An introduction to Illumina Sequencing

Daniel P. Depledge

Systems Virologist
Assistant Professor, Department of Medicine

Biomedical Informatics: Advanced course offerings

Applied Sequencing Informatics (2021)

- Expanded version of current course
- Lectures and practicums (50:50 split)
- Advanced sequencing analyses using both short- and long-read data

- **Prerequisites**: experience working in HPC environments (i.e. Big Purple) + experience in R-based environments

Overview of October

Getting to grips with Illumina sequencing data

- October 6th: Introduction to Illumina Sequencing [Daniel Depledge]
- October 8th: Introduction to unix/bash/slurm [Mark Grivainis]
- October 13th: Getting to grips with SAMtools and BEDtools [Daniel Depledge]
- October 15th: Assignment #1 presentations

A crash course in differential gene expression analysis

- October 20th: Introduction to RNA-Seq [Daniel Depledge]
- October 22nd: Introduction to Differential Gene Expression analysis [Daniel Depledge]
- October 27th: Advanced unix/bash/slurm [Mark Grivainis]
- October 29th: Assignment #2 presentations

Setting the scene

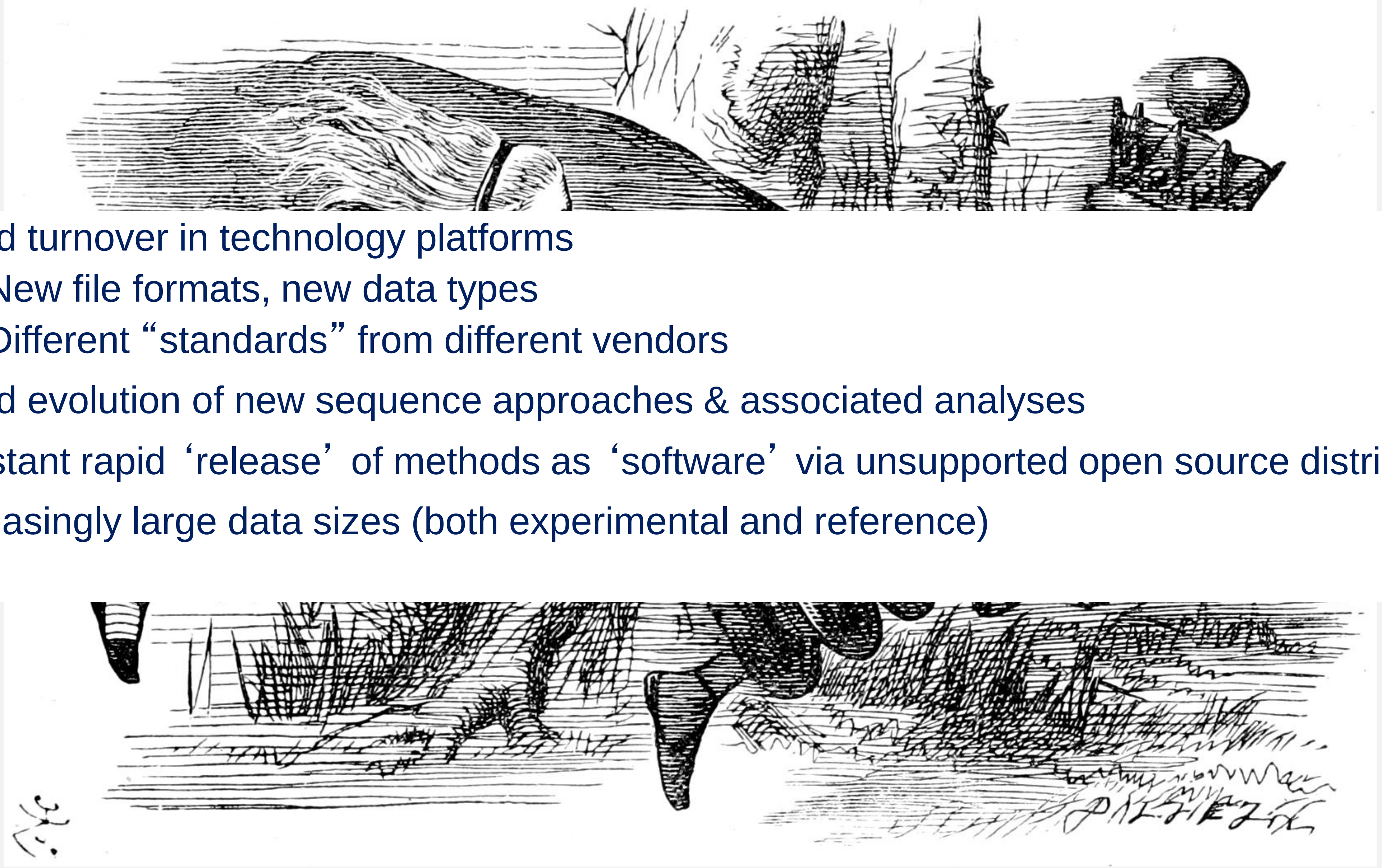
High-throughput sequencing (HTS) is fundamentally changing how we approach science

- HTS is a readout for many different types of laboratory experiments
- Clinical and basic science investigators from all areas of biology can make use of this technology
- Many (most?) are completely naïve about bioinformatics
- Decreasing sequencing costs = increasing use for routine assays + technical innovation + novel applications

Sequencing informatics is a bottleneck!

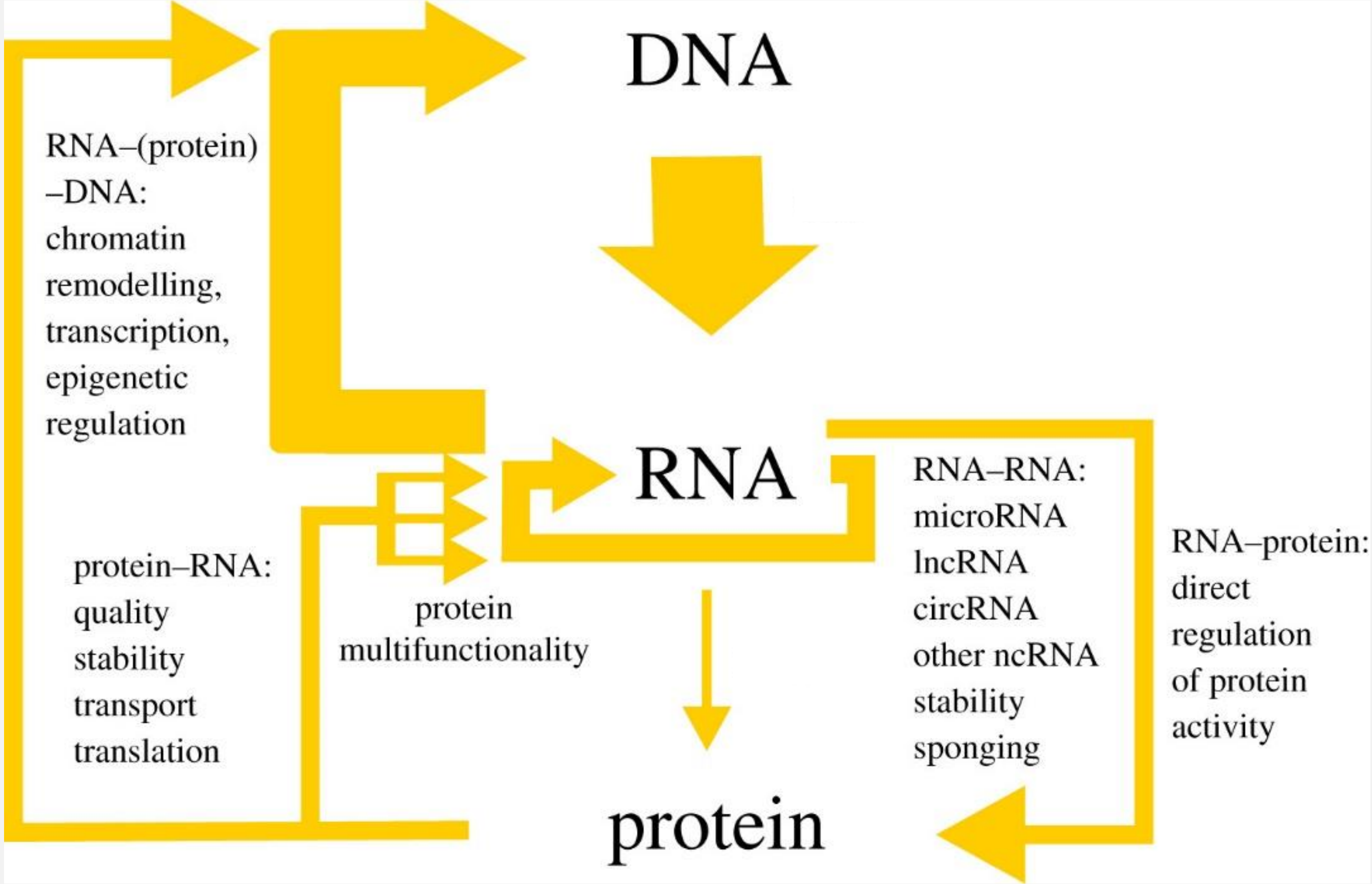
- Sequencing is a commodity – easy to outsource
- Sequencing informatics is the essential point of the science
- Data analysis and discovery of meaning in raw results
- Increasing data throughput = increasing time and cost of analysis

Staying in the game…

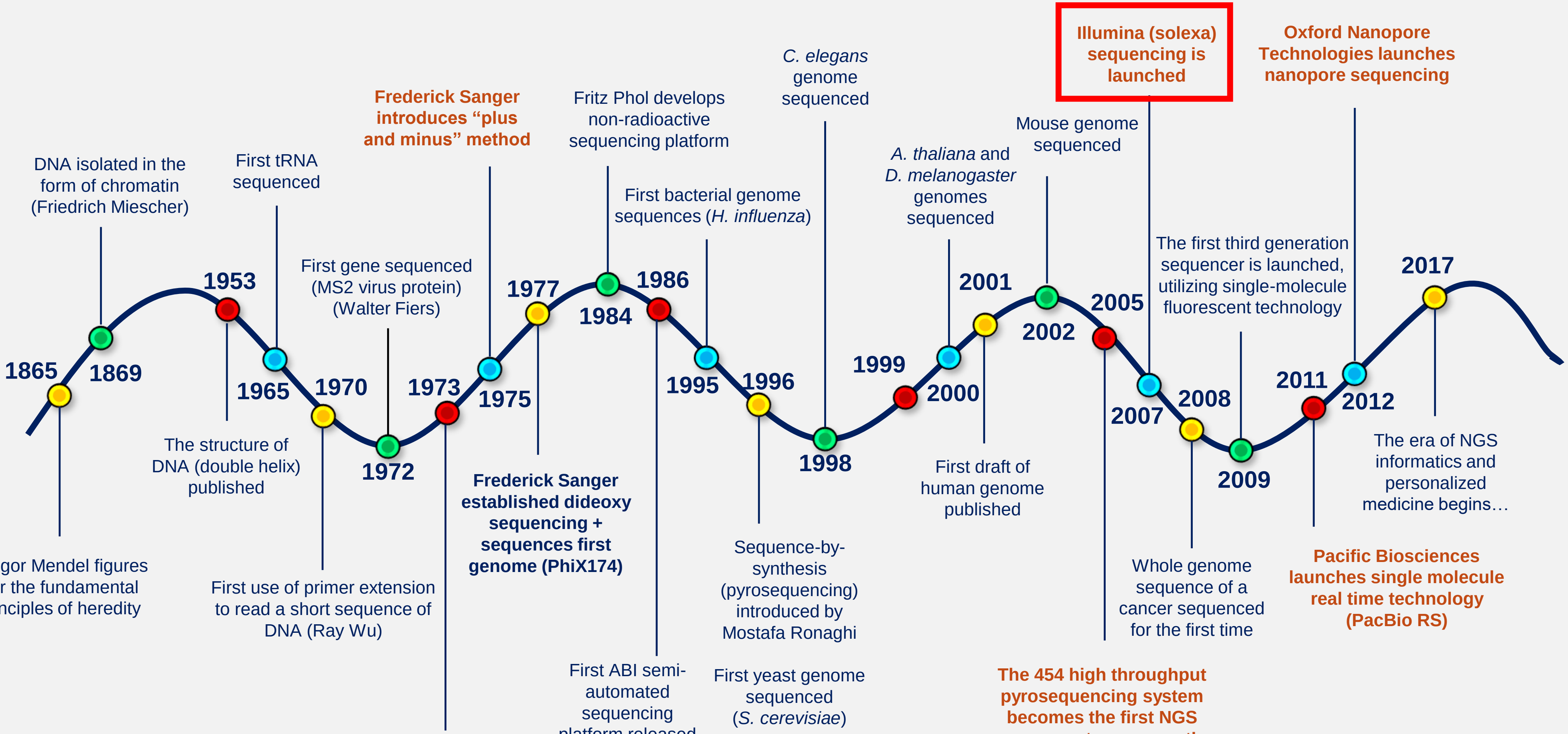


- Rapid turnover in technology platforms
  - New file formats, new data types
  - Different “standards” from different vendors
- Rapid evolution of new sequence approaches & associated analyses
- Constant rapid ‘release’ of methods as ‘software’ via unsupported open source distribution
- Increasingly large data sizes (both experimental and reference)

Why we sequence?


DNA
RNA
protein
RNA–(protein) –DNA: chromatin remodelling, transcription, epigenetic regulation
protein–RNA: quality stability transport translation
protein multifunctionality
RNA–RNA: microRNA lncRNA circRNA other ncRNA stability sponging
RNA–protein: direct regulation of protein activity

An abridged history of sequencing



1865 – Gregor Mendel figures our the fundamental principles of heredity

1869 – DNA isolated in the form of chromatin (Friedrich Miescher)

1953 – The structure of DNA (double helix) published

1965 – First tRNA sequenced

1970 – First use of primer extension to read a short sequence of DNA (Ray Wu)

1972 – First gene sequenced (MS2 virus protein) (Walter Fiers)

1973 – DNA sequencing through chemical cleavage (Walter Gilbert & Allan Maxam)

1975 – **Frederick Sanger introduces “plus and minus” method**

1977 – **Frederick Sanger established dideoxy sequencing + sequences first genome (PhiX174)**

1984 – Fritz Phol develops non-radioactive sequencing platform

1986 – First ABI semi-automated sequencing platform released

1995 – First bacterial genome sequences (*H. influenza*)

1996 – Sequence-by-synthesis (pyrosequencing) introduced by Mostafa Ronaghi

First yeast genome sequenced (*S. cerevisiae*)

ABI release first commercial capillary sequencing platform

1998 – *C. elegans* genome sequenced

1999

2000 – *A. thaliana* and *D. melanogaster* genomes sequenced

2001 – First draft of human genome published

2002 – Mouse genome sequenced

2005 – **The 454 high throughput pyrosequencing system becomes the first NGS sequencer to come on the market**

2007 – **Illumina (solexa) sequencing is launched**

2008 – Whole genome sequence of a cancer sequenced for the first time

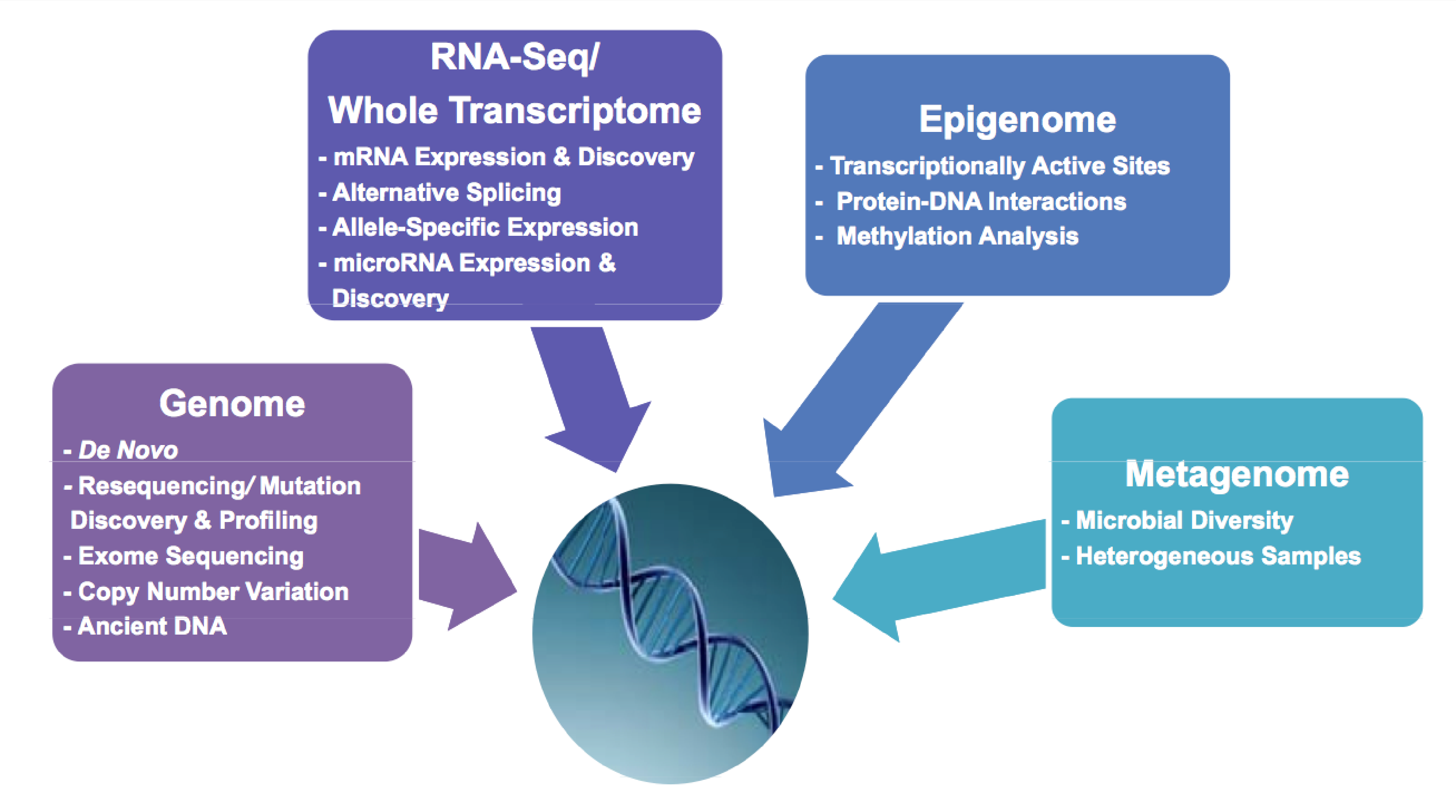
2009 – The first third generation sequencer is launched, utilizing single-molecule fluorescent technology

2011 – **Pacific Biosciences launches single molecule real time technology (PacBio RS)**

2012 – **Oxford Nanopore Technologies launches nanopore sequencing**

2017 – The era of NGS informatics and personalized medicine begins…

The rise of high-throughput sequencing



RNA-Seq/ Whole Transcriptome

- mRNA Expression & Discovery
- Alternative Splicing
- Allele-Specific Expression
- microRNA Expression & Discovery

Epigenome

- Transcriptionally Active Sites
- Protein-DNA Interactions
- Methylation Analysis

Genome

- *De Novo*
- Resequencing/ Mutation Discovery & Profiling
- Exome Sequencing
- Copy Number Variation
- Ancient DNA

Metagenome

- Microbial Diversity
- Heterogeneous Samples

The principles of generating a short-read sequencing library

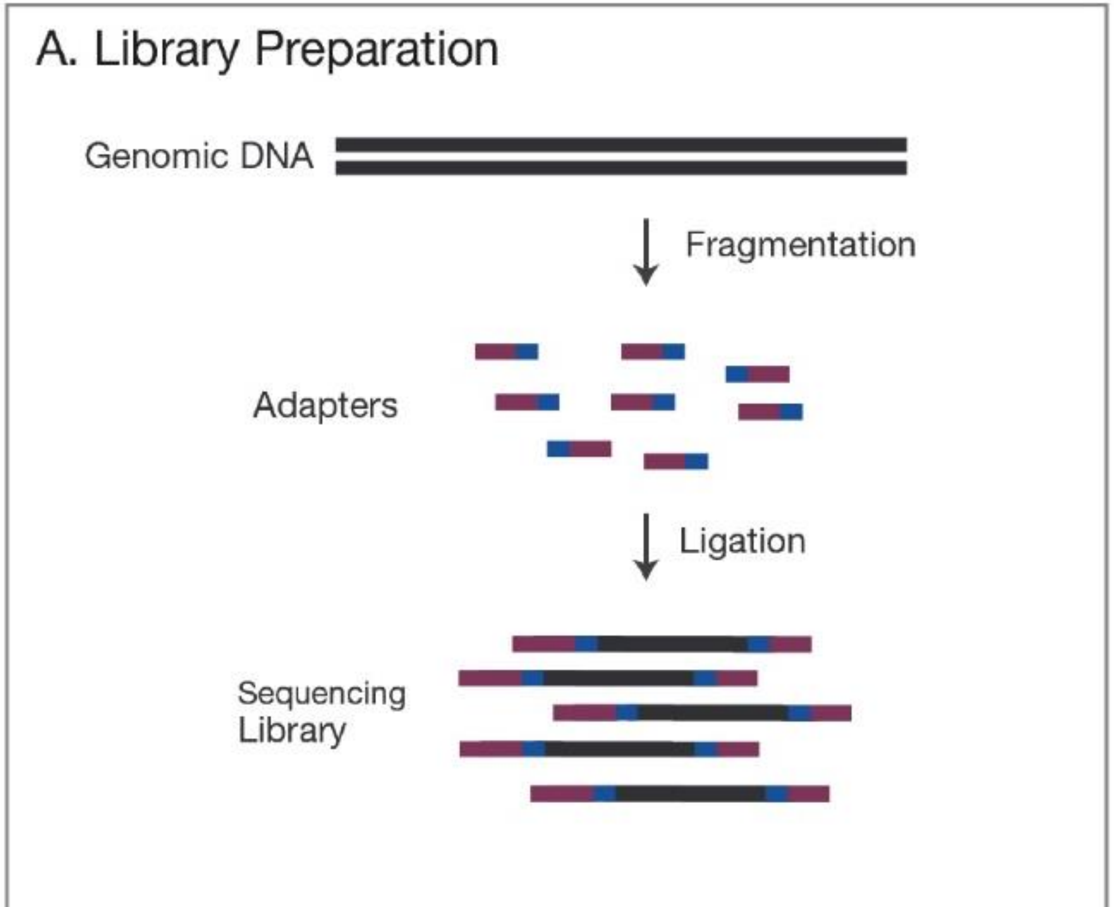
1. Capture DNA or RNA of interest
   - cDNA must be synthesized from RNA

2. Fragment DNA/cDNA to produce fragments of 150-300 nt
   - Acoustic sonication (random shearing) is favoured
   - Alternative strategies include use of transposases or targeted ligation

3. Repair ends and ligate adapter sequences

4. PCR amplification to enrich for fragments with correct ligation
   - PCR primes of sequences in adapters

5. Sequence

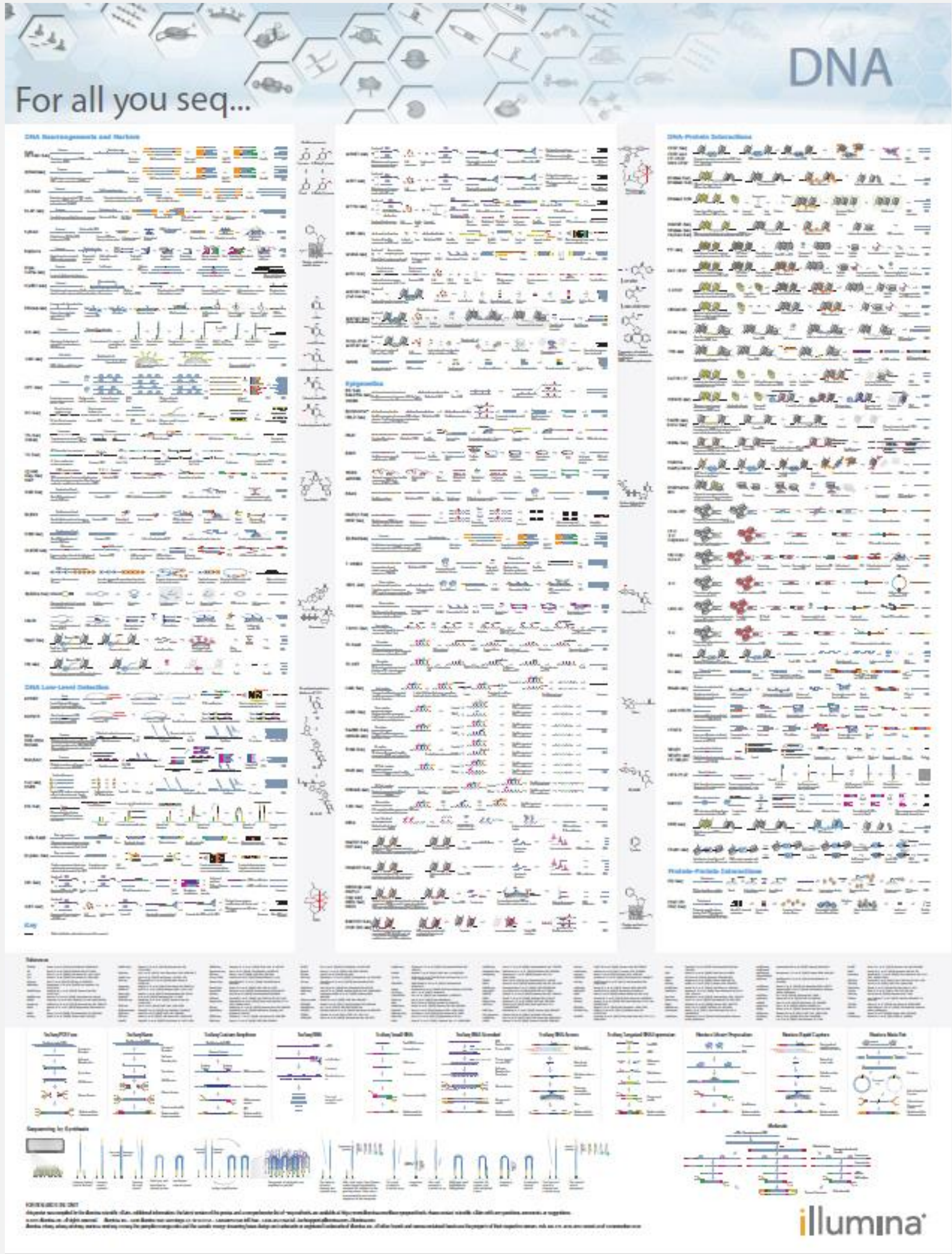


NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

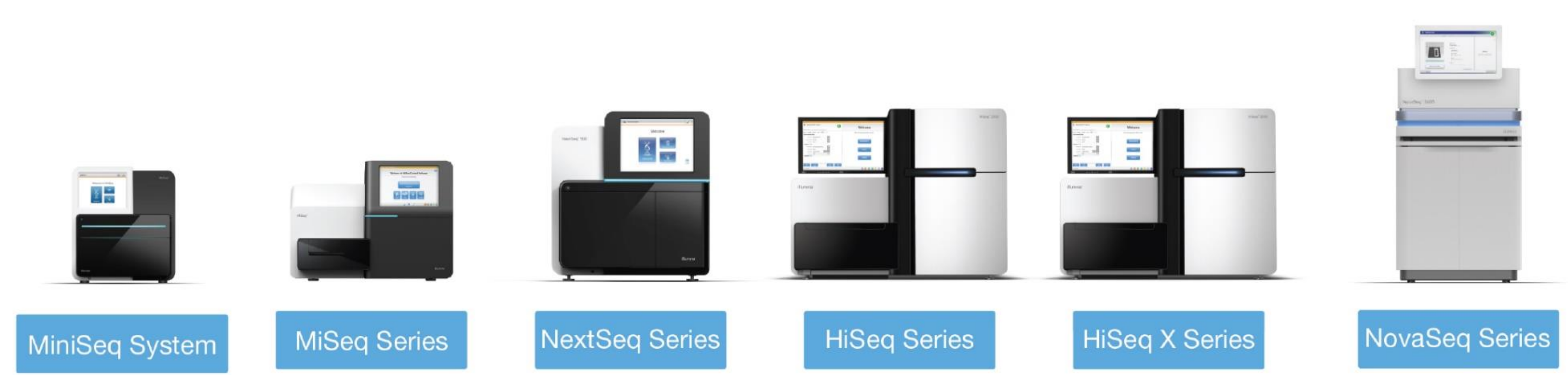
The incredible versatility of Illumina sequencing

- ***Hundreds*** of distinct Illumina-based methods for DNA & RNA sequencing at global (bulk) or single-cell level
- Most all of these methods require tweaks and special considerations when performing informatics analyses

https://www.illumina.com/science/sequencing-method-explorer.html

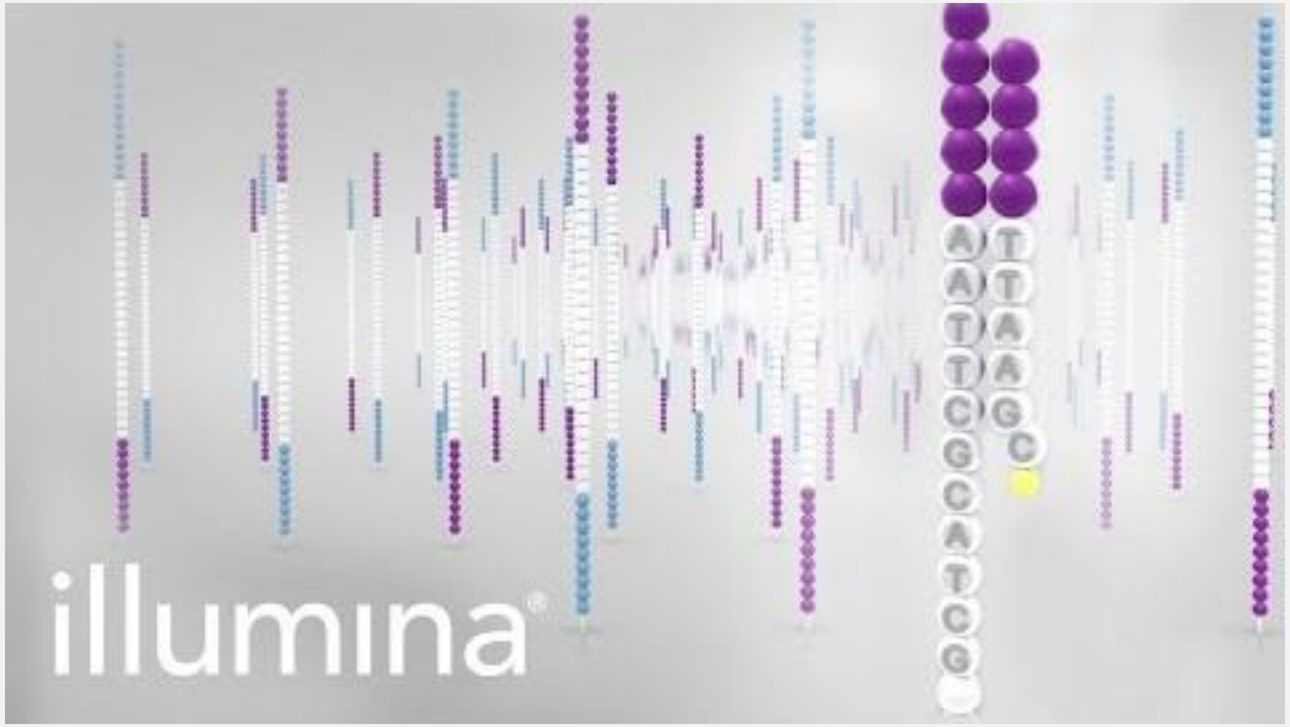


Illumina sequencing platforms



8 Gb 15 Gb 120 Gb 1 Tb 2 Tb 6 Tb

Illumina: sequencing by synthesis

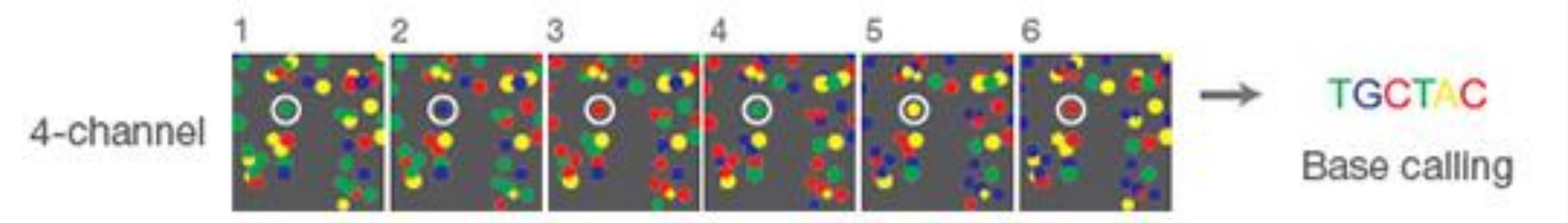


https://youtu.be/fCd6B5HRaZ8

Converting signal into bases

- Basecalling is the process of converting raw signal into basecalls (A, T, G, C)

FASTA format – two lines, no quality information



An example of Illumina basecalling

- Individual basecalls are assigned probability values based on how closely a raw signal matches the expected signal

- Most sequencing analysis pipelines make use of these basecall probability values to inform quality control

Defining basecall qualities: the Phred system

$$\varepsilon = 10^{-\frac{Q_{Phred}}{10}}$$

$$Q_{Phred} = -10 \cdot \log_{10}(\varepsilon)$$

E is the *Error Probability (*probability that a base call is wrong

Q: *Phred Quality Score*

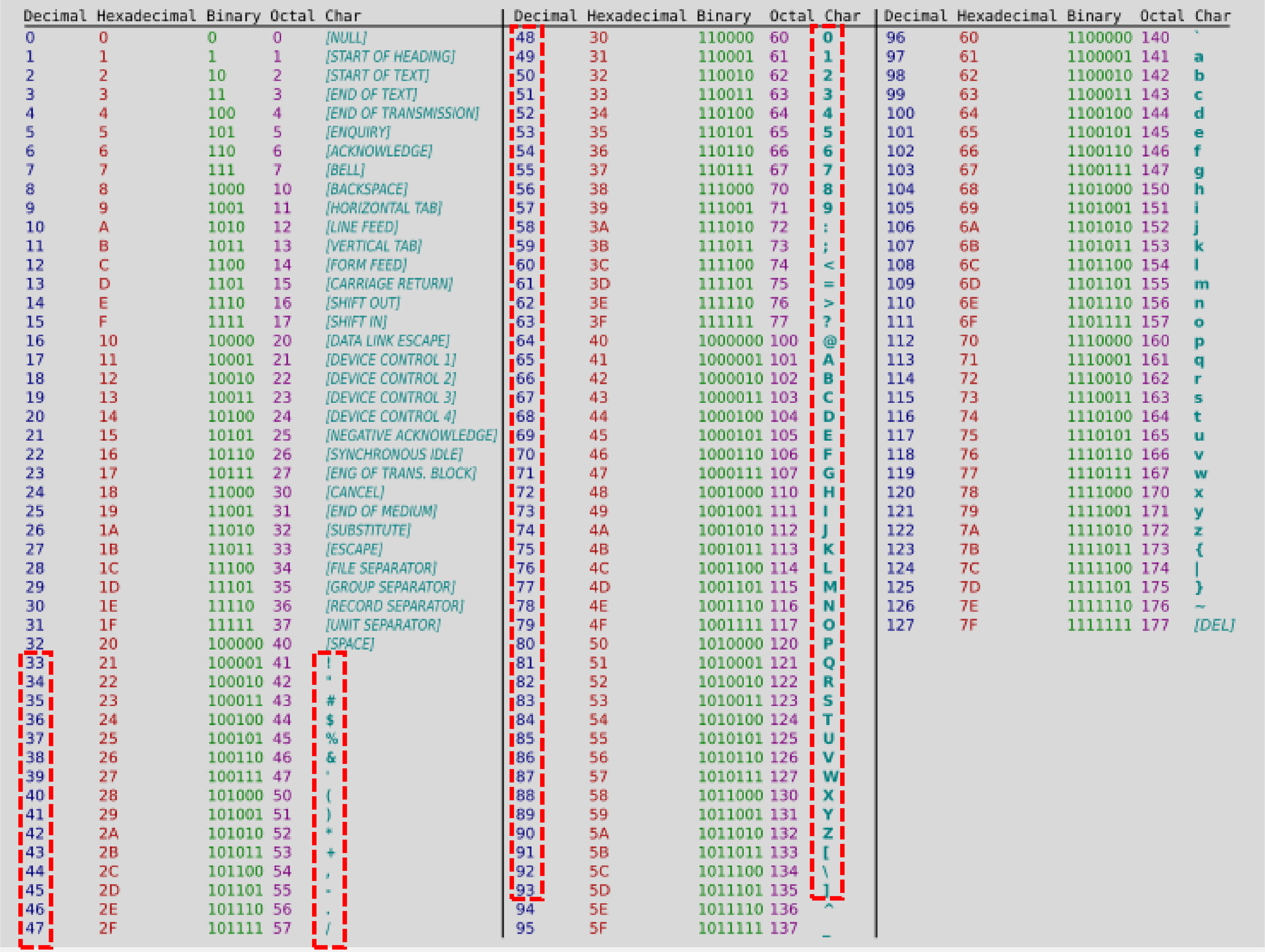
Q	***Probability of incorrect basecall***	***Basecall accuracy***
60	1 in 1000000	99.9999%
50	1 in 100000	99.999%
40	1 in 10000	99.99%
30	1 in 1000	99.9%
20	1 in 100	99%
10	1 in 10	90%

ASCII tables

- Phred Q score for a given basecall is stored as a single (ASCII) character (save space)

- ASCII tables allow translation between numerical value and characters
  - 33 = !
  - 35 = #
  - 40 = (

- Quality score is derived from Decimal – 32
  - 33 = ! = 1
  - 35 = # = 3
  - 40 = (= 8



Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char	Decimal	Hexadecimal	Binary	Octal	Char
0	0	0	0	[NULL]	48	30	110000	60	0	96	60	1100000	140	`
1	1	1	1	[START OF HEADING]	49	31	110001	61	1	97	61	1100001	141	a
2	2	10	2	[START OF TEXT]	50	32	110010	62	2	98	62	1100010	142	b
3	3	11	3	[END OF TEXT]	51	33	110011	63	3	99	63	1100011	143	c
4	4	100	4	[END OF TRANSMISSION]	52	34	110100	64	4	100	64	1100100	144	d
5	5	101	5	[ENQUIRY]	53	35	110101	65	5	101	65	1100101	145	e
6	6	110	6	[ACKNOWLEDGE]	54	36	110110	66	6	102	66	1100110	146	f
7	7	111	7	[BELL]	55	37	110111	67	7	103	67	1100111	147	g
8	8	1000	10	[BACKSPACE]	56	38	111000	70	8	104	68	1101000	150	h
9	9	1001	11	[HORIZONTAL TAB]	57	39	111001	71	9	105	69	1101001	151	i
10	A	1010	12	[LINE FEED]	58	3A	111010	72	:	106	6A	1101010	152	j
11	B	1011	13	[VERTICAL TAB]	59	3B	111011	73	;	107	6B	1101011	153	k
12	C	1100	14	[FORM FEED]	60	3C	111100	74	<	108	6C	1101100	154	l
13	D	1101	15	[CARRIAGE RETURN]	61	3D	111101	75	=	109	6D	1101101	155	m
14	E	1110	16	[SHIFT OUT]	62	3E	111110	76	>	110	6E	1101110	156	n
15	F	1111	17	[SHIFT IN]	63	3F	111111	77	?	111	6F	1101111	157	o
16	10	10000	20	[DATA LINK ESCAPE]	64	40	1000000	100	@	112	70	1110000	160	p
17	11	10001	21	[DEVICE CONTROL 1]	65	41	1000001	101	A	113	71	1110001	161	q
18	12	10010	22	[DEVICE CONTROL 2]	66	42	1000010	102	B	114	72	1110010	162	r
19	13	10011	23	[DEVICE CONTROL 3]	67	43	1000011	103	C	115	73	1110011	163	s
20	14	10100	24	[DEVICE CONTROL 4]	68	44	1000100	104	D	116	74	1110100	164	t
21	15	10101	25	[NEGATIVE ACKNOWLEDGE]	69	45	1000101	105	E	117	75	1110101	165	u
22	16	10110	26	[SYNCHRONOUS IDLE]	70	46	1000110	106	F	118	76	1110110	166	v
23	17	10111	27	[ENG OF TRANS. BLOCK]	71	47	1000111	107	G	119	77	1110111	167	w
24	18	11000	30	[CANCEL]	72	48	1001000	110	H	120	78	1111000	170	x
25	19	11001	31	[END OF MEDIUM]	73	49	1001001	111	I	121	79	1111001	171	y
26	1A	11010	32	[SUBSTITUTE]	74	4A	1001010	112	J	122	7A	1111010	172	z
27	1B	11011	33	[ESCAPE]	75	4B	1001011	113	K	123	7B	1111011	173	{
28	1C	11100	34	[FILE SEPARATOR]	76	4C	1001100	114	L	124	7C	1111100	174	|
29	1D	11101	35	[GROUP SEPARATOR]	77	4D	1001101	115	M	125	7D	1111101	175	}
30	1E	11110	36	[RECORD SEPARATOR]	78	4E	1001110	116	N	126	7E	1111110	176	~
31	1F	11111	37	[UNIT SEPARATOR]	79	4F	1001111	117	O	127	7F	1111111	177	[DEL]
32	20	100000	40	[SPACE]	80	50	1010000	120	P					
33	21	100001	41	!	81	51	1010001	121	Q					
34	22	100010	42	"	82	52	1010010	122	R					
35	23	100011	43	#	83	53	1010011	123	S					
36	24	100100	44	$	84	54	1010100	124	T					
37	25	100101	45	%	85	55	1010101	125	U					
38	26	100110	46	&	86	56	1010110	126	V					
39	27	100111	47	'	87	57	1010111	127	W					
40	28	101000	50	(	88	58	1011000	130	X					
41	29	101001	51	)	89	59	1011001	131	Y					
42	2A	101010	52	*	90	5A	1011010	132	Z					
43	2B	101011	53	+	91	5B	1011011	133	[					
44	2C	101100	54	,	92	5C	1011100	134	\					
45	2D	101101	55	-	93	5D	1011101	135	]					
46	2E	101110	56	.	94	5E	1011110	136	^					
47	2F	101111	57	/	95	5F	1011111	137	_					

Pitfalls of the Phred scoring system

- Based on empirical properties of the data (intensity of cluster, signal-to-noise ratio), combined with observations of actual error rates for known standard samples

- The calculation method is essentially arbitrary (varies by technology), and changes with every iteration of software, chemistry, and hardware on the sequencing machine

- Q scores currently use more data storage space (8 bits) than the bases (2 bits)

The FASTA format – two lines of simplicity

FASTA format – two lines, no quality information

```
(1) >@SRR350953.5|MENDEL_0047_FC62MN8AAXX:1:1:1646:938|length=152

(2) NTCTTTTTCTTTCCTCTTTTGCCAACTTCAGCTAAATAGGAGCTACACTGATTAGGCAGAAACTTGATTAACAG
    GGCTTAAGGTAACCTTGTTGTAGGCCGTTTTGTAGCACTCAAAGCAATTGGTACCTCAACTGCAAAAGTCCTTG
    GCCC
(3) >@SRR350953.5|MENDEL_0047_FC62MN8AAXX:1:1:1934:042|length=152

(4) NTCTTTTACAACCAGCGAGCGACTATCGAGCGCGTCGTAGCGTACGATCGTAAATAGCTGATCGATGCTAGCTA
    GCTAGCGCGATCATCTTTCCTCTAGCACTCAAAGCAATTAGCTACACTGATTAGGCAGAAACTTGATTAACAGG
    GCCT
(5) >HEADER

(6) SEQUENCE
```

Note that header line always start with **>**

The FASTQ format – layering basecall quality information

FASTQ format – four lines, quality information encoded

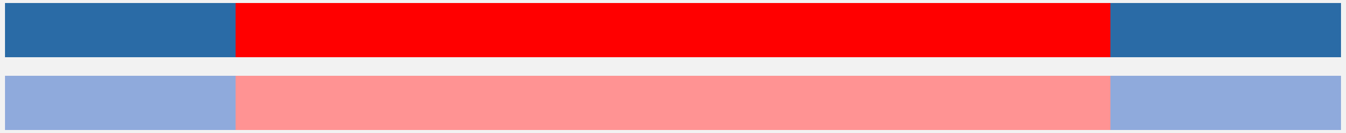
```
(1) @SRR350953.5 MENDEL_0047_FC62MN8AAXX:1:1:1646:938 length=152

(2)NTCTTTTTCTTTCCTCTTTTGCCAACTTCAGCTAAATAGGAGCTACACTGATTAGGCAGAAACTTGATTAACAGGGCTTAAG
   GTAACCTTGTTGTAGGCCGTTTTGTAGCACTCAAAGCAATTGGTACCTCAACTGCAAAAGTCCTTG

(3) +SRR350953.5 MENDEL_0047_FC62MN8AAXX:1:1:1646:938 length=152

(4)+50000222C@@@@@22::::8888898989::::::<<<:<<<<<<:<<<<::<<:::::<<<<<:<:<<<IIIIIGFEEG
   GGGGGGII@IGDGBGGGGGGDDIIGIIEGIGG>GGGGGGDGGGGGIIHIIBIIIGIIIHIIIIGII
```

Note: Header line starts with @

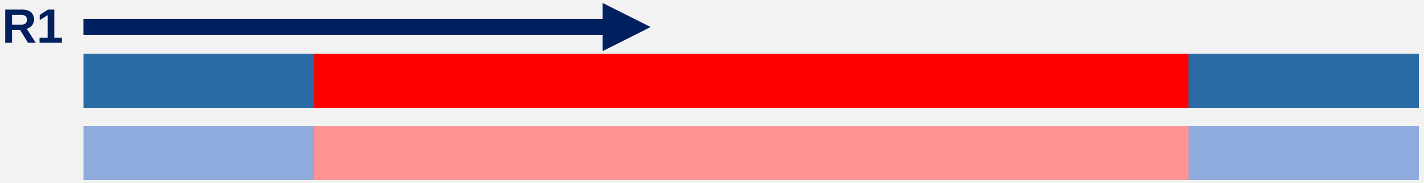
Note: Third line is redundant and typically only contains a single + character

Single vs. paired-end reads (Illumina)

Illumina library sequence fragment with ligated adapters

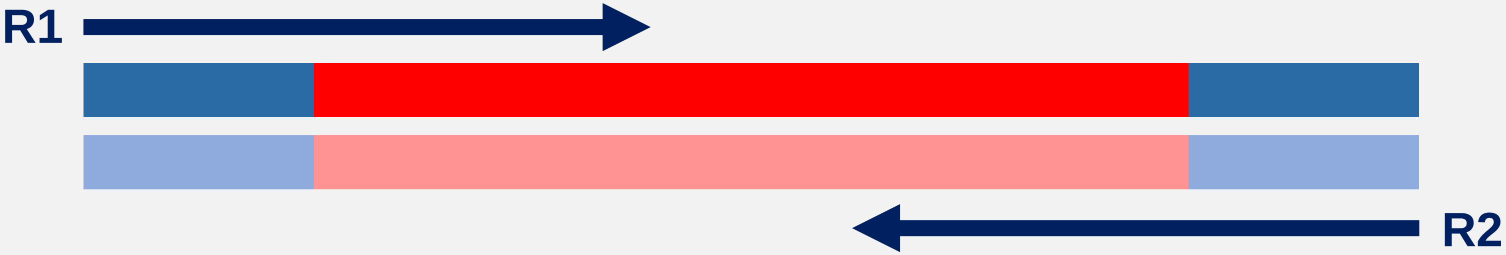


Single end sequencing (Illumina)



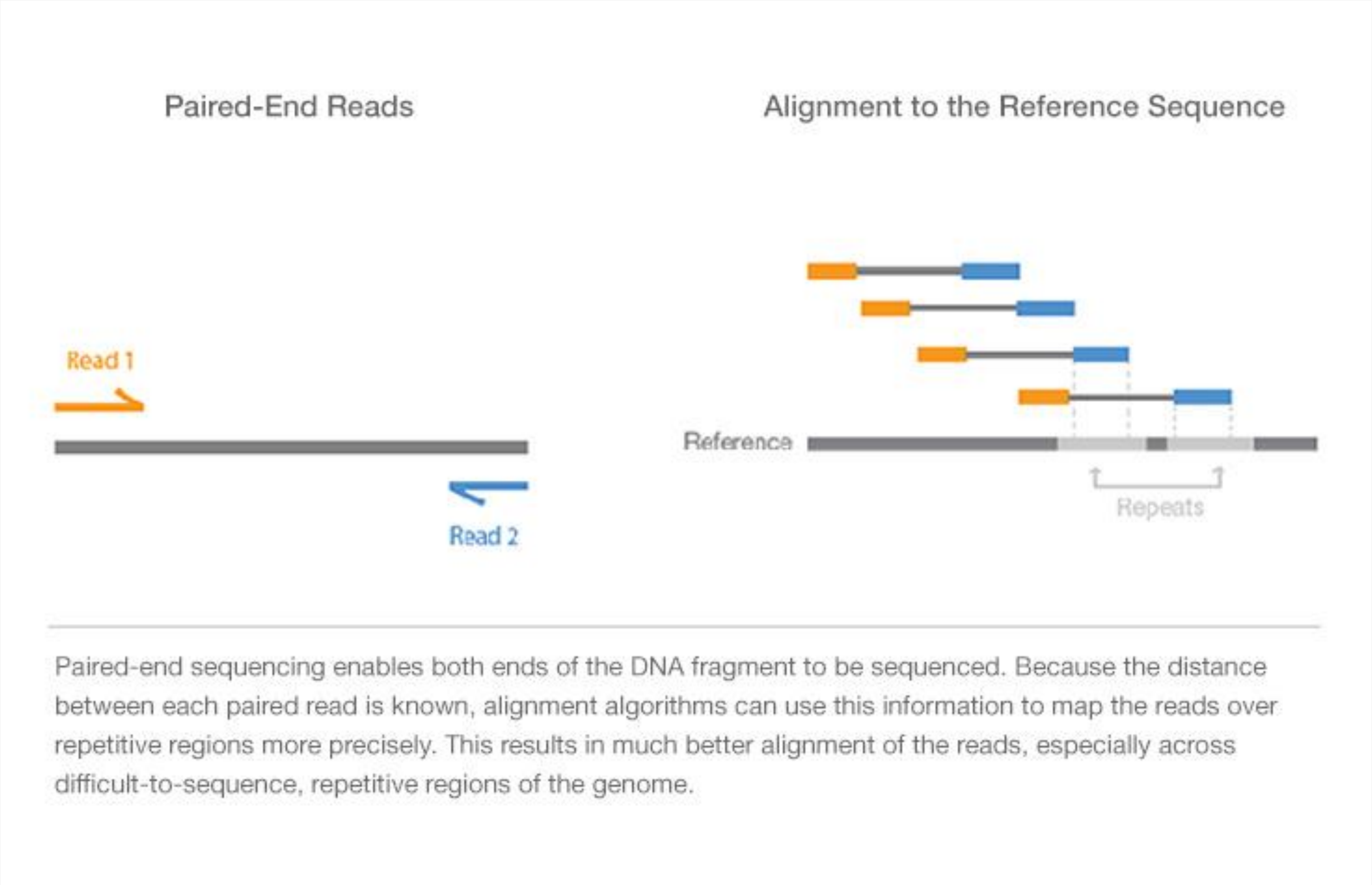
One read per fragment

Paired end sequencing (Illumina)



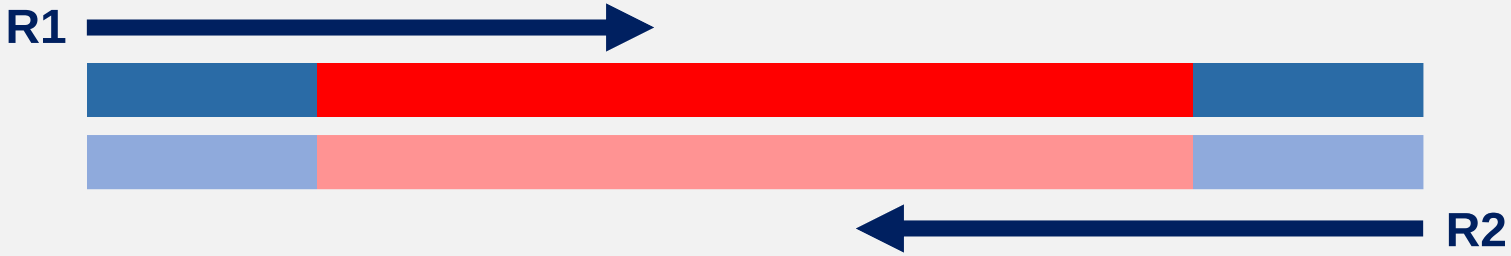
Two reads per fragment

Why paired-end reads?

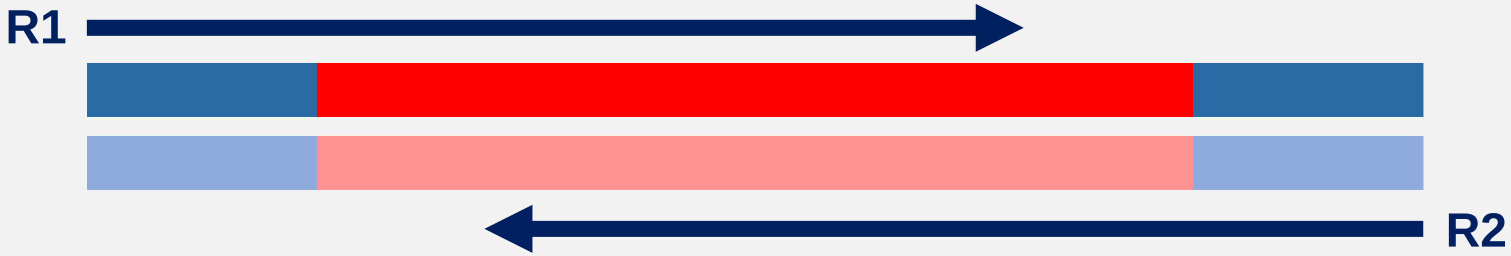


Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Challenges in sequencing short fragments (Illumina)

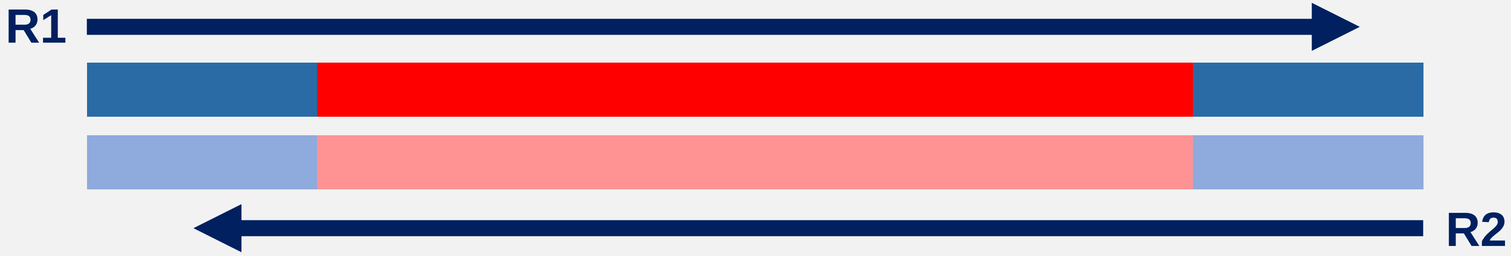


R1 & R2 sequences do not overlap



Partial overlap of R1 & R2

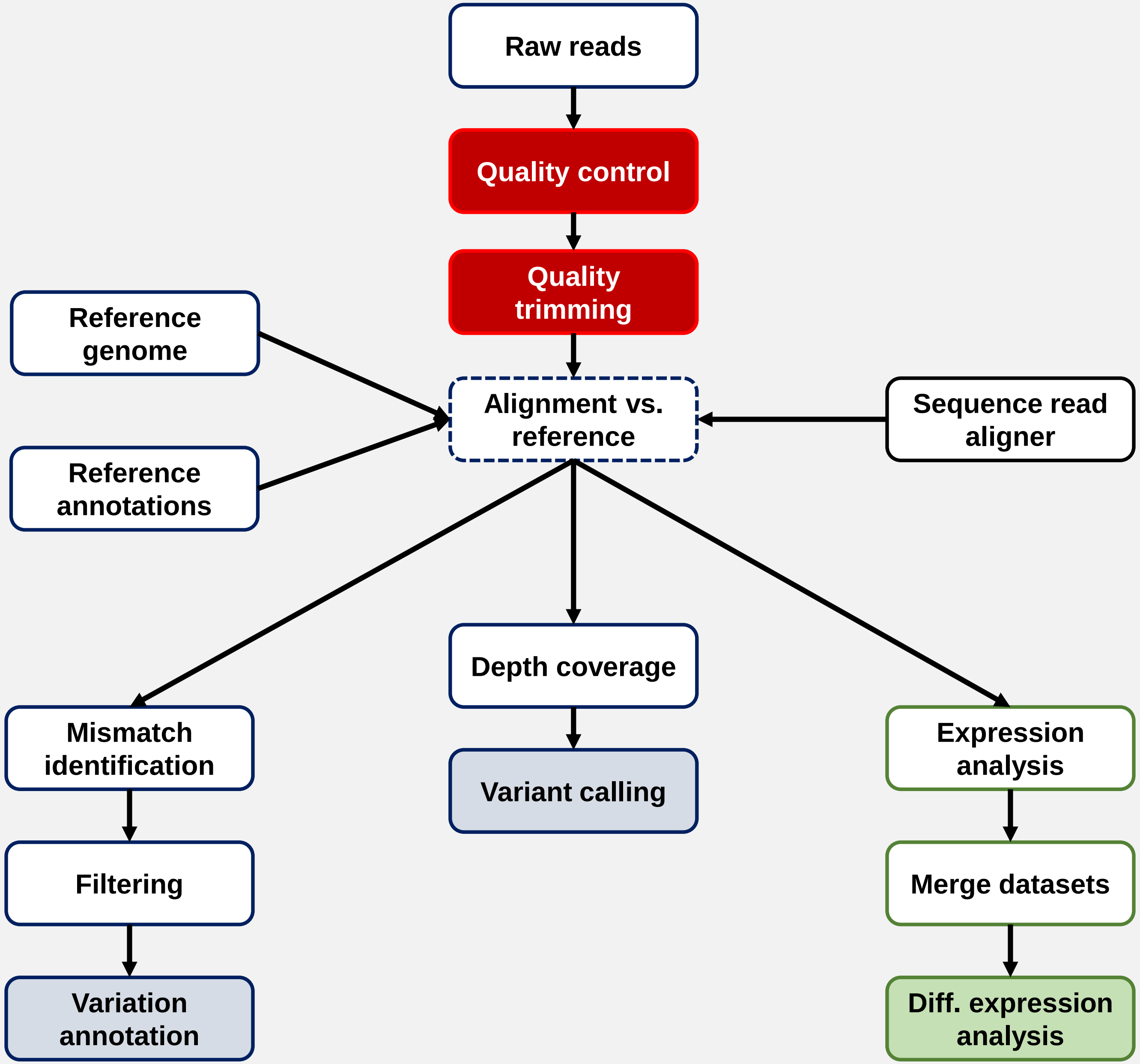
- impacts depth calculation and variant calling)
- remove overlap from one sequence



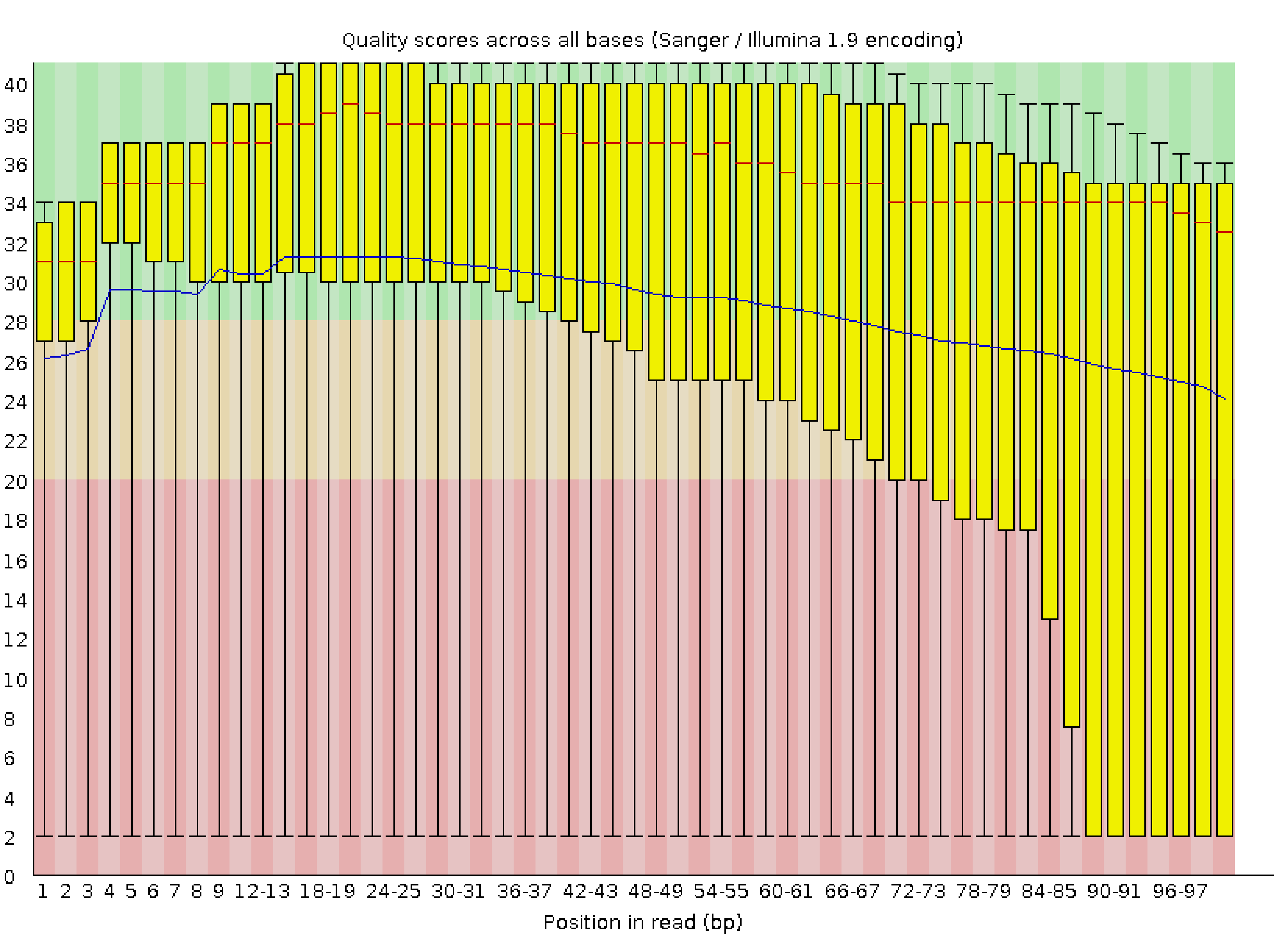
R1 & R2 overlap + adapter read through

- Trim adapter sequence at 3’ end
- Remove overlap from one sequence

Standard workflow



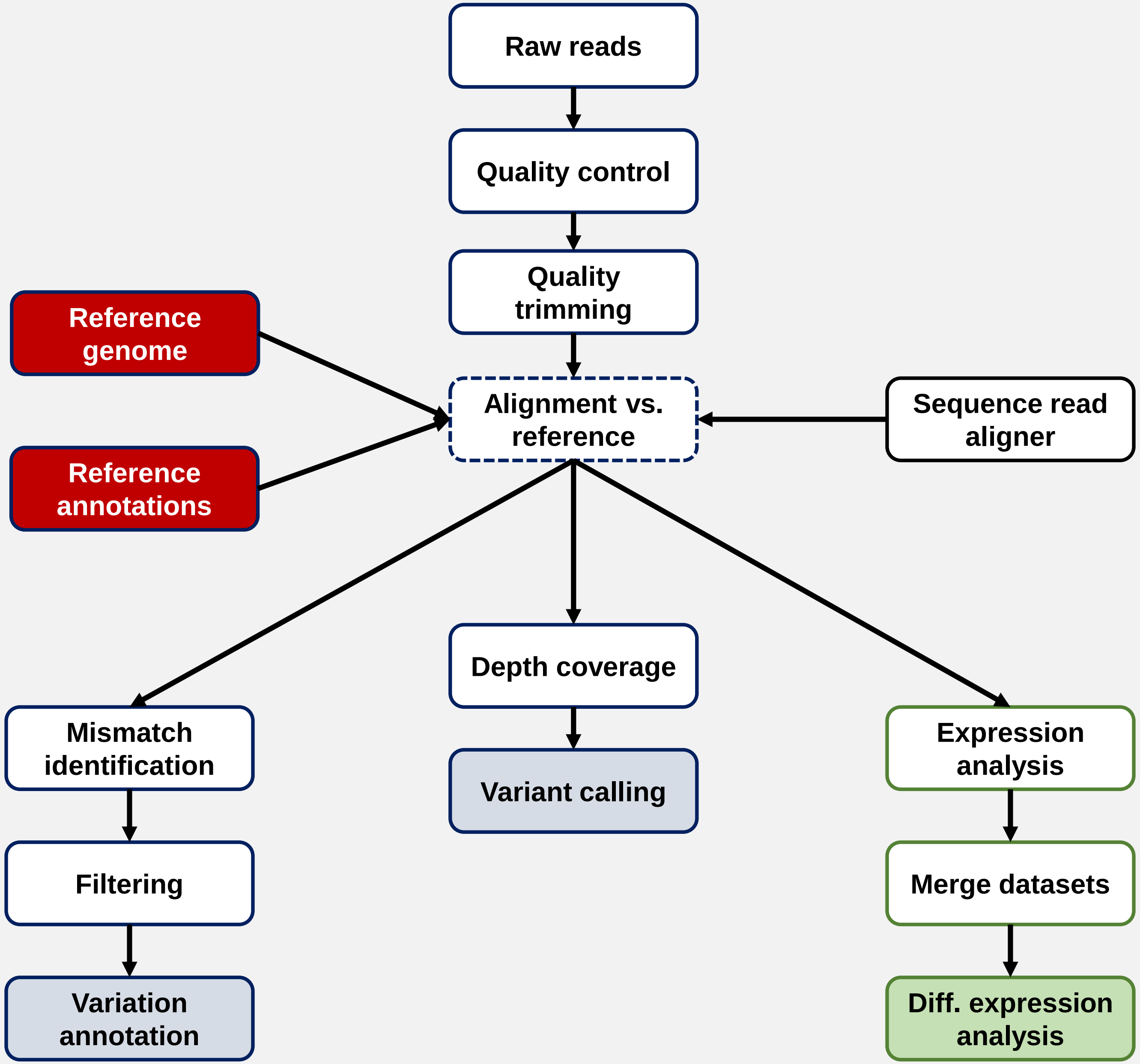
QC of sequence datasets



QC of sequence datasets

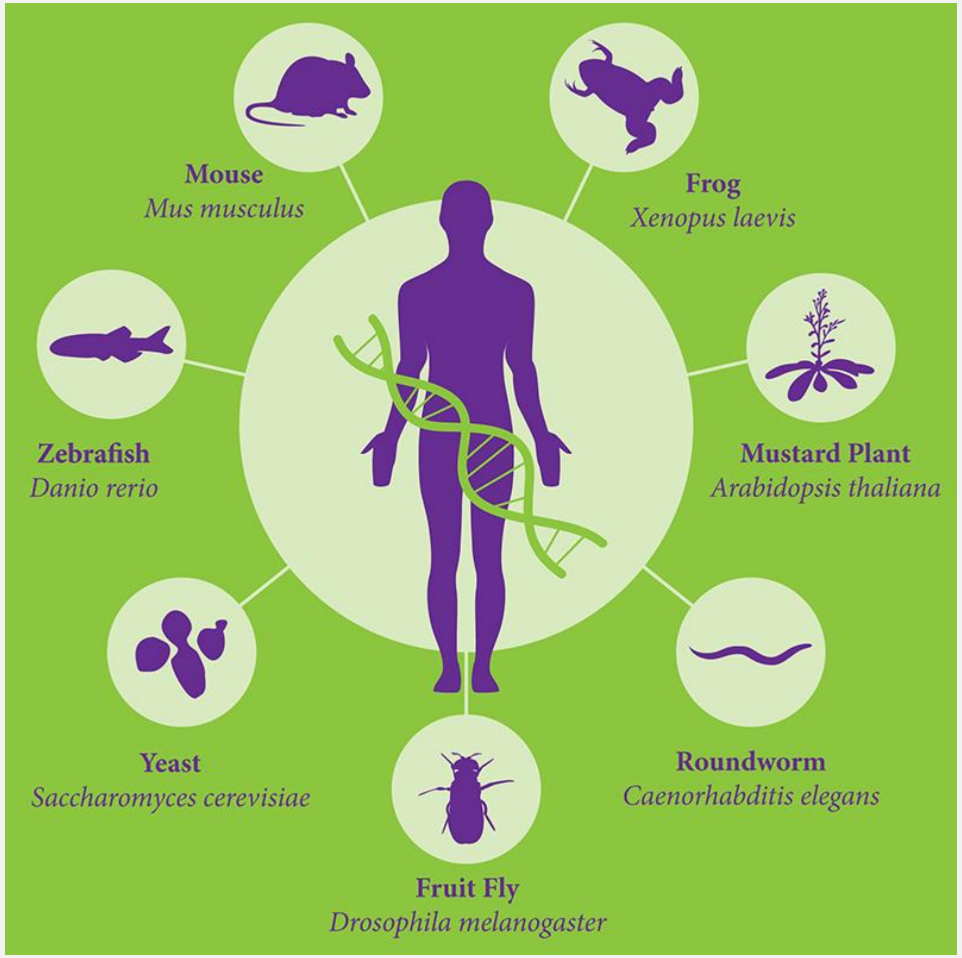
- Used to determine whether sequence reads are ‘compromised’
- Compromised reads are unlikely to align to a target correctly (or at all) and count as wasted/lost data
- Most compromised reads contain adapter sequences and/or low quality basecalls
- Multiple tools exist to examine datasets and trim sequence data to maximize data usage
- Optimal combination remains FASTQC + Trim Galore (feat. CutAdapt)
  - **Top tip for using Trim Galore is to set –q 30 as a flag!**

Standard workflow

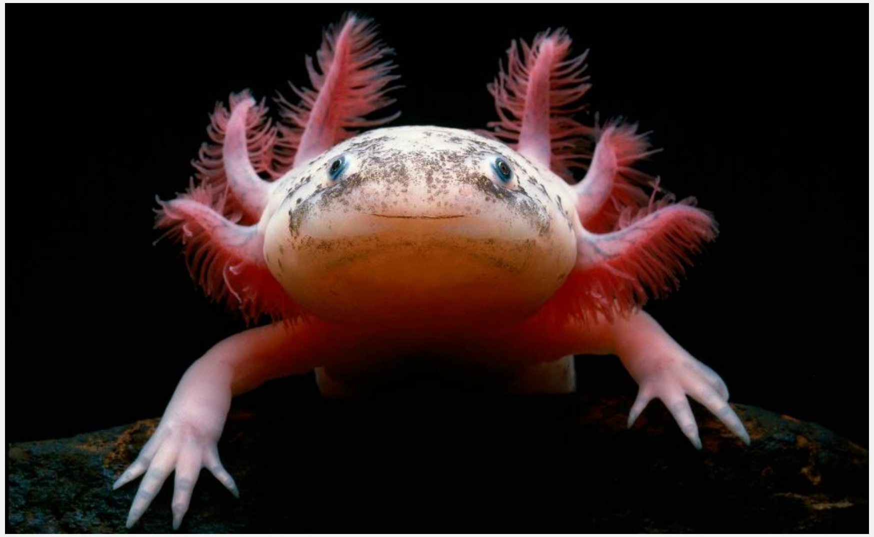


Model vs. non-model organisms

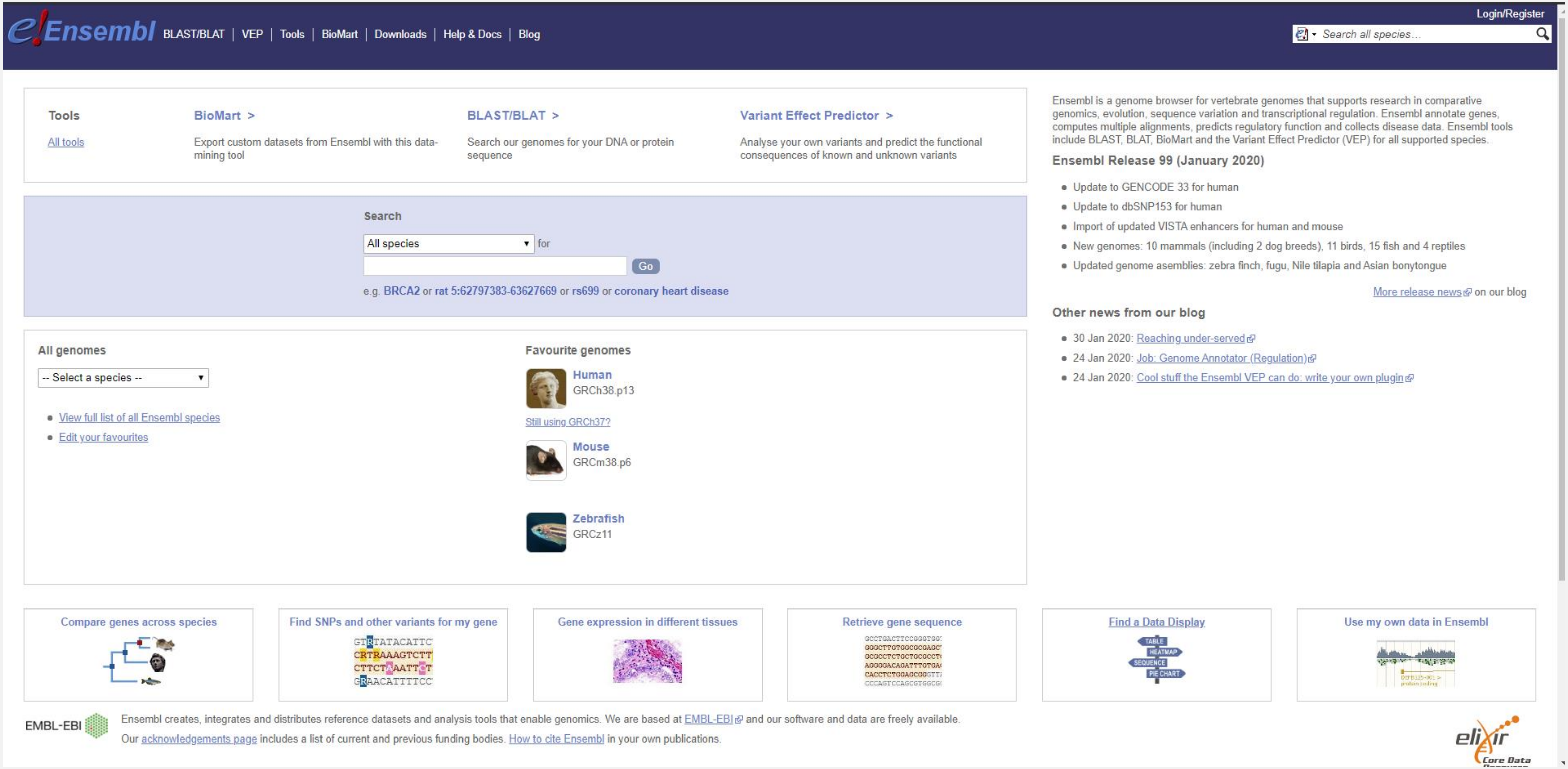
A model organism is a non-human species that is extensively studied to understand particular biological phenomena, with the expectation that discoveries made in the model organism will provide insight into the workings of other organisms.



Non-model organisms are organisms that have not been selected by the research community for extensive study either for historic reasons, or because they lack the features that make model organisms easy to investigate (e.g. they cannot grow in the laboratory, have a long life cycle, low fecundity or poor genetics).



Reference genomes and where to find them (model organisms)



Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
Login/Register
Search all species…

Tools

All tools

BioMart >

Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >

Search our genomes for your DNA or protein sequence

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

Search

All species ▾ for

Go

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

All genomes

-- Select a species -- ▾

- View full list of all Ensembl species
- Edit your favourites

Favourite genomes

Human
GRCh38.p13

Still using GRCh37?

Mouse
GRCm38.p6

Zebrafish
GRCz11

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 99 (January 2020)

- Update to GENCODE 33 for human
- Update to dbSNP153 for human
- Import of updated VISTA enhancers for human and mouse
- New genomes: 10 mammals (including 2 dog breeds), 11 birds, 15 fish and 4 reptiles
- Updated genome asemblies: zebra finch, fugu, Nile tilapia and Asian bonytongue

More release news on our blog

Other news from our blog

- 30 Jan 2020: Reaching under-served
- 24 Jan 2020: Job: Genome Annotator (Regulation)
- 24 Jan 2020: Cool stuff the Ensembl VEP can do: write your own plugin

Compare genes across species

Find SNPs and other variants for my gene

Gene expression in different tissues

Retrieve gene sequence

Find a Data Display

Use my own data in Ensembl

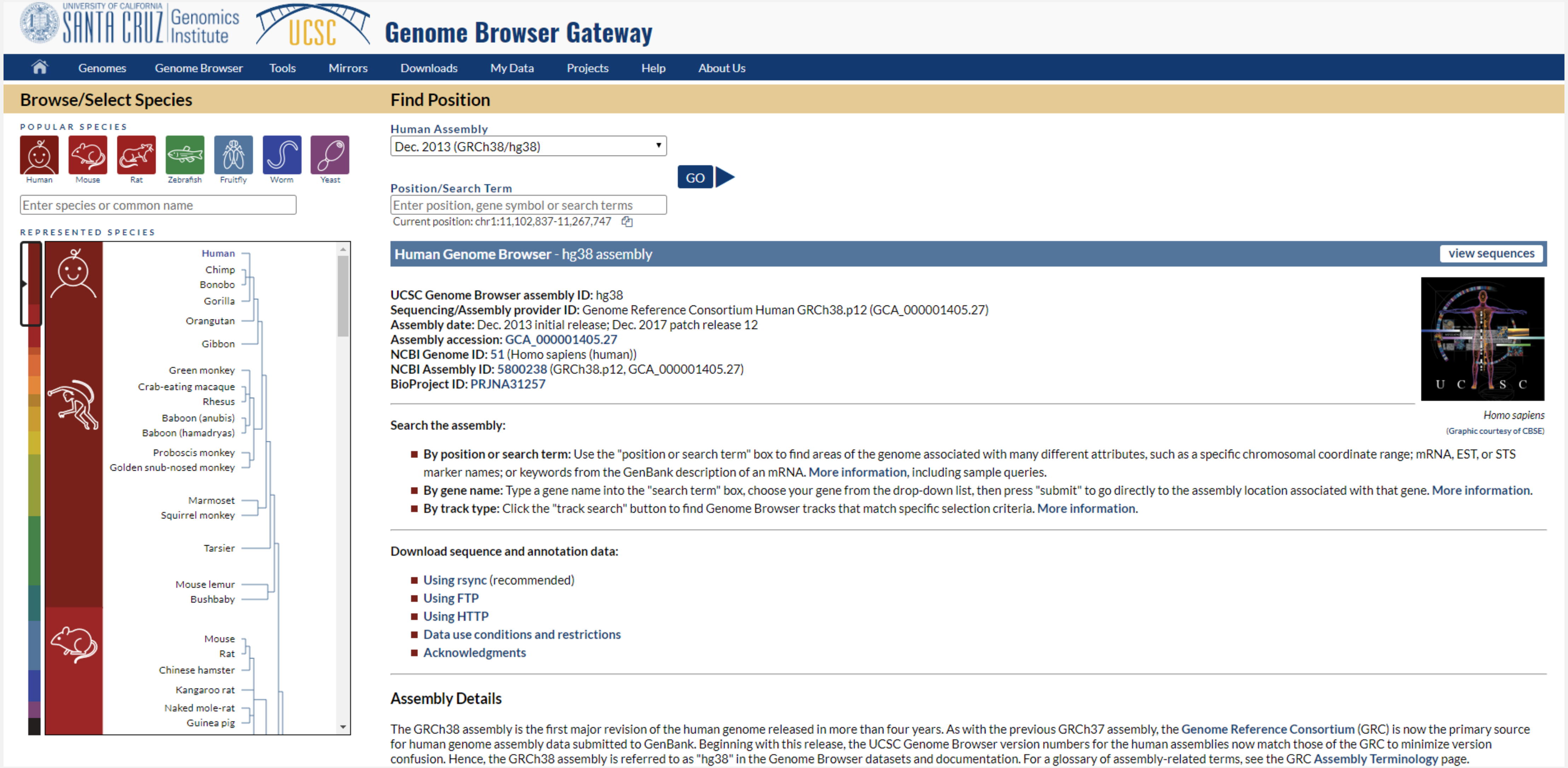
Ensembl creates, integrates and distributes reference datasets and analysis tools that enable genomics. We are based at EMBL-EBI and our software and data are freely available.

Our acknowledgements page includes a list of current and previous funding bodies. How to cite Ensembl in your own publications.

http://www.ensembl.org/

275 genomes [Aardvark – Zig Zag Eel]

Reference genomes and where to find them (model organisms)



UCSC Genome Browser assembly ID: hg38
Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p12 (GCA_000001405.27)
Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 12
Assembly accession: GCA_000001405.27
NCBI Genome ID: 51 (Homo sapiens (human))
NCBI Assembly ID: 5800238 (GRCh38.p12, GCA_000001405.27)
BioProject ID: PRJNA31257

Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. More information, including sample queries.
- **By gene name:** Type a gene name into the "search term" box, choose your gene from the drop-down list, then press "submit" to go directly to the assembly location associated with that gene. More information.
- **By track type:** Click the "track search" button to find Genome Browser tracks that match specific selection criteria. More information.

Download sequence and annotation data:

- Using rsync (recommended)
- Using FTP
- Using HTTP
- Data use conditions and restrictions
- Acknowledgments

Assembly Details

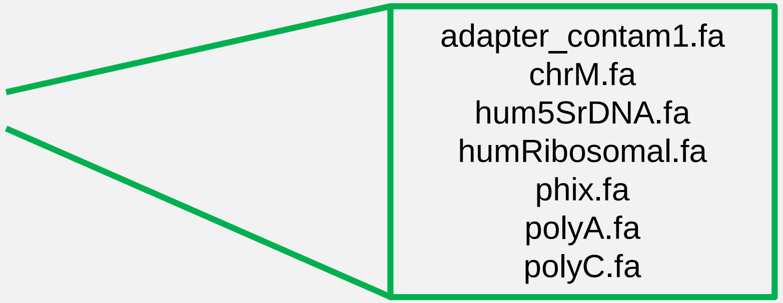
The GRCh38 assembly is the first major revision of the human genome released in more than four years. As with the previous GRCh37 assembly, the Genome Reference Consortium (GRC) is now the primary source for human genome assembly data submitted to GenBank. Beginning with this release, the UCSC Genome Browser version numbers for the human assemblies now match those of the GRC to minimize version confusion. Hence, the GRCh38 assembly is referred to as "hg38" in the Genome Browser datasets and documentation. For a glossary of assembly-related terms, see the GRC Assembly Terminology page.

https://genome.ucsc.edu/cgi-bin/hgGateway

Reference genomes and where to find them

FASTA and pre-compiled INDEXES

- Homo_sapiens/UCSC/hg38/Sequence/AbundantSequences
- Homo_sapiens/UCSC/hg38/Sequence/BowtieIndex
- Homo_sapiens/UCSC/hg38/Sequence/Bowtie2Index
- Homo_sapiens/UCSC/hg38/Sequence/BWAIndex
- Homo_sapiens/UCSC/hg38/Sequence/Chromosomes (individual fasta for each chromosome)
- Homo_sapiens/UCSC/hg38/Sequence/WholeGenomeFasta (Single fasta w/ all chromosome)



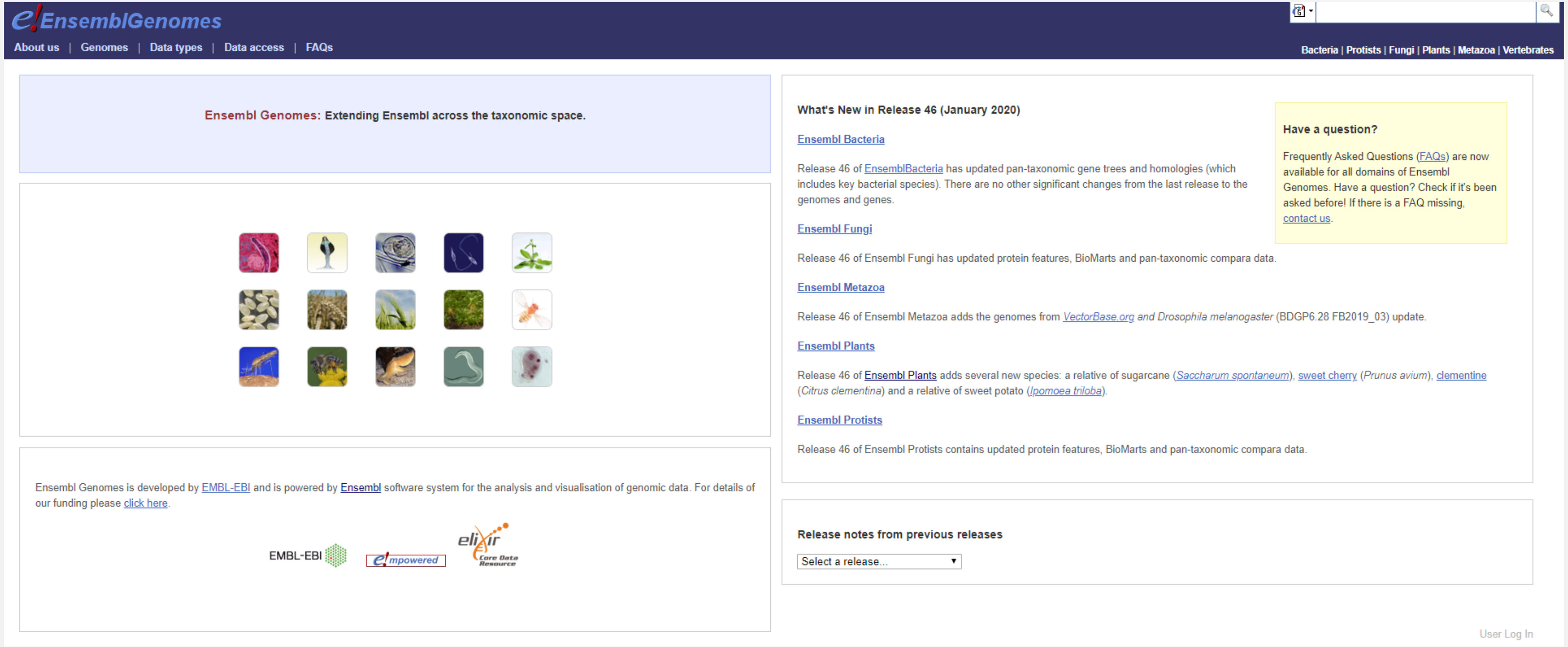
adapter_contam1.fa
chrM.fa
hum5SrDNA.fa
humRibosomal.fa
phix.fa
polyA.fa
polyC.fa

Gene annotation files (GTF and small RNA FASTA)

- Homo_sapiens/UCSC/hg38/Annotation/Genes
- Homo_sapiens/UCSC/hg38/Annotation/Genes.gencode
- Homo_sapiens/UCSC/hg38/Annotation/SmallRNA

Top tip: watch out for versioning! hg38 is the latest available version of the human genome: 2013 (hg38) vs 2009 (hg19)

Reference genomes and where to find them (non-model organisms)



EnsemblGenomes
About us | Genomes | Data types | Data access | FAQs
Bacteria | Protists | Fungi | Plants | Metazoa | Vertebrates

Ensembl Genomes: Extending Ensembl across the taxonomic space.

Ensembl Genomes is developed by EMBL-EBI and is powered by Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please click here.

EMBL-EBI
e!mpowered
elixir Core Data Resource

What's New in Release 46 (January 2020)

Ensembl Bacteria

Release 46 of EnsemblBacteria has updated pan-taxonomic gene trees and homologies (which includes key bacterial species). There are no other significant changes from the last release to the genomes and genes.

Ensembl Fungi

Release 46 of Ensembl Fungi has updated protein features, BioMarts and pan-taxonomic compara data.

Ensembl Metazoa

Release 46 of Ensembl Metazoa adds the genomes from *VectorBase.org and Drosophila melanogaster* (BDGP6.28 FB2019_03) update.

Ensembl Plants

Release 46 of Ensembl Plants adds several new species: a relative of sugarcane (*Saccharum spontaneum*), sweet cherry (*Prunus avium*), clementine (*Citrus clementina*) and a relative of sweet potato (*Ipomoea triloba*).

Ensembl Protists

Release 46 of Ensembl Protists contains updated protein features, BioMarts and pan-taxonomic compara data.

Have a question?

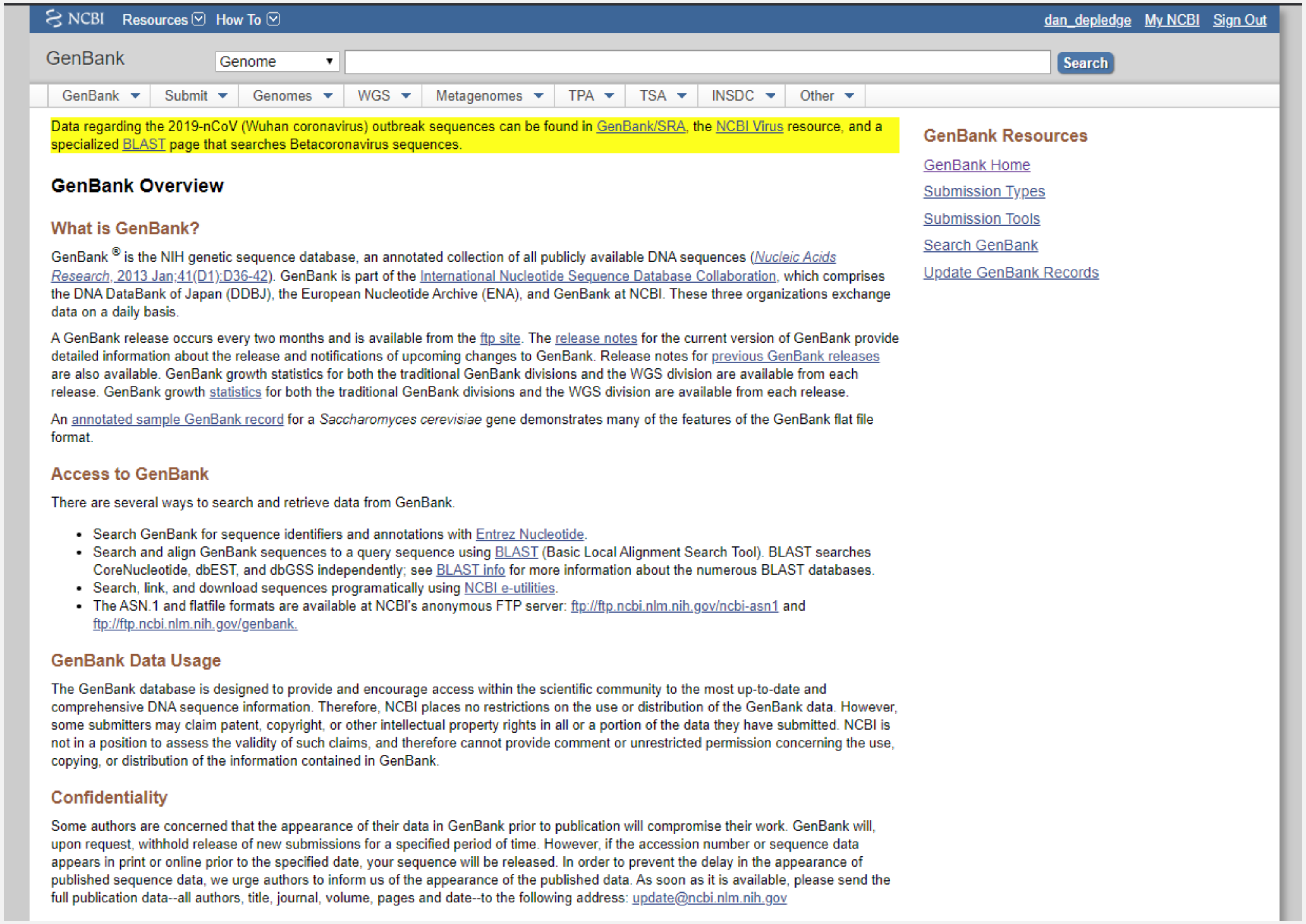
Frequently Asked Questions (FAQs) are now available for all domains of Ensembl Genomes. Have a question? Check if it's been asked before! If there is a FAQ missing, contact us.

Release notes from previous releases

http://ensemblgenomes.org/

50,000+ genomes

Reference genomes and where to find them (non-model organisms)



Data regarding the 2019-nCoV (Wuhan coronavirus) outbreak sequences can be found in GenBank/SRA, the NCBI Virus resource, and a specialized BLAST page that searches Betacoronavirus sequences.

GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see BLAST info for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using NCBI e-utilities.
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank.

GenBank Data Usage

The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.

Confidentiality

Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work. GenBank will, upon request, withhold release of new submissions for a specified period of time. However, if the accession number or sequence data appears in print or online prior to the specified date, your sequence will be released. In order to prevent the delay in the appearance of published sequence data, we urge authors to inform us of the appearance of the published data. As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following address: update@ncbi.nlm.nih.gov

GenBank Resources

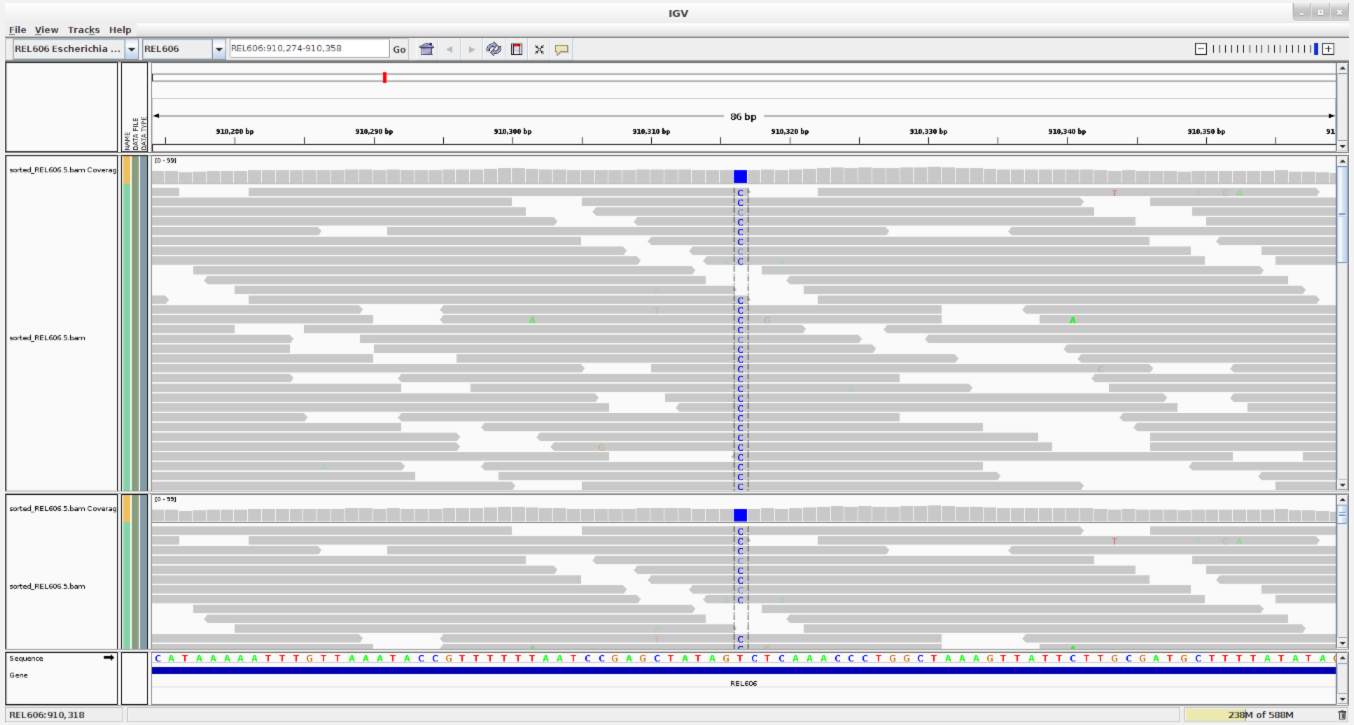
- GenBank Home
- Submission Types
- Submission Tools
- Search GenBank
- Update GenBank Records

https://www.ncbi.nlm.nih.gov/genbank/ - only if you are desperate (non-model organism)

Challenges of visualizing big and small data

- Why use visual inspection?
  - Simplest way to troubleshoot
  - Helps to confirm effectiveness of alignment strategy – are your parameters causing problems?
  - Inspect read alignments at interesting locations (e.g. SNPs, transcription termini, antisense transcription)

- Limitations of visual inspection
  - BAM files are often too big to load into memory (unless using a very high spec computer)
  - Window of analysis is often too small or insufficiently detailed
  - Screen grabs make for lousy images (not publication quality)



REL606 Escherichia ...

REL606

REL606:910,274-910,358

REL606:910,318

238M of 588M

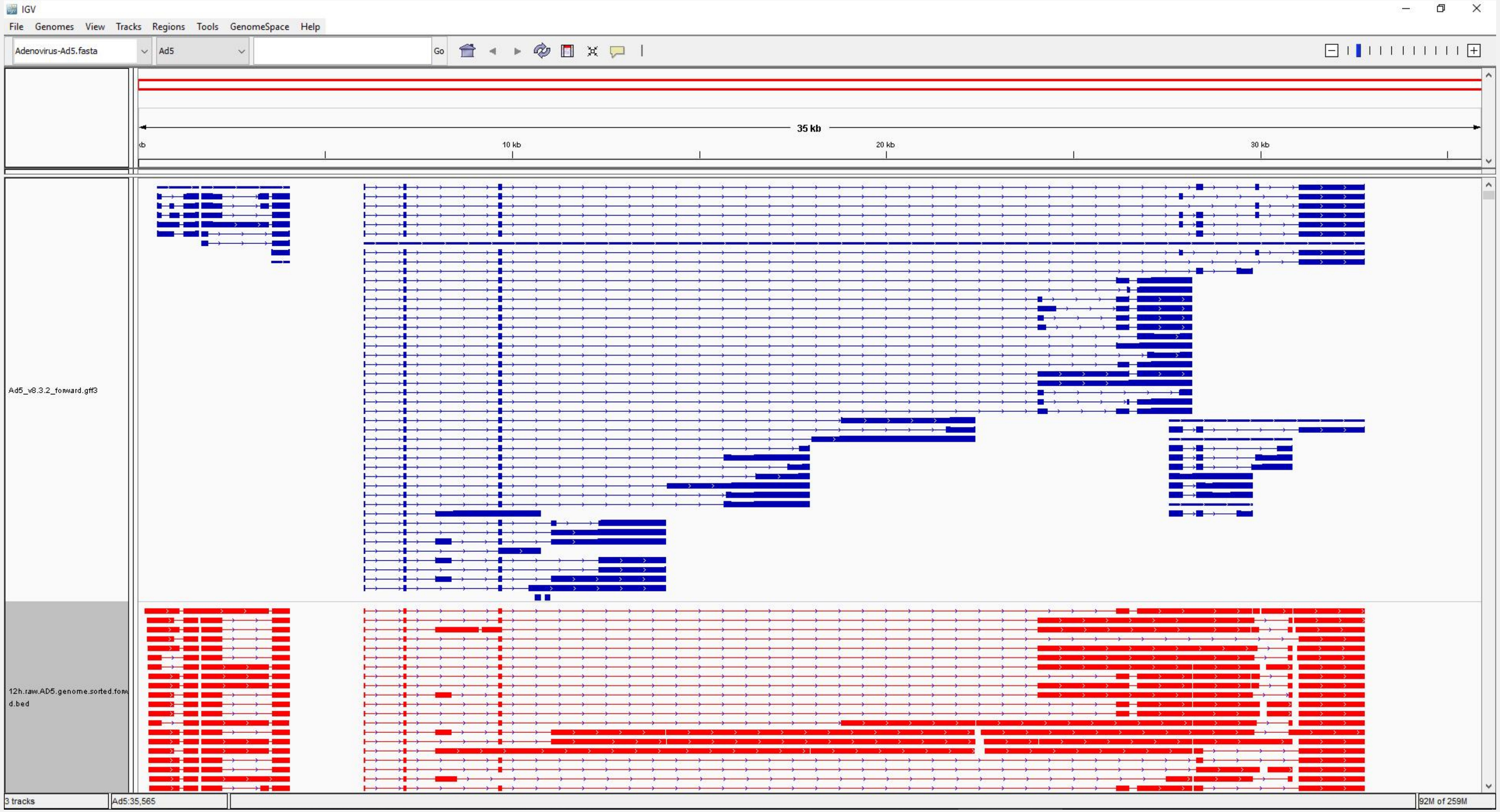
IGV
File
View
Tracks
Help
Go
NAME
DATA FILE
DATA TYPE
86 bp
910,280 bp
910,290 bp
910,300 bp
910,310 bp
910,320 bp
910,330 bp
910,340 bp
910,350 bp
[0 - 59]
sorted_REL606.5.bam Coverag
sorted_REL606.5.bam
Sequence
Gene
C A T A A A A A T T T G T T A A A T A C C G T T T T T T A A T C C G A G C T A T A G T C T C A A A C C C T G G C T A A A G T T A T T C T T G C G A T G C T T T T T A T A T A
REL606

List of alignment viewers

- IGV
- UCSC Genome Browser
- Artemis
- Ugene
- Tablet
- tview (SAMtools)
  - Literally and hilariously text based

The Integrative Genomics Viewer (IGV)

- Remains the simplest (i.e. user-friendly) solution for both model and custom genomes



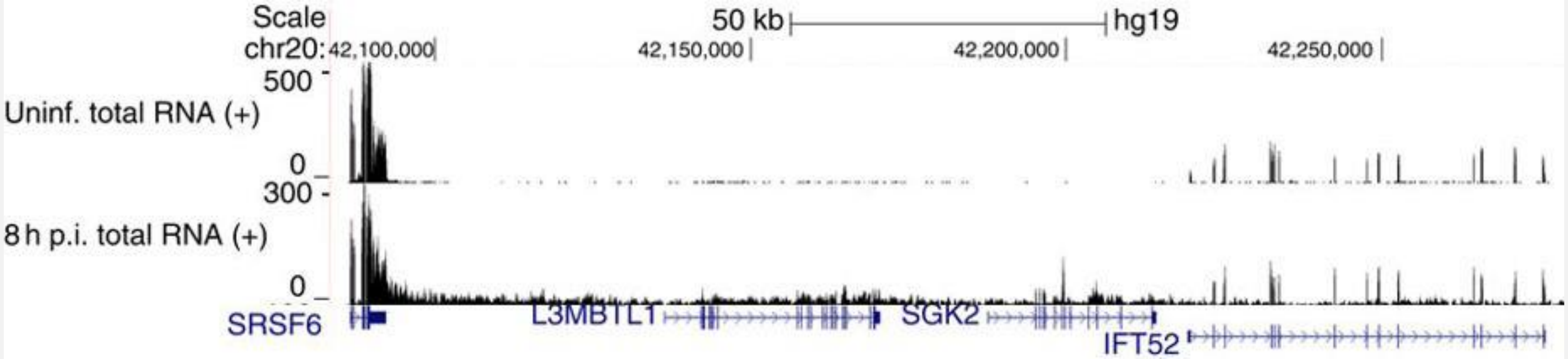
Annotation track

Data track

Gviz

- Pros
  - Simple to learn (requires R)
  - Great at both high level and granular levels
  - Can generate publication quality images
  - Compatible with multiple file formats (BAM/BED)
  - Plenty of support available via google / forums / developers
  - Good support for *popular* genomes

- Cons
  - Tricky to get working with non-model genomes



Getting to grips with Gviz

General manual

- https://bioconductor.org/packages/release/bioc/vignettes/Gviz/inst/doc/Gviz.pdf

Useful tutorial #1

- http://www.sthda.com/english/wiki/gviz-visualize-genomic-data

Useful tutorial #2

- https://davetang.org/muse/2013/10/03/using-gviz/

Assignment #1

Investigating the cellular response to exogenous dsDNA

Introduction

Double-stranded DNA (dsDNA) in the cytosol of human cells stimulates the type 1 interferon (IFN) response, a component of innate immunity that is active against invading pathogens and many cancers. Over the course of Assignment #1 and Assignment #2, we will examine the host genes that are transcriptionally regulated upon detection of invading dsDNA.

Assignment #1 will focus on (1) finding and downloading stranded paired-end RNA-Seq datasets from a recently published study, (2) performing basic QC and alignment of these datasets, and (3) visualizing the read coverage across several regions of the genome. The aim is to become familiar with a range of common tools used in the processing of NGS data.

Note that these alignments will be carried over into Assignment #2 in which you will need to undertake a typical differential gene expression analysis using read counts generated from the aligned data.

Assignment #1

1. **Download six datasets (3 x dsDNA 12 hr bioreps and 3 x CTRL 12 hr bioreps) from the SRA. These are associated with the BioProject ID PRJNA451188.**
   - List of SRA IDs: SRR7049616, SRR7049615, SRR7049609, SRR7049610, SRR7049611, SRR7049612
   - Hint #1 – use sra-tools on BigPurple to download data
   - Hint #2 – Use the SRA run selector to get a simple overview of datasets and to filter for those you are interested in
   - Hint #3 – ensure that each dataset downloaded comprises two files, one with the forward reads (R1) and one with the reverse (R2).

2. **Examine dataset using FASTQC and perform adapter + quality trimming with TrimGalore**
   - Examine all downloaded files
   - This requires using TrimGalore and piping the output into FastQC (BigPurple)
   - Remember to run TrimGalore in paired-end mode and consider appropriate Phred score for trimming

3. **Align all datasets to the human genome (UCSC HG38 version) using a spliced aligner**
   - This requires downloading the correct copy of the human genome and aligning paired-end datasets against it in a sensible manner (i.e. bowtie2, bbmap)
   - Note, all six datasets must be aligned (separately)

3. **Visualize read coverage across the following loci: IFN1B, IFIT2, and ISG15**
   - This requires loading alignment data (BAM, BED, BIGWIG) into local installations of IGV and/or using the Gviz package to make useful plots.
   - Remember to consider strandedness in alignments

Assignment #1

Useful notes

- Downloading a single dataset from the SRA can take several hours. Consider the use of parallelization.

- Alignment is a non-trivial process. Think carefully about the software and parameters you use (i.e. read the manuals!)

- IGV is a simple way to look at read coverage but does not produce publication quality images…

- Gviz requires more time/patience/fiddling but produces publication quality images and is far more flexible than IGV.

Strandedness refers to the fact that individual RNA-Seq reads can be assigned to a specific DNA strand. In a genic context, one would expect RNA-Seq reads for an expressed gene to predominantly align to the strand encoding that gene and to only align to the exonic regions of that gene

Thank you for your attention

Questions?