
JOHN DEERE
8R

NEXT GENERATION SEQUENCING (NGS)

Andreas Gisel

Institute of Biomedical Technologies (ITB)
National Research Council (CNR)
Bari, Italy
andreas.gisel@ba.itb.cnr.it

International Institute of Tropical Agriculture (IITA)
Ibadan, Nigeria
a.gisel@cgiar.org

Bioinformatics specialist at

International Institute of Tropical Agriculture, Nigeria

Institute for Biomedical Technologies, Italy

Novartis SA, Pharmaceutical Company, Switzerland

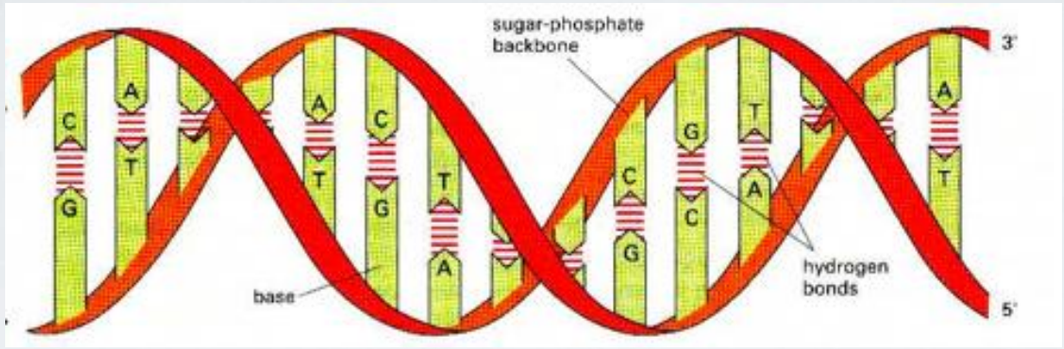
Trained as Molecular Biologist

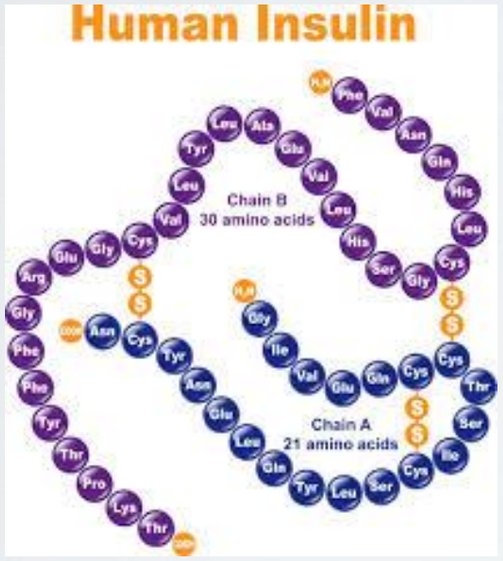
Novartis SA, Pharmaceutical Company, Switzerland

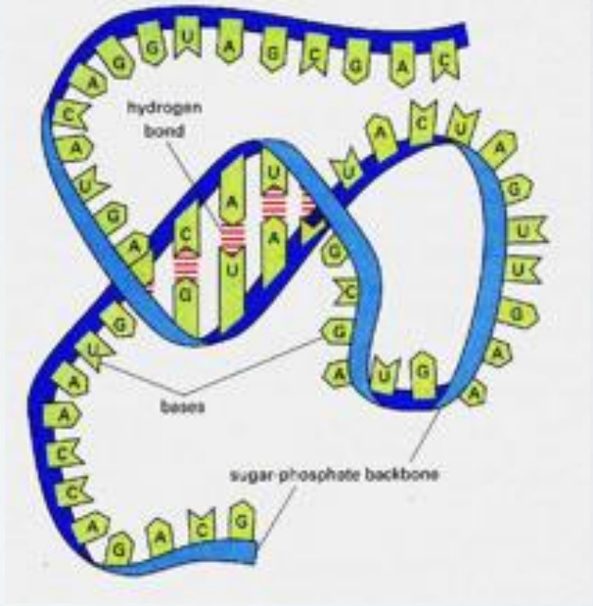
University of California, Berkeley

Federal Institute of Technology, Switzerland

What can you sequence?







SEQUENCING

First fully sequenced bio-sequence

- amini acid of insulin (51aa) 1955

First fully sequence nucleic acid

- tRNA (75nt) 1965

First DNA

- Bacteriophage (5375nt) 1977

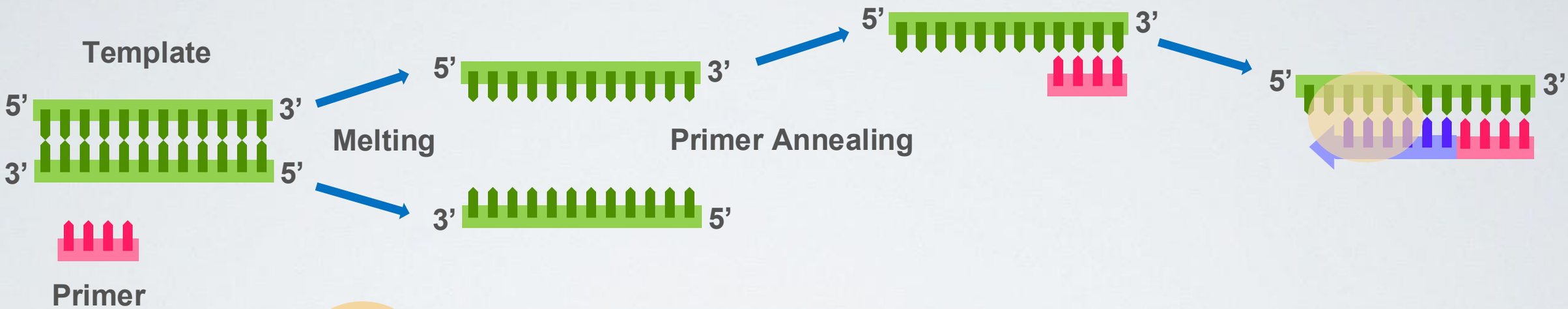
DNA sequencing

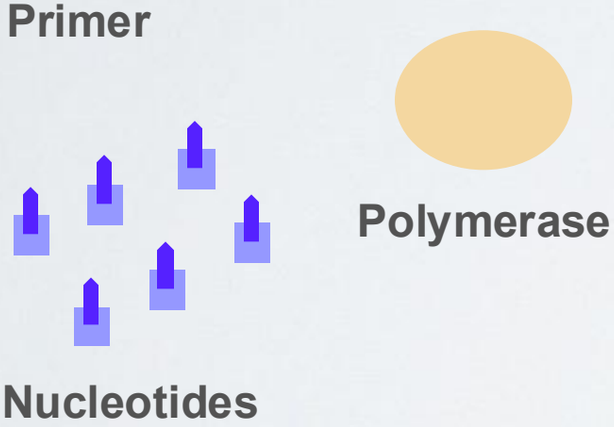
- Sanger sequencing technology (1975)

SEQUENCING

Polymerase Chain Reaction (PCR)

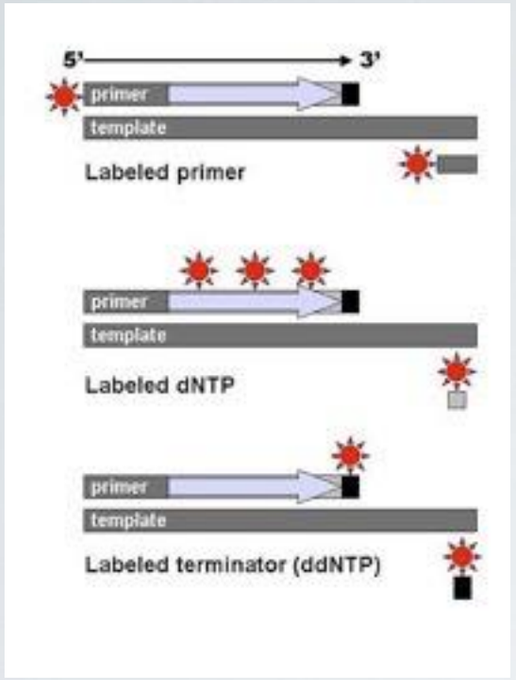
To amplify identical copies of DNA fragments

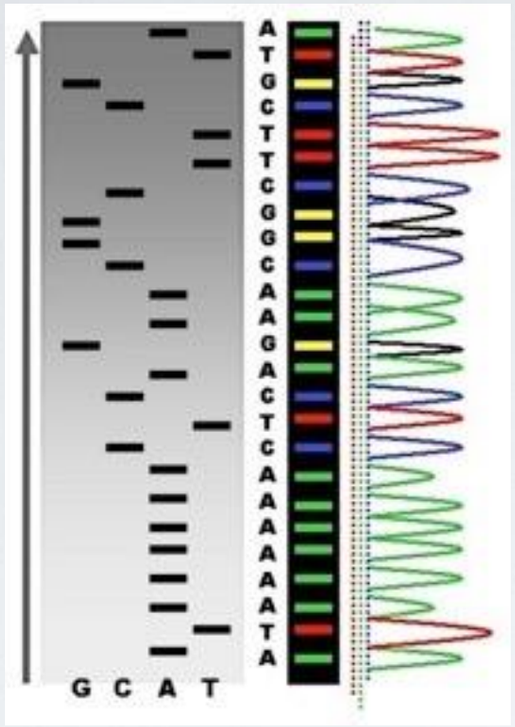


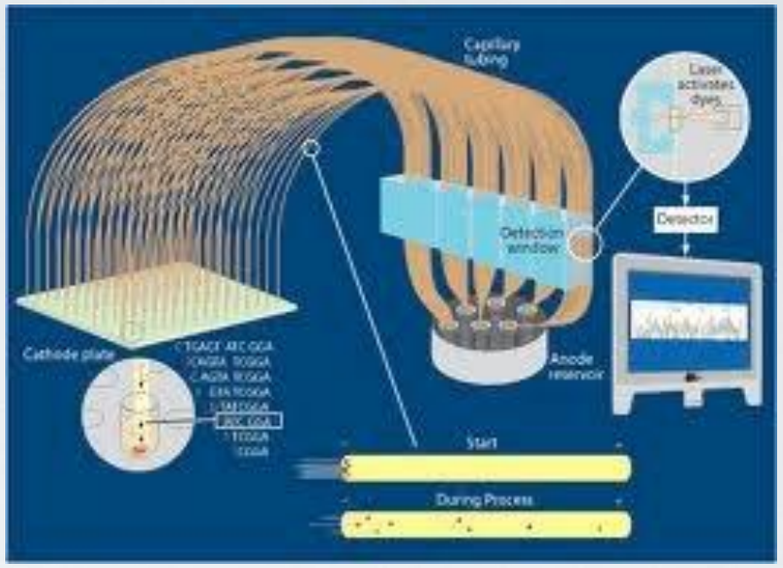


PCR was invented in 1983 by American biochemist Kary Mullis at Cetus Corporation. Mullis and biochemist Michael Smith, who had developed other essential ways of manipulating DNA, were jointly awarded the Nobel Prize in Chemistry in 1993.

sequencing by chain-termination method



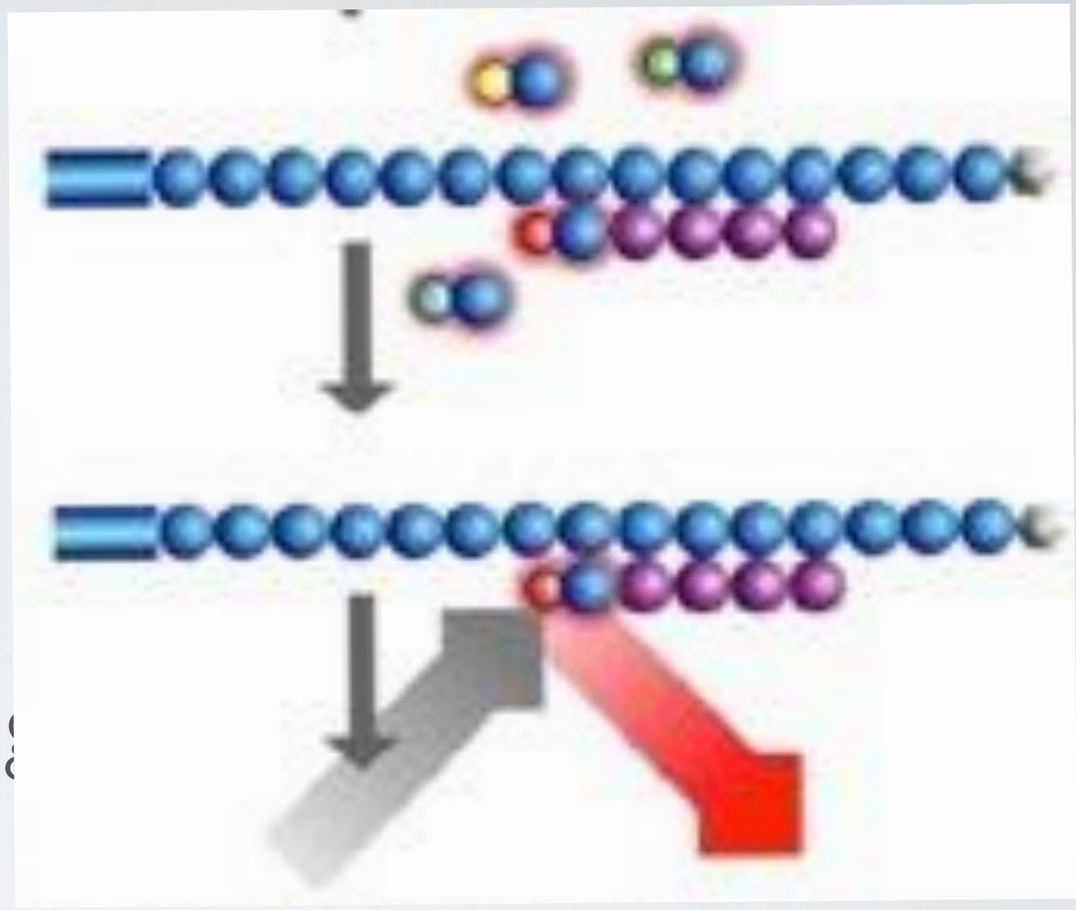




- DNA sequencing by capillary electrophoresis
- 384 reactions in parallel
- sequences up to 1000nt

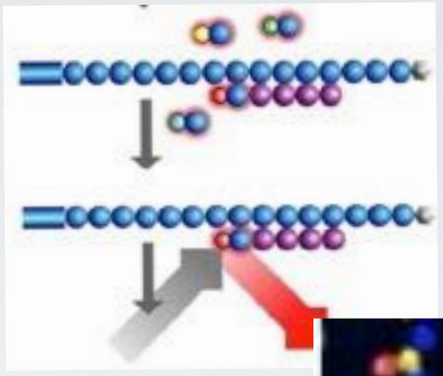
SANGER TECHNOLOGY

- Sequencing by synthesis
- highly parallelized sequencing
- Paired-end sequencing

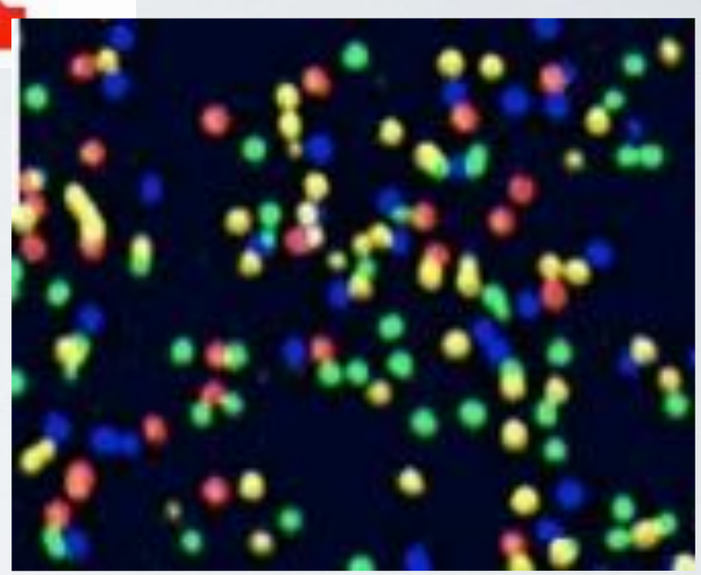


NEXT GENERATION SEQUENCING

- Sequencing by synthesis



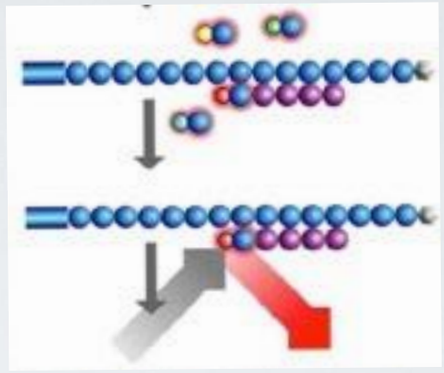
- highly parallelized sequencing



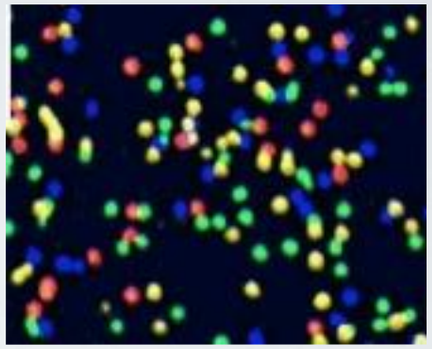
- Paired-end sequencing

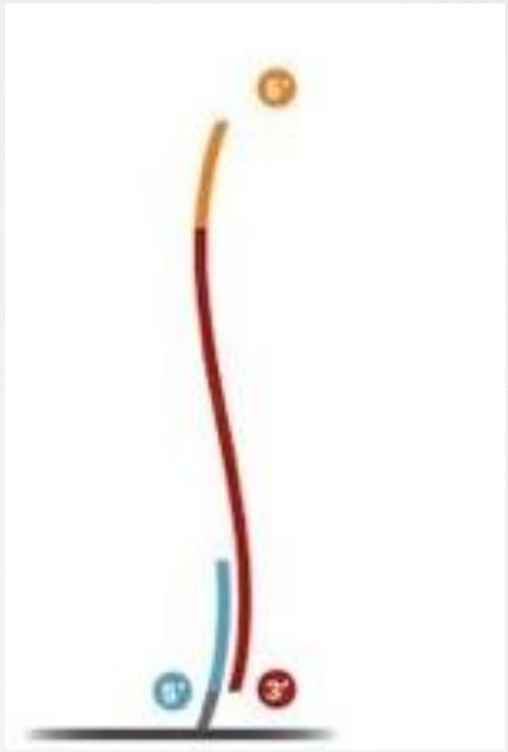
NEXT GENERATION SEQUENCING

- Sequencing by synthesis



- highly parallelized sequen





- Paired-end sequencing

NEXT GENERATION SEQUENCING

Amplification steps

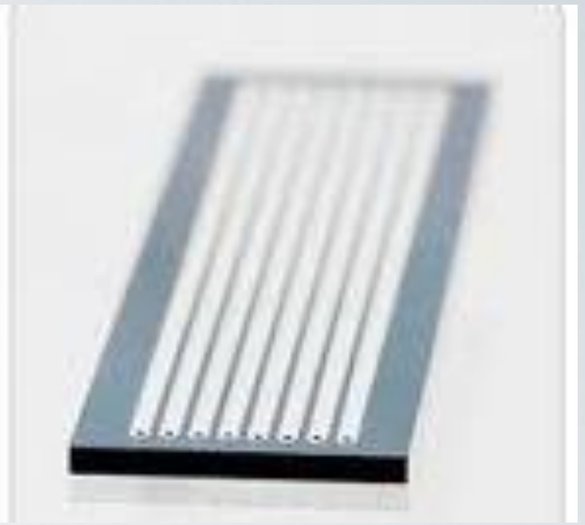
- 454 Roche
- **Illumina**
- SOLiD Applied Biosystems

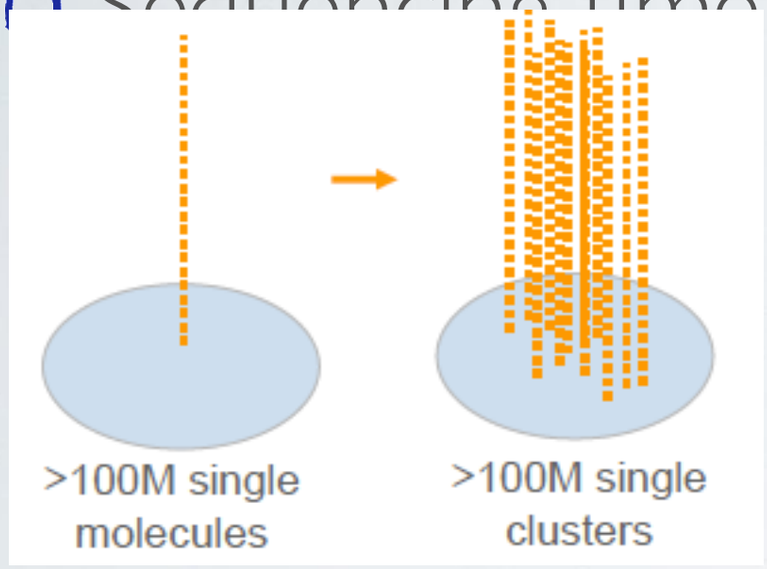
Single molecule

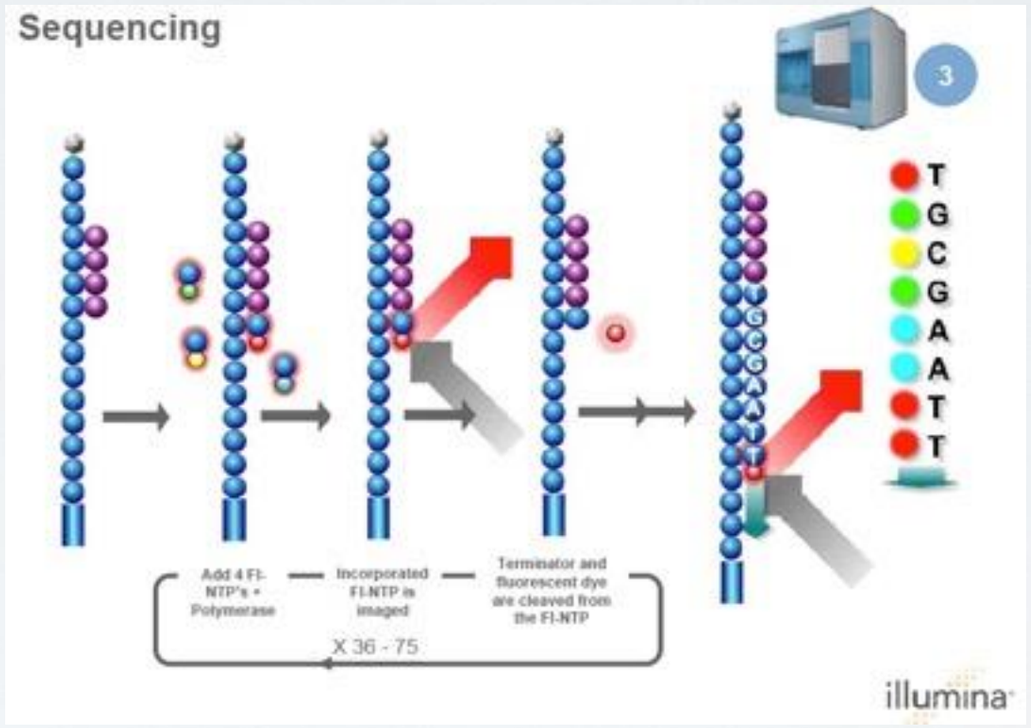
- **Pacific BioSciences**
- **Ion Torrent**
- **Nanopore**

TECHNOLOGIES

- DNA template immobilized to a flow cell
- Cluster formation by Bridge PCR
- Sequencing on flow cell (26B reads)
- Sequencing by synthesis (protected nts)
- Sequencing by cycling
- Sequencing time in the range of days







ILLUMINA

- sequencing length up to 300nt – optimal 150nt
- up to 26B sequences per flow cell - high coverage
- 1.2 – 8000 Gb per run
-

Key specifications	iSeq 100 System	MiniSeq System	MiSeq System	MiSeq i100 Series	NextSeq 550 System	NextSeq 1000 and 2000 Systems
Max output per flow cell	1.2 Gb [b]	7.5 Gb [c]	15 Gb [d]	30 Gb [a]	120 Gb [c]	540 Gb [e]
Run time (range)[e]	~9.5–19 hr	~5–24 hr	~5.5–56 hr	~4–15.5 hr	~11–29 hr	~8–44 hr
Max reads per run (single reads)	4M [ab]	25M [c]	25M [d]	100M [a]	400M [c]	1.8B [e]
Max read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 300 bp	2 × 150 bp	2 × 300 bp

- sequencing length up to 300nt – optimal 150nt
- up to 26B sequences per flow cell - high coverage
- 1.2 – 8000 Gb per run
-

New

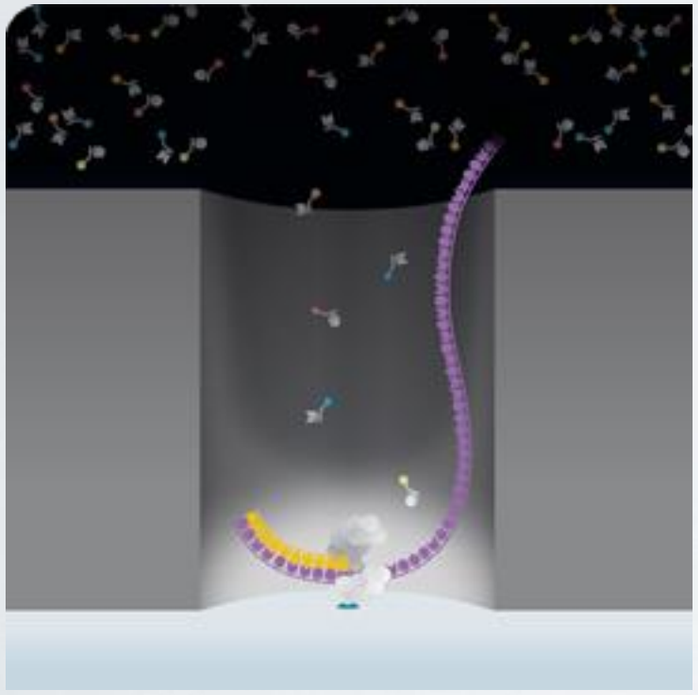
Key specifications	 NextSeq 1000 and 2000 Systems	 NovaSeq 6000 System	 NovaSeq X Series
Max output per flow cell	540 Gb[a]	3 Tb[b]	8 Tb[c]
Run time (range)[d]	~8–44 hr	~13–44 hr	~17–48 hr
Max reads per run (single reads)	1.8B[a]	10B (single flow cell)[b] 20B (dual flow cells)	26B (single flow cell)[c] 52B (dual flow cells)[c,e]
Max read length	2 × 300 bp	2 × 250 bp	2 × 150 bp

Max read length	2 × 150 bp	2 × 150 bp	2 × 300 bp	2 × 300 bp	2 × 150 bp	2 × 300 bp

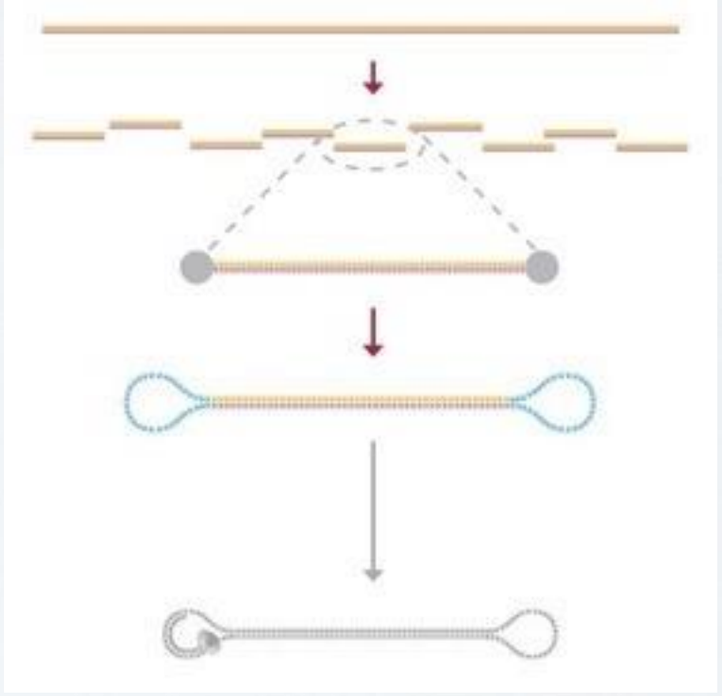
- Sequence fragmentation
- Adaptor ligation
- DNA template immobilization
- PCR amplification
- Real-time sequencing (by synthesis / cycling)
- huge amount of short sequence reads
- high coverage
- difficulties with assembling

SUMMARY

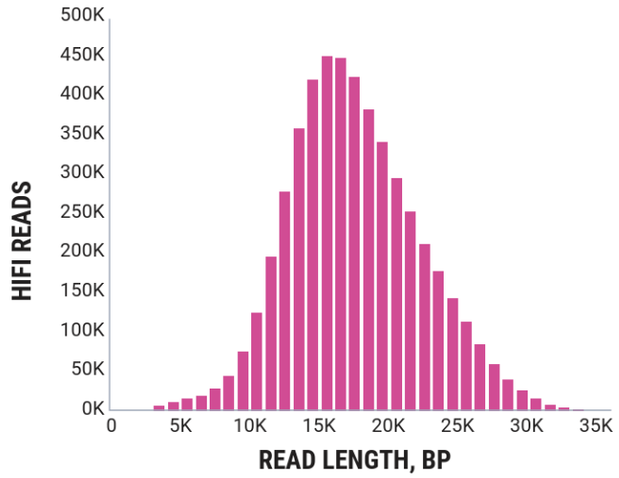
- Polymerase immobilized in a nano well
- NO amplification (true single molecule sequencing)
- Sequencing on flow cell (max 1M reads per cell)
- Sequencing by synthesis (fluorescence)
- No cycling
- Read length up to 25000nt average >19000nt
- Fast sample preparation and sequencing (24h)

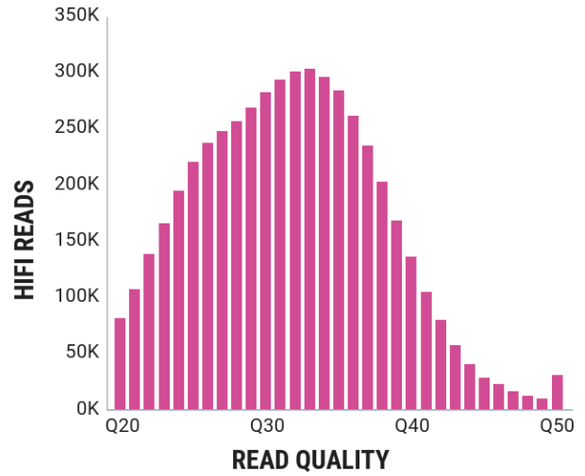


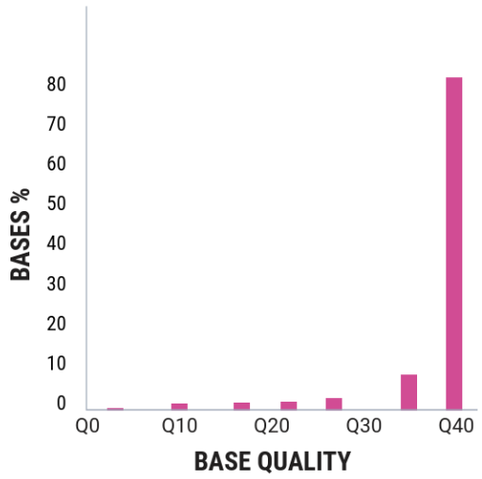
Zero-mode waveguide



PACIFIC BIOSCIENCES

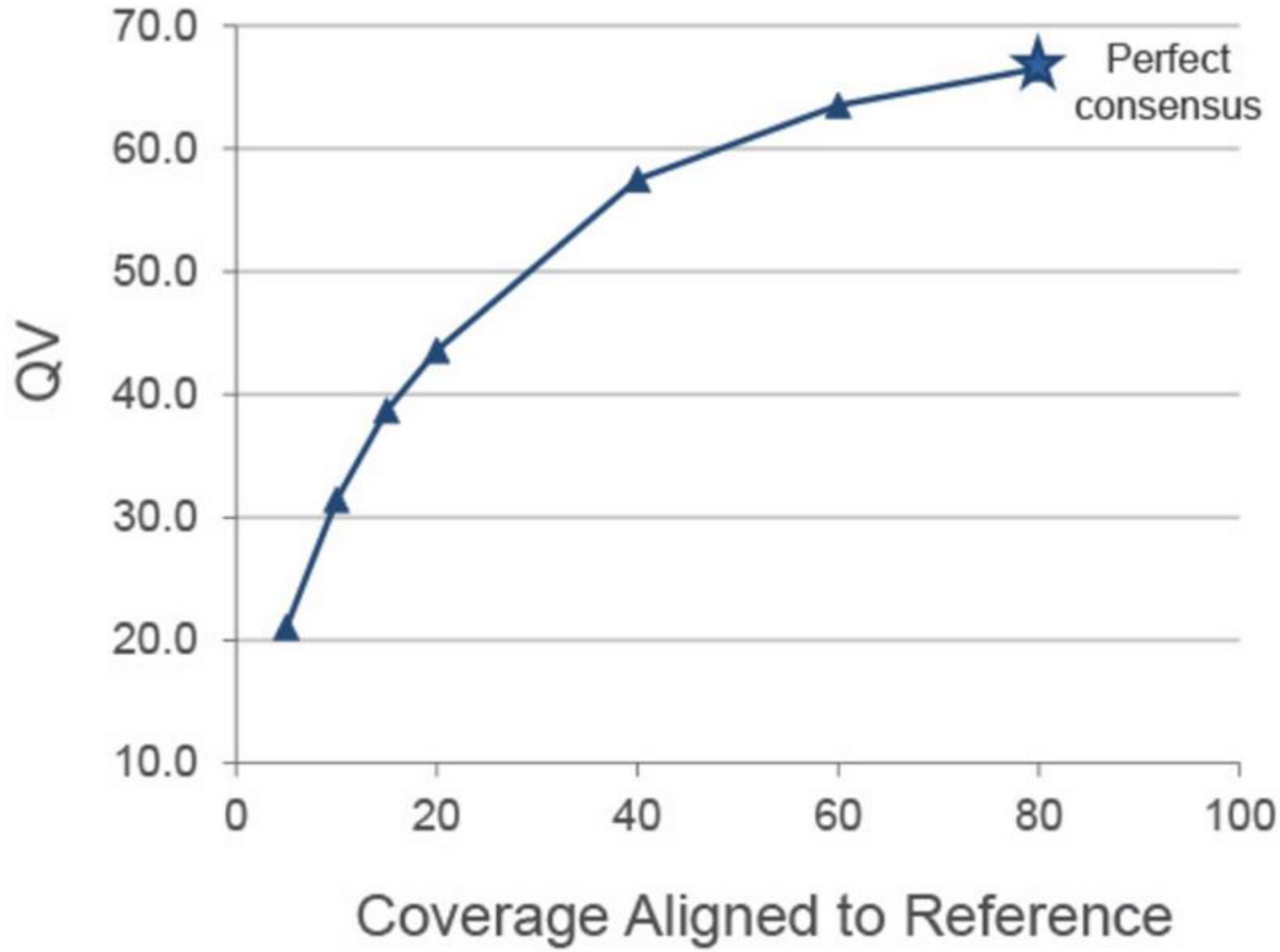






Data shown is for a single Revio SMRT Cell for HG002/GM24385 sequenced with SPRQ chemistry.[5]

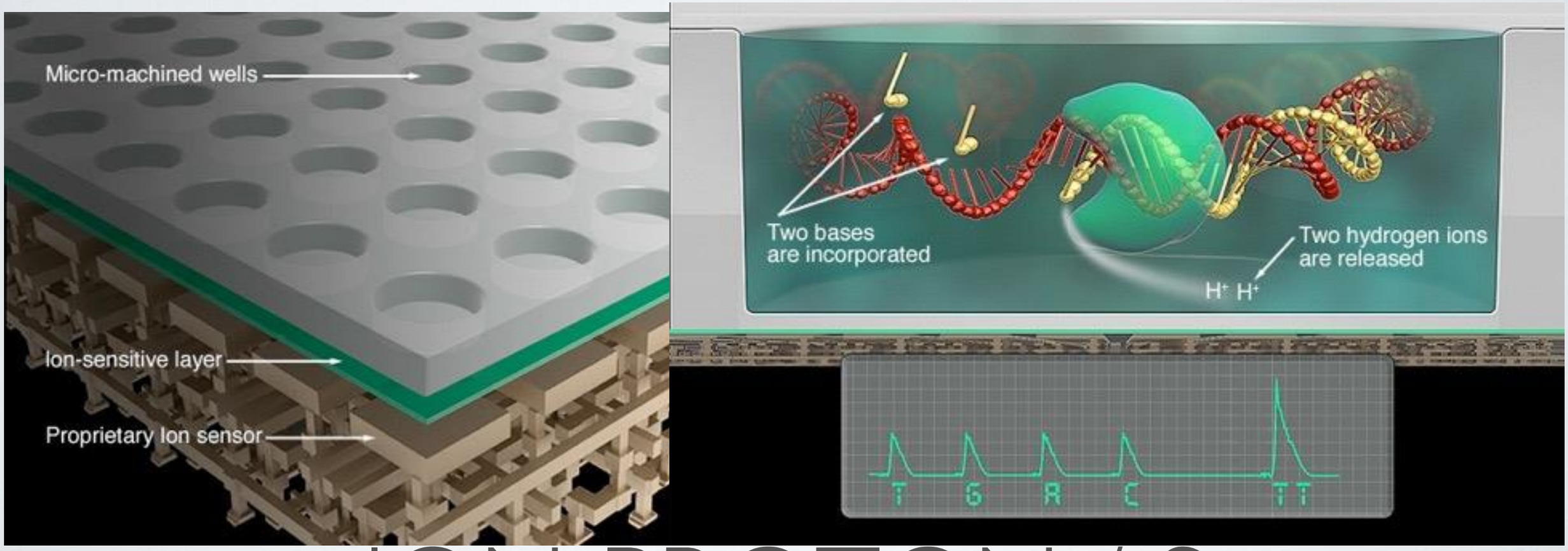
PACIFIC BIOSCIENCES



SMRT Sequencing accuracy is a function of coverage and chemistry. The random distribution of single-pass errors allows consensus accuracy to rapidly build with coverage, and can even lead to perfect consensus as shown above. The P6-C4 chemistry estimates shown here are based on multiple bacterial genomes.

PACIFIC BIOSCIENCES

- DNA template immobilized in a nano well (semiconductor)
- NO amplification (true single molecule sequencing)
- Sequencing on flow cell (40 - 80M reads)
- Sequencing by synthesis (H+ release)
- Read length up to 400
- Fast sample preparation and sequencing (8h)



ION PROTON / S

- Immobilized in a nano well (semiconductor)
- NO amplification (true single molecule sequencing)
- Sequencing on flow cell (40 - 80M reads)
- se)

The Ion Proton™ System

Rapid, high-throughput benchtop sequencing

The Ion Proton™ System minimizes the high cost and complexity of high-throughput sequencing, bringing it to your research lab, on your budget, and on your schedule. This system enables high-quality exome and transcriptome sequencing in just a few days.

Ion Proton™ System performance specifications with Ion PI™ Chip	
Output	Up to 10 Gb
Reads	60–80 million reads
Read length	Up to 200-base reads
Run time	2–4 hours

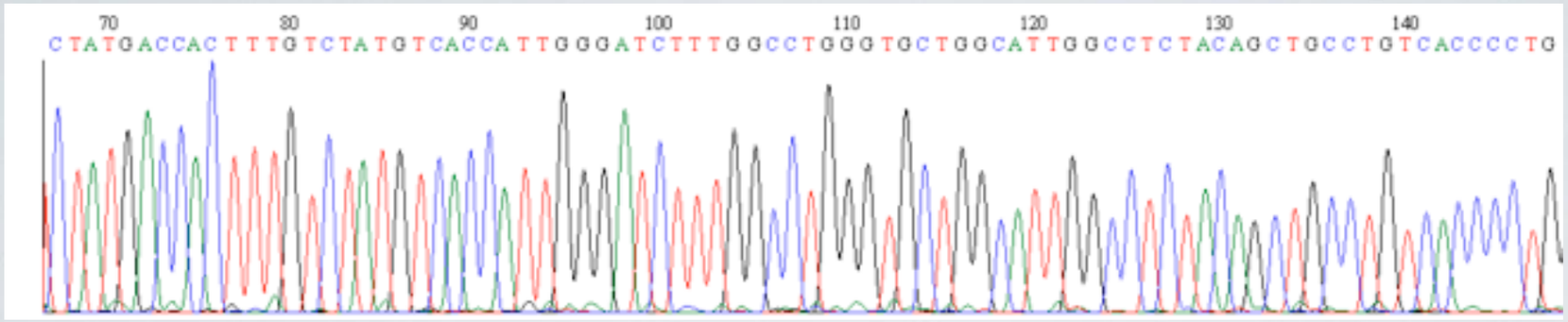
		Ion S5 System: Simple workflow for panels, microbes, exomes, and transcriptomes			Ion S5 XL System: Simple, rapid workflow for panels, microbes, exomes, and transcriptomes		
		Ion 520 Chip	Ion 530 Chip	Ion 540 Chip	Ion 520 Chip	Ion 530 Chip	Ion 540 Chip
Reads		3–5 million	15–20 million	60–80 million	3–5 million	15–20 million	60–80 million
Output*	200 bp	0.6–1 Gb	3–4 Gb	10–15 Gb	0.6–1 Gb	3–4 Gb	10–15 Gb
	400 bp	1.2–2 Gb	6–8 Gb	—	1.2–2 Gb	6–8 Gb	—
Run times	200 bp	2.5 hr	2.5 hr	2.5 hr	2.5 hr	2.5 hr	2.5 hr
	400 bp	4 hr	4 hr	—	4 hr	4 hr	—

ION PROTON / S

- **FASTQ (*.fastq)**

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+
IIIIIIIIIIIIIIIIIIAIIIIIIIIII&;II&,I
```

DATA FORMAT

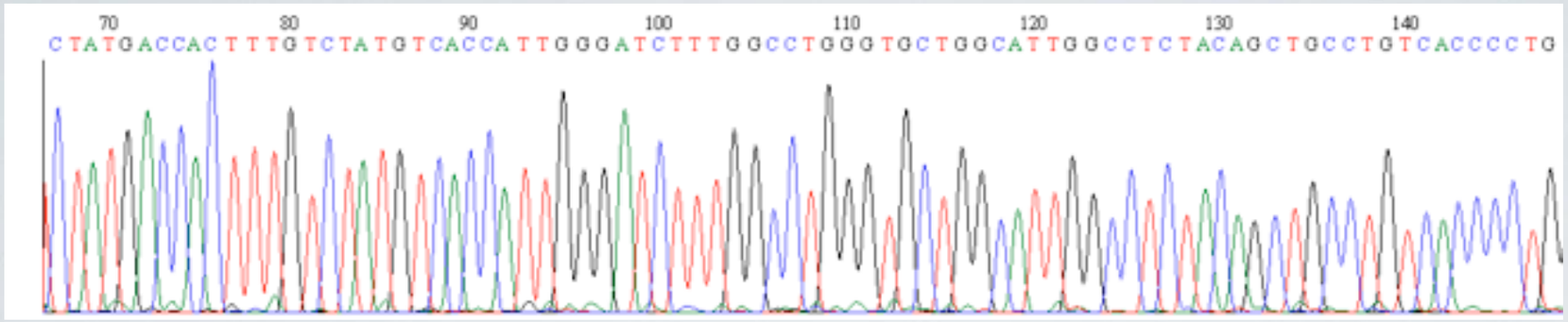


Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+
IIIIIIIIIIIIIIIIAIIIIIIIII&;II&,I
```

QUALITY SCORE



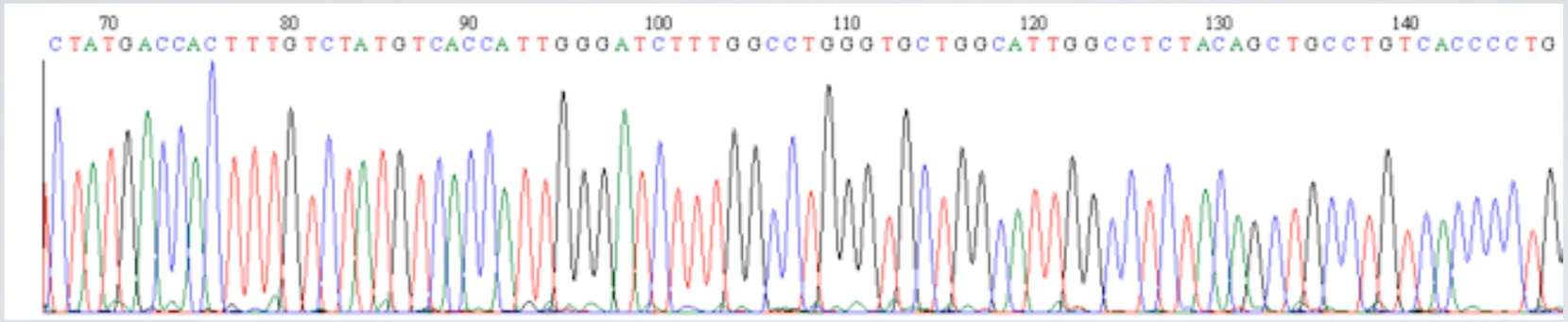
Quality scores are cu
chromatogram; well-k

They range from 0 to
by ASCII code.

```
@HWI-EAS
TTAATCTA
+
IIIIIIII
```

Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char	Dec	Hex	Char
0	00	NUL	25	19	EM	51	33	3	77	4D	M	103	67	g
1	01	SOH	26	1A	SUB	52	34	4	78	4E	N	104	68	h
2	02	STX	27	1B	ESC	53	35	5	79	4F	O	105	69	i
3	03	ETX	28	1C	FS	54	36	6	80	50	P	106	6A	j
4	04	EOT	29	1D	GS	55	37	7	81	51	Q	107	6B	k
5	05	ENQ	30	1E	RS	56	38	8	82	52	R	108	6C	l
6	06	ACK	31	1F	US	57	39	9	83	53	S	109	6D	m
7	07	BEL	32	20	space	58	3A	:	84	54	T	110	6E	n
8	08	BS	33	21	!	59	3B	;	85	55	U	111	6F	o
9	09	HT	34	22	"	60	3C	<	86	56	V	112	70	p
10	0A	LF	35	23	#	61	3D	=	87	57	W	113	71	q
11	0B	VT	36	24	$	62	3E	>	88	58	X	114	72	r
12	0C	FF	37	25	%	63	3F	?	89	59	Y	115	73	s
13	0D	CR	38	26	&	64	40	@	90	5A	Z	116	74	t
14	0E	SO	39	27	'	65	41	A	91	5B	[	117	75	u
15	0F	SI	40	28	(	66	42	B	92	5C	\	118	76	v
16	10	DLE	41	29	)	67	43	C	93	5D	]	119	77	w
17	11	DC1	42	2A	*	68	44	D	94	5E	^	120	78	x
18	12	DC2	43	2B	+	69	45	E	95	5F	_	121	79	y
19	13	DC3	44	2C	,	70	46	F	96	60	`	122	7A	z
20	14	DC4	45	2D	-	71	47	G	97	61	a	123	7B	{
21	15	NAK	46	2E	.	72	48	H	98	62	b	124	7C	|
22	16	SYN	47	2F	/	73	49	I	99	63	c	125	7D	}
23	17	ETB	48	30	0	74	4A	J	100	64	d	126	7E	~
24	18	CAN	49	31	1	75	4B	K	101	65	e	127	7F	DEL
			50	32	2	76	4C	L	102	66	f			

Q



Quality scores are currently calculated to reliably call bases from a Sanger chromatogram; well-known as Phred scores.

They range from 0 to 93 (Illumina 0 - 40), even though rarely exceed 60; represented by ASCII code.

```
@HWI-EAS255_4_FC2010Y_1_43_110_790
TTAATCTACAGAATAGATAGCTAGCATATATTT
+
IIIIIIIIIIIIIIIIAIIIIIIIII&;
```

Sanger / Illumina quality code (Phred): ASCII character c

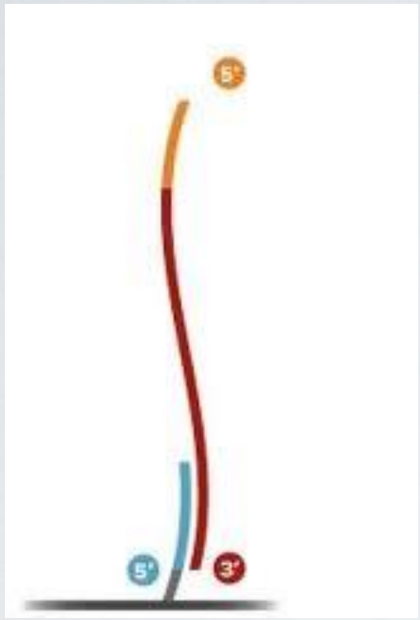
Quality Value	Error Probability	Probability Called Base is Correct
10	0.1	0.9
20	0.01	0.99
30	0.001	0.999
40	0.0001	0.9999

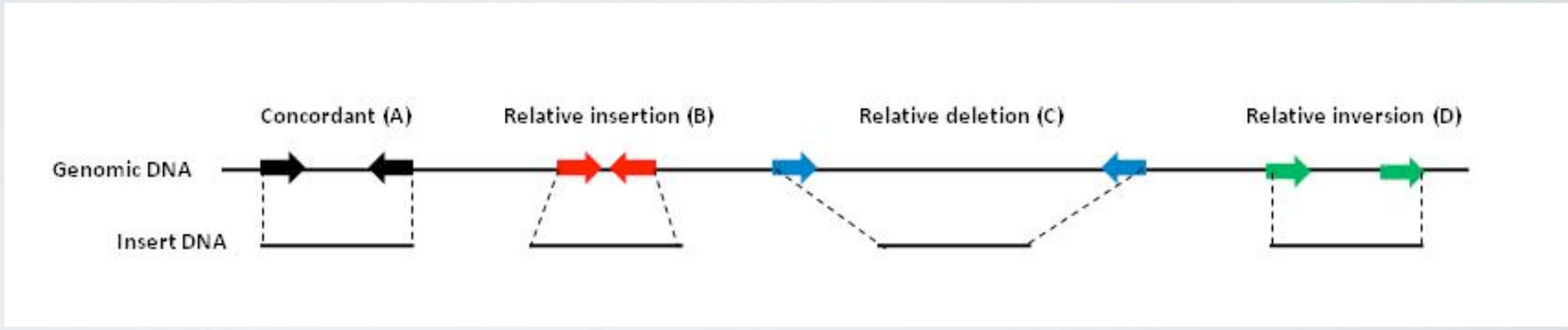
Illumina: `I` (ASCII 73) => 73 - 33 = 40
Illumina: `&` (ASCII 38) => 38 - 33 = 5

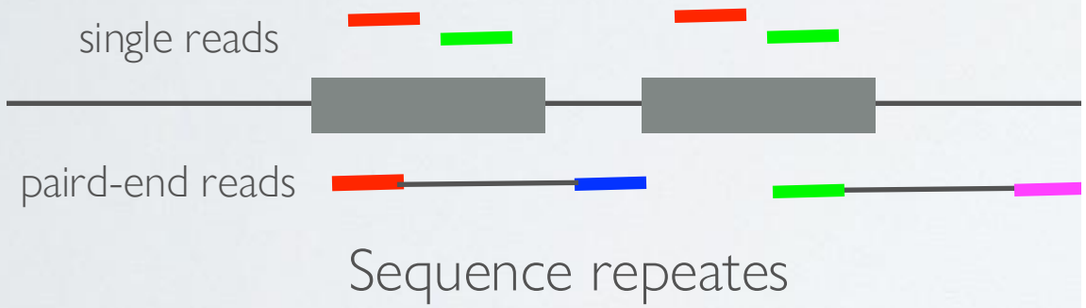
$q=-10\log_{10}(p)$

QUALITY SCORE

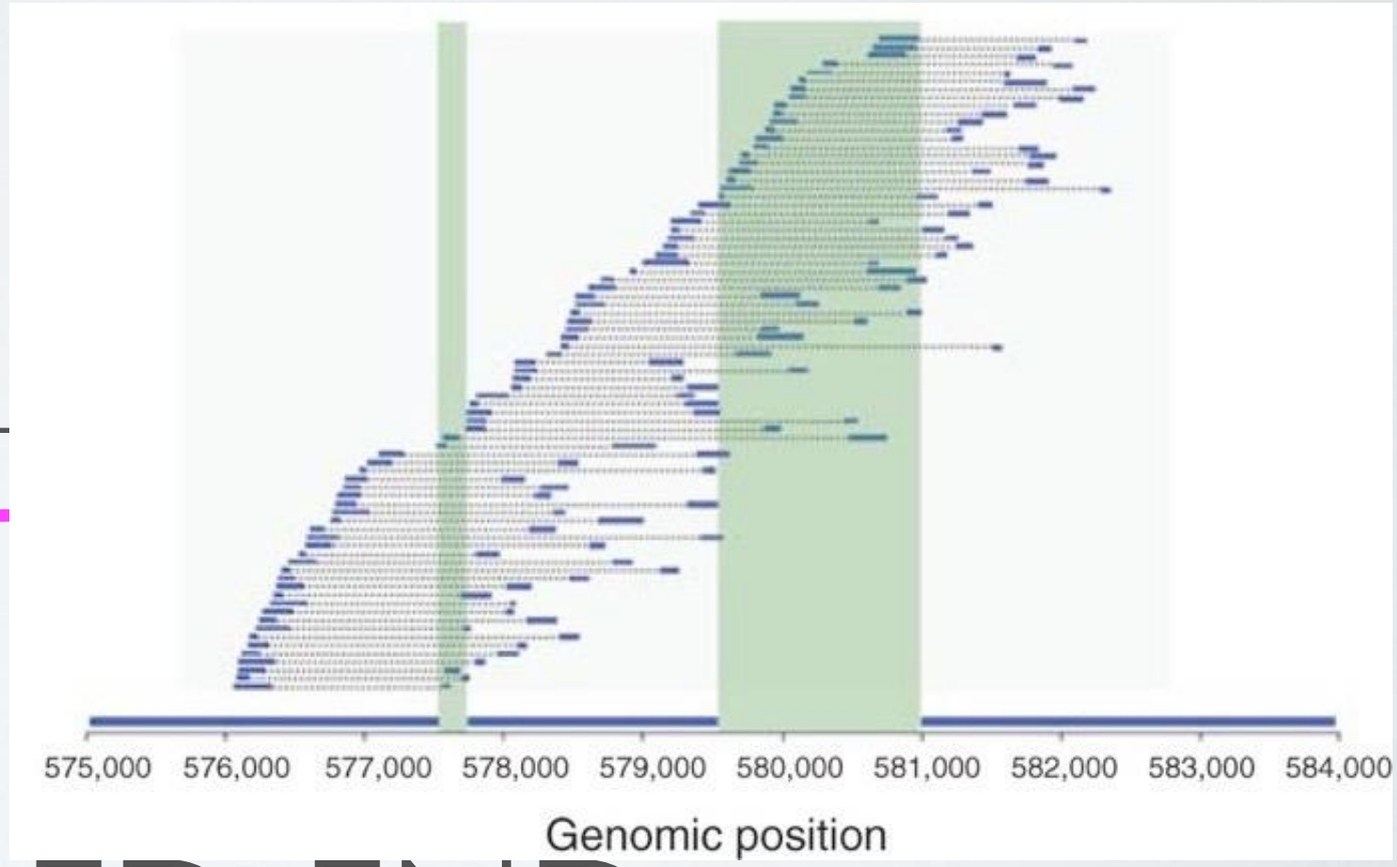
Sequencing of the two end of the same DNA fragment







Sequence repeates



PAIRED-END

Sample Prep

Whole genome
- Resequencing → sequence variations such as SNP, CNV, inserts, deletions, reversions
- De-novo → new genomes
- Targeted → sequence variations with higher coverage
- Metagenomics → environmental studies, community studies

Transcriptome
- RNA-Seq → transcriptomics, splicing variants,
- DGE → digital gene expression
- Small RNA → non-coding RNA research
- miRNA → miRNA induced regulations

Regulation
- Methylation → epigenetics, methylation induced regulations
- ChIP-Seq → protein DNA interactions such as TFBS, histon, polymerase

APPLICATIONS

- Quality control

- Mapping
- Assembly
- Digital gene expression

DATA ANALYSIS

- Quality control
  - Quality score distribution
  - Sequence size distribution
  - Sequence coverage
  - Adaptor search
- Mapping
- Assembly
- Digital gene expression

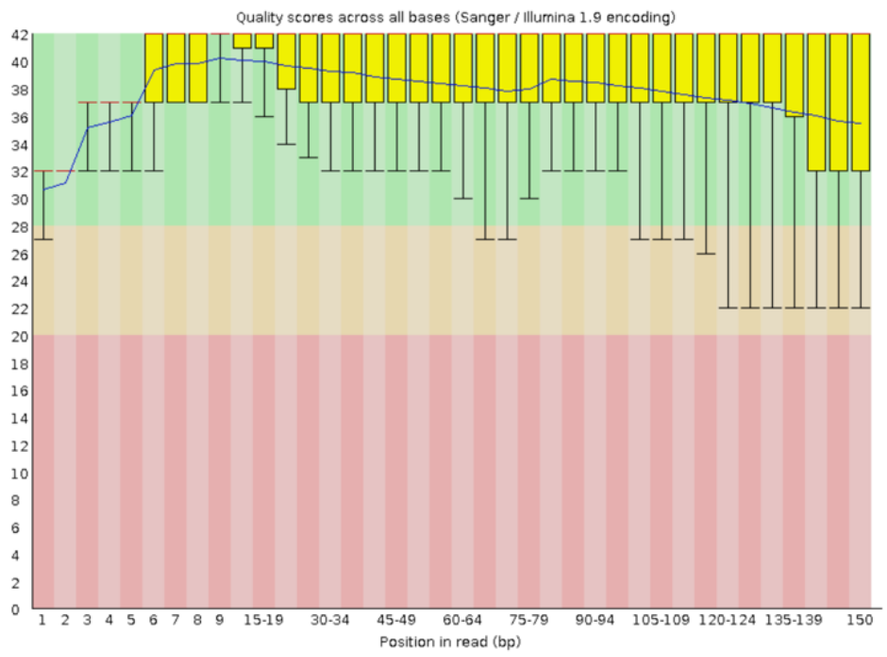
DATA ANALYSIS

- Quality control
  - Quality score distribution
  - Sequence size distribu

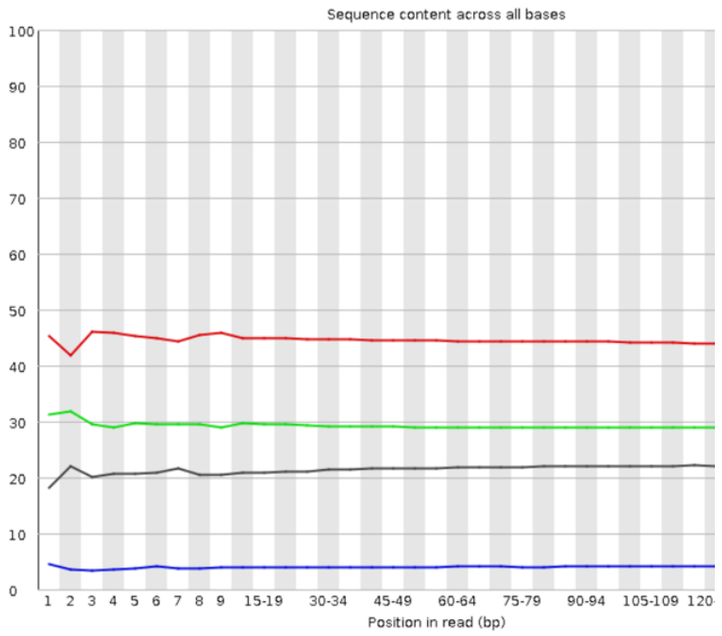
Basic Statistics

Measure	Value
Filename	FCHF2VWALXX_L1_WHCASztdHAAADEAAPEI-53_1.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	15344560
Sequences flagged as poor quality	0
Sequence length	150
%GC	26

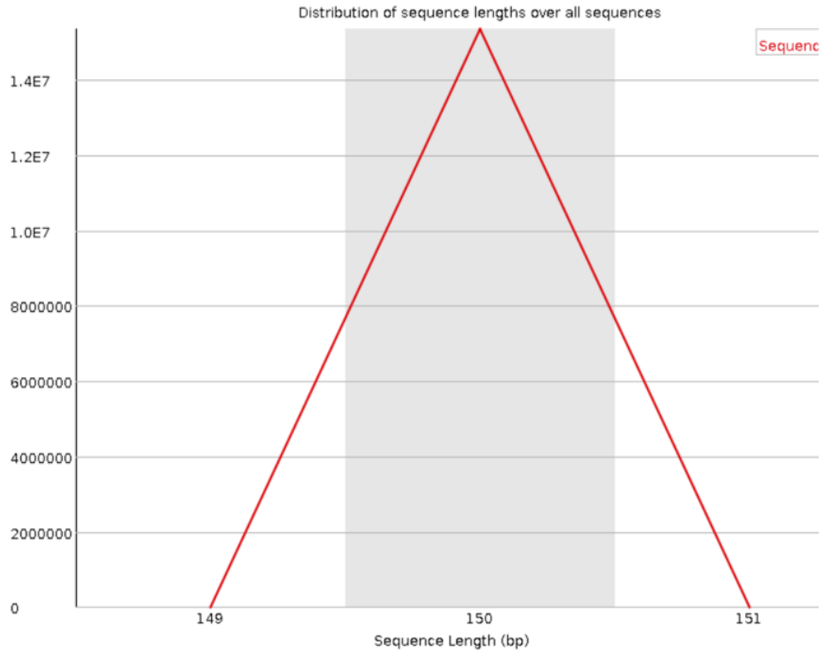
Per base sequence quality



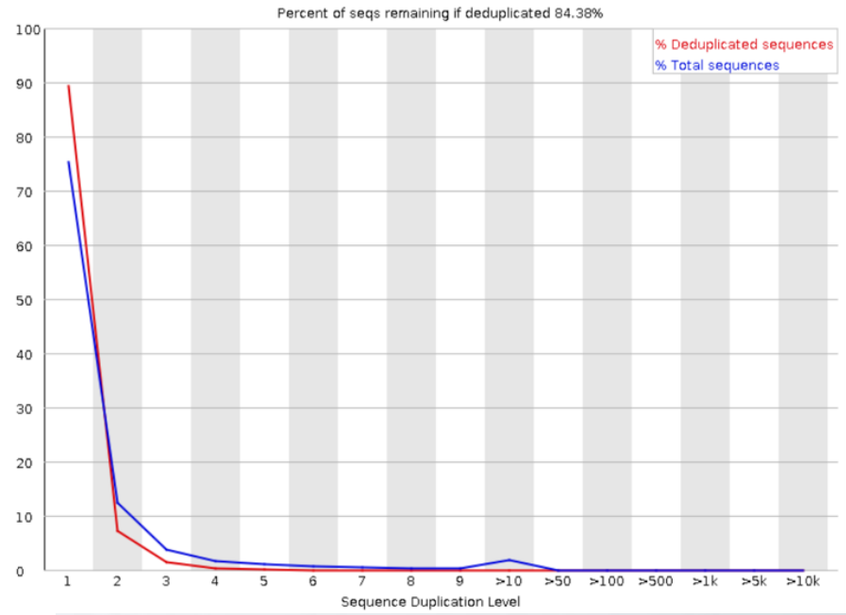
Per base sequence content



Sequence Length Distribution



Sequence Duplication Levels

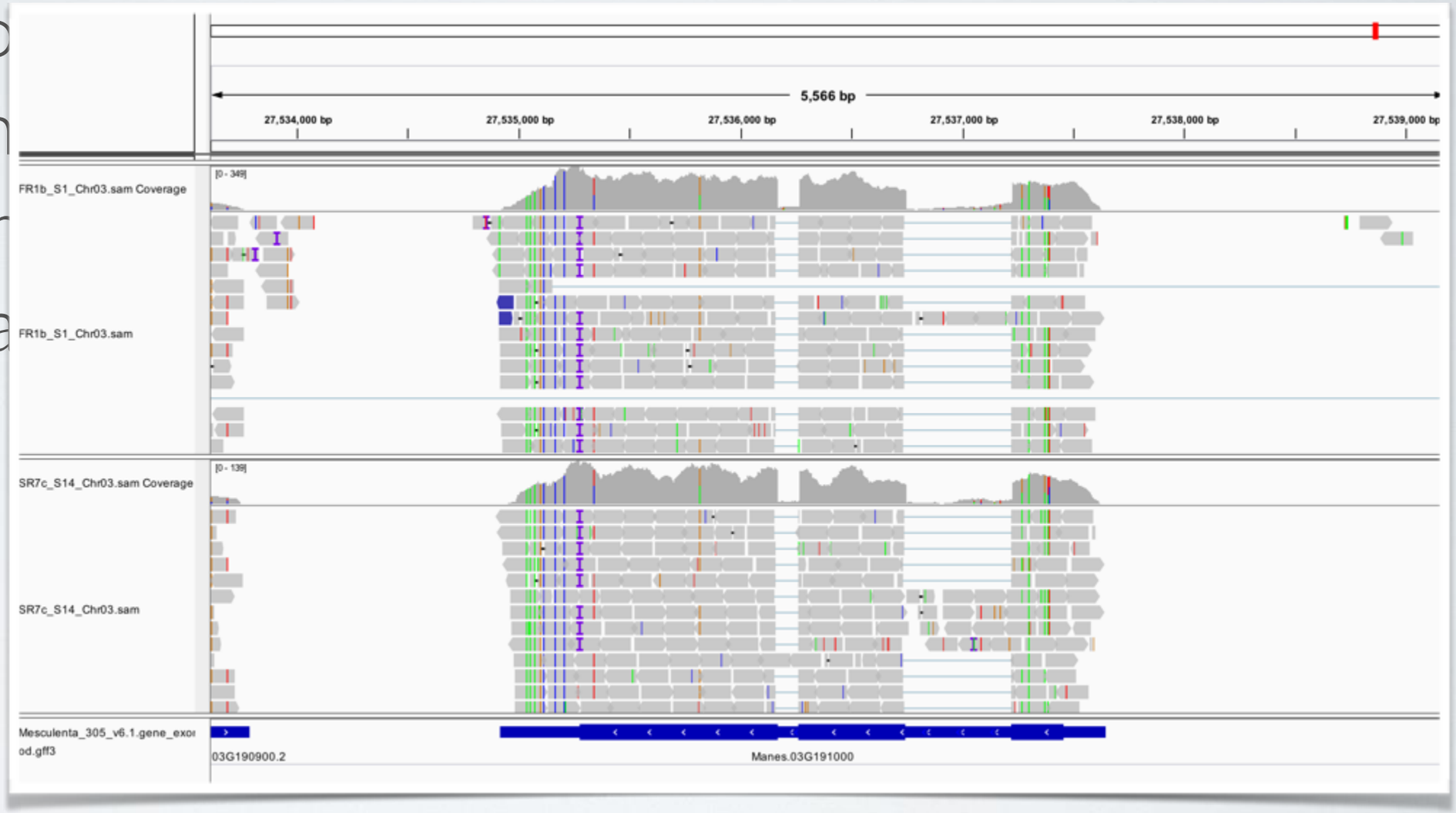


- Quality control

- Mapping
  - Aligning short sequences on a reference
- Assembly
- Digital gene expression

DATA ANALYSIS

- Quality control
- Mapp
  - Align
- Assem
- Digita



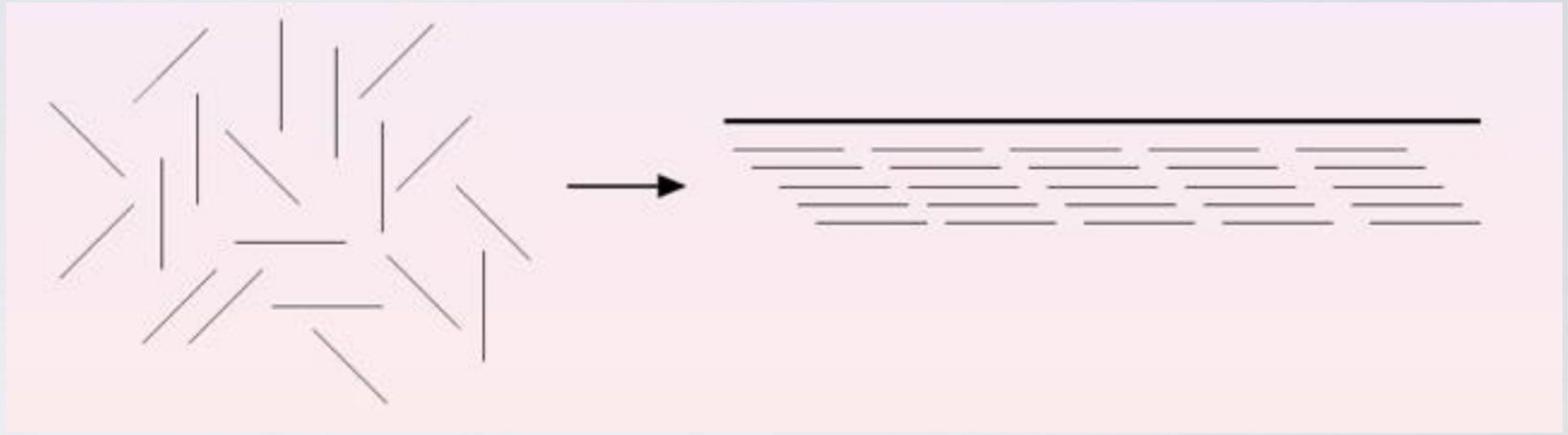
DATA ANALYSIS

- Quality control

- Mapping
- Assembly
- Digital gene expression
- Visualisation

DATA ANALYSIS

- Quality control



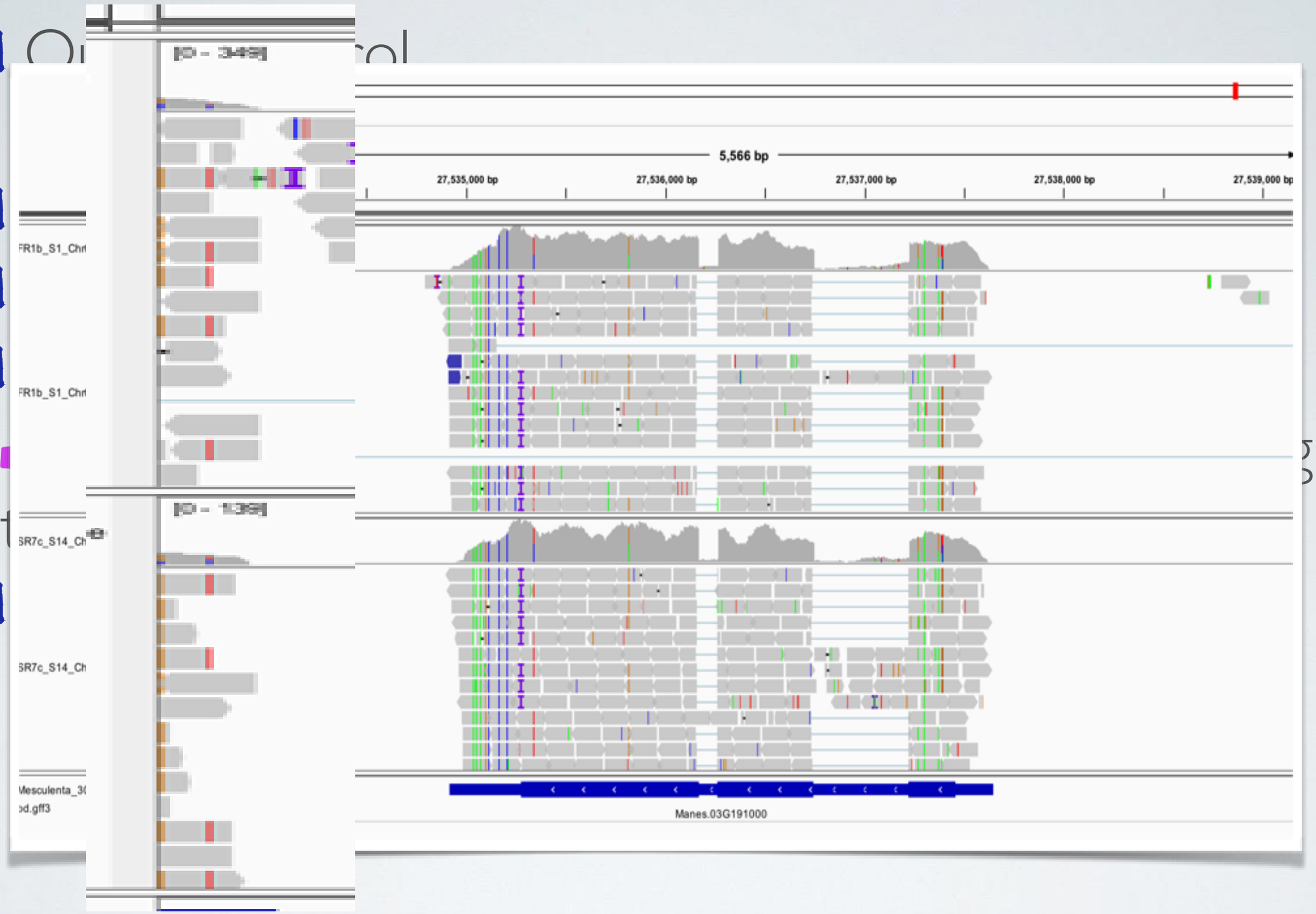
- Mapping
- Assembly
  - Assemble sequencing reads to recover the sequence in investigation
- Digital gene expression
- Visualisation

DATA ANALYSIS

- Quality control

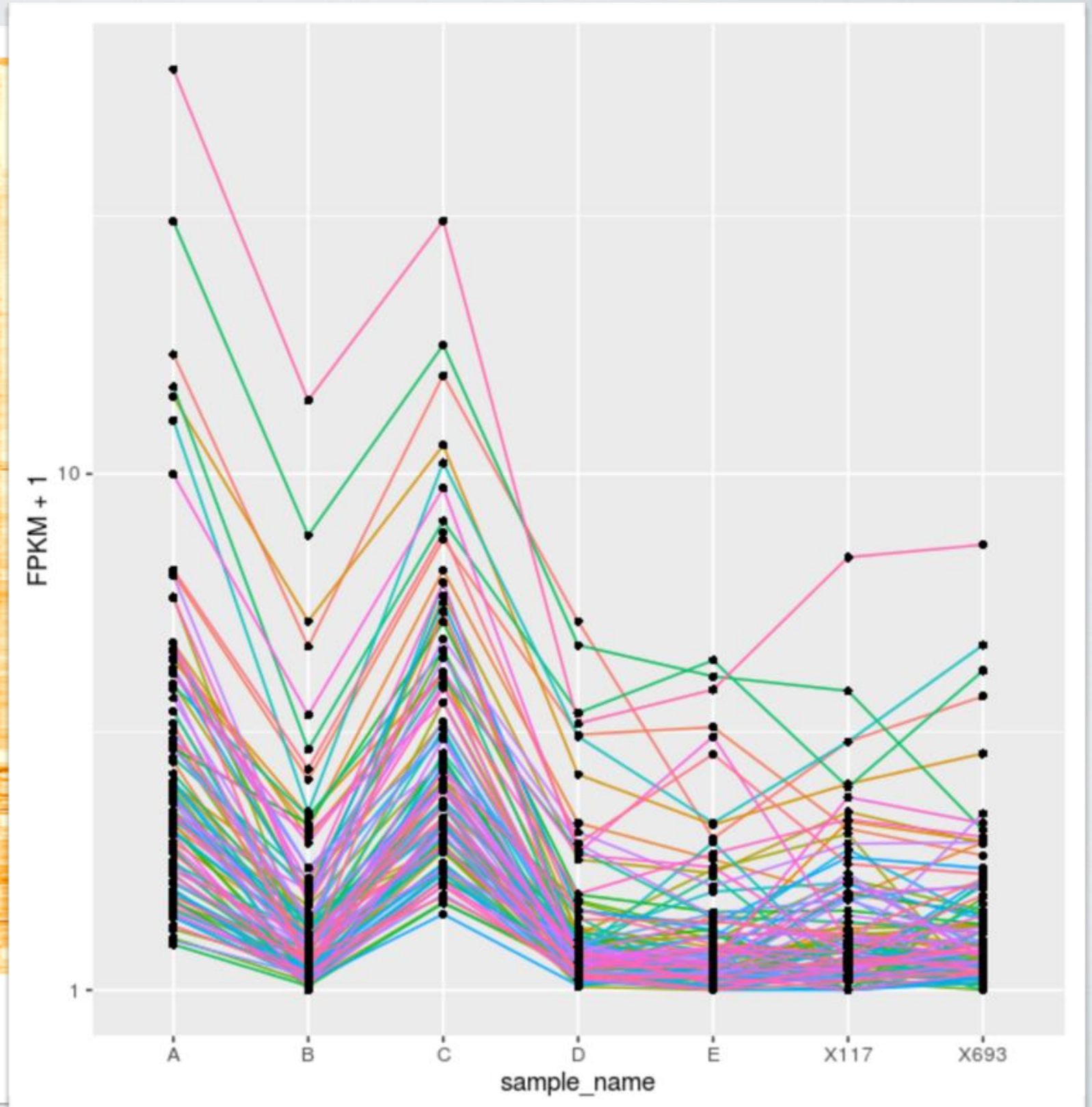
- Mapping
- Alignment
- Digital gene expression
  - Compare differentially expressed sequences using there frequency witin the data set
- Visualisation

DATA ANALYSIS



DATA ANALYSIS

Quality control



- Quality control

- Mapping
- Assembly
- Digital gene expression
- Visualisation

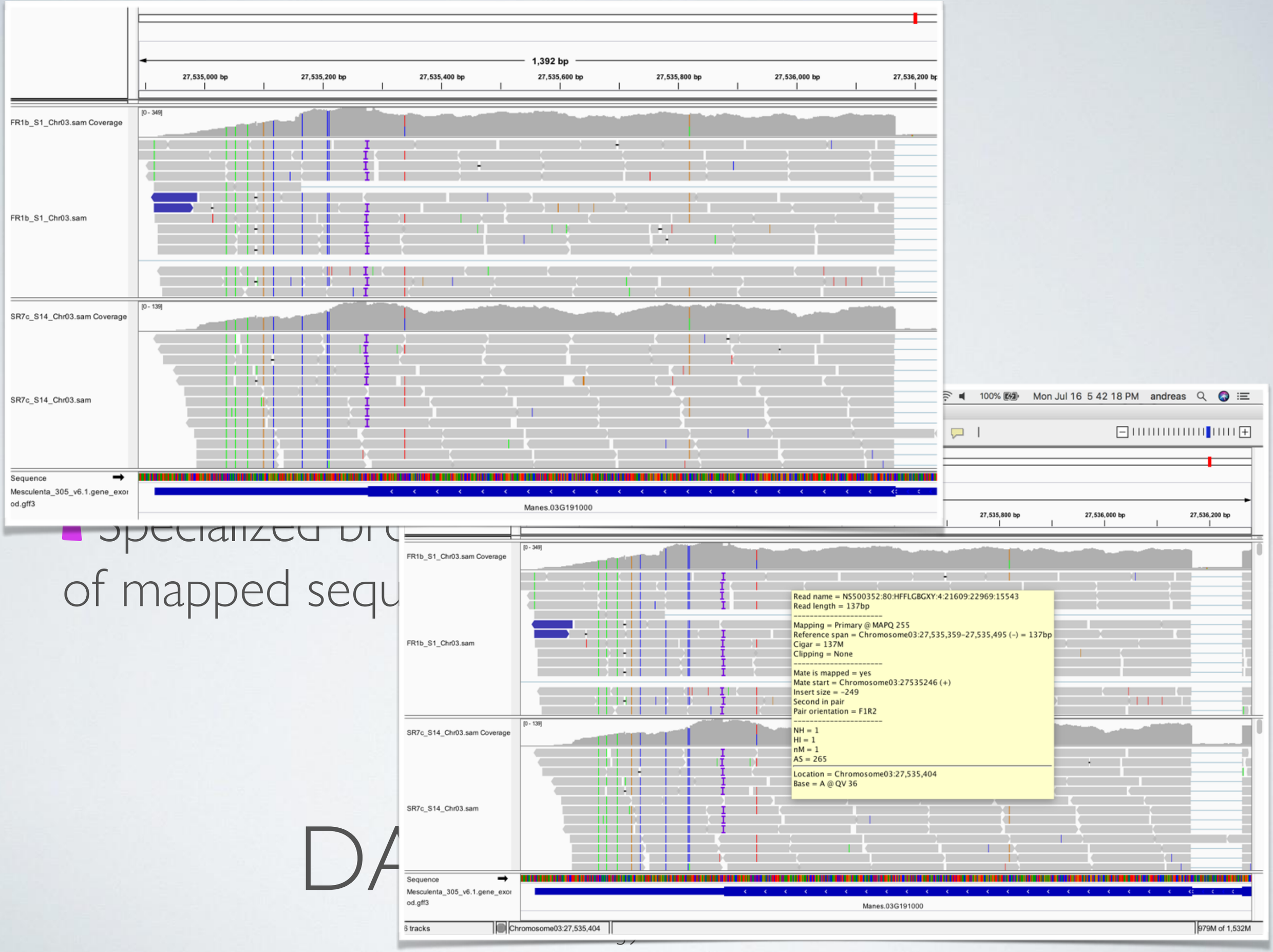
- Specialized browsers to visualize the vast amount of mapped sequences

DATA ANALYSIS

Specialized bro

of mapped sequ

DA



- Quality control

- Mapping
- Assembly
- Digital gene expression
- Visualisation

TOOLS

- Quality control
  - in most cases incorporated in sequencing platform software
  - FASTQC / MultiQC
  - Trimmomatic
  - fastp
  - TrimGalore
- Mapping
- Assembly
- Digital gene expression
- Visualisation

TOOLS

- Quality control

- Mapping
  - read indexing with hash table
  - genome indexing with hash table
  - genome indexing with suffix array
  - SAM/BAM format
  - http://lh3lh3.users.sourceforge.net/NGSalign.shtml
- Assembly
- Digital gene expression
- Visualisation

TOOLS

- Quality control

- Mapping
- Assembly
  - Greedy
  - Overlap Layout Consensus (OLC)
  - de Bruijn graph based
  - https://en.wikipedia.org/wiki/De_novo_sequence_assemblers

- Digital gene expression
- Visualisation

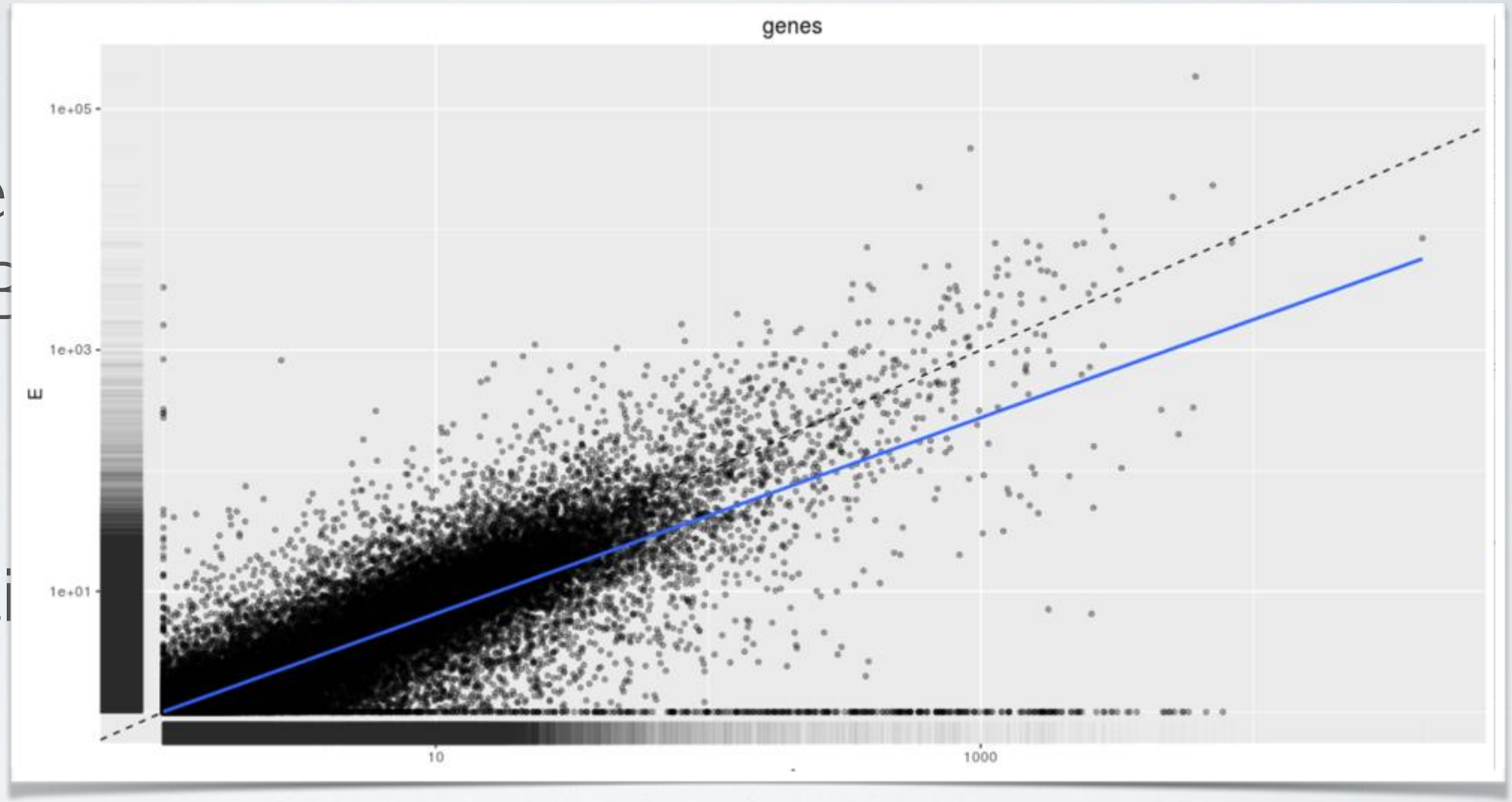
TOOLS

- Quality control

- Mapping
- Assembly
- Digital gene expression
  - DESeq, BaySeq, edgeR are R package to analyse count data from high-throughput sequencing assays such as RNA-Seq and test for differential expression.
- Visualization

TOOLS

- Quality control
- Mapping
- Assembly
- Digital ge
  - DESeq, E
    data from
    RNA-Seq
- Visualizati



TOOLS

- Quality control

- Mapping
- Assembly
- Digital gene expression
- Visualization
  - http://lh3lh3.users.sourceforge.net/NGSaInview.shtml

TOOLS

Tablet (http://bioinf.scri.ac.uk/tablet/)

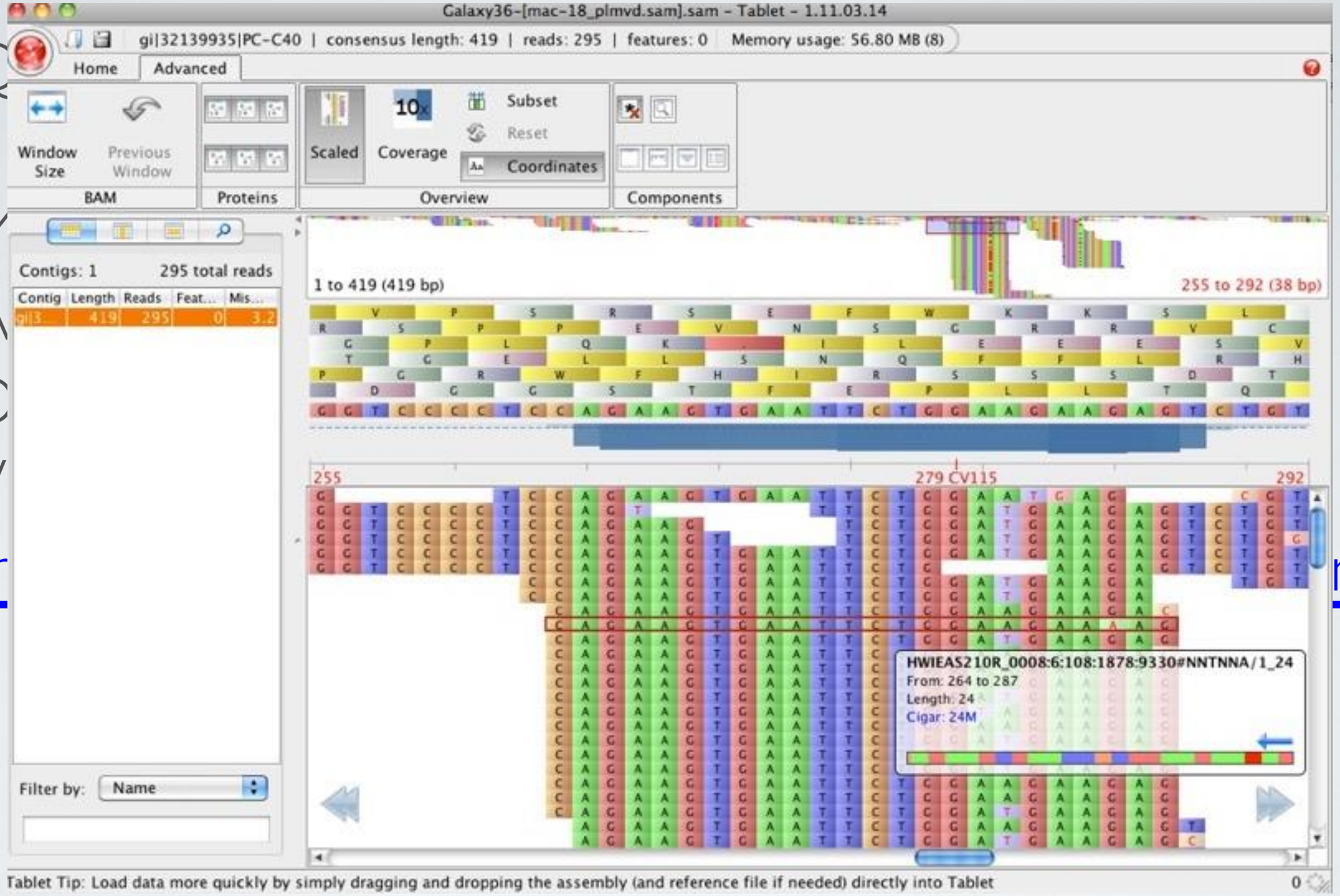
- Quality control

- Mapping
- Assembly
- Digital gene expression

- Visualization
  - http://lh3lh3.users.sourceforge.net/NGSaInview.shtml

TOOLS

Tablet (http://bioinf.scri.ac.uk/tablet/)



TOOLS

JBrowse (https://www.jbrowse.org/jb2/)

- Quality control

- Mapping
- Assembly
- Digital gene expression

- Visualization
  - http://lh3lh3.users.sourceforge.net/NGSaInview.shtml

TOOLS

FILE ADD TOOLS HELP Comprehensive user guide SHARE JBrowse

volvox

ctgA 10,000 20,000 30,000 40,000

ctgA:14,427..14,534 108bp

ctgA 14,440 14,460 14,480 14,500 14,520

volvox-sorted.bam (ctgA)

26

0

Sorted by base pair

FILE HELP New Session 2020-07-07T01:10:31.349Z JBrowse

text filter

show only regions with data

	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	1	564464	122419_0	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=MT;END=3916;ST +;STRANDS2=4,7,4,7;RE=11;AF=0.305556
2	1	565469	116060_1	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=MT;END=4919;ST +;STRANDS2=11,19,11,19;RE=30;AF=0.882353
3	1	567234	116060_1	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=MT;END=6706;ST
4	1	568528	116060_1	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=MT;END=7975;ST
5	1	569820	116060_1	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=MT;END=9269;ST
6	1	3393896	235_2	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=15;END=797316
7	1	9121445	596_2	N	<TRA>		PASS	PRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=14;END=93713178;
8	1	9121448	124893	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=14;END=937124 -;STRANDS2=0,12,12,0;RE=12;AF=0.705882
9	1	17009787	122457	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=X;END=2014409
10	1	17011394	849_3	N	<TRA>		PASS	IMPRECISE;SVMETHOD=Snifflesv1.0.3;CHR2=X;END=2014570

273 rows

IGV (https://software.broadinstitute.org/software/igv/)

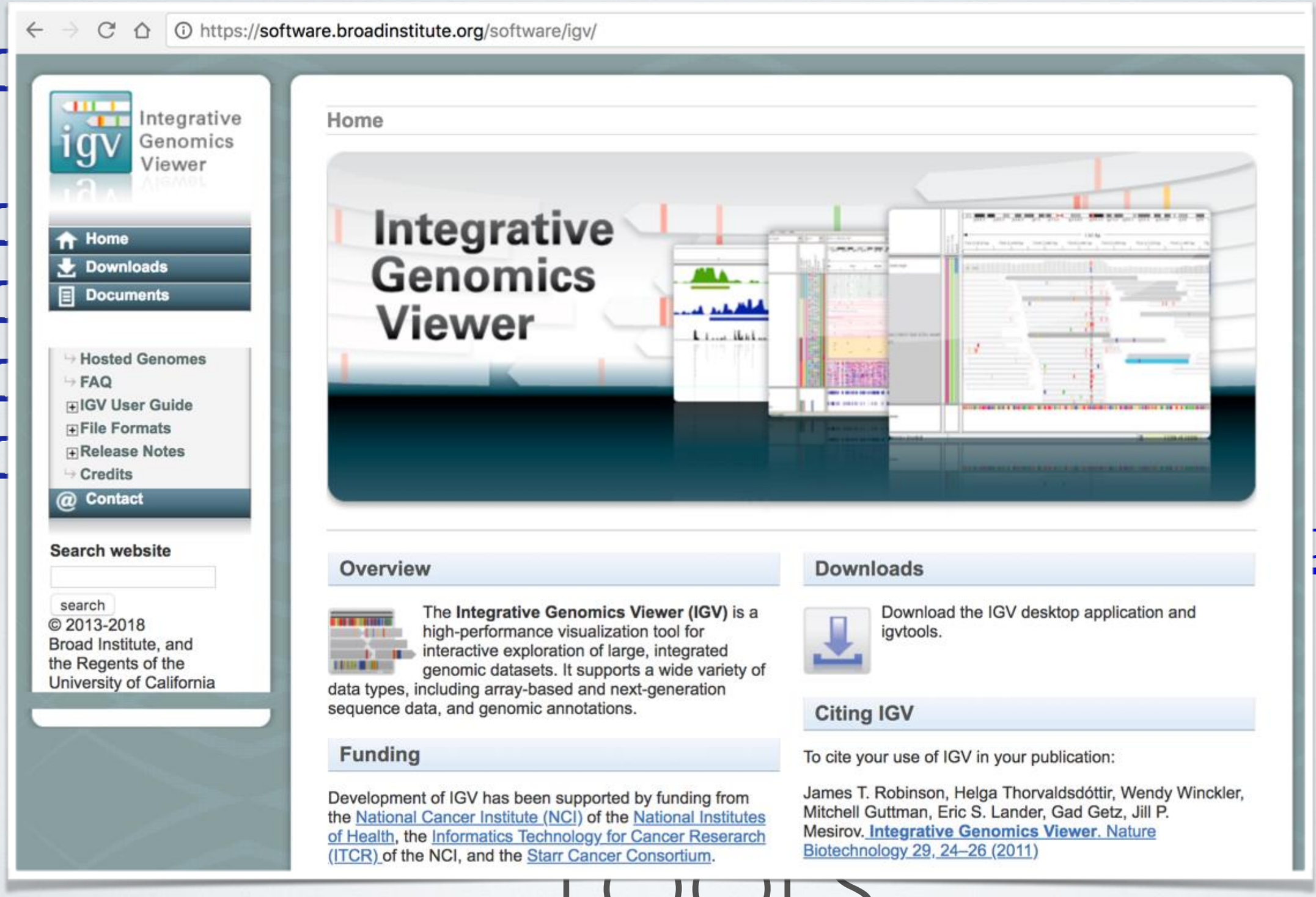
- Quality control

- Mapping
- Assembly
- Digital gene expression

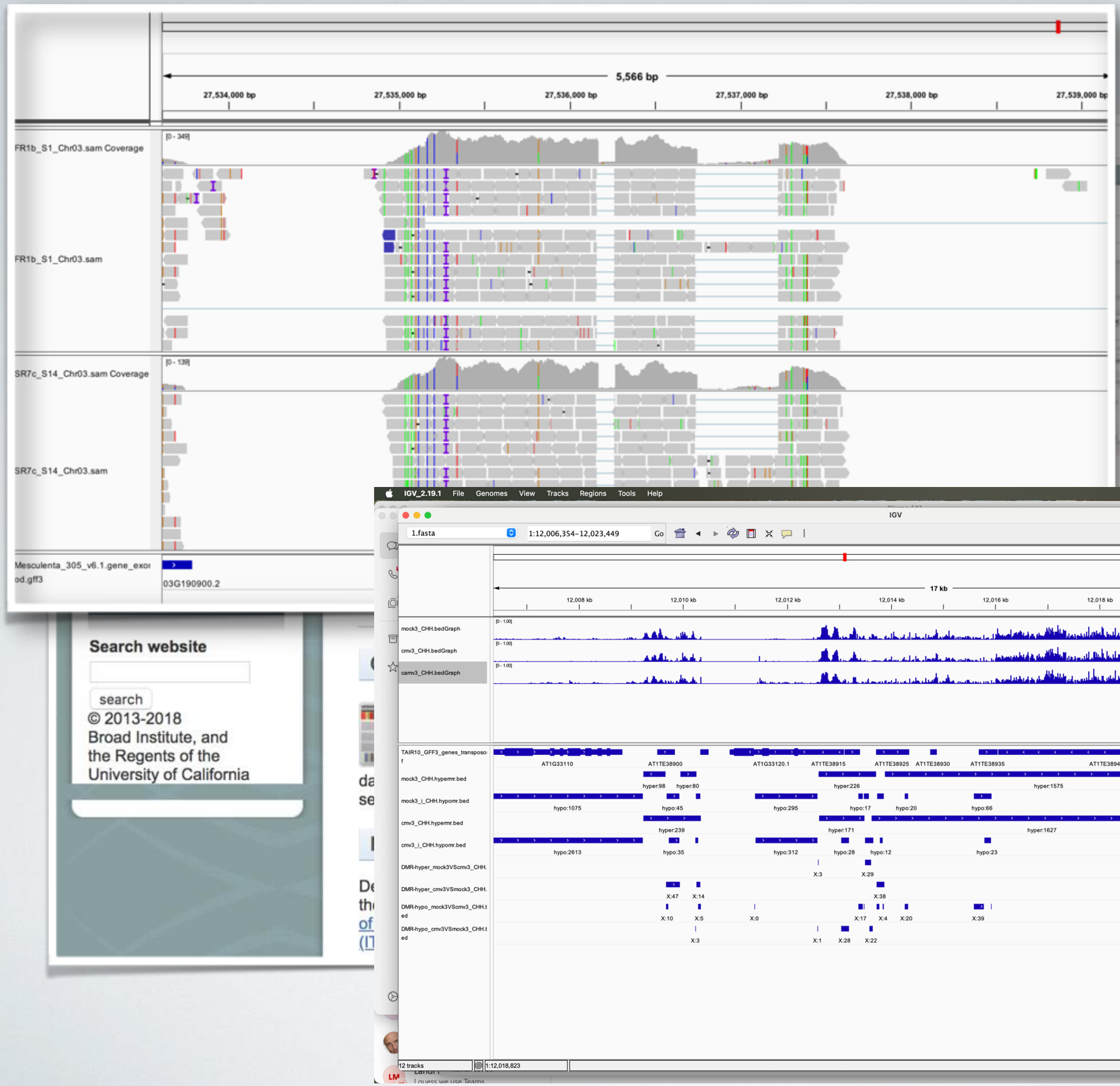
- Visualization
  - http://lh3lh3.users.sourceforge.net/NGSaInview.shtml

TOOLS

IGV (https://software.broadinstitute.org/software/igv/)



TOOLS


5,566 bp
27,534,000 bp
27,535,000 bp
27,536,000 bp
27,537,000 bp
27,538,000 bp
27,539,000 bp
FR1b_S1_Chr03.sam Coverage
[0 - 349]
FR1b_S1_Chr03.sam
SR7c_S14_Chr03.sam Coverage
[0 - 139]
SR7c_S14_Chr03.sam
IGV_2.19.1 File Genomes View Tracks Regions Tools Help
Tue Apr 8 16:13
IGV
1.fasta
1:12,006,354-12,023,449
Go
17 kb
12,008 kb
12,010 kb
12,012 kb
12,014 kb
12,016 kb
12,018 kb
12,020 kb
12,022 kb
mock3_CHH.bedGraph
cmv3_CHH.bedGraph
camv3_CHH.bedGraph
TAIR10_GFF3_genes_transposon
mock3_CHH.hypermr.bed
mock3_i_CHH.hypomr.bed
cmv3_CHH.hypermr.bed
cmv3_i_CHH.hypomr.bed
DMR-hyper_mock3VScmv3_CHH.
DMR-hyper_cmv3VSmock3_CHH.
DMR-hypo_mock3VScmv3_CHH.bed
DMR-hypo_cmv3VSmock3_CHH.bed
12 tracks
1:12,018,823
6,994M of 8,589M
Mesculenta_305_v6.1.gene_exon
od.gff3
03G190900.2
Search website
search
© 2013-2018
Broad Institute, and
the Regents of the
University of California

http://www.ebi.ac.uk/ena/

EMBL-EBI home | Services | Research | Training | About us | EMBL-EBI

European Nucleotide Archive

Enter text search terms — Search

Examples: histone, BN000065

Enter accession — View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | About | Support

We recommend that you subscribe to the ENA-announce mailing list for updates on ENA services.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

Submit | Search | Support

Latest ENA news

ENA launches Data Hubs Portal **Mar 26, 2025, 11:00:22 AM**

The Data Hubs Portal is a new interface that enables users to setup and manage pre-release and/or public Data Hubs at the ENA.

Read more >

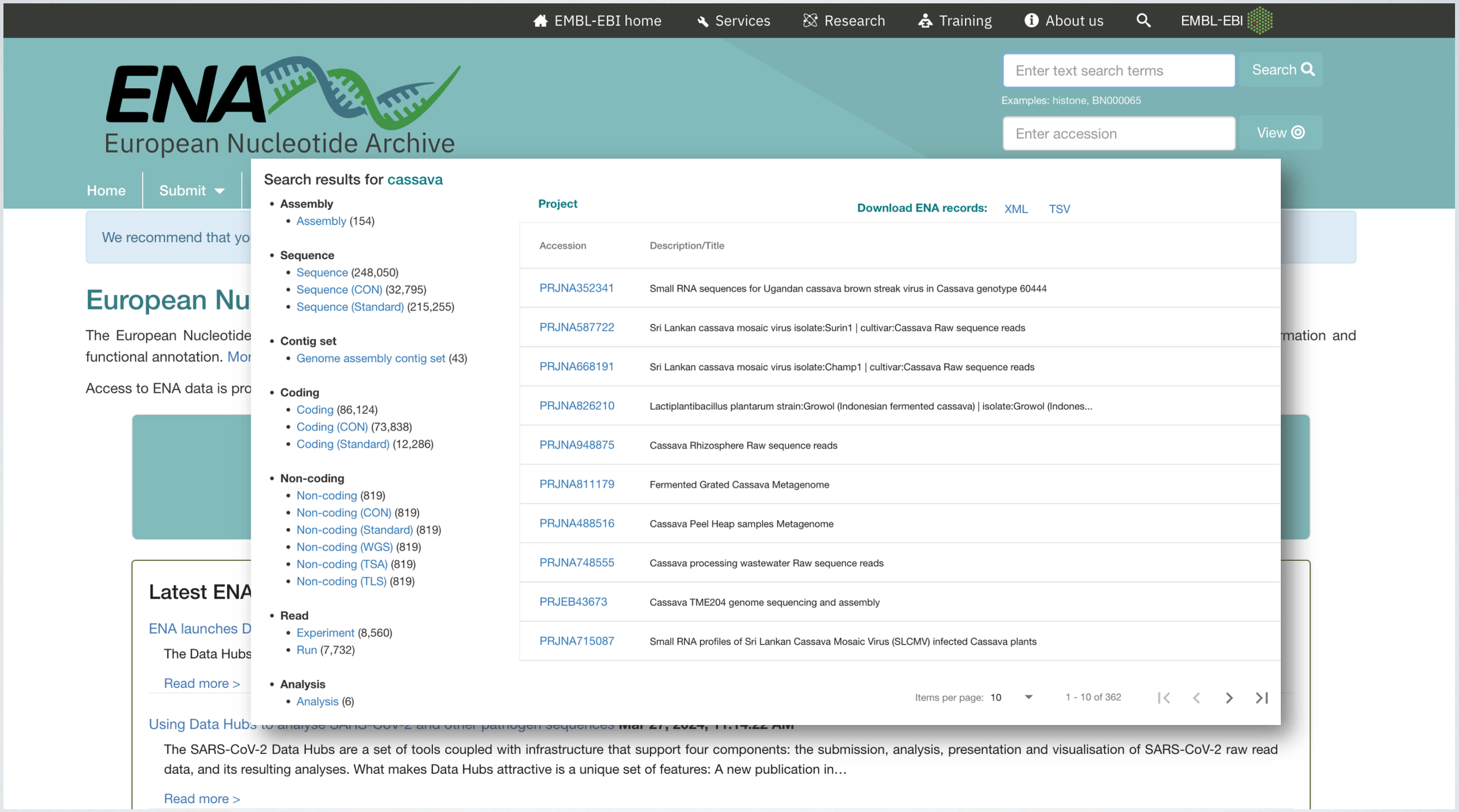
Using Data Hubs to analyse SARS-CoV-2 and other pathogen sequences **Mar 27, 2024, 11:14:22 AM**

The SARS-CoV-2 Data Hubs are a set of tools coupled with infrastructure that support four components: the submission, analysis, presentation and visualisation of SARS-CoV-2 raw read data, and its resulting analyses. What makes Data Hubs attractive is a unique set of features: A new publication in…

Read more >

DATA REPOSITORY

http://www.ebi.ac.uk/ena/



DATA REPOSITORY

THANKS!!!