Introduction to the analysis of next generation sequencing data

Mik Black, University of Otago
Cristin Print, The University of Auckland

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 New Zealand License. To view a copy of this license, visit
http://creativecommons.org/licenses/by-nc-sa/3.0/nz/
or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.



Overview

- Introduction:
  - Summary of NGS technologies
  - Data formats and quality assessment
- Alignment
  - Tools and formats
  - Visualization
- RNA-seq analysis

CC BY NC SA

Technology – NextGenSeq

- Over the past few years, the "Next Generation" of sequencing technologies has emerged.
- Three major players:
  - Roche: GS-FLX, GS-Junior
  - Applied Biosystems: SOLiD 4
  - Illumina: Genome Analyzer / HiSeq2000

CC BY NC SA

Roche GS-FLX (454)

- 10 hours per run
- 1.3 million reads per run
- Read length of ~400bp
- Generates ~500 Mb per run



CC BY NC SA

Roche GS-Junior

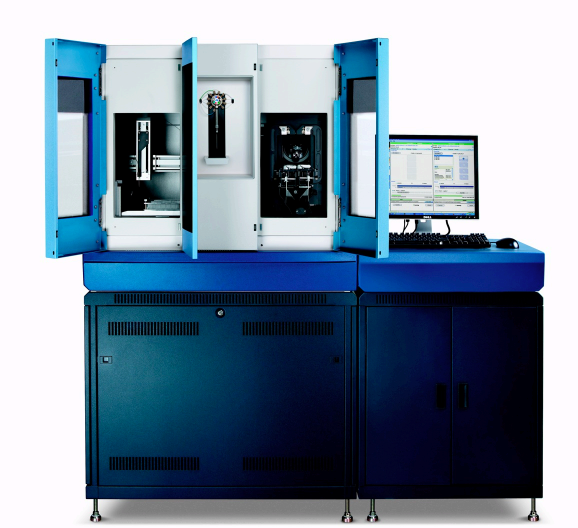


- 10 hours per run
- 100,000 reads per run
- Read length of ~400bp
- Generates ~ 35 Mb per run (filtered)

cc BY NC SA

Applied Biosystems SOLiD 4

- 4-16 days per run (35bp SE vs 50bp MP)
- 1.4 billion reads/run
- Read length of 50bp (x2)
- Generates ~ 100Gbp per run

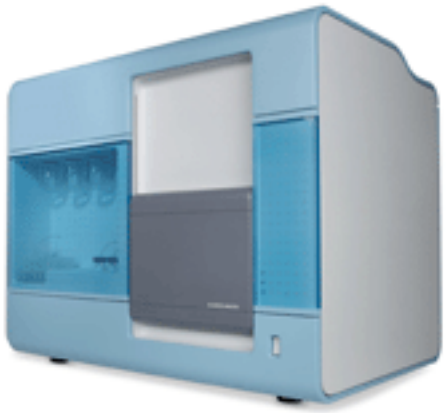


Library Type	Read Length	Days/Run	Total Tags/Run	Mappable Data
Mate-Paired	2 x 35 bp	8–9	>1.4 B	50–70 GB
	2 x 50 bp	12–16	>1.4 B	80–100 GB
Paired-End	50 bp x 25 bp	11–13	>1.4 B	55–70 GB
Fragment	1 x 35 bp	3.5–4.5	>700 M	25–35 GB
	1 x 50 bp	6–8	>700 M	40–50 GB

cc BY NC SA

Illumina Genome Analyzer IIx

- 5 days per run
- 250 million reads/run
- Read length of 100bp (x2) (variable)
- Generates ~ 25 Gb per run
- Accuracy rate: >98.5%



CC BY NC SA

Illumina HiSeq2000

- 8 days per run
- 1 billion reads/run
- Read length of 100bp (x2)
- Generates ~ 200 Gb per run

Read Length	Run Time	Output
1 × 35 bp	~1.5 days	26–35 Gb
2 × 50 bp	~4 days	75–100 Gb
2 × 100 bp	~8 days	150–200 Gb

*Sequencing output generated with a PhiX library and cluster densities between 260,000–347,000 clusters/mm^2 that pass filtering on a HiSeq 2000.

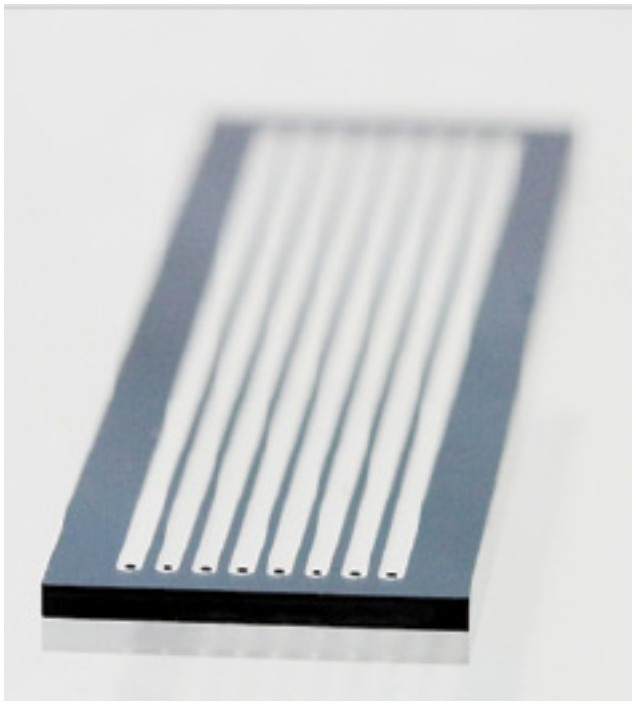
Throughput
Up to 25 Gb per day for a 2 × 100 bp run.

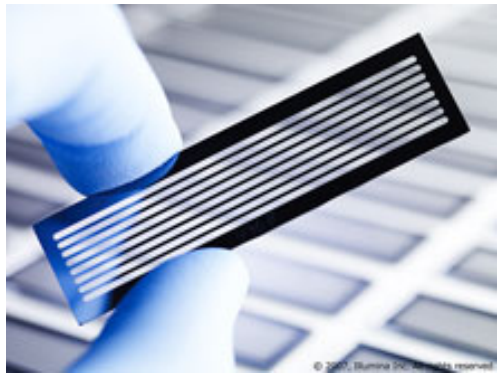


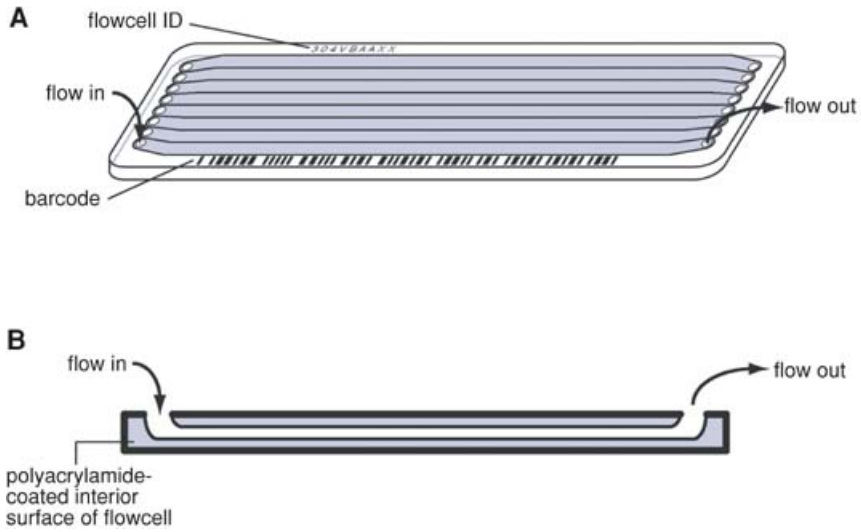
cc BY NC SA

Illumina flow cell

www.illumina.com





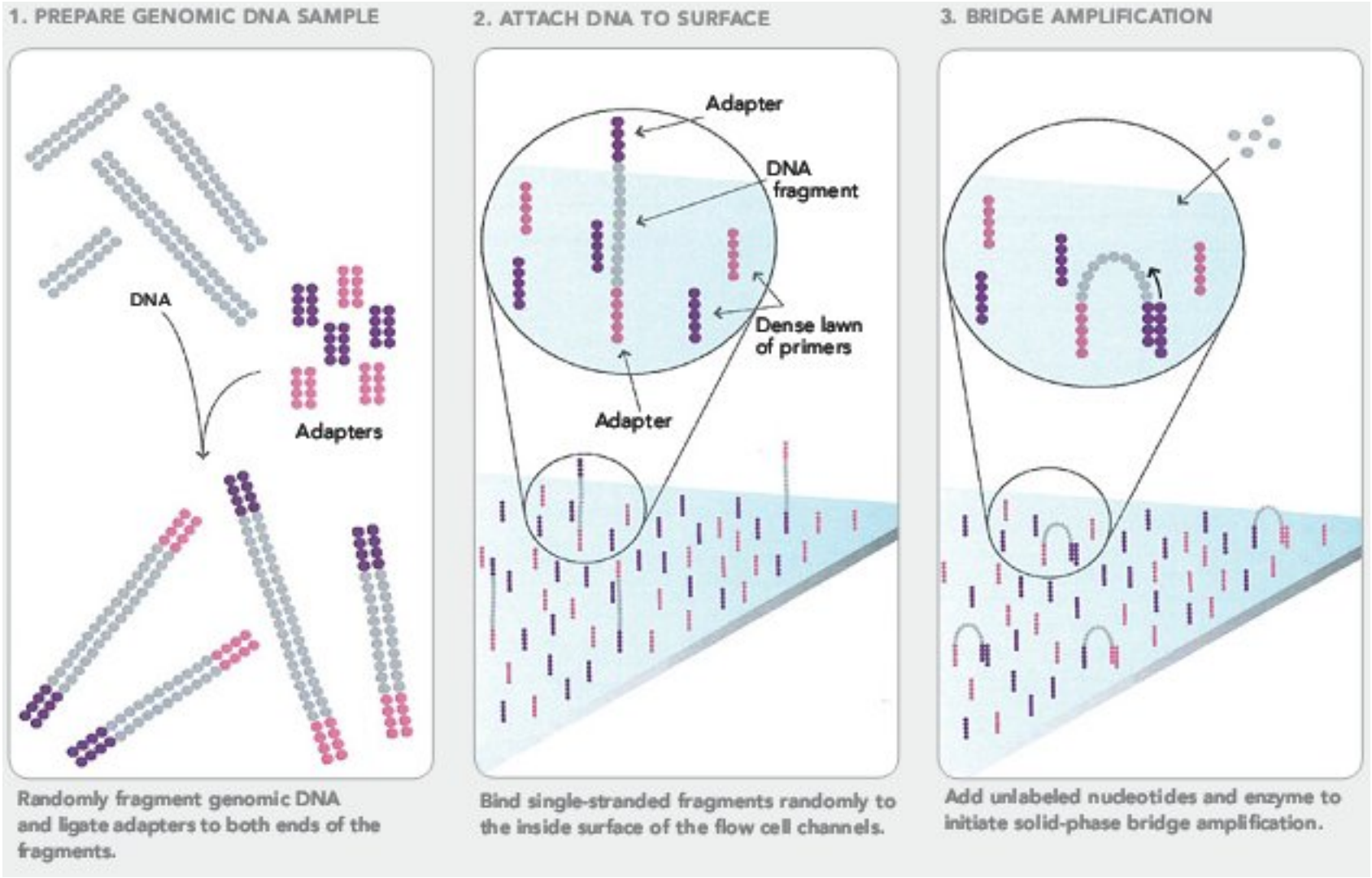


Quail et al. Current protocols in human genetics (2009) vol. Chapter 18 pp. Unit 18.2

CC BY NC SA

Illumina sequencing

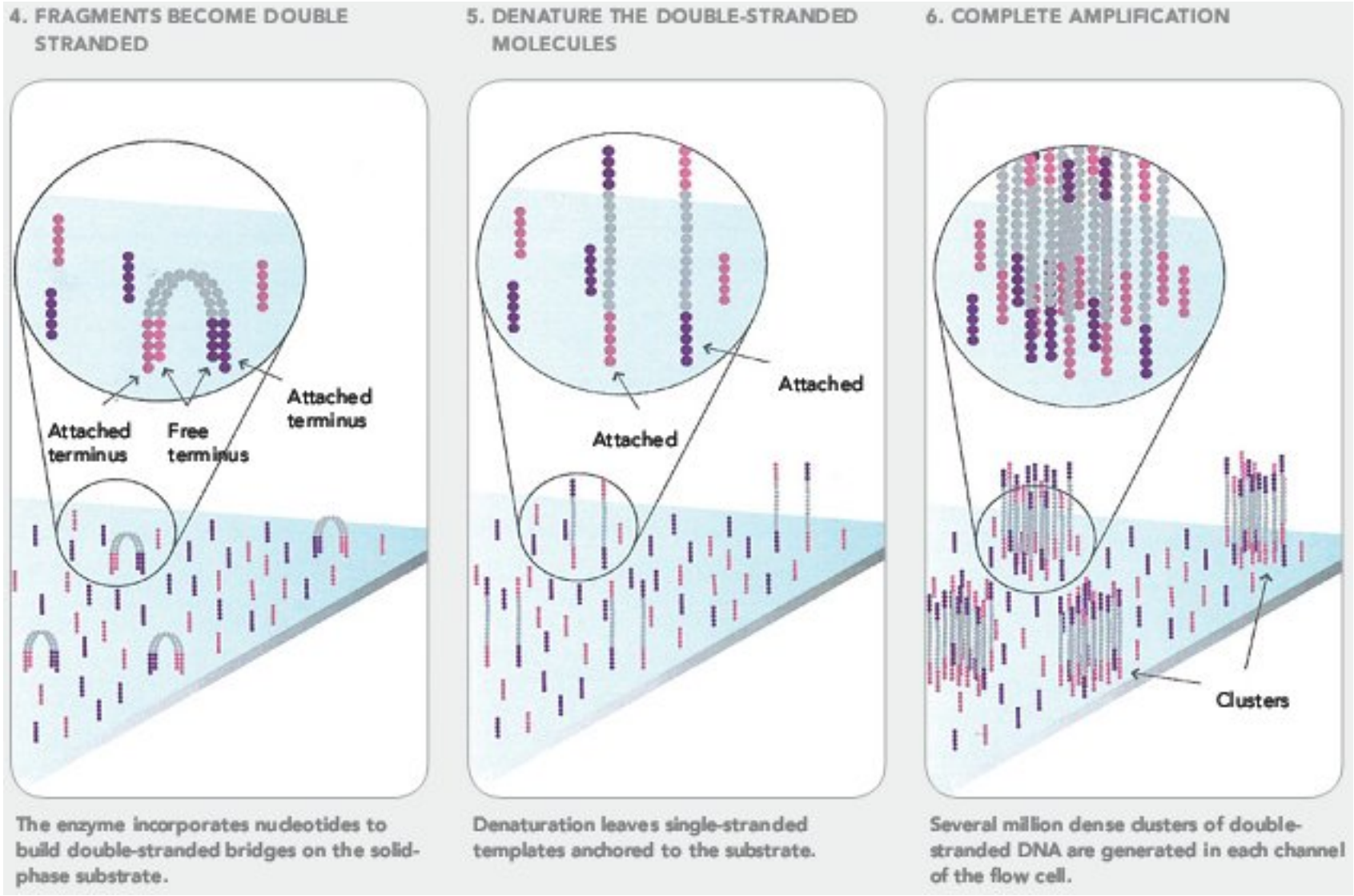
www.illumina.com



cc BY NC SA

www.illumina.com

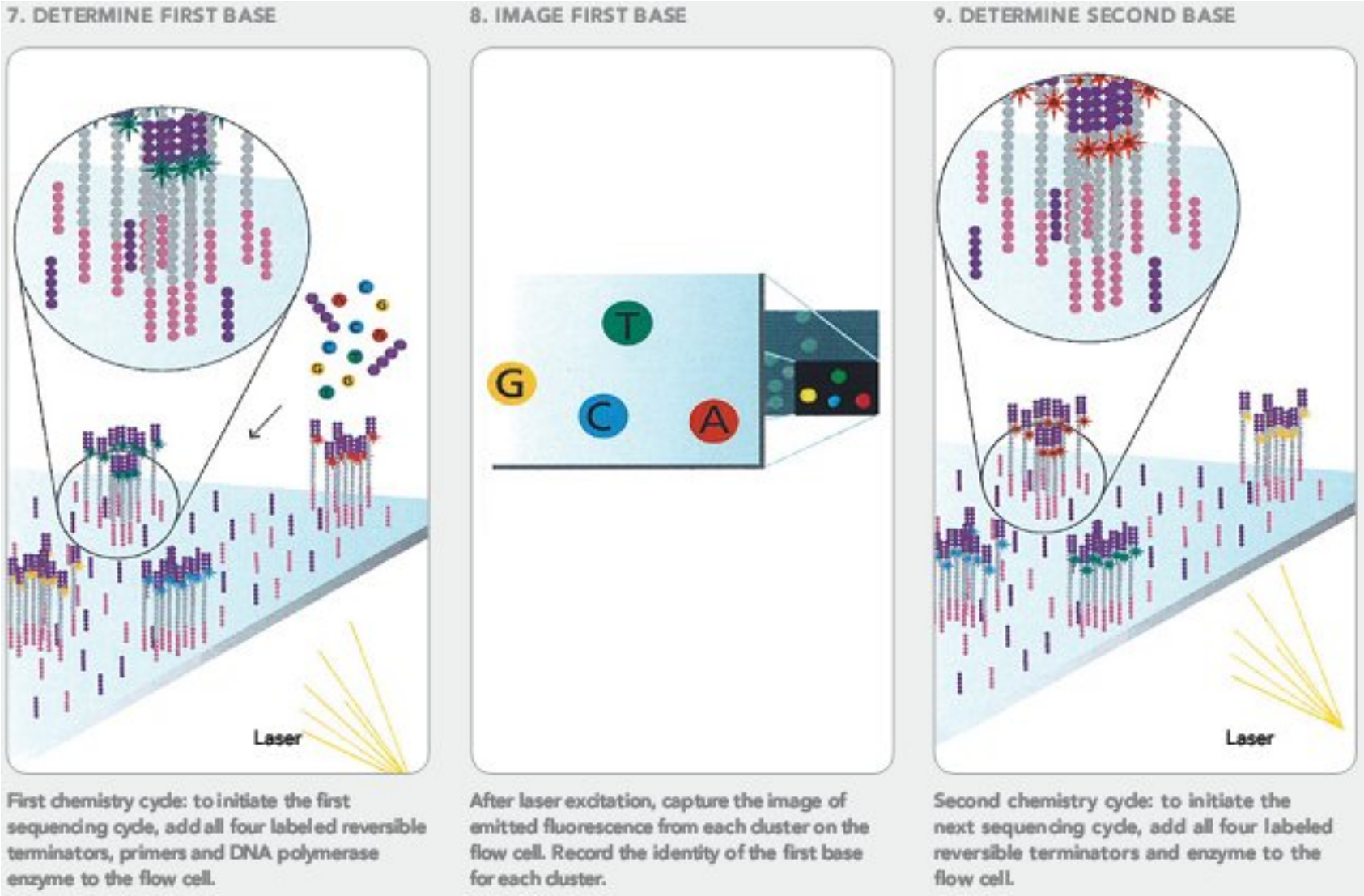
Illumina sequencing



CC BY NC SA

Illumina sequencing

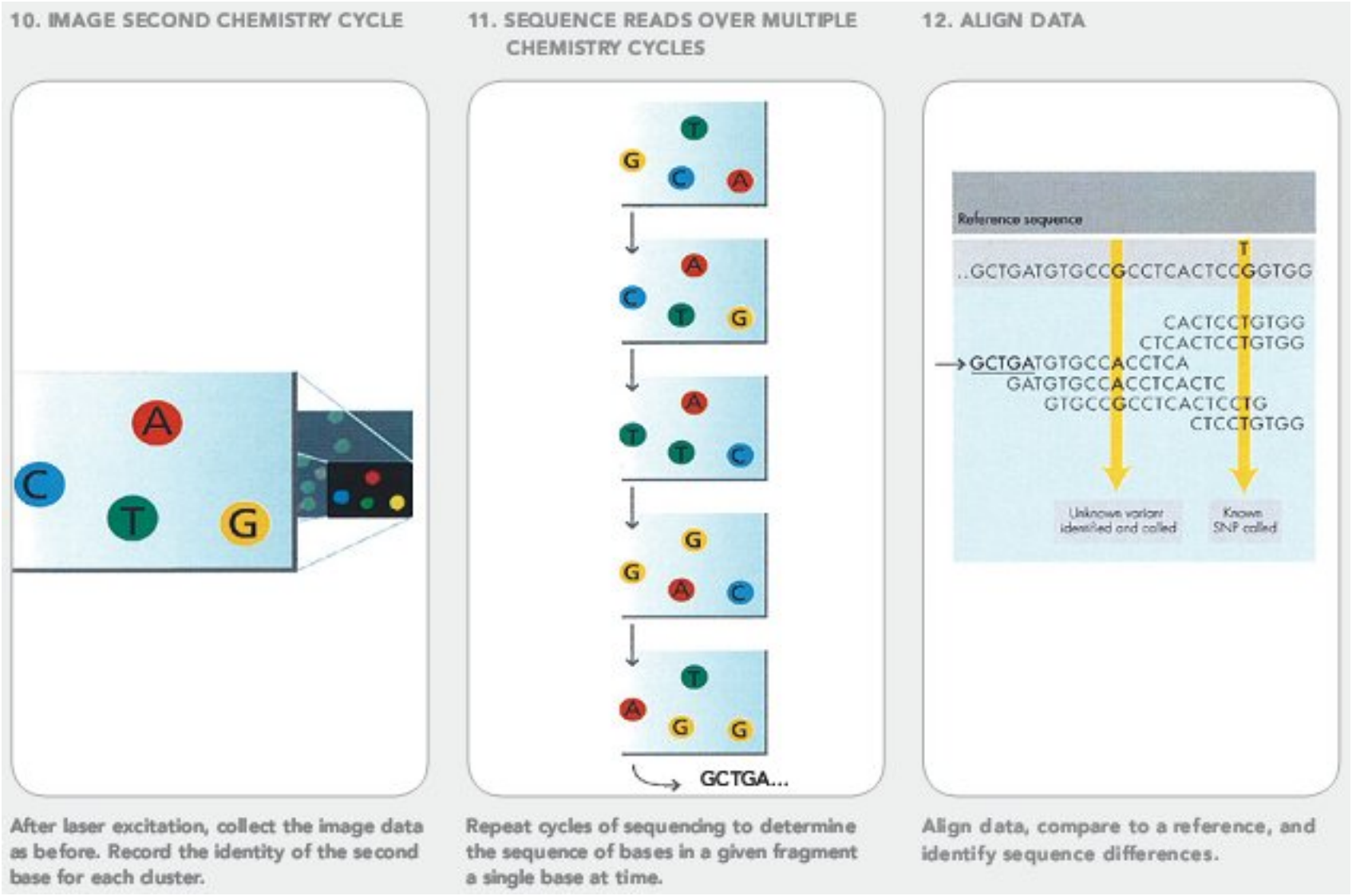
www.illumina.com



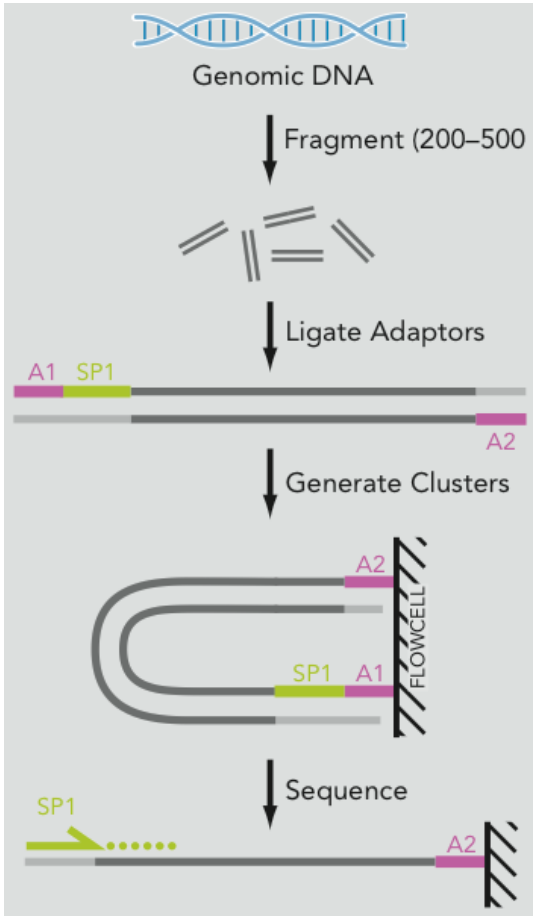
cc BY NC SA

Illumina sequencing

www.illumina.com



cc BY NC SA

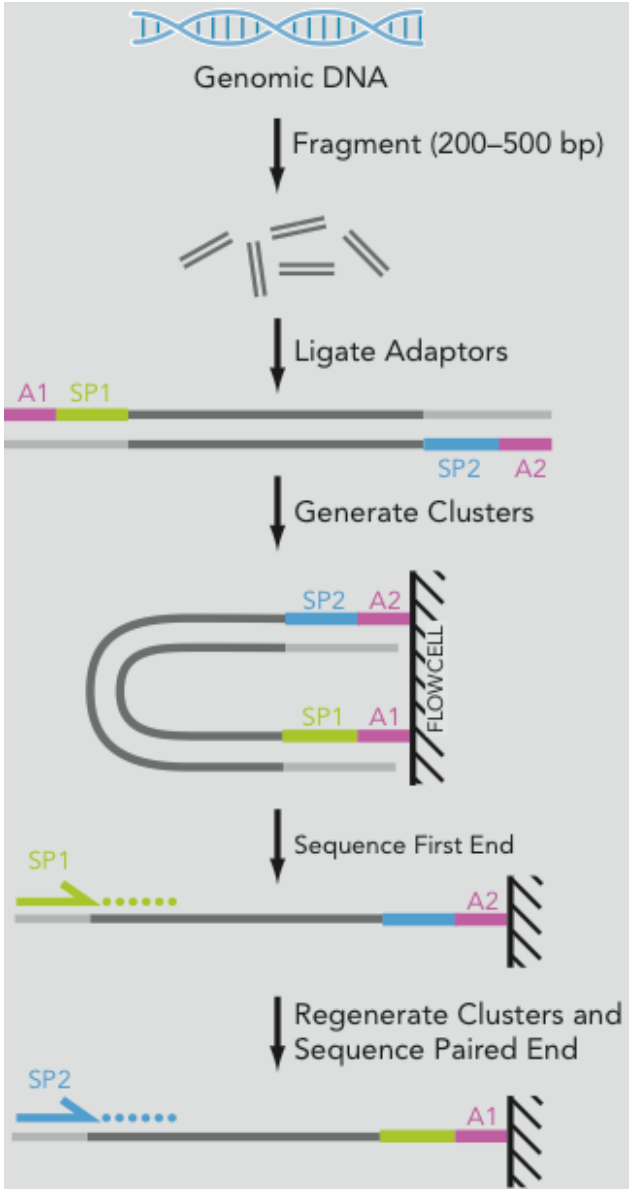


Single-end sequencing

- Size select DNA fragments and add adapters (A1 and A2).
- A1 also has sequencing primer (SP1) attached.
- Sequencing occurs on A2 fragment.

http://www.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf

cc BY NC SA

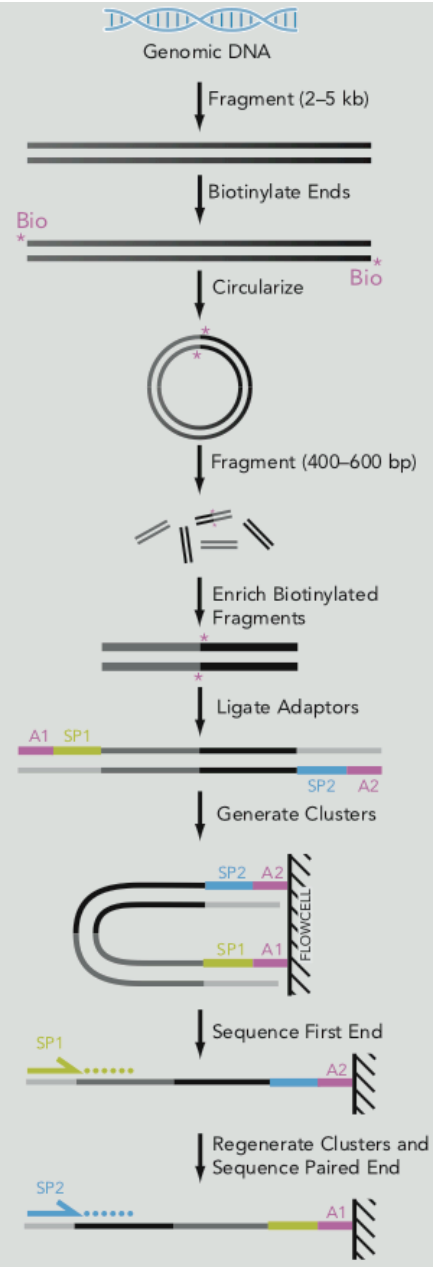


Paired-end sequencing

- Size select DNA fragments and add adapters and primers (A1+SP1 and A2+SP2).
- Sequencing occurs on both fragments.
- The distance between the sequenced read pair is called the "insert size".

http://www.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf

cc BY NC SA



Mate-pair sequencing

- Label ends of long DNA fragment.
- Circularize and fragment again.
- Enrich (amplify) the biotin labeled fragments.
- Proceed as for paired-end reads (basically "mate-pairs" are paired ends with a long insert size).

http://www.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf

cc BY NC SA

Nature Reviews Genetics | AOP, published online 8 December 2009; doi:10.1038/nrg2626



APPLICATIONS OF NEXT-GENERATION SEQUENCING

Sequencing technologies — the next generation

*Michael L. Metzker**‡

CC BY NC SA

Experiments

- DNA-seq: de novo, resequencing
- RNA-seq: mRNA, ncRNA, smRNA…
- ChIP-seq: Chromatin ImmunoPrecipitation
- Methyl-seq: methylated DNA (epigenome)

CC BY NC SA

Assessing quality: phred scores

Phred quality scores were originally developed by the program Phred to help in the automation of DNA sequencing in the Human Genome Project. Phred quality scores are assigned to each base call in automated sequencer traces.[1][2] Phred quality scores have become widely accepted to characterize the quality of DNA sequences, and can be used to compare the efficacy of different sequencing methods. Perhaps the most important use of Phred quality scores is the automatic determination of accurate, quality-based consensus sequences.

http://en.wikipedia.org/wiki/Phred_quality_score

$$Q = -10\log_{10} P$$

P=error probability of a given base call

Ewing B, Green P. (1998): Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8(3):186-194.

cc BY NC SA

Assessing quality: phred scores

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

http://en.wikipedia.org/wiki/Phred_quality_score

Can use ASCII to represent quality scores by adding 33 to the phred score (although Illumina scores use an offset of 64) and converting to ASCII.

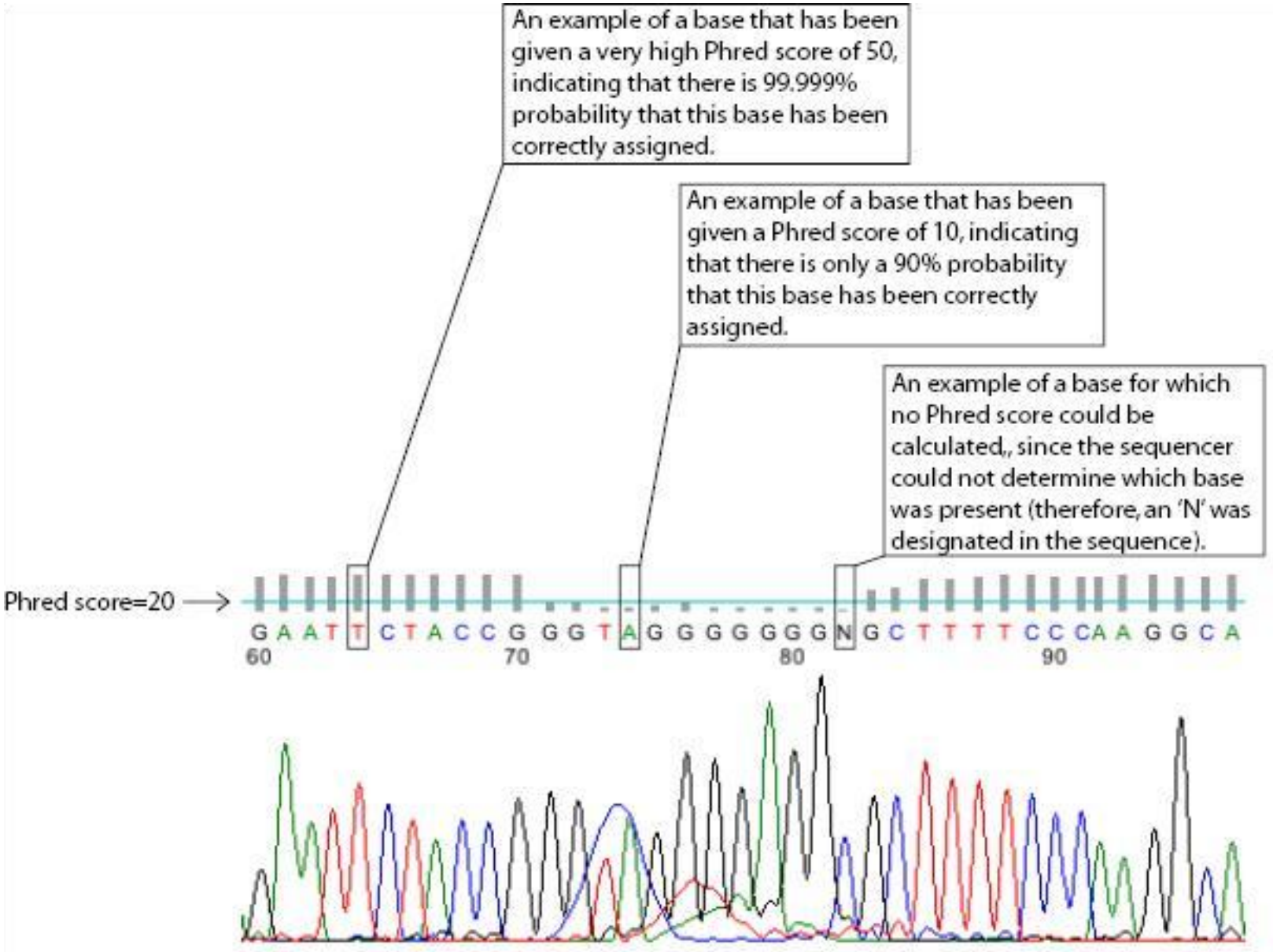
- Illumina quality score of 40 becomes 40+64=104: "h"

cc BY NC SA

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

Source: www.LookupTables.com

CC BY NC SA



http://en.wikipedia.org/wiki/Phred_quality_score

cc BY NC SA

Data format

- The FASTQ format allows the storage of both sequence and quality information for each read.
- This is a compact text-based format that has become the *de facto* standard for storing data from next generation sequencing experiments.

cc BY NC SA

Fastq format

```
@HWUSI-EAS582_157:6:1:1:1501/1
NCACAGACACACACGAACACACAAAGACATGCCCATATGAAGAT
+
%.7786867:7785568587465750588873/347777476035
@HWUSI-EAS582_157:6:1:1:1606/1
NCTGGCACCTTGATTTTGGACTTCCCAGCCTCCAGAACTGTGAG
+
%1948988888798988366898888648998788898888588
@HWUSI-EAS582_157:6:1:1:453/1
NCTGCTTGCACCCCTGAAGTCACTGATCACATTTCAGGGTCACC
+
%/868998988888867668888986644788988413488885
@HWUSI-EAS582_157:6:1:1:1844/1
NGATTGACATTGGCAAAGAGGACAACTGATTGCAAACTTCACAC
+
%-7;::::::;86499;75574586::635:62687666887879
@HWUSI-EAS582_157:6:1:1:1707/1
NAGGCTCAGGCGCACGGCCTACATCGTCGCTGTCGGCCAAGGGG
+
```

CC BY NC SA

FASTQ format

```
@HWUSI-EAS582_157:6:1:1:1501/1
NCACAGACACACACGAACACACAAAGACATGCCCATATGAAGAT
+
%.7786867:7785568587465750588 73/347777476035
@HWUSI-EAS582_157:6:1:1:1606/1
NCTGGCACCTTGATTTTGGACTTCCCAGCCTCCAGAACTGTGAG
+
%1948988888798988366898888648998788898888588
@HWUSI-EAS582_157:6:1:1:453/1
NCTGCTTGCACCCCTGAAGTCACTGATCACATTTCAGGGTCACC
+
%/8689989888888676688889866447889884134888 85
@HWUSI-EAS582_157:6:1:1:1844/1
NGATTGACATTGGCAAAGAGGACAACTGATTGCAAACTTCACAC
+
%-7;:::::;86499;75574586::635:62687666887879
@HWUSI-EAS582_157:6:1:1:1707/1
NAGGCTCAGGCGCACGGCCTACATCGTCGCTGTCGGCCAAGGGG
+
```

"Read" (sequence)

Quality scores (phred-33)

Illumina sequence identifiers

Sequences from the Illumina software use a systematic identifier:

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

http://en.wikipedia.org/wiki/FASTQ_format

cc BY NC SA

FastQC

- Simple java-based tool for quality assessment of next generation sequencing data.
- Takes FASTQ file as input and generates multiple QC plots.
- No ability to customize or interact with plots.

cc BY NC SA

FastQC



FastQC

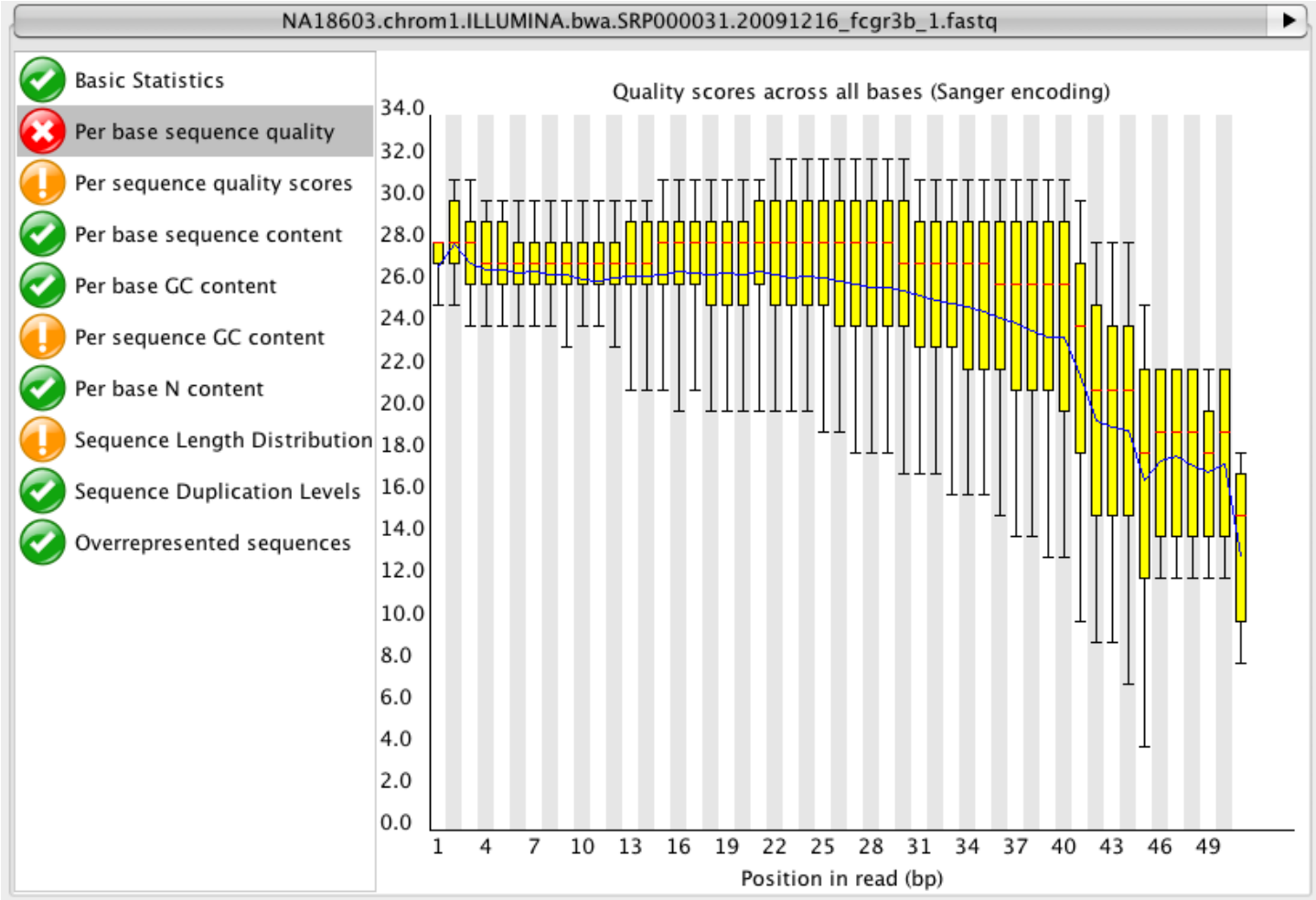
Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

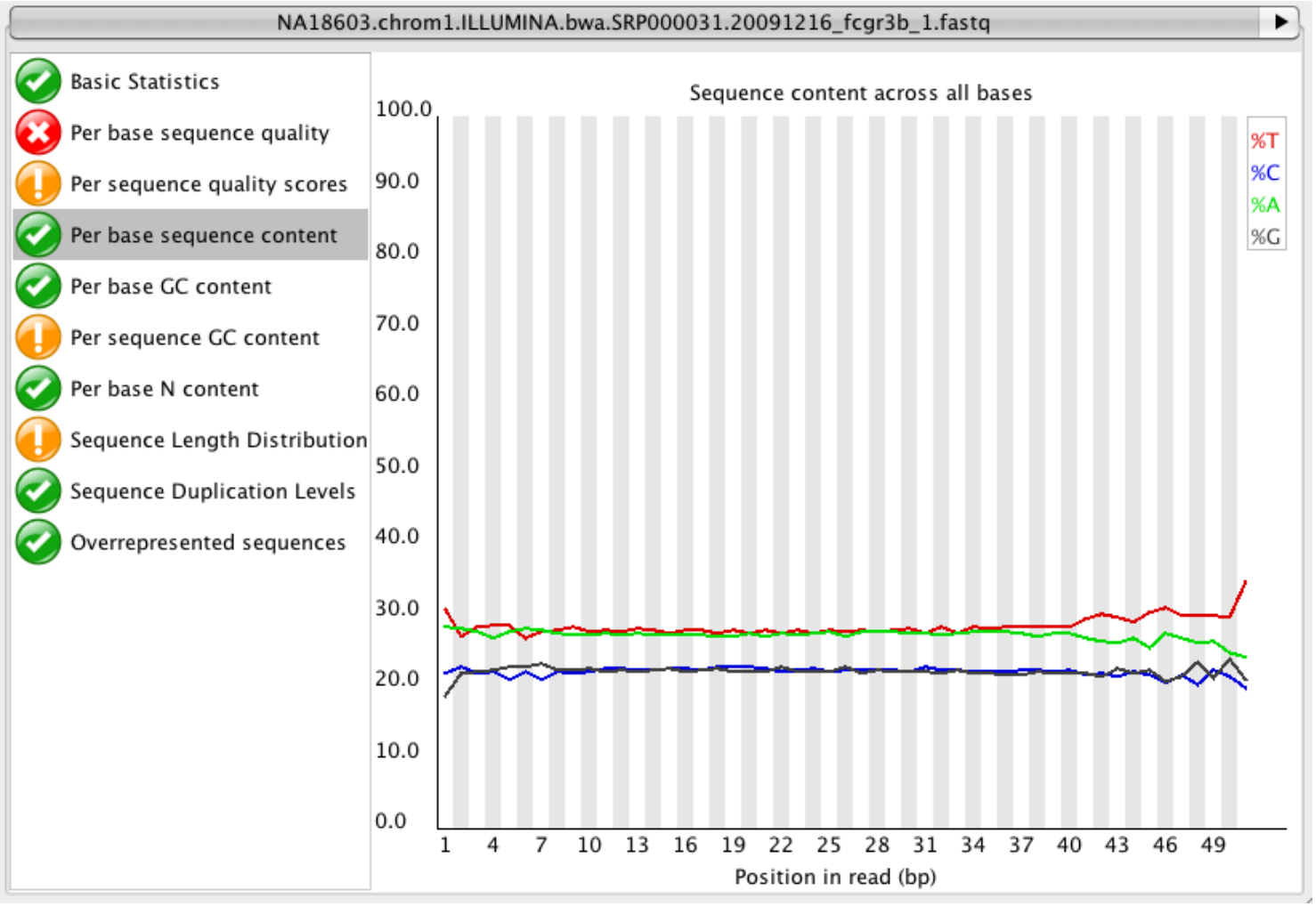
cc BY NC SA

FastQC



CC BY NC SA

FastQC



CC BY NC SA

Assembly

- *De novo* genome assembly based on short read data is a major computational task.
- A number of specialized tools exist:
  - ABySS, gap4, Geneious, Mira, Newbler, SSAKE, SOAPdenovo, Velvet….

Sense from sequence reads: methods for alignment and assembly

Paul Flicek & Ewan Birney

NATURE METHODS SUPPLEMENT | VOL.6 NO.11s | NOVEMBER 2009

cc BY NC SA

Goecks *et al. Genome Biology* 2010, **11**:R86
http://genomebiology.com/2010/11/8/R86



SOFTWARE Open Access

Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences

Jeremy Goecks[1], Anton Nekrutenko[2*], James Taylor[1*], The Galaxy Team

- Galaxy provides a web-based application for the analysis of sequence data
- Includes tools for the analysis and manipulation of NGS data
- Simple and extensible interface

CC BY NC SA

http://galaxy.psu.edu/

Galaxy

use | see | develop

Use it now!

Galaxy allows you to do analyses you cannot do anywhere else without the need to install or download anything. You can analyze multiple alignemnts, compare genomic annotations, profile metagenomic samples and much much more...

- Go to our public site and start working.
- Watch screencasts to get an idea of what we mean.
- Go to our wiki page to read about technical details.



Galaxy = collaboration and reproducibility

Best of all, Galaxy's history system provides a complete analyses record that can be shared. Every history is an analysis workflow, which can be used to reproduce the entire experiment...

- **History is an analysis record** | Every step of your analyses is recorded in Galaxy's history system. You can have any number of histories saved. This way you can go back to your analyses anytime.
- **Share your analyses** | Alice works at Penn State, while Bob suffers from the terrible San Diego climate. Alice wants Bob to see her analyses. Alice clicks the "share" link and enters Bob's e-mail address. Now Alice's history is visible to Bob (see "Sharing history" screencast).
- **Now your results are reproducible!** | When publishing results, replace "the data were analyzed using a collection of in-house scripts" with a URL pointing to Galaxy's history. Your reviewers will have no further questions. That's reproducible genomics!

Designed for biologists and developers

Yep, sometimes you can mix water with oil...

- **Biologists** | Use our public site to access popular sources of data like the UCSC Table Browser. Run analyses right on the spot using a variety of integrated tools. Your results are never deleted and can be easily shared with others.
- **Developers** | Galaxy is an easy-to-use, open-source, scalable framework for tool and data integration. Stop wasting time writing interfaces and get your tools used by biologists! Galaxy includes everything you need to get started, so download and start integrating!

cc BY NC SA

Citing Galaxy

http://bitbucket.org/galaxy/galaxy-central/wiki/Citations

If you use or extend Galaxy in your published work, please cite the following publications:

Goecks, J, Nekrutenko, A, Taylor, J and The Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* 2010 Aug 25;11(8):R86.

Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J. "Galaxy: a web-based genome analysis tool for experimentalists". *Current Protocols in Molecular Biology*. 2010 Jan; Chapter 19:Unit 19.10.1-21.

Other Publications about Galaxy

Please cite the appropriate publication if using the specific features of Galaxy described

Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A and The Galaxy Team. "Manipulation of FASTQ data with Galaxy". *Bioinformatics*. 2010 Jul 15;26(14):1783-5.

Bock C, Von Kuster G, Halachev K, Taylor J, Nekrutenko A, Lengauer T. "Web-based analysis of (epi-) genome data using EpiGRAPH and Galaxy". In "Methods in Molecular Biology: Genetic Variation: Methods And Protocols" (ed. M.R. Barnes and G. Breen). Humana Press, Towata, NJ. 2010 Feb.

Koskovsky Pond S, Wadhawan S, Chiaromonte F, Ananda G, Chung W, Taylor J, and Nekrutenko A. "Windshield splatter analysis with the Galaxy metagenomic pipeline". *Genome Research*. 2009 Nov; 19(11):2144-53.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. "28-Way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Research*. 2007 Dec; 17(12):1797-808.

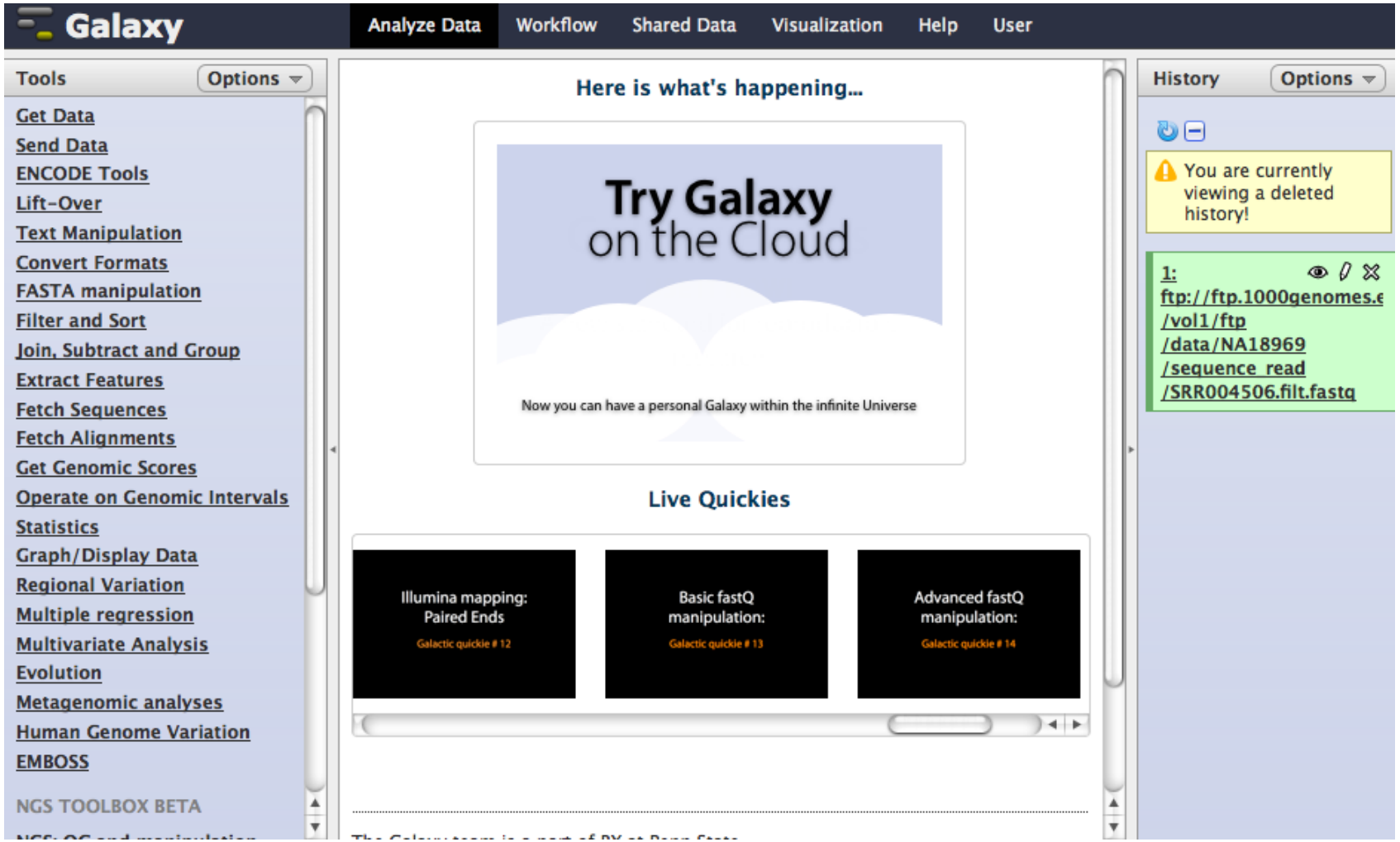
Taylor J, Schenk I, Blankenberg D, and Nekrutenko, A. "Using Galaxy to Perform Large-Scale Interactive Data Analysis". *Current Protocols in Bioinformatics*. 2007 Sep; 19:10.5.1-10.5.25.

Blankenberg D, Taylor J, Schenk I, He J, Zhang Y, Ghent M, Veeraraghavan N, Albert I, Miller W, Makova K, Hardison RC, Nekrutenko A. "A framework for collaborative analysis of ENCODE data: Making large-scale analyses biologist-friendly". *Genome Research*. 2007 Jun; 17(6):960-4.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." *Genome Research*. 2005 Oct; 15(10):1451-5.

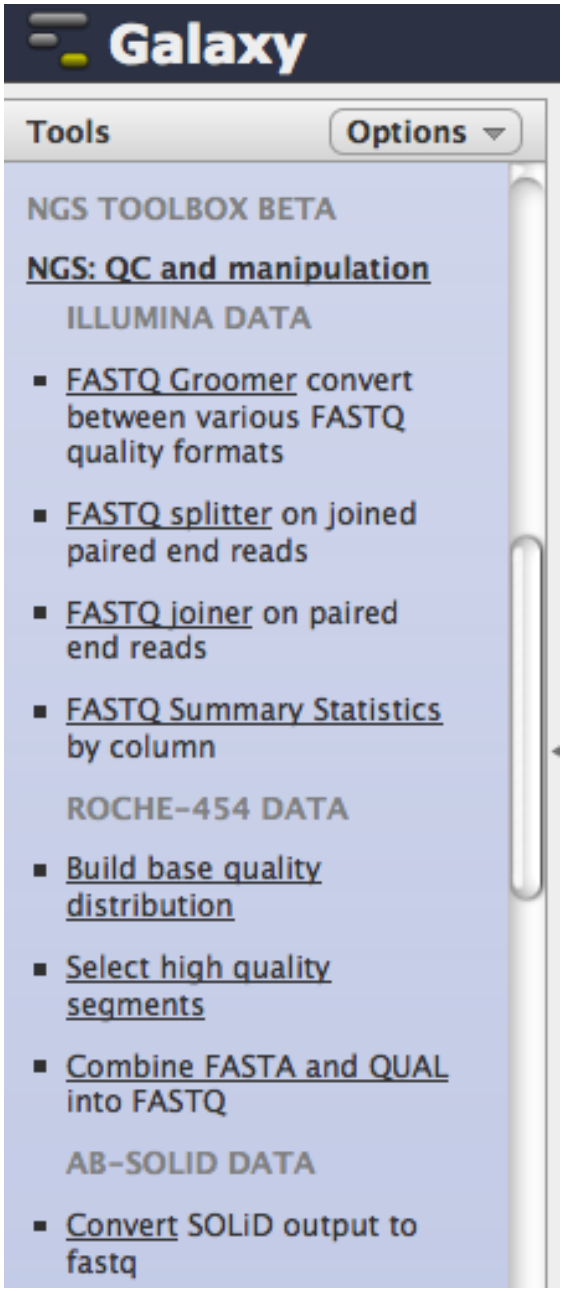
CC BY NC SA

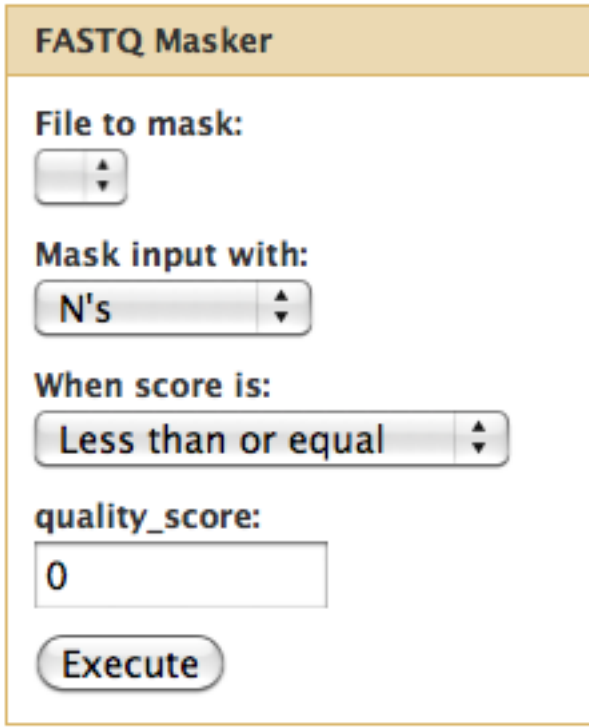
http://main.g2.bx.psu.edu/



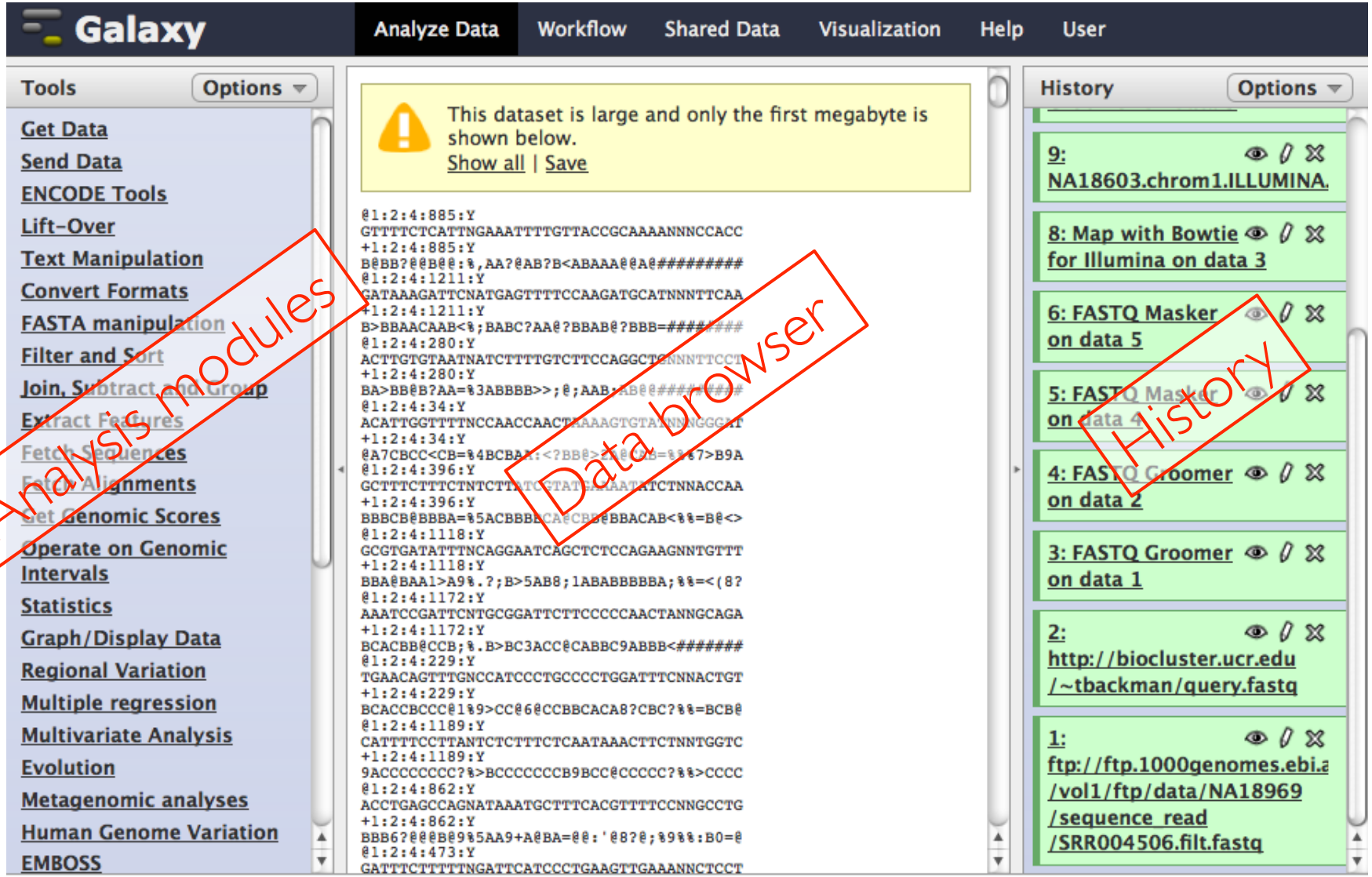
CC BY NC SA


Galaxy
Tools
Options
NGS TOOLBOX BETA
NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling

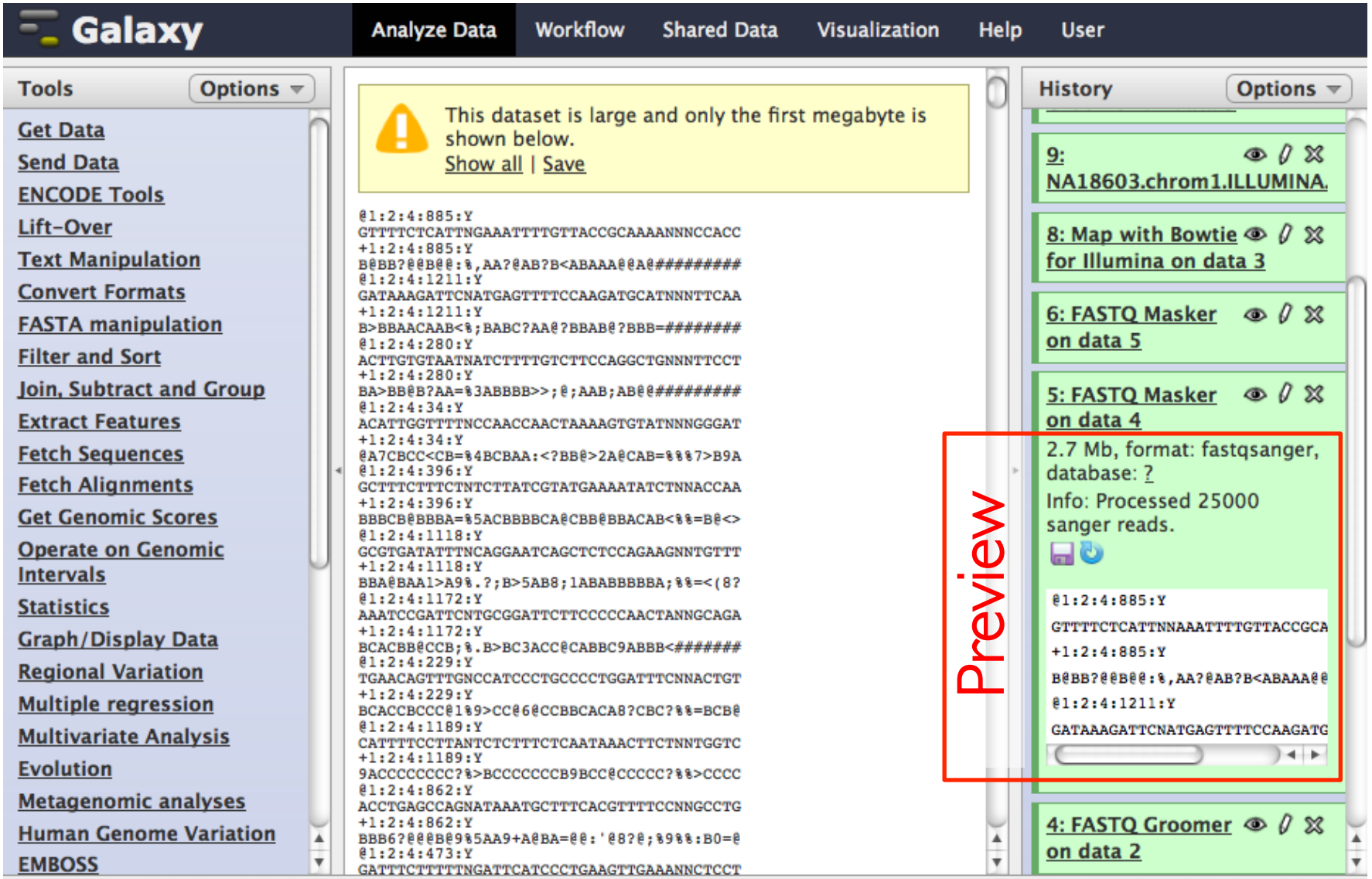

Galaxy
Tools
Options
NGS TOOLBOX BETA
NGS: QC and manipulation
ILLUMINA DATA
FASTQ Groomer convert between various FASTQ quality formats
FASTQ splitter on joined paired end reads
FASTQ joiner on paired end reads
FASTQ Summary Statistics by column
ROCHE-454 DATA
Build base quality distribution
Select high quality segments
Combine FASTA and QUAL into FASTQ
AB-SOLID DATA
Convert SOLiD output to fastq


FASTQ Masker
File to mask:
Mask input with:
N's
When score is:
Less than or equal
quality_score:
0
Execute

CC BY NC SA


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
This dataset is large and only the first megabyte is shown below.
Show all | Save
@1:2:4:885:Y
GTTTTCTCATTNGAAATTTTGTTACCGCAAAANNNCCACC
+1:2:4:885:Y
B@BB?@@B@@:%,AA?@AB?B<ABAAA@@A@#########
@1:2:4:1211:Y
GATAAAGATTCNATGAGTTTTCCAAGATGCATNNNTTCAA
+1:2:4:1211:Y
B>BBAACAAB<%;BABC?AA@?BBAB@?BBB=########
@1:2:4:280:Y
ACTTGTGTAATNATCTTTTGTCTTCCAGGCTGNNNTTCCT
+1:2:4:280:Y
BA>BB@B?AA=%3ABBBB>>;@;AAB:BB@@########
@1:2:4:34:Y
ACATTGGTTTTNCCAACCAACTAAAGTGTANNNGGGAT
+1:2:4:34:Y
@A7CBCC<CB=%4BCBAA:<?BB@>A@@CB=%%%7>B9A
@1:2:4:396:Y
GCTTTCTTTCTNTCTTATCTTATAAAATATCTNNACCAA
+1:2:4:396:Y
BBBCB@BBBA=%5ACBBBBCA@CB@@BBACAB<%%=B@<>
@1:2:4:1118:Y
GCGTGATATTTNCAGGAATCAGCTCTCCAGAAGNNTGTTT
+1:2:4:1118:Y
BBA@BAA1>A9%.?;B>5AB8;1ABABBBBBA;%%=<(8?
@1:2:4:1172:Y
AAATCCGATTCNTGCGGATTCTTCCCCCAACTANNGCAGA
+1:2:4:1172:Y
BCACBB@CCB;%.B>BC3ACC@CABBC9ABBB<#######
@1:2:4:229:Y
TGAACAGTTTGNCCATCCCTGCCCCTGGATTTCNNACTGT
+1:2:4:229:Y
BCACCBCCC@1%9>CC@6@CCBBCACA8?CBC?%%=BCB@
@1:2:4:1189:Y
CATTTTCCTTANTCTCTTTCTCAATAAACTTCTNNTGGTC
+1:2:4:1189:Y
9ACCCCCCCC?%>BCCCCCCCB9BCC@CCCCC?%%>CCCC
@1:2:4:862:Y
ACCTGAGCCAGNATAAATGCTTTCACGTTTTCCNNGCCTG
+1:2:4:862:Y
BBB6?@@@B@9%5AA9+A@BA=@@:'@8?@;%9%%:B0=@
@1:2:4:473:Y
GATTTCTTTTTNGATTCATCCCTGAAGTTGAAANNCTCCT
History
Options
9: NA18603.chrom1.ILLUMINA.
8: Map with Bowtie for Illumina on data 3
6: FASTQ Masker on data 5
5: FASTQ Masker on data 4
4: FASTQ Groomer on data 2
3: FASTQ Groomer on data 1
2: http://biocluster.ucr.edu/~tbackman/query.fastq
1: ftp://ftp.1000genomes.ebi.a/vol1/ftp/data/NA18969/sequence_read/SRR004506.filt.fastq
Analysis modules
Data browser
History

cc BY NC SA


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS
This dataset is large and only the first megabyte is shown below.
Show all | Save
@1:2:4:885:Y
GTTTTCTCATTNGAAATTTTGTTACCGCAAAANNNCCACC
+1:2:4:885:Y
B@BB?@@B@@:%,AA?@AB?B<ABAAA@@A@#########
@1:2:4:1211:Y
GATAAAGATTCNATGAGTTTTCCAAGATGCATNNNTTCAA
+1:2:4:1211:Y
B>BBAACAAB<%;BABC?AA@?BBAB@?BBB=########
@1:2:4:280:Y
ACTTGTGTAATNATCTTTTGTCTTCCAGGCTGNNNTTCCT
+1:2:4:280:Y
BA>BB@B?AA=%3ABBBB>>;@;AAB;AB@@#########
@1:2:4:34:Y
ACATTGGTTTTNCCAACCAACTAAAAGTGTATNNNGGGAT
+1:2:4:34:Y
@A7CBCC<CB=%4BCBAA:<?BB@>2A@CAB=%%%7>B9A
@1:2:4:396:Y
GCTTTCTTTCTNTCTTATCGTATGAAAATATCTNNACCAA
+1:2:4:396:Y
BBBCB@BBBA=%5ACBBBBCA@CBB@BBACAB<%%=B@<>
@1:2:4:1118:Y
GCGTGATATTTNCAGGAATCAGCTCTCCAGAAGNNTGTTT
+1:2:4:1118:Y
BBA@BAA1>A9%.?;B>5AB8;1ABABBBBBA;%%=<(8?
@1:2:4:1172:Y
AAATCCGATTCNTGCGGATTCTTCCCCCAACTANNGCAGA
+1:2:4:1172:Y
BCACBB@CCB;%.B>BC3ACC@CABBC9ABBB<#######
@1:2:4:229:Y
TGAACAGTTTGNCCATCCCTGCCCCTGGATTTCNNACTGT
+1:2:4:229:Y
BCACCBCCC@1%9>CC@6@CCBBCACA8?CBC?%%=BCB@
@1:2:4:1189:Y
CATTTTCCTTANTCTCTTTCTCAATAAACTTCTNNTGGTC
+1:2:4:1189:Y
9ACCCCCCCC?%>BCCCCCCCB9BCC@CCCCC?%%>CCCC
@1:2:4:862:Y
ACCTGAGCCAGNATAAATGCTTTCACGTTTTCCNNGCCTG
+1:2:4:862:Y
BBB6?@@@B@9%5AA9+A@BA=@@:'@8?@;%9%%:B0=@
@1:2:4:473:Y
GATTTCTTTTTNGATTCATCCCTGAAGTTGAAANNCTCCT
History
Options
9: NA18603.chrom1.ILLUMINA.
8: Map with Bowtie for Illumina on data 3
6: FASTQ Masker on data 5
5: FASTQ Masker on data 4
2.7 Mb, format: fastqsanger, database: ?
Info: Processed 25000 sanger reads.
@1:2:4:885:Y
GTTTTCTCATTNNAAATTTTGTTACCGCA
+1:2:4:885:Y
B@BB?@@B@@:%,AA?@AB?B<ABAAA@@
@1:2:4:1211:Y
GATAAAGATTCNATGAGTTTTCCAAGATG
4: FASTQ Groomer on data 2
Preview

CC BY NC SA

2: http://biocluster.ucr.edu/~tbackman/query.fastq

99,975 lines, format: qualillumina, database: ?

Info: uploaded fastqsolexa file

```
@1:2:4:885:Y
GTTTTCTCATTNGAAATTTTGTTACCGCAAAANNNCCACC
+1:2:4:885:Y
B@BB?@@B@@:%,AA?@AB?B<ABAAA@@A@#########
@1:2:4:1211:Y
GATAAAGATTCNATGAGTTTTCCAAGATGCATNNNTTCAA
```

Import fastq data

"Groom" imported data for use in other modules

Mask low quality bases with N's

4: FASTQ Groomer on data 2

99,975 lines, format: qual, database: ?

Info: Groomed 25000 sanger reads into sanger reads. Based upon quality and sequence, the input data is valid for: sanger

Input ASCII range: '#'(35) – 'C'(67)

Input decimal range: 2 – 34

```
@1:2:4:885:Y
GTTTTCTCATTNGAAATTTTGTTACCGCAAAANNNCCACC
+1:2:4:885:Y
B@BB?@@B@@:%,AA?@AB?B<ABAAA@@A@#########
@1:2:4:1211:Y
GATAAAGATTCNATGAGTTTTCCAAGATGCATNNNTTCAA
```

5: FASTQ Masker on data 4

2.7 Mb, format: fastqsanger, database: ?

Info: Processed 25000 sanger reads.

```
@1:2:4:885:Y
GTTTTCTCATTNNAAATTTTGTTACCGCAAANNNNNNNNN
+1:2:4:885:Y
B@BB?@@B@@:%,AA?@AB?B<ABAAA@@A@#########
@1:2:4:1211:Y
GATAAAGATTCNATGAGTTTTCCAAGATGCATNNNNNNNN
```

CC BY NC SA

Alignment

- If a "reference" genome exists for the organism you are sequencing, reads can be "aligned" to the reference.
- This involves finding the place in the reference genome that each read matches to.
- Due to high sequence similarity within members of the same species, most reads *should* map to the reference.

cc BY NC SA

Tools for generating alignments

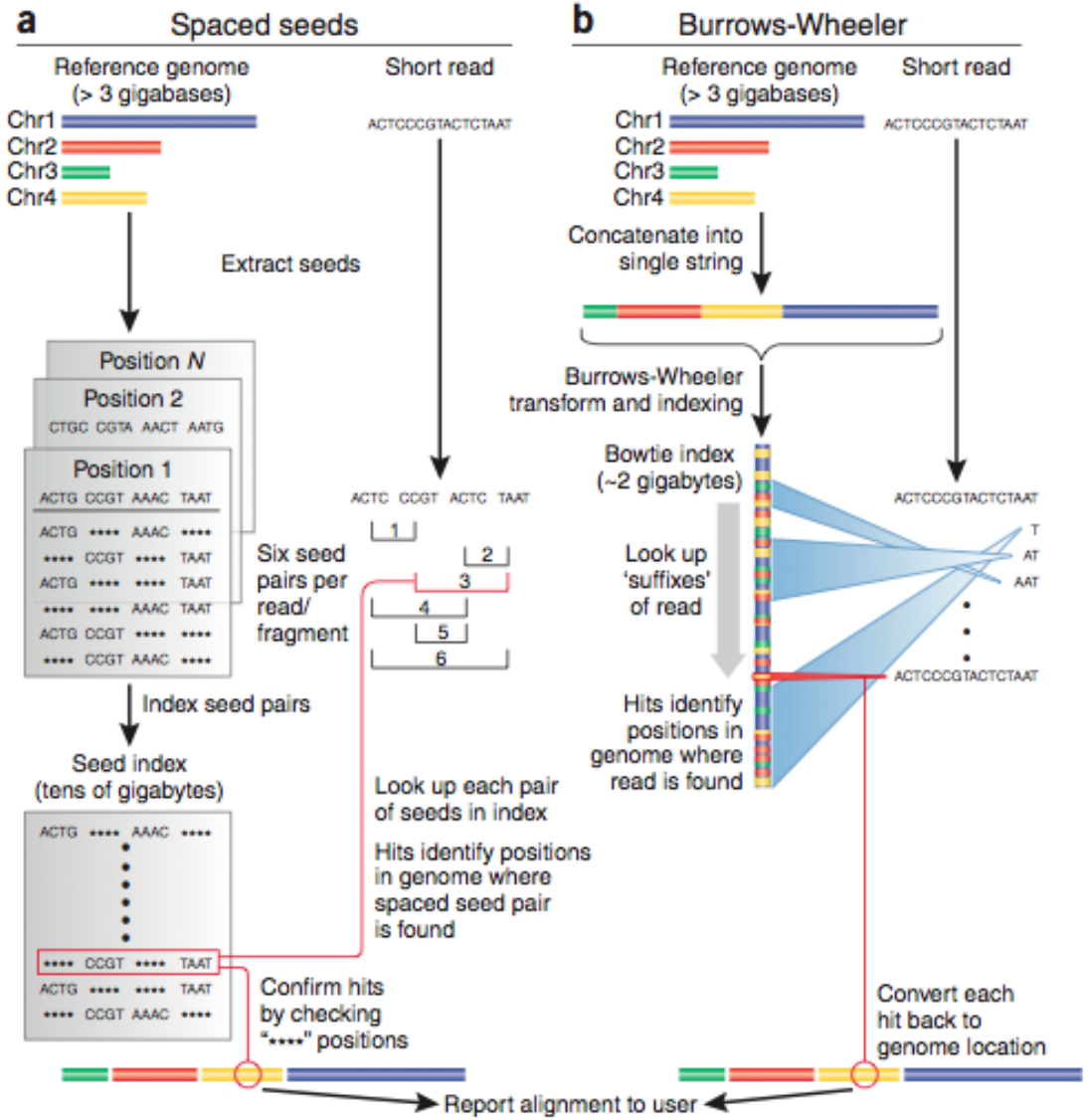
- There are MANY software packages available for aligning data from next generation sequencing experiments.
- Two of the most popular are:
  - BWA: http://bio-bwa.sourceforge.net
  - Bowtie: http://bowtie-bio.sourceforge.net
- Both utilize the "Burrows-Wheeler Transform."

cc BY NC SA

How to map billions of short reads onto genomes

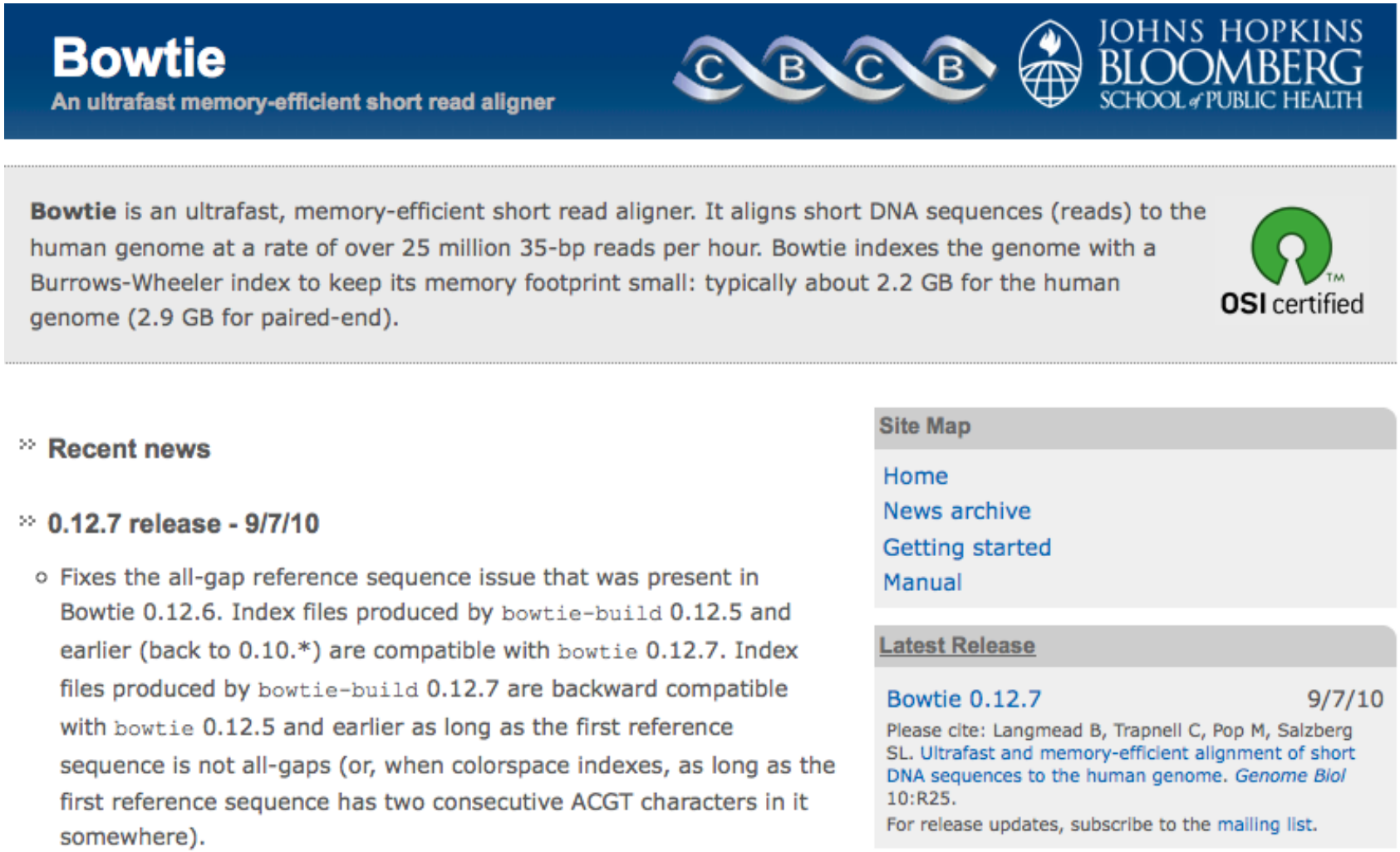
VOLUME 27 NUMBER 5 MAY 2009 NATURE BIOTECHNOLOGY

Cole Trapnell & Steven L Salzberg



CC BY NC SA

http://bowtie-bio.sourceforge.net/index.shtml



CC BY NC SA

http://bowtie-bio.sourceforge.net/index.shtml

Open Access

Software

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Address: Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742, USA.

Correspondence: Ben Langmead. Email: langmead@cs.umd.edu

Published: 4 March 2009

Genome ***Biology*** 2009, **10**:R25 (doi:10.1186/gb-2009-10-3-r25)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/3/R25

Received: 21 October 2008
Revised: 19 December 2008
Accepted: 4 March 2009

© 2009 Langmead *et al.*; licensee BioMed Central Ltd.
This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

http://bio-bwa.sourceforge.net/

Burrows-Wheeler Aligner

Home

Introduction

Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and BWA-SW. The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kbp. Both algorithms do gapped alignment. They are usually more accurate and faster on queries with low error rates. Please see the BWA manual page for more information.

FAQ

How can I cite BWA?

The short read alignment component (bwa-short) has been published:

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60. [PMID: 19451168]

If you use BWA-SW, please cite:

Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. Bioinformatics, Epub. [PMID: 20080505]

BWA:

- SF project page
- SF download page
- Mailing list
- BWA maual page
- Release notes

Links:

- SAMtools
- MAQ

cc BY NC SA

http://bio-bwa.sourceforge.net/

BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 14 2009, pages 1754–1760
doi:10.1093/bioinformatics/btp324

Sequence analysis

Fast and accurate short read alignment with Burrows–Wheeler transform

Heng Li and Richard Durbin*

Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK

Received on February 20, 2009; revised on May 6, 2009; accepted on May 12, 2009

Advance Access publication May 18, 2009

Associate Editor: John Quackenbush

CC BY NC SA

Alignment formats

- SAM (Sequence Alignment/Map) format has become the *de facto* standard for storing alignment data.
- BAM is a binary version of SAM allowing more efficient storage.

CC BY NC SA

BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 16 2009, pages 2078–2079
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

CC BY NC SA

http://samtools.sourceforge.net/

SAMtools

SOURCEFORGE.NET®

Home

Introduction

SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments. SAM aims to be a format that:

- Is flexible enough to store all the alignment information generated by various alignment programs;
- Is simple enough to be easily generated by alignment programs or converted from existing alignment formats;
- Is compact in file size;
- Allows most of operations on the alignment to work on a stream without loading the whole alignment into memory;
- Allows the file to be indexed by genomic position to efficiently retrieve all reads aligning to a locus.

SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

General Information

SAM Format Specification
SF Project Page
SF Download Page
Mailing Lists
SVN Browse
Related Software
FAQ

SAMtools in C

General Introduction
Manual Page
Pileup Format
Multisample pileup
Consensus/Indel Calling
Text Alignment Viewer
API Documentation

cc BY NC SA

http://samtools.sourceforge.net/

SAMtools

- SAMtools provides a command line interface for manipulation of SAM/BAM formatted data.
- Open source and multi-platform (R package available: Rsamtools).
- Able to:
  - Extract reads from specific genomic region
  - Operate on remote files
  - Much more….

cc BY NC SA

SAM format

```
ERR005646.11088674      147     1       161099954       60
54M     =       161099742       -266
TTTTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACT
89=<;@>EECFCBBFFCAEFBGB=FFFC?@AB@G=FFB@CABABA?A@<>>=;=
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646  OQ:Z:D?
FFEEEFFFFFFFFFEFFDFECFFFE;EEEEFCFFEEEEFEFECEEC=E;EF
ERR005646.5518024       147     1       161099956       60
54M     =       161099847       -163
TTCTGAACAGGGATGATATTTGTAATTTCATAGAATTAAGAGATATCTGACTCT  :
68=<A@@A???AB?A>ABBB>@CABCAAA>B@BAB@BA@A@A@A@=A=A=>;<
XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0
XO:i:0  XG:i:0  MD:Z:54 RG:Z:ERR005646
OQ:Z:ECEEEEEEEEDEEEEE>EEEEEEEEEEEEE@EEEEEBEEEEEEEEECCCBEEEE
```

CC BY NC SA

http://samtools.sourceforge.net/

SAM format

Type	Tag	Description
HD - header	VN*	File format version.
	SO	Sort order. Valid values are: *unsorted*, *queryname* or *coordinate*.
	GO	Group order (full sorting is not imposed in a group). Valid values are: *none*, *query* or *reference*.
SQ - Sequence dictionary	SN*	Sequence name. Unique among all sequence records in the file. The value of this field is used in alignment records.
	LN*	Sequence length.
	AS	Genome assembly identifier. Refers to the reference genome assembly in an unambiguous form. Example: HG18.
	M5	MD5 checksum of the sequence in the uppercase (gaps and space are removed)
	UR	URI of the sequence
	SP	Species.
RG - read group	ID*	Unique read group identifier. The value of the ID field is used in the RG tags of alignment records.
	SM*	Sample (use pool name where a pool is being sequenced)
	LB	Library
	DS	Description
	PU	Platform unit (e.g. lane for Illumina or slide for SOLiD); should be a full, unambiguous identifier
	PI	Predicted median insert size (maybe different from the actual median insert size)
	CN	Name of sequencing center producing the read.
	DT	Date the run was produced (ISO 8601 date or date/time).
	PL	Platform/technology used to produce the read.
PG - Program	ID*	Program name
	VN	Program version
	CL	Command line
CO - comment		One-line text comments

CC BY NC SA

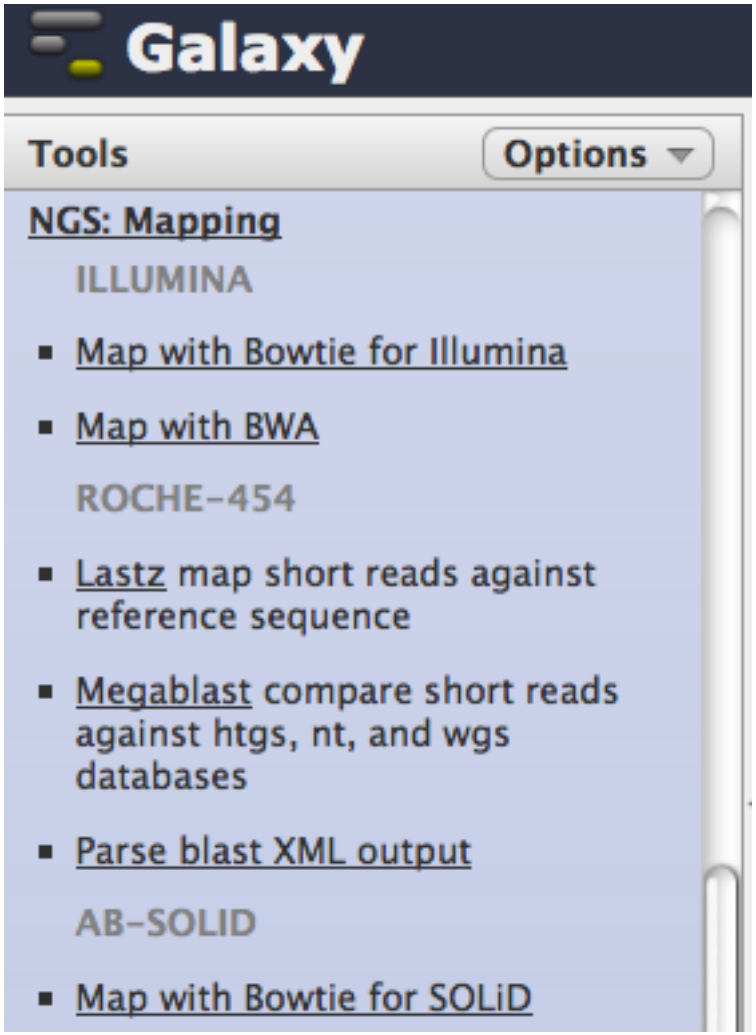
Alignment using BWA and SAMtools

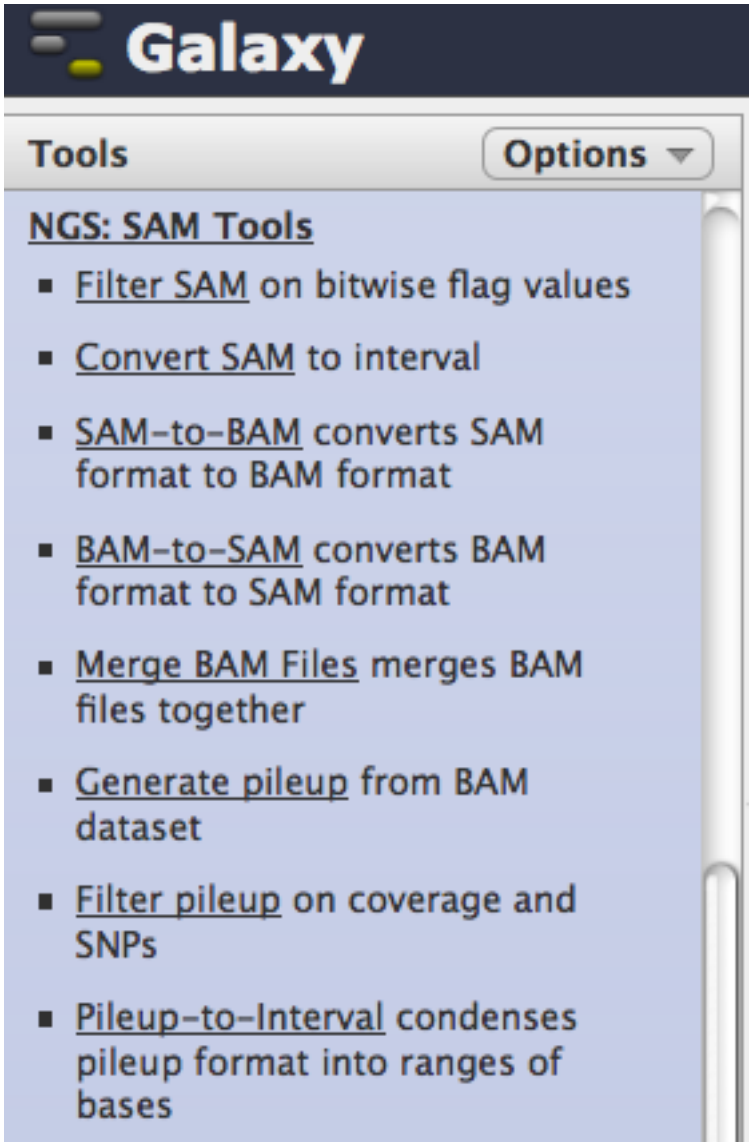
```
# download a test reference genome (TAIR9 Chromosome 1)
wget http://biocluster.ucr.edu/~tbackman/genome.fasta
# download some test Illumina reads from Arabidopsis
wget http://biocluster.ucr.edu/~tbackman/query.fastq
# index reference genome
bwa index -a is genome.fasta
# perform alignment
bwa aln genome.fasta query.fastq > output.sai
# generate SAM formatted alignment output
bwa samse -n -1 genome.fasta output.sai query.fastq > output.sam
# use samtools to generate bam file
samtools view -bS -o output.bam  output.sam
# sort entries in bam file
samtools sort output.bam output.sorted
# index reads in bam file
samtools index output.sorted.bam
```

Code from: http://manuals.bioinformatics.ucr.edu/home/ht-seq

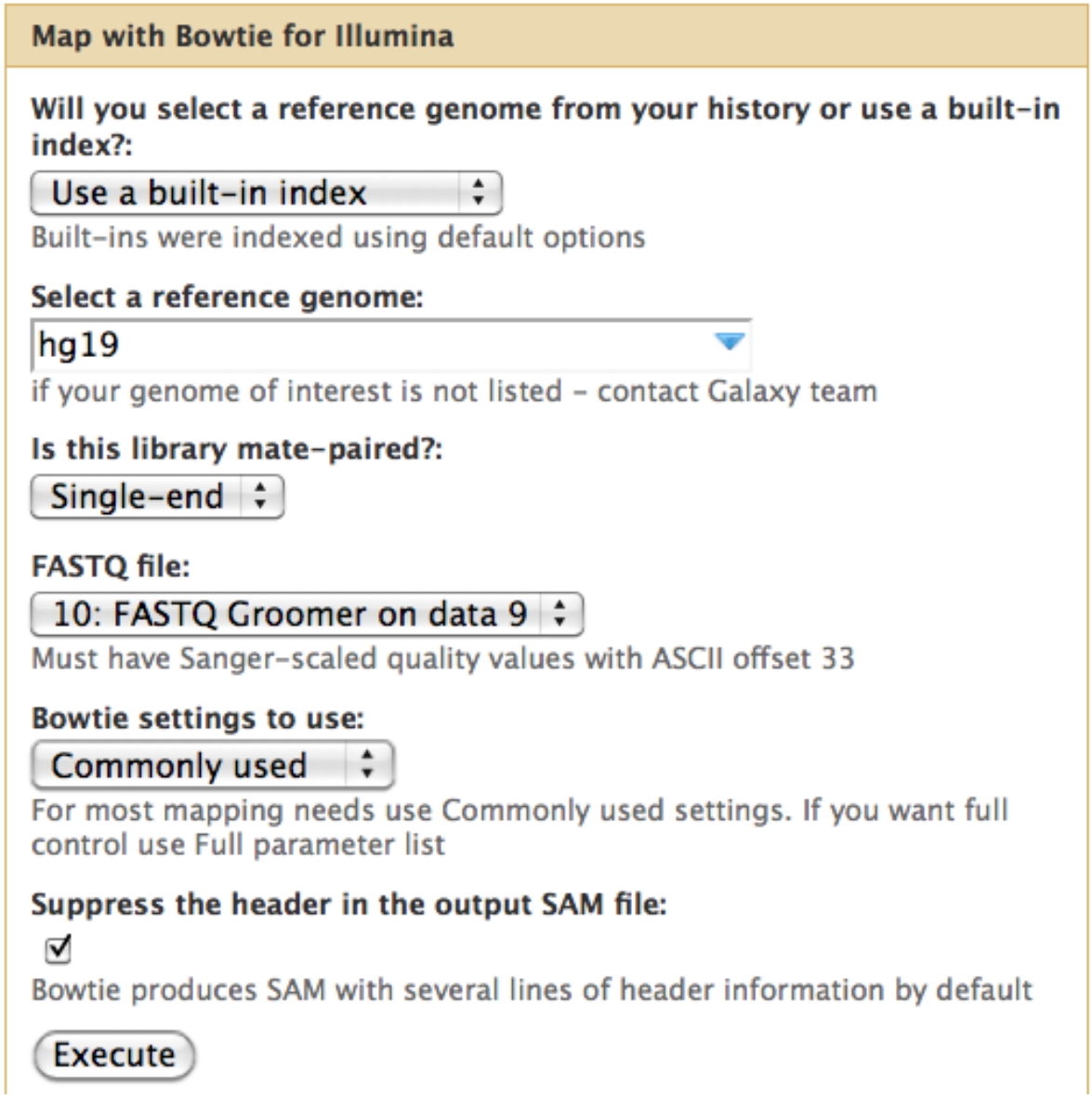
cc BY NC SA

Alignment using Galaxy

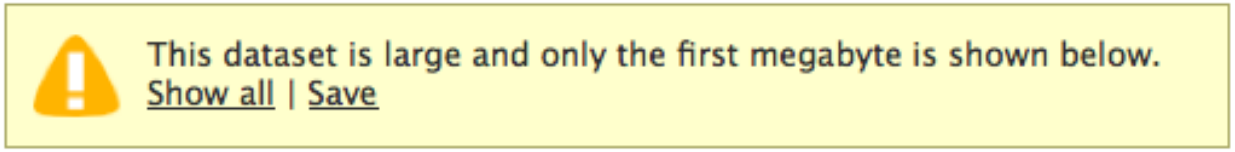




CC BY NC SA


Map with Bowtie for Illumina
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 Galaxy team
Is this library mate-paired?:
Single-end
FASTQ file:
10: FASTQ Groomer on data 9
Must have Sanger-scaled quality values with ASCII offset 33
Bowtie settings to use:
Commonly used
For most mapping needs use Commonly used settings. If you want full control use Full parameter list
Suppress the header in the output SAM file:
Bowtie produces SAM with several lines of header information by default
Execute

cc BY NC SA



ERR000494.7831411/1	16	chr1	161100152	255	41M	*
ERR000494.8751557/1	16	chr1	161100143	255	41M	*
ERR000492.3331028/1	16	chr1	161100168	255	41M	*
ERR000492.9521233/1	16	chr1	161100122	255	41M	*
ERR000495.9873465/1	0	chr1	161100019	255	41M	*
ERR000066.3373915/1	0	chr1	161100029	255	41M	*
ERR000345.1515240/1	0	chr1	161100102	255	41M	*
ERR000493.8919447/1	0	chr1	161100287	255	41M	*
ERR000319.9062286/1	0	chr1	161100196	255	45M	*
ERR000497.3212156/1	16	chr1	161100336	255	41M	*
ERR000496.4165020/1	16	chr1	161100373	255	41M	*
ERR000494.7623510/1	0	chr1	161100301	255	41M	*
ERR000345.855241/1	16	chr1	161100365	255	41M	*
ERR000103.3112982/1	0	chr1	161100128	255	51M	*
ERR000494.8355511/1	16	chr1	161100494	255	41M	*
ERR000319.1861126/1	0	chr1	161100274	255	45M	*
ERR000497.2982726/1	16	chr1	161100521	255	41M	*
ERR000066.2996155/1	0	chr1	161100442	255	41M	*
ERR000496.7289044/1	0	chr1	161100421	255	41M	*
ERR000491.4951810/1	0	chr1	161100502	255	41M	*
ERR000066.5352541/1	16	chr1	161100560	255	41M	*
ERR000491.5216042/1	0	chr1	161100553	255	41M	*
ERR000495.3611499/1	0	chr1	161100553	255	41M	*
ERR000103.216589/1	16	chr1	161100543	255	51M	*
ERR000066.3726692/1	0	chr1	161100639	255	41M	*
ERR000491.1427824/1	16	chr1	161100637	255	41M	*
ERR000493.8067934/1	16	chr1	161100763	255	41M	*
ERR000492.6208053/1	16	chr1	161100796	255	41M	*
ERR000494.5036239/1	16	chr1	161100807	255	41M	*
ERR000066.2768591/1	16	chr1	161100826	255	41M	*
ERR000491.4175381/1	16	chr1	161100871	255	41M	*
ERR000495.1300023/1	16	chr1	161100869	255	41M	*
ERR000497.8598006/1	16	chr1	161100915	255	41M	*
ERR000103.2165391/1	16	chr1	161100929	255	51M	*
ERR000491.5416160/1	16	chr1	161101029	255	41M	*
ERR000103.2567170/1	16	chr1	161101074	255	51M	*
ERR000066.3574991/1	0	chr1	161100976	255	41M	*
ERR000493.9351568/1	0	chr1	161100968	255	41M	*
ERR000319.6698579/1	16	chr1	161101093	255	45M	*

CC BY NC SA

Visualization

- Many (many!) genome browsers available:
  - UCSC Genome Browser
  - Ensembl
  - Gbrowse
  - 1000 Genomes Browser
  - Integrative Genomics Viewer (IGV)
  - …

cc BY NC SA

http://www.broadinstitute.org/igv

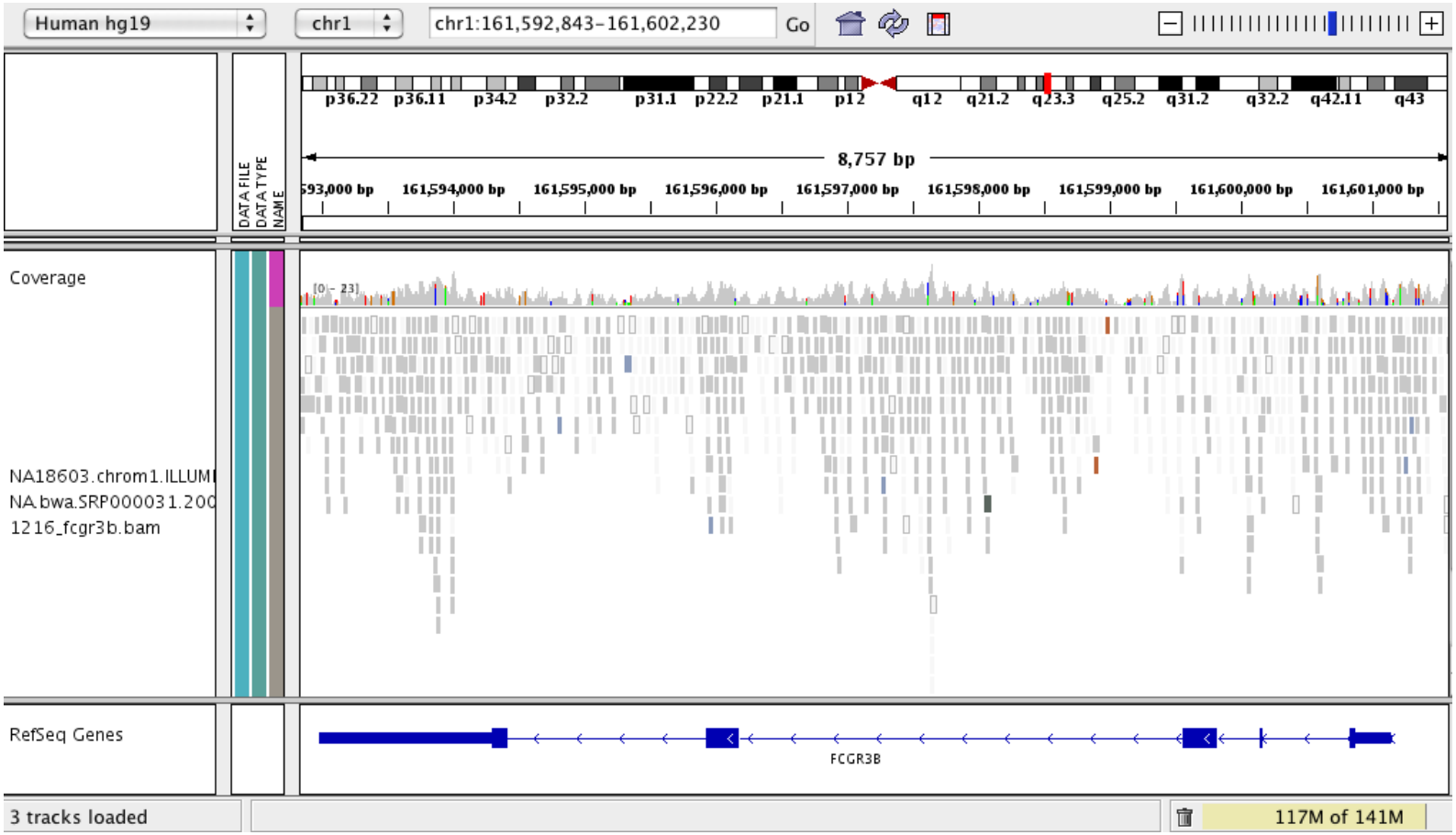
Visualization: IGV

- Developed at the Broad Institute (MIT)
- Wide variety of data types:
  - Sequence alignments
  - Microarrays (SNP, CNV, expression…)
  - Genomic annotations

cc BY NC SA

http://www.broadinstitute.org/igv

IGV with NGS data



CC BY NC SA

http://www.broadinstitute.org/igv

IGV with NGS data



CC BY NC SA

SeqMonk



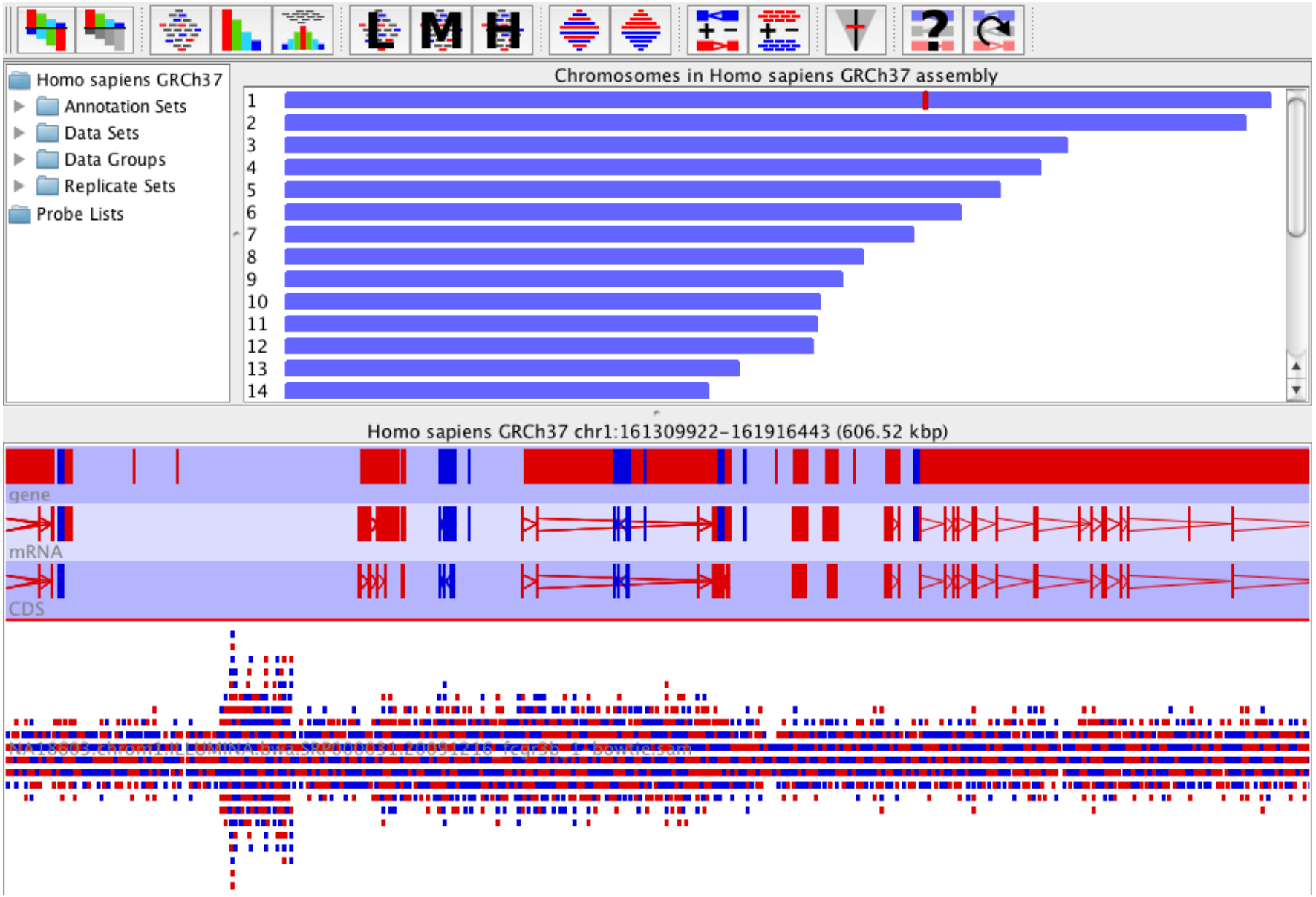
SeqMonk

Function	A tool to visualise and analyse high throughput mapped sequence data
Language	Java
Requirements	A suitable Java Runtime Environment
Code Maturity	Stable. This code is being widely used for real work, although SeqMonk is still under active development.
Code Released	Yes, under GPL v3 or later.
Initial Contact	Simon Andrews

Download Now

http://www.bioinformatics.bbsrc.ac.uk/projects/seqmonk/

cc BY NC SA


L
M
H
Homo sapiens GRCh37
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
Chromosomes in Homo sapiens GRCh37 assembly
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Homo sapiens GRCh37 chr1:161309922-161916443 (606.52 kbp)
gene
mRNA
CDS

CC BY NC SA

RNA-seq analysis

- RNA-seq is rapidly gaining ground on microarray technology in terms of popularity:
  - Sequence and align RNA fragments
  - Generate counts for genes/exons/regions
  - Perform comparative analysis (e.g., differential expression)
- Some pitfalls: e.g., transcript length bias…

CC BY NC SA

RNA-seq walkthrough

- View aligned data in SeqMonk
- Generate counts for each gene
- Export and perform differential expression analysis via limma in GenePattern

CC BY NC SA

OPEN ACCESS Freely available online PLoS GENETICS

Novel Low Abundance and Transient RNAs in Yeast Revealed by Tiling Microarrays and Ultra High–Throughput Sequencing Are Not Conserved Across Closely Related Yeast Species

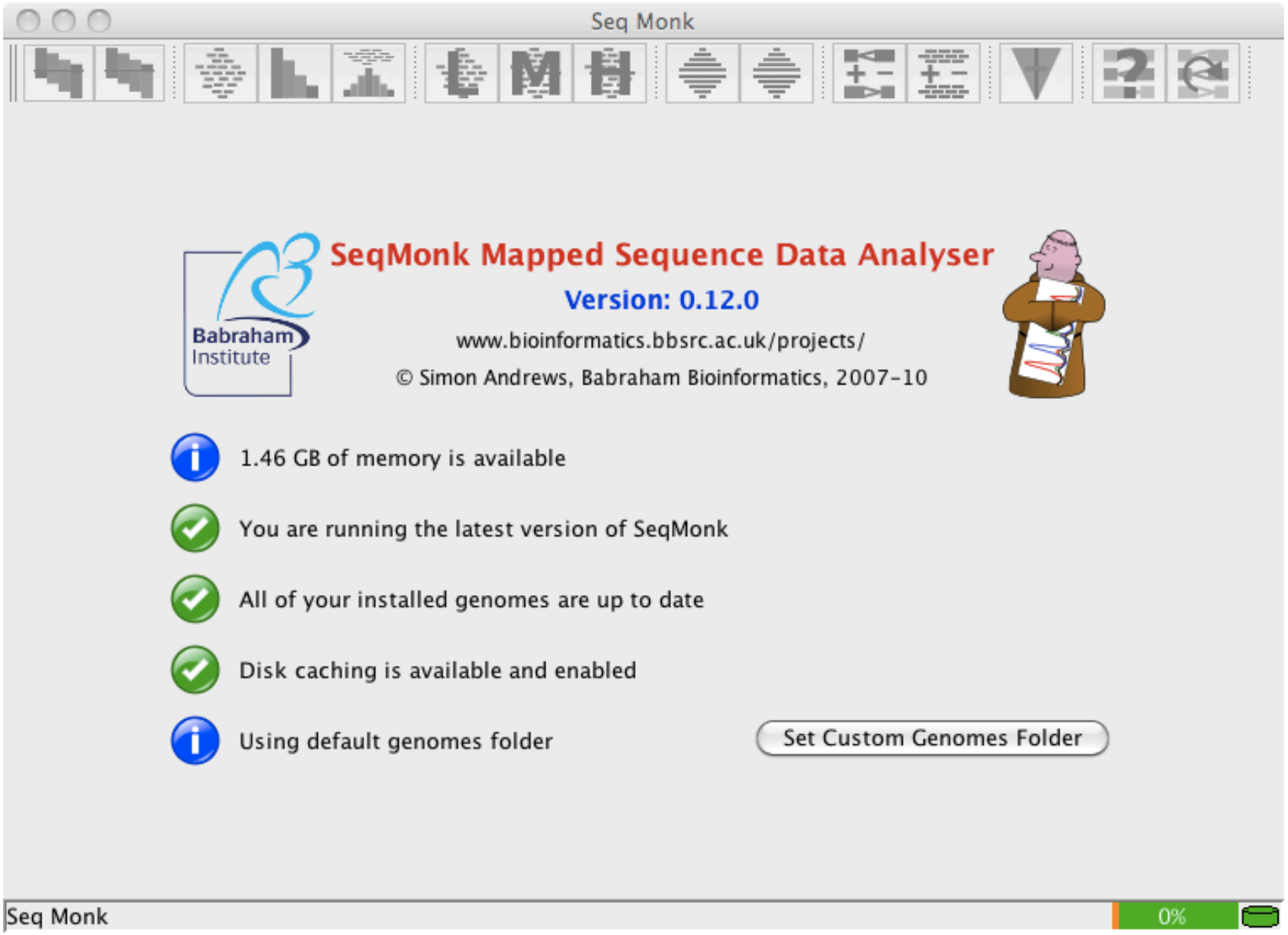
Albert Lee[1☯], Kasper Daniel Hansen[2☯], James Bullard[2☯], Sandrine Dudoit[2], Gavin Sherlock[1*]

1 Department of Genetics, Stanford University, Stanford, California, United of States of America, 2 Division of Biostatistics, School of Public Health, University of California Berkeley, Berkeley, California, United States of America

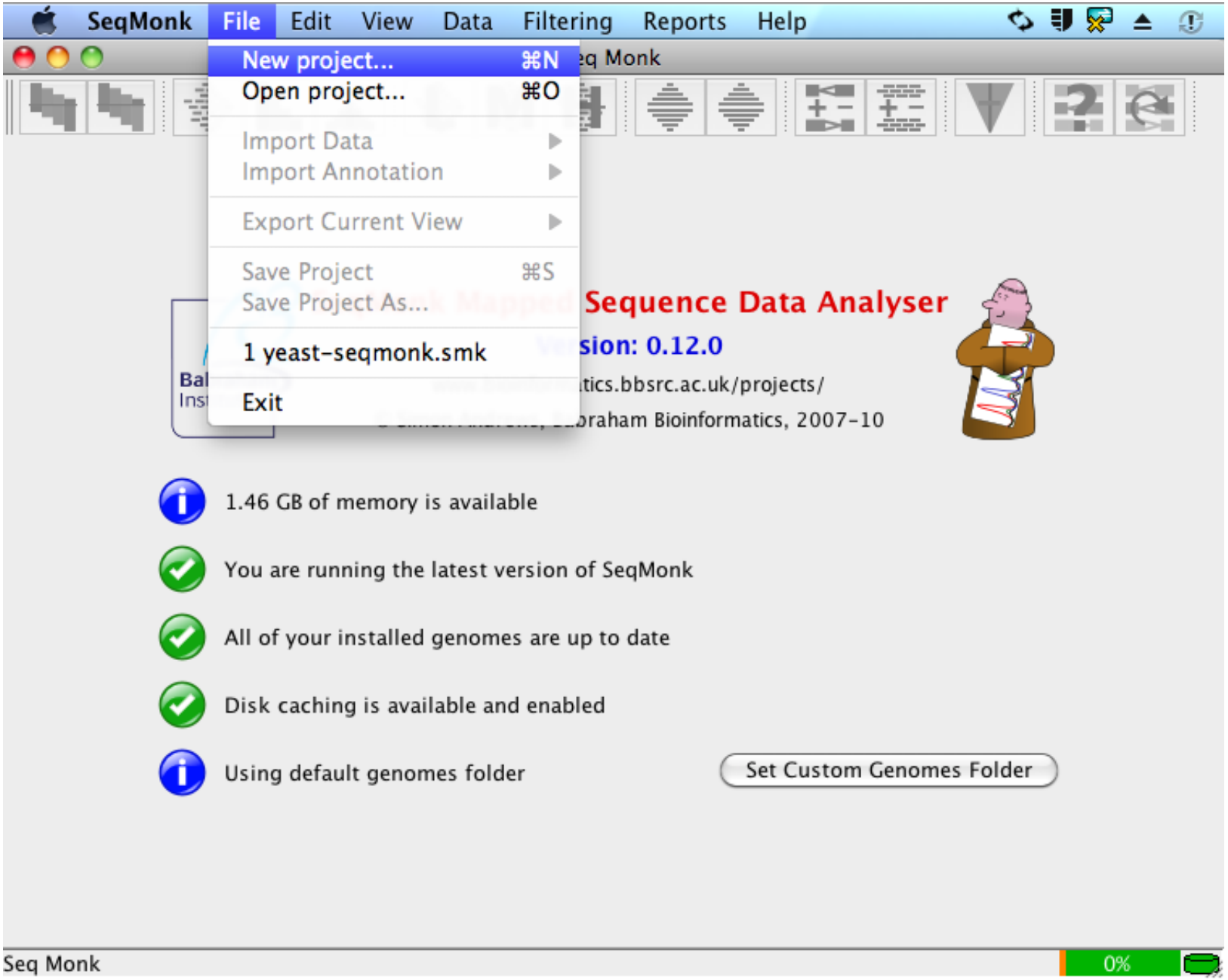
PLoS Genetics | www.plosgenetics.org December 2008 | Volume 4 | Issue 12 | e1000299

- Data obtained from yeast RNA-seq experiment
- Wild-type *versus* RNA degradation mutants
- Subset of data (chromosome 1)
- Six samples (3 WT / 3 MT)

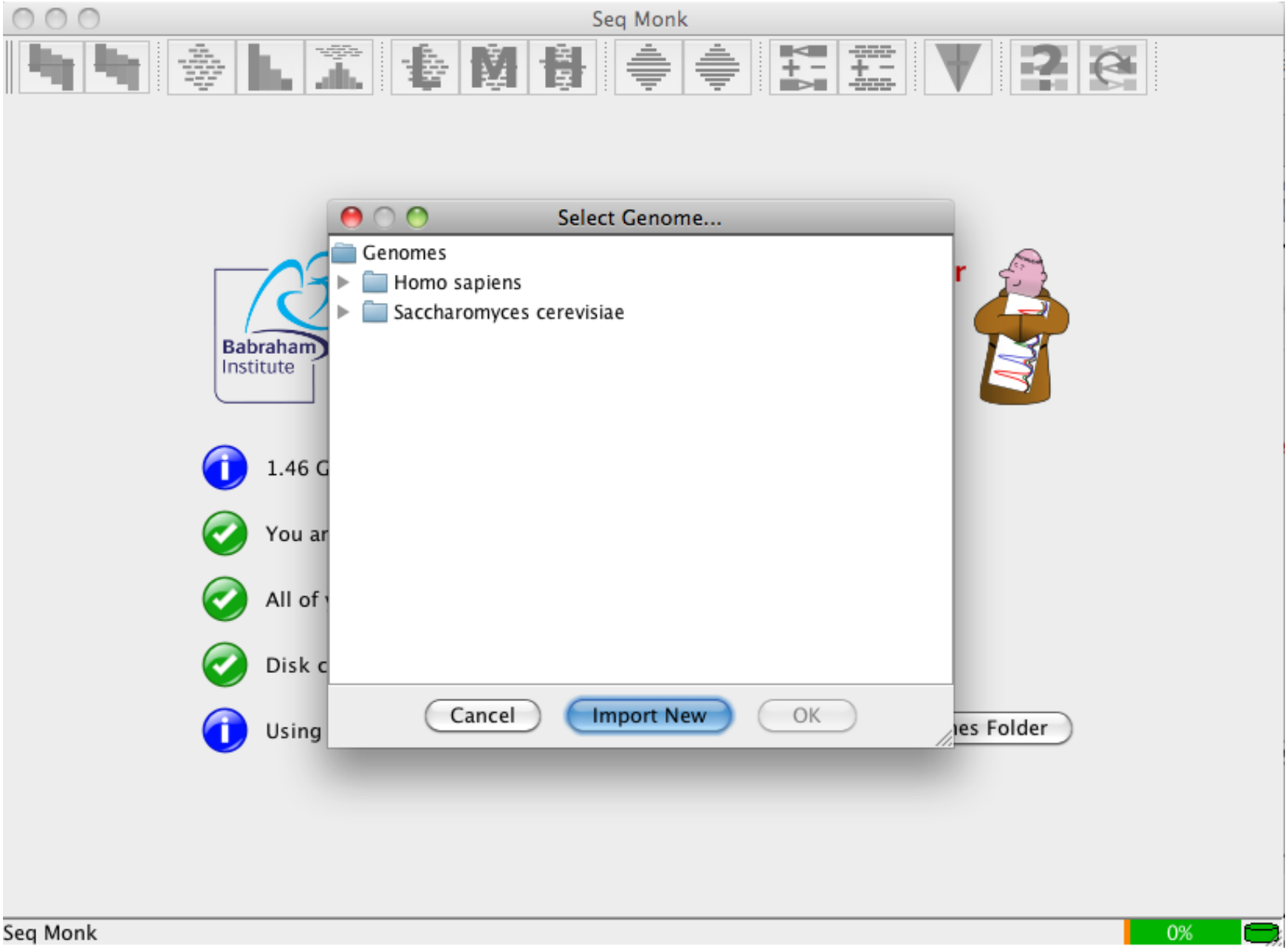
CC BY NC SA


Seq Monk
SeqMonk Mapped Sequence Data Analyser
Version: 0.12.0
www.bioinformatics.bbsrc.ac.uk/projects/
© Simon Andrews, Babraham Bioinformatics, 2007-10
Babraham Institute
1.46 GB of memory is available
You are running the latest version of SeqMonk
All of your installed genomes are up to date
Disk caching is available and enabled
Using default genomes folder
Set Custom Genomes Folder
Seq Monk
0%

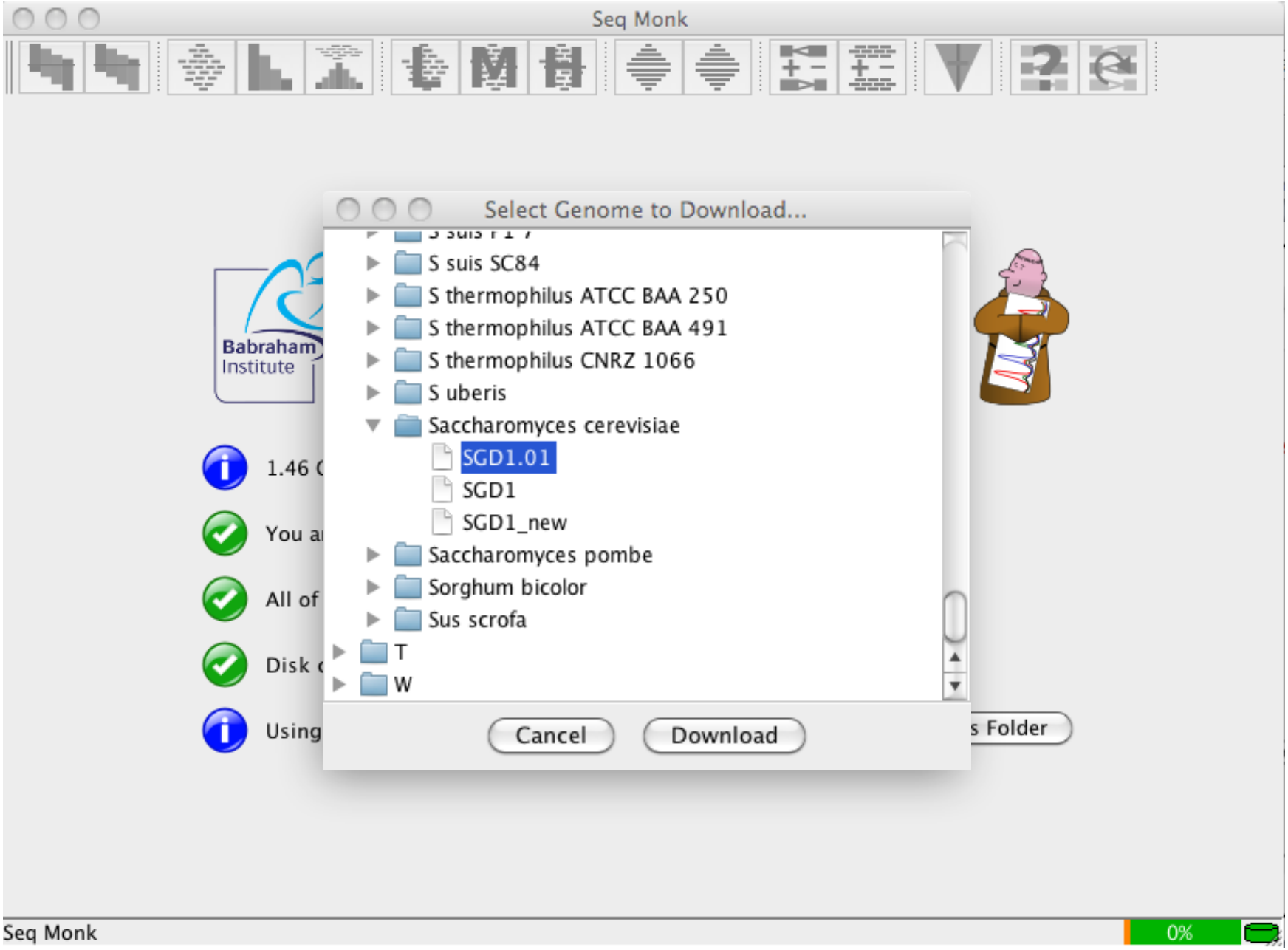
cc BY NC SA


SeqMonk
File
Edit
View
Data
Filtering
Reports
Help
New project... ⌘N
Open project... ⌘O
Import Data
Import Annotation
Export Current View
Save Project ⌘S
Save Project As...
1 yeast-seqmonk.smk
Exit
Sequence Data Analyser
sion: 0.12.0
tics.bbsrc.ac.uk/projects/
braham Bioinformatics, 2007-10
1.46 GB of memory is available
You are running the latest version of SeqMonk
All of your installed genomes are up to date
Disk caching is available and enabled
Using default genomes folder
Set Custom Genomes Folder
Seq Monk
0%

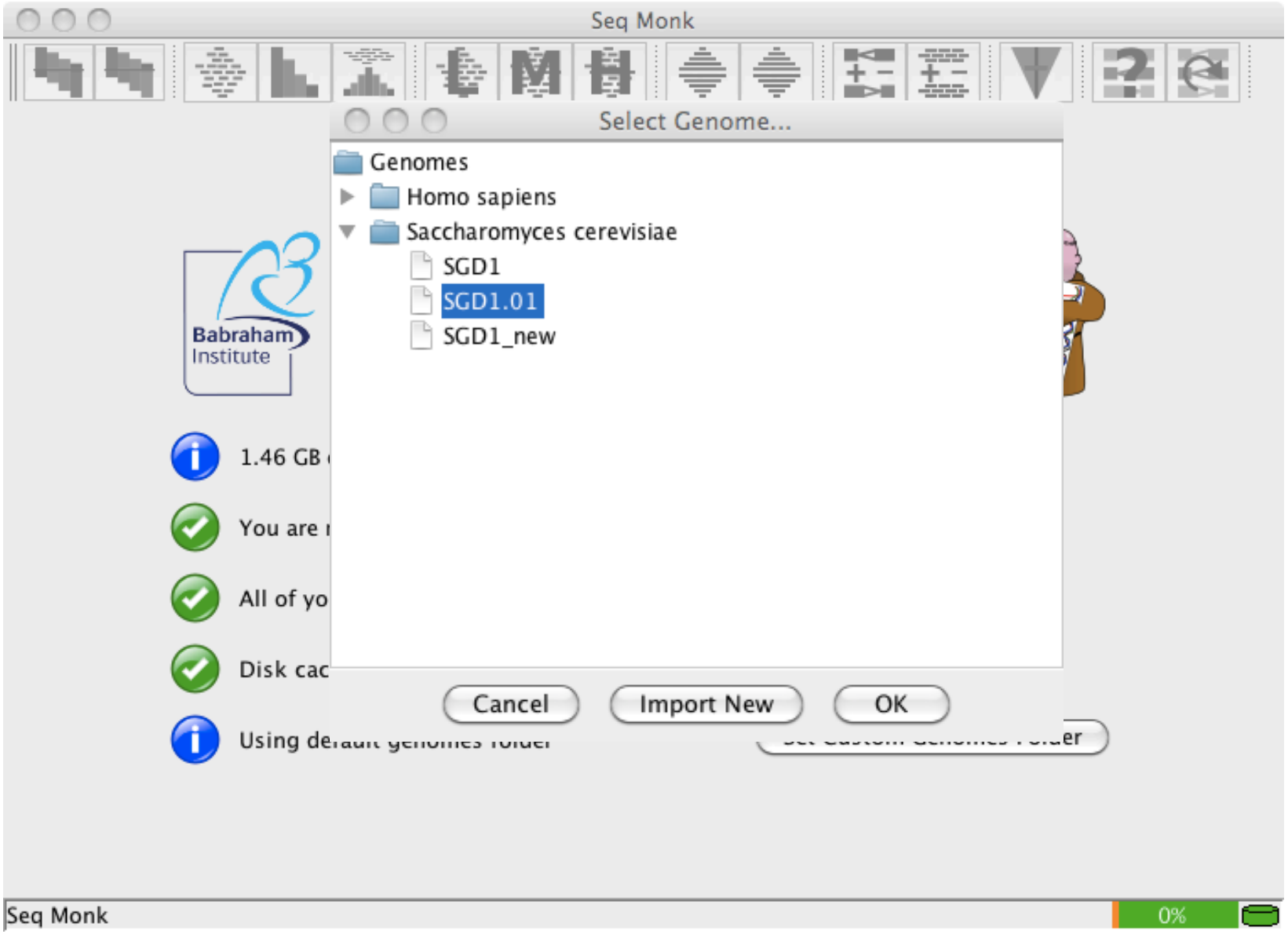
CC BY NC SA


Seq Monk
Select Genome...
Genomes
Homo sapiens
Saccharomyces cerevisiae
Cancel
Import New
OK
Babraham Institute
Seq Monk
0%

CC BY NC SA


Seq Monk
Select Genome to Download...
S suis SC84
S thermophilus ATCC BAA 250
S thermophilus ATCC BAA 491
S thermophilus CNRZ 1066
S uberis
Saccharomyces cerevisiae
SGD1.01
SGD1
SGD1_new
Saccharomyces pombe
Sorghum bicolor
Sus scrofa
T
W
Cancel
Download
Babraham Institute
Seq Monk
0%

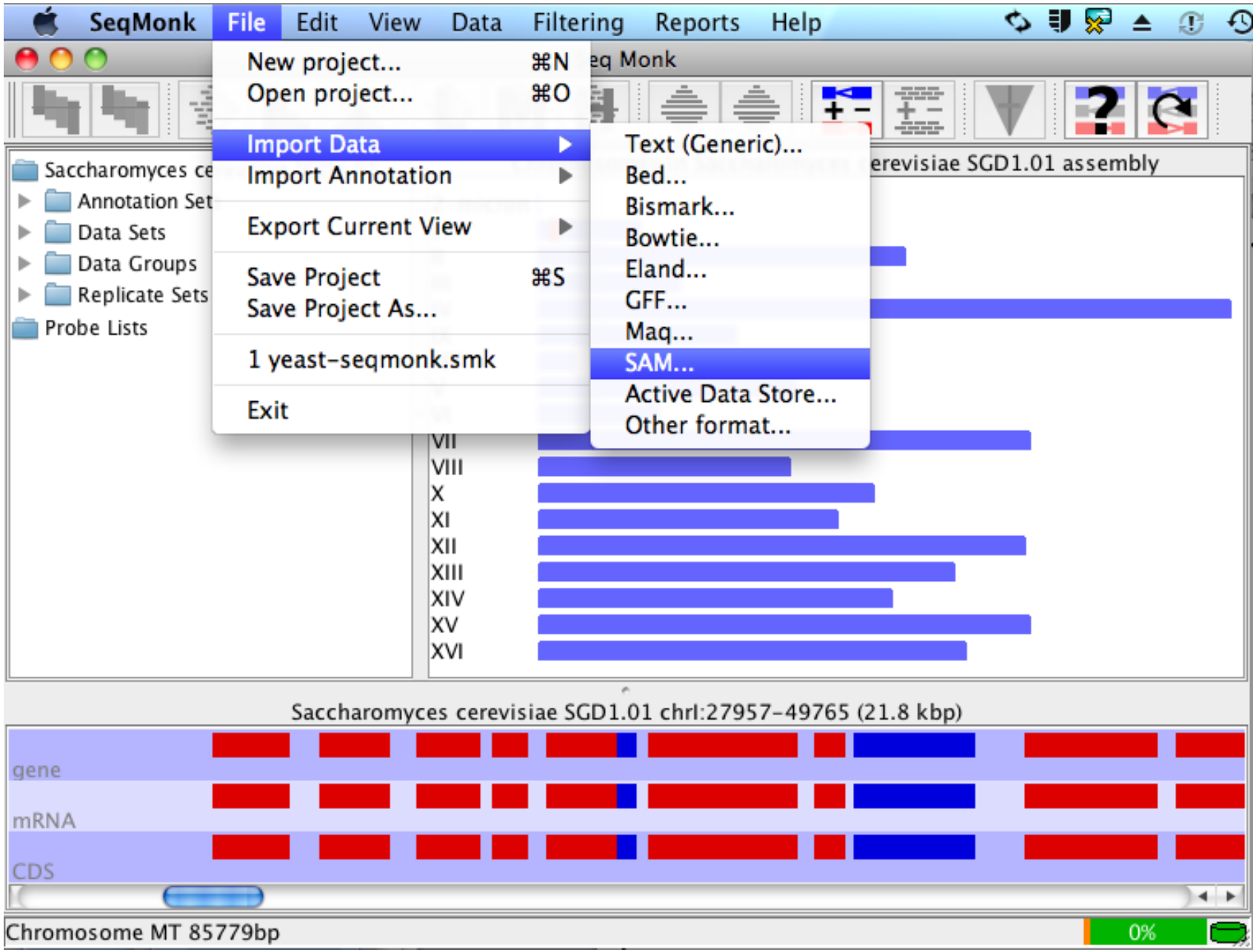
CC BY NC SA


Seq Monk
Select Genome...
Genomes
Homo sapiens
Saccharomyces cerevisiae
SGD1
SGD1.01
SGD1_new
Babraham Institute
1.46 GB
You are
All of yo
Disk cac
Using de
Cancel
Import New
OK
Seq Monk
0%

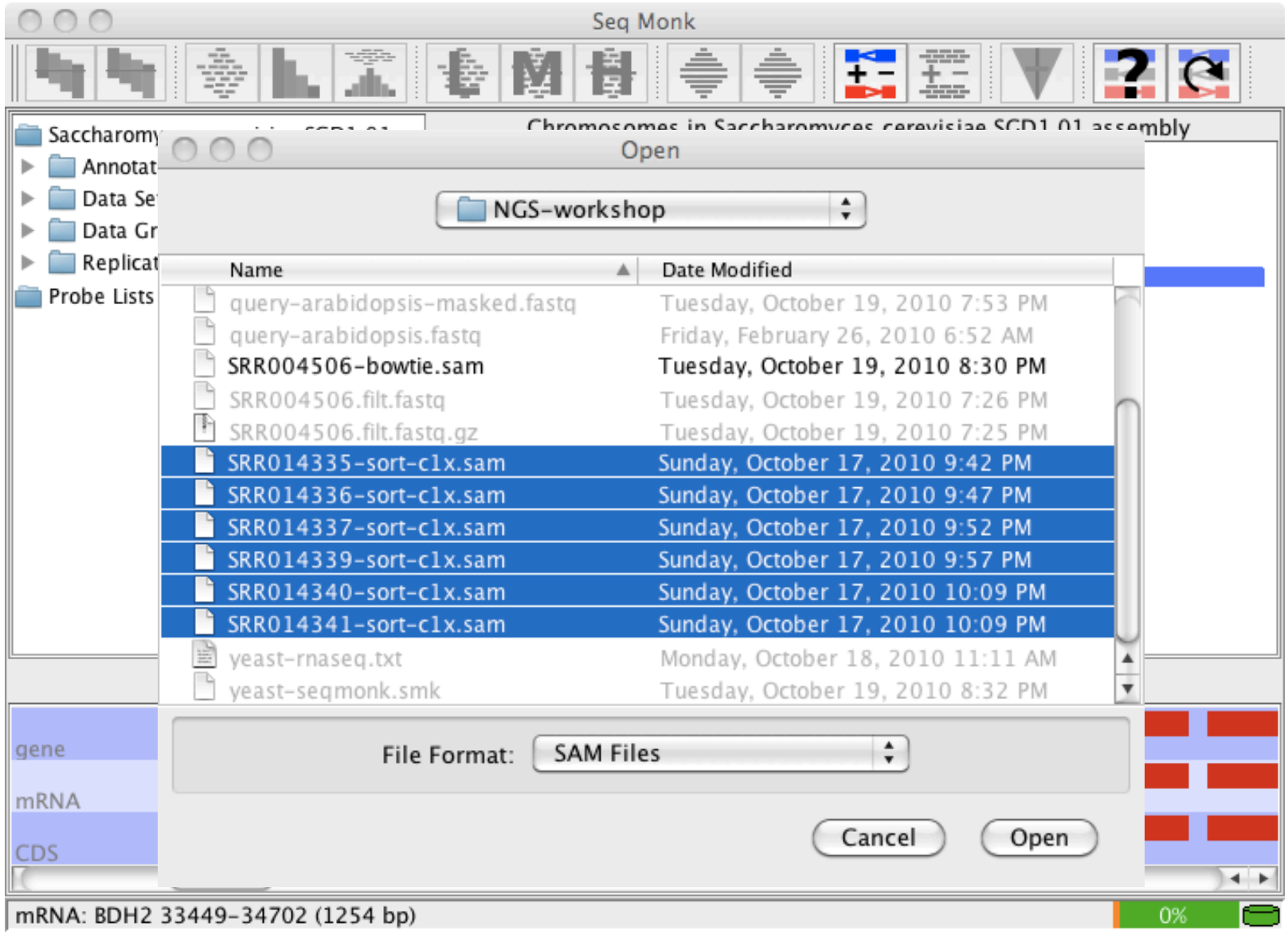
CC BY NC SA


Seq Monk
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
Chromosomes in Saccharomyces cerevisiae SGD1.01 assembly
2-micron
I
II
III
IV
IX
MT
V
VI
VII
VIII
X
XI
XII
XIII
XIV
XV
XVI
Saccharomyces cerevisiae SGD1.01 chrI:27957-49765 (21.8 kbp)
gene
mRNA
mRNA:BDH2
CDS
mRNA: BDH2 33449-34702 (1254 bp)
0%

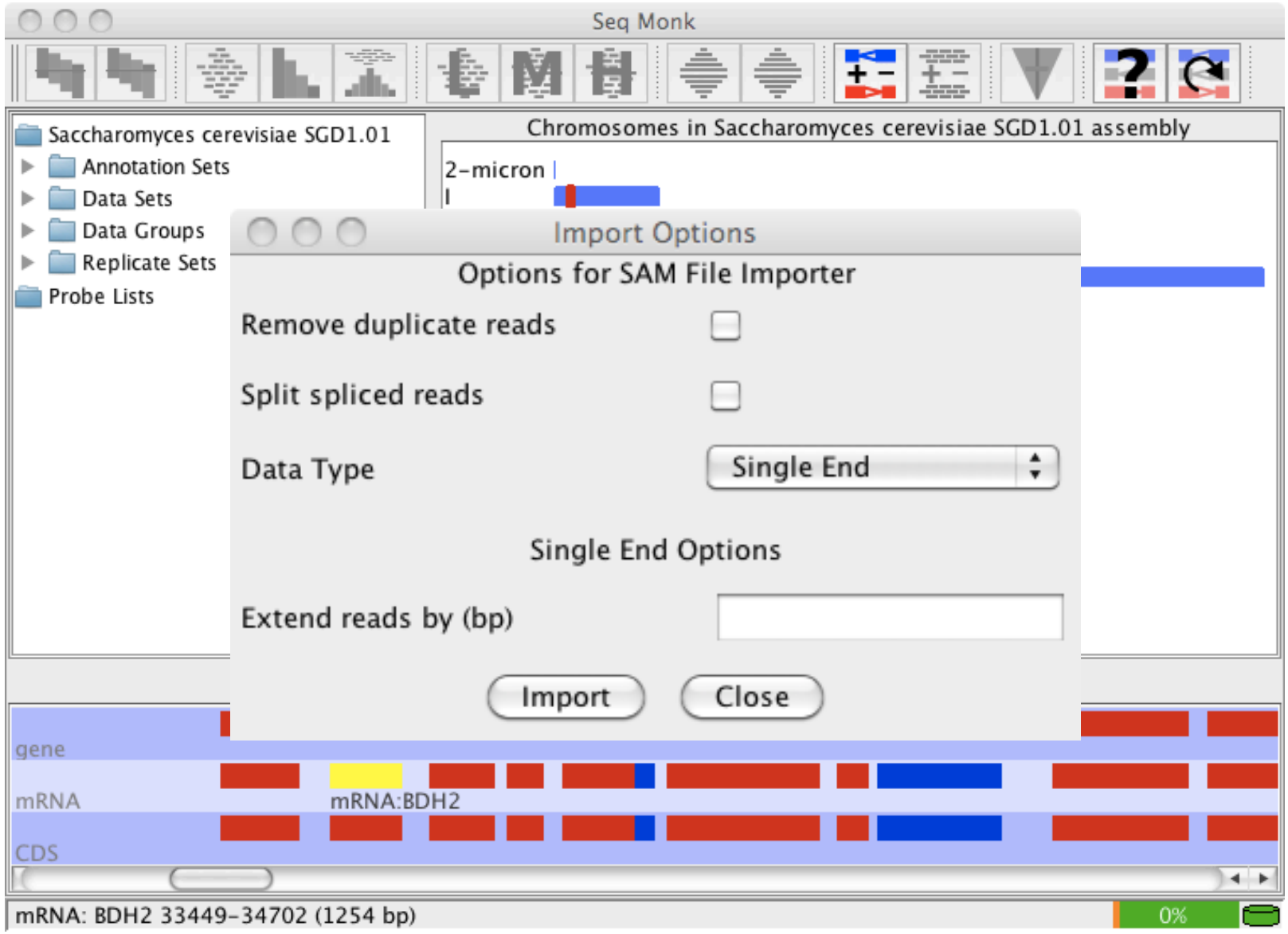
CC BY NC SA


SeqMonk
File
Edit
View
Data
Filtering
Reports
Help
New project... ⌘N
Open project... ⌘O
Import Data
Import Annotation
Export Current View
Save Project ⌘S
Save Project As...
1 yeast-seqmonk.smk
Exit
Text (Generic)...
Bed...
Bismark...
Bowtie...
Eland...
GFF...
Maq...
SAM...
Active Data Store...
Other format...
Saccharomyces ce
Annotation Set
Data Sets
Data Groups
Replicate Sets
Probe Lists
VII
VIII
X
XI
XII
XIII
XIV
XV
XVI
Saccharomyces cerevisiae SGD1.01 chrI:27957-49765 (21.8 kbp)
gene
mRNA
CDS
Chromosome MT 85779bp
0%

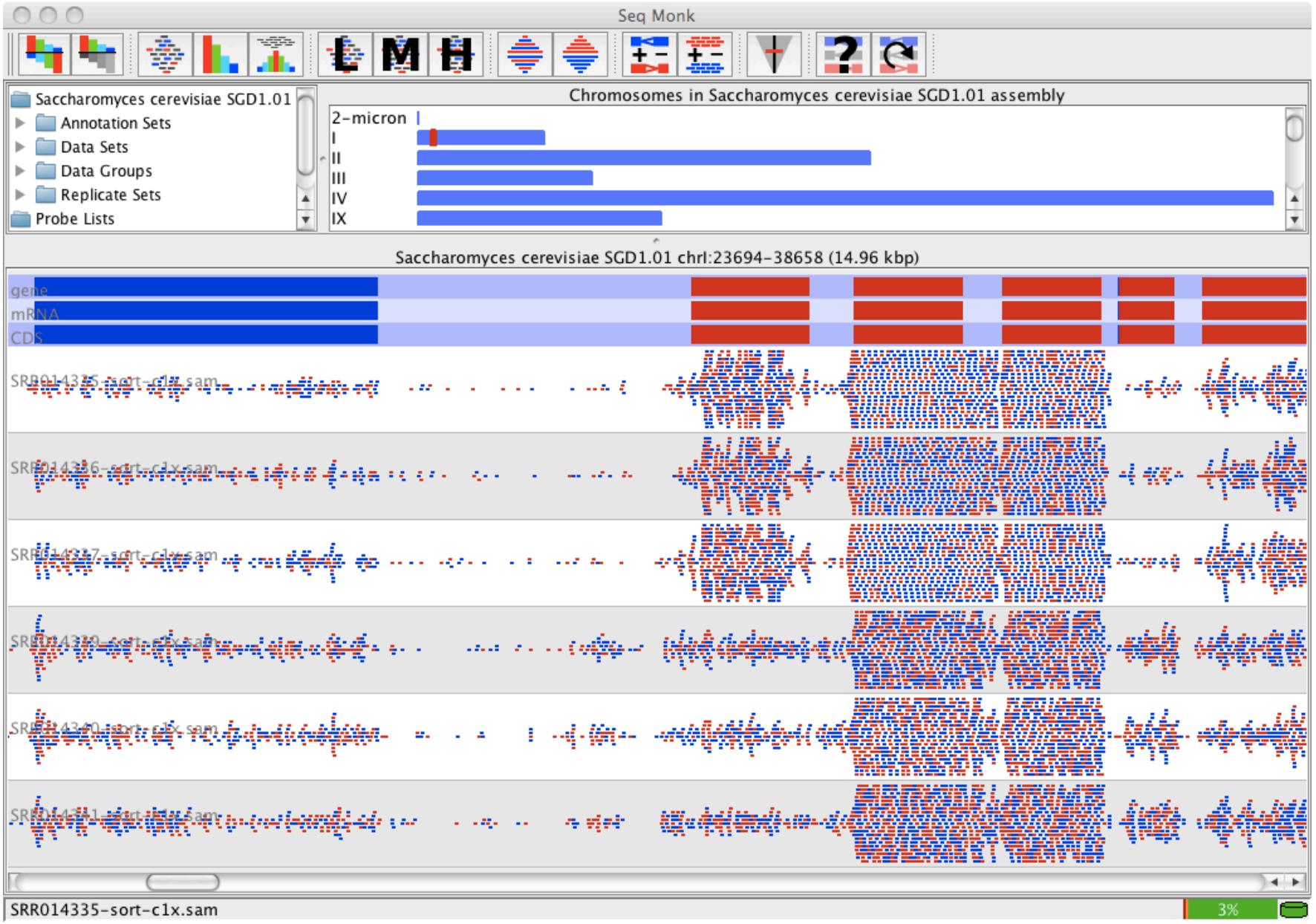
cc BY NC SA


Seq Monk
Open
NGS-workshop
Name
Date Modified
query-arabidopsis-masked.fastq
Tuesday, October 19, 2010 7:53 PM
query-arabidopsis.fastq
Friday, February 26, 2010 6:52 AM
SRR004506-bowtie.sam
Tuesday, October 19, 2010 8:30 PM
SRR004506.filt.fastq
Tuesday, October 19, 2010 7:26 PM
SRR004506.filt.fastq.gz
Tuesday, October 19, 2010 7:25 PM
SRR014335-sort-c1x.sam
Sunday, October 17, 2010 9:42 PM
SRR014336-sort-c1x.sam
Sunday, October 17, 2010 9:47 PM
SRR014337-sort-c1x.sam
Sunday, October 17, 2010 9:52 PM
SRR014339-sort-c1x.sam
Sunday, October 17, 2010 9:57 PM
SRR014340-sort-c1x.sam
Sunday, October 17, 2010 10:09 PM
SRR014341-sort-c1x.sam
Sunday, October 17, 2010 10:09 PM
yeast-rnaseq.txt
Monday, October 18, 2010 11:11 AM
yeast-seqmonk.smk
Tuesday, October 19, 2010 8:32 PM
File Format:
SAM Files
Cancel
Open
Probe Lists
gene
mRNA
CDS
mRNA: BDH2 33449-34702 (1254 bp)
0%

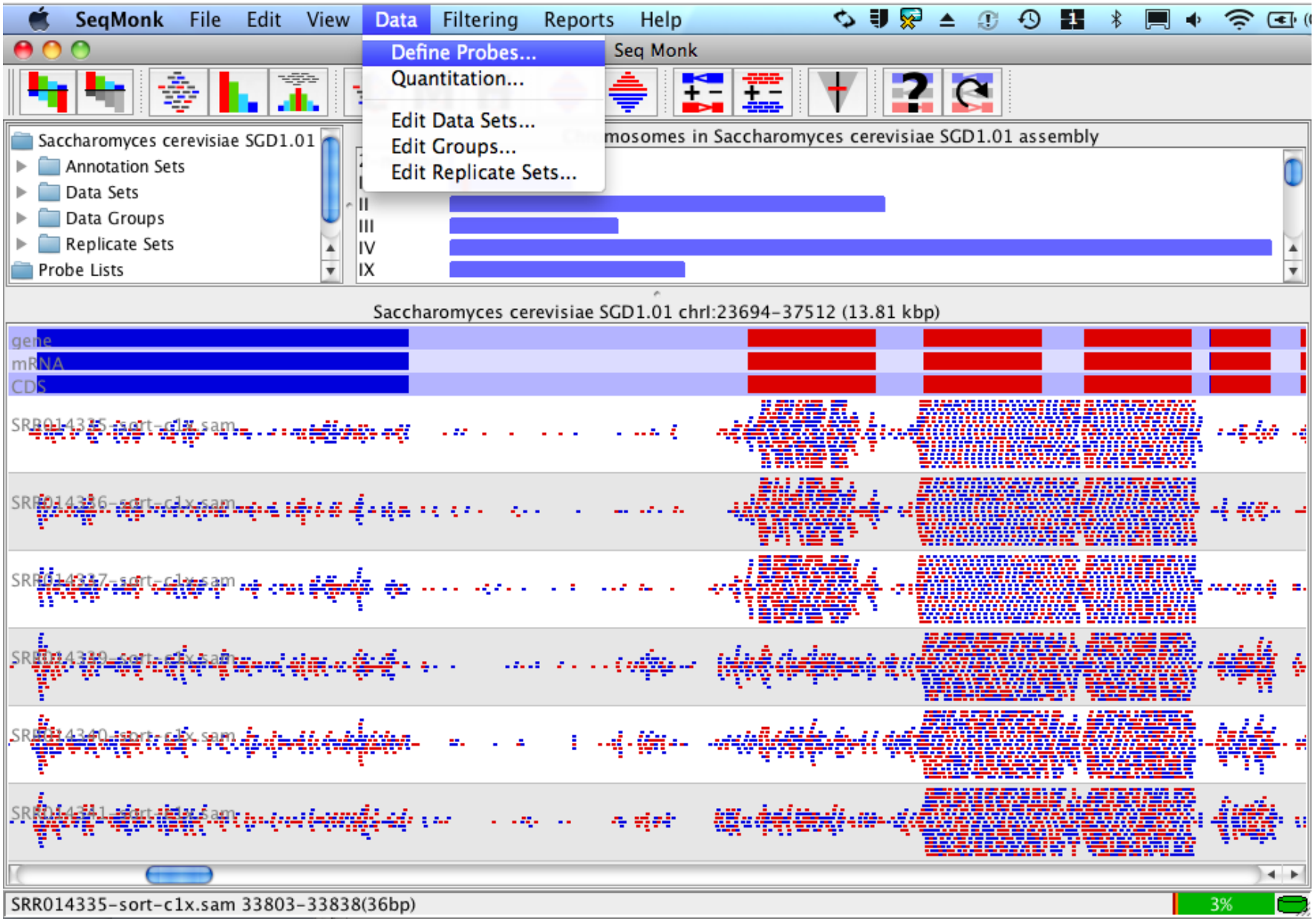
cc BY NC SA


Seq Monk
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
Chromosomes in Saccharomyces cerevisiae SGD1.01 assembly
2-micron
Import Options
Options for SAM File Importer
Remove duplicate reads
Split spliced reads
Data Type
Single End
Single End Options
Extend reads by (bp)
Import
Close
gene
mRNA
mRNA:BDH2
CDS
mRNA: BDH2 33449-34702 (1254 bp)
0%

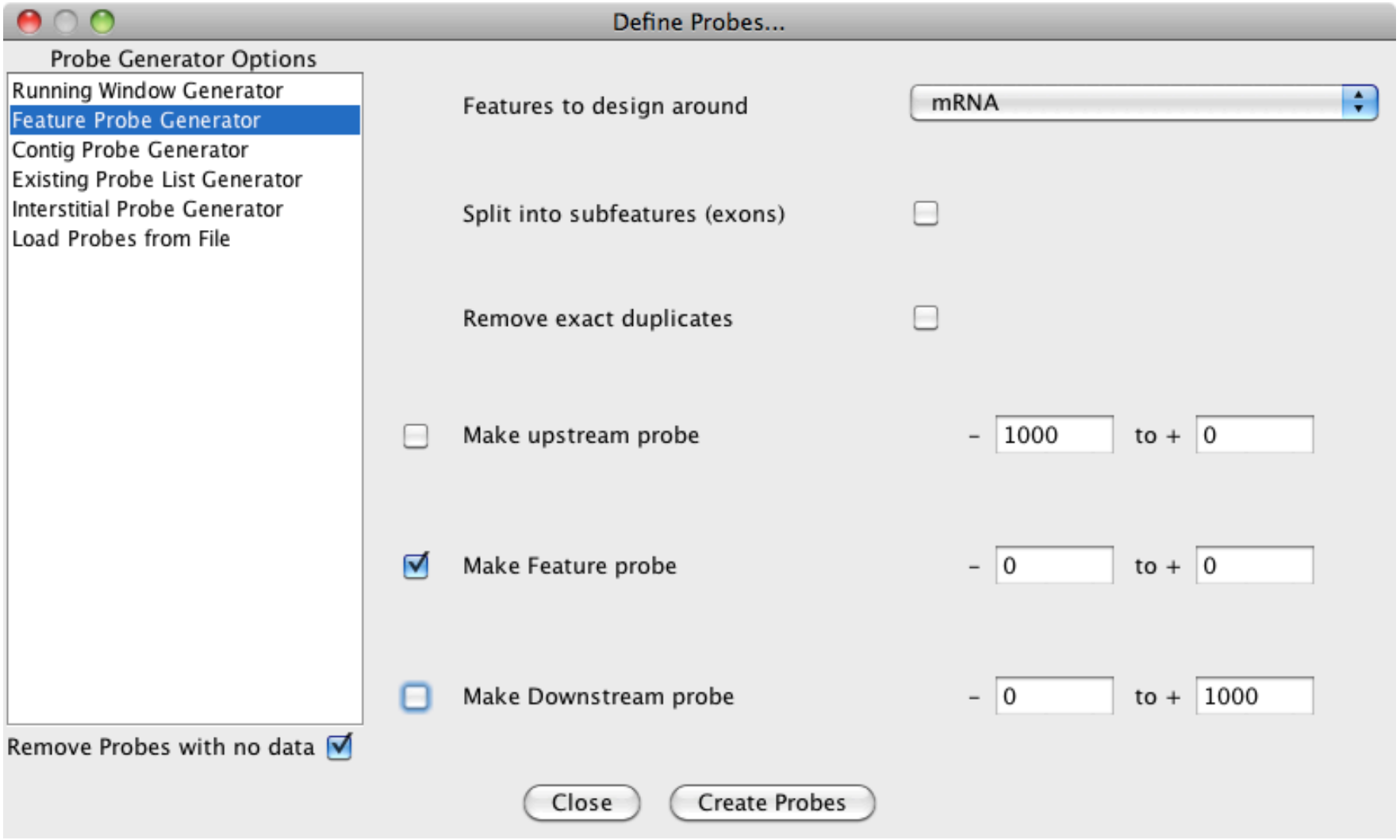
CC BY NC SA


Seq Monk
L
M
H
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
Chromosomes in Saccharomyces cerevisiae SGD1.01 assembly
2-micron
I
II
III
IV
IX
Saccharomyces cerevisiae SGD1.01 chrI:23694-38658 (14.96 kbp)
gene
mRNA
CDS
SRR014335-sort-c1x.sam
SRR014336-sort-c1x.sam
SRR014337-sort-c1x.sam
SRR014340-sort-c1x.sam
SRR014340-sort-c1x.sam
SRR014341-sort-c1x.sam
SRR014335-sort-c1x.sam
3%

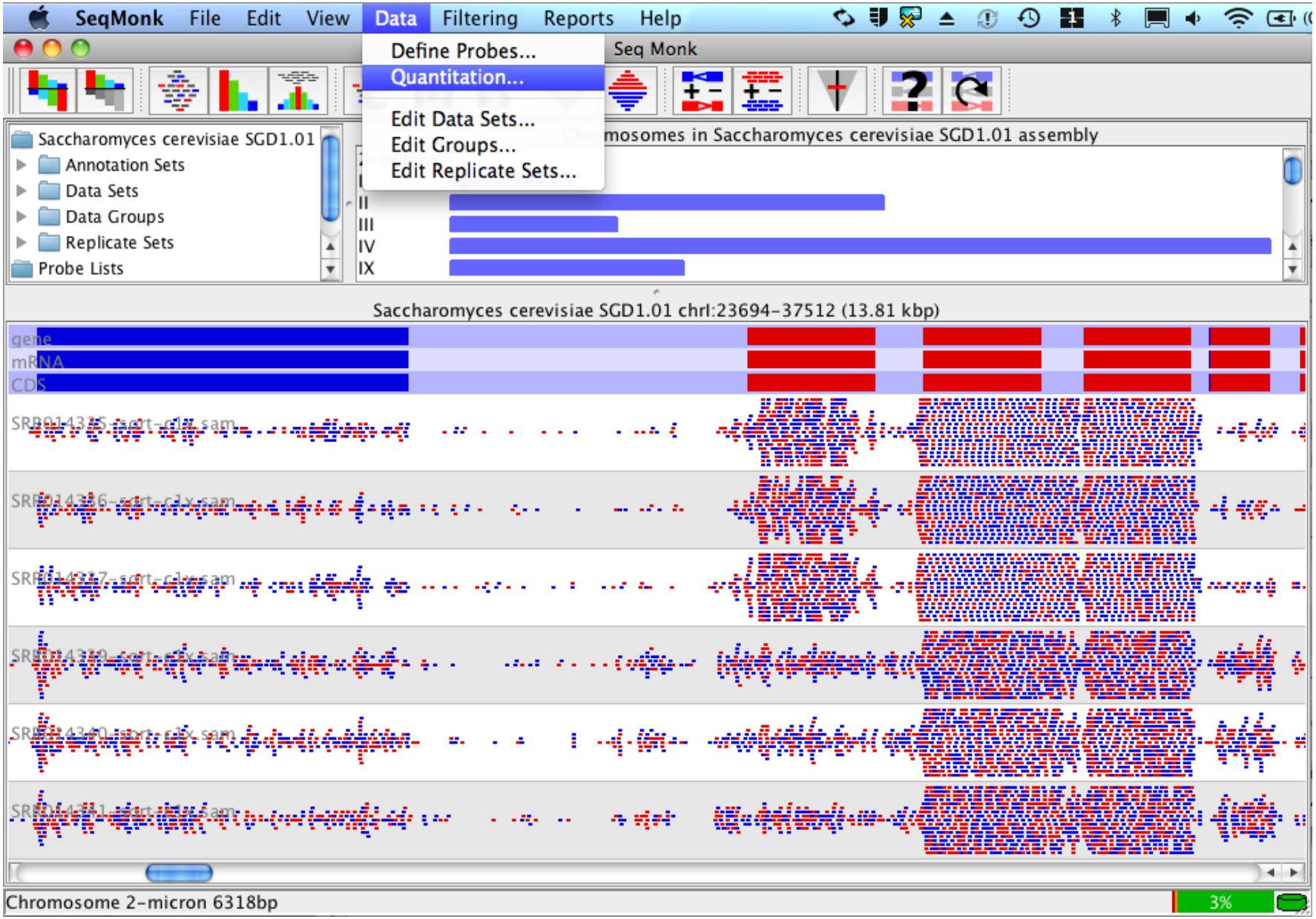
CC BY NC SA


SeqMonk File Edit View Data Filtering Reports Help
Define Probes...
Quantitation...
Edit Data Sets...
Edit Groups...
Edit Replicate Sets...
Seq Monk
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
mosomes in Saccharomyces cerevisiae SGD1.01 assembly
II
III
IV
IX
Saccharomyces cerevisiae SGD1.01 chrI:23694-37512 (13.81 kbp)
gene
mRNA
CDS
SRR014335-sort-c1x.sam 33803-33838(36bp)
3%

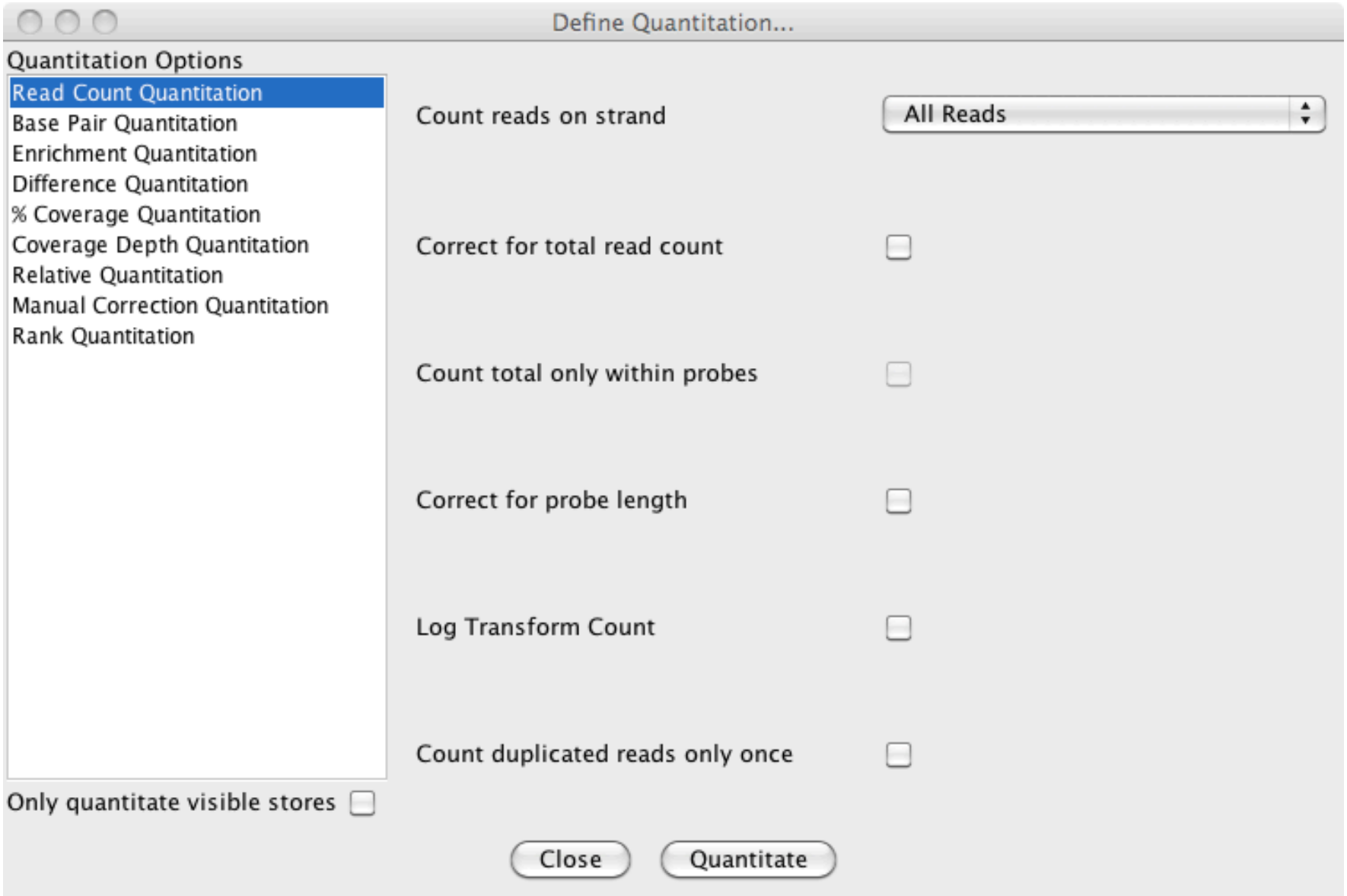
CC BY NC SA


Define Probes...
Probe Generator Options
Running Window Generator
Feature Probe Generator
Contig Probe Generator
Existing Probe List Generator
Interstitial Probe Generator
Load Probes from File
Features to design around
mRNA
Split into subfeatures (exons)
Remove exact duplicates
Make upstream probe
- 1000 to + 0
Make Feature probe
- 0 to + 0
Make Downstream probe
- 0 to + 1000
Remove Probes with no data
Close
Create Probes

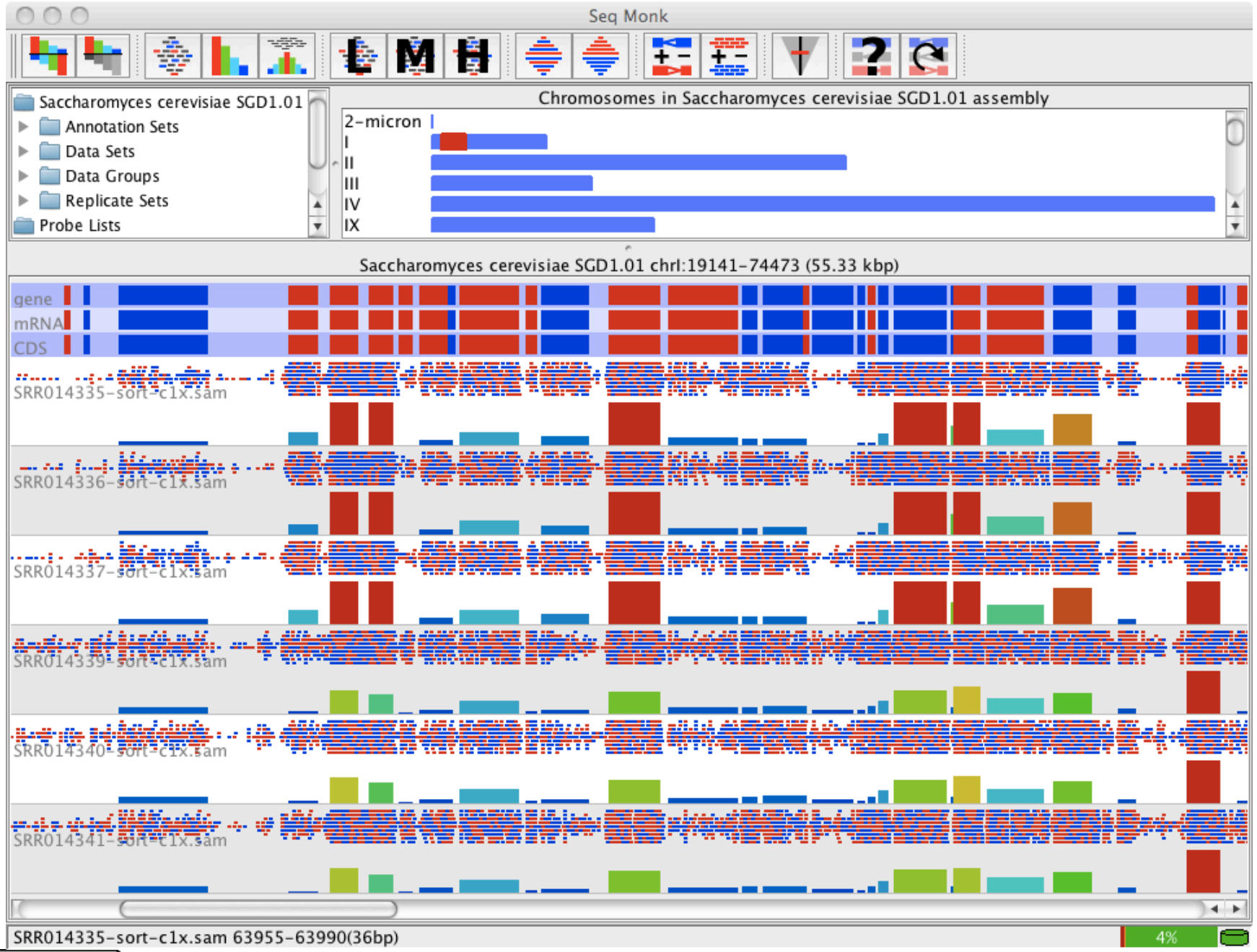
cc BY NC SA


SeqMonk File Edit View Data Filtering Reports Help
Define Probes...
Quantitation...
Edit Data Sets...
Edit Groups...
Edit Replicate Sets...
Seq Monk
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
mosomes in Saccharomyces cerevisiae SGD1.01 assembly
II
III
IV
IX
Saccharomyces cerevisiae SGD1.01 chrI:23694-37512 (13.81 kbp)
gene
mRNA
CDS
Chromosome 2-micron 6318bp
3%

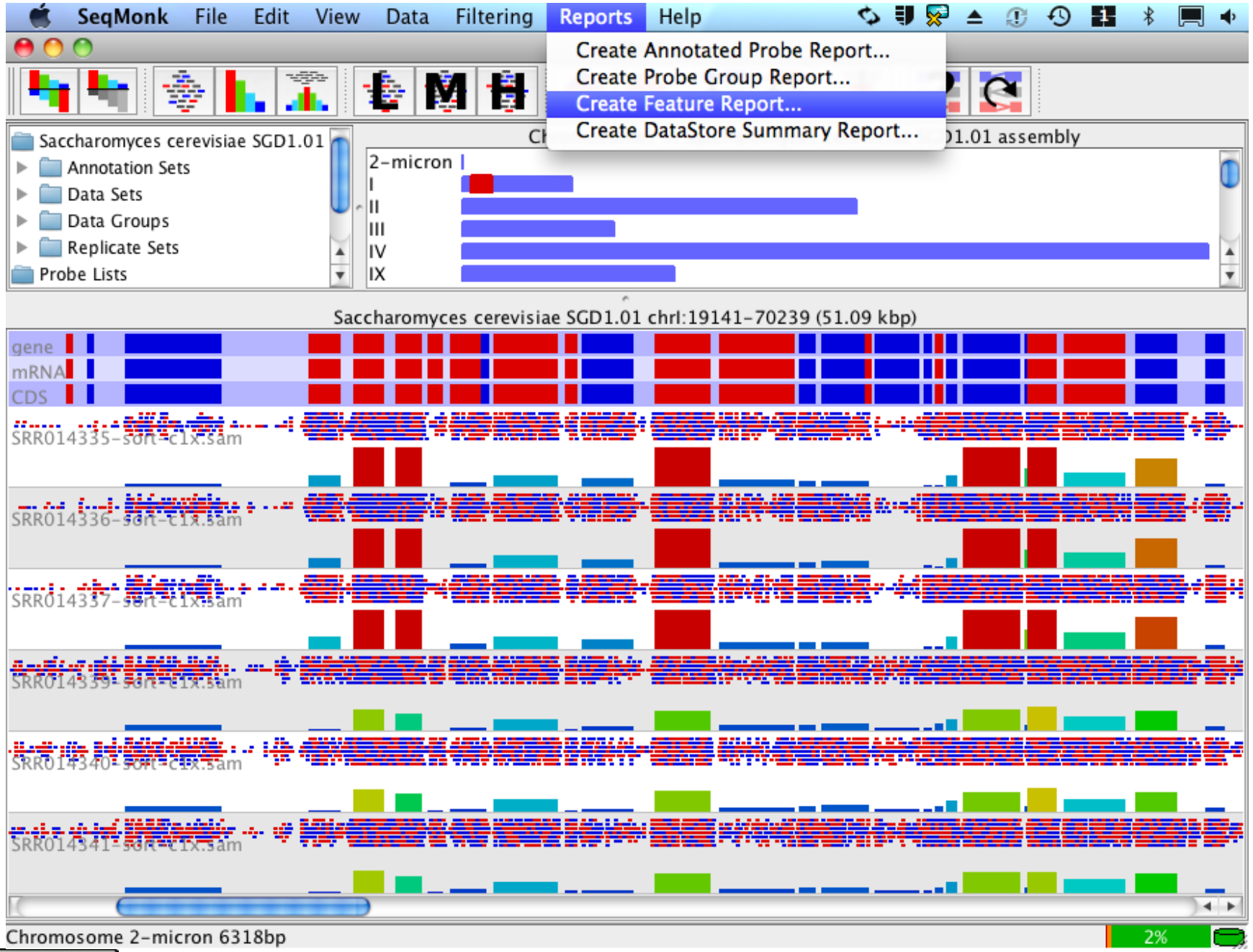
CC BY NC SA


Define Quantitation...
Quantitation Options
Read Count Quantitation
Base Pair Quantitation
Enrichment Quantitation
Difference Quantitation
% Coverage Quantitation
Coverage Depth Quantitation
Relative Quantitation
Manual Correction Quantitation
Rank Quantitation
Count reads on strand
All Reads
Correct for total read count
Count total only within probes
Correct for probe length
Log Transform Count
Count duplicated reads only once
Only quantitate visible stores
Close
Quantitate

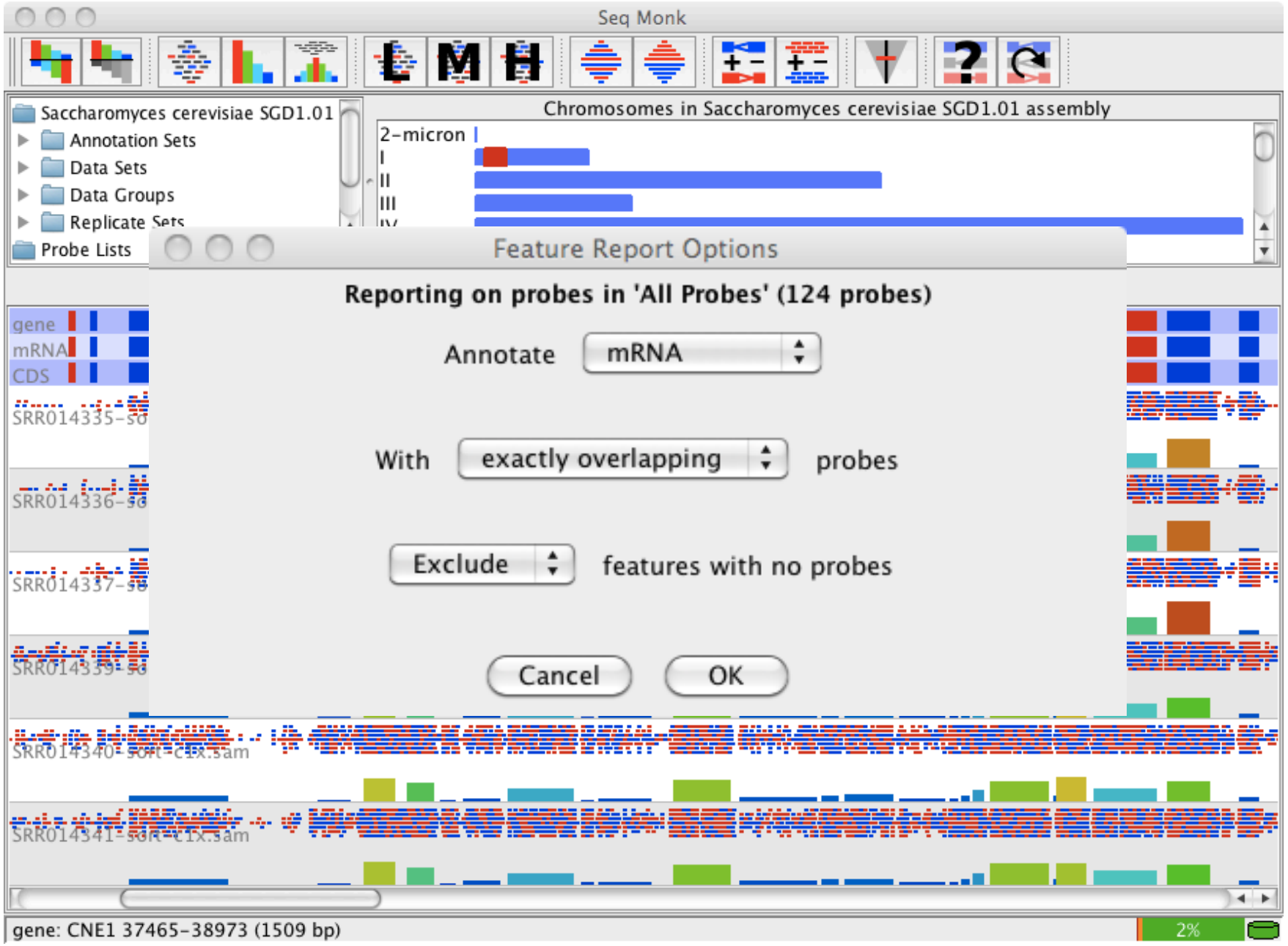
cc BY NC SA


Seq Monk
L M H
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
Chromosomes in Saccharomyces cerevisiae SGD1.01 assembly
2-micron
I
II
III
IV
IX
Saccharomyces cerevisiae SGD1.01 chrI:19141-74473 (55.33 kbp)
gene
mRNA
CDS
SRR014335-sort-c1x.sam
SRR014336-sort-c1x.sam
SRR014337-sort-c1x.sam
SRR014339-sort-c1x.sam
SRR014340-sort-c1x.sam
SRR014341-sort-c1x.sam
SRR014335-sort-c1x.sam 63955-63990(36bp)
4%

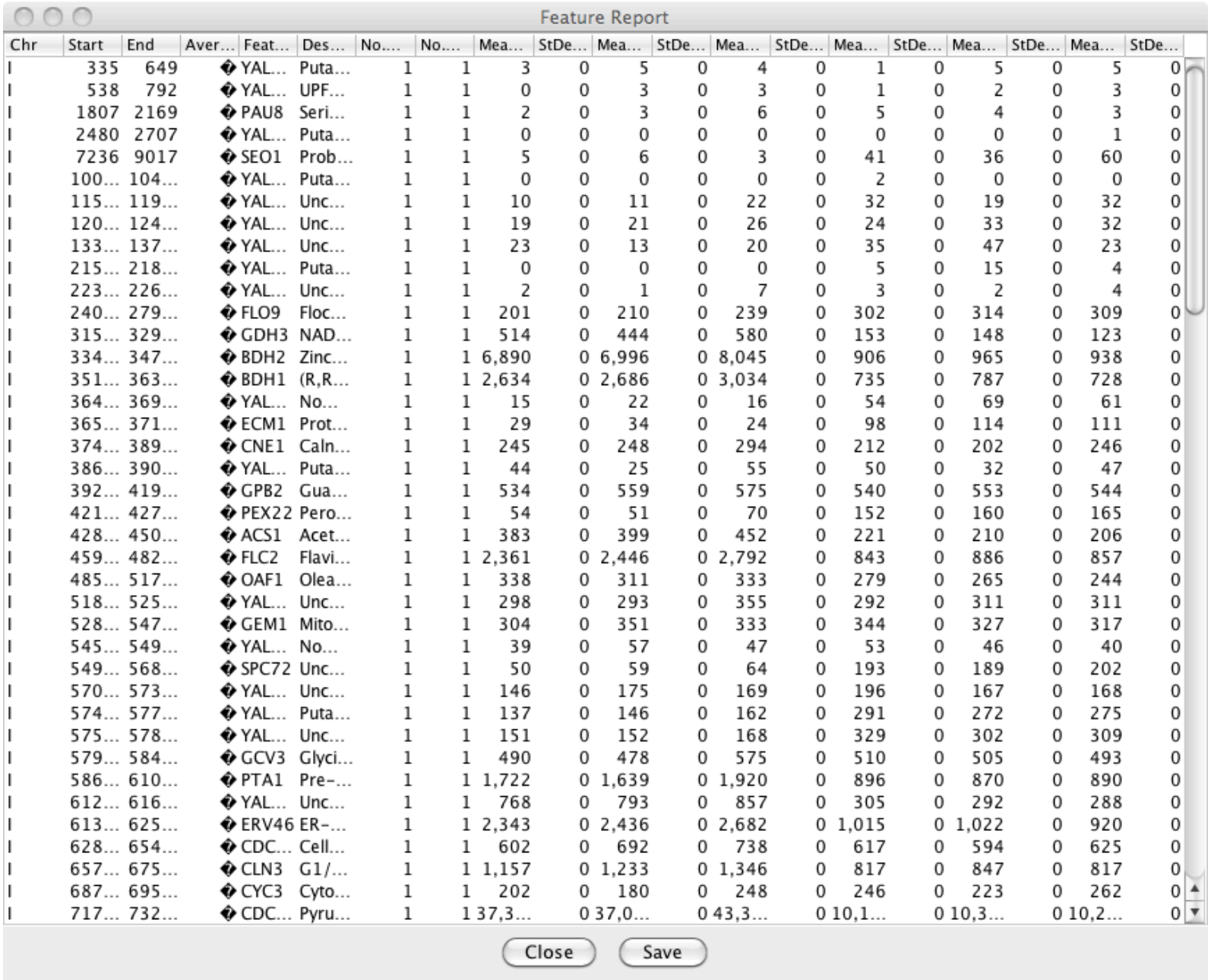
cc BY NC SA


SeqMonk
File
Edit
View
Data
Filtering
Reports
Help
Create Annotated Probe Report...
Create Probe Group Report...
Create Feature Report...
Create DataStore Summary Report...
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
2-micron
I
II
III
IV
IX
Saccharomyces cerevisiae SGD1.01 chrI:19141-70239 (51.09 kbp)
gene
mRNA
CDS
SRR014335-sort-c1x.sam
SRR014336-sort-c1x.sam
SRR014337-sort-c1x.sam
SRR014339-sort-c1x.sam
SRR014340-sort-c1x.sam
SRR014341-sort-c1x.sam
Chromosome 2-micron 6318bp
2%

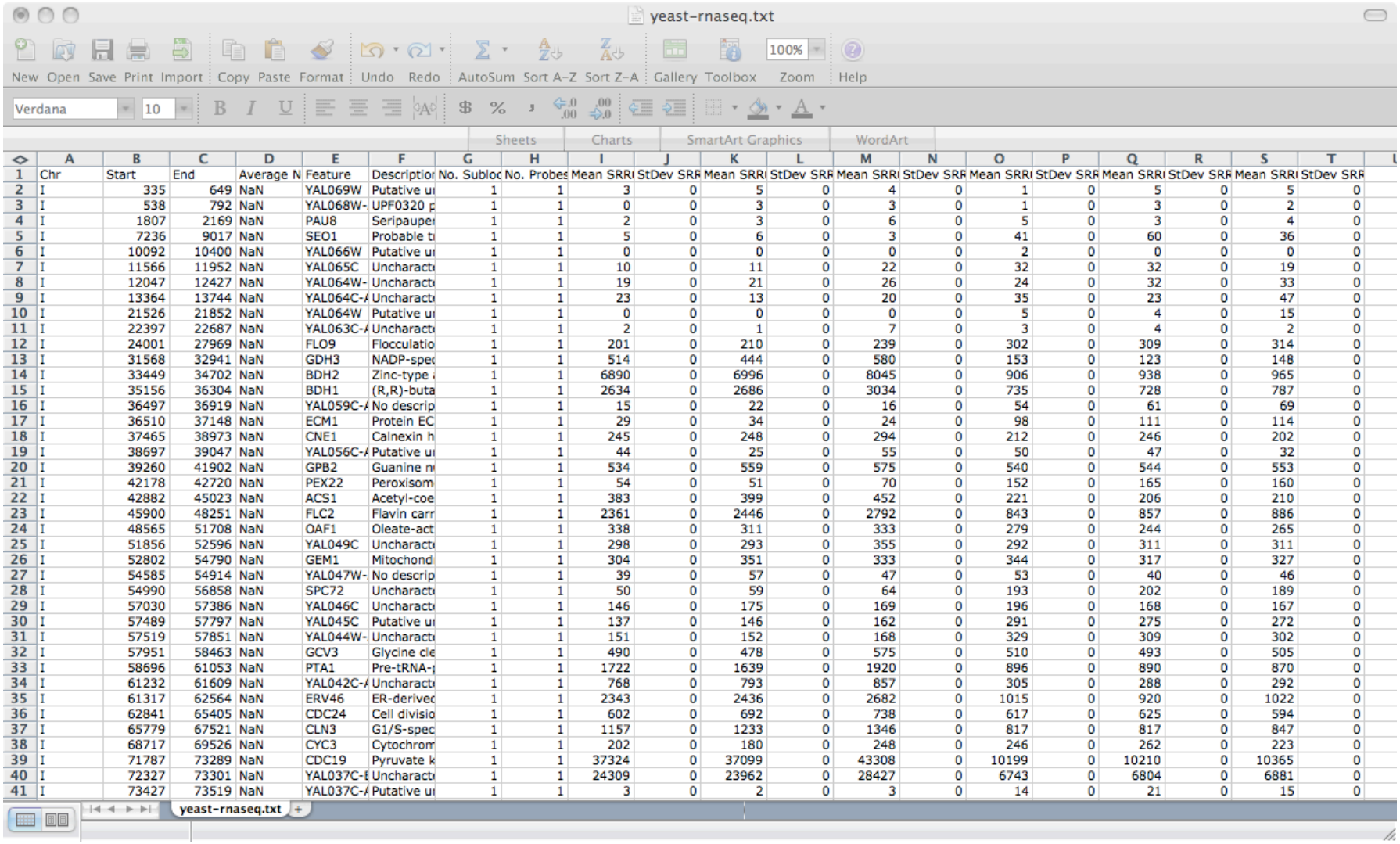
cc BY NC SA


Seq Monk
L M H
Saccharomyces cerevisiae SGD1.01
Annotation Sets
Data Sets
Data Groups
Replicate Sets
Probe Lists
Chromosomes in Saccharomyces cerevisiae SGD1.01 assembly
2-micron
I
II
III
IV
Feature Report Options
Reporting on probes in 'All Probes' (124 probes)
Annotate
mRNA
With
exactly overlapping
probes
Exclude
features with no probes
Cancel
OK
gene
mRNA
CDS
gene: CNE1 37465-38973 (1509 bp)
2%

CC BY NC SA


Feature Report
Chr Start End Aver... Feat... Des... No.... No.... Mea... StDe... Mea... StDe... Mea... StDe... Mea... StDe... Mea... StDe... Mea... StDe...
I 335 649 YAL... Puta... 1 1 3 0 5 0 4 0 1 0 5 0 5 0
I 538 792 YAL... UPF... 1 1 0 0 3 0 3 0 1 0 2 0 3 0
I 1807 2169 PAU8 Seri... 1 1 2 0 3 0 6 0 5 0 4 0 3 0
I 2480 2707 YAL... Puta... 1 1 0 0 0 0 0 0 0 0 0 0 1 0
I 7236 9017 SEO1 Prob... 1 1 5 0 6 0 3 0 41 0 36 0 60 0
I 100... 104... YAL... Puta... 1 1 0 0 0 0 0 0 2 0 0 0 0 0
I 115... 119... YAL... Unc... 1 1 10 0 11 0 22 0 32 0 19 0 32 0
I 120... 124... YAL... Unc... 1 1 19 0 21 0 26 0 24 0 33 0 32 0
I 133... 137... YAL... Unc... 1 1 23 0 13 0 20 0 35 0 47 0 23 0
I 215... 218... YAL... Puta... 1 1 0 0 0 0 0 0 5 0 15 0 4 0
I 223... 226... YAL... Unc... 1 1 2 0 1 0 7 0 3 0 2 0 4 0
I 240... 279... FLO9 Floc... 1 1 201 0 210 0 239 0 302 0 314 0 309 0
I 315... 329... GDH3 NAD... 1 1 514 0 444 0 580 0 153 0 148 0 123 0
I 334... 347... BDH2 Zinc... 1 1 6,890 0 6,996 0 8,045 0 906 0 965 0 938 0
I 351... 363... BDH1 (R,R... 1 1 2,634 0 2,686 0 3,034 0 735 0 787 0 728 0
I 364... 369... YAL... No... 1 1 15 0 22 0 16 0 54 0 69 0 61 0
I 365... 371... ECM1 Prot... 1 1 29 0 34 0 24 0 98 0 114 0 111 0
I 374... 389... CNE1 Caln... 1 1 245 0 248 0 294 0 212 0 202 0 246 0
I 386... 390... YAL... Puta... 1 1 44 0 25 0 55 0 50 0 32 0 47 0
I 392... 419... GPB2 Gua... 1 1 534 0 559 0 575 0 540 0 553 0 544 0
I 421... 427... PEX22 Pero... 1 1 54 0 51 0 70 0 152 0 160 0 165 0
I 428... 450... ACS1 Acet... 1 1 383 0 399 0 452 0 221 0 210 0 206 0
I 459... 482... FLC2 Flavi... 1 1 2,361 0 2,446 0 2,792 0 843 0 886 0 857 0
I 485... 517... OAF1 Olea... 1 1 338 0 311 0 333 0 279 0 265 0 244 0
I 518... 525... YAL... Unc... 1 1 298 0 293 0 355 0 292 0 311 0 311 0
I 528... 547... GEM1 Mito... 1 1 304 0 351 0 333 0 344 0 327 0 317 0
I 545... 549... YAL... No... 1 1 39 0 57 0 47 0 53 0 46 0 40 0
I 549... 568... SPC72 Unc... 1 1 50 0 59 0 64 0 193 0 189 0 202 0
I 570... 573... YAL... Unc... 1 1 146 0 175 0 169 0 196 0 167 0 168 0
I 574... 577... YAL... Puta... 1 1 137 0 146 0 162 0 291 0 272 0 275 0
I 575... 578... YAL... Unc... 1 1 151 0 152 0 168 0 329 0 302 0 309 0
I 579... 584... GCV3 Glyci... 1 1 490 0 478 0 575 0 510 0 505 0 493 0
I 586... 610... PTA1 Pre-... 1 1 1,722 0 1,639 0 1,920 0 896 0 870 0 890 0
I 612... 616... YAL... Unc... 1 1 768 0 793 0 857 0 305 0 292 0 288 0
I 613... 625... ERV46 ER-... 1 1 2,343 0 2,436 0 2,682 0 1,015 0 1,022 0 920 0
I 628... 654... CDC... Cell... 1 1 602 0 692 0 738 0 617 0 594 0 625 0
I 657... 675... CLN3 G1/... 1 1 1,157 0 1,233 0 1,346 0 817 0 847 0 817 0
I 687... 695... CYC3 Cyto... 1 1 202 0 180 0 248 0 246 0 223 0 262 0
I 717... 732... CDC... Pyru... 1 1 37,3... 0 37,0... 0 43,3... 0 10,1... 0 10,3... 0 10,2... 0
Close
Save

CC BY NC SA



yeast-rnaseq.txt

New Open Save Print Import Copy Paste Format Undo Redo AutoSum Sort A-Z Sort Z-A Gallery Toolbox 100% Zoom Help

Verdana 10

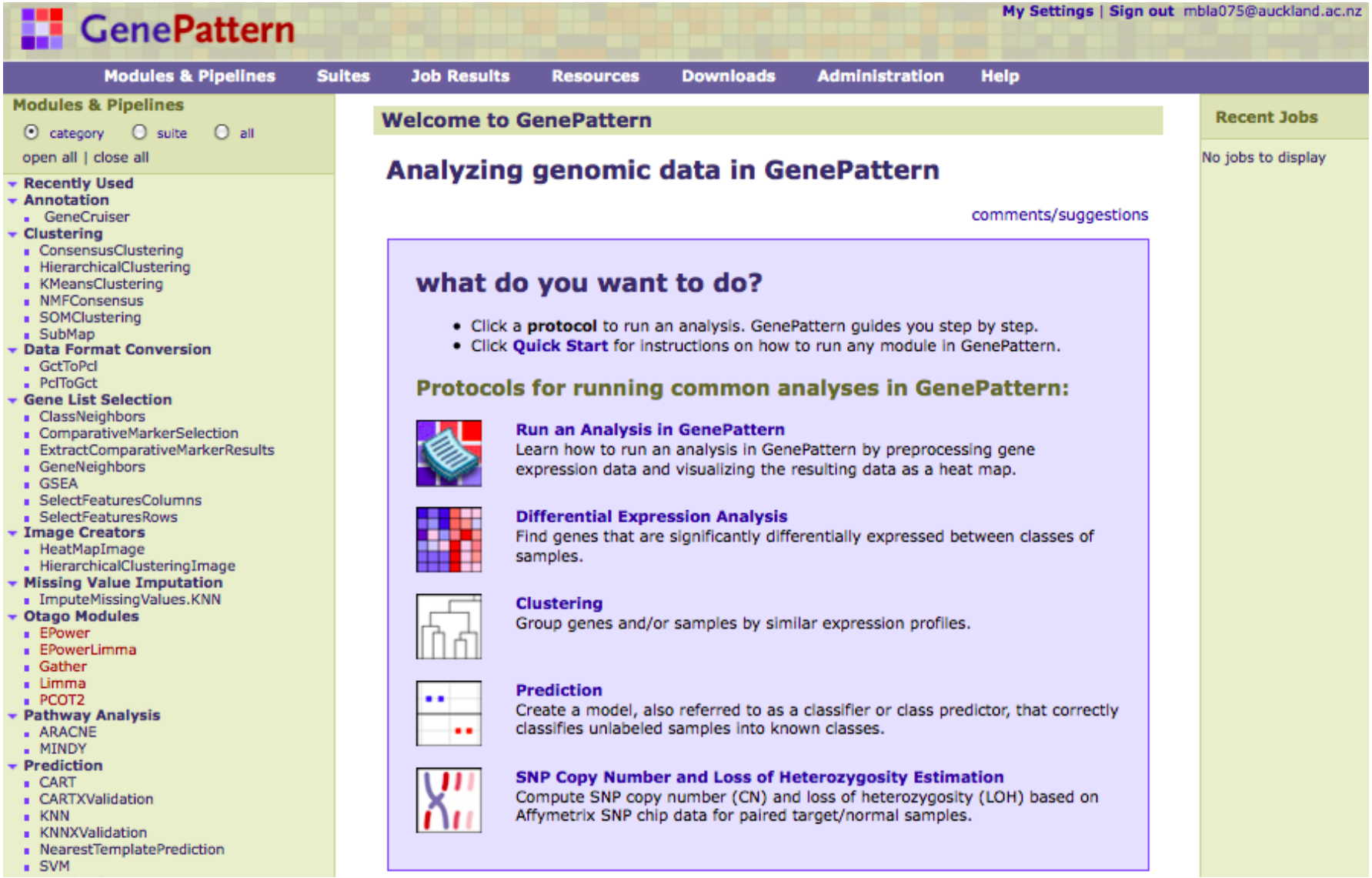
Sheets Charts SmartArt Graphics WordArt

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Chr	Start	End	Average N	Feature	Description	No. Subloc	No. Probes	Mean SRR	StDev SRR	Mean SRR	StDev SRR	Mean SRR	StDev SRR	Mean SRR	StDev SRR	Mean SRR	StDev SRR	Mean SRR	StDev SRR
2	I	335	649	NaN	YAL069W	Putative u	1	1	3	0	5	0	4	0	1	0	5	0	5	0
3	I	538	792	NaN	YAL068W-	UPF0320 p	1	1	0	0	3	0	3	0	1	0	3	0	2	0
4	I	1807	2169	NaN	PAU8	Seripauper	1	1	2	0	3	0	6	0	5	0	3	0	4	0
5	I	7236	9017	NaN	SEO1	Probable t	1	1	5	0	6	0	3	0	41	0	60	0	36	0
6	I	10092	10400	NaN	YAL066W	Putative u	1	1	0	0	0	0	0	0	2	0	0	0	0	0
7	I	11566	11952	NaN	YAL065C	Uncharacte	1	1	10	0	11	0	22	0	32	0	32	0	19	0
8	I	12047	12427	NaN	YAL064W-	Uncharacte	1	1	19	0	21	0	26	0	24	0	32	0	33	0
9	I	13364	13744	NaN	YAL064C-A	Uncharacte	1	1	23	0	13	0	20	0	35	0	23	0	47	0
10	I	21526	21852	NaN	YAL064W	Putative u	1	1	0	0	0	0	0	0	5	0	4	0	15	0
11	I	22397	22687	NaN	YAL063C-A	Uncharacte	1	1	2	0	1	0	7	0	3	0	4	0	2	0
12	I	24001	27969	NaN	FLO9	Flocculatio	1	1	201	0	210	0	239	0	302	0	309	0	314	0
13	I	31568	32941	NaN	GDH3	NADP-spec	1	1	514	0	444	0	580	0	153	0	123	0	148	0
14	I	33449	34702	NaN	BDH2	Zinc-type a	1	1	6890	0	6996	0	8045	0	906	0	938	0	965	0
15	I	35156	36304	NaN	BDH1	(R,R)-buta	1	1	2634	0	2686	0	3034	0	735	0	728	0	787	0
16	I	36497	36919	NaN	YAL059C-A	No descrip	1	1	15	0	22	0	16	0	54	0	61	0	69	0
17	I	36510	37148	NaN	ECM1	Protein EC	1	1	29	0	34	0	24	0	98	0	111	0	114	0
18	I	37465	38973	NaN	CNE1	Calnexin h	1	1	245	0	248	0	294	0	212	0	246	0	202	0
19	I	38697	39047	NaN	YAL056C-A	Putative u	1	1	44	0	25	0	55	0	50	0	47	0	32	0
20	I	39260	41902	NaN	GPB2	Guanine n	1	1	534	0	559	0	575	0	540	0	544	0	553	0
21	I	42178	42720	NaN	PEX22	Peroxisom	1	1	54	0	51	0	70	0	152	0	165	0	160	0
22	I	42882	45023	NaN	ACS1	Acetyl-coe	1	1	383	0	399	0	452	0	221	0	206	0	210	0
23	I	45900	48251	NaN	FLC2	Flavin carr	1	1	2361	0	2446	0	2792	0	843	0	857	0	886	0
24	I	48565	51708	NaN	OAF1	Oleate-act	1	1	338	0	311	0	333	0	279	0	244	0	265	0
25	I	51856	52596	NaN	YAL049C	Uncharacte	1	1	298	0	293	0	355	0	292	0	311	0	311	0
26	I	52802	54790	NaN	GEM1	Mitochond	1	1	304	0	351	0	333	0	344	0	317	0	327	0
27	I	54585	54914	NaN	YAL047W-	No descrip	1	1	39	0	57	0	47	0	53	0	40	0	46	0
28	I	54990	56858	NaN	SPC72	Uncharacte	1	1	50	0	59	0	64	0	193	0	202	0	189	0
29	I	57030	57386	NaN	YAL046C	Uncharacte	1	1	146	0	175	0	169	0	196	0	168	0	167	0
30	I	57489	57797	NaN	YAL045C	Putative u	1	1	137	0	146	0	162	0	291	0	275	0	272	0
31	I	57519	57851	NaN	YAL044W-	Uncharacte	1	1	151	0	152	0	168	0	329	0	309	0	302	0
32	I	57951	58463	NaN	GCV3	Glycine cle	1	1	490	0	478	0	575	0	510	0	493	0	505	0
33	I	58696	61053	NaN	PTA1	Pre-tRNA-p	1	1	1722	0	1639	0	1920	0	896	0	890	0	870	0
34	I	61232	61609	NaN	YAL042C-A	Uncharacte	1	1	768	0	793	0	857	0	305	0	288	0	292	0
35	I	61317	62564	NaN	ERV46	ER-derived	1	1	2343	0	2436	0	2682	0	1015	0	920	0	1022	0
36	I	62841	65405	NaN	CDC24	Cell divisio	1	1	602	0	692	0	738	0	617	0	625	0	594	0
37	I	65779	67521	NaN	CLN3	G1/S-spec	1	1	1157	0	1233	0	1346	0	817	0	817	0	847	0
38	I	68717	69526	NaN	CYC3	Cytochrom	1	1	202	0	180	0	248	0	246	0	262	0	223	0
39	I	71787	73289	NaN	CDC19	Pyruvate k	1	1	37324	0	37099	0	43308	0	10199	0	10210	0	10365	0
40	I	72327	73301	NaN	YAL037C-B	Uncharacte	1	1	24309	0	23962	0	28427	0	6743	0	6804	0	6881	0
41	I	73427	73519	NaN	YAL037C-A	Putative u	1	1	3	0	2	0	3	0	14	0	21	0	15	0

yeast-rnaseq.txt

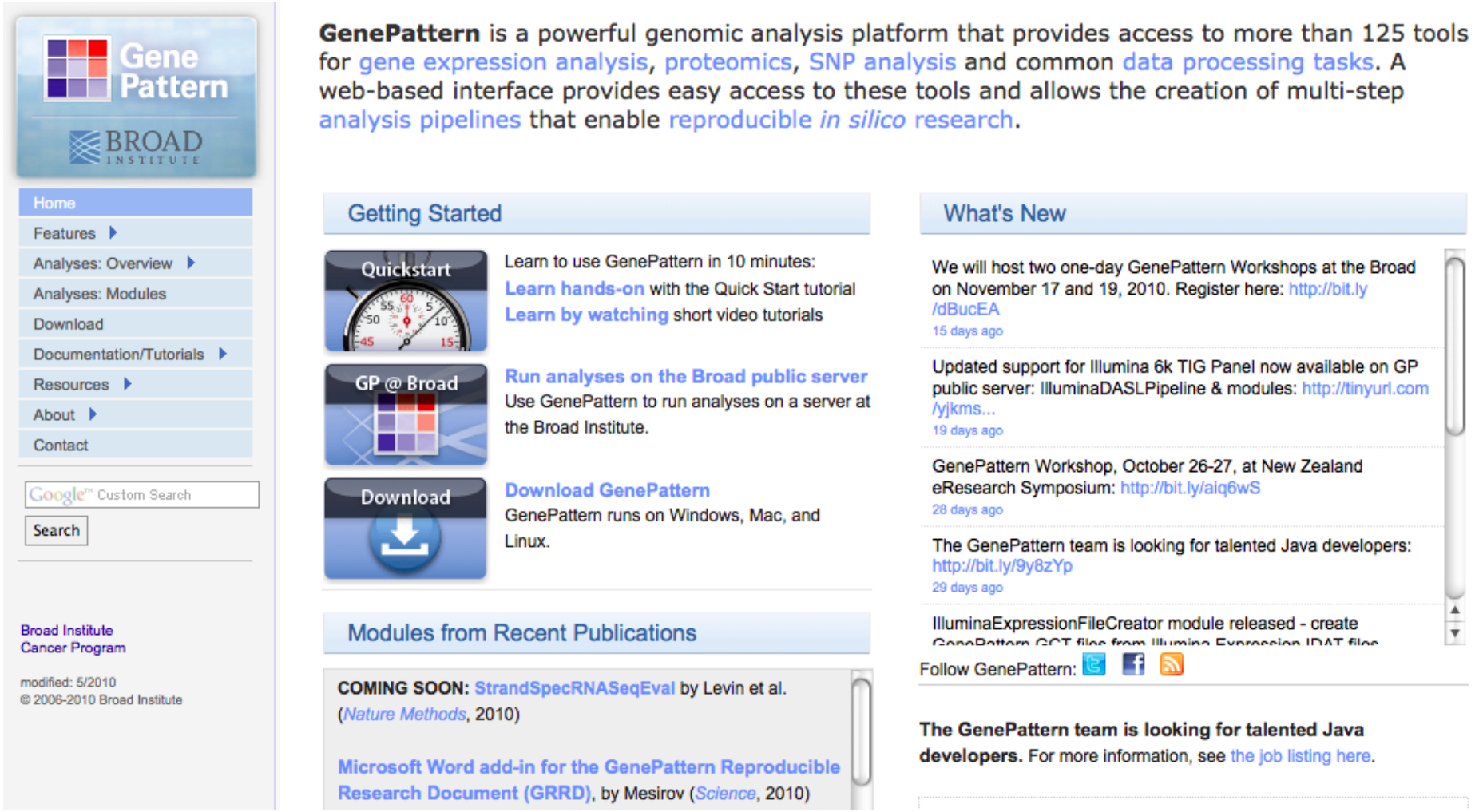
CC BY NC SA

http://genepattern.auckland.ac.nz



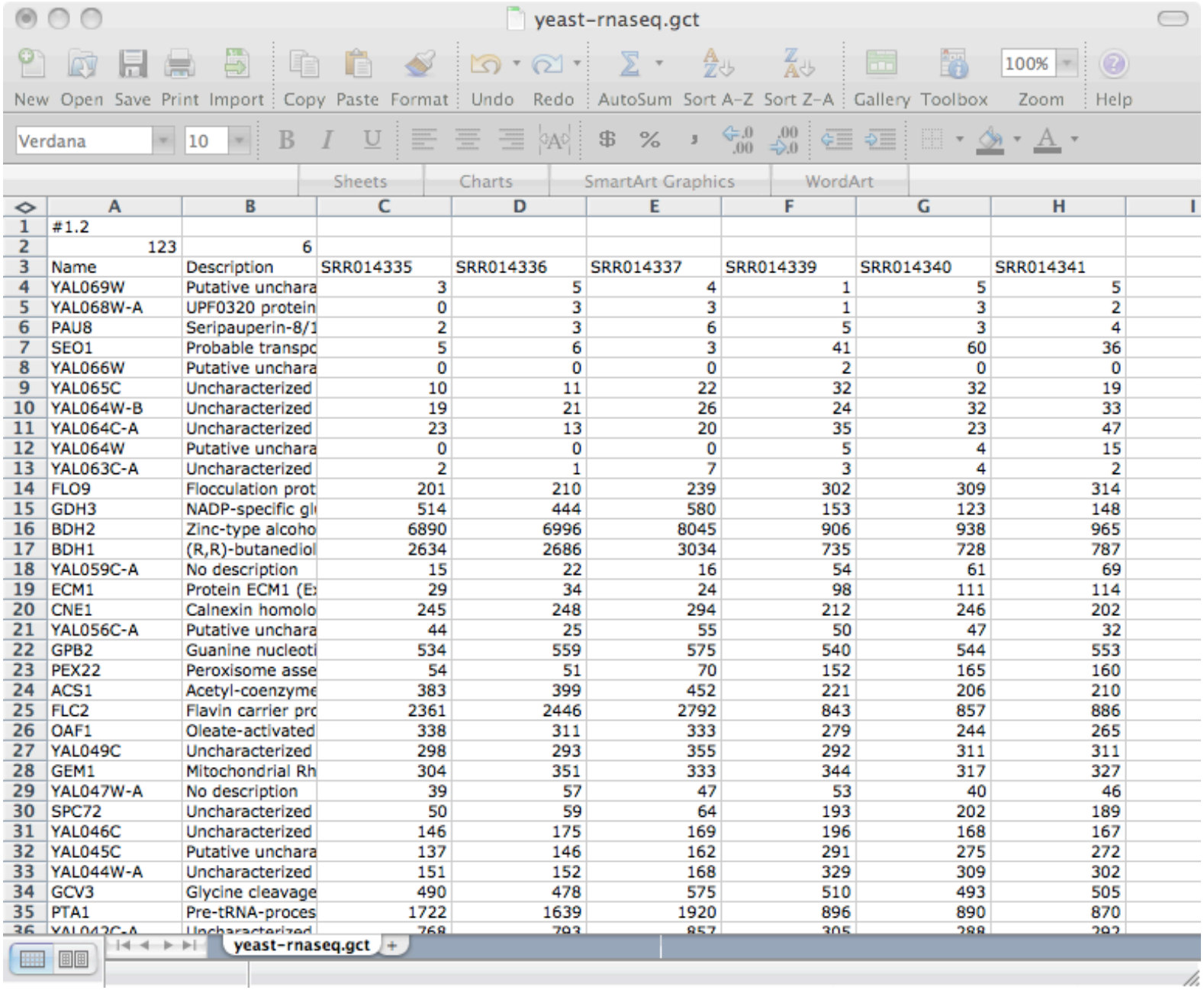
cc BY NC SA

http://www.broadinstitute.org/cancer/software/genepattern/



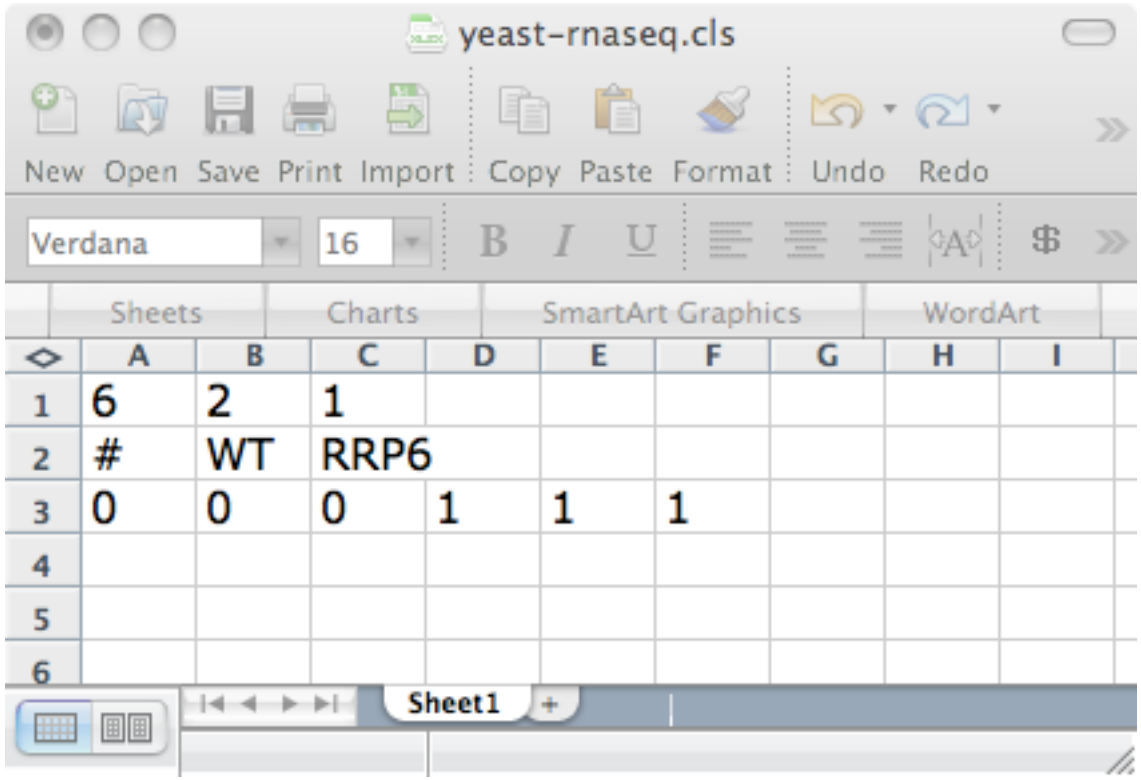
Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP
GenePattern 2.0 Nature Genetics 38 no. 5 (2006): pp500-501

CC BY NC SA

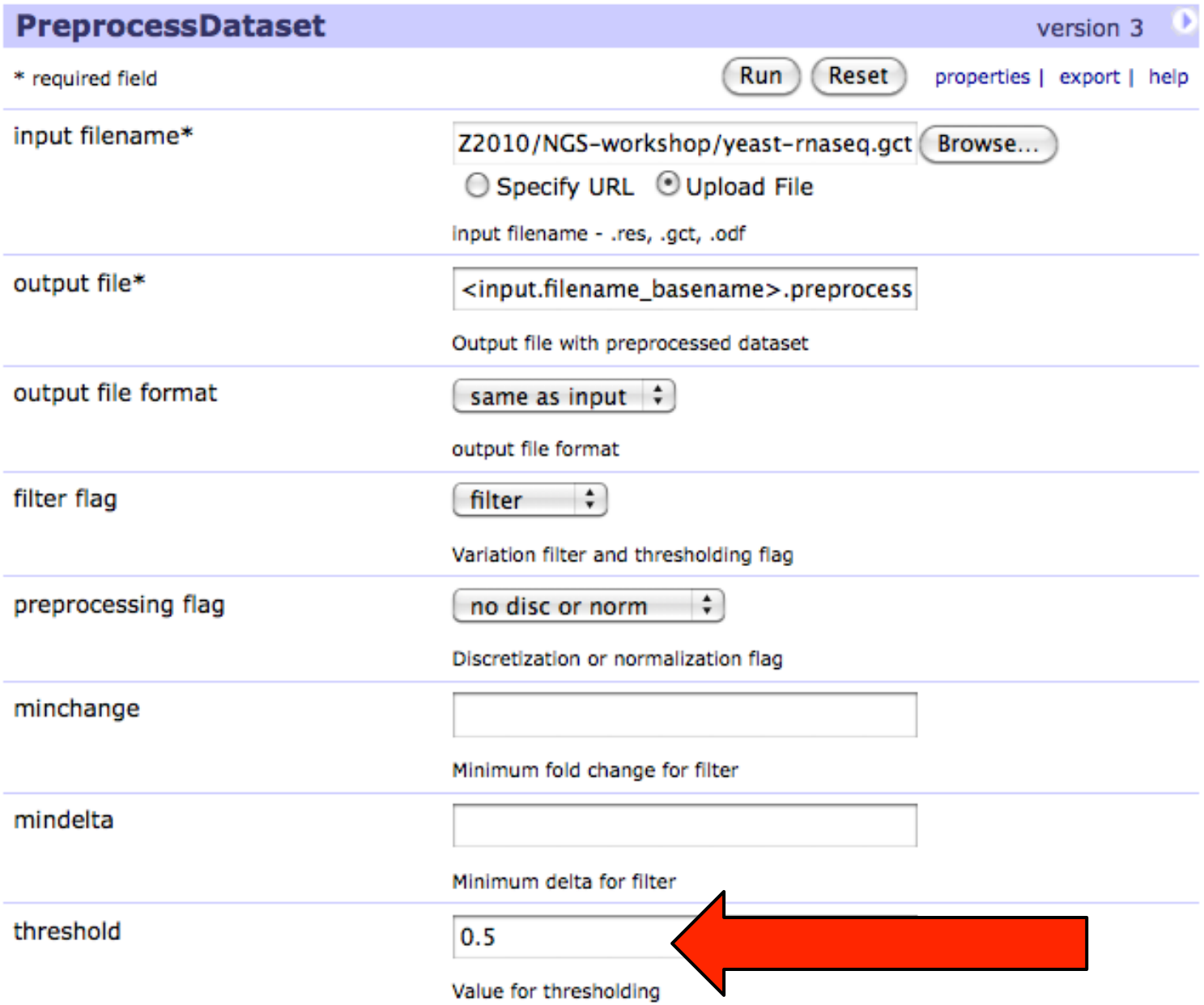


	A	B	C	D	E	F	G	H	I
1	#1.2								
2	123	6							
3	Name	Description	SRR014335	SRR014336	SRR014337	SRR014339	SRR014340	SRR014341	
4	YAL069W	Putative unchara	3	5	4	1	5	5	
5	YAL068W-A	UPF0320 protein	0	3	3	1	3	2	
6	PAU8	Seripauperin-8/1	2	3	6	5	3	4	
7	SEO1	Probable transpc	5	6	3	41	60	36	
8	YAL066W	Putative unchara	0	0	0	2	0	0	
9	YAL065C	Uncharacterized	10	11	22	32	32	19	
10	YAL064W-B	Uncharacterized	19	21	26	24	32	33	
11	YAL064C-A	Uncharacterized	23	13	20	35	23	47	
12	YAL064W	Putative unchara	0	0	0	5	4	15	
13	YAL063C-A	Uncharacterized	2	1	7	3	4	2	
14	FLO9	Flocculation prot	201	210	239	302	309	314	
15	GDH3	NADP-specific gl	514	444	580	153	123	148	
16	BDH2	Zinc-type alcoho	6890	6996	8045	906	938	965	
17	BDH1	(R,R)-butanediol	2634	2686	3034	735	728	787	
18	YAL059C-A	No description	15	22	16	54	61	69	
19	ECM1	Protein ECM1 (E:	29	34	24	98	111	114	
20	CNE1	Calnexin homolo	245	248	294	212	246	202	
21	YAL056C-A	Putative unchara	44	25	55	50	47	32	
22	GPB2	Guanine nucleoti	534	559	575	540	544	553	
23	PEX22	Peroxisome asse	54	51	70	152	165	160	
24	ACS1	Acetyl-coenzyme	383	399	452	221	206	210	
25	FLC2	Flavin carrier pro	2361	2446	2792	843	857	886	
26	OAF1	Oleate-activated	338	311	333	279	244	265	
27	YAL049C	Uncharacterized	298	293	355	292	311	311	
28	GEM1	Mitochondrial Rh	304	351	333	344	317	327	
29	YAL047W-A	No description	39	57	47	53	40	46	
30	SPC72	Uncharacterized	50	59	64	193	202	189	
31	YAL046C	Uncharacterized	146	175	169	196	168	167	
32	YAL045C	Putative unchara	137	146	162	291	275	272	
33	YAL044W-A	Uncharacterized	151	152	168	329	309	302	
34	GCV3	Glycine cleavage	490	478	575	510	493	505	
35	PTA1	Pre-tRNA-proces	1722	1639	1920	896	890	870	
36	YAL042C-A	Uncharacterized	768	793	857	305	288	292	

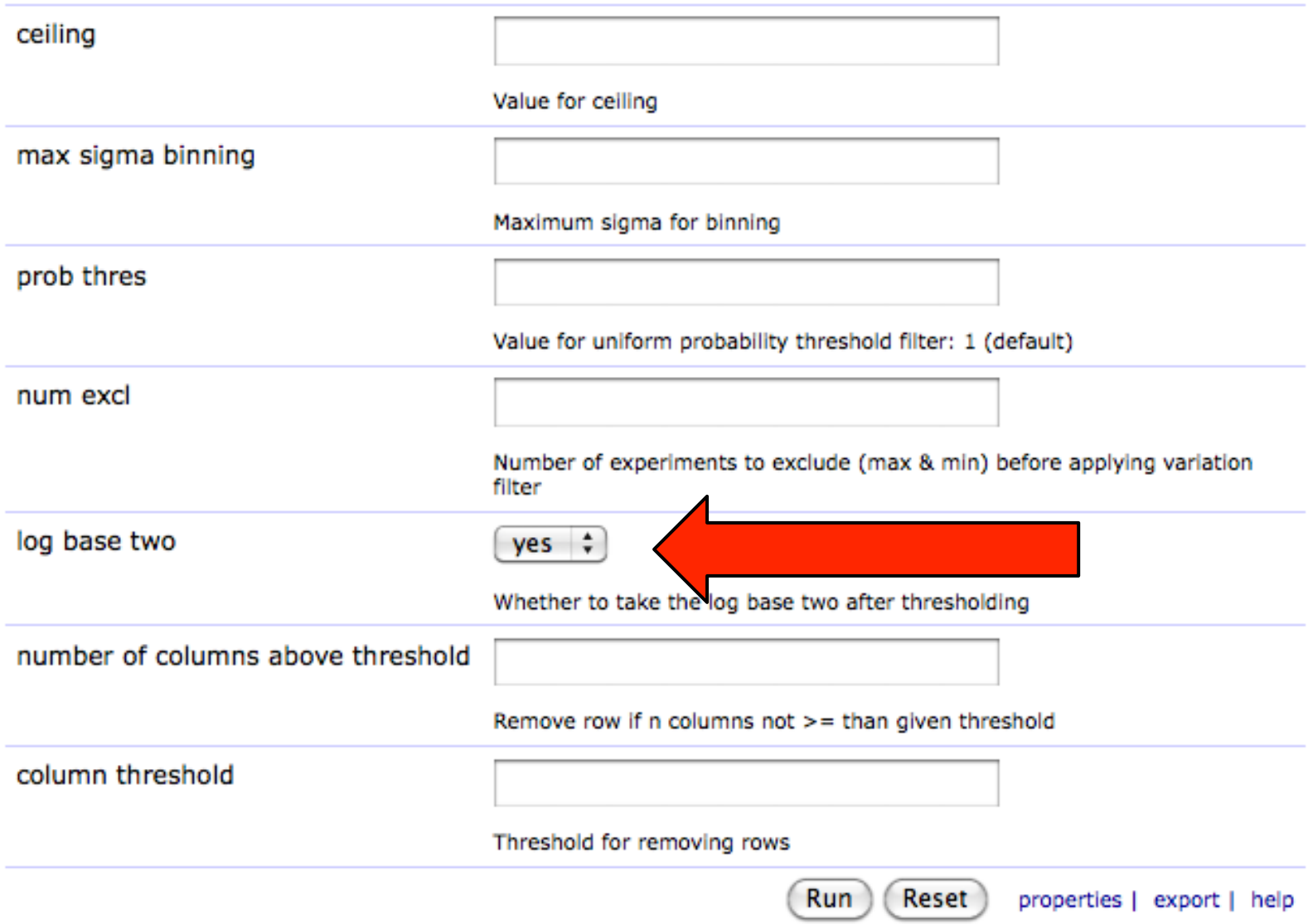
CC BY NC SA


yeast-rnaseq.cls
New Open Save Print Import Copy Paste Format Undo Redo
Verdana 16
Sheets Charts SmartArt Graphics WordArt
A B C D E F G H I
1 6 2 1
2 # WT RRP6
3 0 0 0 1 1 1
4
5
6
Sheet1

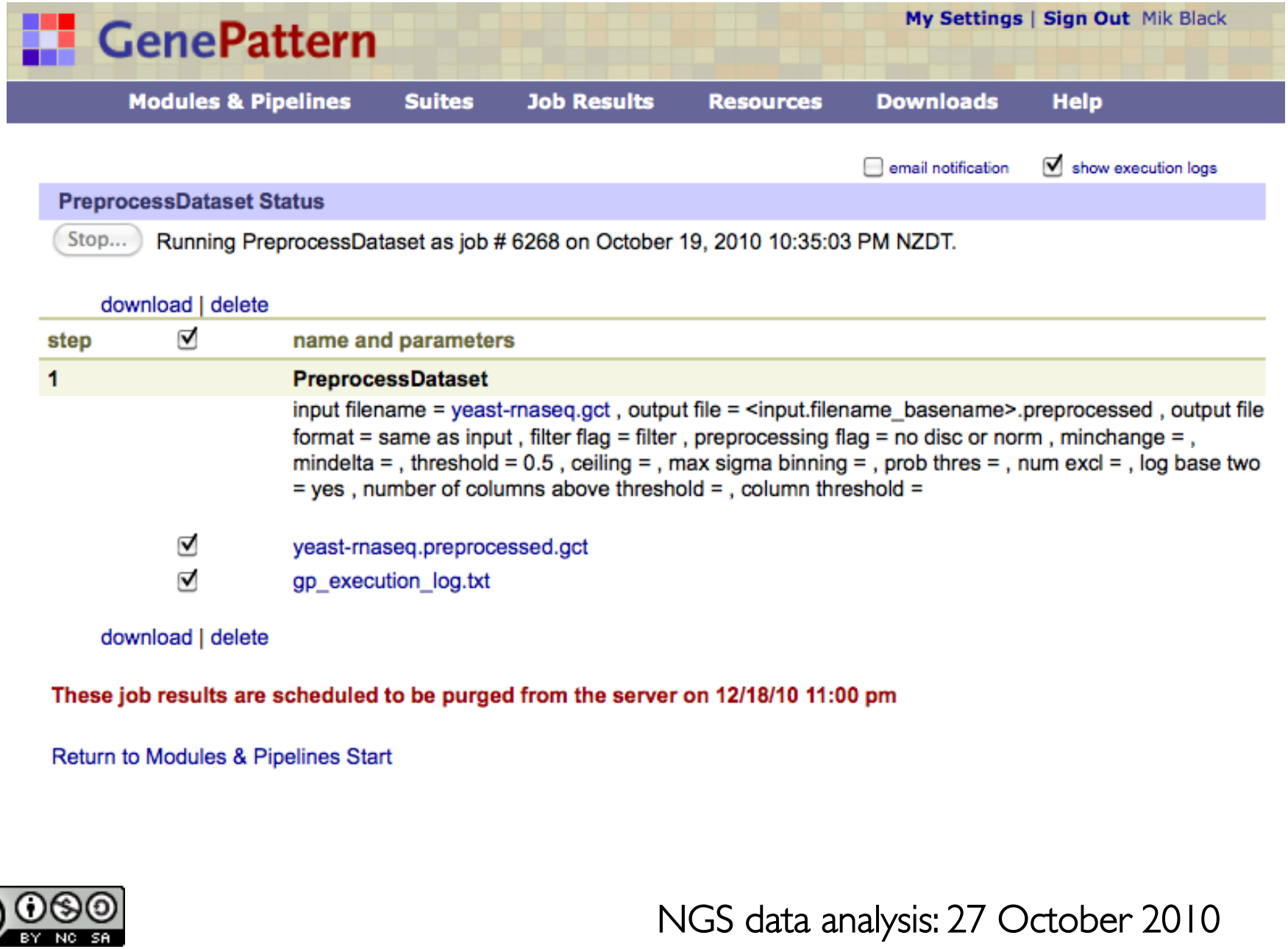
cc BY NC SA


PreprocessDataset
version 3
* required field
Run
Reset
properties | export | help
input filename*
Z2010/NGS-workshop/yeast-rnaseq.gct
Browse...
Specify URL
Upload File
input filename - .res, .gct, .odf
output file*
<input.filename_basename>.preprocess
Output file with preprocessed dataset
output file format
same as input
output file format
filter flag
filter
Variation filter and thresholding flag
preprocessing flag
no disc or norm
Discretization or normalization flag
minchange
Minimum fold change for filter
mindelta
Minimum delta for filter
threshold
0.5
Value for thresholding

cc BY NC SA


ceiling
Value for ceiling
max sigma binning
Maximum sigma for binning
prob thres
Value for uniform probability threshold filter: 1 (default)
num excl
Number of experiments to exclude (max & min) before applying variation filter
log base two
yes
Whether to take the log base two after thresholding
number of columns above threshold
Remove row if n columns not >= than given threshold
column threshold
Threshold for removing rows
Run
Reset
properties | export | help

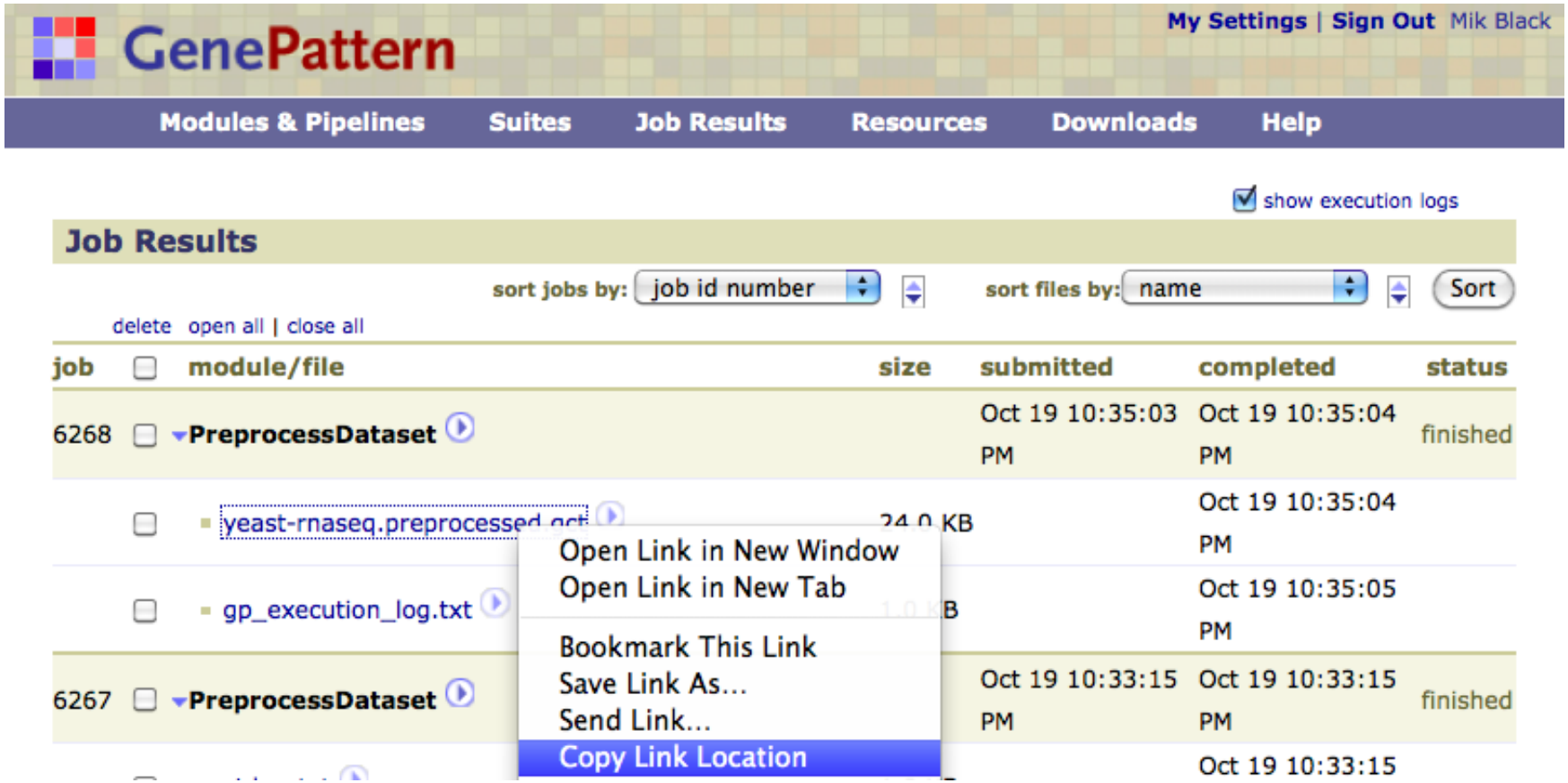
cc BY NC SA


GenePattern
My Settings | Sign Out Mik Black
Modules & Pipelines
Suites
Job Results
Resources
Downloads
Help
email notification
show execution logs
PreprocessDataset Status
Stop...
Running PreprocessDataset as job # 6268 on October 19, 2010 10:35:03 PM NZDT.
download | delete
step
name and parameters
1
PreprocessDataset
input filename = yeast-rnaseq.gct , output file = <input.filename_basename>.preprocessed , output file format = same as input , filter flag = filter , preprocessing flag = no disc or norm , minchange = , mindelta = , threshold = 0.5 , ceiling = , max sigma binning = , prob thres = , num excl = , log base two = yes , number of columns above threshold = , column threshold =
yeast-rnaseq.preprocessed.gct
gp_execution_log.txt
download | delete
These job results are scheduled to be purged from the server on 12/18/10 11:00 pm
Return to Modules & Pipelines Start

CC BY NC SA


GenePattern
My Settings | Sign Out Mik Black
Modules & Pipelines
Suites
Job Results
Resources
Downloads
Help
show execution logs
Job Results
sort jobs by: job id number
sort files by: name
Sort
delete open all | close all
job
module/file
size
submitted
completed
status
6268
PreprocessDataset
Oct 19 10:35:03 PM
Oct 19 10:35:04 PM
finished
yeast-rnaseq.preprocessed.gct
24.0 KB
Oct 19 10:35:04 PM
gp_execution_log.txt
1.0 KB
Oct 19 10:35:05 PM

CC BY NC SA


GenePattern
My Settings | Sign Out Mik Black
Modules & Pipelines
Suites
Job Results
Resources
Downloads
Help
show execution logs
Job Results
sort jobs by: job id number
sort files by: name
Sort
delete open all | close all
job
module/file
size
submitted
completed
status
6268
PreprocessDataset
Oct 19 10:35:03 PM
Oct 19 10:35:04 PM
finished
yeast-rnaseq.preprocessed.gct
24.0 KB
Oct 19 10:35:04 PM
gp_execution_log.txt
Oct 19 10:35:05 PM
6267
PreprocessDataset
Oct 19 10:33:15 PM
Oct 19 10:33:15 PM
finished
Oct 19 10:33:15
Open Link in New Window
Open Link in New Tab
Bookmark This Link
Save Link As...
Send Link...
Copy Link Location

cc BY NC SA

Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 3

Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments

Gordon K. Smyth*

limma

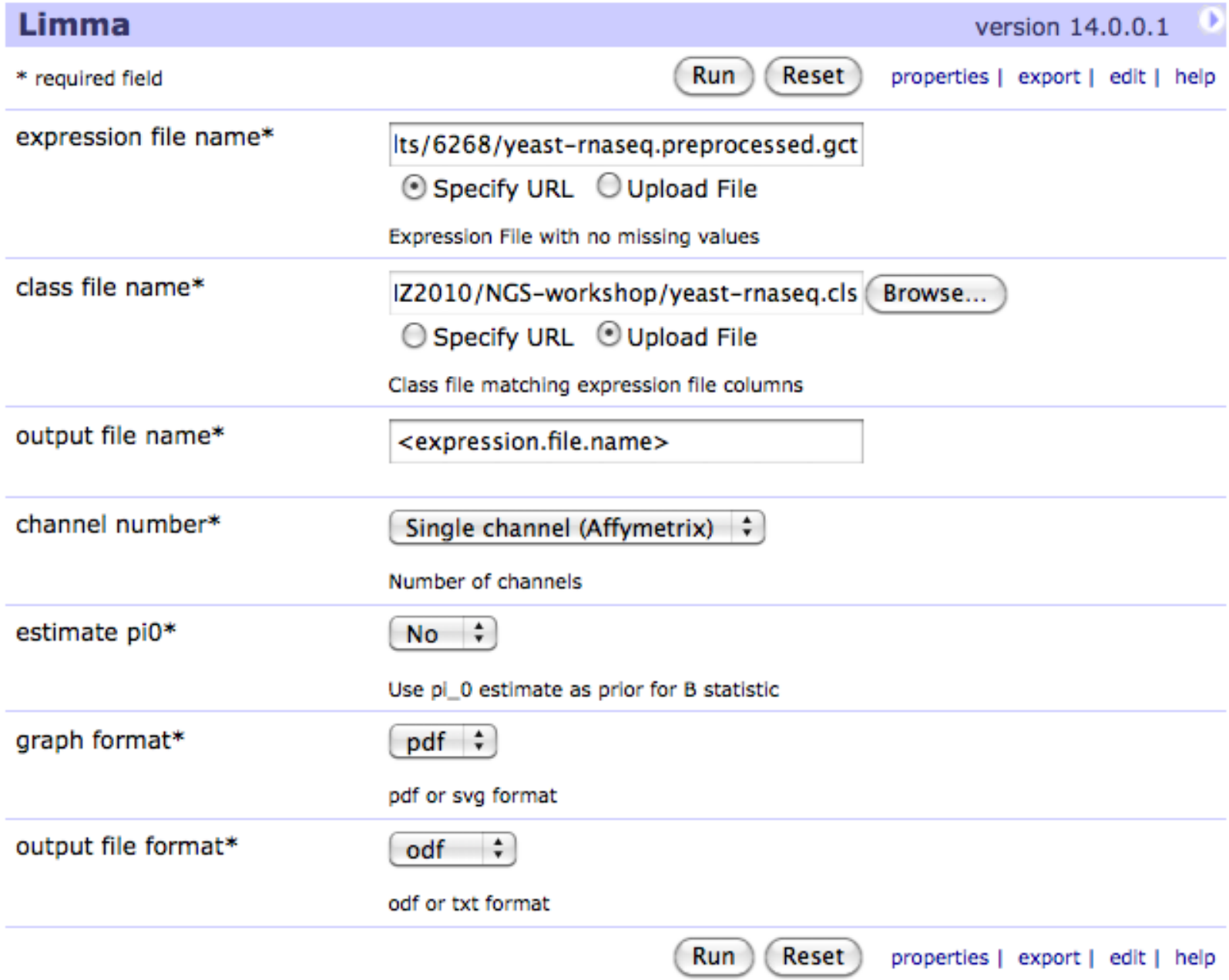
Linear Models for Microarray Data

Data analysis, linear models and differential expression for microarray data.

Author	Gordon Smyth with contributions from Matthew Ritchie, Jeremy Silver, James Wettenhall, Natalie Thorne, Mette Langaas, Egil Ferkingstad, Marcus Davy, Francois Pepin, Dongseok Choi, Di Wu, Alicia Oshlack, Carolyn de Graaf, Yifang Hu, Wei Shi and Belinda Phipson.
Maintainer	Gordon Smyth

http://www.bioconductor.org/packages/release/bioc/html/limma.html

cc BY NC SA


Limma
version 14.0.0.1
* required field
Run
Reset
properties | export | edit | help
expression file name*
lts/6268/yeast-rnaseq.preprocessed.gct
Specify URL
Upload File
Expression File with no missing values
class file name*
IZ2010/NGS-workshop/yeast-rnaseq.cls
Browse...
Specify URL
Upload File
Class file matching expression file columns
output file name*
<expression.file.name>
channel number*
Single channel (Affymetrix)
Number of channels
estimate pi0*
No
Use pi_0 estimate as prior for B statistic
graph format*
pdf
pdf or svg format
output file format*
odf
odf or txt format
Run
Reset
properties | export | edit | help

CC BY NC SA



email notification show execution logs

Limma Status

Stop... Running Limma as job # 6274 on October 19, 2010 10:57:19 PM NZDT.

download | delete

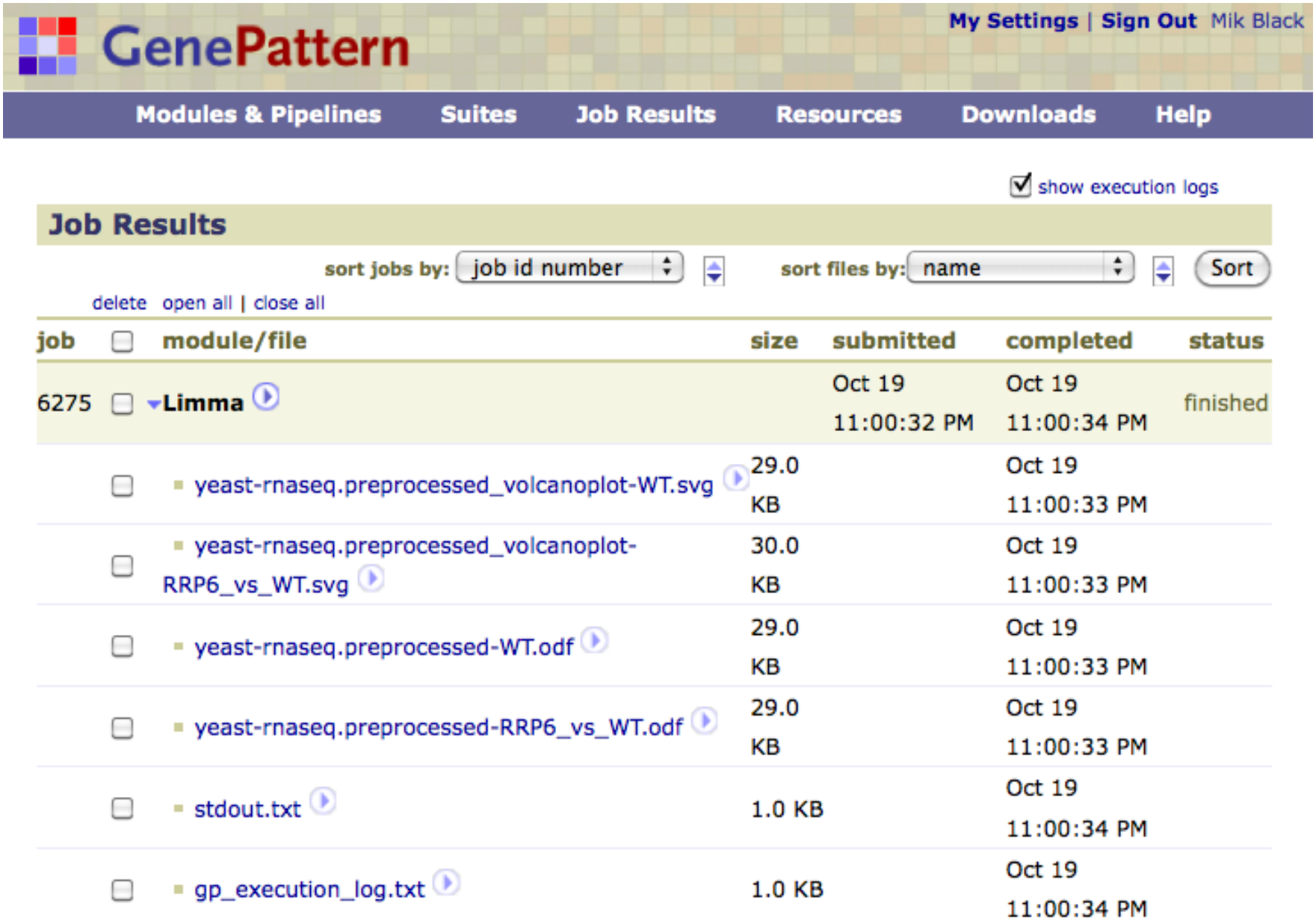
step		name and parameters
1		**Limma**
		expression file name = http://bioanalysis.otago.ac.nz/gp/jobResults/6268/yeast-rnaseq.preprocessed.gct , class file name = yeast-rnaseq.cls , output file name = <expression.file.name> , channel number = Single channel (Affymetrix) , estimate pi0 = No , graph format = pdf , output file format = odf
	☑	yeast-rnaseq.preprocessed-WT.odf
	☑	yeast-rnaseq.preprocessed-RRP6_vs_WT.odf
	☑	yeast-rnaseq.preprocessed_volcanoplots.pdf
	☑	stdout.txt
	☑	gp_execution_log.txt

download | delete

These job results are scheduled to be purged from the server on 12/18/10 11:00 pm

Return to Modules & Pipelines Start

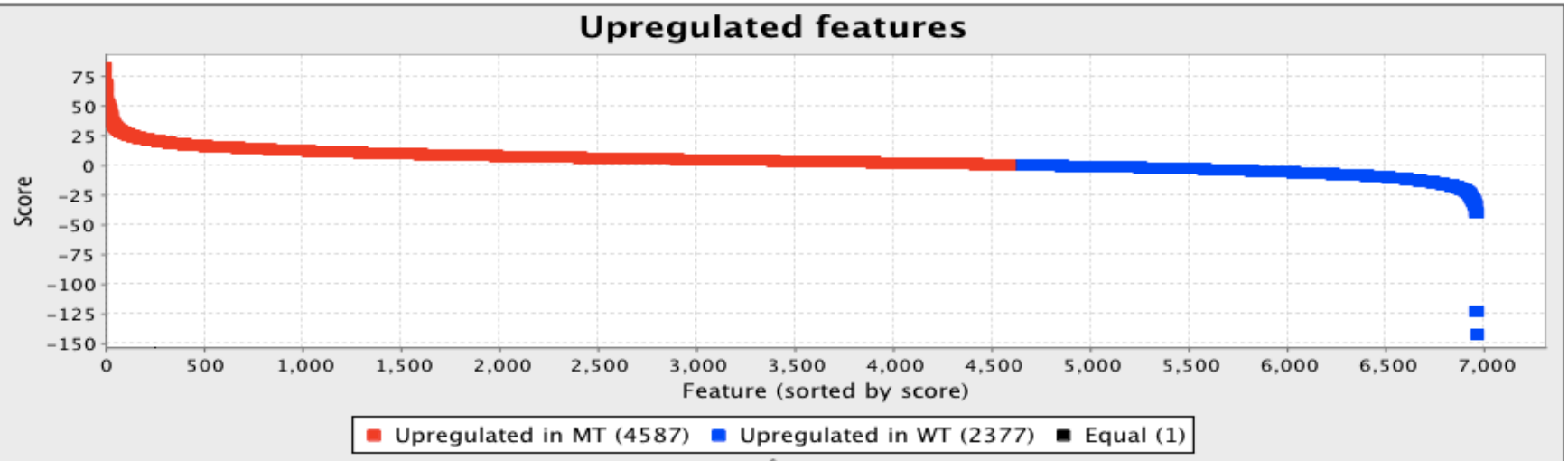
CC BY NC SA


GenePattern
My Settings | Sign Out Mik Black
Modules & Pipelines
Suites
Job Results
Resources
Downloads
Help
show execution logs
Job Results
sort jobs by: job id number
sort files by: name
Sort
delete open all | close all
job
module/file
size
submitted
completed
status
6275
Limma
Oct 19 11:00:32 PM
Oct 19 11:00:34 PM
finished
yeast-rnaseq.preprocessed_volcanoplot-WT.svg
29.0 KB
Oct 19 11:00:33 PM
yeast-rnaseq.preprocessed_volcanoplot-RRP6_vs_WT.svg
30.0 KB
Oct 19 11:00:33 PM
yeast-rnaseq.preprocessed-WT.odf
29.0 KB
Oct 19 11:00:33 PM
yeast-rnaseq.preprocessed-RRP6_vs_WT.odf
29.0 KB
Oct 19 11:00:33 PM
stdout.txt
1.0 KB
Oct 19 11:00:34 PM
gp_execution_log.txt
1.0 KB
Oct 19 11:00:34 PM

CC BY NC SA

Comparative Marker Selection – yeast-rnaseq-allchroms.preprocessed-MT_vs_WT.odf

File Edit View GeneCruiser Window



Row	Rank	Upregulate...	Feature	Description	Score	Feature P	FDR(BH)	Q Value	Bonferroni	Fold Change	B statistic
1	1	WT	PAT1	PAT1	-142.193	0	0	0	0	-10.608	16.968
2	2	WT	RRP6	RRP6	-123.194	0	0	0	0	-8.608	16.615
3	3	MT	YJR129C	YJR129C	81.955	0	0	0	0	6.163	15.217
4	4	MT	YMR013W-A	YMR013...	71.05	0	0	0	0	5.87	14.587
5	5	MT	YJL047C-A	YJL047C-A	68.042	0	0	0	0	4.4	14.382
6	6	MT	YMR194C-B	YMR194...	65.374	0	0	0	0	4.238	14.187
7	7	MT	YGL088W	YGL088W	59.436	0	0	0	0	6.182	13.703
8	8	MT	RLP7	RLP7	54.551	0	0	0	0	4.197	13.245
9	9	MT	CSN9	CSN9	54.039	0	0	0	0	3.612	13.193
10	10	MT	SAS10	SAS10	52.639	0	0	0	0	3.53	13.048
11	11	MT	RPA34	RPA34	52.574	0	0	0	0	4.058	13.041
12	12	MT	URA7	URA7	51.366	0	0	0	0	3.136	12.911
13	13	MT	LIN1	LIN1	51.304	0	0	0	0	6.362	12.904
14	14	MT	YER158W-A	YER158...	50.545	0	0	0	0	3.119	12.82
15	15	MT	YBL028C	YBL028C	50.041	0	0	0	0	4.006	12.763
16	16	MT	VMA22	VMA22	50.04	0	0	0	0	3.266	12.763
17	17	MT	LOC1	LOC1	49.847	0	0	0	0	3.846	12.741
18	18	MT	YPR092W	YPR092W	49.708	0	0	0	0	4.345	12.725
19	19	MT	UTP20	UTP20	49.441	0	0	0	0	3.878	12.694
20	20	MT	GLO4	GLO4	47.742	0	0	0	0	4.473	12.493
21	21	MT	MMS2	MMS2	47.447	0	0	0	0	3.347	12.457
22	22	MT	RRP14	RRP14	47.354	0	0	0	0	4.358	12.445

Features: 6965

cc BY NC SA

Langmead *et al. Genome Biology* 2010, **11**:R83
http://genomebiology.com/content/11/8/R83



SOFTWARE **Open Access**

Cloud-scale RNA-sequencing differential expression analysis with Myrna

Ben Langmead, Kasper D Hansen, Jeffrey T Leek*

Abstract

As sequencing throughput approaches dozens of gigabases per day, there is a growing need for efficient software for analysis of transcriptome sequencing (RNA-Seq) data. Myrna is a cloud-computing pipeline for calculating differential gene expression in large RNA-Seq datasets. We apply Myrna to the analysis of publicly available data sets and assess the goodness of fit of standard statistical models. Myrna is available from http://bowtie-bio.sf.net/myrna.

CC BY NC SA