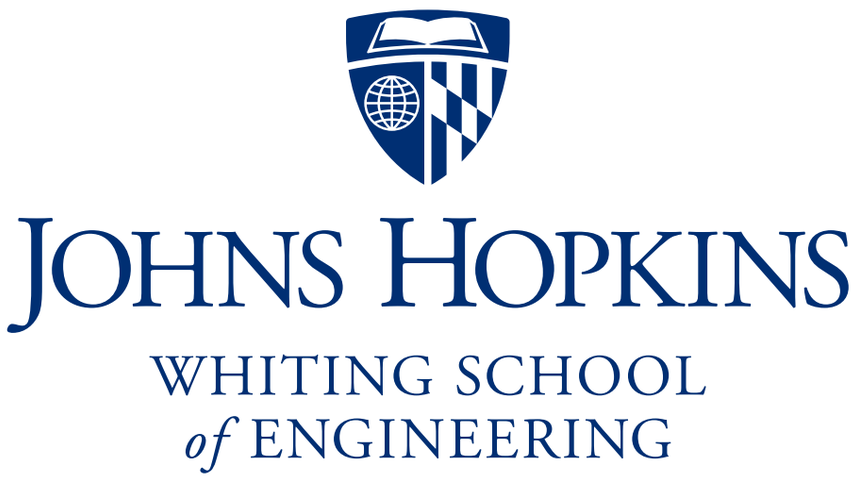
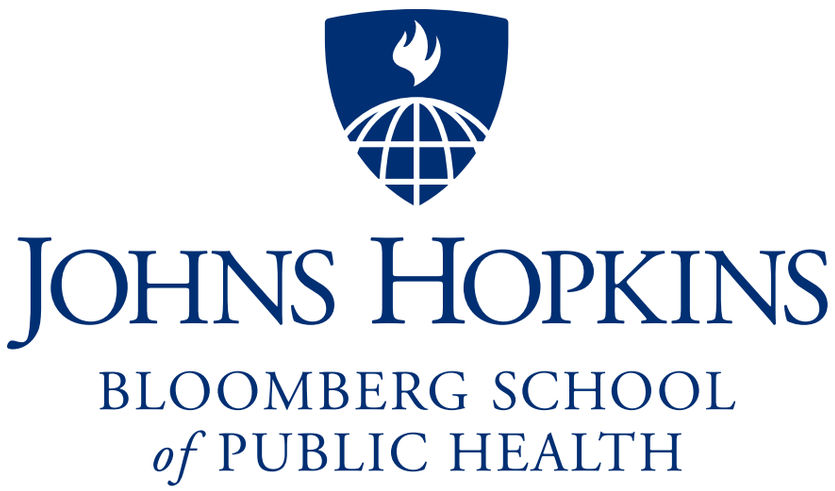
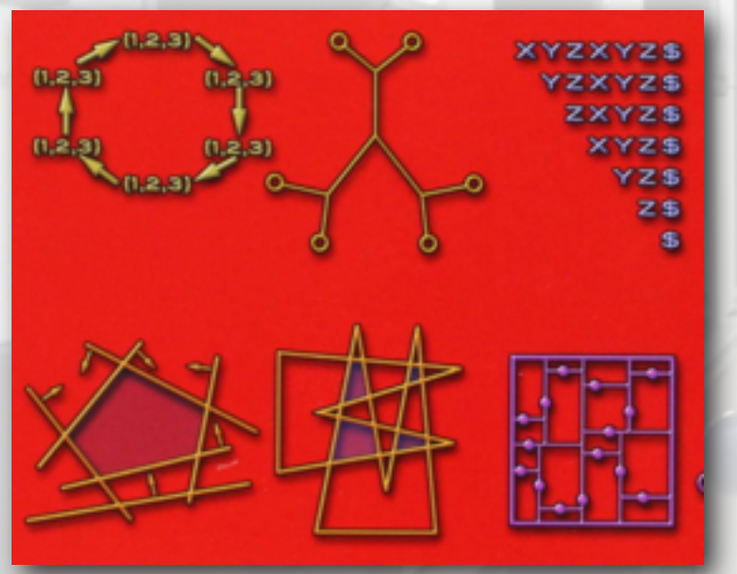
Navigating tens of thousands of RNA-seq datasets with recount, SciServer & Jupyter

Ben Langmead
Assistant Professor, Computer Science
langmea@cs.jhu.edu

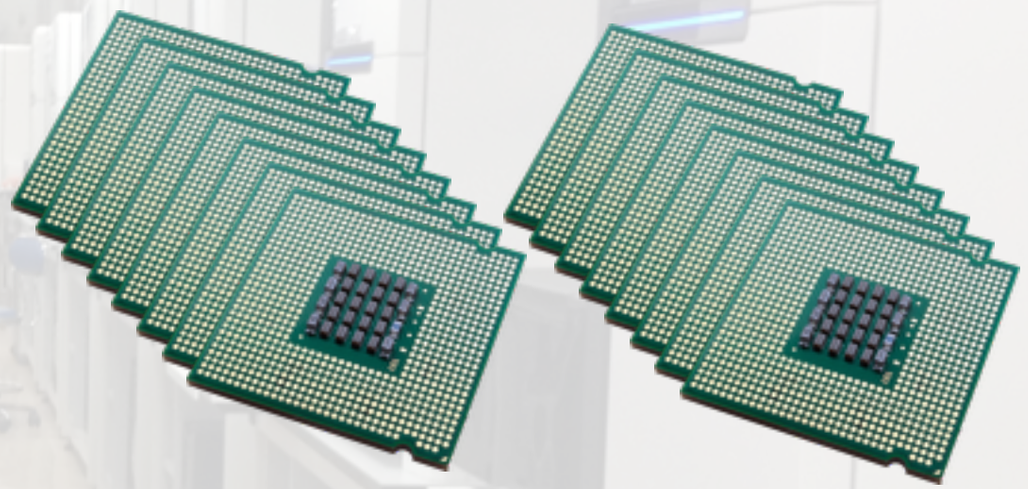
IDIES Symposium, October 21 2016






(1,2,3)
(1,2,3)
(1,2,3)
(1,2,3)
(1,2,3)
(1,2,3)
XYZXYZ$
YZXYZ$
ZXYZ$
XYZ$
YZ$
Z$
$

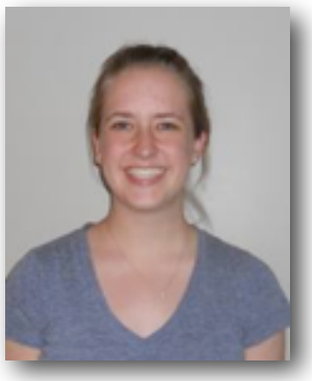
+



Abhinav Nellore

Chris Wilks

Jacob Pritt



Alyssa Frazee

Leo Collado Torres

Kasper Hansen

Jeff Leek

Margaret Taub

Shannon Ellis

Kai Kammers

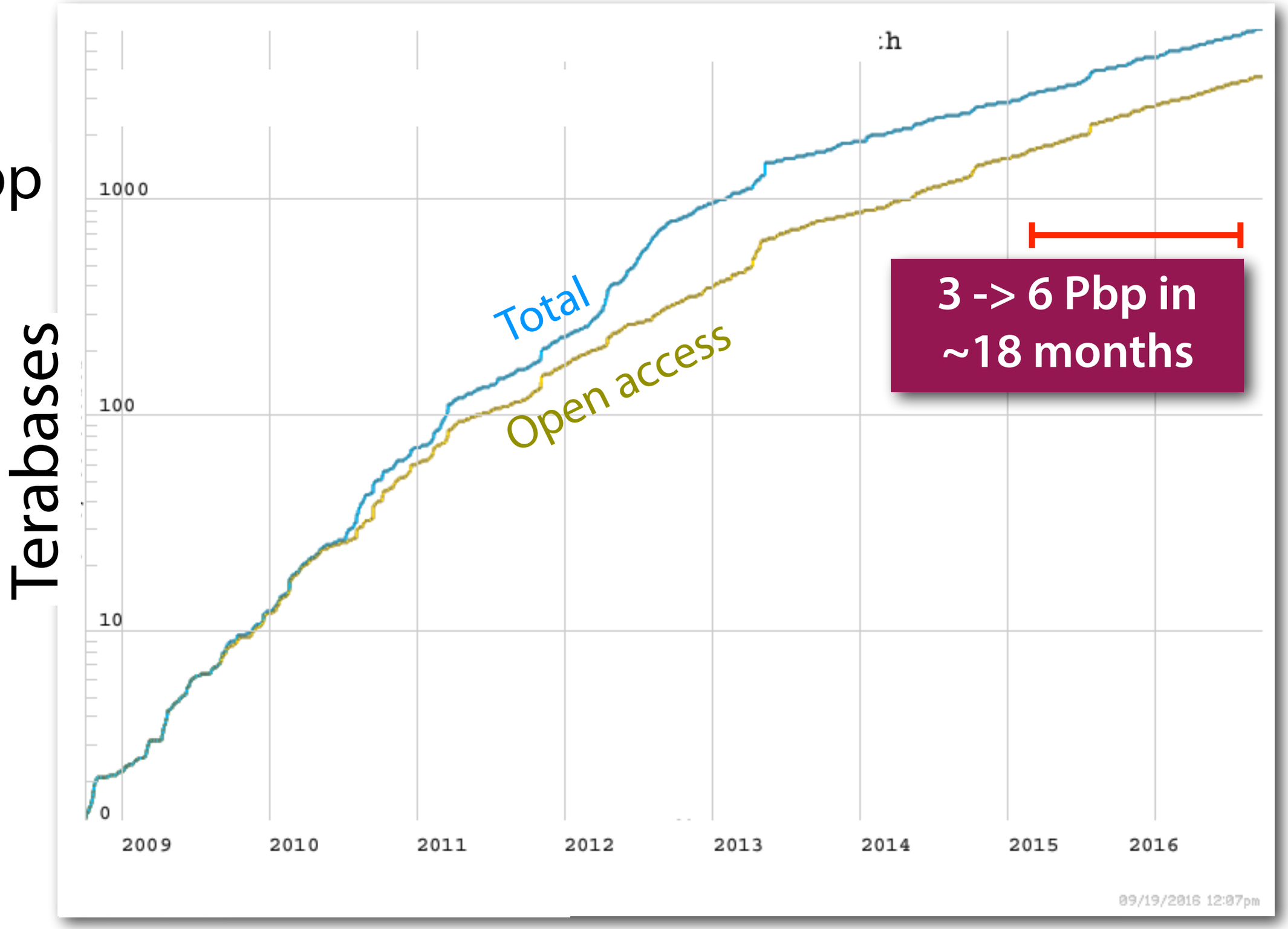
Jamie Morton

José Alquicira-Hernández

Andrew Jaffe

Rail-RNA and recount teams

Sequence Read Archive (SRA) growth




hadoop

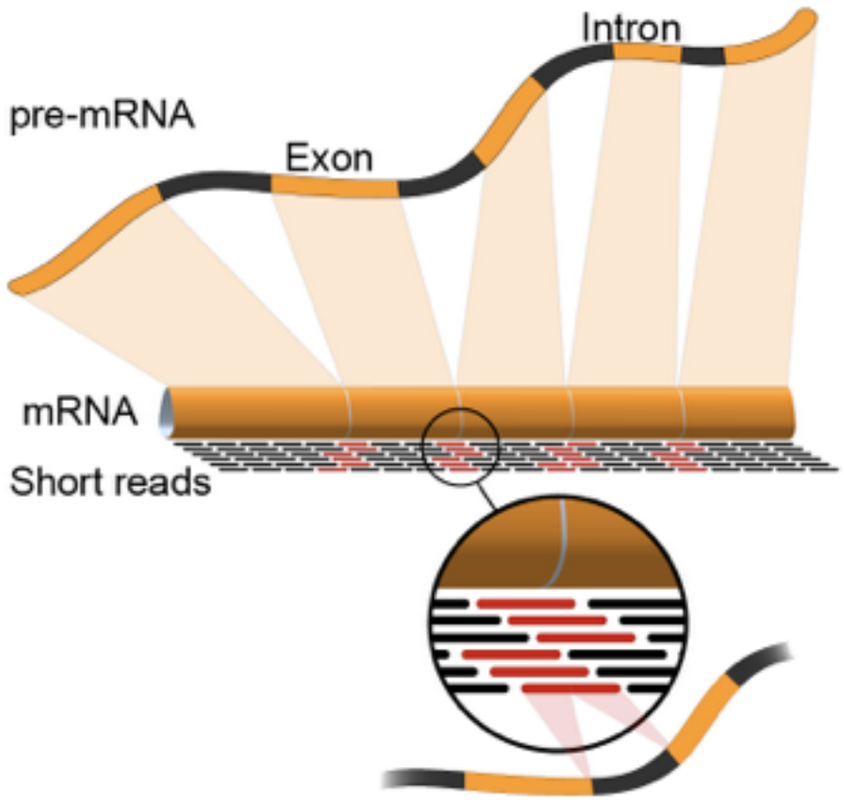

amazon
web services™

RA database
1000
100
10
0
Size, terabases
2009
2010
2011
2012
2013
2014
2015
2016
Total bases
Open access bases
09/19/2016 12:07pm

AMAZON S3
EMR
AMAZON S3
Elastic MapReduce

Rail-RNA

Thank you: IDIES Seed grant



Abhinav Nellore

Jeff Leek

Website: http://rail.bio, Paper: http://bit.ly/rail-aa

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*. 2016 Sep 4.

Rail-RNA

- Analyzed **~50,000** human RNA-seq samples with Rail-RNA; about **150 Tbp**
- Rapid: input to results in 2 weeks
- Repeatable: http://github.com/nellore/runs

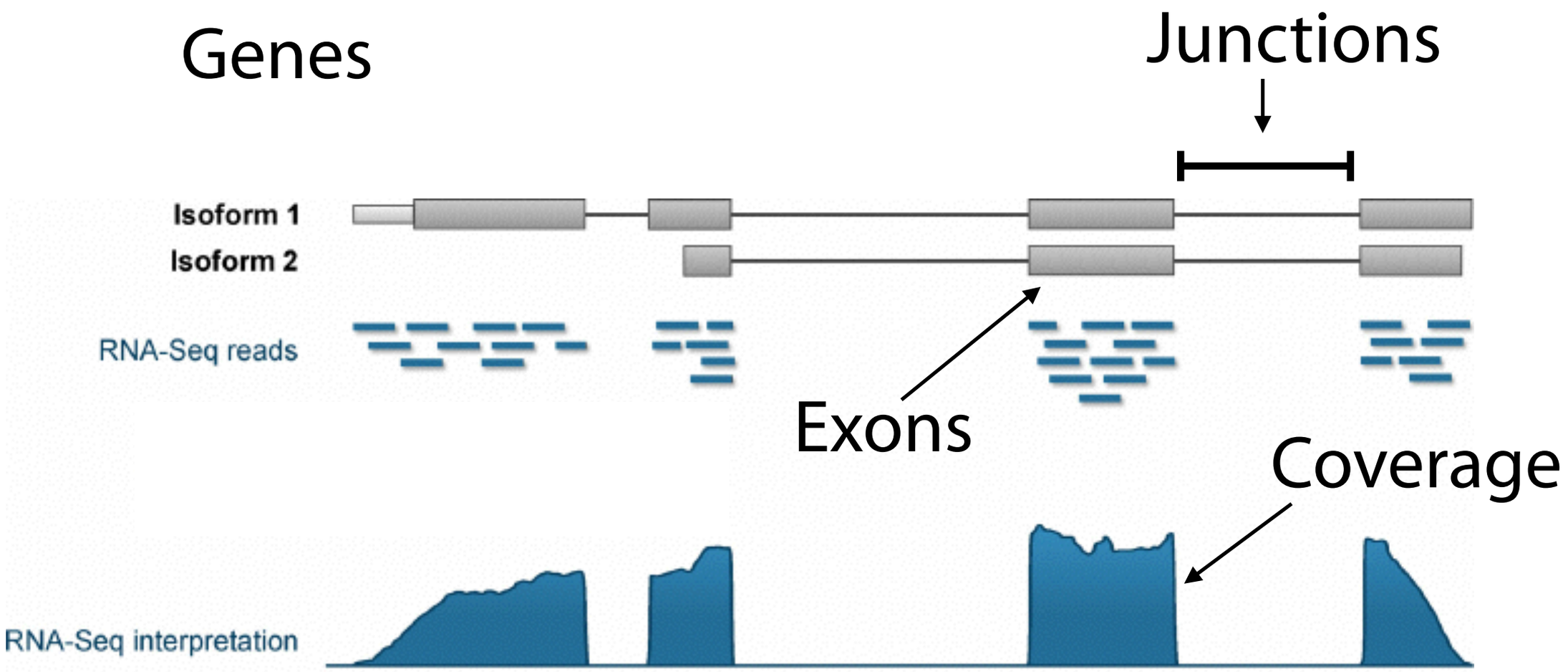
  (Exact commands we used to run on AWS)
- Inexpensive: ~ $1.40 / sample

  (Compare to sequencing costs)

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*. 2016 Sep 4.

recount

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



Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. **recount: A large-scale resource of analysis-ready RNA-seq expression data**. bioRxiv doi: 10.1101/068478.

recount

- Shiny-app front-end:

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

- Over 6 TB of data hosted at SciServer
- SciServer Compute lets users to work with locally-hosted data in Jupyter notebook

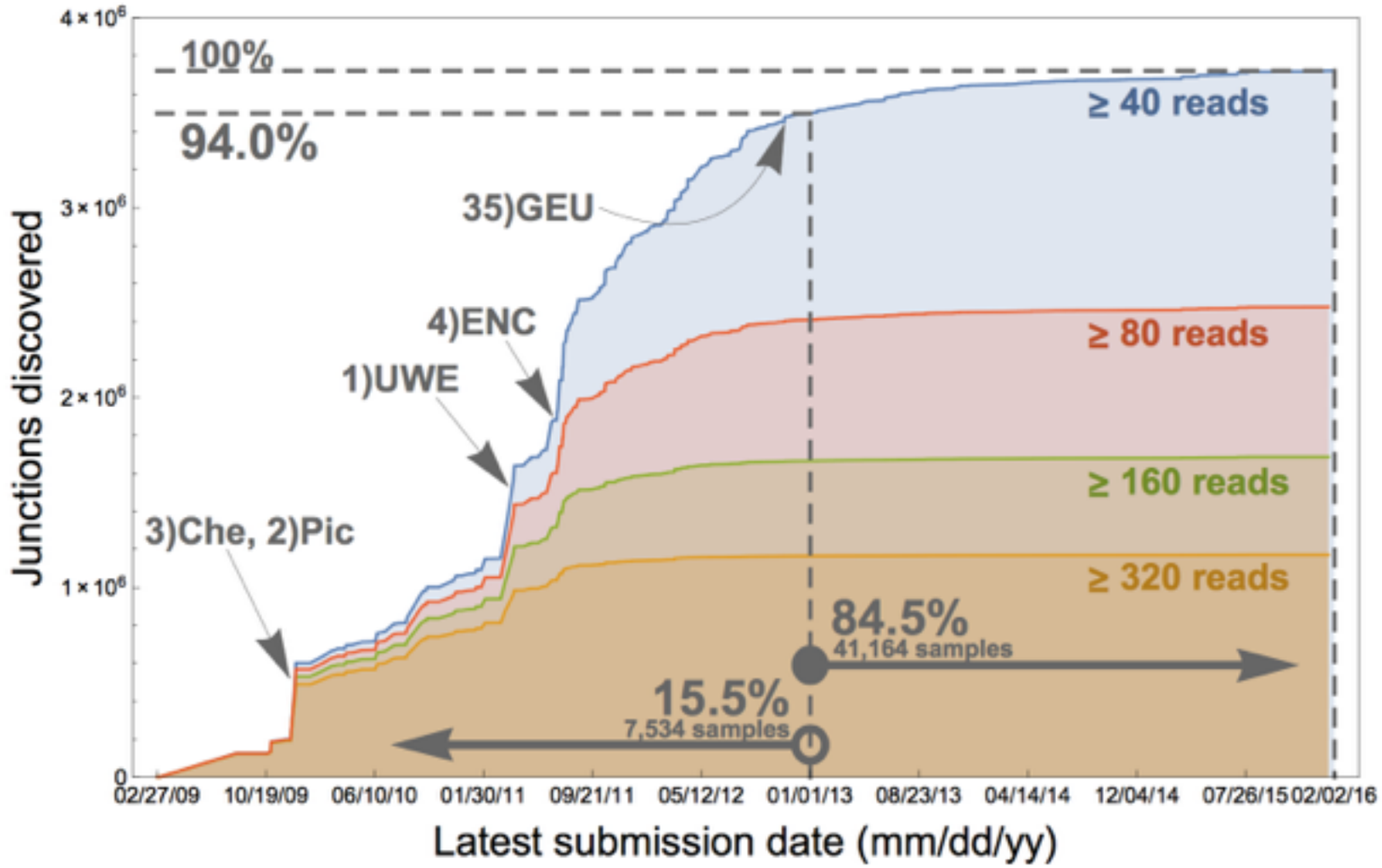
  http://compute.sciserver.org/dashboard/

- Preprint & Bioconductor 3.4 package available

Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. **recount: A large-scale resource of analysis-ready RNA-seq expression data**. bioRxiv doi: 10.1101/068478.

recount

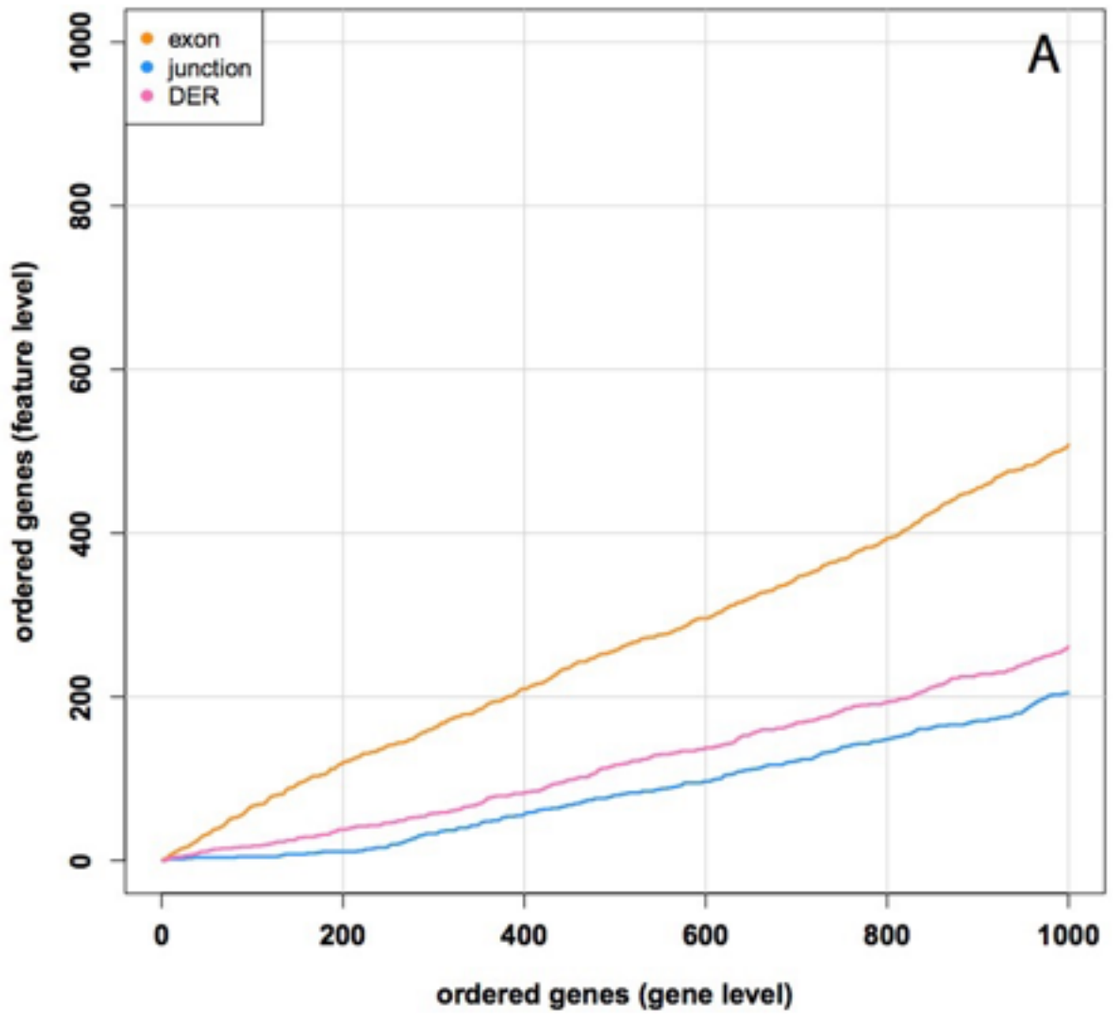
- Discovery of novel splicing events has leveled off

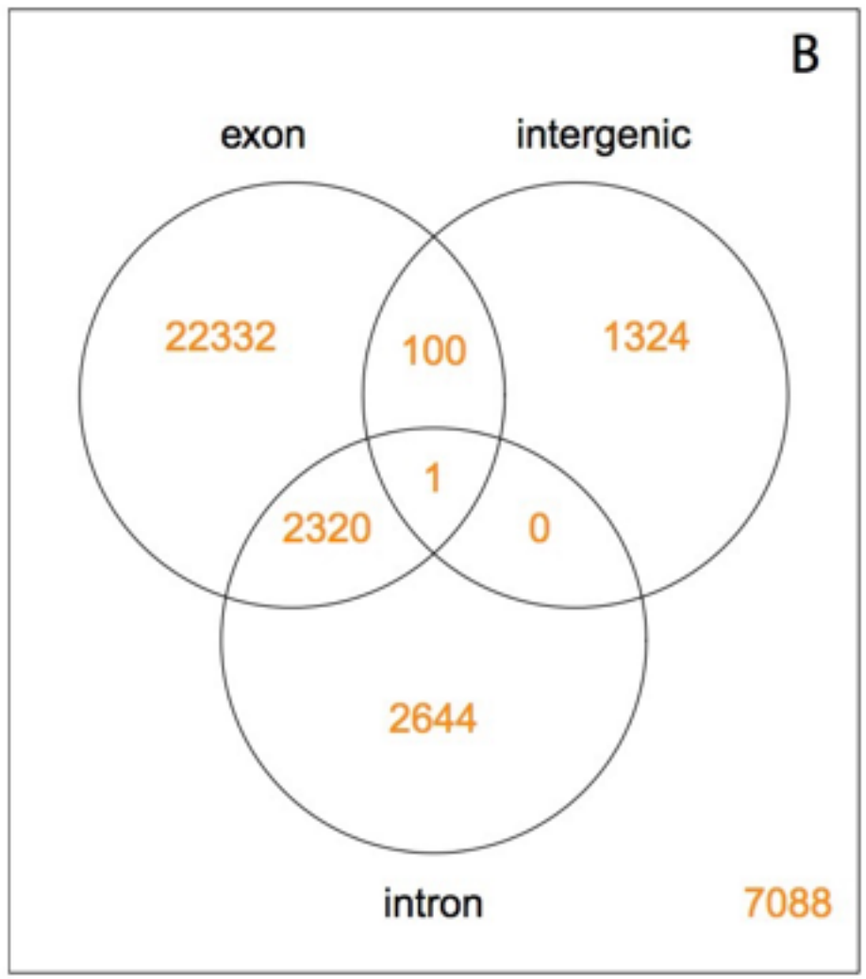


Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. **recount: A large-scale resource of analysis-ready RNA-seq expression data**. bioRxiv doi: 10.1101/068478.

recount

- Distinct summaries tell complementary stories about differential expression

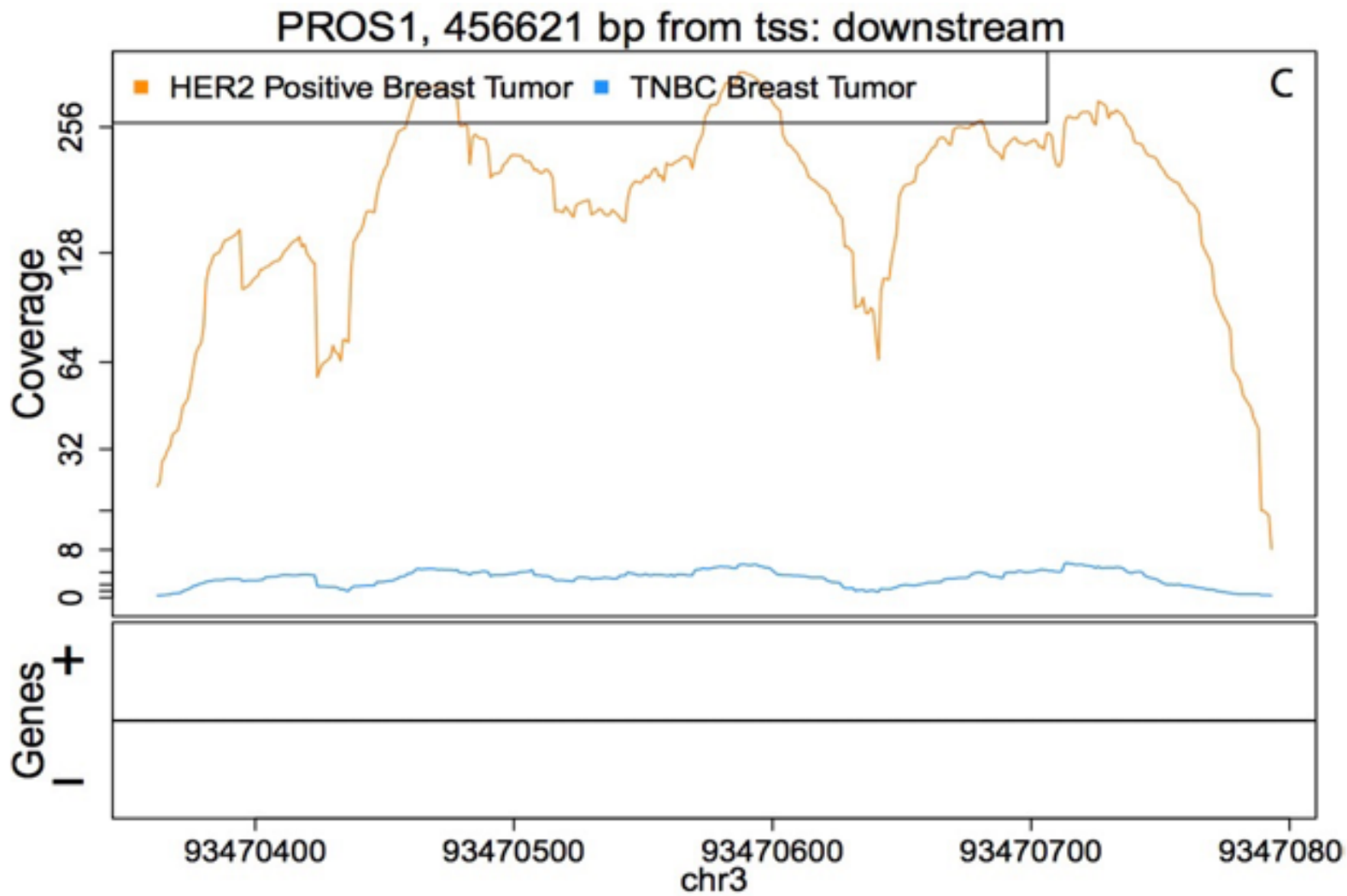




Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. **recount: A large-scale resource of analysis-ready RNA-seq expression data**. bioRxiv doi: 10.1101/068478.

recount

- Some differential expression is outside of any known-transcribed area



Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, Jaffe AE, Langmead B, Leek JT. **recount: A large-scale resource of analysis-ready RNA-seq expression data**. bioRxiv doi: 10.1101/068478.

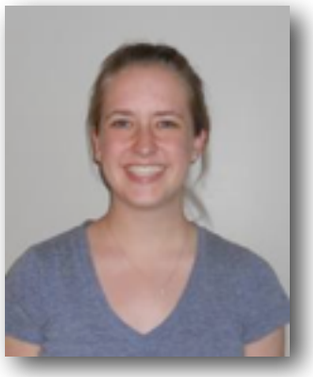
Brief demo

Abhinav Nellore

Leo Collado Torres

Chris Wilks

Jacob Pritt



Alyssa Frazee

Kasper Hansen

Jeff Leek

Margaret Taub

Shannon Ellis

Kai Kammers

Jamie Morton

José Alquicira-Hernández

Andrew Jaffe

- NIH R01GM118568
- NSF CAREER IIS-1349906
- Sloan Research Fellowship
- IDIES Seed Funding program
- Amazon Web Services

Thank you:

IDIES Seed funding

SciServer

SciServer Compute

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

langmead-lab.org, @BenLangmead