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

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

Genomics

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

Reflections on Galaxy
Molecular Biology Primer

Molecular Biology Primer

Molecular Biology Primer

Central Dogma of Molecular Biology: Eukaryotic Model

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
- ...

8
Trends in Genomics (1)

![Cost per Q20 Megabase ($) graph](image)

- Moore's Law
- $/MQ20

14,000X
Trends in Genomics (2)

Cost and Growth of Bases

- Billions of bases
- Cost per million base pairs of sequence (log scale)

SOURCE: NCBI

- 2000
- 2001
- 2002
- 2003
- 2004
- 2005
- 2006
- 2007
- 2008
- 2009

Cost ($)

- $10,000
- $1,000
- $100
- $10
- $1

GenBank
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...

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?
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

Galaxy

Map with Bowtie for Illumina

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

- Use 2 built-in index

Built-ins were indexed using default options

Select a reference genome:

- mm9

- If your genome of interest is not listed – contact Galaxy team

Is this library mate-paired?

- Paired-end

Forward FASTQ file:

- 1: 118 PE.1 Reads

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file:

- 1: 118 PE.1 Reads

Must have Sanger-scaled quality values with ASCII offset 33

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/ -fr/ -fr):

- FR (for Illumina)

Bowtie settings to use:

- Commonly used

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:

- Yes

Bowtie produces SAM with several lines of header information by default

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.
Connecting 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 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

Galaxy

Analyze Data Workflow Shared Data Visualization Help User

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.1 Reads Groomed, Trimmed
5: E18 PE.1 Reads Groomed, Trimmed
4: E18 PE.2 Reads Groomed, Trimmed
3: E18 PE.1 Reads Groomed
2: E18 PE.2 Reads
1: E18 PE.1 Reads

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

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

Accessibility
Connecting 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 expression

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
Connecting 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 a specific criteria
- Operate on Genomes
  - Intersect the intersections of queries
  - Subtract the intersections of queries
  - Merge the overlaps of a query
  - Compound query

NGS: SAM Tools
- Filter SAM on quality values
- Convert SAM to BAM format
- BAM to SAM format
- 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
Connecting Users with Tools

Accessibility
A Tool in Galaxy

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%)

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%)


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%)

Metadata = Reproducibility
Automatic Metadata

Map with Bowtie for Illumina

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

Built-ins were indexed using default options

Select a reference genome:
- mm9

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

Is this library mate-paired?
- Paired-end

Forward FASTQ file:
- 5: ELB PE.1 Reads Graded, Trimmed
  Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file:
- 6: ELB PE.2 Reads Graded, Trimmed
  Must have Sanger-scaled quality values with ASCII offset 33

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

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:
- Yes
  Bowtie produces SAM with several lines of header information by default

Execute
User Metadata

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.

Find variants with coverage >= 30 and quality score >= 20.

<table>
<thead>
<tr>
<th>Chrom</th>
<th>Start</th>
<th>End</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr10</td>
<td>6882036</td>
<td>6882037</td>
<td>A</td>
<td>A</td>
<td>107</td>
</tr>
<tr>
<td>chr10</td>
<td>14243075</td>
<td>14243076</td>
<td>G</td>
<td>G</td>
<td>96</td>
</tr>
<tr>
<td>chr10</td>
<td>14243079</td>
<td>14243080</td>
<td>C</td>
<td>C</td>
<td>106</td>
</tr>
<tr>
<td>chr10</td>
<td>14465082</td>
<td>14465083</td>
<td>T</td>
<td>K</td>
<td>173</td>
</tr>
<tr>
<td>chr10</td>
<td>14465083</td>
<td>14465084</td>
<td>G</td>
<td>K</td>
<td>144</td>
</tr>
<tr>
<td>chr10</td>
<td>14465084</td>
<td>14465085</td>
<td>T</td>
<td>T</td>
<td>117</td>
</tr>
</tbody>
</table>
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
Galaxy Workflows

- **Tool**: Upload File
  - History items created: 1: E18 PE.1 Reads

- **Tool**: FASTQ Groomer
  - History items created: 2: E18 PE.2 Reads

- **Tool**: FASTQ Trimmer
  - History items created: 3: E18 PE.1 Reads Groomed
  - History items created: 4: E18 PE.2 Reads Groomed

- **Tool**: Map with Bowtie for Illumina
  - History items created: 5: E18 PE.1 Reads Groomed, Trimmed
  - History items created: 6: E18 PE.2 Reads Groomed, Trimmed

- **Tool**: SAM-to-BAM
  - History items created: 7: Map with Bowtie for Illumina on data 6 and data 5

- **Tool**: Generate pileup
  - History items created: 8: SAM-to-BAM on data 7
  - History items created: 9: Generate pileup on data 8
Galaxy Workflows

Reproducibility
Sharing, Collaborating, and Publishing with Galaxy
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 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.q2.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 History 'Variant Analysis for Sample E18'

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

Dataset

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

Annotation
- Forward reads from sample E18.
- Reverse reads from sample E18.
- Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3.
- Groom reads to convert quality scores from Solexa 1.0 to Solexa 1.3.
- Trim reads from 3' end to remove low-quality nts.
- Trim reads from 3' to remove low-quality nts.
- Map paired-end reads with default parameters.
- Need to convert Bowtie SAM to BAM so that pileup analysis can be performed.
- Pileup analysis with default parameters.
- Find variants with coverage >= 30.
- Filter pileup to find variants where the consensus base is different than the reference base.
- UCSC mm9 RefSeq Genes.
- Variants with consensus different that occur in RefSeq genes.
<table>
<thead>
<tr>
<th>Name</th>
<th>Annotation</th>
<th>Owner</th>
<th>Community Rating</th>
<th>Community Tags</th>
<th>Last Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galaxy vs MEGAN</td>
<td>Comparison of Galaxy vs. MEGAN pipeline.</td>
<td>aun1</td>
<td>★★★★★</td>
<td>metagenomics</td>
<td>Mar 19, 2010</td>
</tr>
<tr>
<td>metagenomic analysis</td>
<td></td>
<td>aun1</td>
<td>★★★★★</td>
<td>metagenomics</td>
<td>Mar 19, 2010</td>
</tr>
<tr>
<td>SM_118608R</td>
<td>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.</td>
<td>fischerlab</td>
<td>★★★★★</td>
<td>snp, pileup</td>
<td>Apr 19, 2010</td>
</tr>
<tr>
<td>Variant Analysis for Sample E18</td>
<td>Perform a pileup analysis with default parameters to identify variants in sample E18.</td>
<td>jgoecks</td>
<td>★★★★★</td>
<td>bowtie, demo, sample</td>
<td>2 minutes ago</td>
</tr>
<tr>
<td>get longest exon</td>
<td></td>
<td>henri</td>
<td>★★★★★</td>
<td>chr22, longest, marc, exon, human, workshop</td>
<td>Sep 02, 2010</td>
</tr>
<tr>
<td>FASTA to Tabular Test</td>
<td></td>
<td>jj</td>
<td>★★★★★</td>
<td></td>
<td>Aug 26, 2010</td>
</tr>
<tr>
<td>EKLF</td>
<td></td>
<td>yzc109</td>
<td>★★★★★</td>
<td></td>
<td>Aug 24, 2010</td>
</tr>
</tbody>
</table>
Galaxy Pages

A web-based, interactive medium for presenting all aspects of an analysis: data, methods, and results
Variant Analysis of Embryonic Mouse Brain Tissue

Jeremy Goecks, Anton Nekutenko, 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 4,336,964 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:

References


Galaxy Pages

Published Pages | jgoecks | Variant Analysis for sample E18

The initial analysis produced support for 27,742 possible variants. Of these possible variants, there are 5,623 where (a) the consensus base— as determined by the MAQ alignment—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 trimmed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastq-canger. 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 trimming and aligning, 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
Galaxy Pages
Galaxy Pages

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 (i) the consensus base—as determined by the MAC model—differs from the reference base and (ii) read coverage at the base is 3x or greater. Of these potential variants, 2,796 occur in known RefSeq Genes. These potential variants are:

**Method**

In the first step of this analysis, the reads were grommed 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:

**References**

Open: [http://main.g2.bx.psu.edu/history/imp?h=3e06bb566011b012](http://main.g2.bx.psu.edu/history/imp?h=3e06bb566011b012) in a new tab
Galaxy Pages

Filter pileup

Select dataset:
9: Generate pileup on data 8

which contains:
1. Pileup with tee columns (with consensus)
2. 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

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
Galaxy Pages

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 gnomed to convert their quality scores from Solexa 1.0 to Solexa 1.3/Fastqiang. Next, the reads were trimmed from 38ph to 27ph 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 30x reads. The complete analysis is contained in this history.

References


Open "http://main.g2.bx.psu.edu/workflow/impId=18d1645527905027" in a new tab
Creating a Page

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 m/m9 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, 2196 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/Fastqtagger. 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 30x 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

Creating a Page

Galaxy

Page Editor | Title: Variant Analysis for sample E18

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 5,615 where (a) the consensus base—as determined by the MAQ model—differs from the reference base and (b) read coverage at the base is 10x 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 pilot analysis using SAMtools [3] was then performed and was filtered to identify variants supported by 10+ 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


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

A. Users in different labs

B. Isolated Galaxy instance(s)

C. Dense data center

SaaS

Galaxy CloudMan

IaaS

Internet
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
Configure Your Instance
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.

**Status**

- **Cluster name:** Heteroplasy 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?
Welcome to Galaxy on the Cloud

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

Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application allows you to manage this instance of Galaxy CloudMan. Your previous data store has been reconnected. Once the cluster has initialized, use the controls below to add and remove ‘worker’ nodes for running jobs.

Status

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

Cluster status log

Autoscaling Configuration

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.

Autoscaling is simple, just specify the cluster size limits you want to work within and use your cluster as you normally do. The cluster will not automatically shrink to less than the minimum number of worker nodes you specify and it will never grow larger than the maximum number of worker nodes you specify.

While respecting the set limits, if there are more jobs than the cluster can comfortably process at a given time autoscaling will automatically add compute nodes if there are cluster nodes sitting idle at the end of an hour autoscaling will terminate those nodes reducing the size of the cluster and your cost.

Once turned on, the cluster size limits respected by autoscaling can be adjusted or autoscaling can be turned off.
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?

http://en.wikipedia.org/wiki/Project_triangle
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
**Genome Browser**
- physical depiction of data
- visually identify correlations
- find interesting regions, features

**Galaxy**
- tool integration framework
- heavy focus on usability
- sharing, publication framework

**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
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