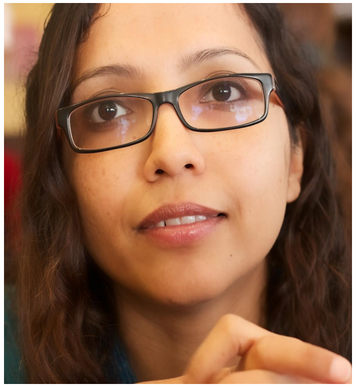
Stanford Galaxy Workshop (plus RNA-seq!)

Bioinformatics processing without coding

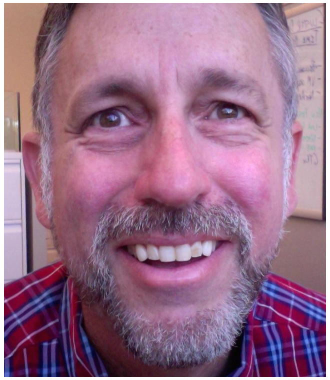
Agenda for today

- Keith: Brief introduction to Galaxy
- Jennifer: Hands-on Galaxy workshop
- Ramesh: RNA-seq Pipeline in Galaxy

Bioinformatics Team



Somalee Datta, PhD
Director



Keith Bettinger, MS
Sr Bioinformatician

Ramesh Nair, PhD
Sr Bioinformatician

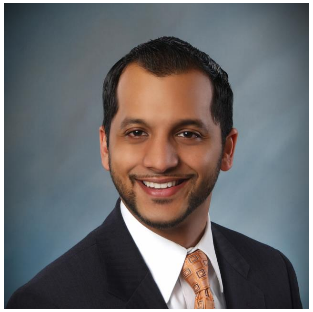
Alex Chekholko, MS
Systems Admin

Nathan Hammond, PhD
Software Developer

Amin Zia, PhD
Staff Scientist

Nathaniel Watson, MS
Bioinformatician

Isaac Liao, PhD
Software Developer



Denis Salins, BS
Software Developer

Bioinformatics Team

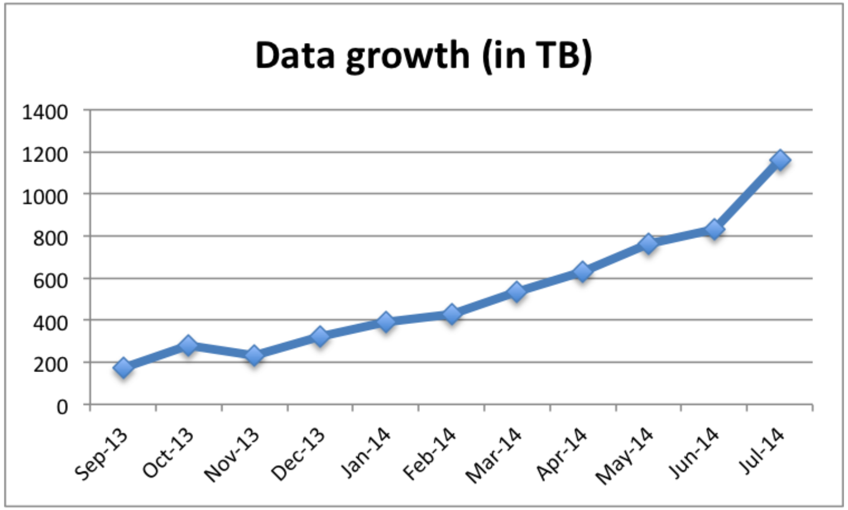
Bioinformaticians for Big Data Genomics
(supporting grants totalling more than $30M)

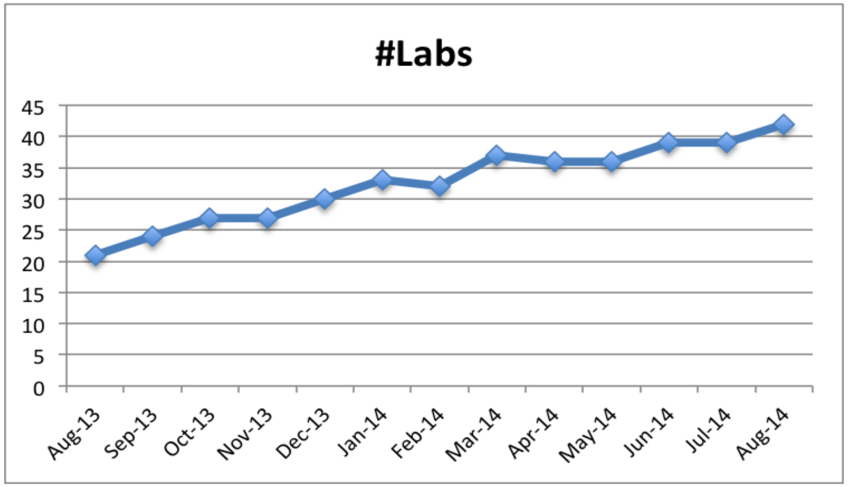
- Stanford Clinical Genomics Service
- Stanford Sequencing Service Center
- Bioinformatics Service Center
- ENCODE
- CIRM Stem Cell Center of Excellence
- iPOP
- VA Million Veteran Program

SCG Cluster

Stanford's Cluster for Big Data Genomics

- ~50 Labs & 450+ users
- ~1200 cores
- 3 Pb+ of storage
- dbGaP compliant





SCG Cluster

Advisory Committee

Labs

Artandi ✦ Ashley ✦ Assimes
Baker ✦ Barna ✦ Bassik
Batzoglou ✦ Bhutani ✦ Blau
Brunet ✦ Bustamante ✦ Butte
Cherry ✦ Cho ✦ Coller ✦ Fuller
Kundaje ✦ Li ✦ Merker ✦ Mignot
Montgomery ✦ Petrov ✦ Pringle
Pritchard ✦ Quertermous
Rosenberg ✦ Sabatti ✦ Sage
Saltzman ✦ Sattely ✦ Sherlock
Singh ✦ Skotheim ✦ Snyder
Steinmetz ✦ Sweet-Cordero
Tang ✦ Urban ✦ Whittemore
Winkelmann ✦ Winslow ✦ Wong
Wu

Challenges in using SCG Cluster

- Need to learn how to code!
- Command-line interface makes data management too abstract
- Difficult to share pipelines and data
- Analysis and visualization tools are scattered

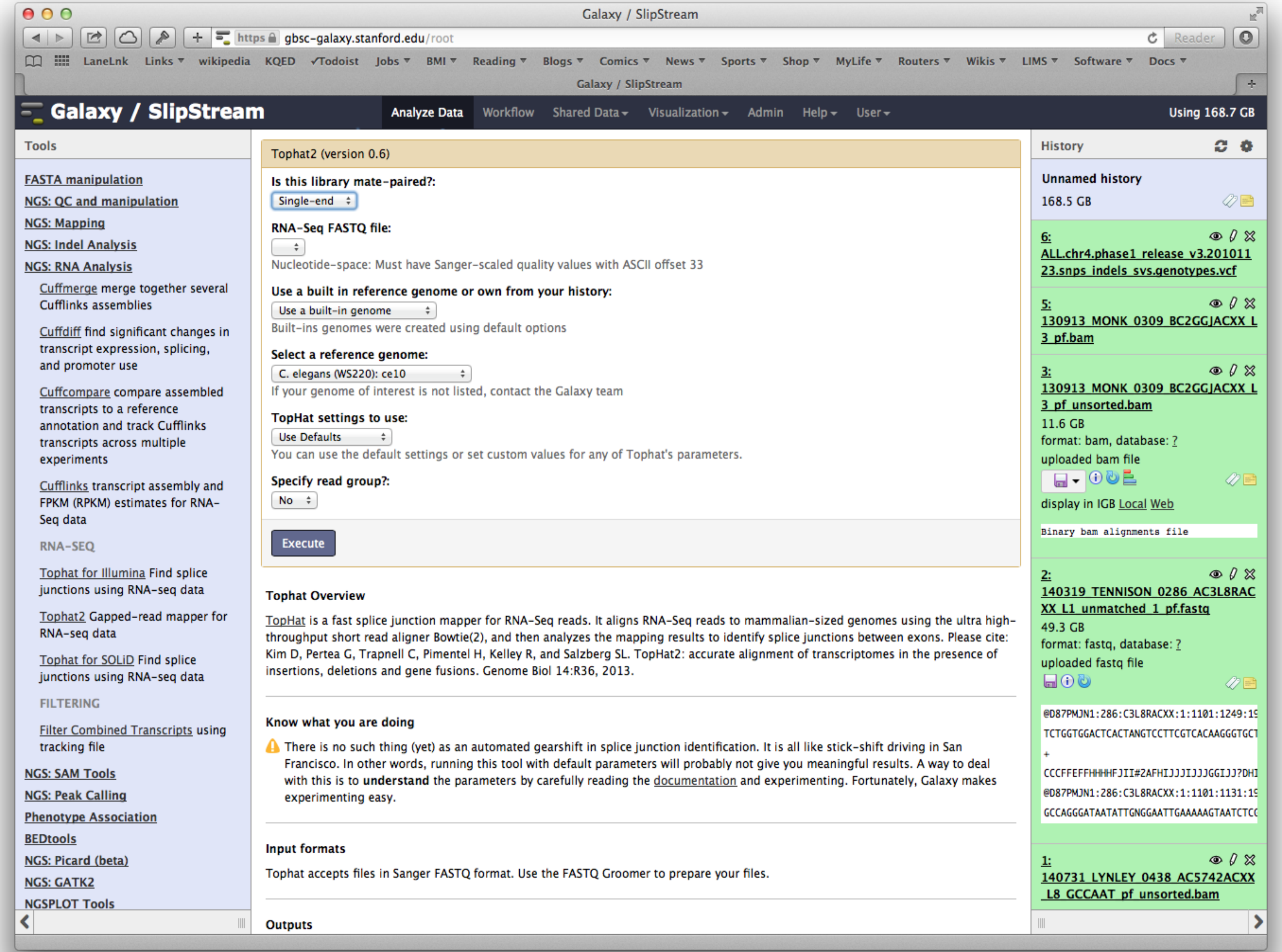
What is Galaxy?

Web interface to bioinformatic analyses

- Point-and-click execution
- Preinstalled suite of tools (extendable)
- Graphical pipeline builder (workflows)
- Visualizations

Why use Galaxy?

- Analyze bioinformatics data without learning to code
- Run standard analyses repeatably
- Easily create new pipeline flows
- Publish tools, pipelines, and data to community for easy sharing



FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: Indel Analysis
NGS: RNA Analysis

Cuffmerge merge together several Cufflinks assemblies

Cuffdiff find significant changes in transcript expression, splicing, and promoter use

Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments

Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data

RNA-SEQ

Tophat for Illumina Find splice junctions using RNA-seq data

Tophat2 Gapped-read mapper for RNA-seq data

Tophat for SOLiD Find splice junctions using RNA-seq data

FILTERING

Filter Combined Transcripts using tracking file

NGS: SAM Tools
NGS: Peak Calling
Phenotype Association
BEDtools
NGS: Picard (beta)
NGS: GATK2
NGSPLOT Tools

Tophat2 (version 0.6)

Is this library mate-paired?:
Single-end

RNA-Seq FASTQ file:

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Use a built in reference genome or own from your history:
Use a built-in genome
Built-ins genomes were created using default options

Select a reference genome:
C. elegans (WS220): ce10
If your genome of interest is not listed, contact the Galaxy team

TopHat settings to use:
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.

Specify read group?:
No

Execute

Tophat Overview

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie(2), and then analyzes the mapping results to identify splice junctions between exons. Please cite: Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, and Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36, 2013.

Know what you are doing

⚠ There is no such thing (yet) as an automated gearshift in splice junction identification. It is all like stick-shift driving in San Francisco. In other words, running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to **understand** the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.

Input formats

Tophat accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prepare your files.

Outputs

Unnamed history
168.5 GB

6: ALL.chr4.phase1 release v3.20101123.snps indels svs.genotypes.vcf

5: 130913 MONK 0309 BC2GGJACXX L3 pf.bam

3: 130913 MONK 0309 BC2GGJACXX L3 pf unsorted.bam
11.6 GB
format: bam, database: ?
uploaded bam file
display in IGB Local Web

```
Binary bam alignments file
```

2: 140319 TENNISON 0286 AC3L8RACXX L1 unmatched 1 pf.fastq
49.3 GB
format: fastq, database: ?
uploaded fastq file

```
@D87PMJN1:286:C3L8RACXX:1:1101:1249:19
TCTGGTGGACTCACTANGTCCTTCGTCACAAGGGTGCT
+
CCCFFEFFHHHHFJII#2AFHIJJJIJJJGGIJJ?DHI
@D87PMJN1:286:C3L8RACXX:1:1101:1131:19
GCCAGGGATAATATTGNGGAATTGAAAAAGTAATCTCC
```

1: 140731 LYNLEY 0438 AC5742ACXX L8 GCCAAT pf unsorted.bam

SLIPSTREAM GALAXY APPLIANCE

SLIPSTREAM GALAXY EDITION
APPLIANCE

Galaxy made easy.

Powerful dedicated desktop server pre-configured with a fully operational production instance of Galaxy

www.bioteam.net/slipstream/galaxy-edition

BIOTEAM
Enabling Science

WHO IS BIOTEAM?

Galaxy

Over a Decade of Life Sciences IT Consulting

- Staffed by **scientists** forced to learn IT to get research done
- Reduce the barrier to entry into data analysis by simplifying accessibility to Galaxy
- OFFICIAL APPLIANCE PROVIDER FOR THE GALAXY PROJECT

Thanks to...

- BioTeam
  - Server
  - Training
- Intel
  - Lunch!
- Dean Ann Arvin / CTO Ruth Marinshaw
  - Funding for Galaxy Server

intel Look Inside.™

Accelerate Science.
Translate Results.
Deliver Today.

Follow us on Twitter:
@IntelHealth

Now, on to Jennifer...

Enjoy the workshop!

HARDWARE SPECIFICATIONS	
CPU	**2x** Intel® Xeon® Processor E5-2690, 8-core (16 cores total)
Memory	**12x** 32 GB RDIMM (384 GB)
Storage	**7x** 3TB SAS 6 Gbps HDD (16 TB usable) **1x** 100GB SSD
Network	Dual Gigabit network adaptor
Power	Dual redundant power supplies