How is DNA Packaged into Chromosomes?

Chromatin - the unit of analysis of the chromosome

- Reflects the general structure of the chromosome
- ***Basic structure shared by all eukaryotic chromosomes***

Important observations from the *Age of Cytogentics*

- Used dyes to stain to develop a ***karyotype*** of the species
- **Chromosome number** and **"organization"**

The first genomics science!!!!

G-banding

- Giemsa stain binds to phosphate groups
  - Differentially stained regions observed in all species

Treat with trypsin; then stain
****Dark bands: A-T rich**
****Light bands: G-C rich**

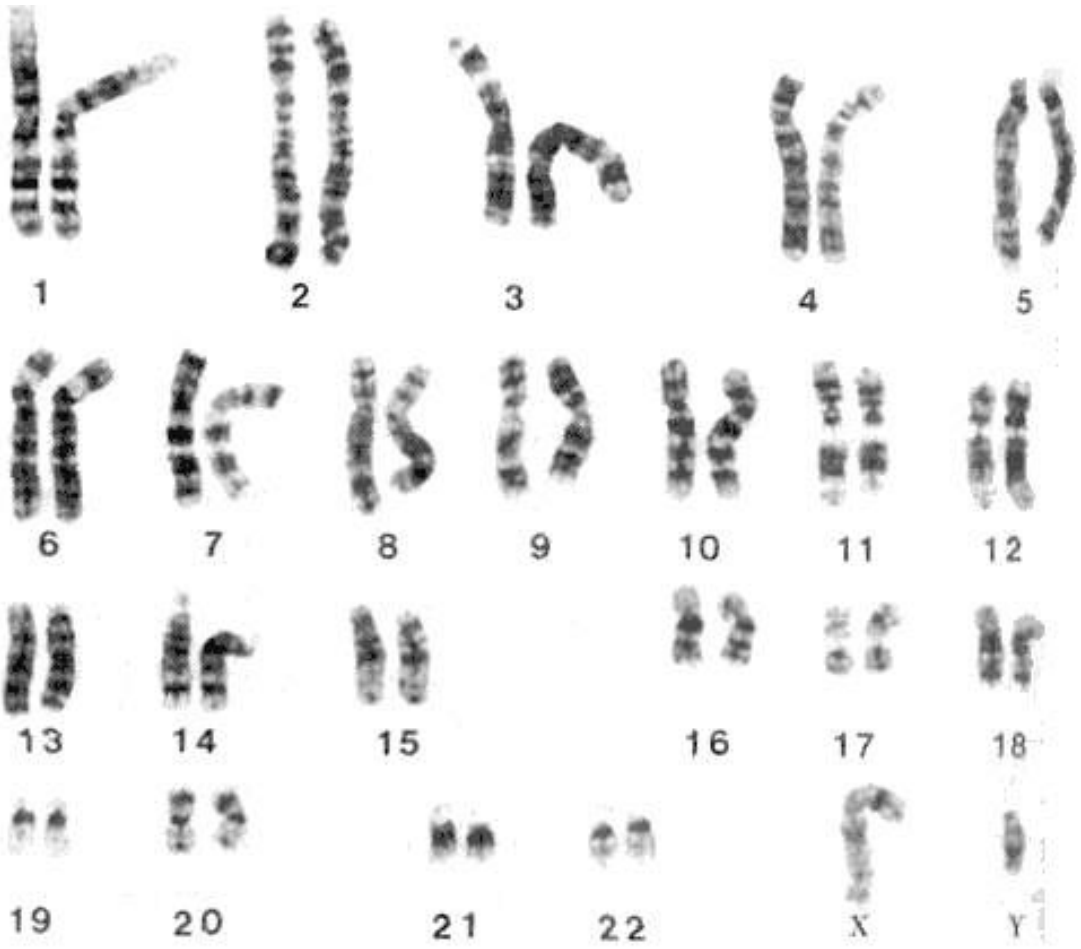
Euchromatin

- **Lightly-stained** regions are euchromatin
- Contain **single-copy genes**
  - Genetically-active DNA (generally)
    - *High rates of recombination*

Heterochromatin

- **Darkly-stained** regions are heterochromatic
- Contain **repetitive sequences**
  - Genetically inactive (generally)
    - *Low rates of recombination*
- Thought to contain genes of high functional importance

Human G-banding



Human G-banding number

- Used to identify cytogenetic location human mutations

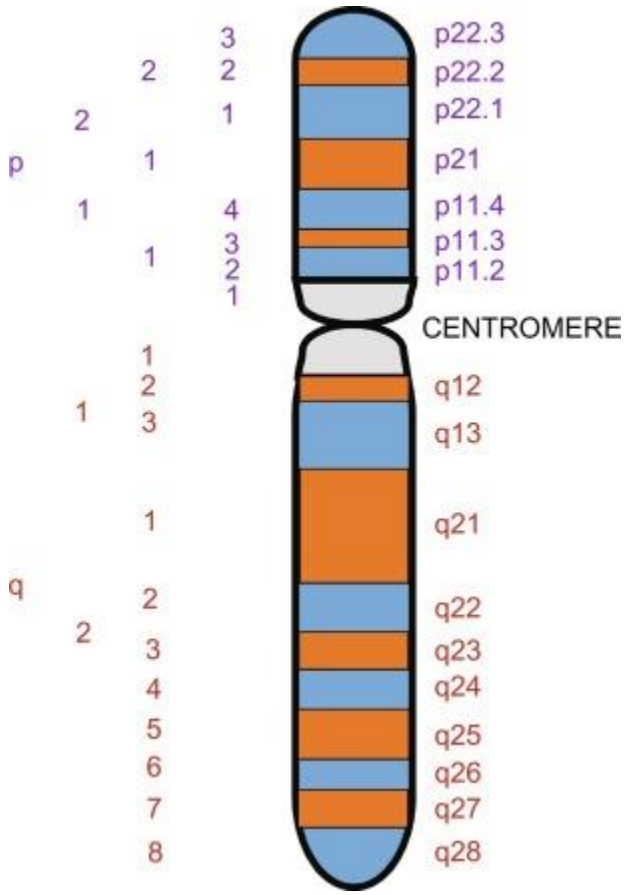


Fig. 13.4. An example of G-banding pattern on X-chromosome. The short arm is p, and the long arm is q. Each arm is divided into larger regions that are further subdivided into smaller bands and interbands. **from:** Chang-Hui Shen (2019) Molecular Diagnosis of Chromosomal Disorders ***in*** Diagnostic Molecular Biology.

Human Chromosome Karyotype Numbering

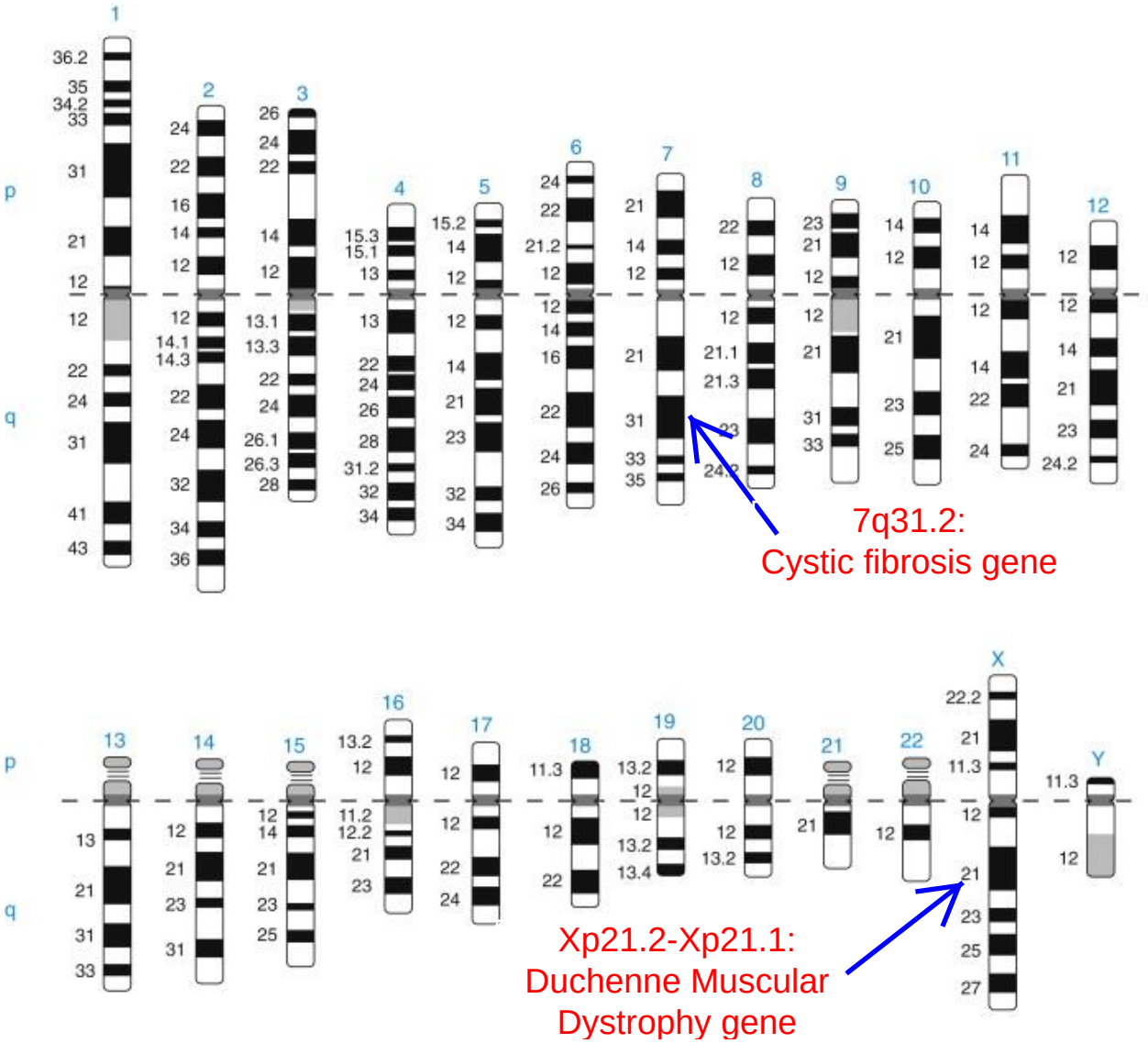


FIGURE 11.2. Ideogram showing G-banding patterns of human metaphase chromosomes. About 400 bands are observed per haploid set. Centromeres are indicated by the dark, gray regions separating the short (p) arms from the long (q) arms. **from**: Leon E. Rosenberg, Diane Drobnis Rosenberg (2012) Chromosome Abnormalities ***in*** Human Genes and Genomes

Location of Human Genes of Interest

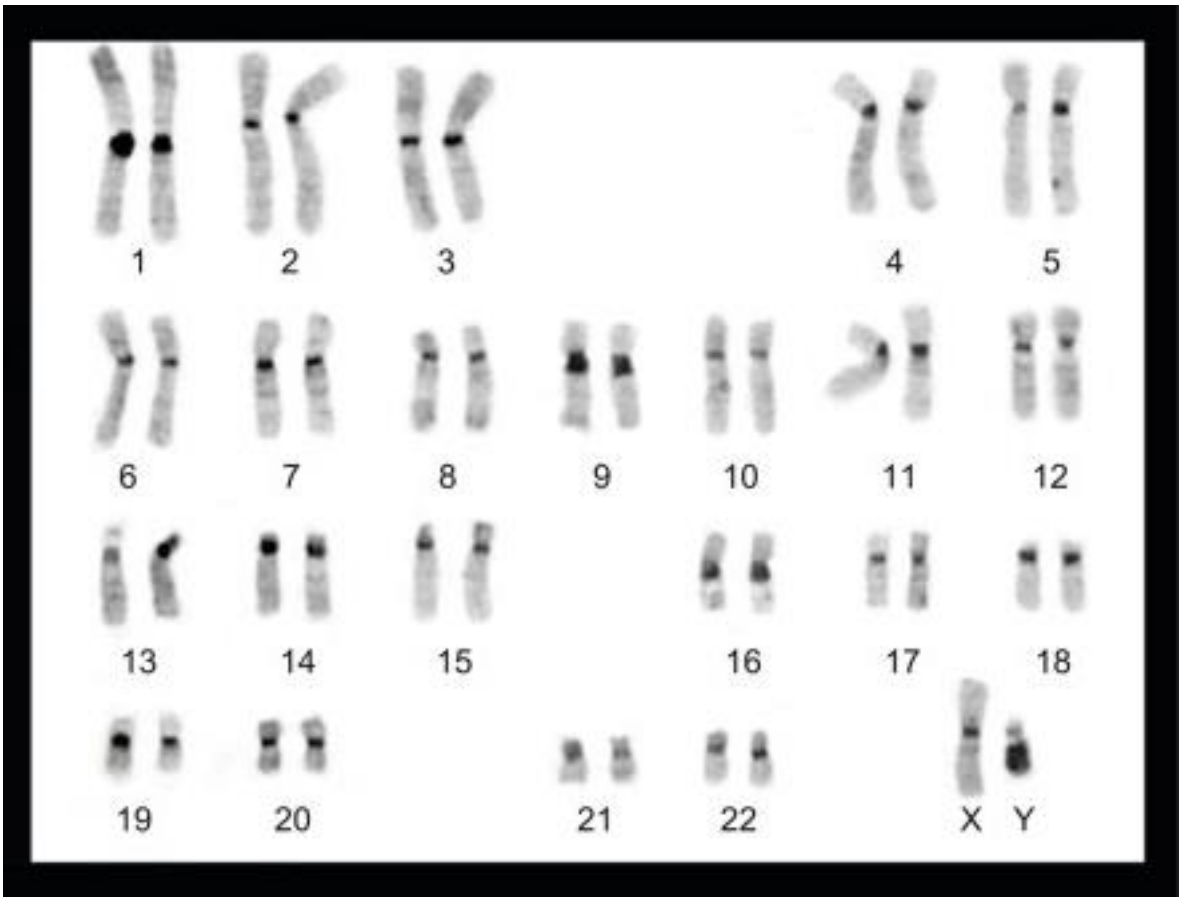
- 7q31.2: Cystic Fibrosis gene
- Xp21.2 – Xp21.1: Duchenne Muscular Dystrophy gene
    - Dystrophin – largest human gene

Alkaline denaturation of chromosomes; then stain

****Specific to heterochromatic regions**

C-banding

- Stains centromeres of chromosomes
    - Differentially stained regions observed in all species



From: Chang-Hui Shen (2019) Molecular Diagnosis of Chromosomal Disorders in Diagnostic Molecular Biology

Wheat, an allohexaploid species.
Three subgenomes, A, B, and D!!!

Hexaploid Wheat C-banding Detects Evolutionary Differences

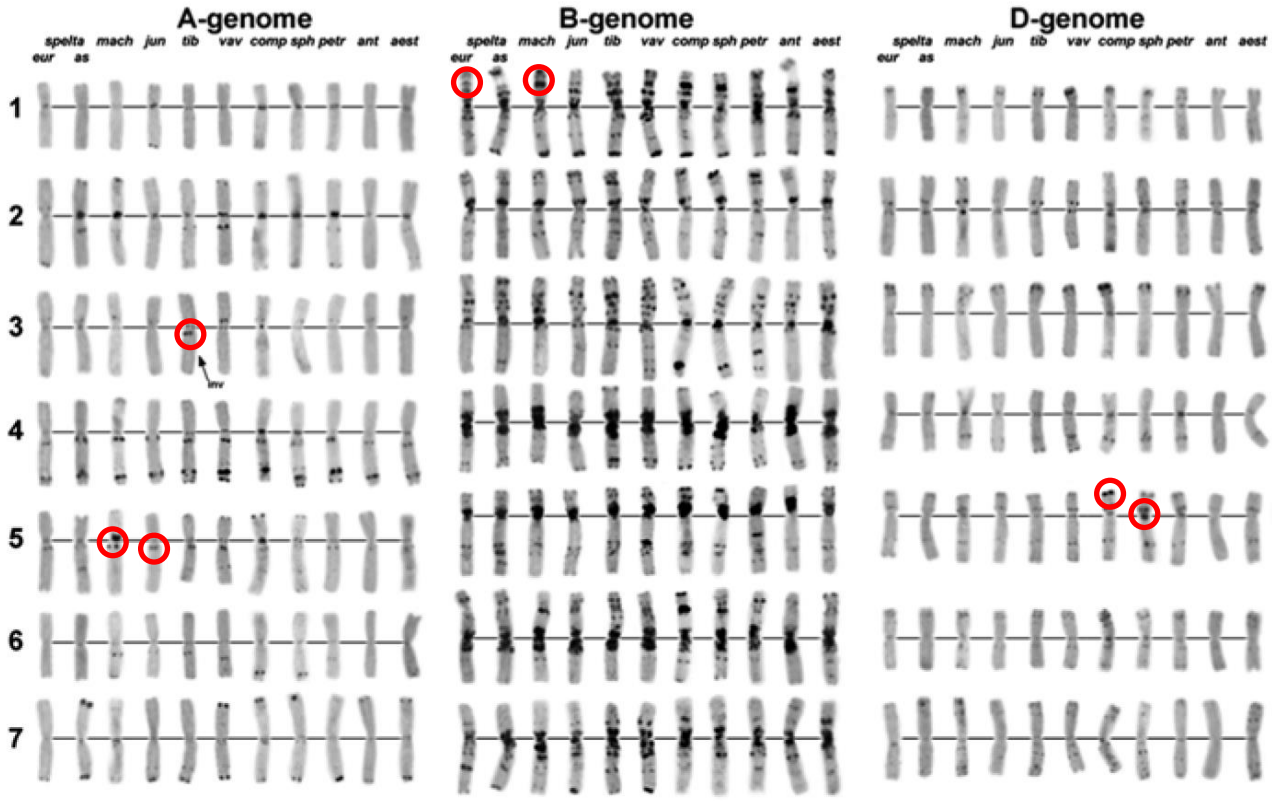


Fig. 9.2 Karyotype diversity in hexaploid wheat species of the Emmer group; splt- T. spelta L., eur – European type, as – Asian type, mach – T. macha Dekapr. et Menabde, jun – T. aesticum ssp. junnanense , tib – T. aestivum ssp. tibetianum Shao, vav – T. vavilovii (Thum.) Jakubz., comp – T. compactum Host, sph – T. sphaerococcum Perciv., petr – T. petropavlovskyi Udacz. et Migush., ant – T. antiquorum Heer ex Udacz, aest – T. aestivum L. em Thell. 1–7 – homoeologous groups. Chromosomal rearrangements are arrowed. **from:** Badaeva et al (2015) Chromosomal Changes over the Course of Polyploid Wheat Evolution and Domestication ***in*** Advances in Wheat Genetics: from Genome to Field.

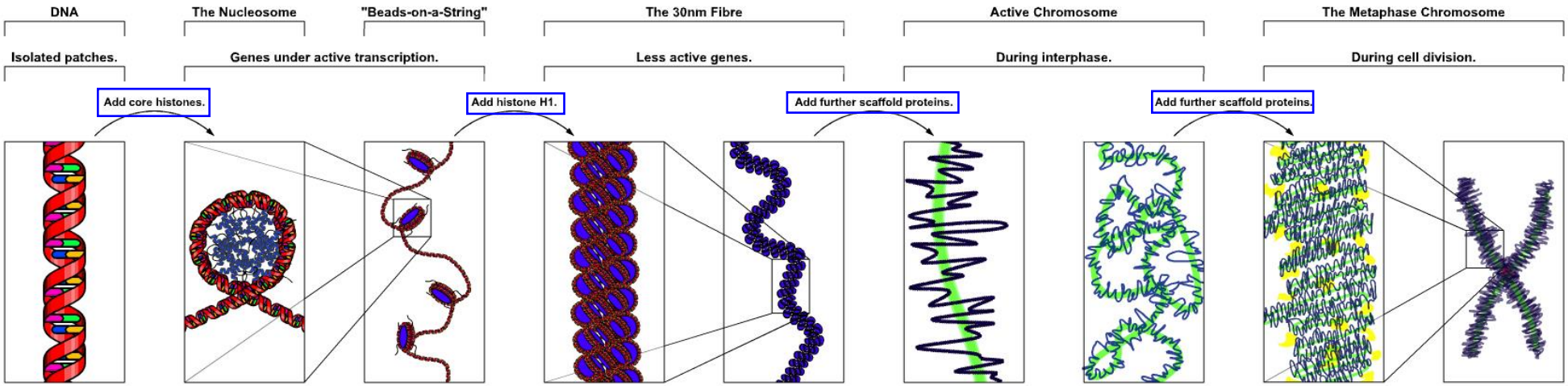
Problem of Great Biological Importance

- Length of DNA is far greater than the size of the nucleus
  - ***How does all of that chromosome DNA packed into the nucleus???***
- DNA has to be condensed in some manner
  - **So, how is the DNA condensed?**

Packing ratio

- Degree to which DNA is condensed
- **Packing ratio** = the length of DNA divided by the length into which it is packaged
- Human Chromosome 22 Example
  - Contains 4.6 x 10^7 bp of DNA
    - About 10 times the genome size of *E. coli*
  - 14,000 µm long when fully extended DNA
    - During mitosis chromosome 22 is 2 µm long
      - **Packing ratio = 7000 (14,000/2).**

Chromatin Packaging Structures



Nucleosome

- Simplest packaging structure of DNA
    - Found in **ALL** eukaryotic chromosomes
        - **Protein/DNA complex**
            - **Packing Ratio = 6**

Histones

- **Protein component of histones**
    - Octamer of proteins
        - **Two copies of four histone proteins**
            - H2A, H2B, H3 and H4
    - **Highly conserved sequences**

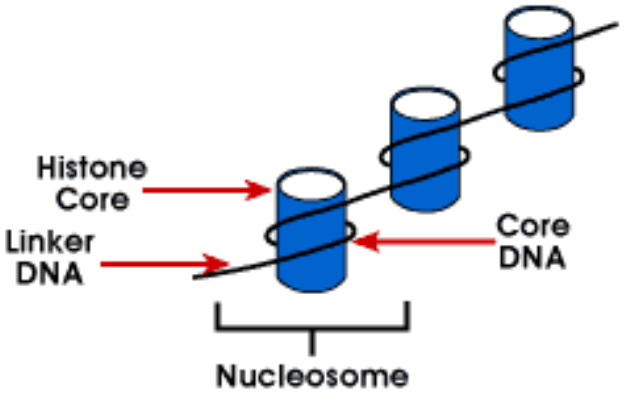
Core DNA

- **146 bp** wrapped around a protein complex
    - **Length is invariant in eukaryotes**
- **Two loops** wrap the histone complex
    - DNA sequences 80 bp apart can be brought into close proximity
- Structure causes **negative supercoiling**

Linker DNA

- **DNA located between histone octamer complex**
    - Variable length from 8 to 114 base pairs
        - Species specific
- Linker DNA length associated with
    - **Developmental stage** of the organism
    - **Specific regions** of the genome

Nucleosome Structure



30 nm fiber

- **Second level** of organization of the chromatin
- Solenoid structure
  - **Six nucleosomes per turn**
- **Stability** of 30 nm fiber requires
  - **Histone H1**
    - Strip H1 from chromatin structure falls apart
- **Packing ratio = 40**

700 nm Structure

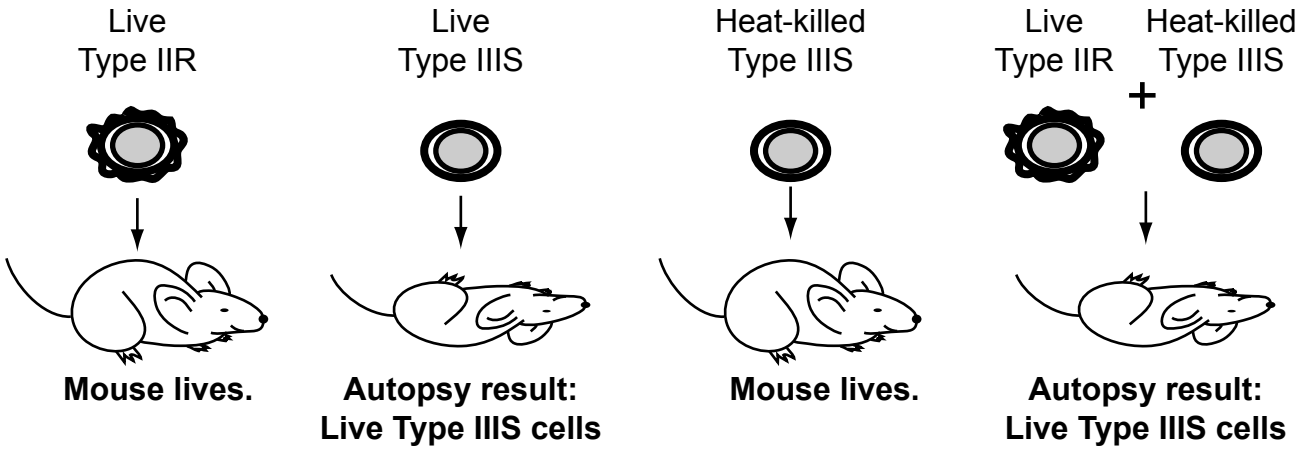
- **Final structural level** of chromatin
- Seen in the metaphase chromosome.
  - Result of **extensive looping** of the DNA in the chromosome
- Final packing ratio
  - **1,000 in interphase chromosomes**
  - **10,000 in mitotic chromosomes**

Griffith and the Transforming Principle

This is the CLASSIC EXPERIMENT that set the stage for the discovery that DNA was the genetic material.

A. The Concept

The experiments of Griffith and Avery, MacLeod and McCarty are closely related. Griffith developed the concept of the ***transforming principle***. The prinicple was able to transform a non-pathogenic bacteria into a pathogenic strain. Changing phenotype is one of the characteristics of the hereditary material. Griffith called the factor that changed the phenotype the tranforming principle. Avery, McCarty, and MacLeod performed a series of experiments that demonstrated ***the hereditary materials was DNA***..



Fred Griffith's experiments provided the experimental platform for Avery, McCarty, and MacLeold to prove the DNA was the genetic material. He worked with the pathogenic bacteria Streptococcus pneumoniae that is lethal to mice. But not all types of the bacteria all lethal: type R is non-lethal, whereas type S is lethal. In addition, there are type II an III strains of the bacteria. Each of these can be either R or S. So a Type IIIS strain is lethal, whereas a type IIR is non-lethal.

Griffith was able to show that if you heat kill a Type IIIS strain and injected it into the mouse, the mouse lived. But if you mixed the heat-killed type IIIS material with live type IIR bacteria, the mouse would die. Furthermore, the autopsy showed that the mouse became infected with the Type IIIS strain. These meant that some material from the Type IIIS strain was taken up by the Type IIR strain to convert it into the Type IIIS strain. Griffith termed the material the ***transforming principle***.

One feature of the genetic material is its ability to control phenotype. In Griffith's experiment, the bacterial strains have several phenotypes. The R types are not only non-lethal, and they have a rough (R) appearance on a blood agar plate. The S type are distinct from the R type: they are lethal and have a smooth morphology on the plates. The S types have a polysaccharide capsule that is lacking in the R types. Each capsule type is distinguished using antibodies; the type II capsule is antigenically distinct from the type III. The trans-formation from type II to type III and the conversion of type R to S are each distinct phenotypic changes. Therefore if the chemical nature of the transforming principle could be determined, then we would know the nature of the genetic material. Avery, MacLeod and McCarty found the answer.

Figure 1. The experiment of Griffith that demonstrated the concept of the transforming principle.

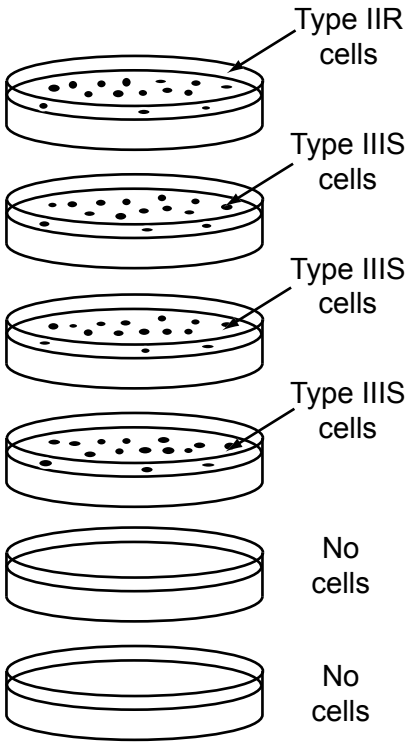
The concept of a *TRANSFORMING PRINCIPLE* is directly related to the term *TRANSGENIC* and the principle of *GENETIC COMPLEMENTATIION*.

Avery, MacLeod and McCarty: DNA Is The Genetic Material

A. The Concept

Avery, MacLeod and McCarty extended the work of Griffith. They used his system, but rather than working with the mice they only studied the bacterial phenotypes relative to the material from the dead type IIIS. They performed careful analysis and proved that DNA, and not protein or RNA, was the genetic material.

Type IIR Cells	Heat-killed IIIS Cells	Type IIIS antibody	Enzyme
+			
+	+	+	
+	+	+	Protease
+	+	+	RNase
+	+	+	DNase
	+		



DNase cuts the DNA; thus DNA is the GENETIC MATERIAL!!!

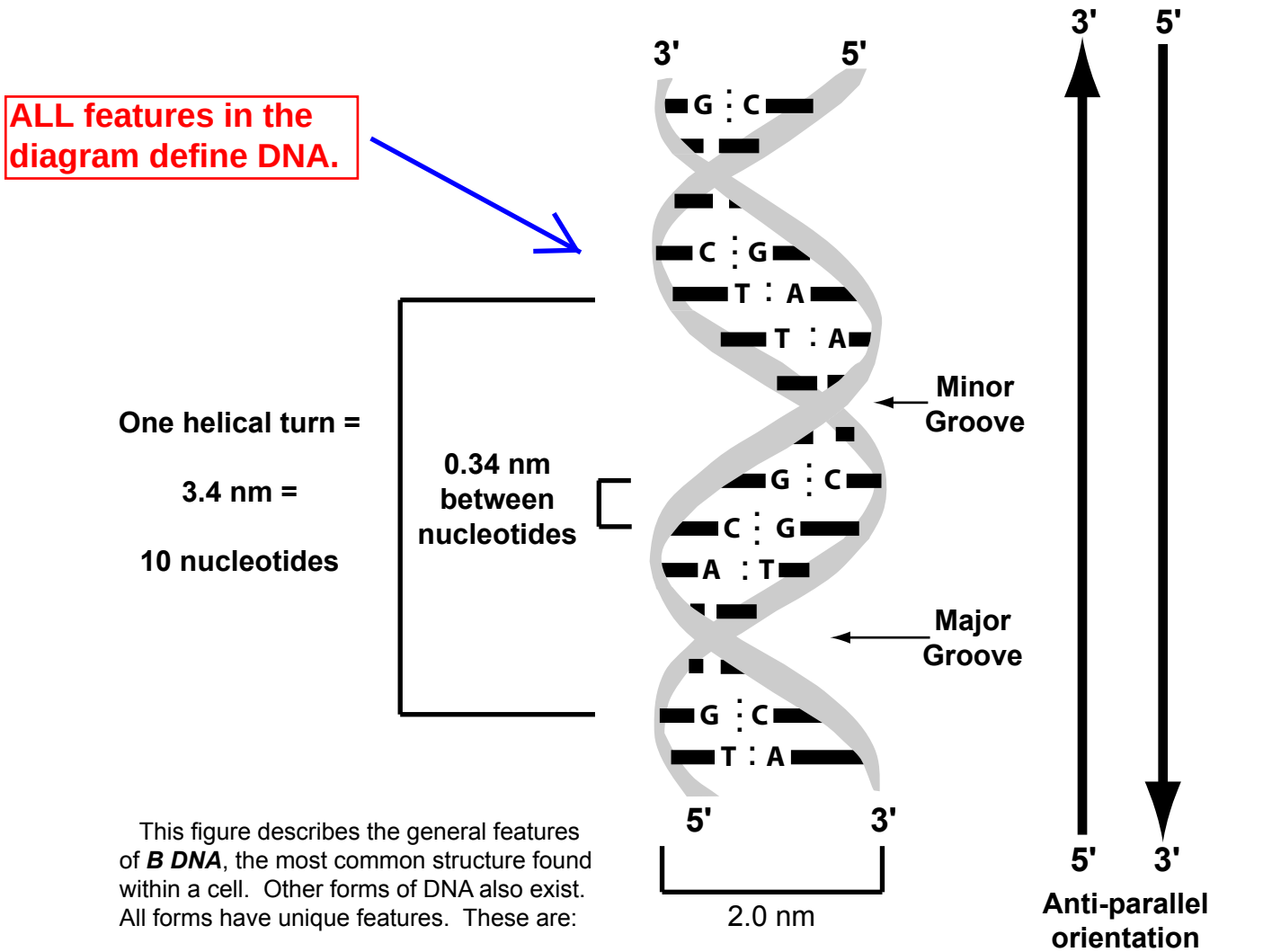
Rather than work with mice, Avery, MacLeod and McCarty used the phenotype of the *Streptococcus pneumoniae* cells expressed on blood agar. To ensure, a few potentially live cells did not escapte the heat treatment, they also precipitated those cells out of culture using an antibody to the type IIR cells. Finally, they included an enzyme treatment of the the material from the heat-killed cells. Each of these enzyme destroyed either proteins (protease), RNA (RNase), or DNA (DNase). These are the three main components of the heat-killed cells. As you can see above, the only treatement that prevented the conversion of the type IIR cells to type IIIS was DNase. This demonstrated conclusively that DNA was the transforming principle and the heredity chemical of life.

Figure 2. The experiment of Avery, MacLeod and McCarty that demonstrated that DNA was the genetic material.

DNA Structure

A. The Concept

DNA has a regular structure. It's orientation, width, width between nucleotides, length and number of nucleotides per helical turn is constant. All of these features were described by Watson and Crick. Adenine is always opposite thymine, and cytosine is always oppostie guanine. The two strands are held together by hydrogen bonds: two bonds between adeninine and thymine and three bonds between guanine and cytosine.



This figure describes the general features of ***B DNA***, the most common structure found within a cell. Other forms of DNA also exist. All forms have unique features. These are:

Form	Helix Direction	Nucleotides per turn	Helix Diameter
A	Right	11	2.3 nm
B	Right	10	2.0 nm
Z	Left	12	1.8 nm

Z-DNA Function
*Transitional state during transcription
**Not found in nucleosomes
*Opens up DNA
**Allows access to transciption factors

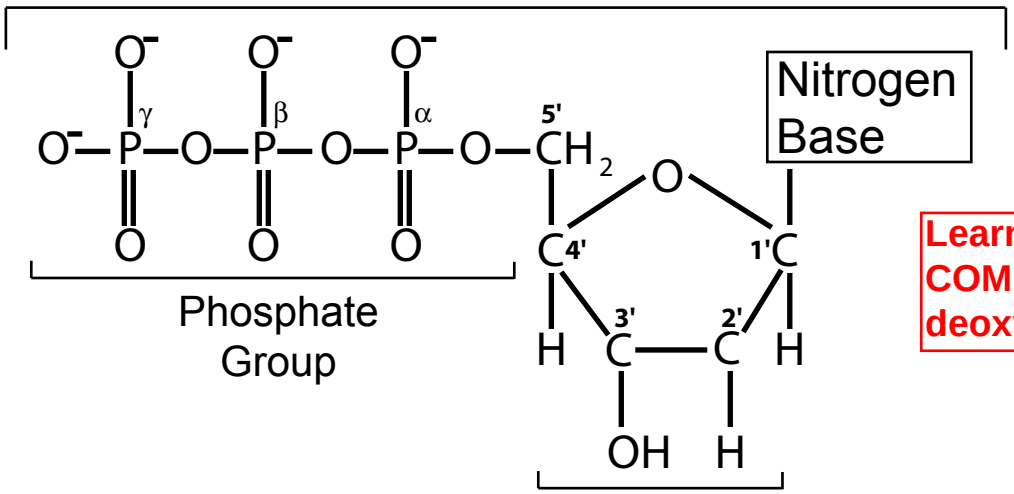
Figure 3. The structure of common DNA molecules.

Deoxyribonucleotide Structure

A. The Concept

DNA is a string of deoxyribonucleotides. These consist of three different components. These are the ***dexoyribose sugar***, a ***phosphate group***, and a ***nitrogen base***. Variation in the nitrogen base composition distingushes each of the four deoxyribonucleotides.

Basic deoxyribonucleotide components



Learn the three COMPONENTS of deoxyribonucleotides.

The basic building block is the ***deoxyribose sugar***. This sugar is distinguished because it contains a hydrogen (H) atom at the number 2' carbon. Normal ribose has a hydorxyl (-OH) group at this position.

Attached to the 5' carbon is a triphosphate group. This group is important because in a DNA chain it undergoes a reaction with the 3' OH group to produce polydeoxynucleotide.

The final feature of the molecule is a ***nitrogen base***. These are attached to the 1' carbon. Four bases are possilbe. Two pyrimidines (thymine and cytosine) and two purines (adenine and guanine). The double stranded DNA molecule is held together by hyrodgen bonds. Pairing involves specific atoms in each base. Adenine pairs with the thymine, and guanine pairs with cytosine. These pairings and the atoms involved are shown to the right.

You have probaly heard of ATP, the energy molecule. It is the deoxyribonucleotide to which adenine is attached. This molecule serves two very important functions in biological organisms.

Learn the NITROGEN BASES and how they PAIR.

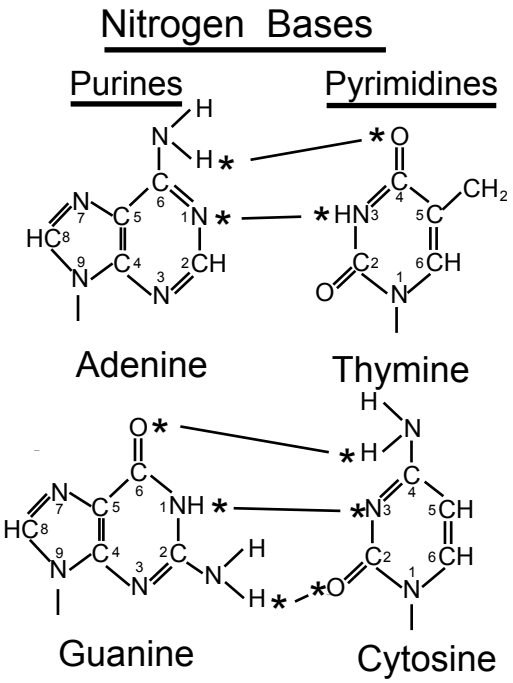
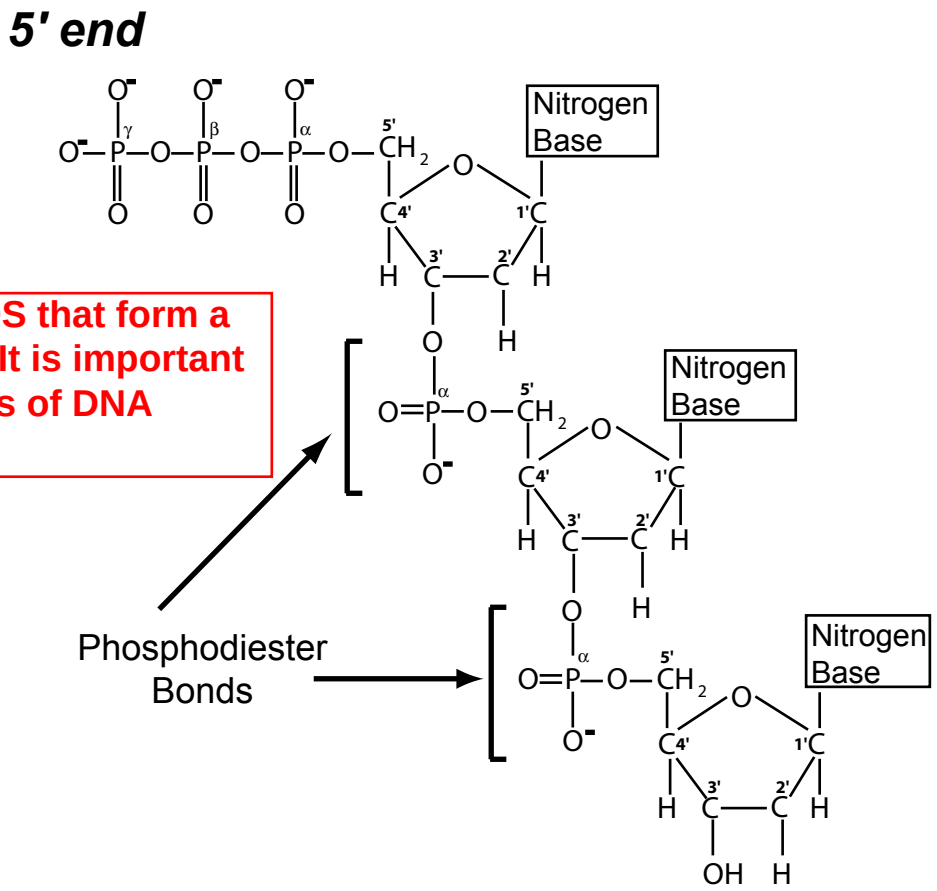


Figure 4. The structure of deoxyribonucleotides and base pairing among N bases.

A Single Strand Molecule of DNA

A. The Concept

Each strand of the double-stranded DNA molecule has the same basic structure. It is a series of [redacted] deoxyribonucleotides linked together by phophodiester bonds.



Know the BONDS that form a DNA molecule. It is important for the principles of DNA sequencing.

DNA is a polynucleotide. It consists of a series of deoxyribonucleotides that are joined by phosphodiester bonds. This bond joins the [redacted] phosphate group to the 3' carbon of the deoxyribose sugar.

Each strand is complementary to the opposite strand. If one strand has an adenine at a position, its anti-parallel[redacted] strand would have a thymine at the the corresponding position. Likewise, guanine and cytosine would be complementary.

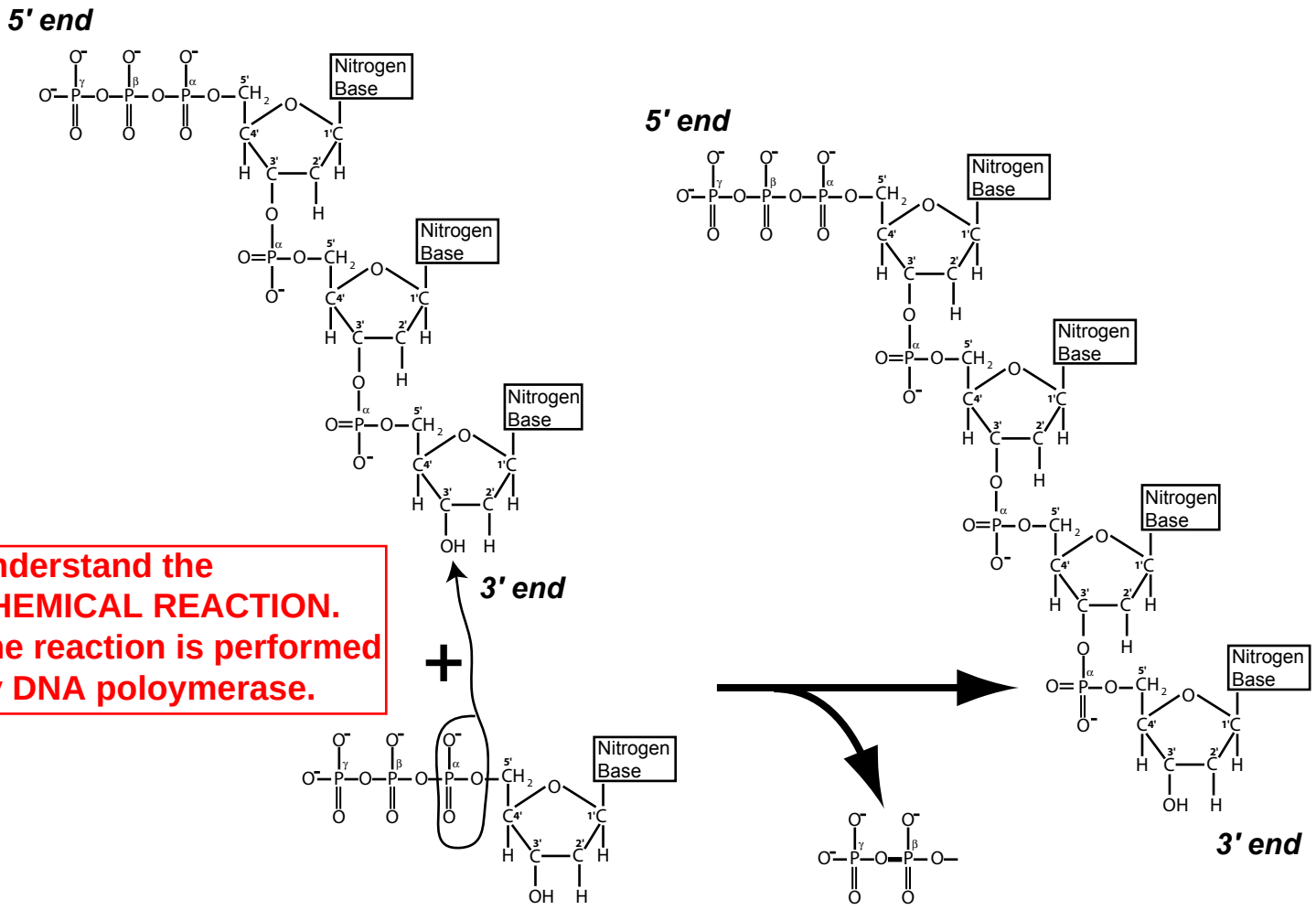
Fig. 5. The single strand structure of DNA.

Why study DNA chain elongation???
DNA sequencing is a technological adaptation of DNA elongation!!!

Making a Phosphodiester Bond/ Growing the DNA Chain

A. The Concept

The addition of a new nucleotide to a DNA molcule creates a phosphodiester bond. This requires the DNA chain that is being elongated and a deoxyribo-nucleotide.



Understand the CHEMICAL REACTION. The reaction is performed by DNA poloymerase.

(Pyrophosphate) and H+ molecule

Phosphodiester bonds are formed when a new dideoxynucleotide is added to a growing DNA molecule. During the reaction, a condensation reaction occurs between the α phosphate of the nucleotide and the hyroxyl group attached to the 3' carbon. This reaction is performed by the enzyme DNA polymerase. This is also an energy requiring reaction. The energy is provided by the breaking of the high-energy phophate bond in the nucleotide. This results in the release of a pyrophosphate molecule.

Figure 6. The formation of the phosphodiester bond that grows the DNA chain.

H+ is a bi-product of the reaction. The H+ generation is monitoried in the ION TORRENT sequencing technology.

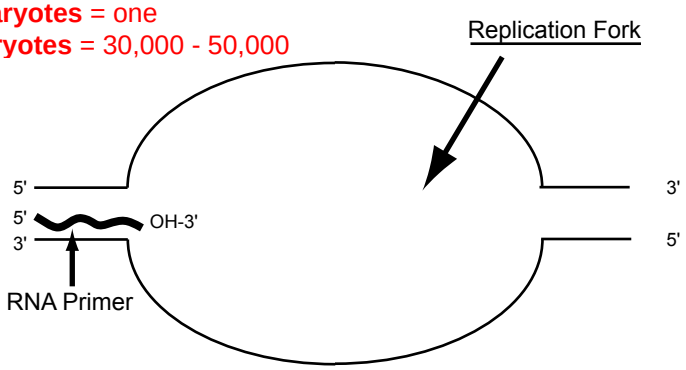
Steps of DNA Replication (Part 1)

A. The Concept

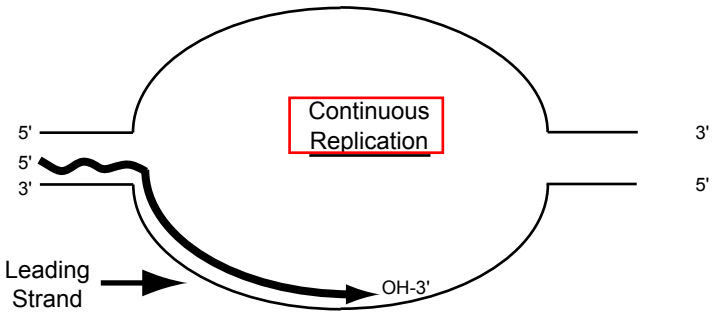
DNA replication is an essential biological process. It's primary function is to produce new DNA for cell division. The process has several distinct steps that are important to understand. The factors that are absolute requirements for DNA replication to begin are a ***free 3'-OH group*** and a ***DNA template***. A RNA primer provides the free 3'-OH group. The DNA to be replicated serves as the template. It is important to remember that ***all*** DNA replication proceeds in the 5'-3' direction.

1. The replication fork is formed; RNA primer added.

Prokaryotes = one
Eukaryotes = 30,000 - 50,000

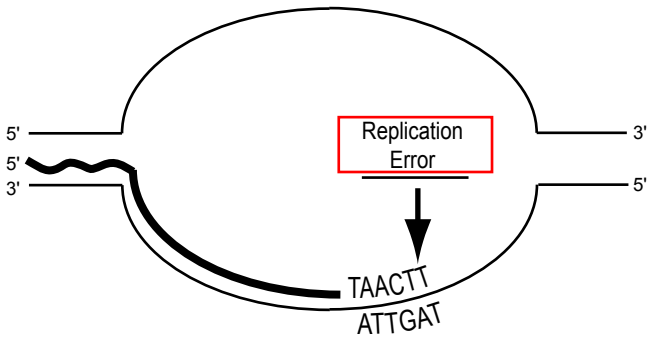


2. DNA is replicated by the 5'-3' synthesis function of DNA polymerase using the leading strand in a continuous manner.

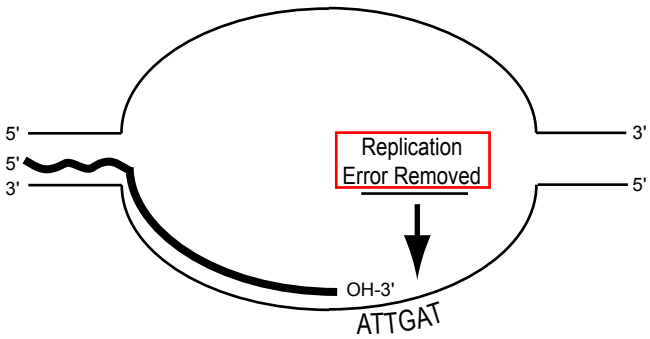


DNA Pol III

3. An error occurs during DNA replication.



4. The DNA replication error is removed by 3'-5' exonuclease function of DNA polymerase.



DNA Pol III

Notes on *E. coli* replication:

DNA Polymerase I and III. Pol III is the primary replicase enzyme that performs the elongation of the DNA strand. It adds nucleotides first to the RNA primer and then grows the chain by creating the phosphodiester bonds. It also has a 3'-5' proofreading (exonulcease) function that removes incorreclty incorporated nucleotides. DNA Pol I also has the 5'-3' replicase function, but it is primarily used to fill the gaps in the replicated DNA that occur when the RNA primer is removed. This enzyme also has a 5'-3' exonuclease function that is used to remove the RNA primer.

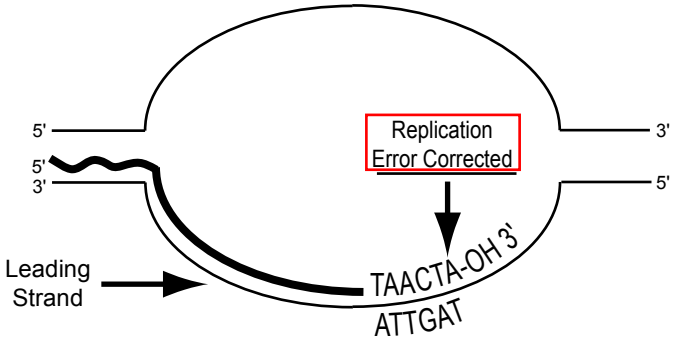
Figure 7. The steps of DNA replication.

DNA Polymerase III Function
1. 5'-3' elongation
2. 3'-5' proof reading

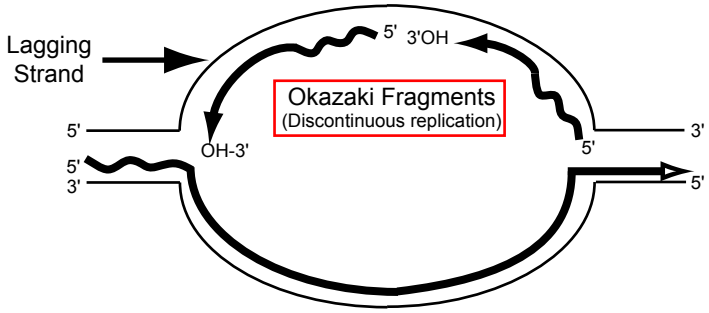
DNA Polymerase I Function
1. 5'-3' exonuclease remove RNA primer
2. 5'-3' replicase to fill in gap

Steps of DNA Replication
(Part 2)

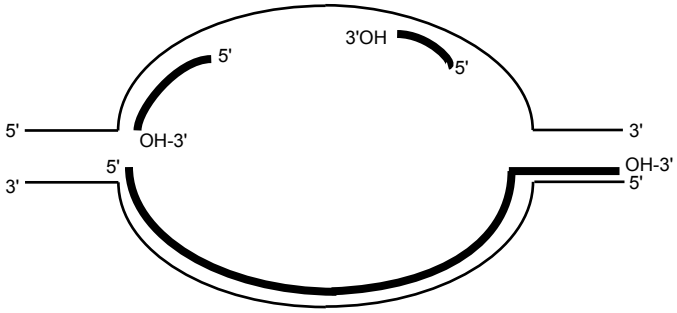
5. The DNA replication error is corrected.



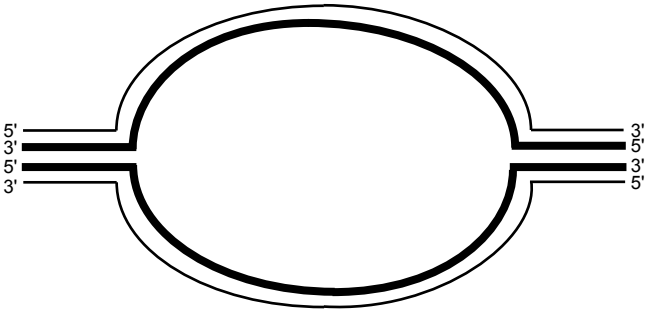
6. Meanwhile, Okazaki fragments are synthesized using the lagging strand in a discontinuous manner while the leadng strand is completed simultaneously.



7. The RNA primers are removed by 5'-3 exonuclease function of DNA polymerase.



8.. Replication is completed by the filling in the gaps by DNA polymerae and DNA ligase.



Notes on replication:

Okazaki fragments: Both prokaryotic and eukaryotic DNA replication proceed in the 5'-3' direction. This poses a problem because the replication fork only moves in that one direction. The problem relates to what is called the ***lagging strand***. It must be replicated in a direction that is opposite of the direction of the replication fork. This problem was solved by the discovery of Okazaki fragments (named after the person who discovered the process). In contrast to the ***leading strand***, in which DNA is replicated as a single molecule in a ***continuous*** manner, DNA is replicated in a ***disocontinuous*** manner on the lagging strand. Each fragment requires a RNA primer, and DNA Pol III in *E. coli* makes short stretches of DNA. These fragments are then stitched together when the primer is removed, and the strands are completed by the action of DNA Pol I and ligase.

Figure 7 (cont.). The steps of DNA replication.

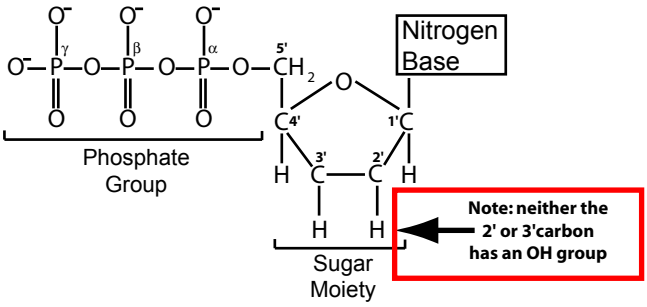
Other Enzymes
1. DNA Polymerase I
2. Ligase

Chain Termination Sequencing: the Sanger Technique

A. The Concept

DNA sequencing is the most important technique of genomics. By collecting the sequence of genes and genomes we begin to understand the raw material of phenotype development. The most common DNA sequencing technique is called ***chain termination sequencing*** or the ***Sanger technique*** (named after the person who created it). It is called chain termination because the incorporation of a ***dideoxynucleotide*** terminates the replication process because this nucleotide lacks the required 3'-OH group.

a. A dideoxynucleotide



b. The reaction reagents

DNA template
sequencing primer
dNTPs
ddNTPs (low concentration)
DNA polymerase
salts

c. The sequencing reaction result: fragments that differ by one nucleotide in length

Template → A T T C G G A T C C T T A A

Primer →

5' T A A G C C T A G G A A T T - H 3'
5' T A A G C C T A G G A A T - H 3'
5' T A A G C C T A G G A A - H 3'
5' T A A G C C T A G G A - H 3'
5' T A A G C C T A G G - H 3'
5' T A A G C C T A G - H 3'
5' T A A G C C T A - H 3'
5' T A A G C C T - H 3'
5' T A A G C C - H 3'
5' T A A G C - H 3'
5' T A A G - H 3'
5' T A A - H 3'
5' T A - H 3'
5' T - H 3'

The SANGER TECHNIQUE is STILL used extensively today!!! Primarily used to sequence targeted genes or regions of DNA!!!

◯ = Terminal dideoxynucleotides; extension stops

When a dideoxynucleotide is inserted, the DNA replication process terminates because dideoxynucleotides do not have the necessary free 3' hydroxyl group required for the addition of additional nucleotides. This results in fragments that differ by one nucleotide in length.

Figure 8. The chain termination (Sanger) DNA sequencing technique.

Gel-based Detection of DNA Sequences

A. The concept

Four DNA sequencing reactions are performed. Each contains only one of the four dideoxynucleotides. Each reaction is added to a single lane on the gel. Since one of the dNTPs is radioactive, the gel in which the fragments are separated, can be used to expose an x-ray film and read the sequence.

a. The sequencing products

Four separate sequencing reactions performed;

Different ddNTPs used

Reaction with ddATP

A T T C G G A T C C T T A A
5' T A A G C C T A G G A A -H 3'
5' T A A G C C T A G G A -H 3'
5' T A A G C C T A -H 3'
5' T A A -H 3'
5' T A -H 3'

Reaction with ddTTP

A T T C G G A T C C T T A A
5' T A A G C C T A G G A A T T -H 3'
5' T A A G C C T A G G A A T -H 3'
5' T A A G C C T -H 3'
5' T -H 3'

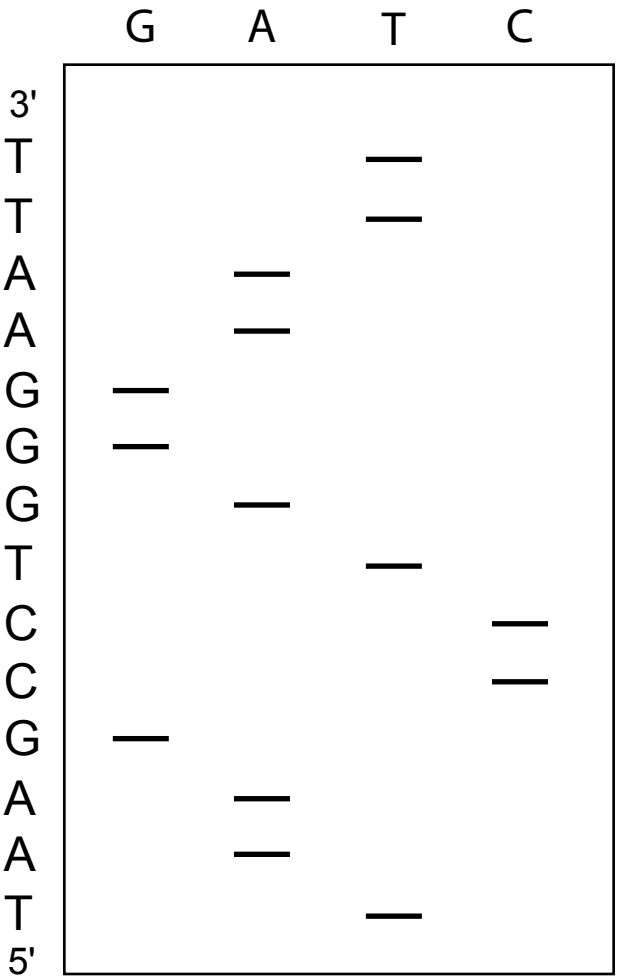
Reaction with ddGTP

A T T C G G A T C C T T A A
5' T A A G C C T A G G -H 3'
5' T A A G C C T A G -H 3'
5' T A A G -H 3'

Reaction with ddCTP

A T T C G G A T C C T T A A
5' T A A G C C -H 3'
5' T A A G C -H 3'

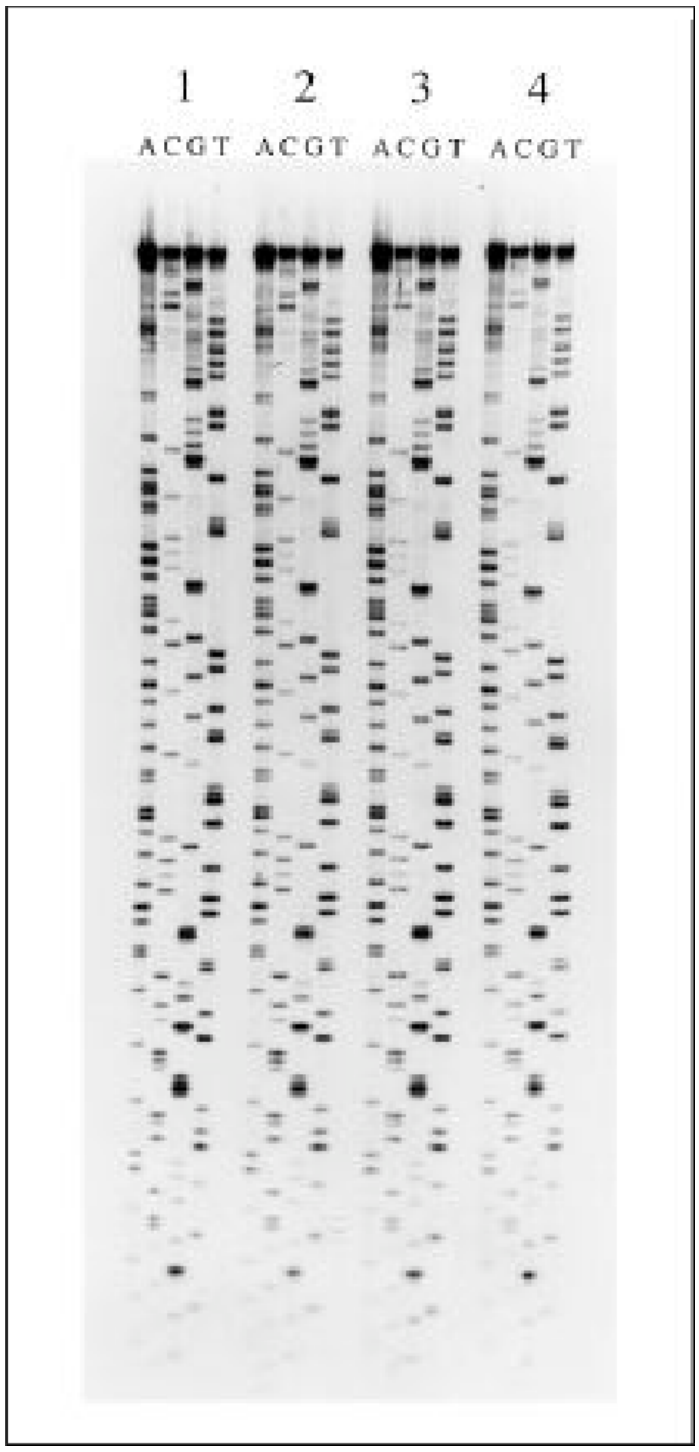
b. The sequencing gel



The sequencing reactions are separated on a polyacrylamide gel. This gel separates the fragments based on size. The shorter fragments run further, the longer fragments run a shorter distance. This allows the scientists to read the sequence in the 5'-3' direction going from the bottom to the top of the gel.

Figure 9. Gel-based detection of DNA sequencing products.

DNA Autoradiogram



Fluorescent Sequencing and Laser Detection

A. The Concept

Rather than using four different reactions, each with a single dideoxynucleotide, the advent of fluorescently labeled dideoxynucleotide enabled 1) the sequencing reaction to be performed in a single tube, and 2) the fragment could be detected by laser technology. Originally, the products were separated in a polyacrylamide gel prior to laser detection. The introduction of capillary electrophoresis, coupled with laser detection enabled the detection of up to 96 products at a time.

B. The Reaction Products and Analysis

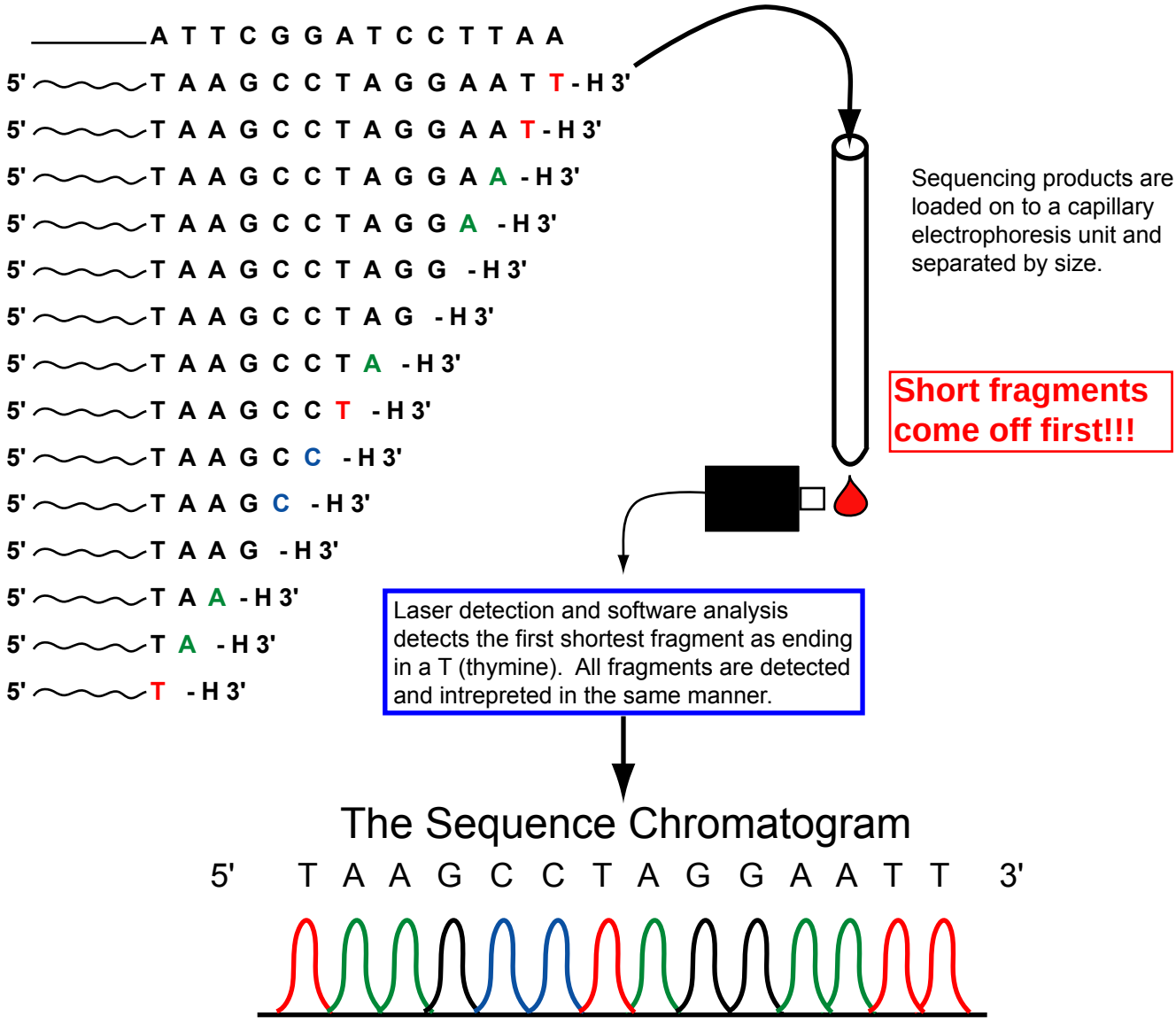
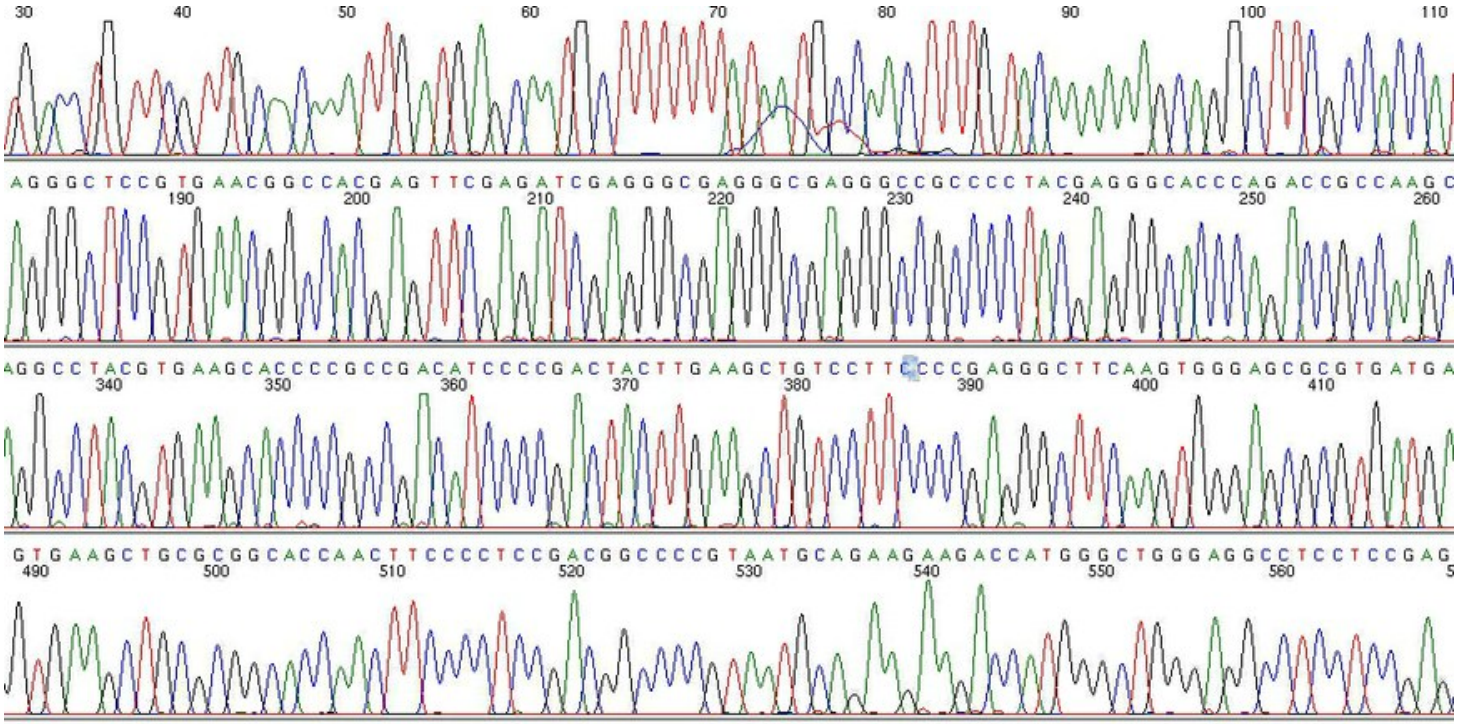


Figure 10. The fluorescent sequencing and laser detectiion process of DNA sequencing.

Output from Automated DNA Sequencer

Sanger sequencing throughput
Originally
****96 samples per ~two hours**
Then
****384 samples per ~two hours**
Fragment Length
****500-750 nt**



NO ONE ever uses this type of chromatogram. All of the ANALYSIS is performed using SOFTWARE PACKAGES.

MASSIVELY PARALLEL SEQUENCING
The underlying principle of all modern genome sequencing projects.

What Was Needed for All New Approaches

Reducing Cost

- How: Parallel sequencing
  - Large number of sequencing reactions occurring simultaneously
    - Requires high density reactions matrix
      - Many reactions in a small space
      - Miniaturization of reaction unit or space
  - Reduce reagent cost
    - Accomplished when above factors achieved

Throughput

- Many reactions occurring simultaneously
  - Current Sanger macrocapillary system
    - 96-384 reactions per machine
    - Sequencing centers have 30-60 machines
    - ***New approaches must have significantly greater throughput***

Sequence Accuracy Must Be Maintained

- Sanger procedure highly accurate
  - Well understood Phred scores reported
    - ***New systems will require quantifiable accuracy scores***

Quality score
** >Q20 requirement

Completeness

- Read length issue
  - Sanger technology with capillary detection
  - 500-700 nt
    - Allows for assembly into
      - Contigs
      - Supercontigs ← Now called **Scaffolds**
- Emerging technologies
  - Length requirement
  - Must be long enough to align accurately
    - 25-100 nt read length

Original goal

- Sufficient for resequencing with a reference genome

- Whole genome sequencing
  - 100 nt (or longer) needed for smaller genomes
  - Other advances needed for larger genomes

Today the principle read lengths are:
***Illumina = 150 bp (standard)**
***PacBio = 20 kb (CCS or HiFi)**
= 60-80 kb (CLR)
***Oxford Nanopore Technology (ONT)**
= up to Megabase

Notice the EVOLUTION of sequencing by the NUMBER OF MACHINES in the sequencing facility.

How Large Scale Sequencing Has Changed Over Time From a Centers Perspective

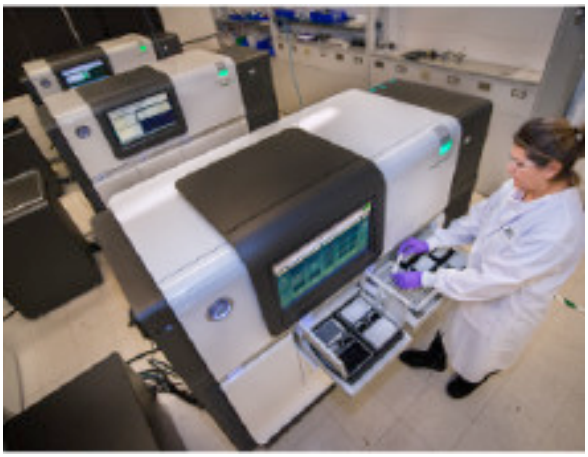
Then: DOE/JGI Sanger Sequencing Equipment Room

Two rooms
**32 96 sample machines
**32 384 sample machines

Recently: DOE/JGI Illumina GAII Equipment Room

Now: DOE/JGI Illumina HiSEQ Equipment Room

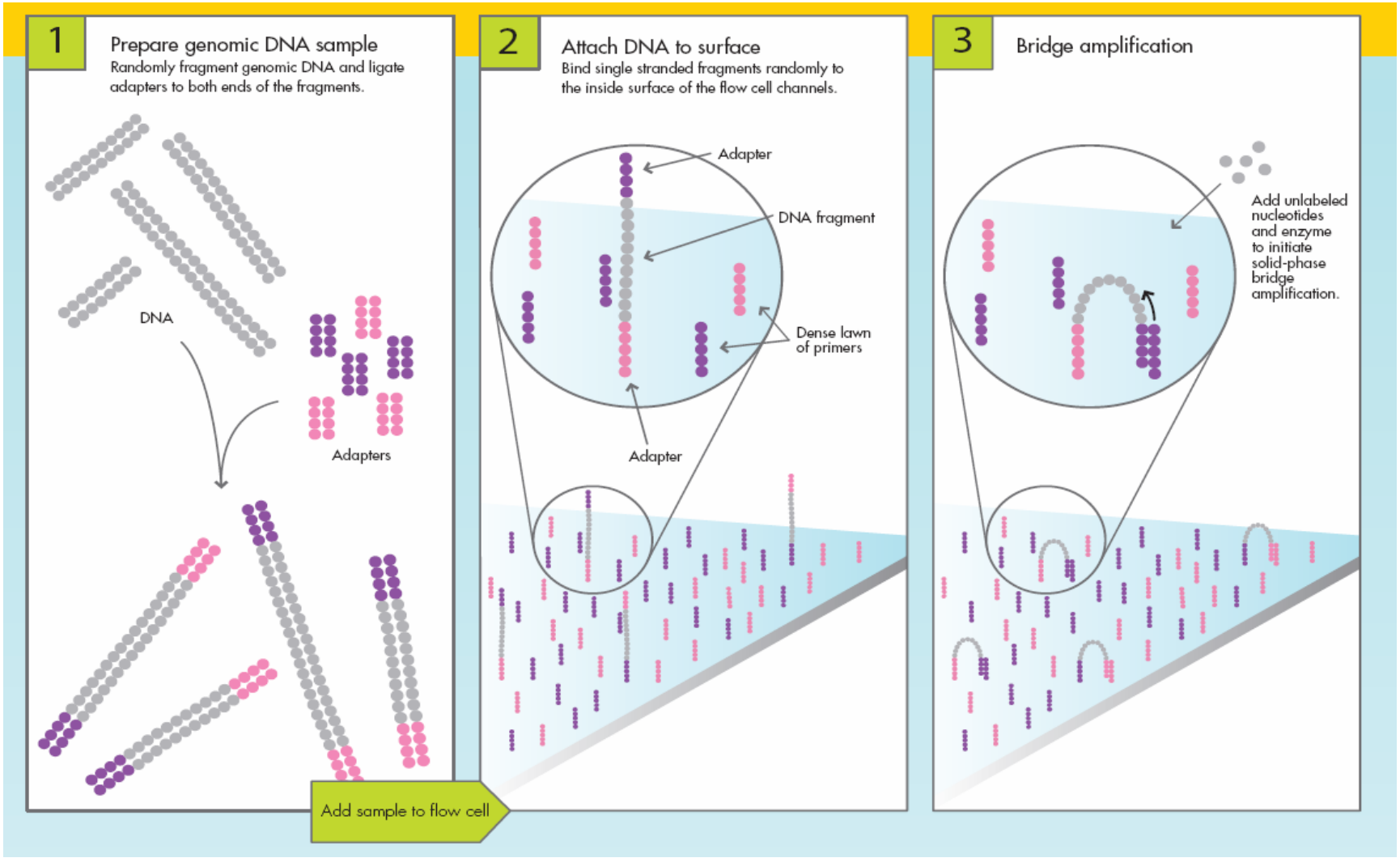
Now: DOE/JGI PacBIO Equipment Room



TODAY, Illumina is the MARKET LEADER in high throughput sequencing.

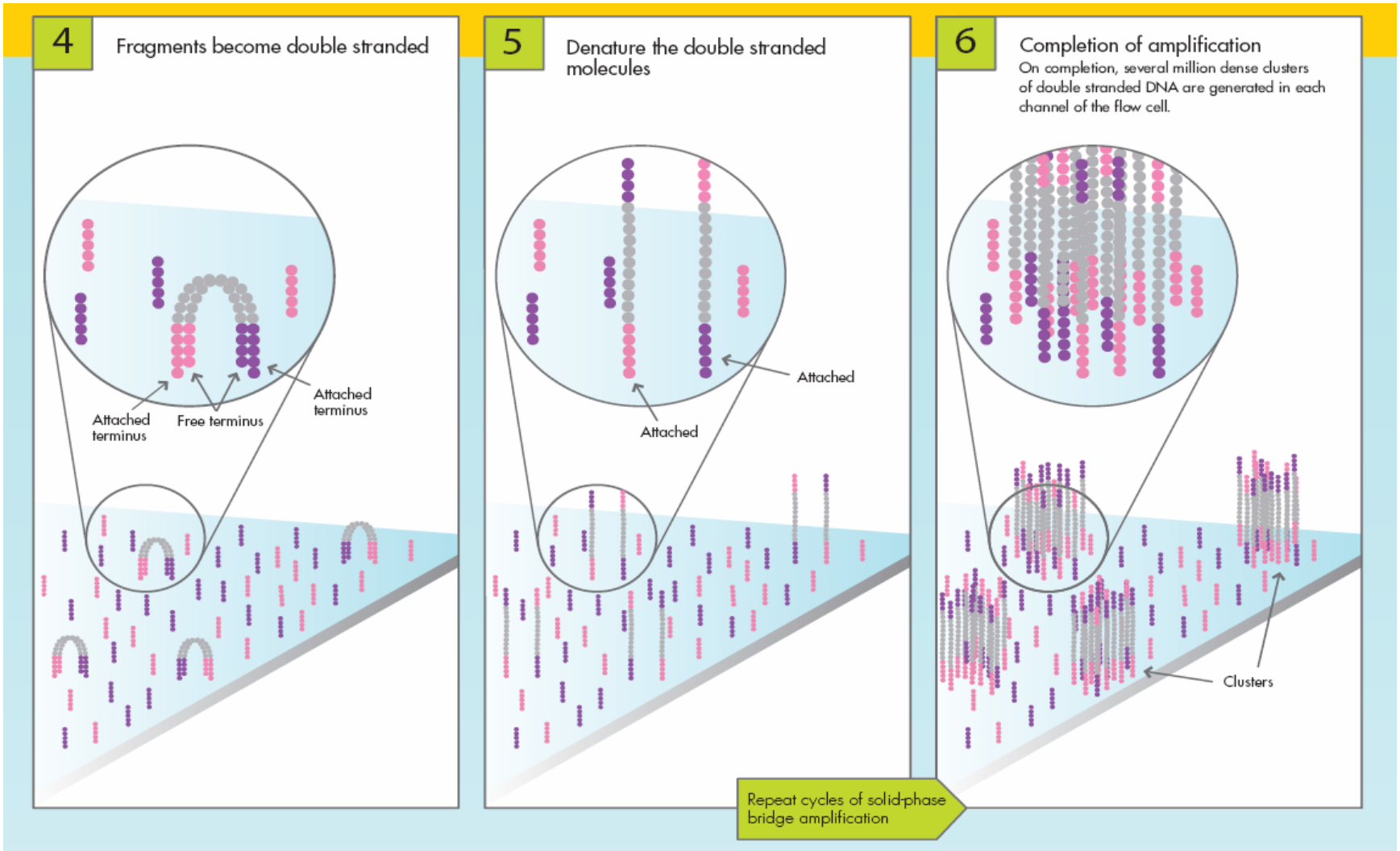
Illumina Sequencing by Synthesis Technology

Founded 1998

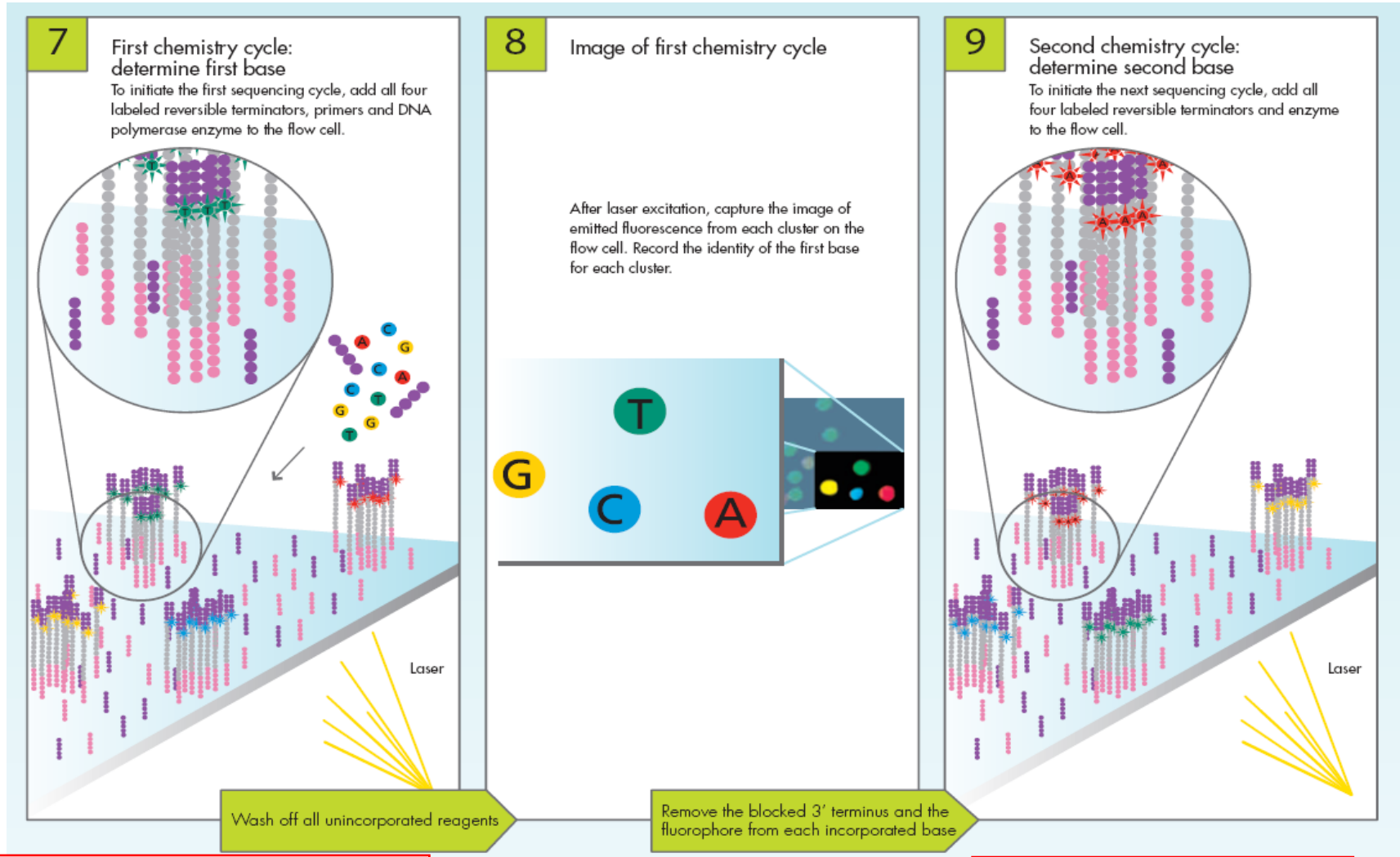


A SINGLE STRAND molecule is bound to the flow cell.

BRIDGE AMPLIFICATION:
***Steps 3-6**



EVERY MOLECULE IN THE CLUSTER IS AN IDENTICAL TEMPLATE FOR SEQUENCING!!



The first base is added to the template of each cluster with a blocker that prevents other bases from being added.

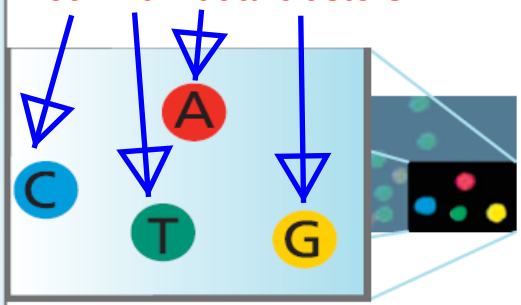
A PICTURE is taken of the flow cell; the color emitted determines the base added to the cluster. The blocker is removed.

The LAST TWO steps are repeated until the desired read length is reached.

10 Image of second chemistry cycle is captured by the instrument

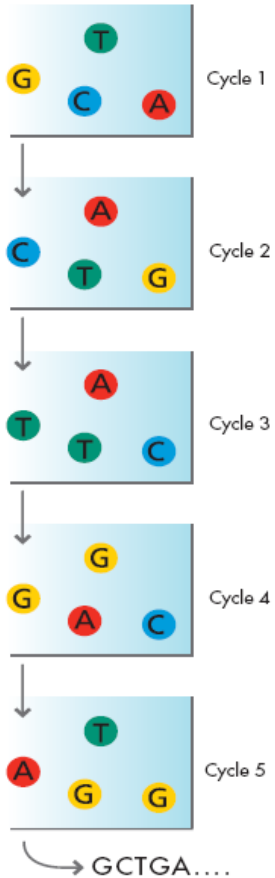
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

Four individual clusters!!

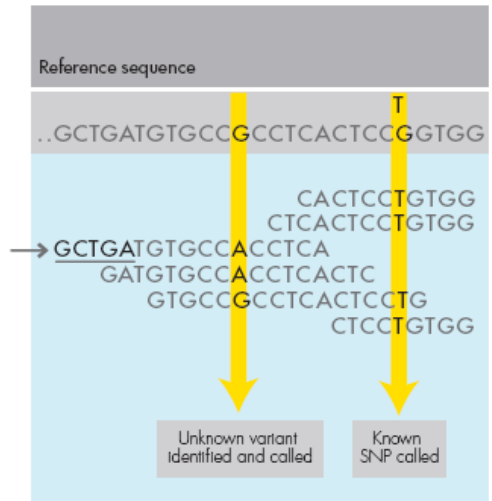


11 Sequence read over multiple chemistry cycles

Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.



12 Align the new data to a reference and identify sequence differences



Notice the change in data OUTPUT over the different generations of the machines.

Illumina Sequencers Over Time: Today's Workhorse

**

Illumina GAII [Maximum (Max) output: 25 gigabases (Gb)]

Bean Genome
~550 Mb (=0.55 Gb)
~45 Bean genomes of data collected

**

Illumina HiSeq 2500 (Max output; 500 Gb; Rapid Run Mode: 150 Gb)

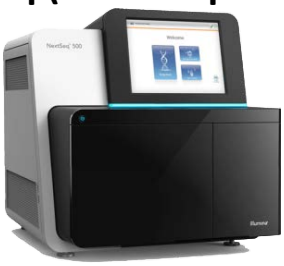
~900 Bean genomes of data collected

Today's Illumina Models

(Mostly chemistry and reads per flow cell differences)

**

Illumina NextSeq (Max output: 120 gigabases)



**

Illumina HiSeq X10 (Max output: 1.8 Tb) GENOMES ONLY

**

Illumina HiSEQ 4000 (Max output: 1.5 Tb) Most other sequencing

**

Illumina NovaSeq (Next Generation; 2017 release; Max output: 1.5 Tb)

The NovaSeq 6000 is the principle machine used TODAY for most HIGH THROUGHPUT Illumina sequencing in sequencing centers!!!

PACBIO: Single molecule sequencing
*****SECOND TOOL in modern genome sequencing**

Single Polymerase Real Time DNA Sequencing

Developed by Pacific Biosciences **Founded 2004**
Sequences occurs at the rate of ***10 nt per second***

Principle

Read the details here on your own after going over the images and watching the lecture. This is more for your in-depth knowledge than exam material.

Reaction Cell

- A single DNA polymerase is immobilized on the bottom of a reaction cell
  - Reaction cell called a ZMW (Zero-mode waveguide)
- Φ29 DNA polymerase is used
  - Fast single subunit enzyme. 100 bp/sec
- Each sequencing plate contains ~3000 individual cells
  - Each holds only a single DNA molecule

TODAY
****25 million cells**
****In practice not all used**

Chemistry

- A phospholinked dNTP is used
  - Each dNTP contains a different fluorophore
- During sequence
  - A single labeled dNTP enters the polymerase
  - dNTP held in place shortly
  - Fluorescence signal is emitted in the ZMW for a short period of time
  - dNTP leaves and new dNTP enters

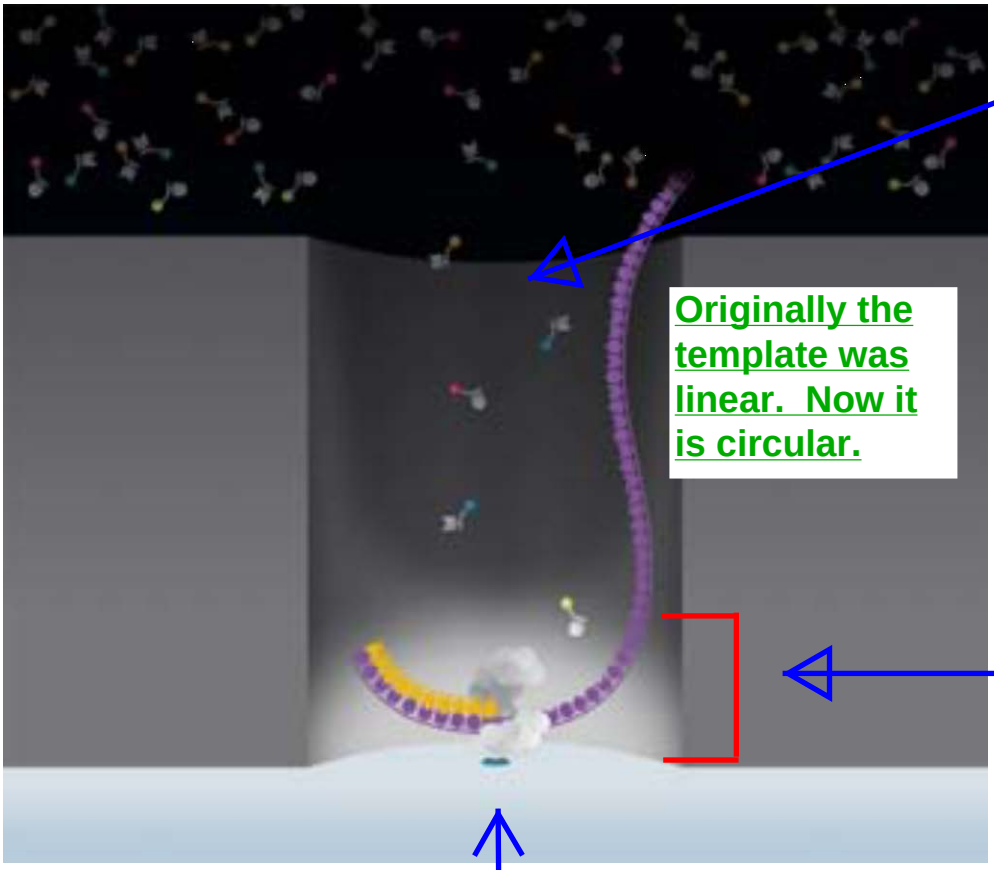
Detection and sequence determination

- Fluorescence signals for each ZMW collected
  - Data is collected as a movie of the sequential signals
    - Each individual signal is measured as a short pulse of light
  - Successive fluorescence signal data is collected
  - DNA sequence of single molecule is determined by sequence of light pulses

Images and Notes Below From:

Pacific Biosciences Technology Backgrounder (11/24/2008)

Title: Pacific Biosciences Develops Transformative DNA Sequencing Technology: Single Molecule Real Time (SMRT) DNA Sequencing

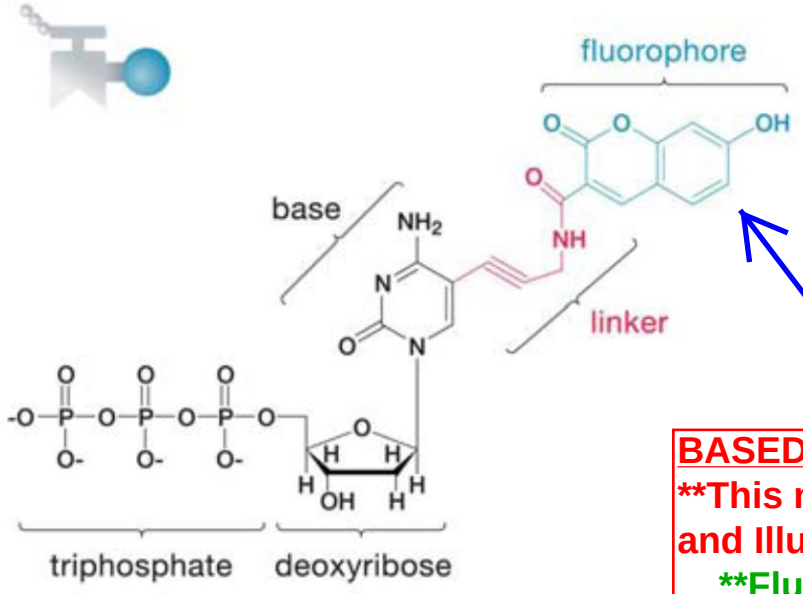


ZMW (Zero-mode waveguide) with Φ29 DNA polymerase and DNA template

ZMW is the sequencing reaction well.

A single DNA molecule HELD IN PLACE by the DNA polymerase enayme.
**Different from Illumina sequencing where the DNA template is bound to the flow cell

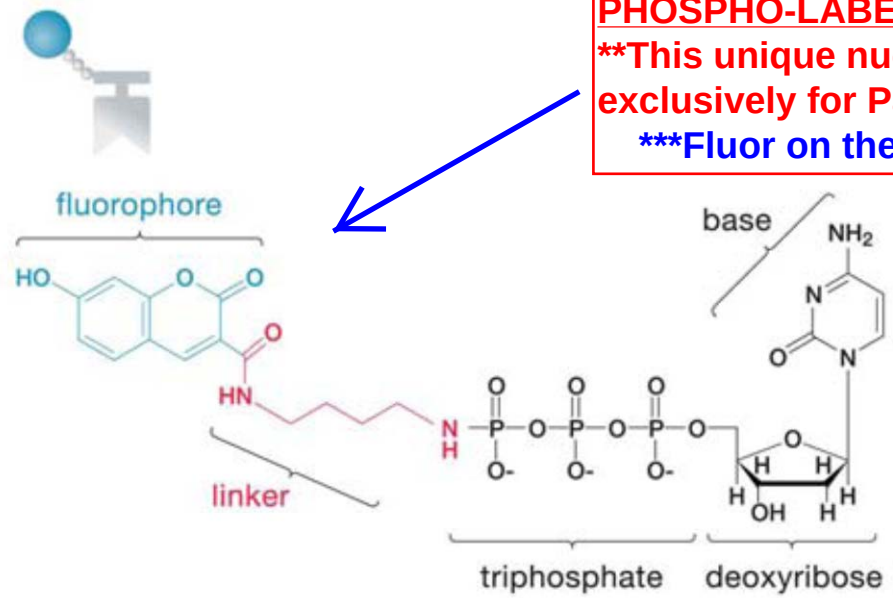
Base-labeled dNTP

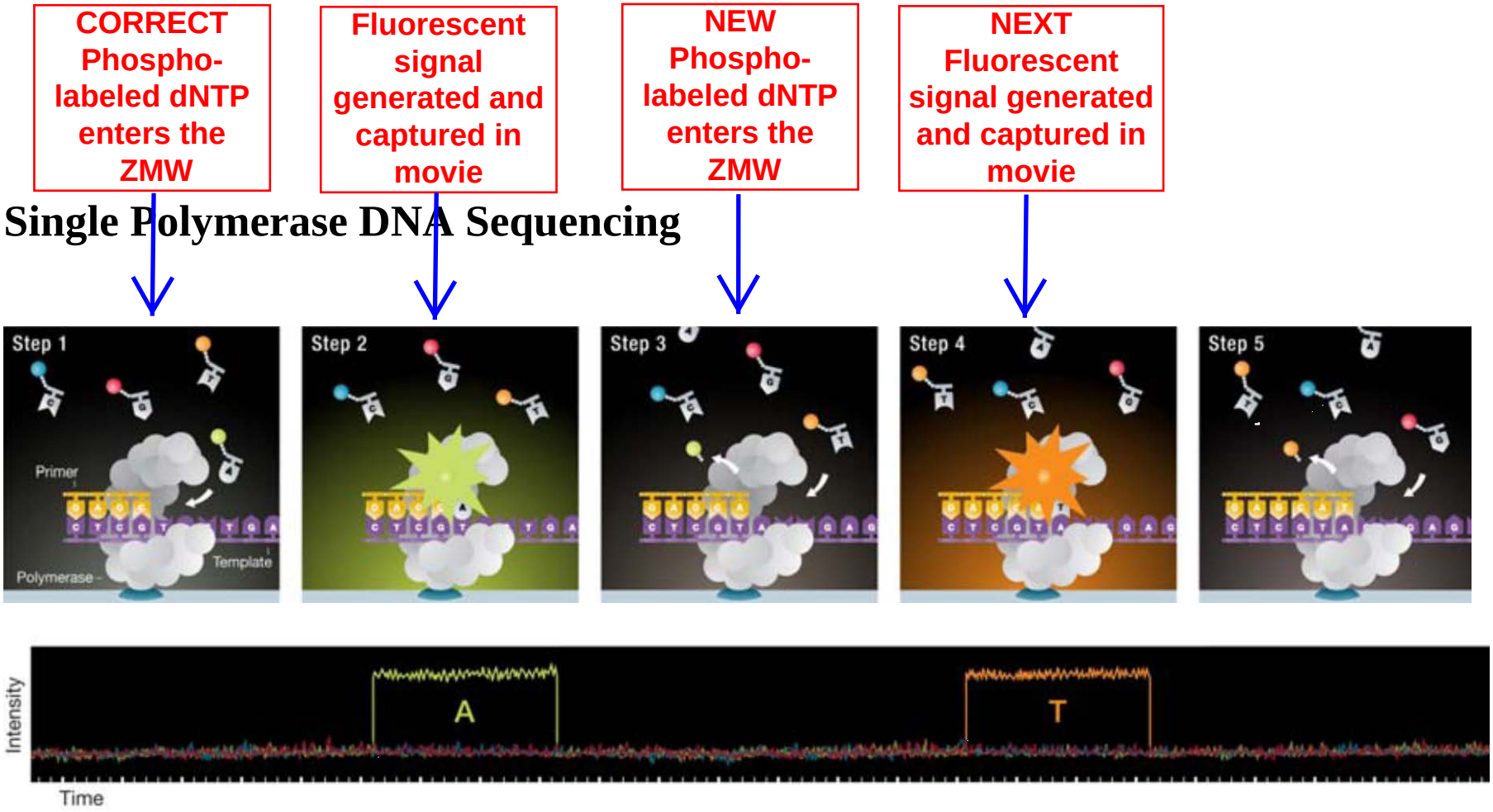


BASED-LABELED dNTP
****This nucleotide is used for Sanger and Illumina sequencing protocols.**
****Fluor on the N-BASE**

Phospho-labeled dNTP

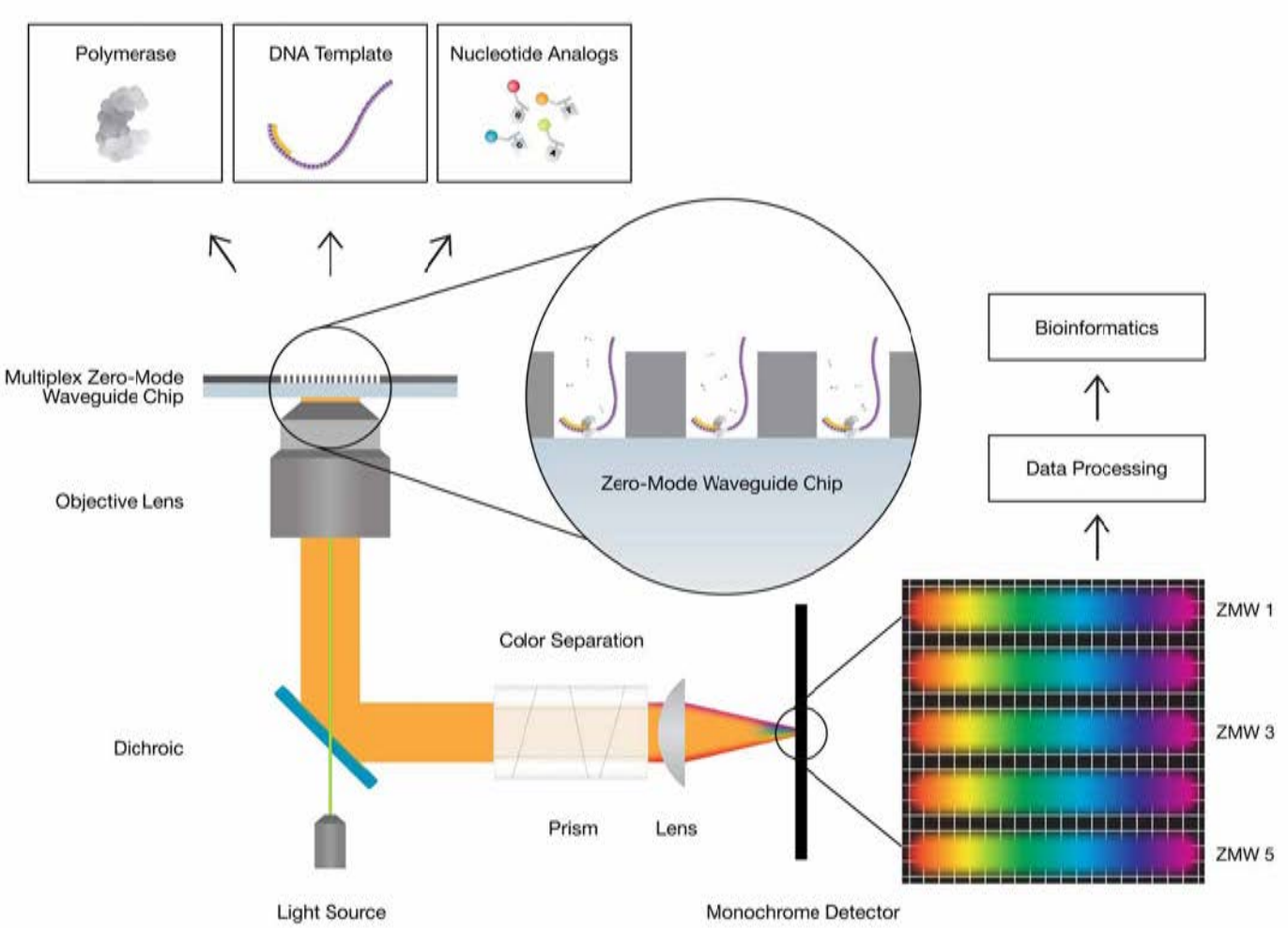
PHOSPHO-LABELED dNTP
****This unique nucleotide is used exclusively for PacBio sequencing.**
*****Fluor on the PHOSPHATE GROUP**





Step 1: Fluorescent phospholinked labeled nucleotides are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing a bright flash of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Step 4-5: The process repeats.

This shows that a movie is made for EACH of the ZMWs.



Pac Bio Newest Technology
****SEQUEL II System**
****8 million wells**
****Revio System**
****25 million wells**
HudsonAlpha Institute of Biotechnology Data (March 2020)
****Hi-fidelity CCS (Circular Consensus Sequencing) Mode**
****20 kilobases (kb) read length (up to 30kb)**
****Yield = 25 gigabases (gb) per SMRT cell**
****or 45 bean genomes**
****Long read CLR (Continuous Long Read)**
****30 kb read length (up to 60kb)**
****Yield = 120 gb per SMRT cell**
****or 218 bean genomes**
CCS IS PREFERRED BECAUSE OF READ ACCURACY!!!!

Newest PacBio System
Revio
****2023 Release**
****25 million ZMW wells**
****15x increased output**

This system is NOT based on DNA replication!!
****ONT reads the native DNA sequence!!**

Oxford Nanopore Technology

https://www.youtube.com/watch?v=E9-Rm5AoZGw

Concept **(Founded 2005)**

- Disruption of current flow through nanopore is distinctive for each nucleotide

Tools of Nanopore

- Substrate
  - Electrically resistant membrane
- Complex
  - Protein nanopore embedded in the membrane

Potential applied to membrane

- Current flows only through the aperture of the nanopore
- Molecules that flow through the nanopore cause a characteristic change in the current flow
  - Measuring the disruption allows the molecule to be identified

Processive enzyme
**Enzyme moves along DNA molecule
**Polymerase, helicase, nuclease??
**Trade secret about which enzyme ONT uses
**Continuing to develop new enzymes

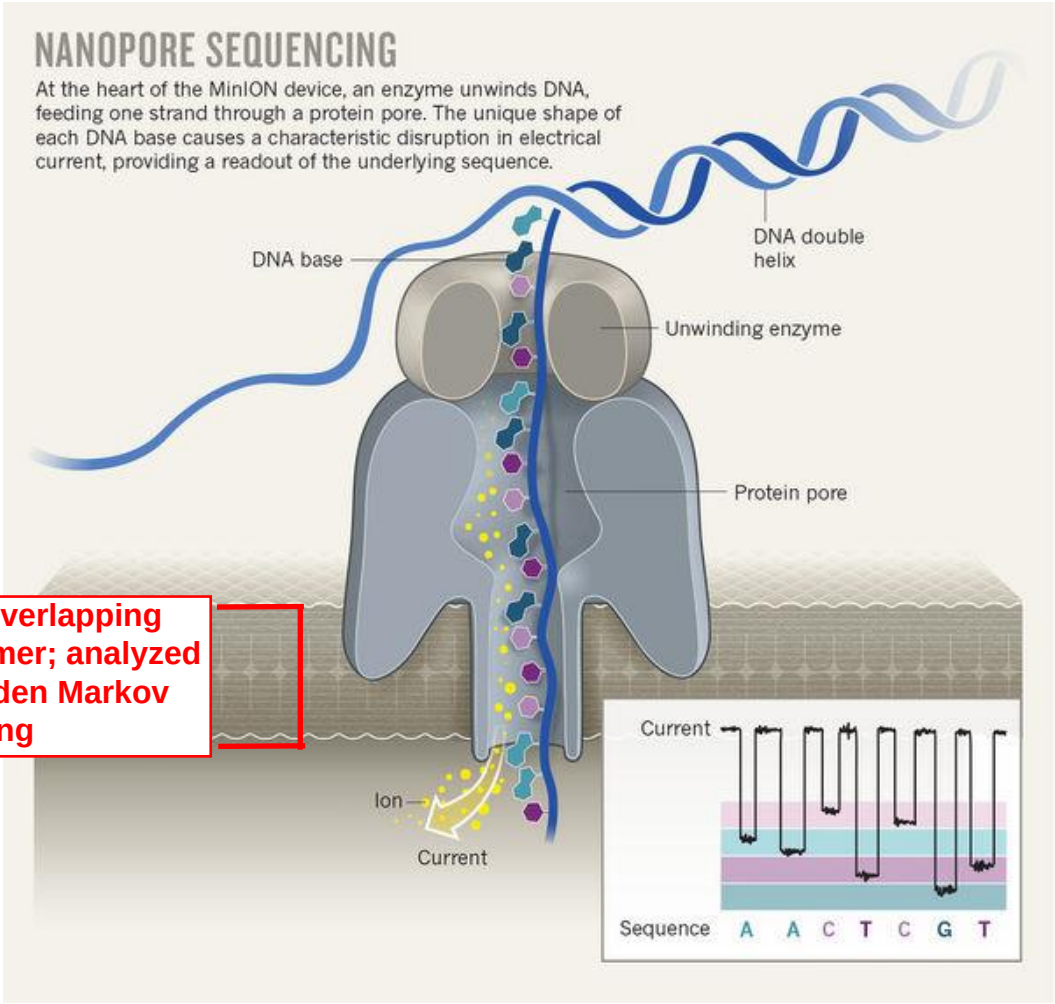
Nanopore uses a

- Strand sequencing method
  - A processive enzyme is bound to the DNA to be sequenced
    - DNA strand pulled through the nanopore by the enzyme
      - One base at a time
- Read length
  - 100s of kilobases
- Both strands can be read
  - DNA preparation creates a hairpin on one strand
    - Second strand read after first strand finished

Maximum read length observed:
**2 Mb = 2,000 kilobases

from $1,760 | from $1,000 | $49,955 | $285,000

	Instrument				
	Flongle	**MinION Mk 1B**	**GridION X5**	**PromethION (1 flow cell)**	**PromethION (48 flow cells)**
Run time	1 min - 16 hrs	1 min - 48 hrs	1 min - 48 hrs	1 min - 48 hrs	1 min - 48 hrs
Yield (Theoretical)	Up to 3.3 Gb	Up to 40 Gb	Up to 200 Gb	Up to 315 Gb	**10.5 Tb**
Current yield	NA	Up to 30Gb	Up to 150Gb	Up to 150 Gb	NA
Number of channels	Up to 126	Up to 512	Up to 2,560	Up to 3,000	Up to 144,000



Read overlapping 6bp k-mer; analyzed by Hidden Markov Modeling

Flongle

MinION

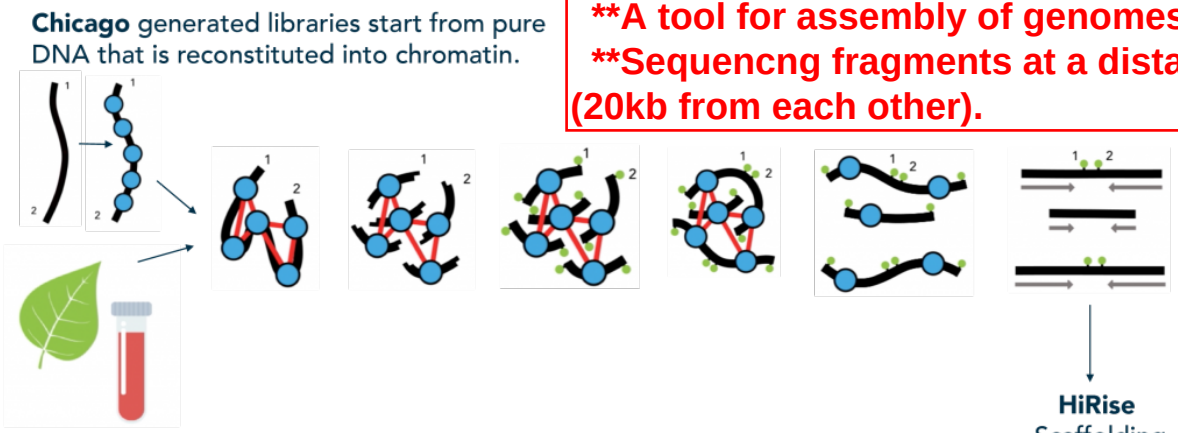
GridION

PromethION

Proximity Ligation Sequencing
Major Tool for Genome Assembly

Dovetail Genomics Sequencing

Hi-C LIBRARY SEQUENCING
****A tool for assembly of genomes**
****Sequencng fragments at a distance (20kb from each other).**



from: https://dovetailgenomics.com/technology/

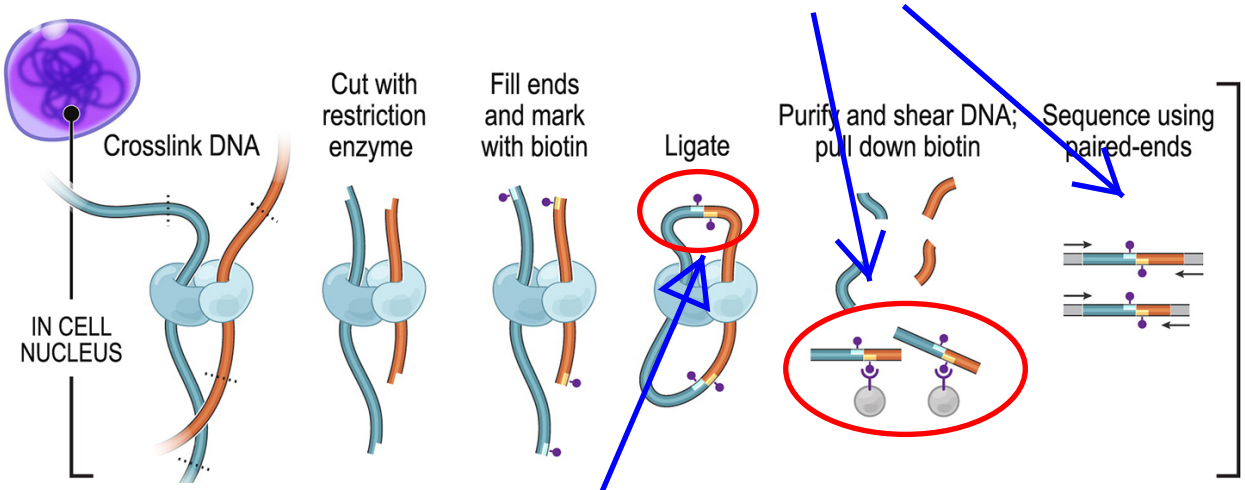
Hi-C linking

- Based on links between natural interactions within a chromosome
  - Regions of the chromosome are associated via chromatin
  - Based on principal that DNA has a 3-D confirmation in the cell
  - 3D configuration occurs because controlling elements that regulate a gene's expression are not always immediately adjacent to coding region of the gene

Chicago **Not currently used to aid genome assembly**

- An artificial linking procedure
- When used with Hi-C, the Hi-C derived relationships can be confirmed

These sequence reads are 20kb apart. That are used to LINK SCAFFOLDS during assembly.



from: http://science.sciencemag.org/content/326/5950/289/tab-pdf

Two regions far apart now linked

Hi-C procedure

1. Crosslink the cells using formaldehyde to stick chromosomes together
2. Isolate "crosslinked" DNA bound with chromatin
3. Digest DNA with six-cutter restriction enzyme
4. Fill ends and add biotin to end
5. Ligate ends and pull down molecules with biotin procedure
6. Sequence pull down library using Illumina paired-end protocol

Assembly

- Long distance relationships can be used during assembly
- Distances between ends are typically >20Kb
- Data can be used in the final steps of assembly.

Common bean
560 Mb genome
$10,000 for basic sequence data

General Steps That Apply To ALL Massively Parallel DNA Sequencing Systems

1. Isolate DNA

- Care is needed to ensure the DNA is of uniform high quality

2. Fractionate DNA into appropriate size for specific sequencing system

- Length will vary depending on the read length you will be generating

3. Amplify individual DNA fragments that will be sequenced

- This could be in a reaction emulsion bead (Roche 454) or reaction matrix (Illumina or Pacific Biological Science [PacBio])

4. Load DNA samples onto DNA sequencing matrix

- The matrix can be a solid chip with individual wells (Roche 454, PacBio) or a chip with sequencing oligonucleotides (Illumina)

5. Perform sequencing reactions

- Varies from system to system

6. Collect DNA sequence data for each read

- Varies from system to system

RNA sequence data need for gene modeling
**MULTIPLE TISSUES ARE SEQUENCED
WHY?
**Genes are expressed in a temporal (time) and spatial (tissue) manner

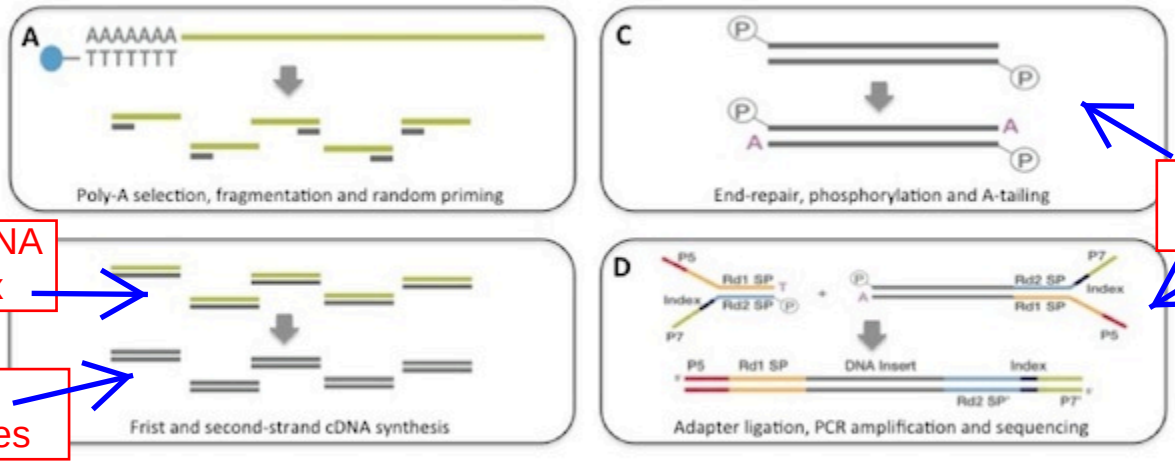
Sequencing the Expressed Portion of the Genome

10 day leaf
10 day root
19 day leaf
19 day root
10 day stem
19 day stem
Flower buds
Petals
Pods
Seed coats
**4 stages

- Genes are expressed in a the following manners
  - Tissue-specific (where)
  - Temporal specific (when)
  - Quantitatively (how much)
- Transcriptomics
  - The study of gene expression
- Massively parallel sequencing has changed the study of the transcriptome
  - All the genes at a specific place or time can be accurately quantified
- Procedure
  - RNA-seq or massively parallel RNA sequencing
    - Very powerful
    - Can monitor expression of even rarely expressed genes

RNA-seq costs ~$10K for a genome project

Illumina Tru-Seq RNA-seq protocol



Paired-end sequencing

mRNA/cDNA complex

cDNA molecules

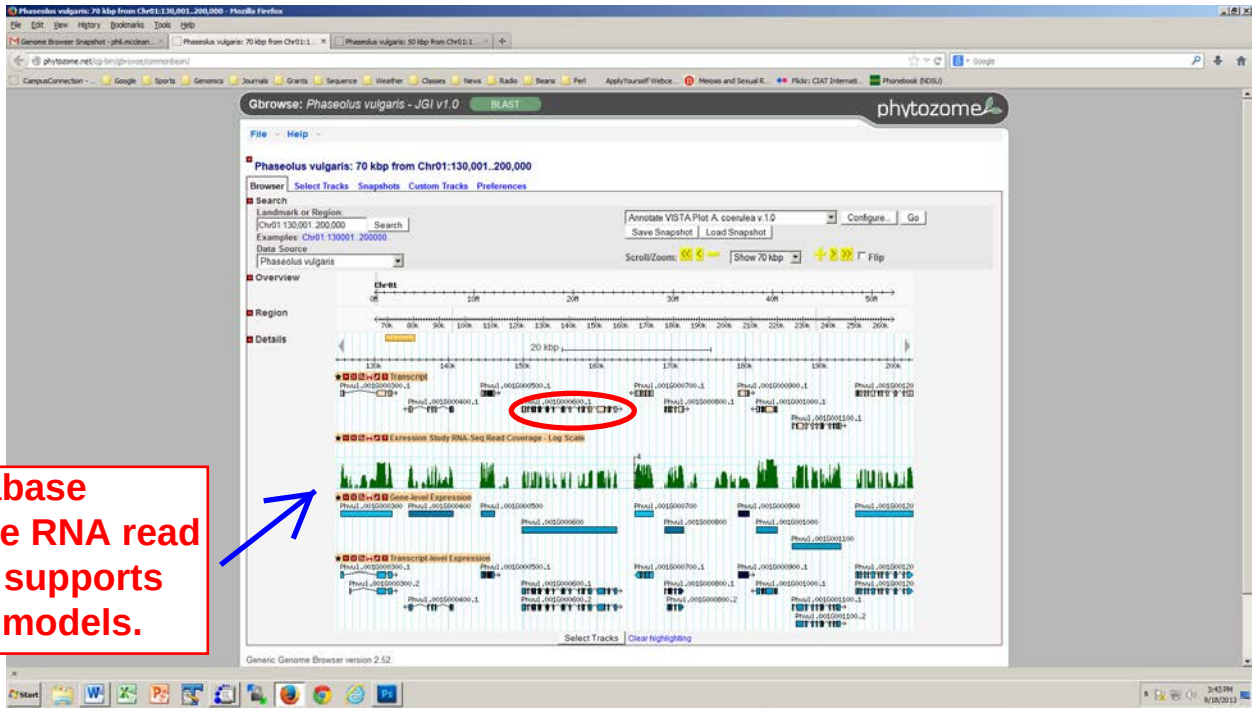
Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

RNA-seq procedure

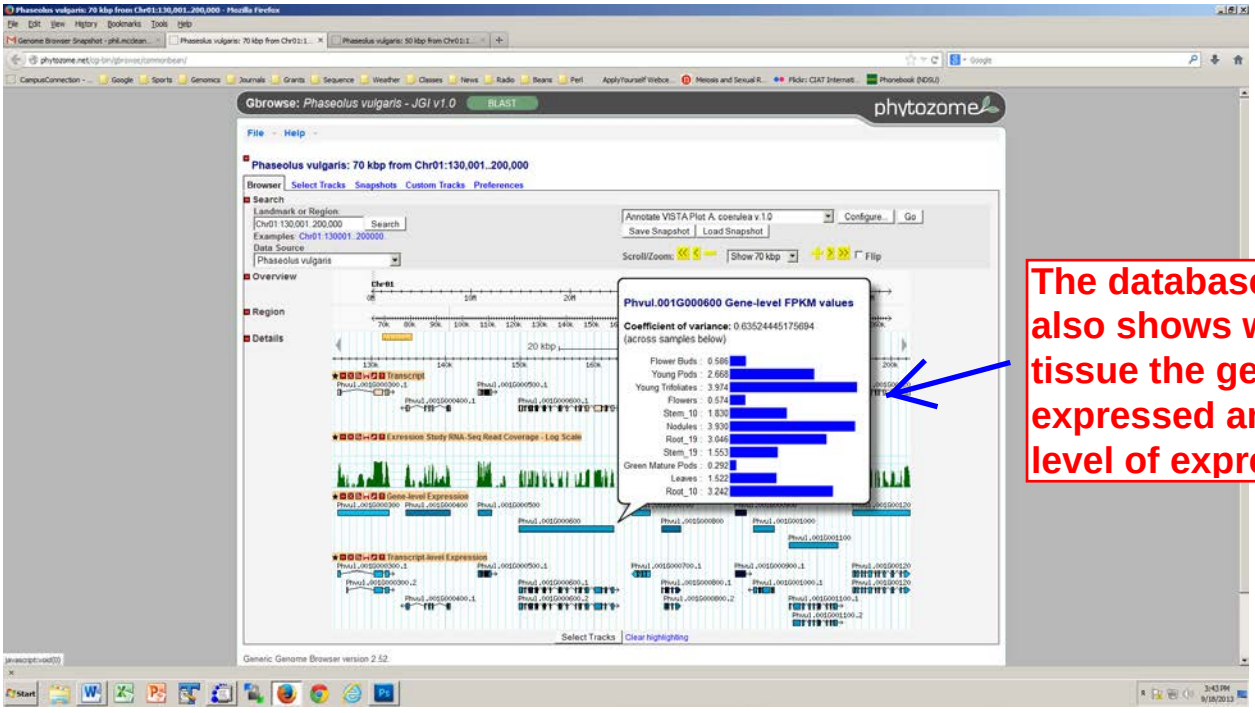
1. Isolate RNA from target tissue
2. Select mRNA using poly-T primers
   - Based on principle that all mRNA in eukaryotes have a poly-tail
3. Perform first and second strand cDNA (copy DNA) synthesis to convert mRNA into cDNA
4. Prepare cDNA for sequencing by adding appropriate sequencing adaptors
5. Sequence the cDNA pool using a massively parallel technology
6. Align reads against a reference genome and quantify

Assembly of mRNA reads into transcripts
****RNA-seq data represent expressed mRNAs = genes**
**Overlapping reads are identified by PERTRAN
Assembled transcripts are used to discover genes during **GENE MODELING

Aligning RNA-seq Data to the Reference Sequence



This database shows the RNA read data that supports the gene models.



The database also shows which tissue the gene is expressed and the level of expression.

This is the approach that was used for SEQUENCING PLANT GENOMES until ~2017. It required the sequencing of cloned fragments of different sizes. Paired-end reads were collected

Plant Genome Sequencing

Traditional Sanger Sequencing Genome Sequencing Approach

1. Create sequencing libraries of different insert sizes

- 2kb
  - Bulk of sequencing is performed on these libraries
- 10kb
  - Used for linking contigs during assembly
- 40kb
  - Used to link larger contigs assembly
- Bacterial artificial chromosomes 100-150 kb
  - Used to link ever larger contigs assembly

2. Paired-end sequencing data collected for libraries

3. Contigs created by looking for overlapping reads

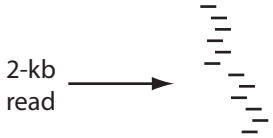
4. Contigs assembled based on homology to 10kb, 40kb and BAC sequence data; these large assemblies are called **scaffolds**

5. Pseudochromosomes assembled based on homology of scaffolds to the markers located on a high-density genetic map

Scaffold Assembly

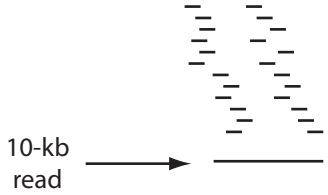
Building a Scaffold Using Paired-end Reads of Different Sized Sequences

Step 1: Build a contig with overlapping 2-kb paired-end reads

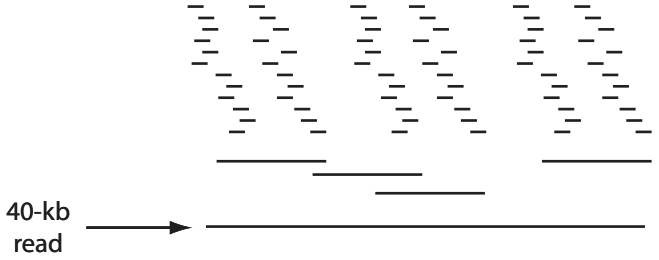


This ASSEMBLY approach is the TRADITIONAL method. Paired-end read data from libraries based on fragments of different sizes (10kb, 40kb, 100kb) is used to convert read data into assembled scaffolds.

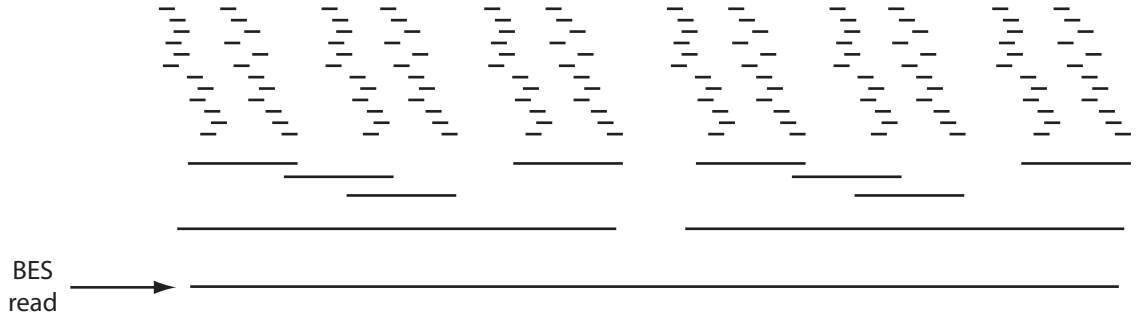
Step 2: Link two contigs with 10-kb paired-end reads



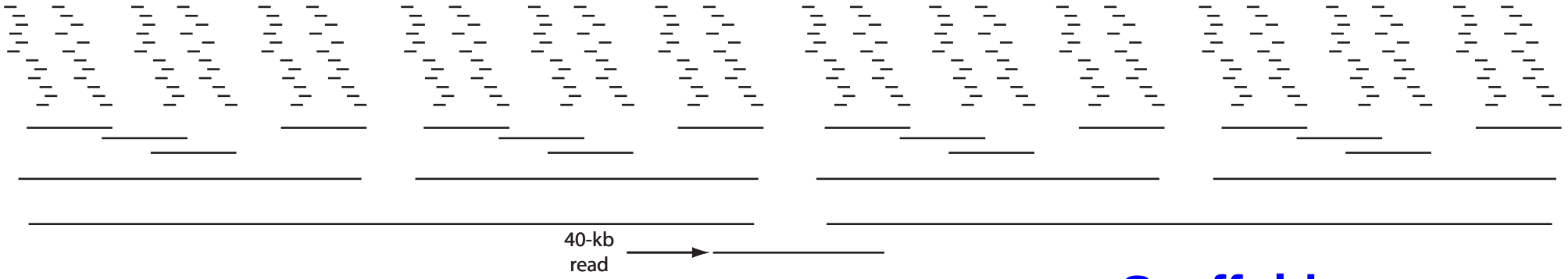
Step 3: Link three 10-kb contigs with 40-kb paired-end reads



Step 4: Link two 40-kb contigs with 100-kb BAC end sequences (BES)



Step 5: Here link two100-kb BAC sized contigs with a 40-kb paired-end read; other sized reads can also be used for this linking

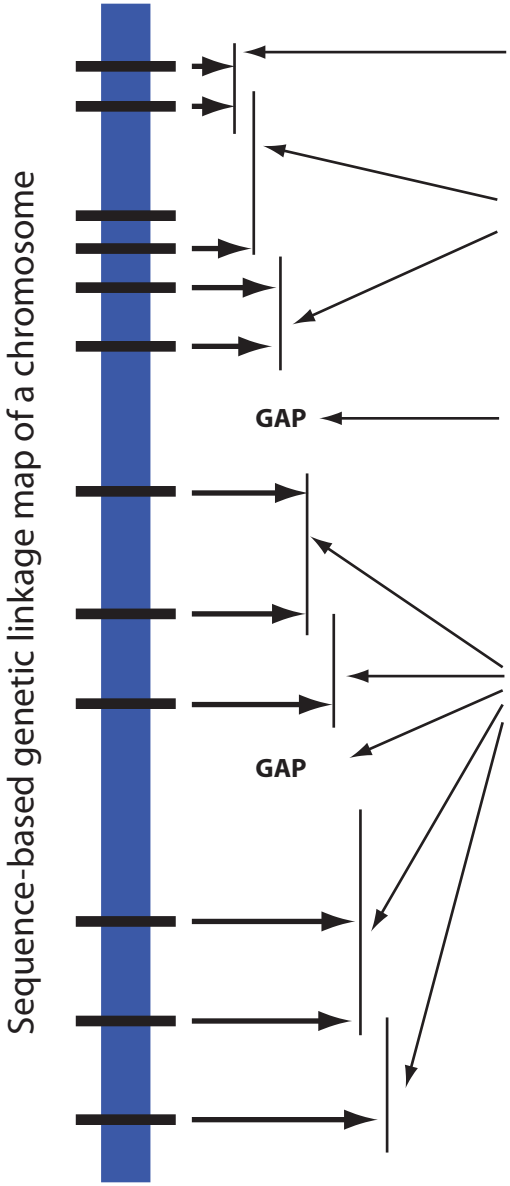


Scaffold

Step 6: Continue linking larger blocks of sequences until the block can not be linked with another block. This block is defined as a scaffold.

Genome Assembly

Linking Scaffolds to a Dense Genetic Map



Step 1: Place scaffold relative to sequence complementarity of marker

Step 2: Sequentially place other scaffolds relative to complementarity of markers

Step 3: If no scaffold is complementary to a marker, a gap is inserted relative to the sequence of genetic map. These are represented as "Ns" in the sequence.

GAP

Max # of Ns = 10,000

Step 4: Repeat steps 1-3 until a chromosome length sequence is developed. The overlapping sequences of each of the linked scaffolds defines a pseudochromosome.

GAP

Hi-C library sequencing

A
A
T
G
C
T
C
T
A
C
N
N
N
N
A
A
T
T
G
C
T
N
N
N
C
A
T
G
G
C
T
A
A
T
T

This figure represents assembling PSEUDOCHROMOSOMES

****Based on by linking scaffolds to a molecular marker map**

****The sequence of the markers provides a accurate data for the organization of the scaffolds**

REMEMBER genetic data is still the most useful data for assembly

****It is directly related to recombination events.**

How Reference Genome Assemlies are developed today

Modern Long Read PacBio Sequencing Genome Sequencing Approach (~2017-2020)

1. Create 20kb insert libraries

2. Sequence with PacBio single molecule technology
   - Reads generally 10-15 kb in length

3. Add short read (150bp) paired end data to correct for inherent PacBio errors

4. Assembly reads into contigs
   - Contigs MUCH longer than with Sanger sequencing

5. Scaffolds developed by long-range scaffolding methods
   - BioNano restriction enzyme mapping
   - Hi-C cross-linked DNA library sequencing
   - 10X linked read sequencing

This method was discontinued by the company!!!

6. Pseudochromosomes assembled based on homology of scaffolds to the markers located on a high-density genetic map

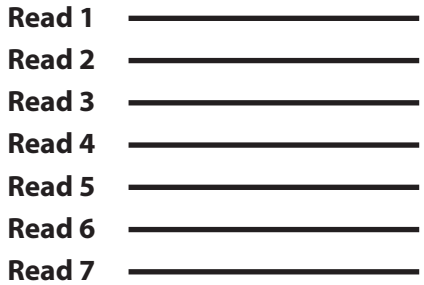
Preferred methods now!!!

Now (2020-current) with HiFi (CCS) mode reads

****HiFi (CCS) reads lead to contigs that are much longer**

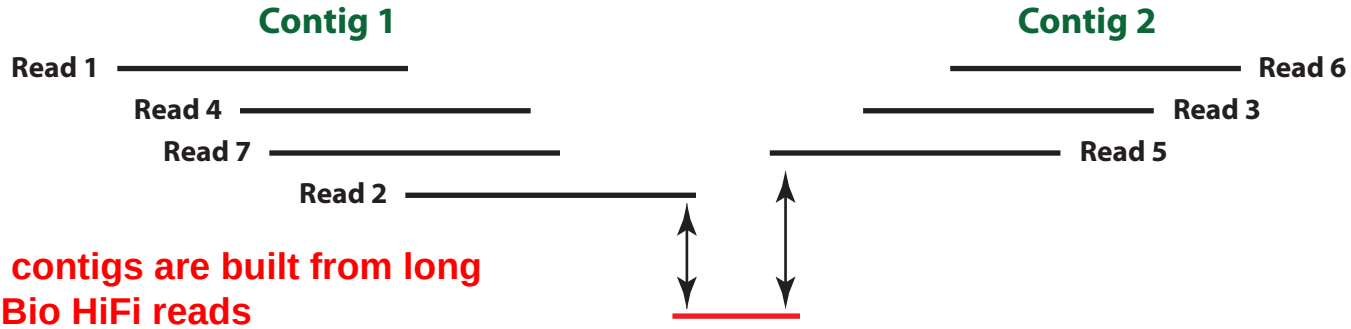
****Hi-C data is used to assembly contigs into pseudochromosomes**

Chromosome Segment Assembly



PacBio Hi-F CCS Reads
How can these be ordered?
**Hi-C reads

1. CCS Reads Aligned
****Contigs defined**



The contigs are built from long PacBio HiFi reads
****Not the short reads used in older Illumina-based projects**

2. Hi-C Read links two contigs

Product: Assembled Segment of Pseudochromosome

N50 and L50: Measures of the Quality of Genomes

Contig

- An aligned group of reads that represent one section of the genome
    - No missing sequence data

Scaffolds

- Groups of contigs that define a section of the genome
    - Larger than contigs
    - Can contain gaps (missing sequence) that are filled in with Ns
    - Number of scaffolds is always smaller than the number of contigs

Pseudochromosome

- Group of scaffolds that represent one chromosome of the species

N50

- The number of contigs (or scaffolds) whose collective distance equals 50% of the genome length
    - This is a **NUMBER**

The SMALLER the number the BETTER the genome.

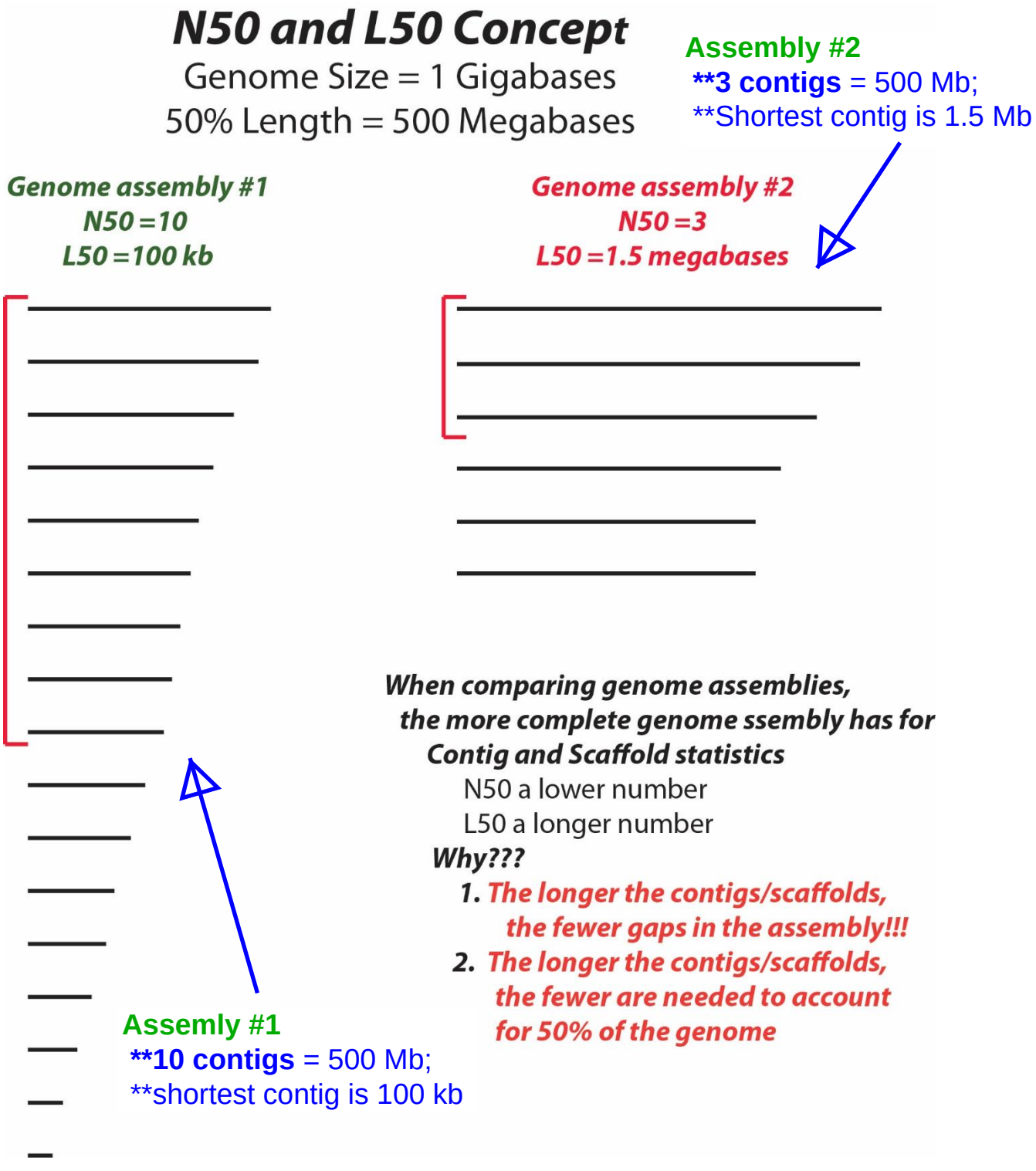
L50

- The length of the smallest contig (or scaffolds), of the collection of the contigs (or scaffolds)that comprise the set of N50 contigs (or scaffolds)
    - This is a **LENGTH**

The LARGER the number the BETTER the genome.

IMPORTANT NOTE
Today, the L50 length is almost always reported as the N50

Graphic Illustration of N50/L50 Concept



Measuring the Completeness of A Genome Assembly Based on Gene Content

BUSCO: A measure of the completeness of the genome based on single copy orthologs

BUSCO software

- **B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs
    - Evolutionary based approach
    - **What percentage of your single copy genes** found in your genome annotation are found in a database set of genes (n=~400) found in other species in your high-order biological taxa
- Uses a **subset (n~400)** of nearly universal single copy orthologs for each database
    - Highly represented in major biological taxa for comparative purposes
- Multiple databases (Busco v.5.7) are available for each major biological lineage
    - Plants (n=9)
        - e.g. Eudicots, Fabales, Brassicales, Liliopsida, Poales,
    - Vertebrates (n=15)
    - Arthropods (m=8)
    - Fungi (n=24)
    - Bacteria (n=83)
    - Viruses (n=27)
- Data set is based Orthodb database release 10
    - Orthodb
        - Database of orthologous protein-coding genes
        - Data set specific to each major biological taxa

What is an Ortholog?

- Genes in multiple species that are evolutionarily related by descent
    - A gene found throughout a **biological lineage**
- Example for plants:
    - Genes related to photosynthesis

What is a Paralog

- Genes within a species that are evolutionarily related by duplication
    - A gene found only in that species
- Example for a plant species:
    - Disease resistance genes

Evolution of Genome Sequencing

Effect of Evolving Sequencing Technologies

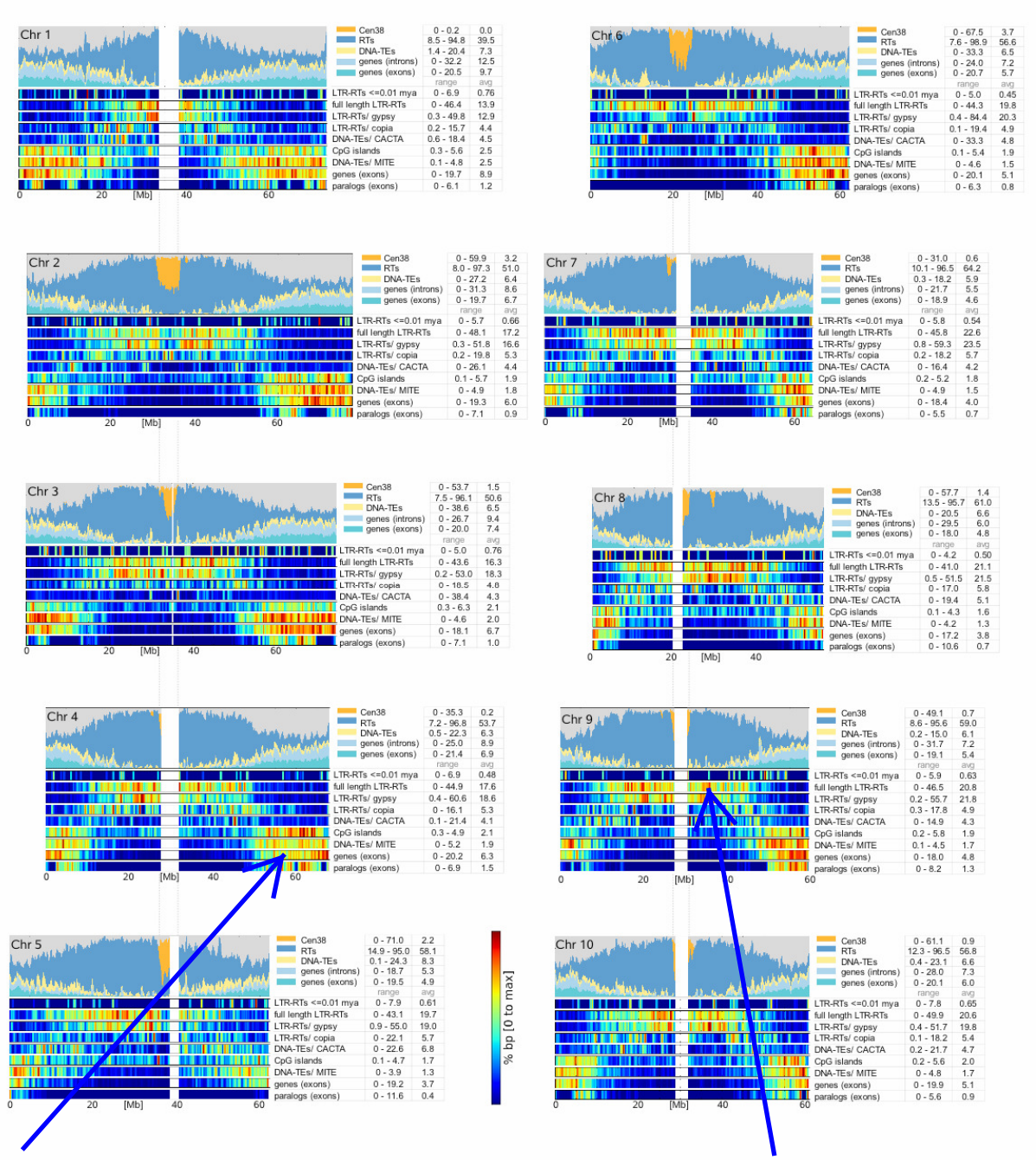
Common Bean Genome Examples

Genotype, version, release date	**G19833 v1 (2014)**	**G19833 v2 (2015)**	**UI 111 v1 (2019)**	**Labor Ovalle v1 (2020)**	**5-593 v1 (2022)**
Primary Technology	**Roche 454**	**PacBio RSII**	**PacBio Sequel II**	**PacBio Sequel II CCS**	**PacBio Sequel II CLR**
Average read length	-	-	8.5 kb	19.6 kb	20.2 kb
Coverage	19.2X	83.2X	141.5X	51.2X	135.0X
Main genome scaffold total	708	478	58	15	13
Main genome contig total	41,391	1,044	167	36	27
Main genome scaffold sequence total	521.1 Mb	537.2 Mb	554.9 Mb	571.9 Mb	572.2 Mb
Main genome contig sequence total	472.5 Mb (9,3% gap)	531.6 Mb (1.1% gap)	553.8 Mb (0.2% gap)	571.7 Mb (0.0 % gap)	572.1 Mb (0.0 % gap)
Main genome scaffold N50/L50	5/50.4 Mb	5/49.7 Mb	5/51.0 Mb	5/55.5 Mb	5/54.8 Mb
Main genome contig N50/L50	3,273/39.5 kb	73/1.9 Mb	28/8.5 Mb	9/20.5 Mb	7/33.5 Mb
# Protein coding genes	27,197	27,433	27,385	27,218	27,065

Remember:

Contig N50 (again formerly L50) is the critical statistic that measure assembly completeness

DISTRIBUTION of GENES and REPEATS in Sorghum genome.
**Typical of most eukaryotic genomes



Most GENES are located at the ends of chromosomes

Most LTR REPEATs are located in the heterochromatic region of chromosomes

Reference: Paterson et al (2009) Nature 457(7229):551

Genome Resequencing

Goal

- Discover variation in a population

How?

- Resequence many individuals
- 10x – 40x, depending on the goal

Types of variants

Also called SNVs
= single nucleotide variants

- SNPs
  - Single nucleotide differences among a population
- Indels
  - Typically short in length
  - 1 to 50 nt
- Copy Number Variants (CNVs)
  - No clear definition
    - Depends on the research group
  - Often considered >1000 nt
    - Can be just 50 nt