INSTITUTE OF PLANT SCIENCES

Sant'Anna
School of Advanced Studies – Pisa

Advanced Genomics

MPS and Illumina technology

First Generation

Sanger Sequencing
Maxam and Gilbert
Sanger Chain-termination

- Infer nucleotide identity using dNTPs then visualize with electrophoresis
- 500-1000 bp fragments

Second Generation
(Next Generation Sequencing)

454, Solexa,
Ion Torrent
Illumina

- High throughput from the parallelization of sequencing reactions
- ~50-500 bp fragments

Third Generation

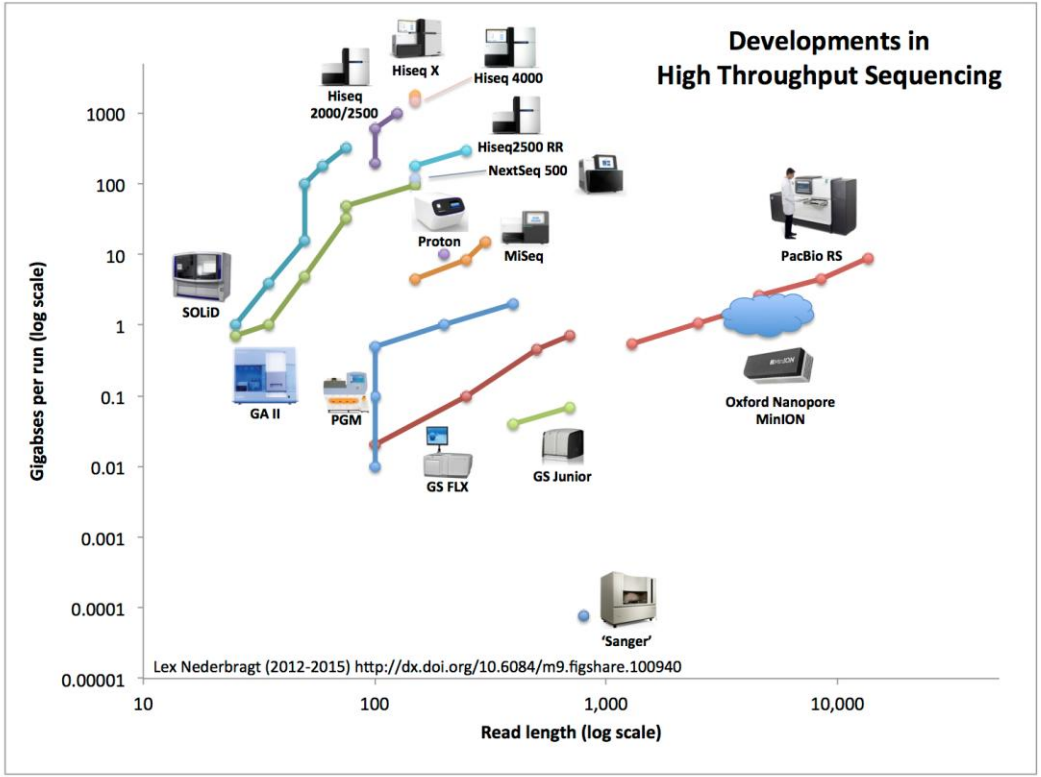
PacBio
Oxford Nanopore

- Sequence native DNA in real time with single-molecule resolution
- Tens of kb fragments, on average

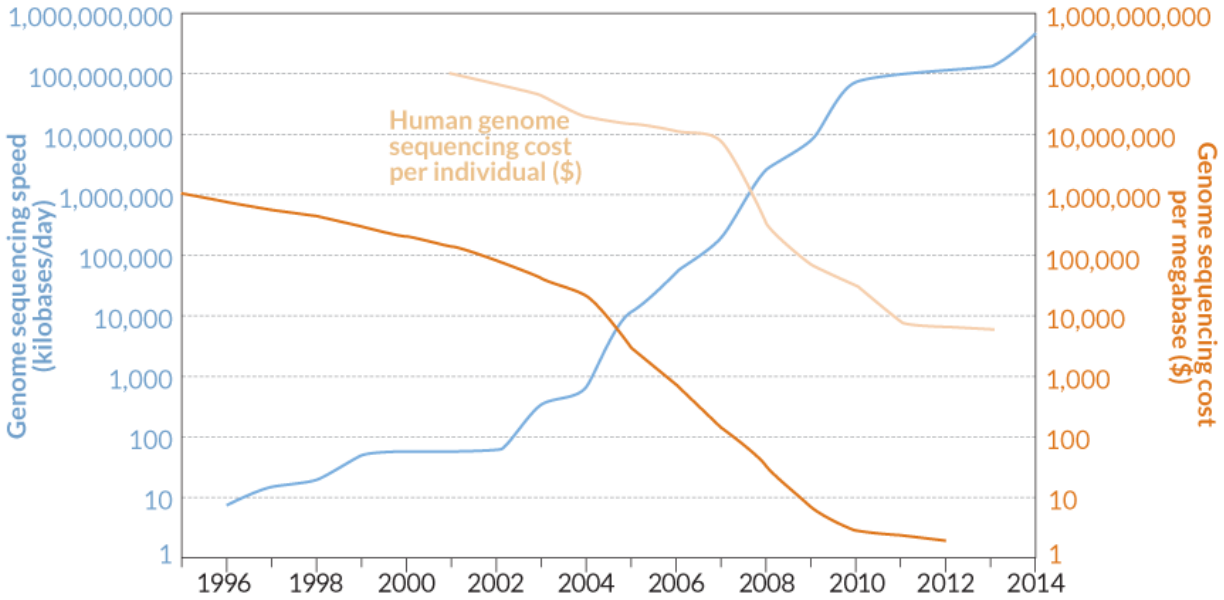
Short-read sequencing

Long-read sequencing

An ever-expanding array of sequencing tools



There is a tight relation between throughput and costs



Grandfather and Grandson

Platfrom	ABI 3730	HiSeq2500 HO (v4)
Method	Sanger	Illumina
Capacity (bp/hour)	76.000	7.000.000.000
Cost (€/Gb)	1.250.000	32

~90.000 ABI3730 vs 1 HiSeq2500

Massive parallel sequencing

- AKA Next Generation Sequencing (NGS) or second generation sequencing
- A step forward towards overcoming the bottleneck of data generation with Sanger method
- Use of miniaturized and parallelized platforms to sequence billions of fragements at once
- The idea is to start from spatially separated, clonally amplified DNA fragments in flowcells (no electrophoresis, no chain termination)

Key competing technologies

- Pyrosequencing (Roche)
  - Bioluminescence method that measures the release of inorganic pyrophosphate by proportionally converting it into visible light
  - DNA polymerase runs on single additions of a dNTP in limiting amounts. Upon incorporation of the complementary dNTP, DNA polymerase extends the primer and pauses
  - The order and intensity of the light peaks are recorded as flowgrams, which reveal the underlying DNA sequence
- Dye sequencing (Illumina-Solexa)
  - Continuous incorporation of reversible-termination and dye-labelled nucleotides during DNA synthesis
  - A camera takes images of the fluorescently labeled nucleotides
  - The dye, along with the terminal 3' blocker, is chemically removed from the DNA, allowing for the next cycle to begin

Illumina

NovaSeq X Series, Sept 2022

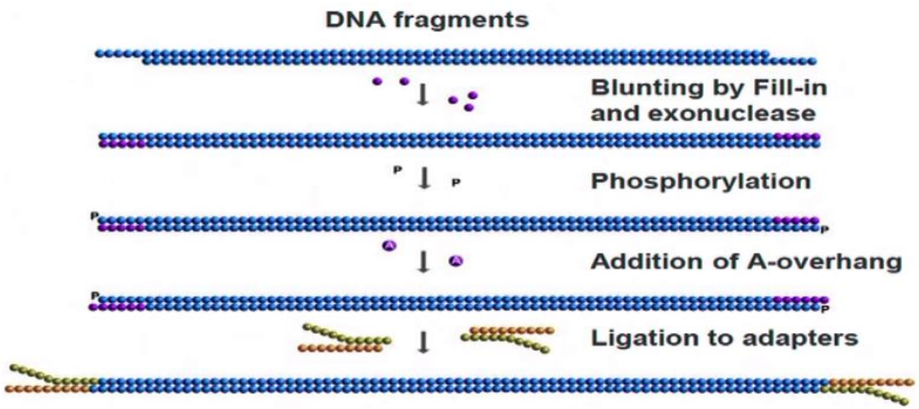
- Sequencing by synthesis
- Multiple single molecules (library) are attached to a surface (lane or beads)
- An amplification phase is used to increment number of molecules, necessary to increase signal-to-background ratio
- Sequence information is derived by imaging of fluorescence

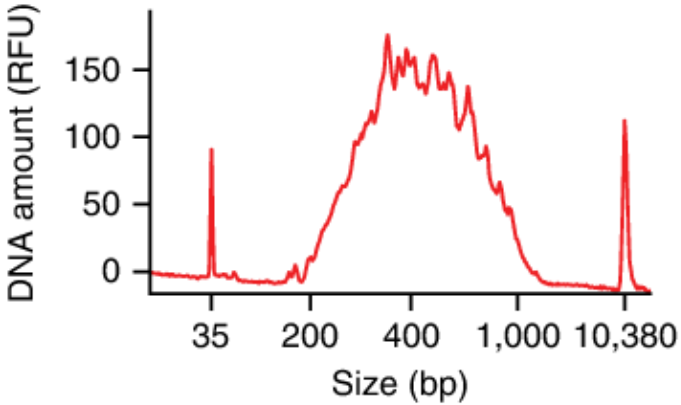
Key steps:

- Library preparation
- Cluster generation (amplification)
- Sequencing

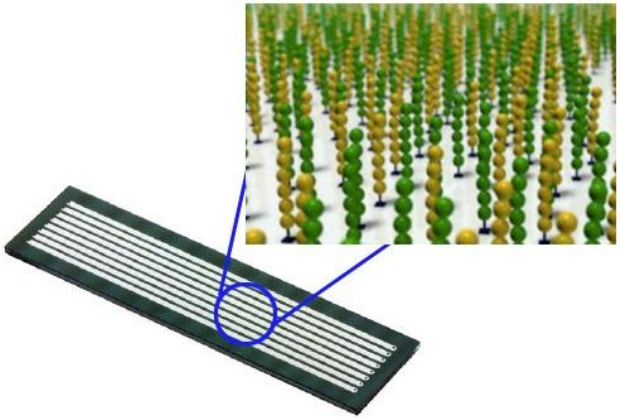
Library preparation

- Fragmenting DNA by sonication (or enzymatic reaction)
- Making DNA ends blunt
- Ligating syntethic DNA adapter (tipically with a barcode)
- Quantification of the library
- Loading on the sequencing instrument (whole library or subset)



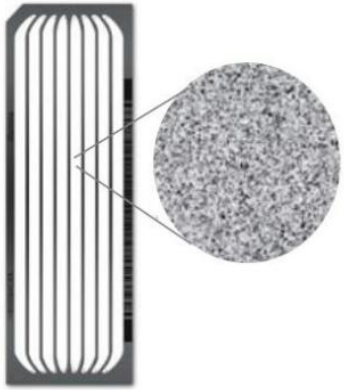


At the core of Illumina sequencing: the flow cell



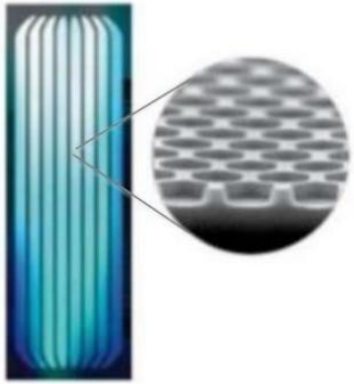
Random Flow Cell

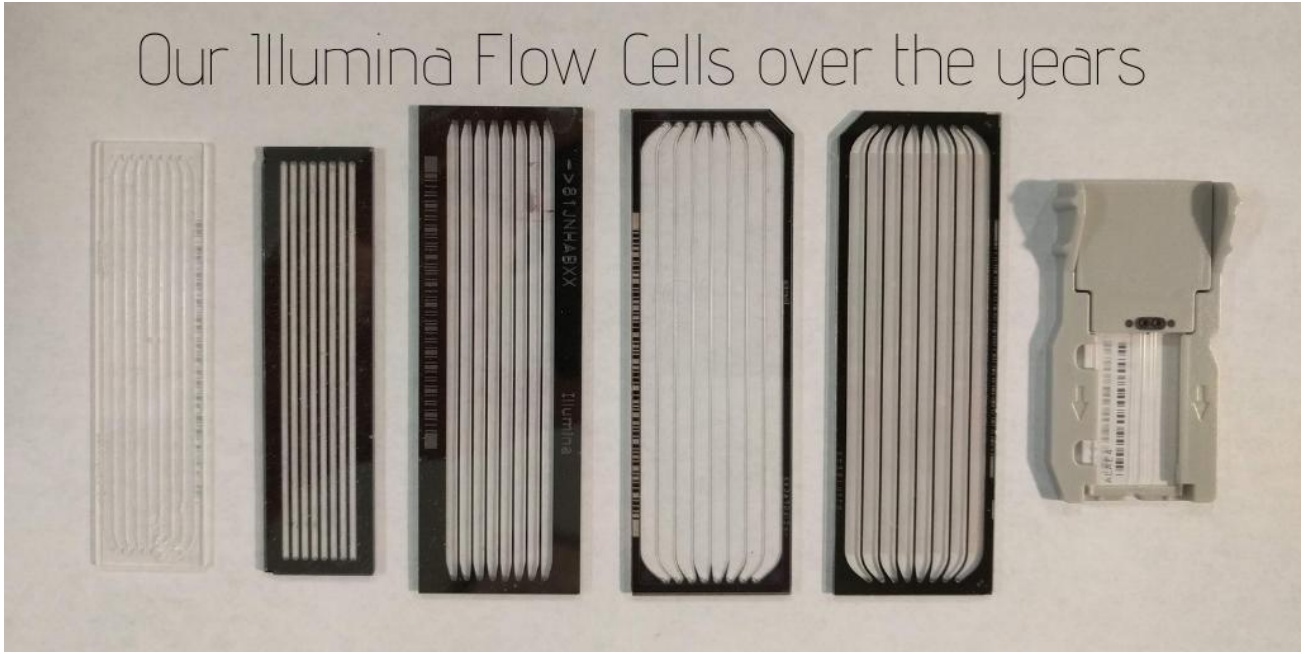
- HiSeq 2500, MiSeq, NextSeq, MiniSeq
- Randomly spaced clusters
- Variable Insert Sizes
- Lower Duplication Rates

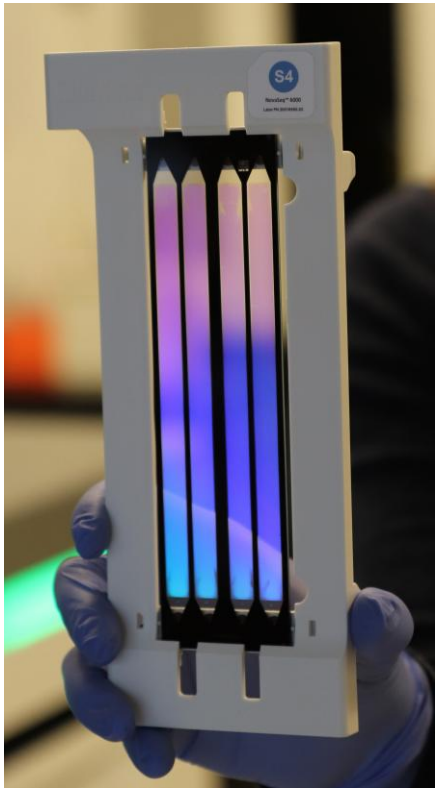


Patterned Flow Cell

- HiSeq 3K/4K/X, NovaSeq, iSeq 100
- Defined size and spacing
- Increased Cluster density
- Simplified imaging



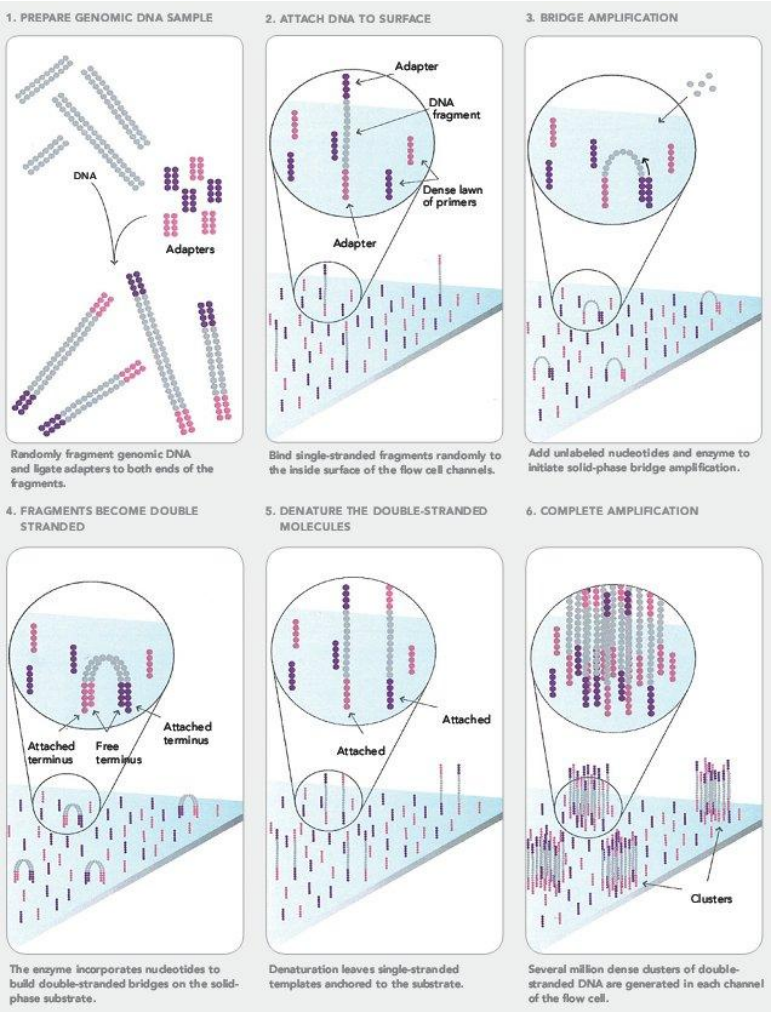


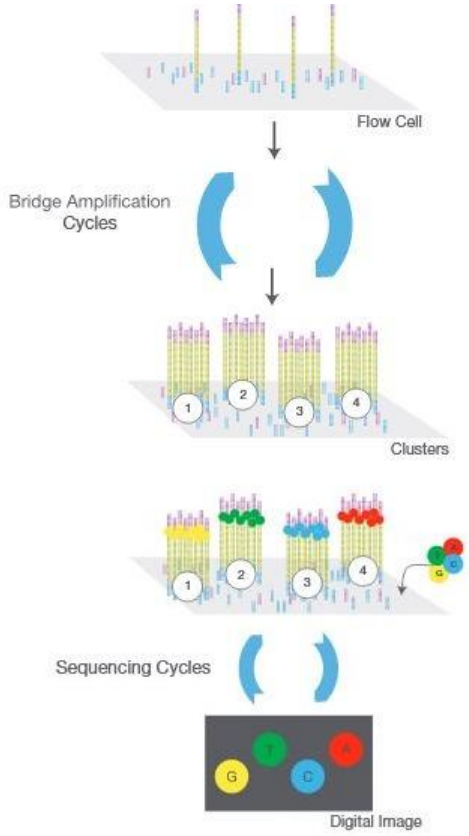


Novaseq 6000

Cluster generation

- The library is loaded on the target surface, to which it gets attached thanks to oligos complementary to adapters
- The position is random – it doesn't matter, since you have barcodes that you will eventually read in the generated sequence
- Bridge amplification is performed to increase the number of molecules





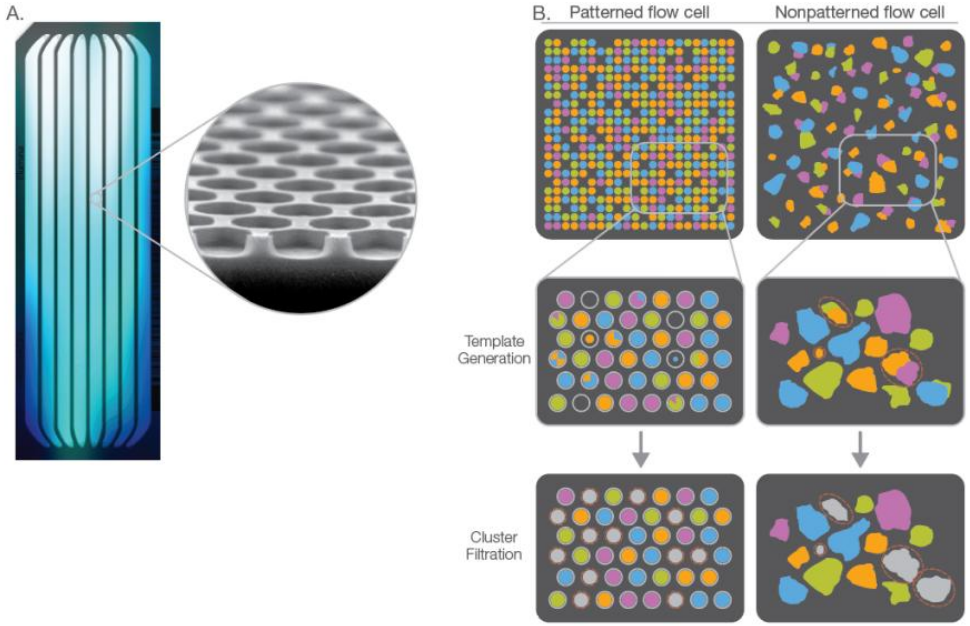
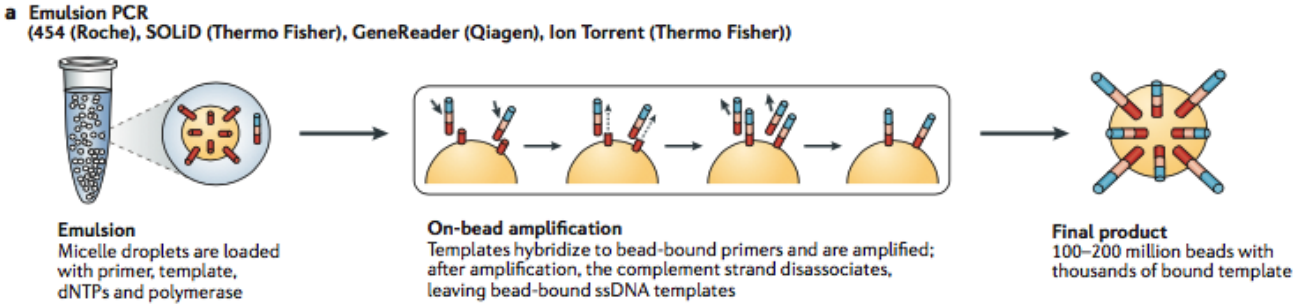


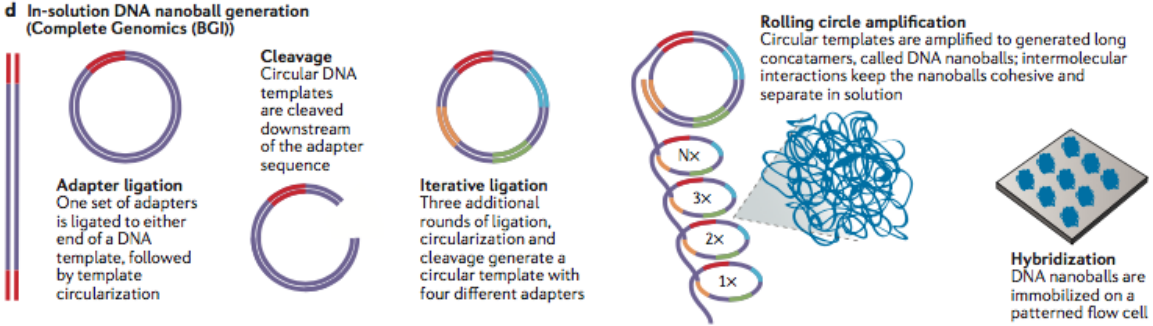
Figure 7: Clusters passing filter on patterned and nonpatterned flow cells—A patterned flow cells with nanowells etched into its surface (A). With nonpatterned flow cells, poor quality or dim clusters are filtered during template generation (B). With patterned flow cells, empty wells and suboptimal clusters are filtered during the later stage of chastity filtration, which leads to a lower %PF metric (B).

Alternative clustering strategies

- Bead-based

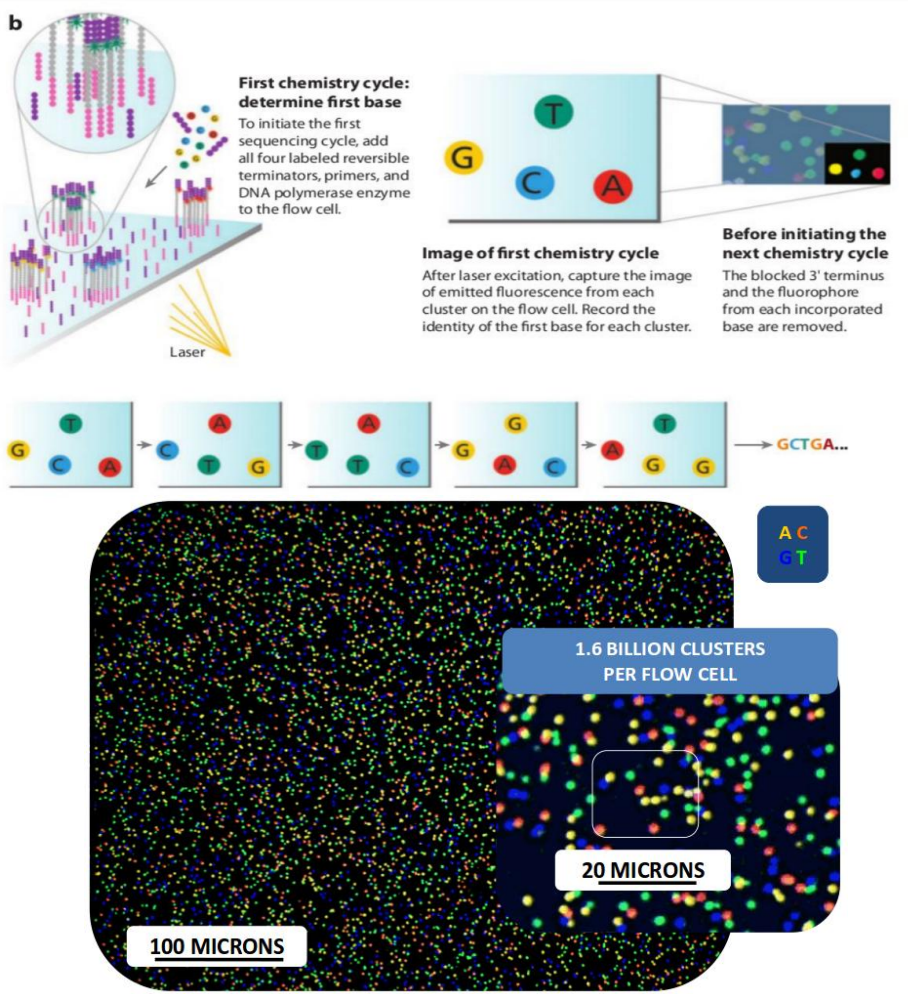


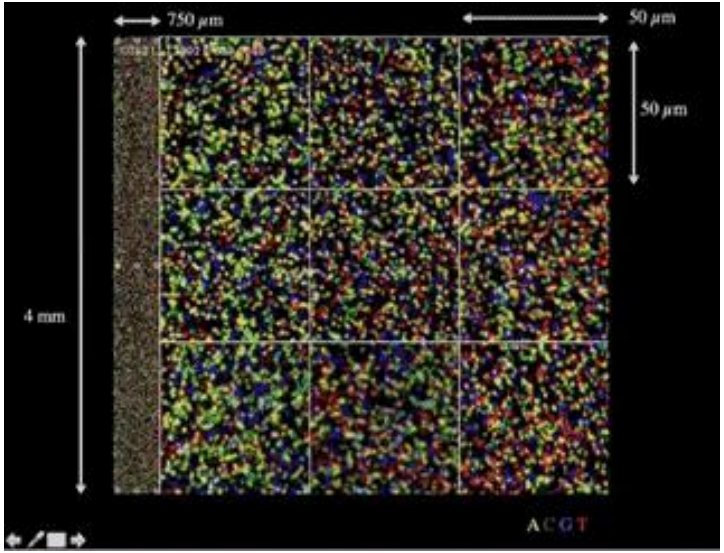
- DNA nanoball generation

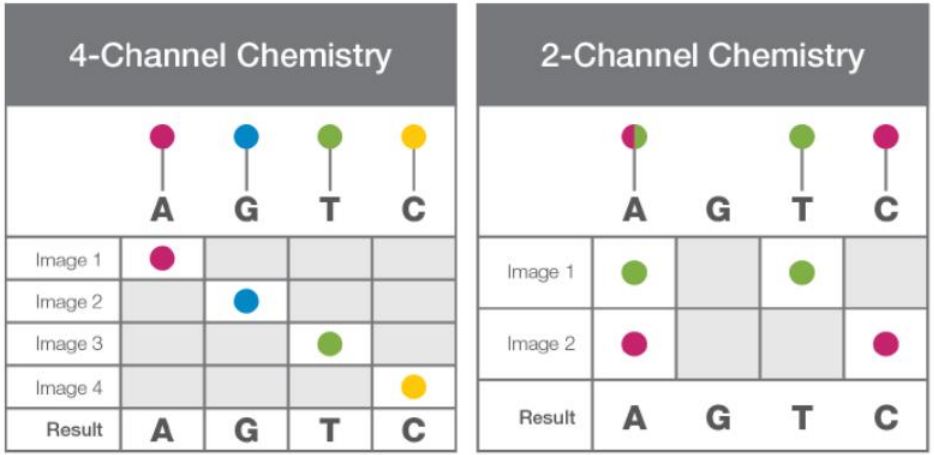


Sequencing by synthesis

- A polymerase reaction is used to extend fragments from free primers
- Nt are labeled with different colours (2-channels or 4-channel chemistry)
- When a Nt is incorporated, it emits the corresponding light color
- A laser captures the light emission and maps it to a cluster




750 μm
50 μm
50 μm
4 mm
A C G T



Four-channel SBS

- Bases are identified using four different fluorescent dyes, one for each base and four images per sequencing cycle

Two-channel SBS

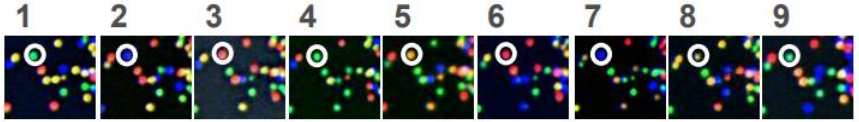
- Simplified nucleotide detection by using two fluorescent dyes and two images to determine all four base calls

Four Channels SBS:

- MiSeq

Two Channels SBS:

- MiniSeq, NextSeq, NovaSeq
- Accelerates sequencing and data processing times



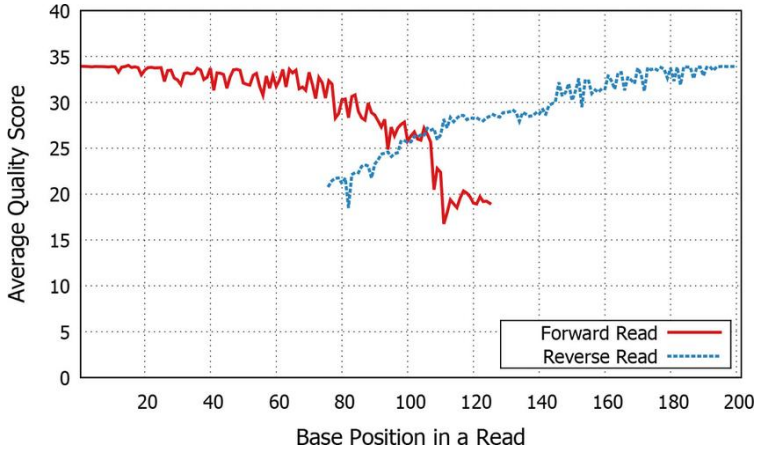
→ TGATCAGCT

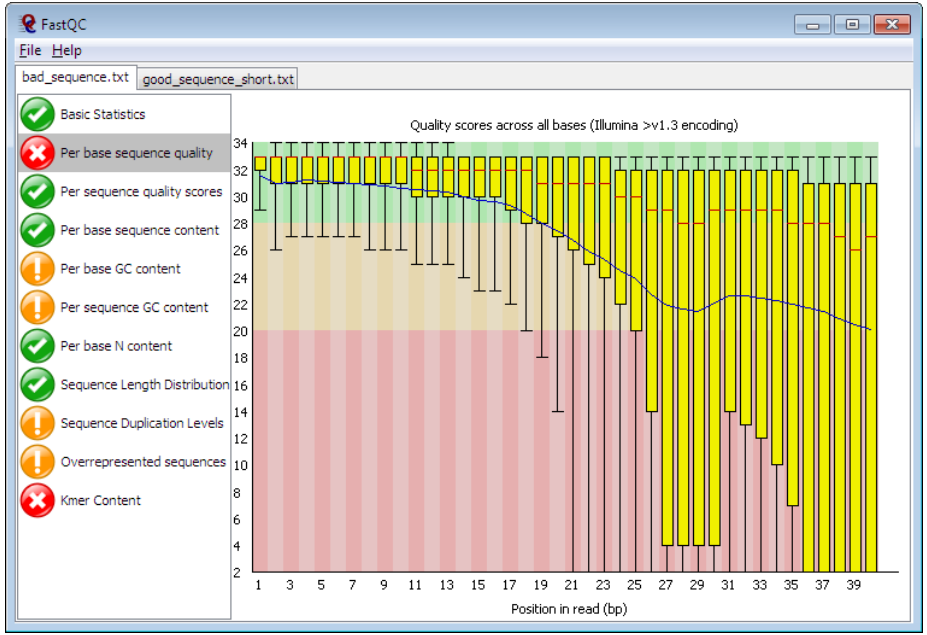
- In any sequencing technology, PCR amplify the individual DNA fragments once they have hybridized to flowcells or beads
- you end up with *both* strands of DNA. If you were to read both of the strands from their respective 3′ ends at once, you’d be getting two different sequences.
- To avoid this problem, sequencing technologies ligate non-complementary adapters to the 3′ and 5′ ends of DNA fragments so that the primer for one adapter only begins synthesis on one strand and not on its complement.

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Why short reads?

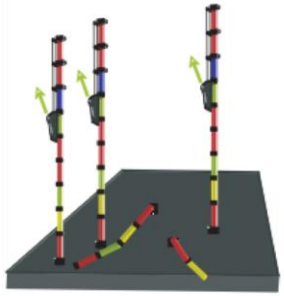
- For reasons related to chemistry, quality is tipically lower at the end of the read

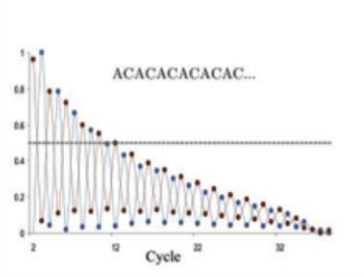




- Tools exist to check many quality parameters and rise red flags

Loss of quality at 3’



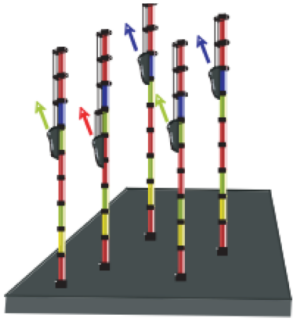


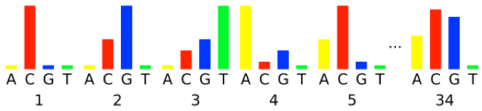
Signal decay

- As number of cycles incrases, fluorophores decay and clusters have uneven amplification

Loss of sinchronicity

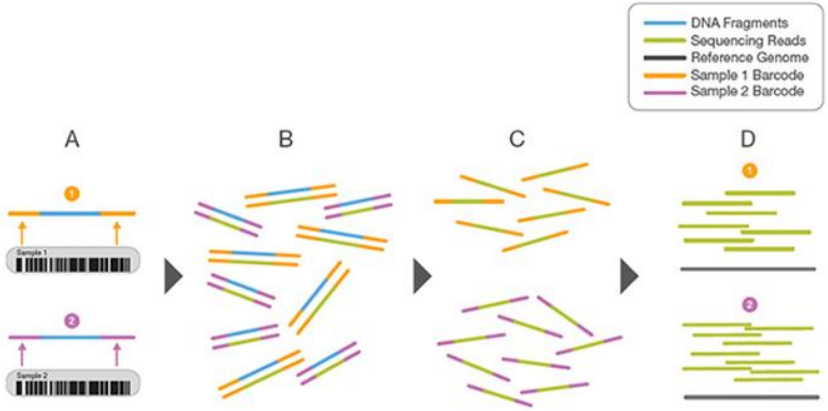
- As amplification progresses, reads in the cluster loose sinchronicity and the signal becomes blurred



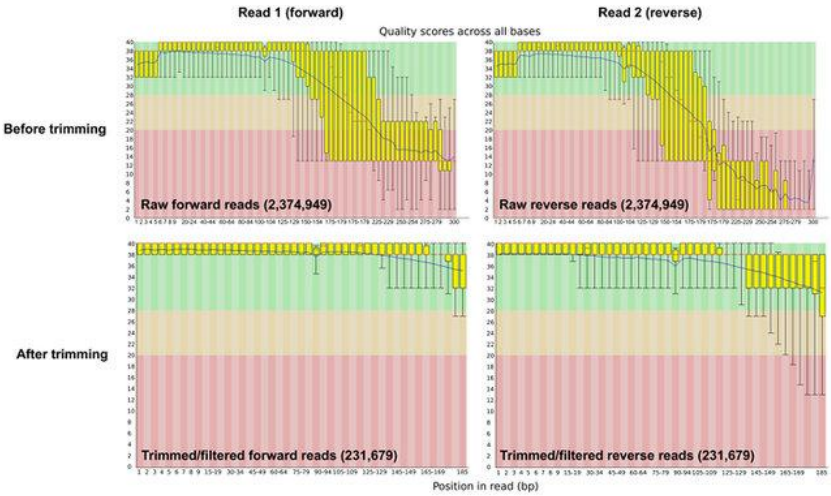


Raw reads need to be trimmed

- In silico approaches are used to remove adapters, sort reads by barcode, and drop low quality nucleotides

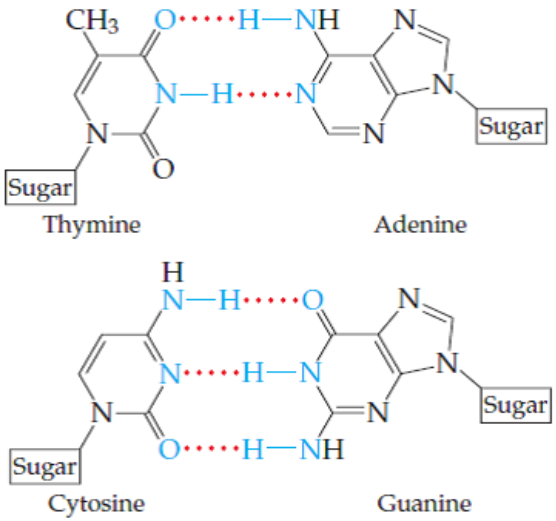


A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.



The issue of GC bias

- Being based on clonal amplification, Illumina sequencing suffers from bias caused by CG content in PCR steps
- GC rich regions can be biologically important (e.g. promoters and coding regions)
- GC abundance is heterogeneously distributed throughout the genome and is frequently correlated with functionality



Huttener *et al. BMC Ecol Evo* (2021) 21:176
https://doi.org/10.1186/s12862-021-01905-7

BMC Ecology and Evolution

RESEARCH ARTICLE Open Access

Sequencing refractory regions in bird genomes are hotspots for accelerated protein evolution

R. Huttener[1†], L. Thorrez[1,2†], T. In't Veld[1†], M. Granvik[1], L. Van Lommel[1], E. Waelkens[3], R. Derua[3], K. Lemaire[1], L. Goyvaerts[1], S. De Coster[1], J. Buyse[4] and F. Schuit[1*]

Abstract

Background: Approximately 1000 protein encoding genes common for vertebrates are still unannotated in avian genomes. Are these genes evolutionary lost or are they not yet found for technical reasons? Using genome landscapes as a tool to visualize large-scale regional effects of genome evolution, we reexamined this question.

Results: On basis of gene annotation in non-avian vertebrate genomes, we established a list of 15,135 common vertebrate genes. Of these, 1026 were not found in any of eight examined bird genomes. Visualizing regional genome effects by our sliding window approach showed that the majority of these "missing" genes can be clustered to 14 regions of the human reference genome. In these clusters, an additional 1517 genes (often gene fragments) were underrepresented in bird genomes. The clusters of "missing" genes coincided with regions of very high GC content, particularly in avian genomes, making them "hidden" because of incomplete sequencing. Moreover, proteins encoded by genes in these sequencing refractory regions showed signs of accelerated protein evolution. As a proof of principle for this idea we experimentally characterized the mRNA and protein products of four "hidden" bird genes that are crucial for energy homeostasis in skeletal muscle: *ALDOA*, *ENO3*, *PYGM* and *SLC2A4*.

Conclusions: A least part of the "missing" genes in bird genomes can be attributed to an artifact caused by the difficulty to sequence regions with extreme GC% ("hidden" genes). Biologically, these "hidden" genes are of interest as they encode proteins that evolve more rapidly than the genome wide average. Finally we show that four of these "hidden" genes encode key proteins for energy metabolism in flight muscle.

Keywords: Avian genomes, Evolution, Accelerated, Sequencing artifacts, Transcript landscapes, Missing genes, GLUT4, *SLC2A4*, *ENO3*, *ALDOA*, *PYGM*

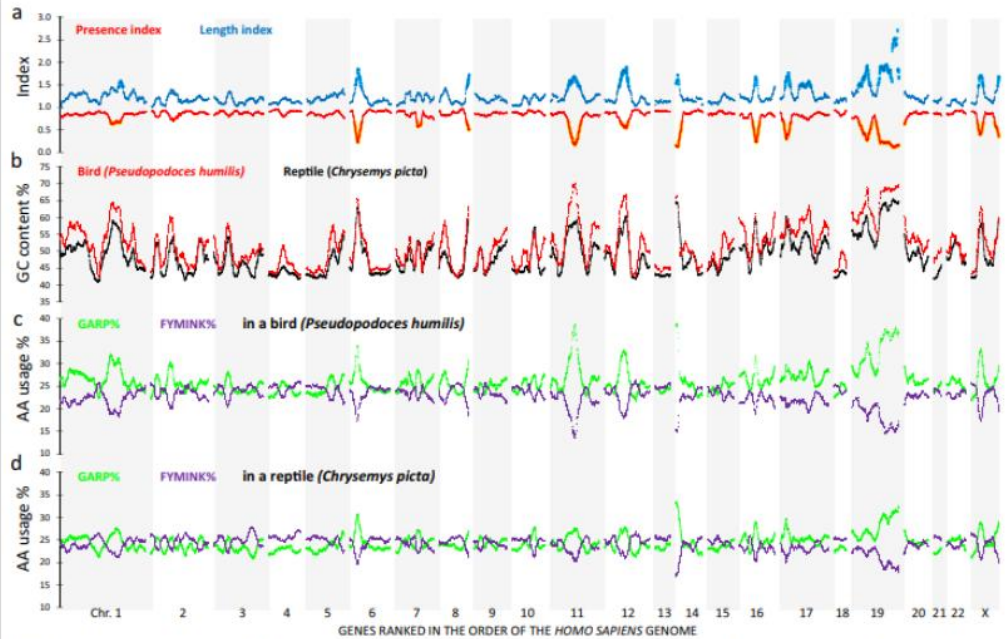
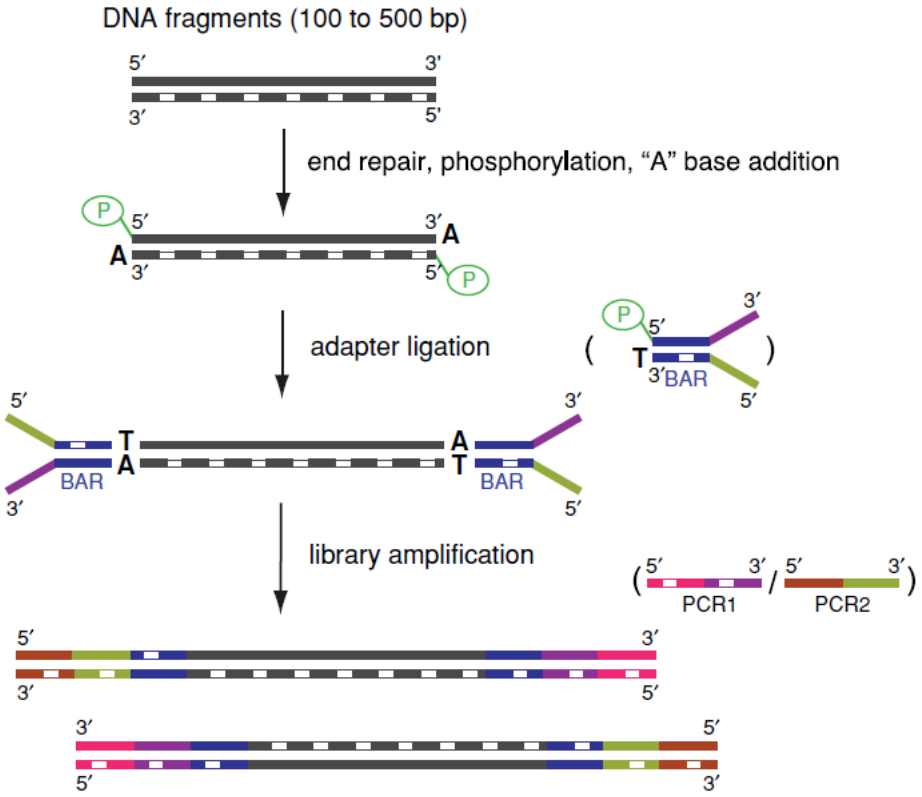
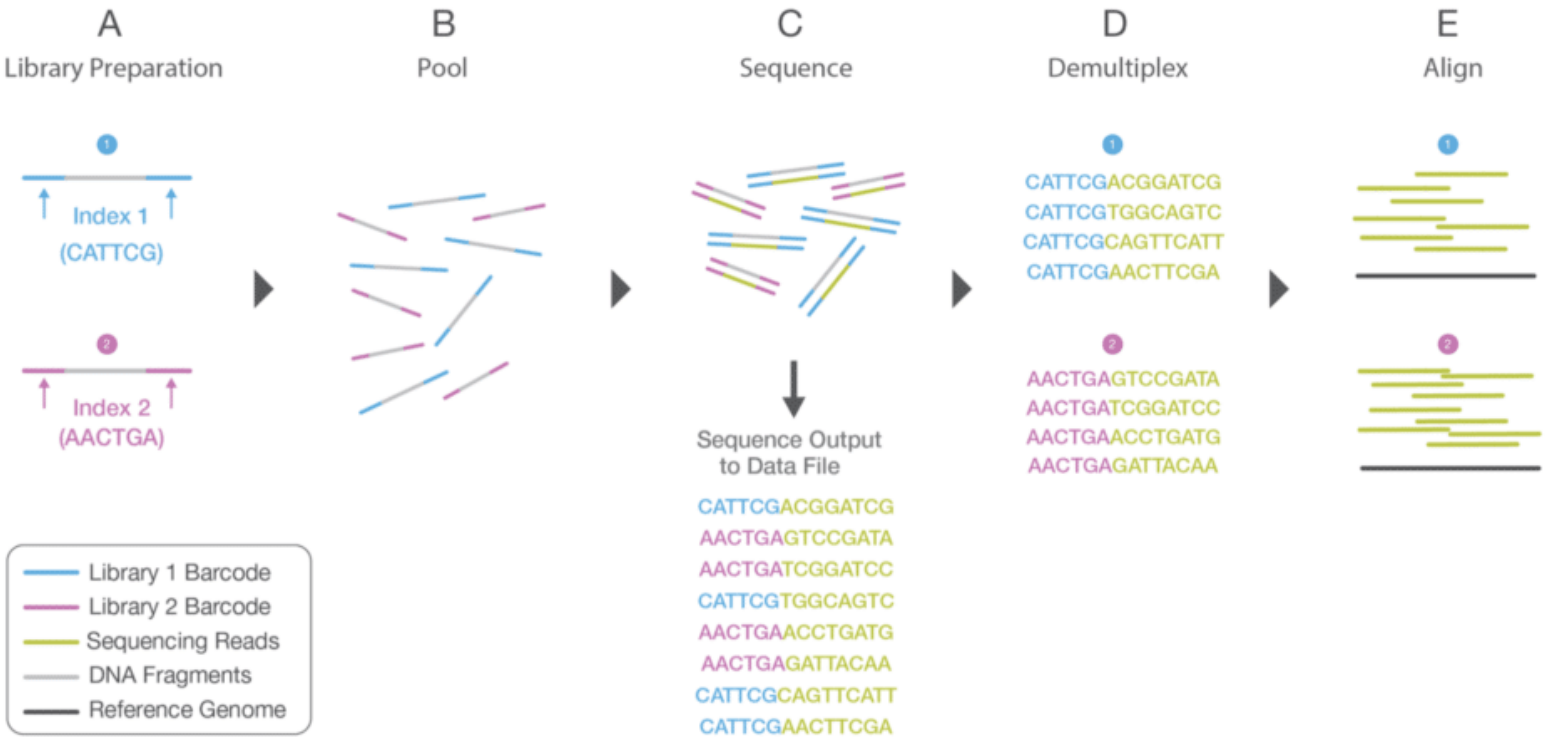


Fig. 1 Avian and non-avian reptilian landscapes of protein encoding genes. A set of 15,135 common vertebrate genes was sorted in the order of the human reference genome, alternating grey/white bars represent the different chromosomes. A sliding window of a centered gene and its 100 neighbors was taken to calculate the regional genomic average for each variable. **a** Presence index (red) and length index (blue) of the genes in the eight avian genomes. The areas in orange dots define the genes where the presence index is below the threshold of 0.70. In light blue it is shown where the length index is higher than the threshold of 1.46. In panel **b**, we have displayed the GC content of mRNA transcripts of the best annotated of the 4 studied non-avian reptiles (black, *Chrysemys picta*) and eight studied birds (red, *Pseudopodoces humilis*). The highest peaks of GC content are often seen in areas of a low presence index. **c** and **d** Landscapes of the cumulative presence of GARP% (encoded by GC-rich codons, green) or FYMINK% (encoded by AU-rich codons, purple) in the *Pseudopodoces humilis* (**c**) and *Chrysemys picta* (**d**) genome. The amount of GARP% and GC content are strongly correlated (R = 0.92 for *Chrysemys picta* and R = 0.91 for *Pseudopodoces humilis*)

Optimizing throughput by multiplexing

- The sequencing throughput is an overkill for many applications
- Lots of short reads can be produced in a single run
- Individuals can be bulked with multiplexed sequencing to fractionate the throughput

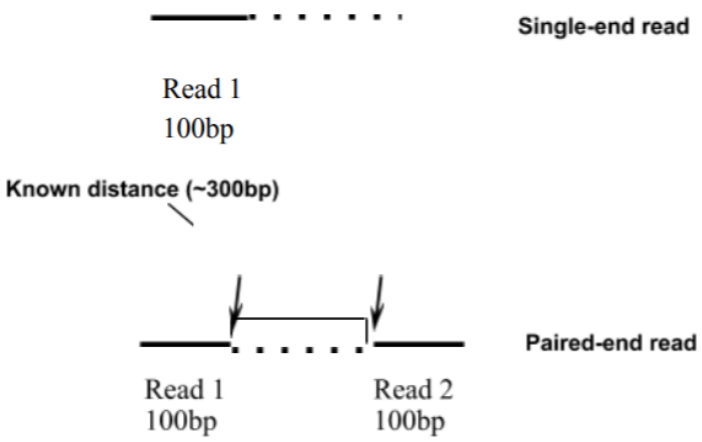


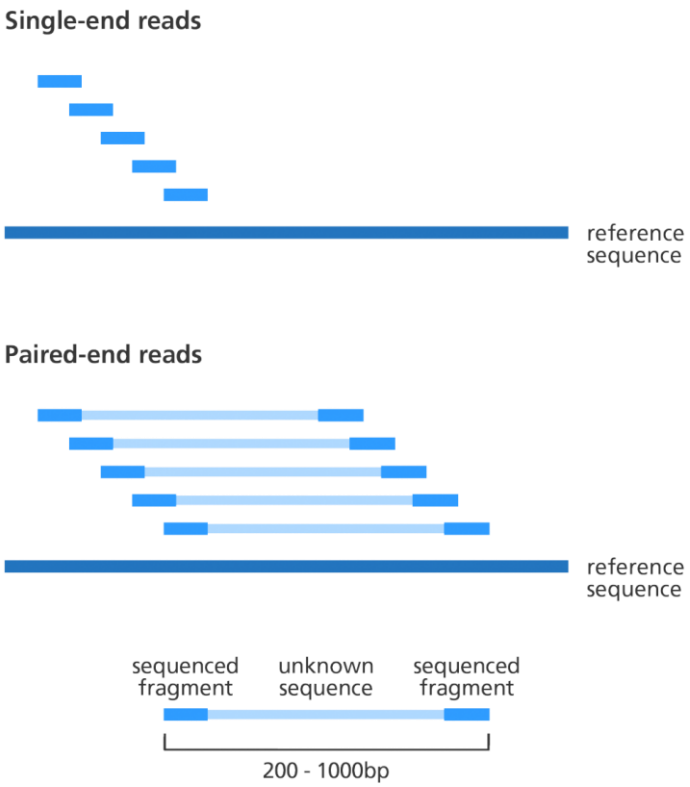


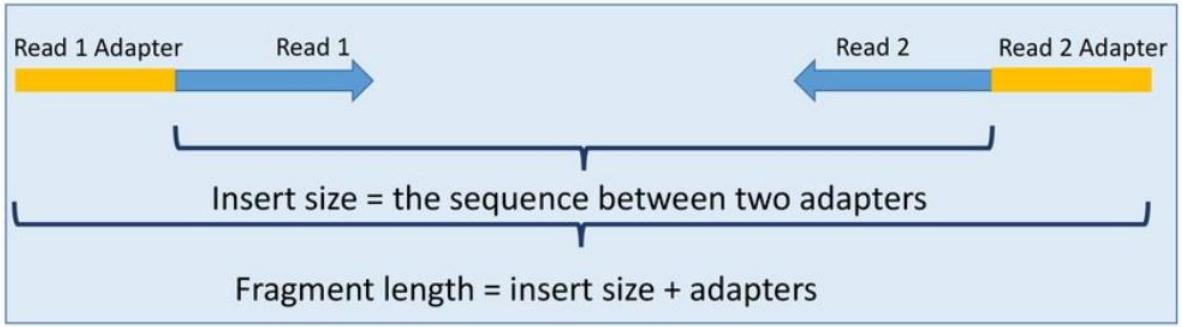
- Barcodes are designed of different lenghts, and at least two Nts away (to avoid misassignments due to sequencing error)

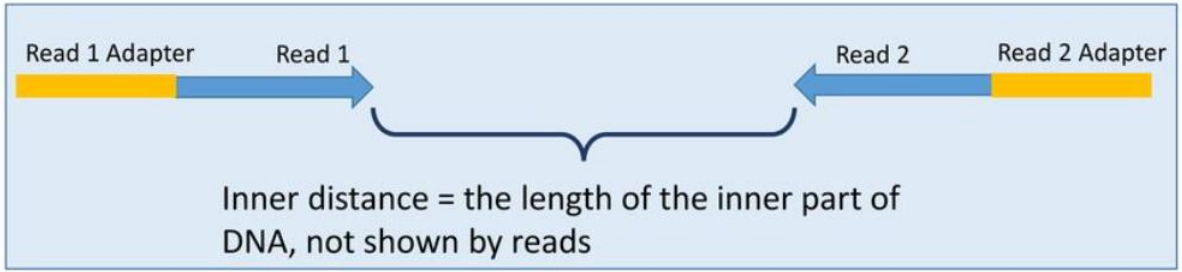
Optimizing sequencing range via paired ended sequencing

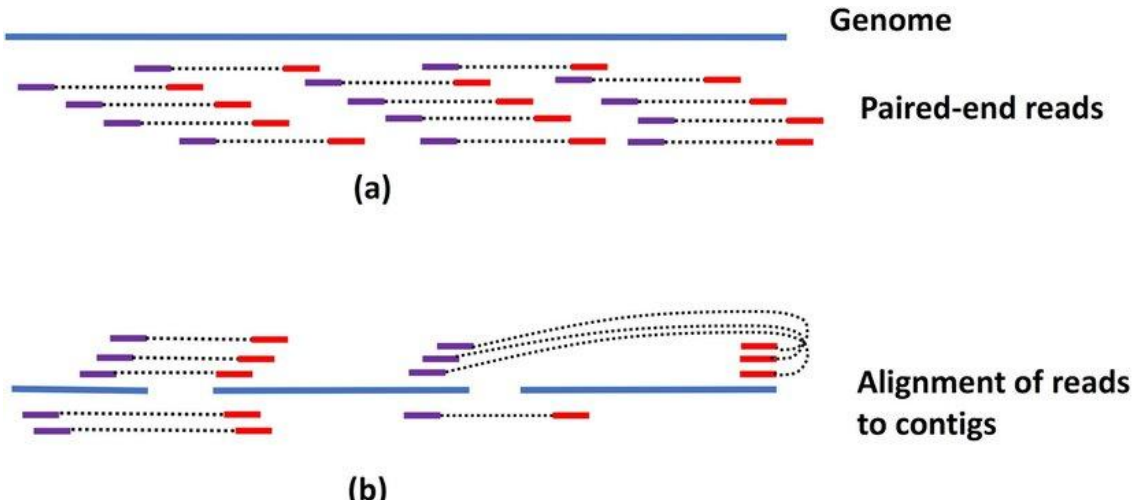
- Read lenght may be extended using paired ended sequencing, i.e. including a fragment of unknown sequence but known size in between two reads
- It remains challenging to reconstruct complex genomes

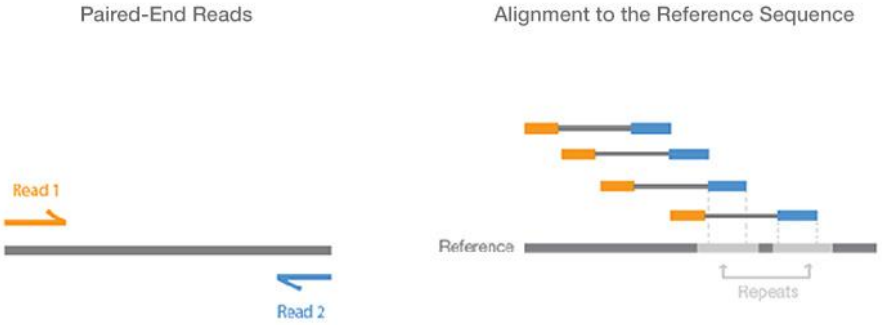





Read 1 Adapter
Read 1
Read 2
Read 2 Adapter
Insert size = the sequence between two adapters
Fragment length = insert size + adapters


Read 1 Adapter
Read 1
Read 2
Read 2 Adapter
Inner distance = the length of the inner part of DNA, not shown by reads





- Paired ended sequencing makes it easier to identify alignment issues and to overcome complexity (e.g. repeated regions)

NGS pros

- No *E.coli* subcloning
    - No cloning bias
    - Easier library preparation
    - Lower robotics, also for large genomes sequencing projects
- Each sequence comes from a single DNA molecule
    - Quantification possible by ‘digital counting’
    - Huge dinamic range
    - Rare variants detection
- Revolutionary cost decrease and very fast data production
- Non only DNA sequencing but a lot of different applications
    - DNA-seq
    - RNA-seq
    - CHIP-seq
    - amplicon-seq
    - target resequencing
    - BS-seq
    - Etc...

NGS cons

- Shorter reads with respect to Sanger sequencing
  - Illumina technology can reach now 300bp paired-end (e.g. 600bp/fragment)
  - Third and fourth generation sequencers will produce very long reads (Kbp)
- Big investments in terms of IT infrastructures (storage and RAM)
    - Tbytes of data are produced per run
    - Analysis pipelines out of the instrument
- Complex bioinformatics data analysis
    - Scripting in real time
    - The technology is going faster than the human capacity to give to the data a biological significance