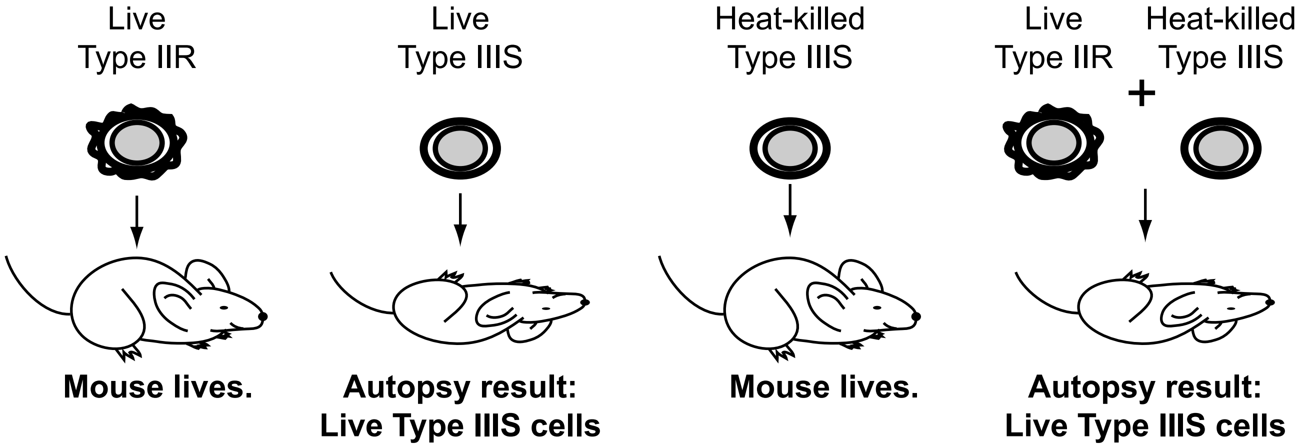
Griffith and the Transforming Principle

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 transformation 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*.

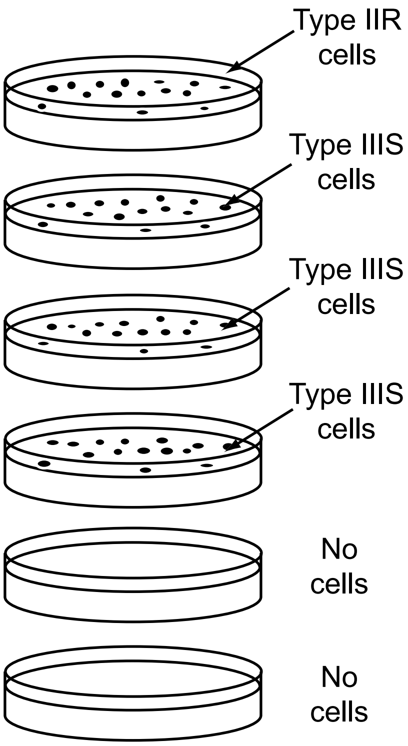
Avery, MacLeod and McCarty: DNA Is The Genetic Material

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 IIR Antibody	Enzyme
+			
+	+	+	
+	+	+	Protease
+	+	+	RNase
+	+	+	DNase
	+		

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



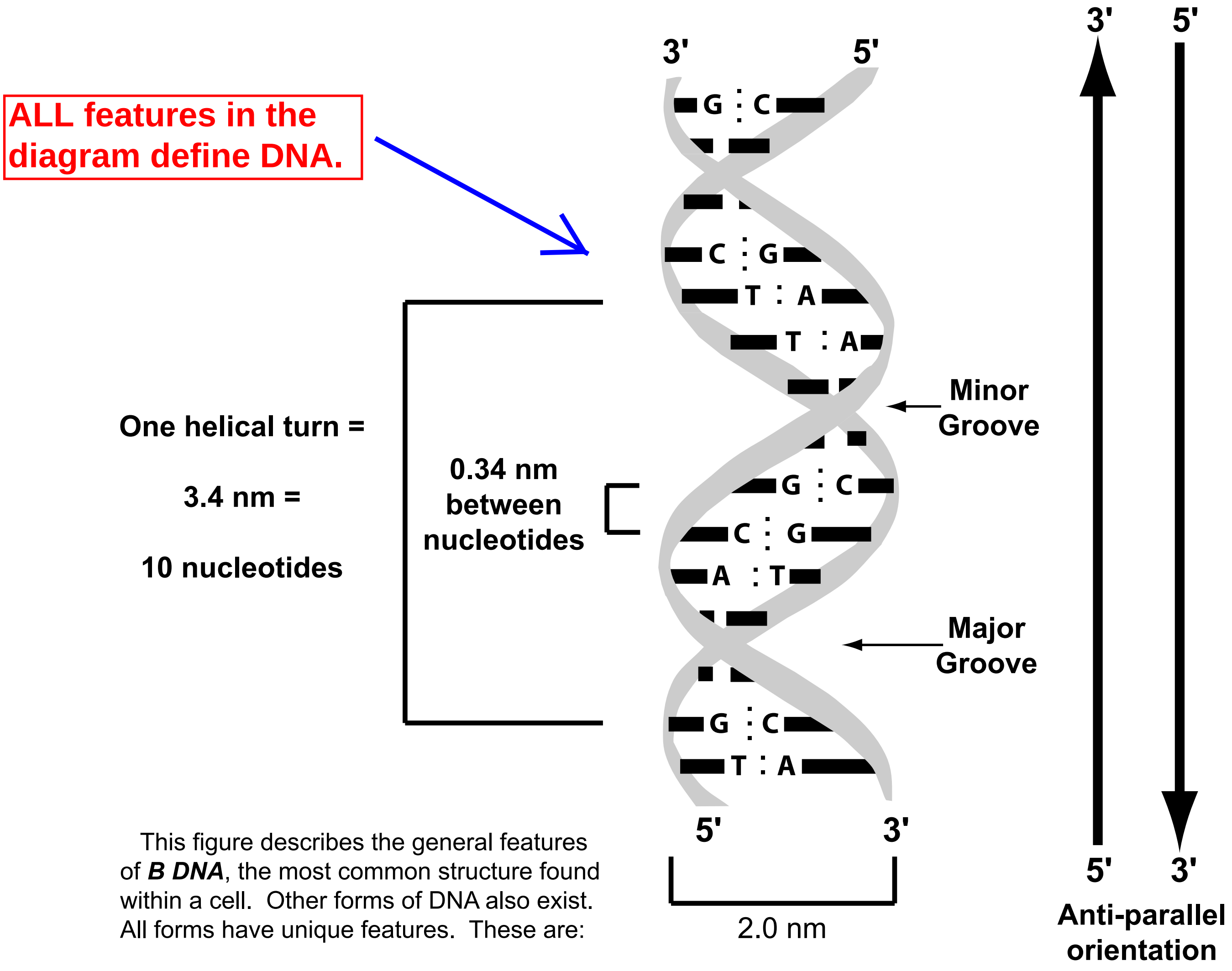
Rather than work with mice, Avery, MacLeod and McCarty used the phenotype of the *Streptococcus pneumoniae* cells expressed on blood agar. To ensure, a few potentially live cells did not escape the heat treatment, they also precipitated those cells out of culture using an antibody to the type IIR cells. 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

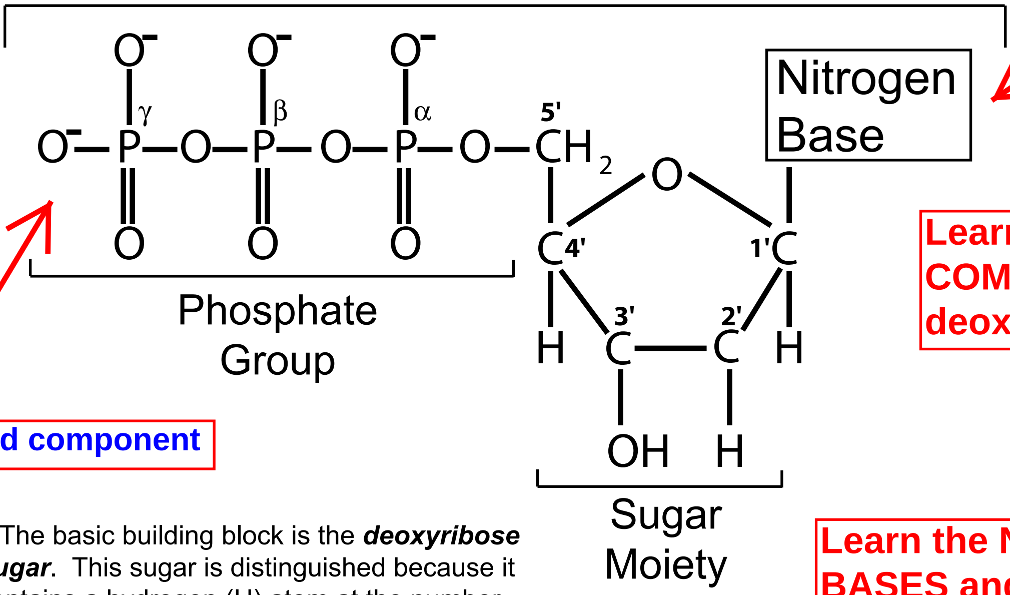
Figure 3. The structure of common DNA molecules.

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



Sanger and Illumina labeled component

PacBio labeled component

Learn the two 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.

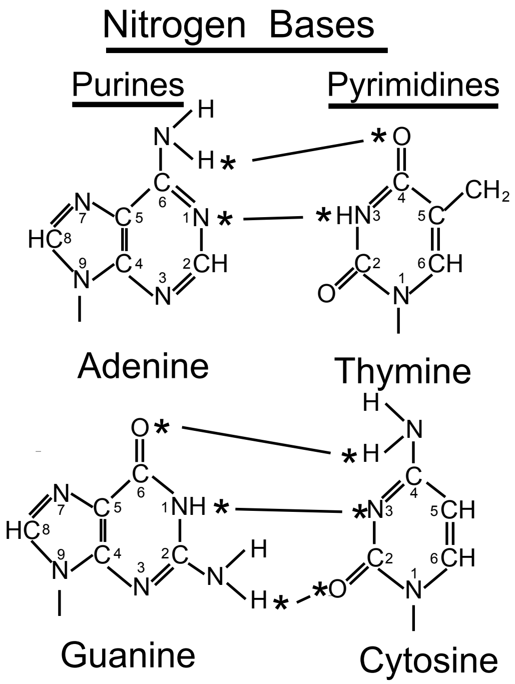
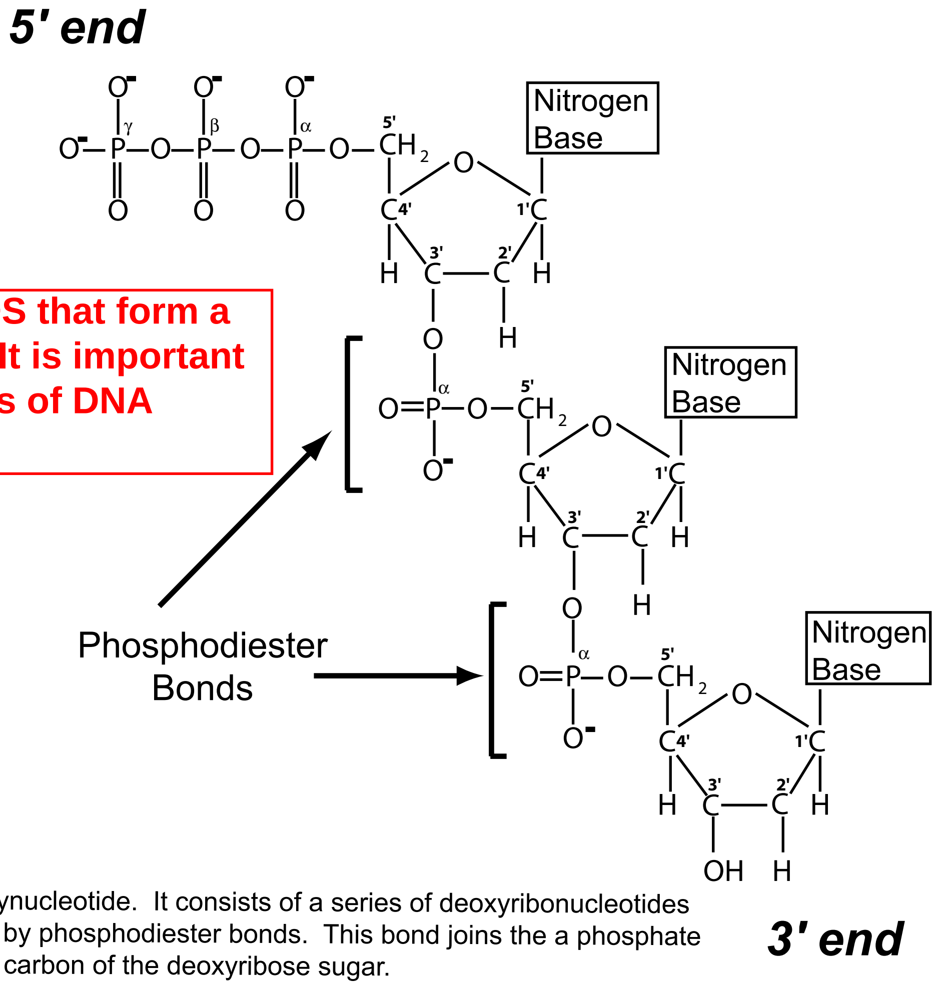


Figure 4. The structure of deoxyribonucleotides and base pairing among N bases.

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 series of 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 a phosphate group to the 3' carbon of the deoxyribose sugar.

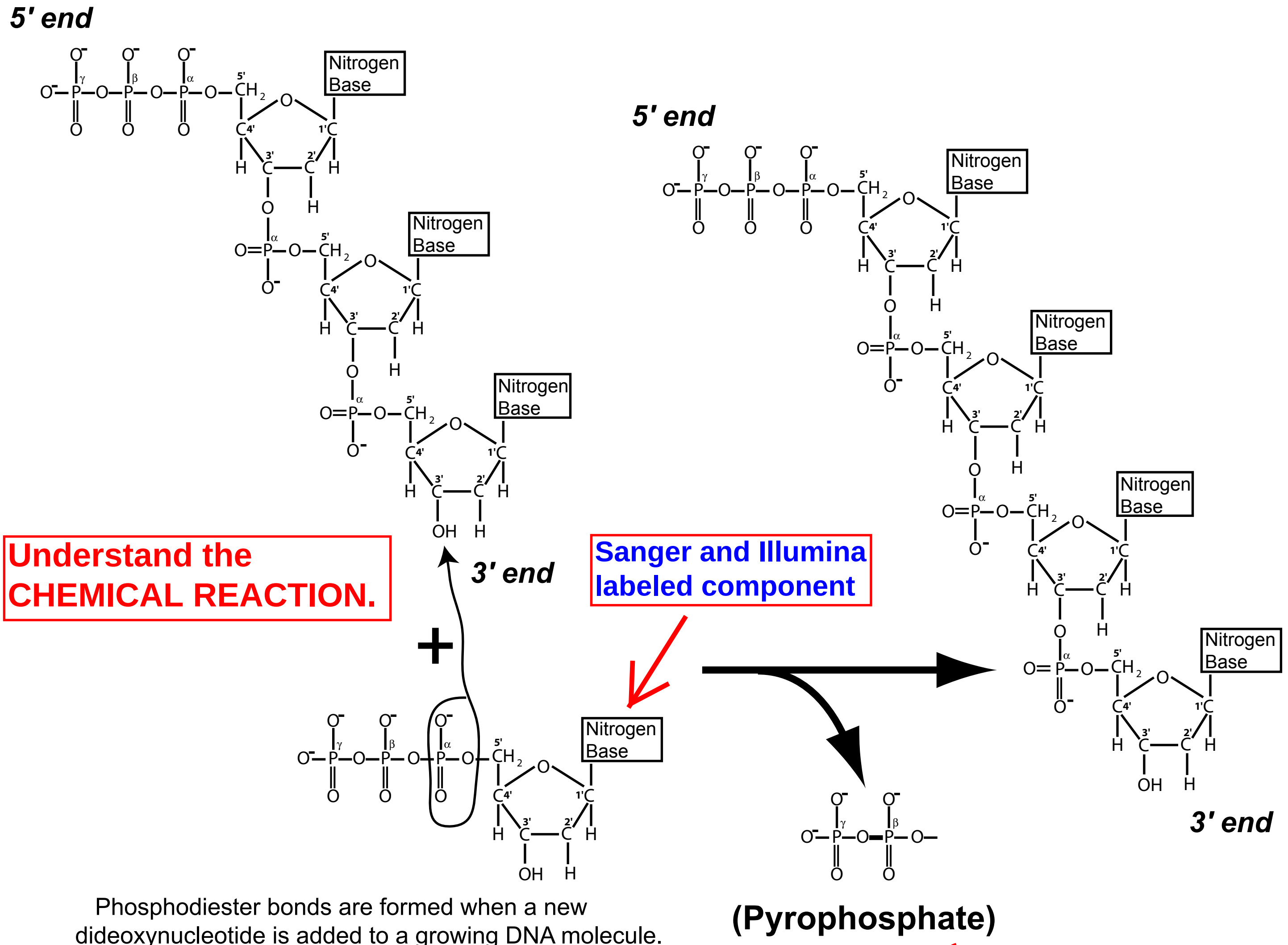
Each strand is complementary to the opposite strand. If one strand has an adenine at a position, its anti-parallele strand would have a thymine at the the corresponding position. Likewise, guanine and cytosine would be complementary.

Fig. 5. The single strand structure of DNA.

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 deoxyribonucleotide.



Understand the CHEMICAL REACTION.

Sanger and Illumina labeled component

PacBio labeled component

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.

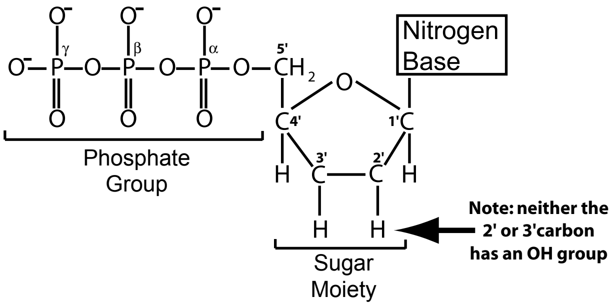
Figure 6. The formation of the phosphodiester bond that grows the DNA chain.

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

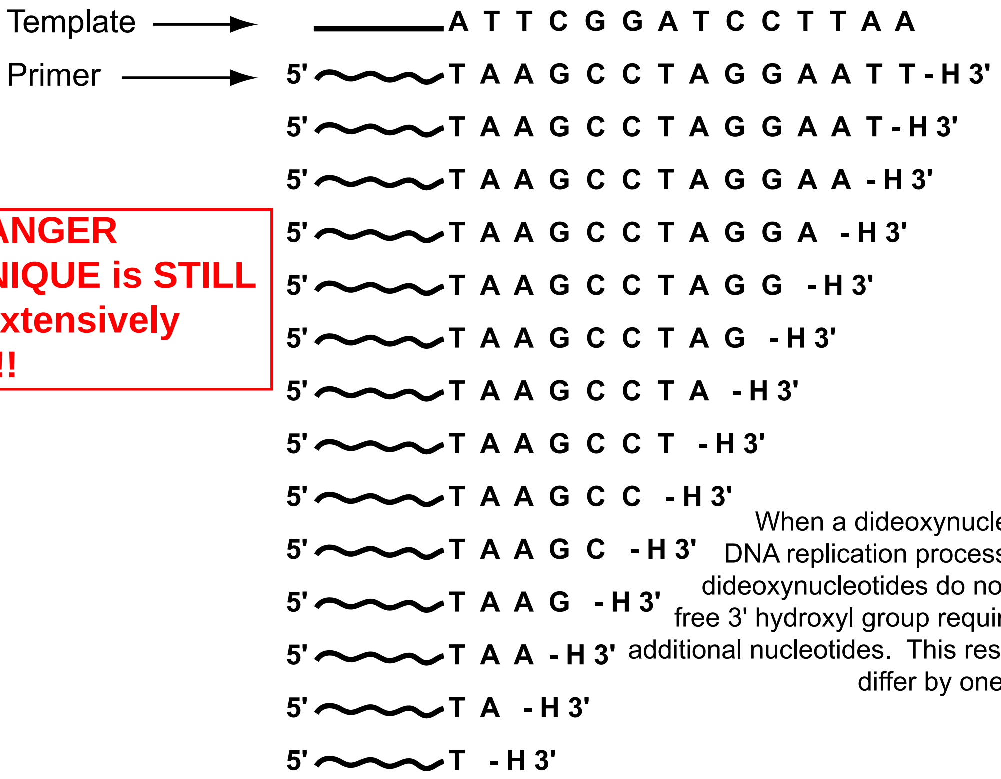


Note: neither the 2' or 3'carbon has an OH group

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 → ATTCGGATCCTTAA
Primer → 5' TAAGCCTAGGAATT-H 3'
5' TAAGCCTAGGAAT-H 3'
5' TAAGCCTAGGAA -H 3'
5' TAAGCCTAGGA -H 3'
5' TAAGCCTAGG -H 3'
5' TAAGCCTAG -H 3'
5' TAAGCCTA -H 3'
5' TAAGCCT -H 3'
5' TAAGCC -H 3'
5' TAAGC -H 3'
5' TAAG -H 3'
5' TAA -H 3'
5' TA -H 3'
5' T -H 3'

The SANGER TECHNIQUE is STILL used extensively today!!!

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.

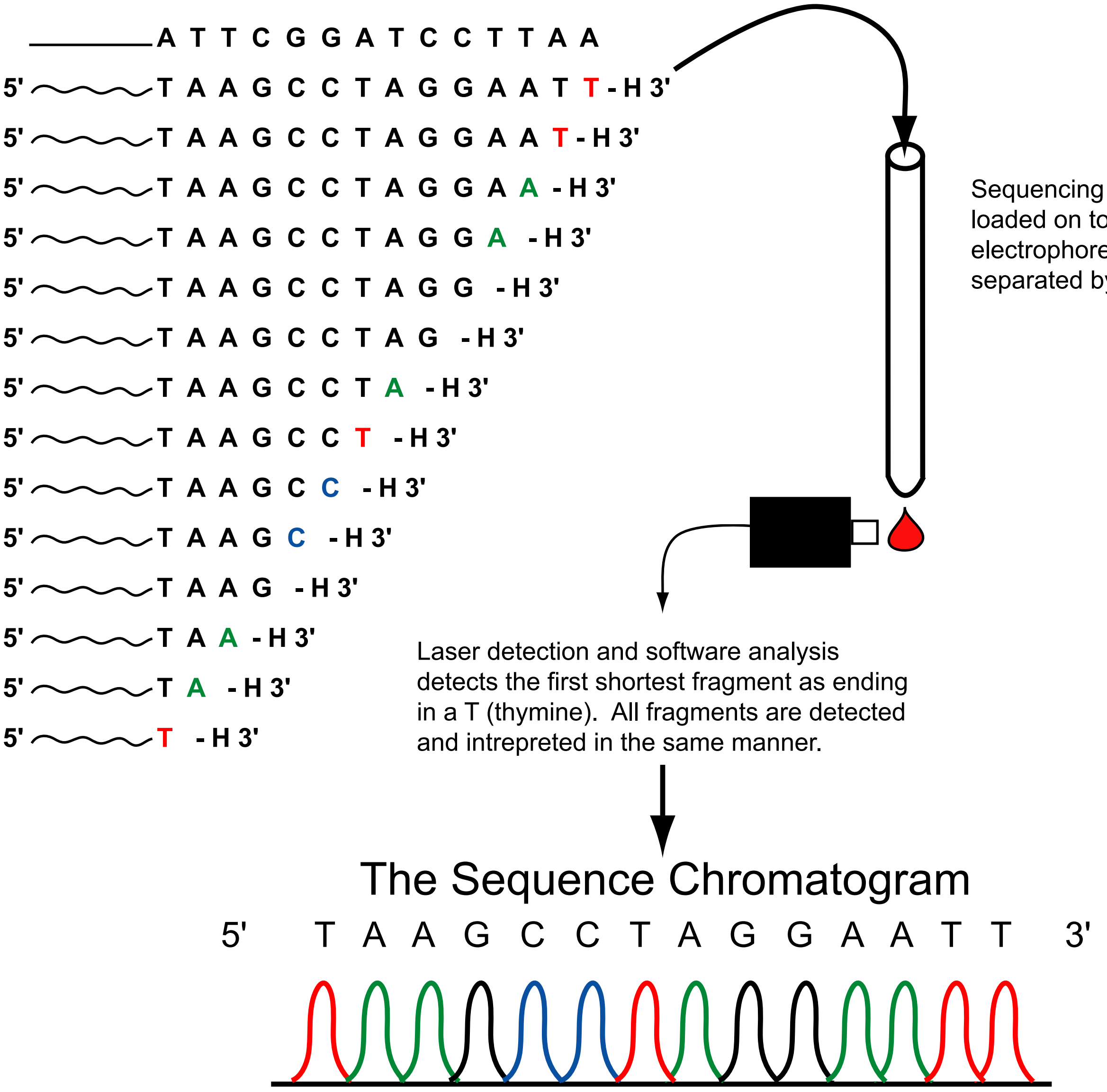
Figure 8. The chain termination (Sanger) DNA sequencing technique.

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


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 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'
5' T A A G C C T A G G A A T T 3'

Sequencing products are loaded on to a capillary electrophoresis unit and separated by size.

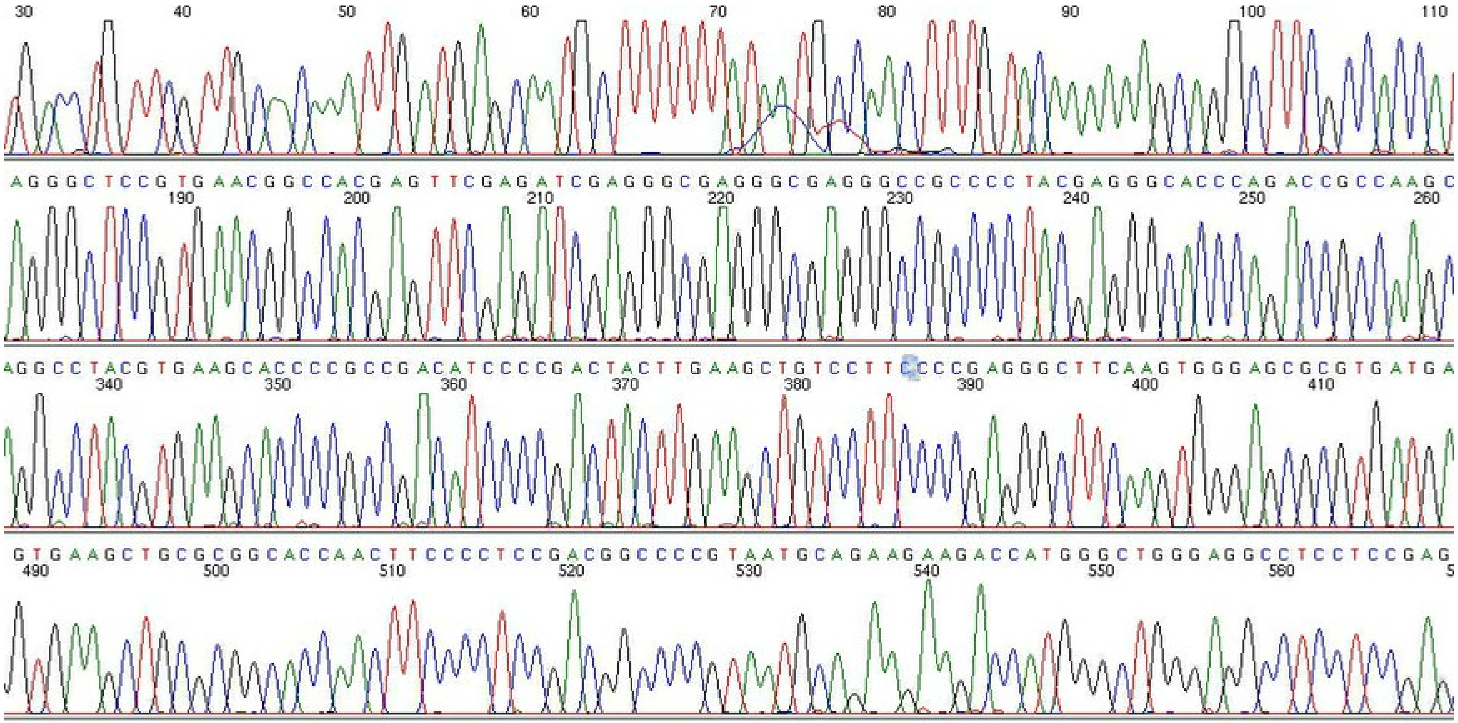
Laser detection and software analysis detects the first shortest fragment as ending in a T (thymine). All fragments are detected and intrepreted in the same manner.

The Sequence Chromatogram

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

Output from Automated DNA Sequencer

Photo: John J. Cardamone, Jr.



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

What Was Needed for All New Approaches

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

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***

Completeness

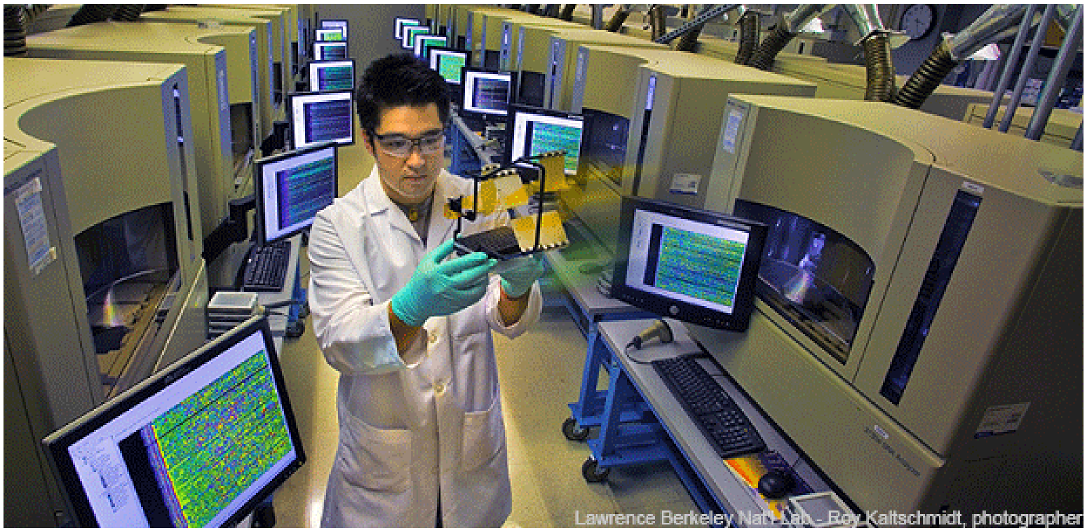
- Read length issue
  - Sanger technology with capillary detection
  - 500-700 nt
    - Allows for assembly into
      - Contigs
      - Supercontigs
- Emerging technologies
  - Length requirement
  - Must be long enough to align accurately
    - 25-100 nt read length
      - Sufficient for resequencing 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 = 150bp**
***PacBio = 20-60 kb.**

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

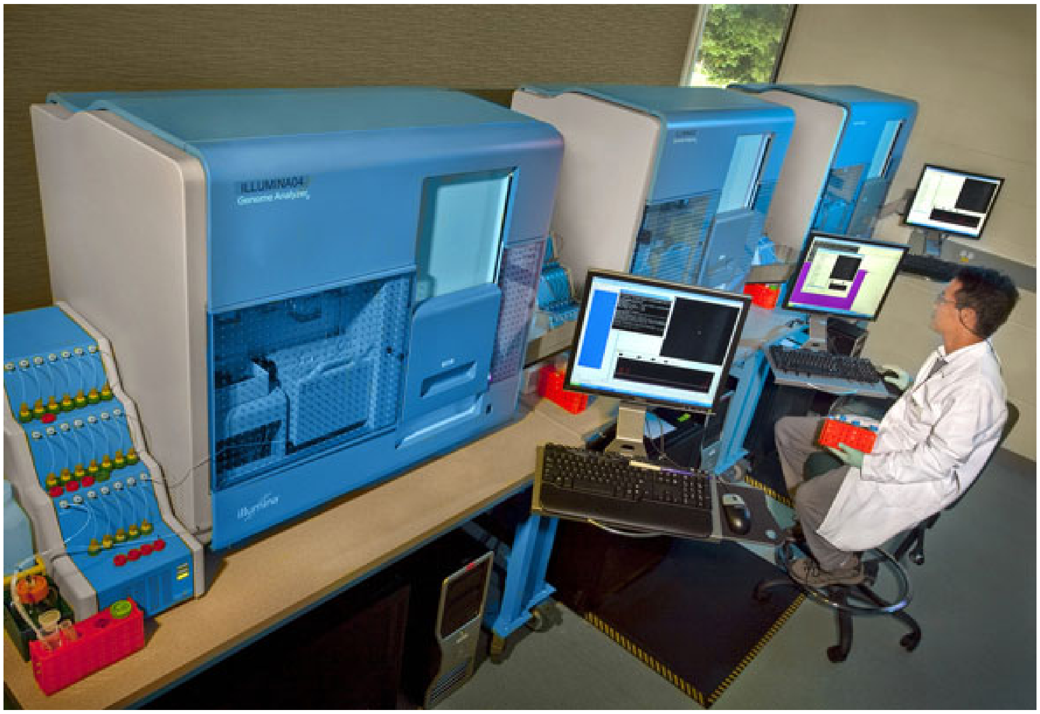
How Large Scale Sequencing Has Changed Over Time From a Centers Perspective

Then: DOE/JGI Sanger Sequencing Equipment Room



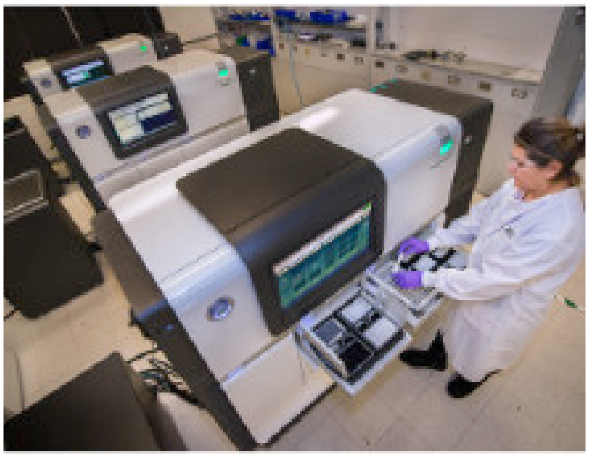
Lawrence Berkeley Nat'l Lab - Roy Kaltschmidt, photographer

Recently: DOE/JGI Illumina GAII Equipment Room



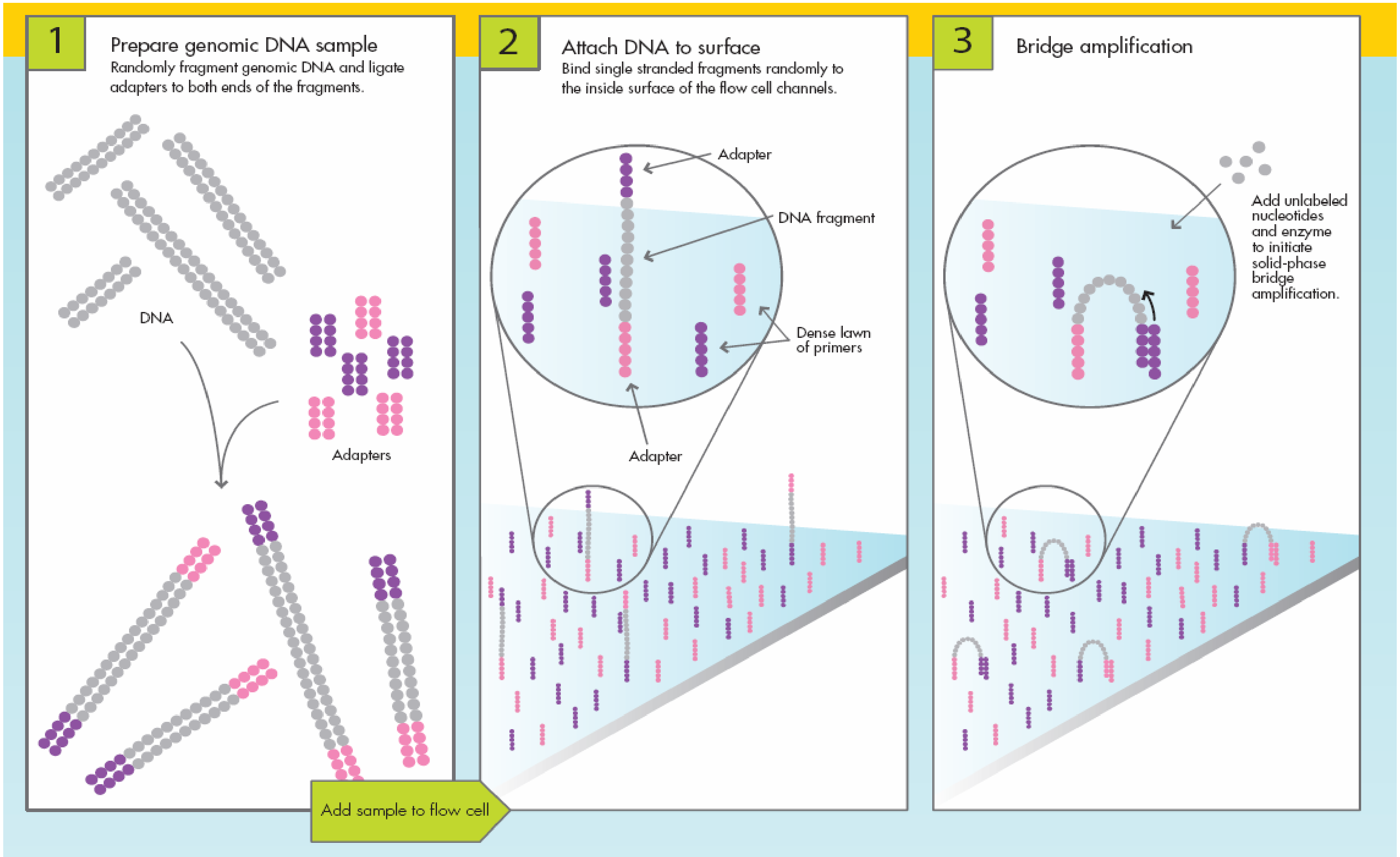
Now: DOE/JGI Illumina HiSEQ Equipment Room

Now: DOE/JGI PacBIO Equipment Room



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

Illumina Sequencing by Synthesis Technology



1 Prepare genomic DNA sample

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

2 Attach DNA to surface

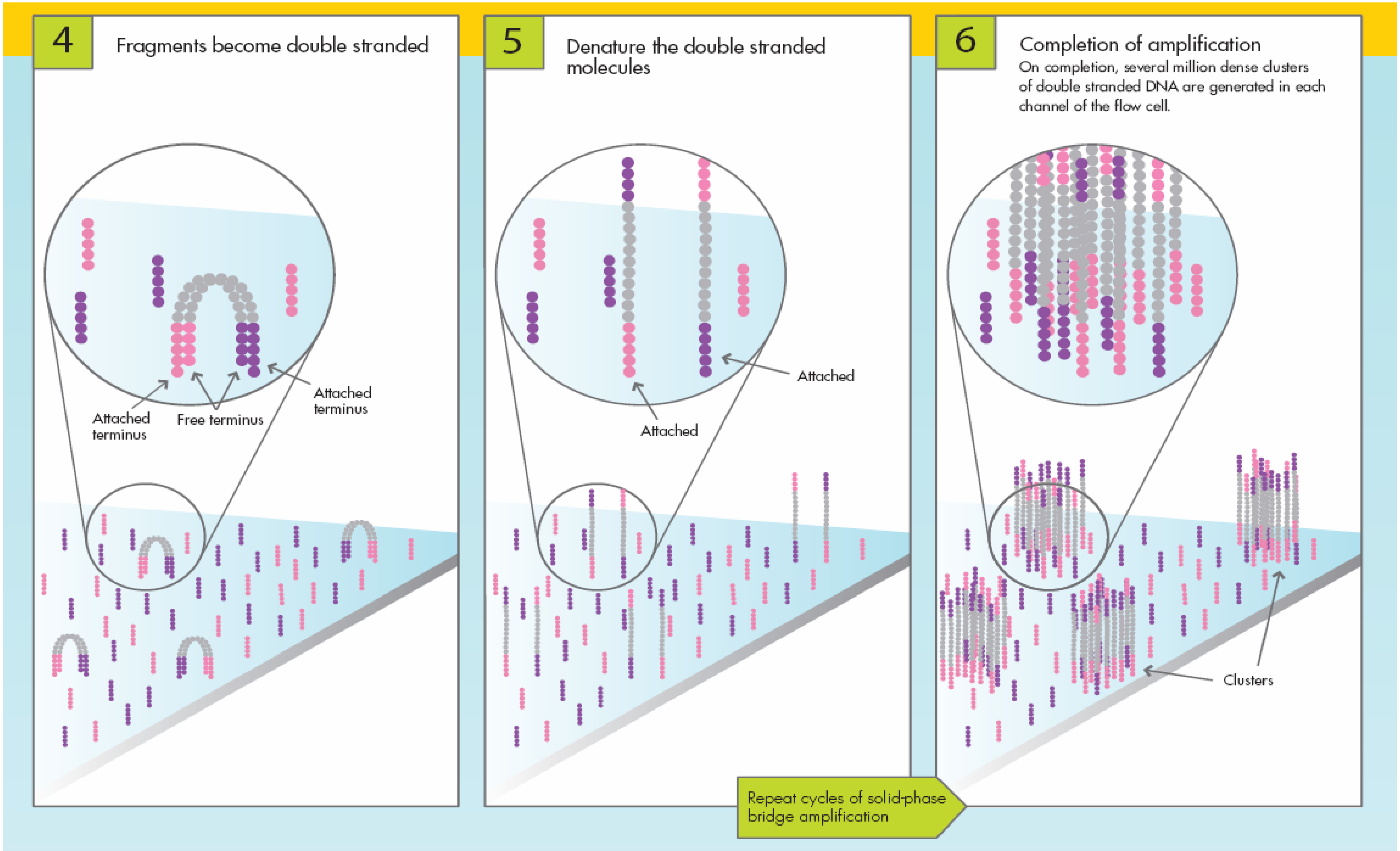
Bind single stranded fragments randomly to the inside surface of the flow cell channels.

3 Bridge amplification

DNA
Adapters
Add sample to flow cell
Adapter
DNA fragment
Dense lawn of primers
Adapter
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

A SINGLE STRAND molecule is bound to the flow cell.

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



4 Fragments become double stranded

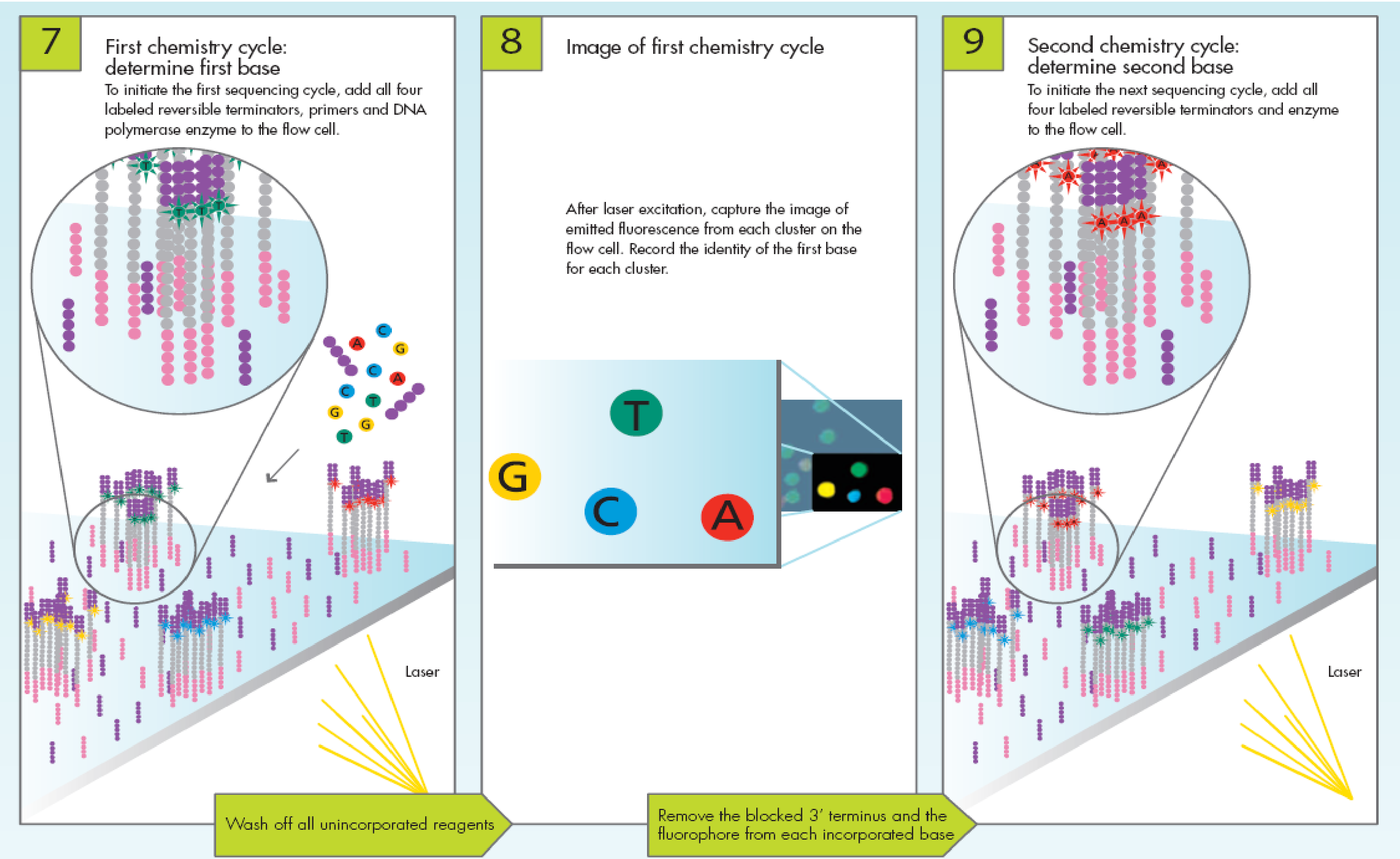
5 Denature the double stranded molecules

6 Completion of amplification

On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell.

Repeat cycles of solid-phase bridge amplification

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



7 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.

Wash off all unincorporated reagents

8 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.

Remove the blocked 3' terminus and the fluorophore from each incorporated base

9 Second chemistry cycle: determine second base

To initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

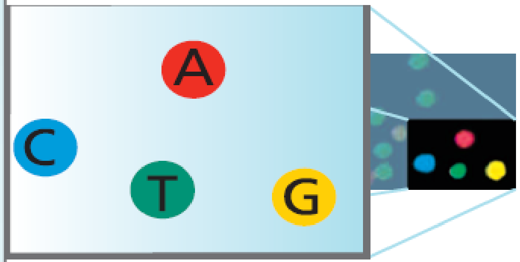
The first base is added to the template of each cluster with a blocker that prevents other bases from being added.

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.

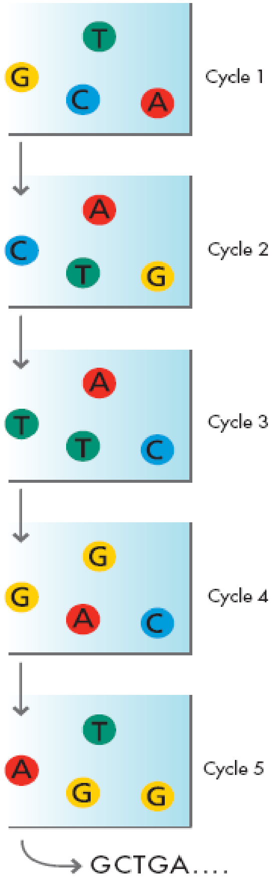
10 Image of second chemistry cycle is captured by the instrument

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

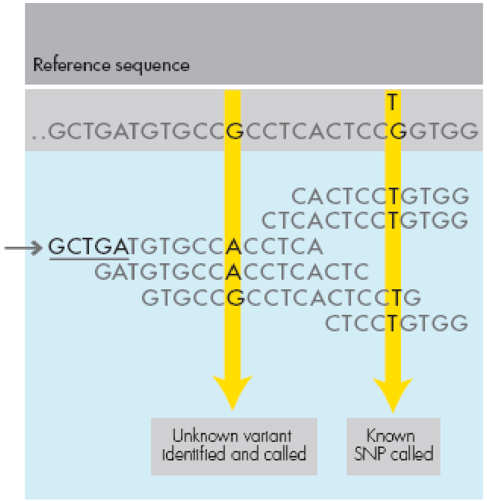


11 Sequence read over multiple chemistry cycles

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

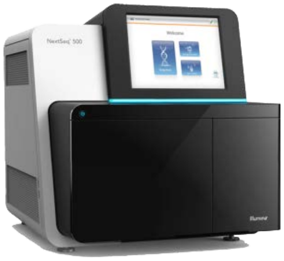
Illumina GAII [Maximum (Max) output: 25 gigabases (Gb)]

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

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: 6 Tb)

The NovaSeq 6000 is the principle machine used TODAY for HIGH THROUGHPUT sequencing!!!

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

Single Polymerase Real Time DNA Sequencing

Developed by Pacific Biosciences
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.
- Each sequencing plate contains ~3000 individual cells
    - Each holds only a single DNA molecule

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

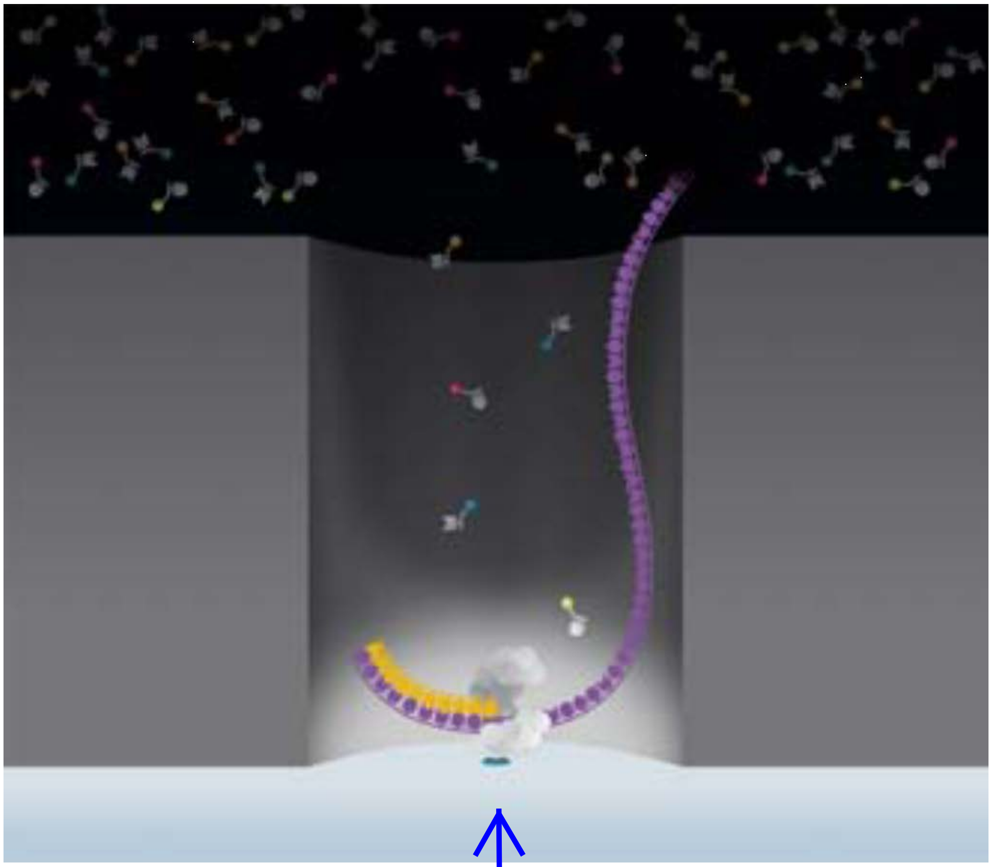
Detection and sequence determination

- 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

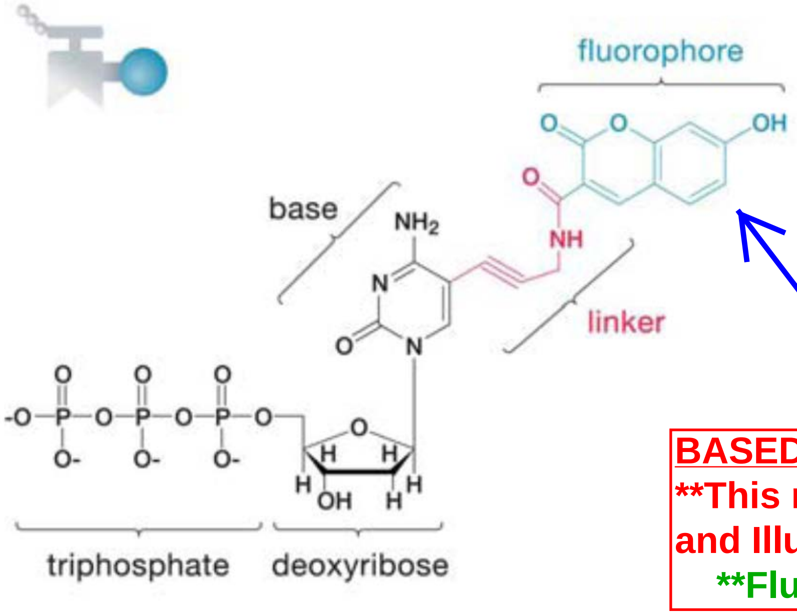


ZMW (Zero-mode waveguide) with Φ29 DNA polymerase and DNA template

ZMW is the sequencing reaction well.

A single DNA molecule HELD IN PLACE by the DNA polymerase enayme.

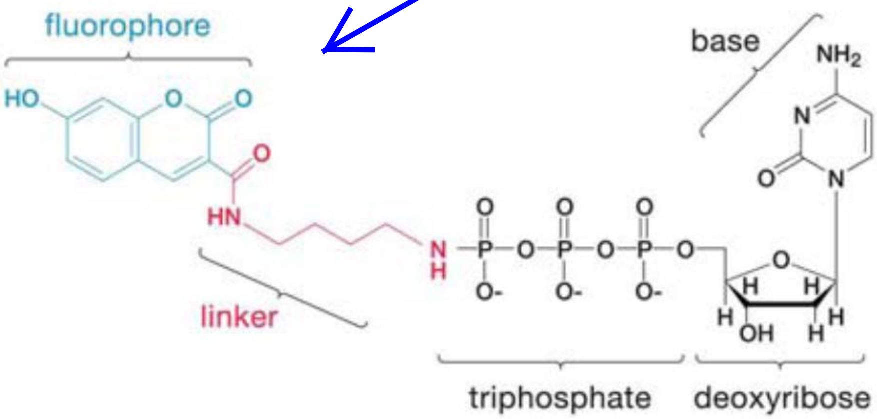
Base-labeled dNTP



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**



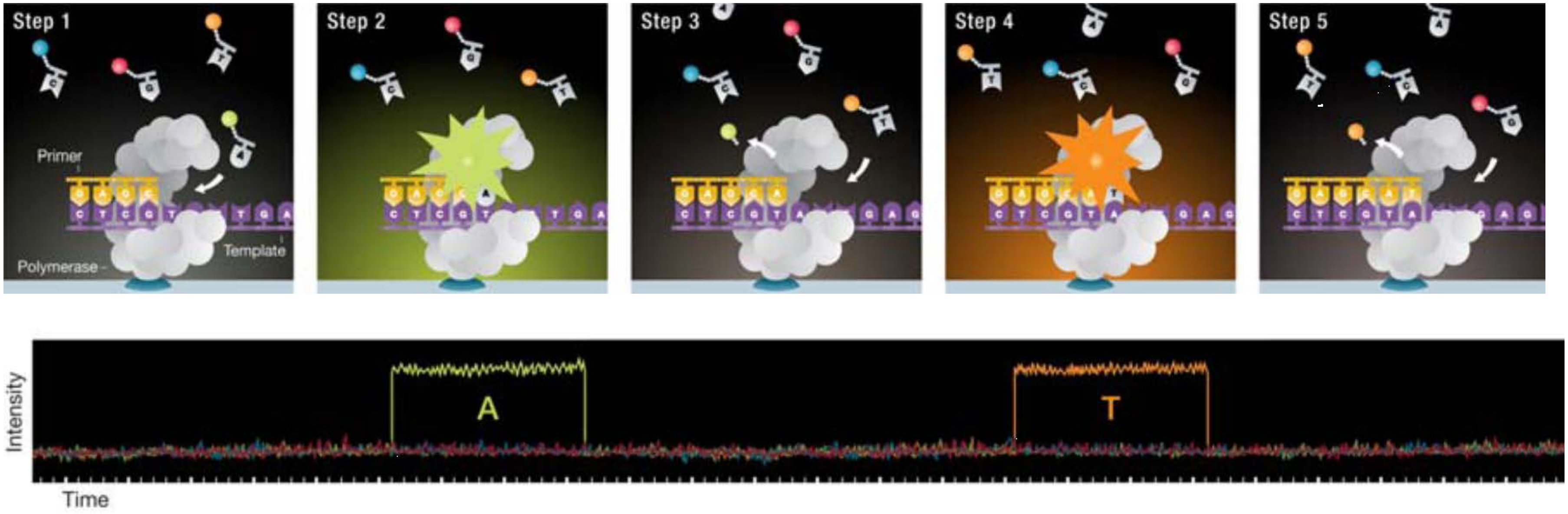
CORRECT Phospho-labeled dNTP enters THE ZMW

Fluorescent signal generated and captured in movie

NEW Phospho-labeled dNTP enters THE ZMW

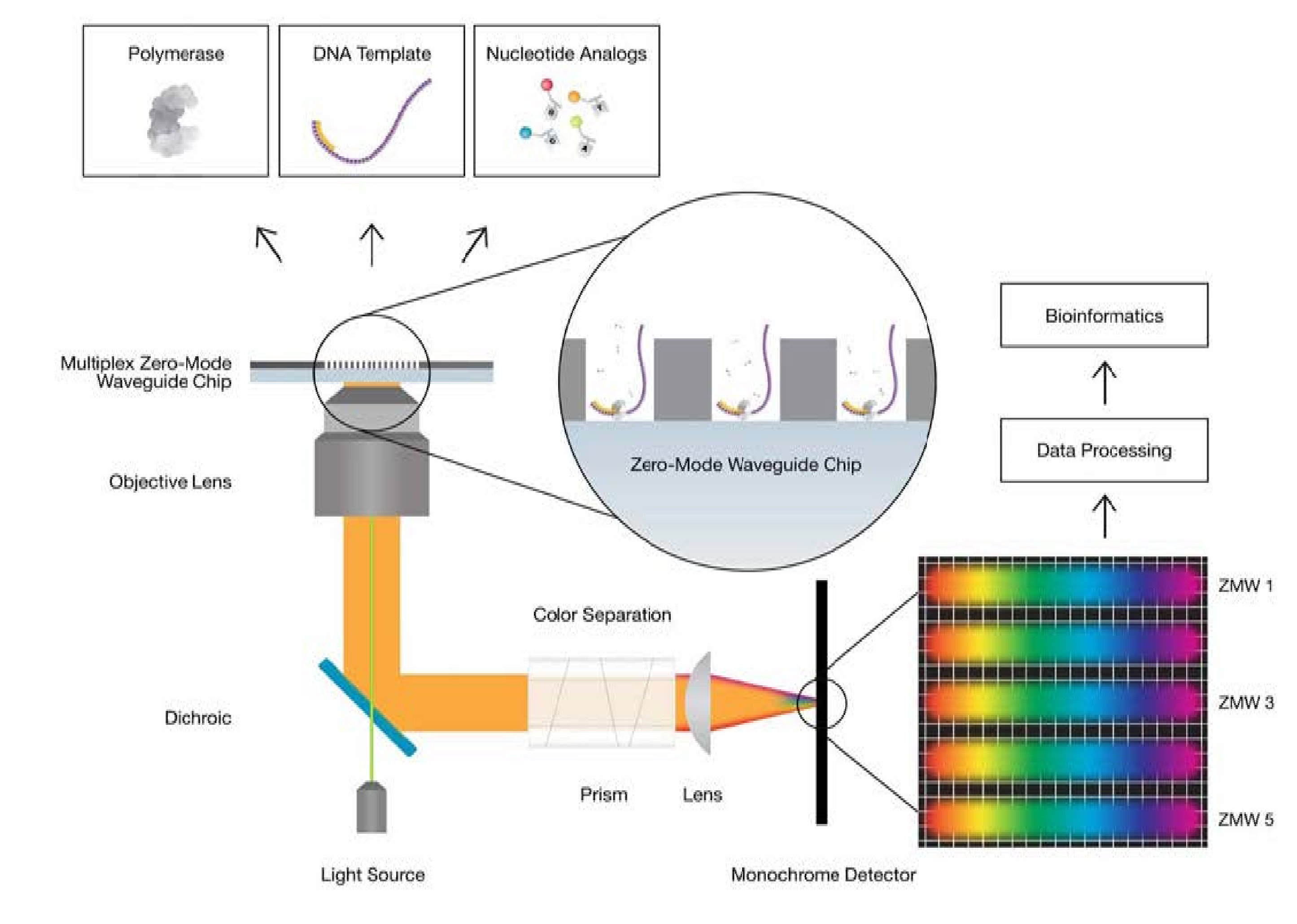
NEXT Fluorescent signal generated and captured in movie

Single Polymerase DNA Sequencing



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

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

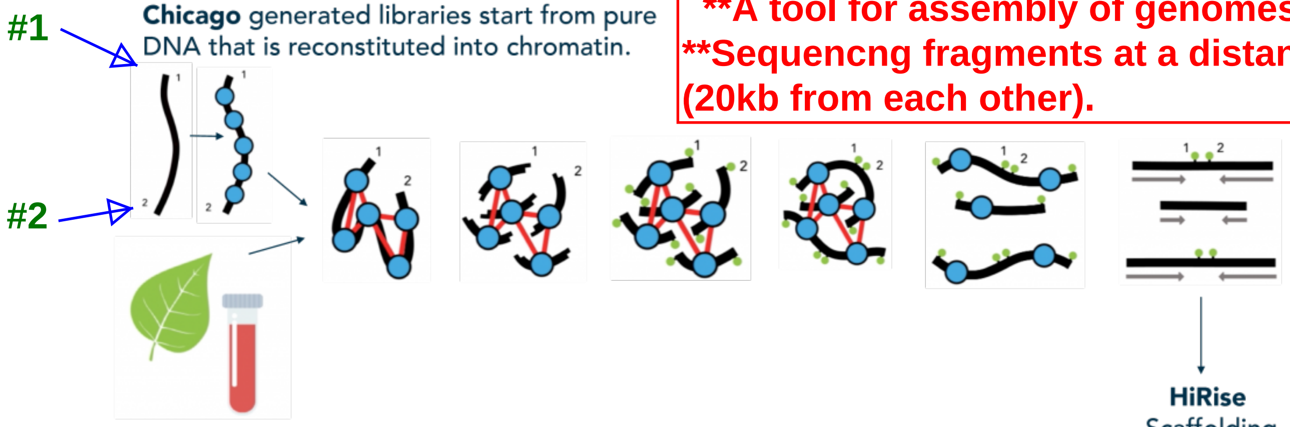
CCS IS PREFERRED BECAUSE OF READ ACCURACY!!!!

Goal: use proximal sequence data for assembly
**Region 1 and 2 brought together
**Sequences obtained
**Proximity of these sequences helps assembly

Dovetail Genomics Sequencing

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

Chicago generated libraries start from pure DNA that is reconstituted into chromatin.



Dovetail Hi-C generated libraries start from tissue or cell culture and endogenous chromatin is extracted after fixation.

Sequence 1 and 2 data collected

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

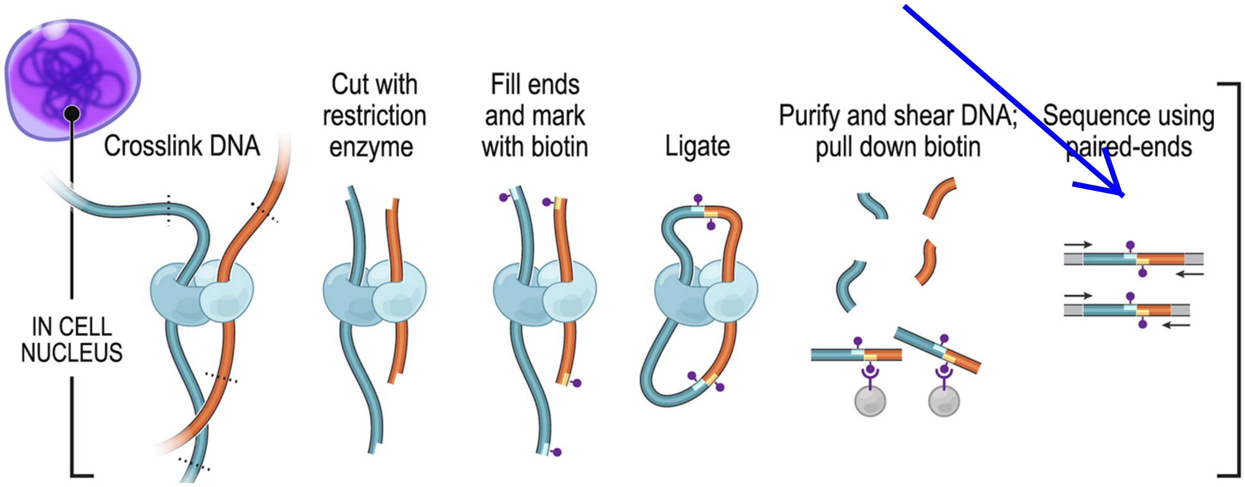
Hi-C linking

- 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

- 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.


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

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

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.

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**

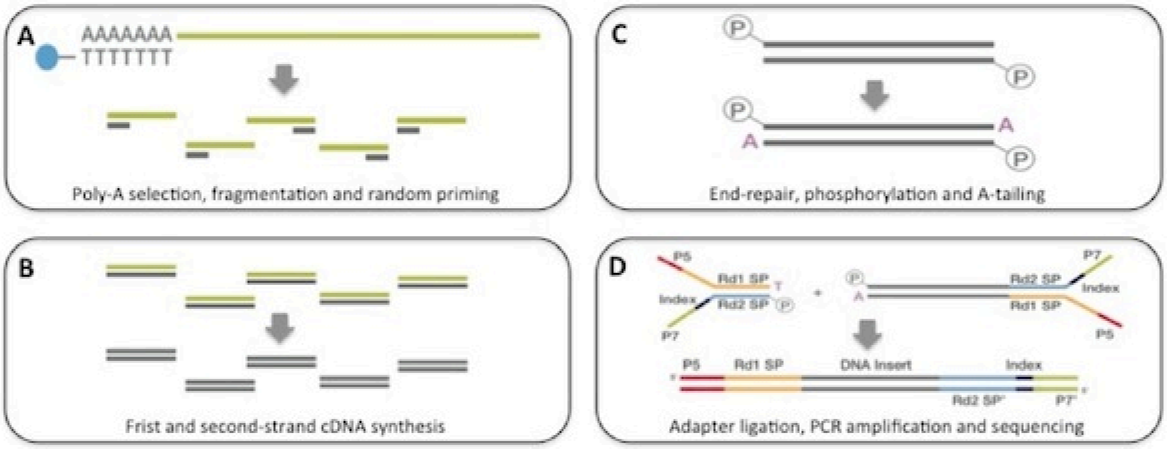
WHY?

****Genes are expressed in a temporal (time) and spatial (tissue) manner**

Sequencing the Expressed Portion of the Genome

- 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

Illumina Tru-Seq RNA-seq protocol



Poly-A selection, fragmentation and random priming

End-repair, phosphorylation and A-tailing

Frist and second-strand cDNA synthesis

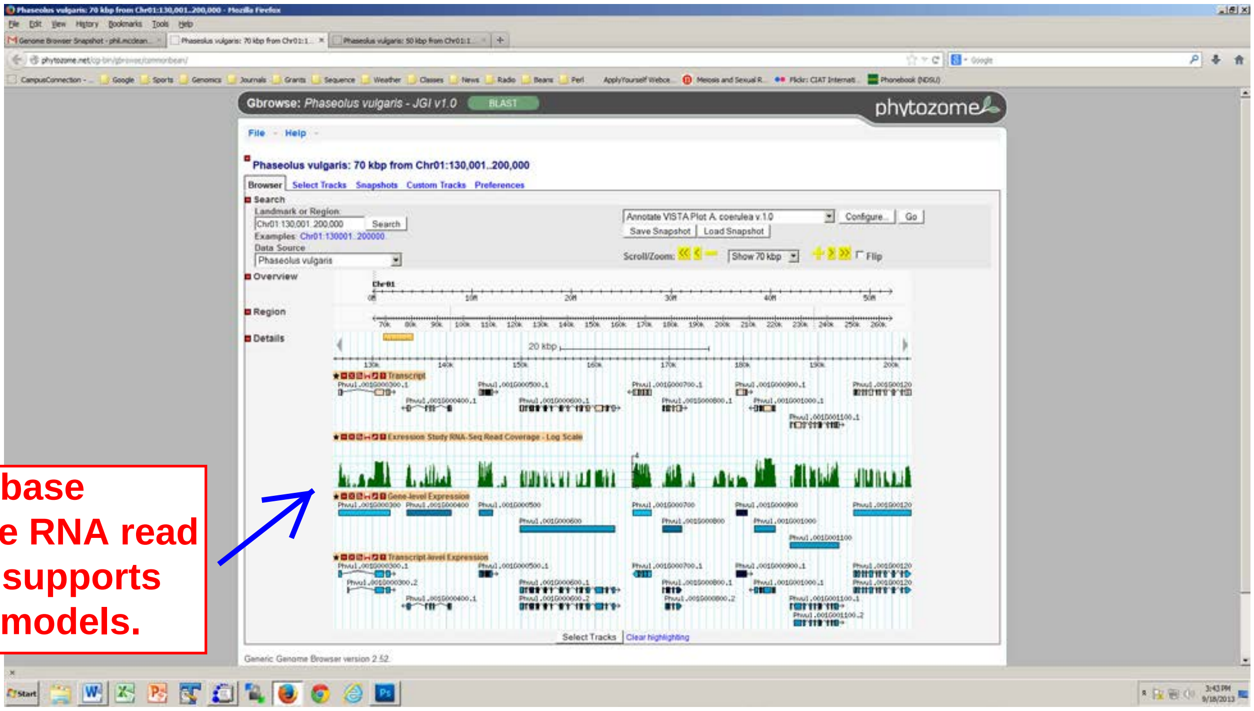
Adapter ligation, PCR amplification and sequencing

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

Aligning RNA-seq Data to the Reference Sequence



Gbrowse: *Phaseolus vulgaris - JGI v1.0*

Phaseolus vulgaris: 70 kbp from Chr01:130,001..200,000

Landmark or Region: Chr01:130,001..200,000

Examples: Chr01:130001..200000

Data Source: Phaseolus vulgaris

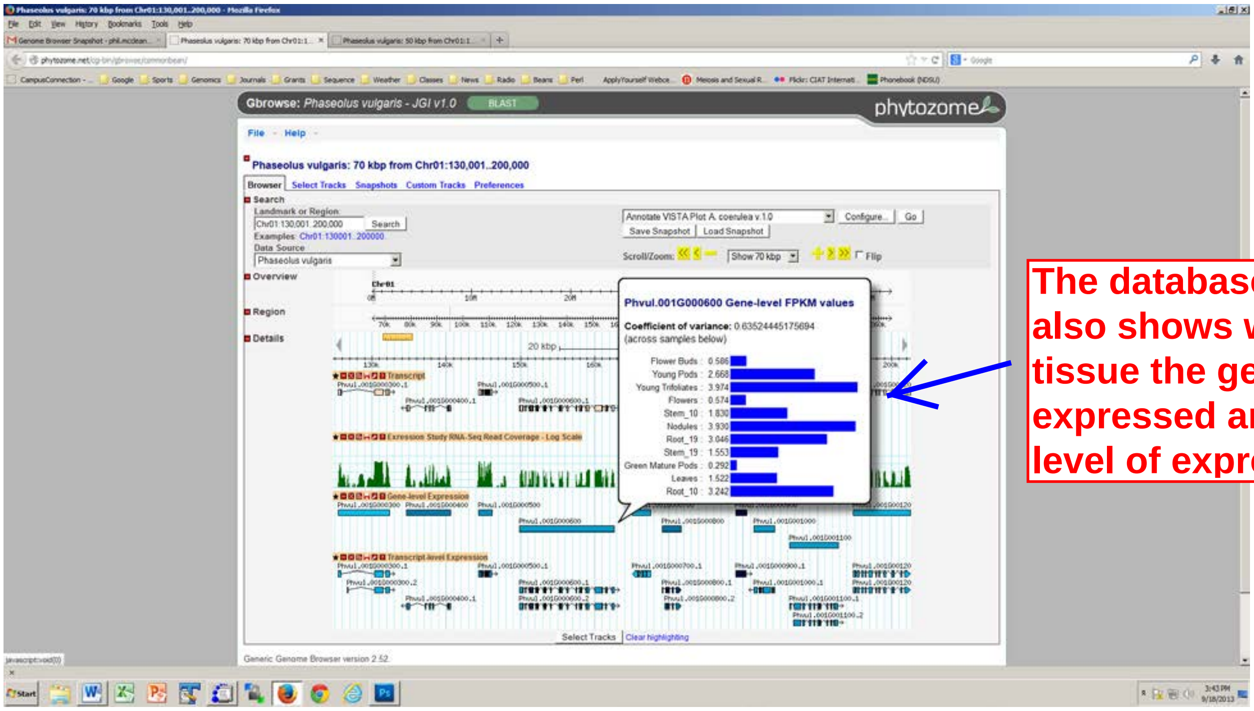
Annotate VISTA Plot A. coerulea v.1.0

Scroll/Zoom: Show 70 kbp

Expression Study RNA-Seq Read Coverage - Log Scale

Generic Genome Browser version 2.52

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



Phvul.001G000600 Gene-level FPKM values

Coefficient of variance: 0.63524445175694
(across samples below)

Flower Buds :	0.586
Young Pods :	2.668
Young Trifoliates :	3.974
Flowers :	0.574
Stem_10 :	1.830
Nodules :	3.930
Root_19 :	3.046
Stem_19 :	1.553
Green Mature Pods :	0.292
Leaves :	1.522
Root_10 :	3.242

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 fragments of different sizes.

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
  - 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

Modern Long Read PacBio Sequencing Genome Sequencing Approach

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

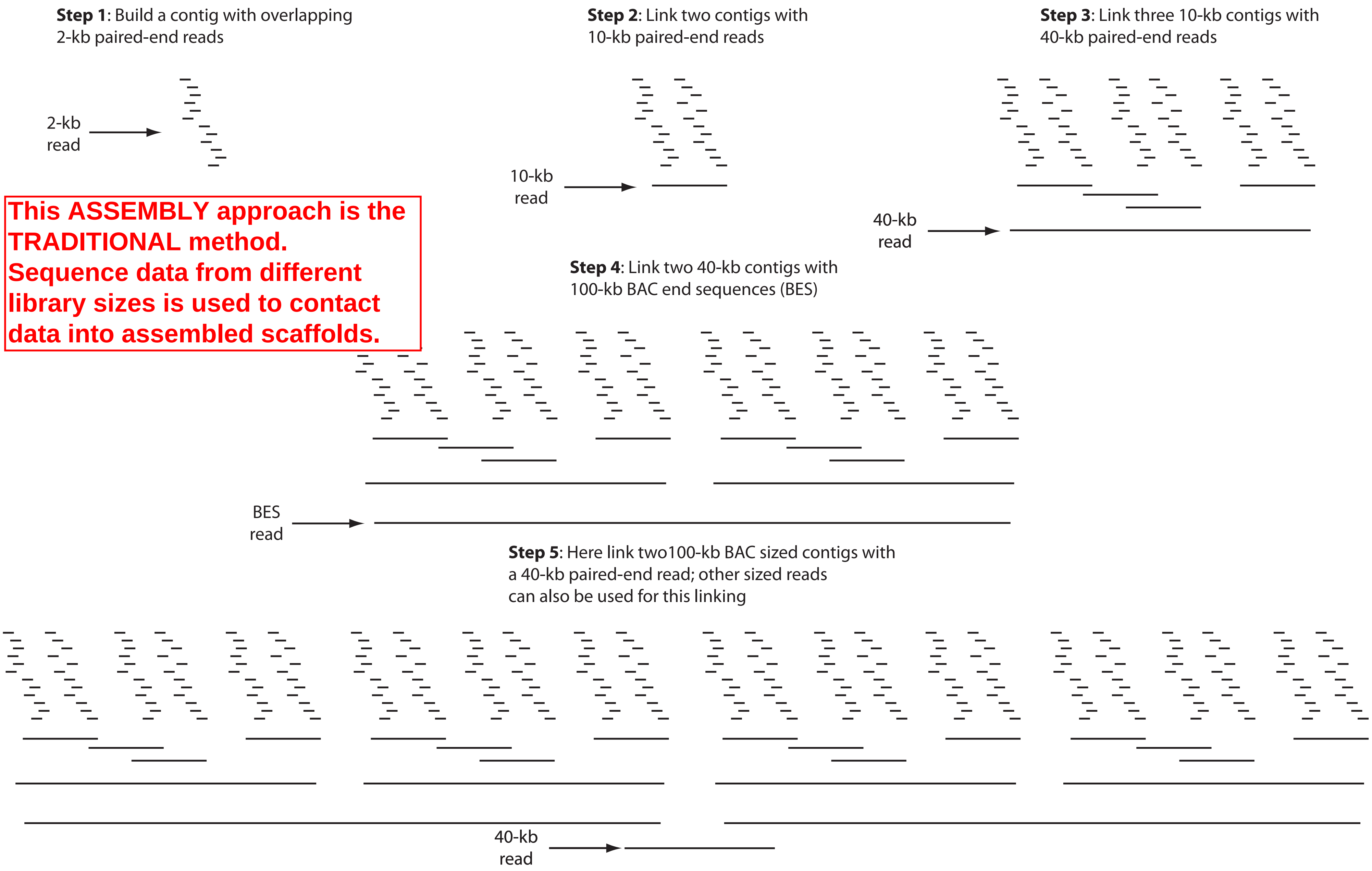
4. Assembly reads into contigs
   - Contigs MUCH longer than with Sanger sequencing

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

6. 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

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

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

This ASSEMBLY approach is the TRADITIONAL method. Sequence data from different library sizes is used to contact data into assembled scaffolds.

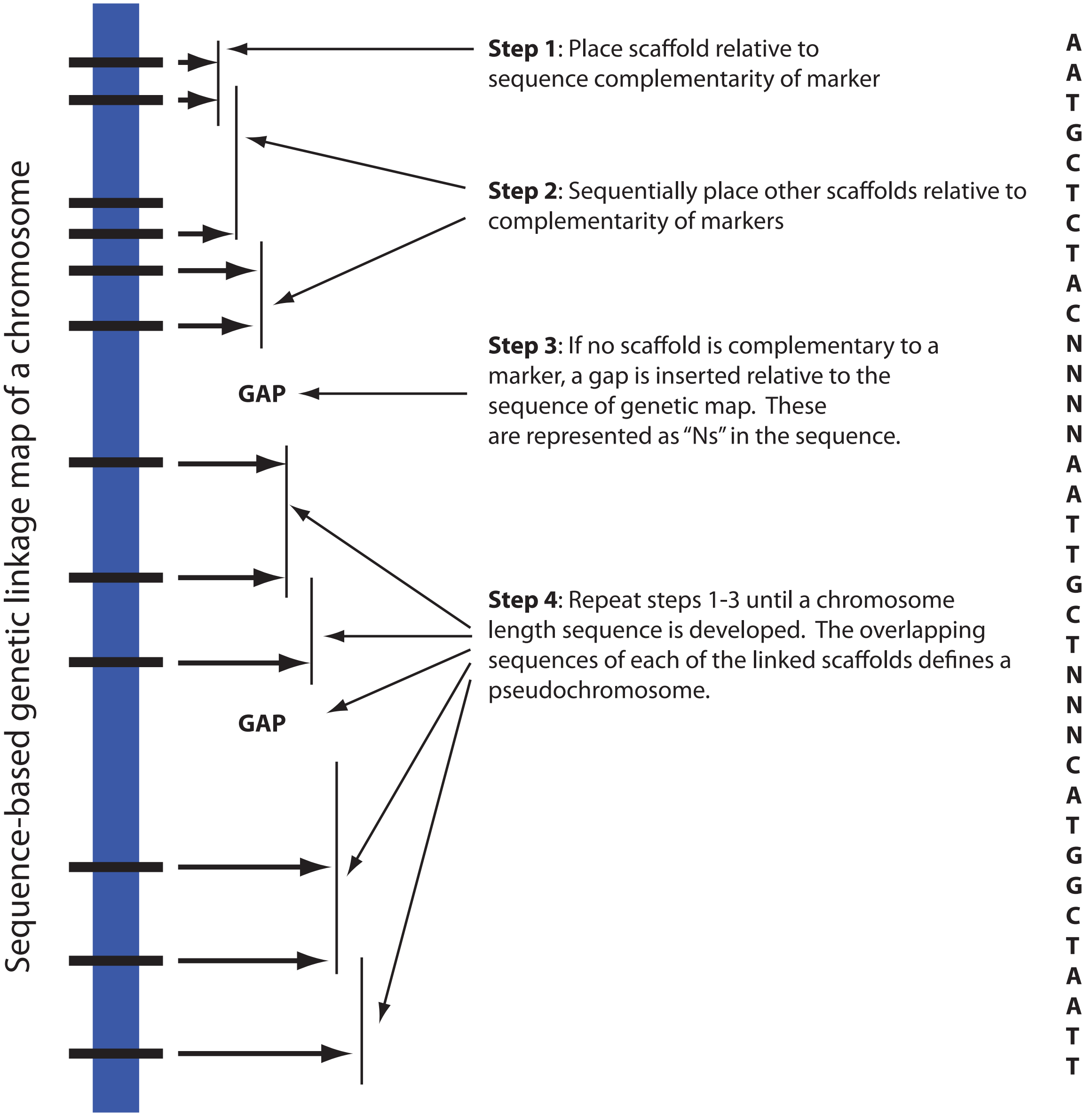
Step 4: Link two 40-kb contigs with 100-kb BAC end sequences (BES)

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

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


Sequence-based genetic linkage map of a chromosome
GAP
GAP
AATGCTCTACNNNNAATTGCTNNNCATGGCTAATT

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.

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.

This figure represents assembling PSEUDOCHROMOSOMES by linking scaffolds using marker locations. The sequence of the markers provides a accurate data for the organization of the scaffolds. REMEMBER that genetic data is still the most useful data for assembly. It is directly related to recombination events.

Phaseolus vulgaris

Summary Genome Sequencing and Assembly

Short read production information

- Sequence technology: Sanger, Roche 454, Illumina
- Number of libraries: 21 (15 paired, 6 unpaired)
- Total Reads: 49,214,786 (10,696,722 successful paired-end reads; 2.3% failed)
- Coverage: 21.02x total (18.64X linear, 3.38X paired-end)

Long read production information

- PacBio technology
- 83.2x sequence coverage
- Illumina data from short read project added to PacBio data

The data COMPARES the experimental methods used to develop a reference genome based on SHORT or LONG reads,

Estimated genome coverage from Kew Gardens C-value Database

- *P. vulgaris* = 0.6 picograms
  - 1 pg = 978 megabases
    - *P. vulgaris* **= 586.8 Mb**

Coverage

- Short read
  - 521.1 Mb/586.8 Mb = **88.8% coverage**
- Long read
  - 537.2/586.8 Mb = 91.5% **coverage**

IMPORTANT: Long read genomes provide better genome coverage.

Summary information	**Short read**	**Long read**
Main genome scaffold total	708	478
Main genome contig total	41,391	1,044
Main genome scaffold sequence total	521.1 Mb	537.2 Mb
Main genome contig sequence total	472.5 Mb (9.3% gap)	531.6 Mb (1.1% gap)
Main genome scaffold N50/L50	5/50.4 Mb	5/49.7 Mb
Main genome contig N50/L50	3,273/39.5 Mb	73/1.9 Mb
Number of scaffolds > 50 Kb	28	87
% main genome in scaffolds >50 Kb	99.3%	99.1%

Best STATISTICS to compare quality of genomes.

Loci

27,433 total loci containing 36,995 protein-coding transcripts

Alternative Transcripts

9,562 total alternatively spliced transcripts

Probably an UNDERESTIMATE; More tissues are needed for a better estimate.

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 **<u>NUMBER</u>**

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 **<u>LENGTH</u>**

The LARGER the number the BETTER the genome.

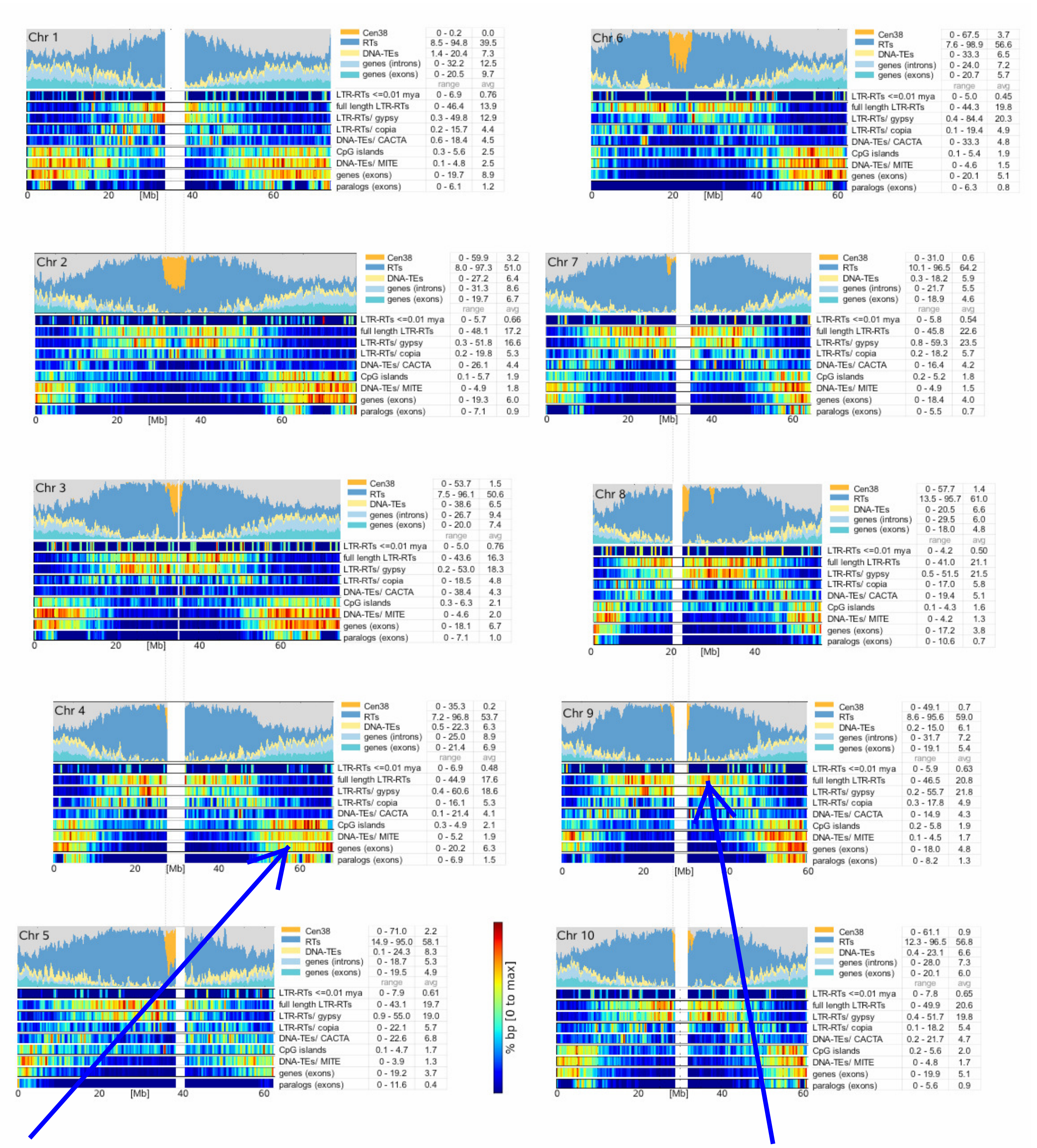
IMPORTANT NOTE

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

Species name	Common name	Genotype	Year	Publication	Technical method	# Chrom	Est. genome size/assembled size (Mb)	Repeat content (%)	Chrom size range (Mb)	# genes/ transcripts	Contig N50/L50 (#/kb)	Scaffold N50/L50 (#/kb)	Genome duplication history
Arabidopsis thaliana	Arabidopsis	Columbia	2000	Nature 408:796	HSS/S	5	125/135	20[1]	18-29	27,416/ 35,386			Eudicot 3x + Brassicaceae (2x+2x)
Oryza sativa	Rice	Nipponbare	2005	Nature 436:793	HSS/S	12	430/371	45[1]	23-43	39,049/ 49,061			Poales (2x+2x)
Populus trichocarpa	Poplar	Nisqually 1	2006	Science 313:1596	WGS/S	19	485/423	40[1]	11-36	41,335/ 73,013	??/126	??/3,100	Eudicot 3x + (2x)
Vitus vinifera	Grape	PN40024	2007	Nature 449:463	WGS/S	19	475/487	22[1]	10-22	/ 26,346	??/126	??/2,065	Eudicot 3x
Carica papaya	Papaya	Sunup	2008	Nature 452:991	WGS/S	9	372/370	52		27,332/ 27,996	??/11	??/1,000	Eudicot 3x
Sorghum bicolor	Sorghum	BTx623	2009	Nature 457:551	WGS/S	10	818/727[2]	63[1]	50-70	33,032/ 39,441	958/195	6/62,400	Poales (2x+2x)
Zea mays	Maize	B73	2009	Science 326:1112	HSS/S	10	/3,234	84	150-301	39,475/ 137,208			Poales (2x+2x) + (2x)
Cucumis sativus	Cucumber	9930	2009	Nat Genet 41:1275	WGS/S,I	7	??/244	22[1]		21,491/ 32,528	??/227	??/1,140	Eudicot 3x
Glycine max	Soybean	Williams 82	2010	Nature 463:178	WGS/S	20	1115/978	57	37-62	56,044/ 88,647	1,492/189	10/47,800	Eudicot 3x + Legume 2x + (2x)
B. distachyon	Brachypodium	Bd21	2010	Nature 463:763	WGS/S	5	272/275	28	25-75	26,552/ 31,029	252/348	3/59,300	Poales (2x+2x)
Ricinus communis	Castor bean	Hale	2010	Nat Biotech 28:951	WGS/S, 454	10	320/326	~50		31,237/??	??/21	??/497	Eudicot 3x
Malus x domestica	Apple	Golden Delcious	2010	Nat Genet 42:833	WGS/S	17	742/604	36	21-47	63,538/ 63,541	16,171/13	102/1,542	Eudicot 3x + Rosaceae 2x
Jatropha curcas	Jatropha		2010	DNA Res 18:65	WGS/S		380/285	37		40,929/??	??/4		
Theobroma cacao	Cocao	B97-61/B2	2011	Nat Genet 43:101	WGS/S, 454, I	10	430/362	24	12-31	29,452/ 44,405		??/5,624	Eudicot 3x
Fragaria vesca	Strawberry	H4x4	2011	Nat Genet 43:109	WGS/S, 454, I, So	7	240/220	23		32,831/??		??/1,300	Eudicot 3x
Arabidopsis lyrata	Lyrata	MN47	2011	Nat Genet 43:476	WGS/S	8	??/207	30	19-33	32,670/??	1,309/5,200		Eudicot 3x + Brassicaceae (2x+2x)
Phoenix dactylifera	Date palm	Khalas	2011	Nat Biotech 29:521	WGS/I	18	658/381	29		28,890/??	??/6	??/30	
Solanum tuberosum	Potato	DM1-3 516 R44	2011	Nature 475:189	WGS/S, 454, I, So	12	844/727	62		35,119/ 51,472	6,446/31	121/1,782	Eudicot 3x + Solanaceae 3x
Thellungiella parvula	Thellungiella		2011	Nat Genet 43:913	WGS/454, I	7	160/137	8		30,419/??		8/5,290	
Cucumis sativus	Cucumber	B10	2011	PLoS ONE 6:e22728	WGS/S, 454	7	??/323			26,587/??	??/23	??/323	Eudicot 3x
Brassica rapa	Cappage	Chiifu-401-42	2011	Nat Genet 43:1035	WGS/I	10	??/283	40		41,174/??	2,778/27	39/1,971	Brassicaceae 2x + (2x)
Cajanus cajan	Pigeon pea	ICPL 87119		Nat Biotech 30:83	WGS/S, I	11	808/606	52	10-48	40,071	7815/23	380/516	Eudicot 3x + Legume 2x
Medicago truncatula	Medicago		2011	Nature 480:520	WGS/S, 454, I	8	454/384		35-57	44,135/ 45,888		53/1270	Eudicot 3x + Legume 2x
Setaria italica	Foxtail millet	Yugu 1	2012	Nat Biotech 30:555	WGS/S	9	451/406	40	24-48	35,471/ 40,599	982/126	4/47,300	

Species name	**Common name**	**Genotype**	**Year**	**Publication**	**Technical method**	**# Chrom**	**Est. genome size/assembled size (Mb)**	**Repeat content (%)**	**Chrom size range (Mb)**	**# genes/ transcripts**	**Contig N50/L50 (#/kb)**	**Scaffold N50/L50 (#/kb)**	**Genome duplication history**
Solanum lycopersicon	Tomato	Heinz 1706	2012	Nature 485:635	WGS/S,So	12	900/760	63	45-65	34,727/??			Eudicot 3x + Solanaceae 3x
Linum usitatissimum	Flax	CDC Bethune	2012	Pl Journal 72:461	WGS/I	15	373/318	24		43,484	4,427/20	132/693	Eudicot 3x + (2x)
Musa acuminata	Banana	DH-Pahang, ITC1511	2012	Nature 488:213	WGS/S, 454, I	11	??/523	44	22-35	36,542	/43	/1,311	Zingiberales 2x + (2x + 2x)
Gossypium raimondii	Cotton (B genome diploid)		2012	Nat Genet 44:1098	WGS/I	13	775/567	57	25-69	40,976/??	4,918/45	2,284/95	Eudicot 3x + Gossypium 2x
Azadirachta indica	Neem	Local tree	2012	BMC Genomics 13:464	WGS/I		??/364	13		20,169/??	??/0.7	??/452	
Gossypium raimondii	Cotton (D genome diploid)		2012	Nature 492:423	WGS/S, 454, I	13	880/738	61	35-70	37,505/ 77,267	1596/136	6/62,200	Eudicot 3x + Gossypium 2x
Prunus mume	Chinese plum	2 genotypes	2012	Nature Communications 3:1318	WGS/I	8	??/237	45		31,390/??	2009/32	120/578	
Pyrus bretschneideri	Pear		2013	Genome Research	HSS+WGS/I	17	528/512	53	11-43	42,812/??	??/36	??698	Eudicot 3x + Rosaceae 2x
Cirtullus lanatus	Watermelon	97103	2013	Nat Genet 45:51	WGS/I	11	425/354	45	24-34	24,828/??	??/26	??/2380	Eudicot 3x
Morus notabilis	Mulberry		2013	Nature Communications 4:2445	WGS/I	7	357/330	47		29,338/??	2,638/34	245/390	Eudicot 3x
Phaseolus vulgaris	Common bean	G19833	2014	Nat Genet (in press)	WGS/S, 454,I	11	587/521	45	32-60	27,197/ 31,688	3,273/40	5/50	Eudicot 3x + Legume 2x

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

Genome Resequencing

Goal

- Discover variation in a population

How?

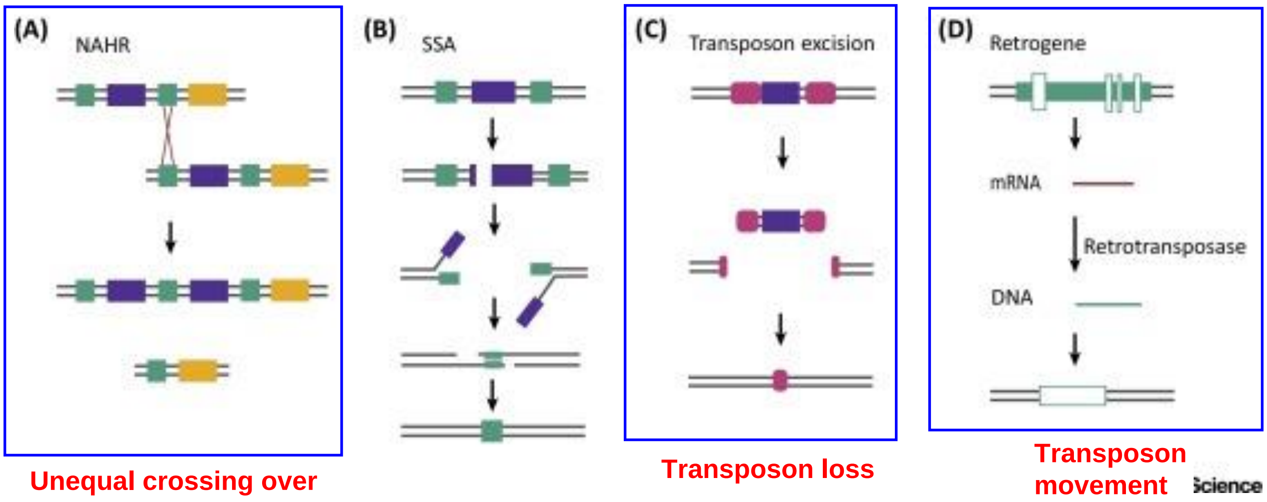
- Resequence many individuals
- 10x – 40x, depending on the goal

Types of 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

Lye and Purugannan (2019) Trends in Plant Science 24:352

Examples of Copy Number Variants Types



Unequal crossing over

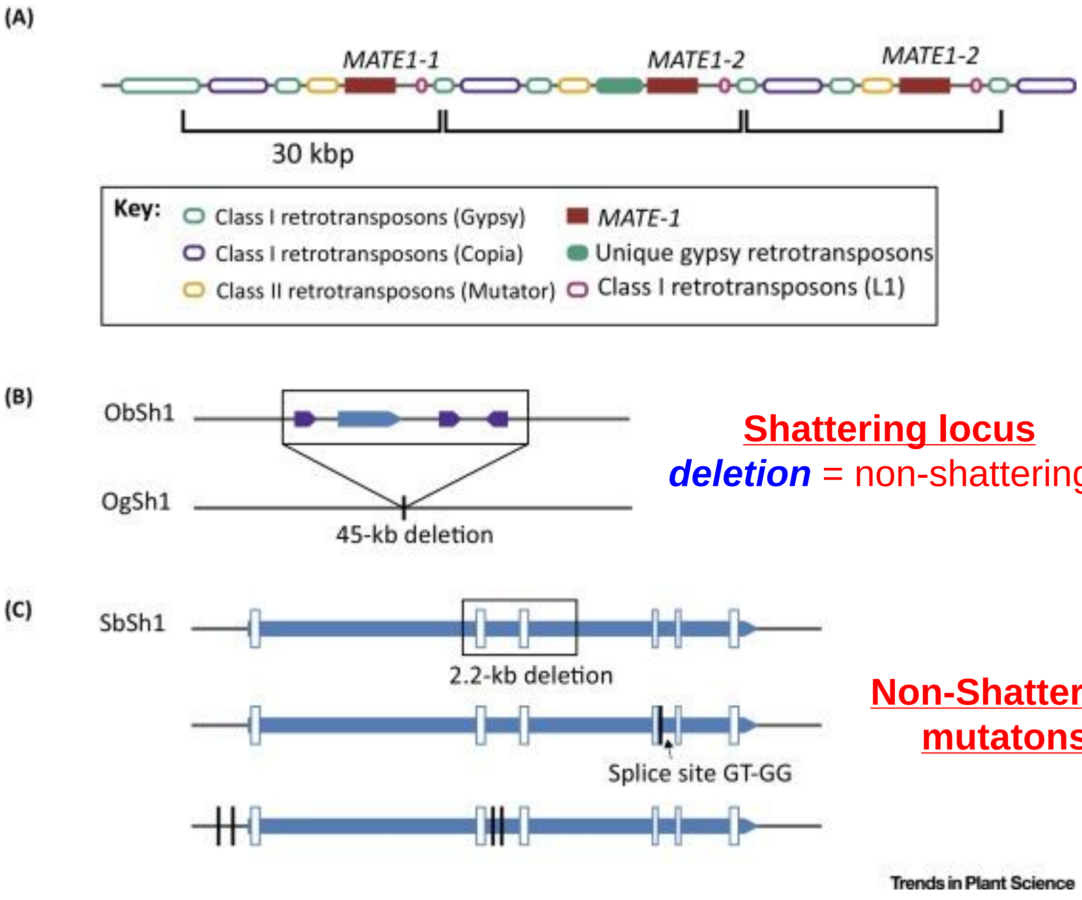
Transposon loss

Transposon movement

Figure 1. Mechanisms of Copy Number Variation (CNV) Formation. **(A)** ***Nonallelic homologous recombination*** (NAHR; unequal crossing over): during a recombination-based double-strand break (DSB) repair, a direct repeat, represented in green, is used as homology and incorrectly pairs during crossing over, this causes a reciprocal deletion and duplication of sequence between the repeats (purple). In this scenario, the resulting CNV break point is flanked by tracts of homologous sequence. **(B)** ***Single-strand annealing (SSA)***. During double-strand break repair, the 5′ stands are resected to expose complimentary sequences either side of the break (green). Although this is similar to the microhomology-mediated end joining repair pathway, SSA requires longer tracts of homology, typically >30 base pairs (bp). This can result in significant deletions of intervening sequence (purple**). (C) Transposon excision**. Transposons (pink ovals) flank a unique sequence (purple). Both transposons excise simultaneously, removing the unique sequence with them, and can result in a deletion. **(D) Retro-gene formation**. Retrotransposon activity causes insertion of a coding sequence into the genome (gene is shown in green with white boxes representing introns). mRNA (red) from the gene is reverse transcribed to DNA. This DNA can be occasionally inserted into the genome and become a retrogene, a copy of the original gene lacking introns (green box). These genes can be inserted into another gene, creating a chimeric gene, or become under control of different promoter sequences and take on a new expression regime.

Duplicated region

Examples of CNVs that Change Phenotype


(A)
MATE1-1
MATE1-2
MATE1-2
30 kbp
Key:
Class I retrotransposons (Gypsy)
Class I retrotransposons (Copia)
Class II retrotransposons (Mutator)
MATE-1
Unique gypsy retrotransposons
Class I retrotransposons (L1)
(B)
ObSh1
OgSh1
45-kb deletion
(C)
SbSh1
2.2-kb deletion
Splice site GT-GG
Trends in Plant Science

Shattering locus

deletion = non-shattering

Non-Shattering mutatons

Figure 2. Examples of Copy Number Variations (CNVs). ***(A)* Multidrug and toxic compound extrusion 1 *(*MATE1*) locus in maize.*** A 30-kb region containing transposable elements and the *MATE1* gene is triplicated in tandem. The filled red boxes represent each copy of *MATE1.* One copy contains an additional unique gypsy retrotransposon (filled teal). The outlined boxes represent other classes of retrotransposons that are part of the duplicated region. ***(B) The* shattering1 *(*Sh1*) locus in* Oryza barthii *(*ObSh1*) and* Oryza glaberrima *(*OgSh1*).*** A 45-kb region including the *Sh1* gene (blue), a YABBY transcription factor, and three additional genes (purple) is deleted in domesticated *O. glaberrima* relative to *O. barthii*. This deletion is polymorphic in domesticated populations. ***(C) Three haplotypes of* Sh1 *locus (7758 bp) in nonshattering* Sorghum bicolor *(Sb) relative to the wild, shattering,* Sorghum virgatum *sequence.*** From top to bottom: a 2.2-kb deletion including two exons, a SNP polymorphism at a splice site, and four SNP variants, two upstream of the transcription start site and two in an intron. Each of these haplotypes is present in nonshattering domesticated species, indicating that CNV is one of multiple mutations that may be causing the loss-of-function trait. Adapted from [67] (A), [102] (B), and [52] (C).