Introduction to mapping/alignment with short read sequence data

Ian Dworkin & Chris Chandler

KBS June 8th 2011

The book!!!!

- Practical Computing for Biologists

  Steven Haddock and Casey Dunn 2010
- Sinauer

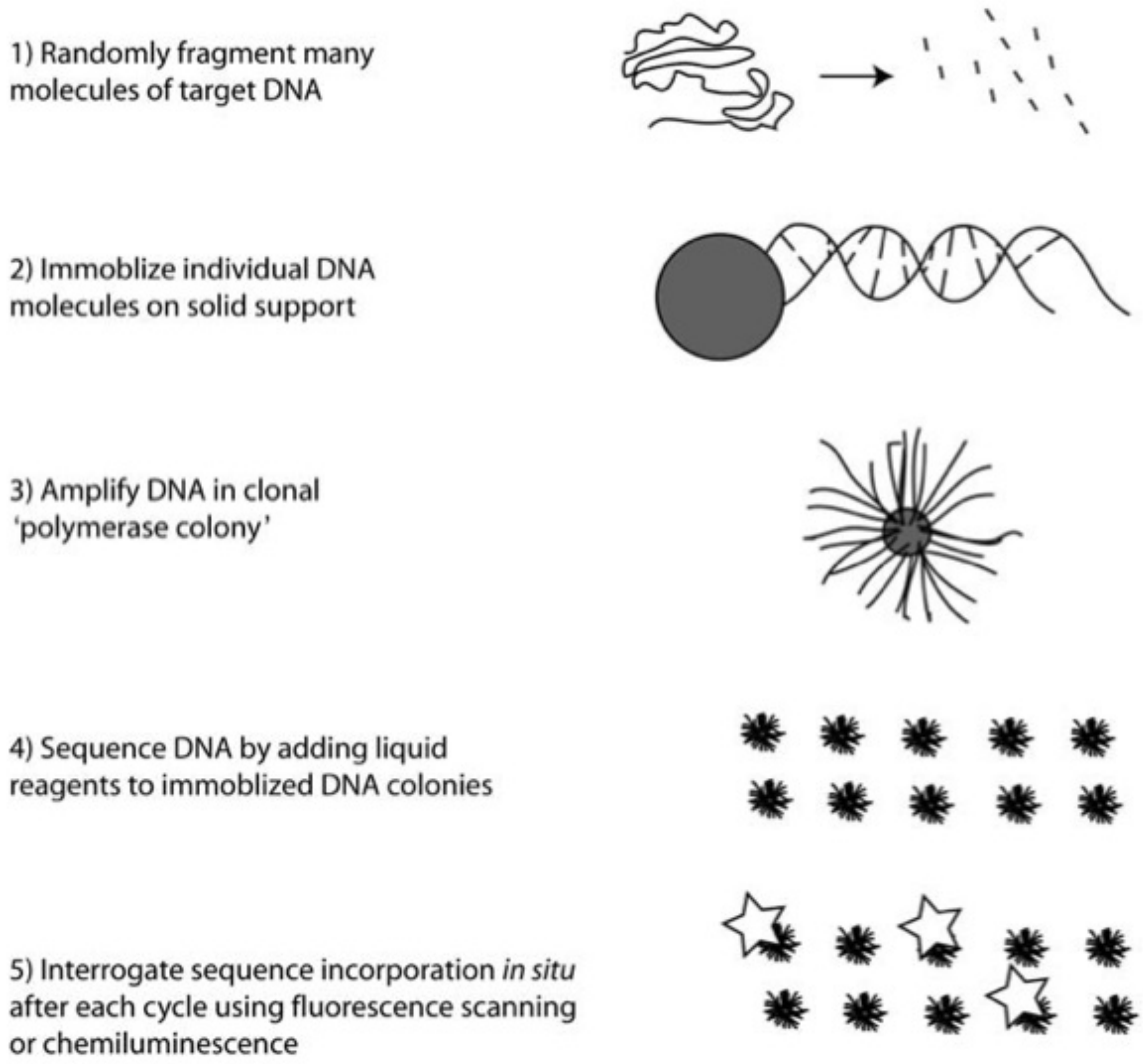
goals

- Brief introduction to 2nd and 3rd generation sequencing platforms.
- Discuss how “aligning” or mapping reads occurs.
- Some basics of data structures needed.
- Options.
- Performance of gapped vs ungapped alignments. Paired end vs single end.

What makes these sequencing technologies "next"?

- Large amount of sequencing possible, in a (relatively) short time and relatively cheaply.
- Infrastructure for the technologies is quite different than Sanger.
- Different chemistries.
- Library based sequencing (instead of individual unique amplicons).
- SCALING!
- Shorter read lengths (for now).

Some general principles (Illumina/ Roche454/ABISolid)



1) Randomly fragment many molecules of target DNA

2) Immoblize individual DNA molecules on solid support

3) Amplify DNA in clonal 'polymerase colony'

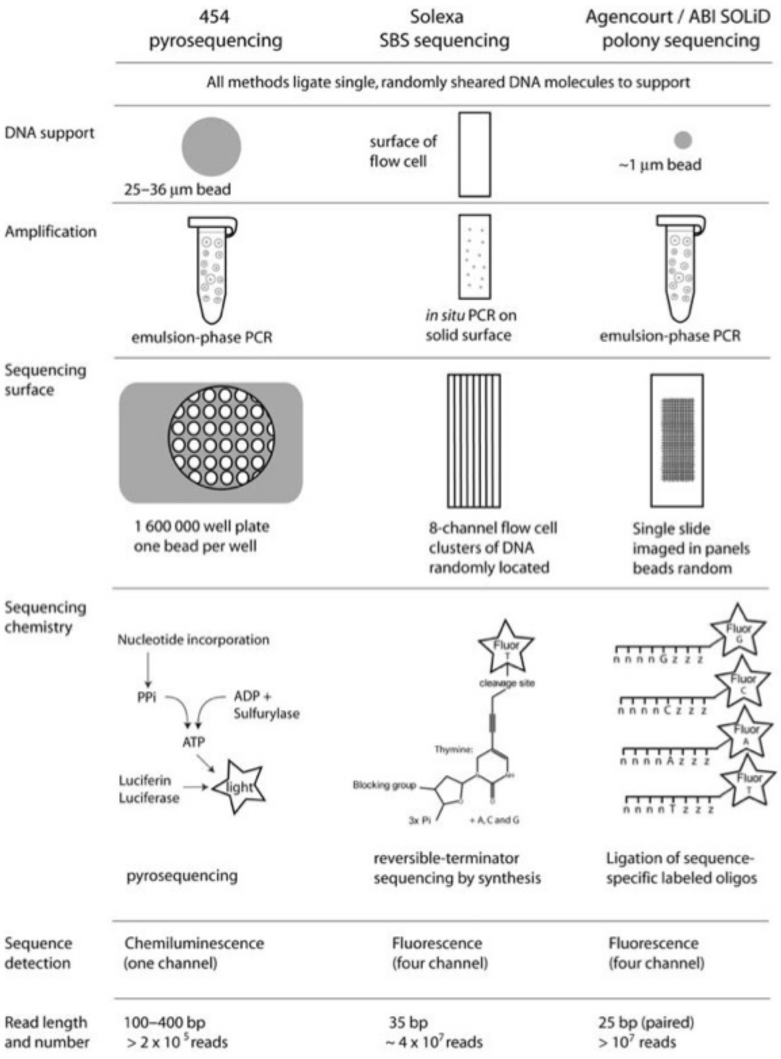
4) Sequence DNA by adding liquid reagents to immoblized DNA colonies

5) Interrogate sequence incorporation *in situ* after each cycle using fluorescence scanning or chemiluminescence

Fig. 2 A generalized description of the steps common to next-generation genome sequencing technologies. All these technologies involve genomic DNA random fragmentation, immobilization of single molecules on a solid support (a bead or planar solid surface), amplification by PCR, and subsequent *in situ* biochemical interrogation of the template DNA at each base in turn.

Hudson 2008

A few general differences..

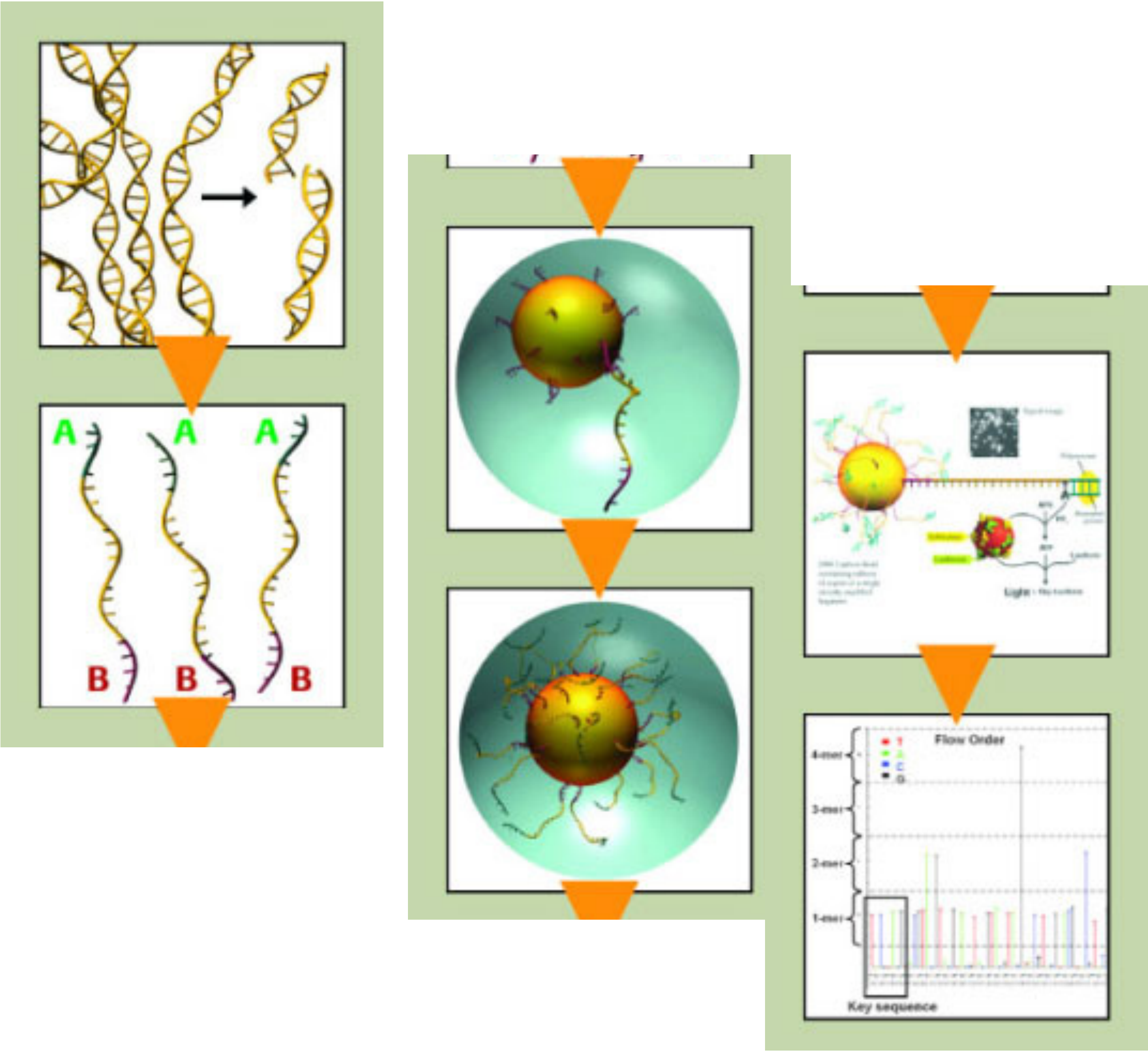


	454 pyrosequencing	Solexa SBS sequencing	Agencourt / ABI SOLiD polony sequencing
	All methods ligate single, randomly sheared DNA molecules to support		
DNA support	25–36 μm bead	surface of flow cell	~1 μm bead
Amplification	emulsion-phase PCR	*in situ* PCR on solid surface	emulsion-phase PCR
Sequencing surface	1 600 000 well plate one bead per well	8-channel flow cell clusters of DNA randomly located	Single slide imaged in panels beads random
Sequencing chemistry	pyrosequencing	reversible-terminator sequencing by synthesis	Ligation of sequence-specific labeled oligos
Sequence detection	Chemiluminescence (one channel)	Fluorescence (four channel)	Fluorescence (four channel)
Read length and number	100–400 bp $> 2 \times 10^5$ reads	35 bp $\sim 4 \times 10^7$ reads	25 bp (paired) $> 10^7$ reads

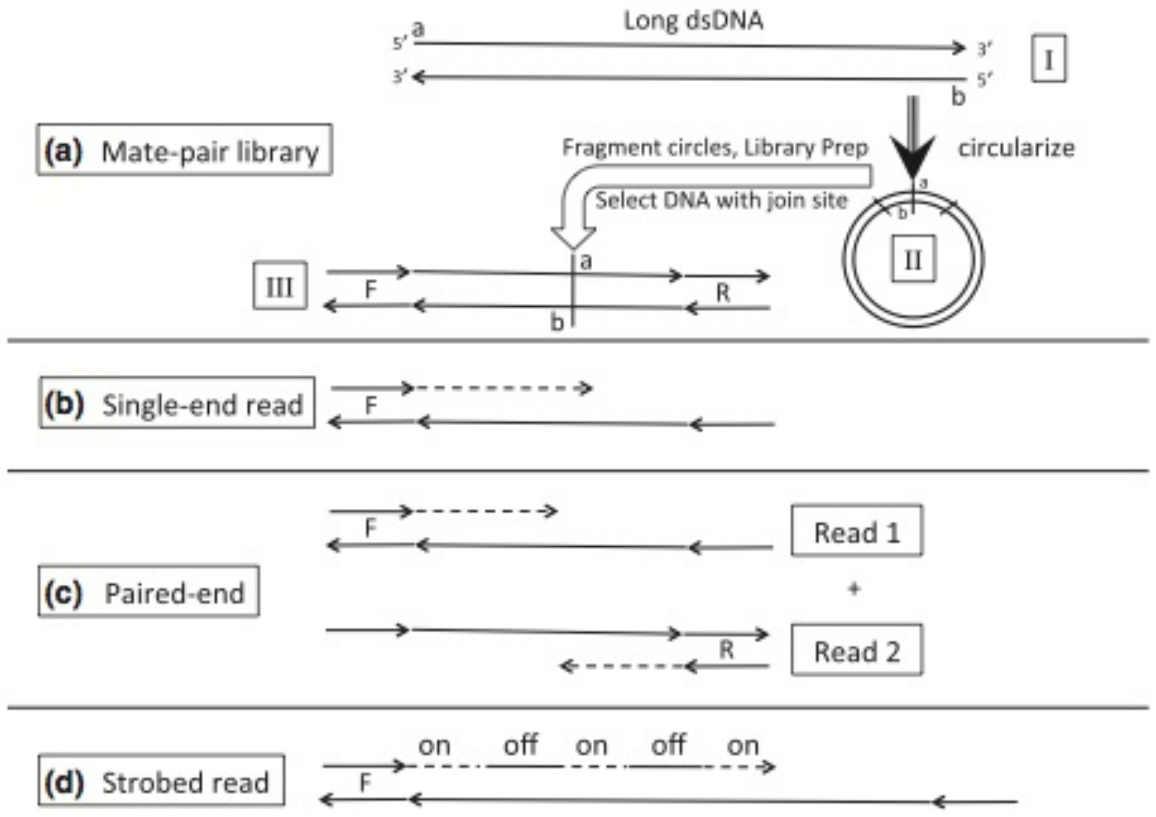
Fig. 3 A description of the key features of, and differences between, the three commercially available next-generation sequencing methods. The major steps in each procedure are arranged in the order in which they are performed by the operator or sequencing instrument. All three technologies share a common workflow, but differ greatly in the type of solid support used and the chemistry used to interrogate the DNA base pairs.

Hudson 2008

http://www.454.com/products-solutions/how-it-works/index.asp

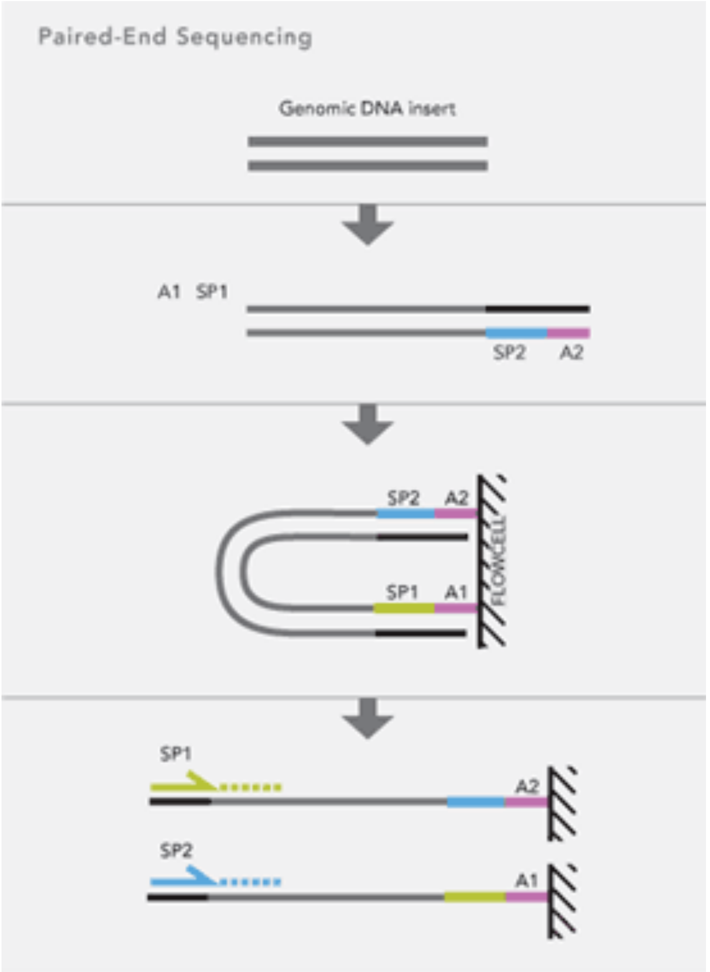


Among the many useful things these technologies allow is to "overcome" the short read length by doing multiple short reads from the same PCR amplicon.



Glenn 2011

Illumina Paired end



Paired-End Sequencing

Adapter (A1 and A2) with sequencing primer sites (SP1 and SP2) are ligated onto DNA fragments. Template clusters are formed on the flow cell by bridge amplification and then sequenced by synthesis from the paired primers sequentially.

Inserts on the order of 200-500bp

http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn

The Roche454 FLX (or whatever version)…

- The mate pairs allow for larger inserts of different sizes.
- Currently ranging between 3-12kb (??).
- However with this approach the actual length of the read from each side is effectively cut in half (not true for PE with Illumina).

The technologies are changing so fast..

- We are all trying to keep up with latest developments.
- The good news is everything is getting better and cheaper very quickly.
- However it is still very important to pick the appropriate technology for your particular application (i.e. 454 for assembly, but Illumina for resequencing, chipSeq, RNAseq…).

Comparison among platforms

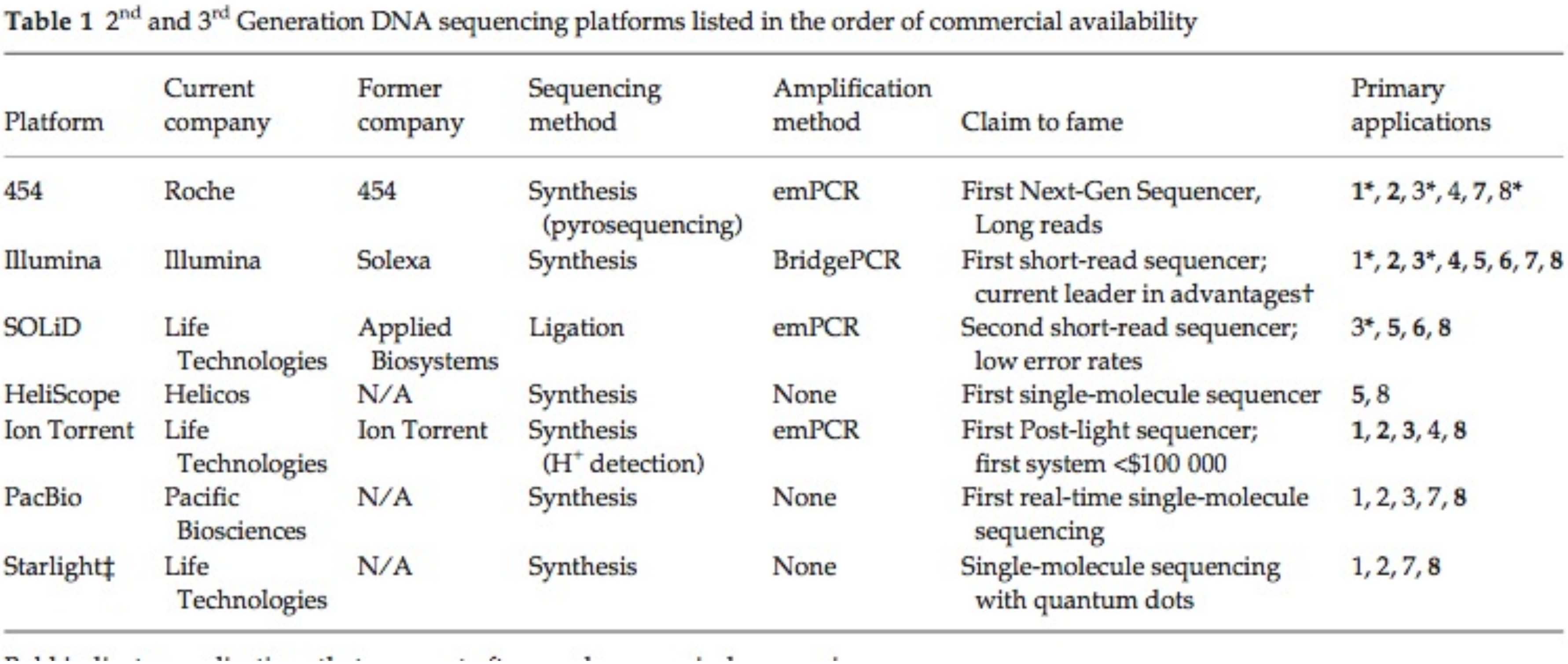


Table 1 2nd and 3rd Generation DNA sequencing platforms listed in the order of commercial availability

Platform	Current company	Former company	Sequencing method	Amplification method	Claim to fame	Primary applications
454	Roche	454	Synthesis (pyrosequencing)	emPCR	First Next-Gen Sequencer, Long reads	1*, 2, 3*, 4, 7, 8*
Illumina	Illumina	Solexa	Synthesis	BridgePCR	First short-read sequencer; current leader in advantages†	1*, 2, 3*, 4, 5, 6, 7, 8
SOLiD	Life Technologies	Applied Biosystems	Ligation	emPCR	Second short-read sequencer; low error rates	3*, 5, 6, 8
HeliScope	Helicos	N/A	Synthesis	None	First single-molecule sequencer	5, 8
Ion Torrent	Life Technologies	Ion Torrent	Synthesis (H^+ detection)	emPCR	First Post-light sequencer; first system <$100 000	1, 2, 3, 4, 8
PacBio	Pacific Biosciences	N/A	Synthesis	None	First real-time single-molecule sequencing	1, 2, 3, 7, 8
Starlight‡	Life Technologies	N/A	Synthesis	None	Single-molecule sequencing with quantum dots	1, 2, 7, 8

Glenn 2011

Comparisons among platforms

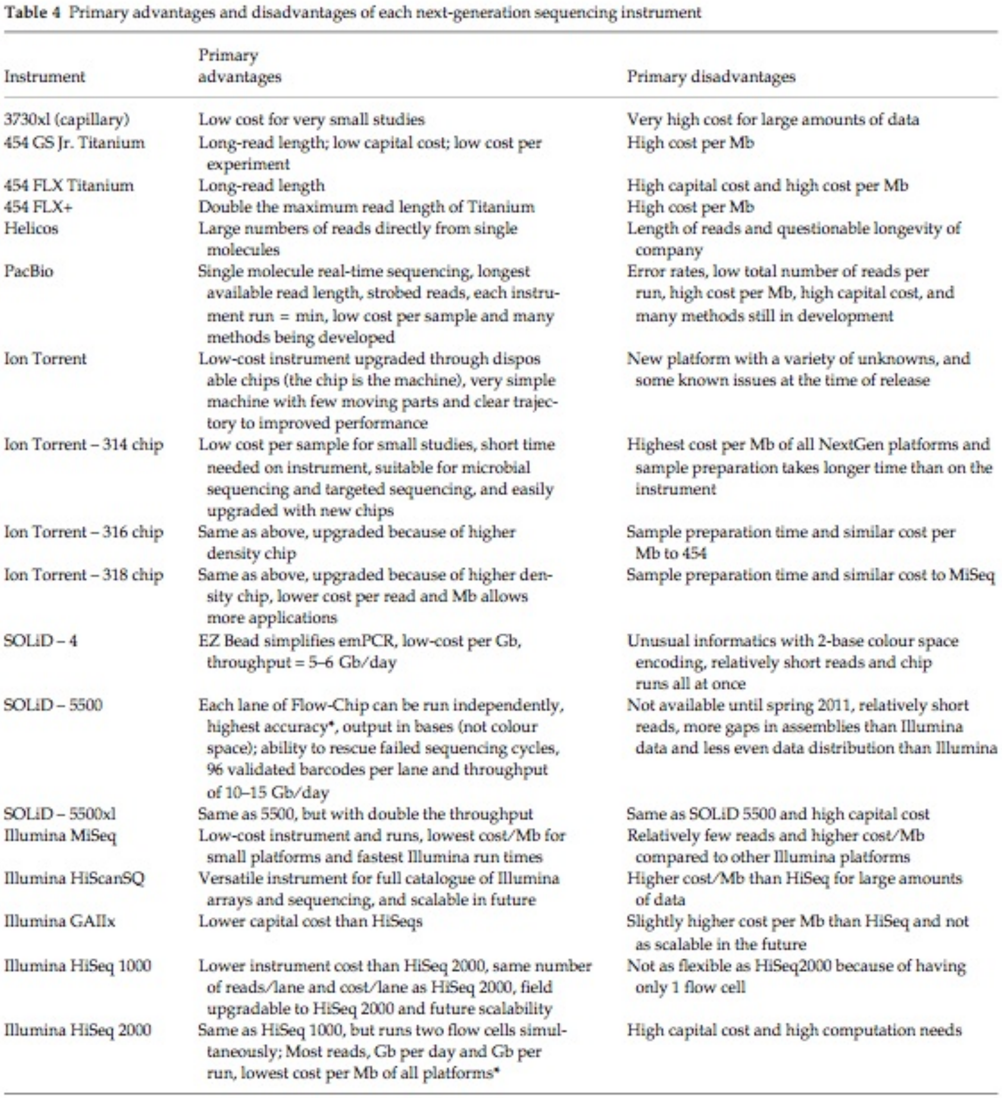


Table 4 Primary advantages and disadvantages of each next-generation sequencing instrument

Instrument	Primary advantages	Primary disadvantages
3730xl (capillary)	Low cost for very small studies	Very high cost for large amounts of data
454 GS Jr. Titanium	Long-read length; low capital cost; low cost per experiment	High cost per Mb
454 FLX Titanium	Long-read length	High capital cost and high cost per Mb
454 FLX+	Double the maximum read length of Titanium	High cost per Mb
Helicos	Large numbers of reads directly from single molecules	Length of reads and questionable longevity of company
PacBio	Single molecule real-time sequencing, longest available read length, strobed reads, each instrument run = min, low cost per sample and many methods being developed	Error rates, low total number of reads per run, high cost per Mb, high capital cost, and many methods still in development
Ion Torrent	Low-cost instrument upgraded through disposable chips (the chip is the machine), very simple machine with few moving parts and clear trajectory to improved performance	New platform with a variety of unknowns, and some known issues at the time of release
Ion Torrent – 314 chip	Low cost per sample for small studies, short time needed on instrument, suitable for microbial sequencing and targeted sequencing, and easily upgraded with new chips	Highest cost per Mb of all NextGen platforms and sample preparation takes longer time than on the instrument
Ion Torrent – 316 chip	Same as above, upgraded because of higher density chip	Sample preparation time and similar cost per Mb to 454
Ion Torrent – 318 chip	Same as above, upgraded because of higher density chip, lower cost per read and Mb allows more applications	Sample preparation time and similar cost to MiSeq
SOLiD – 4	EZ Bead simplifies emPCR, low-cost per Gb, throughput = 5–6 Gb/day	Unusual informatics with 2-base colour space encoding, relatively short reads and chip runs all at once
SOLiD – 5500	Each lane of Flow-Chip can be run independently, highest accuracy*, output in bases (not colour space); ability to rescue failed sequencing cycles, 96 validated barcodes per lane and throughput of 10–15 Gb/day	Not available until spring 2011, relatively short reads, more gaps in assemblies than Illumina data and less even data distribution than Illumina
SOLiD – 5500xl	Same as 5500, but with double the throughput	Same as SOLiD 5500 and high capital cost
Illumina MiSeq	Low-cost instrument and runs, lowest cost/Mb for small platforms and fastest Illumina run times	Relatively few reads and higher cost/Mb compared to other Illumina platforms
Illumina HiScanSQ	Versatile instrument for full catalogue of Illumina arrays and sequencing, and scalable in future	Higher cost/Mb than HiSeq for large amounts of data
Illumina GAIIx	Lower capital cost than HiSeqs	Slightly higher cost per Mb than HiSeq and not as scalable in the future
Illumina HiSeq 1000	Lower instrument cost than HiSeq 2000, same number of reads/lane and cost/lane as HiSeq 2000, field upgradable to HiSeq 2000 and future scalability	Not as flexible as HiSeq2000 because of having only 1 flow cell
Illumina HiSeq 2000	Same as HiSeq 1000, but runs two flow cells simultaneously; Most reads, Gb per day and Gb per run, lowest cost per Mb of all platforms*	High capital cost and high computation needs

Glenn 2011

Ok… I now have 0.25Tb of sequence data… what do we do first

- For most applications, the unique short reads by themselves are of little use.

- In general most people try to "piece the data together" in some fashion.
- In general this is either de-novo assembly, or syntenic assembly (i.e. aligning/mapping reads).
- Hybrid approaches are becoming more common.

Syntenic assembly

- **Syntenic** assembly *assumes* that you have some sort of **reference** genome that you can use as a scaffold to build your genome (or transcripts).
- You make an assumption that the locations of your new sequence reads are syntenic with that of the reference genome?
- How could this possibly go wrong?

Assumptions/issues with syntenic assembly?

- Insertion/deletions (relative to reference genome).
- High levels of sequence variation (relative to reference).
- Low complexity DNA (repeats)
- Recombination
- Re-arrangements
- Gene duplication
- Quality of the reference genome
- (reference is incomplete, and order may be incorrect).

So how do we map reads?

- We can start where we left off yesterday..
- As CTB described, Blast begins with a “seed” of length 11 (default).
- Blast generates a database of the reference genome data (genomic DNA, transcripts etc) of 11bp sequences, and looks for exact (?) matches.
- It then takes all matches (which is a small subset) and used the Smith-Waterman algorithm (CTB will explain later) to find the “best” matches.
- This needs to be done as SMA and other approaches are slow.

Why not use Blast for NGS data?

- Too Slow!
- “homology” over short regions can be difficult for short reads.

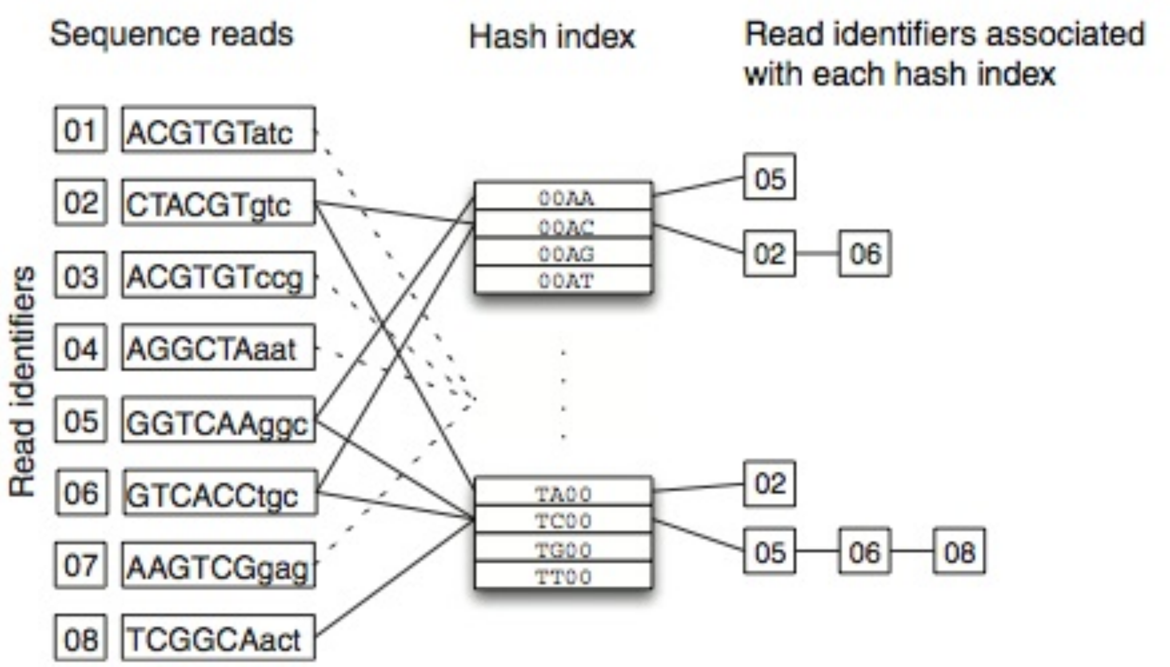
NGS alignment/mapping

- Most approaches still use this two step process of generating a subset of possible matches followed by a (slow) refinement step.
- The major differences is how they go about doing this.

Hash tables

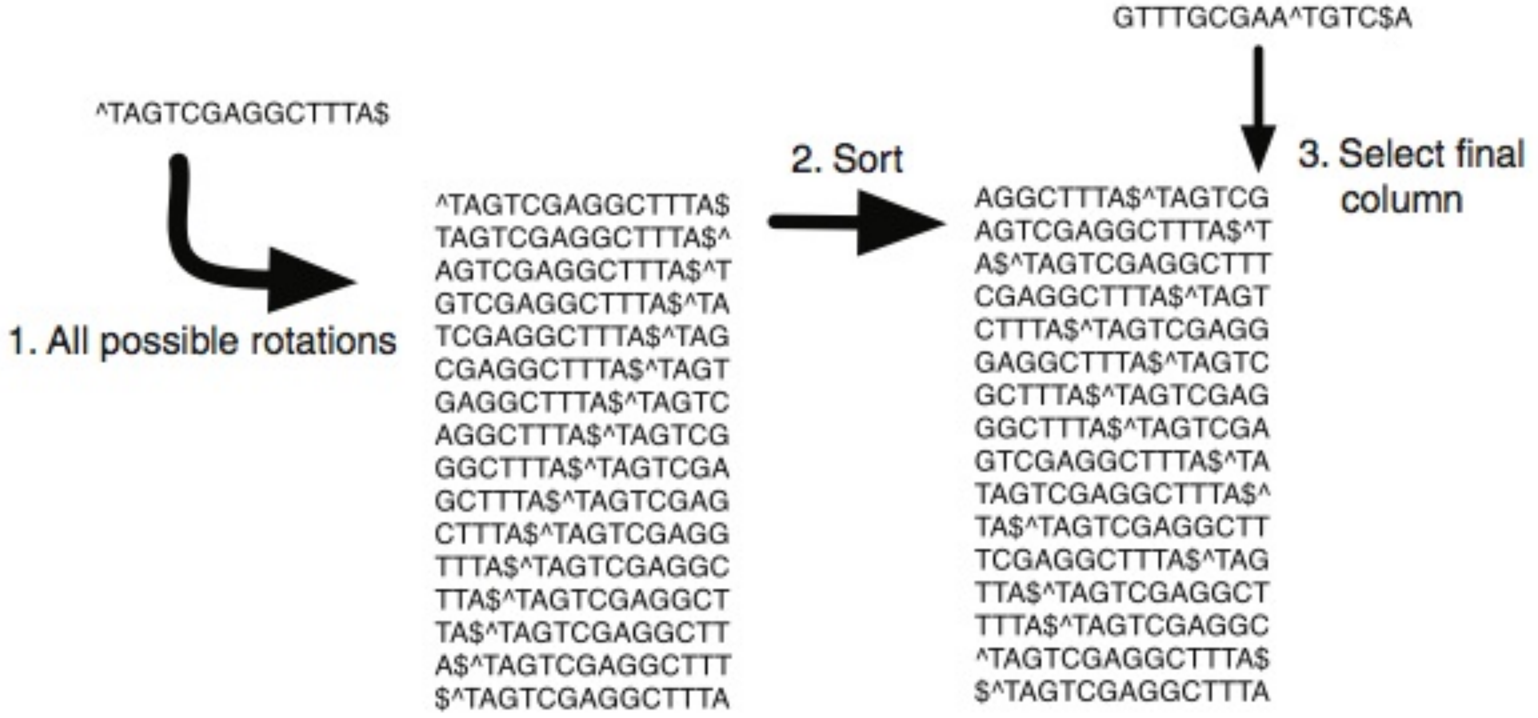
- Some of software uses **hash tables**.
- These are also called **associative arrays**, **maps** and in python **dictionaries**.
- The basic idea is that instead of some sort of numerical indexing, the data structure (the hash table) uses a key-value pairing system.
- You used one yesterday in python {}.
- Pairs = {}

Hash table.



Flickey & Birney 2009

Burrows Wheeler.



^TAGTCGAGGCTTTA$

1. All possible rotations

```
^TAGTCGAGGCTTTA$
TAGTCGAGGCTTTA$^
AGTCGAGGCTTTA$^T
GTCGAGGCTTTA$^TA
TCGAGGCTTTA$^TAG
CGAGGCTTTA$^TAGT
GAGGCTTTA$^TAGTC
AGGCTTTA$^TAGTCG
GGCTTTA$^TAGTCGA
GCTTTA$^TAGTCGAG
CTTTA$^TAGTCGAGG
TTTA$^TAGTCGAGGC
TTA$^TAGTCGAGGCT
TA$^TAGTCGAGGCTT
A$^TAGTCGAGGCTTT
$^TAGTCGAGGCTTTA
```

2. Sort

```
AGGCTTTA$^TAGTCG
AGTCGAGGCTTTA$^T
A$^TAGTCGAGGCTTT
CGAGGCTTTA$^TAGT
CTTTA$^TAGTCGAGG
GAGGCTTTA$^TAGTC
GCTTTA$^TAGTCGAG
GGCTTTA$^TAGTCGA
GTCGAGGCTTTA$^TA
TAGTCGAGGCTTTA$^
TA$^TAGTCGAGGCTT
TCGAGGCTTTA$^TAG
TTA$^TAGTCGAGGCT
TTTA$^TAGTCGAGGC
^TAGTCGAGGCTTTA$
$^TAGTCGAGGCTTTA
```

GTTTGCGAA^TGTC$A

3. Select final column

^TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG$ Genomic sequence

GGTTGGTCGGATTCGGAATCACGGAAAATT^AGATTCC$G Transform

Flickey & Birney 2009

Among the many software options…

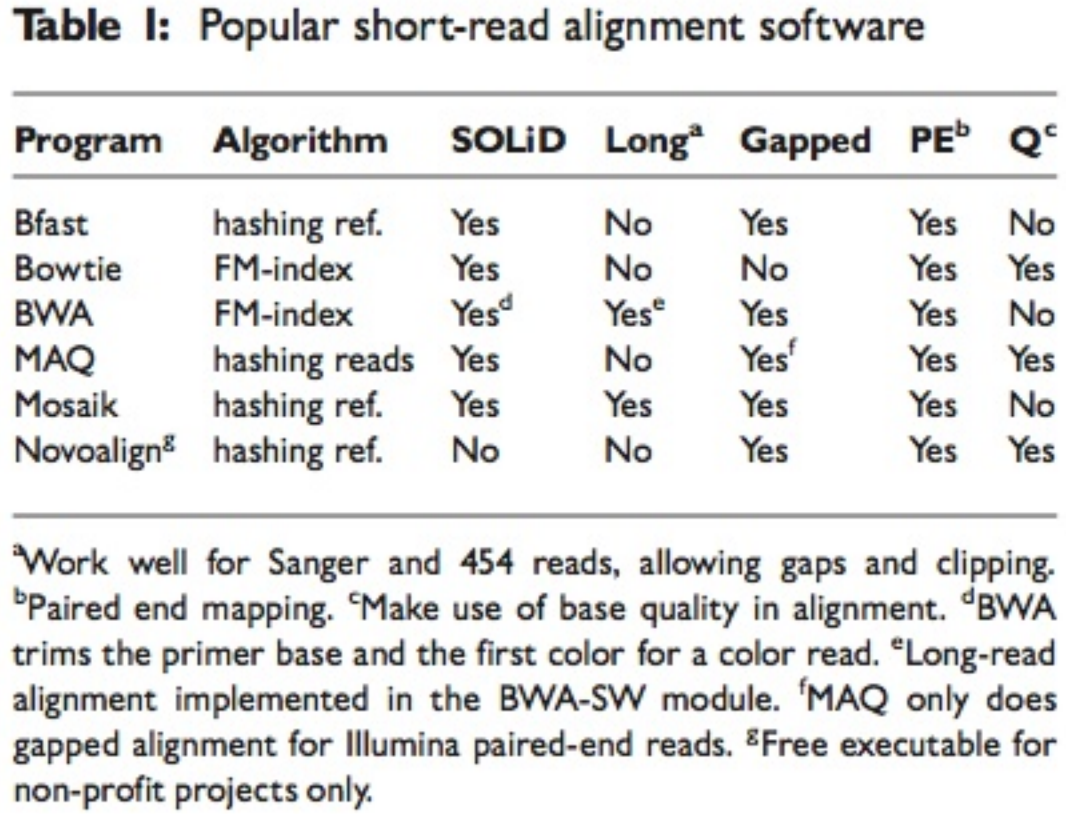


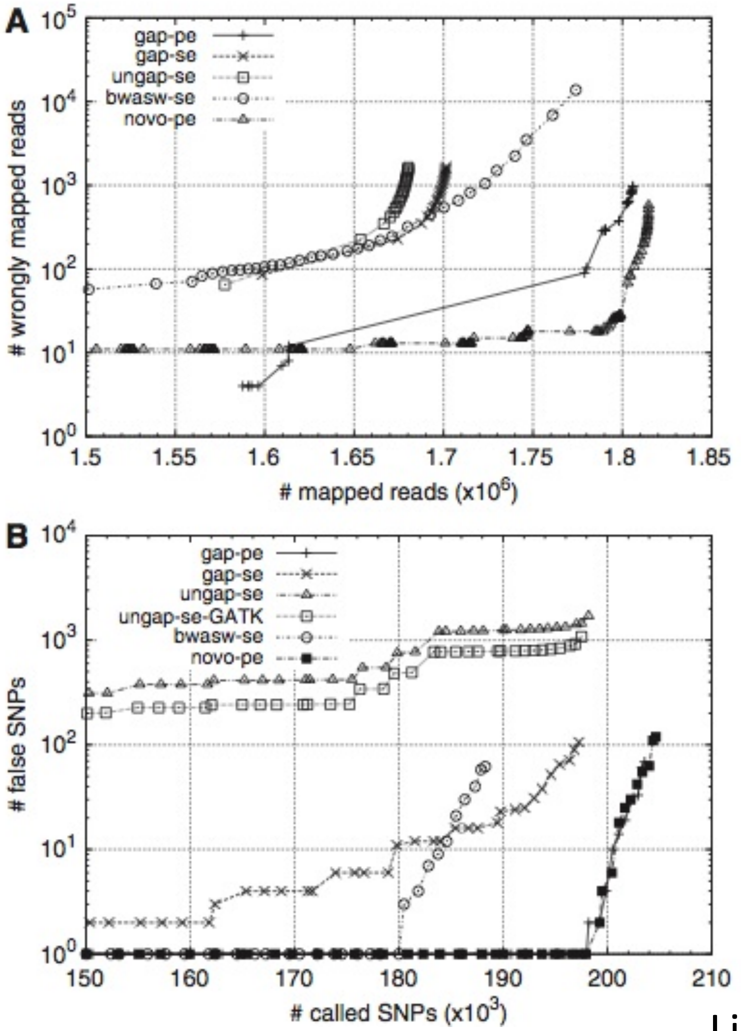
Table 1: Popular short-read alignment software

Program	Algorithm	SOLiD	Long[a]	Gapped	PE[b]	Q[c]
Bfast	hashing ref.	Yes	No	Yes	Yes	No
Bowtie	FM-index	Yes	No	No	Yes	Yes
BWA	FM-index	Yes[d]	Yes[e]	Yes	Yes	No
MAQ	hashing reads	Yes	No	Yes[f]	Yes	Yes
Mosaik	hashing ref.	Yes	Yes	Yes	Yes	No
Novoalign[g]	hashing ref.	No	No	Yes	Yes	Yes

[a] Work well for Sanger and 454 reads, allowing gaps and clipping. [b] Paired end mapping. [c] Make use of base quality in alignment. [d] BWA trims the primer base and the first color for a color read. [e] Long-read alignment implemented in the BWA-SW module. [f] MAQ only does gapped alignment for Illumina paired-end reads. [g] Free executable for non-profit projects only.

Li & homer 2011

Some comparisons among gapped/ ungapped & Single vs. Paired end (se/



Li & homer 2011

How long a list?

- http://seqanswers.com/forums/showthread.php?t=43
- http://seqanswers.com/wiki/Software
- http://seqanswers.com/wiki/Software/list
- http://en.wikipedia.org/wiki/List_of_sequence_alignment_software

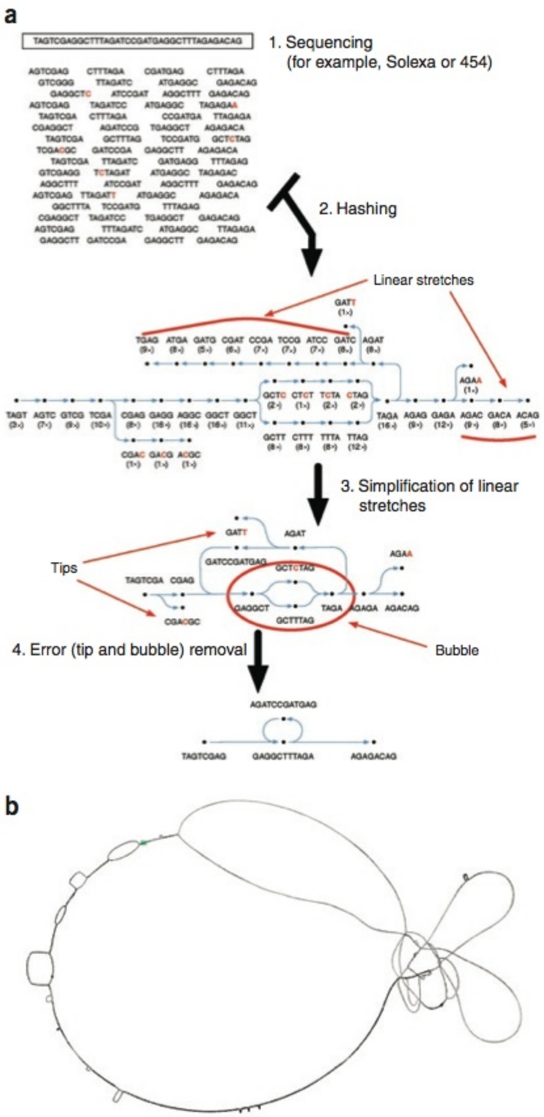
References

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly Nature Methods, 6(11 Suppl), S6–S12. doi:10.1038/nmeth.1376

Glenn, T. C. (2011). Field guide to next-generation DNA sequencers Molecular Ecology Resources. doi:10.1111/j.1755-0998.2011.03024.x

Hudson, M. E. (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology Molecular Ecology Resources, 8(1), 3–17. doi:10.1111/j.1471-8286.2007.02019.x

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing Briefings in bioinformatics, 11(5), 473–483. doi:10.1093/bib/bbq015



Flickey & Birney 2009