Galaxy for Next Generation Sequencing

初探次世代序列分析平台

蘇聖堯

2013/9/12

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 team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University.

➧ The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

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.

Available Tools Pre-installed 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
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
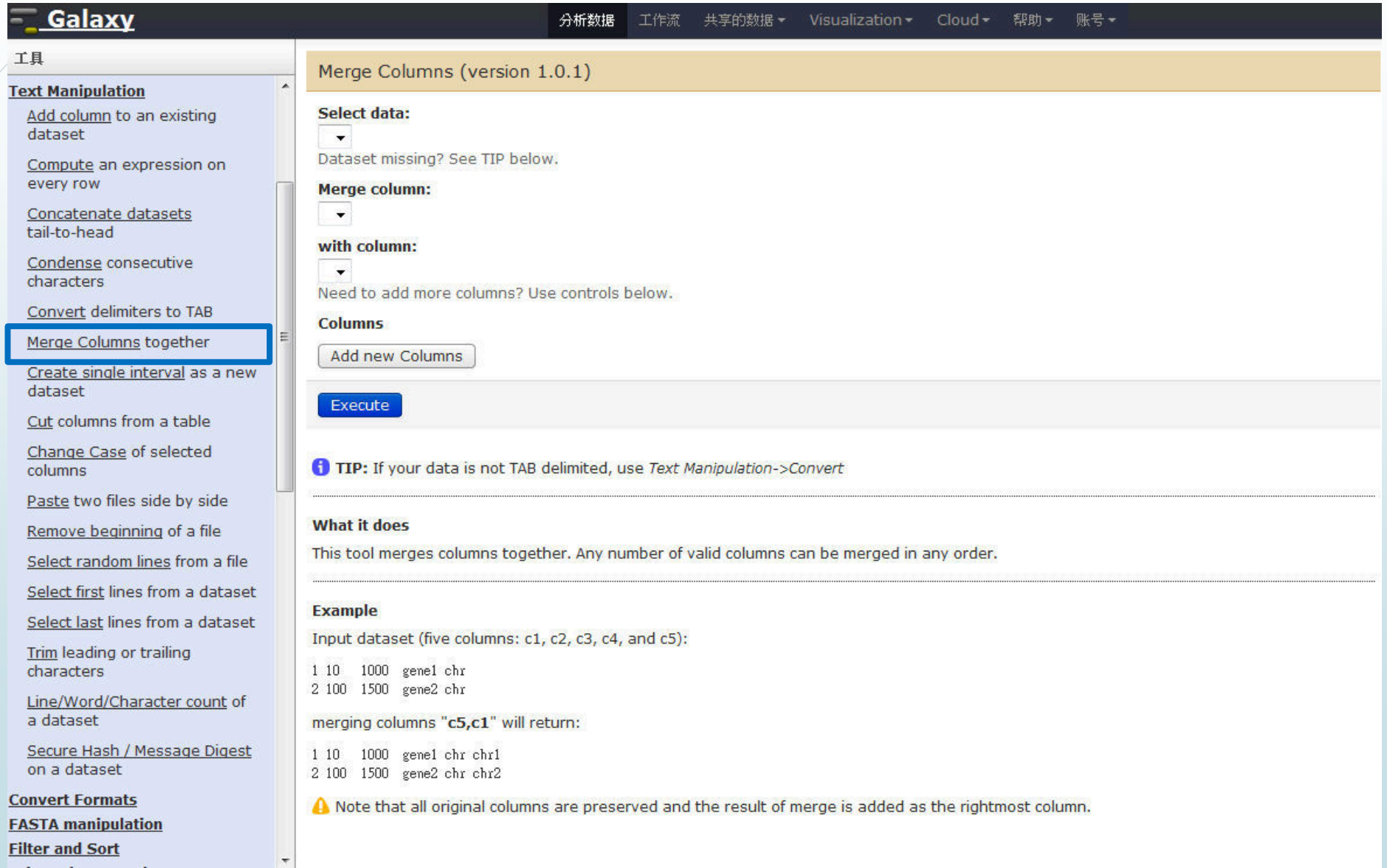VCF Tools

Portal of Galaxy

Computer Use Experience

➡ Traditional:

command mode execution

➡ Modern:

click , select file & **Execute**

Have You Experience These Commands On Linux?

- 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   combines two files based on the matching content lines found

- paste   merge contents of two files side by side

- split   split file into smaller files

Text Manipulation

Galaxy Can Help Analyses On

➥ 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)

Learning Garden

http://wiki.galaxyproject.org/Learn

Choice

CloudMan   SLIPSTREAM APPLIANCE GALAXY EDITION *Galaxy made easy.*

	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

Getting Started with Galaxy CloudMan

This page provides a step-by-step instructions on how to start your own instance of Galaxy on Amazon Web Services (AWS) Elastic Compute Cloud (EC2). More general information and instructions about Galaxy CloudMan can be found here.

AWS
Get Started
Capacity Planning
AMIs
↑ CloudMan

目錄

1. Screencast
2. Step 1: One Time Amazon S
3. Step 2: Starting a Master In
4. Step 3: Galaxy CloudMan W
5. Step 4: Use Galaxy as you r
6. Step 5: Shutting Down

Step 1: One Time Amazon Setup

1. Because AWS services implement pay-as-you-go access model for compute resources, it is necessary for every user of the service to *register with Amazon*. You will need a credit card to register. (You can apply for a AWS Education Grant after you register).
2. Once your account has been approved by Amazon (note that this may take up to one business day), *log into the EC2 AWS Management Console* and set your AWS Region to *US East (Virginia)*. This is the only region Galaxy CloudMan is fully supported in at this time (see screenshot 1.2).

Screencast

Here's a screencast that walks through the process of setting up your own Galaxy cloud cluster. Note, the screencast skips one step, detailed setting up of the Inbound TCP rules. See "Inbound Rules" below.

SlipStream

SlipS

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

Focus

The SlipS

consulting

INTRO

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.

GALAXY EDITION

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**.

Where to run Galaxy

- Main

  http://main.g2.bx.psu.edu/

- Public accessible servers

  http://wiki.galaxyproject.org/PublicGalaxyServers

- Local

  http://your-local-site/

Public Accessible Servers

Publicly Accessible Galaxy Servers

Galaxy'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>

Andromeda

- *Link:*
    Andromeda
- *Domain/Purpose:*
    This is a fully populated Galaxy instance.
- *Comments:*
    Andromeda is hosted at the SURFsara High Performance Computing (HPC) cloud. The installation is supported by Enis Afgan (CloudMan project) and Mattias de Hollander (NIOO).
- *Support:*

Andromeda: A cloud-based Galaxy

Data Upload

Data Format

➡ FASTA sequence format

➡ NGS file formats

    ➡ fastq, sam, bam

➡ UCSC file format specifications

    ➡ bed, wig, gtf, gff

Keep In Mind The Fundamental Concept

➡ Raw data (generated from sequencer): FASTQ

➡ After short read mapping 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
YYYYYYYYYYYYYYYYYYWYWYYSU
```

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
```

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.

```
Coor      12345678901234   567890123456789012345678901234567890012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1          TTAGATAAAGGATA*CTG
+r002          aaaAGATAA*GGATA
+r003        gcctaAGCTAA
+r004                        ATAGCT..............TCAGC
-r003                              ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

The corresponding SAM format is:

```
@HD  VN:1.5  SO:coordinate
@SQ  SN:ref  LN:45
r001   163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003     0 ref  9 30 5S6M        *  0   0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M     *  0   0 ATAGCTTCAGC        *
r003 2064 ref 29 17 6H5M         *  0   0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001    83 ref 37 30 9M          =  7 -39 CAGCGGCAT          * NM:i:1
```

More About SAM Format

test_mRNA_3_187_51   99   test_chromosome
53   255   75M   =   163   0   TACTATTTGACTAGACTGGAGGCGCTTG
CGACTGAGCTAGGACGTGCCACTACGGGGATGACGACTCGGACTACG
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII   NM:i:2

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 POSistion
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

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

Text Manipulation

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_+
GTAGGAATCGCAGCGCCAGCGGTTGCAAG
>hg17_chr7_127485994_127486166_+
GCCCAAGAAGCCCATCCTGGGAAGGAAAATGCATTGGGGAACCCTGTGCG
GATTCTTGTGGCTTTGGCCCTATCTTTTCTATGTCCAAGCTGTGCCCATC
CAAAAAGTCCAAGATGACACCAAAACCCTCATCAAGACAATTGTCACCAG
GATCAATGACATTTCACACACG
>hg17_chr7_127486011_127486166_+
TGGGAAGGAAAATGCATTGGGGAACCCTGTGCGGATTCTTGTGGCTTTGG
CCCTATCTTTTCTATGTCCAAGCTGTGCCCATCCAAAAAGTCCAAGATGA
CACCAAAACCCTCATCAAGACAATTGTCACCAGGATCAATGACATTTCAC
ACACG
```

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

```
chr7   127475281   127475310   NM_000230   0   +   GTAGGAATCGCAGCGCCAGCGGTTGCAAG
chr7   127485994   127486166   NM_000230   0   +   GCCCAAGAAGCCCATCCTGGGAAGGAAAATGCATTGGGGAACCCTGTGCGGATTCTTGTGGCTTTGG
chr7   127486011   127486166   D49487   0   +   TGGGAAGGAAAATGCATTGGGGAACCCTGTGCGGATTCTTGTGGCTTTGGCCCTATCTTTTCTATGTCCAAGCTG
```

Filter and Sort

Convert Formats

Reverse Complement

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
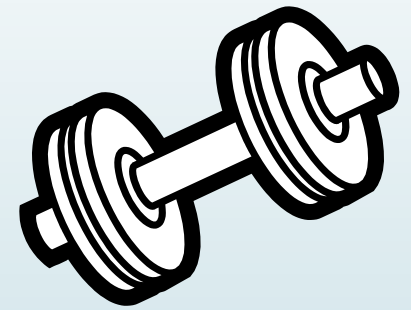
FastQC: NGS Quality Control

✅ **Per base sequence quality**
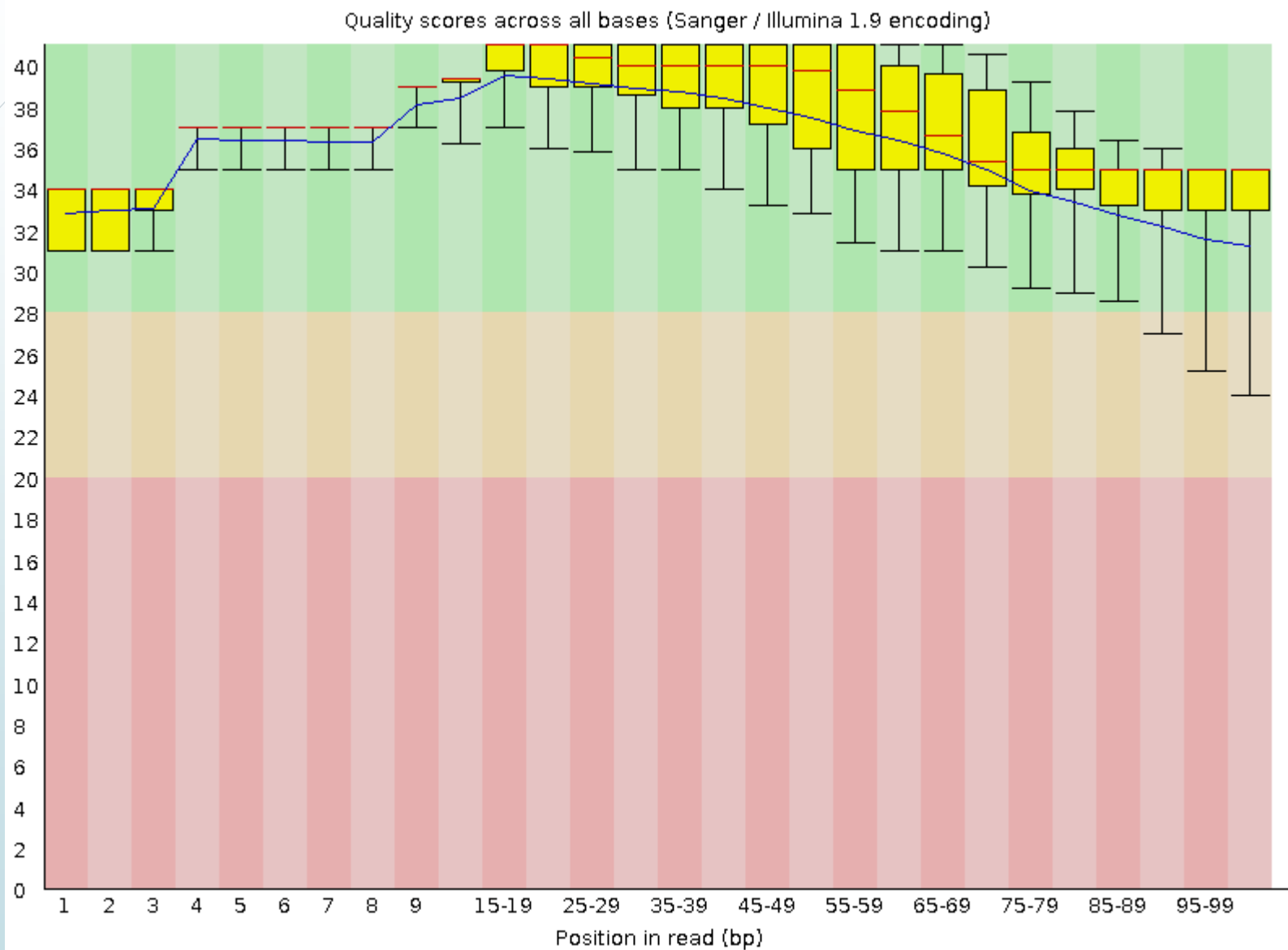
✅ Per sequence quality scores

Quality score distribution over all sequences

FASTQ Trimmer

工具 ⚙
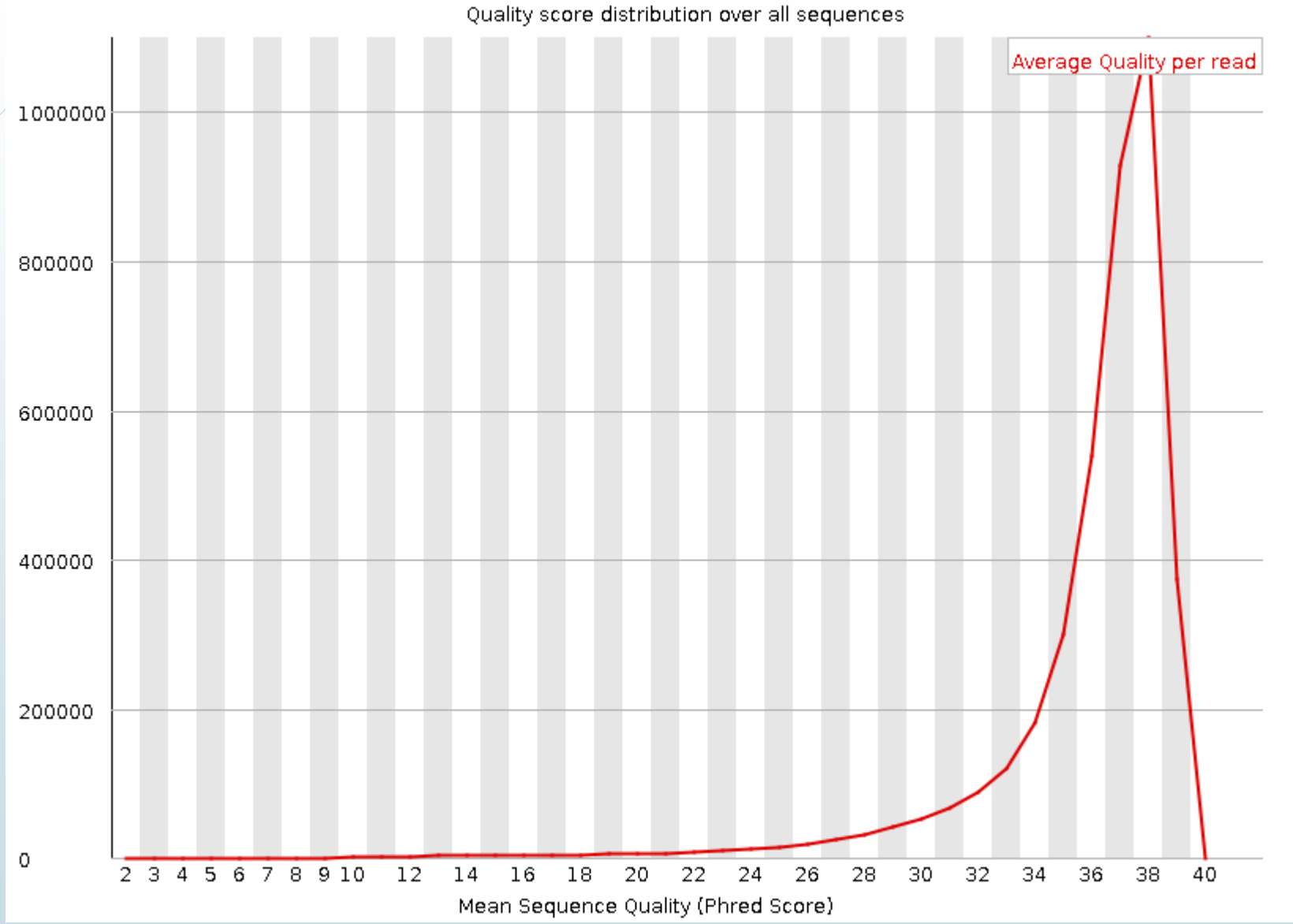
GENERIC FASTQ MANIPULATION

- 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 paire end reads
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTY TOOLKIT FOR

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
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA
+
=@@.@;B-%?8>CBA@>7@7BBCA4-48%<;;%<B@
```

And you set absolute offsets of 2 and 9:

```
@Some FASTQ Sanger Read
ATATGTNCTCACTGATAAGTGGATA
+
@.@;B-%?8>CBA@>7@7BBCA4-4
```
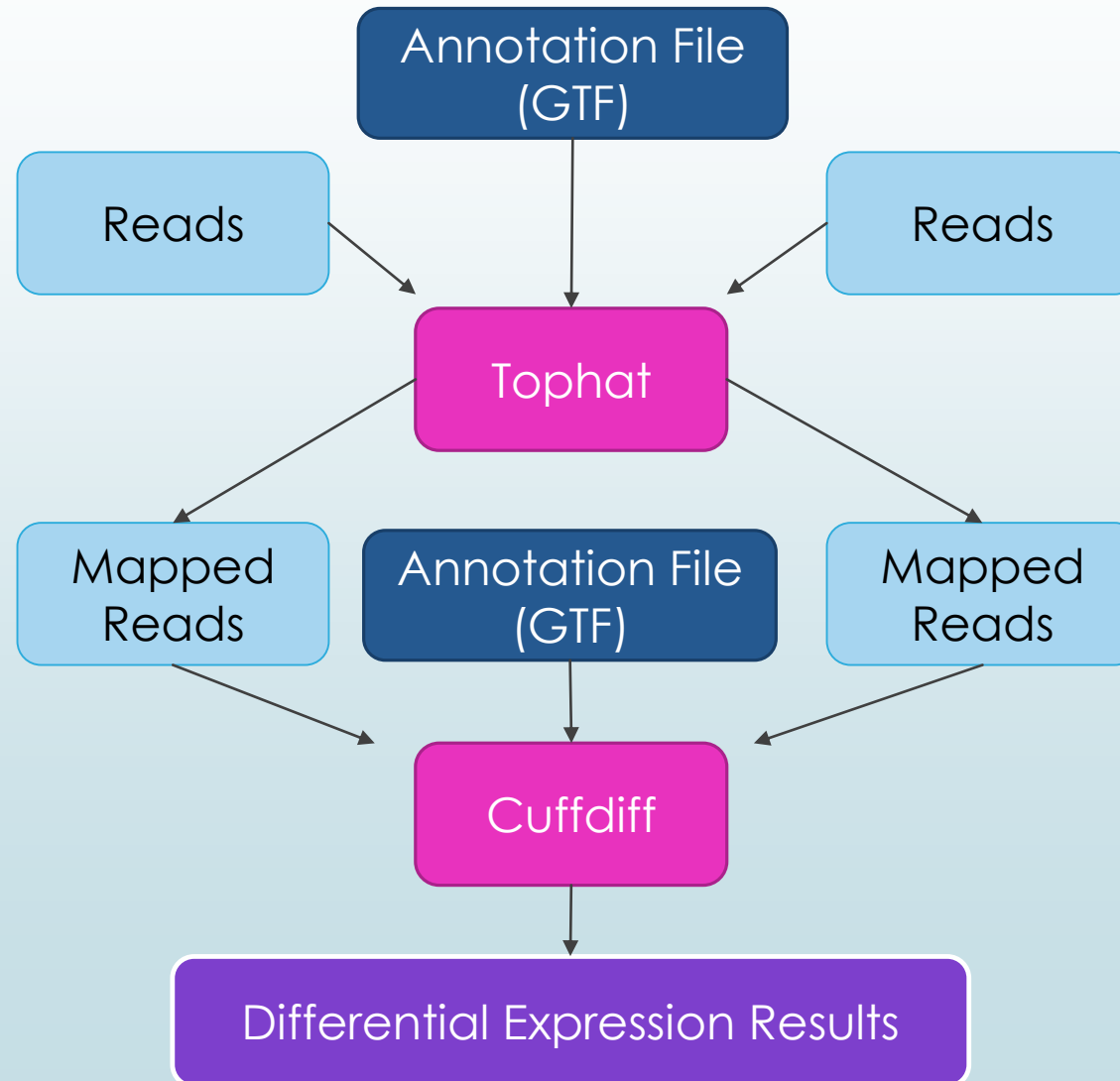
Or you set percent offsets of 6% and 20% (corresponds to absolute offsets of 2,7 for a read length of 36):

```
@Some FASTQ Sanger Read
ATATGTNCTCACTGATAAGTGGATATN
+
@.@;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
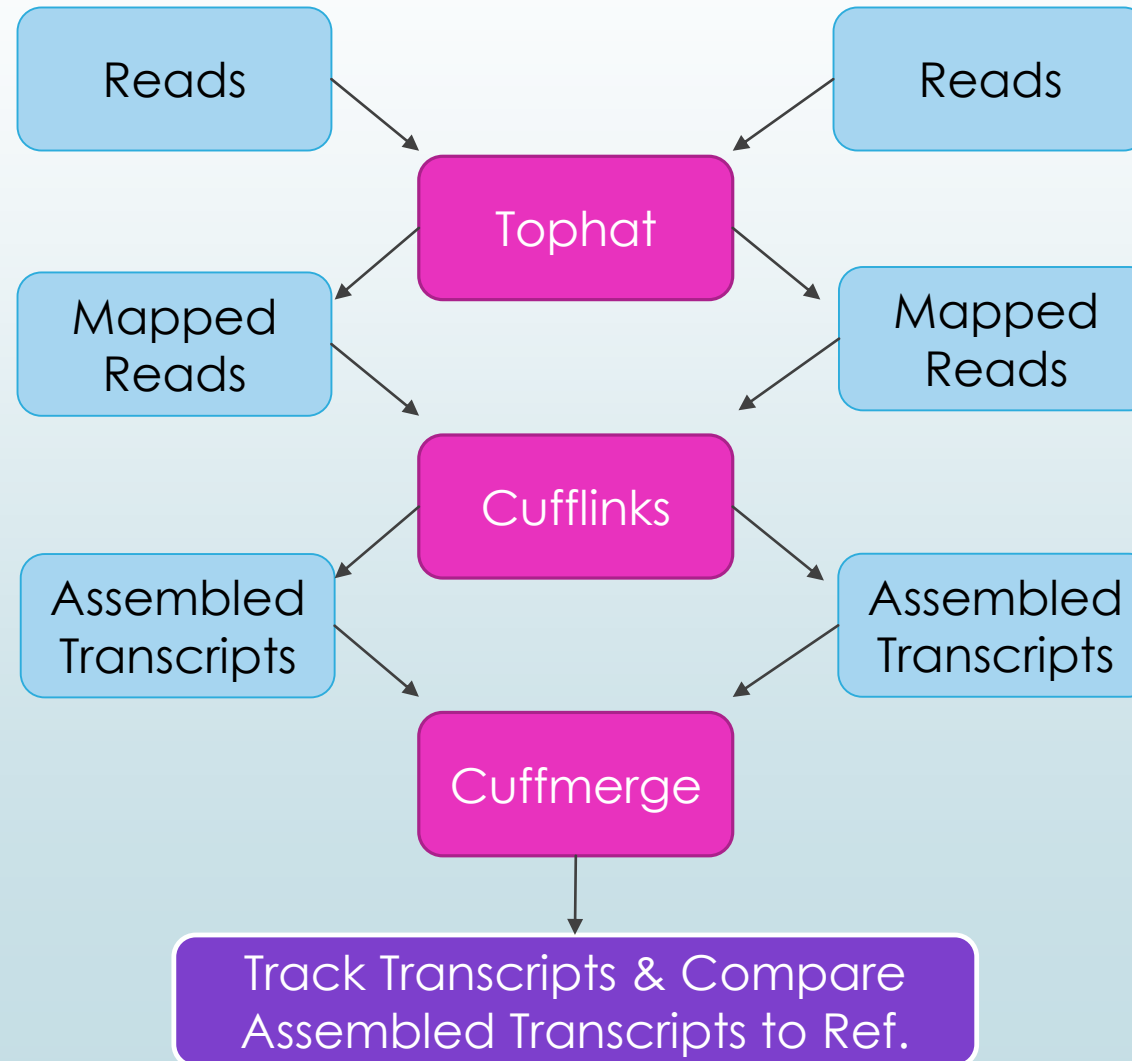
Transcript Assembly and Differential Expression Analysis

Map The Reads (Tophat)

Assemble Transcripts (Cufflinks)

Merge Assemblies (Cuffmerge)

Identify Significant Changes (Cuffdiff)

Differential Expression Results

test_id	gene_id	gene	locus	sample_1	sample_2	status	val
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	TTLL10	chr1:1109285-1133313	q1	q2	NOTEST	
TCONS_00000018	XLOC_000015	TTLL10	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

Published Workflows

search name, annotation, owner, and tags

Advanced Search

Galaxy

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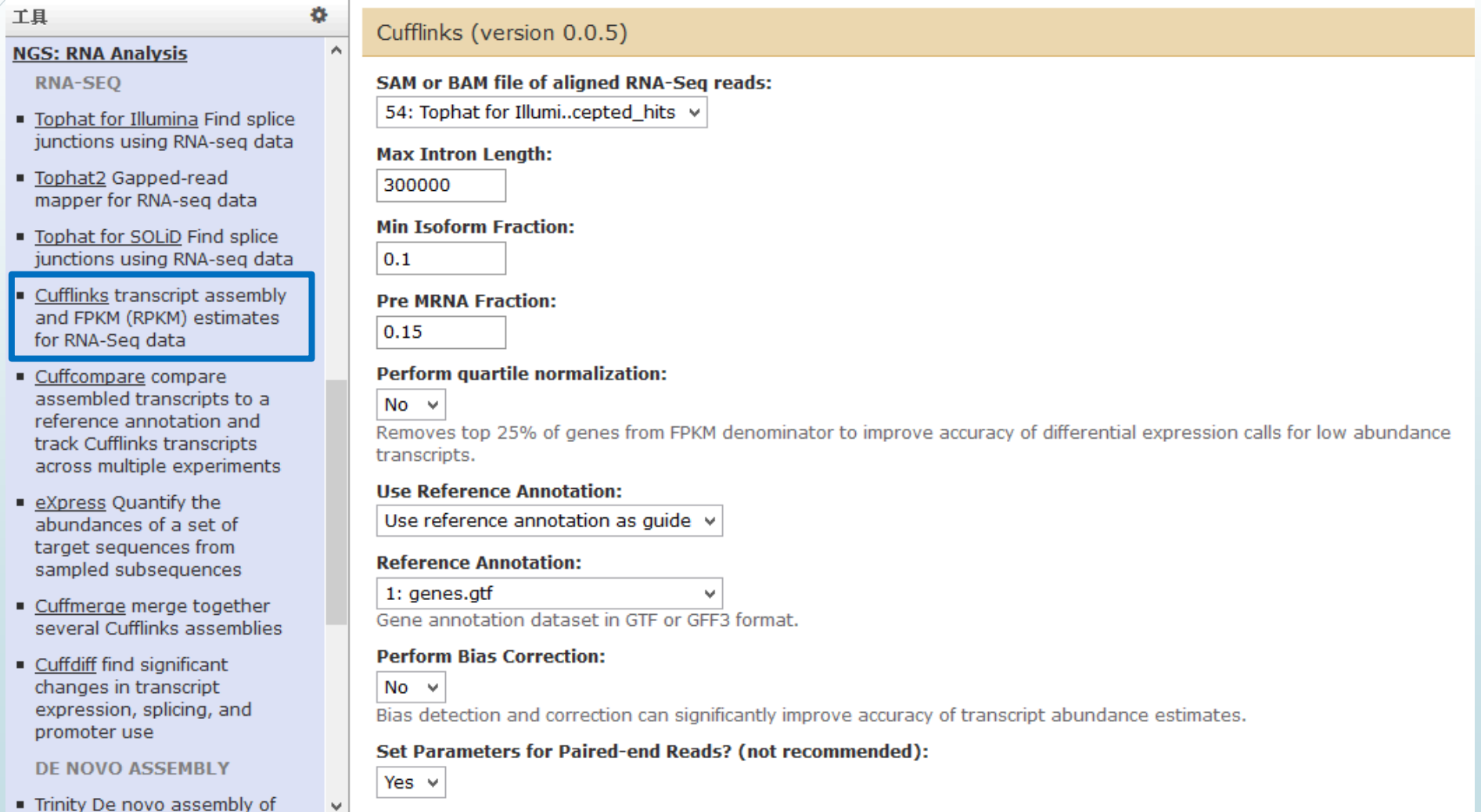
RNA-Seq FASTQ file
Output dataset 'output' from step 5

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

Using 5.6 Gb

Vorkflow

orkflows
l workflows
orkflows by daniel

☆☆☆☆☆

(average)

☆☆☆☆☆

Run Existing Workflow

iGenome

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

Take Home Message

Galaxy is very powerful!

Its user-friendly interface allows biologists to perform complex genomic analyses. (RNA-seq, ChIP-seq, SNP analysis, etc)

The use of computing resources should be taken into account.

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 for Next Generation Sequencing

初探