



Introduction to Transcriptomics Analysis

Class 02 - Next Generation Sequencing applied to RNA-Seq



INSTRUCTOR:

Aureliano Bombarely
Department of Bioscience
Università degli Studi di Milano
aureliano.bombarely@unimi.it

Outline of Topics

1. Basics about genetics and sequencing
2. First steps: Pre-NGS era
3. Short read sequencing technologies
4. Long read sequencing technologies
5. Common file formats
6. Transcriptome sequencing



Outline of Topics

1. Basics about genetics and sequencing
2. First steps: Pre-NGS era
3. Short read sequencing technologies
4. Long read sequencing technologies
5. Common file formats
6. Transcriptome sequencing



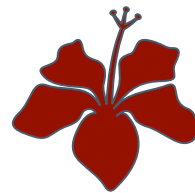
1. Basics about genetics and sequencing

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



1. Basics about genetics and sequencing

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



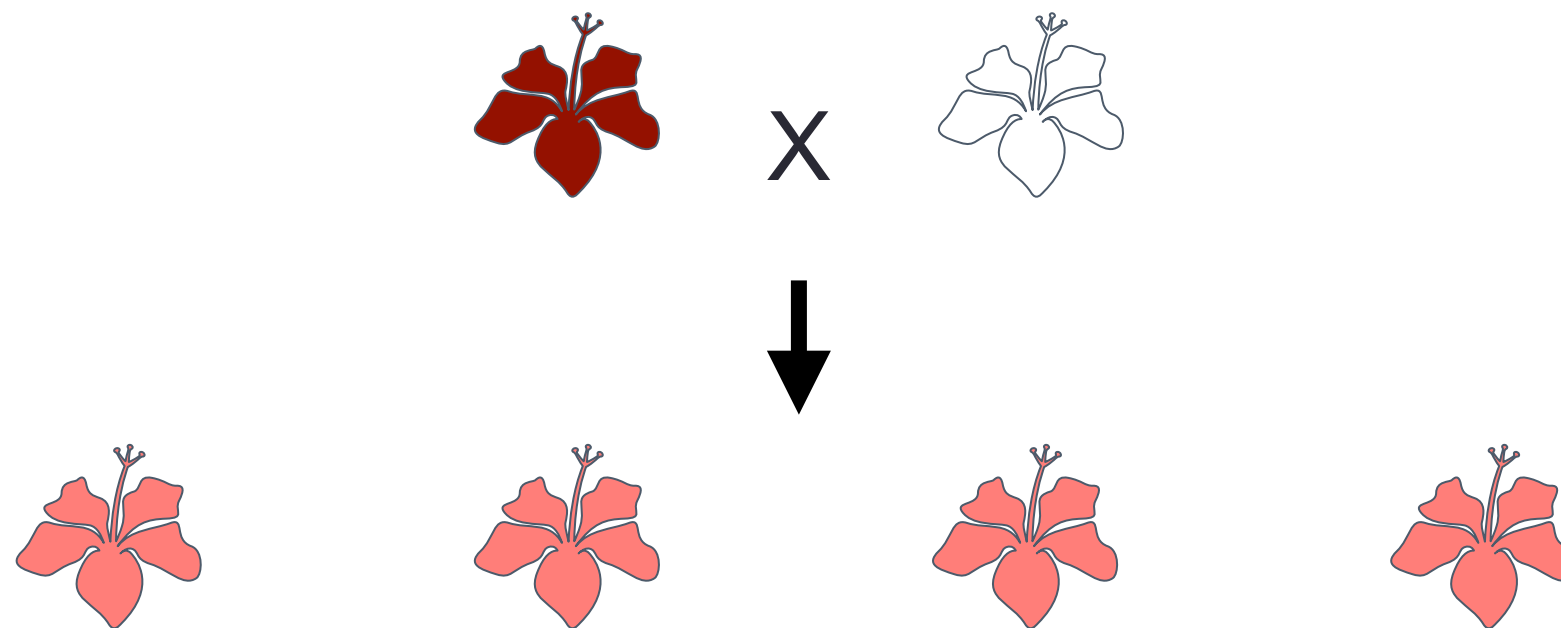
1. Basics about genetics and sequencing

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



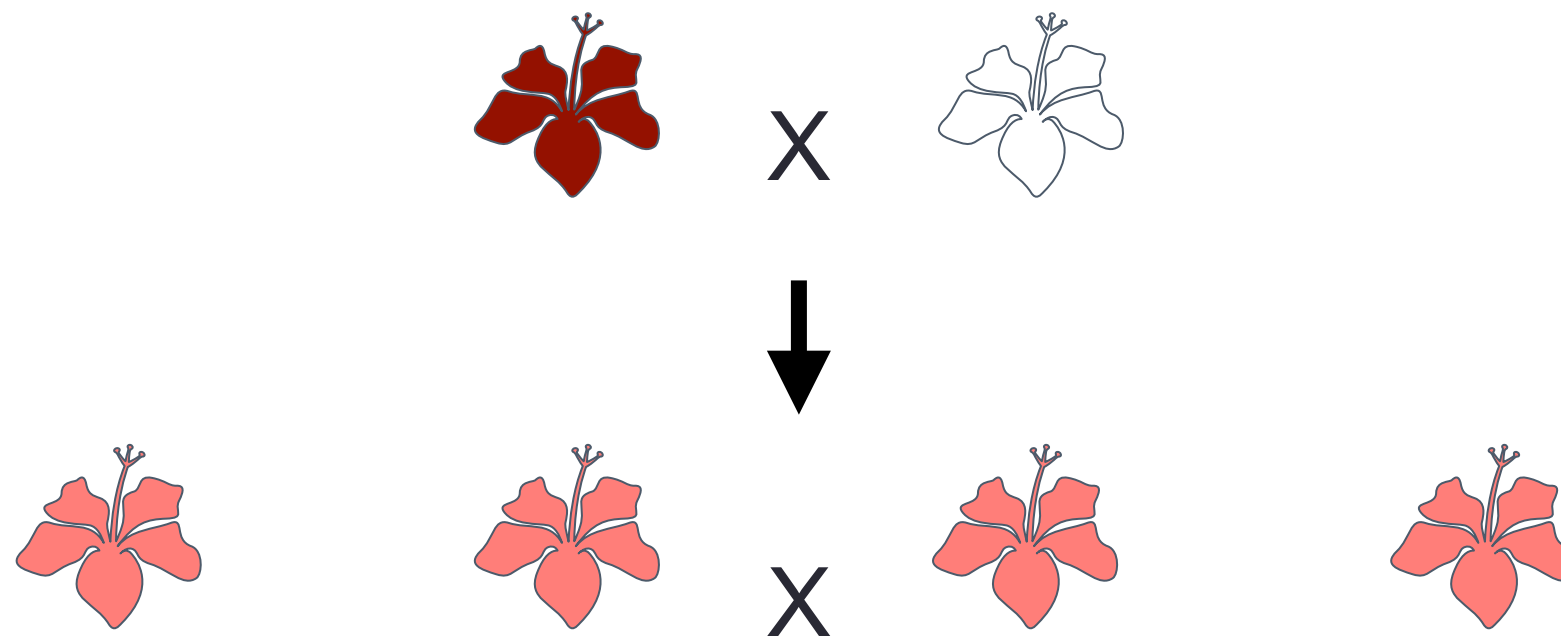
1. Basics about genetics and sequencing

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



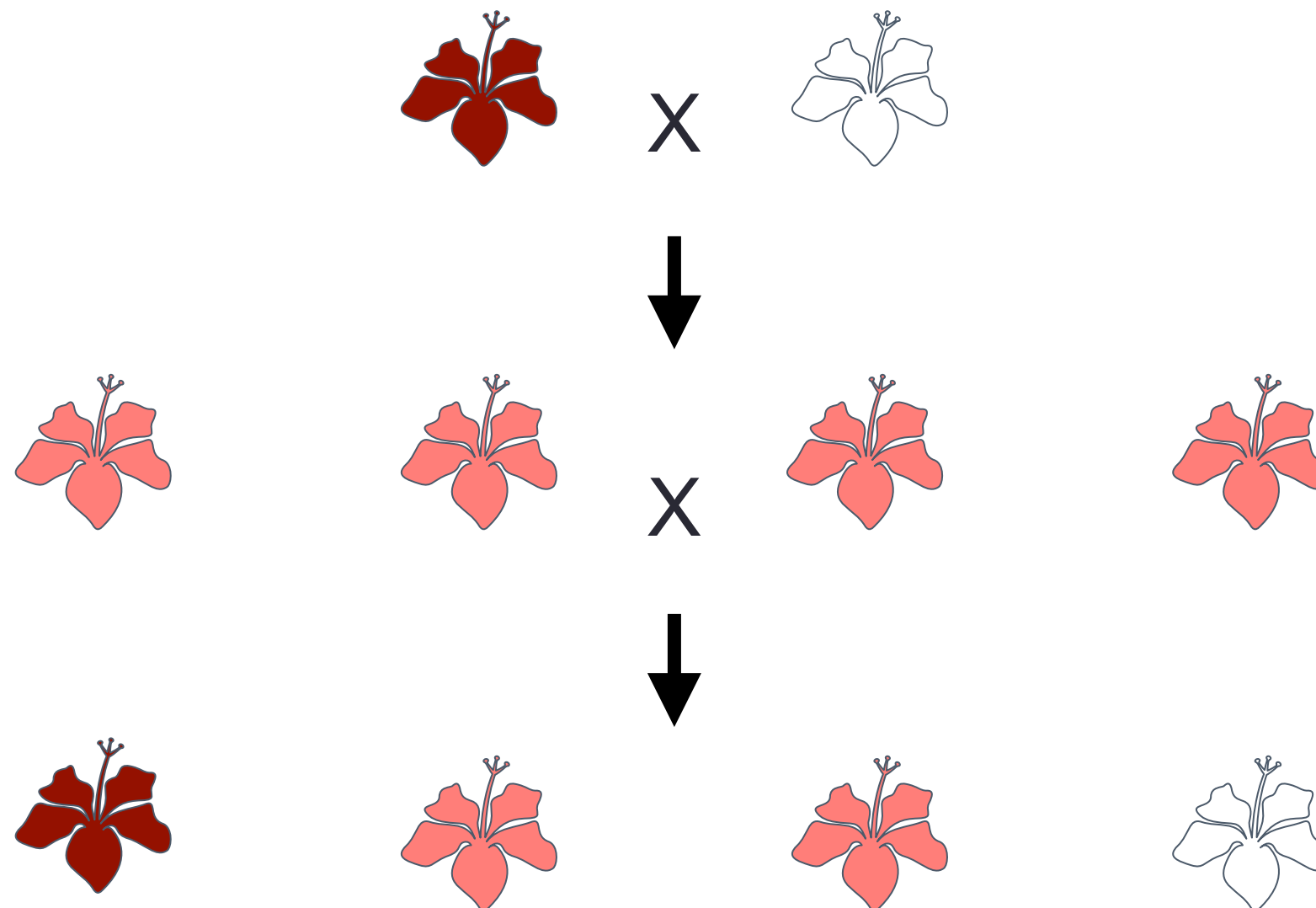
1. Basics about genetics and sequencing

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



1. Basics about genetics and sequencing

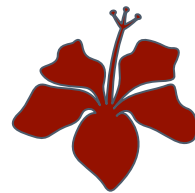
Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



1. Basics about genetics and sequencing

Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms

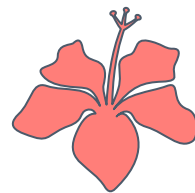
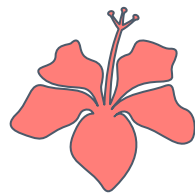
Sources of Variation



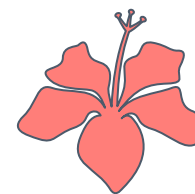
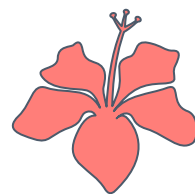
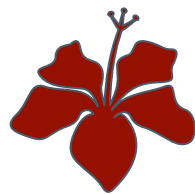
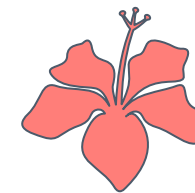
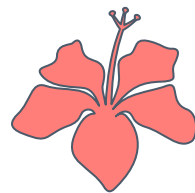
X



Mechanisms of Inherence



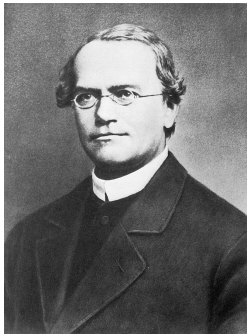
X



1. Basics about genetics and sequencing

Mendel’s Law of Inherence

Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866



Material: Pea plants with different traits

Characteristics of pea plants Gregor Mendel used in his inheritance experiments						
Seeds		Flower colour	Pod		Stem	
form	cotyledons		form	colour	position of inflorescences	size
 round roundish	 yellow	 white	 full	 yellow	 axial	 long
 wrinkled	 green	 violett-red	 constricted between the seeds	 green	 terminal	 short

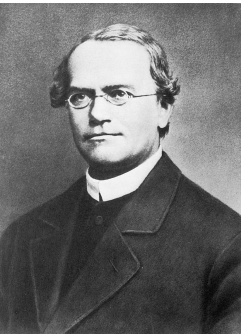
Methods: Cross the parental lines (P) to obtain its progeny (F1) and cross them with themselves to obtain a new progeny (F2). Measure the frequency of each trait in the progeny.



1. Basics about genetics and sequencing

Mendel's Law of Inherence

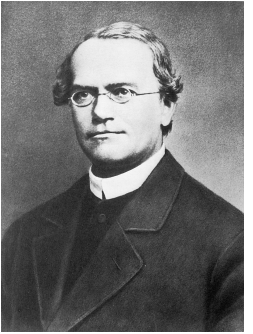
Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866



1. Basics about genetics and sequencing

Mendel's Law of Inherence

Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866



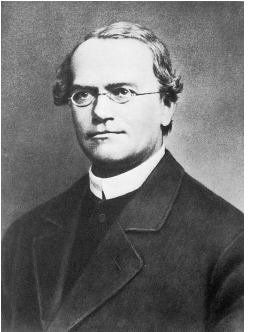
1. Law of dominance and uniformity
2. Law of segregation
3. Law of independent assortment



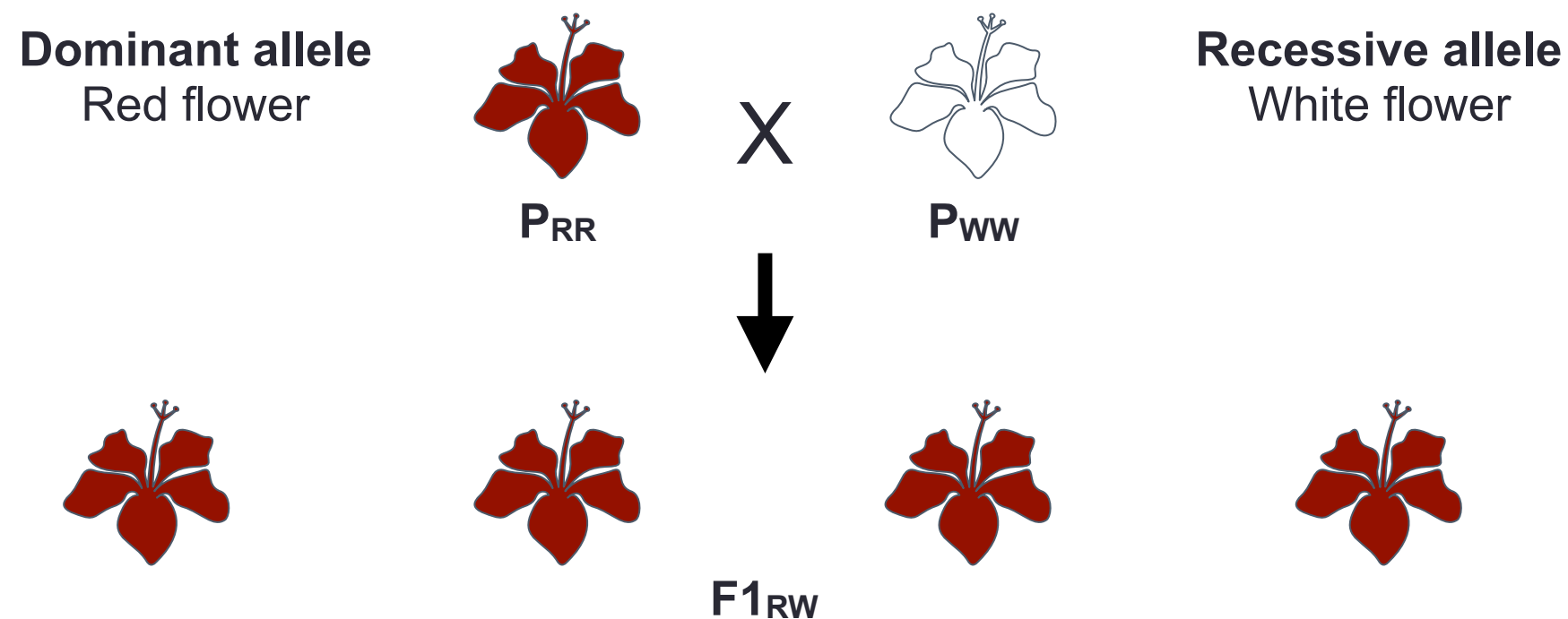
1. Basics about genetics and sequencing

Mendel's Law of Inherence

Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866



1. Law of dominance and uniformity



The dominant allele will present uniformity for the F1 generation

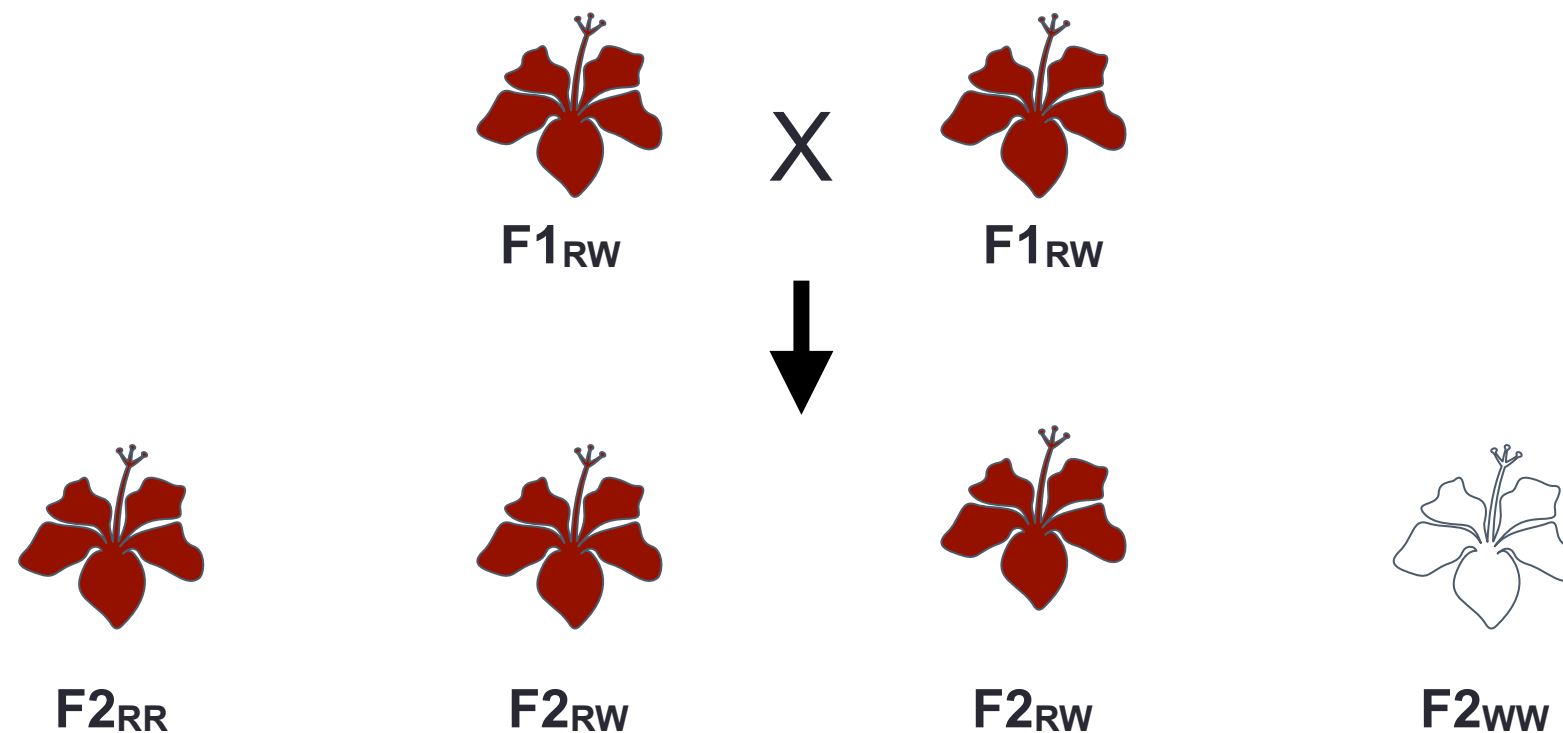
1. Basics about genetics and sequencing

Mendel's Law of Inherence

Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866



2. Law of segregation



		pollen ♂	
		B	b
pistil ♀	B	BB	Bb
	b	Bb	bb

The recessive allele will reappear in the F₂ generation in a 3:1 proportion (dominant:recessive)

Alleles segregate in the progeny



1. Basics about genetics and sequencing

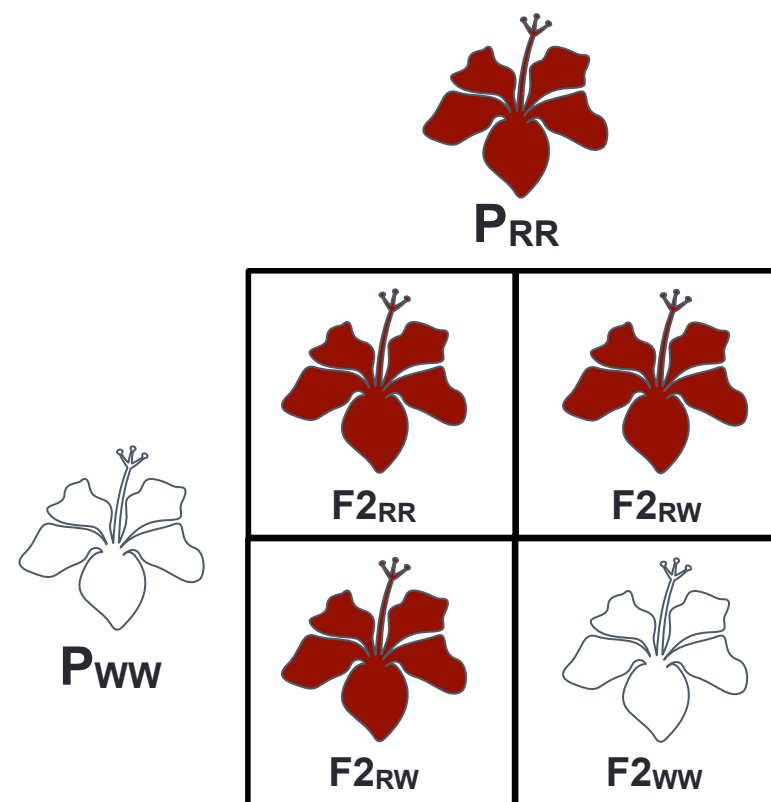
Mendel's Law of Inherence

Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866

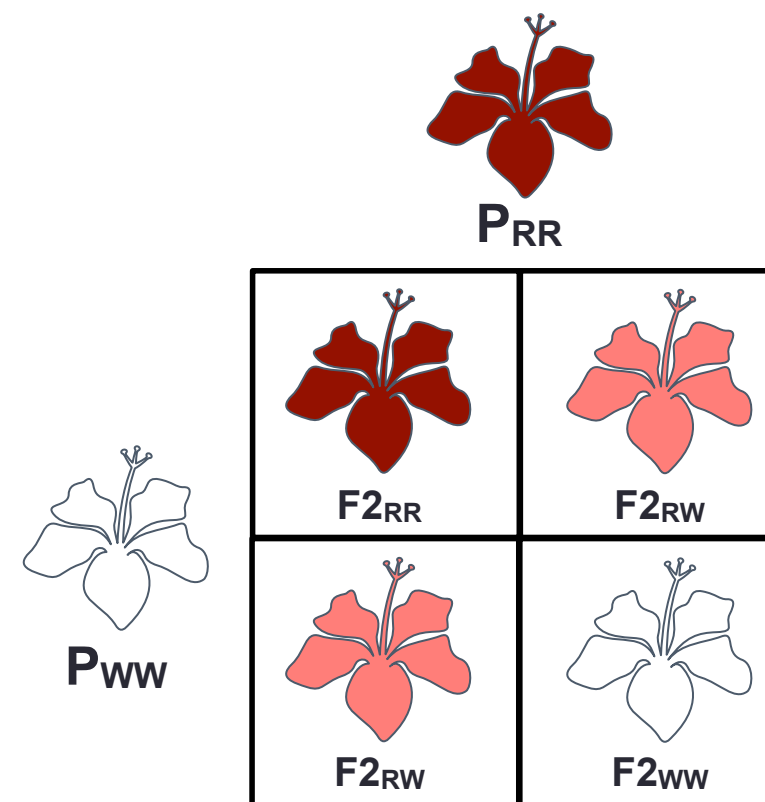


2. Law of segregation

Alleles segregate in the progeny



Dominance
3:1

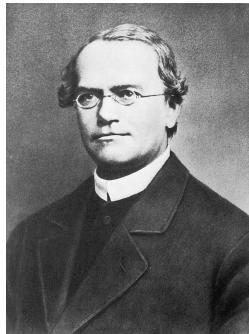


Co-dominance
1:2:1

1. Basics about genetics and sequencing

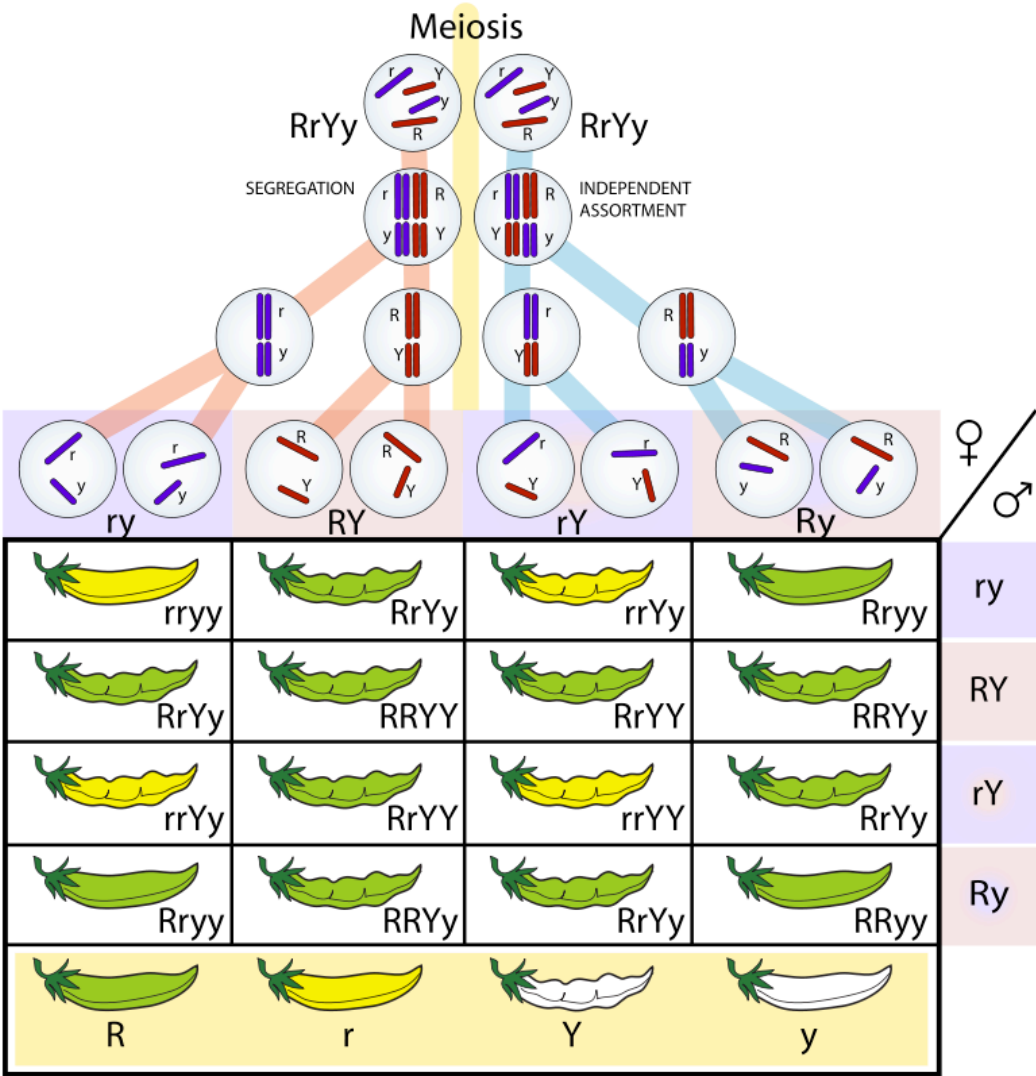
Mendel’s Law of Inherence

Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866



3. Law of independent assortment

Alleles for separate traits are passed independently of one another



Independent Dominance
9:3:3:1



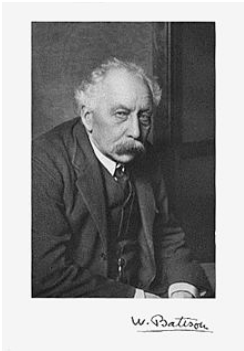
3:1 Green/Yellow (R/r)
3:1 Wrinkle/plain (Y/y)



1. Basics about genetics and sequencing

Mendel’s Law of Inherence

3. Law of independent assortment



Bateson



Saunders



Punnett

but... there are deviations from Mendelian segregations

Flower color: Purple/Red
Seed shape: Long/Round

GENETIC LINKAGE

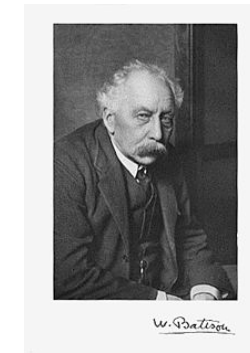
Bateson, Saunders, and Punnett experiment

Phenotype and genotype	Observed	Expected from 9:3:3:1 ratio
Purple, long ($P_L_$)	284	216
Purple, round (P_ll)	21	72
Red, long ($ppL_$)	21	72
Red, round ($ppll$)	55	24

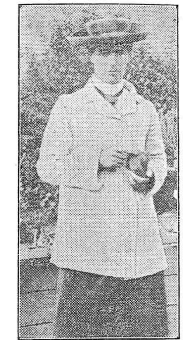


1. Basics about genetics and sequencing

Allele definition



Bateson

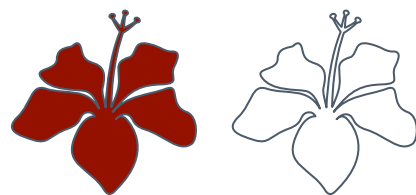


Saunders

Allele is a **variant form** of a given gene, meaning it is one of two or more versions of a known mutation at the same place on a chromosome



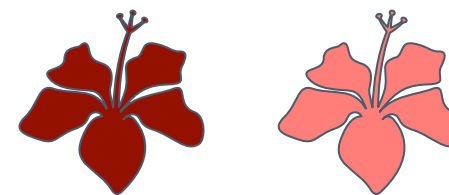
Most individuals have **two sets of chromosomes (diploid)**, so they can have **one (homozygous state)** or **two alleles (heterozygous state)** of the same gene.



RR

WW

Homozygous state



RW

RW

Heterozygous state

1. Basics about genetics and sequencing

Allele definition

Calculation of allele frequencies from genotype frequencies [\[edit \]](#)

The actual frequency calculations depend on the [ploidy](#) of the species for autosomal genes.

Haploids [\[edit \]](#)

The frequency (p) of an allele **A** is the fraction of the number of copies (i) of the **A** allele and the population or sample size (N), so

$$p = i/N.$$

Diploids [\[edit \]](#)

If $f(\mathbf{AA})$, $f(\mathbf{AB})$, and $f(\mathbf{BB})$ are the frequencies of the three genotypes at a locus with two alleles, then the frequency p of the **A**-allele and the frequency q of the **B**-allele in the population are obtained by counting alleles.^[2]

$$p = f(\mathbf{AA}) + \frac{1}{2}f(\mathbf{AB}) = \text{frequency of A}$$

$$q = f(\mathbf{BB}) + \frac{1}{2}f(\mathbf{AB}) = \text{frequency of B}$$

Because p and q are the frequencies of the only two alleles present at that locus, they must sum to 1. To check this:

$$p + q = f(\mathbf{AA}) + f(\mathbf{BB}) + f(\mathbf{AB}) = 1$$

$$q = 1 - p \text{ and } p = 1 - q$$

If there are more than two different allelic forms, the frequency for each allele is simply the frequency of its homozygote plus half the sum of the frequencies for all the heterozygotes in which it appears.

(For 3 alleles see [Allele § Allele and genotype frequencies](#))

Allele frequency can always be calculated from [genotype frequency](#), whereas the reverse requires that the [Hardy–Weinberg conditions](#) of random mating apply.

Example [\[edit \]](#)

Consider a locus that carries two alleles, **A** and **B**. In a diploid population there are three possible genotypes, two [homozygous](#) genotypes (**AA** and **BB**), and one [heterozygous](#) genotype (**AB**). If we sample 10 individuals from the population, and we observe the genotype frequencies

1. freq (**AA**) = 6
2. freq (**AB**) = 3
3. freq (**BB**) = 1

then there are $6 \times 2 + 3 = 15$ observed copies of the **A** allele and $1 \times 2 + 3 = 5$ of the **B** allele, out of 20 total chromosome copies. The frequency p of the **A** allele is $p = 15/20 = 0.75$, and the frequency q of the **B** allele is $q = 5/20 = 0.25$.



1. Basics about genetics and sequencing

Genetic Linkage and Genetic Maps

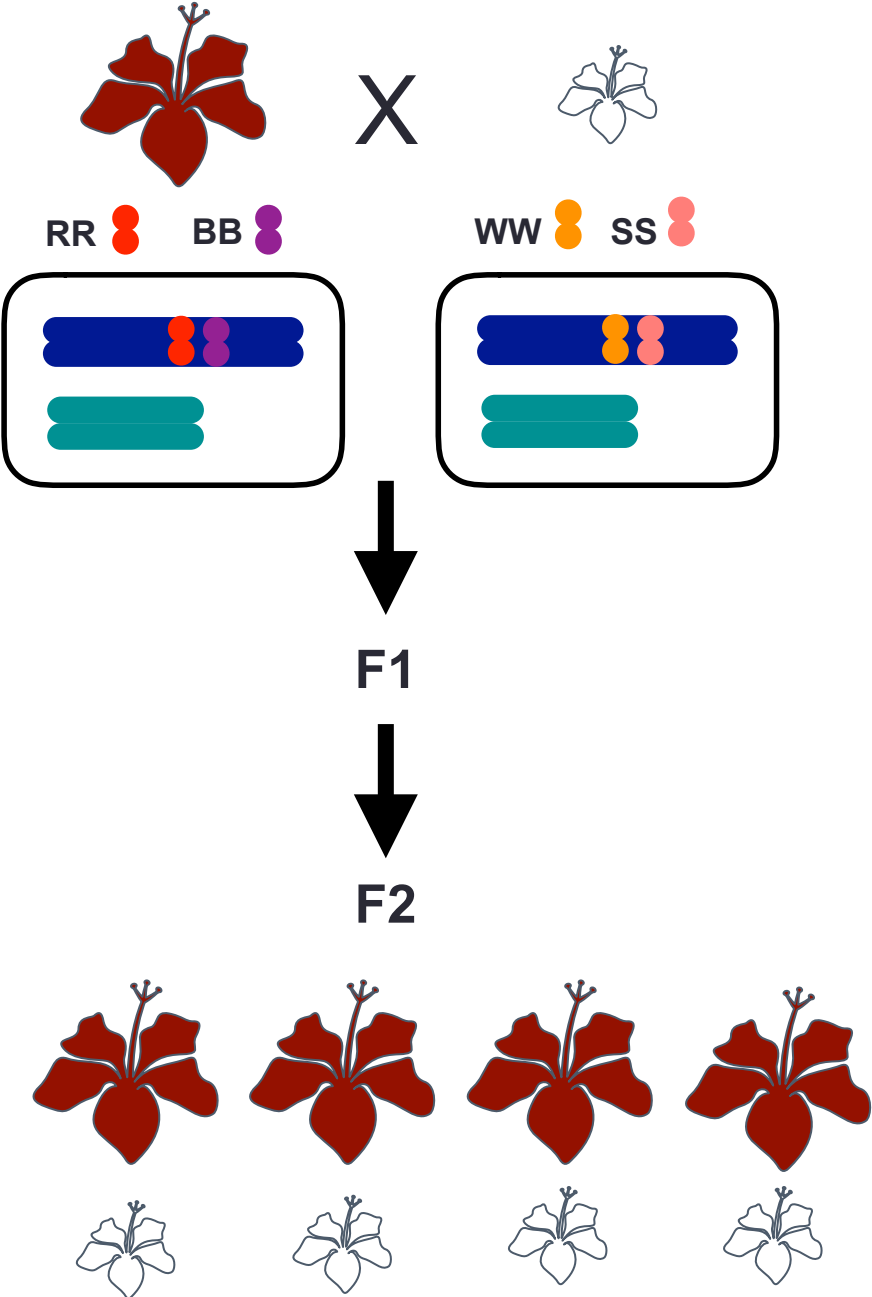
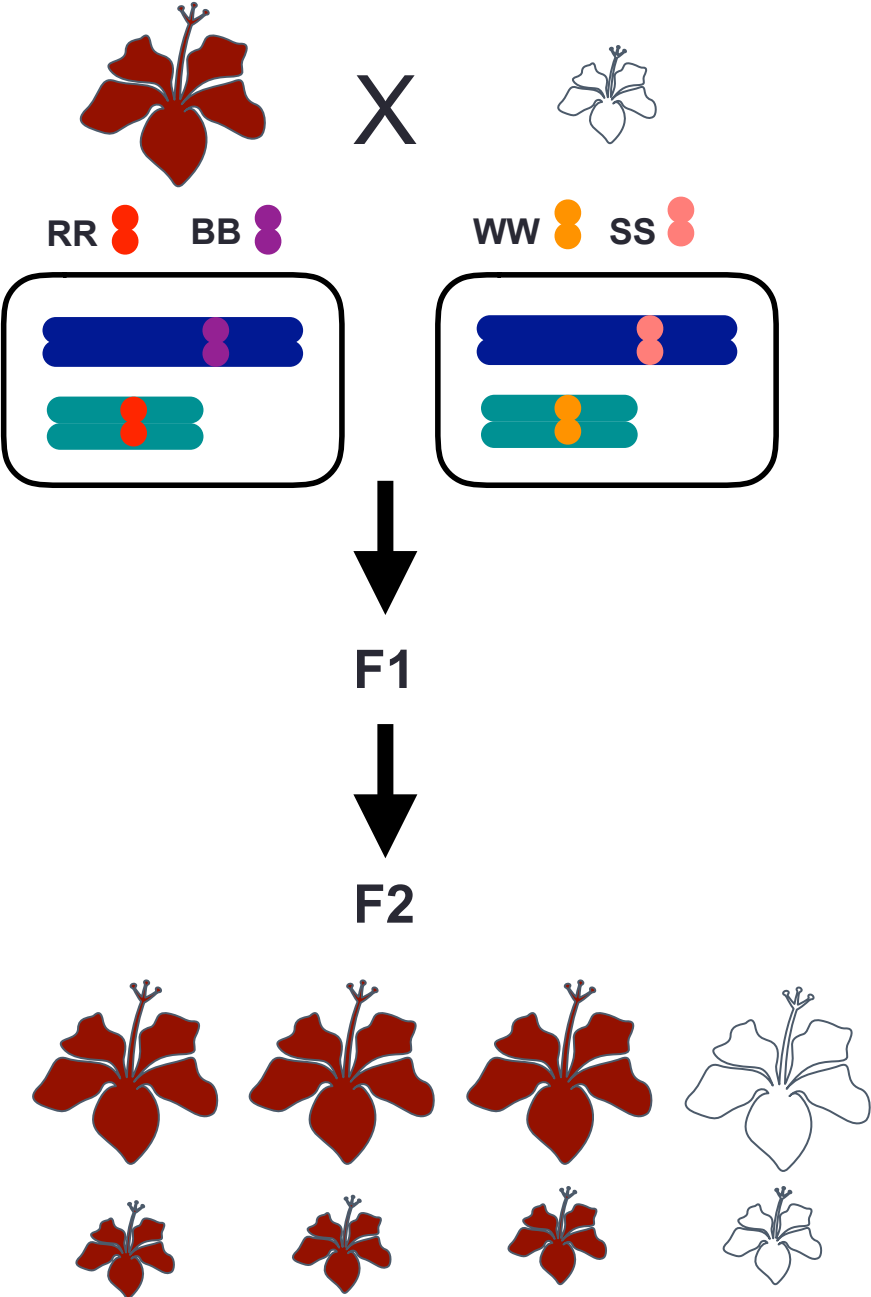


Morgan



Sturtevant

Genetic linkage is the tendency of **DNA sequences that are close together** on a chromosome to **be inherited together** during the meiosis phase of sexual reproduction



1. Basics about genetics and sequencing

Genetic Linkage and Genetic Maps

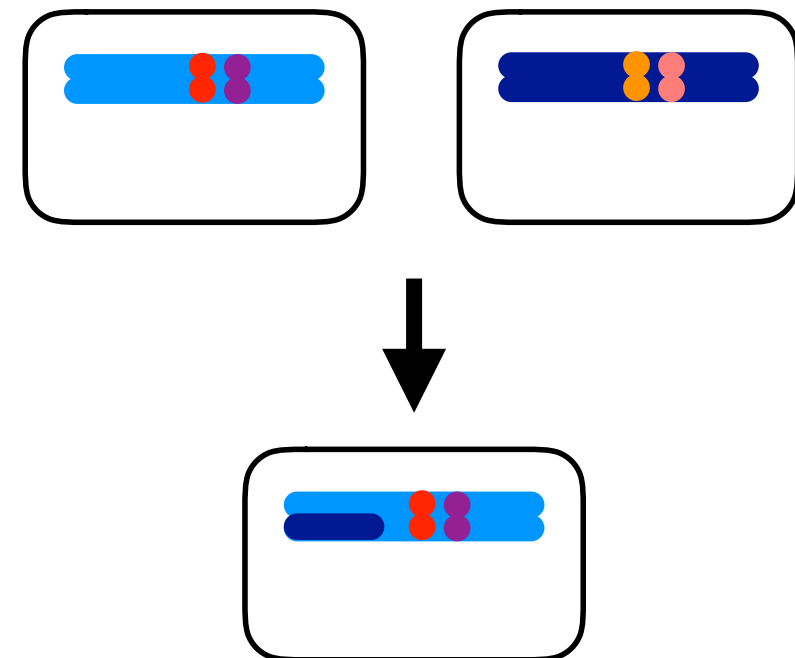
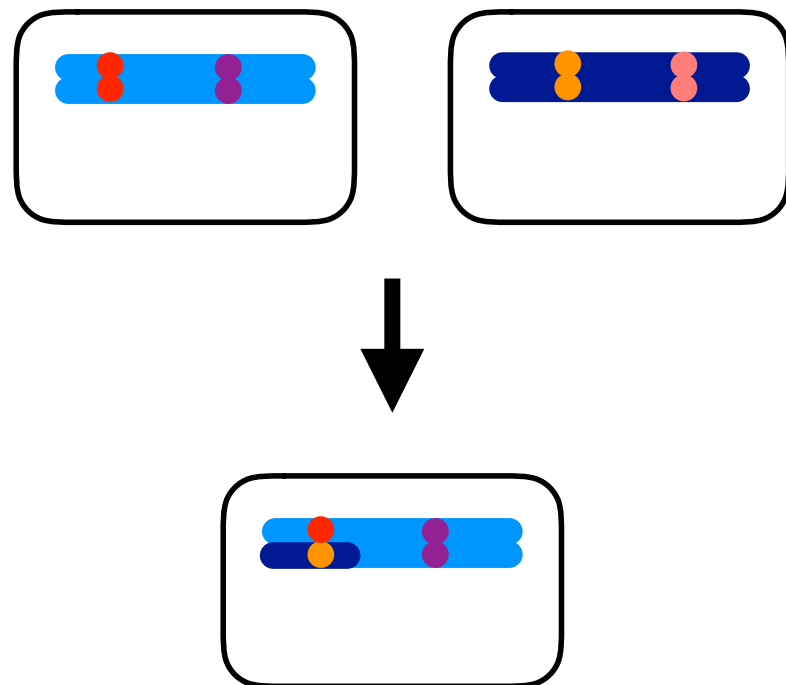


Morgan



Sturtevant

Homologous recombination is a type of genetic recombination in which nucleotide sequences are exchanged between two similar or identical molecules of double-stranded or single-stranded nucleic acids



1. Basics about genetics and sequencing

Genetic Linkage and Genetic Maps

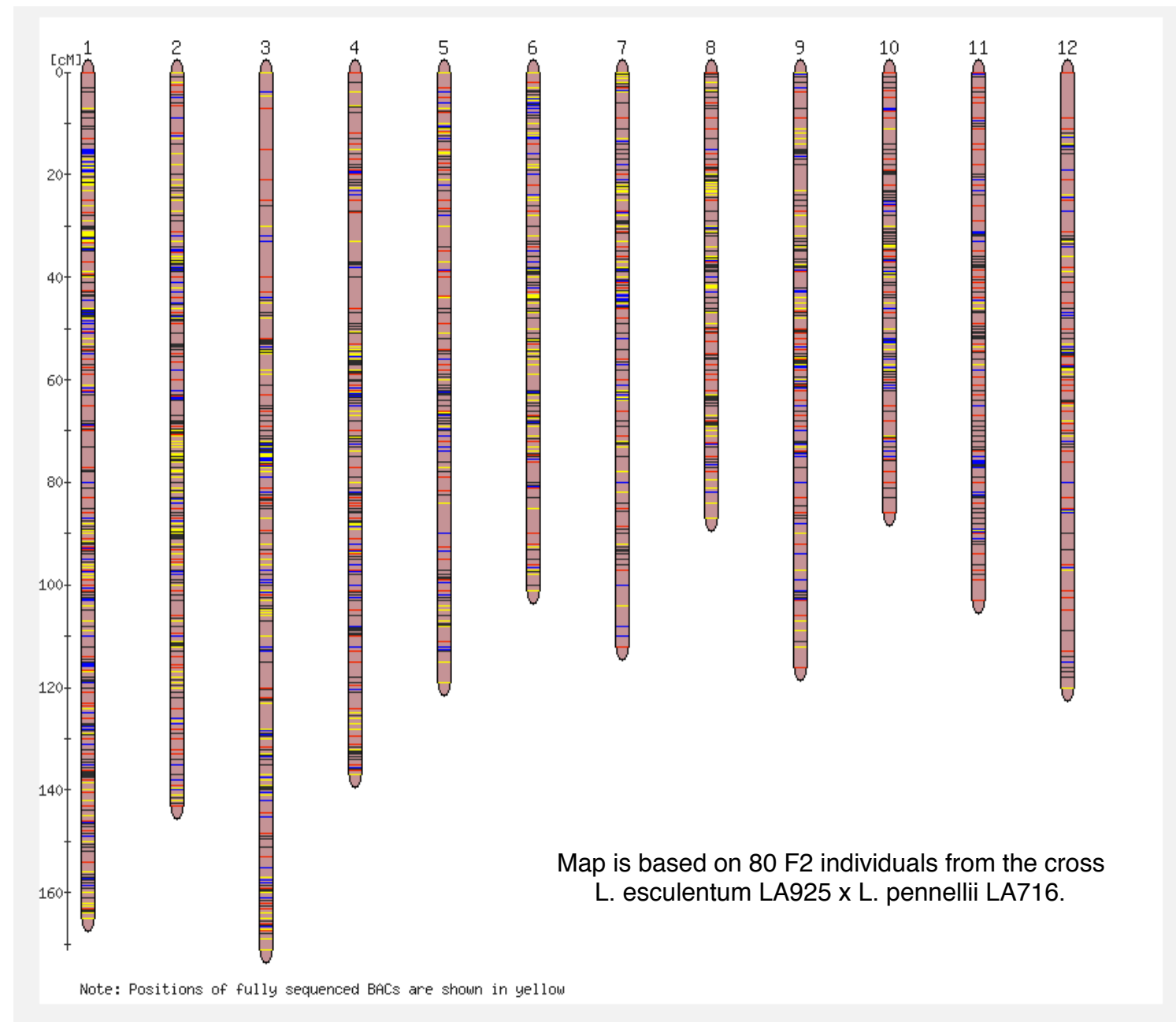
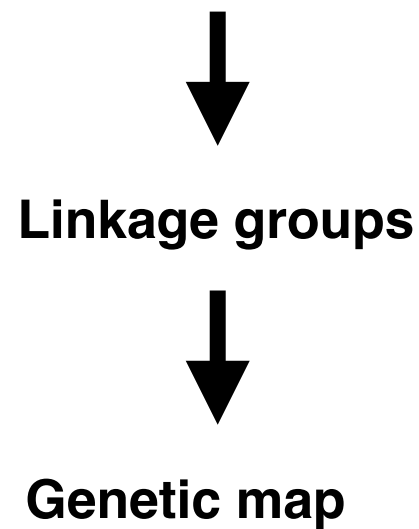


Morgan



Sturtevant

Linkage analysis is a genetic method that searches for chromosomal segments that **cosegregate**



1. Basics about genetics and sequencing

Genetic Marker

A **genetic marker** is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.

- RFLP (or [Restriction fragment length polymorphism](#))
- SSLP (or [Simple sequence length polymorphism](#))
- AFLP (or [Amplified fragment length polymorphism](#))
- RAPD (or [Random amplification of polymorphic DNA](#))
- VNTR (or [Variable number tandem repeat](#))
- SSR [Microsatellite](#) polymorphism, (or [Simple sequence repeat](#))
- SNP (or [Single nucleotide polymorphism](#))
- STR (or [Short tandem repeat](#))
- SFP (or [Single feature polymorphism](#))
- DArT (or [Diversity Arrays Technology](#))
- RAD markers (or [Restriction site associated DNA markers](#))



1. Basics about genetics and sequencing

Genetic Marker

A **genetic marker** is a gene or **DNA sequence** with a known location on a **chromosome** that can be used to identify individuals or **species**. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.

Microsatellite

———— **ATATATAT** ————

Reference

———— **ATATATAT** ————

Sample 1

———— **ATATATAT** ————

———— **AT** ————

Sample 2

———— **ATATATA** ————

———— **AT** ————

Sample 3

———— **AT** ————



1. Basics about genetics and sequencing

Genetic Marker

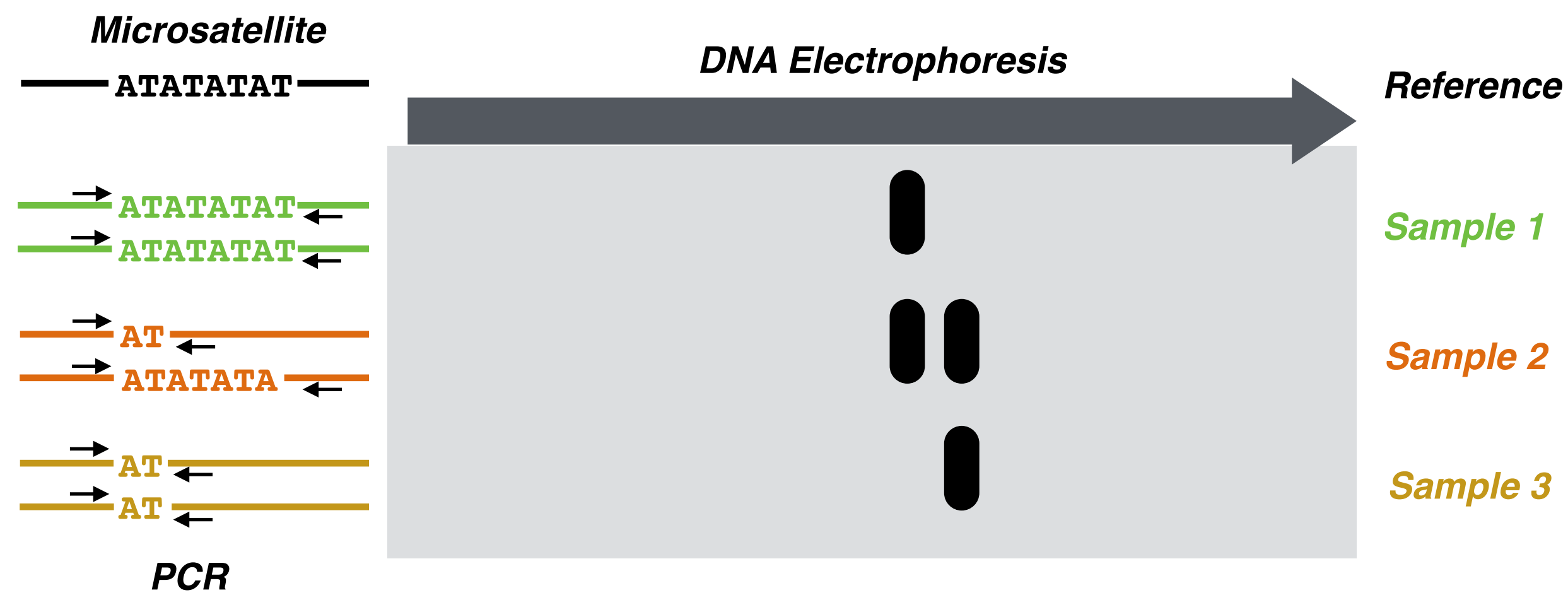
A **genetic marker** is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.



1. Basics about genetics and sequencing

Genetic Marker

A **genetic marker** is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.



1. Basics about genetics and sequencing

Genetic Marker

A **genetic marker** is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.

<i>Microsatellite</i>		<i>SNP</i>						<i>Reference</i>
———— ATATATAT ————		C ————						
———— ATATATAT ————	———— C ————	chr1	1001	C	G	...	0 / 0	<i>Sample 1</i>
———— ATATATAT ————	———— C ————							
———— AT ————	———— G ————	chr1	1001	C	G	...	0 / 1	<i>Sample 2</i>
———— ATATATA ————	———— C ————							
———— AT ————	———— G ————	chr1	1001	C	G	...	1 / 1	<i>Sample 3</i>
———— AT ————	———— G ————							



1. Basics about genetics and sequencing

Locus

Locus (plural loci) is a specific, **fixed position on a chromosome** where a particular gene or genetic marker is located



Tomato locus sun

[Download GMOD XML](#) | [Note to Editors](#) | [Annotation guidelines](#)

New Edit Delete

Locus	
Locus name	sun
Symbol	<i>sun</i>
Gene activity	putative calmodulin binding protein
Description	Belongs to the IQ67 domain family of proteins. Controls elongated and pointed fruit shape. Note that this locus is not present in the Heinz 1706 reference sequence (Heinz 1706 has round fruits). The most closely related sequence in the reference is Solyc10g079240, the presumptive origin of the transposition.
Chromosome	7
Arm	short

Locus synonyms 1: SUN [\[Add/Remove\]](#)

sun is a [TGRC gene](#)
Locus editors: [Esther van der Knaap](#)
[log-in to request editor privileges]
Created on: 2006-05-10 Last updated on: 2012-11-15 by [Lukas Mueller](#)

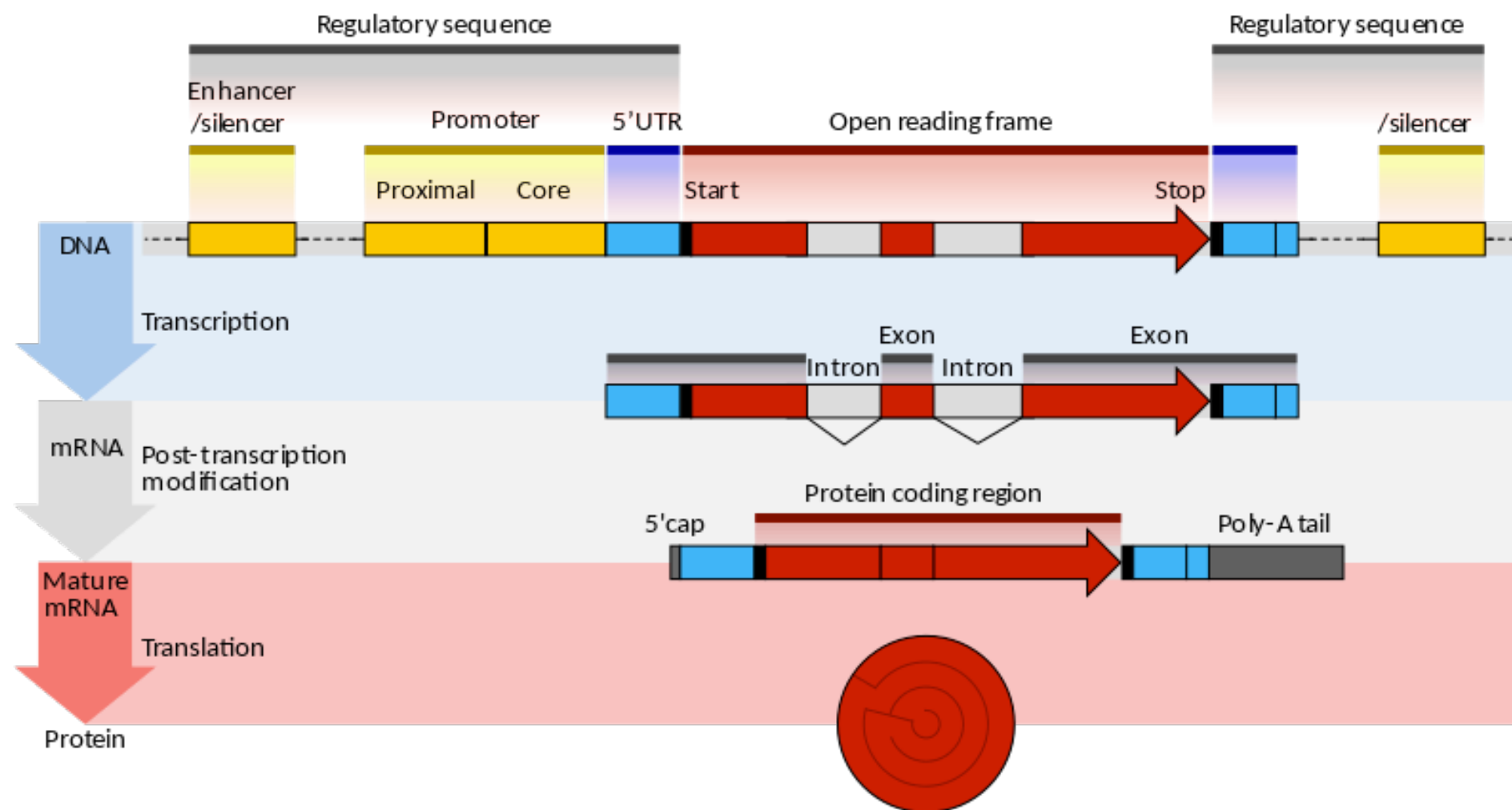
Tomato-EXPEN 2000



1. Basics about genetics and sequencing

Gene

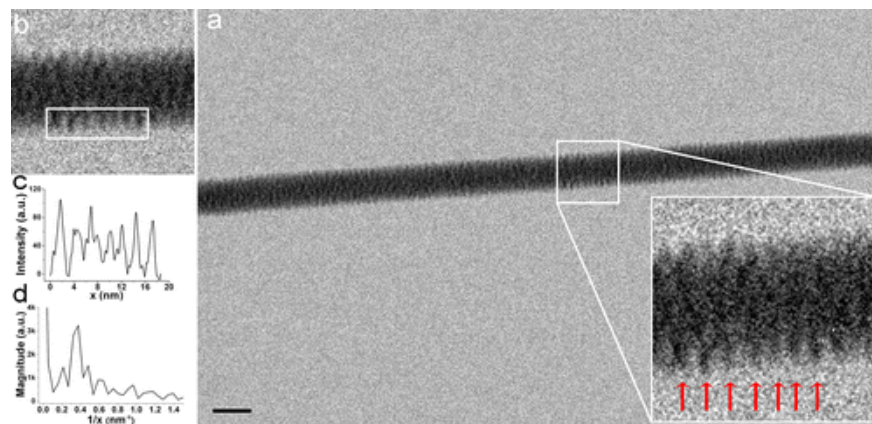
Gene is a **sequence of nucleotides** in DNA or RNA that encodes the synthesis of a gene product, either RNA or protein



1. Basics about genetics and sequencing

DNA sequencing is the process of determining the ***precise order of nucleotides within a DNA molecule***. It includes any method or technology that is used to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA.

https://en.wikipedia.org/wiki/DNA_sequencing



ATGCGCGTCGCGGTGAAT

(Gentile et al. Nano Lett., 2012, 12 (12), pp 6453–6458)

1. Basics about genetics and sequencing

To know more about basic genetics...

<https://learn.genetics.utah.edu/content/basics/>

 **Learn.Genetics**
GENETIC SCIENCE LEARNING CENTER

DONATE

[Home](#) / [Basic Genetics](#)

Basic Genetics

View Teach.Genetics for Classroom Materials

MESSAGE FROM LEARN.GENETICS

Learn.Genetics visitors,

We're asking for your help. For over 20 years, the Learn.Genetics website has provided engaging, multimedia educational materials at *no cost*.

Learn.Genetics is one of the most-used science websites. Tens of millions of visitors come to our site each year to find the science and health information they're looking for.

If Learn.Genetics is useful to you, please take a moment to [donate](#) – even a few dollars from each of our visitors would add up to a significant amount!

Your support will help us keep Learn.Genetics free and available to everyone. It will also help us develop new content for you.

Please help us keep Learn.Genetics going!

Thank you,
The Genetic Science Learning Center team – creators of Learn.Genetics

TOUR OF BASIC GENETICS



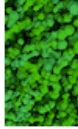
WHAT ARE TRAITS?
Explore traits, the characteristics that make us unique.

video



WHAT ARE DNA & GENES?
Get to know the molecule that holds the instructions for building every living thing.

video



WHAT ARE PROTEINS?
Learn about proteins, the molecules we're made of.

video



WHAT IS INHERITANCE?
Learn how traits pass from parents to offspring.

video



WHAT IS MUTATION?
Take a look at how variation occurs.

video

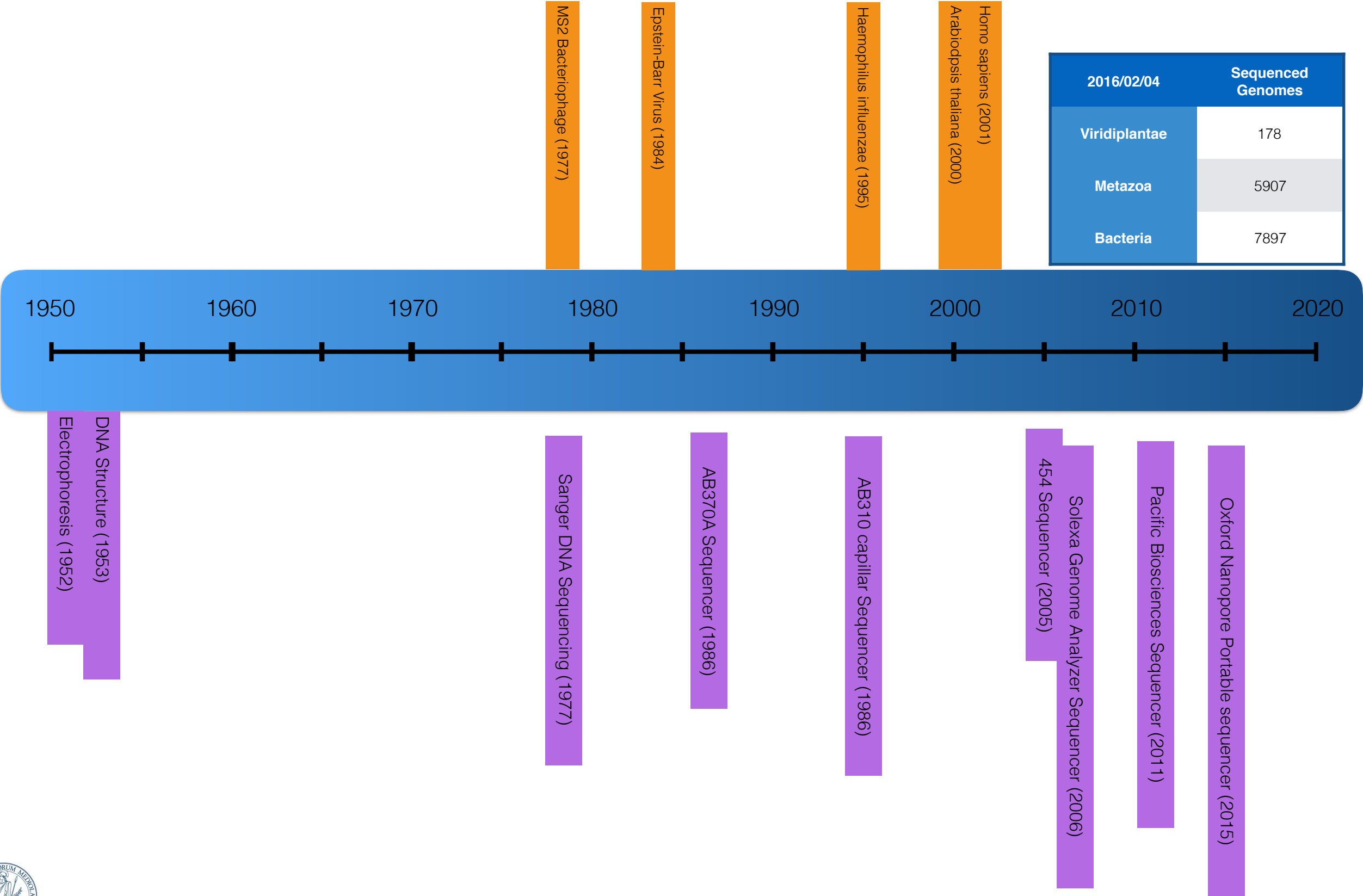


Outline of Topics

1. Basics about genetics and sequencing
- 2. First steps: Pre-NGS era**
3. Short read sequencing technologies
4. Long read sequencing technologies
5. Common file formats
6. Transcriptome sequencing



2. First steps: Pre-NGS era



2. First steps: Pre-NGS era

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

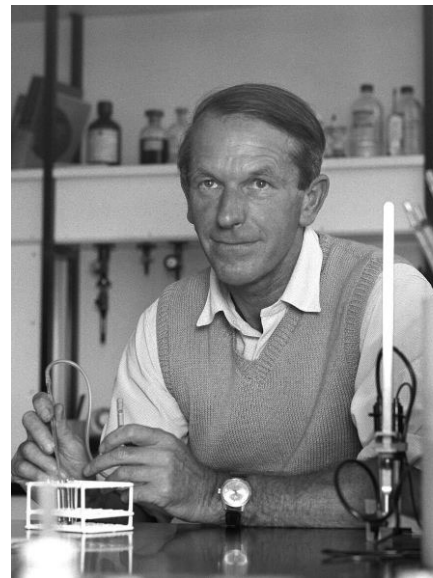
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕ X174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977



Frederick Sanger (1918-2013)
Twice awarded with the Nobel Prize of Chemistry

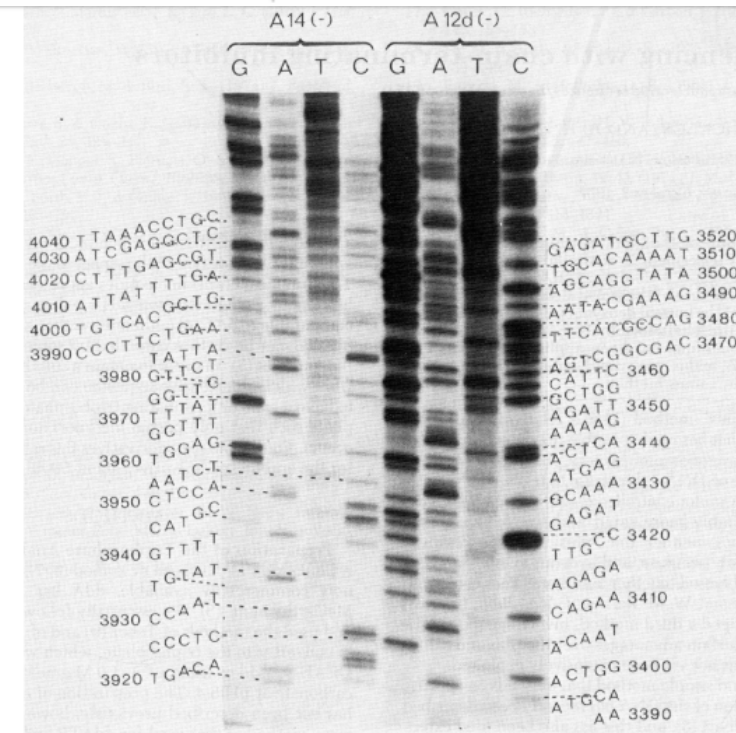


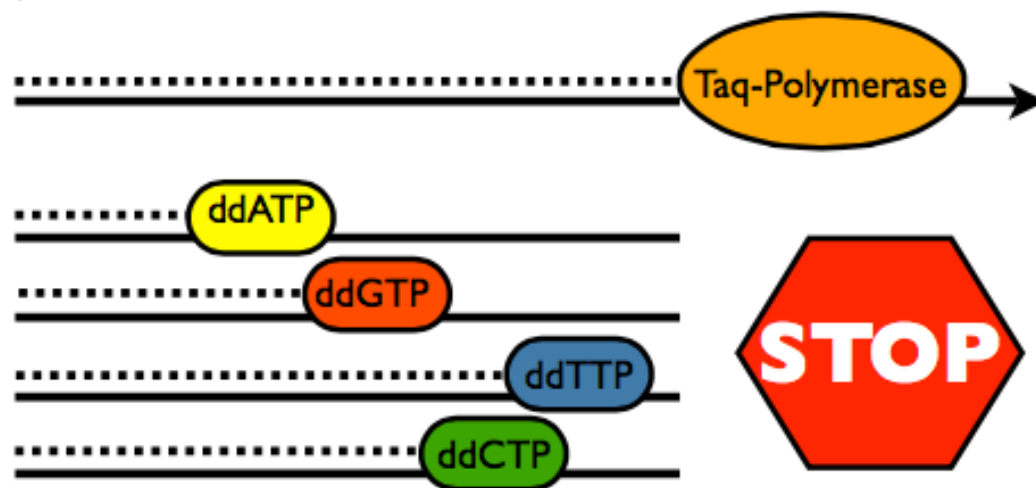
FIG. 1. Autoradiograph of the acrylamide gel from the sequence determination using restriction fragments A12d and A14 as primers on the complementary strand of ϕ X174 DNA. The inhibitors used were (left to right) ddGTP, ddATP, ddTTP, and araCTP. Electrophoresis was on a 12% acrylamide gel at 40 mA for 14 hr. The top 10 cm of the gel is not shown. The DNA sequence is written from left to right and upwards beside the corresponding bands on the radioautograph. The numbering is as given in ref. 2.



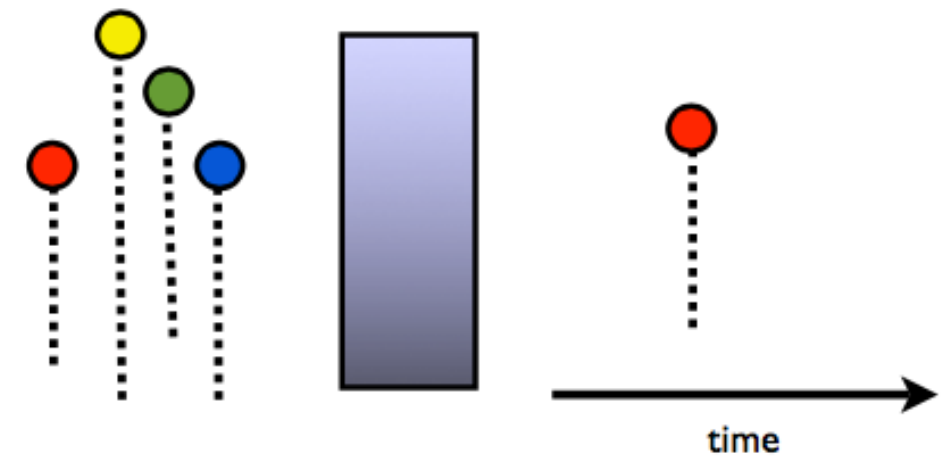
2. First steps: Pre-NGS era

Sanger DNA sequencing

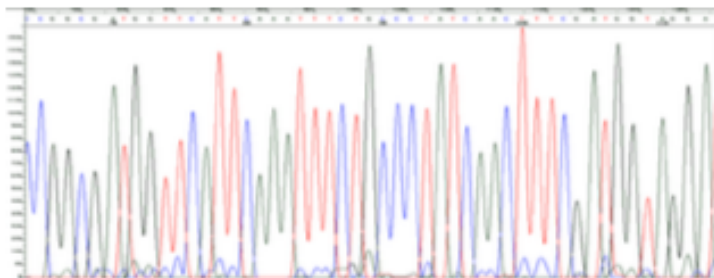
1) PCR with ddNTPs



2) Chromatographic Separation



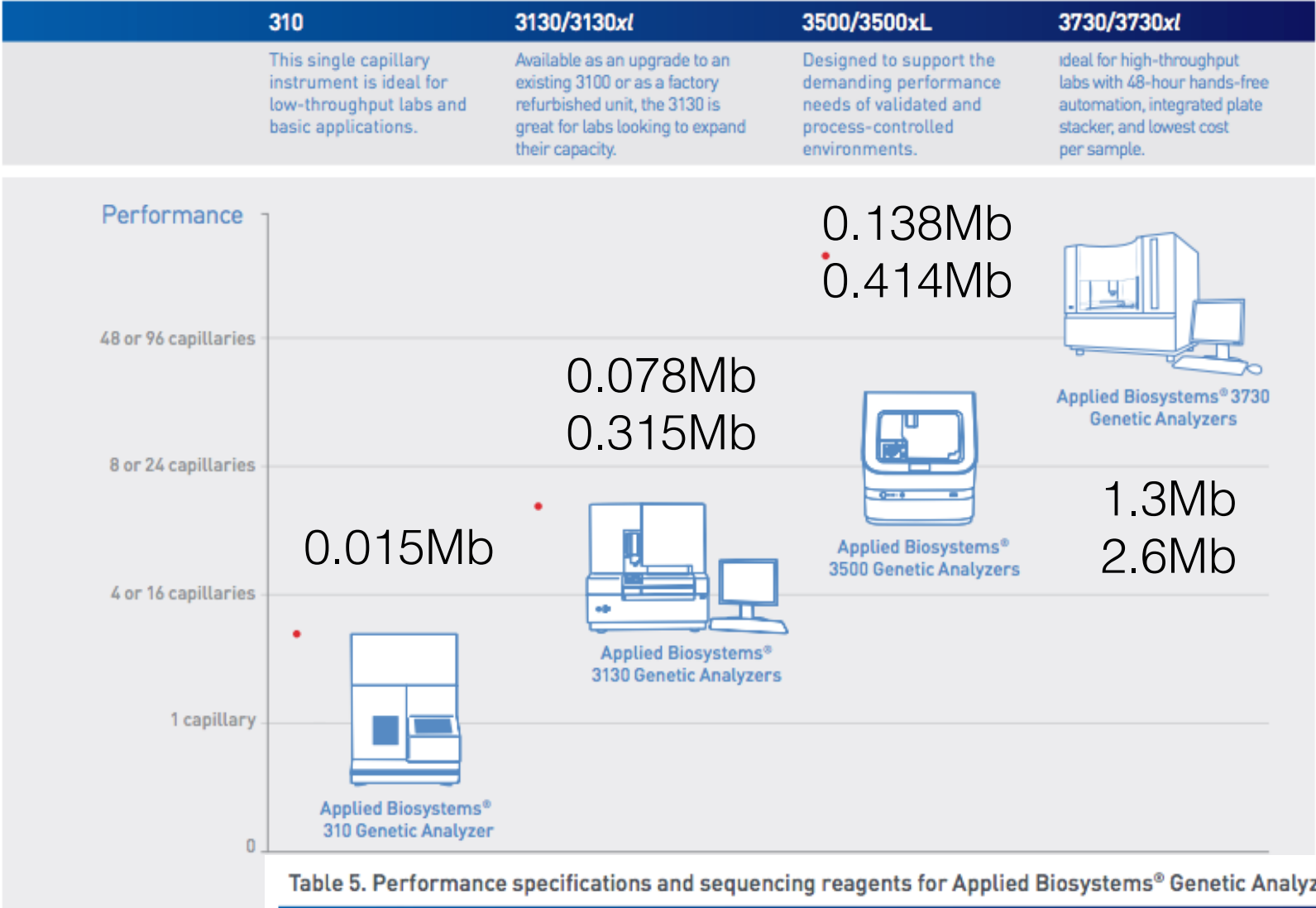
3) Chromatogram Read



GTCACCCCTGAAT

2. First steps: Pre-NGS era

Sanger DNA sequencing systems



Error Rate
0.1%

Error Type
substitution



2. First steps: Pre-NGS era

Next Generation Sequencing vs Sanger

Next Generation Sequencing	Sanger
DNA libraries need to be prepared	Fragment amplification
Direct nucleotide detection based in different methods	Physical fragment separation for detection
Millions to billions of reads	Thousands of reads
Variable size (short and long technologies)	400 to 900 bp read length
Variable error rate	Very low error rate
Quantitative comparison	Semicomparative comparison



2. First steps: Pre-NGS era

Sequencing

Technology	Read length (bp)	Accuracy	Reads/Run	Time/Run	Cost/Mb
Applied Bio 3730XL (Sanger)	400 - 900	99.9%	384	4 h (12 runs/day)	US\$2,400
Roche 454 GS FLX (Pyrosequencing)	700 Single/Pairs	99.9%	1,000,000	24h	US\$10
Illumina HiSeq4000 (Seq. by synthesis)	75-250 Single/Pairs	99%	5,000,000,000	24 to 120 h	\$0.05 to \$0.15
Illumina MiSeq (Seq. by synthesis)	50-300 Single/ Pairs	99%	44,000,000	24 to 72 h	US\$0.17
SOLiD 4 (Seq. by ligation)	25-50 Single/Pairs	99.9%	1,400,000,000	168 h	US\$0.13
ION Torrent (Seq. by semiconductor)	170-400 Single	98%	80,000,000	2 h	US\$2
Pacific Biosciences Sequel (SMRT)	14,000 Single	85% (99.9%)	1,600,000	4 h	US\$0.6
Oxford N. Minion (Nanopore sequencing)	10,000 Single	62% (96%)	4,400,000	48 h	US\$0.02

Short
reads

Long
reads



Outline of Topics

1. Basics about genetics and sequencing
2. First steps: Pre-NGS era
- 3. Short read sequencing technologies**
4. Long read sequencing technologies
5. Common file formats
6. Transcriptome sequencing



3. Short read sequencing technologies

Libraries types

★ Library types (orientations):

- Single reads

F →

- Pair ends (PE) (150-800 bp insert size)

F → ← R

Illumina

- Mate pairs (MP) (2-40 Kb insert size)

R ← → F

Illumina

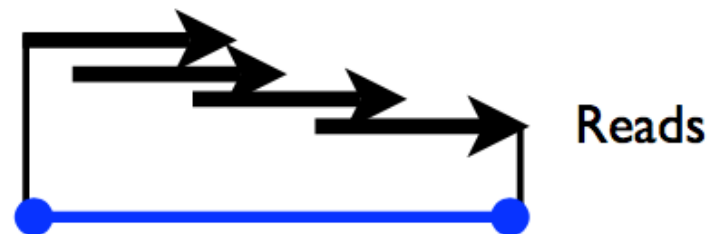
F → ← R

454/Roche

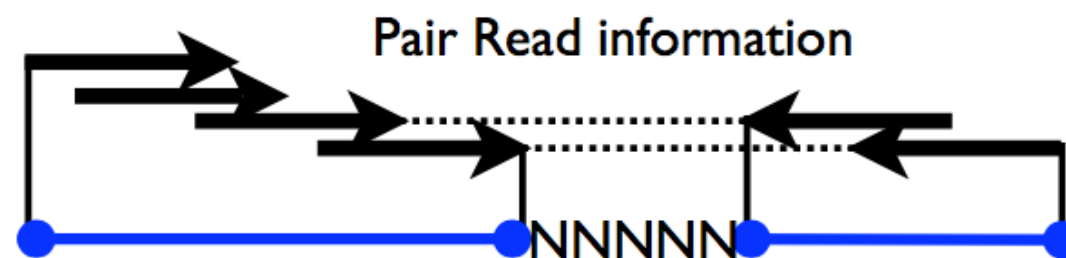
3. Short read sequencing technologies

Libraries types

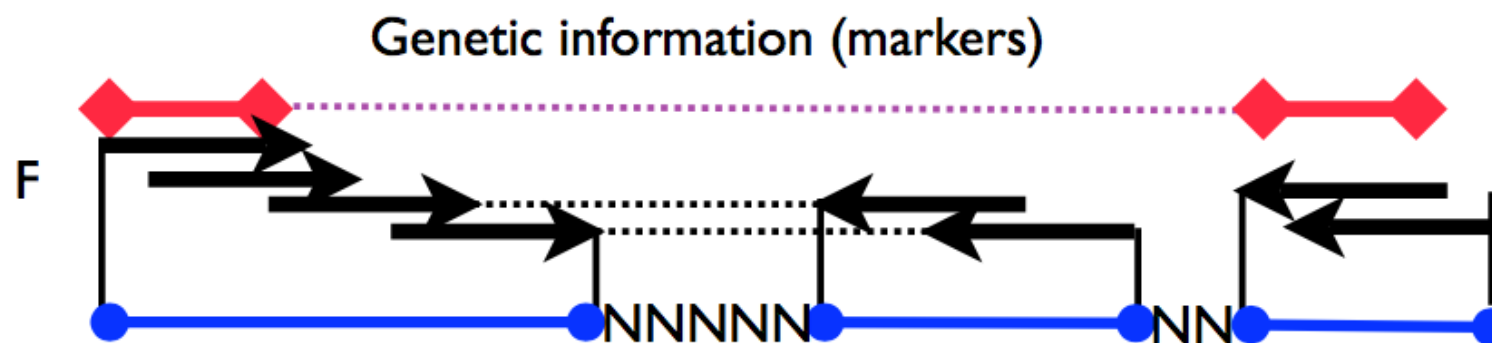
- Why is important the pair information ?
 - *novο* assembly:



Consensus sequence
(Contig)



Scaffold
(or Supercontig)



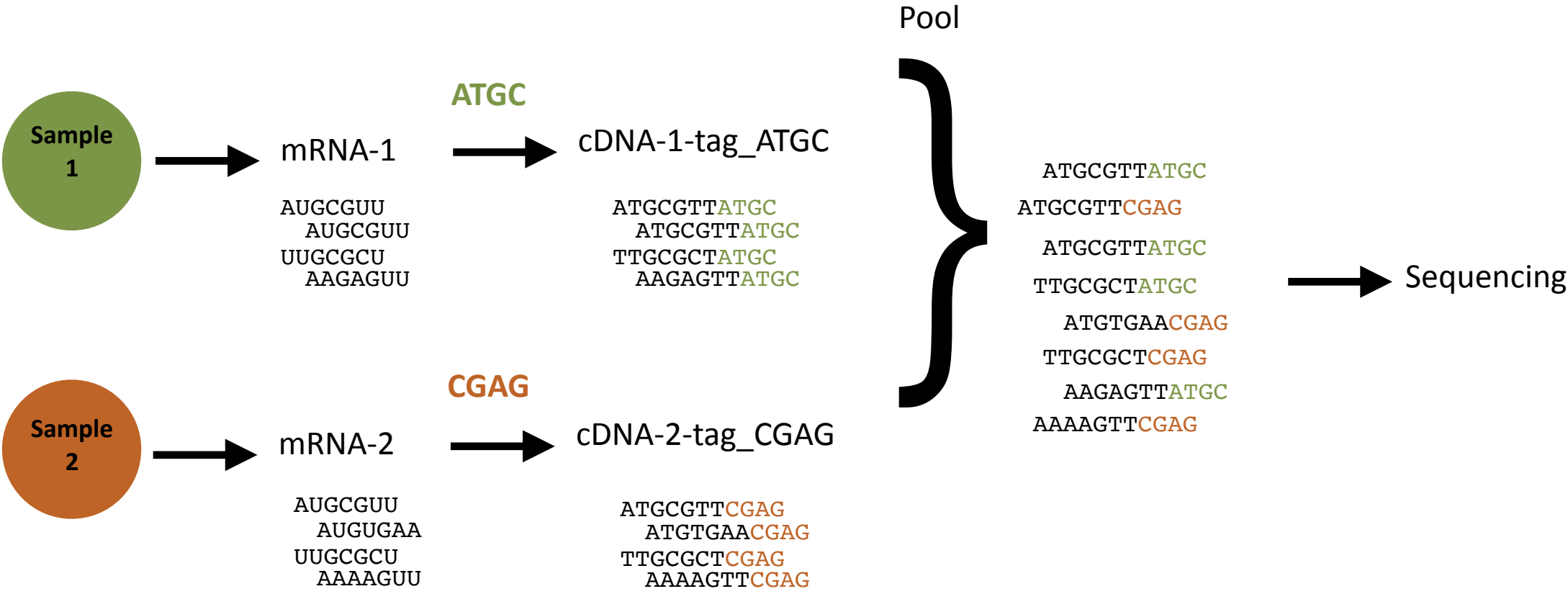
Pseudomolecule
(or ultracontig)

3. Short read sequencing technologies

Libraries types

Multiplexing

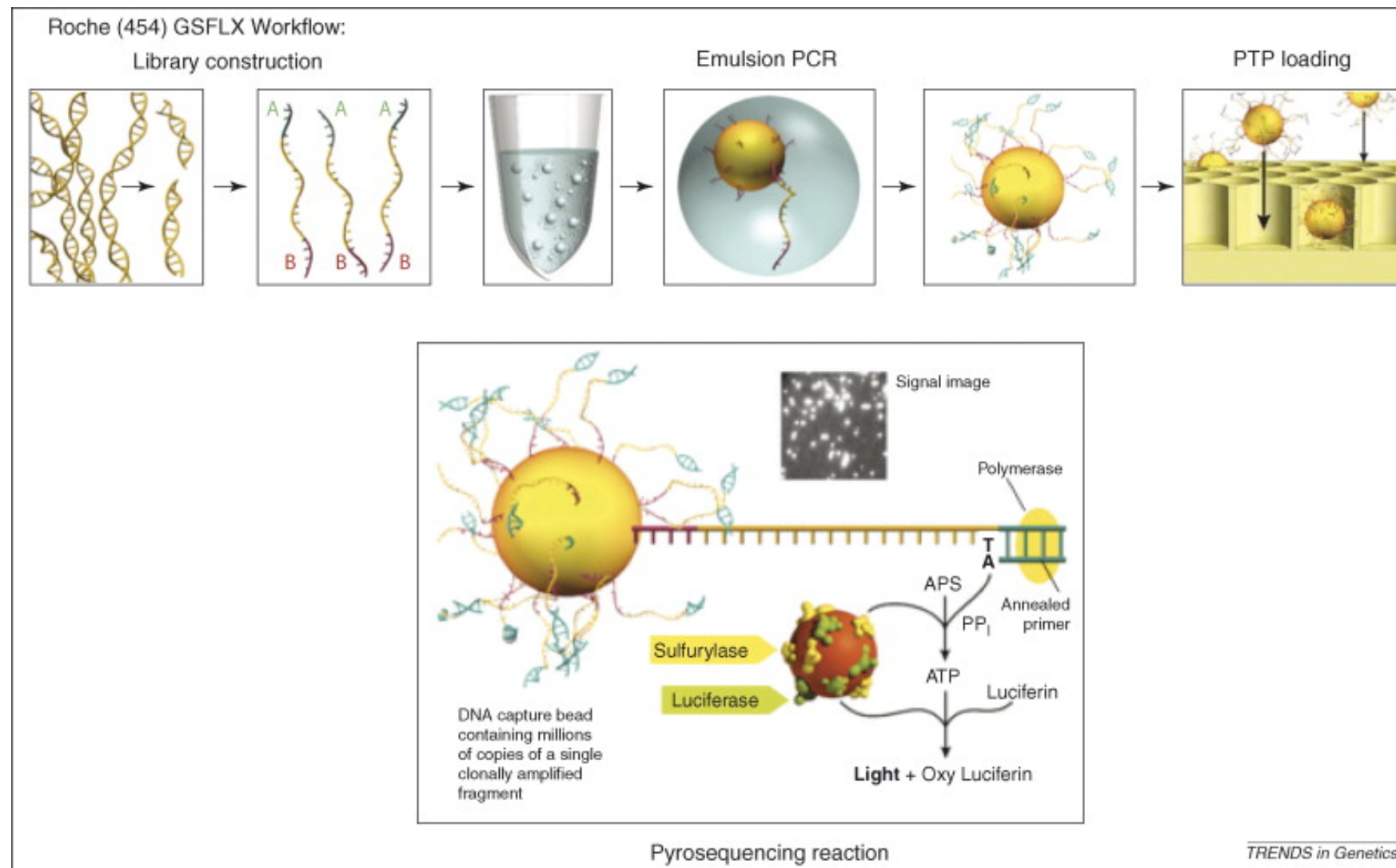
Use of DNA tags (4-7 bp) to identify samples in the same sequencing lane, cell or sector.



3. Short read sequencing technologies

Sequencing systems: 454

Pyrosequencing technology



3. Short read sequencing technologies

Sequencing systems: 454

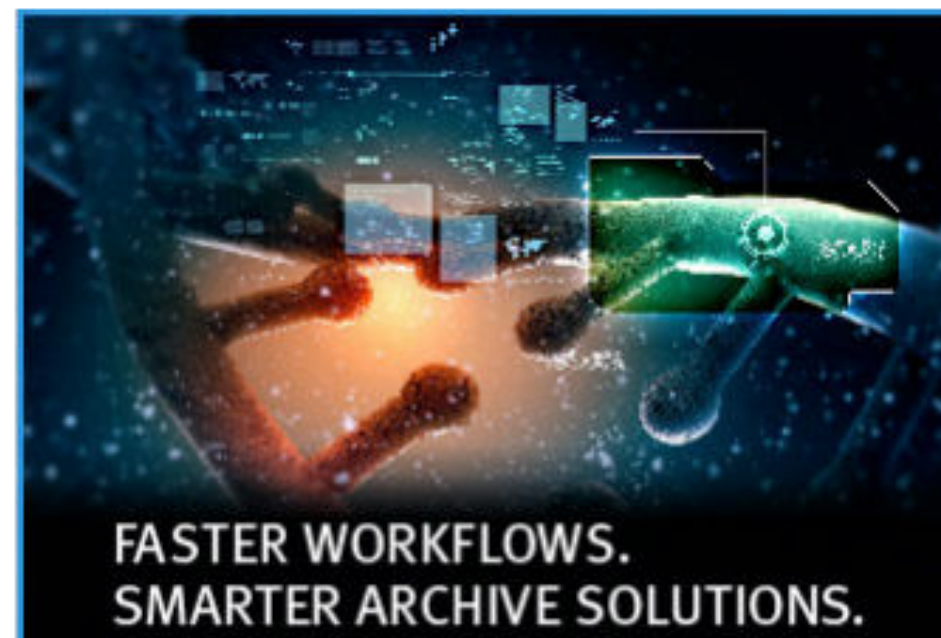
http://www.bio-itworld.com/BioIT_Article.aspx?id=131053



Six Years After Acquisition, Roche Quietly Shuttters 454

By Bio-IT World Staff

October 16, 2013 | This month, Roche began the process of closing its wholly-owned subsidiary 454 Life Sciences, a once-dominant player in next-generation sequencing, and laying off the company's 130 employees. Manufacturing of 454 sequencers will continue through 2015, and the sequencers will continue to be serviced through mid-2016; the layoffs will be phased over this period. This announcement follows a series of downsizing measures from Roche in the area of genetic sequencing over the past year.



3. Short read sequencing technologies

Sequencing systems: Illumina

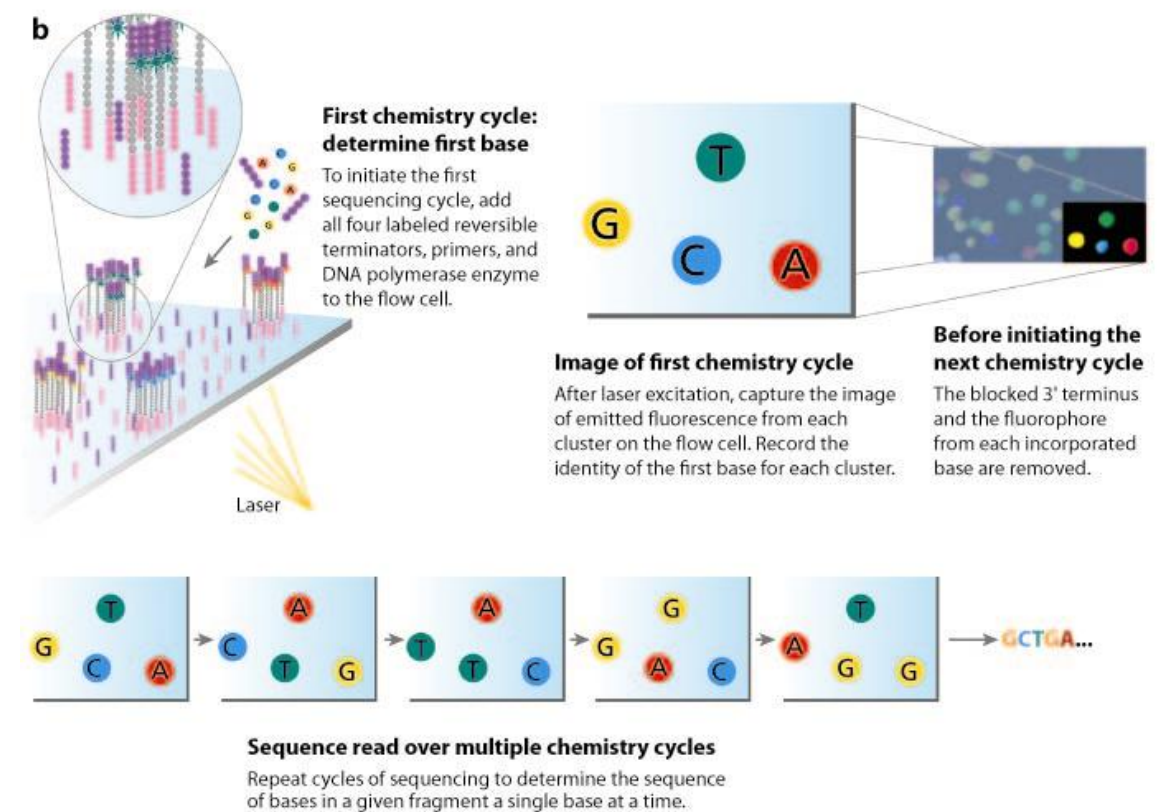
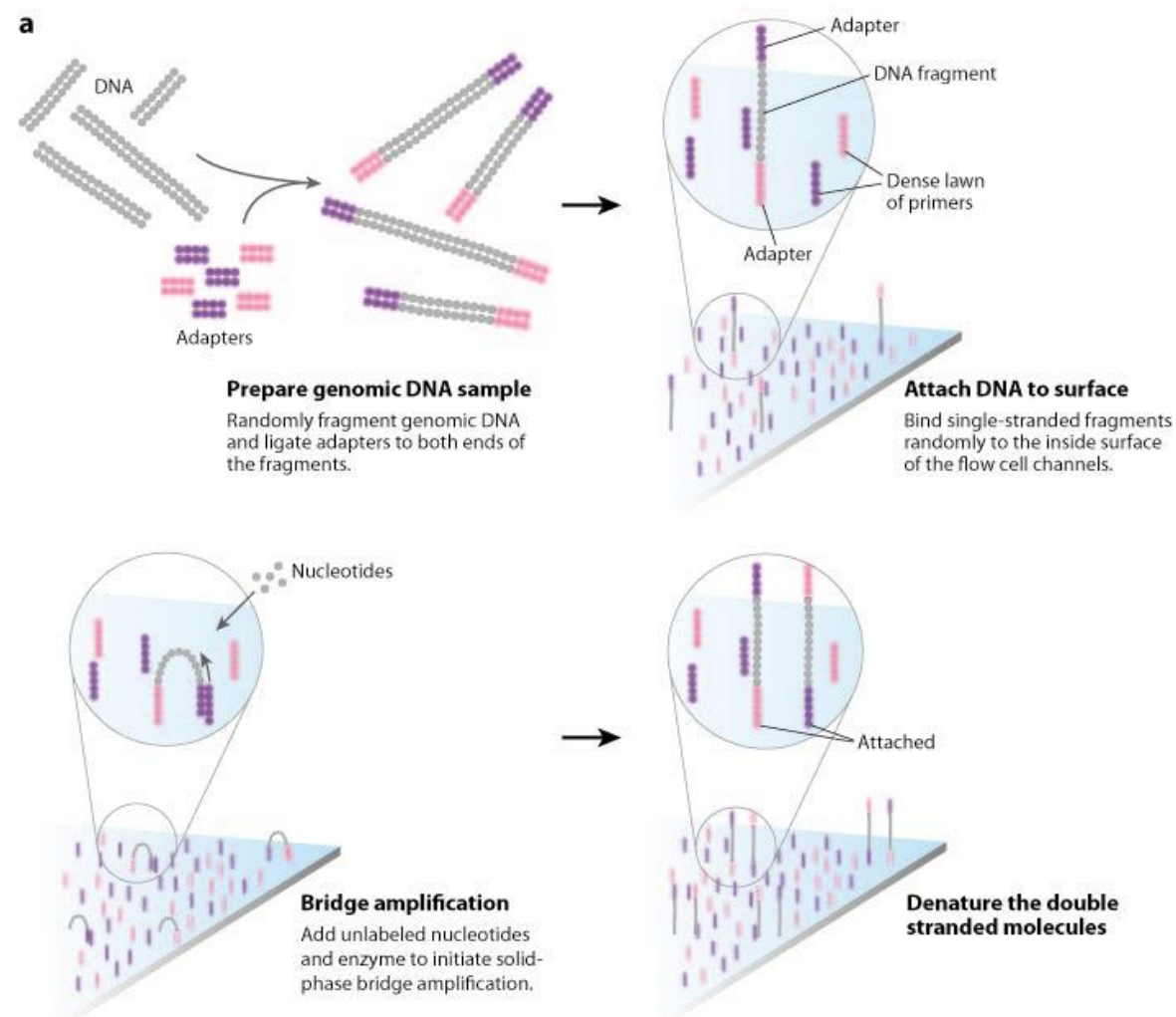
<http://www.illumina.com/>



3. Short read sequencing technologies

Sequencing systems: Illumina

Sequence by Synthesis technology



3. Short read sequencing technologies

Sequencing systems: Illumina

<http://www.illumina.com/techniques/sequencing/dna-sequencing.html#>



3. Short read sequencing technologies

Sequencing systems: Illumina

<https://www.illumina.com/systems/sequencing-platforms.html>

Benchtop systems



MiniSeq System



MiSeq Series +



NextSeq Series +

Run Time	4–24 hours	4–55 hours	12–30 hours
Maximum Output	7.5 Gb	15 Gb	120 Gb
Maximum Reads Per Run	25 million	25 million*	400 million
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp

Production-scale systems



NextSeq Series +



HiSeq Series +



HiSeq X Series†



NovaSeq 6000
System +

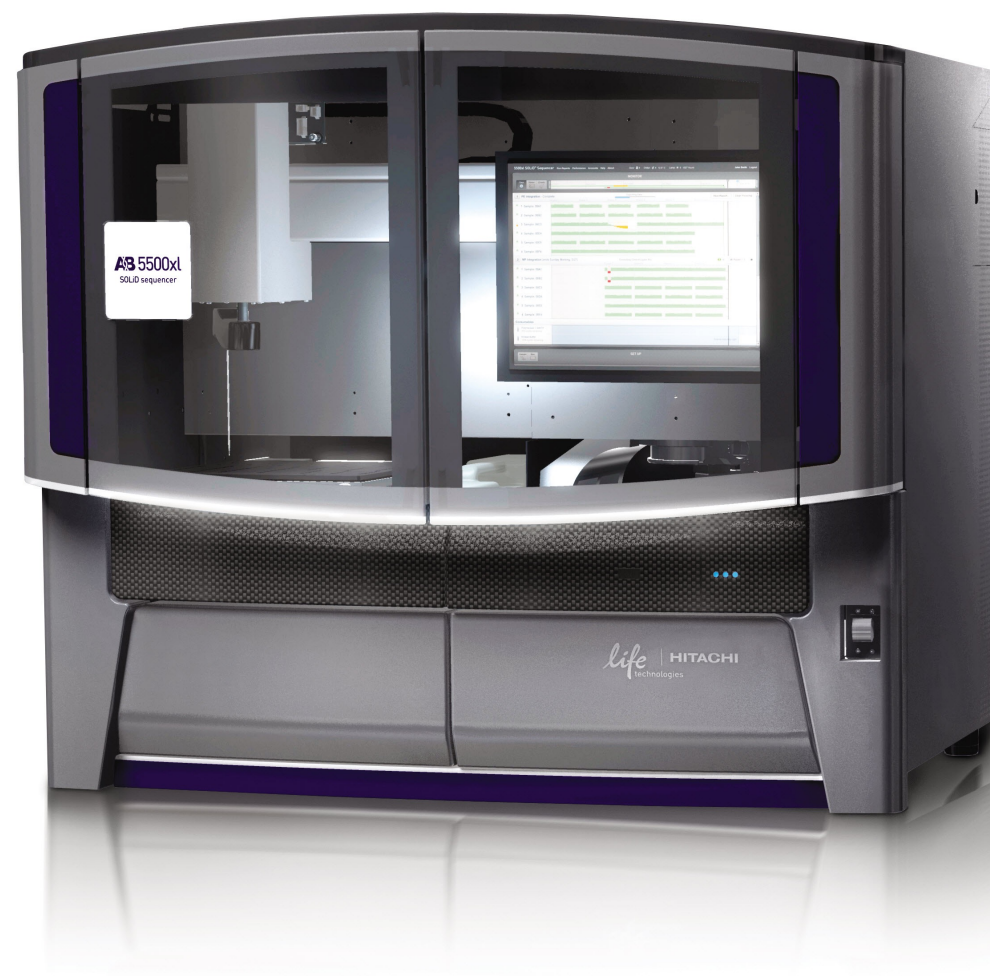
Run Time	12–30 hours	< 1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	< 3 days	16–36 hours (Dual S2 flow cells) 44 hours (Dual S2 flow cells)
Maximum Output	120 Gb	1500 Gb	1800 Gb	6000 Gb [§]
Maximum Reads Per Run	400 million	5 billion	6 billion	20 billion ^l
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp



3. Short read sequencing technologies

Sequencing systems: SOLiD

<https://products.appliedbiosystems.com>



3. Short read sequencing technologies

Sequencing systems: SOLiD

Sequence by Ligation technology

Overview of SOLiD™ Sequencing Chemistry

Library Preparation

1. Prepare one of the two types of libraries (Figure 1) for SOLiD™ System sequencing—fragment or mate-paired. Your choice of library depends on the application you're performing and the information you desire from your experiments.

[View Larger Image](#)

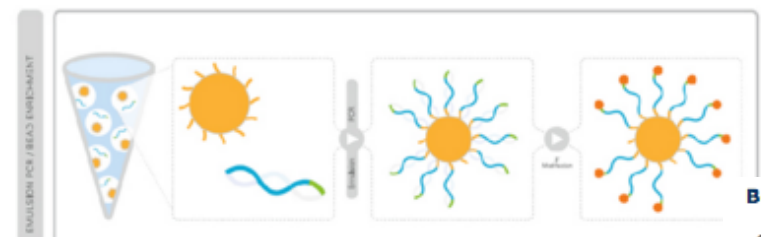


Figure 1

Emulsion PCR/Bead Enrichment

2. Prepare clonal bead populations (Figure 2) in microreactors containing template, PCR reaction components, beads, and primers.
3. After PCR, denature the templates and perform bead enrichment to separate beads with extended templates from undesired beads. The template on the selected beads undergoes a 3' modification to allow covalent attachment to the slide.

[View Larger Image](#)



Bead Deposition

4. Deposit 3' modified beads onto a glass slide (Figure 3). During bead loading, deposition chambers enable you to segment a slide into one, four, or eight sections. A key advantage of the system is the ability to accommodate increasing densities of beads per slide, resulting in a higher level of throughput from the same system.

Figure 2

[View Larger Image](#)

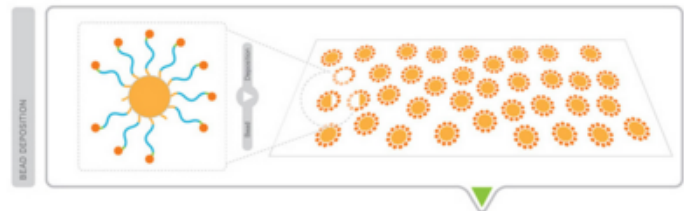
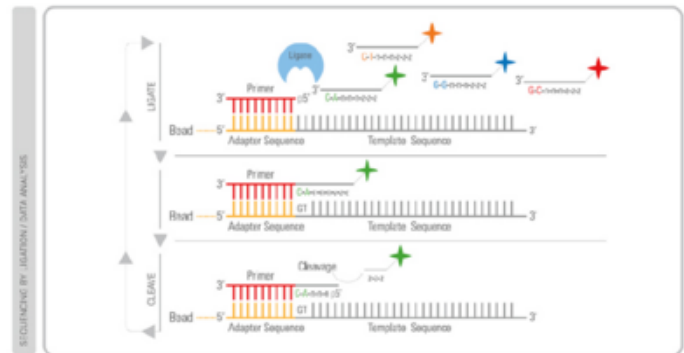


Figure 3

Sequencing by Ligation

5. Primers hybridize to the P1 adapter sequence on the templated beads (Figure 4).
6. A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction.
7. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length.
8. Following a series of ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles.

[View Larger Image](#)



3. Short read sequencing technologies

Sequencing systems: SOLiD

Primer Reset

- Five rounds of primer reset are completed for each sequence tag (Figure 5). Through the primer reset process, virtually every base is interrogated in two independent ligation reactions by two different primers.

For example, the base at read position 5 is assayed by primer number 2 in ligation cycle 2 and by primer number 3 in ligation cycle 1 (see figure at right). This dual interrogation is fundamental to the unmatched accuracy characterized by the SOLiD™ System.

Exact Call Chemistry

- Up to 99.99% accuracy is achieved with the Exact Call Chemistry Module by sequencing with an additional primer using a multi-base encoding scheme.

[View Larger Image](#)

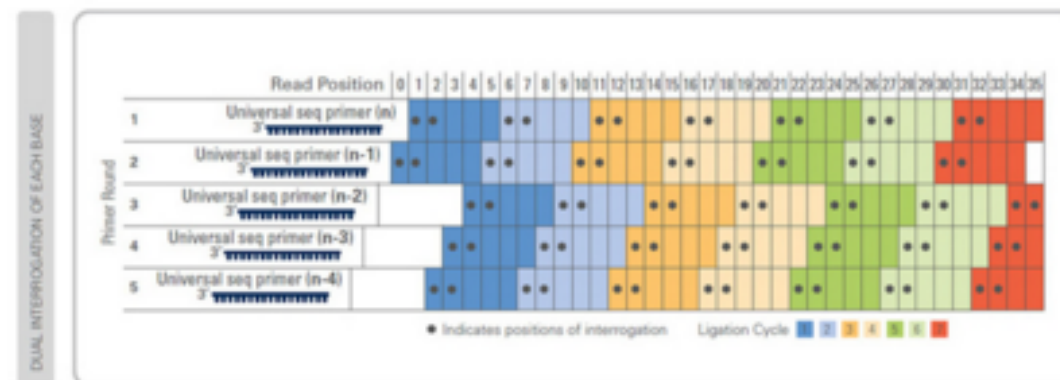


Figure 5

3. Short read sequencing technologies

Sequencing systems: SOLiD

System and features	5500 System (1.0 µm microbeads)	5500xl System (1.0 µm microbeads)	5500xl System (0.75 µm nanobeads available 2nd Half 2011 ¹)
Pay-Per-Lane Sequencing (PPL-Seq™)	Reagent consumption engineered independently for each lane; users pay only for reagent consumables in the active lanes when performing a partial run.		
Application-Per-Lane Sequencing	Independent FlowChip lanes allow you to configure read length of chemistry for each lane enabling multiple applications in a single run.		
System Accuracy with Exact Call Chemistry (ECC) Module²	Up to 99.99%		
Multiplexing	96 barcodes for both RNA and DNA applications		
Independent lanes	1–6 (1 FlowChip)	1–12 (2 FlowChips)	1–12 (2 FlowChips)
Throughput^{3,4}	7–9 Gb/day	10–15 Gb/day	>20 Gb/day
Exomes/run⁵	Up to 8 exomes	Up to 16 exomes	Up to 24 exomes
Transcriptomes/run⁶	Up to 8 transcriptomes	Up to 16 transcriptomes	Coming in 2nd Half 2011
Human genome/run⁷	Up to 1 genome (30X average coverage)	Up to 2 genomes (30X average coverage)	Coming in 2nd Half 2011
Maximum read lengths	Mate-paired: 2 x 60 bp Paired-end: 75 bp x 35 bp Fragment: 75 bp	Mate-paired: 2 x 60 bp Paired-end: 75 bp x 35 bp Fragment: 75 bp	Fragment: 50 bp
Sequencing run type	Yield and run times for 1 lane		
PE 50 bp x 5 bp^{5,8}	1 exome, 2 days		
PE 50 bp x 35 bp^{6,8}	1 transcriptome, 3.5 days		
MP 60 bp x 60 bp⁸	1 human genome (4–5X average coverage), 7 days		



3. Short read sequencing technologies

Sequencing systems: SOLiD

<http://media.invitrogen.com.edgesuite.net/ab/applications-technologies/solid/solid-5500.html>



3. Short read sequencing technologies

Sequencing systems: Ion Torrent

<https://products.appliedbiosystems.com>

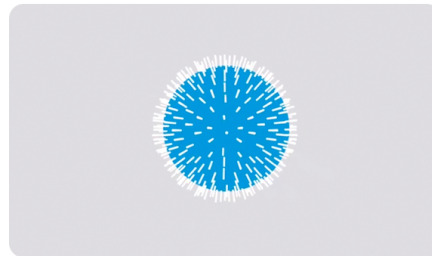


3. Short read sequencing technologies

Sequencing systems: Ion Torrent

Sequence by Semiconductor technology

Copy DNA

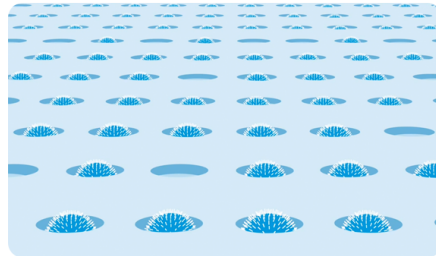


A sample of DNA is cut into millions of fragments, and each fragment is attached to its own bead

The fragment is copied until it covers the bead

This automated process produces millions of beads covered with millions of different fragments

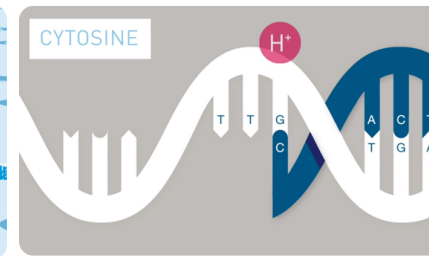
Load chip



The beads are then flowed across the chip, each being deposited into a well

Then the chip is flooded with one of the four nucleotides

Incorporate nucleotide



If the next base on the DNA strand is complementary to this nucleotide, a nucleotide will be incorporated and a hydrogen ion will be released

The hydrogen ion changes the pH of the solution in the well

Detect and call



An ion-sensitive layer beneath the well measures that pH change and converts it to voltage

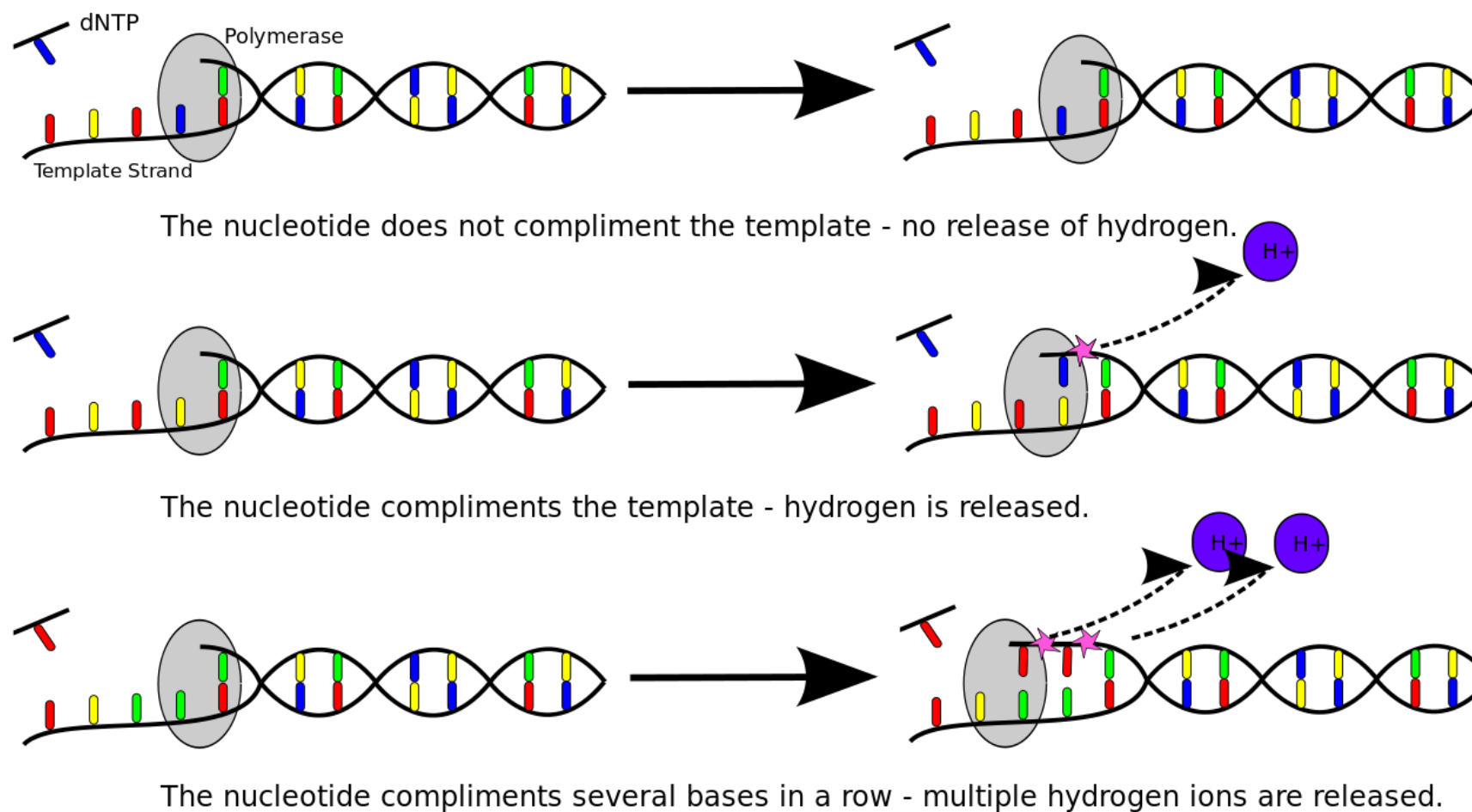
This voltage change is recorded, indicating the nucleotide has been incorporated and the base is called

This process happens simultaneously in millions of wells

3. Short read sequencing technologies

Sequencing systems: Ion Torrent

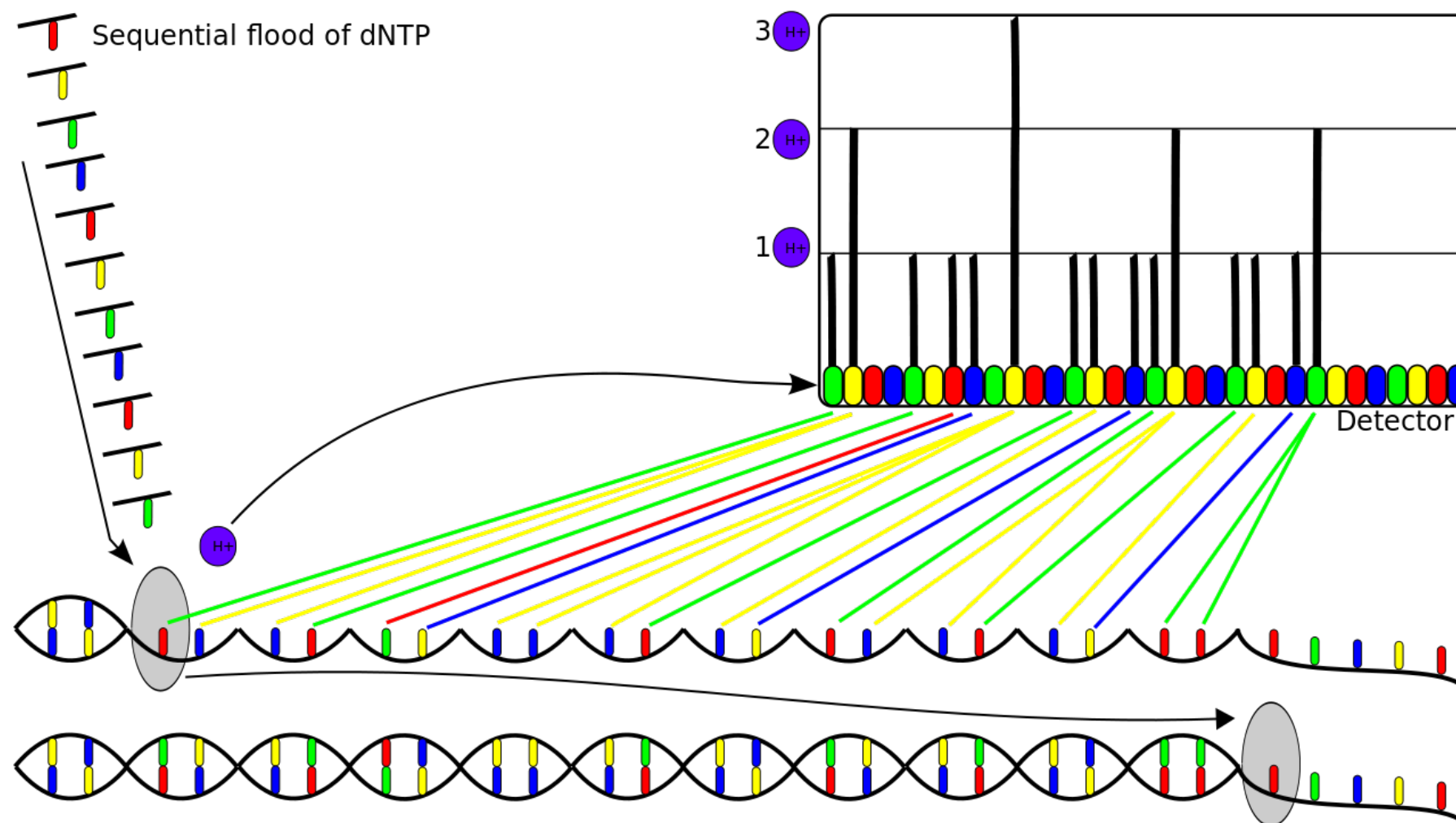
Sequence by Semiconductor technology



3. Short read sequencing technologies

Sequencing systems: Ion Torrent

Sequence by Semiconductor technology



Outline of Topics

1. Basics about genetics and sequencing
2. First steps: Pre-NGS era
3. Short read sequencing technologies
- 4. Long read sequencing technologies**
5. Common file formats
6. Transcriptome sequencing



4. Long read sequencing technologies

Sequencing systems: Pacific Biosystems (PacBio)

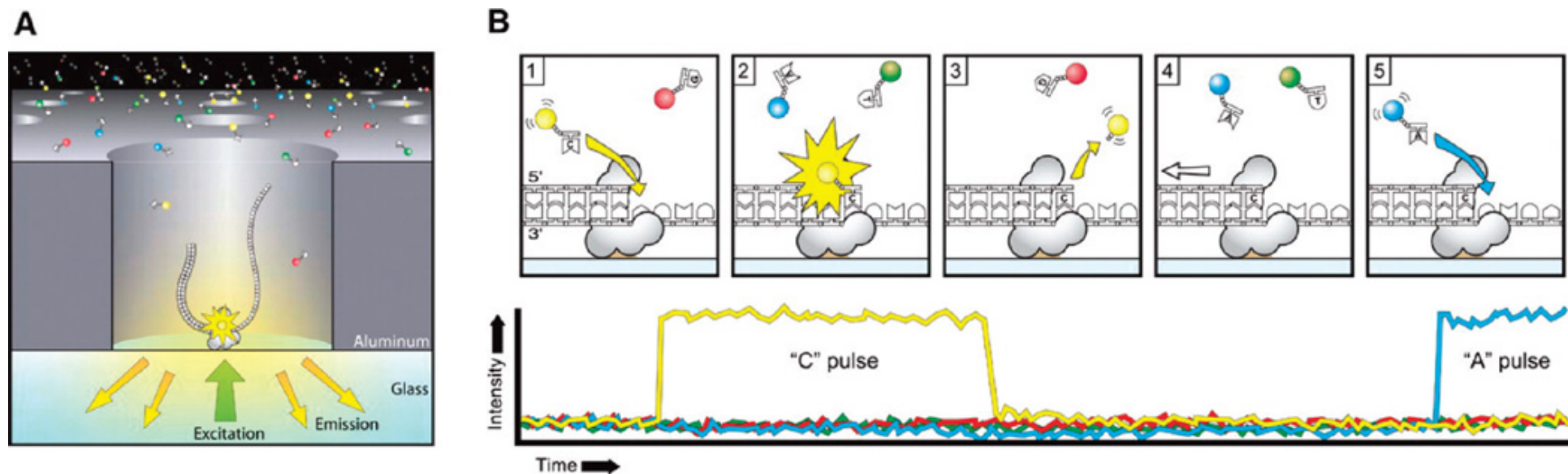
<http://www.pacb.com/>



4. Long read sequencing technologies

Sequencing systems: Pacific Biosystems (PacBio)

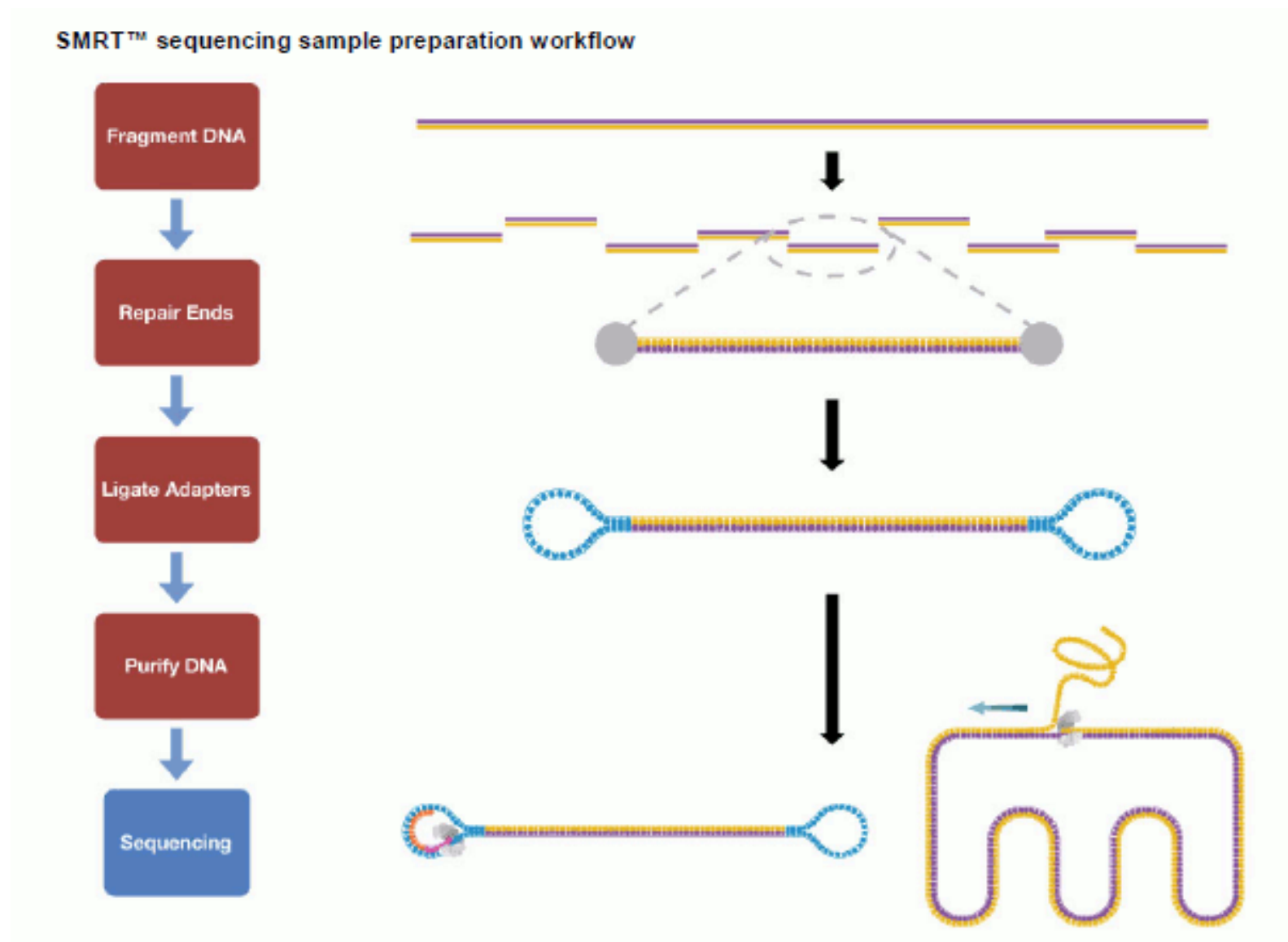
Single Molecule Real Time (SMRT) technology



4. Long read sequencing technologies

Sequencing systems: Pacific Biosystems (PacBio)

Single Molecule Real Time (SMRT) technology



4. Long read sequencing technologies

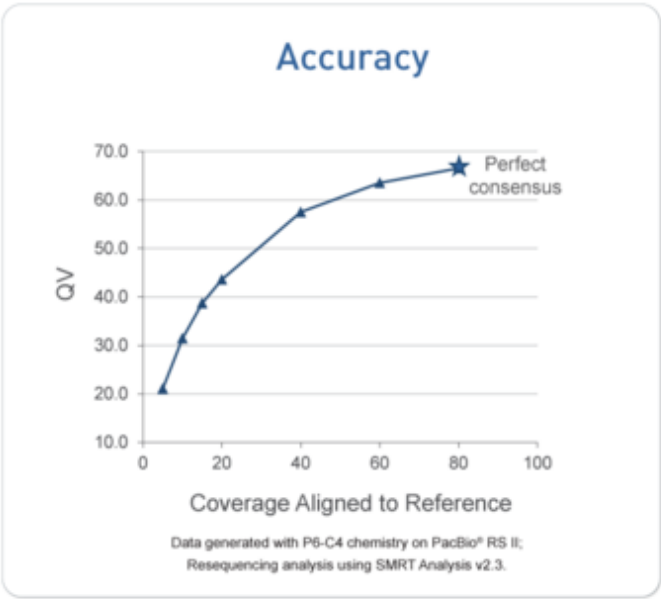
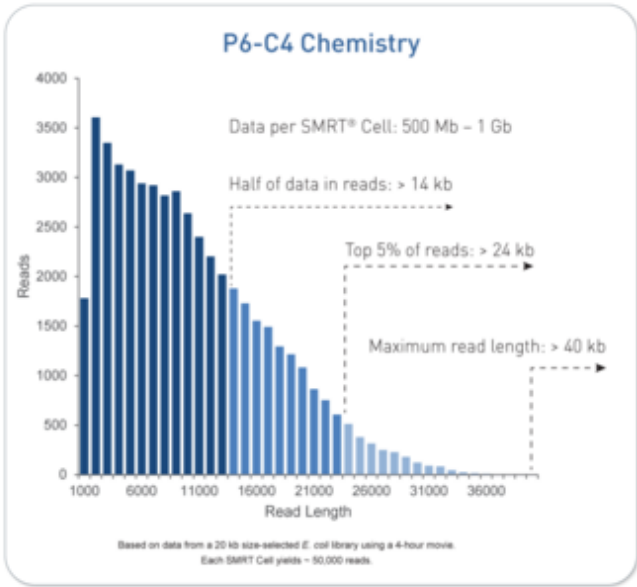
Sequencing systems: Pacific Biosystems (PacBio)

Single Molecule Real Time (SMRT) technology

PacBio RS II smrtcell specifications (P6-C4 chemistry and 4-hours movie) *

Total Yield	Reads	Average Read Length	Average Subread Length
>500 Mb/smrtcell	>50,000/smrtcell	~10 kb	~7 kb *

* Specification based on a 15-20 kb insert library. Average subread length can vary depending on insert size and DNA input quality.



4. Long read sequencing technologies

Sequencing systems: Oxford Nanopore (ON)

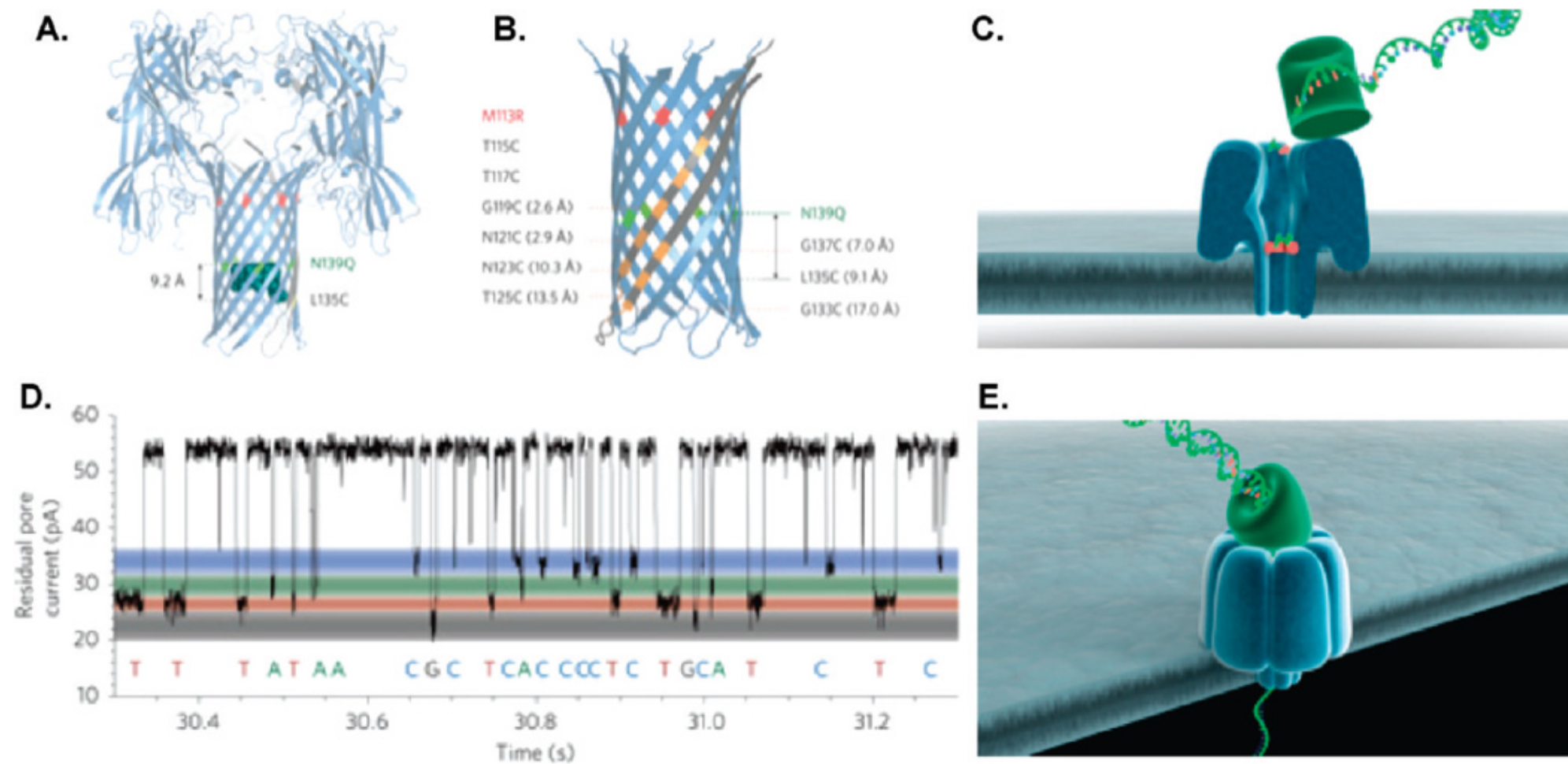
<https://www.nanoporetech.com/>



4. Long read sequencing technologies

Sequencing systems: Oxford Nanopore (ON)

Sequence by Nanopore technology



4. Long read sequencing technologies

Sequencing systems: Oxford Nanopore (ON)

Sequence by Nanopore technology

Specifications

	MinION	PromethION	
			
	Mk 1 MinION	Single PromethION Flow Cell	PromethION (48 Flow Cells)
			

System Operation

Run time ⁴	1 minute - 48 hours	1 minute - 48 hours	1 minute - 48 hours
Flow cell lifetime ⁴	~72hrs	>= 72hrs	>= 72hrs
Time to first usable read (data available in real time)	2 minutes	2 minutes	2 minutes
Number of reads at 10Kb at standard speed (70bps) ⁴	Up to 600,000	N/A	N/A
Number of reads at 10kb in Fast Mode (500bps) ⁴	Up to 4.4M	Up to 26M	Up to 1250M
Read Length	Read length = fragment length Longest reported between 230-300 Kilobases (1D)	Read length = fragment length Longest reported between 230-300 Kilobases (1D)	Read length = fragment length Longest reported between 230-300 Kilobases (1D)
1D Yield ⁵ at 70 bps in 48 hours	Up to 6 Gb	N/A	N/A
1D Yield ⁵ at 500 bps in 48 hours	up to 42 Gb	up to 256 Gb	up to 12 Tb
Base calling accuracy ⁶	up to 96%	up to 96%	up to 96%



4. Long read sequencing technologies

Sequencing systems: Oxford Nanopore (ON)

Original Article

Assessing the performance of the Oxford Nanopore Technologies MinION

T. Laver^{a, 1}, J. Harrison^{a, 1}, P.A. O'Neill^{a, b}, K. Moore^{a, b}, A. Farbos^{a, b}, K. Paszkiewicz^{a, b}, D.J. Studholme^a

[Show more](#)

doi:10.1016/j.bdc.2015.02.001

[Get rights and content](#)

Open Access funded by Biotechnology and Biological Sciences Research Council

Under a Creative Commons license

Abstract

The Oxford Nanopore Technologies (ONT) MinION is a new sequencing technology that potentially offers read lengths of tens of kilobases (kb) limited only by the length of DNA molecules presented to it. The device has a low capital cost, is by far the most portable DNA sequencer available, and can produce data in real-time. It has numerous prospective applications including improving genome sequence assemblies and resolution of repeat-rich regions. Before such a technology is widely adopted, it is important to assess its performance and limitations in respect of throughput and accuracy. In this study we assessed the performance of the MinION by re-sequencing three bacterial genomes, with very different nucleotide compositions ranging from 28.6% to 70.7%: the high G + C strain was underrepresented in the sequencing reads. We estimate the error rate of the MinION (after base calling) to be 38.2%. Mean and median read lengths were 2 kb and 1 kb respectively, while the longest single read was 98 kb.



4. Long read sequencing technologies

Sequencing systems: Oxford Nanopore (ON)

LARGE-SCALE BIOLOGY ARTICLE

***De novo* Assembly of a New *Solanum pennellii* Accession Using Nanopore Sequencing**

Maximilian H.-W. Schmidt^{1,§}, Alexander Vogel^{1,§}, Alisandra K. Denton^{1,§}, Benjamin Istace², Alexandra Wormit¹, Henri van de Geest^{3,+}, Marie E. Bolger⁴, Saleh Alseekh⁵, Janina Maß⁴, Christian Pfaff⁴, Ulrich Schurr⁴, Roger Chetelat⁶, Florian Maumus⁷, Jean-Marc Aury², Sergey Koren⁸, Alisdair R. Fernie⁵, Dani Zamir⁹, Anthony M. Bolger¹, Björn Usadel^{1,4,*}

¹Institute for Botany and Molecular Genetics, BioEconomy Science Center, RWTH Aachen University, Aachen, Germany

²Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Genoscope, 2 rue Gaston Crémieux, 91057 Evry, France

³Wageningen Plant Research, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

⁴Institute for Bio- and Geosciences (IBG-2: Plant Sciences), Forschungszentrum Jülich, Jülich, Germany

⁵Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, Potsdam-Golm, Germany

⁶C. M. Rick Tomato Genetics Resource Center, Department of Plant Sciences, University of California, Davis, California 95616

⁷URGI, INRA, Université Paris-Saclay, 78026 Versailles, France



Outline of Topics

1. Basics about genetics and sequencing
2. First steps: Pre-NGS era
3. Short read sequencing technologies
4. Long read sequencing technologies
- 5. Common file formats**
6. Transcriptome sequencing



5. Common file formats

I. FASTA

FASTA format is a text based file format that store three different types: DNA, RNA or protein sequences. Used to represent the information for sequences for genomes, mRNA's, cDNA's, miRNA's...



5. Common file formats

II. FASTQ

FASTQ format is a text based file format that store usually DNA sequences. It contains information about the sequencing QUALITY of each nucleotide.

*ID line always starts
with “@”*

sequence can be one or more lines

One line ID

quality line always starts with “+”

@GWNJ-0957:89:GW170928504:7:1101:2757:1309 1:N:0:NCGTCCC
TATCTAAGTATTTGATTAATGATAGATGACGATGGAGAAATATAATCTACTTTTTTAAGTCCCTCATTTTC
TTTCTCCATCTTTCTTTTTTTATTACTCCCATTTGTTCCCCAT
+
AAAAAFFJJFJJFJJAAAAFJJJ<FJJJJJJJJJJ7<7<<<JJJJJJFFJJJAFJF-7<-7AFJJFJJJ
JJJJJAJJFJFJ<7<-7A-7FAFJA777777<7-7--7--7
@GWNJ-0957:89:GW170928504:7:1101:3549:1309 1:N:0:NCGTCCC
ACCATTTCATTATTTTTTTTATTAGTCTTTATTACTTTTCTTCCTTCCTTCTGAAATACTGCTATTGTACA
TAAACAAAATGATCTACTTAAAAATAAAACAAATTTAAAA
+
AAA-AAJJFJJAAJAA-7AFJJ-7-<<-<AJJ--<J-<-<---77F7-A---A7--
<777<7<7<<F-77F<J<JJ7F7AFF77<77<7777<77<---7---77---7---

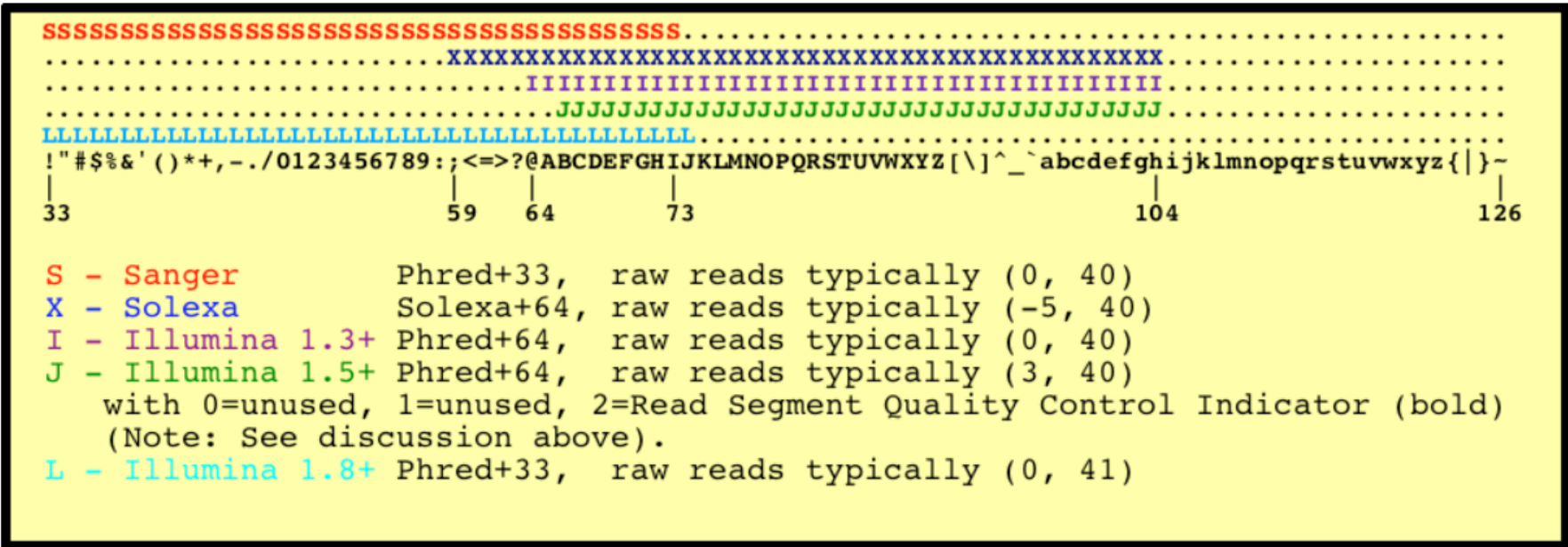
One quality character per nucleotide. Each character code a number from 0-41 (Illumina v1.8+).



5. Common file formats

II. FASTQ

FASTQ format is a text based file format that store usually DNA sequences. It contains information about the sequencing QUALITY of each nucleotide.



Phred score of a base is: $Q_{\text{phred}} = -10 \log_{10}(e)$

Q=15	e=0.03	(min. used Sanger)
Q=20	e=0.01	(min. used 454 and Illumina)
Q=30	e=0.001	(standard used 454)



5. Common file formats

III. SAM/BAM

SAM (and its binary form BAM) format is designed to store read mapping information to a reference. It has 11 columns.

Coord
ref

12345678901234 5678901234567890123456789012345

AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1
+r002
+r003
+r004
-r003
-r001/2

TTAGATAAAGGATA*CTG
aaaAGATAA*GGATA
gcctaAGCTAA
ATAGCT.....TCAGC

Alignment

Sam file

r001	163	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA	*
r003	0	ref	9	30	5H6M	*	0	0	AGCTAA	* NM:i:1
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*
r003	16	ref	29	30	6H5M	*	0	0	TAGGC	* NM:i:0
r001	83	ref	37	30	9M	=	7	-39	CAGCGCCAT	*

Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL



5. Common file formats

III. SAM/BAM

SAM (and its binary form BAM) format is designed to store read mapping information to a reference. It has 11 columns.

Coord
ref

12345678901234 5678901234567890123456789012345

AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1
+r002
+r003
+r004
-r003
-r001/2

TTAGATAAAGGATA*CTG
aaaAGATAA*GGATA
gcctaAGCTAA
ATAGCT.....TCAGC

Alignment

Sam file

r001	163	ref	7	30	8M2I4M1D3M	=	37	39	TTAGATAAAGGATACTG	*
r002	0	ref	9	30	3S6M1P1I4M	*	0	0	AAAAGATAAGGATA	*
r003	0	ref	9	30	5H6M	*	0	0	AGCTAA	* NM:i:1
r004	0	ref	16	30	6M14N5M	*	0	0	ATAGCTTCAGC	*
r003	16	ref	29	30	6H5M	*	0	0	TAGGC	* NM:i:0
r001	83	ref	37	30	9M	=	7	-39	CAGCGCCAT	*

Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL



5. Common file formats

III. SAM/BAM

SAM (and its binary form BAM) format is designed to store read mapping information to a reference. It has 11 columns.

The 2nd column: FLAG defines the status of the read mapping.

Col	Field	Bit	Description
1	QNAME	0x1	template having multiple segments in sequencing
2	FLAG	0x2	each segment properly aligned according to the aligner
3	RNAME	0x4	segment unmapped
4	POS	0x8	next segment in the template unmapped
5	MAPQ	0x10	SEQ being reverse complemented
6	CIGAR	0x20	SEQ of the next segment in the template being reversed
7	RNEXT	0x40	the first segment in the template
8	PNEXT	0x80	the last segment in the template
9	TLEN	0x100	secondary alignment
10	SEQ	0x200	not passing quality controls
11	QUAL	0x400	PCR or optical duplicate

- ▶ Flag = 4 means 0x4 read unmapped
- ▶ Flag = 16 means 0x10 read reverse strand
- ▶ Flag = 83 means 0x1 read paired, 0x2 read mapped proper pair, 0x10 read reverse strand and 0x40 first in pair

<http://picard.sourceforge.net/explain-flags.html>



5. Common file formats

III. SAM/BAM

SAM (and its binary form BAM) format is designed to store read mapping information to a reference. It has 11 columns.

The 2nd column: FLAG defines the status of the read mapping.

Col	Field	Bit	Description
1	QNAME	0x1	template having multiple segments in sequencing
2	FLAG	0x2	each segment properly aligned according to the aligner
3	RNAME	0x4	segment unmapped
4	POS	0x8	next segment in the template unmapped
5	MAPQ	0x10	SEQ being reverse complemented
6	CIGAR	0x20	SEQ of the next segment in the template being reversed
7	RNEXT	0x40	the first segment in the template
8	PNEXT	0x80	the last segment in the template
9	TLEN	0x100	secondary alignment
10	SEQ	0x200	not passing quality controls
11	QUAL	0x400	PCR or optical duplicate

- ▶ Flag = 4 means 0x4 read unmapped
- ▶ Flag = 16 means 0x10 read reverse strand
- ▶ Flag = 83 means 0x1 read paired, 0x2 read mapped proper pair, 0x10 read reverse strand and 0x40 first in pair

<http://picard.sourceforge.net/explain-flags.html>

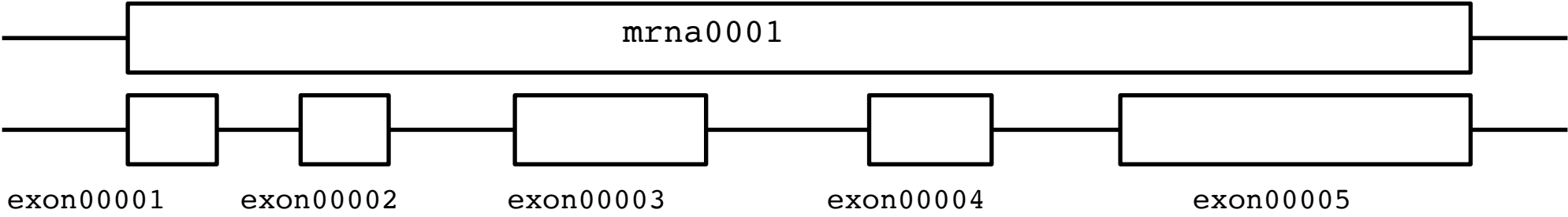


5. Common file formats

IV. GFF3

GFF3 is a text-based file with 9 columns. It is designed to store genomic features (e.g. genes, exons, repetitive elements...) information. More information at <http://gmod.org/wiki/GFF3>.

##gff-version 3								
ctg13	.	mRNA	1300	9000	.	+	.	ID=mrna0001;Name=GDR1
ctg13	.	exon	1300	1500	.	+	.	ID=exon00001;Parent=mrna0001
ctg13	.	exon	1600	1800	.	+	.	ID=exon00002;Parent=mrna0001
ctg13	.	exon	3000	3900	.	+	.	ID=exon00003;Parent=mrna0001
ctg13	.	exon	5000	5500	.	+	.	ID=exon00004;Parent=mrna0001
ctg13	.	exon	7000	9000	.	+	.	ID=exon00005;Parent=mrna0001
seqid	source	type	start	end	score	strand	phase	attributes



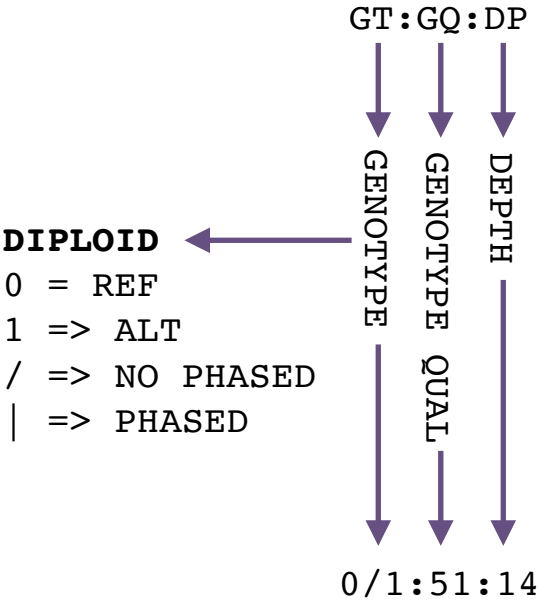
5. Common file formats

V. VCF

VCF is a text-based file with 8 fixed columns and one extra per sample for the multisample files. It contains metadata at the beginning of the file as “#” explaining the different fields

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	
20	1370	rs01	G	A	29	PASS	DP=14;AF=0.5	GT:GQ:DP	0/1:51:14	E.g. 1
20	1730	.	T	A	3	q10	DP=11;AF=0.2	GT:GQ:DP	0/1:58:11	E.g. 2
20	2121	rs02	A	G,T	67	PASS	DP=10;AF=0.5	GT:GQ:DP	1/2:23:10	E.g. 3
20	6781	.	T	.	47	PASS	DP=13;AF=1	GT:GQ:DP	1/1:56:13	E.g. 4

- E.g. 1 is a biallelic heterozygous SNP.
- E.g. 2 is a biallelic heterozygous SNP with low quality, probably because the mapping.
- E.g. 3 is a non-biallelic heterozygous SNP.
- E.g. 4 is a biallelic homozygous Deletion



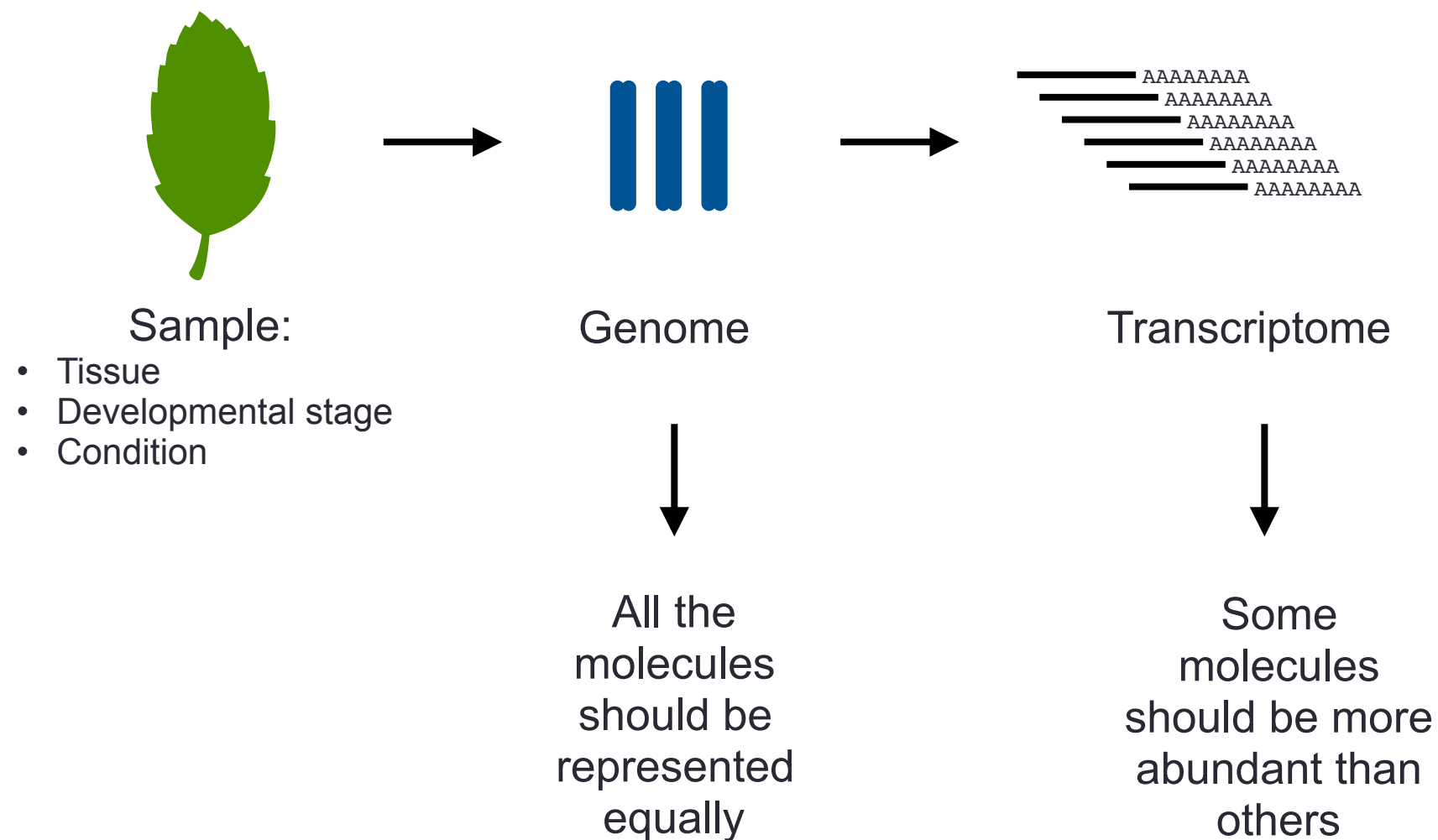
Outline of Topics

1. Basics about genetics and sequencing
2. First steps: Pre-NGS era
3. Short read sequencing technologies
4. Long read sequencing technologies
5. Common file formats
- 6. Transcriptome sequencing**



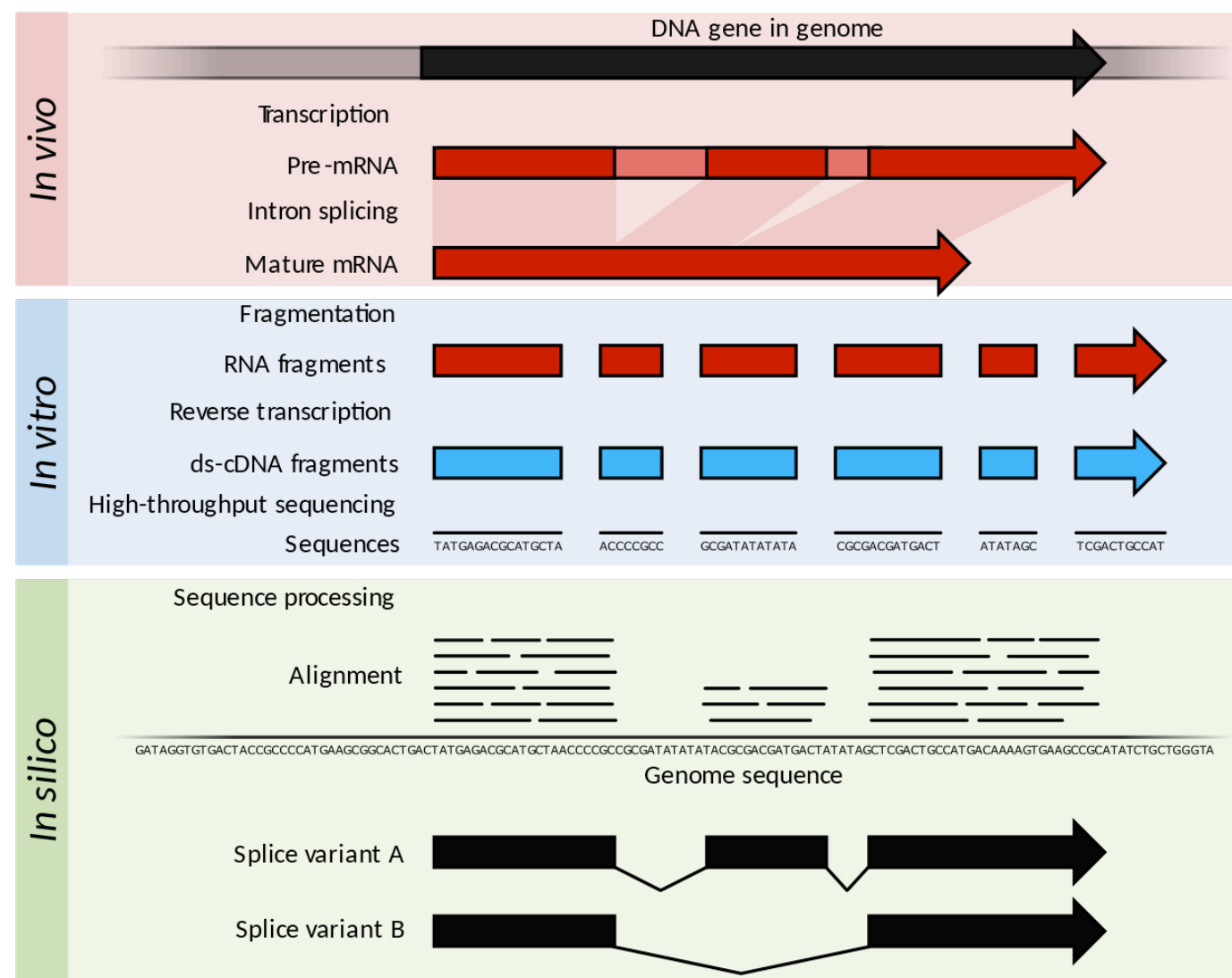
6. Transcriptome Sequencing

Transcriptome: Set of genes transcribed in a specific moment. The transcribed genome for a specific cell type/tissue at some specific developmental stage and under some specific conditions



6. Transcriptome Sequencing

RNA-Seq (named as an abbreviation of "RNA sequencing") is a particular technology-based sequencing technique which uses **next-generation sequencing (NGS)** to reveal the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome



6. Transcriptome Sequencing

RNA-Seq studies have several functions:

- Different RNA populations can be analysed such as total RNA, mRNA, lncRNA, small RNA such miRNA...
- The RNA can be analysed at different moments such as at the end of the transcription or during the transcription using ribosomal profiling.
- Different forms can be analysed (alternative splicing).
- The analysis can be qualitative (identity of the transcripts) or quantitative (measure the amount of transcripts).



6. Transcriptome Sequencing

Transcriptome library preparation has some important differences with the genomic DNA sequencing.

- The RNA has to be isolated. The RNA population depends of the specific moment in which is extracted. When the RNA integrity is affected, low represented transcripts may disappear.
- The molecules of interest has to be depleted (e.g. mRNA filtering out rRNA and tRNA).
- In most of the cases the mRNA has to be transformed in cDNA before sequencing, which introduce some biases such as the ones depending of the retrotranscription efficiency. Some technologies such as Oxford Nanopore can be used to sequence directly the mRNA.

