



Galaxy: An Open Platform for Data Intensive Biomedical Research

線上次世代序列分析平台

蘇聖堯

2014/9/17



LAB OF System Biology & Network Biology

中央研究院資訊科學研究所

@iis, Academia Sinica, TAIWAN

系統生物學與網路生物學實驗室

What's Galaxy?



Bringing Developers And Biologists Together. Reproducible Science Is Our Goal

- An open, **web-based platform** for data intensive biomedical research.
- Whether on this free public server or your own instance, you can **perform**, **reproduce**, and **share complete analyses**.
- The Galaxy Project is supported by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and John Hopkins University.

Use Galaxy



Use [project's free server](#) or other public servers

Get Galaxy



Install [locally](#) or [in the cloud](#) or get [Galaxy on SlipStream](#)

Learn Galaxy



[Screencasts](#), [Galaxy 101](#), ...

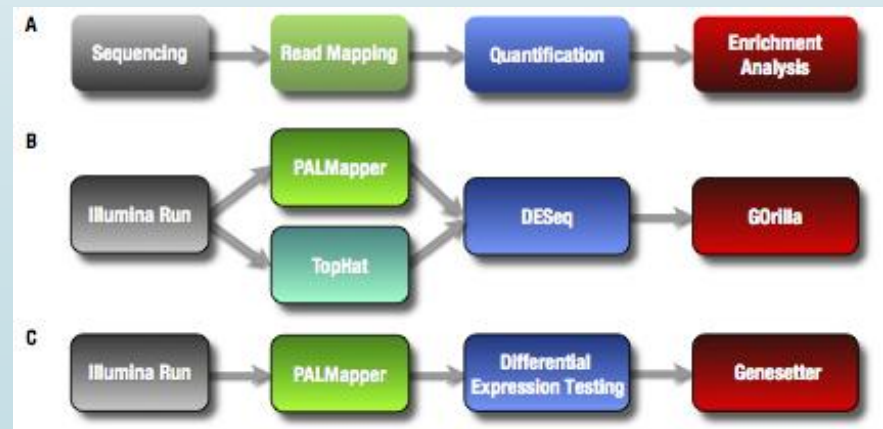
Get Involved



[Mailing lists](#), [Tool Shed](#), [wiki](#)

More About Galaxy

- ▶ A platform/interface for popular NGS software.
- ▶ A data integration and analysis framework for biomedical research. It allows nearly any tool that can be run from the command line to be integrated into it.
- ▶ NO need of programming experience.
- ▶ Keeps track of all the steps performed and results throughout the analysis.



Pre-installed tools in Galaxy

Allows biologists to perform complex genomic analyses

- Analyze multiple alignments
- Compare genomic annotations
- Profile metagenomic samples
- Examine human genomic variation
- Operate on next generation sequencing data

.....

Get Data	
Send Data	
ENCODE Tools	
Lift-Over	
Text Manipulation	
Filter and Sort	
Join, Subtract and Group	
Convert Formats	
Extract Features	
Fetch Sequences	
Fetch Alignments	
Get Genomic Scores	
Operate on Genomic Intervals	
Statistics	
Wavelet Analysis	
Graph/Display Data	
Regional Variation	
Multiple regression	
Multivariate Analysis	
Evolution	
Motif Tools	
Multiple Alignments	
Metaomic analyses	
FASTA manipulation	
	NGS: QC and manipulation
	NGS: Picard (beta)
	NGS: Methylation Mapping
	NGS: Mapping
	NGS: Indel Analysis
	NGS: RNA Analysis
	NGS: SAM Tools
	NGS: GATK Tools (beta)
	NGS: Peak Calling
	NGS: Simulation
	SNP/WGA: Data; Filters
	SNP/WGA: QC; LD; Plots
	SNP/WGA: Statistical Models
	Phenotype Association
	VCF Tools

Where to run Galaxy

- ▶ Main

<https://usegalaxy.org/>

- ▶ Public accessible servers

<https://wiki.galaxyproject.org/PublicGalaxyServers>

- ▶ Galaxy on Amazon (Cloud, charged by Credit card)

- ▶ 國網中心



<http://alps1.nchc.org.tw/galaxy>

- ▶ NTU Galaxy (Limited to NTU IP)

- ▶ Local installation (Your own machine/ Server)

bio-linux? Or Visit our Website to download Live-DVD with myBLAST and ELN (<http://eln.iis.sinica.edu.tw>)



Public Accessible Servers

050+

Public Galaxy Servers
and counting

Publicly Accessible Galaxy Servers

The Galaxy Project's public server (UseGalaxy.org, [Main](#)) can meet many needs, but it is not suitable for everything (see [Choices](#) for why) and cannot possibly scale to meet the entire world's needs.

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 such public or semi-public Galaxy servers.

To add your public Galaxy server to this list, please either just add it (*hey, it's a wiki*), or contact Galaxy Outreach <outreach AT galaxyproject DOT org>.

General Purpose Servers

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.

Andromeda

• Links:

- [Andromeda server](#)
- Andromeda was the featured topic at the [March 2013 GalaxyAdmins Meetup](#). Includes slides and video.
- [GCC2013 Poster and Lightning talk: Andromeda: NBIC Galaxy at Surfsara's HPC cloud](#)

• Domain/Purpose:

- This is a fully populated Galaxy instance.

• Comments:

- As of 2014/01/01:
 - "Due to funding issue, the NBIC Galaxy server is running now with very limited support and maintenance as of January 1st, 2014. We hope this is temporary but please be aware that your analysis will be not performed at an optimal speed and most questions will not be answered."
- Andromeda is hosted at the [SURFsara High Performance Computing \(HPC\) cloud](#).



Computation Power limited

1. General Purpose Servers

1. Andromeda
2. Biomina
3. CBIB Galaxy
4. DBCLS Galaxy
5. Galaxy Main
6. Galaxy Test
7. GeneNetwork
8. Genboree
9. GigaGalaxy
10. GVL QLD
11. GVL Tutorial
12. INRA-URGI
13. NELLY

2. Domain Servers

1. ballaxy
2. CAPER
3. Cistrome Analysis Pipeline
4. CNIC.DarwinTree
5. CoSSci
6. Galaxy-P
7. Galaxy PGTB (Virtual Biodiversity Lab)
8. Genomic Hyperbrowser
9. Gene Ontology (GO)
10. Globus Genomics Proteomics
11. Image Analysis and Processing Toolkit
12. Nebula
13. Oqtans
14. Orione
15. OSDDlinux LiveGalaxy
16. PopGenIE
17. RepeatExplorer

Portal of Galaxy

The screenshot displays the Galaxy web portal interface. At the top, there is a navigation bar with the Galaxy logo, a search bar, and user information (Using 5.6 Gb). A green notification box at the top center contains a checkmark and the text: "This is the Galaxy server, packaged for Bio-Linux. To customize this page edit /etc/galaxy/static/welcome.html then sudo restart galaxy." Below this, a central workflow diagram titled "WWFSMD? grow noody appendages..." is shown. The workflow includes steps like "Input dataset", "Filter", "Join", "Group", "Sort", "Join from Query", and "Select Rows". The URL "usegalaxy.org" is displayed below the workflow. To the right, a "History" panel lists recent jobs, including "35: Tophat for Illumina on data 17, data 16, and data 31: accepted hits" and "28: Tophat for Illumina on data 17 and data 16: accepted hits". A left sidebar contains a "Tools" menu with various categories like "Get Data", "Send Data", "ENCODE Tools", etc.

Commands on Linux for File Processing

- ▶ tail tail tophat_out_SRR039999_1/accepted_hits.sam
- ▶ head head tophat_out_SRR039999_1/accepted_hits.sam
- ▶ cat cat file1 file2 > file3
- ▶ sort sort file
- ▶ diff diff file1.sam file2.sam
- ▶ sed sed '1,2d' tophat_out_SRR039999_1/accepted_hits.sam
- ▶ awk awk '{print \$1 "\t" \$3}' tophat_out/accepted_hits.sam
- ▶ join join combines two files based on the matching content lines found
- ▶ paste paste merge contents of two files side by side
- ▶ split split file into smaller files

Text Manipulation

Galaxy 分析数据 工作流 共享的数据 Visualization Cloud 帮助 账号

工具

Text Manipulation

- [Add column](#) to an existing dataset
- [Compute](#) an expression on every row
- [Concatenate datasets](#) tail-to-head
- [Condense](#) consecutive characters
- [Convert](#) delimiters to TAB
- Merge Columns together**
- [Create single interval](#) as a new dataset
- [Cut](#) columns from a table
- [Change Case](#) of selected columns
- [Paste](#) two files side by side
- [Remove beginning](#) of a file
- [Select random lines](#) from a file
- [Select first](#) lines from a dataset
- [Select last](#) lines from a dataset
- [Trim](#) leading or trailing characters
- [Line/Word/Character count](#) of a dataset
- [Secure Hash / Message Digest](#) on a dataset

Convert Formats

FASTA manipulation

Filter and Sort

Merge Columns (version 1.0.1)

Select data:
Dataset missing? See TIP below.

Merge column:

with column:
Need to add more columns? Use controls below.

Columns
Add new Columns

Execute

TIP: If your data is not TAB delimited, use *Text Manipulation->Convert*

What it does
This tool merges columns together. Any number of valid columns can be merged in any order.

Example
Input dataset (five columns: c1, c2, c3, c4, and c5):
1 10 1000 gene1 chr
2 100 1500 gene2 chr

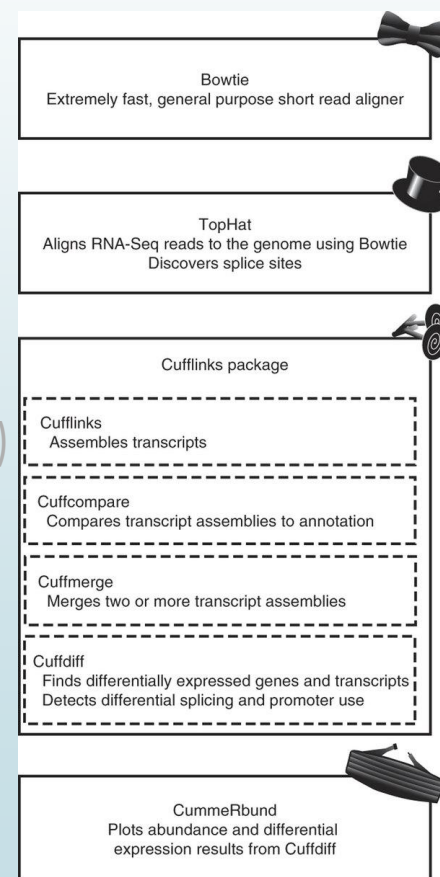
merging columns "c5,c1" will return:
1 10 1000 gene1 chr chr1
2 100 1500 gene2 chr chr2

Note that all original columns are preserved and the result of merge is added as the rightmost column.

Galaxy Can Help Analyze Data

- RNA-seq (Tuxedo pipeline)
- ChIP-seq/ChIP-chip
- Variant detection
- Mass Spectrometry based proteomics (Galaxy-P)
- Multivariate GWAS analysis (GWIS)
- Prediction of metagenomic taxonomy (MGTAXA)
- Gene Ontology
- Protein synthesis (GWIPS-viz)

.....



Playground for Self-Learning

Galaxy Wiki 登入 | 搜尋: 標題 內文

Learn Locked History Actions



<http://wiki.galaxyproject.org/Learn>

Learn Galaxy

目錄

- 1. [Galaxy 101](#)
- 2. [Screencasts](#)
- 3. [Shared Pages, Histories & Workflows](#)
- 4. [Other Tutorials](#)
- 5. [Datasets](#)
- 6. [Tools](#)
- 7. [Visualization](#)
- 8. [User Accounts](#)
- 9. [Other](#)

There are many approaches to learning how to use Galaxy. The most popular is probably to just dive in and use it. Galaxy is simple enough to use that you can do many analyses just by exploring the interface. However, you may miss much of the power this way.

Watch the short [Learn](#) screencast for a learning resource overview.

Galaxy 101

Walking through the [Galaxy 101](#) exercise will show you the ins and outs of using Galaxy. This includes loading data (from UCSC in this example), using genome builds, the tool interface, filtering, sorting, and combining datasets, generating statistics, and Galaxy's History, Workflow and [sharing](#) support.

Galaxy 101 is available in several formats. You can start with either the [Galaxy Page](#) or the [screencast](#).

- Learn
- Screencasts
- FAQ
- Interval Ops
- Datasets
- Pages
- Share
- FTP Upload
- Accounts
- Support
- Security
- Search



Use Galaxy

- [Use Main \(about\)](#)
- [Use Others!](#) • [Learn](#)
- [Share](#) • [Search](#)

Communication

- [Support](#) • [News](#)
- [Events](#) • [Twitter](#)
- [Mailing Lists \(search\)](#)

Deploy Galaxy

- [Get Galaxy](#) • [Cloud](#)
- [Admin](#) • [Tool Config](#)
- [Tool Shed](#) • [Search](#)



Contribute

- [Tool Shed](#) • [Share](#)

Galaxy Events/ Training Programs



Date	Topic/Event	Venue/Location	Contact
September 6-10	At least one tutorial, a panel of European Public Galaxy Instances, and 5 posters	European Conference on Computational Biology (ECCB'14) , Strasbourg, France	Presenters
September 11	<i>Tools integration on Galaxy</i>	Galaxy User Group Grand Ouest, Rennes, France	Cyril Monjeaud, Yvan Le Bras
September 15	<i>Fourth GUGGO meeting</i>	Galaxy User Group Grand Ouest, Rennes, France	Cyril Monjeaud, Yvan Le Bras
September 19	<i>The Great GigaScience and Galaxy (G3) Workshop</i>	The University of Melbourne, Melbourne, Australia	Nick Wong <nwon AT unime1b DOT edu.au>, Ross Lazarus
September 23-25	<i>Analisi dati Next Generation Sequencing con Galaxy</i>	Cagliari, Italy	CRS4 <ngs-courses@crs4.it>
September 24	<i>Introduction to Galaxy - Data Manipulation and Visualisation</i>	University of Cambridge, United Kingdom	Anne Pajon, Jing Su
September 25	<i>RADseq analysis using STACKS on Galaxy</i>	Galaxy User Group Grand Ouest, Rennes, France	Yvan Le Bras, Cyril Monjeaud
September 30	<i>CIR Interactive Workshop - Introduction to bioinformatics analysis with Galaxy application</i>	RBI, Zagreb, Croatia	Enis Afgan
September 30 - October 2	Galaxy Training and Demo Day	2014 Swiss-German Galaxy Tour with events in Bern, Switzerland and Freiburg, Germany	<div style="background-color: black; color: white; padding: 5px; text-align: center;"> https://wiki.galaxyproject.org/Events </div> Hans-Rudolf Hotz and Bjoern Gruening
	<i>(second Swiss) Galaxy Workshop</i>		
	<i>German Galaxy Developers Day</i>		

Platform Choice for Running Galaxy



	Main	Local	Cloud	Appliance	Other
Your data sets are moderately sized	Yes	Yes	Yes	Yes	?
Your computational requirements are moderate	Yes	Yes	Yes	Yes	?
You want to share your Galaxy objects with others	Yes	Yes	Yes	Yes	?
All needed Tools are installed on Main.	Yes	?	Yes	Yes	?
Your data sets are very large	No	?	Yes	Yes	?
Your computational requirements are very large	No	?	Yes	Yes	?
You have absolute data security requirements	No	Yes	Yes	Yes	?
No network transfer of data	No	Yes	No	Yes	Yes

CloudMan: Galaxy on Cloud



CloudMan

目錄

1. [About Galaxy on the cloud](#)
2. [Instantiating a Galaxy instance on the Amazon cloud](#)
3. [Detailed steps](#)
4. [Galaxy AMIs](#)
5. [Determining the size of your cloud cluster](#)
6. [Customizing your cloud cluster](#)
7. [Notes](#)
8. [Presentations](#)
9. [Publications](#)

Note: There are several choices for using Galaxy. This page describes installing Galaxy on a *cloud infrastructure* using CloudMan (see below). For other options, see [Choices](#) and [Cloud](#).

About Galaxy on the cloud

With sporadic availability of data, individuals and labs may have a need to, over a period of time, process greatly variable amounts of data. Such variability in data volume imposes variable requirements on availability of compute resources used to process given data.

<https://wiki.galaxyproject.org/CloudMan>

enabled Galaxy to be instantiated on [cloud computing](#) infrastructures, primarily [Amazon Elastic Compute Cloud \(EC2\)](#). An instance of Galaxy on the cloud behaves just like a local instance of Galaxy except that it offers the benefits of cloud computing resource availability and [pay-as-you-go](#) resource ownership model. Having simple access to Galaxy on the cloud enables as many instances of Galaxy to be acquired and started as is needed to process given data. Once the need subsides, those instances can be released as simply as they were acquired. With such a paradigm, one pays only for the resources they need and use while all the other concerns and costs are eliminated. To see how much using Amazon cloud might cost, you can use the [AWS cost calculator](#). When calculating the total cost, in addition to the EC2 instance, you will have EBS volumes associated with your cluster. There are a total of three EBS volumes associated with each Galaxy cluster: your data volume (size is decided by you when setting up the cluster, say 100GB to begin with), tools volume (10GB), and indices volume (700GB). (Note, the indices volume can be greatly reduced if you don't need all the genome data).

CloudMan

- Customize
- Get Started w AWS
- User Data
- Capacity Planning
- HTCondor
- Hadoop

Appliance for Galaxy: SlipStream

You are here: Home » SlipStream Appliance

TOOLS	TASK	DATA	RUN-TIME
Bowtie 2	Mapping whole human genome	204 million paired-end 100bp Illumina reads	2 Hours 44 Minutes
SAMTools	SAM-BAM conversion	127GB SAM (41GB resulting BAM)	2 Hours 7 Minutes
TopHat 2	RNA-Seq mapping	24 million 100bp Illumina reads	1 Hours 24 Minutes
Cufflinks 2	Differential Expression Analysis	4.3 GB SAM File	0 Hours 11 Minutes



The SlipStream Appliance: Galaxy Edition offers a powerful dedicated resource for data analysis. It reduces the IT and administrative burden of running a production instance of Galaxy. It offers a powerful dedicated resource and, like the Galaxy platform, is designed to lower the barrier to entry into data analysis.



SlipStream Galaxy is a hardware appliance consists of **16** Intel cores, **100 GB** of solid state drive, **384 GB** of memory, and **16 TB** of usable storage space. Galaxy is pre-installed and configured. The appliance sells for under **\$20,000**.

Data Upload

Galaxy 分析数据 工作流 共享的数据 帮助 User

工具

search tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- EBI SRA ENA SRA
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
浏览... 未選擇檔案

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size
Your FTP upload directory contains no files.	

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at localhost using your Galaxy credentials (email address and password).

Convert spaces to tabs:
 Yes
Use this option if you are entering intervals by hand.

Genome:
Human Feb. 2009 (GRCh37/hg19) (hg19)

1. Upload file

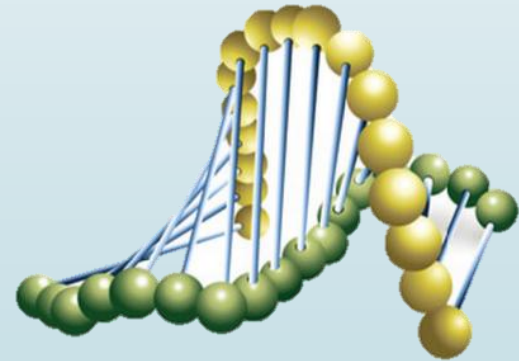
2. Use the URL

3. via FTP

Large Data (>2G) takes time

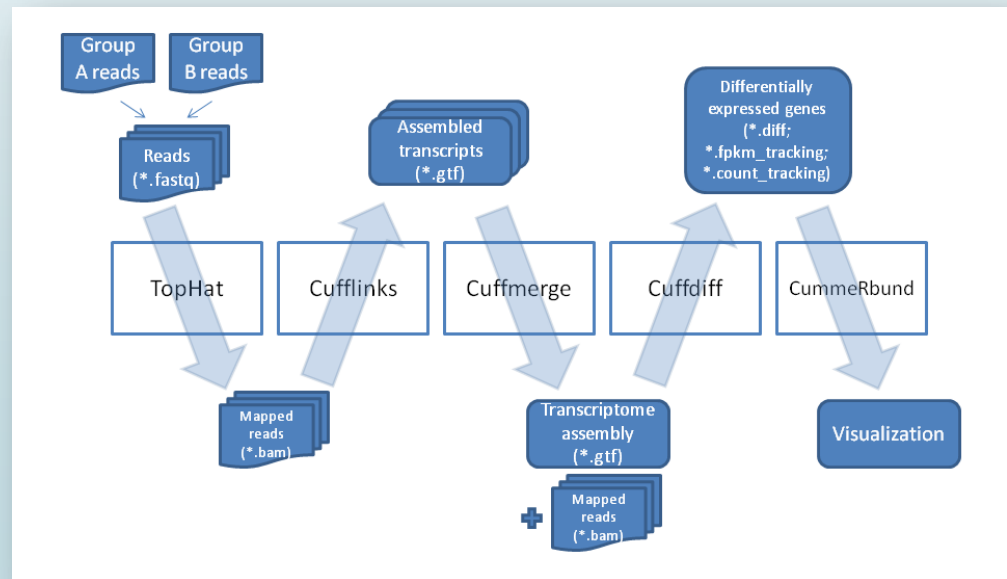
Data Format

- ▶ FASTA sequence format
- ▶ NGS file formats
 - ▶ fastq, sam, bam
- ▶ UCSC file format specifications
 - ▶ bed, wig, gtf, gff



Some Basic Concepts should be Kept In Mind

- ▶ Raw data (generated from sequencer): FASTQ
- ▶ Output of NGS read alignment tools (BWA, Bowtie): SAM/BAM
- ▶ Annotation file for genome browser: GTF, WIG, BED



FASTQ

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores

```
@SLXA-B3_649_FC8437_R1_1_1_610_79  
GATGTGCAATACCTTTGTAGAGGAA  
+SLXA-B3_649_FC8437_R1_1_1_610_79  
YYYYYYYYYYYYYYYYYYYYWYWYYSU
```

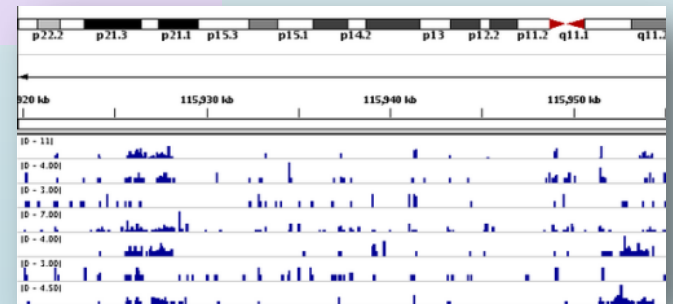
Line 1 begins with a '@' character and is followed by a sequence identifier.
Line 2 is the raw sequence letters.
Line 3 begins with a '+' character.
Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

WIG

The **wiggle (WIG) format** is for display of dense, continuous data such as GC percent, probability scores, and transcriptome data

```
variableStep chrom=chr2  
300701 12.5  
300702 12.5  
300703 12.5  
300704 12.5  
300705 12.5
```

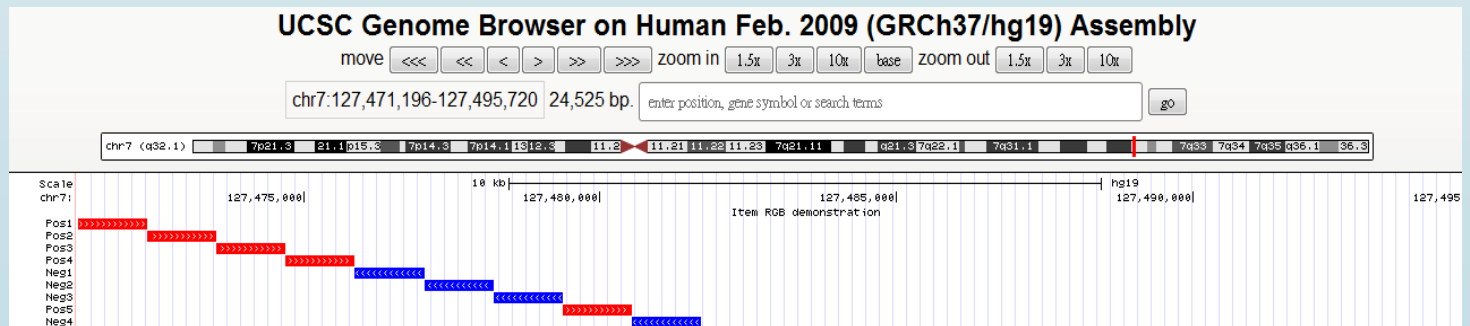
```
variableStep chrom=chr2 span=5  
300701 12.5
```



BED

BED format provides a flexible way to define the data lines that are displayed in an annotation track

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2
itemRgb="On"
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```



GFF/GTF (Gene Transfer Format)

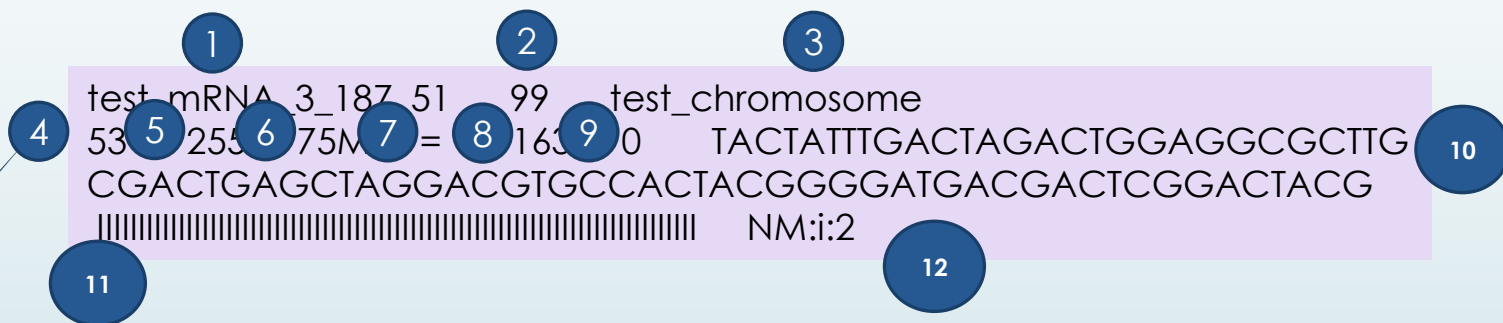
GFF format General Feature Format is a format for describing genes and other features associated with DNA, RNA and Protein sequences.

1 2 3 4 5 6 7 8 9
chr1 unknown exon 27951600 27951662 . - . gene_id "FGR";
transcript_id "NM_001042747"; gene_name "FGR"; p_id "P20191"; tss_id "TSS3342";

1. seqname - Must be a chromosome or scaffold.
2. source - The program that generated this feature.
3. feature - The name of this type of feature. Some examples of standard feature types are "CDS", "start_codon", "stop_codon", and "exon".
4. start - The starting position of the feature in the sequence. The first base is numbered 1.
5. end - The ending position of the feature (inclusive).
6. score - A score between 0 and 1000. If there is no score value, enter ".".
7. strand - Valid entries include '+', '-', or '.' (for don't know/care).
8. frame - If the feature is a coding exon, frame should be a number between 0-2 that represents the reading frame of the first base. If the feature is not a coding exon, the value should be '.'.
9. group - All lines with the same group are linked together into a single item.

SAM

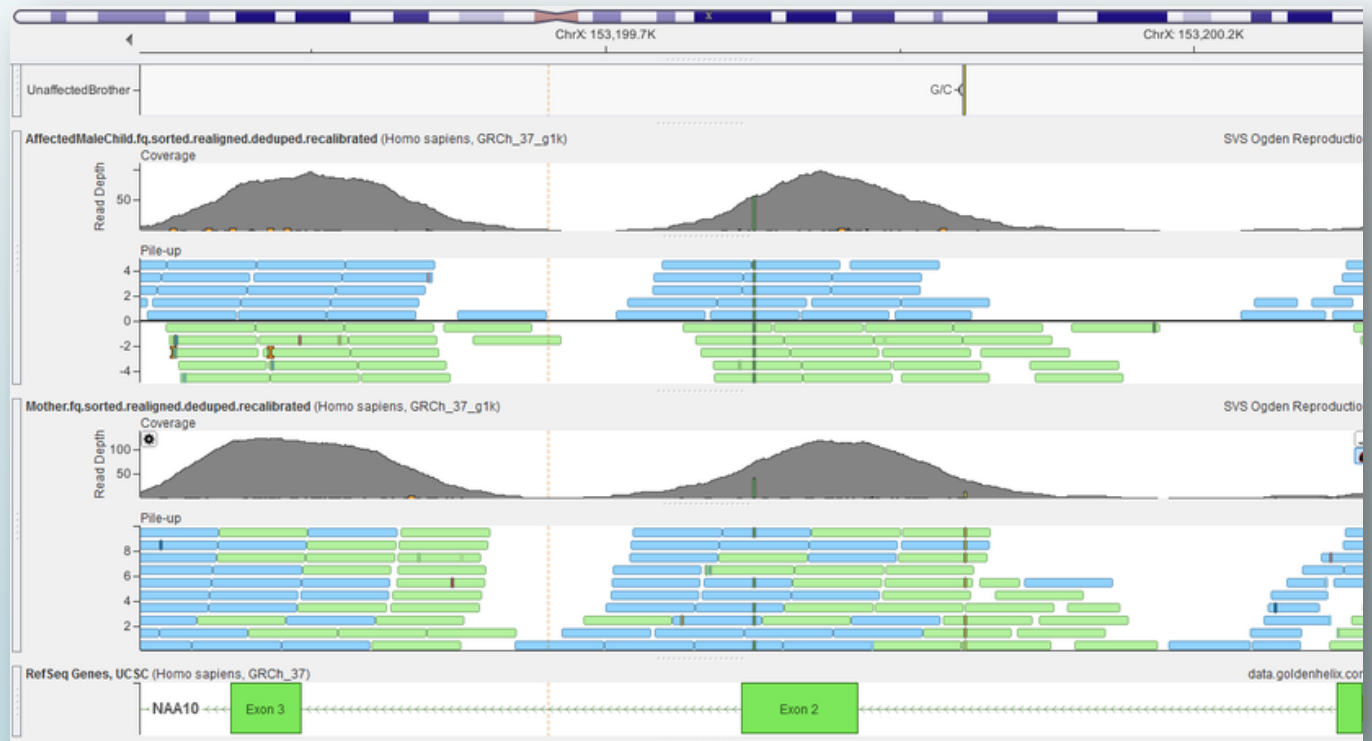
SAM format data is output from aligners that read FASTQ files and assign the sequences to a position with respect to a known reference genome.



Col	Field	Description
1	QNAME	Query (pair) NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIAGR	extended CIGAR string
7	MRNM	Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS	1-based Mate POSition
9	ISIZE	Inferred insert SIZE
10	SEQ	query SEQUENCE on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

BAM

Stores the same data as SAM file in a **compressed**, indexed, binary form.



NGS: SAM Tools: BAM ↔ SAM

工具

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format**
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- MPileup SNP and indel caller

SAM-to-BAM (version 1.1.2)

Choose the source for the reference list:
Locally cached

SAM File to Convert:
↓

Execute

What it does
This tool uses the [SAMTools](#) toolkit to produce an indexed BAM file based on a sorted input SAM file.

工具

NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format**
- Merge BAM Files merges BAM files together
- MPileup SNP and indel caller

BAM-to-SAM (version 1.0.3)

BAM File to Convert:
54: Tophat for Illumi..cepted_hits

Include header in output:

Execute

What it does
This tool uses the [SAMTools](#) toolkit to produce a SAM file from a BAM file.

Basic Modules

- Data I/O : Get Data & Send Data
- Text Manipulation
- Convert Formats
- Statistics
- Display Data
- NGS Analysis : QC, Mapping, RNA Analysis, Methylation Mapping, Peak Calling
- SNP Analysis

[Get Data](#)
[Send Data](#)
[ENCODE Tools](#)
[Lift-Over](#)
[Text Manipulation](#)
[Filter and Sort](#)
[Join, Subtract and Group](#)
[Convert Formats](#)
[Extract Features](#)
[Fetch Sequences](#)
[Fetch Alignments](#)
[Get Genomic Scores](#)
[Operate on Genomic Intervals](#)
[Statistics](#)
[Wavelet Analysis](#)
[Graph/Display Data](#)
[Regional Variation](#)
[Multiple regression](#)
[Multivariate Analysis](#)
[Evolution](#)
[Motif Tools](#)
[Multiple Alignments](#)
[Metagenomic analyses](#)
[FASTA manipulation](#)

[NGS: QC and manipulation](#)
[NGS: Picard \(beta\)](#)
[NGS: Methylation Mapping](#)
[NGS: Mapping](#)
[NGS: Indel Analysis](#)
[NGS: RNA Analysis](#)
[NGS: SAM Tools](#)
[NGS: GATK Tools \(beta\)](#)
[NGS: Peak Calling](#)
[NGS: Simulation](#)
[SNP/WGA: Data; Filters](#)
[SNP/WGA: QC; LD; Plots](#)
[SNP/WGA: Statistical Models](#)
[Phenotype Association](#)
[VCF Tools](#)

Must Know Your Data Types Very Well

工具

NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat2 Gapped-read mapper for RNA-seq data
- Tophat for SOLiD Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- eXpress Quantify the abundances of a set of target sequences from sampled subsequences
- Cuffmerge merge together several Cufflinks assemblies
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

DE NOVO ASSEMBLY

- Trinity De novo assembly of

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:
41: Galaxy5-[brain_2...fastqsanger
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options

Select a reference genome:
hg19
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Paired-end

RNA-Seq FASTQ file:
41: Galaxy5-[brain_2...fastqsanger
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:
20

TopHat settings to use:
Full parameter list
Use the Full parameter list to change default settings.

Library Type:
FR Unstranded

TopHat will treat the reads as strand specific. Every read alignment will have an XS attribute tag. Consider supplying library type options below to select the correct RNA-seq protocol.

Text Manipulation

工具

Text Manipulation

- Add column to an existing dataset
- Compute an expression on every row
- Concatenate datasets tail-to-head
- Cut columns from a table
- **Merge Columns together**
- Convert delimiters to TAB
- Create single interval as a new dataset
- Change Case of selected columns
- Paste two files side by side
- Remove beginning of a file
- Select random lines from a file
- Select first lines from a dataset
- Select last lines from a dataset
- Trim leading or trailing characters
- Line/Word/Character count of a dataset

74: Filter on data 70
Dataset missing? See TIP below.

Merge column:
c1

with column:
c1

Need to add more columns? Use controls below.

Columns
Add new Columns

Execute

TIP: If your data is not TAB delimited, use *Text Manipulation*->*Convert*

What it does
This tool merges columns together. Any number of valid columns can be merged in any order.

Example
Input dataset (five columns: c1, c2, c3, c4, and c5):
1 10 1000 gene1 chr
2 100 1500 gene2 chr

merging columns "c5,c1" will return:
1 10 1000 gene1 chr chr1
2 100 1500 gene2 chr chr2

Warning: Note that all original columns are preserved and the result of merge is added as the rightmost column.

Extract Genomic DNA

工具

Extract Features

Fetch Sequences

- Extract Genomic DNA using coordinates from assembled/unassembled genomes

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Wavelet Analysis

Graph/Display Data

Regional Variation

Multiple regression

Multivariate Analysis

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

NGS: QC and manipulation

NGS: Picard (beta)

NGS: Methylation Mapping

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

by two separate tools.

What it does

This tool uses coordinate, strand, and build information to fetch genomic DNAs in FASTA or interval format. If strand is not defined, the default value is "+".

Example

If the input dataset is:

```
chr7 127475281 127475310 NM_000230 0 +
chr7 127485994 127486166 NM_000230 0 +
chr7 127486011 127486166 D49487 0 +
```

Extracting sequences with **FASTA** output data type returns:

```
>hg17_chr7_127475281_127475310_+
GTAGGAATCGCAGCGCCAGCGGTGCAAG
>hg17_chr7_127485994_127486166_+
GCCCAAGAAAGCCCATCCTGGGAAGGAAAATGCATTGGGGAACCCCTGTGCG
GATTCTTGTGGCTTTGGCCCTATCTTTTCTATGTCCAAGCTGTGCCCATC
CAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAG
GATCAATGACATTTACACACG
>hg17_chr7_127486011_127486166_+
TGGGAAGGAAAATGCATTGGGGAACCCCTGTGCGGATTCTTGTGGCTTTGG
CCCTATCTTTTCTATGTCCAAGCTGTGCCCATCAAAAAGTCCAAGATGA
CACCAAAACCCTCATCAAGACAATTGTCACCAGGATCAATGACATTTAC
ACACG
```

Extracting sequences with **Interval** output data type returns:

```
chr7 127475281 127475310 NM_000230 0 + GTAGGAATCGCAGCGCCAGCGGTGCAAG
chr7 127485994 127486166 NM_000230 0 + GCCCAAGAAAGCCCATCCTGGGAAGGAAAATGCATTGGGGAACCCCTGTGCGGATTCTTGTGGCTTTGG
chr7 127486011 127486166 D49487 0 + TGGGAAGGAAAATGCATTGGGGAACCCCTGTGCGGATTCTTGTGGCTTTGGCCCTATCTTTTCTATGTCCAAGCTG
```

Filter and Sort

工具

Text Manipulation

Filter and Sort

- Filter data on any column using simple expressions
- Sort data in ascending or descending order
- Select lines that match an expression

GFF

- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions
- Filter GTF data by attribute values list

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Filter (version 1.1.0)

Filter:

74: Filter on data 70

Dataset missing? See TIP below.

With following condition:

c1=='chr22'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

Execute

⚠ Double equal signs, ==, must be used as "equal to" (e.g., `c1 == 'chr22'`)

ℹ **TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

ℹ **TIP:** If your data is not TAB delimited, use *Text Manipulation->Convert*

Syntax

The filter tool allows you to restrict the dataset using simple conditional statements.

Columns are referenced with **c** and a **number**. For example, `c1` refers to the first column of a tab-delimited file. Make sure that multi-character operators contain no white space (e.g., `<=` is valid while `< =` is not valid)

When using 'equal-to' operator **double equal sign** '==' **must be used** (e.g., `c1=='chr1'`)

Non-numerical values must be included in single or double quotes (e.g., `c6=='+'`)

Filtering condition can include logical operators, but **make sure operators are all lower case** (e.g., `(c1!='chrX' and c1!='chrY')` or not `c6=='+'`)

Convert Formats

工具 

Convert Formats

- [AXT to concatenated FASTA](#)
Converts an AXT formatted file to a concatenated FASTA alignment
- [AXT to FASTA](#) Converts an AXT formatted file to FASTA format
- [AXT to LAV](#) Converts an AXT formatted file to LAV format
- **[BED-to-GFF converter](#)**
- [FASTA-to-Tabular](#) converter
- [GFF-to-BED](#) converter
- [LAV to BED](#) Converts a LAV formatted file to BED format
- [Maf to BED](#) Converts a MAF formatted file to the BED format
- [MAF to Interval](#) Converts a MAF formatted file to the Interval format
- [MAF to FASTA](#) Converts a MAF formatted file to FASTA format
- [Tabular-to-FASTA](#) converts tabular file to FASTA format
- [FASTQ to FASTA](#) converter

BED-to-GFF (version 2.0.0)

Convert this query:

60: (as bed) Cuffmerge on data..transcripts ▾

Execute

What it does

This tool converts data from BED format to GFF format (scroll down for format description).

Example

The following data in BED format:

```
chr28 346187 388197 BC114771 0 + 346187 388197 0 9 144,81,115,63,155,96,134,105,112, 0,24095,261
```

Will be converted to GFF (**note** that the start coordinate is incremented by 1):

```
##gff-version 2
##bed_to_gff_converter.py

chr28 bed2gff mRNA 346188 388197 0 + . mRNA BC114771;
chr28 bed2gff exon 346188 346331 0 + . exon BC114771;
chr28 bed2gff exon 370283 370363 0 + . exon BC114771;
chr28 bed2gff exon 372378 372492 0 + . exon BC114771;
chr28 bed2gff exon 377194 377256 0 + . exon BC114771;
chr28 bed2gff exon 378319 378473 0 + . exon BC114771;
chr28 bed2gff exon 379722 379817 0 + . exon BC114771;
chr28 bed2gff exon 383182 383315 0 + . exon BC114771;
chr28 bed2gff exon 387981 388085 0 + . exon BC114771;
chr28 bed2gff exon 388086 388197 0 + . exon BC114771;
```


Reverse Complement

The screenshot shows the Galaxy web interface for the 'Reverse-Complement' tool. On the left, a sidebar lists various tools, with 'Reverse-Complement' highlighted in a blue box. The main panel shows the tool's configuration, including a dropdown menu for the library set to '42: http://hgdownload..es/chr19.fa'. Below the configuration is an 'Execute' button. The tool's description, 'What it does', states that it reverse-complements each sequence in a library and also reverses quality scores if the input is FASTQ. An 'Example' section shows the transformation of an input FASTQ sequence: 'TGTCTGTAGCCTCNTCCTTGTAATTCAAAGNNGGTA' is converted to 'TTACAAGGANAGAGGCTACAGACA'. The output FASTQ file is also shown, with the reverse-complemented sequence and its quality scores. At the bottom, a note states that the tool is based on 'FASTX-toolkit' by Assaf Gordon.

TGTCTGTAGCCTC**N**TCCTTGTA
→ TTACAAGGAN**N**GAGGCTACAGACA

Pipeline for RNA-Seq Analysis

1

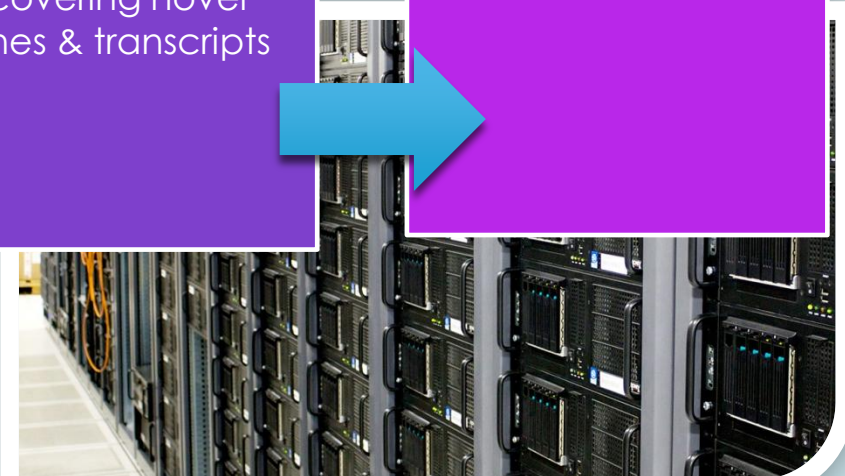
- Library construction
- Sequencing
- Quality control
- Sequence Trimming

2

- Transcript assembly w/ w/o reference
- Calculate abundances
- Identifying DEGs
- Discovering novel genes & transcripts

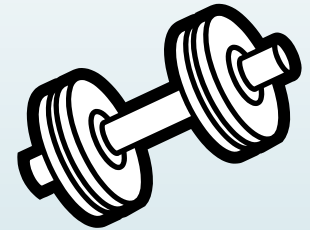
3

- Pathway & functional analysis
- Data Integration
- Visualization



Exercise

- Adrenal & brain tissues RNA-seq data (Illumina BodyMap 2.0)
- Know its reads quality (Trim reads)
- Map the reads
- Assemble and analyze transcripts
- Identify all novel splice junctions and transcript isoforms
- Find loci that exhibit differences in TSS and splicing



FastQC: NGS Quality Control

工具

FASTQ manipulation

NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

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

ROCHE-454 DATA

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

AB-SOLID DATA

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

L7_GA120-6_NoIndex_L007_R1_001.fastq FastQC Report

FastQC Report
Wed 4 Sep 2013
L7_GA120-6_NoIndex_L007_R1_001.fastq

Summary

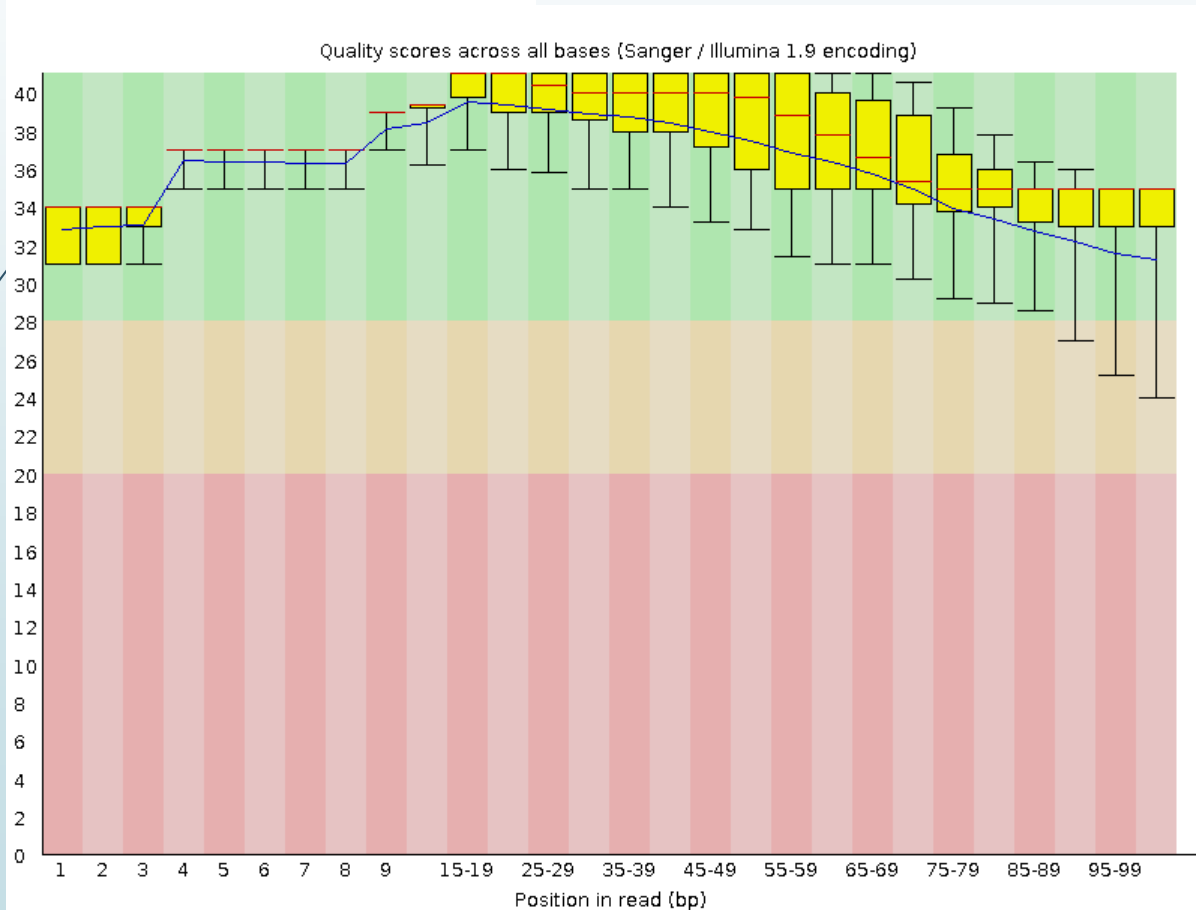
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	L7_GA120-6_NoIndex_L007_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	4000000
Filtered Sequences	0
Sequence length	100
%GC	49

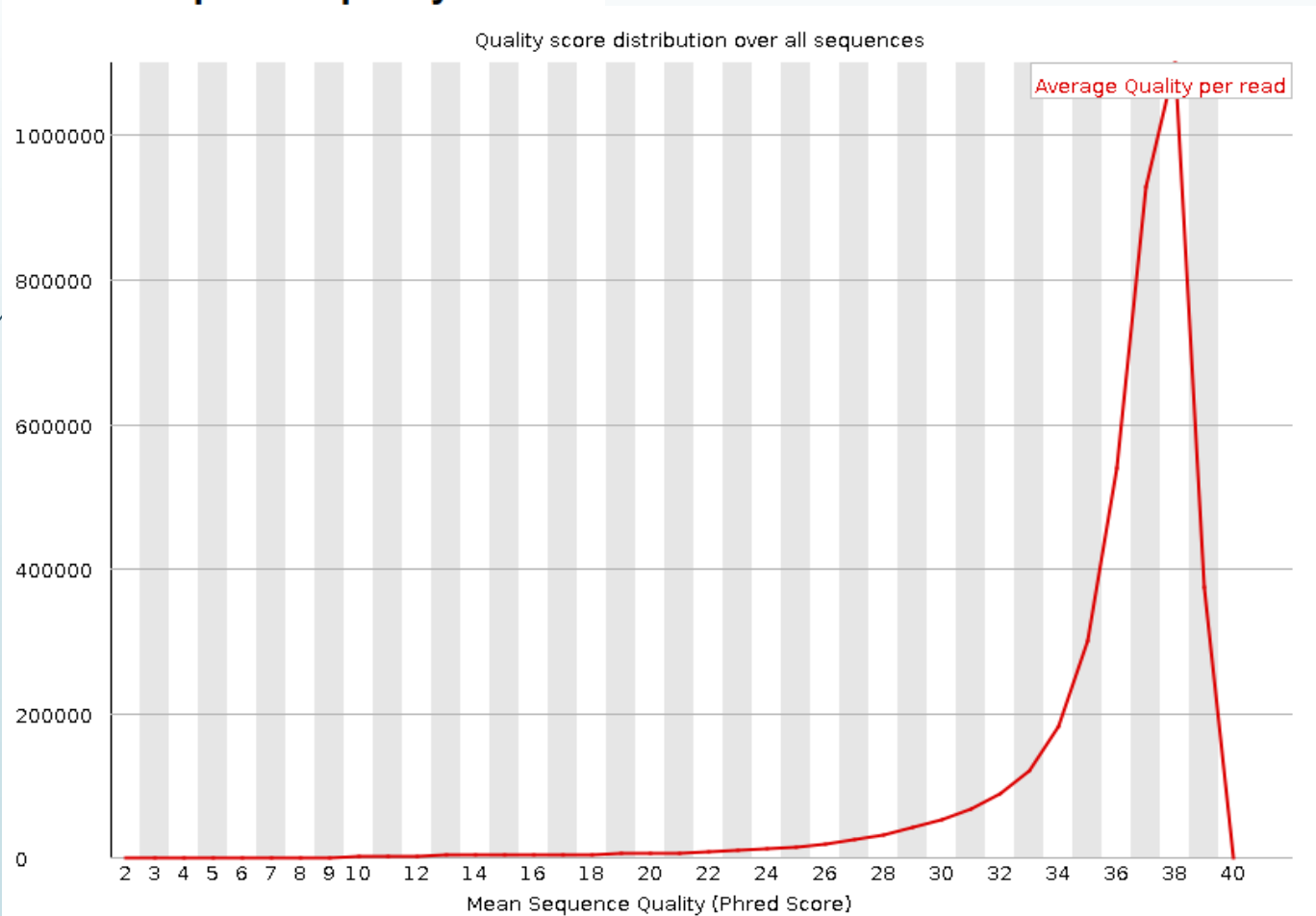
Visualize the Quality

✓ Per base sequence quality



Distribution of Quality Score

✓ Per sequence quality scores



FASTQ Trimmer

工具

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column**
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- FASTQ interlacer on paired end reads
- FASTQ de-interlacer on paired end reads
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTQ Trimmer (version 1.0.0)

FASTQ File:
41: Galaxy5-[brain_2...fastqsanger

Define Base Offsets as:
Absolute Values
Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

This tool allows you to trim the ends of reads.
You can specify either absolute or percent-based offsets. Offsets are calculated, starting at 0, from the respective end to be trimmed. When using the percent-based method, offsets are rounded to the nearest integer.
For example, if you have a read of length 36:

```
@Some FASTQ Sanger Read  
CAATATGINCTACTGATAAGTGGATATNAGCNCCA  
+  
=@@.@;B-%?8>CBA@>7@7BBCA4-48%<;;%<B@
```

And you set absolute offsets of 2 and 9:

```
@Some FASTQ Sanger Read  
ATATGINCTACTGATAAGTGGATA  
+  
@.@;B-%?8>CBA@>7@7BBCA4-4
```

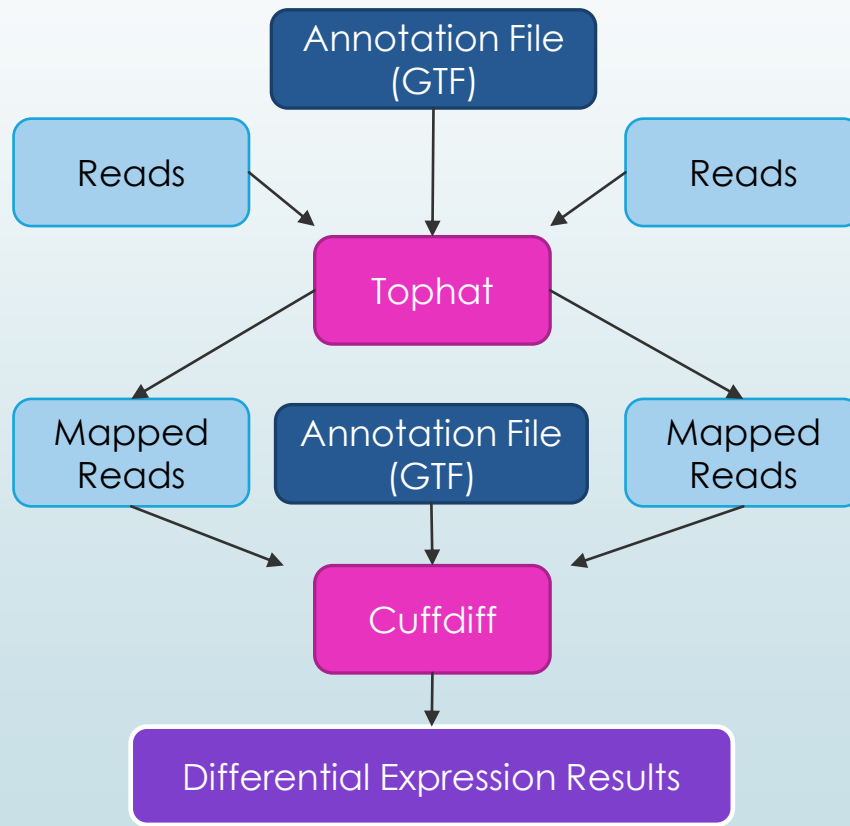
Or you set percent offsets of 6% and 20% (corresponds to absolute offsets of 2,7 for a read length of 36):

```
@Some FASTQ Sanger Read  
ATATGINCTACTGATAAGTGGATATN  
+  
@.@;B-%?8>CBA@>7@7BBCA4-48%
```

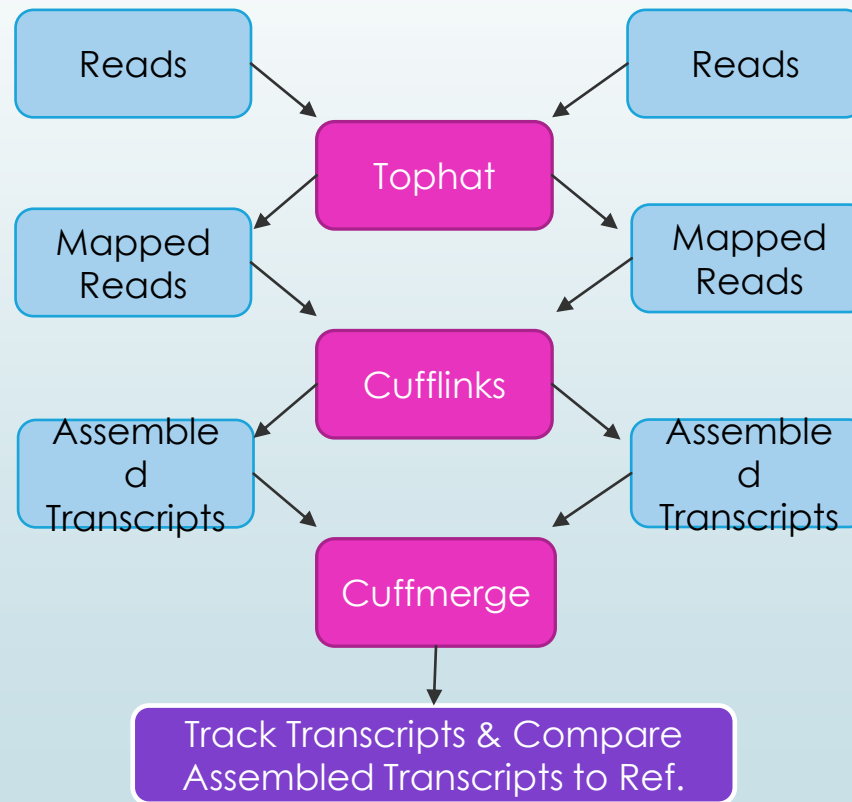
RNA-seq Analysis: Tuxedo Tools

Tool	Description
Bowtie	Ultrafast short read aligner
Tophat	Aligns RNA-seq reads to the genome using Bowtie Discovers splice sites
Cufflinks	Assembles transcripts
Cuffcompare	Compares your assembled transcripts to a reference annotation Tracks Cufflinks transcripts across multiple experiments
Cuffmerge	Merges two or more transcript assemblies
Cuffdiff	Finds significant changes in transcript expression, splicing, and promoter use

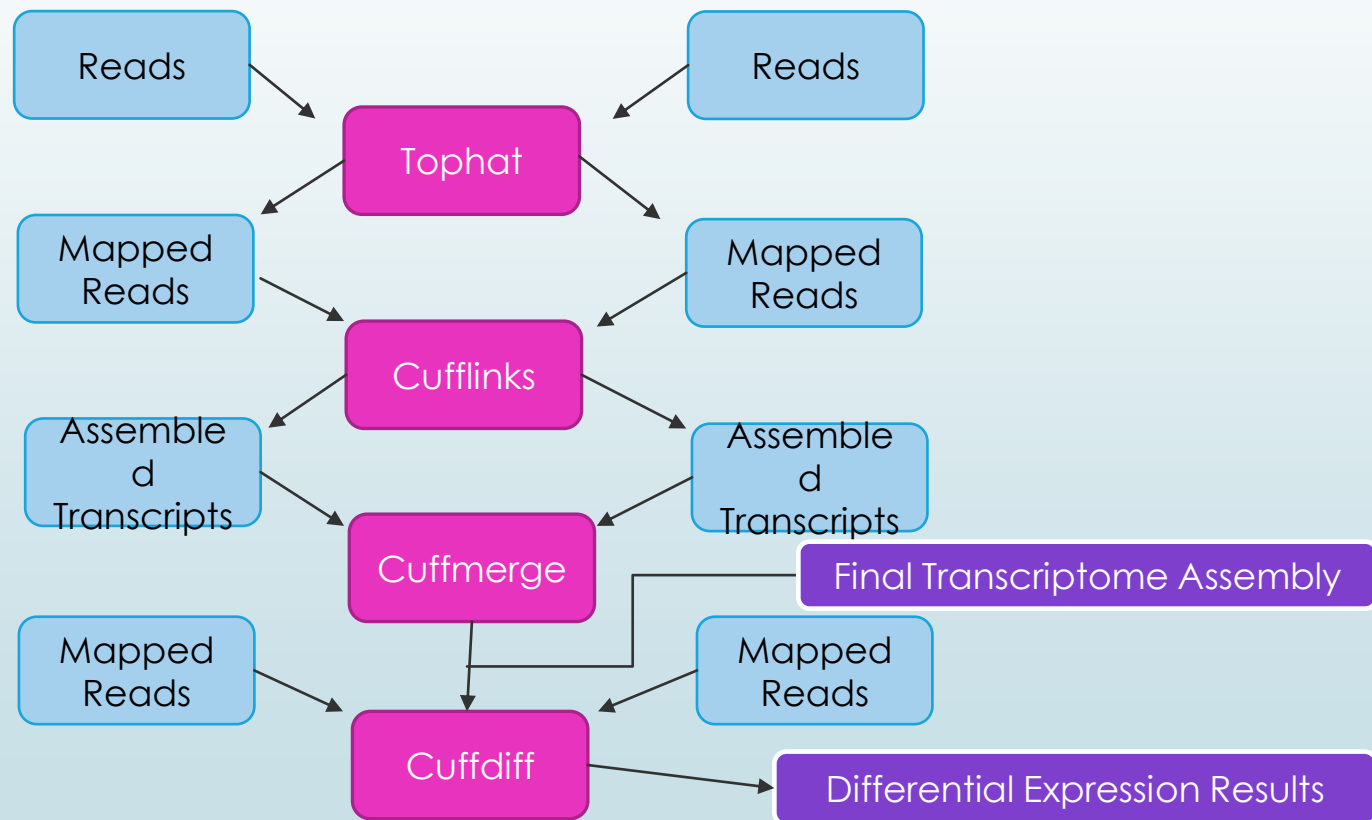
Differential Expression Analysis



Transcript Assembly and Transcript Comparison



Transcript Assembly and Differential Expression Analysis



Map The Reads (Tophat)

工具

- NGS: Mapping
- NGS: Indel Analysis
- NGS: RNA Analysis
 - RNA-SEQ
 - Tophat for Illumina** Find splice junctions using RNA-seq data
 - Tophat2 Gapped-read mapper for RNA-seq data
 - Tophat for SOLiD Find splice junctions using RNA-seq data
 - Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
 - Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
 - eXpress Quantify the abundances of a set of target sequences from sampled subsequences
 - Cuffmerge merge together several Cufflinks assemblies
 - Cuffdiff find significant

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:
40: Galaxy4-[brain_1...fastqsanger ▼
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:
Use one from the history ▼
Built-ins were indexed using default options

Select the reference genome:
42: http://hgdownload..es/chr19.fa ▼

Is this library mate-paired?:
Paired-end ▼

RNA-Seq FASTQ file:
41: Galaxy5-[brain_2...fastqsanger ▼
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:
110

TopHat settings to use:
Default settings ▼
Use the Full parameter list to change default settings.

Execute

Assemble Transcripts (Cufflinks)

工具

NGS: RNA Analysis

- RNA-SEQ
 - [Tophat for Illumina](#) Find splice junctions using RNA-seq data
 - [Tophat2](#) Gapped-read mapper for RNA-seq data
 - [Tophat for SOLiD](#) Find splice junctions using RNA-seq data
 - [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data**
 - [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
 - [eXpress](#) Quantify the abundances of a set of target sequences from sampled subsequences
 - [Cuffmerge](#) merge together several Cufflinks assemblies
 - [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use
- DE NOVO ASSEMBLY
 - [Trinity](#) De novo assembly of

Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:
54: Tophat for Illumi...cepted_hits ▾

Max Intron Length:
300000

Min Isoform Fraction:
0.1

Pre mRNA Fraction:
0.15

Perform quartile normalization:
No ▾
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:
Use reference annotation as guide ▾

Reference Annotation:
1: genes.gtf ▾
Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction:
No ▾
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):
Yes ▾

Merge Assemblies (Cuffmerge)

工具

NGS: RNA Analysis

RNA-SEQ

- [Tophat for Illumina](#) Find splice junctions using RNA-seq data
- [Tophat2](#) Gapped-read mapper for RNA-seq data
- [Tophat for SOLiD](#) Find splice junctions using RNA-seq data
- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- [eXpress](#) Quantify the abundances of a set of target sequences from

Cuffmerge (version 0.0.5)

GTF file produced by Cufflinks:

57: Cufflinks on data..transcripts ▾

Additional GTF Input Files

Add new Additional GTF Input Files

Use Reference Annotation:

Yes ▾

Reference Annotation:

1: genes.gtf ▾

Make sure your annotation file is in GTF format and that Galaxy knows that your file is GTF--not GFF.

Use Sequence Data:

No ▾

Use sequence data for some optional classification functions, including the addition of the p_id attribute required by Cuffdiff.

Execute

Identify Significant Changes (Cuffdiff)

工具

NGS: RNA Analysis

- RNA-SEQ
 - Tophat for Illumina Find splice junctions using RNA-seq data
 - Tophat2 Gapped-read mapper for RNA-seq data
 - Tophat for SOLiD Find splice junctions using RNA-seq data
 - Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
 - Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
 - eXpress Quantify the abundances of a set of target sequences from sampled subsequences
 - Cuffmerge merge together several Cufflinks assemblies
 - Cuffdiff find significant changes in transcript expression, splicing, and promoter use**
- DE NOVO ASSEMBLY
 - Trinity De novo assembly of

Cuffdiff (version 0.0.5)

Transcripts:
60: Cuffmerge on data..transcripts
A transcript GTF file produced by cufflinks, cuffcompare, or other source.

Perform replicate analysis:
No
Perform cuffdiff with replicates in each group.

SAM or BAM file of aligned RNA-Seq reads:
50: Tophat for Illumi..cepted_hits

SAM or BAM file of aligned RNA-Seq reads:
54: Tophat for Illumi..cepted_hits

False Discovery Rate:
0.05
The allowed false discovery rate.

Min Alignment Count:
10
The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

Perform quartile normalization:
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Perform Bias Correction:
No
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Differential Expression Results

test_id	gene_id	gene	locus	sample_1	sample_2	status
TCONS_00000001	XLOC_000001	DDX11L1	chr1:11873-29370	q1	q2	NOTEST
TCONS_00000002	XLOC_000002	OR4F5	chr1:69090-70008	q1	q2	NOTEST
TCONS_00000003	XLOC_000003	LOC100132062	chr1:323891-328581	q1	q2	NOTEST
TCONS_00000004	XLOC_000003	LOC100133331	chr1:323891-328581	q1	q2	NOTEST
TCONS_00000005	XLOC_000004	OR4F3	chr1:367658-368597	q1	q2	NOTEST
TCONS_00000006	XLOC_000005	LOC643837	chr1:763063-789740	q1	q2	NOTEST
TCONS_00000007	XLOC_000006	SAMD11	chr1:861120-894679	q1	q2	NOTEST
TCONS_00000008	XLOC_000007	KLHL17	chr1:895966-901099	q1	q2	NOTEST
TCONS_00000009	XLOC_000008	PLEKHN1	chr1:901876-910484	q1	q2	NOTEST
TCONS_00000010	XLOC_000008	PLEKHN1	chr1:901876-910484	q1	q2	NOTEST
TCONS_00000011	XLOC_000009	ISG15	chr1:948846-949919	q1	q2	NOTEST
TCONS_00000012	XLOC_000010	AGRN	chr1:955502-991499	q1	q2	NOTEST
TCONS_00000013	XLOC_000011	LOC254099	chr1:1072396-1079434	q1	q2	NOTEST
TCONS_00000014	XLOC_000012	MIR200B	chr1:1102483-1102578	q1	q2	NOTEST
TCONS_00000015	XLOC_000013	MIR200A	chr1:1103242-1103332	q1	q2	NOTEST
TCONS_00000016	XLOC_000014	MIR429	chr1:1104384-1104467	q1	q2	NOTEST
TCONS_00000017	XLOC_000015	TLL10	chr1:1109285-1133313	q1	q2	NOTEST
TCONS_00000018	XLOC_000015	TLL10	chr1:1109285-1133313	q1	q2	NOTEST
TCONS_00000019	XLOC_000016	B3GALT6	chr1:1167628-1170420	q1	q2	NOTEST
TCONS_00000020	XLOC_000017	SCNN1D	chr1:1215815-1227409	q1	q2	NOTEST
TCONS_00000021	XLOC_000017	SCNN1D	chr1:1215815-1227409	q1	q2	NOTEST
TCONS_00000022	XLOC_000018	PUSL1	chr1:1243993-1260067	q1	q2	NOTEST
TCONS_00000023	XLOC_000019	GLTPD1	chr1:1260142-1264276	q1	q2	NOTEST
TCONS_00000024	XLOC_000020	TAS1R3	chr1:1266725-1269844	q1	q2	NOTEST
TCONS_00000025	XLOC_000021	LOC148413	chr1:1334909-1342693	q1	q2	NOTEST
TCONS_00000026	XLOC_000022	TMEM88B	chr1:1361507-1363167	q1	q2	NOTEST
TCONS_00000027	XLOC_000023	VWA1	chr1:1370902-1378262	q1	q2	NOTEST
TCONS_00000028	XLOC_000023	VWA1	chr1:1370902-1378262	q1	q2	NOTEST

历史

71: Cuffdiff on data 50, data 54, and data 60: transcript FPKM tracking

70: Cuffdiff on data 50, data 54, and data 60: transcript differential expression testing

69: Cuffdiff on data 50, data 54, and data 60: gene FPKM tracking

68: Cuffdiff on data 50, data 54, and data 60: gene differential expression testing

67: Cuffdiff on data 50, data 54, and data 60: TSS groups FPKM tracking

66: Cuffdiff on data 50, data 54, and data 60: TSS groups differential expression testing

65: Cuffdiff on data 50, data 54, and data 60: CDS FPKM tracking

64: Cuffdiff on data 50, data 54, and data 60: CDS FPKM differential expression testing

Publish Your Workflow

Galaxy 分析

Published Workflows

search name, annotation, owner, and tags

Advanced Search

Galaxy 分析

Published Workflows | [daniel](#) | RNA-seq Analysis Exercise 20130912

Galaxy Workflow 'RNA-seq Analysis Exercise 20130912'

Step

Step 1: Input dataset

Input Dataset
select at runtime

Step 2: Input dataset

Input Dataset
select at runtime

Step 3: Input dataset

Input Dataset
select at runtime

Step 4: Input dataset

Input Dataset
select at runtime

Step 5: Input dataset

Step 8: Tophat for Illumina Using 5.6 Gb

RNA-Seq FASTQ file
Output dataset 'output' from step 4

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

Select the reference genome
Output dataset 'output' from step 6

Is this library mate-paired?
Paired-end

RNA-Seq FASTQ file
Output dataset 'output' from step 5

Mean Inner Distance between Mate Pairs
110

TopHat settings to use
Default settings

Step 9: Cufflinks


SAM or BAM file of aligned RNA-Seq reads
Output dataset 'accepted_hits' from step 7

Max Intron Length
300000

Min Isoform Fraction
0.1

Pre MRNA Fraction

Workflow Using 5.6 Gb


orkflows
[orkflows](#)
orkflows by daniel

(average) ★★★★★
★★★★★

none

Run Existing Workflow

Your workflows

Name

RNA-seq Analysis Exercise 20130912

imported: RNA-seq Analysis Exercise 20130912

RNA-seq

Workflows shared with you by

No workflows have been shared with you.

Other options

Configure your workflow menu

Running workflow "RNA-seq Analysis Exercise 20130912"

Expand All

Collapse

Step 1: Input dataset

Input Dataset

60: Cuffmerge on data..transcripts

type to filter

Step 2: Input dataset

Input Dataset

41: Galaxy5-[brain_2...fastqsanger

type to filter

Step 3: Input dataset

Input Dataset

41: Galaxy5-[brain_2...fastqsanger

type to filter

Step 4: Input dataset

Input Dataset

41: Galaxy5-[brain_2...fastqsanger

type to filter

Step 5: Input dataset

illumina® Log in to get personalized account information. Quick Order View Cart

Contact Us MyIllumina Tools

APPLICATIONS SYSTEMS CLINICAL SERVICES SCIENCE **SUPPORT** COMPANY

Support » Sequencing » Sequencing Software » **iGenomes** Follow us:

Ready-To-Use Reference Sequences and Annotations

The iGenomes are a collection of reference sequences and annotation files for commonly analyzed organisms. The files have been downloaded from Ensembl, NCBI, or UCSC, and chromosome names have been changed to be simple and consistent with their download source. Each iGenome is available as a compressed file that contains sequences and annotation files for a single genomic build of an organism.

For more information, see the [iGenomes Overview](#) and [Change Log](#).

Species	Source	Build(s)			
<i>Arabidopsis thaliana</i>	Ensembl	TAIR10	TAIR9		
	NCBI	TAIR10	build9.1		
<i>Bacillus_cereus</i> strain ATCC 10987	NCBI	2003-02-13			
<i>Bacillus_subtilis</i> strain 168	Ensembl	EB2			
<i>Bos taurus</i> (Cow)	Ensembl	UMD3.1	Btau_4.0		
	NCBI	UMD_3.1	Btau_4.6.1	Btau_4.2	
	UCSC	bosTau7	bosTau6	bosTau4	
<i>Caenorhabditis elegans</i>	Ensembl	WBcel215	WS210		
	NCBI	WS195	WS190		
	UCSC	ce10	ce6		
<i>Canis familiaris</i> (Dog)	Ensembl	CanFam3.1	BROADD2		
	NCBI	build3.1	build2.1		
	UCSC	canFam3	canFam2		
<i>Drosophila melanogaster</i>	Ensembl	BDGP5	BDGP5.25		
	NCBI	build5.41	build5.3	build5	build4.1
	UCSC	dm3			

http://support.illumina.com/sequencing/sequencing_software/igenome.ilmn

UCSC Genome Resource

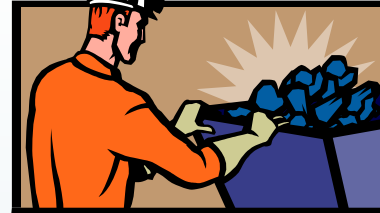
Human Genome

Dec. 2013 (hg38, GRCh38)

- [Full data set](#)
- [Data set by chromosome](#)
- [Annotation database](#)
- [Protein database for hg38](#)
- [LiftOver files](#)
- Pairwise Alignments
 - [Human/Chimp \(panTro4\)](#)
 - [Human/Rhesus \(rheMac3\)](#)
 - [Human/Mouse \(mm10\)](#)
 - [Human/Rat \(rn5\)](#)
 - [Human/Dog \(canFam3\)](#)
 - [Human/Opossum \(monDom5\)](#)
- Multiple Alignments
 - [Multiple alignments of 7 vertebrate genomes with Human](#)
 - [Conservation scores for alignments of 7 vertebrate genomes with Human](#)
 - [Basewise conservation scores \(phyloP\) of 7 vertebrate genomes with Human](#)
 - [FASTA alignments of 7 vertebrate genomes with Human for CDS regions](#)

<http://hgdownload.cse.ucsc.edu/downloads.html#human>

Literature



- ▶ Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.
Genome Biology 11, R86 (2010)
- ▶ Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.
Nature Protocols 7, 562–578 (2012)
- ▶ Full-length transcriptome assembly from RNA-Seq data without a reference genome.
Nature Biotechnology 29, 644–652 (2011)



Thank you !

Galaxy: An Open Platform for Data Intensive
Biomedical Research