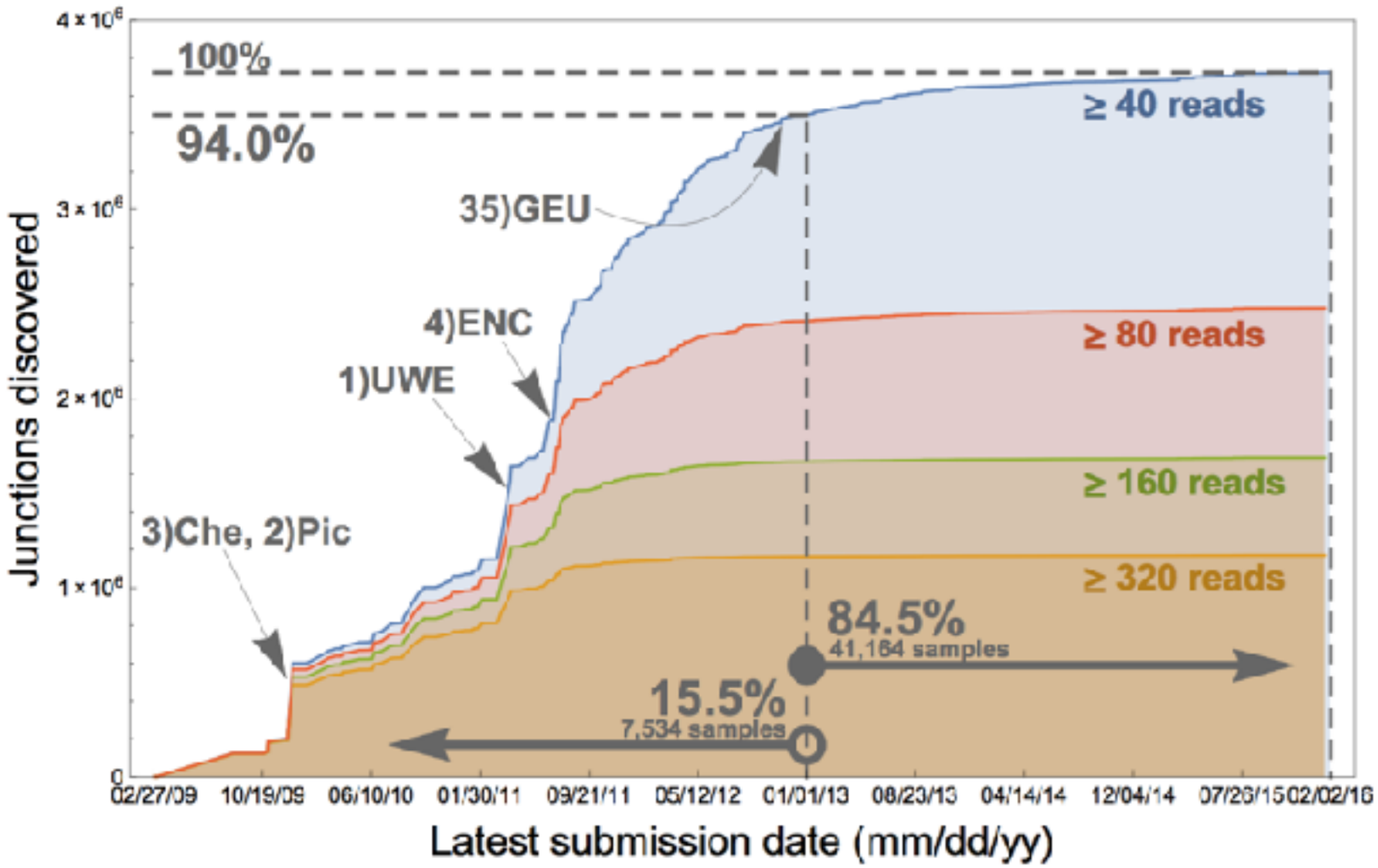
- As big as bigWigs & 1-2 orders of magnitude smaller than sorted BAMs
- Usable with Cufflinks, StringTie

Jacob Pritt

Pritt J, Langmead B. Boiler: lossy compression of RNA-seq alignments using coverage vectors. Nucleic Acids Res. 2016 Sep 19;44(16):e133.

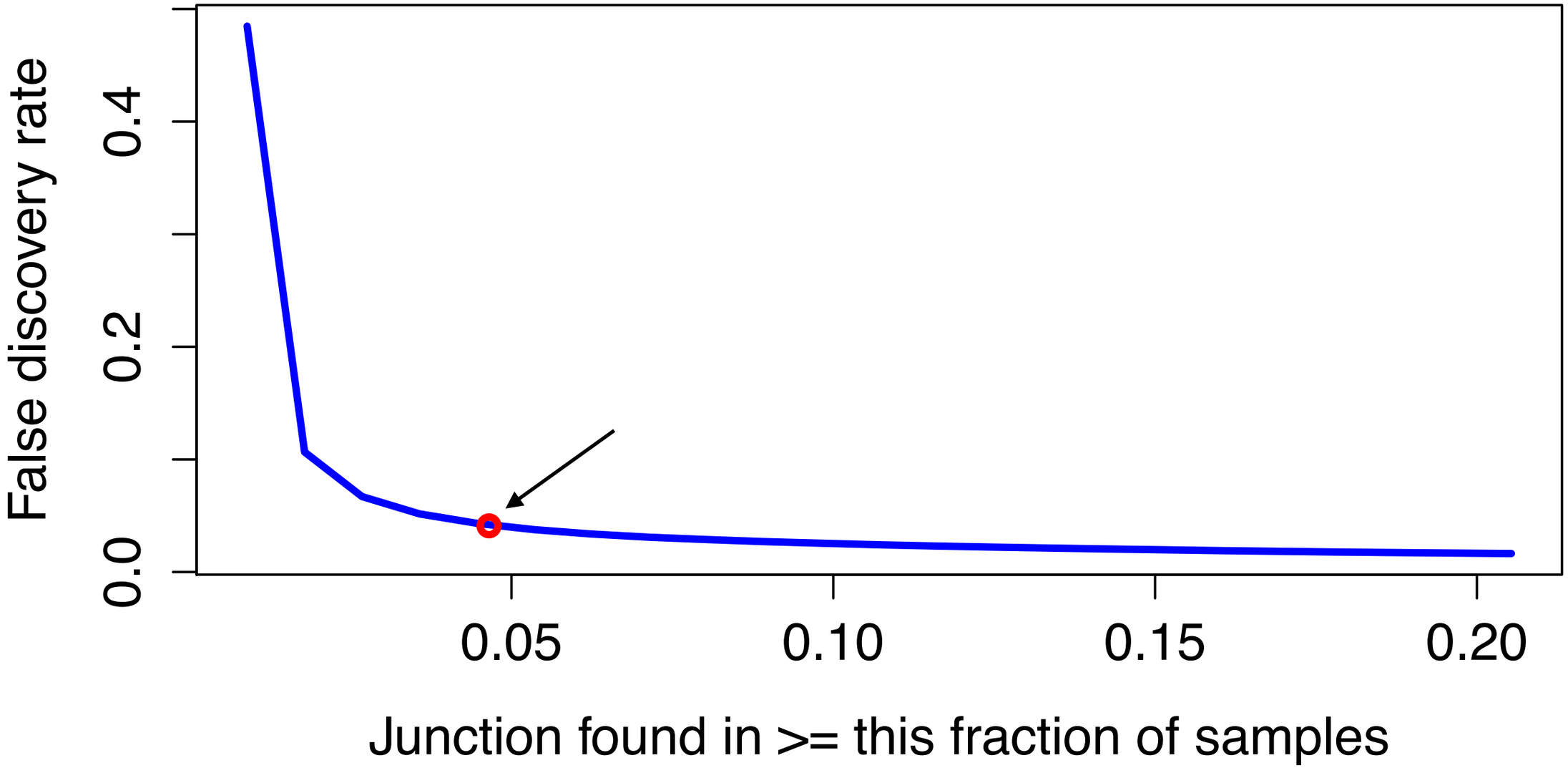
Intropolis

- Discovery of novel splicing events has leveled off
- Good time to put effort into a more complete annotation

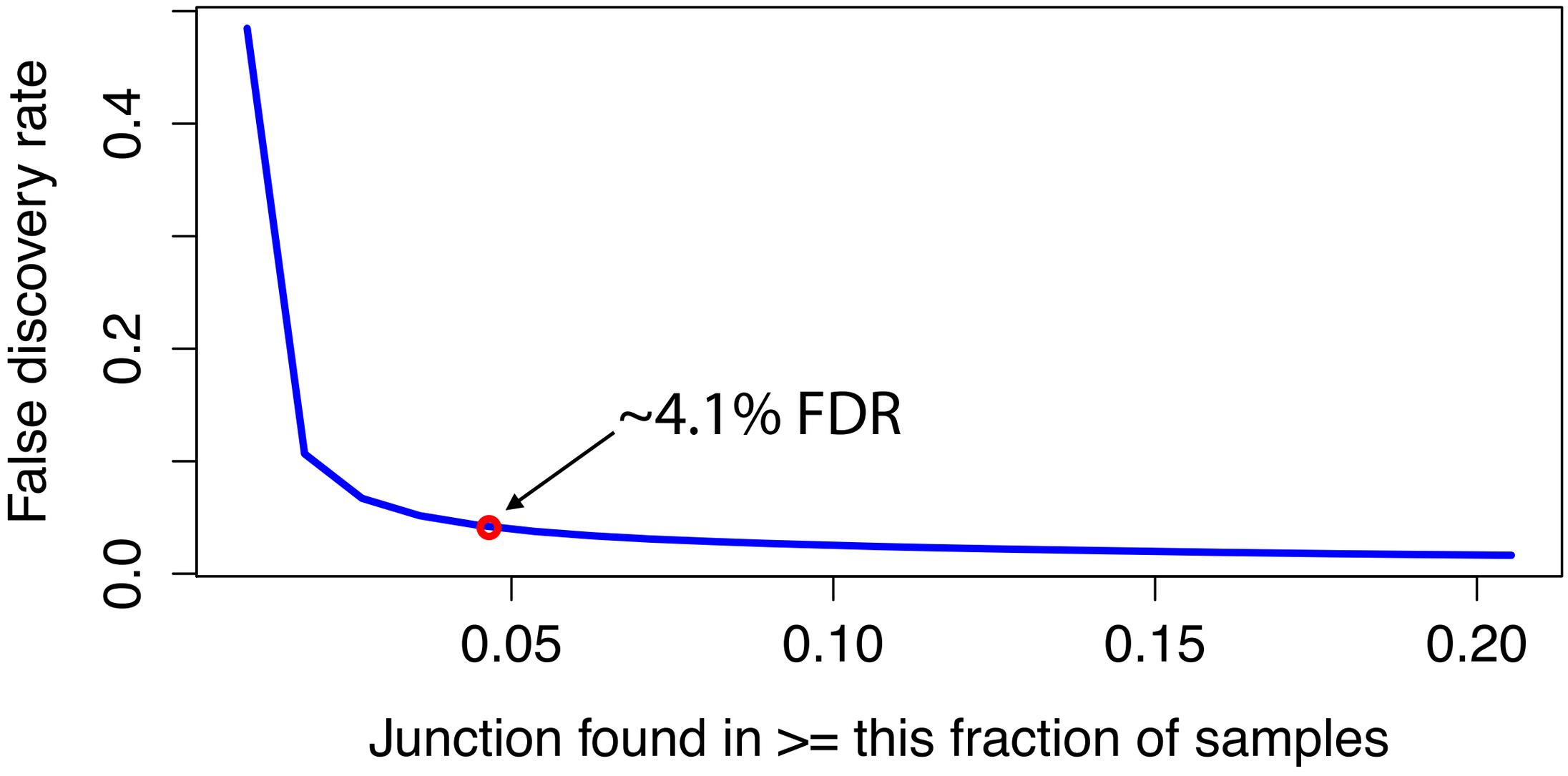


Nellore A, *et al*. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 2016 Dec 30;17(1):266.

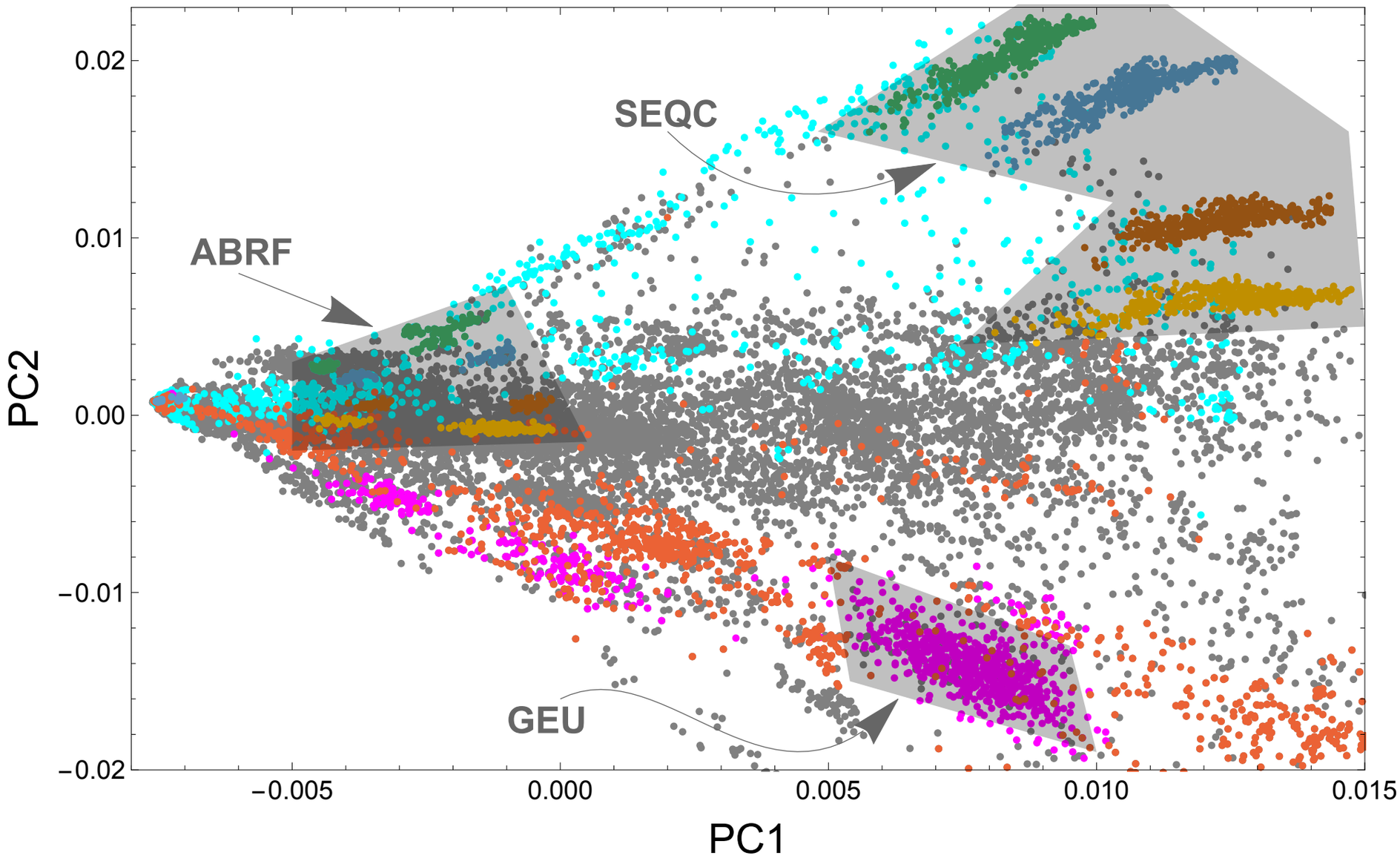
First−pass junction−call FDR estimated from 112 GEUVADIS−like simulations



First–pass junction–call FDR estimated from 112 GEUVADIS–like simulations



Intropolis



Nellore A, *et al*. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol. 2016 Dec 30;17(1):266.

Pass 1: align to genome, make junction calls

Reads:

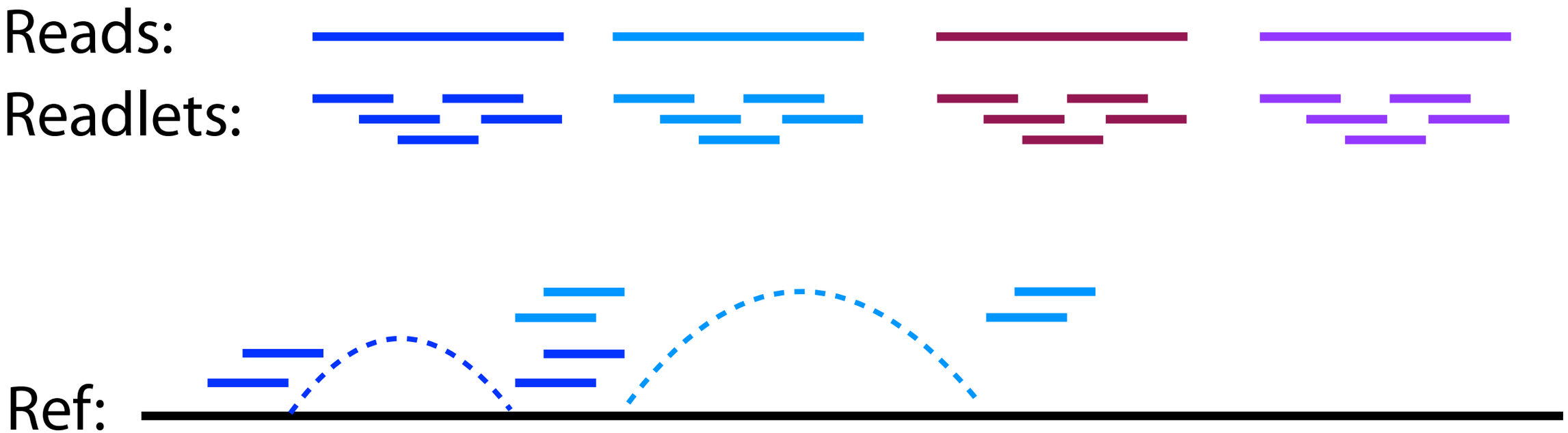
Pass 1: align to genome, make junction calls

Reads:

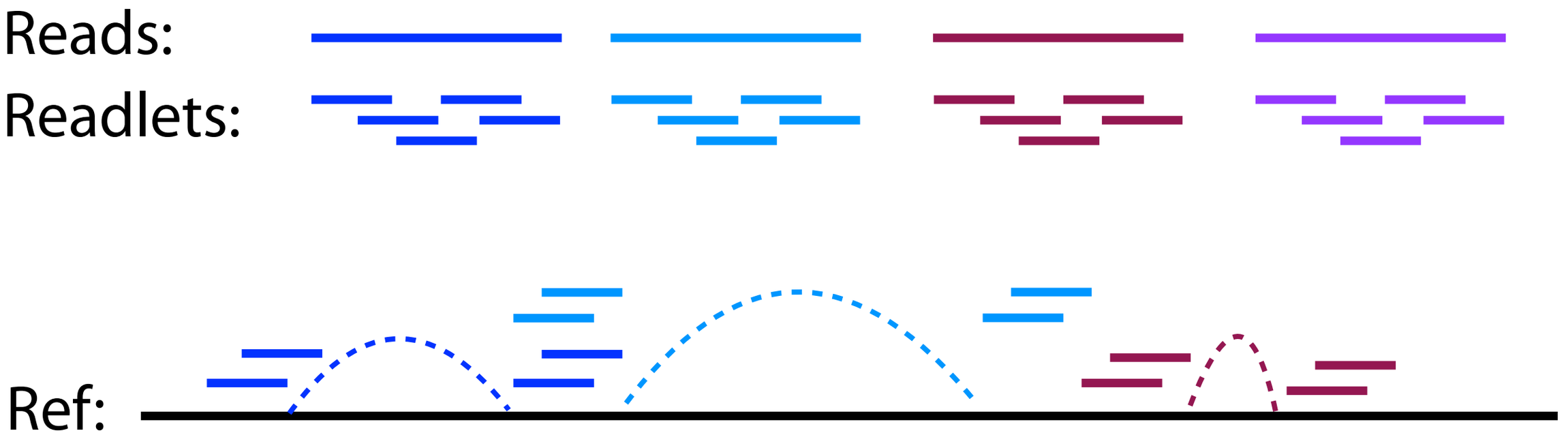
Readlets:

Pass 1: align to genome, make junction calls

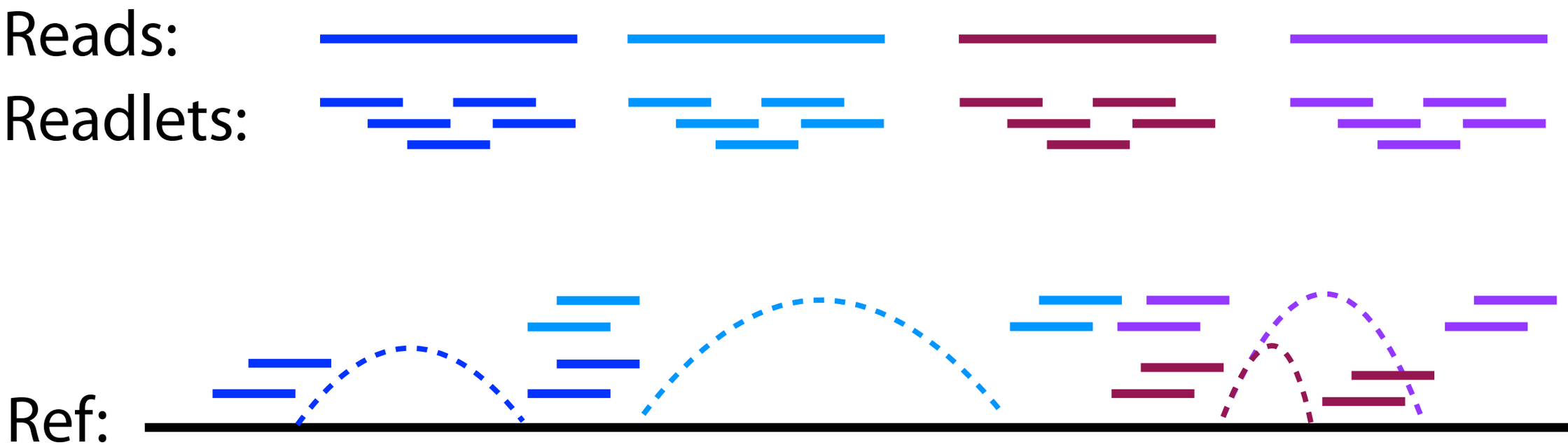
Pass 1: align to genome, make junction calls



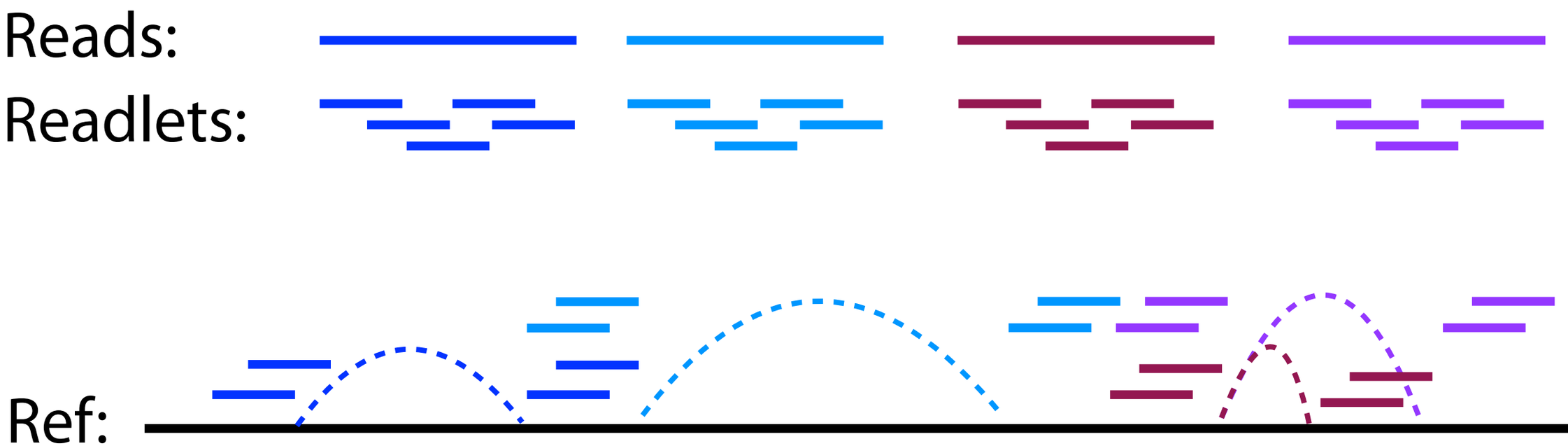
Pass 1: align to genome, make junction calls



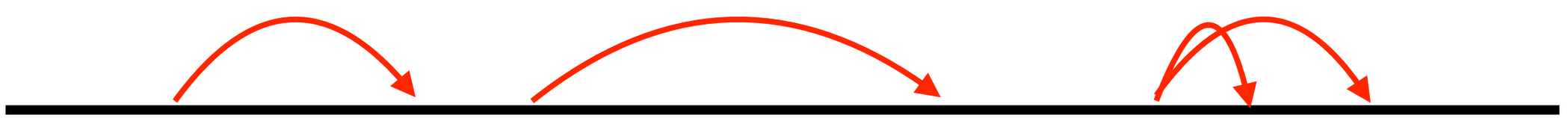
Pass 1: align to genome, make junction calls



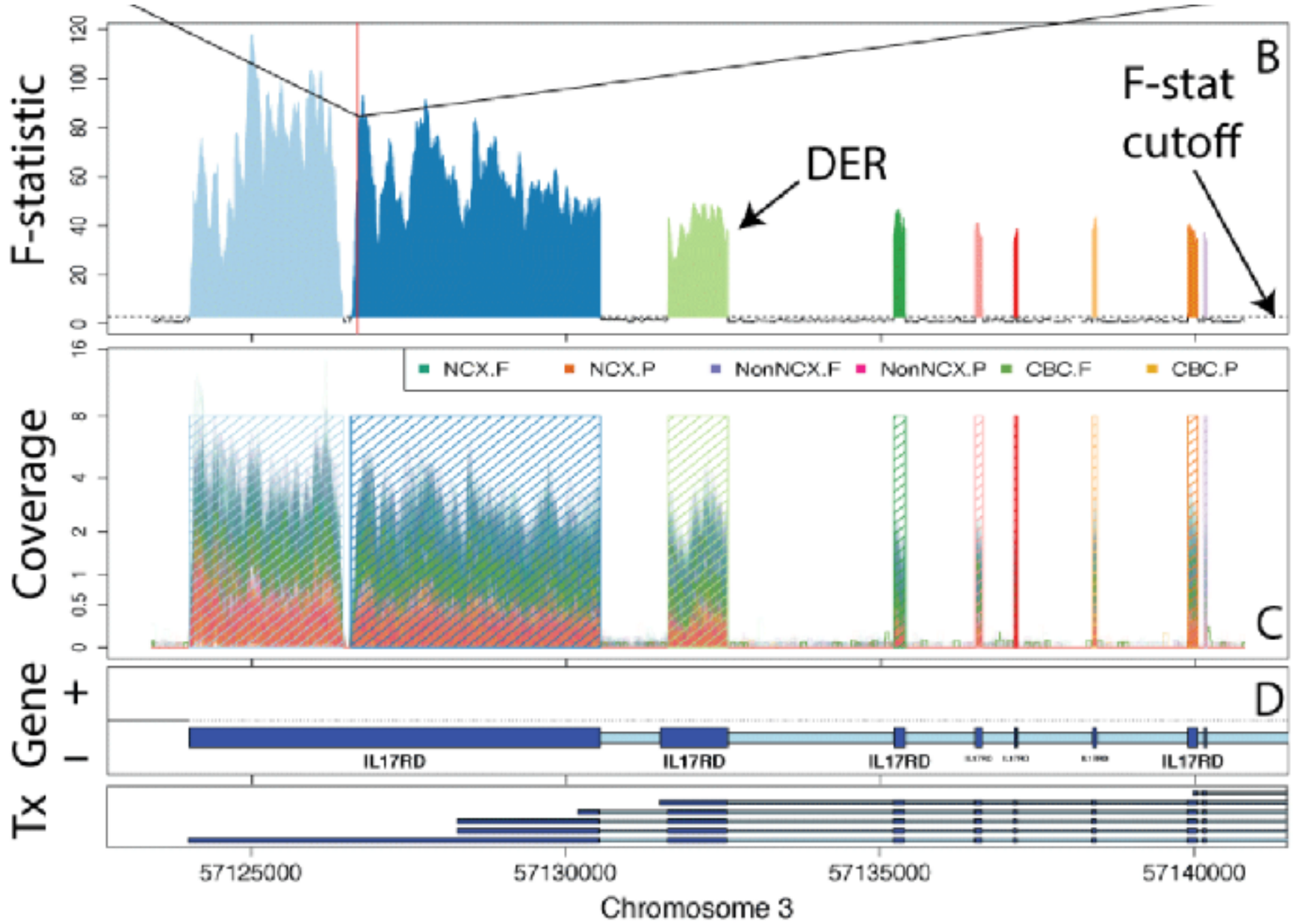
Pass 1: align to genome, make junction calls



Pass 2: re-align to genome *with putative junctions*



A third way



Collado-Torres L, Nellore A, Frazee AC, Wilks C, Love MI, Langmead B, Irizarry RA, Leek JT, Jaffe AE. Flexible expressed region analysis for RNA-seq with derfinder. Nucleic Acids Res. 2017 Jan 25;45(2):e9.

Indexing raw sequencing data

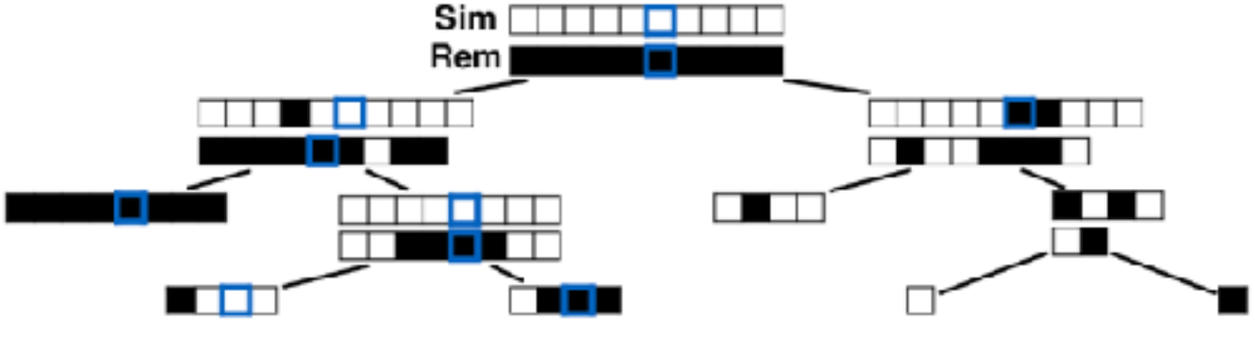


Image from Split SBT paper

Sequence Bloom Trees. Solomon B, Kingsford C. Fast search of thousands of short-read sequencing experiments. *Nat Biotechnol*. 2016 Mar;34(3):300-2.

Solomon B, Kingsford C. Improved Search of Large Transcriptomic Sequencing Databases Using Split Sequence Bloom Trees. J Comput Biol. 2018 Mar 12.

Sun C, Harris RS, Chikhi R, Medvedev P. AllSome Sequence Bloom Trees. *J Comput Biol*. 2018 May;25(5):467-479.

Mantis. Ferdman, M., Johnson, R., & Patro, R. Mantis: A Fast, Small, and Exact Large-Scale Sequence-Search Index. In *Research in Computational Molecular Biology* (p. 271). Springer.

BIGSI: Bradley, P., den Bakker, H., Rocha, E., McVean, G., & Iqbal, Z. (2017). Real-time search of all bacterial and viral genomic data. *bioRxiv*, 234955.

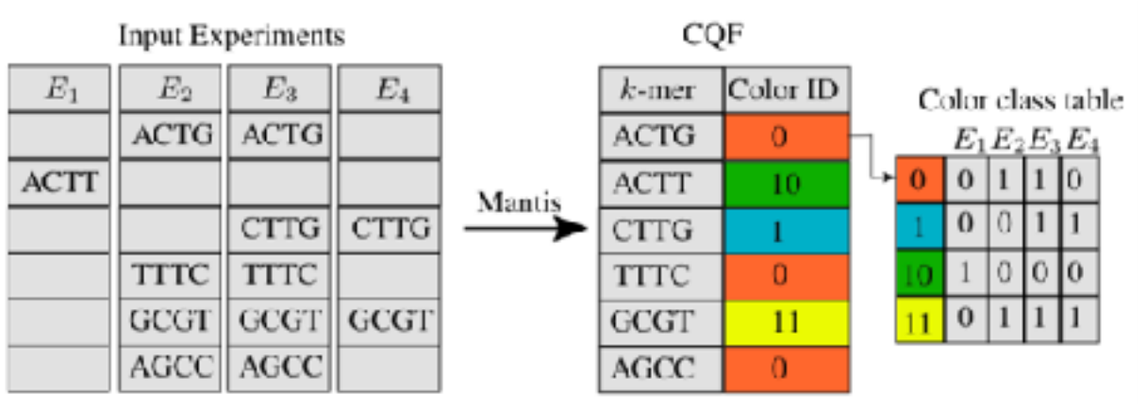
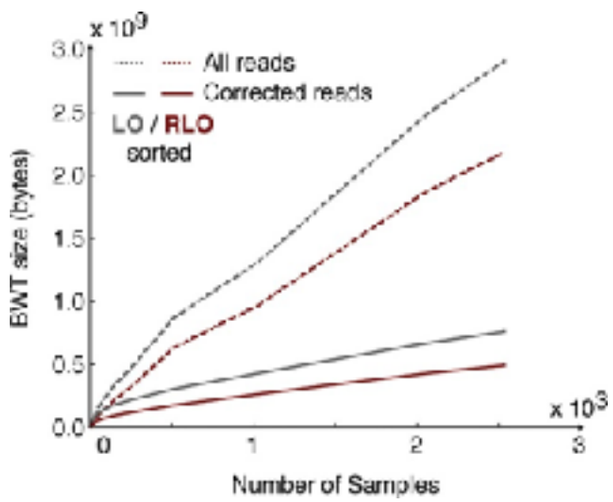
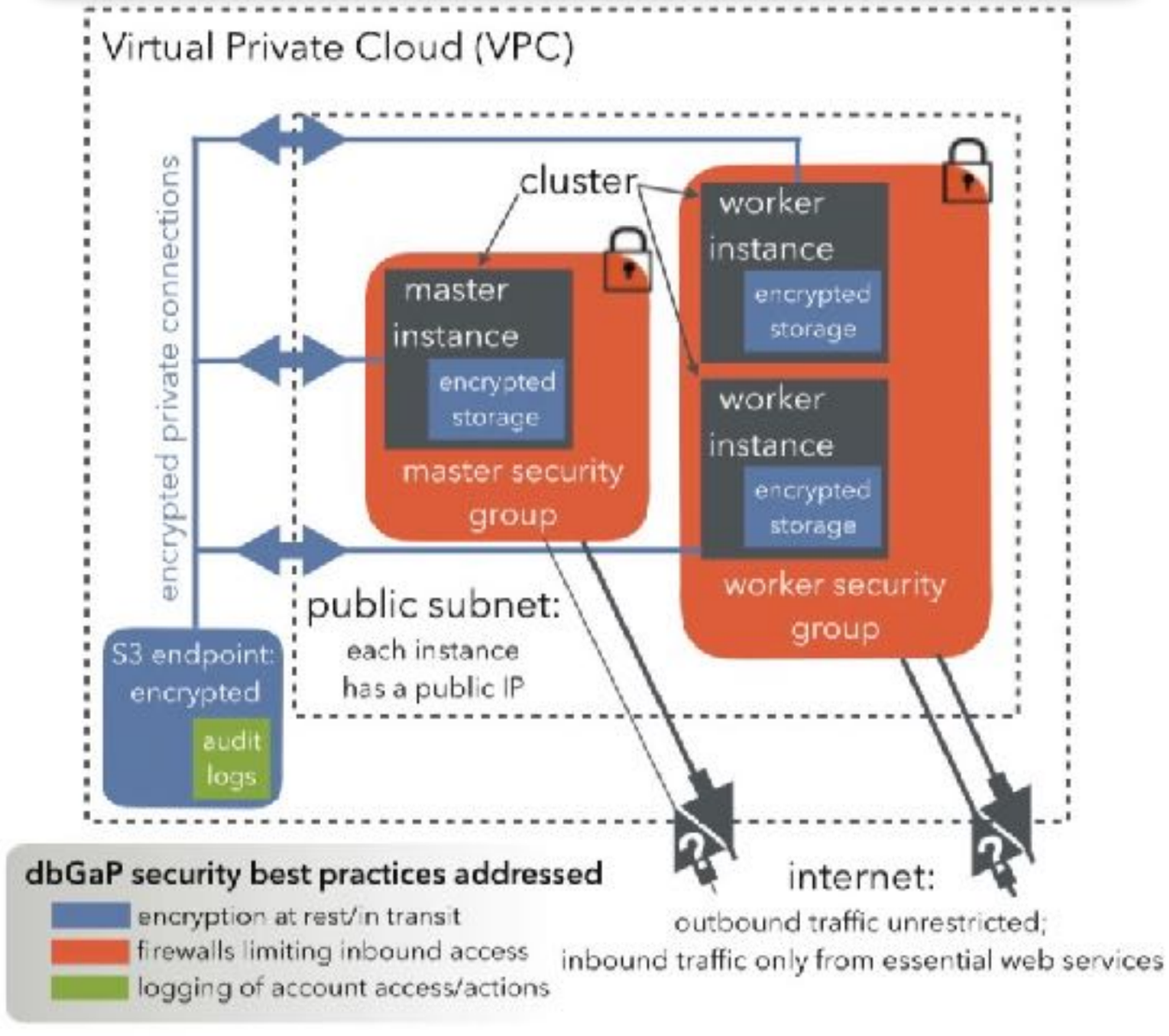


Image from Mantis paper



1000 Genomes FM Index: Dolle DD, Liu Z, Cotten M, Simpson JT, Iqbal Z, Durbin R, McCarthy SA, Keane TM. Using reference-free compressed data structures to analyze sequencing reads from thousands of human genomes. Genome Res. 2017 Feb;27(2):300-309.

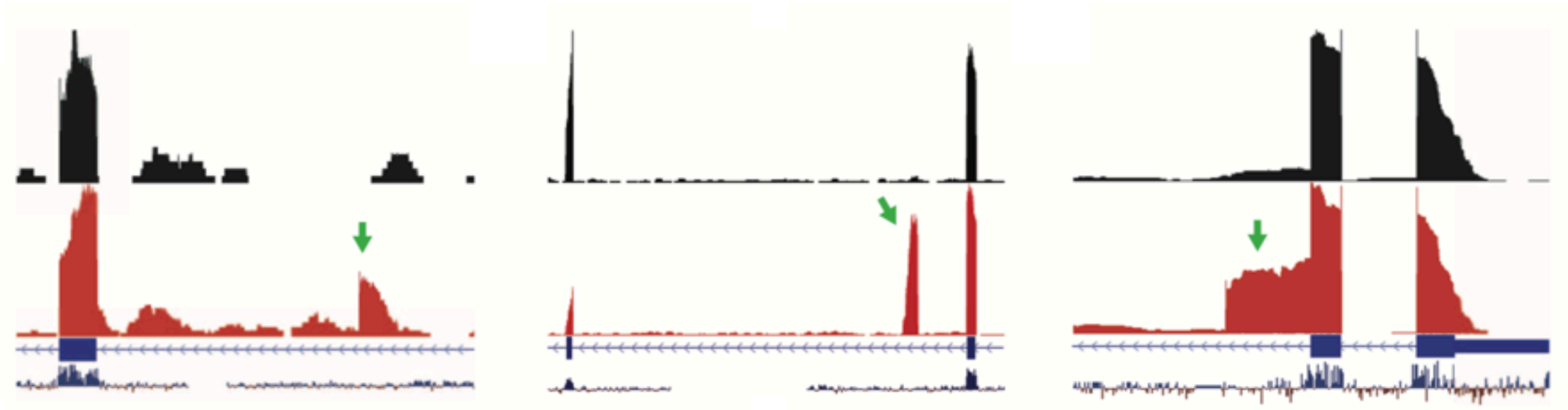
Rail-dbGaP



http://docs.rail.bio/dbgap/

Nellore A, Wilks C, Hansen KD, Leek JT, Langmead B. Rail-dbGaP: analyzing dbGaP-protected data in the cloud with Amazon Elastic MapReduce. *Bioinformatics*. 2016 Aug 15;32(16):2551-3.

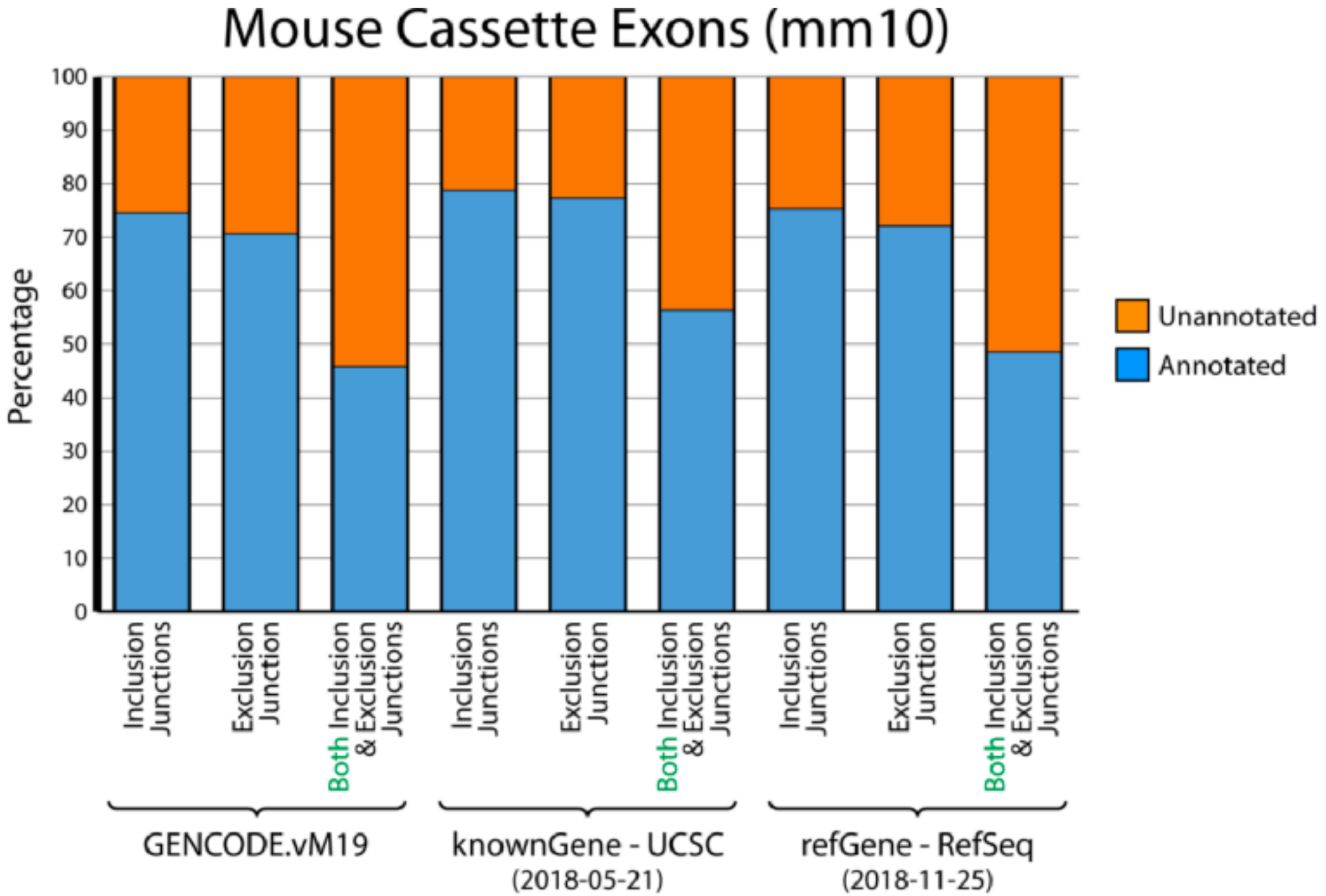
Splicing factors



Gene	Mouse (mm9) Genomic Sequence
Hgsnat	GUCAGUAAAUACACUCUGCUUAUAUUCGGGCAGUAUAAAAUGGUGAAG...ACAAGAUUGCUUUUAGUAAGAAAUAUGAUUGUGUCUCGUGUGUGUGUGUG
Cluh	CGGCCCUGCCACCUUCUGGUUCUCUUGUCCUAGACGUUCCUGUGCGGC...AGAGCGCAAGUUGGGGUGUGUGUGUGUGUGUGUGUUAUUGGGGUGGGUUGUC
Sars	AUAAAGGGAAGAGAUUUUACUUUUUUCAAACAGGGUCUCAUGUAGUCC...ACAGCAGGCUUAAAGGUAUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUU
Ptcd2	GUGGUCAAAUACCUACUGUUUUUGUCUGUGCAGUGACCUGAUUGCCCA...AGGAGUUUAUGAAGGGUGUGUGUGUGAGUGCAUGUGUGUUUGUGUGUGUG
Adnp2	GCAAAAAGAAAAGAAAAGUGACCUCUGUUUUAGGAGAAAAAGAAAGCA...UUCUCUUGACUGACAGUGAGUUUGUGUGUGUGUGUGUGUGUGUGUGUGUG
Usp15	UGUCUUAAUUUUUUUUCCUUCAUAAAAUAAAAGUAUCCUAGUGUACUC...AAACAGAAUUUAGAGGUGUGUGUGUGUGUGUGUGUGUCUGUGUGUUGU
Synj2bp	UGUGUGUGUGUGUGUGUGUGUGUAUGUGUGUAGGGUGAAAUGUACACA...UAAGACUUUUAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
Adipor2	UGUGUGUAUGACAUGUGUGUCUUAUGCUUUCAGAGGCUUGAAGAGGGU...AUUGUGAAUCACCAUGUAGGUGCUGGGACUUGAACCCAGGUUCUCUGUAA
Smg5	GUGUAUCUGUGUACGUGUGUGUCUGUGUUUCAGAAUGGAGAGAUGAAG...CUCCUCUGGCCUCAGGUGUGUGUGUGUGUGCUGUUUCAAAGCAUUUGCCC
A230046K03Rik	UGUCAUCUGAUUGUUGUGUGUGUGUUGCUAGAGAGAACCCAAACUG...AAGGGACAGUGACAGGUAAGGACAACAUAGUUGCCAGCAGUGUGGACCTT
Tecpr1	AGCCCCUUGCCAGCUCCUUCUUCCAUUUCCUAGGAGCAUUGUGGGUGU...GUGGAUAGACUGAUGGUGUGUGUCUAGAUGGUGGUGGUGGGUAUAUGAAU
Spata7	UGCUCAUGUUUCUAUUUGUGUUUAAAAUUUUAGGUGUCUAUAUUGGGG...CAAGUGUGUAUUCAGGUGCGUAUGAAUGUGUGUGUAUGCAUAUAGAAGUC
Wbscr22	UGACCAGAAUGAAGCCCGGAAAUACGUUCGCAAGUGAGGAGAGCAUGG...AAUUUGGGAGUGUGUGUGAGUGUGUGUGUGUGUGUGUGUGUGUGUGUG
Sptbn4	UUUUUUUUUUUUUUUUUUUCCUGGAGGGAGGAGGAGGAUCUUGUGACU...AUUACCUUAGCUUGGGUGUGUGUGUGUGUGUGUGUGUGUGUGUGUGCU
	5′ Upstream (Intron) — 5′ Cryptic Exon ··· 3′ Cryptic Exon — 3′ Downstream (Intron)

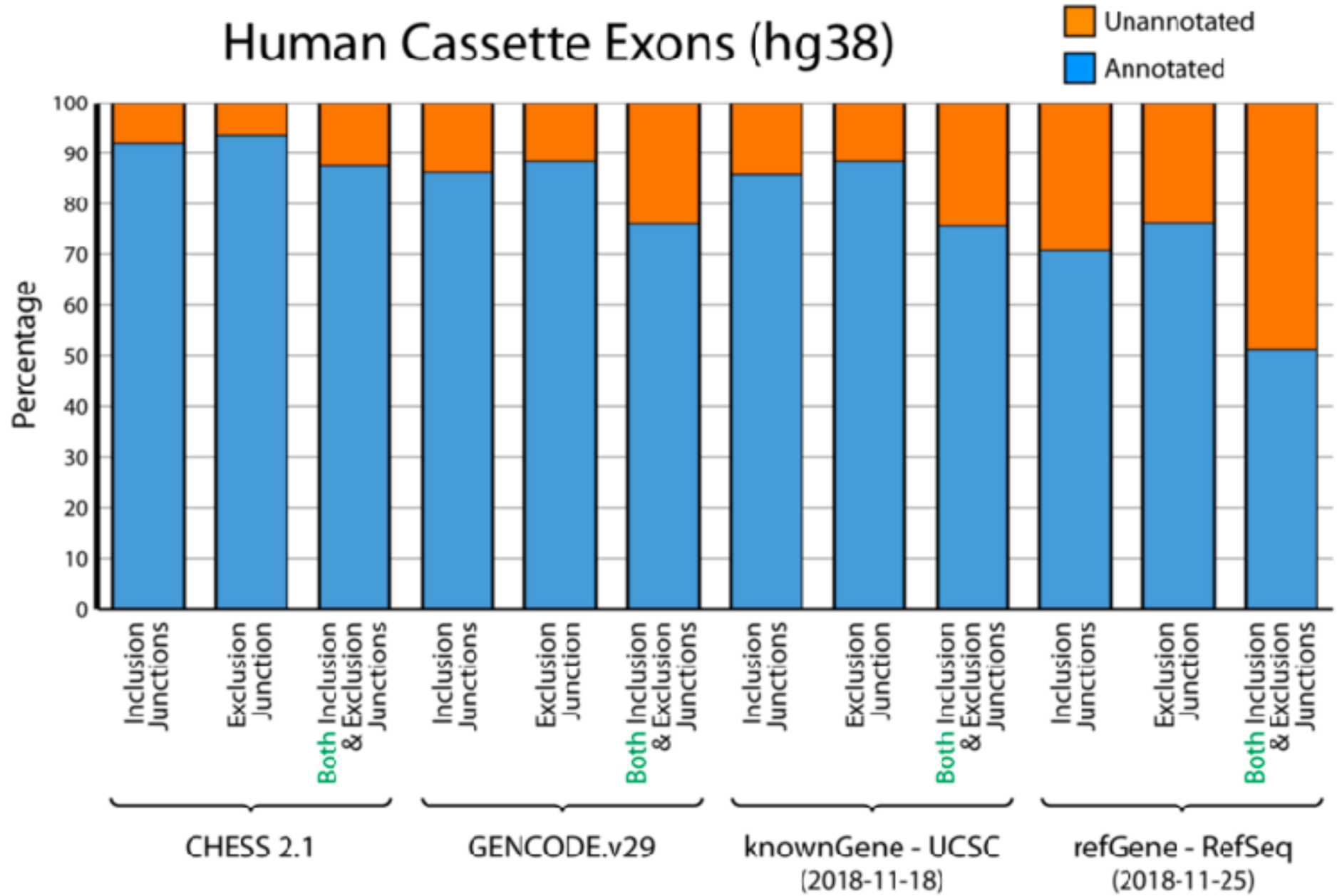
Ling JP, Pletnikova O, Troncoso JC, Wong PC. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*. 2015 Aug 7;349(6248):650-5.

Splicing annotation



Ling JP, Wilks C, Charles R, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, Blackshaw S. ASCOT identifies key regulators of photoreceptor-specific splicing. bioRxiv doi:*10.1101/501882.*

Splicing annotation



Ling JP, Wilks C, Charles R, Ghosh D, Jiang L, Santiago CP, Pang B, Venkataraman A, Clark BS, Nellore A, Langmead B, Blackshaw S. ASCOT identifies key regulators of photoreceptor-specific splicing. bioRxiv doi:*10.1101/501882.*