

mRNA-Seq

- Basic processing
- Read mapping (shown here, but optional. May due if time allows)
 - *Tophat*
- Gene expression estimation
 - *cufflinks*
 - Confidence intervals
- Gene expression changes (*separate use case*)
 - Sample groups
 - *cuffdiff*

Use case of RNA-Seq tools

- 2 breast cancer cell lines
 - Joe Gray 51 breast cancer cell lines panel
 - Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, Fevr T, Clark L, Bayani N, Coppe JP, Tong F, Speed T, Spellman PT, DeVries S, Lapuk A, Wang NJ, Kuo WL, Stilwell JL, Pinkel D, Albertson DG, Waldman FM, McCormick F, Dickson RB, Johnson MD, Lippman M, Ethier S, Gazdar A, Gray JW. “A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes.” ***Cancer Cell***. 2006 Dec;10(6):515-27.

Evaluating Gene Expression Differences

Sample	Luminal/Basal	ER status	PR status	Her2/ERBB2 status
BT474	Luminal	+	-	+
HCC1143	BasalA	-	-	-

Run Cufflinks in The Genboree Workbench

System/Network Data QC and Pre-processing Genome Transcriptome Cistrome Epigenome Metagenome Visual

Welcome to the Genboree Workbench

**Populate Input Data with the accepted_hits.bam file.
Populate Output Targets with your destination database**

Data Selector

Refresh Data Filter: Select a filter...

- Epigenome Informatics Workshop (May 2012)
 - Epigenome ToolSet Demo Input Data
 - Databases
 - Binding Sites Demo
 - Brain
 - Breast
 - Breast 450K
 - Disease Epigenome
 - MeDIP and GSEA
 - Peak Calling Demo
 - RNA-Seq Tool Demo
 - All Annotations in Database
 - Tracks
 - Lists & Selections
 - SampleSets
 - Samples
 - Files
 - TopHat-BT474_accepted_hits.bam**
 - mRNA.subset
 - TopHat-HCC1143_accepted_hits.bam
 - Queries

Details

Attribute	Value
Download	Click to Download File
Group	Epigenome ToolSet Demo Input Data
Database	RNA-Seq Tool Demo
Description	
Name	TopHat-

Input Data

↑ ↓ ✕ 📄

TopHat-BT474_accepted_hits.bam

Output Targets

↑ ↓ ✕ 📄

GenboreeUser_database

Drag

Select Transcriptome → Analyze RNA-Seq Data → Assemble and Measure Transcripts by Cufflinks

The screenshot displays the Genboree web interface. At the top, a navigation bar includes tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome, and Metagenome. A red arrow points from the text 'Select Transcriptome → Analyze RNA-Seq Data → Assemble and Measure Transcripts by Cufflinks' to the 'Transcriptome' tab. Below the navigation bar, a 'Welcome' message is visible. The 'Data Selector' section on the left shows a tree view of data sources, including 'www.genboree.org', 'Atlas Tools Access', 'EDACC', 'Epigenome Informatics Workshop (May 2012)', 'Epigenome ToolSet Demo Input Data', 'Epigenomics Roadmap Repository', and 'GenboreeUser_group'. The 'GenboreeUser_group' is expanded, showing 'Databases' and 'GenboreeUser_database'. The 'GenboreeUser_database' is further expanded, showing 'All Annotations in Database', 'Tracks', 'Lists & Selections', 'SampleSets', 'Samples', 'Files', 'Queries', 'Projects', 'GMT_Tutorial', and 'JonathanMill_Lab'. The 'Transcriptome' dropdown menu is open, showing options: 'Map Reads and Splice Junctions by TopHat', 'Assemble and Measure Transcripts by Cufflinks', and 'Detect Transcription Changes by Cuffdiff'. The 'Assemble and Measure Transcripts by Cufflinks' option is selected, and a tooltip is displayed. The tooltip text reads: 'Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.' Below the tooltip, the 'Input Data' section shows 'TopHat-BT474_accepted_hits.bam' and the 'Output Targets' section shows 'GenboreeUser_database'.

Cufflinks: Trapnell et al, “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation” *Nature Biotechnology*, 28 (5), May, 2010.

Tool Settings

Assemble and Measure Transcripts by Cufflinks

+ Tool Overview

Input Files:

Database: RNA-Seq Tool Demo
Group: Epigenome ToolSet Demo Input Data
File: TopHat-BT474_accepted_hits.bam

Output Database:

Database: GenboreeUser_database Group: GenboreeUser_group

Settings

Analysis Name Cufflinks-2013-3-2-14:16:17

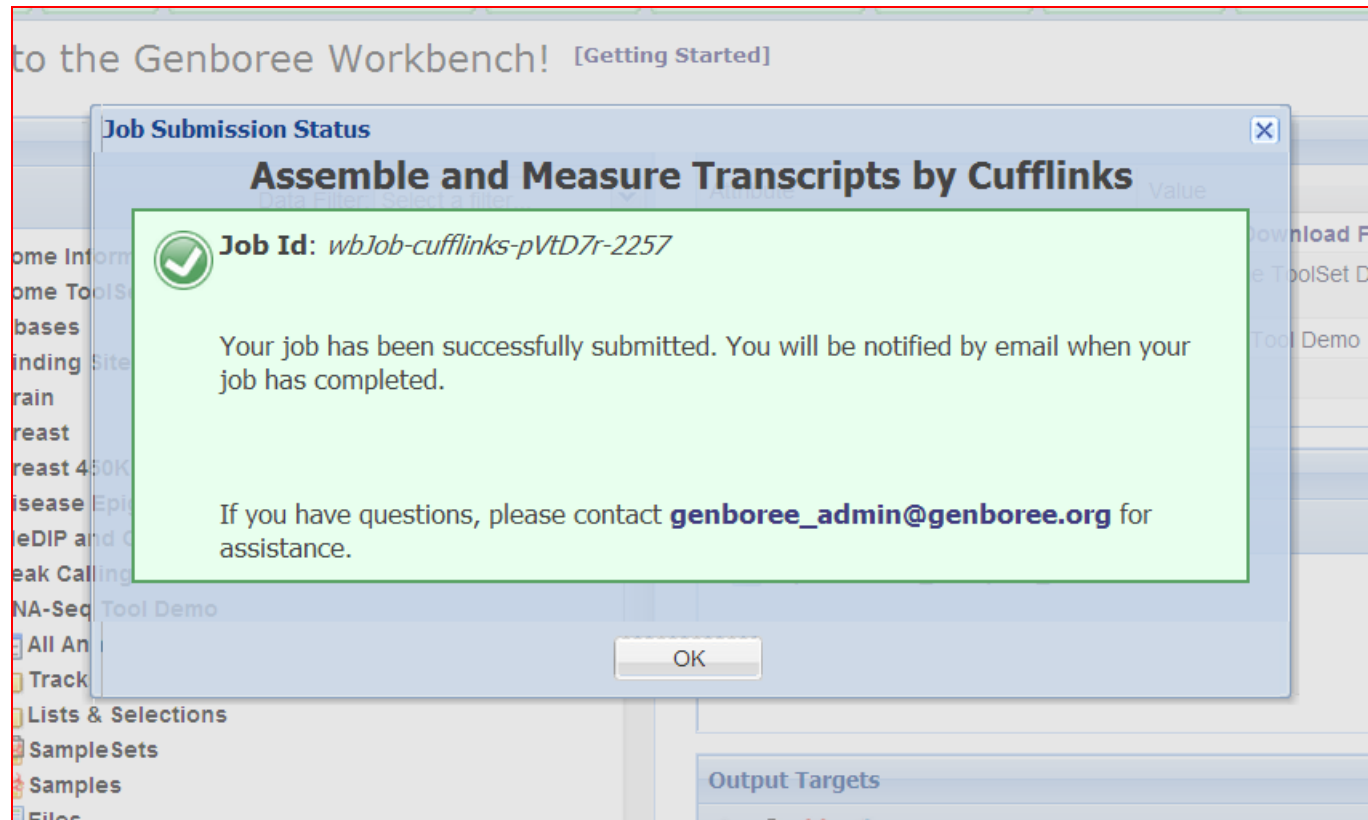
Mask File ☐

Multi Read Correct ☐

+ Advanced Abundance Estimation Options:
+ Advanced Assembly Options:
+ Advanced Reference Annotation Guided Assembly Options

Submit Cancel

-Use default settings
-Click Submit



You will receive an email with the following message when your job is finished:

Hello Genboree User,

Your Assemble and Measure Transcripts by Cufflinks job completed successfully.

Job Summary:

JobID - wbJob-cufflinks-pVtD7r-2257

Additional Info:

Database: 'GenboreeUser_database'

Group: 'GenboreeUser_group'

You can download result files from the 'Cufflinks-2013-3-2-14:16:17' folder under the 'Cufflinks' directory.

- The Genboree Team

Cufflinks files are deposited in the Data Selector in your user group.

- Expand "Databases"
- Expand "Files"
- Expand "Cufflinks"
- Click on the file of interest to highlight it in Details
- Download onto desktop (example output on next slide)

The screenshot displays the Genboree Data Selector interface. On the left, the 'Data Selector' pane shows a hierarchical tree of data categories. The 'Files' category is expanded, and the 'Cufflinks' sub-category is further expanded, revealing a folder named 'Cufflinks-2013-3-2-14:16:17'. Inside this folder, the file 'genes.fpkms_tracking.withGeneName.xls' is selected and highlighted. A red arrow points from this file to the 'Click to Download File' link in the 'Details' pane on the right. The 'Details' pane shows a table with attributes and values for the selected file. Below the table, there are sections for 'Input Data' and 'Output Targets', each with icons for upload, download, delete, and refresh.

System/Network

Welcome to the

Visualiz

Data Selector

Refresh Data Filter: Select a filter...

- All Annotations in Database
 - Tracks
 - Lists & Selections
 - SampleSets
 - Samples
 - Files
 - EpigenomeSlice
 - EpigenomicExpHeatmap
 - MACS
 - Raw Data Files
 - Spark - Results
 - Cufflinks
 - Cufflinks-2013-3-2-14:16:17
 - genes.fpkms_tracking.withGeneName.xls
 - isoforms.fpkms_tracking.withGeneName.xls
 - jobFile.json
 - raw
 - transcripts.withGeneName.gtf
 - EpigenomeCompLIMMA
 - Epigenome_Limma

Details

Attribute	Value
Download	Click to Download File
Group	GenboreeUser_group
Database	GenboreeUser_database
Description	
Name	Cufflinks/Cufflinks-2013-3-2-14:16:17/genes.fpkms_tracking

Input Data

Icons: Upload, Download, Delete, Refresh

Output Targets

Icons: Upload, Download, Delete, Refresh

Cufflinks output file: Genes.fpkm_tracking.withGeneName.Cufflinks

The gene_name
Attribute of the reference GTF
record for this transcript

Fragments Per
Kilobase of exon model per Million
mapped fragments

Lower and upper limits of 95%
FPKM confidence interval

	A	B	C	D	E	F	G	H
1	tracking_id	gene_Name	gene_id	locus	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
2	NR_026818	FAM138A	NR_026818	chr1:34610-36081	0	0	0	OK
3	NM_001005484	OR4F5	NM_001005484	chr1:69090-70008	0	0	0	OK
4	NR_039983	LOC729737	NR_039983	chr1:134772-140566	2.53257	2.11926	2.94588	OK
5	NR_046018	DDX11L1	NR_046018	chr1:11873-14408	0	0	0	OK
6	NR_024540	WASH7P	NR_024540	chr1:14361-29370	8.29061	6.90868	9.67254	OK
7	NM_001005221	OR4F29	NM_001005221	chr1:367658-368597	0	0	0	OK
8	NR_028322	LOC100132287	NR_028322	chr1:323891-328581	0	0	0	OK
9	NR_028327	LOC100133331	NR_028327	chr1:323891-328581	1.37383	1.02514	1.72252	OK
10	NM_001005221	OR4F29	NM_001005221	chr1:621095-622034	0	0	0	OK
11	NR_028327	LOC100133331	NR_028327	chr1:661138-665731	1.75903	1.36669	2.15137	OK
12	NR_033908	LOC100288069	NR_033908	chr1:700244-714068	5.58483	4.26695	6.90271	OK
13	NR_024321	LINC00115	NR_024321	chr1:761585-762902	0.408129	0.0430874	0.773171	OK
14	NR_015368	LOC643837	NR_015368	chr1:763063-789740	4.20914	3.14288	5.27541	OK
15	NR_027055	FAM41C	NR_027055	chr1:803450-812182	0.205156	0	0.427136	OK
16	NR_026874	LOC100130417	NR_026874	chr1:852952-854817	0	0	0	OK
17	NM_198317	KLHL17	NM_198317	chr1:895966-901099	3.14807	2.45698	3.83916	OK

Note: 6 columns have been removed from the spreadsheet since they are empty.

Those columns are: class_code, nearest_ref_id, gene_short_name, tss_id, length, and coverage.

Mapping with TopHat

- Process data subset of one of the cell lines
 - ***BT474***
 - Visualization in Genboree Browser and UCSC Browser

Run TopHat in The Genboree Workbench

The screenshot displays the Genboree Workbench interface. At the top, a navigation bar includes tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome, and Metagenome. The 'Transcriptome' tab is active, showing a workflow diagram with steps: 'Map Reads and Splice Junctions by TopHat', 'Assemble and Measure Transcripts by Cufflinks', and 'Detect Transcription Changes by Cuffdiff'. A tooltip for the first step explains that TopHat is a fast splice junction mapper for RNA-Seq reads, using Bowtie for alignment and identifying splice junctions between exons.

On the left, the 'Data Selector' panel shows a tree view of 'Epigenome ToolSet Demo Input Data'. Under the 'Files' folder, two fastq files are selected: 'BT474.subset.1.fastq.gz' and 'BT474.subset.2.fastq.gz'. A red dashed arrow points from these files to the 'Input Data' field in the workflow diagram.

On the right, the 'Output Targets' panel shows a list of targets, with 'GenboreeUser_database' selected.

Red text boxes provide instructions:

- Populate Input Data with two fastq files
- Populate Output Targets with your destination database
- Select Transcriptome → Analyze RNA-Seq Data → Map Reads and Splice Junctions by TopHat

TopHat will map your reads by running Bowtie, and deposit results in a TopHat Files Folder in your destination database

Tool Settings

Map Reads and Splice Junctions by Tophat ?

Tool Overview

Input Files:

Database: RNA-Seq Tool Demo
 Group: Epigenome ToolSet Demo Input Data
 File: mRNA.subset/BT474.subset.1.fastq.gz
 Database: RNA-Seq Tool Demo
 Group: Epigenome ToolSet Demo Input Data
 File: mRNA.subset/BT474.subset.2.fastq.gz

Output Location:

Database: GenboreeUser_database Group: GenboreeUser_group

Tophat Settings

Analysis Name: BT474-subset_TopHat-2013-3

Upload Results ? ☒

Track Name: BT474 : mRNA

Auto-Determine Mate Inner Dist And Stdev ? ☒

Min Anchor Length: 8

Splice Mismatches: 0

Min Intron Length: 70

Max Intron Length: 70

Max Insertion Length: 3

Max Deletion Length: 3

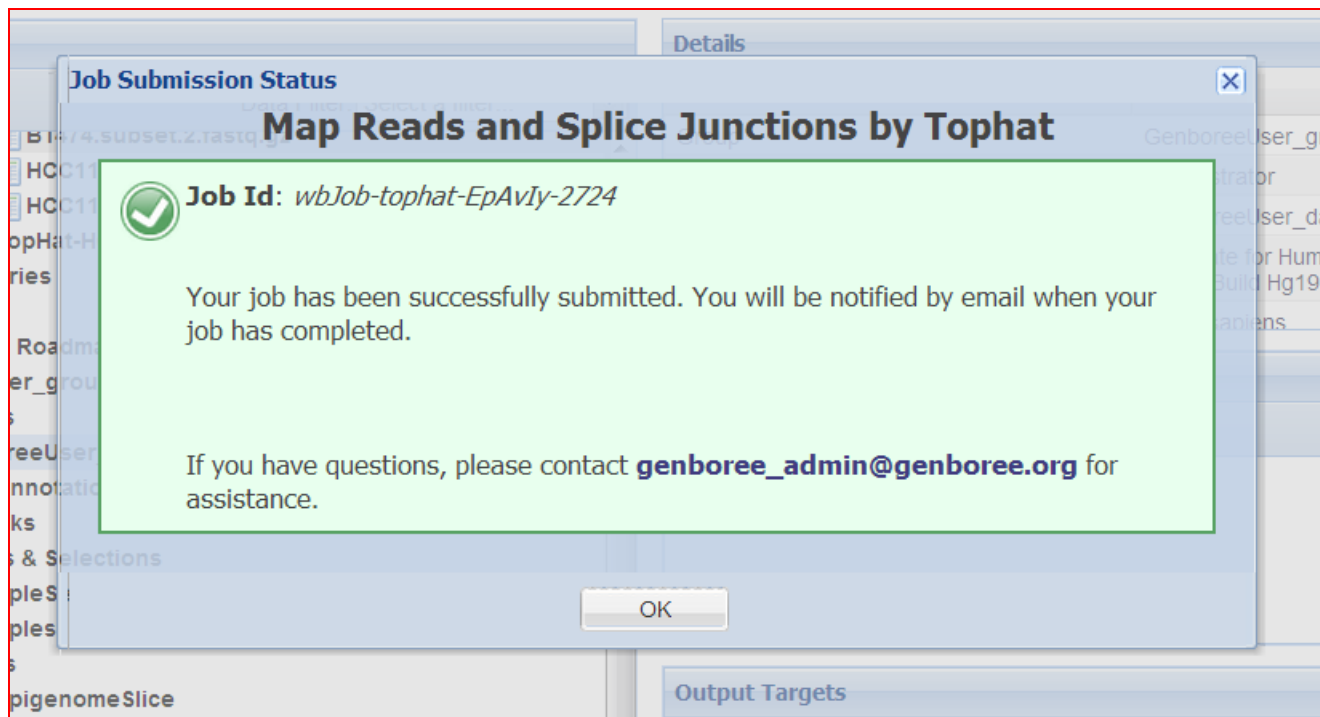
Coverage Search ? ☐

Advanced Settings:

/workbench.jsp#

is built & maintained by the Bioinformatics Research Laboratory

Track Name will appear after checking "Upload Results". The output tracks will bear this name when they appear in your destination database and when visualized in the browser



You will receive an email with the following message when your job is finished:

Hello Genboree User,

Your Map Reads and Splice Junctions by Tophat job completed successfully.

Job Summary:

JobID - wbJob-tophat-EpAvly-2724

Additional Info:

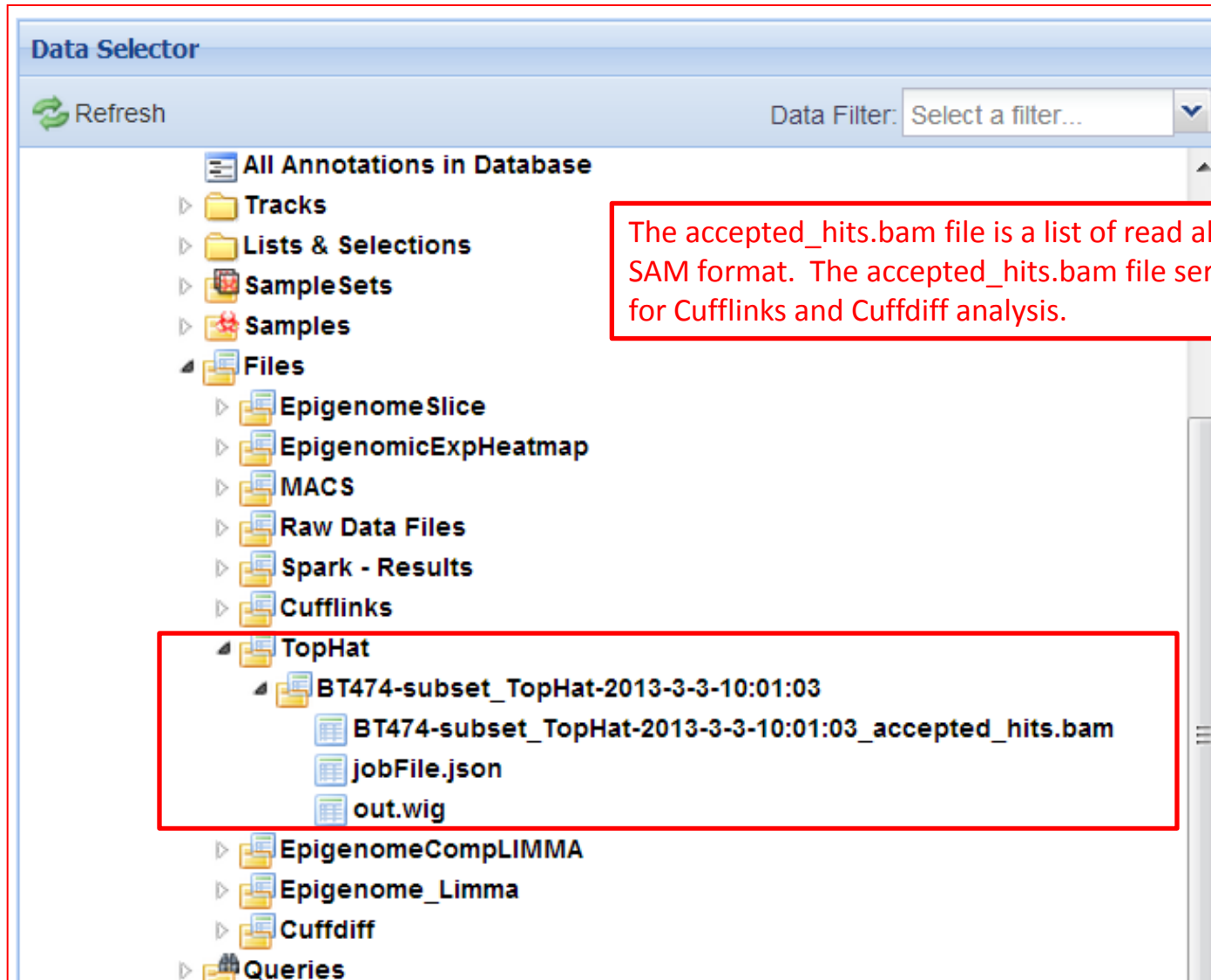
Database: 'GenboreeUser_database'

Group: 'GenboreeUser_group'

You can download result files from the 'BT474-subset_TopHat-2013-3-3-10:01:03' folder under the 'TopHat' directory.

- The Genboree Team

TopHat Output Files in the Data Selector



Home Workbench **Browser** Profile ▾ Groups ▾ Projects ▾ Databases ▾ Tools ▾ Log Out Help

GENBOREE

System/Network ▾ Data ▾ CC and Pre-processing ▾ Genome ▾ Transcriptome ▾ Cistrome ▾ Epigenome ▾

Welcome to the Genboree Workbench! [Getting Started]

Data Selector
Refresh Data Filter: Select a filter... ▾

- Epigenomics Roadmap Repository
 - GenboreeUser_group
 - Databases
 - GenboreeUser_database
 - All Annotations in Database
 - Tracks
 - Lists & Selections
 - SampleSets
 - Samples
 - Files
 - EpigenomeSlice
 - EpigenomicExpHeatmap
 - MACS
 - Raw Data Files
 - Spark - Results
 - Cumlinks
 - TopHat
 - BT474-subset_TopHat-2013-3-3-10:01:03
 - EpigenomeCompLIMMA
 - Epigenome_Limma
 - Cuffdiff

Details

Attribute	Value
-----------	-------

Output Targets

We now wish to visualize the TopHat results in the context of genomic and/or epigenomic data via the Genboree Browser and the UCSC Browser

Tell the Genboree Browser What You Wish to View

Select your Group and Database. Click "View"
("chr1" is default, and can be changed (next slides))

The screenshot displays the Genboree Browser interface. At the top, there are navigation tabs: Home, Workbench, Browser (selected), and another partially visible tab. Below these are links: GetDNA, Full URL, Download, Style Setup, Link Manager, Track Manager, and Class Manager.

The main configuration area includes:

- Group:** A dropdown menu set to "GenboreeUser_group" and a **Role:** field set to "ADMINISTRATOR". An **Email Group** button is to the right.
- Database:** A dropdown menu set to "GenboreeUser_database".
- Assembly:** A dropdown menu set to "hg19".
- Entry Point:** A dropdown menu set to "chr1".
- From:** A text field containing "117,997,344".
- To:** A text field containing "131,253,276".
- A **View** button.

Below the configuration fields are navigation controls:

- Extend:** A range selector showing "2,000" with left and right arrows.
- Navigation buttons: Home, Previous, First, Previous, Next, Last, and Search.
- Zoom In:** Buttons for 1.5x, 2x, 3x, 5x, 10x, and Base.
- Zoom Out:** Buttons for 1.5x, 2x, 3x, 5x, 10x, and Full.
- A horizontal scale bar with a red box highlighting a specific position.

The main display area shows a genomic track for "chr1" with coordinates from 120,000,000 to 130,000,000. A green bar at the top of the track is labeled "DRAG HERE TO SELECT". Below this, several tracks are visible:

- BMC:H3K4me3:** A track showing signal intensity with vertical bars.
- BT474-subset:mRNA:** A track showing a single mRNA annotation with a score of 4004.0.
- ESCs:Rad21_Nanog:** A track showing a single annotation with a score of 1.0000.
- Read:Density_BodyS:** A track showing a single annotation.
- Read:Density_Indiv:** A track showing a single annotation.

On the right side of the track display, there are labels for the tracks: "test", "Track 'BT474-subset:mRNA' (8,132,748 bp with scores)", "Track 'ESCs:Rad21_Nanog' (218 annotations)", "Track 'Read:Density_BodySite' (1,552 annotations)", and "Track 'Read:Density_Individual' (248 annotations)".

View TopHat Results in the Genboree Browser

Change Entry Point to "chr17" and the coordinates as shown. Click View.

The screenshot displays the Genboree Browser interface. At the top, the Genboree logo and Baylor College of Medicine (BCM) logo are visible. Below the navigation bar, the user is logged in as 'GenboreeUser_group' with the role of 'ADMINISTRATOR'. The main search area shows the following parameters: Database: GenboreeUser_database, Assembly: hg19, Entry Point: chr17, From: 37,849,089, To: 37,892,079. A 'View' button is present. Below the search area, there are controls for 'Extend' (set to 2,000) and 'Zoom In' (1.5x, 2x, 3x, 5x, 10x, Base) and 'Zoom Out' (1.5x, 2x, 3x, 5x, 10x, Full). A 'Search' button is also present. The main display area shows a genomic track for chr17. The track is labeled 'chr17' and 'BMC:H3K4me3'. Below the track, there are several tracks: 'BT474-subset:mRNA' (8,132,748 bp with scores), 'Cyto:Band' (862 annotations), 'Gene:RefSeq' (345,623 annotations), 'CCDS:Genes' (244,393 annotations), and 'Known:Gene' (733,510 annotations). The 'BT474-subset:mRNA' track shows a peak at 125.00. The 'Cyto:Band' track shows a band at 17q12. The 'Gene:RefSeq' track shows a gene structure with exons and introns. The 'CCDS:Genes' track shows a gene structure with exons and introns. The 'Known:Gene' track shows a gene structure with exons and introns.

Visualize TopHat output in the UCSC Genome Browser

System/Network Data QC and Pre-processing Genome Transcriptome Cistrome Epigenome

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter

- www.genboree.org
 - Atlas Tools Access
 - EDACC
 - Epigenome Informatics Workshop (May 2012)
 - Epigenome ToolSet Demo Input Data
 - Epigenomics Roadmap Repository
 - GenboreeUser_group
 - Databases
 - GenboreeUser_database**

Input Data

GenboreeUser_database

Output Targets

You will need to generate BigWig files to view the tracks in the UCSC Browser.

-Drag the database containing the tracks of interest into Input Data

Select Data → Tracks → Utilities → Generate BigWig Files

The screenshot shows the Genomics Workbench interface. The top navigation bar includes tabs for System/Network, Data, QC and Pre-processing, Genome, Transcriptome, Cistrome, Epigenome, and Metabolome. The 'Data' tab is active, displaying a 'Welcome to Genomics Workbench! [Getting Started]' message. On the left, the 'Data Selector' panel shows a tree view of data sources, including 'www.genboree.org' and 'GenboreeUser_group'. The 'Databases' menu is open, showing options like 'Entity Lists', 'Entrypoints', 'Files', 'Projects', 'Samples & Sample Sets', and 'Tracks'. The 'Tracks' option is highlighted, and a sub-menu is open showing 'Import', 'Utilities', 'Download Track', 'Copy/Move Tracks', 'Rename Tracks', 'Generate BigWig Files', 'Generate BigBed Files', and 'Coverage Computation'. The 'Generate BigWig Files' option is highlighted, and a tooltip is displayed: 'Generate BigWig Files. Create Bigwig files for existing tracks. Creating these is required for viewing tracks in the UCSC browser.' The 'Details' panel on the right shows user information for 'GenboreeUser_group' and 'GenboreeUser_database'. The 'Output Targets' panel at the bottom shows icons for adding, removing, and saving targets.

-Select the tracks for which you wish to generate igWig Files

-Click Submit

Tool Settings

Generate BigWig files

Tool Overview

All tracks belong to:

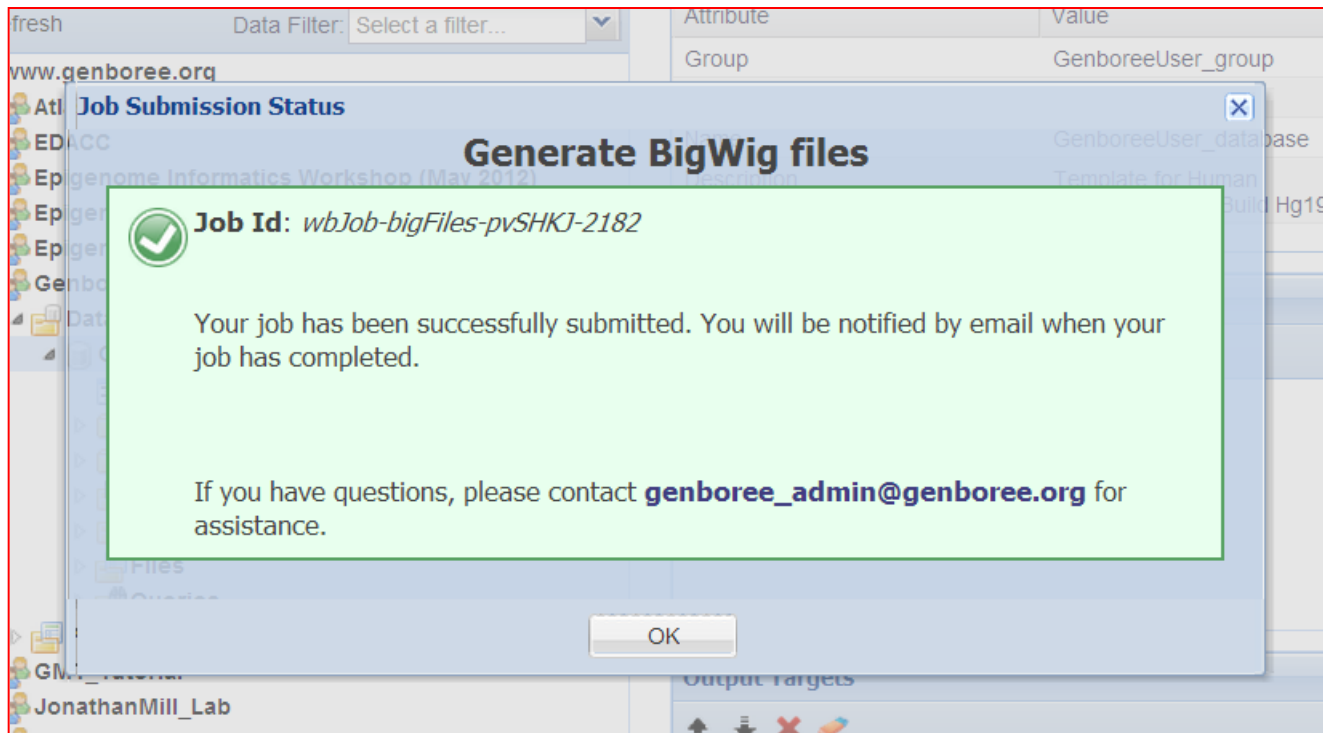
Database: *GenboreeUser_database* **Group:** *GenboreeUser_group*

Generate BigWig Files

Track	BigWig File
BMC:H3K4me3	<input type="checkbox"/> no file
BT474-subset:mRNA	<input checked="" type="checkbox"/> no file
EpigenomicLimmaComp:Analysis	<input type="checkbox"/> no file
ESCs:Rad21_Nanog	<input type="checkbox"/> no file
Read:Density_BodySite	<input type="checkbox"/> no file
Read:Density_Individual	<input type="checkbox"/> no file

Clear All

Submit Cancel



You will receive an email with the following message when your job is finished:

Hello Genboree User,

Your Generate BigWig Files job completed successfully.

Job Summary:

JobID - wbJob-bigFiles-pvSHKJ-2182

Additional Info:

You can use the following links to either download the big* files or visualize the data in the UCSC genome browser (if the database has been unlocked)

BT474-subset:mRNA:

Download bigWig file:

http://www.genboree.org/REST/v1/grp/GenboreeUser_group/db/GenboreeUser_database/trk/BT474-subset%3AmRNA/bigWig?gbKey=eb8oft3v

Use this link to view the track in the UCSC browser.

http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hgt.customText=http%3A%2F%2Fwww.genboree.org%2FREST%2Fv1%2Fgrp%2FGenboreeUser_group%2Fdb%2FGenboreeUser_database%2Ftrk%2FBT474-subset%253AmRNA%3FgbKey%3Deb8oft3v%26format%3Ducsc_browser%26ucscType%3DbigWig

-Your Database is secure, and will need to be unlocked before the viewing the BigWig files in the UCSC Browser

-Drag your database with the tracks of interest into Output Targets

The screenshot displays the UCSC Genome Browser interface. At the top, there are tabs for 'System/Network', 'Data', 'QC and Pre-processing', 'Genome', 'Transcriptome', 'Cistrome', and 'Epigenome'. The 'Data' tab is active, showing a 'Data Selector' on the left with a 'Refresh' button and a tree view of various data sources. The 'Databases' section is expanded, showing a list of databases including 'GenboreeUser_group' and 'GenboreeUser_database'. A context menu is open over the 'Databases' section, with the 'Unlock/Lock Database' option highlighted. A red arrow points from this menu option to the 'Output Targets' section at the bottom right, where the 'GenboreeUser_database' is being dragged. A red dashed arrow also points from the 'GenboreeUser_database' in the 'Data Selector' to the 'Output Targets' section. A tooltip for 'Unlock/Lock Database' is visible, stating: 'Unlocking the database allows you to expose database resources to the public. Once unlocked, you can use the key to view data without requiring authentication.'

System/Network Data QC and Pre-processing Genome Transcriptome Cistrome Epigenome

Welcome to

Data Selector

Refresh

www.genboree.org

Atlas Tool

EDACC

Epigenome

Epigenome ToolSet Demo Input Data

Epigenomics Roadmap Repository

GenboreeUser_group

Databases

GenboreeUser_database

All Annotations in Database

Tracks

Lists & Selections

SampleSets

Samples

Files

Queries

Projects

GMT_Tutorial

JonathanMill_Lab

paithank_group

Public

Create Database

Rename Database

Delete Database

Edit Database Info

Clone/Copy Database

Unlock/Lock Database

Publish/Retract Database

Unlock/Lock Database

Unlocking the database allows you to expose database resources to the public. Once unlocked, you can use the key to view data without requiring authentication.

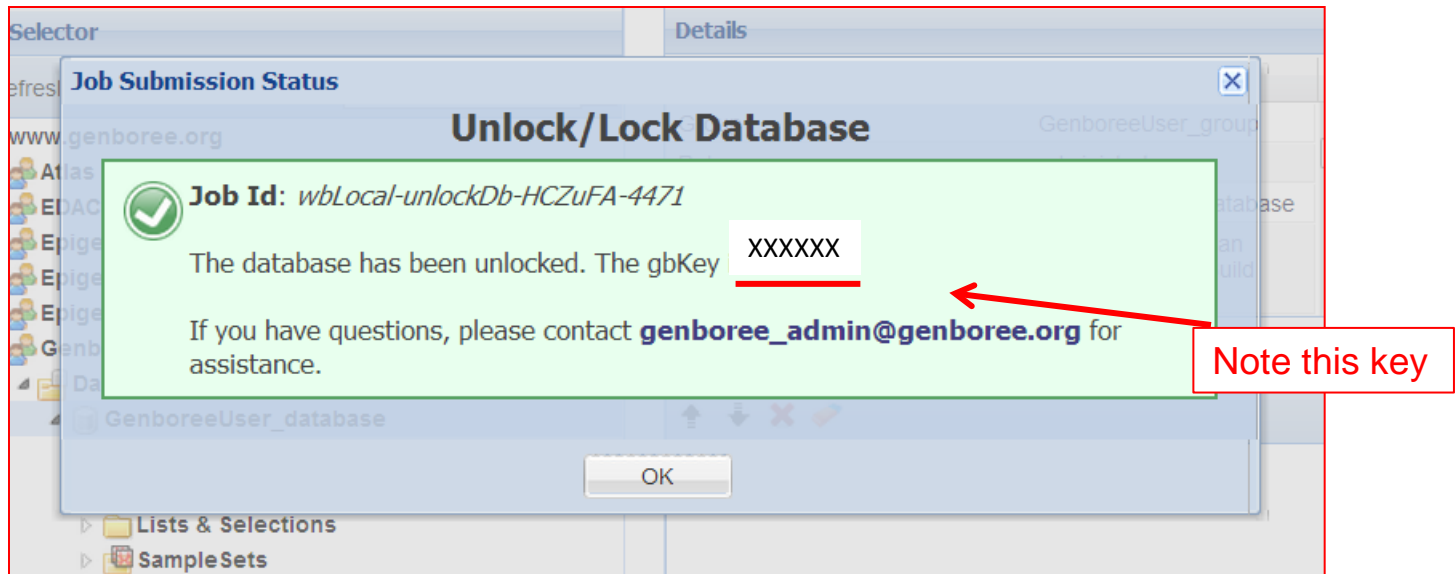
Input Data

Output Targets

GenboreeUser_database

-Select Data → Databases → Unlock/Lock Database





System/Network Data QC and Pre-processing Genome Transcriptome Cistrome Epigenome Metagenome Visualization

Welcome to the Genboree Workbench! [Getting Started]

Data Selector

Refresh Data Filter: Select a filter...

- www.genboree.org
 - Atlas Tools Access
 - EDACC
 - Epigenome Informatics Workshop (May 2012)
 - Epigenome ToolSet Demo Input Data
 - Epigenomics Roadmap Repository
 - GenboreeUser_group
 - Databases
 - GenboreeUser_database**

Details

Attribute	Value
Group	GenboreeUser_group
Role	administrator
Name	GenboreeUser_database
Description	Template for Human Genome, UCSC Build Hg19
Species	Homo sapiens

Input Data

GenboreeUser_database

Output Targets

View Track Grid
View Sample Grid
Tabular Annotation Viewer
Launch UCSC Genome Browser

- Populate Input Data with the database containing the tracks of interest
- Select Visualization → Launch UCSC Genome Browser

Tool Settings

Launch UCSC Genome Browser

+ Tool Overview

All tracks belong to:

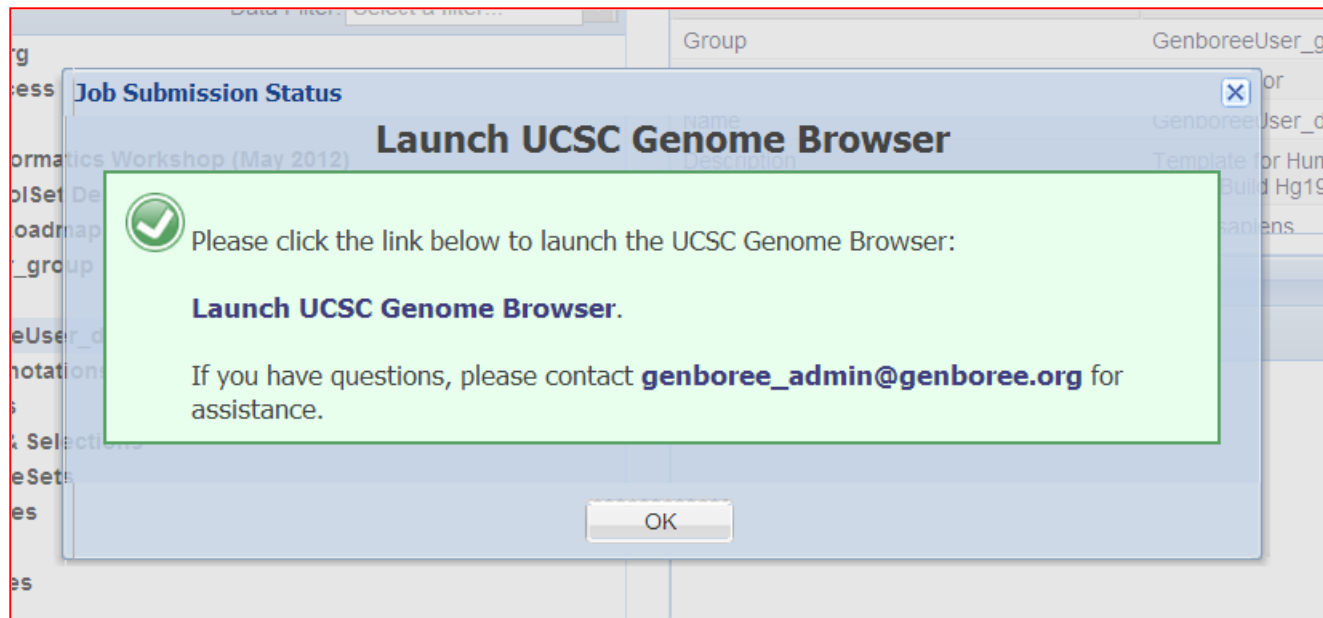
Database: *GenboreeUser_database* **Group:** *GenboreeUser_group*

Launch UCSC Browser

Track	BigWig	BigBed
BMC:H3K4me3	no file	no file
BT474-subset:mRNA	<input checked="" type="checkbox"/>	no file
CCDS:Genes	no file	no file
Cyto:Band	no file	no file
EpigenomicLimmaComp:Analysis	no file	no file
ESCs:Rad21_Nanog	no file	no file
Gene:RefSeq	no file	no file
Known:Gene	no file	no file
Read:Density_BodySite	no file	no file
Read:Density_Individual	no file	no file

-Indicate which tracks you wish to visualize

-Click Submit



Tophat output in the Context of the UCSC Genome Browser

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

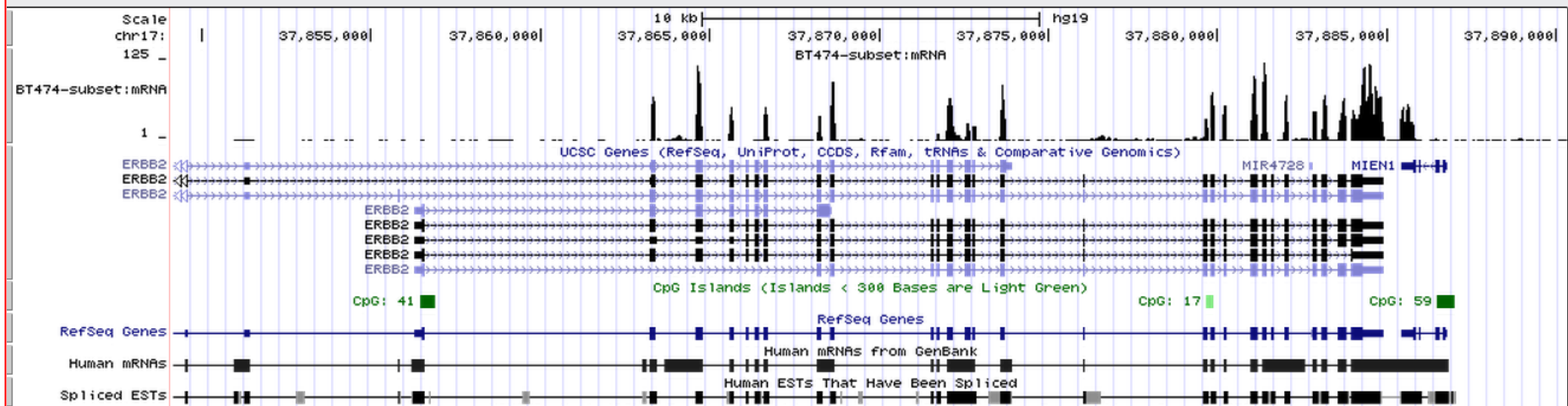
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

chr17:37,849,151-37,890,319 41,169 bp.

enter position, gene symbol or search terms

go

chr17 (q12) p13.3 p13.2 p13.1 17p12 17p11.2 17q11.2 17q12 q21.31 17q22 23.2 24.2 q24.3 q25.1 17q25.3



Exercise plan

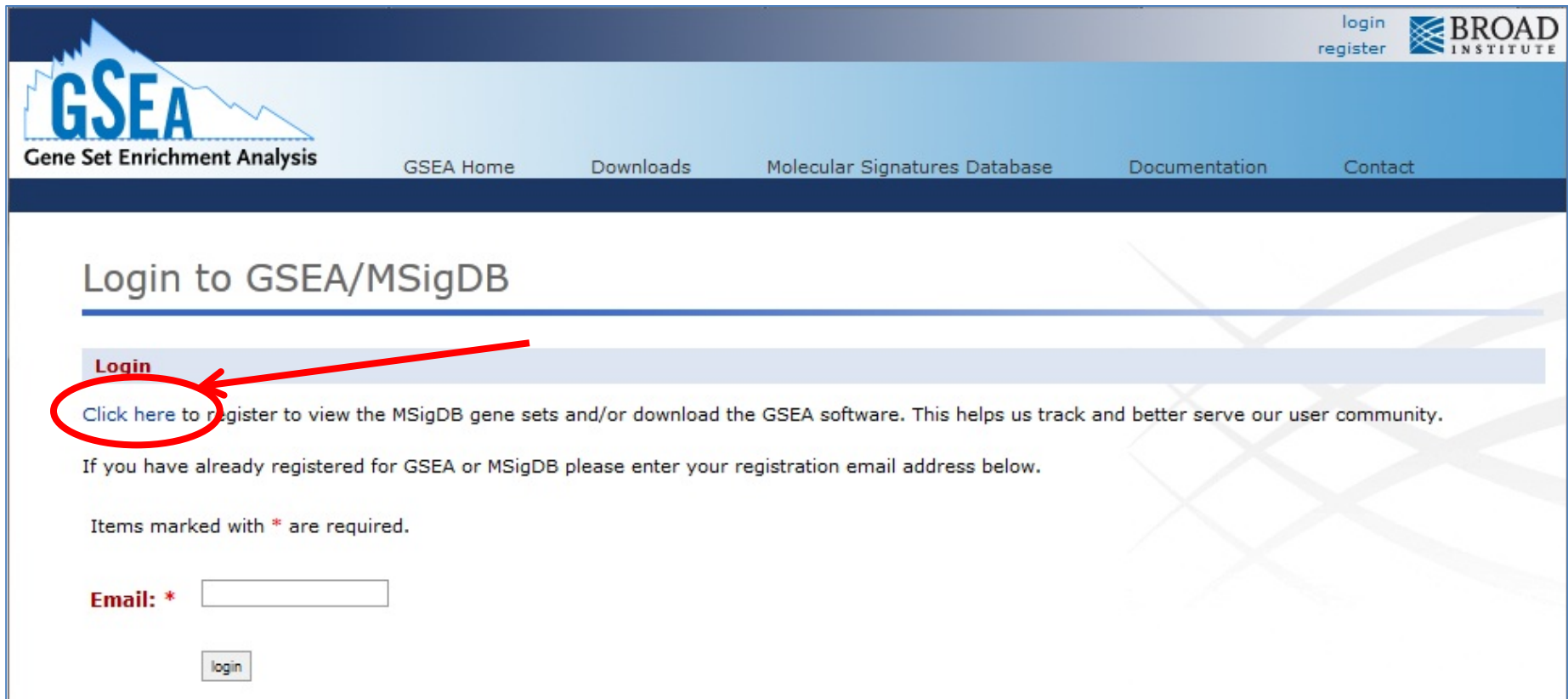
- Process subset of one of the cell lines
 - BT474
 - Visualization in Genboree Browser and UCSC Browser
 - *Gene enrichment via GSEA/MSigDB*

GSEA/MSigDB

- **Gene Set Enrichment Analysis**
 - Subramanian, Tamayo, et al. 2005, PNAS 102, 15545-15550
 - Mootha, Lindgren, et al. 2003, Nat Genet 34, 267-273
- **Molecular Signatures Database**
 - Subramanian, Tamayo, et al. 2005, PNAS 102, 15545-15550
- Exposed as a web service
- Future plan
 - integrate into the Epigenome Toolset

Register to MSigDB

<http://www.broadinstitute.org/gsea/login.jsp>



GSEA
Gene Set Enrichment Analysis

login
register

BROAD
INSTITUTE

GSEA Home Downloads Molecular Signatures Database Documentation Contact

Login to GSEA/MSigDB

Login

[Click here](#) to register to view the MSigDB gene sets and/or download the GSEA software. This helps us track and better serve our user community.

If you have already registered for GSEA or MSigDB please enter your registration email address below.

Items marked with * are required.

Email: *

Register to MSigDB

GSEA/MSigDB Registration and License Agreement

Instructions to obtain GSEA software and/or MSigDB gene sets. Please Read carefully.

1. Fill in the form below.
2. The software and gene sets are freely available to individuals in academic and private institutions. There are no licensing fees.
3. Source code is freely available.
4. Read the license agreement and make sure you agree with the terms of the agreement.
If so, click the 'I Agree button' at the end of the form and you will be transferred to the GSEA download page.

Items marked with * are required.

Name: *

Email: *

(You will receive a registration notification email.)

Organization: *

Country: *

Join mailing list:

☒ notify me of GSEA updates

(You will receive a confirmation email. Reply to join the list.)

Comments:

**GSEA and MSigDB
license agreements:**

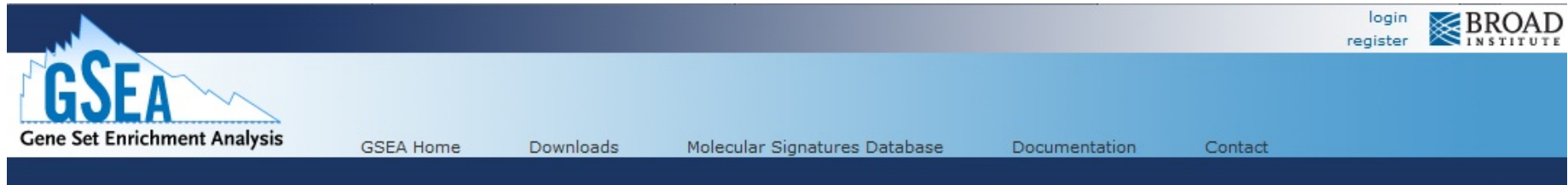
***** GSEA/MSigDB LICENSE AGREEMENT *****

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SINGLE USER LICENSE AGREEMENT FOR INTERNAL RESEARCH PURPOSES
ONLY

This Agreement is made between Massachusetts Institute of
Technology with a principal address at 77
Massachusetts Avenue, Cambridge, MA 02139 ("MIT") and the
subscriber above ("LICENSEE"), and is effective at the date the
downloading is completed and proper registration/licensing

Login to MSigDB

<http://www.broadinstitute.org/gsea/login.jsp>



Login to GSEA/MSigDB

Login

Click [here](#) to register to view the MSigDB gene sets and/or download the GSEA software. This helps us track and better serve our user community.

If you have already registered for GSEA or MSigDB please enter your registration email address below.

Items marked with * are required.

Email: *

Use MSigDB

Investigate Gene Sets

Gain further insight into the biology behind a gene set by using the following tools:

- ▶ **compute overlaps** with other gene sets in MSigDB ([more...](#))
- ▶ **display the gene set expression profile** based on a selected compendium of expression data ([more...](#))
- ▶ **categorize** members of the gene set by gene families ([more...](#))

Gene Identifiers

Compute Overlaps

- ☐ C1: positional gene sets [?](#)
- ☐ C2: curated gene sets [?](#)
- ☐ CGP: chemical and genetic perturbations [?](#)
- ☐ CP: canonical pathways [?](#)
 - ☐ CP:BIOCARTA: BioCarta gene sets [?](#)
 - ☐ CP:KEGG: KEGG gene sets [?](#)
 - ☐ CP:REACTOME: Reactome gene sets [?](#)
- ☐ C3: motif gene sets [?](#)
- ☐ MIR: microRNA targets [?](#)
- ☐ TFT: transcription factor targets [?](#)
- ☐ C4: computational gene sets [?](#)
 - ☐ CGN: cancer gene neighborhoods [?](#)
 - ☐ CM: cancer modules [?](#)
- ☐ C5: GO gene sets [?](#)
 - ☐ BP: GO biological process [?](#)
 - ☐ CC: GO cellular component [?](#)
 - ☐ MF: GO molecular function [?](#)

show top 10 genesets

compute overlaps

Compendia expression profiles

- ☒ Human tissue compendium (Novartis)
- ☐ Global Cancer Map (Broad Institute)
- ☐ NCI-60 cell lines (National Cancer Institute)

display expression profile

Gene families

show gene families

Gene Identifier Platform

GENE SYMBOL

Gene expression differences

Filter by “significant” column

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	test_id	gene_Nar	gene_id	gene	locus	sample	sample	status	value_1	value_2	log2(fol	test_sta	p_value	q_value	significant
5	NM_000001	ACADS	NM_000001	-	chr12:121	Luminal	BasalA	OK	0.181132	9.56262	5.72229	-5.28258	1.27E-07	5.70E-06	yes
10	NM_000002	ADA	NM_000002	-	chr20:432	Luminal	BasalA	OK	0.046568	12.6184	8.08196	-5.02623	5.00E-07	1.95E-05	yes
32	NM_000004	AR	NM_000004	-	chrX:6676	Luminal	BasalA	OK	10.932	0.008774	-10.2831	6.59281	4.32E-11	4.09E-09	yes
41	NM_000005	ATP7B	NM_000005	-	chr13:525	Luminal	BasalA	OK	7.02049	0.417952	-4.07016	4.16209	3.15E-05	0.000777	yes
51	NM_000006	C3	NM_000006	-	chr19:667	Luminal	BasalA	OK	0.038313	48.1341	10.295	-10.0105	0	0	yes
88	NM_000100	CYBA	NM_000100	-	chr16:887	Luminal	BasalA	OK	32.1228	0.168313	-7.57631	4.75509	1.98E-06	6.67E-05	yes
91	NM_000100	CYP1B1	NM_000100	-	chr2:3829	Luminal	BasalA	OK	5.49719	33.7556	2.61836	-3.01935	0.002533	0.030879	yes
195	NM_000201	ITGA6	NM_000201	-	chr2:1732	Luminal	BasalA	OK	3.16834	29.6013	3.22386	-3.59329	0.000327	0.006	yes
199	NM_000201	JAG1	NM_000201	-	chr20:106	Luminal	BasalA	OK	0.530239	54.1496	6.67416	-6.70849	1.97E-11	2.02E-09	yes
254	NM_000201	NPC1	NM_000201	-	chr18:210	Luminal	BasalA	OK	5.22127	31.5541	2.59535	-2.96999	0.002978	0.035004	yes
290	NM_000300	CTSA	NM_000300	-	chr20:445	Luminal	BasalA	OK	59.9431	3.58534	-4.06341	4.16583	3.10E-05	0.000766	yes
327	NM_000300	SOX9	NM_000300	-	chr17:701	Luminal	BasalA	OK	4.77218	61.8855	3.69688	-4.09221	4.27E-05	0.001003	yes
392	NM_000401	HSD17B1	NM_000401	-	chr17:407	Luminal	BasalA	OK	27.5426	3.35674	-3.03653	3.10584	0.001897	0.024754	yes
401	NM_000401	KRT17	NM_000401	-	chr17:397	Luminal	BasalA	OK	0.180647	587.103	11.6662	-10.4385	0	0	yes
403	NM_000401	KRT5	NM_000401	-	chr12:529	Luminal	BasalA	OK	0.27407	625.903	11.1572	-10.2628	0	0	yes
414	NM_000401	NOTCH3	NM_000401	-	chr19:152	Luminal	BasalA	OK	11.6804	176.97	3.92134	-4.11637	3.85E-05	0.000926	yes

Copy “official”
gene symbol

Use MSigDB

Gene Identifiers

CA2
HMBS
ITGA6
JAG1
LAMA3
NF1
KRT17
NOTCH3
APRT
F12
TGFB1
ADORA2A
ALDH1A3
TSPO
CRAT
GSTT1
HTR1D
OXTR
CD44
CD44
SMAD5
MTUS1
ATP5C1
ZNF787
PTRH1
ARHGEF35
MAGED1
RHBDF2
TCF4

Gene Identifier Platform

GENE SYMBOL

Compute Overlaps

- ☐ C1: positional gene sets ?
- ☒ C2: curated gene sets ?
 - ☒ CGP: chemical and genetic perturbations ?
 - ☒ CP: canonical pathways ?
 - ☒ CP:BIOCARTA: BioCarta gene sets ?
 - ☒ CP:KEGG: KEGG gene sets ?
 - ☒ CP:REACTOME: Reactome gene sets ?
- ☐ C3: motif gene sets ?
 - ☐ MIR: microRNA targets ?
 - ☐ TFT: transcription factor targets ?
- ☒ C4: computational gene sets ?
 - ☒ CGN: cancer gene neighborhoods ?
 - ☒ CM: cancer modules ?
- ☒ C5: GO gene sets ?
 - ☒ BP: GO biological process ?
 - ☒ CC: GO cellular component ?
 - ☒ MF: GO molecular function ?

show top 20 genesets

compute overlaps

Select
gene sets

Number of gene
sets returned

Diff
expressed
genes

Use MSigDB

Compute Overlaps for Selected Genes






Converted 701 submitted identifiers into 599 gene symbols. [click here](#) for details.

Collections	# Overlaps Shown	# Gene Sets in Collections	# Genes in Comparison (n)	# Genes in Collections (N)
C2, C4, C5	10	5607	599	22684





Click the gene set name to see the gene set page. Click the number of genes [in brackets] to download the list of genes.

Color bar shading from light green to black, where lighter colors indicate more significant p values (< 0.05) and black indicates less significant p values (≥ 0.05).

Export: [Excel](#)

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p value ?
NUYTEN_NIPP1_TARGETS_DN [777]	Genes down-regulated in PC3 cells (prostate cancer) after knockdown of NIPP1 [Gene ID=5511] by RNAi.	67		0 e ⁰
SMID_BREAST_CANCER_BASAL_DN [713]	Genes down-regulated in basal subtype of breast cancer samples.	65		0 e ⁰
SMID_BREAST_CANCER_LUMINAL_B_DN [599]	Genes down-regulated in the luminal B subtype of breast cancer.	63		0 e ⁰
CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_NCE_5 [482]	The 'group 5 set' of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 but not ERBB2 [Gene ID=2099, 2064].	54		0 e ⁰
SMID_BREAST_CANCER_BASAL_UP [676]	Genes up-regulated in basal subtype of breast cancer samples.	78		0 e ⁰
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN [456]	Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like	59		0 e ⁰

Use MSigDB

SMID_BREAST_CANCER_BASAL_DN [713]	Genes down-regulated in basal subtype of breast cancer samples.	65		0 e^0
SMID_BREAST_CANCER_LUMINAL_B_DN [599]	Genes down-regulated in the luminal B subtype of breast cancer.	63		0 e^0
SMID_BREAST_CANCER_BASAL_UP [676]	Genes up-regulated in basal subtype of breast cancer samples.	78		0 e^0
CHARAFE_BREAST_CANCER_LUMINAL_VS_BASAL_DN [456]	Genes down-regulated in luminal-like breast cancer cell lines compared to the basal-like ones.	59		0 e^0

Enrichments for gene sets differentiating luminal vs basal breast cancer cells