Bachelor's Degree in Bioinformatics
Course 2023-2024
52115 - Algorithms for sequence analysis in Bioinformatics (ASAB)

ESCI upf.
School of International Studies
BelongBecome

Description

Name of the course: Algorithms for sequence analysis in Bioinformatics (ASAB)
Academic year: **2023/2024**
Year: 2nd
Term: 2nd
Code: 52325

Number of credits: 4 credits

Teaching language: English

Lecturers:

- Arnau Cordomí
- Fernando Cruz

Timetable and Room: Official calendar, regular SIGMA updates

Assessment

NO MIDTERM EXAM
This Year !!

Assessment	Weight	Description
Continuous Assessments	30%	Exercices during theory classes or seminars
Group Project [1]	20%	Group project **shared with** the subject of Clustering Methods and Algorithms in Genomics **(CMAG)**
Final exam [2]	50%	Exam (Theory + Practical) Friday, 22/03/2023 (15:00-17:15) ? **Minimum grade ≥ 4**

[1] **36 students registered in ASAB**

[2] **Retake Exam** - second opportunity **for those doing the Final exam**.

Sessions

- **Punctuality and Respect (e.g. back row)**

- 2 hours with 10 minutes break (2x55 minutes blocks)

- **2 Teachers**
  - Arnau Codormí (coordinator 2nd course)
  - Fernando Cruz

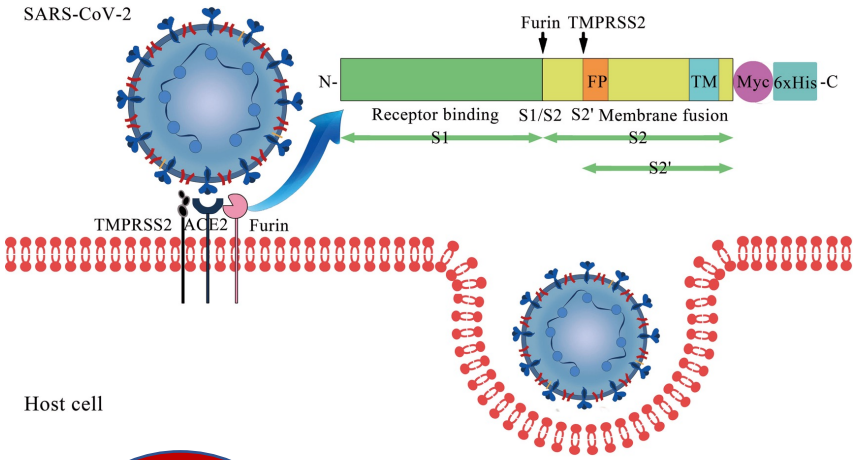
- **36 students ->2 seminar groups**
    - Group 101 (19)
    - Group 102 (17)

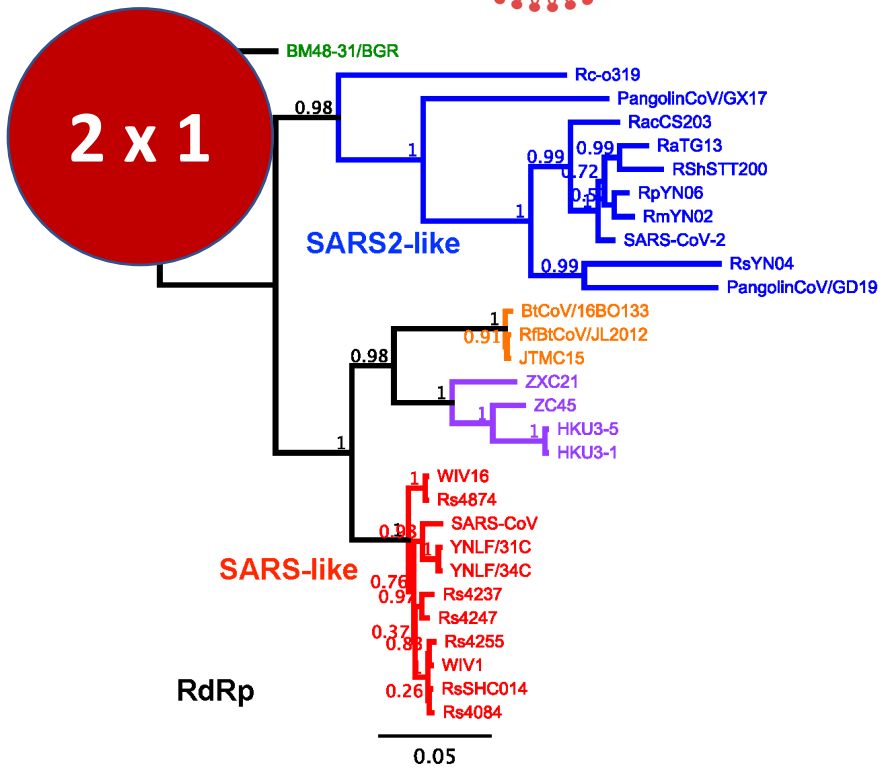
- **Practical part:**
  - Directly **related with Theory**
  - Some sessions combining **Theory & Exercises**
  - Evaluate **Assignments**

- **Assistance and Group Project are Mandatory**

Final Project (Mandatory)

- **Groups of 4-5 students (in principle)**
- **Likely related with Infectious Diseases**
  (e.g. sequencing data & genomics of Covid-19/SARS-Cov2)
- **Possible projects & groups** announced in **February**
- Presentations of **10 minutes + 5-7 minutes** for **questions**
- **Possibility final project shared with CMAG**
- Wednesday, **06/03/2024 (ASAB+CMAG)**
  - **15:00-17:00**
  - **17:30-19:30**
- Exact Details can vary depending on circumstances





Extracurricular activities

- **Participation in Oral Communication Workshop**
  (already started 14th of January 2024!)
- Stregthen your **presentation skills** will be good for the **Group Project**
- **<u>Bonus of 0.5 points</u>** added to individual **Group project qualification**

Theoretical Session 1
Sequencing and Computational Genomics

ESCI upf.
School of International Studies
BelongBecome

Date: 15/01/2024, 15:00-17:00

Room: 61.226

Teacher: **Fernando Cruz** (CNAG)

fernando.cruz@prof.esci.upf.edu

Bachelor's Degree in Bioinformatics
Course 2023-2024
52115 - Algorithms for Sequence Analysis in Bioinformatics (**ASAB**)

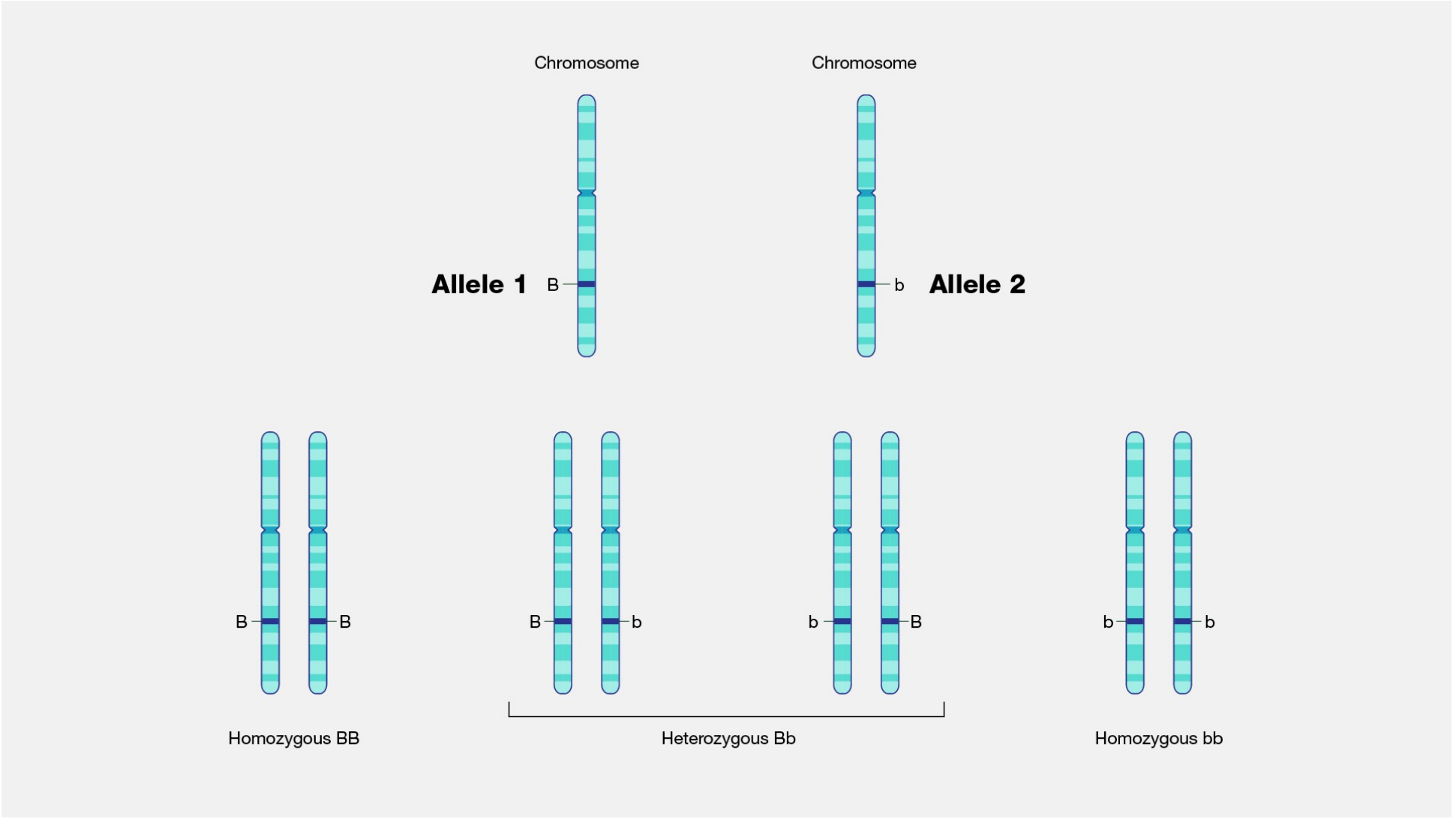
Sequencing and Computational Genomics

1. **Evolution**
2. **Mutations and Genomic Variation**
3. **Massive Parallel Sequencing**
4. **Sequence comparison**
5. **File Formats for Sequencing Data**

1. Evolution

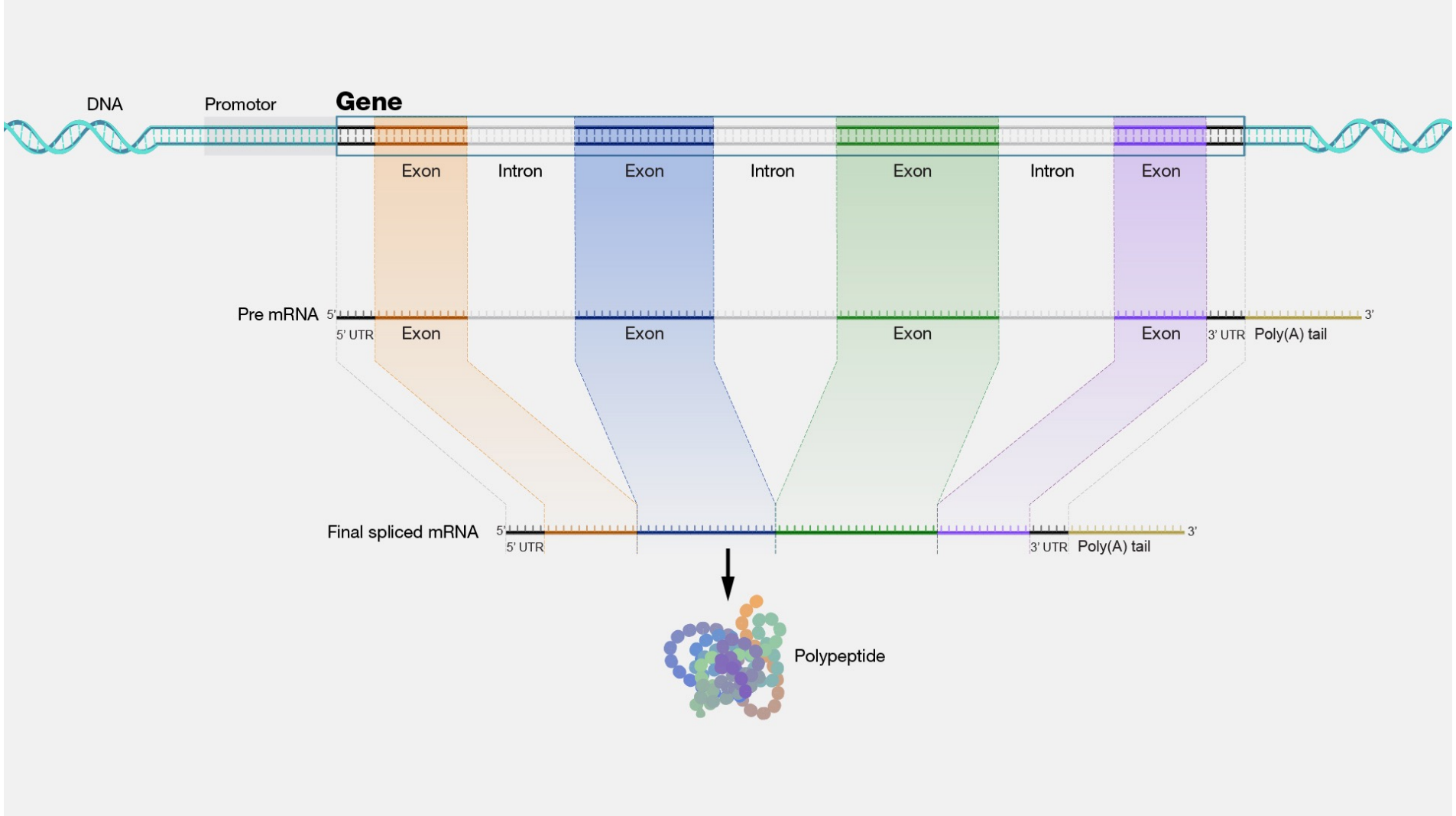
1.1. Introduction: Key Terms in Genetics

Allele: is one of two or more *versions* of DNA sequence (a single base or a segment) at a given genomic location.



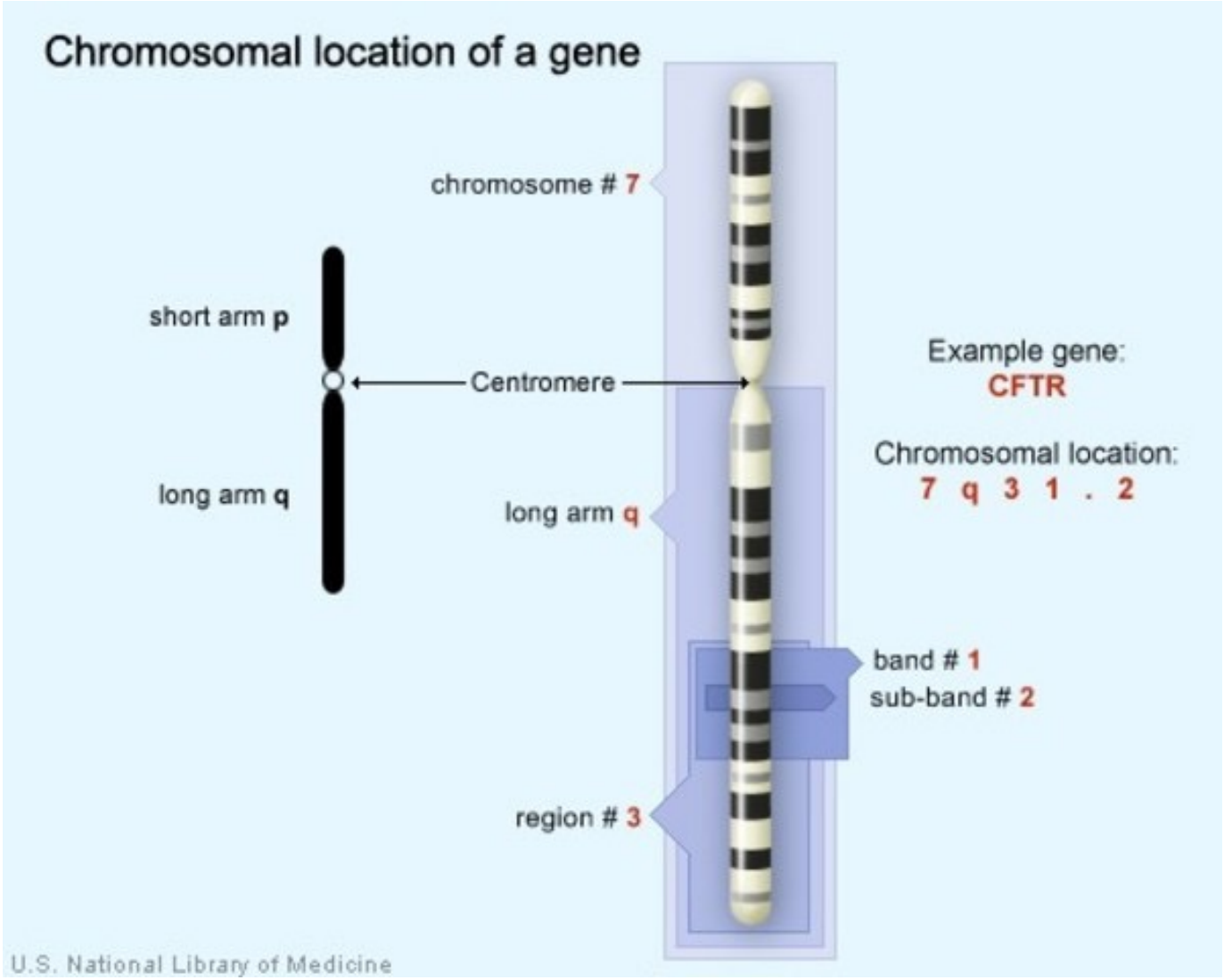
1.1. Introduction: Key Terms in Genetics

Gene: DNA sequences that contain the information needed to specify physical and biological *traits* (protein-coding or RNA genes)



1.1. Introduction: Key Terms in Genetics

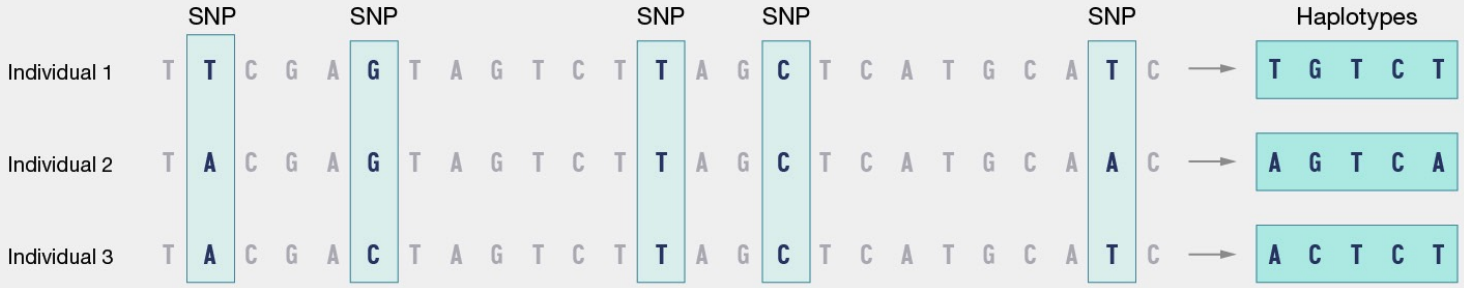
Locus: is the specific *physical location* of a gene or other DNA sequence on a chromosome, like a genetic street address. (pl. Loci)



NCBI Genetics Glossary https://www.genome.gov/genetics-glossary

1.1. Introduction: Key Genetic Terms

Haplotype: A haplotype refers to *a set of DNA variants* along a single chromosome that tend to be inherited together. They tend to be inherited together because they are close to each other on the chromosome

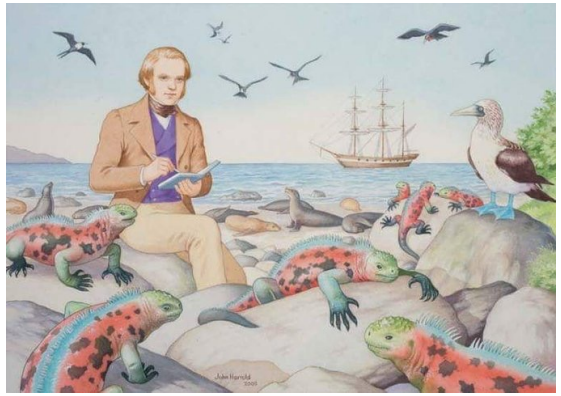


1.1. Introduction: Key Genetic Terms

- **Allele**: is one of two or more *versions* of DNA sequence (a single base or a segment) at a given genomic location. — Variation
- **Gene**: DNA sequences that contain the information needed to specify physical and biological *traits* (protein-coding or RNA genes) — Function
- **Locus**: is the specific *physical location* of a gene or other DNA sequence on a chromosome, like a genetic street address. (pl. Loci) — Location
- **Haplotype**: A haplotype refers to *a set of DNA variants* along a single chromosome that tend to be inherited together. They tend to be inherited together because they are close to each other on the chromosome — Combinations

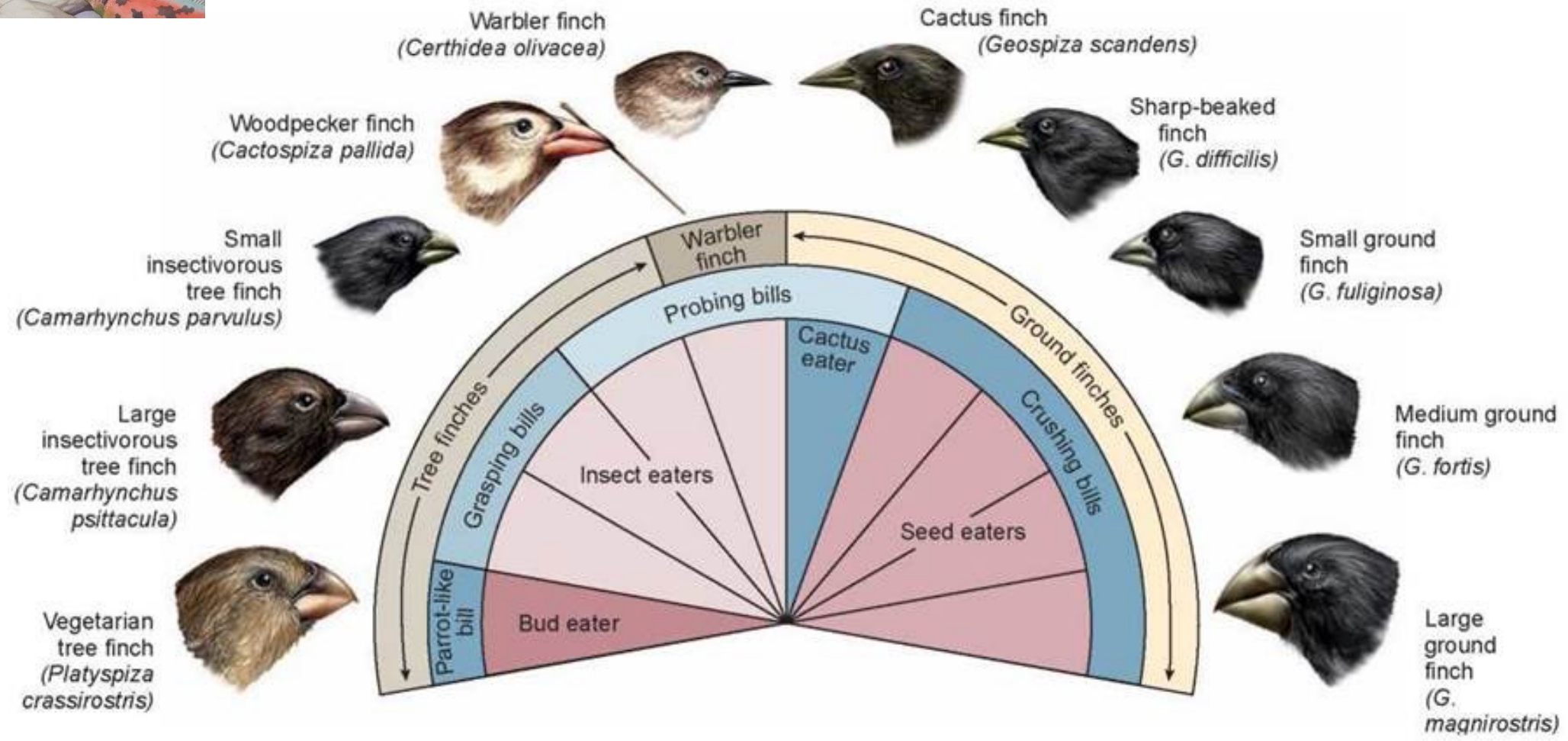
NCBI Genetics Glossary https://www.genome.gov/genetics-glossary

1.2. Evolution: organisms evolve adapting to different environments and situations

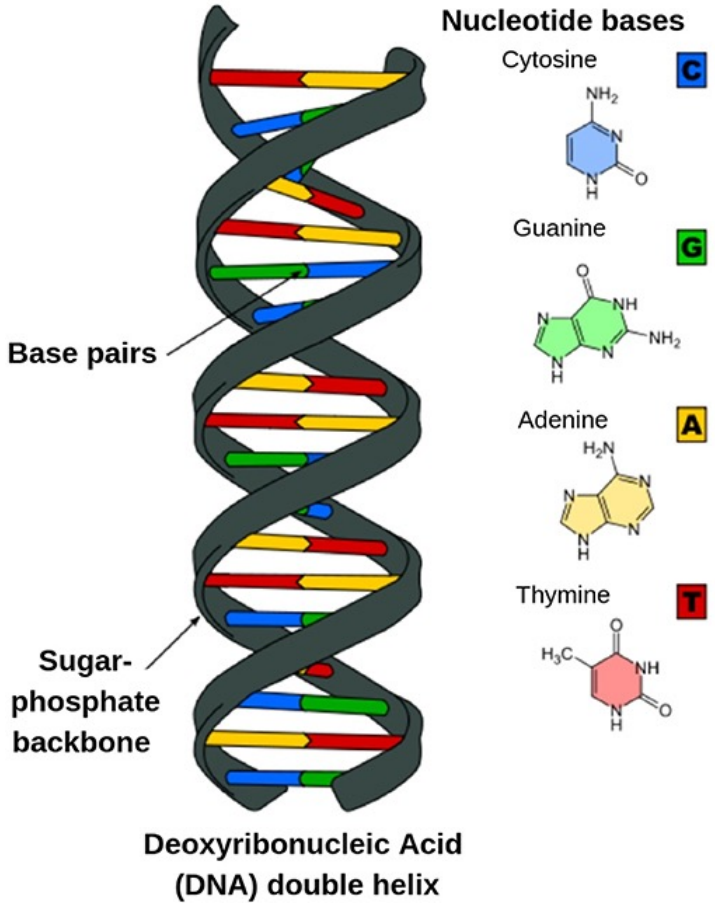


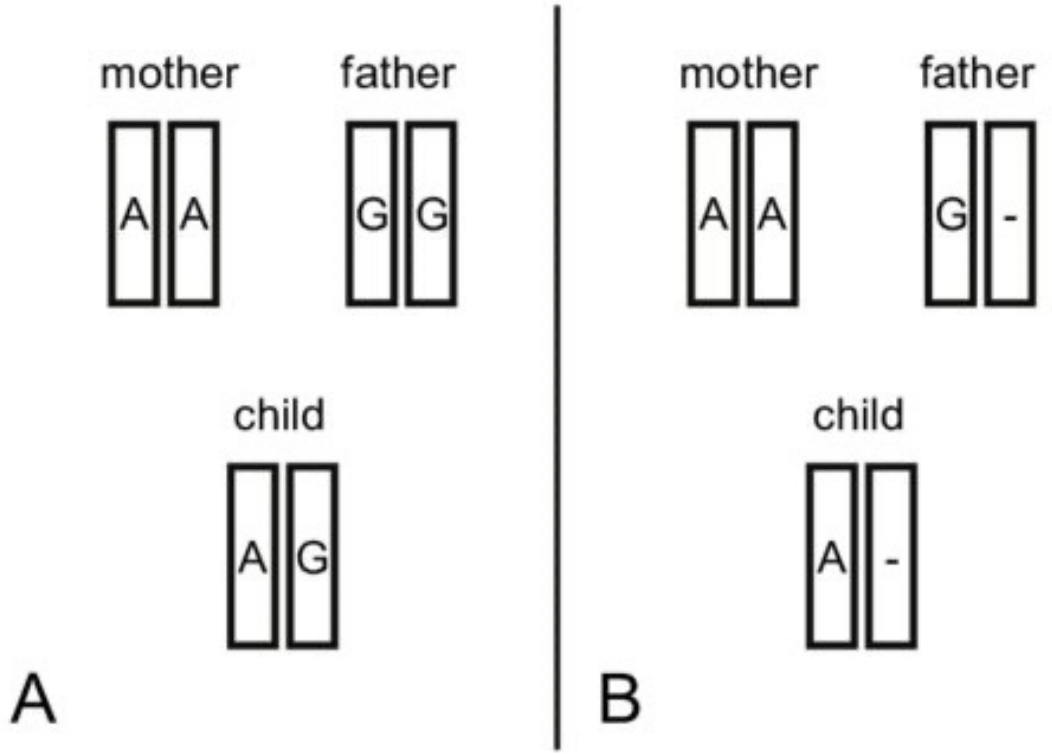
On September 15, 1835, Charles Darwin arrived to the Galapagos Islands

*Darwin realized that **finches** has specialized **traits** that were **"selected" over time** by the **environment** they lived in **and** the **foods** they ate*



1.3. Evolution: changes are recorded at DNA level and should be inheritable

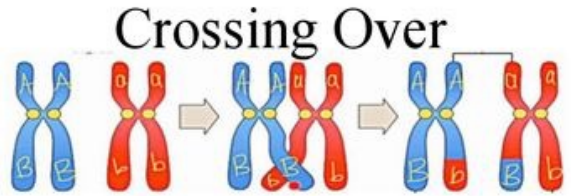




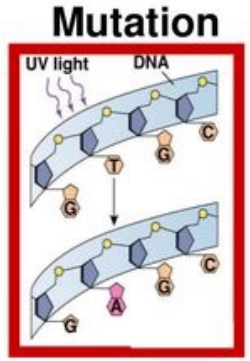
1.4. Evolutionary Forces Act at Population Level

Mechanisms of Evolution

- Recombination
- Mutations
- Natural Selection
- Gene Flow
- Genetic Drift

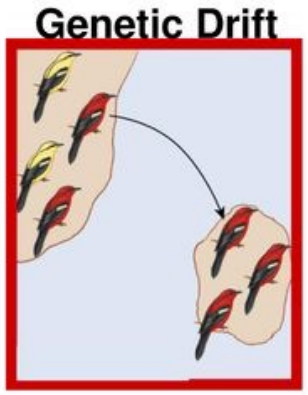


Exchange of genetic material between chromosomes (change combinations)



Substitution of bases generates variation

Variants selected based on their “fitness”



Frequency of variants also varies due to random processes (gamete generation and migration)

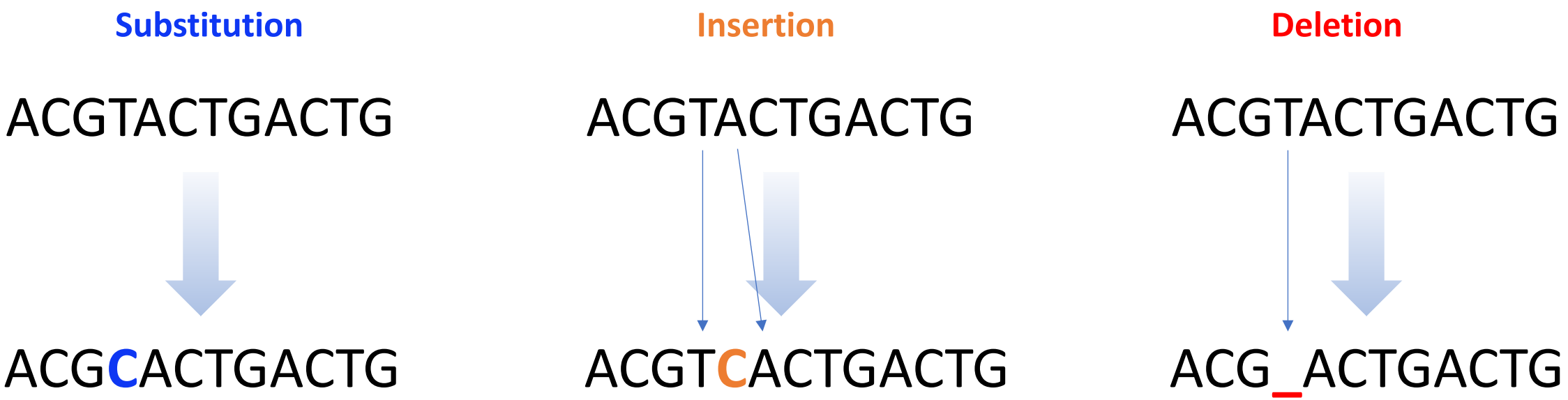
2. Mutations and Genomic Variants

2.1. Point Mutations

Small scale mutations affecting to a single nucleotide.

Most frequent mutations

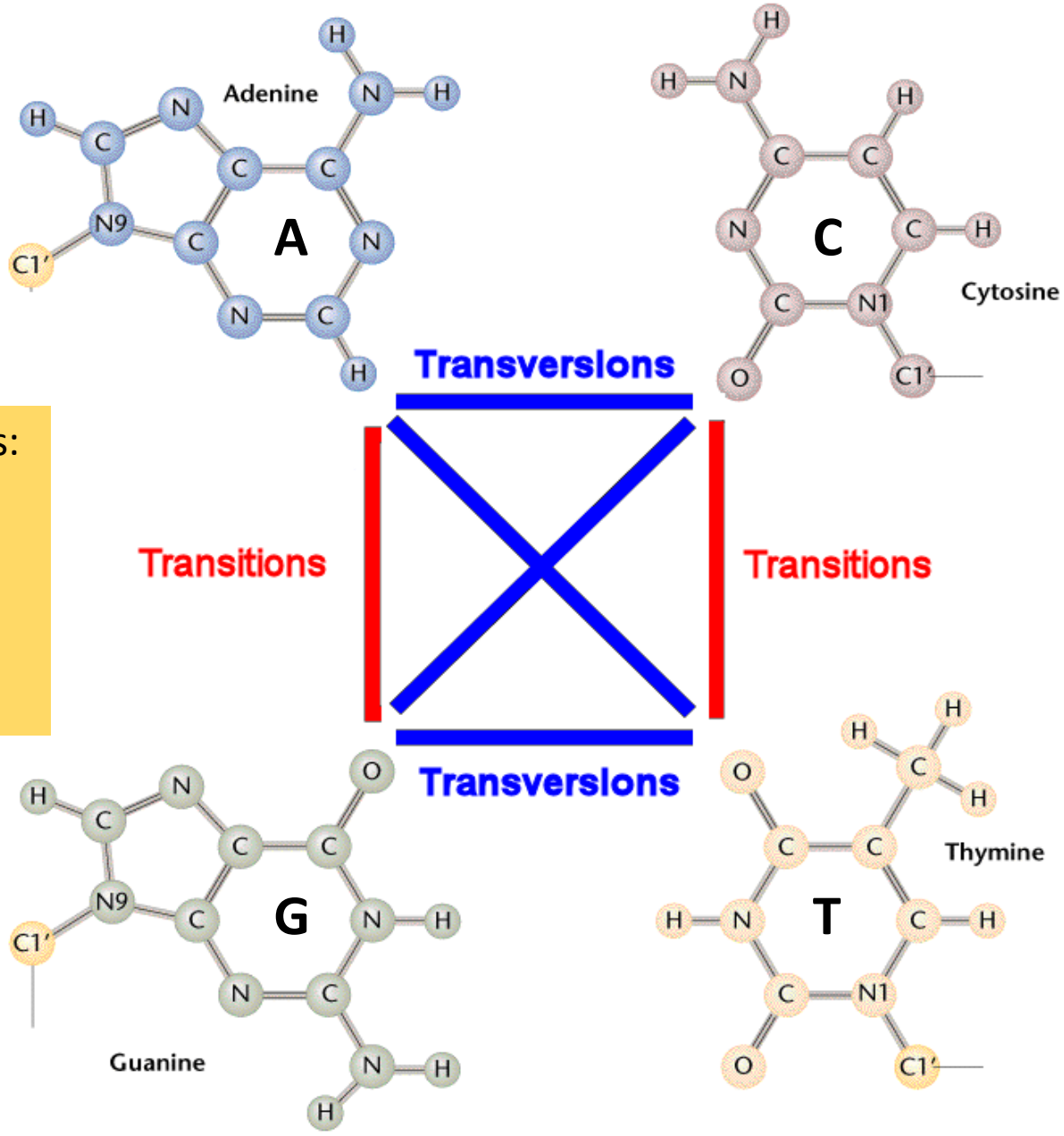
If they consist on a replacement they are known as *substitutions*.



"Substitution" in evolutionary genetics, occurs only when a mutation becomes fixed in a population

"Substitution mutation" is just a point mutation replacing a nucleotide

2.2. Point Mutations: Classified by Nucleotide Change



Mnemonic for purines:

"*Pure* **A**s **G**old" [EN]

"AGUA *Pura"* [ES]

Transitions are **more frequent** than **Transversions**

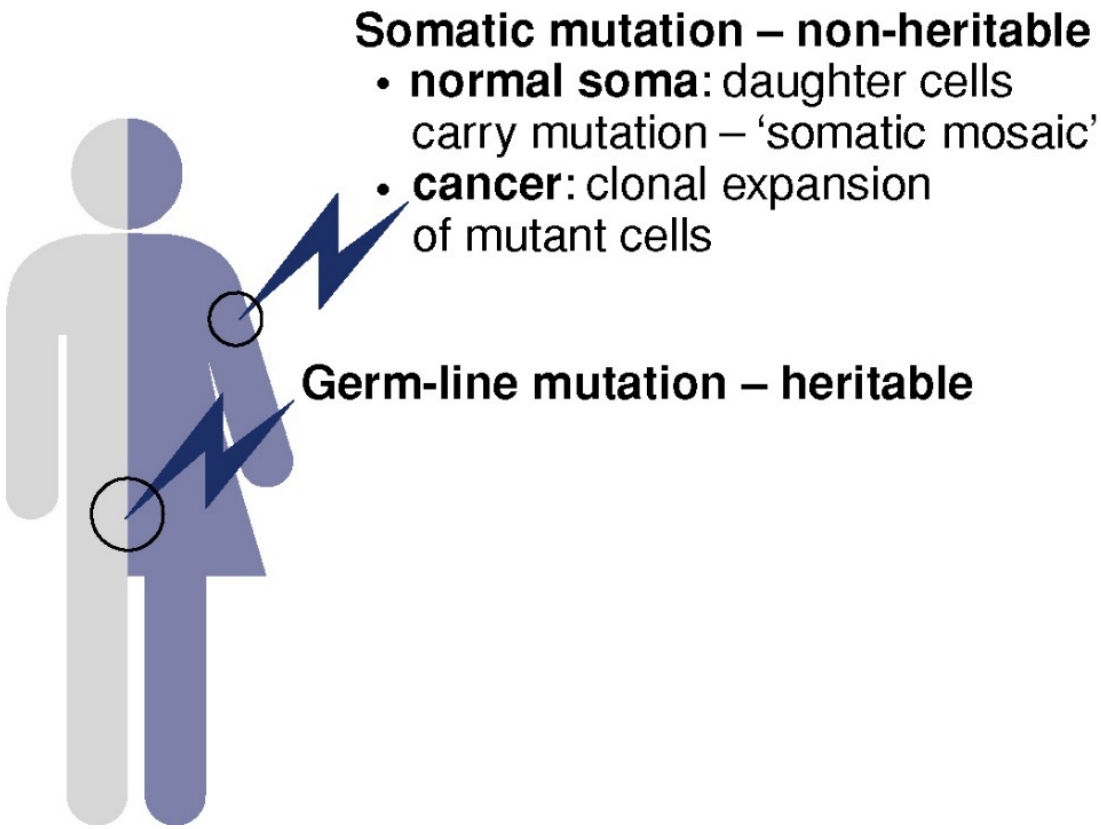
Transition/transversion bias in mammals

- 2 in non coding regions
- 3-5 in protein-coding genes (*)

(* Rosenberg et al 2003
DOI: 10.1093/molbev/msg11)

2.3. Point Mutations: Classified by Target Tissue

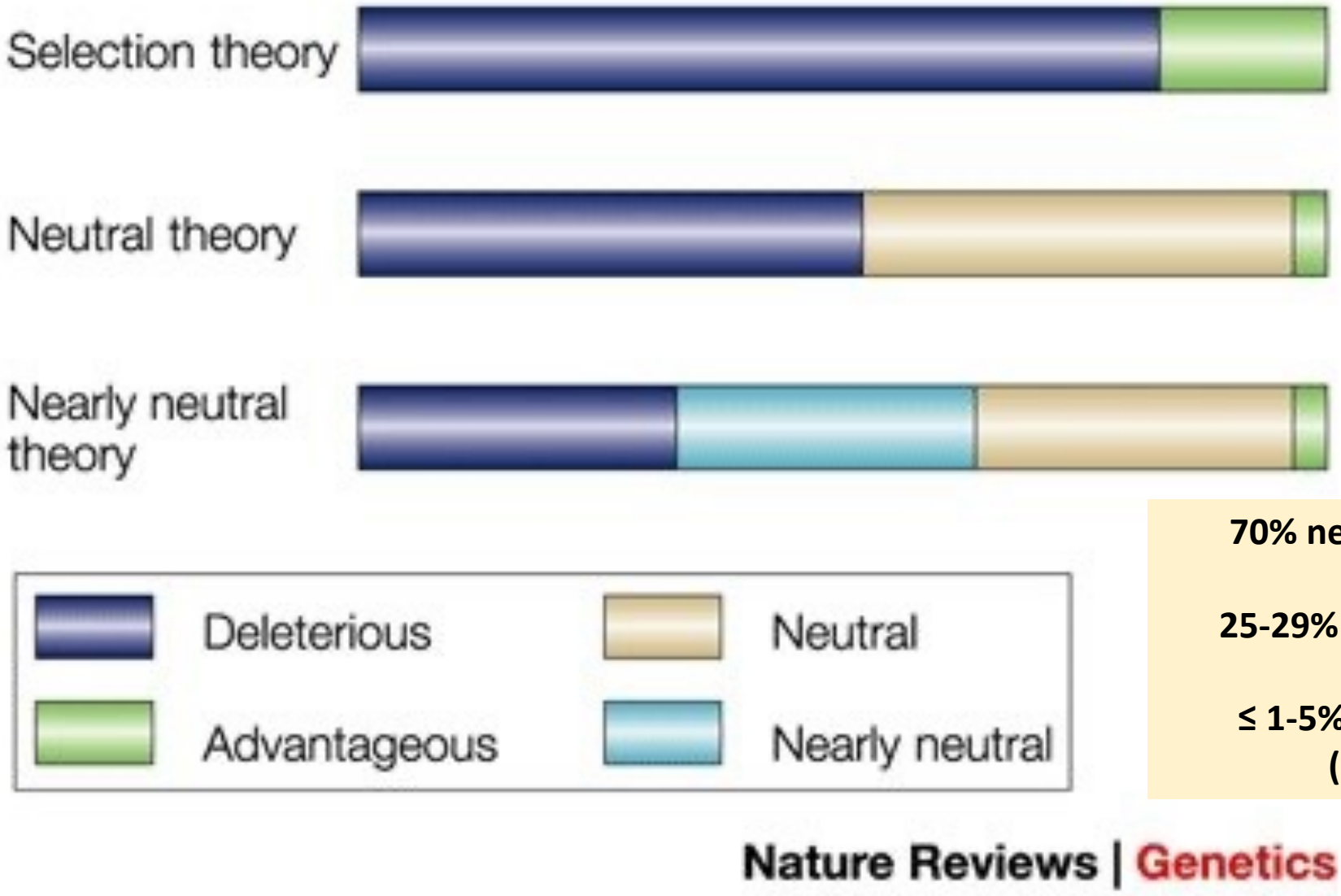
To be inherited they need to occur in the germline, somatic mutation affect just to the cell or tissue were they occur



ACGTACTGACTG

Substitution

ACGGACTGACTG



70% neutral (or nearly)

25-29% negative effects

≤ 1-5% positive effect (adaptive)

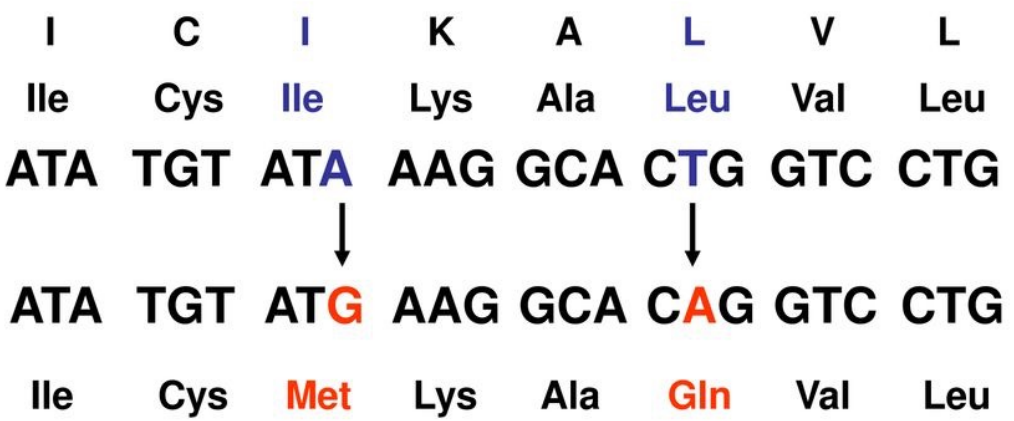
Bromham, L., Penny, D. The modern molecular clock. *Nat Rev Genet* **4,** 216–224 (2003). https://doi.org/10.1038/nrg1020

2.4. Point Mutations in Genes

Synonymous Substitution

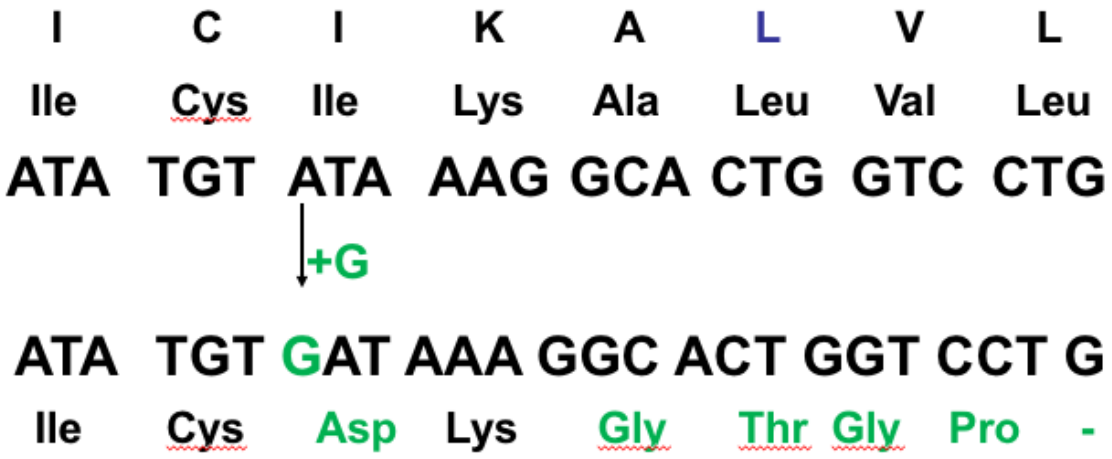
Presumably neutral in effect.

Nonsynomymous Substitution

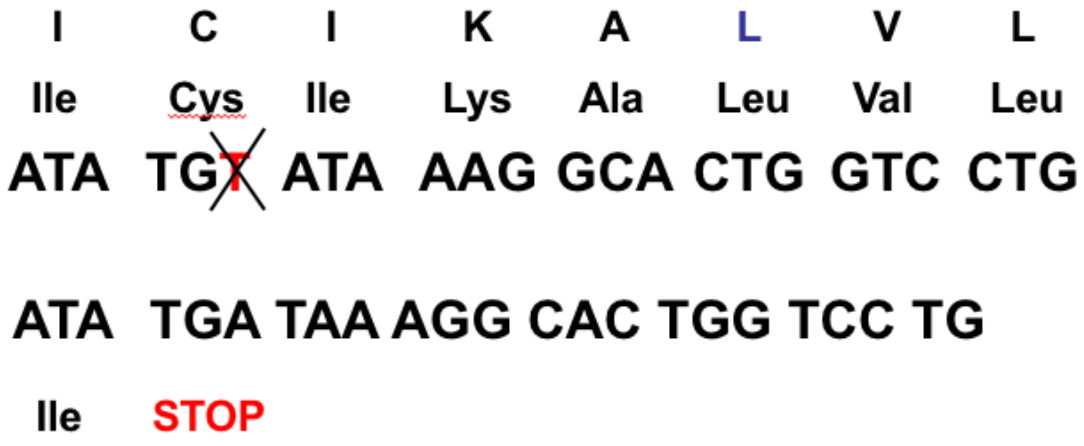


Maybe deleterious or advantageous effect.

Frameshift Mutations



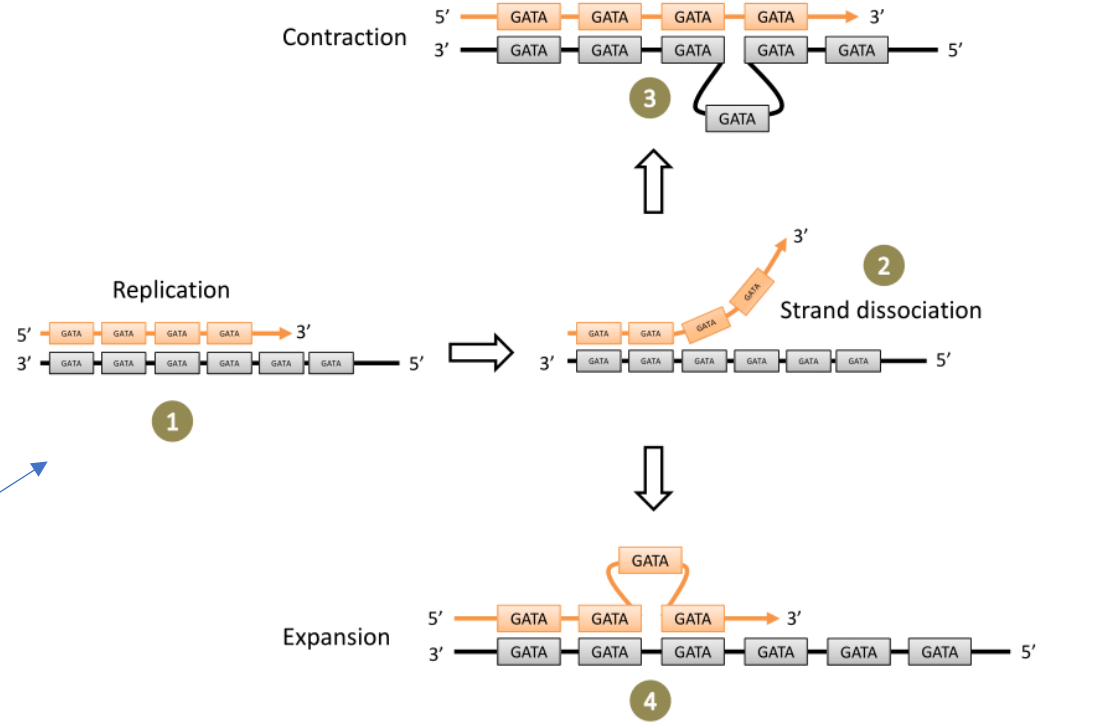
Insertion altering the Protein

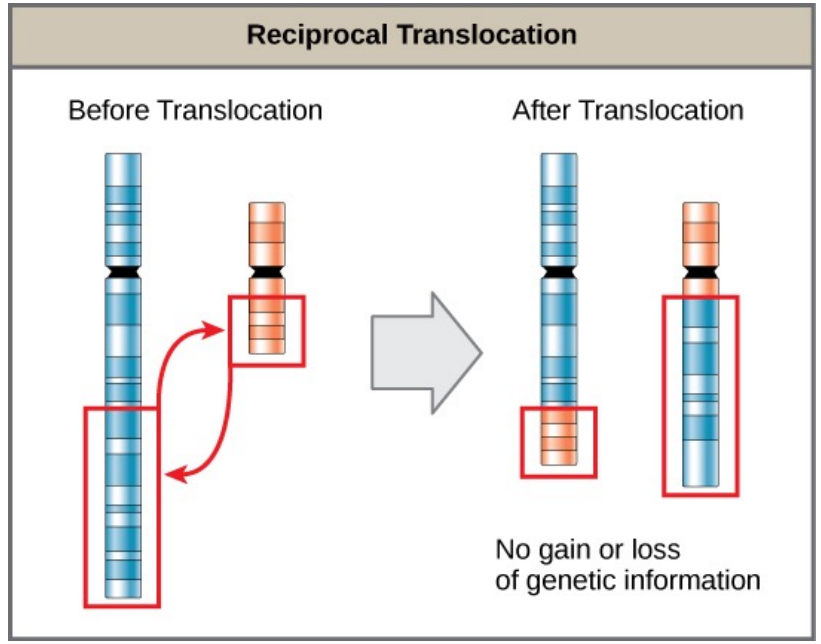


Deletion causing Premature Stop Codon

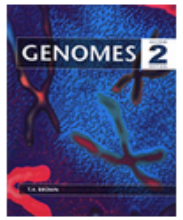
2.5. Other Mutations

- Less frequent than point mutations
- Involving several nucleotides
- Polymerase Slippage – e.g. microsatellites or SSRs
- Non homologous recombinations – e.g. Inversions/Translocations





From: Chapter 14, Mutation, Repair and Recombination



Genomes. 2nd edition.
Brown TA.
Oxford: Wiley-Liss; 2002.

2.6. Genomic Variants

Sequence Variants

- Single Nucleotide Polymorphisms (**SNPs**)
- Insertions/Deletions (**Indels**)

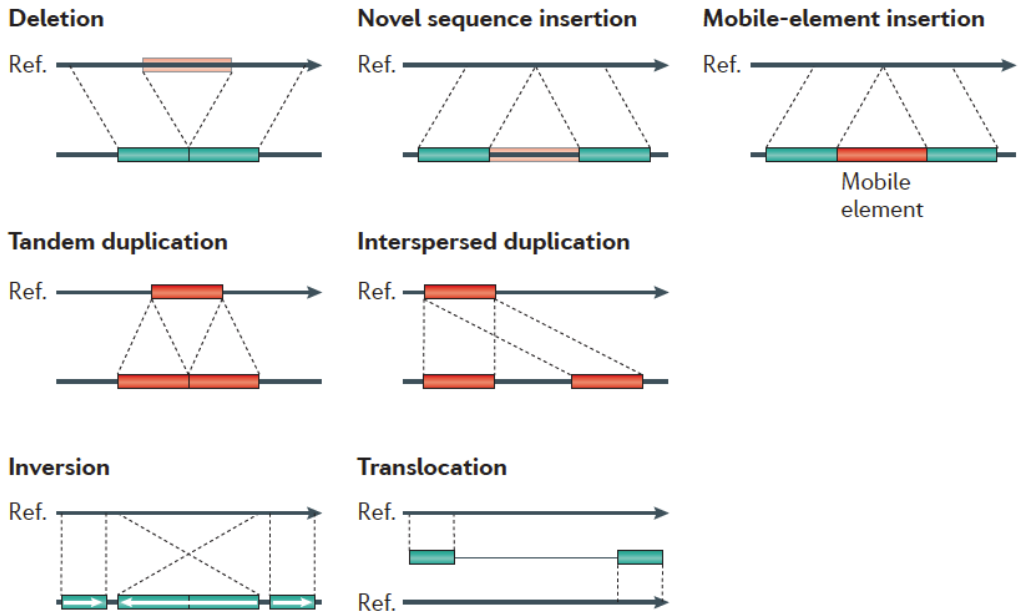
TTATGG
TTGTGG

TTATGG
TT-TGG

<= 50bp in length

Structural Variants

- Segmental Duplications
- Translocations
- Inversions
- Large Indels



> 50bp in length

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-376.

SNP

Sequence Variants

- Single Nucleotide Polymorphisms (SNPs)

TTATGG
TTGTGG

<= 50bp in length

Definition
A single nucleotide polymorphism, or SNP (pronounced "snip"), is a variation at a single position in a DNA sequence among individuals. Recall that the DNA sequence is formed from a chain of four nucleotide bases: A, C, G, and T. If **more than 1%** of a population does not carry the same nucleotide at a specific position in the DNA sequence, then this variation can be classified as a SNP.
http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295

Genotypes
AA Homozygous
AG Heterozygous
GG Homozygous

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-376.

SNV

Sequence Variants

- Single Nucleotide **Variants** (SNVs)

TTATGG
TTGTGG

<= 50bp in length

Definition

A single nucleotide **variant**, is a variation at a single nucleotide position in a DNA sequence **without any limitation on its frequency in the population**. Is a most comprehensive term (includes *de novo* mutations, tumor variants and very low frequency variants…).

All the SNPs are SNVs but not all the SNVs are SNPs!!!

Genotypes

AA Homozygous
AG Heterozygous
GG Homozygous

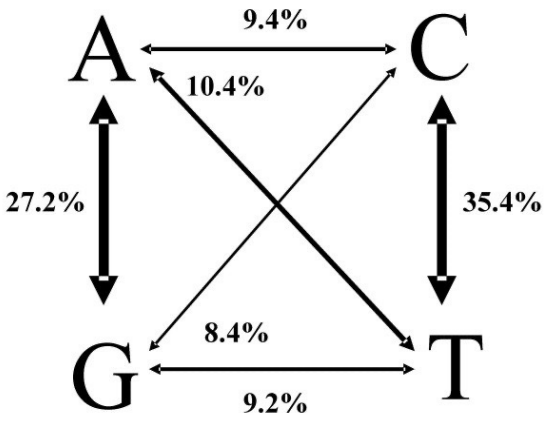
Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-376.

KEEP IN MIND

All of this (the kind of mutation, their location in the genome, their effect on fitness…) will determine:

- **Substitution rates**
- **Bioinformatic models** and values to build subtitution **matrices**

Substitution matrices are usually seen in the context of **amino acid** or **DNA sequence alignments**.

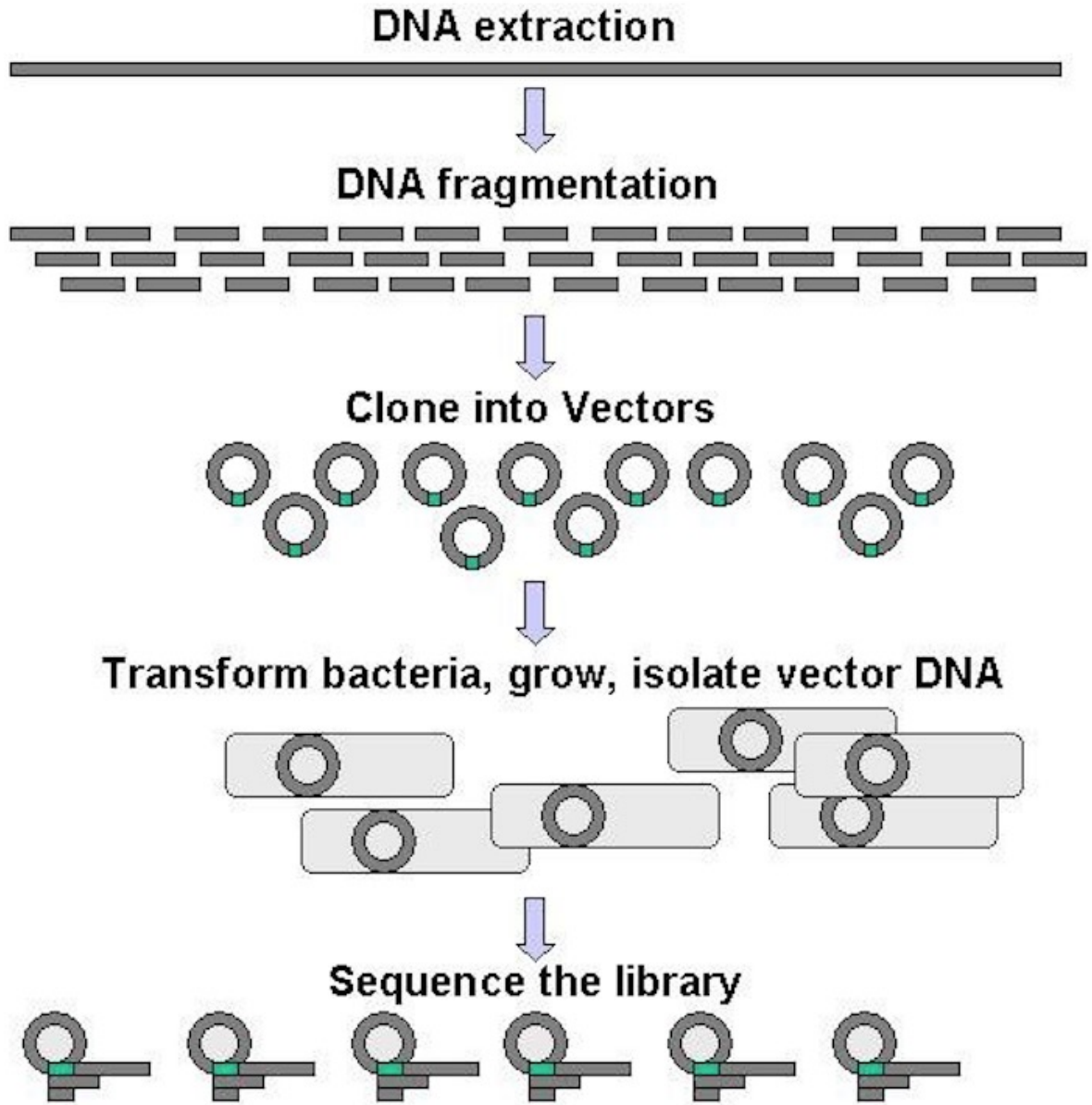


SNP data *A. funestus* 2007
(https://doi.org/10.1186/1471-2164-8-5)

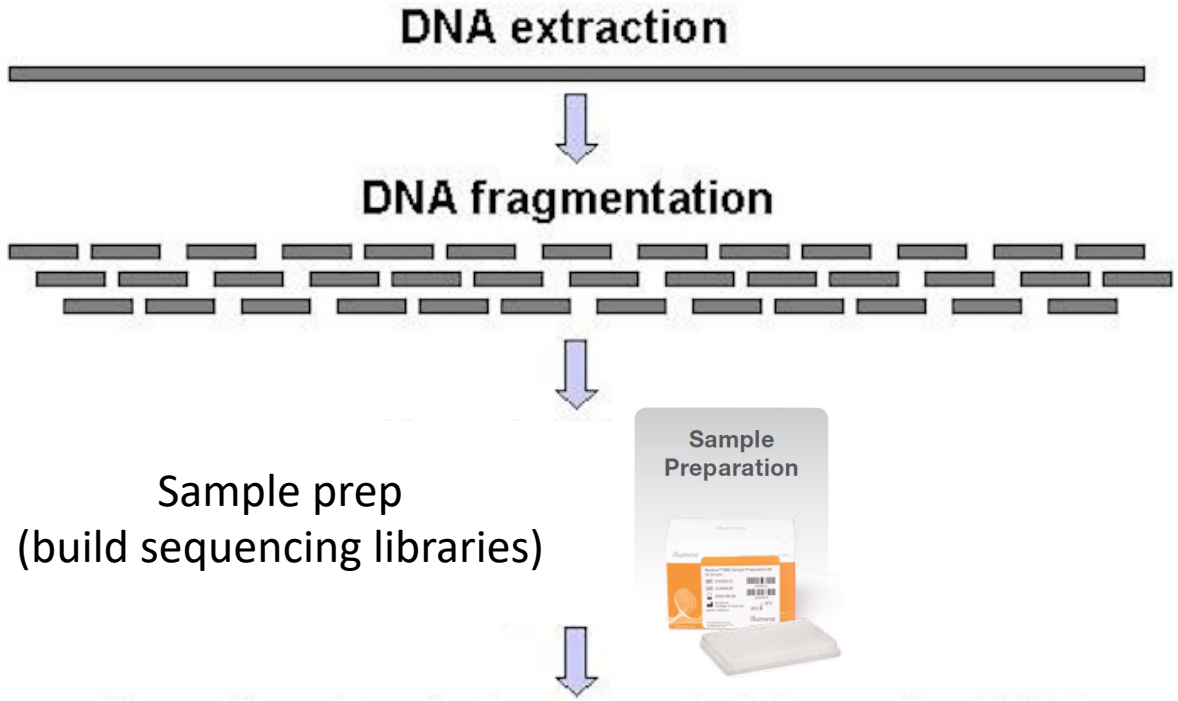
	A	C	T	G
A	-	0.094	0.104	0.272
C	0.094	-	0.354	0.084
T	0.104	0.354	-	0.092
G	0.272	0.084	0.092	-

3. Sequencing Technologies

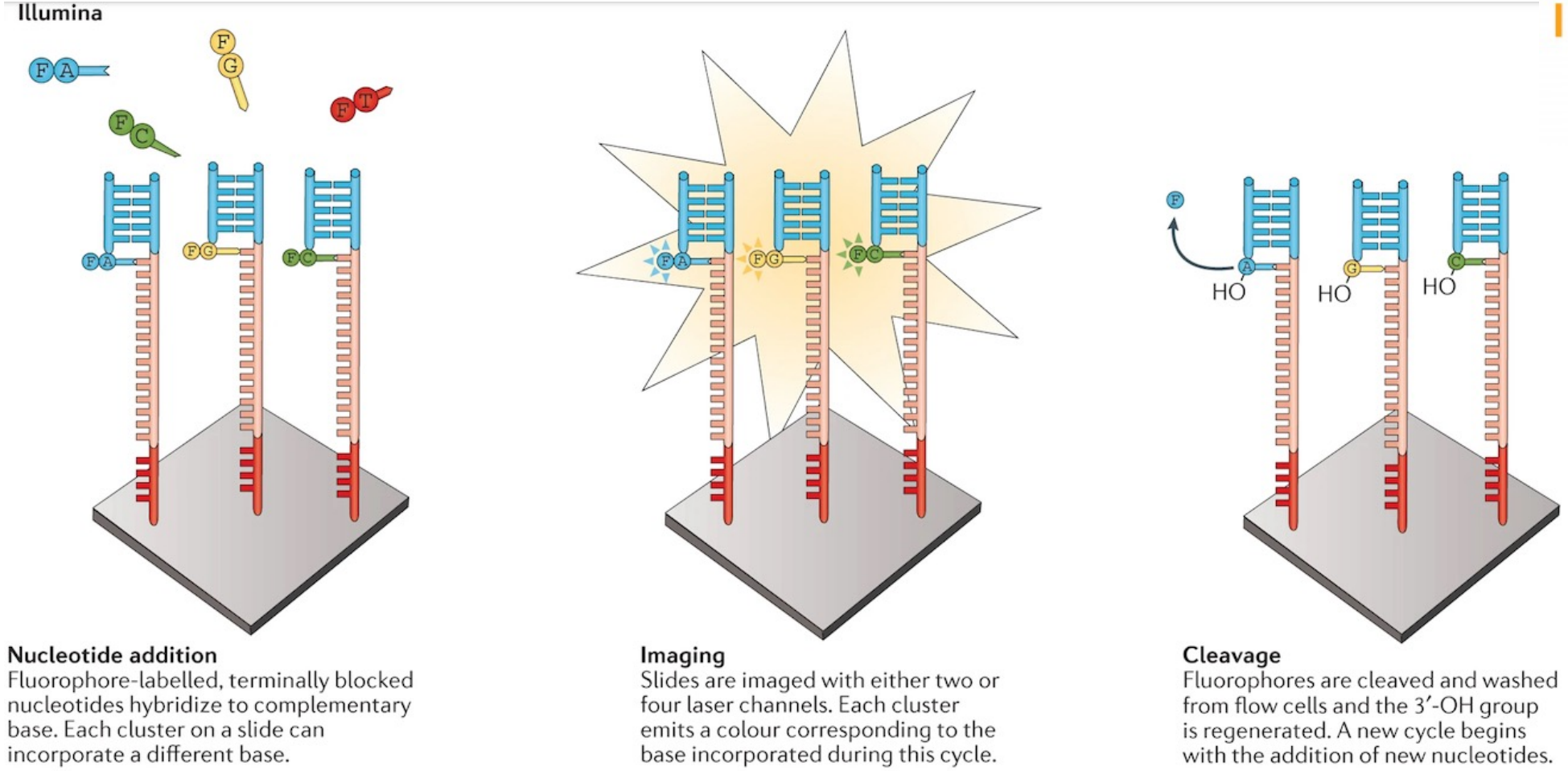
Whole Genome Shotgun (WGS)



Whole Genome Shotgun (WGS)



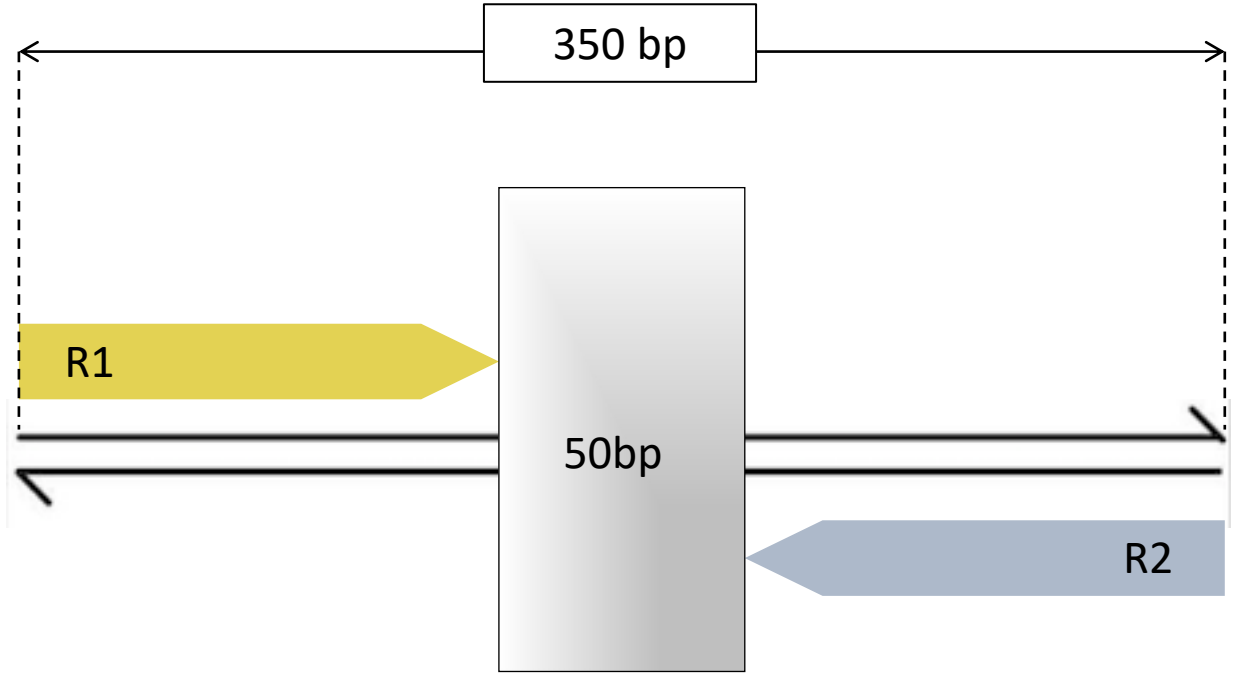
Illumina - Short Read Sequencing



Sequencing by synthesis Method

Goodwin, S., McPherson, J. D., and McCombie, W. R. 2016, Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics.*

Paired End (PE) Reads Illumina



Short Reads, usually 150 bp

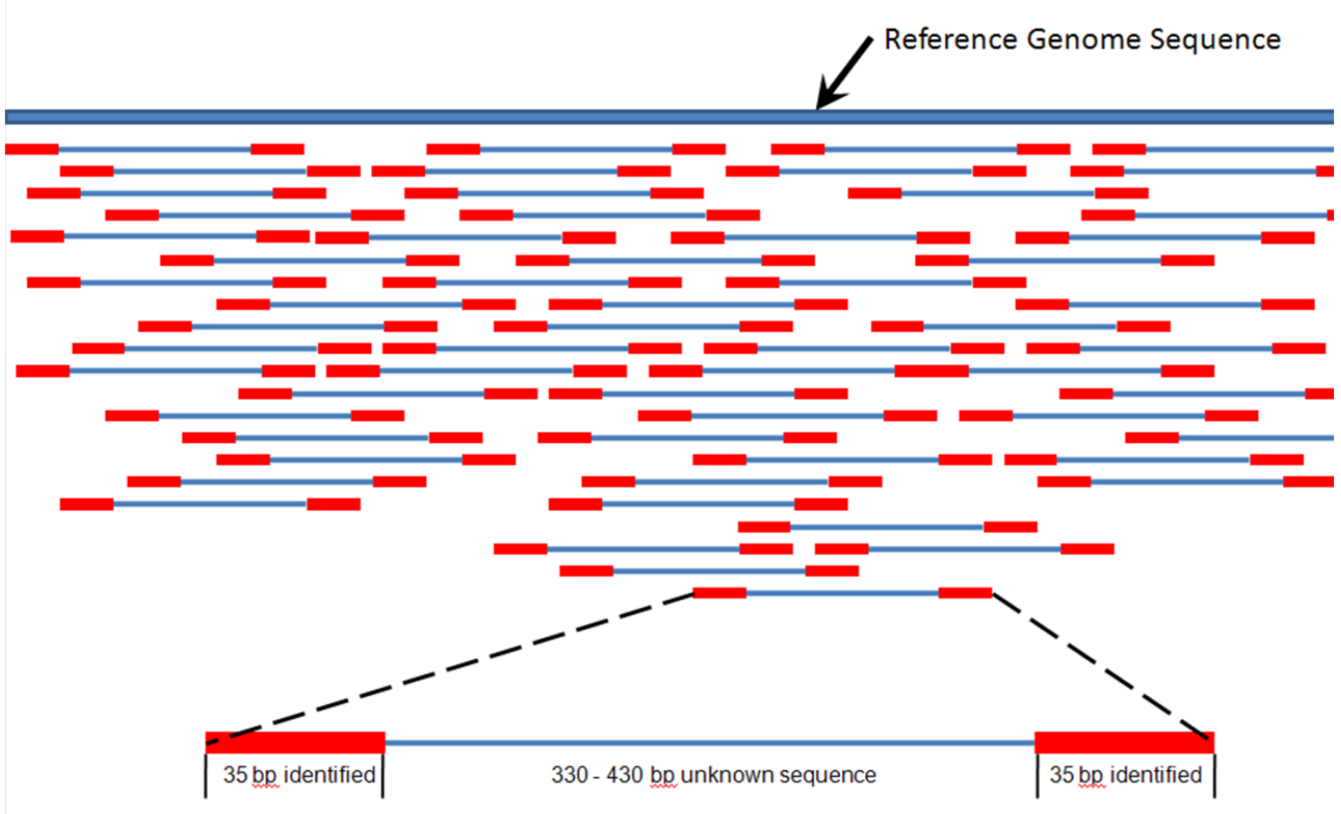
Fidelity depends on **Error Rate**

Genomic fragments (DNA template) usually 250-800 bp

Fragment and Insert Size:
http://thegenomefactory.blogspot.com.es/2013/08/paired-end-read-confusion-library.html

Illumina Sequencing Data

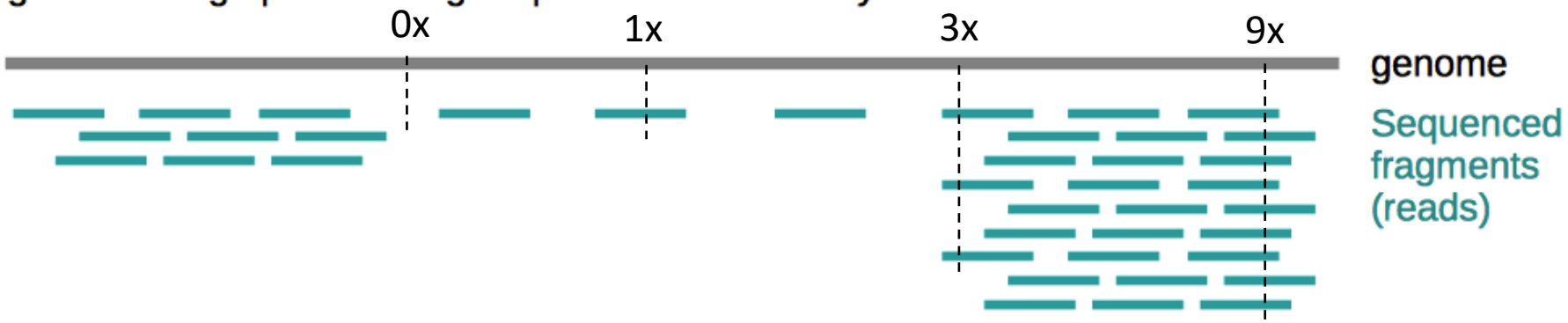
- >300 Million Reads from different parts of the genome
- Yield, the total base pairs produced, is high normally ≥ 30 Gb



The coverage

represents the number of times a base of the sample genome (or target region) is read during sequencing.

A higher coverage provides higher power for data analysis.



Coverage = Total base pairs/Genome Length

Illumina "HiSeq" Instruments

Product	HiSeq 2500		HiSeq 3000	HiSeq 4000	HiSeq X Five†	HiSeq X Ten†
Description	Power and efficiency for large-scale genomics		Maximum throughput and lowest cost for production-scale genomics		Maximum throughput and lowest cost for population- and production-scale human WGS	
Key methods	Production-scale genome, exome, transcriptome sequencing, and more				Population-scale human WGS	
Run mode	Rapid run	High-output	—	—	—	—
Flow cells processed per run	1 or 2	1 or 2	1	1 or 2	1 or 2	1 or 2
Output range	10–300 Gb	50–1000 Gb	125–750 Gb	125–1500 Gb	900–1800 Gb	900–1800 Gb
Run time	7–60 hours	< 1–6 days	< 1–3.5 days	< 1–3.5 days	< 3 days	< 3 days
Reads per flow cell‡	300 million	2 billion	2.5 billion	2.5 billion	3 billion	3 billion
Maximum read length	2 × 250 bp	2 × 125 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Human Genome Coverage	94x	312x	235x	468x	562x	562x

Adapted from:

https://www.illumina.com/

http://www.molecularecologist.com/next-gen-fieldguide-2014/

Base Calling Errors

Illumina Typically 0.05-0.1%

REF ATGGTTTTTTTTTTTGGCTCTGCTTGTTGGCCCTT**A**TGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTGTCGGGATTGTGA

```
1                    GGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
2                      CTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGT
3                  TTTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTA
4                      TCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
5            TTTTTTTTTTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAAC
6                    GCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTG
7                    GGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
8                      TCTGCTTGTTGGCCCTTGTGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTG
9                  TTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAA
10                         GTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTGTCGGGAT
```

Error Rate = 1 error / (100bp x 10 reads) = 0.001

Base Calling Errors

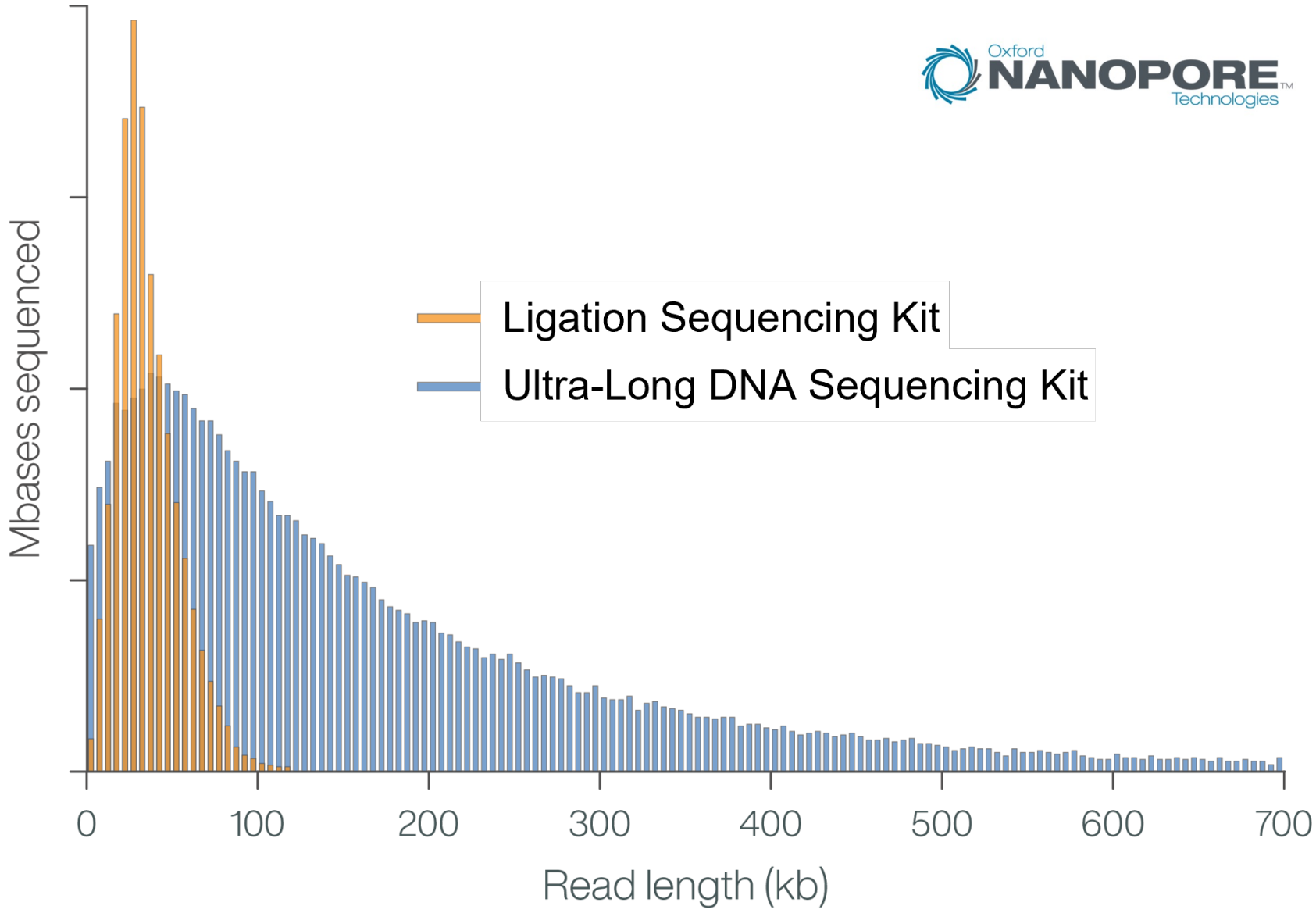
Illumina Typically 0.05-0.1%

REF ATGGTTTTTTTTTTTGGCTCTGCTTGTTGGCCCTT**A**TGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTGTCGGGATTGTGA

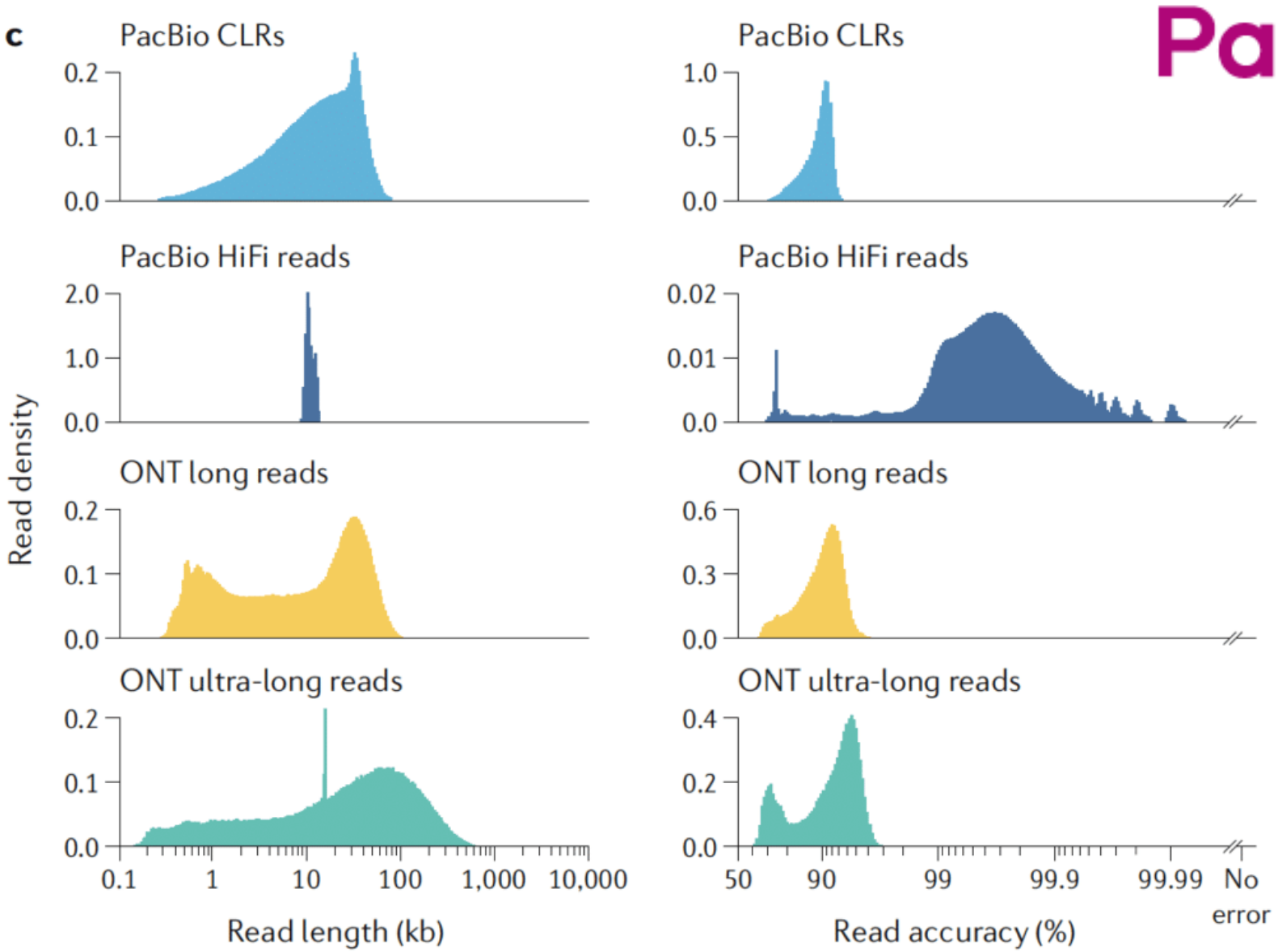
```
1                    GGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
2                      CTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGT
3                  TTTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTA
4                      TCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
5          TTTTTTTTTTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAAC
6                    GCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTG
7                    GGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
8                      TCTGCTTGTTGGCCCTTGTGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTG
9                  TTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAA
10                           GTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTGTCGGGAT
11                   GGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
12                     CTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGT
13                 TTTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTA
14                     TCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
15             TTTTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCT
16                   GCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTG
17                   GGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGT
18                     TCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTG
19                 TTGGCTCTGCTTGTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAA
20                           GTTGGCCCTTATGGCTCAACATTATTCAATCATTAATATTTACGGCTATTAGTCCGAGTATTATCACAACGGGCCGAAACCGGCCTAAGTGTGTCGGGAT
```

Error Rate = 1 error / (100bp x 20 reads) = 0.0005

Long Read Sequencing Technologies

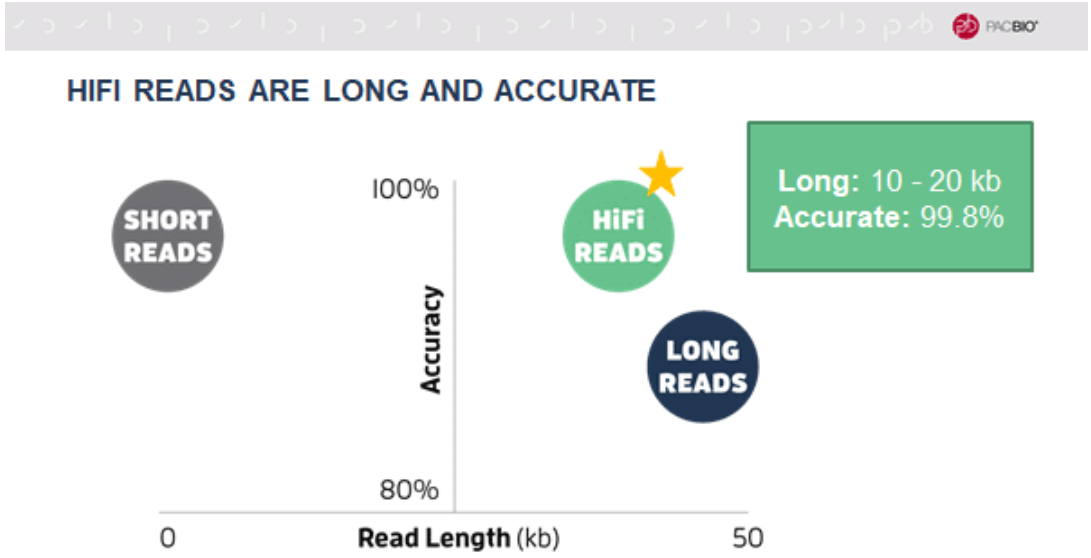


Long Read Sequencing Technologies

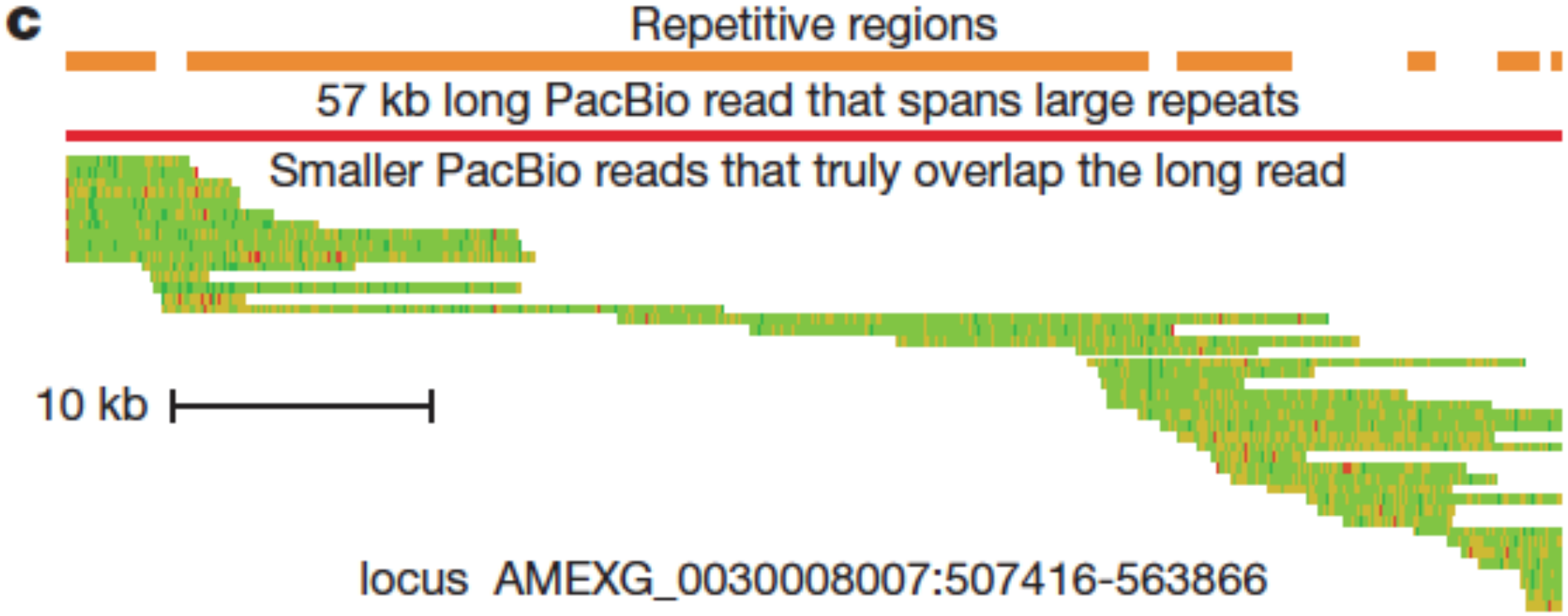


Long Read Sequencing Technologies

- Average **Read Lengths 15Kb**
- **Error rates 10% -** Pacbio CLR and Oxford Nanopore Technologies (ONT)
- **ONT Q20** 1% errros 99% Accuracy
- **Pacbio Hifi** Q=30 ~ 0.1% Error, 99.9% Accuracy



Long Reads: Spanning Long Repetitive Regions



Nowoshilow S, Schloissnig S, Fei J-F, Dahl A, Pang AWC, Pippel M, Winkler S, Hastie AR, Young G, Roscito JG, et al: **The axolotl genome and the evolution of key tissue formation regulators.** *Nature* 2018, **554:**50.

HiC - chromosome conformation capture technologies

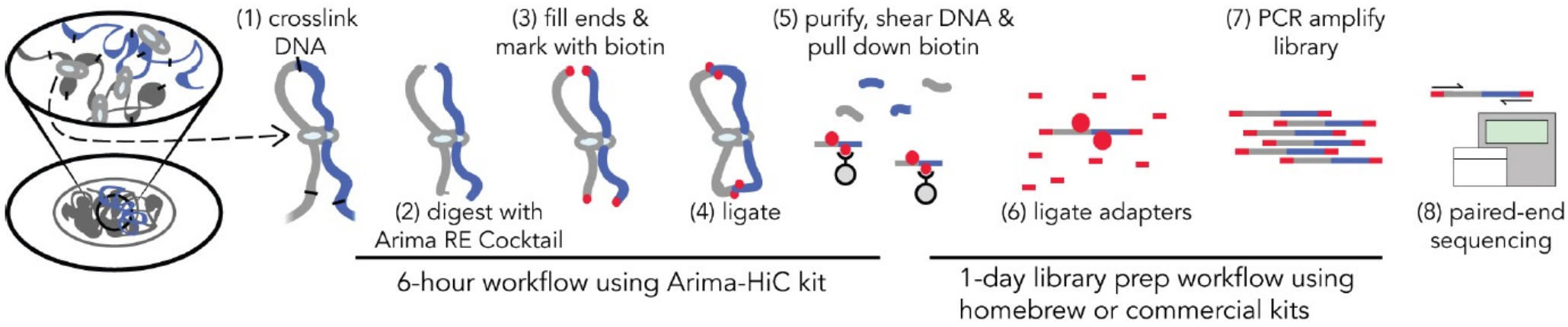
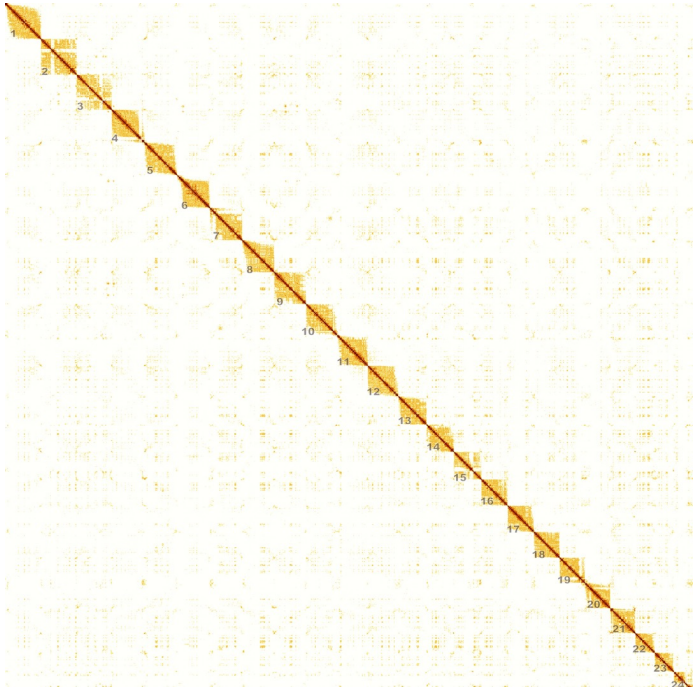


Figure 2: The Arima-HiC workflow results in ligated and biotinylated DNA that is PCR-amplified and prepared as a library using a multitude of library prep kits with appropriate adapters for paired-end Illumina sequencing.



Raor 2n=48

Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, et al. 2009, Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, **326**, 289–93.

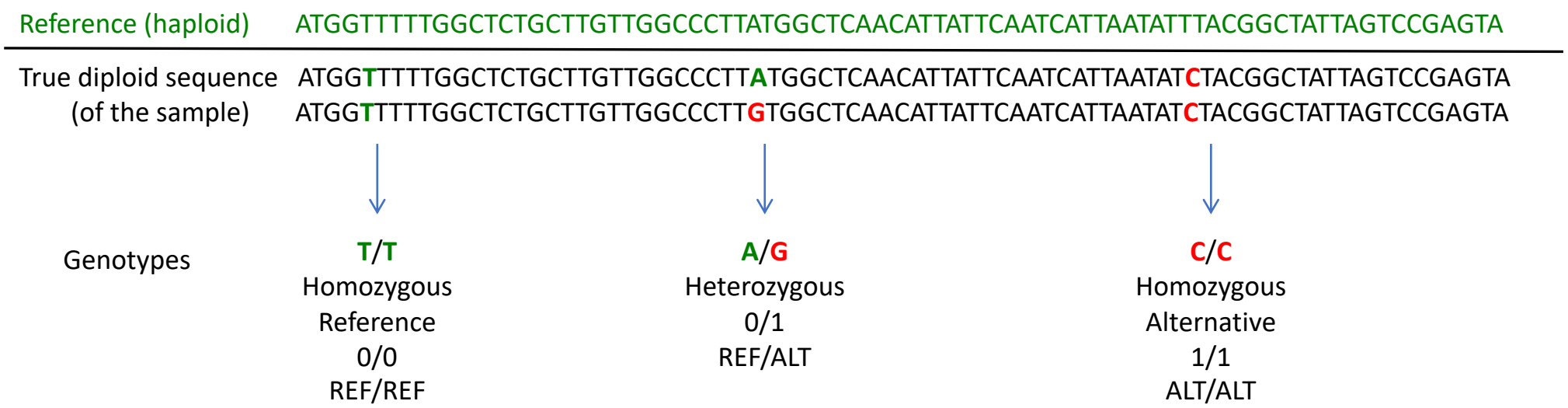
KEEP IN MIND

- You´ll work with “reads” not real DNA sequence
- Reads are versions of DNA sequences with certain accuracy
- **Short reads** are more accurate (**lower error rate**). This is good for most applications.
- **Long Reads** cover **larger stretches** so they are useful to:
  - Resolve **Repeats**
  - Resolve **Complex Variants**
  - Improve **Genome Assembly Contiguity**

4. Sequence Comparison

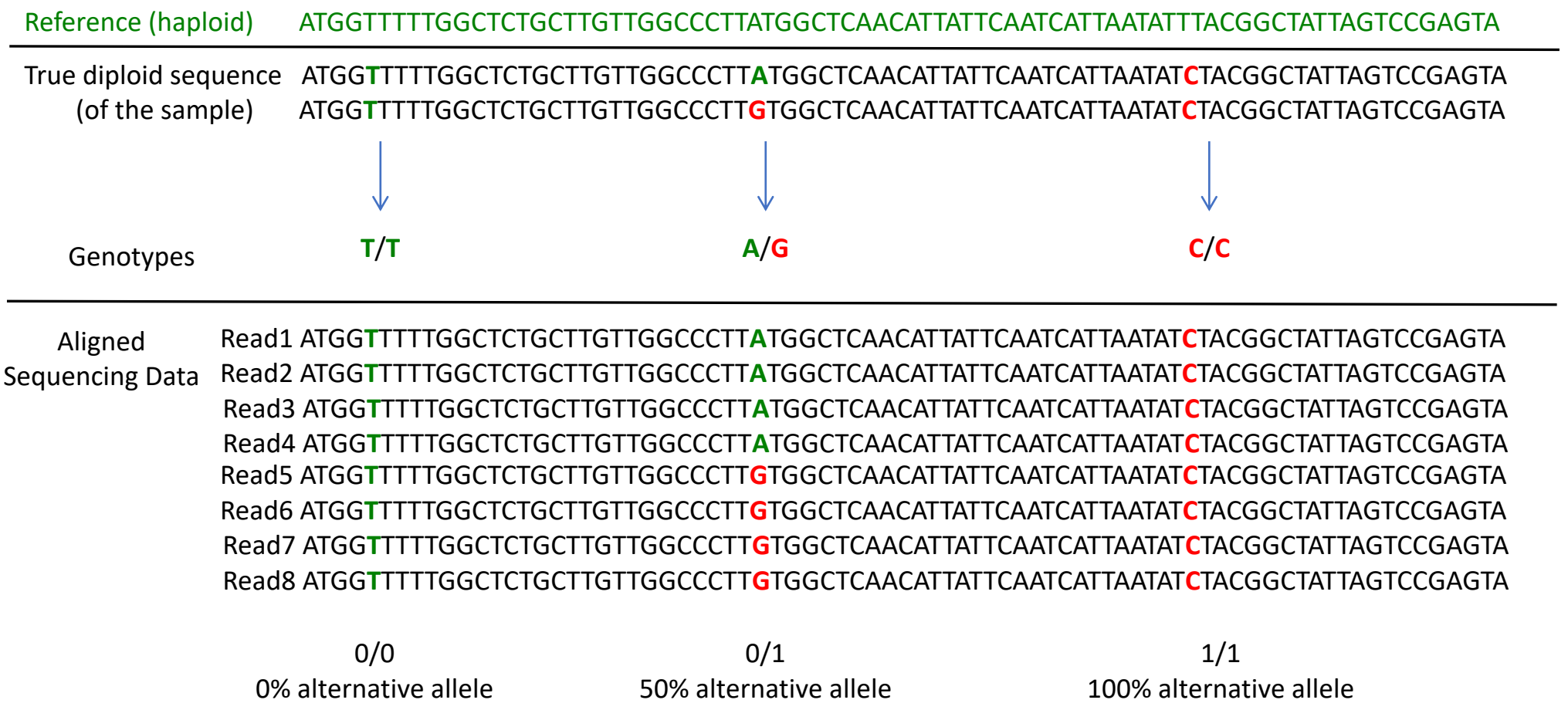
Variant Calling - from Sequence Alignments to Genomic Variants

Genetic difference identified by comparison to an haploid reference:



Variant Calling - from Sequence Alignments to Genomic Variants

Genetic difference identified by comparison to an haploid reference:



KEEP IN MIND

- **Alignments are essential** for Sequence comparison they will determine:
  - Biological **inferences**
  - **Reliabilty** of the **Analyses**:
    - Classification of DNA samples, etc.
    - Transcription levels (RNAseq)
    - Variant Calling and Genotyping
    - Etc..

5. File Formats for Sequencing Data

Fasta

- Structure
  - Description line
    - >
    - ID (optional)
    - Description (optional)
  - Sequence
- Common uses
  - Genome reference sequence
  - Sanger sequencing
- Example

```
>sequenceName Comments about the sequence len=120
ACTGACTGACACTGACTGACACTGACTGACACTGACTGACACTGACTGACACTGACTGAC
ACTGACTGACACTGACTGACACTGACTGACACTGACTGACACTGACTGACACTGACTGAC
```

Fastq

- Structure
  - Description line
    - @
    - ID (optional)
    - Description (optional)
  - Sequence
  - Optional line
    - +
    - ID (optional)
    - Description (optional)
  - Quality
- Common uses
  - Produced by most NGS sequencers
- Example

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*(((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Phred Quality Scores

Phred quality scores are logarithmically linked to error probabilities:

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

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

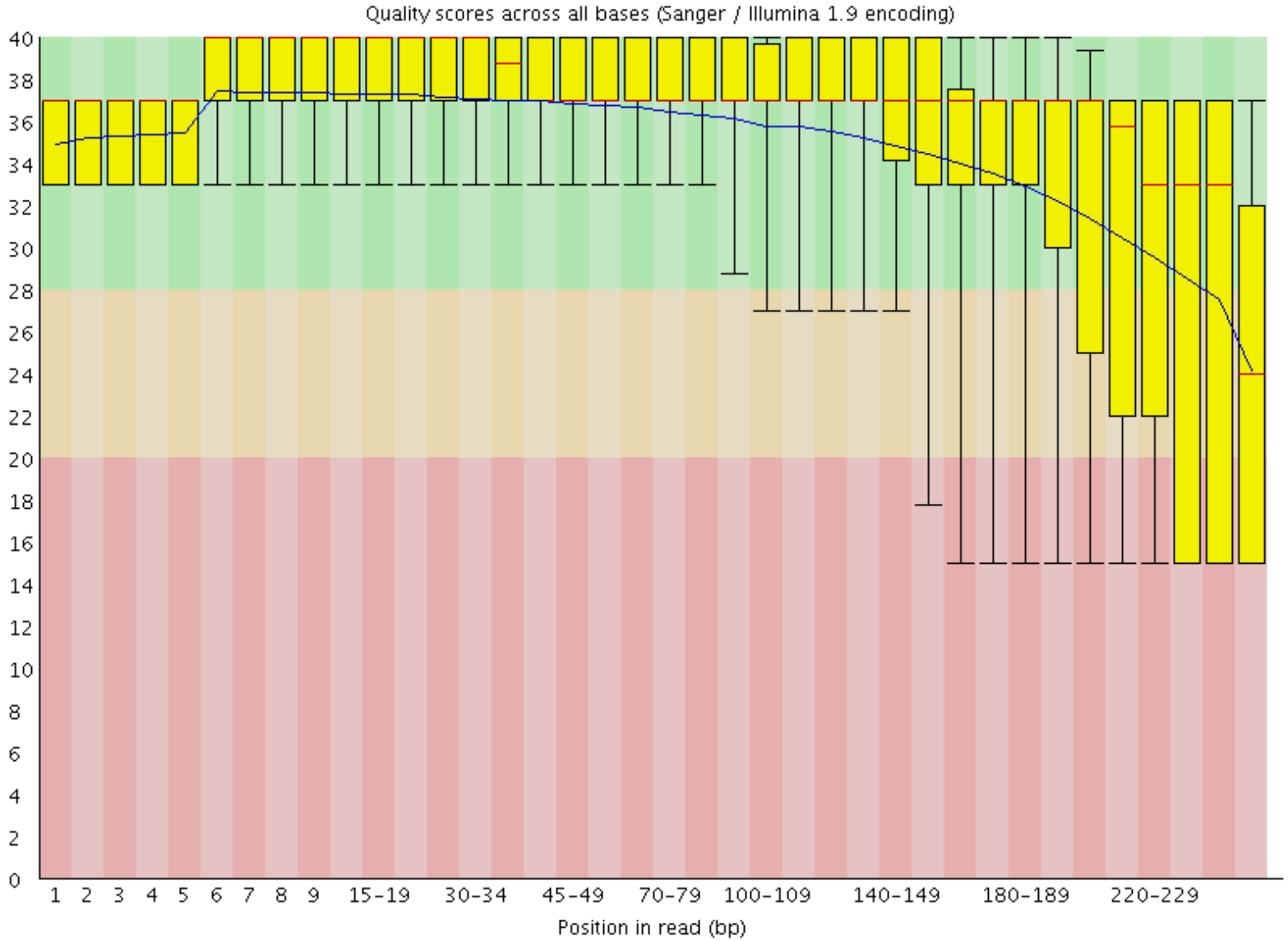
$$P = 10^{(-Q/10)}$$

In .fastq and .bam files, phred scores are represented by matching ASCII characters:

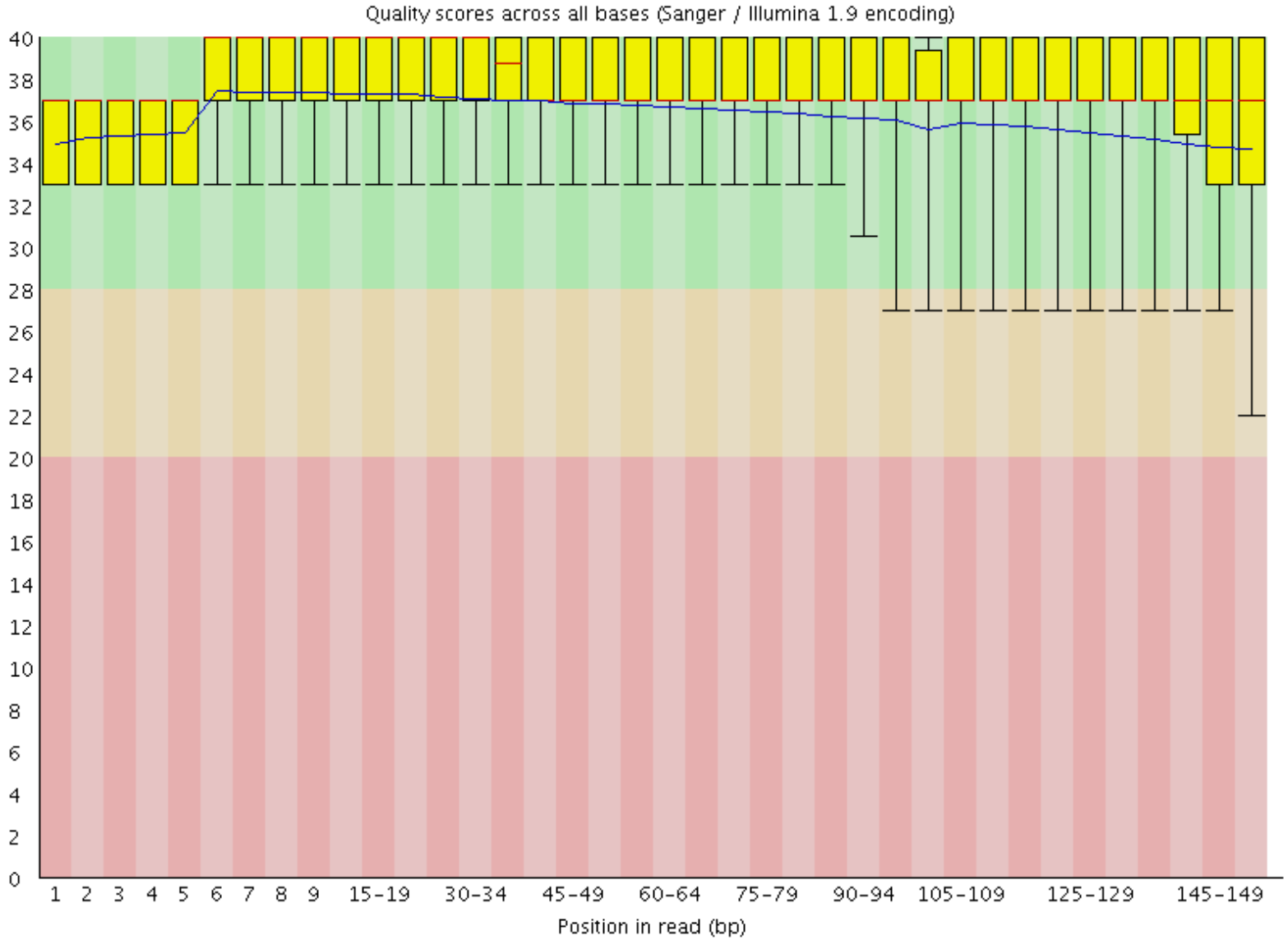
```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ......................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
  ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
  ............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
  ..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |       |                              |                     |
 33                        59   64      73                             104                   126
  0.....................26...31.......40
                         -5....0........9.............................40
                              0........9.............................40
                                 3.....9.............................40
  0.2...................26...31........41

 S - Sanger        Phred+33,  raw reads typically (0, 40)
 X - Solexa        Solexa+64, raw reads typically (-5, 40)
 I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
 J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
     with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
     (Note: See discussion above).
 L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

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

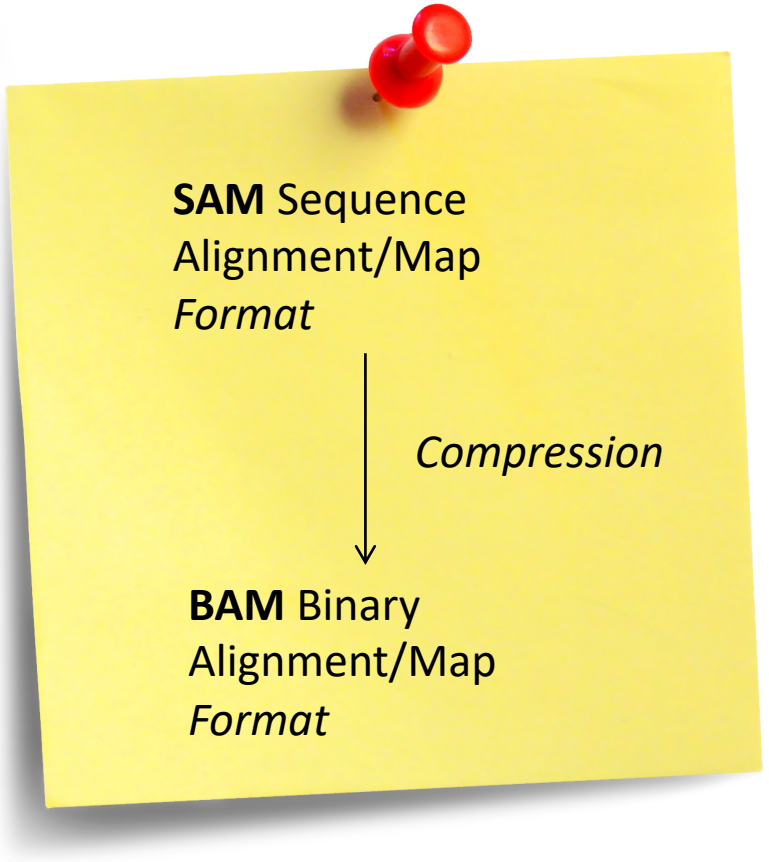
250 bp


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
15-19
30-34
45-49
70-79
100-109
140-149
180-189
220-229
Position in read (bp)

150 bp


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
15-19
30-34
45-49
60-64
75-79
90-94
105-109
125-129
145-149
Position in read (bp)

Bam/Sam Format



Samtools is open source and is available here: http://samtools.sourceforge.net/
Detailed information on the .sam/.bam format standards:
http://samtools.sourceforge.net/SAMv1.pdf

Bam/Sam

```
Coor      12345678901234  5678901234567890123456789012345
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1         TTAGATAAAGGATA*CTG
+r002         aaaAGATAA*GGATA
+r003       gcctaAGCTAA
+r004                     ATAGCT..............TCAGC
-r003                           ttagctTAGGC
-r001/2                                      CAGCGGCAT
```

```
@HD VN:1.5 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
```

.bam files: One of the standard alignment formats: The header

.bam files contain one read with its mapping coordinates per line
.bam files are binary and can be viewed using the program samtools:

Inspect the alignments

```
samtools view -h sample.bam
```

```
@HD VN:1.0 GO:none SO:coordinate

@SQ SN:chrM LN:16571 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:d2ed829b8a1628d16cbeee88e88e39eb
@SQ SN:chr1 LN:249250621 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:1b22b98cdeb4a9304cb5d48026a85128
@SQ SN:chr2 LN:243199373 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:a0d9851da00400dec1098a9255ac712e
@SQ SN:chr3 LN:198022430 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:641e4338fa8d52a5b781bd2a2c08d3c3
@SQ SN:chr4 LN:191154276 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:23dccd106897542ad87d2765d28a19a1
@SQ SN:chr5 LN:180915260 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:0740173db9ffd264d728f32784845cd7
@SQ SN:chr6 LN:171115067 UR:file:/project/production/Genomes/fasta/hsapiens_coordsort_v37.fa M5:1d3a93a248d92a729ee764823acbbc6b
[…]

@RG ID:control PL:ILLUMINA PU:flowcell_lane_index_number LB:library_name SM:sample_barcode

@PG ID:GEM,GEM-mapper,GEM-split-mapper
@PG ID:GATK IndelRealigner VN:1.6-5-g557da77 CL:knownAlleles=[] targetIntervals=CompleteIndelRealigner.intervals
LODThresholdForCleaning=5.0 consensusDeterminationModel=USE_READS entropyThreshold=0.15 maxReadsInMemory=150000
maxIsizeForMovement=3000 maxPositionalMoveAllowed=200 maxConsensuses=30 maxReadsForConsensuses=120
maxReadsForRealignment=20000 noOriginalAlignmentTags=false nWayOut=null generate_nWayOut_md5s=false check_early=false noPGTag=false
keepPGTags=false indelsFileForDebugging=null statisticsFileForDebugging=null SNPsFileForDebugging=null
```

.bam files: One of the standard alignment formats: The records

Alignment mandatory fields

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0, 2^{16} - 1]$	bitwise FLAG
3	RNAME	String	`\*|[:rname:^*=][:rname:]*`	Reference sequence NAME[11]
4	POS	Int	$[0, 2^{31} - 1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8} - 1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[:rname:^*=][:rname:]*`	Reference name of the mate/next read
8	PNEXT	Int	$[0, 2^{31} - 1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31} + 1, 2^{31} - 1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

.bam files: One of the standard alignment formats: The records

.bam files contain one read with its mapping coordinates per line
.bam files are binary and can be viewed using the program samtools:

Inspect the alignments

```
samtools view -h sample.bam
```

QNAME FLAG RNAME POS MAPQ CIGAR RNEXT PNEXT TLEN SEQ QUAL [Optional fields]

HWI-ST699_183:3:2313:15075:99217#25@0 147 chr9 131455900 35 101M = 131455683 -317
TTCTGGTATTCATTTAGGATTTGACGAAACGTTCGAGTCAAACGCAGAATAAAGCCAGCAGGAAGAGGCAGCATGAGGAACCAGAGAGC
TTCTTTACCTGG
CDDDCACEDDDCCCCDDCCDBDDCFFECCHGJIIIJIGIHB@IIGEIJIHFJIGGIIJIIIHHJJJIHGIIJIJJJJIIJIIIHEIIJJJHHHHHHFDDFFCCB
RG:Z:control

HWI-ST537_151:1:2115:10177:85188#22@0 83 chr9 131455701 35 101M = 131455906 305
TATTCATTTAGGATTTGACGAAACGTTCGAGTCAAACGCAGAATAAAGCCAGCAGGAAGAGGCAGCATGAGGAACCAGAGAGCTTCTTTA
CCTGGTTTACT
:CCCDCA>A>ADCCBBB?DE@A;CB@=FEFHEAACFIGGIIJIDJIJIJJIIJJJFIJIGIIIGGJIJIIIGEJIJJJJJJJIJJIHFHHHFFFFF@CC
RG:Z:tumor

HWI-ST537_151:1:2110:9208:83477#22@0 147 chr9 131455909 35 101M = 131455711 -298
TCATTTAGGATTTGACGAAACGTTCGAGTCAAACGCAGAATAAAGCCAGCAGGAAGAGGCAGCATGAGGAACCAGAGAGCTTCTTTACC
TGGTTTACTGAC
DEDDCADDCCDDBB@DDDC=DFFFFHHHHHIIGJJIIJJJJJIJJJJJJIIJJJJJJJJGJJJJJJJJIGAJJJJJJJIJJJJJIHGJHHHHHFFFFFCCC
RG:Z:tumor

VCF

- Structure
  - Metadata
    - ##
    - Fields present in variant lines
  - Header
    - #
    - Mandatory fields (CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO
    - Optional fields (FORMAT, sample1 … sample s)
  - Variant
- Common uses
  - Produced by most NGS Sequences (variant calling)

.vcf files: The Variant Call Format: The header

Variant calling

.vcf files contain one variant position per line
.vcf files can be viewed like text files

```
##fileformat=VCFv4.1
##samtoolsVersion=0.1.19-44428cd
##reference=file:///project/production/Genomes/fasta/hsapiens_coordsort_v37.fa
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="# high-quality ref-forward bases, ref-reverse, alt-forward and alt-reverse
bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##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="# high-quality bases">
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT [s1] … [sn]
```

.vcf files: The Variant Call Format: The records

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT sample_barcode sample_barcode_tumor

chr9 131456174 . G T 221 . DP=58;VDB=1.406345e-01;RPB=7.168422e-01;AF1=0.25;
AC1=1;DP4=16,20,8,11;MQ=35;FQ=222;PV4=1,0.013,1,0.18 GT:PL:DP:SP:GQ
0/0:0,66,255:22:0:68
0/1:254,0,183:33:9:99

chr9 131456385 . GTATATATATATATATA GTATATATATATATATATATA 999 .
INDEL;IS=20,0.689655;DP=57;VDB=3.058426e-
01;AF1=0.9364;AC1=3;DP4=1,1,20,21;MQ=37;FQ=-35.3;
PV4=1,1,1,1 GT:PL:DP:SP:GQ 1/1:255,0,4:22:0:3 1/1:255,63,0:21:0:67

chr9 131456401 . A ATC 120 . INDEL;IS=1,0.031250;DP=59;VDB=3.316126e-
01;AF1=0.5;AC1=2;DP4=8,8,14,11;MQ=36;FQ=123;PV4=0.76,2.2e-06,0.38,1 GT:PL:DP:SP:GQ
0/1:84,0,91:19:0:86 0/1:78,0,181:22:0:81

chr9 131456519 . T G 999 . DP=78;VDB=1.280794e-01;AF1=1;AC1=4;
DP4=0,0,37,26;MQ=35;FQ=-116 GT:PL:DP:SP:GQ 1/1:255,84,0:28:0:99 1/1:255,105,0:35:0:99

Detailed information on the .vcf format standards:
http://samtools.github.io/hts-specs/VCFv4.3.pdf