Real Problems in High Throughput Sequencing

The big picture

- Sequencing companies like Illumina, Biotage, Pyrosequencing, seek to develop technologies that make DNA sequencing cheaper and faster
- Tech battle very much like Intel vs AMD
  - The basic product of each company gets cheaper and cheaper as they make more and more
  - The “race to the $1000 genome” will simply be replaced by the “race to the $10 genome” later
- Long R&D cycles
- Lots of secret and semi-secret technology to keep a competitive edge

Real sequencing is fraught with errors

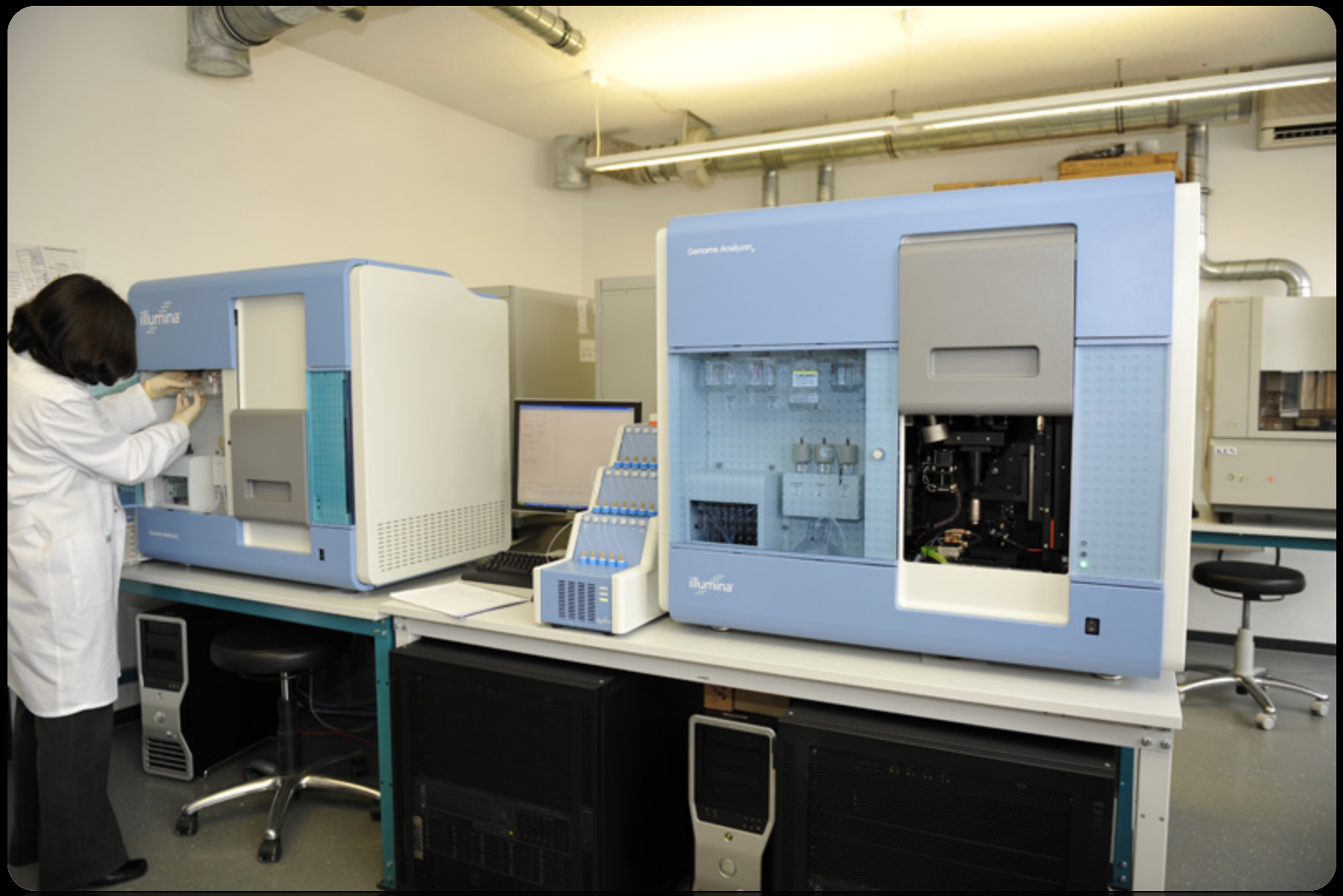
- Insufficient reads
  - In this lecture, you will see why new sequencing technologies will never lack reads.
- Read Errors
  - These errors happen all the time, but handling them requires very high coverage that would make Project 1 too hard.
  - In this lecture, we'll see how they happen, and what can be done to deal with them
- Dephasing
  - Sequencing on some reads runs ahead or behind other reads. We'll see what this is today. This is a hardware problem in Illumina sequencers that has generally been resolved

In this lecture

- We will see how these errors happen in the context of the operation of an illumina genome analysis platform

- We will talk about how these errors get dealt with in the algorithmic assembly phase

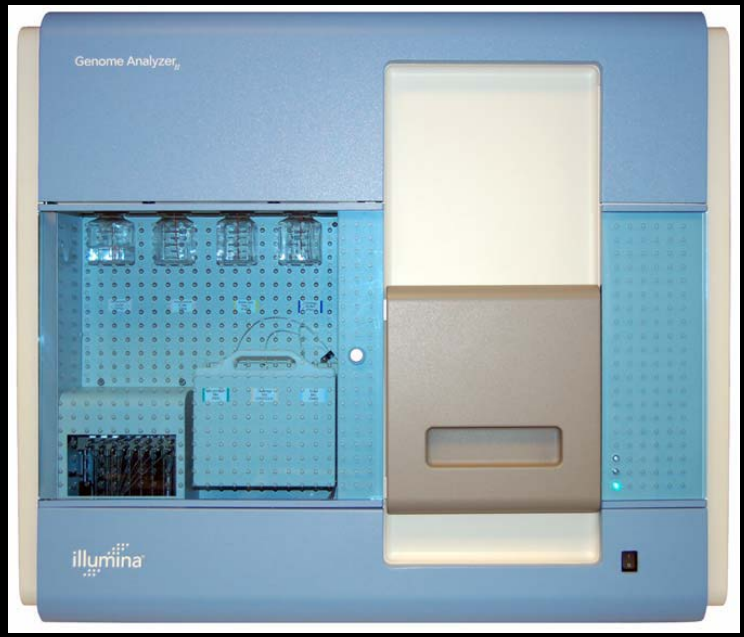
Replacing Reagents in the Genome Analyzer II, a part of a popular sequencing platform made by Illumina.



Solexa (Illumina) Sequencing Capabilities

- Read Performance:

Read Length	Days	Gigabases
35 bp	~2	10 - 12
50 bp	~5	25 - 30
75 bp	~7	37.5 - 18
100 bp	~9.5	54 - 60
150 bp	~14	85 - 95



- Read Performance:
  - The human genome is 2.9 Gigabases, getting 30x coverage in 14 days
  - You wouldn't sequence Homo Sapiens on an Illumina because of repeats: longer reads desired

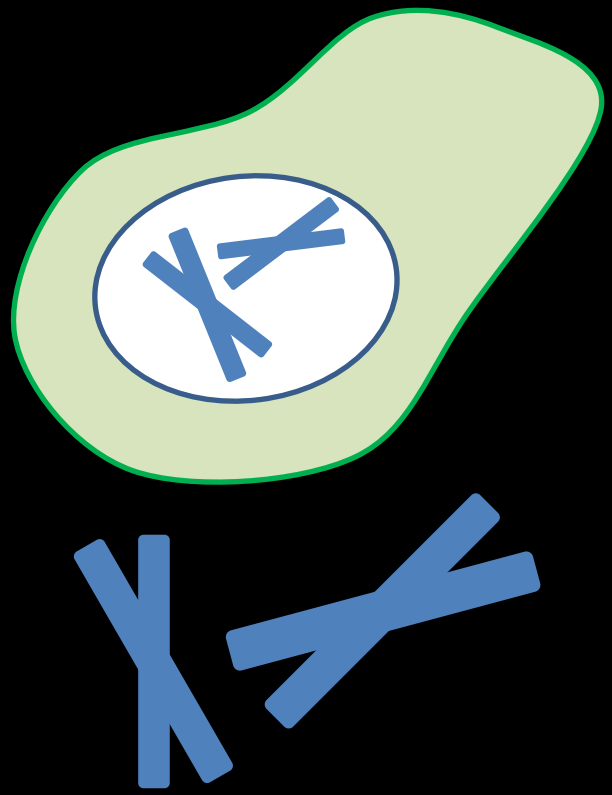
The Illumina sequencing process

- Random fragmentation of genomic DNA
- Addition of short adapters to the DNA fragments
- Immobilization of modified fragments to the flow cell
- Solid phase amplification to generate millions of distinct DNA clusters
- Base-by-base sequencing-by-synthesis using fluorescent reversible terminators
- Assembly

- All of these steps involve lots of carefully developed chemistry that makes this process possible – this took a lot of time to figure out

Start: with a tissue sample

- This is your cheek swab or other living cell sample



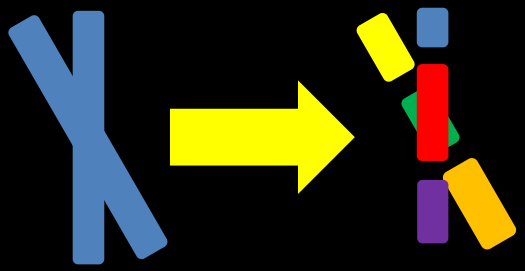
- The cell must be broken up first
  - Grinding
  - Chemical methods that lyse (pop) the cell

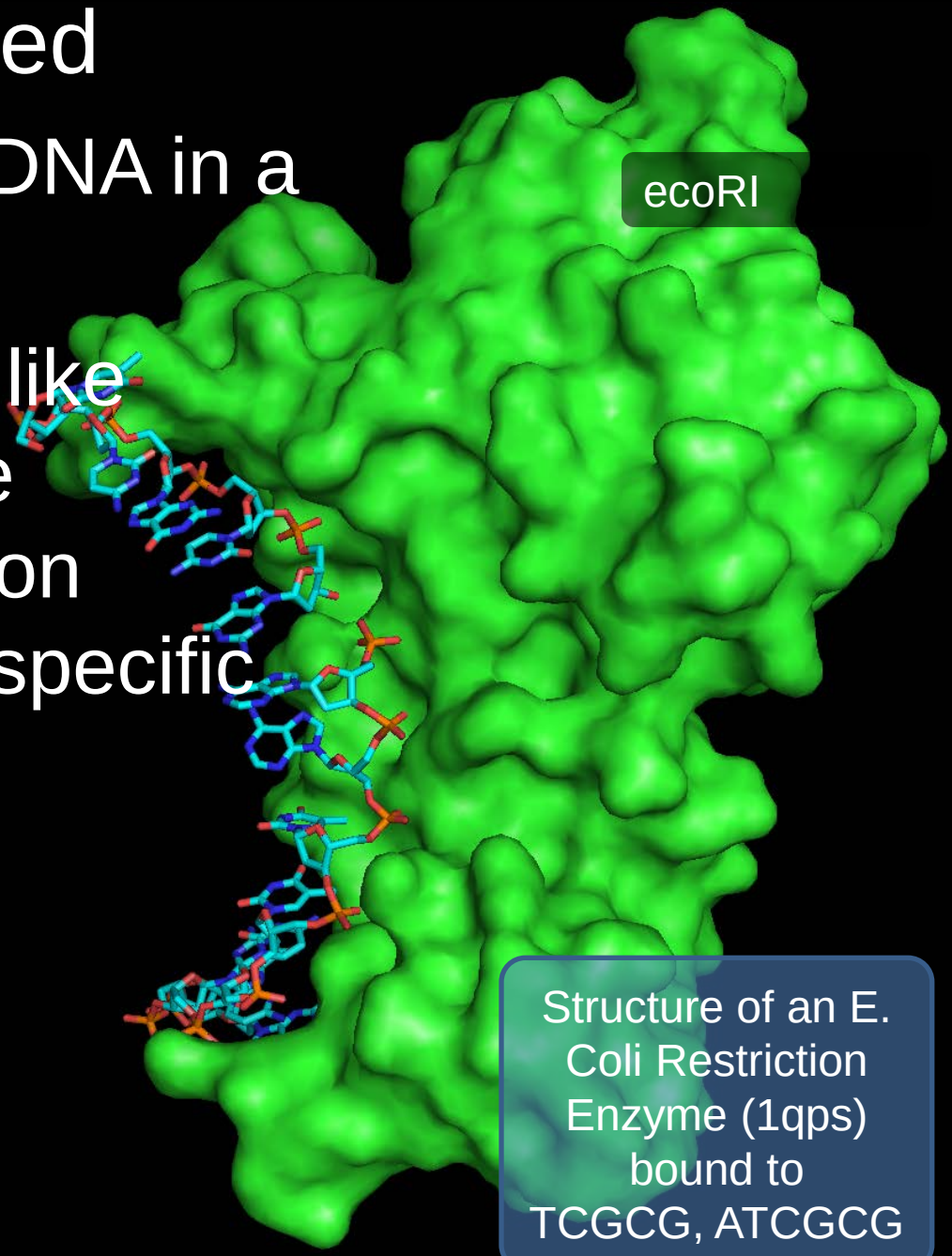
- The DNA can be precipitated out of solution by adding salts and ethanol

- Centrifugation can separate the precipitate from the rest of the cell matter

DNA must be amplified

- There is not enough DNA in a few cheek cells
- Large pieces of DNA like chromosomes can be separated by restriction enzymes that detect specific DNA sequences





Structure of an E. Coli Restriction Enzyme (1qps) bound to TCGCG, ATCGCG

Amplifying DNA

- Each restriction enzyme recognizes a specific sequence
  - For example, ecoRI recognizes:

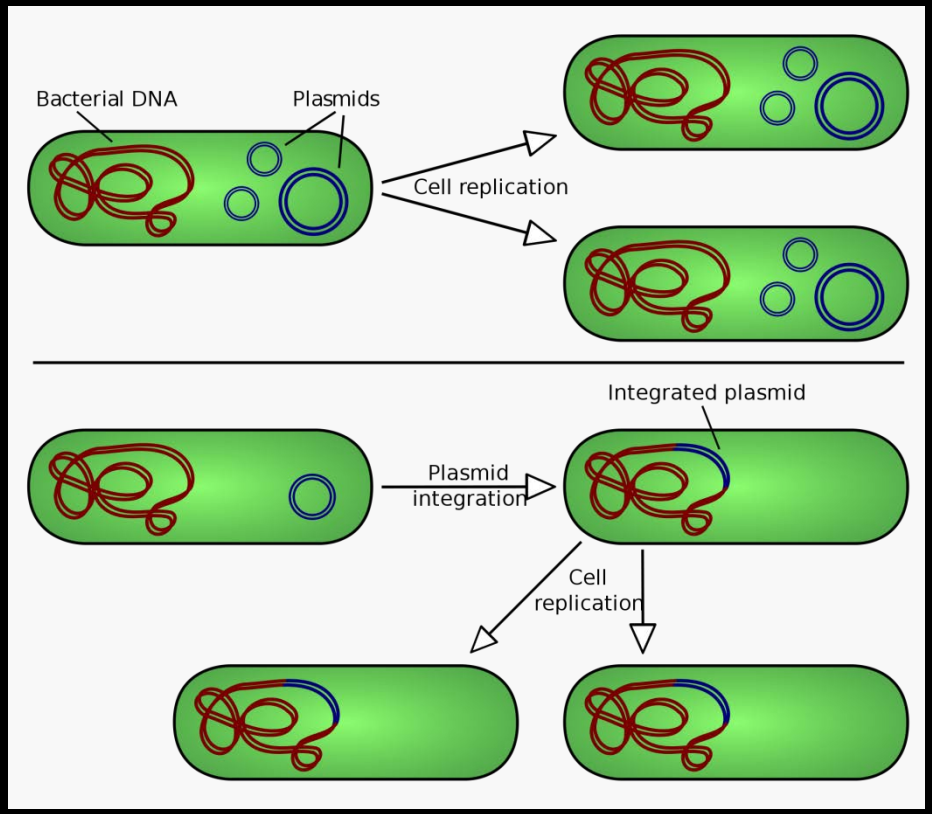
```
5'GAATTC      5'---G   AATTC---3'
3'CTTAAG      3'---CTTAA   G---5'
```

- Restriction enzymes break up double stranded DNA into “sticky ends”

- These DNA are deliberately added into bacterial DNA for amplification

Amplifying DNA

- Each bit of DNA is added to a carefully made DNA loop called a vector or a plasmid.
- The rest of the DNA loop encodes genes that convey antibiotic resistance in bacteria

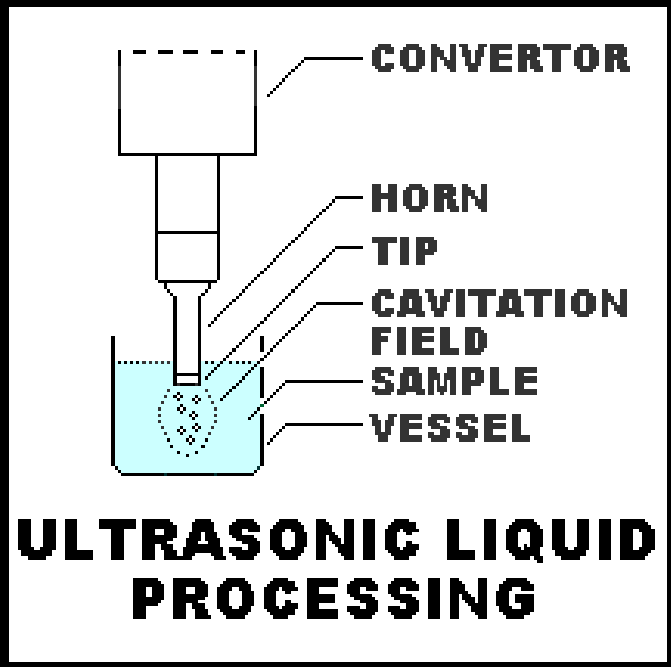


- The plasmids are inserted into bacteria
- The bacteria are exposed to antibiotics
- Bacteria that do not take the plasmid into their genome die
- The rest replicate, making many clones of the DNA

Step 2) Fragmentation

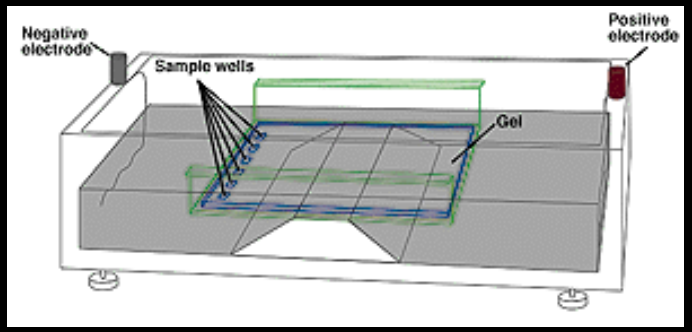
- Now we can destroy all the bacteria and harvest the DNA
- Next the DNA must be fragmented into tiny pieces so that they can be read. You are literally making reads.
- DNA can be fragmented with a probe sonicator, which breaks up the DNA into 200 bp fragments, but not smaller
  - Sound waves shake up DNA
  - Takes about 30 minutes

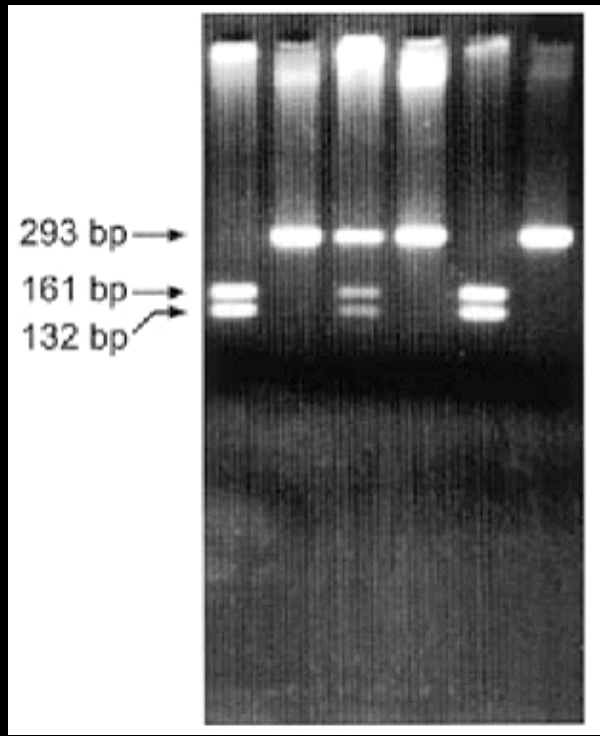
Probe Sonicator



Step 3) Electrophoretic purification

- DNA is negatively charged
  - When in an electric field, DNA is attracted to positive charge
- Sonicated DNA placed in gel
- Electric field is applied to the gel
- The gel's microporous structure slows down bigger molecules more than smaller molecules
  - Takes about two hours
- DNAs of different base lengths separate into bands
- Cut out the appropriate band with a razor
  - All approximately the same length





Gel separation of DNA strands of different lengths

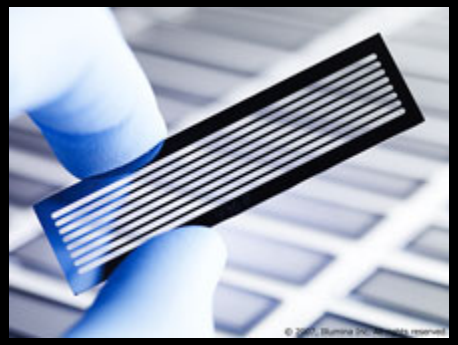
What we've got so far

- Fragmented DNA
  - fragments of generally the right size are kept
- All other components of the cell are gone
  - Two stages of filtration:
    - Centrifugation
    - Electrophoresis
- DNA separated into individual strands
  - No longer in double-strand configuration

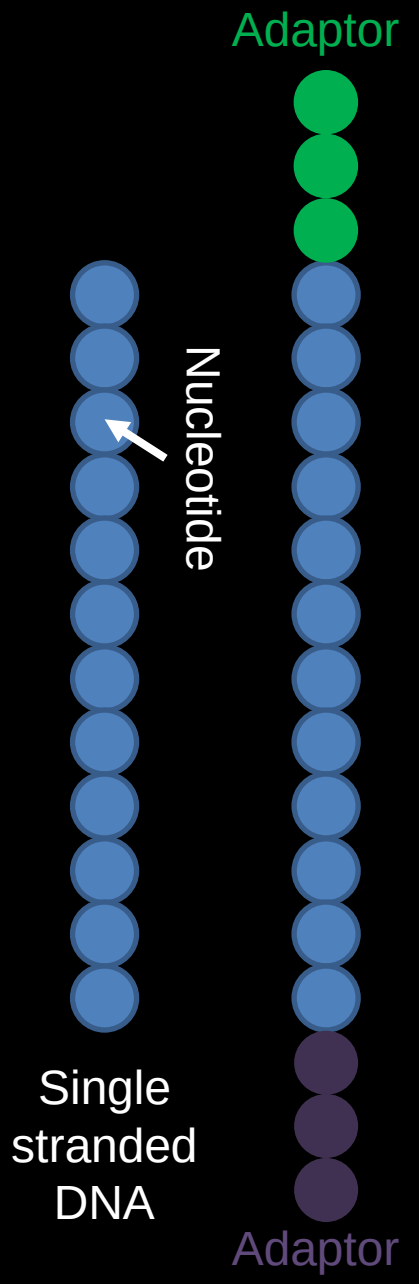
Illumina Cluster Station

Step 4) Adding adaptors

- Specially modified 5’ nucleotides are added to one end of the DNA strands

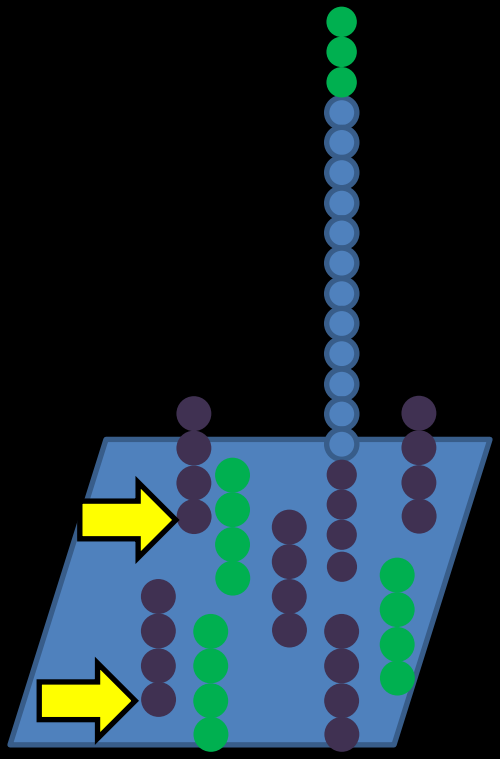


- The adapted DNA strands are placed into the flow cell, a tiny multi-lane chip that allows fluid to flow through from one side to another.
- One of these adaptors is detachable (more later)



Inside the flow cell

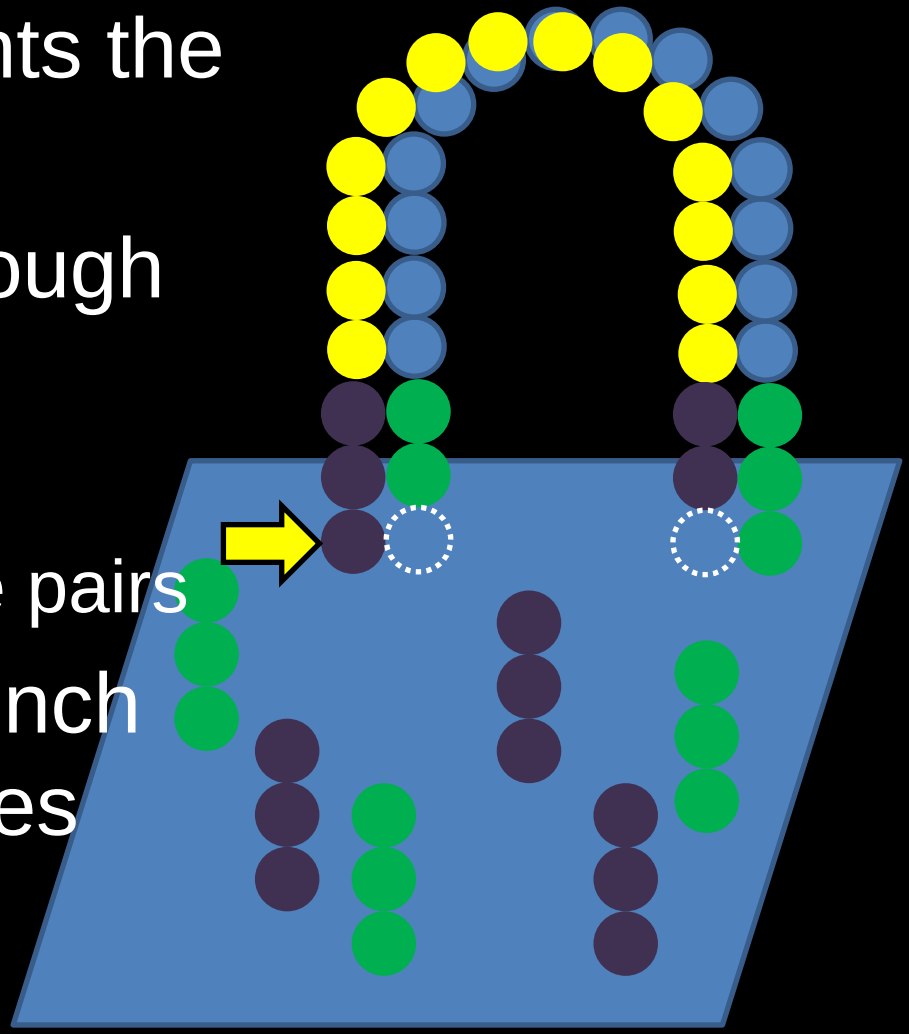
- The floor of the flow cell is covered very densely with both kinds of adaptors
- A special chemical reaction binds the DNA to the floor of the cell with an extra nucleotide



Illumina Cluster Station, interior

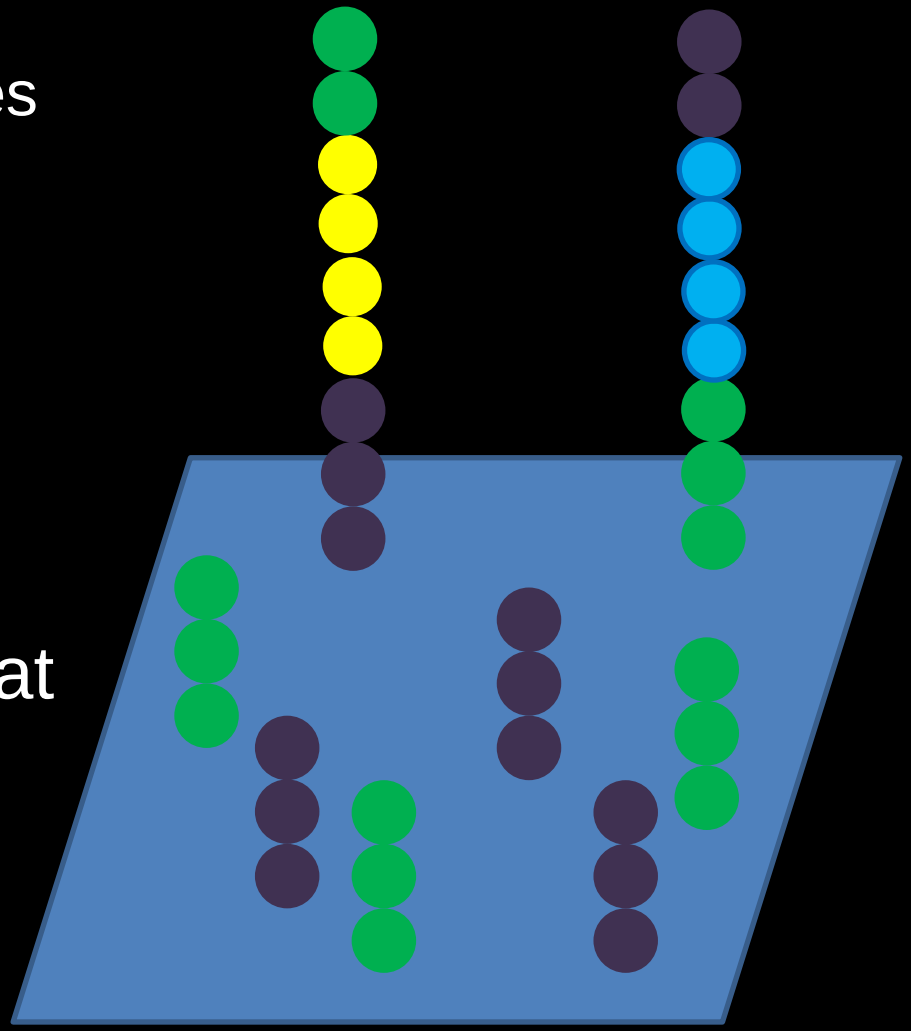
Step 5) “Bridge Extension” (surface PCR)

- The adaptors hybridize onto the opposite adaptor on the plate
- Polymerase complements the nucleotides
- Flowing nucleotides through the flow cell builds new bridges easily
  - Via complementing base pairs
- This creates a whole bunch of complementing bridges



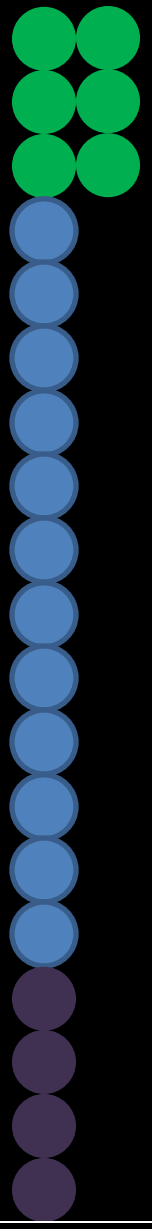
Creating more bridge extension

- Applying some heat and adding formamide straightens the bridges out by releasing the bridges from the flow cell
- separates one end of the bridges
- This process is repeated ususally 35 times.
- The green adaptor can be separated at the end, so that in the end you only have copies (and not complements)



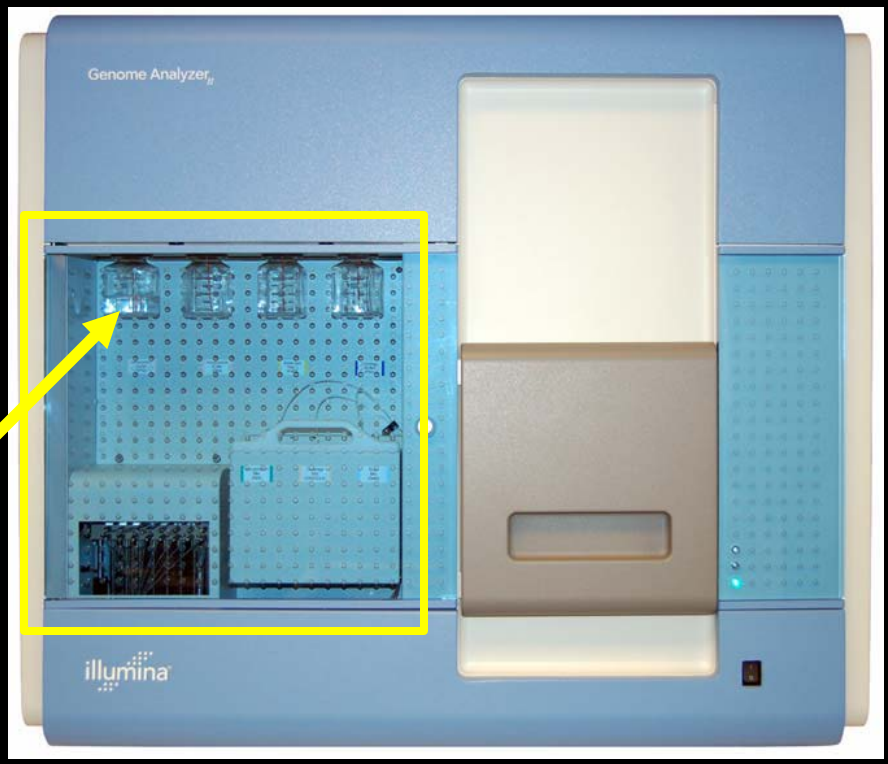
Bridge Extension Makes Detectable signals

- Sequencing by Synthesis
- A second green adaptor is attached for each DNA-bound green adaptor not on the glass.
- The second green adaptor is now an open receptor for nucleotides that complement the vertical sequence



Panels on the GAII

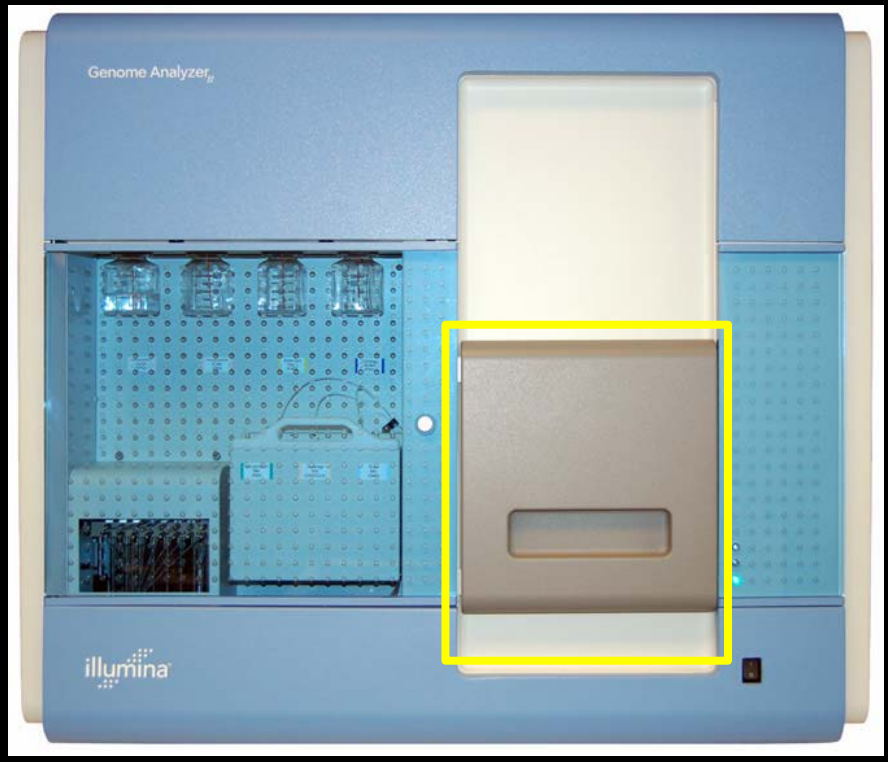
- The reagent processing panel holds connectors for reagent supplies (modified nucleotides, across the top)
- Flow control plumbing for many lanes is handled in the lower left
- Waste chemicals stored in the lower right

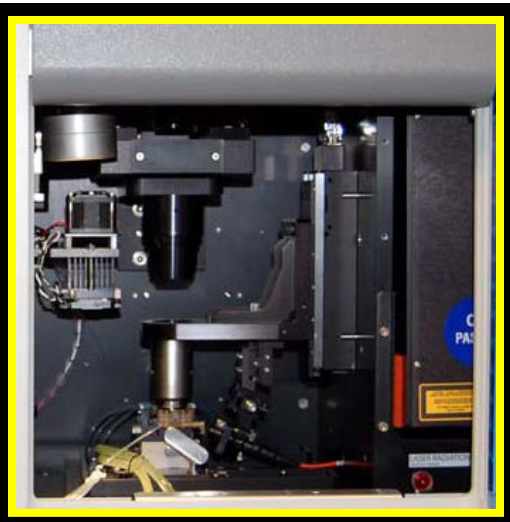


The Illumina Imaging panel

- The imaging panel holds a high resolution camera and flow plumbing.

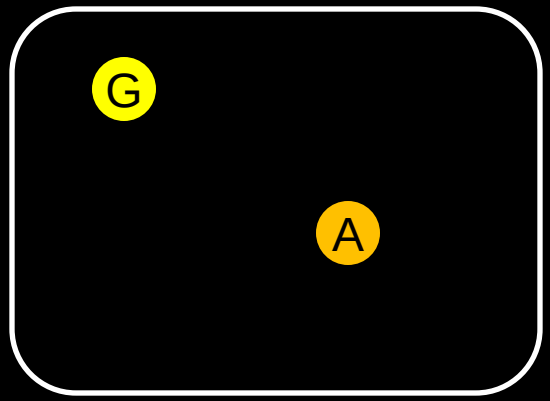
- The flow cell is mounted under the camera and specially modified nucleotide reagents are flowed over the dna in the flow cell

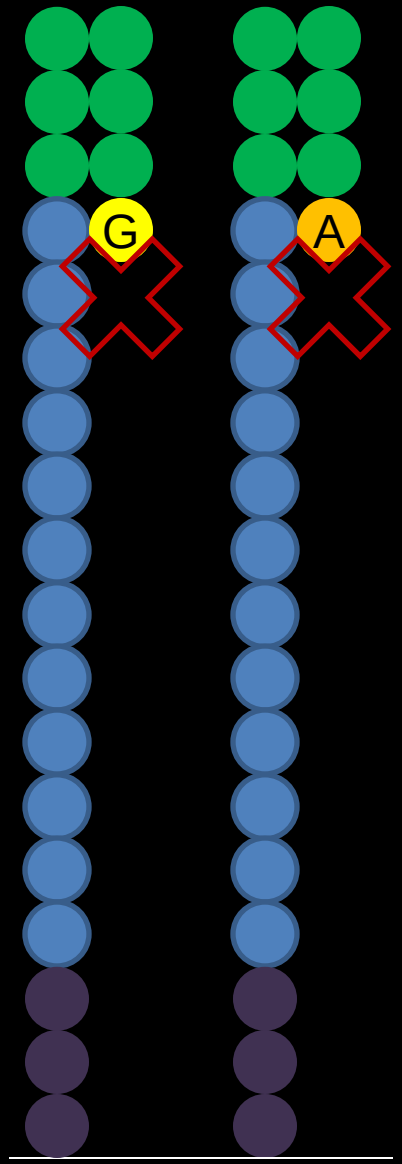




Step 7) Flowing the fluorescent nucleotides

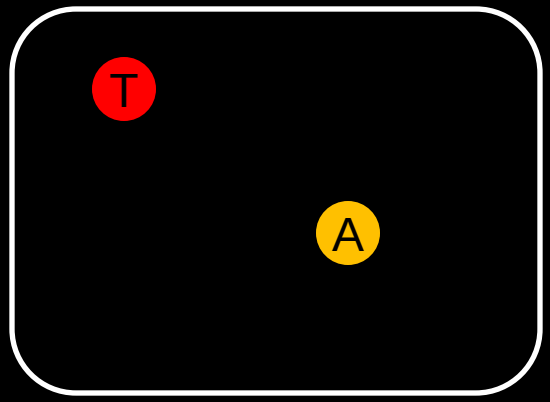
- Four modified nucleotides are washed over the flow cell
  - Nucleotides are fluorescent in a color corresponding to their nucleotide type
  - Specially modified to only extend the 3' end once (Self terminating)
- Once the wash is complete, the camera takes a fluorescent picture:

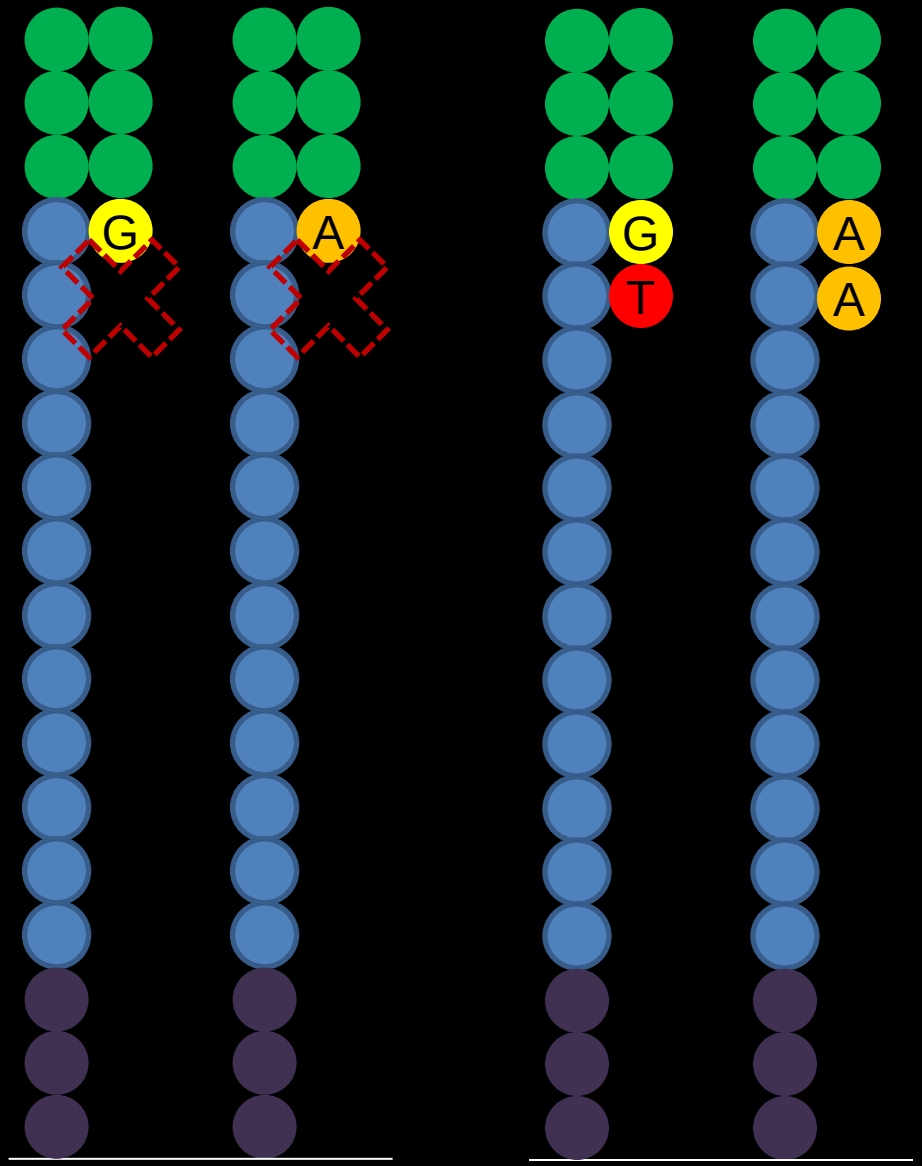




Step 7a) Terminators washed off

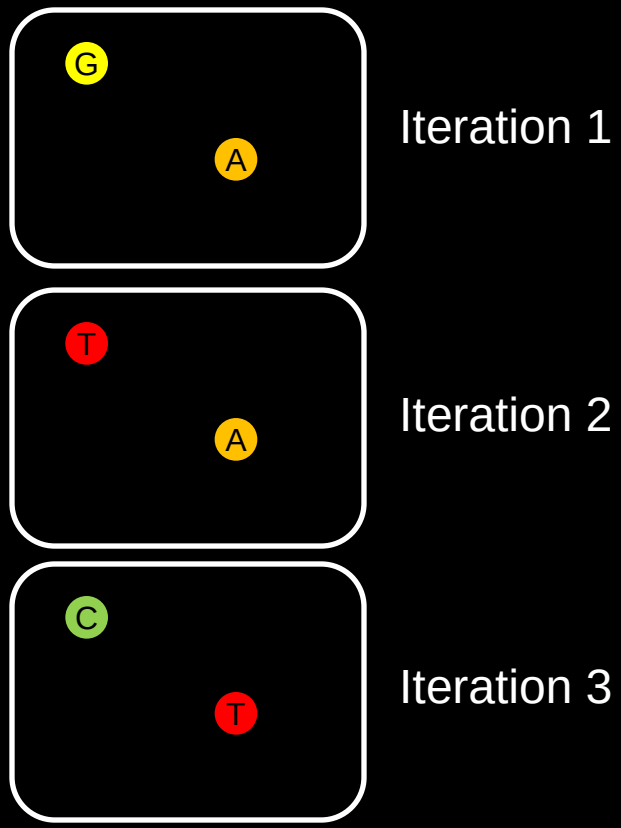
- Since we wish to continue sequencing the DNA, we wash off the chain terminators
- Now we can flow the next set of modified nucleotides over the flow cell:

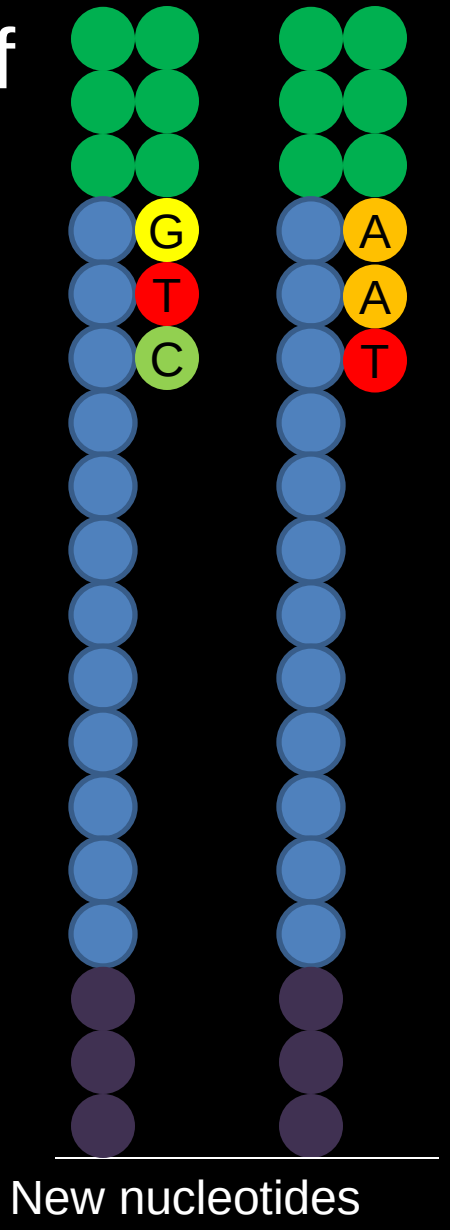




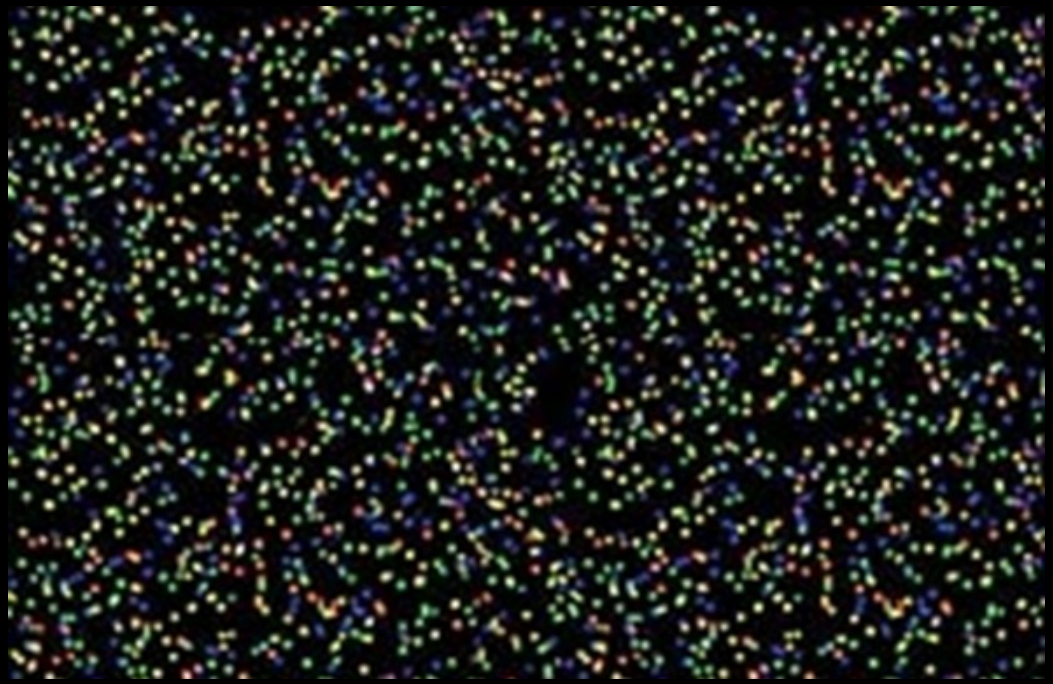
What the camera sees

- As chain terminators are washed off and new nucleotides are added, we get an evolving picture from the camera:





What this actually looks like in real life



- Each read corresponds to a fluorescent blob on the surface of the flow cell
- The reads change as the nucleotide terminators are removed and new nucleotides are added

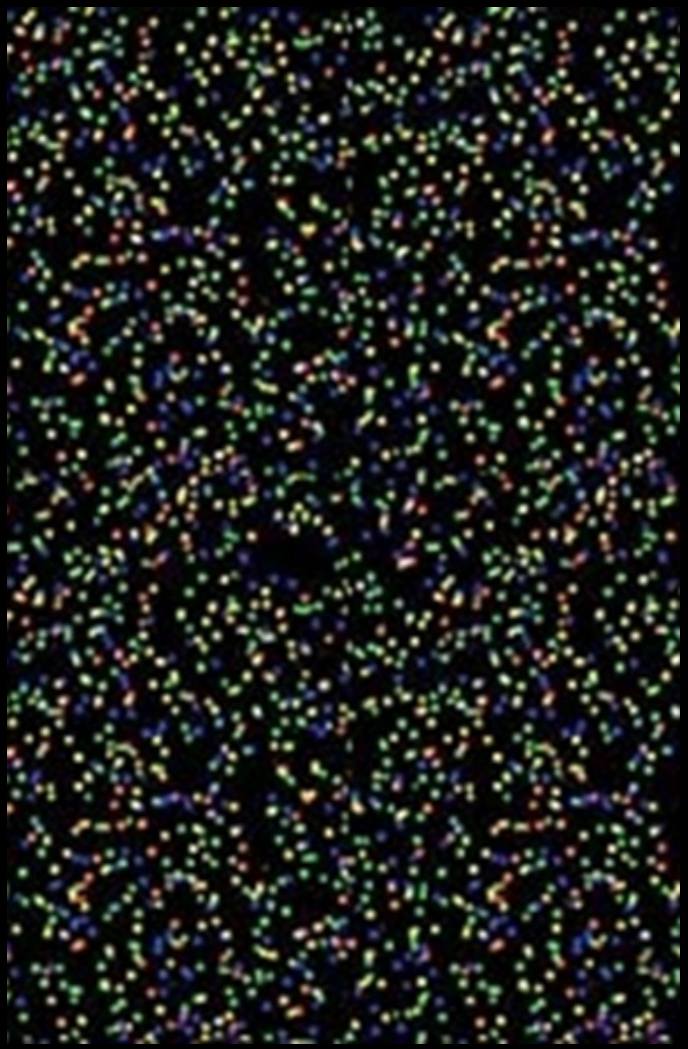
Parallel reads from the camera to the data

- There is nothing mechanical that is done for any read individually.
- The number of reads is dependent only on the size of the flow cell and the accuracy of the camera – so there can be MANY reads

- This is why modern sequencing technology will never be short on the number of reads or the amount of coverage
  - Technologies like this enable the data to be collected in an efficient way

Problems with image-based data collection

- Segmenting reads from the image data is nontrivial
  - Reads can be too faint
  - Reads might overlap with other reads
- Read errors can still happen
  - Sometimes the wrong nucleotide binds to the DNA, the camera cannot tell
  - This just gets registered as a wrong read



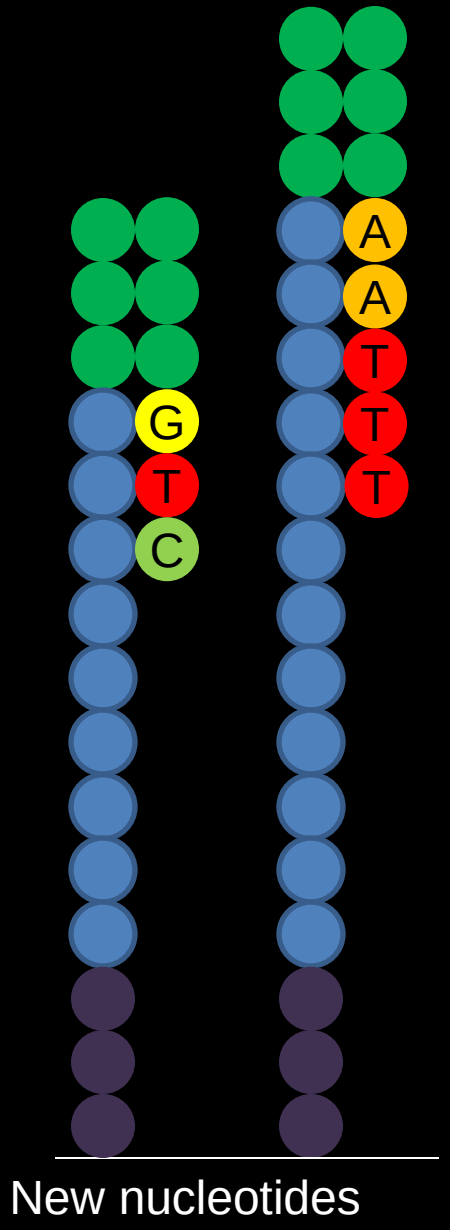
Simple problems that can be dealt with

- Faint and overlapping reads
  - Throw this data away and do not considerate for image analysis
  - There are so many reads anyway that these are not essential

- Dephasing problems
  - This is when the reads are the wrong length because a nucleotide did not stick, or the DNA itself is too short
    - This happens and the dot simply disappears. These get thrown out as well.

A picture of dephasing

- Dephasing can happen because:
  - one of the reads is too short for some reason
    - Note that it does not matter if it is too long
    - We cannot typically read the entire length of the segment (200 bp) because the termination becomes increasingly unstable
  - Several nucleotides are bound at once
    - This happens sometimes and if it happens, you cannot detect it, so the sequence will appear shorter



Incorrect reads

- Incorrect reads are the reason why such high coverage is necessary for modern sequencing
  - When there is disagreement, contigs cannot be logically assembled without an indication that something is an error

Multiple identical reads

- When you have super high coverage, you have multiple identical reads
- You can search for these reads and compare them; if only a tiny minority are slightly different, you can eliminate that read
- To make this possible, you have 6-7 copies of each read, which means for 30 bp reads, you have 210x coverage, easily
- Modern techniques are going for 300-500x coverage

Illumina sequencing performance

- In the beginning of the lecture, we saw that we could sequence the human genome at 10-15x coverage in a few days
- Because of errors in reads, our practical rate is slower, because of all the additional coverage that is necessary for correctness

Read Length	Days	Gigabases
35 bp	~2	10 - 12
50 bp	~5	25 - 30
75 bp	~7	37.5 - 18
100 bp	~9.5	54 - 60
150 bp	~14	85 - 95

Announcement

- Thursday March 1st will be dedicated to question and answer about project 1

- There is no class on Tuesday Feb 28 (Professor Chen at Conference)

- Project is due March 2, midnight

Questions