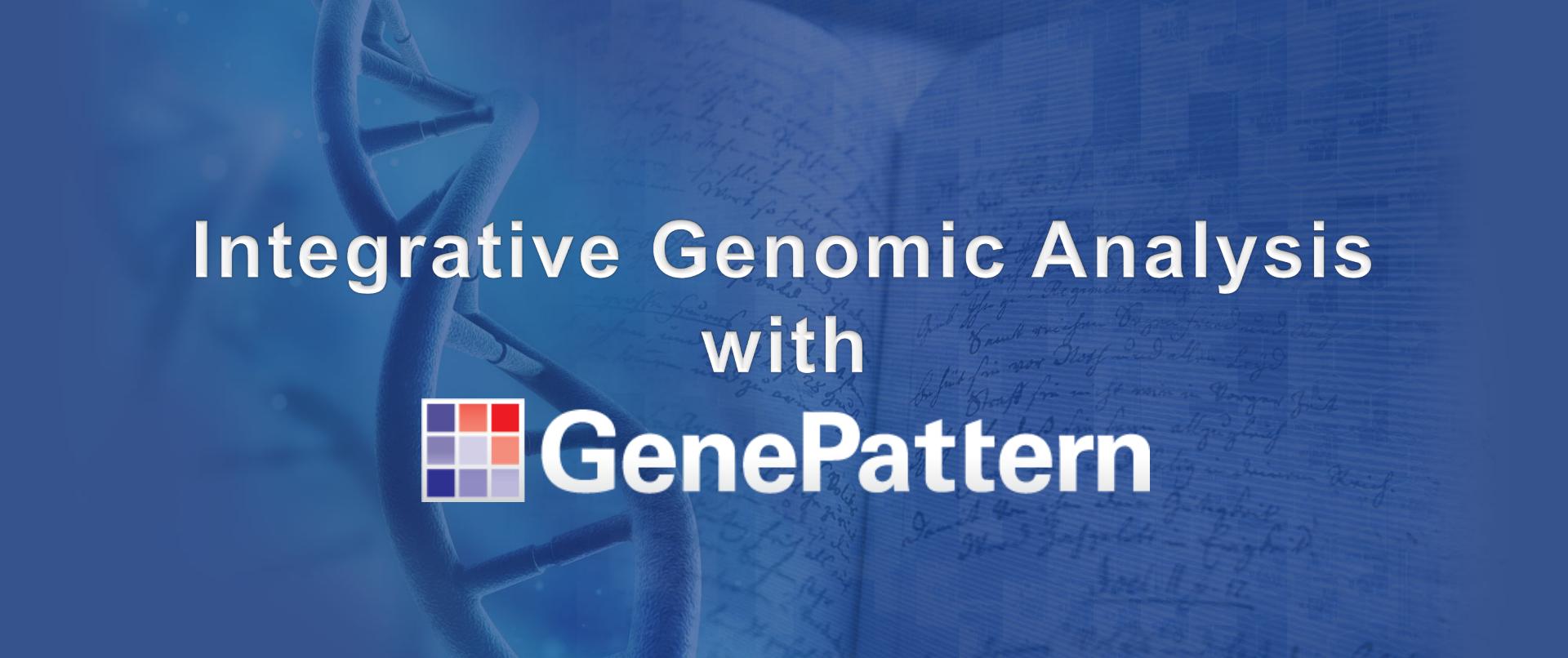


Preliminaries

- Make sure you are connected to the wireless network
- Register and log in at <https://notebook.genepattern.org>
- Make sure you can log in to: <https://cloud.genepattern.org>



Integrative Genomic Analysis with

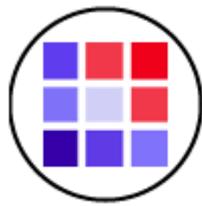


GenePattern

April 30, 2019
Barbara Hill

Agenda

- Overview
- Data Formats
- Running Analyses
- **Break**
- RNA-seq in GenePattern
- IGV
- Other GP Features
- Closing
- Open Q&A



GenePattern Overview

Tools for Bioinformatics



Best-Practices Documentation Blog Forum Events Download

Search

[Back to Tool Docs Index](#)

MuTect2

Call somatic SNPs and indels via local re-assembly of haplotypes

HISAT2

graph-based alignment of next generation sequencing reads to a population of genomes

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools: Reading/writing/indexing/viewing SAM/BAM/CRAM format
BCFTools: Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib: A C library for reading/writing high-throughput sequencing data

Samtools and BCFTools both use HTSlib internally, but these source packages contain their own copies of htseq so they can be built independently.



Bowtie 2
Fast and sensitive read alignment



Home Installation Documentation Examples

1.4. Support Vector Machines

Principal Component Analysis

Picard

build passing

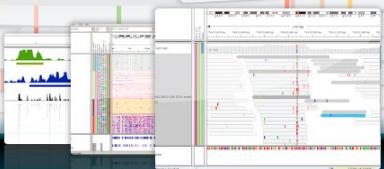
A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Hierarchical Clustering / Dendrograms

Cufflinks

Transcriptome assembly and differential expression analysis for RNA-Seq.

Integrative Genomics Viewer



Burrows-Wheeler Aligner

Introduction

BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms:

NMF: Non-negative Matrix Factorization

What is HAPSEG?

HAPSEG is a probabilistic method to interpret bi-allelic marker data in cancer samples.

MAGECK

Model-based Analysis of Genome-wide CRISPR-Cas9 Knockout

What is RNA-SeQC?

RNA-SeQC is a java program which computes a series of quality control metrics for RNA-seq data.



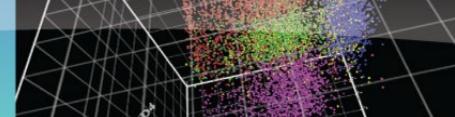
Network Data Integration, Analysis, and Visualization in a Box

Trimmomatic: A flexible read trimming tool for Illumina NGS data



MSigDB
Molecular Signatures Database

FLAME
Flow analysis with Automated Multivariate Estimation



Gene Set Enrichment Analysis
Constellation Map: Downstream visualization and interpretation of gene set enrichment results [version 1; referees: 2 approved]

Problems with bioinformatics tool use and interoperability

- Tools are built using different languages and with different architectural assumptions.
- Each tool has its own installation and operational requirements.
- Tools require (sometimes extensive) Unix knowledge.
- Tools are not designed to communicate with each other.

```
bowtie -a --best --strata -S -m 100 -X 400 --chunkmb 256 --fullref -p 4
Dmel.BDGP5-transcripts \ -1 SRR031714_1.fastq -2 SRR031714_2.fastq |
samtools view -F 0xC -bS - | \ samtools sort -n - ~/Desktop/untreated3-
transcriptome
```

Solution features: Wrapping tools

- “Wrap” tools in a web-based interface
- No installation/running requirements
- No programming required
- Fill out required parameters and provide input files

The screenshot shows a web browser window titled "GenePattern - Bowtie.aligner" at the URL <https://genepattern.broadinstitute.org/gp/pages/index.jsf?lsid=urn:lsid:broad.mit.edu:cancer.software.genepattern.module.analysis.Bowtie.aligner>. The page displays the "Bowtie.aligner" module configuration. The top navigation bar includes links for Modules & Pipelines, Suites, Job Results, Resources, Help, and GenomeSpace. The main content area is titled "Bowtie.aligner" version 4. It provides a brief description of Bowtie2 (v. 2.1.0) as an ultrafast and memory-efficient short read aligner. The configuration form includes fields for "prebuilt bowtie index" (with a dropdown menu and a "Batch" checkbox), "custom bowtie index" (with "Upload File...", "Add Path or URL...", and "Drag Files Here" fields, along with a note about 2GB file upload limit), "input format*" (dropdown menu and "Batch" checkbox), "reads pair 1*" (with "Upload File...", "Add Path or URL...", and "Drag Files Here" fields, along with a note about 2GB file upload limit), "reads pair 2" (with "Upload File...", "Add Path or URL...", and "Drag Files Here" fields, along with a note about 2GB file upload limit), "quality value scale" (dropdown menu set to "Phred" and "Batch" checkbox), "integer quality values" (dropdown menu set to "no" and "Batch" checkbox), and "max reads to align" (text input field and "Batch" checkbox). At the bottom right are "Reset" and "Run" buttons.

Solution features: Tool Repository

- Collection of hundreds of wrapped tools
- Gene expression, sequence variation, proteomics, network analysis, machine learning, flow cytometry, etc.
- Searchable by name, keyword, etc.
- Widely-used community tools, lab-developed tools, utilities
- Users can contribute their own tools

AddNoiseToFCS	Add noise to specified parameters in an FCS data file. Flow Cytometry	Preprocess & Utilities
ApplyGatingML	Apply a Gating-ML file on an FCS data file (gate and/or transform list mode data) Flow Cytometry	Pathway Analysis
AreaChange	Calculates fraction of area under the spectrum that is attributable to signal (area after noise... Proteomics, ProteomicsSuite	
ATARI	Runs ATARI on RNAi reagent-level data. RNAi	Proteomics
BedToGtf	Converts BED files to GFF or GTF format Data Format Conversion	
BlastTrainTest	Sequence similarity classification using BLAST Multi-label Protein Prediction Suite (MiPPS), Prediction	
Bowtie.aligner	Bowtie2 (v. 2.1.0) is an ultrafast and memory-efficient short read aligner. RNA-seq	
BWA.aln	[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]... RNA-seq	
BWA.indexer	[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]... RNA-seq	
CaArray2ImportViewer	Imports data files from CaArray 2.4.1 and creates gct or cls files Visualizer	
CART	Classification and Regression Tree Prediction	
CBS	Segments DNA copy number data into regions of estimated equal copy number using circular binary... SNP Analysis	
ChIPSeq.CreateHeatmap	Generates a heatmap based on the ChIP-Seq signal extracted from a BAM file, according to the... ChIP Seq	

Solution features: Reproducibility

- Record and replay of all analyses
- Retain all versions of code – so results can be reproduced even if code changes
- Chain analyses into “pipelines”, or workflows that can be shared and published

GenePattern vocabulary: Modules

Copy Number
Divide
by Normals

GSEA

Variation
Filter

GISTIC

CBS

k-Nearest
Neighbors

Classification
and
Regression Trees

Support
Vector
Machines

Hierarchical
Clustering

GISTIC

Expression
File
Creator

Metagene
Projection

GenePattern vocabulary: Modules

Copy Number
Divide
by Normals

GISTIC

Classification
and
Regression Trees

GISTIC

GSEA

CBS

Support
Vector
Machines

Expression
File
Creator

Variation
Filter

k-Nearest
Neighbors

**Hierarchical
Clustering**

Metagene
Projection

Hierarchical Clustering

Files

HCL.jar
cluster.exe
ant.jar
gp-modules.jar
Jama-1.0.2.jar

Documentation

HierarchicalClustering.pdf

Parameter descriptions

```
-f <input.filename>
  <log.transform>
  <row.center>
  <row.normalize>
  <column.center>
  <column.normalize>
-u <output.base.name>
-e <column.distance.measure>
-g <row.distance.measure>
-m <clustering.method>
```

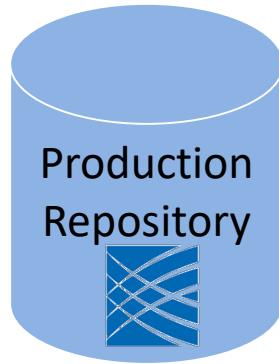
>250 GenePattern Modules, 4/2019

The screenshot shows the GenePattern web interface with a red box highlighting the 'Browse Modules > All Modules' section. The interface includes a navigation bar with links for Modules & Pipelines, Suites, Job Results, Resources, Help, and GenomeSpace. Below the navigation is a search bar and tabs for Modules, Jobs, and Files. A 'Favorite Modules' section allows users to drag modules here. A 'Recent Modules' section lists several viewers: ComparativeMarkerSelectionViewer, FeatureSummaryViewer, HeatMapView, and HierarchicalClusteringViewer. The main area displays a grid of 25 module cards, each with a title, description, and a gear icon for configuration.

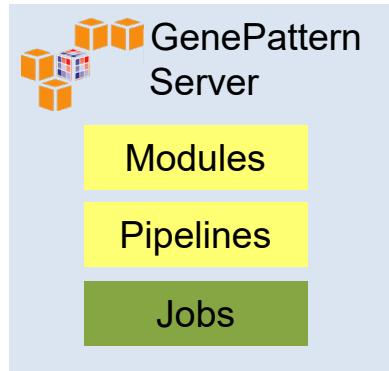
Browse Modules > All Modules	
ABSOLUTE Extracts absolute copy numbers per cancer cell from a mixed DNA population. Use this module... SNP Analysis	ABSOLUTE.review Extracts the absolute copy number per cancer cell from a mixed DNA population. Use this module... SNP Analysis
ABSOLUTE.summarize Summarizes the results from multiple ABSOLUTE runs so that an analyst can manually review the solutions. SNP Analysis	AddFCSEventIndex Adds indexes to events in a Flow Cytometry Standard (FCS) data file. Flow Cytometry
AddFCSParameter Add parameters and their values to a FCS data file Flow Cytometry	AddNoiseToFCS Add noise to specified parameters in an FCS data file. Flow Cytometry
aml.all.pipeline ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline	ApplyGatingML Apply a Gating-ML file on an FCS data file (gate and/or transform list mode data) Flow Cytometry
ARACNE Runs the ARACNE algorithm for reverse engineering cellular networks Pathway Analysis	AreaChange Calculates fraction of area under the spectrum that is attributable to signal (area after noise)... Proteomics , ProteomicsSuite
Arff2Gct Convert an .arff file into a gene pattern .gct / .cls file pair Multi-label Protein Prediction Suite (MiPPS) , Preprocess ...	ATARIS Runs ATARIS on RNAi reagent-level data RNAi
AuDIT Automated Detection of Inaccurate and Imprecise Transitions in MRM Mass Spectrometry Proteomics	BedToGtf Converts BED files to GFF or GTF format Data Format Conversion
Beroukhim.Getz.2007.PNAS.Glioma.GI! pipeline	Birdseed SNP genotyping algorithm that runs on the Affymetrix 500K, SNP5.0, and SNP6.0 platforms SNP Analysis
BirdseedCallRate Computes the call rate of the Birdseed algorithm SNP Analysis	BirdseedDataPreparation Prepare a bsnp file for running Birdseed SNP Analysis
BlastTrainTest Sequence similarity classification using BLAST Multi-label Protein Prediction Suite (MiPPS) , Prediction	BlastXValidation Sequence similarity cross validation prediction using BLAST Multi-label Protein Prediction Suite (MiPPS) , Prediction
Bowtie.aligner Bowtie2 (v. 2.1.0) is an ultrafast and memory-efficient short read aligner. RNA-seq	Bowtie.indexer Builds a Bowtie2 (v. 2.1.0) index from a set of DNA sequences RNA-seq
BWA.aln	BWA.bwasw

How GenePattern works

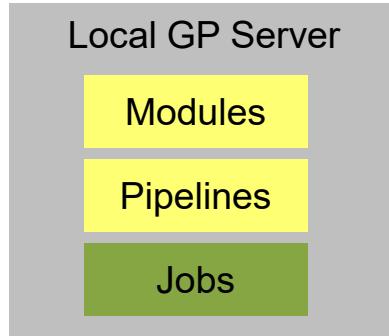
Repositories



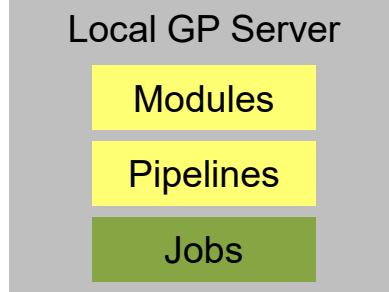
Servers



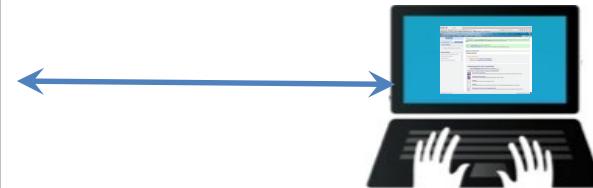
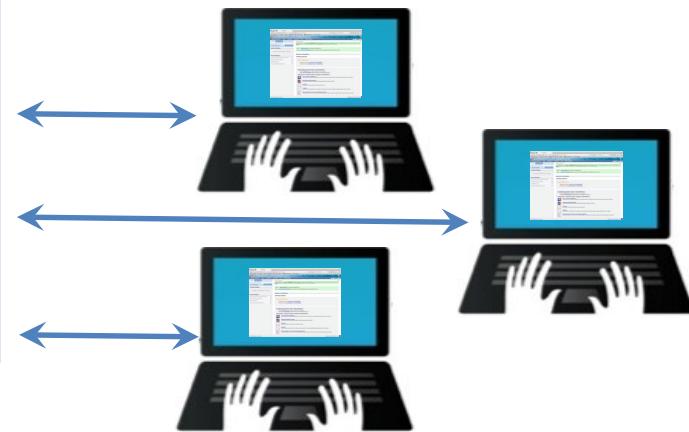
Local GP Server



Local GP Server



GenePattern Users





GenePattern Notebook Environment

GenePattern Notebook Tutorial GenePattern Notebook Tutorial (unsaved changes)

Control Panel Logout gpdemo

File Edit View Insert Cell Kernel Widgets Help

Not Trusted Python 3.6

Markdown Tools

GenePattern Notebook Tutorial

The GenePattern Notebook Environment provides a variety of features for both basic and advanced users. This tutorial will familiarize you with some of its most important features.

All instructions for you to follow will appear in a blue panel like this one.

GenePattern Notebook Introduction Video

Below is a brief video introduction to the GenePattern Notebook Environment. This video introduces many of the basic concepts and features provided by the tool. If you would prefer a more "hands on" introduction, scroll down and follow the subsequent interactive tutorial.

To view the video, click the Play button in the middle of the video cell.

If the video is not visible, highlight the cell below and press the Run Cell (▶) button in the toolbar to see the video.

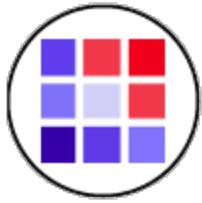
In [3]: `%HTML
<iframe width="854" height="480" src="https://www.youtube.com/embed/r5Km4UPhb1Q" frameborder="0" allowfullscreen></iframe>`

GenePattern Notebook Environment

jupyter Untitled Last Checkpoint: 2 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Help

Test Dataset → Bayesian Predictor of Outcome → Probability of Relapse



Jupyter Notebook

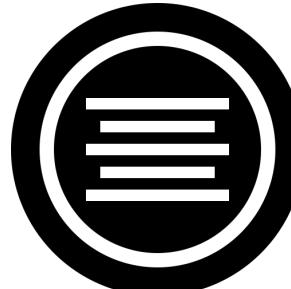
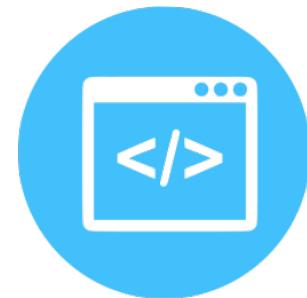
- Popular and well-supported framework for scientific computing
- Ecosystem of available extensions and resources
- Open source

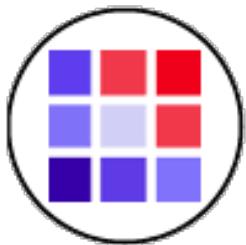




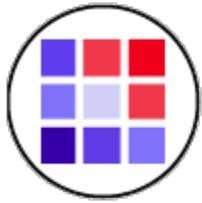
Complete Research Narrative

- Leverages the best of Jupyter and GenePattern
- Interleave text, visualization, graphics and analytical aspects





Exercise



GenePattern Notebook Repository

The screenshot shows the GenePattern Notebook login interface. The background is a blue-toned image of a DNA double helix and some handwritten mathematical or scientific notes. On the left, there's a 'Sign in' form with fields for 'Username' and 'Password', and buttons for 'Sign In' and 'Forgot Password?'. In the center, there's a 'GenePattern Server Status' box containing a link to the 'GenePattern Help Forum'. On the right, a green-highlighted box contains instructions for logging in and a 'Register a New GenePattern Account' button. A note at the bottom of this box directs users to the GenePattern Notebook website for documentation.

GenePattern Notebook

Sign in

Username:

Password:

Sign In Forgot Password?

GenePattern Server Status

For more information please contact us on the [GenePattern Help Forum](#)

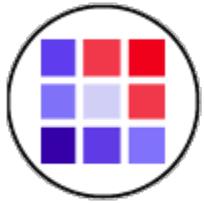
Register GenePattern Account

Log in using your GenePattern public server username and password. If you do not have an account, click the Register Account button below.

Register a New GenePattern Account

Documentation is available on the GenePattern Notebook website.

<https://notebook.genepattern.org>



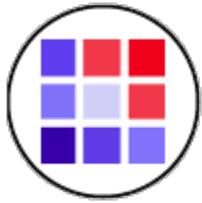
Notebook Workspace

- Lists your private notebooks and associated files.
- Copy, move, rename, delete, download and publish notebooks from here.

The screenshot shows the GenePattern Notebook web application. At the top, there is a dark header bar with the "GenePattern Notebook" logo on the left, a "Logout bhill@broadinstitute.org" link, and a "Control Panel" button on the right. Below the header, there is a navigation bar with three tabs: "Files" (selected), "Running", and "Notebook Library". A message "Select items to perform actions on them." is displayed above the file list. The main area contains a table of files:

	Name	Last Modified	File size
<input type="checkbox"/> 0	/		
<input type="checkbox"/>	GenePattern Notebook Tutorial.ipynb	5 months ago	25.2 kB
<input type="checkbox"/>	GenePattern Python Tutorial.ipynb	5 months ago	17.4 kB

At the top right of the file list, there are buttons for "Upload", "New", and a refresh icon. On the far left, there is a sidebar with a "Files" tab and a small thumbnail preview.

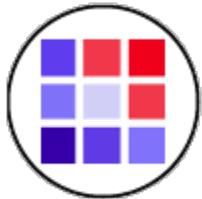


Browse Public Notebooks

- A variety of public notebooks are available in the GenePattern Notebook Library.
- Anyone can make a copy of these notebooks to read, run and reproduce.

The screenshot shows the GenePattern Notebook interface. At the top, there's a navigation bar with tabs for 'Files', 'Running', and 'Notebook Library'. The 'Notebook Library' tab is active. On the left, there's a sidebar with categories like 'Public Notebooks' (with 'Featured' selected), 'All Notebooks', 'Tutorial', 'Workshop', 'Community', 'My Notebooks', and 'Shared Notebooks' (with 'Shared By Me' and 'Shared With Me'). The main area is titled 'Featured Notebooks' and contains a table with the following data:

Notebook	Authors	Updated	Quality
Classification and Prediction - RNAseq Use RNA-seq data with k-Nearest Neighbors (kNN) to build a predictor, use it to classify leukemia subtypes, and assess its accuracy in cross-validation. <small>featured</small>	GenePattern Team	2018-03-20	Release
Classification and Prediction Example of how to use k-Nearest Neighbors (kNN) to build a predictor, use it to classify leukemia subtypes, and assess its accuracy in cross-validation. <small>featured</small>	GenePattern Team	2018-03-20	Release
Differential Expression Analysis Example of using differential expression analysis to find genes that are significantly differentially expressed between classes of samples. <small>featured</small>	GenePattern Team	2018-03-19	Release
Hierarchical Clustering - RNASeq Use RNA-seq data to cluster genes and/or samples agglomeratively, based on how close they are to one another. <small>featured</small>	GenePattern Team	2018-03-19	Release



Run an Analysis Notebook

2019-04-30-BU Molecular Bio – Intro to RNA-seq in GenePattern

The screenshot shows the GenePattern Notebook interface. At the top is a header bar with the title "GenePattern Notebook" and the date "2019-04-04-BU Molecular Bi...". It also shows "Last Checkpoint: an hour ago (autosaved)", "Logout gpdemo", and "Control Panel". Below the header is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". A toolbar below the menu bar contains icons for file operations like Open, Save, Print, and Run, along with "Markdown" and "Tools" dropdowns.

Differential Expression of RNA-Seq data in GenePattern Notebook

Compute differentially expressed genes or transcripts and visualize the results

Before you begin

You must log in to a GenePattern server, in this notebook we will use the **GenePattern Public Server**, hosted in the Amazon cloud.

Instructions

- Sign in to GenePattern by clicking "Login as..." in the dialog that should be displayed below.
 - This will log you in as the same user you used to log into the notebook repository.

GenePattern GPDemo

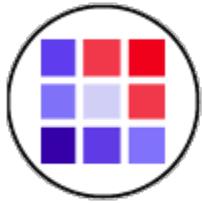
<https://cloud.genepattern.org/gp>



RNA-Seq Differential Analysis Workflow

As we progress through RNA-seq analysis, we are looking to answer several questions. The first question is – what does my data look like? Does it look approximately the way most RNA-seq data looks, or are there significant issues which suggest a problem with the sequencing? Are there biases toward 3' or 5' ends, different bases, different sequences?

Once we've determined your data is within acceptable bounds, then we can proceed with alignment and downstream analyses. Then, we are asking other questions, such as – how did my data align? How many of the reads aligned? Are there visual differences between the phenotypes?



Name, Save & Checkpoint Notebooks

- Name or rename notebooks
- Save or revert to a checkpoint
- Make a duplicate notebook

The screenshot shows the GenePattern Notebook Environment interface. The top navigation bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. The status bar indicates "Not Trusted" and "Python 3.6". The main content area displays a "pattern Notebook Tutorial" with text about its features and a "pattern Notebook Introduction Video" section with a video player.

File

- New Notebook
- Open...
- Make a Copy...
- Save as...
- Rename...
- Save and Checkpoint
- Revert to Checkpoint**
- Print Preview
- Download as
- Publish to Repository
- Share with Collaborators
- Trust Notebook
- Close and Halt

pattern Notebook Tutorial

pattern Notebook Environment provides a variety of features for both basic and advanced users. This tutorial will familiarize you with some of its important features.

Actions for you to follow will appear in a blue panel like this one.

pattern Notebook Introduction Video

Brief video introduction to the GenePattern Notebook Environment. This video introduces many of the basic concepts and features provided by the environment. You may also prefer a more "hands on" introduction, scroll down and follow the subsequent interactive tutorial.

To view the video, click the Play button in the middle of the video cell.

If the video is not visible, highlight the cell below and press the Run Cell (▶) button in the toolbar to see the video.



GenePattern Cells

Authentication Cell

GenePattern Login

GenePattern Server

GenePattern Cloud

GenePattern Username

Username

GenePattern Password

Password

Log into GenePattern Register an Account

Analysis Cell

GenePattern ConvertLineEndings Version 2

Converts line endings to the host operating system's format.

Run

input filename* https://cloud.genepattern.org/gp/jobResults/104715/all_aml_train.preprocessed.gct

The input file (any non-binary file format)

output file* <input.filename_basename>.cvt.<input.filename_extension>

The output file

Run

Job #104716

Completed Submitted by GPDemo on 2019-03-29T18:29:45+00:00

all_aml_train.preprocessed.cvt.gct

gp_execution_log.txt



Authentication Cells

GenePattern Login

GenePattern Server
GenePattern Cloud

GenePattern Username
Username

GenePattern Password
Password

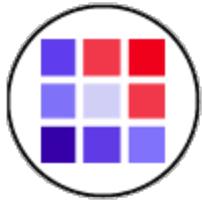
Log into GenePattern **Register an Account**

GenePattern GPDemo <https://cloud.genepattern.org/gp>

-- March 18, 2019 -- Due to recent browser security updates, cracking down on mixed content, some of our viewers are unable to display their content. We are currently only aware of one module (ConstellationMap) being impacted. Please let us know if you discover others. Sincerely, The GenePattern Team Follow the GenePattern team on Twitter, Instagram or Facebook to keep up with the latest news and events or join the conversation in our forum!

Experiencing a bug? Have thoughts on how to make GenePattern Notebook better?
Let us know by leaving feedback.

Leave Feedback



Analysis Cells

- Drag [SRR1039508_1.fastq.gz](#) to the **input file** field.
- Leave the rest of the parameters as default (found in the Advanced section)
- Click **Run**

GenePattern FastQC Version 1

Provides quality control metrics on raw sequence data **Run**

Basic -

input file*

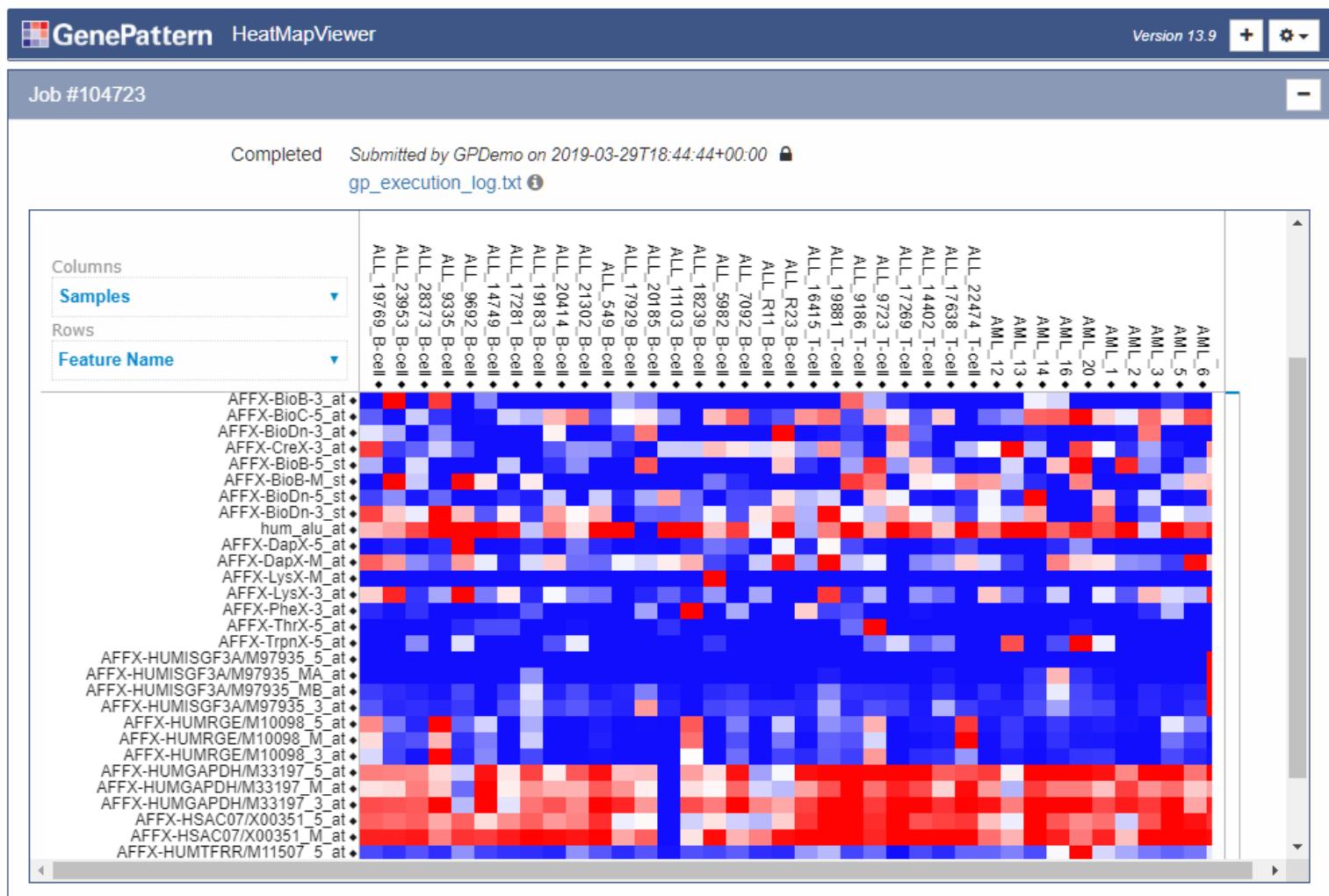
RNA-seq reads file in FASTQ (bz2 and gz compressed files are supported), SAM, or BAM format.

Advanced +

Run

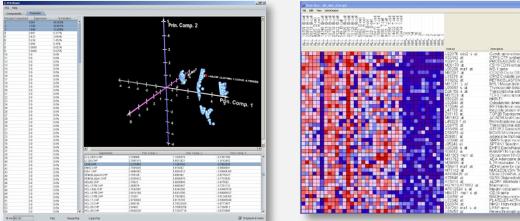
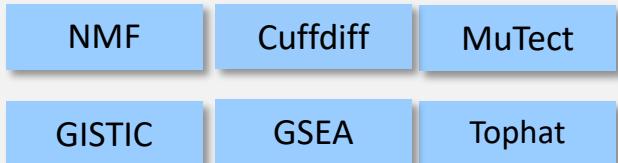


Job/Result Cells



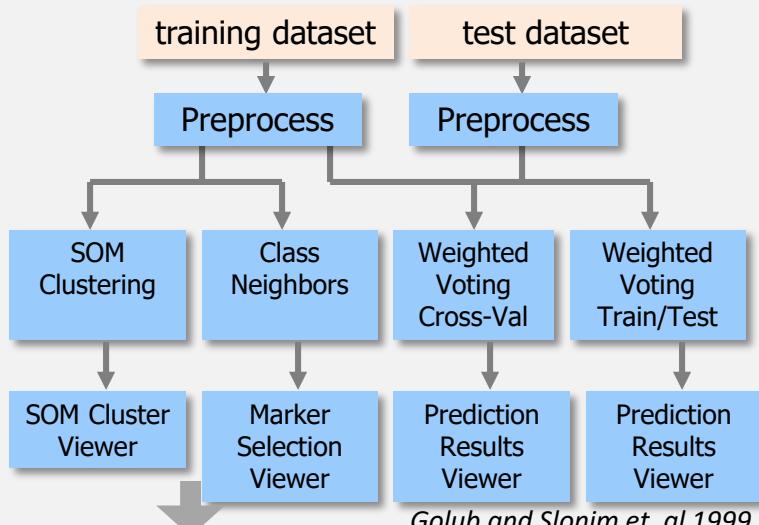
The GenePattern Ecosystem: Architecture

Module Repository



Hundreds of analysis and visualization tools

Pipeline Environment



Golub and Slonim et. al 1999

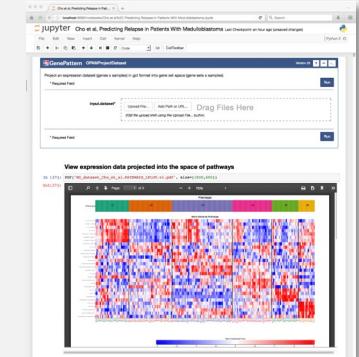


Support for *in silico* reproducible research

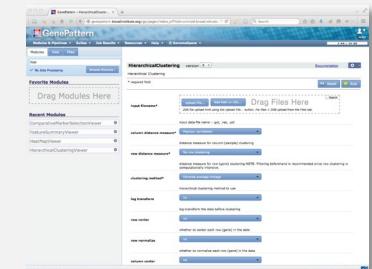
Analysis Engine



Record/replay analyses
Versioning of methods
Web service access



Notebook

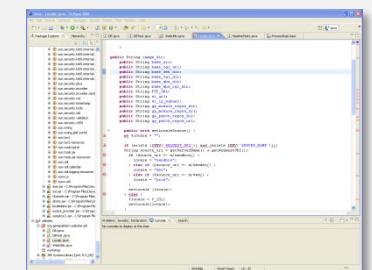


Web

Module Integrator



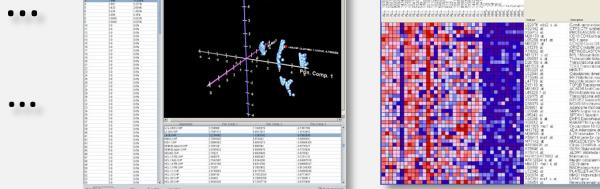
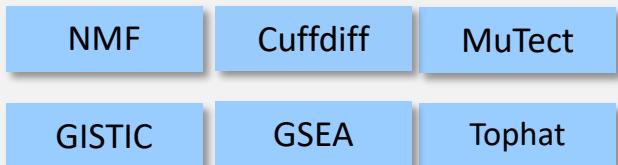
Easy addition of new tools



Programming
Access for all levels of user

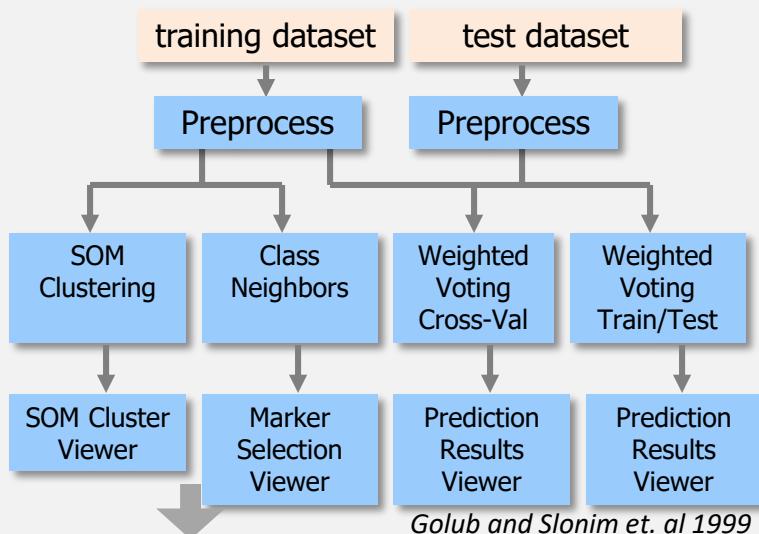
The GenePattern Ecosystem: Architecture

Tool Library

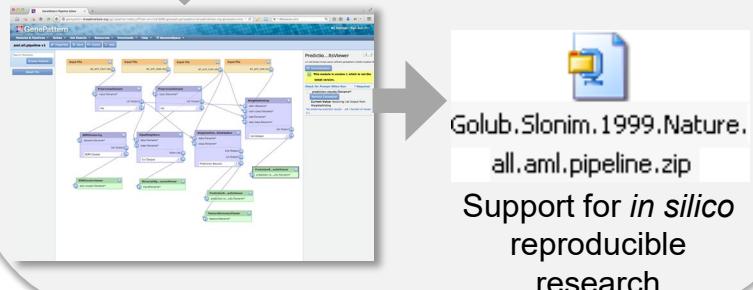


Hundreds of analysis and visualization tools

Workflows



Golub and Slonim et. al 1999

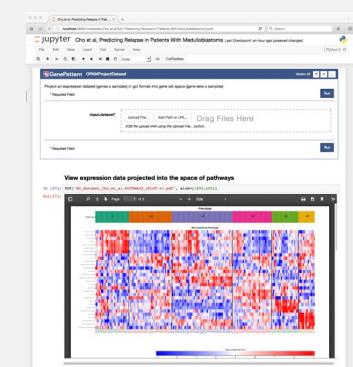


Analysis Engine

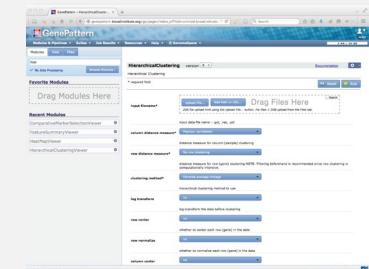


Record/replay analyses
Versioning of methods
Web service access

Access Points

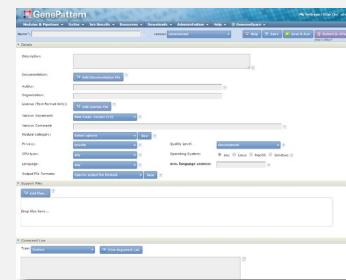


Notebook

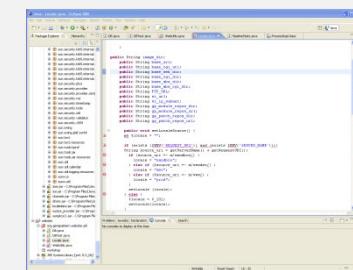


Web

Create your own tool

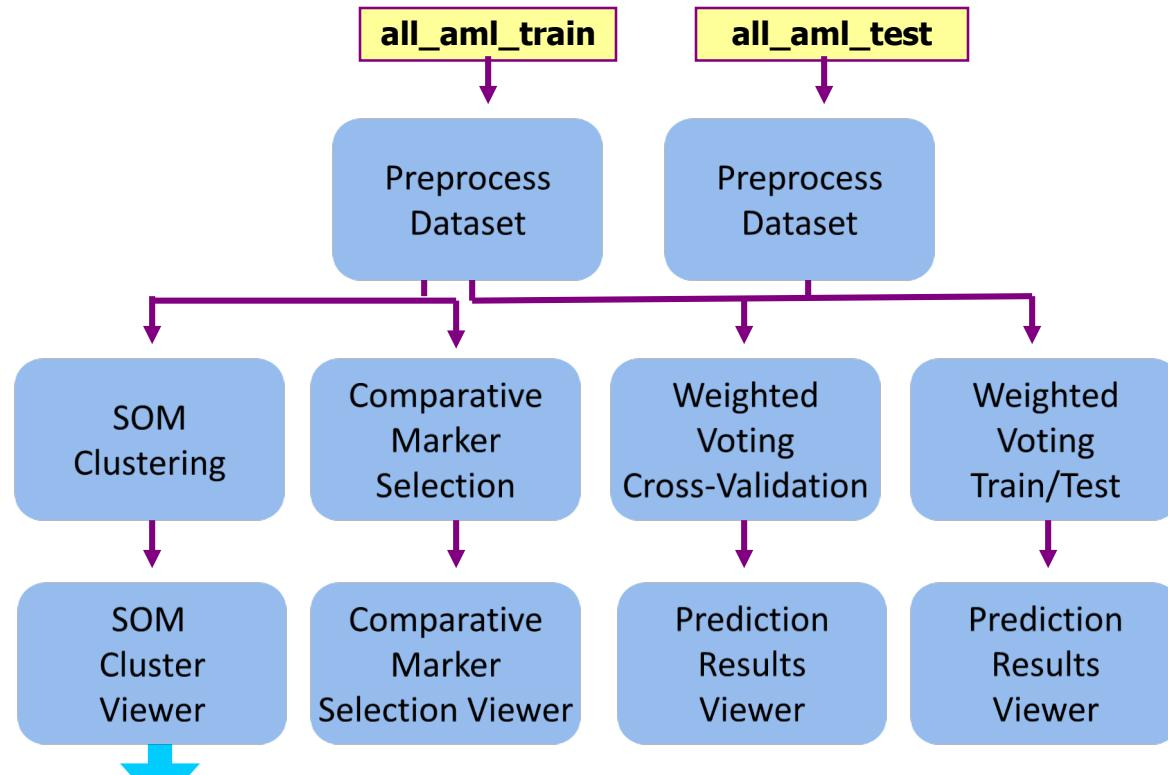


Easy addition of new tools



Programming
Access for all levels of user

GenePattern vocabulary: Pipelines



Golub.Slonim.1999.Nature.
all.aml.pipeline.zip

Pipelines in GenePattern

The screenshot shows the GenePattern web interface. On the left, there's a sidebar with tabs for 'Modules & Pipelines', 'Suites', 'Job Results', and 'Resources'. Below these are sections for 'Favorite Modules' (with a placeholder 'Drag Modules Here') and 'Recent Modules' (including ComparativeMarkerSelectionViewer, FeatureSummaryViewer, HeatMapView, and HierarchicalClusteringViewer). The main content area is titled 'Browse Modules > pipeline' and contains a grid of 18 pipeline entries, each with a title, a brief description, and a gear icon for settings. A red box highlights this grid. To the right of the grid, there's a large green box containing release notes, and below it, two light blue boxes with partial text visible.

Module Title	Description
aml.all.pipeline	ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline
CBSWrapperPipeline	A one step pipeline that runs CBS pipeline
CopyNumberInferencePipeline.Part2of:	A pipeline that runs CopyNumberInferencePipeline.Part2of2 – Part of the pipeline
FLAMEContourViewer.Pipeline	Pipeline which runs the FLAMECounterDataGenerator and the FLAMEViewer pipeline
Golub.Slonim.1999.Nature.all.aml.pipe	ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline
IlluminaDASLPipeline	creates a GenePattern gct file from raw Illumina scan data pipeline
ImmPort_FLOCK_Individual_FCS	ImmPort FLOCK and Individual FCS pipeline pipeline
job212786	describe it here pipeline
job437446	describe it here pipeline
MGED_Reich	test pipeline pipeline
PWRGPTTestAuto_InheritType_Vis	Automated pipeline with file input as stored path (ie saved with the pipeline) and text inputs.... pipeline
Rot13Madness	
Beroukhim.Getz.2007.PNAS.Glioma.GI!	* pipeline
CopyNumberInferencePipeline.Part2of:	Second half of Pipeline for processing SNP 6 data pipeline
CufflinksCuffmergePipeline	[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]Creates...
GetDataSetInSilico	downloads a compressed .tgz file from the Insilico servers and extract it pipeline
Golub_Slonim	ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline
ImmPort_FLOCK_CrossSample	ImmPort FLOCK and CrossSample pipeline pipeline
job212108	describe it here pipeline
job298686	describe it here pipeline
Lu.Getz.Miska.Nature.June.2005.mous	Normal/tumor classifier and kNN prediction of mouse lung samples LuGetzMiska.Nature.2005.Suite, pipeline
ParallelICBS	Runs CBS algorithm on multiple samples in parallel pipeline
RNaseQC_CEGS	pipeline
ScripturePipeline	

GenePattern

Modules & Pipelines Suites Job Results Resources Help GenomeSpace

Modules Jobs Files

Search Modules & Pipelines

No Jobs Processing Browse Modules >

Favorite Modules Drag Modules Here

Recent Modules

ComparativeMarkerSelectionViewer

FeatureSummaryViewer

HeatMapView

HierarchicalClusteringViewer

Browse Modules > pipeline

aml.all.pipeline ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline

CBSWrapperPipeline A one step pipeline that runs CBS pipeline

CopyNumberInferencePipeline.Part2of: A pipeline that runs CopyNumberInferencePipeline.Part2of2 – Part of the pipeline

FLAMEContourViewer.Pipeline Pipeline which runs the FLAMECounterDataGenerator and the FLAMEViewer pipeline

Golub.Slonim.1999.Nature.all.aml.pipeline ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline

IlluminaDASLPipeline creates a GenePattern gct file from raw Illumina scan data pipeline

ImmPort_FLOCK_Individual_FCS ImmPort FLOCK and Individual FCS pipeline pipeline

job212786 describe it here pipeline

job437446 describe it here pipeline

MGED_Reich test pipeline pipeline

PWRGPTTestAuto_InheritType_Vis Automated pipeline with file input as stored path (ie saved with the pipeline) and text inputs.... pipeline

Rot13Madness

Beroukhim.Getz.2007.PNAS.Glioma.GI! * pipeline

CopyNumberInferencePipeline.Part2of: Second half of Pipeline for processing SNP 6 data pipeline

CufflinksCuffmergePipeline [**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors]Creates...

GetDataSetInSilico downloads a compressed .tgz file from the Insilico servers and extract it pipeline

Golub_Slonim ALL/AML methodology, from Golub and Slonim et al., 1999 pipeline

ImmPort_FLOCK_CrossSample ImmPort FLOCK and CrossSample pipeline pipeline

job212108 describe it here pipeline

job298686 describe it here pipeline

Lu.Getz.Miska.Nature.June.2005.mous Normal/tumor classifier and kNN prediction of mouse lung samples LuGetzMiska.Nature.2005.Suite, pipeline

ParallelICBS Runs CBS algorithm on multiple samples in parallel pipeline

RNaseQC_CEGS pipeline

ScripturePipeline

you through the new features. See the release notes for more

2 MB

About GenePattern | Contact Us

©2003-2014 Broad Institute, MIT

Community Activity

- Current version: 3.9.11 rc.4 b210 (2/2019)
- >50,000 registered users
- Open source, BSD-style license
- New Public server runs ~2,500 analyses/week
- GParc: GenePattern community repository
 - ~100 community-contributed methods
 - CRISPR analysis
 - Bisulfite sequencing
 - Flow cytometry
 - RNAi screens

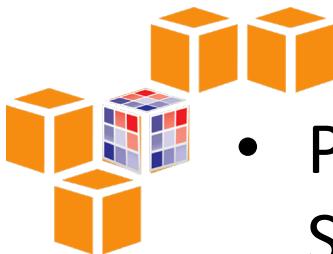
www.genepattern.org

The screenshot shows the GenePattern homepage with a dark blue header featuring the logo and navigation links: Run, Learn, Modules, Analytics, Resources, Contact, Help, and Search. Below the header is a banner with the text "GenePattern A Platform for Reproducible Bioinformatics". There are three circular icons: "Use GenePattern", "GenePattern Basics", and "Community". The main content area has sections for "Features" (describing powerful genomics tools in a user-friendly interface), "New: GenePattern Notebooks" (mentioning the transition to a Jupyter Notebook environment), and "Analysis Pipelines" (describing how GenePattern notebooks can be used to create analysis pipelines). On the right, there's a sidebar with "Blog > GP Updates" and a list of recent posts.

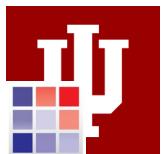
www.gparc.org

The screenshot shows the GParc homepage with a dark blue header featuring the logo and navigation links: Upload a Module, Resources, Contact Us, and Login. Below the header is a banner with the text "GParc A repository and community where users can share and discuss their own GenePattern modules". There are two buttons: "Learn More" and "Upload a module". The main content area has a section titled "Browse Modules" with a search bar and a list of modules. One module listed is "Acgh2Tab v4", which converts acgh files to a tab-delimited format. To the right of the module list is a "Filter By Available Tags" sidebar with several tags listed, including "CRISPR" which is highlighted in yellow. At the bottom of the page is a footer with links to "View All Tags", "Annotation", "Bisulfite Sequencing", "BisulfiteConversion", "Clustering", "ConceptualMarkerSelection", and "ConceptualMarkerConversion".

Availability



- Public server running on Amazon Web Services (cloud.genepattern.org)

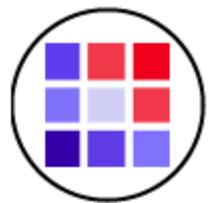


- Public server at Indiana U, backed by Carbonate HPC cluster (gp.indiana.edu)



- Downloadable server (www.genepattern.org)
- Amazon Machine Image (AMI)





Data Formats

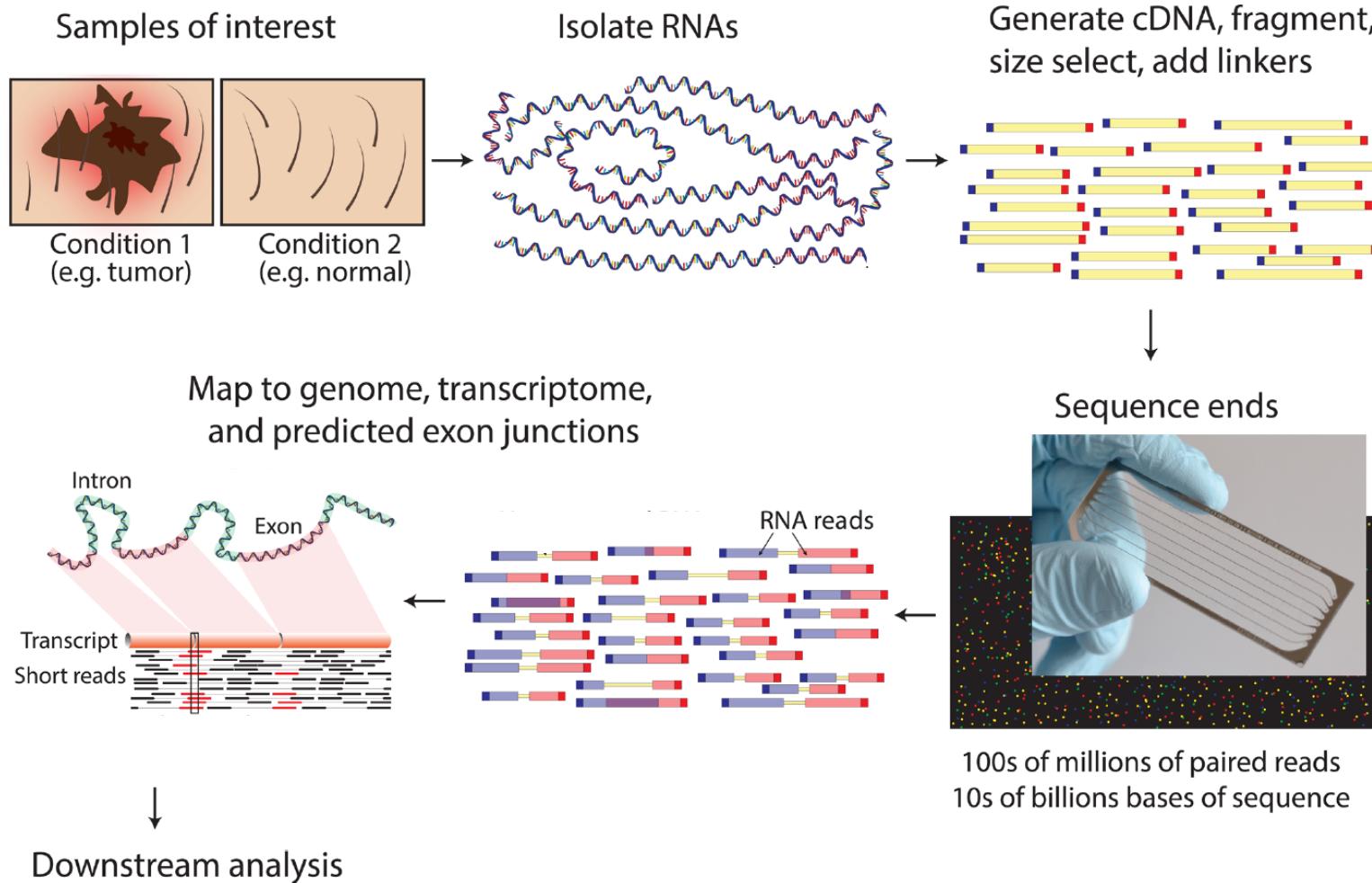
Supported Platforms

Platform	Data type	Module
Illumina	RNA-Seq	RNA-Seq Suite (Bowtie, TopHat, Cufflinks... etc)
	Whole Genome Expression	IlluminaExpressionFileCreator
	6K DASL	IlluminaDASLPipeline
Affymetrix	Gene expression chips	ExpressionFileCreator
	SNP chips	SNPFileCreator
	SNP6 chips	CopyNumberInferencePipeline
Agilent (In development)	microRNA, GeneView	AgilentExpressionFileCreator

TODAY: We will work with RNA-Seq data, sequenced using the Illumina platform

What is RNA-seq?

Snapshot of sequence/quantity of RNA being expressed in a genome.



RNA-seq data types

Sequence data comes in many varieties. FASTQ/BAM files are common.

File format	Description	Human readable?
FASTA	Raw sequence data with identifier	Yes
FASTQ	Raw sequence data with identifier, and quality information about the sequence (Phred quality scores)	Yes
SAM	Identifies alignment of sequences with quality scores against a template	Yes
BAM	Compressed binary version of a SAM file. Can also be un-aligned sequences, e.g. compressed binary version of FASTQ	No

TODAY: Starting with FASTQ files, we align raw reads against a reference genome.

FastQ Format

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a [FASTA](#) title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!''*(((***+))%%%+)(%%%%.1***-+*'')**55CCF>>>>CCCCCCC65
```

The character '!' represents the lowest quality while '~' is the highest. Here are the quality value characters in left-to-right increasing order of quality ([ASCII](#)):

```
!"#$%&'()*+,./0123456789:;<=>?@ABCDEFGHIJKLMNPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

Illumina sequence identifiers [edit]

Sequences from the [Illumina](#) software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

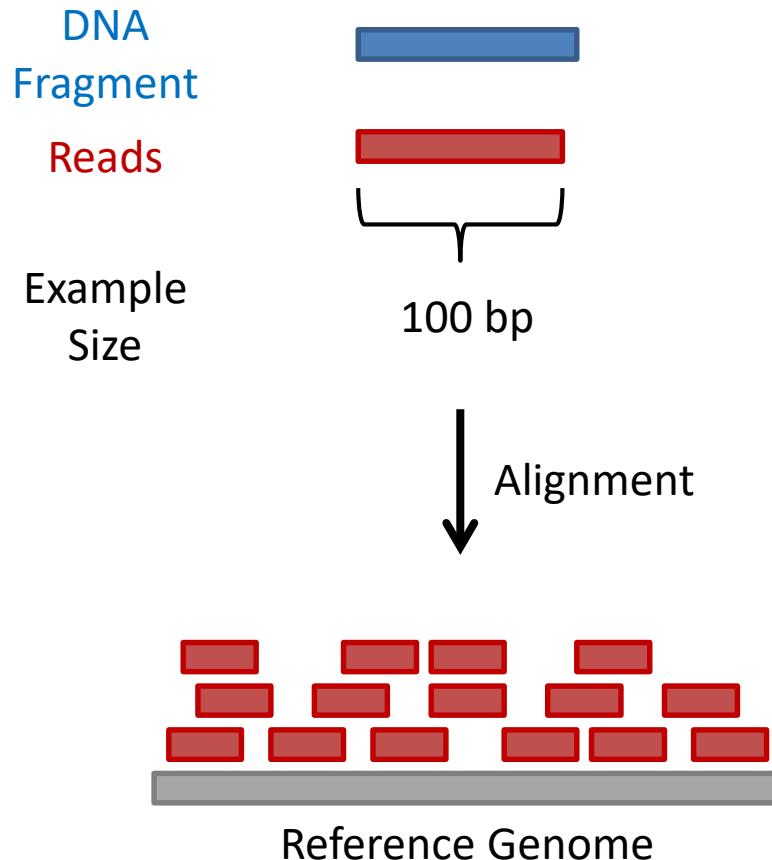
NCBI Sequence Read Archive [edit]

FASTQ files from the [NCBI/EBI](#) Sequence Read Archive often include a description, e.g.

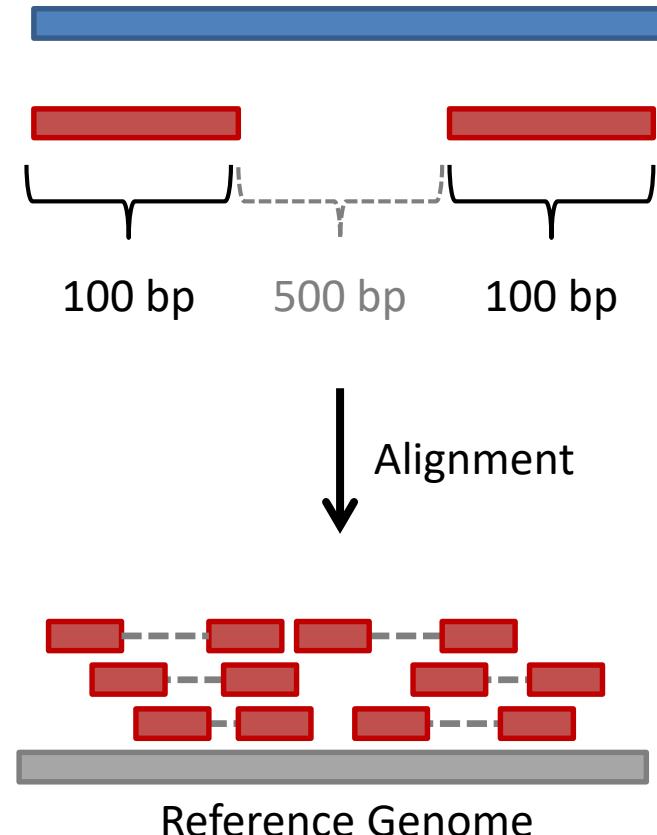
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGCCGCTGCCGATGGCGTCAAATCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

RNA-seq read types

Single-end reads



Paired-end reads



TODAY: We will use paired-end reads.

Importing and Converting Data

Importing Data

SraToFastQ

GEOImporter

MAGEMLImportViewer

MAGETABImportViewer

Converting data

Picard (BAM/SAM)

BedToGtf (RNA-seq annotation)

CsvToFcs (Flow Cytometry)

FcsToCsv (Flow Cytometry)

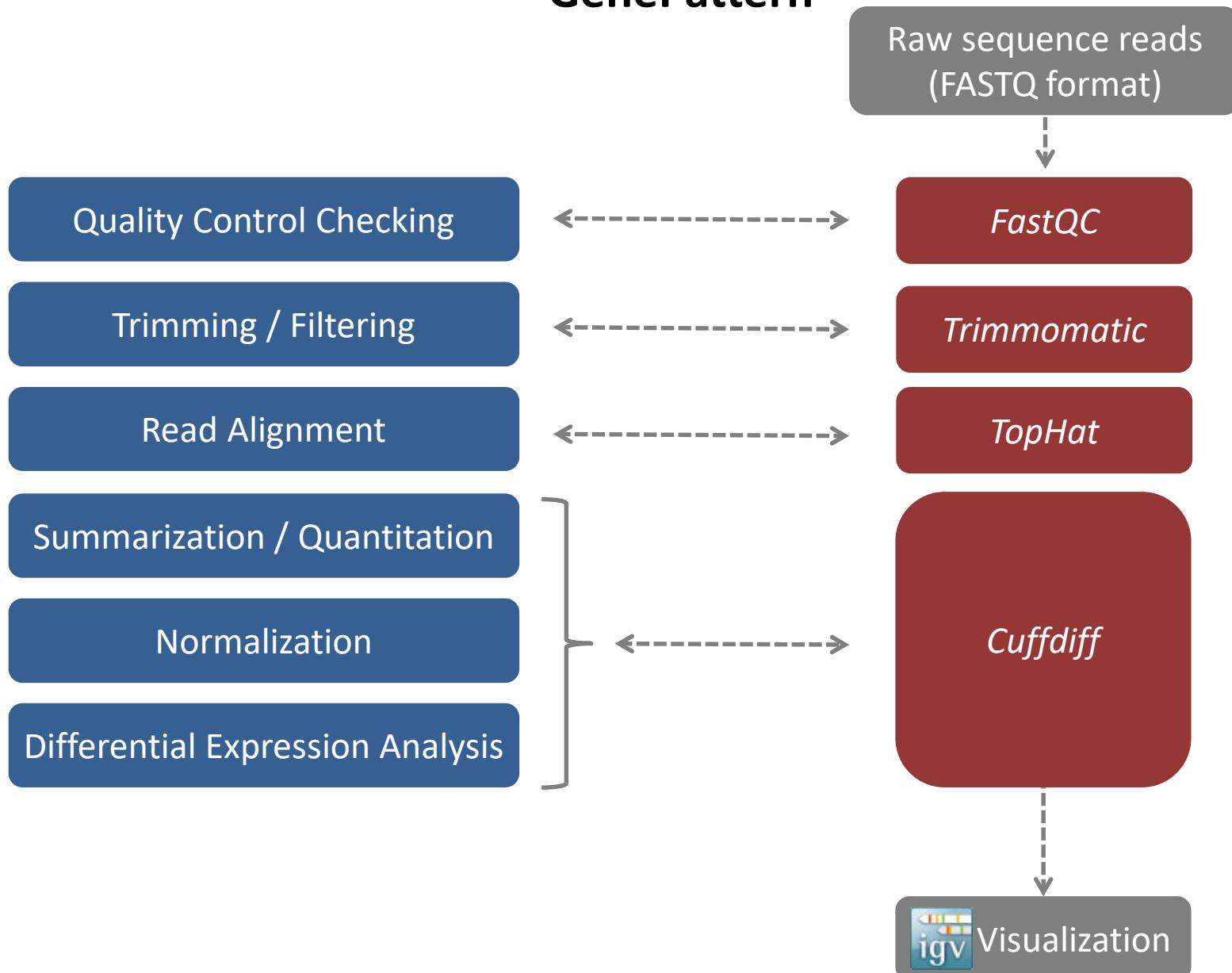
Read_group_trackingToGct (Cufflinks)

MergeHTSeqCounts (HTSeq)

Other data formats

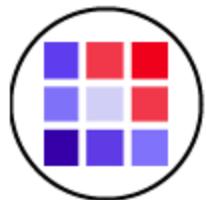
- Many others available
- Manually convert data to required format or post in our forum: genepattern.org/help

Modules Supporting RNA-Seq Differential Expression Workflow in GenePattern



GenePattern modules





Running Analyses

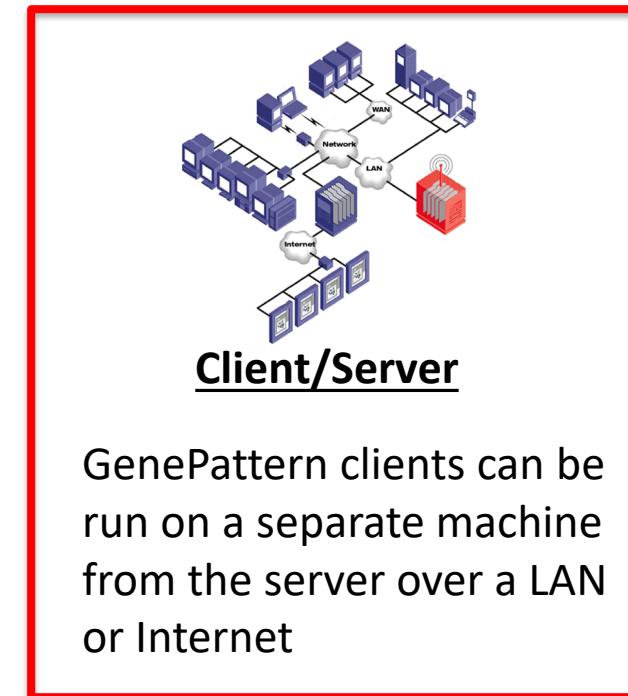
GenePattern configurations

- Use a publicly hosted GenePattern server (Recommended)
- Install your own GenePattern server

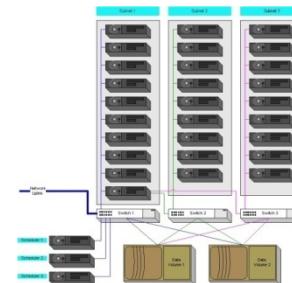


Standalone

GenePattern can be run self contained on a laptop or desktop machine



GenePattern clients can be run on a separate machine from the server over a LAN or Internet



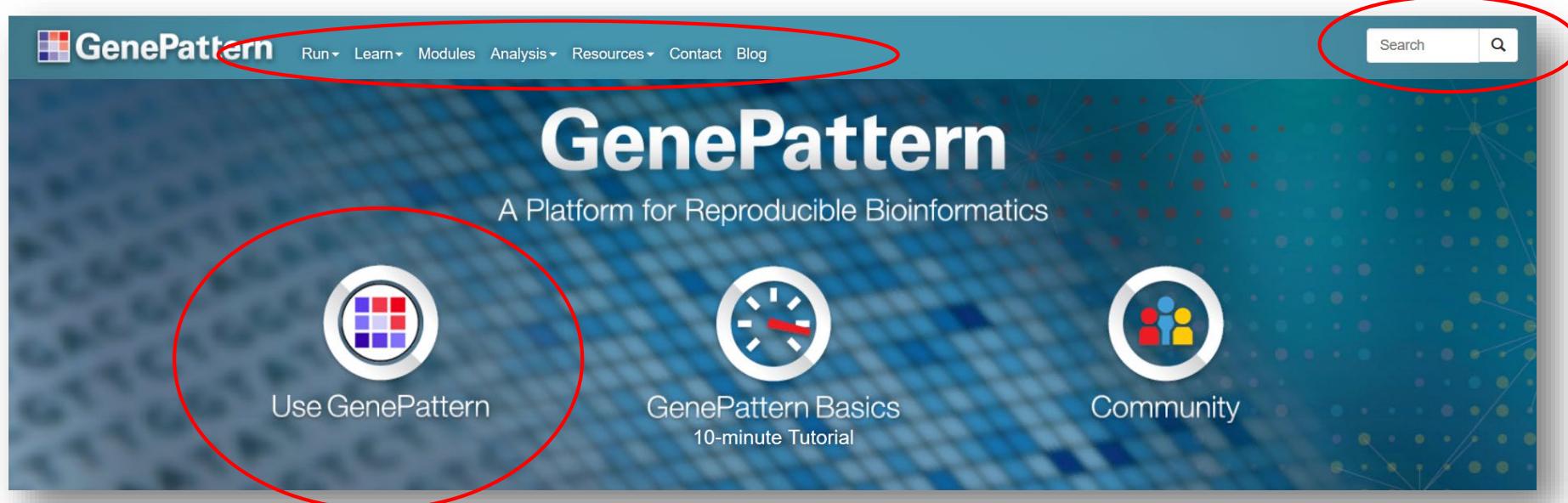
Cluster

GenePattern can be configured to run on a cluster using AWS Batch, GridEngine, PBS/Torque, SLURM, or LSF software

Note: GenePattern server & client are separate software

GenePattern Website

<http://genepattern.org>



GenePattern UI – main page

The screenshot shows the GenePattern main page as it would appear in a web browser. The URL in the address bar is genepattern.broadinstitute.org/gp/pages/index.jsf. The page has a blue header with the GenePattern logo and navigation links for Modules & Pipelines, Suites, Job Results, Resources, Help, and GenomeSpace. A user icon labeled "GPDemo" is in the top right. The left sidebar contains sections for Favorite Modules (TopHat, Cufflinks, ComparativeMarkerSelection, PreprocessDataset, HeatMapView) and Recent Modules (HierarchicalClustering, PreprocessDataset). The main content area features a "Welcome to GenePattern" message, a "Getting Started" section with "New! Web tours" (links to what's new and introductory tour), and a "Analyzing genomic data in GenePattern" section with protocols for analysis, differential expression, clustering, and prediction.

GenePattern

genepattern.broadinstitute.org/gp/pages/index.jsf

Welcome to GenePattern

Getting Started

New! Web tours

- Click here for a tour of [what's new in GenePattern](#).
- Click here for an [introductory tour of GenePattern](#).

Analyzing genomic data in GenePattern

Recent Modules

Favorite Modules

Protocols for running common analyses in GenePattern:

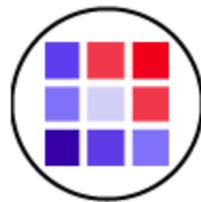
- Run an Analysis in GenePattern**
Learn how to run an analysis in GenePattern by preprocessing gene expression data and visualizing the resulting data as a heat map.
- Differential Expression Analysis**
Find genes that are significantly differentially expressed between classes of samples.
- Clustering**
Group genes and/or samples by similar expression profiles.
- Prediction**
Create a model, also referred to as a classifier or class predictor, that correctly classifies unlabeled samples into known classes.

About GenePattern | Contact Us

©2003-2014 Broad Institute, MIT

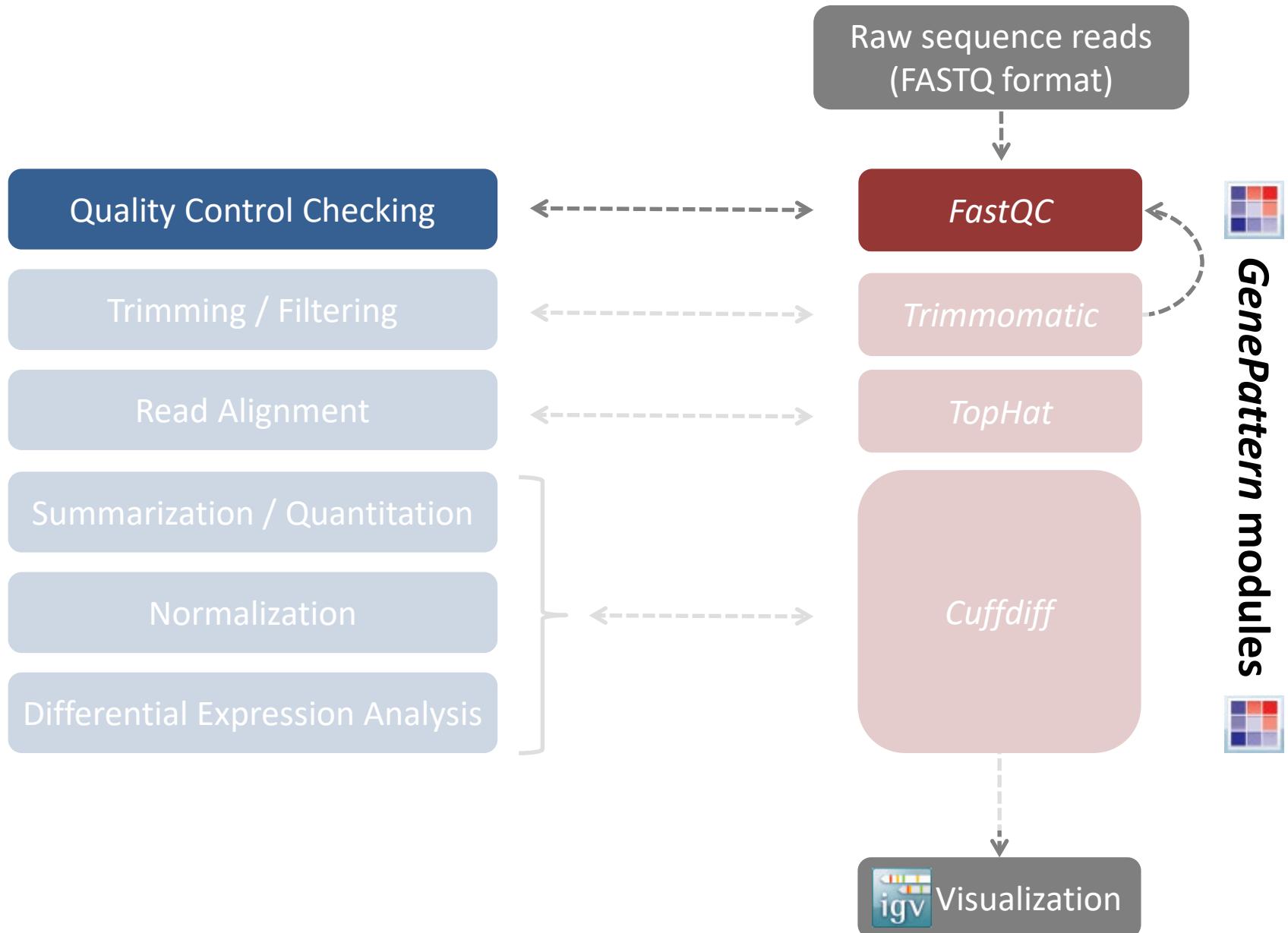
Running analyses

- Selecting a module
- Setting parameter values
- Running the module
- Viewing job status



Exercise

RNA-Seq Differential Expression Analysis Workflow



RNA-Seq Differential Expression Analysis Workflow

Quality Control Checking

Is my raw data of high quality?

Are any bases or sequences over- or under-represented?

Are there any biases in the reads (e.g. 3' end or 5' end)?

Is there any kind of enrichment bias in my samples (e.g. from PCR)?

Dataset: GSE52778, SRP033351

Human Airway Smooth Muscle Transcriptome Changes in Response to Asthma Medications

Himes, Blanca E. et al. "RNA-Seq Transcriptome Profiling Identifies *CRISPLD2* as a Glucocorticoid Responsive Gene That Modulates Cytokine Function in Airway Smooth Muscle Cells." Ed. Jan Peter Tuckermann. *PLoS ONE* 9.6 (2014): e99625. *PMC*.

- HASM cells from four white male donors
- Four treatment conditions:
 - 1) no treatment
 - 2) treatment with a β 2-agonist (i.e. Albuterol, 1 μ M for 18h)
 - 3) treatment with a glucocorticosteroid (i.e. Dexamethasone (Dex), 1 μ M for 18h)
 - 4) simultaneous treatment with a β 2-agonist and glucocorticoid,

What does *FastQC* do?

FastQC evaluates raw sequence data and determines the quality of reads before alignment. It can evaluate FASTQ, SAM, and BAM files.

Quality metric	Description
Basic Statistics	Simple stats (file type, # of sequences, % GC content, etc.)
Per base sequence quality	Boxplots of quality scores (Phred) across all bases at each position
Per sequence quality scores	Determines if a subset of sequences have unusually low quality values
Per base sequence content	Plots the proportion of A/T/C/G at each position in a sequence
Per base GC content	Proportion of GC content at each position in a sequence
Per sequence GC content	Measure of GC content across the length of each sequence
Per base N content	Percentage of N (null) content at each position in a sequence
Sequence Length Distribution	Distribution of fragment sizes
Sequence Duplication Levels	Degree of duplication for every sequence in a library
Overrepresented sequences	List of sequences (e.g. adapters) that are overrepresented in the library
Kmer Content	Enrichment of specific k-mers at each position in a sequence

FastQC results

FA

GenePattern 1657086. FastQC

Submitted by GPDemo on 2018-04-02T17:53:53-04:00 Completed

SRR1039508_1_fastqc.zip ⓘ
SRR1039508_1_fastqc/lcons/error.png ⓘ
SRR1039508_1_fastqc/lcons/fastqc_icon.png ⓘ
SRR1039508_1_fastqc/lcons/tick.png ⓘ
SRR1039508_1_fastqc/lcons/warning.png ⓘ
SRR1039508_1_fastqc/Images/duplication_levels.png ⓘ
SRR1039508_1_fastqc/Images/kmer_profiles.png ⓘ
SRR1039508_1_fastqc/Images/per_base_gc_content.png ⓘ
SRR1039508_1_fastqc/Images/per_base_n_content.png ⓘ
SRR1039508_1_fastqc/Images/per_base_quality.png ⓘ
SRR1039508_1_fastqc/Images/per_base_sequence_content.png ⓘ
SRR1039508_1_fastqc/Images/per_sequence_nc_content.png ⓘ
[SRR1039508_1_fastqc/fastqc_data.txt ⓘ](#)
[SRR1039508_1_fastqc/fastqc_report.html ⓘ](#)
[SRR1039508_1_fastqc/summary.txt ⓘ](#)
gp_execution_log.txt ⓘ

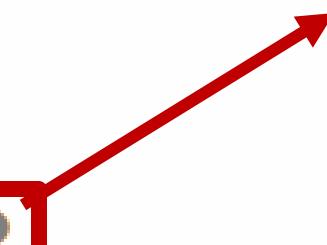
SRR1039508_1_fastqc/fastqc_report.html

[Download File](#)

[Open in New Tab](#)

[Send to Code](#)

[Send to Existing GenePattern Cell](#)



FastQC results

Input FASTQ files are flagged for quality issues. Quality measures can be good (), slightly abnormal (), or very unusual ().

genepattern.broadinstitute.org/gp/jobResults/974693/SRR1039508_1_fastqc/fastqc_report.html

FastQC Report

Mon 22 Sep 2014
SRR1039508_1.fastq.gz

Summary

-  Basic Statistics
-  Per base sequence quality
-  Per sequence quality scores
-  Per base sequence content
-  Per base GC content
-  Per sequence GC content
-  Per base N content
-  Sequence Length Distribution
-  Sequence Duplication Levels
-  Overrepresented sequences
-  Kmer Content

Basic Statistics

Measure	Value
Filename	SRR1039508_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	22935521
Filtered Sequences	0
Sequence length	63
%GC	50

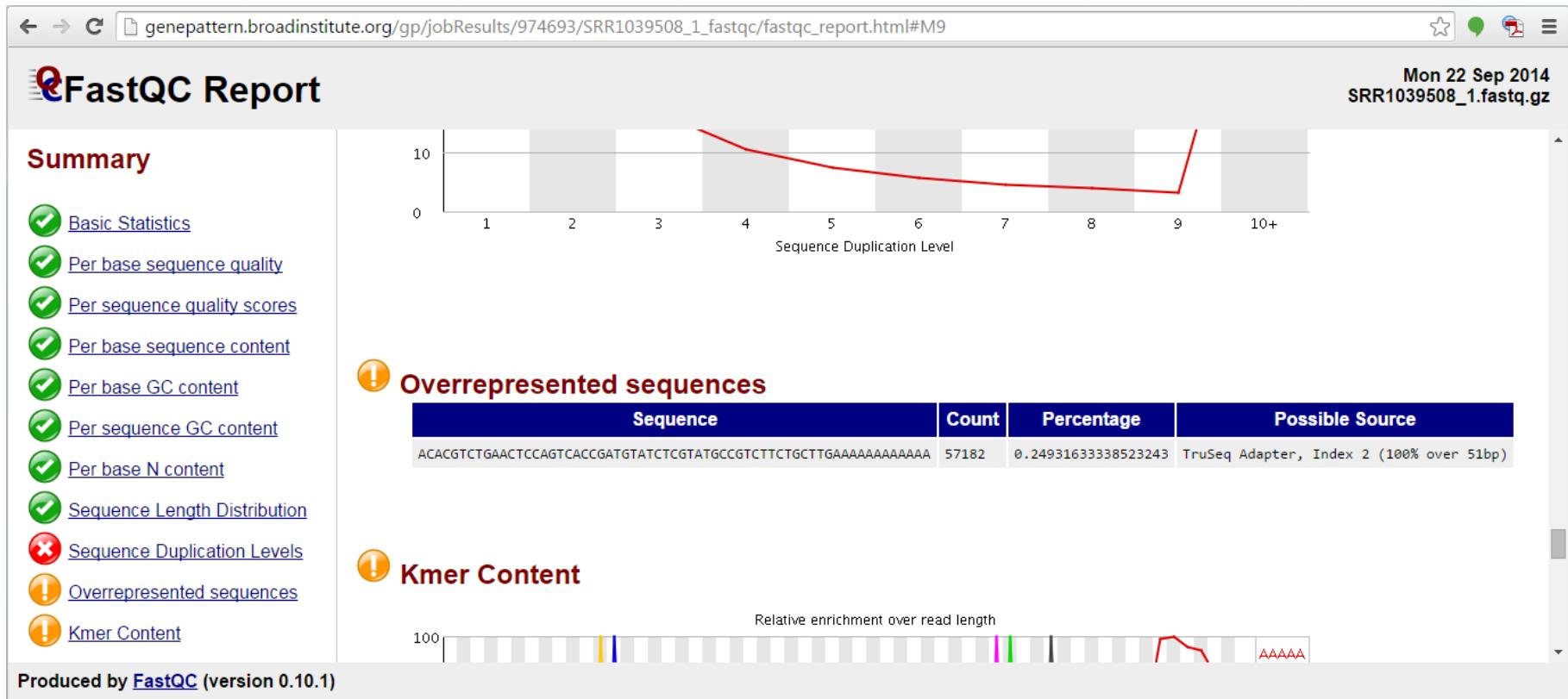
Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

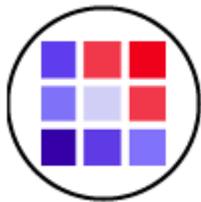
Produced by [FastQC](#) (version 0.10.1)

FastQC results

For example, our file has a small amount of adapter contamination, which is found under the ***Overrepresented Sequences*** section.

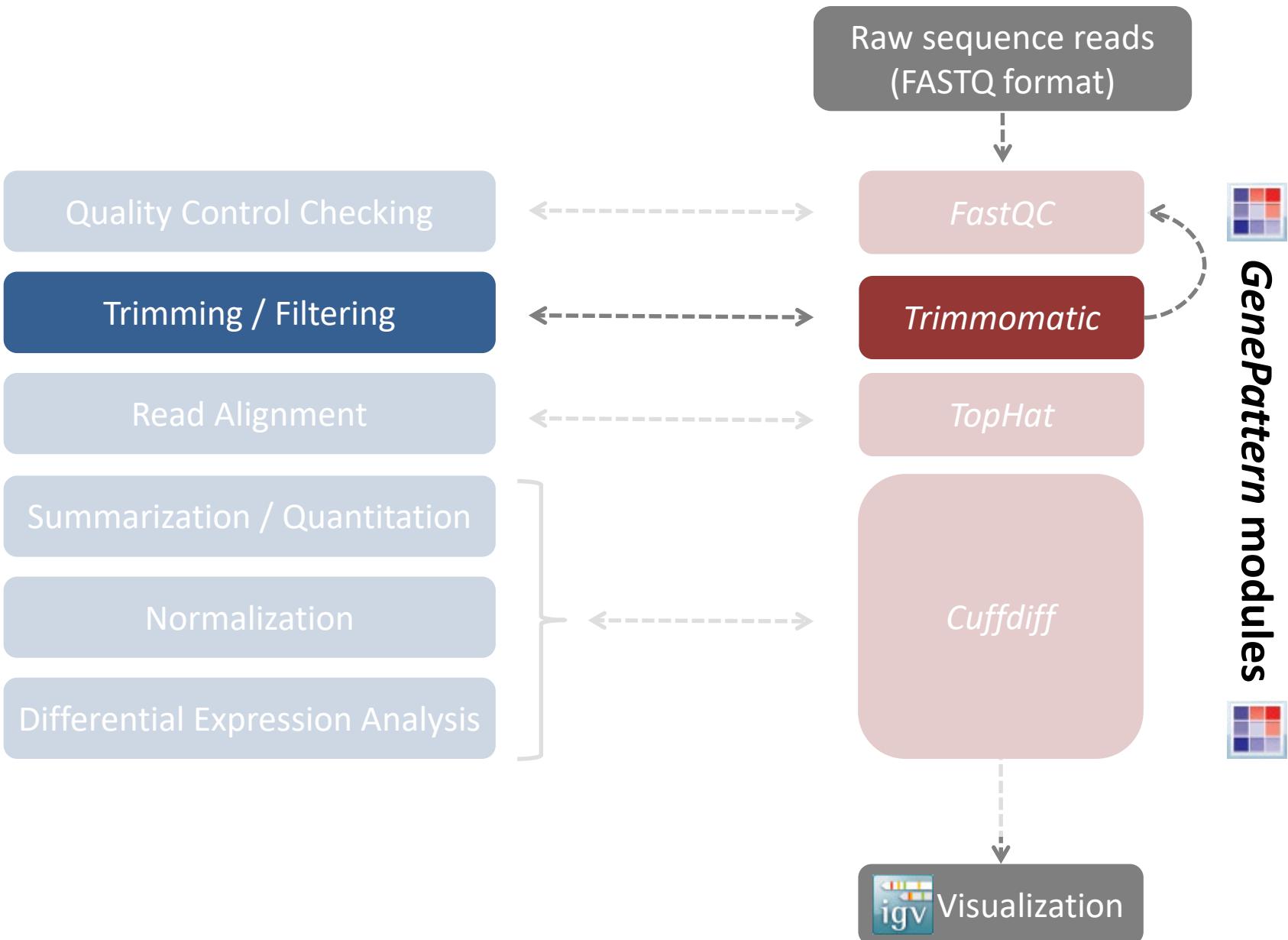


We can see that a TruSeq Adapter sequence is contaminating many of our reads.



RNA-Seq in GenePattern

RNA-Seq Differential Expression Analysis Workflow



Run *Trimmomatic* on paired FASTQ files

We use *Trimmomatic* to trim contaminating adapter sequences. We discard reads if they are below a specified length after trimming.



Output files:

- _1P or _2P** are forward and reverse reads which both passed quality checks ✓
- _1U or _2U** are “unpaired” reads in which one of the pair did not pass ✗

Run *Trimmomatic* on paired FASTQ files FA

Run Trimmomatic on Paired FASTQ files

We use Trimmomatic to trim contaminating adapter sequences. We discard reads if they are below a specified length after trimming.

- Drag [SRR1039508_1.fastq.gz](#) to the **input file 1** field.
- Drag [SRR1039508_2.fastq.gz](#) to the **input file 2** field.
- Collapse the **Basic Input Parameters and Options** section
- Expand the **Adapter Clipping** section
- In order to trim the TruSeq 2 adaptor we must either supply or choose an adaptor clip sequence file. Since TruSeq2 is common, the module provides this for us.
 - Click on the box containing **Add File or URL...**
 - Choose **TruSeq2-PE.fa**
- Set the **adapter clip seed mismatches** to the recommended value of **2**
- Set the **adapter clip palindrome clip threshold** to the recommended value of **40**
- Set the **adapter clip simple clip threshold** to **15** (from the recommended range)
- Allow Trimmomatic to set the adaptor clip min length to 8, which is the default.
- Set **adapter clip keep both reads** to **yes**, as is recommended
- Leave the rest of the parameters as default - starting with defaults is a good idea, if you don't know otherwise.
- Click **Run**

*More information about these parameters and why you might want to change them or add other trimming methods can be found in the Trimmomatic documentation.

Due to time constraints, "prebaked" output has been supplied for you below.

 GenePattern Trimmomatic Version 1.4 - ⚙️

Provides a variety of options for trimming Illumina FASTQ files of adapter sequences and low-quality reads. Run

Basic Input Parameters and Options -

Run *Trimmomatic* on paired FASTQ files FA

- Expand the **Adapter Clipping** section
- In order to trim the TruSeq 2 adaptor we must either supply or choose an adaptor clip sequence file. Since TruSeq2 is common, the module provides this for us.
 - Click on the box containing **Add File or URL...**
 - Choose **TruSeq2-PE.fa**
- Set the **adapter clip seed mismatches** to the recommended value of **2**
- Set the **adapter clip palindrome clip threshold** to the recommended value of **40**
- Set the **adapter clip simple clip threshold** to **15** (from the recommended range)
- Allow Trimmomatic to set the adaptor clip min length to 8, which is the default.
- Set **adapter clip keep both reads** to **yes**, as is recommended
- Leave the rest of the parameters as default - starting with defaults is a good idea, if you don't know otherwise.
- Click **Run**

**More information about these parameters and why you might want to change them or add other trimming methods can be found in the Trimmomatic documentation.*

Due to time constraints, "prebaked" output has been supplied for you below.

GenePattern Trimmomatic Version 1.4

Provides a variety of options for trimming Illumina FASTQ files of adapter sequences and low-quality reads.

Run

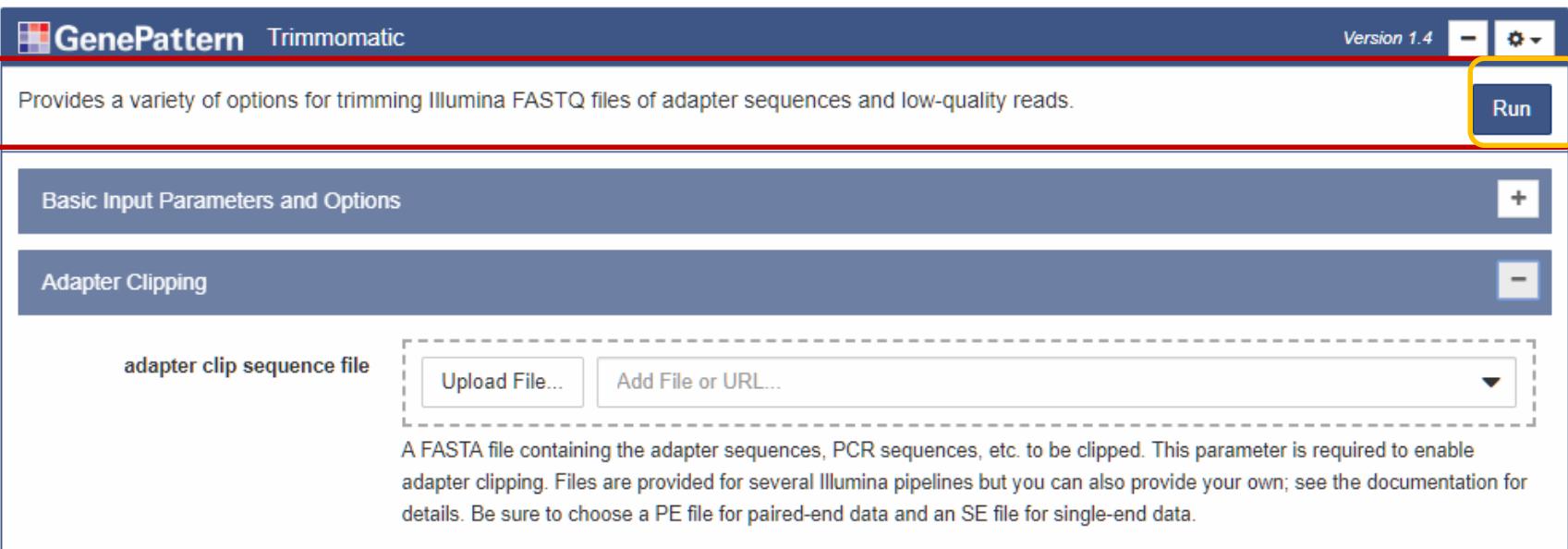
Basic Input Parameters and Options

Adapter Clipping

adapter clip sequence file

Upload File... Add File or URL...

A FASTA file containing the adapter sequences, PCR sequences, etc. to be clipped. This parameter is required to enable adapter clipping. Files are provided for several Illumina pipelines but you can also provide your own; see the documentation for details. Be sure to choose a PE file for paired-end data and an SE file for single-end data.



Trimmomatic results

GenePattern 1657089. Trimmomatic

Submitted by GPDemo on 2018-04-02T17:54:48-04:00

Completed

[cmdline.log](#) ⓘ

[SRR1039508_1P.fastq.gz](#) ⓘ

[SRR1039508_1U.fastq.gz](#) ⓘ

[SRR1039508_2P.fastq.gz](#) ⓘ

[SRR1039508_2U.fastq.gz](#) ⓘ

[stdout.txt](#) ⓘ

[gp_execution_log.txt](#) ⓘ

***Trimmomatic* on paired FASTQ files**

We use ***Trimmomatic*** to trim contaminating adapter sequences.
We discard reads if they are below a specified length after trimming.



Output files:

- _1P or _2P** are forward and reverse reads which both passed quality checks ✓
- _1U or _2U** are “unpaired” reads in which one of the pair did not pass ✗

FastQC re-run results

GenePattern 1657628. FastQC

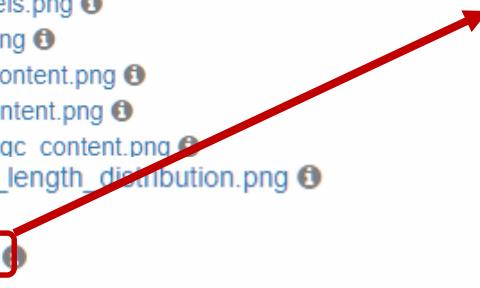
Submitted by GPDemo on 2018-04-03T10:07:49-04:00

Completed

SRR1039508_1P_fastqc.zip ⓘ
SRR1039508_1P_fastqc/licons/error.png ⓘ
SRR1039508_1P_fastqc/licons/fastqc_icon.png ⓘ
SRR1039508_1P_fastqc/licons/tick.png ⓘ
SRR1039508_1P_fastqc/licons/warning.png ⓘ
SRR1039508_1P_fastqc/Images/per_base_quality.png ⓘ
SRR1039508_1P_fastqc/Images/per_base_sequence_content.png ⓘ
SRR1039508_1P_fastqc/Images/per_sequence_quality.png ⓘ
SRR1039508_1P_fastqc/summary.txt ⓘ
SRR1039508_1P_fastqc/Images/duplication_levels.png ⓘ
SRR1039508_1P_fastqc/Images/kmer_profiles.png ⓘ
SRR1039508_1P_fastqc/Images/per_base_gc_content.png ⓘ
SRR1039508_1P_fastqc/Images/per_base_n_content.png ⓘ
SRR1039508_1P_fastqc/Images/per_sequence_ac_content.png ⓘ
SRR1039508_1P_fastqc/Images/sequence_length_distribution.png ⓘ
SRR1039508_1P_fastqc/fastqc_data.txt ⓘ
SRR1039508_1P_fastqc/fastqc_report.html ⓘ
gp_execution_log.txt ⓘ

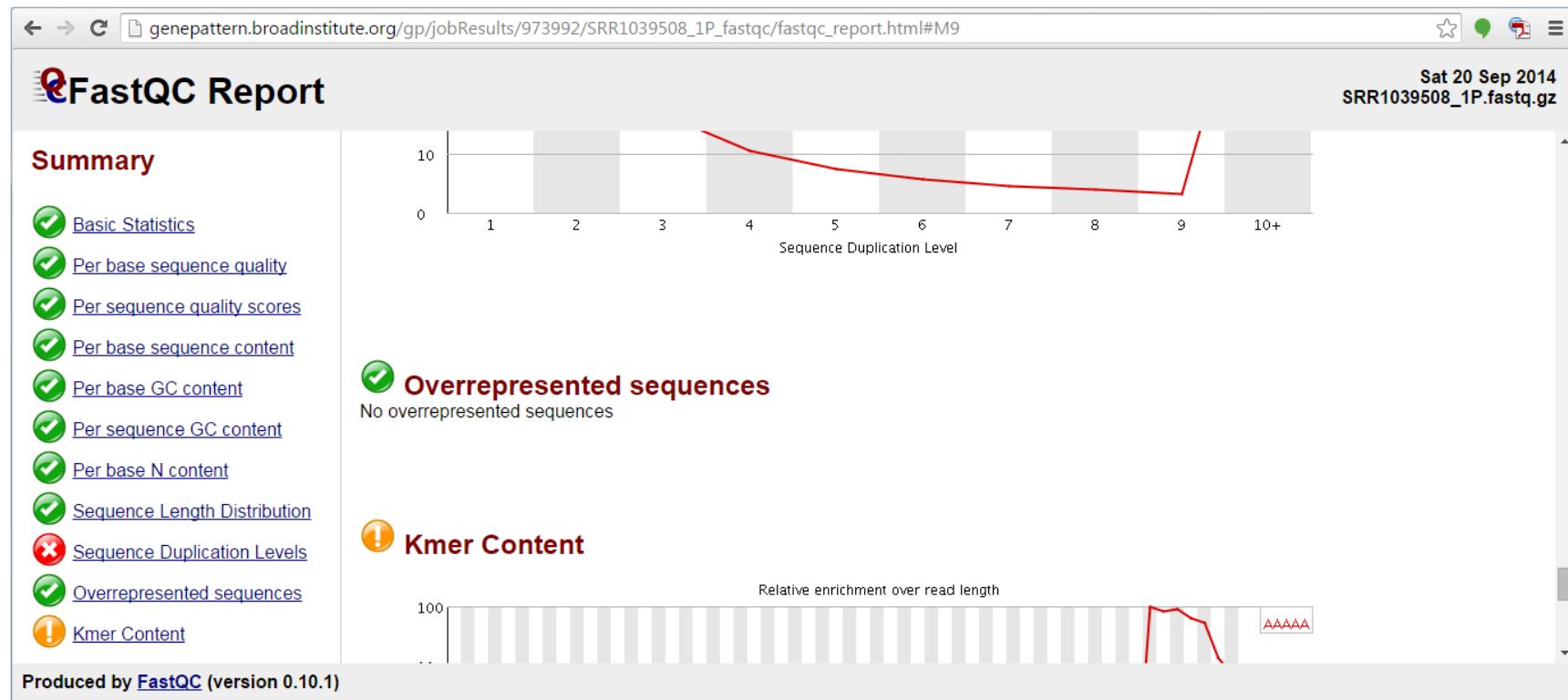
SRR1039508_1P_fastqc/fastqc_report.html

Download File
Open in New Tab
Send to Code
Send to Existing GenePattern Cell

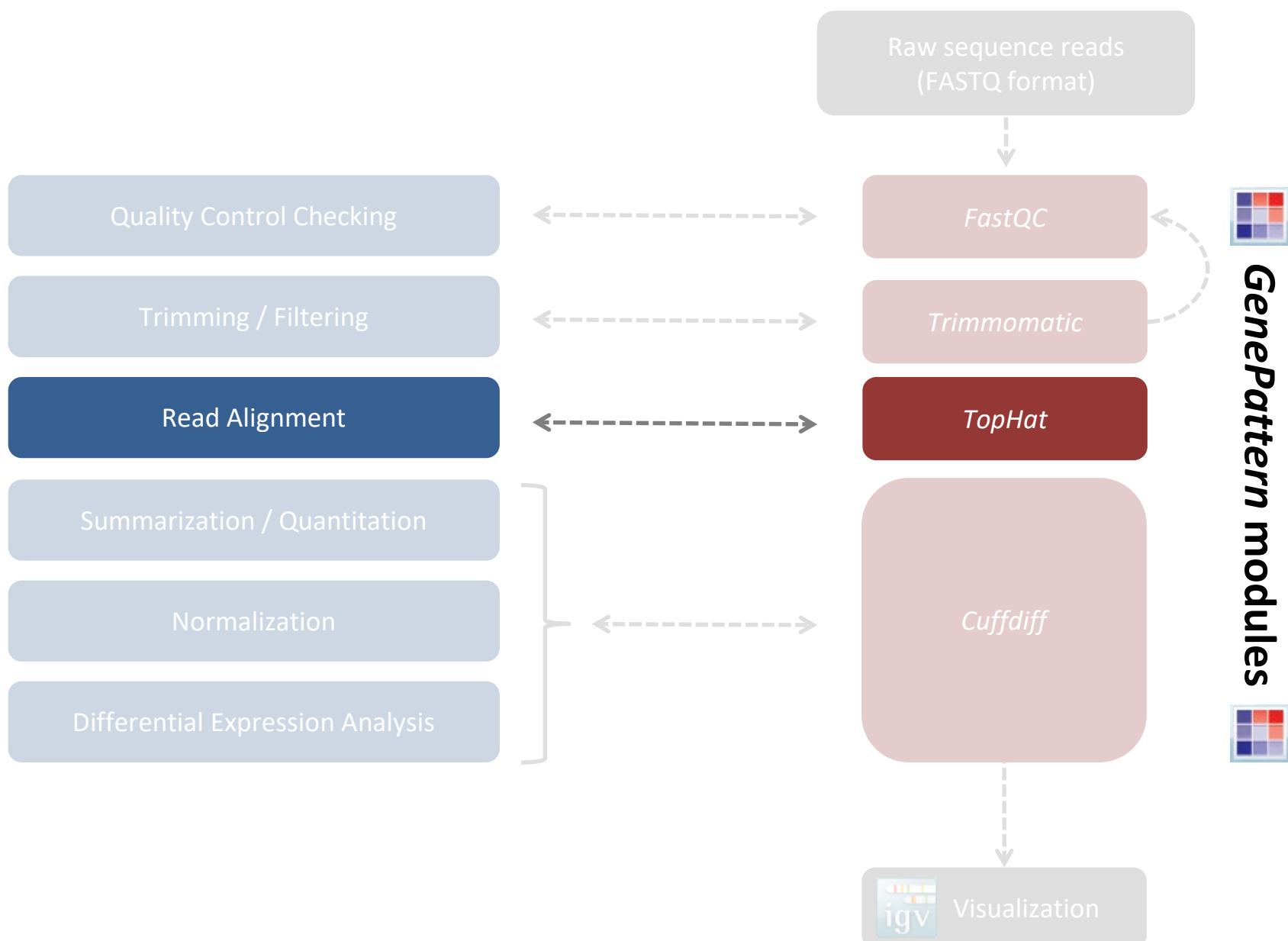


Re-run *FastQC* to see improvement

If we re-run *FastQC* after trimming adapter sequences using *Trimmomatic*, we can see that the quality of the reads has improved.



RNA-Seq Differential Expression Analysis Workflow



RNA-Seq Differential Expression Analysis Workflow

Alignment & Downstream Analyses

How many reads successfully aligned?

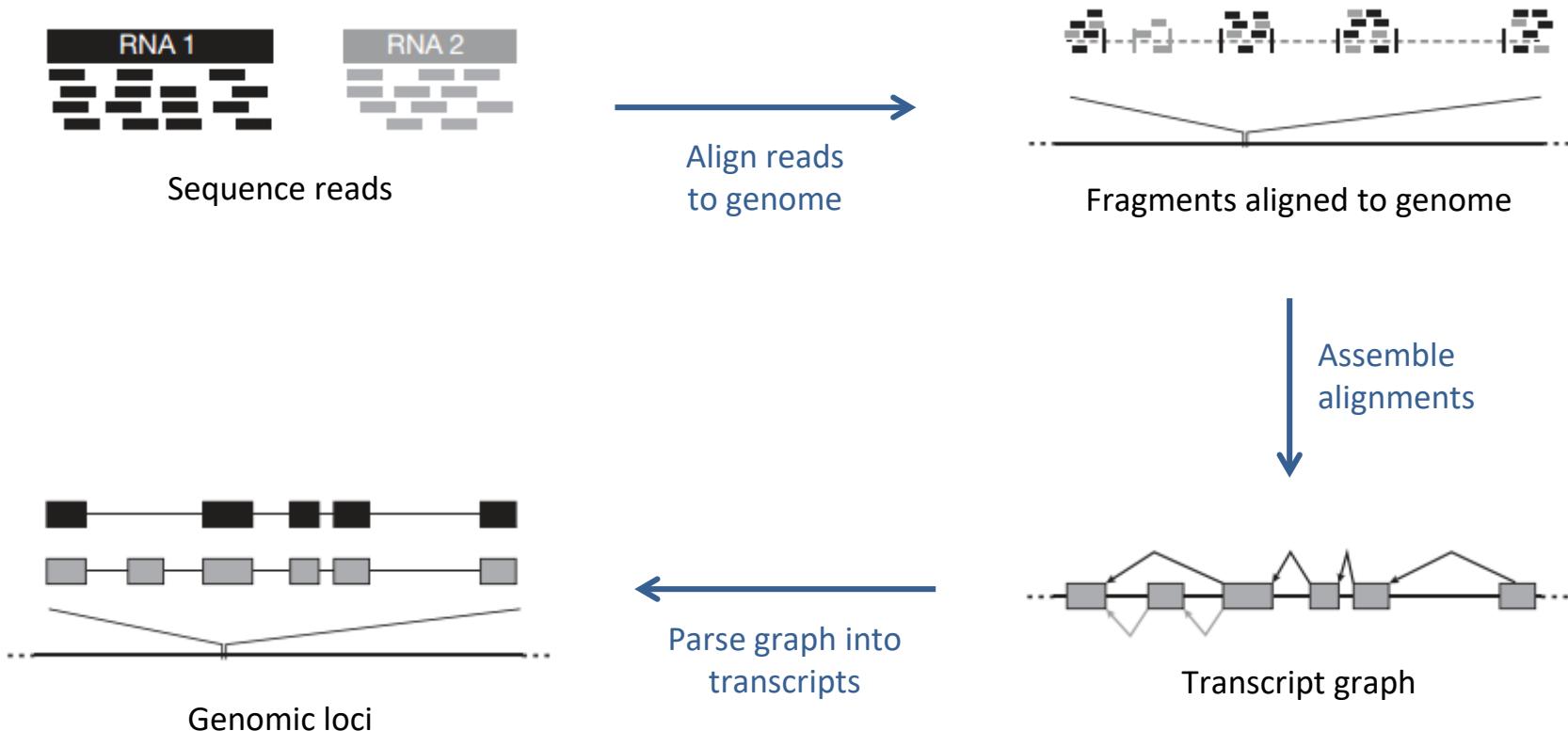
Where did the reads align?

Are there differences in the number of reads for each phenotype?

Can I visualize the differences in phenotypes?

Run *TopHat* to align reads

TopHat is an alignment tool which can be used to align sequence reads to a genome or transcriptome.



Run *TopHat* to align reads

Aligning to the transcriptome with *TopHat* requires either a genome annotation file (**GTF file**), or a **transcriptome index**. We will use the **UCSC hg19** reference annotation.

Step 1. Create a transcriptome index.

To speed up alignment, we create a **transcriptome index** by running the UCSC hg19 reference annotation on its own.

Step 2. Align reads to the transcriptome.

We align the reads to the transcriptome using the transcriptome index created in **Step 1**. We run *TopHat* for each sample (8 times).

This is the least computationally intensive mode to run *TopHat*; however, **alignment will take >1 hour per sample**.

Run *TopHat* to create a transcriptome index

Create Transcriptome Index

Here we just need to run tophat with a bowtie index and a gtf file, to create a transcriptome index that we can use in subsequent alignments.

- Click on the file input for [GTF file](#)
- Select [Homo_sapiens_hg19_UCSC.gtf](#)
- Click on the file input for [bowtie index](#)
 - Scroll to the [FTP Server Files](#) section of the list
 - Select [Homo_sapiens_hg19_UCSC](#)

Due to time constraints, "prebaked" output has been supplied for you below.

GenePattern TopHat Version 8.11 - ⚙️

[**Beta Release** Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] TopHat 2.0.11 is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Run

Basic Parameters -

bowtie index* -
Upload File... Add File or URL...
A zip file or directory containing a Bowtie 2 index.

GTF file -
Upload File... Add File or URL...
A GTF file (v. 2.2 or higher) or GFF3 file containing a list of gene model annotations. TopHat will first extract the transcript sequences and align them to this virtual transcriptome first. Only the reads that do not fully map to the transcriptome will then be mapped on the genome. The reads that did map on the transcriptome will be converted to genomic mappings (spliced as needed) and merged with the novel mappings.

Run *TopHat* to create a transcriptome index

FA

The screenshot shows the GenePattern web interface for job #105854. The top bar indicates the job is completed, submitted by GPDemo on 2019-04-03T16:08:55+00:00. Below this, a list of files generated by the TopHat run is displayed, each with a download icon (blue square with white arrow).

- 105854_tmp/Homo_sapiens_UCSC_hg19.fa
- cmdline.log
- transcriptome_index/Homo_sapiens_hg19_UCSC.1.bt2
- transcriptome_index/Homo_sapiens_hg19_UCSC.4.bt2
- transcriptome_index/Homo_sapiens_hg19_UCSC.3.bt2
- transcriptome_index/Homo_sapiens_hg19_UCSC.fa
- transcriptome_index/Homo_sapiens_hg19_UCSC.fa.tlst
- transcriptome_index/Homo_sapiens_hg19_UCSC.2.bt2
- transcriptome_index/Homo_sapiens_hg19_UCSC.gff
- transcriptome_index/Homo_sapiens_hg19_UCSC.rev.1.bt2
- transcriptome_index/Homo_sapiens_hg19_UCSC.rev.2.bt2
- transcriptome_index/Homo_sapiens_hg19_UCSC.ver
- stdout.txt
- gp_execution_log.txt

Index = the whole transcriptome_index folder:

https://genepattern.broadinstitute.org/gp/jobResults/105854/transcriptome_index

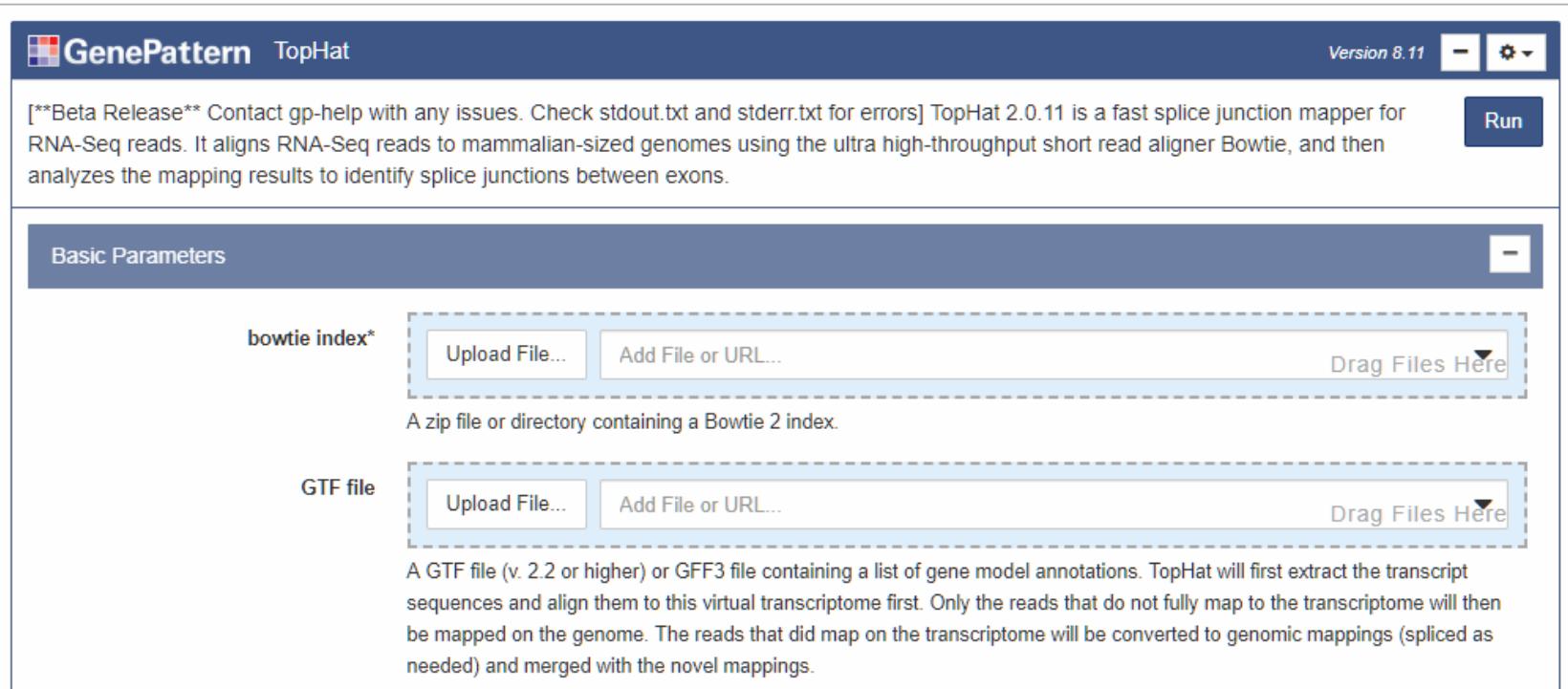
Run *TopHat* to align reads

Align Reads to Transcriptome

This time, we are aligning reads to the transcriptome.

- Choose the same bowtie index as before (`Homo_sapiens_hg19_UCSC`)
- Drag the link to the folder of our transcriptome index, https://cloud.genepattern.org/gp/jobResults/105854/transcriptome_index/, to the **transcriptome index** parameter.
- Click the input field for **reads pair 1** and select `SRR1039508_1P.fastq.gz` from the Trimmomatic output.
- Click the input field for **reads pair 2** and select `SRR1039508_2P.fastq.gz` from the Trimmomatic output.
- Set the **library type** to **Standard Illumina (fr-unstranded)**
- Set **transcriptome only** to **yes**

Due to time constraints, "prebaked" output has been supplied for you below.



The screenshot shows the GenePattern TopHat interface. At the top, it says "GenePattern TopHat Version 8.11". Below that is a descriptive text block about TopHat 2.0.11. The main area is titled "Basic Parameters". It has two sections: "bowtie index*" and "GTF file". Each section has an "Upload File..." button, an "Add File or URL..." button, and a "Drag Files Here" input field. Below each section is a detailed description of the required file types.

Beta Release Contact gp-help with any issues. Check stdout.txt and stderr.txt for errors] TopHat 2.0.11 is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Run

Basic Parameters

bowtie index* Upload File... Add File or URL... Drag Files Here
A zip file or directory containing a Bowtie 2 index.

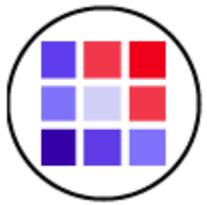
GTF file Upload File... Add File or URL... Drag Files Here
A GTF file (v. 2.2 or higher) or GFF3 file containing a list of gene model annotations. TopHat will first extract the transcript sequences and align them to this virtual transcriptome first. Only the reads that do not fully map to the transcriptome will then be mapped on the genome. The reads that did map on the transcriptome will be converted to genomic mappings (spliced as needed) and merged with the novel mappings.

TopHat results

The important output is the ***.accepted_hits.bam** file. Remember that BAM files are not human readable (we will not be able to view it).

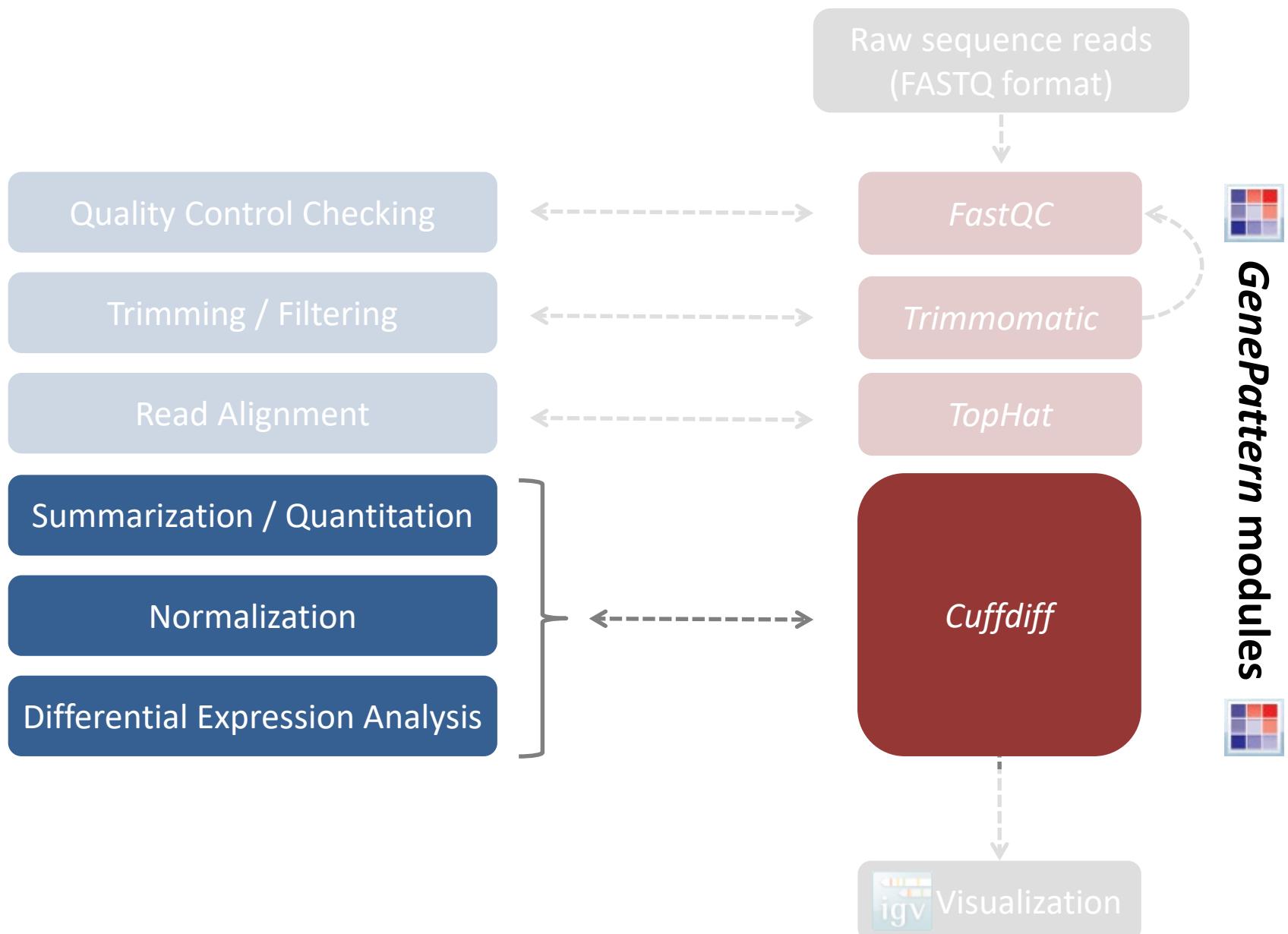
The screenshot shows the GenePattern software interface with the title "GenePattern 1658233. TopHat". The status bar indicates "Completed". The main window lists several output files from the TopHat analysis, including:

- cmdline.log
- 1658233_SRR1039508.prep_reads.info
- 1658233_SRR1039508.align_summary.txt
- 1658233_SRR1039508.deletions.bed
- 1658233_SRR1039508.insertions.bed
- 1658233_SRR1039508.junctions.bed
- 1658233_SRR1039508.accepted_hits.bam** (this file is highlighted with a red box)
- 1658233_SRR1039508.unmapped.bam
- stdout.txt
- gp_execution_log.txt



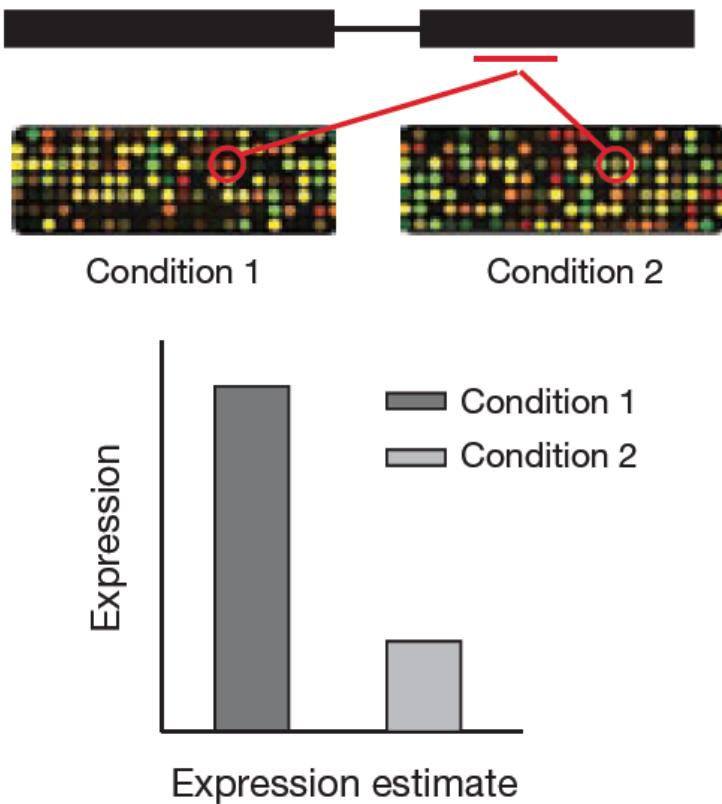
RNA-Seq Differential Expression Analysis

RNA-Seq Differential Expression Analysis Workflow

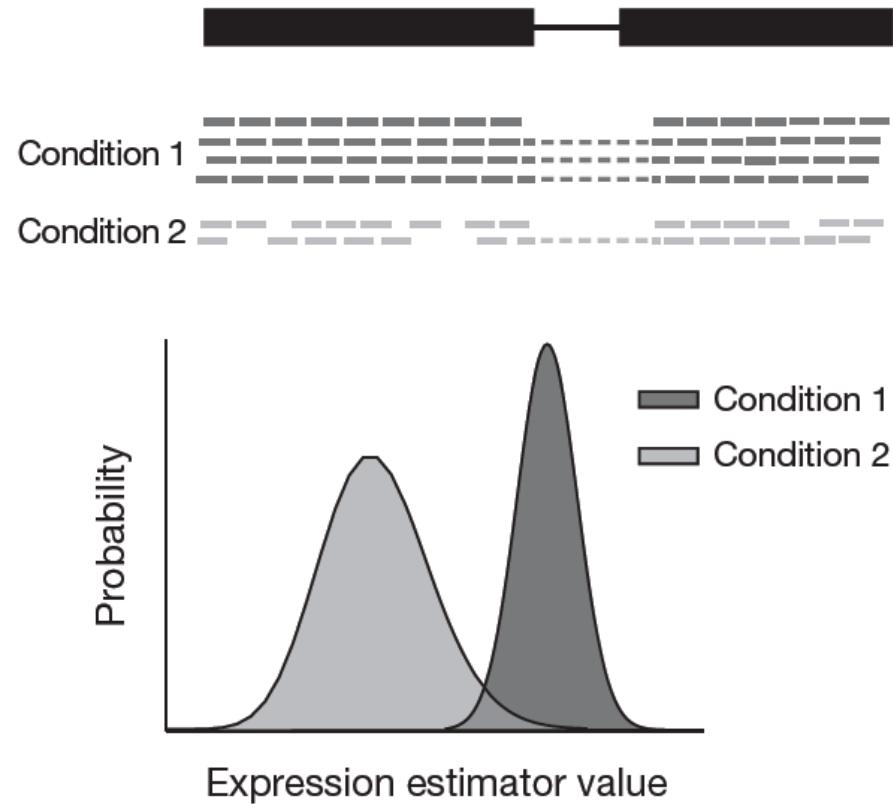


Differential gene expression analysis

Microarray



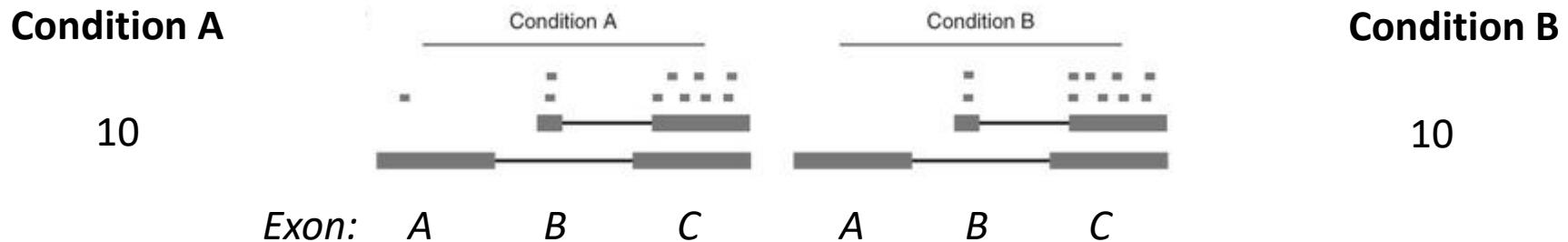
RNA-Seq



Differential gene expression analysis

Summarization/quantitation accurately derives expression values by assigning reads to the most likely transcription template.

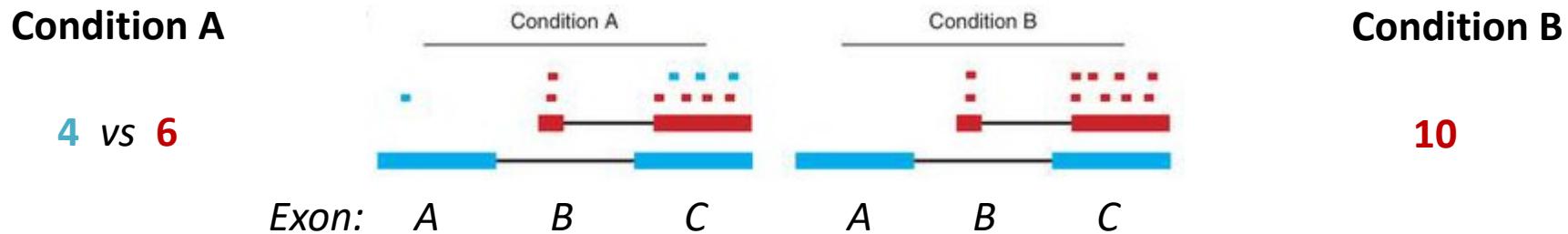
Raw count: map reads to exons and divide by a scaling factor based on the length of the exons



Differential gene expression analysis

Summarization/quantitation accurately derives expression values by assigning reads to the most likely transcription template.

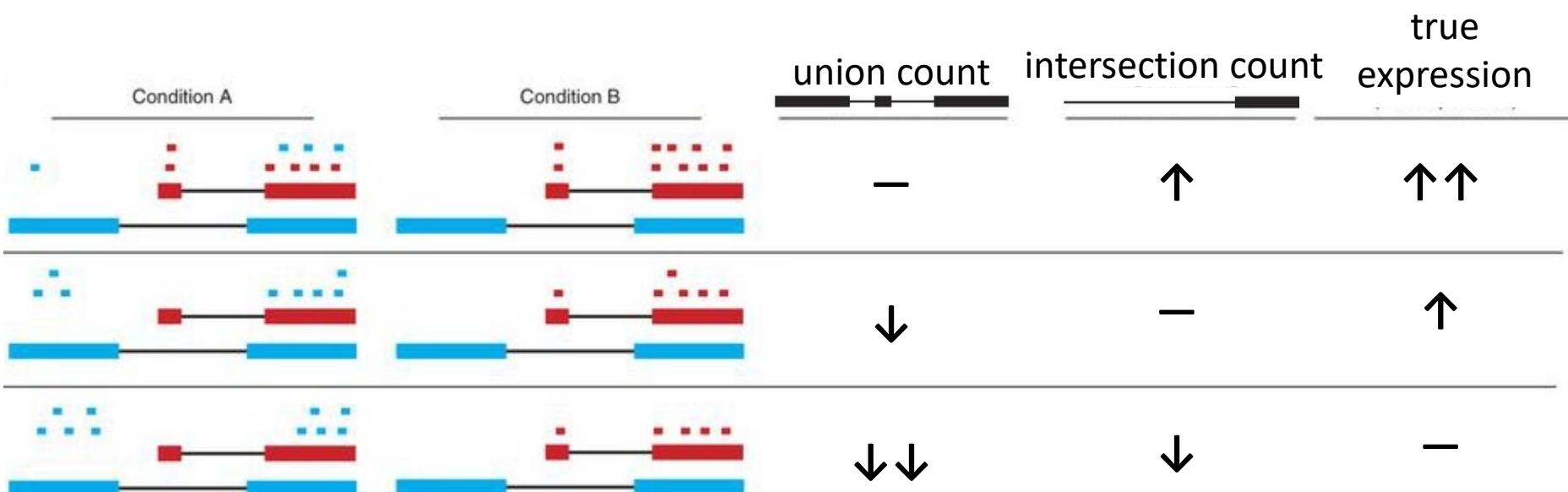
Raw count: map reads to exons and divide by a scaling factor based on the length of the exons



Differential gene expression analysis

Summarization/quantitation accurately derives expression values by assigning reads to the most likely transcription template.

Raw count: map reads to exons and divide by a scaling factor based on the length of the exons

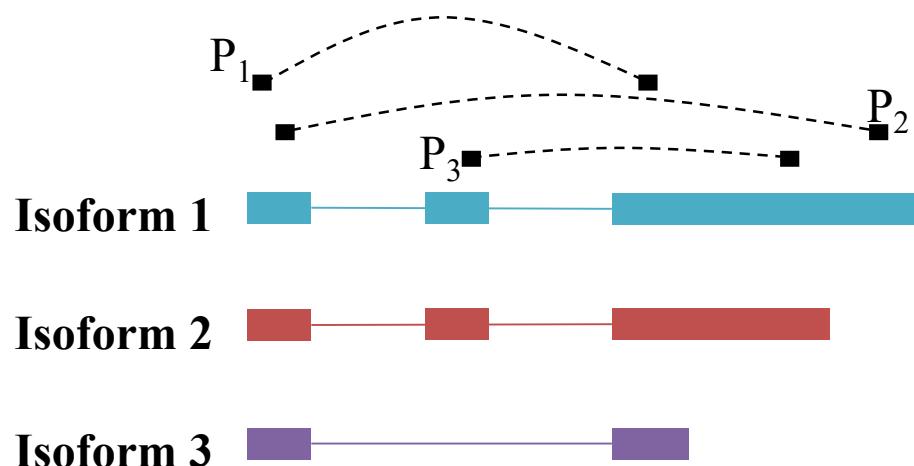


Differential gene expression analysis

Summarization/quantitation accurately derives expression values by assigning reads to the most likely transcription template.

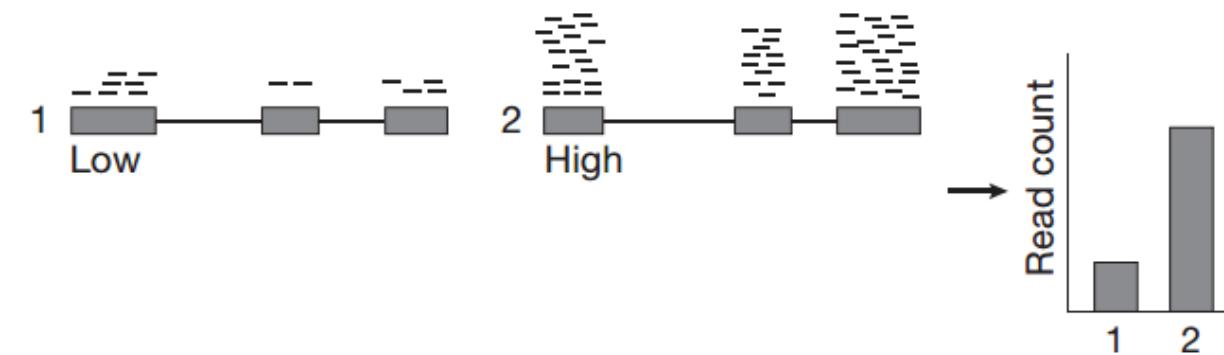
Isoform deconvolution: estimate the expression levels of a gene's alternative isoforms (transcripts) and then sum those transcript-level estimates to derive gene-level expression estimates.

Paired ends increase isoform deconvolution confidence.

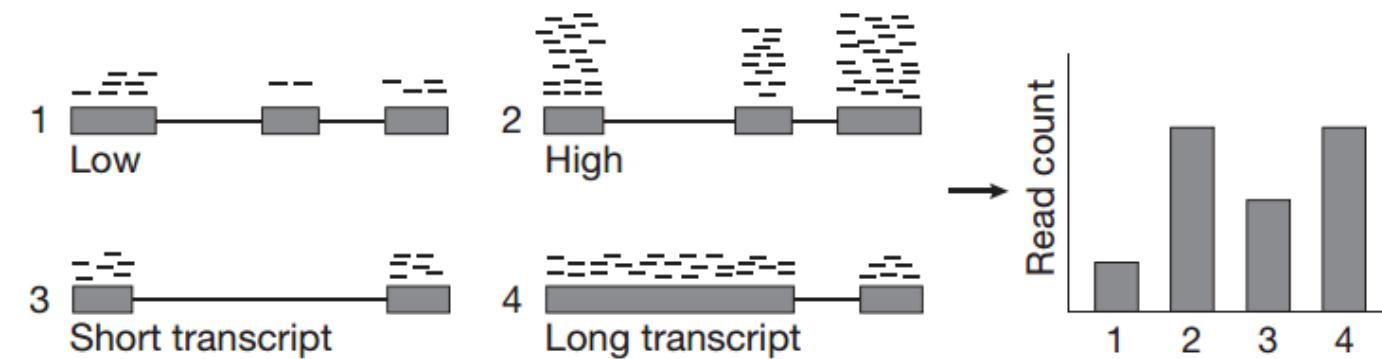


P₁ is from I₁, I₂ or I₃.
P₂ is from I₁.
P₃ is from I₁ or I₂.

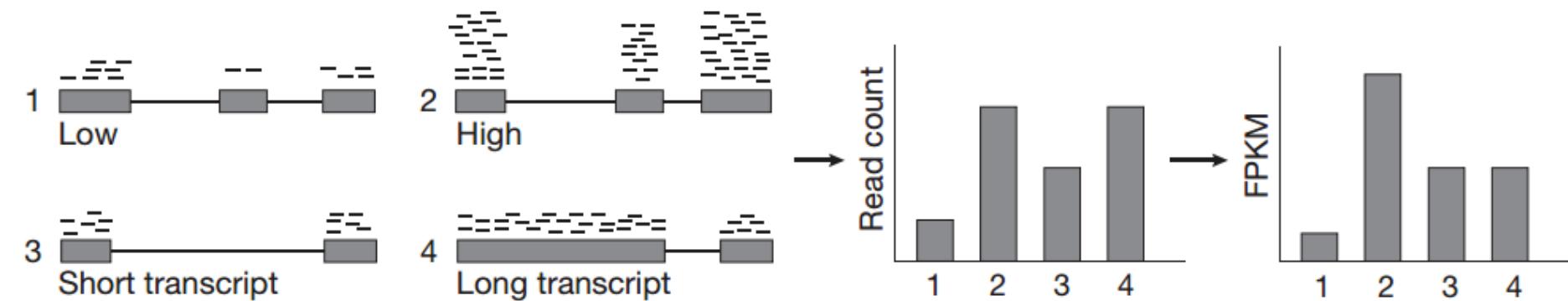
How to quantify expression from reads



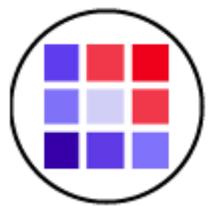
How to quantify expression from reads



How to quantify expression from reads



Fragments Per Kilobase of transcript per Million mapped reads (FPKM), a count of mapped fragments normalized to transcript length.



Exercise

Run *Cuffdiff* to differentiate phenotypes

Cuffdiff summarizes, normalizes, and quantitates aligned read data.
It can also conduct differential expression analysis.

First, we group aligned reads from *TopHat* condition
(untreated vs. dex).

Then, we identify genes which are differentially expressed by
comparing the two conditions.

Run *Cuffdiff* to differentiate phenotypes

FA

Search for “Cuffdiff” under the Modules panel in GenePattern.

The screenshot shows the GenePattern software interface. At the top, there is a navigation bar with links for Modules & Pipelines, Suites, Job Results, Resources, Help, and GenomeSpace. A user profile icon for "RecipeUser" is also present. Below the navigation bar, the main window has tabs for Modules, Jobs, Files, and GenomeSpace. The "Modules" tab is selected. In the search bar, the text "cuffdiff" is typed. The right side of the screen displays the "Cuffdiff" module configuration panel. The panel title is "Cuffdiff version 7". It includes a brief description: "Cufflinks 2.0.2 - Finds significant changes in transcript expression, splicing, and promoter use". There is a note indicating "* required field". On the right side of the panel are "Reset" and "Run" buttons. The main configuration area is titled "Basic Input Parameters and Options" and contains the following fields:

- aligned files***: A section for uploading aligned files. It includes a "Condition" input field, "Upload Files..." and "Add Paths or URLs..." buttons, and a note about a 2GB file upload limit. It also includes an "Add Another Condition" button and a note about maximum group limits.
- GTF file***: A section for selecting a transcript annotation file. It includes "Select a file" and "Upload your own file" buttons, and a "Batch" checkbox.
- frag bias correct**: A section for selecting a reference file for bias detection. It includes "Select a file" and "Upload your own file" buttons, and a "Batch" checkbox.
- time series***: A dropdown menu set to "no". A note below it states: "Analyze the provided samples as a time series, rather than testing for differences between all pairs of samples".

Run *Cuffdiff* to differentiate phenotypes

FA

Define the first condition (e.g. ‘untreated’), then click [Add Path or URL...](#)

Then, navigate to

shared_data>gpTutorial_files>2019_BU_MolBio>untreated,

and choose the *.accepted_hits.bam files of that same condition (e.g. *SRR...08, SRR...12, SRR...16* and *SRR...20*).

The screenshot shows the Cuffdiff web application interface. At the top, it displays 'Cuffdiff version 7' and 'Documentation'. Below this, a message states 'Cufflinks 2.0.2 - Finds significant changes in transcript expression, splicing, and promoter use'. A note indicates '* required field'. On the right, there are 'Reset' and 'Run' buttons.

The main area is titled 'Basic Input Parameters and Options'. It contains a section for 'aligned files*' where a condition is set to 'untreated'. There are buttons for 'Upload Files...', 'Add Paths or URLs...', and a large 'Drag Files Here' area. A note specifies a 2GB file upload limit. Below this, a list of four selected files is shown:

- http://genepattern.broadinstitute.org/gp/jobResults/973995/SRR1039508.accepted_hits.bam
- http://genepattern.broadinstitute.org/gp/jobResults/973997/SRR1039512.accepted_hits.bam
- http://genepattern.broadinstitute.org/gp/jobResults/973999/SRR1039516.accepted_hits.bam
- http://genepattern.broadinstitute.org/gp/jobResults/974001/SRR1039520.accepted_hits.bam

At the bottom, there is a button for 'Add Another Condition' and a note stating '(Maximum Groups Allowed=Unlimited)'. A footer message says 'A set of aligned files grouped by condition'.

Run *Cuffdiff* to differentiate phenotypes

FA

Click **Add Another Condition**. Define the second condition (e.g. 'dex'),
then click **Add Path or URL...**.

Go to **shared_data>gpTutorial_files>2019_BU_MolBio>untreated** and
choose the ***.accepted_hits.bam** files of that same condition
(e.g. **SRR...09, SRR...13, SRR...17** and **SRR...21**).

Basic Input Parameters and Options

These parameters are essential for the analysis.

Condition: **untreated**

Upload Files... **Add Paths or URLs...** **Drag Files Here**

2GB file upload limit using the Upload Files... button. For files > 2GB upload from the Files tab.

aligned files*

Hide Files...(Selected 4 files)

http://genepattern.broadinstitute.org/gp/jobResults/973995/SRR1039508.accepted_hits.bam **X**
http://genepattern.broadinstitute.org/gp/jobResults/973997/SRR1039512.accepted_hits.bam **X**
http://genepattern.broadinstitute.org/gp/jobResults/973999/SRR1039516.accepted_hits.bam **X**
http://genepattern.broadinstitute.org/gp/jobResults/974001/SRR1039520.accepted_hits.bam **X**

Add Another Condition

(Maximum Groups Allowed=Unlimited)

A set of aligned files grouped by condition

Run *Cuffdiff* to differentiate phenotypes

FA

Set the **GTF file** (`Homo_sapiens_hg19_UCSC.gtf`), the **frag bias correct** (`Homo_sapiens_hg19_UCSC.fa`) and the **library type** (`fr-unstranded`) parameters. To run the job you would click  .

GTF file* or Batch 
`Homo_sapiens_hg19_UCSC.gtf` 

A transcript annotation file in GTF/GFF format produced by cufflinks, cuffcompare, cuffmerge, or other source (such as a reference annotation GTF)

frag bias correct or Batch 
`Homo_sapiens_hg19_UCSC.fa` 

Reference (FASTA/FA) for bias detection and correction algorithm

library type `fr-unstranded`  Batch 

The library type used to generate reads. The default is inferred, meaning that no library type information is passed.

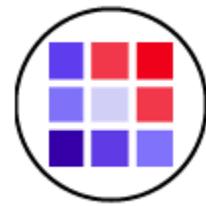
Step 6: Results

Once we determine how the expression of genes/transcripts differs between phenotypes (**untreated** vs. **dex**), what next?

Aligned reads and gene expression results can be viewed with ***IGV***.



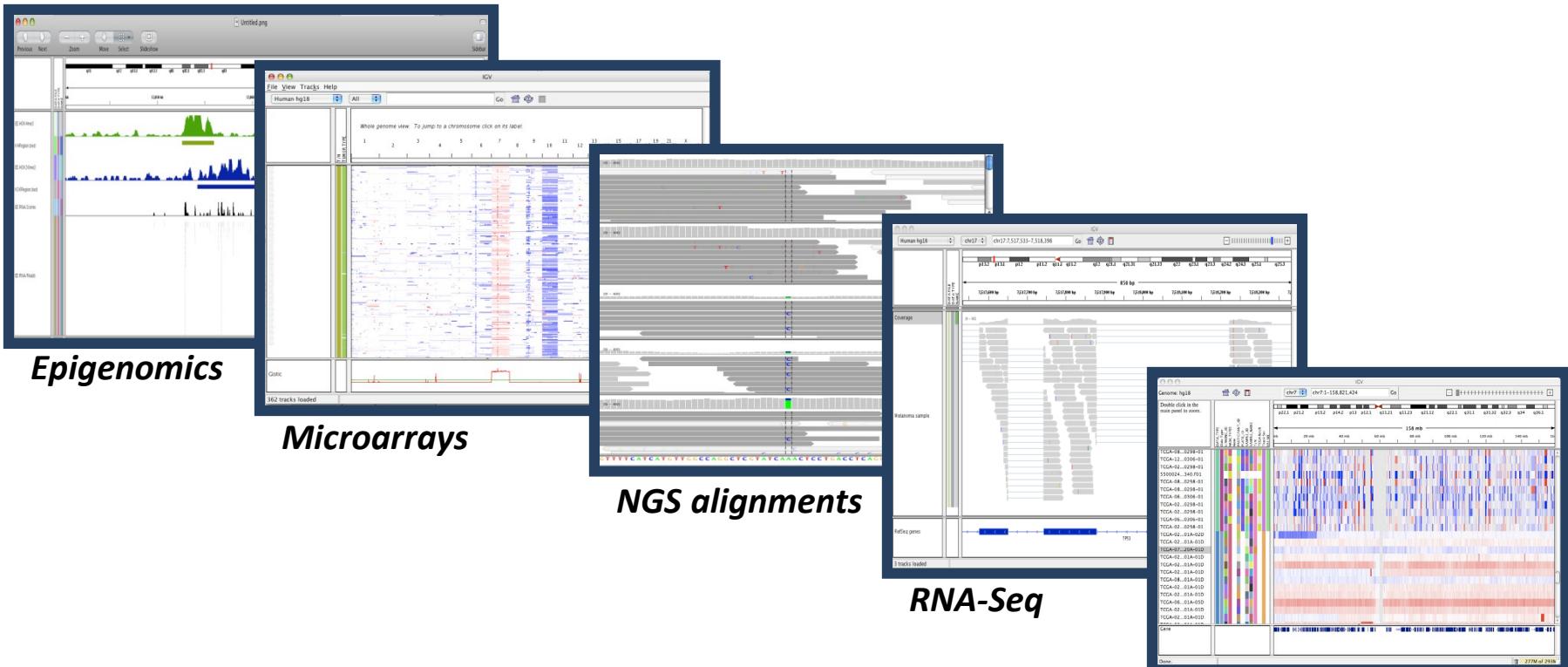
Before we continue we'll take a slight detour to learn how to use ***IGV***.



Break

Integrative Genomics Viewer (IGV)

A desktop application for the interactive visual exploration of integrated genomic datasets

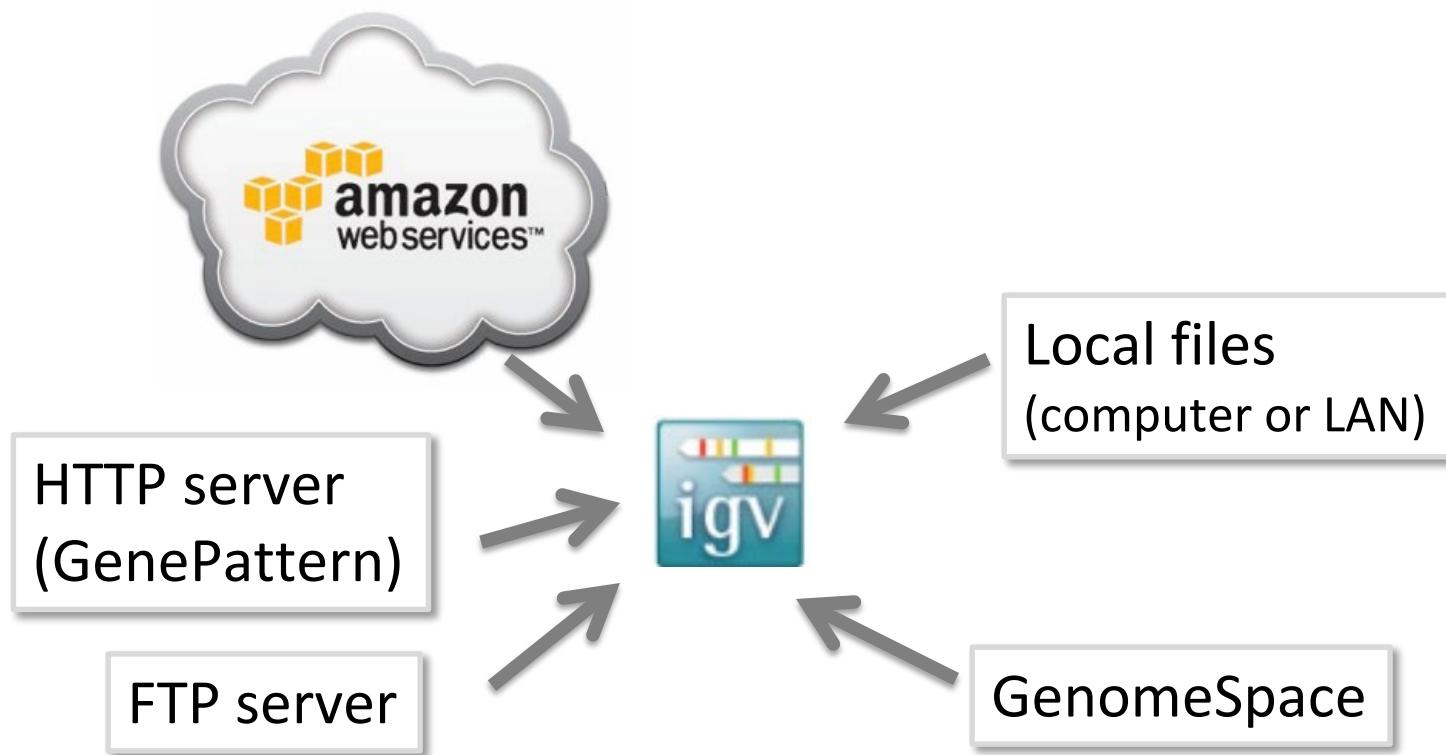


IGV Features

With IGV you can...

- explore large genomic datasets with an intuitive, easy-to-use interface.
- integrate multiple data types with clinical and other sample information.
- view data from multiple sources: local, remote, and “cloud-based”.

IGV Data Sources: How do I load data?



- View **local** files without uploading.
- View **remote** files without downloading the whole dataset.

How to use IGV

The screenshot shows the IGV website with several sections highlighted by red boxes:

- Desktop application**: A button for the Mac application.
- Downloads**: A link in the sidebar menu.
- Mac**: A section for Mac users with a "Download Mac App" button.
- Java Web Start (All Platforms)**: A section for Java Web Start with four "Launch" buttons for different memory configurations: 750 MB, 1.2 GB, 2 GB, and 10 GB.

Note: IGV 2.3.x requires Java 7. Users with Java 6 (JRE 1.6) should first try to upgrade Java to the latest version. If this is not possible you will need to run a 2.2.x version available in the archive.

Next...

- Select a reference genome
- Load data
- Navigate through the data

Viewing next generation sequencing (NGS) data in IGV

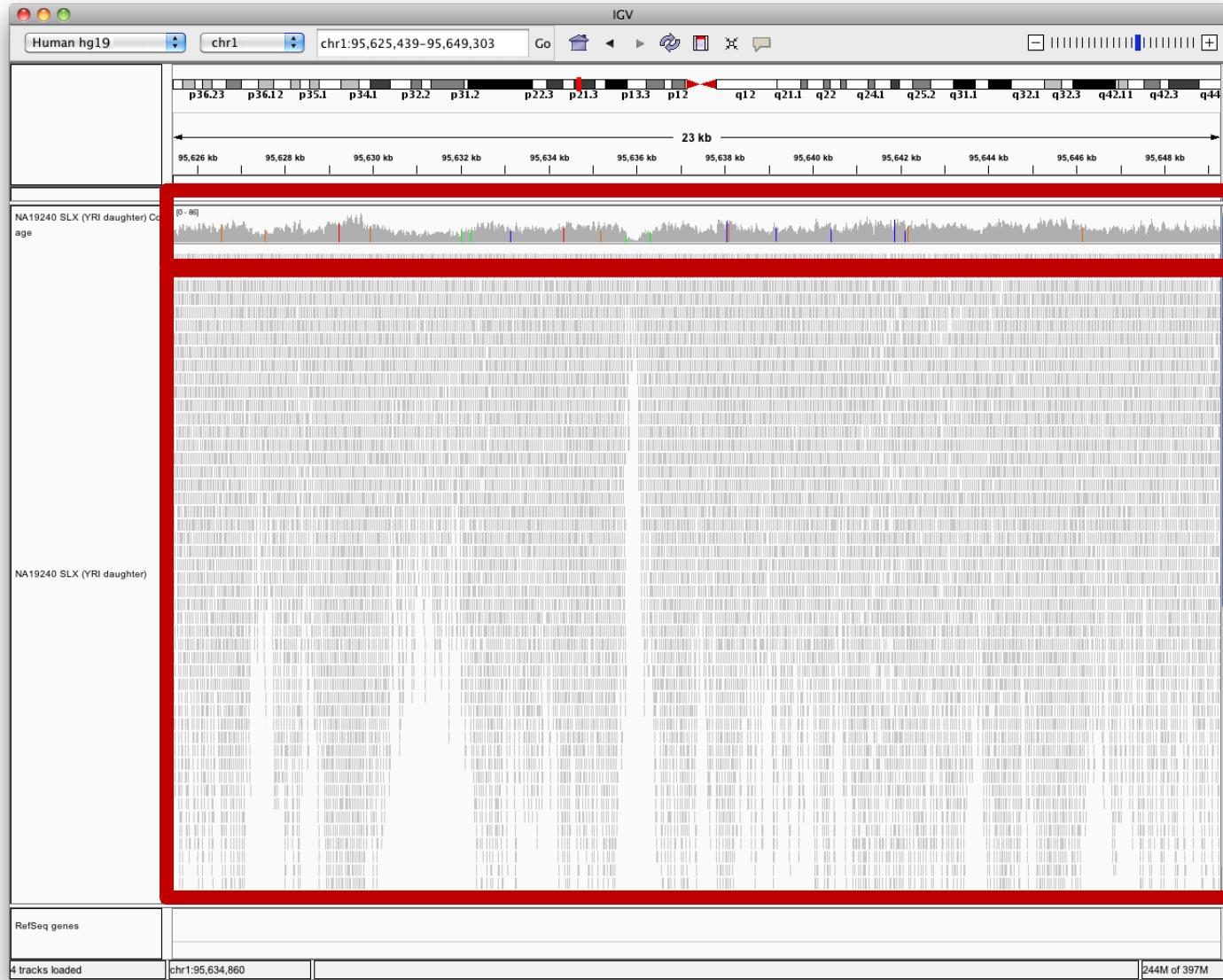
Viewing alignments

Whole chromosome view.



Viewing alignments

Zoom in to view alignments.

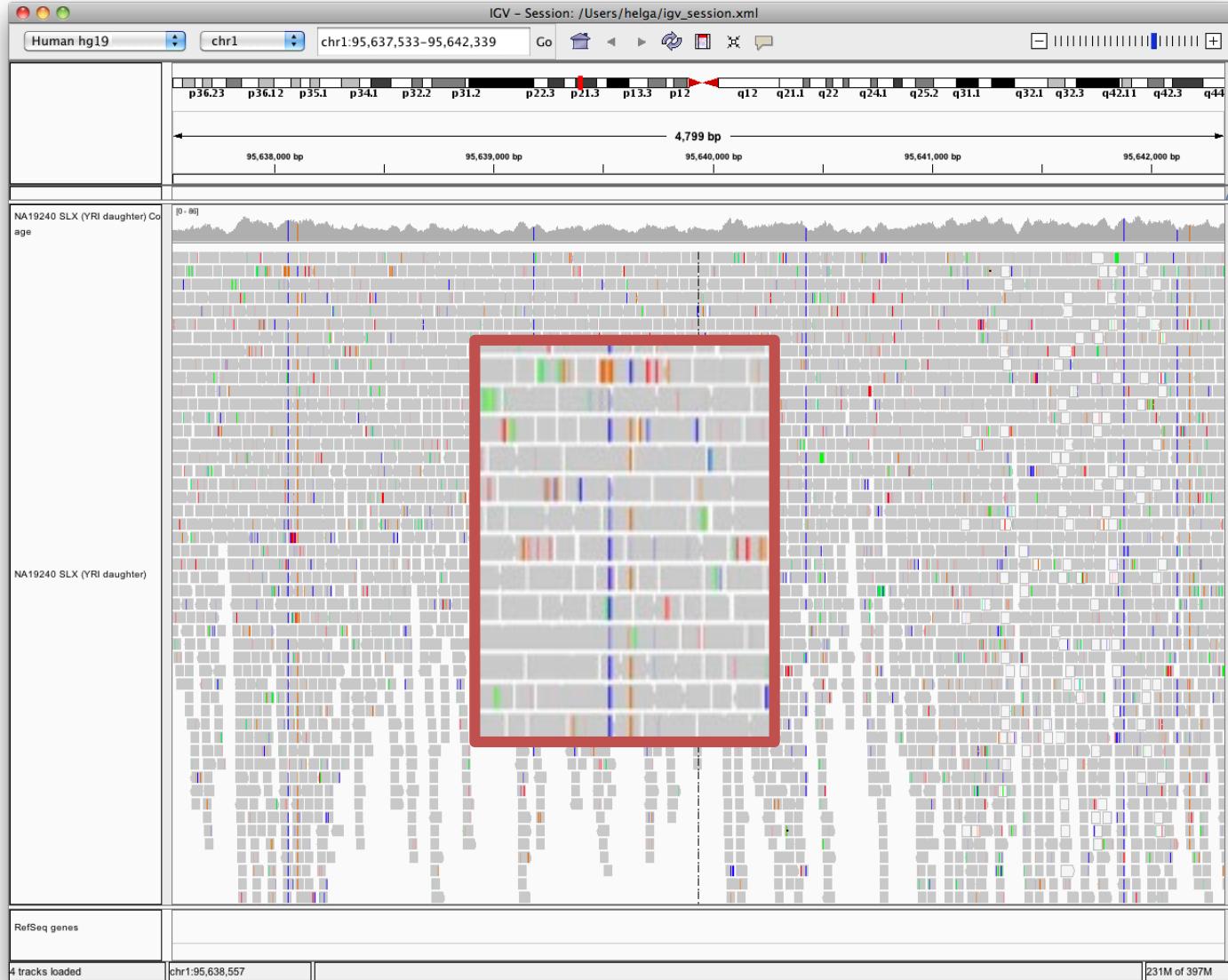


Coverage

Aligned
reads

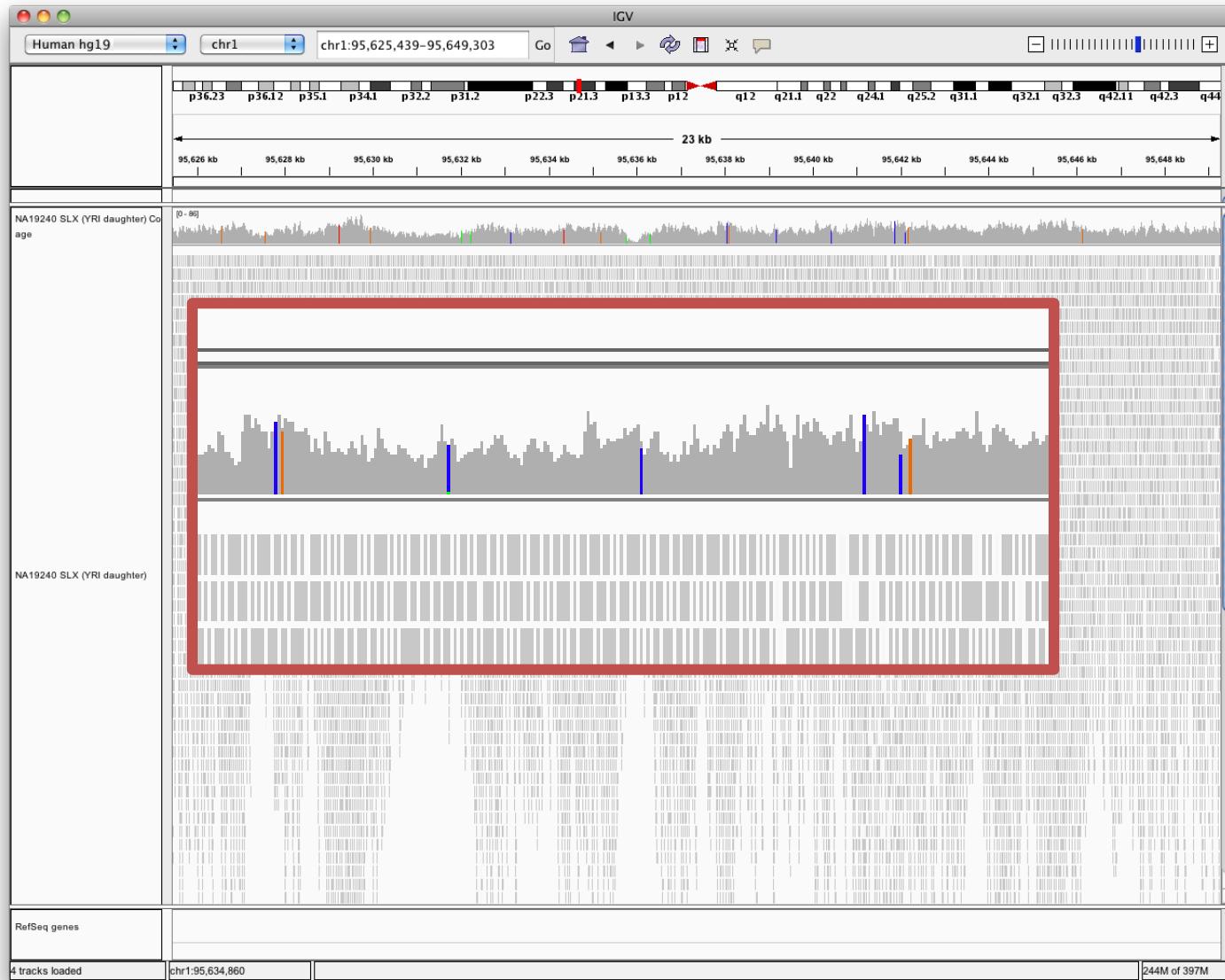
Viewing alignments

Bases that do not match the reference sequence are highlighted by color.



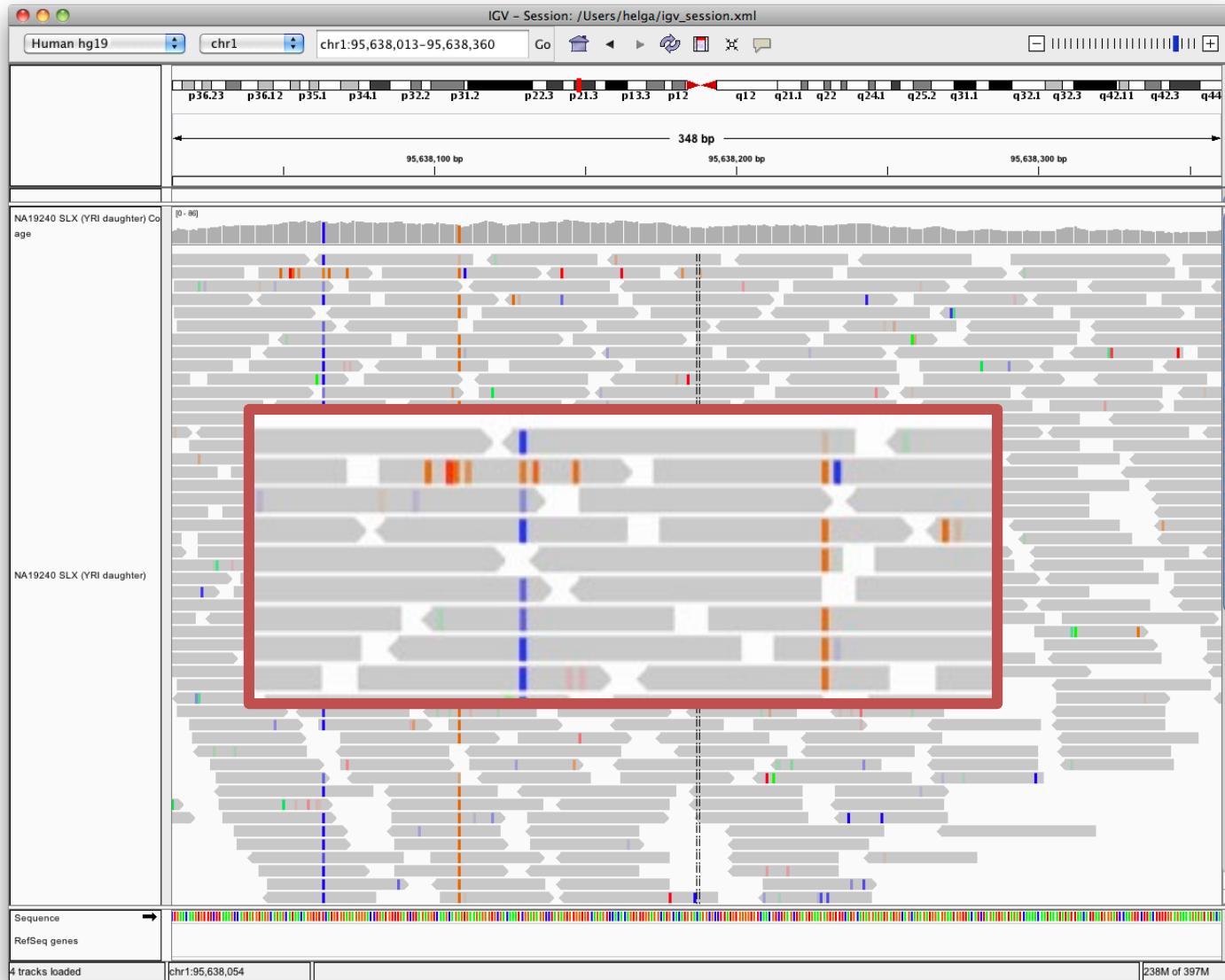
Viewing alignments

Coverage track indicates loci with large number of mismatches.



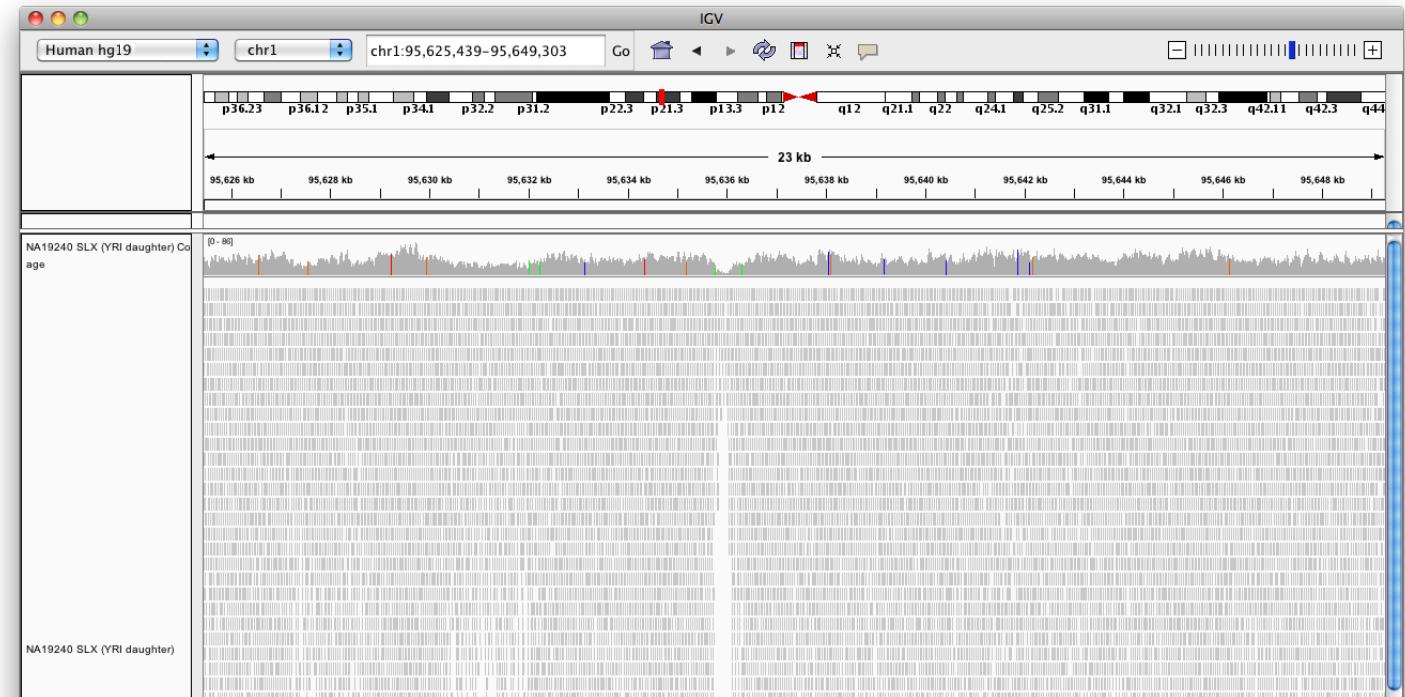
Viewing alignments

Zoom in to see more detail.



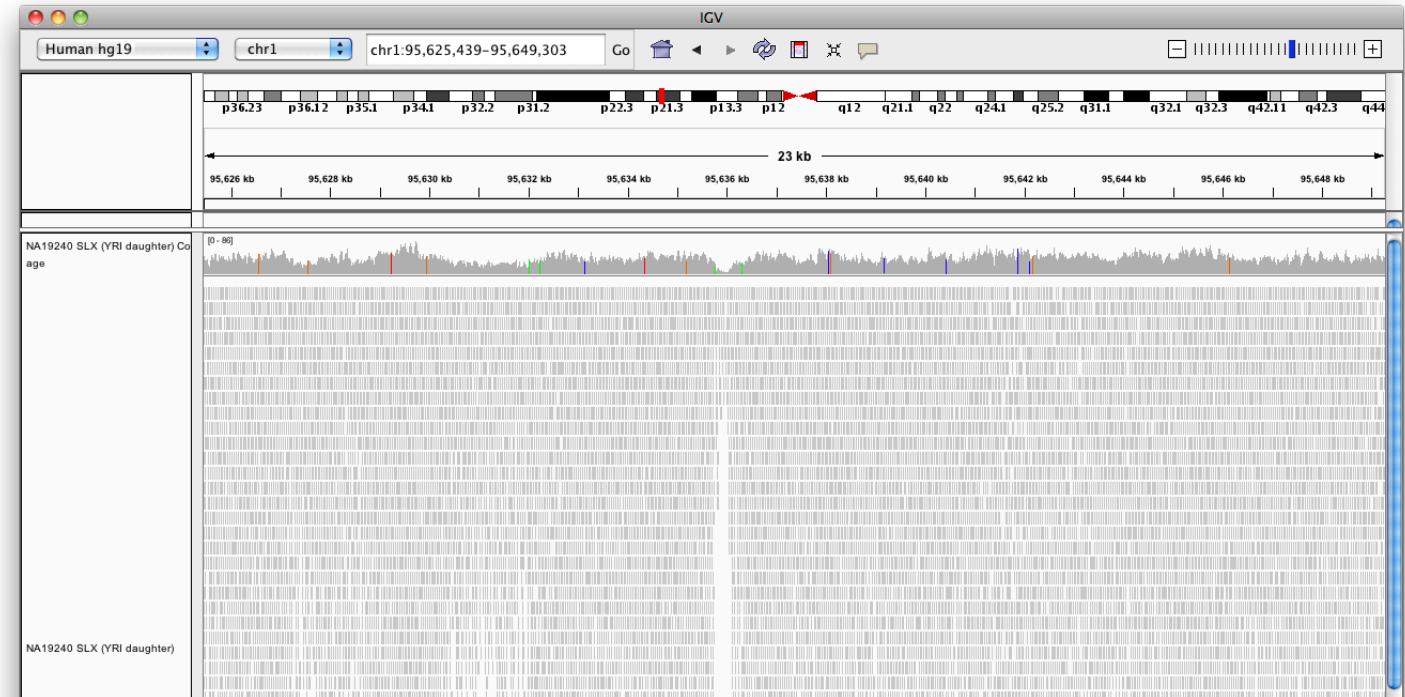
Viewing alignments

How far do you need to zoom in to see the alignments?



Viewing alignments

How far do you need to zoom in to see the alignments? **30kb.**

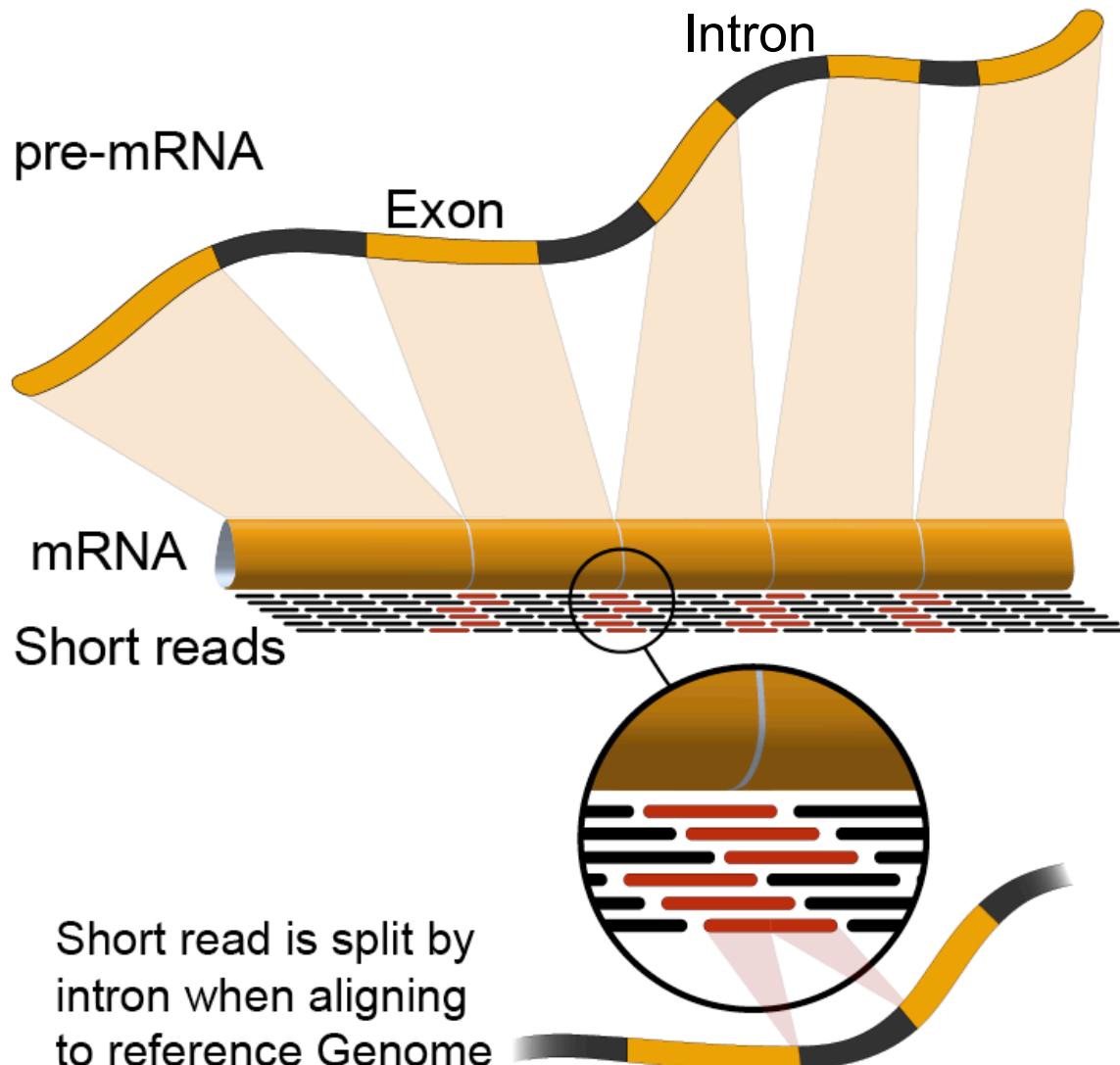


Or you can set another threshold in **View > Preferences > Alignments.**

- Higher values require more memory
- Low coverage files can use higher values
- Very deep coverage files should use lower value

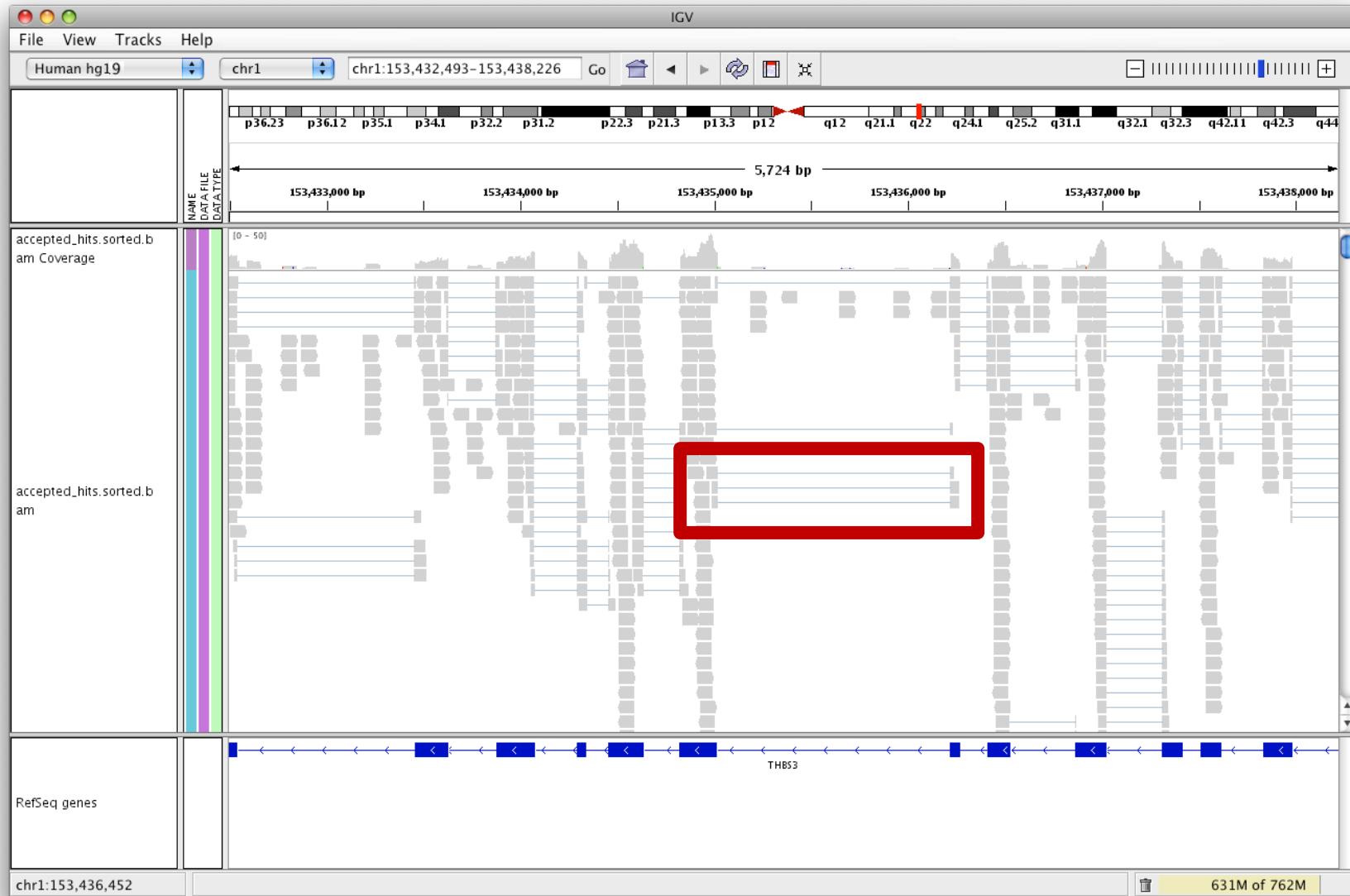
Viewing RNA-seq data in IGV

RNA-seq alignments



RNA-seq alignments

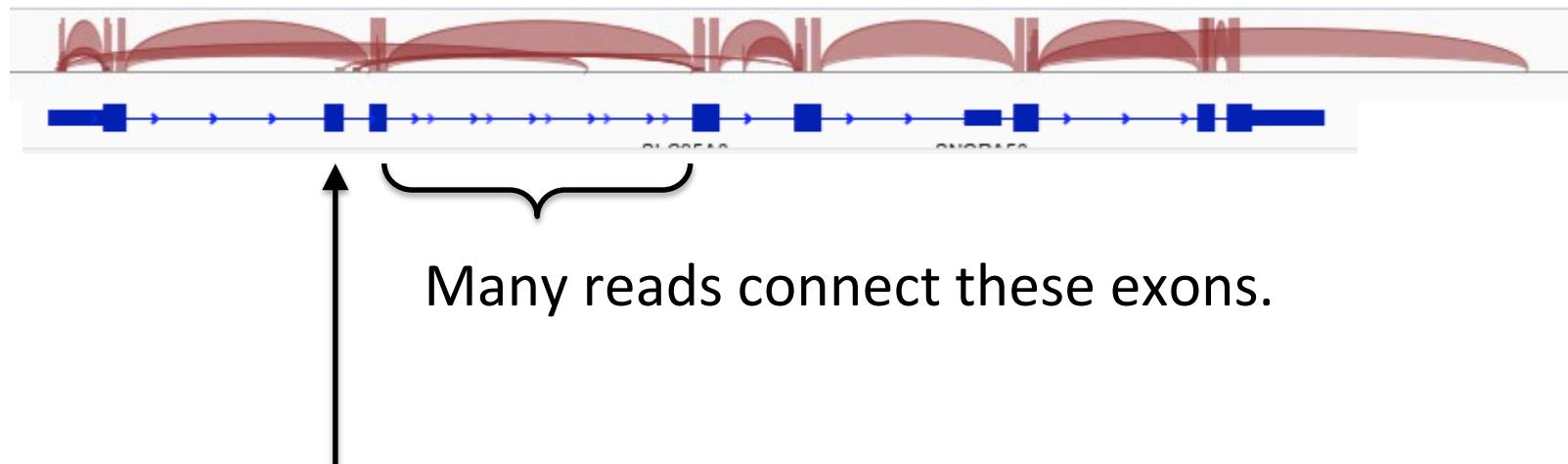
RNA-seq reads can span splice junctions.



RNA-seq alignments

You can display a separate splice junction track which shows how many reads span exon junctions.

Arcs represent reads that span exon junctions. Height is proportional to the number of reads.

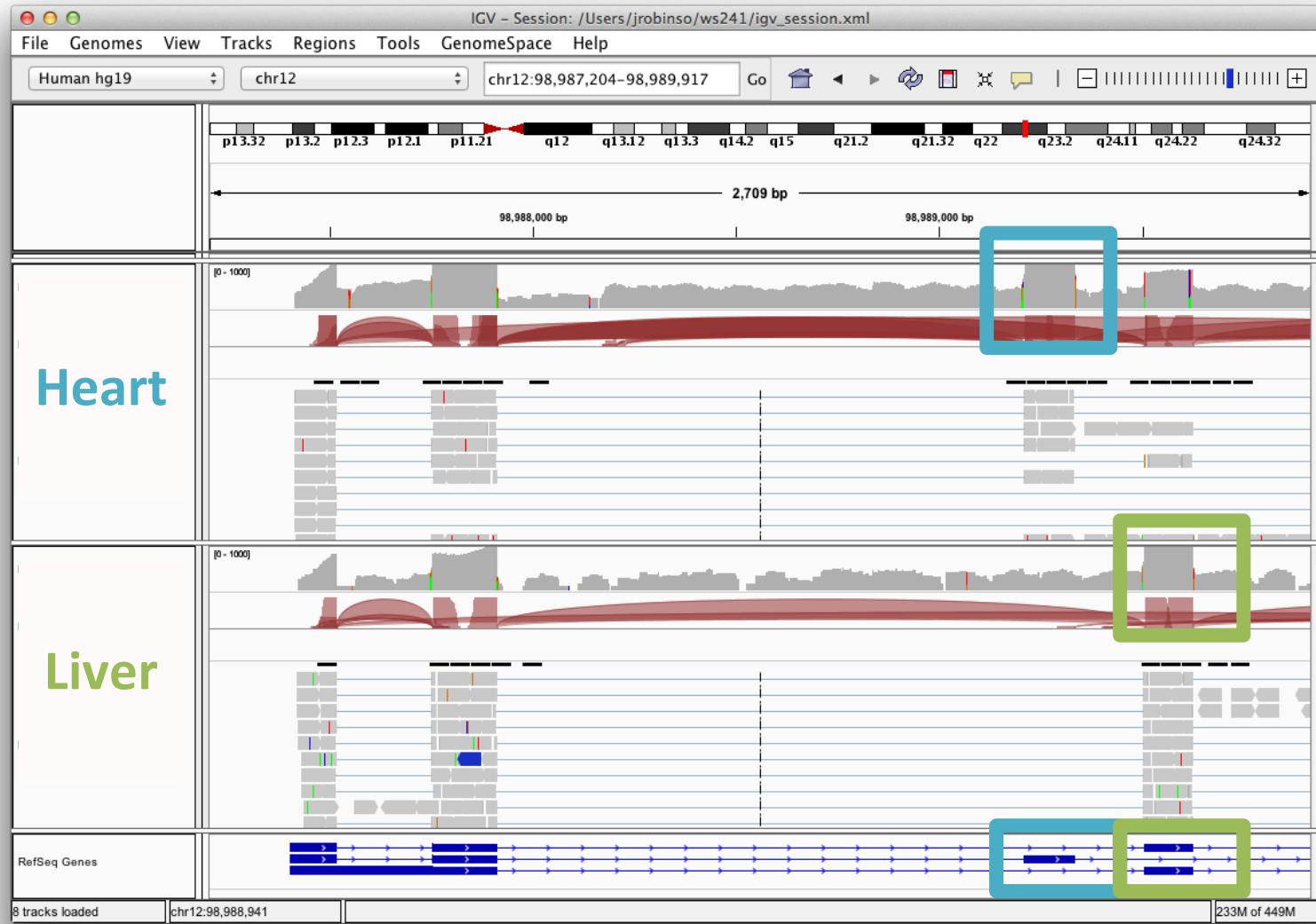


Relatively few span to this exon.

Many reads connect these exons.

RNA-seq alternative splicing

IGV lets you visualize evidence of alternative splicing.



Viewing our aligned reads in IGV



Visualization in *IGV*

Open a web browser (), then navigate to igv.org, and click on **Desktop application**.



For Developers

igv.js	Use igv.js to embed an interactive genome visualization component in your web app.
igv-jupyter	Extension for Jupyter Notebook which integrates igv.js

All IGV software is open source - MIT License.

GitHub repos: [IGV Desktop](#) | [IGV Web App](#) | [igv.js](#) | [igv-jupyter](#) | [Juicebox Web](#)

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. Integrative Genomics Viewer. *Nature Biotechnology* 29, 24–26 (2011)

Visualization in *IGV*

Click on the Downloads button.

The screenshot shows a web browser displaying the IGV software download page at software.broadinstitute.org/software/igv/. The page features a sidebar on the left with links like Home, Downloads (which is highlighted with a red box), and Documents. Below this is a search bar and the Broad Institute logo. The main content area has a large title "Integrative Genomics Viewer" and a preview image of the software's interface showing genomic tracks. At the bottom, there are sections for Overview, Downloads, and Funding.

Home

Downloads

Documents

Hosted Genomes

FAQ

+ IGV User Guide

+ File Formats

+ Release Notes

+ IGV for iPad

Credits

@ Contact

Search website

search

BROAD INSTITUTE

© 2013-2016 Broad Institute

Overview

The **Integrative Genomics Viewer (IGV)** is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Downloads

Download the IGV desktop application and igvtools.

Funding

Development of IGV is made possible by funding from th

Visualization in IGV

Select the appropriate button for your computer.

Install IGV 2.5.x



IGV Mac App

Download and unzip the Mac App Archive, then double-click the IGV application to run it. You can move the app to the *Applications* folder, or anywhere else.



IGV for Windows

Download and run the installer. An IGV shortcut will be created on the Desktop; double-click it to run the application.



IGV for Linux

Download and unzip the Archive. See the downloaded *readme.txt* for further instructions.



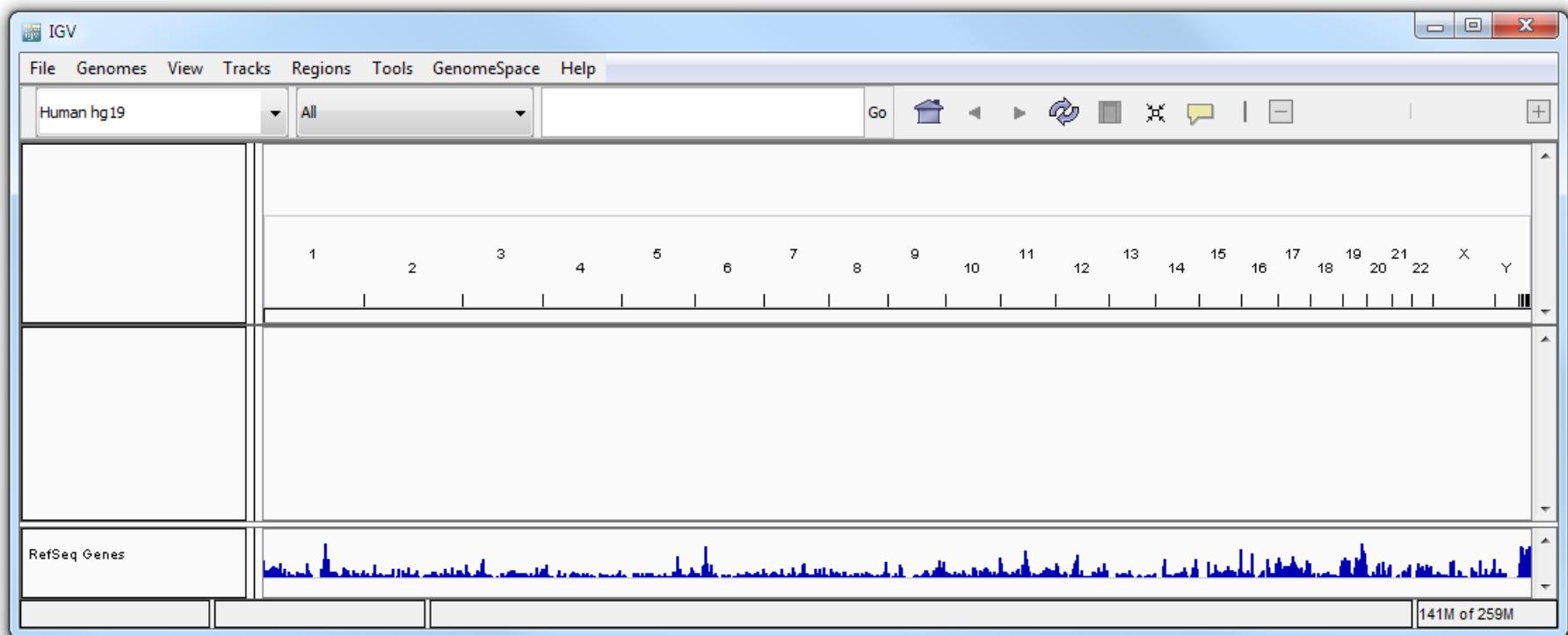
IGV and igvtools to run on the command line (all platforms)

Download and unzip the Archive. **Requires Java 11**. See the downloaded *readme.txt* and *igvtools_readme.txt* for further instructions.

Visualization in *IGV*

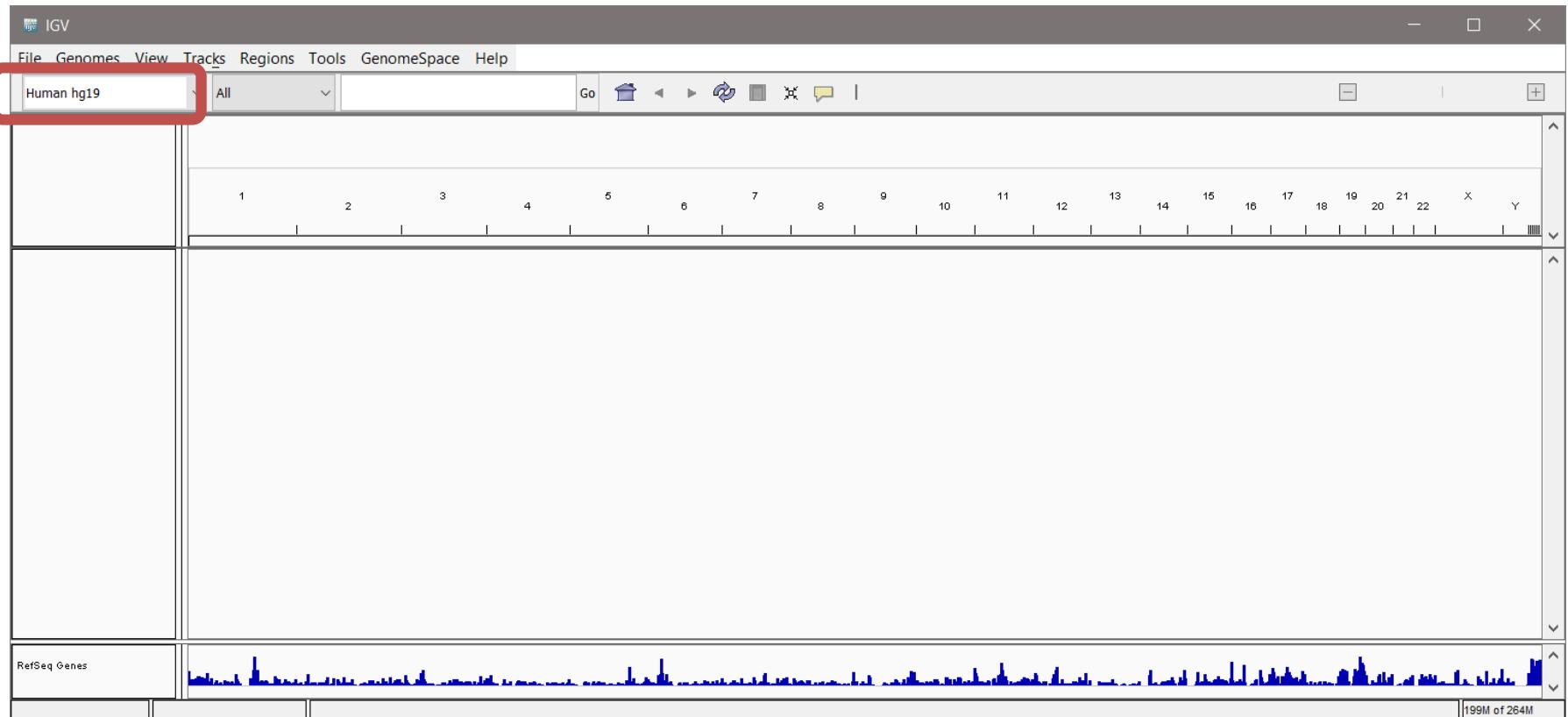
FA

An **executable IGV** file will be downloaded to your computer. Click to launch *IGV*. If a Java security warning pops up make sure to allow IGV access to the internet – it needs this to download genomes.



Visualization in *IGV*

Make sure *IGV* is using the **Human hg19** genome.



Visualization in *IGV*

If our CuffDiff job had run, we would have output like this:
And we would right-click to copy the URL of **genes.fpkm.tracking**.

1657632. **Cuffdiff** ⓘ

Source: Broad production (new)
submitted: Apr 03 10:11:20 AM, completed: Apr 03 10:34:17 PM, size: 106 MB
[Show details](#)

[Edit Sharing...](#) [Show input Parameters](#)

[Comments \(0\)](#)

[Tags \(0\)](#)

- [aligned.files: aligned.files.group.tsv](#)
- [GTF.file: ftp://gftp.broadinstitute.org/modu...n/gtf/Homo_sapiens_hg19_UCSC.gtf](#)
- [frag.bias.correct: ftp://gftp.broadinstitute.org/modu...genome/Homo_sapiens_hg19_UCSC.fa](#)
- [cmdline.log \(2.0 KB\) \(Last modified: Tue Apr 03 16:14:49 EDT 2018\)](#)
- [Homo_sapiens_hg19_UCSC.fa.fai \(1.0 KB\) \(Last modified: Tue Apr 03 16:15:37 EDT 2018\)](#)
- [gene_exp.diff \(2.5 MB\) \(Last modified: Tue Apr 03 22:33:29 EDT 2018\)](#)
- [isoform_exp.diff \(4.5 MB\) \(Last modified: Tue Apr 03 22:33:29 EDT 2018\)](#)
- [tss_group_exp.diff \(3.1 MB\) \(Last modified: Tue Apr 03 22:33:29 EDT 2018\)](#)
- [cds.count_tracking \(1.7 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [cds.diff \(1.5 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [cds_exp.diff \(3.2 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [**genes.fpkm_tracking \(2.4 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)**](#)
- [isoforms.count_tracking \(2.4 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [promoters.diff \(1.8 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [splicing.diff \(2.4 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [tss_groups.count_tracking \(1.7 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [tss_groups.fpkm_tracking \(3 MB\) \(Last modified: Tue Apr 03 22:33:30 EDT 2018\)](#)
- [cds.read_group_tracking \(10.5 MB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [genes.count_tracking \(1.4 MB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [genes.read_group_tracking \(7.9 MB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [isoforms.read_group_tracking \(15.2 MB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [read_groups.info \(2.0 KB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [run.info \(2.0 KB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [tss_groups.read_group_tracking \(10.3 MB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [stdout.txt \(23.1 MB\) \(Last modified: Tue Apr 03 22:33:31 EDT 2018\)](#)
- [gp_execution_log.txt \(1.0 KB\) \(Last modified: Tue Apr 03 22:34:17 EDT 2018\)](#)

Visualization in *IGV*

Since it did not, go to the notebook for this class, and copy the URL of the **genes.fpkm.tracking** file found in the last cell of the notebook.

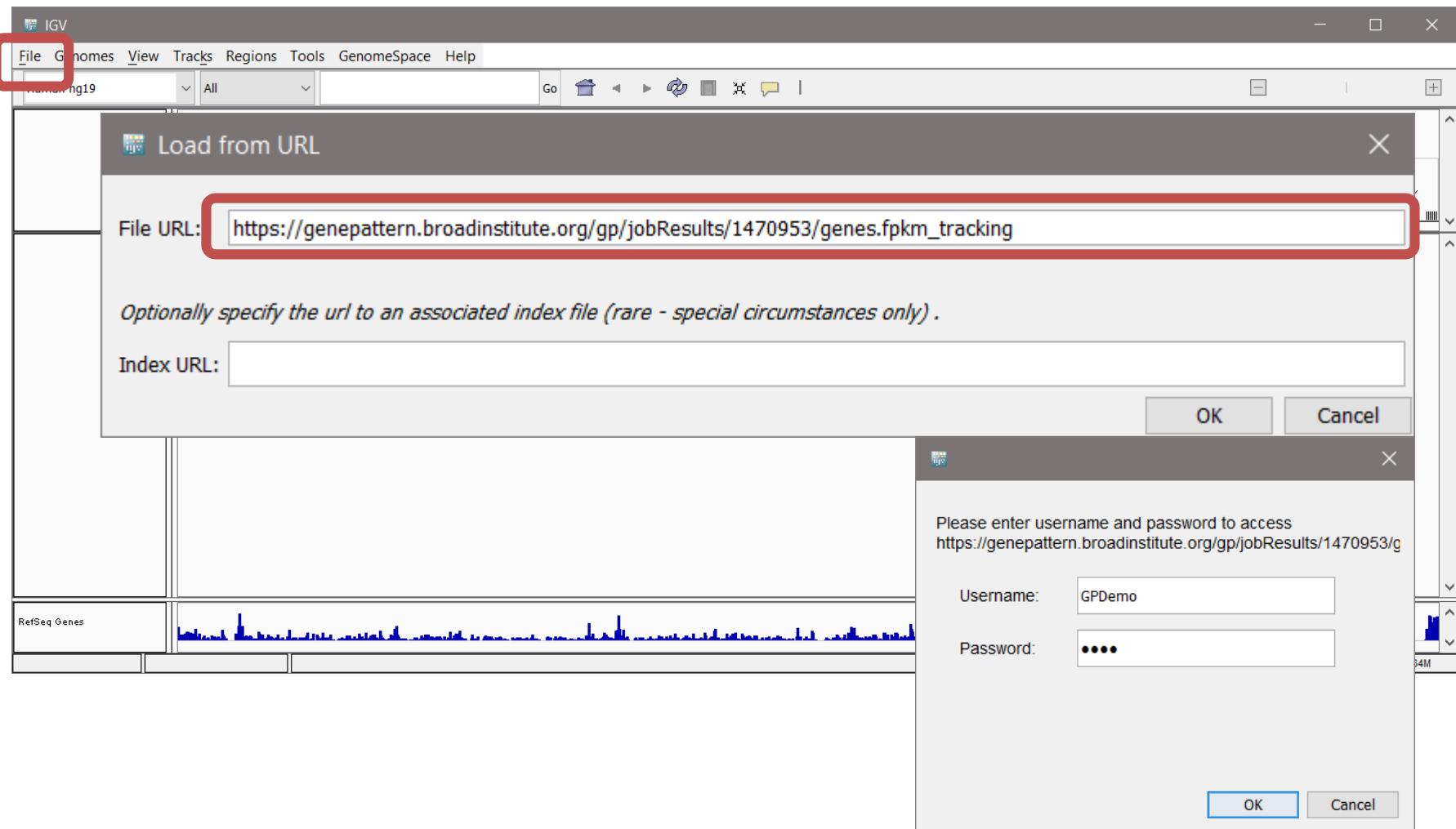
For the next step in this workflow, [CuffDiff](#), - we will use the [GenePattern WebApp](#) - the setting of conditions is currently not available in the GenePattern Notebook extension. This feature is due in an upcoming release of the GenePattern Notebook Extension.
Log in to the GenePattern server with the same username and password you used for this notebook, and you will be able to use all of the jobs you have run in this notebook.

I've included the output of the CuffDiff job here - you can also copy the **genes.fpkm_tracking** file from here to use in IGV

[**genes.fpkm_tracking**](#)
[cmdline.log](#)
[Homo_sapiens_hg19_UCSC.fa.fai](#)
[isoforms.count_tracking](#)
[isoforms.fpkm_tracking](#)
[isoforms.read_group_tracking](#)
[gene_exp.diff](#)
[genes.read_group_tracking](#)
[genes.count_tracking](#)
[cds.count_tracking](#)
[promoters.diff](#)
[read_groups.info](#)
[run.info](#)
[isoform_exp.diff](#)
[cds_exp.diff](#)
[cds.read_group_tracking](#)
[cds.fpkm_tracking](#)
[splicing.diff](#)
[stdout.txt](#)
[cds.diff](#)
[tss_groups.count_tracking](#)
[tss_groups.fpkm_tracking](#)
[tss_groups.read_group_tracking](#)
[tss_group_exp.diff](#)
[gp_execution_log.txt](#)

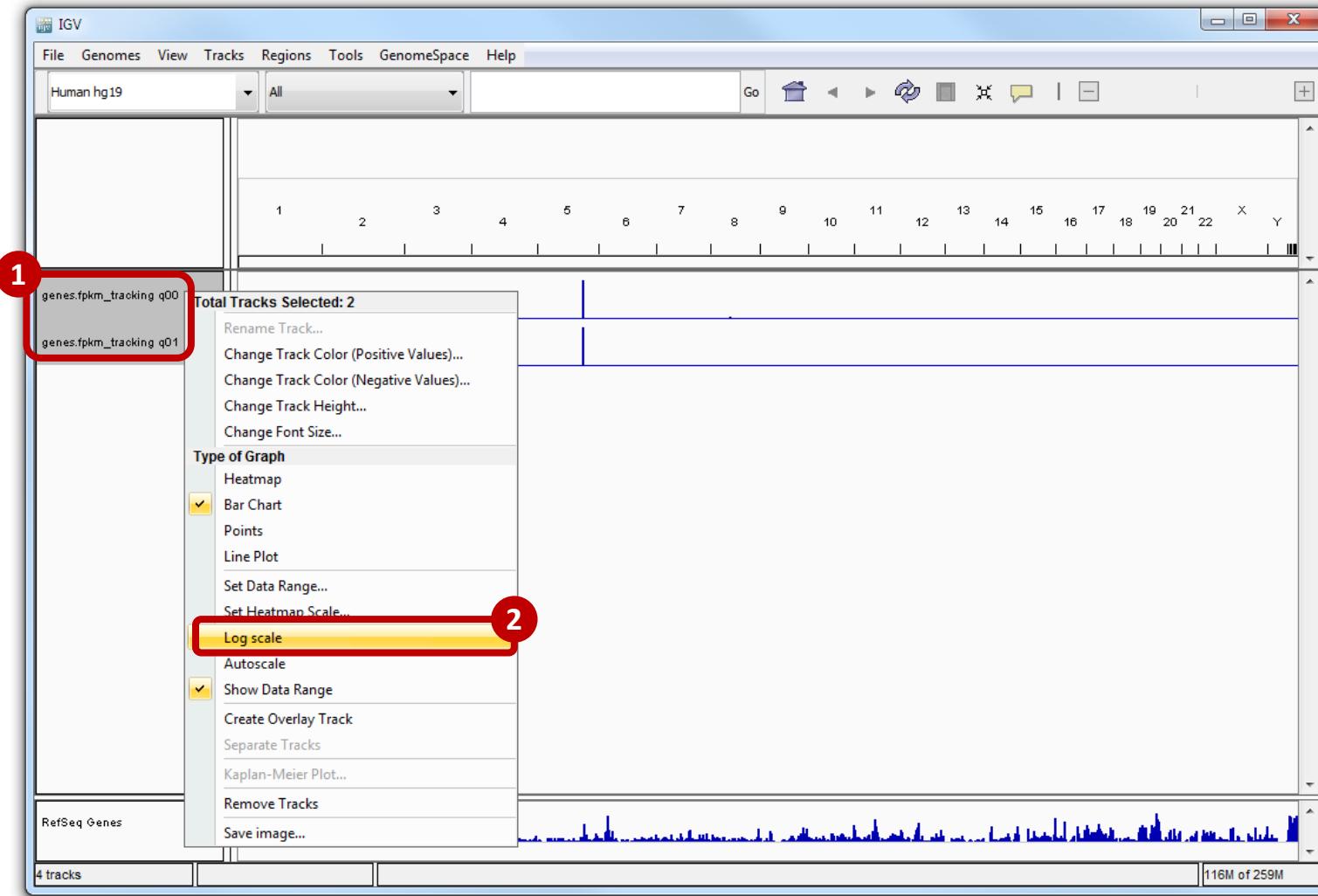
Visualization in IGV

From the IGV *File* menu, you would then paste in the **genes.fpkm.tracking** URL you just copied, click OK. Then enter your GenePattern login and password.



Visualization in IGV – scale matters

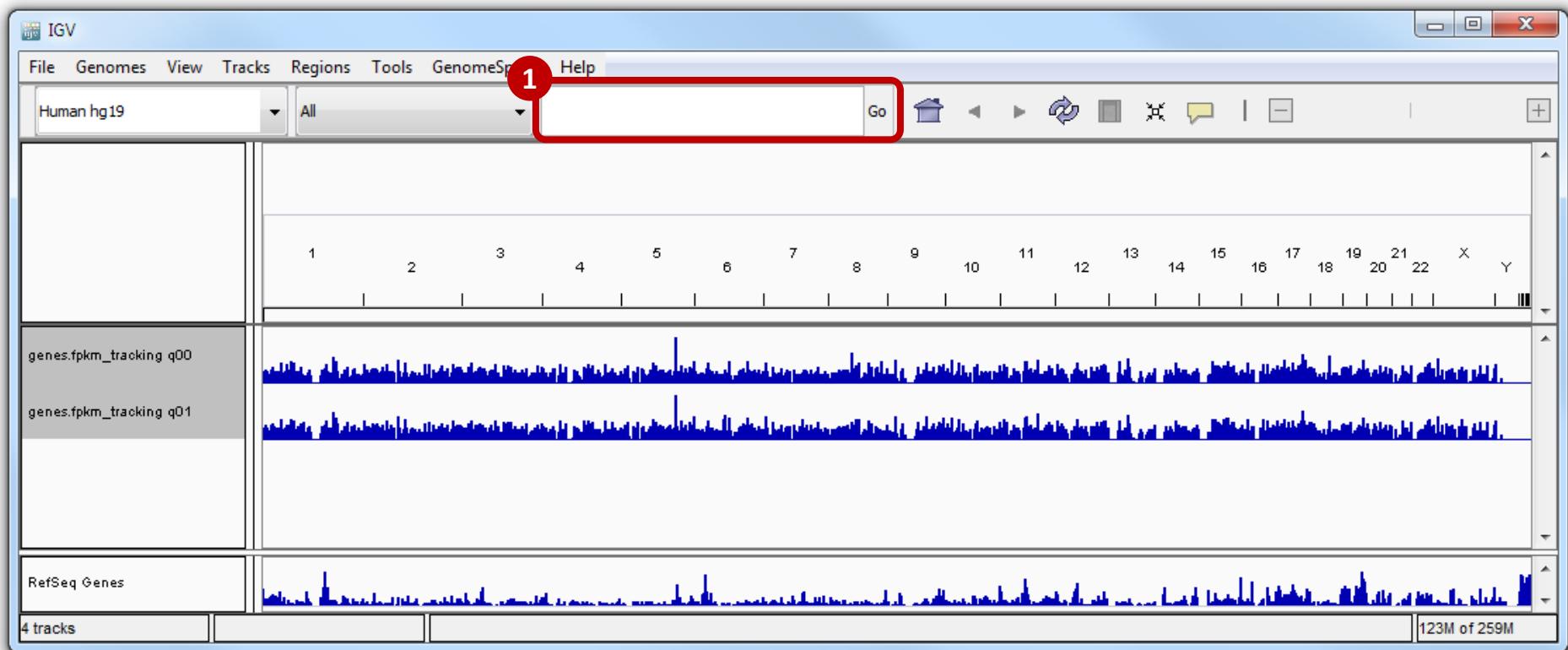
- ① Ctrl+Click the **q00** and **q01** tracks.
- ② Right-click and scroll down, then select **log scale** to correctly display the data.



Visualization in *IGV*

Try looking for differences between *untreated* vs. *dex* gene expression.

- ① Search for **C7** in the search bar.



Visualization in IGV

Example genes: **C7**, CCDC69, DUSP1, FKBP5, GPX3, KLF15, MAOA, SAMHD1, SERPINA3, SPARCL3, TSC22D3, CRISPLD2, C13orf15, PER1



similar expression

dex >> untreated

low expression

Visualization in *IGV*

In addition to looking at the gene expression in the **genes.fpkm_tracking** file, we can look at how the reads stack against transcripts by looking at the aligned reads.

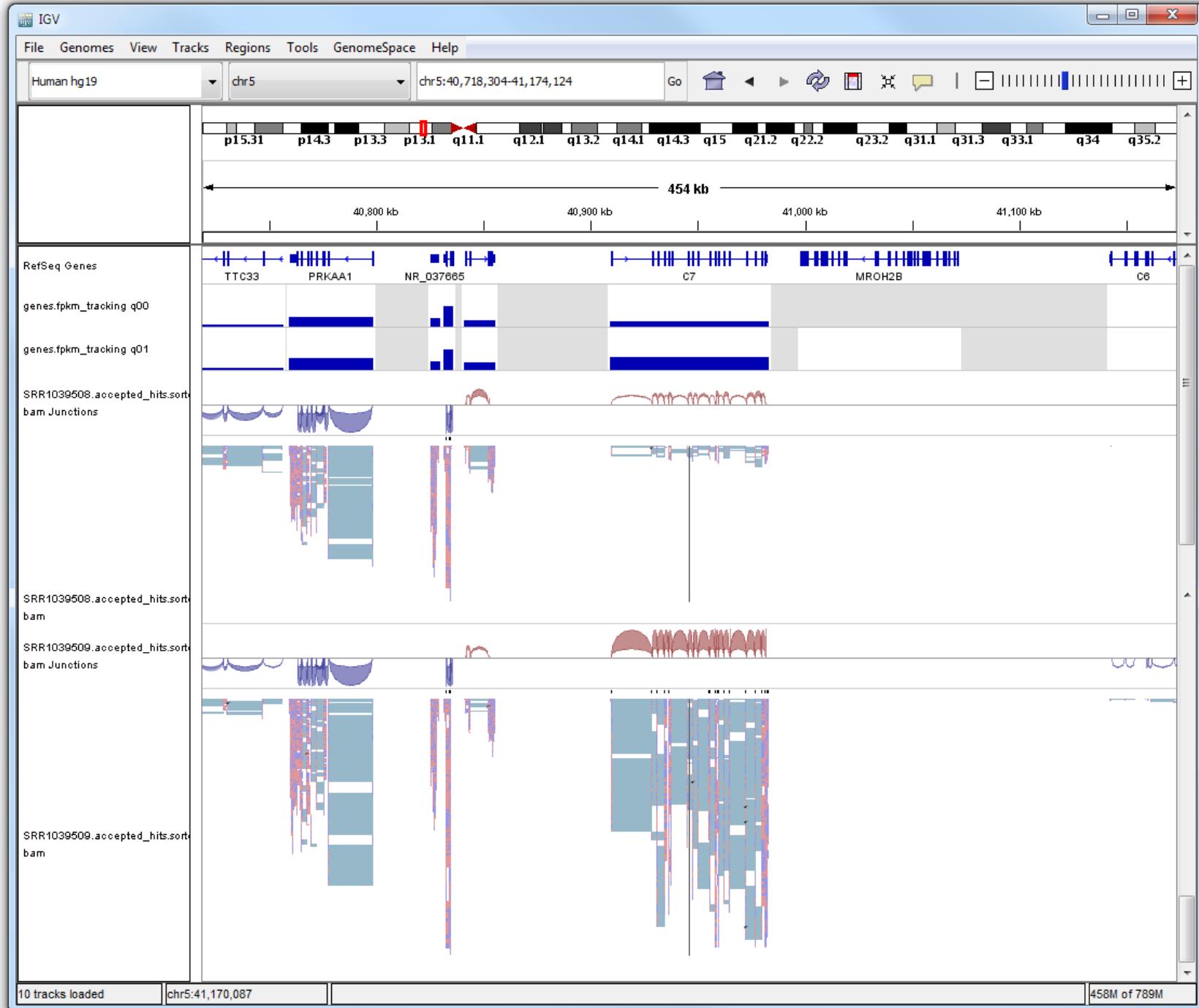
Load the following two files into *IGV*:

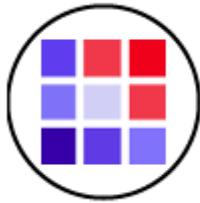
SRR1039508.accepted_hits.sorted.bam *untreated*

SRR1039509.accepted_hits.sorted.bam *dex*

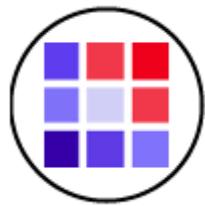
Steps:

1. Open a browser window to
genepattern.datasets.org/?prefix=data/HASM_Asthma_RNA-seq_workshop_files/SortSam_Output/
2. Copy the URL for **SRR1039508.accepted_hits.sorted.bam**
3. In *IGV*, navigate to: **File/Load from URL**
4. Paste in the URL you just copied.
5. Click **Open**.
6. Repeat **Steps 1-4** to load the second file.





Other GenePattern Features



Managing Job Results

Managing job results

- Viewing job results
- Saving and deleting job results
- Sharing job results with others

Viewing job results

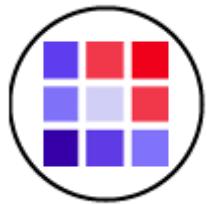
Where are my job results?

Will my jobs stay on the server?

- Job details and result files are, by default, purged from the cloud GenePattern server after 30 days – however this can be configured by the server administrator. Save result files and job information
 - to prevent them from being lost on purge
 - to share them with others

Managing job access

- By default all jobs are private (you + administrator)
- Grant access to your jobs (input files + parameter values + result files).
- Grant read-only access or read-write access
- Share with members of your group, or share with everyone (group = Public)

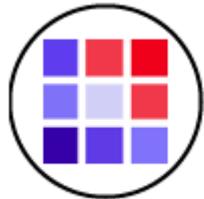


Batching Jobs

Example: run *FastQC* on all pairs of FASTQ files

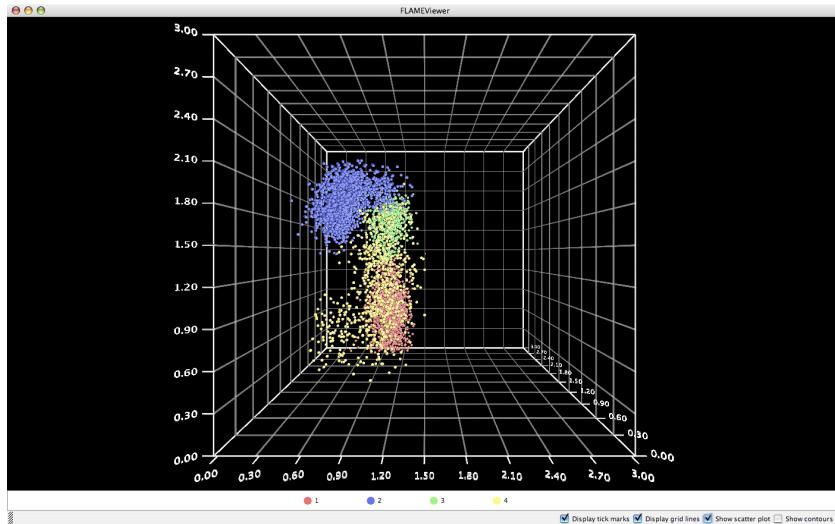
Find the ***.fastq.gz** files in the your folder.
Then click and drag *all* *.fastq.gz to the **input** box.

The screenshot shows the GenePattern web interface with the FastQC module selected. The left sidebar shows a file tree with several folders like 'Public', '2017CEGS', and user-specific folders. In the '2017CEGS' folder, a red box highlights a list of FASTQ files: SRR1039509_1.fastq.gz, SRR1039509_2.fastq.gz, SRR1039512_1.fastq.gz, SRR1039512_2.fastq.gz, SRR1039513_1.fastq.gz, SRR1039513_2.fastq.gz, SRR1039516_1.fastq.gz, SRR1039516_2.fastq.gz, SRR1039517_1.fastq.gz, SRR1039517_2.fastq.gz, SRR1039520_1.fastq.gz, SRR1039520_2.fastq.gz, SRR1039521_1.fastq.gz, and SRR1039521_2.fastq.gz. A red arrow points from this list to the 'input file*' field in the main form. The 'input file*' field has a red border and contains the same list of files. To the right of the field is a 'Batch' checkbox, also highlighted with a red box. The 'Basic' section of the form includes a 'Drag Files Here' area and a note about 2GB file upload limit. Other sections like 'Advanced', 'Memory Settings', 'Comments', and 'Tags' are also visible.

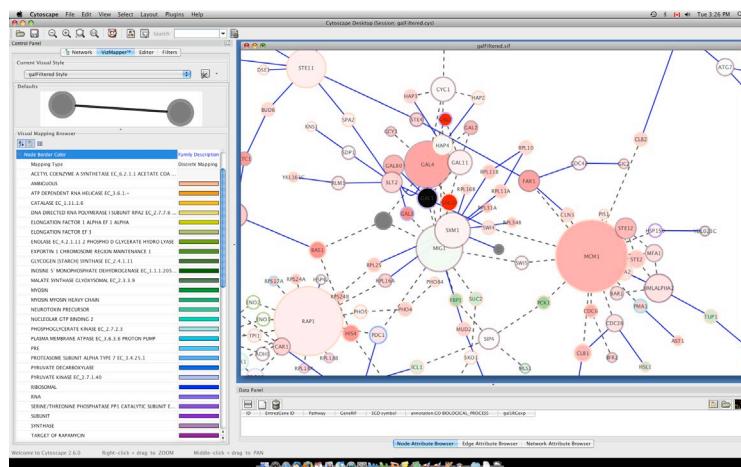


Creating your own Notebooks & Editing existing

Other GenePattern Features

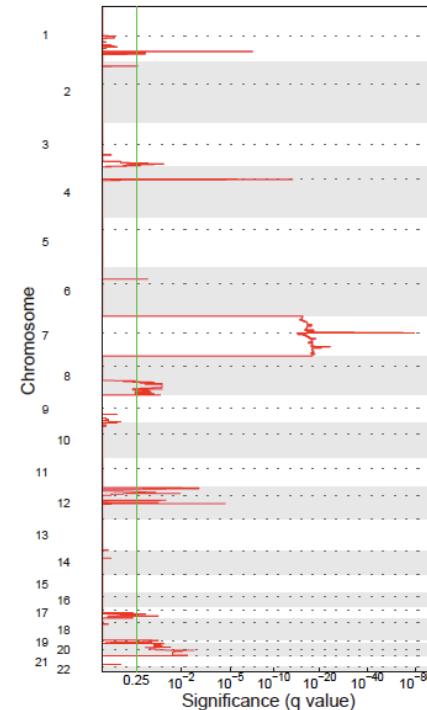


Flow Cytometry



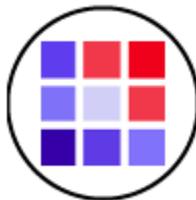
Network Analysis

Amplifications
Glioma (initial dataset, n=141)



Sequence Variation Analysis

Proteomics



Closing

Resources

Public GenePattern server
cloud.genepattern.org

Notebook website
genepattern-notebook.org

GenePattern Notebook
notebook.genepattern.org

Integrative Genomics Viewer (IGV)
www.igv.org

GenePattern Website
genepattern.org

GenePattern Archive (GPArc)
gparc.org

Our Team

Anthony Castanza – San Diego, CA

David Eby - Japan

Barbara Hill – Cambridge, MA

Edwin Juarez – San Diego, CA

Forrest Kim – San Diego, CA

Ted Liefeld – San Diego, CA

Michael Reich – San Diego, CA

Jim Robinson – San Diego, CA

Thorin Tabor – San Diego, CA

Pablo Tamayo – San Diego, CA

Helga Thorvaldsdottir – Cambridge, MA

Douglass Turner – Cambridge, MA

PI

Jill P. Mesirov – San Diego, CA



Keep in touch!

Feature requests, bug reports, and general help
genepattern.org/help

Mailing list to receive GenePattern news.
Sign up at www.genepattern.org/gp_mail.html



@GenePattern

Thank You!