Genomics and Transcriptomics

Class 03 - Next Generation Sequencing

INSTRUCTOR:
Aureliano Bombarely
Department of Bioscience
Universita 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

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

1. Basics about genetics and sequencing

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

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

1. Basics about genetics and sequencing

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

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

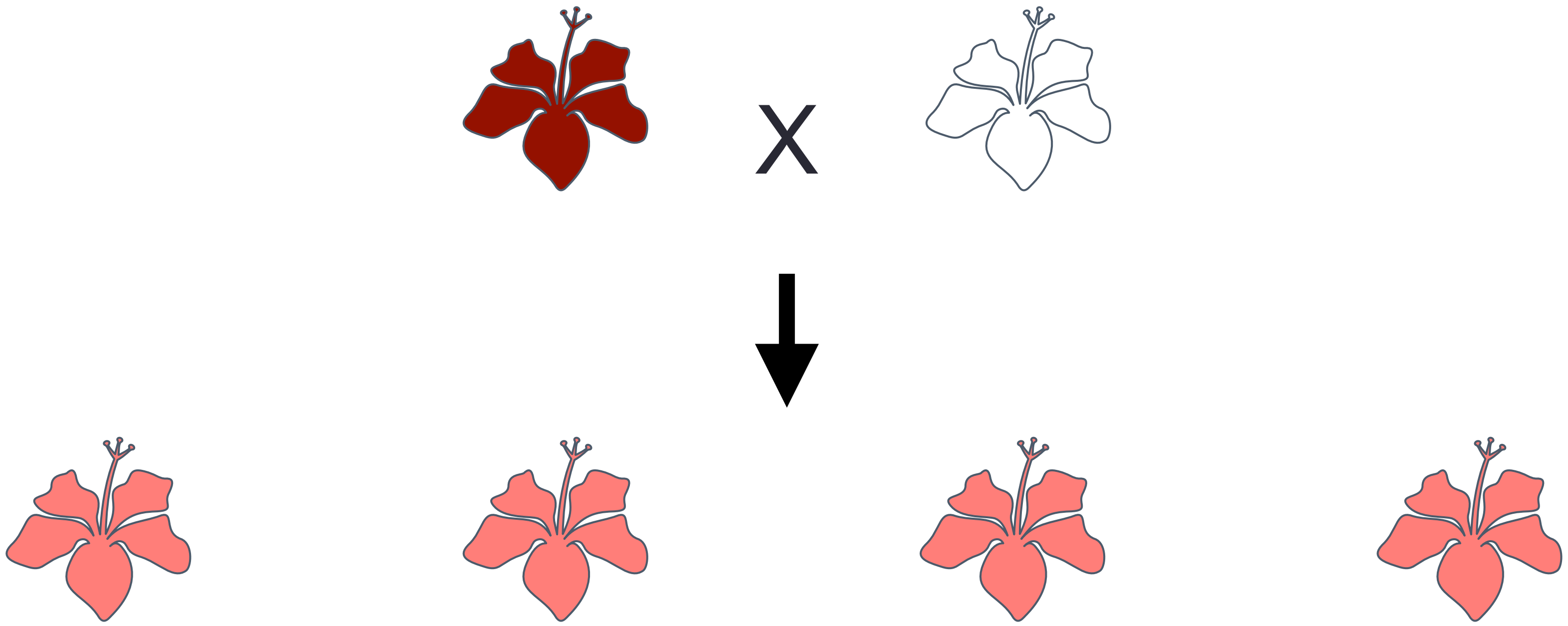
1. Basics about genetics and sequencing

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

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

1. Basics about genetics and sequencing

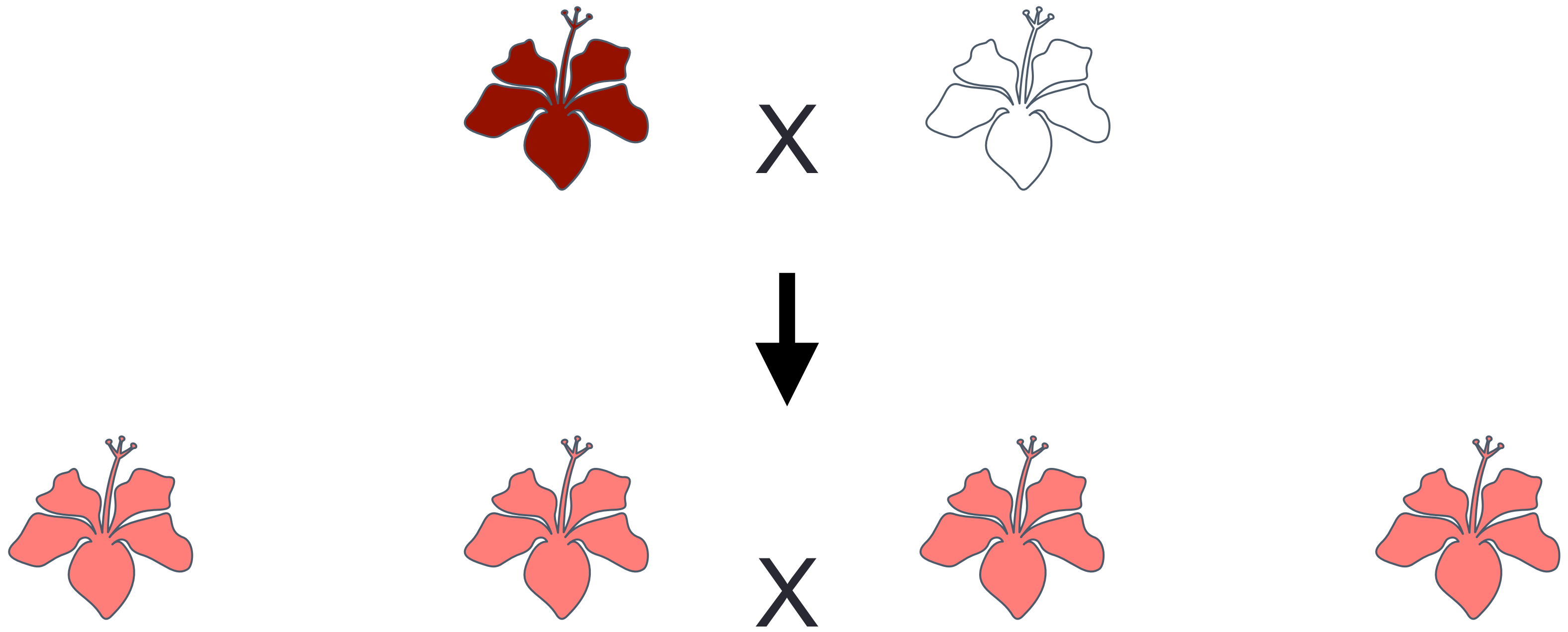
Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



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

1. Basics about genetics and sequencing

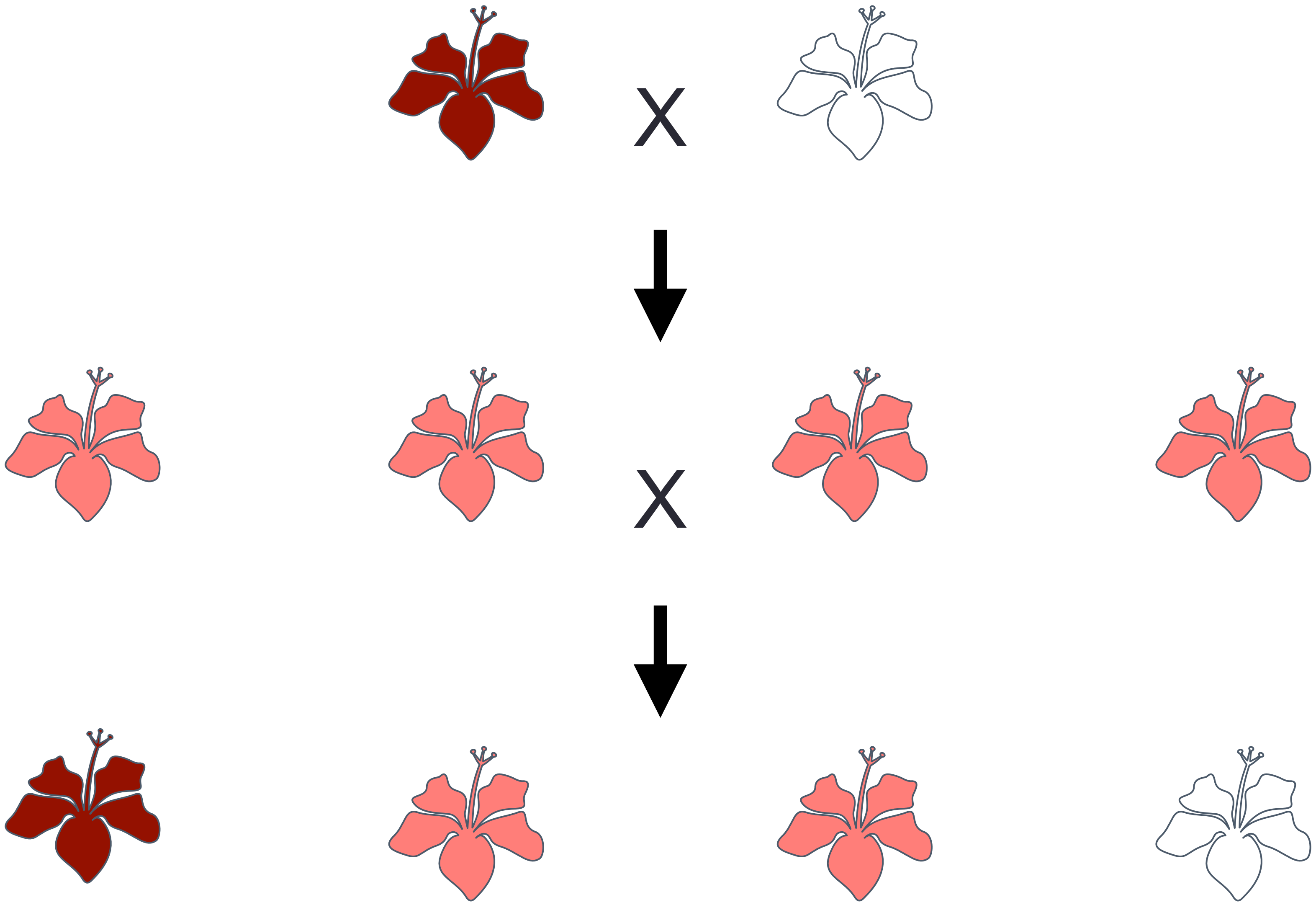
Genetics is a branch of biology concerned with the study of genes, genetic variation, and heredity in organisms



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

1. Basics about genetics and sequencing

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



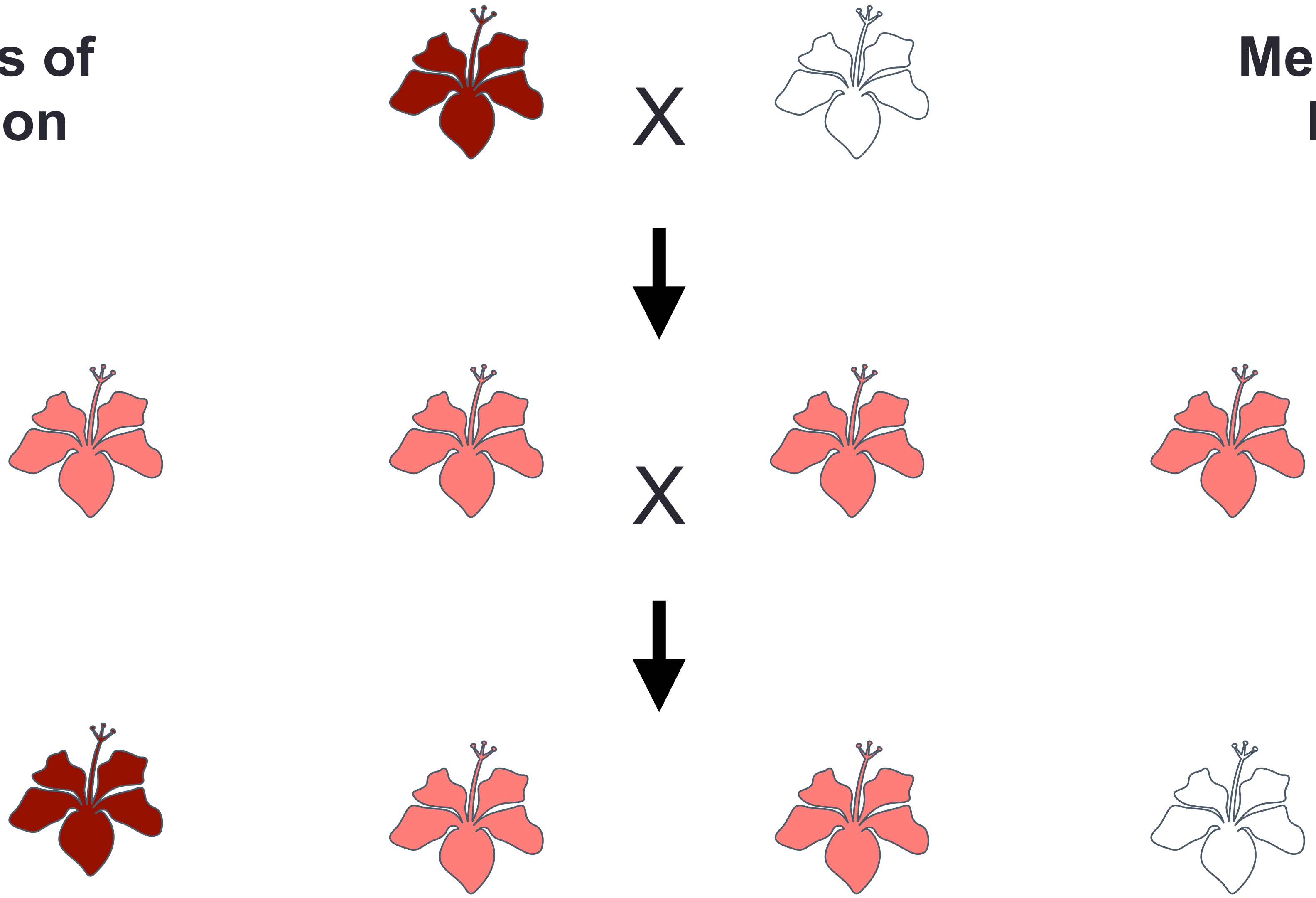
https://en.wikipedia.org/wiki/Genetics

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

Mechanisms of Inherence

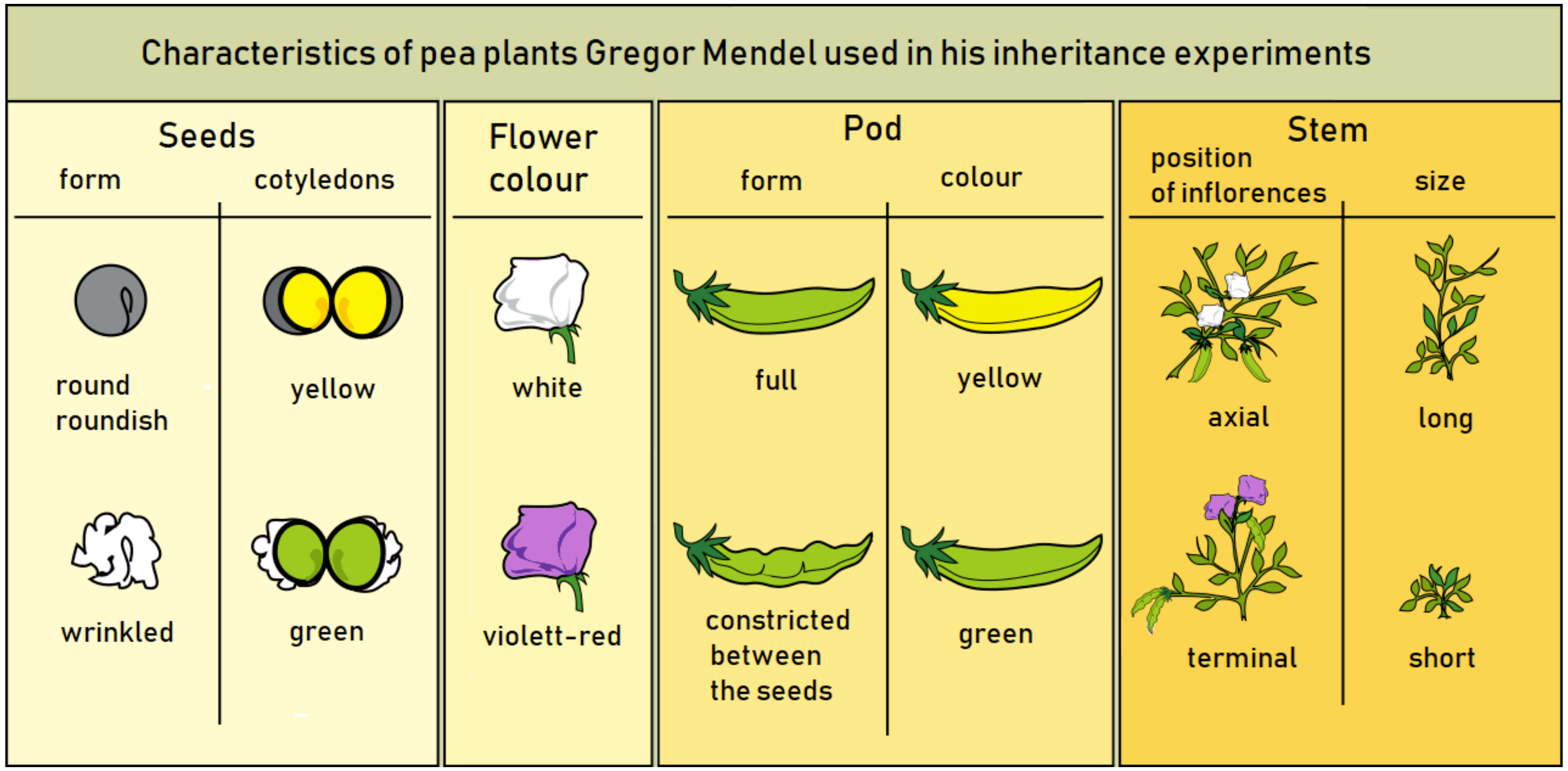


https://en.wikipedia.org/wiki/Genetics

Mendel’s Law of Inherence

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

Material: Pea plants with different traits



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.

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

1. Basics about genetics and sequencing

Mendel’s Law of Inherence

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

X

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

Mendel’s Law of Inherence

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

1. Law of dominance and uniformity

2. Law of segregation

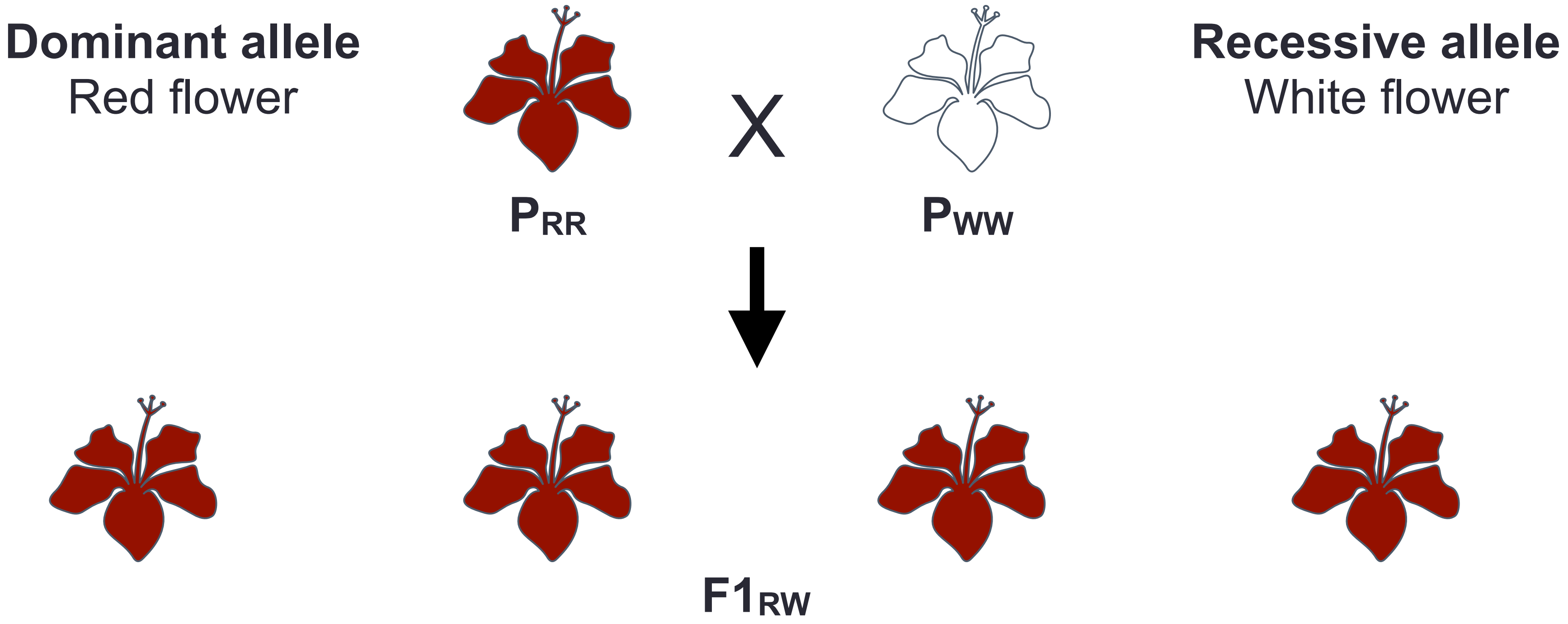
3. Law of independent assortment

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

Mendel’s Law of Inherence

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

1. Law of dominance and uniformity



Dominant allele
Red flower

P_{RR}

X

P_{WW}

Recessive allele
White flower

$F1_{RW}$

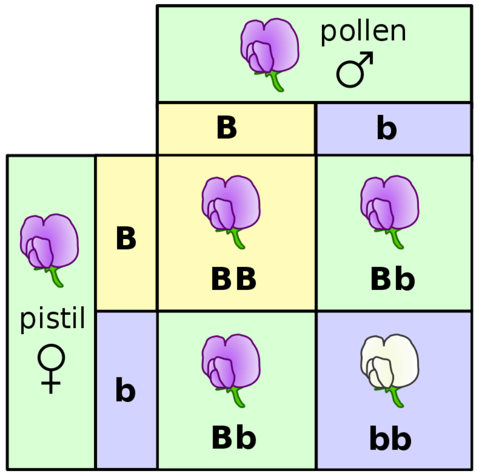
The dominant allele will present uniformity for the F1 generation

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

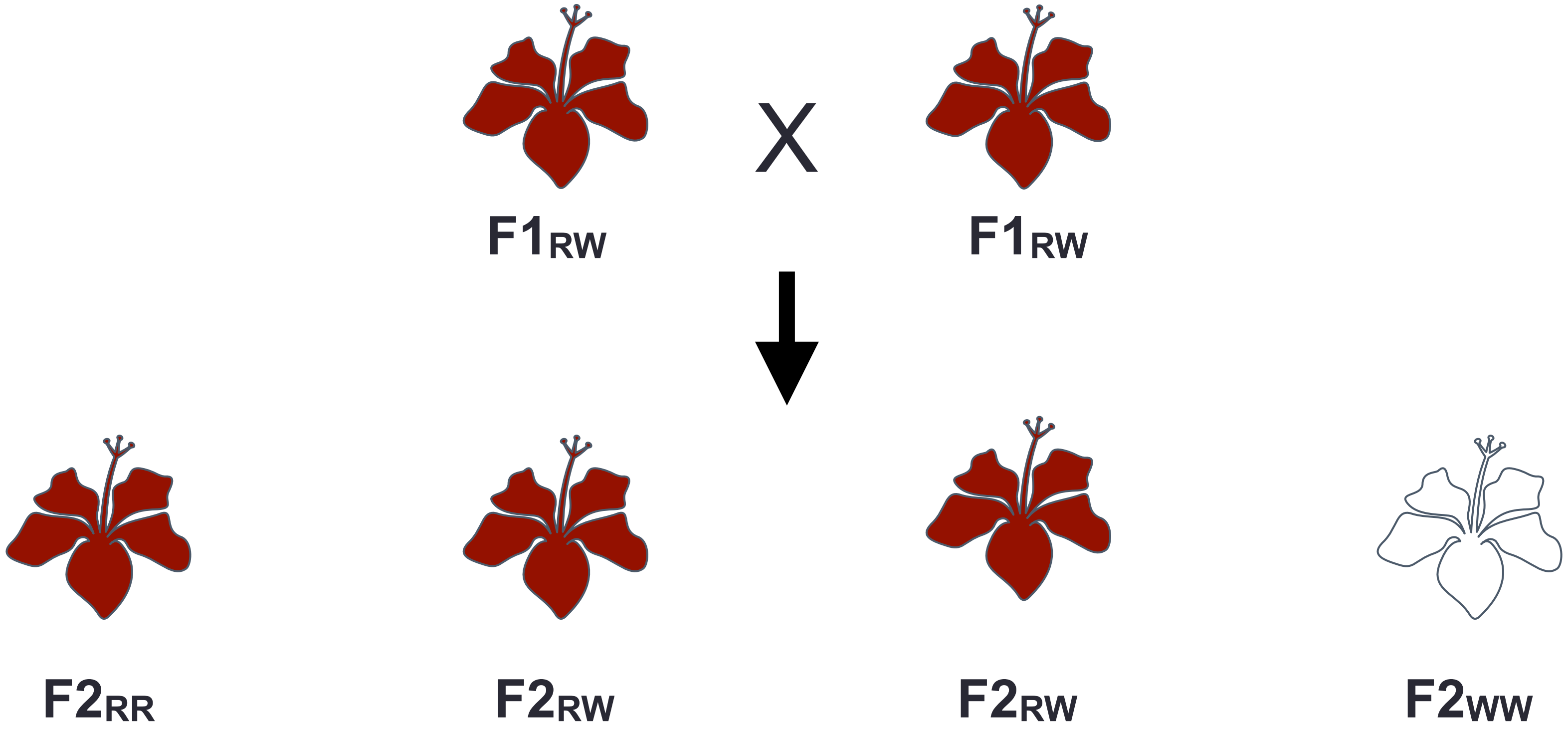
Mendel’s Law of Inherence

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

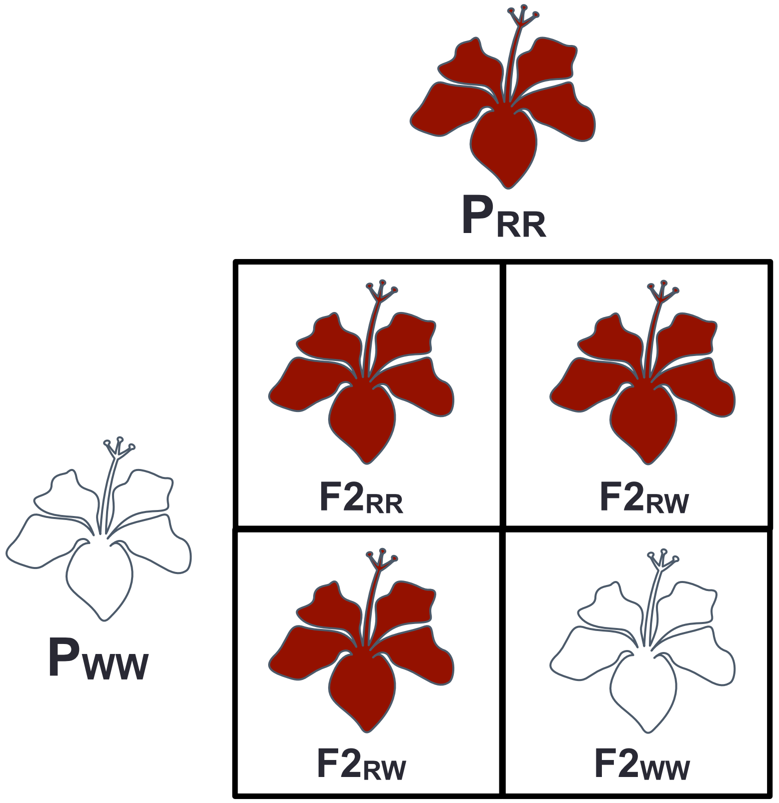
https://en.wikipedia.org/wiki/Gregor_Mendel

Mendel’s Law of Inherence

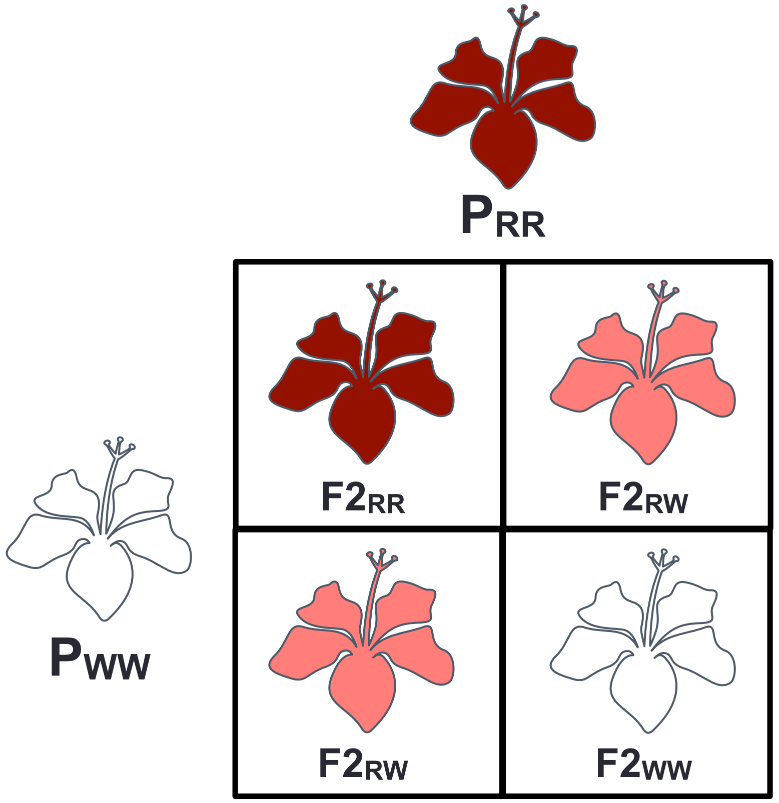
Versuche über Pflanzen-Hybriden
(Experiments on Plant Hybridization)
1866

2. Law of segregation

Alleles segregate in the progeny



Dominance
3:1



Co-dominance
1:2:1

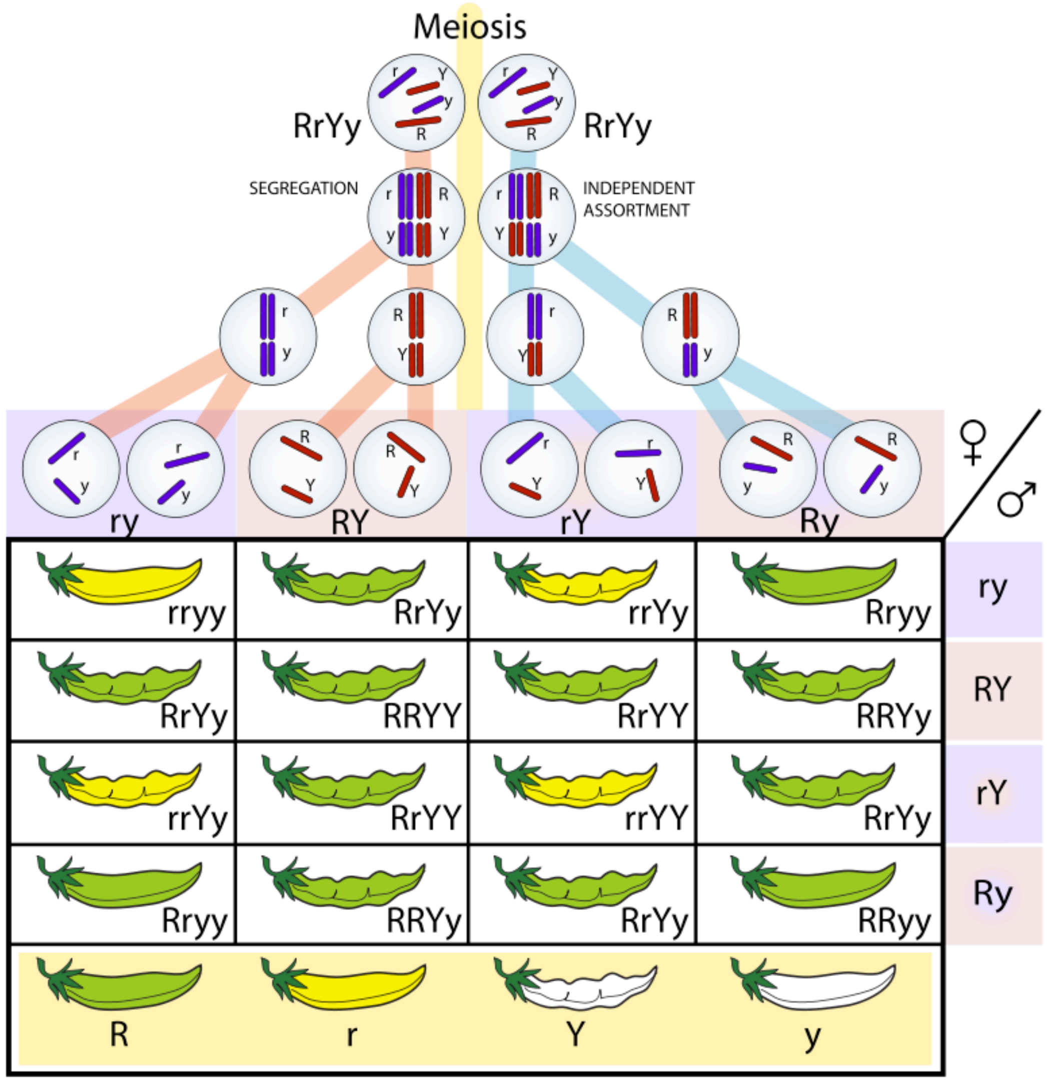
https://en.wikipedia.org/wiki/Gregor_Mendel

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



	ry	RY	rY	Ry
ry	rryy	RrYy	rrYy	Rryy
RY	RrYy	RRYY	RrYY	RRYy
rY	rrYy	RrYY	rrYY	RrYy
Ry	Rryy	RRYy	RrYy	RRyy
	R	r	Y	y

Independent Dominance
9:3:3:1

↓

3:1 Green/Yellow (R/r)

3:1 Wrinkle/plain (Y/y)

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

Mendel's Law of Inherence

Bateson

Saunders

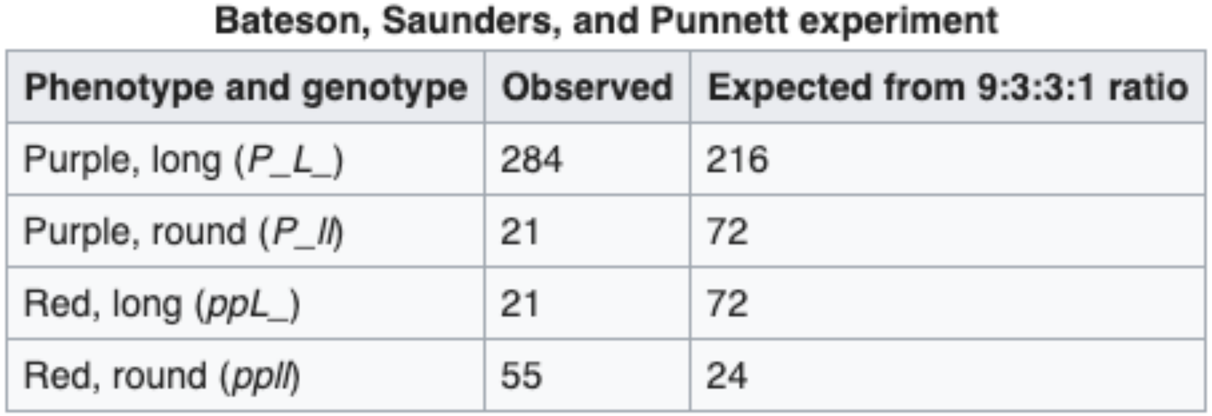
Punnett

3. Law of independent assortment

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

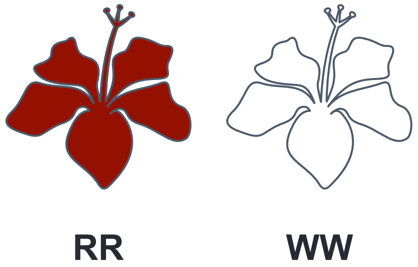
Allele definition

Bateson

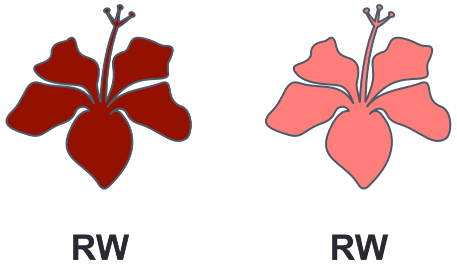
Saunders

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.

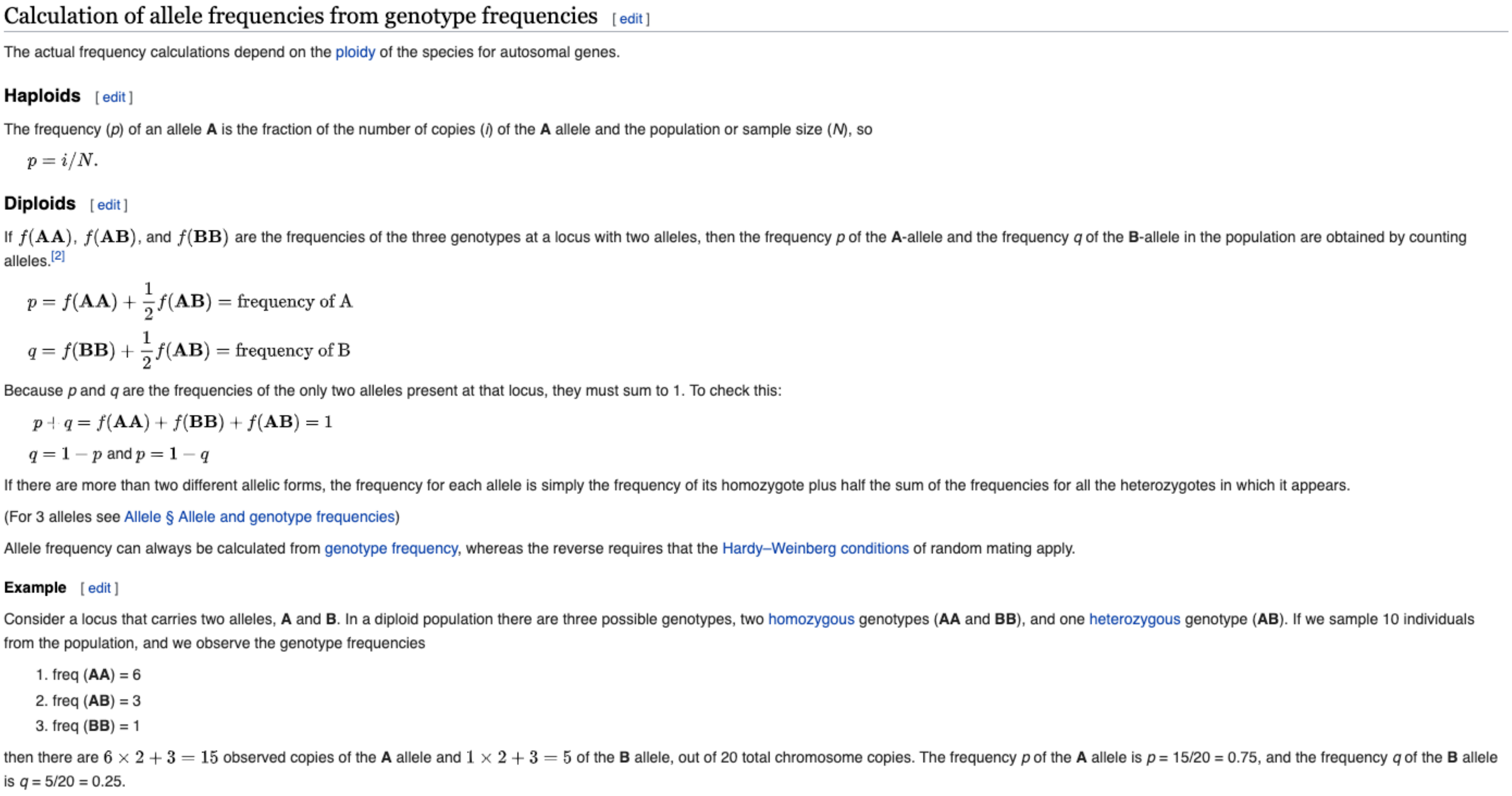


Homozygous state



Heterozygous state

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:

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

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

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.

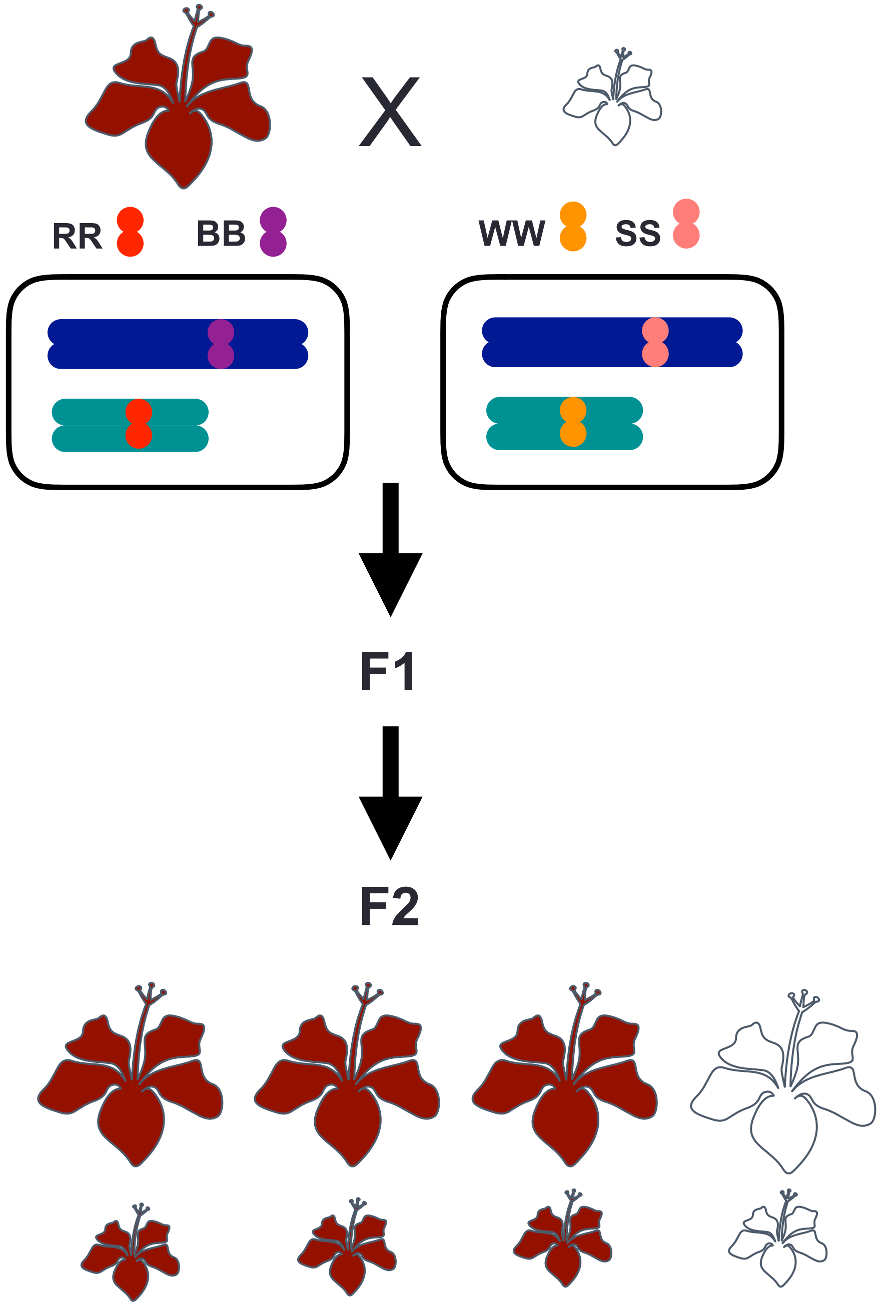
https://en.wikipedia.org/wiki/Allele_frequency

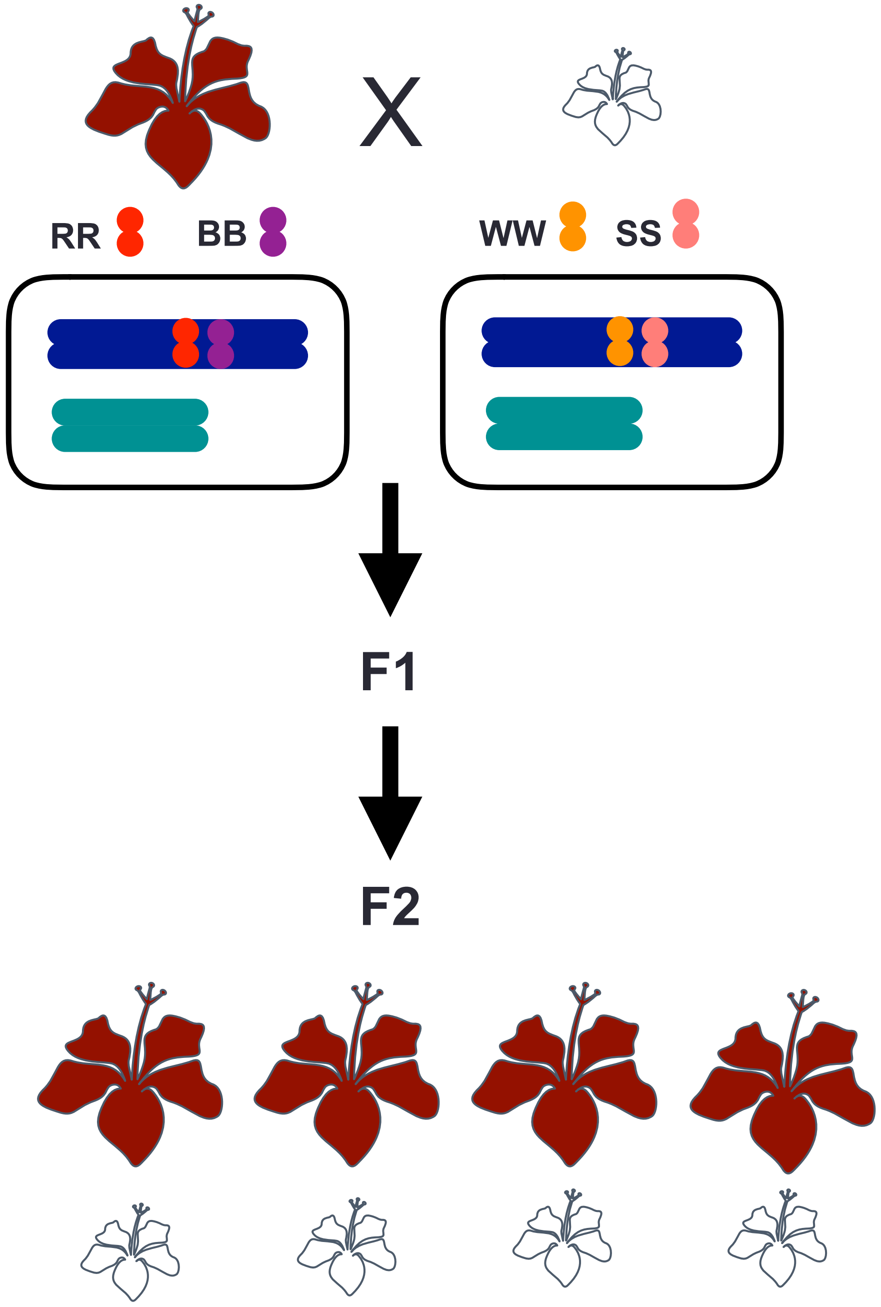
Genetic Linkage and Genetic Maps

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



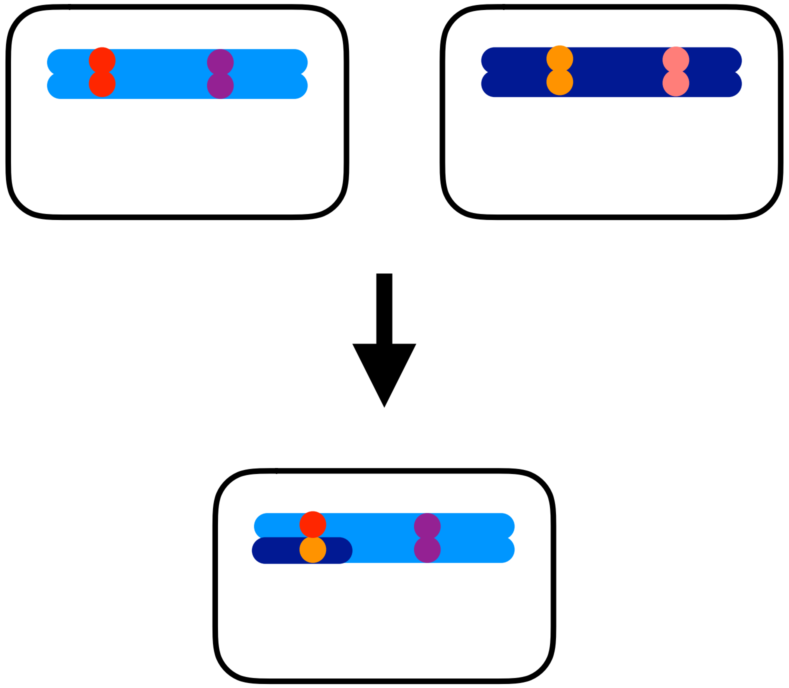


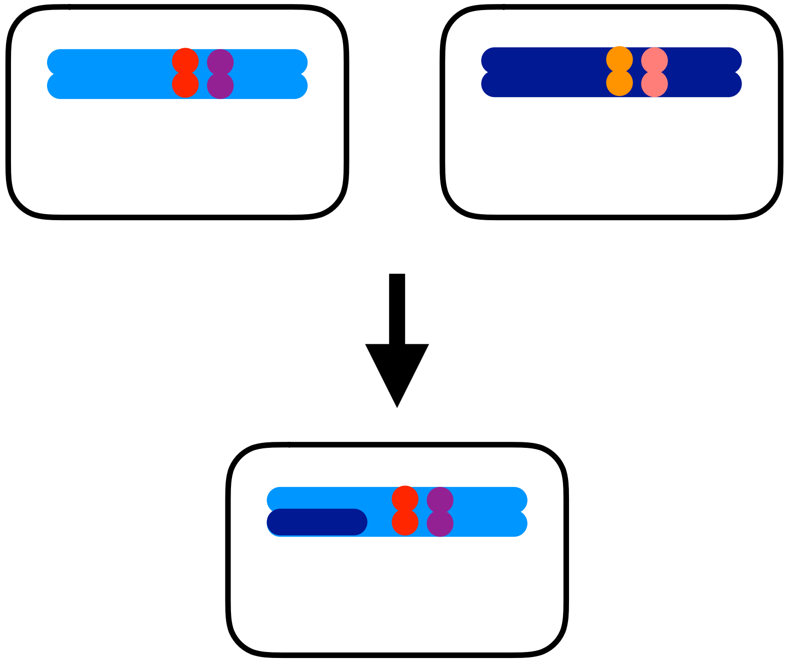
Genetic Linkage and Genetic Maps

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



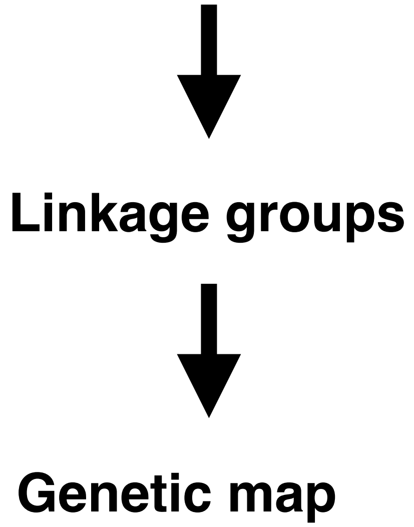


Genetic Linkage and Genetic Maps

Morgan

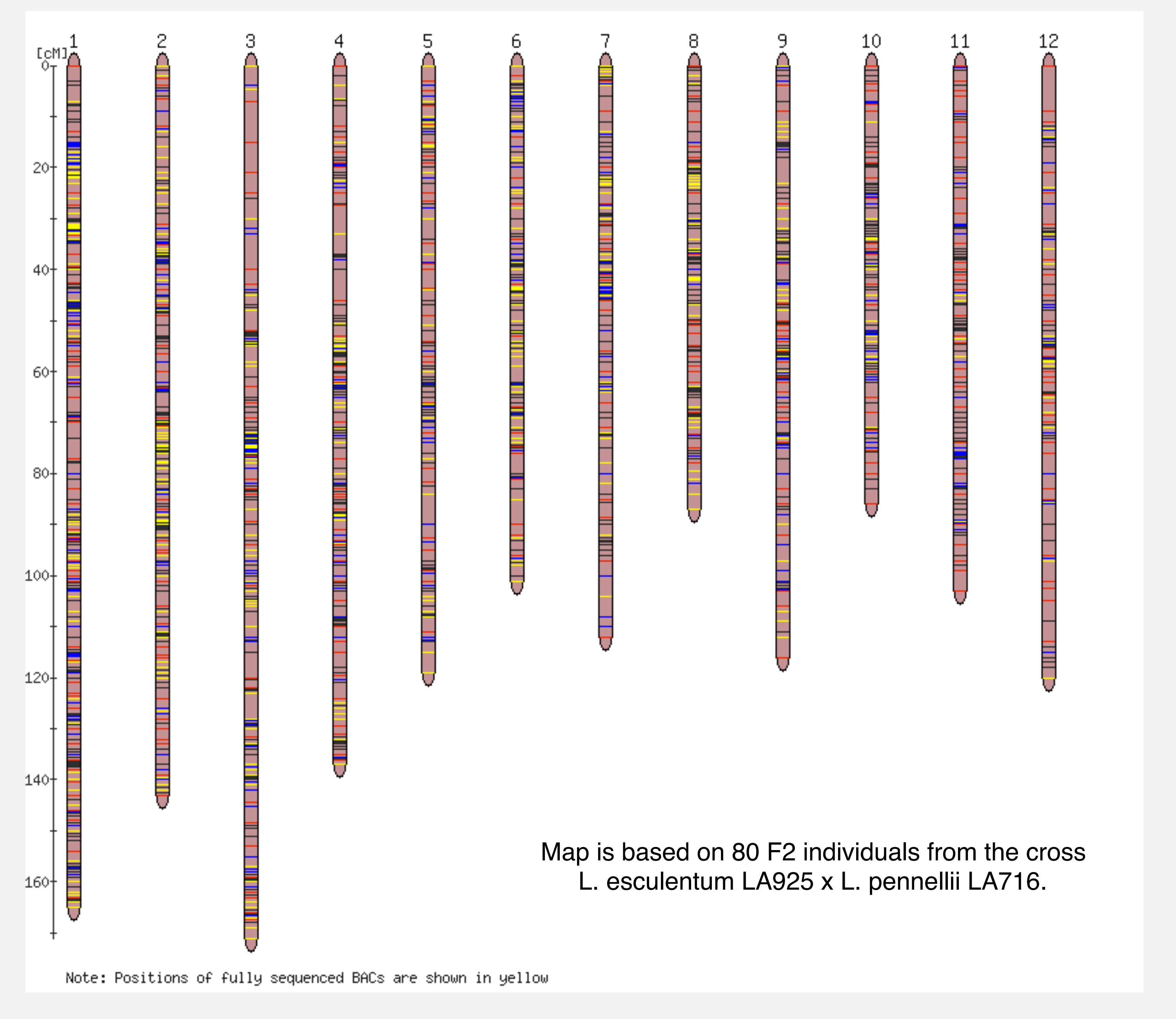
Sturtevant

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



Linkage groups

Genetic map



Map is based on 80 F2 individuals from the cross L. esculentum LA925 x L. pennellii LA716.

Note: Positions of fully sequenced BACs are shown in yellow

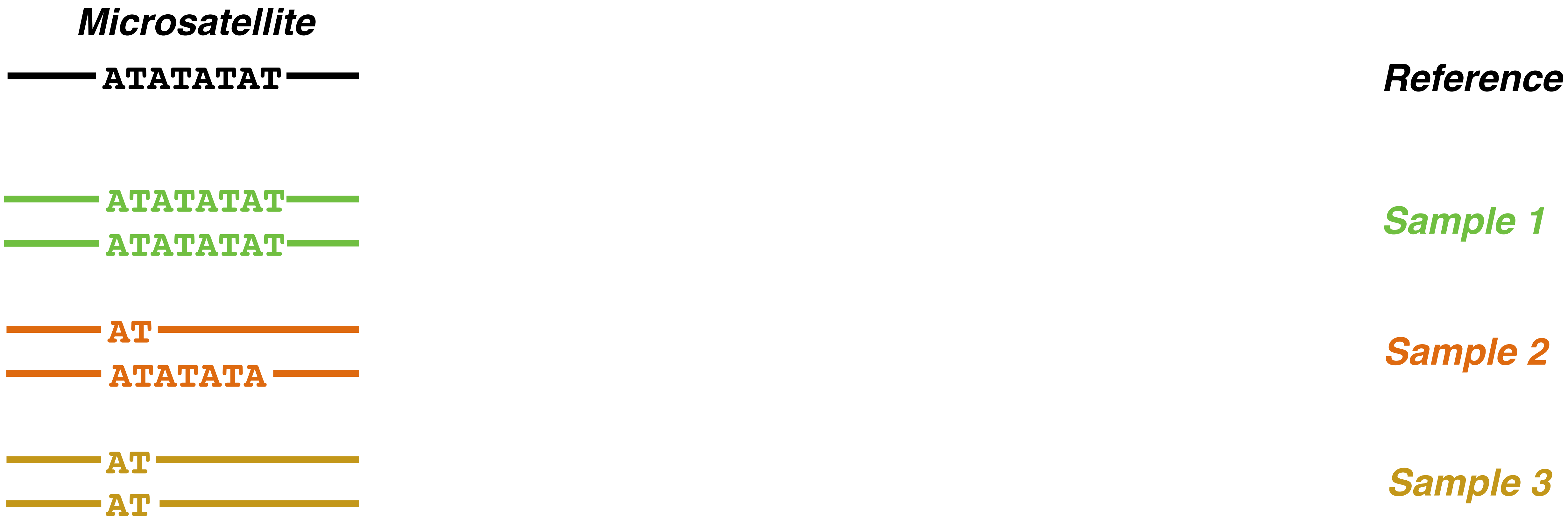
Genetic Marker

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)

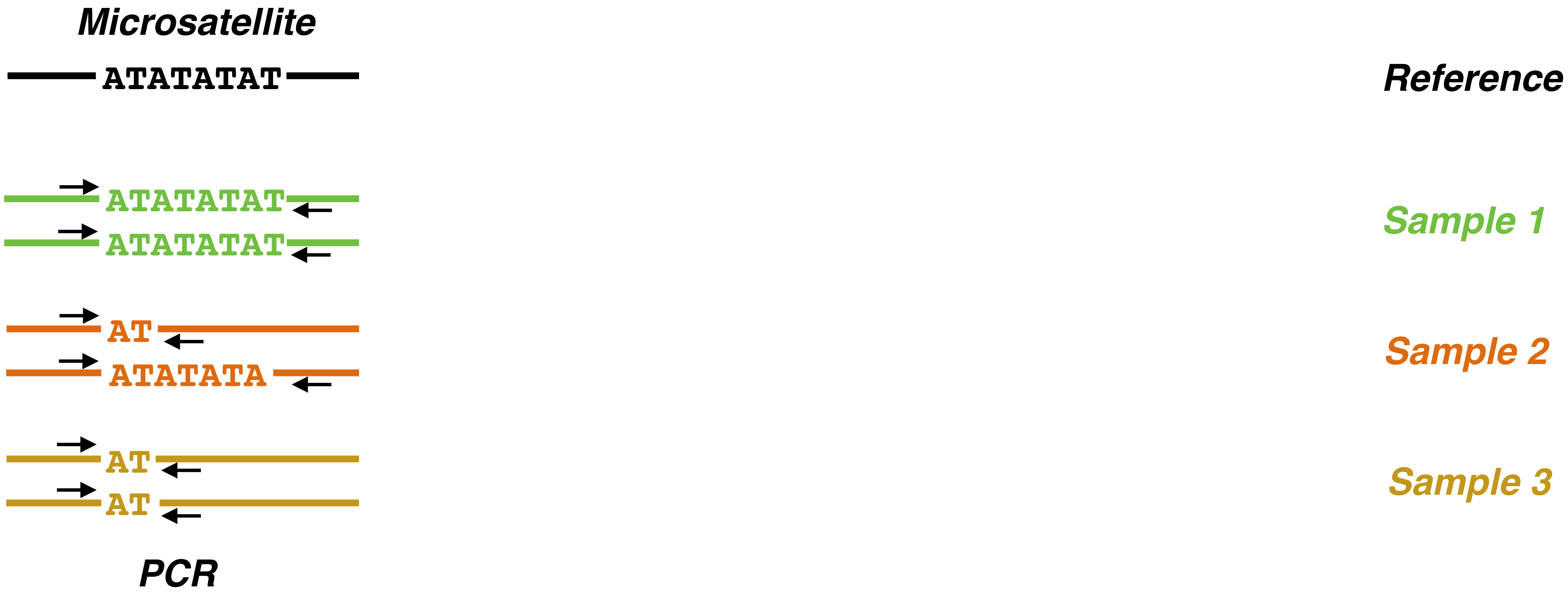
Genetic Marker

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.



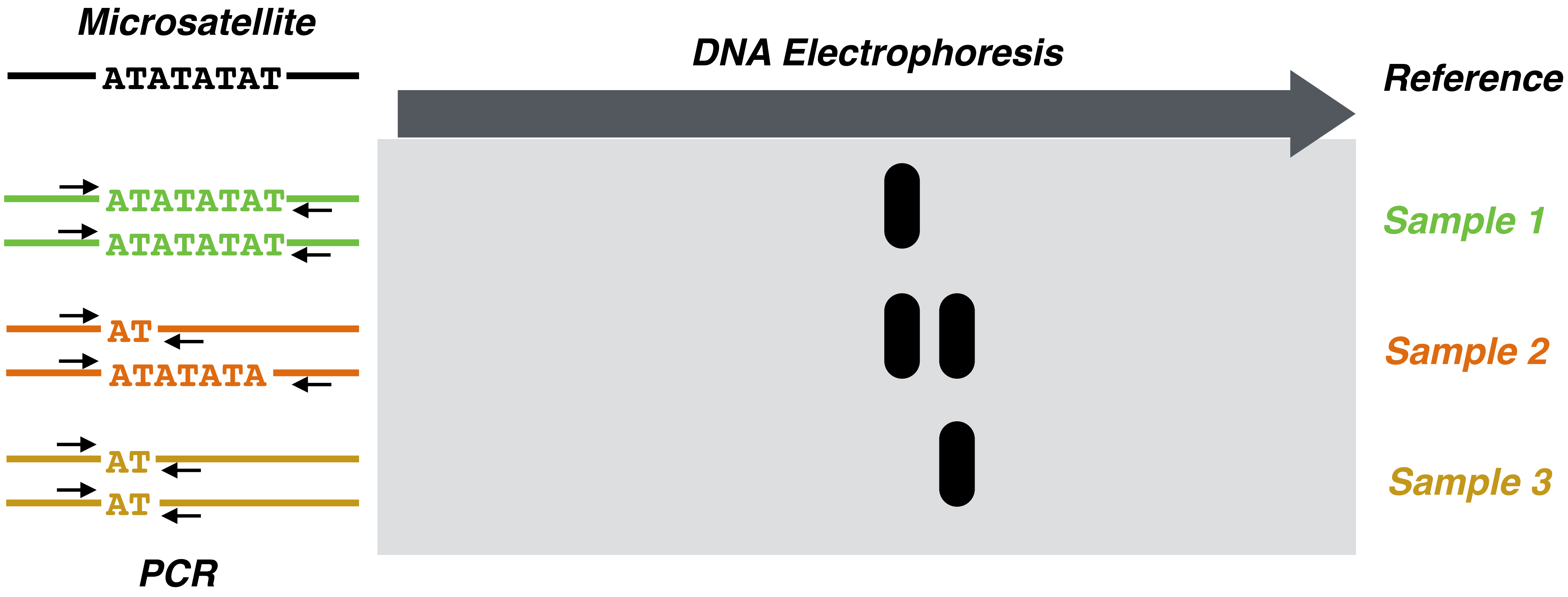
Genetic Marker

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.



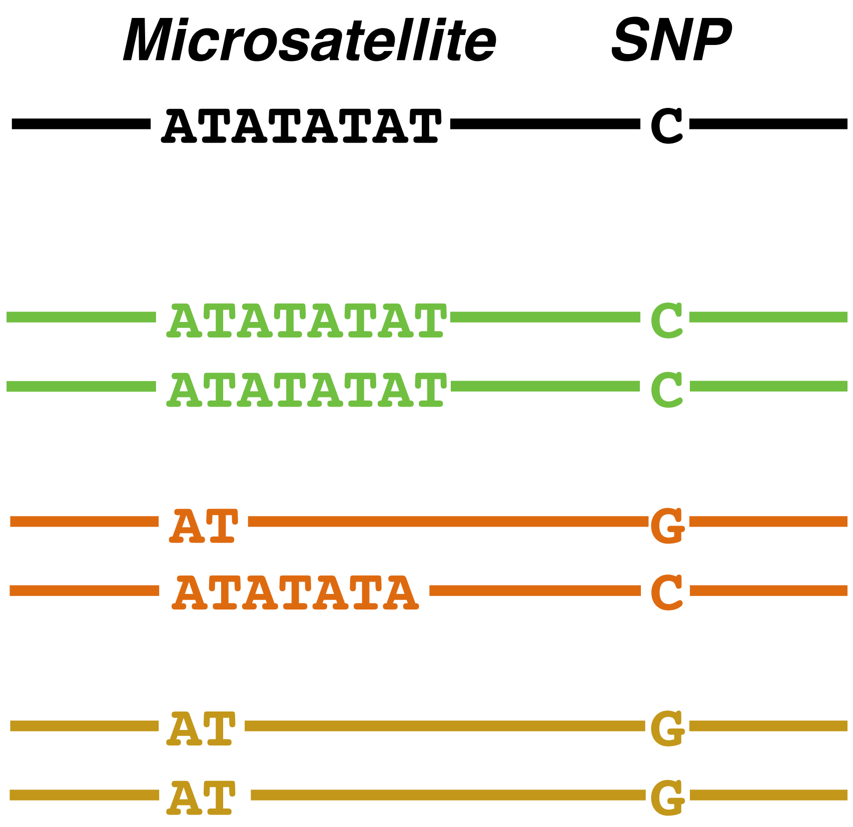
Genetic Marker

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.



Genetic Marker

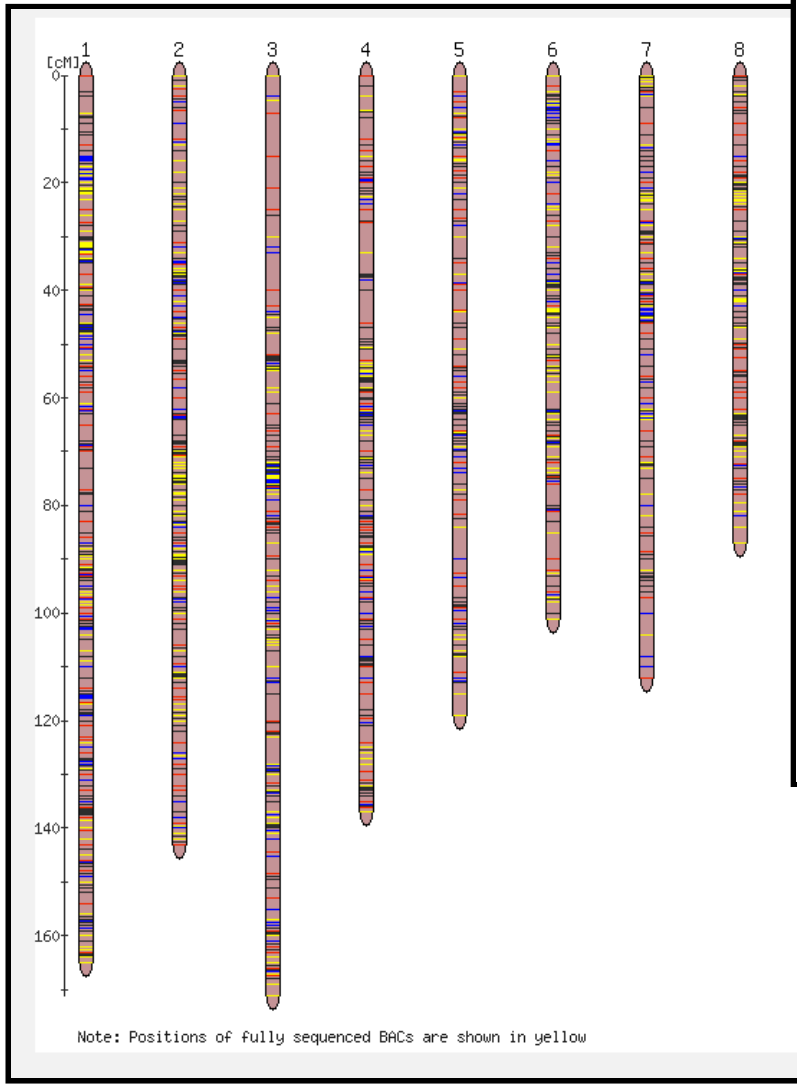
A **genetic marker** is a gene or DNA sequence with a known location on a chromosome that can be used to identify individuals or species. It can be described as a variation (which may arise due to mutation or alteration in the genomic loci) that can be observed.



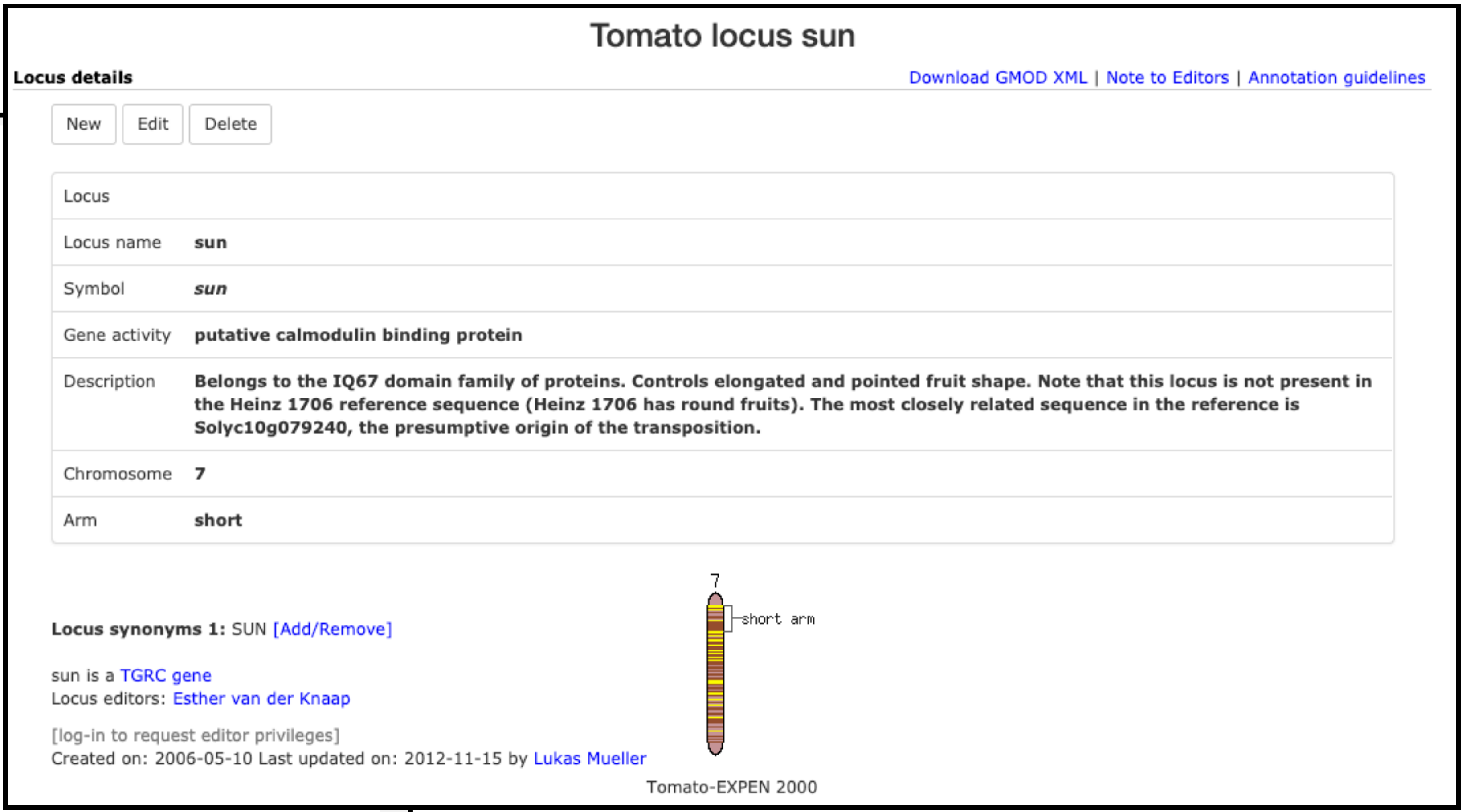
						Reference
chr1	1001	C	G	…	0/0	*Sample 1*
chr1	1001	C	G	…	0/1	*Sample 2*
chr1	1001	C	G	…	1/1	*Sample 3*

Locus

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



Note: Positions of fully sequenced BACs are shown in yellow



Tomato locus sun

Locus details

Download GMOD XML | Note to Editors | Annotation guidelines

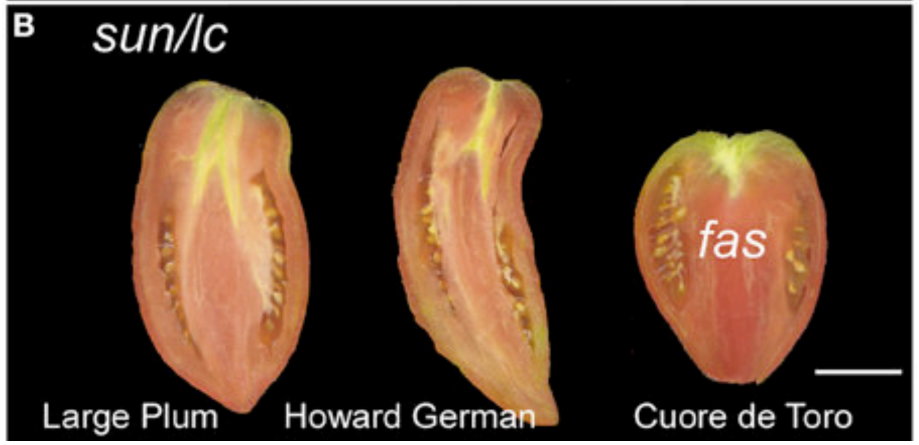
New Edit Delete

Locus	
Locus name	**sun**
Symbol	***sun***
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

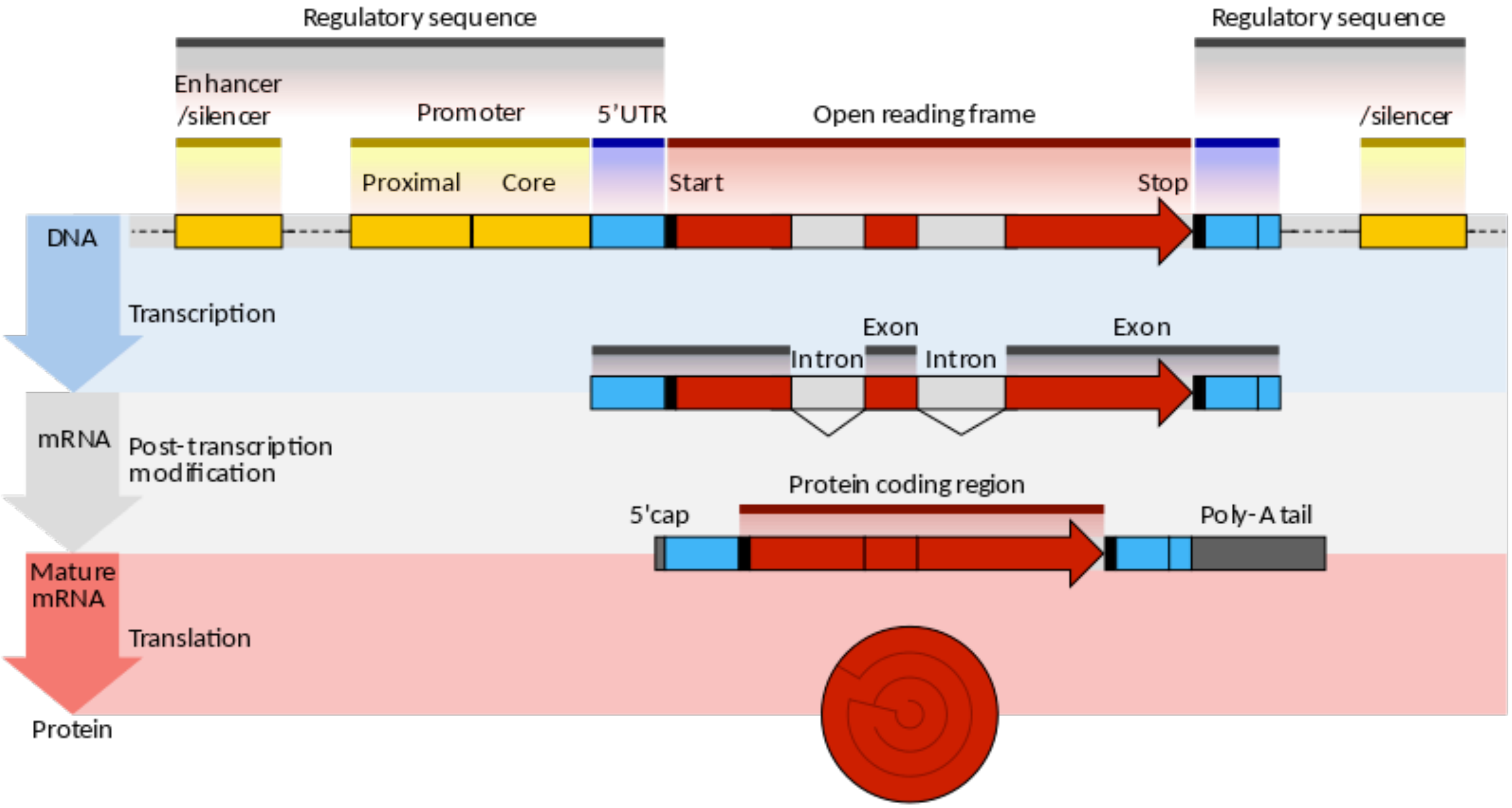
[log-in to request editor privileges]
Created on: 2006-05-10 Last updated on: 2012-11-15 by Lukas Mueller



van der Knaap, Esther, et al. "What lies beyond the eye: the molecular mechanisms regulating tomato fruit weight and shape." *Frontiers in Plant Science* 5 (2014): 227.

Gene

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

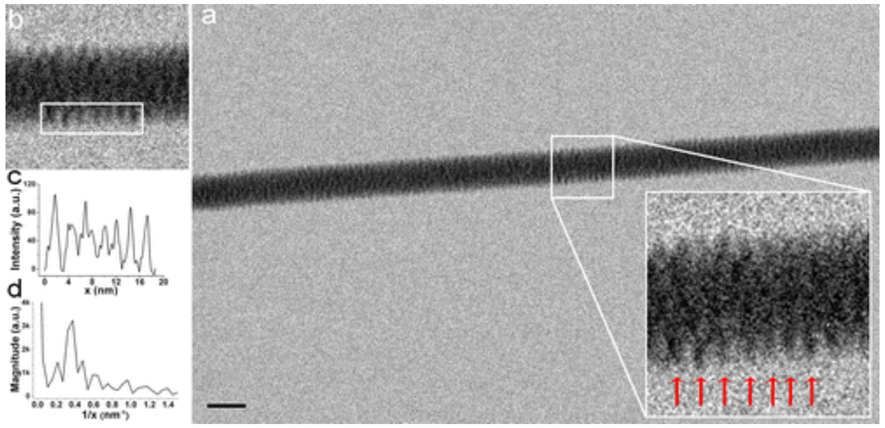


https://en.wikipedia.org/wiki/Gene

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



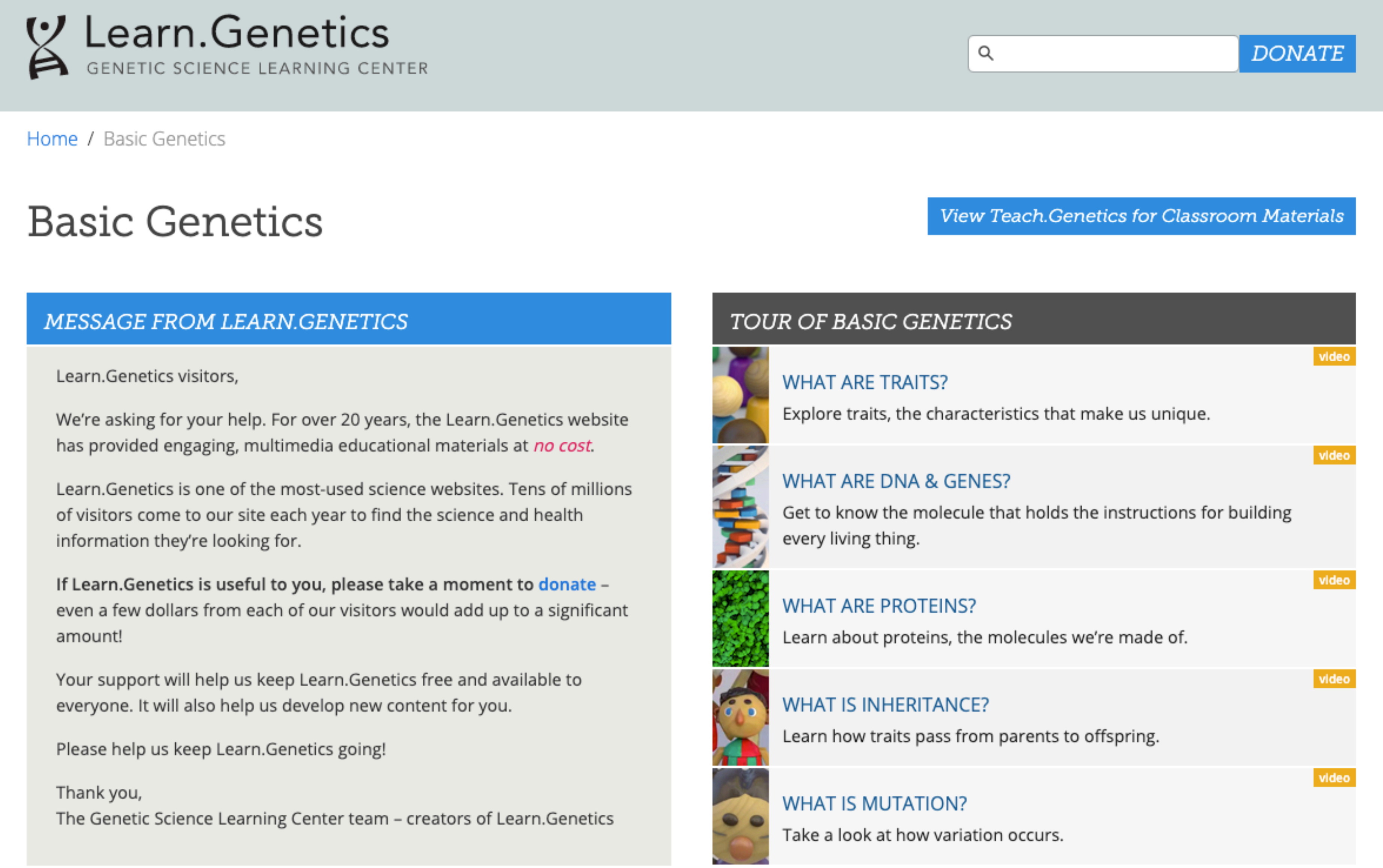
(Gentile et al. Nano Lett., 2012, 12 (12), pp 6453–6458)

ATGCGCGTCGCGGTGAAT

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? video
Explore traits, the characteristics that make us unique.

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

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

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

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

Outline of Topics

1. Basics about genetics and sequencing

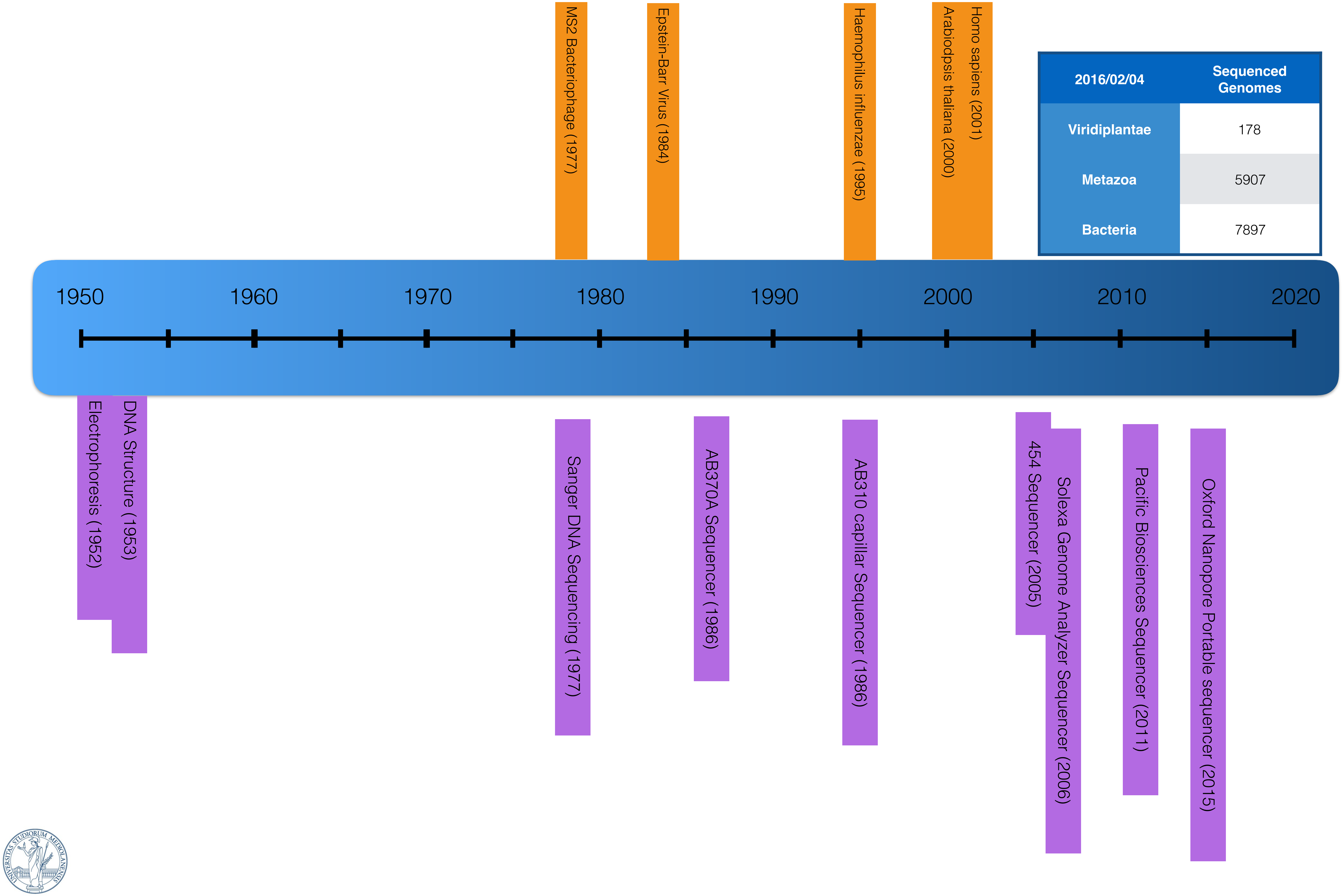
2. First steps: Pre-NGS era

3. Short read sequencing technologies

4. Long read sequencing technologies

5. Common file formats

2. First steps: Pre-NGS era



MS2 Bacteriophage (1977)

Epstein-Barr Virus (1984)

Haemophilus influenzae (1995)

Arabiodpsis thaliana (2000)

Homo sapiens (2001)

1950 1960 1970 1980 1990 2000 2010 2020

Electrophoresis (1952)

DNA Structure (1953)

Sanger DNA Sequencing (1977)

AB370A Sequencer (1986)

AB310 capillar Sequencer (1986)

454 Sequencer (2005)

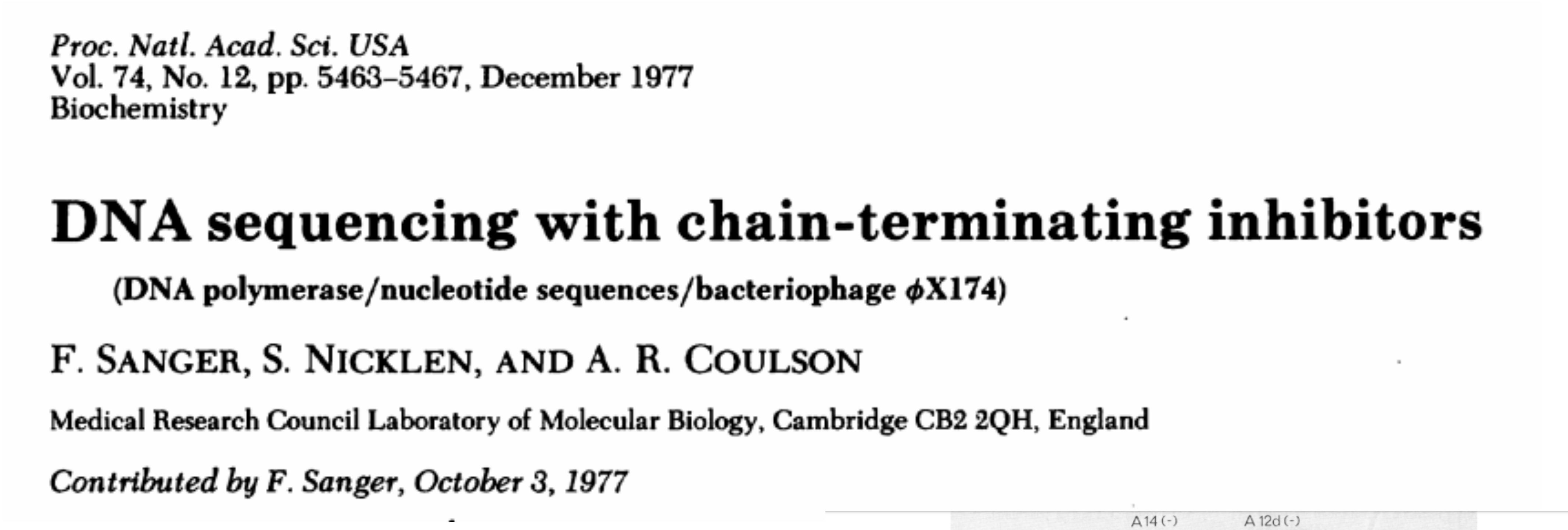
Solexa Genome Analyzer Sequencer (2006)

Pacific Biosciences Sequencer (2011)

Oxford Nanopore Portable sequencer (2015)

2016/02/04	Sequenced Genomes
Viridiplantae	178
Metazoa	5907
Bacteria	7897

2. First steps: Pre-NGS era



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

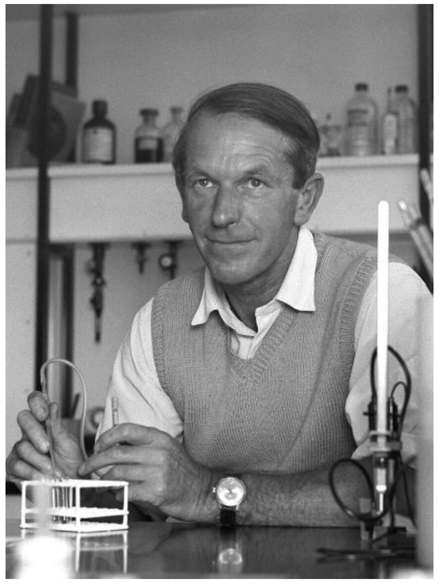
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

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

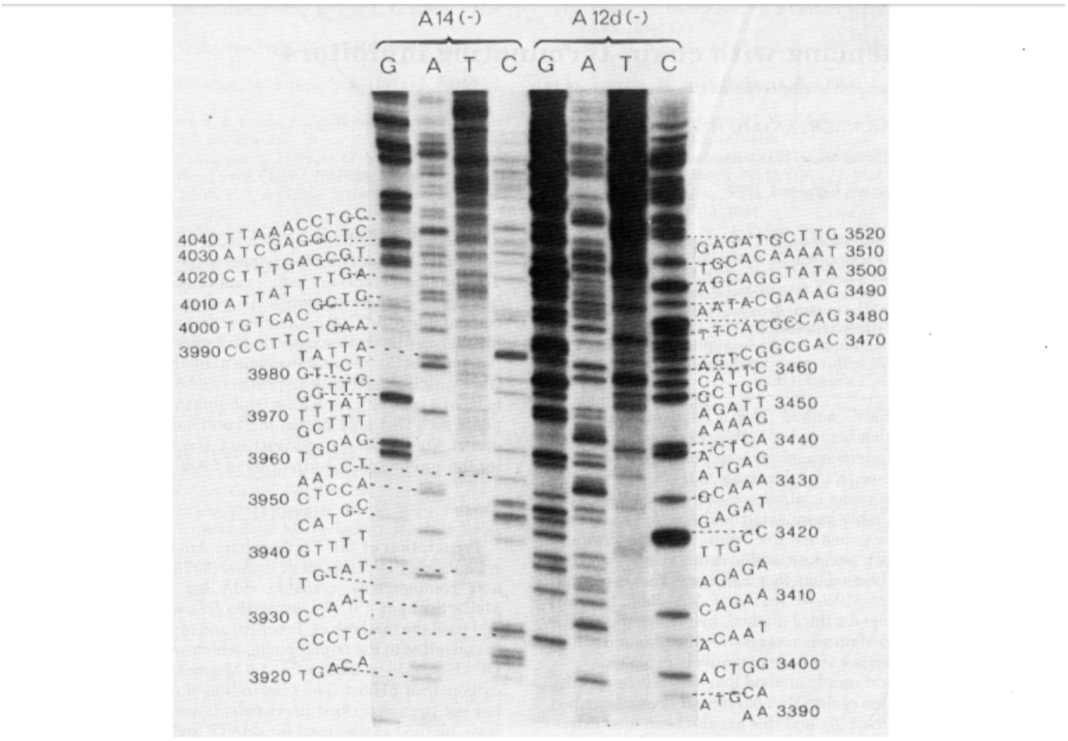
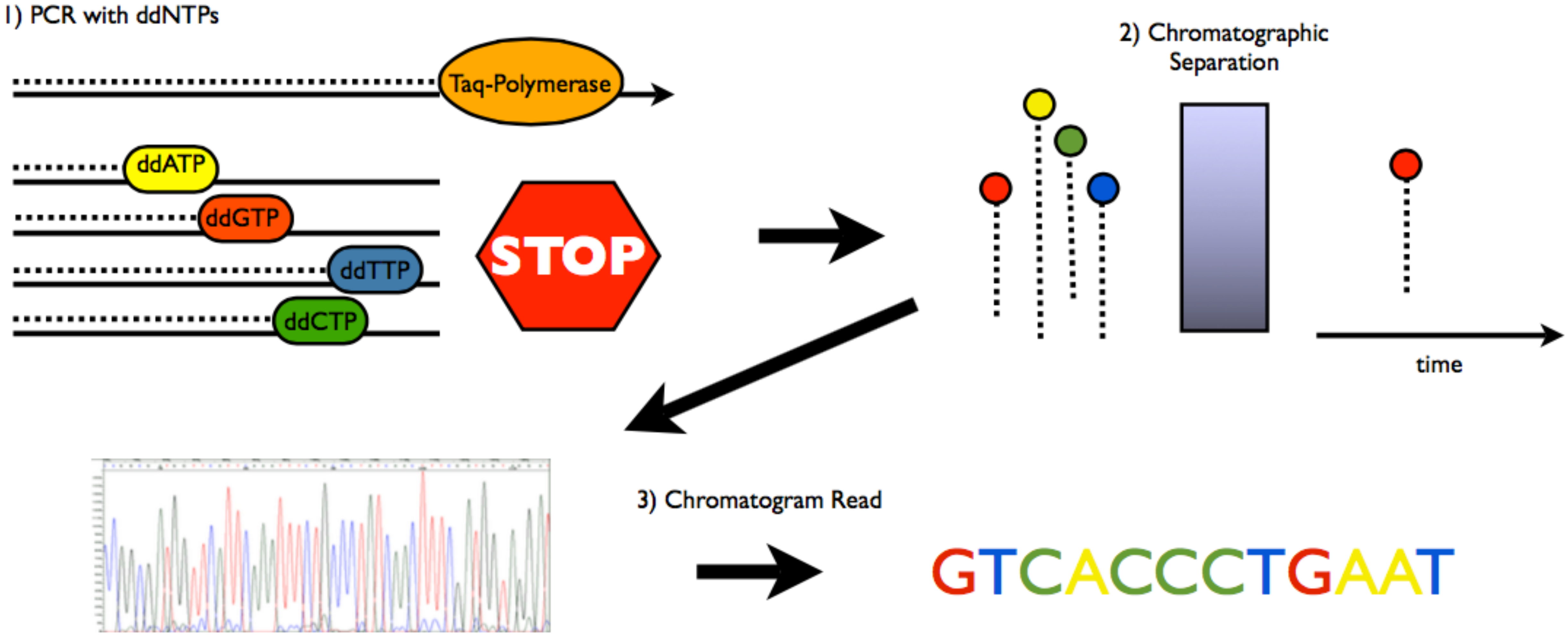


FIG. 1. Autoradiograph of the acrylamide gel from the sequence determination using restriction fragments A12d and A14 as primers on the complementary strand of ϕX174 DNA. The inhibitors used were (left to right) ddGTP, ddATP, ddTTP, and araCTP. Electrophoresis was on a 12% acrylamide gel at 40 mA for 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.

Sanger DNA sequencing



1) PCR with ddNTPs

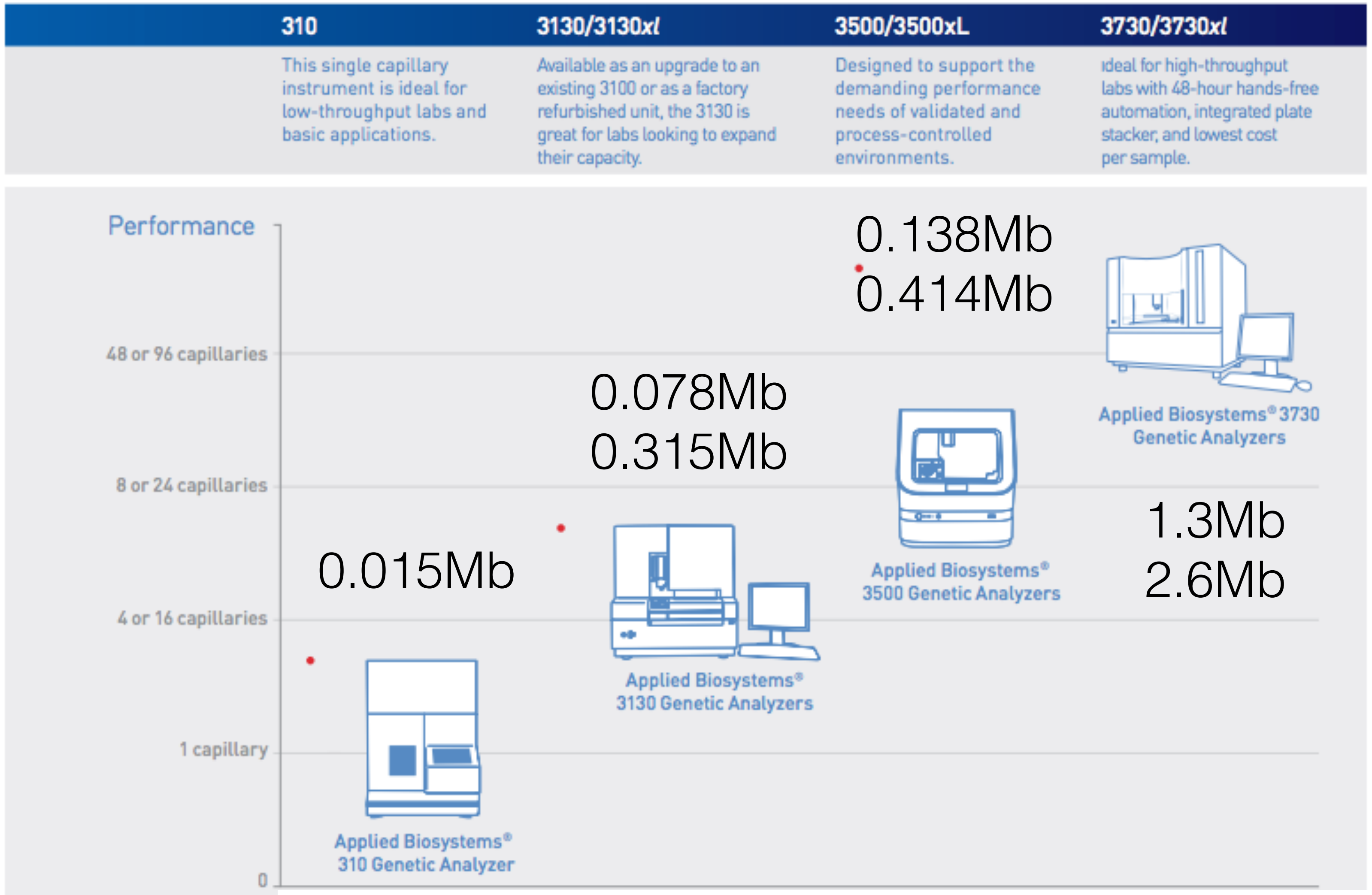
2) Chromatographic Separation

3) Chromatogram Read

GTCACCCTGAAT

2. First steps: Pre-NGS era

Sanger DNA sequencing systems



	310	3130/3130*xl*	3500/3500xL	3730/3730*xl*
	This single capillary instrument is ideal for low-throughput labs and basic applications.	Available as an upgrade to an existing 3100 or as a factory refurbished unit, the 3130 is great for labs looking to expand their capacity.	Designed to support the demanding performance needs of validated and process-controlled environments.	Ideal for high-throughput labs with 48-hour hands-free automation, integrated plate stacker, and lowest cost per sample.

Performance
48 or 96 capillaries
8 or 24 capillaries
4 or 16 capillaries
1 capillary
0
Applied Biosystems® 310 Genetic Analyzer
Applied Biosystems® 3130 Genetic Analyzers
Applied Biosystems® 3500 Genetic Analyzers
Applied Biosystems® 3730 Genetic Analyzers

0.015Mb

0.078Mb
0.315Mb

0.138Mb
0.414Mb

1.3Mb
2.6Mb

Error Rate
0.1%

Error Type
substitution

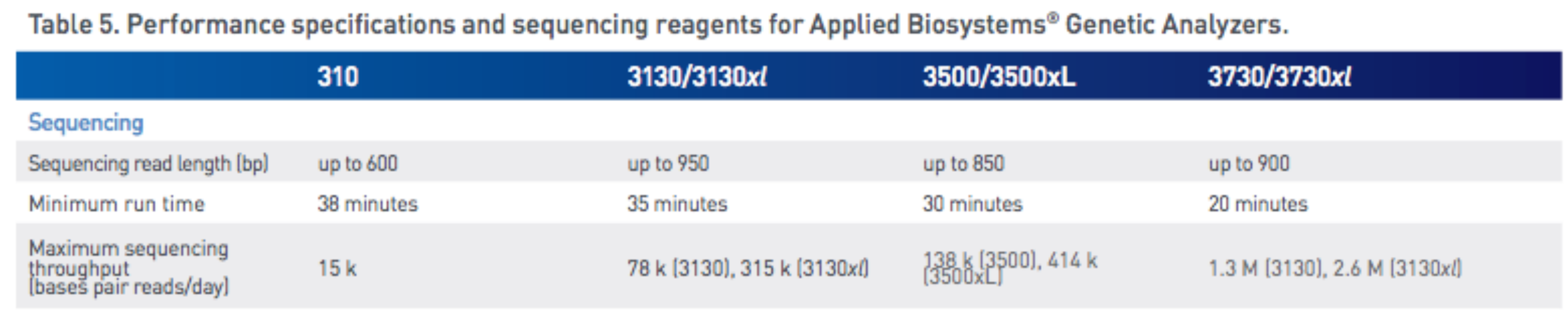


Table 5. Performance specifications and sequencing reagents for Applied Biosystems® Genetic Analyzers.

	310	3130/3130*xl*	3500/3500xL	3730/3730*xl*
Sequencing				
Sequencing read length (bp)	up to 600	up to 950	up to 850	up to 900
Minimum run time	38 minutes	35 minutes	30 minutes	20 minutes
Maximum sequencing throughput (bases pair reads/day)	15 k	78 k (3130), 315 k (3130*xl*)	138 k (3500), 414 k (3500xL)	1.3 M (3130), 2.6 M (3130*xl*)

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

(Mardis E.R. (2013) Annual Review of Analytical Chemistry 6: 287-303)

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
Ilumina 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

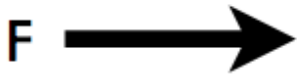
4. Long read sequencing technologies

5. Common file formats

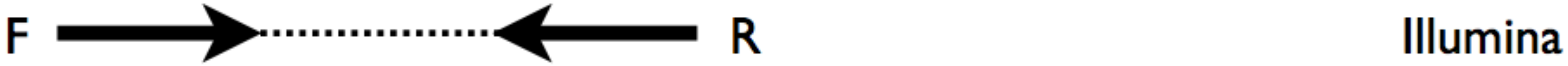
Libraries types

★ Library types (orientations):

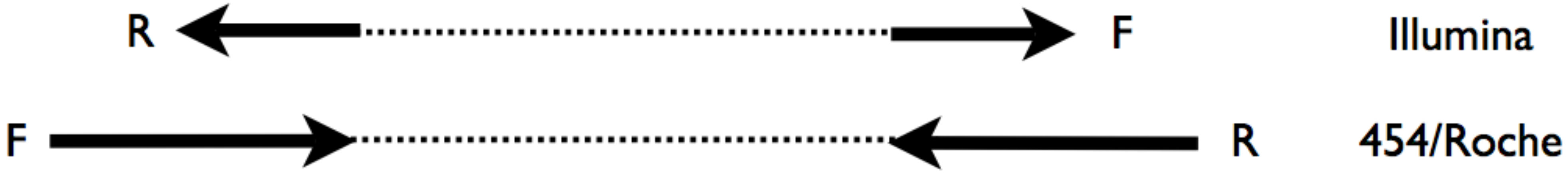
- Single reads



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



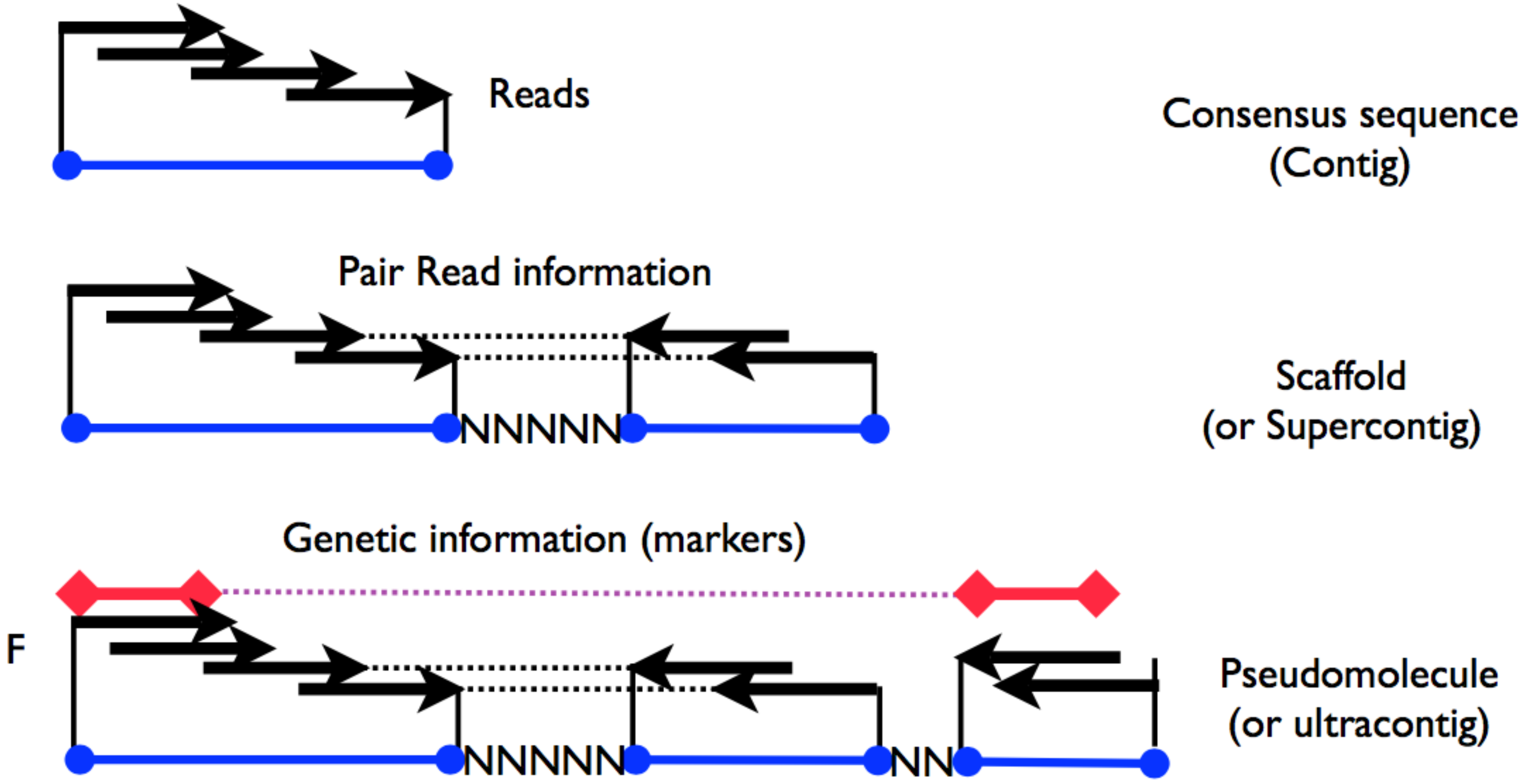
- Mate pairs (MP) (2-40 Kb insert size)



Libraries types

- Why is important the pair information ?

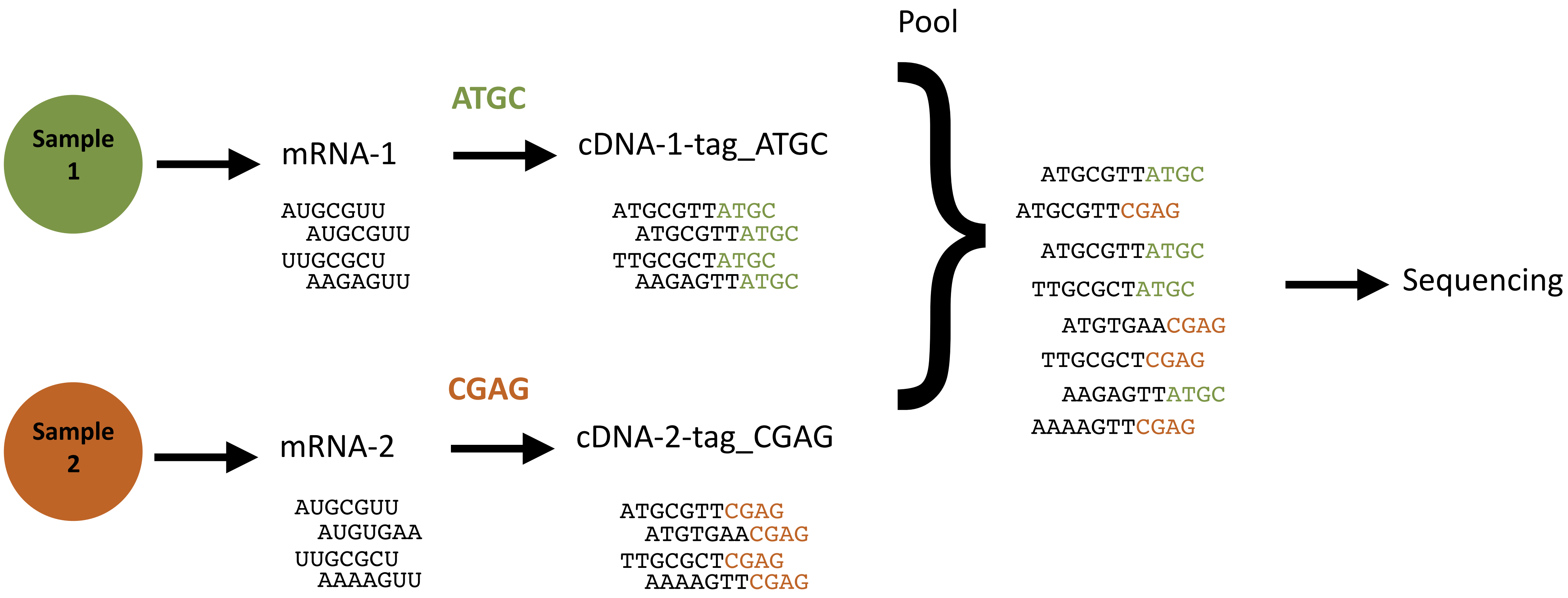
  - *novo* assembly:



Libraries types

Multiplexing

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



Sample 1 → mRNA-1 —ATGC→ cDNA-1-tag_ATGC

AUGCGUU
AUGCGUU
UUGCGCU
AAGAGUU

ATGCGTTATGC
ATGCGTTATGC
TTGCGCTATGC
AAGAGTTATGC

Sample 2 → mRNA-2 —CGAG→ cDNA-2-tag_CGAG

AUGCGUU
AUGUGAA
UUGCGCU
AAAAGUU

ATGCGTTCGAG
ATGTGAACGAG
TTGCGCTCGAG
AAAAGTTCGAG

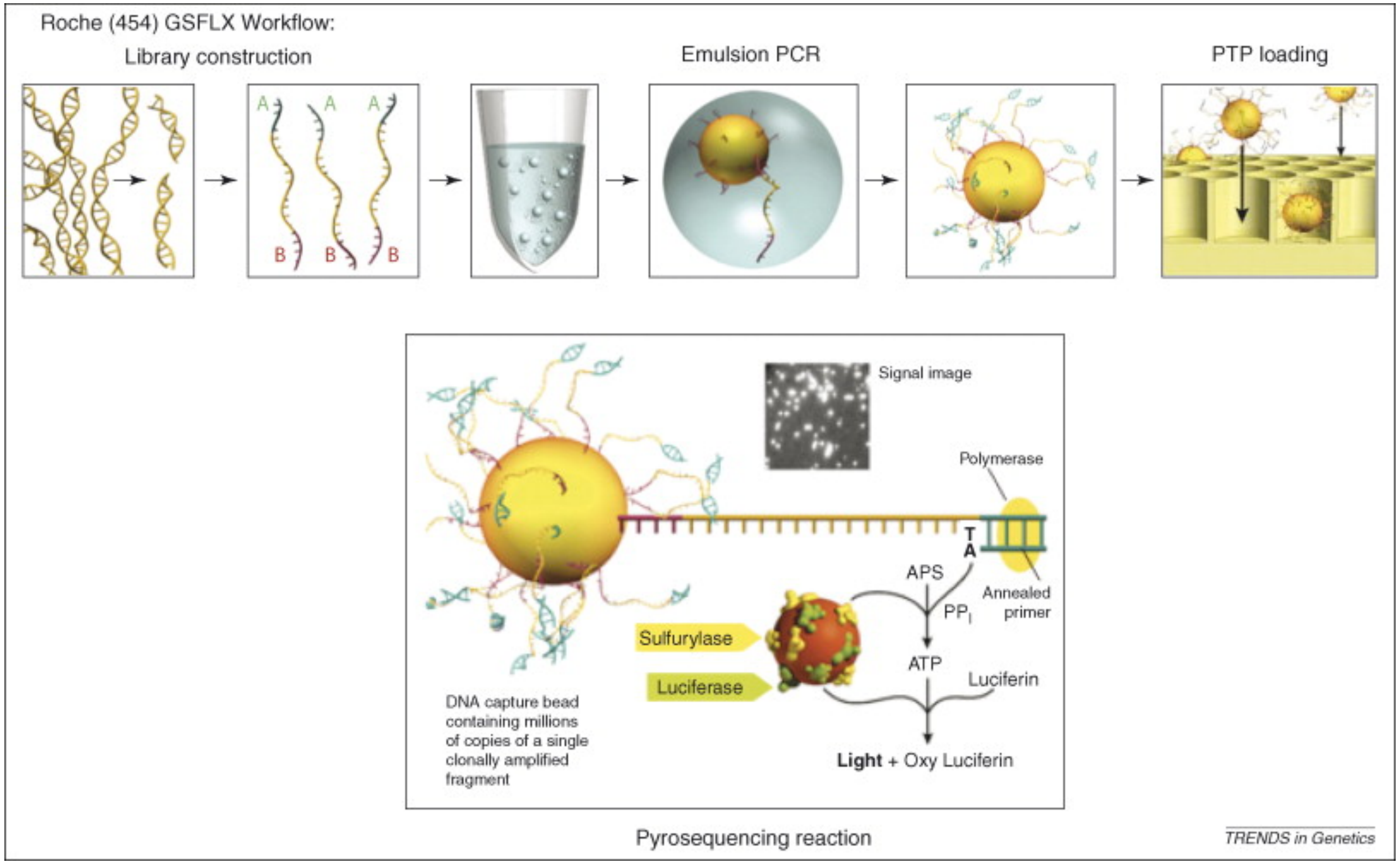
Pool

ATGCGTTATGC
ATGCGTTCGAG
ATGCGTTATGC
TTGCGCTATGC
ATGTGAACGAG
TTGCGCTCGAG
AAGAGTTATGC
AAAAGTTCGAG

→ Sequencing

Sequencing systems: 454

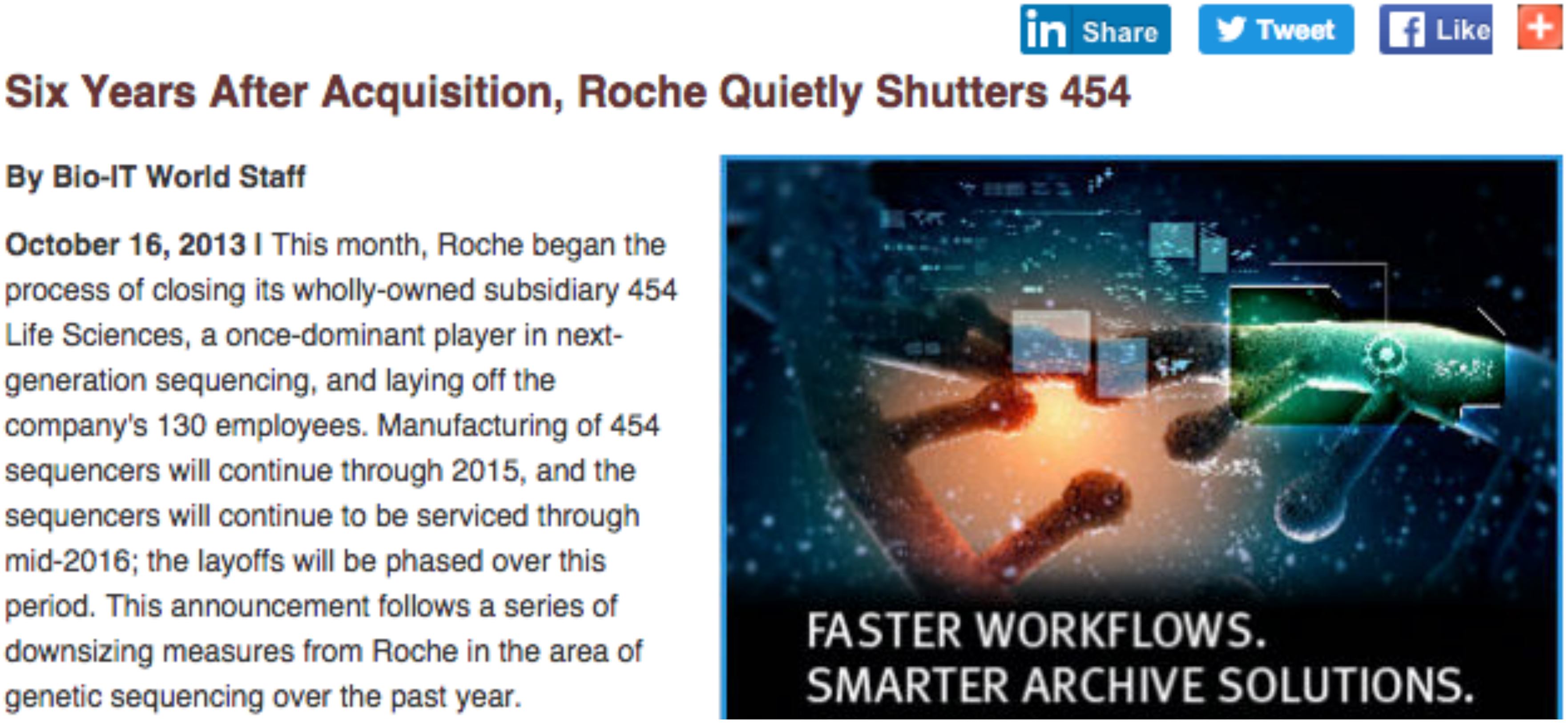
Pyrosequencing technology



(Mardis E.R. (2008) Trends in Genetics 24: 133-141)

Sequencing systems: 454

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



Six Years After Acquisition, Roche Quietly Shutters 454

By Bio-IT World Staff

October 16, 2013 l 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.

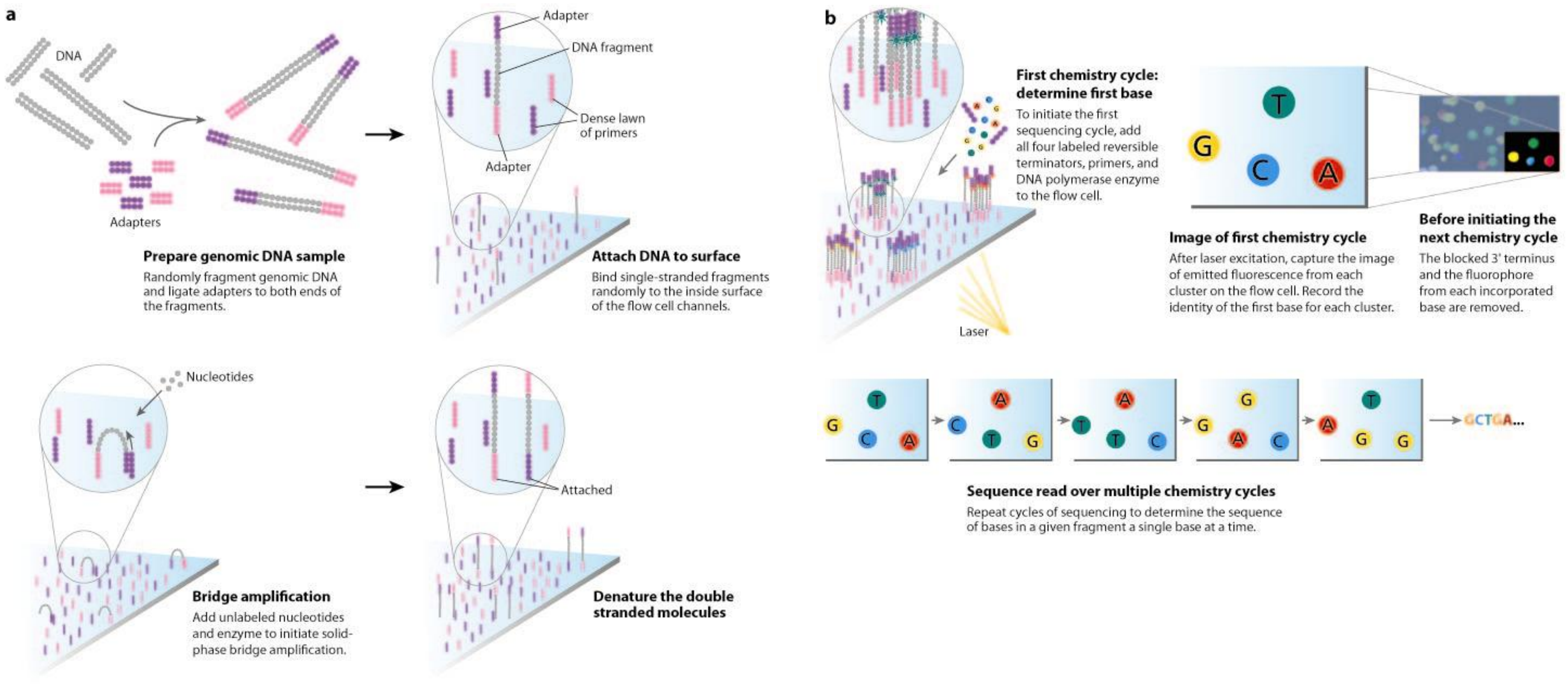
Sequencing systems: Illumina

http://www.illumina.com/

3. Short read sequencing technologies

Sequencing systems: Illumina

Sequence by Synthesis technology



Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

(Mardis E.R. (2013) Annual Review of Analytical Chemistry 6: 287-303)

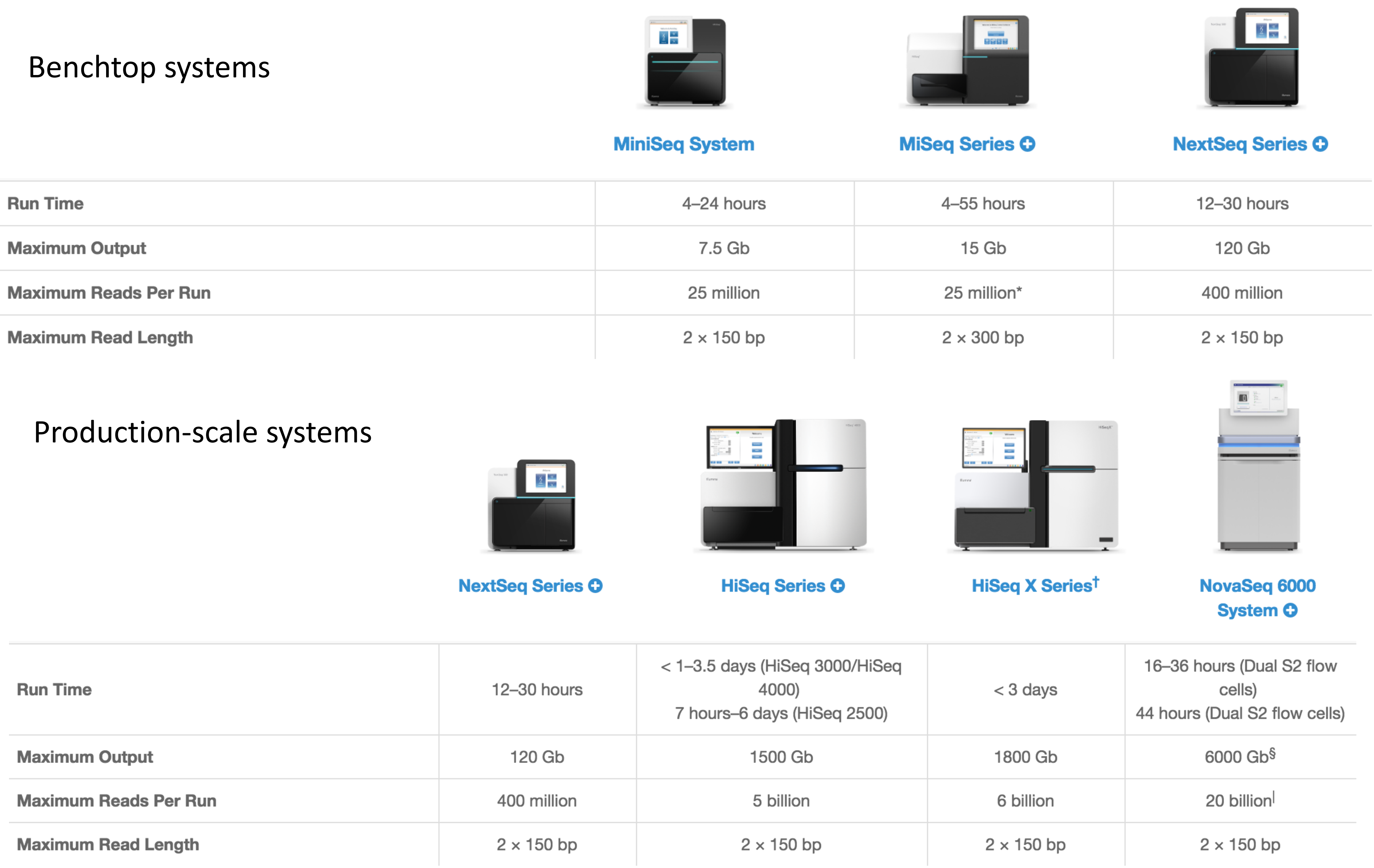
Sequencing systems: Illumina

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[‖]
Maximum Read Length	2 × 150 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp

Sequencing systems: SOLiD

https://products.appliedbiosystems.com

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

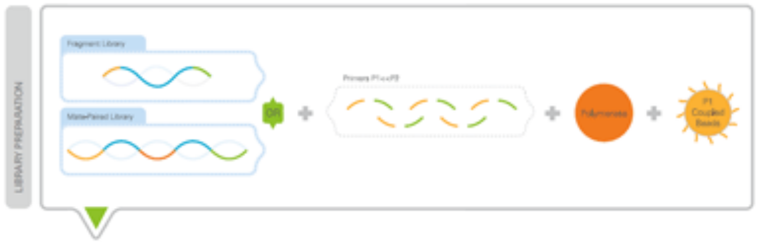


Figure 1

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

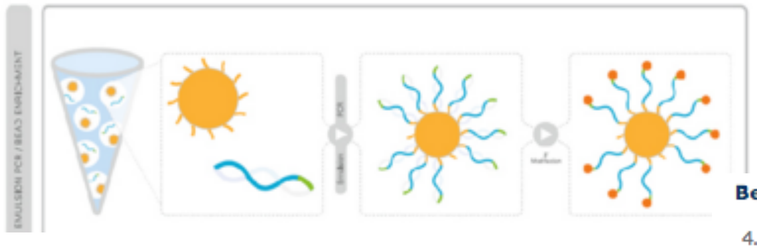


Figure 2

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.

View Larger Image

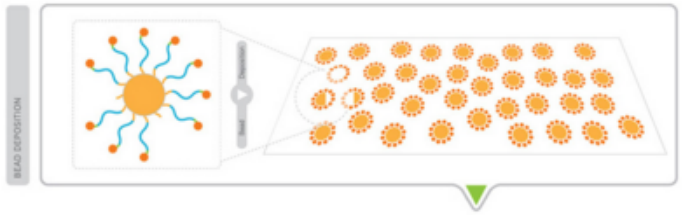
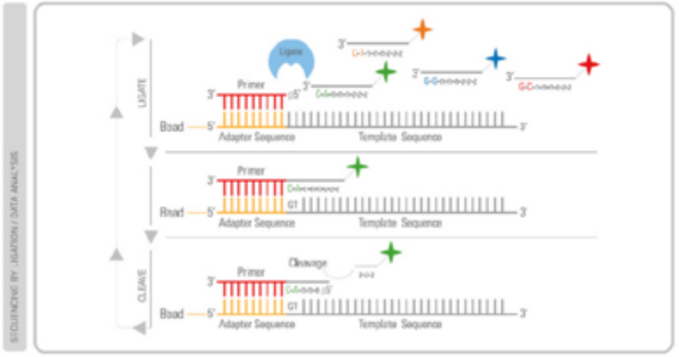


Figure 3

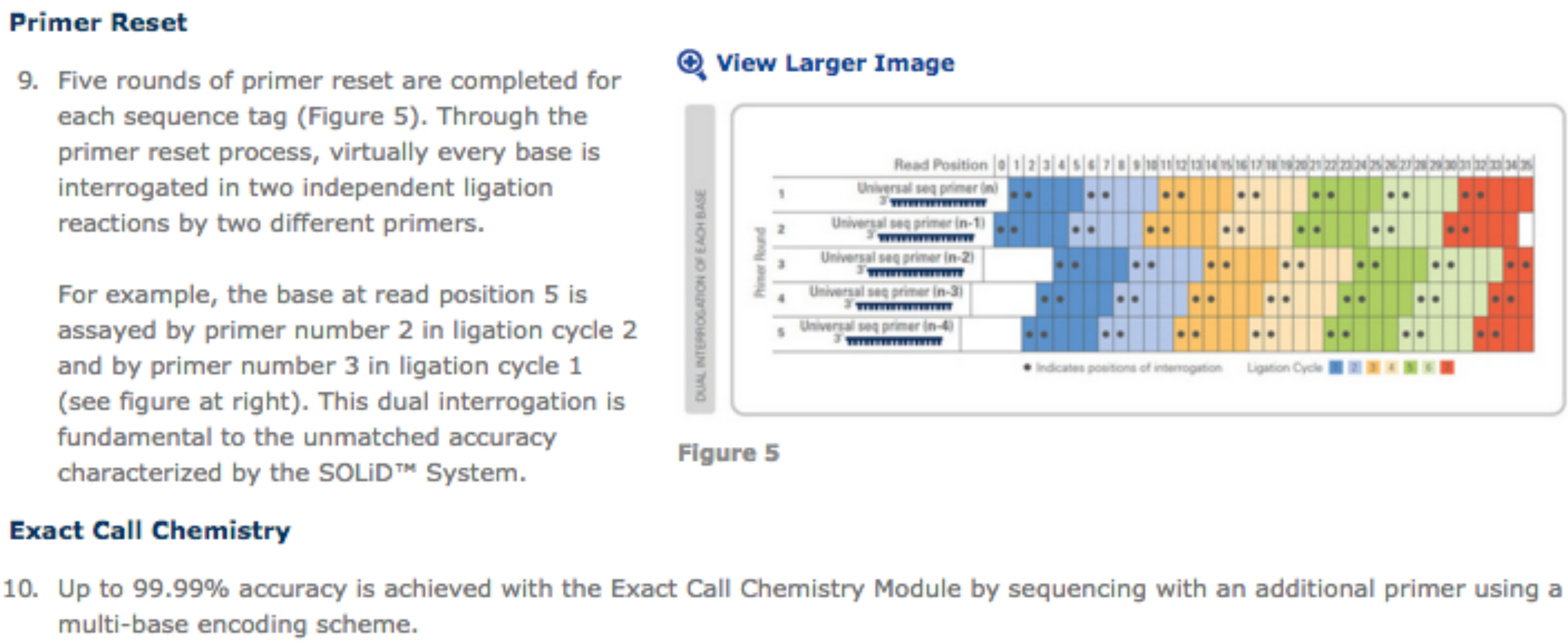
Sequencing by Ligation

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



Sequencing systems: SOLiD



Primer Reset

9. 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.

View Larger Image

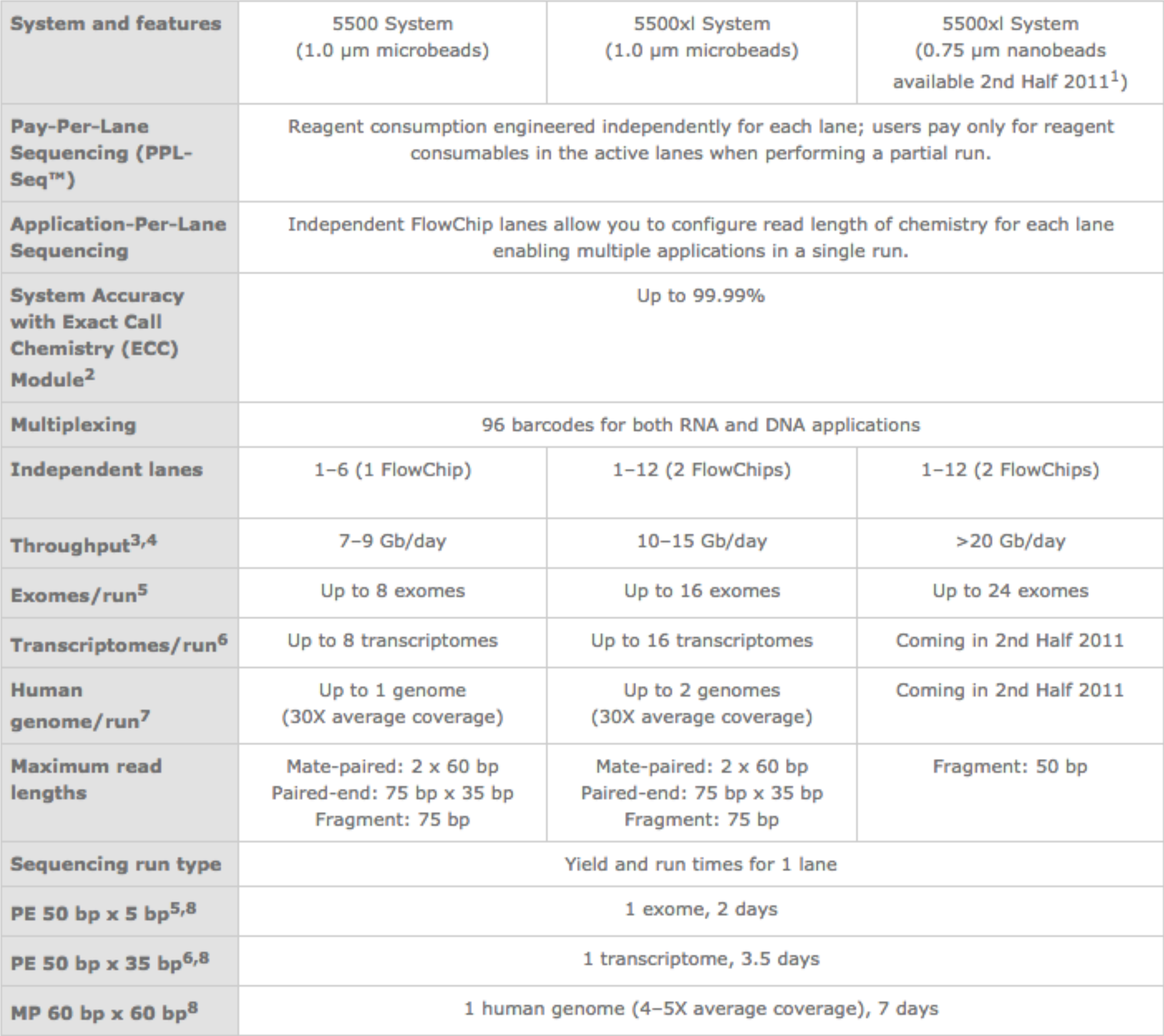
DUAL INTERROGATION OF EACH BASE
Read Position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer Round
1 Universal seq primer (n) 3'
2 Universal seq primer (n-1) 3'
3 Universal seq primer (n-2) 3'
4 Universal seq primer (n-3) 3'
5 Universal seq primer (n-4) 3'
● Indicates positions of interrogation
Ligation Cycle 1 2 3 4 5 6 7

Figure 5

Exact Call Chemistry

10. 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.

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) Module[2]	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/run[5]	Up to 8 exomes	Up to 16 exomes	Up to 24 exomes
Transcriptomes/run[6]	Up to 8 transcriptomes	Up to 16 transcriptomes	Coming in 2nd Half 2011
Human genome/run[7]	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 bp[8]	1 human genome (4–5X average coverage), 7 days		

Sequencing systems: SOLiD

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

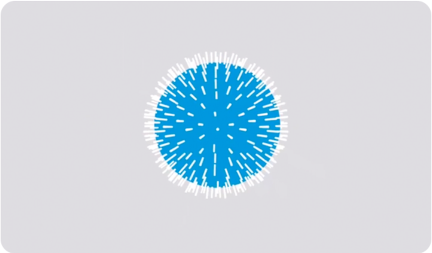
Sequencing systems: Ion Torrent

https://products.appliedbiosystems.com

Sequencing systems: Ion Torrent

Sequence by Semiconductor technology

Copy DNA

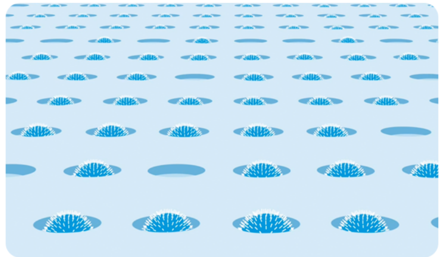


A sample of DNA is cut into millions of fragments, and each fragment is attached to its own bead

The fragment is copied until it covers the bead

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

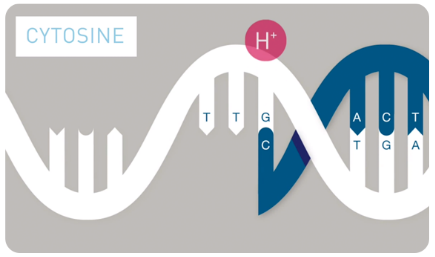
Load chip



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

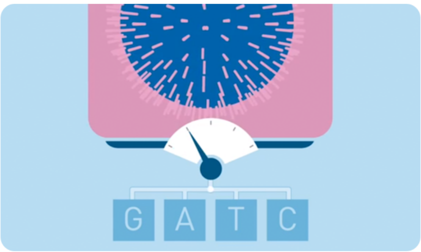
Incorporate nucleotide



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



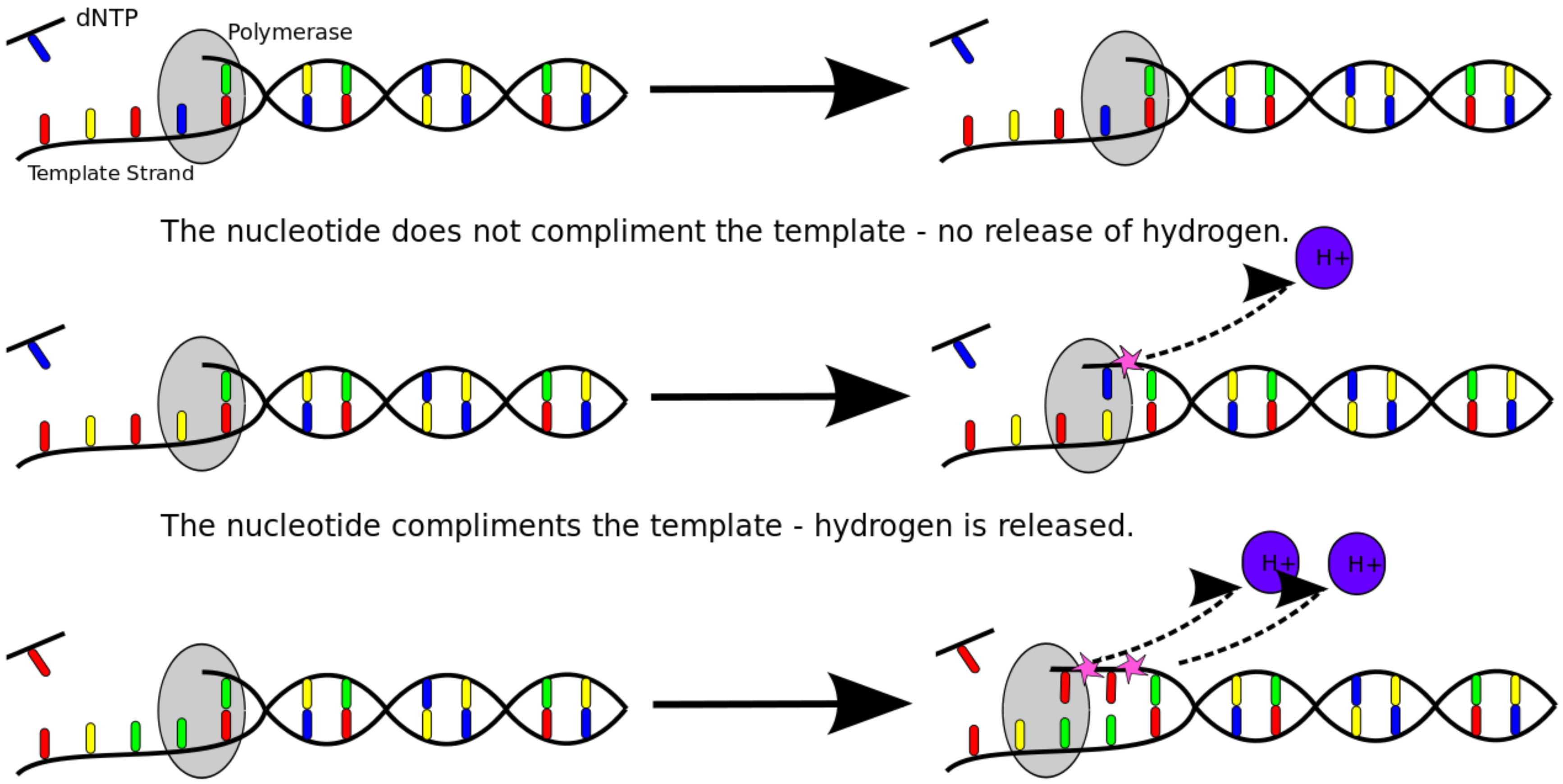
An ion-sensitive layer beneath the well measures that pH change and converts it to voltage

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

This process happens simultaneously in millions of wells

Sequencing systems: Ion Torrent

Sequence by Semiconductor technology



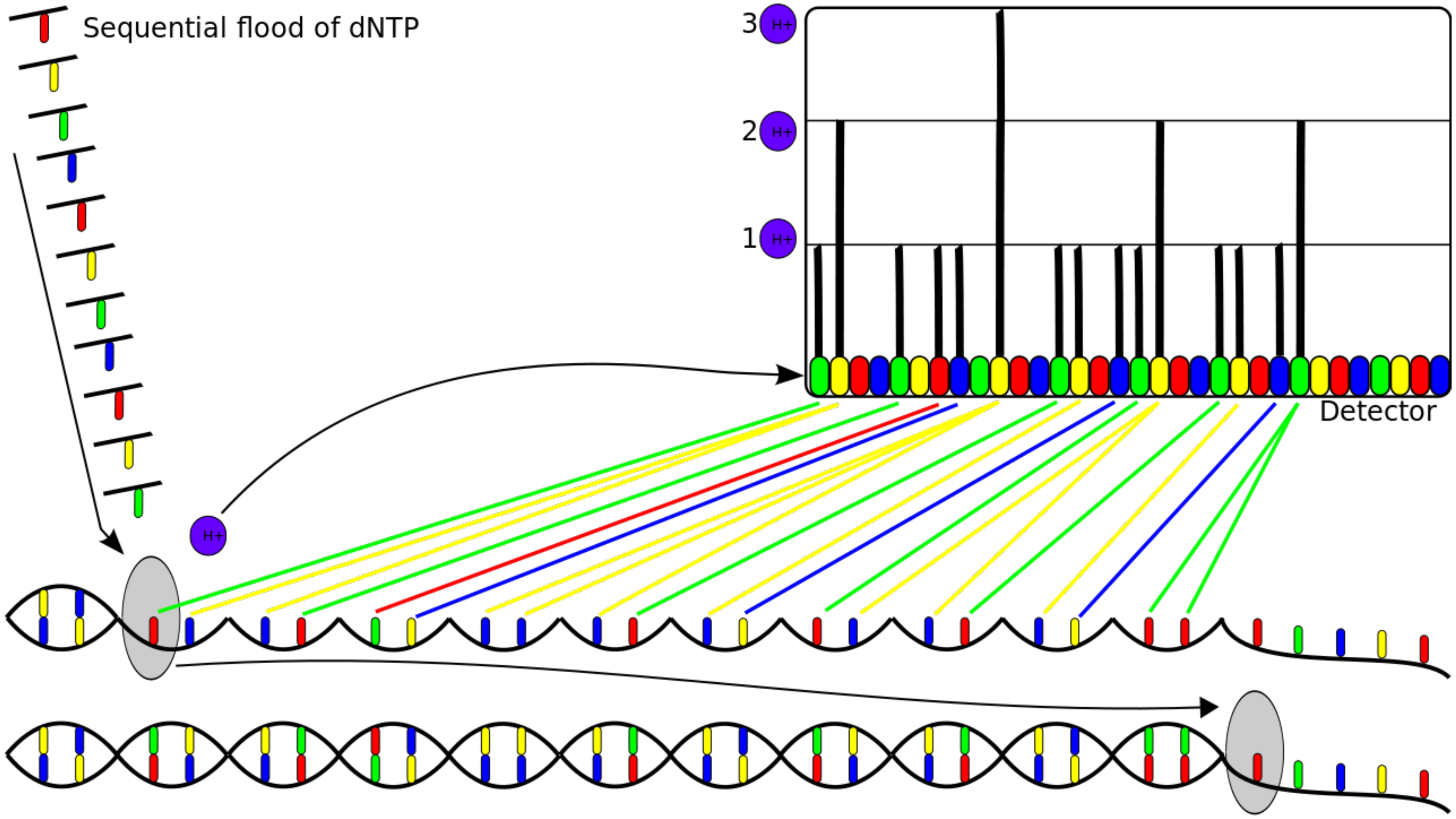
The nucleotide does not compliment the template - no release of hydrogen.

The nucleotide compliments the template - hydrogen is released.

The nucleotide compliments several bases in a row - multiple hydrogen ions are released.

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

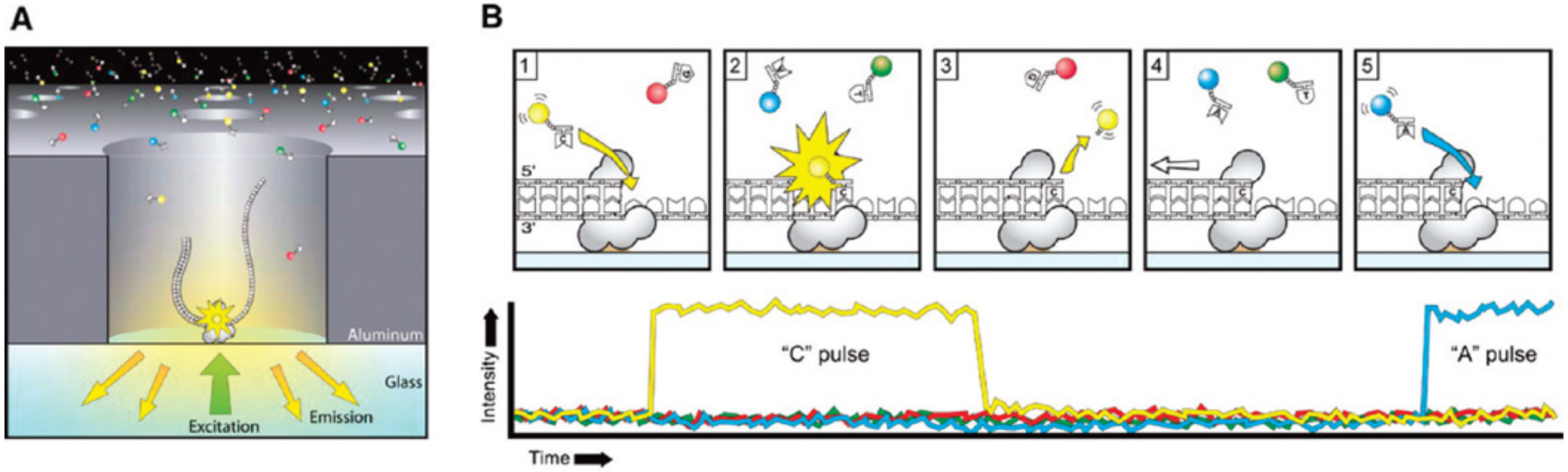
5. Common file formats

Sequencing systems: Pacific Biosystems (PacBio)

http://www.pacb.com/

Sequencing systems: Pacific Biosystems (PacBio)

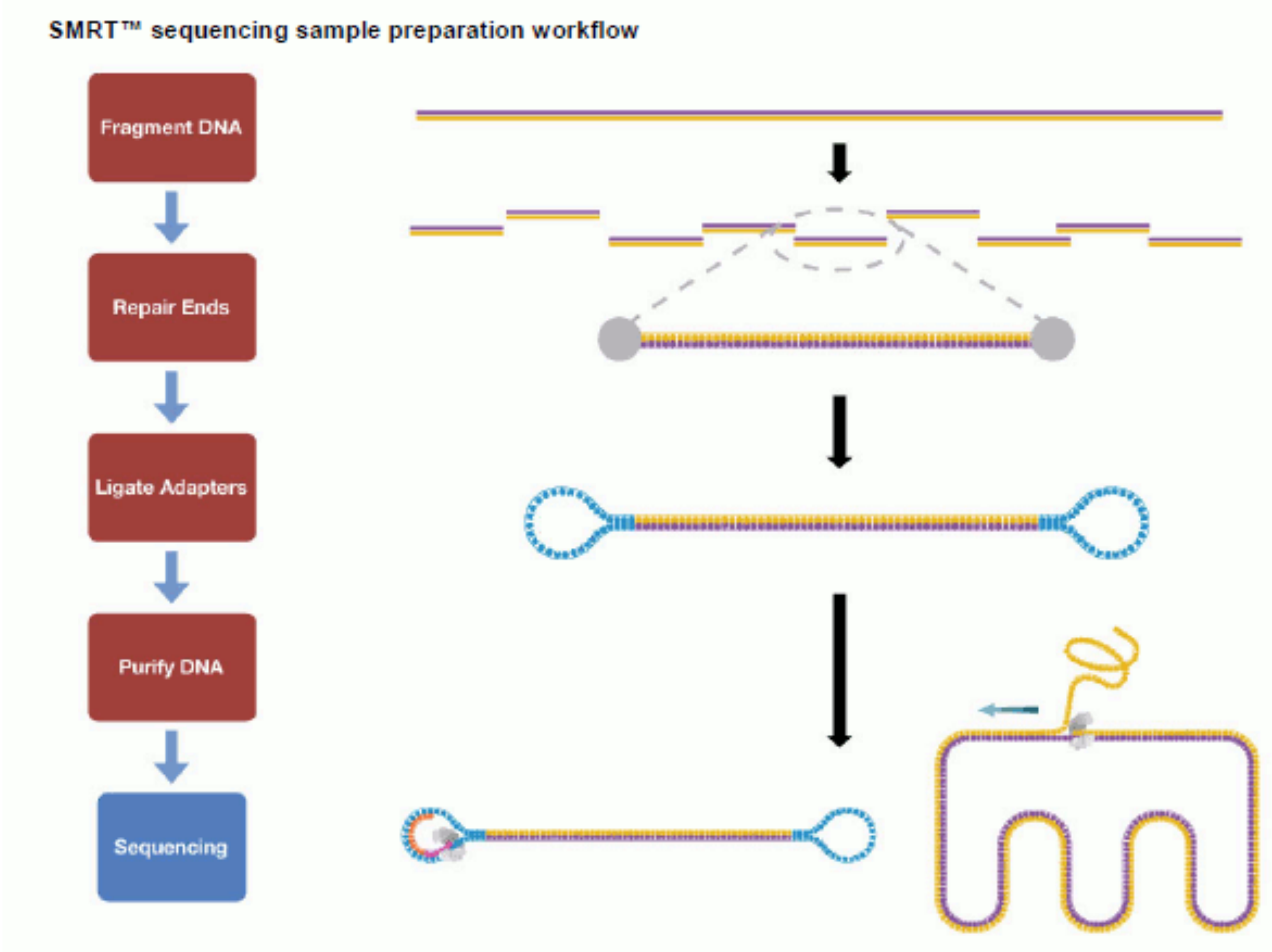
Single Molecule Real Time (SMRT) technology



Niedringhaus et al. Analytical Chemistry 2011

Sequencing systems: Pacific Biosystems (PacBio)

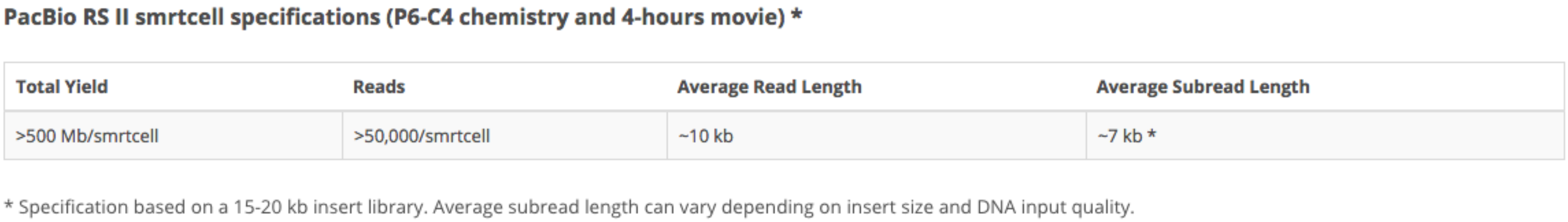
Single Molecule Real Time (SMRT) technology



Niedringhaus et al. Analytical Chemistry 2011

Sequencing systems: Pacific Biosystems (PacBio)

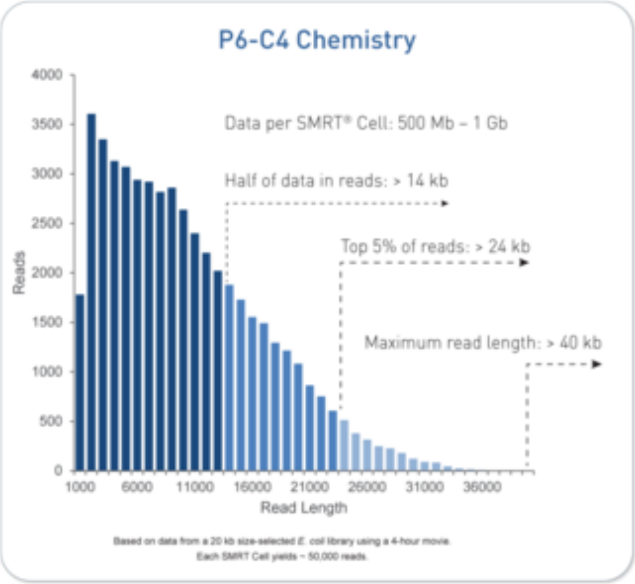
Single Molecule Real Time (SMRT) technology



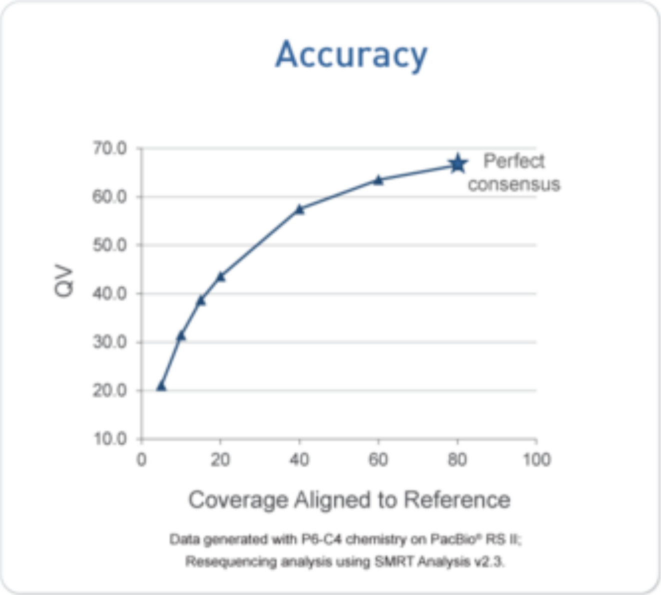
PacBio RS II smrtcell specifications (P6-C4 chemistry and 4-hours movie) *

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.



Based on data from a 20 kb size-selected *E. coli* library using a 4-hour movie.
Each SMRT Cell yields ~ 50,000 reads.

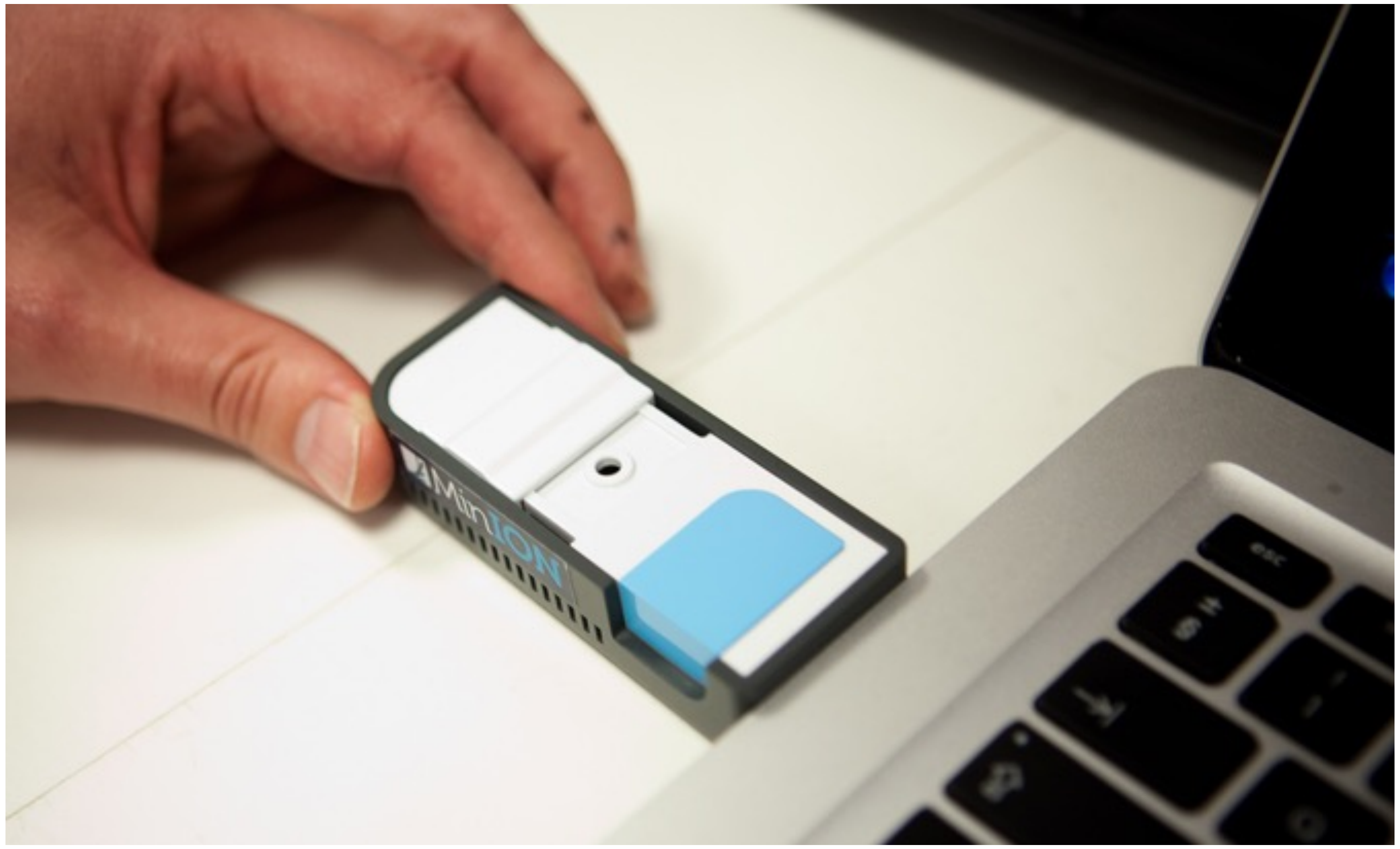


Data generated with P6-C4 chemistry on PacBio® RS II;
Resequencing analysis using SMRT Analysis v2.3.

http://genome.duke.edu/cores-and-services/sequencing-and-genomic-technologies/pacbio

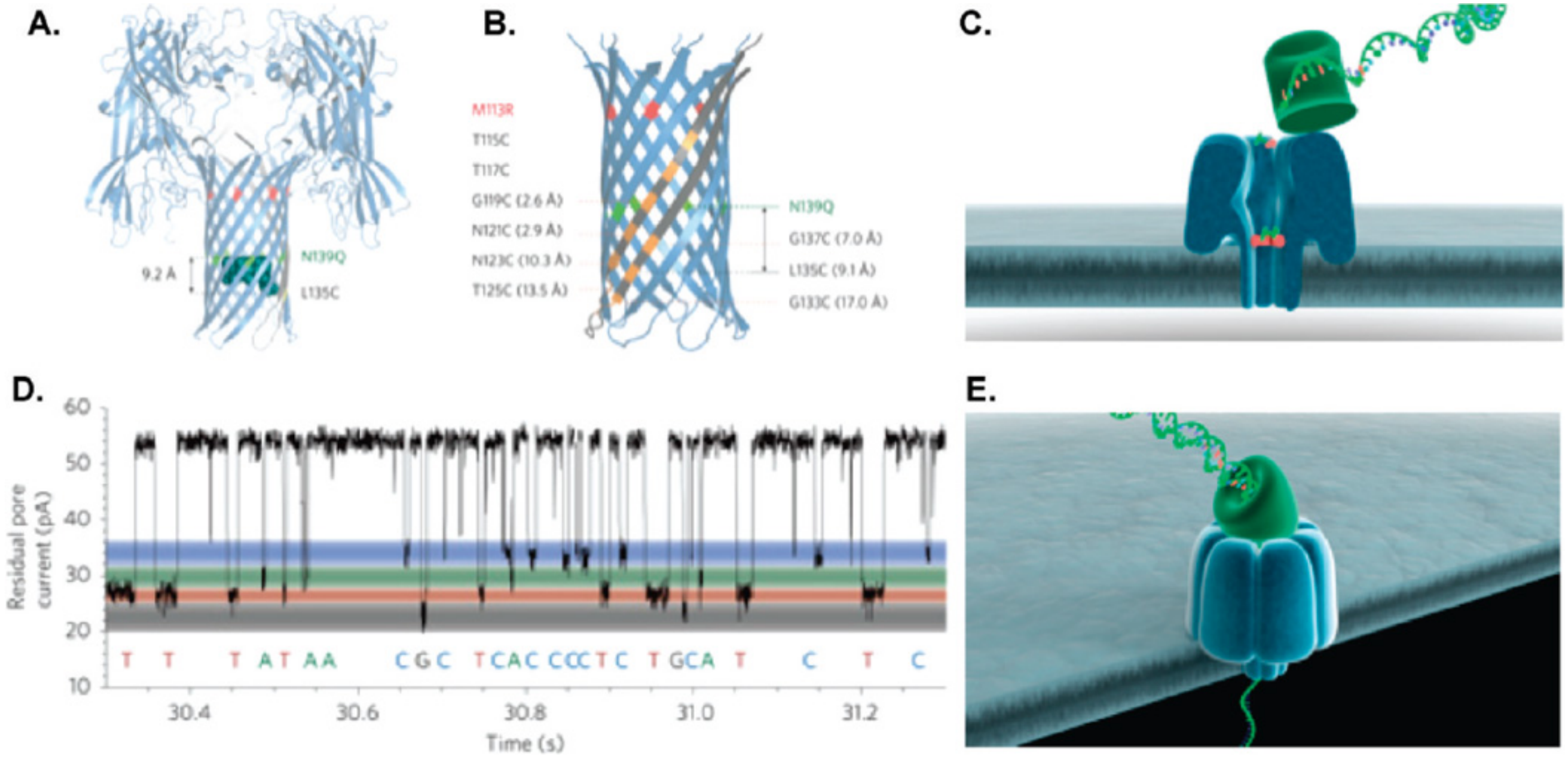
Sequencing systems: Oxford Nanopore (ON)

https://www.nanoporetech.com/



Sequencing systems: Oxford Nanopore (ON)

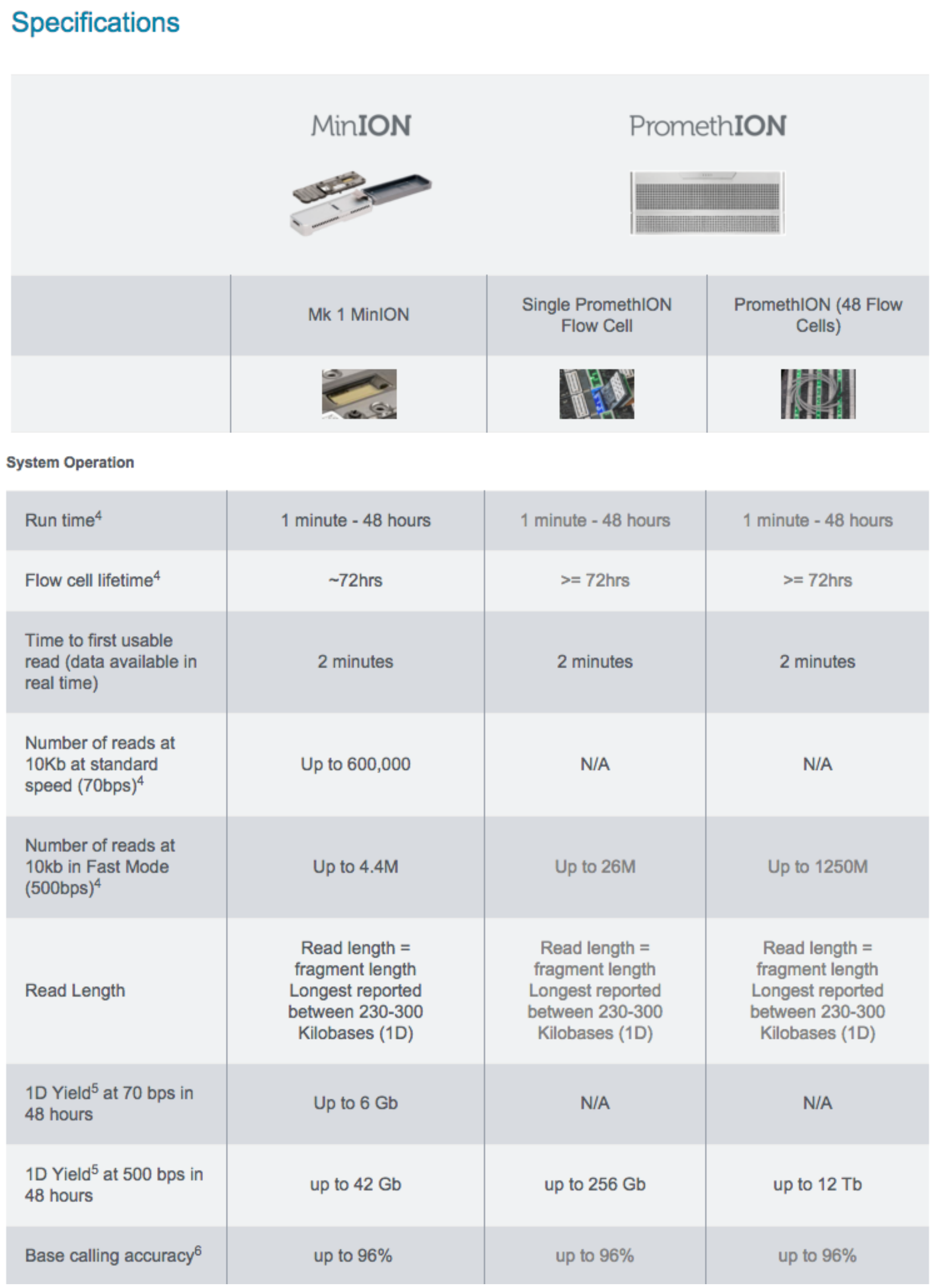
Sequence by Nanopore technology



Niedringhaus et al. Analytical Chemistry 2011

Sequencing systems: Oxford Nanopore (ON)

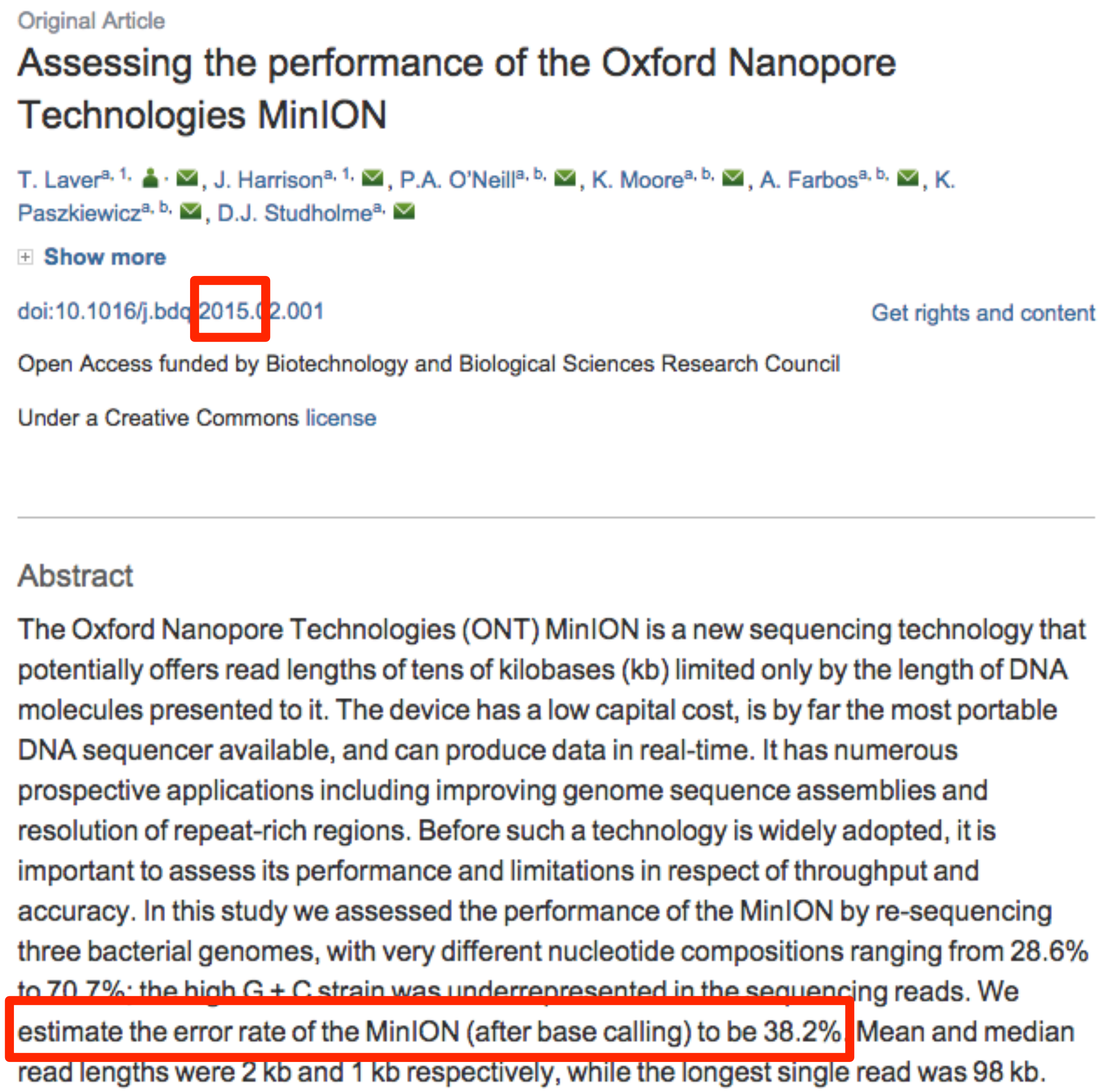
Sequence by Nanopore technology



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%

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.bdq.2015.02.001

Get rights and content

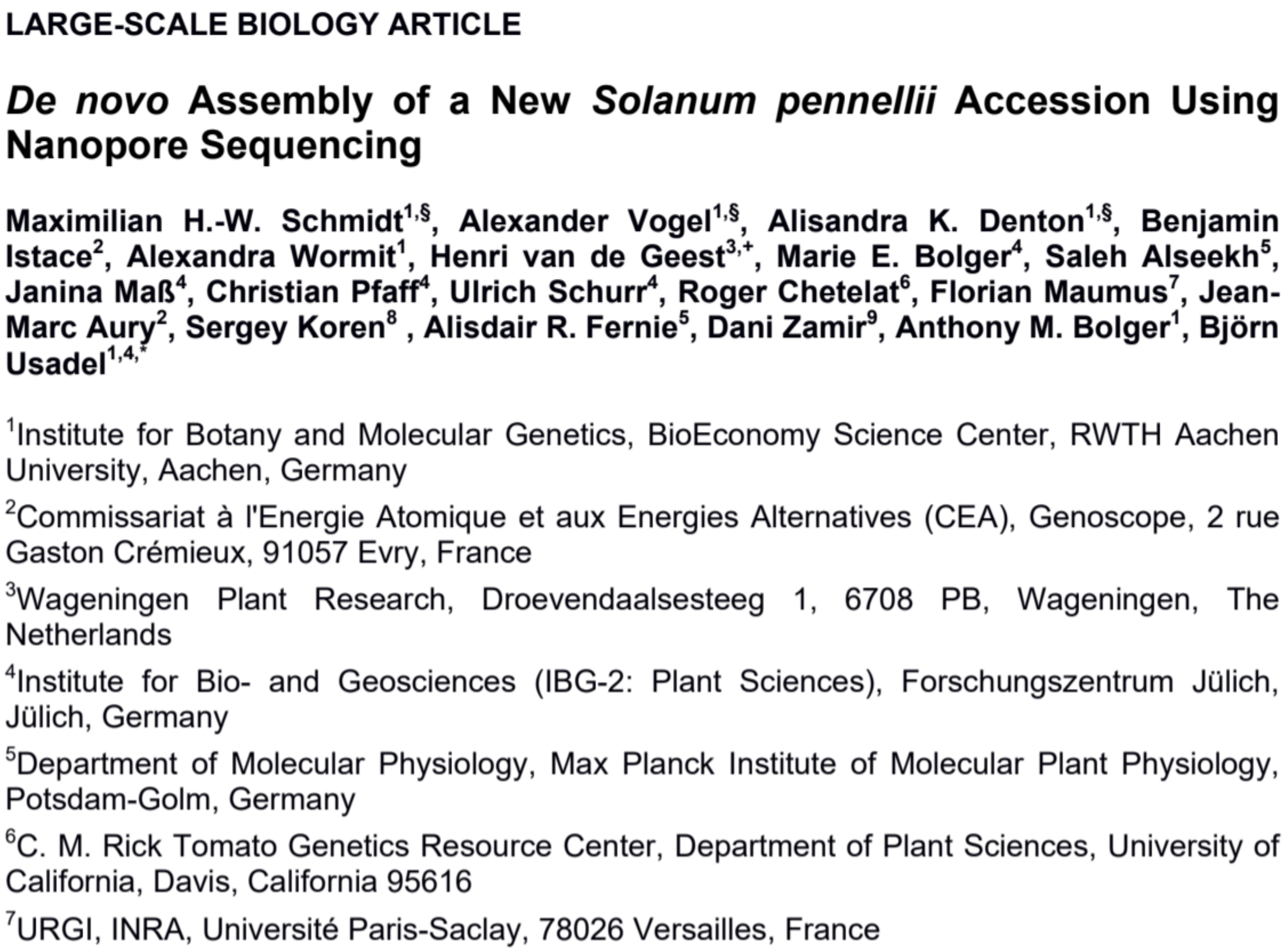
Open Access funded by Biotechnology and Biological Sciences Research Council

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.

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 Istace[2], Alexandra Wormit[1], Henri van de Geest[3,+], Marie E. Bolger[4], Saleh Alseekh[5], Janina Maß[4], Christian Pfaff[4], Ulrich Schurr[4], Roger Chetelat[6], Florian Maumus[7], Jean-Marc Aury[2], Sergey Koren[8] , Alisdair R. Fernie[5], Dani Zamir[9], Anthony M. Bolger[1], Björn Usadel[1,4,*]

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

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

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

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

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

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

[7]URGI, 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

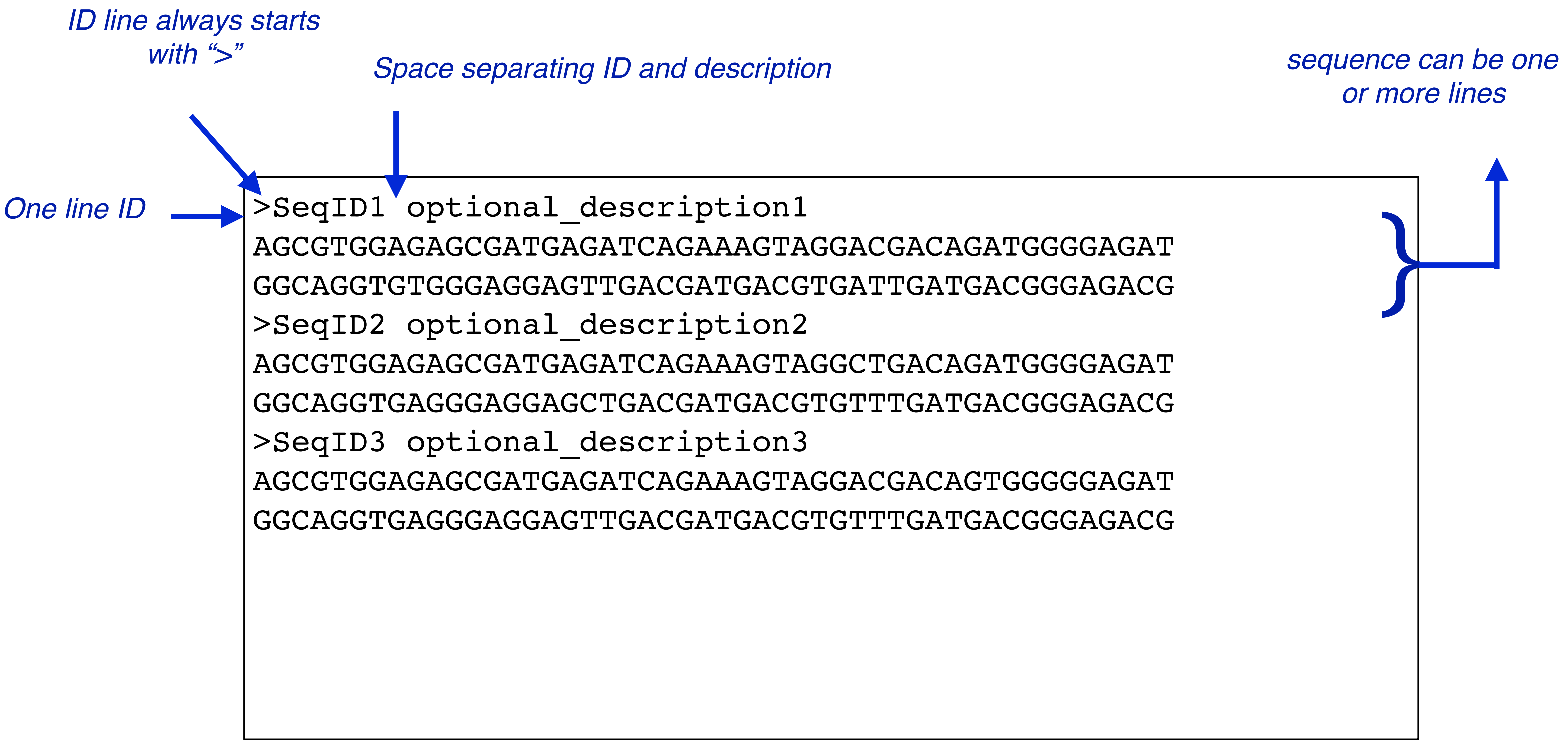
4. Long read sequencing technologies

5. Common file formats

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…



ID line always starts with “>”

Space separating ID and description

sequence can be one or more lines

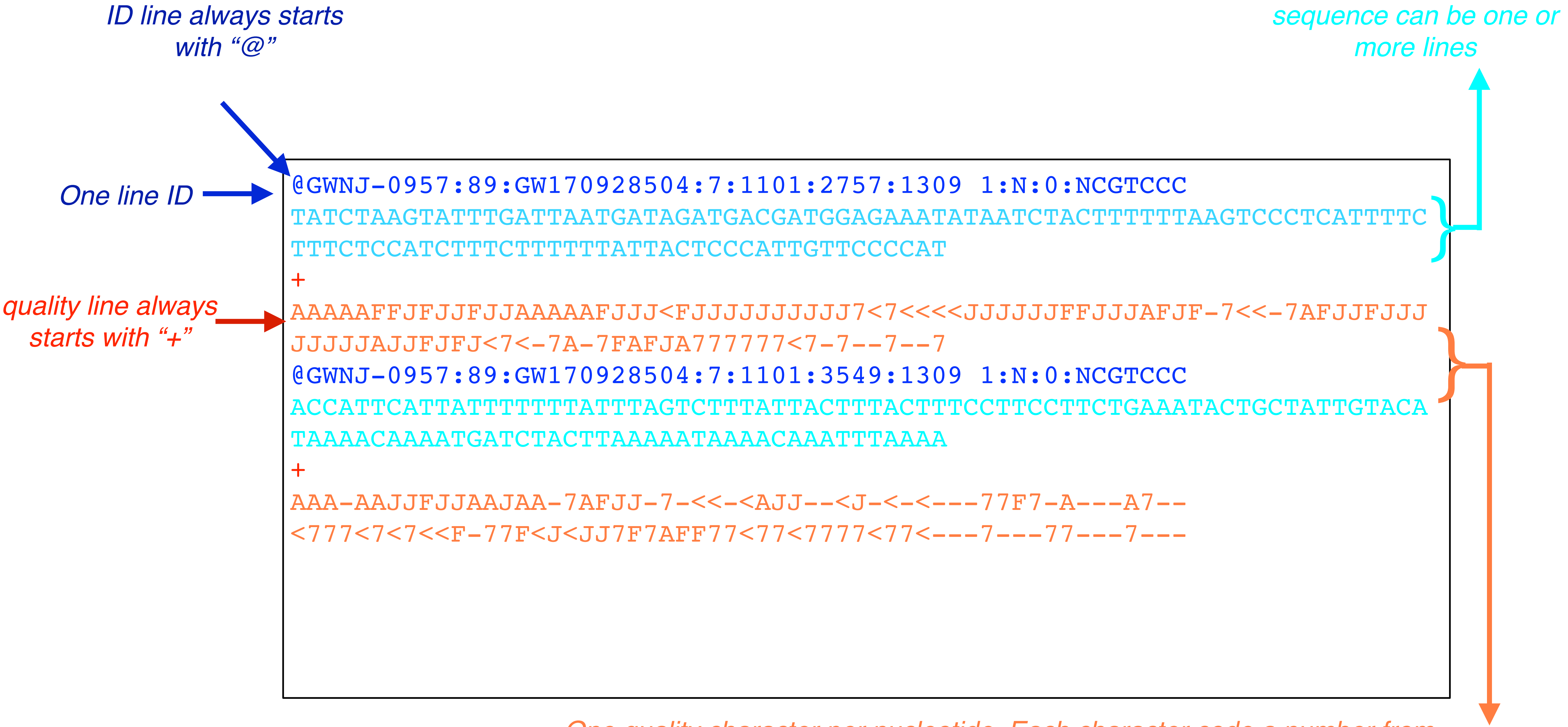
One line ID

```
>SeqID1 optional_description1
AGCGTGGAGAGCGATGAGATCAGAAAGTAGGACGACAGATGGGGAGAT
GGCAGGTGTGGGAGGAGTTGACGATGACGTGATTGATGACGGGAGACG
>SeqID2 optional_description2
AGCGTGGAGAGCGATGAGATCAGAAAGTAGGCTGACAGATGGGGAGAT
GGCAGGTGAGGGAGGAGCTGACGATGACGTGTTTGATGACGGGAGACG
>SeqID3 optional_description3
AGCGTGGAGAGCGATGAGATCAGAAAGTAGGACGACAGTGGGGGAGAT
GGCAGGTGAGGGAGGAGTTGACGATGACGTGTTTGATGACGGGAGACG
```

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 “+”

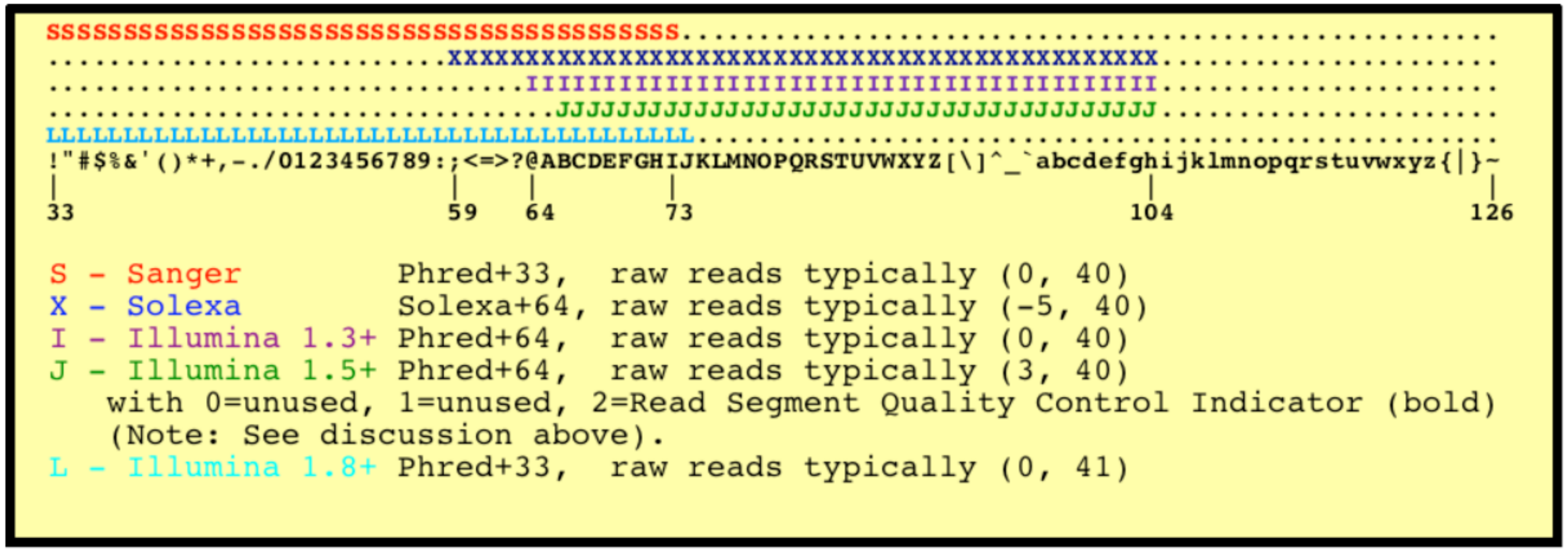
```
@GWNJ-0957:89:GW170928504:7:1101:2757:1309 1:N:0:NCGTCCC
TATCTAAGTATTTGATTAATGATAGATGACGATGGAGAAATATAATCTACTTTTTTAAGTCCCTCATTTTC
TTTCTCCATCTTTCTTTTTTATTACTCCCATTGTTCCCCAT
+
AAAAAFFJFJJFJJAAAAAFJJJ<FJJJJJJJJJJ7<7<<<<JJJJJJFFJJJAFJF-7<<-7AFJJFJJJ
JJJJJAJJFJFJ<7<-7A-7FAFJA777777<7-7--7--7
@GWNJ-0957:89:GW170928504:7:1101:3549:1309 1:N:0:NCGTCCC
ACCATTCATTATTTTTTTATTTAGTCTTTATTACTTTACTTTCCTTCCTTCTGAAATACTGCTATTGTACA
TAAAACAAAATGATCTACTTAAAAATAAAACAAATTTAAAA
+
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.



```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126

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)
```

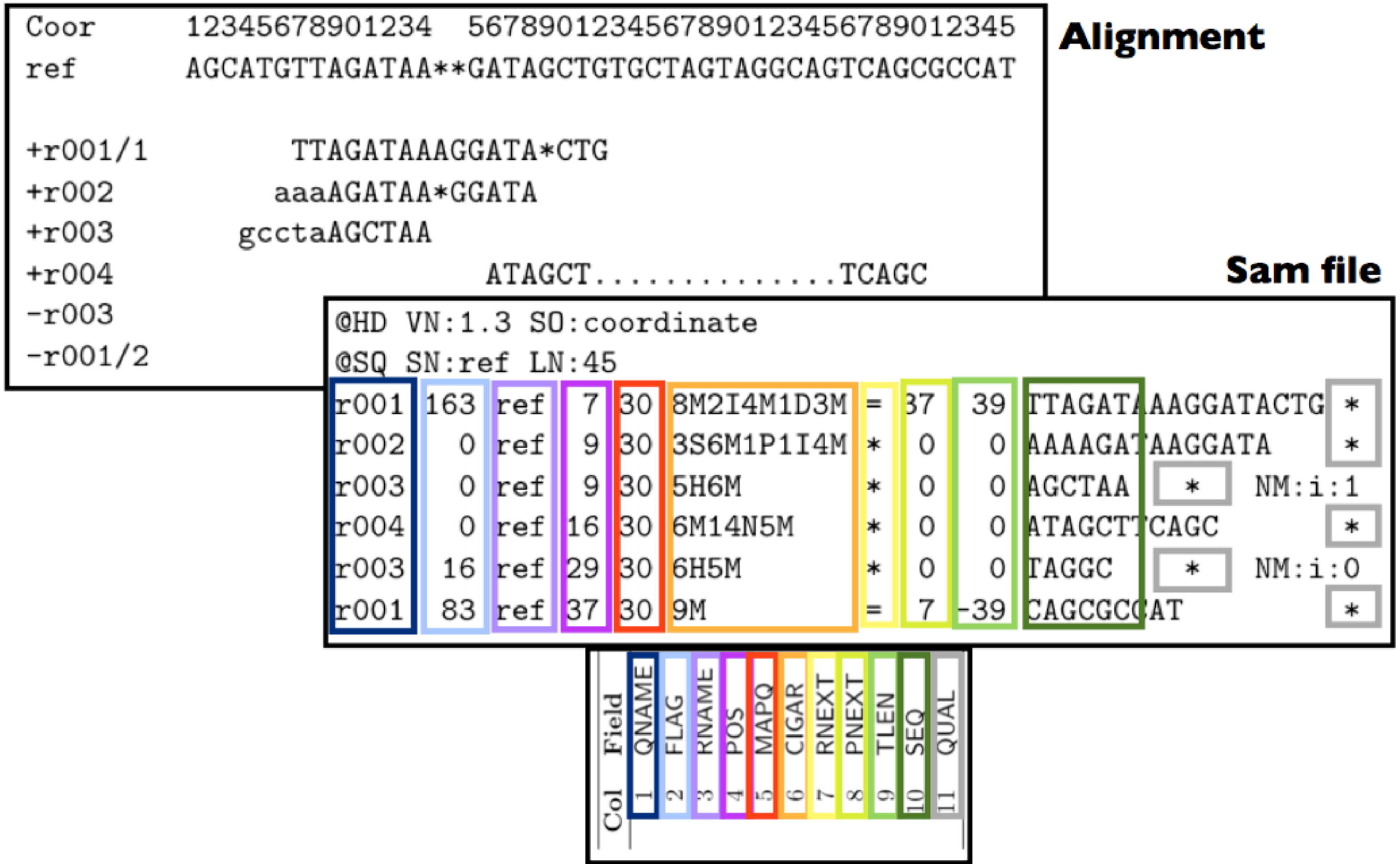
Phred score of a base is: **Qphred=-10 log10 (e)**

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.



Alignment

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

Sam file

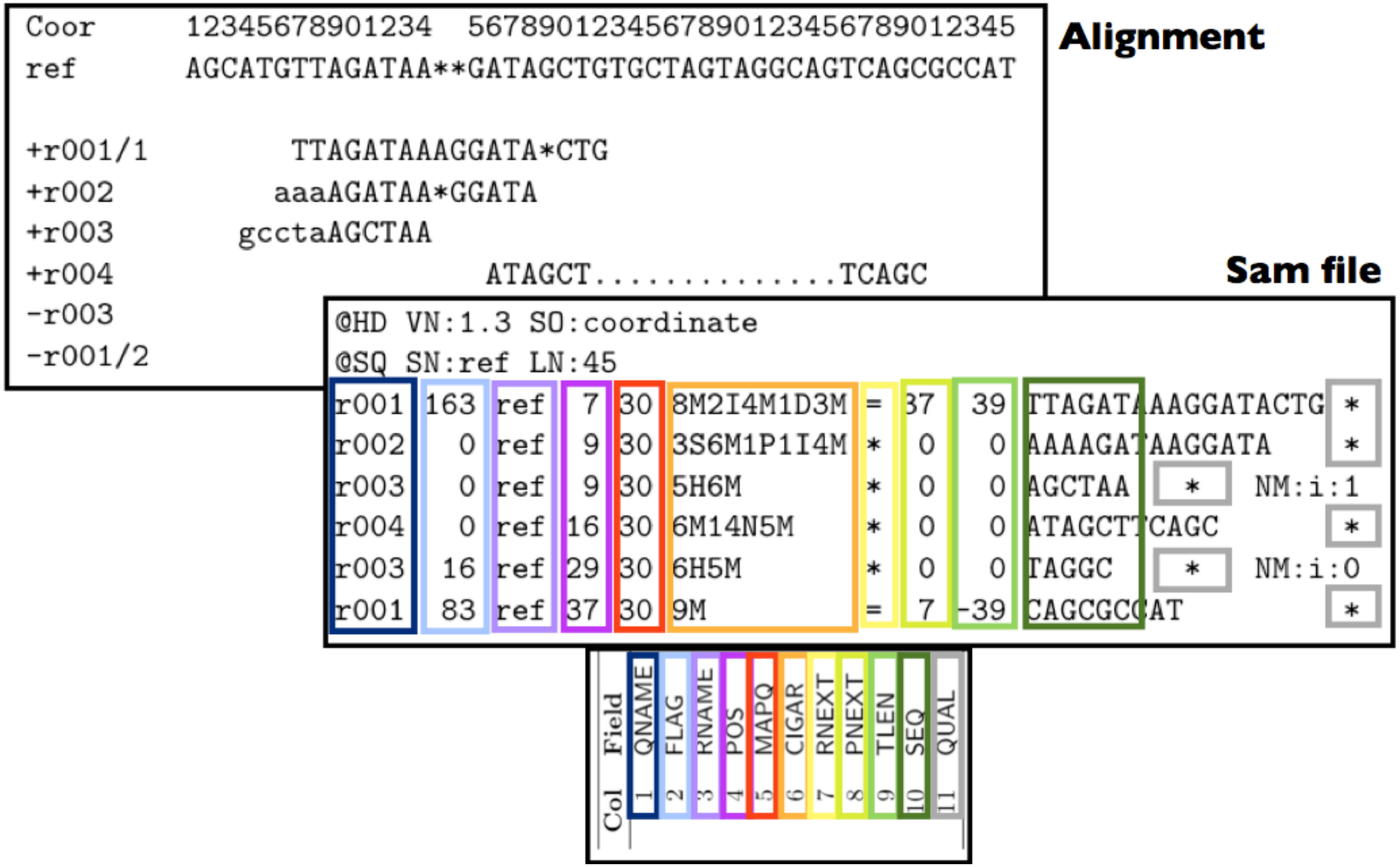
```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
r001 163 ref  7 30 8M2I4M1D3M  = 37  39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M  *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30 5H6M        *  0   0 AGCTAA  *   NM:i:1
r004   0 ref 16 30 6M14N5M     *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 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.



Alignment

```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

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

Sam file

```
@HD VN:1.3 SO:coordinate
@SQ SN:ref LN:45
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         *
```

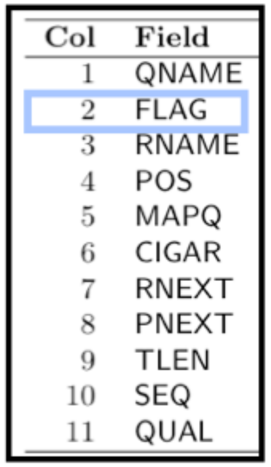
Col	Field
1	QNAME
2	FLAG
3	RNAME
4	POS
5	MAPQ
6	CIGAR
7	RNEXT
8	PNEXT
9	TLEN
10	SEQ
11	QUAL

5. Common file formats

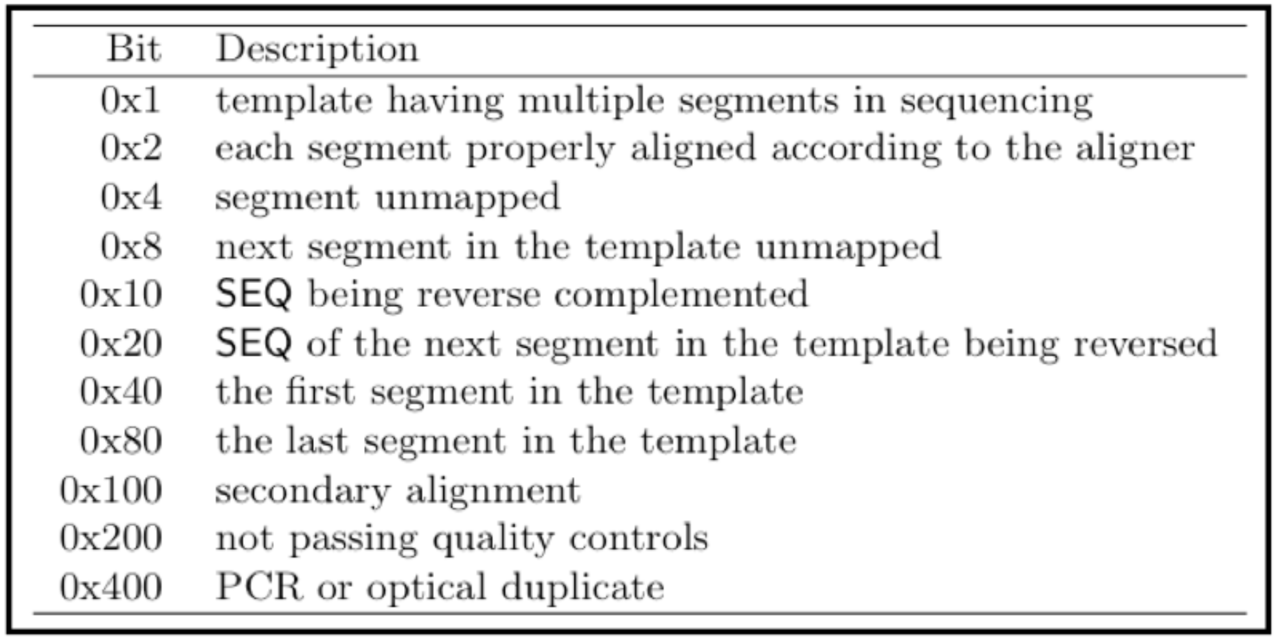
III. SAM/BAM

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

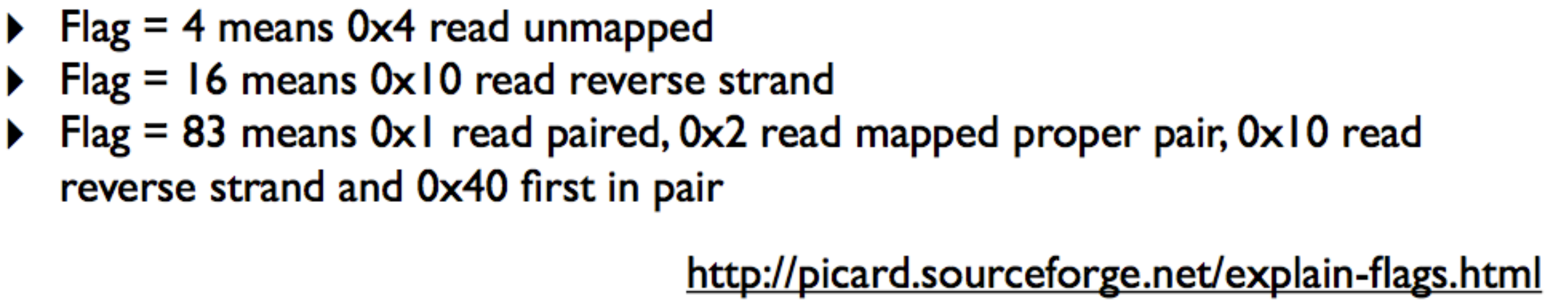
The 2nd column: FLAG defines the status of the read mapping.



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



Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
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

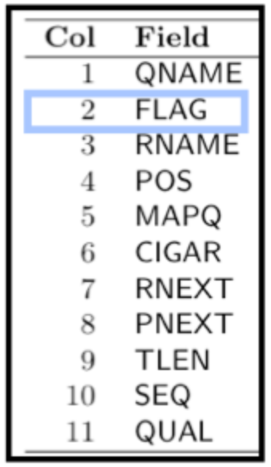
http://picard.sourceforge.net/explain-flags.html

5. Common file formats

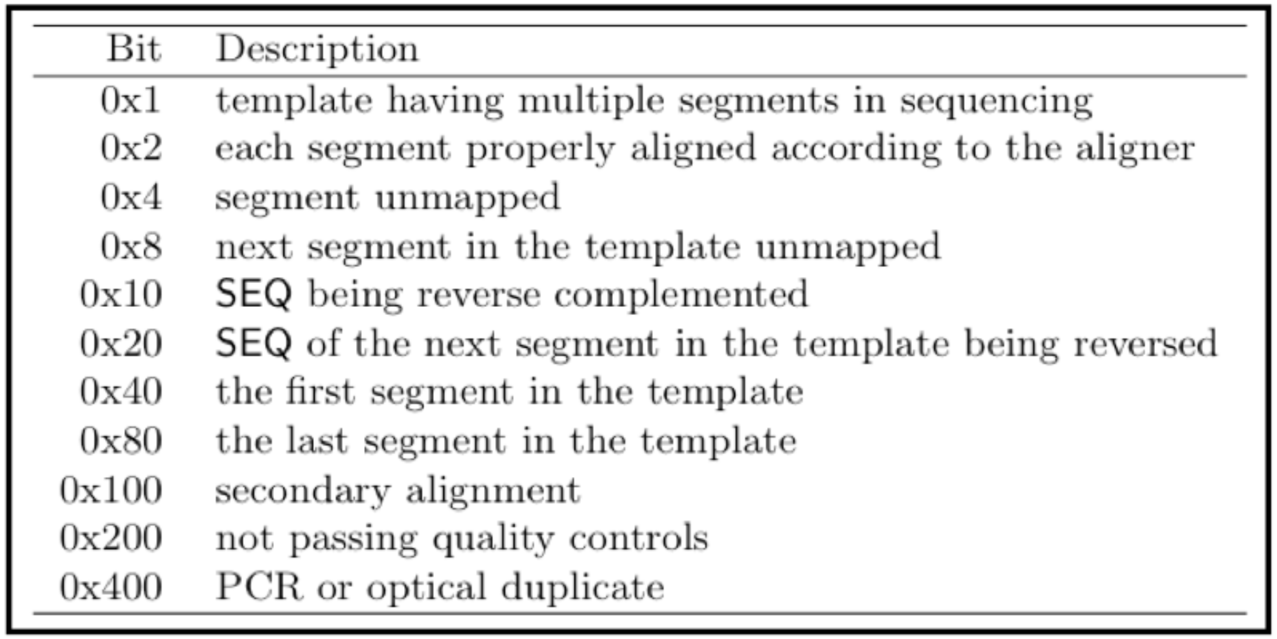
III. SAM/BAM

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

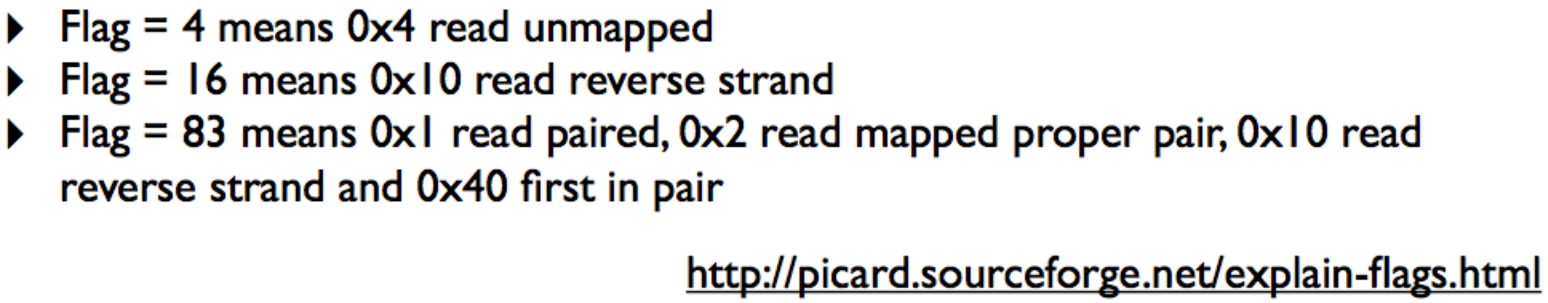
The 2nd column: FLAG defines the status of the read mapping.



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



Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
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

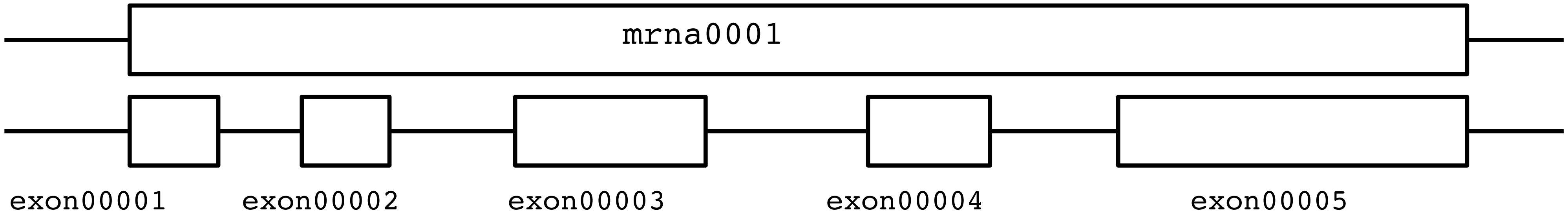
5. Common file formats

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
```

seqid	source	type	start	end	score	strand	phase	attributes
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



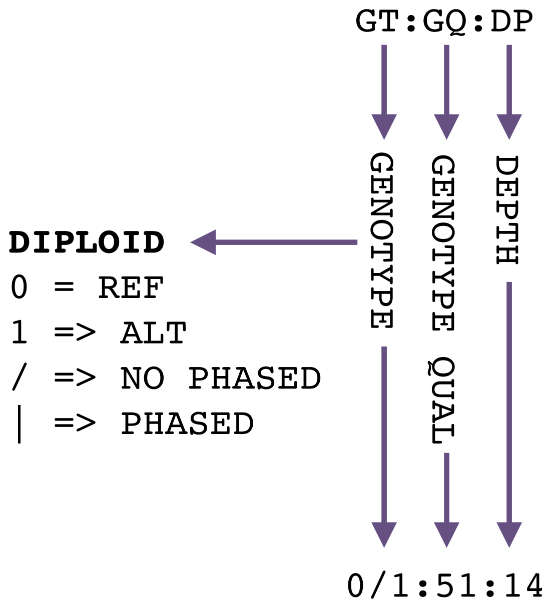
5. Common file formats

V. VCF

VCF is a text-based file with 8 fixed columns and one extra per sample for the multisample files. It contacts 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



Epilogue

NGS Applications

Whole Genome Sequencing

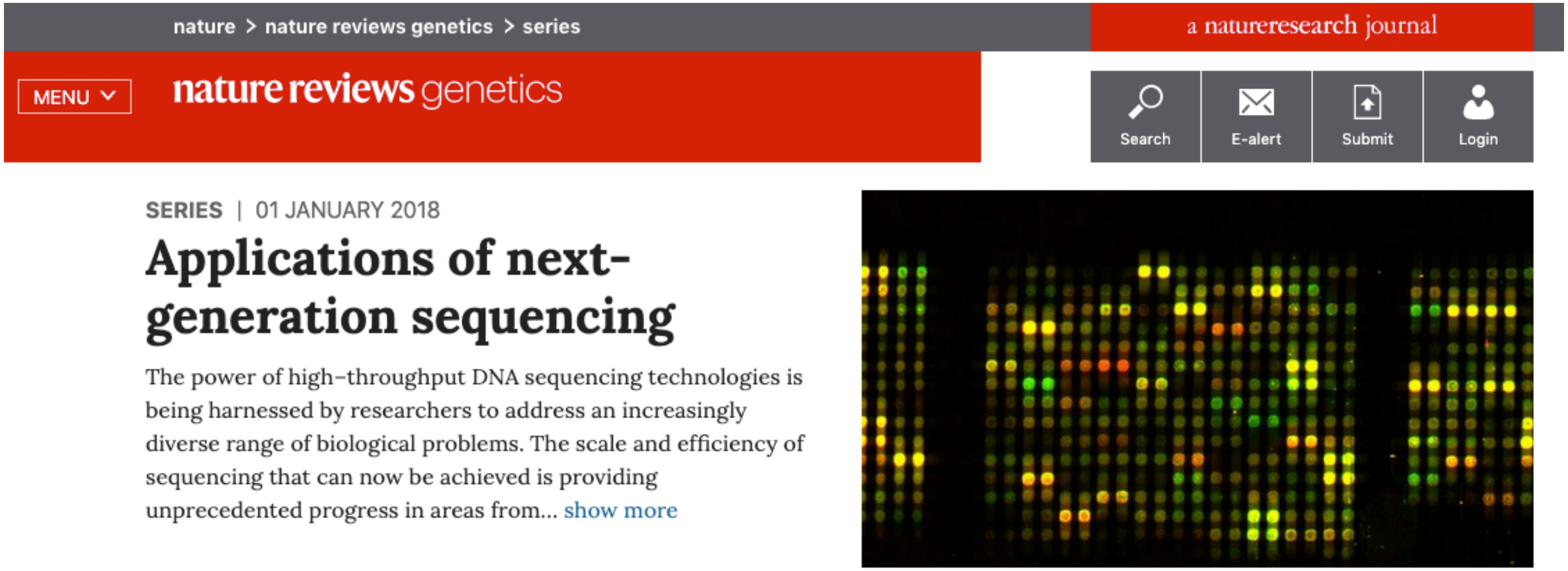
Transcriptome Sequencing

Reduced Representation Approaches

Bisulphite Sequencing

Chromatin Contact Sequencing

Applications



SERIES | 01 JANUARY 2018

Applications of next-generation sequencing

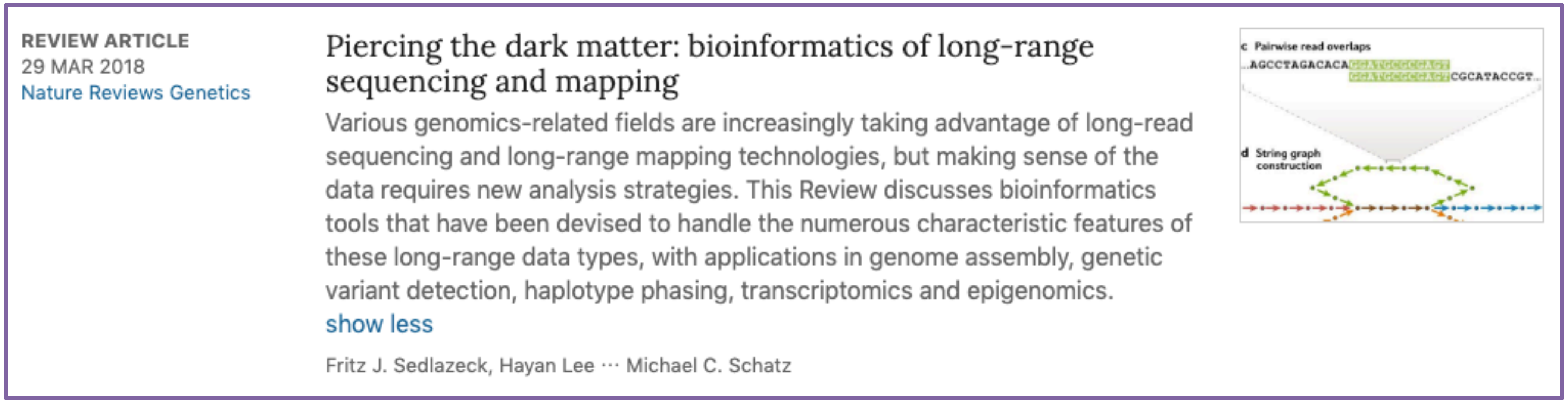
The power of high-throughput DNA sequencing technologies is being harnessed by researchers to address an increasingly diverse range of biological problems. The scale and efficiency of sequencing that can now be achieved is providing unprecedented progress in areas from... show more

Homework:

1. Select one article from group 1 and another one of group 2.
2. Read critically and summarise the article in three tweets.
3. Prepare two questions for each article.
4. Send me (aureliano.bombarely@unimi.it) the tweets and the questions by March 31st by email. Use the Subject “Bibliographic Work Genomics 2020”.
5. I will send you back a question from the list of question that I will produce and you will need to send me the answer by April 7th, 2020.

Applications

Group 1

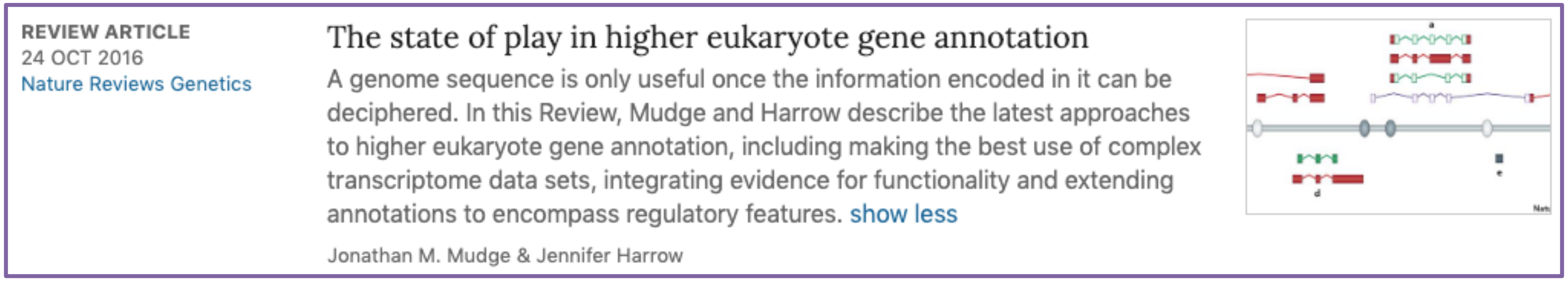


REVIEW ARTICLE
29 MAR 2018
Nature Reviews Genetics

Piercing the dark matter: bioinformatics of long-range sequencing and mapping

Various genomics-related fields are increasingly taking advantage of long-read sequencing and long-range mapping technologies, but making sense of the data requires new analysis strategies. This Review discusses bioinformatics tools that have been devised to handle the numerous characteristic features of these long-range data types, with applications in genome assembly, genetic variant detection, haplotype phasing, transcriptomics and epigenomics. show less

Fritz J. Sedlazeck, Hayan Lee ··· Michael C. Schatz

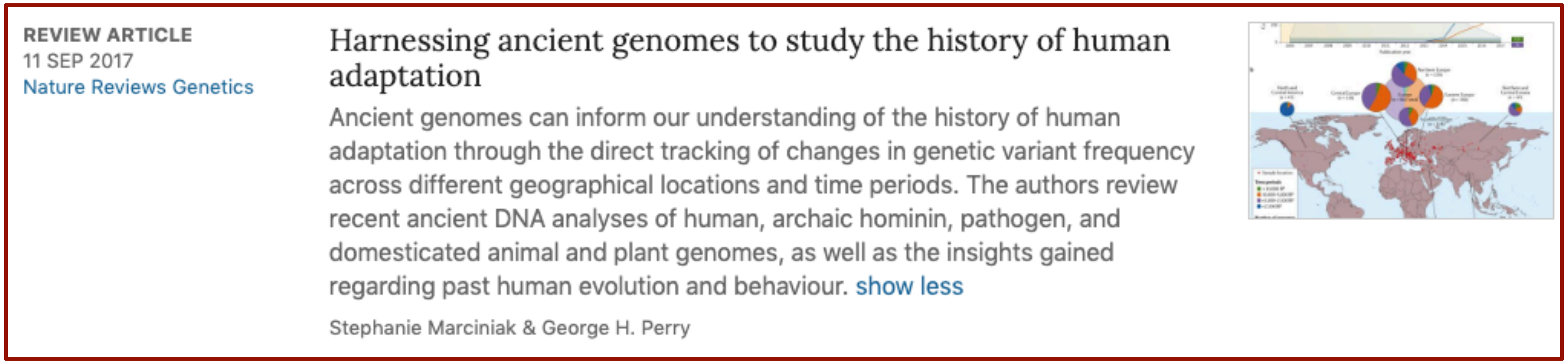


REVIEW ARTICLE
24 OCT 2016
Nature Reviews Genetics

The state of play in higher eukaryote gene annotation

A genome sequence is only useful once the information encoded in it can be deciphered. In this Review, Mudge and Harrow describe the latest approaches to higher eukaryote gene annotation, including making the best use of complex transcriptome data sets, integrating evidence for functionality and extending annotations to encompass regulatory features. show less

Jonathan M. Mudge & Jennifer Harrow



REVIEW ARTICLE
11 SEP 2017
Nature Reviews Genetics

Harnessing ancient genomes to study the history of human adaptation

Ancient genomes can inform our understanding of the history of human adaptation through the direct tracking of changes in genetic variant frequency across different geographical locations and time periods. The authors review recent ancient DNA analyses of human, archaic hominin, pathogen, and domesticated animal and plant genomes, as well as the insights gained regarding past human evolution and behaviour. show less

Stephanie Marciniak & George H. Perry

Applications

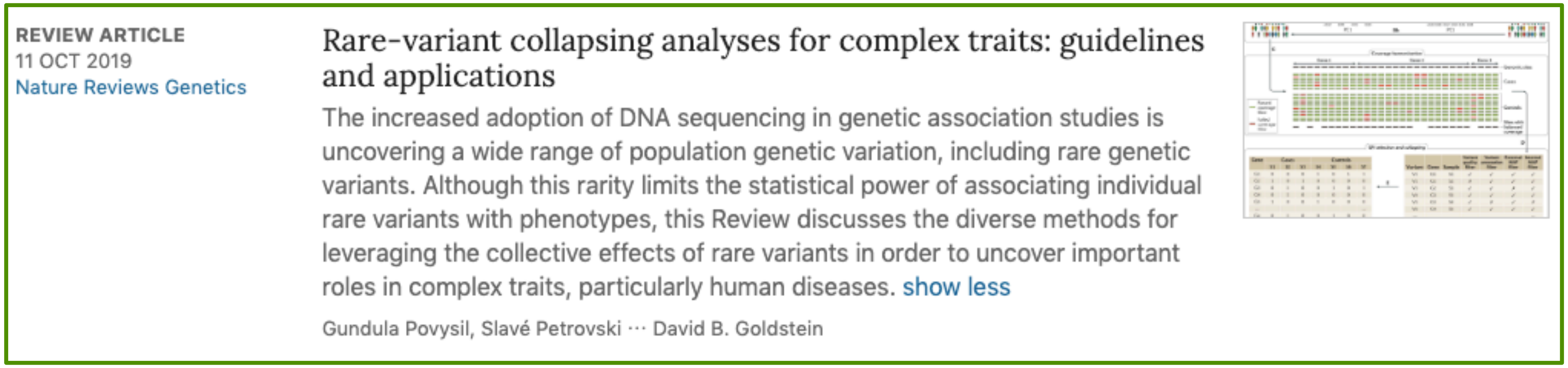
Group 2

REVIEW ARTICLE
13 NOV 2017
Nature Reviews Genetics

Towards a genomics-informed, real-time, global pathogen surveillance system

Next-generation sequencing has the potential to support public health surveillance systems to improve the early detection of emerging infectious diseases. This Review delineates the role of genomics in rapid outbreak response and the challenges that need to be tackled for genomics-informed pathogen surveillance to become a global reality. show less

Jennifer L. Gardy & Nicholas J. Loman



REVIEW ARTICLE
11 OCT 2019
Nature Reviews Genetics

Rare-variant collapsing analyses for complex traits: guidelines and applications

The increased adoption of DNA sequencing in genetic association studies is uncovering a wide range of population genetic variation, including rare genetic variants. Although this rarity limits the statistical power of associating individual rare variants with phenotypes, this Review discusses the diverse methods for leveraging the collective effects of rare variants in order to uncover important roles in complex traits, particularly human diseases. show less

Gundula Povysil, Slavé Petrovski ⋯ David B. Goldstein

REVIEW ARTICLE
19 JUN 2017
Nature Reviews Genetics

Reference standards for next-generation sequencing

Technical errors can hamper the interpretation of next-generation sequencing (NGS) data, which poses a major challenge for the clinical application of this technology. This Review discusses how reference standards circumvent this issue by calibrating NGS measurements and evaluating diagnostic performance of NGS-based genetic tests. show less

Simon A. Hardwick, Ira W. Deveson & Tim R. Mercer