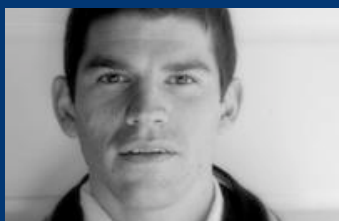


Progress and Challenges in Developing a Web-based Platform for Computational Biomedical Research

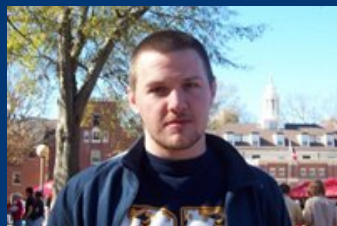
Jeremy Goecks
Depts. of Biology and Math & Computer Science
Emory University



EMORY



Enis Afgan



Dannon Baker



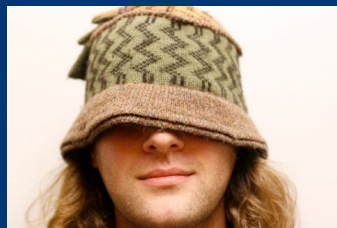
Dave Clements



Jeremy Goecks



Kanwei Li

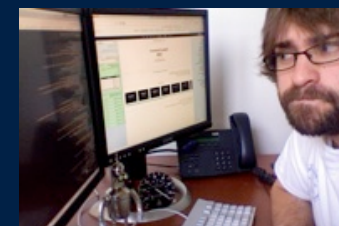


James Taylor

PENNSTATE



Dan Blankenberg



Nate Coraor



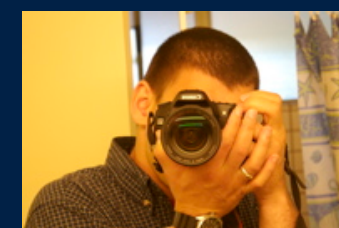
Jennifer Jackson



Greg von Kuster



Kelly Vincent



Anton Nekrutenko

Supported by the **NHGRI** (HG005542, HG004909, HG005133), **NSF** (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health

Overview

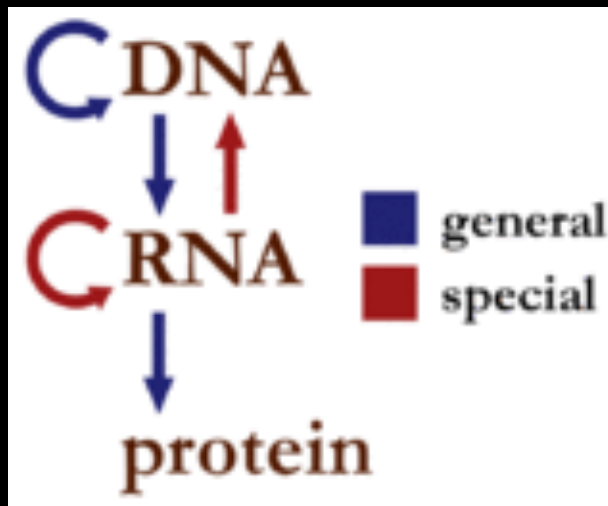
Genomics

Galaxy

- ✦ accessible, reproducible, and transparent science
- ✦ on the cloud
- ✦ visual analytics

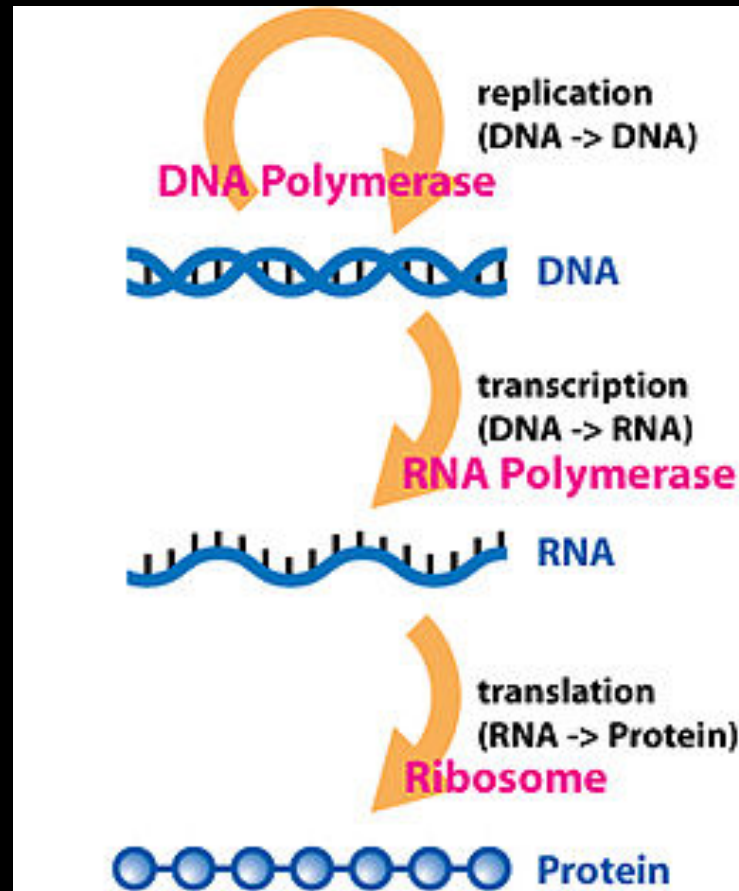
Reflections on Galaxy

Molecular Biology Primer



http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

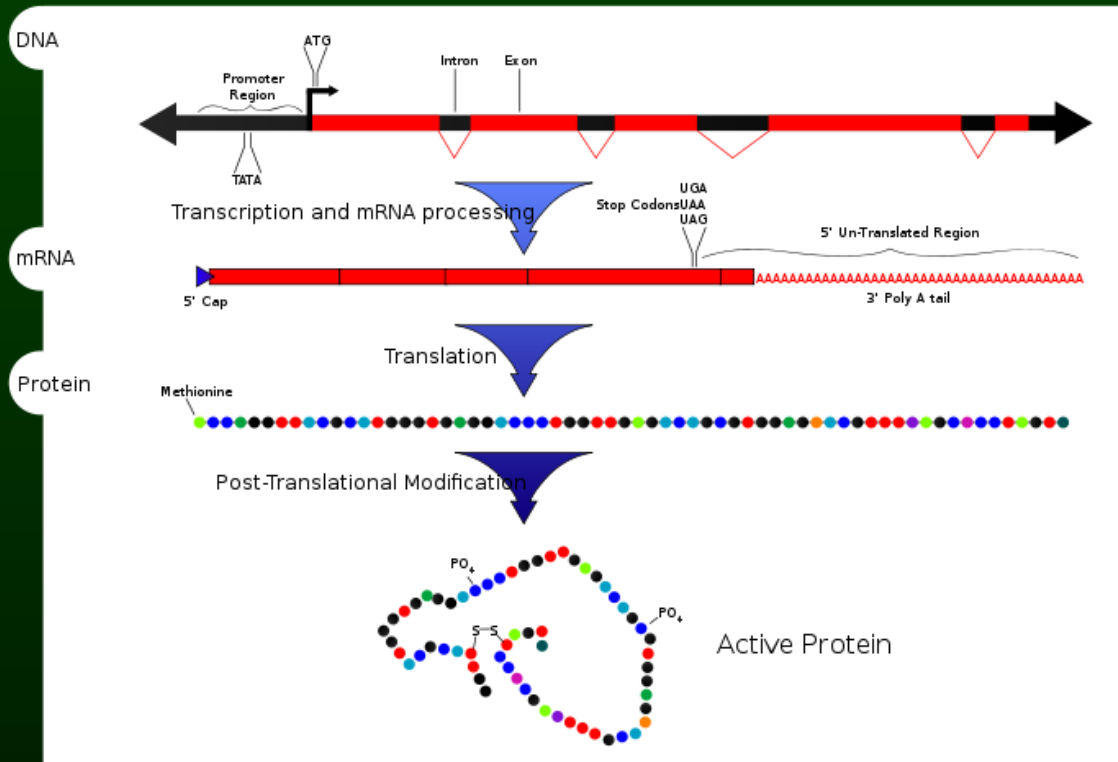
Molecular Biology Primer



http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

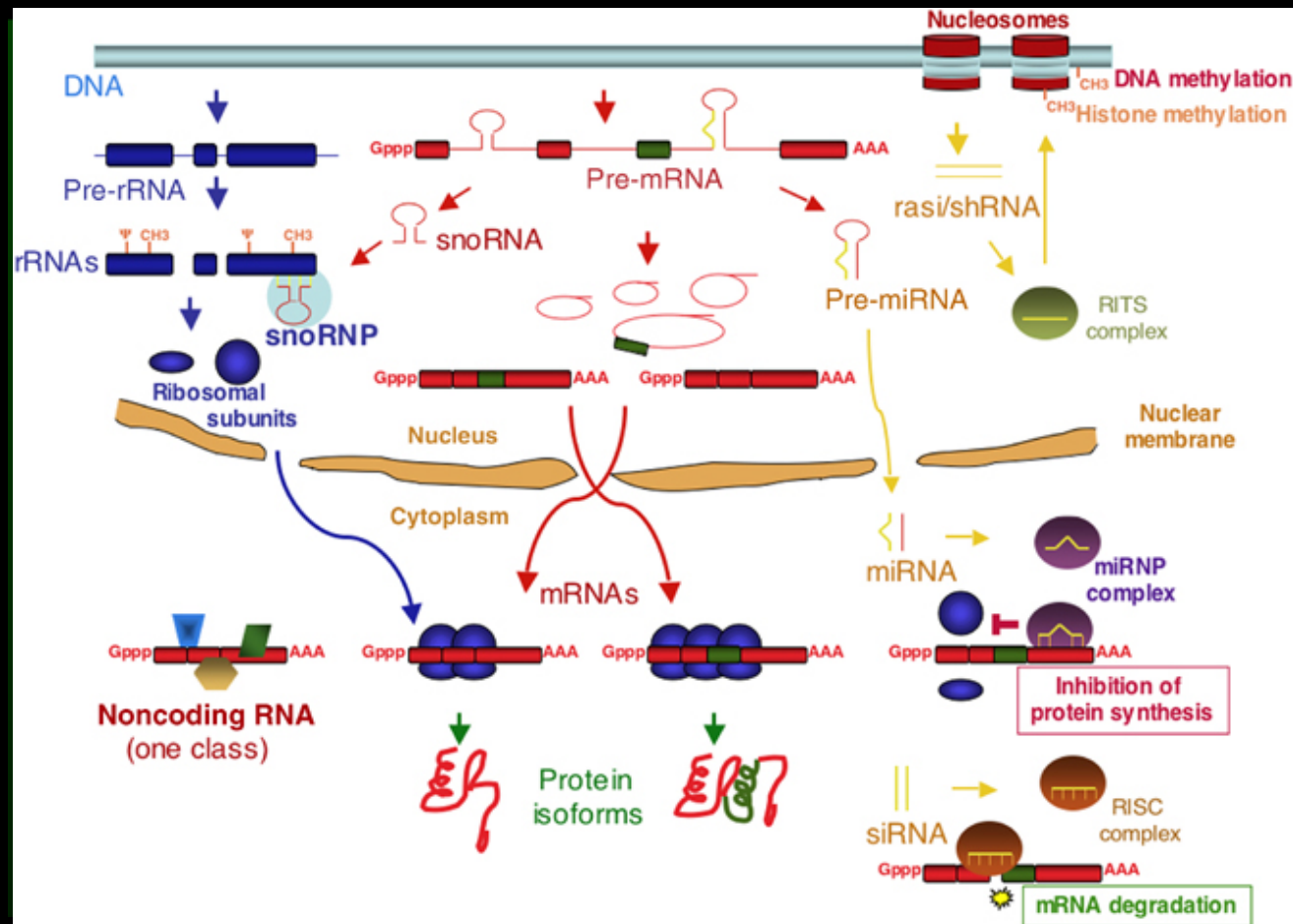
Molecular Biology Primer

Central Dogma of Molecular Biology : Eukaryotic Model



http://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology

Molecular Biology Primer



Goals of Genomics

Identify and annotate all functional genomic elements

- ✦ genes, promoters, enhancers, silencers, epigenetic modifications

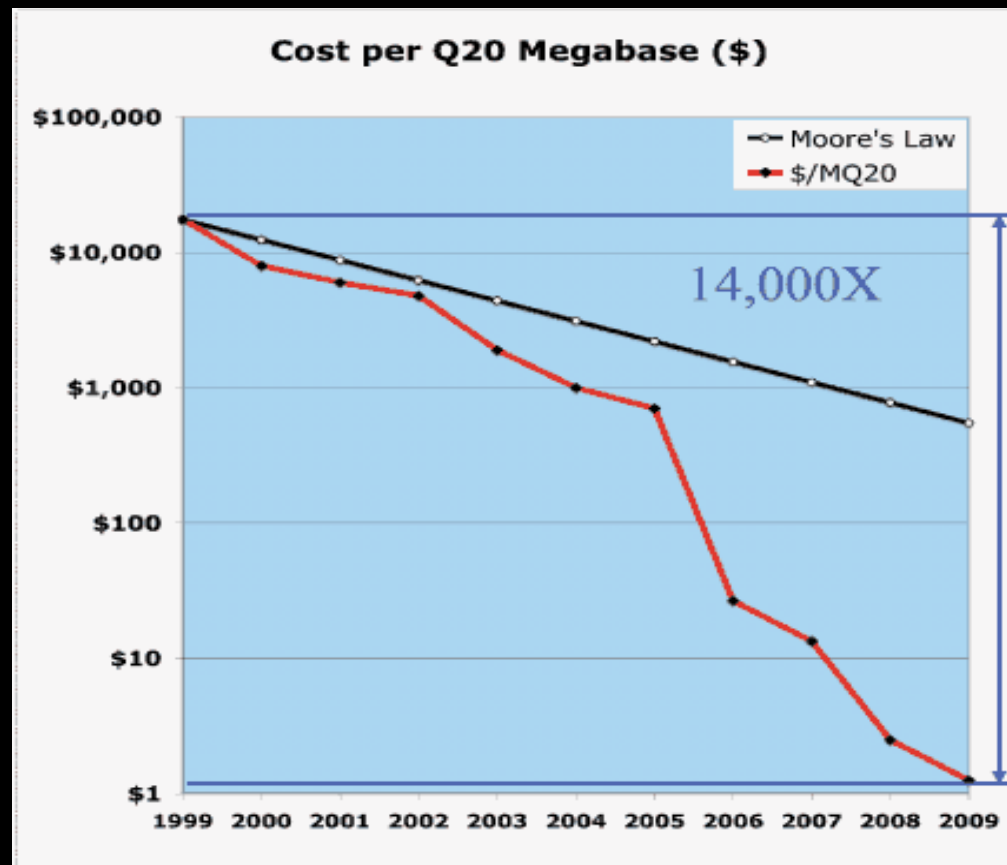
Understand genomic regulation

- ✦ interactions, networks, feedback systems

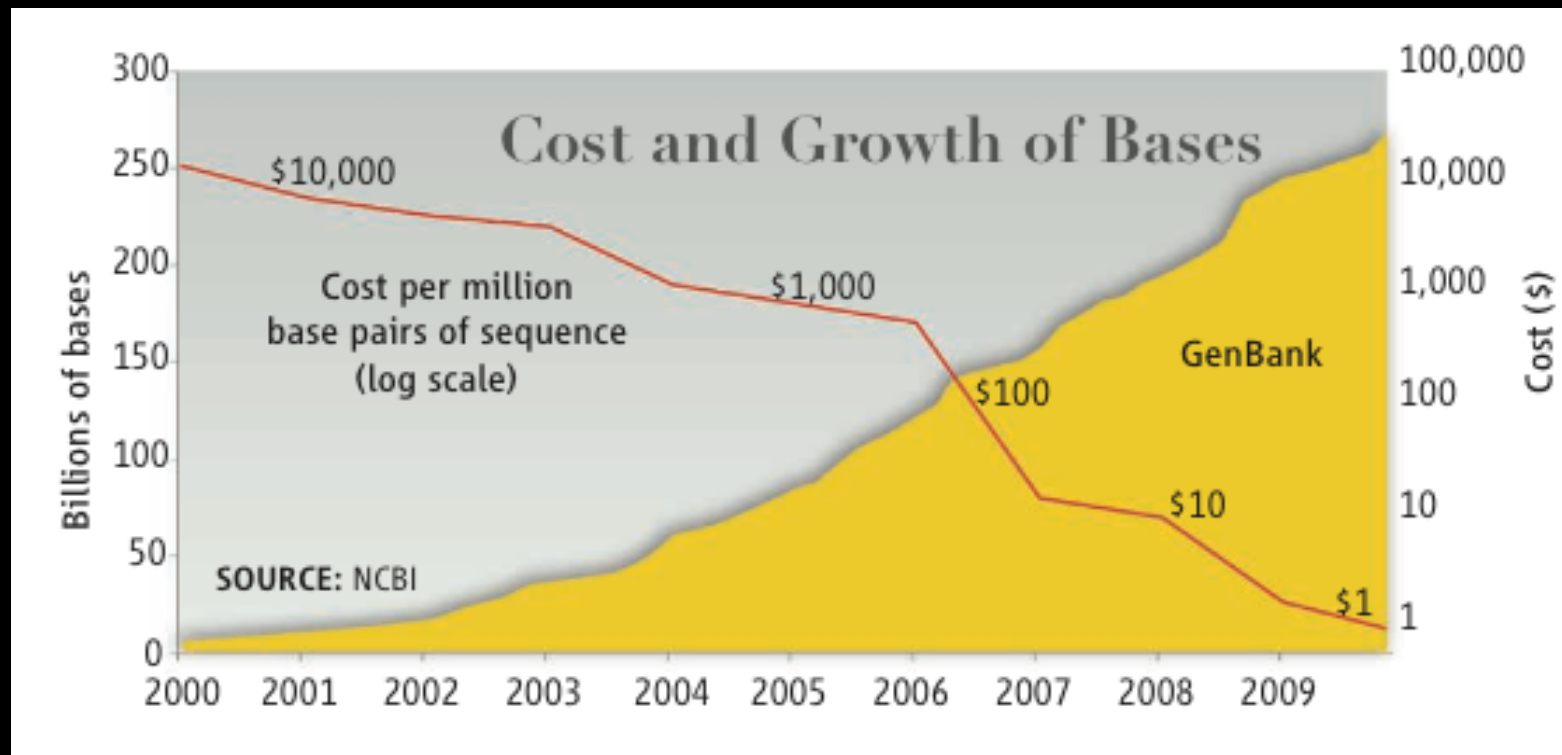
Apply knowledge of genome to address biomedical challenges

- ✦ personalized medicine
- ✦ aging
- ✦ environment interactions
- ✦ pathogen analysis
- ✦ ...

Trends in Genomics (1)



Trends in Genomics (2)





Will Computers Crash Genomics?

New technologies are making sequencing DNA easier and cheaper than ever, but the ability to analyze and store all that data is lagging

you-go service, accessible from one's own desktop, that provides rented time on a large cluster of machines that work together in parallel as fast as, or faster than, a single powerful computer. "Surviving the data deluge means computing in parallel," says Michael

www.sciencemag.org on February 10, 2011

"Will Computers Crash Genomics?", Pennisi, E., *Science*, Feb 11, 2011

Challenges in Genomics

Generating data is easy

- ✦ high-throughput sequencing (HTS) technologies improving rapidly
- ✦ datasets are hundreds of MBs to GBs

Analyzing data is THE bottleneck

- ✦ computation is essential due to dataset size
- ✦ \$1,000 genome, \$1,000,000 interpretation?

<http://www.bio-itworld.com/2010/10/01/interpretation.html>

Using Computation in Science?

Scientists often not trained in computation

Reproducibility hindered by complexity:
systems, scripts, tools, parameters

Collaboration and publishing difficult
because current media do not support
computational artifacts well

Overview

Genomics

Galaxy

- ✦ accessible, reproducible, and transparent science
- ✦ on the cloud
- ✦ visual analytics

Reflections on Galaxy

Galaxy Project: Fundamental Questions

When Biology (or any science) becomes dependent on computational methods:

- ✦ how can those methods best be made **accessible** to scientists?
- ✦ how best to ensure that analyses are **reproducible**?
- ✦ how best to facilitate **transparent communication and reuse** of analyses?

Vision

Galaxy is an open, Web-based platform for accessible, reproducible, and transparent computational biomedical research

Connecting Users with Tools

The screenshot displays the Galaxy web interface at <http://main.g2.bx.psu.edu/>. The main navigation bar includes links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. The left sidebar lists various tool categories such as Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Convert Formats, FASTA manipulation, Filter and Sort, Join, Subtract and Group, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Metagenomic analyses, EMBOS, NGS TOOLBOX BETA, NGS: QC and manipulation, NGS: Mapping, NGS: SAM Tools, NGS: Indel Analysis, NGS: Peak Calling, RGENETICS, SNP/WGA: Data: Filters, SNP/WGA: QC: LD: Plots, SNP/WGA: Statistical Models, and Workflows.

The central panel shows the configuration for the 'Map with Bowtie for Illumina' tool. The configuration includes:

- Will you select a reference genome from your history or use a built-in index?:** (Built-ins were indexed using default options)
- Select a reference genome:** (If your genome of interest is not listed - contact Galaxy team)
- Is this library mate-paired?:**
- Forward FASTQ file:** (Must have Sanger-scaled quality values with ASCII offset 33)
- Reverse FASTQ file:** (Must have Sanger-scaled quality values with ASCII offset 33)
- Maximum insert size for valid paired-end alignments (-X):**
- The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff):**
- Bowtie settings to use:** (For most mapping needs use Commonly used settings. If you want full control use Full parameter list)
- Suppress the header in the output SAM file:** ☒ (Bowtie produces SAM with several lines of header information by default)
- Execute** button

The right sidebar shows the **History** panel with a list of previous jobs:

- Imported: SNP Pileup Analysis for Sample E18
- 15: Variants from sample E18, consensus different, in RefSeq Genes
- 14: UCSC mm9 RefSeq Genes
- 13: Variants from sample E18 where consensus base different than ref. base
- 10: Variants from sample E18
- 9: Generate pileup on data 8
- 8: SAM-to-BAM on data 7
- 7: Map with Bowtie for Illumina on data 6 and data 5
- 6: E18 PE.2 Reads Groomed, Trimmed
- 5: E18 PE.1 Reads Groomed, Trimmed
- 4: E18 PE.2 Reads Groomed
- 3: E18 PE.1 Reads Groomed
- 2: E18 PE.2 Reads
- 1: E18 PE.1 Reads

At the bottom of the central panel, a description of the Bowtie tool is provided:

What it does
Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF FILES

- Extract features from GFF file
- Filter GFF file by attribute using simple expressions
- Filter GFF file by feature count using simple expressions

[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Operate on Genomic Intervals](#)
[Statistics](#)
[Graph/Display Data](#)
[Regional Variation](#)
[Multiple regression](#)
[Multivariate Analysis](#)
[Evolution](#)
[Metagenomic analyses](#)
[EMBOSS](#)

[NGS TOOLBOX BETA](#)
[NGS: QC and manipulation](#)
[NGS: Mapping](#)
[NGS: SAM Tools](#)
[NGS: Indel Analysis](#)
[NGS: Peak Calling](#)

[RGENETICS](#)
[SNP/WGA: Data; Filters](#)
[SNP/WGA: QC; LD; Plots](#)
[SNP/WGA: Statistical Models](#)

[Workflows](#)

Selecting Users with Tools

The screenshot displays the Galaxy web interface for the 'Map with Bowtie for Illumina' tool. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main configuration area contains the following fields and options:

- Reference genome:** A dropdown menu set to 'mm9'. A message below states: 'Your genome of interest is not listed - contact Galaxy team'.
- Paired-end:** A dropdown menu set to 'paired-end'.
- Forward FASTQ file:** A dropdown menu set to '1: E18 PE.1 Reads'. A note below reads: 'Must have Sanger-scaled quality values with ASCII offset 33'.
- Reverse FASTQ file:** A dropdown menu set to '1: E18 PE.1 Reads'. A note below reads: 'Must have Sanger-scaled quality values with ASCII offset 33'.
- Maximum insert size for valid paired-end alignments (-X):** A text input field containing '1000'.
- The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff):** A dropdown menu set to 'FR (for Illumina)'.
- Bowtie settings to use:** A dropdown menu set to 'Commonly used'. A note below reads: 'For most mapping needs use Commonly used settings. If you want full control use Full parameter list'.
- Suppress the header in the output SAM file:** A checked checkbox.
- Execute:** A button to run the tool.

Below the configuration area, a section titled 'What it does' provides a description of the Bowtie tool: 'Bowtie is a short read aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead and Cole Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.'

The right-hand sidebar shows a 'History' panel with a list of previous tool runs, including 'Imported: SNP Pileup Analysis for Sample E18', '15: Variants from sample E18, consensus different, in RefSeq Genes', '14: UCSC mm9 RefSeq Genes', '13: Variants from sample E18 where consensus base different than ref. base', '10: Variants from sample E18', '9: Generate pileup on data 8', '8: SAM-to-BAM on data 7', '7: Map with Bowtie for Illumina on data 6 and data 5', '6: E18 PE.2 Reads Groomed, Trimmed', '5: E18 PE.1 Reads Groomed, Trimmed', '4: E18 PE.2 Reads Groomed', '3: E18 PE.1 Reads Groomed', '2: E18 PE.2 Reads', and '1: E18 PE.1 Reads'.

Selecting Users with Tools

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order

- Select lines that match an

Operate on Genomic Intervals

- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query
- Concatenate two queries into one query
- Base Coverage of all intervals
- Coverage of a set of intervals on second set of intervals
- Complement intervals of a query
- Cluster the intervals of a query
- Join the intervals of two queries side-by-side
- Get flanks returns flanking region/s for every gene
- Fetch closest feature for every interval
- Profile Annotations for a set of genomic intervals

The screenshot shows the Galaxy web interface. The main panel is titled 'Bowtie for Illumina' and contains the following configuration options:

- Select a reference genome from your history or use a built-in index?:
 - in index: [dropdown]
 - are indexed using default options
 - reference genome: [dropdown]
- ome of interest is not listed - contact Galaxy team
- ry mate-paired?: [checkbox]
- STQ file:
 - Reads: [dropdown]
 - anger-scaled quality values with ASCII offset 33
- STQ file:
 - Reads: [dropdown]
 - anger-scaled quality values with ASCII offset 33
- nsert size for valid paired-end alignments (-X): [input]
- am/downstream mate orientation for valid paired-end alignment against the erence strand (--fr/--rf/--ff): [input]
- ina) [dropdown]
- ings to use:
 - used: [dropdown]
- apping needs use Commonly used settings. If you want full control use Full parameter
- he header in the output SAM file:
 - uces SAM with several lines of header information by default

The right sidebar shows a 'History' panel with a list of previous jobs:

- Imported: SNP Pileup Analysis for Sample E18
- 15: Variants from sample E18, consensus different, in RefSeq Genes
- 14: UCSC mm9 RefSeq Genes
- 13: Variants from sample E18 where consensus base different than ref. base
- 10: Variants from sample E18
- 9: Generate pileup on data 8
- 8: SAM-to-BAM on data 7
- 7: Map with Bowtie for Illumina on data 6 and data 5
- 6: E18 PE.2 Reads Groomed, Trimmed
- 5: E18 PE.1 Reads Groomed, Trimmed
- 4: E18 PE.2 Reads Groomed
- 3: E18 PE.1 Reads Groomed
- 2: E18 PE.2 Reads
- 1: E18 PE.1 Reads

Selecting Users with Tools

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order

- Select lines that match an

Operate on Genomic Intervals

- Intersect the intervals of two queries
- Subtract the intervals of two queries
- Merge the overlapping intervals of a query

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases

The screenshot displays the Galaxy web interface. The main panel shows the configuration for the 'Bowtie for Illumina' tool. The configuration includes a dropdown for 'Select a reference genome from your history or use a built-in index?', a text input for 'Index', a dropdown for 'Reference genome', and a checkbox for 'Use paired-end alignments (-X)'. The right sidebar shows a 'History' panel with a list of previous steps, including 'Imported: SNP Pileup Analysis for Sample E18', '15: Variants from sample E18, consensus different, in RefSeq Genes', '14: UCSC mm9 RefSeq Genes', '13: Variants from sample E18 where consensus base different than ref. base', '10: Variants from sample E18', '9: Generate pileup on data 8', '8: SAM-to-BAM on data 7', '7: Map with Bowtie for Illumina on data 6 and data 5', '6: E18 PE.2 Reads Groomed, Trimmed', '5: E18 PE.1 Reads Groomed, Trimmed', '4: E18 PE.2 Reads Groomed', '3: E18 PE.1 Reads Groomed', '2: E18 PE.2 Reads', and '1: E18 PE.1 Reads'. The interface is clean and organized, with a navigation bar at the top and a search bar.

Tools

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match a query
- Operate on Genomes
 - Intersect the intersection of two queries
 - Subtract the intersection of two queries
 - Merge the overlap of a query
- NGS: SAM Tools
 - Filter SAM on values
 - Convert SAM to BAM
 - SAM-to-BAM format to BAM
 - BAM-to-SAM format to SAM
 - Merge BAM files together
 - Generate pileup dataset
 - Filter pileup on coverage and SNPs
 - Pileup-to-Interval condenses pileup format into ranges of bases

Filter pileup

Select dataset:

10: Variants from sample E18

which contains:

Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

Execute

aligner designed to be ultrafast and memory-efficient. It is developed by Ben Langmead. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25.

Filter and Sort

- Filter data on any column using simple expressions

- Sort data in ascending or descending order

- Select lines that match a query

Operate on Genomes

- Intersect the intersection of two queries

- Subtract the intersection of two queries

- Merge the overlap of a query

NGS: SAM Tools

- Filter SAM on values

- Convert SAM to BAM

- SAM-to-BAM format to BAM

- BAM-to-SAM format to SAM

- Merge BAM files together

- Generate pileup dataset

- Filter pileup on coverage and SNPs

- Pileup-to-Interval condenses pileup format into ranges of bases

Filter pileup

Select dataset:

10: Variants from sample E18

which contains:

Pileup with six columns (simple)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:

20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:

3

Pileup lines with coverage lower than this value will be skipped

Only report variants?:

Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:

No

See "Output format" below for explanation

Print total number of differences?:

No

See "Example 3" below for explanation

Print quality and base string?:

Yes

See "Example 4" below for explanation

Execute

History

Options



Variant Analysis for Sample E18

15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes

14: UCSC mm9 RefSeq Genes

13: Filter to get Variants from sample E18 where consensus base different than ref. base

10: Filter pileup to get Variants from sample E18

9: Generate pileup on data 8

8: SAM-to-BAM on data 7

7: Map with Bowtie for Illumina on data 6 and data 5

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed, Trimmed

aligner designed to be ultrafast and memory-efficient. It is developed by Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol.

- Filter data on any column using simple expressions

- Sort data in ascending or descending order

- Select lines that match
- Operate on Genomes

- Intersect the in queries

- Subtract the int queries

- Merge the over of a query

NGS: SAM To

- Filter SAM values

- Convert SA

- SAM-to-BAM
format to

- BAM-to-SAM
format to S

- Merge BAM files together

- Generate p
dataset

- Filter pileup on coverage and SNPs

- Pileup-to-Interval condenses pileup format into ranges of bases



This dataset is large and only the first megabyte is shown below.

Show all | Save

chr10	6882036	6882037	A	A	107	0	60	32	\$. C
chr10	14243075	14243076	G	G	96	0	60	35	t
chr10	14243079	14243080	C	C	106	0	60	35	
chr10	14465082	14465083	T	K	173	176	60	35	GE
chr10	14465083	14465084	G	K	144	144	60	35	
chr10	14465084	14465085	T	T	117	0	60	38	
chr10	14465085	14465086	G	G	70	0	60	38	
chr10	14465257	14465258	C	C	79	0	60	42	
chr10	14465258	14465259	A	A	137	0	60	46	
chr10	14465263	14465264	A	A	136	0	60	61	
chr10	14465366	14465367	A	A	101	0	60	38	g
chr10	14465371	14465372	G	G	137	0	60	50	
chr10	14465410	14465411	G	G	184	0	60	69	\$
chr10	14465447	14465448	T	T	186	0	60	65	\$
chr10	14465456	14465457	G	G	193	0	60	70	
chr10	14465465	14465466	T	T	177	0	60	63	
chr10	14465485	14465486	C	T	129	129	60	34	t
chr10	14465569	14465570	T	T	219	0	60	84	
chr10	14465581	14465582	G	G	240	0	60	84	\$
chr10	14465586	14465587	C	C	248	0	60	82	\$
chr10	14465621	14465622	C	C	134	0	60	49	
chr10	14465658	14465659	C	C	134	0	60	49	
chr10	14465660	14465661	T	T	153	0	60	55	
chr10	14465691	14465692	G	G	128	0	60	42	\$
chr10	14465778	14465779	C	C	89	0	60	34	\$
chr10	14465791	14465792	G	G	104	0	60	33	\$
chr10	14465881	14465882	G	G	110	0	60	41	
chr10	17445088	17445089	A	A	103	0	60	34	
chr10	17445271	17445272	A	A	55	0	60	34	
chr10	17731269	17731270	T	T	113	0	60	42	\$
chr10	19928287	19928288	G	A	135	135	60	36	AA
chr10	19928468	19928469	C	T	132	132	60	35	T
chr10	19928488	19928489	A	A	119	0	60	44	
chr10	19928494	19928495	C	T	138	138	60	37	TT
chr10	19928527	19928528	A	A	134	0	60	45	
chr10	19928538	19928539	G	G	144	0	60	52	\$
chr10	19928543	19928544	A	G	147	147	60	40	G
chr10	19928741	19928742	T	T	80	0	60	30	
chr10	20799826	20799827	G	G	117	0	60	37	
chr10	28750217	28750218	C	T	138	138	60	37	TT
chr10	28750397	28750398	A	C	154	211	60	64	C
chr10	28750401	28750402	A	A	128	0	60	47	\$
chr10	28750423	28750424	C	T	113	113	60	35	T
chr10	28750438	28750439	A	A	95	0	60	36	\$
chr10	28750446	28750447	A	G	165	165	60	46	G
chr10	28750487	28750488	A	A	80	0	60	31	
chr10	28750512	28750513	G	G	220	0	60	72	\$
chr10	28750548	28750549	G	C	255	255	60	97	C
chr10	28750574	28750575	T	T	237	0	60	83	\$
chr10	28750577	28750578	T	T	234	0	60	82	\$
chr10	28750578	28750579	T	T	242	0	60	76	\$
chr10	28750593	28750594	G	G	220	0	60	75	\$
chr10	28750640	28750641	T	C	165	165	60	46	C
chr10	28750746	28750747	G	A	202	202	60	58	AA
chr10	28750766	28750767	A	G	205	205	60	59	G
chr10	28750769	28750770	T	C	175	175	60	49	

aligner designed to be ultrafast and memory-efficient. It is developed by Trapnell. Please cite: Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*.

Accessibility

Options ▾

Analysis for Sample E18

Select to get Variants   
Sample E18, consensus different,
1 Genes

mm9 RefSeq Genes

to get Variants from 18 where consensus base
than ref. base

pileup to get
from sample E18

Generate pileup on data 8   

o-BAM on data 7

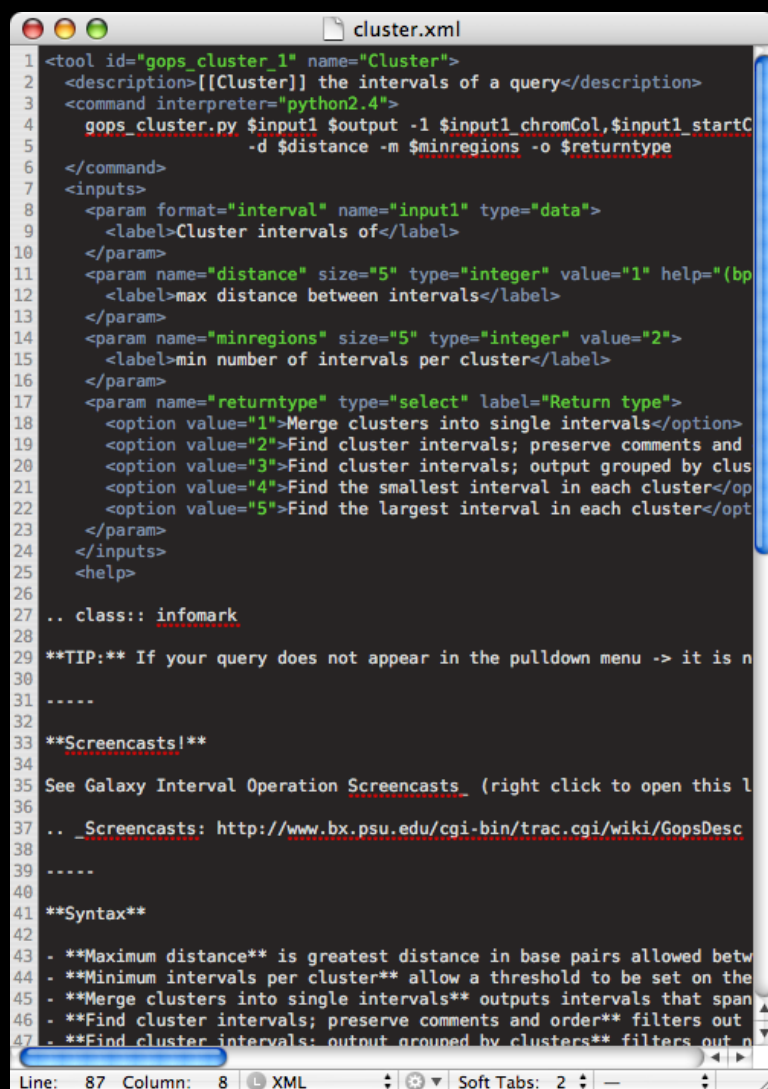
with Bowtie for
on data 6 and data 5

6: E18 PE.2 Reads Groomed, Trimmed

5: E18 PE.1 Reads Groomed,   

Trimmed

A Tool in Galaxy



```

1 <tool id="gops_cluster_1" name="Cluster">
2   <description>[[Cluster]] the intervals of a query</description>
3   <command interpreter="python2.4">
4     gops_cluster.py $input1 $output -l $input1_chromCol,$input1_startC
5     -d $distance -m $minregions -o $returntype
6   </command>
7   <inputs>
8     <param format="interval" name="input1" type="data">
9       <label>Cluster intervals of</label>
10    </param>
11    <param name="distance" size="5" type="integer" value="1" help="(bp
12    <label>max distance between intervals</label>
13    </param>
14    <param name="minregions" size="5" type="integer" value="2">
15    <label>min number of intervals per cluster</label>
16    </param>
17    <param name="returntype" type="select" label="Return type">
18      <option value="1">Merge clusters into single intervals</option>
19      <option value="2">Find cluster intervals; preserve comments and
20      <option value="3">Find cluster intervals; output grouped by clus
21      <option value="4">Find the smallest interval in each cluster</opt
22      <option value="5">Find the largest interval in each cluster</opt
23    </param>
24  </inputs>
25  <help>
26
27  .. class:: infomark
28
29  **TIP:** If your query does not appear in the pulldown menu -> it is n
30
31  ----
32
33  **Screencasts!**
34
35  See Galaxy Interval Operation Screencasts_ (right click to open this l
36
37  .. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc
38
39  ----
40
41  **Syntax**
42
43  - **Maximum distance** is greatest distance in base pairs allowed betw
44  - **Minimum intervals per cluster** allow a threshold to be set on the
45  - **Merge clusters into single intervals** outputs intervals that span
46  - **Find cluster intervals; preserve comments and order** filters out
47  - **Find cluster intervals; output grouped by clusters** filters out n
  
```

Line: 87 Column: 8 XML Soft Tabs: 2

Defined via abstract interface:

- ✦ inputs & outputs
- ✦ parameters
- ✦ how to generate command line

As simple as possible
but allows for rigorous
reasoning

Reproducibility in Genomics

18 *Nat. Genetics* experiments in
microarray gene expression

<50% of reproducible

Problems

- ✦ missing data (38%)
- ✦ missing software, hardware details (50%)
- ✦ missing method, processing details (66%)

Ioannidis, J.P.A. et al. "Repeatability of published microarray gene expression analyses." Nat Genet 41, 149-155 (2009)

Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- ✦ missing data (38%)
- ✦ missing software, hardware details (50%)
- ✦ missing method, processing details (66%)

Ioannidis, J.P.A. et al. "Repeatability of published microarray gene expression analyses." Nat Genet 41, 149-155 (2009)

14 re-sequencing experiments in *Nat. Genetics, Nature, Science*

0% reproducible?

Problems

- ✦ missing primary data (50%)
- ✦ tools unavailable (50%)
- ✦ missing parameter setting, tool versions (100%)

"Devil in the details," Nature, vol. 470, 305-306 (2011).

Metadata = Reproducibility

Automatic Metadata

7: Map with Bowtie for Illumina on data 6 and data 5   

9,073,928 lines, format: sam,
database: mm9
Run this job again
info: sequence file aligned.

1. QNAME	2. FLAG	3. RNAME
HWI-EAS269:3:1:1449:913	99	chr1
HWI-EAS269:3:1:1449:913	147	chr1
HWI-EAS269:3:1:709:832	99	chr1
HWI-EAS269:3:1:709:832	147	chr1
HWI-EAS269:3:1:1422:1087	99	chr1
HWI-EAS269:3:1:1422:1087	147	chr1

Map with Bowtie for Illumina

Will you select a reference genome from your history or use a built-in index?

 Built-ins were indexed using default options

Select a reference genome:

 if your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?:

Forward FASTQ file:

 Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file:

 Must have Sanger-scaled quality values with ASCII offset 33

Maximum insert size for valid paired-end alignments (-X):

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff):

Bowtie settings to use:

 For most mapping needs use Commonly used settings. If you want full control use Full parameter list

Suppress the header in the output SAM file:
☒
 Bowtie produces SAM with several lines of header information by default

User Metadata

History
Options

Variant Analysis for Sample E18

Tags:

snp × pileup × bowtie ×
demo × sample:e18 ×

Annotation / Notes:
Perform a variant analysis with default parameters to identify variants in sample E18 that lie in annotated genes.

10: Variants from sample E18

26,742 regions, format: interval, database: mm9

Info:

Tags:

pileup × sample:e18 ×
snps ×

Annotation:

Find variants with coverage ≥ 30 and quality score ≥ 20 .

| display at UCSC [main](#) | view in [GeneTrack](#) | display at Ensembl [Current](#)

1. Chrom	2. Start	3. End	4	5	6	7
chr10	6882036	6882037	A	A	107	
chr10	14243075	14243076	G	G	96	
chr10	14243079	14243080	C	C	106	
chr10	14465082	14465083	T	K	173	
chr10	14465083	14465084	G	K	144	
chr10	14465084	14465085	T	T	117	

Data Provenance

Datasets are immutable in Galaxy

- ✦ associations between *Dataset* objects and their usage: *HistoryDatasetAssociation*, *LibraryDatasetAssociation*
- ✦ metadata associated with associations, not datasets

Exporting from Galaxy

- ✦ histories and workflows can be exported in JSON format

Metadata

- ✦ Datatype defined in Python code
- ✦ Job/tool metadata has official
- ✦ JSON format in database and when exported

Galaxy Workflows

The screenshot displays the Galaxy web interface at <http://main.g2.bx.psu.edu/>. The main content area shows a dataset view with a table of genomic data. A yellow warning banner at the top of the table states: "This dataset is large and only the first megabyte is shown below." The table columns include chromosome (chr10), coordinates, and various metrics.

On the right side, the "History" panel is open, showing a list of workflows. The workflow "10: V" is selected, and its details are displayed below. The workflow steps are:

- 1. Create
- 2. Generate pileup on data 8
- 3. SAM-to-BAM on data 8
- 4. Map with Bowtie for Illumina on data 6 and data 5

The workflow details show that it consists of 9,073,928 lines, format: sam, database: mm9. The workflow is currently running, and the progress bar indicates that the sequence file is aligned.

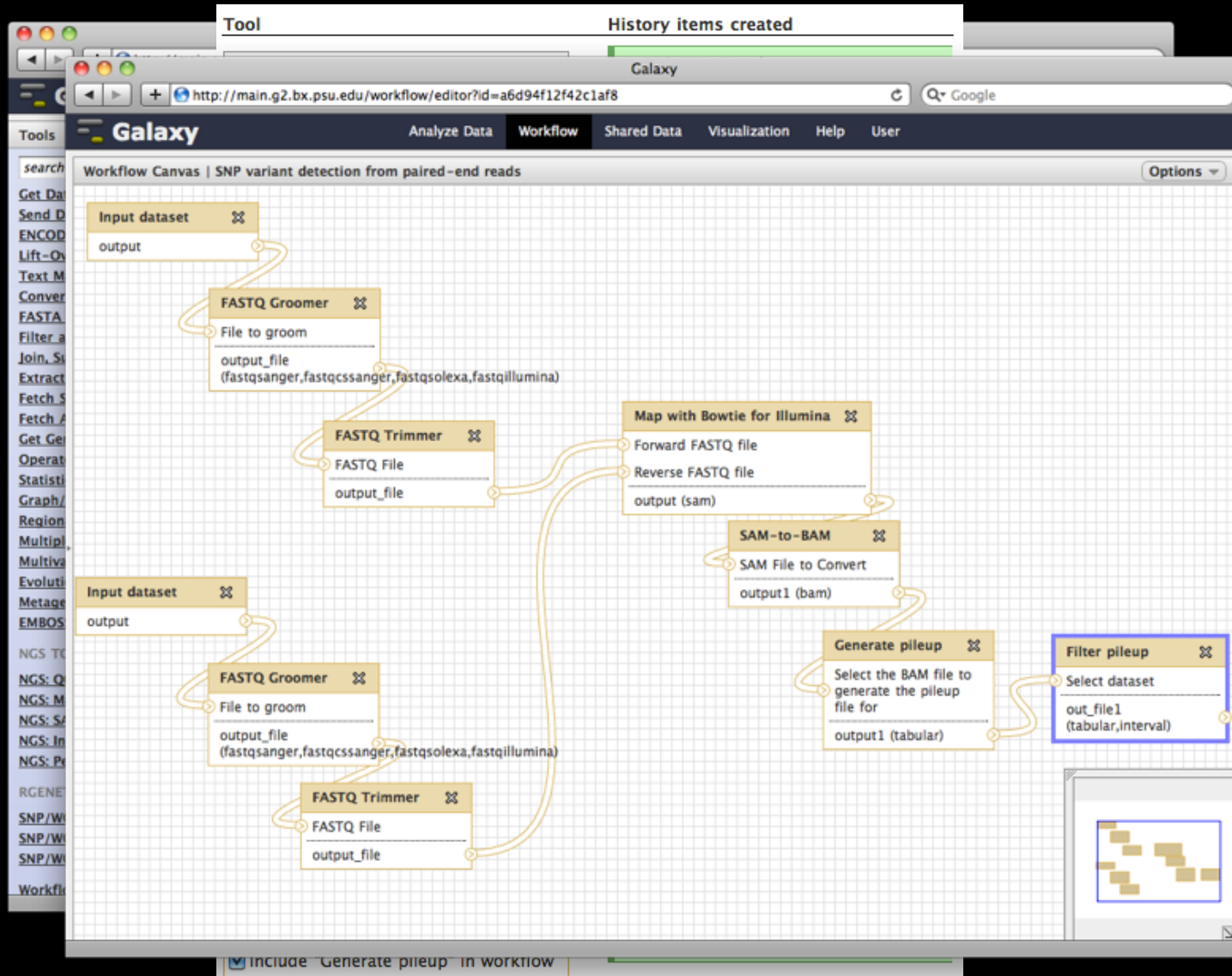
Galaxy Workflows



Tool	History items created
Upload File <i>This tool cannot be used in workflows</i>	1: E18 PE.1 Reads ☒ Treat as input dataset
Upload File <i>This tool cannot be used in workflows</i>	2: E18 PE.2 Reads ☒ Treat as input dataset
FASTQ Groomer ☒ Include "FASTQ Groomer" in workflow	3: E18 PE.1 Reads Groomed
FASTQ Groomer ☒ Include "FASTQ Groomer" in workflow	4: E18 PE.2 Reads Groomed
FASTQ Trimmer ☒ Include "FASTQ Trimmer" in workflow	5: E18 PE.1 Reads Groomed, Trimmed
FASTQ Trimmer ☒ Include "FASTQ Trimmer" in workflow	6: E18 PE.2 Reads Groomed, Trimmed
Map with Bowtie for Illumina ☒ Include "Map with Bowtie for Illumina" in workflow	7: Map with Bowtie for Illumina on data 6 and data 5
SAM-to-BAM ☒ Include "SAM-to-BAM" in workflow	8: SAM-to-BAM on data 7
Generate pileup ☒ Include "Generate pileup" in workflow	9: Generate pileup on data 8



Galaxy Workflows



Sharing, Collaborating, and Publishing with Galaxy

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history is currently restricted so that only you and the users listed below can access it. You can:

Make History Accessible via Link

Generates a web link that you can share with other people so that they can view and import the history.

Make History Accessible and Publish

Makes the history accessible via link (see above) and publishes the history to Galaxy's [Published Histories](#) section, where it is publicly listed and searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

[Back to Histories List](#)

Sharing and Publishing

Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history accessible via link and published.

Anyone can view and import this history by visiting the following URL:

<http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18>

This history is publicly listed and searchable in Galaxy's Published Histories section.

You can:

Unpublish History

Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Disable Access to History via Link and Unpublish

Disables history's link so that it is not accessible and removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

Sharing History with Specific Users

You have not shared this history with any users.

Share with a user

[Back to Histories List](#)

Galaxy | Published History | Variant Analysis for Sample E18

http://main.g2.bx.psu.edu/u/jgoecks/h/variant-analysis-for-sample-e18

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

Published Histories | jgoecks | Variant Analysis for Sample E18

Galaxy History 'Variant Analysis for Sample E18'

[+ Import history](#)

Annotation: Perform a pileup analysis with default parameters to identify variants in sample E18.

Dataset	Annotation
1: E18 PE.1 Reads	Forward reads from sample E18.
2: E18 PE.2 Reads	Reverse reads from sample E18.
3: E18 PE.1 Reads Groomed	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
4: E18 PE.2 Reads Groomed	Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3
5: E18 PE.1 Reads Groomed, Trimmed	Trim reads from 3' end to remove low-quality nts.
6: E18 PE.2 Reads Groomed, Trimmed	Trim reads from 3' to remove low-quality nts.
7: Map with Bowtie for Illumina on data 6 and data 5	Map paired-end reads with default parameters.
8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage ≥ 30 .
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

About this History

Author
jgoecks 

Related Histories
[All published histories](#)
[Published histories by jgoecks](#)

Rating
Community (1 rating, 4.0 average) ★★★★★
Yours ★★★★★

Tags
Community:
snp pileup bowtie demo
sample

Yours:
snp x pileup x bowtie x
demo x sample:e18 x 

Galaxy | Published Workflow | SNP variant detection from paired-end reads

http://main.g2.bx.psu.edu/u/jgoecks/w/snp-variant-detection-from-paired-end-reads

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Published Workflows | jgoecks | SNP variant detection from paired-end reads

Step 6: FASTQ Trimmer

Trim reads to remove low-quality bases.

FASTQ File
Output dataset 'output_file' from step 4

Define Base Offsets as
Absolute Values

Offset from 5' end
0

Offset from 3' end
9

Keep reads with zero length
False

Step 7: Map with Bowtie for Illumina

Map reads using default parameter values.

Will you select a reference genome from your history or use a built-in index?
Use a built-in index

Select a reference genome
/galaxy/data/apiMel3/bowtie_index/apiMel3

Is this library mate-paired?
Paired-end

Forward FASTQ file
Output dataset 'output_file' from step 6

Reverse FASTQ file
Output dataset 'output_file' from step 5

Maximum insert size for valid paired-end alignments (-X)
1000

The upstream/downstream mate orientation for valid paired-end alignment against the forward reference strand (--fr/--rf/--ff)
FR (for Illumina)

Bowtie settings to use
Commonly used

Suppress the header in the output SAM file
True

Step 8: SAM-to-BAM

Convert Bowtie SAM output to BAM format so that pileup can be run.

Choose the source for the reference list
Locally cached

About this Workflow

Author
jgoecks

Related Workflows
[All published workflows](#)
[Published workflows by jgoecks](#)

Rating
Community (0 ratings, 0.0 average)
Yours

Tags
Community: snp bowtie
Yours: snp bowtie

Galaxy | Published Histories

http://main.g2.bx.psu.edu/history/list_published

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Published Histories

search | Advanced Search

Name	Annotation	Owner	Community Rating ↑	Community Tags	Last Updated
Galaxy vs MEGAN	Comparison of Galaxy vs. MEGAN pipeline.	aun1	★★★★★	metagenomics megan galaxy	Mar 19, 2010
metagenomic analysis		aun1	★★★★★	metagenomics galaxy	Mar 19, 2010
SM_1186088	Datasets correspond to our paper published in Science by Peleg et al. entitled : Altered histone acetylation is associated with age-dependent memory impairment. Experiment layout: This history contains 4 datasets in the form of BED files of uniquely mapped reads produced after chip-seq for histone modifications H4K12ac and H3K9ac in mouse hippocampus of 3 months (young) and 16 months (old) mice after fear conditioning. For detailed information please refer to supplementary materials and methods of the respective work by peleg et al.	fischerlab	★★★★★		Apr 19, 2010
Variant Analysis for Sample E18	Perform a pileup analysis with default parameters to identify variants in sample E18.	jgoecks	★★★★★	snp pileup bowtie demo sample	2 minutes ago
get longest exon		henri	★★★★★	chr22 longest marc exon human workshop	Sep 02, 2010
FASTA to Tabular Test		JJ	★★★★★		Aug 26, 2010
EKLF		yzc109	★★★★★		Aug 24, 2010

Open "http://main.g2.bx.psu.edu/history/list_published?sort=rating&f-tags=All" in a new tab

Galaxy Pages

A web-based, interactive medium for presenting all aspects of an analysis: data, methods, and results

Galaxy Pages

The screenshot shows a web browser window displaying a Galaxy page. The browser's address bar shows the URL: <http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18>. The Galaxy logo is in the top left, and navigation links (Analyze Data, Workflow, Shared Data, Visualization, Help, User) are in the top right. The page title is "Variant Analysis of Embryonic Mouse Brain Tissue" by Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team. The "Results" section describes a variant analysis experiment. A green box highlights a dataset: "Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes". The "Method" section describes the analysis steps. A blue box highlights a history item: "Galaxy History | Variant Analysis for Sample E18". An orange box highlights a workflow: "Galaxy Workflow | Variant identification within annotated genes from NGS PE Data". The "References" section lists two papers. On the right, the "About this Page" sidebar shows the author's profile, related pages, and a rating section.

Galaxy | Published Page | Variant Analysis for sample E18

Published Pages | jgoecks | Variant Analysis for sample E18

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

[Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes](#)
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

[Galaxy History | Variant Analysis for Sample E18](#)
Perform a pileup analysis with default parameters to identify variants in sample E18.

Here is a workflow for performing this analysis:

[Galaxy Workflow | Variant identification within annotated genes from NGS PE Data](#)
Identify variants in annotated genes from NGS paired-end data.

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* 106, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).

[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079 (2009).

About this Page

Author
jgoecks

Related Pages
[All published pages](#)
[Published pages by jgoecks](#)

Rating
Community (0 ratings, 0.0 average)
Yours

Tags
Community: none
Yours:

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Published Pages | jgoecks | Variant Analysis for sample E18

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Galaxy History | Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.

8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage ≥ 30 .
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

Here is a workflow for performing this analysis:

Galaxy Workflow | Variant identification within annotated genes from NGS PE Data
Identify variants in annotated genes from NGS paired-end data.

References

About this Page

Author
jgoecks

Related Pages
All published pages
Published pages by jgoecks

Rating
Community (0 ratings, 0.0 average)
Yours: ★★★★★

Tags
Community: none
Yours:

Galaxy Pages

Galaxy | Dataset | Filter to get Variants from sample E18 where consensus base different than ref. base

http://main.g2.bx.psu.edu/u/jgoecks/d/fdcdbcf1207fbd2

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Dataset | Filter to get Variants from sample E18 where consensus base different than ref. base

Galaxy Dataset 'Filter to get Variants from sample E18 where consensus base different than ref. base'

Annotation: Filter pileup to find variants where the consensus base is different than the reference base.

! This dataset is large and only the first megabyte is shown below. | [Show all](#)

chr10	14465082	14465083	T	K	173	176	60	35	GGGGG.GGGGGGGG.GG....
chr10	14465083	14465084	G	K	144	144	60	35	T.....T..TTTT
chr10	14465485	14465486	C	T	129	129	60	34	t\$TTTTTTTTTTTTTTTTTT
chr10	19928287	19928288	G	A	135	135	60	36	AAAAAAAAAAAAAAAAAAAA
chr10	19928468	19928469	C	T	132	132	60	35	T\$tttttTttttttTttTTT
chr10	19928494	19928495	C	T	138	138	60	37	TTTTttttttttttTttTtt
chr10	19928543	19928544	A	G	147	147	60	40	GgGgGgggGGGGGGGGGG
chr10	28750217	28750218	C	T	138	138	60	37	TtttttTttttTttTttTTT
chr10	28750397	28750398	A	C	154	211	60	64	C\$. \$C\$C\$C\$C\$C\$C\$C\$C
chr10	28750423	28750424	C	T	113	113	60	35	T\$ttTttttTttTttTTT
chr10	28750446	28750447	A	G	165	165	60	46	G\$GGGggGggGGggGGGGG
chr10	28750548	28750549	G	C	255	255	60	97	C\$C\$C\$C\$C\$C\$C\$C\$C\$C
chr10	28750640	28750641	T	C	165	165	60	46	CccccccccCCCCCCCCC
chr10	28750746	28750747	G	A	202	202	60	58	AAAAAAAAAAAAAaaaaaaa
chr10	28750766	28750767	A	G	205	205	60	59	G\$g\$G\$g\$G\$gggggggggg
chr10	28750769	28750770	T	C	175	175	60	49	ccccCccCCCCCcCccccC
chr10	28750924	28750925	C	T	182	217	60	64	T\$T\$ttTttTttTttTttT
chr10	28751092	28751093	a	A	147	0	60	123	g\$g\$g\$g\$g\$g\$g\$g
chr10	28751096	28751097	a	A	212	0	60	119	g\$g\$,.....g\$g\$g\$g\$g\$g\$g
chr10	28751114	28751115	g	A	225	235	60	85	aaaa,aaaaaaaaaaaaaaaa
chr10	28751117	28751118	T	A	191	198	60	79	a\$a\$a\$a\$a\$a\$a\$a\$a\$a\$a
chr10	28918972	28918973	c	M	114	114	60	75	\$,\$,\$,\$,\$,\$,\$,\$,\$,\$,\$
chr10	28918975	28918976	a	A	177	0	60	63	,\$,\$,\$,\$,\$,\$,\$,\$,\$,\$
chr10	28918995	28918996	C	M	154	154	60	48	,\$,\$,\$,\$,\$,\$,\$,\$,\$,\$
chr10	33613489	33613490	G	A	82	114	60	30	a\$aaaaaaaaaaaaaaaaaaa
chr10	36721501	36721502	G	K	129	129	60	43	.\$A\$A\$A\$A\$A\$A\$A\$A\$A
chr10	36721507	36721508	C	Y	51	51	60	54	TTTT.....TTTTTTT
chr10	36721695	36721696	T	A	120	120	60	31	.\$.....\$.....\$.....
chr10	36805412	36805413	A	G	126	126	60	33	a\$a\$a\$a\$a\$a\$a\$a\$a\$a\$a
chr10	36805605	36805606	A	G	120	120	60	31	GGGGGGGGGGGGGGGGGG
chr10	36853176	36853177	C	Y	138	138	60	35	ggggggggggggggggggg
chr10	36853265	36853266	A	R	17	17	60	37	T.....T.....
chr10	36854675	36854676	T	K	99	99	60	48	T.....T.....
chr10	36854678	36854679	A	M	94	94	60	48	T.....T.....
chr10	36855347	36855348	a	R	159	159	60	50	T.....T.....
chr10	36855350	36855351	a	R	156	156	60	56	T.....T.....
chr10	36855356	36855357	a	A	134	0	60	49	T.....T.....
chr10	36855367	36855368	g	G	52	0	60	40	T.....T.....
chr10	36855370	36855371	g	G	25	0	60	32	T.....T.....
chr10	36855665	36855666	A	M	11	11	60	39	T.....T.....
chr10	39144591	39144592	G	C	126	126	60	33	T.....T.....
chr10	42722938	42722939	C	Y	63	63	60	34	T.....T.....

About this Dataset

Author
jgoecks



Tags

Community: none

Yours:

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Published Pages | jgoecks | Variant Analysis for sample E18

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Galaxy History | Variant Analysis for Sample E18
Perform a pileup analysis with default parameters to identify variants in sample E18.

8: SAM-to-BAM on data 7	Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
9: Generate pileup on data 8	Pileup analysis with default parameters
10: Filter pileup to get Variants from sample E18	Find variants with coverage >= 30.
13: Filter to get Variants from sample E18 where consensus base different than ref. base	Filter pileup to find variants where the consensus base is different than the reference base.
14: UCSC mm9 RefSeq Genes	UCSC mm9 RefSeq genes.
15: Intersect to get Variants from sample E18, consensus different, in RefSeq Genes	Variants with consensus different that occur in RefSeq genes.

Here is a workflow for performing this analysis:

Galaxy Workflow | Variant Identification within annotated genes from NGS PE Data
Identify variants in annotated genes from NGS paired-end data.

References
Open "http://main.g2.bx.psu.edu/history/imp?id=e0b8bd5d661b10c2" in a new tab

About this Page

Author
jgoecks

Related Pages
All published pages
Published pages by jgoecks

Rating
Community (0 ratings, 0.0 average)
Yours: ★★★★★

Tags
Community: none
Yours:

Galaxy Pages

The screenshot displays the Galaxy web interface. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The left sidebar lists various tools under categories like 'Tools', 'NGS TOOLBOX BETA', and 'RGENETICS'. The main panel shows the 'Filter pileup' tool configuration. The 'Select dataset' dropdown is set to '9: Generate pileup on data 8'. The 'which contains:' dropdown is set to 'Pileup with ten columns (with consensus)'. The 'Do not consider read bases with quality lower than:' is set to '20'. The 'Do not report positions with coverage lower than:' is set to '30'. The 'Only report variants?:' is set to 'Yes'. The 'Convert coordinates to intervals?:' is set to 'Yes'. The 'Print total number of differences?:' is set to 'Yes'. The 'Print quality and base string?:' is set to 'Yes'. The 'Execute' button is visible. The right sidebar shows a 'History' panel with a list of jobs, including '15: Variants from sample E18, consensus different, in RefSeq Genes', '14: UCSC mm9 RefSeq Genes', '13: Variants from sample E18 where consensus base different than ref. base', '10: Variants from sample E18', '9: Generate pileup on data 8', '8: SAM-to-BAM on data Z', '7: Map with Bowtie for Illumina on data 6 and data 5', and '6: E18 PE.2 Reads'. A table of genomic data is also visible in the history panel.

Filter pileup

Select dataset:
9: Generate pileup on data 8

which contains:
Pileup with ten columns (with consensus)

See "Types of pileup datasets" below for examples

Do not consider read bases with quality lower than:
20

No variants with quality below this value will be reported

Do not report positions with coverage lower than:
30

Pileup lines with coverage lower than this value will be skipped

Only report variants?:
Yes

See "Examples 1 and 2" below for explanation

Convert coordinates to intervals?:
Yes

See "Output format" below for explanation

Print total number of differences?:
Yes

See "Example 3" below for explanation

Print quality and base string?:
Yes

See "Example 4" below for explanation

Execute

What it does

Allows one to find sequence variants and/or sites covered by a specified number of reads with bases above a set quality threshold. The tool works on six and ten column pileup formats produced with *samtools pileup* command. However, it also allows you to specify columns in the input file manually. The tool assumes the following:

- the quality scores follow phred33 convention, where input qualities are ASCII characters equal to the Phred quality plus 33.
- the pileup dataset was produced by the *samtools pileup* command (although you

History

15: Variants from sample E18, consensus different, in RefSeq Genes

14: UCSC mm9 RefSeq Genes

13: Variants from sample E18 where consensus base different than ref. base

10: Variants from sample E18
26,742 regions, format: interval,
Run this job again

I display at UCSC main | view in GeneTrack | display at Ensembl Current

1. Chrom	2. Start	3. End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106
chr10	14465082	14465083	T	K	173
chr10	14465083	14465084	G	K	144
chr10	14465084	14465085	T	T	117

9: Generate pileup on data 8

8: SAM-to-BAM on data Z

7: Map with Bowtie for Illumina on data 6 and data 5

6: E18 PE.2 Reads

Open "http://main.g2.bx.psu.edu/tool_runner/rerun?id=1703758" in a new tab

Galaxy Pages

Galaxy | Published Page | Variant Analysis for sample E18

http://main.g2.bx.psu.edu/u/jgoecks/p/variant-analysis-for-sample-e18

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Published Pages | jgoecks | Variant Analysis for sample E18

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

[Galaxy Dataset | Intersect to get Variants from sample E18, consensus different, in RefSeq Genes](#)
Variants with consensus different that occur in RefSeq genes.

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

[Galaxy History | Variant Analysis for Sample E18](#)
Perform a pileup analysis with default parameters to identify variants in sample E18.

Here is a workflow for performing this analysis:

[Galaxy Workflow | Variant identification within annotated genes from NGS PE Data](#)
Identify variants in annotated genes from NGS paired-end data.

Import workflow

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* **106**, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).

[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).

Open "http://main.g2.bx.psu.edu/workflow/imp?id=58d16d45527990b7" in a new tab

About this Page

Author
jgoecks

Related Pages
[All published pages](#)
[Published pages by jgoecks](#)

Rating
Community (0 ratings, 0.0 average) ★★★★★
Yours ★★★★★

Tags
Community: none
Yours:

Creating a Page

The screenshot shows a web browser window with the Galaxy interface. The address bar shows the URL: http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427. The Galaxy logo is in the top left, and navigation links (Analyze Data, Workflow, Shared Data, Visualization, Help, User) are in the top right. The page title is "Page Editor | Title : Variant Analysis for sample E18". The editor toolbar includes buttons for Bold (B), Italic (I), Text color (x²), Background color (x₂), List (ul), Indent (list-item), Outdent (list-item), Link (link), Unlink (unlink), Image (image), and a "Paragraph type" dropdown. There are also buttons for "Insert Link to Galaxy Object" and "Embed Galaxy Object".

Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekrutenko, James Taylor, and The Galaxy Team

Results

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Embedded Galaxy History 'Variant Analysis for Sample E18'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

Here is a workflow for performing this analysis:

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* 106, 12741-12746 (2009).

Open # on this page in a new tab

Creating a Page

The screenshot shows the Galaxy web interface in a browser window. The address bar shows the URL: http://main.g2.bx.psu.edu/page/edit_content?id=d2523e005e1ec427. The Galaxy logo is in the top left, and navigation links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User' are in the top right. The page title is 'Page Editor | Title : Variant Analysis for sample E18'. Below the title bar is a toolbar with various editing tools. The main content area contains the following text:

To demonstrate how Galaxy can support accessible, reproducible, and transparent NGS re-sequencing studies, we performed a simple variant analysis experiment. This experiment identifies variants from a set of 4,536,964 RNA-seq reads obtained from sequencing a sample of mm9 brain tissue from day 18 of embryonic development.

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,625 where (a) the consensus base--as determined by the MAQ model--differs from the reference base and (b) read coverage at the base is 30x or greater. Of these potential variants, 2796 occur in known RefSeq Genes. These potential variants are:

Embedded Galaxy Dataset 'Variants from sample E18, consensus different, in RefSeq Genes'

[Do not edit this block; Galaxy will fill it in with the annotated dataset when it is displayed.]

Method

In the first step of this analysis, the reads were groomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqsanger. Next, the reads were trimmed from 36bp to 27bp to exclude base pairs with low quality scores; see [1] for this step's rationale and parameter choices. After grooming and trimming, the reads were mapped using the short-read mapper Bowtie [2]. A pileup analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 30+ reads. The complete analysis is contained in this history:

Embedded Galaxy History 'Variant Pileup Analysis for Sample E18'

[Do not edit this block; Galaxy will fill it in with the annotated history when it is displayed.]

Here is a workflow for performing this analysis:

Embedded Galaxy Workflow 'SNP identification within annotated genes from NGS PE Data'

[Do not edit this block; Galaxy will fill it in with the annotated workflow when it is displayed.]

References

[1] Han, X. et al. Transcriptome of embryonic and neonatal mouse cortex by high-throughput RNA sequencing. *Proceedings of the National Academy of Sciences* 106, 12741-12746 (2009).

[2] Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).

[3] Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078 -2079 (2009).

AI Opportunity: “Now What?”

What can I do with this dataset?

When should I use this tool?

What should be the next step in my analysis?

What are the “best practice” workflows for my analysis?

“Now What?” Factors

Past work

- ✦ what have I already done? (personal history)
- ✦ what have other people already done? (community history)

Approach

- ✦ exploration vs. focused analysis
- ✦ Google approach vs. scientist approach?

Overview

Genomics

Galaxy

- ✦ accessible, reproducible, and transparent science
- ✦ on the cloud
- ✦ visual analytics

Reflections on Galaxy

Three Ways to Use Galaxy

1. Download and Run Locally
2. Public Website (<http://usegalaxy.org>)
3. Run on the Cloud

1. Download and Run Locally

No configuration needed but everything can be configured

Prominent local installations at:

- ✦ Cold Spring Harbor Lab
- ✦ Dept. of Energy's Joint Genome Institute
- ✦ Harvard School of Public Health
- ✦ U of Texas System
- ✦ U of Oklahoma
- ✦ Netherlands Bioinformatics Center
- ✦ Oxford College

Requires computing resources, technical expertise, and maintenance

2. Galaxy main site (<http://usegalaxy.org>)

Public Website, anybody can use

~500 new users per month, ~100 TB of user data,
~130,000 analysis jobs per month, every month is
our busiest month ever...

Will continue to be maintained and enhanced, but
with limits and quotas

Centralized solution cannot scale to meet data
analysis demands

3. Galaxy on the Cloud

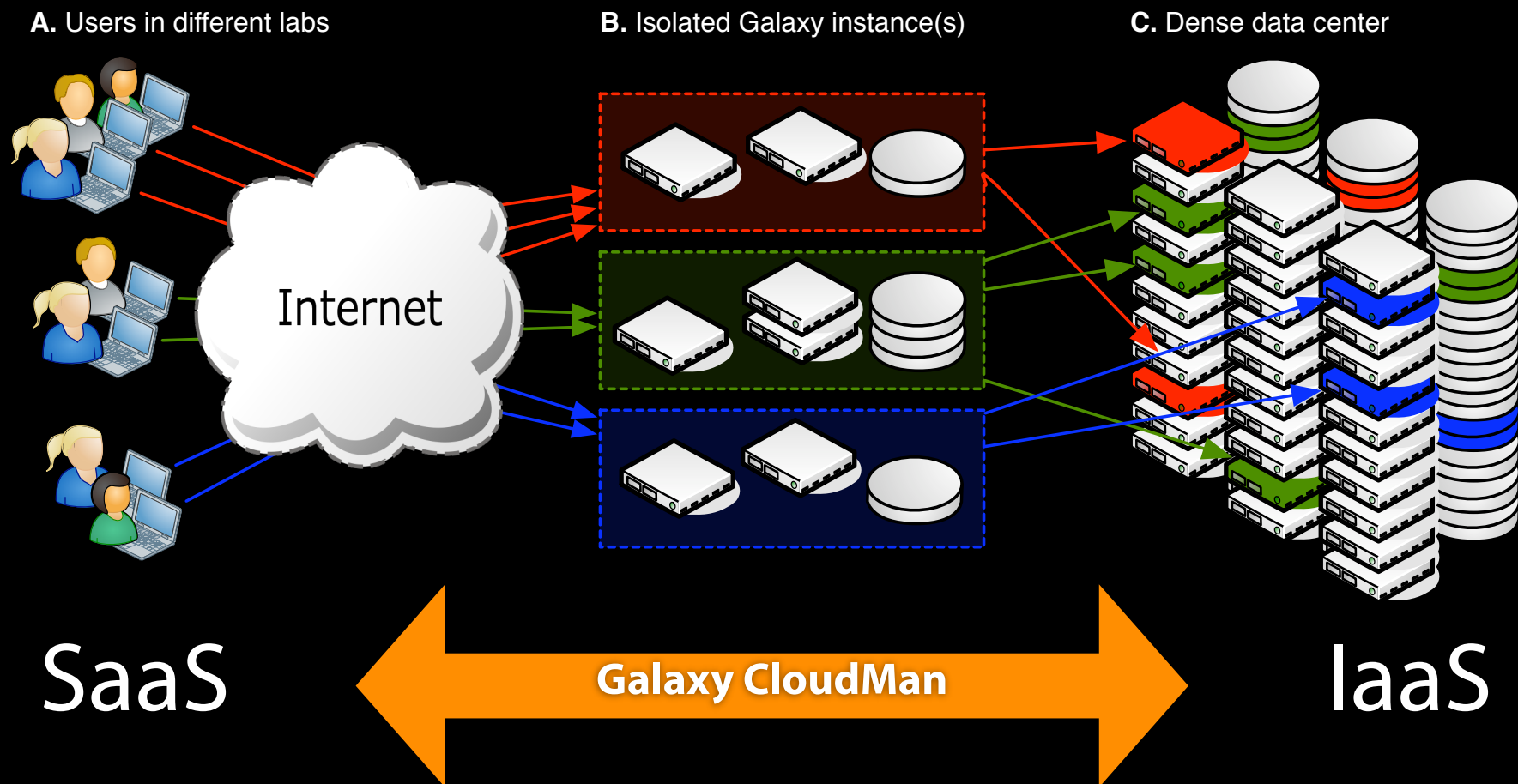
For extended or particular resource needs

- ✦ customization necessary
- ✦ oscillating data volume

Limited informatics expertise or infrastructure

Data production and (no?) sharing

The big picture



Galaxy CloudMan

Complete solution for instantiating, running and scaling cloud resources with automatically configured Galaxy

- ✦ Tools and reference datasets exceed Galaxy Main

No computational expertise needed

- ✦ no configuration needed but completely configurable

Reproducibility ensured

- ✦ Automated configuration for machine image, tools, and data
- ✦ Dynamic but persistent storage

Start an Instance

The screenshot shows the AWS Management Console interface. The top navigation bar includes links for AWS, Products, Developers, Community, Support, and Account. The 'Account' link is highlighted. The main content area displays the 'Amazon EC2 Console Dashboard' for the 'US East' region. The dashboard includes a 'Getting Started' section with a 'Launch Instance' button, a 'My Resources' section showing 0 Running Instances, 0 Elastic IPs, 6 EBS Volumes, and 12 EBS Snapshots, and a 'Service Health' section. The 'Request Instances Wizard' is open, showing the following configuration details:

- CHOOSE AN AMI:** Other Linux AMI ID ami-ed03ed84 (x86_64) Edit AMI
- Number of Instances:** 1
- Availability Zone:** No Preference
- Monitoring:** Disabled
- Instance Type:** Large (m1.large)
- Instance Class:** On Demand Edit Instance Details
- Kernel ID:** Use Default
- Ramdisk ID:** Use Default
- User Data:** testGC1|AKIAJKQI3RT... Edit Advanced Details
- Key Pair Name:** galaxy_keypair Edit Key Pair
- Security Group(s):** default, galaxyWeb Edit Firewall

The wizard includes a 'Launch' button at the bottom right.

Configure Your Instance

The screenshot shows a web browser window with the URL `ec2-50-16-1-149.compute-1.amazonaws.com/cloud`. The page title is "Galaxy Cloudman". A modal dialog box titled "Initial Cluster Configuration" is centered on the screen. The dialog contains a welcome message, a radio button selection for "Start a full Galaxy Cluster", a text input field with "1000" and "GB OK" next to it, a link for "Show more startup options", and a "Start Cluster" button. The background of the page is dimmed, showing sections for "Status" and "Cluster status log".

← → ↻ `ec2-50-16-1-149.compute-1.amazonaws.com/cloud` ☆ 🔑

Galaxy Cloudman Info: [report bugs](#) | [wiki](#) | [screencast](#)

Galaxy Cloudman

Welcome to Galaxy Cloudman. This application will allow you to manage this cluster and the services provided within. To get started, choose the type of cluster you'd like to work with and specify the size of your persistent data storage, if any.

☒ Start a full Galaxy Cluster. Specify initial storage size (in Gigabytes)

GB OK

[Show more startup options](#)

Start Cluster

Cluster status log +

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud instance and the services provided within. If this is your first time running this cluster, you will need to select an initial data volume size. Once the data store is configured, default services will start and you will be able to add and remove additional services as well as 'worker' nodes on which jobs are run.

Terminate cluster

Add nodes ▼

Remove nodes

Access Galaxy

Status

Cluster name: Heteroplasmy study 

Disk status: 50M / 1000G (1%) 

Worker status: Idle: 0 Available: 0 Requested: 0

Service status: Applications  Data 

External Logs: [Galaxy Log](#)



Autoscaling is **off**.
Turn on?

Cluster status log



← → ↻ ec2-50-17-119-106.compute-1.amazonaws.com/root ☆ 🔧

Galaxy Analyze Data Workflow Shared Data Visualization Help User

Tools Options ▾

- [Get Data](#)
- [Send Data](#)
- [ENCODE Tools](#)
- [Lift-Over](#)
- [Text Manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Fetch Alignments](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Wavelet Analysis](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Metagenomic analyses](#)
- [FASTA manipulation](#)
- NGS TOOLBOX BETA
- [NGS: QC and manipulation](#)
- [NGS: Assembly](#)
- [NGS: Mapping](#)
- [NGS: Indel Analysis](#)
- [NGS: Expression Analysis](#)
- [NGS: SAM Tools](#)
- [NGS: Peak Calling](#)
- [Human Genome Variation](#)
- [EMBOSS](#)

Welcome to Galaxy on the Cloud

History Options ▾

 Your history is empty. Click 'Get Data' on the left pane to start

The importance of sharing

Share entire Galaxy CloudMan cluster instances

- ✦ Fully automated solution

Publish an analysis

- ✦ In progress or otherwise

Use CloudMan as PaaS

- ✦ Deploy your own tool and make it available

Snapshot your instance

- ✦ Data
- ✦ Configuration

Scaling the infrastructure with the computation

The image shows a screenshot of the Galaxy Cloudman web interface. The browser address bar indicates the URL `ec2-50-17-119-106.compute-1.amazonaws.com/cloud`. The page title is "Galaxy Cloudman Console".

A modal dialog titled "Autoscaling Configuration" is open, explaining the concept of autoscaling and its benefits. It states: "Autoscaling attempts to automate the elasticity offered by cloud computing for this particular cluster. Once turned on, autoscaling takes over the control over the size of your cluster." It also mentions that the cluster will not shrink below a minimum number of worker nodes and will never grow larger than the maximum number of worker nodes specified.

The main console interface shows the following information:

- Cluster name:** share-an-instance demo
- Disk status:** 84M / 10G (1%)
- Worker status:** Idle: 0 Available: 0 Requested: 0
- Service status:** Applications (green dot) Data (green dot)
- External Logs:** [Galaxy Log](#)

Buttons for "Terminate cluster", "Add nodes", "Remove nodes", and "Access Galaxy" are visible. A "Cluster status log" section is at the bottom.

On the right side of the console, there is a summary box with a 4x4 grid of icons. The first icon is green, and the rest are grey. The text in the box reads: "Autoscaling is on. Turn off? Min nodes: 0 Max nodes: 15 Adjust limits?".

Exercising elasticity

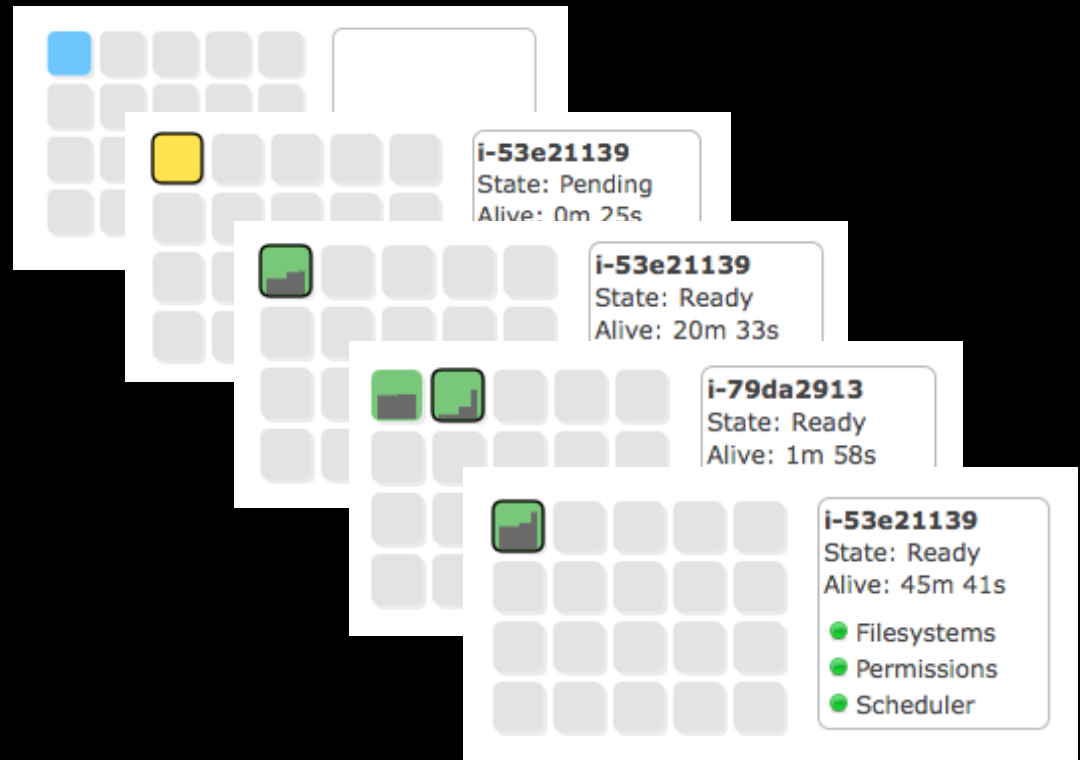
Fixed cluster size

5 nodes Computation time: 9 hrs
 Computation cost: \$20

20 nodes Computation time: 6 hrs
 Computation cost: \$50

Dynamic cluster size

1 to 16
nodes Computation time: 6 hrs
 Computation cost: \$20

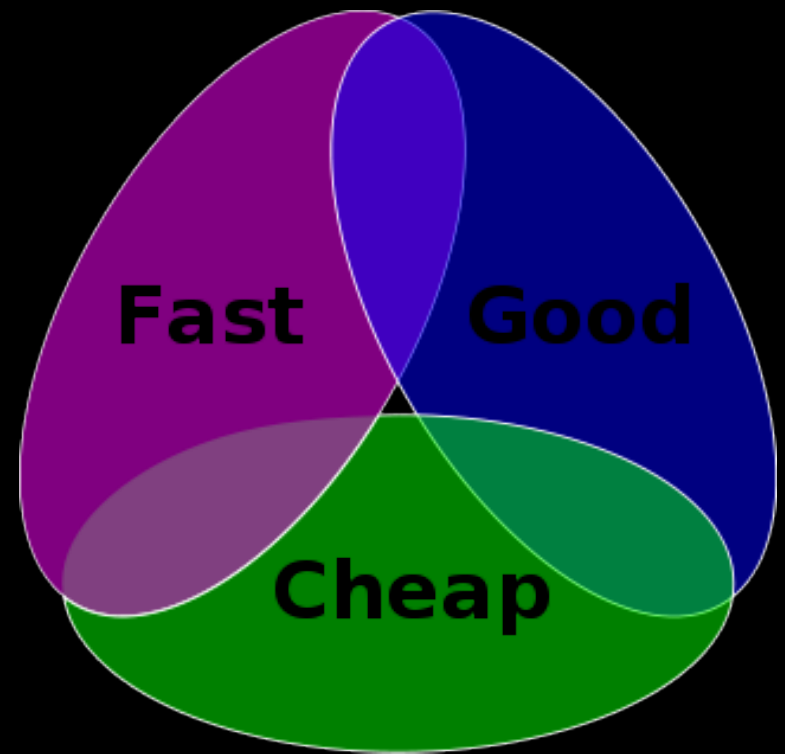


AI Opportunity: Smart Resource Usage

How long will this tool run?

How much will it cost?

How much computing
power do I need for this
analysis?



User Support vs. Automated

Users like (or require) control

- ✦ cognitive models of dynamic computing: parallelization and autoscaling?

Automation requires tool profiling

- ✦ local vs. global, parallelization
- ✦ linear vs. non-linear (e.g. graphs)

Overview

Genomics

Galaxy

- ✦ accessible, reproducible, and transparent science
- ✦ on the cloud
- ✦ **visual analytics**

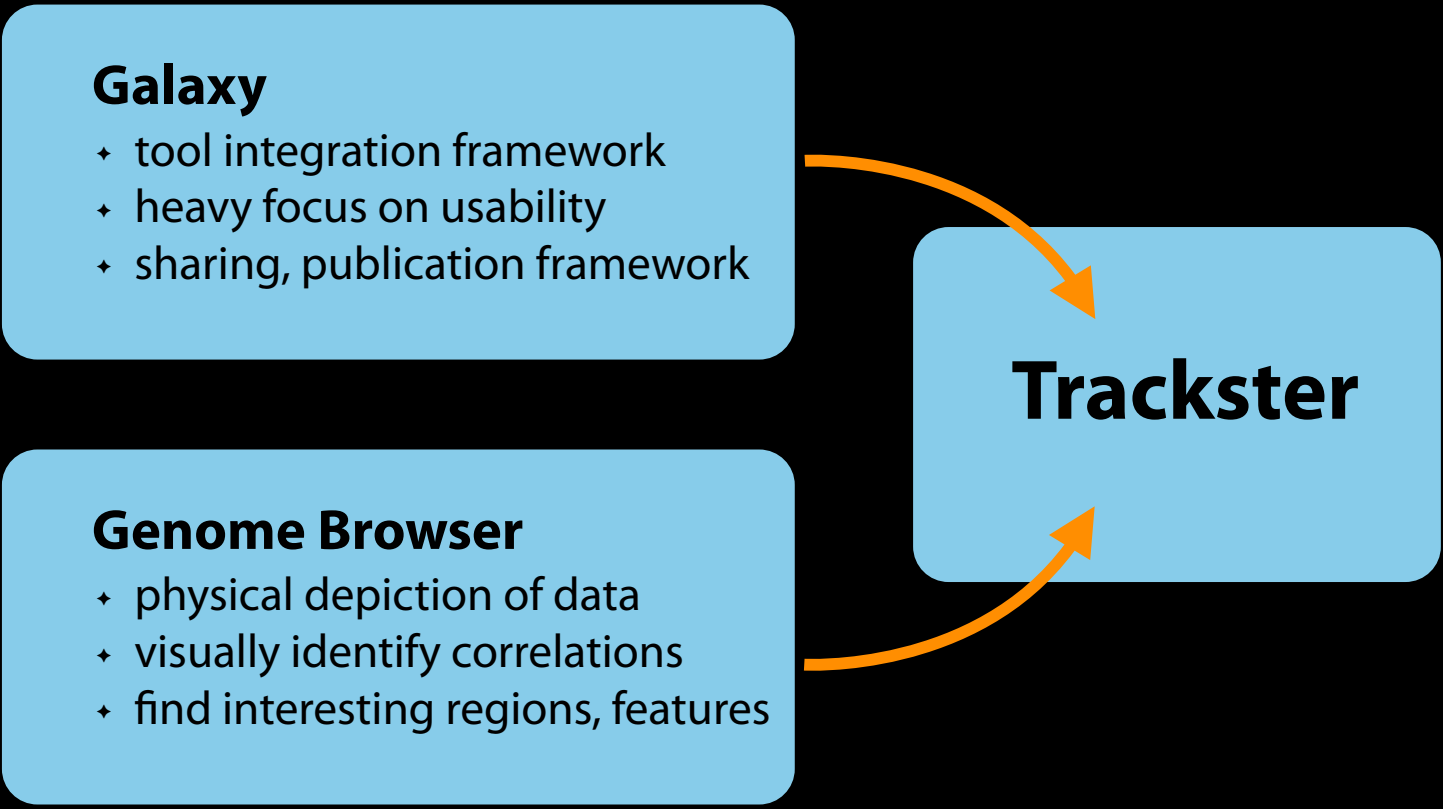
Reflections on Galaxy

Galaxy

- ✦ tool integration framework
- ✦ heavy focus on usability
- ✦ sharing, publication framework

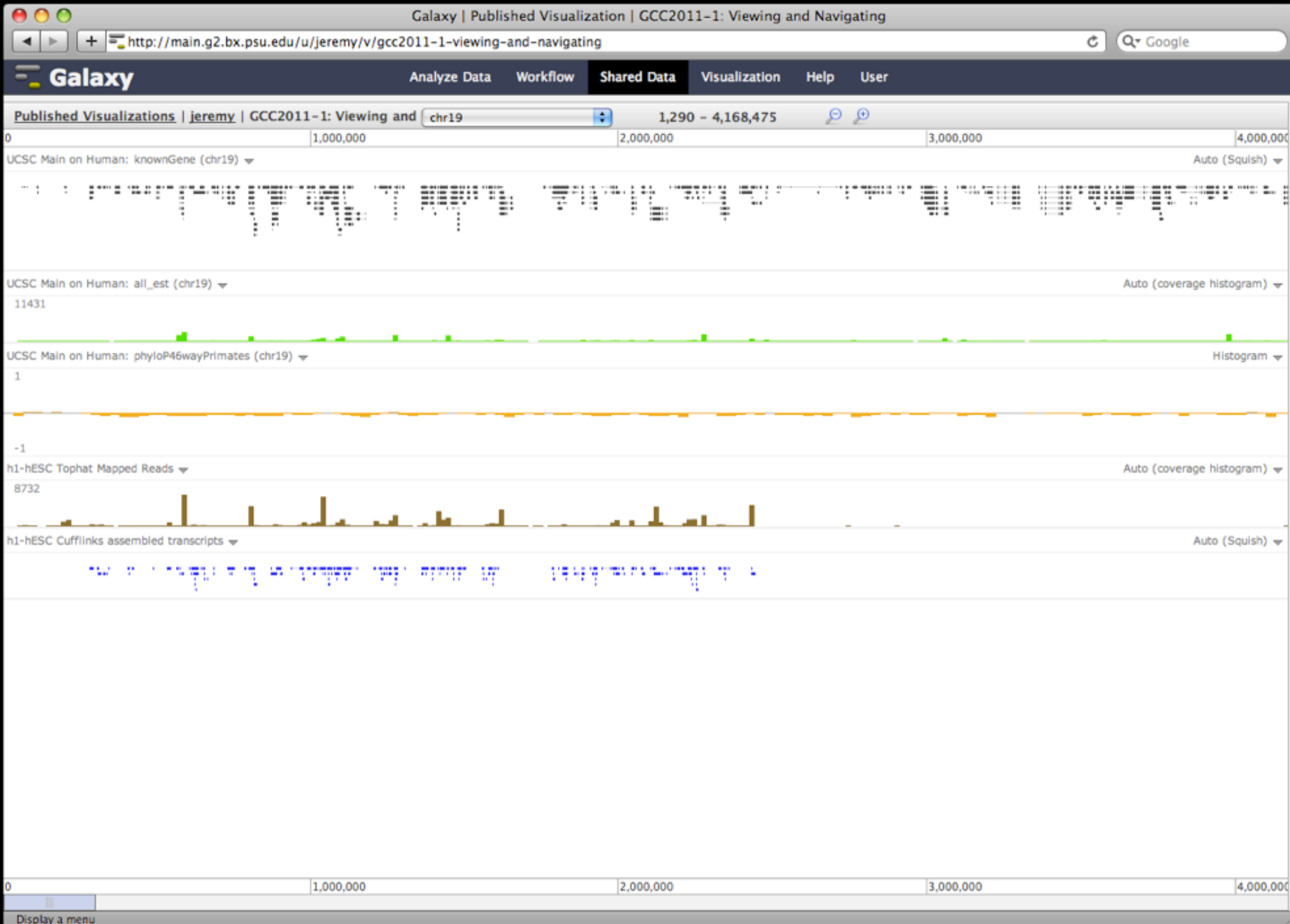
Genome Browser

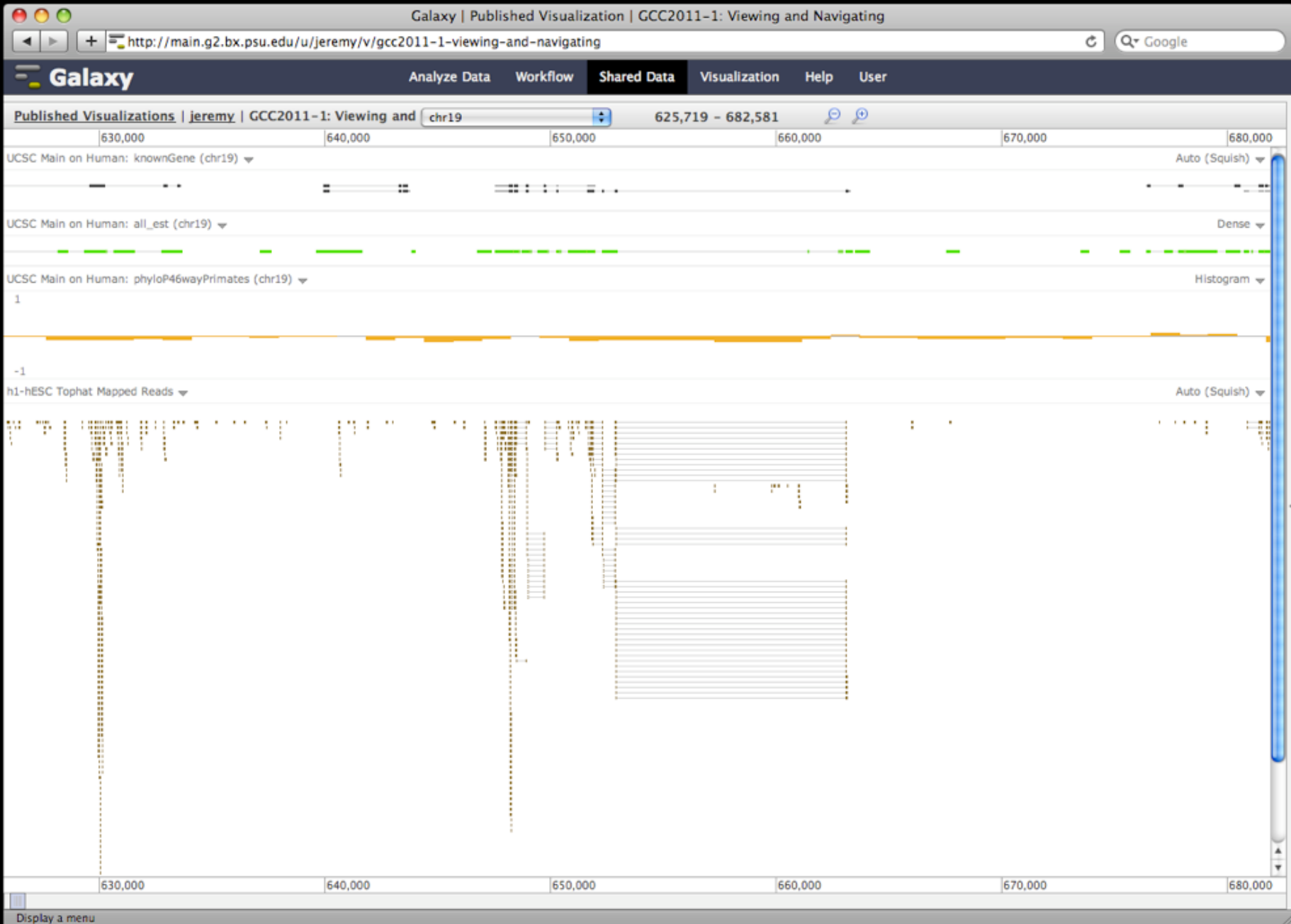
- ✦ physical depiction of data
- ✦ visually identify correlations
- ✦ find interesting regions, features

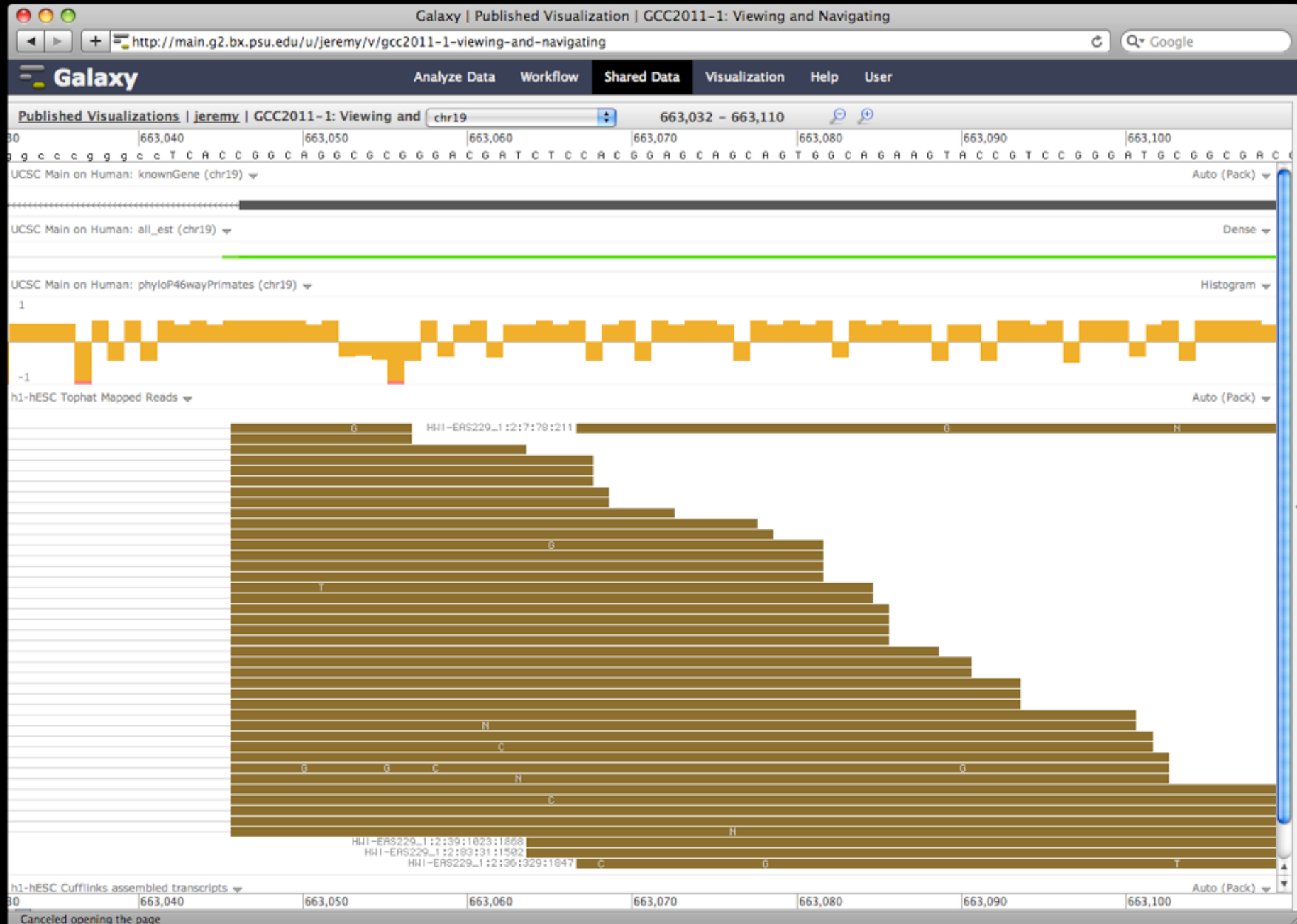


```
graph LR; Galaxy[Galaxy] --> Trackster[Trackster]; GB[Genome Browser] --> Trackster;
```

Trackster







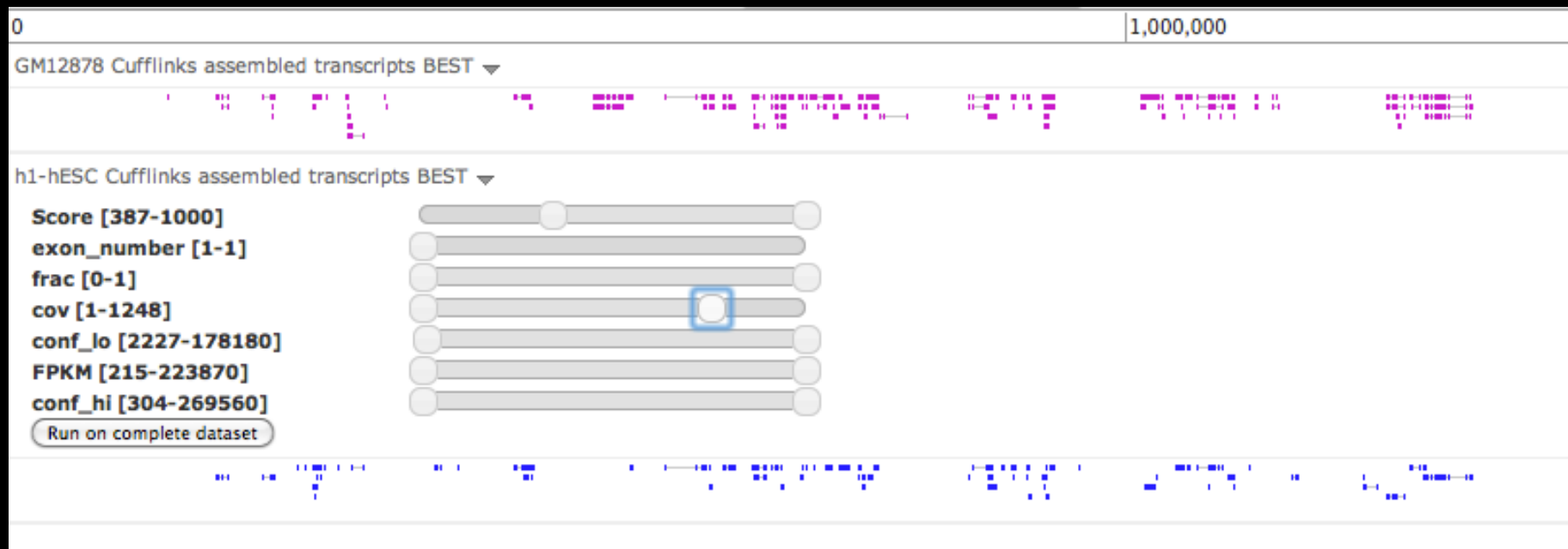
HTS Analysis Challenges

Complex tools, parameter dependent

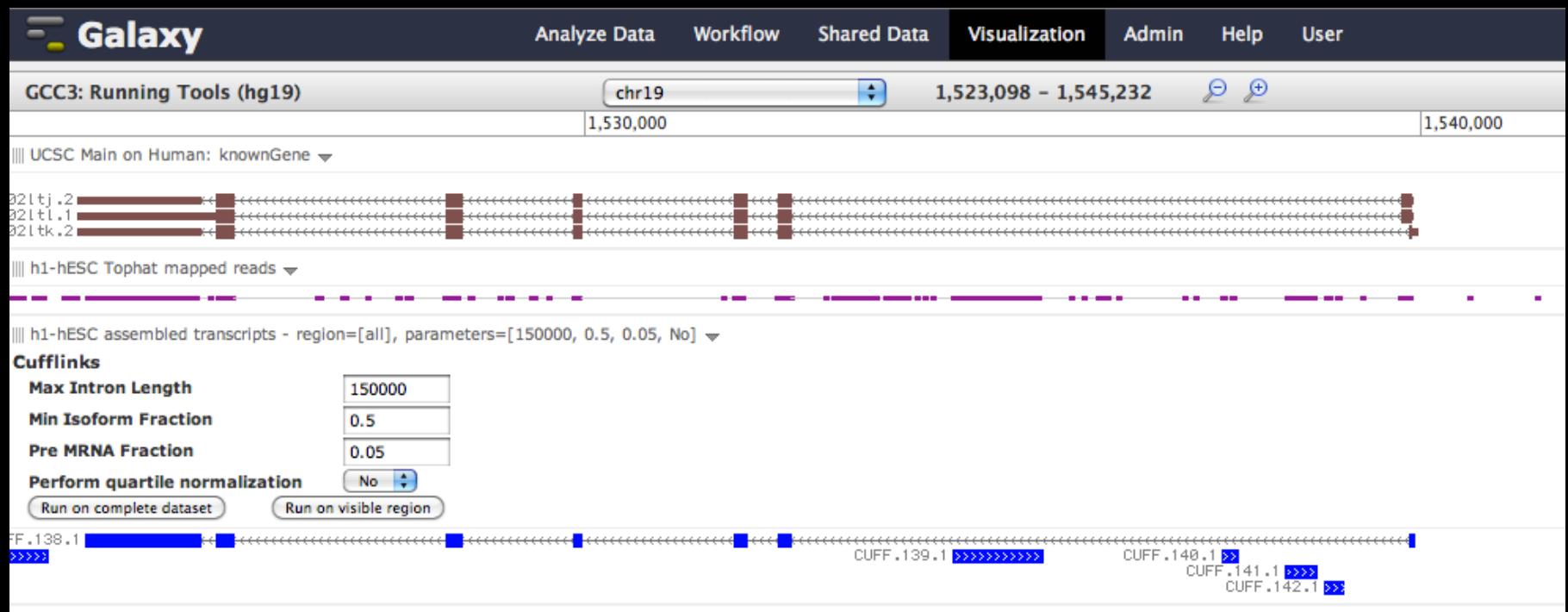
Analysis and visualization not integrated

Want to be able to experiment

Dynamic Filtering



Integrating Tools and Visualization



||| h1-hESC assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, No] ▼

Cufflinks

Max Intron Length

Min Isoform Fraction

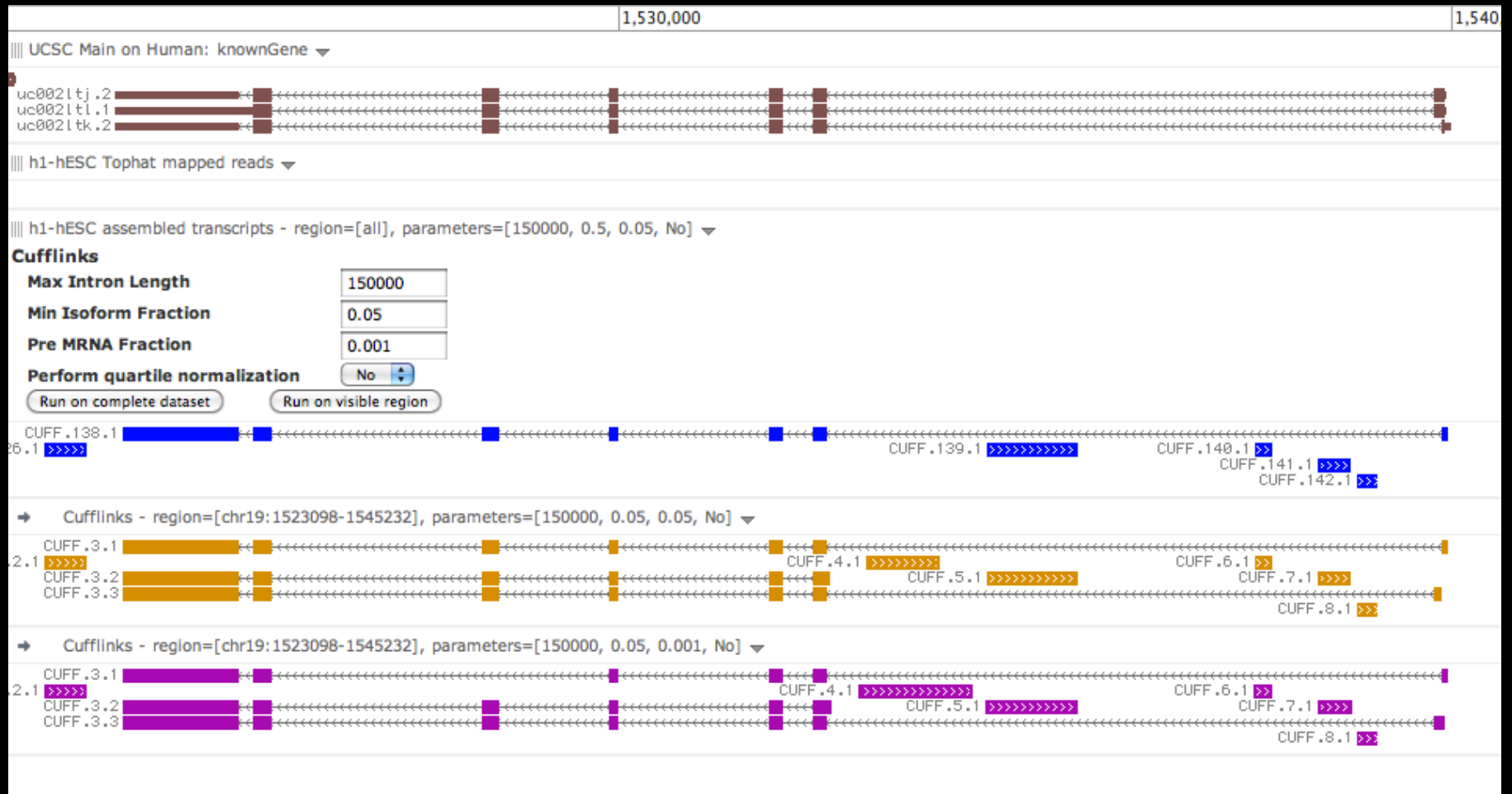
Pre mRNA Fraction

Perform quartile normalization



➔ Cufflinks - region=[chr19:1523098-1545232], parameters=[150000, 0.05, 0.05, No] ▼





Experimentation in Trackster

Tools integrated in visualization environment

Dynamically filter:

- ✦ visually identify features that match ranges
- ✦ on whole dataset

Run tools:

- ✦ quickly on visible region
- ✦ on whole dataset

AI Opportunity: Smart Visual Analytics

What datasets should I include in my visualization?

What are the interesting areas of my visualization?

What tools should I use in my visualization?

Combining Models

User modeling

- ✦ interests + past activities + community activities

Data modeling

- ✦ relationships amongst datasets

“Visual acuity” modeling: what can and should be seen in a visualization

Overview

Genomics

Galaxy

- ✦ accessible, reproducible, and transparent science
- ✦ on the cloud
- ✦ visual analytics

Reflections on Galaxy

Accessibility

Accessibility drives usage

- ✦ transparency and esp. reproducibility are secondary
- ✦ few incentives for reproducible research (for now)

A win-win for everyone

- ✦ tool developers: free GUI, more exposure
- ✦ users: easy to run tools, consistent interface
- ✦ everyone: amplification b/c tools can be chained to create complex analyses

Approach

Computer science research is driven by scientific needs

- ✦ we do biology research ourselves with Galaxy
- ✦ publish largely in biology journals

Don't need do everything, but what is done is done well

- ✦ software engineering, community support are priorities

Galaxy used in and cited by papers in *Science*, *Nature*, *Genome Research*, *Genome Biology*, and more

Challenges Going Forward

Promoting community involvement

- ✦ tools, assays, analyses growing too fast for us alone
- ✦ facilitate community contributions and usage of contributions

Scaling to many, many Galaxies

- ✦ moving objects between Galaxies while ensuring reproducibility
- ✦ enabling users to find useful “stuff”

Novel application areas

- ✦ genomics ideal application area -- what next?

Thanks! Questions, discussion?

<https://bitbucket.org/galaxy/galaxy-central/wiki/Citations>

Public server: <http://usegalaxy.org>

Download and run: <http://getgalaxy.org>

On the cloud: <http://usegalaxy.org/cloud>

jeremy.goecks@emory.edu