# 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

gatk — 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

**DESeq2**

platforms all | downloads top 5% | in Bioc 4.5 years | build ok

DOI: 10.18129/B9.bioc.DESeq2

Differential gene expression analysis based on the negative binomial distribution

**HISAT2**

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

**Classification And Regression Trees for Machine Learning**

**Samtools**

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

| | |
|---|---|
| Samtools | Reading/writing/editing/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 htslib so they can be built independently.

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.

BOW TIE **Bowtie 2**

Fast and sensitive read alignment

scikit learn — Home | Installation | Documentation | Examples

**1.4. Support Vector Machines**

Hierarchical Clustering / Dendrograms

Burrows-Wheeler Aligner

**Cufflinks**

*Transcriptome assembly and differential expression analysis for RNA-Seq.*

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:

**GSEA**

Gene Set Enrichment Analysis

Integrative Genomics Viewer

**NMF: Non-negative Matrix Factorization**

**What is HAPSEG?**

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

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

**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.

**Cytoscape**

Network Data Integration, Analysis, and Visualization in a Box

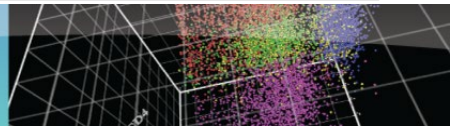
**Trimmomatic: A flexible read trimming tool for Illumina NGS data**

**MSigDB**

Molecular Signatures Database

FLAME

FLow analysis with Automated Multivariate Estimation



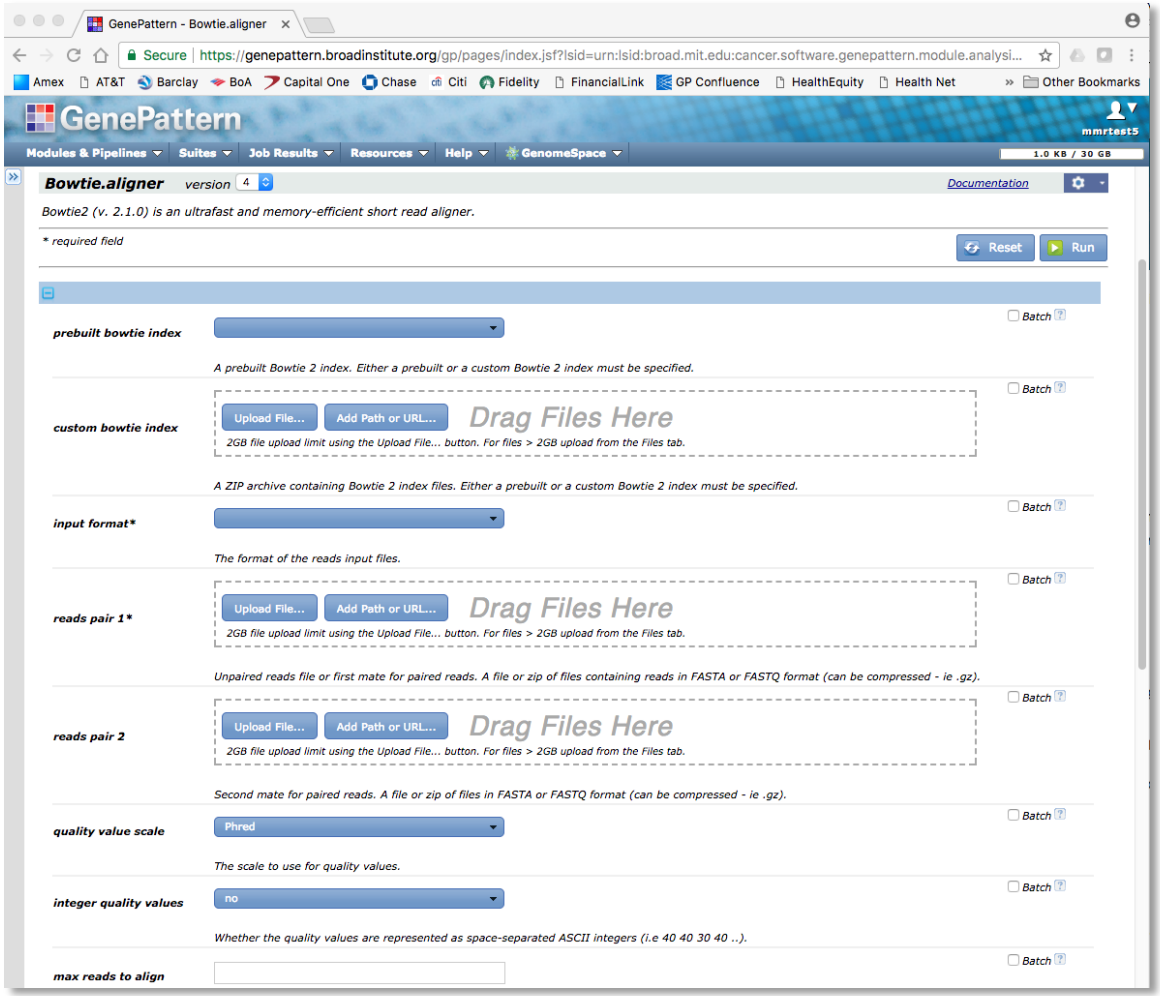
# Problems with bioinformatics tool use and interoperability

- 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 --chunkmbs 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



# 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

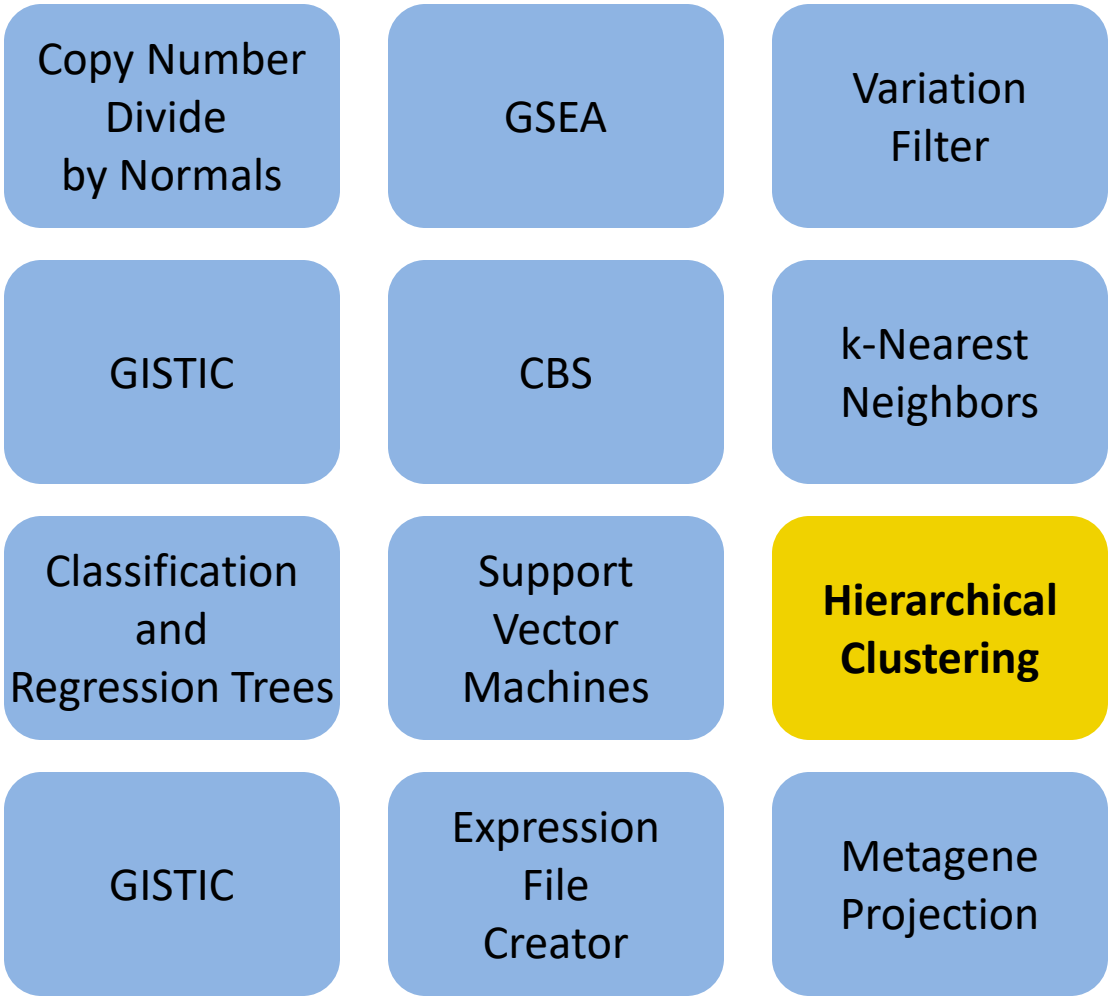
# Solution features: Reproducibility

- 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




**Hierarchical Clustering Files**

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

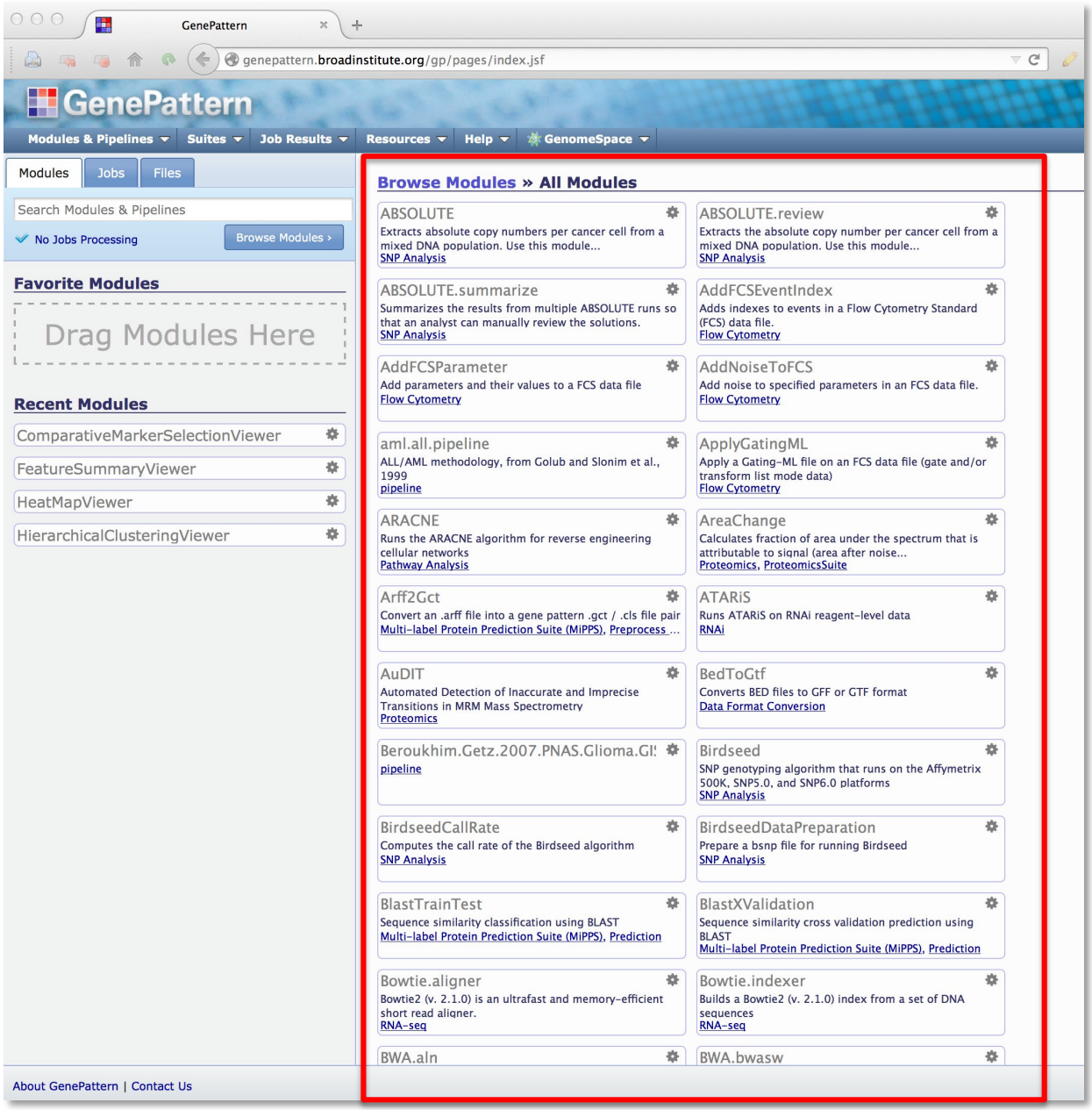
**Documentation**

HierarchicalClustering.pdf

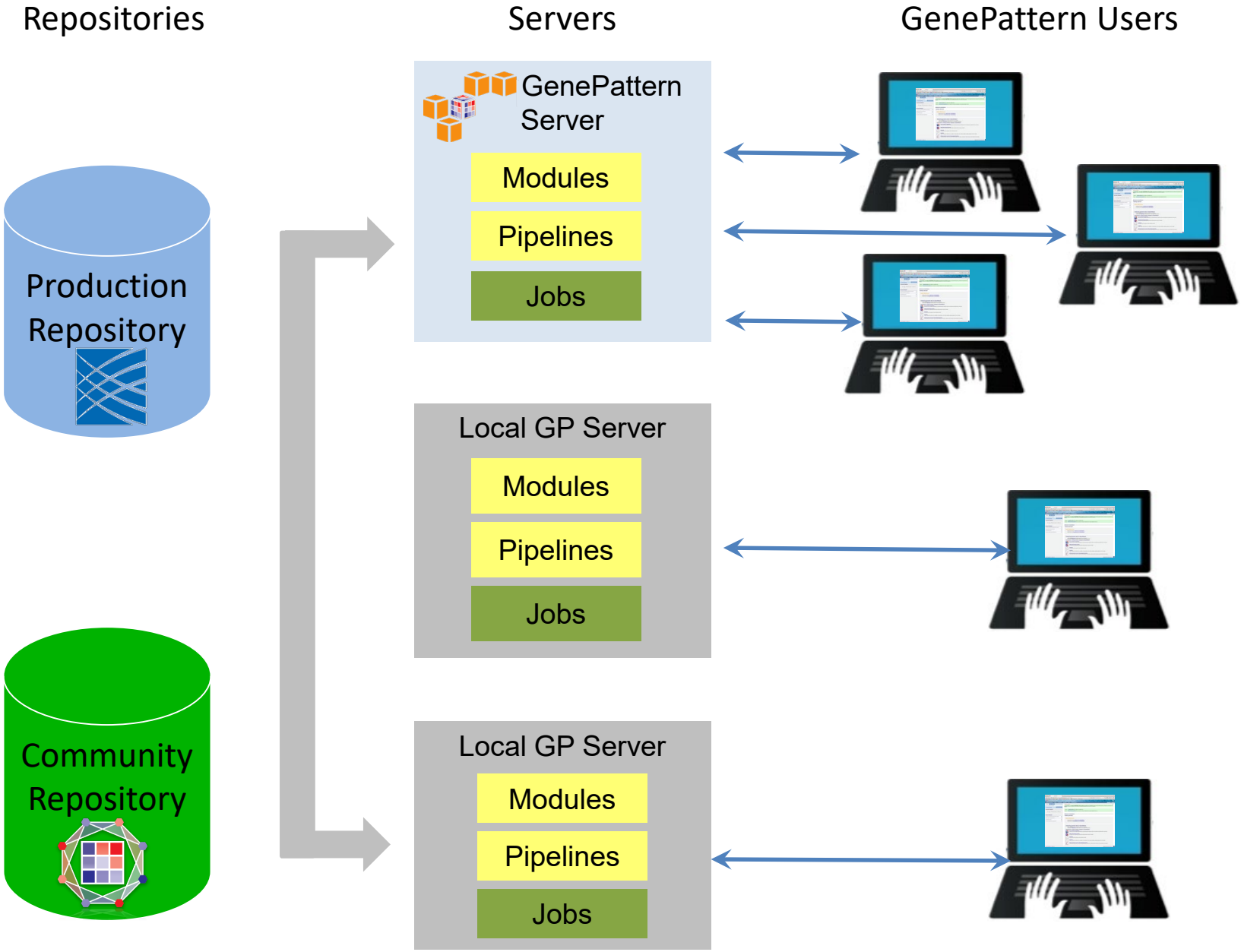
**Parameter descriptions**

-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>

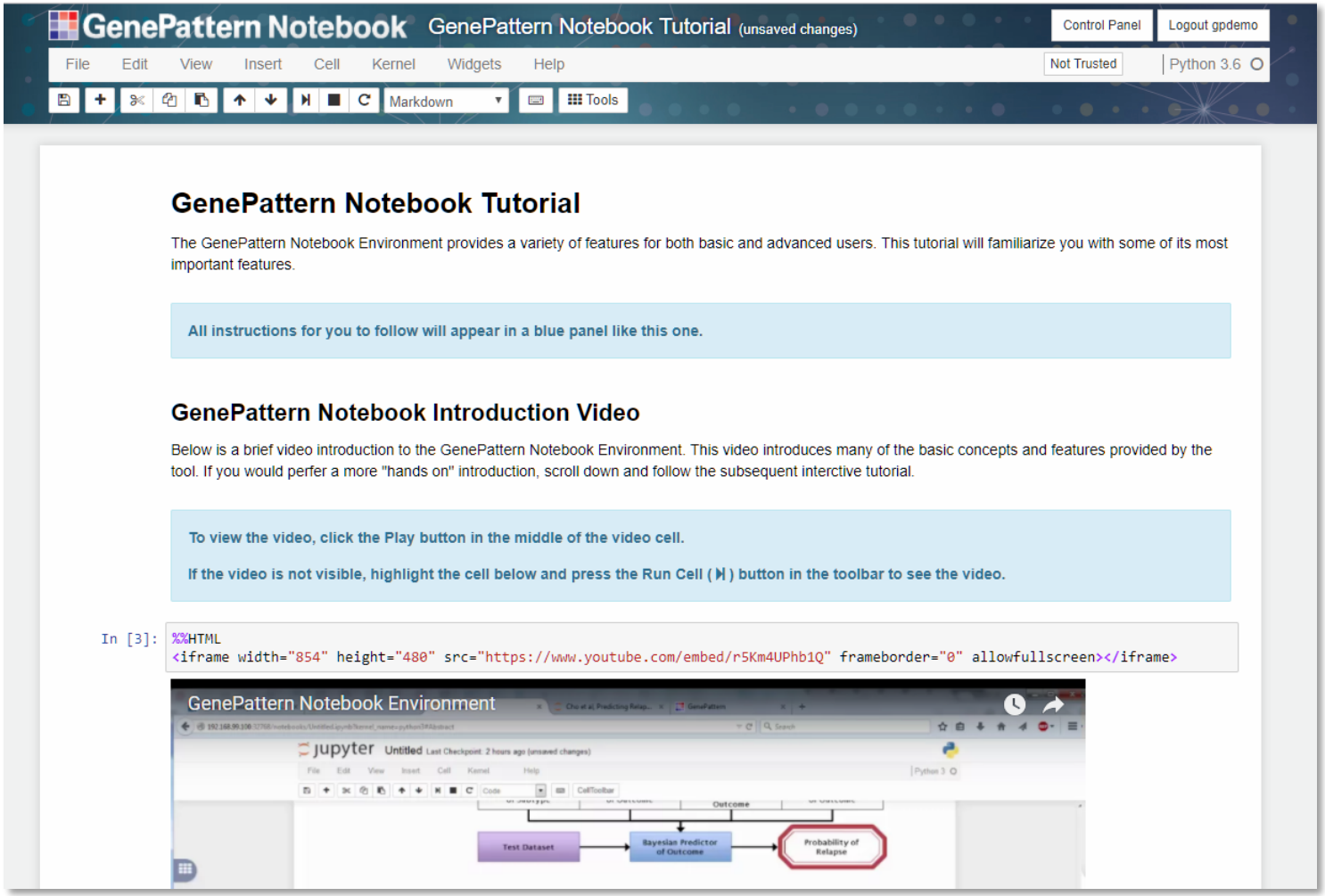
# >250 GenePattern Modules, 4/2019



# How GenePattern works



# GenePattern Notebook Environment



# Jupyter Notebook

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

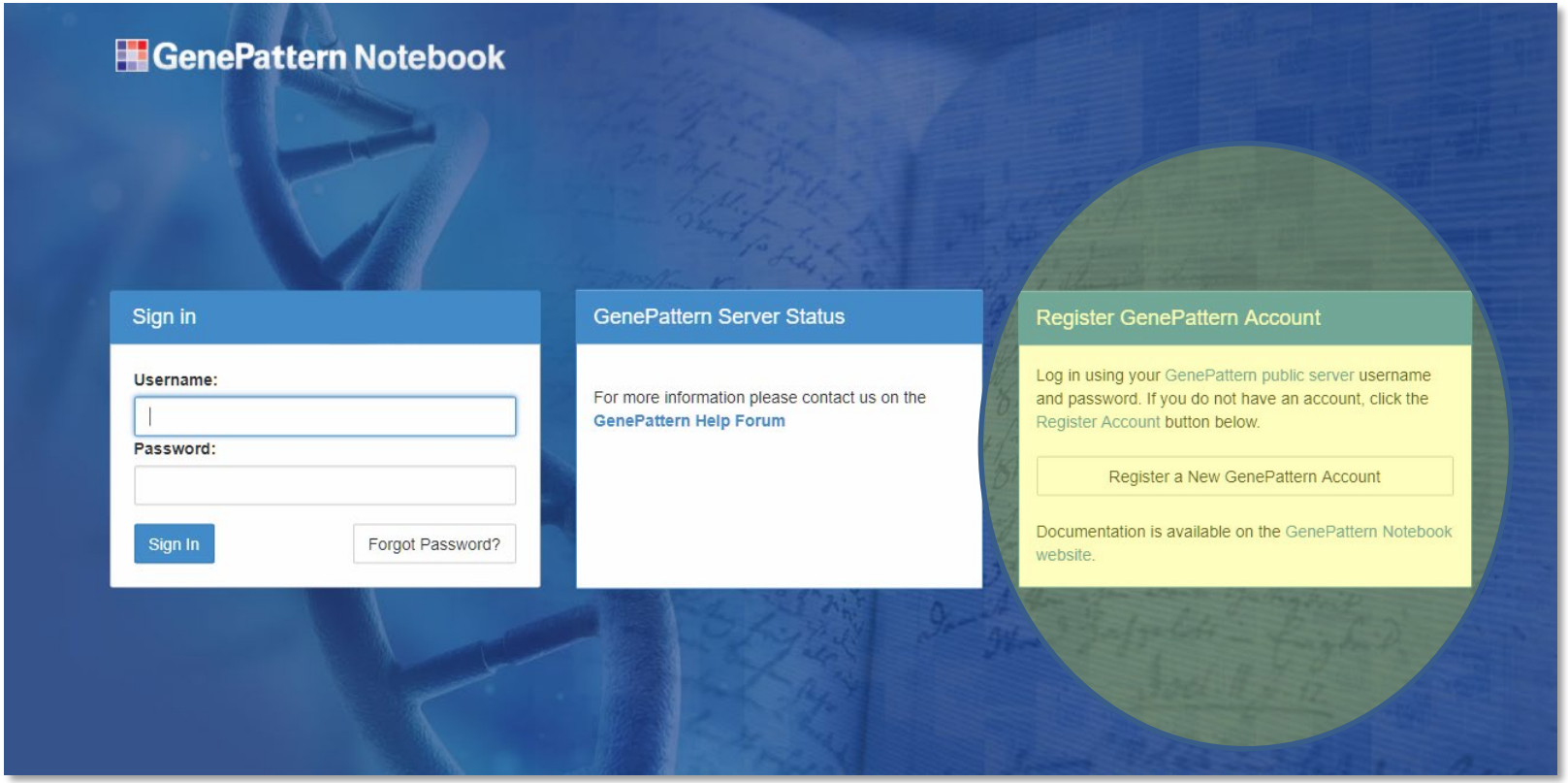
# Complete Research Narrative

- Leverages the best of Jupyter and GenePattern

- Interleave text, visualization, graphics and analytical aspects

# Exercise

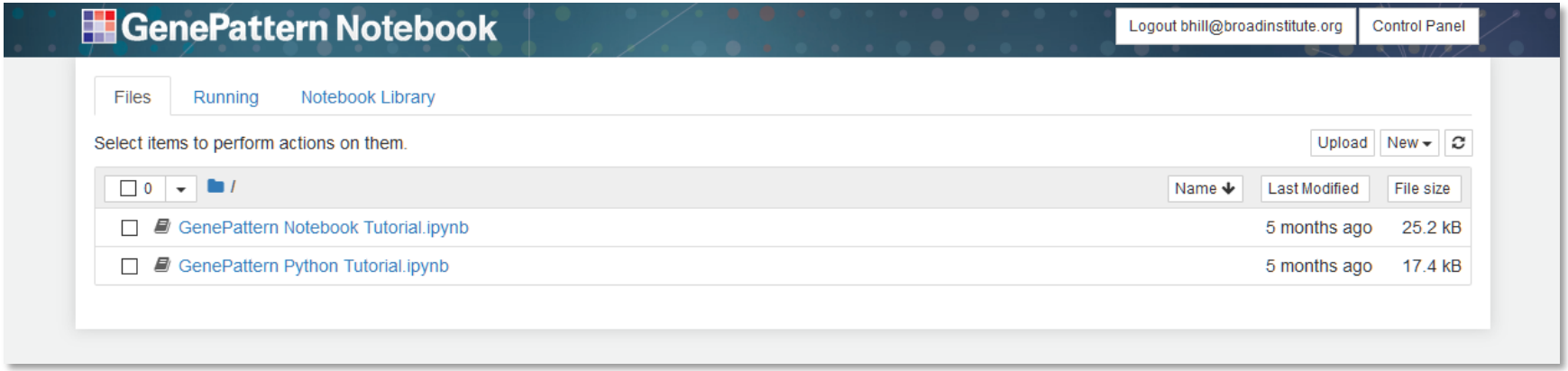
# GenePattern Notebook Repository




https://notebook.genepattern.org

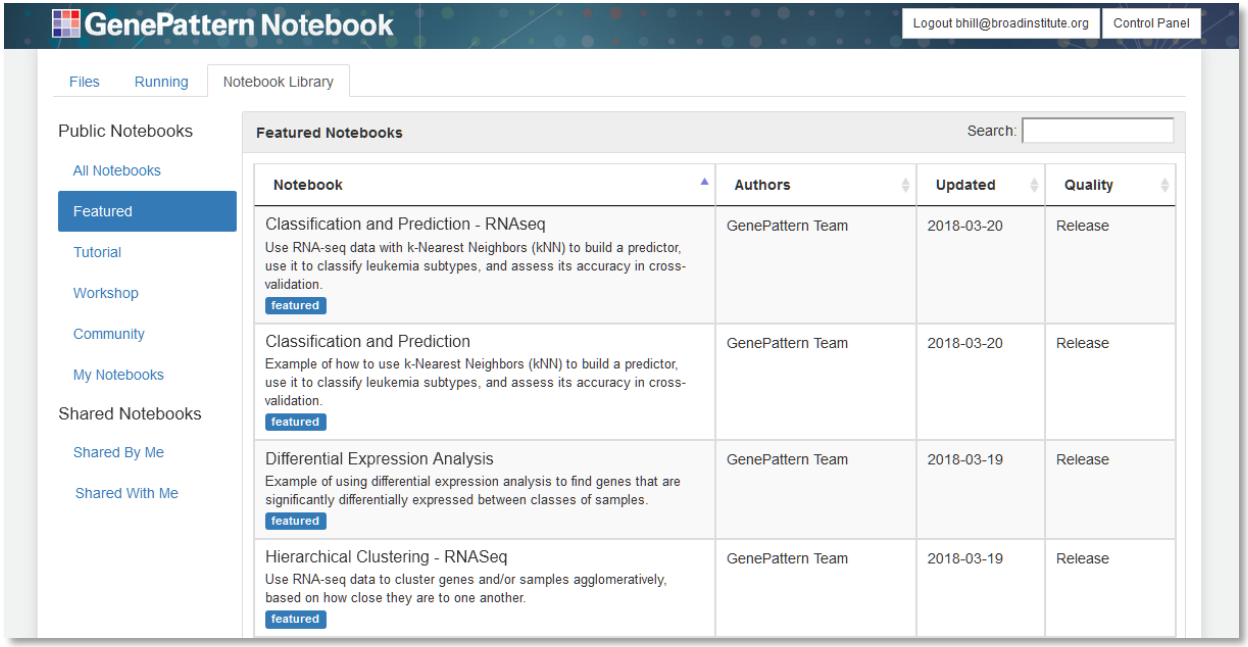
# Notebook Workspace

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



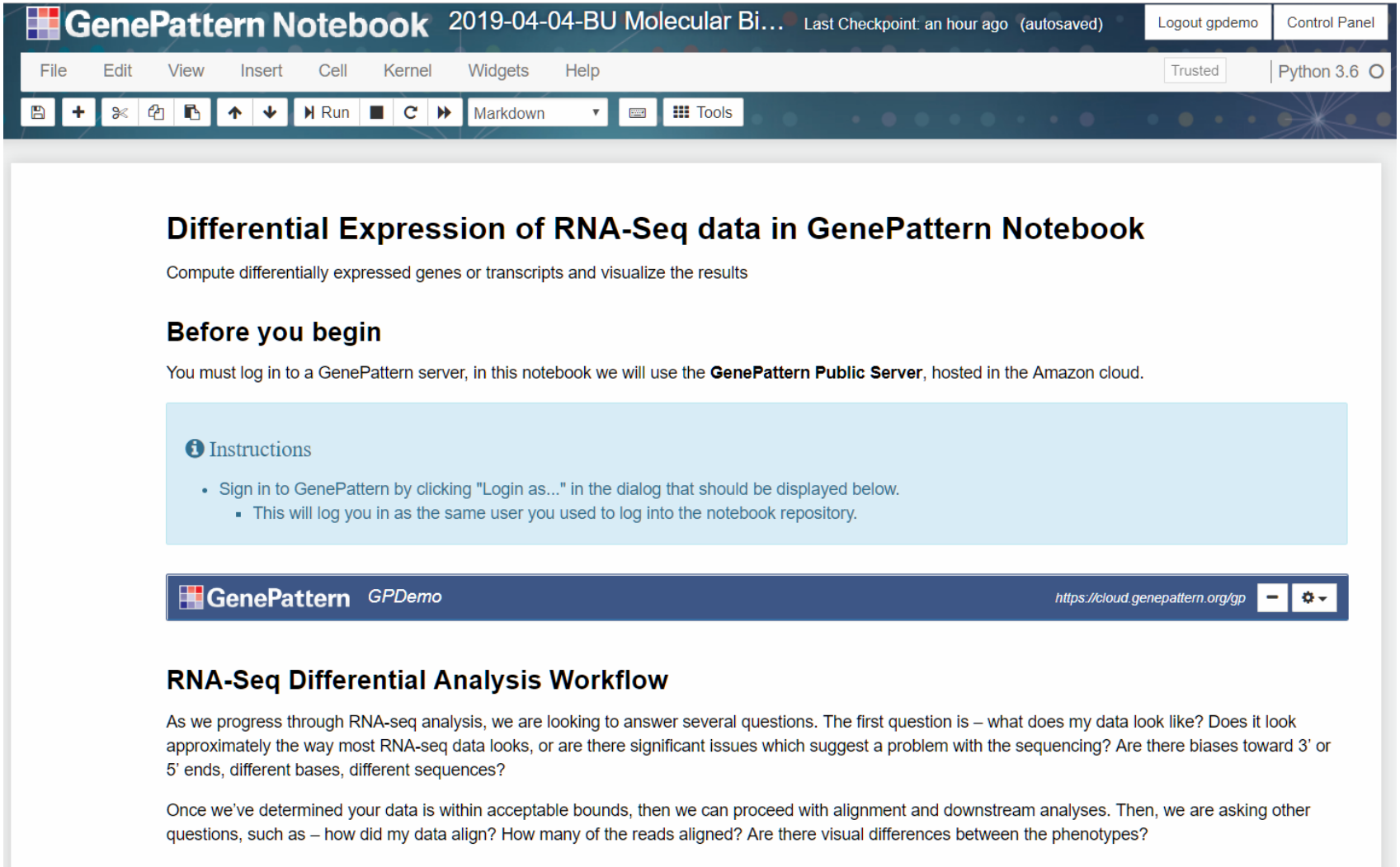
# Browse Public Notebooks

- 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.



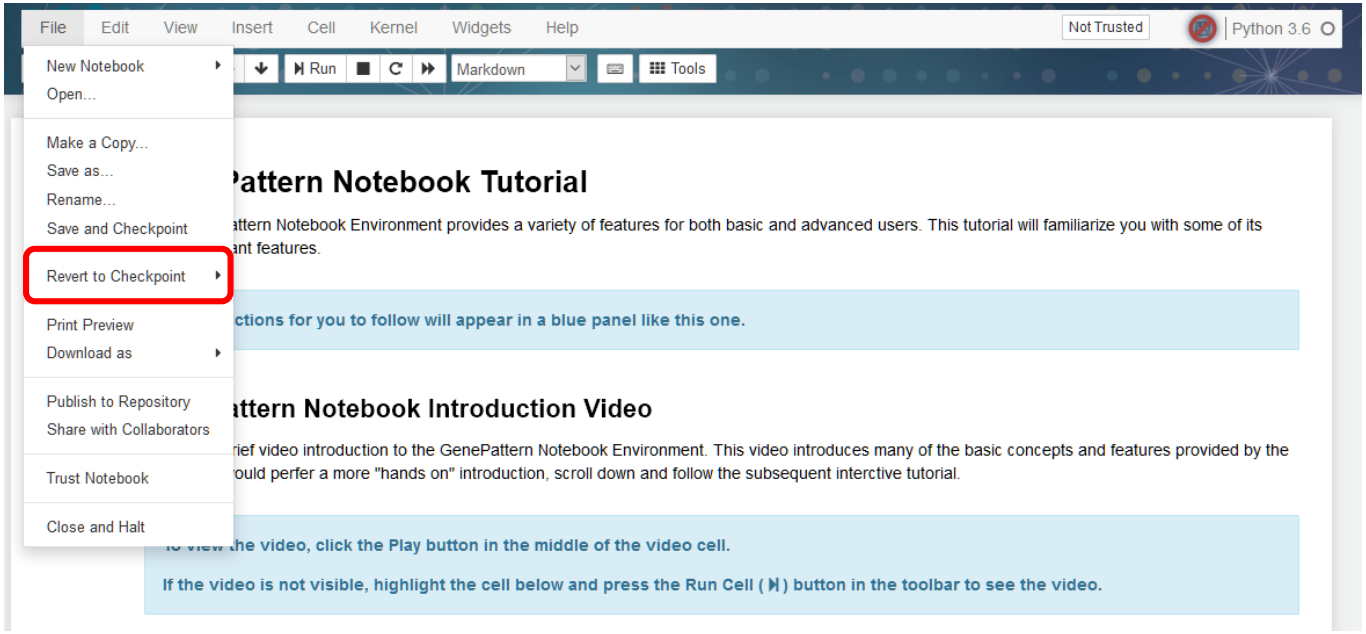
# Run an Analysis Notebook

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



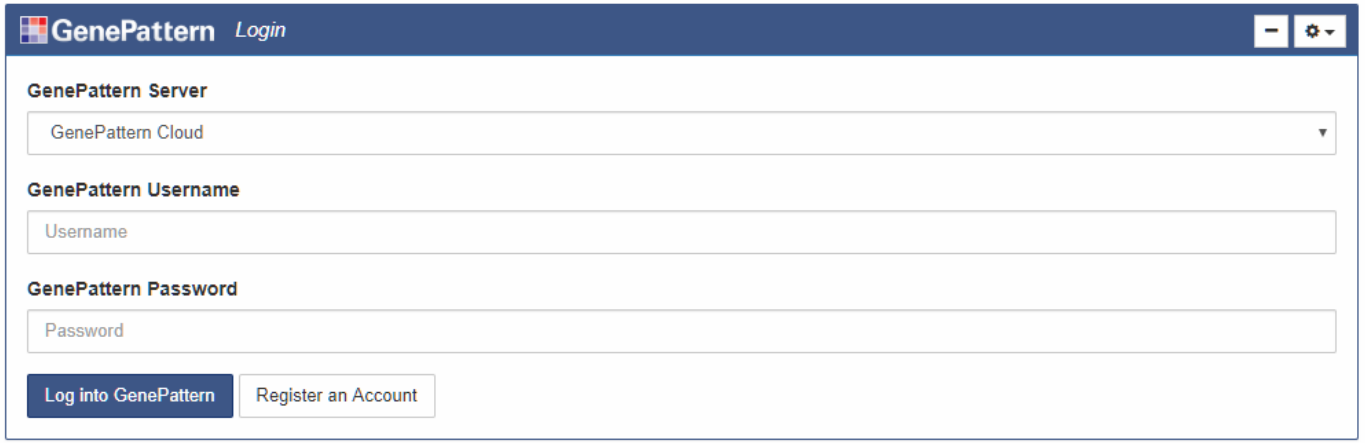
# Name, Save & Checkpoint Notebooks

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

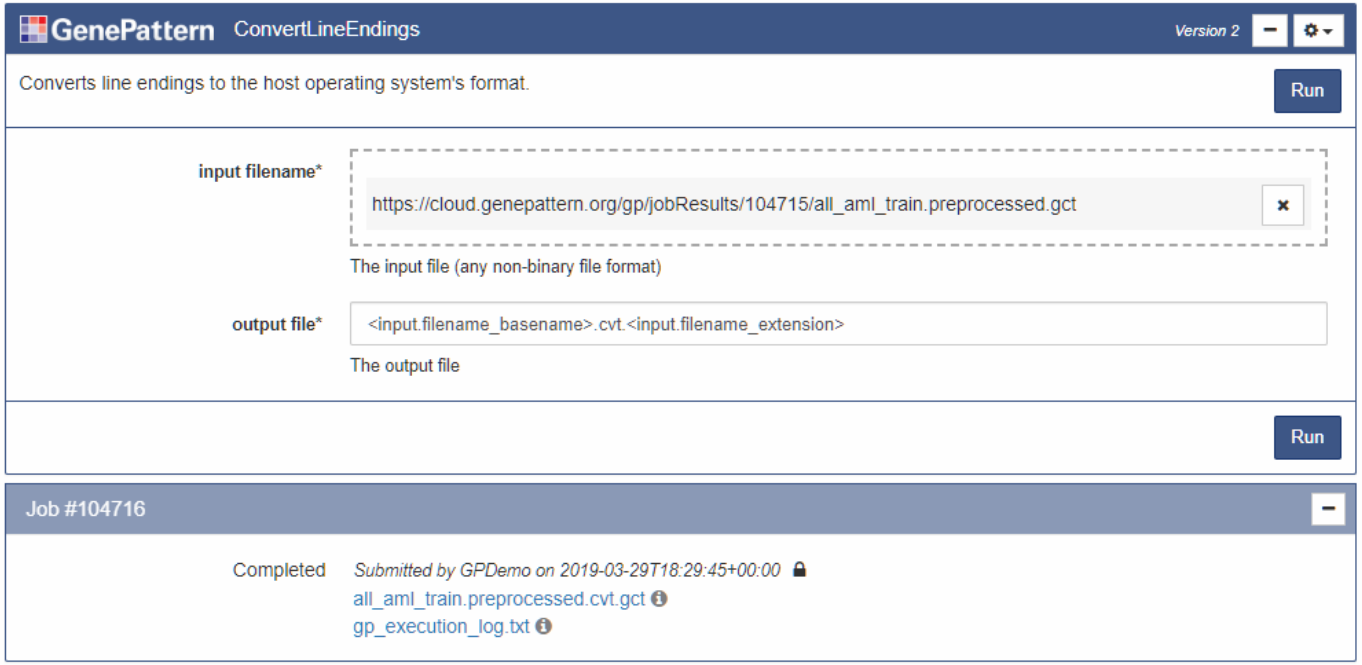


# GenePattern Cells

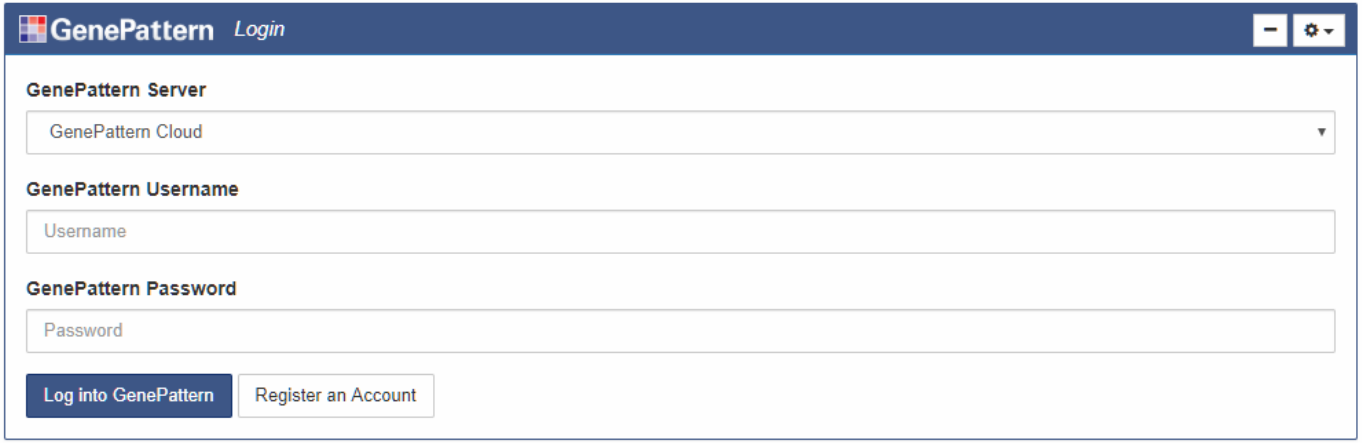
Authentication Cell

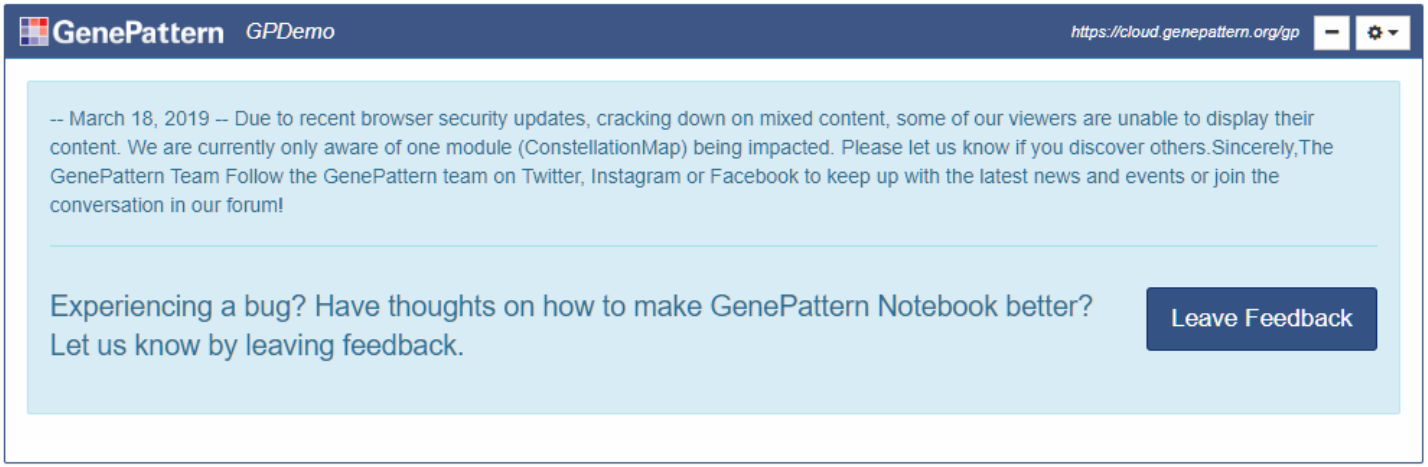



Analysis Cell



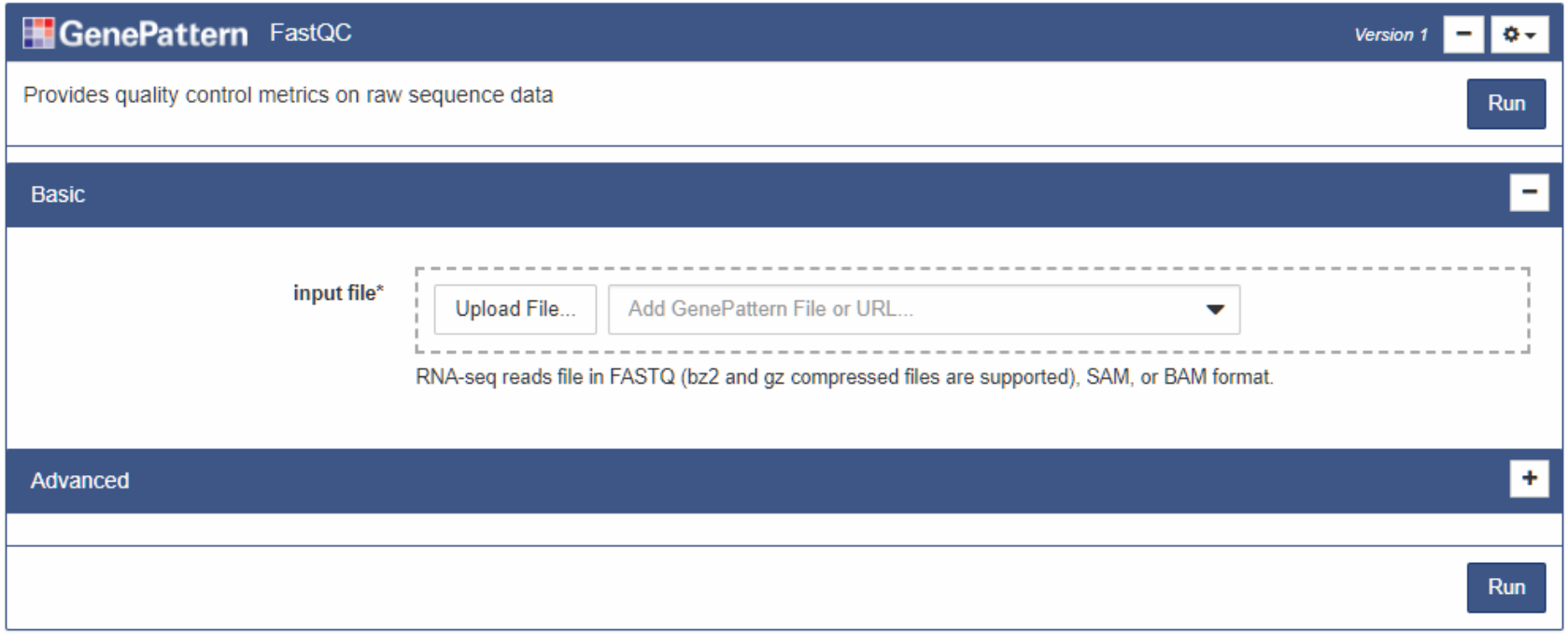
# Authentication Cells



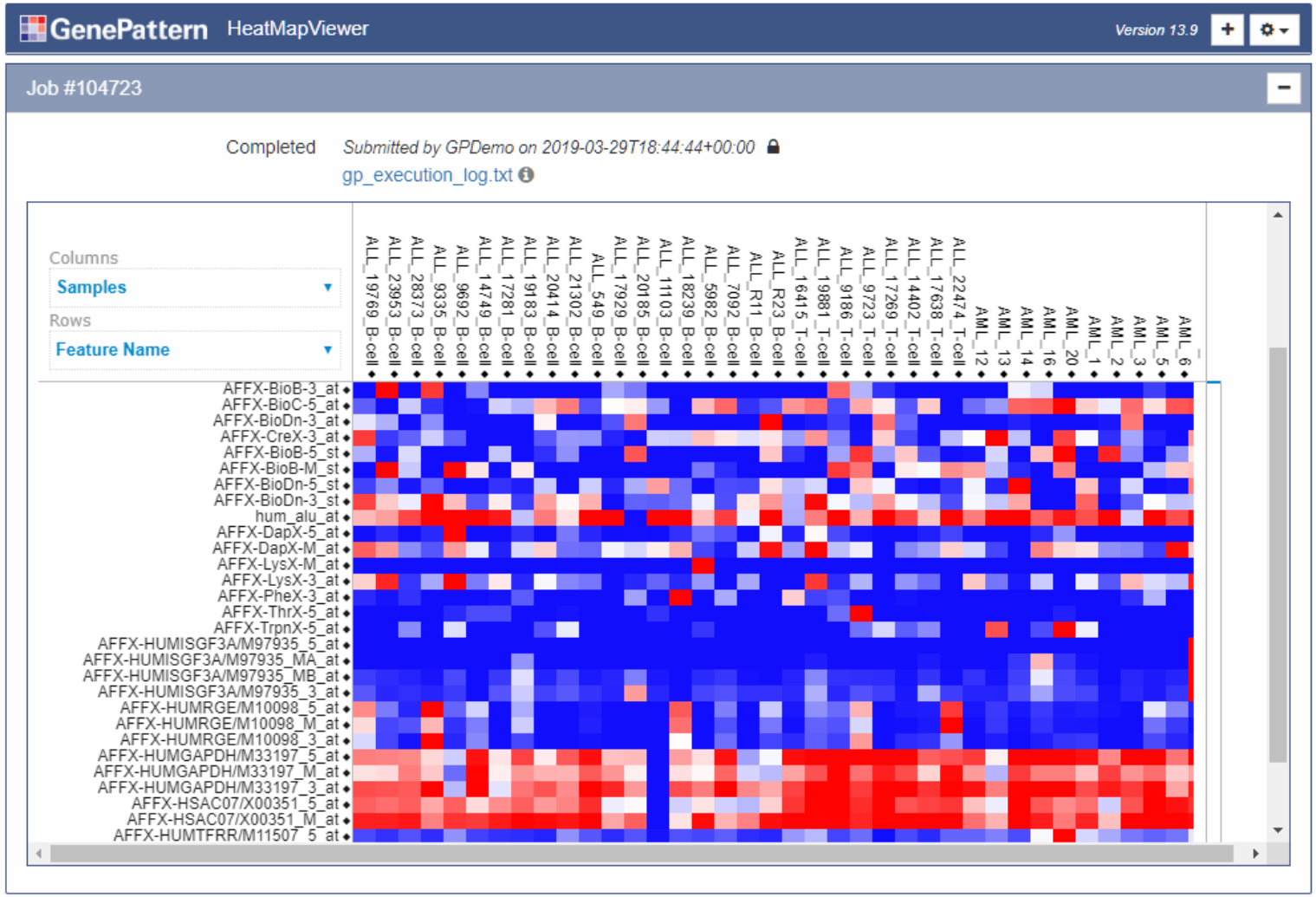



# Analysis Cells

- 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**



# Job/Result Cells

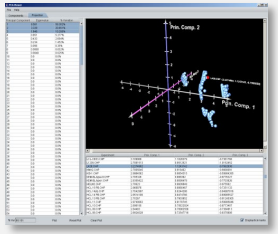


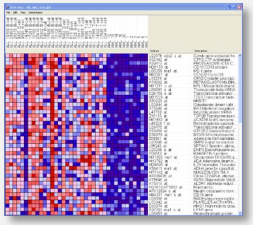
# The GenePattern Ecosystem: Architecture

## Module Repository

NMF | Cuffdiff | MuTect | ...

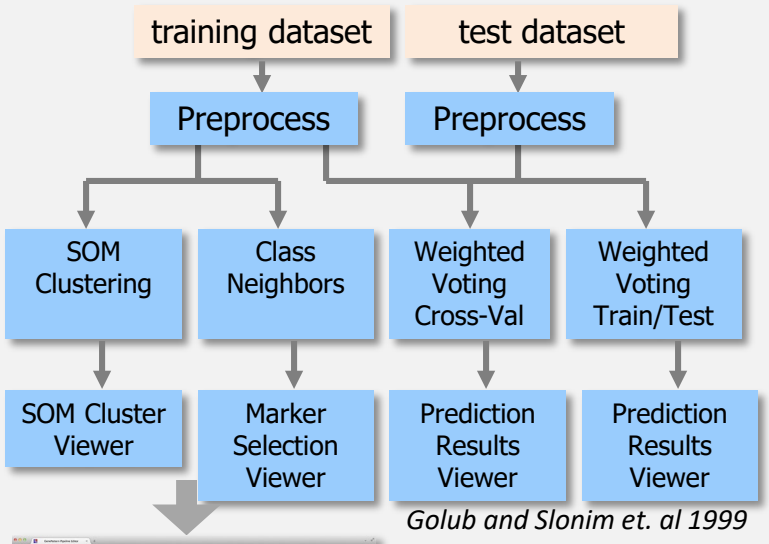
GISTIC | GSEA | Tophat | ...



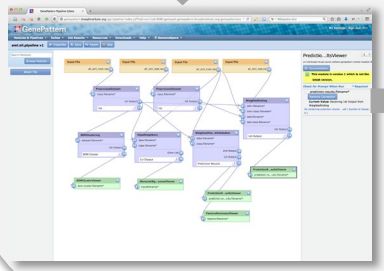


Hundreds of analysis and visualization tools

## Pipeline Environment




*Golub and Slonim et. al 1999*



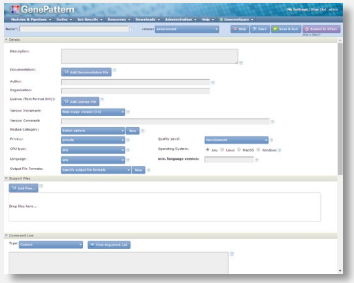
Golub.Slonim.1999.Nature.all.aml.pipeline.zip

Support for *in silico* reproducible research

## Analysis Engine

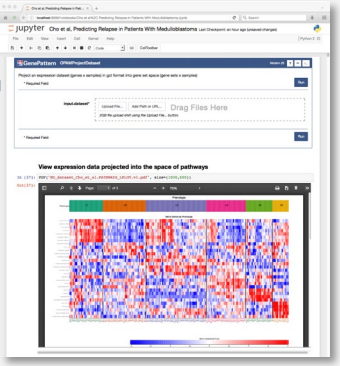
Record/replay analyses
Versioning of methods
Web service access

## Module Integrator

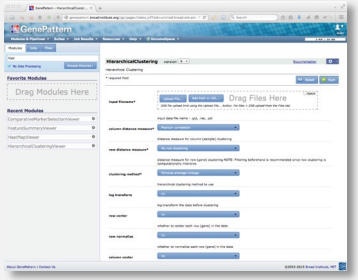


Easy addition of new tools

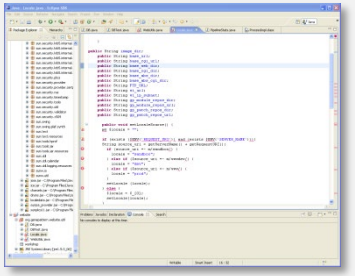
## Clients



Notebook



Web



Programming

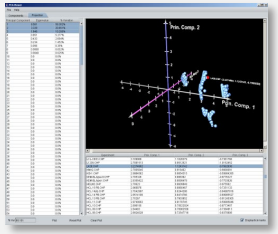
Access for all levels of user

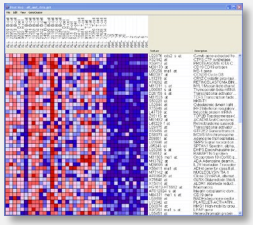
# The GenePattern Ecosystem: Architecture

## Tool Library

NMF | Cuffdiff | MuTect | ...

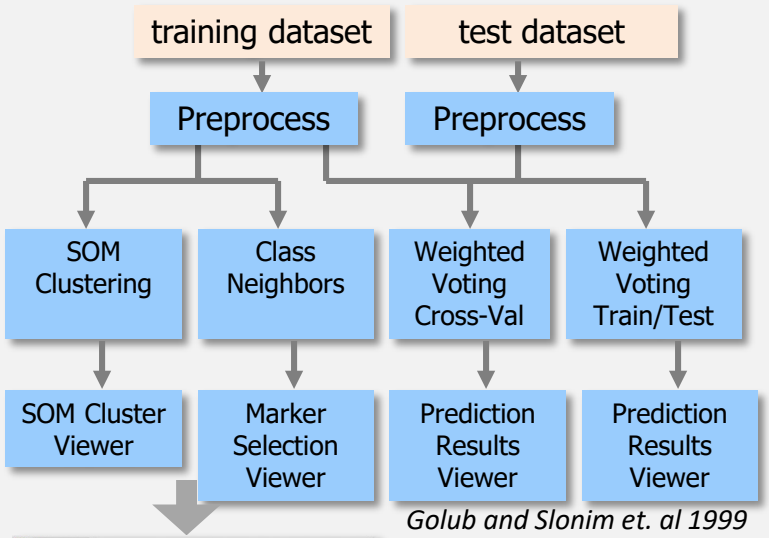
GISTIC | GSEA | Tophat | ...



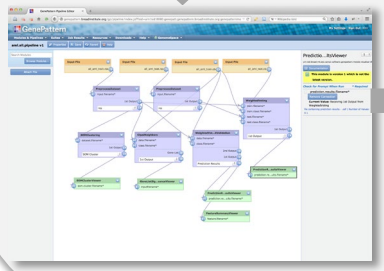


Hundreds of analysis and visualization tools

## Workflows




*Golub and Slonim et. al 1999*



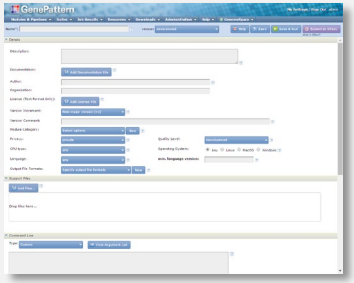
Golub.Slonim.1999.Nature.all.aml.pipeline.zip

Support for *in silico* reproducible research

## Analysis Engine

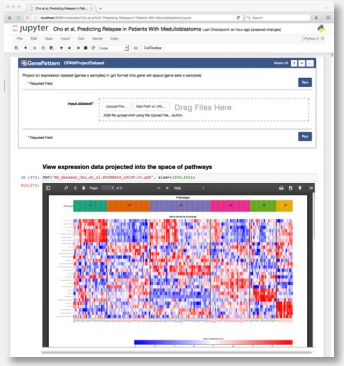
Record/replay analyses
Versioning of methods
Web service access

## Create your own tool

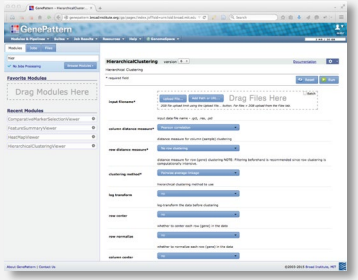


Easy addition of new tools

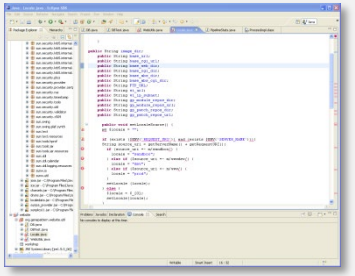
## Access Points



Notebook



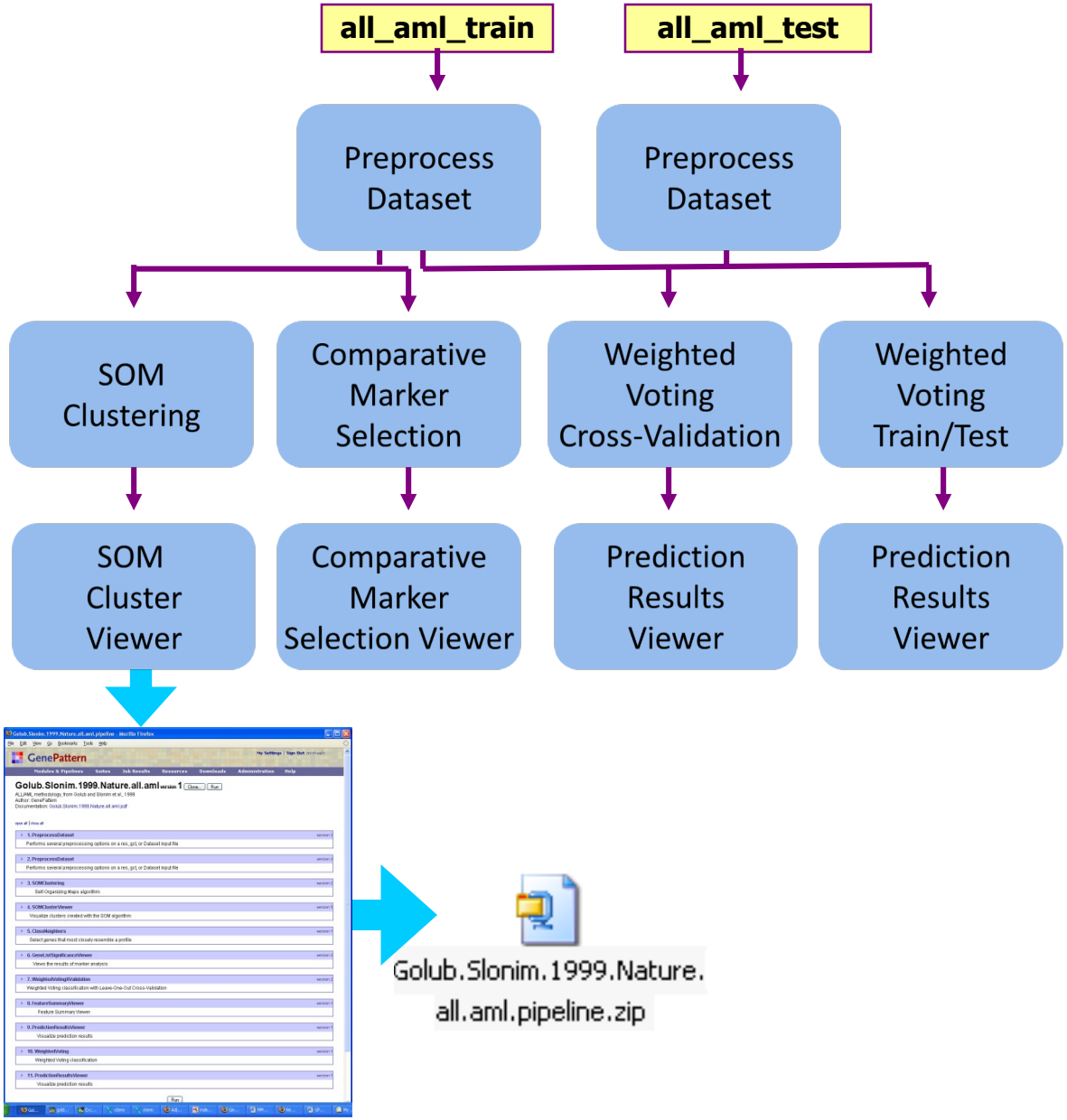
Web



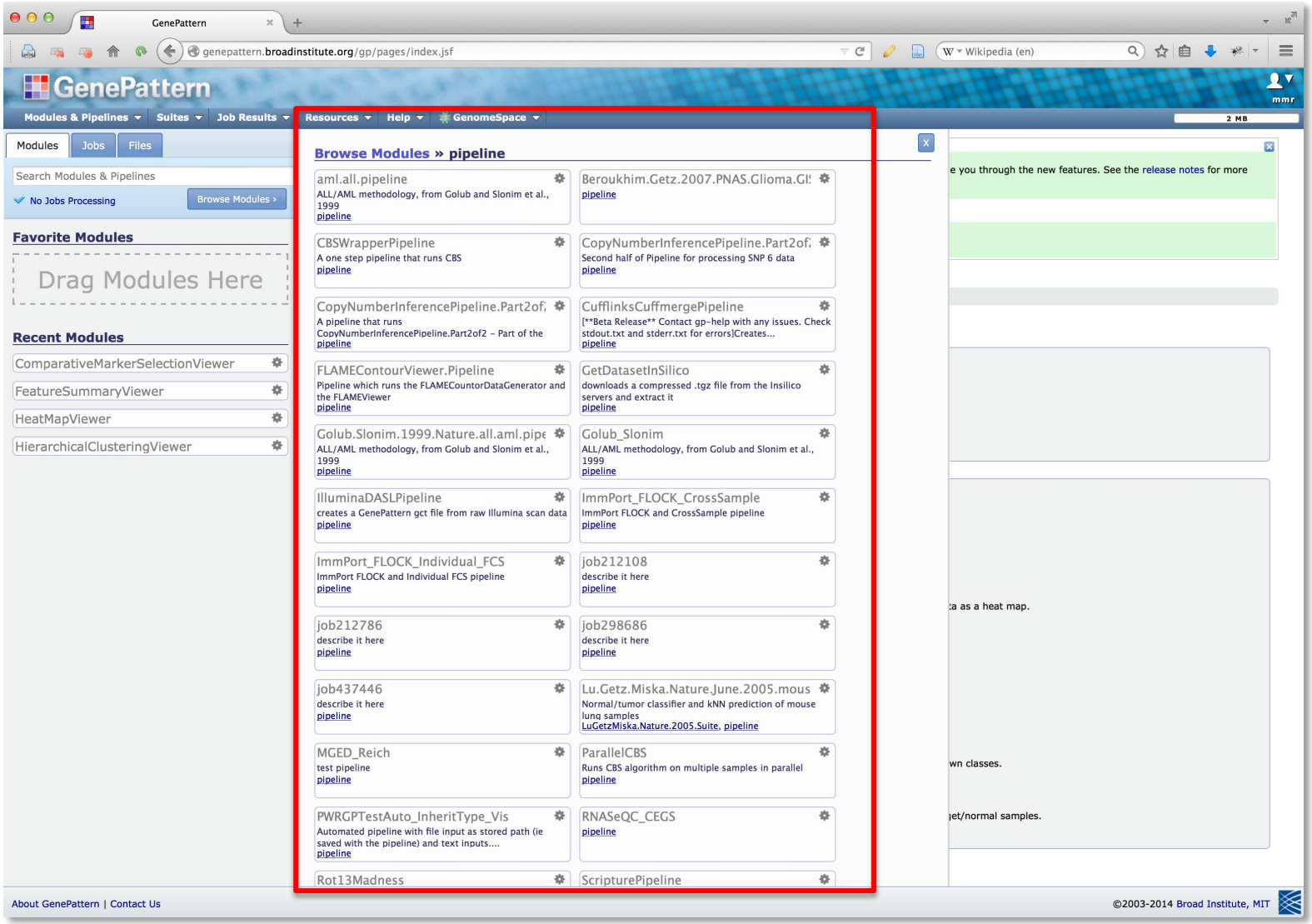
Programming

Access for all levels of user

# GenePattern vocabulary: Pipelines

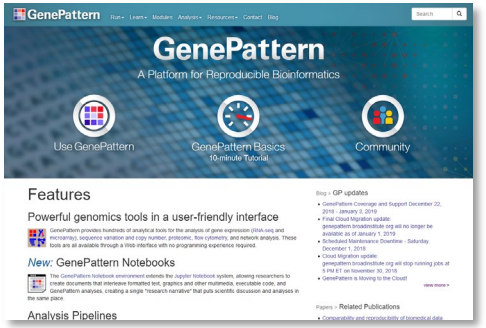


# Pipelines in GenePattern



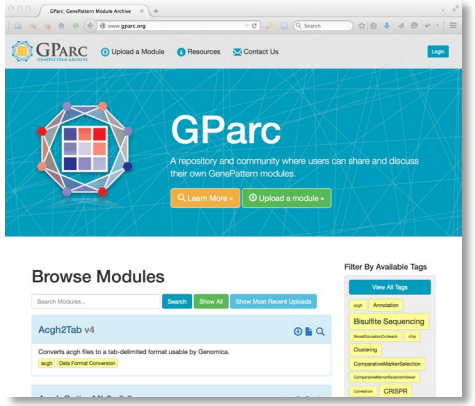
# Community Activity

www.genepattern.org



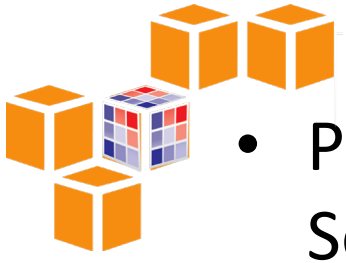

- Current version: 3.9.11 rc.4 b210 (2/2019)
- >50,000 registered users
- Open source, BSD-style license
- New Public server runs ~2,500 analyses/week

www.gparc.org

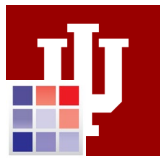



- GParc: GenePattern community repository
  - ~100 community-contributed methods
  - CRISPR analysis
  - Bisulfite sequencing
  - Flow cytometry
  - RNAi screens

# 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

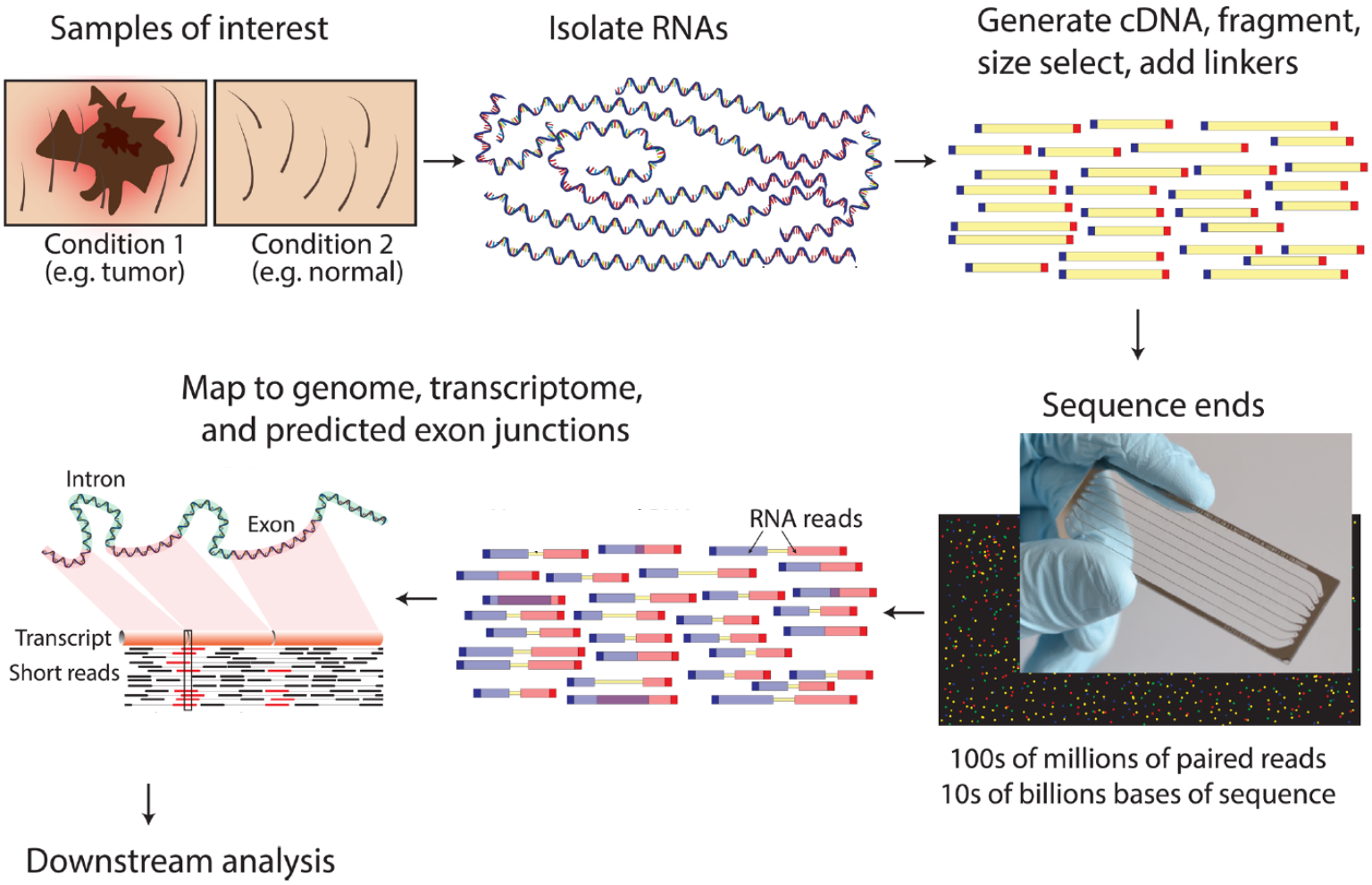
# Supported Platforms

| 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.




Griffith et al. (2015) Doi:10.137/journal.pcbi.1004393.g002

# 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
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).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:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

## Illumina sequence identifiers [ edit ]

Sequences from the Illumina software use a systematic identifier:

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

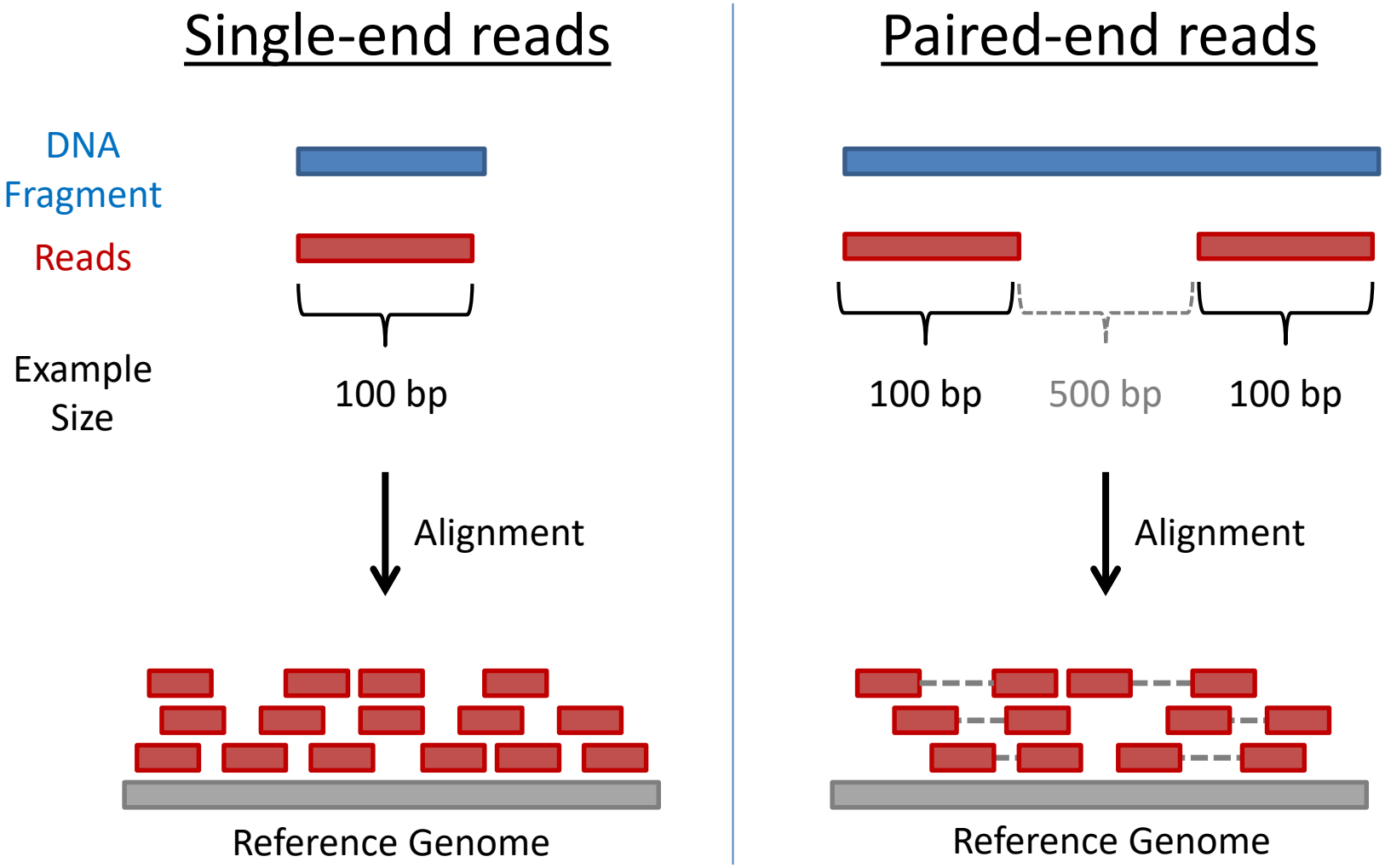
## NCBI Sequence Read Archive [ edit ]

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
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

https://en.wikipedia.org/wiki/FASTQ_format

# RNA-seq read types




**TODAY:** We will use paired-end reads.

# Importing and Converting Data

## Importing Data

SraToFastQ
GEOImporter
MAGEMLImportViewer
MAGETABImportViewer

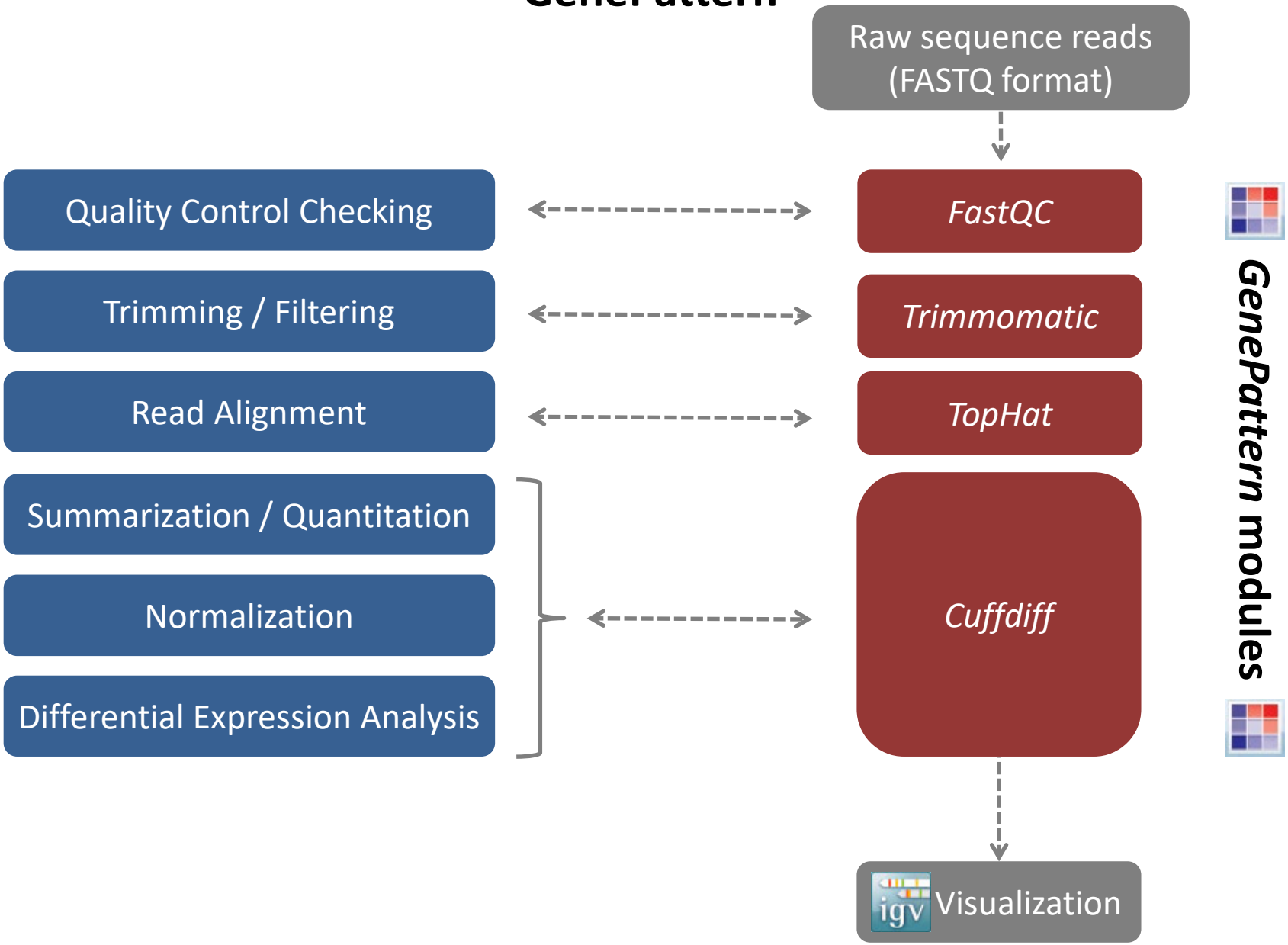
## Converting data

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



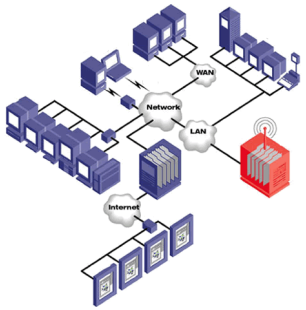
# Running Analyses

# GenePattern configurations

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

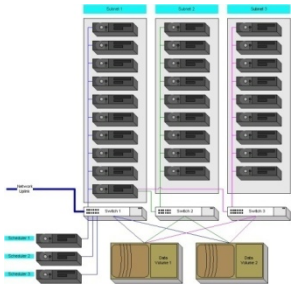
**Standalone**

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




**Client/Server**

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



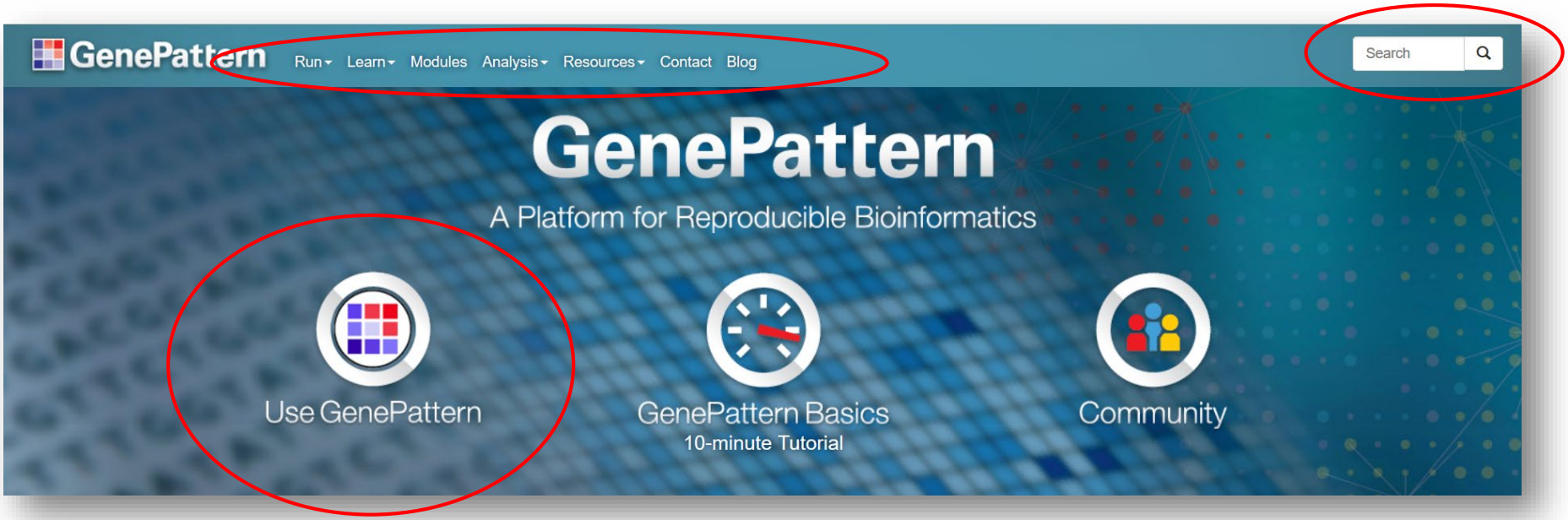
**Cluster**

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

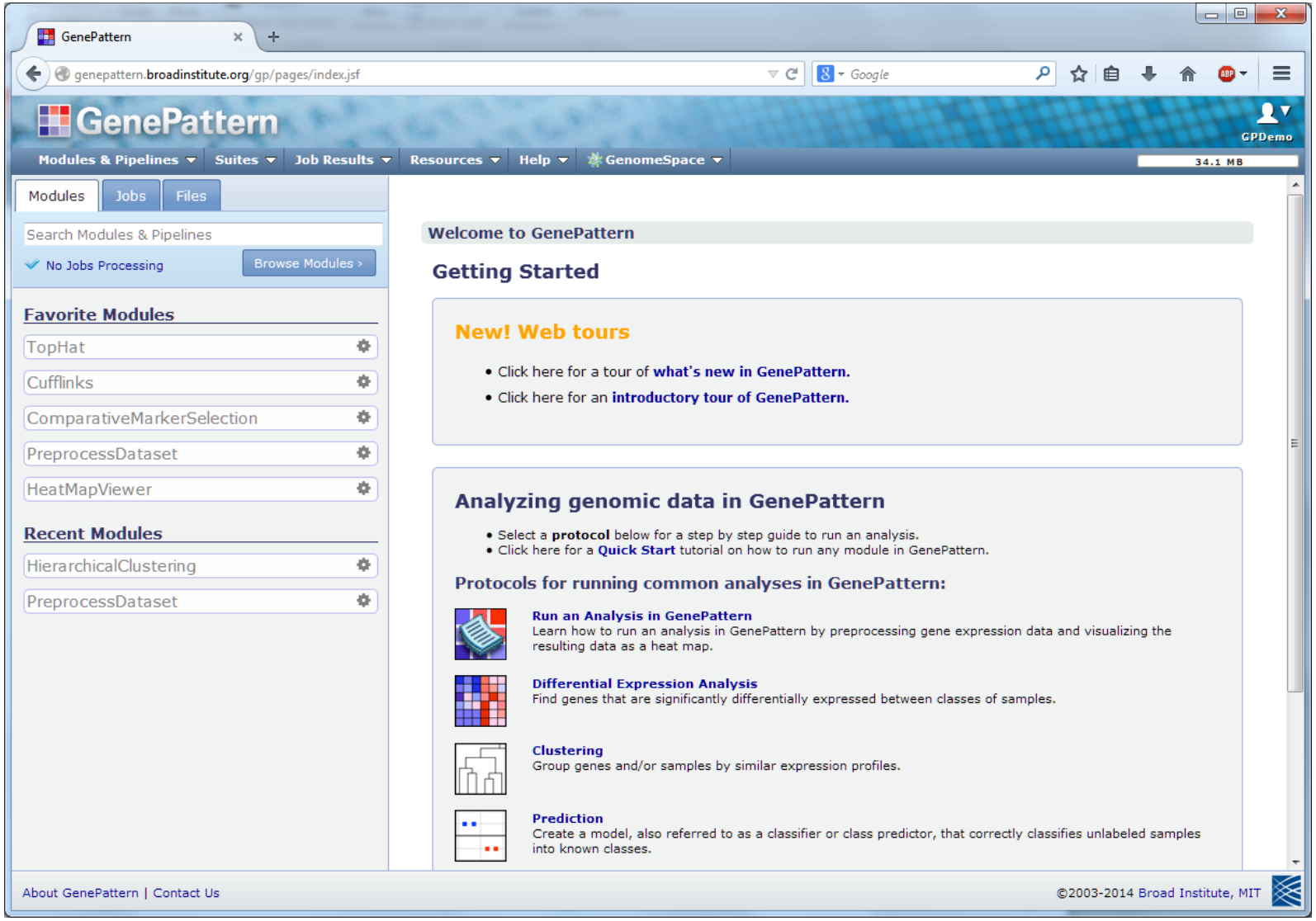
Note: GenePattern server & client are separate software

# GenePattern Website

http://genepattern.org



# GenePattern UI – main page

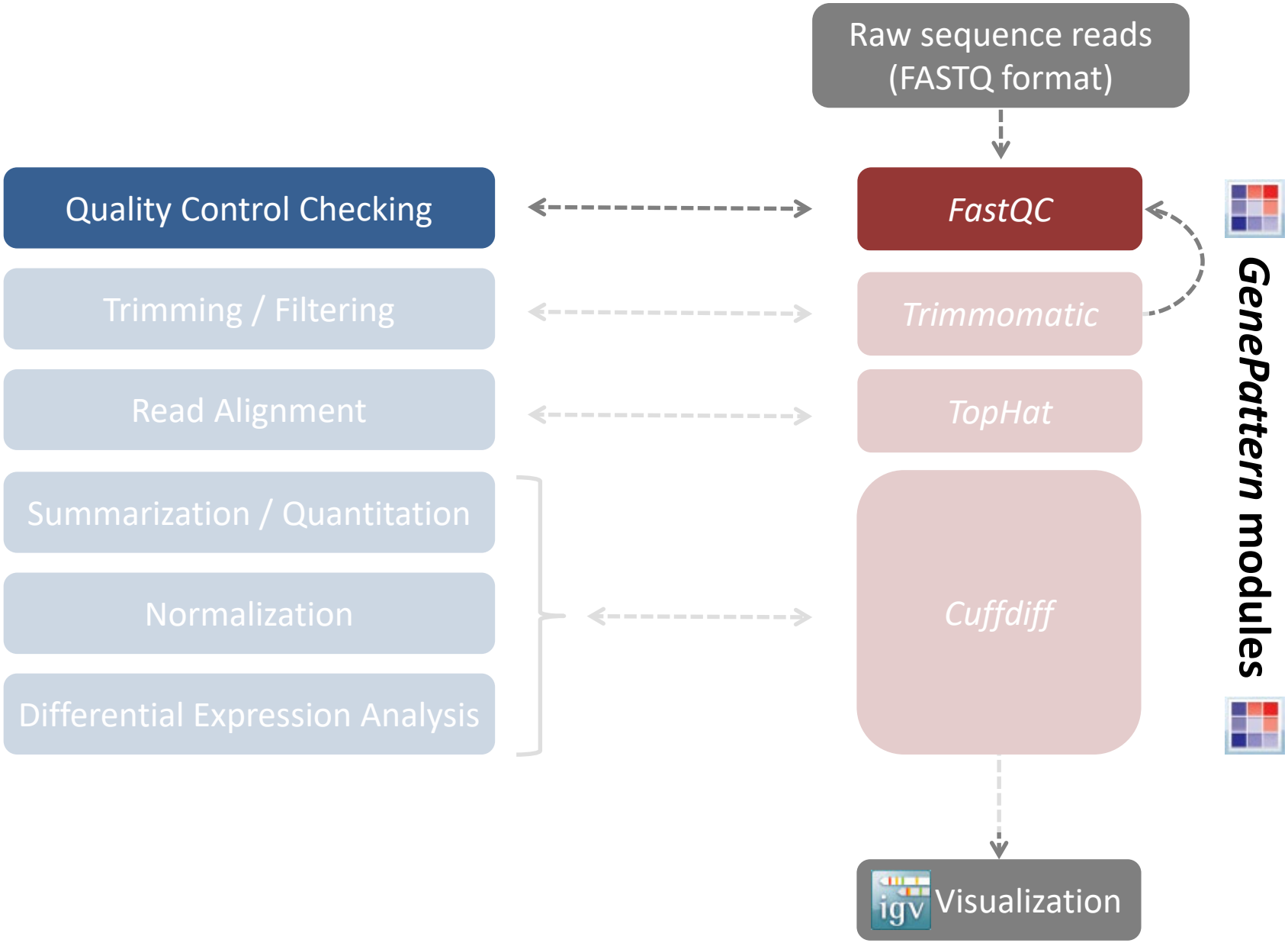


# Running analyses

- 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 β2-agonist (i.e. Albuterol, 1μM for 18h)
  - 3) treatment with a glucocorticosteroid (i.e. Dexamethasone (Dex), 1μM for 18h)
  - 4) simultaneous treatment with a β2-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 |

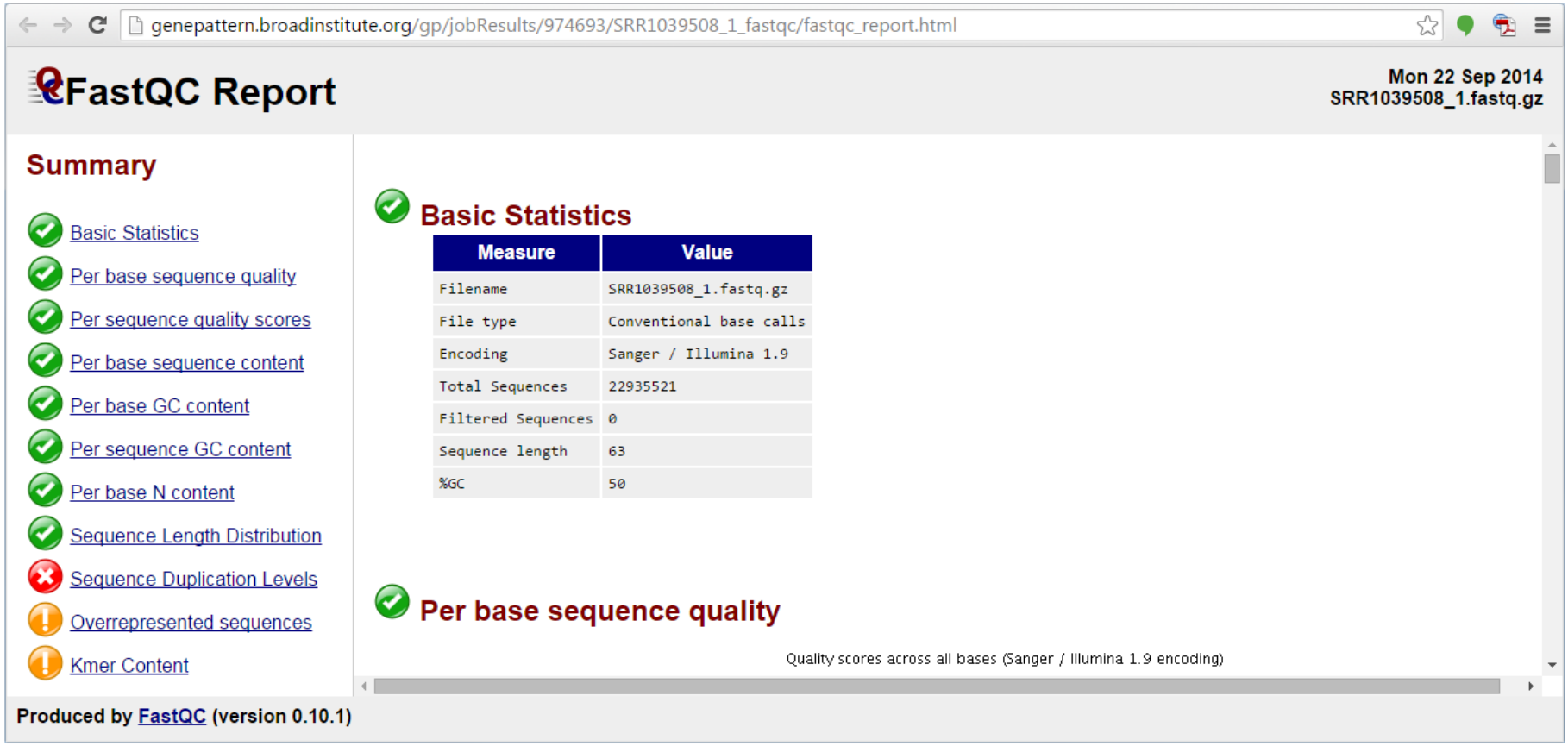
# ***FastQC*** results

FA

# *FastQC* results

FA

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

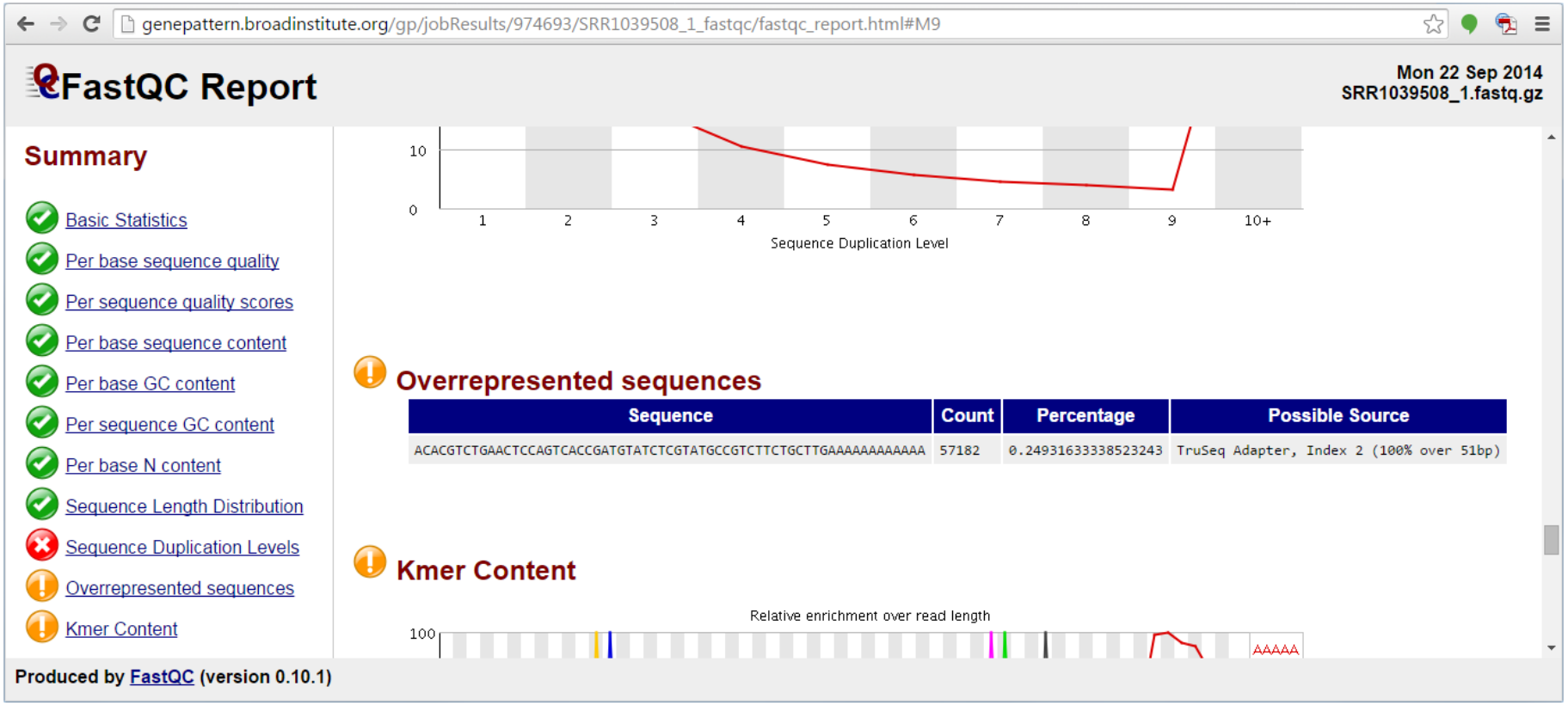



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

# *FastQC* results

FA

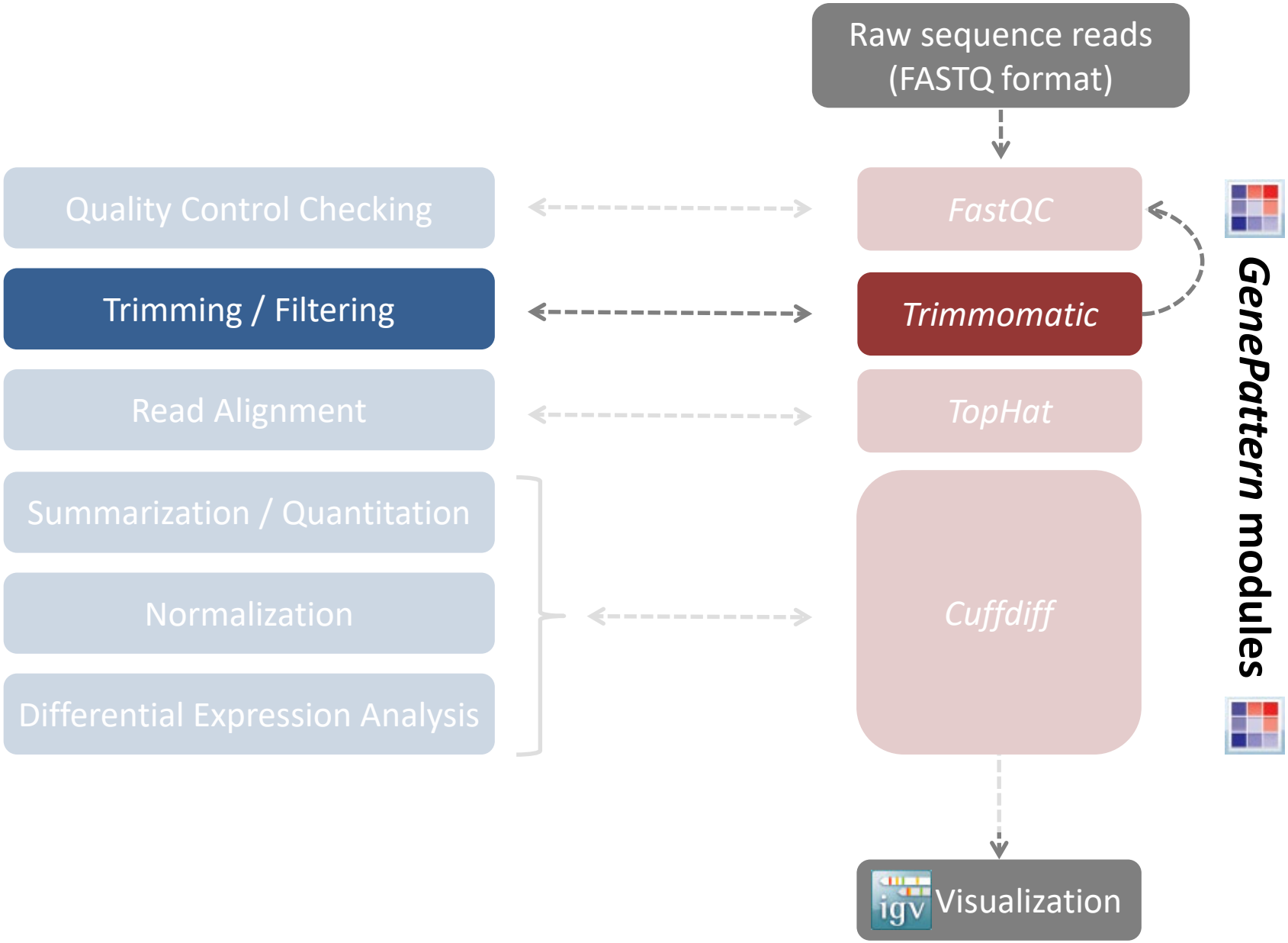
For example, our file has a small amount of adapter contamination, which is found under the ***Overrepresented Sequences*** section.




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

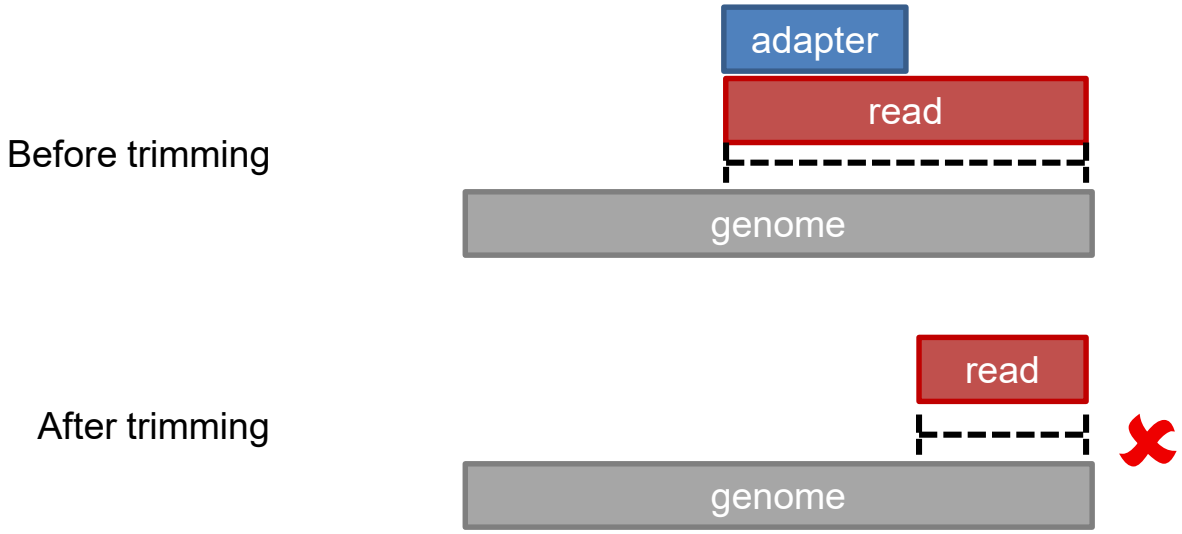
# RNA-Seq in GenePattern

# 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 ✗

http://www.usadellab.org/cms/?page=trimmomatic

# Run *Trimmomatic* on paired FASTQ files FA

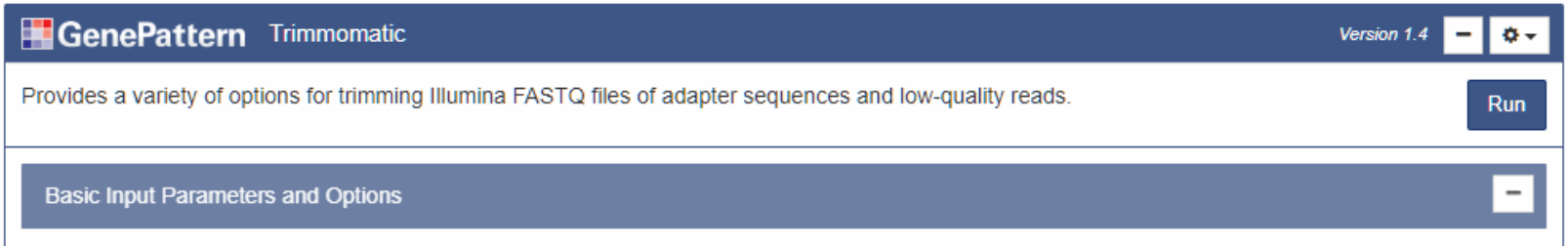
### Run Trimmomatic on Paired FASTQ files

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.

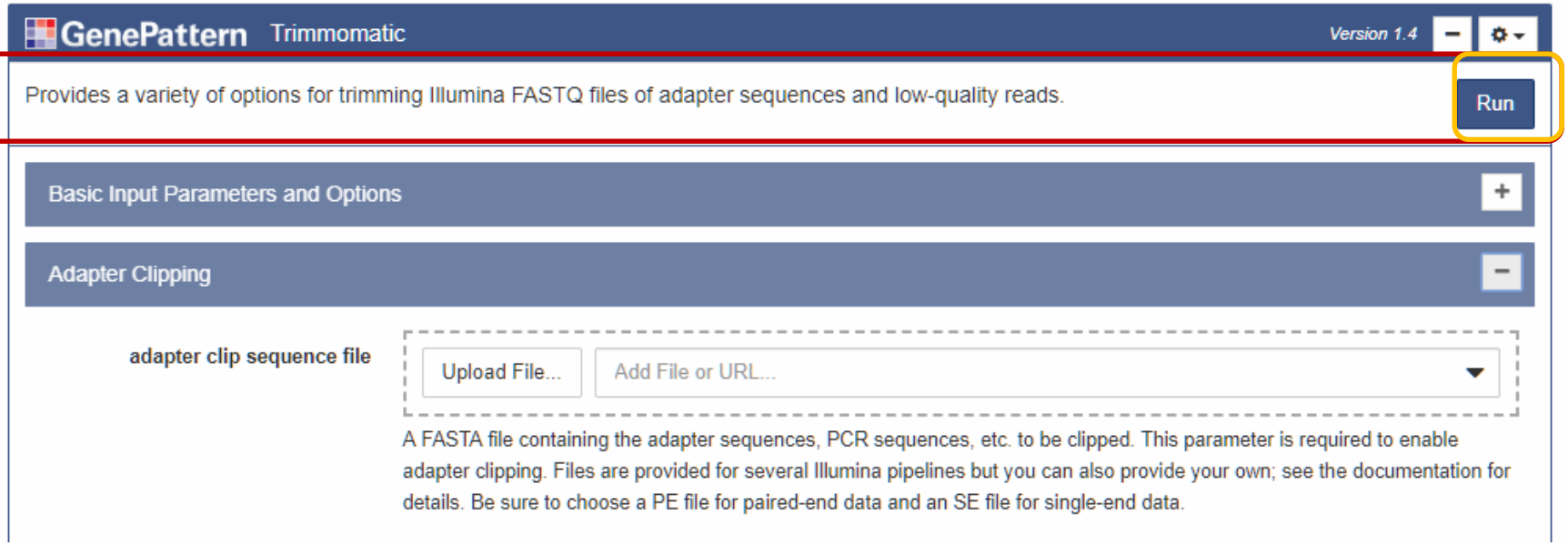


# Run *Trimmomatic* on paired FASTQ files FA

- 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.

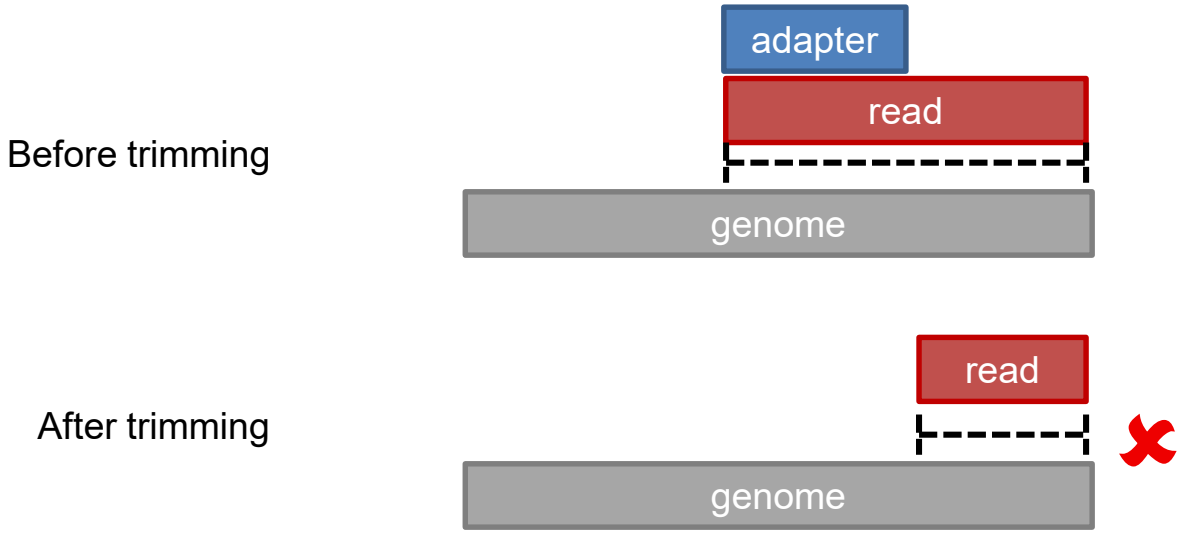


# *Trimmomatic* results

FA

# *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:**

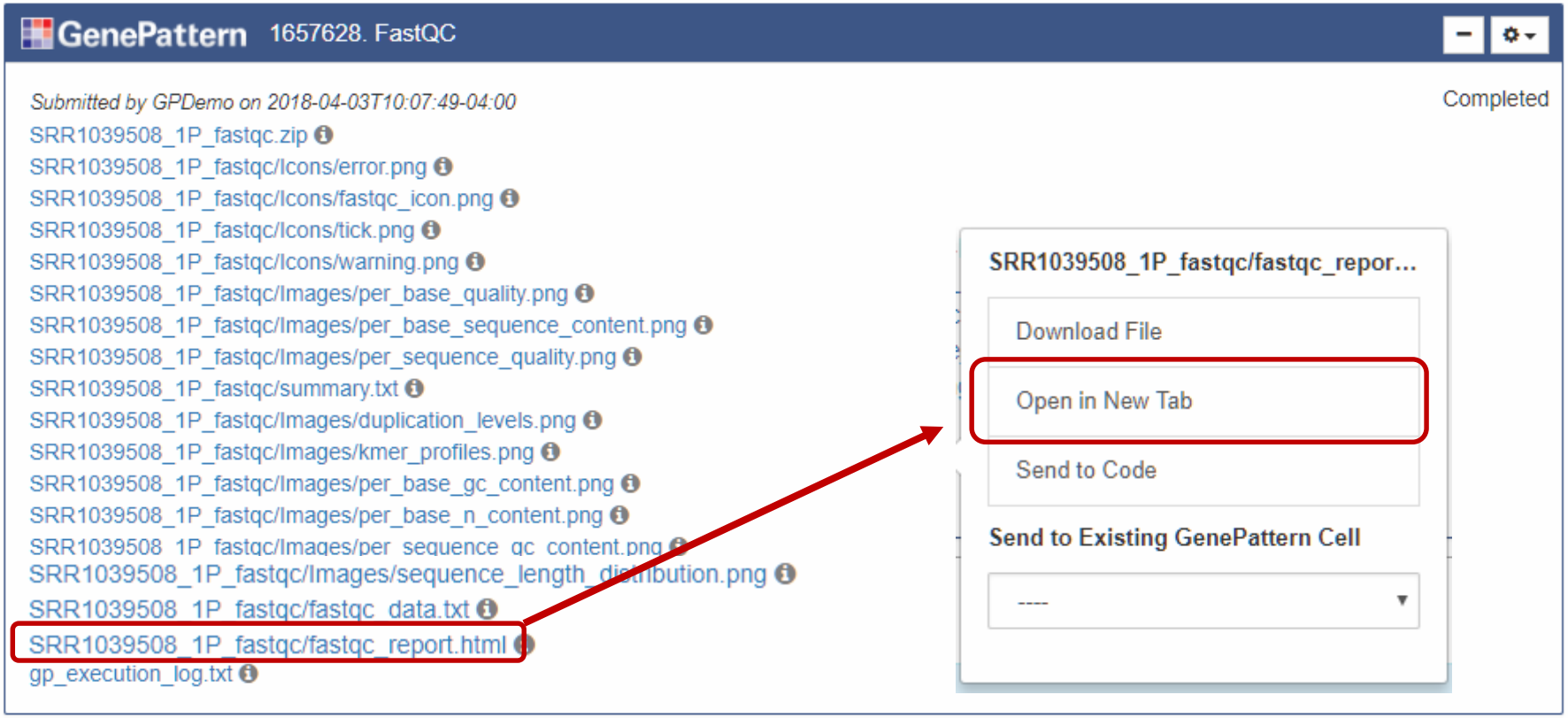
**_1P** or **_2P** are forward and reverse reads which both passed quality checks ✓

**_1U** or **_2U** are “unpaired” reads in which one of the pair did not pass ✗

http://www.usadellab.org/cms/?page=trimmomatic

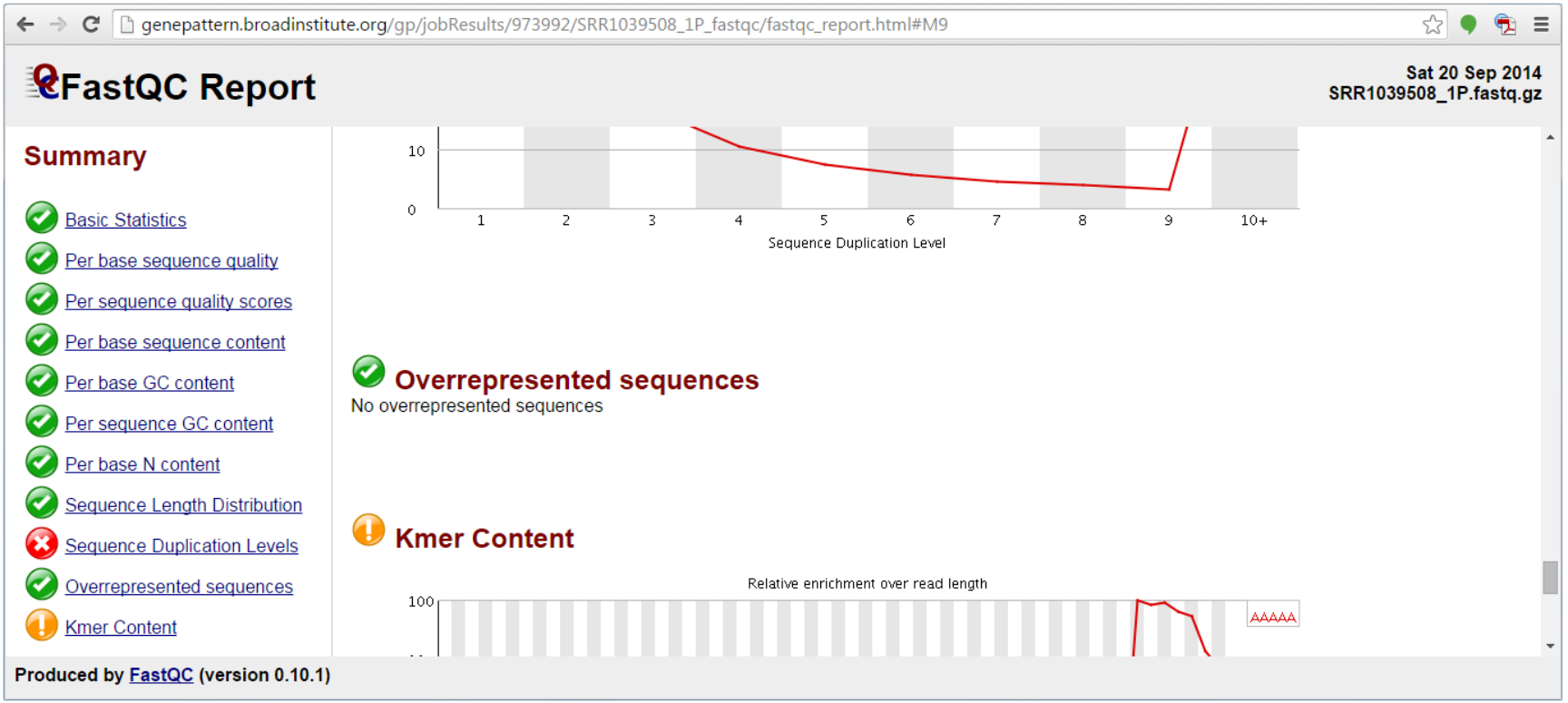
# ***FastQC*** re-run results

FA

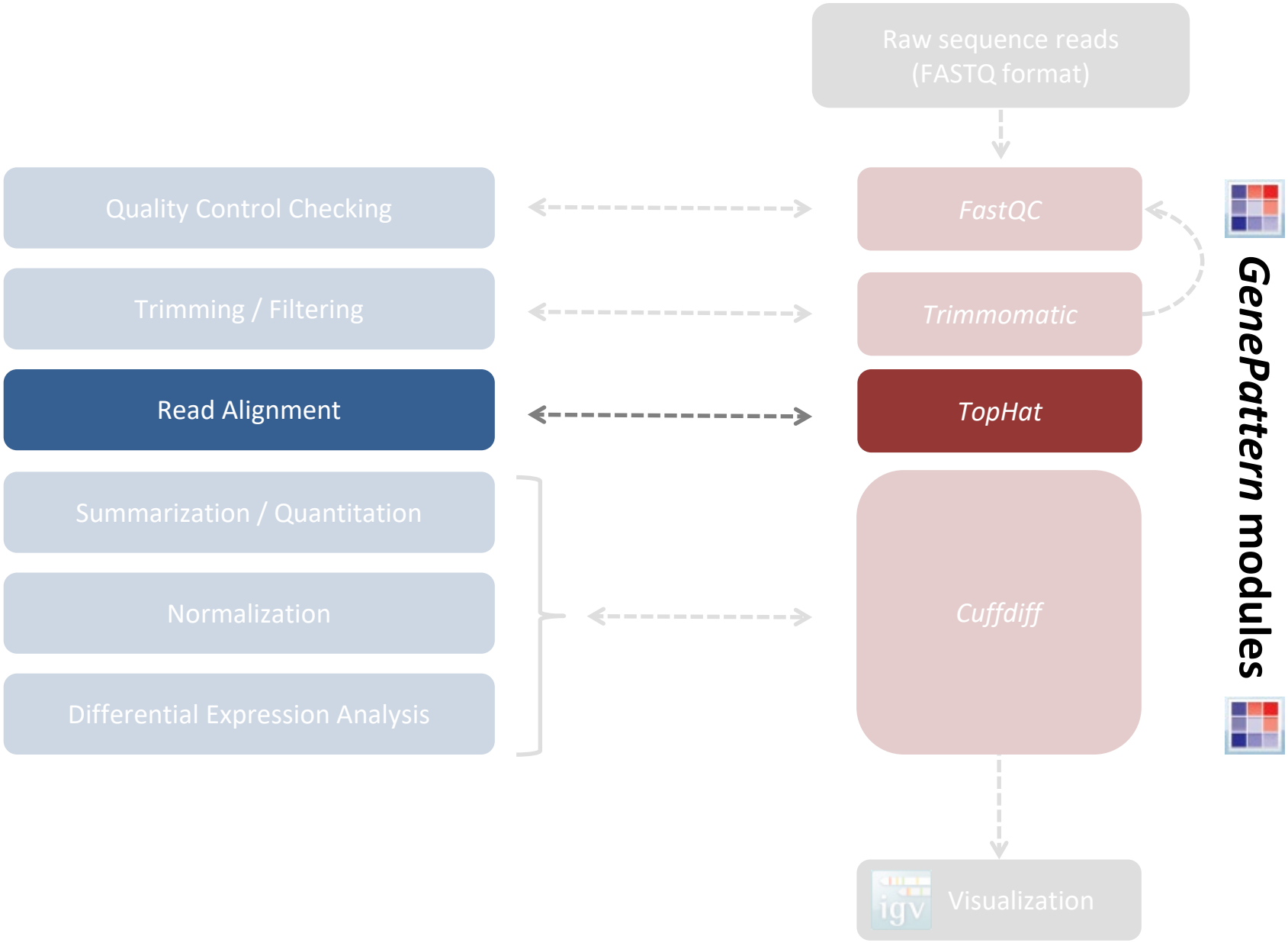


# Re-run ***FastQC*** to see improvement

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



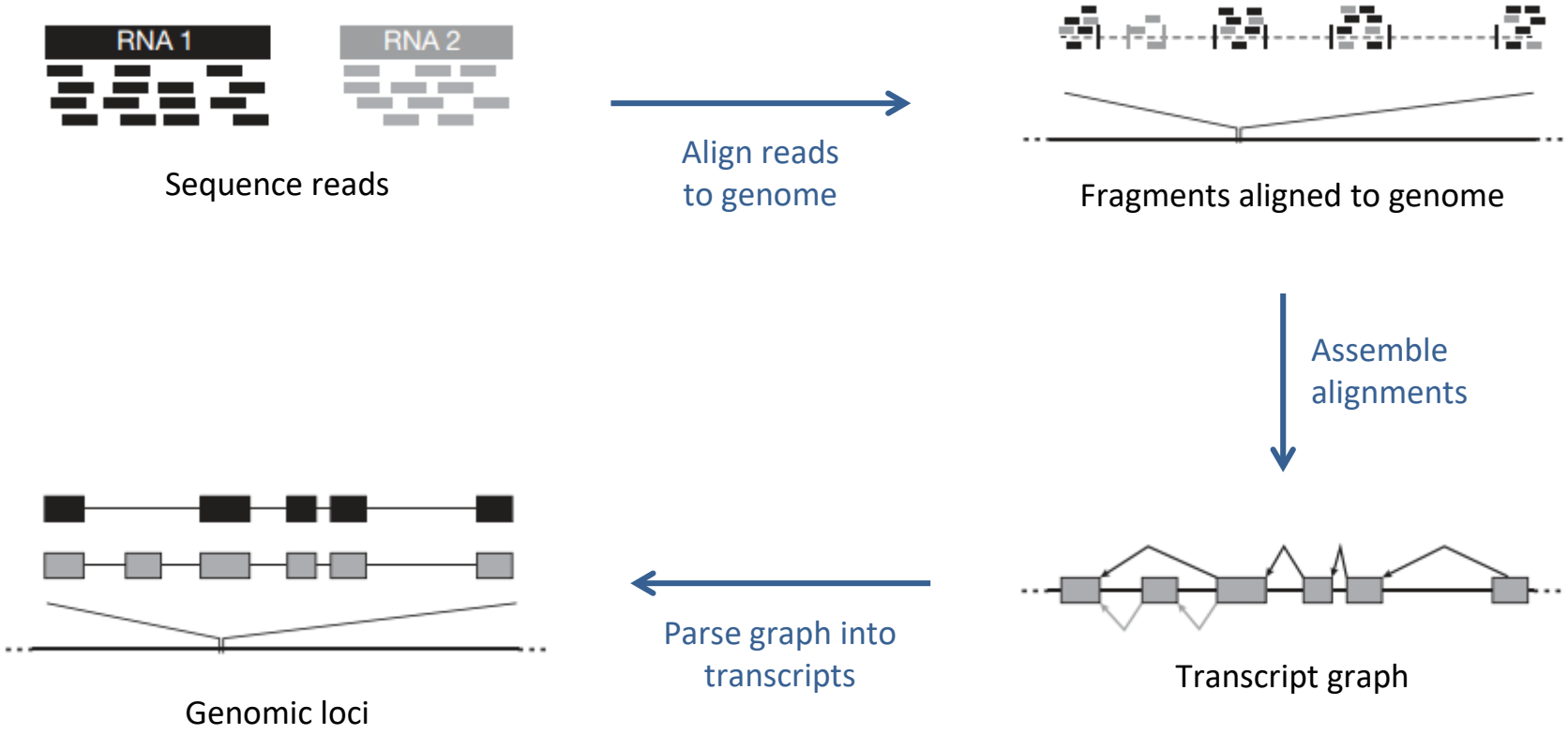
# RNA-Seq Differential Expression Analysis Workflow

**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.




Garber et al. (2011) Doi:10.1038/nmeth.1613

# 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

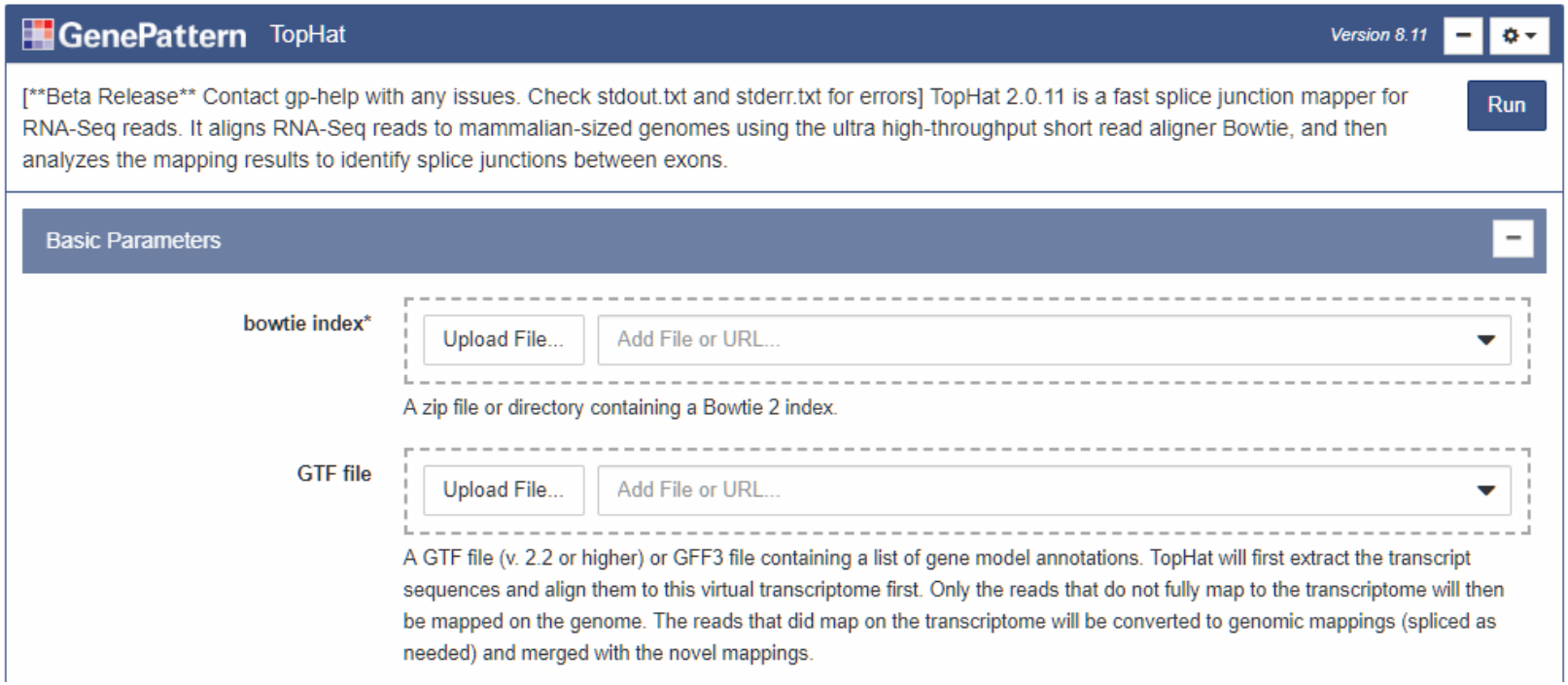
**FA**

### 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.



# Run ***TopHat*** to create a transcriptome index

**FA**


Index = the whole transcriptome_index folder:

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

# Run ***TopHat*** to align reads

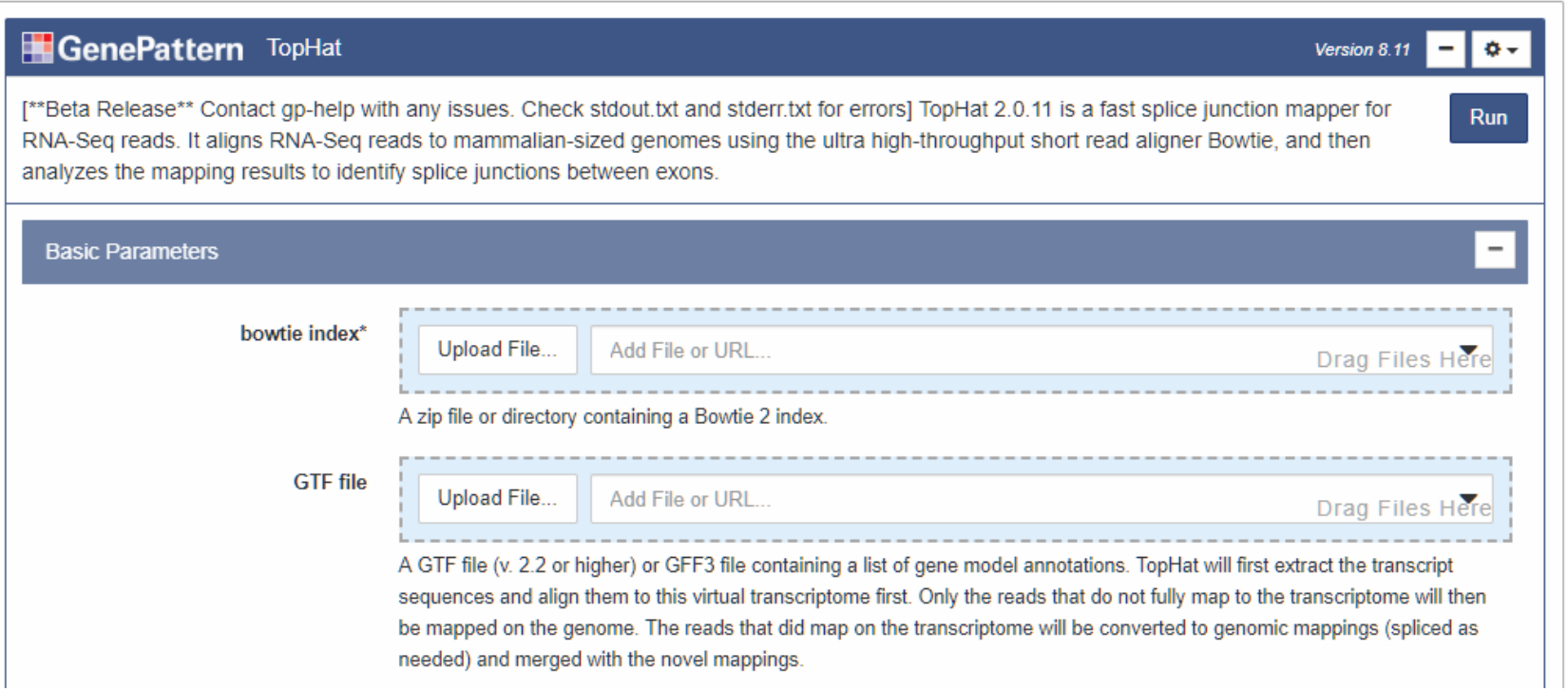
FA

## 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 ouput.
- Click the input field for **reads pair 2** and select **SRR1039508_2P.fastq.gz**from the Trimmomatic ouput.
- 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.



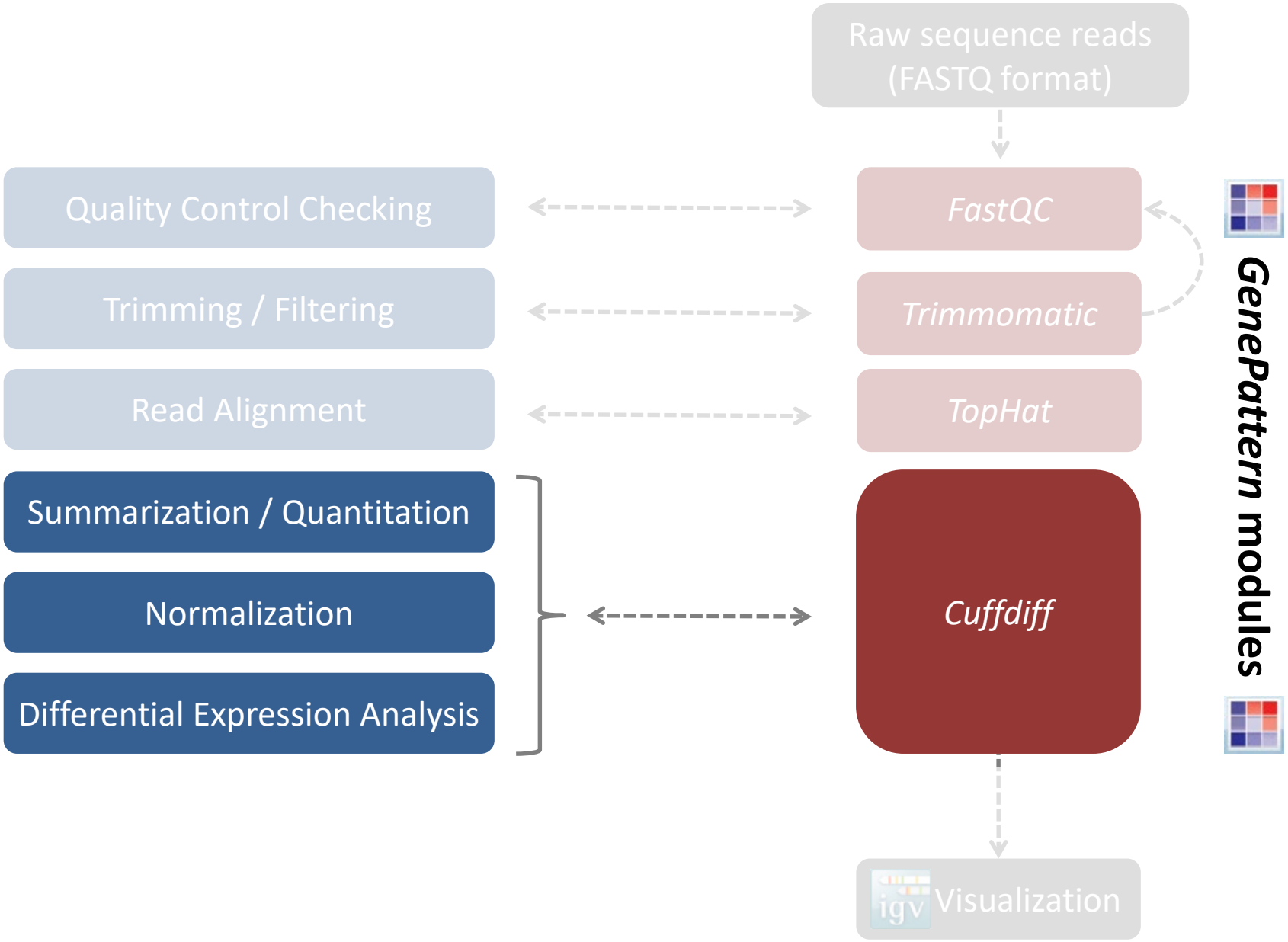
# ***TopHat*** results

FA

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).

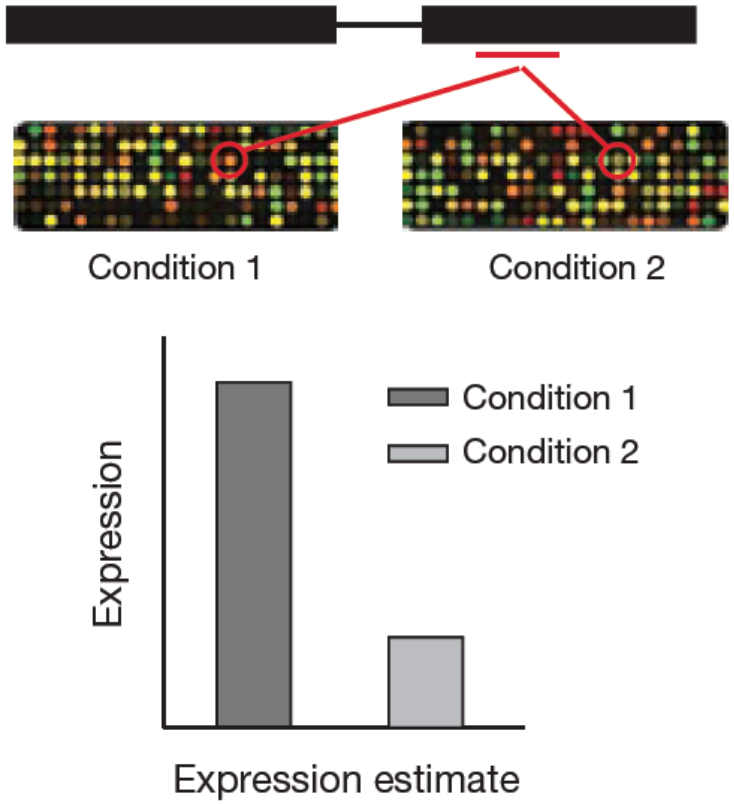
# RNA-Seq Differential Expression Analysis

# RNA-Seq Differential Expression Analysis Workflow

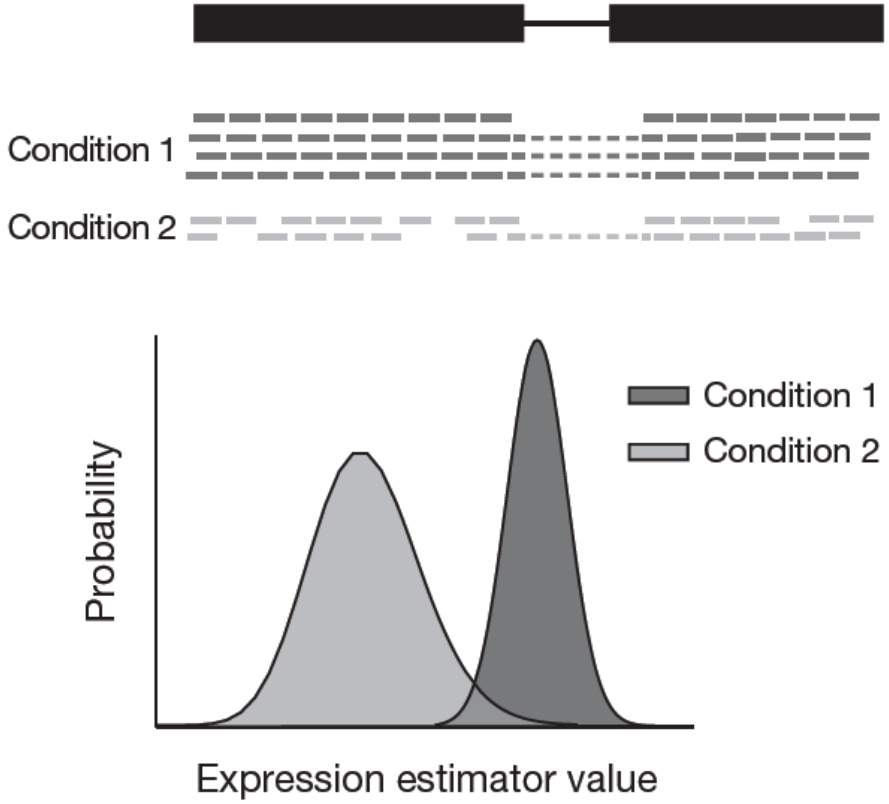


# Differential gene expression analysis

## Microarray




## RNA-Seq




Garber et al. (2011) Doi:10.1038/nmeth.1613

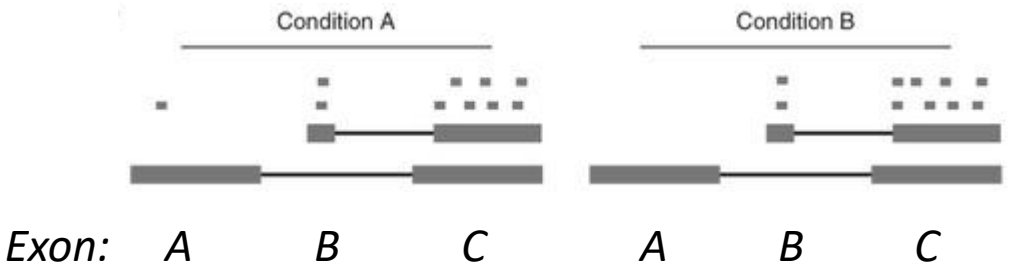
# Differential gene expression analysis

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

**Condition A**

10




**Condition B**

10

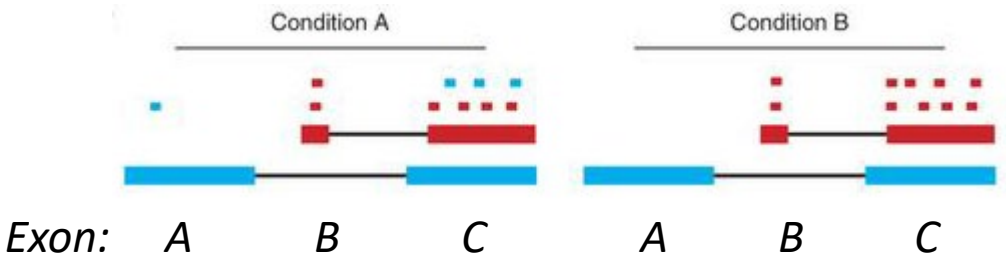
# Differential gene expression analysis

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

**Condition A**

**4** *vs* **6**



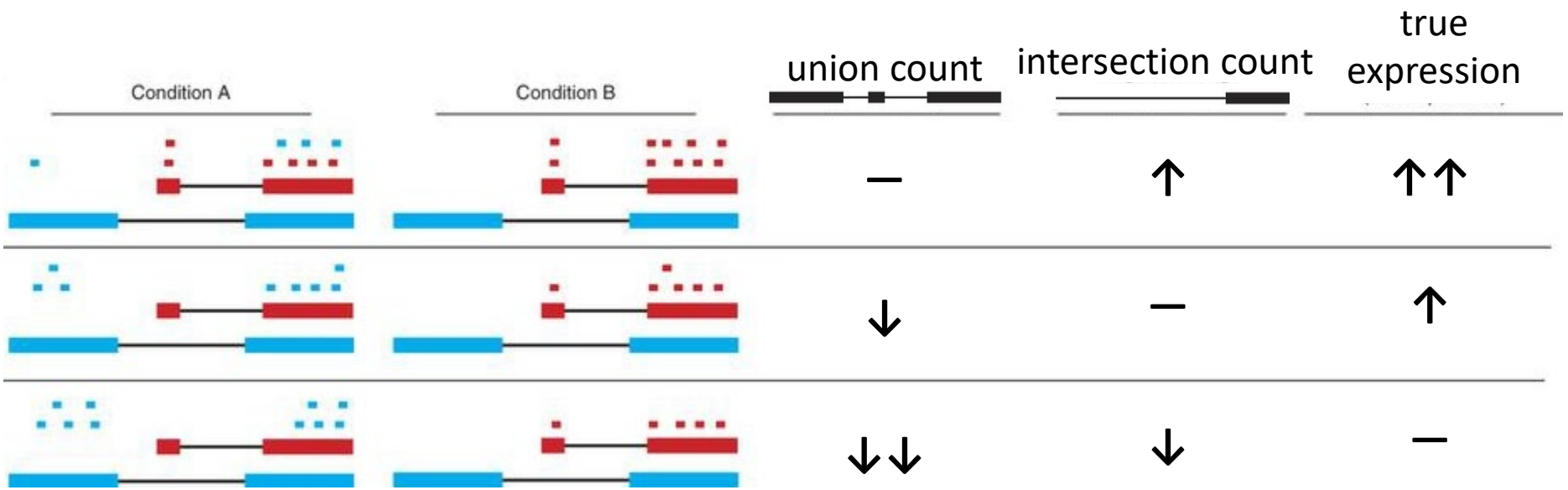

**Condition B**

**10**

# 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

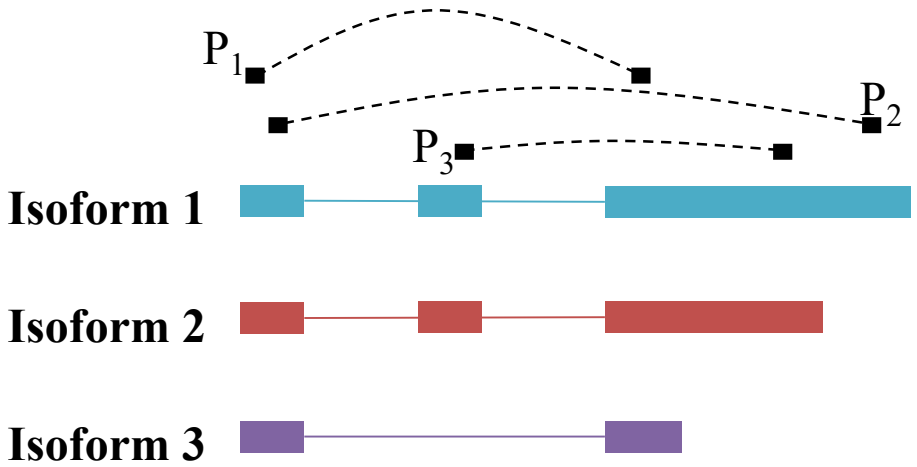


# Differential gene expression analysis

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

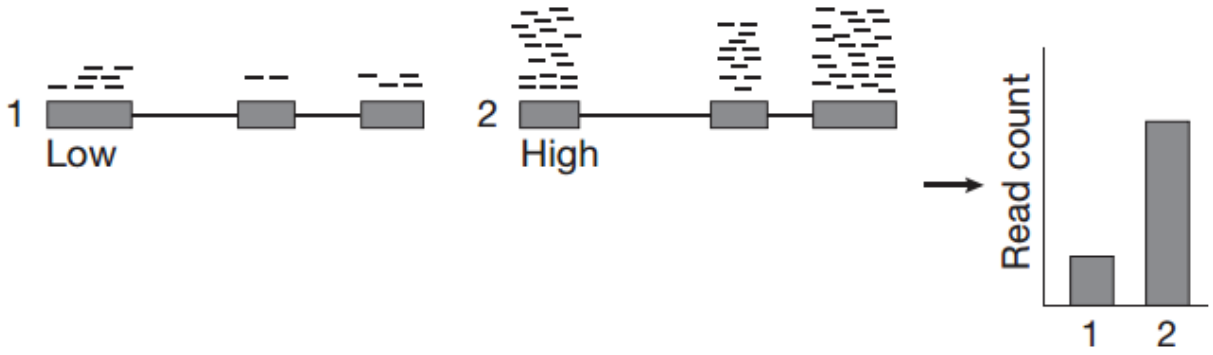
**Isoform deconvolution:** estimate the expression levels of a gene's alternative isoforms (transcripts) and then sum those transcript-level estimates to derive gene-level expression estimates.

**Paired ends** increase isoform deconvolution confidence.



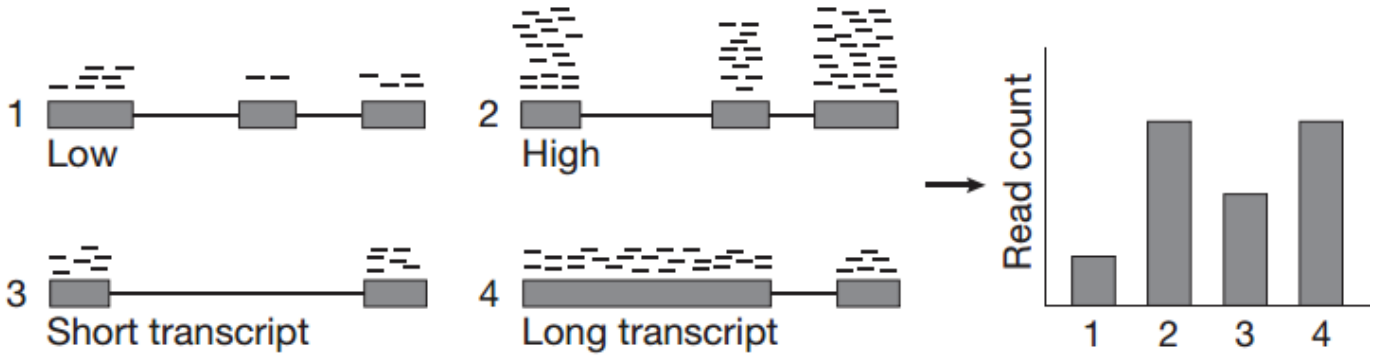

$P_1$ is from $I_1$, $I_2$ or $I_3$.
$P_2$ is from $I_1$.
$P_3$ is from $I_1$ or $I_2$.

# How to quantify expression from reads



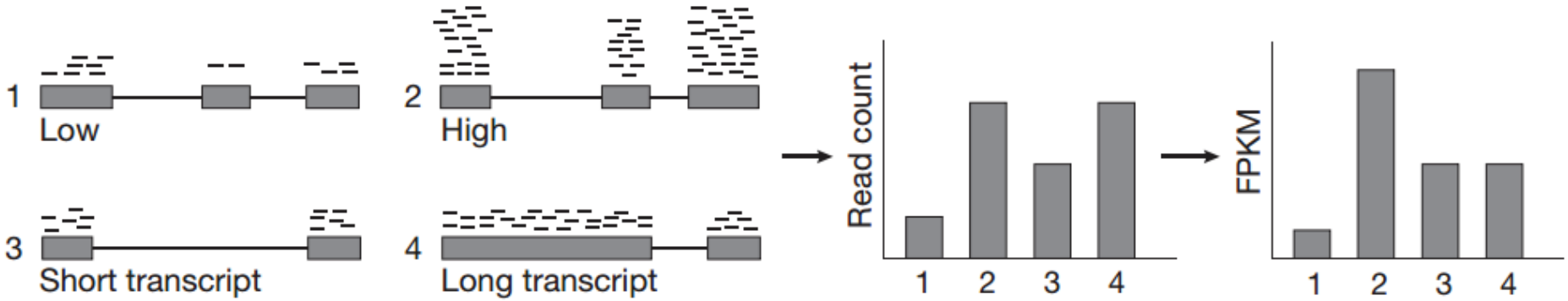

Garber et al. (2011) Doi:10.1038/nmeth.1613

# How to quantify expression from reads




Garber et al. (2011) Doi:10.1038/nmeth.1613

# How to quantify expression from reads




**F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads (**FPKM**), a count of mapped fragments normalized to transcript length.

Garber et al. (2011) Doi:10.1038/nmeth.1613

# Exercise

# Run ***Cuffdiff*** to differentiate phenotypes

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

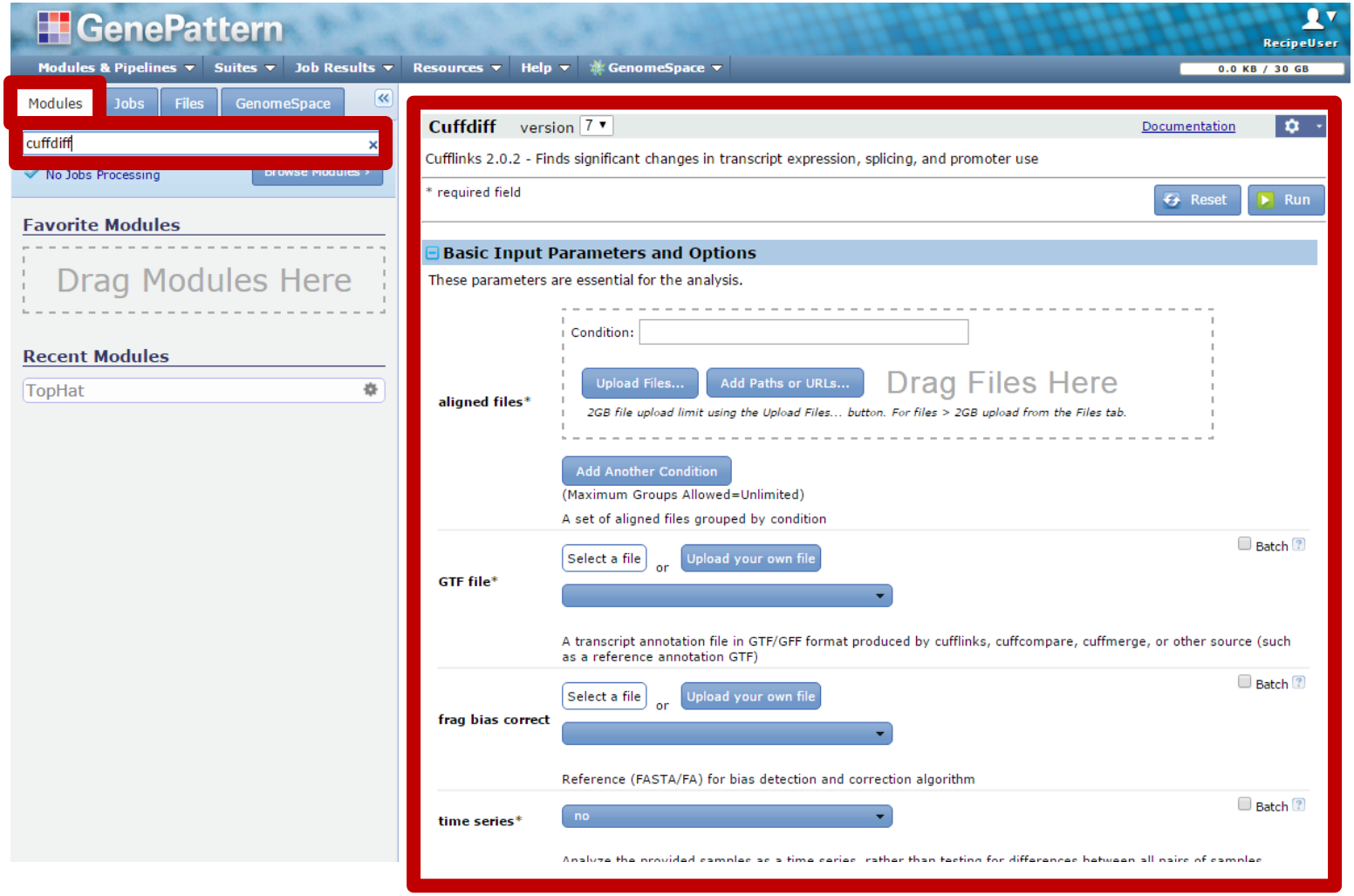
First, we group aligned reads from ***TopHat*** condition (**untreated** vs. **dex**).

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

# Run ***Cuffdiff*** to differentiate phenotypes

FA

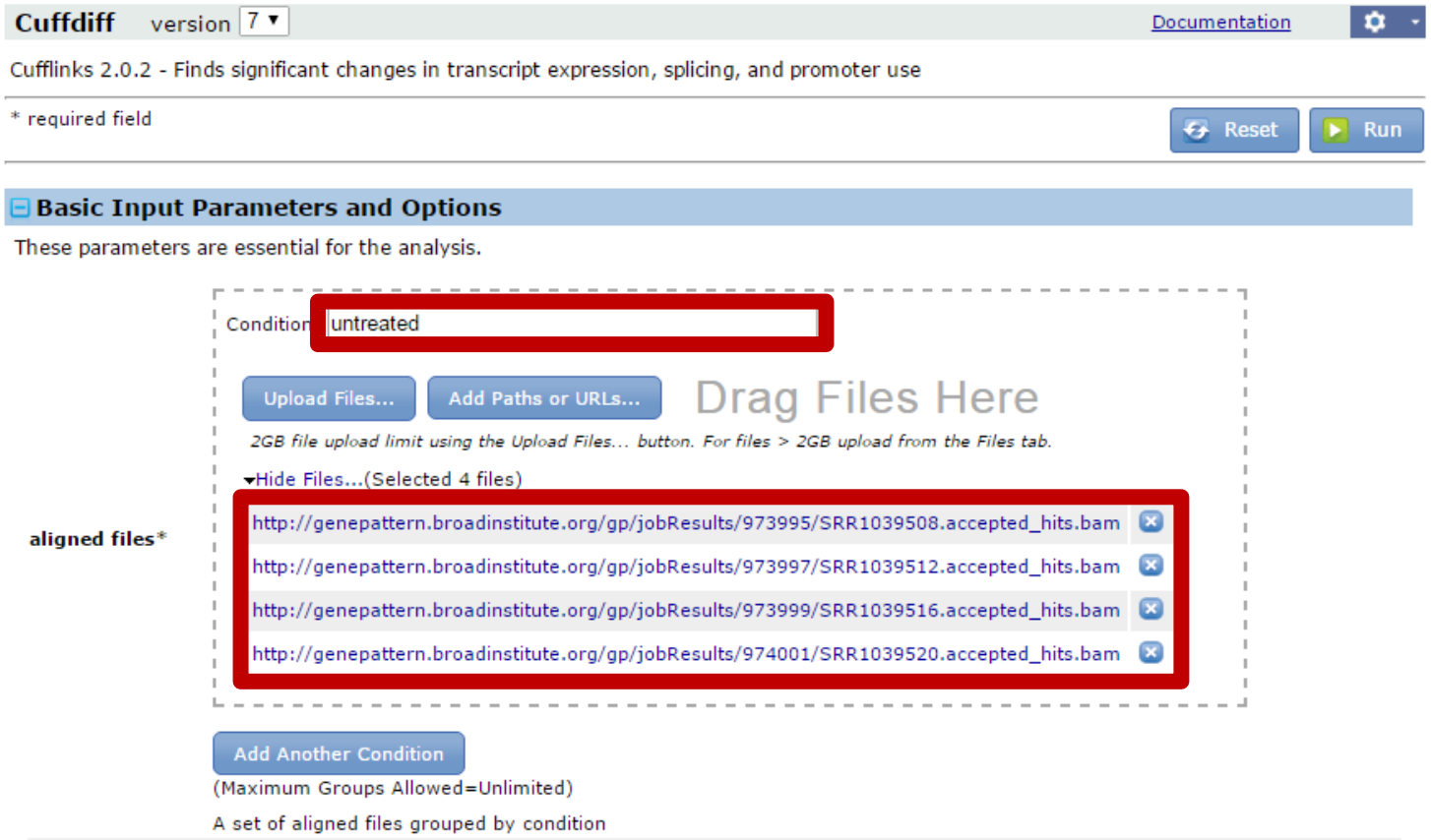
Search for “Cuffdiff” under the Modules panel in GenePattern.



# Run ***Cuffdiff*** to differentiate phenotypes

FA

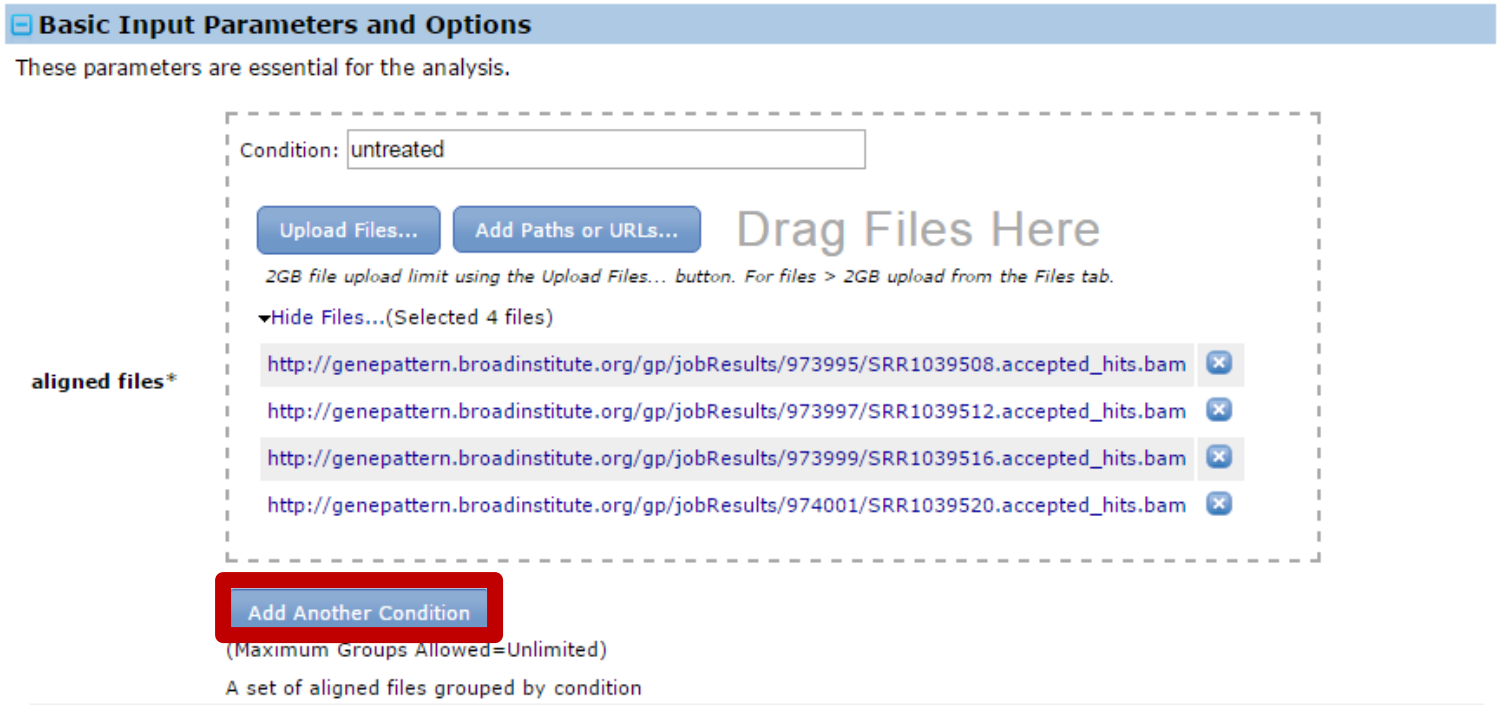
Define the first condition (e.g. '**untreated**'), then click Add Path or URL... Then, navigate to ***shared_data>gp_tutorial_files>2019_BU_MolBio>untreated***, and choose the ***.accepted_hits.bam** files of that same condition (e.g. *SRR…**08***, *SRR…**12***, *SRR…**16*** and *SRR…**20***).



# 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>gp_tutorial_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***).



# Run ***Cuffdiff*** to differentiate phenotypes

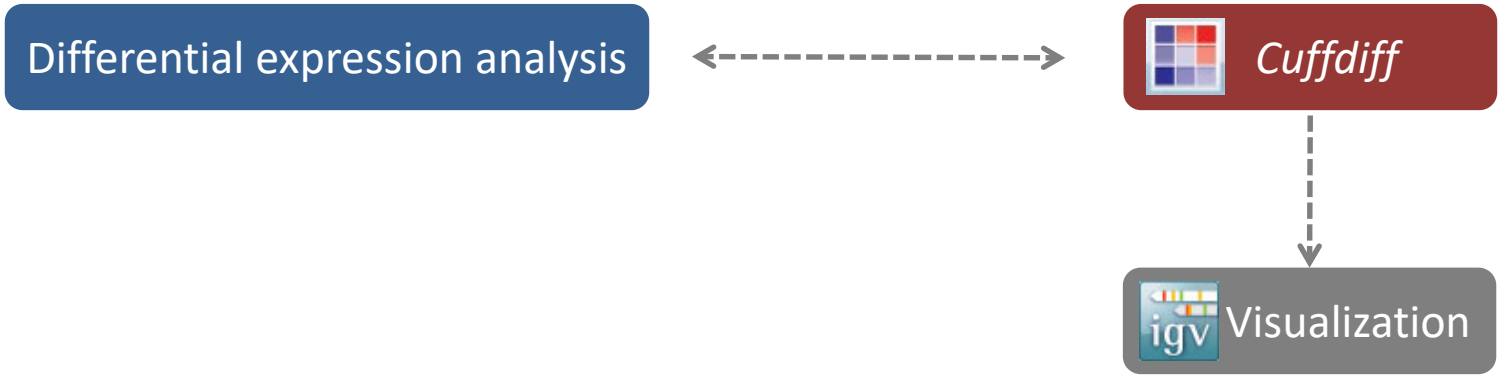
**FA**

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 Run.

# **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***.

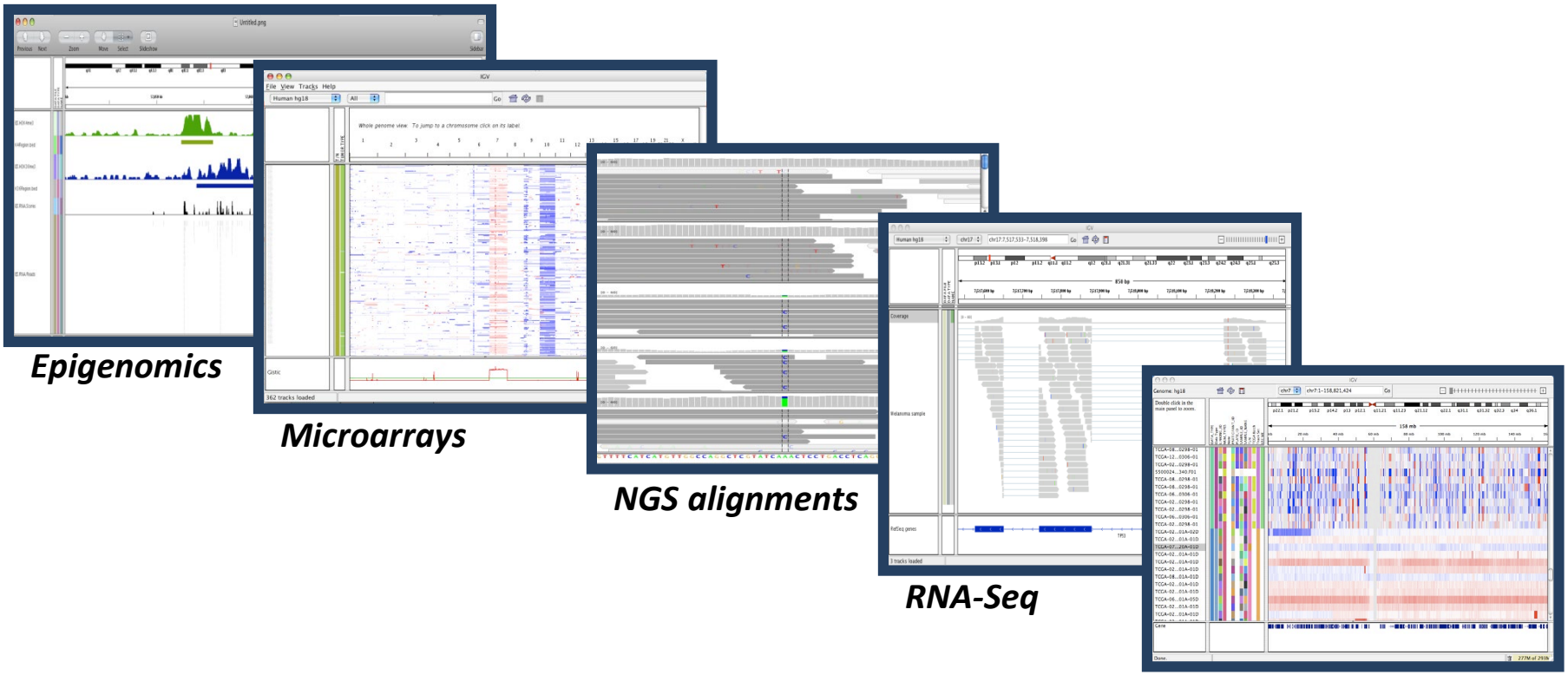



Before we continue we'll take a slight detour to learn how to use ***IGV***.

# Break

# Integrative Genomics Viewer (IGV)

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



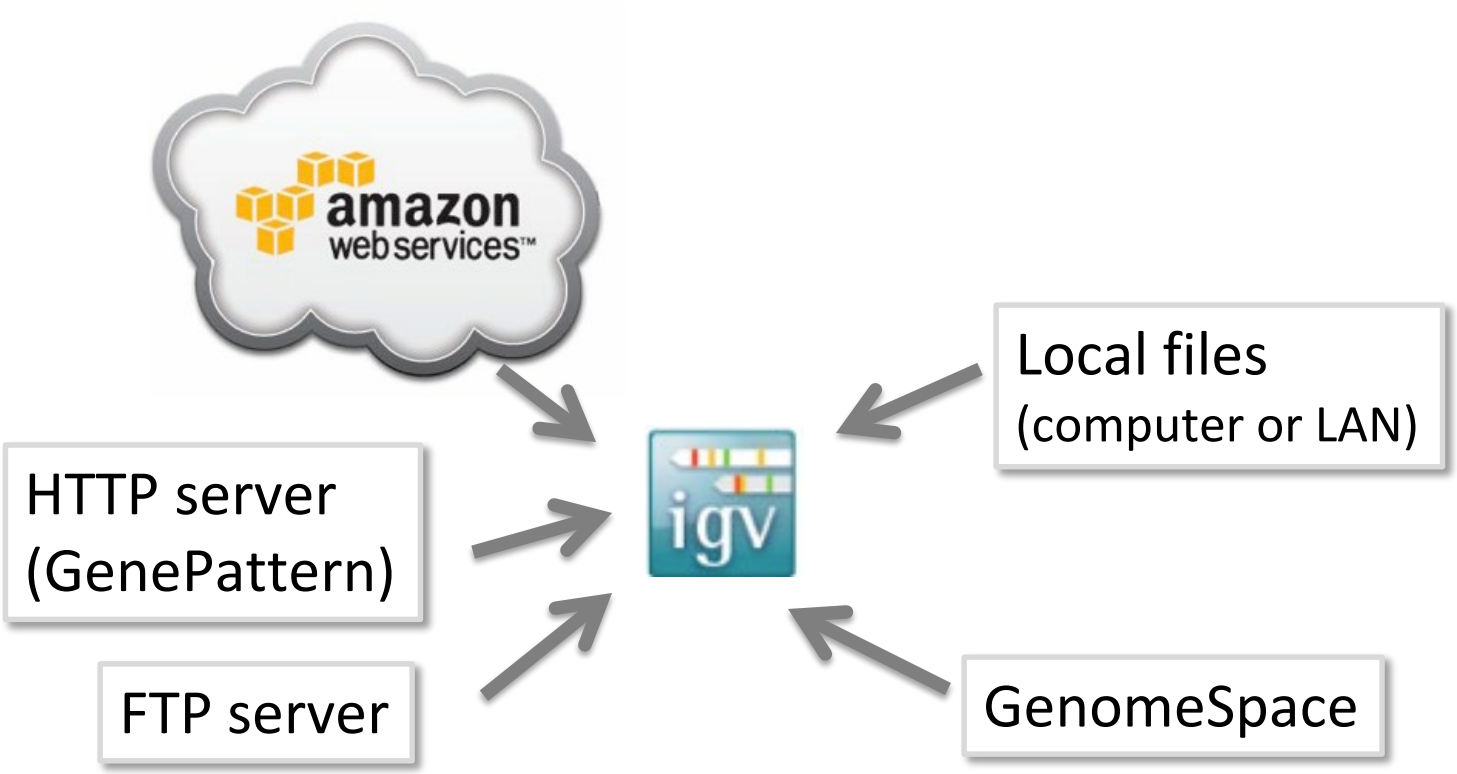

**www.igv.org**

# IGV Features

With IGV you can…

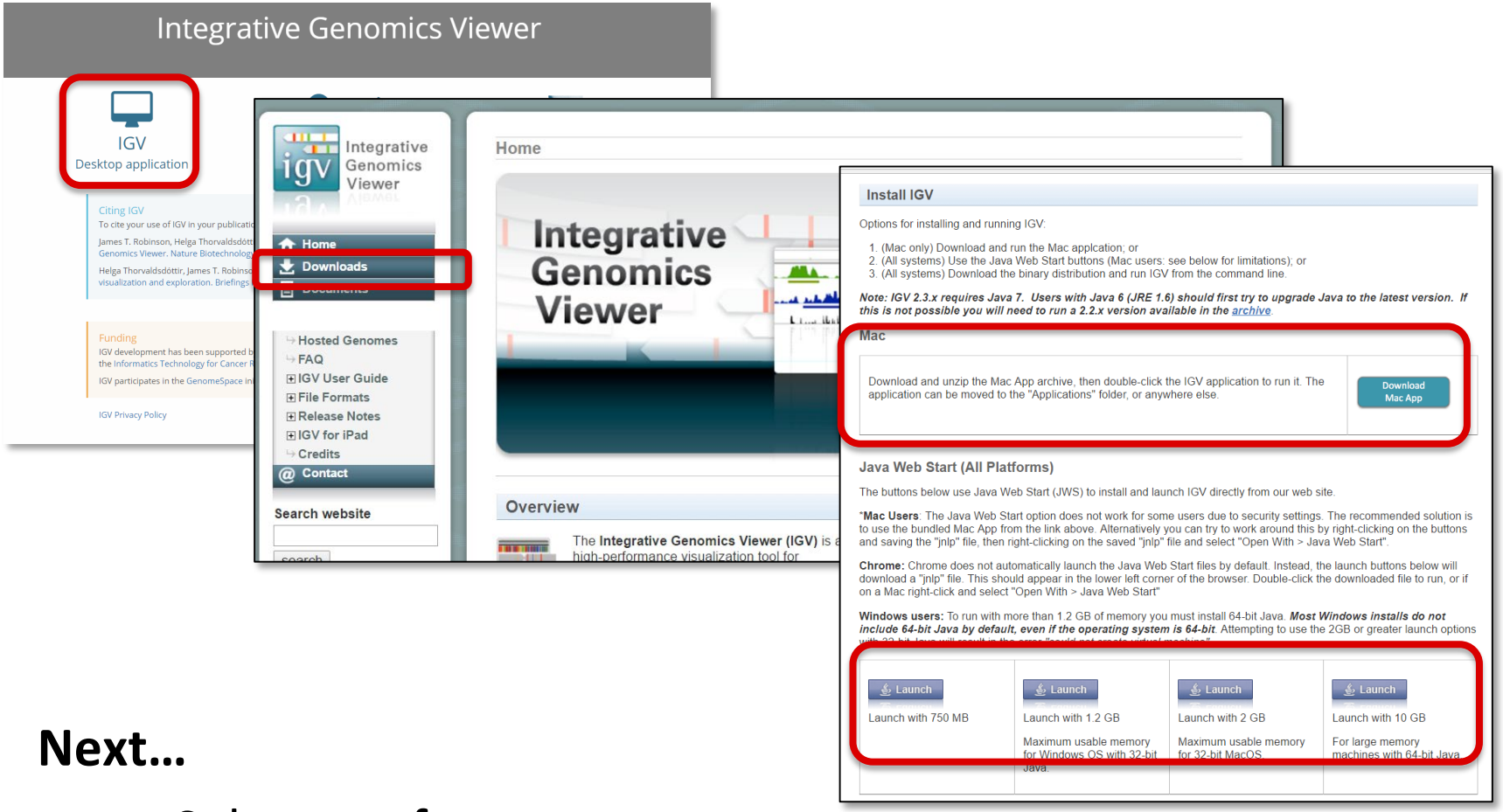
- explore large genomic datasets with an intuitive, easy-to-use interface.
- integrate multiple data types with clinical and other sample information.
- view data from multiple sources: local, remote, and “cloud-based”.

# 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




**Next…**

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

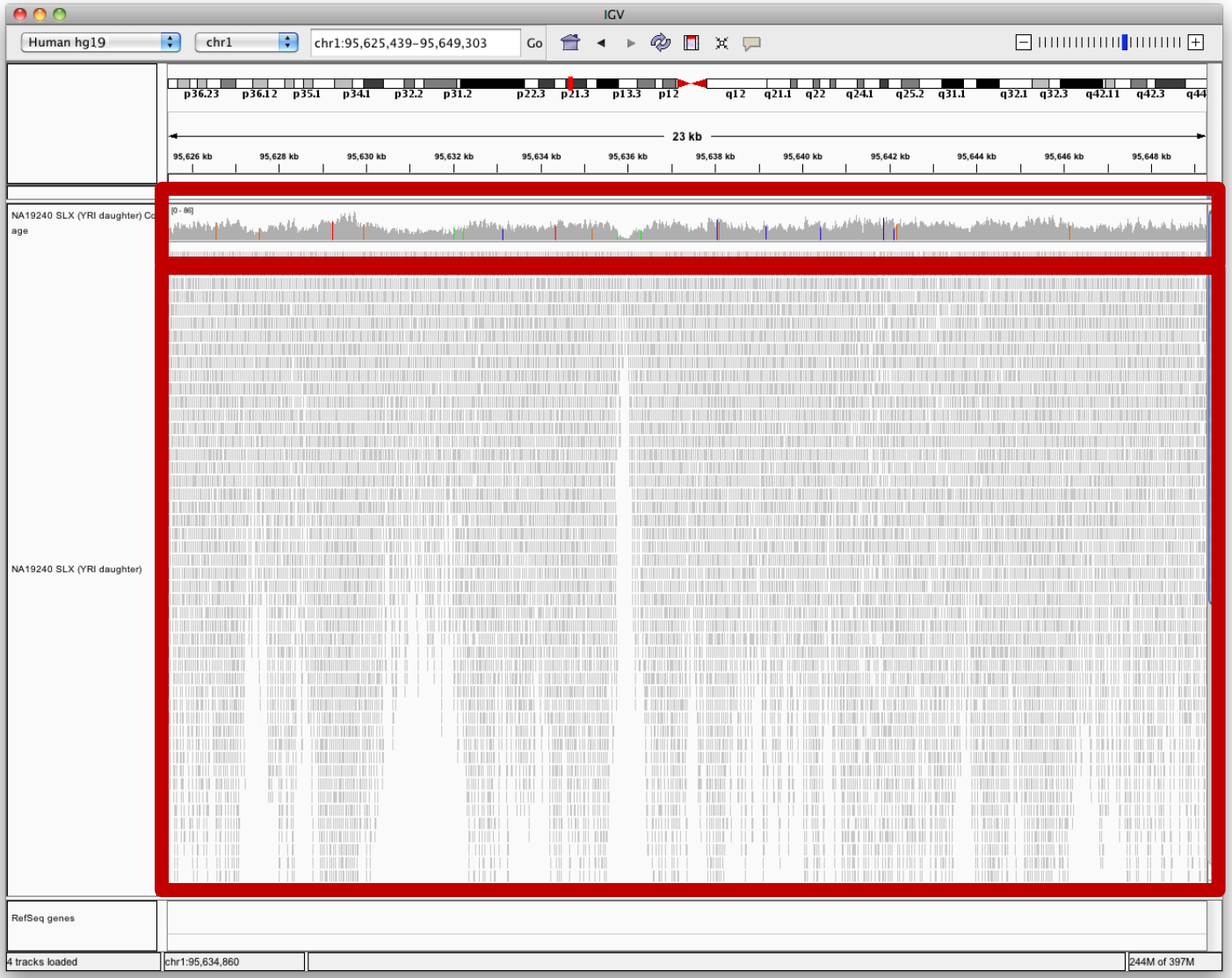
# Viewing next generation sequencing (NGS) data in IGV

# Viewing alignments

Whole chromosome view.

# Viewing alignments

Zoom in to view alignments.

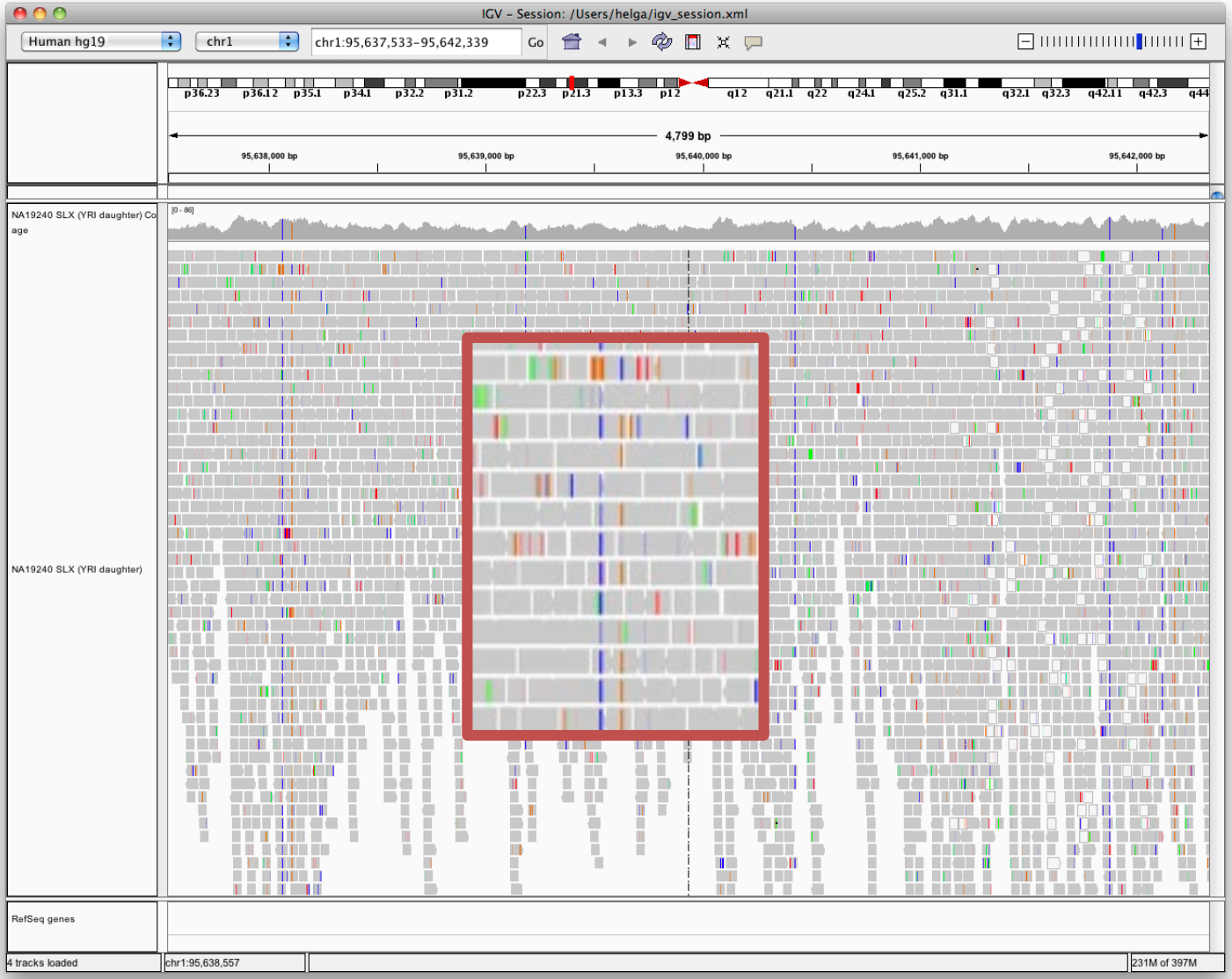



Coverage

Aligned reads

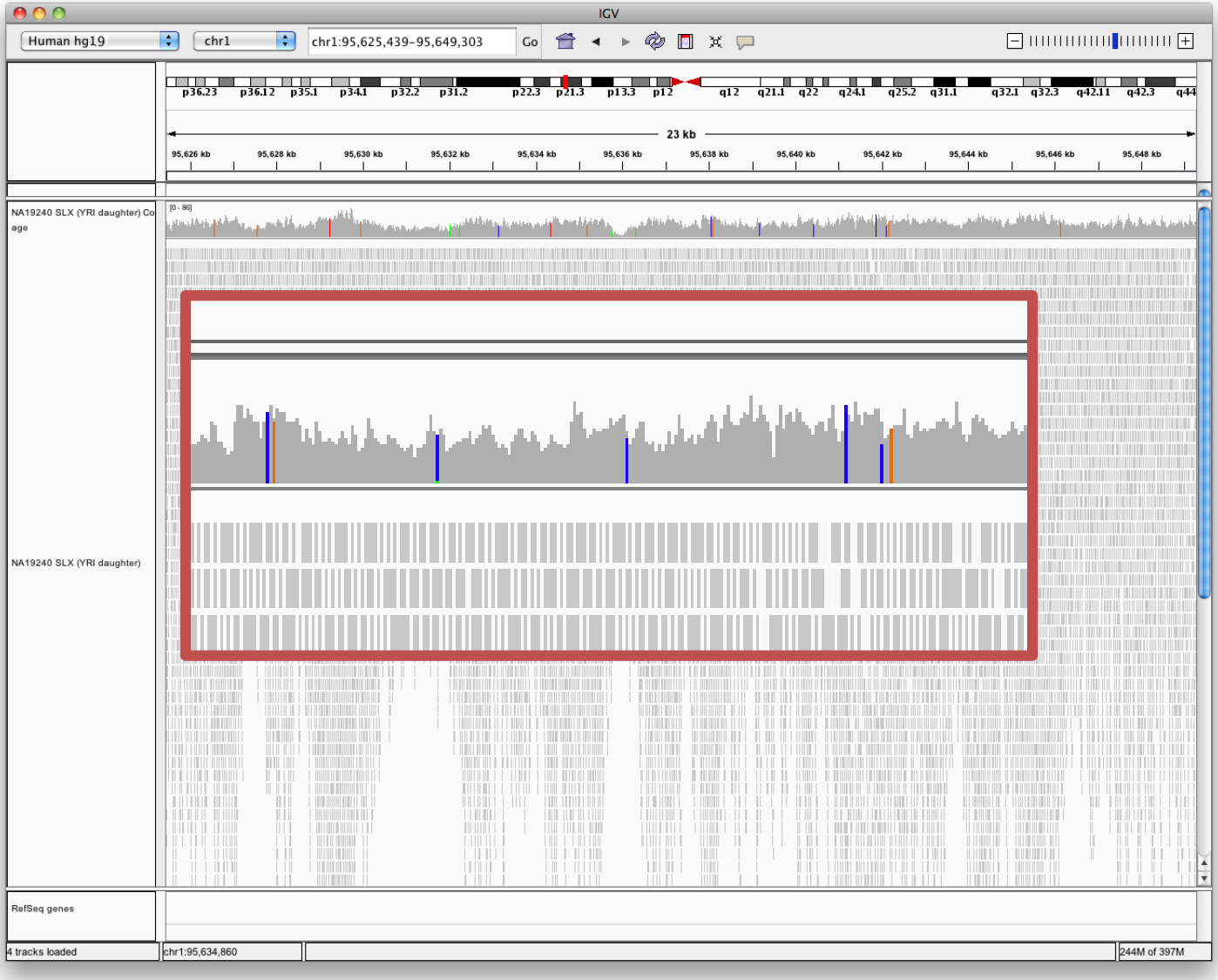
# Viewing alignments

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



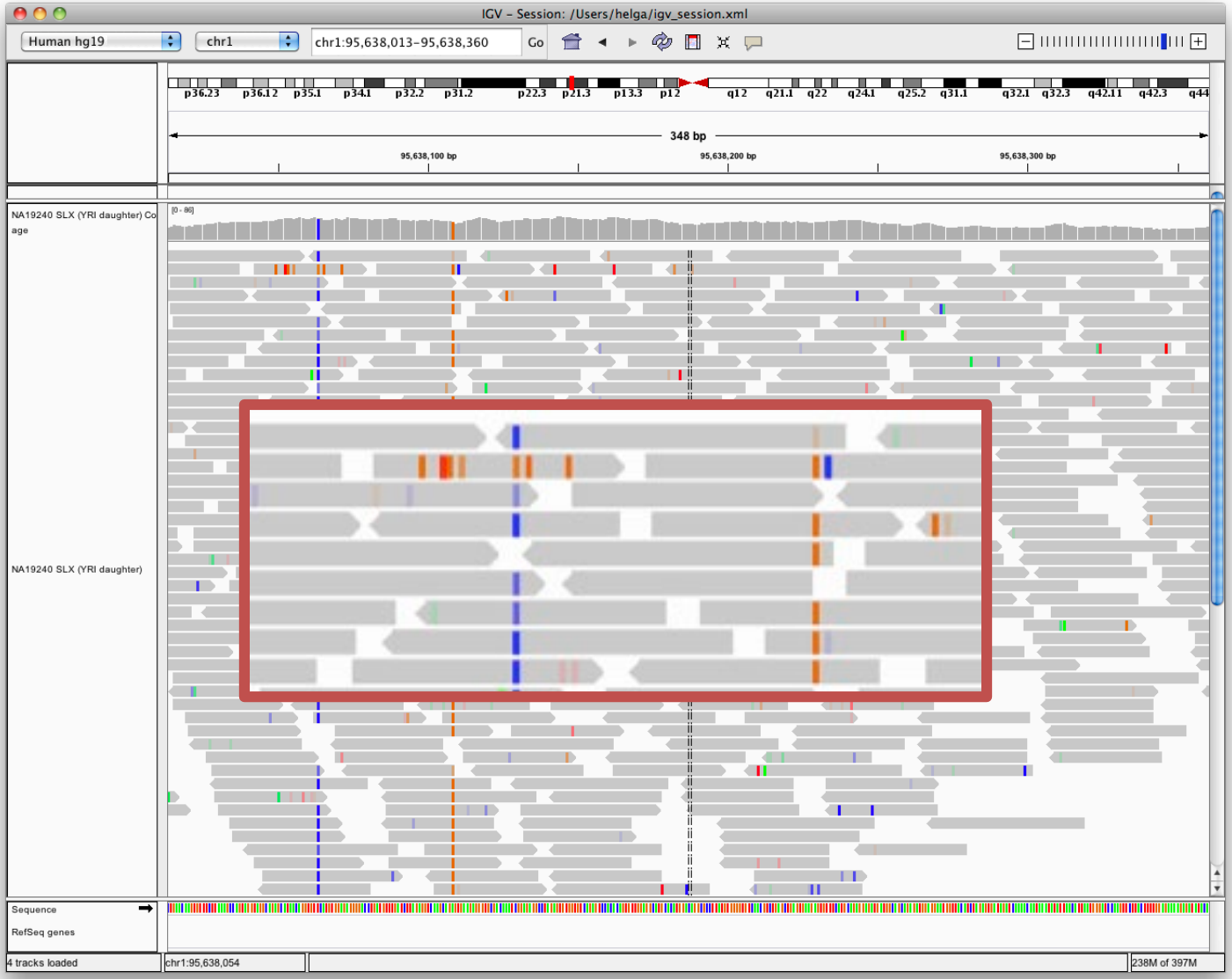
# Viewing alignments

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

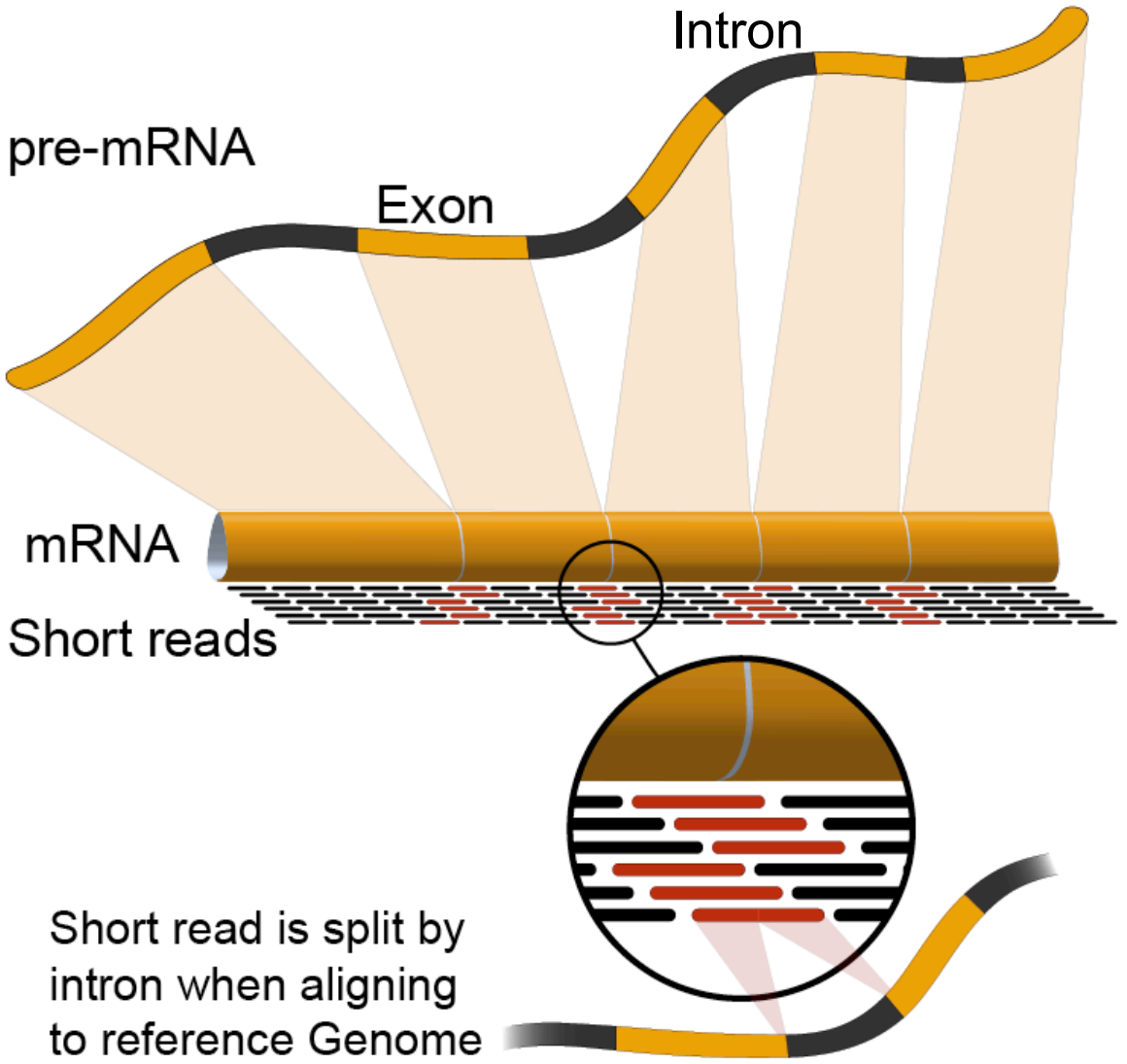
How far do you need to zoom in to see the alignments? **30kb.**


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

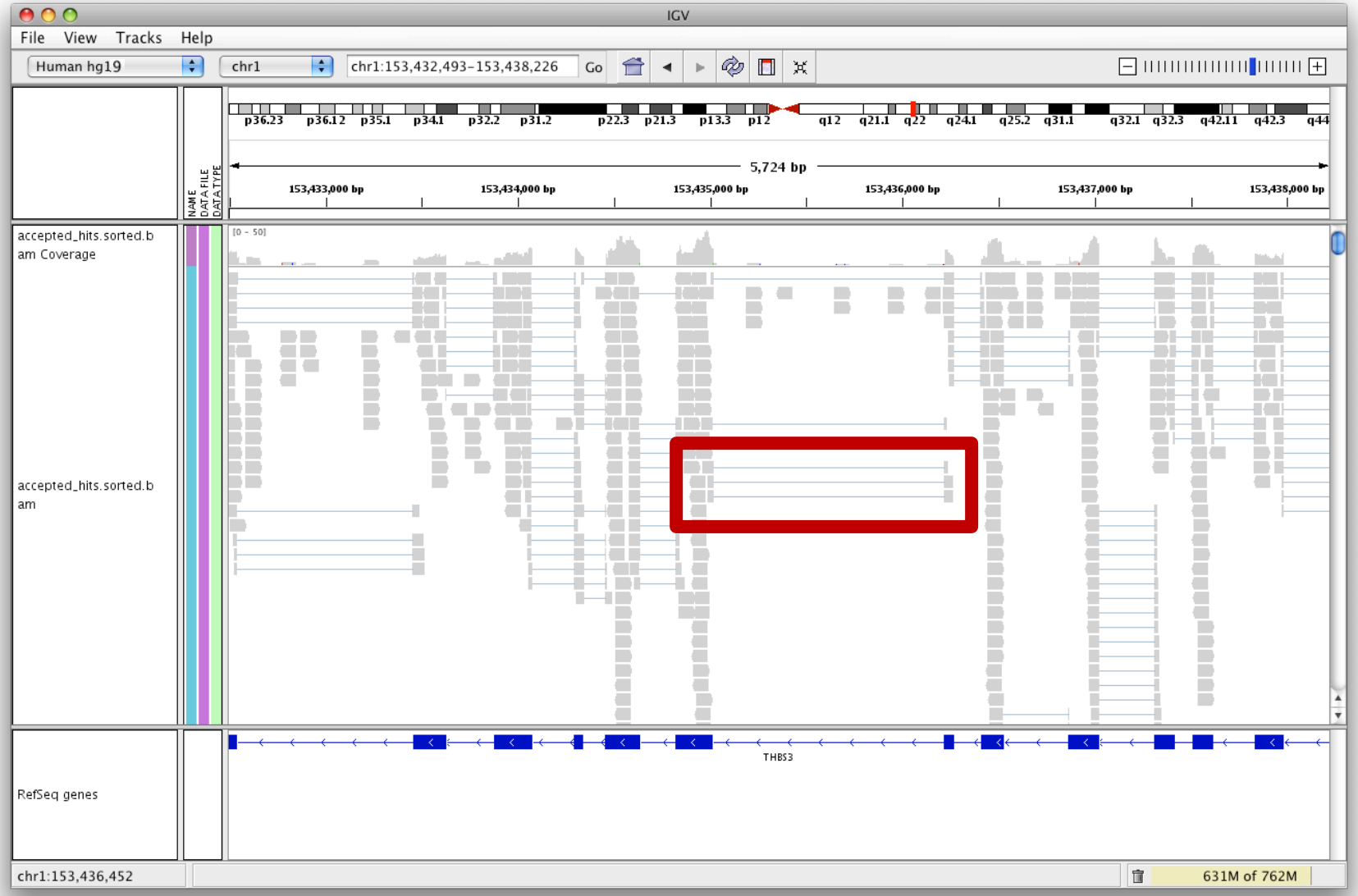
# RNA-seq alignments




http://en.wikipedia.org/wiki/RNA-Seq#Gene_expression

# RNA-seq alignments

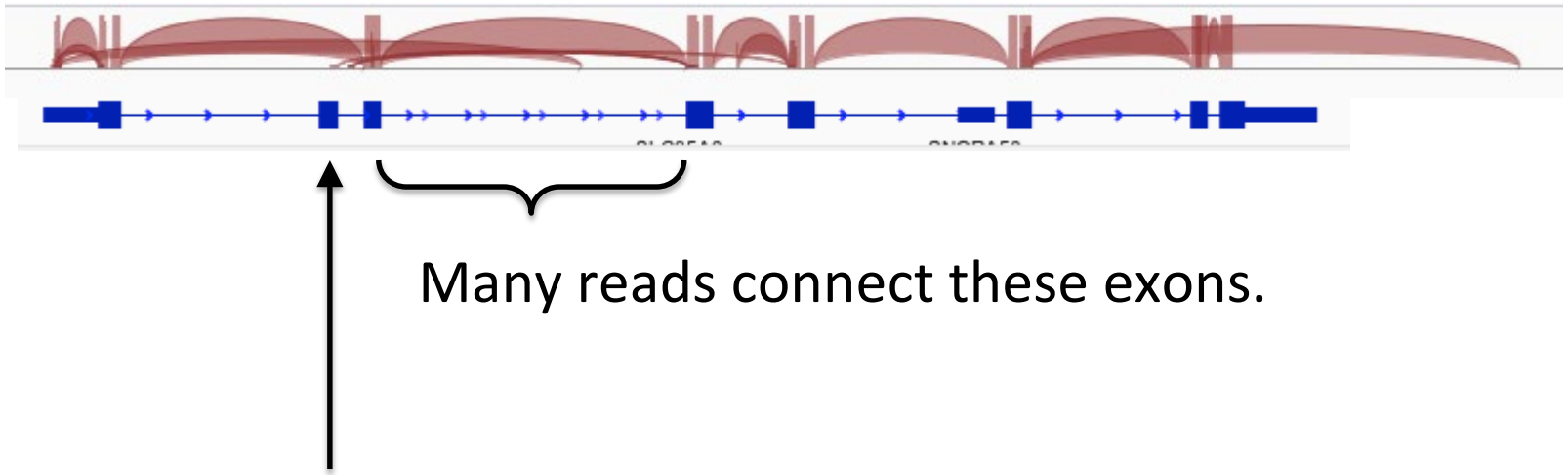
RNA-seq reads can span splice junctions.



# RNA-seq alignments

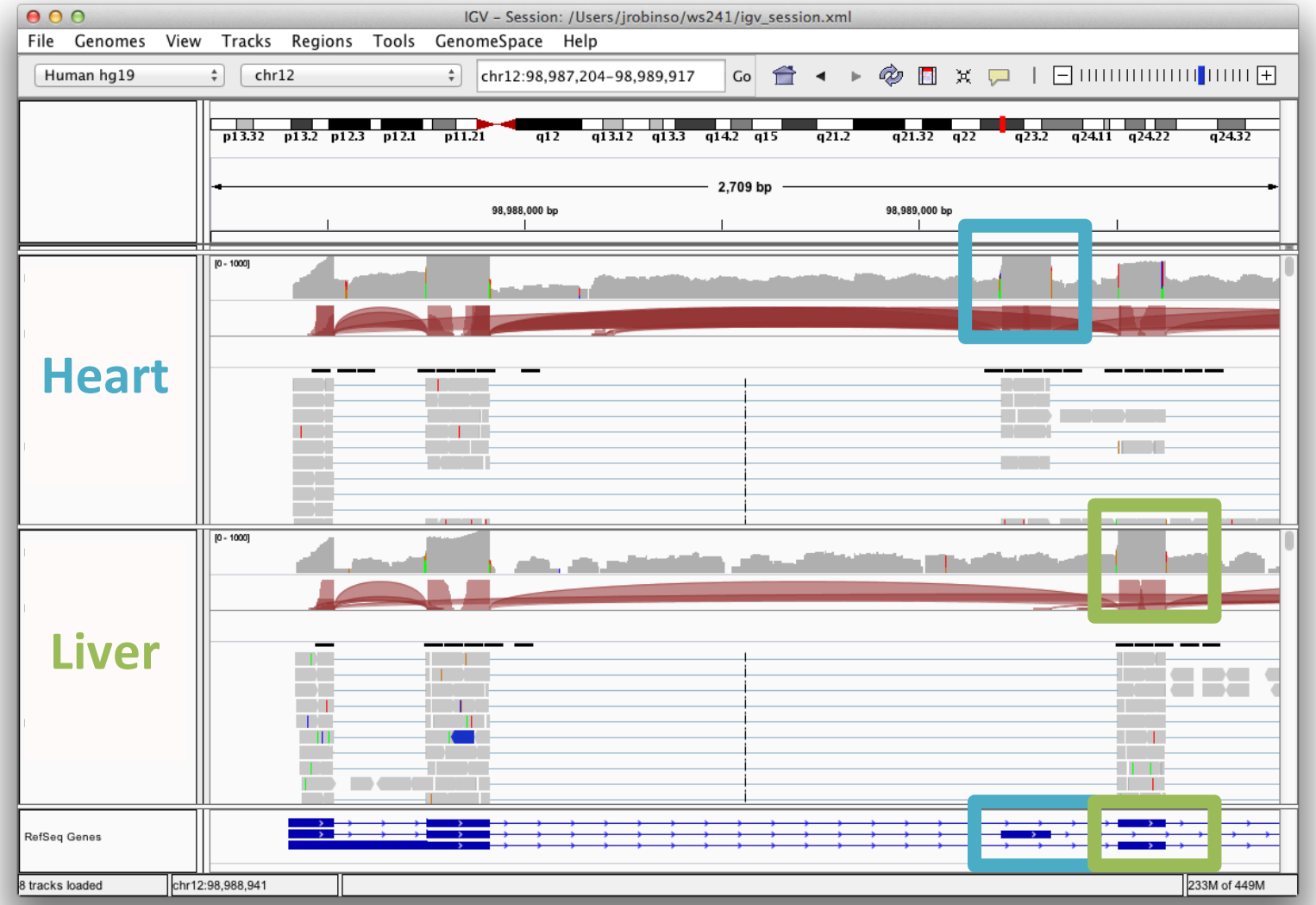
You can display a separate splice junction track which shows how many reads span exon junctions.

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



# RNA-seq alternative splicing

IGV lets you visualize evidence of alternative splicing.



# Viewing our aligned reads in IGV

# Visualization in ***IGV***

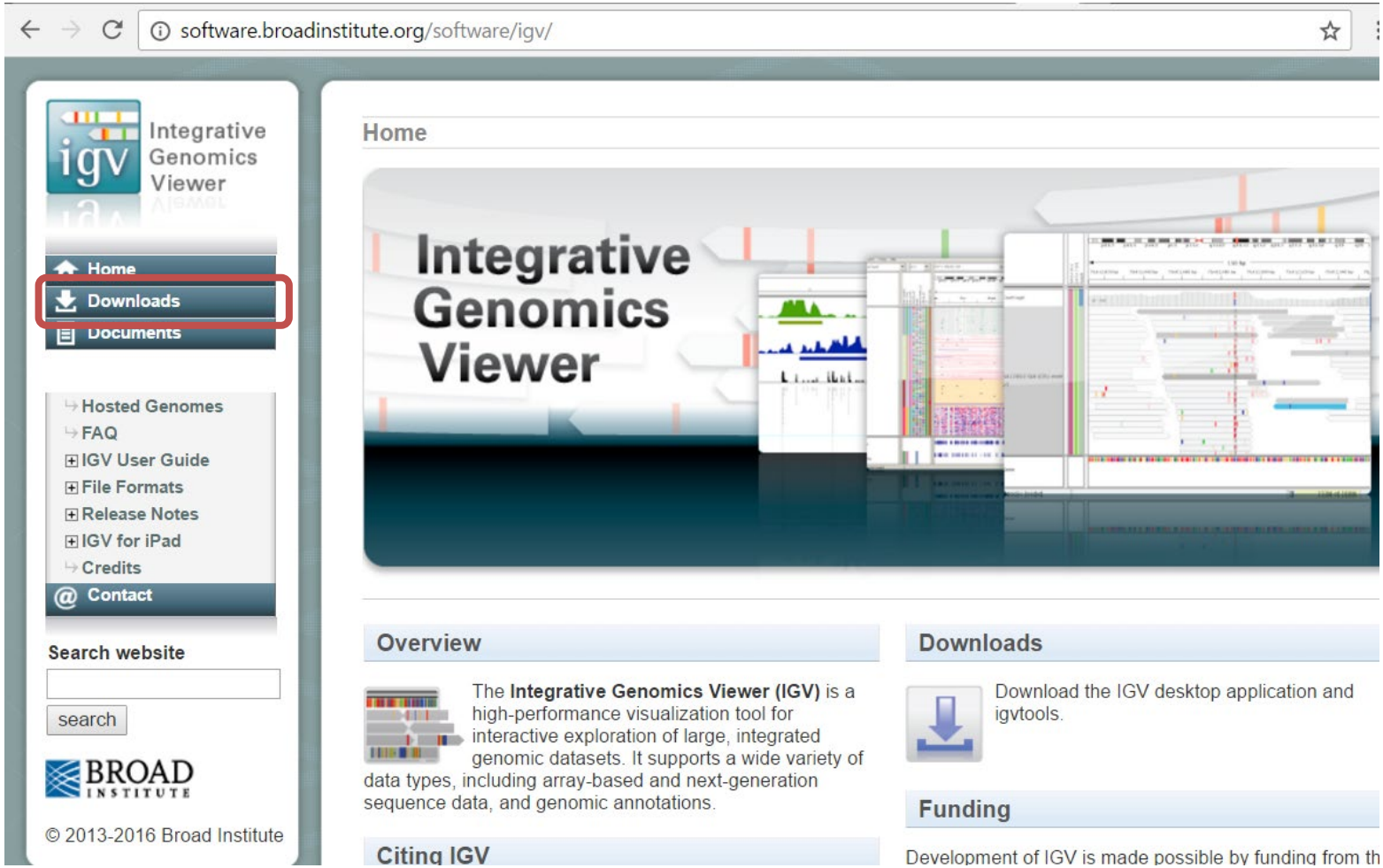
FA

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

# Visualization in ***IGV***

FA

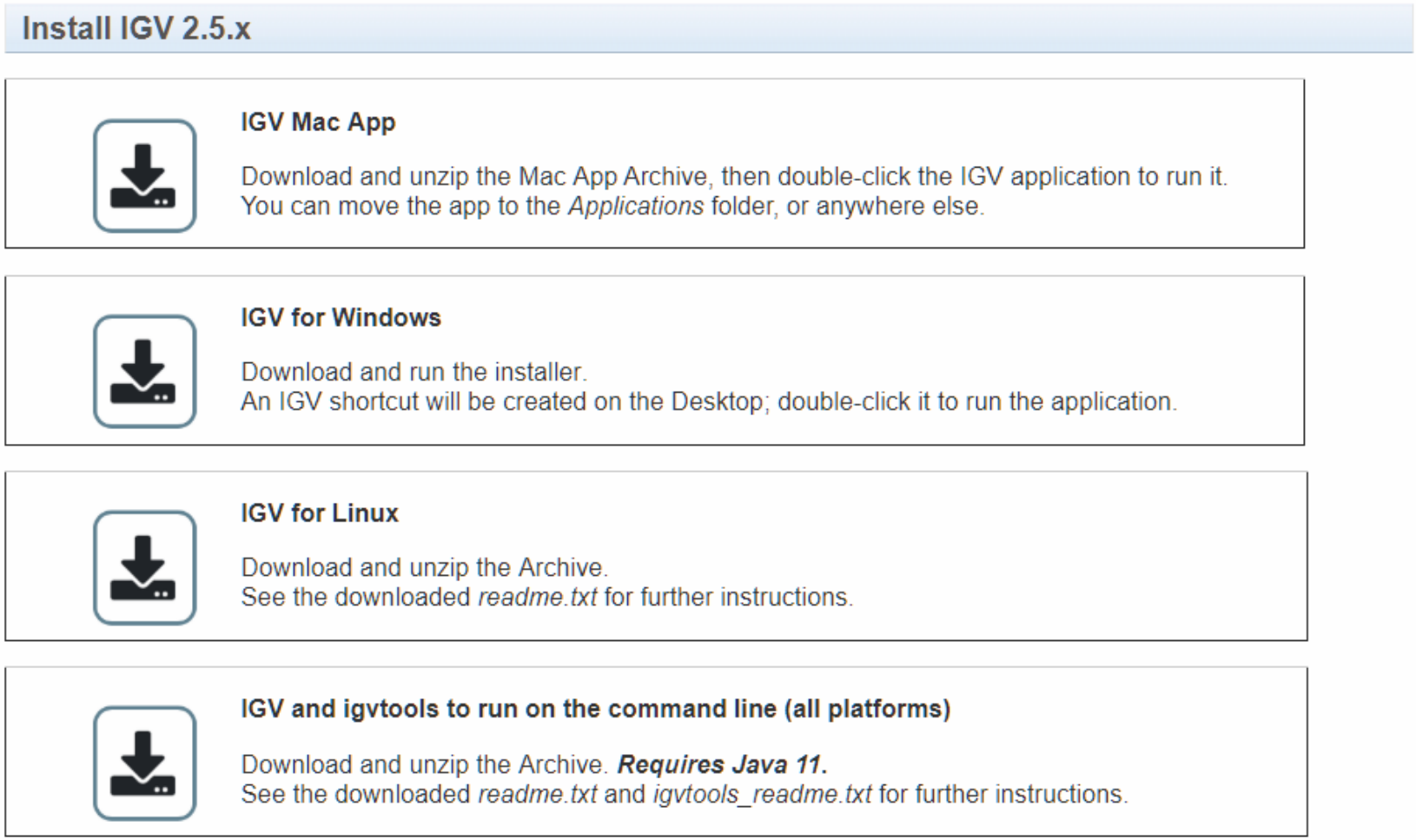
Click on the Downloads button.



# Visualization in ***IGV***

FA

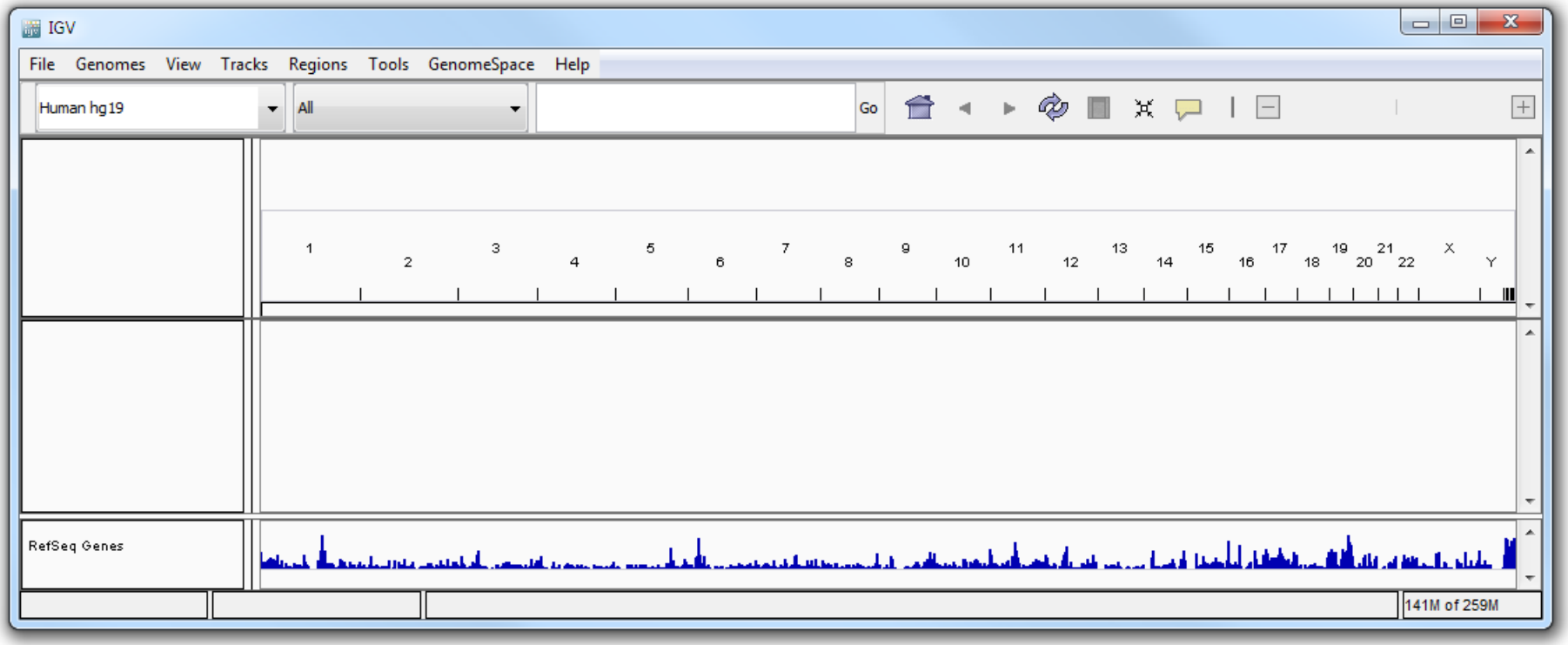
Select the appropriate button for your computer.



# Visualization in ***IGV***

**FA**

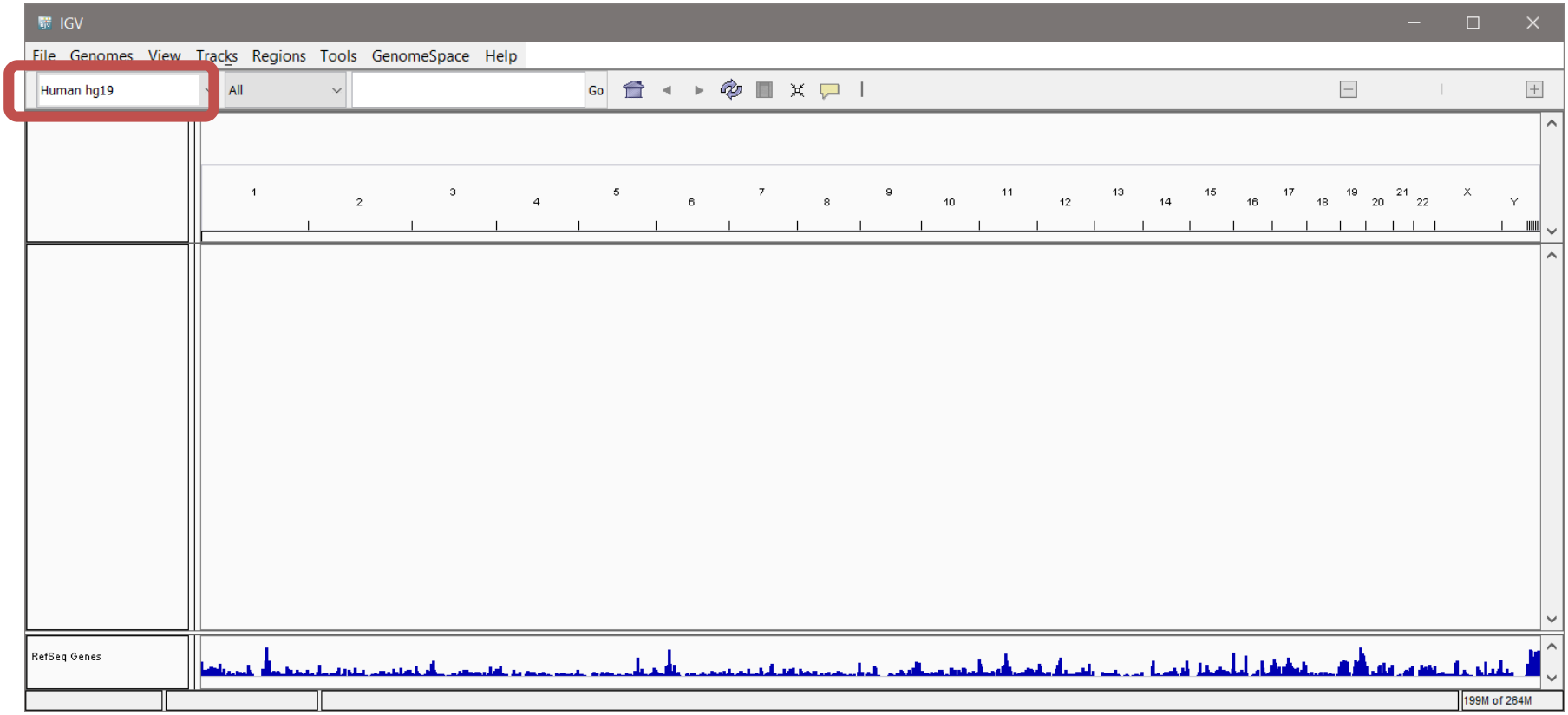
An ***executable IGV*** file will be downloaded to your computer. Click to launch ***IGV***. If a Java security warning pops up make sure to allow IGV access to the internet – it needs this to download genomes.



# Visualization in ***IGV***

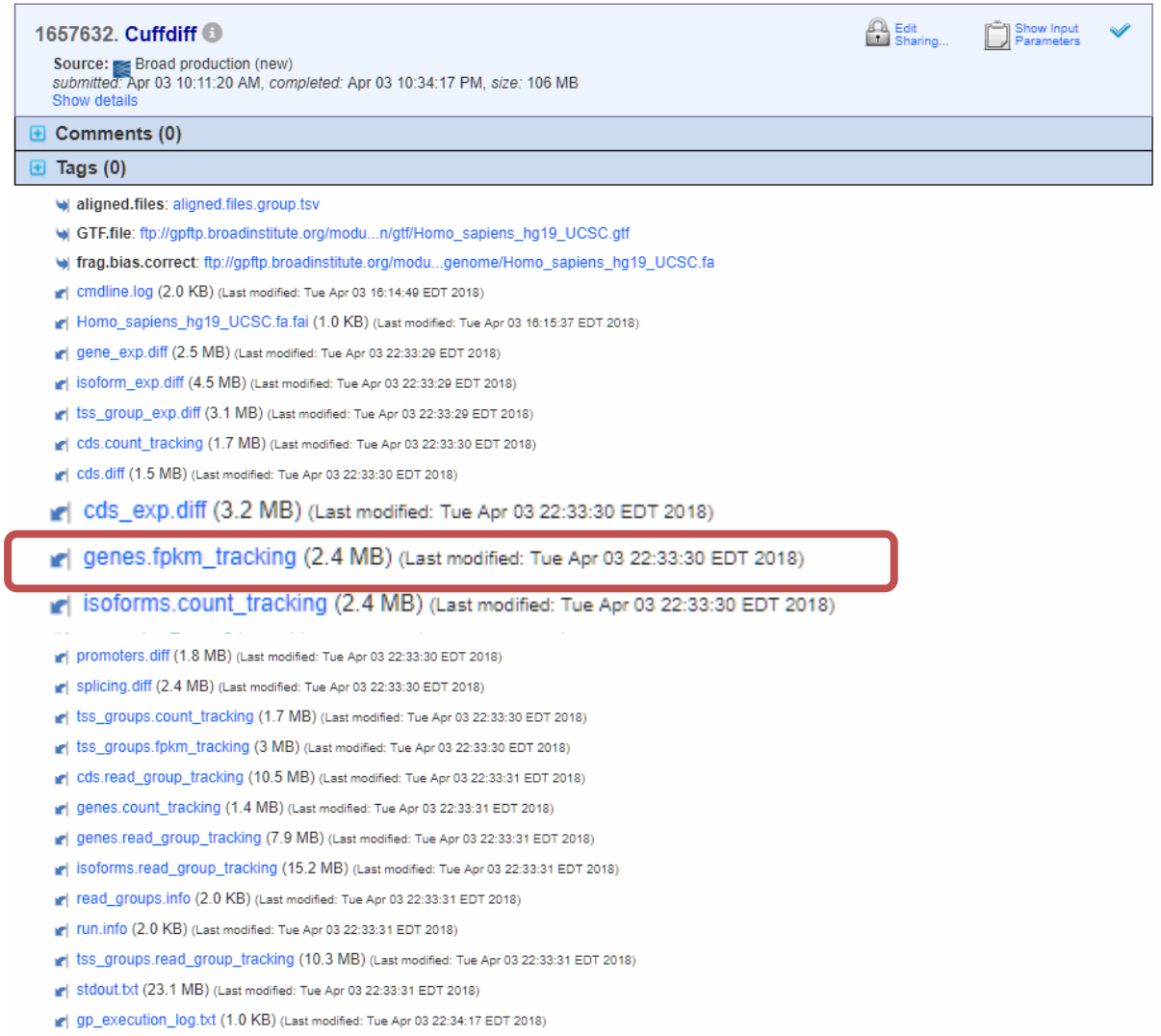
FA

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



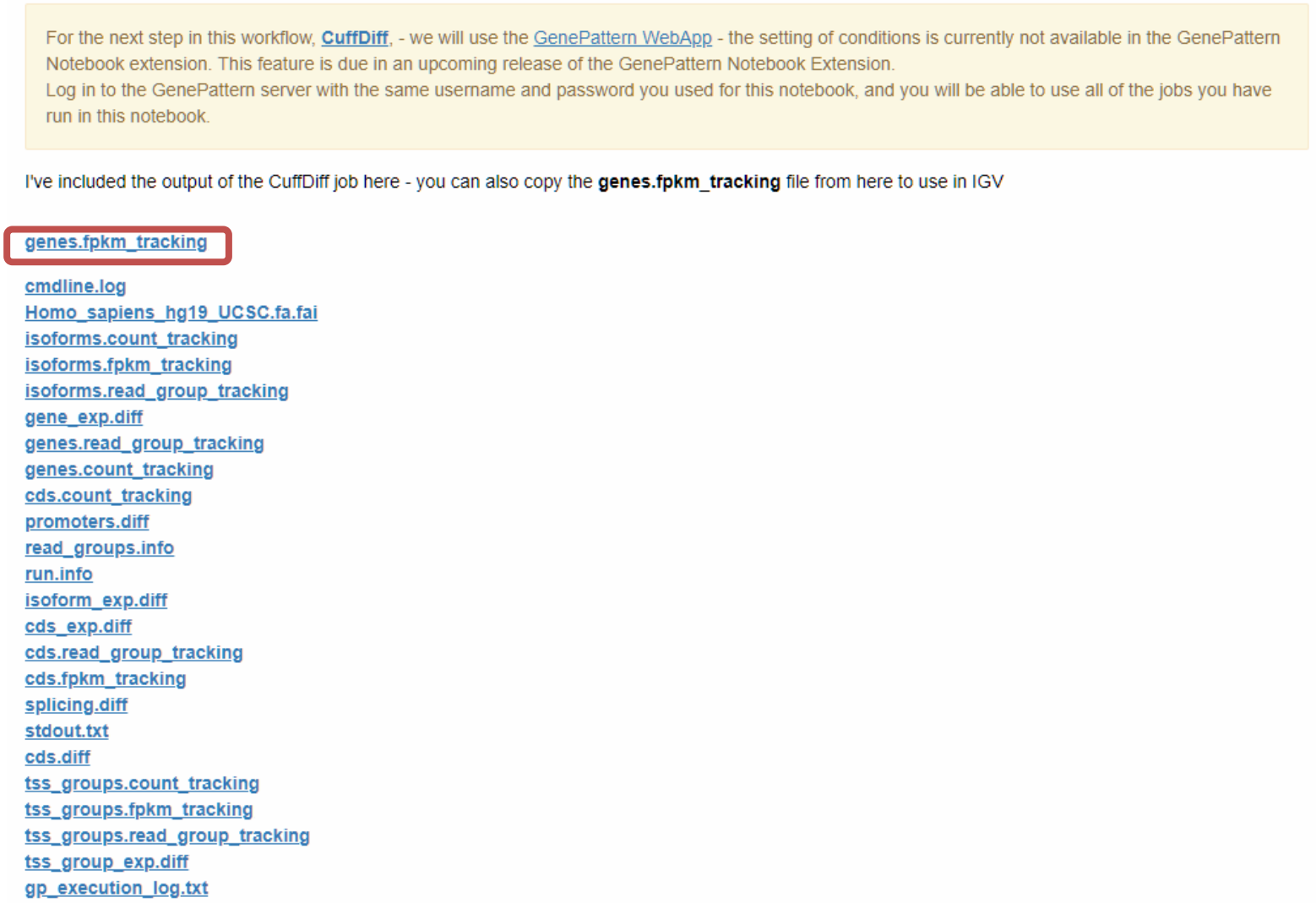
# Visualization in ***IGV***

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**.



# 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.

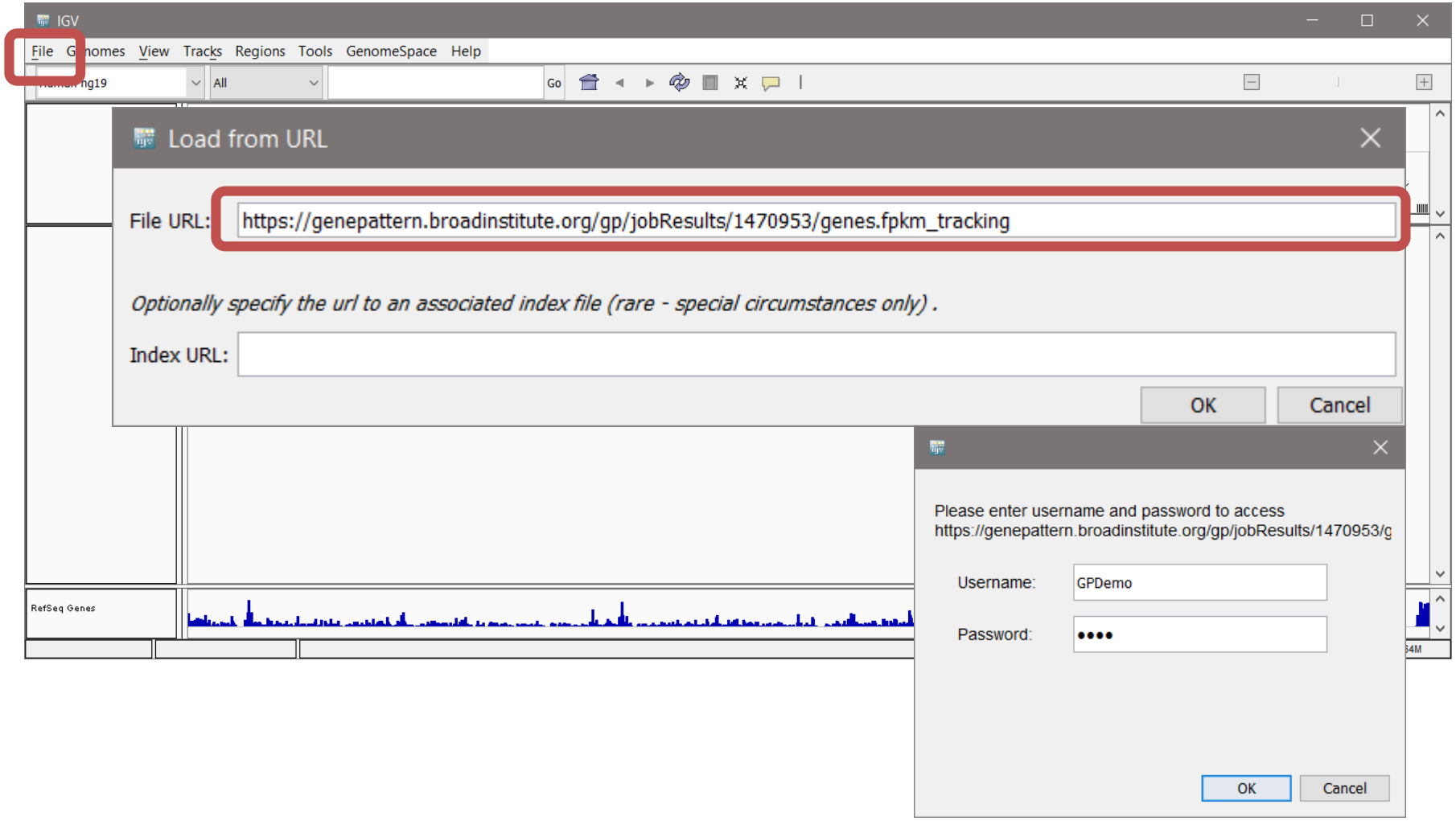
I've included the output of the CuffDiff job here - you can also copy the **genes.fpkm_tracking** file from here to use in IGV

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

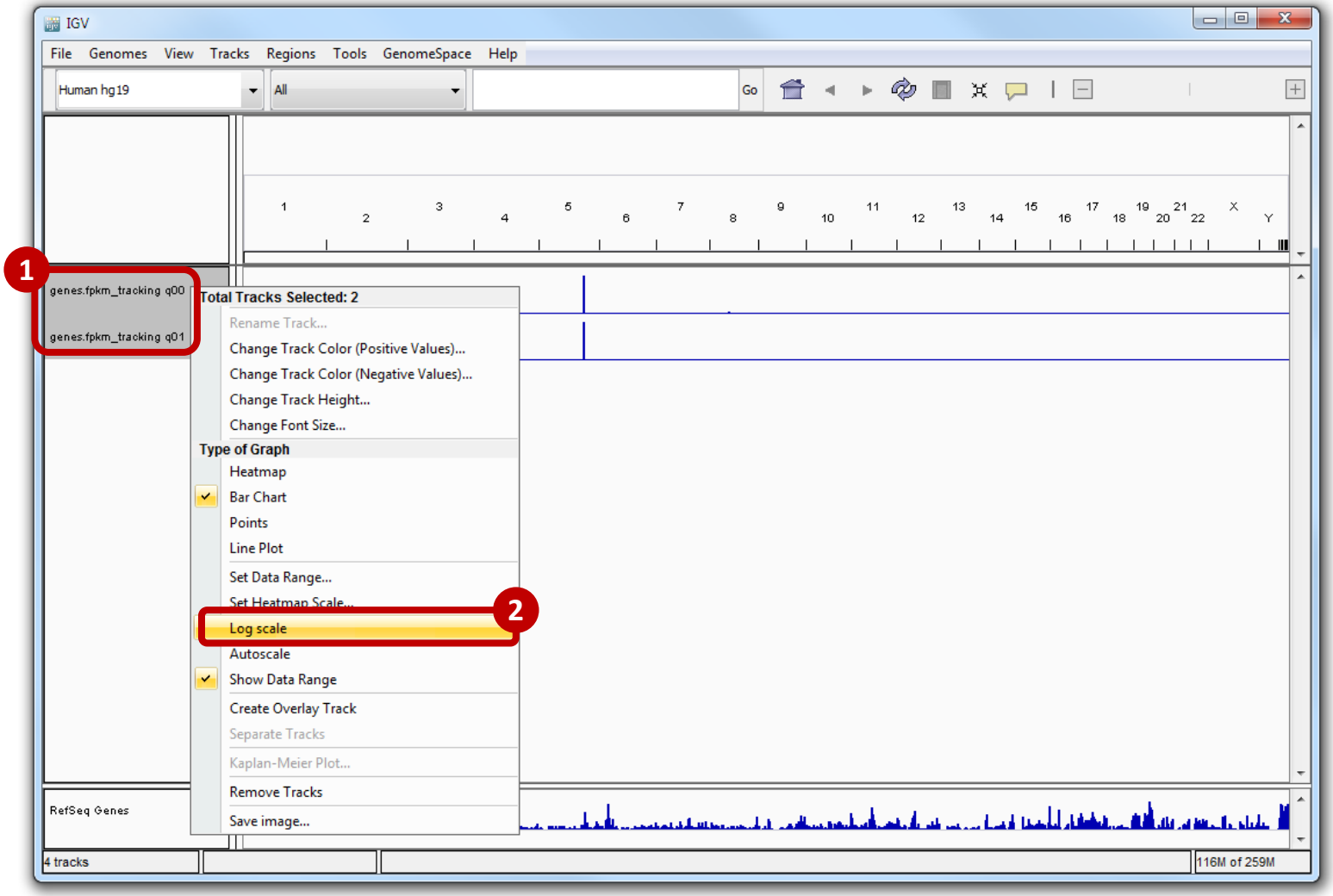
# Visualization in ***IGV***

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***

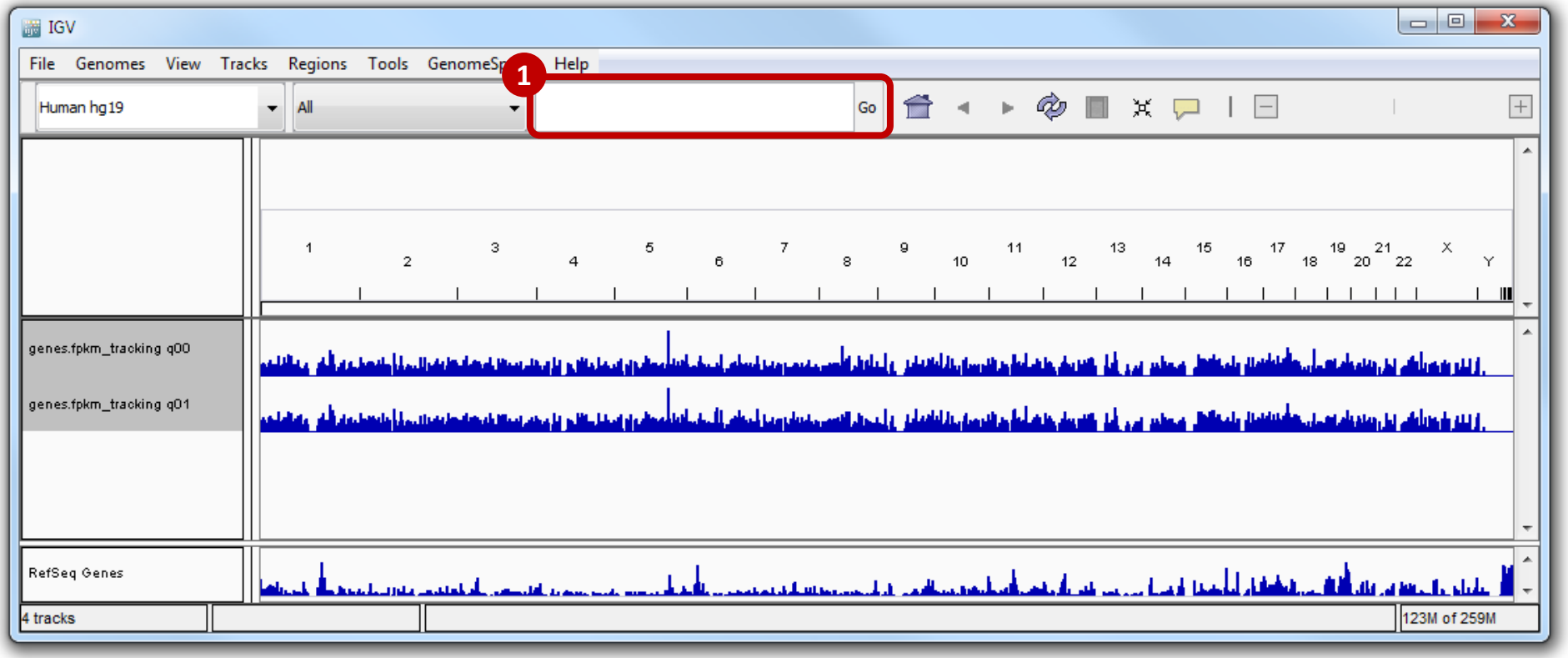
❶ **Ctrl+Click** the **q00** and **q01** tracks. ❷ Right-click and scroll down, then select **log scale** to correctly display the data.



# Visualization in ***IGV***

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

❶ 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***

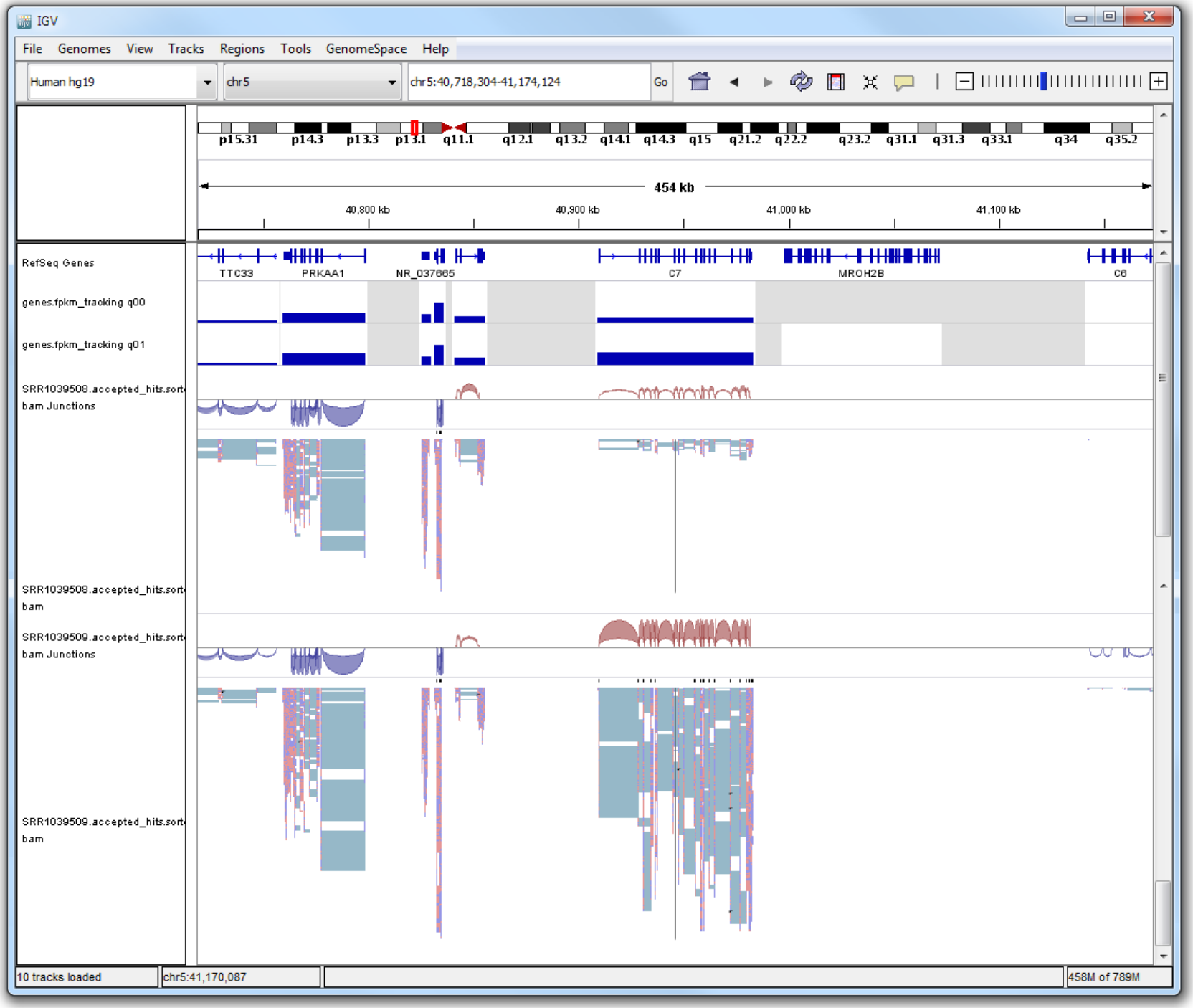
In addition to looking at the gene expression in the **genes.fpkm_tracking** file, we can look at how the reads stack against transcripts by looking at the aligned reads.

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.


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr5
chr5:40,718,304-41,174,124
Go
p15.31 p14.3 p13.3 p13.1 q11.1 q12.1 q13.2 q14.1 q14.3 q15 q21.2 q22.2 q23.2 q31.1 q31.3 q33.1 q34 q35.2
454 kb
40,800 kb
40,900 kb
41,000 kb
41,100 kb
RefSeq Genes
TTC33
PRKAA1
NR_037665
C7
MROH2B
C6
genes.fpkm_tracking q00
genes.fpkm_tracking q01
SRR1039508.accepted_hits.sort
bam Junctions
SRR1039508.accepted_hits.sort
bam
SRR1039509.accepted_hits.sort
bam Junctions
SRR1039509.accepted_hits.sort
bam
10 tracks loaded
chr5:41,170,087
458M of 789M

# 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

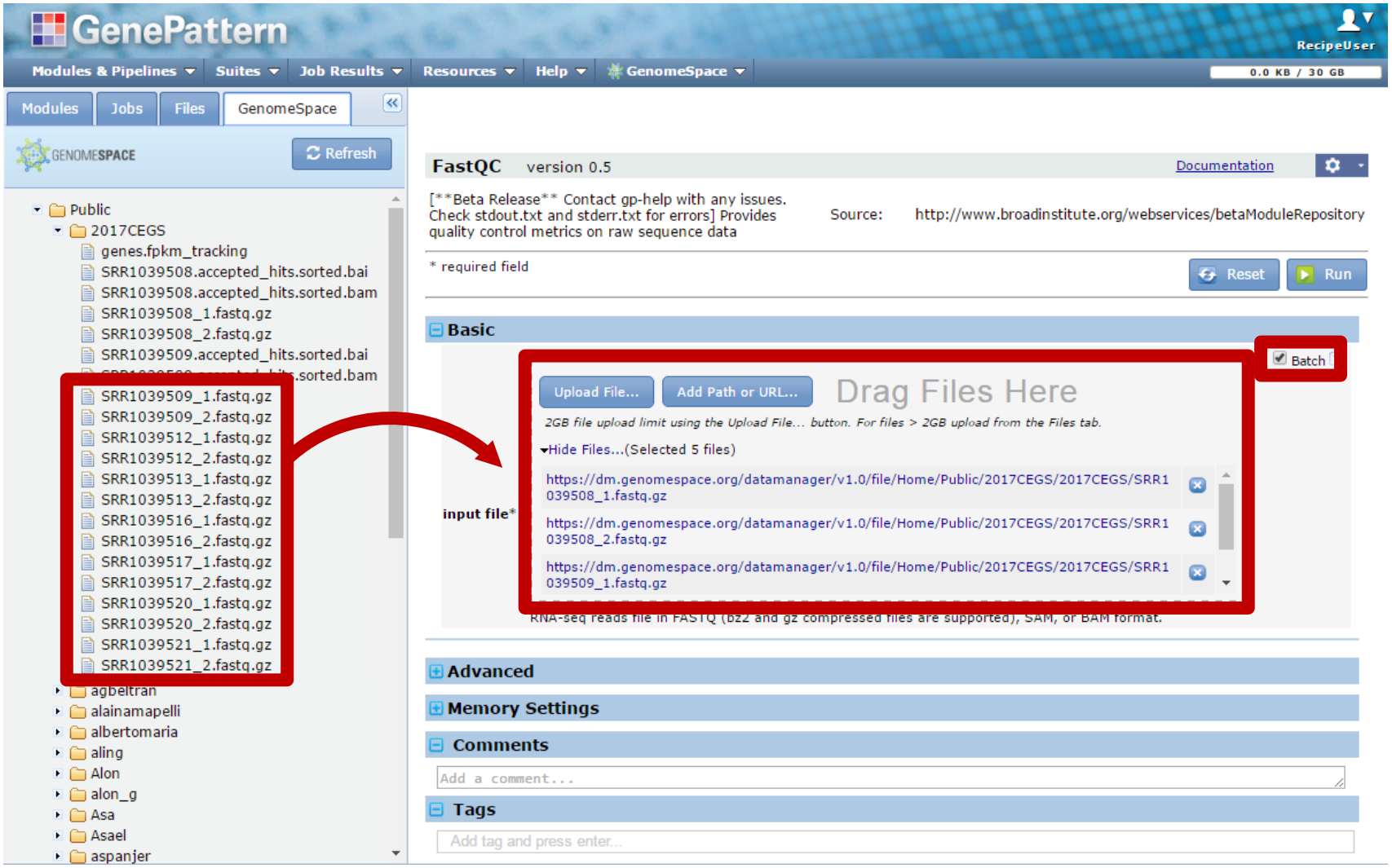
# Managing job access

- 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

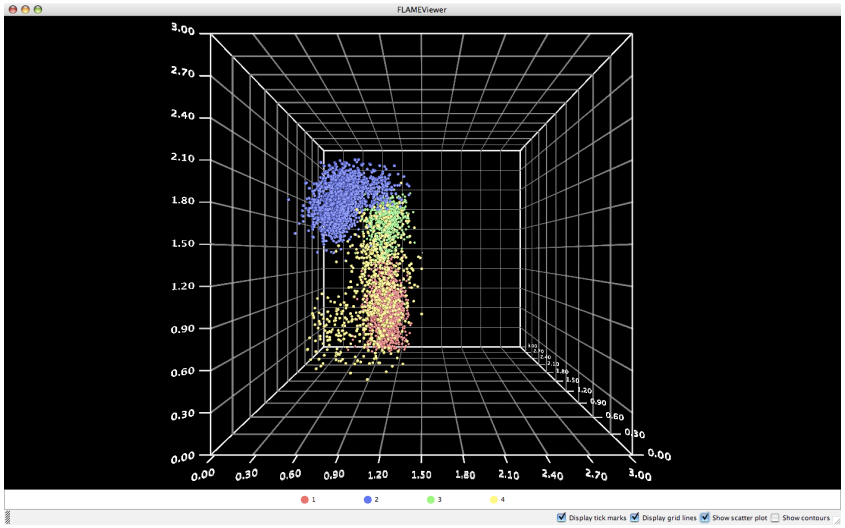
# **Example:** run ***FastQC*** on all pairs of FASTQ files

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

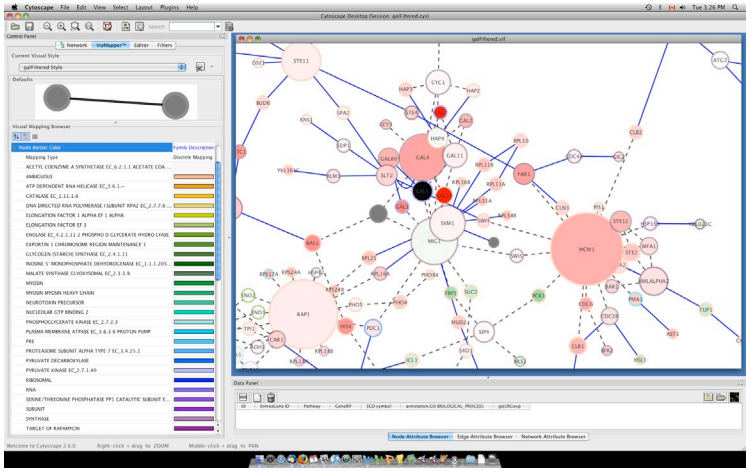


# Creating your own Notebooks & Editing existing

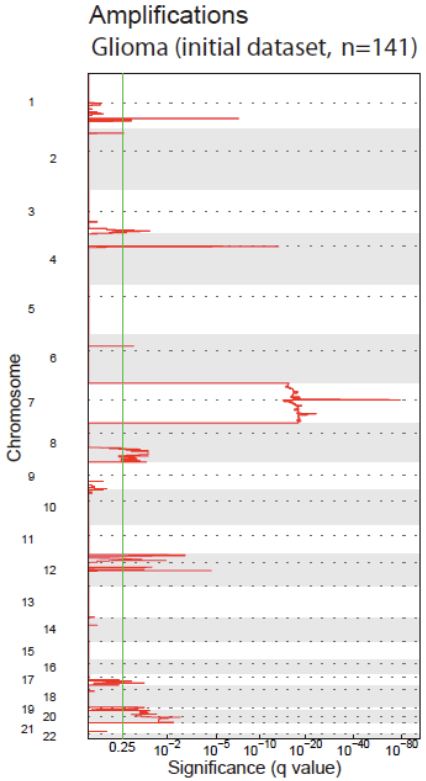
# Other GenePattern Features




Flow Cytometry



Network Analysis




Sequence Variation Analysis

Proteomics

# Closing

# Resources

Public GenePattern server
cloud.genepattern.org

GenePattern Notebook
notebook.genepattern.org

GenePattern Website
genepattern.org

Notebook website
genepattern-notebook.org

Integrative Genomics Viewer (IGV)
www.igv.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 Thorvaldsdottìr – 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!