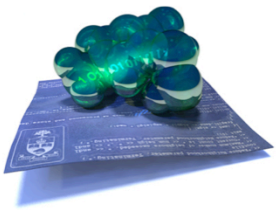
A
Bioinformatics
Course

Genome Sequencing



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

NGS – THE GAME CHANGER

Since publication of the human genome sequence in 2001, sequencing costs have dropped almost a million fold.

Dozens of small genomes can now be sequenced in less than a day.

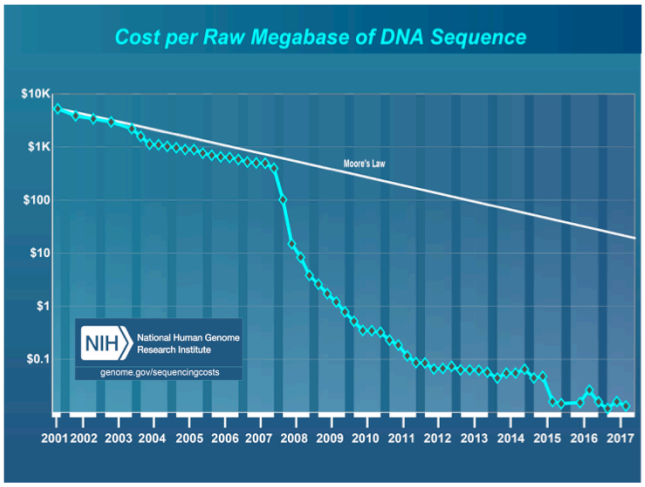
Commodity sequencing can provide personal human genomes for less than a thousand dollars.

Cancer genomics and GWAS by sequencing are now routine.

Novel applications (like RNAseq) have displaced traditional technologies (like microarrays).

Petabytes of data have to be stored and processed.

NGS is disruptive.



NHGRI (http://www.genome.gov/sequencingcosts/)

Sanger sequencing

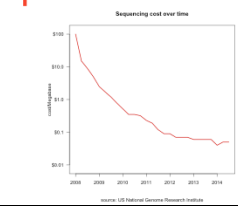
"First generation" – classical

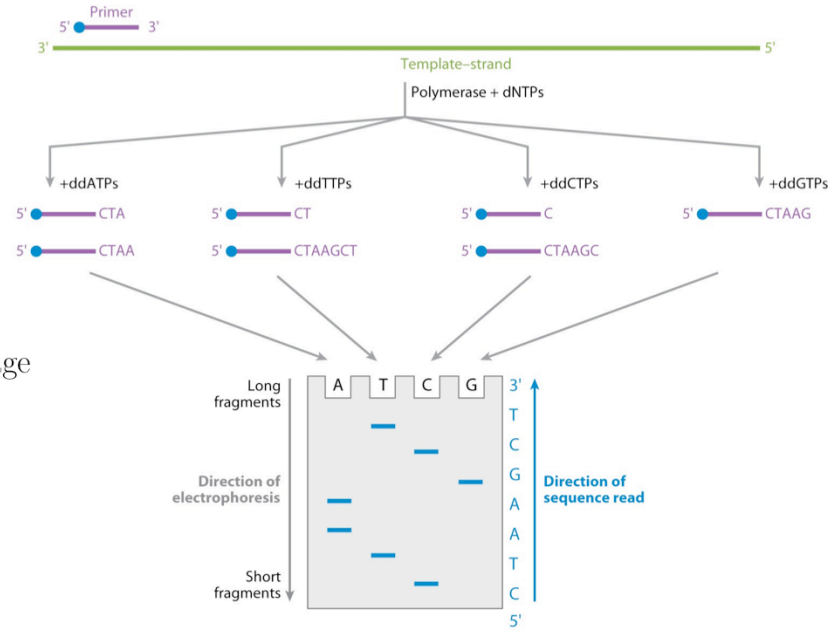
Traditional sequencing:

~800bp of high-quality reads on a capillary sequencer. 96 capillary machines: 2Mb/day

Requires ~6-fold coverage for shotgun sequencing

$6,000/Mb





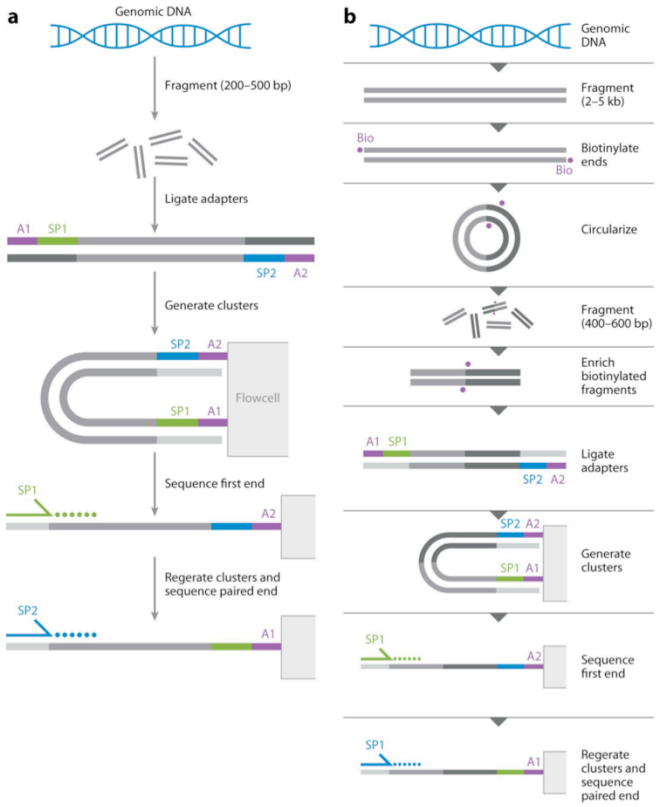
Mardis E.R (2013) Next-Generation Sequencing Platforms. *Ann. Rev. Anal. Chem.* 6:287-303

To sequence DNA, specificity has to be ensured in **two** distinct ways: (i) the reaction needs to be targeted specifically to a **unique location** in the DNA, and (ii) the extension reaction has to provide a **base-specific signal**. All sequencing strategies are subject to these fundamental requirements[1], they solve them in different ways. Sanger sequencing provides *location specificity* with a uniquely matching primer, *signal specificity* with termination nucleotides that control the size of the reaction product in a base-specific manner.

[1] Single-molecule sequencing is an exception. The requirement arises from the need to produce identical signals from multiple molecules, which in turn arises from a need to amplify the signals from individual molecules.

Library Construction

High-throughput sequencing requires high-throughput libraries.



Mardis E.R (2013) Next-Generation Sequencing Platforms. *Ann. Rev. Anal. Chem.* 6:287-303

Shotgun sequencing strategies solve the *location specificity* problem by multiplying individual molecules through PCR after separating them, and sequencing them from known adapters.

PYROSEQUENCING

"Second generation" – parallel

Roche 454 System

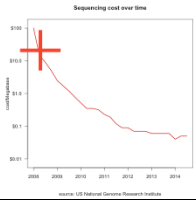
First introduced 2005.

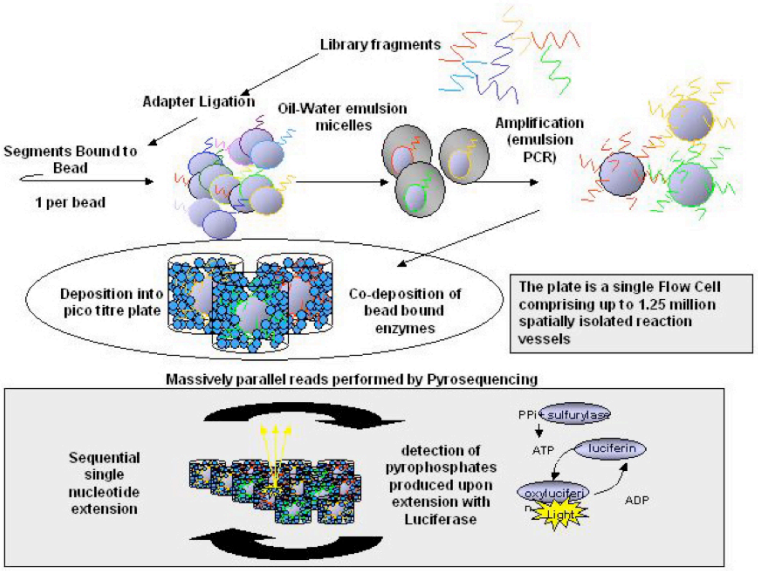
Now largely obsolete.

250-500bp reads, 800 Mb/day

Requires ~10 fold coverage for shotgun sequencing

$25/Mb





Pyrosequencing provided the first massive drop in sequencing costs. The sequence is deduced from the location, the type of nucleotides added, and the intensity of the signal, e.g. GGG creates a three-times stronger signal than G. For long homo-oligomeric sequences this degrades accuracy.

Ion Torrent

"Second generation" – parallel

Sample: single molecules, (µg)

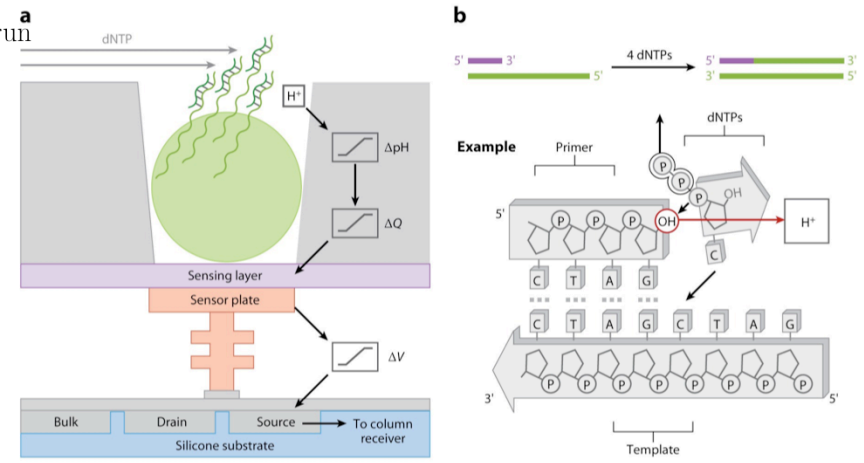
Read length: up to 400bp

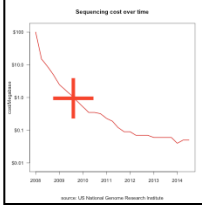
Throughput: 80 million/run

Principle: pH changes in microcompartments

Accuracy: 98%

Cost: $1/Mb





Mardis E.R (2013) Next-Generation Sequencing Platforms. *Ann. Rev. Anal. Chem.* 6:287-303

Ion torrent technology senses pH changes as the reaction proceeds in on-chip microcompartments. Just as in pyrosequencing, homo-oligomeric stretches are inferred from higher signal strengths and errors limit accuracy.

Illumina

"Second generation" – parallel

Sample: amplified library, DNA clusters

Read length: 50 - 300bp

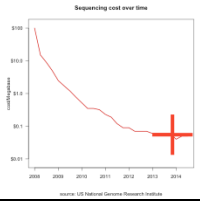
Throughput: 3 billion/run

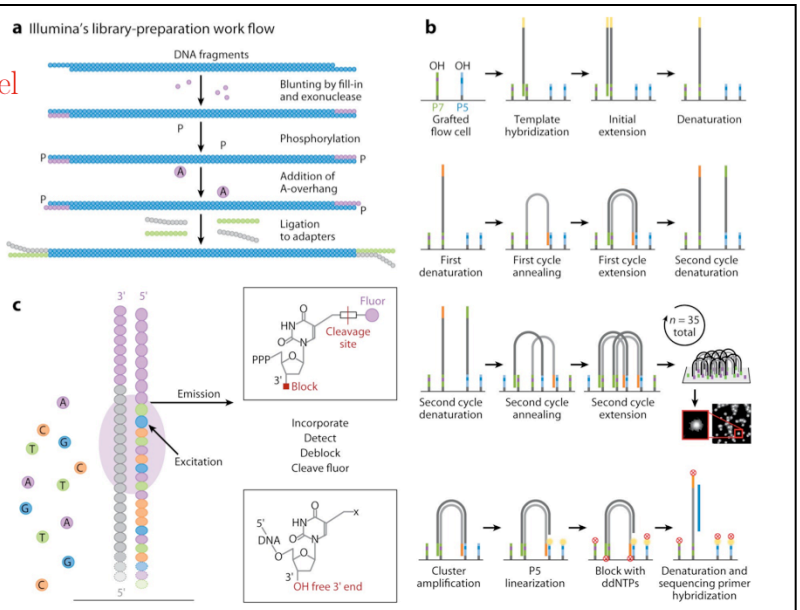
Principle: reversible dye termination

Accuracy: 98%

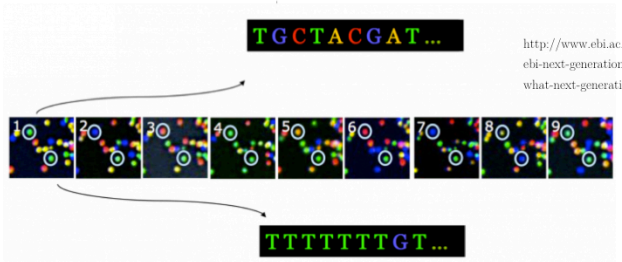
Cost: $0.05/Mb

~90% of current sequencing is done on Illumina platforms.





Mardis E.R (2013) Next-Generation Sequencing Platforms. *Ann. Rev. Anal. Chem.* 6:287-303



http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/illumina-

Illumina technology is the currently most widely used sequencing technology. It is also the technology that poses the most challenges regarding data processing since the data volumes are very, very large, and the short reads require considerable ingenuity for efficient processing. The technology has other applications however, such as RNAseq, which has largely displaced microarrays for expression profiling.

Illumina TruSeq

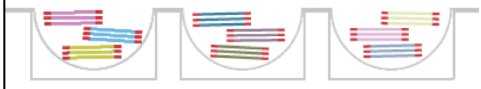
Genomic DNA is fragmented to approximately 10 Kb.
Adapters are ligated to the fragments.

Fragments are sequenced.

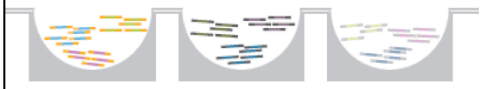
A BaseSpace app is used to assemble the long reads or phase a human genome

TruSeq Long-Read Assembly

TruSeq Phasing Analysis



Fragments are clonally amplified across 384 wells.



Fragments are sheared and labeled with unique indexes.
Fragments from all 384 wells are then pooled,
purified, and size selected.

From : http://products.illumina.com/products/truseq-synthetic-long-read-kit.html

Oxford Nanopore

"Third generation" – single-molecule

Single molecule sequencing (and resequencing) ...

No (theoretical) limit to read-length ...

Virtually no sample preparation ...

No quality degradation with read length ...

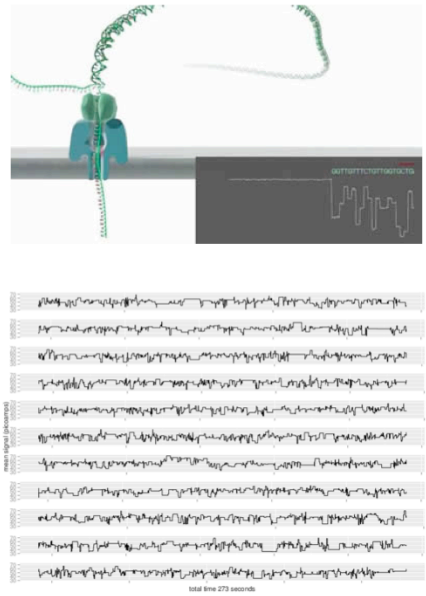
Hand-held systems, $900 ...

1000bases/min/pore, 10s of GB/day/chip, scalable to human genomes in minutes ...

Cost comparable to cheapest current technology (i.e. cents/Mb) ...

First out in the field in April 2014...

Final verdict uncertain due to accuracy issues.



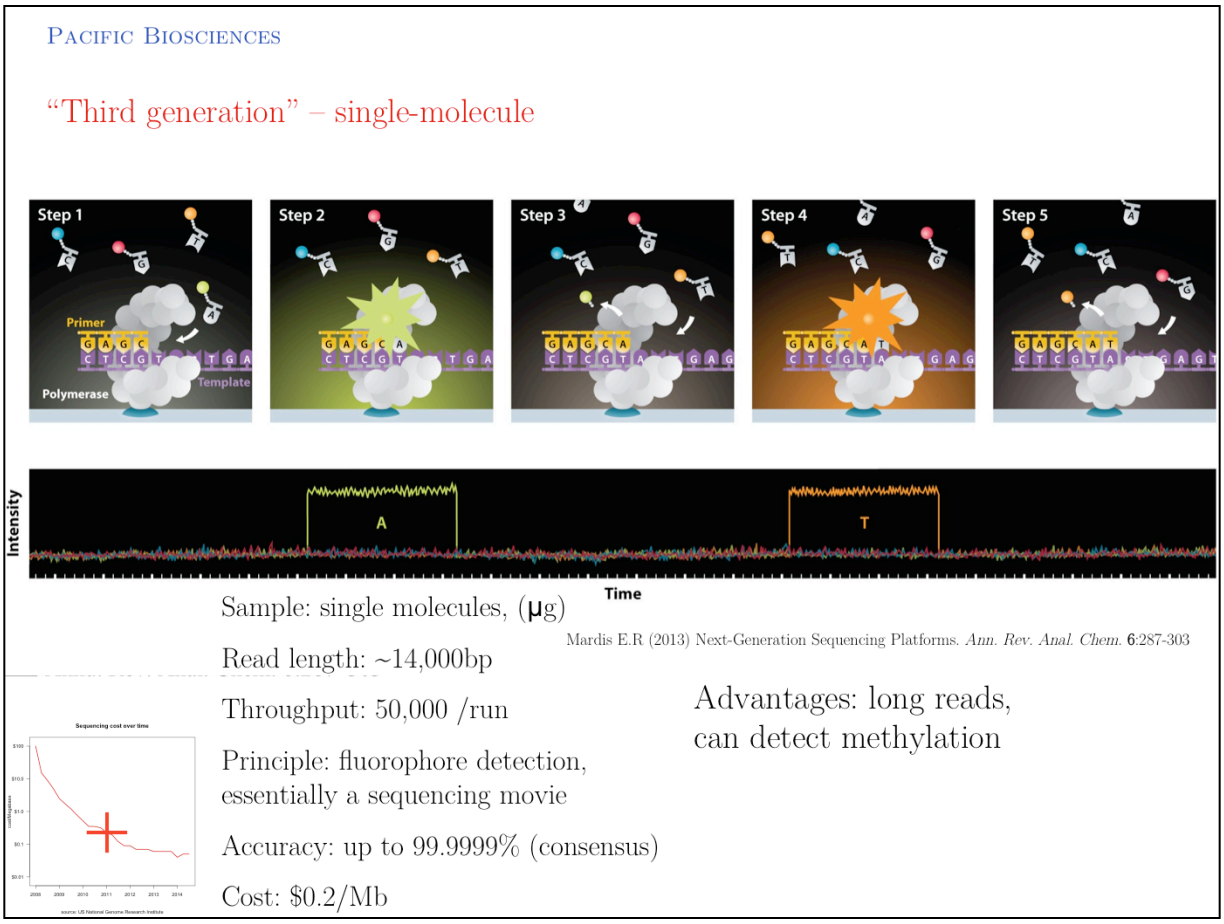
Loman, Nicholas (2014): Wiggle plot showing Oxford Nanopore signal data for a P. aeruginosa read. http://dx.doi.org/10.6084/m9.figshare.1053026

Oxford Nanopore is marketing sequencers at the price of and with the form-factor of an iPhone. It plugs into your laptop computer's USB port...

Reports have it that an entire lambda phage genome can be sequenced in one go. The technology is already impressive, and still maturing. There may be a problem with deletion errors (*i.e.* bases being skipped). Experience in 2014 has shown that these problems are significant and they need to be addressed - but single-molecule sequencing has the potential to displace everything else.

Moreover, many of the information-processing problems of NGS data arise from the need to sequence short reads at high-coverage and to reassemble them. With the very large read-lengths of nanopores, these problems will be obviated. That said, Illumina technology is catching up.

see also:
http://biomickwatson.wordpress.com/2014/09/07/thoughts-on-oxford-nanopores-minion-mobile-dna-sequencer/



The Pacific Biosciences system is essentially a highly-parallel confocal microscope. Since the microscope can restrict illumination to a zeptoliter scale volume, fluorescence of the phosphate-bound flourophore is only detected when the new nucleotide is bound, until it is released. In effect, the instrument acquires a movie of a single polymerase molecule doing its work.

The only disadvantage is the very large and expensive instrument.

PacBio

Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum.

Plant genomes, and eukaryotic genomes in general, are typically repetitive, polyploid and heterozygous, which complicates genome assembly. The short read lengths of early Sanger and current next-generation sequencing platforms hinder assembly through complex repeat regions, and many draft and reference genomes are fragmented, lacking skewed GC and repetitive intergenic sequences, which are gaining importance due to projects like the Encyclopedia of DNA Elements (ENCODE). Here we report the whole-genome sequencing and assembly of the desiccation-tolerant grass *Oropetium thomaeum*. Using only single-molecule real-time sequencing, which generates long (>16 kilobases) reads with random errors, we assembled 99% (244 megabases) of the *Oropetium* genome into 625 contigs with an N50 length of 2.4 megabases. Oropetium is an example of a 'near-complete' draft genome which includes gapless coverage over gene space as well as intergenic sequences such as centromeres, telomeres, transposable elements and rRNA clusters that are typically unassembled in draft genomes. *Oropetium* has 28,466 protein-coding genes and 43% repeat sequences, yet with 30% more compact euchromatic regions it is the smallest known grass genome. The *Oropetium* genome demonstrates the utility of single-molecule real-time sequencing for assembling high-quality plant and other eukaryotic genomes, and serves as a valuable resource for the plant comparative genomics community.

VanBuren, R. *et al.* (2105) Nature. Published online 11 November 2015

The PacBio technology is indeed practical – in this paper it has been used to sequence a plant genome.

PacBio

Single-molecule sequencing of the desiccation-tolerant grass Oropet

Notes:

32 cells, <1 week sequencing time, ~$10,000 in reagents.

Accuracy: 99.99995%

Sequence for all 18 telomeres

Three of the nine centromeric satellites are completely assembled into large inverted repeats spanning 400 kb with a base monomer length of 155 bp, and higher order structures of dimers (310 bp), trimers (465 bp) and tetramers (620 bp)

Estimate around fifty different software packages and tools involved in the data processing.

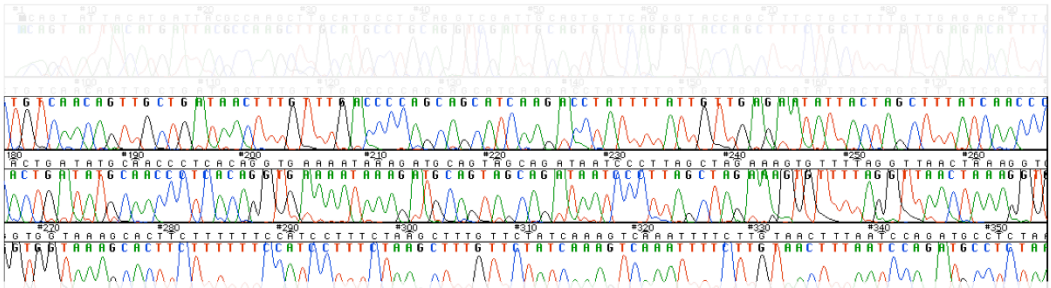
VanBuren, R. *et al.* (2105) Nature. Published online 11 November 2015

Note that the modest cost of reagents needs to be seen in the context of the instrument price which is several hundred thousand dollars.

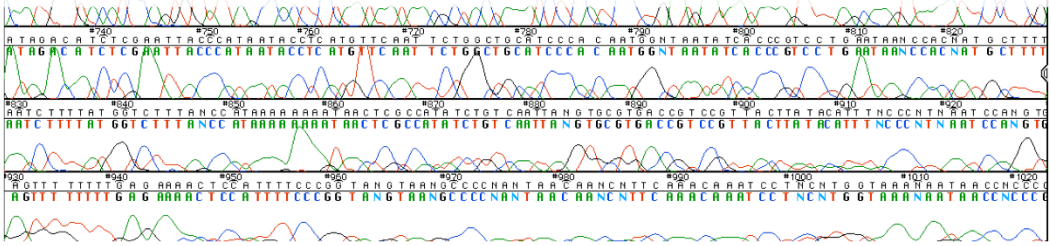
Sequencing data has to be stored and maintained. This includes data, per-experiment metadata and per-base metadata. Sequences need to go through:

1. base calling
2. sequence trimming
3. vector trimming
4. removing contaminants
5. assembly (or variant calling)

Base Calling



Extracting sequence from raw data. Depends on the sequencing technology although traditionally the term is applied to fluorescent dye labelled chromatograms. Employs standard approaches of signal processing and machine learning. All sequencing methods have characteristic error profiles.



FASTQ

A standard format for NGS (and capillary sequencers)

- Line 1 begins with '@' , followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with '+', optionally followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence, and must contain the same number of symbols.

(after Wikipedia)

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Quality values: from 0 to 93 – encoded as ASCII 33 (!) to 126 (~)

```
ASCII codes 33 – 126:
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

30

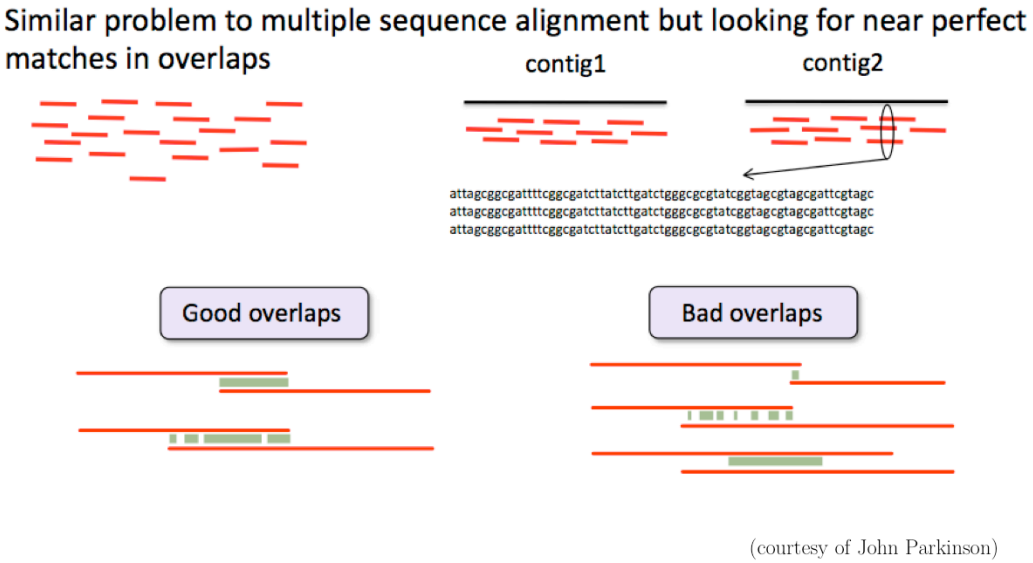
$$Q = -10 \log_{10} P$$

i.e. Q=10: 1 in 10 (90%); Q=30: 1 in 1000 errors (99.9% correct)

Desired quality? Depends ... but >30 ("?") is probably good.

The result of "base calling" is sequence, commonly stored in FASTQ files that store sequence and sequencing quality for further processing.

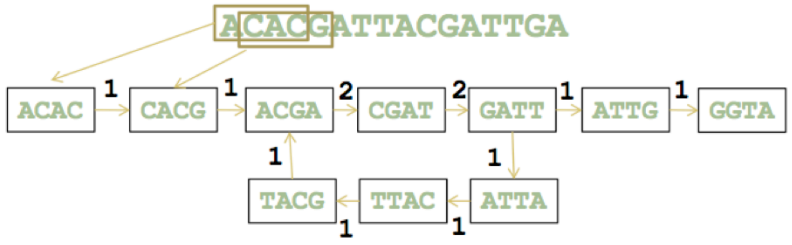
Overlap extension (e,g, Abyss)



(courtesy of John Parkinson)

De Bruijn graphs (e.g. Velvet)

Newer algorithms use de Bruijn graphs in which sequences are represented by overlapping Kmers



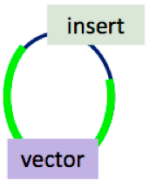
If you can determine the path that spans through these nodes then you should be able to recreate the original sequence

Advantages of the de Bruijn approach include reduction in the amount of memory needed, as redundant sequences need only be represented once, also Kmers are also stored and manipulated more efficiently through the use of hash tables

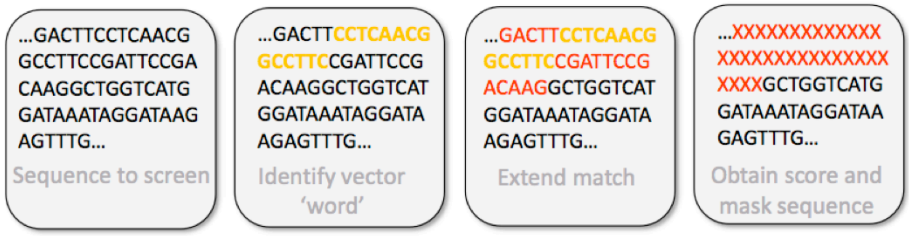
(courtesy of John Parkinson)

Vector Trimming

In addition to removing low quality sequence also need
To remove contaminating vector sequence - several tools now available
- Crossmatch and Lucy (TIGR)
- Requires a database of vector sequences to compare against



Crossmatch uses dynamic programming to identify and build out from minimum word matches (by default 14 bases)



Two major parameters
- word size (exact match that must first be identified, by default set to 14)
- minimum alignment score to trigger a match

Varying these can affect sensitivity

(courtesy of John Parkinson)

SAM (and BAM)

Sequence Alignment Map: a format for storing large nucleotide sequence alignments (often several Gigabytes large!).

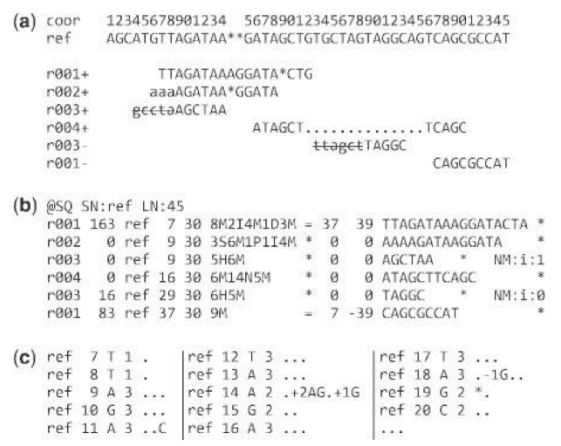
(see: http://samtools.sourceforge.net/)

Reference genome and mapped reads

SAM

(with extended CIGAR strings)

PILEUP

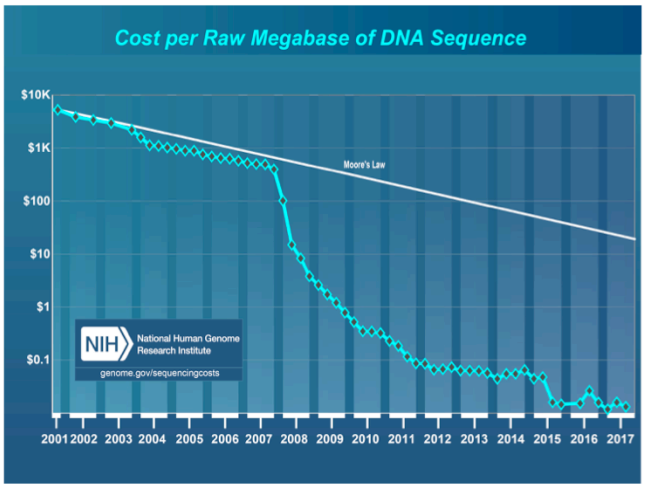


BAM is just Binary (compressed) SAM.

Li *et.al.* (2009) http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/

Storage

NGS sequence production is outpacing Moore's law. This means data storage is an expanding part of total sequencing cost. Moreover, sequence *acquisition* costs end with determining the sequence, sequence *storage* costs continue over time.



NHGRI (http://www.genome.gov/sequencingcosts/)

This (among other aspects) has led to a shift of storage and processing from lab-based infrastructure to commodity storage and compute services in the "cloud".

Genome computing today *is* Cloud computing.

(Google Server Farm)

For example the 1000 Genomes Project (http://www.1000genomes.org/) made its data publically available on the Amazon Web Services (AWS) cloud storage in 2012–about 200TB. But if you want to analyze this, how are you going to read it into your machine? You don't. These analyses actually have to be run on distributed servers as well. The golden age of the desktop bioinformatician may be coming to an end. (See: http://aws.amazon.com/1000genomes/ for access details and how to compute with the data.) It must however be emphasized, that a part of the problem lies in technical issues with the determination process that make it prudent to store raw experimental data - the actual genomes are much more compact. Some advocate therefore not storing the experimental data at all, but only the biological sample–to be resequenced whenever necessary. Indeed, obtaining properly validated and consented human genome samples is a bottleneck in itself.

http://steipe.biochemistry.utoronto.ca/abc

B O R I S . S T E I P E @ U T O R O N T O . C A

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA