Galaxy
for high-throughput sequence data analysis

http://usegalaxy.org
The Galaxy Team

Supported by the National Human Genome Research Institute (HG005542, HG004909, HG005133), the National Science Foundation (DBI-0850103), Penn State University, Emory University, and the Pennsylvania Department of Public Health
Are data intensive techniques accessible to researchers?

- For example, high-throughput sequencing:
  - Increasingly availability of instruments, with different strengths, enabling a huge number of high-throughput functional assays
  - However, making use of these techniques requires sophisticated and computationally intensive approaches
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 facilitate transparent communication of those analysis?
  - How best to ensure that analysis are reproducible?
A crisis in genomics research: 
reproducibility
Key Reproducibility Problems

- **Datasets**: not all available, difficult to access
- **Tools**: inaccessible, hard to record details
- **Publication**: results, data, methods separate
Microarray Experiment Reproducibility

- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible

Problems

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

Galaxy: accessible analysis system
What is Galaxy?

- **A free (for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage

- **Open source software** that makes integrating your own tools and data and customizing for your own site simple
Integrating existing tools into a uniform framework

- Defined in terms of an abstract interface (inputs and outputs)
- In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning
Galaxy analysis interface

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
Automatically tracks every step of every analysis
As well as user-generated metadata and annotation...
Galaxy workflow system

- Workflows can be constructed from scratch or extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis
Analyzing high throughput sequence data with Galaxy

- The Galaxy framework is generic; supporting a new type of analysis is as simple as integrating tools
- Galaxy is well suited to large-scale analysis
  - Allows tools to work with data in native, efficient formats
  - Integrates easily with cluster computing resources
(some) Galaxy tools for sequence data analysis

<table>
<thead>
<tr>
<th>NGS: QC and manipulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILLUMINA DATA</td>
</tr>
<tr>
<td>• FASTQ Groomer convert between various FASTQ quality formats</td>
</tr>
<tr>
<td>• FASTQ splitter on joined paired end reads</td>
</tr>
<tr>
<td>• FASTQ joiner on paired end reads</td>
</tr>
<tr>
<td>• FASTQ Summary Statistics by column</td>
</tr>
<tr>
<td>ROCHE-454 DATA</td>
</tr>
<tr>
<td>• Build base quality distribution</td>
</tr>
<tr>
<td>• Select high quality segments</td>
</tr>
<tr>
<td>• Combine FASTA and QUAL into FASTQ</td>
</tr>
<tr>
<td>AB-SOLID DATA</td>
</tr>
<tr>
<td>• Convert SOLiD output to fastq</td>
</tr>
<tr>
<td>• Compute quality statistics for SOLiD data</td>
</tr>
<tr>
<td>• Draw quality score boxplot for SOLiD data</td>
</tr>
<tr>
<td>GENERIC FASTQ MANIPULATION</td>
</tr>
<tr>
<td>• Filter FASTQ reads by quality score and length</td>
</tr>
<tr>
<td>• FASTQ Trimmer by column</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGS TOOLBOX BETA</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGS: QC and manipulation</td>
</tr>
<tr>
<td>ILLUMINA</td>
</tr>
<tr>
<td>• Map with Bowtie for Illumina</td>
</tr>
<tr>
<td>• Map with BWA for Illumina</td>
</tr>
<tr>
<td>ROCHE-454</td>
</tr>
<tr>
<td>• Lastz map short reads against reference sequence</td>
</tr>
<tr>
<td>• Megablast compare short reads against hgs, nt, and wgs databases</td>
</tr>
<tr>
<td>• Parse blast XML output</td>
</tr>
<tr>
<td>AB-SOLID</td>
</tr>
<tr>
<td>• Map with Bowtie for SOLID</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGS: SAM Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGS: Indel Analysis</td>
</tr>
<tr>
<td>NGS: Peak Calling</td>
</tr>
<tr>
<td>NGS: RNA Analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGS TOOLBOX BETA</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGS: QC and manipulation</td>
</tr>
<tr>
<td>NGS: Mapping</td>
</tr>
<tr>
<td>NGS: SAM Tools</td>
</tr>
<tr>
<td>• Filter SAM on bitwise flag values</td>
</tr>
<tr>
<td>• Convert SAM to interval</td>
</tr>
<tr>
<td>• SAM-to-BAM converts SAM format to BAM format</td>
</tr>
<tr>
<td>• BAM-to-SAM converts BAM format to SAM format</td>
</tr>
<tr>
<td>• Merge BAM Files merges BAM files together</td>
</tr>
<tr>
<td>• Generate pileup from BAM dataset</td>
</tr>
<tr>
<td>• Filter pileup on coverage and SNPs</td>
</tr>
<tr>
<td>• Pileup-to-Interval condenses pileup format into ranges of bases</td>
</tr>
<tr>
<td>• flagstat provides simple stats on BAM files</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGS: SAM Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGS: Indel Analysis</td>
</tr>
<tr>
<td>NGS: Peak Calling</td>
</tr>
<tr>
<td>NGS: RNA Analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGS: SAM Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGS: RNA Analysis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NGS: RNA Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-SEQ</td>
</tr>
<tr>
<td>• Tophat Find splice junctions using RNA-seq data</td>
</tr>
<tr>
<td>• Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data</td>
</tr>
<tr>
<td>• Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments</td>
</tr>
<tr>
<td>• Cuffdiff find significant changes in transcript expression, splicing, and promoter use</td>
</tr>
</tbody>
</table>

FILTERING
Example: Workflow for differential expression analysis of RNA-seq using Tophat/Cufflinks tools
Community of tool developers
<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convert Formats</td>
<td>Tools for converting data formats</td>
<td>5</td>
</tr>
<tr>
<td>Data Source</td>
<td>Tools for retrieving data from external data sources</td>
<td>1</td>
</tr>
<tr>
<td>Fasta Manipulation</td>
<td>Tools for manipulating fasta data</td>
<td>5</td>
</tr>
<tr>
<td>Next Gen Mappers</td>
<td>Tools for the analysis and handling of Next Gen sequencing data</td>
<td>7</td>
</tr>
<tr>
<td>Ontology Manipulation</td>
<td>Tools for manipulating ontologies</td>
<td>1</td>
</tr>
<tr>
<td>SAM</td>
<td>Tools for manipulating alignments in the SAM format</td>
<td>0</td>
</tr>
<tr>
<td>Sequence Analysis</td>
<td>Tools for performing Protein and DNA/RNA analysis</td>
<td>10</td>
</tr>
<tr>
<td>SNP Analysis</td>
<td>Tools for single nucleotide polymorphism data such as WGA</td>
<td>1</td>
</tr>
<tr>
<td>Statistics</td>
<td>Tools for generating statistics</td>
<td>1</td>
</tr>
<tr>
<td>Text Manipulation</td>
<td>Tools for manipulating data</td>
<td>3</td>
</tr>
<tr>
<td>Visualization</td>
<td>Tools for visualizing data</td>
<td>1</td>
</tr>
<tr>
<td>Name</td>
<td>Description</td>
<td>Version</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>AGILE</td>
<td>Quickly match reads to a reference genome or sequence file</td>
<td>1.0.0</td>
</tr>
<tr>
<td>assemblystats</td>
<td>Summarise an assembly (e.g. N50 metrics)</td>
<td>1.0.1</td>
</tr>
<tr>
<td>Divide FASTQ file into paired and unpaired reads</td>
<td>using the read name suffixes</td>
<td>0.0.4</td>
</tr>
<tr>
<td>FastQC</td>
<td>quality control checks on raw sequence data</td>
<td>1.0.0</td>
</tr>
<tr>
<td>Filter FASTA by ID</td>
<td>from a tabular file</td>
<td>0.0.3</td>
</tr>
</tbody>
</table>
This is the latest approved version of this tool suite

**Mothur Metagenomics**

**Tool Id:**
Mothur_toolsuite

**Version:**
1.15.1

**Description:**
Mothur metagenomics commands as Galaxy tools

**User Description:**
Provides galaxy tools for the commands in the Mothur metagenomics package: http://www.mothur.org/wiki/Main_Page

**Uploaded by:**
jjohnson

**Date uploaded:**
about 22 hours ago

**Categories:**
- Sequence Analysis

**Tool Contents**

- Mothur_toolsuite_1.15.1.tar.gz
- mothur/
- mothur/tools/
- mothur/tools/mothur/
- mothur/tools/mothur/split.abund.xml
Data management
Everything can be shared and published
The Galaxy team is a part of BX at Penn State.

This project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, and The Institute for CyberScience at Penn State.

Galaxy build: $Rev 4802:ea7b055efbfa$
Other information about 01Feb2010_In7 CTCF CH12 groomed reads

Term - Cell Type
CH12
The 'Term' should be the shortest recognizable identifier for the cell/tissue type. Please select from the controlled vocabulary listed here:
http://encodewiki.ucsc.edu/EncodeDCC/index.php/Mouse_cell_types (Required)

Description
B-cell lymphoma (GM12878 analog)
Description of the cell type. Please select from the controlled vocabulary listed here:
http://encodewiki.ucsc.edu/EncodeDCC/index.php/Mouse_cell_types (Required)

Target
CTCF
What was the target of the ChIP? Please select from the controlled vocabulary listed here:
http://encodewiki.ucsc.edu/EncodeDCC/index.php/Antibodies (Required)

Lab
Hardison
What is your primary investigators last Name? (Required)

Sample generated by
Cheryl Keller
Who prepared the library? (Optional)

Antibody Name
CTCF
What is the name of the Antibody used in this ChIP? (Optional)

Antibody Manufacturer
Millipore
Who produced the antibody used in this ChIP? (Optional)

Antibody Catalog Number
Making Galaxy your own
Building local Galaxy instances

- Galaxy is designed for local installation and customization
  - Just download and run, completely self-contained
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters
Scale up on your cluster

- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any scheduler that supports DRMAA (most of them)
- It's easy
- But, requires an **existing computational resource** on which to be deployed
Cloud computing / Infrastructure virtualization

- Computing using resources acquired on demand
- Virtual infrastructure allows for (potential) economies of scale, and (definite) improvements to management automation
- Cloud-style deployment provides a solution both for users without dedicated compute resources, and for simplifying deployment and management
Using Amazon EC2: Startup in 3 steps
Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud 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 add and remove additional services as well as 'worker' nodes on which jobs are run.

Status
- Cluster name:
- Disk status:
- Worker status:
- Service status:

Initial Cluster Configuration

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):
  - 100 GB OK
- Show more startup options

Start Cluster
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud 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 add and remove additional services as well as 'worker' nodes on which jobs are run.

### Status

- **Cluster name:** ttt
- **Disk status:** 0 / 0 (0%)
- **Worker status:** Idle: 0 Available: 0 Requested: 0
- **Service status:** Applications: Data

---

[Cluster status log]
Galaxy Cloudman Console

Welcome to Galaxy Cloudman. This application will allow you to manage this cloud 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 add and remove additional services as well as 'worker' nodes on which jobs are run.

Cluster name: ttt
Disk status: 0 / 0 (0%)
Worker status: Idle: 0 / Available: 0
Service status: Application

Status

Add nodes

Number of nodes to start:
4
OK
Type of Node(s):
Same as Master

Start Additional Nodes
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud 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 adding and removing additional services as well as 'worker' nodes on which jobs are run.

**Status**

- **Cluster name:** ttt
- **Disk status:** 50M / 100G (1%)
- **Worker status:** Idle: 0 Available: 0 Requested: 4
- **Service status:** Applications: 4 Data: 4

**Cluster status log**
Welcome to Galaxy on the Cloud

Your history is empty. Click 'Get Data' on the left pane to start.
Can use like any other Galaxy instance, with additional compute nodes acquired and released (automatically) in response to usage.
Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of “worker” nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

- Terminate Galaxy
- Access Galaxy

Scale

+ Add more instances  - Remove idle instances

Status

Cluster name:   James-galaxy-cluster-9May2010-1
Cluster status: Ready
Disk status:    48G / 100G (48%)
Instance status: Idle: 0 Available: 4 Requested: 12

Cluster status log

- Filesystems
- Database
- Scheduler
- Galaxy
Galaxy Cloud Console

The Galaxy cloud console allows you to manage this instance of Galaxy. From here you can start the main Galaxy interface (including an initial set of "worker" nodes on which jobs will be run), as well as add and remove workers while the main interface is running.

- Terminate Galaxy

Scale

- Add more instances
- Remove idle instances

Status

- Cluster name: james-galaxy-cluster-9May2010-1
- Cluster status: Ready
- Disk status: 59G / 100G (59%)
- Instance status: Idle: 6 Available: 12 Requested: 12

Cluster status log
Persistence

- Once analysis is complete, can scale down worker nodes or shutdown the entire analysis interface
- Data, configuration, *et cetera* is stored, and you can start the cluster back up to continue analysis at any time
- Pay for just what you need
Publishing analysis
Sharing and publishing

- All analysis components (datasets, histories, workflows) can be *shared* among Galaxy users and *published*.
- Pages and annotation allow analysis to be augmented with textual content and provided in the form of an integrated document.
Sharing and publishing

All analysis components—datasets, histories, workflows—can be shared among Galaxy users and published. Pages and annotation allow analysis to be augmented with textual content and provided in the form of an integrated document.
<table>
<thead>
<tr>
<th>Title</th>
<th>Annotation</th>
<th>Owner</th>
<th>Community Rating</th>
<th>Community Tags</th>
<th>Last Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChrY 1000 Genomes</td>
<td>A demo workshop project during CSHL course on Computational Genomics Nov 2010</td>
<td>ericy</td>
<td>★★★★★</td>
<td></td>
<td>2 days ago</td>
</tr>
<tr>
<td>Galaxy Exercises</td>
<td>Various exercises for learning about Galaxy</td>
<td>james</td>
<td>★★★★★</td>
<td></td>
<td>5 days ago</td>
</tr>
<tr>
<td>Galaxy 101: The first thing you need to try</td>
<td>An elementary guide to Galaxy</td>
<td>aun1</td>
<td>★★★★★</td>
<td></td>
<td>Nov 03, 2010</td>
</tr>
<tr>
<td>Windshield Splatter</td>
<td>Live supplement for Genome Research windshield splatter paper.</td>
<td>aun1</td>
<td>★★★★</td>
<td>paper</td>
<td>Oct 27, 2010</td>
</tr>
<tr>
<td>Galaxy RNA-seq Analysis Exercise</td>
<td>An exercise that illustrates how to use Galaxy for RNA-seq analyses.</td>
<td>jeremy</td>
<td>★★★★</td>
<td>heteroplasmy</td>
<td>Oct 27, 2010</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>bwa</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>resequencing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>illumina</td>
<td></td>
</tr>
<tr>
<td>heteroplasmy</td>
<td>aun1</td>
<td>★★★★</td>
<td></td>
<td></td>
<td>Oct 26, 2010</td>
</tr>
</tbody>
</table>
Dynamics of mitochondrial heteroplasmy in three families: A fully reproducible re-sequencing study

Hiroki Goto1, Benjamin Dickins2, Enis Afgan3,5, Ian M. Paul4, James Taylor3,5, Kateryna D. Makova1, and Anton Nekrutenko2,5

Correspondence should be addressed to KDM, IT, or AN.

1. How to use this document

This document is a live copy of supplementary materials for the manuscript. It provides access to all the data as well as to exact analyses and workflows discussed in the paper, so you can play with them by re-running, changing parameters, or even applying them to your own sequencing data. To import workflows you must create a Galaxy account (unless you already have one) – a hassle-free procedure where you are only asked for a username and password. To make this even easier, we created several screencasts (very short movies) to help you:

- access our datasets
- re-use workflows listed on this page
- view and import histories listed on this page

In addition, we created two longer screencasts:

- Watch the analysis of one family (F7) from start (Illumina reads) to finish (a list of variable position):
- Watch how the complete analysis can be performed on the Amazon Cloud.

If you experience any problems while using this page, please e-mail our bug report list and we will get back to you.

2. Accessing the Data

All datasets discussed in the paper can be found in two places:

- A Galaxy Library called mtProject:
- An S3 bucket on the Amazon Cloud
M10, M10C2, M15, and M15C2; the workflow 'mt analysis 0.01 strand-specific (fastq single)' was run four times on datasets that lacked PCR replicates: M9 and M4C3; for this we created three separate histories: one for each family. Each history (F4 = Family 4, F7 = Family 7, F11 = Family 11) can be examined in detail and imported below (see a Screencast explaining how to do this):

Each of the histories contain original Illumina datasets and outputs of workflows.

3.3 Generating initial summary datasets

In the previous step we identified variable sites in all samples. Now we need to merge the results by generating reports for each family. To do this we first copied results workflow executions into a new history called "F4–F7–F11 final report" (for explanation on how to copy datasets between histories see this Screencast):

Within this history individual datasets are merged into summaries generated for each family. To be more specific, datasets 1 through 10 were merged into dataset 19 called "F4 summary", datasets 11 – 14 were joined into history item 22 called "F7 summary", and, finally, datasets 15 – 18 were used to generate #24 called "F11 summary". Merging of datasets was performed with "Join, Subtract, and Group -> Column Join" tool. Let's look at datasets "F7 summary" to understand what this means:

Results of heteroplasmy workflow for all individuals of family 7 joined together. You can click in "reri" button above to see the parameters.
the workflow 'mt analysis 0.01 strand-specific (fastq single)' was run four times on datasets that lacked PCR replicates: M9 and M4C3; for this we created three separate histories: one for each family. Each history (F4 = Family 4, F7 = Family 7, F11 = Family 11) can be examined in detail and imported below (see a Screencast explaining how to do this).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1: bM4C1-1</td>
<td>Child M4C1 blood PCR 1</td>
</tr>
<tr>
<td>2: bM4C1-2</td>
<td>Child M4C1 blood PCR 2</td>
</tr>
<tr>
<td>3: cM4C1-1</td>
<td>Child M4C1 cheek PCR 1</td>
</tr>
<tr>
<td>4: cM4C1-2</td>
<td>Child M4C1 cheek PCR 2</td>
</tr>
<tr>
<td>5: bM4C3</td>
<td>Child M4C3 blood PCR (no replicated were performed for this individual)</td>
</tr>
<tr>
<td>6: cM4C3</td>
<td>Child M4C3 cheek PCR (no replicated were performed for this individual)</td>
</tr>
<tr>
<td>7: bMSG-1</td>
<td>Grandmother MSG blood PCR 1</td>
</tr>
</tbody>
</table>
Discovery of human heteroplasmic mtDNA in accessible interface to data

Enis Afgan, Hiroki Goto, Ian Paul, Francesca Choo

Datasets

We analyzed the mitochondrial genome from three cheek swab specimens and from blood at Penn State and H11571; L10796 and H3370. These primers induced errors, each amplification was performed in a child pair + two mtDNA amplification for each cheek.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>File Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1-m-c-1.fastq</td>
<td></td>
</tr>
<tr>
<td>p1-m-c-2.fastq</td>
<td></td>
</tr>
<tr>
<td>p1-m-b-1.fastq</td>
<td></td>
</tr>
</tbody>
</table>

Reads were mapped against the hg19 version of the human genome using bwa. Only those reads aligning exactly once to the mitochondrial genome and having no hits to the nuclear genome were retained. This procedure eliminated potential contamination of our data with reads associated with unmapped reads (our PCR strategy enriched mtDNA but did not eliminate nuclear DNA from the sample: approximately 10-20% of the reads mapped to the nuclear genome and were subsequently eliminated from the analysis).
We analyzed the mitochondrial genome from three mother/child pairs. For each mother and child pair the DNA was collected from cheek swab specimen and from blood at Penn State Medical School. mtDNA was amplified with PCR using two primer sets L2815 and H11571; L10796 and H3370. These primers are originally described in Tanaka et al. (1996). To control for possible PCR-induced errors, each amplification was performed twice. In total we generated 24 Illumina datasets (eight for each mother and child pair - two mtDNA amplification for each cheek swab and blood samples.

Reads were mapped against hg19 version of the human genome using bwa. Only those reads aligning exactly once to the mitochondrial genome and having no hits to the nuclear genome were retained. This procedure eliminated potential contamination of our data with reads associated with nmt (our PCR strategy enriched mt DNA but did not eliminate nuclear DNA from the sample, approximately 10–20% of the reads map to the nuclear genome and were subsequently eliminated from the analysis). Using PCRs replicates for each sample, the following workflow estimates the methodological error rate by comparing mapping results between two amplifications. To do so we identified all sites where in one replicate where there were no deviant reads (all reads contained the same nucleotide; i.e. 1000 'A' bases) but the other contained such sites (e.g., 1000 As and 12 Cs). Dividing the number of deviant reads (12 in this case) by the total read coverage (1012) at such positions gave us error rate of 1.18% (12/1012) at this position.
We analyzed the mitochondrial genome from three mother/child pairs: L17157; L10796 and H3370. These primers are one of the
child pair - two mtDNA amplification for each cheek swab specimen and from blood at Penn State Med.
This reads associated with the mitochondrial genome and having no hits to the
nucleotide (or other) of our data with reads associated with mtDNA (four PCR
sample: approximately 10–20% of the reads mapped to
Using PCR replicates for each sample, the following workflow results between two amplifications. To do so we identify
reads contained the same nucleotide: i.e. 1000 'A' bases
number of deviant reads (12 in this case) by the total bases
(12/1012) at this position.

Step 16: Filter
Filter: Output dataset 'out_file1' from step 14
With following condition
\[ c_{10} = 'chrM' \] and \[ 10 \geq 200 \]

Step 17: Join
Join: Output dataset 'out_file1' from step 15
with
Output dataset 'out_file1' from step 16

for all positions that have sufficient quality to consider in both replicates

Histories resulting from first workflow on each pair: History 'mt replicates pair 1', History 'mt replicates pair 2', History 'mt replicates pair 3'.
The power of Galaxy publishing and sharing

- Galaxy's publishing features facilitate access and reproducibility without any extra leg work.
- One click grants access to the *actual analysis* you performed to generate your original results.
- Not just data access: the full pipeline.
- Annotate each step.
- Anyone can import your work and immediately reproduce or build on it.
Winshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond1,2,6,9, Samir Wadhawan3,6,7, Francesca Chiaromonte4, Guruprasad Ananda1,3, Wen-Yu Chung1,3,8, James Taylor1,5,9, Anton Nekrutenko1,3,9 and The Galaxy Team1

Abstract

How many species inhabit our immediate surroundings? A straightforward collection technique suitable for answering this question is known to anyone who has ever driven a car at highway speeds. The windshields of a moving vehicle are subjected to numerous insect strikes and can be used as a collection device for representative sampling. Unfortunately the analysis of biological material collected in that manner, as with most metagenomic studies, proves to be rather demanding due to the large number of required tools and considerable computational infrastructure. In this study, we use organic matter collected by a

Footnotes

[Supplemental material is available online at http://www.genome.org. All data and tools described in this manuscript can be downloaded or used directly at http://galaxypjeekt.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]

Article published online in advance of print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094508.109.
Try it now: http://usegalaxy.org

Develop and deploy: http://getgalaxy.org