NGS Techniques

Practical workshop on Large-Scale Genomic Data Analyses:
GWAS in structured populations

November 27th, 2018

Selene L. Fernández-Valverde

regRNAlab.github.io

@SelFdz

Learning objectives

In this class we will learn

- How high-throughput (NGS) sequencing technologies arose
- How NGS technologies transformed our capacity to acquire large amounts of genomic information ‘
- Get acquainted with the common NGS techniques available in the market

The sequencing revolution

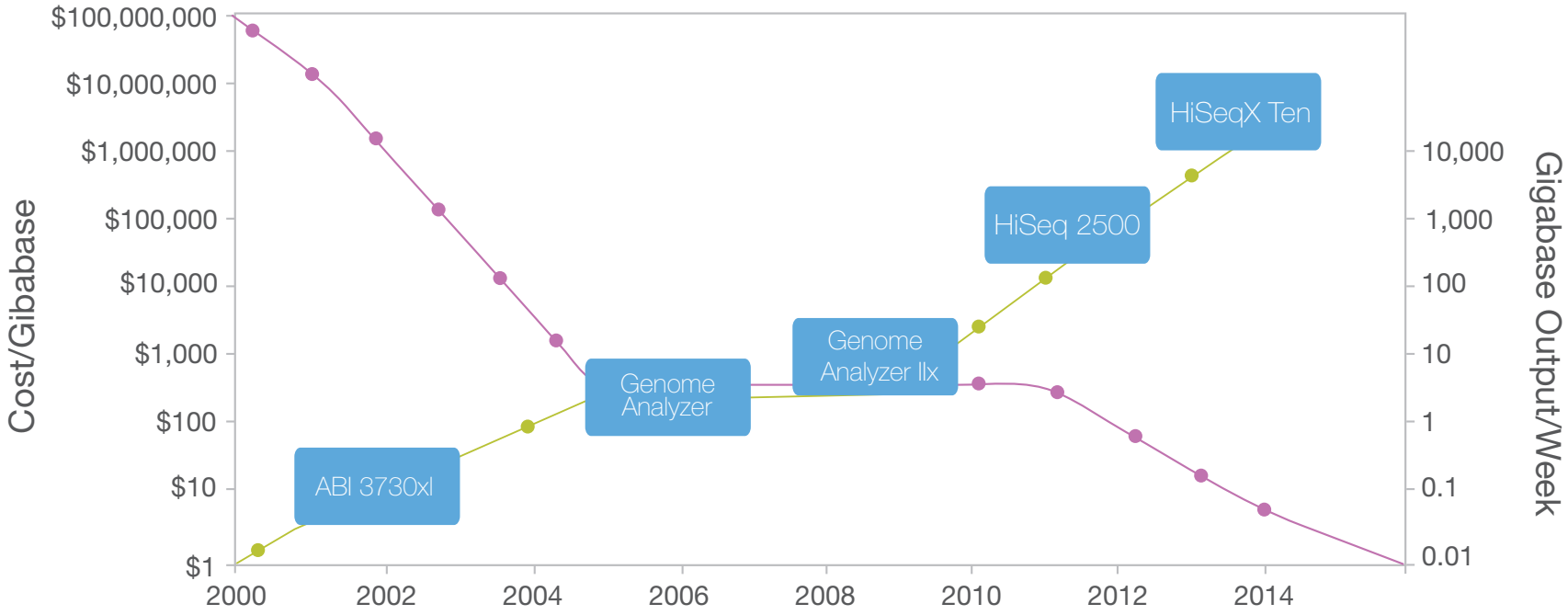


Figure 1: Sequencing Cost and Data Output Since 2000—The dramatic rise of data output and concurrent falling cost of sequencing since 2000. The Y-axes on both sides of the graph are logarithmic.

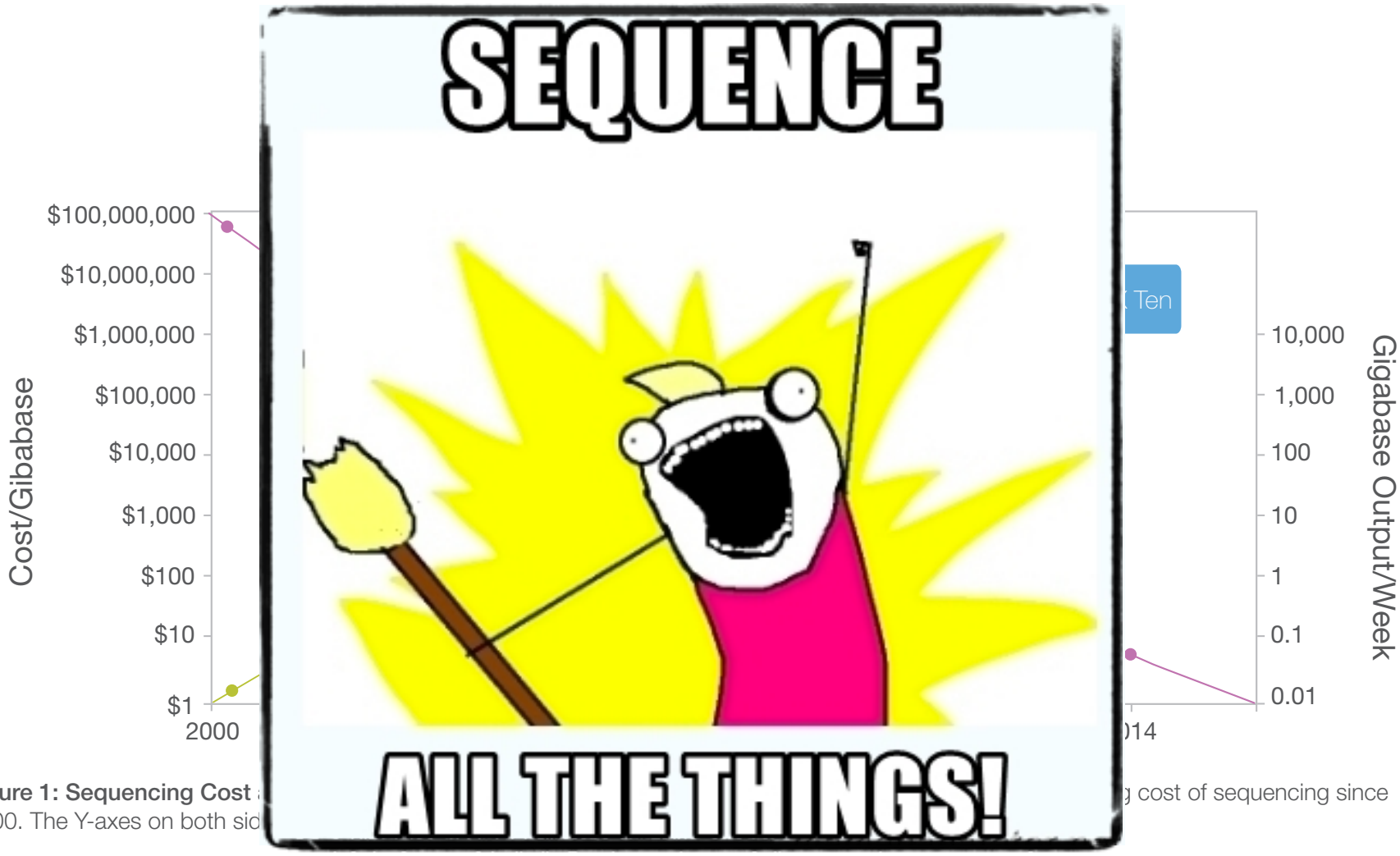
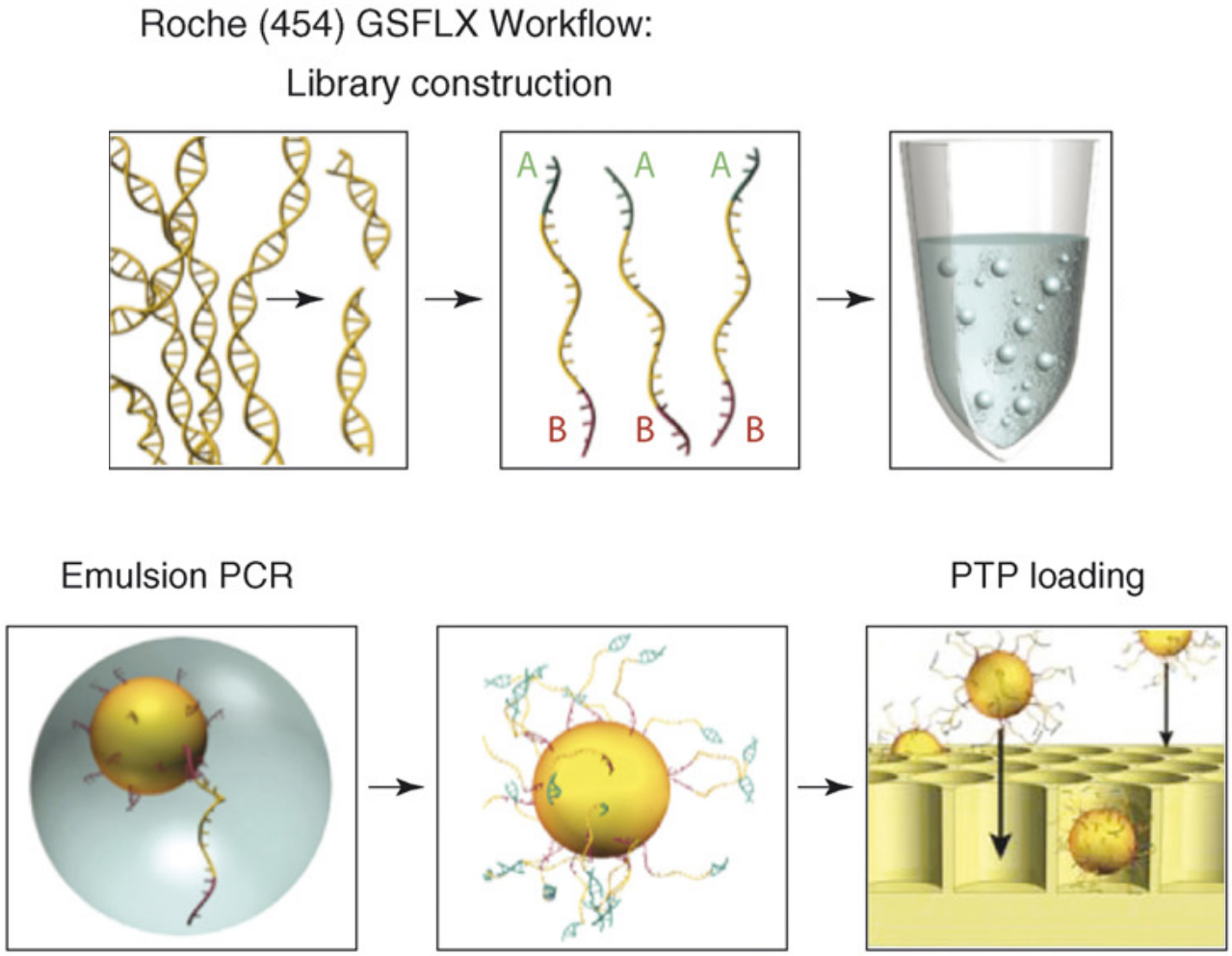


Figure 1: Sequencing Cost g cost of sequencing since 2000. The Y-axes on both sid

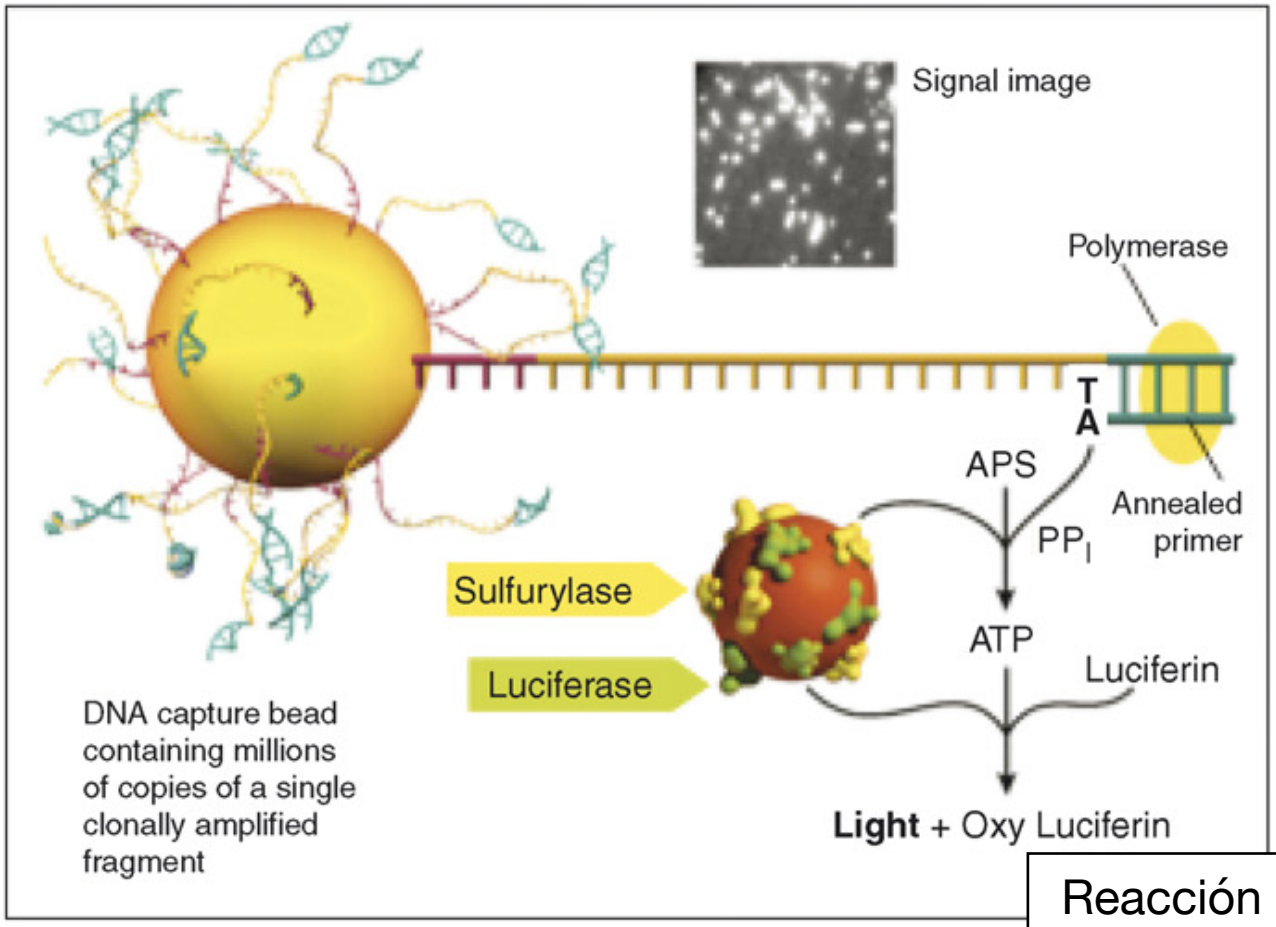
High-throughput sequencing techniques

- **Pyrosequencing**
- **Sequencing by synthesis**
- Sequencing by ligation
- Ion semiconductor
- **Nanopore sequencing**
- **Single Molecule Real Time Sequencing (SMRT)**

Pyrosequencing - 1



Pyrosequencing - 2



Pyrosequencing reaction

Reacción enzimatica chemoluminiscente

Pyrosequencing

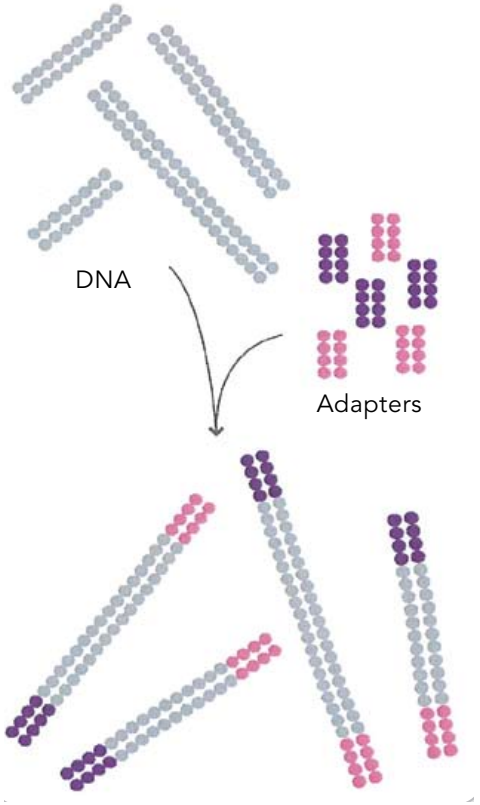
Advantages

- Reasonable cost
- Long sequences (500 nts)

Disadvantages

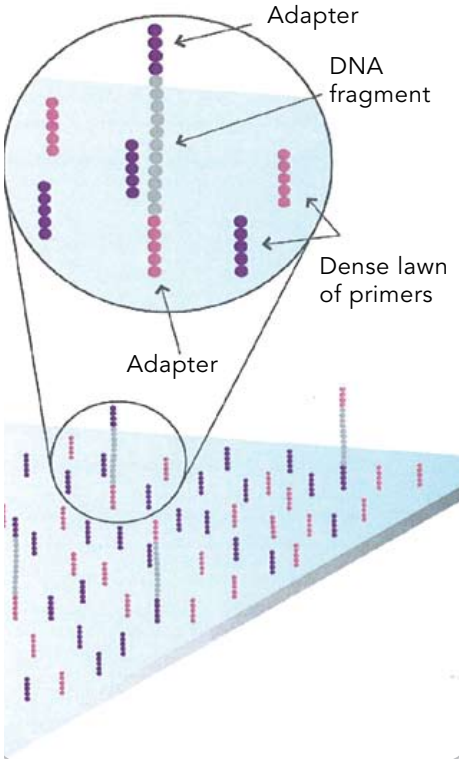
- Few sequences produced
- High number of errors in regions with the same nucleotide (homopolymers)
- With the rise of other technologies and given its high level of errors it was ultimately discontinued

Illumina - sequencing by synthesis - 1



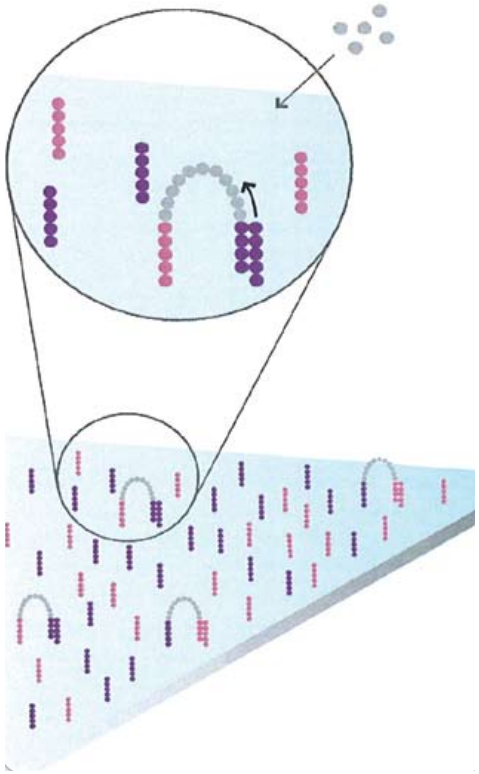
- The process starts by joining adapters to the DNA or RNA fragments that we want to sequence.

Illumina - sequencing by synthesis - 2



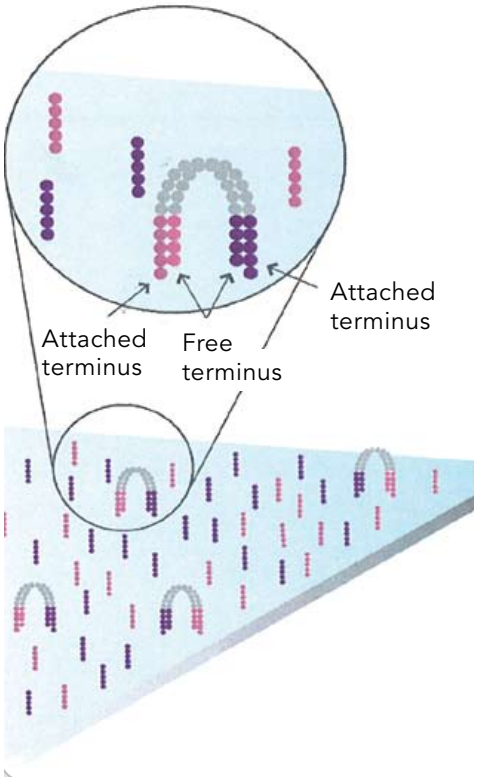
- The templates are immobilized on a flow cell
- In the case of RNA-Seq, complementarity with the adapter is used to synthesize a new cDNA chain in order to preserve information about the directionality of the transcript.

Illumina - sequencing by synthesis - 3



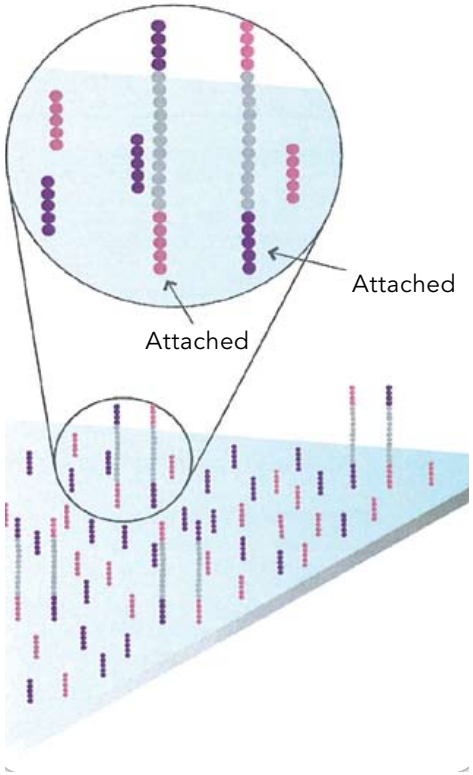
- A chain of DNA complementary to the DNA template is synthesized on the flow cell surface.

Illumina - sequencing by synthesis - 4



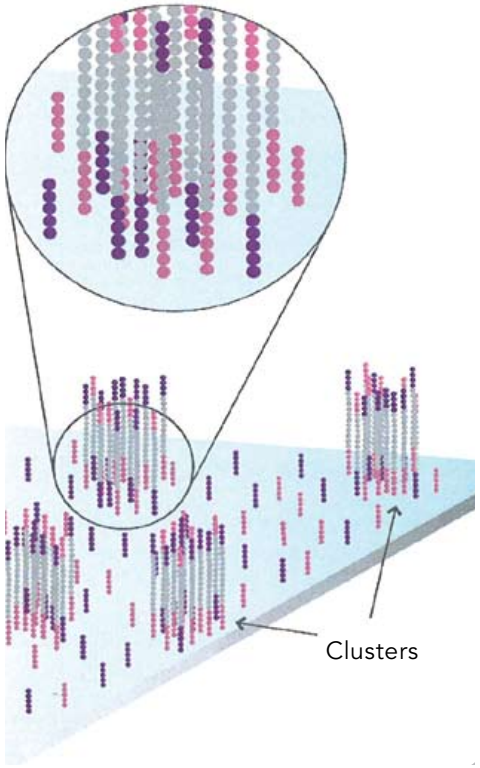
- A chain of DNA complementary to the DNA template is synthesized on the flow cell surface.

Illumina - sequencing by synthesis - 5



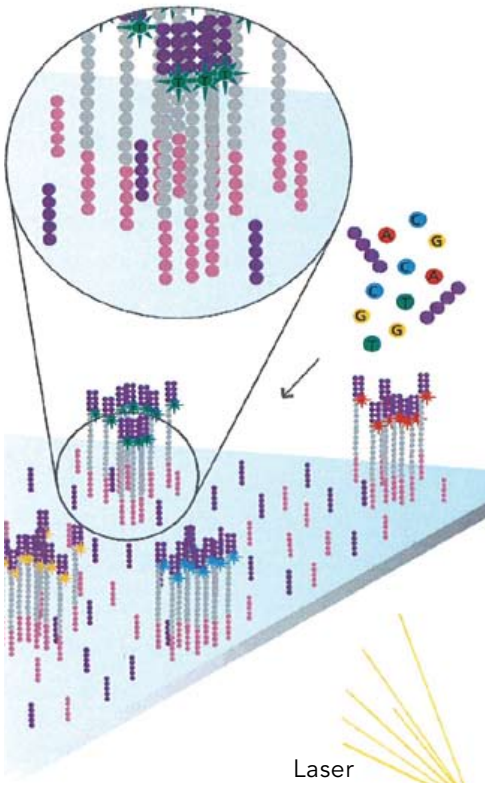
- The templates are separated using high temperature.

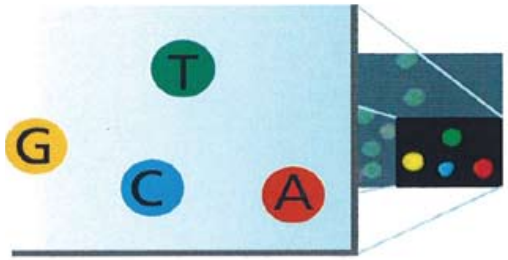
Illumina - sequencing by synthesis - 6



- This process is repeated hundreds of times until generating a "colony" or cluster of identical transcripts.

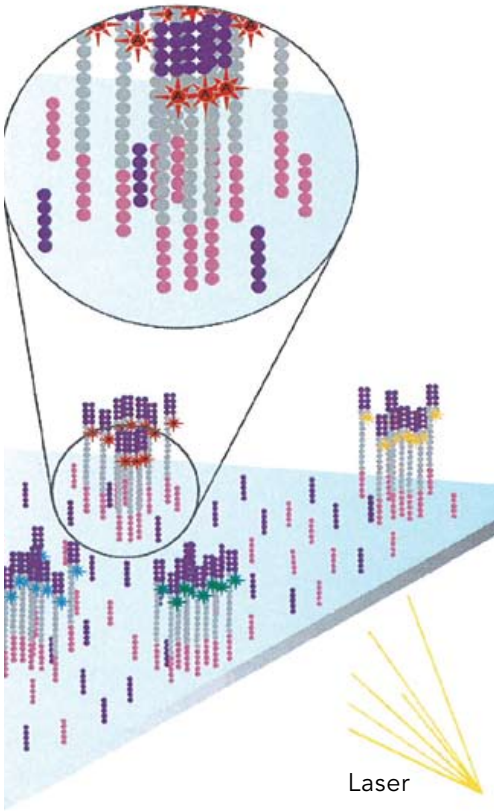
Illumina - sequencing by synthesis - 7

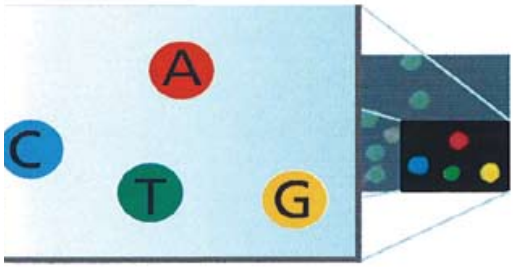




- Primers and fluorescent nucleotides (reversible terminators) are added in order (first A, then T, etc.) along with polymerase. When a nucleotide is incorporated a laser pulse coupled with imaging are used to identify which base was incorporated in each position.

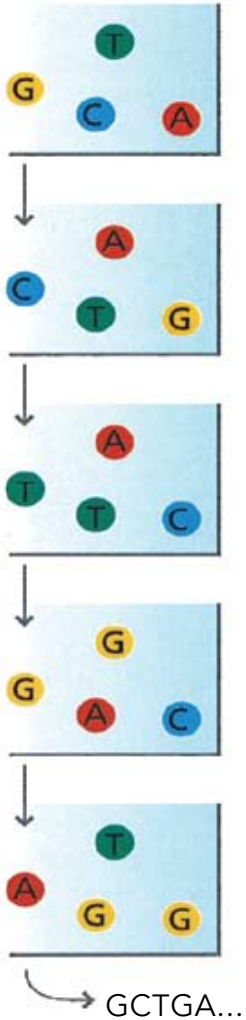
Illumina - sequencing by synthesis - 8

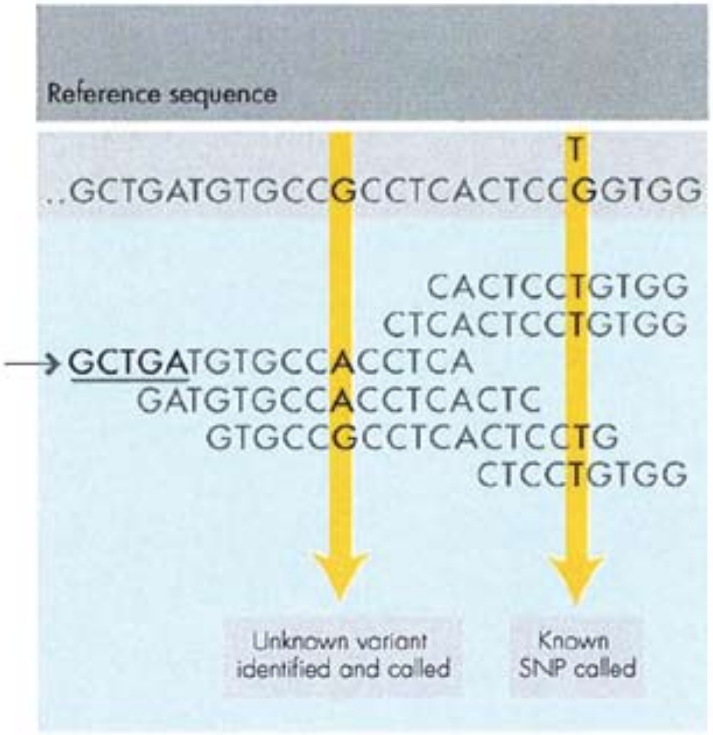




- This process is continued for all bases.

Illumina - sequencing by synthesis - 9





- The images are analyzed spatially to reveal each sequence.

Sequencing by Synthesis

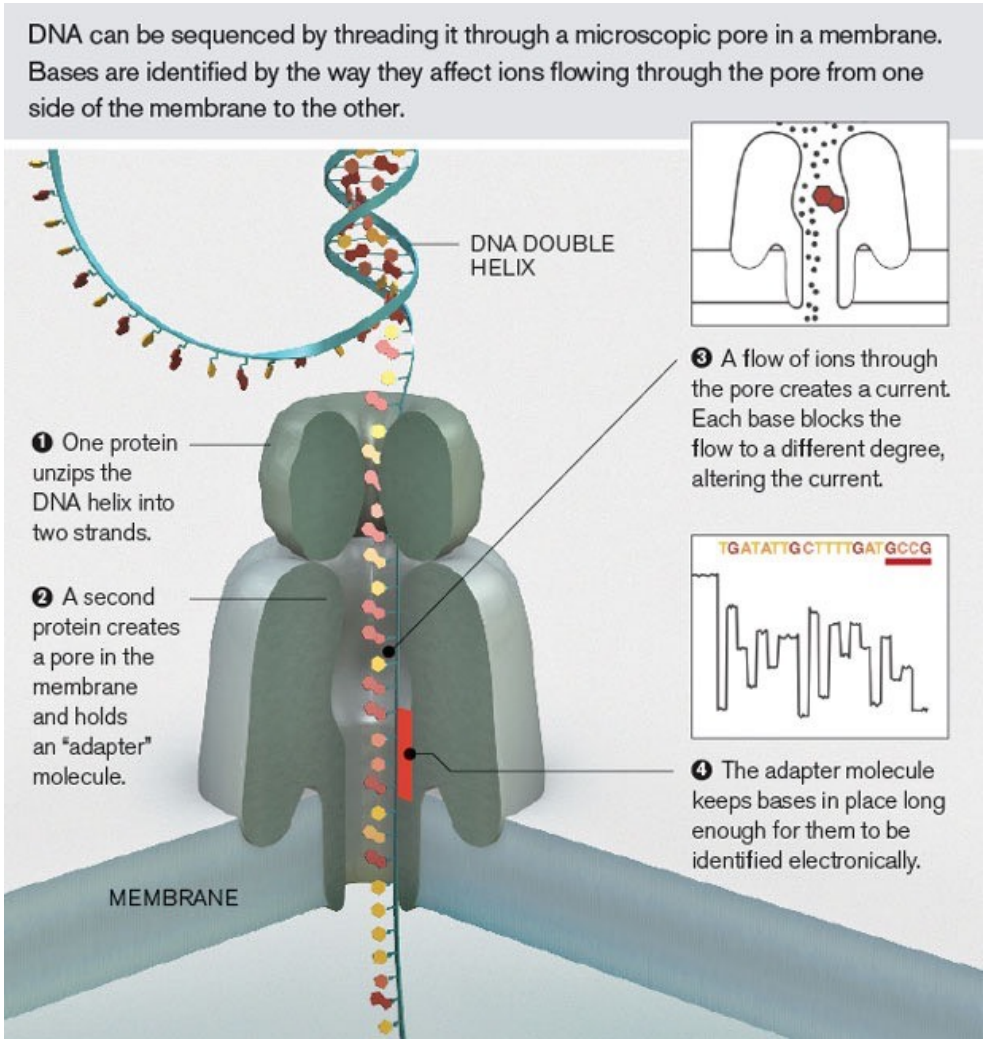
Advantages

- Undoubtedly the leader in the market = strong scientific support network
- Produces large amounts of sequences (Up to 20 billion for NovaSeq)
- Low error rate compared with other technologies

Disadvantages

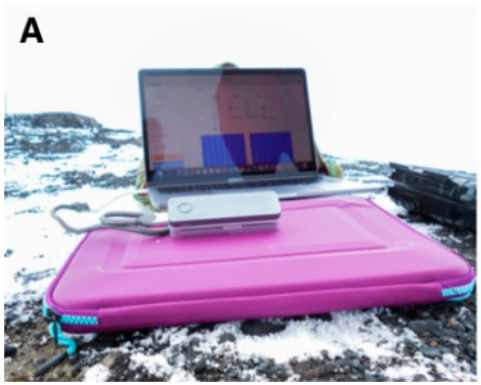
- The sequences are short (150 to 300 bp)
- The cost is high
- Relatively slow sequencing (13–44 hr for NovaSeq)

Nanopore sequencing

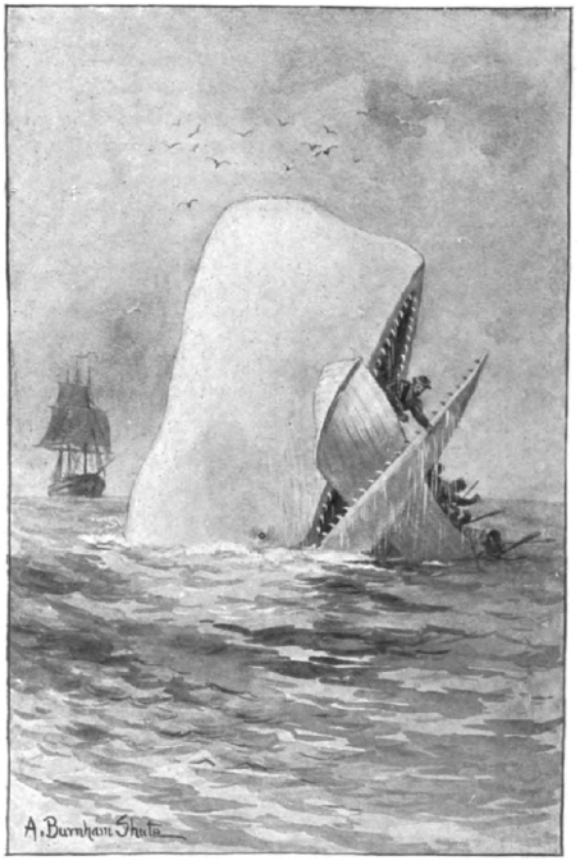


Nanopore sequencing

Kate Rubins



Nanopore whale watching



"Both jaws, like enormous shears, bit the craft completely in twain."

—*Page 510.*

- Nanopore is capable of generating very very long reads or "whales"
- The longest read detected to date has a length of 2,272,580 bases

Nanopore sequencing

Advantages

- Real-time sequencing
- You can stop sequencing when you have enough data
- Very portable - useful for work in difficult areas
- Simple preparation
- Low cost - $ 80 USD per sample

Disadvantages

- High number of errors although they have had a drastic increase in accuracy in the last year
- Pores failed - sequence loss

Sources of error

- There are two main sources of error:
  - **Human error:** mixing of samples (in the laboratory or when the files were received), errors in the protocol
  - **Technical error:** Errors inherent to the platform (e.g., mononucleotide sequences in pyrosequencing) - All platforms have some level of error that must be taken into account when designing the experiment.

Errors in sample preparation

- User error (e.g. mistakenly labeling a sample)
- DNA / RNA degradation by preservation methods
- Contamination with external sequences
- Low amount of DNA start

Errors in library preparation

- User error (e.g. polluting one sample with another, contaminate with previous reactions, errors in the protocol)
- PCR amplification errors
- Bias for primers (binding bias, methylation bias, primer dimers [first dimers])
- Bias for capture (Poly-A, Ribozero)
- Machine errors (misconfiguration, reaction interruption)
- Chimeras
- Index errors, adapter (contamination of adapters, lack of index diversity, incompatible codes (barcodes), overload)

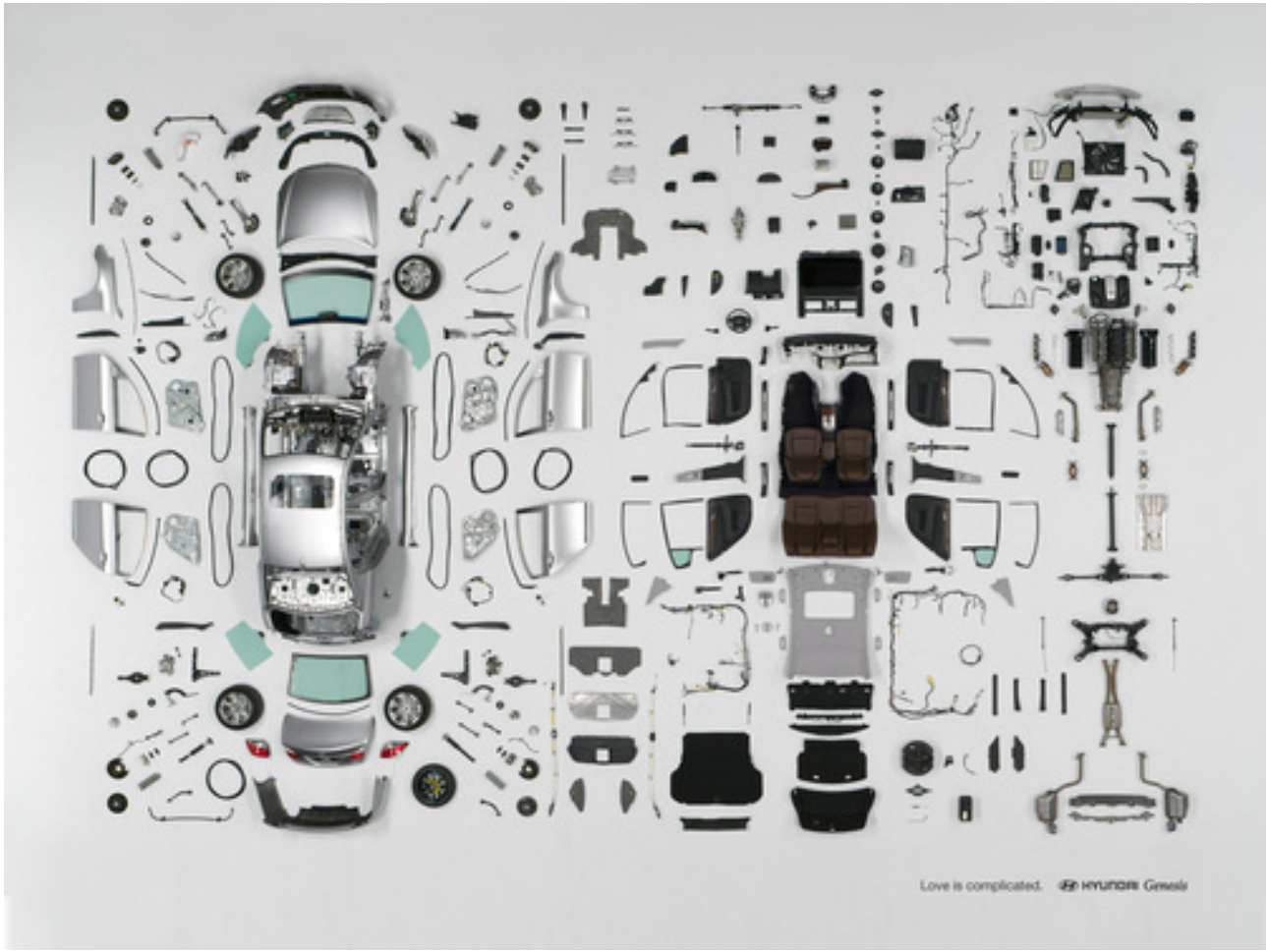
Sequencing and image errors

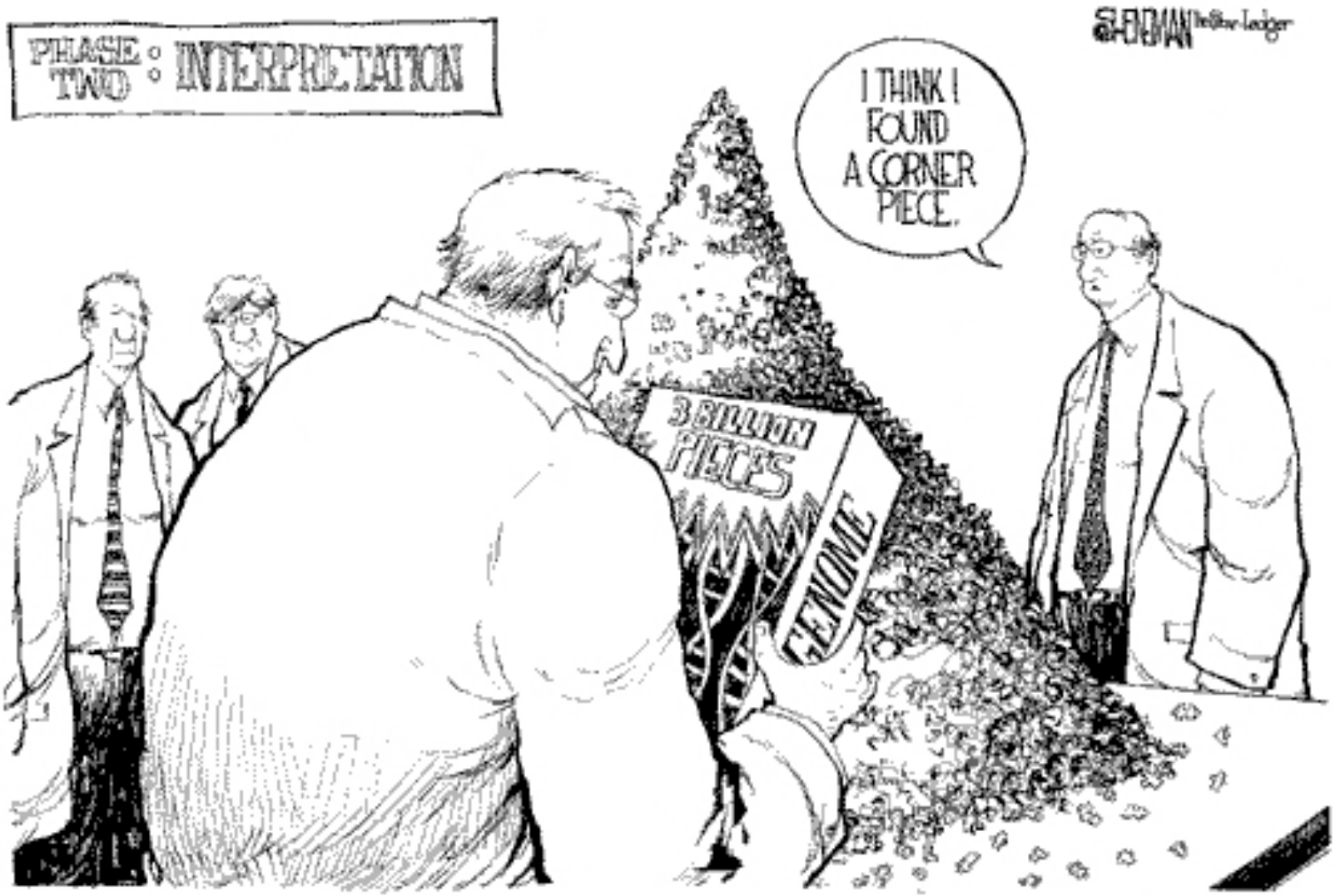
- User error (e.g. cell overload)
- Delay (e.g., incomplete extension, addition of multiple nucleotides)
- Dead fluorophores, damaged nucleotides and overlapping signals
- Context of the sequence (e.g. high GC content, homologous and low complexity sequences, homopolymers).
- Machine errors (e.g. laser, hard disk, programs)
- Chain biases

The challenge - differentiate biological signals from noise/errors

- Negative and positive controls - What do I expect?
- Technical and biological replicas - help determine the noise rate
- Know the types of common errors in a certain platform

Now what?




PHASE TWO: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

Practical - Fastq format and QC of NGS data

https://liz-fernandez.github.io/MxBiobank_NGS/01-quality.html