RNA-Seq in Galaxy: Tuxedo protocol

Igor Makunin, UQ RCC, QCIF

Acknowledgments

Genomics Virtual Lab: gvl.org.au
Galaxy for tutorials: galaxy-tut.genome.edu.au
Galaxy Australia: galaxy-aust.genome.edu.au

Contributors and participants:

RDS Research Data Services

qcif

nectar

QRIS cloud
QRISnews for Leaders in eResearch

EMBL Australia
Bioinformatics Resource

The University of Queensland Australia

RCC Research Computing Centre

QFAB Bioinformatics

genomics virtual lab

Melbourne Bioinformatics
Bioinformatics + data services + infrastructure, for life sciences today

Plan for today

Galaxy

Data types used in RNA-Seq analysis

RNA-Seq practical

Galaxy workflow

High-throughput sequencing

Big scale sequencing

- 100,000,000s sequences, or reads, per experiment
- sequencing of a (*random*) library
- low cost per nucleotide

Popular technologies:

- illumina
- ion / proton
- PacBio

Emerging technologies

- Oxford Nanopore MinION

Analysis of NGS data
Big datasets
Computationally intensive
Dedicated tools and data types
Extensive use of public data

Computational resources

Tools

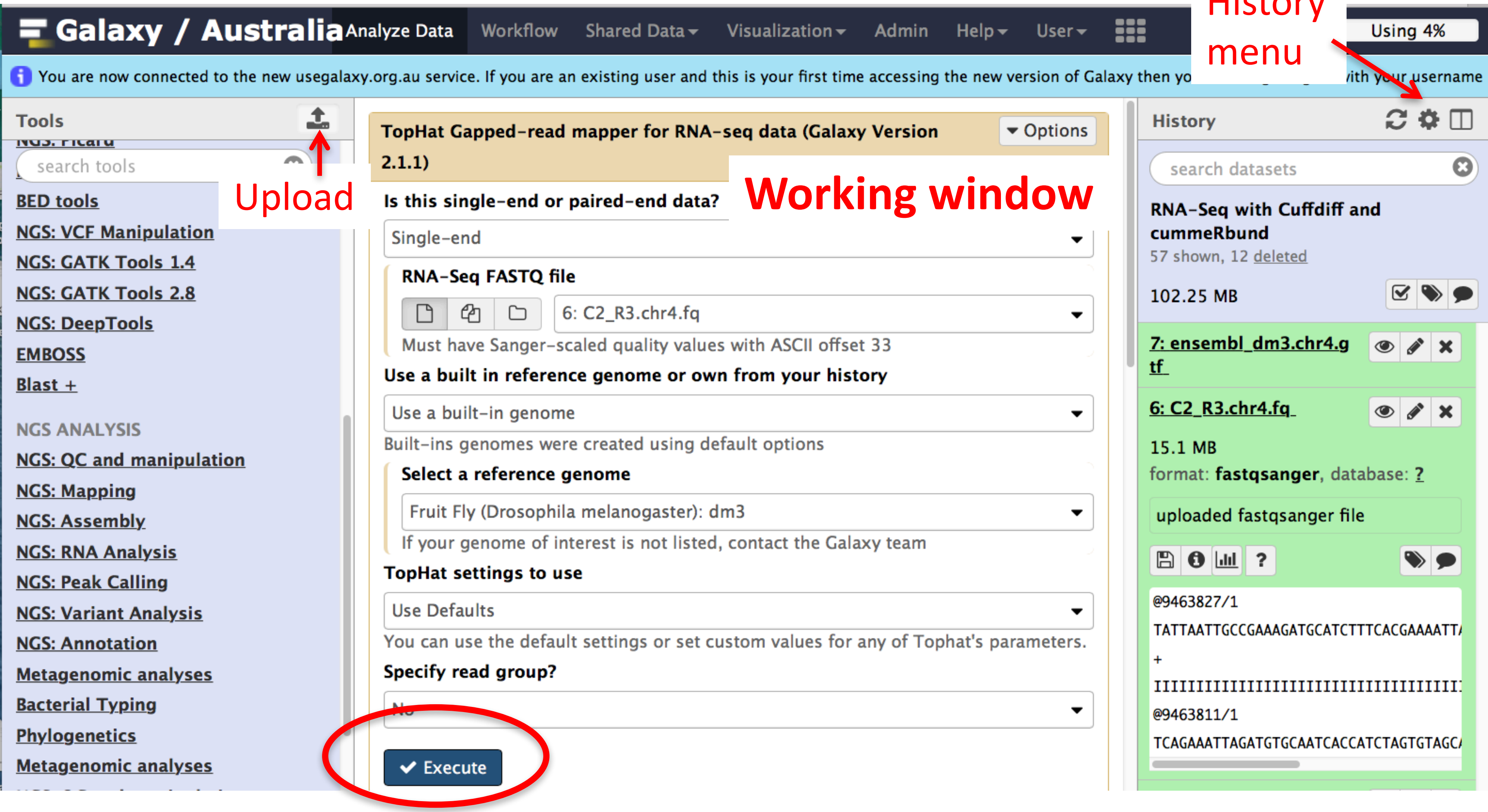
Storage

Public data

Galaxy

Knowledge and skills

Galaxy: how does it look like



You are now connected to the new usegalaxy.org.au service. If you are an existing user and this is your first time accessing the new version of Galaxy then yo ... with your username

TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.1)

Is this single-end or paired-end data?: Single-end

RNA-Seq FASTQ file: 6: C2_R3.chr4.fq

Must have Sanger-scaled quality values with ASCII offset 33

Use a built in reference genome or own from your history: Use a built-in genome

Built-ins genomes were created using default options

Select a reference genome: Fruit Fly (Drosophila melanogaster): dm3

If your genome of interest is not listed, contact the Galaxy team

TopHat settings to use: Use Defaults

You can use the default settings or set custom values for any of Tophat's parameters.

Specify read group?: No

RNA-Seq with Cuffdiff and cummeRbund

57 shown, 12 deleted

102.25 MB

7: ensembl_dm3.chr4.gtf

6: C2_R3.chr4.fq

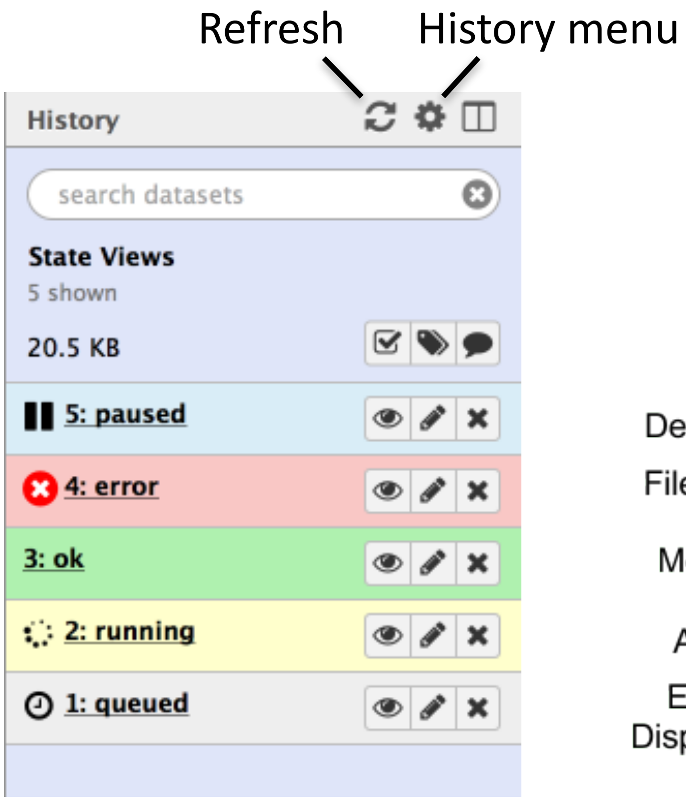
15.1 MB

format: **fastqsanger**, database: ?

uploaded fastqsanger file

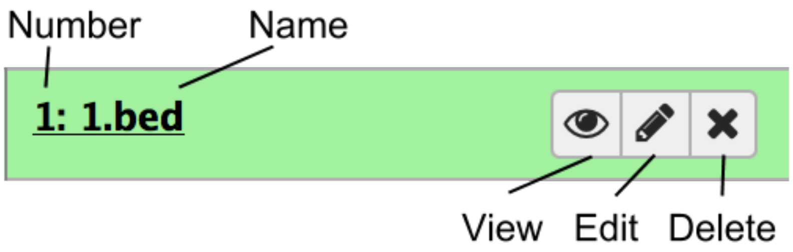
```
@9463827/1
TATTAATTGCCGAAAGATGCATCTTTCACGAAAATT
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@9463811/1
TCAGAAATTAGATGTGCAATCACCATCTAGTGTAGC
```

Galaxy history system



Refresh

History menu



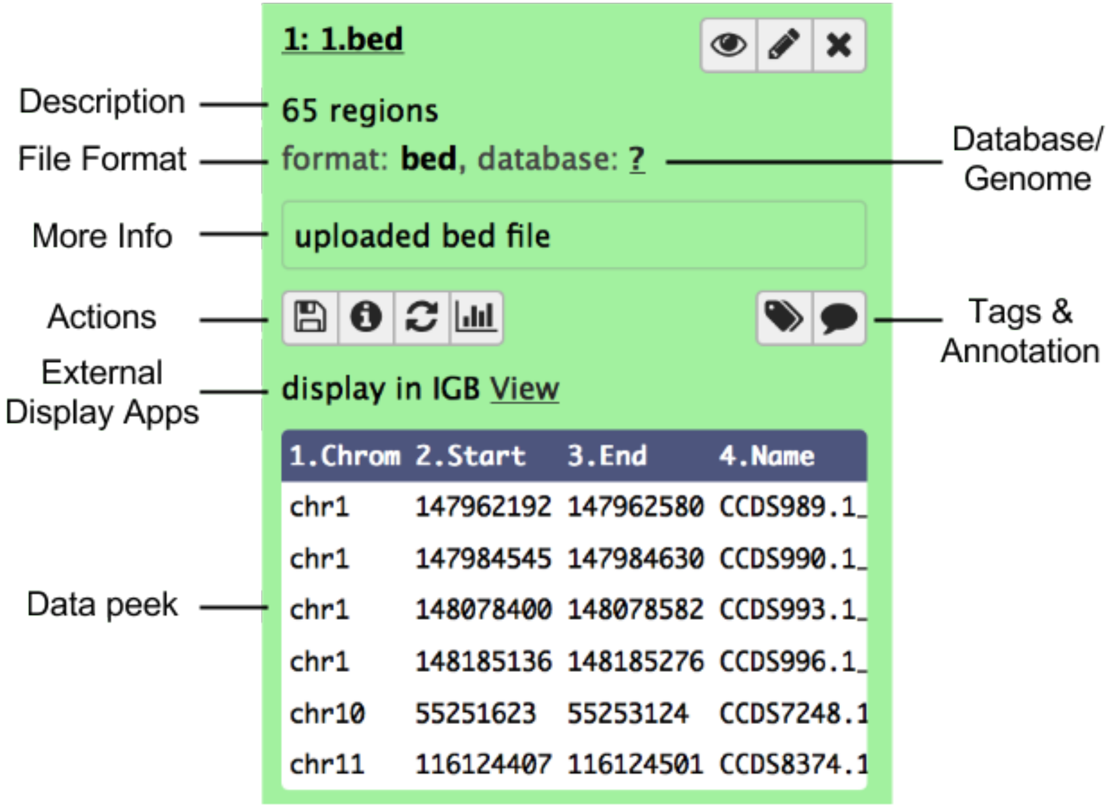
Number

Name

View

Edit

Delete



Description

File Format

More Info

Actions

External Display Apps

Data peek

Database/ Genome

Tags & Annotation

1: 1.bed

65 regions

format: **bed**, database: ?

uploaded bed file

display in IGB View

1.Chrom	2.Start	3.End	4.Name
chr1	147962192	147962580	CCDS989.1_
chr1	147984545	147984630	CCDS990.1_
chr1	148078400	148078582	CCDS993.1_
chr1	148185136	148185276	CCDS996.1_
chr10	55251623	55253124	CCDS7248.1
chr11	116124407	116124501	CCDS8374.1

Source: http://galaxyproject.github.io/training-material/topics/introduction/tutorials/galaxy-intro-history/tutorial.html

Public Galaxy servers

Galaxy servers:
usegalaxy.org
usegalaxy.eu

galaxy-tut.genome.edu.au

galaxy-aust.genome.edu.au

- *Independent registration on every Galaxy server*
- *Different tools, different user policy*
- *Data can be moved between Galaxy servers*

Advantage of the registration:

- *access to histories over long time*
- *multiple histories*
- *ability to use Galaxy from different devices*
- *bigger quotas (on some servers)*
- *ftp*

Galaxy Australia

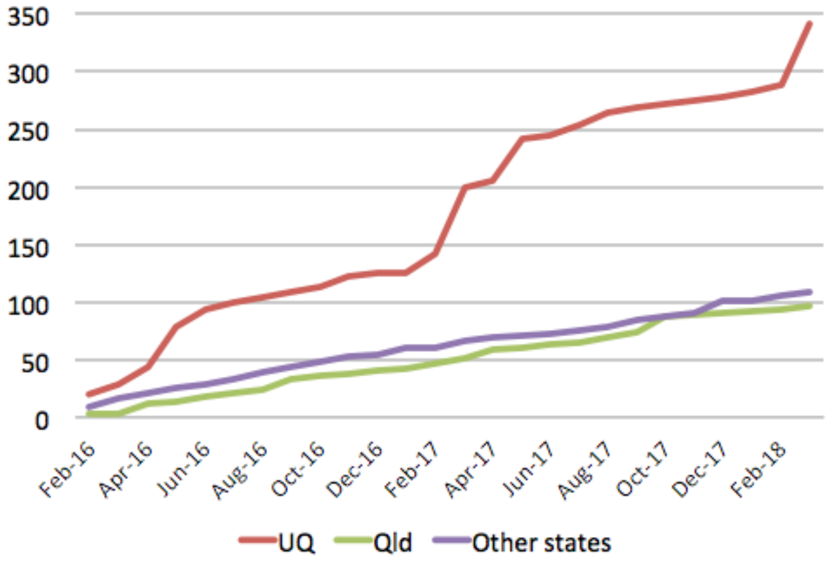
galaxy-aust.genome.edu.au

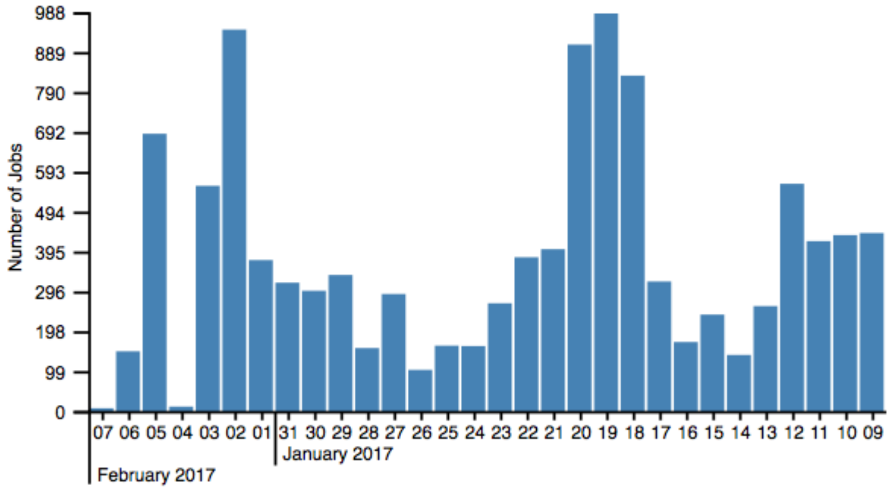
Worker nodes:
16 CPUs, 64 GB RAM

49 Tb Volume storage (user data)

Designed for a genome scale research
>1,600 registered users

Up to 16 CPUs 60 GB RAM per job
Up to 12 concurrent jobs per user
Up to 1 Tb per user

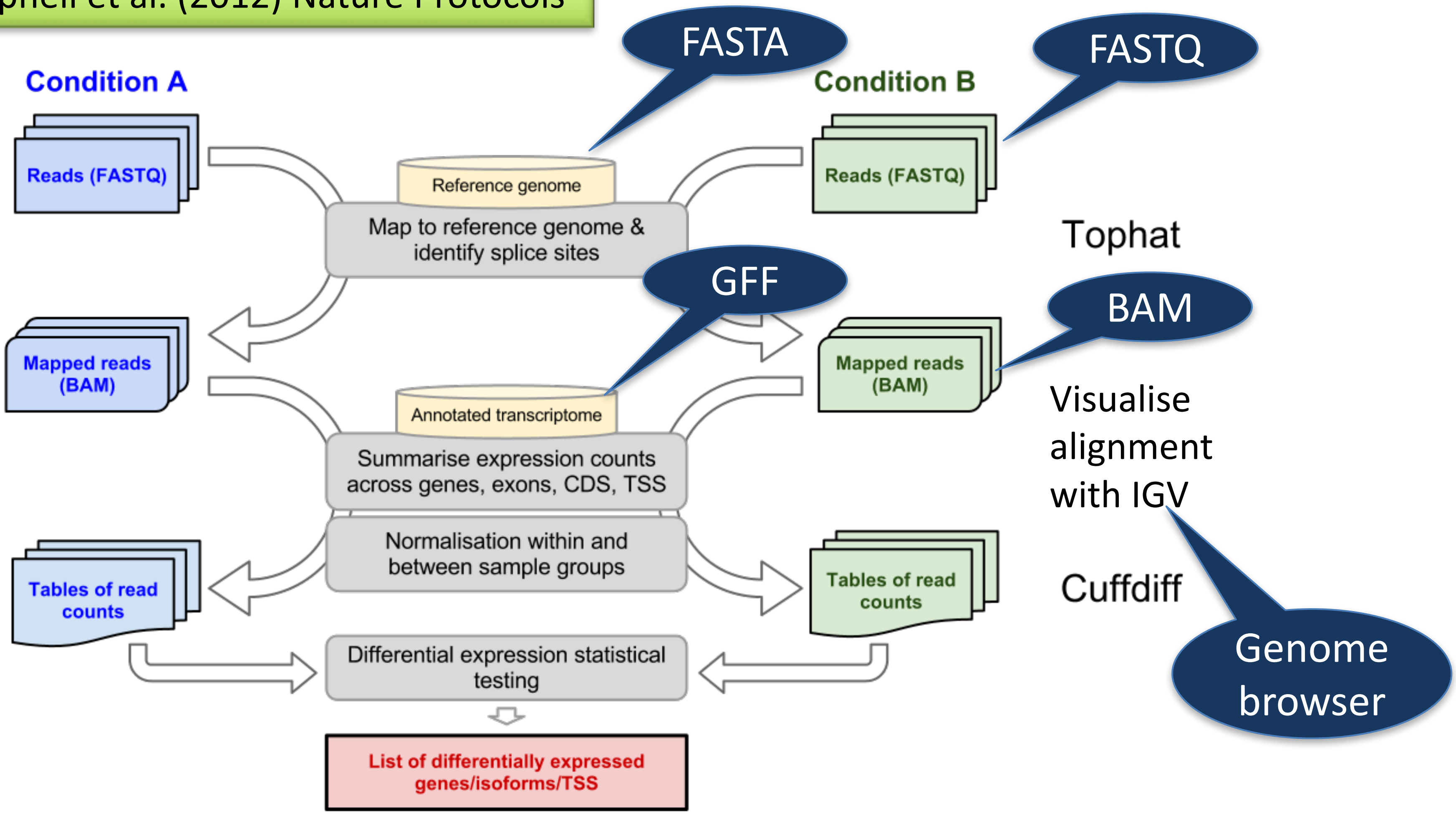

Galaxy-qld: Australian users


Jobs per day

Less jobs on weekends

Tuxedo protocol

GVL Basic RNA-Seq Galaxy tutorial
Trapnell et al. (2012) Nature Protocols

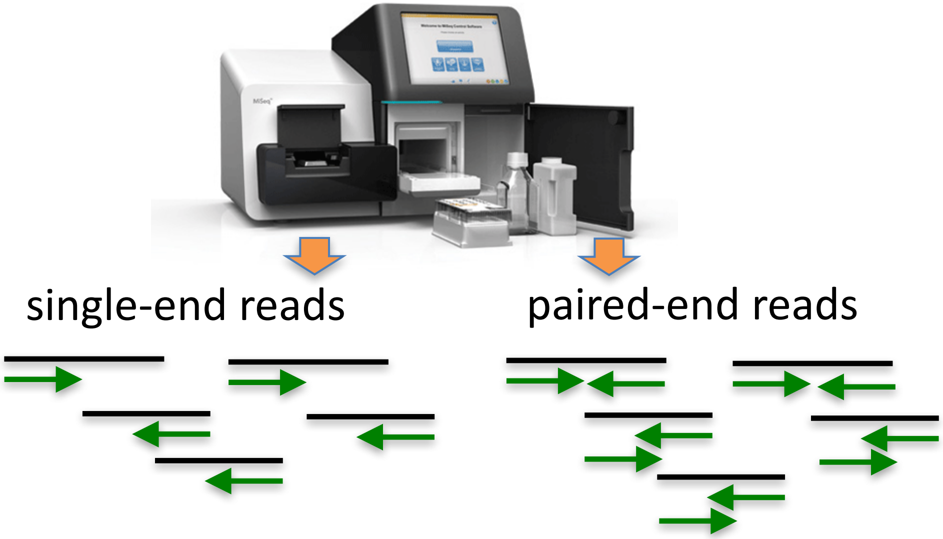

FASTA
FASTQ
Condition A
Condition B
Reads (FASTQ)
Reads (FASTQ)
Reference genome
Map to reference genome & identify splice sites
GFF
BAM
Mapped reads (BAM)
Mapped reads (BAM)
Annotated transcriptome
Summarise expression counts across genes, exons, CDS, TSS
Normalisation within and between sample groups
Tables of read counts
Tables of read counts
Differential expression statistical testing
List of differentially expressed genes/isoforms/TSS
Genome browser

Tophat

Visualise alignment with IGV

Cuffdiff

FASTQ format



single-end reads

paired-end reads

Terminology: *read* is a sequence with quality score values produced by a sequencing machine

Multiple reads in a single FASTQ file
Each read is described by four lines

Common output format: *FASTQ* compressed with gzip, *e.g.* SRR3145_1.fq.gz

```
@SRR3145.19 ILLUMINA-C32_FC:3:1:80:12/1
TAGCAGCACATCATGGTTTACATCGTATGC
+
IIHIDIIIIIIIIIIIIIIHIHIIIIIIDGIB
```

Name always starts with @
Sequence
Always starts with +; may have name
Encoded Phred quality score

FASTQ Phred quality score

A Phred quality score is a measure of the quality of the identification for every nucleotide.

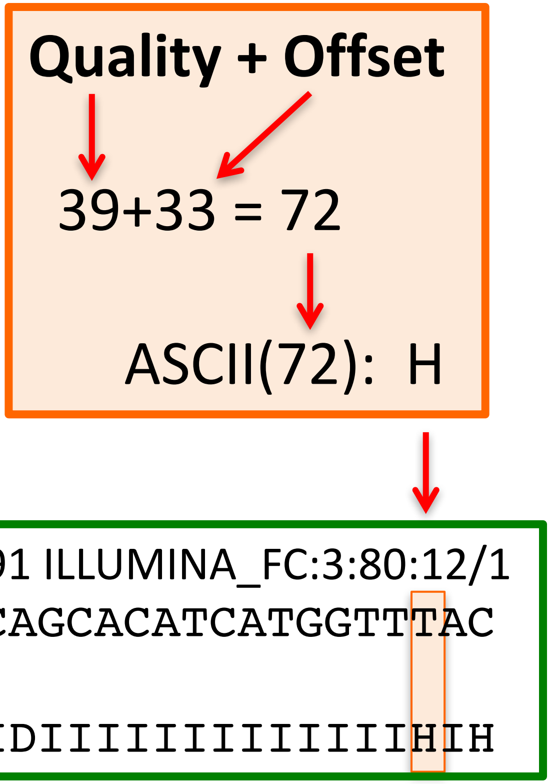
$$\boxed{Q_{\text{sanger}} = -10 \log_{10} p}$$

Range: ~0 to ~40

Phred 10: accuracy 90%
Phred 20: accuracy 99%
Phred 30: accuracy 99.9%
Phred 40: accuracy 99.99%

Values are encoded by characters

Advantage: a single character is used instead of a two-digit number



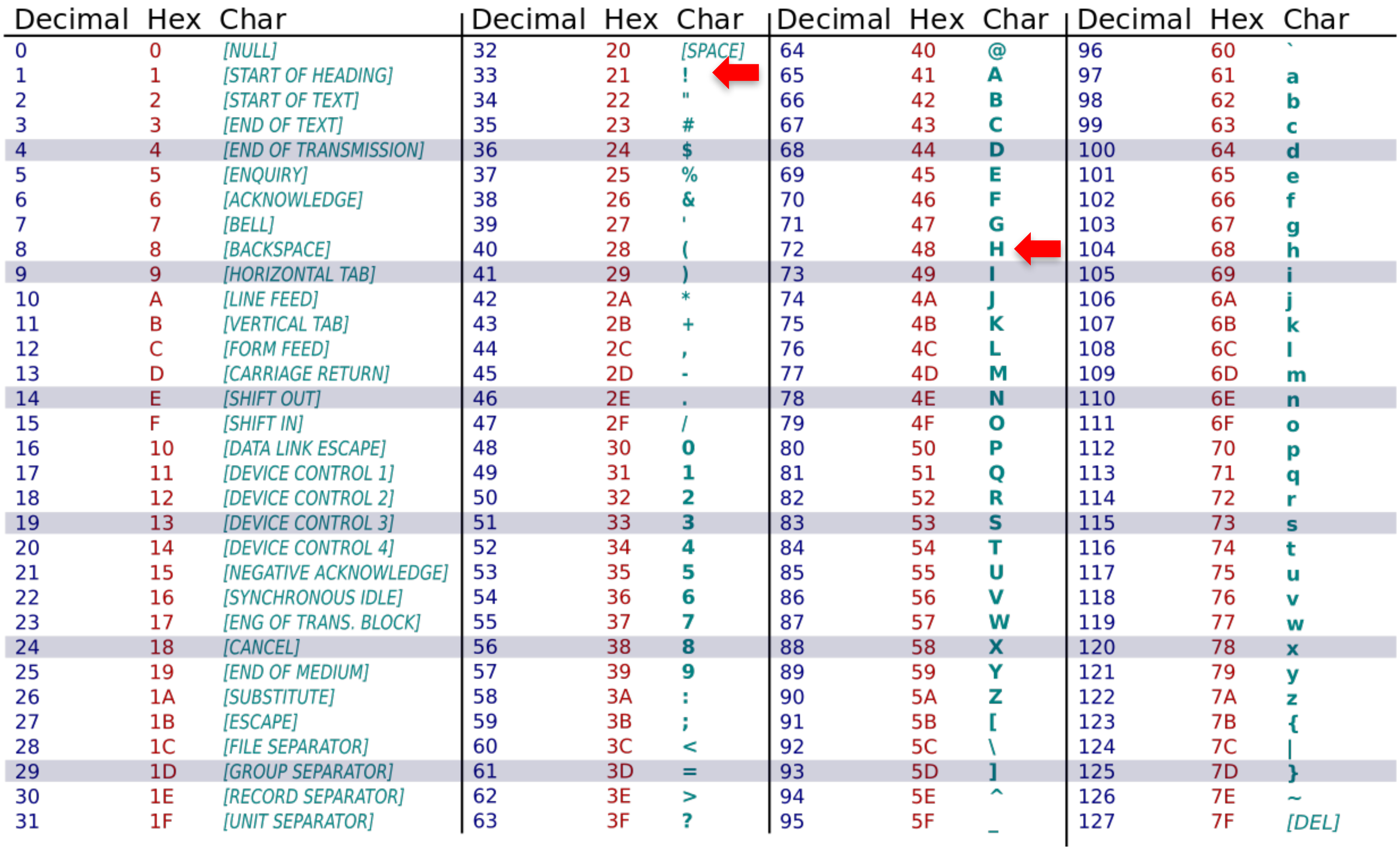
Quality + Offset

39+33 = 72

ASCII(72): H

```
@S391 ILLUMINA_FC:3:80:12/1
TAGCAGCACATCATGGTTTAC
+
IIHIDIIIIIIIIIIIIIHIH
```

ASCII table

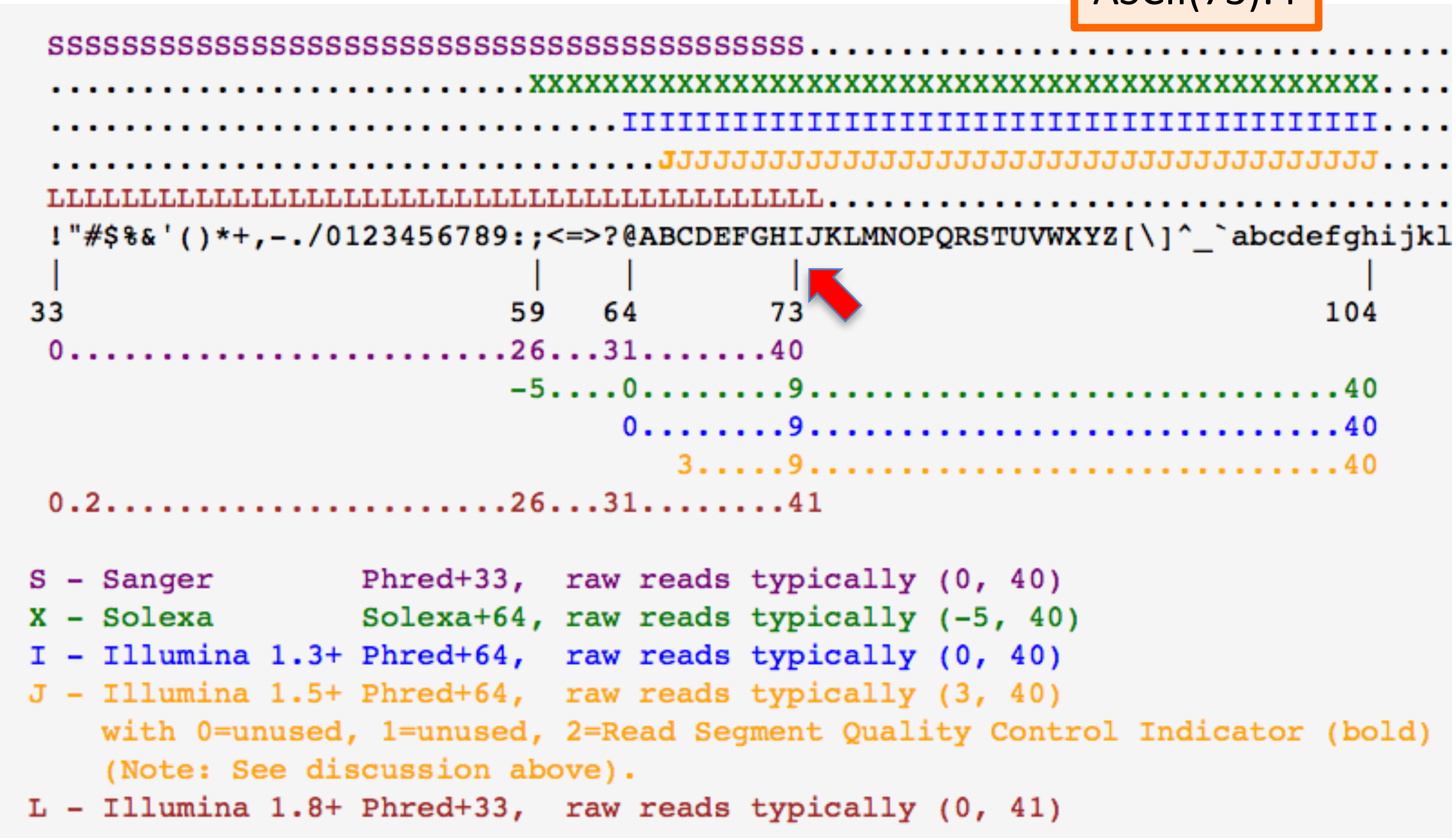


Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

Phred quality score encoding

Offset 33 - Sanger
Offset 64 - old illumina

Qual. = 40
Offset = 33
40+33 = 73
ASCII(73): I



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS...................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX....
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII....
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ....
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL..................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijkl
 |                         |    |        |                              |
33                        59   64       73                            104
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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

FASTQ quality score in Galaxy

Many old illumina datasets have a proprietary data encoding (*offset 64*)
Currently most NGS datasets use the Sanger encoding (*offset 33*)

Galaxy
By default Galaxy assign ‘***fastq***’ data type to uploaded FASTQ files.
In this case the offset is not specified, and many tools do not recognize the data

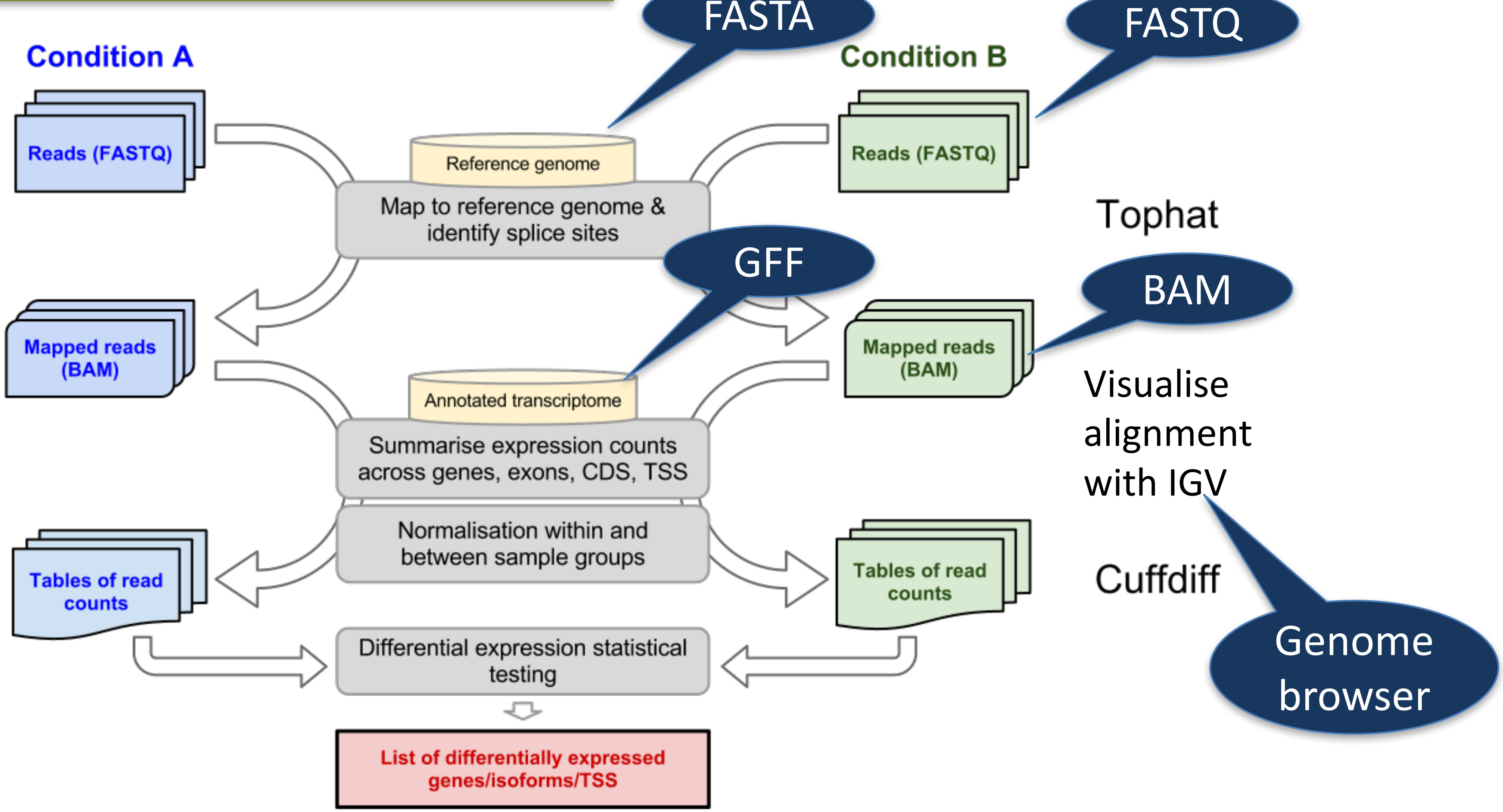
fastqillumina – old illumina quality score encoding (*offset 64*, illumina 1.3+)
fastqsanger – new illumina 1.8+ / Sanger quality score encoding
Some tools in Galaxy now work only with *fastqsanger* datatype

Solution:

- specify *fastqsanger* or *fastqillumina* datatype during upload
- change the format via Attributes > Datatype
- use **NGS: QC and manipulation** > **FASTQ Groomer** tool

Tuxedo protocol

GVL Basic RNA-Seq Galaxy tutorial
Trapnell et al. (2012) Nature Protocols



FASTA

FASTQ

Condition A

Condition B

Reads (FASTQ)

Reads (FASTQ)

Reference genome

Map to reference genome & identify splice sites

Tophat

GFF

BAM

Mapped reads (BAM)

Mapped reads (BAM)

Annotated transcriptome

Visualise alignment with IGV

Summarise expression counts across genes, exons, CDS, TSS

Normalisation within and between sample groups

Tables of read counts

Tables of read counts

Cuffdiff

Genome browser

Differential expression statistical testing

List of differentially expressed genes/isoforms/TSS

Reference genomes

Genome Reference Consortium: … a consensus representation of the genome.

FASTA format

The human reference sequence GRCh37 (hg19) contains the mitochondrial genome, 22 autosomes, chrX, chrY, 9 haplotype chromosomes, 39 unplaced contigs, and 20 unlocalized contigs.

Genomes are big. GRCh38.p10 total non-N bases: 3,080,585,178

Genomes may have many assembly versions (releases, build): mm9, mm10

Use the same assembly version for the reference sequence and gene annotations.

Order of sequences / contigs might be important for some tools.

“chr1” and “1” are not identical for some tools.

http://hgdownload.cse.ucsc.edu/gbdb/hg19/html/description.html

Gene annotations

Coordinate-based: linked to a particular genome assembly, *e.g.*, hg19

GFF (General Feature Format) format consists of **one line per feature**, each containing 9 columns of data, plus optional track definition lines.
Popular versions: GTF(=GFF2), GFF3
tab-separated fields

The first line must be a comment that identifies the version

```
##gff-version 3
ctg123  .  mRNA   1300  9000  .  +  .  ID=mrna0001;Name=sonichedgehog
ctg123  .  exon   1300  1500  .  +  .  ID=exon00001;Parent=mrna0001
ctg123  .  exon   1050  1500  .  +  .  ID=exon00002;Parent=mrna0001
ctg123  .  exon   3000  3902  .  +  .  ID=exon00003;Parent=mrna0001
ctg123  .  exon   5000  5500  .  +  .  ID=exon00004;Parent=mrna0001
ctg123  .  exon   7000  9000  .  +  .  ID=exon00005;Parent=mrna0001
```

seqid **type** **start** **end** **strand** **attributes**

source **score** **phase**

both are 1-based

'0', '1' or '2'

http://asia.ensembl.org/info/website/upload/gff3.html

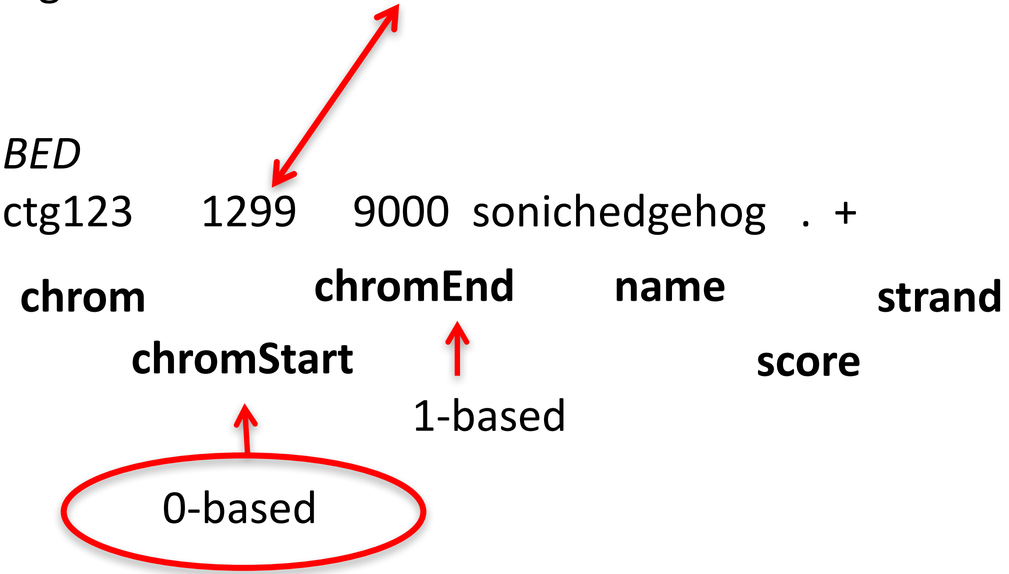
Intervals

Coordinate-based: linked to a particular genome assembly, *e.g.*, hg19

BED format, up to 12 columns of data (UCSC Table Browser), plus optional track header lines.
tab-separated fields

GFF3

```
##gff-version 3
ctg123  .  mRNA     1300    9000  .  +  .  ID=mrna0001;Name=sonichedgehog
```



BED

```
ctg123    1299    9000  sonichedgehog   .  +
```

chrom **chromStart** **chromEnd** **name** **score** **strand**

0-based

1-based

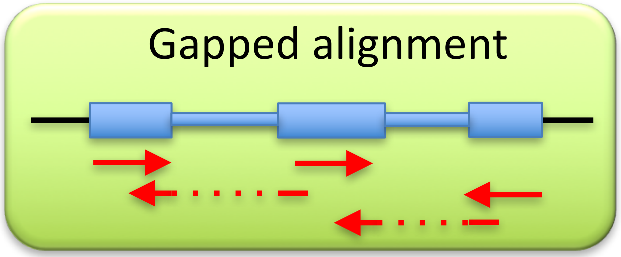
Aligners

Aligners map reads to a reference sequence.

Aligners use proprietary index files for mapping.

```
bwa index hg19.fa
```

Only for BWA   Only for hg19



Galaxy-qld provides indices for several genome assemblies (hg19, hg38, mm9, mm10 *etc.*)

Galaxy users also can use a custom reference sequence for alignment. In this situation the aligner creates indices in a temporary working directory for every job.

Contact Galaxy-qld admins if you plan to run many alignment jobs with a custom genome. We can add genome indices to the server.

Alignments: SAM and BAM

50x coverage of the human genome with read length 100 bp:
~1,500,000,000 reads
Compressed size of such alignment can be > 100 Gb.

SAM: *Sequence Alignment/Map*. Plain text format.
BAM: binary (compressed) version of the alignment format.

SAM coordinates are 1-based
BAM coordinates are 0-based

BAMs are indexed for rapid access. Useful for alignment visualization.

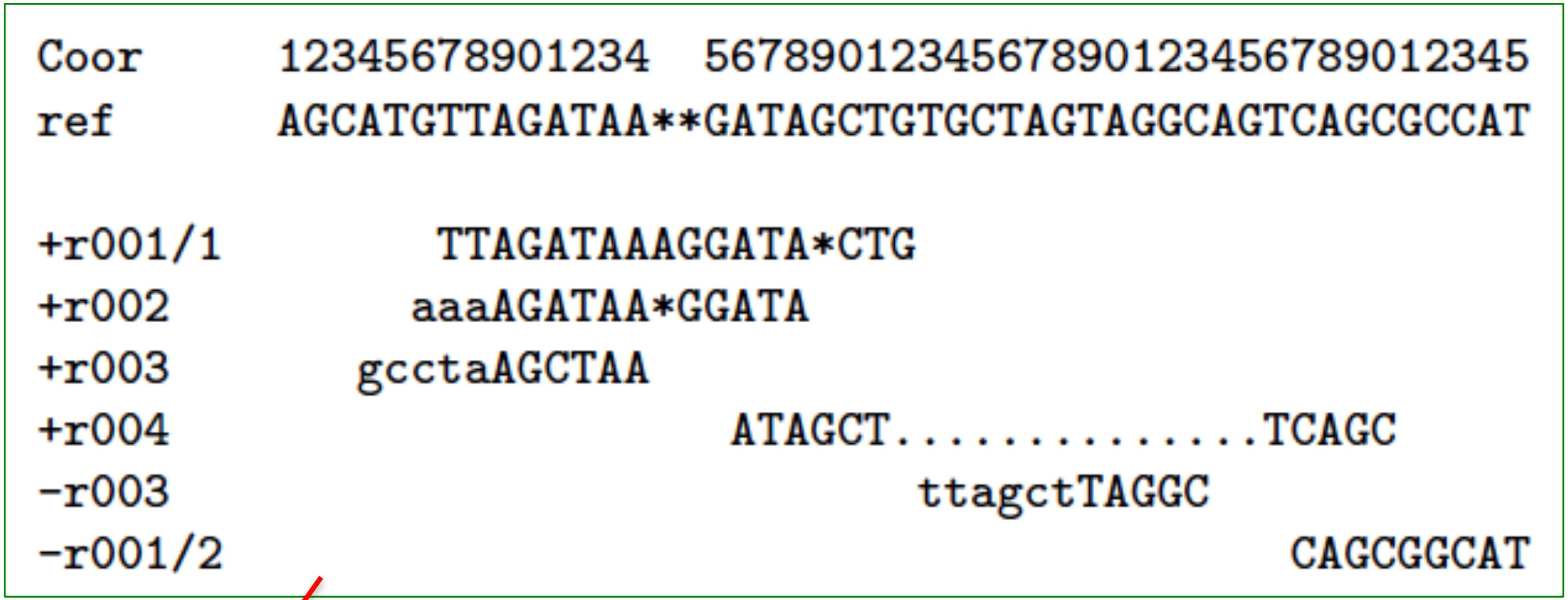
It is always good to have a header!

Read groups

Can handle multiple samples in alignment

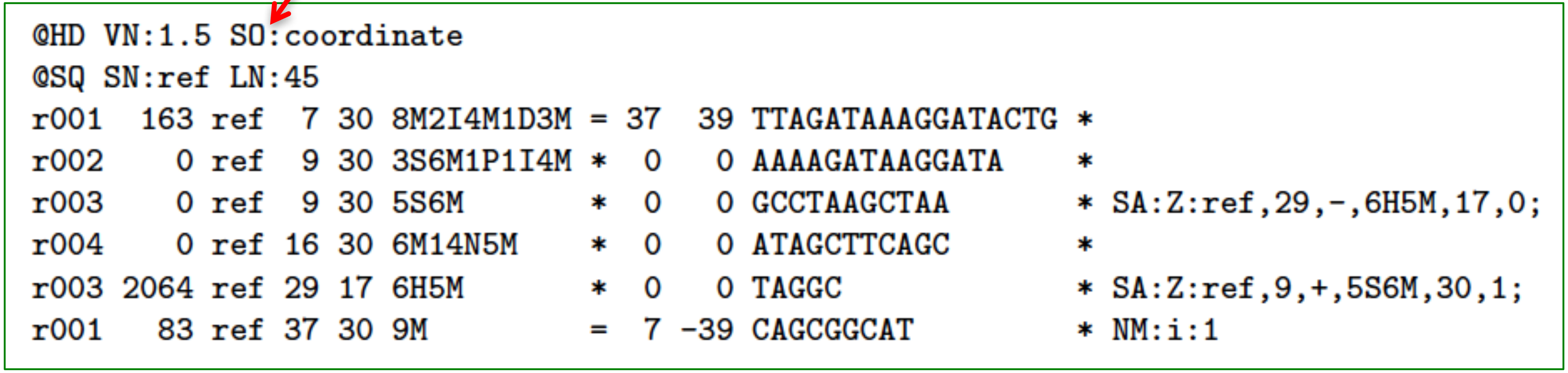
```
@HD     VN:1.0   SO:queryname
@RG     ID:igGroup    SM:igSmpl    LB:igL1   PL:ILLUMINA
@SQ     SN:chr2L      LN:23011544
@PG     ID:TopHat     VN:2.0.14
    CL:/mnt/galaxy/tools/tophat/2.0.14/iuc/package_tophat_2_0_14/536f7b
b5616d/bin/tophat --num-threads 5 ....
```

SAM format



```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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



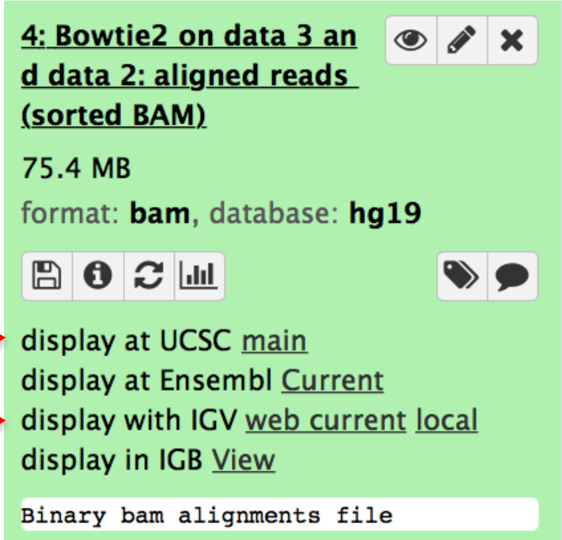
```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC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

11 mandatory columns and optional fields with the TAG:TYPE:VALUE format

Visualization of BAMs

Galaxy servers can act as a track hub

Galaxy ⇄ igv



4: Bowtie2 on data 3 and data 2: aligned reads (sorted BAM)

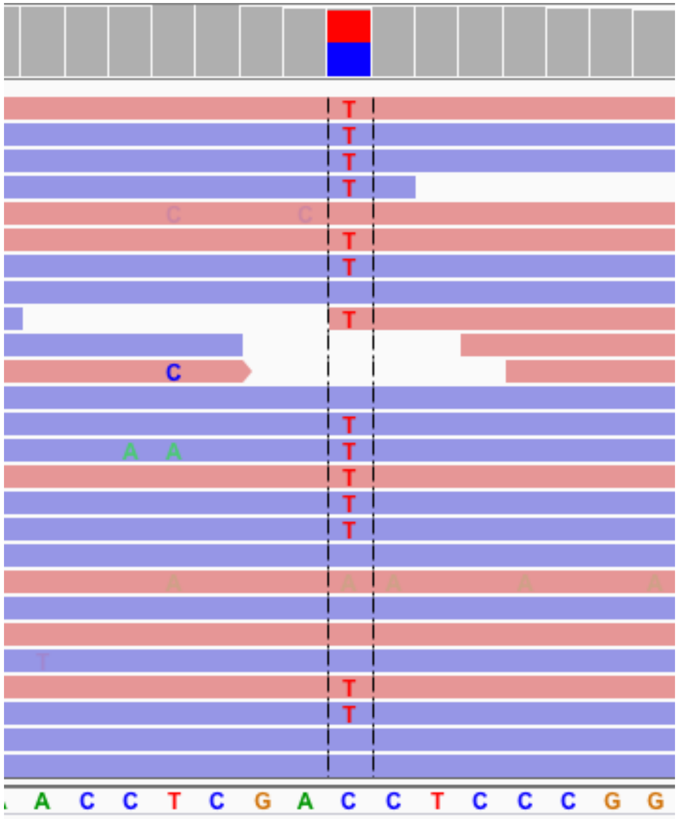
75.4 MB

format: **bam**, database: **hg19**

display at UCSC main
display at Ensembl Current
display with IGV web current local
display in IGB View

```
Binary bam alignments file
```

Alignment on IGV

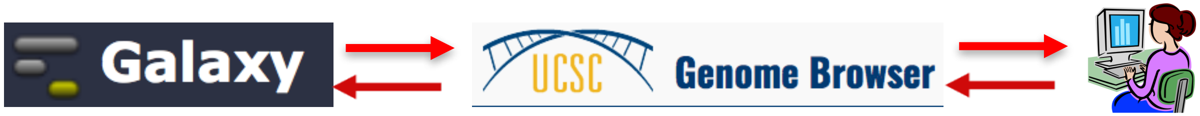


It is possible to add multiple tracks:
BAMs, gene annotations, known variants…

Genome browsers

Integrative Genomics Viewer, IGV
Efficient genome viewer developed by the Broad Institute.
Installable on personal computers.
Can add a custom genome.

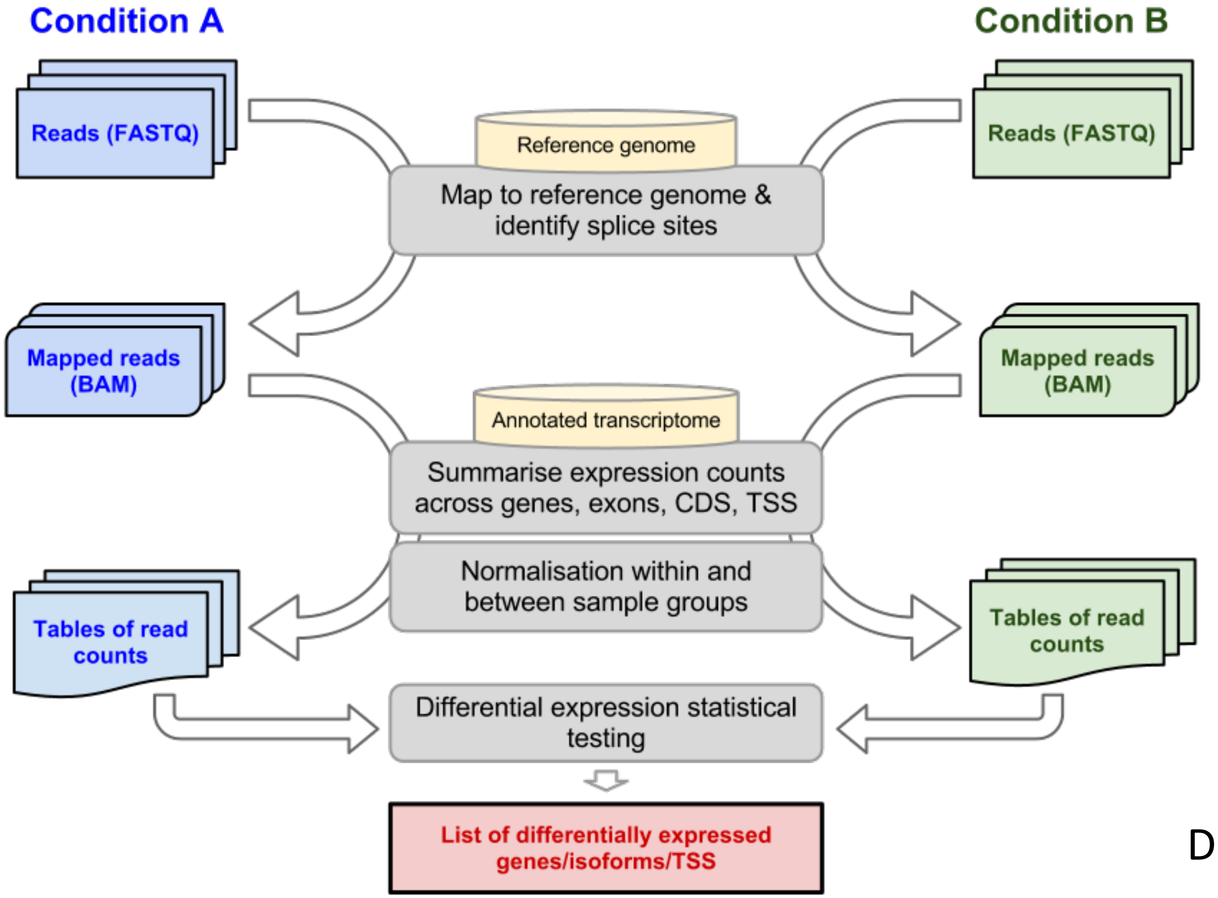
UCSC Genome Browser
A big server in the US.
Table Browser for data analysis (intersection)
Support data export to Galaxy
Custom sessions (can save your tracks)
liftOver tool
Public track hubs



RNA-Seq with the Cufflinks package

GVL Basic RNA-Seq Galaxy tutorial
Trapnell et al. (2012) Nature Protocols

D. melanogaster
Two conditions
Three replicates
Data for chr4


Condition A
Condition B
Reads (FASTQ)
Reads (FASTQ)
Reference genome
Map to reference genome & identify splice sites
Mapped reads (BAM)
Mapped reads (BAM)
Annotated transcriptome
Summarise expression counts across genes, exons, CDS, TSS
Normalisation within and between sample groups
Tables of read counts
Tables of read counts
Differential expression statistical testing
List of differentially expressed genes/isoforms/TSS

Tophat

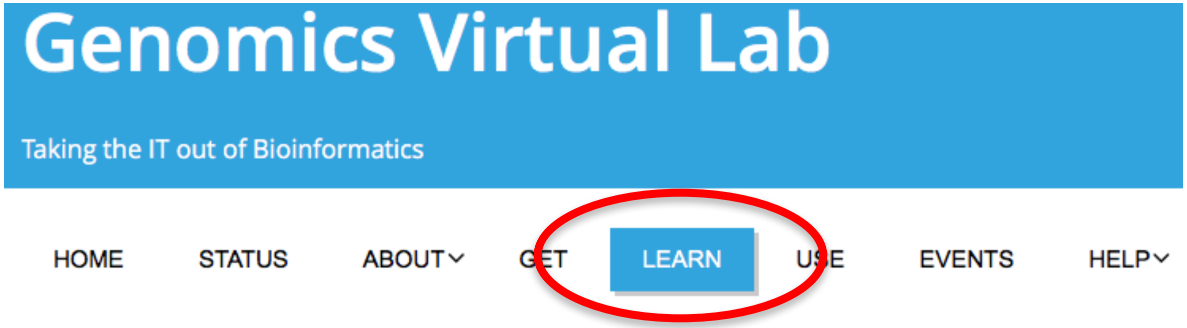
Visualise alignments

Cuffdiff

Data manipulation

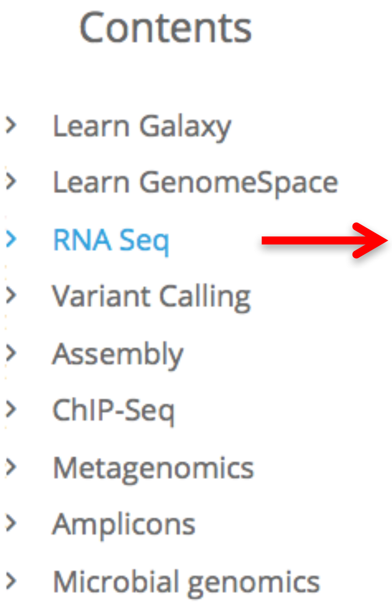
Setup for the workshop

GVL website:
gvl.org.au



Genomics Virtual Lab

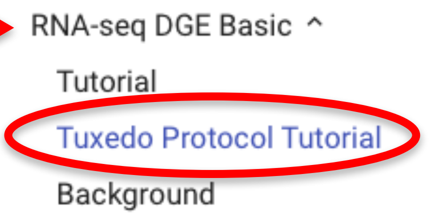
Taking the IT out of Bioinformatics



Contents

- Learn Galaxy
- Learn GenomeSpace
- RNA Seq
- Variant Calling
- Assembly
- ChIP-Seq
- Metagenomics
- Amplicons
- Microbial genomics

Basic Galaxy tutorial

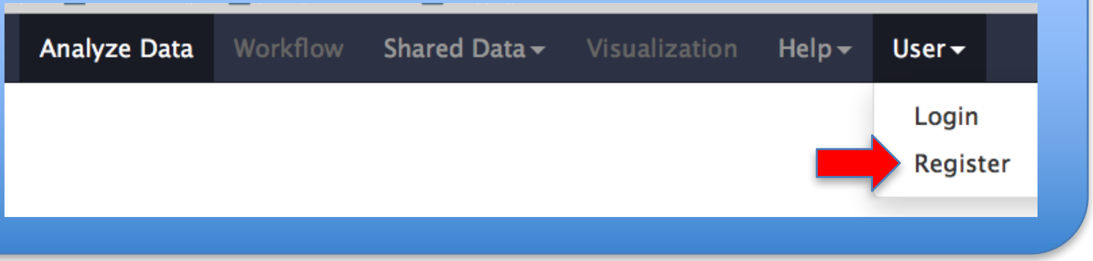


RNA-seq DGE Basic

- Tutorial
- Tuxedo Protocol Tutorial
- Background

3

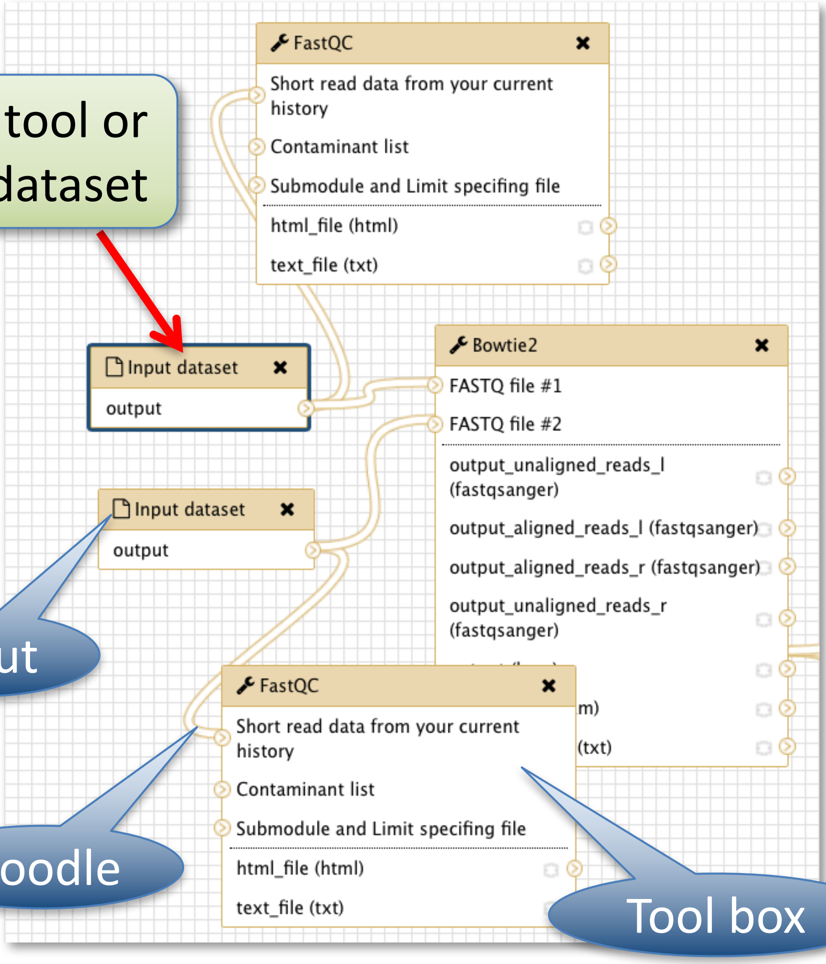
Register on Galaxy-tut: galaxy-tut.genome.edu.au



Galaxy is a workflow engine

A Galaxy workflow is a series of tools and dataset actions that run in sequence as a batch operation

Select tool or input dataset



FastQC ✖

Short read data from your current history

Contaminant list

Submodule and Limit specifing file

html_file (html)

text_file (txt)

Input dataset ✖

output

Bowtie2 ✖

FASTQ file #1

FASTQ file #2

output_unaligned_reads_l (fastqsanger)

output_aligned_reads_l (fastqsanger)

output_aligned_reads_r (fastqsanger)

output_unaligned_reads_r (fastqsanger)

m)

(txt)

Input dataset ✖

output

FastQC ✖

Short read data from your current history

Contaminant list

Submodule and Limit specifing file

html_file (html)

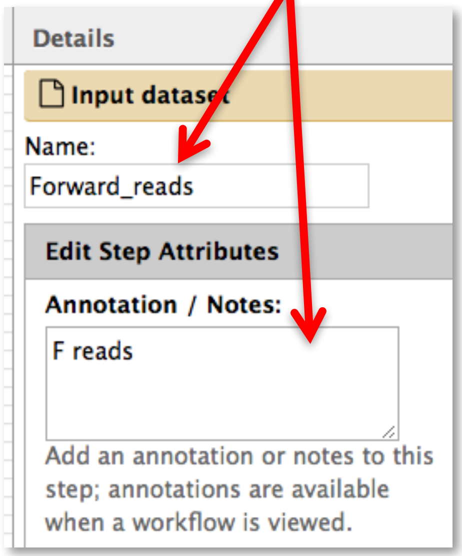
text_file (txt)

Input

Noodle

Tool box

Add name, comments



Details

Input dataset

Name:

Forward_reads

Edit Step Attributes

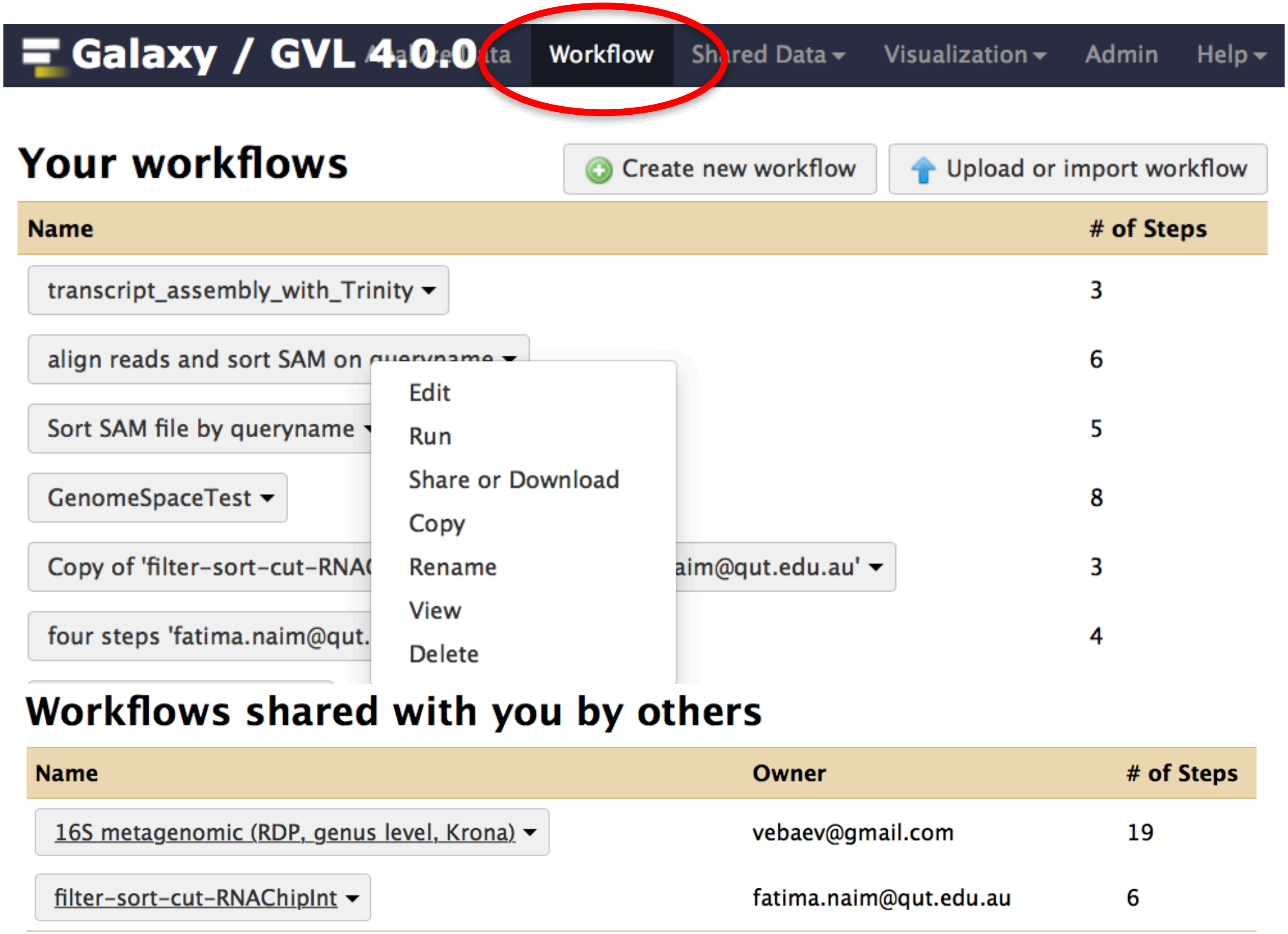
Annotation / Notes:

F reads

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

Email notification

Galaxy workflow



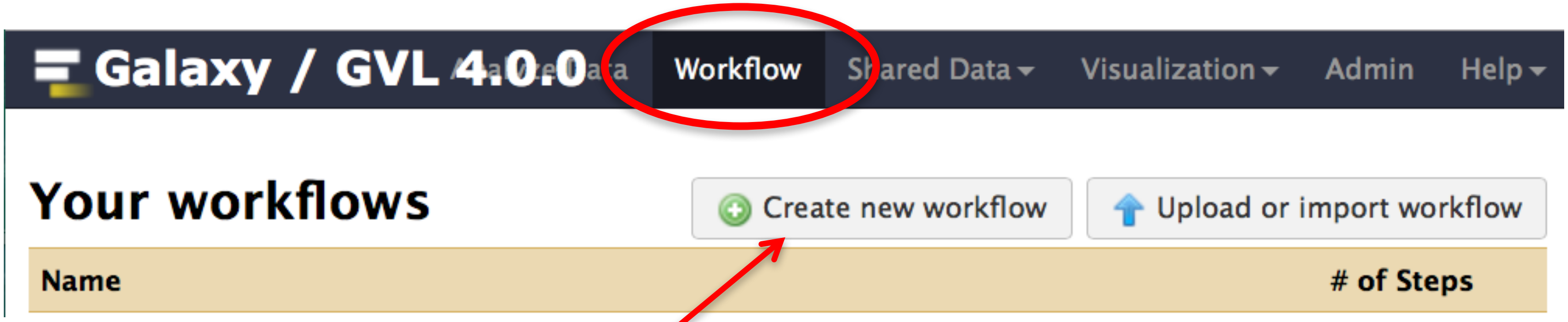
Your workflows

Name	# of Steps
transcript_assembly_with_Trinity	3
align reads and sort SAM on queryname	6
Sort SAM file by queryname	5
GenomeSpaceTest	8
Copy of 'filter-sort-cut-RNA aim@qut.edu.au'	3
four steps 'fatima.naim@qut.	4

Workflows shared with you by others

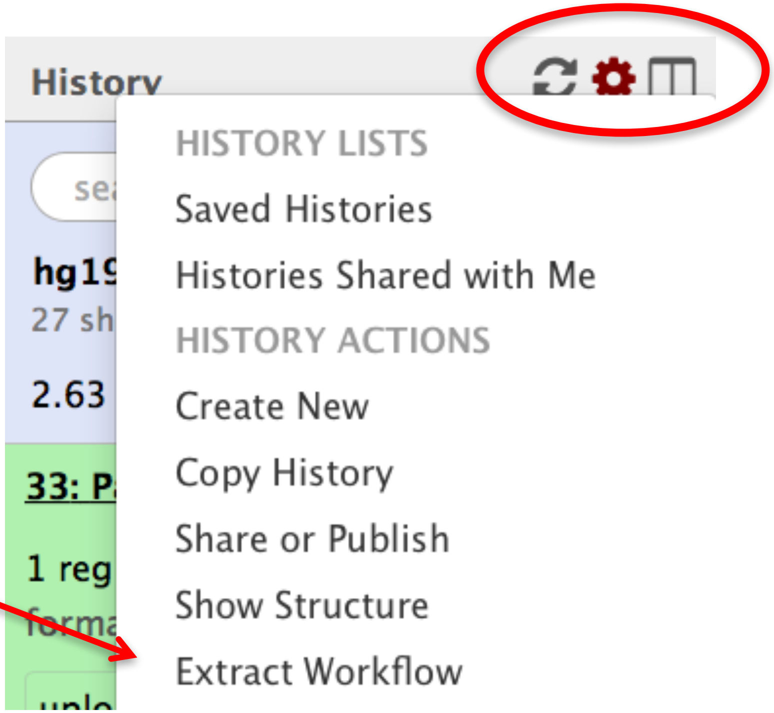
Name	Owner	# of Steps
16S metagenomic (RDP, genus level, Krona)	vebaev@gmail.com	19
filter-sort-cut-RNAChipInt	fatima.naim@qut.edu.au	6

Create a Galaxy workflow



From scratch

From history



Exercise

We will create a Galaxy workflow for RNA-Seq analysis without replicates: *tophat2 > Cuffdiff > Filter*

Acknowledgments

Genomics Virtual Lab: gvl.org.au
Galaxy for tutorials: galaxy-tut.genome.edu.au
Galaxy Australia: galaxy-aust.genome.edu.au

Contributors and participants:

RDS Research Data Services

qcif

nectar

QRIS cloud
QRISnews for Leaders in eResearch

EMBL Australia
Bioinformatics Resource

The University Of Queensland Australia

RCC Research Computing Centre

QFAB Bioinformatics

genomics virtual lab

Melbourne Bioinformatics
BIOINFORMATICS + DATA SERVICES + INFRASTRUCTURE, FOR LIFE SCIENCES TODAY