SPAAM

Standards, Precautions & Advances in Ancient Metagenomics

Lecture 1A: Introduction to NGS Data

James A. Fellows Yates

Overview

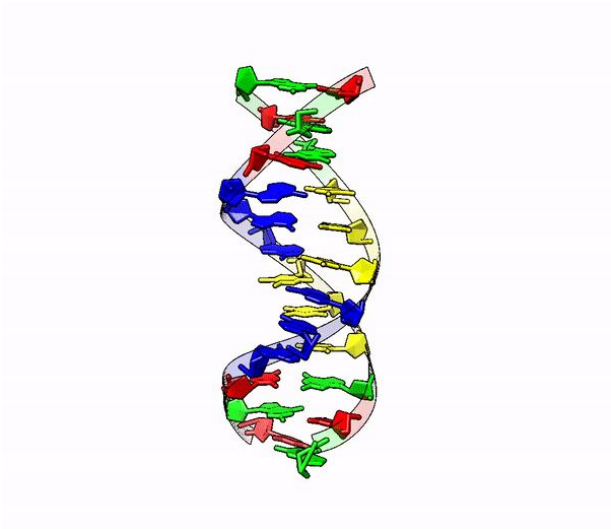
1. Describe basics of DNA

2. Introduce what DNA sequencing is

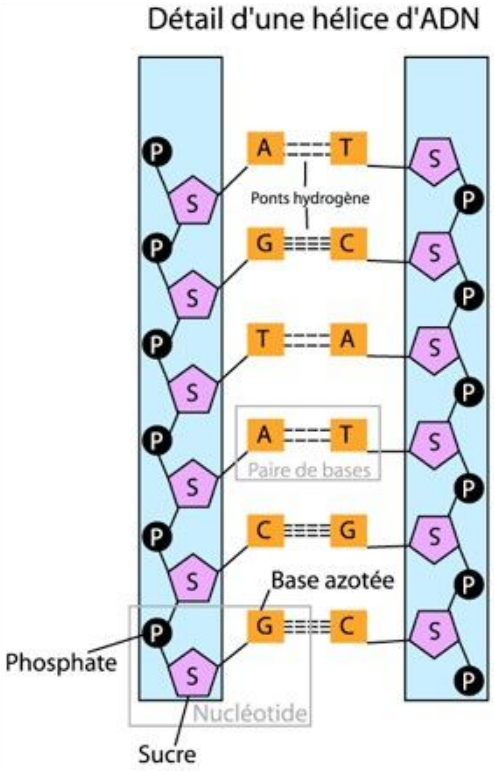
3. Explain how Illumina **NGS** sequencing **data** is generated

What is DNA?

What is DNA?



Erin Rod, CC BY 4.0, via Wikimedia Commons



Pradana Aumars, CC BY-SA 4.0, via Wikimedia Commons

The Rules

Four nucleotides

- Pyrimidines: **C**ytosine, **T**hymine
- Purines: **G**uanine, **A**denine

Base pairing: one pyrimidine with one purine

- *C* with *G* (think: CGI)
- *A* with *T* (think: AT-AT walker)

Complementary

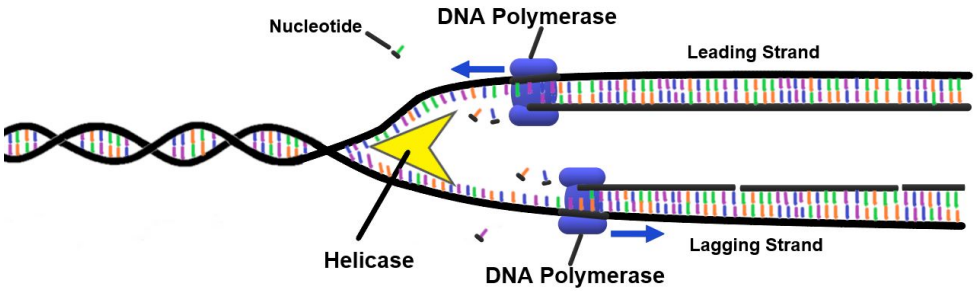
- *C* on one strand, *G* on the other (or v.v.)
- *A* on one strand, *T* on the other (or v.v.)

Screenshot from The Empire Strikes Back
Source: © 1980, 1997, 2004 Lucasfilm Ltd. All Rights Reserved.
(Wikipedia)

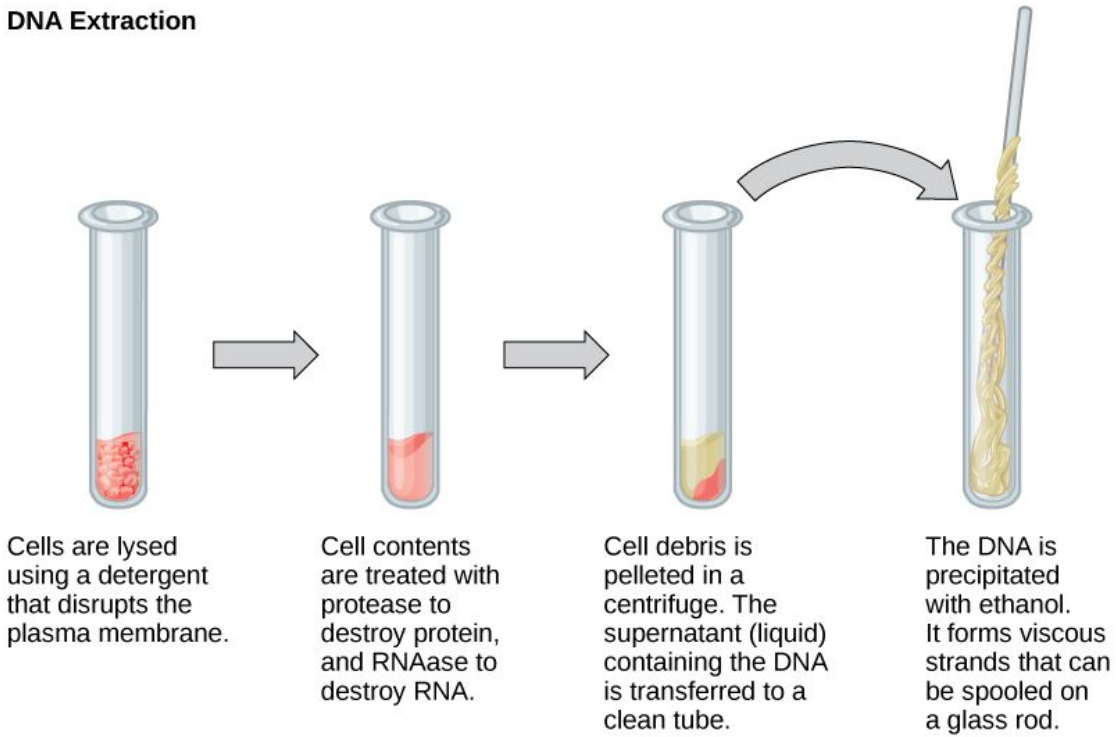
The Rules

Make copy of a DNA strand a.k.a. Replication

- Unwind the DNA
- Separate the strands
- Make new strand: find a C, get new G (etc)
  - via a polymerase (taken from ‘free’ nucleotides’)



How do we get DNA?

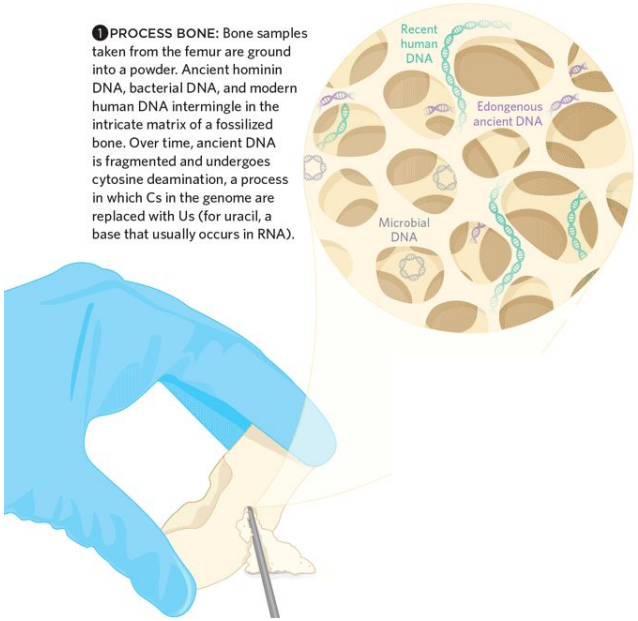


What about ancient DNA?

Basically the same!

Except: aDNA molecules are **degraded**

- Fragmented (short molecules)
- Damaged (modified nucleotides)
- Contamination (aDNA in soup of modern DNA)



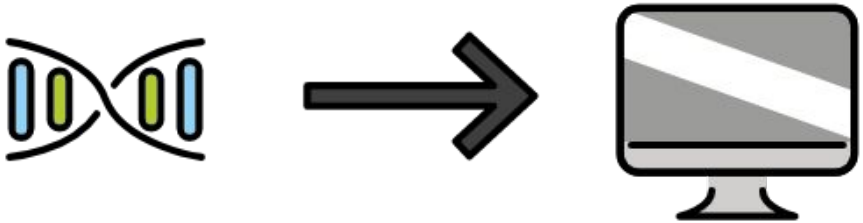
Introduction to DNA Sequencing

What is Sequencing?

Converting the chemical nucleotides of a DNA molecule

to

ACTG on your computer screen

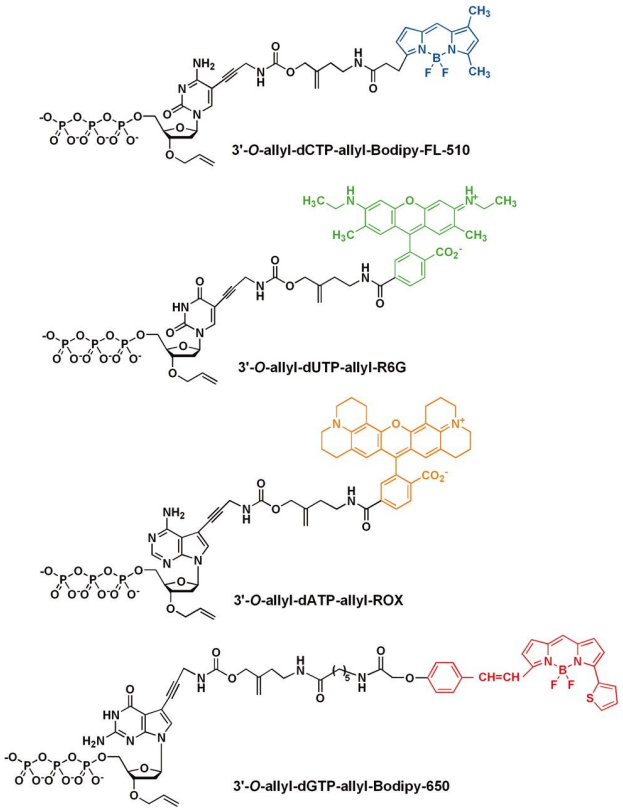


How does it work?

Replicate a strand, but add complementary **fluorophore-modified nucleotide**, one colour per base

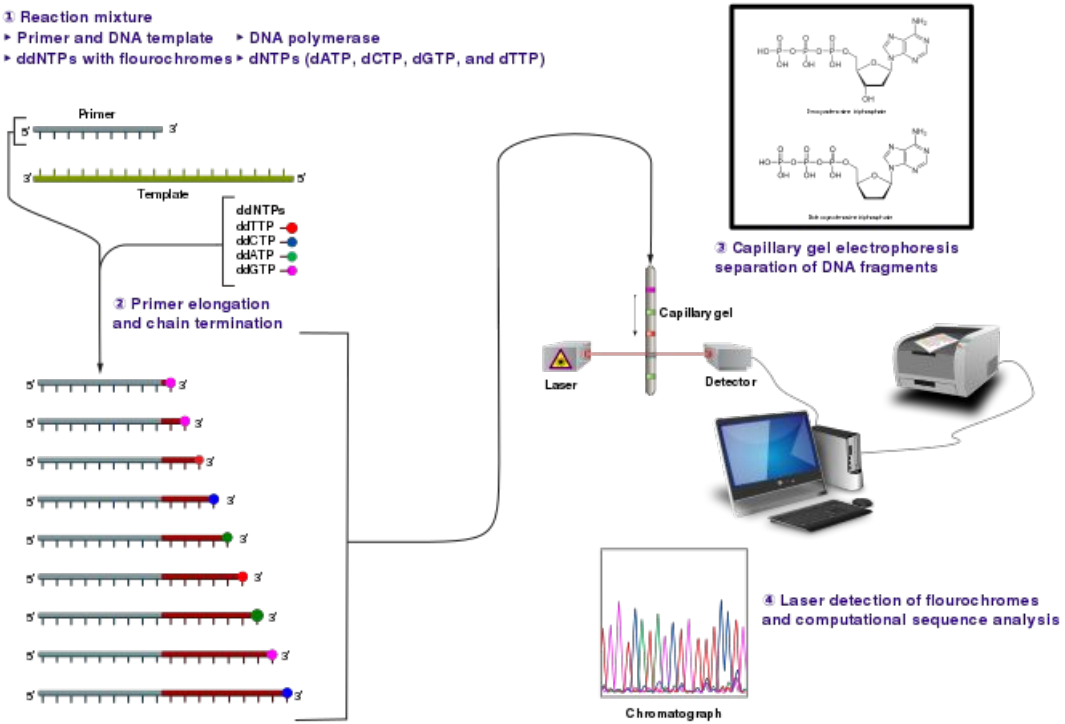
A G T C

Fire mah lazer, and record the colour! Rinse and repeat!

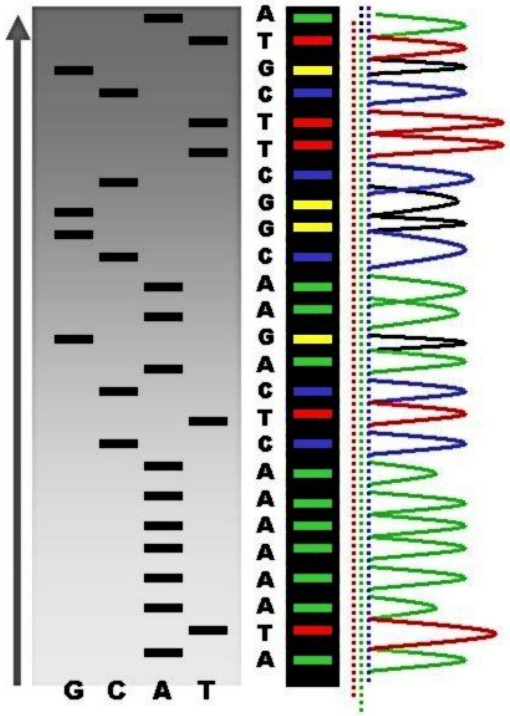


Ju et al. (2006) PNAS DOI: 10.1073/pnas.0609513103

Sanger Sequencing



Estevezj, CC-SA-BY 3.0, via Wikimedia Commons



Abizar, CC-SA-BY 3.0, via Wikimedia Commons

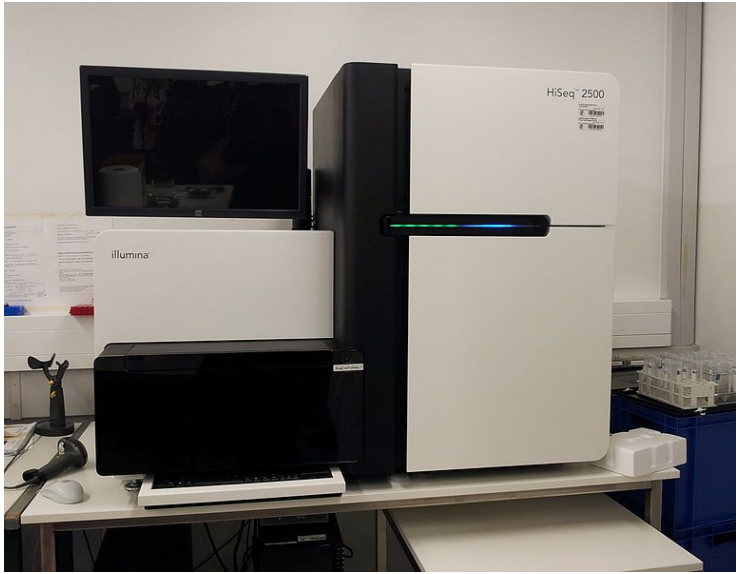
What is NGS?

Historically: Sanger sequencing

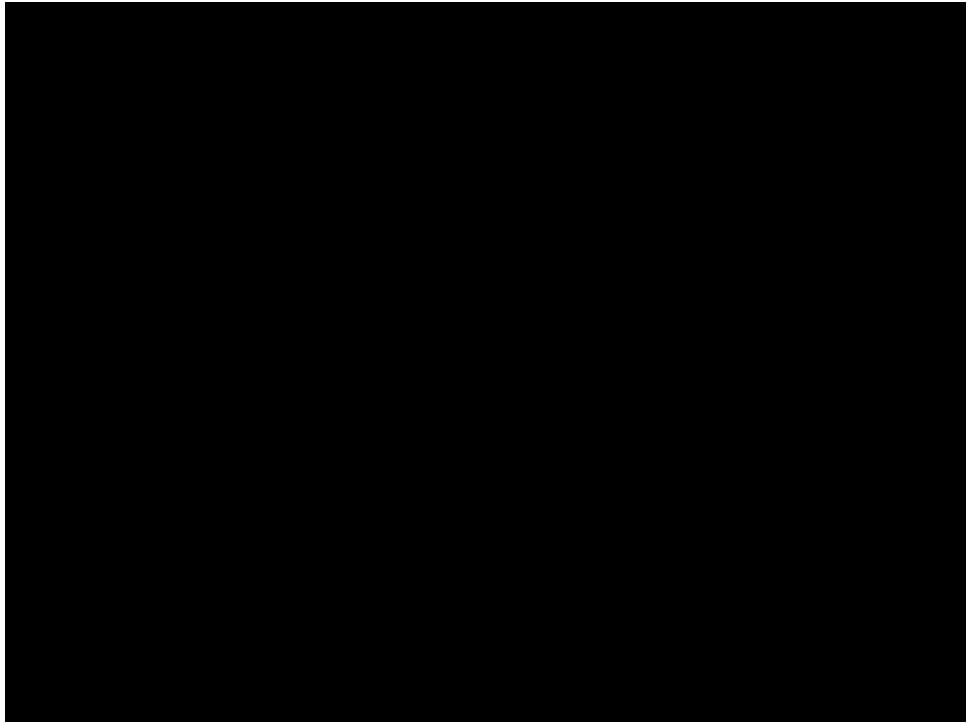
- Slow, expensive, resource hungry

"Next Generation Sequencing"

- Sequence billions of DNA molecules at once!
- Fast and cheap!
- Market leader: Illumina (others: PacBio, IonTorrent)
- *Really more 'second' generation now - see Oxford Nanopore*



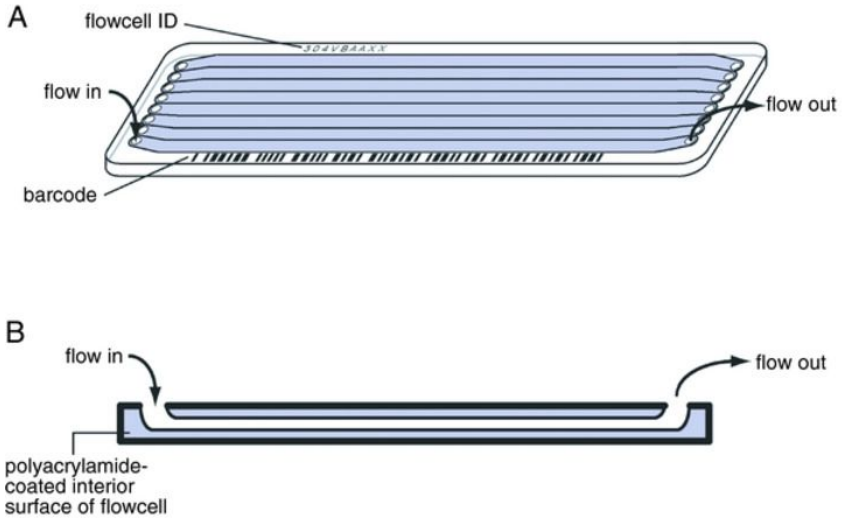
How does it work?



Via Gfycat (BlackGreedyAurochs)

Where does this happen?

On a ‘flow cell’: glass slide with synthetic DNA ‘lawn’



Bronner et al. (2013) Current Protocols in Human Genetics, DOI: (10.1002/0471142905.hg1802s79)

Where does this happen?

But how do you get your DNA to attach to the lawn (and not get lost)?

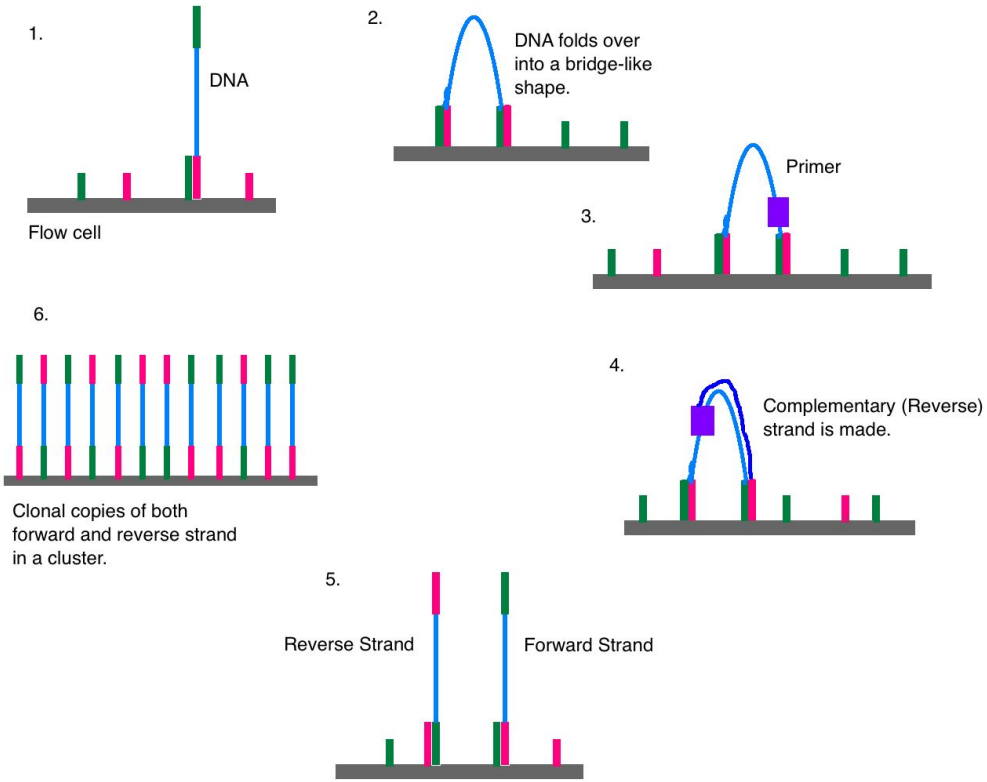
Convert it to library:

- Add adapters: bind to the 'lawn' of the flow cell
- Add priming sites: where enzymes start copying DNA
- Add indexes: sample-specific barcode

Illumina DNA Construct

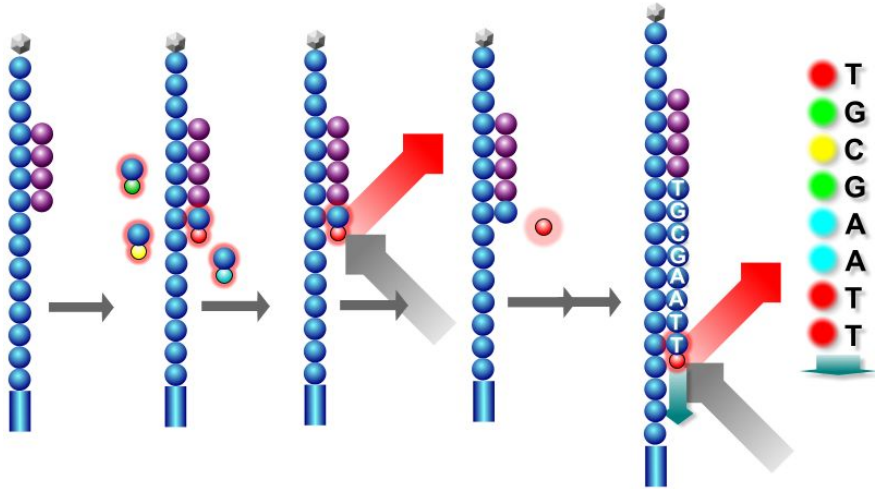
```
GGTGATACGGCGACCACCACaccgacGGCCCTACACGACGCTCTTCCGATCTXXXXXXAGCACACGTCTGGGCTCCAGTCACgacactaCCGTCTTCTGCTTG

||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||

TTACTATGCCGCTGGTGGTGtggctgttGGGATGTGCTGCGAGGGGGCTAGAXXXXXXTCGTGTGCAGACTTGAGGTCAGTGctgtgatGGCAGGGGACGGGC

[Adapter/Index Primer] [Index]   [Target primer]     [Target] [Target primer]       [Index] [Adapter/Index Primer]
```


Clustering

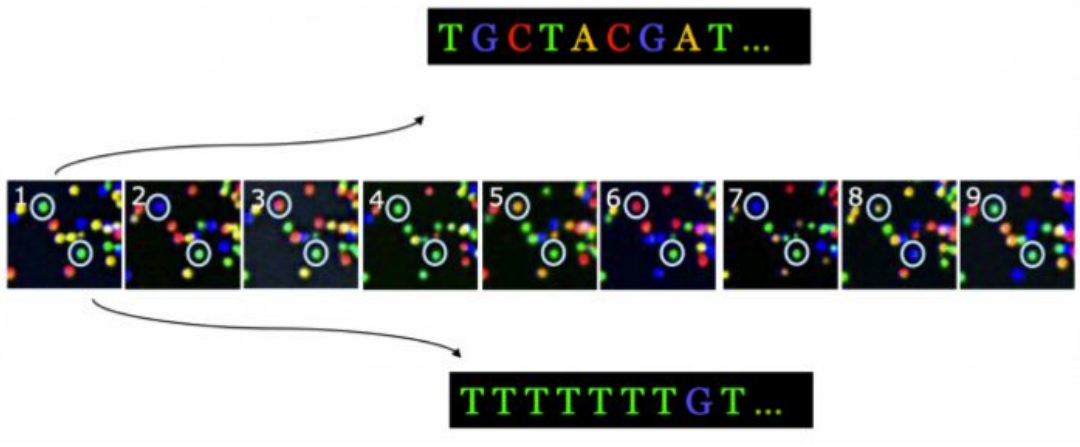


Sequencing-By-Synthesis

1. Add fluorescent nucleotides (complementary will bind)
2. Wash away unbound nucleotides
3. Fire laser & take photo
4. Remove fluorophore
5. Back to 1 ⤴️ [x50, x75 or x125 times, a.k.a. cycles]

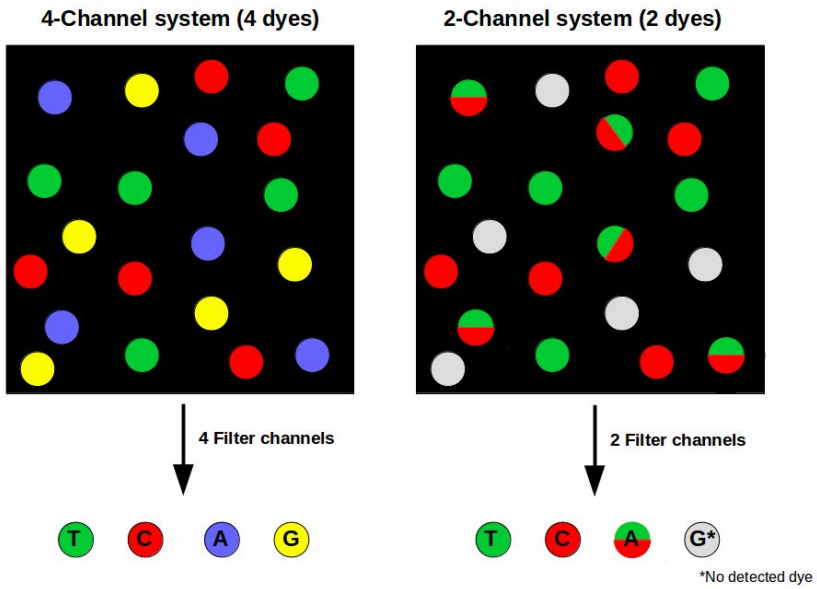


What does this look like?



Remember: this is happening millions of times at once!

Caveat: two colour chemistry



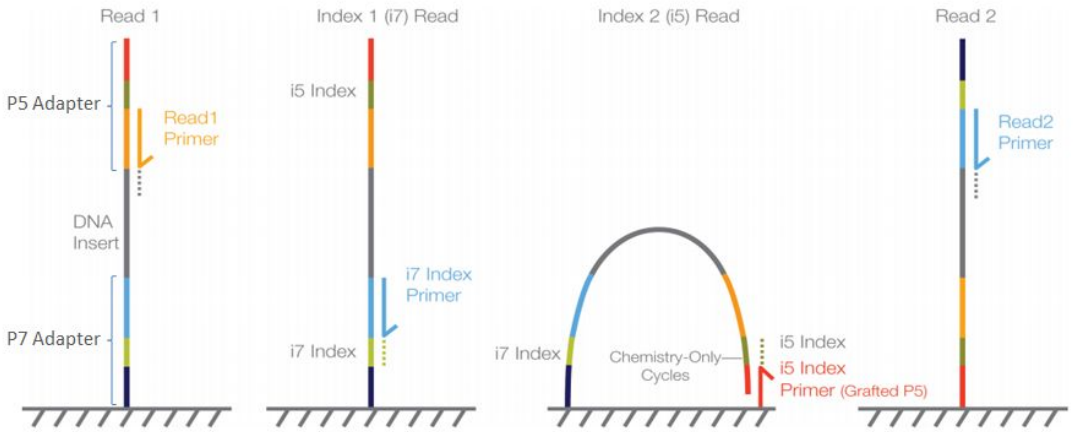
https://www.ecseq.com/support/ngs/do_you_have_two_colors_or_four_colors_in_Illumina

Improving Quality

- Over time, imaging reagents get ‘tired’ and more errors occur
  - Bases sometimes don’t bind, or multiple == clusters ‘desynced’
  - Base-quality: machine calculates probability it got the ‘right’ nucleotide for each photo
- ‘Dead’ base call: typically reported as N
- How to improve or correct?

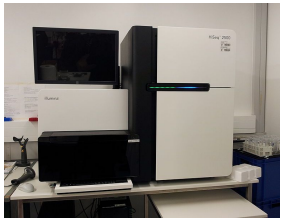
Improving Quality

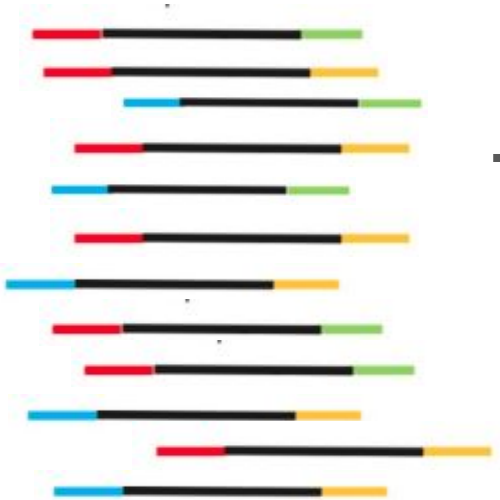
- Improvement: **paired-end sequencing**
  - Get order of nucleotides by sequencing from one end
  - Get reverse order of nucleotides - sequence other end!
  - Bonus: sequence more of read longer than cycles



Biological to Computational Sequences

Demultiplexing

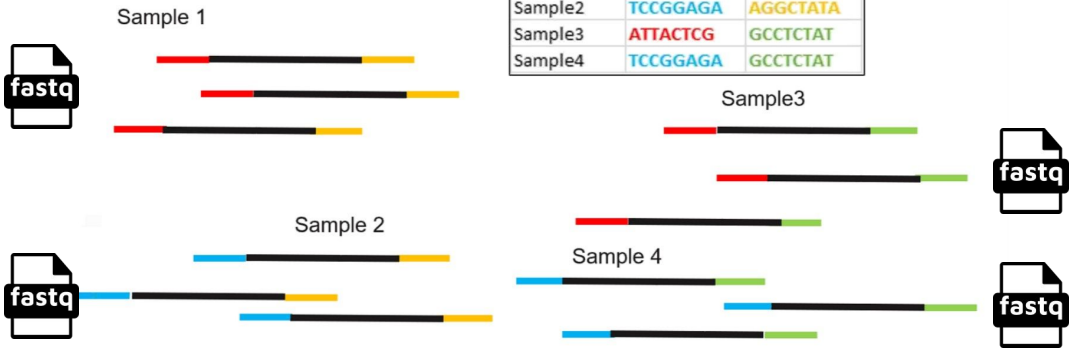




Demultiplexing

- **Assigns clusters to a sample, based on the cluster's index sequence which is provided in the sample sheet**

[Data]		
Sample_ID	index	index2
Sample1	ATTACTCG	AGGCTATA
Sample2	TCCGGAGA	AGGCTATA
Sample3	ATTACTCG	GCCTCTAT
Sample4	TCCGGAGA	GCCTCTAT



10

For Research Use Only. Not for use in diagnostic procedures.

illumina

FASTQ File

> *FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity -* **Wikipedia**

FASTQ File

Example (files can be gigabytes in size!)

```
@K00233:37:HGHLYBBXX:3:1101:2646:1121 1:N:0:NACGCATC+NGCTGGTG
NCGCATGAGCCGCCTGTATCAGGCGCTGATCGGGCCGGGCATTGCAGTTGGGATAGATCGGGGGAGCACACGTCTG
+
#A7F<<GG<JFJFJJJJJJFFJJJJJJJAFFJFJJJJJJJFJAFFFJAJFJJ<FJJJJJFFF<FFA--FFFJJJJJ
@K00233:37:HGHLYBBXX:3:1101:4655:1121 1:N:0:NACGCATC+NGCTGGTG
NATGCATGACAGGAGGTGAGGGCATTTTCCAGATTTTCAGGCTGCGACCTTGAGCATCTTTCGCCGCTTCCAGCAC
+
#GG-<FFFF7JFF7JJJJJFJJ<JJJJJA7FJJJJJJJFF<JFF<J7-<FJJJJFJFFJJJGGGGFFJJ--AJAJJ
```

```
@ <read id, e.g. machine ID, location on flowcell> <extra metadata>
  <DNA sequence; Note: N = base couldn't be called!>
+ <a separator>
  <base quality scores for each nucleotide in sequence>
```

Quality score:

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
0.2.....................26...31........41
```


Recap

- DNA molecules essentially:
  - Made up of nucleotides (ACTG)
  - Two strands: complementary base pairs (C-G, A-T)
  - Modern DNA is long, aDNA is: short
- NGS Sequencing:
  - Massively multiplexed: millions DNA molecules at once
  - Add adapters to bind to a glass slide (lawn)
  - Make new strand, adding fluorescent nucleotides
  - Fire laser at each nucleotide and take photo
  - Desyncing of clusters result in lower base-quality scores over time
  - Improve by paired-end sequencing

Considerations for Ancient Metagenomics

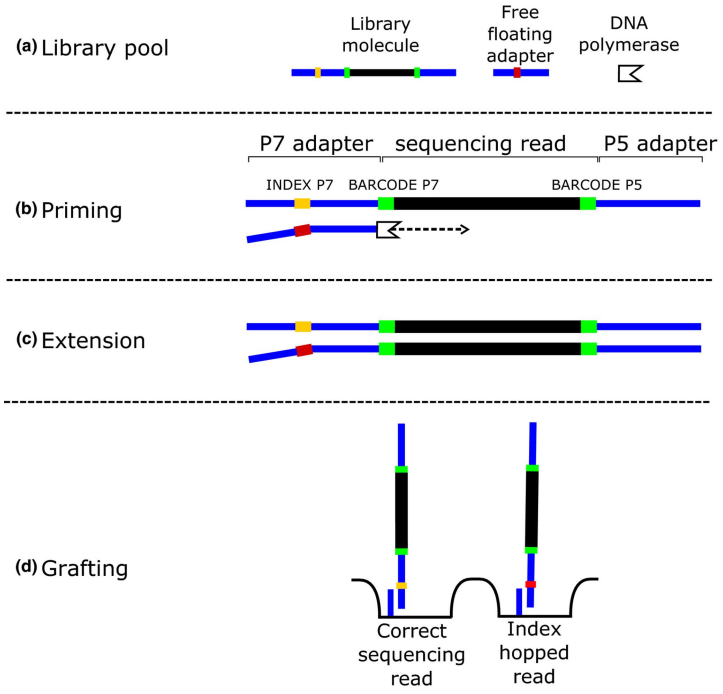
Low DNA preservation

Low template DNA: risk of PCR duplicates

- Can inflate counts
- Reduces number of sequenced reads
  - Duplicates ‘compete’ for sequencing slots over unique reads

Index Hopping

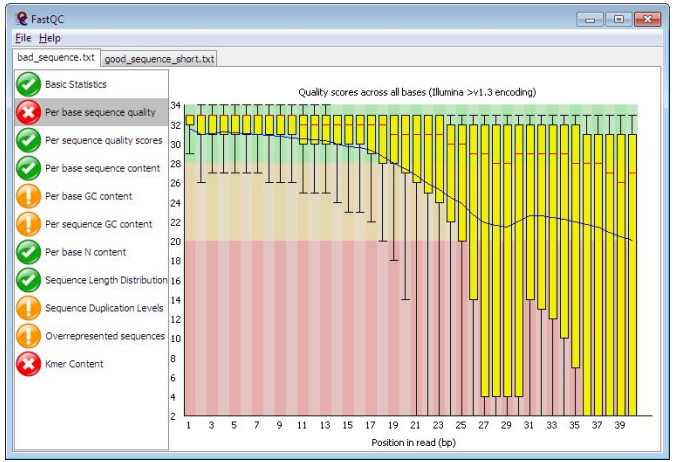
- Challenge when multiplex sequencing
  - Most often in patterned flow cells (HiSeq X/NovaSeq); caused by free-floating index primers
  - Chimeric index combinations: insert 'taxa' from other samples in your sample
  - Scenario: e.g. mixture of capture and shotgun samples on one run!



Van der Valk (2019) Mol. Eco. Res.

Sequencing Errors

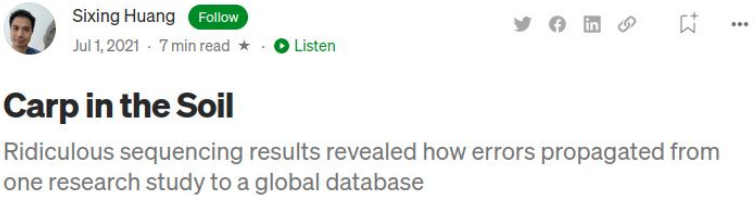
- Sequencing errors
  - Wrong assignment during taxonomic profiling (against wrong genome)
  - Reduces chance of overlap during *de novo* assembly
  - Incorrect variant calling (low coverage)



Dirty Genomes

- Adapters
  - Dirty genomes in databases
  - False-positive assignments

Ezidor under GNU 1.2 via Wikimedia Commons



Sixing Huang Follow
Jul 1, 2021 · 7 min read ★ · Listen

Carp in the Soil

Ridiculous sequencing results revealed how errors propagated from one research study to a global database

Garbage in, garbage out. But first you need to know what garbage looks like.

We need to stop making this simple f*cking mistake

23RD MARCH 2016 / BIOMICKWATSON / 4 COMMENTS

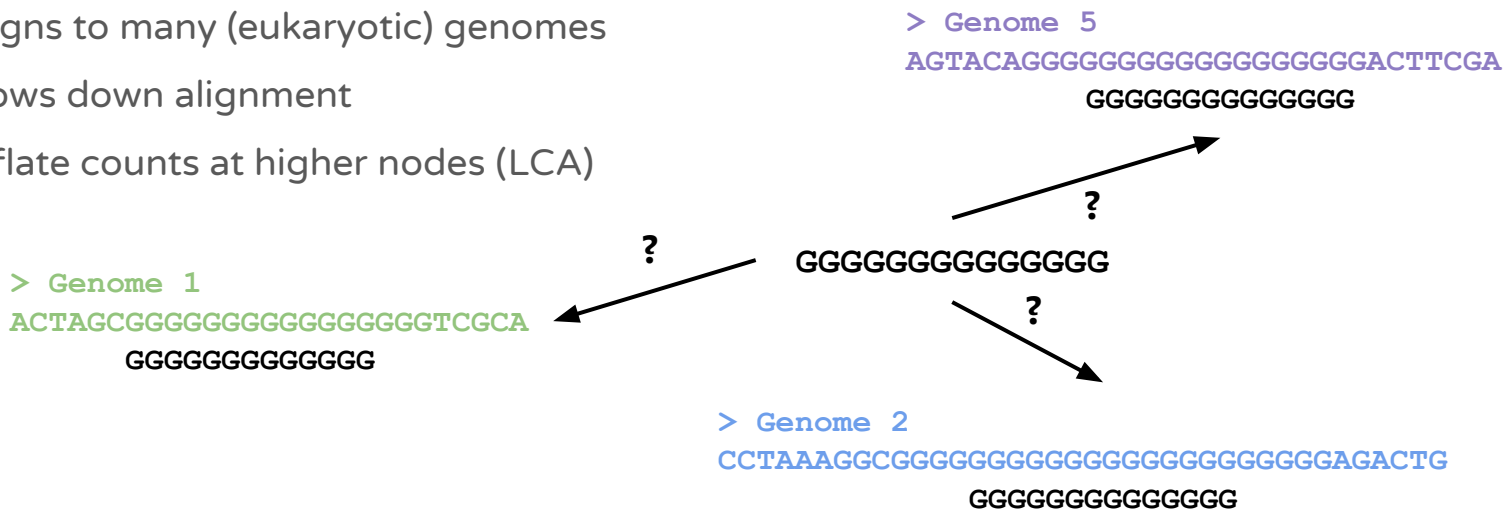
I'm not perfect. Not in any way. I am sure if anyone was so inclined, they could work their way through my research with clinical forensic attention-to-detail and uncover all sorts of mistakes. The same will be true for any other scientist, I expect. We're human and we make mistakes.

However, there is one mistake in bioinformatics that is so common, and which has been around for so long, that it's really annoying when *it keeps happening:*

It turns out the Carp genome is full of Illumina adapters.

Low sequence diversity

- Low sequence diversity reads / polyG tails
  - unspecific
  - reduces number of classified reads
  - aligns to many (eukaryotic) genomes
  - slows down alignment
  - inflate counts at higher nodes (LCA)



Recap

Check for

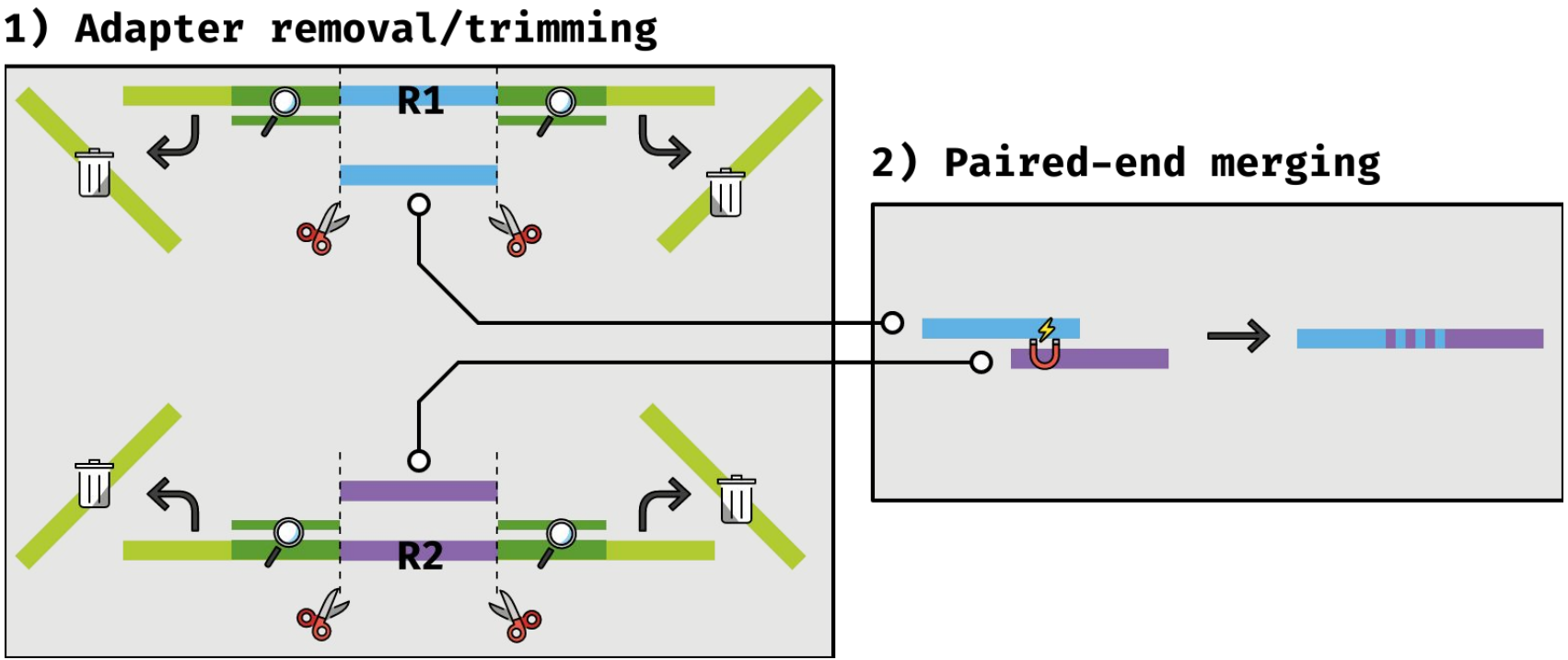
- Duplication rate
- Index Hopping
- Sequencing error
- Adapters
- Low-sequence diversity reads

Questions!
(then tech support)!

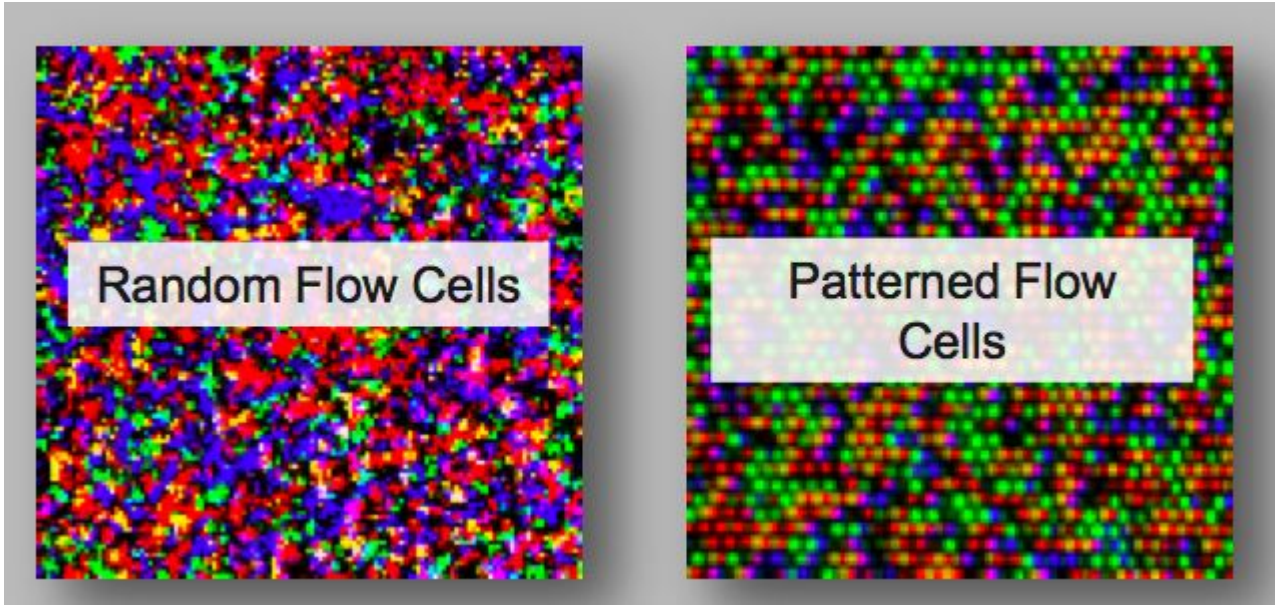
How to evaluate sequencing quality?

FastQC Documentation

Adapter Removal



Random vs Patterned Flow Cells



https://core-genomics.blogspot.com/2016/01/almost-everything-you-wanted-to-know.html