Enabling reproducible research with Galaxy

Galaxy Team + Anton Nekrutenko

Discovery of human heteroplasmic sites enabled by an accessible interface to cloud-computing infrastructure

Galaxy Team + Anton Nekrutenko

Galaxy Cloud

Enis Afgan

Dannon Baker

Galaxy Pages

Jeremy Goecks

Heteropasmy

Hiroki Goto

Ian Paul

Benjamin Dickins

Galaxy Team

Guru Ananda

Dan Blankenberg

Nate Coraor

Ramkrishna Chakrabarty

Greg von Kuster

Kanwei Li

Kelly Vincent

Custodians

Kateryna Makova

Taylor/Nekrutenko

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

I want you to remember
exactly four (4) things:

http://usegalaxy.org

http://usegalaxy.org/**galaxy101**

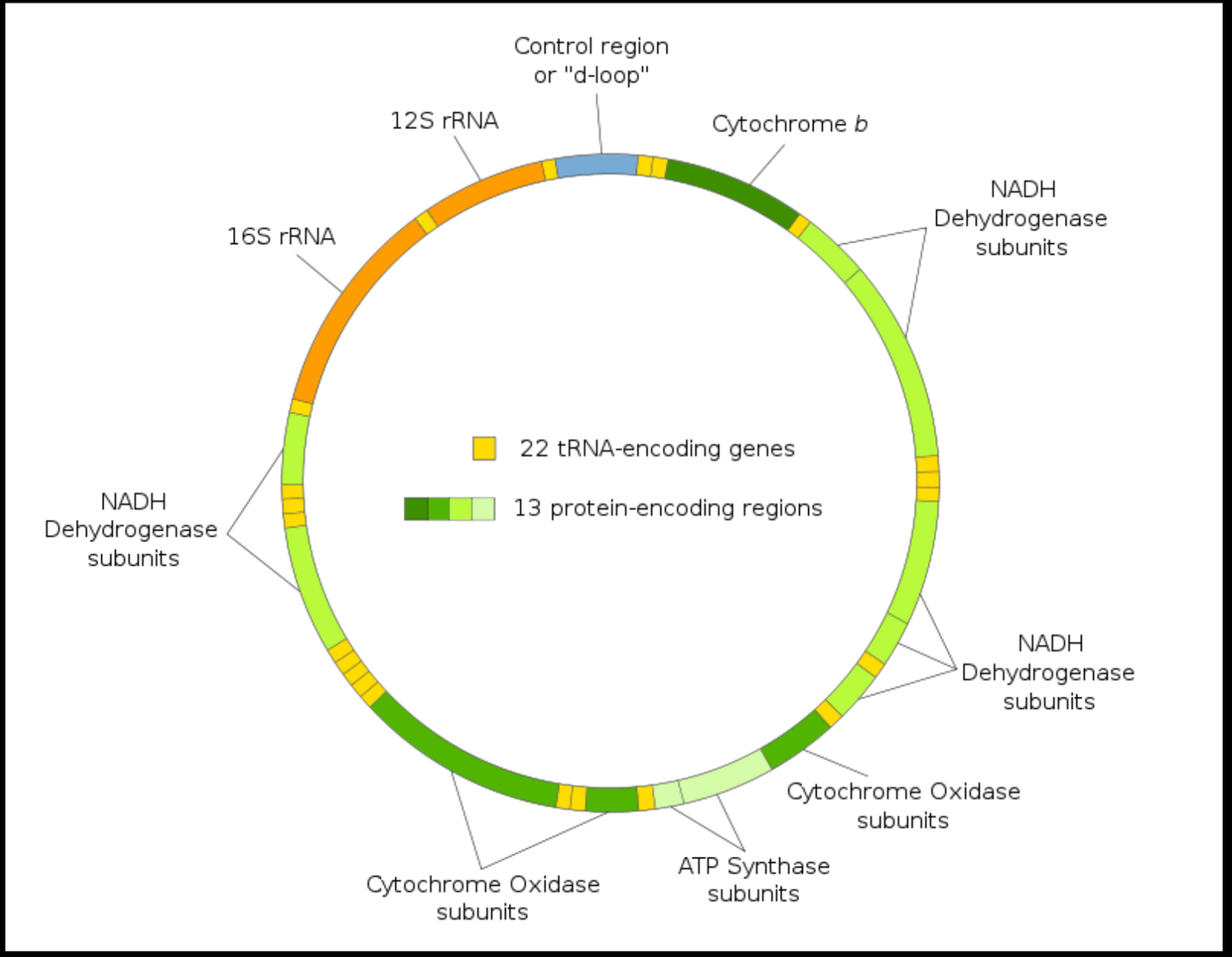
http://usegalaxy.org/**cloud**

http://getgalaxy.org

Setup: I will have three windows open

- This presentation
- A brower pointing to http://usegalaxy.org
- A browser pointing to a cloud instance

The human mitochondrial genome



Mitochondrial heteroplasmy

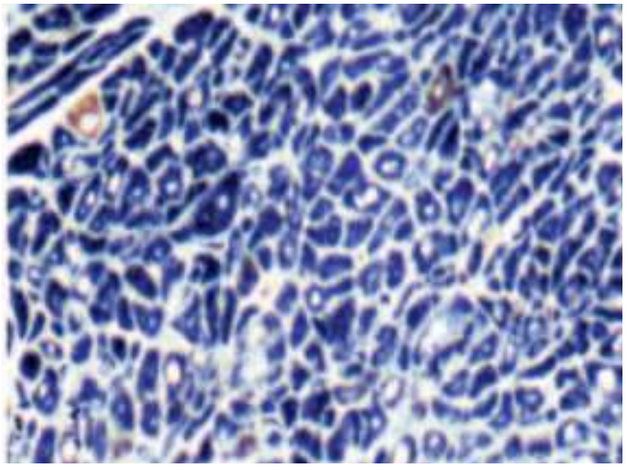
- Typical human cells have ~100s of mitochondria, each with ~10s of copies of the mitochondrial genome
- Heteroplasmy refers to variation among the mitochondrial genomes within a cell or individual

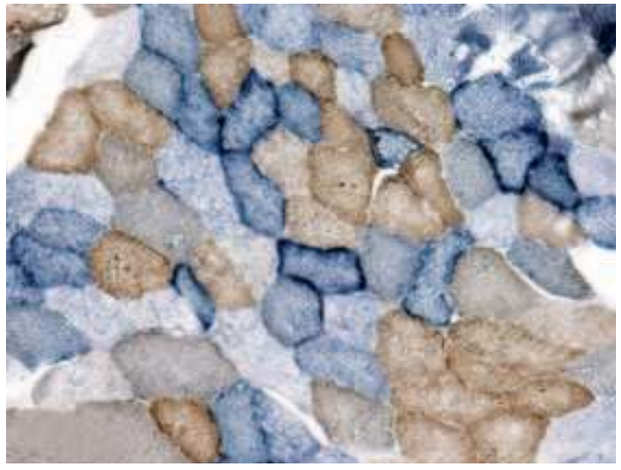
Heteroplasmy and disease

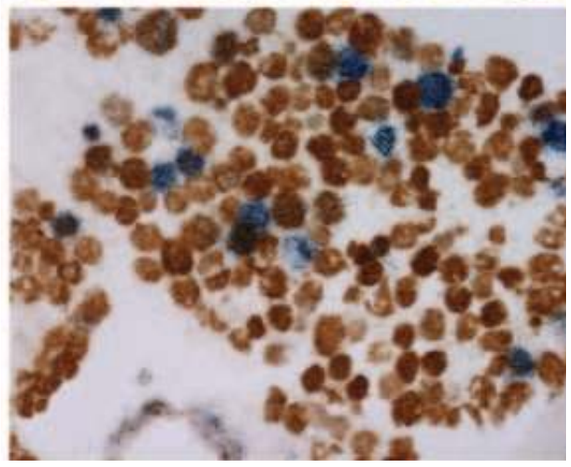
- Mitochondrial mutations are implicated in hundreds of diseases, including common metabolic and neurological disorders
- Many disease causing variants are heteroplasmic, and clinical manifestation depends on levels
  - “Threshold effect”

Brown = Cytochrome c oxidase positive, Blue = COX negative

Homoplasmic tRNA mutation (hypertrophic cardiomyopathy)	Heteroplasmic tRNA mutation (skeletal muscle)	Expanding somatic mutation (extraocular muscle)

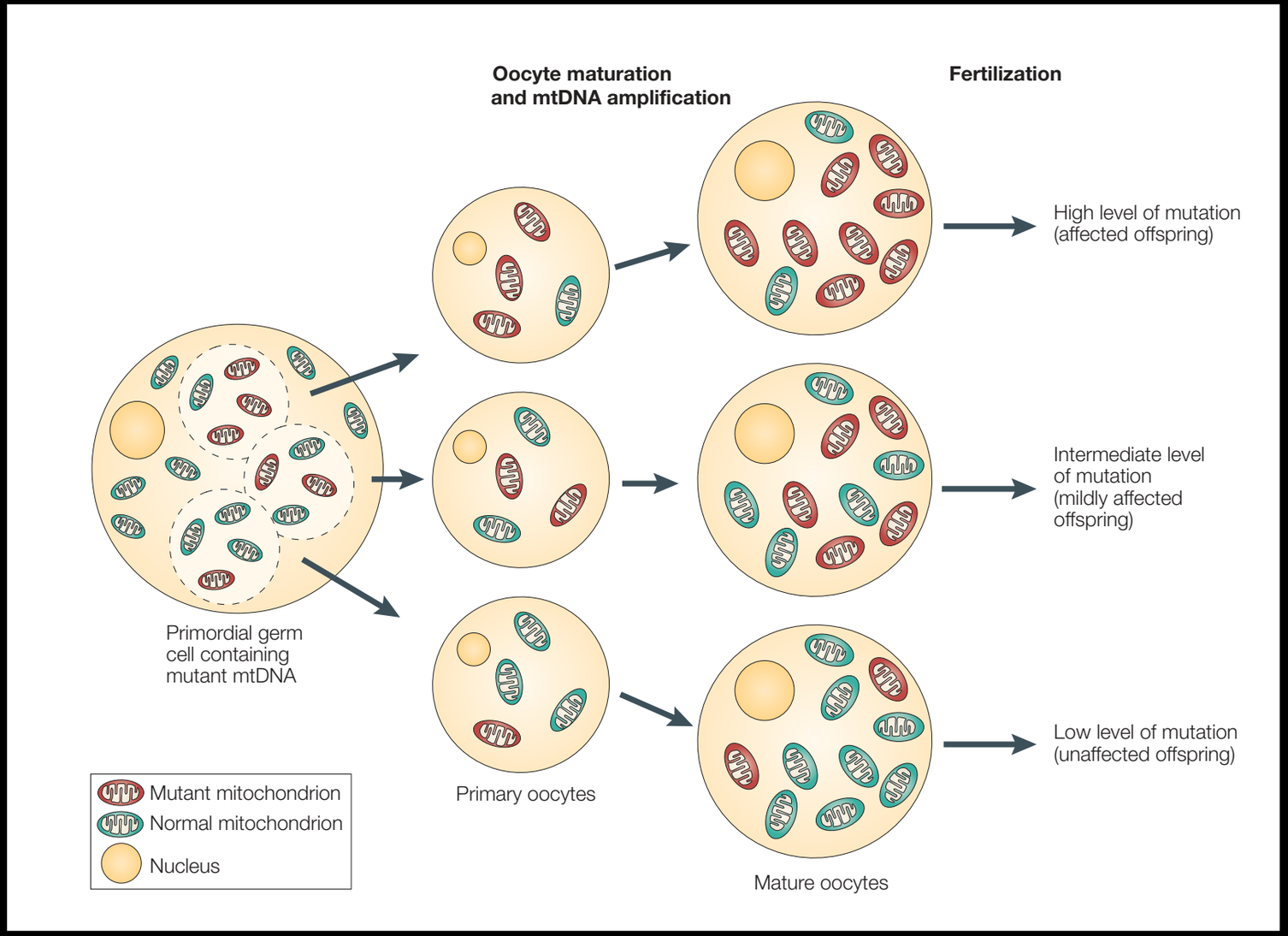






(Taylor RW and Turnbull, Nature Reviews genetics, 2005)

mtDNA bottleneck during oogenesis



(Taylor RW and Turnbull, Nature Reviews genetics, 2005)

Mitochondrial genome

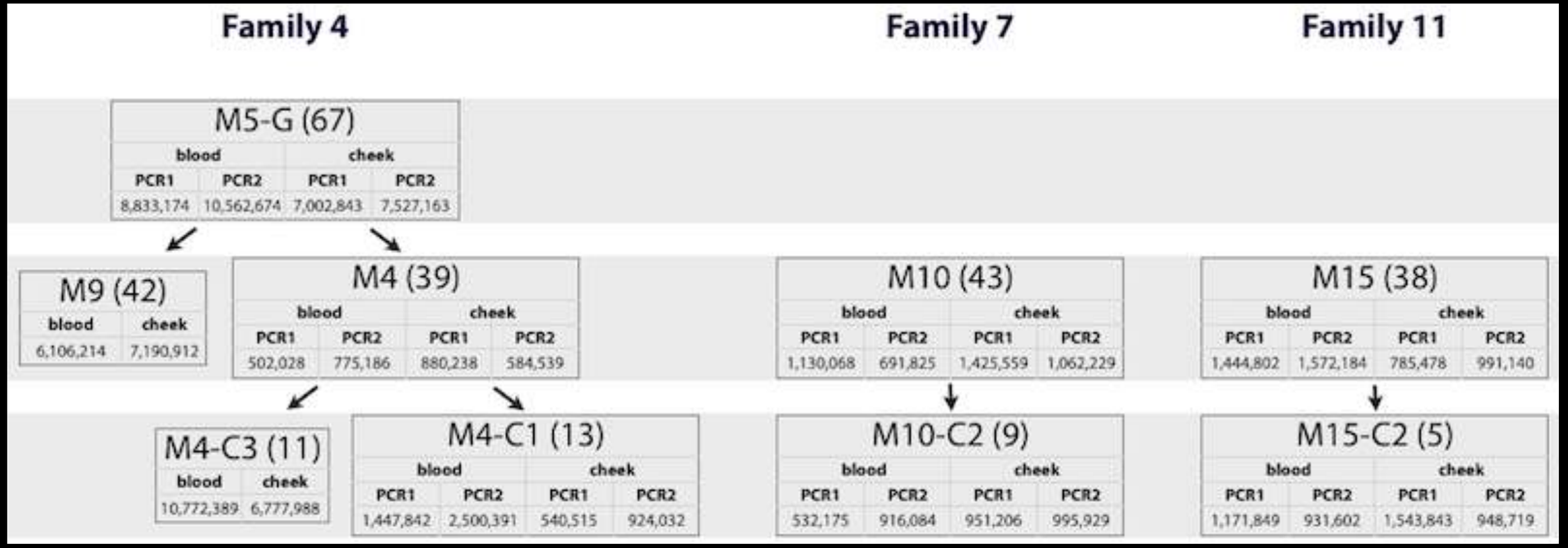
- Multiple genome copies per mitochondria, multiple mitochondria per cell
- Variation exists between genome copies in a cell or individual (heteroplasmy)
- Levels of heteroplasmy can change by sampling in cell division
- Bottleneck in oogenesis can greatly increase these changes in the germline

Measuring frequencies and transmissions of heteroplasmy with deep resequencing

Goals of this study

- Be able to *reliably* identify heteroplasmic sites that occur at very low frequencies (<< 10%)
- Be able to distinguish somatic mutations from germline mutations

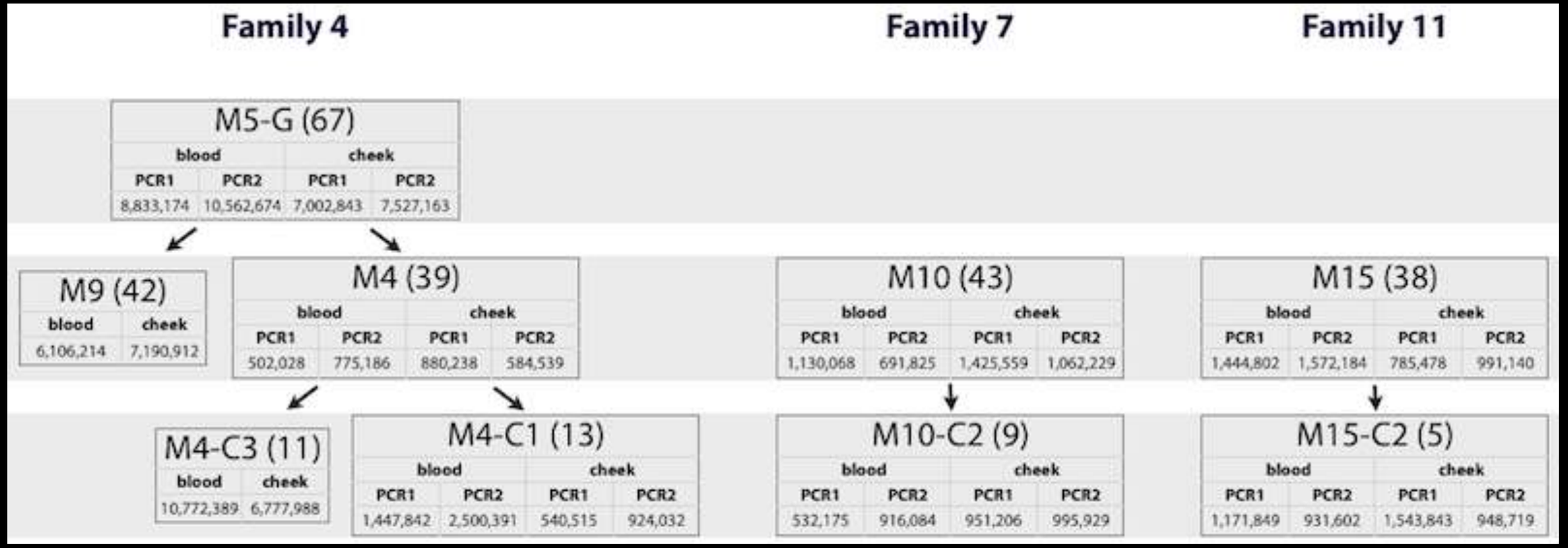
Pilot study: Nine individuals from three families



Strategy

- For each individual, samples taken from two tissues (cheek and blood)
- mtDNA selectively PCR amplified: two independent amplifications for each individual / tissue pair
- High-throughput sequencing using the Illumina platform

Pilot study: Nine individuals from three families



28 Illumina datasets, some multiplexed, 50 or 76bp reads, Performed by Sequensys

Producing the sequence was easy enough...

Now what?

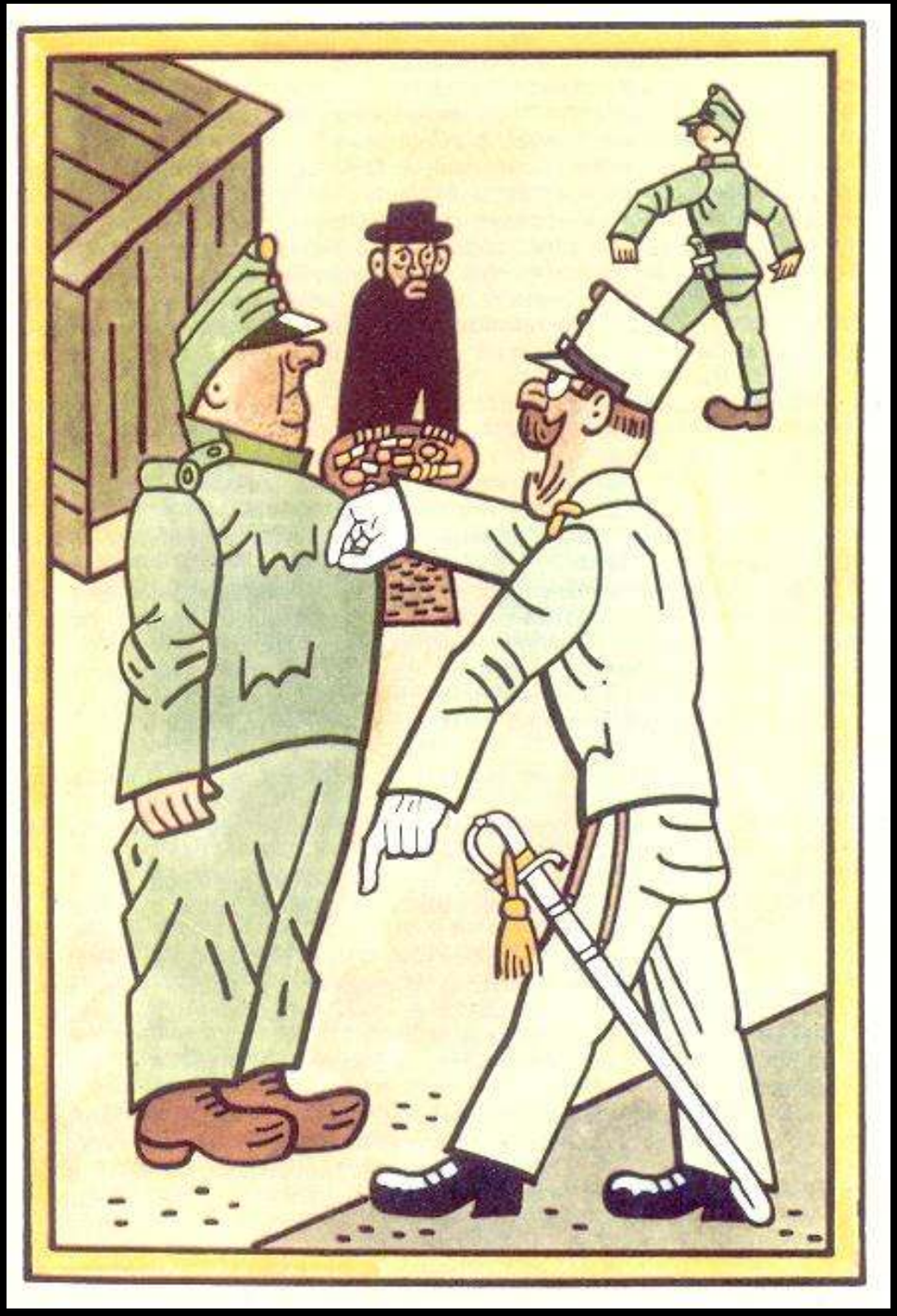
As science becomes increasingly dependent on computation:

- How can methods best be made **accessible** to scientists?
- How to facilitate **transparent** communication of analyses?
- How best to ensure that analyses 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

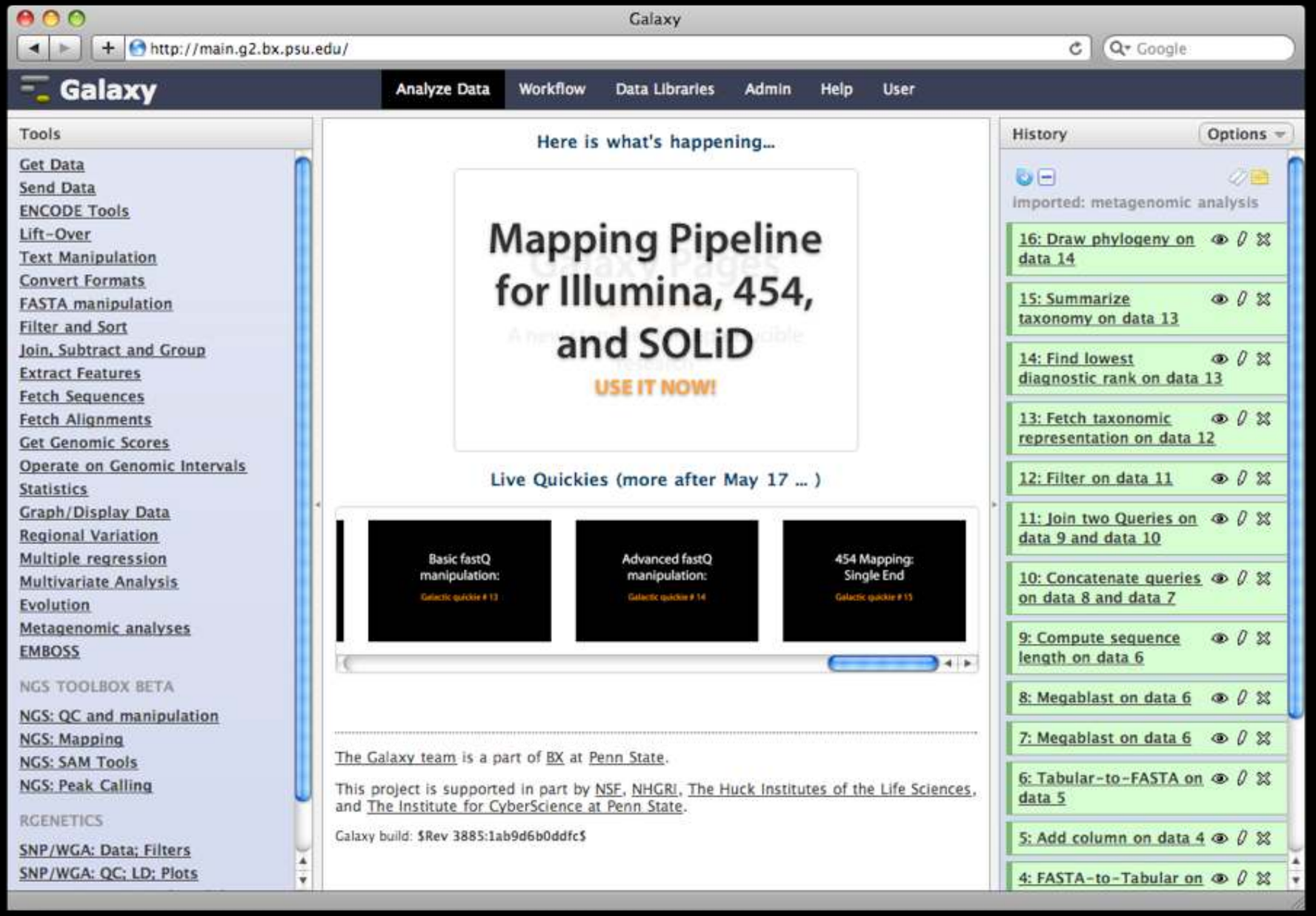
- 18 Nat. Genetics microarray gene expression experiments
- Less than 50% reproducible
- Problems
  - missing data (38%)
  - missing software, hardware details (50%)
  - missing method, processing details (66%)

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

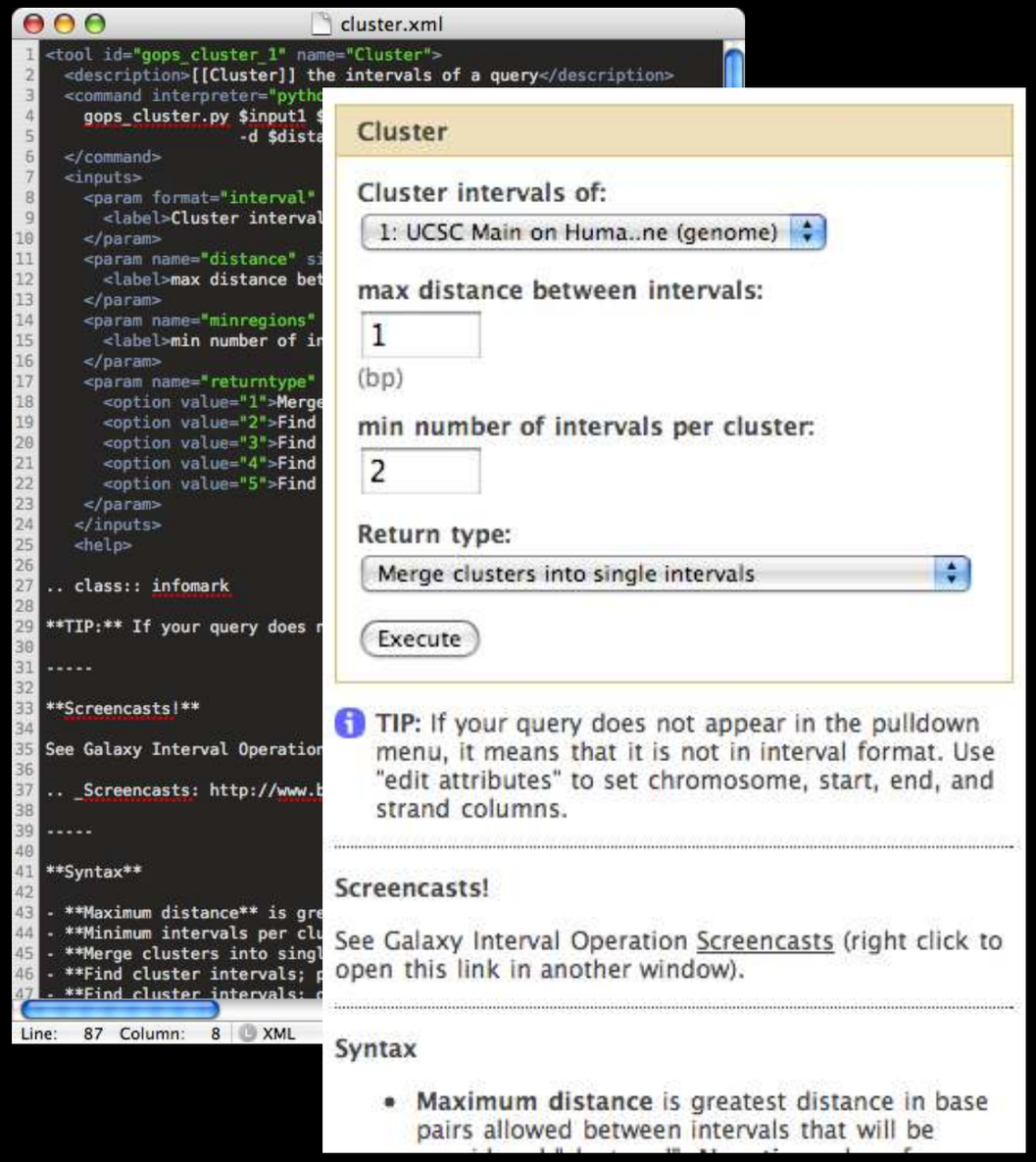
NGS Re-sequencing Experiment Reproducibility

- 14 re-sequencing experiments in Nat. Genetics, Nature, and Science (2010)
- 0% reproducible?
- Problems
  - limited access to primary data (50%)
  - some or all tools unavailable (50%)
  - settings & versions not provided (100%)

Galaxy: accessible analysis system

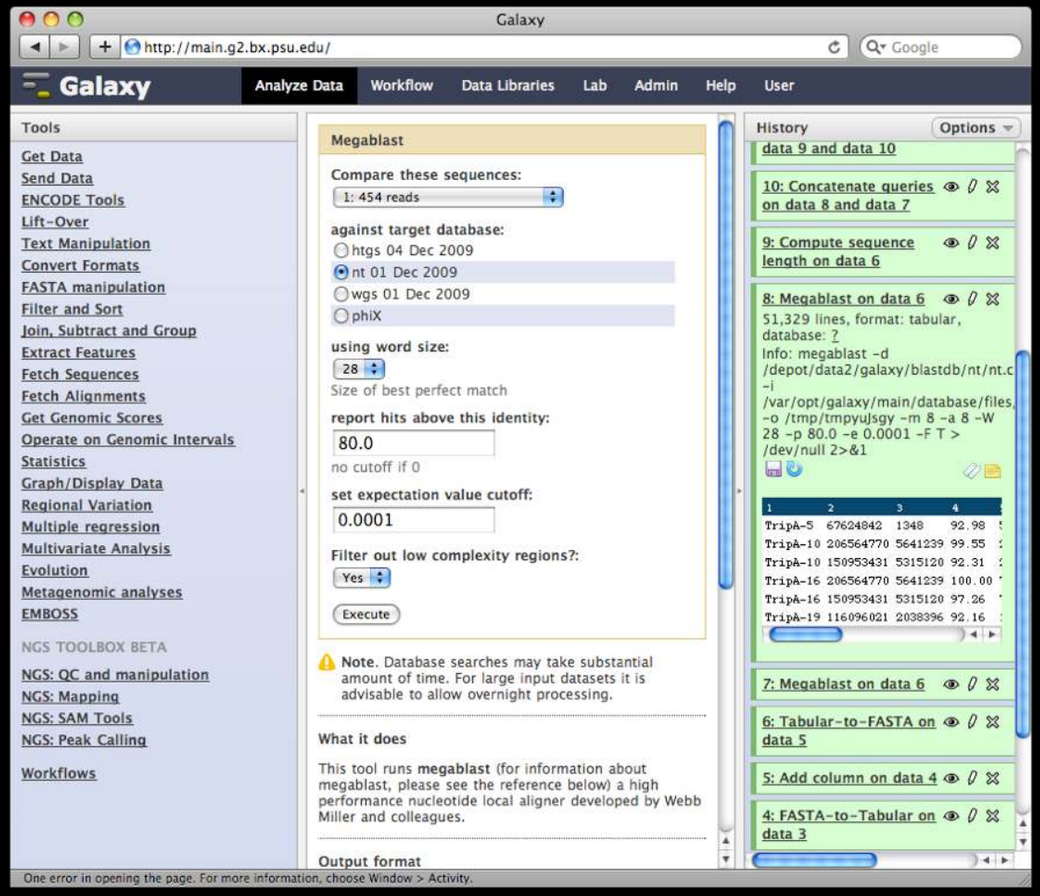


Integrating existing tools into a uniform framework



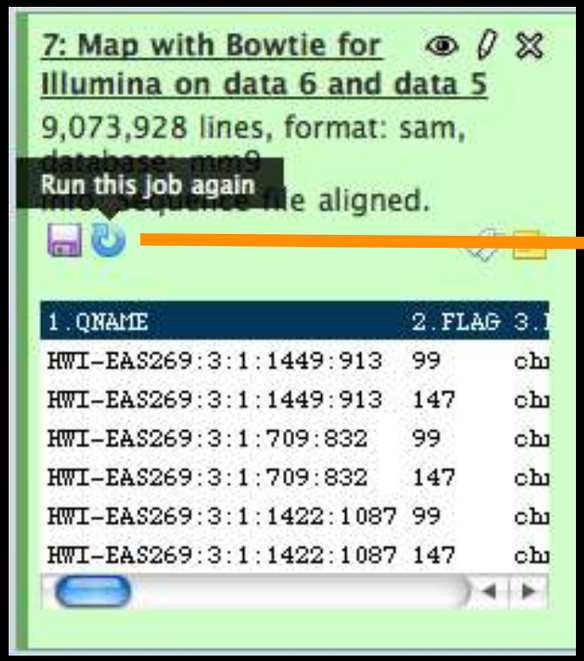
- Defined in terms of an abstract interface (inputs and outputs)
  - In practice, mostly command line tools, a declarative XML description of the interface, how to generate a command line
- Designed to be as easy as possible for tool authors, while still allowing rigorous reasoning

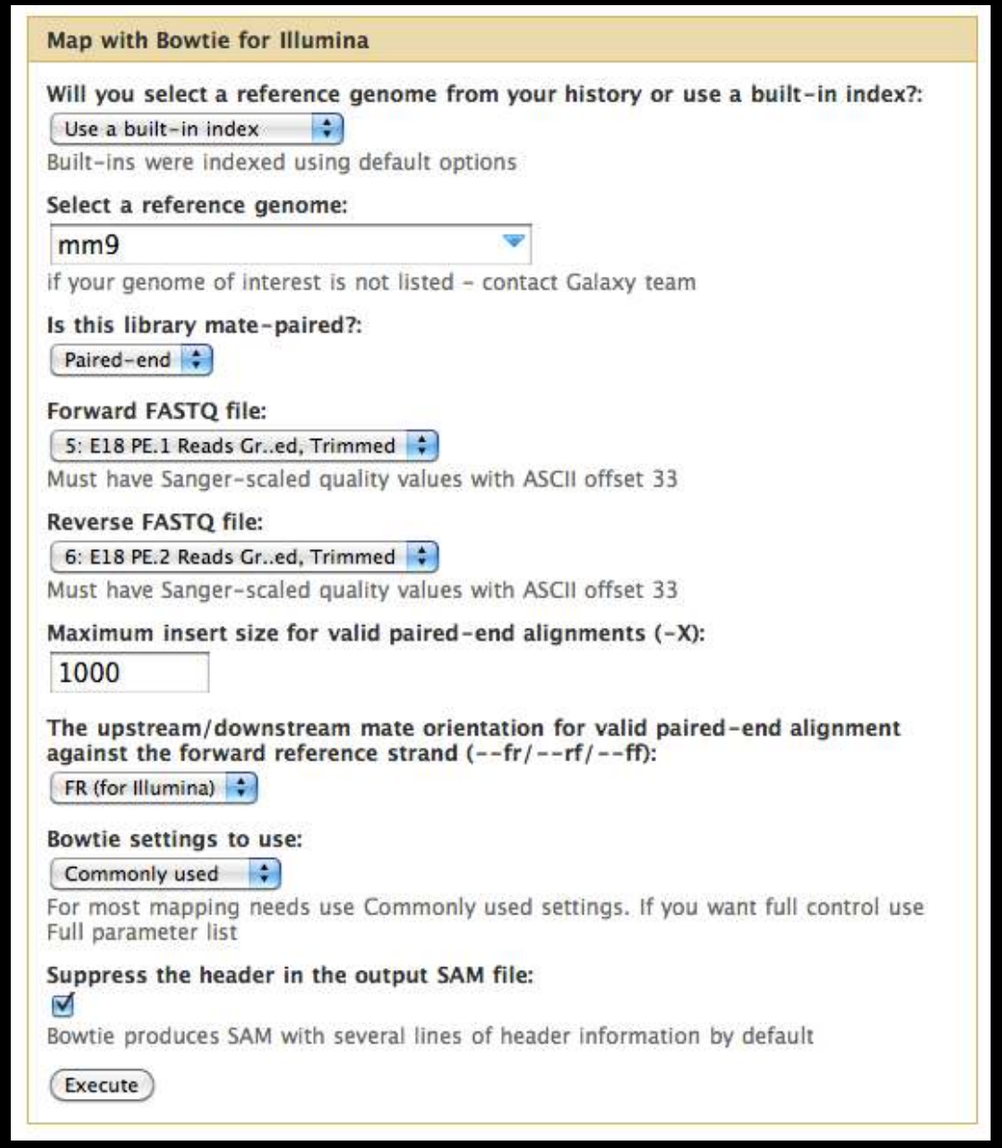
Galaxy analysis interface



- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses
- Exact parameters of a step can always be inspected, and easily rerun

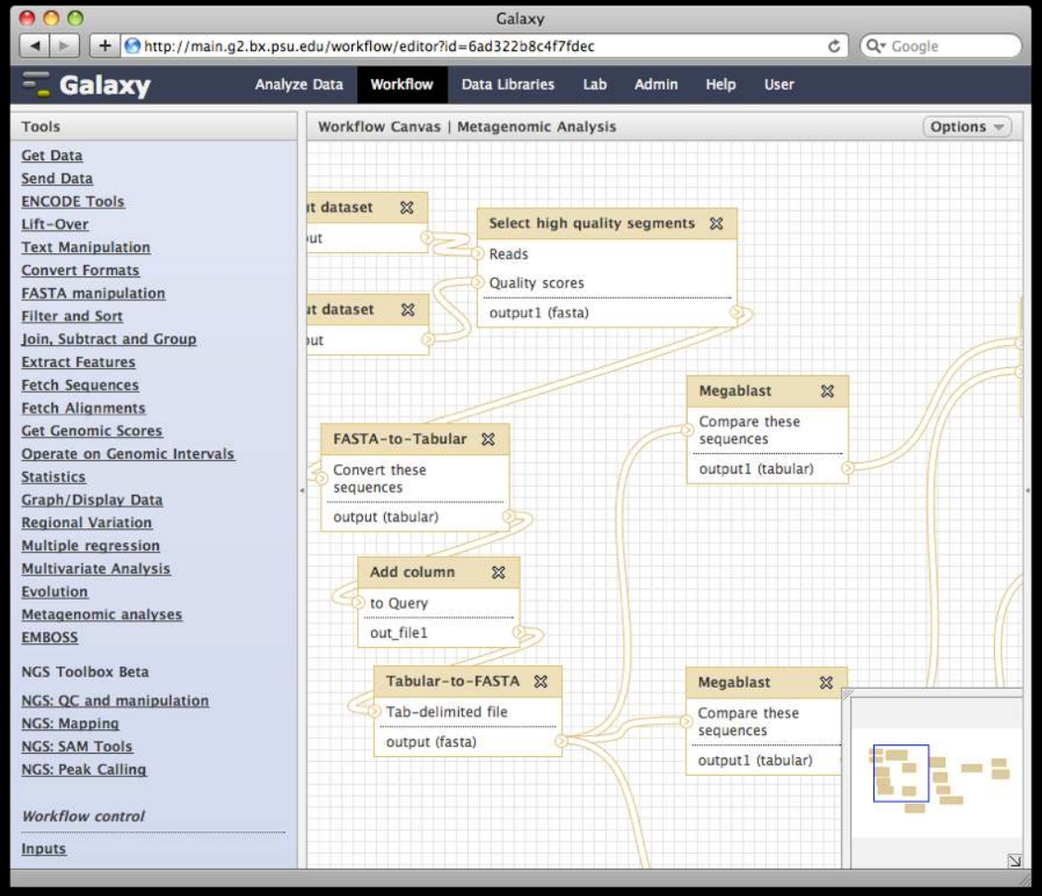
Automatically tracks every step of every analysis





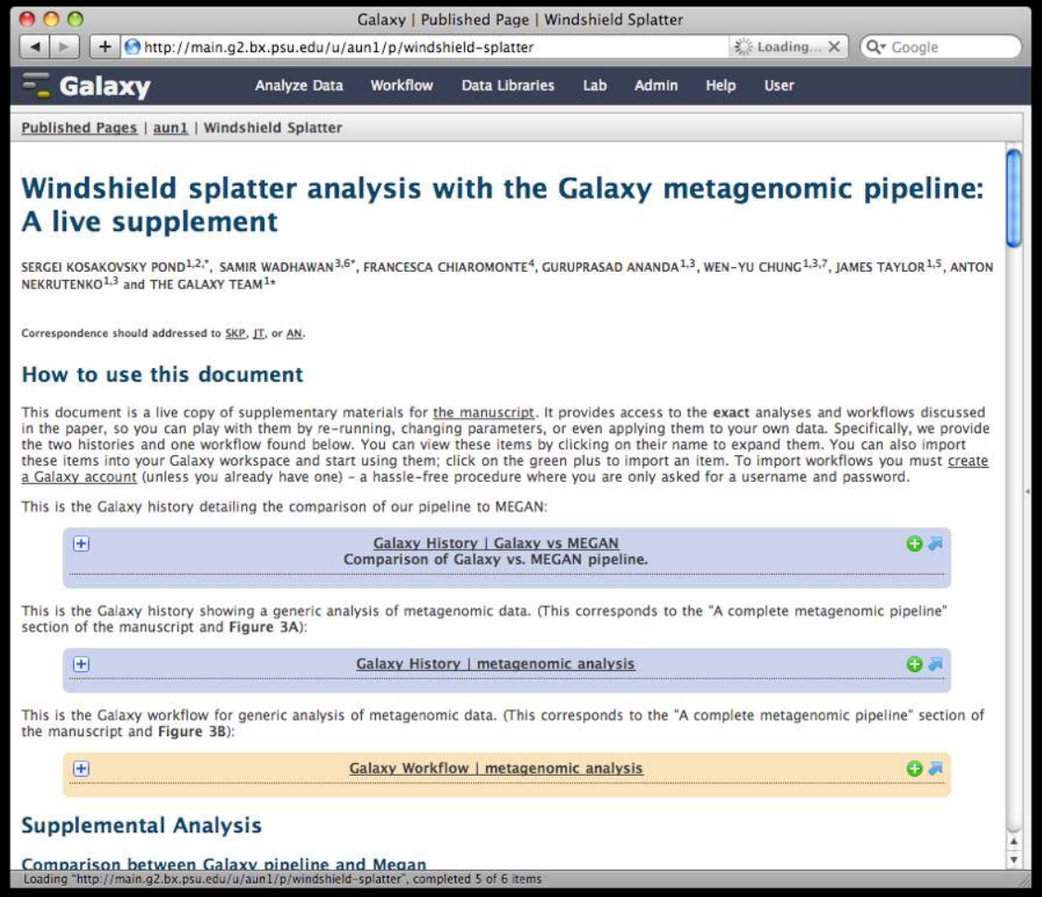
As well as user-generated metadata and annotation...

Galaxy workflow system



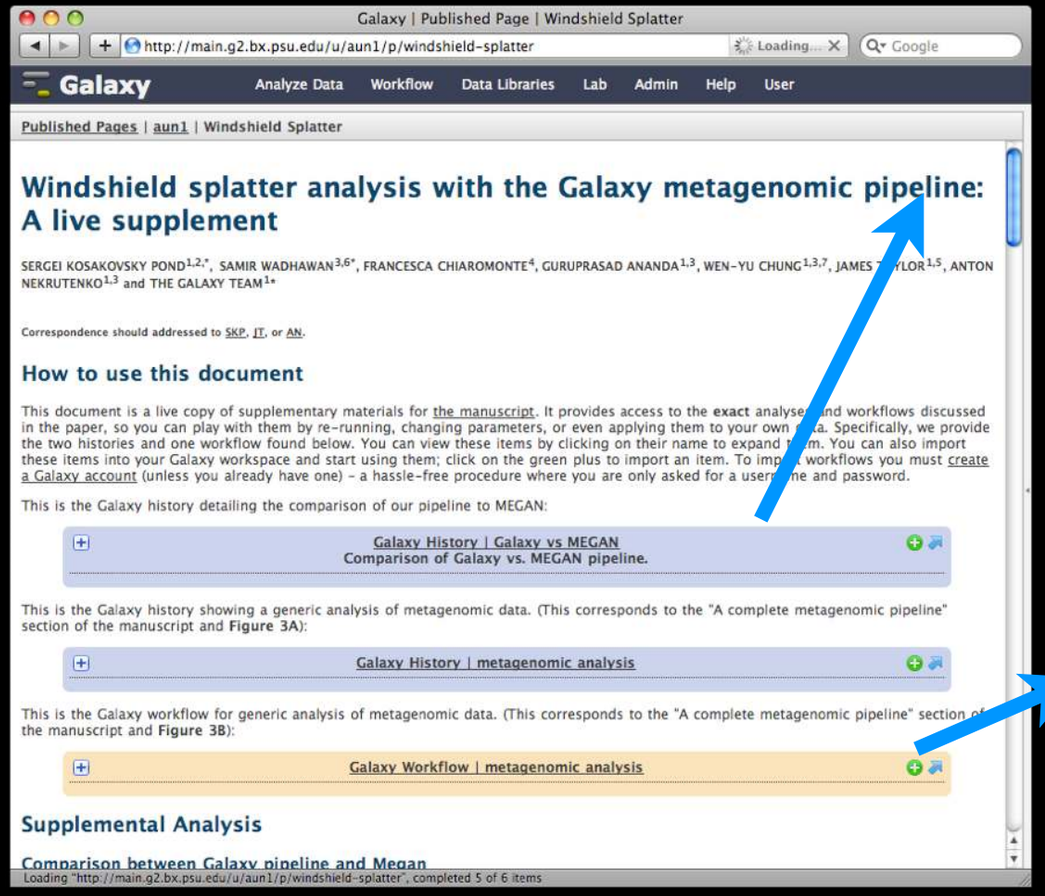
- Workflows can be constructed from scratch *or* extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis

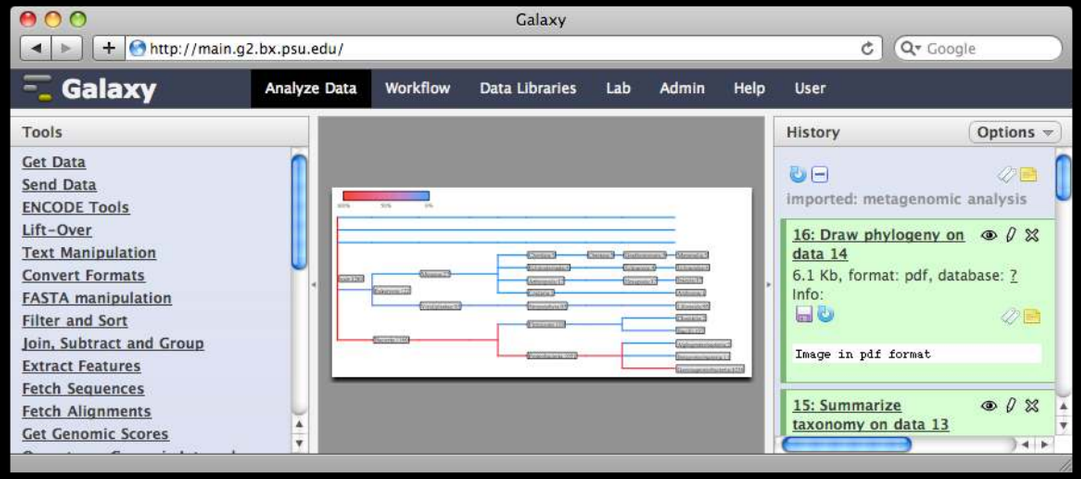
Sharing and publishing

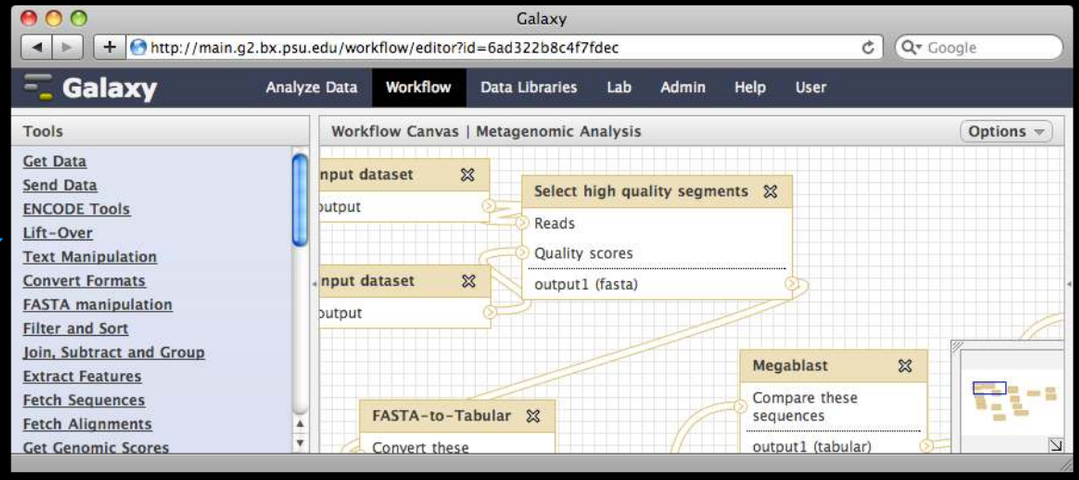


- All analysis components (datasets, histories, workflows) can be *shared* among Galaxy users and *published*
- Pages and annotation allow analysis to be augmented with textual content and provided in the form of an integrated document

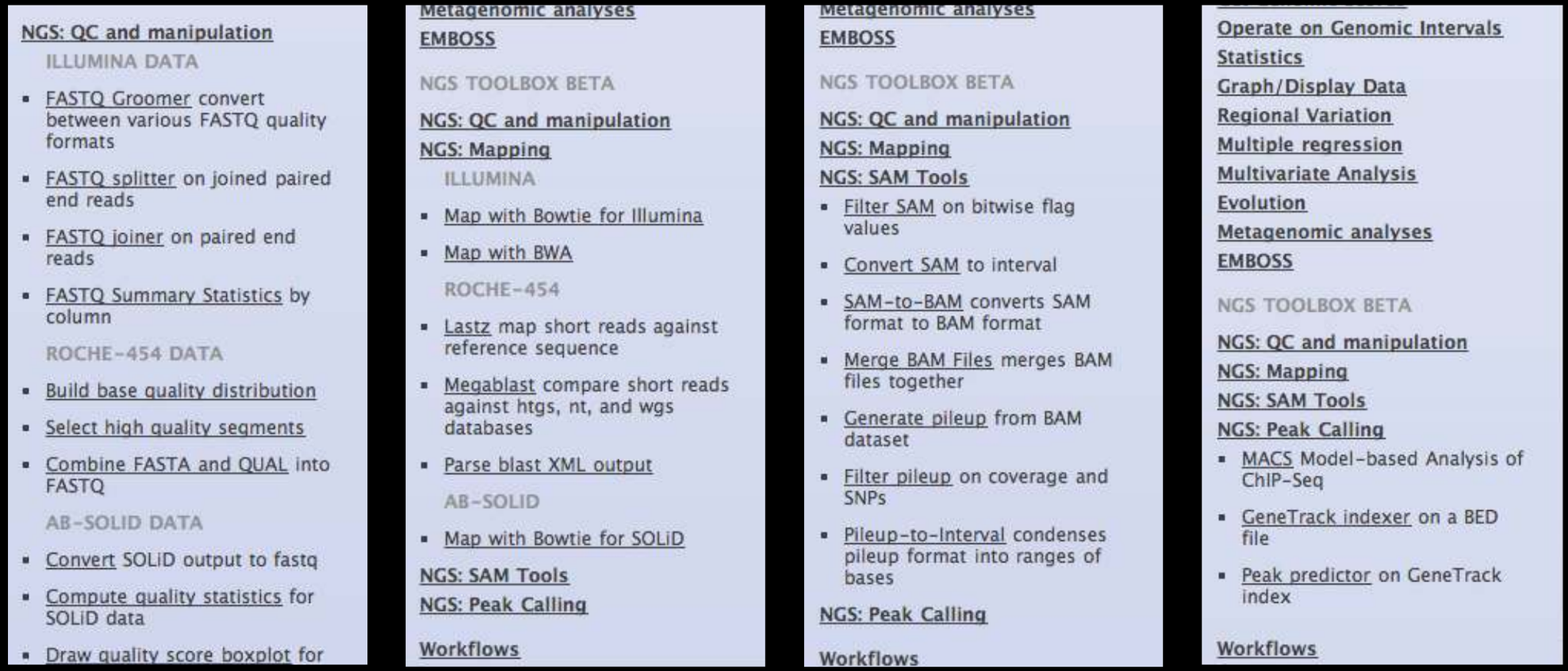
Sharing and publishing







Could use the Galaxy web site for this data...



However, we would be sharing **compute availability** and **bandwidth** with other users, and need to upload our potential **private data**

Could use a local Galaxy instance...

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (unix) system
  - Run jobs on existing compute clusters
- But, requires an **existing computational resource** on which to be deployed

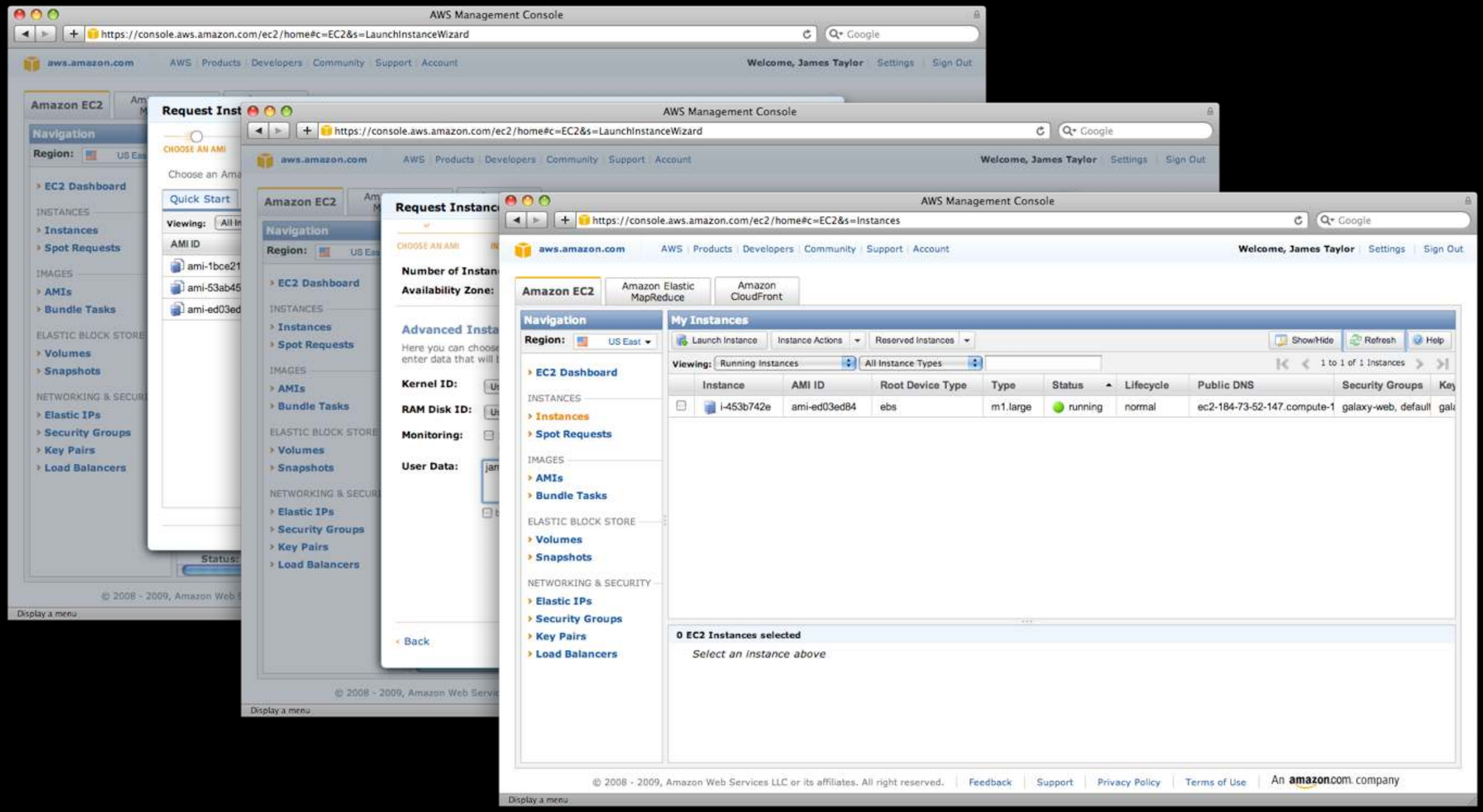
Welcome to **Galaxy** on the Cloud(s)

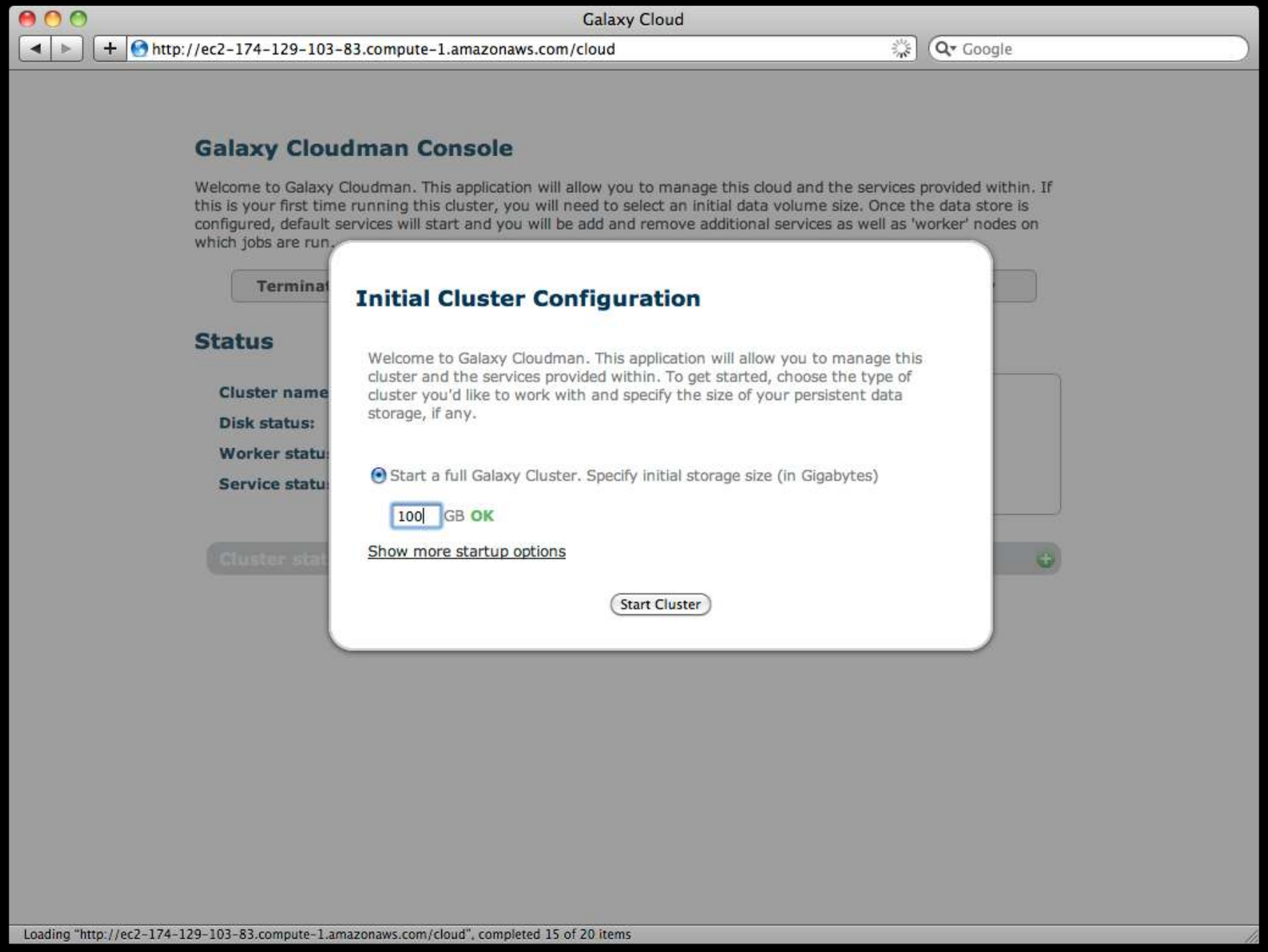
http://usegalaxy.org/cloud

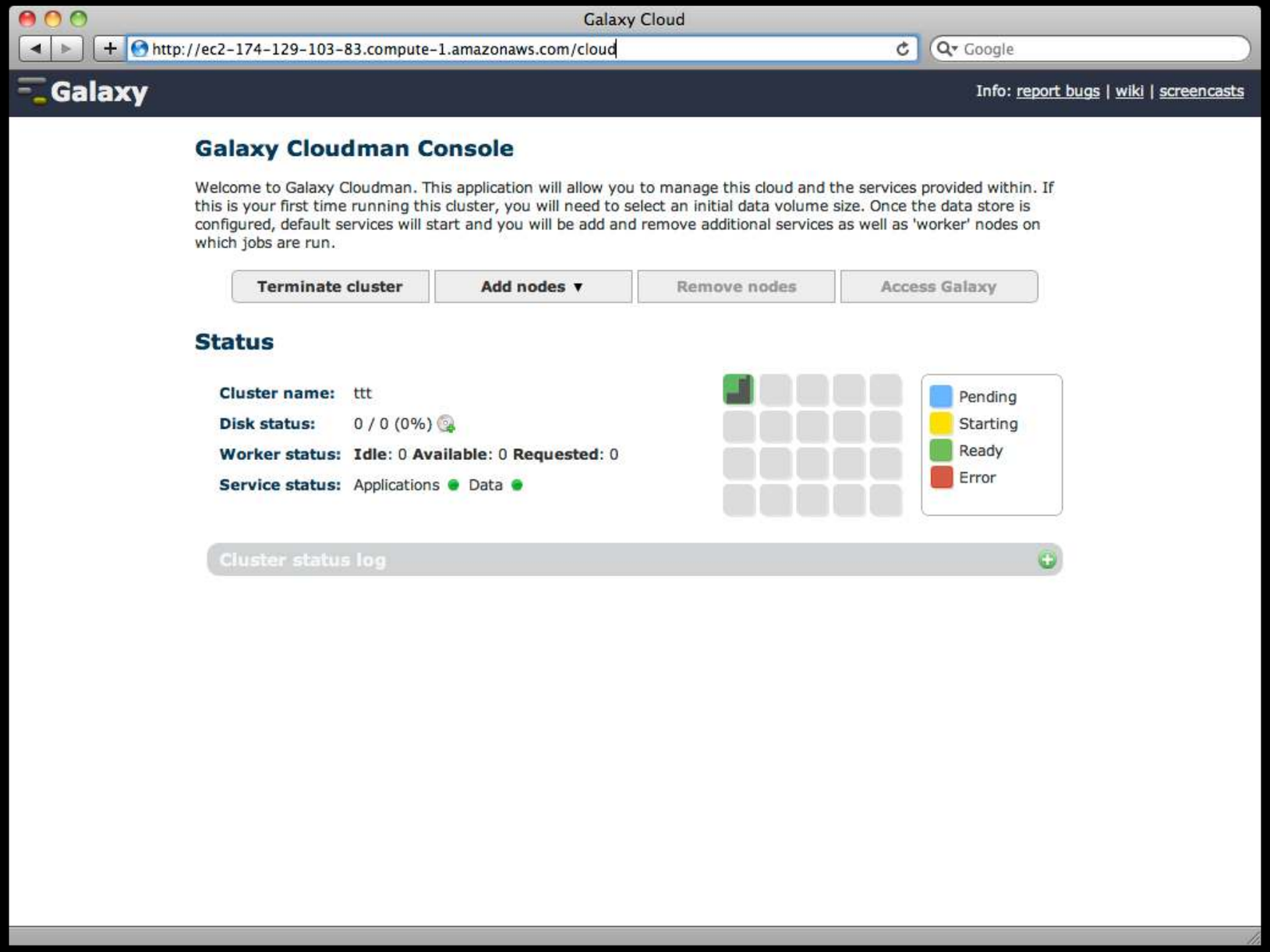
Cloud computing

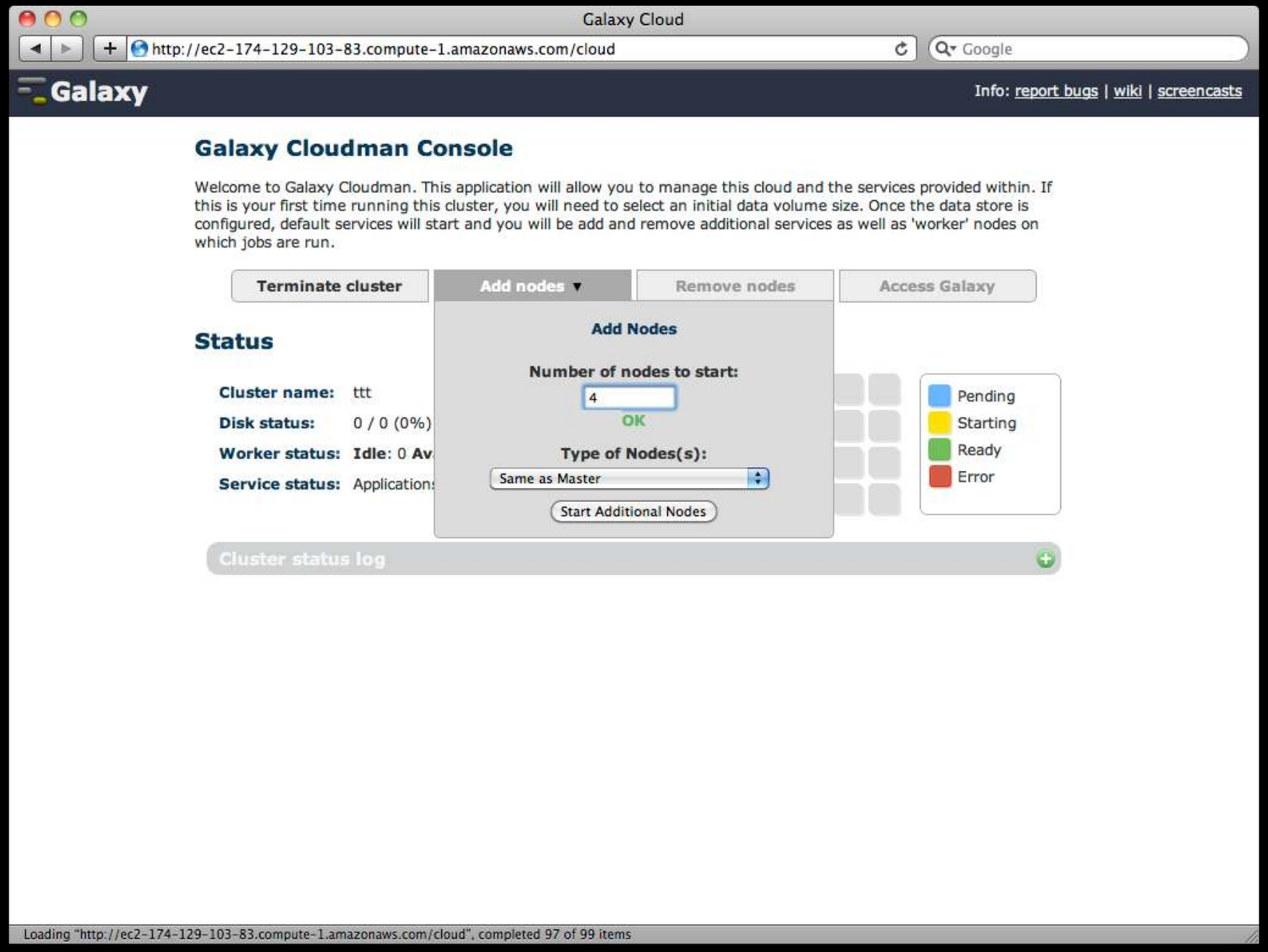
- Computing using resources acquired on demand
- Spectrum from infrastructure as a service (virtual machines, e.g. Amazon EC2) to software as a service (e.g. Google docs)
- Goal for Galaxy: deliver the provider independence of an IaaS based solution, while approaching the ease of use of a SaaS based solution

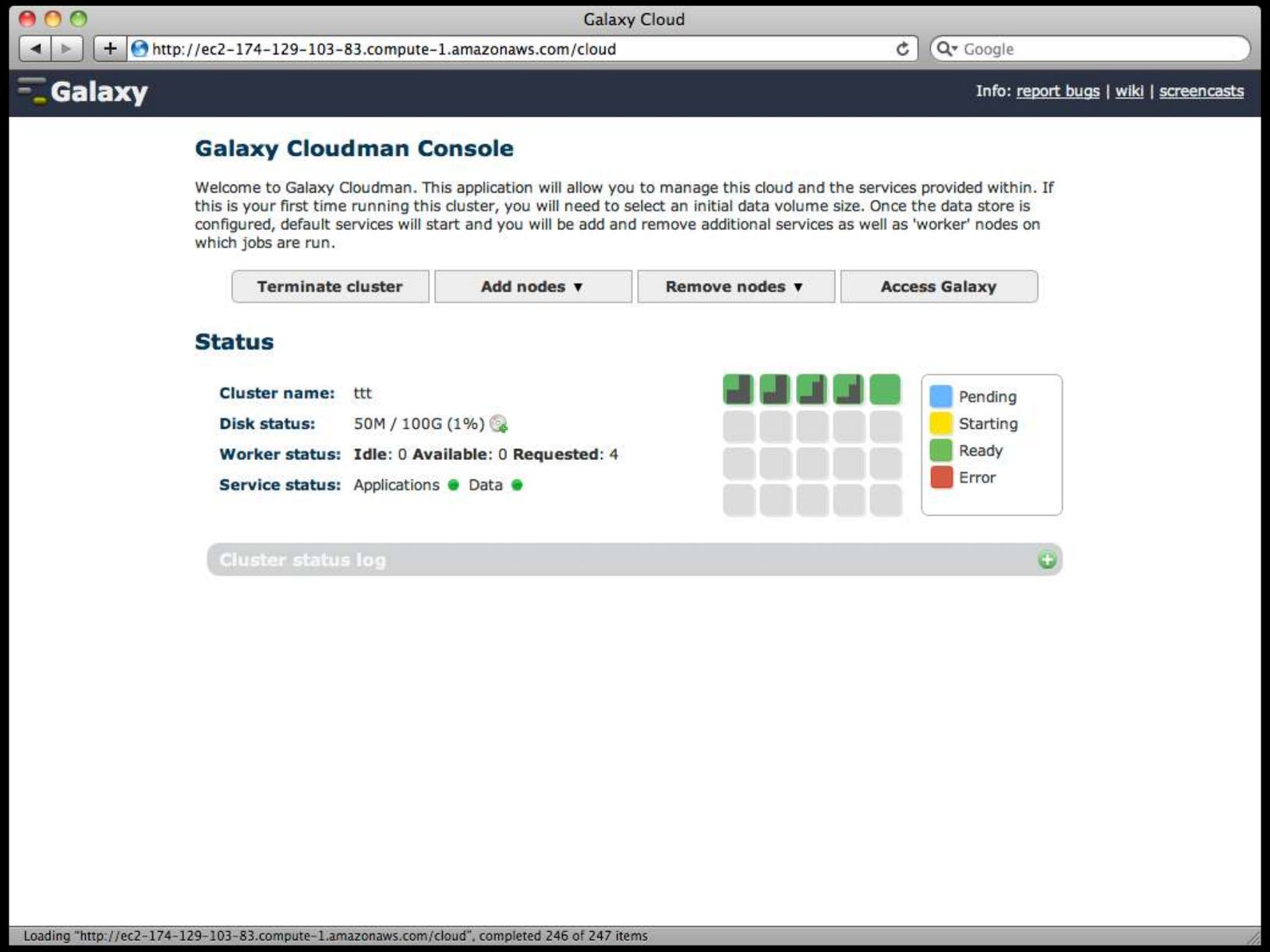
Using Amazon EC2: Startup in 3 steps

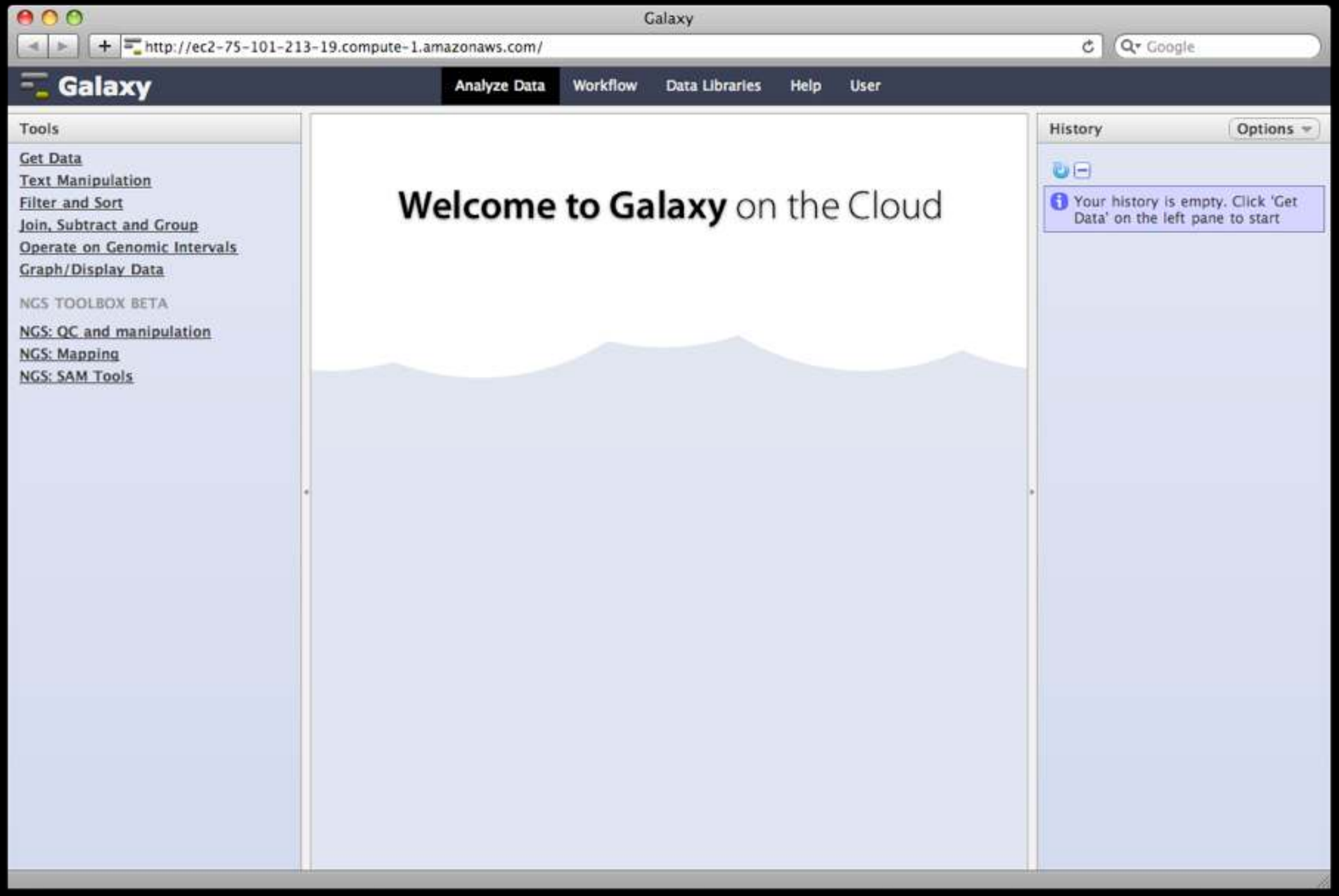



Galaxy Cloud
http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud
Google
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this cluster, you will need to select an initial data 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.
Terminat
Status
Cluster name
Disk status:
Worker statu
Service statu
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
Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 15 of 20 items


Galaxy Cloud
http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud
Google
Galaxy
Info: report bugs | wiki | screencasts
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.
Terminate cluster
Add nodes ▾
Remove nodes
Access Galaxy
Status
Cluster name: ttt
Disk status: 0 / 0 (0%)
Worker status: Idle: 0 Available: 0 Requested: 0
Service status: Applications Data
Pending
Starting
Ready
Error
Cluster status log

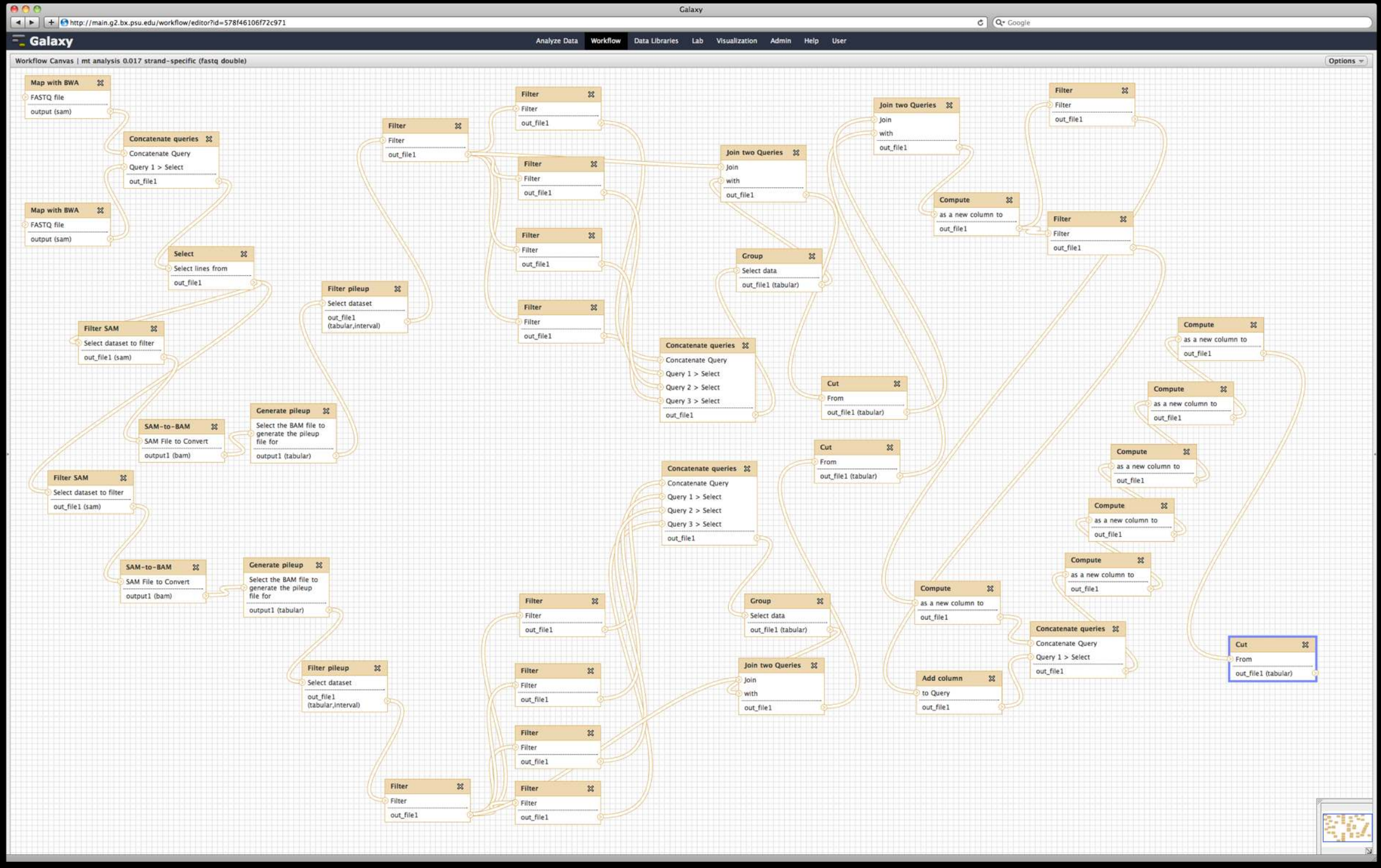

Galaxy Cloud
http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud
Google
Galaxy
Info: report bugs | wiki | screencasts
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this 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.
Terminate cluster
Add nodes ▾
Remove nodes
Access Galaxy
Add Nodes
Number of nodes to start:
4
OK
Type of Nodes(s):
Same as Master
Start Additional Nodes
Status
Cluster name: ttt
Disk status: 0 / 0 (0%)
Worker status: Idle: 0 Av
Service status: Application
Pending
Starting
Ready
Error
Cluster status log
Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 97 of 99 items

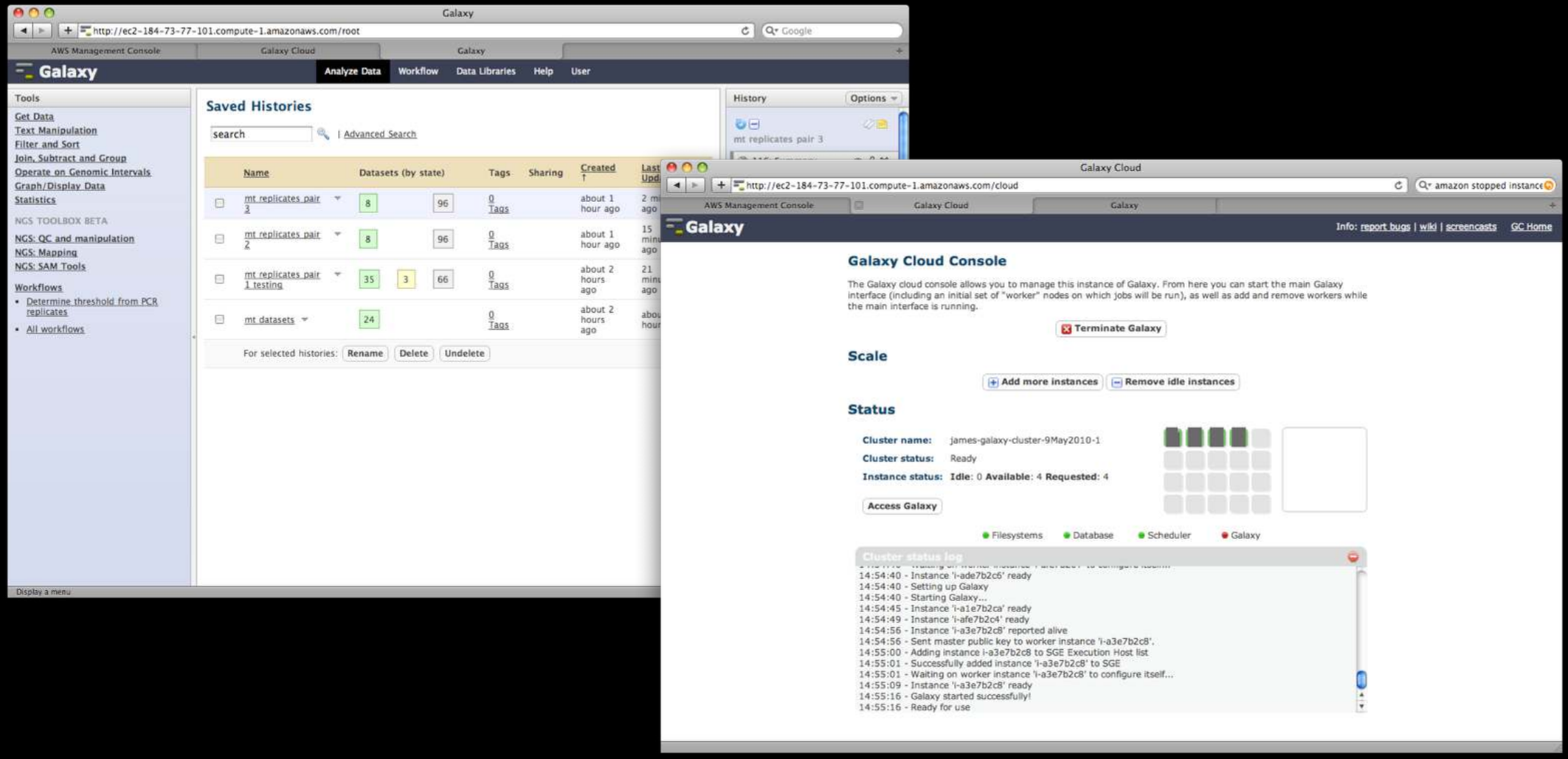

Galaxy Cloud
http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud
Google
Galaxy
Info: report bugs | wiki | screencasts
Galaxy Cloudman Console
Welcome to Galaxy Cloudman. This application will allow you to manage this cloud and the services provided within. If this is your first time running this 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.
Terminate cluster
Add nodes ▾
Remove nodes ▾
Access Galaxy
Status
Cluster name: ttt
Disk status: 50M / 100G (1%)
Worker status: Idle: 0 Available: 0 Requested: 4
Service status: Applications Data
Pending
Starting
Ready
Error
Cluster status log
Loading "http://ec2-174-129-103-83.compute-1.amazonaws.com/cloud", completed 246 of 247 items


Galaxy
http://ec2-75-101-213-19.compute-1.amazonaws.com/
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
Text Manipulation
Filter and Sort
Join, Subtract and Group
Operate on Genomic Intervals
Graph/Display Data
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Welcome to Galaxy on the Cloud
History
Options
Your history is empty. Click 'Get Data' on the left pane to start

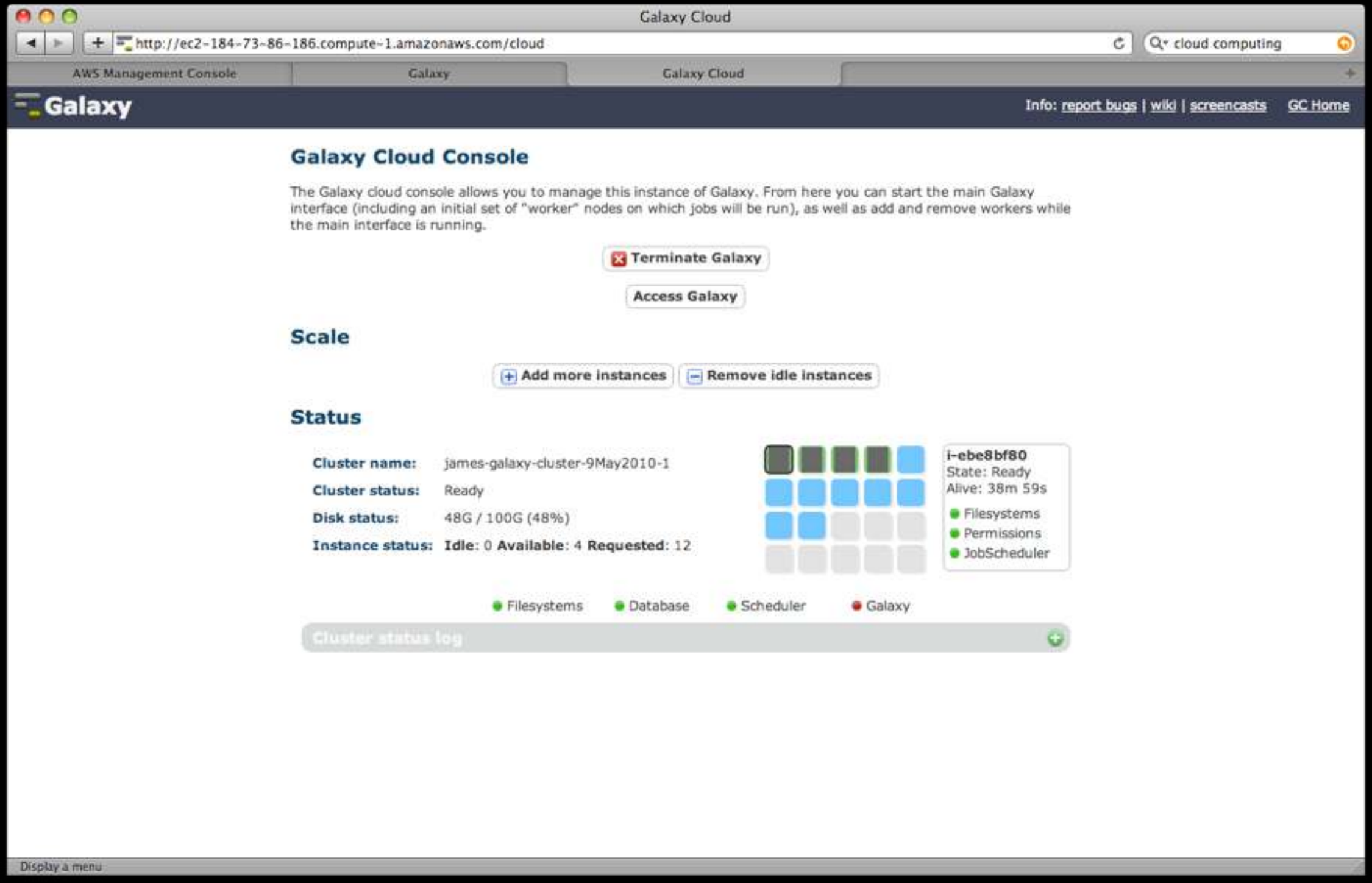
Analysis defined as a Galaxy workflow...

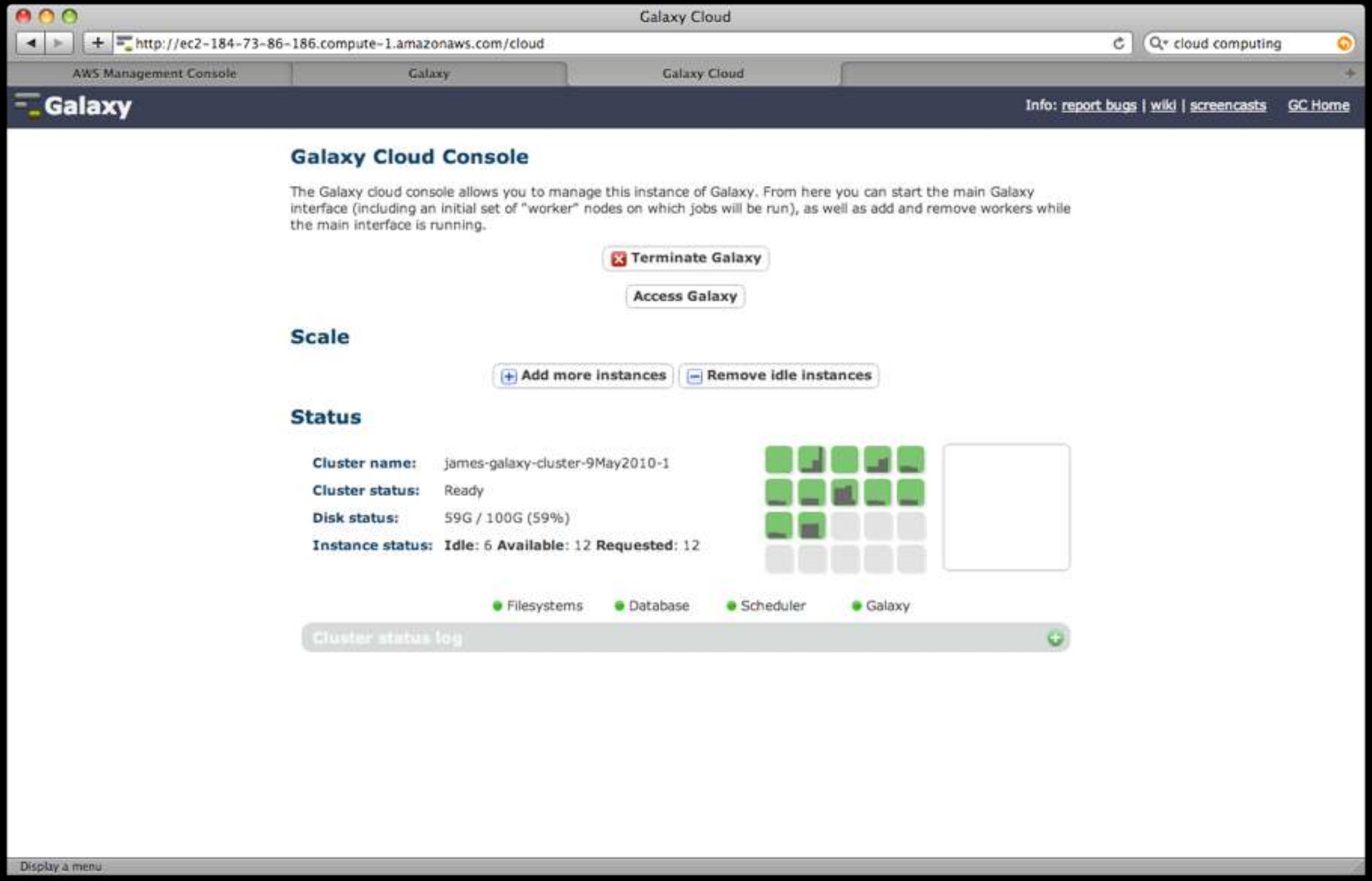
(created by example, and extracted for reuse across individuals)


Galaxy
http://main.g2.bx.psu.edu/workflow/editor?id=578f46106f72c971
Galaxy
Analyze Data
Workflow
Data Libraries
Lab
Visualization
Admin
Help
User
Workflow Canvas | mt analysis 0.017 strand-specific (fastq double)
Options
Map with BWA
FASTQ file
output (sam)
Concatenate queries
Concatenate Query
Query 1 > Select
out_file1
Map with BWA
FASTQ file
output (sam)
Select
Select lines from
out_file1
Filter SAM
Select dataset to filter
out_file1 (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Generate pileup
Select the BAM file to generate the pileup file for
output1 (tabular)
Filter SAM
Select dataset to filter
out_file1 (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Generate pileup
Select the BAM file to generate the pileup file for
output1 (tabular)
Filter pileup
Select dataset
out_file1 (tabular,interval)
Filter pileup
Select dataset
out_file1 (tabular,interval)
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Join two Queries
Join
with
out_file1
Join two Queries
Join
with
out_file1
Group
Select data
out_file1 (tabular)
Group
Select data
out_file1 (tabular)
Concatenate queries
Concatenate Query
Query 1 > Select
Query 2 > Select
Query 3 > Select
out_file1
Concatenate queries
Concatenate Query
Query 1 > Select
Query 2 > Select
Query 3 > Select
out_file1
Cut
From
out_file1 (tabular)
Cut
From
out_file1 (tabular)
Join two Queries
Join
with
out_file1
Compute
as a new column to
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Add column
to Query
out_file1
Concatenate queries
Concatenate Query
Query 1 > Select
out_file1
Cut
From
out_file1 (tabular)



After running first set of workflows, many jobs waiting to run but cluster is completely utilized...


Galaxy Cloud
http://ec2-184-73-86-186.compute-1.amazonaws.com/cloud
cloud computing
AWS Management Console
Galaxy
Galaxy Cloud
Galaxy
Info: report bugs | wiki | screencasts
GC Home
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
i-ebe8bf80
State: Ready
Alive: 38m 59s
Filesystems
Permissions
JobScheduler
Filesystems
Database
Scheduler
Galaxy
Cluster status log
Display a menu


Galaxy Cloud
http://ec2-184-73-86-186.compute-1.amazonaws.com/cloud
cloud computing
AWS Management Console
Galaxy
Galaxy Cloud
Galaxy
Info: report bugs | wiki | screencasts
GC Home
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: 59G / 100G (59%)
Instance status: Idle: 6 Available: 12 Requested: 12
Filesystems
Database
Scheduler
Galaxy
Cluster status log
Display a menu

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

So, what is that workflow actually doing?

Determining a good criteria for identifying heteroplasmic sites

1. Using PCR Replicates

- Errors in initial PCR sequencing could appear as very low frequency heteroplasmic sites
- Experimental design specifically allows addressing errors in PCR amplification prior to sequencing
- Comparisons between PCR replicates (sample by sample) determined the maximum observed variation was less than 1%
- Thus, at a detection threshold of 1% or greater, no variation between PCR duplicates

2. Using simulations

- The smallest dataset consiss of ~502,028 reads, with a median coverage of 1,170x
- Using this coverage, simulared datasets with different minor allele frequencies (0.001, 0.01, 0.05, 0.1), and sequencing error rates (0.001, 0.01, 0.02, 0.05)
- With a minor allele frequency of 0.01, and an error rate of 0.001, false positive and false negative rates were 0.

3. Strand bias filtering

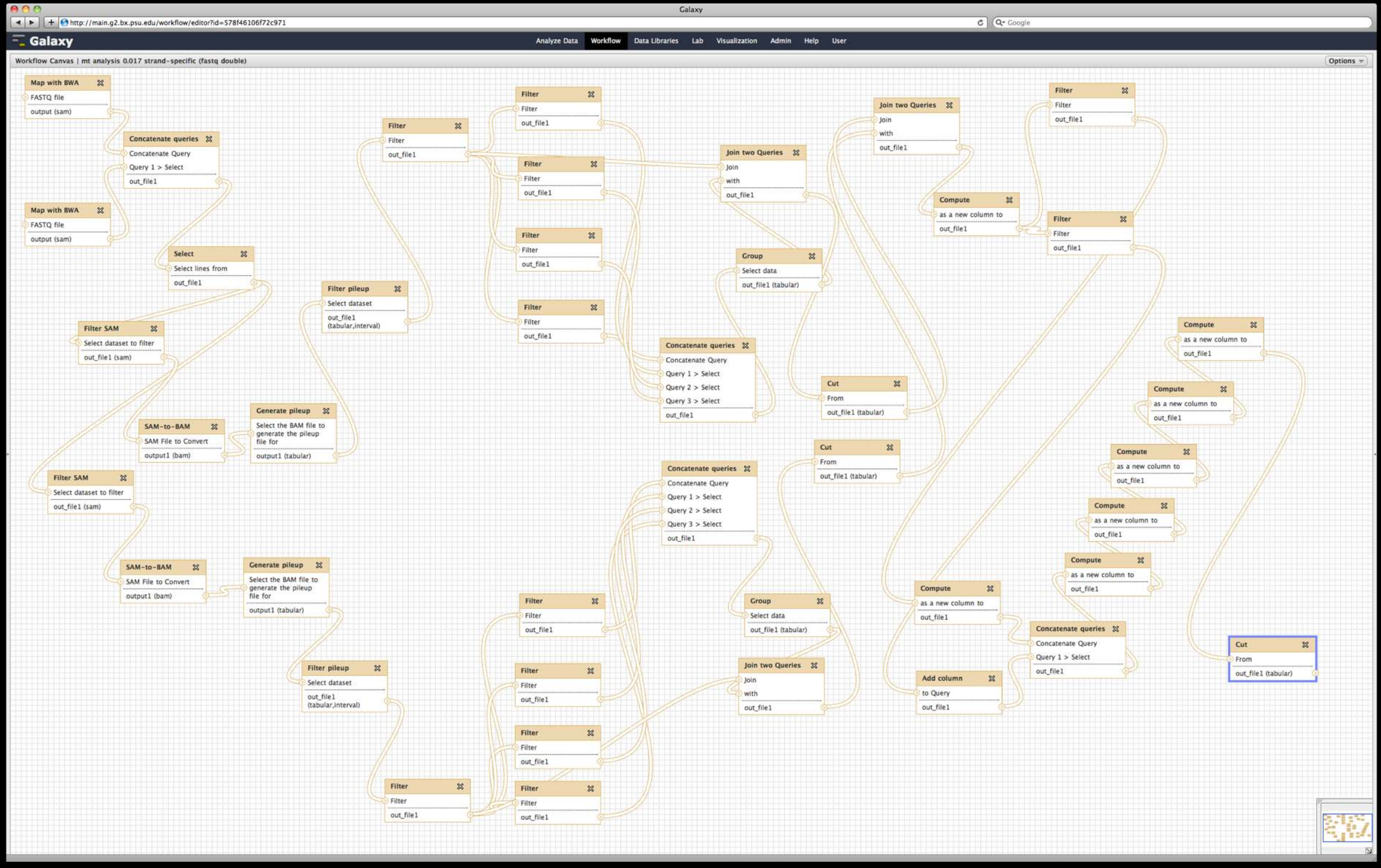
- Illumina sequencing has specific biases, tends toward some specific misidentifications (e.g. A to C).
- Real variants should not show any bias in reads from one strand vs. the other
- Filtering for variants seen in reads aligning to both strands should eliminate these problems (Li et al, AJHG, 2010).

4. Sequencing a clonal specimen

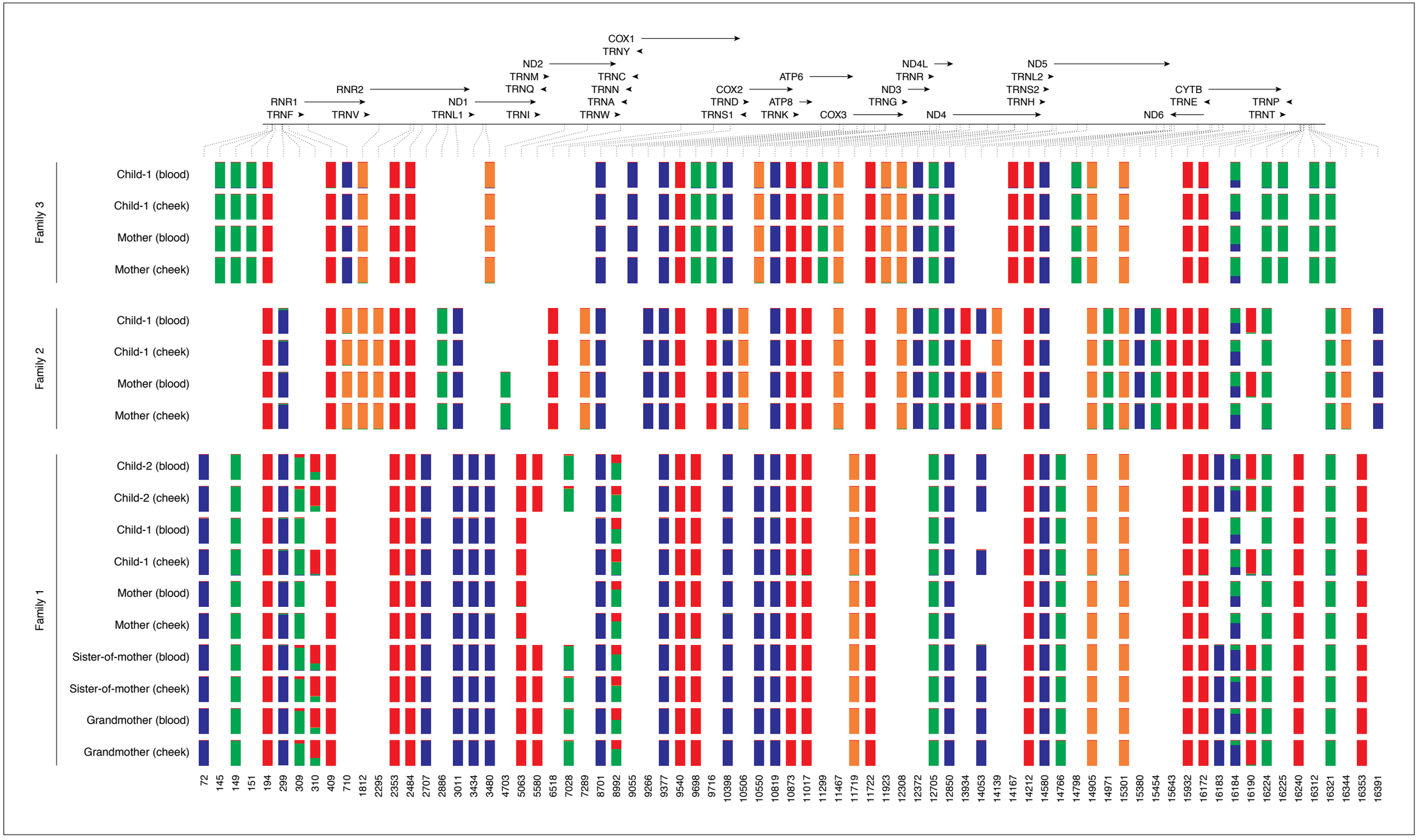
- Sequence a DNA molecule which should have no variation at all, all observed variation will be technical error
- Sequenced plasmid pUC18 to median 1,157,250x with minimum site coverage of 19,382x
- All positions contained at least 2 variant reads, but only one over the simulation threshold of 0.01
- At this one site, this level of variation was not supported on both strands

Heteroplasmy detection workflow

- Map reads to hg19 genome using BWA (Li and Durbin, 2009), and remove non-mt matches
- Pool PCR replicates (no power was gained by treating them separately)
- Split reads by strand
- Identify sites with variation for a given base greater than the detection threshold of 0.02 on both strands (twice our estimated detection level to be conservative)

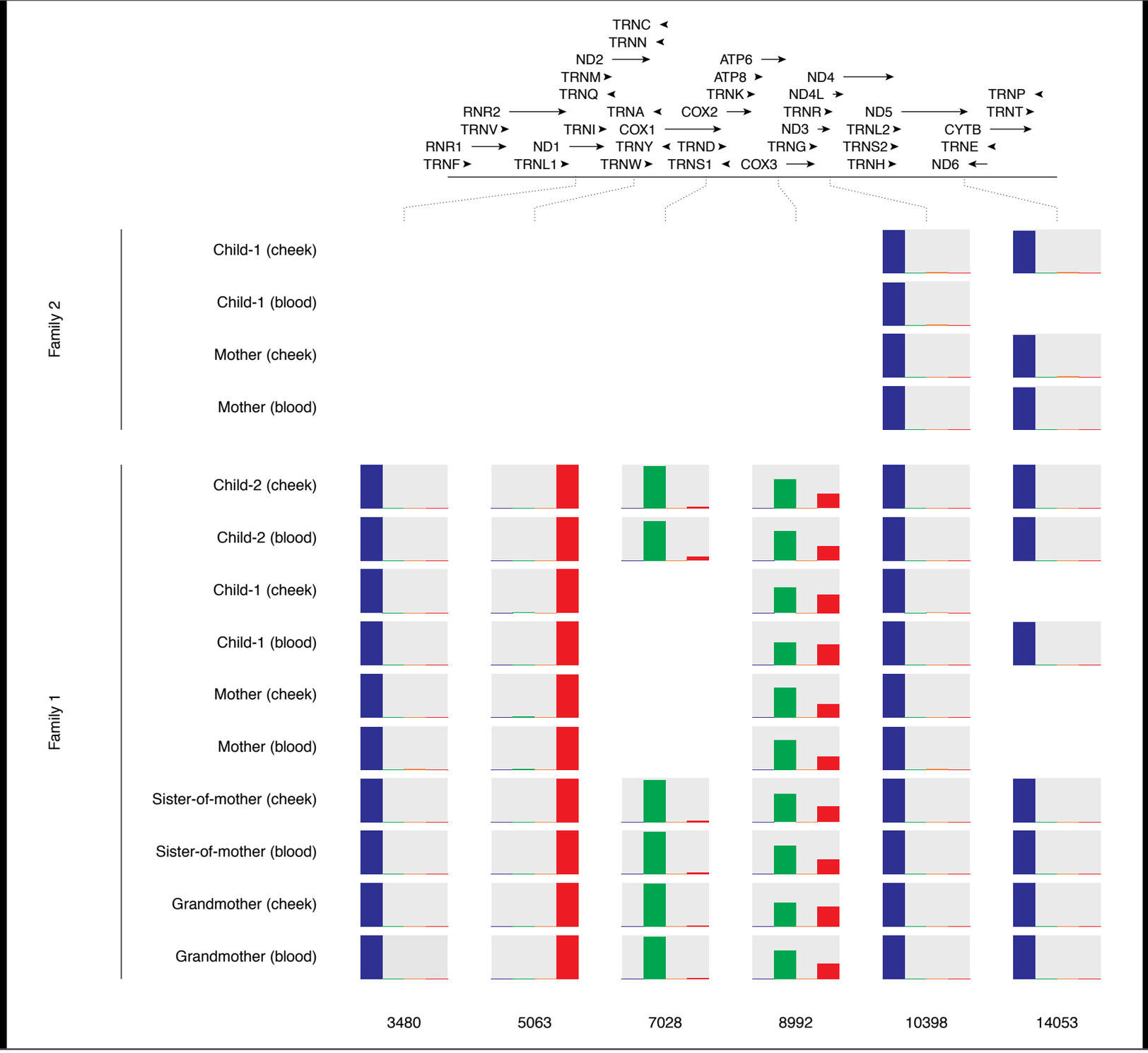

Galaxy
http://main.g2.bx.psu.edu/workflow/editor?id=578f46106f72c971
Galaxy
Analyze Data
Workflow
Data Libraries
Lab
Visualization
Admin
Help
User
Workflow Canvas | mt analysis 0.017 strand-specific (fastq double)
Options
Map with BWA
FASTQ file
output (sam)
Concatenate queries
Concatenate Query
Query 1 > Select
out_file1
Map with BWA
FASTQ file
output (sam)
Select
Select lines from
out_file1
Filter SAM
Select dataset to filter
out_file1 (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Generate pileup
Select the BAM file to generate the pileup file for
output1 (tabular)
Filter SAM
Select dataset to filter
out_file1 (sam)
SAM-to-BAM
SAM File to Convert
output1 (bam)
Generate pileup
Select the BAM file to generate the pileup file for
output1 (tabular)
Filter pileup
Select dataset
out_file1 (tabular,interval)
Filter pileup
Select dataset
out_file1 (tabular,interval)
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Join two Queries
Join
with
out_file1
Join two Queries
Join
with
out_file1
Group
Select data
out_file1 (tabular)
Group
Select data
out_file1 (tabular)
Concatenate queries
Concatenate Query
Query 1 > Select
Query 2 > Select
Query 3 > Select
out_file1
Concatenate queries
Concatenate Query
Query 1 > Select
Query 2 > Select
Query 3 > Select
out_file1
Cut
From
out_file1 (tabular)
Cut
From
out_file1 (tabular)
Join two Queries
Join
with
out_file1
Compute
as a new column to
out_file1
Filter
Filter
out_file1
Filter
Filter
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Compute
as a new column to
out_file1
Add column
to Query
out_file1
Concatenate queries
Concatenate Query
Query 1 > Select
out_file1
Cut
From
out_file1 (tabular)

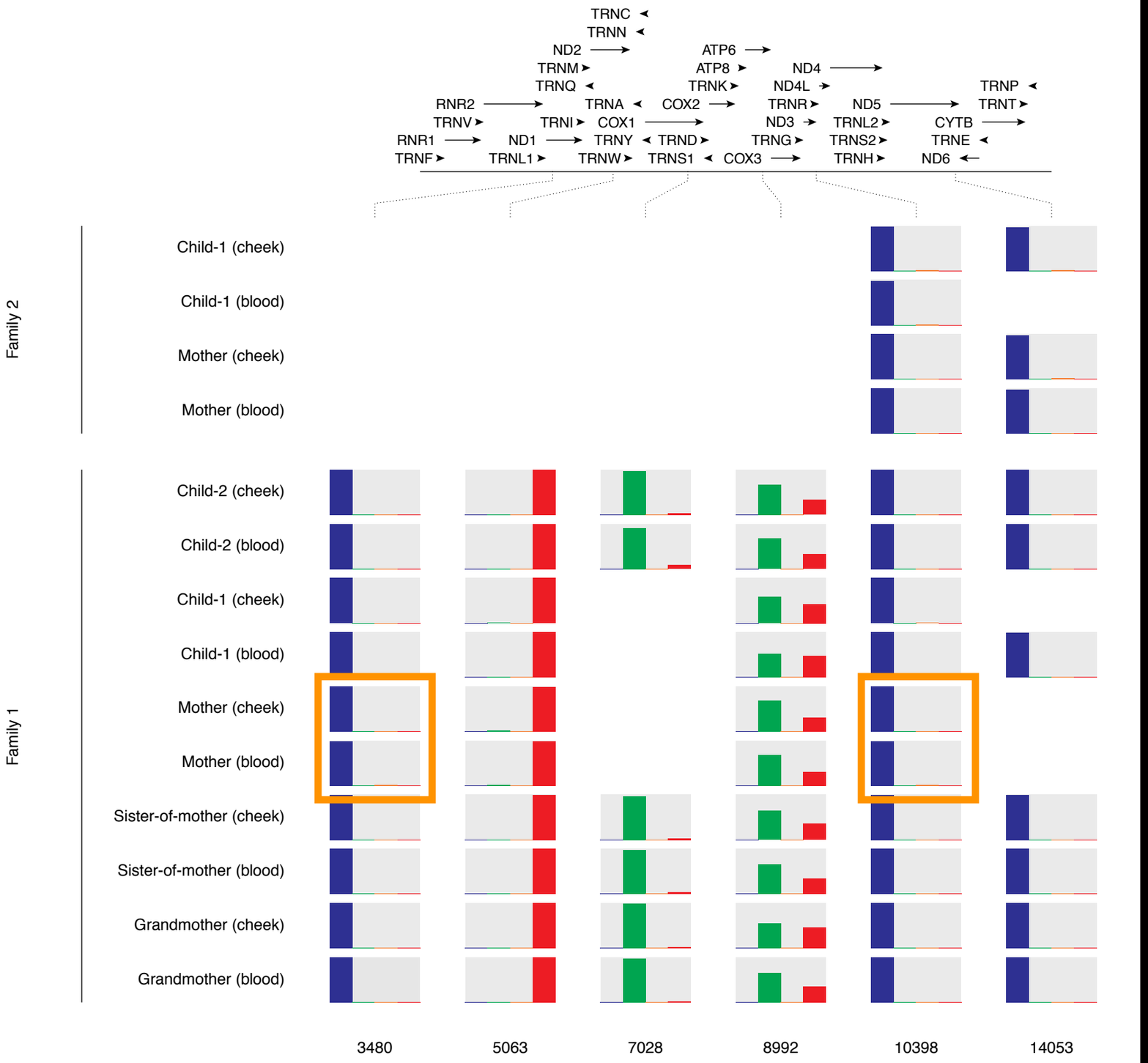
~1 hour and ~$20 later... results!


COX1
TRNY
ND2
TRNM
TRNQ
TRNC
TRNN
TRNA
TRNW
RNR2
RNR1
TRNF
TRNV
ND1
TRNL1
TRNI
ATP6
COX2
TRND
TRNS1
ATP8
TRNK
COX3
ND4L
TRNR
ND3
TRNG
ND4
ND5
TRNL2
TRNS2
TRNH
CYTB
TRNE
ND6
TRNP
TRNT
Family 3
Child-1 (blood)
Child-1 (cheek)
Mother (blood)
Mother (cheek)
Family 2
Child-1 (blood)
Child-1 (cheek)
Mother (blood)
Mother (cheek)
Family 1
Child-2 (blood)
Child-2 (cheek)
Child-1 (blood)
Child-1 (cheek)
Mother (blood)
Mother (cheek)
Sister-of-mother (blood)
Sister-of-mother (cheek)
Grandmother (blood)
Grandmother (cheek)
72
145
149
151
194
299
309
310
409
710
1812
2295
2353
2484
2707
2886
3011
3434
3480
4703
5063
5580
6518
7028
7289
8701
8992
9055
9266
9377
9540
9698
9716
10398
10506
10550
10819
10873
11017
11299
11467
11719
11722
11923
12308
12372
12705
12850
13934
14053
14139
14167
14212
14580
14766
14798
14905
14971
15301
15380
15454
15643
15932
16172
16183
16184
16190
16224
16225
16240
16312
16321
16344
16353
16391

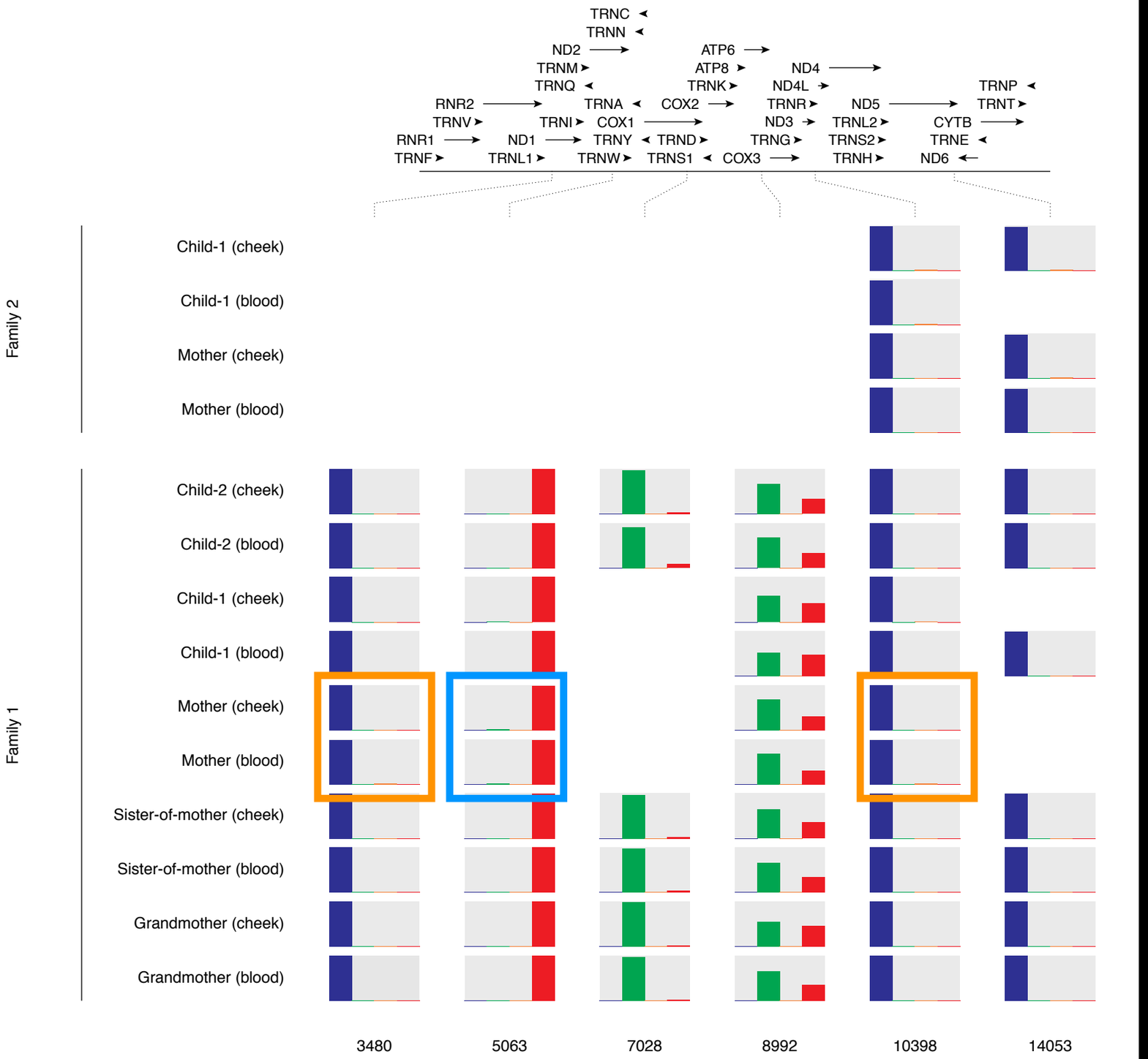
Heteroplasmies in the three families

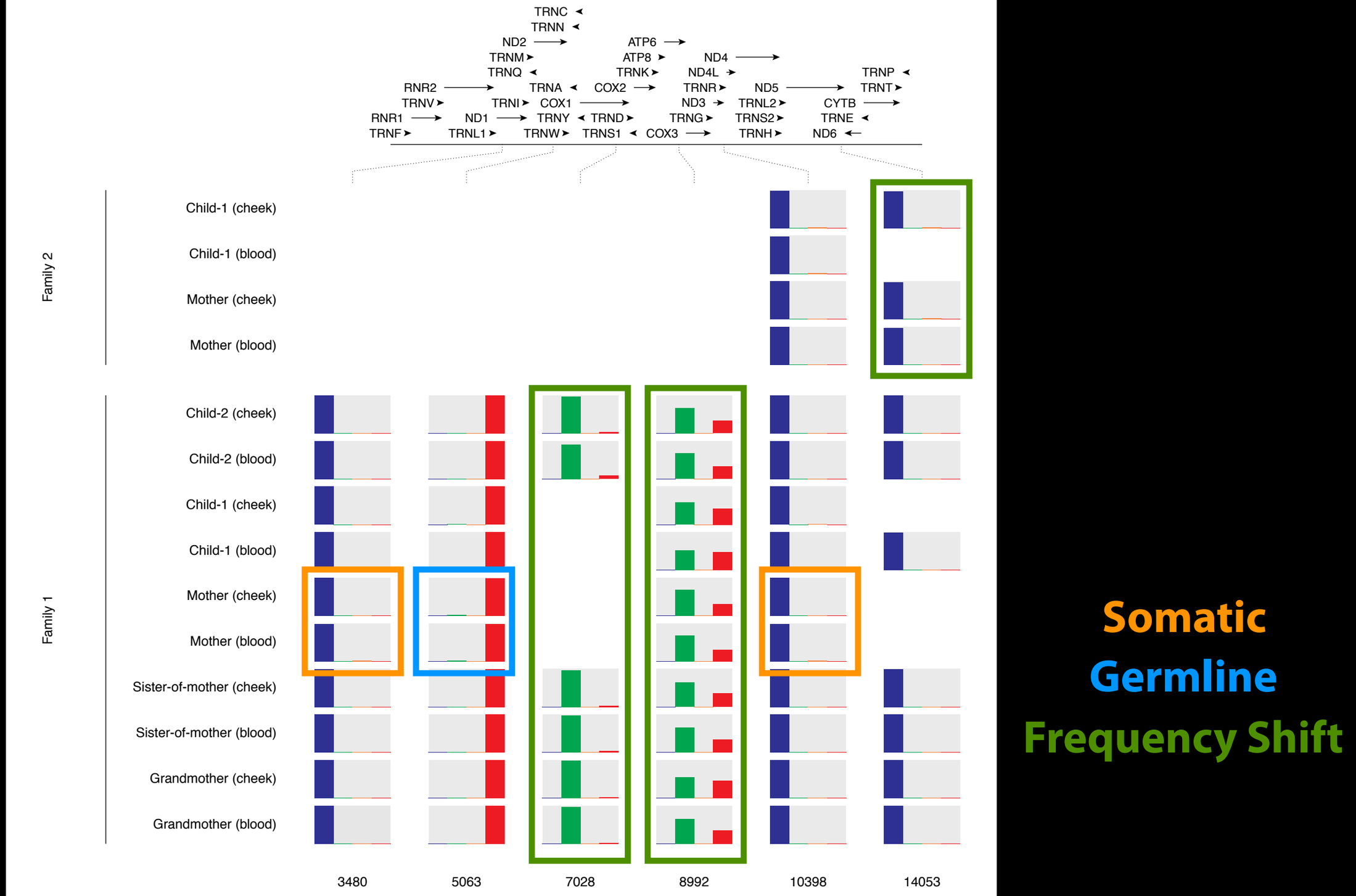
- Of 75 sites showing variation between at least one individual and the reference, 6 show evidence of heteroplasmy
  - Two sites in low complexity regions excluded

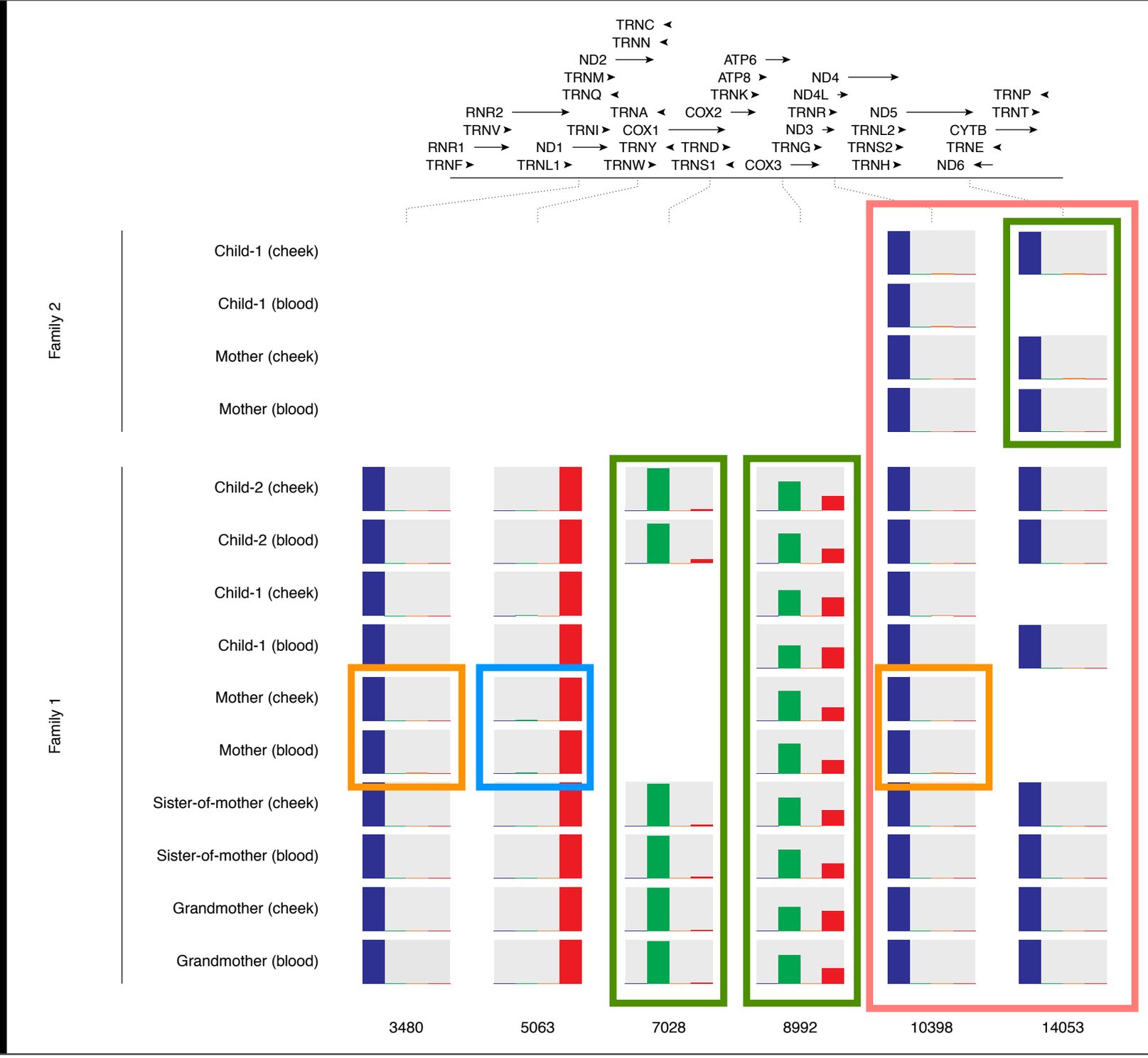

TRNC
TRNN
ND2
ATP6
TRNM
ATP8
ND4
TRNQ
TRNK
ND4L
TRNP
RNR2
TRNA
COX2
TRNR
ND5
TRNT
TRNV
TRNI
COX1
ND3
TRNL2
CYTB
RNR1
ND1
TRNY
TRND
TRNG
TRNS2
TRNE
TRNF
TRNL1
TRNW
TRNS1
COX3
TRNH
ND6
Family 2
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Family 1
Child-2 (cheek)
Child-2 (blood)
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Sister-of-mother (cheek)
Sister-of-mother (blood)
Grandmother (cheek)
Grandmother (blood)
3480
5063
7028
8992
10398
14053


TRNC
TRNN
ND2
ATP6
TRNM
ATP8
ND4
TRNQ
TRNK
ND4L
TRNP
RNR2
TRNA
COX2
TRNR
ND5
TRNT
TRNV
TRNI
COX1
ND3
TRNL2
CYTB
RNR1
ND1
TRNY
TRND
TRNG
TRNS2
TRNE
TRNF
TRNL1
TRNW
TRNS1
COX3
TRNH
ND6
Family 2
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Family 1
Child-2 (cheek)
Child-2 (blood)
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Sister-of-mother (cheek)
Sister-of-mother (blood)
Grandmother (cheek)
Grandmother (blood)
3480
5063
7028
8992
10398
14053

Somatic


TRNC
TRNN
ND2
ATP6
TRNM
ATP8
ND4
TRNQ
TRNK
ND4L
TRNP
RNR2
TRNA
COX2
TRNR
ND5
TRNT
TRNV
TRNI
COX1
ND3
TRNL2
CYTB
RNR1
ND1
TRNY
TRND
TRNG
TRNS2
TRNE
TRNF
TRNL1
TRNW
TRNS1
COX3
TRNH
ND6
Family 2
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Family 1
Child-2 (cheek)
Child-2 (blood)
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Sister-of-mother (cheek)
Sister-of-mother (blood)
Grandmother (cheek)
Grandmother (blood)
3480
5063
7028
8992
10398
14053
Somatic
Germline


TRNC
TRNN
ND2
ATP6
TRNM
ATP8
ND4
TRNQ
TRNK
ND4L
TRNP
RNR2
TRNA
COX2
TRNR
ND5
TRNT
TRNV
TRNI
COX1
ND3
TRNL2
CYTB
RNR1
ND1
TRNY
TRND
TRNG
TRNS2
TRNE
TRNF
TRNL1
TRNW
TRNS1
COX3
TRNH
ND6
Family 2
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Family 1
Child-2 (cheek)
Child-2 (blood)
Child-1 (cheek)
Child-1 (blood)
Mother (cheek)
Mother (blood)
Sister-of-mother (cheek)
Sister-of-mother (blood)
Grandmother (cheek)
Grandmother (blood)
3480
5063
7028
8992
10398
14053
Somatic
Germline
Frequency Shift



Somatic

Germline

Frequency Shift

Nonsynonymous

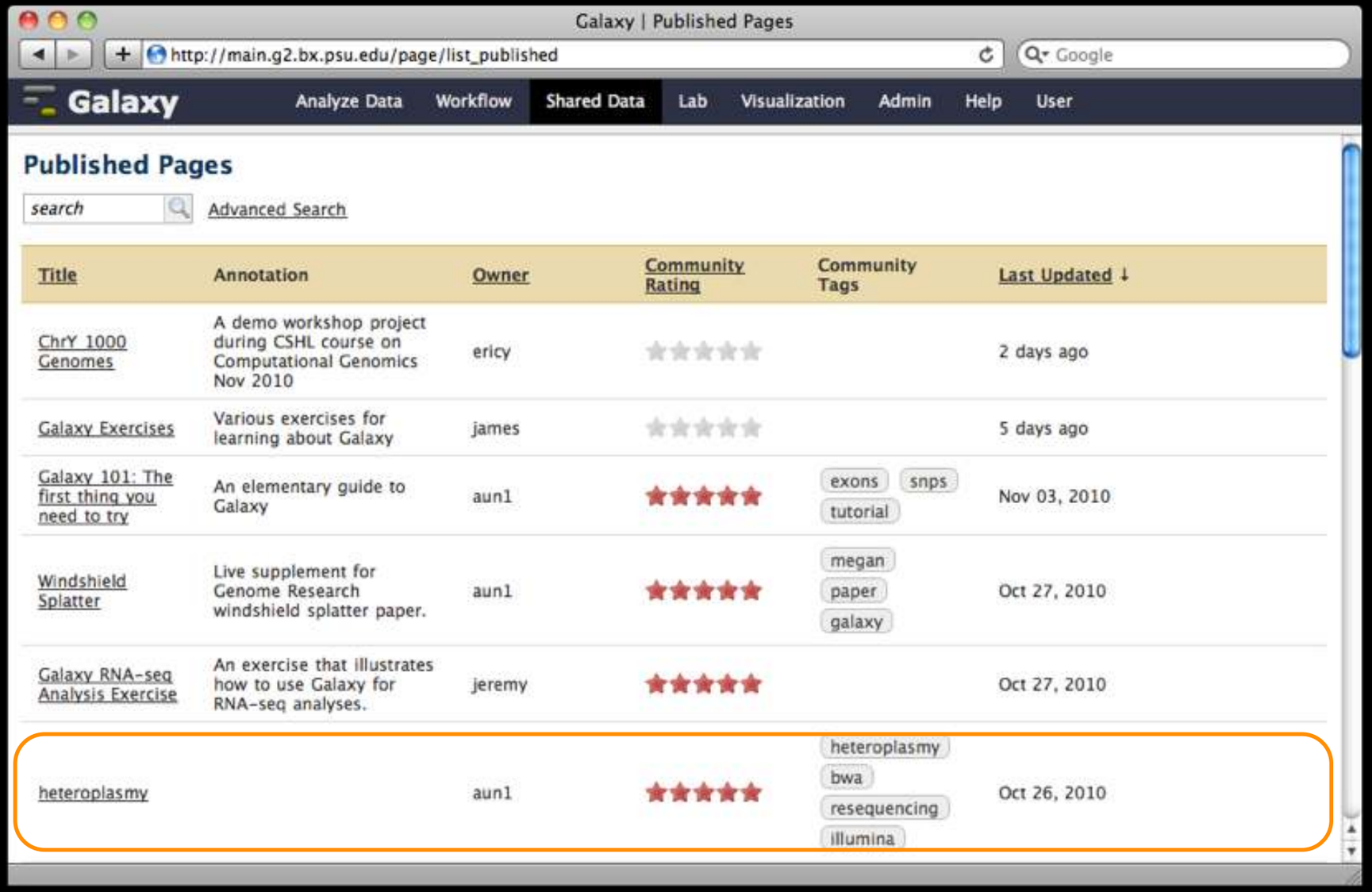
Some conclusions...

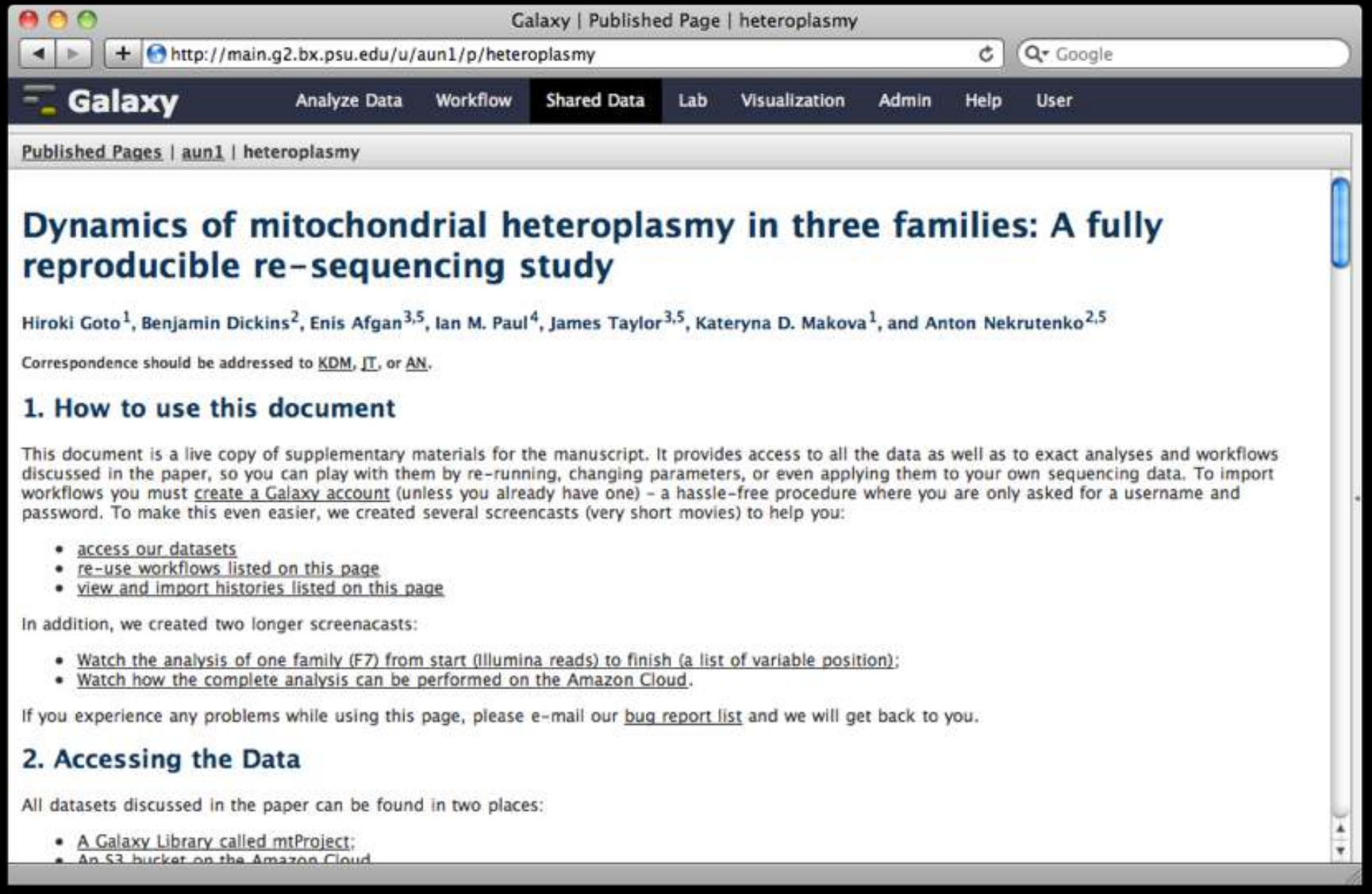
- Heteroplasmy is relatively infrequent, even with the ability to detect low frequency variation
- Heteroplasmy frequencies change through transmission events: e.g. for heteroplasmic site, allele frequencies were significantly different for all but one transmission

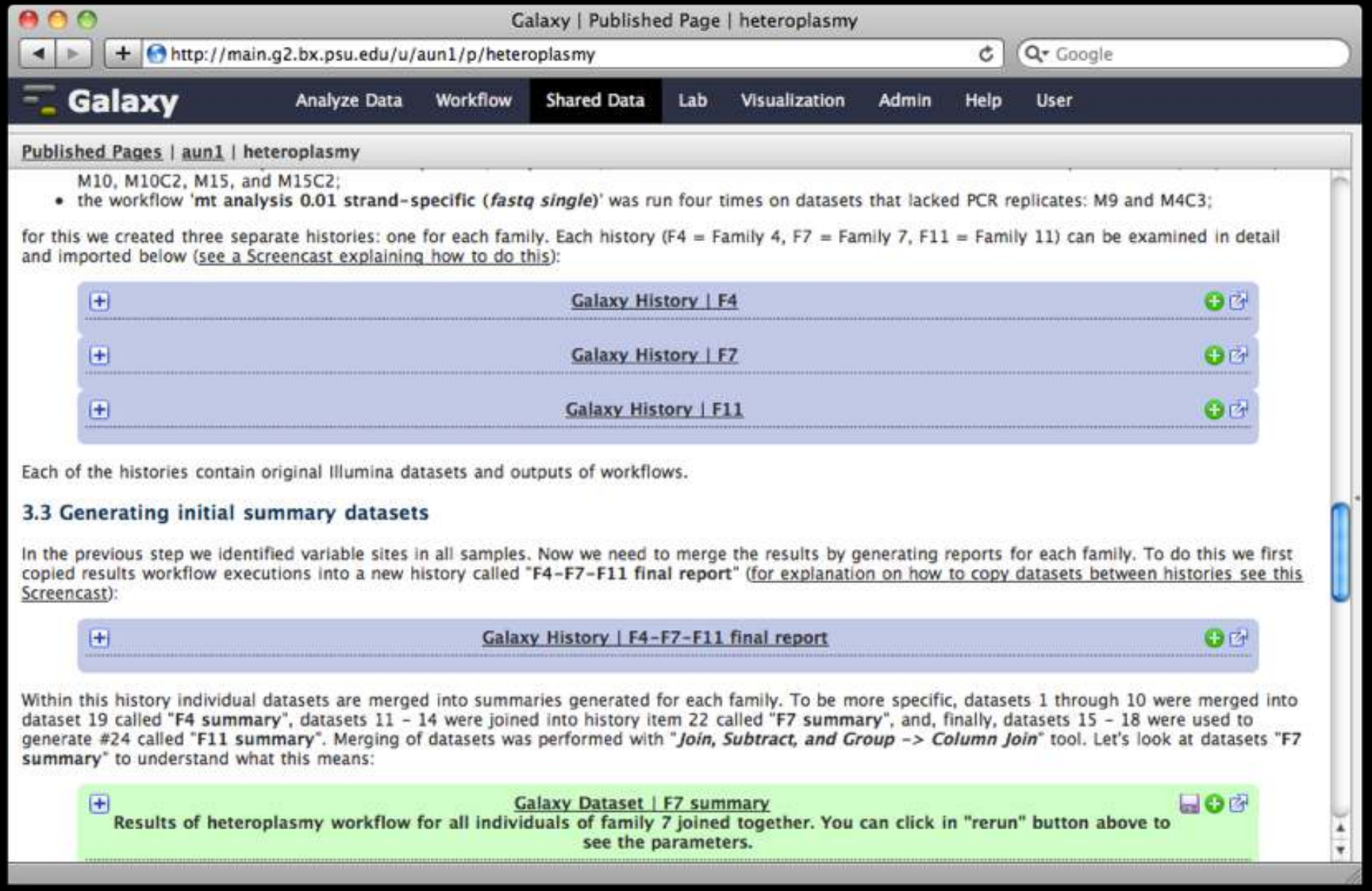
Do you trust me?

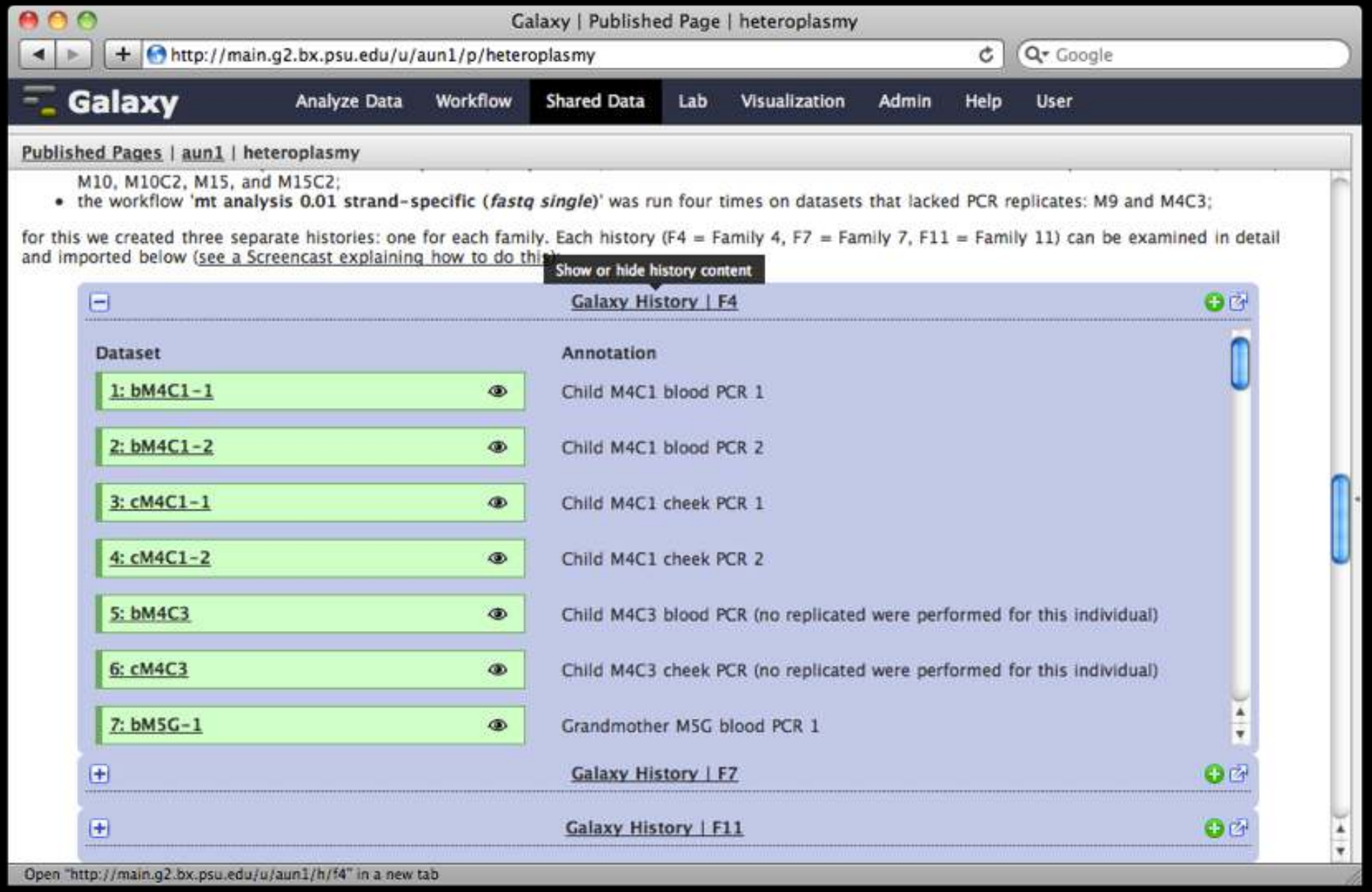
You shouldn't have to...

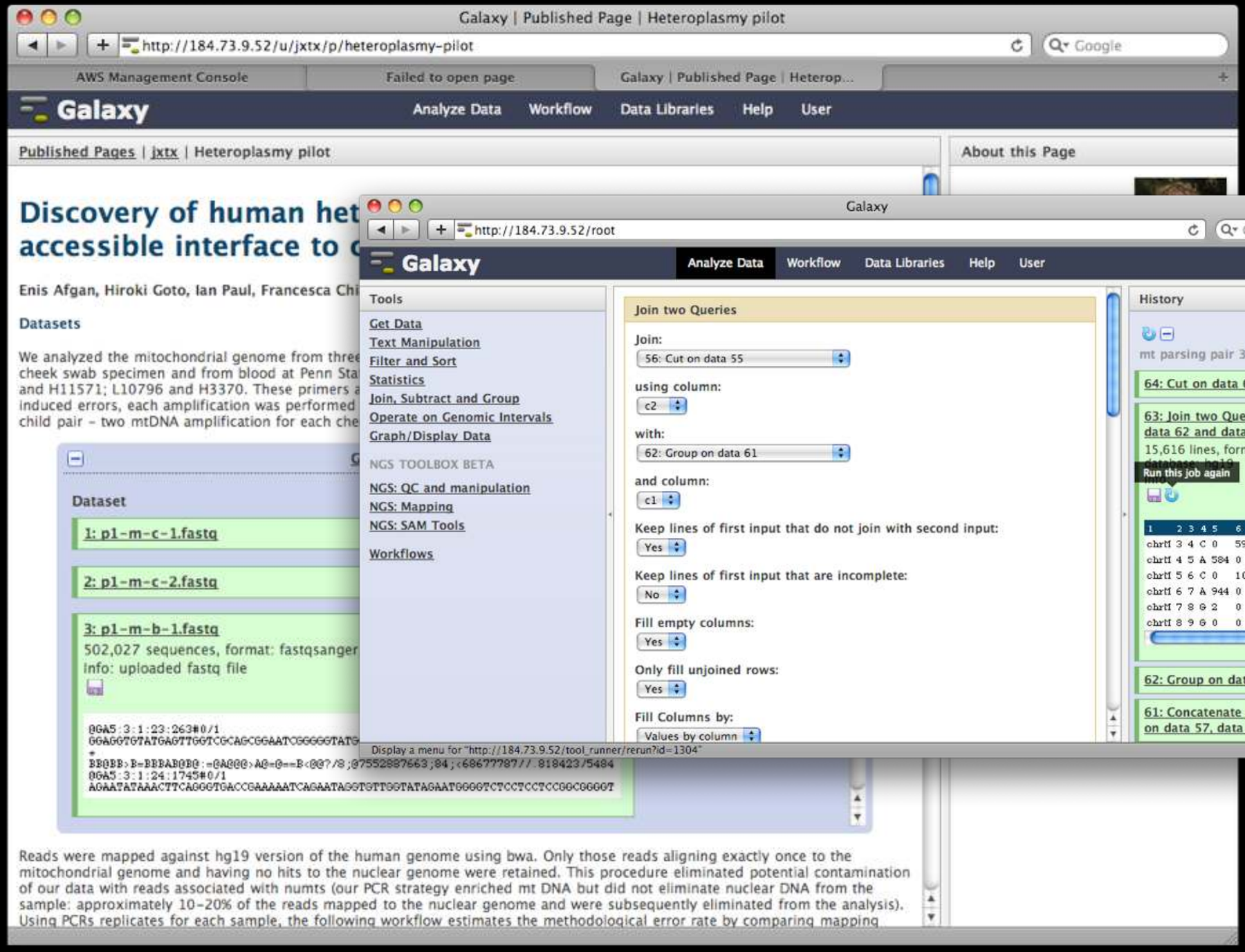
http://usegalaxy.org/heteroplasmy


Galaxy | Published Pages
http://main.g2.bx.psu.edu/page/list_published
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages
search
Advanced Search
Title
Annotation
Owner
Community Rating
Community Tags
Last Updated ↓
ChrY 1000 Genomes
A demo workshop project during CSHL course on Computational Genomics Nov 2010
ericy
2 days ago
Galaxy Exercises
Various exercises for learning about Galaxy
james
5 days ago
Galaxy 101: The first thing you need to try
An elementary guide to Galaxy
aun1
exons
snps
tutorial
Nov 03, 2010
Windshield Splatter
Live supplement for Genome Research windshield splatter paper.
aun1
megan
paper
galaxy
Oct 27, 2010
Galaxy RNA-seq Analysis Exercise
An exercise that illustrates how to use Galaxy for RNA-seq analyses.
jeremy
Oct 27, 2010
heteroplasmy
aun1
heteroplasmy
bwa
resequencing
illumina
Oct 26, 2010


Galaxy | Published Page | heteroplasmy
http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages | aun1 | heteroplasmy
Dynamics of mitochondrial heteroplasmy in three families: A fully reproducible re-sequencing study
Hiroki Goto[1], Benjamin Dickins[2], Enis Afgan[3,5], Ian M. Paul[4], James Taylor[3,5], Kateryna D. Makova[1], and Anton Nekrutenko[2,5]
Correspondence should be addressed to KDM, JT, 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 screenacasts:
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;


Galaxy | Published Page | heteroplasmy
http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages | aun1 | heteroplasmy
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):
Galaxy History | F4
Galaxy History | F7
Galaxy History | F11
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):
Galaxy History | F4-F7-F11 final report
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:
Galaxy Dataset | F7 summary
Results of heteroplasmy workflow for all individuals of family 7 joined together. You can click in "rerun" button above to see the parameters.


Galaxy | Published Page | heteroplasmy
http://main.g2.bx.psu.edu/u/aun1/p/heteroplasmy
Google
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Admin
Help
User
Published Pages | aun1 | heteroplasmy
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)
Show or hide history content
Galaxy History | F4
Dataset
Annotation
1: bM4C1-1
Child M4C1 blood PCR 1
2: bM4C1-2
Child M4C1 blood PCR 2
3: cM4C1-1
Child M4C1 cheek PCR 1
4: cM4C1-2
Child M4C1 cheek PCR 2
5: bM4C3
Child M4C3 blood PCR (no replicated were performed for this individual)
6: cM4C3
Child M4C3 cheek PCR (no replicated were performed for this individual)
7: bM5G-1
Grandmother M5G blood PCR 1
Galaxy History | F7
Galaxy History | F11
Open "http://main.g2.bx.psu.edu/u/aun1/h/f4" in a new tab


Galaxy | Published Page | Heteroplasmy pilot
http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot
Google
AWS Management Console
Failed to open page
Galaxy | Published Page | Heterop...
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Published Pages | jxtx | Heteroplasmy pilot
About this Page
Discovery of human het
accessible interface to
Enis Afgan, Hiroki Goto, Ian Paul, Francesca Chi
Datasets
We analyzed the mitochondrial genome from three
cheek swab specimen and from blood at Penn Sta
and H11571; L10796 and H3370. These primers a
induced errors, each amplification was performed
child pair - two mtDNA amplification for each che
Dataset
1: p1-m-c-1.fastq
2: p1-m-c-2.fastq
3: p1-m-b-1.fastq
502,027 sequences, format: fastqsanger
info: uploaded fastq file
@GA5:3:1:23:263#0/1
GGAGGTGTATGAGTTGGTCGCAGCGGAATCGGGGGTATG
+
BB@BB>B=BBBAB@B@:=@A@@@>A@=@==B<@@?/8;@7552887663;84;<686777877/.818423/5484
@GA5:3:1:24:1745#0/1
AGAATATAAACTTCAGGGTGACCGAAAAATCAGAATAGGTGTTGGTATAGAATGGGGTCTCCTCCTCCGGCGGGGT
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 numts (our PCR strategy enriched mt DNA 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). Using PCRs replicates for each sample, the following workflow estimates the methodological error rate by comparing mapping
Galaxy
http://184.73.9.52/root
Google
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
Text Manipulation
Filter and Sort
Statistics
Join, Subtract and Group
Operate on Genomic Intervals
Graph/Display Data
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Workflows
Join two Queries
Join:
56: Cut on data 55
using column:
c2
with:
62: Group on data 61
and column:
c1
Keep lines of first input that do not join with second input:
Yes
Keep lines of first input that are incomplete:
No
Fill empty columns:
Yes
Only fill unjoined rows:
Yes
Fill Columns by:
Values by column
Display a menu for "http://184.73.9.52/tool_runner/rerun?id=1304"
History
Options
mt parsing pair 3
64: Cut on data 63
63: Join two Queries on data 62 and data 56
15,616 lines, format: tabular,
Run this job again
62: Group on data 61
61: Concatenate queries on data 57, data 60, and others

Galaxy | Published Page | Heteroplasmy pilot

http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot

AWS Management Console | Galaxy | Published Page | Heterop...

Galaxy | Analyze Data | Workflow | Data Libraries | Help | User

Published Pages | jxtx | Heteroplasmy pilot

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

Galaxy History | mt datasets

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 numts (our PCR strategy enriched mt DNA 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). 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 the rate of 1.18% (12/1012) at this position.

Galaxy Workflow | Determine threshold from PCR replicates

c1=='chrM' and c10 >= 200

Step 16: Filter

Filter
Output dataset 'out_file1' from step 14

With following condition
c1=='chrM' and c10 >= 200

Replicate 2: Keep only positions that map to chrM and have quality adjusted coverage greater than 200

Step 17: Join

Join
Output dataset 'out_file1' from step 15

with
Output dataset 'out_file1' from step 16

Create a joined file containing the pileup information 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

Display a menu

About this Page

Author

jxtx

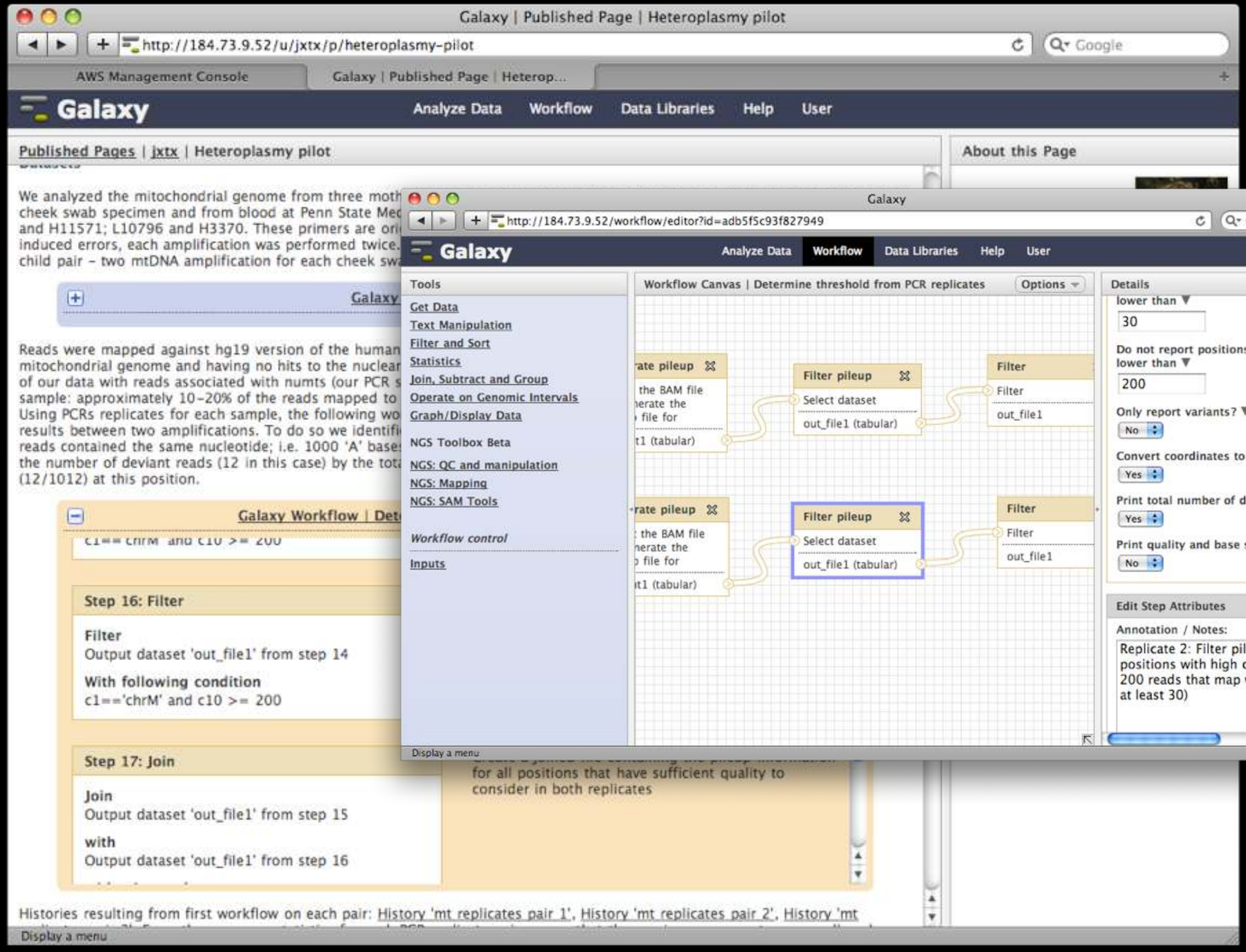
Related Pages

All published pages
Published pages by jxtx

Tags

Community:
cloud | heteroplasmy | ngs

Yours:
heteroplasmy × | cloud × | ngs ×


Galaxy | Published Page | Heteroplasmy pilot
http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot
AWS Management Console
Galaxy | Published Page | Heterop...
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Published Pages | jxtx | Heteroplasmy pilot
About this Page
We analyzed the mitochondrial genome from three moth
cheek swab specimen and from blood at Penn State Me
and H11571; L10796 and H3370. These primers are ori
induced errors, each amplification was performed twice.
child pair - two mtDNA amplification for each cheek sw
Galaxy
Reads were mapped against hg19 version of the human
mitochondrial genome and having no hits to the nuclear
of our data with reads associated with numts (our PCR s
sample: approximately 10-20% of the reads mapped to
Using PCRs replicates for each sample, the following wo
results between two amplifications. To do so we identifi
reads contained the same nucleotide; i.e. 1000 'A' base
the number of deviant reads (12 in this case) by the tot
(12/1012) at this position.
Galaxy Workflow | Det
c1=='chrM' and c10 >= 200
Step 16: Filter
Filter
Output dataset 'out_file1' from step 14
With following condition
c1=='chrM' and c10 >= 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
Display a menu
Galaxy
http://184.73.9.52/workflow/editor?id=adb5f5c93f827949
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Tools
Get Data
Text Manipulation
Filter and Sort
Statistics
Join, Subtract and Group
Operate on Genomic Intervals
Graph/Display Data
NGS Toolbox Beta
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
Workflow control
Inputs
Workflow Canvas | Determine threshold from PCR replicates
Options
Filter pileup
Select dataset
out_file1 (tabular)
Filter
Filter
out_file1
Details
lower than
30
Do not report positions with coverage lower than
200
Only report variants?
No
Convert coordinates to intervals?
Yes
Print total number of differences?
Yes
Print quality and base string?
No
Edit Step Attributes
Annotation / Notes:
Replicate 2: Filter pileup for positions with high coverage (over 200 reads that map with quality of at least 30)
Display a menu

Forward...

- Heteroplasmy
  - We can accurately identify heteroplasmy at low frequencies
  - Deeper sequencing, indels, linkage
  - Many more individuals, low incidence means we need to observe many more transmissions to estimate bottlenecks
- Galaxy
  - Long term archival and migration of analysis artifacts:

  - Cloud: autoscaling, more parallelism, instance migration

I want you to remember

exactly four (4) things:

http://usegalaxy.org

http://usegalaxy.org/**galaxy101**

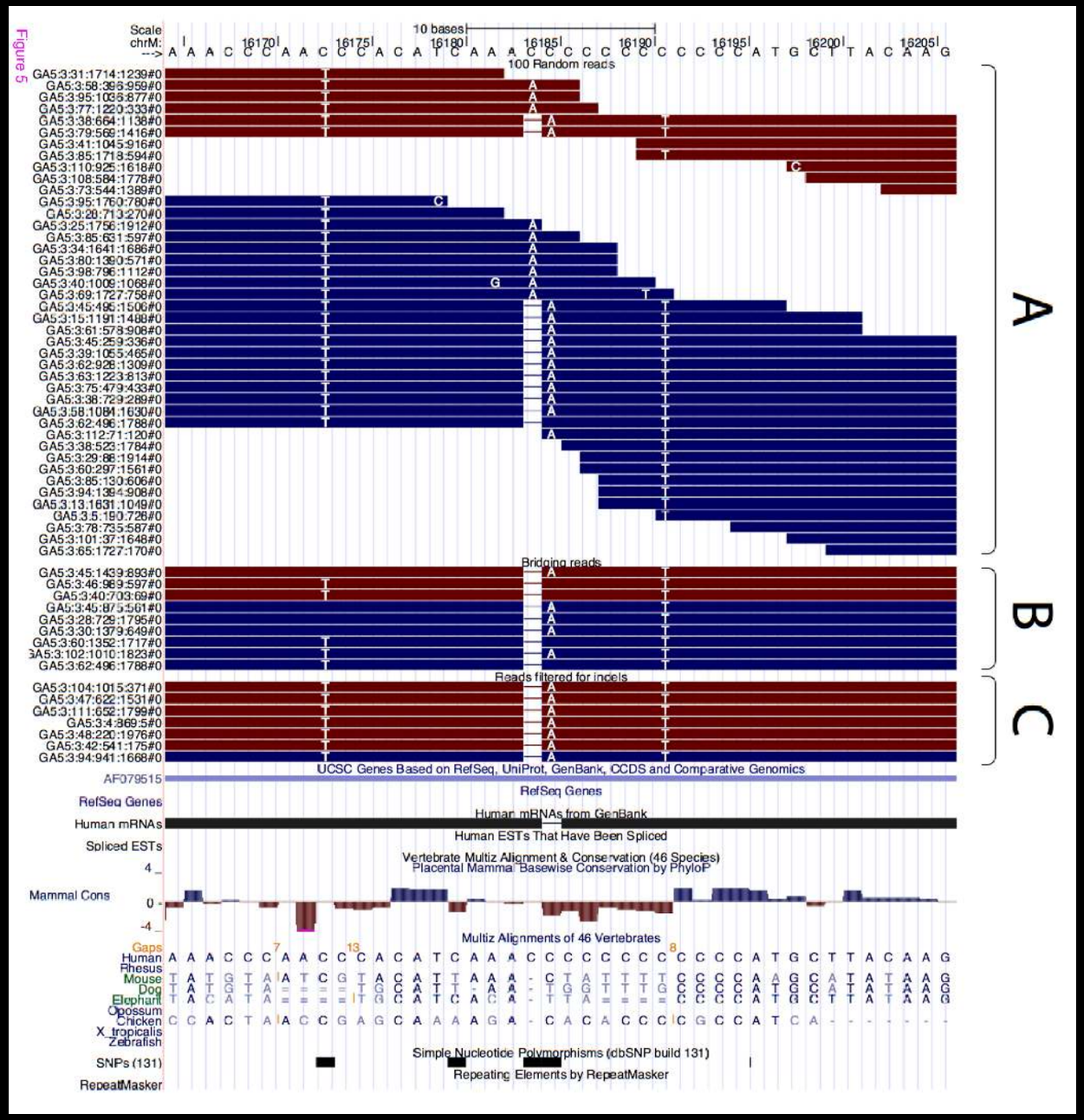
http://usegalaxy.org/**cloud**

http://getgalaxy.org

http://usegalaxy.org/

(end)

Erroneous Heteroplasmy at low complexity regions



- Site 16,190:
  - Appears to show A/C variation
  - However, this is almost certainly alignment error
  - Requiring 10 high quality aligning bases on either side reveals a fixed C deletion flanked by a fixed A substitution

LETTERS

Heteroplasmic mitochondrial DNA mutations in normal and tumour cells

Yiping He[1], Jian Wu[1], Devin C. Dressman[1], Christine Iacobuzio-Donahue[2], Sanford D. Markowitz[3], Victor E. Velculescu[1], Luis A. Diaz Jr[1], Kenneth W. Kinzler[1], Bert Vogelstein[1] & Nickolas Papadopoulos[1]

The presence of hundreds of copies of mitochondrial DNA (mtDNA) in each human cell poses a challenge for the complete characterization of mtDNA genomes by conventional sequencing technologies[1]. Here we describe digital sequencing of mtDNA genomes with the use of massively parallel sequencing-by-synthesis approaches. Although the mtDNA of human cells is considered to be homogeneous, we found widespread heterogeneity (heteroplasmy) in the mtDNA of normal human cells. Moreover, the frequency of heteroplasmic variants varied considerably between different tissues in the same individual. In addition to the variants identified in normal tissues, cancer cells harboured further homoplasmic and heteroplasmic mutations that could also be detected in patient plasma. These studies provide insights into the nature and variability of mtDNA sequences and have implications for mitochondrial processes during embryogenesis, cancer biomarker development and forensic analysis. In particular, they demonstrate that individual humans are characterized by a complex mixture of related mitochondrial genotypes rather than a single genotype.

Mitochondria are crucial in many basic cellular processes, and as a result of its unique maternal inheritance pattern and relatively high mutation rate, mtDNA is often used in studies of evolutionary biology and population genetics. These same attributes, combined with the high copy number of mtDNA in cells, makes mtDNA a favoured substrate for forensic analysis[2]. In typical human cells, there are between 50 and hundreds of mitochondria per cell and five to ten copies of mtDNA per mitochondrion[1]. The presence of multiple copies of mtDNA per cell leaves open the possibility that all the copies are not identical. Many studies have shown that mtDNA is homoplasmic in normal cells; that is, all of the mtDNA copies are identical not only in an individual cell but also among cells. However, there is apparently a low level of heteroplasmy in the mtDNA of various species, including humans[3–14]. To evaluate this issue further, we have used massively parallel sequencing-by-synthesis approaches to thoroughly characterize the mtDNA of normal and neoplastic human cells.

Two sets of PCR primers, each resulting in amplicons of about 650 base pairs (bp) in length, were designed to cover the mtDNA genome (Fig. 1a). Sequencing libraries for Illumina GAII made from the PCR products of normal colonic mucosa DNA (patient 1) yielded 8.5 million tags that matched the mitochondrial genome. Each mtDNA base was sequenced, on average, 16,700 times and fewer than 11 bases (0.07% of the 16,569 bp in the mtDNA genome) were represented fewer than 1,000 times (Supplementary Fig. 1a).

This high coverage permitted us to identify heteroplasmic variants even when they were relatively rare—theoretically, when present in as few as 1 per 10,000 mitochondrial genomes. However, errors that had accumulated during the PCR and sequencing steps limited the sensitivity achieved. Control libraries made from PCR products of nuclear DNA demonstrated that the average proportion of mutations per base was 0.058%, with a standard deviation of 0.057%, and no base was mutated at greater than 0.82% frequency (Supplementary Information). We therefore made the very conservative assumption that all variants present in excess of twice this value (1.6%) represented true heteroplasmies rather than sequencing artefacts. Using these criteria, we detected 28 homoplasmic alleles and 8 heteroplasmic alleles in this sample of normal colonic mucosa (patient 1). Homoplasmic alleles were defined as any allele not present in the standard mtDNA reference sequence of humans but present in more than 98.4% (that is, 100% − 1.6%) of the mtDNA sequences analysed. All homoplasmic alleles identified in patient 1 had previously been identified in normal individuals. The less frequent (minor) alleles at the heteroplasmic sites represented 1.6–29.7% of the total alleles at that site (Table 1). All (100%) of these eight heteroplasmic alleles were listed as normal variants in mtDNA databases, whereas only 3,601

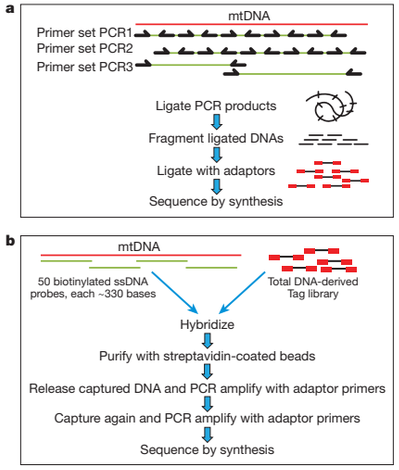


Figure 1 | Sequencing strategy. a, PCR amplification for mtDNA enrichment. **b**, Capture-based method for mtDNA enrichment.

[1]The Ludwig Center for Cancer Genetics and Therapeutics and The Howard Hughes Medical Institute at The Johns Hopkins Kimmel Cancer Center, Baltimore, Maryland 21231, USA. [2]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, Maryland 21231, USA. [3]Departments of Medicine, and Ireland Cancer Center, at Case Western Reserve University and Case Medical Center of University Hospitals of Cleveland, and The Howard Hughes Medical Institute, Cleveland, Ohio 44106, USA.

ARTICLE

Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes

Mingkun Li,[1,*] Anna Schönberg,[1] Michael Schaefer,[1] Roland Schroeder,[1] Ivane Nasidze,[1] and Mark Stoneking[1,*]

Heteroplasmy, the existence of multiple mtDNA types within an individual, has been previously detected by using mostly indirect methods and focusing largely on just the hypervariable segments of the control region. Next-generation sequencing technologies should enable studies of heteroplasmy across the entire mtDNA genome at much higher resolution, because many independent reads are generated for each position. However, the higher error rate associated with these technologies must be taken into consideration to avoid false detection of heteroplasmy. We used simulations and phiX174 sequence data to design criteria for accurate detection of heteroplasmy with the Illumina Genome Analyzer platform, and we used artificial mixtures and replicate data to test and refine the criteria. We then applied these criteria to mtDNA sequence reads for 131 individuals from five Eurasian populations that had been generated via a parallel tagged approach. We identified 37 heteroplasmies at 10% frequency or higher at 34 sites in 32 individuals. The mutational spectrum does not differ between heteroplasmic mutations and polymorphisms in the same individuals, but the relative mutation rate at heteroplasmic mutations is significantly higher than that estimated for all mutable sites in the human mtDNA genome. Moreover, there is also a significant excess of nonsynonymous mutations observed among heteroplasmies, compared to polymorphism data from the same individuals. Both mutation-drift and negative selection influence the fate of heteroplasmies to determine the polymorphism spectrum in humans. With appropriate criteria for avoiding false positives due to sequencing errors, next-generation technologies can provide novel insights into genome-wide aspects of mtDNA heteroplasmy.

Introduction

The mtDNA genome remains one of the most widely studied DNA segments in humans. It is particularly useful for studying population and evolutionary genetics because of its abundance in human cells, its uniparental, nonrecombining mode of inheritance, and its high mutation rate compared to that of the nuclear genome.[1] Although each individual is typically characterized by a single mtDNA type, in fact each individual is a population of mtDNA genomes, and the presence of multiple mtDNA types within an individual is termed heteroplasmy.

Although little noted at the time, the first report of heteroplasmy in humans was in 1983, involving a study of a noncoding region of human mtDNA from 11 human placentas.[2] Heteroplasmy has been investigated most often in correlation with mitochondrial disease, aging, and cancer.[3–6] To date, more than 400 mtDNA mutations have been associated with human disease, and most were observed in heteroplasmic states, with pathogenic mutations coexisting with normal mitochondrial genomes.[7] This suggests that the heteroplasmic level is of particular interest, as the disease phenotype becomes evident only when the percentage of mutant molecules exceeds a critical threshold value. Although this value differs for different mutations and in different tissues, it is usually in the range of 70%~90%.[8,9]

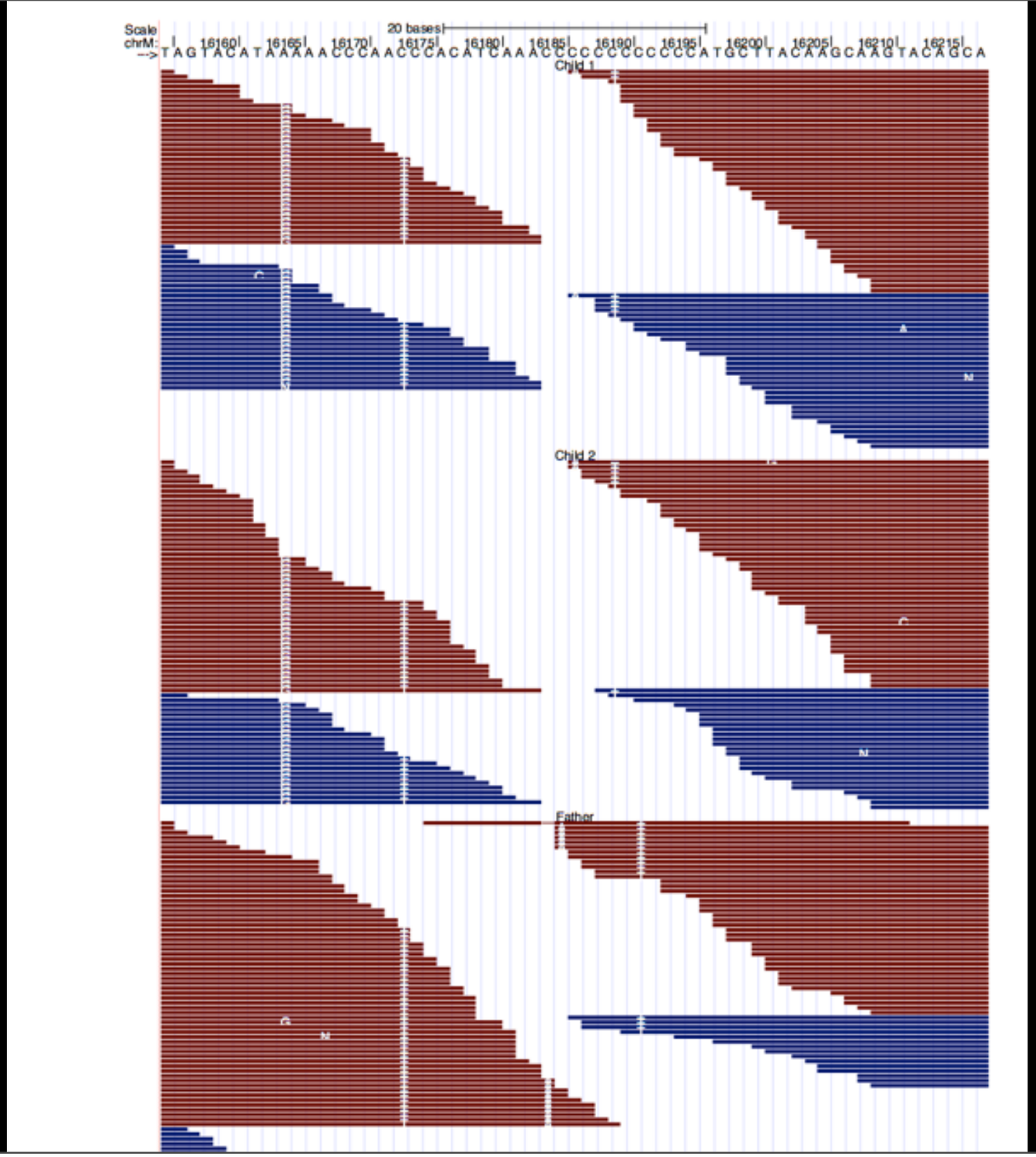
Originally, heteroplasmy was believed to be quite rare in healthy individuals,[10,11] but subsequent studies found many non-disease-related heteroplasmies.[12–15] Moreover, heteroplasmy has started to play an important role in some forensic investigations.[16,17] Thus, heteroplasmy can also be a useful genetic marker. Regarding heteroplasmy as the intermediate stage between the generation of mutations and the fixation of mutations in the individual or cell, it represents polymorphisms within the populations of mitochondrial genomes in one cell or tissue. Thus, it can be a potential resource for studying the mutational pattern, possible role of natural selection, and even the existence of recombination in mtDNA.[18] For example, de novo mtDNA mutations in cancer tissues preferentially locate at the same positions as ancient variants in the human phylogeny, indicating similar selective constraints.[19] Understanding the basis, extent, and forces influencing the occurrence and subsequent fate of heteroplasmic mtDNA mutations is one of the principal challenges facing scientists and clinicians in the field of mitochondrial genetics.

A variety of techniques have been employed for heteroplasmy detection, including Sanger capillary sequencing,[13] high-performance liquid chromatography (HPLC),[20] pyrosequencing,[21,22] SnaPshot,[23] high-resolution melt (HRM) profiling,[24] a temporal temperature gradient gel electrophoresis (TTGE) strategy,[25] the Invader assay,[26] an amplification refractory mutation system,[27] and surveyor nuclease.[28] However, all of these methods have disadvantages, including the following: for some methods, the candidate heteroplasmic position needs to be defined first; the method may not allow determination of the actual heteroplasmic position; the level of heteroplasmy cannot be

[1]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, D04103 Leipzig, Germany
*Correspondence: mingkun_li@eva.mpg.de (M.L.), stoneking@eva.mpg.de (M.S.)
DOI 10.1016/j.ajhg.2010.07.014.

Comparisons with other studies

- He et al (Nature) 2010: Using a 1.6% detection threshold from PCR replicates, identified 40 heteroplasmies in 10 individuals, concluding heteroplasmy is far more frequent than expected
- Li et al (AJHG) 2010: Using a 10% detection threshold (and strand filtering, simulation, ...) identified 37 heteroplasmies in 131 individuals, significantly less than He et al.
    - If we consider one individual from each family, to avoid relatedness, our results support this (p=0.44)



- He et al. call two heteroplasmic sites in this low-complexity region
- No reads (from their data) span the region
- Almost certainly misalignment of a single fixed variant

The developer's dilemma

- You have written a command-line script to analyze genomic data and you want to share it with command-line averse colleagues

The Galaxy solution

- Integrate the script as a new Tool into your own Galaxy server
- Steps:
  - Obtain and install Galaxy source code (GetGalaxy.org)
  - Write an XML file describing the inputs and outputs and how to execute the script

Cluster

Cluster intervals of: 6: UCSC Main on Human: knownGene

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

TIP: If your query does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns

---

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be considered "clustered". **Negative** values for distance are allowed, and are useful for clustering intervals that overlap.
- **Minimum intervals per cluster** allow a threshold to be set on the minimum number of intervals to be considered a cluster. Any area with less than this minimum will not be included in the ouput.
- **Merge clusters into single intervals** outputs intervals that span the entire cluster.
- **Find cluster intervals; preserve comments and order** filters out non-cluster intervals while maintaining the original ordering and comments in the file.
- **Find cluster intervals; output grouped by clusters** filters out non-cluster intervals, but outputs the cluster intervals so that they are grouped together. Comments and original ordering in the file are lost.

Example

cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
                   -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <param format="interval" name="input1" type="data">
      <label>Cluster intervals of</label>
    </param>
    <param name="distance" size="5" type="integer" value="1" help="(bp
      <label>max distance between intervals</label>
    </param>
    <param name="minregions" size="5" type="integer" value="2">
      <label>min number of intervals per cluster</label>
    </param>
    <param name="returntype" type="select" label="Return type">
      <option value="1">Merge clusters into single intervals</option>
      <option value="2">Find cluster intervals; preserve comments and
      <option value="3">Find cluster intervals; output grouped by clus
      <option value="4">Find the smallest interval in each cluster</op
      <option value="5">Find the largest interval in each cluster</opt
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does not appear in the pulldown menu -> it is n

-----

**Screencasts!**

See Galaxy Interval Operation Screencasts_ (right click to open this l

.. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc

-----

**Syntax**

- **Maximum distance** is greatest distance in base pairs allowed betw
- **Minimum intervals per cluster** allow a threshold to be set on the
- **Merge clusters into single intervals** outputs intervals that span
- **Find cluster intervals; preserve comments and order** filters out
- **Find cluster intervals; output grouped by clusters** filters out n
```

Line: 87 Column: 8 XML Soft Tabs: 2

Develop: tools

Cluster

Cluster intervals of: 6: UCSC Main on Human: knownGene

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

Execute

TIP: If your query does not appear in the pulldown menu -> it is not in interval format. Use "edit attributes" to set chromosome, start, end, and strand columns

---

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Maximum distance** is greatest distance in base pairs allowed between intervals that will be considered "clustered". **Negative** values for distance are allowed, and are useful for clustering intervals that overlap.
- **Minimum intervals per cluster** allow a threshold to be set on the minimum number of intervals to be considered a cluster. Any area with less than this minimum will not be included in the ouput.
- **Merge clusters into single intervals** outputs intervals that span the entire cluster.
- **Find cluster intervals; preserve comments and order** filters out non-cluster intervals while maintaining the original ordering and comments in the file.
- **Find cluster intervals; output grouped by clusters** filters out non-cluster intervals, but outputs the cluster intervals so that they are grouped together. Comments and original ordering in the file are lost.

Example

cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
                   -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <param format="interval" name="input1" type="data">
      <label>Cluster intervals of</label>
    </param>
    <param name="distance" size="5" type="integer" value="1" help="(bp
      <label>max distance between intervals</label>
    </param>
    <param name="minregions" size="5" type="integer" value="2">
      <label>min number of intervals per cluster</label>
    </param>
    <param name="returntype" type="select" label="Return type">
      <option value="1">Merge clusters into single intervals</option>
      <option value="2">Find cluster intervals; preserve comments and
      <option value="3">Find cluster intervals; output grouped by clus
      <option value="4">Find the smallest interval in each cluster</op
      <option value="5">Find the largest interval in each cluster</opt
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does not appear in the pulldown menu -> it is n

-----

**Screencasts!**

See Galaxy Interval Operation Screencasts_ (right click to open this l

.. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc

-----

**Syntax**

- **Maximum distance** is greatest distance in base pairs allowed betw
- **Minimum intervals per cluster** allow a threshold to be set on the
- **Merge clusters into single intervals** outputs intervals that span
- **Find cluster intervals; preserve comments and order** filters out
- **Find cluster intervals; output grouped by clusters** filters out n
```

Line: 87 Column: 8 XML Soft Tabs: 2

Develop: tools

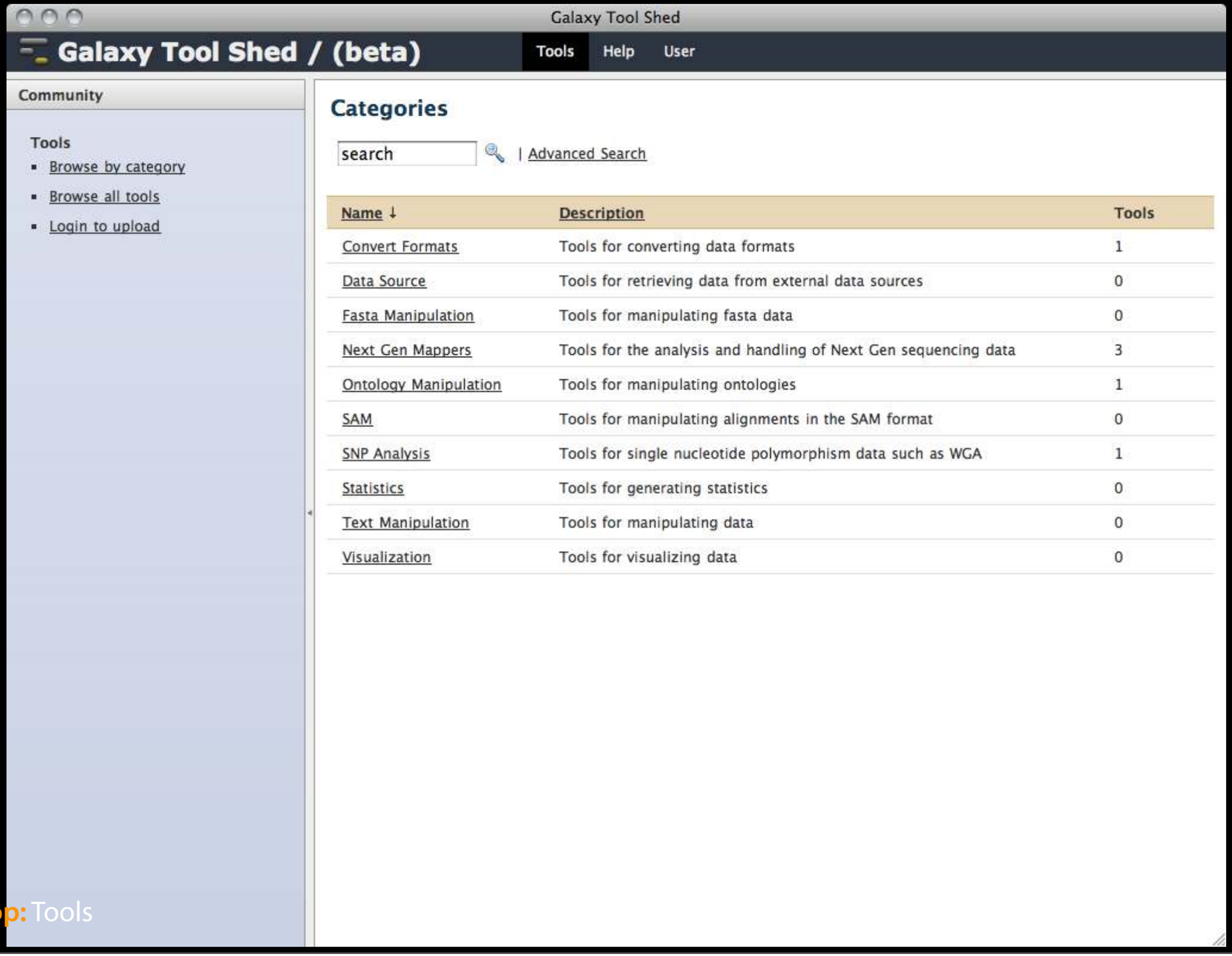
cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startCol,$input1_endCol
                   -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <param format="interval" name="input1" type="data">
      <label>Cluster intervals of</label>
    </param>
    <param name="distance" size="5" type="integer" value="1" help="(bp)">
      <label>max distance between intervals</label>
    </param>
    <param name="minregions" size="5" type="integer" value="2">
      <label>min number of intervals per cluster</label>
    </param>
    <param name="returntype" type="select" label="Return type">
      <option value="1">Merge clusters into single intervals</option>
      <option value="2">Find cluster intervals; preserve comments and order</option>
      <option value="3">Find cluster intervals; output grouped by clusters</option>
      <option value="4">Find the smallest interval in each cluster</option>
      <option value="5">Find the largest interval in each cluster</option>
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does not appear in the pulldown menu -> it is not in interval fo

-----

**Screencasts!**

See Galaxy Interval Operation Screencasts_ (right click to open this link in another wi

.. _Screencasts: http://www.bx.psu.edu/cgi-bin/trac.cgi/wiki/GopsDesc

-----

**Syntax**

- **Maximum distance** is greatest distance in base pairs allowed between intervals tha
- **Minimum intervals per cluster** allow a threshold to be set on the minimum number o
- **Merge clusters into single intervals** outputs intervals that span the entire clust
- **Find cluster intervals; preserve comments and order** filters out non-cluster inter
- **Find cluster intervals; output grouped by clusters** filters out non-cluster interv
```

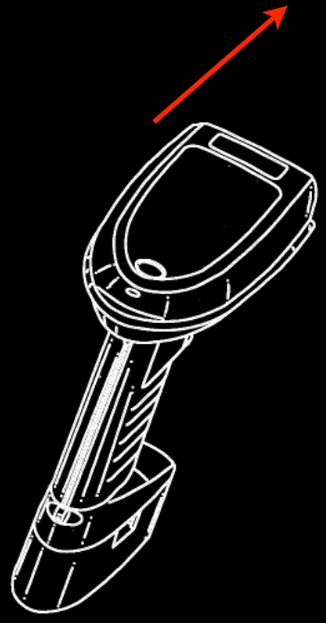
Column: 45 XML Soft Tabs: 2

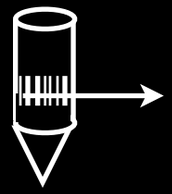
} Template for generating command line from parameter values

Develop Tools



Develop: Tools

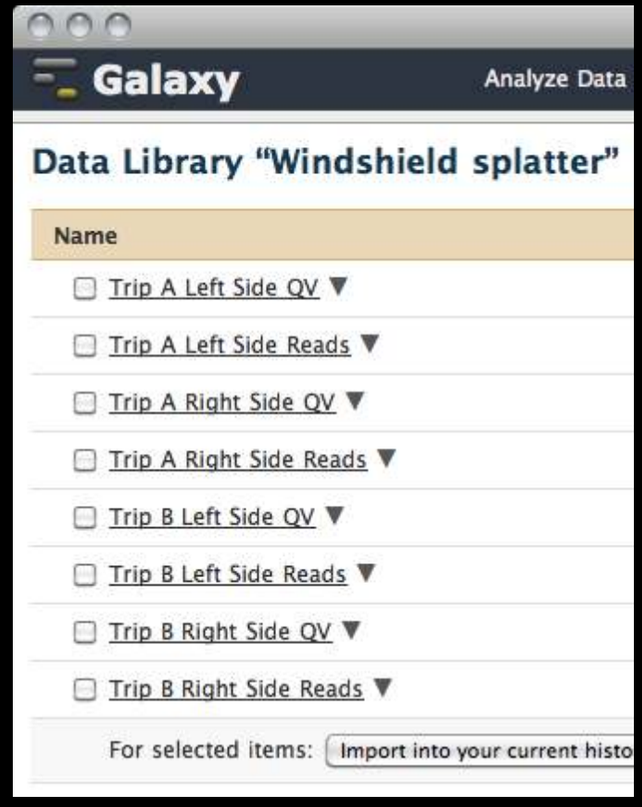




PACIFIC BIOSCIENCES™

illumina®

ion torrent
by life technologies™



Lab: Sequencing Request Tracking

Run Galaxy at your site

- Galaxy is designed for local installation and customization
  - Easily integrate new tools
  - Easy to deploy and manage on nearly any (Unix-like) system

Lab: Local Resources

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

Lab: Local Resources