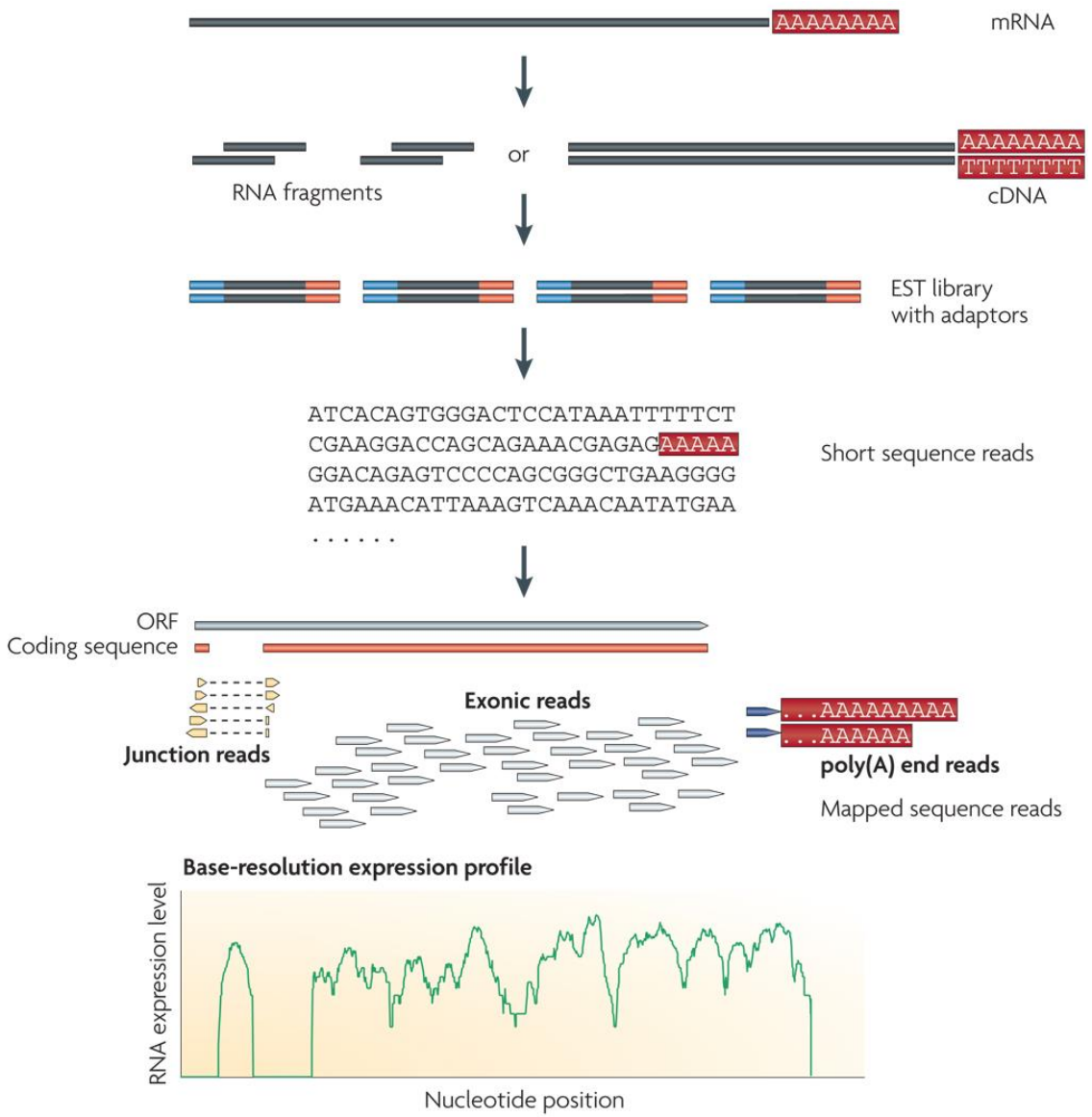
RNA-Seq Differential Gene Expression analysis (Galaxy Server)

GBIO0002

Archana Bhardwaj
University of Liege

Typical RNA-Seq Experiment



Published in final edited form as:
Nat Rev Genet. 2009 January ; 10(1): 57–63. doi:10.1038/nrg2484.

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, **Mark Gerstein**, and **Michael Snyder**
Zhong Wang and Michael Snyder are at the Department of Molecular, Cellular and Developmental Biology, and Mark Gerstein is at the Department of Molecular, Biophysics and Biochemistry, Yale University, 219 Prospect Street, New Haven, Connecticut 06520, USA.

Abstract

RNA-Seq is a recently developed approach to transcriptome profiling that uses deep-sequencing technologies. Studies using this method have already altered our view of the extent and complexity of eukaryotic transcriptomes. RNA-Seq also provides a far more precise measurement of levels of transcripts and their isoforms than other methods. This article describes the RNA-Seq approach, the challenges associated with its application, and the advances made so far in characterizing several eukaryote transcriptomes.

The transcriptome is the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition. Understanding the transcriptome is essential for interpreting the functional elements of the genome and revealing the molecular constituents of cells and tissues, and also for understanding development and disease. The key aims of transcriptomics are: to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications; and to quantify the changing expression levels of each transcript during development and under different conditions.

Mapping and quantifying mammalian transcriptomes by RNA-Seq

Ali Mortazavi[1,2], Brian A Williams[1,2], Kenneth McCue[1], Lorian Schaeffer[1] & Barbara Wold[1]

We have mapped and quantified mouse transcriptomes by deeply sequencing them and recording how frequently each gene is represented in the sequence sample (RNA-Seq). This provides a digital measure of the presence and prevalence of transcripts from known and previously unknown genes. We report reference measurements composed of 41–52 million mapped 25-base-pair reads for poly(A)-selected RNA from adult mouse brain, liver and skeletal muscle tissues. We used RNA standards to quantify transcript prevalence and to test the linear range of transcript detection, which spanned five orders of magnitude. Although >90% of uniquely mapped reads fell within known exons, the remaining data suggest new and revised gene models, including changed or additional promoters, exons and 3′ untranscribed regions, as well as new candidate microRNA precursors. RNA splice events, which are not readily measured by standard gene expression microarray or serial analysis of gene expression methods, were detected directly by mapping splice-crossing sequence reads. We observed 1.45×10^5 distinct splices, and alternative splices were prominent, with 3,500 different genes expressing one or more alternate internal splices.

The mRNA population specifies a cell's identity and helps to govern its present and future activities. This has made transcriptome approaches to large-scale RNA analysis are serial analysis of gene expression (SAGE)[4,5] and related methods such as massively parallel signature sequencing (MPSS)[6], which use DNA sequencing of previously cloned tags 17–25 base pairs (bp) from terminal 3′ (or 5′) sequence tags. These sequence tags are then identified by informatic mapping to mRNA reference databases or, for longer tag lengths, to the source genome. A strength of SAGE and SAGE-like methods is that they produce digital counts of transcript abundance, in contrast to the analog-style signals obtained from fluorescent dye–based microarrays. However, SAGE-family assays provide no information about splice isoforms or new gene discovery, and fully comprehensive measurements of lower-abundance-class RNAs have not been achieved owing to cost and technology constraints. Expressed sequence tag (EST) sequencing of cloned cDNAs has long been the core method for reference transcript discovery[7–9]. It has both qualitative and quantitative limitations, imposed partly by historic sequencing capacity and cost issues, and more crucially by bacterial cloning constraints that affect which sequences are represented and how sequence-complete each clone is. Recently, dense whole-genome tiling microarrays have been developed and applied to transcriptomes for measuring expression and for transcript discovery[10–14]. In contrast to expression arrays, these tiling arrays can discover new genes and exons, but they require large amounts of input RNA and have

What Can You Actually Do With RNA-Seq?

✓**RNA-seq is a powerful and versatile tool published widely over the last few years.**

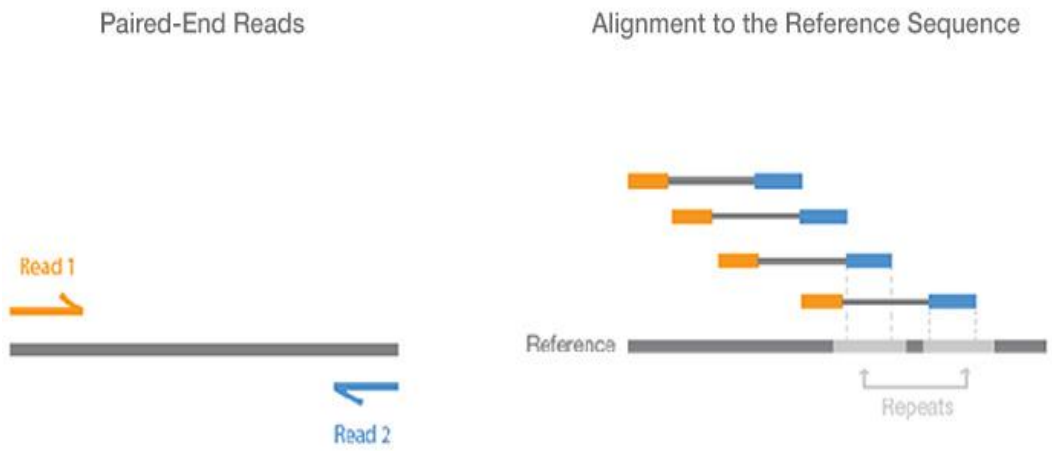
✓**RNA-seq used to investigate complex diseases and find new genes for functional analysis.**

✓**RNA-seq used in one of the study to look at the conservation of RNA Polymerase III binding in mammals.**

✓**RNA-seq and microarray-based capture used to identify and characterize rare transcripts, which are normally undetectable.**

Paired end sequence

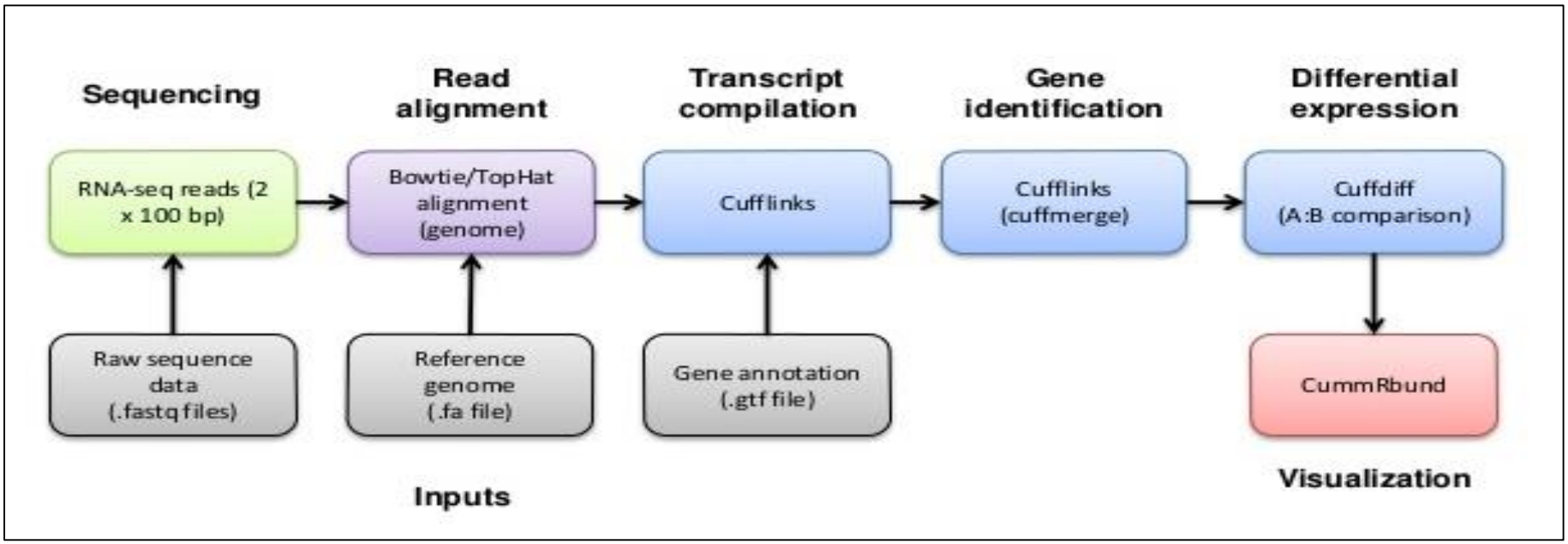
The term 'paired ends' refers to the two ends of the same DNA molecule. So you can sequence one end, then turn it around and sequence the other end. The two sequences you get are 'paired end reads'.



Paired-end RNA sequencing (RNA-Seq) enables discovery applications such as detecting gene fusions in cancer and characterizing novel splice isoforms.

Protocol for RNA Seq Data Analysis

- ✓ RNA Seq analysis is multi step procedure.
- ✓ Different tools are required at each step.



✓We will use one of the galaxy server to perform RNASeq Data analysis

Galaxy Community

Galaxy COMMUNITY HUB | Use | Community | Education | Deploy & Develop | Support | Search Galaxy

Publicly Accessible Galaxy Servers

The Galaxy Project's public server (usegalaxy.org) can meet many needs, but it is not suitable for everything (see Choices for why). Fortunately the Galaxy Community is helping out by installing Galaxy at their institutions and then making those installations either publicly available or open to their organizations or community. This page lists publicly accessible Galaxy servers. To be included here a server must be accessible to any academic researcher anywhere in the world. Some servers may require logins and enforce quotas.

If you maintain a public instance of Galaxy it is recommended to sign up for the public servers mailing list to receive security fixes with priority.

To add your public Galaxy server to this list describe the server here and we'll post it this directory.

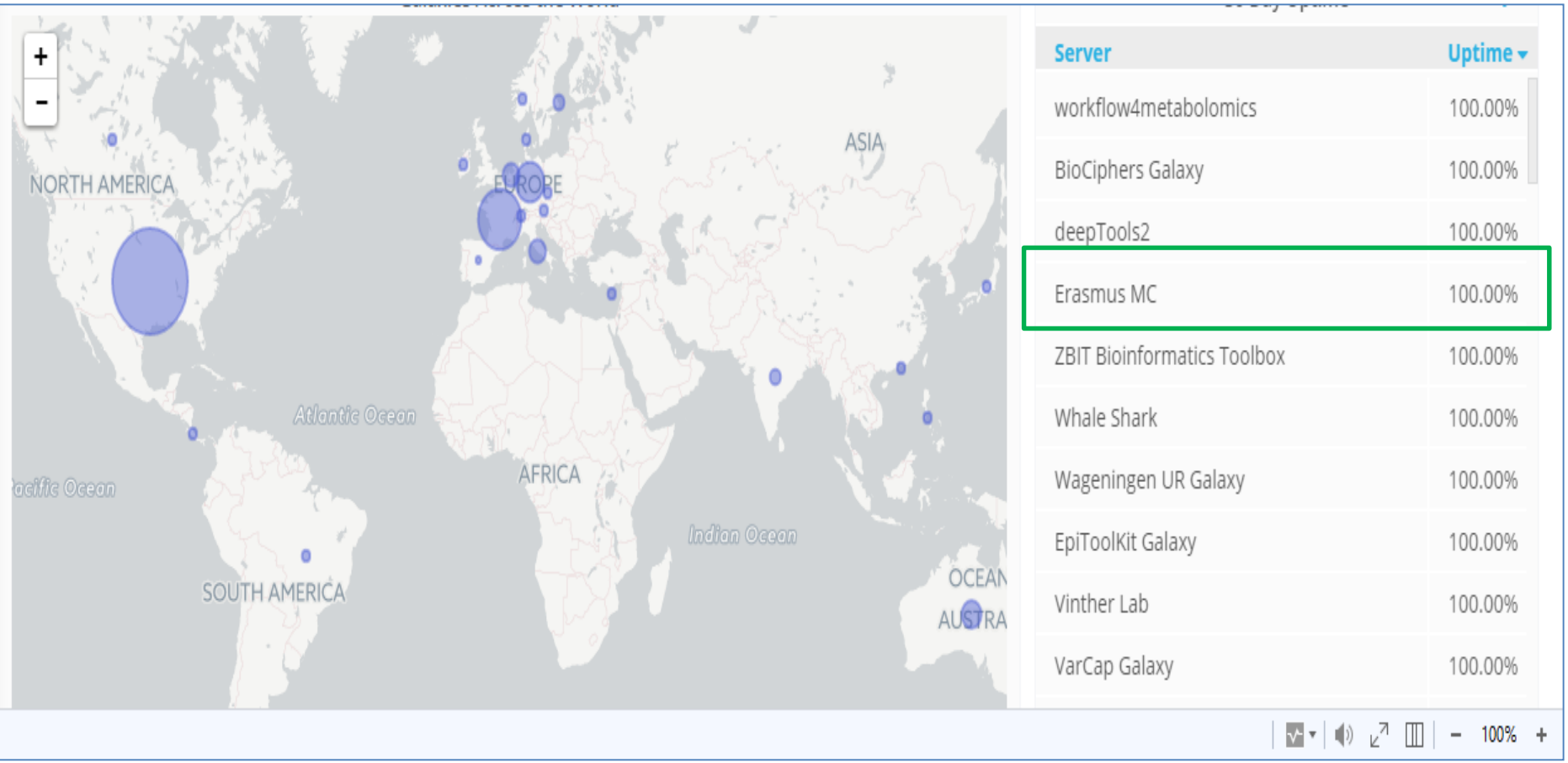
Galaxy Services

There are also a number of Galaxy services that make Galaxy available to research communities, sometimes restricted on a geographical or domain basis. See the Galaxy services list.

Public ToolSheds

In addition to the the main Galaxy and Test ToolSheds, several groups have made their tools available through publicly accessibly ToolSheds.

Galaxies across the World



GALAXY Server : Why to Use

✓The Galaxy Project's public server (usegalaxy.org) can meet many needs

✓General Purpose / Genomics Galaxy Servers

✓Domain Specific Galaxy Servers

✓Tool Publishing Galaxy Servers

Genomics Galaxy Server

These servers implement a broad range of tools and aren't specific to any part of the tree of life, or to any specific type of analysis. These are servers you can use when want to do general genomic analysis.

Name	Links	Summary
ABiMS	ABiMS Galaxy Request an account	General purpose genomics analysis, featuring many standard tools plus many additional tools. Howerver, we are specialized in RNASeq with reference and RNASeq denovo
Biomina	Biomina Galaxy	A general purpose Galaxy instance that includes most standard tools for DNA/RNA sequencing, plus extra tools for panel resequencing, variant annotation and some tools for Illumina SNParray analysis.
CBiB Galaxy	CBiB Galaxy	A general purpose Galaxy instance that includes EMBOSS (a software analysis package for molecular biology) and fibronectin (diversity analysis of synthetic libraries of a Fibronectin domain).
DBCLS Galaxy	DBCLS Galaxy	Adds text mining tools, DBCLS DBSearch Tools, semantic web tools
Erasmus MC	Erasmus MC Bioinformatics Galaxy Server	General purpose genomics analysis, featuring many standard tools plus many additional tools.
GalaxEast	GalaxEast Request an account	Integrative 'omics data analysis
Galaxy Main	Main	The Galaxy Project free public server; biomedical research
Galaxy Test	Galaxy Test	Beta version of Galaxy Main
Galaxy@GenOuest	Galaxy@GenOuest Request a GenOuest account	A general purpose Galaxy server Includes tools developped by Dyliss and GenScale bioinformatics research teams in Rennes, France.
Galaxy@Pasteur	Galaxy@Pasteur	General purpose genomics analysis server.
Galaxy@PRABI	Galaxy@PRABI PRABI Galaxy Tool Shed	Includes bioinformatics tools developed by the research teams working in the perimeter of the PRABI core facility, including *kissplice/kissDE, TETools, SEX-DETector,* and *priam*.
GigaGalaxy	GigaGalaxy	Standard Galaxy tools set plus SOAPdenovo and SOAPsnp for *de novo* assembly and SNP calling.
GVL MEL	Galaxy Melbourne	General purpose Galaxy based on the Genomics Virtual Lab platform.
GVL QLD	Genomics Virtual Lab GVL-QLD	General purpose Galaxy based on the Genomics Virtual Lab platform.
GVL Tutorial	Genomics Virtual Lab GVLTut	Small Galaxy for Training purposes. Loaded with Histories and Tools for Next Gen Sequencing tutorials.

Domain specific Galaxy Server

Domain servers specialize in either a particular branch of the tree of life or in particular types of analysis. However, within their specializations, domain servers offer a wide variety of tools.

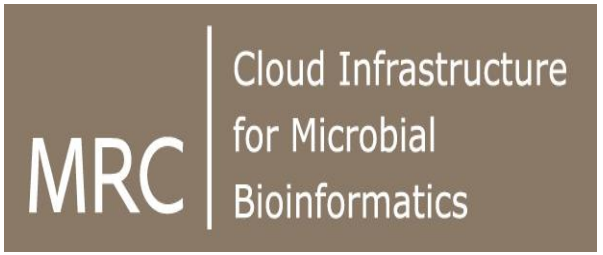
Name	Links	Summary
ballaxy	ballaxy Galaxy server ballaxy using Docker	Hosts the BALL (Biochemical Algorithms Library) Project tools, i.e. computer aided drug design and molecular modelling based on protein and ligand structure data.
BIPAA (BioInformatics Platform for Agroecosystem Arthropods)	BIPAA Galaxy Server BIPAA home page	Insect genomics (aphids, parasitoid wasps, lepidopterans)
Center for Phage Technology (CPT)	Center for Phage Technology (CPT) Galaxy Server CPT home page	Phage biology and automated annotation.
Cistrome Analysis Pipeline	Cistrome Analysis Pipeline	ChIP-chip/seq and gene expression data
CoSSci	CoSSci Complex Social Science Gateway	Tools for solving Galton's problem in Comparative Research and complex network problems in Social Science.
Dintor	Dintor: Data Integrator Tool Suite	GWA and NGS tools and modules for functional annotation of genes and gene products
Galaxy Integrated Omics (GIO)	GIO Server	Proteomics Informed by Transcriptomics (PIT) methodology, and selection of surrogate peptides for targeted proteomics.
Galaxy PGTB (Virtual Biodiversity Lab)	PGTB Galaxy - Virtual Biodiversity Lab Plateforme Genome Transcriptome de Bordeaux	This is a standard Galaxy instance implemented with specific tools for Biodiversity (Biodiversity Virtual Lab) and NGS (Ion Torrent from the PGTB facility) analysis.
Galaxy-CEFAP	Galaxy-CEFAP	Galaxy-CEFAP offers a set of tools to perform RNA-Seq and miRNA analysis.
Galaxy-P	Use Galaxy-P	Galaxy-P is a multiple 'omics' data analysis platform with particular emphasis on mass spectrometry based proteomics. Galaxy-P is developed at the University of Minnesota, deployed at the Minnesota Supercomputing Institute.
Genomic Hyperbrowser	Genomic Hyperbrowser	statistical methodology and computing power to handle a variety of biological inquires on genomic datasets
GrAPPA	Graph Algorithms Pipeline	GrAPPA is a web-based interface constructed on the Galaxy framework for graph theoretical tools. It contains novel

Galaxy Services : Example

- ✓ Geography Based
- ✓ Domain Based

Australia: Genomics Virtual Lab (GVL)

Canada: GenAP



United Kingdom: CLIMB

Poland: PL-Grid

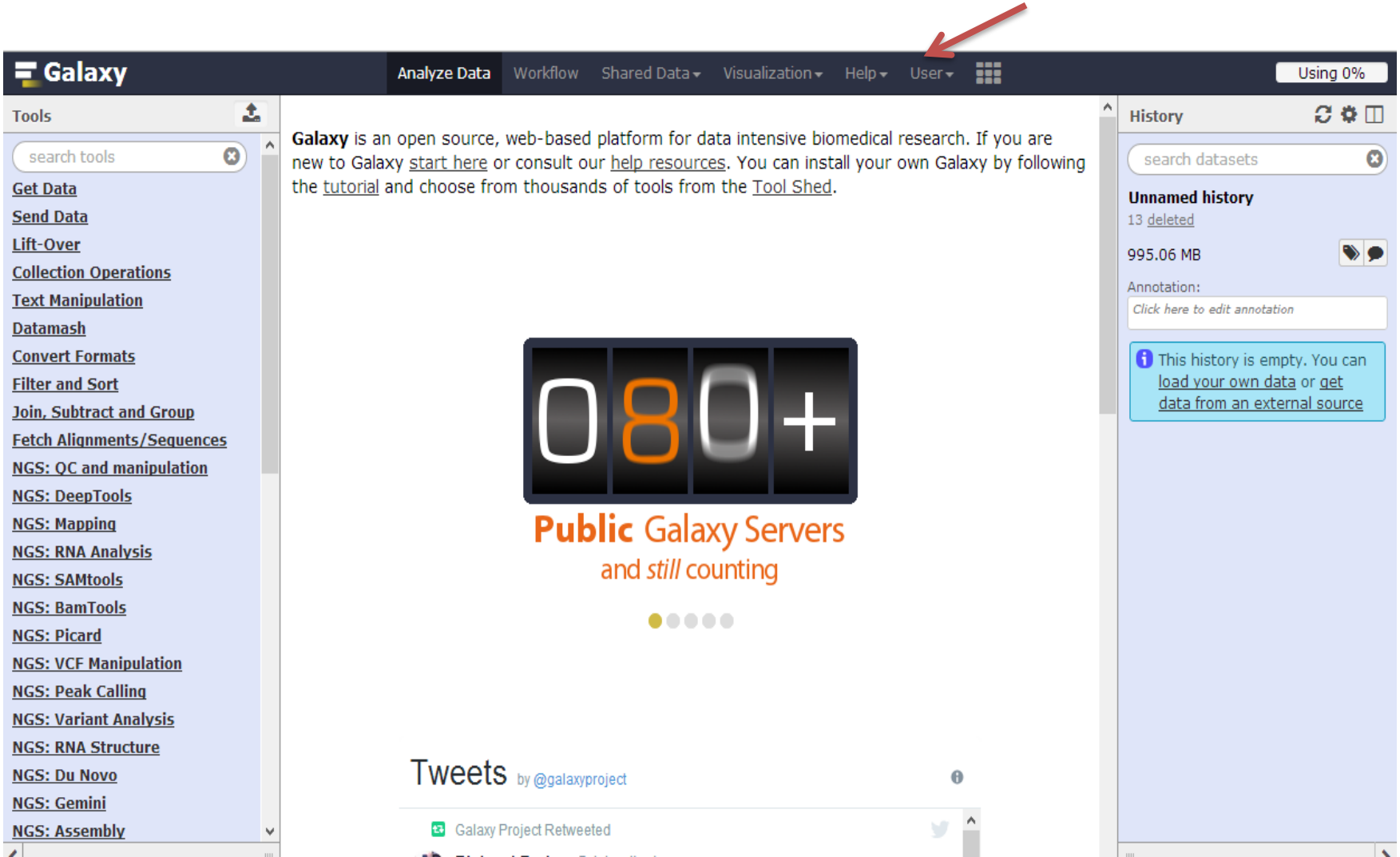
Norway: NeLS

Cancer: Cancer Computer

Galaxy Main Tool Shed

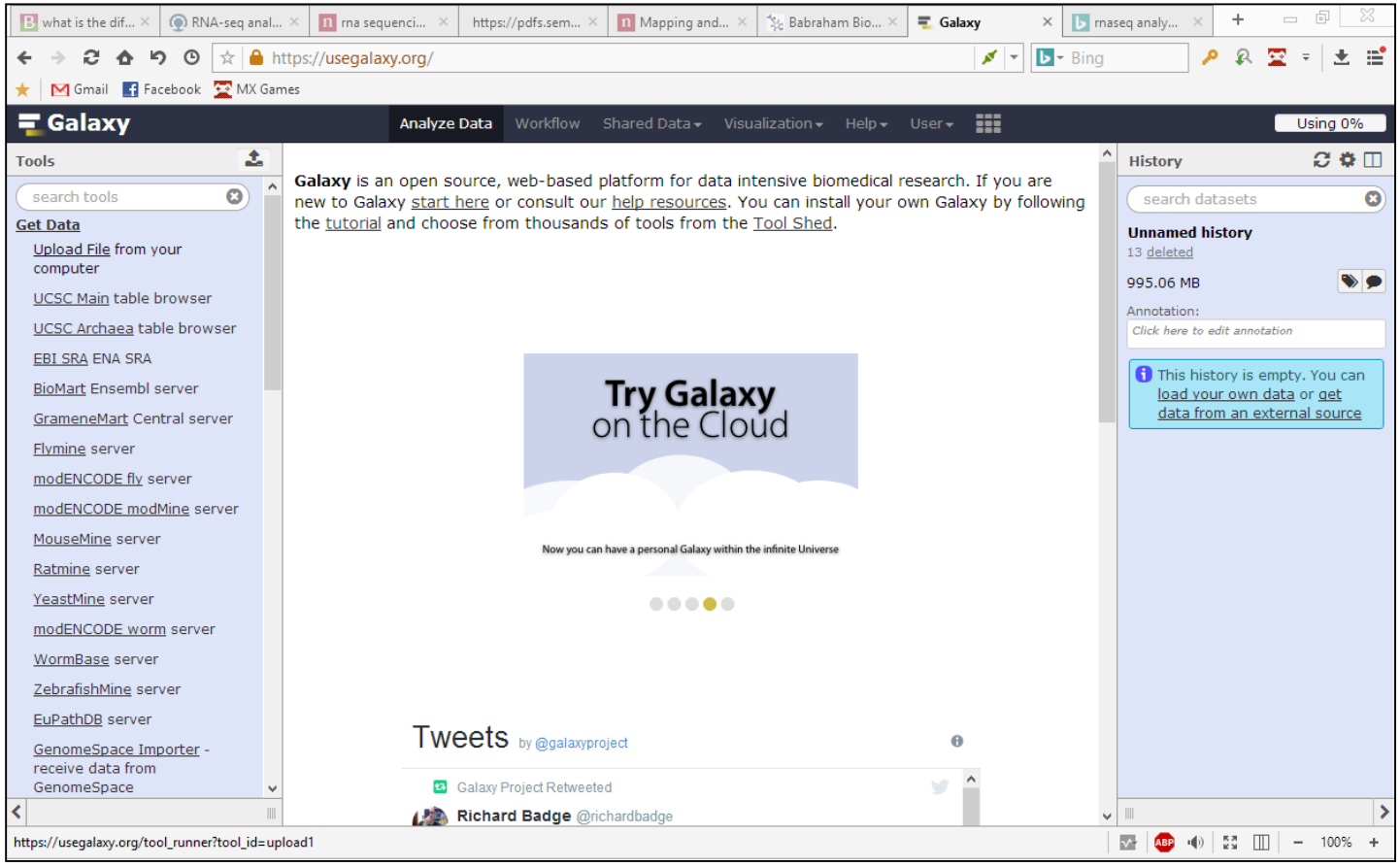
Let Us Use Public GALAXY Server

Go to https://usegalaxy.org/ and create login



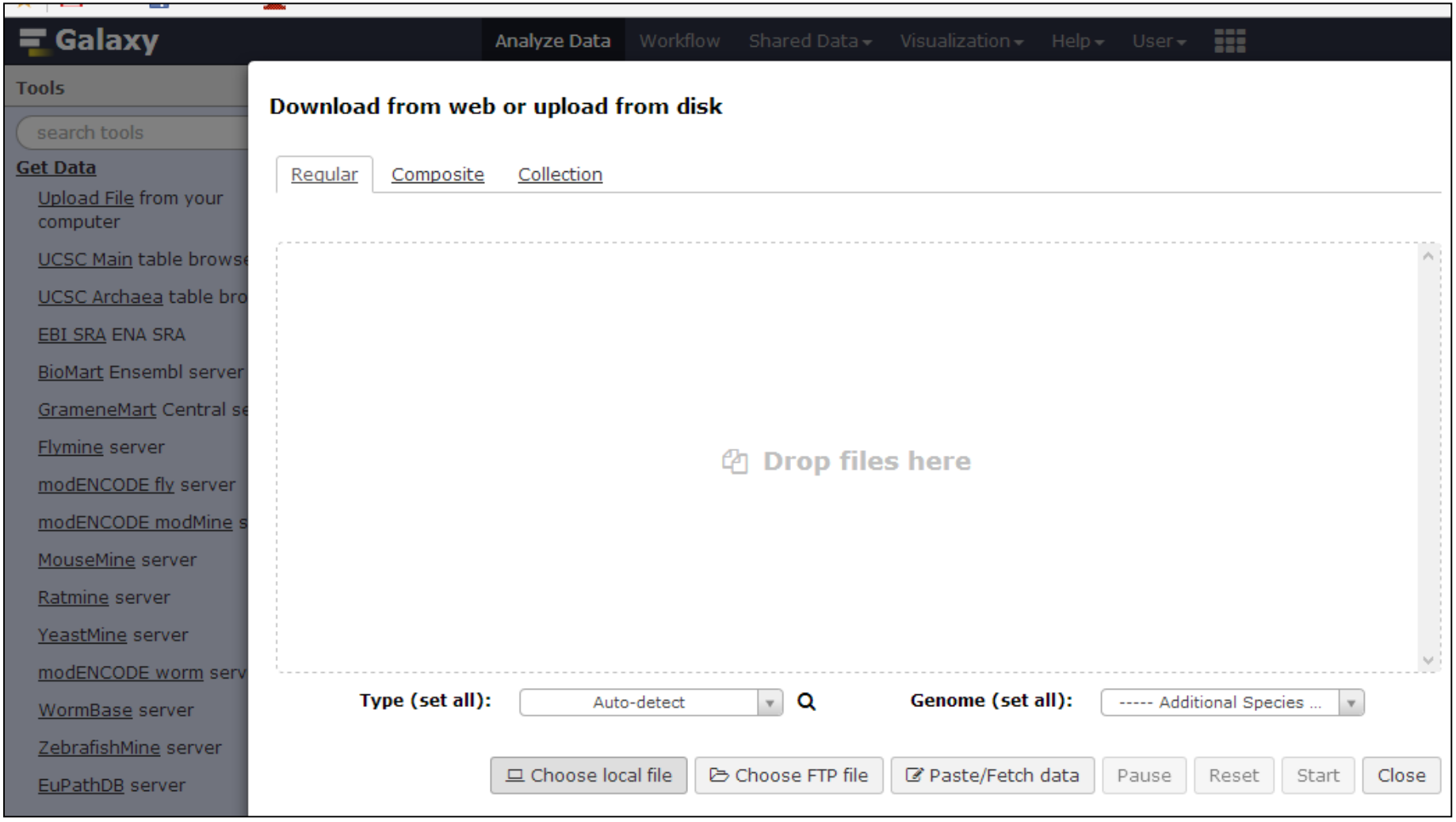
GALAXY Server : Upload Data (I)

✓Click on Get Data and select Upload File from your computer. Download samples files from course website.



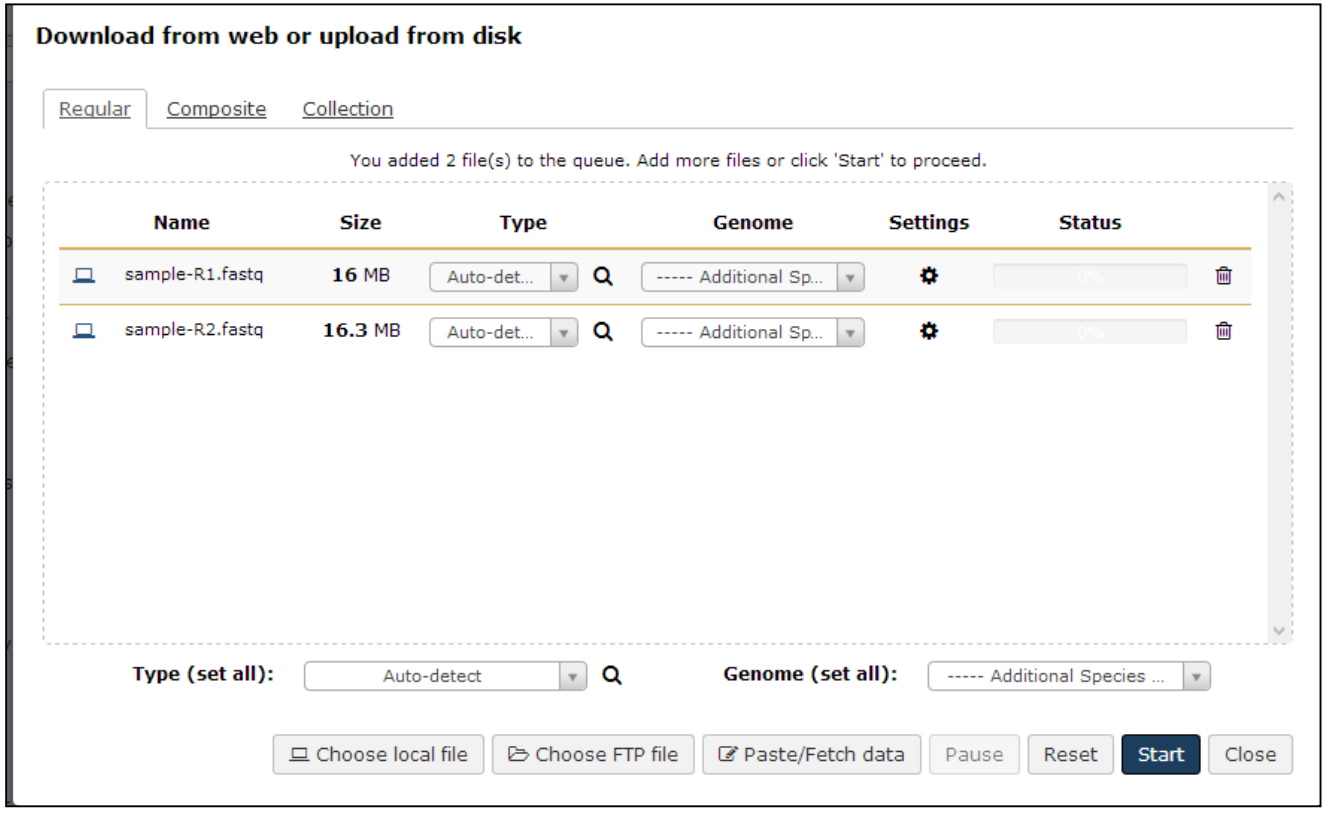
GALAXY Server : Upload Data (II)

New Window will appear. Now, Click option “Choose local file”



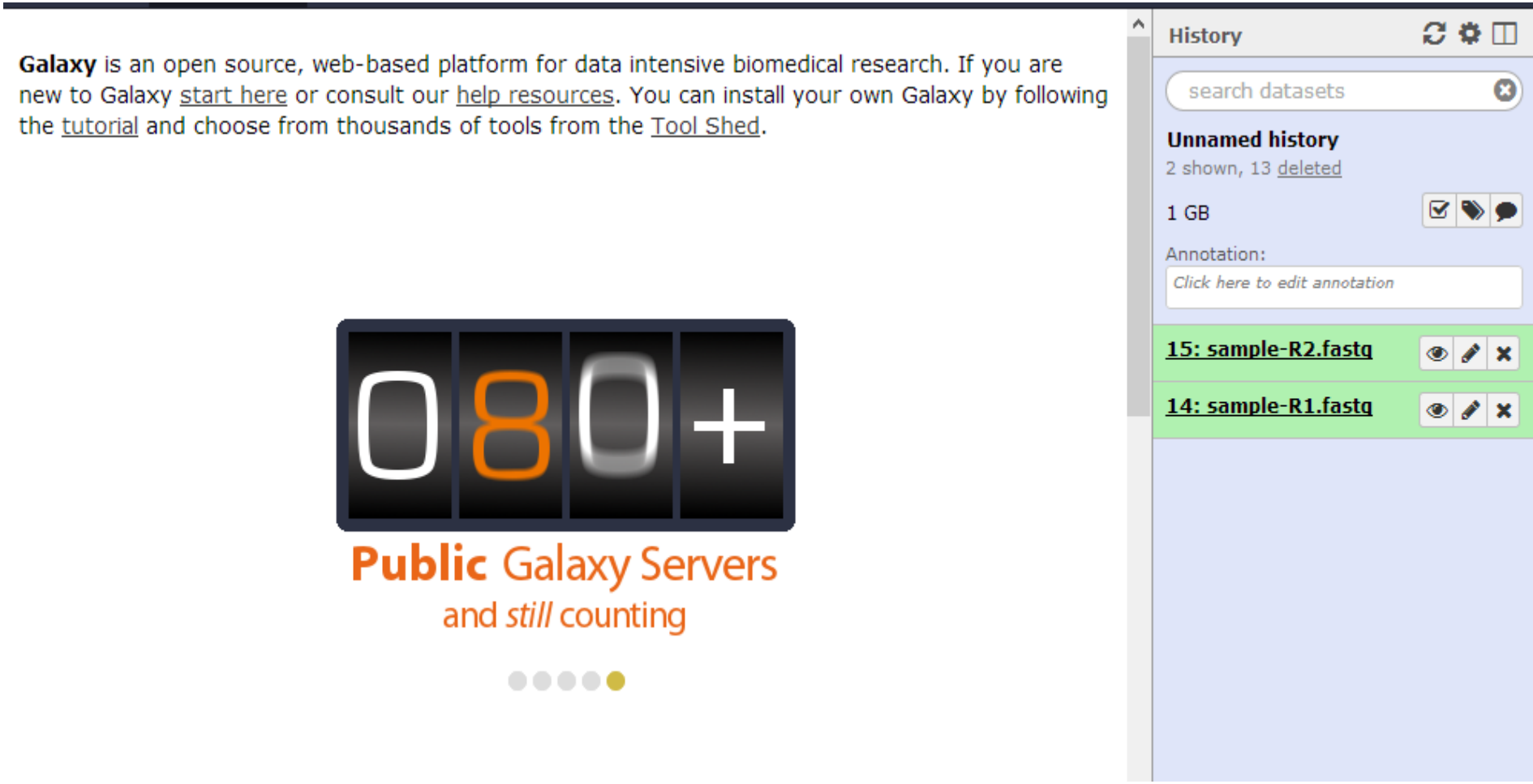
GALAXY Server : Upload Data (III)

✓Now, Click option “Start”. It will upload file to server.



✓Now wait for 10-20 seconds.

✓**Files will be uploaded successfully and appears with green colour .**



✓Galaxy consist of collection of Multiple Tools . Today's session, We will use

✓NGS: QC and manipulation

✓NGS Mapping

✓NGS:RNA analysis

✓SAMtools

Protocol for RNA Seq Data Analysis

1. Pre-processing

2. Quality Filtration

3. Mapping or assembly

4. Expression analysis

Quality Assessment

✓**It is important to check the quality of your sequenced reads**

✓**FASTQC: free program that reports quality profile of reads**

Quality Assessment

✓**Modern high throughput sequencers can generate hundreds of millions of sequences in a single run.**

✓**Before analysing this sequence to draw biological conclusions you should always perform some simple quality control checks to ensure that**

(I) the raw data looks good and

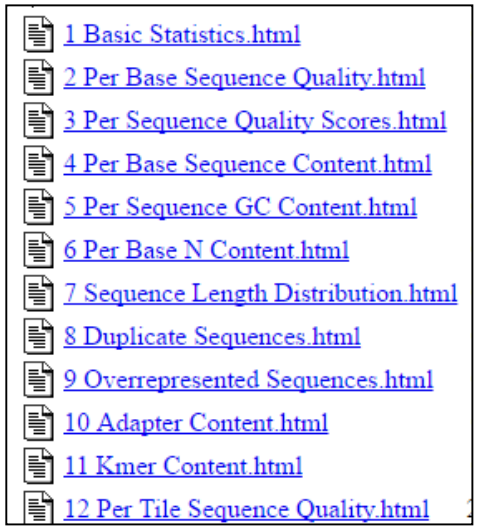
(II) there are no problems or biases in your data

which may affect how you can usefully use it.

FASTQC tool

✓Providing a quick overview to tell you in which areas there may be problems

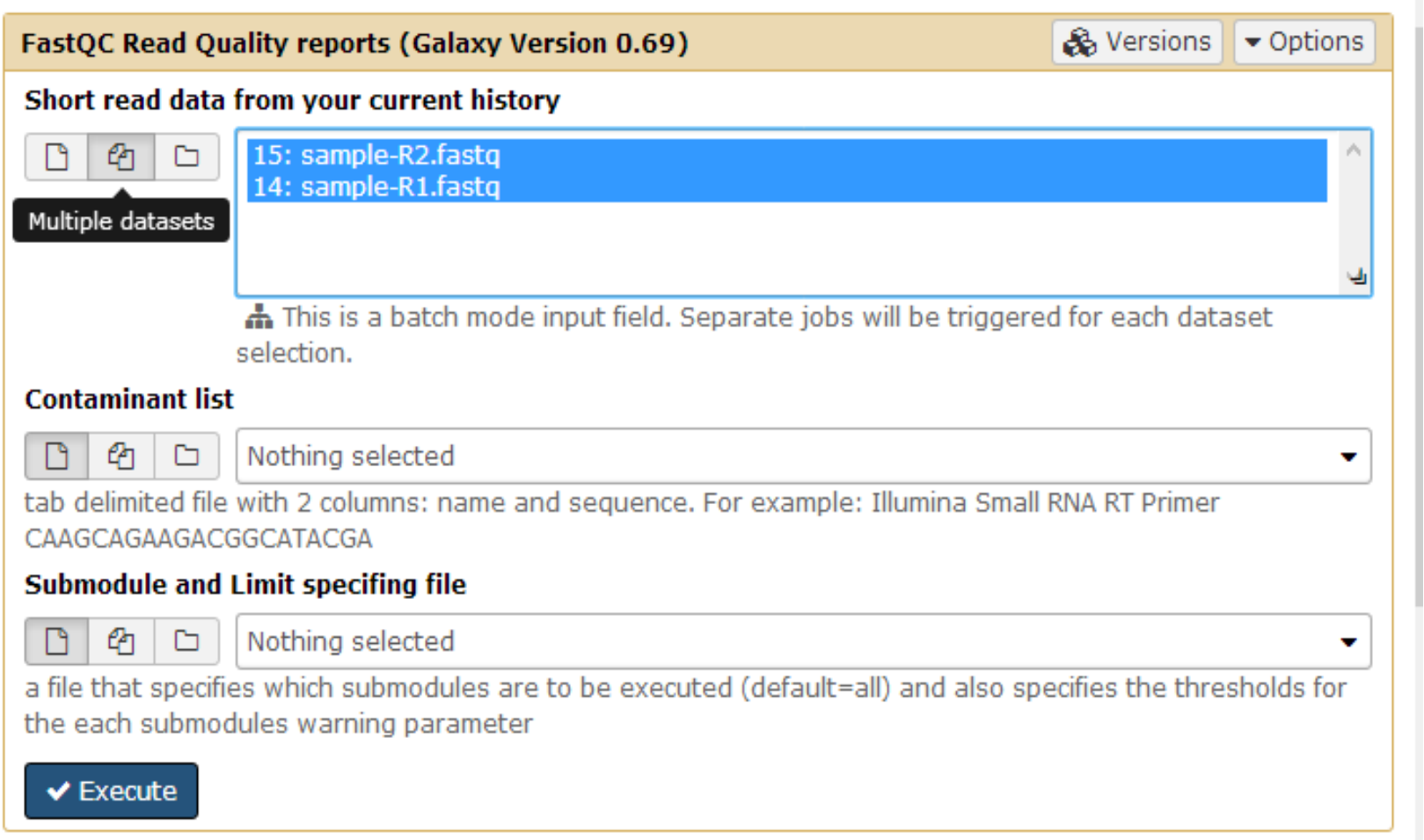
✓Summary graphs and tables to quickly assess your data



✓Double click on NGS: QC and manipulation

✓Select application Fastqc in Galaxy

✓Select Multiple Dataset to run multiple files and press “Execute”

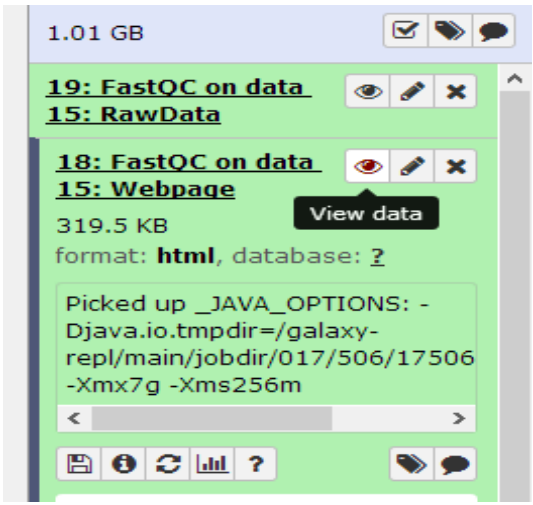


✓You will get two types of output files :

(I) Raw data – It consist of text description

(II) Web page – It consist of detail graphical representation of your fastq data.

✓Click on “eye” symbol to view output files.



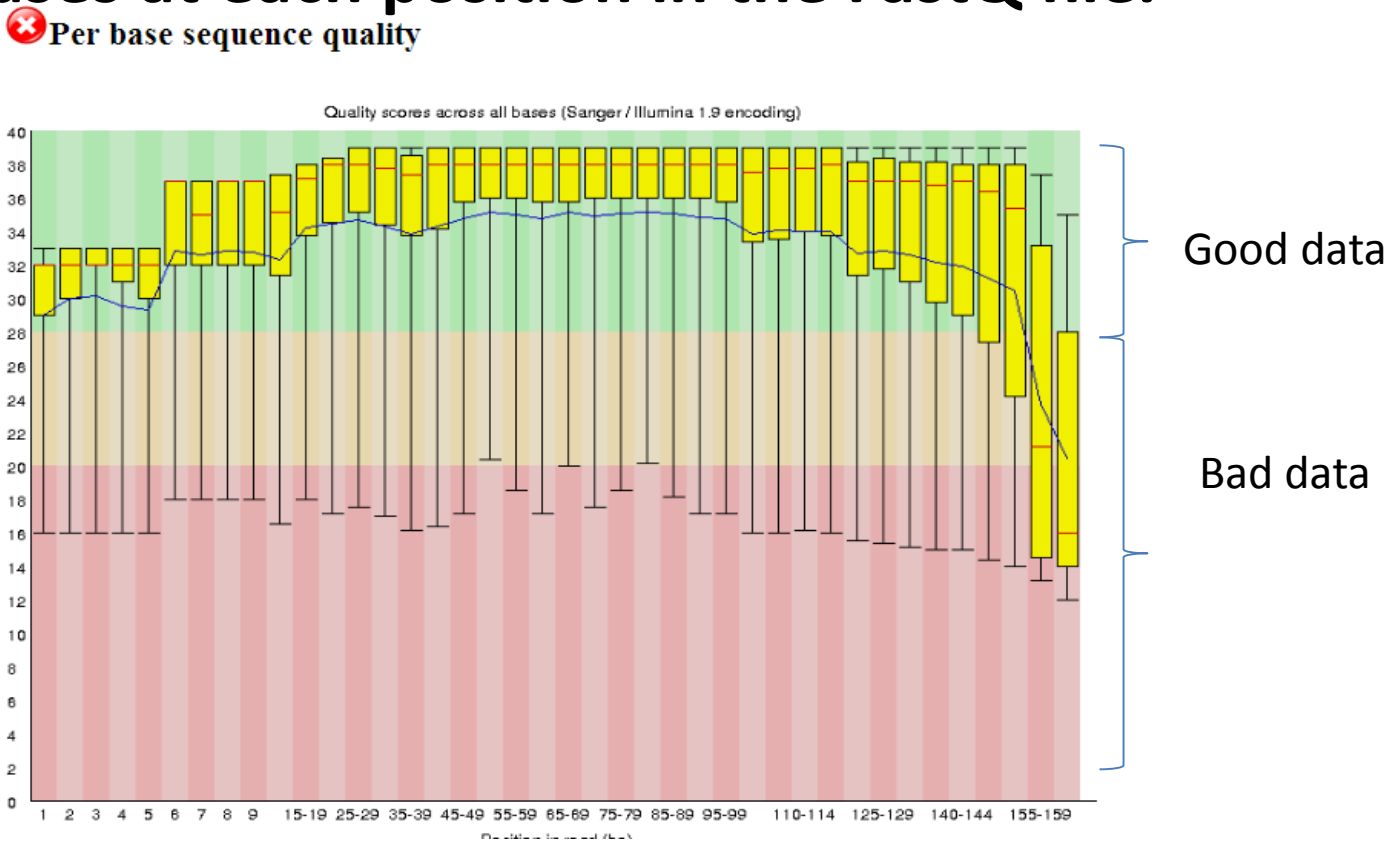
The left hand side of the main interactive display or the top of the HTML report shows a summary of the modules as normal (green tick), slightly abnormal (orange triangle) or very unusual (red cross).

Basic Sample Statistics

Measure	Value
Filename	sample-R2_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	43435
Sequences flagged as poor quality	0
Sequence length	154-160
%GC	47

Per Base Sequence Quality

This view shows an overview of the range of quality values across all bases at each position in the FastQ file.



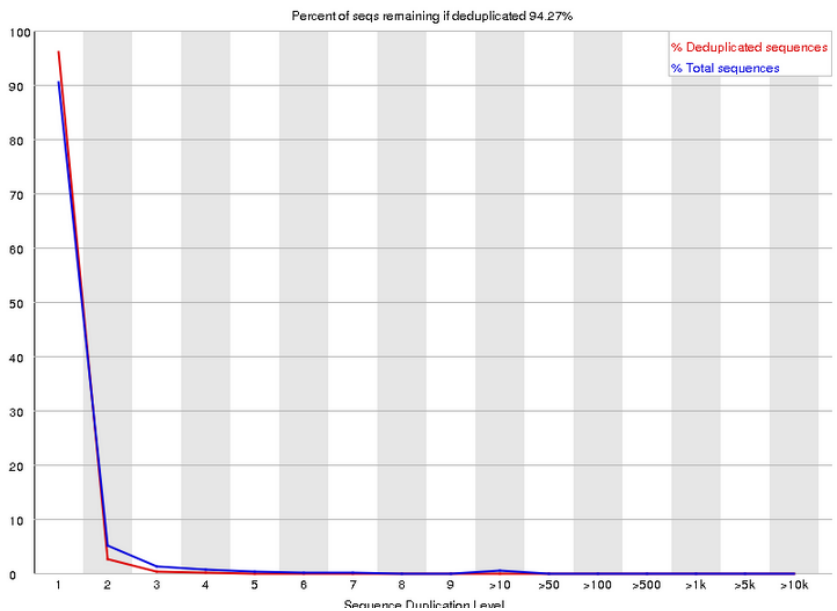
✓**We must consider threshold of Quality : Q30 or above . This graph indicate we need to perform filtration on our data.**

✓**By looking at figure, we can say that there is problem in bases in position 140-150. It can be fixed during quality filtration step.**

Duplicate Sequences

✓A low level of duplication may indicate a very high level of coverage of the target sequence

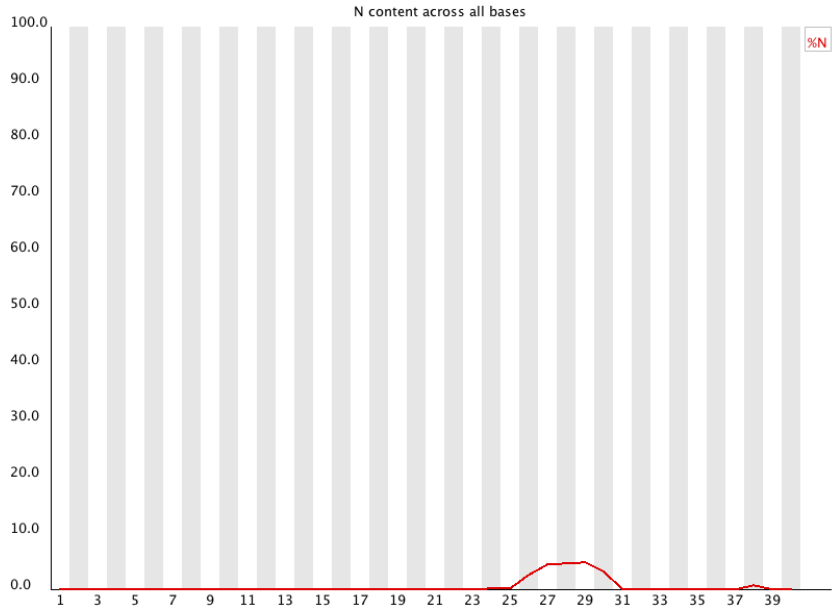
✓A high level of duplication is more likely to indicate some kind of enrichment bias (eg PCR over amplification).



✓High duplication could affect the mapping efficiency and bias your interpretation.

Per Base N Content

✓If a sequencer is unable to make a base call with sufficient confidence then it will normally substitute an N rather than a conventional base] call.



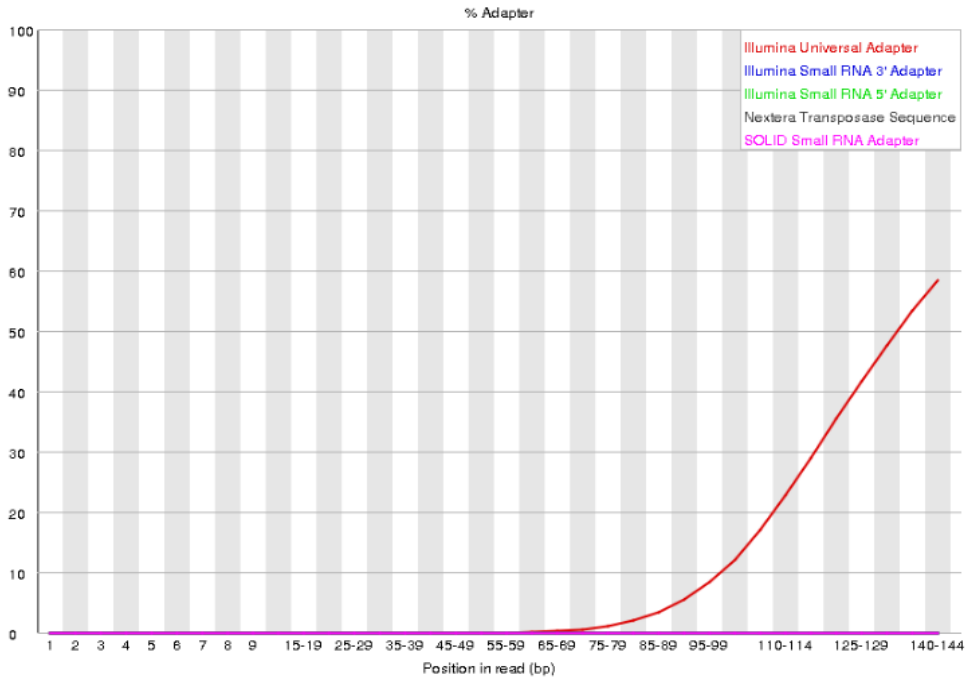
✓This module plots out the percentage of base calls at each position for which an N was called.

Adapter : Trimming

- Universal adapter
- DNA Fragment of Interest
- Index Adapter
- 6 Base index region

Adapter Content

✓If we know the adapter sequence, we can trim it using Trimmomatic tool.



✓To get the adapter sequence information, one can contact person who performed the sequencing and can get full detail of “Adapter sequences”.

Protocol for RNA Seq Data Analysis

1. Pre-processing

2. Quality Filtration

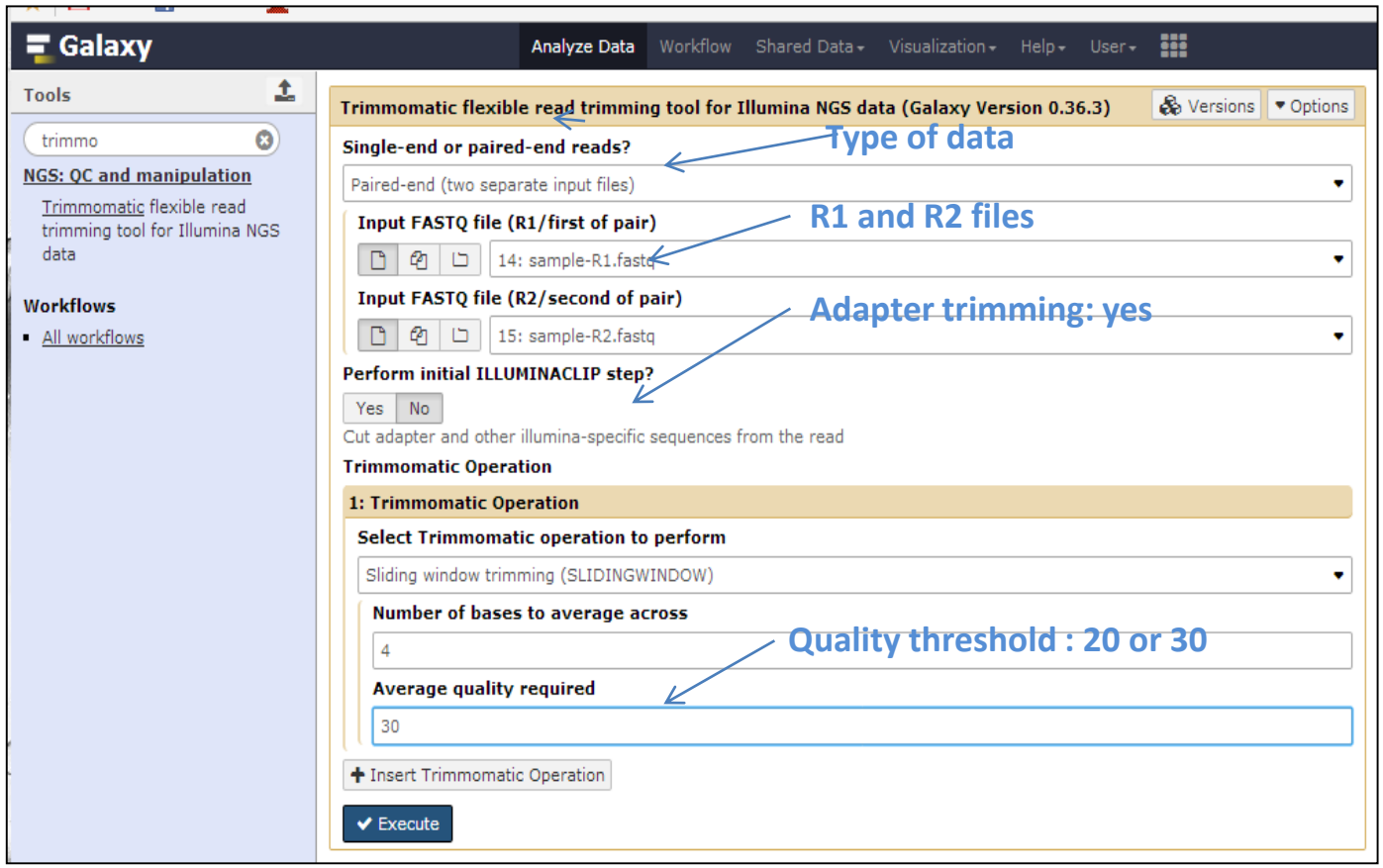
3. Mapping or assembly

4. Expression analysis

Quality Filtration

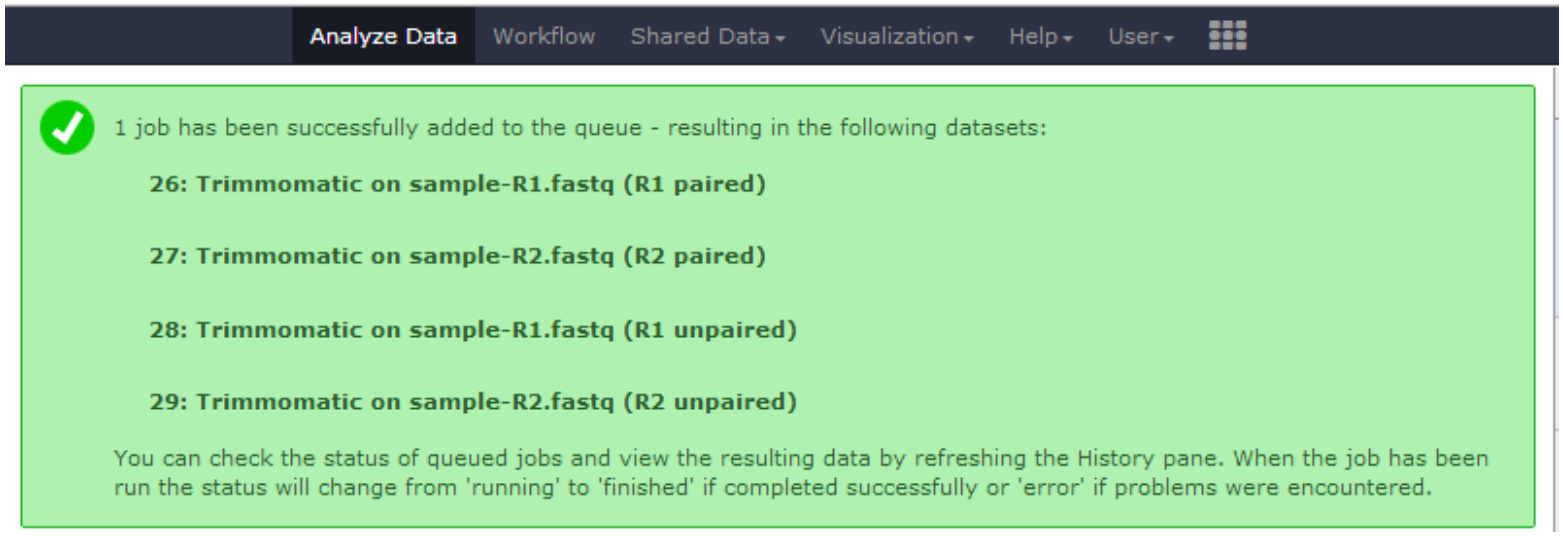
✓Goals : To improve the quality of Data

Trimmomatic :



Quality Filtration

✓**Trimmomatic will produce four output files.**

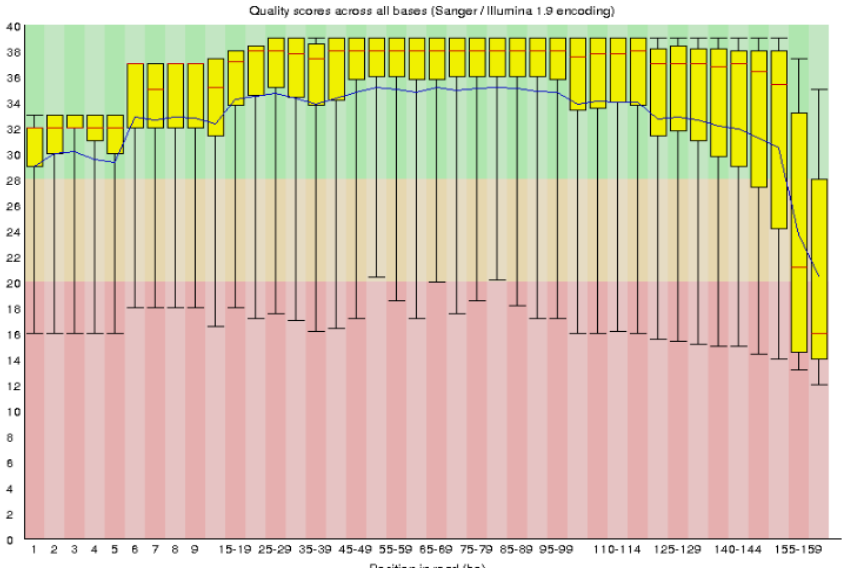


✓**For next analysis, we will consider only R1 paired and R2 paired data While unpaired reads will be discarded.**

✓**Rerun the Fastqc on paired end R1 and R2 paired end files and check statistical output.**

Let us Do Comparison of dataset
Before and After Quality filtration

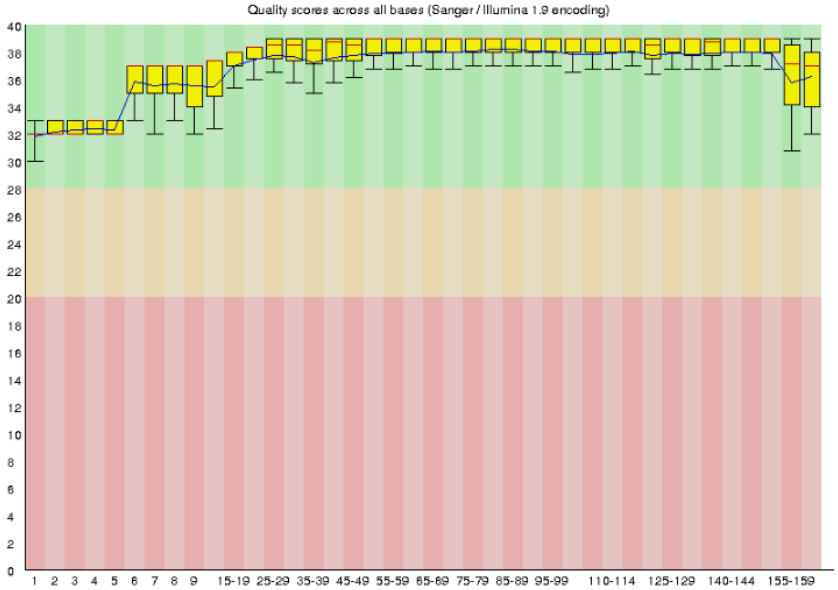
Per base sequence quality



Before Quality filtration :

Bad Data

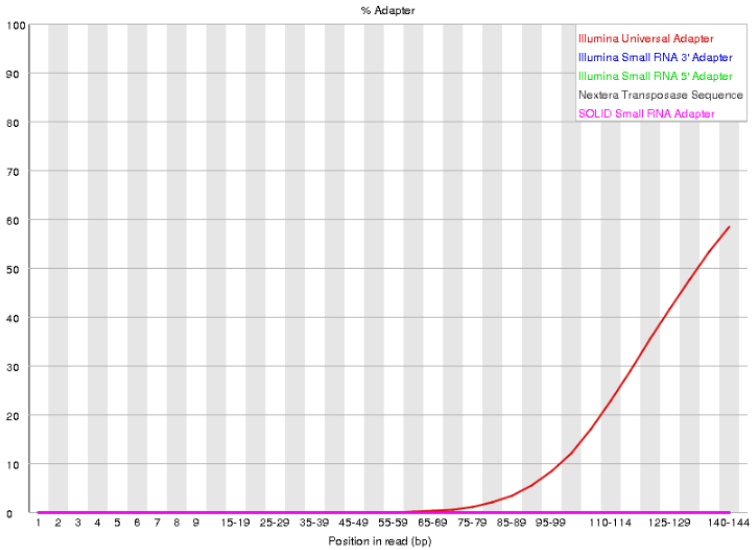
Per base sequence quality



After Quality filtration

: Good Data

Adapter Content



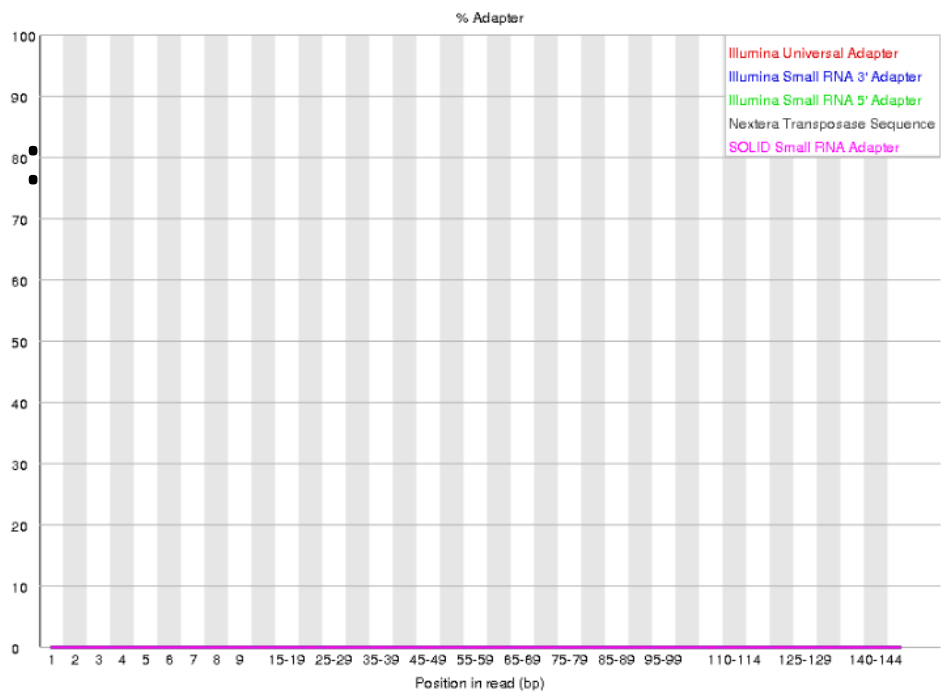
Before Quality filtration :

Adapter contamination

Before Quality filtration :

No Adapter contamination

Adapter Content



Questions ?

Protocol for RNA Seq Data Analysis

1. Pre-processing

2. Quality Filtration

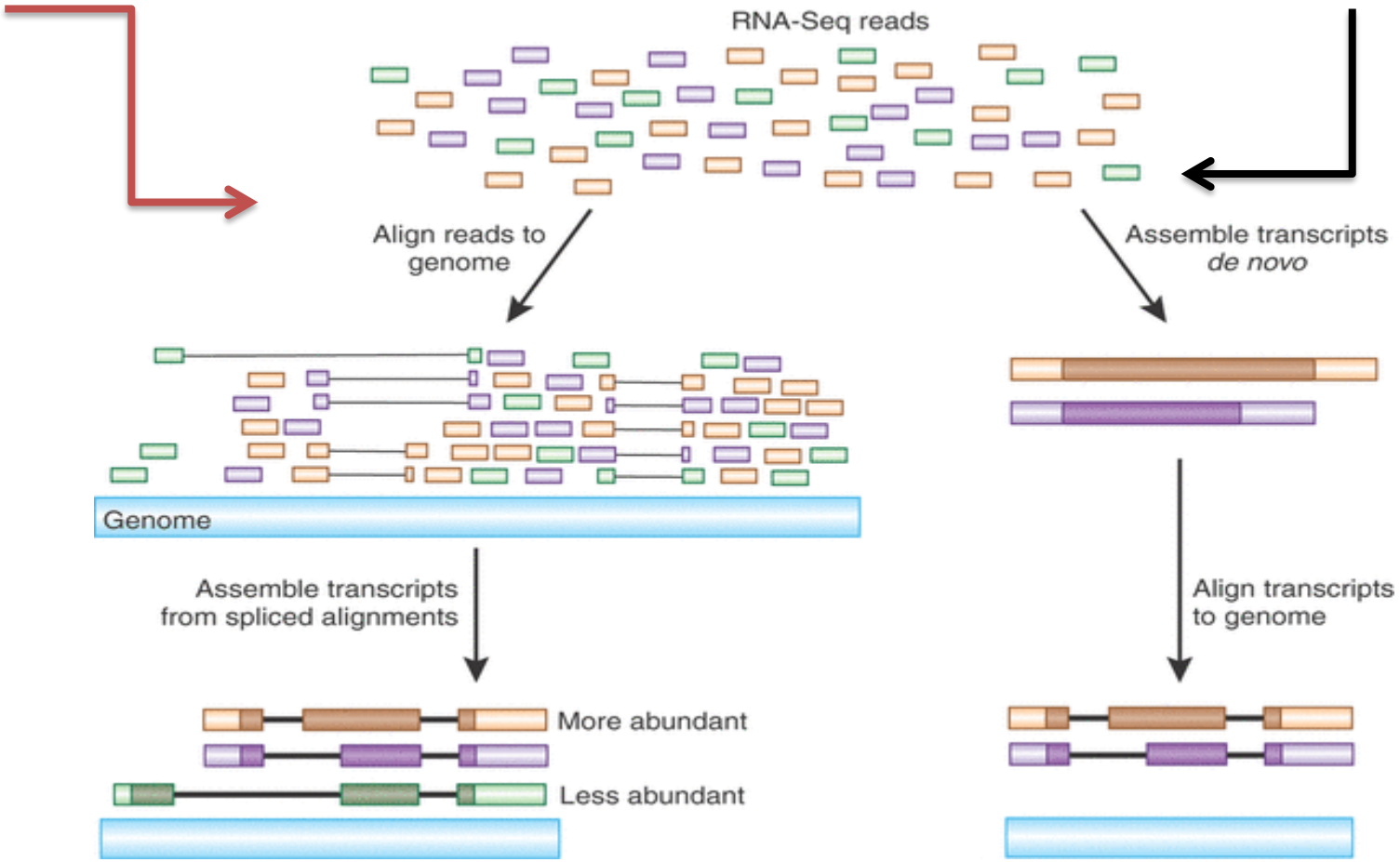
3. Mapping or assembly

4. Expression analysis

How to decide : Mapping or assembly?

If reference genome is available

If reference genome is not available



Mapping tool: Bowtie

✓**Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.**

✓**It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes.**

✓**Bowtie supports gapped, local, and paired-end alignment modes.**

Bowtie : End to End Alignment

The following is an "end-to-end" alignment because it involves all the characters in the read. Such an alignment can be produced by Bowtie 2 in either end-to-end mode or in local mode.

```
Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG

Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

Bowtie : Local Alignment

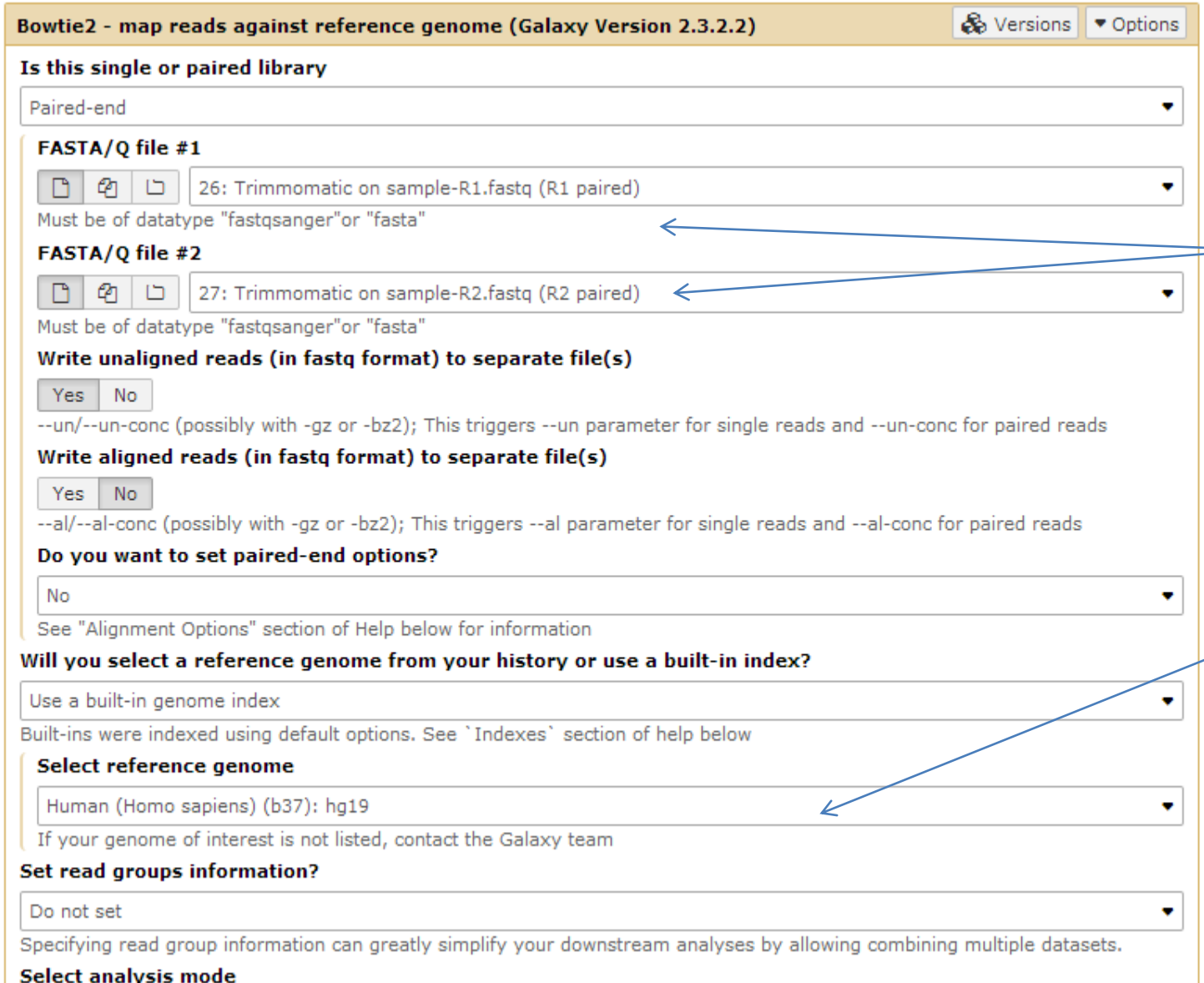
The following is a "local" alignment because some of the characters at the ends of the read do not participate. In this case, 4 characters are omitted (or "soft trimmed" or "soft clipped") from the beginning and 3 characters are omitted from the end. This sort of alignment can be produced by Bowtie 2 only in local mode.

```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG

Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

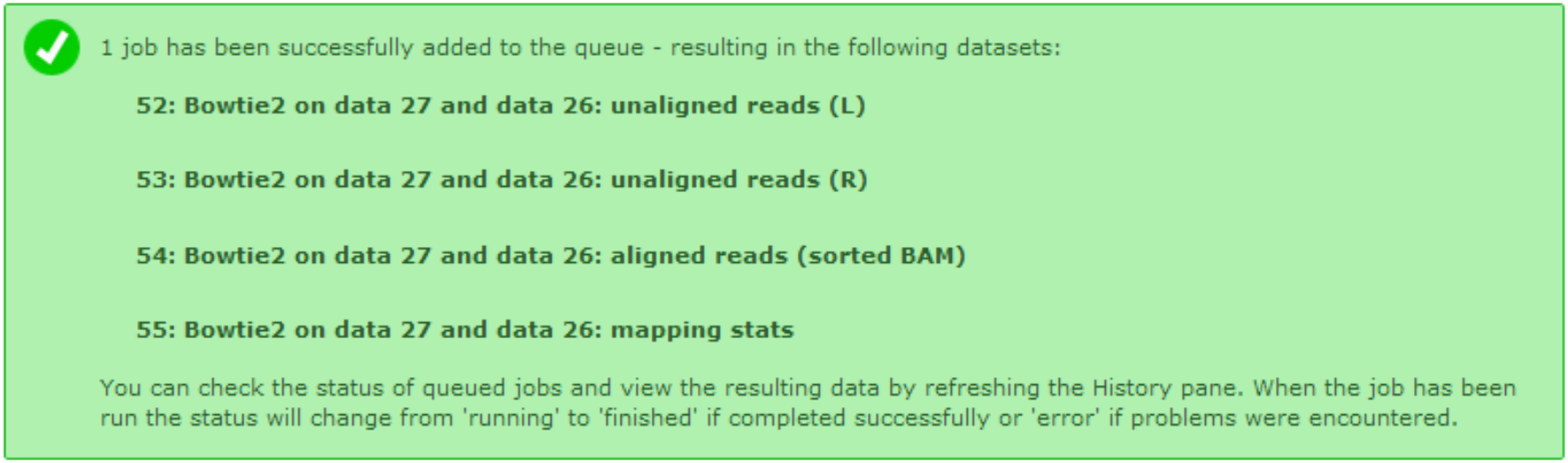
✓**Mapping quality: higher = more unique ☺**

Reference Mapping : Bowtie



Output from trimmomatic

Species name : Human



✓**It will produce the multiple output files. BAM file consist of complete mapping information which stores the same data in a compressed, indexed, binary form.**

✓**The SAM Format is a text format for storing sequence data in a series of tab delimited ASCII columns.**

Mapping Statistics

```
7106 reads; of these:
  7106 (100.00%) were paired; of these:
    2828 (39.80%) aligned concordantly 0 times
    2360 (33.21%) aligned concordantly exactly 1 time
    1918 (26.99%) aligned concordantly >1 times
    ----
    2828 pairs aligned concordantly 0 times; of these:
      324 (11.46%) aligned discordantly 1 time
    ----
    2504 pairs aligned 0 times concordantly or discordantly; of these:
      5008 mates make up the pairs; of these:
        2949 (58.89%) aligned 0 times
        720 (14.38%) aligned exactly 1 time
        1339 (26.74%) aligned >1 times
79.25% overall alignment rate
```

Uniquely mapped

Multi mapped

✓**Uniquely mapped – Reads mapped one time over the reference genome**

✓**Multi mapped - Reads mapped more than one time over the reference genome**

Which Information is in SAM & BAM

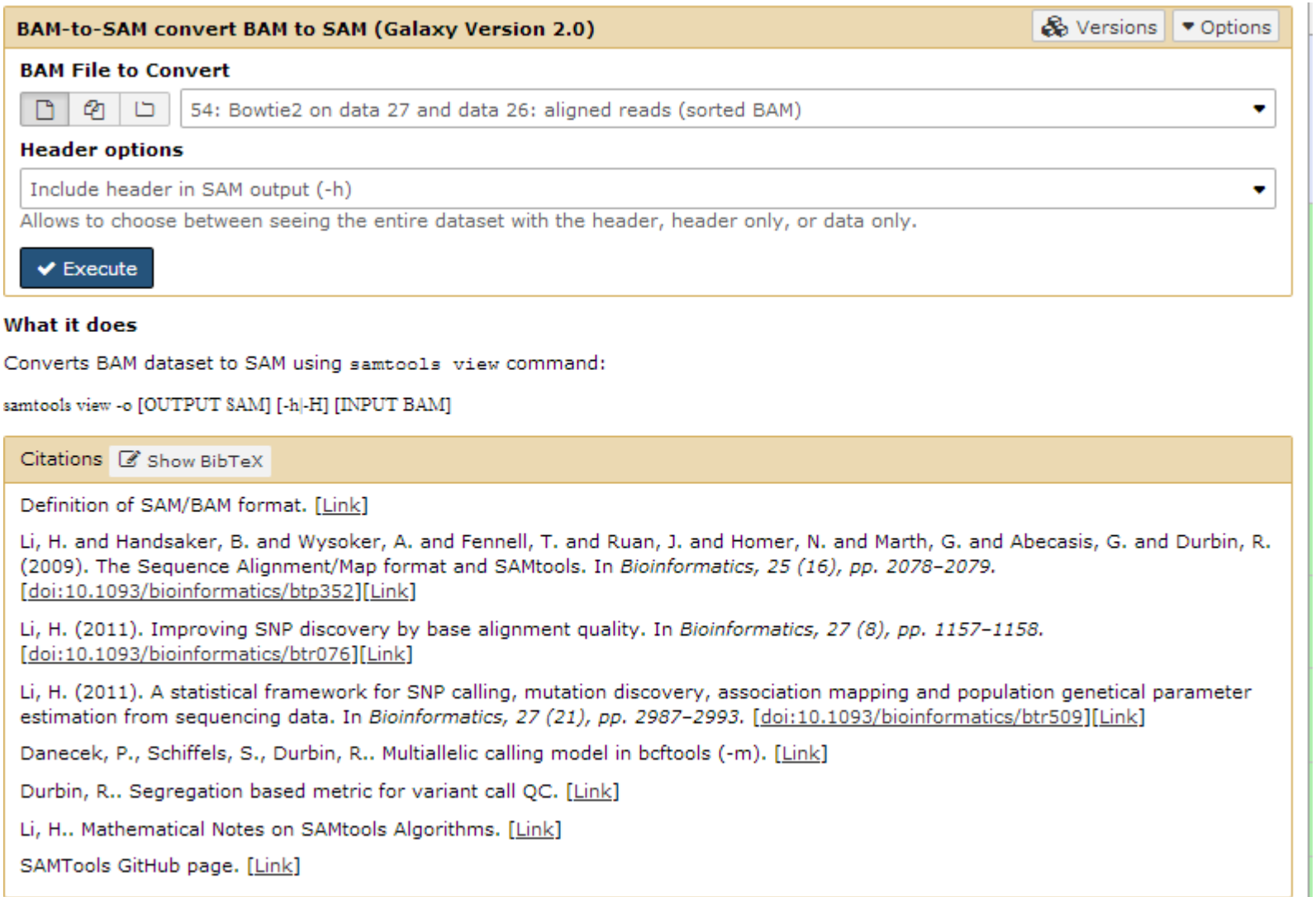
✓**Both SAM & BAM files contain an optional header section followed by the alignment section.**

✓**The header section may contain information about the entire file.**

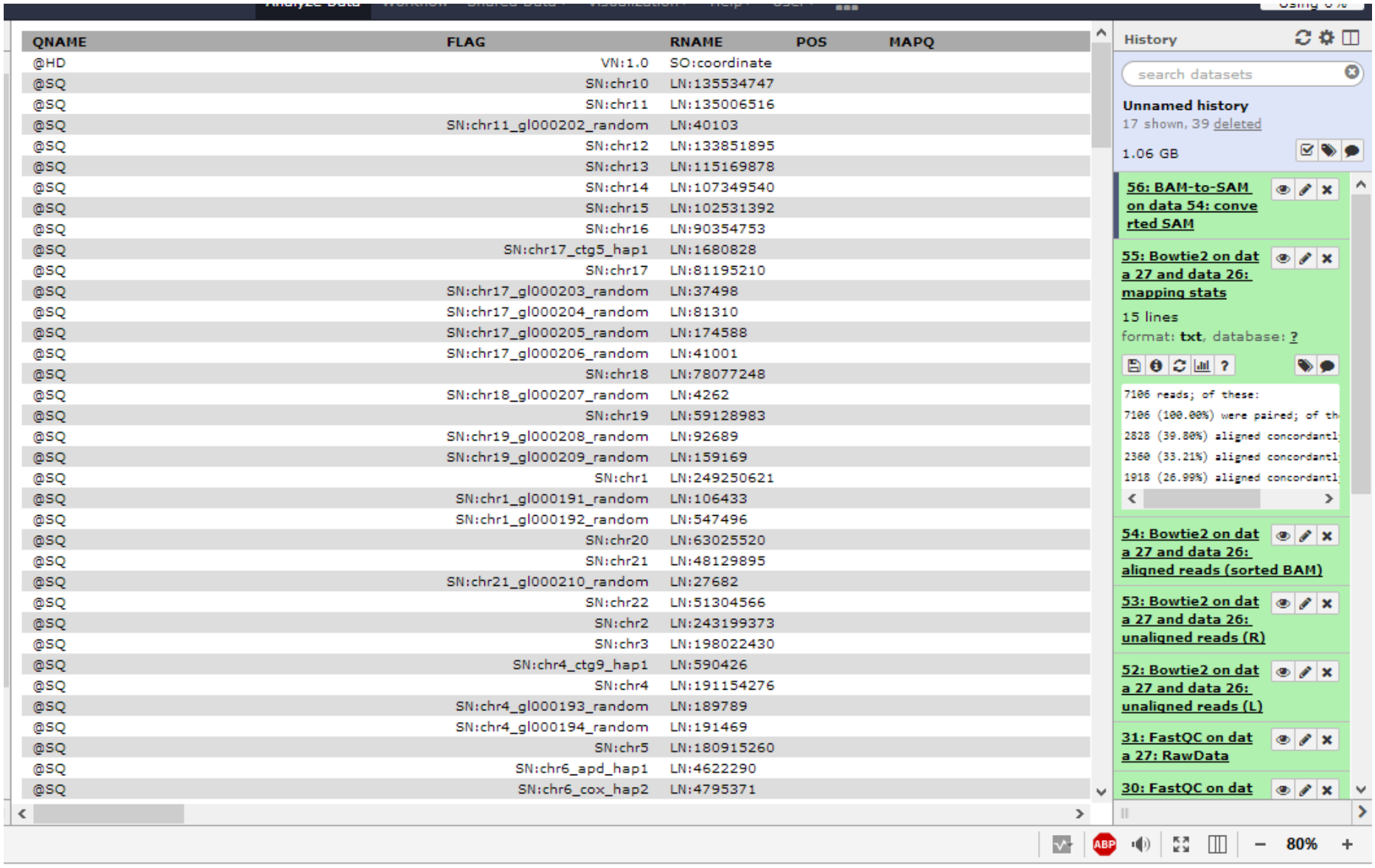
✓**The alignment section contains the information for each sequence about where/how it aligns to the reference genome.**

Let us convert BAM to SAM

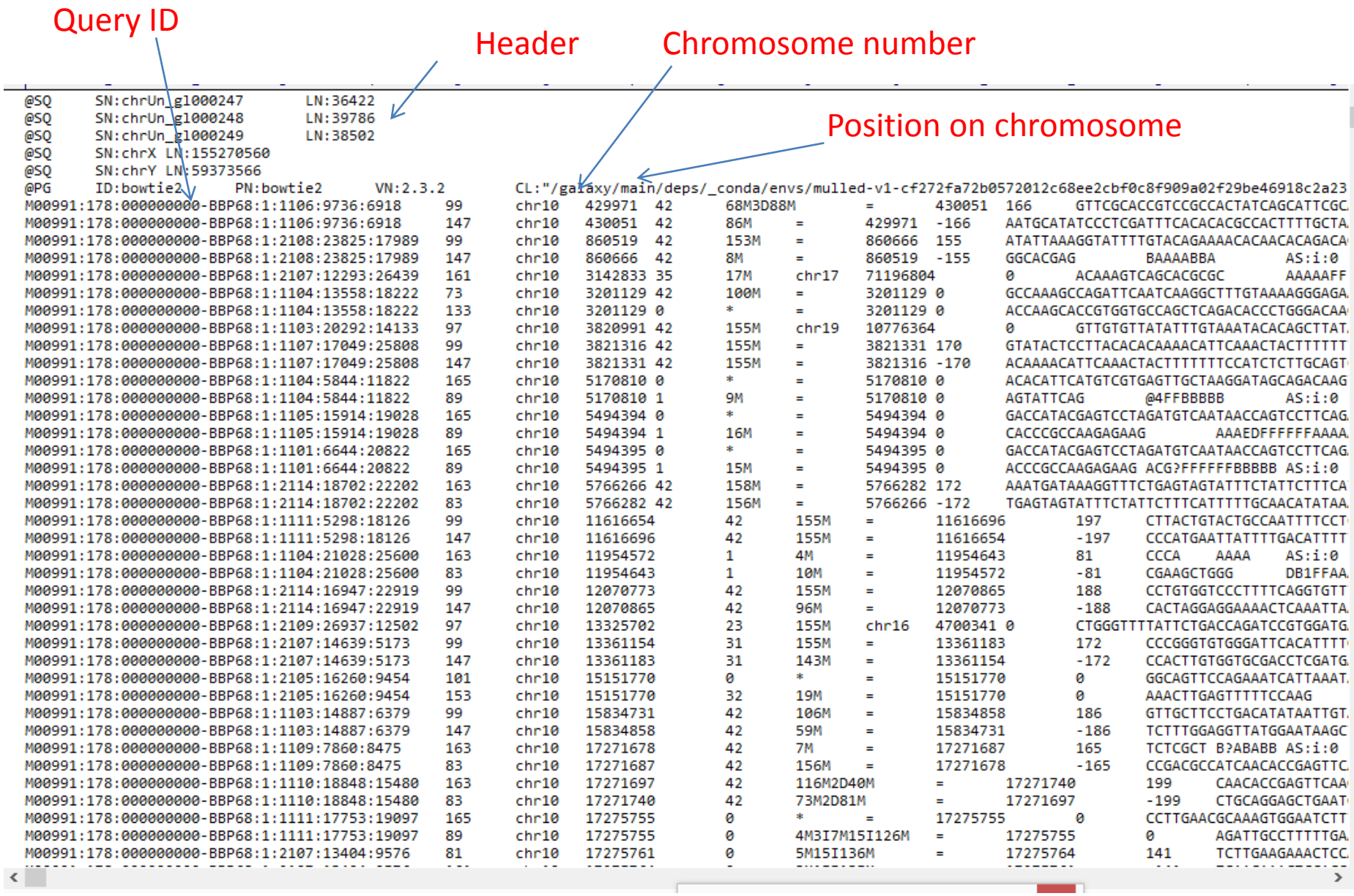
✓Select BAM to SAM tool under samtools



SAM File : Mapping information



QNAME	FLAG	RNAME	POS	MAPQ
@HD	VN:1.0	SO:coordinate		
@SQ	SN:chr10	LN:135534747		
@SQ	SN:chr11	LN:135006516		
@SQ	SN:chr11_gl000202_random	LN:40103		
@SQ	SN:chr12	LN:133851895		
@SQ	SN:chr13	LN:115169878		
@SQ	SN:chr14	LN:107349540		
@SQ	SN:chr15	LN:102531392		
@SQ	SN:chr16	LN:90354753		
@SQ	SN:chr17_ctg5_hap1	LN:1680828		
@SQ	SN:chr17	LN:81195210		
@SQ	SN:chr17_gl000203_random	LN:37498		
@SQ	SN:chr17_gl000204_random	LN:81310		
@SQ	SN:chr17_gl000205_random	LN:174588		
@SQ	SN:chr17_gl000206_random	LN:41001		
@SQ	SN:chr18	LN:78077248		
@SQ	SN:chr18_gl000207_random	LN:4262		
@SQ	SN:chr19	LN:59128983		
@SQ	SN:chr19_gl000208_random	LN:92689		
@SQ	SN:chr19_gl000209_random	LN:159169		
@SQ	SN:chr1	LN:249250621		
@SQ	SN:chr1_gl000191_random	LN:106433		
@SQ	SN:chr1_gl000192_random	LN:547496		
@SQ	SN:chr20	LN:63025520		
@SQ	SN:chr21	LN:48129895		
@SQ	SN:chr21_gl000210_random	LN:27682		
@SQ	SN:chr22	LN:51304566		
@SQ	SN:chr2	LN:243199373		
@SQ	SN:chr3	LN:198022430		
@SQ	SN:chr4_ctg9_hap1	LN:590426		
@SQ	SN:chr4	LN:191154276		
@SQ	SN:chr4_gl000193_random	LN:189789		
@SQ	SN:chr4_gl000194_random	LN:191469		
@SQ	SN:chr5	LN:180915260		
@SQ	SN:chr6_apd_hap1	LN:4622290		
@SQ	SN:chr6_cox_hap2	LN:4795371		




```
@SQ     SN:chrUn_gl000247       LN:36422
@SQ     SN:chrUn_gl000248       LN:39786
@SQ     SN:chrUn_gl000249       LN:38502
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:bowtie2      PN:bowtie2      VN:2.3.2        CL:"/galaxy/main/deps/_conda/envs/mulled-v1-cf272fa72b0572012c68ee2cbf0c8f909a02f29be46918c2a23
M00991:178:000000000-BBP68:1:1106:9736:6918     99      chr10   429971  42      68M3D88M        =       430051  166     GTTCGCACCGTCCGCCACTATCAGCATTCGC
M00991:178:000000000-BBP68:1:1106:9736:6918     147     chr10   430051  42      86M     =       429971  -166    AATGCATATCCCTCGATTTCACACACGCCACTTTTGCTA
M00991:178:000000000-BBP68:1:2108:23825:17989   99      chr10   860519  42      153M    =       860666  155     ATATTAAAGGTATTTTGTACAGAAAACACAACACAGACA
M00991:178:000000000-BBP68:1:2108:23825:17989   147     chr10   860666  42      8M      =       860519  -155    GGCACGAG        BAAAABBA        AS:i:0
M00991:178:000000000-BBP68:1:2107:12293:26439   161     chr10   3142833 35      17M     chr17   71196804        0       ACAAAGTCAGCACGCGC       AAAAAFF
M00991:178:000000000-BBP68:1:1104:13558:18222   73      chr10   3201129 42      100M    =       3201129 0       GCCAAAGCCAGATTCAATCAAGGCTTTGTAAAAGGGAGA
M00991:178:000000000-BBP68:1:1104:13558:18222   133     chr10   3201129 0       *       =       3201129 0       ACCAAGCACCGTGGTGCCAGCTCAGACACCCTGGGACAA
M00991:178:000000000-BBP68:1:1103:20292:14133   97      chr10   3820991 42      155M    chr19   10776364        0       GTTGTGTTATATTTGTAAATACACAGCTTAT
M00991:178:000000000-BBP68:1:1107:17049:25808   99      chr10   3821316 42      155M    =       3821331 170     GTATACTCCTTACACACAAAACATTCAAACTACTTTTTT
M00991:178:000000000-BBP68:1:1107:17049:25808   147     chr10   3821331 42      155M    =       3821316 -170    ACAAAACATTCAAACTACTTTTTTTCCATCTCTTGCAGT
M00991:178:000000000-BBP68:1:1104:5844:11822    165     chr10   5170810 0       *       =       5170810 0       ACACATTCATGTCGTGAGTTGCTAAGGATAGCAGACAAG
M00991:178:000000000-BBP68:1:1104:5844:11822    89      chr10   5170810 1       9M      =       5170810 0       AGTATTCAG       @4FFBBBBB       AS:i:0
M00991:178:000000000-BBP68:1:1105:15914:19028   165     chr10   5494394 0       *       =       5494394 0       GACCATACGAGTCCTAGATGTCAATAACCAGTCCTTCAG
M00991:178:000000000-BBP68:1:1105:15914:19028   89      chr10   5494394 1       16M     =       5494394 0       CACCCGCCAAGAGAAG        AAAEDFFFFFFAAAA
M00991:178:000000000-BBP68:1:1101:6644:20822    165     chr10   5494395 0       *       =       5494395 0       GACCATACGAGTCCTAGATGTCAATAACCAGTCCTTCAG
M00991:178:000000000-BBP68:1:1101:6644:20822    89      chr10   5494395 1       15M     =       5494395 0       ACCCGCCAAGAGAAG ACG?FFFFFFBBBBB AS:i:0
M00991:178:000000000-BBP68:1:2114:18702:22202   163     chr10   5766266 42      158M    =       5766282 172     AAATGATAAAGGTTTCTGAGTAGTATTTCTATTCTTTCA
M00991:178:000000000-BBP68:1:2114:18702:22202   83      chr10   5766282 42      156M    =       5766266 -172    TGAGTAGTATTTCTATTCTTTCATTTTTGCAACATATAA
M00991:178:000000000-BBP68:1:1111:5298:18126    99      chr10   11616654        42      155M    =       11616696        197     CTTACTGTACTGCCAATTTTCCT
M00991:178:000000000-BBP68:1:1111:5298:18126    147     chr10   11616696        42      155M    =       11616654        -197    CCCATGAATTATTTTGACATTTT
M00991:178:000000000-BBP68:1:1104:21028:25600   163     chr10   11954572        1       4M      =       11954643        81      CCCA    AAAA    AS:i:0
M00991:178:000000000-BBP68:1:1104:21028:25600   83      chr10   11954643        1       10M     =       11954572        -81     CGAAGCTGGG      DB1FFAA
M00991:178:000000000-BBP68:1:2114:16947:22919   99      chr10   12070773        42      155M    =       12070865        188     CCTGTGGTCCCTTTTCAGGTGTT
M00991:178:000000000-BBP68:1:2114:16947:22919   147     chr10   12070865        42      96M     =       12070773        -188    CACTAGGAGGAAAACTCAAATTA
M00991:178:000000000-BBP68:1:2109:26937:12502   97      chr10   13325702        23      155M    chr16   4700341 0       CTGGGTTTTATTCTGACCAGATCCGTGGATG
M00991:178:000000000-BBP68:1:2107:14639:5173    99      chr10   13361154        31      155M    =       13361183        172     CCCGGGTGTGGGATTCACATTTT
M00991:178:000000000-BBP68:1:2107:14639:5173    147     chr10   13361183        31      143M    =       13361154        -172    CCACTTGTGGTGCGACCTCGATG
M00991:178:000000000-BBP68:1:2105:16260:9454    101     chr10   15151770        0       *       =       15151770        0       GGCAGTTCCAGAAATCATTAAAT
M00991:178:000000000-BBP68:1:2105:16260:9454    153     chr10   15151770        32      19M     =       15151770        0       AAACTTGAGTTTTTCCAAG
M00991:178:000000000-BBP68:1:1103:14887:6379    99      chr10   15834731        42      106M    =       15834858        186     GTTGCTTCCTGACATATAATTGT
M00991:178:000000000-BBP68:1:1103:14887:6379    147     chr10   15834858        42      59M     =       15834731        -186    TCTTTGGAGGTTATGGAATAAGC
M00991:178:000000000-BBP68:1:1109:7860:8475     163     chr10   17271678        42      7M      =       17271687        165     TCTCGCT B?ABABB AS:i:0
M00991:178:000000000-BBP68:1:1109:7860:8475     83      chr10   17271687        42      156M    =       17271678        -165    CCGACGCCATCAACACCGAGTTC
M00991:178:000000000-BBP68:1:1110:18848:15480   163     chr10   17271697        42      116M2D40M       =       17271740        199     CAACACCGAGTTCAA
M00991:178:000000000-BBP68:1:1110:18848:15480   83      chr10   17271740        42      73M2D81M        =       17271697        -199    CTGCAGGAGCTGAAT
M00991:178:000000000-BBP68:1:1111:17753:19097   165     chr10   17275755        0       *       =       17275755        0       CCTTGAACGCAAAGTGGAATCTT
M00991:178:000000000-BBP68:1:1111:17753:19097   89      chr10   17275755        0       4M3I7M15I126M   =       17275755        0       AGATTGCCTTTTTGA
M00991:178:000000000-BBP68:1:2107:13404:9576    81      chr10   17275761        0       5M15I136M       =       17275764        141     TCTTGAAGAAACTCC
```

✓Millions of reads mapped to genome.

✓ Is it possible to analyse it manually ?

✓ Answer is NO

✓To estimate expression , we needed another tool.

✓In 2010, Trapnell et al. published cufflinks and made the transcript abundance an easy task.

nature biotechnology

LETTERS

Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

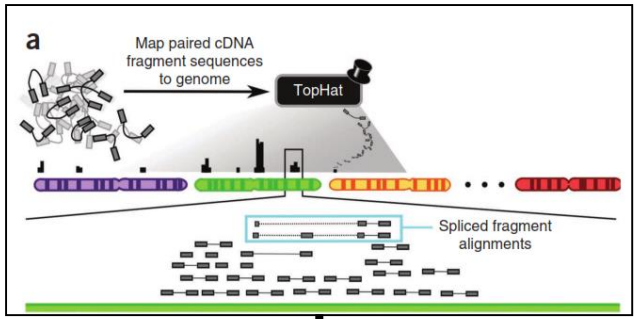
Cole Trapnell[1–3], Brian A Williams[4], Geo Pertea[2], Ali Mortazavi[4], Gordon Kwan[4], Marijke J van Baren[5], Steven L Salzberg[1,2], Barbara J Wold[4] & Lior Pachter[3,6,7]

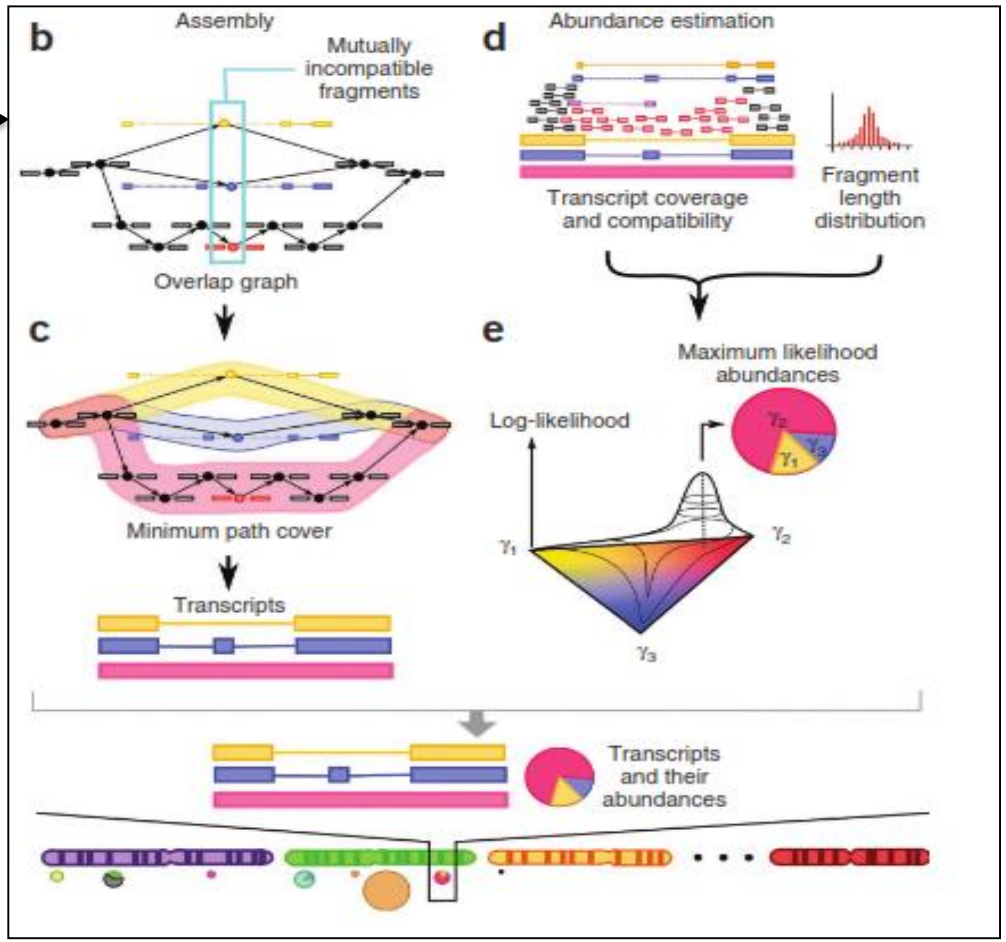
High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation[1–3]. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that

(75 bp in this work versus 25 bp in our previous work) and pairs of reads from both ends of each RNA fragment can reduce uncertainty in assigning reads to alternative splice variants[12]. To produce useful transcript-level abundance estimates from paired-end RNA-Seq data, we developed a new algorithm that can identify complete novel transcripts and probabilistically assign reads to isoforms.

For our initial demonstration of Cufflinks, we performed a time course of paired-end 75-bp RNA-Seq on a well-studied model of skeletal muscle development, the C2C12 mouse myoblast cell line[13] (see Online Methods). Regulated RNA expression of key transcription factors drives myogenesis, and the execution of the differentiation process involves changes in expression of hundreds of genes[14,15]. Previous studies have not measured global transcript isoform expression; however, there are well-documented expression changes at the whole-gene level for a set of marker genes in this system. We aimed to

Cufflink : Assembly and Abundance Estimation





Protocol for RNA Seq Data Analysis

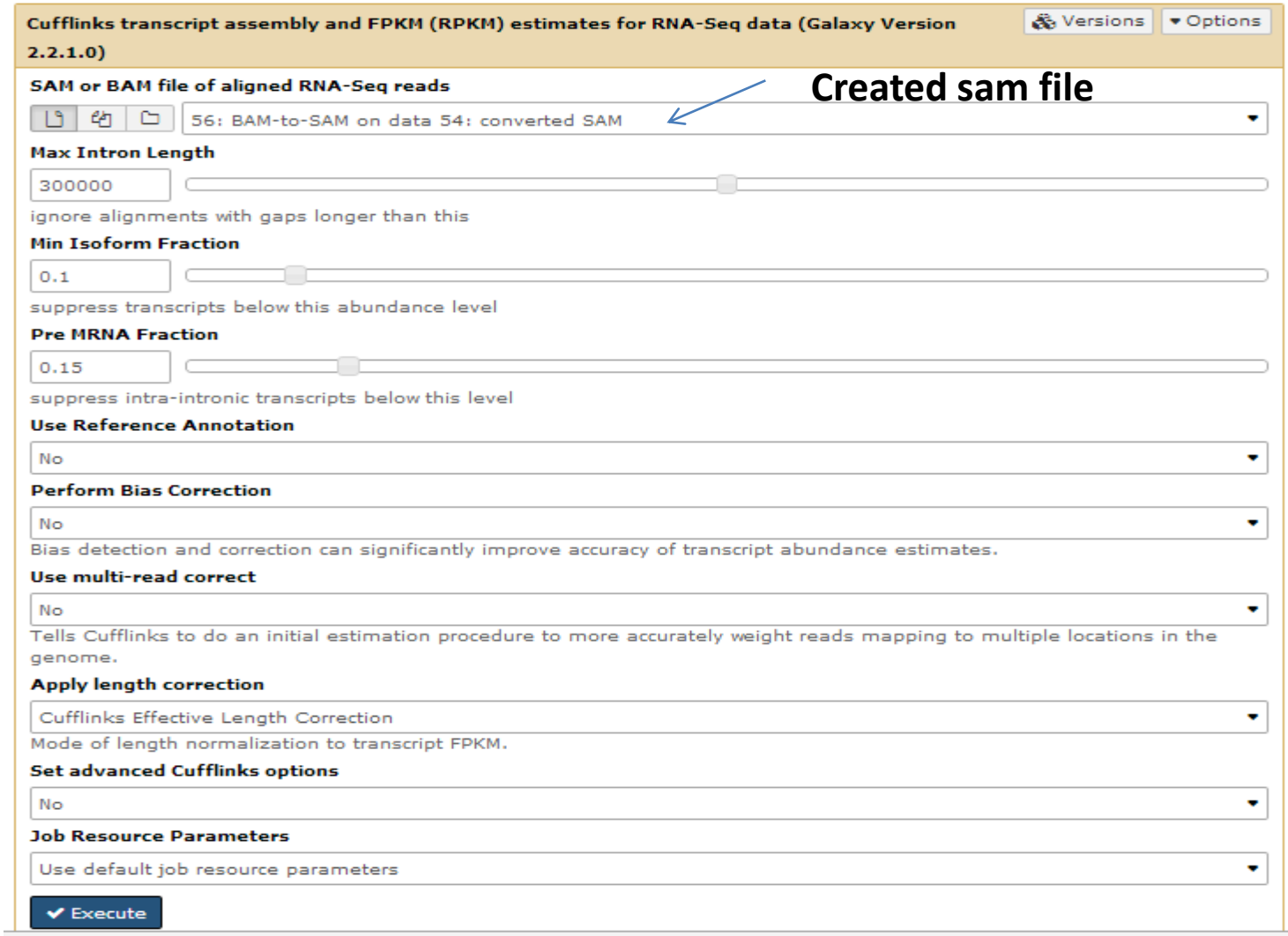
1. Pre-processing

2. Quality Filtration

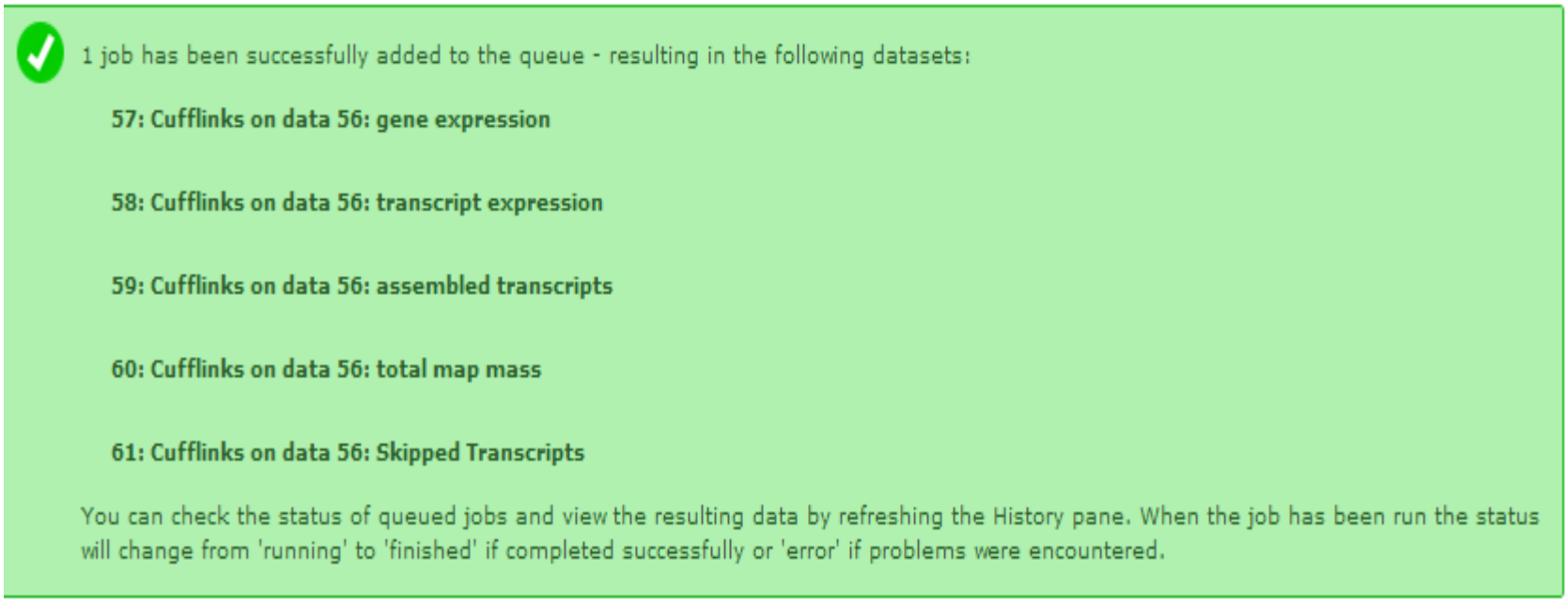
3. Mapping or assembly

4. Expression analysis

Let us run CUFFLINKS to estimate the expression of genes on genomes



You will get 5 output files.



Different Method of Abundance Estimation

✓**Counts per million : Reads counts scaled by the number of fragments you sequenced (N) times one million.**

$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

✓**Transcripts per million (TPM) is a measurement of the proportion of transcripts in your pool of RNA.**

$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6$$

✓**FPKM is a unit of expression. FPKM is simply a unit of expression**

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\widetilde{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l}_i N} \cdot 10^9$$

✓**If you have FPKM, you can easily compute TPM:**

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j}\right) \cdot 10^6$$

Transcript Expressions

1	2	3	4	5	6	7	8	9	10
tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM
CUFF.1.1	-	-	CUFF.1	-	-	chr11:1016561-1017335	774	9.82598	8303.04
CUFF.2.1	-	-	CUFF.2	-	-	chr11:1017439-1018092	653	6.64209	5489.47
CUFF.3.1	-	-	CUFF.3	-	-	chr11:62292242-62293305	1063	3.05145	2641.01
CUFF.4.1	-	-	CUFF.4	-	-	chr11:65272989-65273355	366	32.6074	27409.5
CUFF.5.1	-	-	CUFF.5	-	-	chr11:65266580-65270418	3838	15.0753	12927.1
CUFF.6.1	-	-	CUFF.6	-	-	chr11:65270666-65272939	2273	15.1561	13692.1
CUFF.7.1	-	-	CUFF.7	-	-	chr14:106207848-106208692	844	4.63947	4182.81
CUFF.8.1	-	-	CUFF.8	-	-	chr15:45007636-45007912	276	43.8713	35014.3
CUFF.9.1	-	-	CUFF.9	-	-	chr17:19091226-19091547	321	138.008	116298
CUFF.10.1	-	-	CUFF.10	-	-	chr1:28835082-28835270	188	127.939	105533
CUFF.11.1	-	-	CUFF.11	-	-	chr22:23243039-23243586	547	4.52043	5972.5
CUFF.12.1	-	-	CUFF.12	-	-	chr2:89156745-89157197	452	12.2365	11515.3
CUFF.13.1	-	-	CUFF.13	-	-	chr3:185135517-185136470	953	4.33674	3577.22
CUFF.14.1	-	-	CUFF.14	-	-	chr3:195507749-195508870	1121	2.01657	2699.64
CUFF.15.1	-	-	CUFF.15	-	-	chr3:195508955-195510544	1589	1.60876	2386.61
CUFF.16.1	-	-	CUFF.16	-	-	chr3:195510927-195512268	1341	3.58167	4358.11
CUFF.17.1	-	-	CUFF.17	-	-	chr3:195512377-195514082	1705	1.95635	2615.64
CUFF.18.1	-	-	CUFF.18	-	-	chr7:100550701-100551060	359	17.9467	16894
CUFF.19.1	-	-	CUFF.19	-	-	chr9:135894829-135895508	679	9.04212	7354.52
CUFF.20.1	-	-	CUFF.20	-	-	chrM:2037-2817	780	3.91122	3929.93
CUFF.22.1	-	-	CUFF.22	-	-	chrX:73047134-73047924	790	27.3414	22828.6
CUFF.21.1	-	-	CUFF.21	-	-	chrX:73062299-73062927	628	6.87826	5810.22
CUFF.23.1	-	-	CUFF.23	-	-	chrX:73069152-73069629	477	10.7957	9727.26
CUFF.24.1	-	-	CUFF.24	-	-	chrX:139865623-139866556	933	6.06793	5936.29

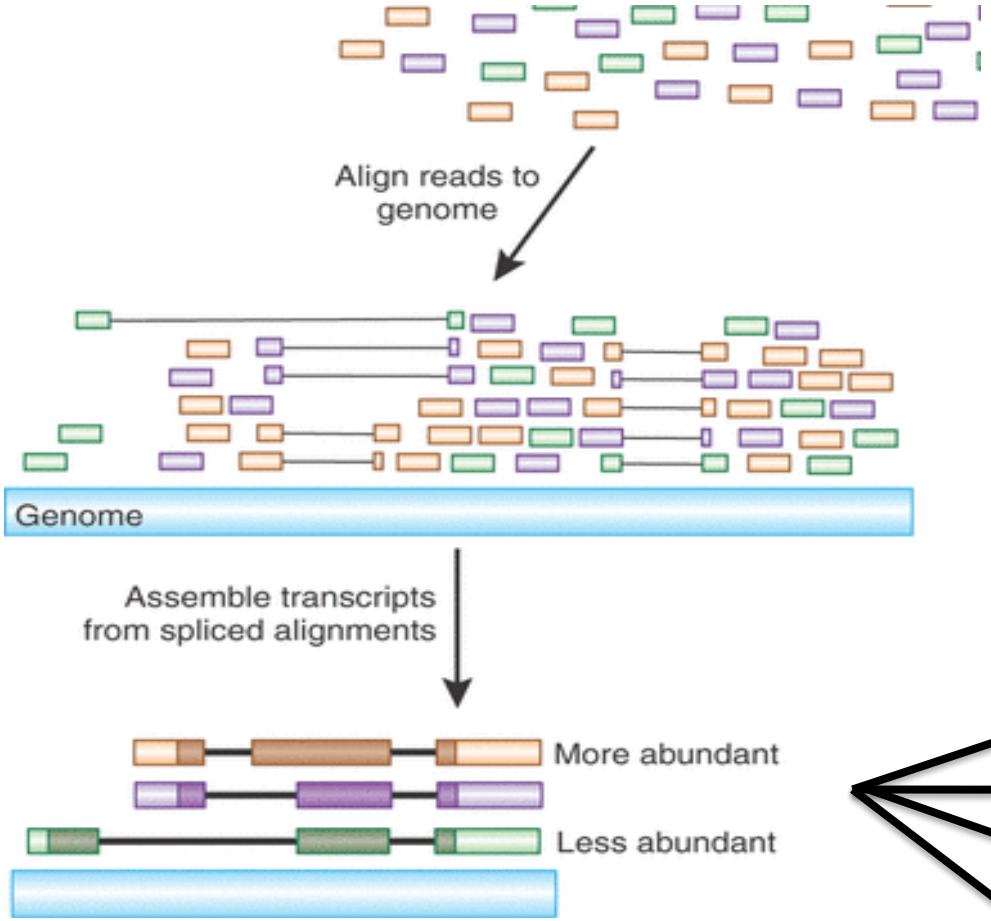
.

Length of transcript

Chromosomal position on genome

Expression in term of FPKM

What if We have two different samples ??



Differential Expression

Differential Isofom Abundance

SNPs , Indels

RNA editing

Sample 2: Start Analysis

✓Upload Sample2 in galaxy server

✓ Run Fastqc (Is there any issue in Quality plot ?)

✓Run Trimmomatic to fix if you find any issue

✓Run bowtie2 (Can you see mapping statistics ? If yes, how many reads mapped)

✓Run cufflink (what is minimum and maximum transcript expression?)

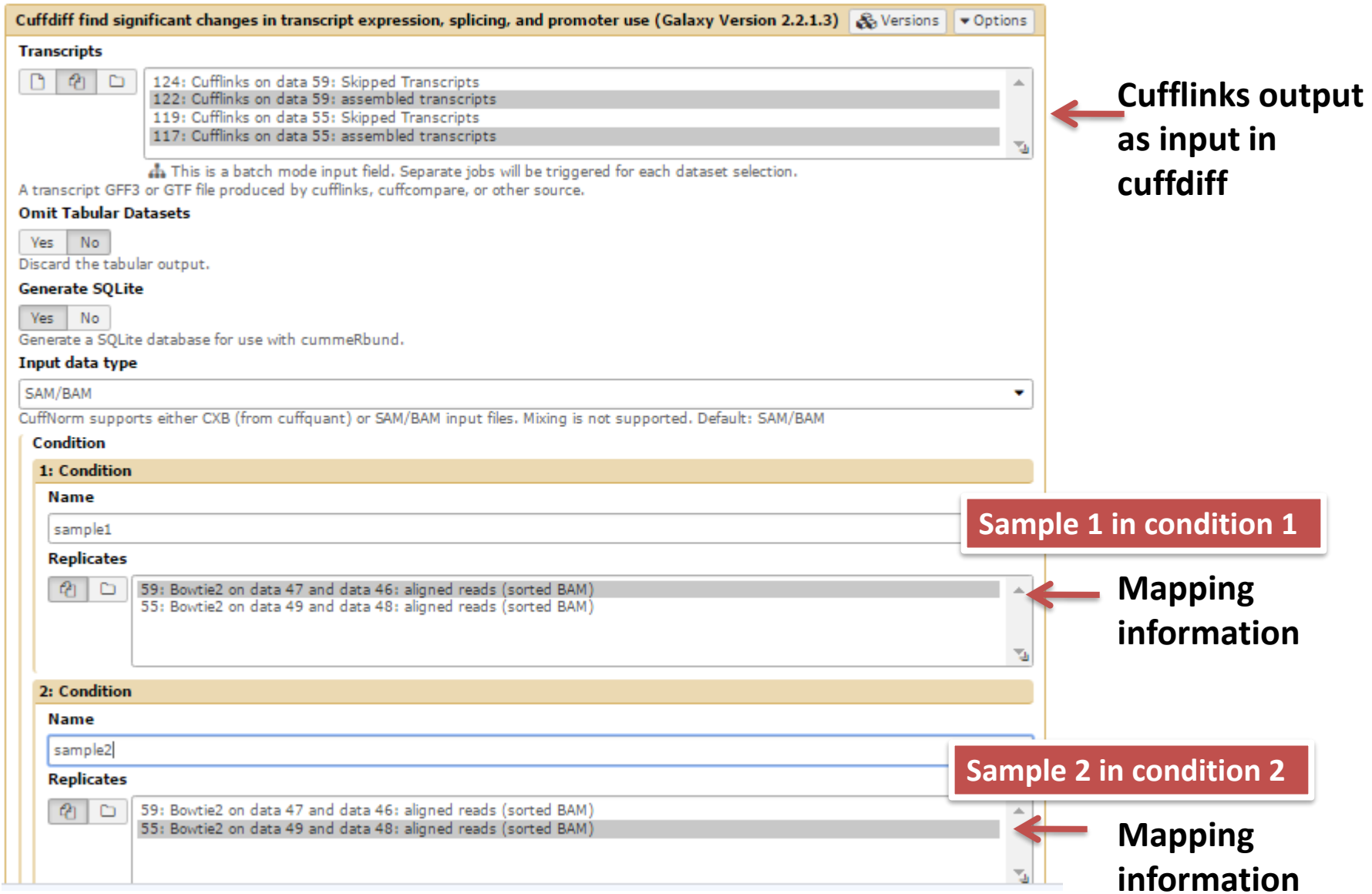
Differentially expression in two different conditions

✓**Cuffdiff is a highly accurate tool for performing sample comparisons, and can tell you which genes are up- or down-regulated between two or more conditions.**

✓**Go to NGS RNA Analysis in galaxy web server.**

✓**Select cuffdiff tool.**

✓**Select assembled transcript as input (output from cufflinks tool).**


Cuffdiff find significant changes in transcript expression, splicing, and promoter use (Galaxy Version 2.2.1.3)
Versions
Options
Transcripts
124: Cufflinks on data 59: Skipped Transcripts
122: Cufflinks on data 59: assembled transcripts
119: Cufflinks on data 55: Skipped Transcripts
117: Cufflinks on data 55: assembled transcripts
This is a batch mode input field. Separate jobs will be triggered for each dataset selection.
A transcript GFF3 or GTF file produced by cufflinks, cuffcompare, or other source.
Omit Tabular Datasets
Yes
No
Discard the tabular output.
Generate SQLite
Yes
No
Generate a SQLite database for use with cummeRbund.
Input data type
SAM/BAM
CuffNorm supports either CXB (from cuffquant) or SAM/BAM input files. Mixing is not supported. Default: SAM/BAM
Condition
1: Condition
Name
sample1
Replicates
59: Bowtie2 on data 47 and data 46: aligned reads (sorted BAM)
55: Bowtie2 on data 49 and data 48: aligned reads (sorted BAM)
2: Condition
Name
sample2
Replicates
59: Bowtie2 on data 47 and data 46: aligned reads (sorted BAM)
55: Bowtie2 on data 49 and data 48: aligned reads (sorted BAM)
Cufflinks output as input in cuffdiff
Sample 1 in condition 1
Mapping information
Sample 2 in condition 2
Mapping information

Cuffdiff Output : FPKM tracking files

isoforms.fpkm_tracking	Transcript FPKMs
genes.fpkm_tracking	Gene FPKMs. Tracks the summed FPKM of transcripts sharing each gene_id
cds.fpkm_tracking	Coding sequence FPKMs. Tracks the summed FPKM of transcripts sharing each p_id, independent of tss_id
tss_groups.fpkm_tracking	Primary transcript FPKMs. Tracks the summed FPKM of transcripts sharing each tss_id

Cuffdiff Output : differential files

isoform_exp.diff	Transcript-level differential expression.
gene_exp.diff	Gene-level differential expression. Tests differences in the summed FPKM of transcripts sharing each gene_id
tss_group_exp.diff	Primary transcript differential expression. Tests differences in the summed FPKM of transcripts sharing each tss_id
cds_exp.diff	Coding sequence differential expression. Tests differences in the summed FPKM of transcripts sharing each p_id independent of tss_id

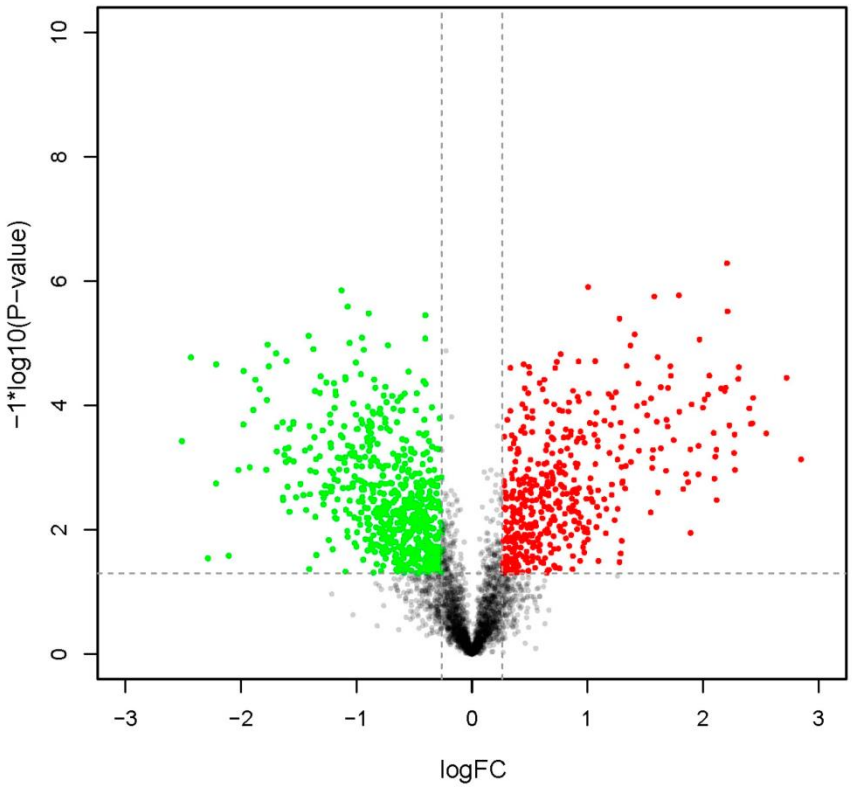
Cuffdiff Output :

1	2	3	4	5	6	7	8	9	10	11	12
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value
CUFF1	CUFF1	-	chr10:98510037-98510664	C1	C2	OK	9493.98	32599.5	1.77976	1.48669	0.17545
CUFF10	CUFF10	-	chr12:6619384-6619710	C1	C2	OK	25232.2	72273.2	1.5182	1.41818	0.1831
CUFF11	CUFF11	-	chr12:49521782-49522637	C1	C2	OK	3189.87	5058.44	0.665194	0.593004	0.57295
CUFF12	CUFF12	-	chr12:125396257-125398338	C1	C2	OK	10291.6	10108.2	-0.0259507	-0.0195277	0.9848
CUFF13	CUFF13	-	chr12:133402275-133402691	C1	C2	OK	14701.8	29691.4	1.01406	0.808465	0.40285
CUFF14	CUFF14	-	chr14:106090797-106091148	C1	C2	OK	22101.4	22363.8	0.0170267	0.0176527	0.93665
CUFF15	CUFF15	-	chr14:106173453-106173905	C1	C2	OK	13261	1561.22	-3.08644	-0.0862077	0.25195
CUFF16	CUFF16	-	chr14:106174096-106174509	C1	C2	OK	12801.5	22623.4	0.821506	0.65991	0.5265
CUFF17	CUFF17	-	chr14:106207785-106208145	C1	C2	OK	70913.6	62761.8	-0.176176	-0.147622	0.887
CUFF18	CUFF18	-	chr14:106109504-106110274	C1	C2	OK	9954.7	7797.09	-0.352442	-0.242097	0.8281
CUFF19	CUFF19	-	chr14:106208210-106208574	C1	C2	OK	31595.9	71169.3	1.17152	0.8854	0.4042
CUFF2	CUFF2	-	chr11:1016562-1018587	C1	C2	OK	5138.32	18878.9	1.8774	1.47427	0.1785
CUFF20	CUFF20	-	chr14:106209105-106209429	C1	C2	OK	49411.3	140200	1.50457	1.16794	0.2812
CUFF21	CUFF21	-	chr14:106110802-106111119	C1	C2	OK	23214.4	59909.3	1.36776	1.05483	0.32665
CUFF22	CUFF22	-	chr14:106235622-106235928	C1	C2	OK	28273.8	4183.93	-2.75654	-0.0769921	0.2652
CUFF23	CUFF23	-	chr15:45007610-45007908	C1	C2	OK	56496.2	187100	1.72758	1.33425	0.2249
CUFF24	CUFF24	-	chr15:82664618-82665097	C1	C2	OK	11063.8	26506.2	1.26048	1.15576	0.2548
CUFF25	CUFF25	-	chr15:83040991-83041657	C1	C2	OK	5466.52	16061.9	1.55495	1.41648	0.192
CUFF26	CUFF26	-	chr16:2812452-2814215	C1	C2	NOTEST	2353.05	1611.13	-0.546457	0	1
CUFF27	CUFF27	-	chr16:2815866-2817216	C1	C2	OK	2859.11	0	-inf	-nan	0.00075
CUFF28	CUFF28	-	chr16:21413502-21415549	C1	C2	OK	3972.75	388.56	-3.35393	-0.23512	0.2134
CUFF29	CUFF29	-	chr16:21415834-21416578	C1	C2	OK	11614.4	3411.63	-1.76737	-1.32884	0.1962
CUFF3	CUFF3	-	chr11:1265350-1265887	C1	C2	OK	7079.54	5677.35	-0.318437	-0.254093	0.8104
CUFF30	CUFF30	-	chr16:21845954-21848154	C1	C2	OK	3665.06	896.135	-2.03205	-1.56799	0.14375
CUFF31	CUFF31	-	chr16:21848746-21849079	C1	C2	OK	18252.1	6470.03	-1.49622	-0.104869	0.44675
CUFF32	CUFF32	-	chr16:22544931-22547789	C1	C2	OK	11608.2	2150.91	-2.43213	-1.71478	0.13965
CUFF33	CUFF33	-	chr16:29494914-29497219	C1	C2	OK	2996.75	850.891	-1.81635	-1.37141	0.17895
CUFF34	CUFF34	-	chr16:30234372-30235341	C1	C2	OK	3511.86	0	-inf	-nan	0.00125
CUFF35	CUFF35	-	chr16:30235457-30237120	C1	C2	OK	2936.07	246.171	-3.57615	-0.0998629	0.27625
CUFF36	CUFF36	-	chr16:51680166-51680519	C1	C2	OK	17105.6	27534.1	0.686749	0.622949	0.5407
CUFF37	CUFF37	-	chr17:18965231-18965486	C1	C2	OK	127516	89551.2	-0.50989	-0.370282	0.71965
CUFF38	CUFF38	-	chr17:18967179-18967437	C1	C2	OK	60815.1	38804.1	-0.648222	-0.536013	0.5706
CUFF39	CUFF39	-	chr17:19015668-19015938	C1	C2	OK	94638.2	71390	-0.4067	-0.299543	0.7727
CUFF4	CUFF4	-	chr11:61732071-61732368	C1	C2	OK	28564.7	18460	-0.629832	-0.367058	0.65485
CUFF40	CUFF40	-	chr17:19091317-19091593	C1	C2	OK	2.28429e+06	1.58952e+06	-0.523156	-0.530726	0.5979
CUFF41	CUFF41	-	chr17:43591221-43592815	C1	C2	NOTEST	2349.71	775.744	-1.59883	0	1
CUFF42	CUFF42	-	chr17:43595213-43595732	C1	C2	OK	8195.18	1205.06	-2.76567	-0.0772472	0.2652
CUFF43	CUFF43	-	chr17:43595898-43596821	C1	C2	OK	2879.33	0	-inf	-nan	0.0057
CUFF44	CUFF44	-	chr17:78318471-78319080	C1	C2	OK	9415.36	922.399	-3.35155	-0.0936056	0.2514
CUFF45	CUFF45	-	chr1:28833875-28834105	C1	C2	OK	264558	270134	0.0300931	0.0260127	0.97805

Cuffdiff Output :

- ✓**Count genes showing log2 => 2 , known as unregulated genes** (increase in expression of a gene in Condition A as compared to B).
- ✓**Count genes showing log2 < 2 , known as down regulated genes** (Decrease in expression of a gene in Condition A as compared to B).

By looking into figure, can you tell
What are unregulated genes (colour)?
What are down regulated (colour)?



Questions ?