Accessible, transparent, reproducible analysis with **Galaxy**

Try it now:
http://usegalaxy.org

Develop and deploy:
http://getgalaxy.org

Try it now:
http://usegalaxy.org

Develop and deploy:
http://getgalaxy.org

EMORY

Enis Afgan

Dannon Baker

Jeremy Goecks

Dave Clements

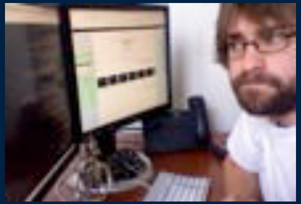
Kanwei Li

James Taylor

PENNSTATE.

1855

Dan Blankenberg


Nate Coraor

Jennifer Jackson

Greg von Kuster

Kelly Vincent

Anton Nekrutenko

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

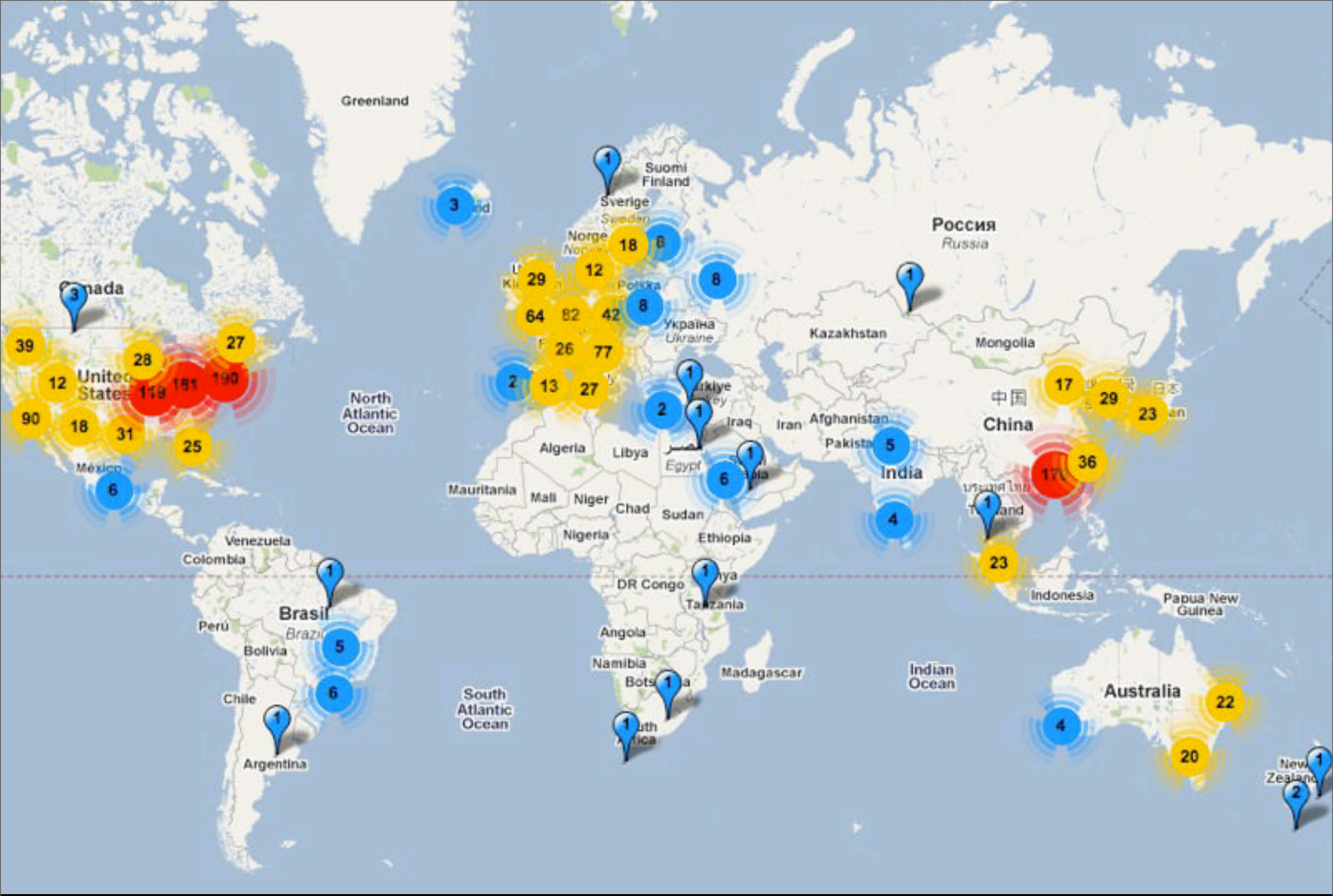
Biology *has been* rapidly transformed into a data intensive science

Illumina Hi-Seq: ~25-50 GB per day, $16k-$20k per run
Greater than 1Mb per dollar
With multiplexing, as little as $100 per sample.

454 GS / Junior: 40-400Mb runs, but read lengths pushing 1kb

Ion Torrent PGM: 10Mb-1Gb runs, 200-400bp reads, 2 hour runtime, $500!

PacBio RS: Direct single molecule sequencing, only 35k reads, but long read lengths, 30 minute runs!



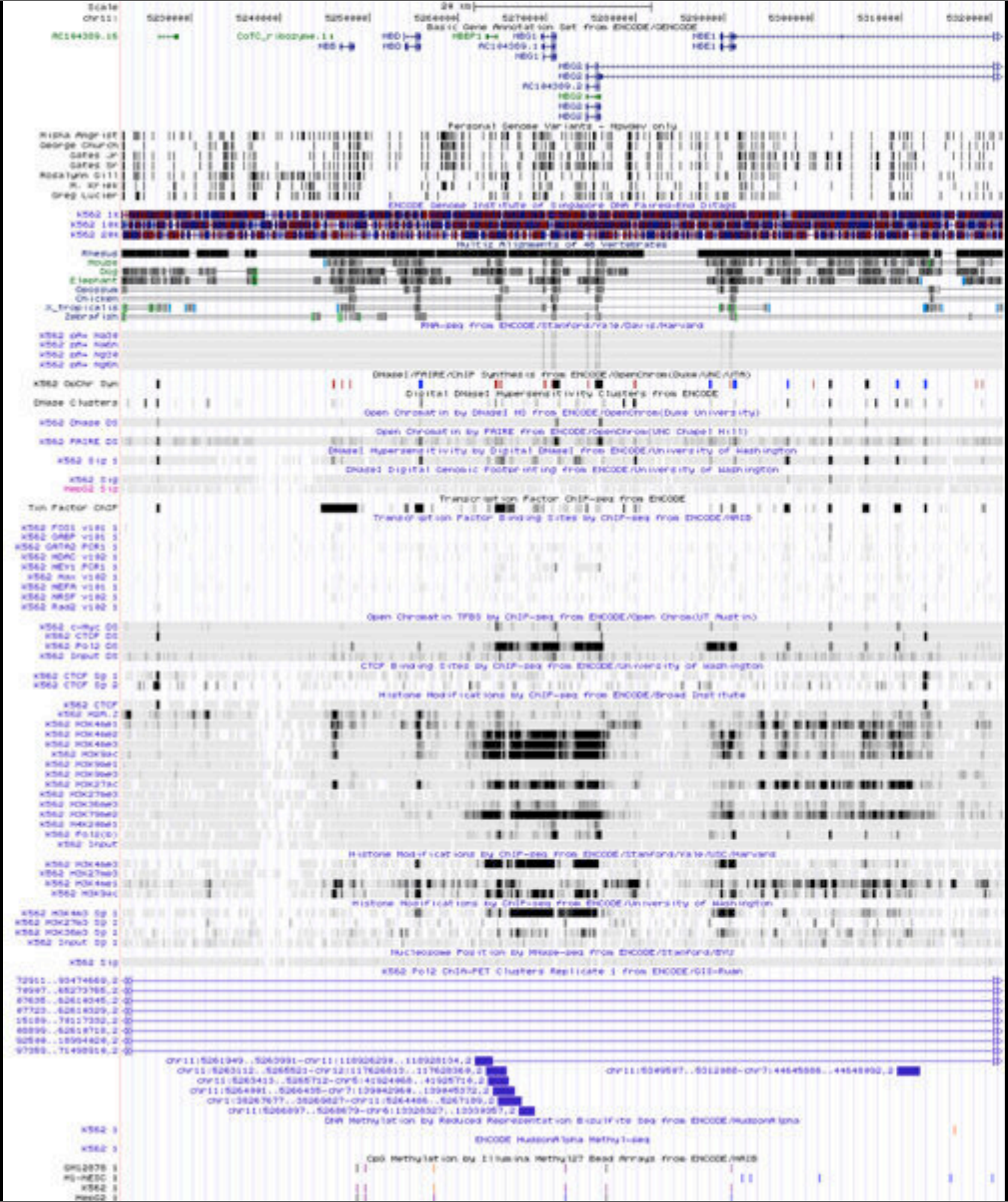
(http://pathogenomics.bham.ac.uk/hts/)

We can turn many **functional annotation** problems into **sequencing** problems

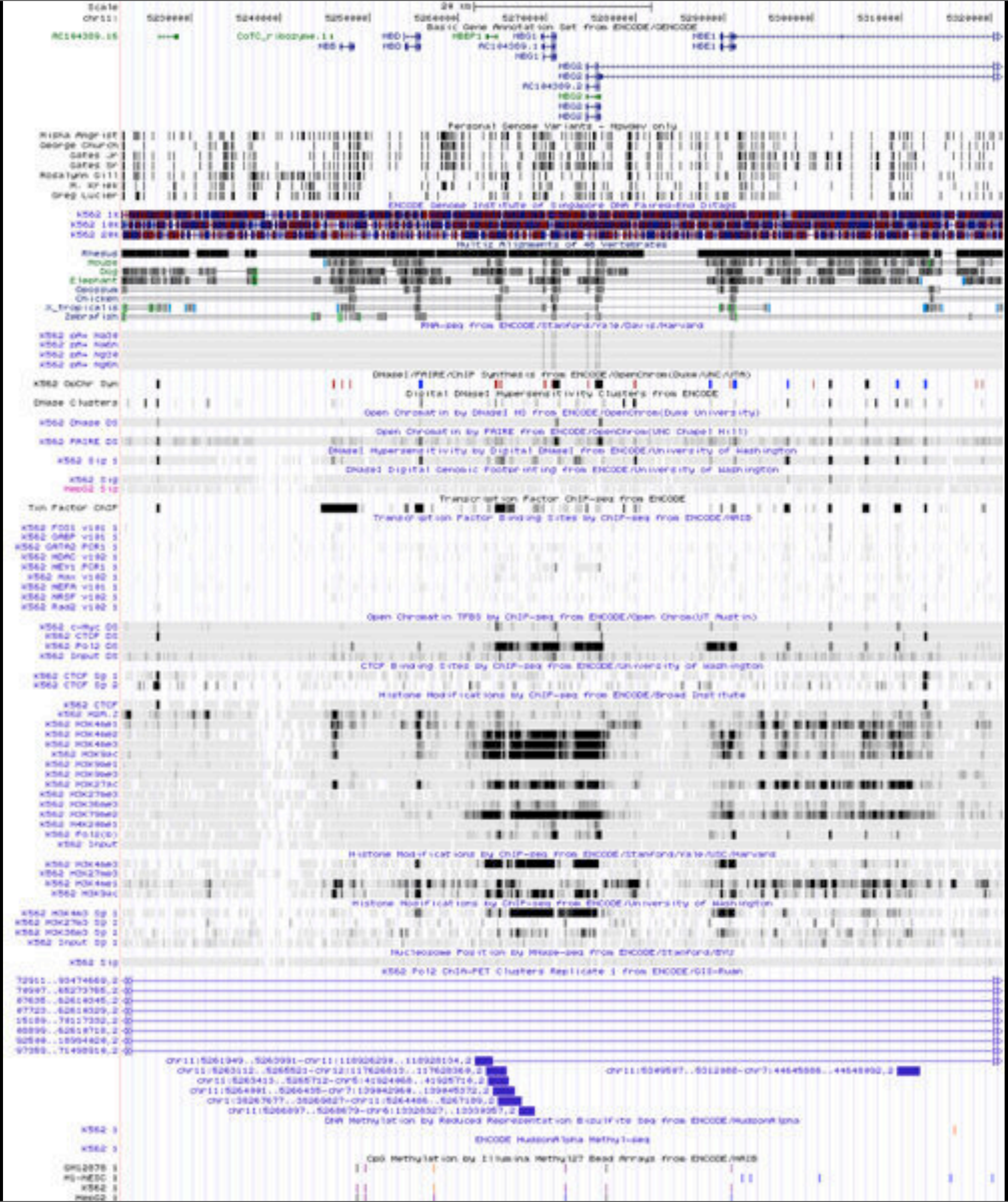
An individual genome is relatively static

Transcript levels, epigenomic modifications, and chromatin structure vary based on cell type, time, condition, ...

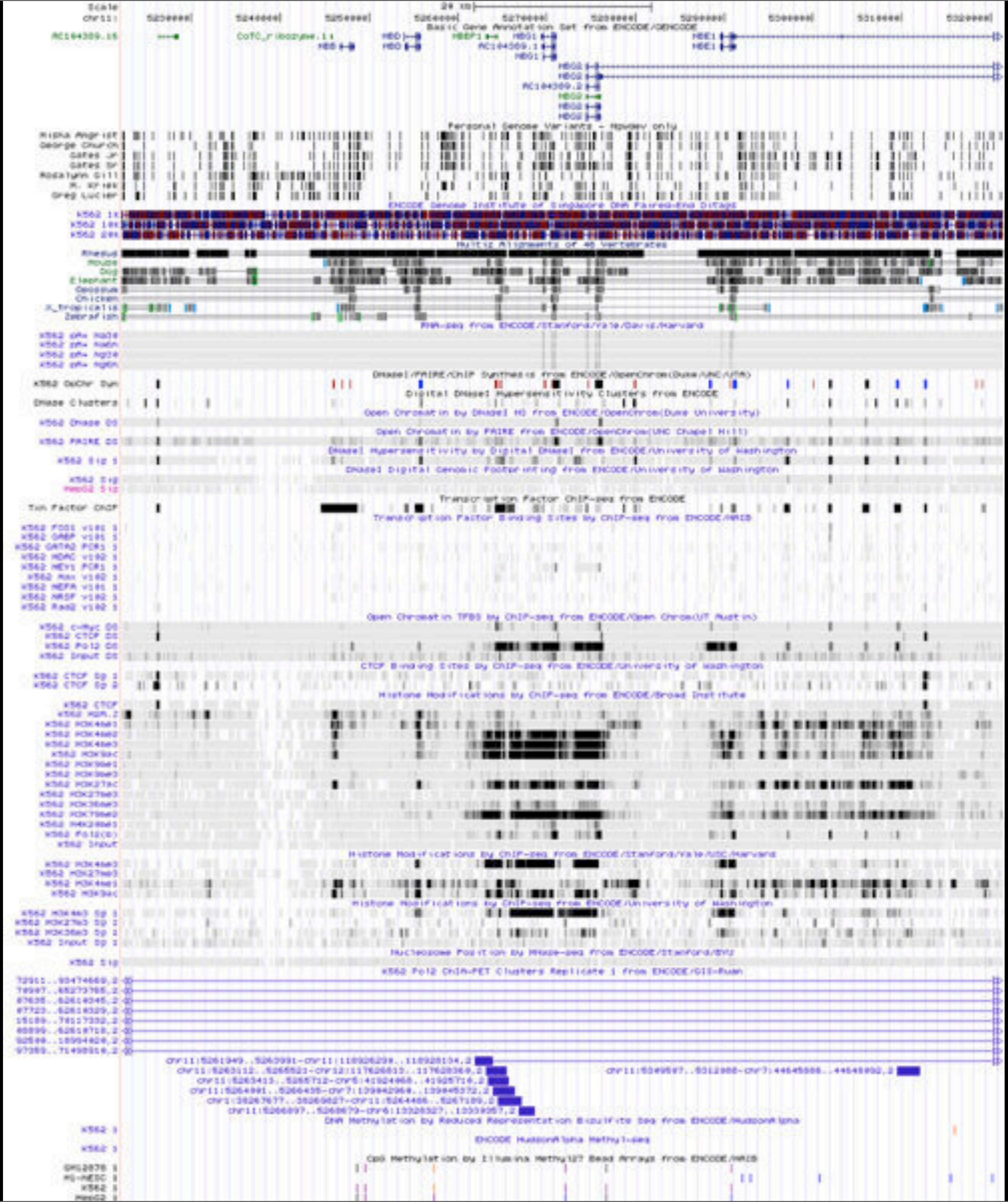
Enormous potential for data generation


Scale
chr11:
Basic Gene Annotation Set from ENCODE/GENCODE
HBB
HBD
HBG1
HBG2
HBE1
Personal Genome Variants - Hgwdev only
Misha Angrist
George Church
Gates Jr
Gates Sr
Rosalynn Gill
Greg Lucier
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
Digital DNaseI Hypersensitivity Clusters from ENCODE
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
Transcription Factor ChIP-seq from ENCODE
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
Histone Modifications by ChIP-seq from ENCODE/University of Washington
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
ENCODE HudsonAlpha Methyl-seq
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

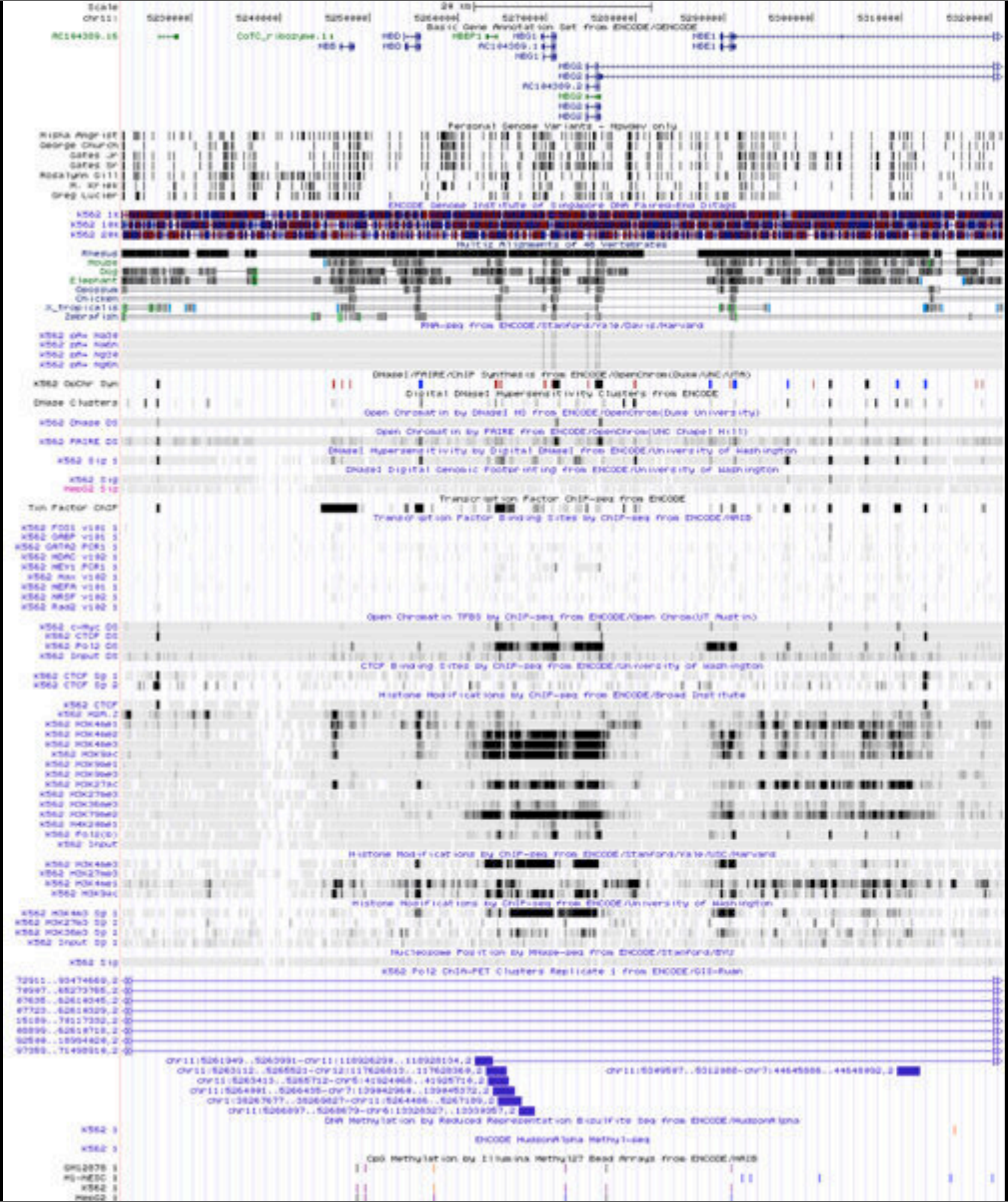
Resequencing


Scale
chr11:
20 kb
Basic Gene Annotation Set from ENCODE/GENCODE
AC104389.15
HBB
HBD
HBBP1
HBG1
HBG2
HBE1
Personal Genome Variants - Hgwdev only
Alpha Angrist
George Church
Gates Jr
Gates Sr
Rosalynn Gill
M. Kriek
Greg Lucier
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
K562 1K
K562 10K
K562 20K
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
Histone Modifications by ChIP-seq from ENCODE/University of Washington
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
ENCODE HudsonAlpha Methyl-seq
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

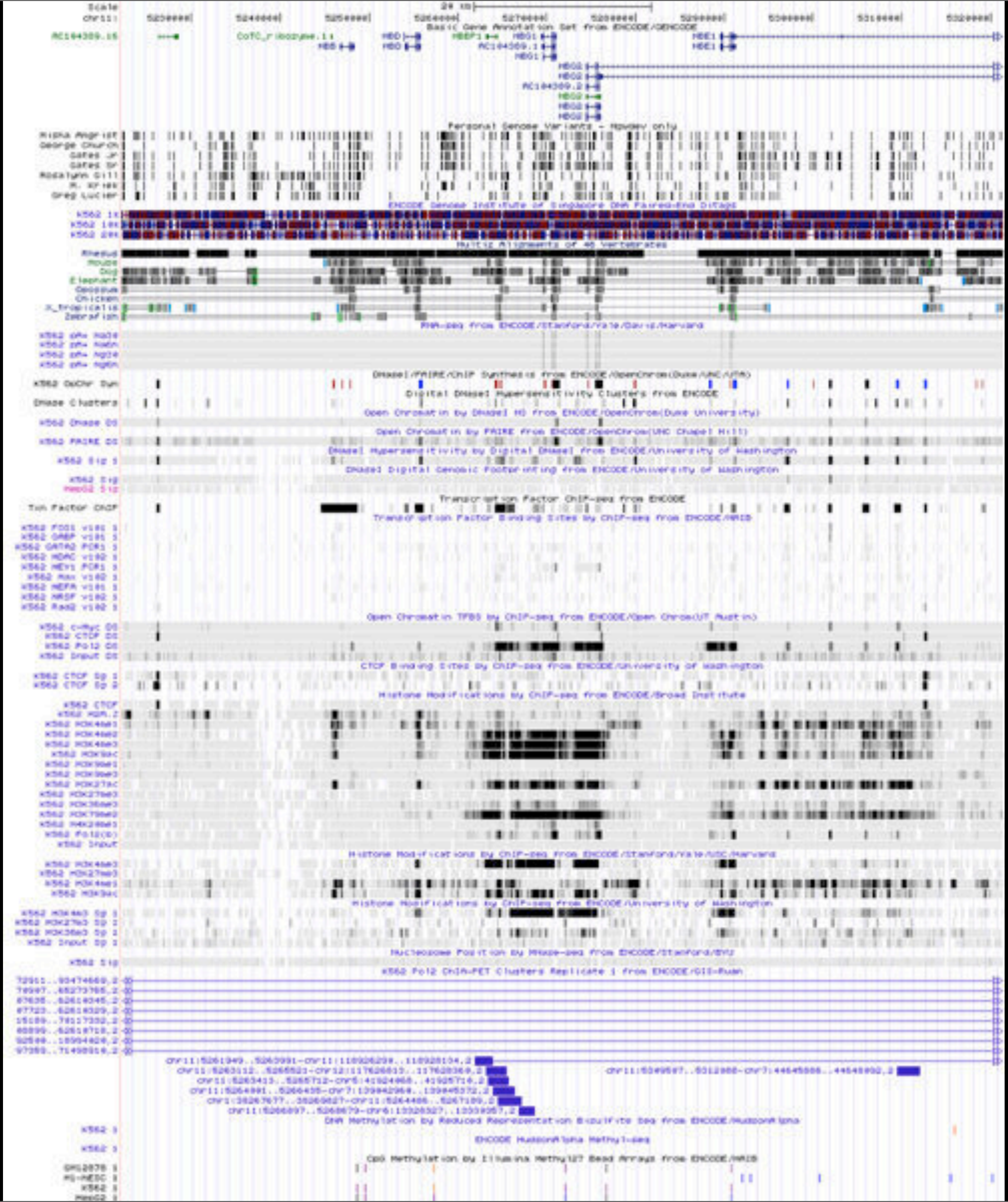
} De-novo genome sequening


Scale
chr11:
Basic Gene Annotation Set from ENCODE/GENCODE
HBB
HBD
HBG1
HBG2
HBE1
Personal Genome Variants - Hgvdev only
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
Digital DNaseI Hypersensitivity Clusters from ENCODE
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
Transcription Factor ChIP-seq from ENCODE
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
Histone Modifications by ChIP-seq from ENCODE/University of Washington
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
ENCODE HudsonAlpha Methyl-seq
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

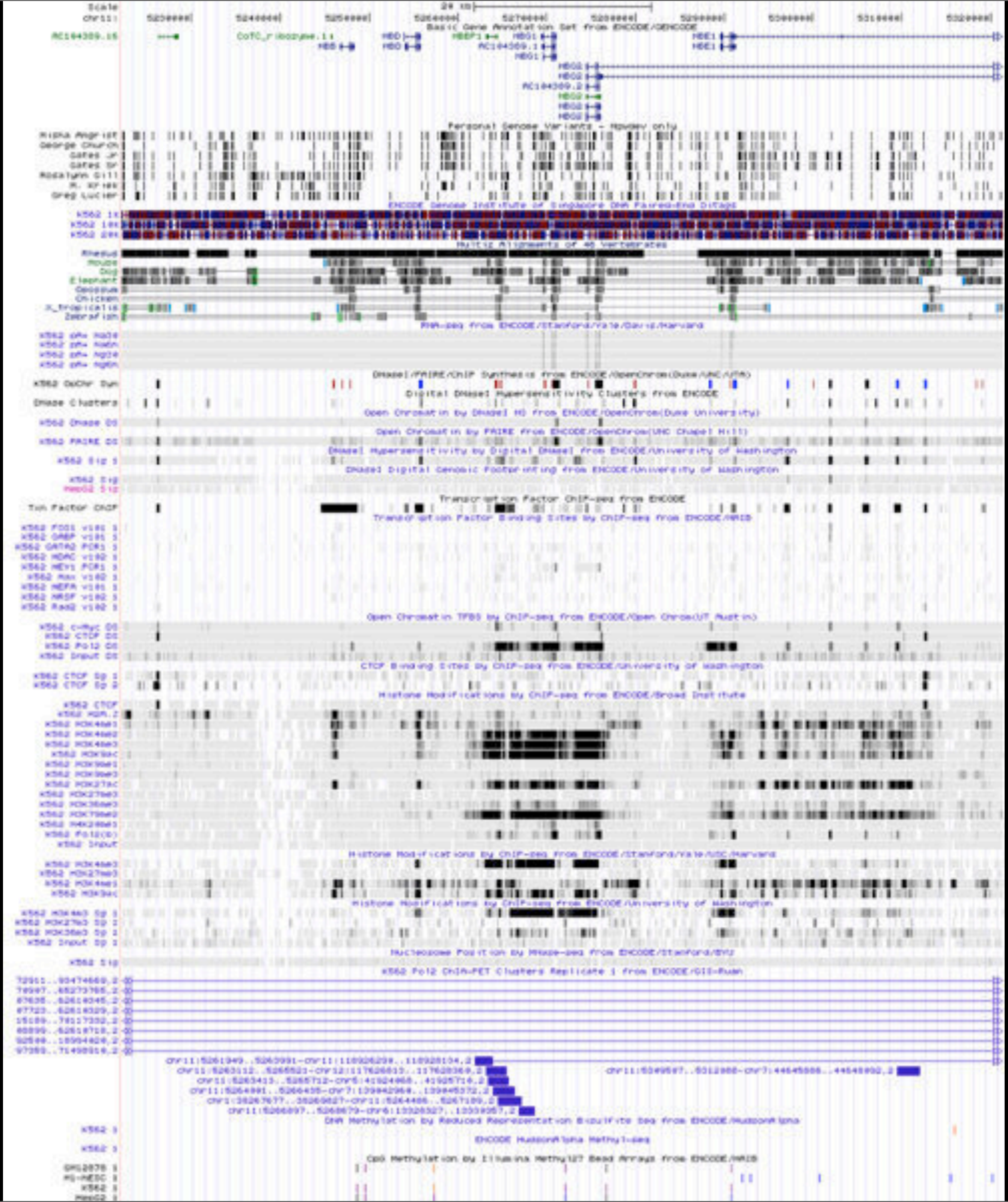
} Direct RNA sequencing


Scale
chr11:
Basic Gene Annotation Set from ENCODE/GENCODE
HBD
HBB
HBBP1
HBG1
HBG2
HBE1
Personal Genome Variants - HgVar only
Alpha Angrist
George Church
Gates Jr
Gates Sr
Rosalynn Gill
M. Kriek
Greg Lucier
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
K562
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
K562 OpChr Syn
Digital DNaseI Hypersensitivity Clusters from ENCODE
DNase Clusters
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
K562 DNase DS
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
K562 FAIRE DS
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
K562 Sig
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
K562 Sig
HepG2 Sig
Transcription Factor ChIP-seq from ENCODE
Txn Factor ChIP
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
K562 c-Myc DS
K562 CTCF DS
K562 Pol2 DS
K562 Input DS
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
K562 CTCF Sp 1
K562 CTCF Sp 2
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
K562 CTCF
K562 H2A.Z
K562 H3K4me1
K562 H3K4me2
K562 H3K4me3
K562 H3K9ac
K562 H3K9me1
K562 H3K9me3
K562 H3K27ac
K562 H3K27me3
K562 H3K36me3
K562 H3K79me2
K562 H4K20me1
K562 Pol2(b)
K562 Input
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
K562 H3K4me3
K562 H3K27me3
K562 H3K4me1
K562 H3K9ac
Histone Modifications by ChIP-seq from ENCODE/University of Washington
K562 H3K4me3 Sp 1
K562 H3K27me3 Sp 1
K562 H3K36me3 Sp 1
K562 Input Sp 1
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Sig
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
K562
ENCODE HudsonAlpha Methyl-seq
K562
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB
GM12878
HG-hESC
K562
HepG2

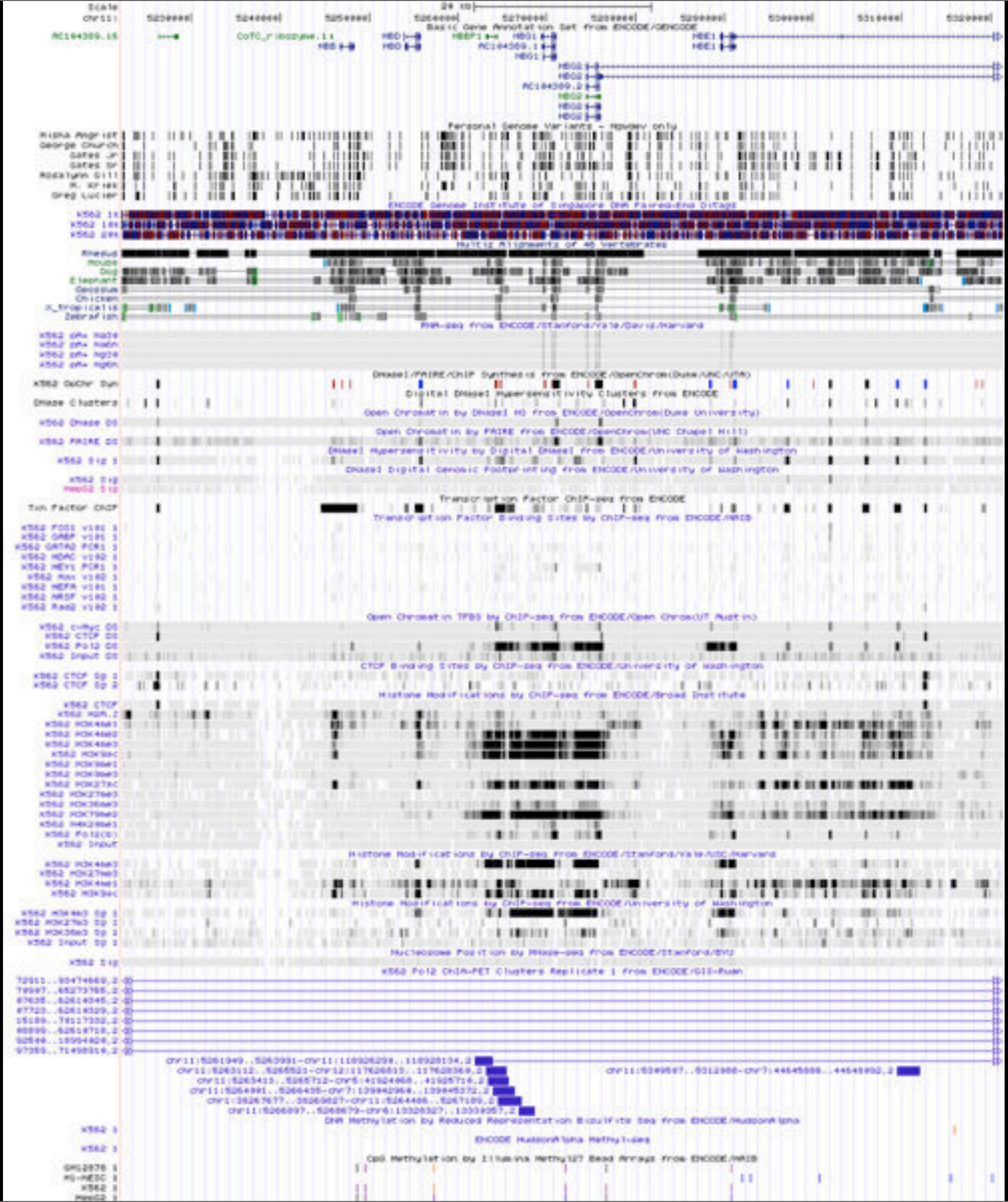
} Open Chromatin assays (DNase, FAIRE)


Scale
chr11:
Basic Gene Annotation Set from ENCODE/GENCODE
HBB
HBD
HBBP1
HBG1
HBG2
HBE1
AC104389.1
AC104389.2
Personal Genome Variants
Misha Angrist
George Church
Gates Jr
Gates Sr
Rosalynn Gill
Greg Lucier
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
Digital DNaseI Hypersensitivity Clusters from ENCODE
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
Transcription Factor ChIP-seq from ENCODE
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
Histone Modifications by ChIP-seq from ENCODE/University of Washington
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
ENCODE HudsonAlpha Methyl-seq
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

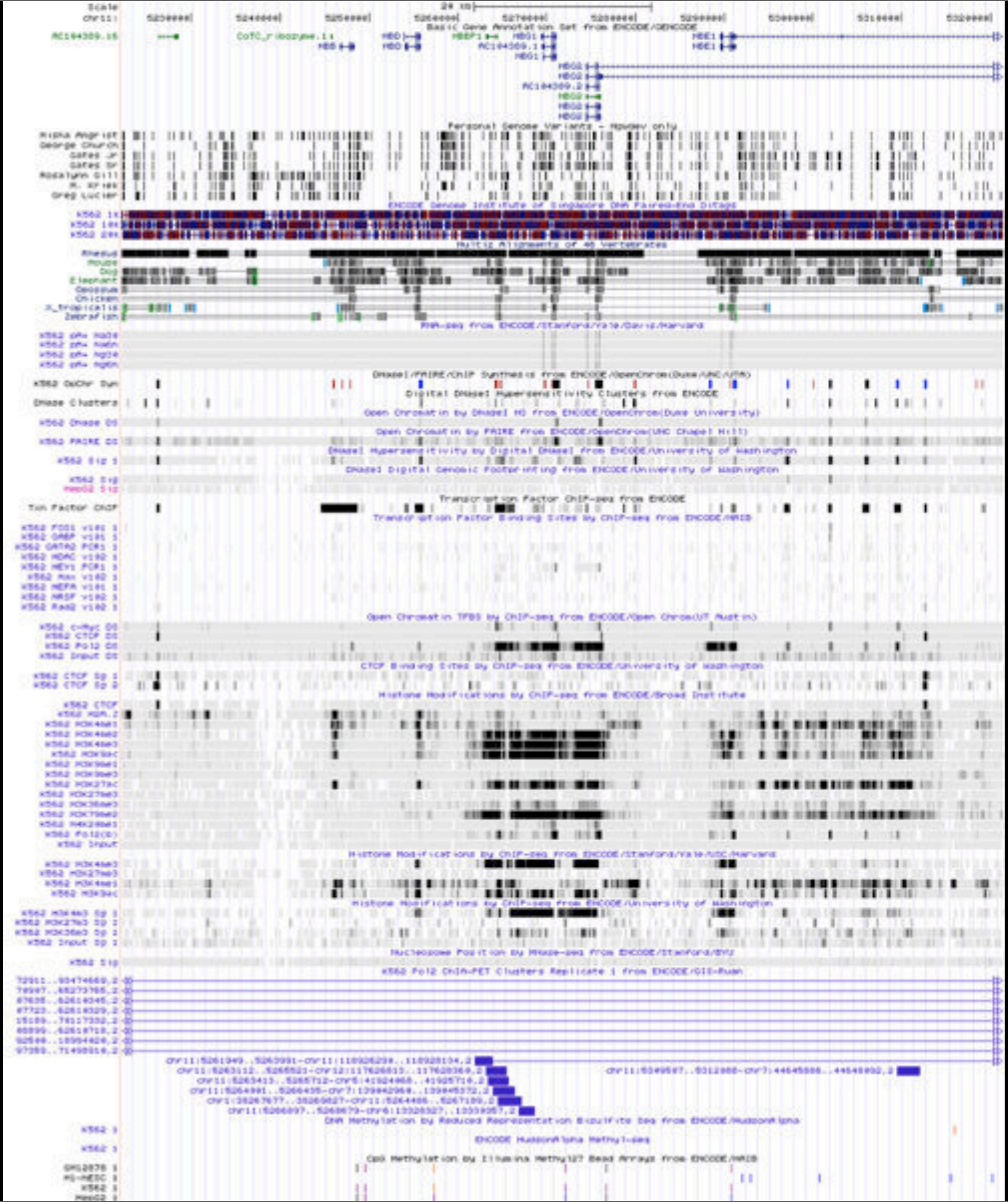
} Transcription factors (ChIP-seq)


Scale
chr11:
Basic Gene Annotation Set from ENCODE/GENCODE
HBB
HBD
HBG1
HBG2
HBE1
Personal Genome Variants
ENCODE Genome Institute of Singapore DNA Paired-End Ditags
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)
Digital DNaseI Hypersensitivity Clusters from ENCODE
Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)
Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)
DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington
DNaseI Digital Genomic Footprinting from ENCODE/University of Washington
Transcription Factor ChIP-seq from ENCODE
Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB
Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)
CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard
Histone Modifications by ChIP-seq from ENCODE/University of Washington
Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan
DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
ENCODE HudsonAlpha Methyl-seq
CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

} Histones variants
(ChIP-seq, MNase-seq)


Scale
chr11:
Basic Gene Annotation Set from ENCODE/GENCODE
HBB
HBD
HBG1
HBG2
HBE1
Personal Genome Variants - Hgwdev only
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard
Digital DNaseI Hypersensitivity Clusters from ENCODE
Transcription Factor ChIP-seq from ENCODE
Histone Modifications by ChIP-seq from ENCODE/Broad Institute
K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan

} Long range interactions (5C, Hi-C, ChIA-PET



Scale 20 kb
chr11:

Basic Gene Annotation Set from ENCODE/GENCODE

Personal Genome Variants - Hgvdev only

ENCODE Genome Institute of Singapore DNA Paired-End Ditags

Multiz Alignments of 46 Vertebrates

RNA-seq from ENCODE/Stanford/Yale/Davis/Harvard

DNaseI/FAIRE/ChIP Synthesis from ENCODE/OpenChrom(Duke/UNC/UTA)

Digital DNaseI Hypersensitivity Clusters from ENCODE

Open Chromatin by DNaseI HS from ENCODE/OpenChrom(Duke University)

Open Chromatin by FAIRE from ENCODE/OpenChrom(UNC Chapel Hill)

DNaseI Hypersensitivity by Digital DNaseI from ENCODE/University of Washington

DNaseI Digital Genomic Footprinting from ENCODE/University of Washington

Transcription Factor ChIP-seq from ENCODE

Transcription Factor Binding Sites by ChIP-seq from ENCODE/HAIB

Open Chromatin TFBS by ChIP-seq from ENCODE/Open Chrom(UT Austin)

CTCF Binding Sites by ChIP-seq from ENCODE/University of Washington

Histone Modifications by ChIP-seq from ENCODE/Broad Institute

Histone Modifications by ChIP-seq from ENCODE/Stanford/Yale/USC/Harvard

Histone Modifications by ChIP-seq from ENCODE/University of Washington

Nucleosome Position by MNase-seq from ENCODE/Stanford/BYU

K562 Pol2 ChIA-PET Clusters Replicate 1 from ENCODE/GIS-Ruan

DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha

ENCODE HudsonAlpha Methyl-seq

CpG Methylation by Illumina Methyl27 Bead Arrays from ENCODE/HAIB

} Methylation
(Bisulfite-seq)

Investigators across nearly all areas of biology can take advantage of these techniques

Investigator driven data production replacing large community data production projects

This "**democratization of sequencing**" has not yet been matched by democratization of analysis infrastructure, burden is largely on the investigator

However, making sense of this data *requires* sophisticated methods

How can these methods be made **accessible** to scientists?

How do we facilitate **transparent** communication of analyses?

How do we ensure that analyses are **reproducible**?

A crisis in genomics research: **reproducibility**

Microarray Experiment Reproducibility

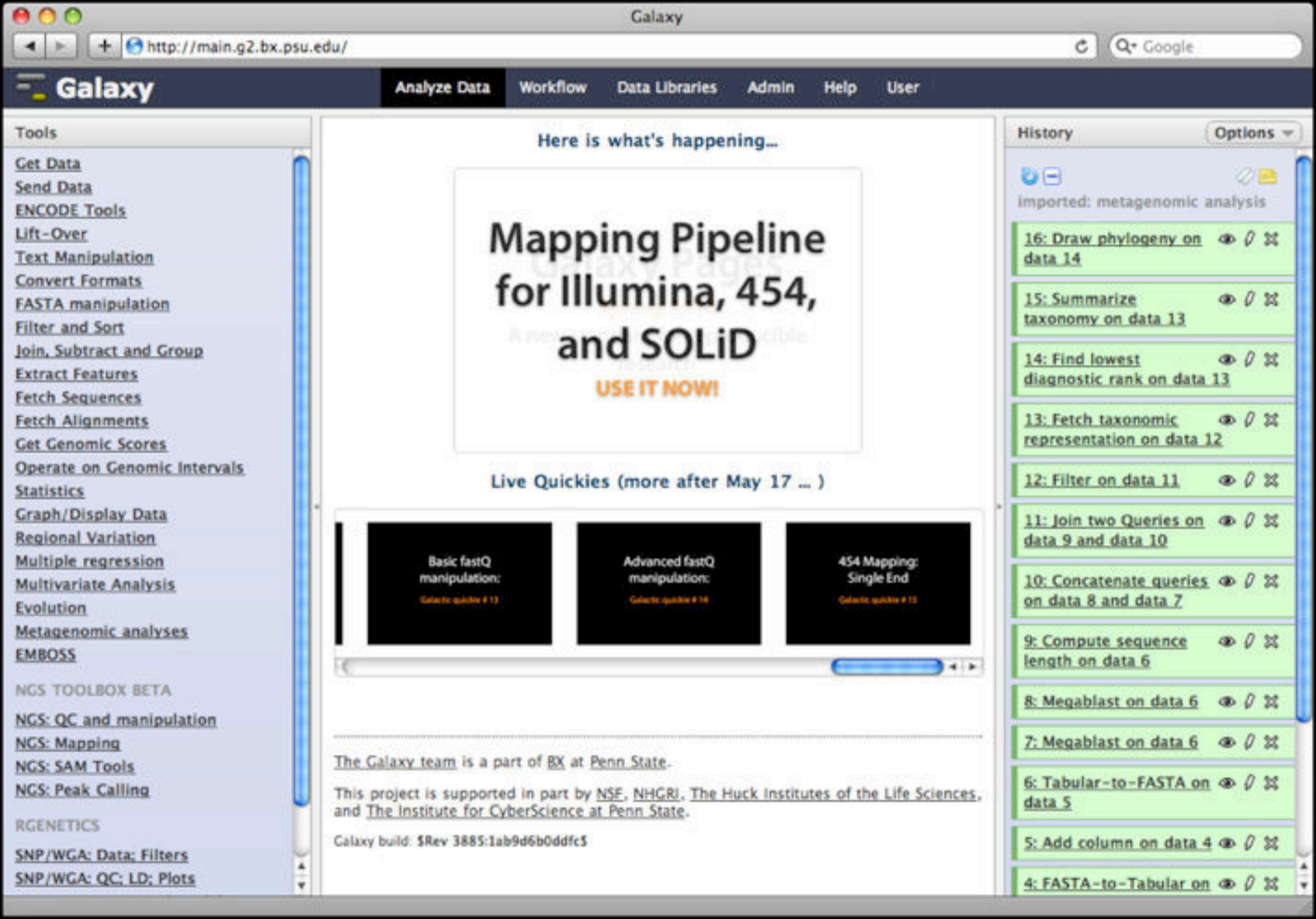
- 18 Nat. Genetics microarray gene expression experiments
- **Less than 50% reproducible**
- Problems
    - missing data (38%)
    - missing software, hardware details (50%)
    - missing method, processing details (66%)

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



Here is what's happening...

Mapping Pipeline for Illumina, 454, and SOLiD

USE IT NOW!

Live Quickies (more after May 17 ...)

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: $Rev 3885:1ab9d6b0ddfc$

imported: metagenomic analysis

16: Draw phylogeny on data 14

15: Summarize taxonomy on data 13

14: Find lowest diagnostic rank on data 13

13: Fetch taxonomic representation on data 12

12: Filter on data 11

11: Join two Queries on data 9 and data 10

10: Concatenate queries on data 8 and data 7

9: Compute sequence length on data 6

8: Megablast on data 6

7: Megablast on data 6

6: Tabular-to-FASTA on data 5

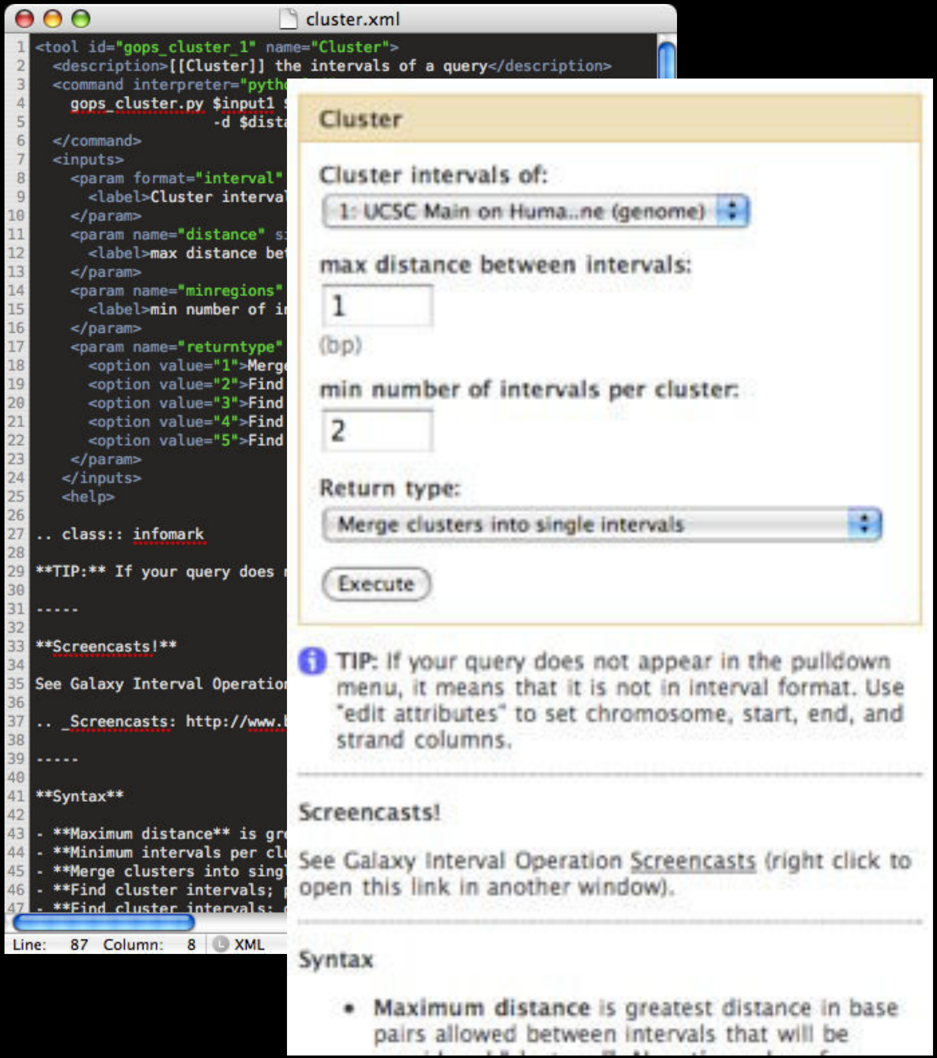
5: Add column on data 4

4: FASTA-to-Tabular on

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



cluster.xml

```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="pytho
    gops_cluster.py $input1 $
                    -d $dista
  </command>
  <inputs>
    <param format="interval"
      <label>Cluster interval
    </param>
    <param name="distance" si
      <label>max distance bet
    </param>
    <param name="minregions"
      <label>min number of in
    </param>
    <param name="returntype"
      <option value="1">Merge
      <option value="2">Find
      <option value="3">Find
      <option value="4">Find
      <option value="5">Find
    </param>
   </inputs>
   <help>

.. class:: infomark

**TIP:** If your query does n

-----

**Screencasts!**

See Galaxy Interval Operation

.. _Screencasts: http://www.b

-----

**Syntax**

- **Maximum distance** is gre
- **Minimum intervals per clu
- **Merge clusters into singl
- **Find cluster intervals; p
- **Find cluster intervals; c
```

Cluster

Cluster intervals of: 1: UCSC Main on Huma..ne (genome)

max distance between intervals: 1 (bp)

min number of intervals per cluster: 2

Return type: Merge clusters into single intervals

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

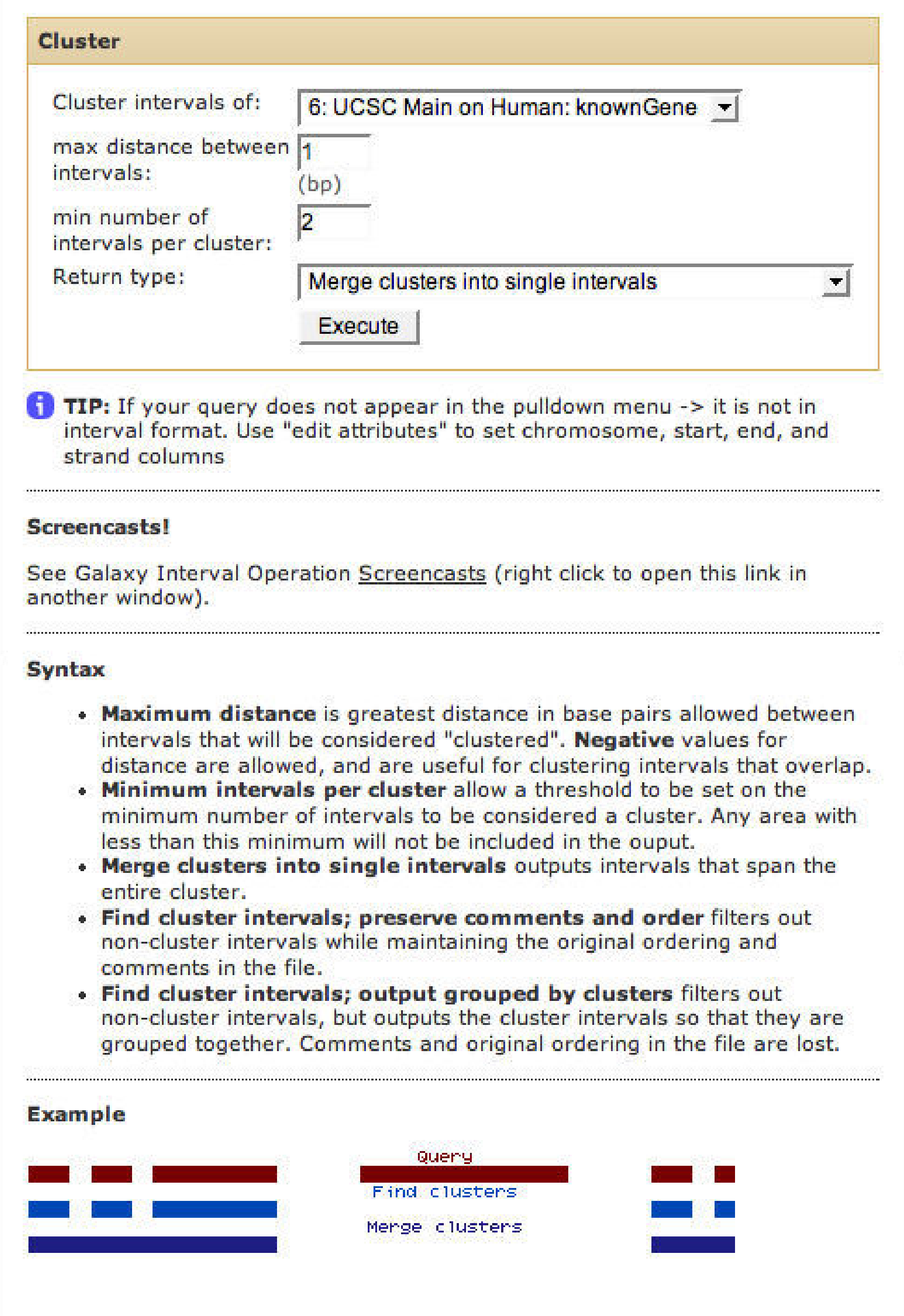
Screencasts!

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

- 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



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 ▾

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

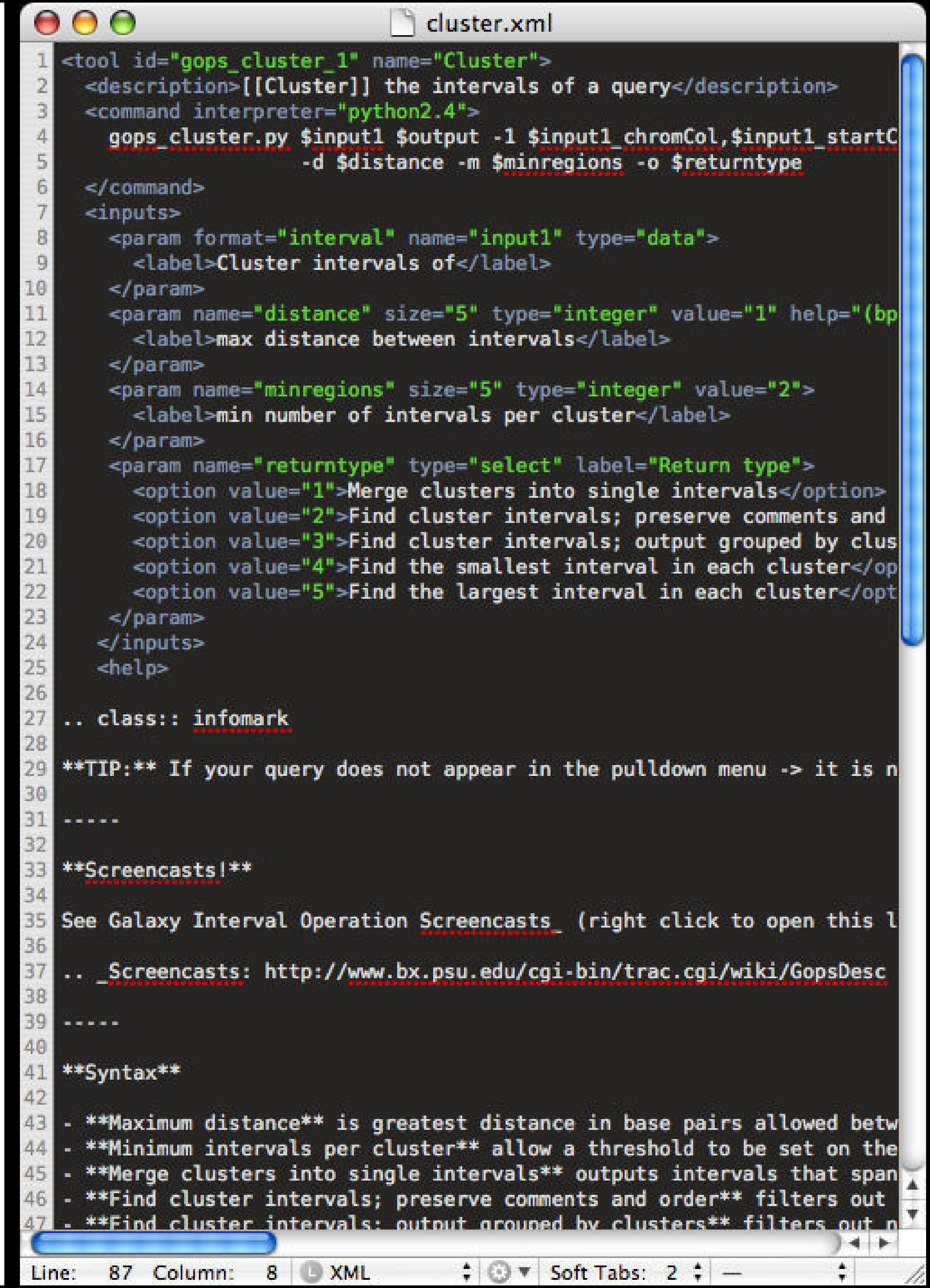
Screencasts!

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.

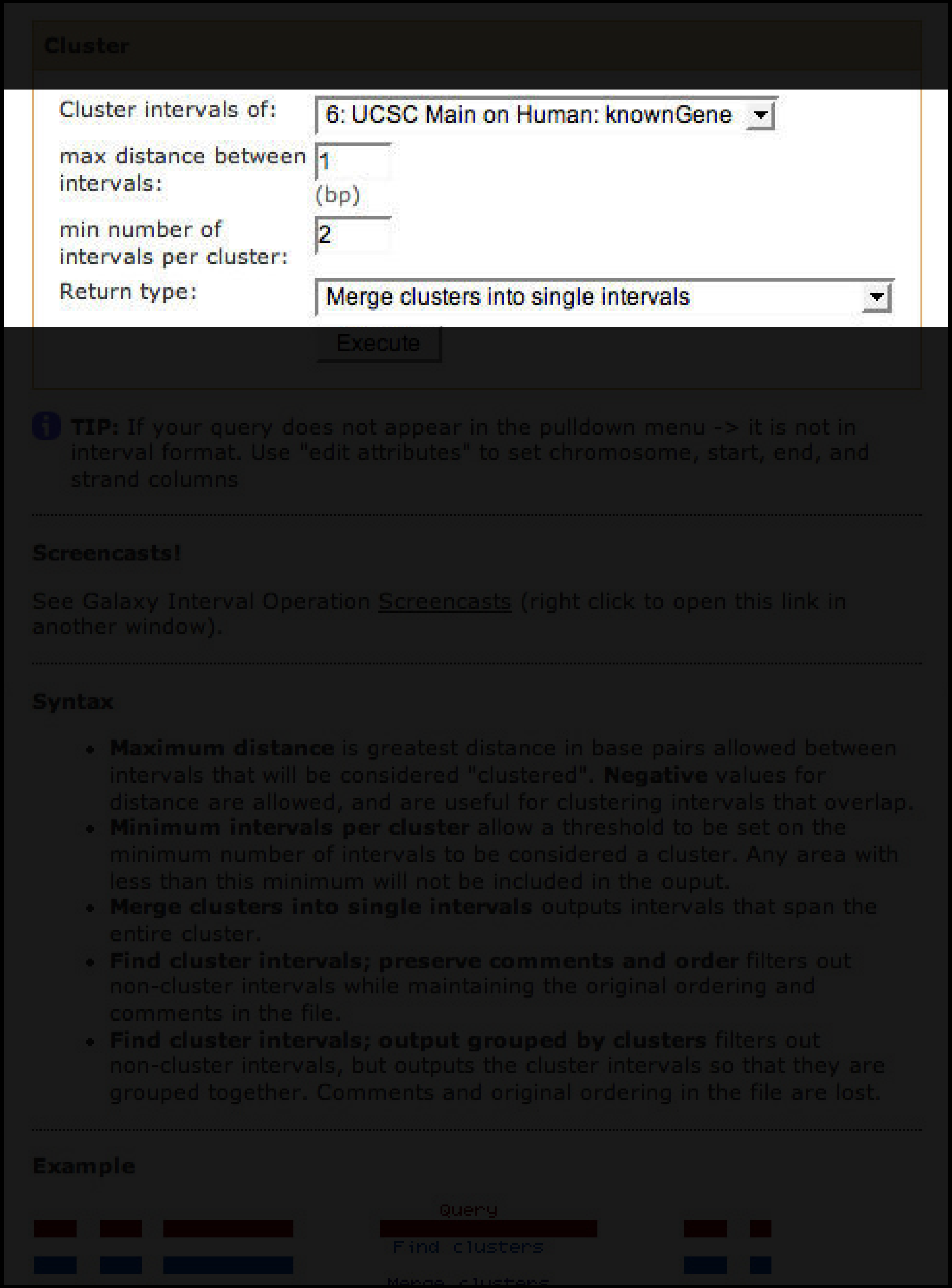
Example



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



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

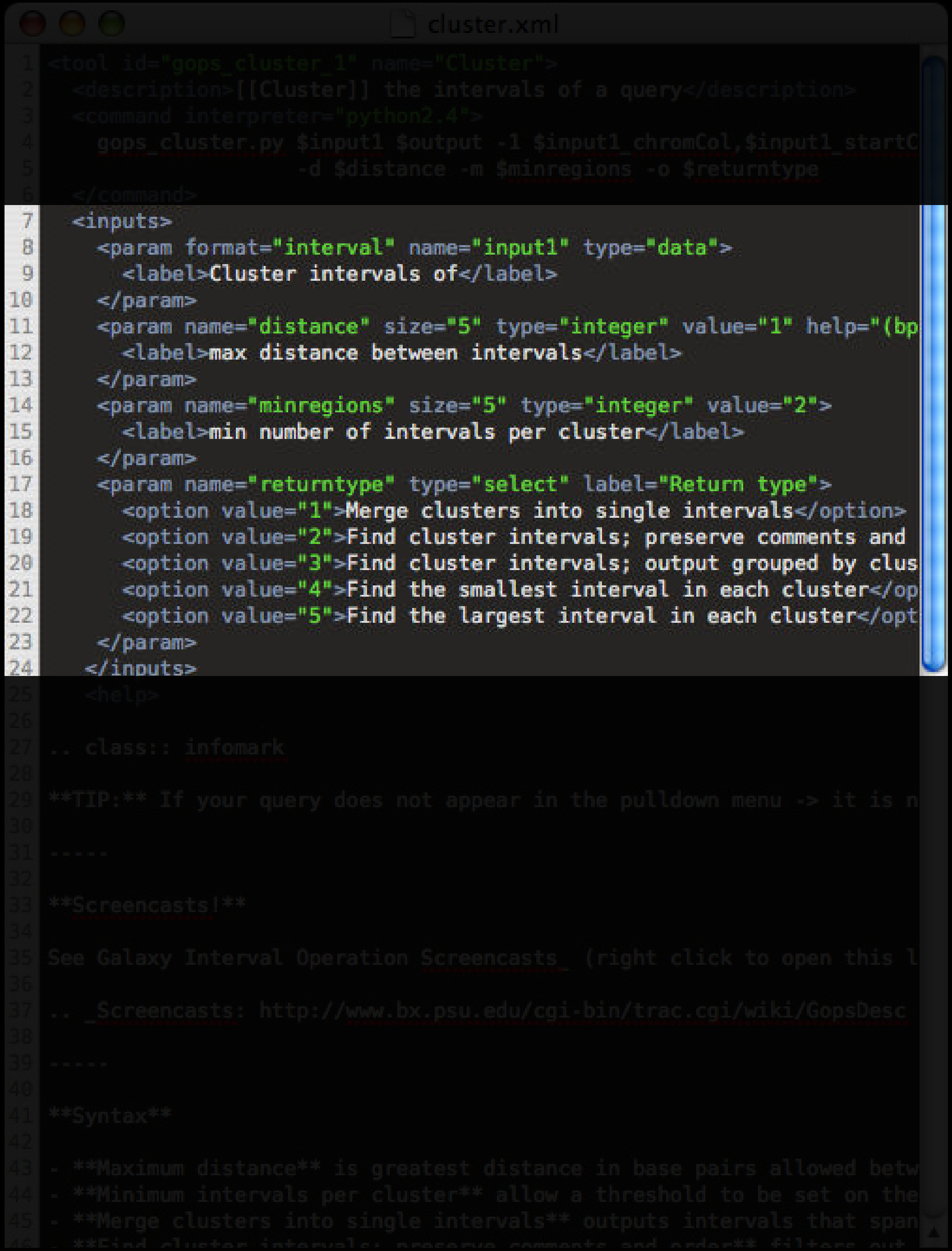
Screencasts!

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.

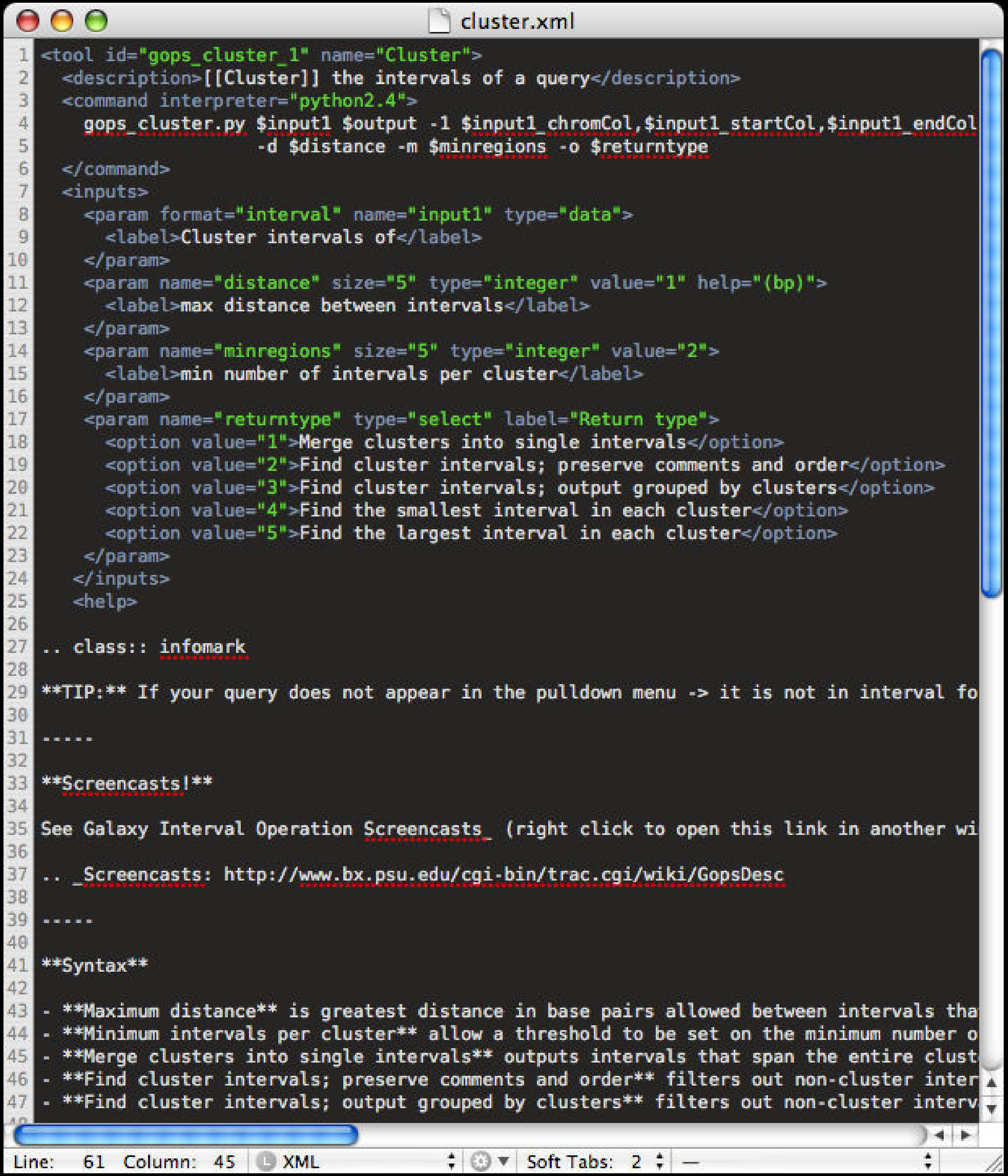
Example



cluster.xml

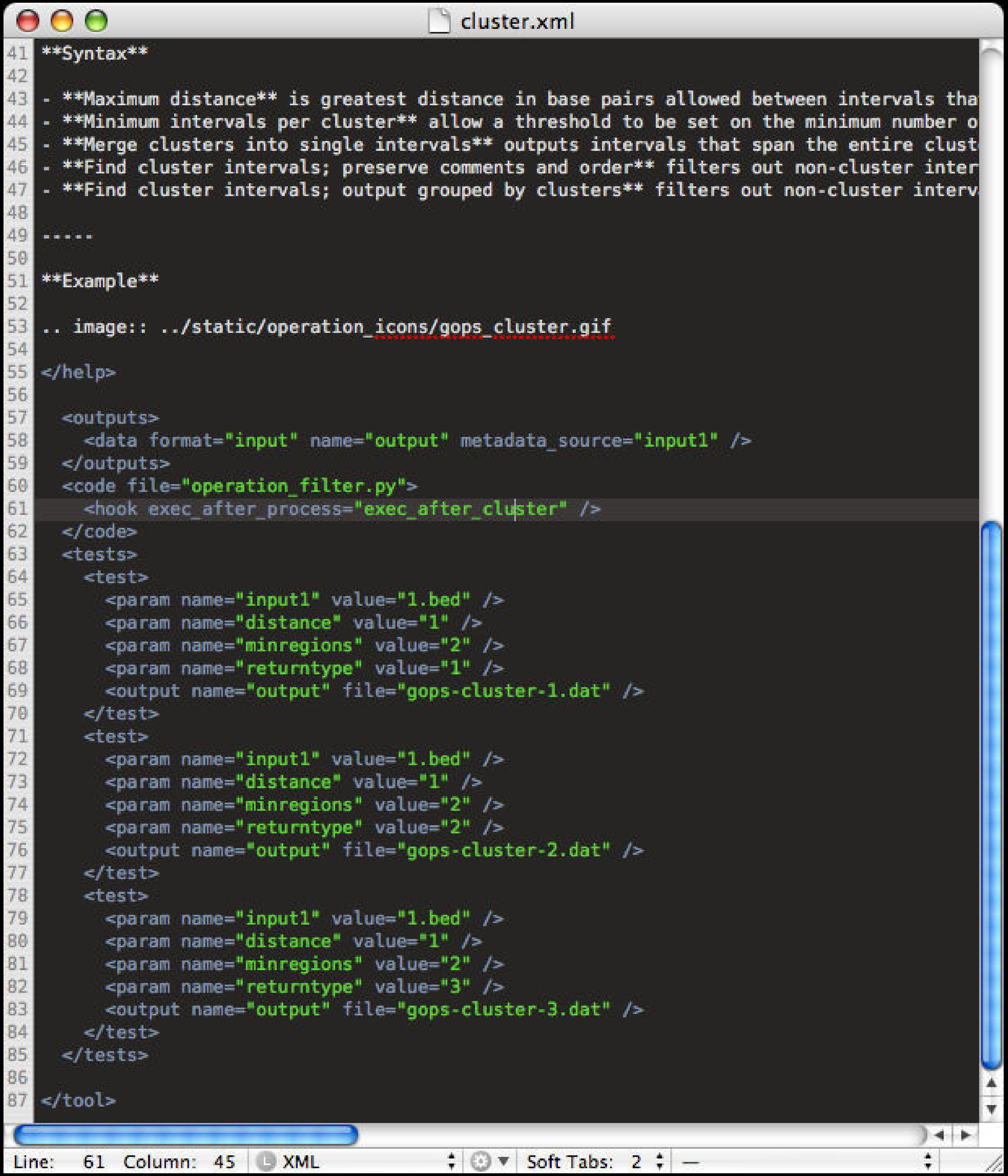
```
<tool id="gops_cluster_1" name="Cluster">
  <description>[[Cluster]] the intervals of a query</description>
  <command interpreter="python2.4">
    gops_cluster.py $input1 $output -1 $input1_chromCol,$input1_startC
                    -d $distance -m $minregions -o $returntype
  </command>
  <inputs>
    <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
```

HTML inputs generated from abstract parameter description



```
<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
```

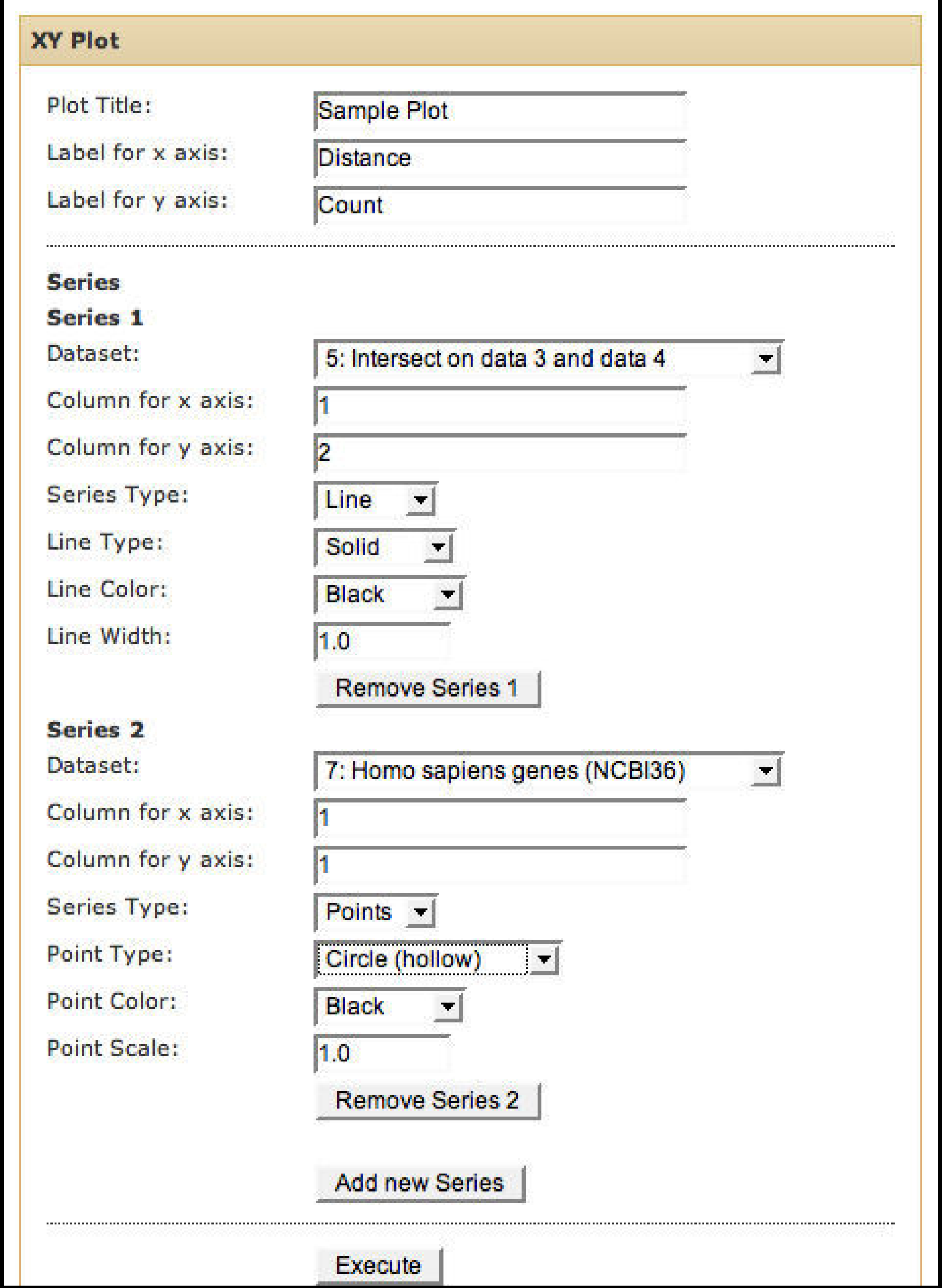
} Template for generating command line from parameter values



```
**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

-----

**Example**

.. image:: ../static/operation_icons/gops_cluster.gif

</help>

  <outputs>
    <data format="input" name="output" metadata_source="input1" />
  </outputs>
  <code file="operation_filter.py">
    <hook exec_after_process="exec_after_cluster" />
  </code>
  <tests>
    <test>
      <param name="input1" value="1.bed" />
      <param name="distance" value="1" />
      <param name="minregions" value="2" />
      <param name="returntype" value="1" />
      <output name="output" file="gops-cluster-1.dat" />
    </test>
    <test>
      <param name="input1" value="1.bed" />
      <param name="distance" value="1" />
      <param name="minregions" value="2" />
      <param name="returntype" value="2" />
      <output name="output" file="gops-cluster-2.dat" />
    </test>
    <test>
      <param name="input1" value="1.bed" />
      <param name="distance" value="1" />
      <param name="minregions" value="2" />
      <param name="returntype" value="3" />
      <output name="output" file="gops-cluster-3.dat" />
    </test>
  </tests>

</tool>
```

Line: 61 Column: 45 XML Soft Tabs: 2

Functional tests to be run with the “full stack” in place



XY Plot

Plot Title: Sample Plot

Label for x axis: Distance

Label for y axis: Count

Series

Series 1

Dataset: 5: Intersect on data 3 and data 4

Column for x axis: 1

Column for y axis: 2

Series Type: Line

Line Type: Solid

Line Color: Black

Line Width: 1.0

Series 2

Dataset: 7: Homo sapiens genes (NCBI36)

Column for x axis: 1

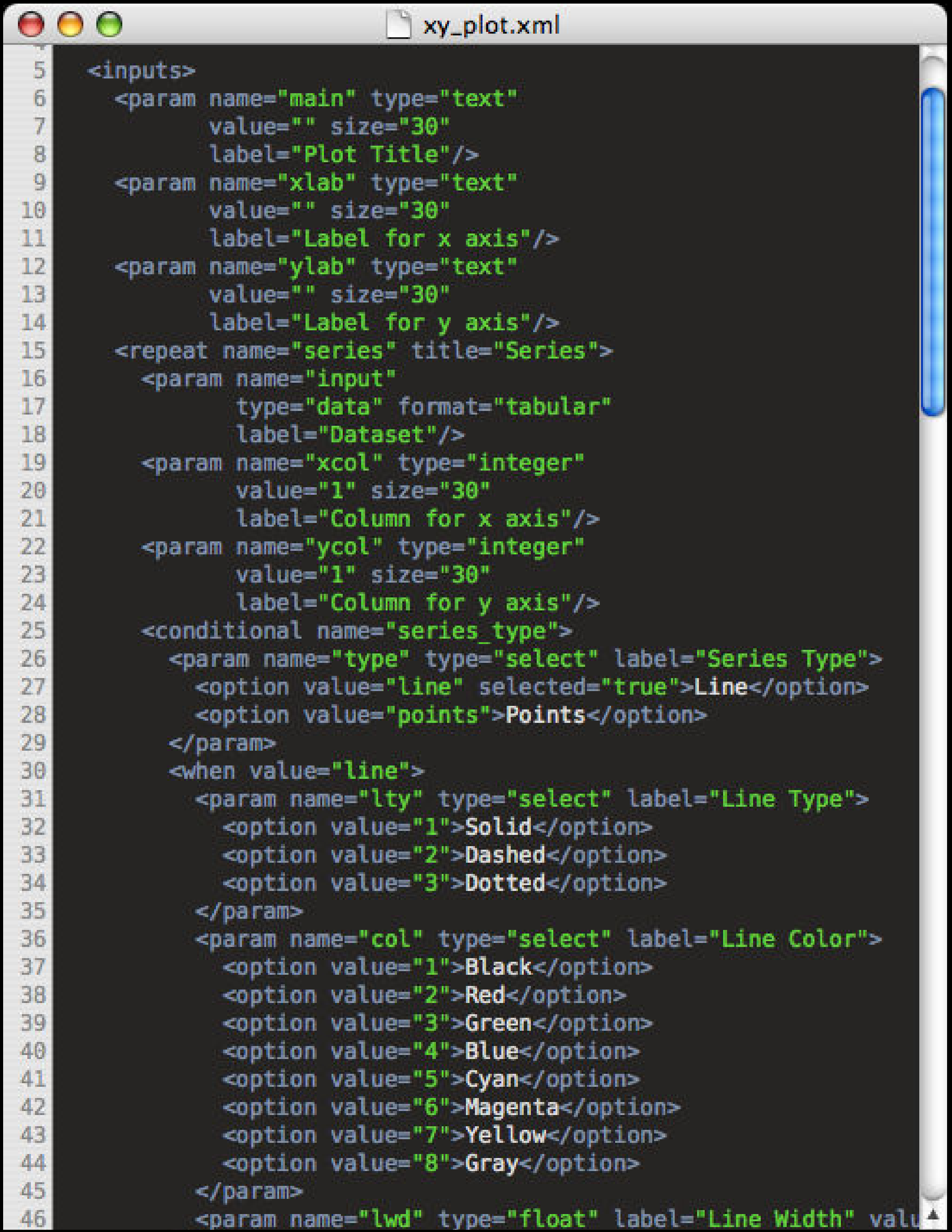
Column for y axis: 1

Series Type: Points

Point Type: Circle (hollow)

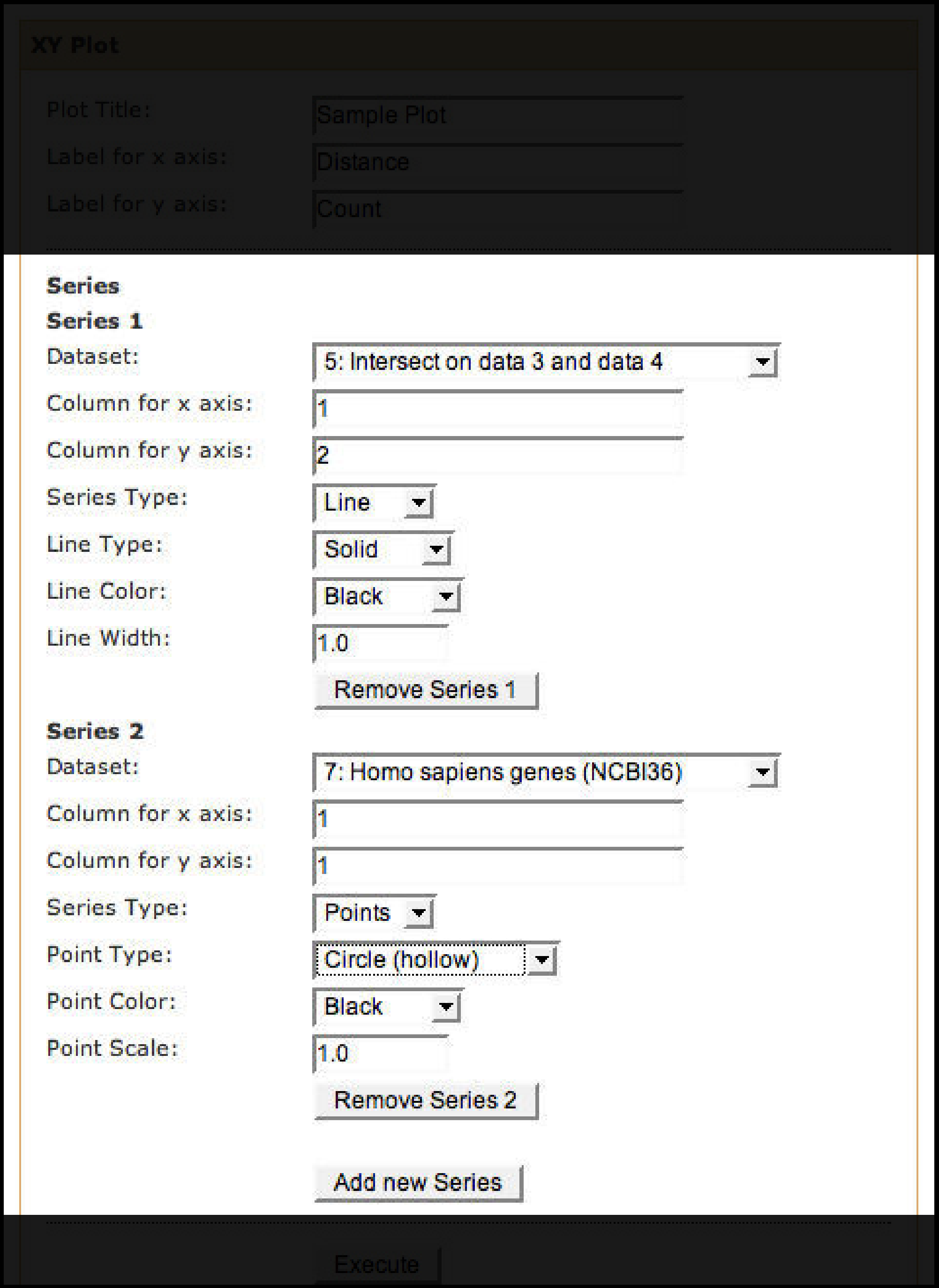
Point Color: Black

Point Scale: 1.0



```
<inputs>
  <param name="main" type="text"
         value="" size="30"
         label="Plot Title"/>
  <param name="xlab" type="text"
         value="" size="30"
         label="Label for x axis"/>
  <param name="ylab" type="text"
         value="" size="30"
         label="Label for y axis"/>
  <repeat name="series" title="Series">
    <param name="input"
           type="data" format="tabular"
           label="Dataset"/>
    <param name="xcol" type="integer"
           value="1" size="30"
           label="Column for x axis"/>
    <param name="ycol" type="integer"
           value="1" size="30"
           label="Column for y axis"/>
    <conditional name="series_type">
      <param name="type" type="select" label="Series Type">
        <option value="line" selected="true">Line</option>
        <option value="points">Points</option>
      </param>
      <when value="line">
        <param name="lty" type="select" label="Line Type">
          <option value="1">Solid</option>
          <option value="2">Dashed</option>
          <option value="3">Dotted</option>
        </param>
        <param name="col" type="select" label="Line Color">
          <option value="1">Black</option>
          <option value="2">Red</option>
          <option value="3">Green</option>
          <option value="4">Blue</option>
          <option value="5">Cyan</option>
          <option value="6">Magenta</option>
          <option value="7">Yellow</option>
          <option value="8">Gray</option>
        </param>
        <param name="lwd" type="float" label="Line Width" valu
```

Much more complex interfaces can be defined



XY Plot

Plot Title: Sample Plot

Label for x axis: Distance

Label for y axis: Count

Series

Series 1

Dataset: 5: Intersect on data 3 and data 4

Column for x axis: 1

Column for y axis: 2

Series Type: Line

Line Type: Solid

Line Color: Black

Line Width: 1.0

Series 2

Dataset: 7: Homo sapiens genes (NCBI36)

Column for x axis: 1

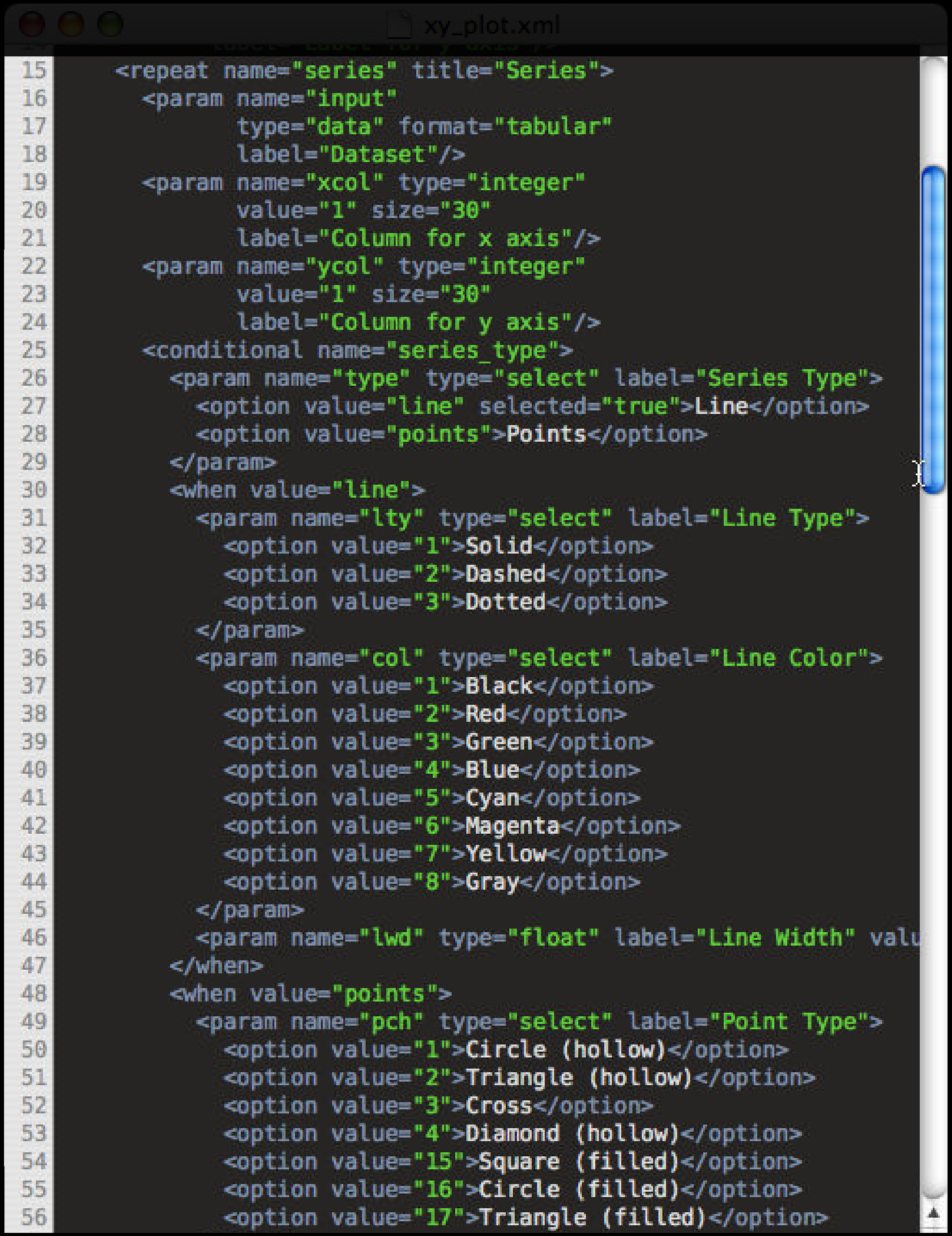
Column for y axis: 1

Series Type: Points

Point Type: Circle (hollow)

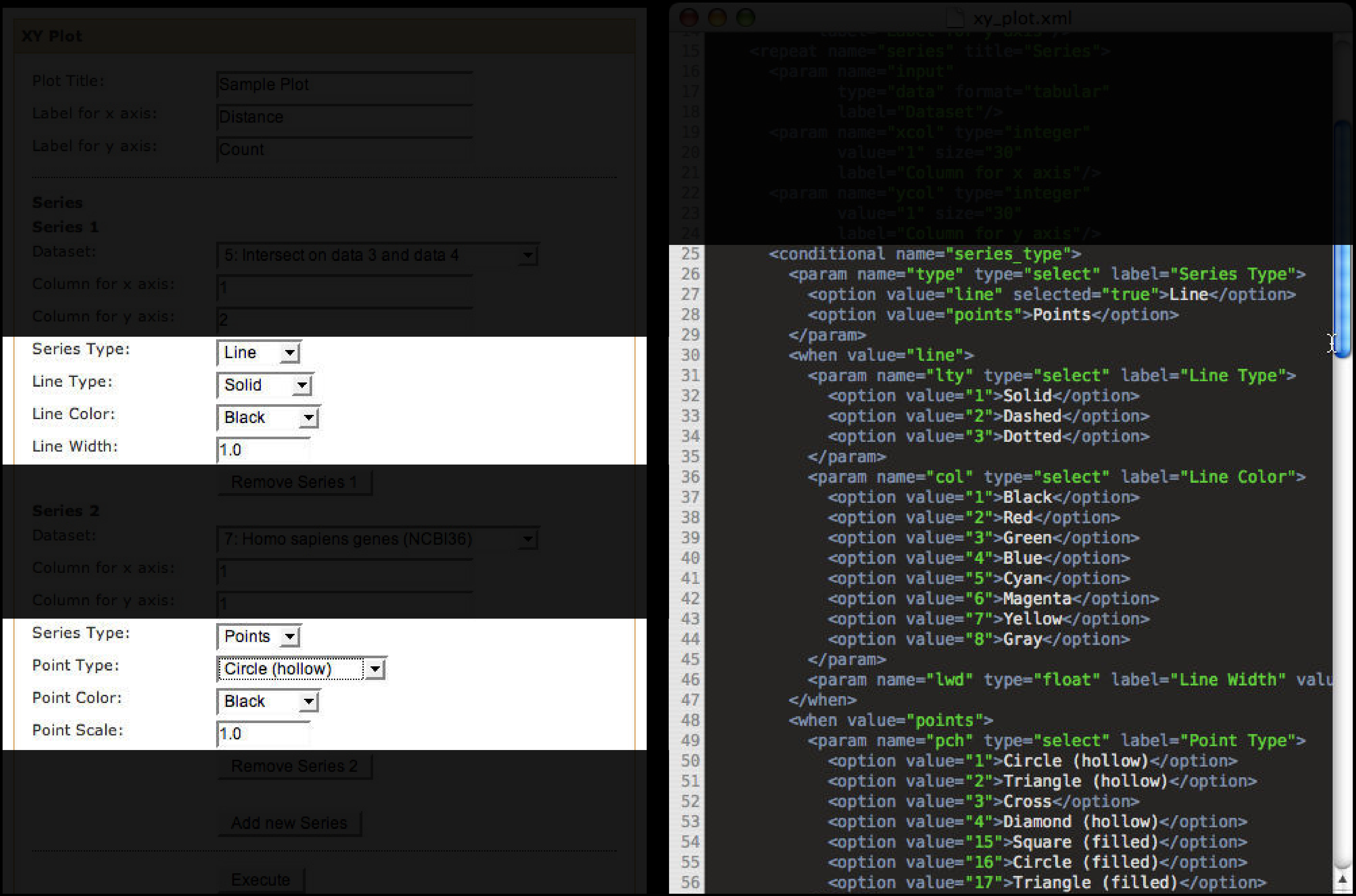
Point Color: Black

Point Scale: 1.0



```
    <repeat name="series" title="Series">
      <param name="input"
             type="data" format="tabular"
             label="Dataset"/>
      <param name="xcol" type="integer"
             value="1" size="30"
             label="Column for x axis"/>
      <param name="ycol" type="integer"
             value="1" size="30"
             label="Column for y axis"/>
      <conditional name="series_type">
        <param name="type" type="select" label="Series Type">
          <option value="line" selected="true">Line</option>
          <option value="points">Points</option>
        </param>
        <when value="line">
          <param name="lty" type="select" label="Line Type">
            <option value="1">Solid</option>
            <option value="2">Dashed</option>
            <option value="3">Dotted</option>
          </param>
          <param name="col" type="select" label="Line Color">
            <option value="1">Black</option>
            <option value="2">Red</option>
            <option value="3">Green</option>
            <option value="4">Blue</option>
            <option value="5">Cyan</option>
            <option value="6">Magenta</option>
            <option value="7">Yellow</option>
            <option value="8">Gray</option>
          </param>
          <param name="lwd" type="float" label="Line Width" valu
        </when>
        <when value="points">
          <param name="pch" type="select" label="Point Type">
            <option value="1">Circle (hollow)</option>
            <option value="2">Triangle (hollow)</option>
            <option value="3">Cross</option>
            <option value="4">Diamond (hollow)</option>
            <option value="15">Square (filled)</option>
            <option value="16">Circle (filled)</option>
            <option value="17">Triangle (filled)</option>
```

Repeating groups of parameters



XY Plot

Plot Title: Sample Plot

Label for x axis: Distance

Label for y axis: Count

Series

Series 1

Dataset: 5: Intersect on data 3 and data 4

Column for x axis: 1

Column for y axis: 2

Series Type: Line

Line Type: Solid

Line Color: Black

Line Width: 1.0

Series 2

Dataset: 7: Homo sapiens genes (NCBI36)

Column for x axis: 1

Column for y axis: 1

Series Type: Points

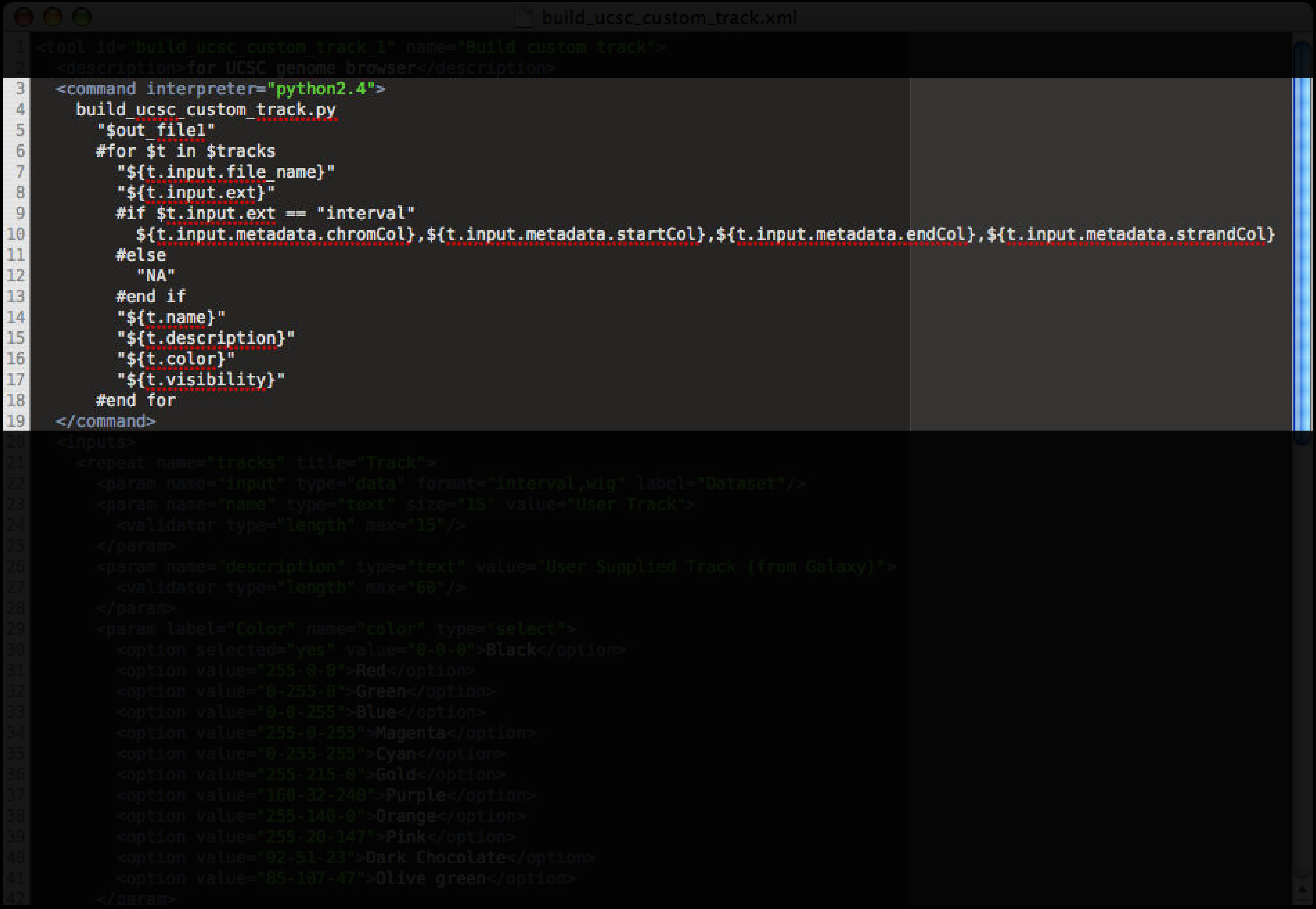
Point Type: Circle (hollow)

Point Color: Black

Point Scale: 1.0

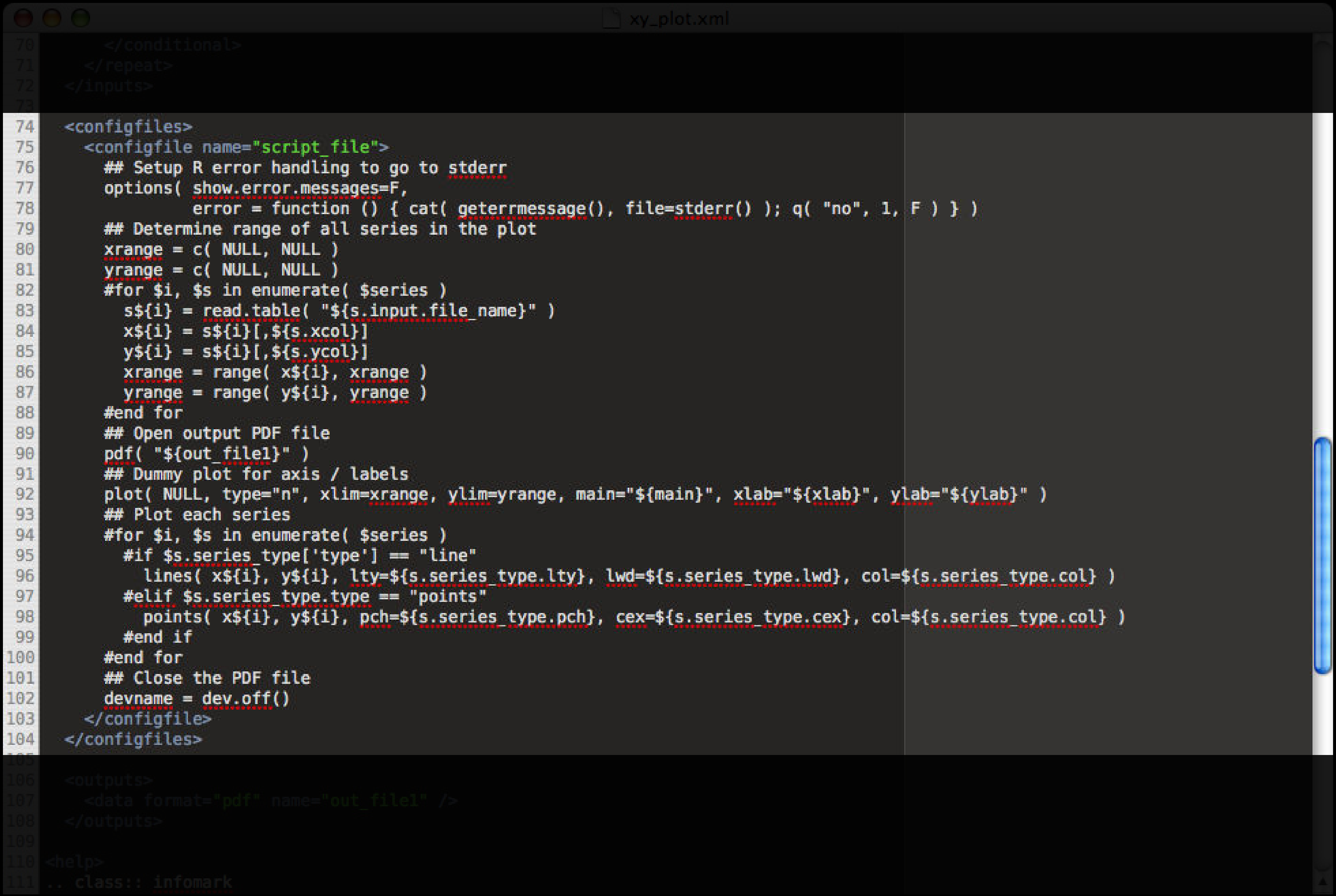
```
  <repeat name="series" title="Series">
    <param name="input"
       type="data" format="tabular"
       label="Dataset"/>
    <param name="xcol" type="integer"
       value="1" size="30"
       label="Column for x axis"/>
    <param name="ycol" type="integer"
       value="1" size="30"
       label="Column for y axis"/>
    <conditional name="series_type">
      <param name="type" type="select" label="Series Type">
        <option value="line" selected="true">Line</option>
        <option value="points">Points</option>
      </param>
      <when value="line">
        <param name="lty" type="select" label="Line Type">
          <option value="1">Solid</option>
          <option value="2">Dashed</option>
          <option value="3">Dotted</option>
        </param>
        <param name="col" type="select" label="Line Color">
          <option value="1">Black</option>
          <option value="2">Red</option>
          <option value="3">Green</option>
          <option value="4">Blue</option>
          <option value="5">Cyan</option>
          <option value="6">Magenta</option>
          <option value="7">Yellow</option>
          <option value="8">Gray</option>
        </param>
        <param name="lwd" type="float" label="Line Width" valu
      </when>
      <when value="points">
        <param name="pch" type="select" label="Point Type">
          <option value="1">Circle (hollow)</option>
          <option value="2">Triangle (hollow)</option>
          <option value="3">Cross</option>
          <option value="4">Diamond (hollow)</option>
          <option value="15">Square (filled)</option>
          <option value="16">Circle (filled)</option>
          <option value="17">Triangle (filled)</option>
```

Conditional groups, grouping constructs can be nested



```
<tool id="build_ucsc_custom_track_1" name="Build custom track">
  <description>for UCSC genome browser</description>
  <command interpreter="python2.4">
    build_ucsc_custom_track.py
      "$out_file1"
      #for $t in $tracks
        "${t.input.file_name}"
        "${t.input.ext}"
        #if $t.input.ext == "interval"
          ${t.input.metadata.chromCol},${t.input.metadata.startCol},${t.input.metadata.endCol},${t.input.metadata.strandCol}
        #else
          "NA"
        #end if
        "${t.name}"
        "${t.description}"
        "${t.color}"
        "${t.visibility}"
      #end for
  </command>
  <inputs>
    <repeat name="tracks" title="Track">
      <param name="input" type="data" format="interval,wig" label="Dataset"/>
      <param name="name" type="text" size="15" value="User Track">
        <validator type="length" max="15"/>
      </param>
      <param name="description" type="text" value="User Supplied Track (from Galaxy)">
        <validator type="length" max="60"/>
      </param>
      <param label="Color" name="color" type="select">
        <option selected="yes" value="0-0-0">Black</option>
        <option value="255-0-0">Red</option>
        <option value="0-255-0">Green</option>
        <option value="0-0-255">Blue</option>
        <option value="255-0-255">Magenta</option>
        <option value="0-255-255">Cyan</option>
        <option value="255-215-0">Gold</option>
        <option value="160-32-240">Purple</option>
        <option value="255-140-0">Orange</option>
        <option value="255-20-147">Pink</option>
        <option value="92-51-23">Dark Chocolate</option>
        <option value="85-107-47">Olive green</option>
      </param>
```

Template language for building complex command lines



```
      </conditional>
    </repeat>
  </inputs>

  <configfiles>
    <configfile name="script_file">
      ## Setup R error handling to go to stderr
      options( show.error.messages=F,
               error = function () { cat( geterrmessage(), file=stderr() ); q( "no", 1, F ) } )
      ## Determine range of all series in the plot
      xrange = c( NULL, NULL )
      yrange = c( NULL, NULL )
      #for $i, $s in enumerate( $series )
        s${i} = read.table( "${s.input.file_name}" )
        x${i} = s${i}[,${s.xcol}]
        y${i} = s${i}[,${s.ycol}]
        xrange = range( x${i}, xrange )
        yrange = range( y${i}, yrange )
      #end for
      ## Open output PDF file
      pdf( "${out_file1}" )
      ## Dummy plot for axis / labels
      plot( NULL, type="n", xlim=xrange, ylim=yrange, main="${main}", xlab="${xlab}", ylab="${ylab}" )
      ## Plot each series
      #for $i, $s in enumerate( $series )
        #if $s.series_type['type'] == "line"
          lines( x${i}, y${i}, lty=${s.series_type.lty}, lwd=${s.series_type.lwd}, col=${s.series_type.col} )
        #elif $s.series_type.type == "points"
          points( x${i}, y${i}, pch=${s.series_type.pch}, cex=${s.series_type.cex}, col=${s.series_type.col} )
        #end if
      #end for
      ## Close the PDF file
      devname = dev.off()
    </configfile>
  </configfiles>

  <outputs>
    <data format="pdf" name="out_file1" />
  </outputs>

<help>
.. class:: infomark
```

Or additional configuration files, scripts, ...

As data sizes grow, increasingly important to be able to express within tool parallelism

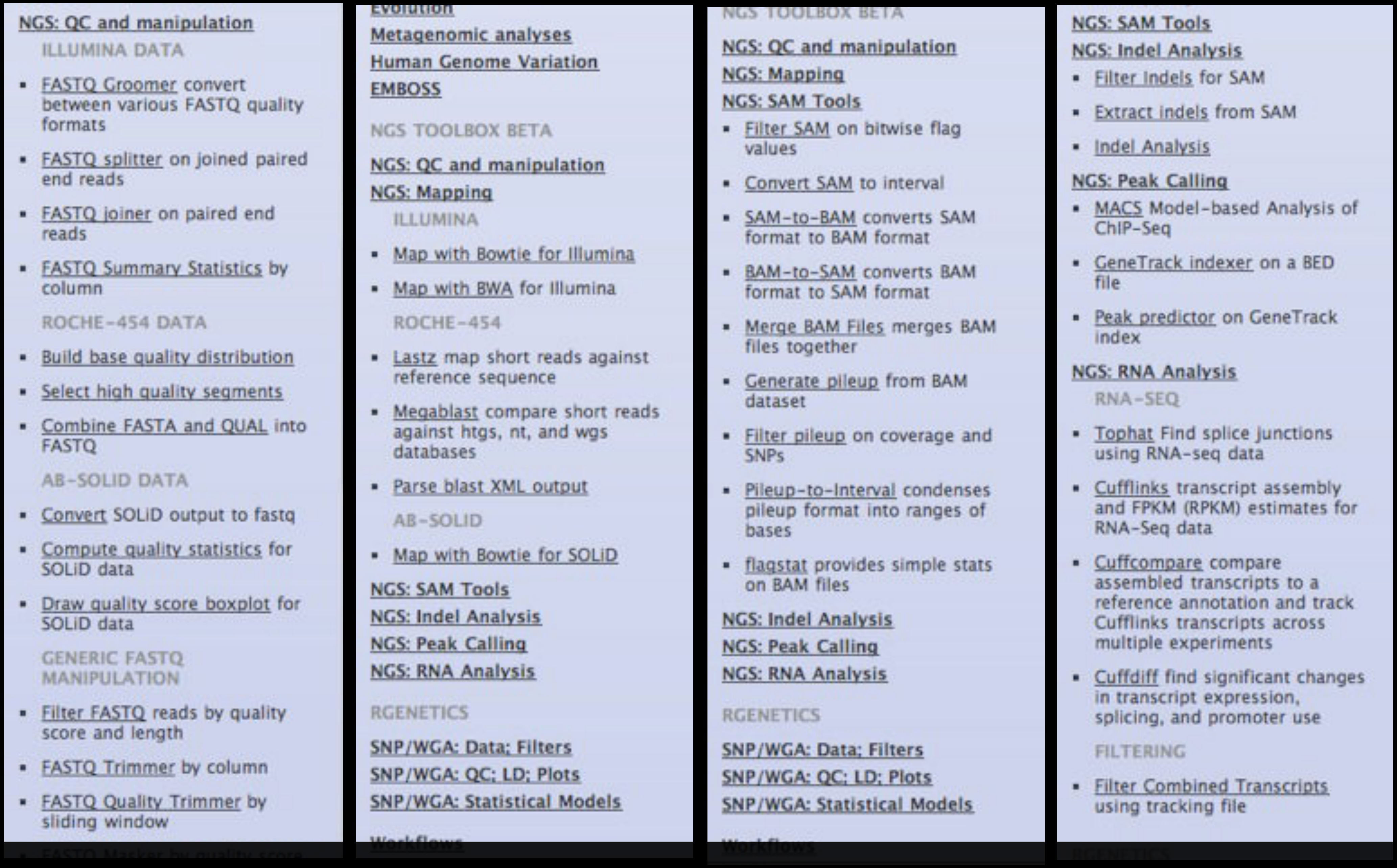
Naturally parallel (split/join) constructs can be specified in configuration

Parallel environments (MPI) can be used, but management delegated to underlying resources

Ongoing work to support more complex scenarios

Customization extends beyond tools

- Everything in the Galaxy framework is either configuration driven or pluggable (or both)
- Tools conventionally extended through configuration, but new tool types can be added
- Datatypes added through configuration, or plugin classes for advanced functionality
- *Nothing inherently specific to genomics!*



NGS: QC and manipulation

ILLUMINA DATA

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score

Evolution

Metagenomic analyses

Human Genome Variation

EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

ILLUMINA

- Map with Bowtie for Illumina
- Map with BWA for Illumina

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD

NGS: SAM Tools

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Workflows

NGS TOOLBOX BETA

NGS: QC and manipulation

NGS: Mapping

NGS: SAM Tools

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

NGS: Indel Analysis

NGS: Peak Calling

NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters

SNP/WGA: QC; LD; Plots

SNP/WGA: Statistical Models

Workflows

NGS: SAM Tools

NGS: Indel Analysis

- Filter Indels for SAM
- Extract indels from SAM
- Indel Analysis

NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

RNA-SEQ

- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

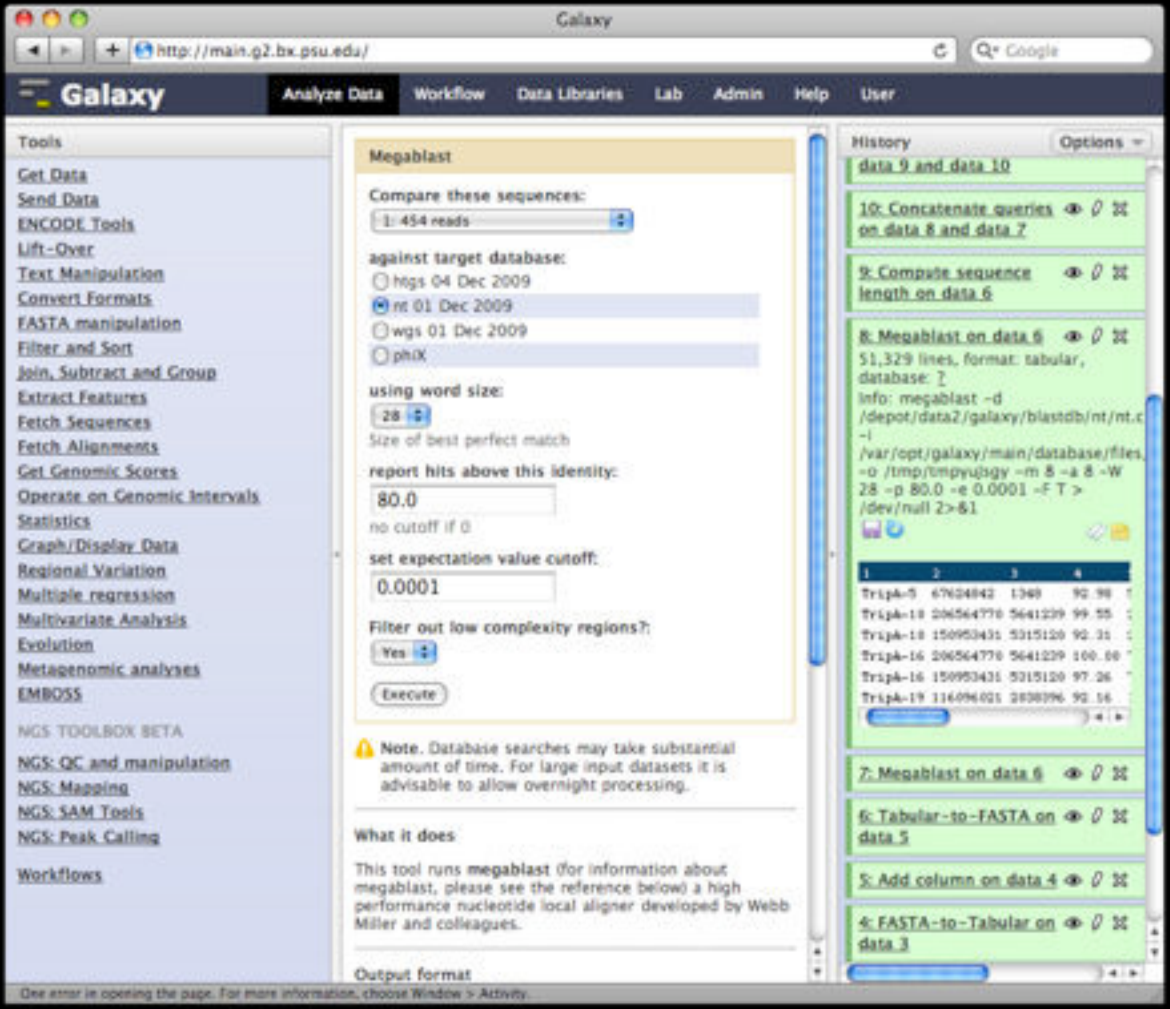
- Filter Combined Transcripts using tracking file

RGENETICS

Dozens of tools for different NGS applications packaged with Galaxy

Analysis environment

Galaxy analysis interface



Megablast

Compare these sequences: 1: 454 reads

against target database:

○ htgs 04 Dec 2009
● nt 01 Dec 2009
○ wgs 01 Dec 2009
○ phiX

using word size: 28

Size of best perfect match

report hits above this identity: 80.0

no cutoff if 0

set expectation value cutoff: 0.0001

Filter out low complexity regions?: Yes

Note. Database searches may take substantial amount of time. For large input datasets it is advisable to allow overnight processing.

What it does

This tool runs **megablast** (for information about megablast, please see the reference below) a high performance nucleotide local aligner developed by Webb Miller and colleagues.

Output format

One error in opening the page. For more information, choose Window > Activity.

data 9 and data 10

10: Concatenate queries on data 8 and data 7

9: Compute sequence length on data 6

8: Megablast on data 6

51,329 lines, format: tabular, database: ?

Info: megablast -d /depot/data2/galaxy/blastdb/nt/nt.c -i /var/opt/galaxy/main/database/files/ -o /tmp/tmpyujsgy -m 8 -a 8 -W 28 -p 80.0 -e 0.0001 -F T > /dev/null 2>&1

1	2	3	4
TripA-5	67624842	1360	92.98
TripA-10	206564770	5641239	99.55
TripA-10	150953431	5315120	92.31
TripA-16	206564770	5641239	100.00
TripA-16	150953431	5315120	97.26
TripA-19	116096021	2838396	92.16

7: Megablast on data 6

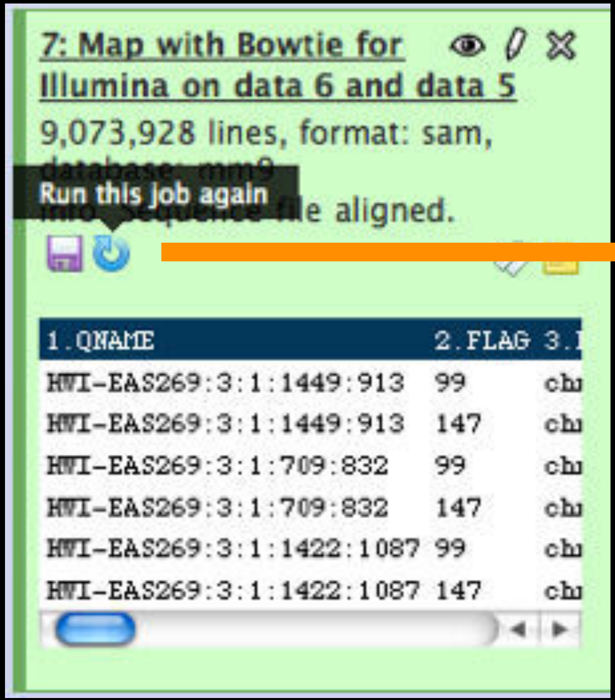
6: Tabular-to-FASTA on data 5

5: Add column on data 4

4: FASTA-to-Tabular on data 3

- Consistent tool user interfaces automatically generated
- History system facilitates and tracks multistep analyses

Automatically and transparently tracks every step of every analysis

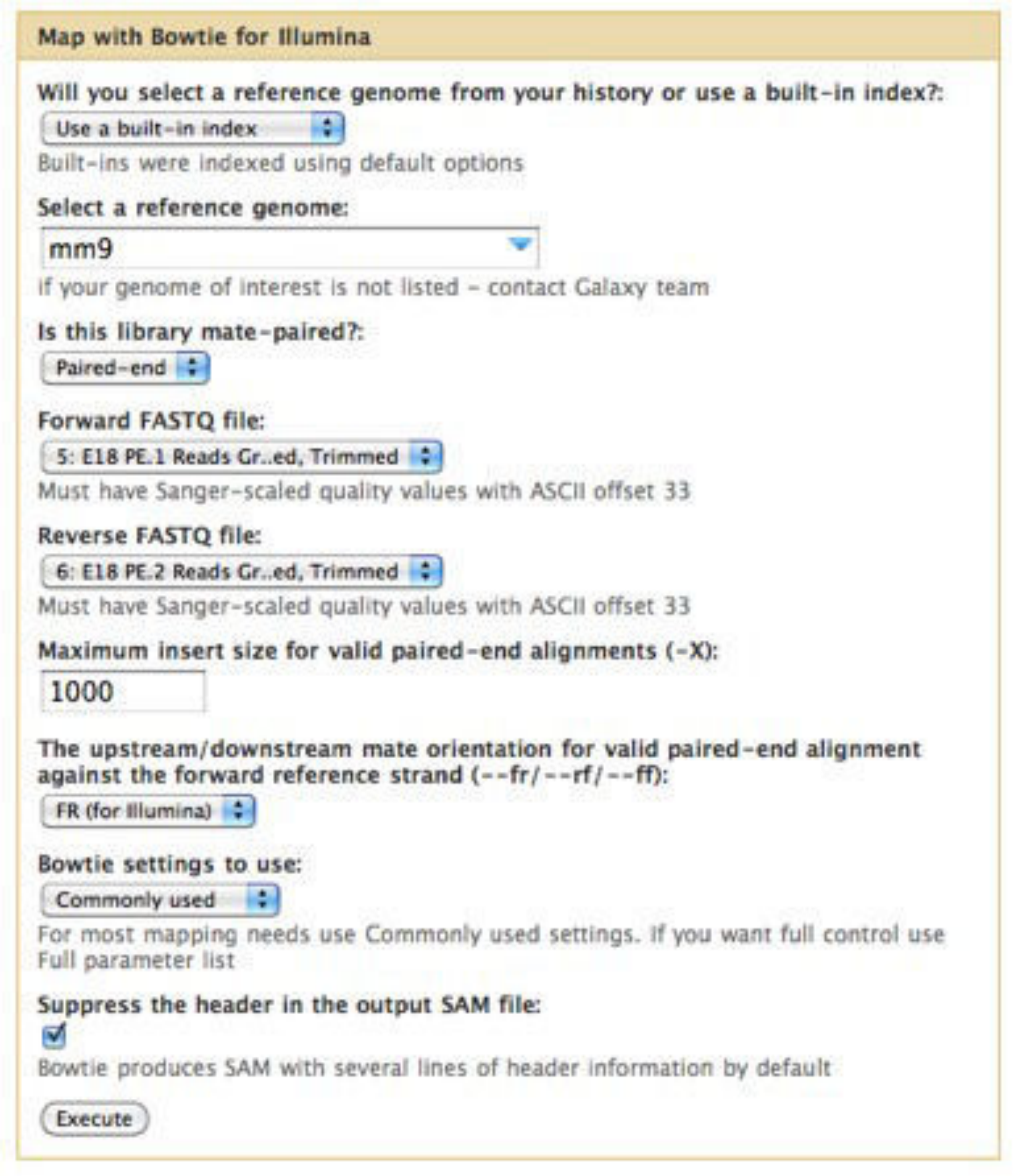


7: Map with Bowtie for Illumina on data 6 and data 5

9,073,928 lines, format: sam, database: mm9

Info: Sequence file aligned.

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



Map with Bowtie for Illumina

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

Built-ins were indexed using default options

Select a reference genome: mm9

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

Is this library mate-paired?: Paired-end

Forward FASTQ file: 5: E18 PE.1 Reads Gr..ed, Trimmed

Must have Sanger-scaled quality values with ASCII offset 33

Reverse FASTQ file: 6: E18 PE.2 Reads Gr..ed, Trimmed

Must have Sanger-scaled quality values with ASCII offset 33

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

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

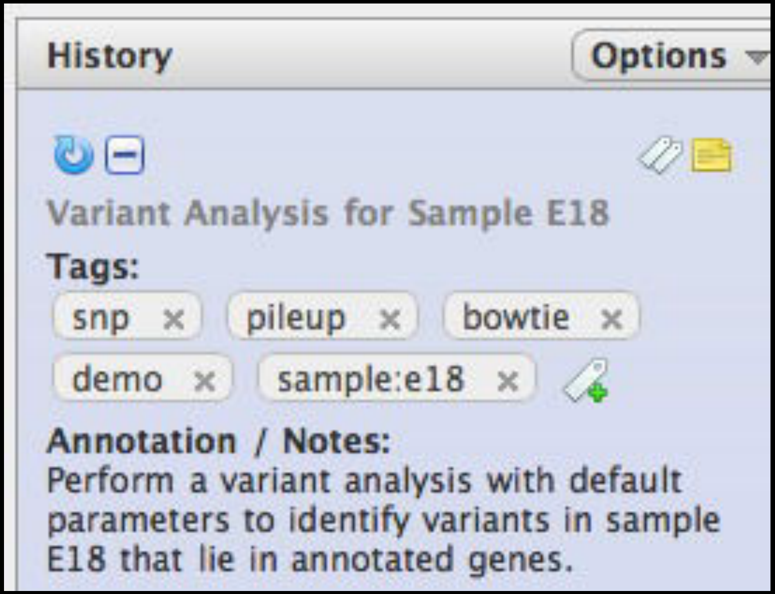
Bowtie settings to use: Commonly used

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

Suppress the header in the output SAM file: ☑

Bowtie produces SAM with several lines of header information by default

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

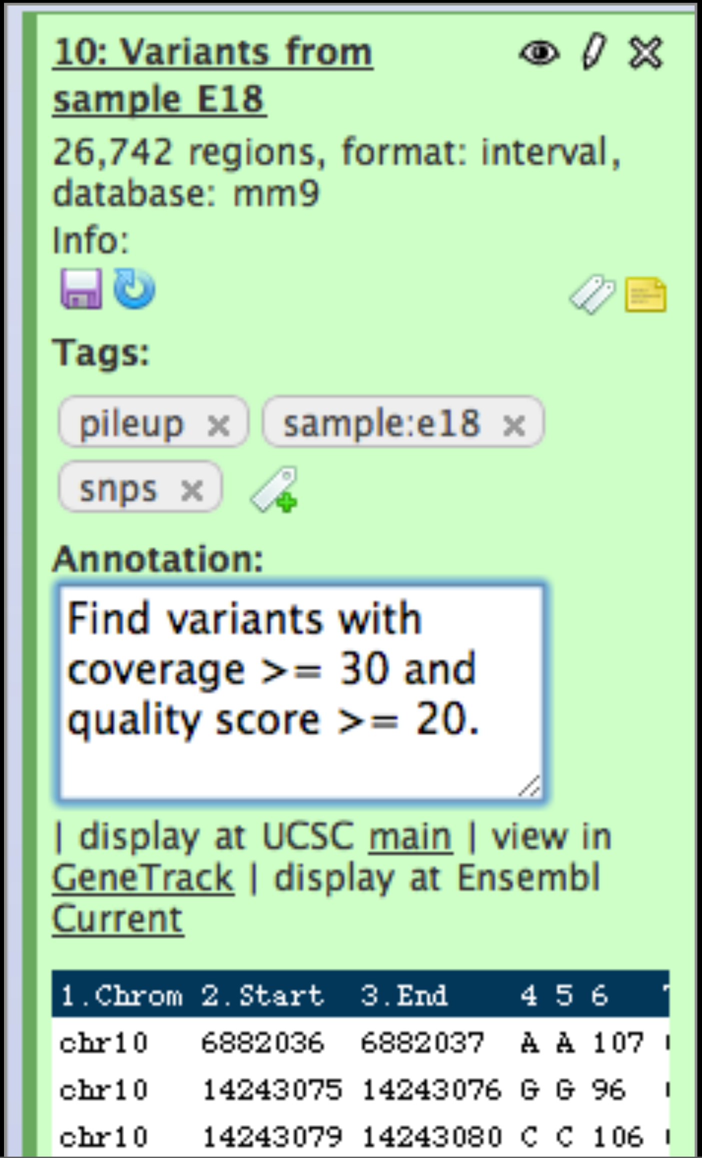


Variant Analysis for Sample E18

Tags:

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

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


10: Variants from sample E18

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

Info:

Tags:

pileup × sample:e18 × snps ×

Annotation:

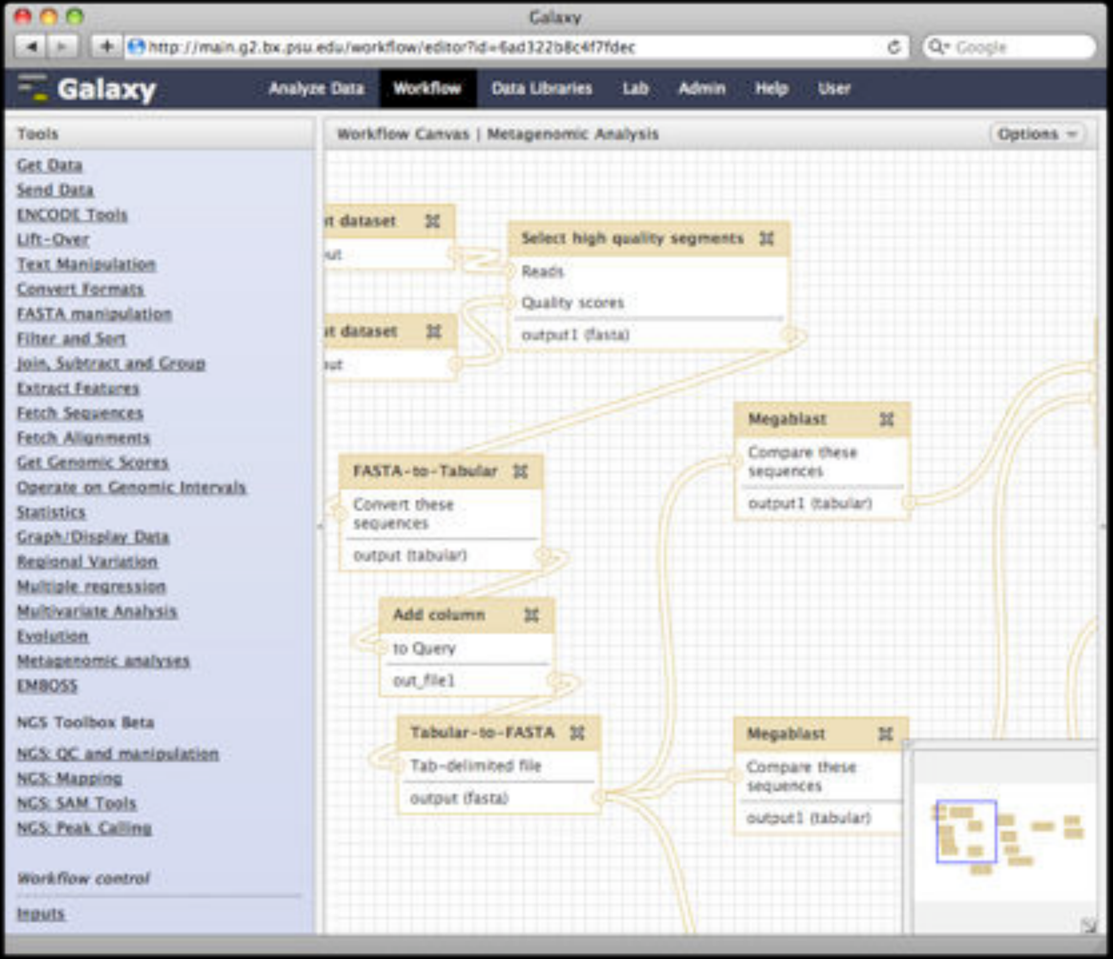
Find variants with coverage >= 30 and quality score >= 20.

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

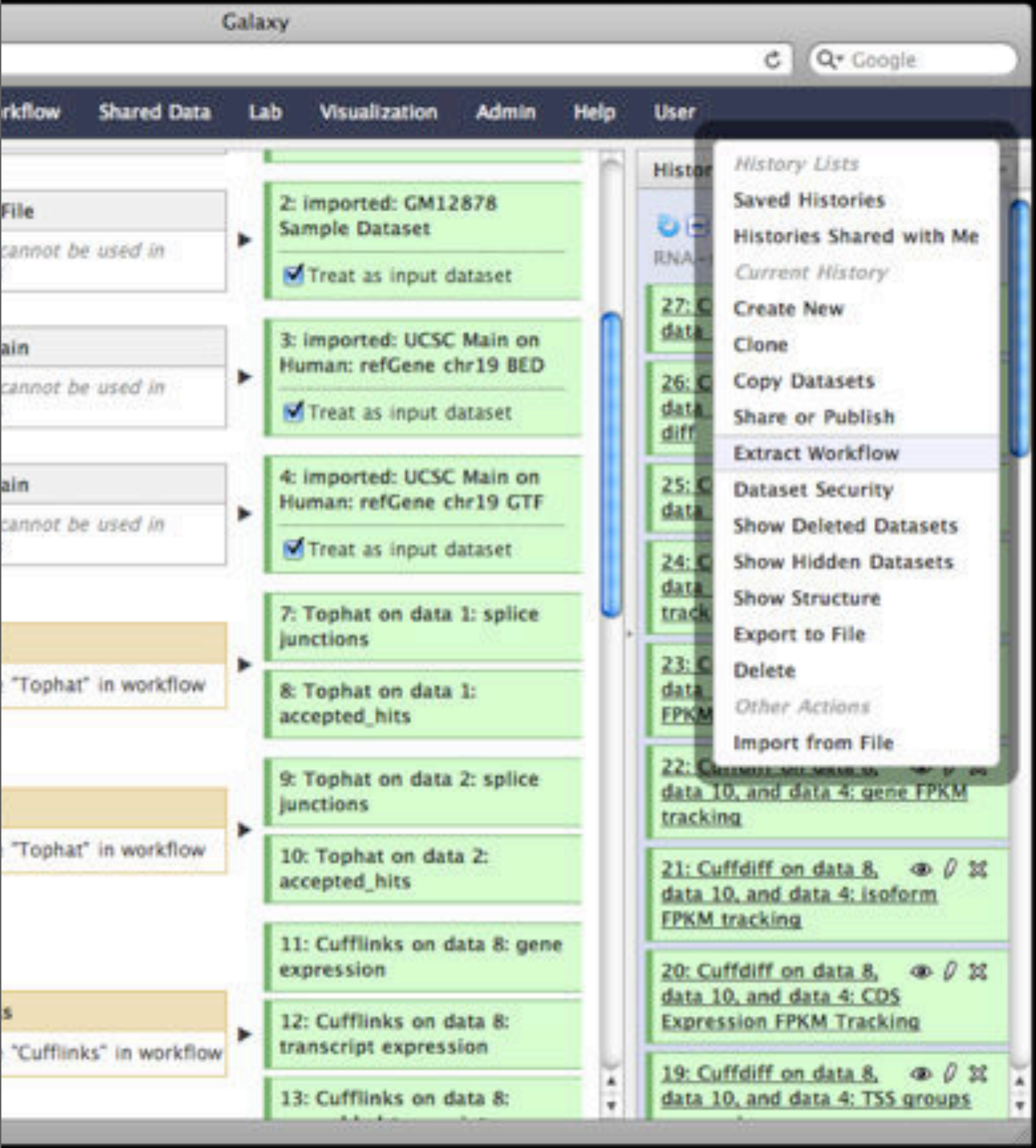
1.Chrom	2.Start	3.End	4	5	6
chr10	6882036	6882037	A	A	107
chr10	14243075	14243076	G	G	96
chr10	14243079	14243080	C	C	106

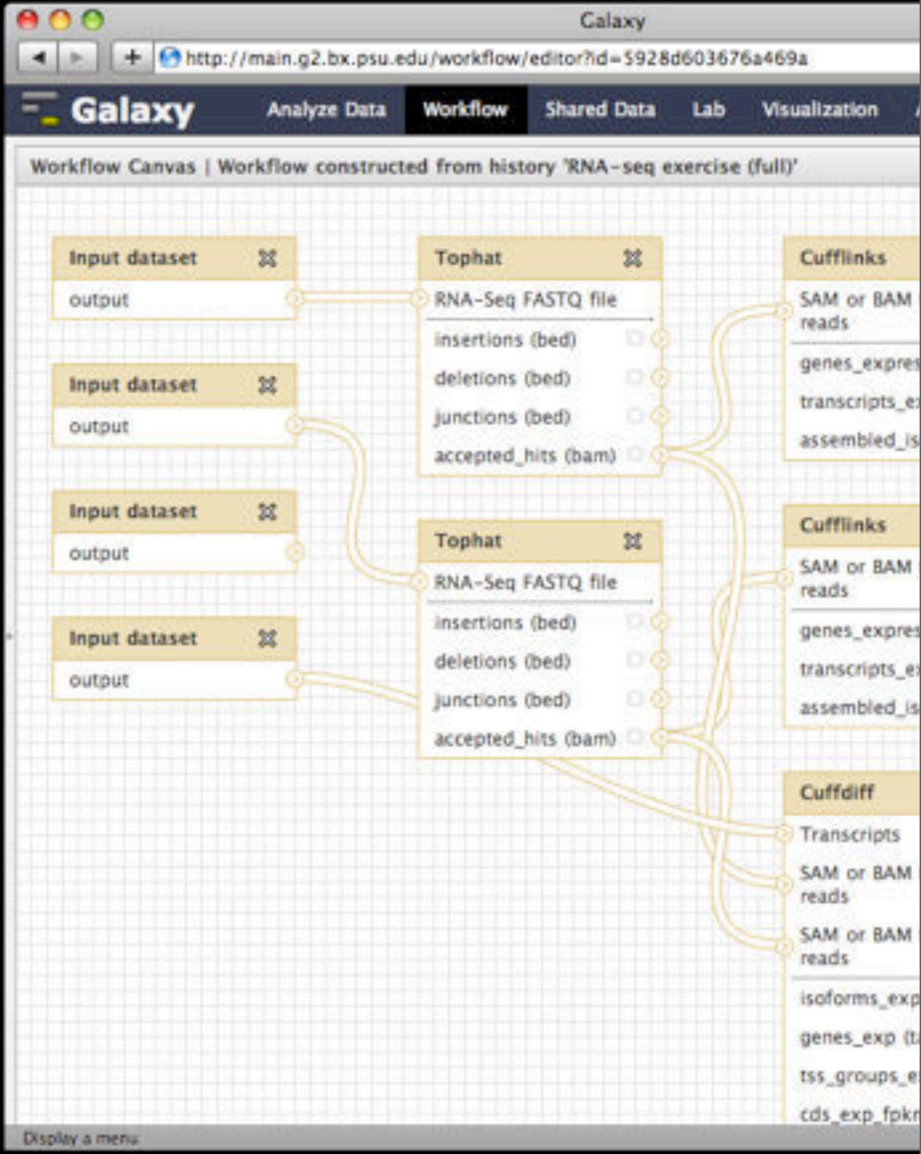
Workflows

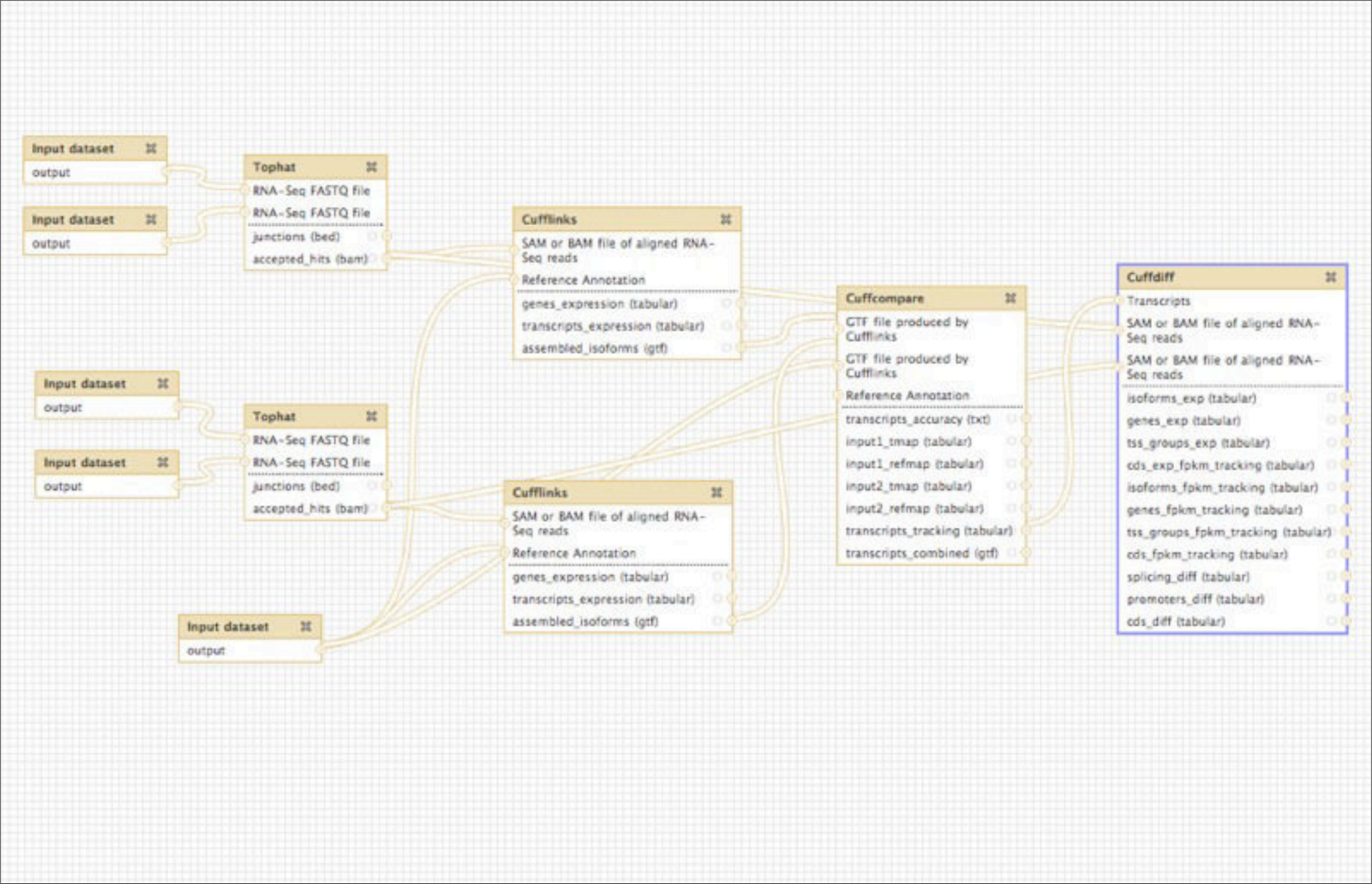
Galaxy workflow system



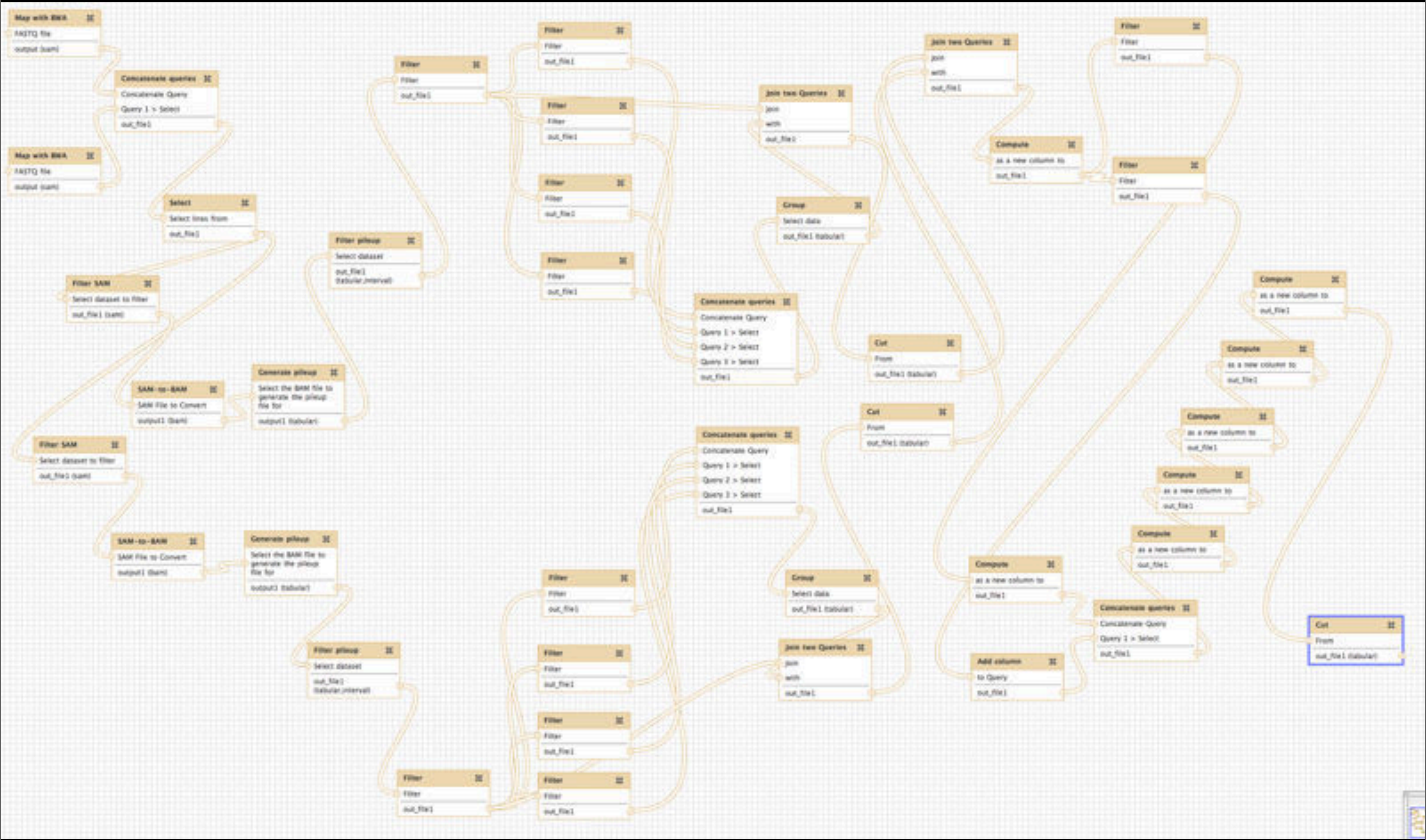
- Workflows can be constructed from scratch *or* extracted from existing analysis histories
- Facilitate reuse, as well as providing precise reproducibility of a complex analysis


Galaxy
Google
rkflow
Shared Data
Lab
Visualization
Admin
Help
User
File
cannot be used in
ain
cannot be used in
ain
cannot be used in
2: imported: GM12878 Sample Dataset
☑ Treat as input dataset
3: imported: UCSC Main on Human: refGene chr19 BED
☑ Treat as input dataset
4: imported: UCSC Main on Human: refGene chr19 GTF
☑ Treat as input dataset
7: Tophat on data 1: splice junctions
"Tophat" in workflow
8: Tophat on data 1: accepted_hits
9: Tophat on data 2: splice junctions
"Tophat" in workflow
10: Tophat on data 2: accepted_hits
11: Cufflinks on data 8: gene expression
12: Cufflinks on data 8: transcript expression
"Cufflinks" in workflow
13: Cufflinks on data 8:
Histor
RNA-
27: C data
26: C data diff
25: C data
24: C data track
23: C data FPKM
22: Cuffdiff on data 8, data 10, and data 4: gene FPKM tracking
21: Cuffdiff on data 8, data 10, and data 4: isoform FPKM tracking
20: Cuffdiff on data 8, data 10, and data 4: CDS Expression FPKM Tracking
19: Cuffdiff on data 8, data 10, and data 4: TSS groups
History Lists
Saved Histories
Histories Shared with Me
Current History
Create New
Clone
Copy Datasets
Share or Publish
Extract Workflow
Dataset Security
Show Deleted Datasets
Show Hidden Datasets
Show Structure
Export to File
Delete
Other Actions
Import from File


Galaxy
http://main.g2.bx.psu.edu/workflow/editor?id=5928d603676a469a
Galaxy
Analyze Data
Workflow
Shared Data
Lab
Visualization
Workflow Canvas | Workflow constructed from history 'RNA-seq exercise (full)'
Input dataset
output
Input dataset
output
Input dataset
output
Input dataset
output
Tophat
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)
Tophat
RNA-Seq FASTQ file
insertions (bed)
deletions (bed)
junctions (bed)
accepted_hits (bam)
Cufflinks
SAM or BAM reads
genes_expres
transcripts_e
assembled_is
Cufflinks
SAM or BAM reads
genes_expres
transcripts_e
assembled_is
Cuffdiff
Transcripts
SAM or BAM reads
SAM or BAM reads
isoforms_exp
genes_exp (t
tss_groups_e
cds_exp_fpkm
Display a menu



Example: Workflow for differential expression analysis of RNA-seq using Tophat/ Cufflinks tools



Example: Diagnosing low-frequency heterosplasmic sites in two tissues from the same individual

Galaxy deployment models

Galaxy main site (http://usegalaxy.org)

- Public web site, anybody can use
- ~500 new users per month, ~100 TB of user data, ~130,000 analysis jobs per month, every month is our busiest month ever...
- Will continue to be maintained and enhanced, but with limits and quotas
- Centralized solution cannot scale to meet data analysis demands

Local Galaxy instances (http://getgalaxy.org)

- 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 existing resources

- Move intensive processing (tool execution) to other hosts
- Frees up the application server to serve requests and manage jobs
- Utilize existing resources
- Supports any batch scheduler that supports DRMAA (most of them)
- All levels of job running and scheduling are pluggable

CLUSTER RESOURCES™
TORQUE

GRID ENGINE

Platform Computing

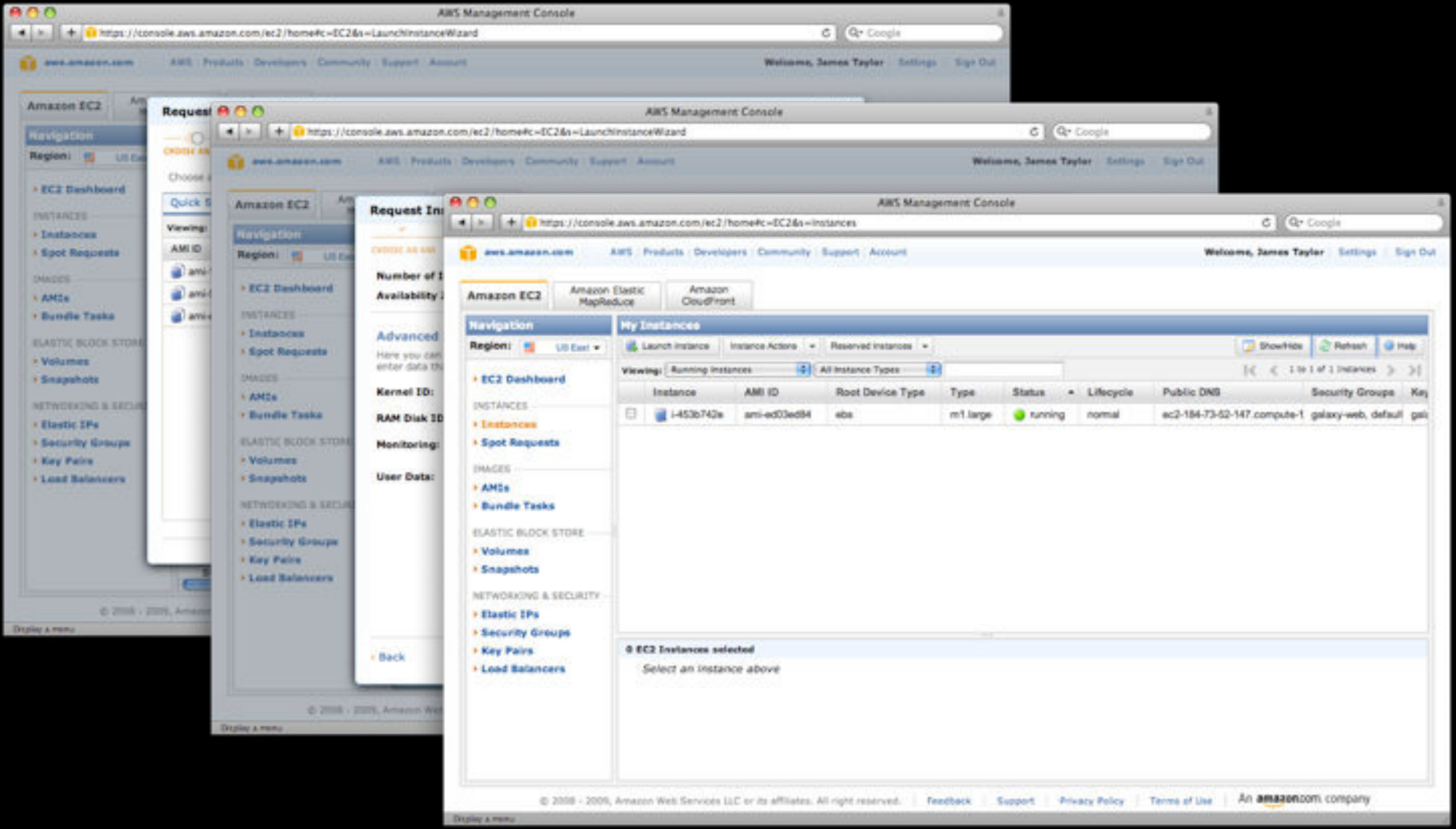
DRMAA
Distributed Resource Management Application API — www.drmaa.org

Galaxy Cloud (http://usegalaxy.org/cloud)

- On-demand resource acquisition fits well with the irregular resource needs of many labs working with sequence data

- Our goal is to approach the ease of use of a "software as a service" solution while maintaining the flexibility and control of an infrastructure based solution

Using Amazon EC2: Startup in 3 steps



AWS Management Console

https://console.aws.amazon.com/ec2/home#c=EC2&s=LaunchInstanceWizard

https://console.aws.amazon.com/ec2/home#c=EC2&s=Instances

Welcome, James Taylor

My Instances

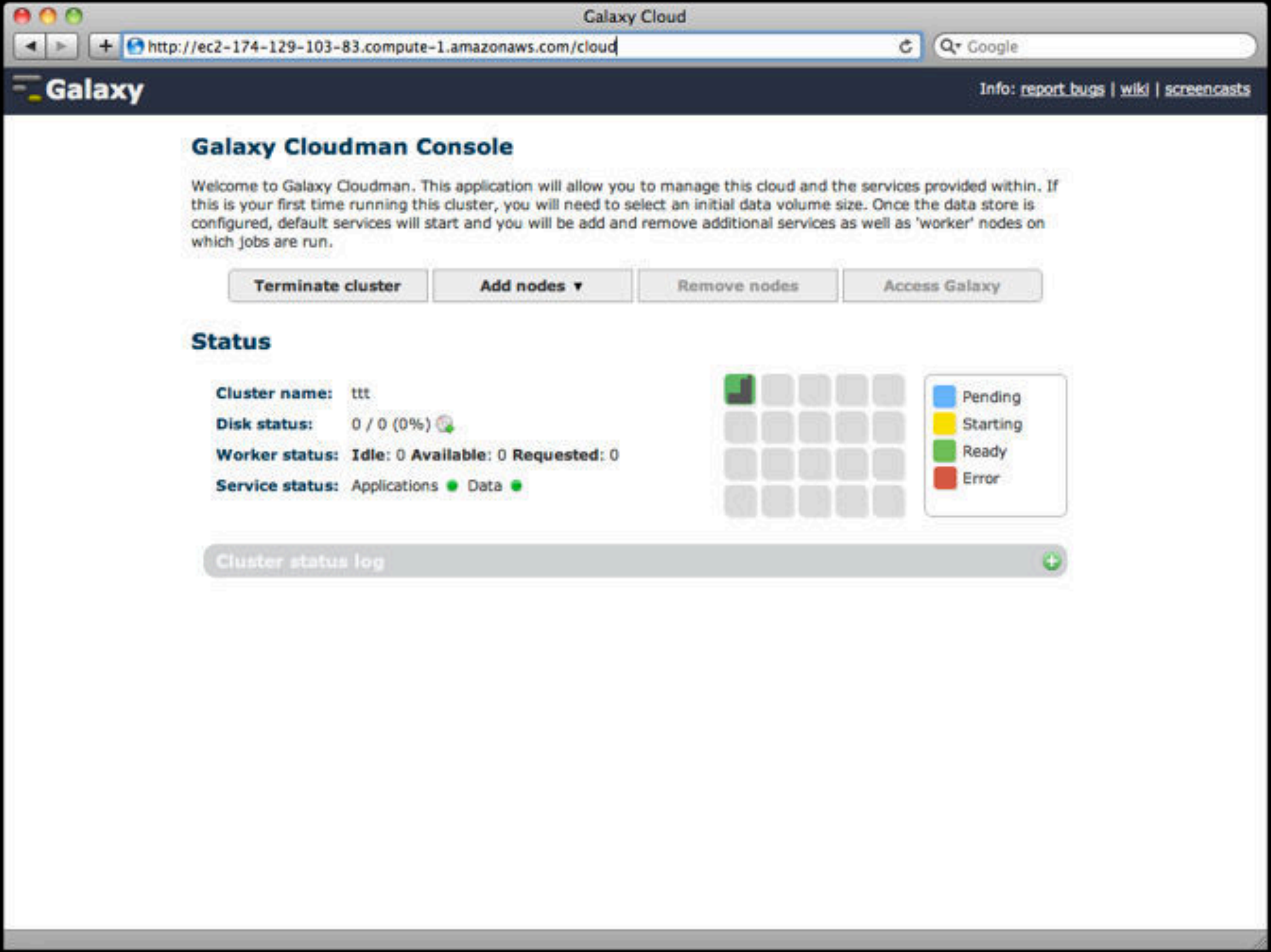
Viewing: Running Instances · All Instance Types · 1 to 1 of 1 Instances

	Instance	AMI ID	Root Device Type	Type	Status	Lifecycle	Public DNS	Security Groups	Key
□	i-453b742e	ami-ed03ed84	ebs	m1.large	● running	normal	ec2-184-73-52-147.compute-1	galaxy-web, default	gal

0 EC2 Instances selected

Select an instance above

© 2008 - 2009, Amazon Web Services LLC or its affiliates. All right reserved.



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.

Status

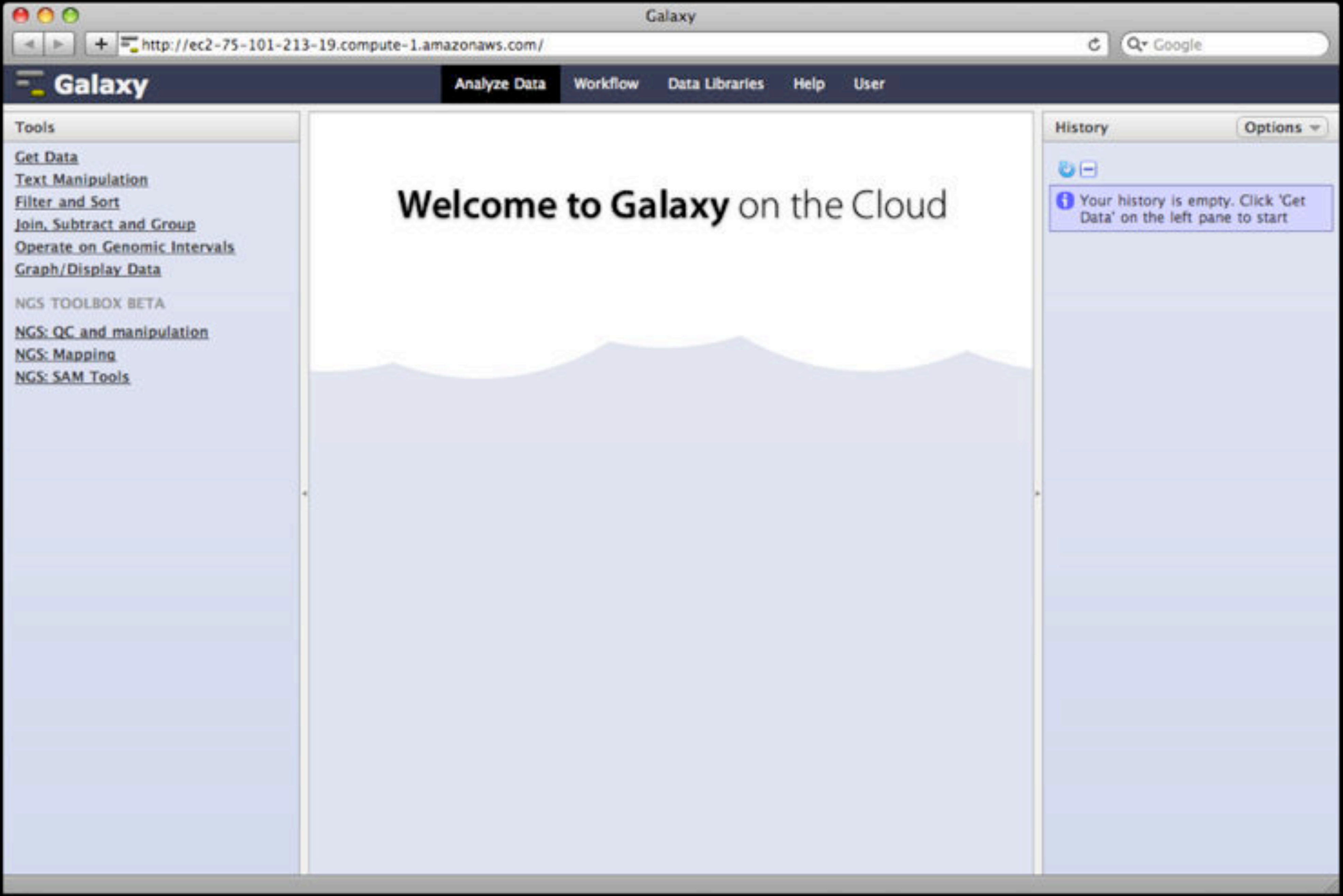
Cluster name: ttt

Disk status: 0 / 0 (0%)

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

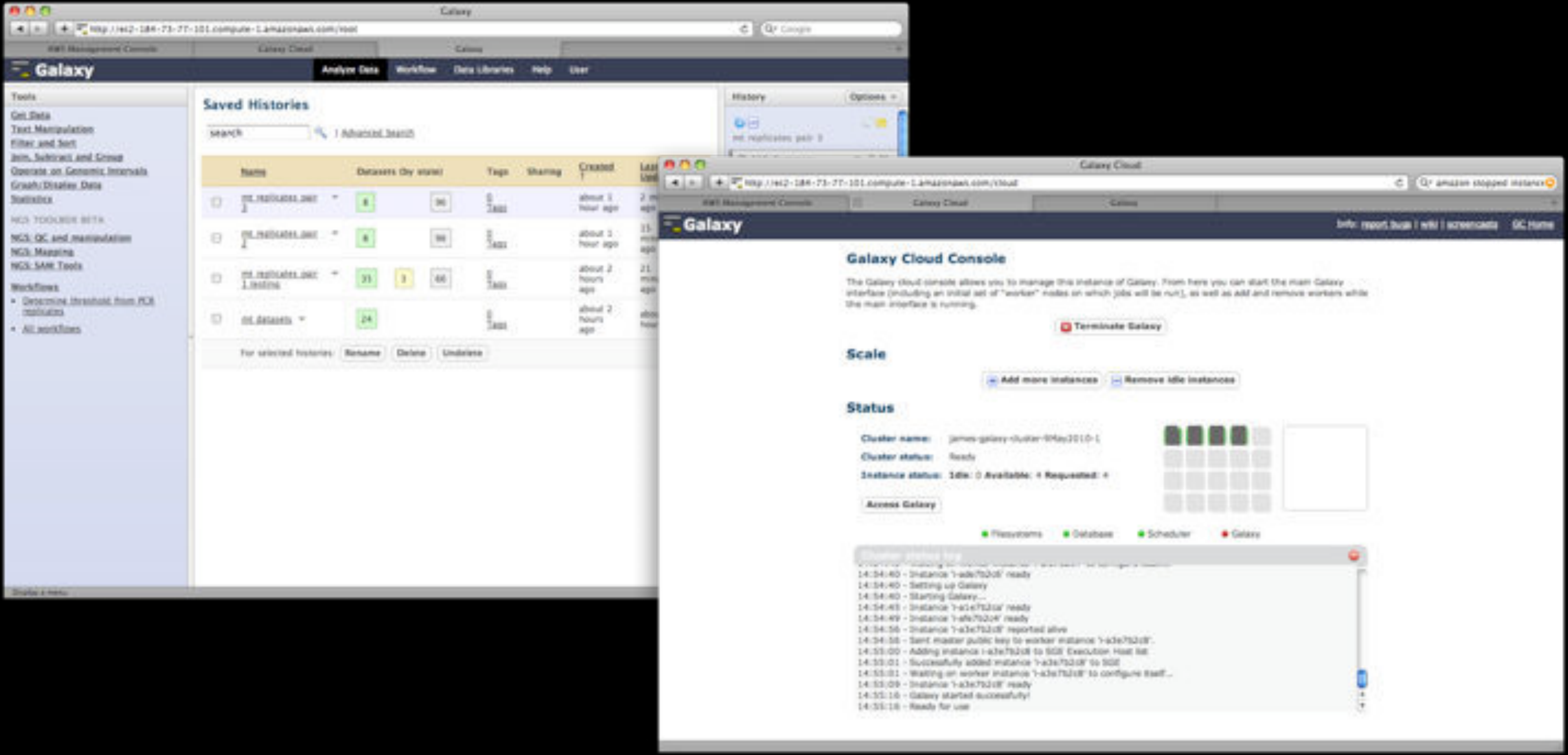
Service status: Applications ● Data ●

🟦 Pending
🟨 Starting
🟩 Ready
🟥 Error



Welcome to Galaxy on the Cloud

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



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.

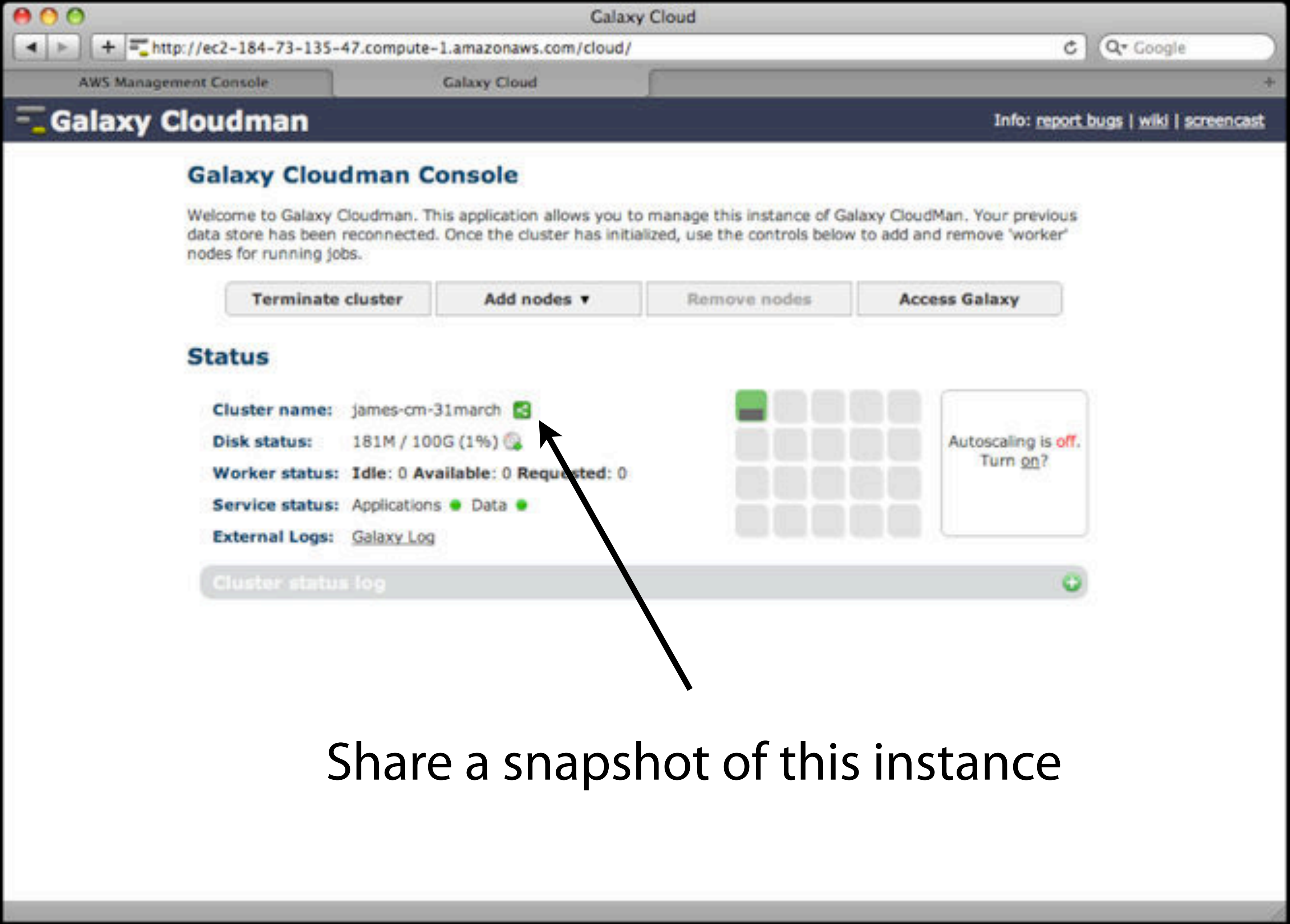
Cluster name: james-galaxy-cluster-9May2010-1

Cluster status: Ready

Instance status: **Idle:** 0 **Available:** 4 **Requested:** 4

```
14:54:40 - Instance 'i-ade7b2d6' ready
14:54:40 - Setting up Galaxy
14:54:40 - Starting Galaxy...
14:54:45 - Instance 'i-a1e7b2ca' ready
14:54:49 - Instance 'i-afe7b2c4' ready
14:54:56 - Instance 'i-a3e7b2c8' reported alive
14:54:58 - Sent master public key to worker instance 'i-a3e7b2c8'.
14:55:00 - Adding instance i-a3e7b2c8 to SGE Execution Host list
14:55:01 - Successfully added instance 'i-a3e7b2c8' to SGE
14:55:01 - Waiting on worker instance 'i-a3e7b2c8' to configure itself...
14:55:09 - Instance 'i-a3e7b2c8' ready
14:55:16 - Galaxy started successfully!
14:55:16 - Ready for use
```

Can use like any other Galaxy instance, with additional compute nodes acquired and released (*automatically*) in response to usage



Galaxy Cloud

http://ec2-184-73-135-47.compute-1.amazonaws.com/cloud/

Galaxy Cloudman Console

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

Status

Cluster name: james-cm-31march

Disk status: 181M / 100G (1%)

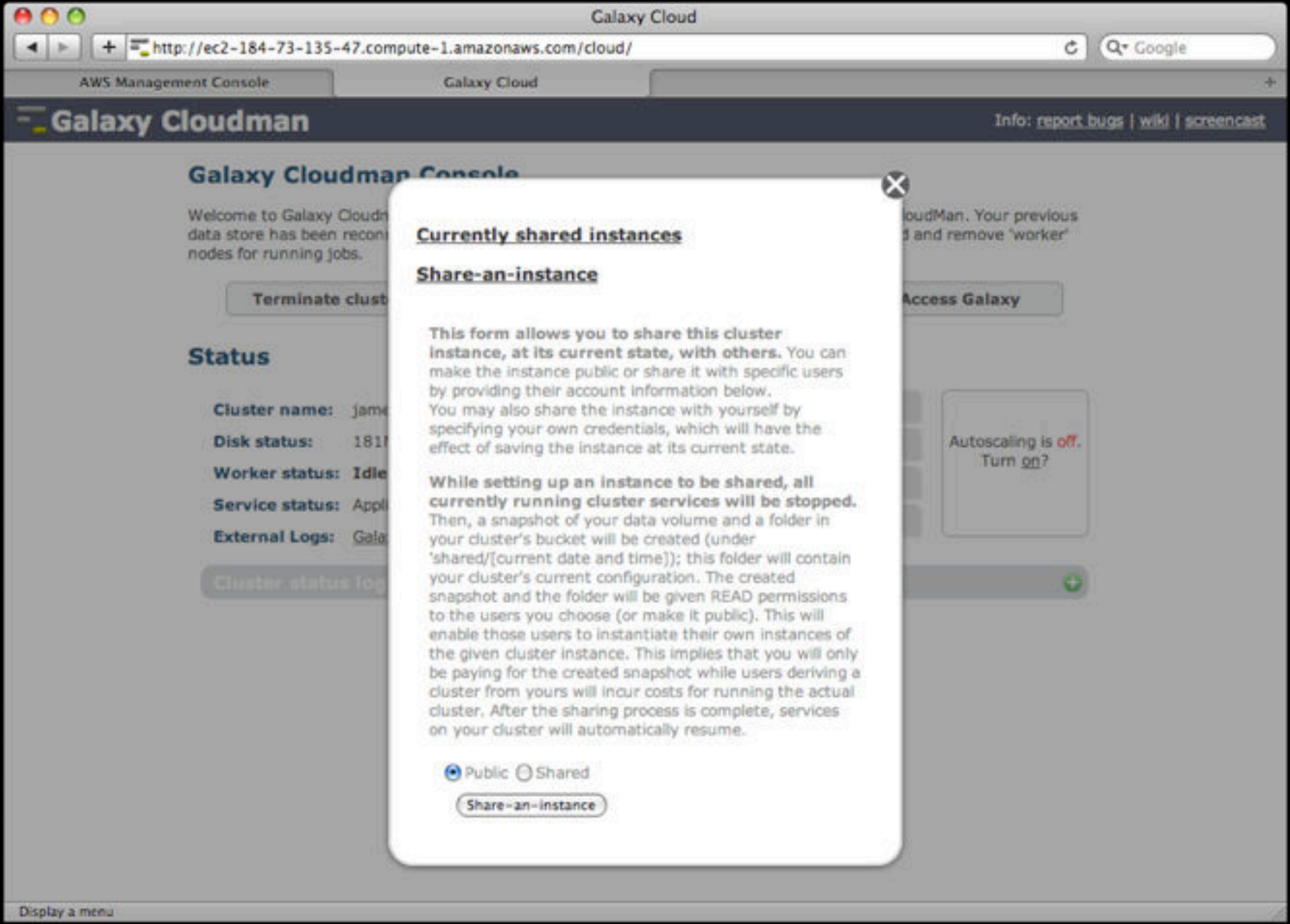
Worker status: **Idle**: 0 **Available**: 0 **Requested**: 0

Service status: Applications ● Data ●

External Logs: Galaxy Log

Autoscaling is off. Turn on?

Share a snapshot of this instance



Currently shared instances

Share-an-instance

This form allows you to share this cluster instance, at its current state, with others. You can make the instance public or share it with specific users by providing their account information below.
You may also share the instance with yourself by specifying your own credentials, which will have the effect of saving the instance at its current state.

While setting up an instance to be shared, all currently running cluster services will be stopped. Then, a snapshot of your data volume and a folder in your cluster's bucket will be created (under 'shared/[current date and time]); this folder will contain your cluster's current configuration. The created snapshot and the folder will be given READ permissions to the users you choose (or make it public). This will enable those users to instantiate their own instances of the given cluster instance. This implies that you will only be paying for the created snapshot while users deriving a cluster from yours will incur costs for running the actual cluster. After the sharing process is complete, services on your cluster will automatically resume.

● Public ○ Shared

Tool installation and configuration, image creation, etc, all **completely automated and extensible**

Cloud instances include all tools available in main Galaxy *and more*

Same automation approach can be used for configuring tool dependencies for a local Galaxy

VM image with just tools available, currently at http://s3.amazonaws.com/usegalaxy/UseGalaxy.ova

Why we love clouds and cloud-like things:

Reasonably cost effective and efficient
(elasticity + autoscaling definitely save money)

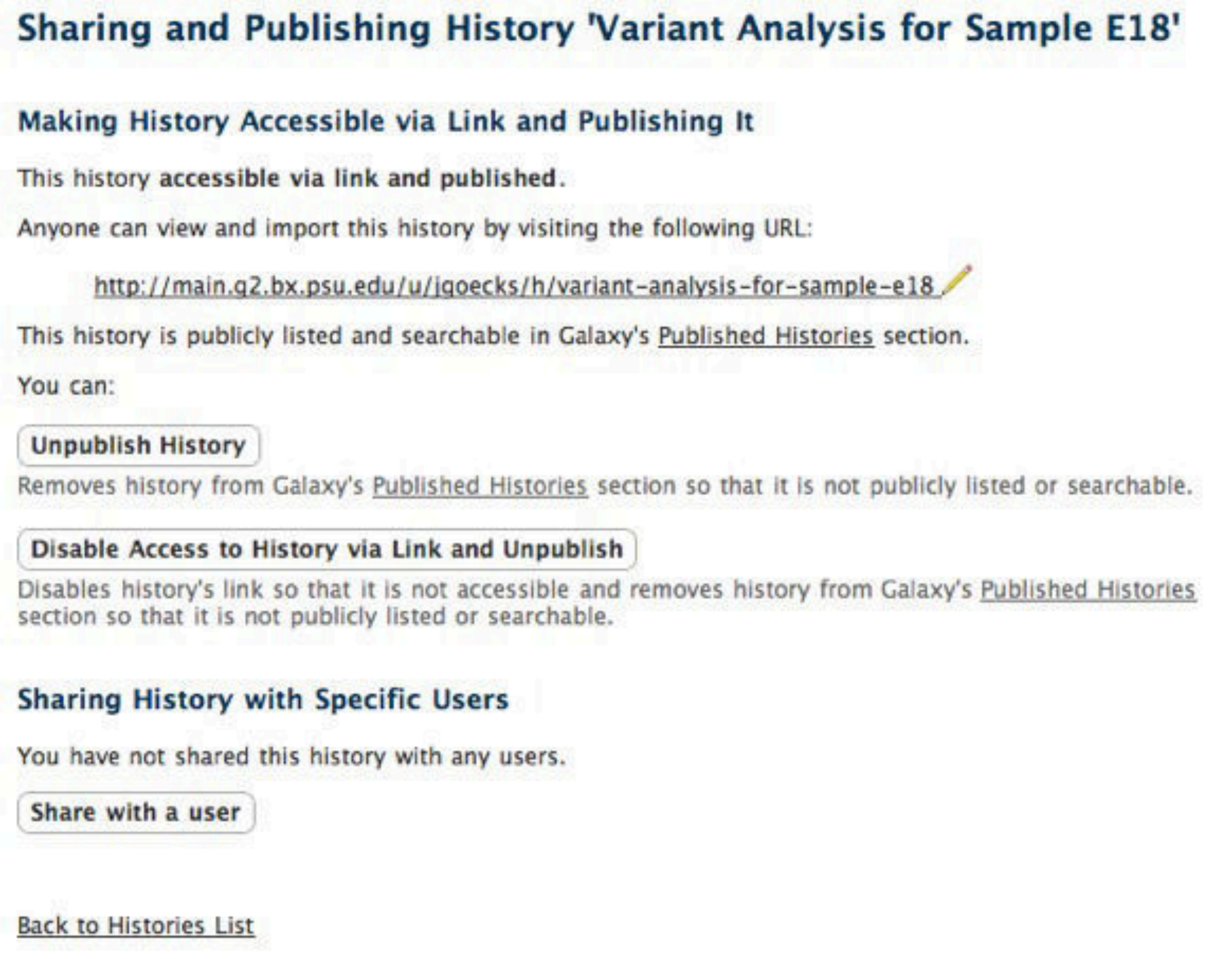
Analysis costs are more directly quantifiable

Infrastructure as an abstraction + standard APIs for provisioning reduces risk of vendor lock-in

Virtualization makes *so* many things easier

Publishing and sharing

Everything can be shared



Sharing and Publishing History 'Variant Analysis for Sample E18'

Making History Accessible via Link and Publishing It

This history **accessible via link and published**.

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

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

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

You can:

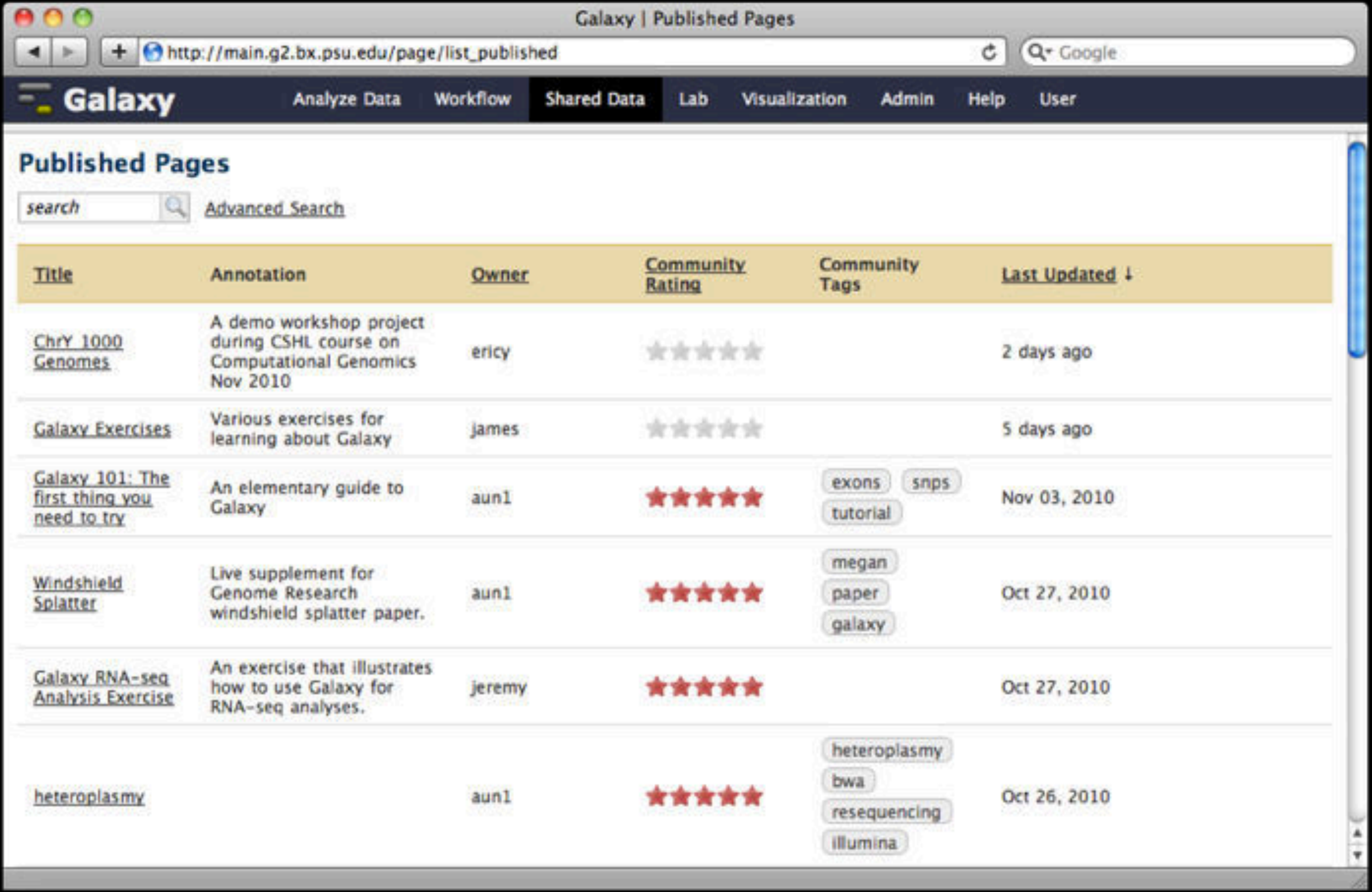
Removes history from Galaxy's Published Histories section so that it is not publicly listed or searchable.

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

Sharing History with Specific Users

You have not shared this history with any users.

Back to Histories List

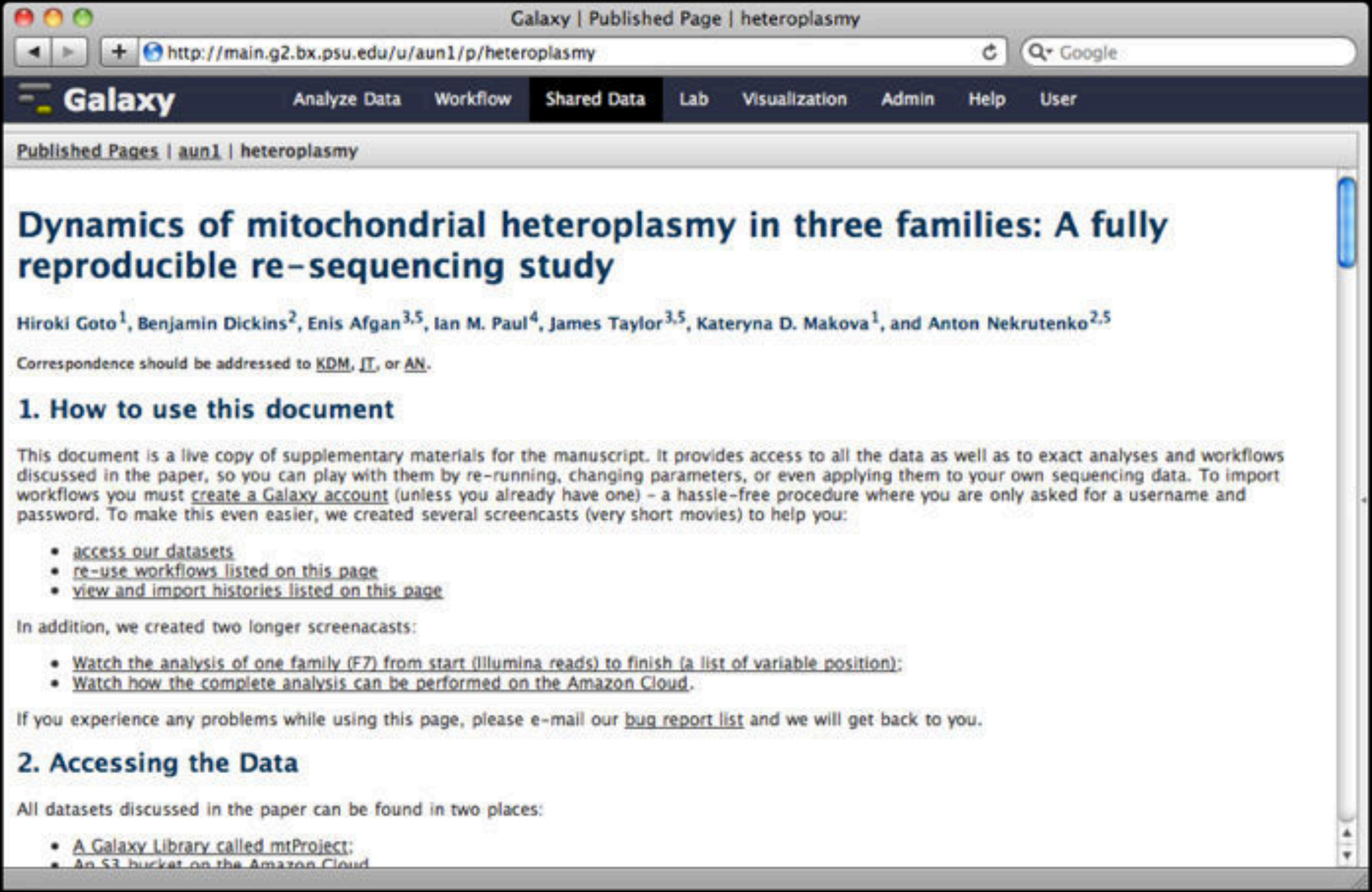


Published Pages

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

Pervasive search allows others to find published items of interest


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.

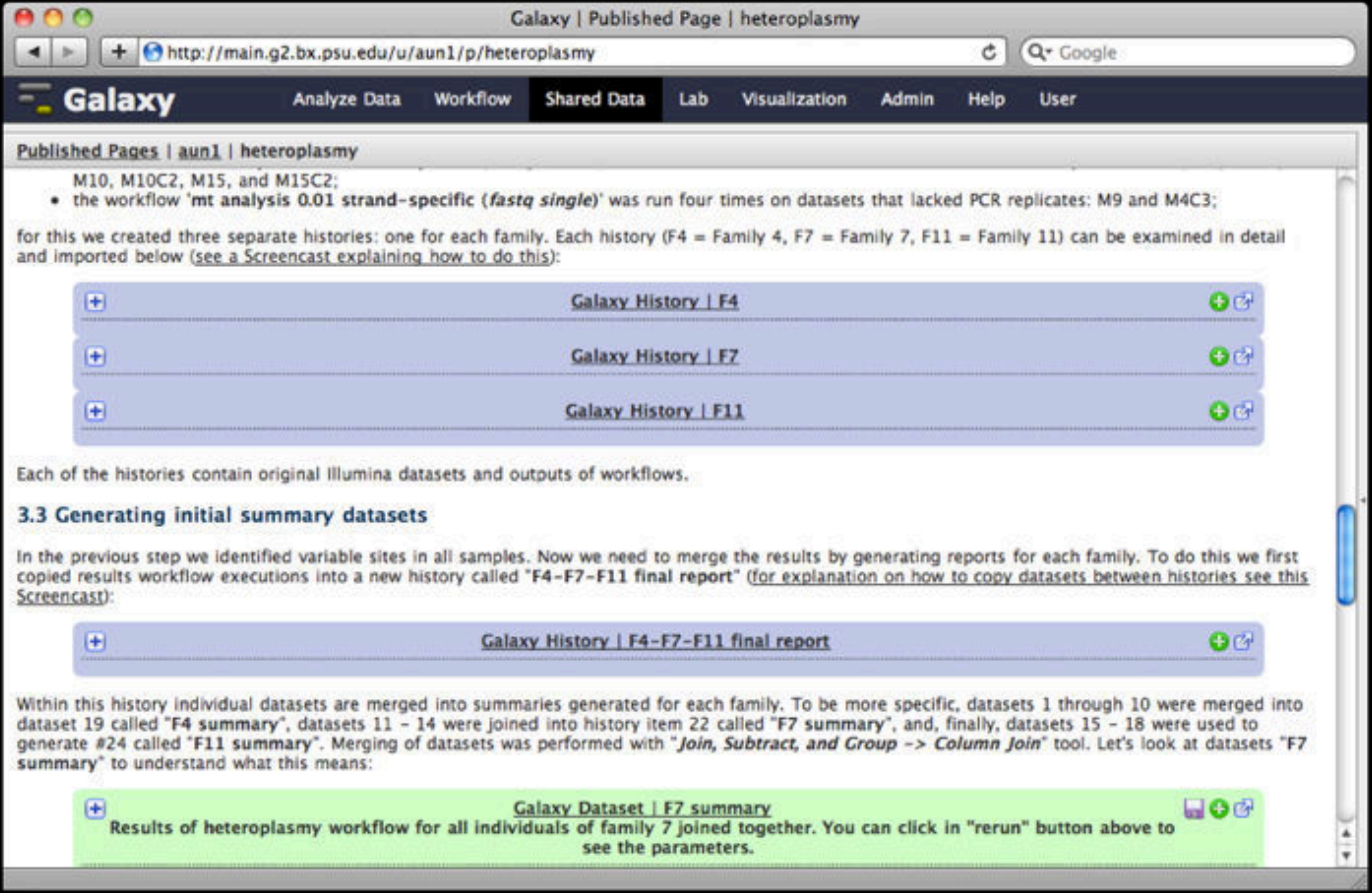
If you experience any problems while using this page, please e-mail our bug report list and we will get back to you.

2. Accessing the Data

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

- A Galaxy Library called mtProject;
- An S3 bucket on the Amazon Cloud

Galaxy Page for a recent study on mitochondrial heteroplasmy



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):

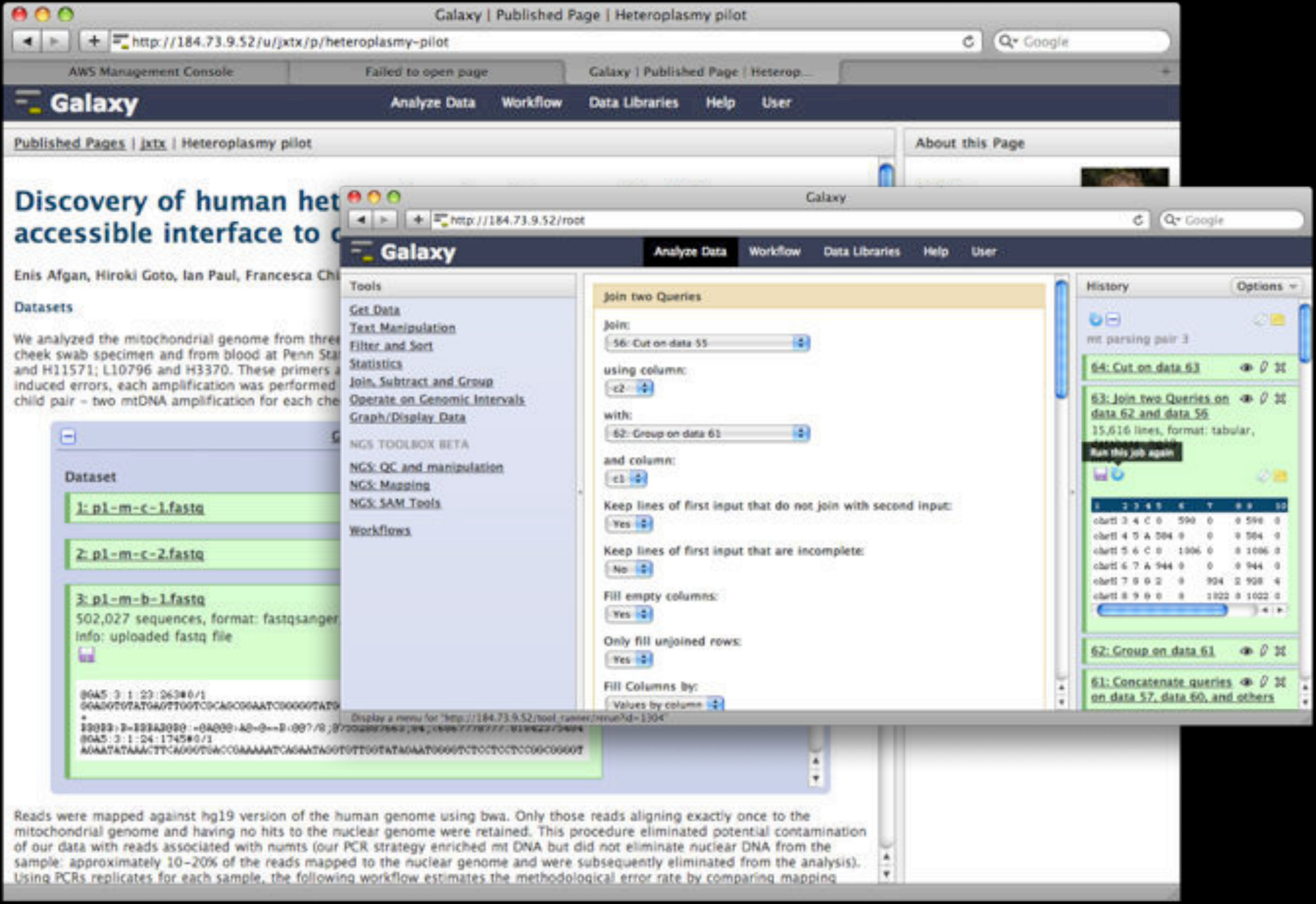
Galaxy History | F4-F7-F11 final report

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.

Actual histories and datasets directly accessible from the text



Galaxy | Published Page | Heteroplasmy pilot

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

Published Pages | jxtx | Heteroplasmy pilot

About this Page

Discovery of human het
accessible interface to c

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
+
...
@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

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

mt parsing pair 3

64: Cut on data 63

63: Join two Queries on data 62 and data 56
15,616 lines, format: tabular,

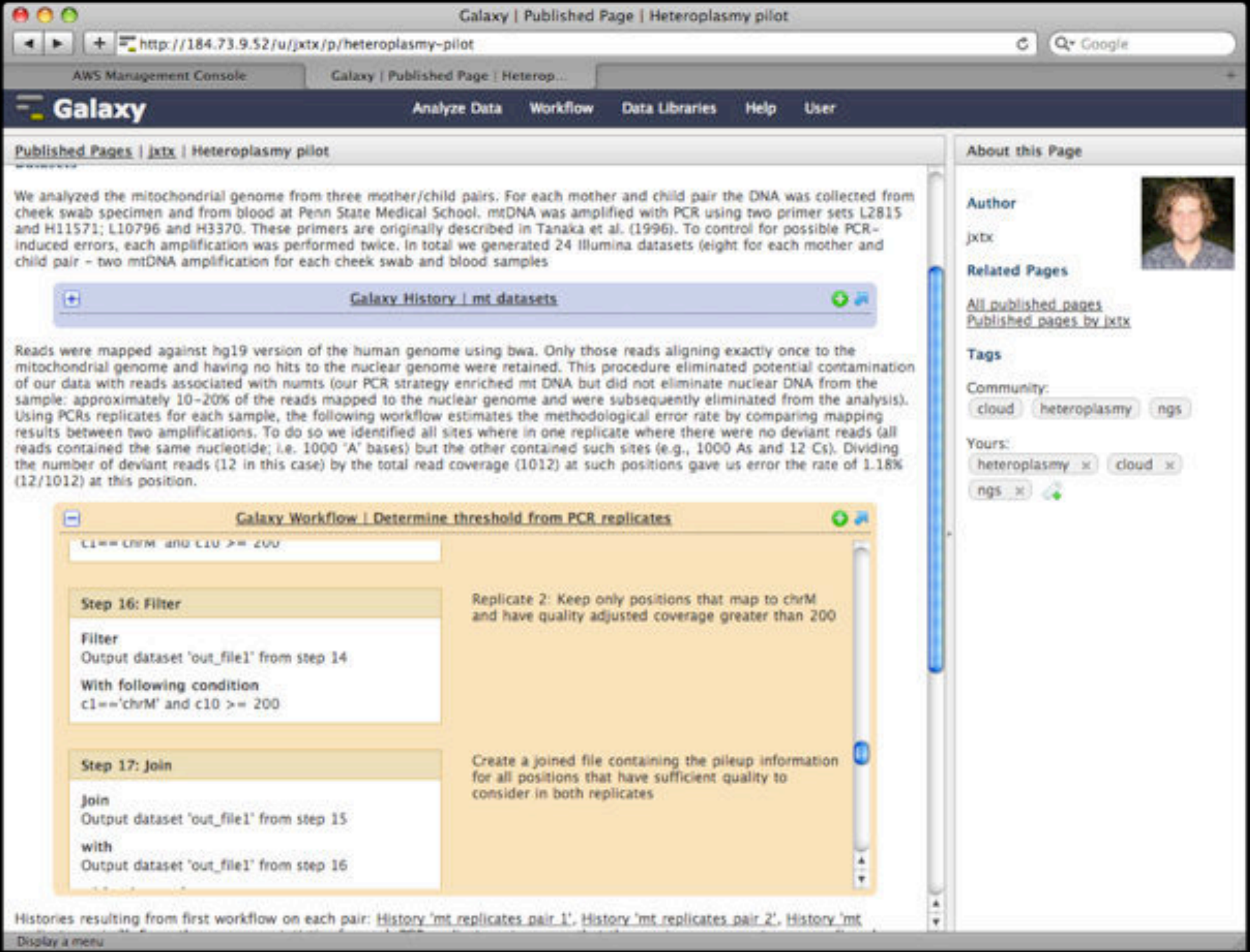
Run this job again

62: Group on data 61

61: Concatenate queries on data 57, data 60, and others

Display a menu for "http://184.73.9.52/tool_runner/rerun?id=1304"

Histories can be imported and the exact parameters inspected


Galaxy | Published Page | Heteroplasmy pilot
http://184.73.9.52/u/jxtx/p/heteroplasmy-pilot
Google
AWS Management Console
Galaxy | Published Page | Heterop...
Galaxy
Analyze Data
Workflow
Data Libraries
Help
User
Display a menu

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

About this Page

Author

jxtx

Related Pages

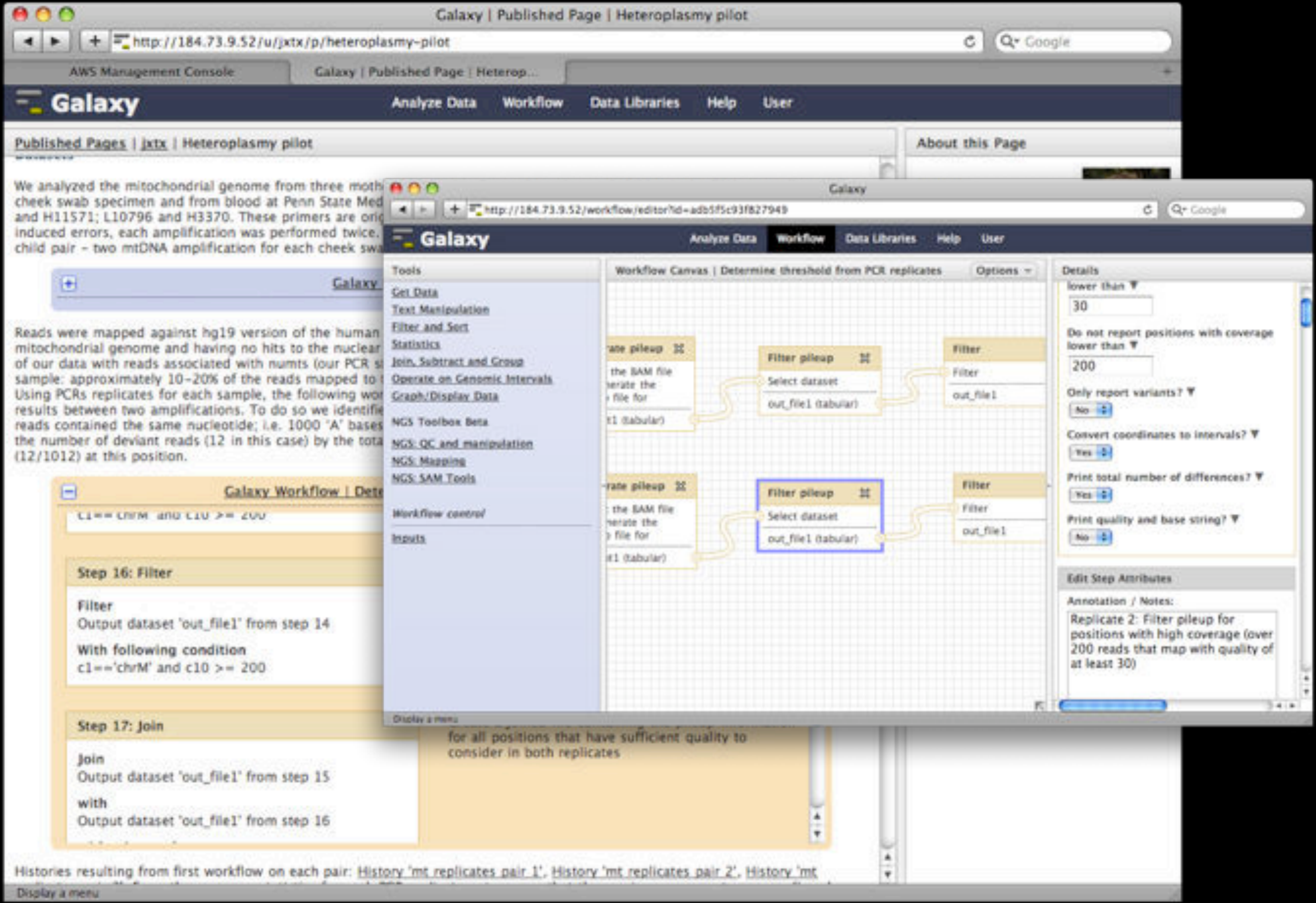
All published pages
Published pages by jxtx

Tags

Community:
cloud heteroplasmy ngs

Yours:
heteroplasmy × cloud × ngs ×

Workflows and other entities can also be embedded



We analyzed the mitochondrial genome from three moth
cheek swab specimen and from blood at Penn State Med
and H11571; L10796 and H3370. These primers are orig
induced errors, each amplification was performed twice.
child pair – two mtDNA amplification for each cheek swa

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 t
Using PCRs replicates for each sample, the following wor
results between two amplifications. To do so we identifie
reads contained the same nucleotide; i.e. 1000 'A' bases
the number of deviant reads (12 in this case) by the tota
(12/1012) at this position.

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

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

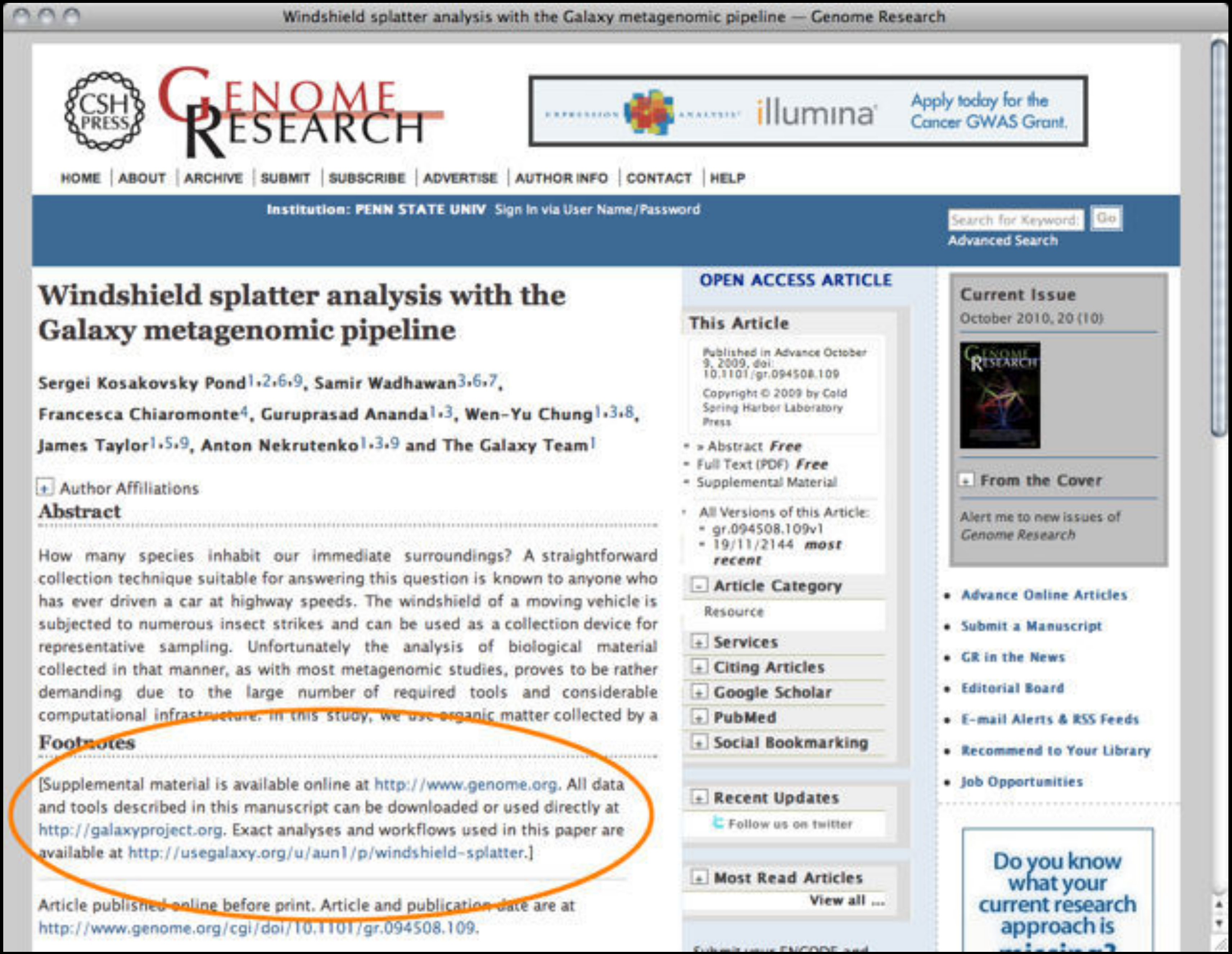
Print quality and base string? No

Annotation / Notes:
Replicate 2: Filter pileup for positions with high coverage (over 200 reads that map with quality of at least 30)

And imported for inspection, verification, and reuse

The power of Galaxy publishing

- 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


Windshield splatter analysis with the Galaxy metagenomic pipeline — Genome Research
CSH PRESS
GENOME RESEARCH
HOME | ABOUT | ARCHIVE | SUBMIT | SUBSCRIBE | ADVERTISE | AUTHOR INFO | CONTACT | HELP
Institution: PENN STATE UNIV Sign In via User Name/Password
Search for Keyword: Go
Advanced Search

Windshield splatter analysis with the Galaxy metagenomic pipeline

Sergei Kosakovsky Pond[1,2,6,9], Samir Wadhawan[3,6,7], Francesca Chiaromonte[4], Guruprasad Ananda[1,3], Wen-Yu Chung[1,3,8], James Taylor[1,5,9], Anton Nekrutenko[1,3,9] and The Galaxy Team[1]

⊞ Author Affiliations

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 windshield of a moving vehicle is 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://galaxyproject.org. Exact analyses and workflows used in this paper are available at http://usegalaxy.org/u/aun1/p/windshield-splatter.]

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.094508.109.

OPEN ACCESS ARTICLE

This Article

Published in Advance October 9, 2009, doi: 10.1101/gr.094508.109

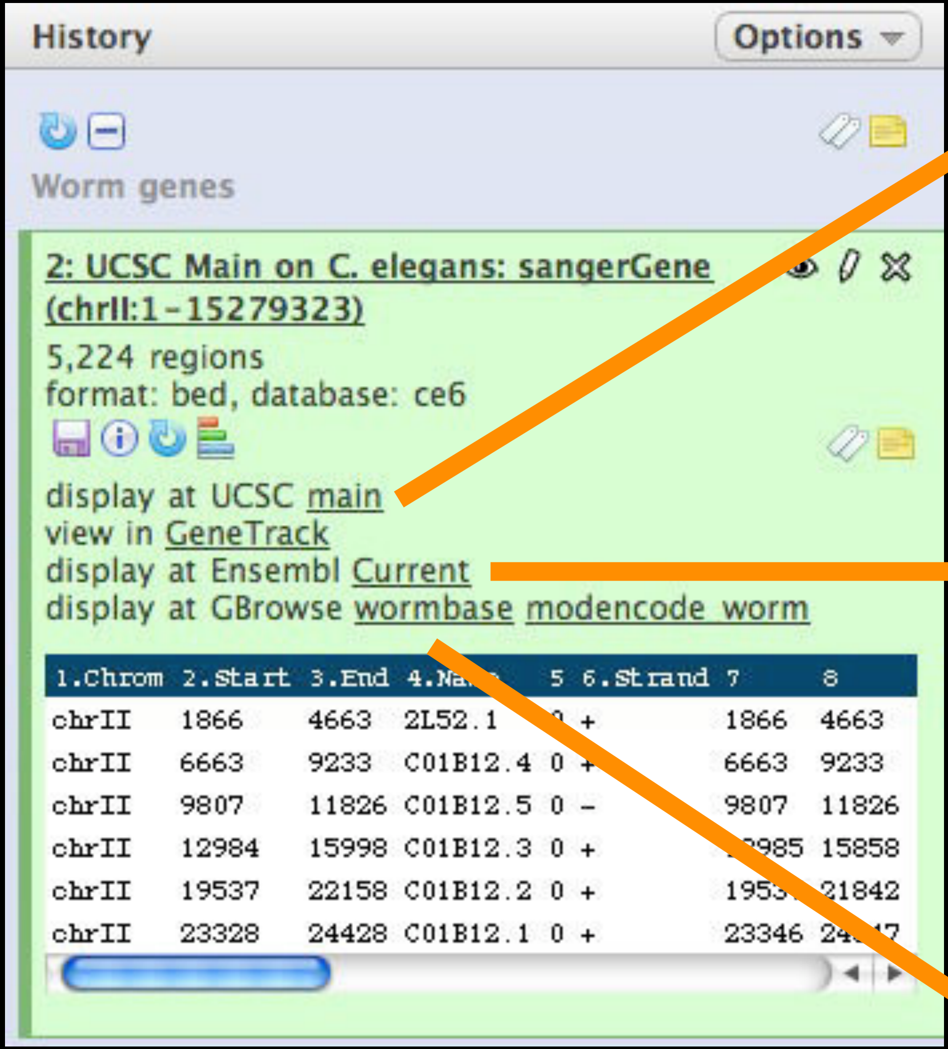
- » Abstract ***Free***
- Full Text (PDF) ***Free***
- Supplemental Material

- All Versions of this Article:
  - gr.094508.109v1
  - 19/11/2144 ***most recent***

Current Issue

October 2010, 20 (10)

Visualization



History

Options

Worm genes

2: UCSC Main on C. elegans: sangerGene (chrII:1-15279323)

5,224 regions

format: bed, database: ce6

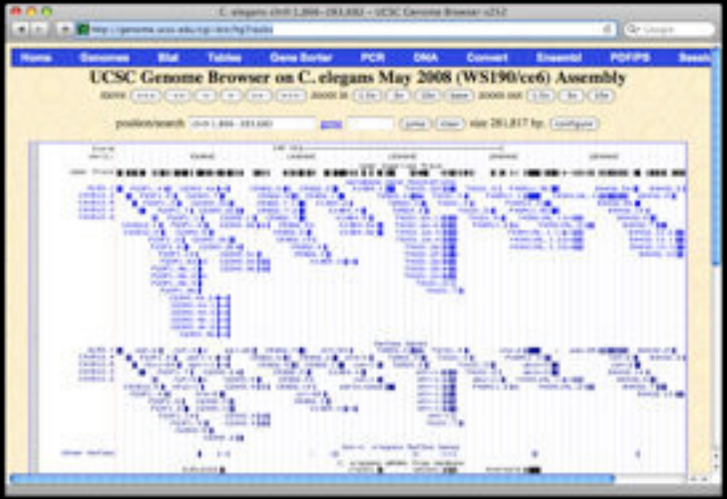
display at UCSC main

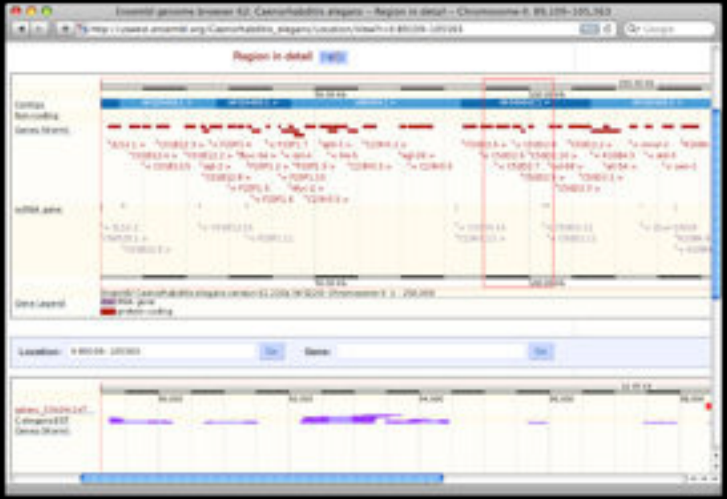
view in GeneTrack

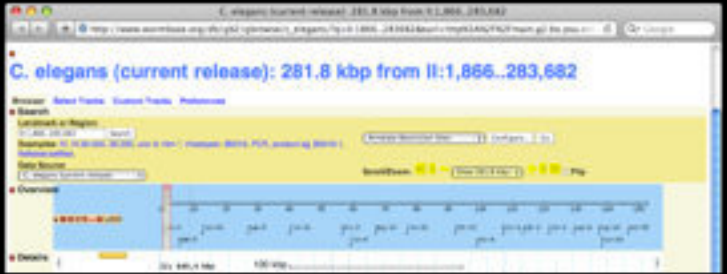
display at Ensembl Current

display at GBrowse wormbase modencode worm

1.Chrom	2.Start	3.End	4.Name	5	6.Strand	7	8
chrII	1866	4663	2L52.1	0	+	1866	4663
chrII	6663	9233	C01B12.4	0	+	6663	9233
chrII	9807	11826	C01B12.5	0	-	9807	11826
chrII	12984	15998	C01B12.3	0	+	12985	15858
chrII	19537	22158	C01B12.2	0	+	19537	21842
chrII	23328	24428	C01B12.1	0	+	23346	24417


UCSC Genome Browser on C. elegans May 2008 (WS190/ce6) Assembly




C. elegans (current release): 281.8 kbp from II:1,866..283,682

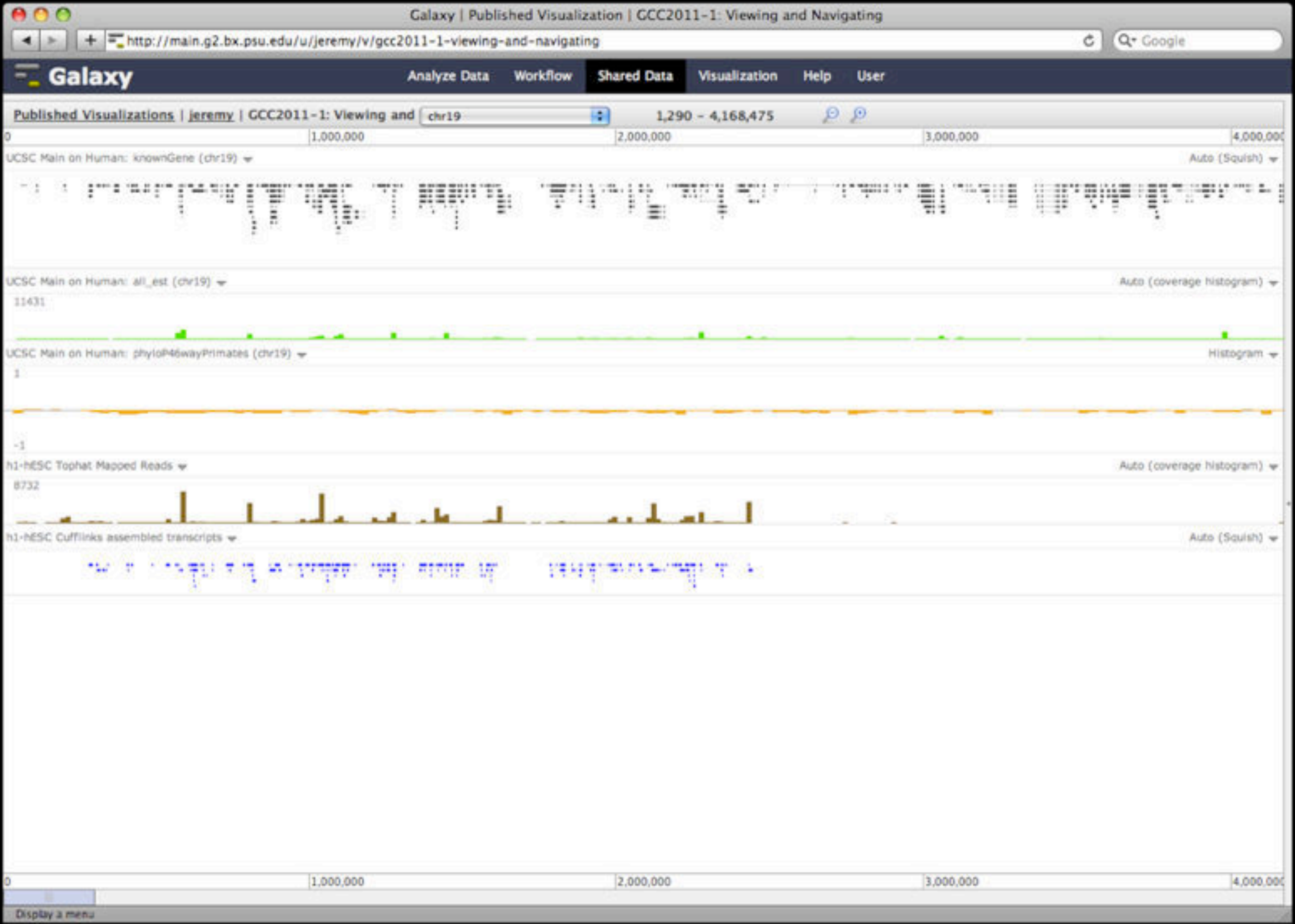
Integration with many existing browsers (extensible)

Visualization and analytics: Galaxy Track Browser

Entirely web standards based to support sharing, communicating, and collaborating around visualizations

Dynamic and responsive

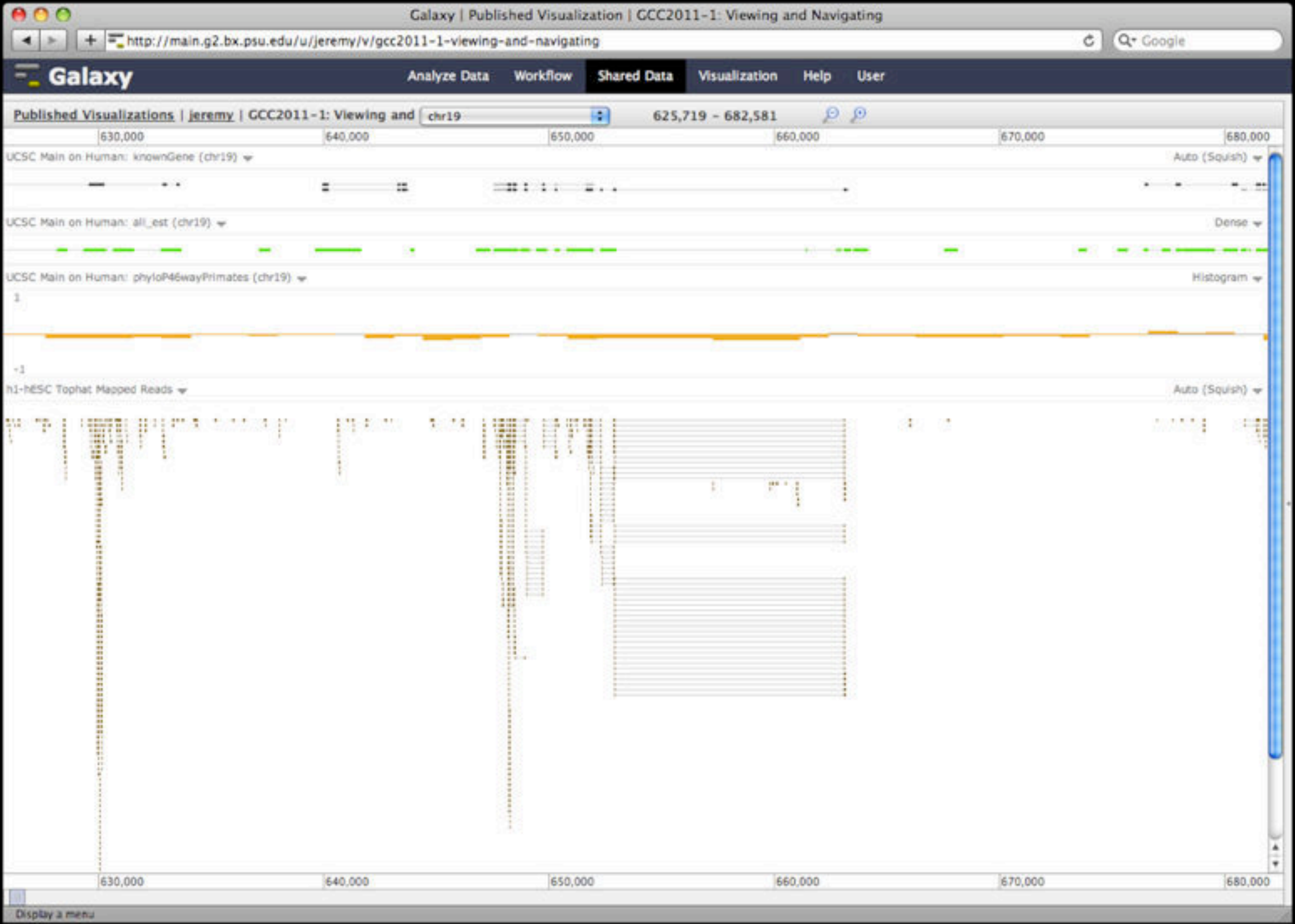
Open source and extremely extensible



Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-1: Viewing and
chr19
1,290 - 4,168,475
0
1,000,000
2,000,000
3,000,000
4,000,00
UCSC Main on Human: knownGene (chr19)
Auto (Squish)
UCSC Main on Human: all_est (chr19)
Auto (coverage histogram)
11431
UCSC Main on Human: phyloP46wayPrimates (chr19)
Histogram
1
-1
h1-hESC Tophat Mapped Reads
Auto (coverage histogram)
8732
h1-hESC Cufflinks assembled transcripts
Auto (Squish)
0
1,000,000
2,000,000
3,000,000
4,000,00
Display a menu



Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
630,000
640,000
650,000
660,000
670,000
680,000
1
-1
Display a menu

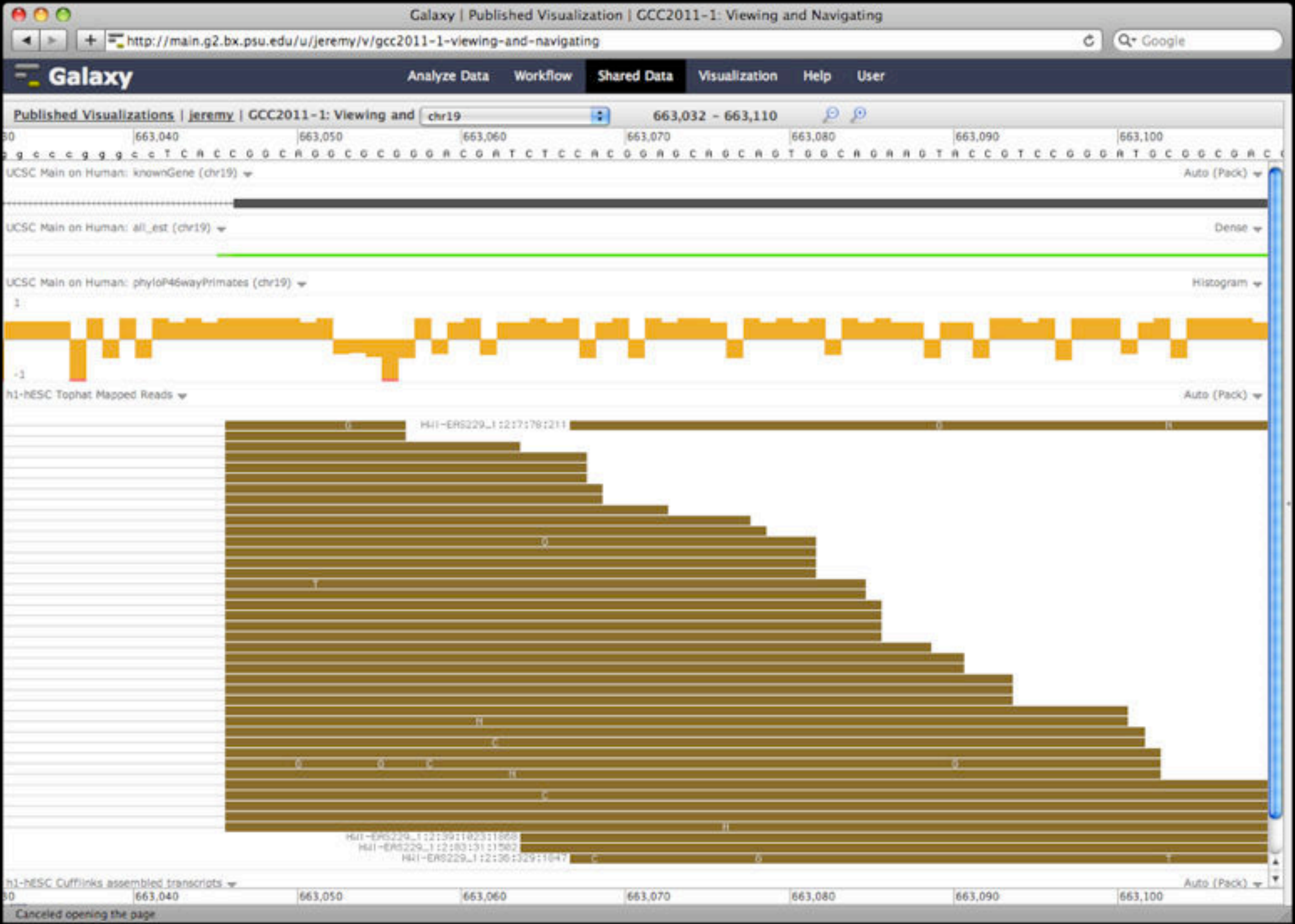
Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 625,719 - 682,581

UCSC Main on Human: knownGene (chr19) Auto (Squish)

UCSC Main on Human: all_est (chr19) Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) Histogram

h1-hESC Tophat Mapped Reads Auto (Squish)



Galaxy | Published Visualization | GCC2011-1: Viewing and Navigating

http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-1-viewing-and-navigating

Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User

Published Visualizations | jeremy | GCC2011-1: Viewing and chr19 663,032 - 663,110

663,040
663,050
663,060
663,070
663,080
663,090
663,100
g g c c c g g g c c T C A C C G G C A G G C G C G G G A C G A T C T C C A C G G A G C A G C A G T G G C A G A A G T A C C G T C C G G G A T G C G G C G A C

UCSC Main on Human: knownGene (chr19) — Auto (Pack)

UCSC Main on Human: all_est (chr19) — Dense

UCSC Main on Human: phyloP46wayPrimates (chr19) — Histogram

h1-hESC Tophat Mapped Reads — Auto (Pack)

HWI-EAS229_1:2:17:78:211
HWI-EAS229_1:2:39:1023:1868
HWI-EAS229_1:2:83:31:1502
HWI-EAS229_1:2:36:329:1047

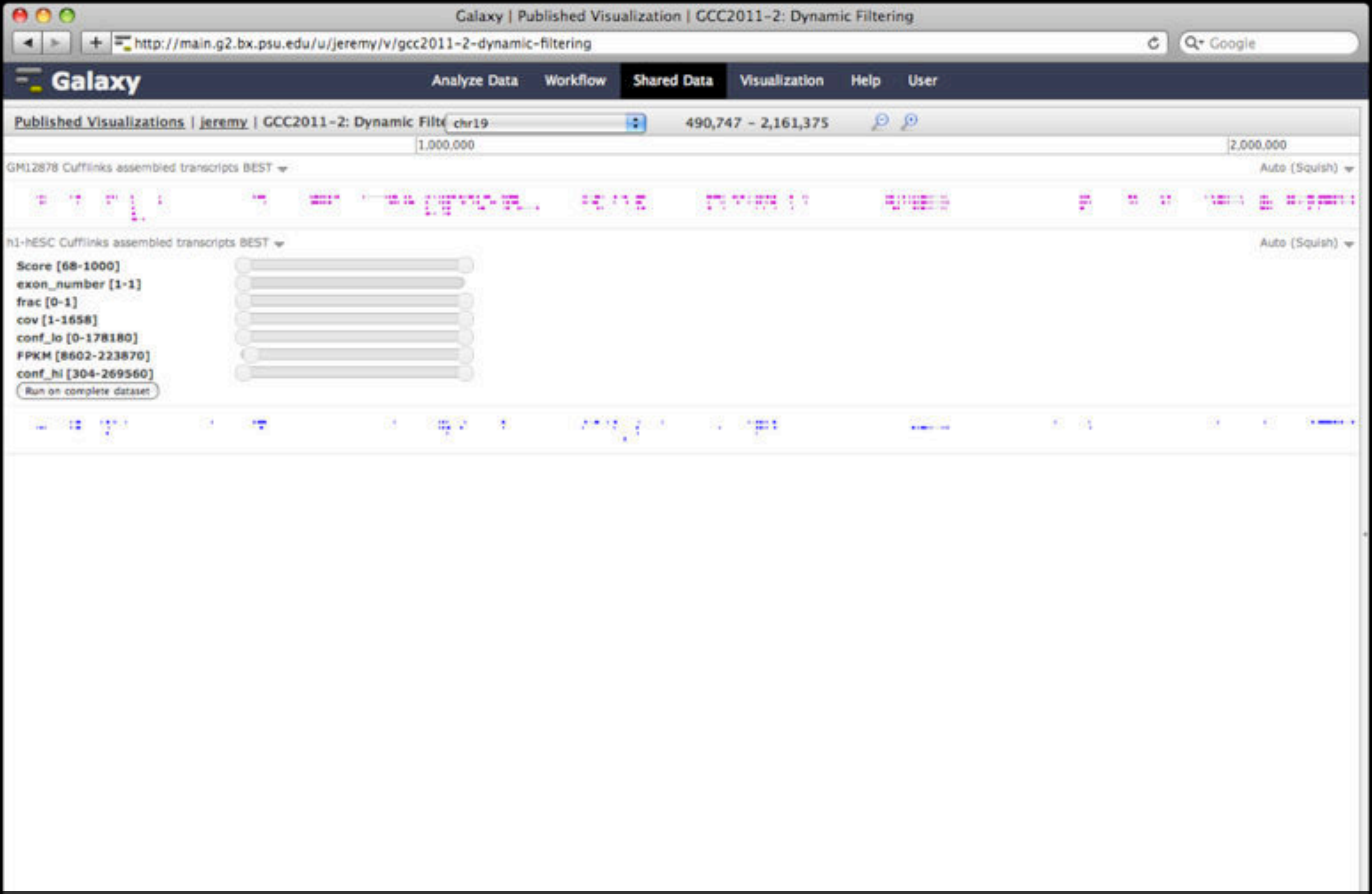
h1-hESC Cufflinks assembled transcripts — Auto (Pack)

Canceled opening the page

With increasingly complex tools, more experimentation with parameters is necessary, visual feedback aids exploration

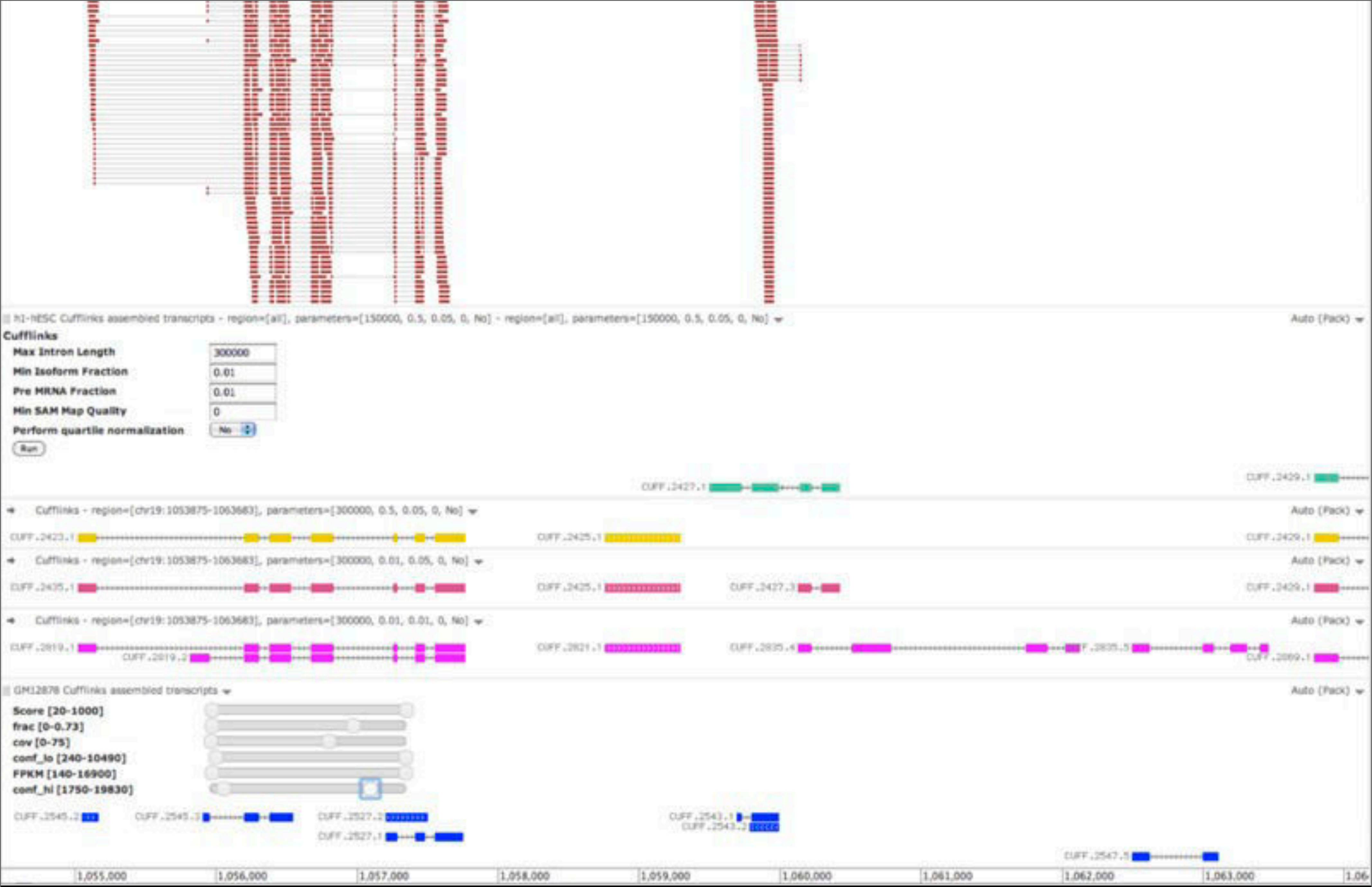
Galaxy already provides a very sound model for abstracting interfaces to analysis tools

Existing tool framework can be leveraged for **visual analytics**


Galaxy | Published Visualization | GCC2011-2: Dynamic Filtering
http://main.g2.bx.psu.edu/u/jeremy/v/gcc2011-2-dynamic-filtering
Google
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Published Visualizations | jeremy | GCC2011-2: Dynamic Filte
chr19
490,747 - 2,161,375
1,000,000
2,000,000
GM12878 Cufflinks assembled transcripts BEST
Auto (Squish)
h1-hESC Cufflinks assembled transcripts BEST
Auto (Squish)
Run on complete dataset

Score [68-1000]
exon_number [1-1]
frac [0-1]
cov [1-1658]
conf_lo [0-178180]
FPKM [8602-223870]
conf_hi [304-269560]

Dynamic filtering on element properties (here, FPKM for putative transcripts)



h1-hESC Cufflinks assembled transcripts - region=[all], parameters=[150000, 0.5, 0.05, 0, No] - region=[all], parameters=[150000, 0.5, 0.05, 0, No]

Cufflinks

Max Intron Length 300000

Min Isoform Fraction 0.01

Pre MRNA Fraction 0.01

Min SAM Map Quality 0

Perform quartile normalization No

Cufflinks - region=[chr19:1053875-1063683], parameters=[300000, 0.5, 0.05, 0, No]

Cufflinks - region=[chr19:1053875-1063683], parameters=[300000, 0.01, 0.05, 0, No]

Cufflinks - region=[chr19:1053875-1063683], parameters=[300000, 0.01, 0.01, 0, No]

GM12878 Cufflinks assembled transcripts

Score [20-1000]

frac [0-0.73]

cov [0-75]

conf_lo [240-10490]

FPKM [140-16900]

conf_hi [1750-19830]

Modifying Cufflinks parameters and locally reassembling

Arbitrary visualization types supported
(but not implemented)

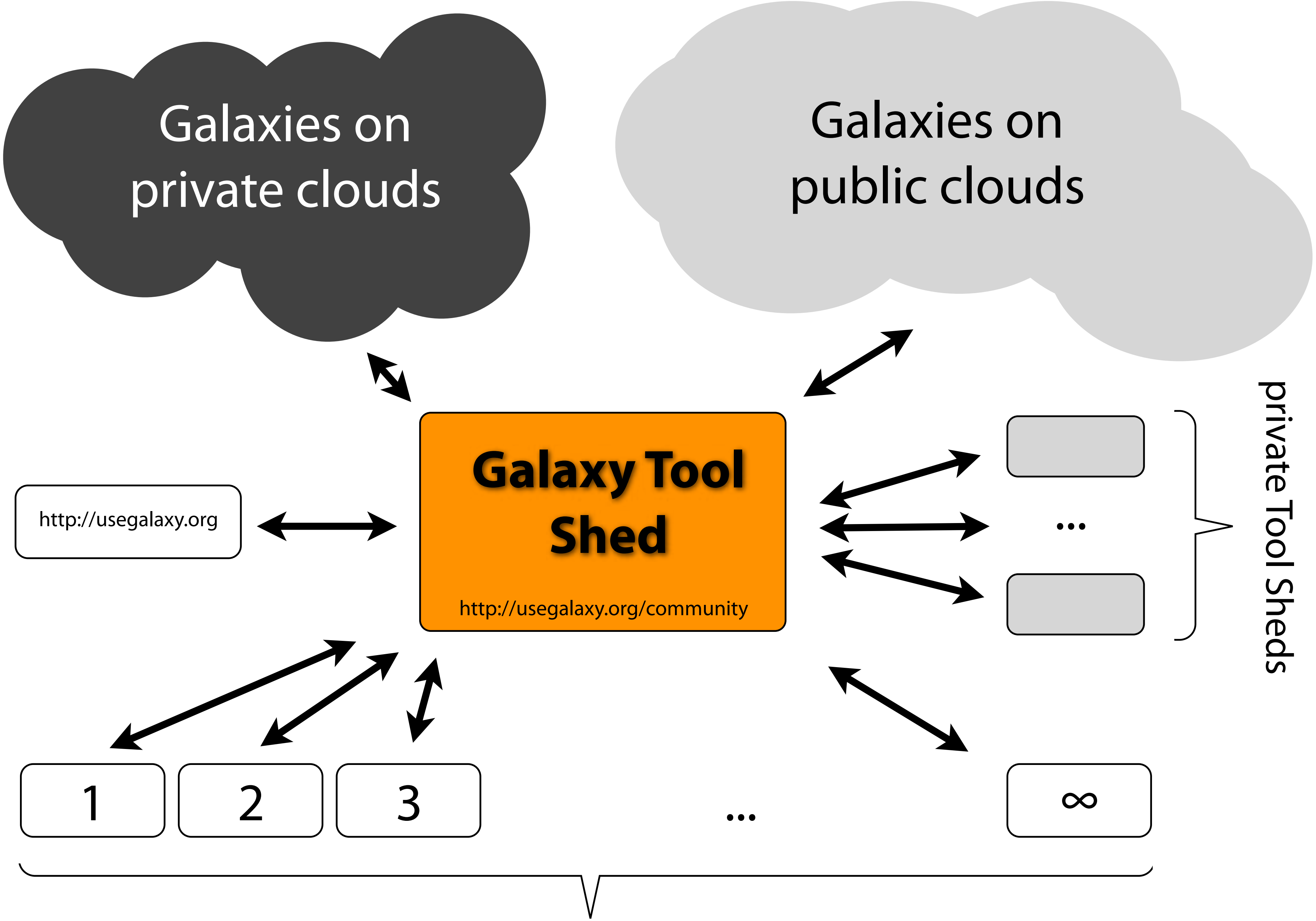
Access to tools and visual analytics specific features
(e.g. local computation using global models)
can be used by new visualization types

Scaling Galaxy: two distinct problems

- So much data, not enough infrastructure.
  - Solution, encourage local Galaxy instances, cloud Galaxy, support increasingly decentralized model, *improve access to exiting resources*
- So many tools and workflows, not enough manpower
  - Focus on building infrastructure to allow community to integrate and share tools, workflows, and best practices

Galaxy toolshed vision

- Allow users to share “suites” containing tools, datatypes, workflows, sample data, and automated installation scripts for tool dependencies
- Version controlled
- Community annotation, rating, comments, review
- Dependency resolution
- Integration with Galaxy instances to automate tool installation and updates



Galaxies on private clouds

Galaxies on public clouds

http://usegalaxy.org

Galaxy Tool Shed

http://usegalaxy.org/community

...

private Tool Sheds

1

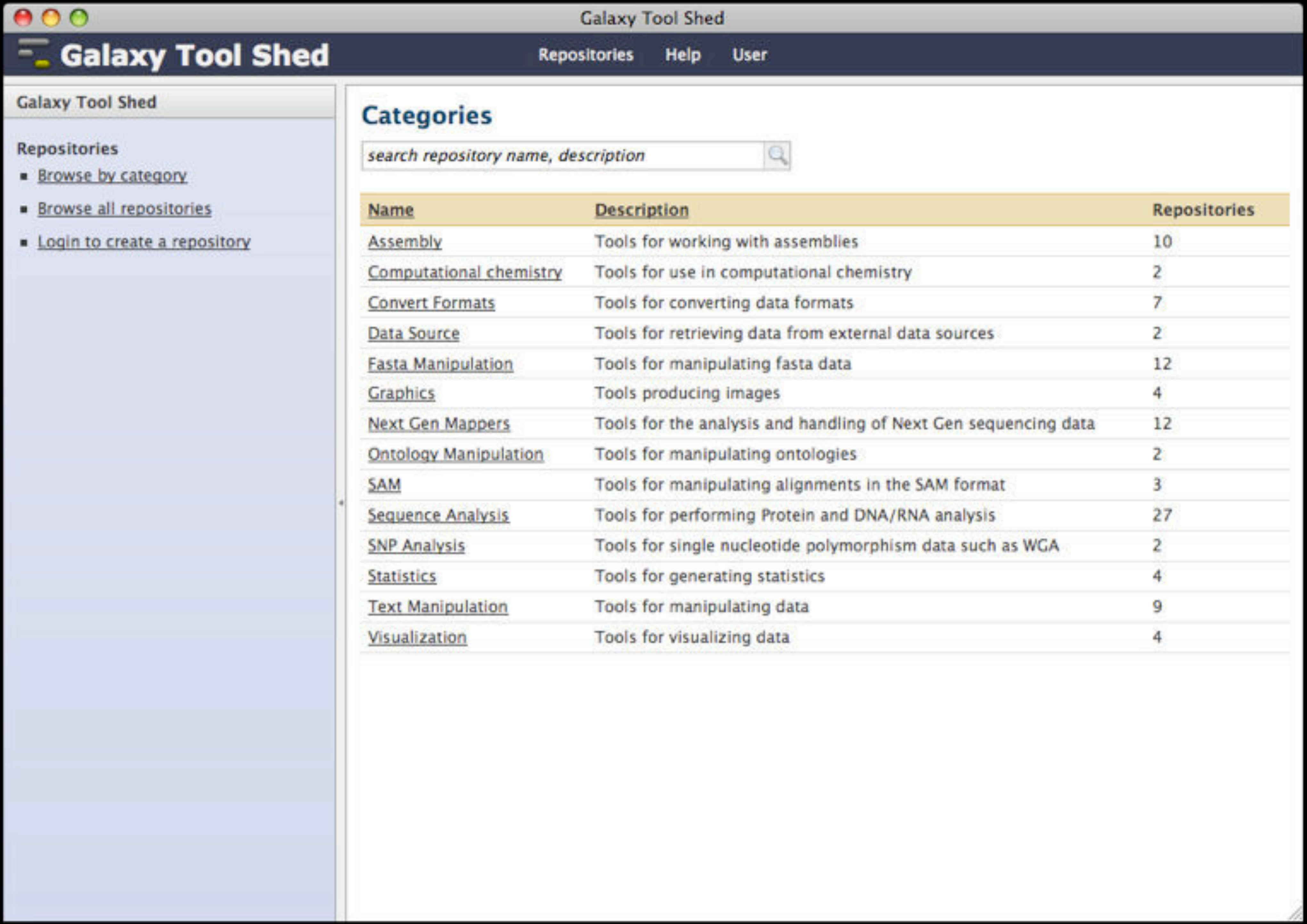
2

3

...

∞

private Galaxy installations



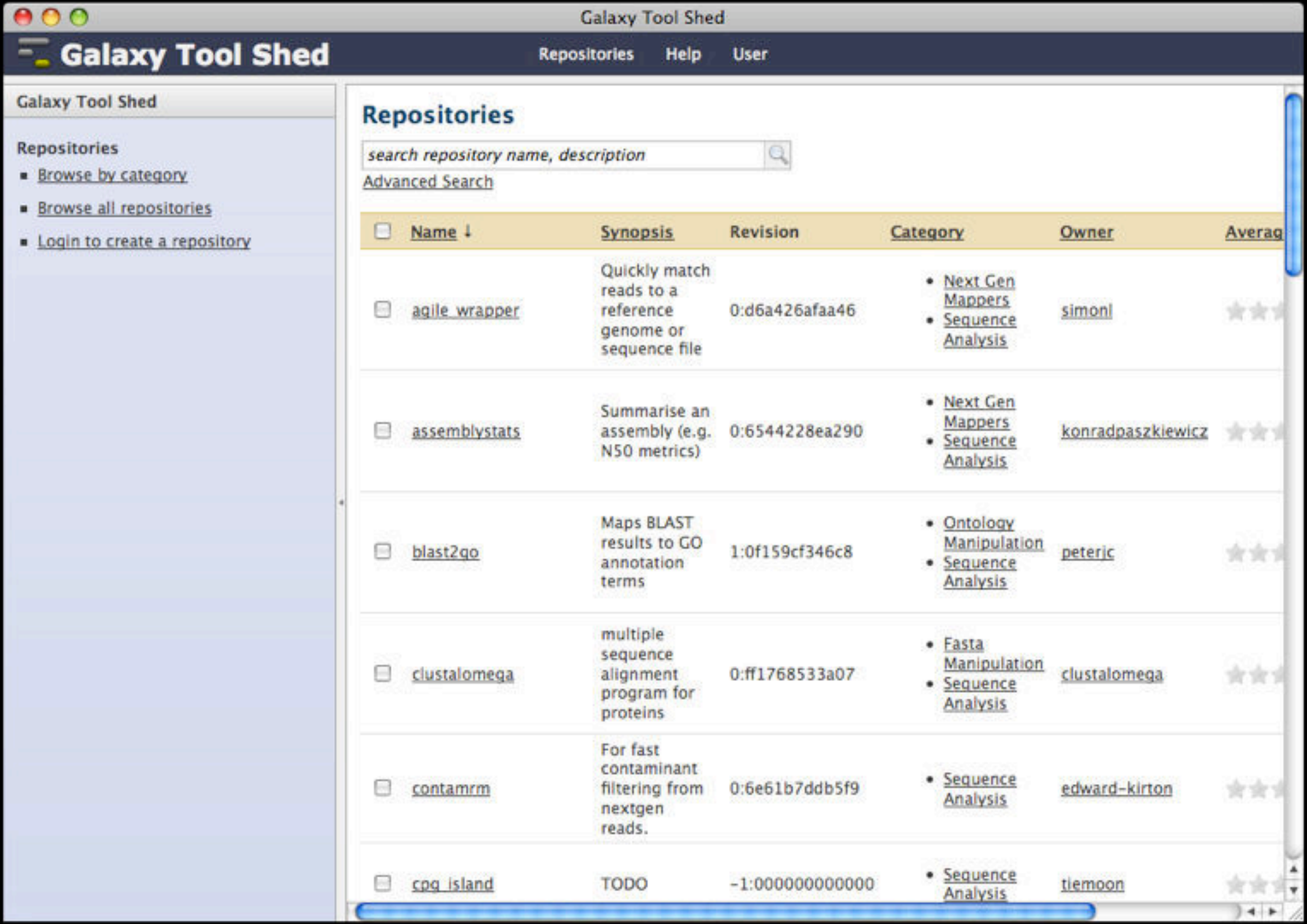
Repositories

- Browse by category
- Browse all repositories
- Login to create a repository

Categories

search repository name, description

Name	Description	Repositories
Assembly	Tools for working with assemblies	10
Computational chemistry	Tools for use in computational chemistry	2
Convert Formats	Tools for converting data formats	7
Data Source	Tools for retrieving data from external data sources	2
Fasta Manipulation	Tools for manipulating fasta data	12
Graphics	Tools producing images	4
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data	12
Ontology Manipulation	Tools for manipulating ontologies	2
SAM	Tools for manipulating alignments in the SAM format	3
Sequence Analysis	Tools for performing Protein and DNA/RNA analysis	27
SNP Analysis	Tools for single nucleotide polymorphism data such as WGA	2
Statistics	Tools for generating statistics	4
Text Manipulation	Tools for manipulating data	9
Visualization	Tools for visualizing data	4



Repositories

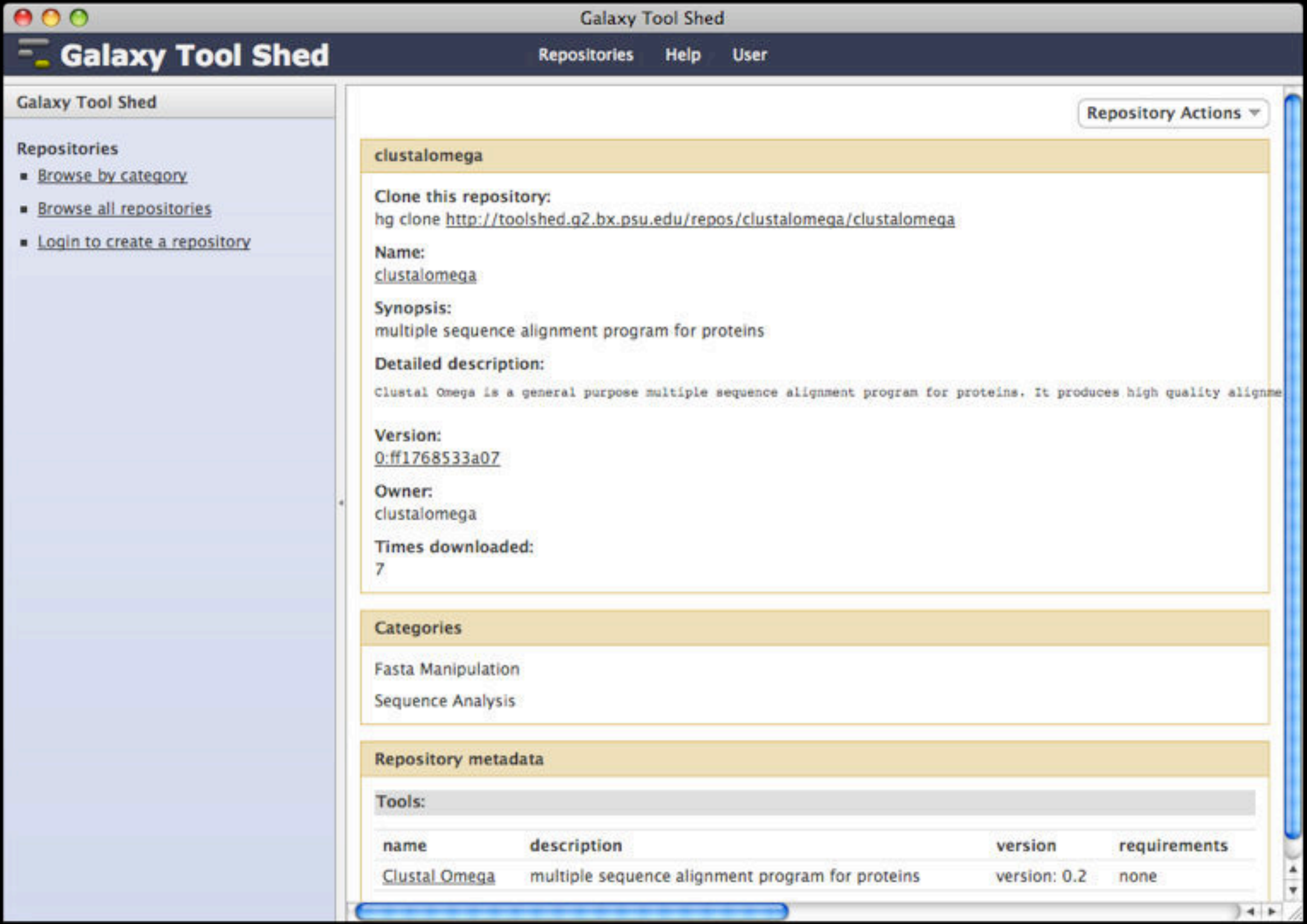
- Browse by category
- Browse all repositories
- Login to create a repository

Repositories

search repository name, description

Advanced Search

□	Name ↓	Synopsis	Revision	Category	Owner	Averag
□	agile_wrapper	Quickly match reads to a reference genome or sequence file	0:d6a426afaa46	• Next Gen Mappers • Sequence Analysis	simonl	☆☆☆
□	assemblystats	Summarise an assembly (e.g. N50 metrics)	0:6544228ea290	• Next Gen Mappers • Sequence Analysis	konradpaszkiewicz	☆☆☆
□	blast2go	Maps BLAST results to GO annotation terms	1:0f159cf346c8	• Ontology Manipulation • Sequence Analysis	peterjc	☆☆☆
□	clustalomega	multiple sequence alignment program for proteins	0:ff1768533a07	• Fasta Manipulation • Sequence Analysis	clustalomega	☆☆☆
□	contamrm	For fast contaminant filtering from nextgen reads.	0:6e61b7ddb5f9	• Sequence Analysis	edward-kirton	☆☆☆
□	cpg_island	TODO	-1:000000000000	• Sequence Analysis	tiemoon	☆☆☆



clustalomega

Clone this repository:
hg clone http://toolshed.g2.bx.psu.edu/repos/clustalomega/clustalomega

Name:
clustalomega

Synopsis:
multiple sequence alignment program for proteins

Detailed description:

```
Clustal Omega is a general purpose multiple sequence alignment program for proteins. It produces high quality alignme
```

Version:
0:ff1768533a07

Owner:
clustalomega

Times downloaded:
7

Categories

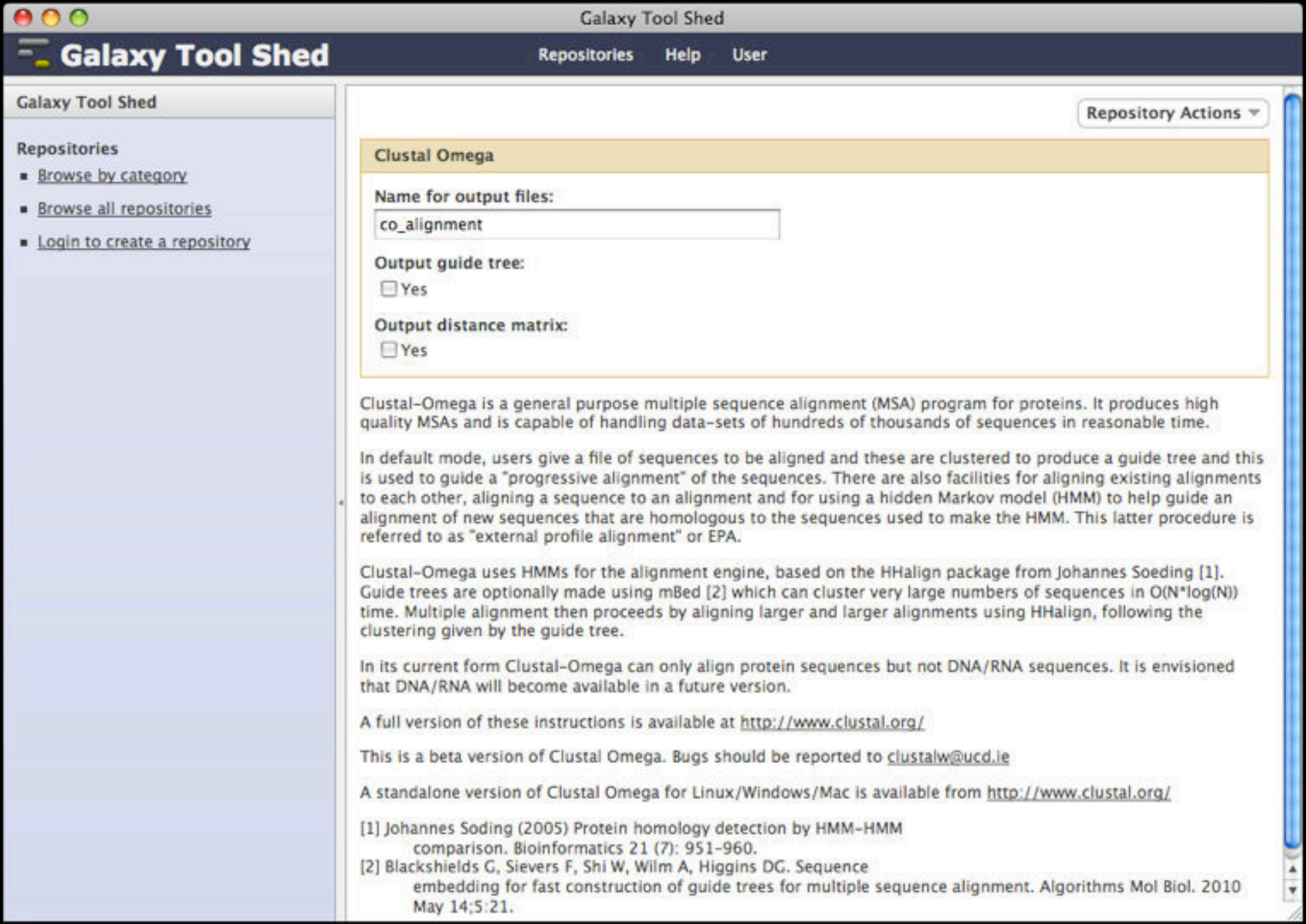
Fasta Manipulation

Sequence Analysis

Repository metadata

Tools:

name	description	version	requirements
Clustal Omega	multiple sequence alignment program for proteins	version: 0.2	none



Clustal Omega

Name for output files:

co_alignment

Output guide tree:

□ Yes

Output distance matrix:

□ Yes

Clustal-Omega is a general purpose multiple sequence alignment (MSA) program for proteins. It produces high quality MSAs and is capable of handling data-sets of hundreds of thousands of sequences in reasonable time.

In default mode, users give a file of sequences to be aligned and these are clustered to produce a guide tree and this is used to guide a "progressive alignment" of the sequences. There are also facilities for aligning existing alignments to each other, aligning a sequence to an alignment and for using a hidden Markov model (HMM) to help guide an alignment of new sequences that are homologous to the sequences used to make the HMM. This latter procedure is referred to as "external profile alignment" or EPA.

Clustal-Omega uses HMMs for the alignment engine, based on the HHalign package from Johannes Soeding [1]. Guide trees are optionally made using mBed [2] which can cluster very large numbers of sequences in O(N*log(N)) time. Multiple alignment then proceeds by aligning larger and larger alignments using HHalign, following the clustering given by the guide tree.

In its current form Clustal-Omega can only align protein sequences but not DNA/RNA sequences. It is envisioned that DNA/RNA will become available in a future version.

A full version of these instructions is available at http://www.clustal.org/

This is a beta version of Clustal Omega. Bugs should be reported to clustalw@ucd.ie

A standalone version of Clustal Omega for Linux/Windows/Mac is available from http://www.clustal.org/

[1] Johannes Soding (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7): 951-960.
[2] Blackshields G, Sievers F, Shi W, Wilm A, Higgins DG. Sequence embedding for fast construction of guide trees for multiple sequence alignment. Algorithms Mol Biol. 2010 May 14;5:21.

Some future challenges

- Capturing and automatically deploying tool dependencies, automatic tool acquisition in Galaxy instances

- Better interfaces for highly parallel analysis (e.g. running the same workflow across 192 individuals)

- Various workflow engine improvements, partial data streaming, combined experimental/ computational workflows

Try it now:

http://usegalaxy.org

Develop and deploy:

http://getgalaxy.org

Join us, contact me at:

james@jamestaylor.org

Opportunities for collaboration, positions for postdocs, researchers, software engineers