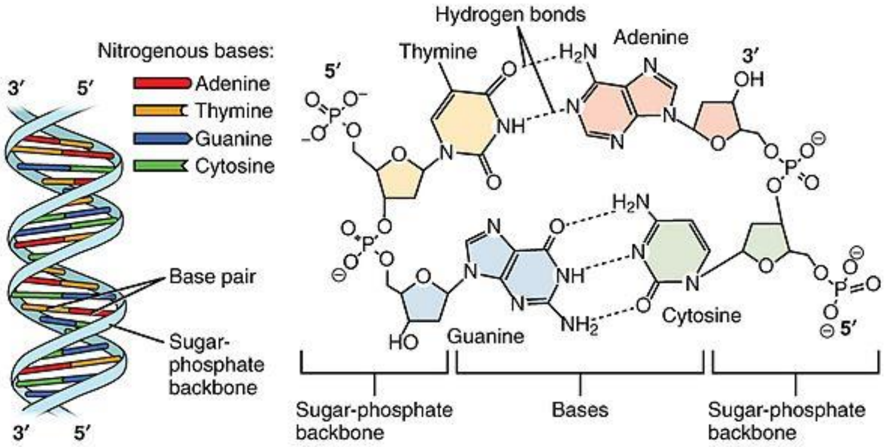
Next Generation Sequencing

Next Generation Sequencing

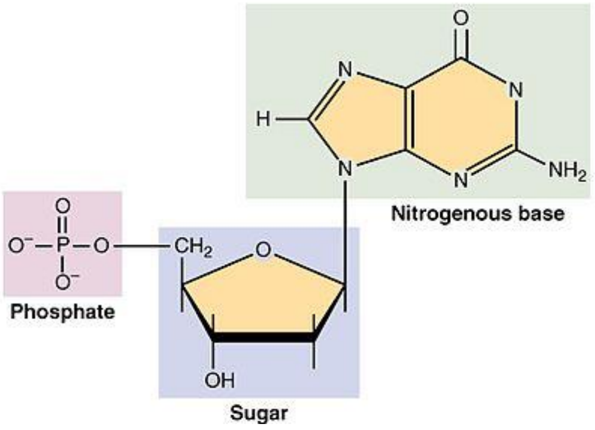
- Back to the basics
- “Previous generation” sequencing: Sanger
- NGS technologies
- Library preparation
- Data Analysis

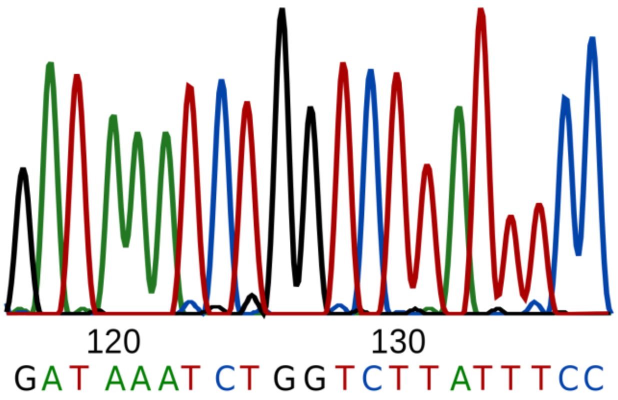
Back to the basics

DNA



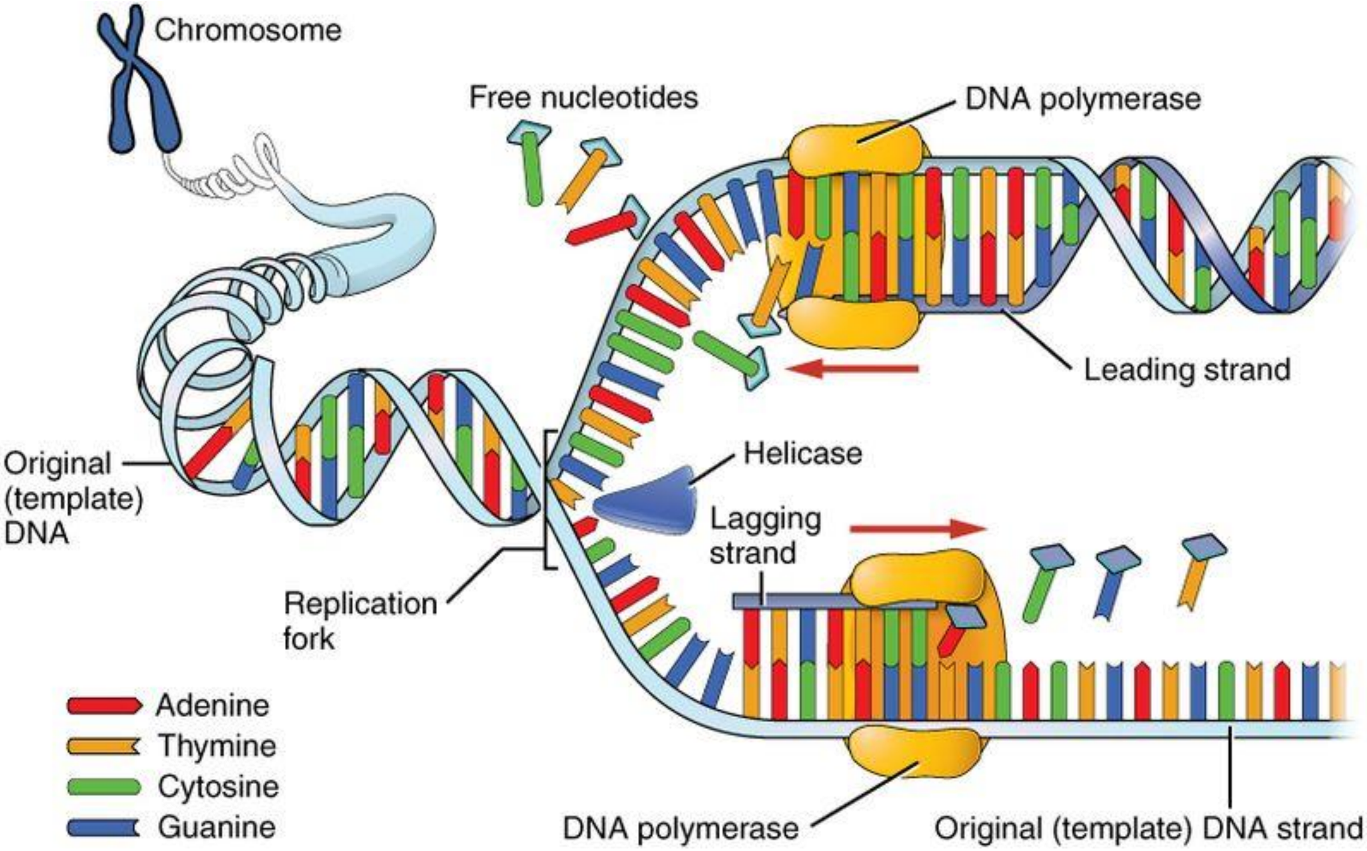
DNA Sequencing: determining the sequence of nucleotides (As, Ts, Cs, and Gs) in a piece of DNA



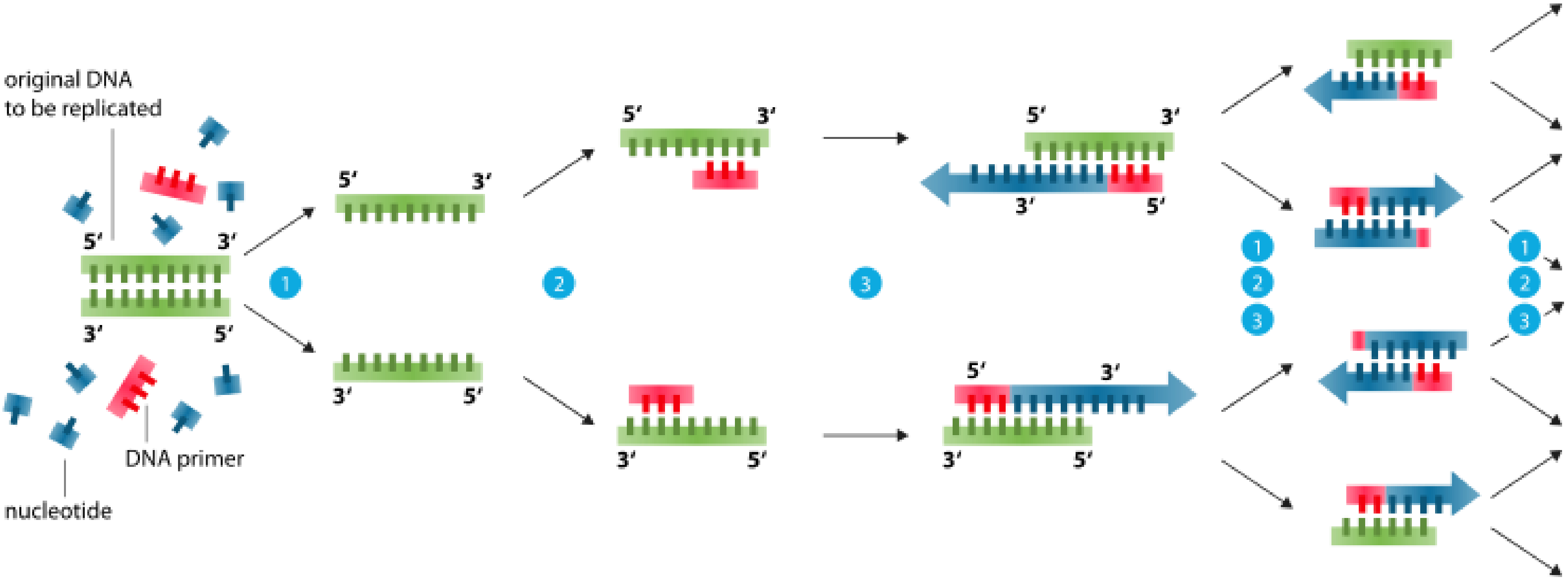


(From now on: nucleotide = base)

DNA replication



Polymerase Chain Reaction (PCR)



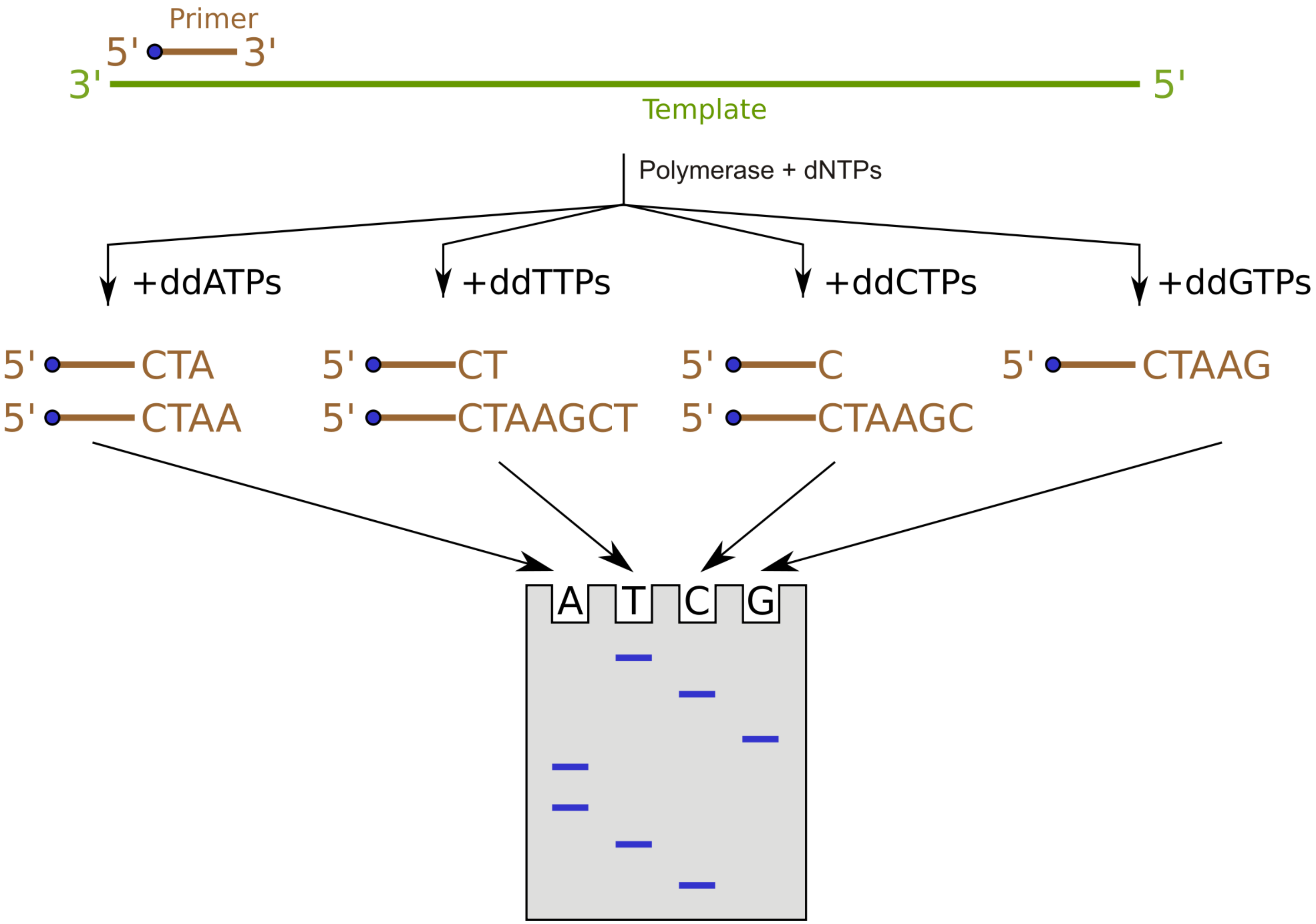
1 **Denaturation** at 94-96°C

2 **Annealing** at ~68°C

3 **Elongation** at ca. 72 °C

Sanger sequencing

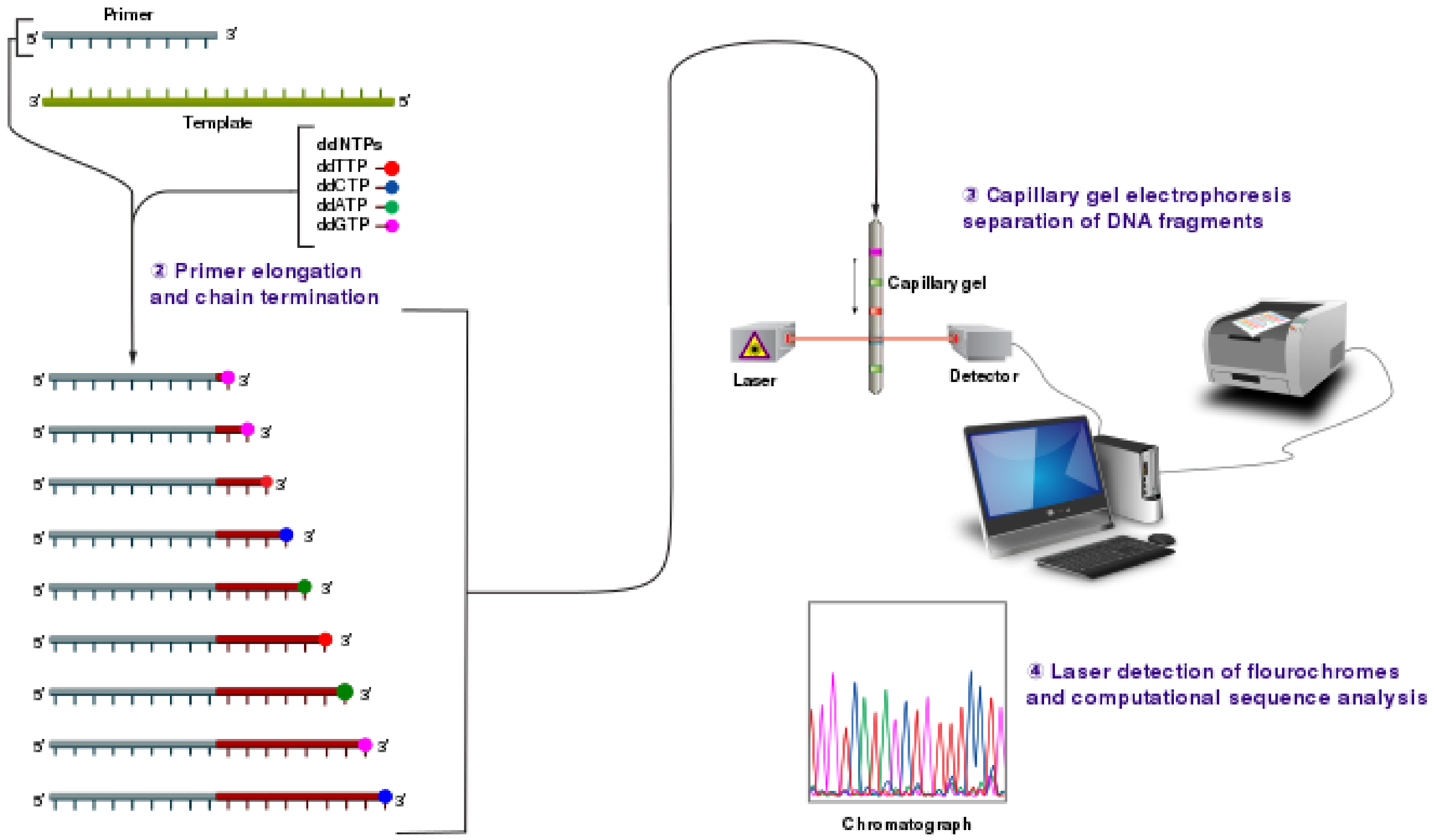
Sanger sequencing in the old times



Sanger sequencing today

① Reaction mixture

- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)



② Primer elongation and chain termination

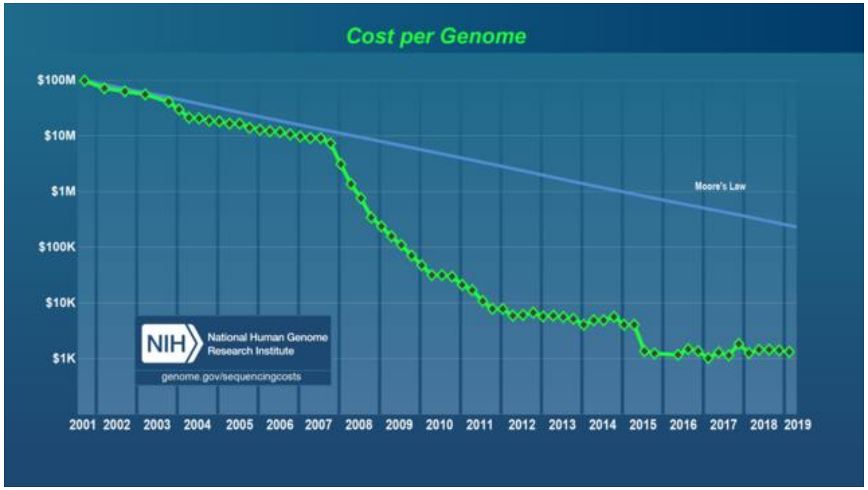
③ Capillary gel electrophoresis separation of DNA fragments

④ Laser detection of flourochromes and computational sequence analysis

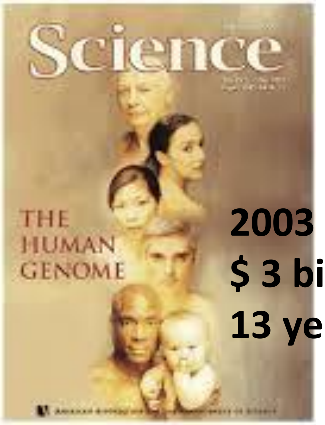
NGS

Next Generation Sequencing

- Highly **parallel**: it can sequence millions of fragments simultaneously per run. (With Sanger only one DNA fragment at a time)
  - Faster
  - Cheaper
  - Less DNA required
  - Higher throughput



From National Human Genome Research Institute (NHGRI)



2003
$ 3 billion
13 years

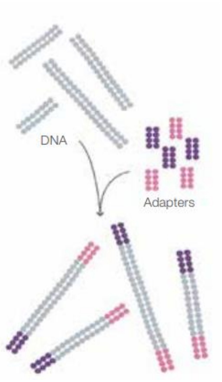
NGS technologies

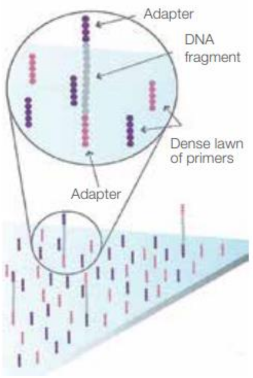
- Short reads: “Second Generation Sequencing”
  - Illumina
  - IonTorrent

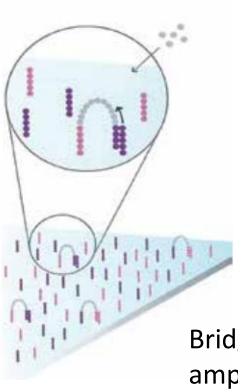
- Long reads: “Third Generation Sequencing”
  - PacBio
  - Oxford Nanopore

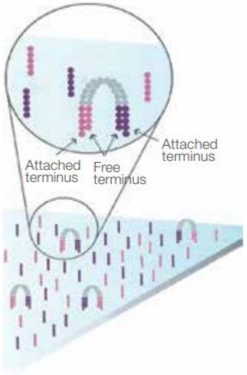
How do the different technologies work?

Illumina: bridge amplification, sequencing by synthesis, imaging

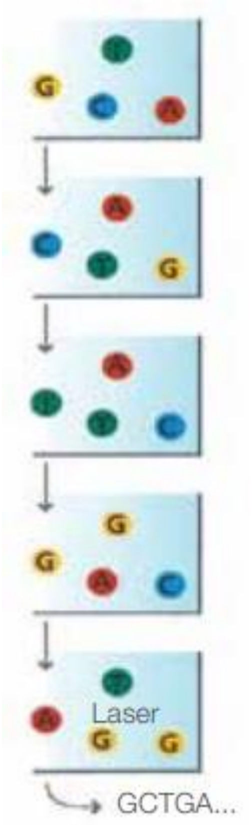


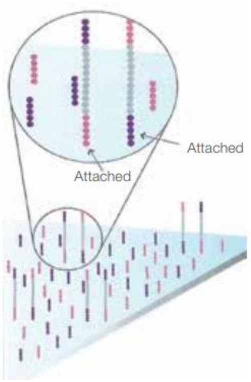


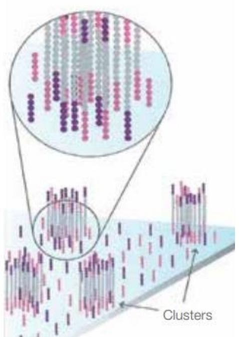




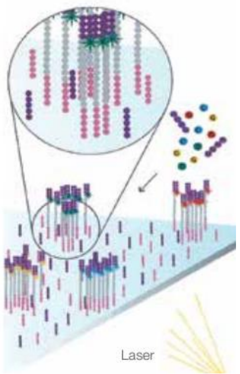
Bridge amplification

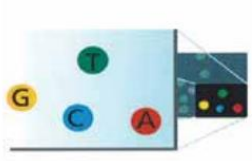






Cluster generation



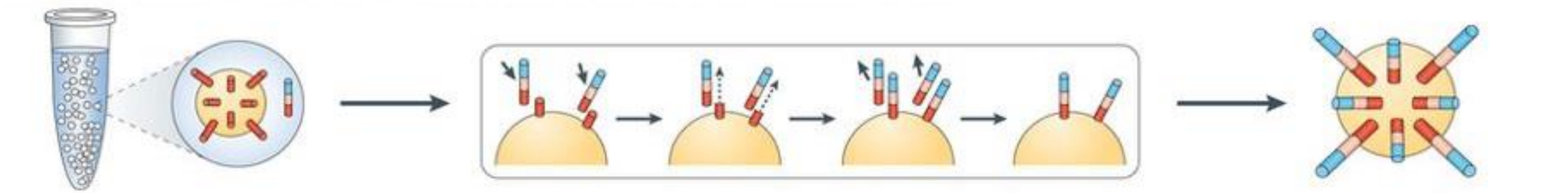


Sequencing by synthesis

https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf

https://www.youtube.com/watch?v=womKfikWlxM

IonTorrent: Emulsion PCR, sequencing by synthesis, changes on pH

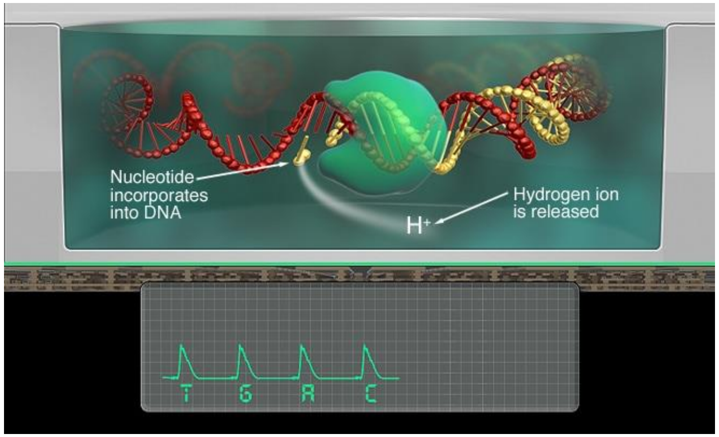


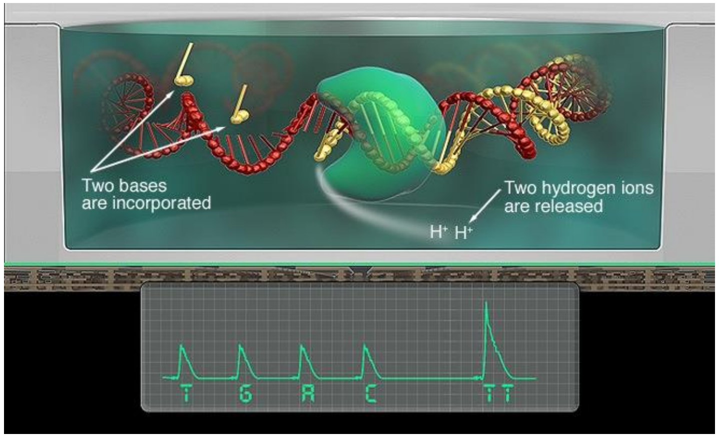
Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates

Final product
100–200 million beads with thousands of bound template

Nature Reviews Genetics 17, 333–351 (2016)

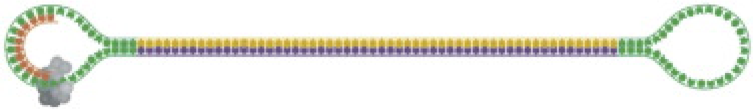




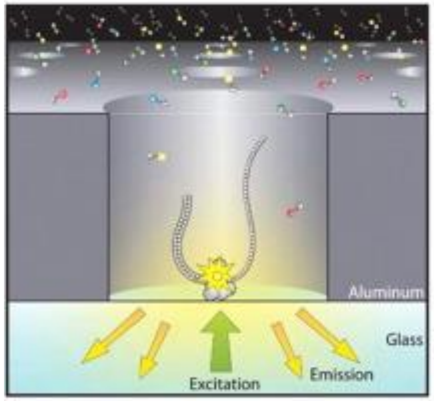
https://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html

https://www.youtube.com/watch?v=WYBzbxIfuKs

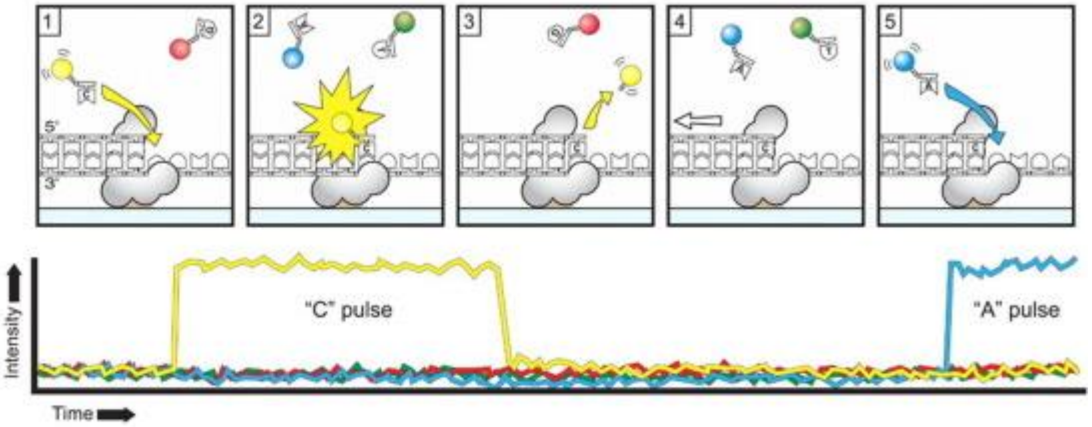
PacBio: Single Molecule Real Time sequencing Sequencing by synthesis, imaging.



Two hairpin adapters circularize the dsDNA



Polymerase is attached to the chamber, template DNA is used to incorporate fluorescent labelled nucleotides.

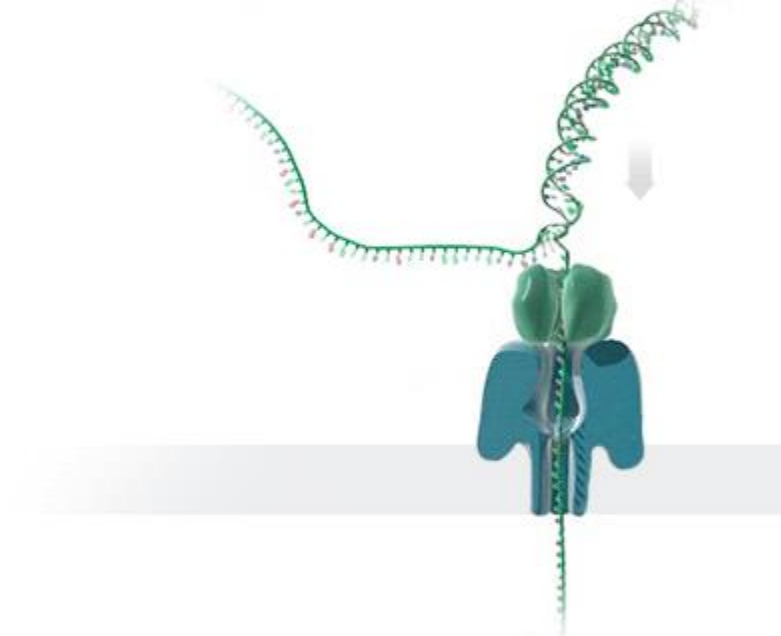


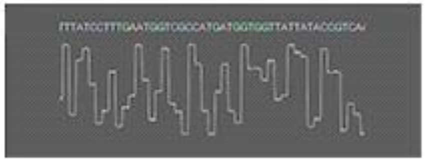
Light emission is recorded

https://www.youtube.com/watch?v=WMZmG00uhwU&feature=youtu.be

Genomics, Proteomics & Bioinformatics 13, 278-289 (2015)

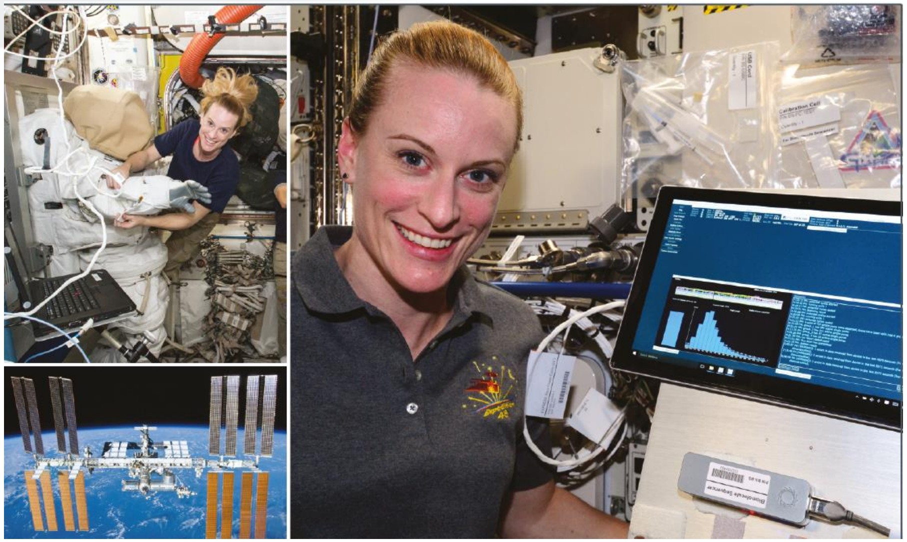
Oxford Nanopore: changes in current as DNA passes through a pore





MinION: so small it can be taken on a space mission

https://nanoporetech.com/how-it-works



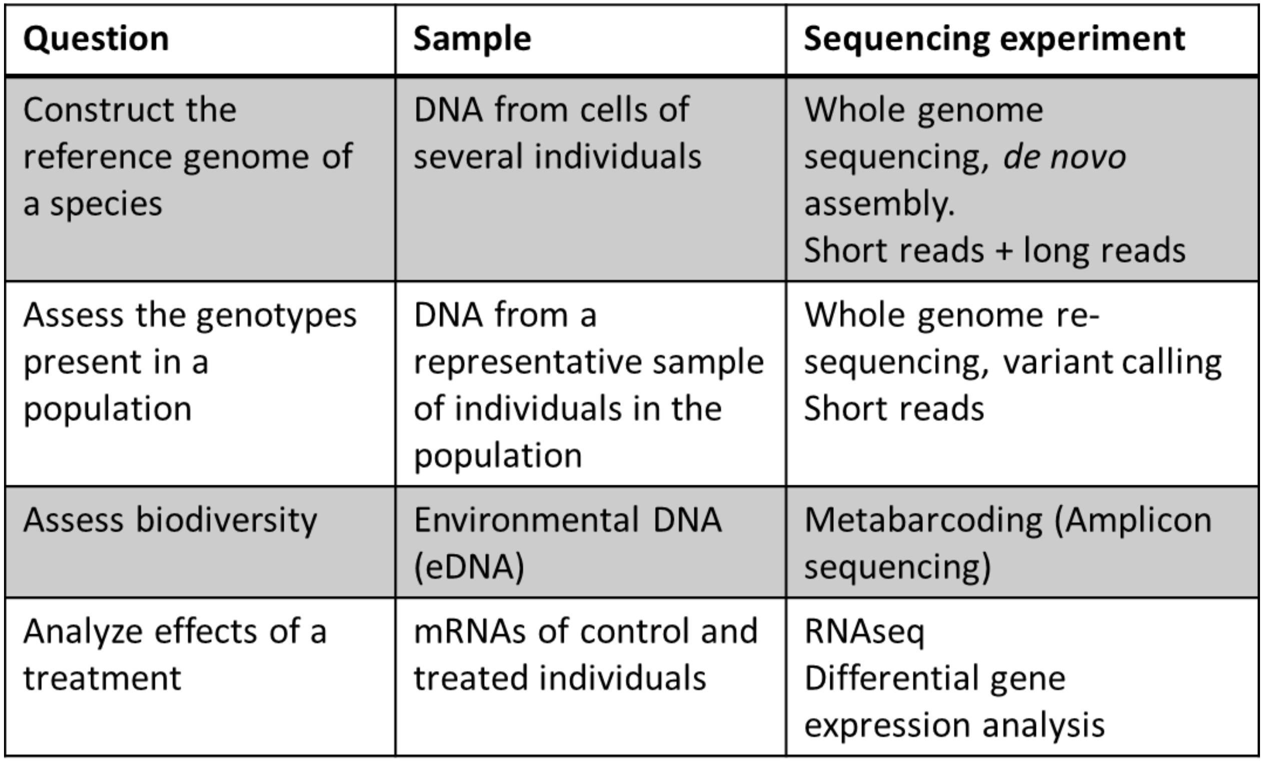
Kate Rubins on the ISS, 2016

Workflow

1) Experimental design: what is the question and what’s the best way of answering it?
2) Sampling.
3) Library preparation: prepare the DNA (or RNA) to be sequenced. Protocol depends on the sample, the platform that we are using, and the question that we want to answer.
4) Sequencing: the machine does it and returns a list of all sequences (reads) in FASTQ format.
5) Analysis: bioinformatics pipelines, depend on the sample and the question.

Experimental design

- All the other steps are going to depend on this
- Careful designing saves time and money (and saves us from frustration!)
- Examples:



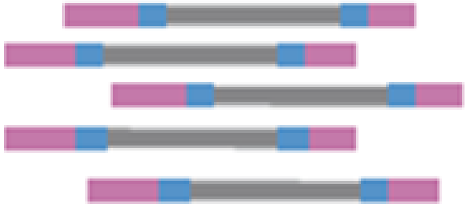
Question	Sample	Sequencing experiment
Construct the reference genome of a species	DNA from cells of several individuals	Whole genome sequencing, *de novo* assembly. Short reads + long reads
Assess the genotypes present in a population	DNA from a representative sample of individuals in the population	Whole genome re-sequencing, variant calling Short reads
Assess biodiversity	Environmental DNA (eDNA)	Metabarcoding (Amplicon sequencing)
Analyze effects of a treatment	mRNAs of control and treated individuals	RNAseq Differential gene expression analysis

Library preparation

For Illumina sequencing

Library preparation for Illumina sequencing

Library for sequencing:
Short fragments of DNA with indexes and adapters attached



Adapters: 30-50bp fragments that contain primer sites for amplification and are required to link the fragment with the slide

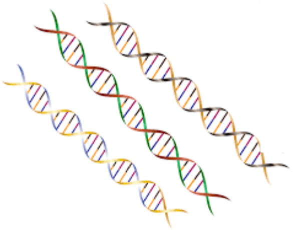
Indexes: 8-10bp fragments with a unique sequence. They are used to distinguish samples run at the same time

Library preparation protocol will depend on the application

Library preparation for Illumina sequencing

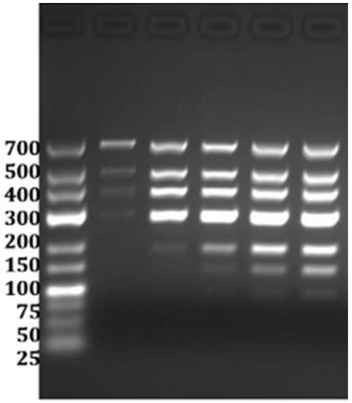
Genome sequencing

DNA extraction

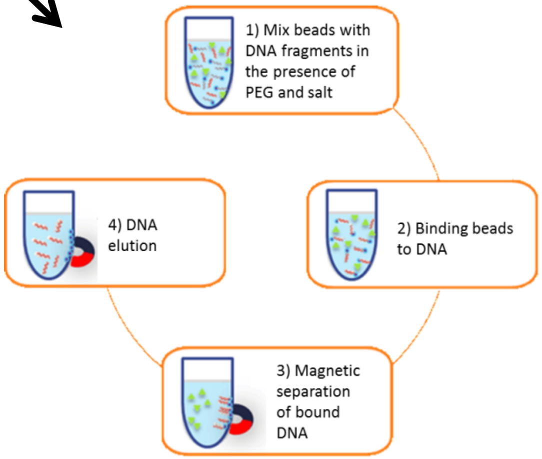


DNA fragmentation:

- Sonication
- Enzymatic reaction



Fragment size selection

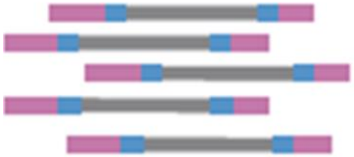


1) Mix beads with DNA fragments in the presence of PEG and salt

2) Binding beads to DNA

3) Magnetic separation of bound DNA

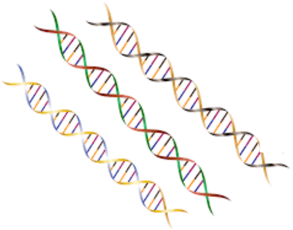
4) DNA elution



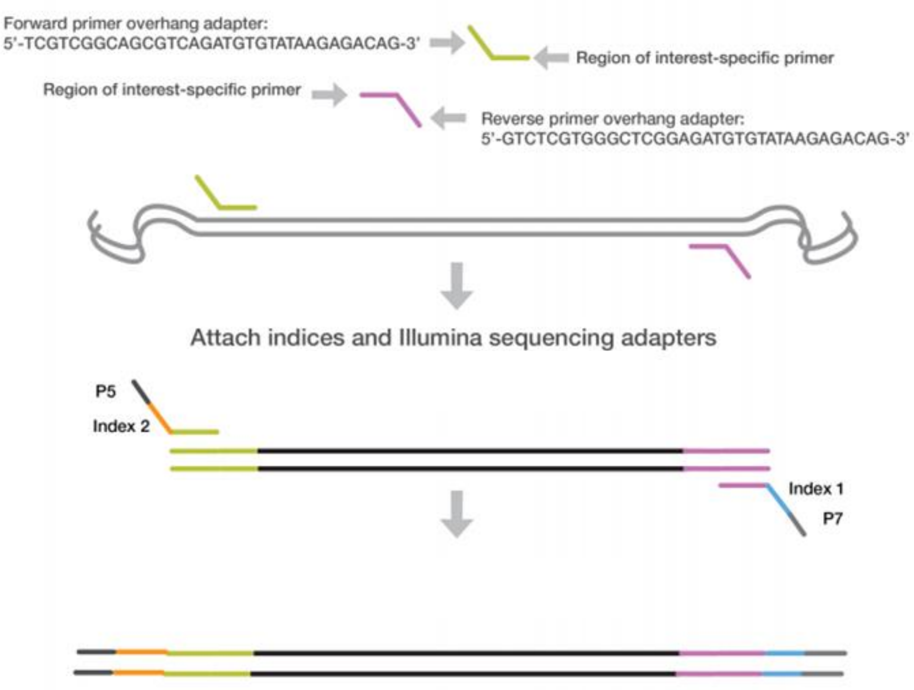
Addition of adapters and indexes

Library preparation for Illumina sequencing

Amplicon sequencing

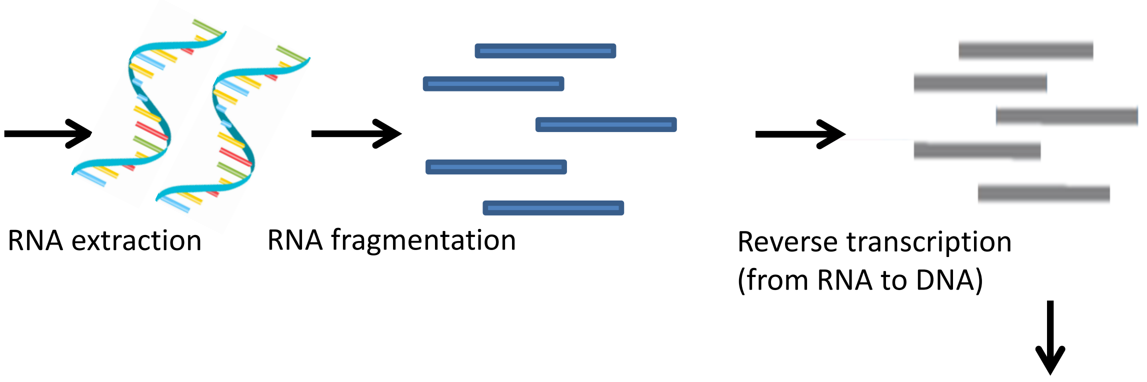


DNA extraction


Forward primer overhang adapter:
5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3'
Region of interest-specific primer
Region of interest-specific primer
Reverse primer overhang adapter:
5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3'
Attach indices and Illumina sequencing adapters
P5
Index 2
Index 1
P7

Library preparation for Illumina sequencing

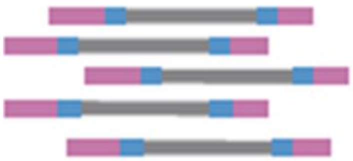
RNA sequencing



RNA extraction

RNA fragmentation

Reverse transcription
(from RNA to DNA)



Addition of adapters and indexes

https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf

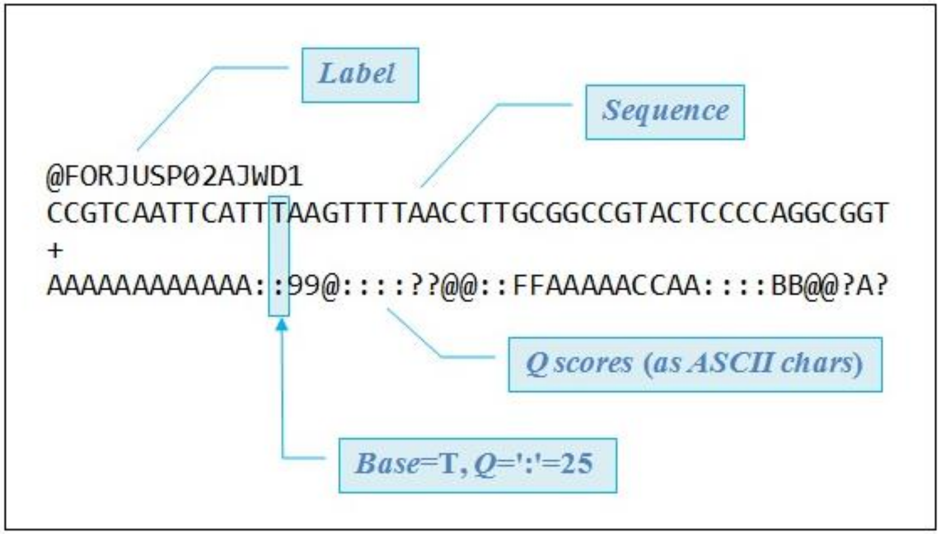
Data Analysis

Data Analysis

- Pre-processing: from raw reads to “clean” reads
- Quality check
- Alignment: Mapping to a reference (or assembly *de novo*)
- Extracting information from the sequences
- Annotation: extracting biological information

Raw reads: FASTQ files

- Text file for storing the sequence and its corresponding quality scores.
- Four lines:
  - The sequence name. It starts with the character ‘@’
  - The sequence itself
  - The character ‘+’
  - Phred quality scores represented as ASCII characters



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

Phred Quality Score

- Indicates the probability that a given base is incorrectly determined (called) by the sequencer
- $Q = -10 \log_{10} P$

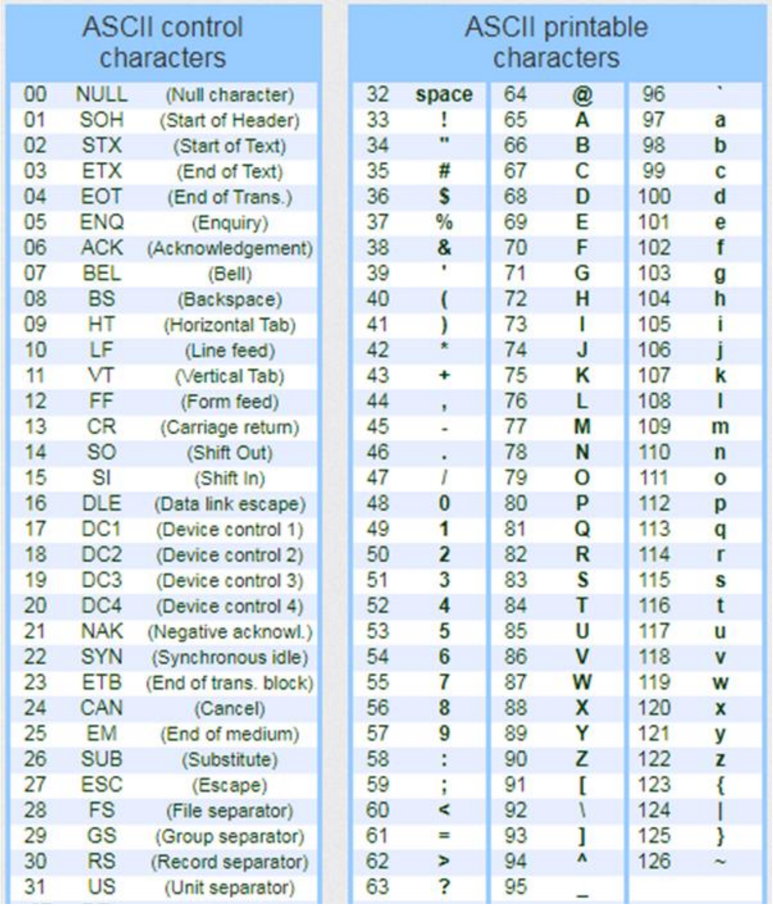
  (where P is the probability of calling the base incorrectly)

Q10 = incorrect base 1/10	(90% accuracy)
Q20 = incorrect base 1/100	(99% accuracy)
Q30 = incorrect base 1/1000	(99.9% accuracy)
Q40 = incorrect base 1/10000	(99.99% accuracy)

- Encoded by ASCII characters

ASCII code

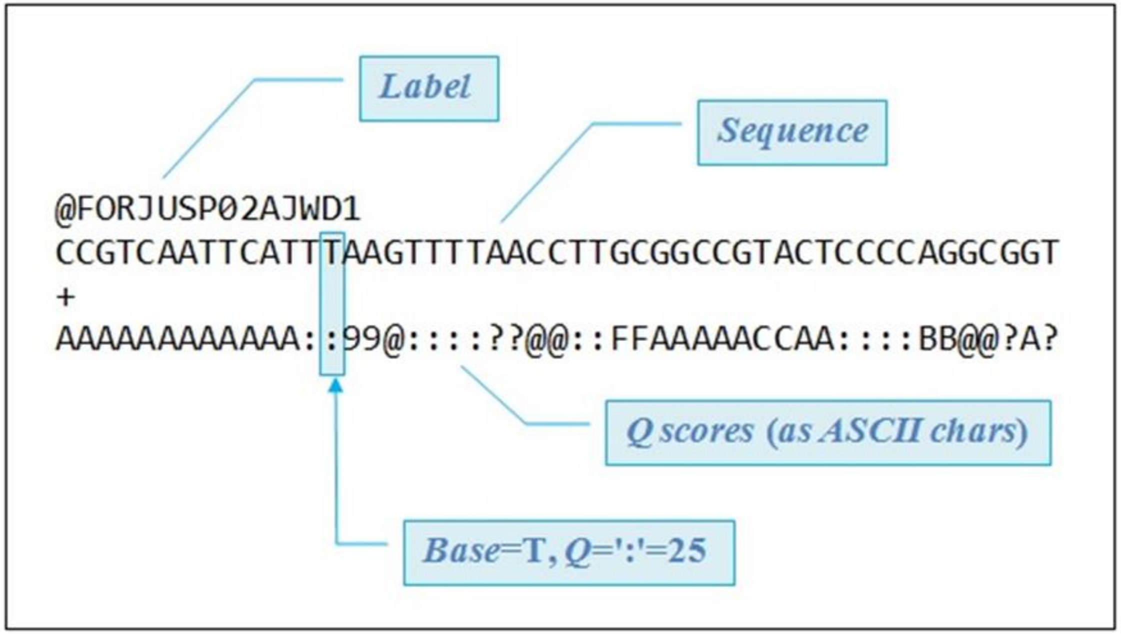
- American Standard Code for Information Interchange
- It’s a code for representing text in computers
- The first 32 characters are unprintable control codes
- We have to subtract 33 to get the Phred score (Q)



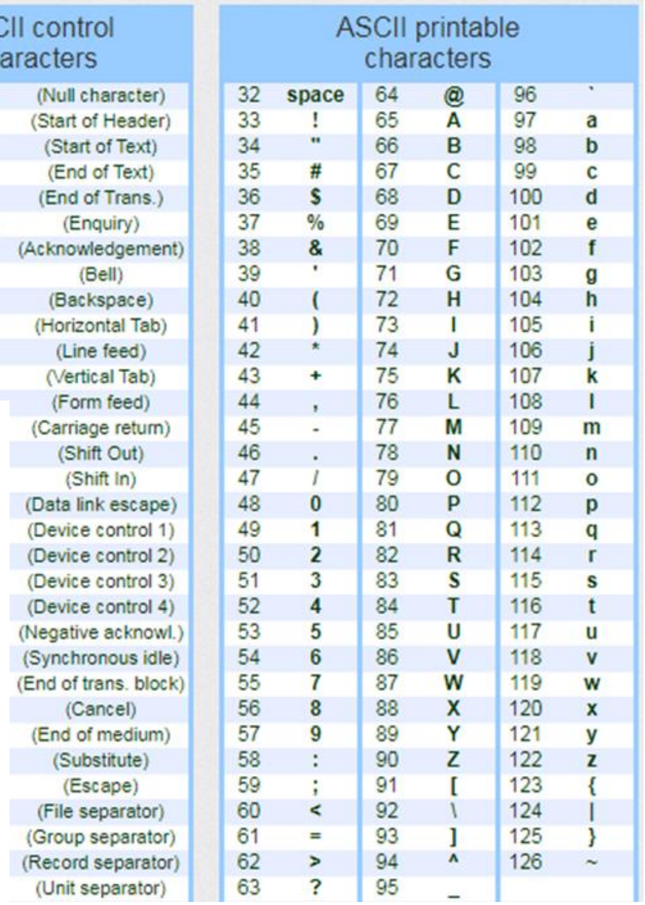
ASCII control characters			ASCII printable characters					
00	NULL	(Null character)	32	space	64	@	96	`
01	SOH	(Start of Header)	33	!	65	A	97	a
02	STX	(Start of Text)	34	"	66	B	98	b
03	ETX	(End of Text)	35	#	67	C	99	c
04	EOT	(End of Trans.)	36	$	68	D	100	d
05	ENQ	(Enquiry)	37	%	69	E	101	e
06	ACK	(Acknowledgement)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

ASCII code

- American Standard Code for Information Interchange
- It’s a code for representing text in computers
- The first 32 characters are unprintable control codes
- We have to subtract 33 to get the Phred score (Q)



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

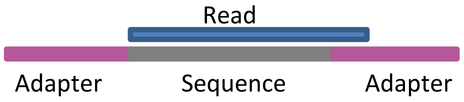


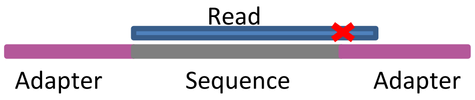
ASCII control characters		
00	NULL	(Null character)
01	SOH	(Start of Header)
02	STX	(Start of Text)
03	ETX	(End of Text)
04	EOT	(End of Trans.)
05	ENQ	(Enquiry)
06	ACK	(Acknowledgement)
07	BEL	(Bell)
08	BS	(Backspace)
09	HT	(Horizontal Tab)
10	LF	(Line feed)
11	VT	(Vertical Tab)
12	FF	(Form feed)
		(Carriage return)
		(Shift Out)
		(Shift In)
		(Data link escape)
		(Device control 1)
		(Device control 2)
		(Device control 3)
		(Device control 4)
		(Negative acknowl.)
		(Synchronous idle)
		(End of trans. block)
		(Cancel)
		(End of medium)
		(Substitute)
		(Escape)
		(File separator)
		(Group separator)
		(Record separator)
		(Unit separator)

ASCII printable characters					
32	space	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{
60	<	92	\	124	|
61	=	93	]	125	}
62	>	94	^	126	~
63	?	95	_		

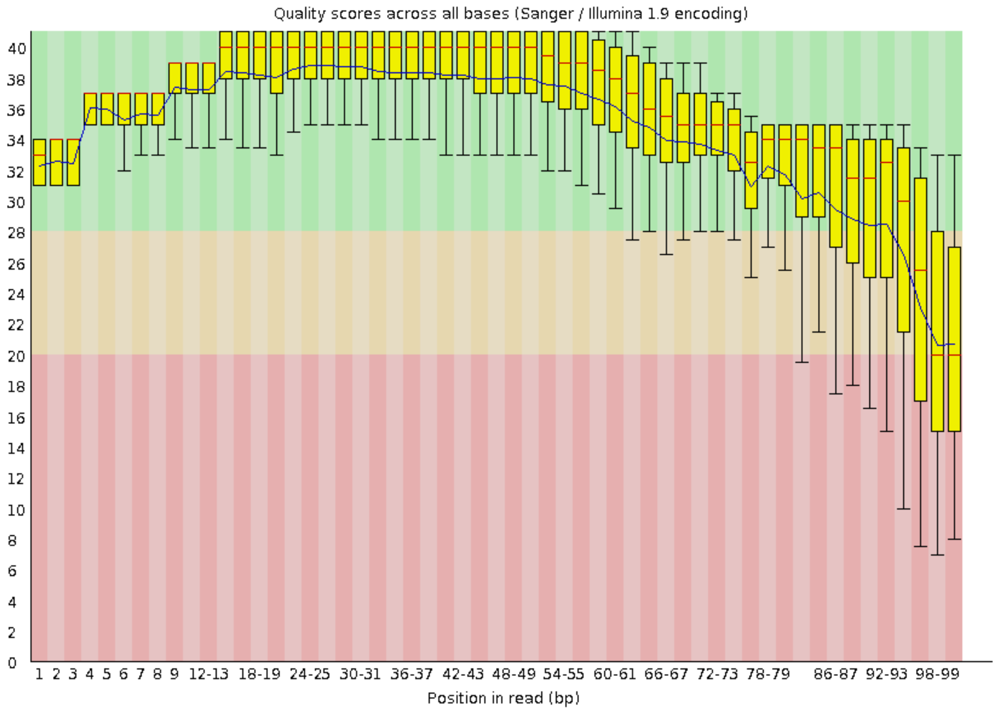
Data pre-processing and quality check

- Remove adapters





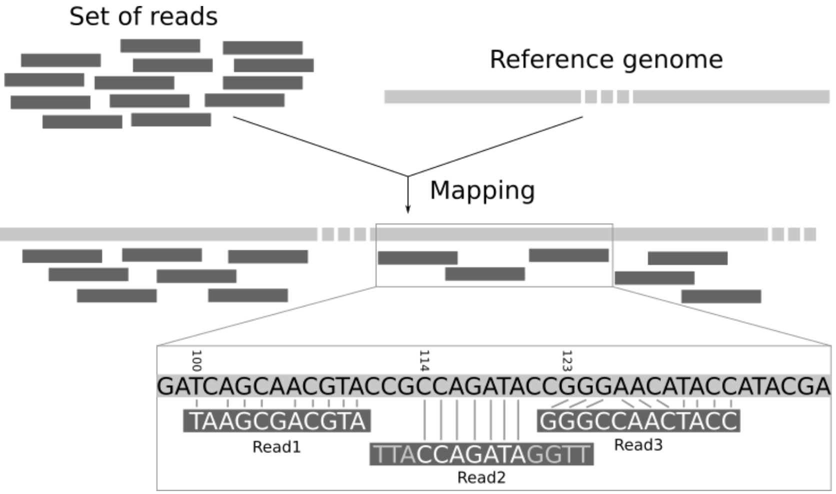
- Remove reads that are too short
- Remove low quality reads
- Check quality
- Trim reads



```
Trimmomatic
FastQC
```

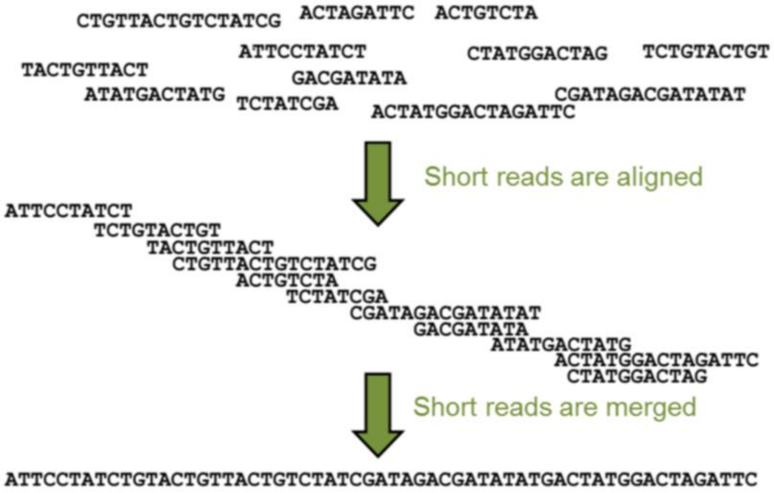
Alignment

Mapping to a reference



BWA
Bowtie2
SOAP

de novo assembly

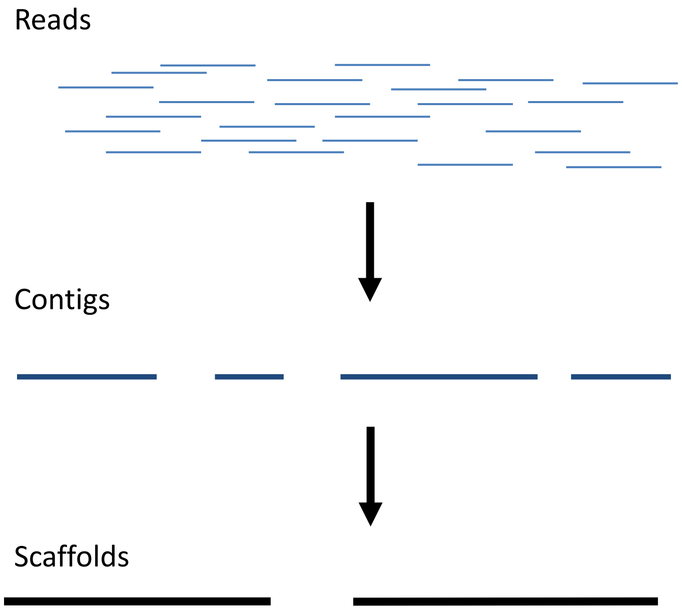


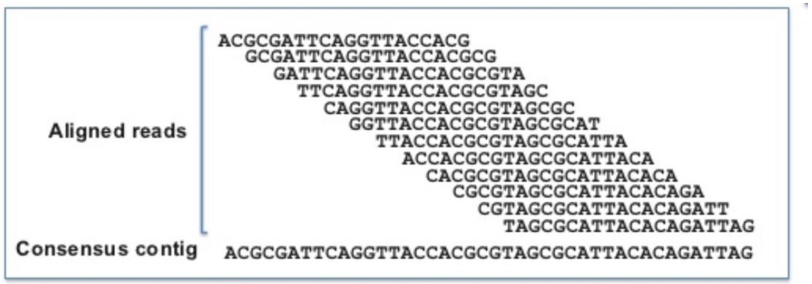
Consensus sequence

Velvet
Trinity
SPAdes
ABySS

We align the reads to reconstruct the original sequences

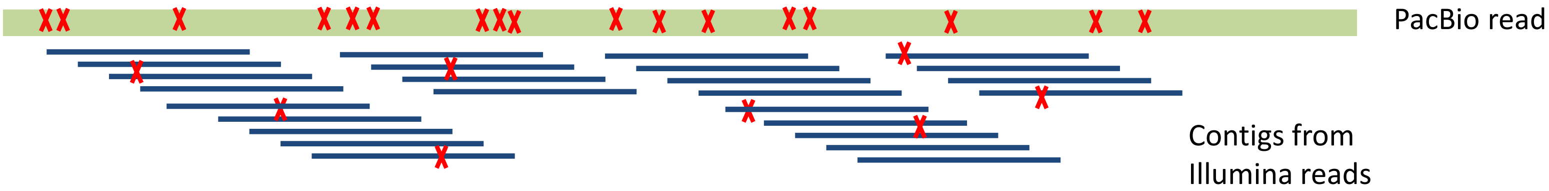
de novo assembly





Problems

- Many short sequences
- Repeats
- Sequencing errors



Some important concepts

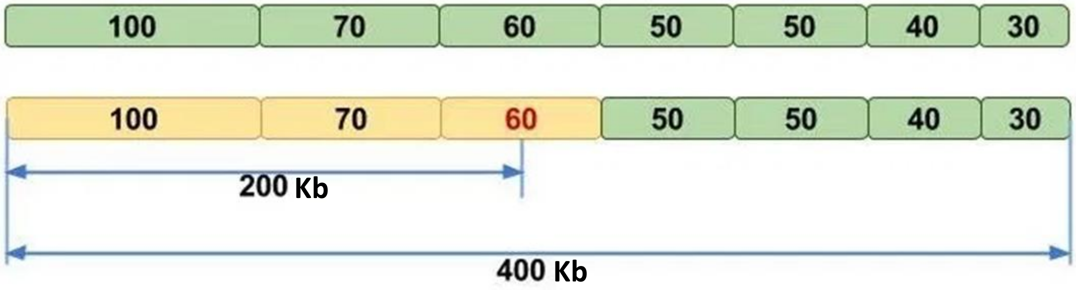
- **Coverage**: average number of reads that include a given nucleotide in the reconstructed sequence

```
Read 1: ATCGTACGAATGCCGTAGTCTGATC
Read 2:    GTACGAATGCCGTAGTCTGATCTACGATC
Read 3:           TGCCGTAGTCTGATCTACGATCATGCGT
Read 4:                 AGTCTGATCTACGATCATGCGTGTA
        11122222223333334444444443333333222222111
```

C = N * L / G

C = average coverage
N =number of reads
L = average read length
G = genome size

- **N50**: measure of the contiguity of an assembly. Given a set of contigs, the N50 is defined as the sequence length of the shortest contig at 50% of the total genome length.



Contigs sorted by length (Kb)

N50 = 60 Kb

Output files

Mapping to a reference

- SAM/BAM files

- Stats files
- Others

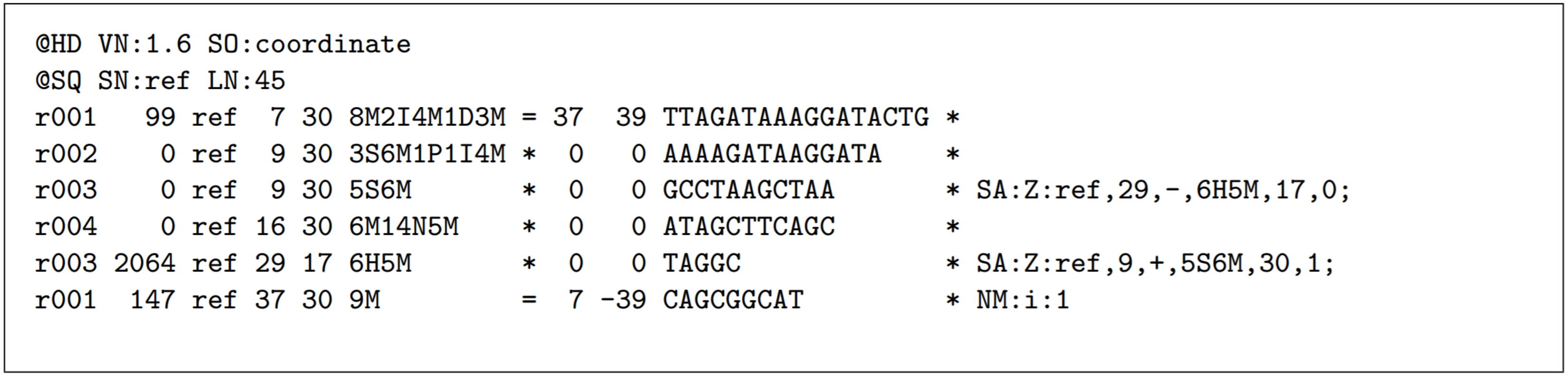
***de novo* assembly**

- Fasta files

- Stats files
- Others

Mapping: SAM/BAM files

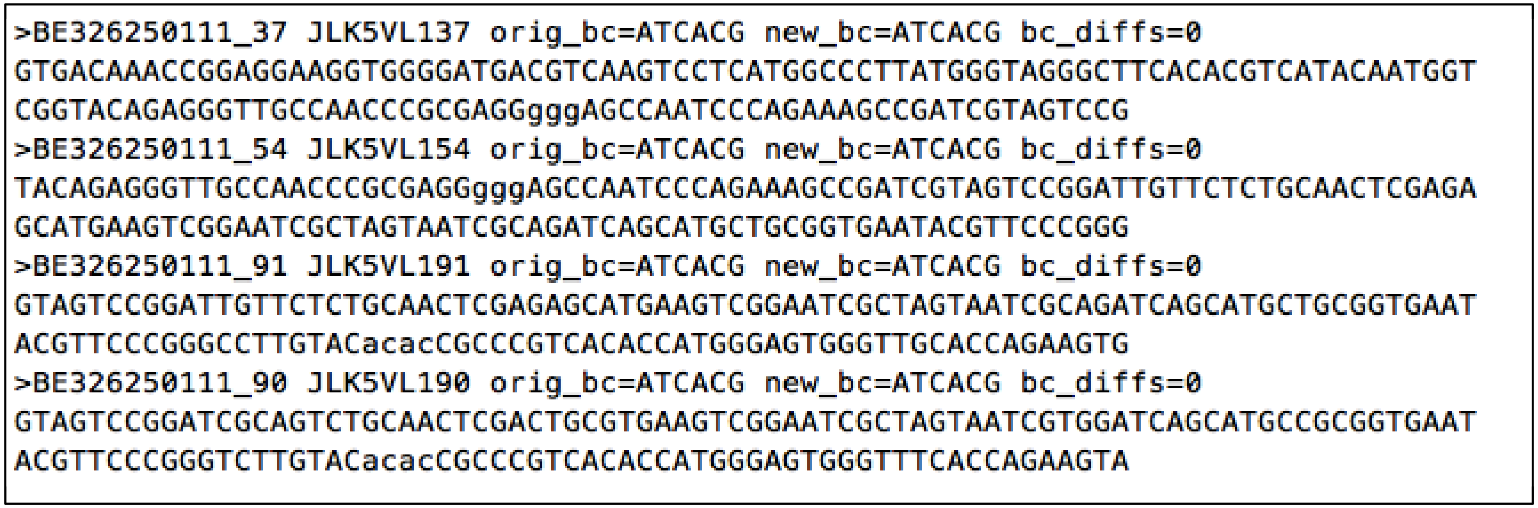
- SAM: Sequence Alignment Map. (BAM: Binary Alignment Map (not human readable))
- Tab delimited text file
- Contains alignment information of short reads mapped against reference sequences.
- Two sections:
  - Header section: contains information about the sample
  - Alignment: contains location and qualities for all the reads. Eleven mandatory columns (QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL) plus optional columns



```
@HD VN:1.6 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Fasta file

- Text file for nucleotide (or peptide) sequences
- Two sections:
  - Header: starts with a ‘>’ character followed by an optional sequence identifier
  - Sequence



```
>BE326250111_37 JLK5VL137 orig_bc=ATCACG new_bc=ATCACG bc_diffs=0
GTGACAAACCGGAGGAAGGTGGGGATGACGTCAAGTCCTCATGGCCCTTATGGGTAGGGCTTCACACGTCATACAATGGT
CGGTACAGAGGGTTGCCAACCCGCGAGGgggAGCCAATCCCAGAAAGCCGATCGTAGTCCG
>BE326250111_54 JLK5VL154 orig_bc=ATCACG new_bc=ATCACG bc_diffs=0
TACAGAGGGTTGCCAACCCGCGAGGgggAGCCAATCCCAGAAAGCCGATCGTAGTCCGGATTGTTCTCTGCAACTCGAGA
GCATGAAGTCGGAATCGCTAGTAATCGCAGATCAGCATGCTGCGGTGAATACGTTCCCGGG
>BE326250111_91 JLK5VL191 orig_bc=ATCACG new_bc=ATCACG bc_diffs=0
GTAGTCCGGATTGTTCTCTGCAACTCGAGAGCATGAAGTCGGAATCGCTAGTAATCGCAGATCAGCATGCTGCGGTGAAT
ACGTTCCCGGGCCTTGTACacacCGCCCGTCACACCATGGGAGTGGGTTGCACCAGAAGTG
>BE326250111_90 JLK5VL190 orig_bc=ATCACG new_bc=ATCACG bc_diffs=0
GTAGTCCGGATCGCAGTCTGCAACTCGACTGCGTGAAGTCGGAATCGCTAGTAATCGTGGATCAGCATGCCGCGGTGAAT
ACGTTCCCGGGTCTTGTACacacCGCCCGTCACACCATGGGAGTGGGTTTCACCAGAAGTA
```

Extracting information from sequences

- Example: Variant calling:
  - Compare sequences with a reference and find differences: SNPs (Single Nucleotide Polymorphisms), Indels (Insertions/deletions), SV (Structural Variants).
  - Before that we have to "clean" the alignment data: BAM refinement:
    - Local realignment: improves the alignment, specially around indels
    - Base quality recalibration: re-evaluates the probability of a wrong call at each position in each read
    - Remove PCR duplicates
    - OUTPUT: SAM/BAM files

```
GATK
Picard
```

Extracting information from sequences

- Example: Variant calling:
  - Compare sequences with a reference and find differences:

<u>SNPs and indels</u>

```
Reference          ATGTAGTCCGTAG
SNP variant        ATGTAGTACGTAG

Reference          ATGTAGTCCGTAG
Insertion variant  ATGTAGTGCCGTAG

Reference          ATGTAGTCCGTAG
Deletion variant   ATGTAGT-CGTAG
```

GATK
SAMtools

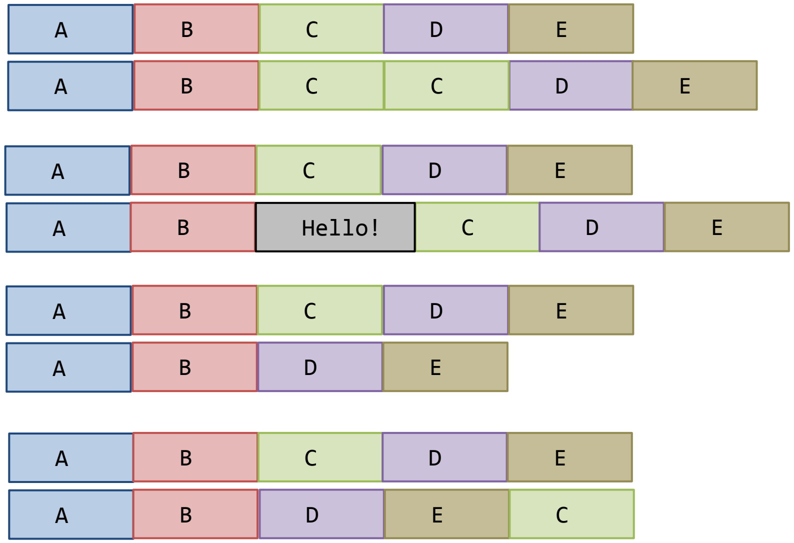
<u>Structural variants: > 50pb</u>

Reference
Duplication variant

Reference
Insertion variant

Reference
Deletion variant

Reference
Translocation variant

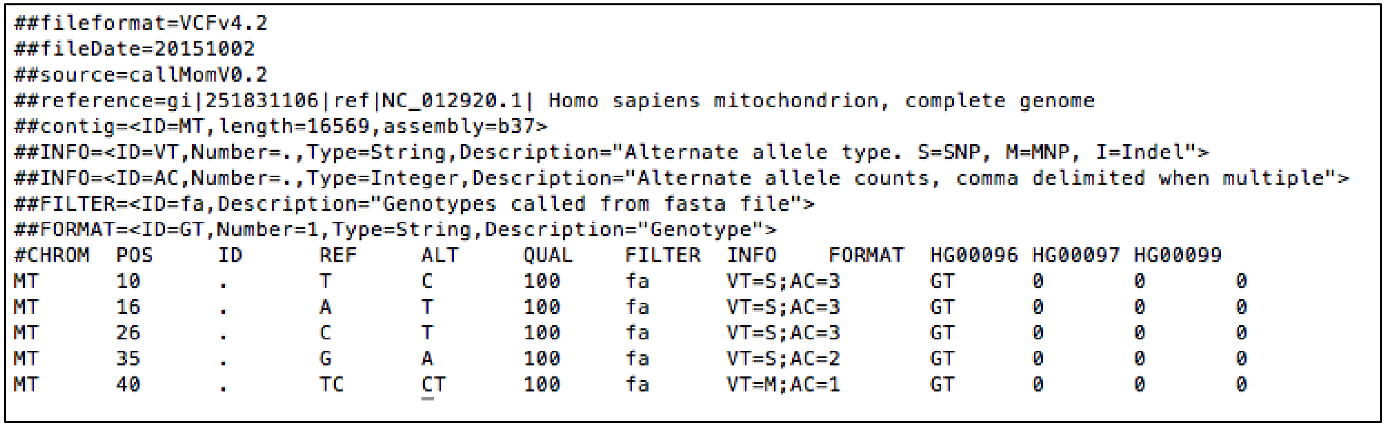


Pindel
GRIDSS

  - Output: VCF (Variant Call Format) files

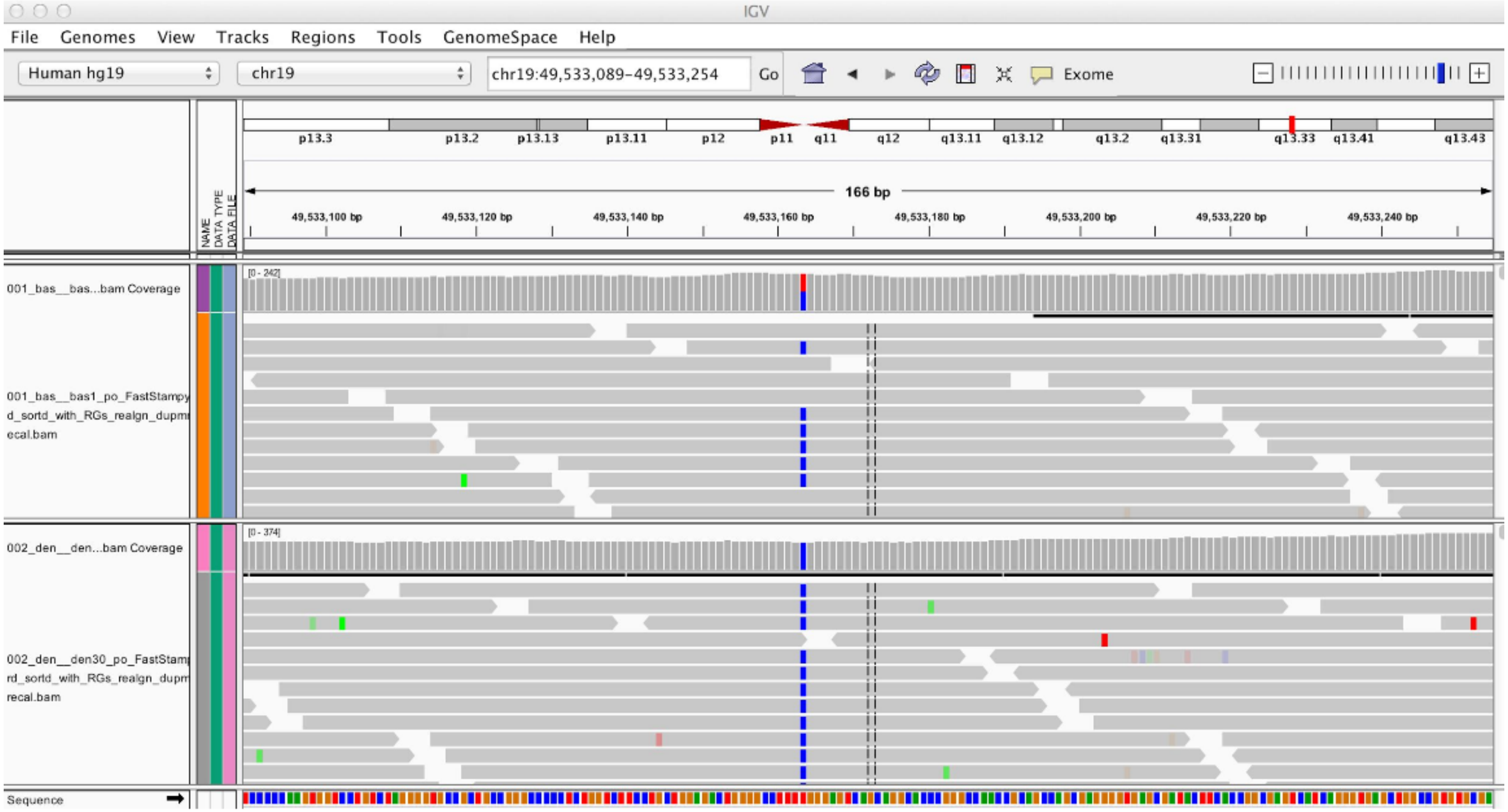
VCF file

- VCF: Variant Call Format
- Tab delimited text file
- Stores information about gene sequence variations
- Three sections:
  - Metadata: lines commencing with '##'. Describe the body of the file
  - Header line: starts with '#'. Names the 8 fixed, mandatory columns (CHROM POS ID REF ALT QUAL FILTER INFO) plus optional columns: FORMAT and sample columns
  - Data lines: contain information (corresponding to header columns) about a position in the genome



```
##fileformat=VCFv4.2
##fileDate=20151002
##source=callMomV0.2
##reference=gi|251831106|ref|NC_012920.1| Homo sapiens mitochondrion, complete genome
##contig=<ID=MT,length=16569,assembly=b37>
##INFO=<ID=VT,Number=.,Type=String,Description="Alternate allele type. S=SNP, M=MNP, I=Indel">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Alternate allele counts, comma delimited when multiple">
##FILTER=<ID=fa,Description="Genotypes called from fasta file">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  HG00096 HG00097 HG00099
MT      10      .       T       C       100     fa      VT=S;AC=3       GT      0       0       0
MT      16      .       A       T       100     fa      VT=S;AC=3       GT      0       0       0
MT      26      .       C       T       100     fa      VT=S;AC=3       GT      0       0       0
MT      35      .       G       A       100     fa      VT=S;AC=2       GT      0       0       0
MT      40      .       TC      CT      100     fa      VT=M;AC=1       GT      0       0       0
```

Data visualization


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr19
chr19:49,533,089-49,533,254
Go
Exome
p13.3 p13.2 p13.13 p13.11 p12 p11 q11 q12 q13.11 q13.12 q13.2 q13.31 q13.33 q13.41 q13.43
166 bp
49,533,100 bp
49,533,120 bp
49,533,140 bp
49,533,160 bp
49,533,180 bp
49,533,200 bp
49,533,220 bp
49,533,240 bp
NAME
DATA TYPE
DATA FILE
001_bas__bas...bam Coverage
[0 - 242]
001_bas__bas1_po_FastStampy
d_sortd_with_RGs_realgn_dupm
ecal.bam
002_den__den...bam Coverage
[0 - 374]
002_den__den30_po_FastStam
rd_sortd_with_RGs_realgn_dupm
recal.bam
Sequence

IGV (Integrative Genomics Viewer)

Annotation

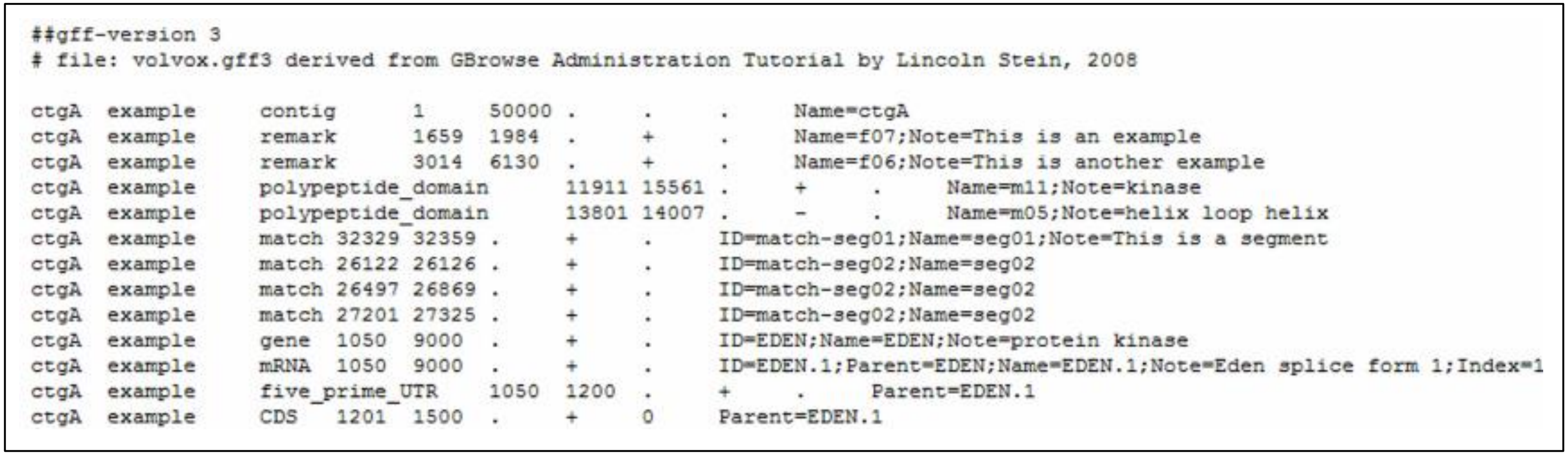
- Extract the relevant biological information from the sequences data
- What's the relevant information?: it depends on the experiment
  - Whole genome sequencing (WGS):
    - Structural annotation: identification of genomic elements (where are the genes located, the non coding regions, etc.)
    - Functional annotation: what do these genes do?
  - Genome re-sequencing:
    - Are the variants associated with disease?
    - Are some variants more frequent in a given population?
  - RNAseq:
    - What do the differentially expressed genes do?
  - Etc
- Output: GFF files

Comparing with DATABASES

```
(lots of different bioinformatics resources, depending on the sample (animal, plant,
bacteria…), the data (WGS, genome re-sequencing, RNAseq…), and the question)
```

GFF file

- GFF: General Feature Format
- Tab delimited text file
- Used for describing genes and other features of DNA, RNA and protein sequences
- One line per feature, 9 columns of data (seqname, source, feature, start, end, score, strand, frame, attribute)



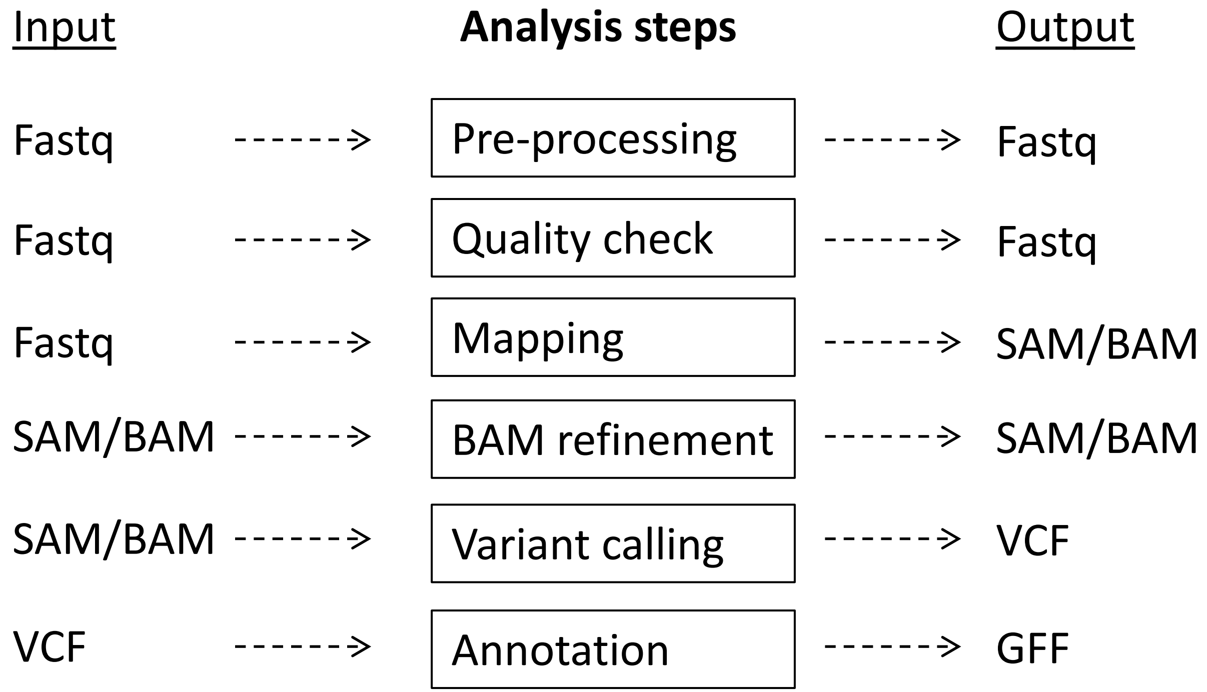
```
##gff-version 3
# file: volvox.gff3 derived from GBrowse Administration Tutorial by Lincoln Stein, 2008

ctgA  example     contig      1     50000 .     .     .     Name=ctgA
ctgA  example     remark      1659  1984  .     +     .     Name=f07;Note=This is an example
ctgA  example     remark      3014  6130  .     +     .     Name=f06;Note=This is another example
ctgA  example     polypeptide_domain      11911 15561 .     +     .     Name=m11;Note=kinase
ctgA  example     polypeptide_domain      13801 14007 .     -     .     Name=m05;Note=helix loop helix
ctgA  example     match 32329 32359 .     +     .     ID=match-seg01;Name=seg01;Note=This is a segment
ctgA  example     match 26122 26126 .     +     .     ID=match-seg02;Name=seg02
ctgA  example     match 26497 26869 .     +     .     ID=match-seg02;Name=seg02
ctgA  example     match 27201 27325 .     +     .     ID=match-seg02;Name=seg02
ctgA  example     gene  1050  9000  .     +     .     ID=EDEN;Name=EDEN;Note=protein kinase
ctgA  example     mRNA  1050  9000  .     +     .     ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Note=Eden splice form 1;Index=1
ctgA  example     five_prime_UTR    1050  1200  .     +     .     Parent=EDEN.1
ctgA  example     CDS   1201  1500  .     +     0     Parent=EDEN.1
```

Summary

- Experimental design
- Sample extraction
- Library preparation
- Sequencing
- Data analysis

Example analysis summary (variant calling)



Input	Analysis steps	Output
Fastq	Pre-processing	Fastq
Fastq	Quality check	Fastq
Fastq	Mapping	SAM/BAM
SAM/BAM	BAM refinement	SAM/BAM
SAM/BAM	Variant calling	VCF
VCF	Annotation	GFF

Visualization

Inputs

- SAM
- VCF
- GFF

Questions?