

Analysis of high-throughput sequencing data using Galaxy platform

Igor Makunin
UQ RCC

Centre for Digital Scholarship, the UQ library; May 9, 2018

High-throughput sequencing, or NGS

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



Knowledge and skills

Galaxy: how does it look like

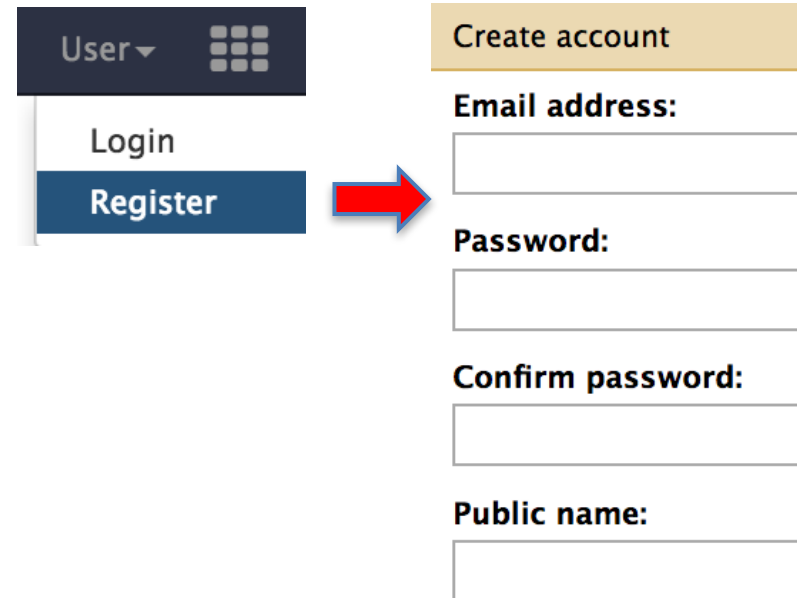
Galaxy is a web-based platform for analysis of genome-scale data

The screenshot displays the Galaxy web interface with the following components and annotations:

- Top menu:** A horizontal bar at the top containing "Galaxy / GVL 4.0.0", "Analyze Data", "Workflow", "Shared Data", "Visualization", "User", and a "Using 10%" indicator.
- Tools:** A vertical sidebar on the left lists various tool categories such as "Statistics", "Graph/Display Data", "NGS COMMON TOOLSETS", "FASTA manipulation", "NGS: Picard", "NGS: SAM Tools", "BED tools", "BEDtools2", "NGS: VCF Manipulation", "NGS: GATK Tools 1.4", "NGS: GATK Tools 2.8", "EMBOSS", "NGS ANALYSIS", "NGS: QC and manipulation", "NGS: Mapping", "NGS: Assembly", and "NGS: RNA Analysis".
- Upload:** An annotation with a red arrow pointing to the upload icon in the top left of the tool configuration area.
- Working window:** The central area showing the configuration for the "Tophat Gapped-read mapper for RNA-seq data (Galaxy Tool Version 0.9)". It includes fields for "Is this single-end or paired-end data?" (set to "Single-end"), "RNA-Seq FASTQ file" (set to "1: C1_R1.chr4.fq"), "Use a built in reference genome or own from your history" (set to "Use a built-in genome"), "Select a reference genome" (set to "D. melanogaster Apr. 2006 (BDGP R5/dm3) (dm3)"), "TopHat settings to use" (set to "Use Defaults"), and "Specify read group?" (set to "No"). A red circle highlights the "Execute" button at the bottom of this window.
- History:** A vertical sidebar on the right showing a list of previous analyses, including "11: Tophat on data 1: splice junctions", "10: Tophat on data 1: deletions", "9: Tophat on data 1: srtions", "8: Tophat on data 1: a lign_summary", "7: ensembl_dm3.chr4. gtf", "6: C2_R3.chr4.fq", and "5: C2_R2.chr4.fq". A red arrow points to the "History" menu icon in the top right of this sidebar.

Why Galaxy?

- Simple intuitive platform
- Public servers with pre-installed tools and storage
- Built-in public data, *eg* aligner indices
- Direct import from public repositories
- 1000s tools are available
- Data visualisation options
- Data sharing
- Big community
- Easy registration



The diagram illustrates the user registration process. On the left, a dark blue header bar contains a 'User' dropdown menu and a grid icon. Below this, a white box contains 'Login' and a dark blue box contains 'Register'. A red arrow points from the 'Register' button to a registration form on the right. The form has a yellow header bar with 'Create account'. It includes input fields for 'Email address:', 'Password:', 'Confirm password:', and 'Public name:'.

User ▼

Login

Register

Create account

Email address:

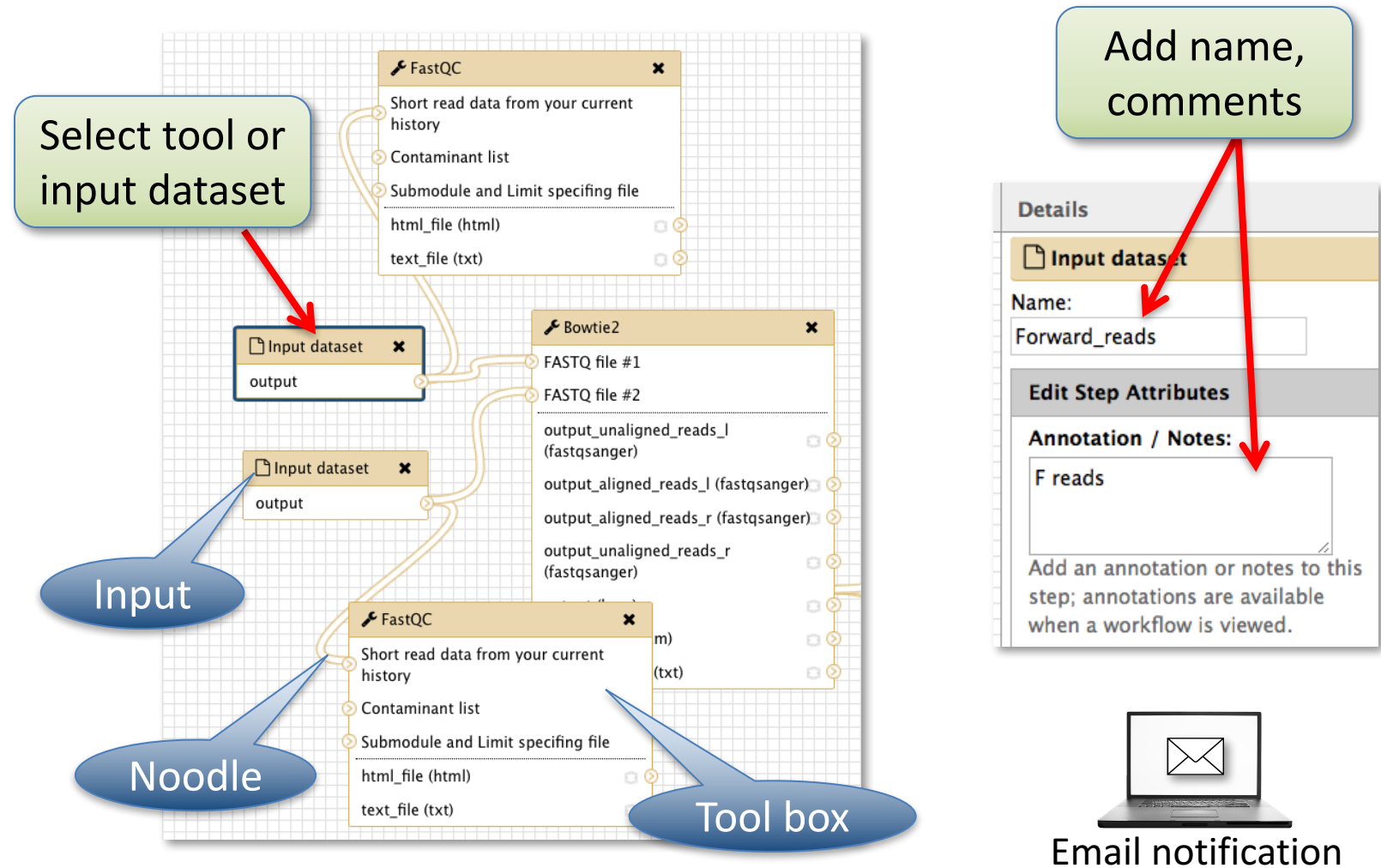
Password:

Confirm password:

Public name:

Galaxy is a workflow engine

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




Galaxy tool shed

New tools can be installed by Galaxy admins from Galaxy tool sheds.

The main tool shed: toolshed.g2.bx.psu.edu

Test tool shed: testtoolshed.g2.bx.psu.edu

 **Galaxy Tool Shed**

Repositories Groups Help▼ User▼

5257 valid tools on Oct 18, 2017

Search

- [Search for valid tools](#)
- [Search for workflows](#)

Valid Galaxy Utilities

- [Tools](#)
- [Custom datatypes](#)
- [Repository dependency definitions](#)
- [Tool dependency definitions](#)

Repositories by Category

<u>Name</u>	<u>Description</u>
Assembly	Tools for working with assemblies
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.
Combinatorial Selections	Tools for combinatorial selection

Public Galaxy servers

Galaxy servers:

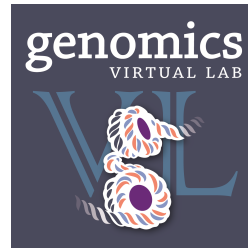
usegalaxy.org

usegalaxy.eu

galaxy-tut.genome.edu.au

galaxy-qld.genome.edu.au

(Galaxy Australia)



- *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*

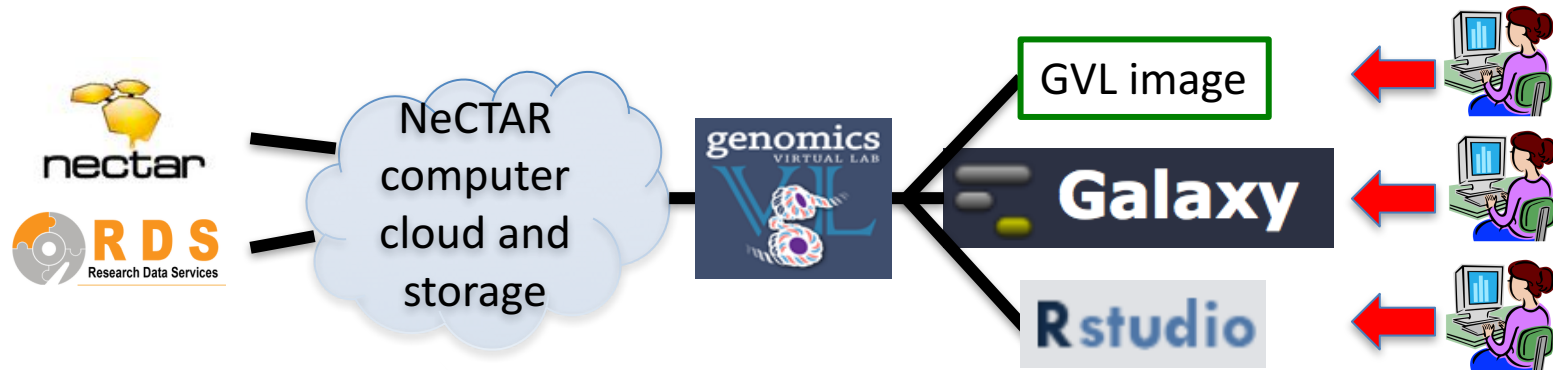
Genomics Virtual Laboratory

The GVL project was started in 2012

Analysis of nextGen sequencing data is a bottleneck (*infrastructure, skills*)

Genomics Virtual Lab: take the IT out of Bioinformatics

- DIY bioinformatics environment (*advanced users*)
- web-based resources (*biologists-friendly*)
- tutorials and training materials: gvl.org.au



GVL advantages:

- ★ - public resource (*no charges to users*)
- ★ - available immediately to anyone

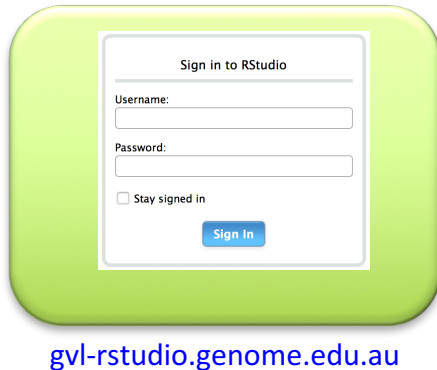
Afgan *et al.* Genomics Virtual Laboratory: a practical bioinformatics workbench for the cloud. PLoS One. 2015 Oct 26;10(10):e0140829. doi: 10.1371/journal.pone.0140829

GVL activities in Brisbane

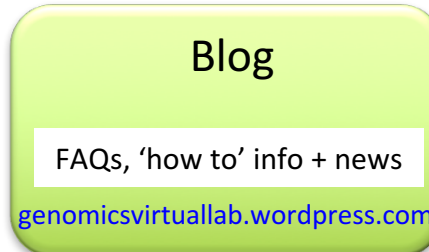
Services



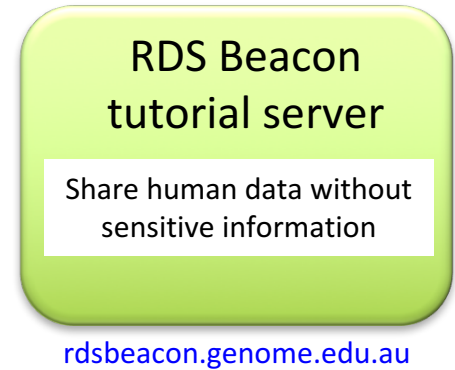
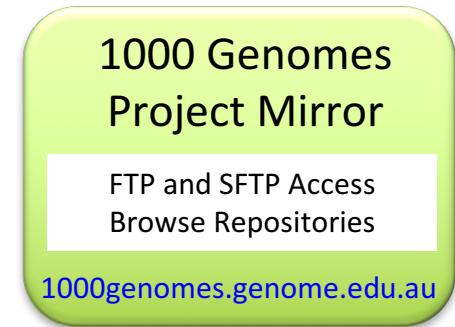
GVL RStudio server



User engagement




Data





Sponsors:



Galaxy Australia

 Master node 16 CPUs, 64 GB RAM

 Worker nodes:
16 CPUs, 64 GB RAM

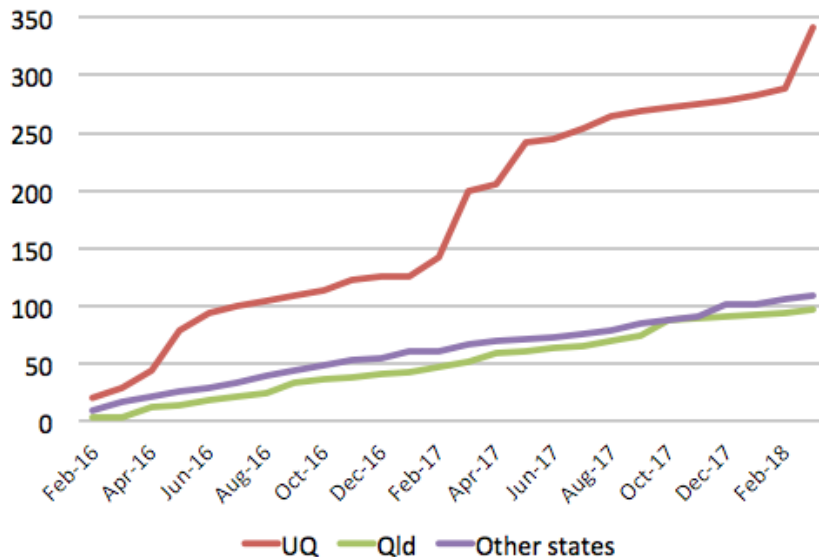
 49 Tb Volume storage (user data)
1 Tb Volume storage for indices

galaxy-qld.genome.edu.au

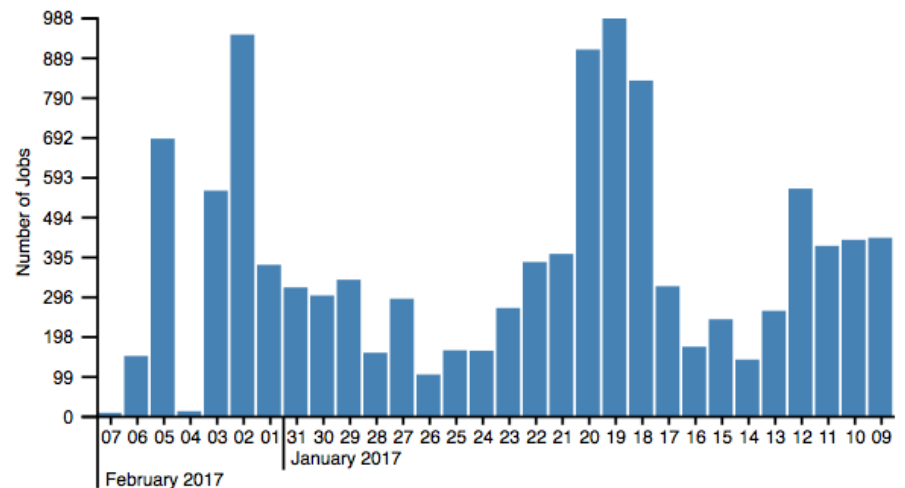


QRISnews for Leaders in eResearch

Galaxy-qld: Australian users



Jobs per day



Less jobs on weekends

Tools

Genomics Virtual Lab

Taking the IT out of Bioinformatics

HOME

STATUS

ABOUT▼

GET

LEARN

USE

EVENTS

HELP▼

GVL Galaxy in Queensland:

galaxy-qld.genome.edu.au/galaxy

Tools:

- BWA, bowtie2
- Velvet, SPAdes
- Trinity
- tophat2, RNA_STAR, HiSAT2
- DESeq, edgeR, Cufflinks, StringTie
- GATK2, variant detection tools
- Metagenomics tools
- MACS2, SPP
- SAMtools
- Picard
- deepTools

Topics:

- ✓ RNA-Seq
- ✓ ChIP-Seq
- ✓ Variant detection
- ✓ Genome assembly
- ✓ Transcriptome
- ✓ Metagenomics

User data and quotas

- Registered users: 100 Gb
 - Australian users: 600 Gb
 - UQ users: 1 Tb
-
- No external backup for user data
 - Download results as soon as convenient
 - Delete and purge unneeded datasets and temporary files

We do not endorse:

- a long term data storage on the server
- multiple registrations



Public data on Galaxy

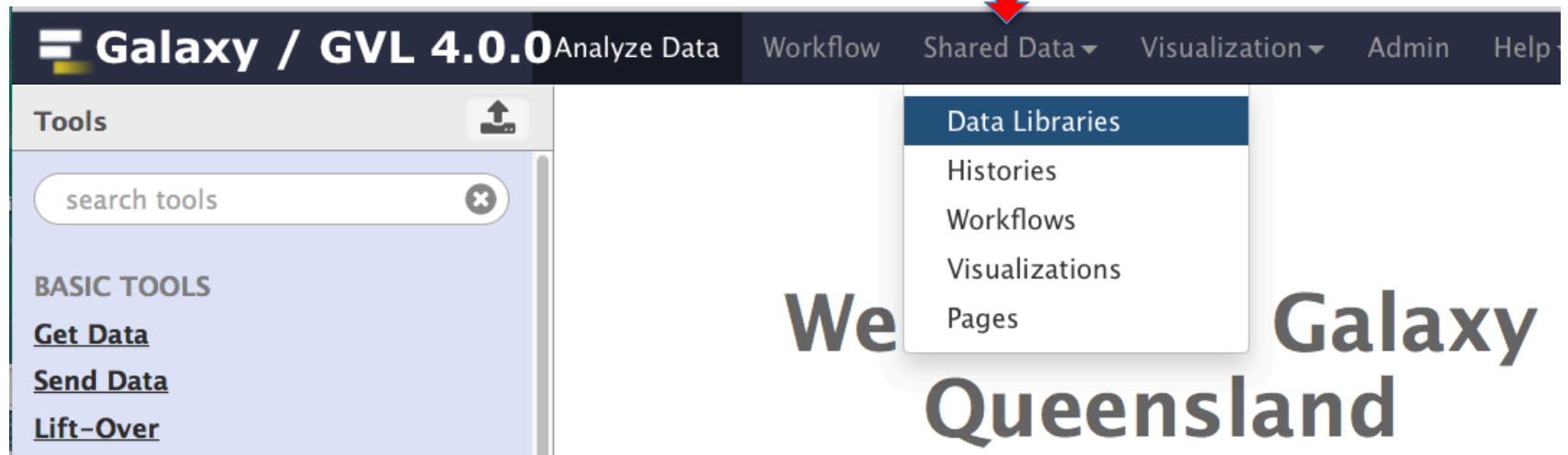
Genome indices
and assemblies

BLAST database

SnEff databases

Data libraries

Files imported from Data Libraries are not counted towards user quota.
We add public data on demand from users.



The screenshot shows the Galaxy GVL 4.0.0 interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', and 'Help'. A red arrow points to the 'Shared Data' dropdown menu, which is open and displays a list of options: 'Data Libraries', 'Histories', 'Workflows', 'Visualizations', and 'Pages'. The 'Data Libraries' option is highlighted in blue. On the left side, there is a 'Tools' panel with a search bar and a list of basic tools including 'Get Data', 'Send Data', and 'Lift-Over'.

We Galaxy
Queensland

Support for Galaxy-qld



Igor Makunin
UQ RCC

User support,
training



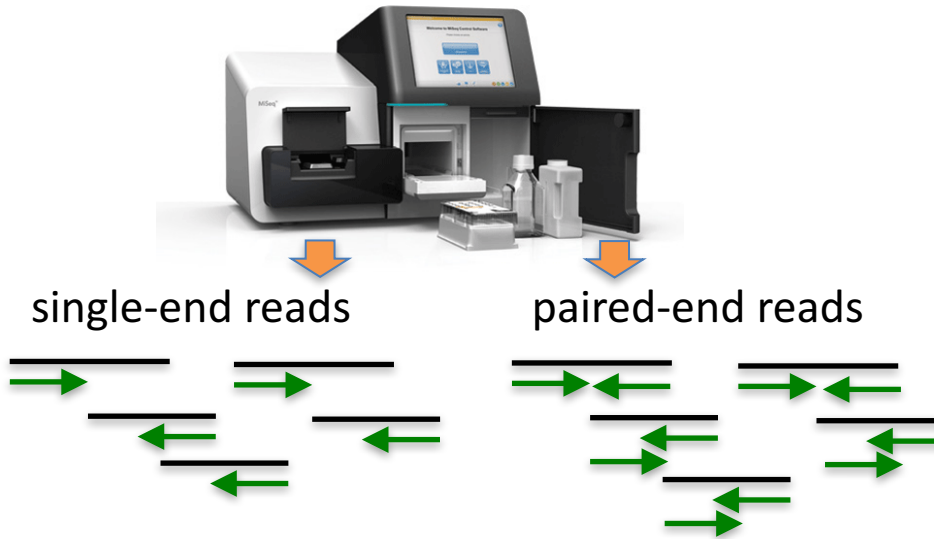
Derek Benson
UQ RCC & IMB

System
administrator

GVL-Qld announcements: twitter.com/GVL_QLD
GVL-Qld blog: genomicsvirtuallab.wordpress.com

GVL FAQ page at gvl.org.au/faq
genomicsvirtuallab.wordpress.com/getting-started

FASTQ format



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

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

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

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

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.

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

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
+
IIHIDIIIIIIIIIIIIIIIHIIH
```


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

```

SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX...
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII...
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ...
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNPOQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzkl
|                                     |               |                               |
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)

```

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

Acknowledgments and useful links

Genomics Virtual Lab: gvl.org.au

Galaxy for tutorials: galaxy-tut.genome.edu.au

Galaxy Australia: galaxy-qld.genome.edu.au

Contributors and participants:



QRISnews for Leaders in eResearch



Bioinformatics Resource



BIOINFORMATICS



BIOINFORMATICS • DATA SERVICES • INFRASTRUCTURE, FOR LIFE SCIENCES TODAY

Galaxy demo: RNA-Seq analysis

Import from a data library

Mapping RNA-Seq reads to a reference genome using tophat2 aligner

Alignment visualisation with Integrative Genomics Viewer

Identification of differentially expressed genes using Cuffdiff

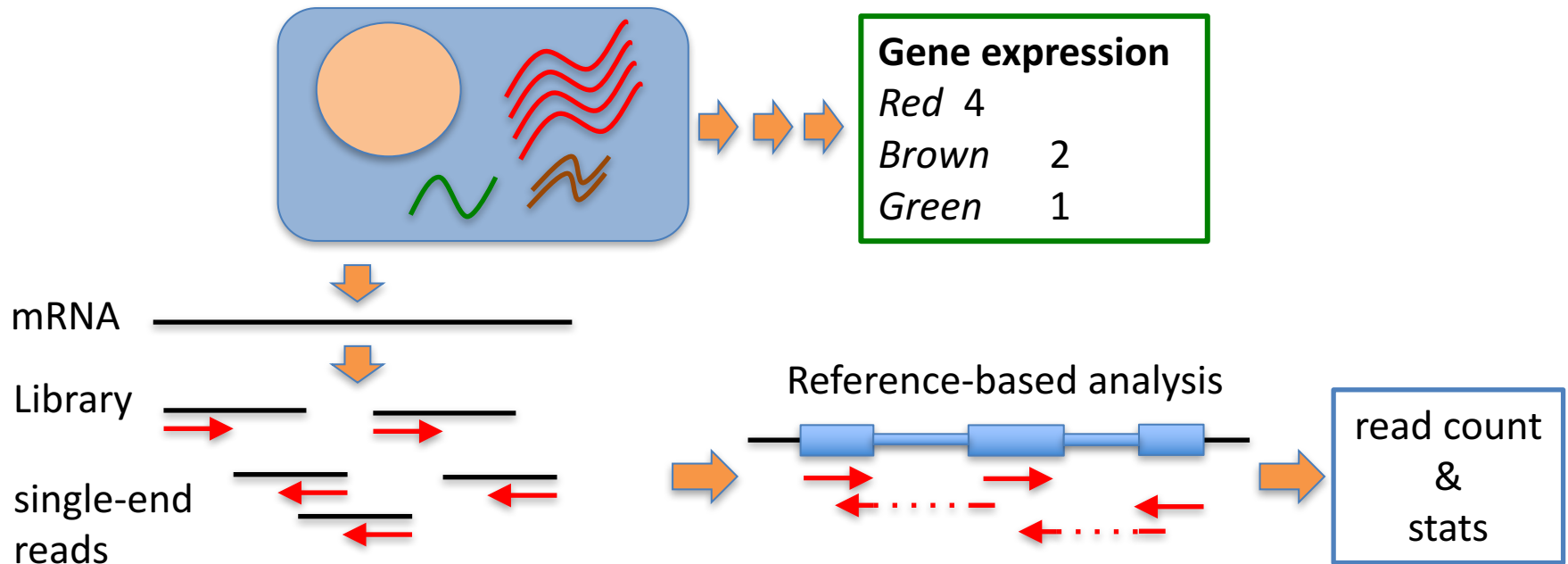
Data filtering

Galaxy workflow

Differential gene expression analysis

NextGen sequencing data can be used for analysis of gene expression on a genome scale.

Assumption: number of reads mapped to a gene correlates with the transcript abundance.



RNA-Seq with the Cufflinks package

