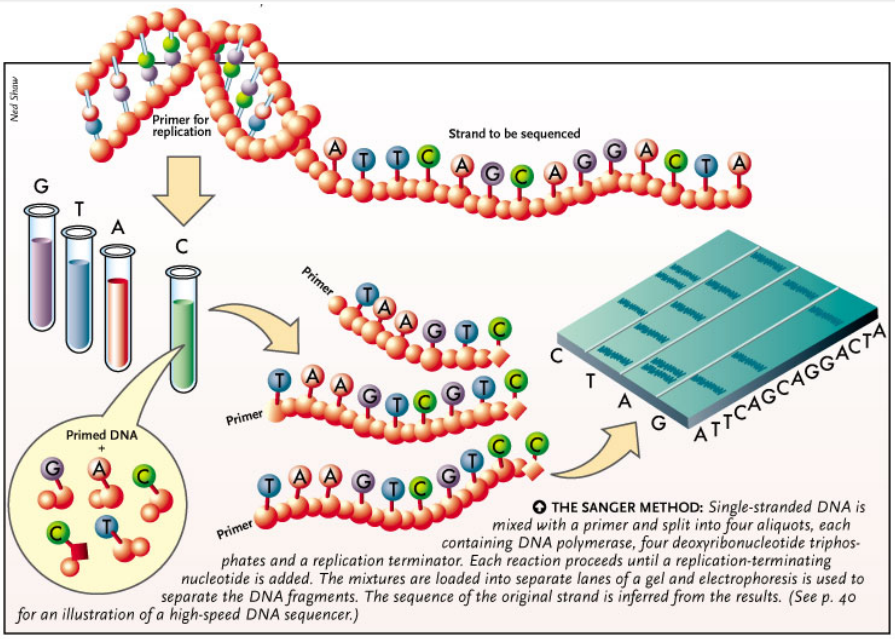
A Primer to Sequencing

Sanger (1977) sequencing



- In 4 tubes, a template, primer, polymerase, the 4 nucleotides ATP, CTP, GTP, and TTP are added, plus one out of the 4 possible ddNTPs, where the DNA-polymerization cannot continue if built in.
- The polymerization reaction is started and whenever a ddNTP is incorporated, a copy exists, exactly ending with that nucleotide. . . and many will be produced ending at different bases in the sequence.
- After the reaction, the 4 probes will be run on a gel, where they will separate according to their length, shorter ones migrating faster.
- From order of bars in the 4 lane on the gel, one can determine the sequence of the original template.

Sanger (1977) sequencing

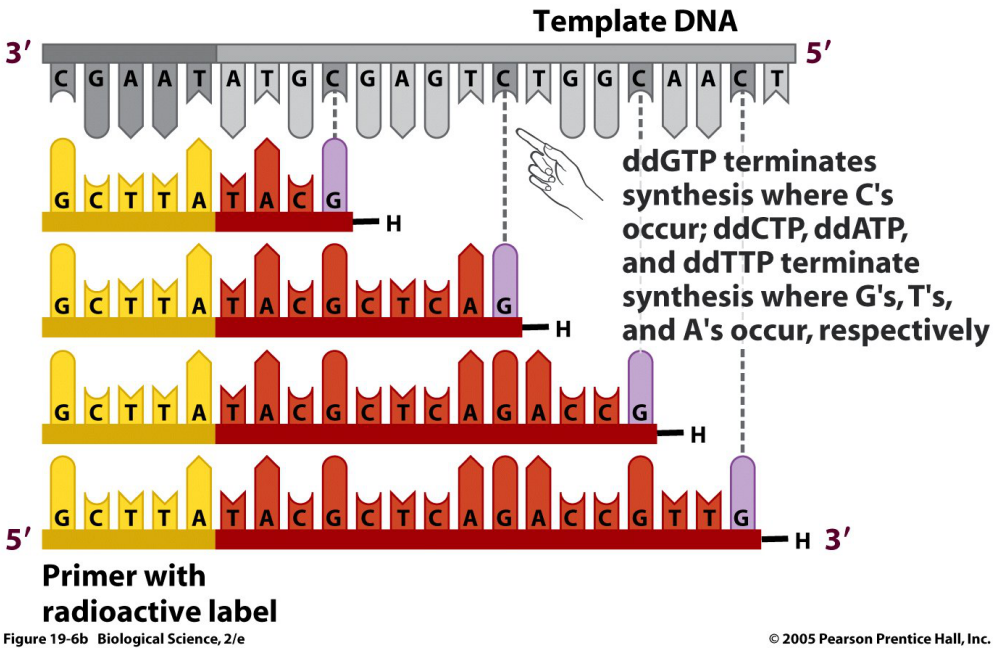


Figure 19-6b Biological Science, 2/e

Sanger (1977) sequencing

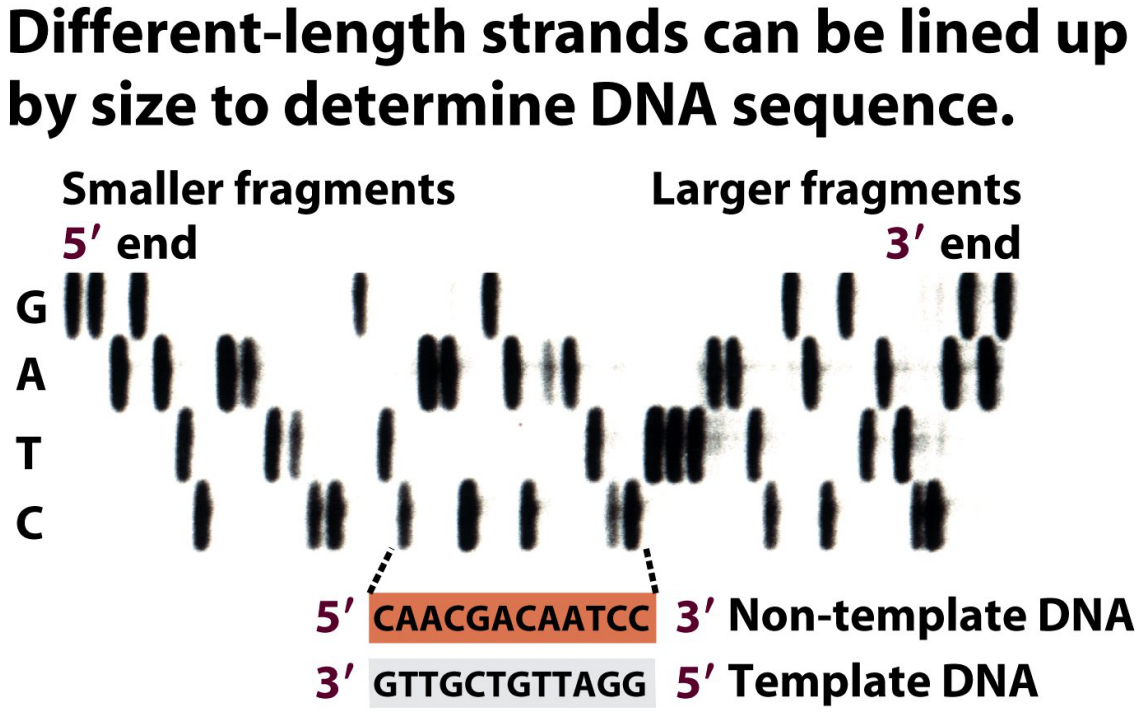
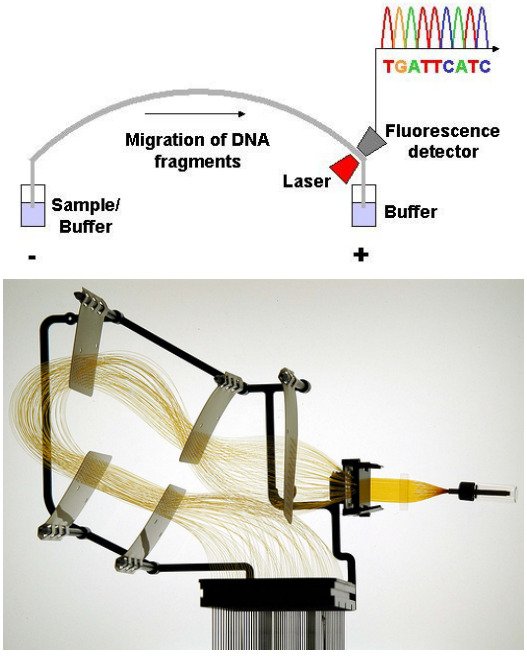
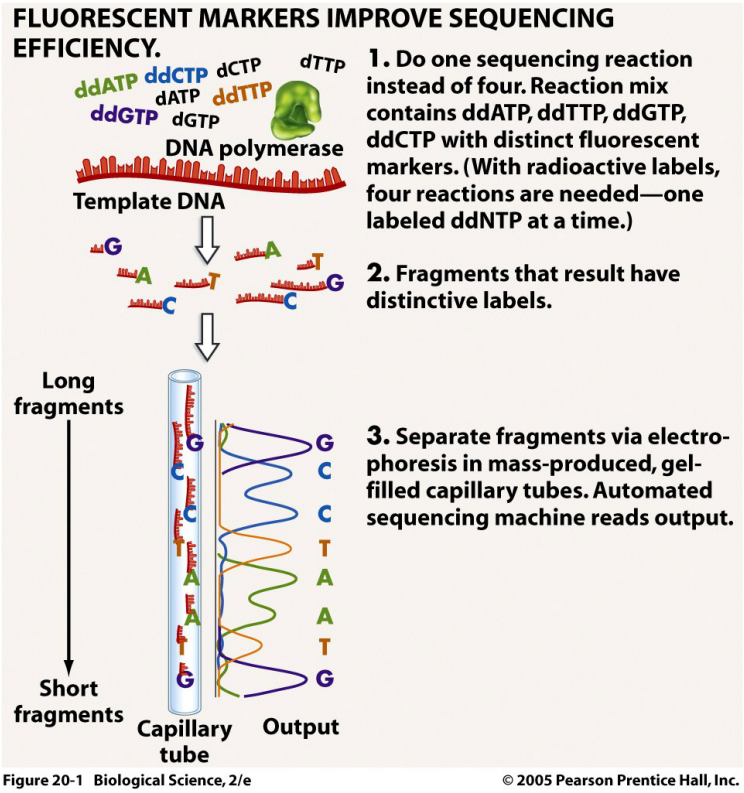


Figure 19-6c Biological Science, 2/e

Automated Sanger sequencing (ABI 3130xl)

Automated Sanger sequencing





Some Specs for Automated Sanger sequencing

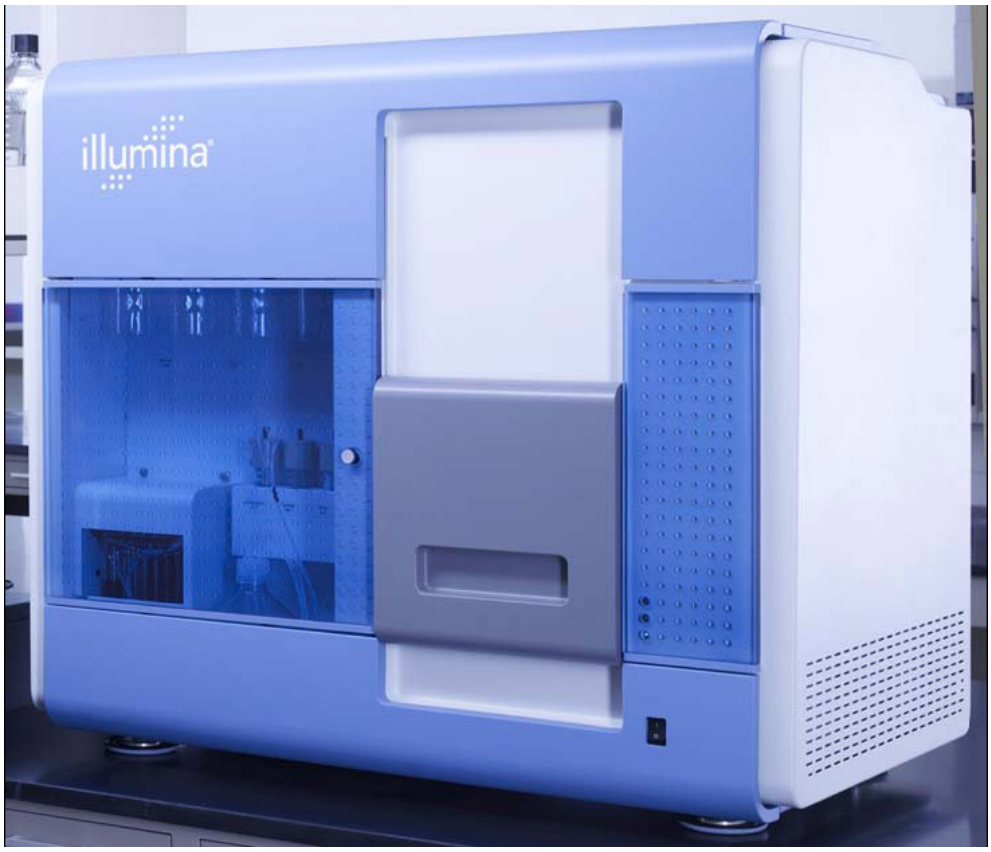
Sequencer	reads/run	length/read	time/run	error (%, type)
3730 capillary	96	650-1000	2h	$0.1 - 1\%$, substitutions

according to Glenn (2011/2016)

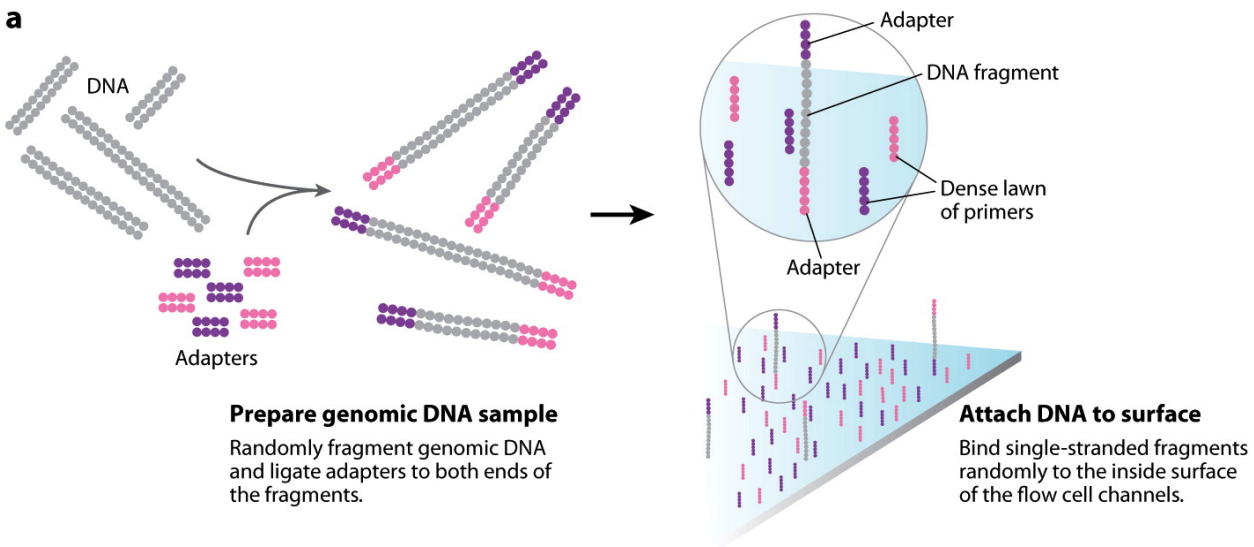
Next-generation sequencing

The 2nd Generation
(not Sanger based)

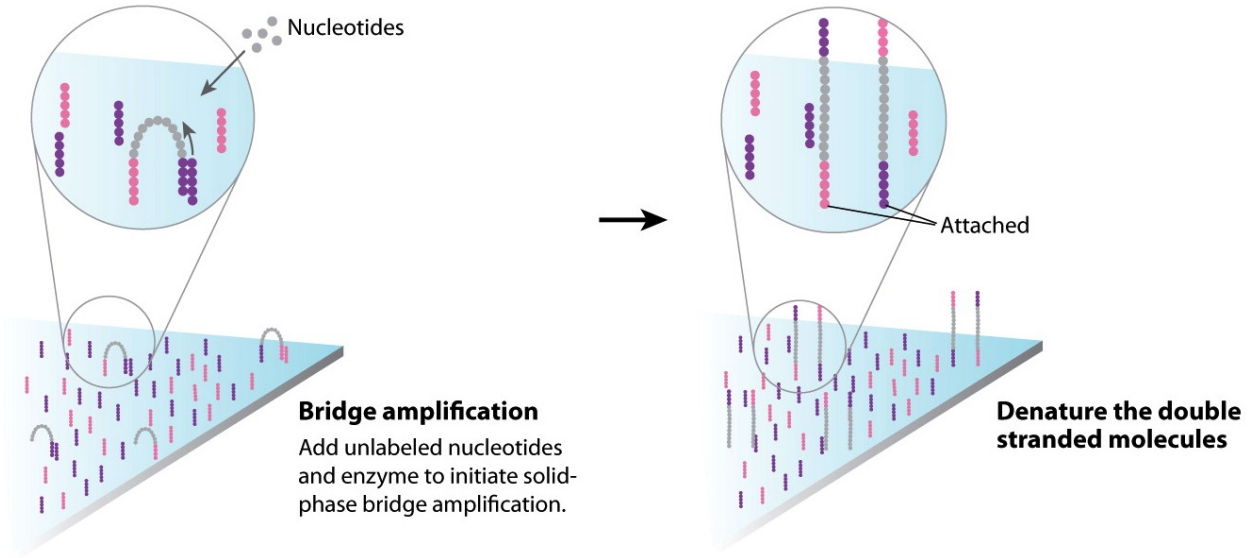
Illumina



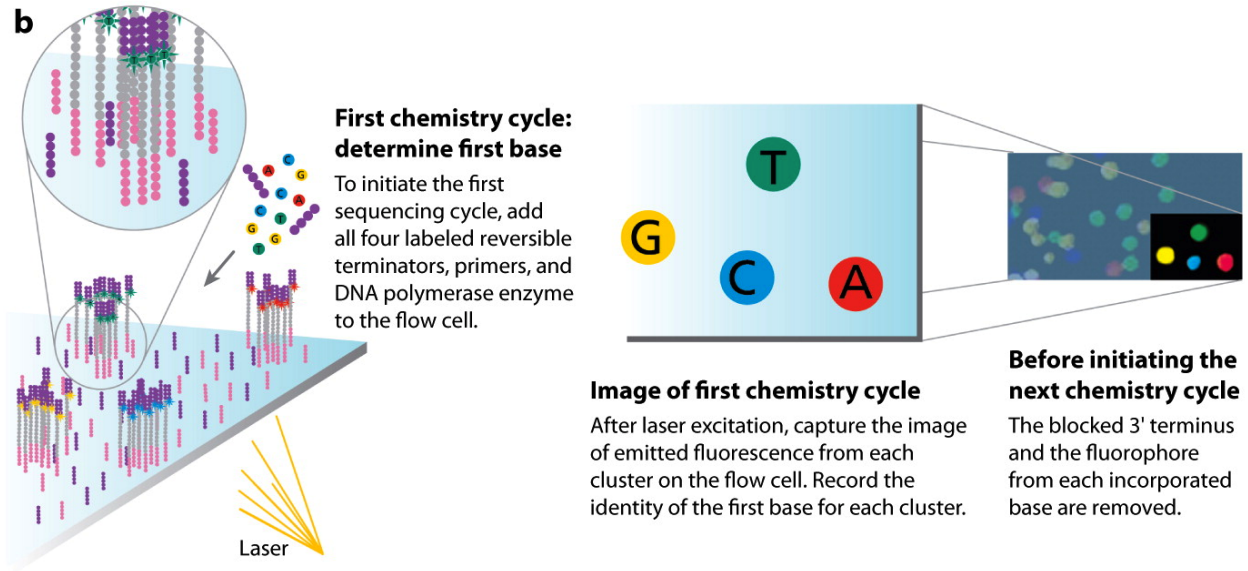
Illumina - a next-generation sequencing method (1)



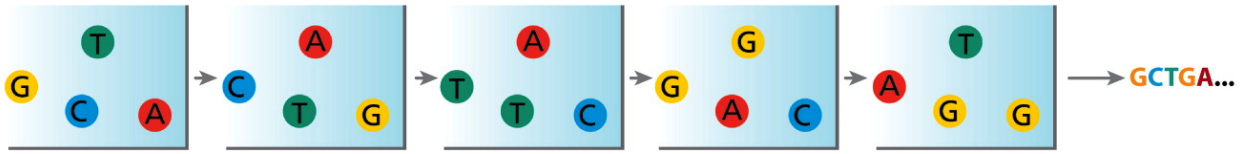
Illumina - a next-generation sequencing method (2)



Illumina - a next-generation sequencing method (3)



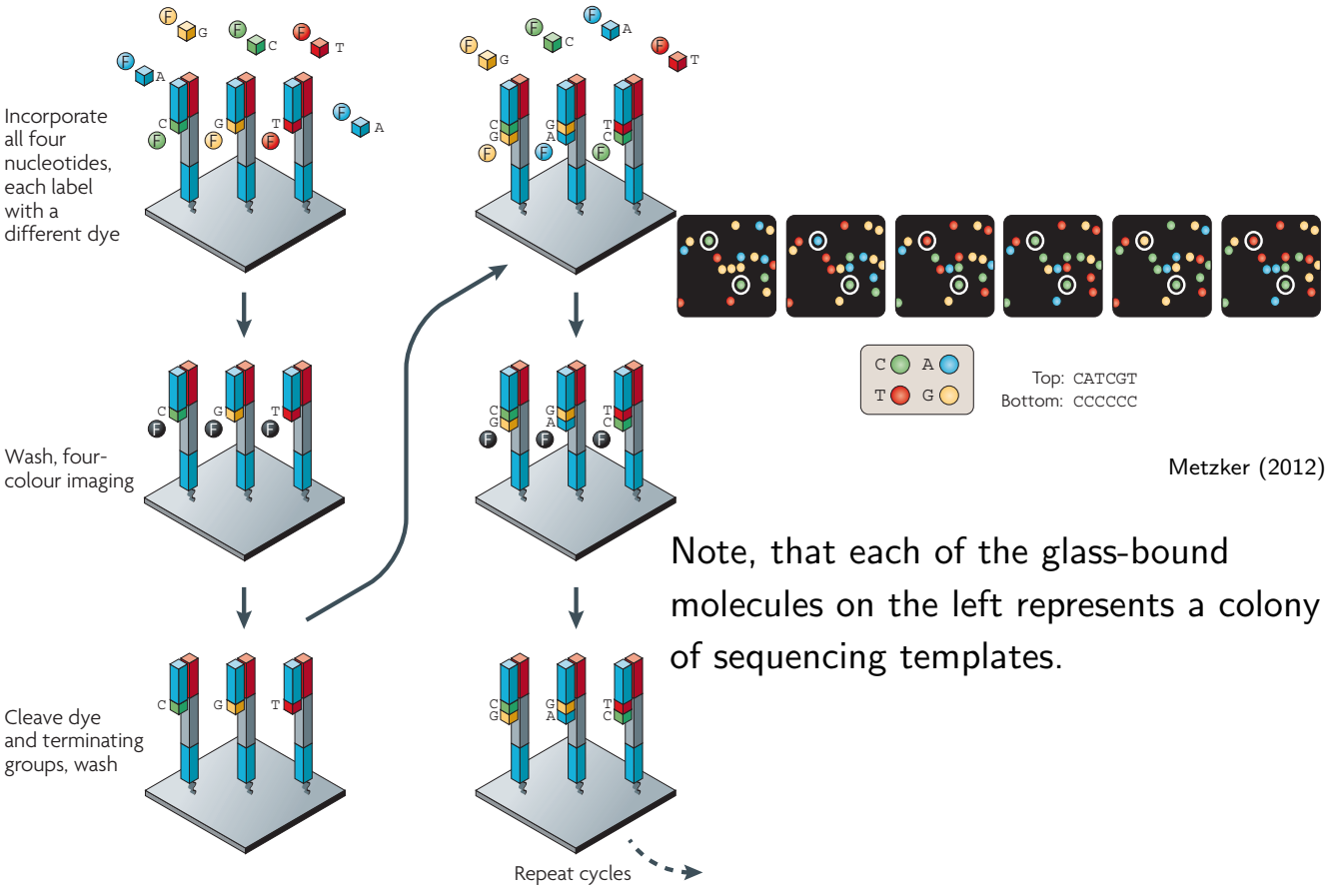
Illumina - a next-generation sequencing method (4)



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.

Illumina - getting the sequence



Metzker (2012)

Note, that each of the glass-bound molecules on the left represents a colony of sequencing templates.

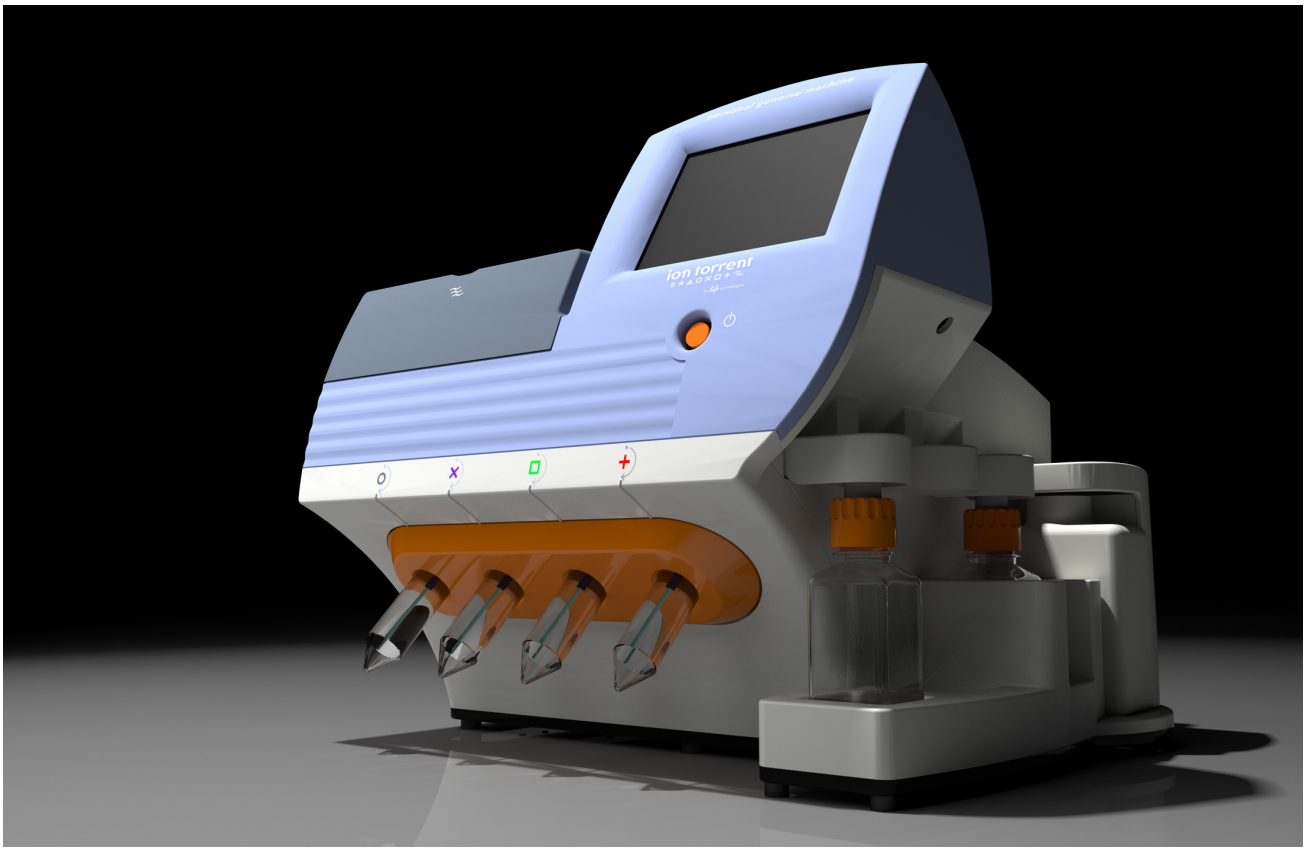
Some Specs for Illumina sequencing

Sequencer	reads/run	length/read	time/run	error (%, type)
MiSeq v3	25mio	$2x75^{PE} - 2x300^{PE}$	21-56h	0.1%, substitutions
HiSeq 4000	2500mio	$50^{SE} - 2x150^{PE}$	1-3.5d	same

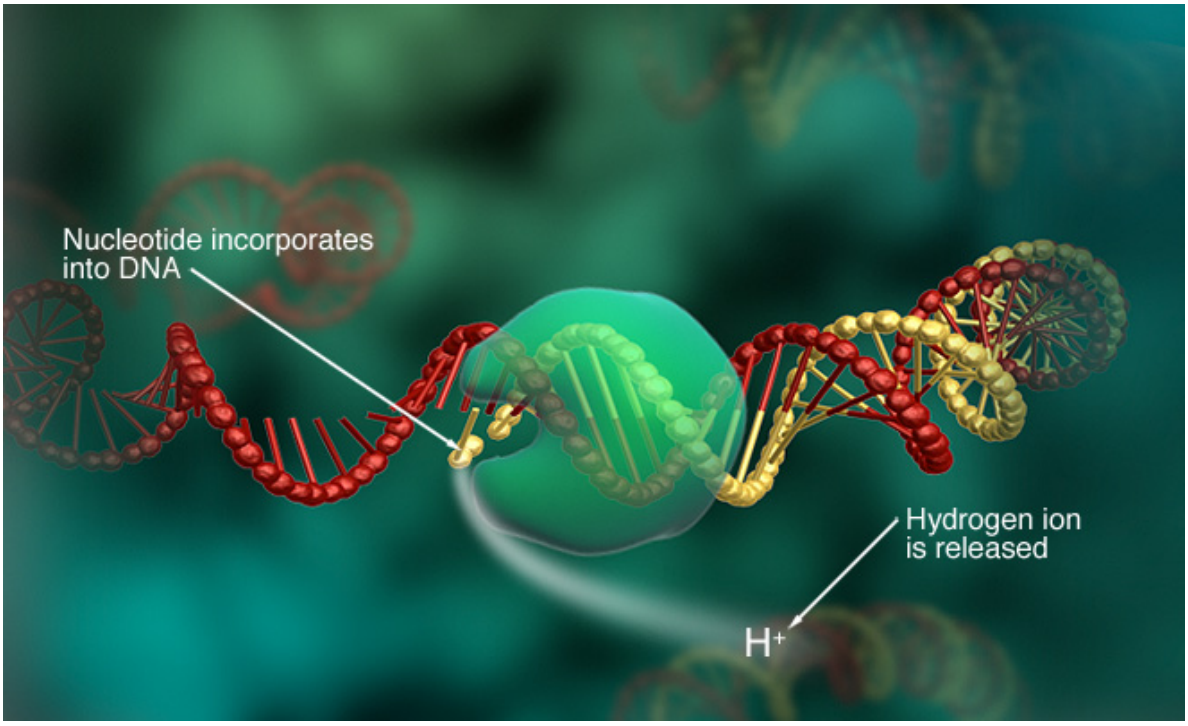
PE - 2 paired-end reads, SE - single-end reads

according to Glenn (2011/2016)

Ion Torrent Sequencing

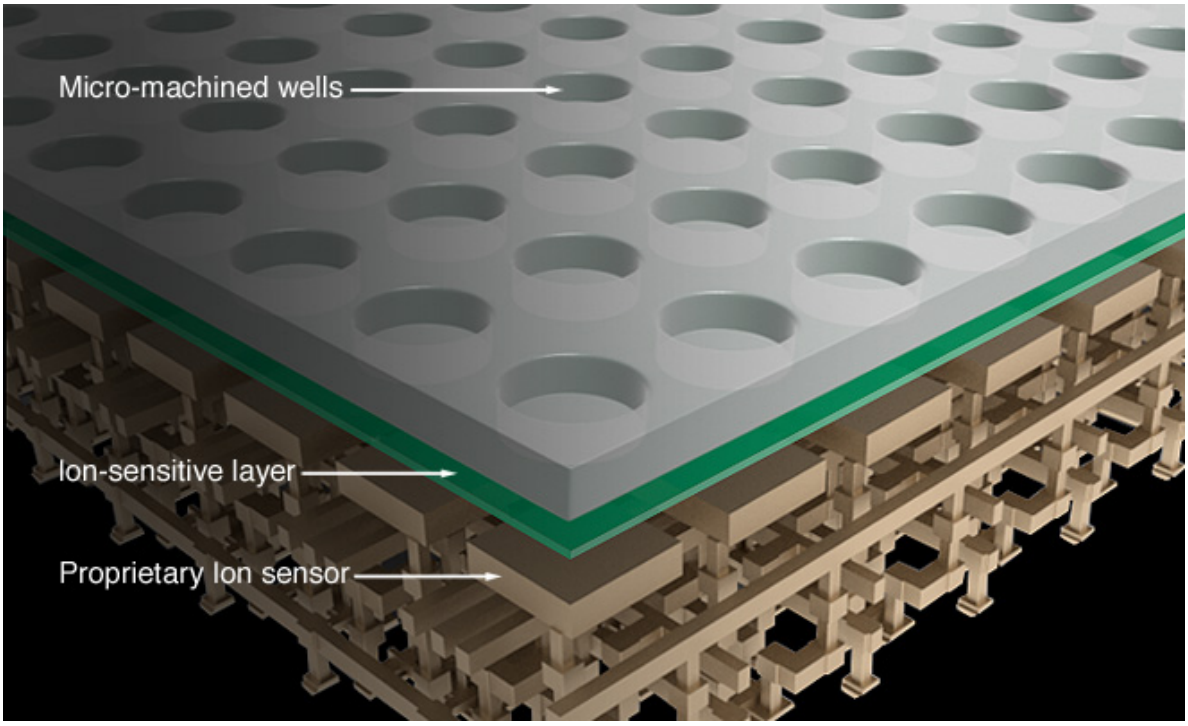


Ion Torrent Sequencing (1)



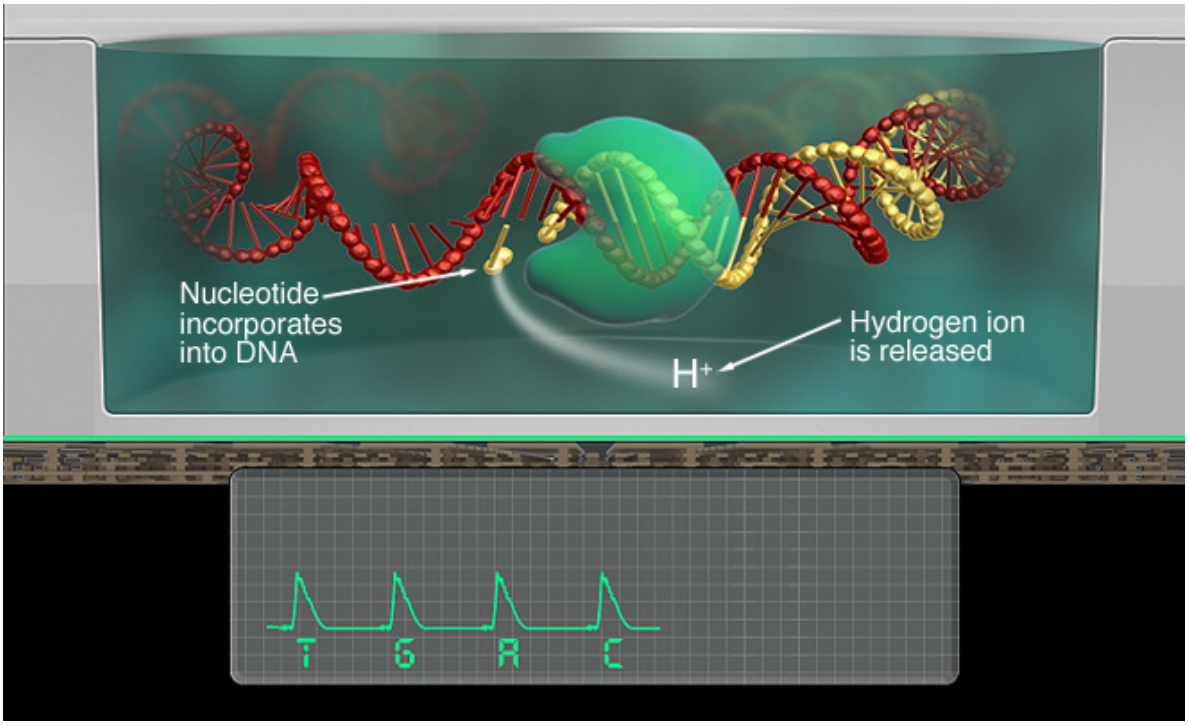
www.iontorrent.com

Ion Torrent Sequencing (2)



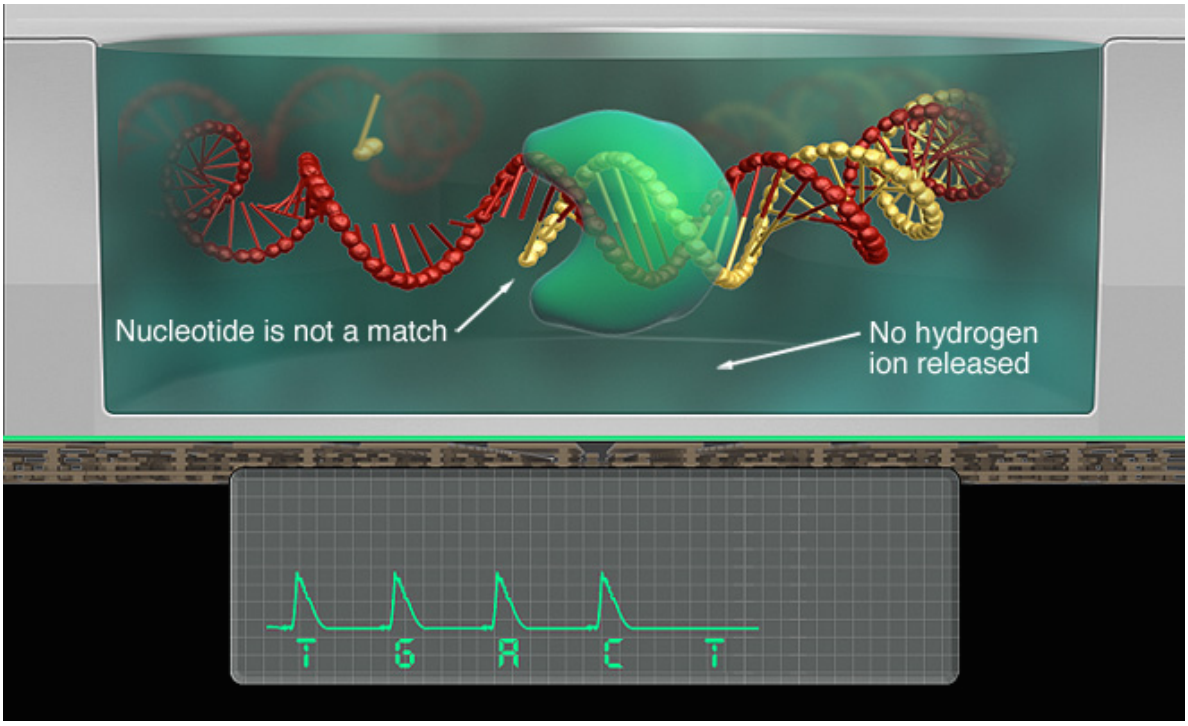
www.iontorrent.com

Ion Torrent Sequencing (3)



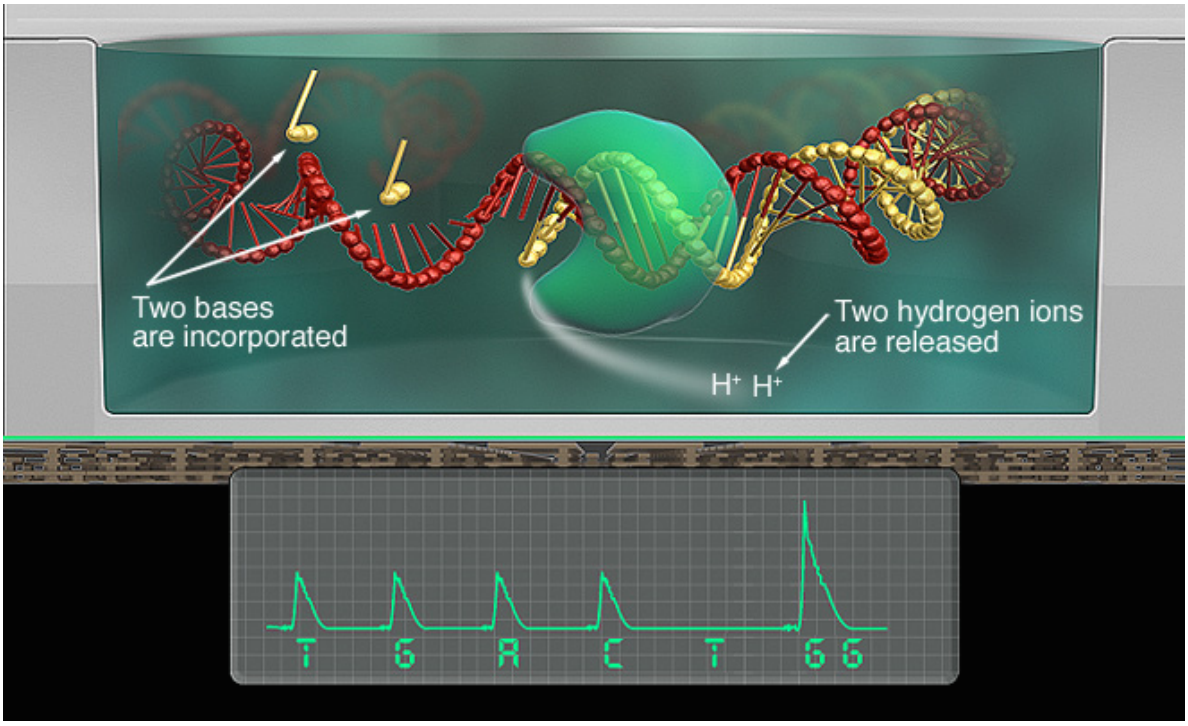
www.iontorrent.com

Ion Torrent Sequencing (4)



www.iontorrent.com

Ion Torrent Sequencing (5)



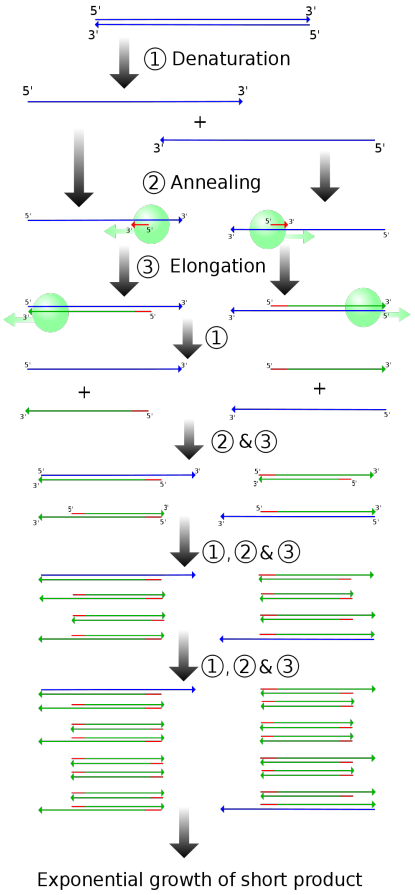
www.iontorrent.com

Some Specs for Ion Torrent sequencing

Sequencer	reads/run	length/read	time/run	error (%, type)
PGM 312 v2	5.5mio	400	7h	$\sim$ 1%, indels
S5 540	80mio	200	2.5h	same

according to Glenn (2011/2016)

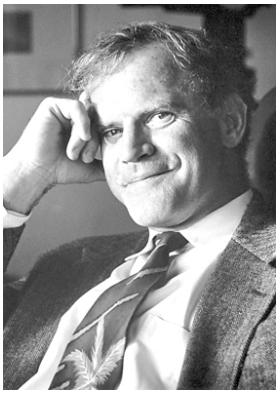
Excursion: DNA Amplification by PCR



(Wikipedia)

1. separate the DNA strands (denaturation)
2. add primers and let them anneal to the templates
3. polymerases elongate the primers producing complementary strands
4. repeating steps 1-3 the part between (and including) the primers is amplified exponentially.

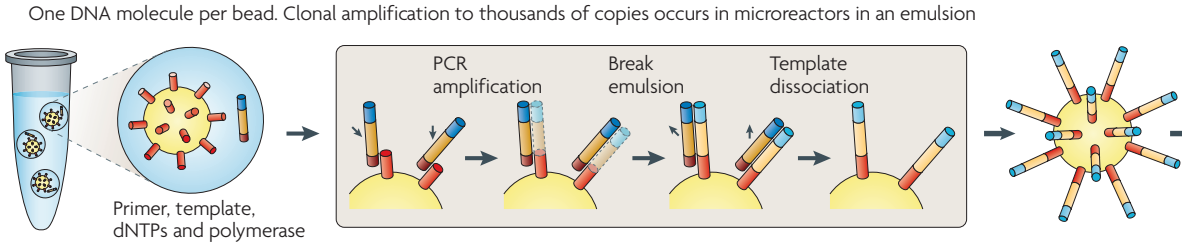
For the efficient *Polymerase Chain Reaction* (PCR) method Kary Mullis got the Nobel Prize (1993).



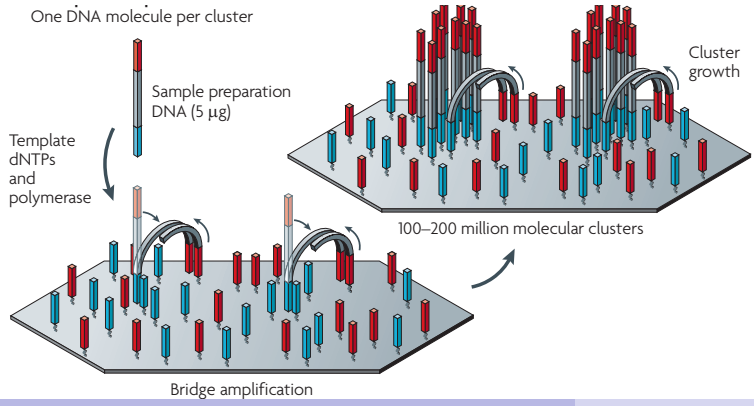
Amplification steps

PCR variants are typically used to amplify the DNA first:

- **emulsion PCR** (used by Roche/454, ABI SOLiD, Ion Torrent)



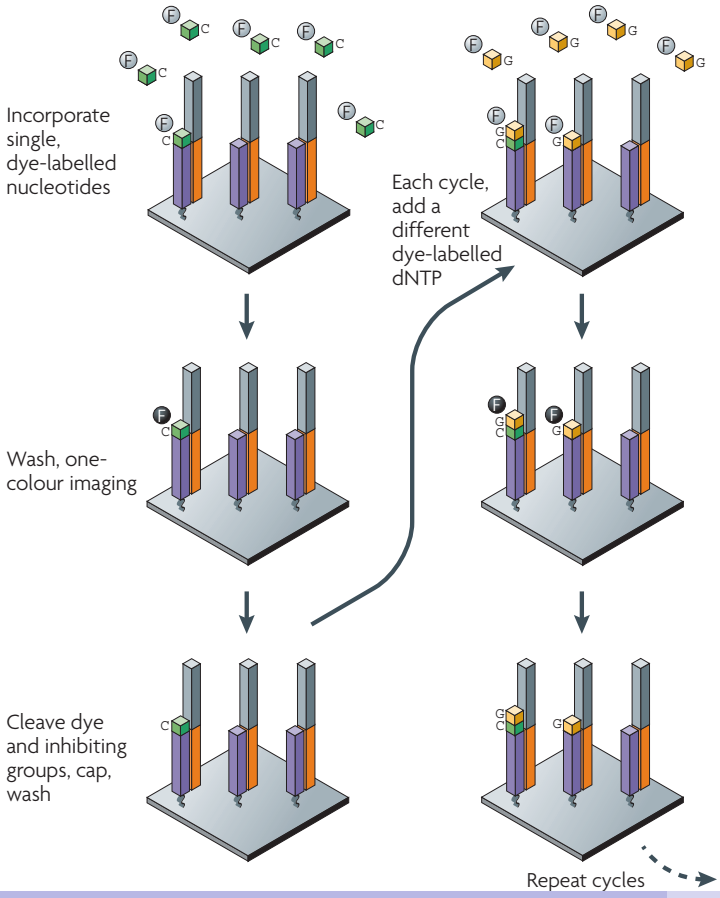
- **solid phase** or **bridge PCR** (used by Illumina)

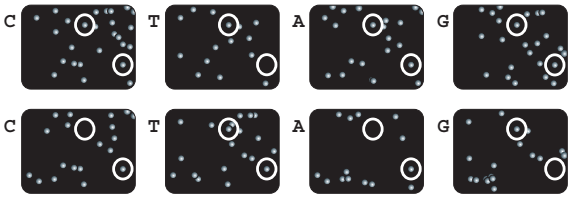


The 3rd Generation
(no amplification step)

Helicos Helioscope (discontinued)
the first single-molecule sequencer

Helioscope - the first single-molecule sequencer (discontinued)



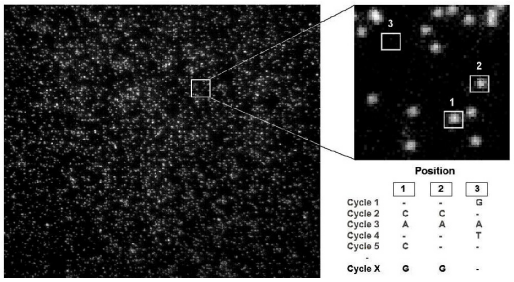


Top: CTAGTG
Bottom: CAGCTA

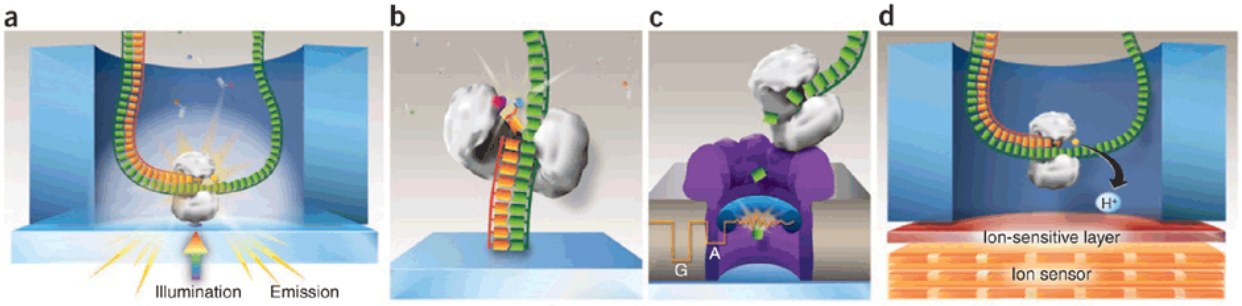
Metzker (2012)

Note, that each of the glass-bound molecules on the left represents a single sequencing template.

A real sequencing image:



Examples for the 3rd Generation



Munroe + Harris (2010)

a PacBio SMRT (single-molecule real-time) DNA sequencing

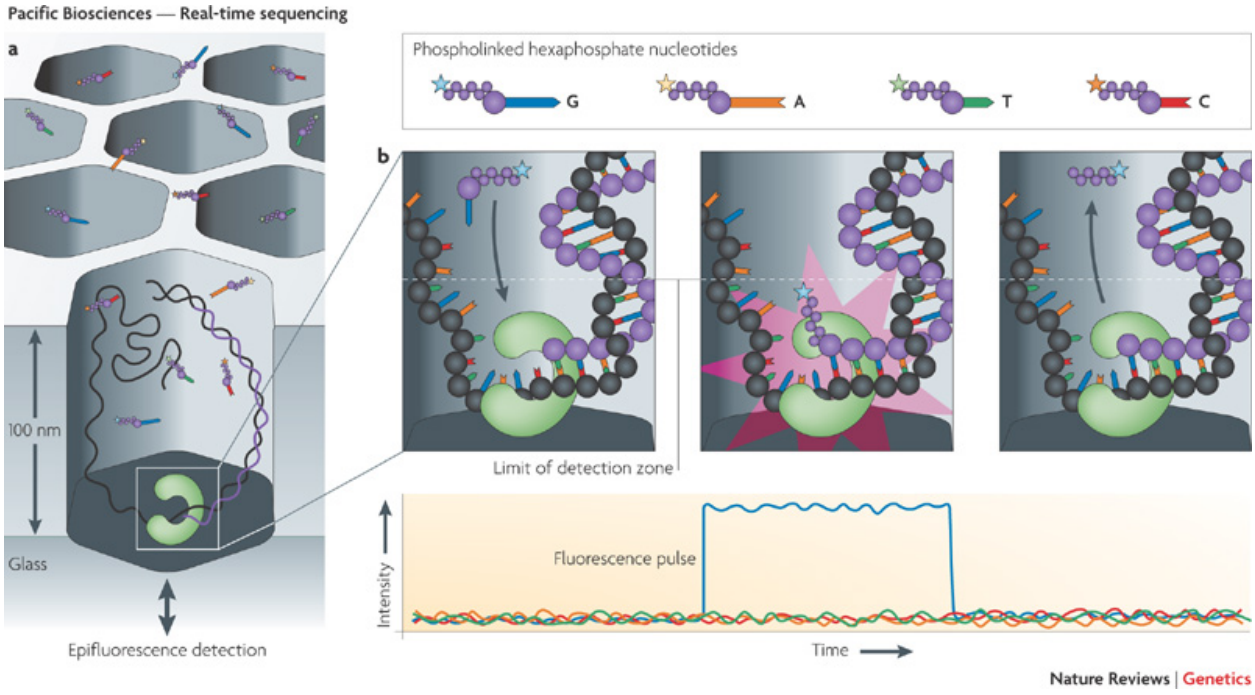
b Life Technologies FRET (fluorescence resonance energy-transfer technology) sequencing platform

c Oxford Nanopore

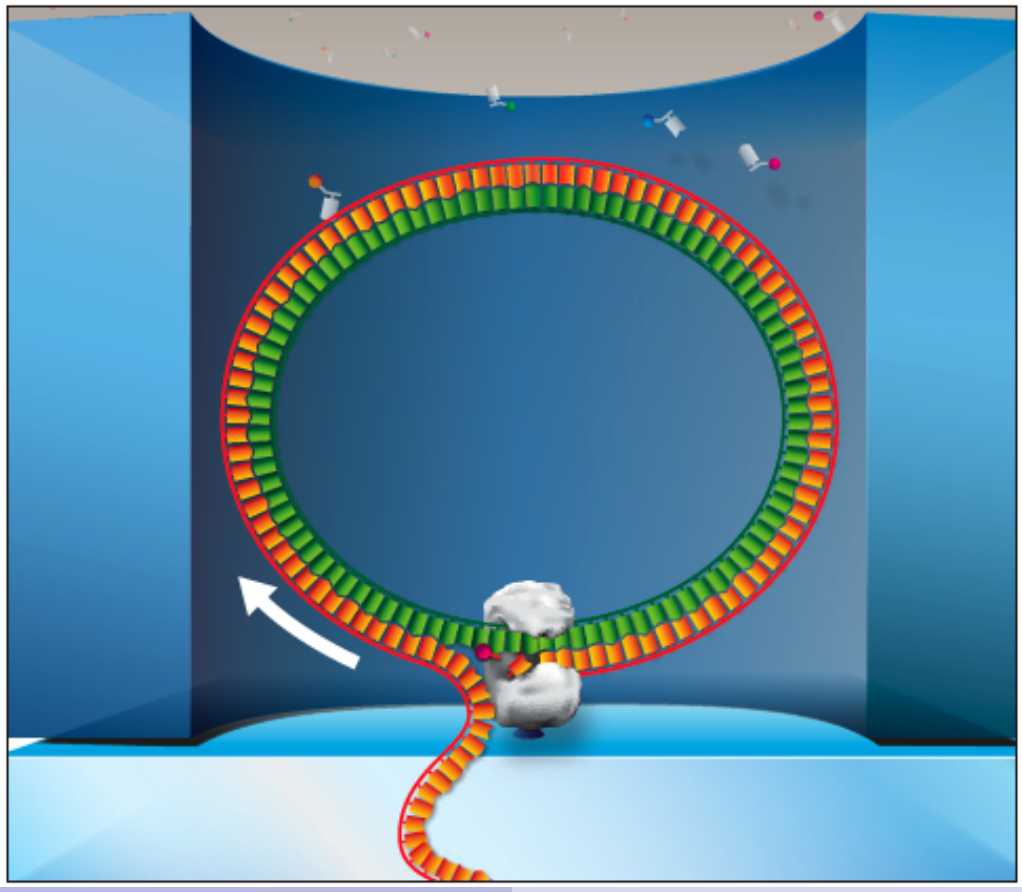
d Single molecule + Ion Torrent

Pacific Biosciences (PacBio)
single molecule real time sequencing (SMRT)

PacBio single molecule sequencing



PacBio sequencing cycles



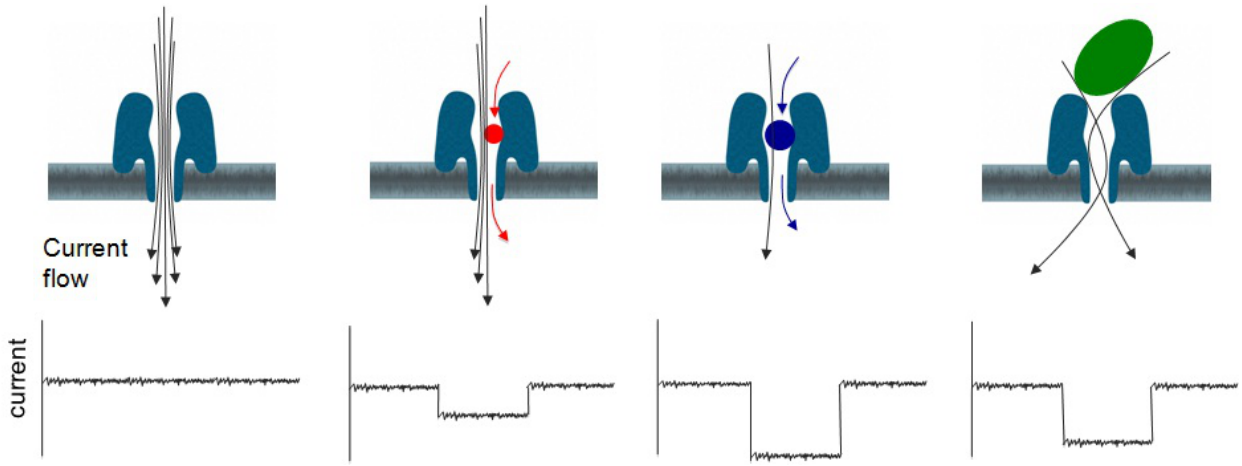
Some Specs for PacBio SMRT sequencing

Sequencer	reads/run	length/read	time/run	error (%, type)
RS II	55k	12000	up to 6h	13% raw, indels
Sequel	385k	12000	up to 6h	

according to Glenn (2011/2016)

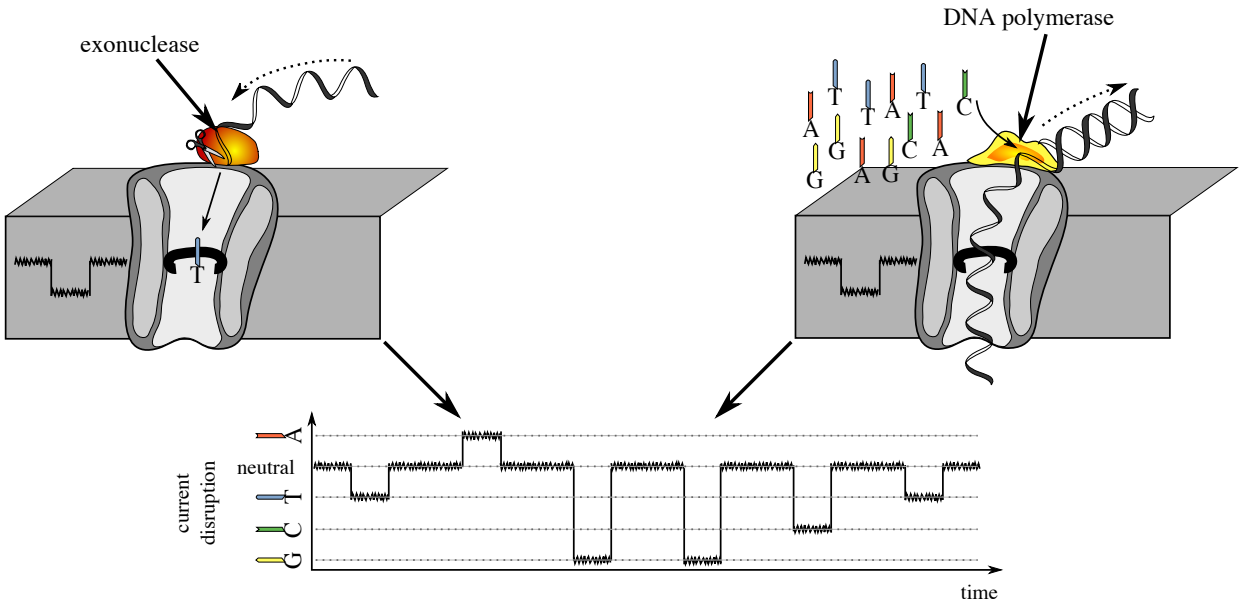
Oxford Nanopore PromethION and MinION

Nanopore sequencing



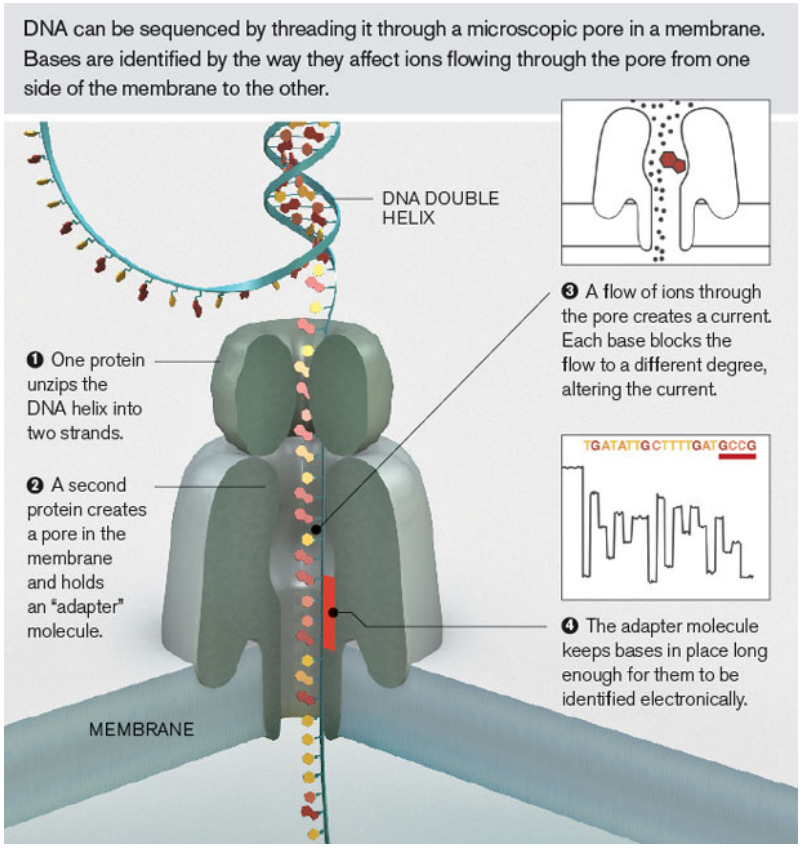
www.nextbigfuture.com

Nanopore sequencing - directions through the pore



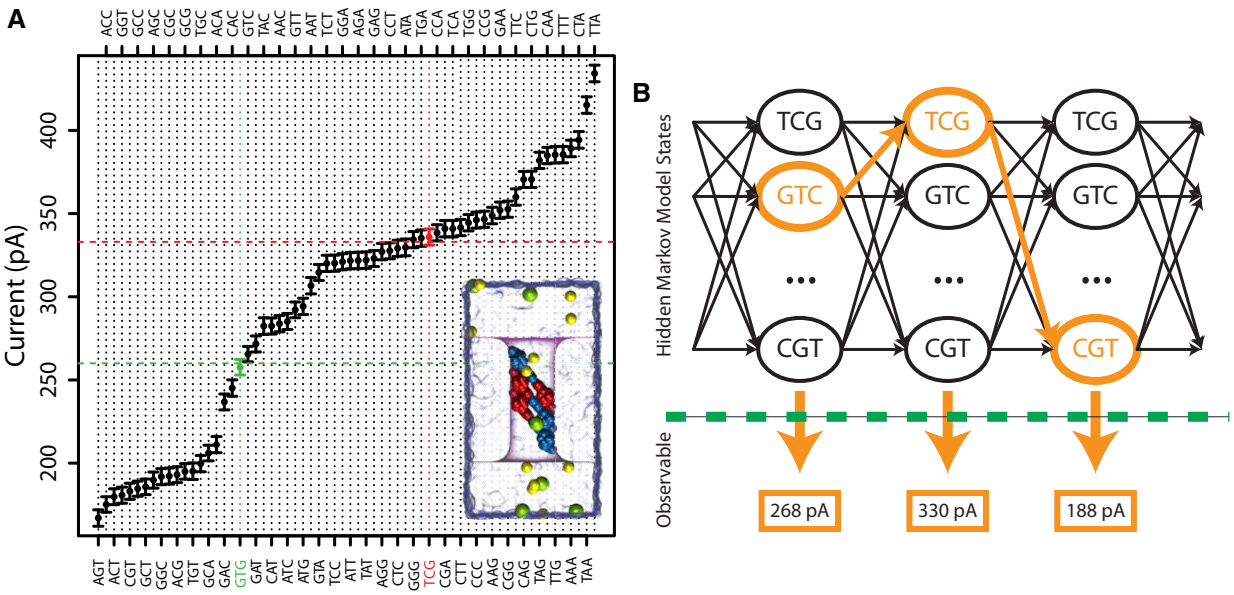
template: TAGGCT (www.wikimedia.org)

Nanopore sequencing



www.technologyreview.com

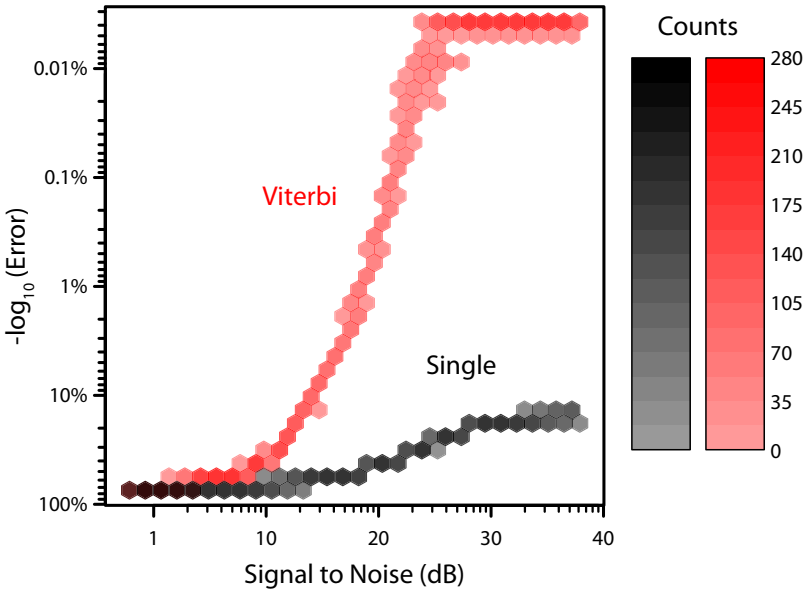
Nanopore base calling with HMMs



(Timp et al. 2012)

- The signal change is influenced by the 3-4 contiguous nucleotides in the center of the pore and the signals can be similar.
- Thus, HMMs are used to deterine the best sequence from the signal, e.g., applying the Viterbi algorithm.

Nanopore Base calling with HMMs



(Timp et al. 2012)

Viterbi shows much lower error rates compared to calling the nucleotides one-by-one.

Oxford Nanopore MinION

The technique can be miniaturized (picture: P. Rescheneder)

Some Specs for Oxford Nanopore sequencing

Sequencer	reads/run	length/read	time/run	error (%, type)
MinION	0.6-4.4mio	10000	varies	4%, deletions
PromethION	26mio/flowcell	10000	varies	same

according to Glenn (2011/2016)

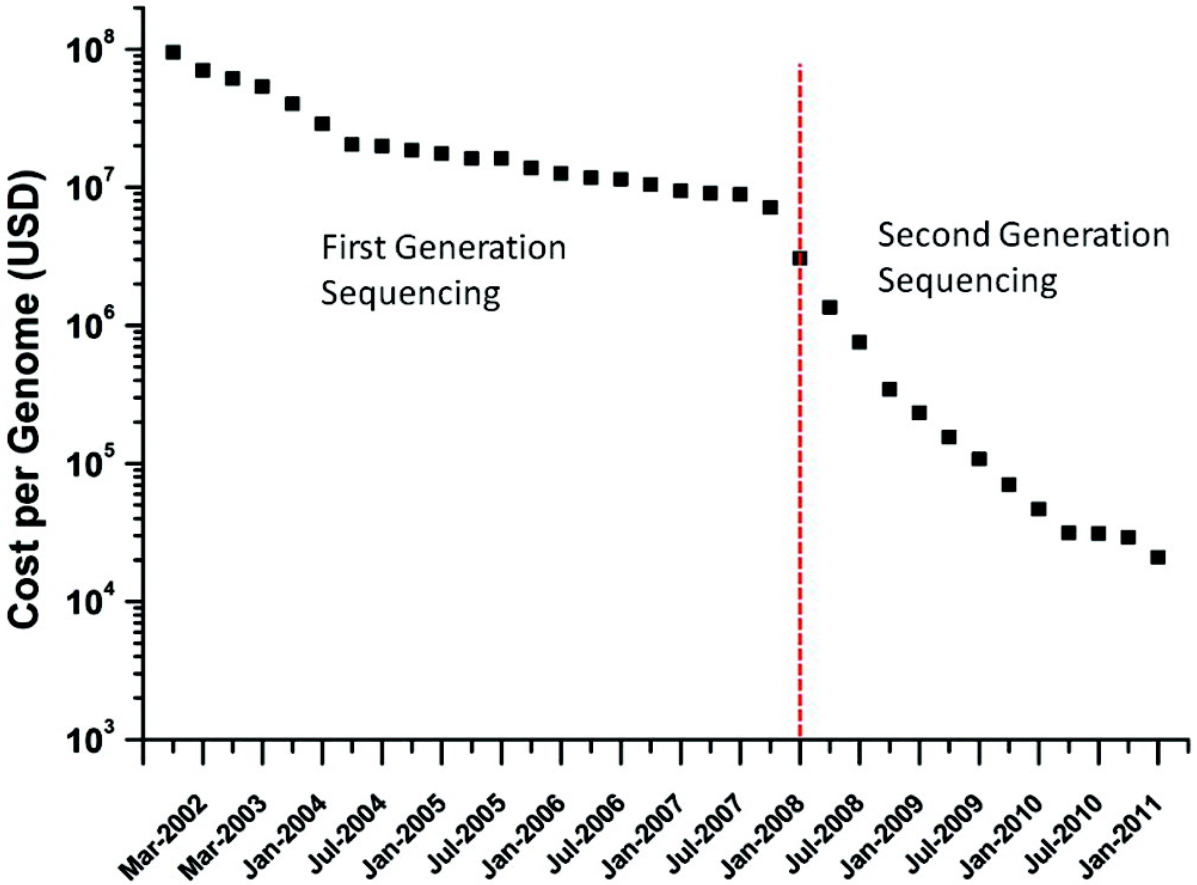
Generations of Sequencing

1. 1st Generation:
   - mainly Sanger-based techniques
   - Maxam-Gilbert (discontinued)
2. 2nd Generation (a.k.a. next-generation sequencing or NGS)
   New techniques, usually incorporating an amplification step.
   - Roche/454 Pyrosequencing (discontinued)
   - Illumina
   - ABI SOLiD (Sequencing by Oligonucleotide LIgation and Detection) (discontinued)
   - Ion Torrent
3. 3rd Generation
   Techniques which aim at avoiding the amplification step, single molecule sequencing.
   - Helicos (discontinued)
   - PacBio (Pacific Biosciences)
   - Oxford Nanopore

Sanger vs. NextGen Sequencing

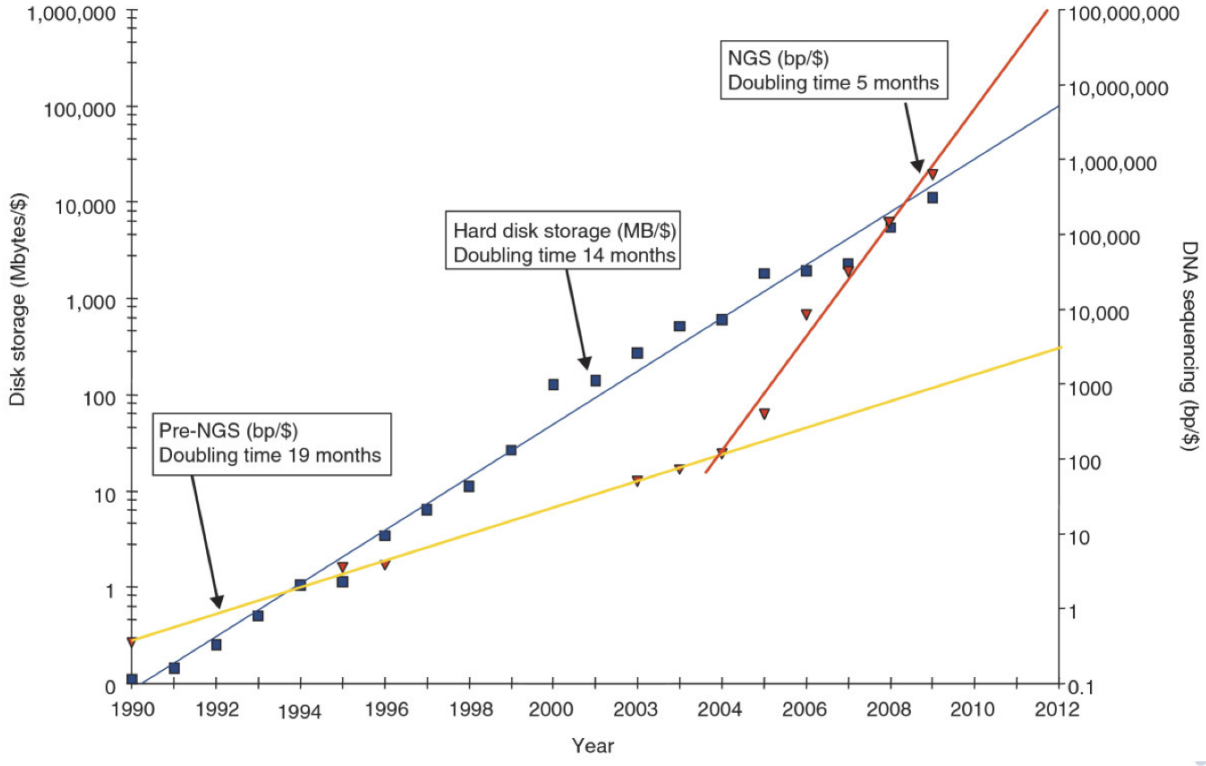
- Sanger can produce (high-quality) read for about 96 samples with read-length up to 1000bp.
- NextGen Sequencing methods like Illumina can produce only read-lengths between 36 and 300, however, millions of them.
- This results in much larger numbers of bases per run and is much cheaper than Sanger, but the short read-lengths pose severe problem in the later processing.

Sequencing Costs



Sequencing Costs vs. Storage Costs

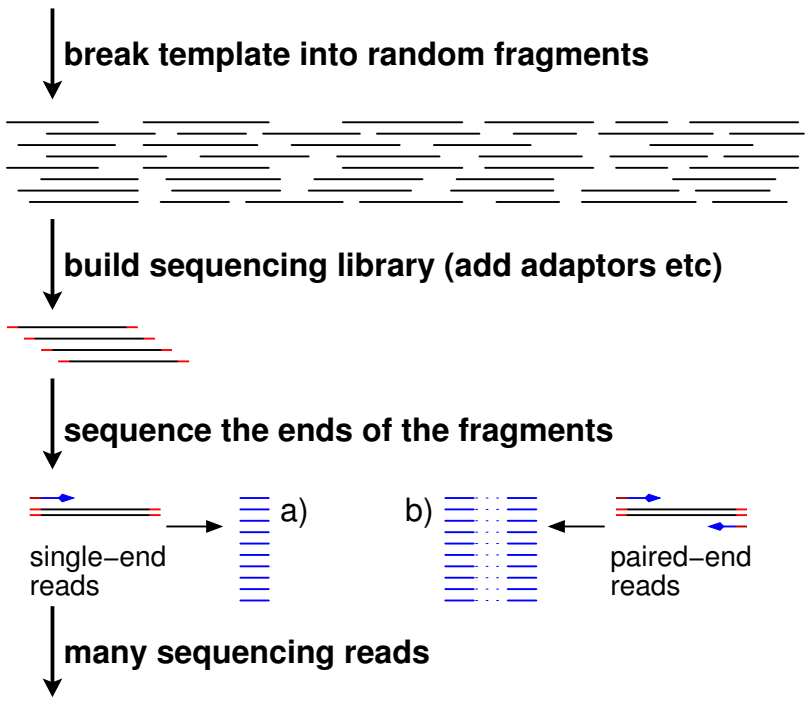
NextGen Sequencing a Game-Changer



Genome Sequencing, Assembly and NGS Data

Shotgun sequencing

template (e.g. genome)



Aim: reconstruct template sequence from reads

– reference–guided assembly (map against reference)
– de–novo assembly (reconstruct directly from reads)

Reference-based Genome Assembly

- If a reference genome of the sequenced species exists (or a relatively close taxonomic relative), we can use it to guide the assembly.
- The reads are mapped to the reference genome using approximative search algorithms.
- The closer the reference is to the sequenced genome, the easier is the mapping and assembly.
- From mapped contiguous reads we construct consensus sequences - the contigs.

Why re-sequencing

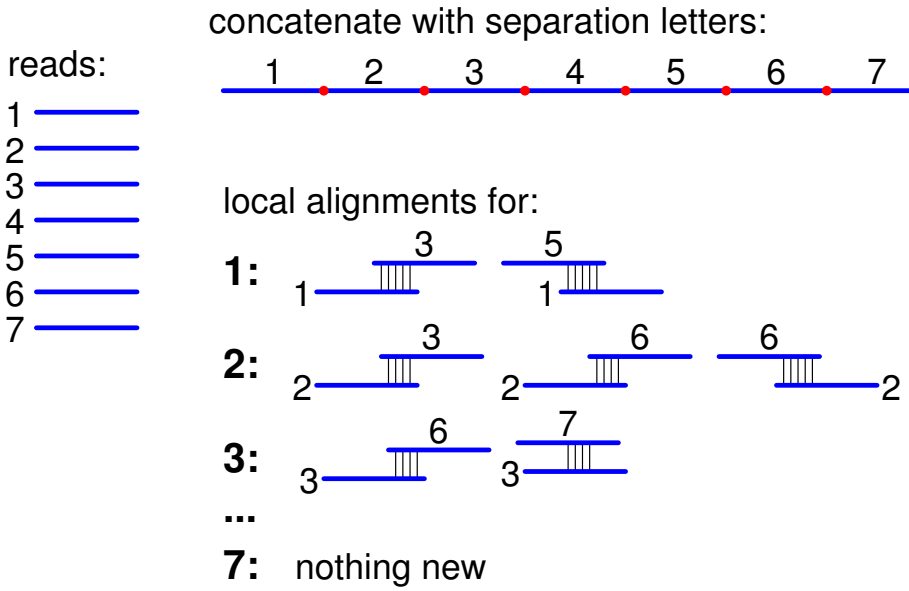
Why re-sequencing if a close reference genome already exists?

- Typically one does not re-sequence exactly the same individual the reference originated from.
- Usually one uses the reference to find
  - the differences in an individual carrying a disease (e.g. personalized medicine),
  - the characteristic changes in a new infectious virus (epidemiology),
  - the abundance of alleles in a population (population genetics),
  - or just to make the assembly of the (yet unassembled) genome of a related species a little bit easier.

de novo Genome Assembly

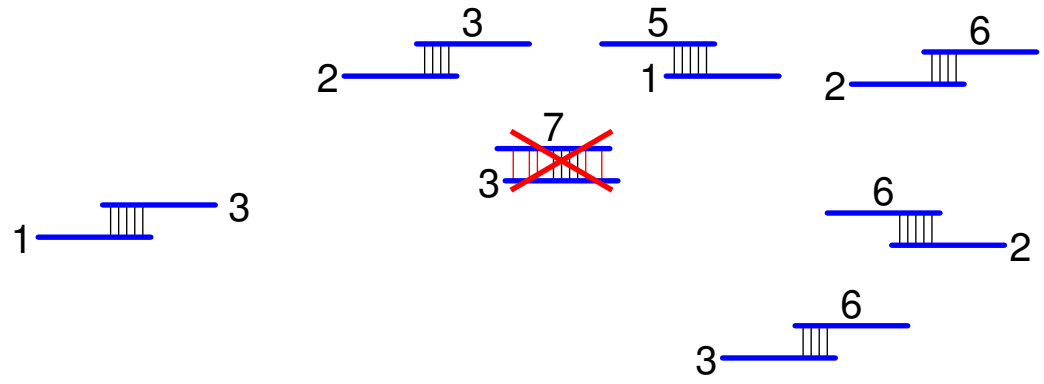
- If no reference genome exists, assembling the sequenced genome is much harder.
- We have to find overlapping reads to stitch them together to longer and longer contigs.

Overlap Layout Consensus (as in CAP3): Overlap search



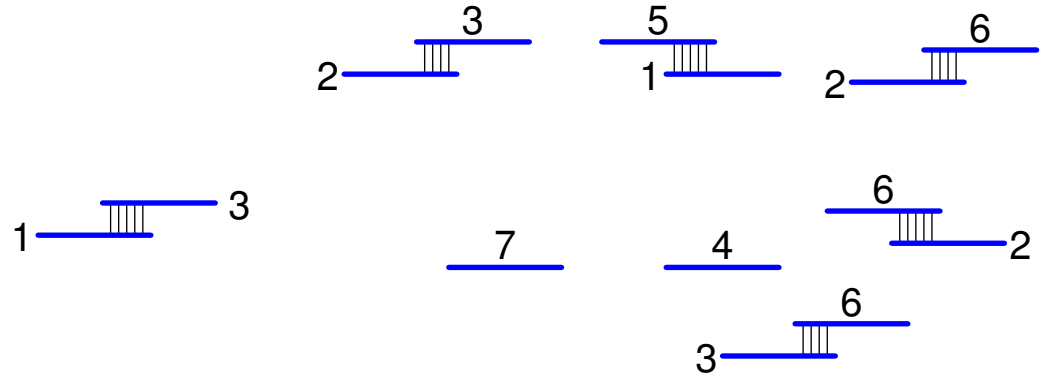
1. concatenate reads (separated by a separation character)
2. identify candidate overlaps (local alignments of reads against the concatenated string)
   - discard the trivial matches (i.e. read i matches itself)
   - each pair only once (result of i vs j should be identical to j vs i)

Overlap Layout Consensus (as in CAP3): Filtering



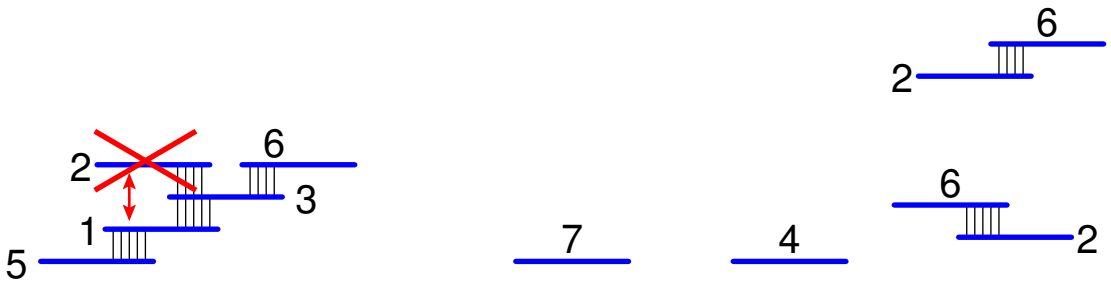
3. remove poor quality reads
4. compute global alignment for high quality pairs.
5. evaluate alignments due to
   1. minimum length
   2. minimum identity
   3. minimum similarity
   4. number of high-quality (true) mismatches
6. remove pairs that do not match thresholds 5.1-5.4.

Overlap Layout Consensus (as in CAP3): Contig building



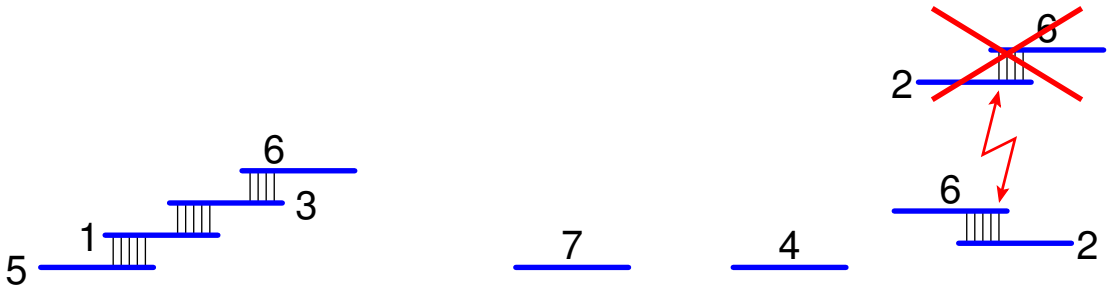
7. add all reads without overlaps

Overlap Layout Consensus (as in CAP3): Contig building



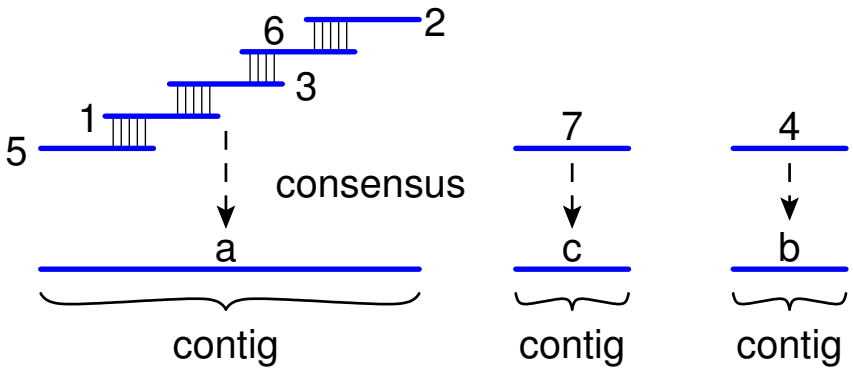
7. add all reads without overlaps
8. generate a general layout using overlapping reads (ordered with decreasing overlap scores)
9. check for incompatibilities
    1. in the layout (remove greedily from layout)

Overlap Layout Consensus (as in CAP3): Contig building



7. add all reads without overlaps
8. generate a general layout using overlapping reads (ordered with decreasing overlap scores)
9. check for incompatibilities
    1. in the layout (remove greedily from layout)
    2. between overlap candidates (remove candidate greedily)

Overlap Layout Consensus (as in CAP3): Contig building



7. add all reads without overlaps
8. generate a general layout using overlapping reads (ordered with decreasing overlap scores)
9. check for incompatibilities
   1. in the layout (remove greedily from layout)
   2. between overlap candidates (remove candidate greedily)
10. construct consensus sequence for each contig

Overlap Layout Consensus (as in CAP3): Scaffolding

(now, assume that all above contigs were constructed from many shorter reads.)

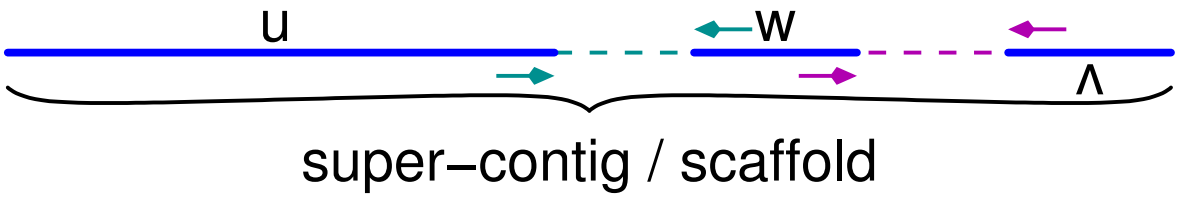
11. the set of contigs can often be extended
12. using additional information like paired-end reads (if available)
13. order contigs to bring matching paired-ends next to each other

Overlap Layout Consensus (as in CAP3): Scaffolding

(now, assume that all above contigs were constructed from many shorter reads.)

11. the set of contigs can often be extended
12. using additional information like paired-end reads (if available)
13. order contigs to bring matching paired-ends next to each other
14. orientate contigs according to the paired-ends

Overlap Layout Consensus (as in CAP3): Scaffolding

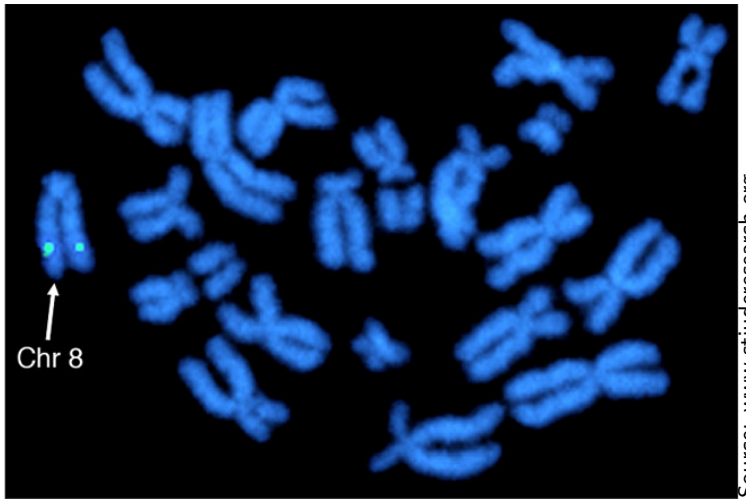


(now, assume that all above contigs were constructed from many shorter reads.)

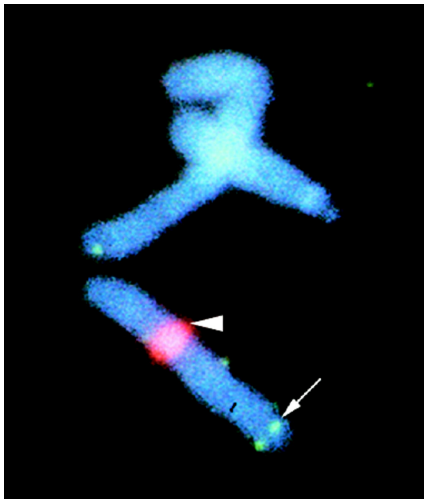
11. the set of contigs can often be extended
12. using additional information like paired-end reads (if available)
13. order contigs to bring matching paired-ends next to each other
14. orientate contigs according to the paired-ends
15. fill the gaps with N's according to the insert sizes used when preparing the sequencing library
16. the joined contigs are called super-contigs or scaffolds

Assembly Completeness and Contig Location

- usually it is not possible to easily assemble each chromosome into a single contig or scaffold
  (e.g. due to repeats, low quality regions, too low read coverage)
- thus, it can be important to locate scaffolds in the genome using, e.g., FISH (fluorescence in-situ hybridization) with genetic markers.



Source: www.stjuderesearch.org



Source: Westbrook et al. (2008)
red chromosome marker, green probe

Assembly from 2nd and 3rd generation sequencing reads

- CAP3 has been developed for Sanger sequencing reads.
- NGS reads are typically shorter and come in huge numbers.
- Thus, also the overlaps are short, producing false positives easily.

- Assembly of NGS data works along the same principles.
- However they have to employ more elaborate methods to deal with the amount of data, the short overlaps and to efficiently detect false positive overlaps.
- A number of such tools apply approaches like *de Bruijn graphs*.