Scalable Software for Analyzing Large Collections of RNA Sequencing Data

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

IDIES Symposium
Johns Hopkins University, October 17, 2014

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Improving sequencers

Source: www.illumina.com

GA II
1.6 billion nt/day
(2008)

GA IIx
5 billion nt/day
(2009)

HiSeq 2000
25 billion nt/day
(2010)

HiSeq 2500
120 billion nt/day
(2012)

HiSeq 2000
75 billion nt/day
(2011)

HiSeq X
600 billion nt/day
(2014)

1 Finding Clues in Genes of "Exceptional Responders", by Gina Kolata. New York Times, October 8, 2014.

2 This Bizarre Organism Builds Itself a New Genome Every Time It Has Sex, by Greg Miller. Wired, September 17, 2014.

3 Fighting Poisons With Bacteria, by Carina Storrssept. New York Times, September 15, 2014.

4 Studying Ebola, Then Dying From It, by Pardis Sabeti. New York Times, September 5, 2014.

5 Tuberculosis Is Newer Than Thought, Study Says, by Carl Zimmer. New York Times, August 20, 2014.

6 Cancer and the Secrets of Your Genes, by Theodora Ross. New York Times, August 16, 2014.

7 One of a Kind, By Seth Mnookin. The New Yorker, July 21, 2014.

8 Searching for Answers in Very Old DNA, by Claudia Dreifus. New York Times, June 23, 2014.

More at: http://www.cs.jhu.edu/~langmea/poppress.shtml

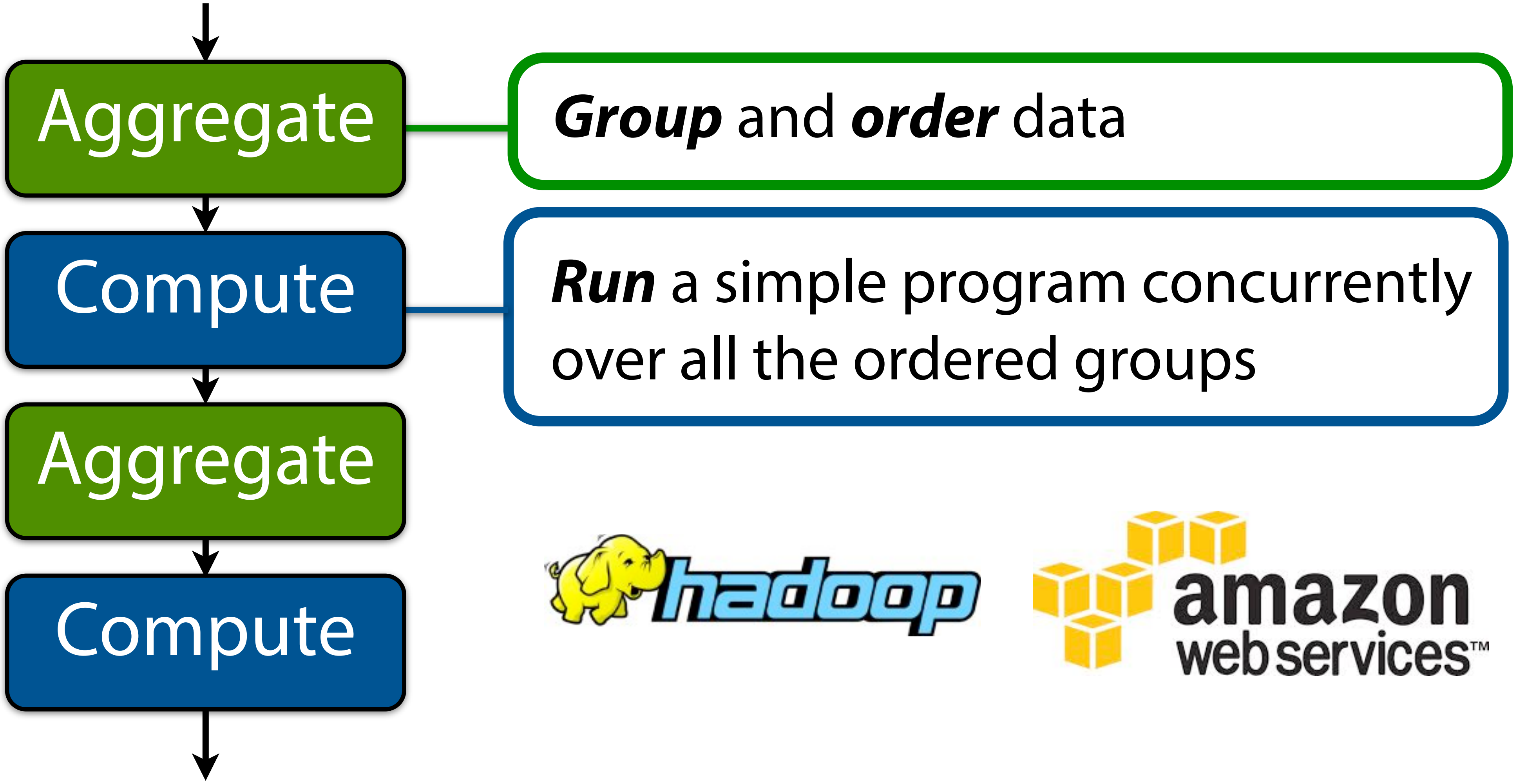
Big projects

Study	**Approx # samples**
ENCODE	100
GEUVADIS	465
Depression Genes Network	950
TCGA	>2,000
GTEx	>10,000

Why study big public datasets?

- To make discoveries missed by original authors
- To combine datasets in new ways
- To add power to a smaller experiment
- As proving ground for new methods

MapReduce: aggregate, compute, repeat



Aggregate

Compute

Aggregate

Compute

Group and ***order*** data

Run a simple program concurrently over all the ordered groups

hadoop

amazon web services™

Myrna

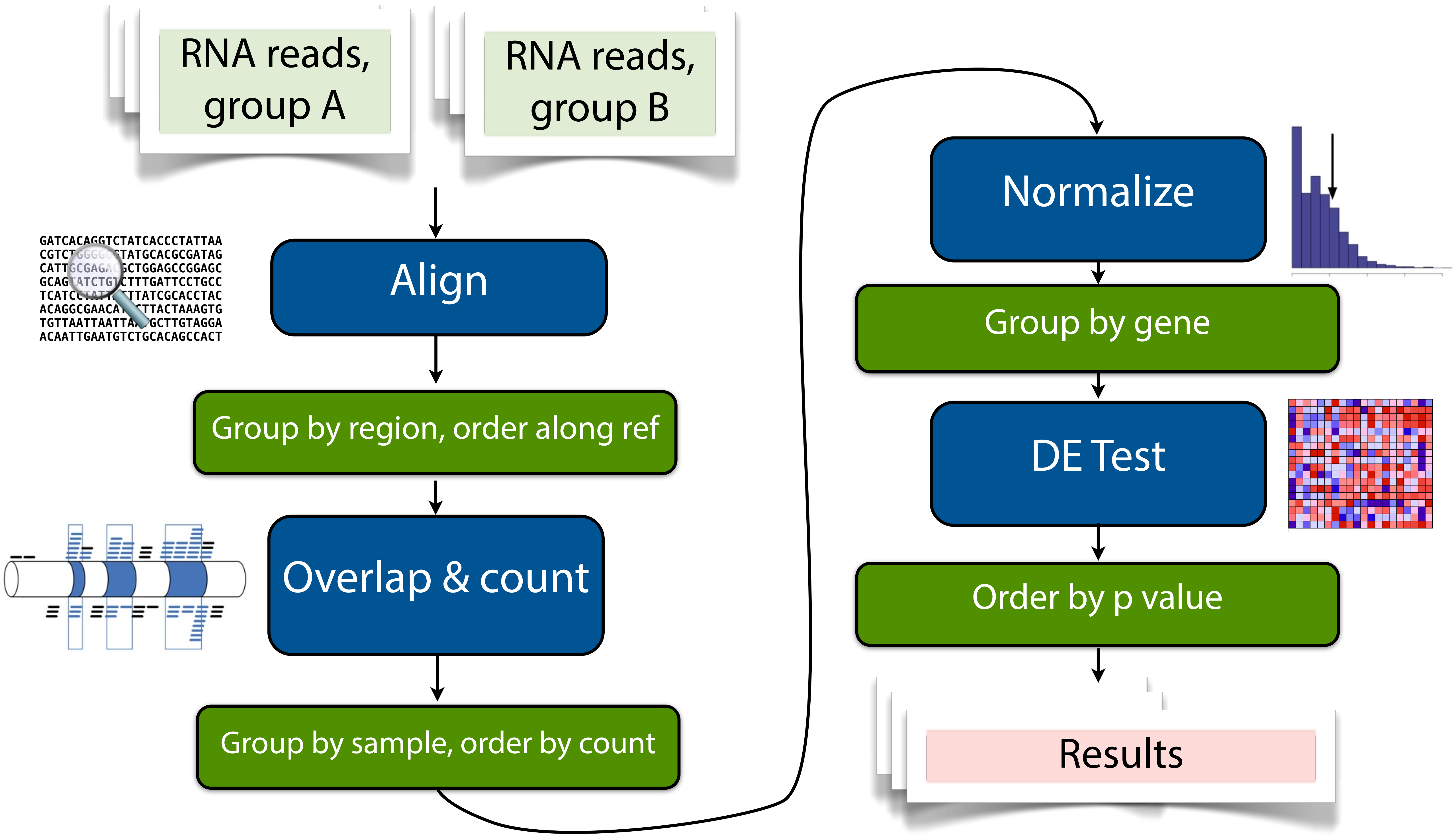
Jeff Leek

Kasper Hansen

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Langmead B, Hansen KD, Leek JT. **Cloud-scale RNA-sequencing differential expression analysis with Myrna**. *Genome Biol*. 2010;11(8):R83. doi:10.1186/gb-2010-11-8-r83

Myrna design



Langmead B, Hansen KD, Leek JT. **Cloud-scale RNA-sequencing differential expression analysis with Myrna**. *Genome Biol*. 2010;11(8):R83. doi:10.1186/gb-2010-11-8-r83

Myrna results

Amazon Elastic MapReduce, c1.xlarge instances

	# Samples	Input size (GB, gzipped)	# CPUS	Wall clock time	Analysis cost	Cost per input GB
Pickrell et al	69	42	320	1h:38m	$65.60	$1.56
GEUVADIS	465	1,068*	600	19h:08m	$364.50	**$0.34**

Pickrell et al data transfer: ~$12, ~1hr

GEUVADIS data transfer: ~$34, ~14hr

* analyzing mate 1

Pickrell, et al. **Understanding mechanisms underlying human gene expression variation with RNA sequencing**. *Nature* 464.7289 (2010): 768-772.

GEUVADIS: Lappalainen T, et al. **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature*. 2013 Sep 15. doi: 10.1038/nature12531.

ReCount: digested RNA-seq using Myrna

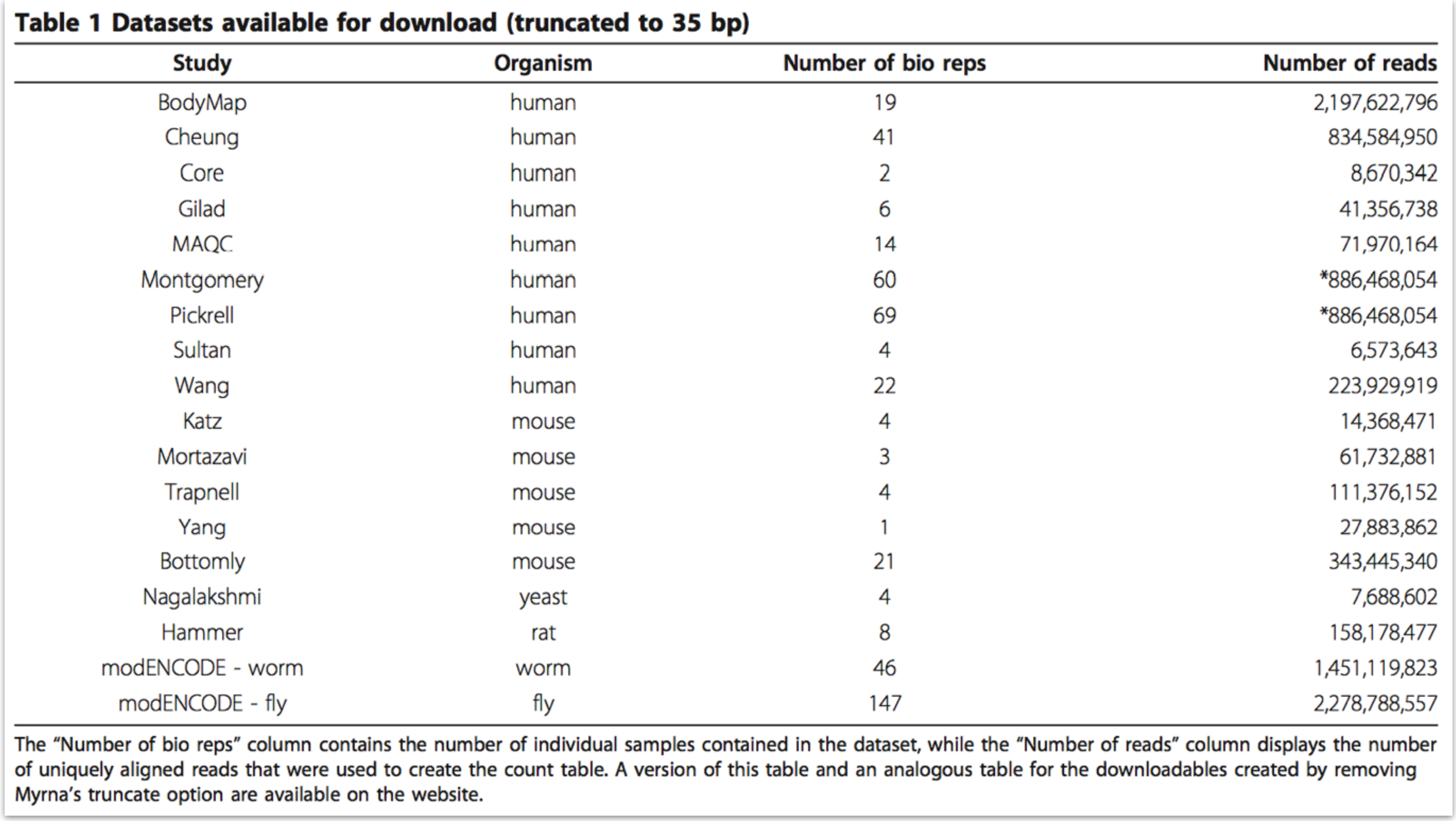
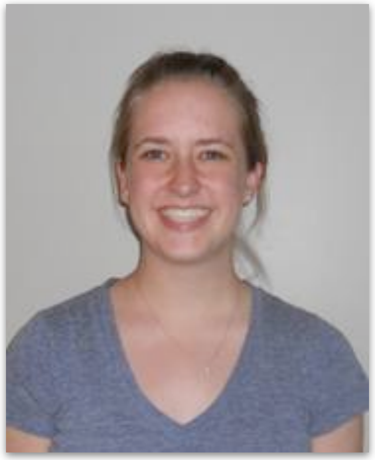


Table 1 Datasets available for download (truncated to 35 bp)

Study	Organism	Number of bio reps	Number of reads
BodyMap	human	19	2,197,622,796
Cheung	human	41	834,584,950
Core	human	2	8,670,342
Gilad	human	6	41,356,738
MAQC	human	14	71,970,164
Montgomery	human	60	*886,468,054
Pickrell	human	69	*886,468,054
Sultan	human	4	6,573,643
Wang	human	22	223,929,919
Katz	mouse	4	14,368,471
Mortazavi	mouse	3	61,732,881
Trapnell	mouse	4	111,376,152
Yang	mouse	1	27,883,862
Bottomly	mouse	21	343,445,340
Nagalakshmi	yeast	4	7,688,602
Hammer	rat	8	158,178,477
modENCODE - worm	worm	46	1,451,119,823
modENCODE - fly	fly	147	2,278,788,557

The "Number of bio reps" column contains the number of individual samples contained in the dataset, while the "Number of reads" column displays the number of uniquely aligned reads that were used to create the count table. A version of this table and an analogous table for the downloadables created by removing Myrna's truncate option are available on the website.



Alyssa Frazee

- Normalized gene-count tables encompassing 18 different published studies, 475 samples, >8 billion RNA-seq reads

Frazee AC, Langmead B, Leek JT. **ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets**. *BMC Bioinformatics*. 2011 Nov 16;12:449. doi: 10.1186/1471-2105-12-449.

Rail-RNA

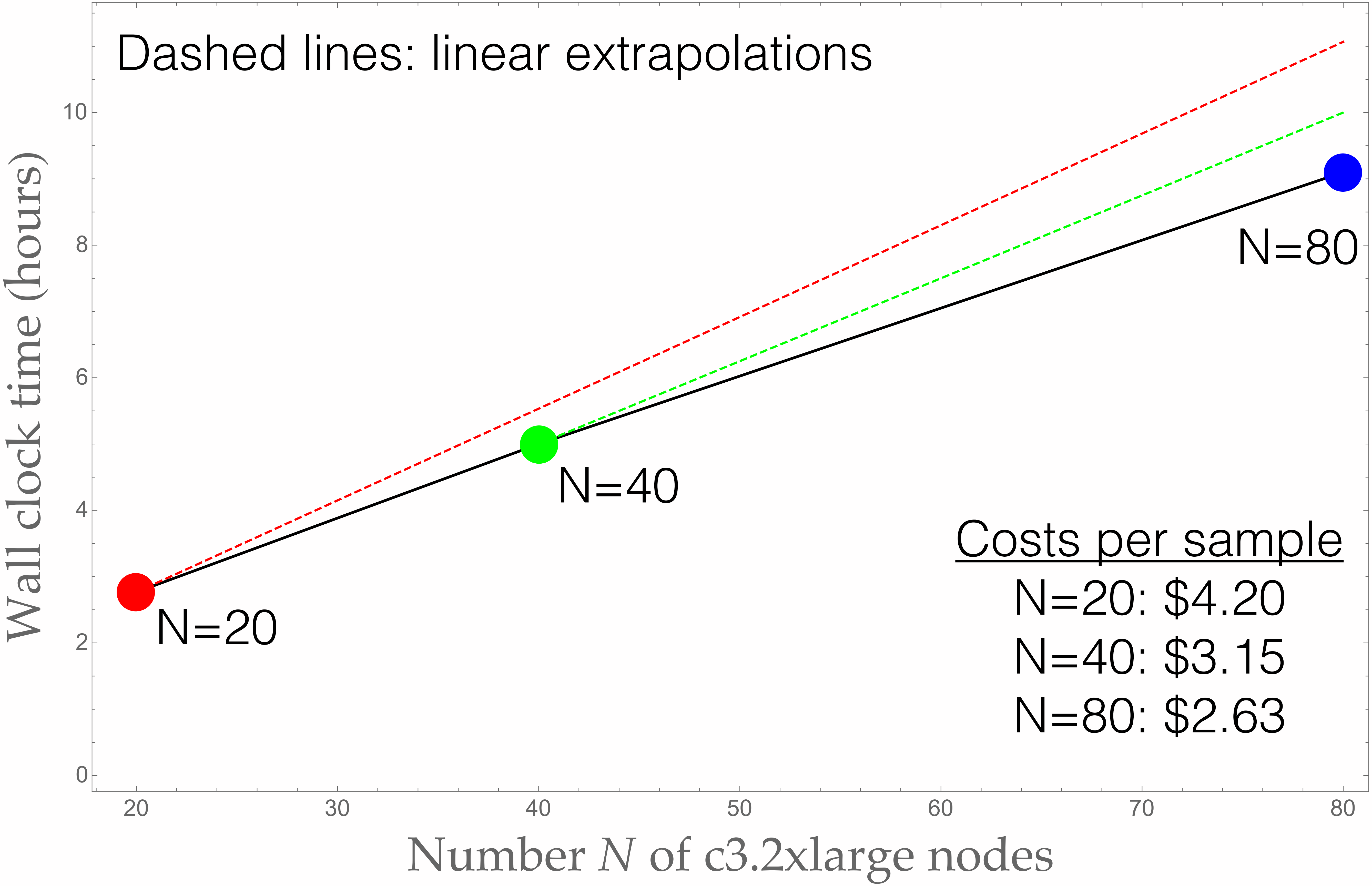
- Bring scalability closer to frontier of RNA-seq analysis
- Benefit maximally from analyzing many samples at once

Abhinav Nellore

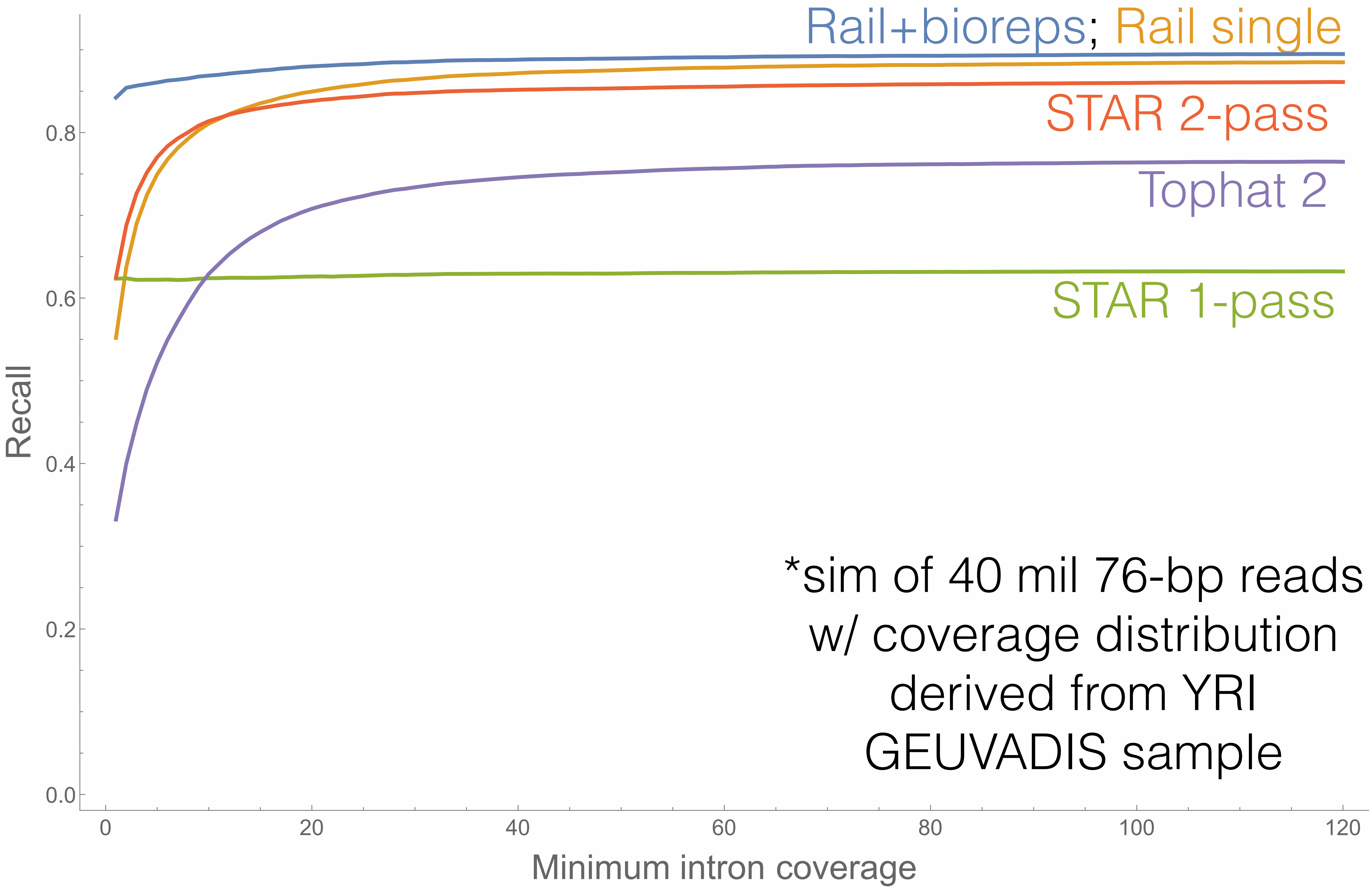
Jacob Pritt

Jeff Leek

On 40 c3.2xlarge EC2 machines



Recall of instances where intron is overlapped by read; no annotation provided*



*sim of 40 mil 76-bp reads w/ coverage distribution derived from YRI GEUVADIS sample

Thank you

Abhinav Nellore

Rail

Jacob Pritt

Rail

Kasper Hansen

Myrna

Jeff Leek

Rail

Myrna

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

Contact:

Email: langmea@cs.jhu.edu
Web: www.langmead-lab.org
Twitter: @BenLangmead