Université de Lille

Egid European Genomic Institute for Diabetes

Communauté d'Universités et Établissements Lille Nord de France

Analysis of NGS-OMICS data in humans

Derhourhi Mehdi
mehdi.derhourhi@cnrs.fr

UMR 1283 8199
Fonctional (Epi)genomics and Mechanisms of Type 2 Diabetes and Related Diseases

LIGAN GENOMIC PLATFORM

Summary

1 - History of sequencing

2 - How does it works ?

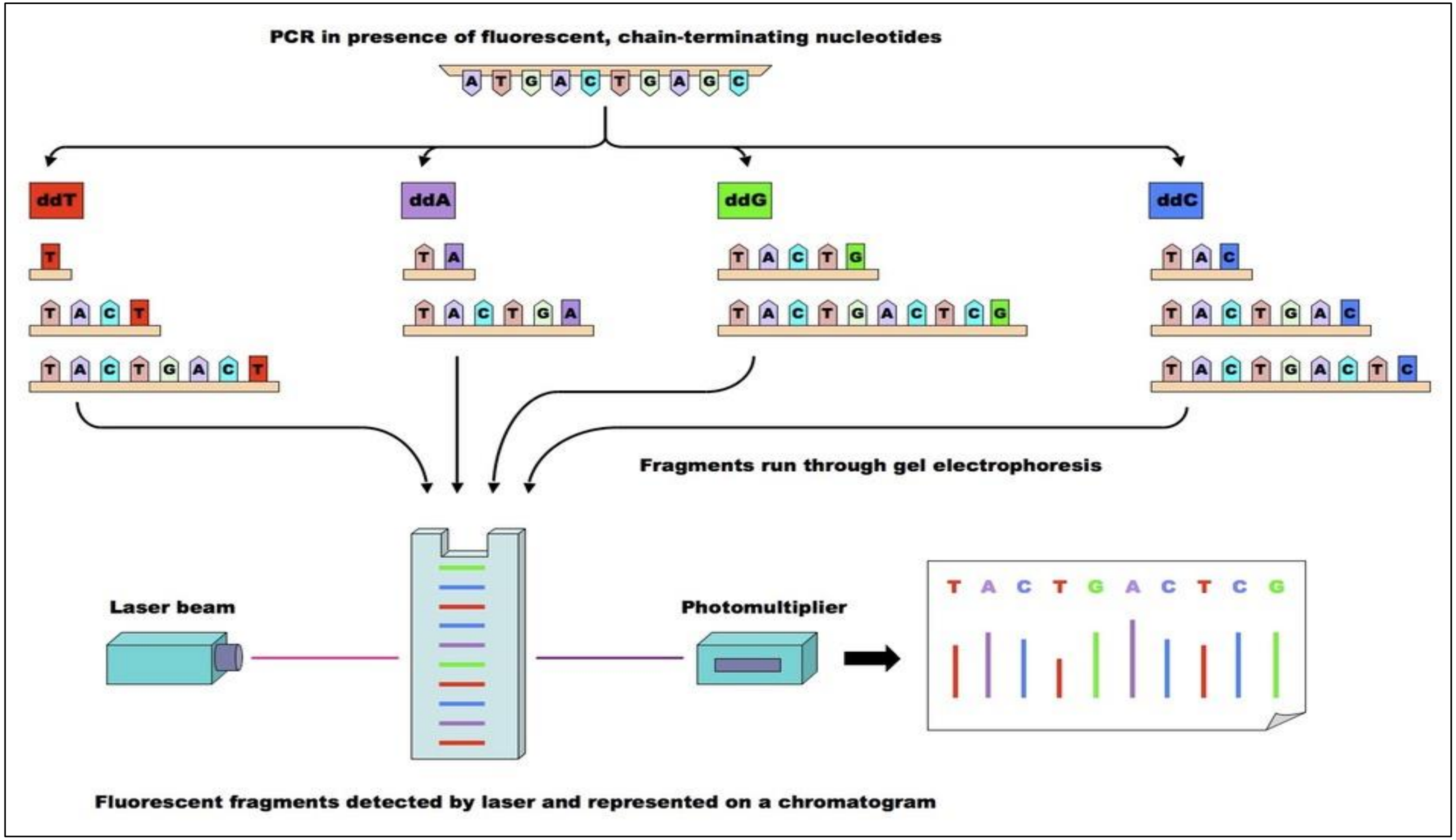
3 - Focus on 4 sequencing methods and analysis

Sanger Sequencing

→ Near 1970

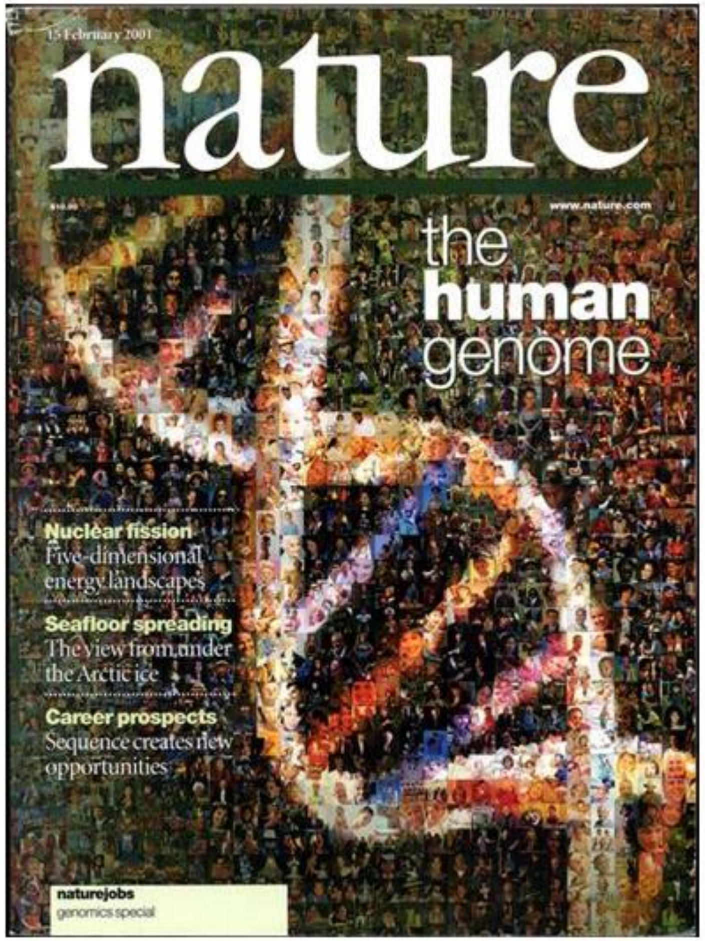
→ Use of fluorescent didesoxyribonuleotides (ddNTP)

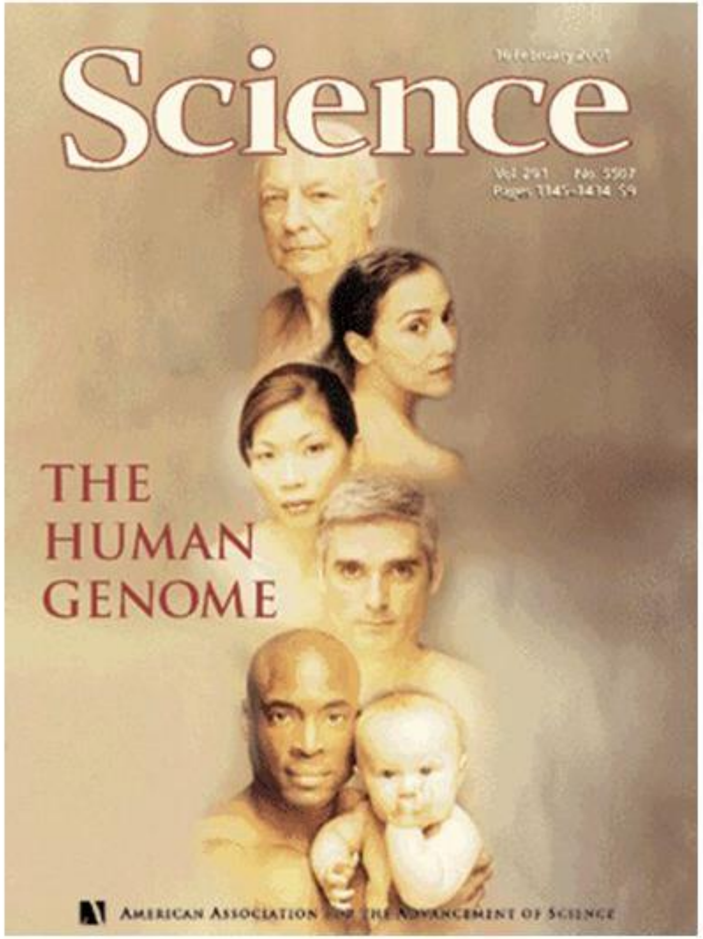
→ Low Throughput


PCR in presence of fluorescent, chain-terminating nucleotides
A T G A C T G A G C
ddT
ddA
ddG
ddC
T
T A C T
T A C T G A C T
T A
T A C T G A
T A C T G
T A C T G A C T C G
T A C
T A C T G A C
T A C T G A C T C
Fragments run through gel electrophoresis
Laser beam
Photomultiplier
T A C T G A C T C G
Fluorescent fragments detected by laser and represented on a chromatogram

Human Genome Project

- → Project aiming to sequence the whole human genome, representing 3 billions bases
- → Massive use of Sanger technology
- → Started in 1990
- → 3 billions dollars cost
- → Finished in 2004





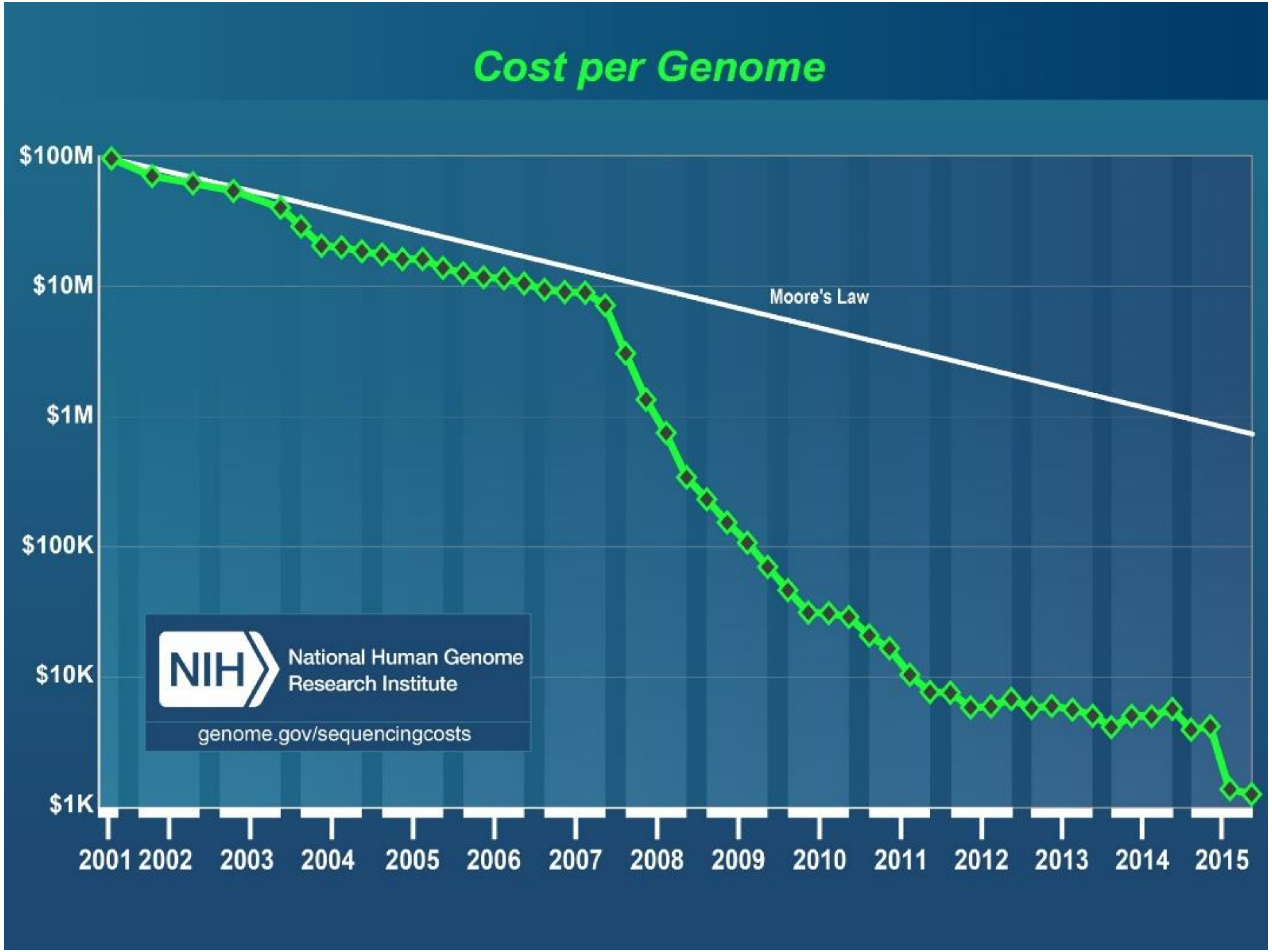
Next Generation Sequencing

New sequencing methods aiming to replace Sanger sequencing started developping in the 80’s, and arrived on the market after 2000.

The throughput widely increased compared to Sanger sequencing

→ « High throughput sequencing »

It’s now possible to sequence a full human genome for nearly 600 dollars in few days.

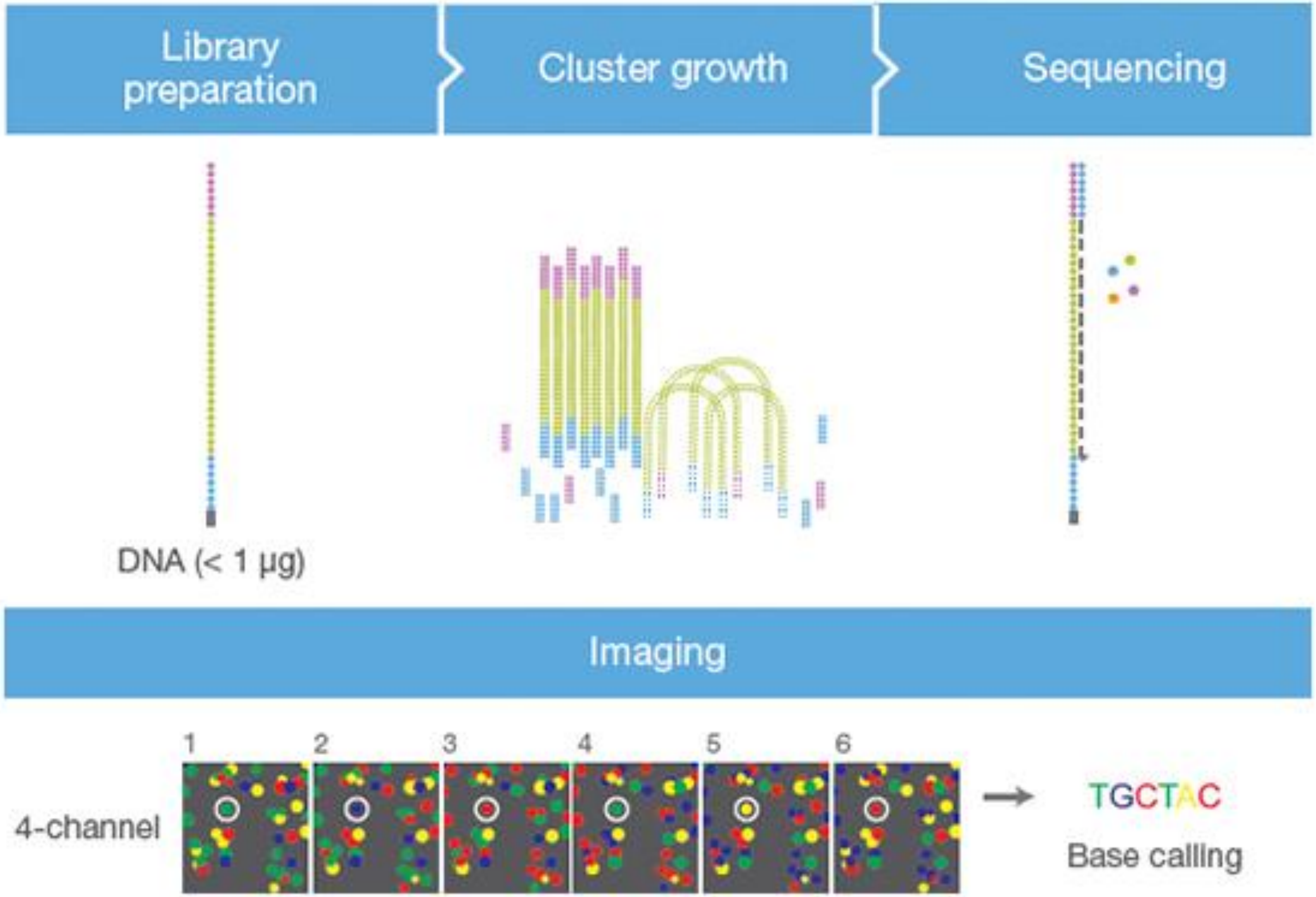


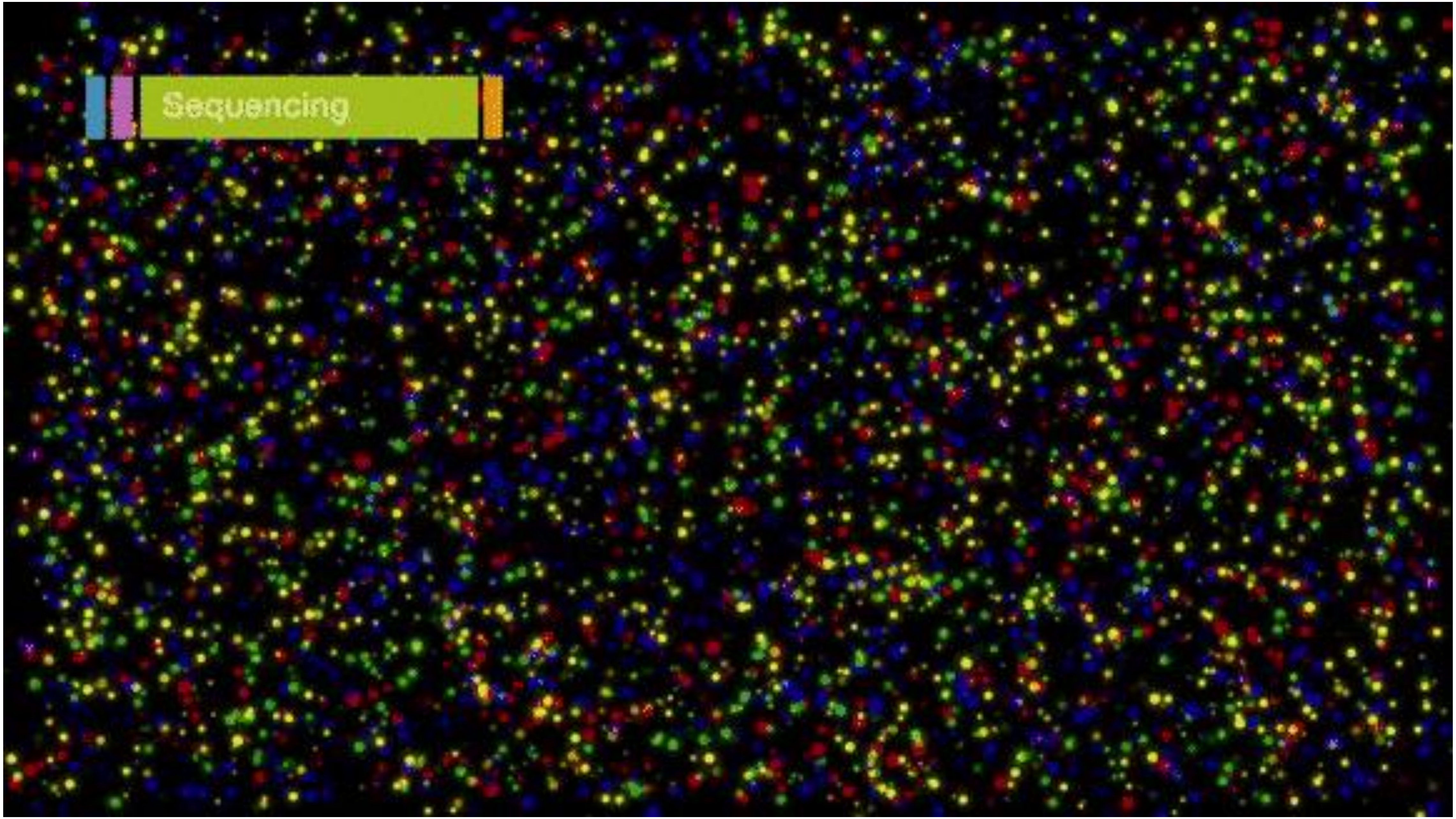
Illumina SBS

The Illumina company is now leader in the field of high throughput sequencing (or NGS for Next Generation Sequencing) with their Sequencing By Synthesis technology (SBS)

→ This technology is based on the use of a polymerase incorporating fluorescent nucleotides

→ Each time a nucleotide is incorporated, a laser excites fluorophores which emit light, then a picture is taken





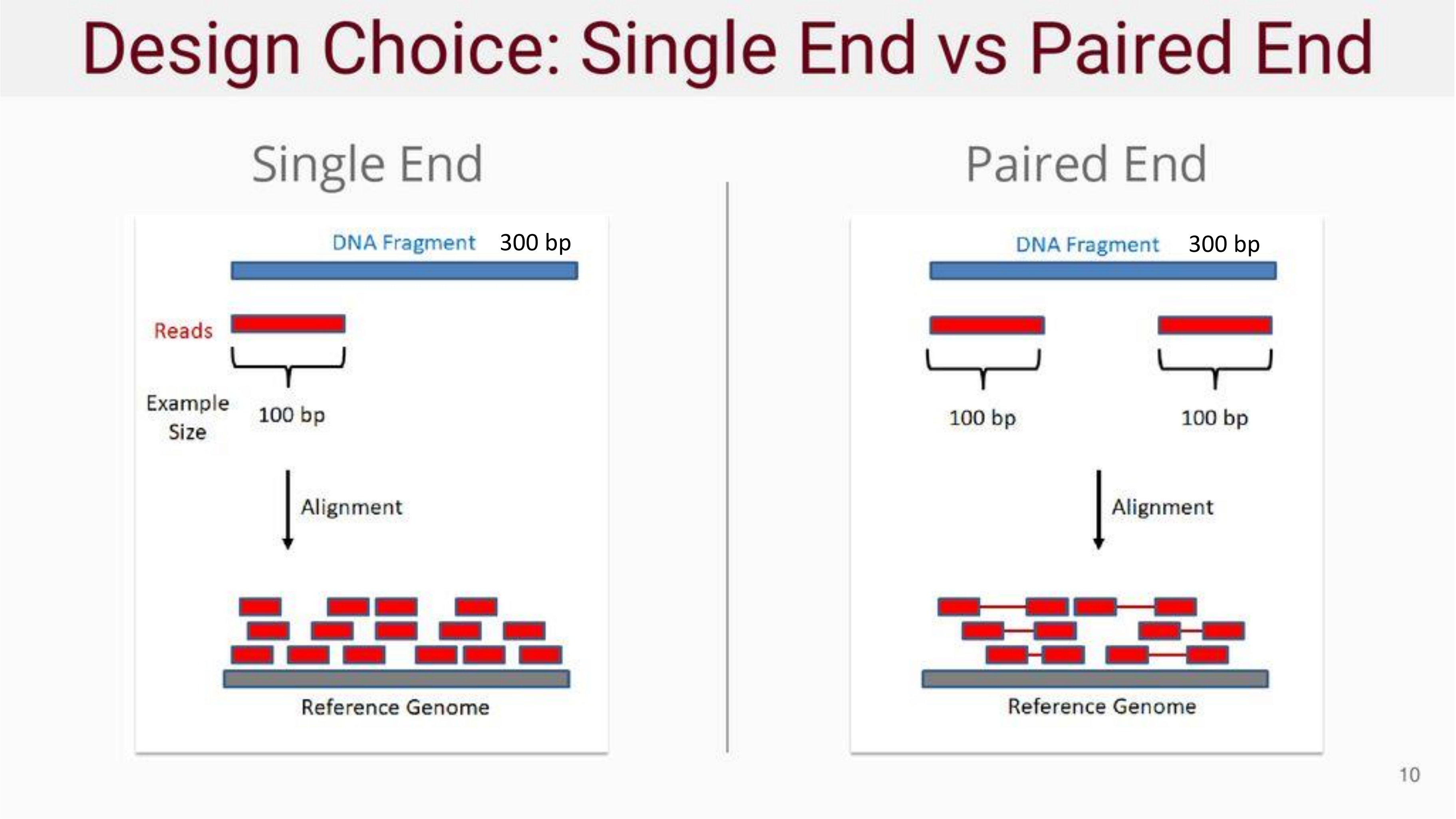
https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

Single end / Paired end

Two different sequencing types with Illumina :

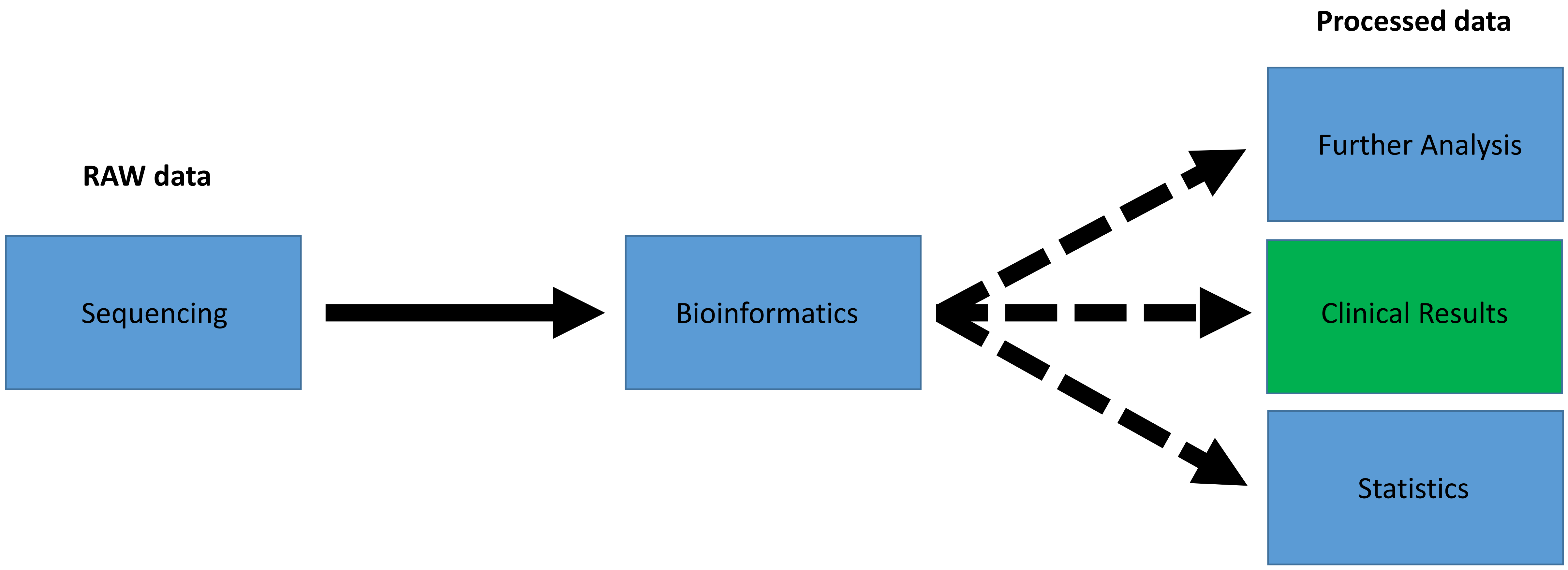
-Single End :
Only one end of each DNA fragment is sequenced
→ Cheaper but less informative

-Paired End
Both ends of each DNA fragment is sequenced
→ More expensive but more informative

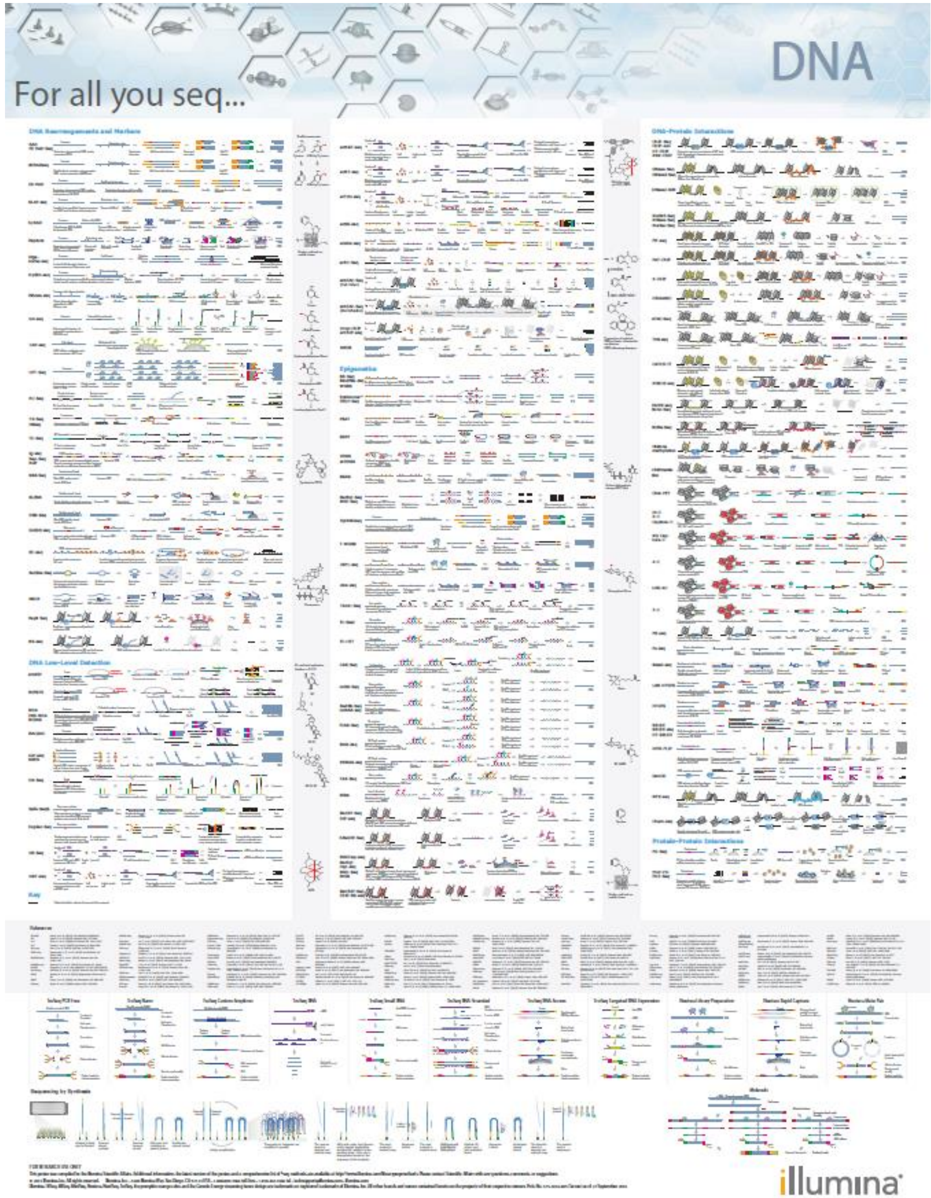


Design Choice: Single End vs Paired End

Bioinformatics



Sequencing methods



Hundreds of different methods

The choice depends of the question asked

The analysis depends of the method and the question

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

Focus on 4 sequencing methods

DNA-seq

Determine the genome sequence

→ In clinical context, find variants which could explain diseases

RNA-seq

Determine the genes expressed and their expression level

→ In clinical context, find unexpressed or over expressed genes which could explain diseases

ChIP-seq

Determine the interactions between DNA and proteins

Hi-C

Determine the interactions between DNA and DNA

Focus on 4 sequencing methods

DNA-seq

Determine the genome sequence

→ In clinical context, find variants which could explain diseases

RNA-seq

Determine the genes expressed and their expression levels

→ In clinical context, find unexpressed or over expressed genes which could explain diseases

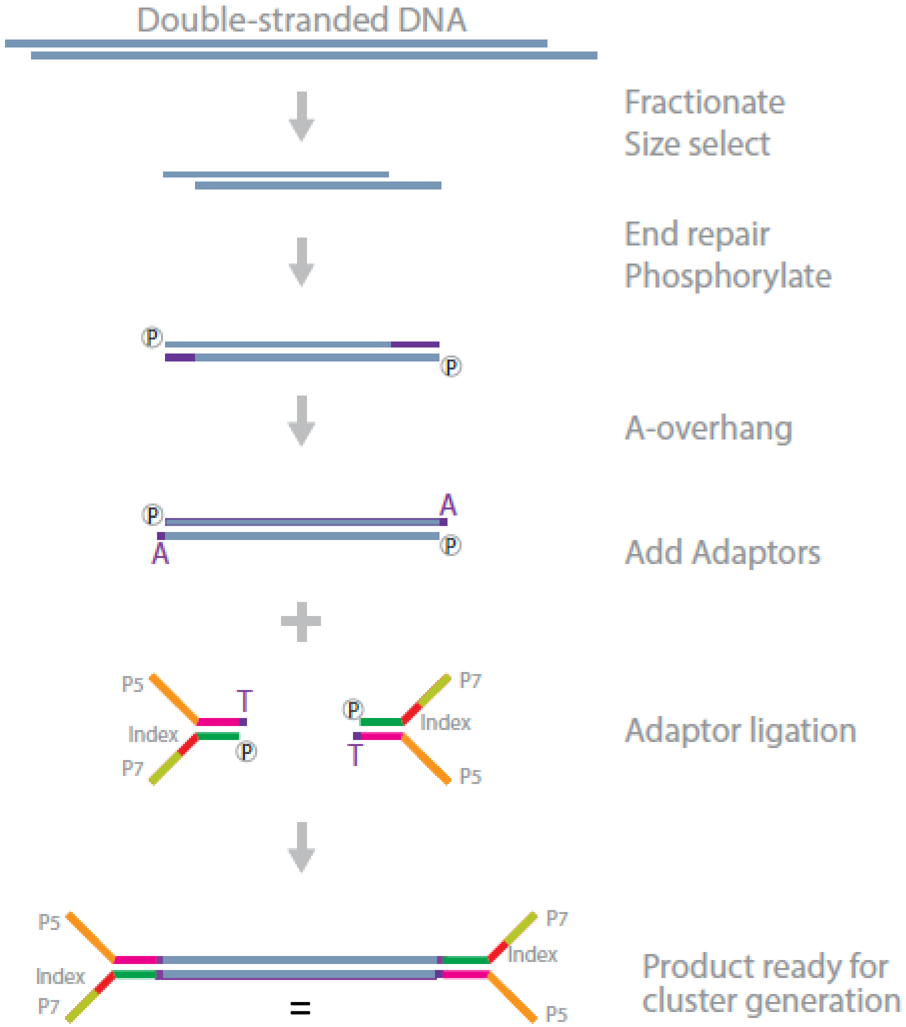
ChIP-seq

Determine the interactions between DNA and proteins

Hi-C

Determine the the interactions between DNA and DNA

DNA sequencing



Fractionate
Size select

End repair
Phosphorylate

A-overhang

Add Adaptors

Adaptor ligation

Product ready for
cluster generation

=

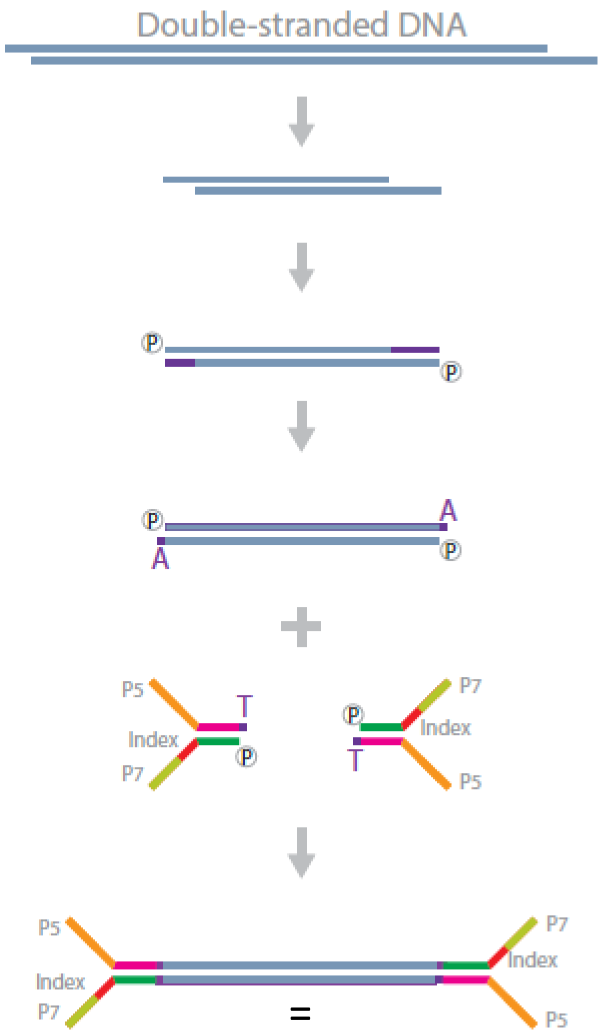
Library

=

All the DNA of a sample

Whole Genome Sequencing

DNA sequencing



Fractionate
Size select

End repair
Phosphorylate

A-overhang

Add Adaptors

Adaptor ligation

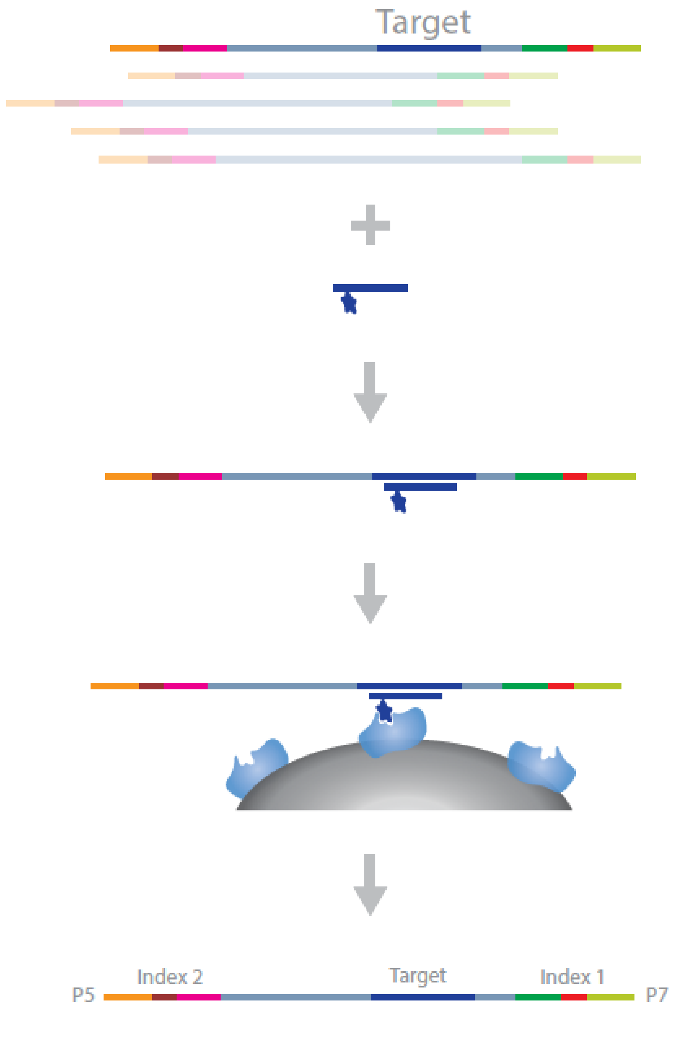
Product ready for
cluster generation

=

Library

=

All the DNA of a sample



Denatured and
pooled fragments
from Nextera library

Biotinylated
target probe
for exome

Hybridize probes
to targets

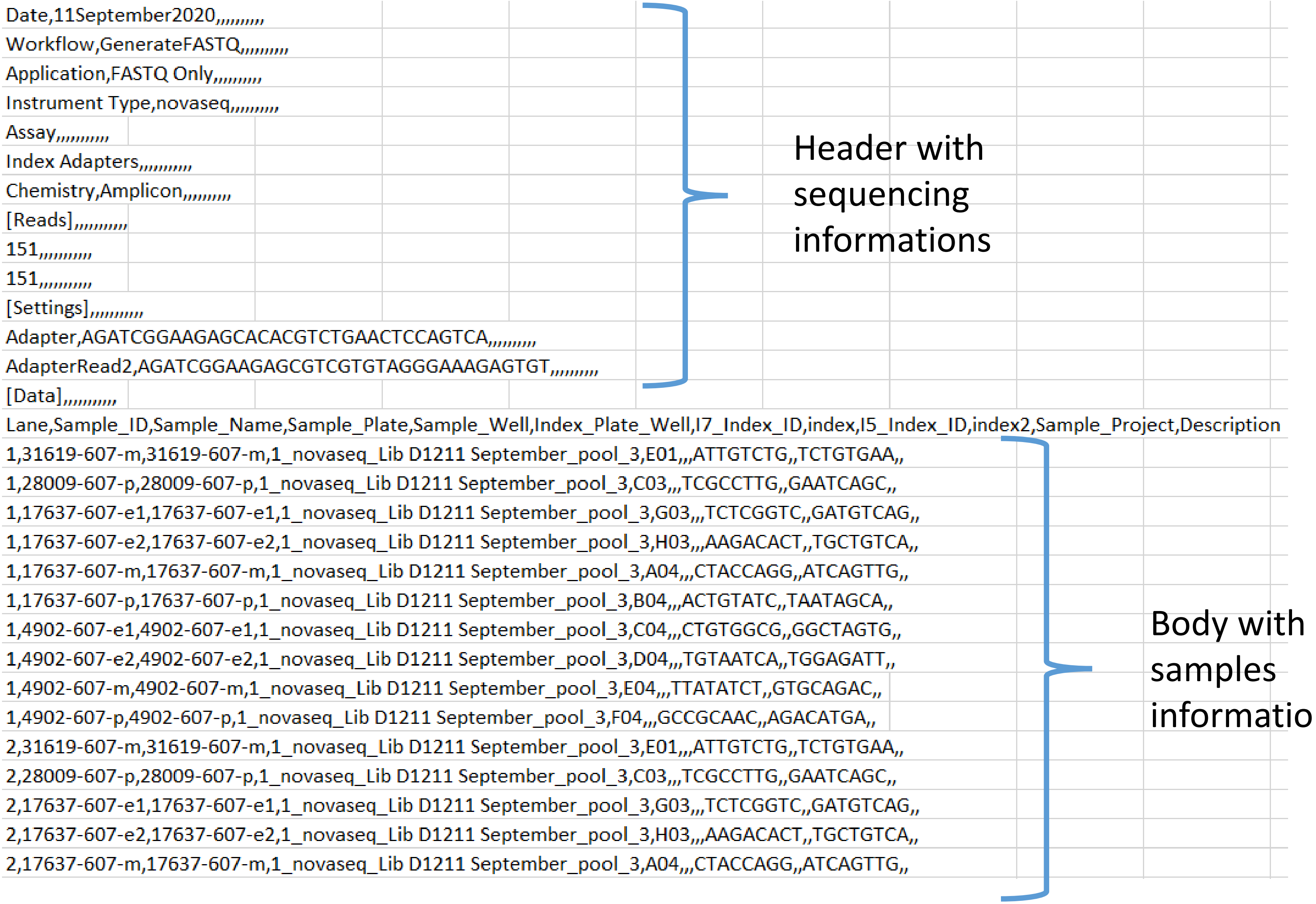
Capture on
magnetic beads

Elute

Product ready for
cluster generation

Whole Exome Sequencing (1-2% of genome)

DNA sequencing : SampleSheet



```
Date,11September2020,,,,,,,,,,
Workflow,GenerateFASTQ,,,,,,,,,,
Application,FASTQ Only,,,,,,,,,,
Instrument Type,novaseq,,,,,,,,,,
Assay,,,,,,,,,,,
Index Adapters,,,,,,,,,,,
Chemistry,Amplicon,,,,,,,,,,
[Reads],,,,,,,,,,,
151,,,,,,,,,,,
151,,,,,,,,,,,
[Settings],,,,,,,,,,,
Adapter,AGATCGGAAGAGCACACGTCTGAACTCCAGTCA,,,,,,,,,,
AdapterRead2,AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT,,,,,,,,,,
[Data],,,,,,,,,,,
Lane,Sample_ID,Sample_Name,Sample_Plate,Sample_Well,Index_Plate_Well,I7_Index_ID,index,I5_Index_ID,index2,Sample_Project,Description
1,31619-607-m,31619-607-m,1_novaseq_Lib D1211 September_pool_3,E01,,,ATTGTCTG,,TCTGTGAA,,
1,28009-607-p,28009-607-p,1_novaseq_Lib D1211 September_pool_3,C03,,,TCGCCTTG,,GAATCAGC,,
1,17637-607-e1,17637-607-e1,1_novaseq_Lib D1211 September_pool_3,G03,,,TCTCGGTC,,GATGTCAG,,
1,17637-607-e2,17637-607-e2,1_novaseq_Lib D1211 September_pool_3,H03,,,AAGACACT,,TGCTGTCA,,
1,17637-607-m,17637-607-m,1_novaseq_Lib D1211 September_pool_3,A04,,,CTACCAGG,,ATCAGTTG,,
1,17637-607-p,17637-607-p,1_novaseq_Lib D1211 September_pool_3,B04,,,ACTGTATC,,TAATAGCA,,
1,4902-607-e1,4902-607-e1,1_novaseq_Lib D1211 September_pool_3,C04,,,CTGTGGCG,,GGCTAGTG,,
1,4902-607-e2,4902-607-e2,1_novaseq_Lib D1211 September_pool_3,D04,,,TGTAATCA,,TGGAGATT,,
1,4902-607-m,4902-607-m,1_novaseq_Lib D1211 September_pool_3,E04,,,TTATATCT,,GTGCAGAC,,
1,4902-607-p,4902-607-p,1_novaseq_Lib D1211 September_pool_3,F04,,,GCCGCAAC,,AGACATGA,,
2,31619-607-m,31619-607-m,1_novaseq_Lib D1211 September_pool_3,E01,,,ATTGTCTG,,TCTGTGAA,,
2,28009-607-p,28009-607-p,1_novaseq_Lib D1211 September_pool_3,C03,,,TCGCCTTG,,GAATCAGC,,
2,17637-607-e1,17637-607-e1,1_novaseq_Lib D1211 September_pool_3,G03,,,TCTCGGTC,,GATGTCAG,,
2,17637-607-e2,17637-607-e2,1_novaseq_Lib D1211 September_pool_3,H03,,,AAGACACT,,TGCTGTCA,,
2,17637-607-m,17637-607-m,1_novaseq_Lib D1211 September_pool_3,A04,,,CTACCAGG,,ATCAGTTG,,
```

Header with sequencing informations

Body with samples informations

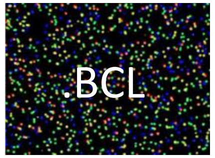
Text file to give when launching the sequencing :

Contain information about **sequencing parameters** and **samples index and lane**

→ **Necessary for demultiplexing step**

DNA sequencing : Demultiplexing

Illumina sequencing

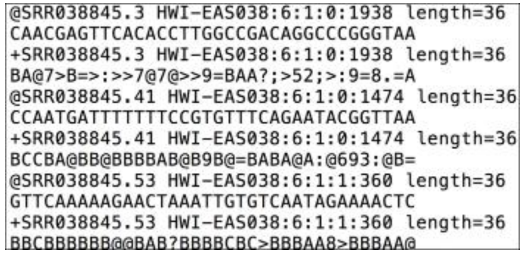


BCL File : compressed picture (Illumina format)

→ Need to be converted for further treatment
→ Need to be separated by sample (one file for all the samples)

Demultiplexing

Ex : BCL2Fastq software



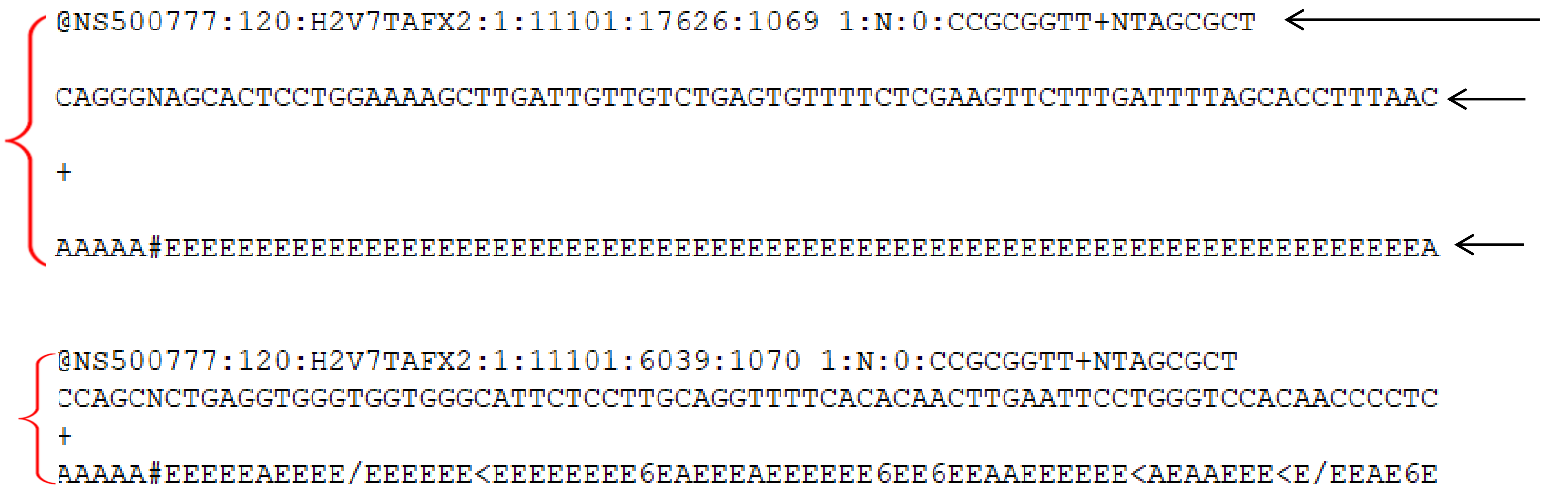
```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

.fastq

Fastq file : text raw file of sequencing

→ Contain raw "reads"
→ Like pieces from a puzzle
→ One file per sample

DNA sequencing : Fastq



Read 1

```
@NS500777:120:H2V7TAFX2:1:11101:17626:1069 1:N:0:CCGCGGTT+NTAGCGCT

CAGGGNAGCACTCCTGGAAAAGCTTGATTGTTGTCTGAGTGTTTTCTCGAAGTTCTTTGATTTTAGCACCTTTAAC

+

AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEA
```

Read information

Read sequence

Quality of each nucleotide

Read 2

```
@NS500777:120:H2V7TAFX2:1:11101:6039:1070 1:N:0:CCGCGGTT+NTAGCGCT
CCAGCNCTGAGGTGGGTGGTGGGCATTCTCCTTGCAGGTTTTCACACAACTTGAATTCCTGGGTCCACAACCCCTC
+
AAAAA#EEEEEAEEEE/EEEEEE<EEEEEEEE6EAEEEAEEEEEE6EE6EEAAEEEEEE<AEAAEEE<E/EEAE6E
```

…

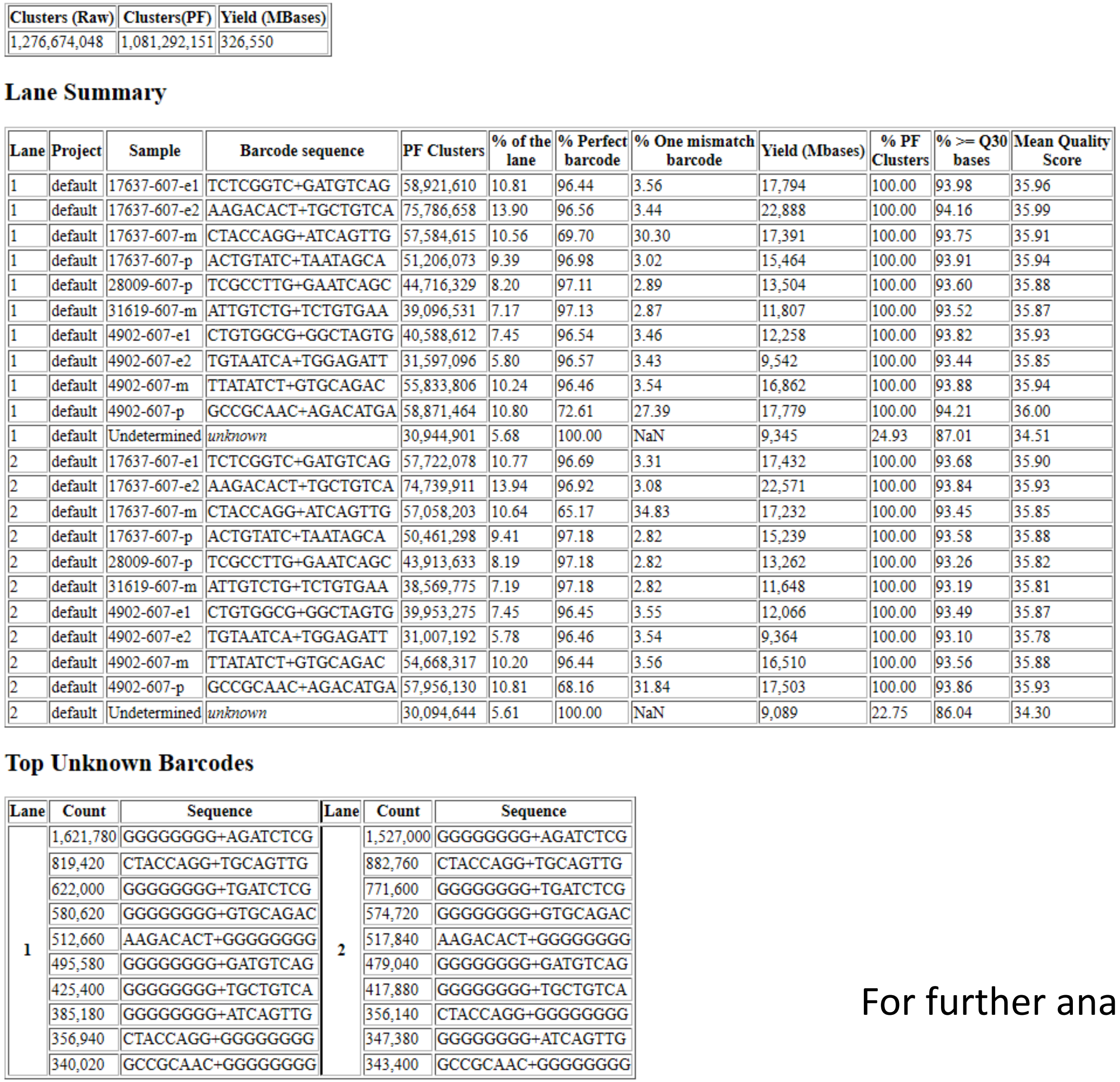
Read n

Same file without quality information

==

Fasta

DNA sequencing : Fastq



Clusters (Raw)	Clusters(PF)	Yield (MBases)
1,276,674,048	1,081,292,151	326,550

Lane Summary

Lane	Project	Sample	Barcode sequence	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
1	default	17637-607-e1	TCTCGGTC+GATGTCAG	58,921,610	10.81	96.44	3.56	17,794	100.00	93.98	35.96
1	default	17637-607-e2	AAGACACT+TGCTGTCA	75,786,658	13.90	96.56	3.44	22,888	100.00	94.16	35.99
1	default	17637-607-m	CTACCAGG+ATCAGTTG	57,584,615	10.56	69.70	30.30	17,391	100.00	93.75	35.91
1	default	17637-607-p	ACTGTATC+TAATAGCA	51,206,073	9.39	96.98	3.02	15,464	100.00	93.91	35.94
1	default	28009-607-p	TCGCCTTG+GAATCAGC	44,716,329	8.20	97.11	2.89	13,504	100.00	93.60	35.88
1	default	31619-607-m	ATTGTCTG+TCTGTGAA	39,096,531	7.17	97.13	2.87	11,807	100.00	93.52	35.87
1	default	4902-607-e1	CTGTGGCG+GGCTAGTG	40,588,612	7.45	96.54	3.46	12,258	100.00	93.82	35.93
1	default	4902-607-e2	TGTAATCA+TGGAGATT	31,597,096	5.80	96.57	3.43	9,542	100.00	93.44	35.85
1	default	4902-607-m	TTATATCT+GTGCAGAC	55,833,806	10.24	96.46	3.54	16,862	100.00	93.88	35.94
1	default	4902-607-p	GCCGCAAC+AGACATGA	58,871,464	10.80	72.61	27.39	17,779	100.00	94.21	36.00
1	default	Undetermined	*unknown*	30,944,901	5.68	100.00	NaN	9,345	24.93	87.01	34.51
2	default	17637-607-e1	TCTCGGTC+GATGTCAG	57,722,078	10.77	96.69	3.31	17,432	100.00	93.68	35.90
2	default	17637-607-e2	AAGACACT+TGCTGTCA	74,739,911	13.94	96.92	3.08	22,571	100.00	93.84	35.93
2	default	17637-607-m	CTACCAGG+ATCAGTTG	57,058,203	10.64	65.17	34.83	17,232	100.00	93.45	35.85
2	default	17637-607-p	ACTGTATC+TAATAGCA	50,461,298	9.41	97.18	2.82	15,239	100.00	93.58	35.88
2	default	28009-607-p	TCGCCTTG+GAATCAGC	43,913,633	8.19	97.18	2.82	13,262	100.00	93.26	35.82
2	default	31619-607-m	ATTGTCTG+TCTGTGAA	38,569,775	7.19	97.18	2.82	11,648	100.00	93.19	35.81
2	default	4902-607-e1	CTGTGGCG+GGCTAGTG	39,953,275	7.45	96.45	3.55	12,066	100.00	93.49	35.87
2	default	4902-607-e2	TGTAATCA+TGGAGATT	31,007,192	5.78	96.46	3.54	9,364	100.00	93.10	35.78
2	default	4902-607-m	TTATATCT+GTGCAGAC	54,668,317	10.20	96.44	3.56	16,510	100.00	93.56	35.88
2	default	4902-607-p	GCCGCAAC+AGACATGA	57,956,130	10.81	68.16	31.84	17,503	100.00	93.86	35.93
2	default	Undetermined	*unknown*	30,094,644	5.61	100.00	NaN	9,089	22.75	86.04	34.30

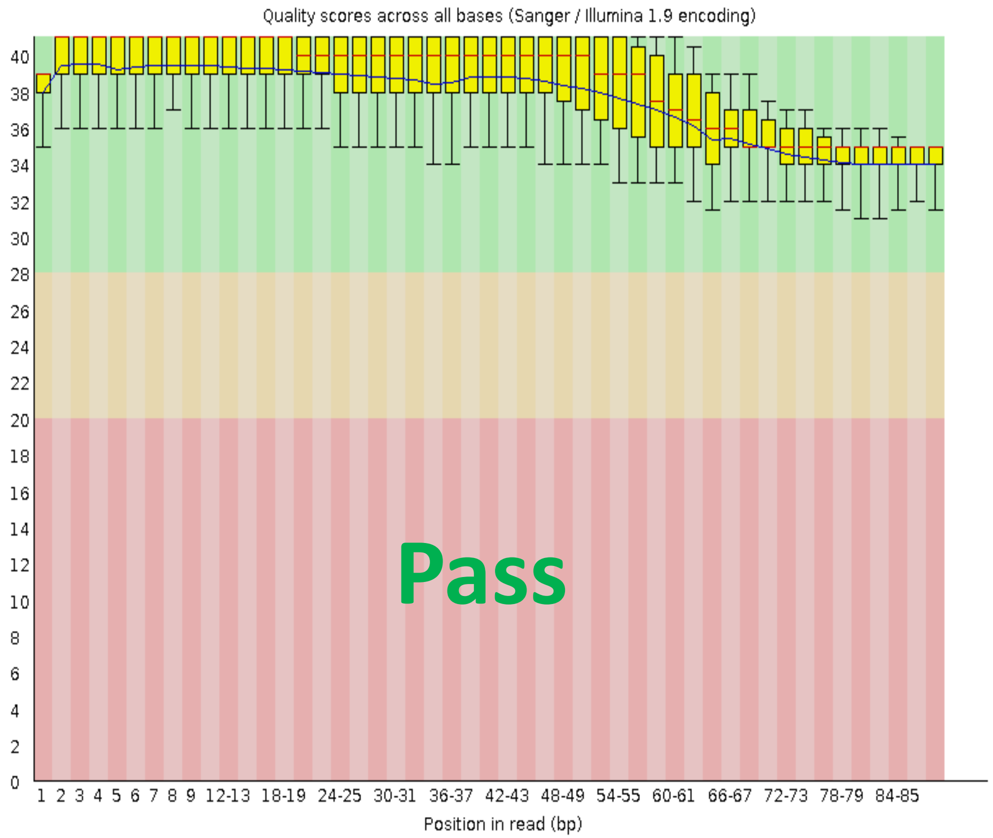
Top Unknown Barcodes

Lane	Count	Sequence	Lane	Count	Sequence
1	1,621,780	GGGGGGGG+AGATCTCG	2	1,527,000	GGGGGGGG+AGATCTCG
	819,420	CTACCAGG+TGCAGTTG		882,760	CTACCAGG+TGCAGTTG
	622,000	GGGGGGGG+TGATCTCG		771,600	GGGGGGGG+TGATCTCG
	580,620	GGGGGGGG+GTGCAGAC		574,720	GGGGGGGG+GTGCAGAC
	512,660	AAGACACT+GGGGGGGG		517,840	AAGACACT+GGGGGGGG
	495,580	GGGGGGGG+GATGTCAG		479,040	GGGGGGGG+GATGTCAG
	425,400	GGGGGGGG+TGCTGTCA		417,880	GGGGGGGG+TGCTGTCA
	385,180	GGGGGGGG+ATCAGTTG		356,140	CTACCAGG+GGGGGGGG
	356,940	CTACCAGG+GGGGGGGG		347,380	GGGGGGGG+ATCAGTTG
	340,020	GCCGCAAC+GGGGGGGG		343,400	GCCGCAAC+GGGGGGGG

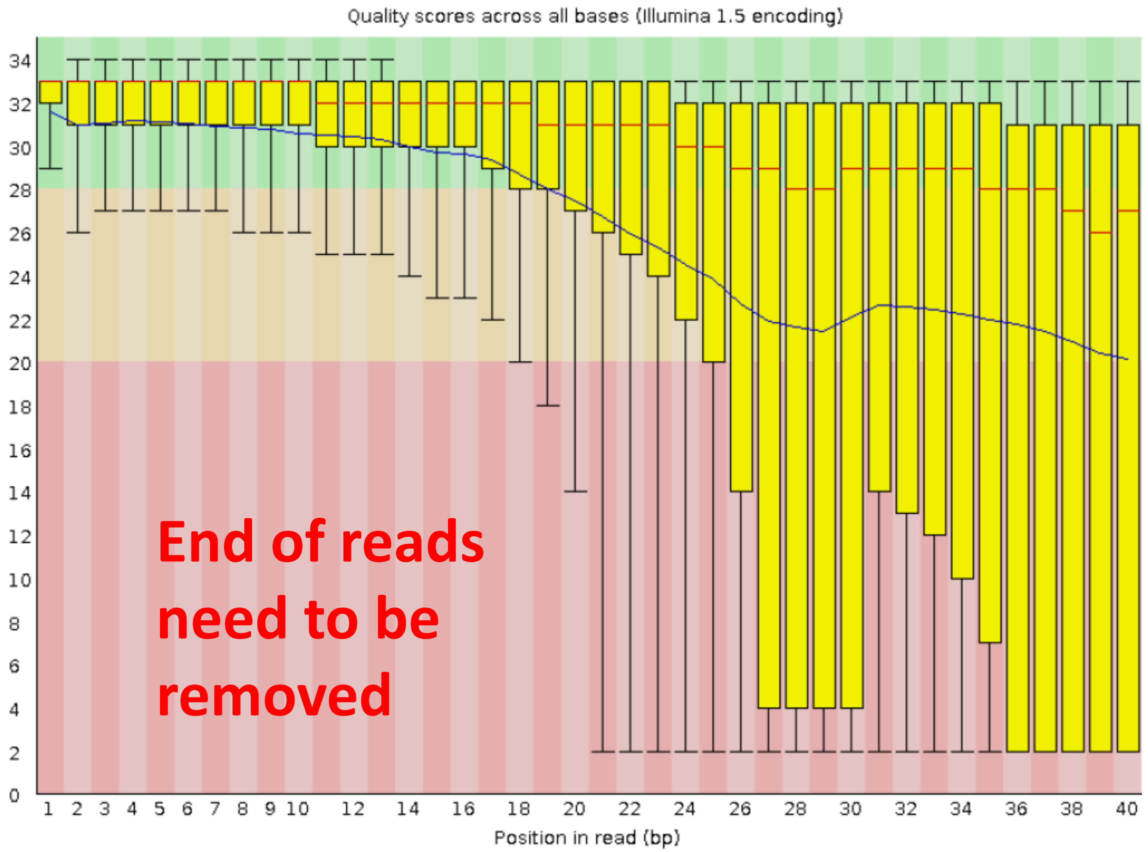
Demultiplexing report gives informations about sequencing quality and quantity

For further analysis, the reads need to be filtered : **the quality control step**

DNA sequencing : Quality Control



Pass



End of reads need to be removed

Quality control tells you what happened during your sequencing and for example allow to check if some reads or parts of read need to be remove before going further

For further analysis, the reads need to be ordered : **the alignment step**

Quality control software like FastQC

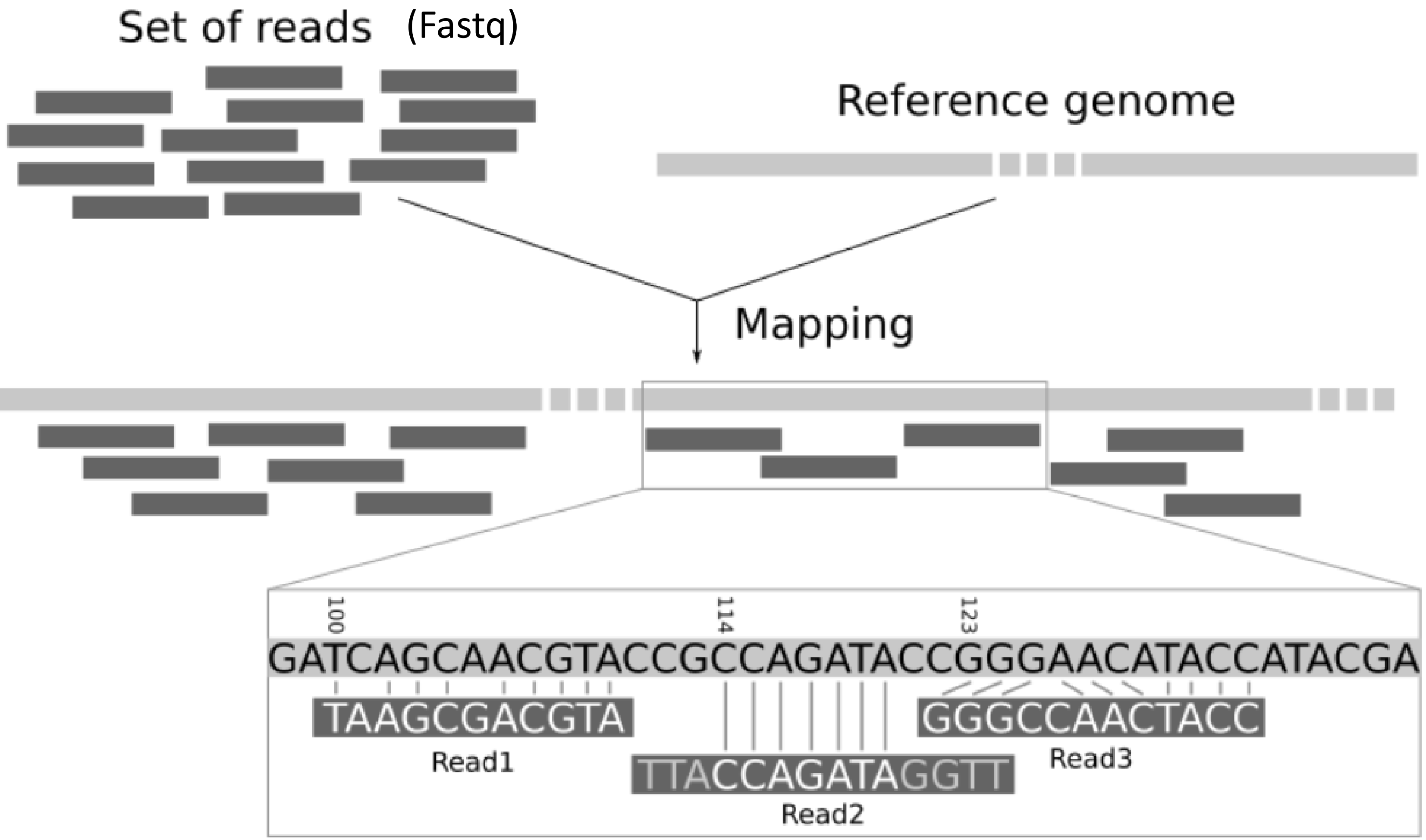
DNA sequencing : Alignment

With human samples, we can use a known **reference genome**.

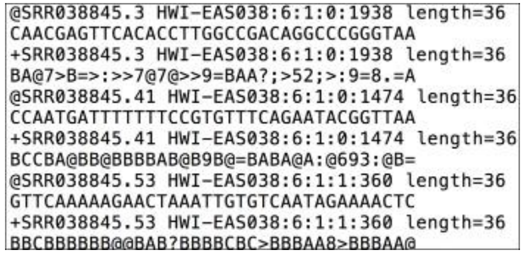
The alignment software try each position of the genome with each read and select the best one

Each nucleotide **match** or **mismatch** increase or decrease a global score.

Different alignment software can be used, each one working differently.



DNA sequencing : Alignment



```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

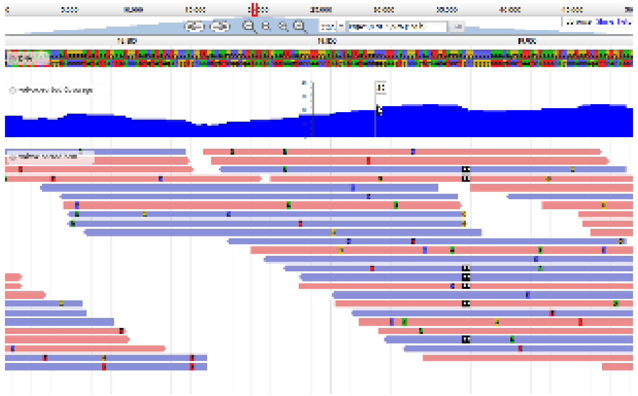
.fastq

Fastq file : text raw file of sequencing

→ Contain raw "reads"
→ Like pieces from a puzzle

Alignment

Ex : BWA software



.sam / .bam / .cram

Bam file : text file containing sequence, quality, and position over a genome

→ Reads can be visualized over a genome
→ Can be used to determine the DNA sequence of a patient

DNA sequencing : BAM file

QNAME	FLAG	CHR	POS	MAPQ	CIGAR	CHRNEXT	PNEXT	TLEN	SEQ	QUAL	Info	
A00554:29:2:2150:6479:12336	99	chr1	3206995	100	76M	=	3207040	121	CTCCCAGGAATCCATTGG	FFFFFFFFFFFFF:	NH:i:2	Read 1
A00554:29:2:1212:15691:25848	99	chr1	3207262	100	56M6121N20M	=	3213478	170	GTTGGATTAATTAACTGCA	FFFFFFFFFFFFF	NH:i:1	Read 2
												...
												Read n

```
REF:AGCTAGCATCGTGTCGCCCGTCTAGCATACGCATGATCGACTGTCAGCTAGTCAGACTAGTC

Read:         GTGTAACCC................................TCAGAATA
```

CIGAR : 9M32N8M ➔ 9 match 32 pb intron 8 match

Bam file : file containing sequence, quality, and position over a genome

→ Reads can be visualized over a genome

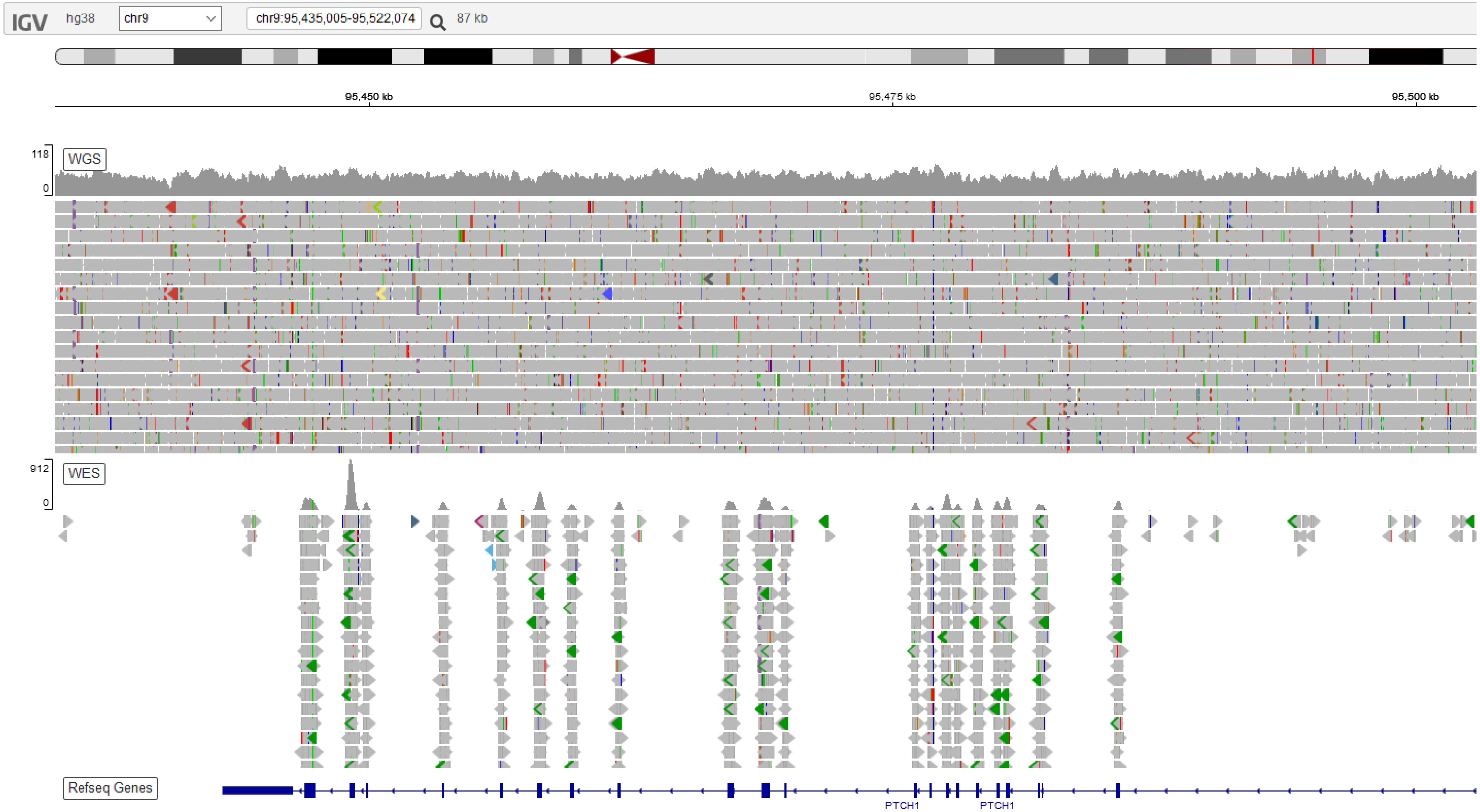
→ Can be used to determine the DNA sequence of a patient

➔ No information about variations compared to reference genome

DNA sequencing : WGS / WES

Whole Genome Sequencing

Whole Exome Sequencing

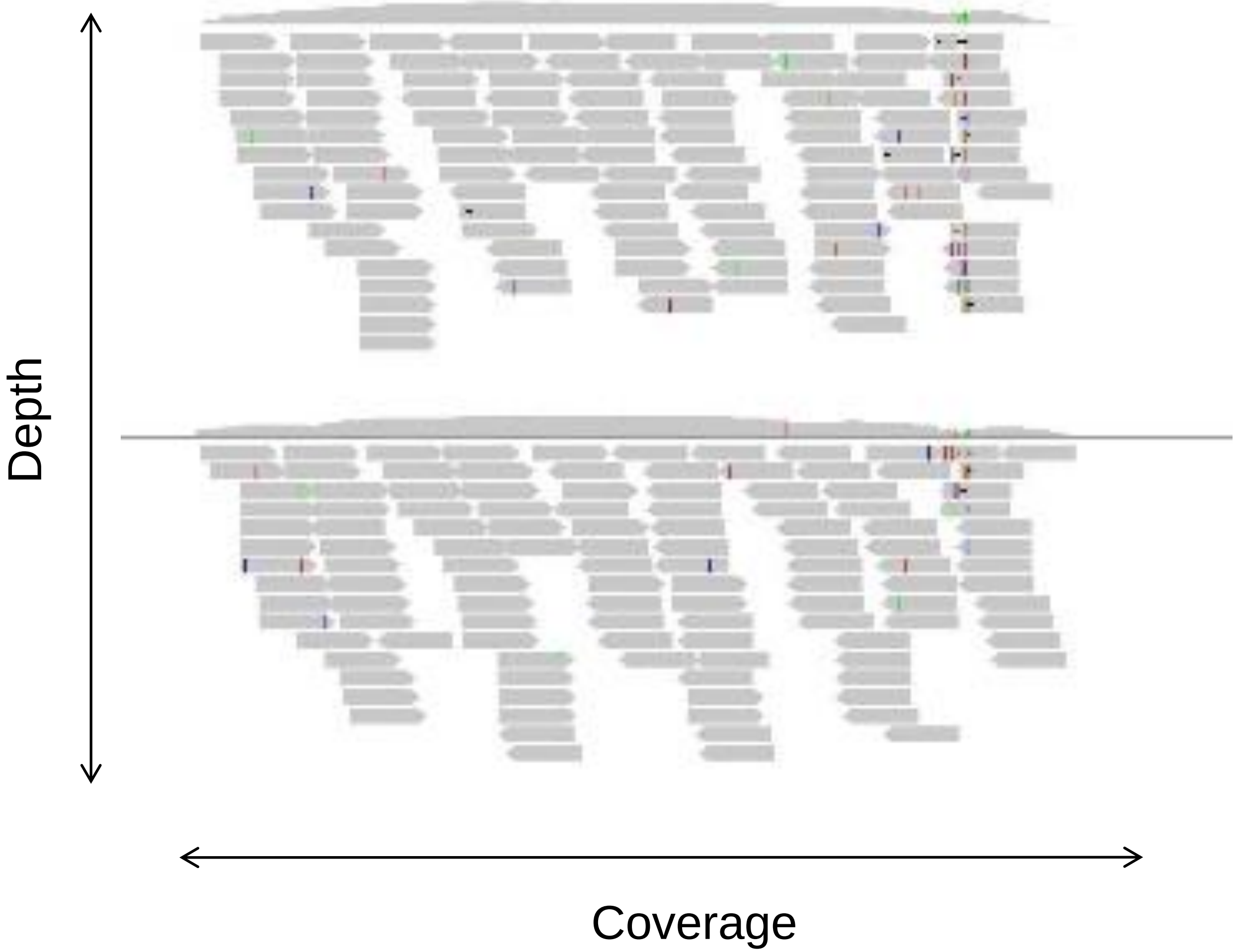


igv.org

DNA sequencing : Alignment

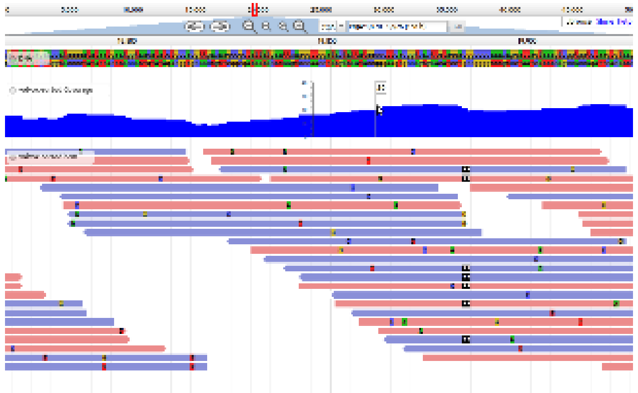
- **Coverage**
Proportion of the genome with a depth of at least X reads (mostly expressed in %)

- **Depth**
Number of reads at one position of the genome (numeric value)



➔ No information about variations compared to reference genome

DNA sequencing : Variant calling



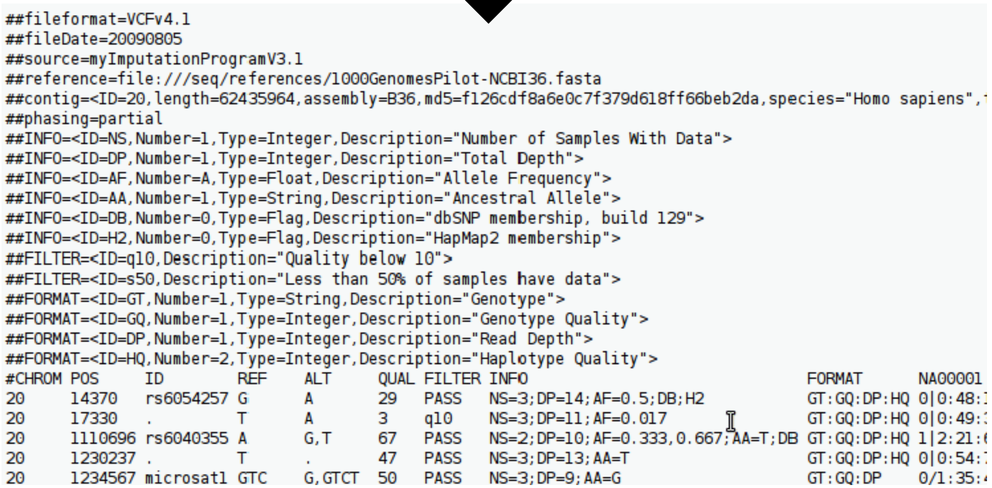
.sam / .bam / .cram

Bam file : text file containing sequence, quality, and position over a genome

→ Reads can be visualized over a genome
→ Can be used to determine the DNA sequence of a patient

Variant Calling

Ex : GATK software



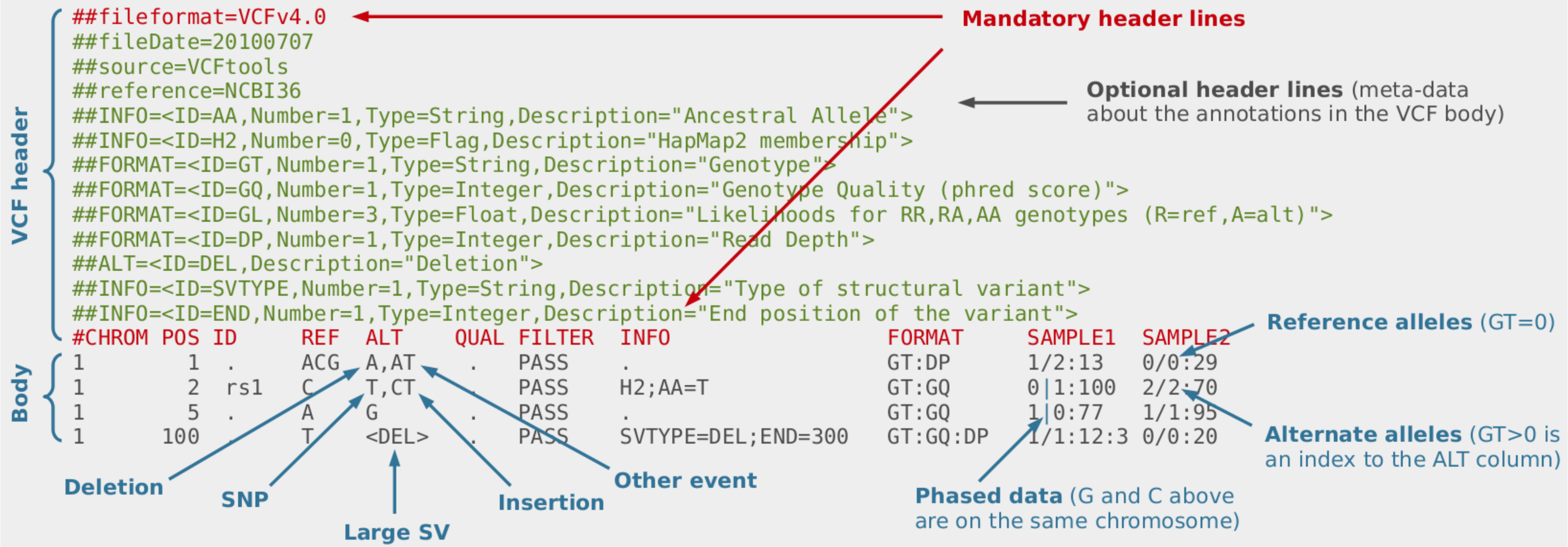
```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:
```

.VCF file

VCF file: text file containing variations found in sample compared to reference genome

→ Used to look for potential pathogenic variations
→ Can contain point variations or insertion / deletion

DNA sequencing : VCF file



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID    REF  ALT    QUAL FILTER INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1 .     ACG  A,AT    .   PASS   .                    GT:DP      1/2:13    0/0:29
1        2 rs1   C    T,CT    .   PASS   H2;AA=T              GT:GQ      0|1:100   2/2:70
1        5 .     A    G       .   PASS   .                    GT:GQ      1|0:77    1/1:95
1      100 .     T    <DEL>   .   PASS   SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

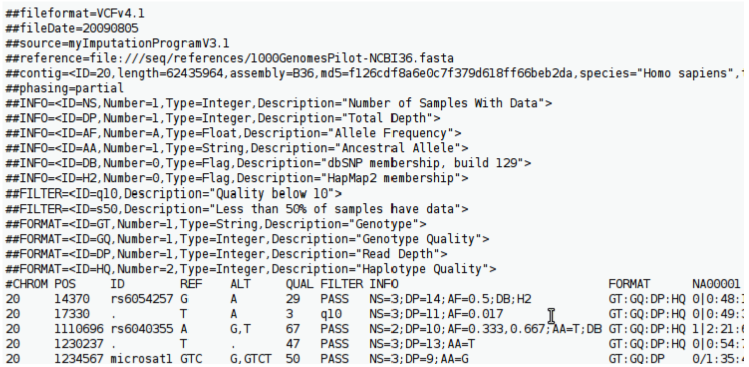
Two parts in VCF file:

-**The Header** : Information about how the vcf was made and it's content

-**The Body** : Information about each variation found (one line each)

→ Missing supplementary informations to conclude about pathogenicity of the variants

DNA sequencing : Annotation step



```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:
```

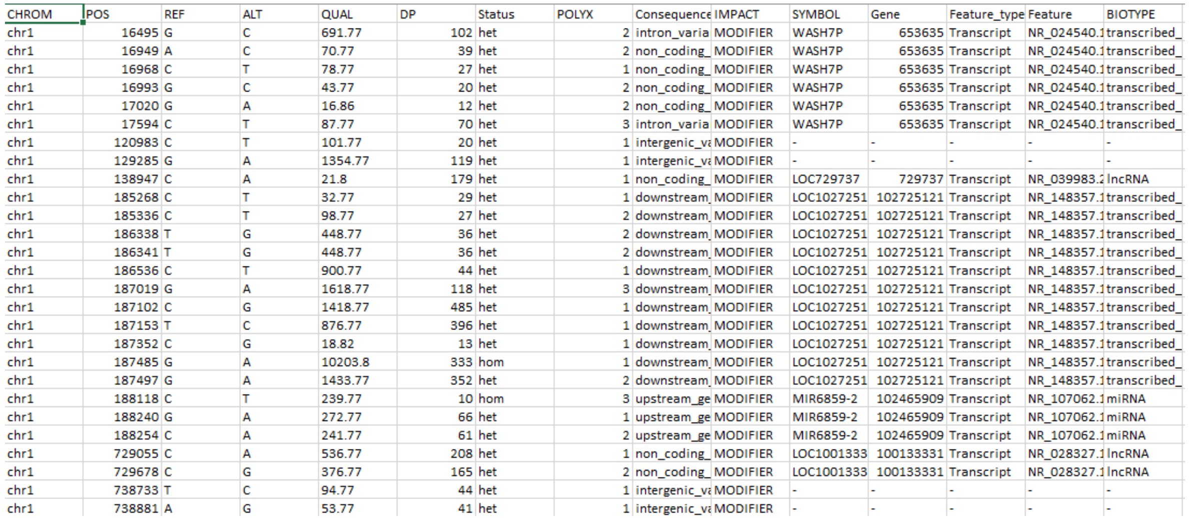
.VCF file

VCF file: text file containing variations found in sample compared to reference genome

→ Used to look for potential pathogenic variations
→ Can contain point variations or insertion / deletion

Annotation

Ex : VEP software (with public or private databases)



CHROM	POS	REF	ALT	QUAL	DP	Status	POLYX	Consequence	IMPACT	SYMBOL	Gene	Feature_type	Feature	BIOTYPE
chr1	16495	G	C	691.77	102	het	2	intron_varia	MODIFIER	WASH7P	653635	Transcript	NR_024540.1	transcribed_
chr1	16949	A	C	70.77	39	het	2	non_coding_	MODIFIER	WASH7P	653635	Transcript	NR_024540.1	transcribed_
chr1	16968	C	T	78.77	27	het	1	non_coding_	MODIFIER	WASH7P	653635	Transcript	NR_024540.1	transcribed_
chr1	16993	G	C	43.77	20	het	2	non_coding_	MODIFIER	WASH7P	653635	Transcript	NR_024540.1	transcribed_
chr1	17020	G	A	16.86	12	het	2	non_coding_	MODIFIER	WASH7P	653635	Transcript	NR_024540.1	transcribed_
chr1	17594	C	T	87.77	70	het	3	intron_varia	MODIFIER	WASH7P	653635	Transcript	NR_024540.1	transcribed_
chr1	120983	C	T	101.77	20	het	1	intergenic_va	MODIFIER	-	-	-	-	-
chr1	129285	G	A	1354.77	119	het	1	intergenic_va	MODIFIER	-	-	-	-	-
chr1	138947	C	A	21.8	179	het	1	non_coding_	MODIFIER	LOC729737	729737	Transcript	NR_039983.2	lncRNA
chr1	185268	C	T	32.77	29	het	1	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	185336	C	T	98.77	27	het	2	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	186338	T	G	448.77	36	het	2	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	186341	T	G	448.77	36	het	2	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	186536	C	T	900.77	44	het	1	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	187019	G	A	1618.77	118	het	3	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	187102	C	G	1418.77	485	het	1	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	187153	T	C	876.77	396	het	1	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	187352	C	G	18.82	13	het	1	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	187485	G	A	10203.8	333	hom	1	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	187497	G	A	1433.77	352	het	2	downstream_	MODIFIER	LOC1027251	102725121	Transcript	NR_148357.1	transcribed_
chr1	188118	C	T	239.77	10	hom	3	upstream_ge	MODIFIER	MIR6859-2	102465909	Transcript	NR_107062.1	miRNA
chr1	188240	G	A	272.77	66	het	1	upstream_ge	MODIFIER	MIR6859-2	102465909	Transcript	NR_107062.1	miRNA
chr1	188254	C	A	241.77	61	het	2	upstream_ge	MODIFIER	MIR6859-2	102465909	Transcript	NR_107062.1	miRNA
chr1	729055	C	A	536.77	208	het	1	non_coding_	MODIFIER	LOC1001333	100133331	Transcript	NR_028327.1	lncRNA
chr1	729678	C	G	376.77	165	het	2	non_coding_	MODIFIER	LOC1001333	100133331	Transcript	NR_028327.1	lncRNA
chr1	738733	T	C	94.77	44	het	1	intergenic_va	MODIFIER	-	-	-	-	-
chr1	738881	A	G	53.77	41	het	1	intergenic_va	MODIFIER	-	-	-	-	-

Annotated .VCF file

VCF file: text file containing variations found in sample compared to reference genome & supplementary information about these variations (ex :gene name / frequency / AA change …

→ Contain necessary information to assess pathogenicity
→ Annotations depends of the data and the question asked

Focus on 4 sequencing methods

DNA-seq

Determine the genome sequence

→ In clinical context, find variants which could explain diseases

RNA-seq

Determine the genes expressed and their expression level

→ In clinical context, find unexpressed or over expressed genes which could explain diseases

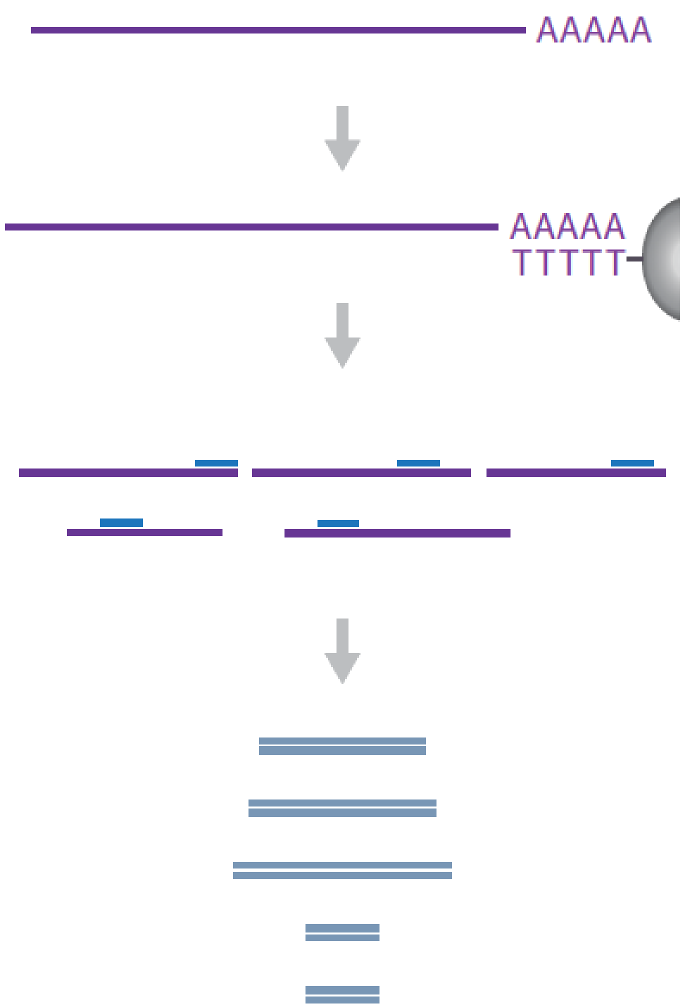
ChIP-seq

Determine the interactions between DNA and proteins

Hi-C

Determine the interactions between DNA and DNA

RNA sequencing



mRNA

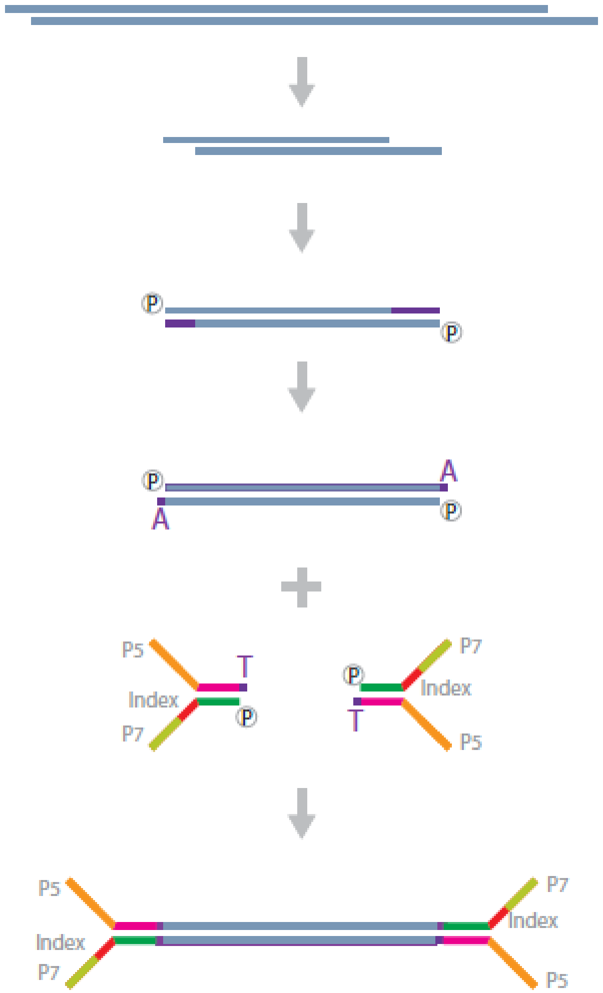
polyA select

Fragment

Random hexamer

First and second strand synthesis

Double-stranded DNA



Fractionate
Size select

End repair
Phosphorylate

A-overhang

Add Adaptors

Adaptor ligation

Product ready for cluster generation

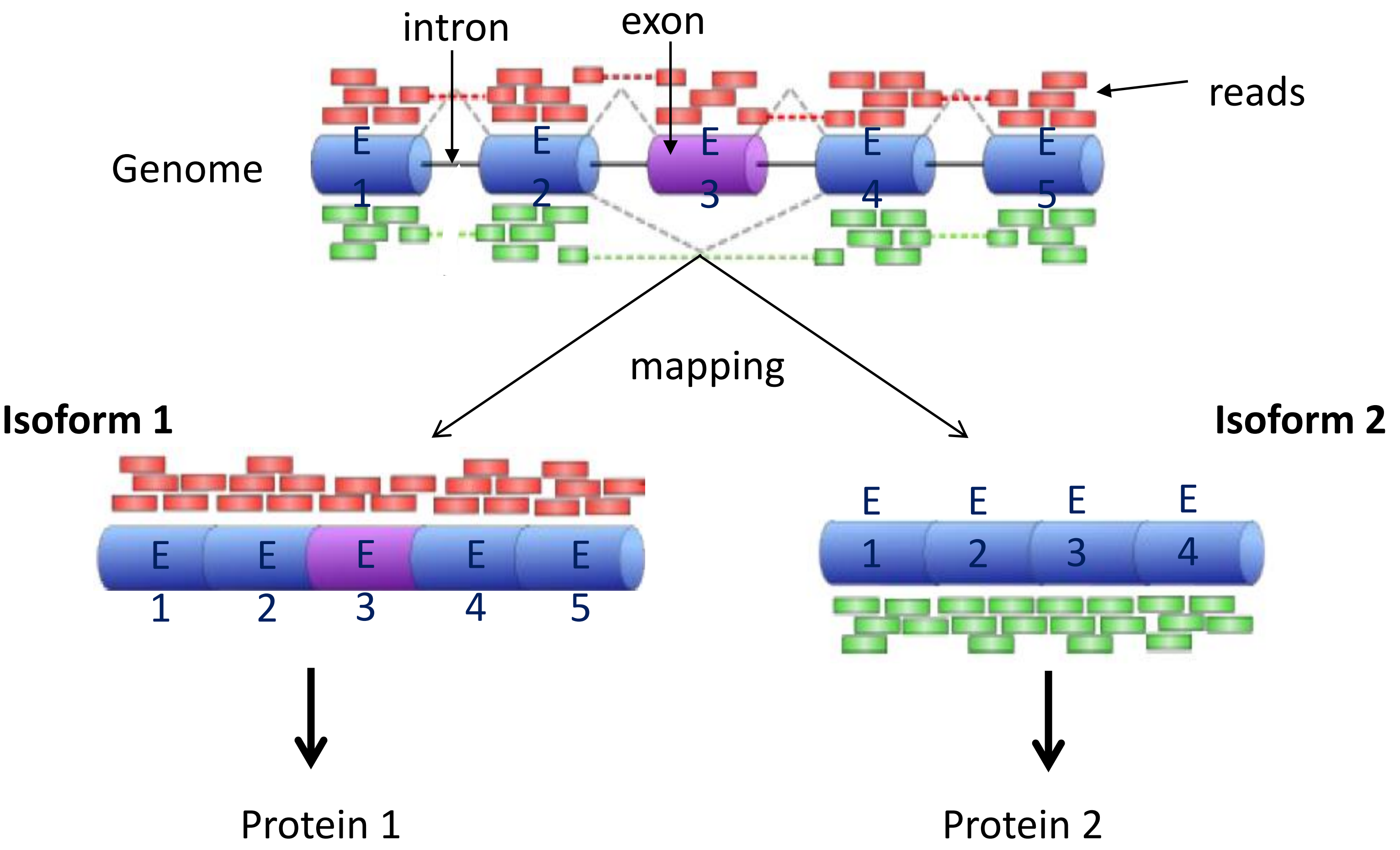
RNA sequencing

Demultiplexing step identical to DNAseq

Alignment step specific to RNA-seq
→ Need to deal with RNA **splicing**

RNA specialised aligner can easily "split" to take in account splicing site

(the penalty of splitting isn't important to the final alignment score of the read)
Ex : STAR



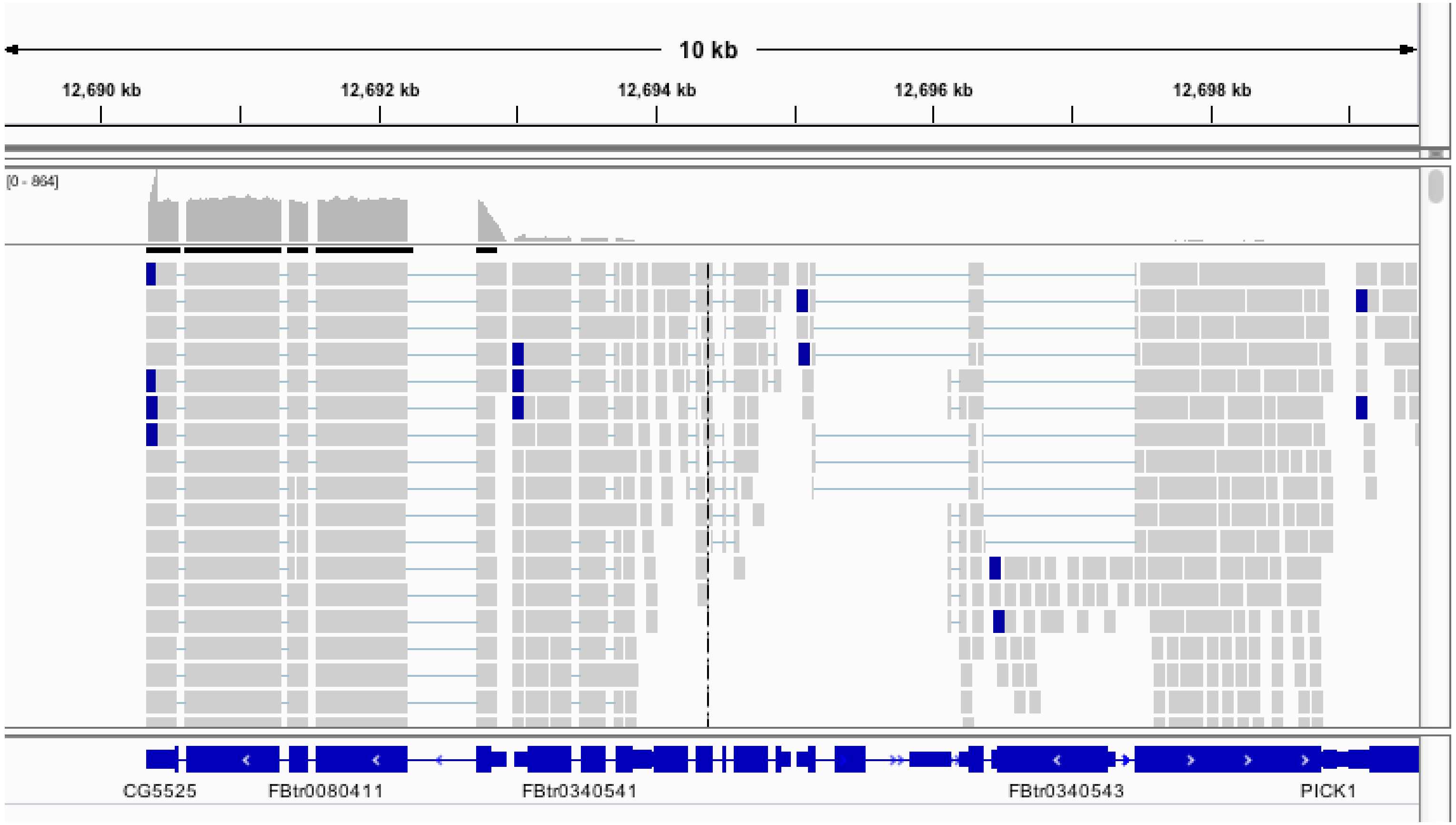
RNA sequencing : splicing

Demultiplexing step identical to DNAseq

Alignment step specific to RNA-seq
→ Need to deal with RNA **splicing**

RNA specialized aligners can easily "split" the reads to take in account splicing sites

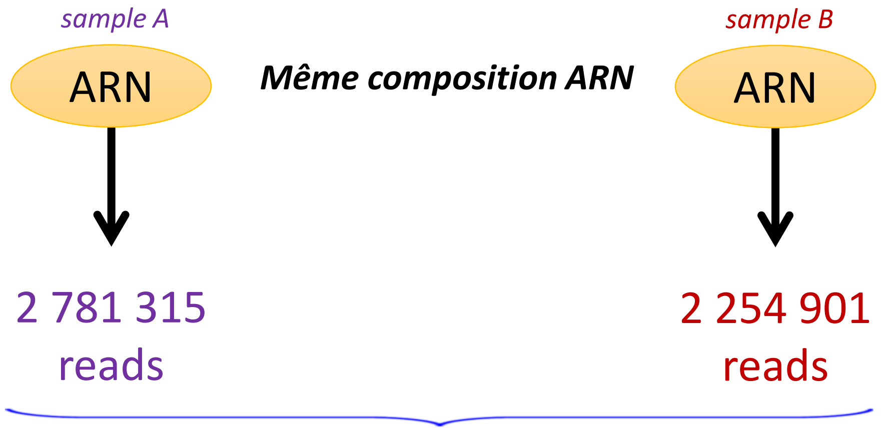
(the penalty of splitting isn't important to the final alignment score of the read)
Ex : STAR



The reads can now be counted to determine the expression level of genes → **Counting step**

RNA sequencing : counting step bias

Sample number of read effect:



sample A

ARN

2 781 315 reads

Même composition ARN

sample B

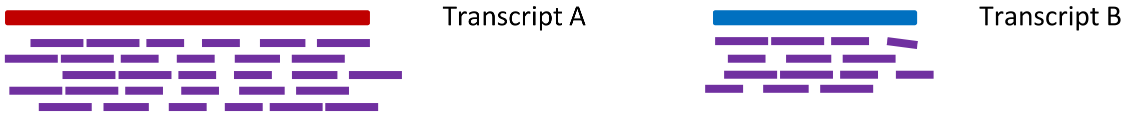
ARN

2 254 901 reads

Artificialy 1.2334 fois more of each RNA fragment for sample A has been sequenced, despite a real DNA quantity per cell identical between the two samples

Gene length effect :

With a same expression level, a long transcript will have more reads



If differential expression based on raw counts, it will create a biais

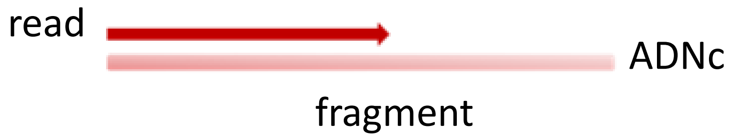
⇒ Correction of these biais needed

⇒ →**Normalisation step**

RNA sequencing : Normalisation methods

1) **RPKM** (**R**eads **P**er **K**ilobase **M**illion)

Single End => read = fragment

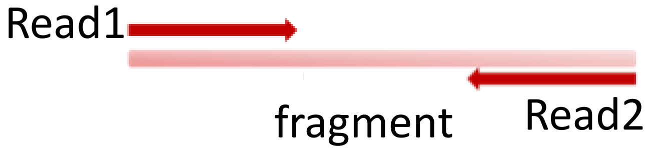


$$\frac{\text{Total read count of sample}}{1\ 000\ 000}$$ **= RPM** =>1st Normalisation/sequencing depth

$$\frac{\textbf{RPM}}{\text{Gene length in kb}}$$ **= RPKM** =>2nd Normalisation/gene length

2) **FPKM** (**F**ragments **P**er **K**ilobase **M**illion)

Paired End => R1/R2 = fragment



Only difference between **RPKM** et **FPKM**
=> **FPKM** count only one read per fragment

RNA sequencing : counting step bias

3) **TPM** (Transcripts Per Kilobase Million)

$\dfrac{\text{Total read count of gene}}{\text{longueur de gène kilobase}}$ **= RPK** => 1st Normalisation/gene length

$\dfrac{\textbf{RPK}}{1\ 000\ 000}$ **= TPM** => 2nd Normalisation/Depth

RNA sequencing : counting step bias

Exemples

RPKM (**R**eads **P**er **K**ilobase **M**illion)

gene_id	Rep1 **counts**	Rep2 **counts**	Rep3 **counts**
ENSMUSG1 (2kb)	10	12	30
ENSMUSG2 (4kb)	20	25	60
ENSMUSG3 (1kb)	5	8	15
ENSMUSG4 (10kb)	0	0	1
Total reads	35	45	106
/10	3,5	4,5	10,6

1) Normalisation over read depth

gene_id	Rep1 **RPM**	Rep2 **RPM**	Rep3 **RPM**
ENSMUSG1 (2kb)	2,85	2,66	2,83
ENSMUSG2 (4kb)	5,71	5,55	5,66
ENSMUSG3 (1kb)	1,42	1,77	1,41
ENSMUSG4 (10kb)	0	0	0,094

2) Normalisation / gene length /kb

gene_id	Rep 1 RPKM	Rep 2 RPKM	Rep 3 RPKM
ENSMUSG1 (2kb)	1,42	1,33	1,41
ENSMUSG2 (4kb)	1,42	1,38	1,41
ENSMUSG3 (1kb)	1,42	1,77	1,41
ENSMUSG4 (10kb)	0	0	0,009
Total	4,26	4,48	4,239

RNA sequencing : counting step bias

before

gene_id	Rep1 **counts**	Rep2 **counts**	Rep3 **counts**
ENSMUSG1 (2kb)	10	12	30
ENSMUSG2 (4kb)	20	25	60
ENSMUSG3 (1kb)	5	8	15
ENSMUSG4 (10kb)	0	0	1

after

gene_id	Rep 1 RPKM	Rep 2 RPKM	Rep 3 RPKM
ENSMUSG1 (2kb)	1,42	1,33	1,41
ENSMUSG2 (4kb)	1,42	1,38	1,41
ENSMUSG3 (1kb)	1,42	1,77	1,41
ENSMUSG4 (10kb)	0	0	0,009

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
ENSMUSG00000000001	ENSMUST00000000001	3262.00	3098.67	3467.00	29.14	26.19
ENSMUSG00000000003	ENSMUST00000000003,ENSMUST00000114041	799.50	636.17	0.00	0.00	0.00
ENSMUSG00000000028	ENSMUST00000000028,ENSMUST00000096990,ENSMUST00000115585	1900.86	1737.53	308.00	4.62	4.15
ENSMUSG00000000031	ENSMUST00000132294,ENSMUST00000136359,ENSMUST00000140716,ENSMUST00000149974,ENSMUST00000152754	2170.51	2007.18	88.00	1.14	1.03

Example of counting file whith count per gene

Focus on 4 sequencing methods

DNA-seq

Determine the genome sequence

→ In clinical context, find variants which could explain diseases

RNA-seq

Determine the genes expressed and their expression level

→ In clinical context, find unexpressed or over expressed genes which could explain diseases

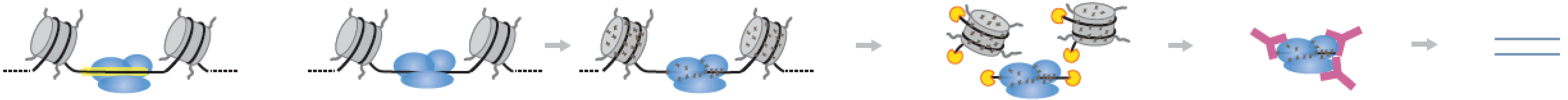
ChIP-seq

Determine the interactions between DNA and proteins

Hi-C

Determine the the interactions between DNA and DNA

ChIP-seq



Chromatin immune precipitation (ChIP-Seq), High-throughput chromatin immunoprecipitation (HT-ChIP))

DNA-protein complex

Crosslink proteins and DNA

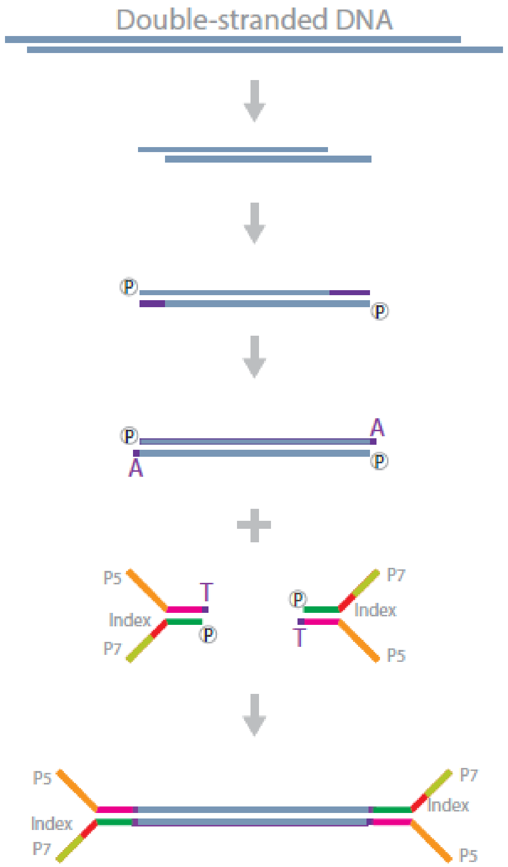
Sample fragmentation

Exonuclease digestion

Immunoprecipitate

DNA extraction

DNA



Fractionate
Size select

End repair
Phosphorylate

A-overhang

Add Adaptors

Adaptor ligation

Product ready for cluster generation

ChIP-seq

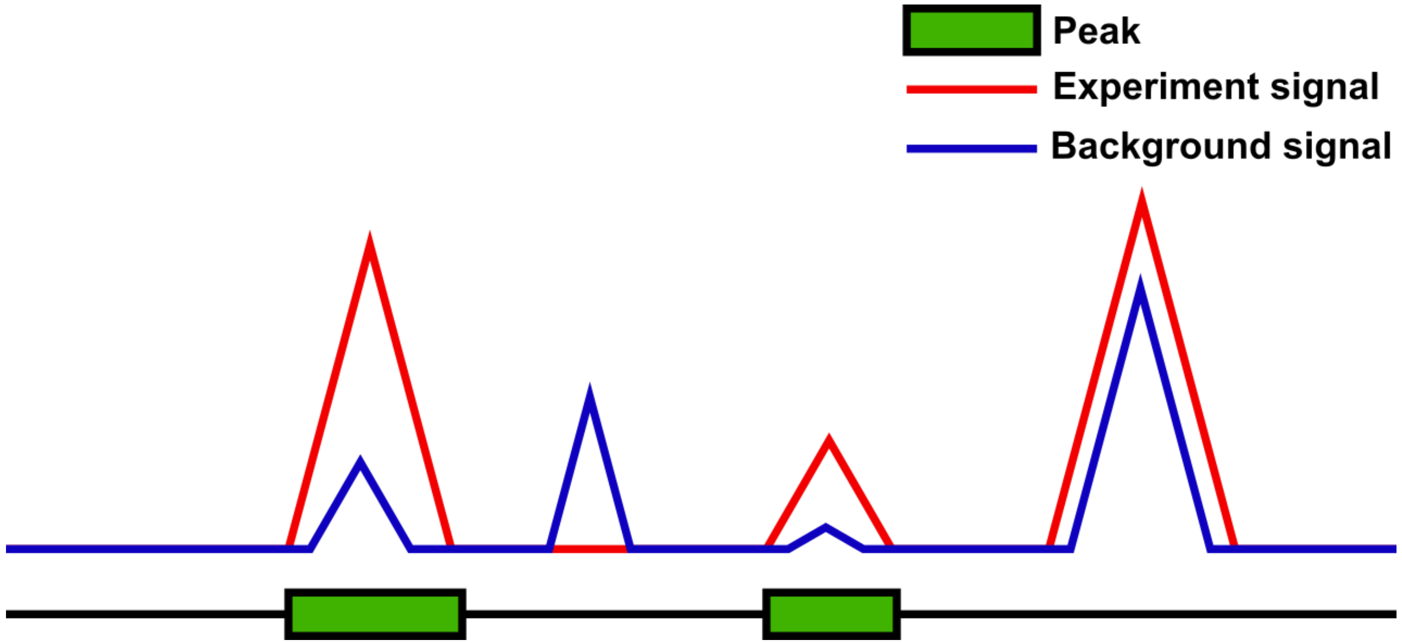
Demultiplexing step identical to DNAseq

Alignment step identical to DNAseq

Next step is called **Peak Calling**

Sofware example : MACS2

You compare the signal (read depth) of you sample of interest with a control condition (like non antibody captured dna from your sample) to keep only signal specific to your sample



https://www.researchgate.net/figure/The-transcription-factor-and-histone-modification-landscape-for-inactive-and-active_fig2_277780300

ChIP-seq

Can be used with every DNA biding protein, as long as you can catch it with antibody

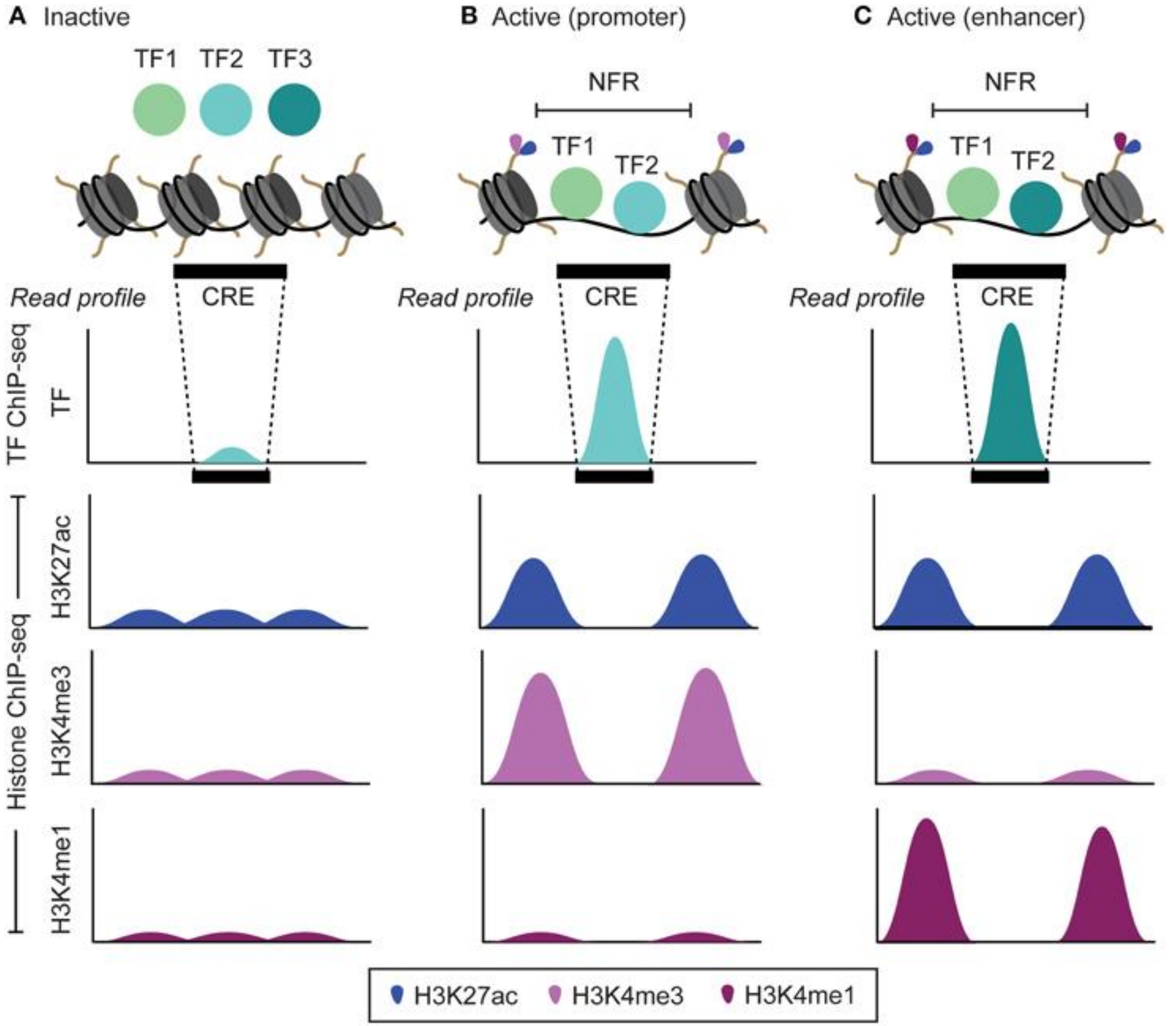
For example, with **histone marks** H3K4me1 / H3K4me3 / H3K27ac, used to determine **promoter / enhancer** position

H3K4me1 + H3K27ac = enhancer

H3K4me3 + H3K27ac = promoter

You can also compare the effect of a treatment by differentially analyzing a treated vs non treated sample

Chromatin state



https://www.researchgate.net/figure/The-transcription-factor-and-histone-modification-landscape-for-inactive-and-active_fig2_277780300

Focus on 4 sequencing methods

DNA-seq

Determine the genome sequence

→ In clinical context, find variants which could explain diseases

RNA-seq

Determine the genes expressed and their expression level

→ In clinical context, find unexpressed or over expressed genes which could explain diseases

ChIP-seq

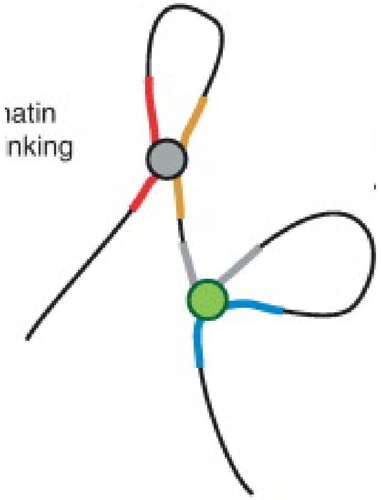
Determine the interactions between DNA and proteins

Hi-C

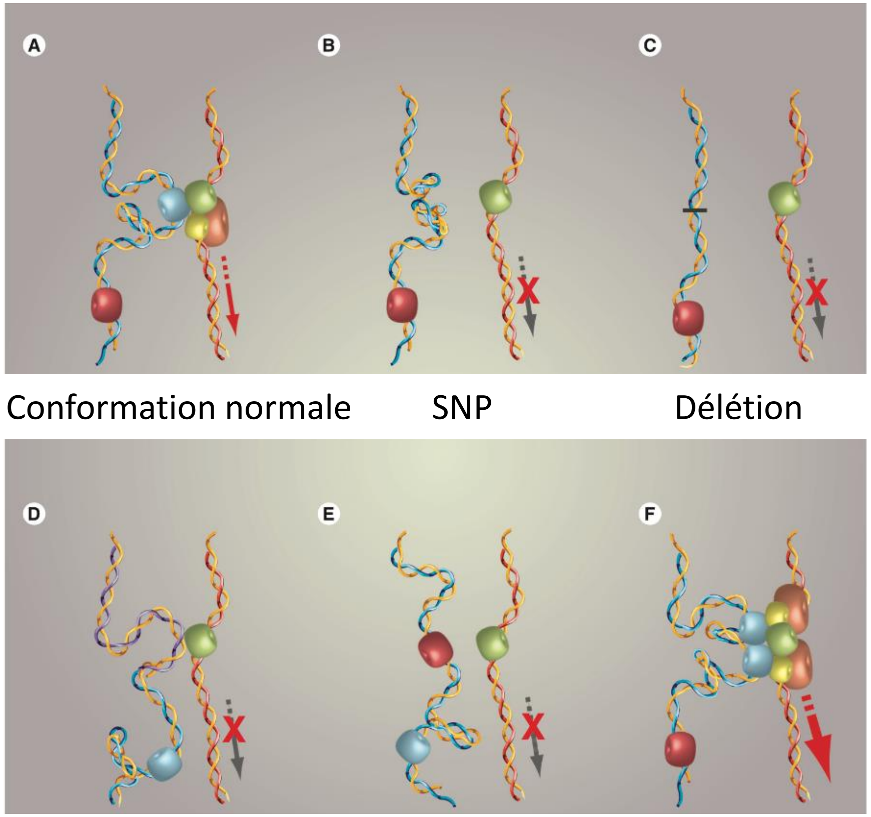
Determine the interactions between DNA and DNA

Hi-C

Chromatin is forming loops in the nucleus, putting in close proximity linearly spaced parts of the genome



Montavon *et al.*, 2012



Conformation normale SNP Délétion

Insertion Inversion CNV

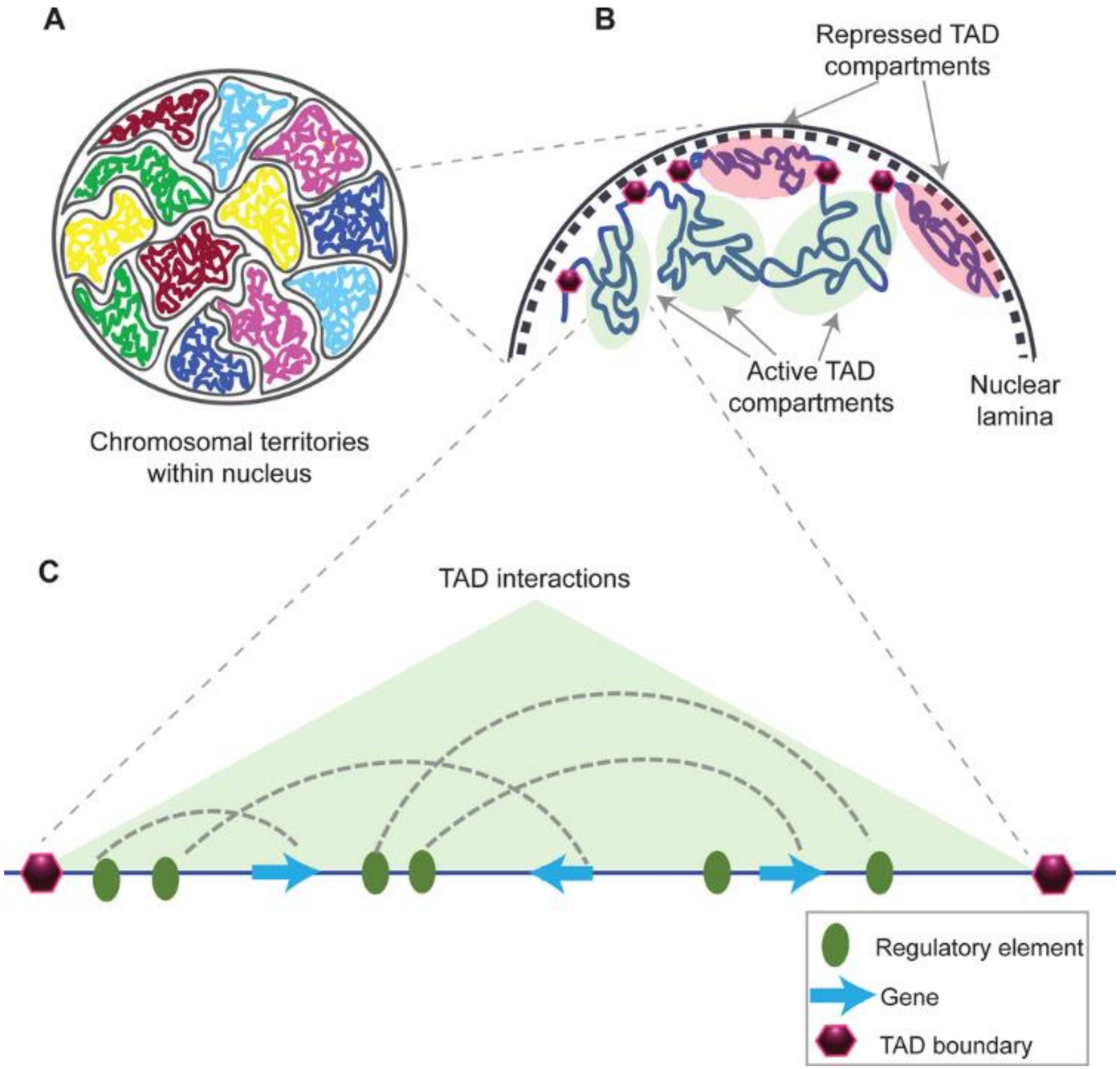
Adapté de Crutchley *et a*l, *Biomarkers Med.* 2010

Possible effect of DNA events over these loops

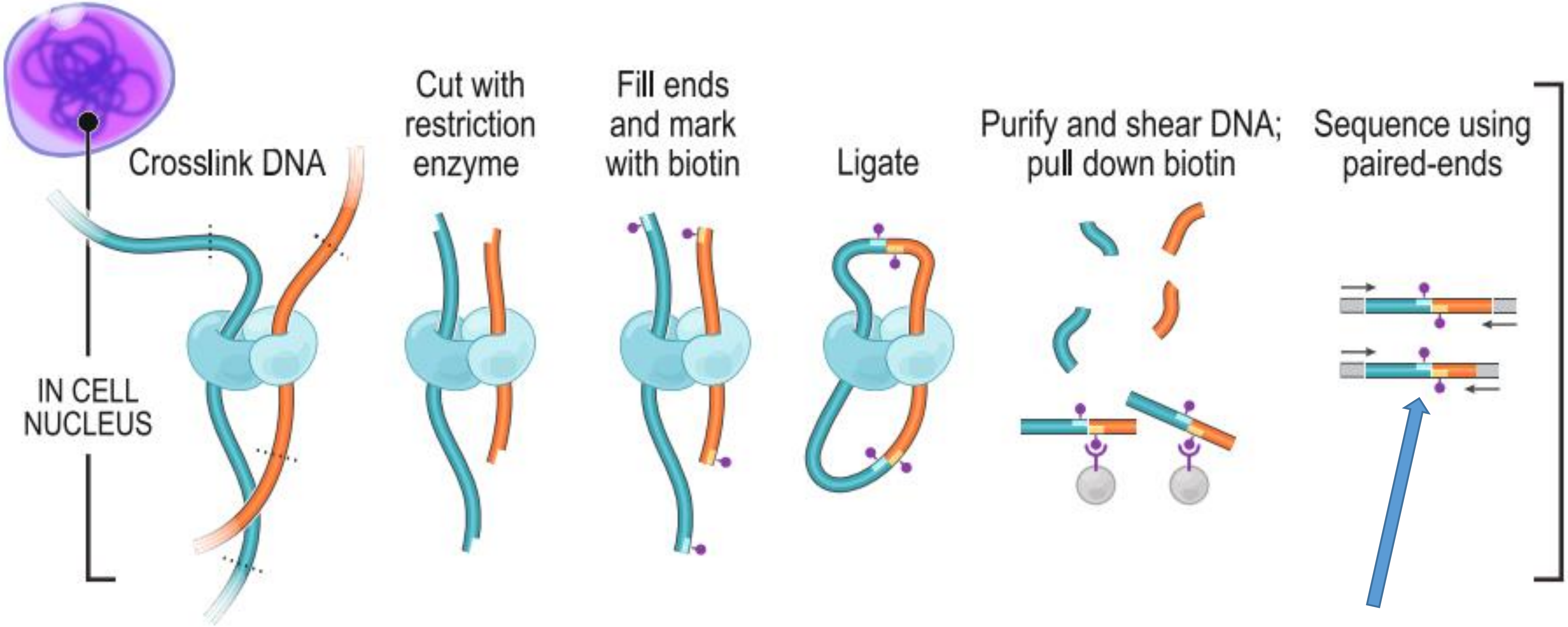
Hi-C

DNA spatial interactions could explain some effects of gwas identified variants, or effects of some regulatory regions linearly spaced from their regulated regions

Notion of **TAD** (Topologicaly Associated Domain)
→ Parts of the genome where DNA is in close proximity



Hi-C


Crosslink DNA
IN CELL NUCLEUS
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Creation of hybrids reads

https://data.4dnucleome.org/experiment-types/in-situ-hi-c/

Hi-C

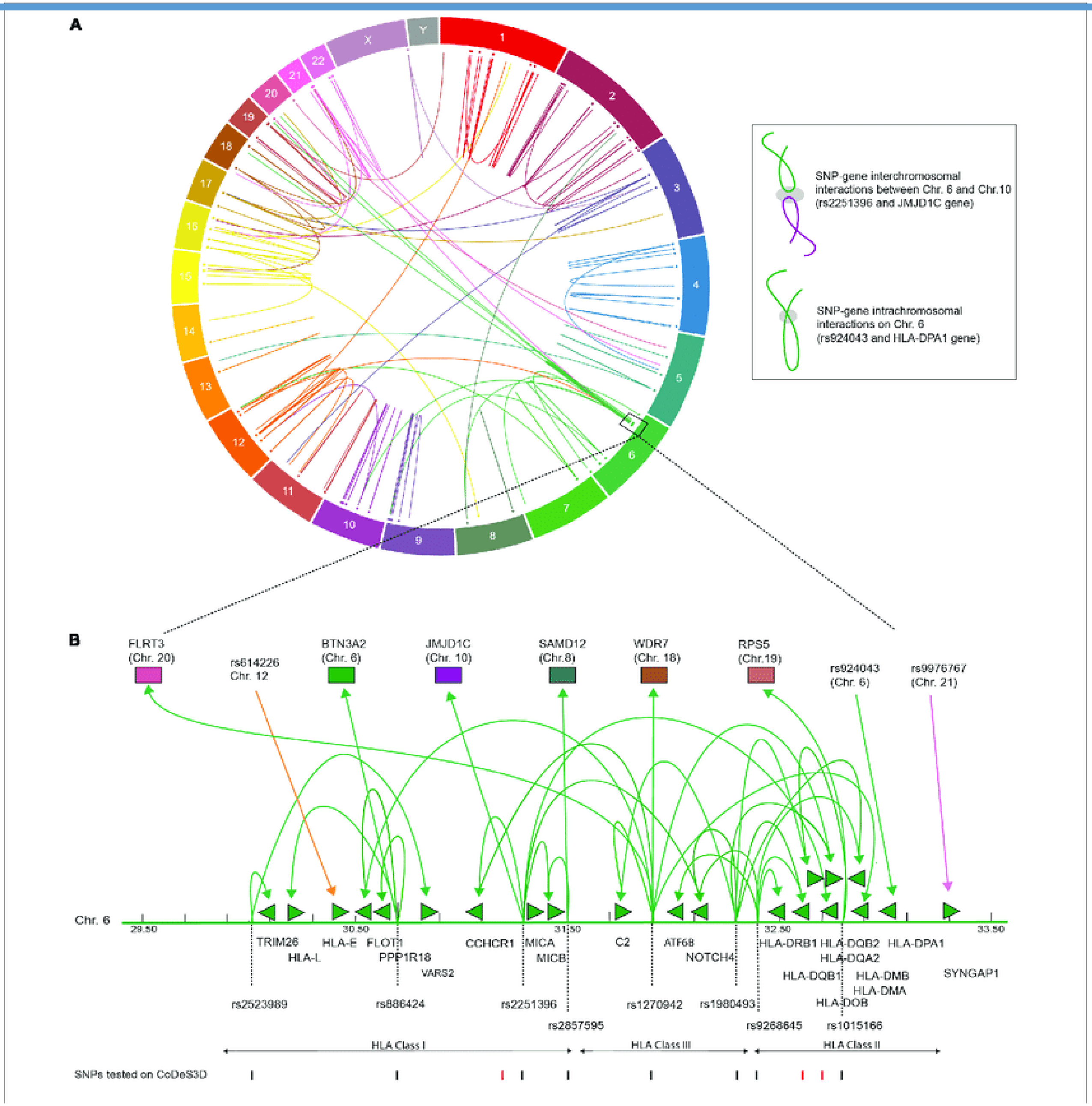
Demultiplexing step identical to DNAseq

Alignment step identical to DNAseq

The two reads from a pair correspond to different parts of the genome

Next step will compare mapping sites of the two parts of each read
(ex software : Hi-C pro)

The results is text file with position of the two reads of a fragment, which can be used to create **maps of chromatin interaction**
(ex software : Circos)

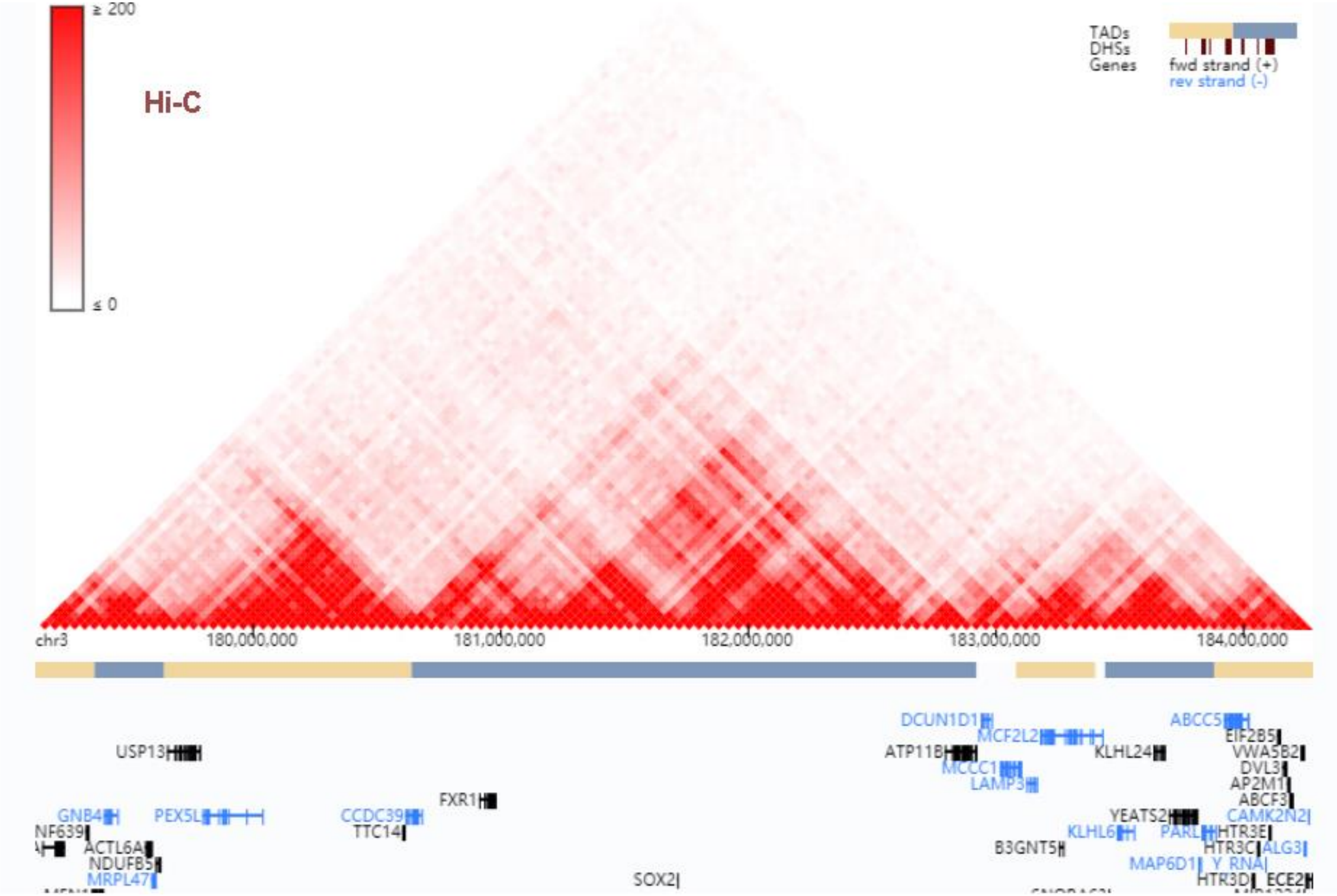


https://www.researchgate.net/figure/T1D-associated-SNPs-form-an-integrated-gene-regulatory-network-A-Circos-plot-showing_fig1_329097399

Hi-C

Other common reprentation

Contact map

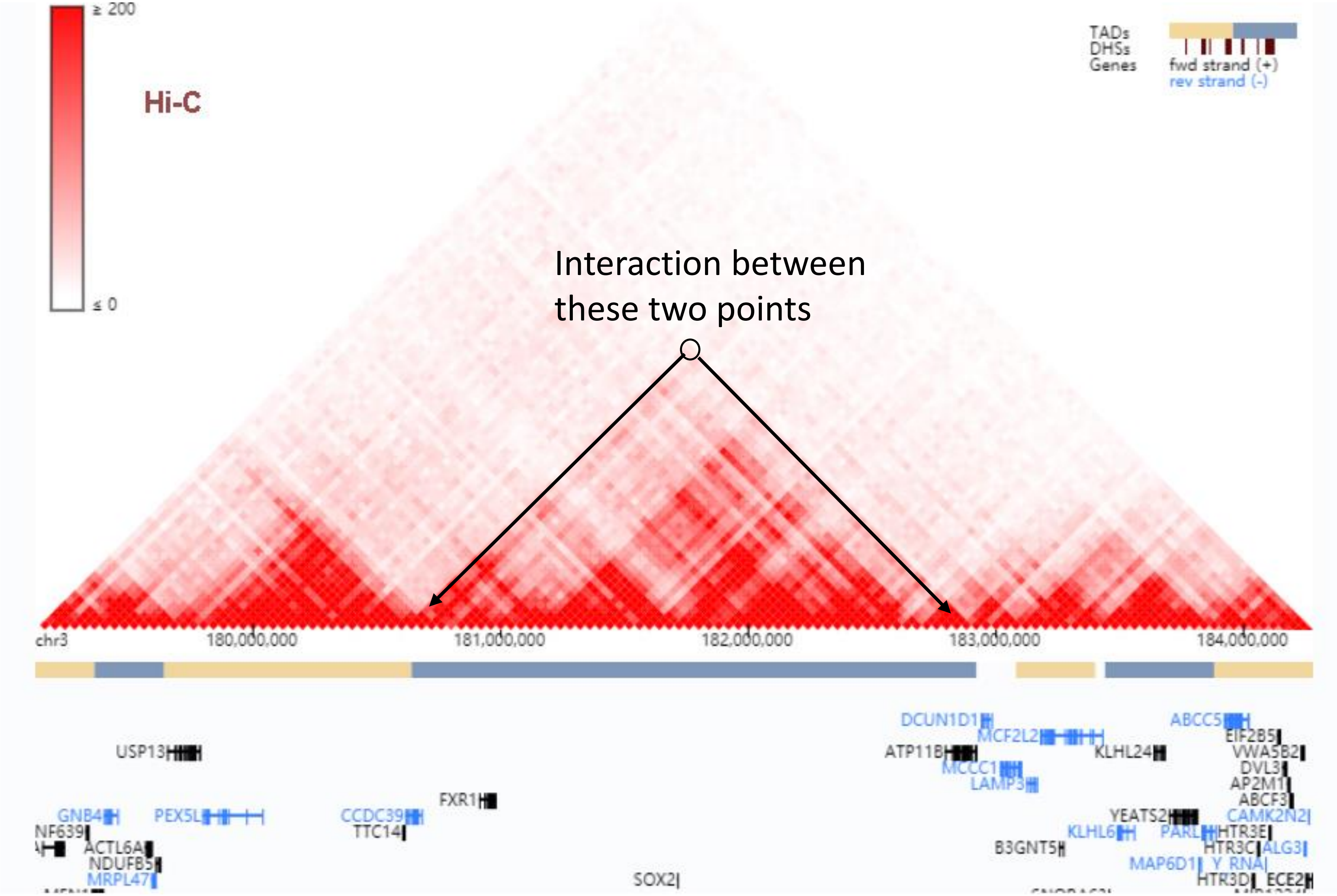


http://promoter.bx.psu.edu/hi-c/

Hi-C

Other common reprentation

Contact map



Interaction between these two points

http://promoter.bx.psu.edu/hi-c/