Second- (“Next”-) Generation Sequencing

Platforms and practical considerations

Based on a presentation by Björn Nystedt, SciLifeLab, Stockholm

Sequencing DNA

DNA fragments

↓

Size selection

↓

[PCR amplification] ← Single-molecule detection avoids this (Helicos, PacBiosciences, etc)

↓

Sequencing by synthesis (polymerase) or ligation (ligase)

↓

[Photodetection] ← Monochrome
1 color/nucleotide
1 color/di-nucleotide
(not in IonTorrent)

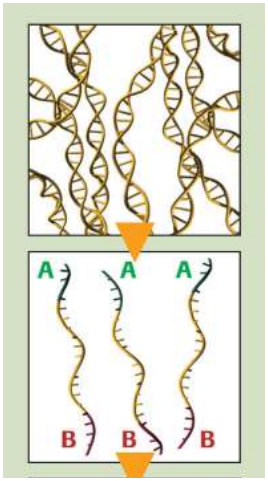
↓

Basecalling/colorspace

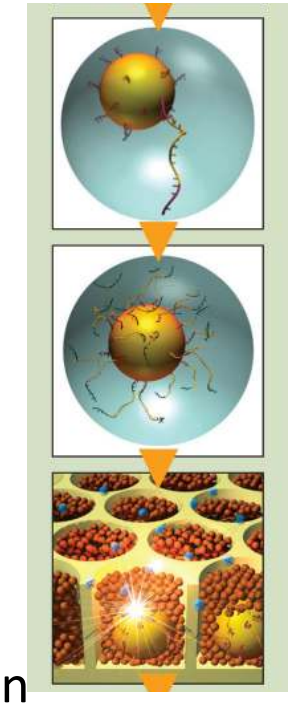
↓

Analysis

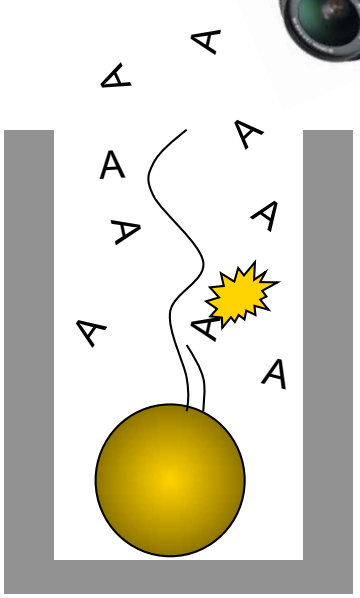
454



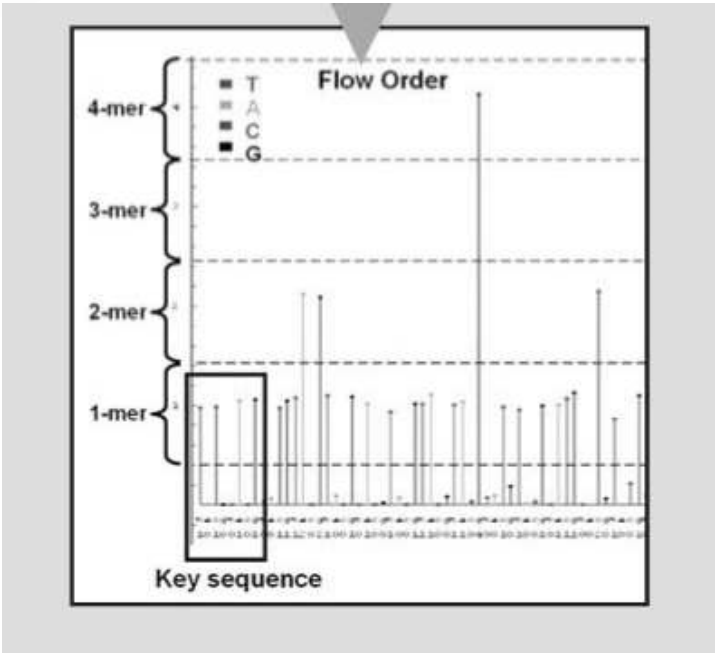
Fragmentation
(>500 bp)



Emulsion PCR
(1 bead/well)



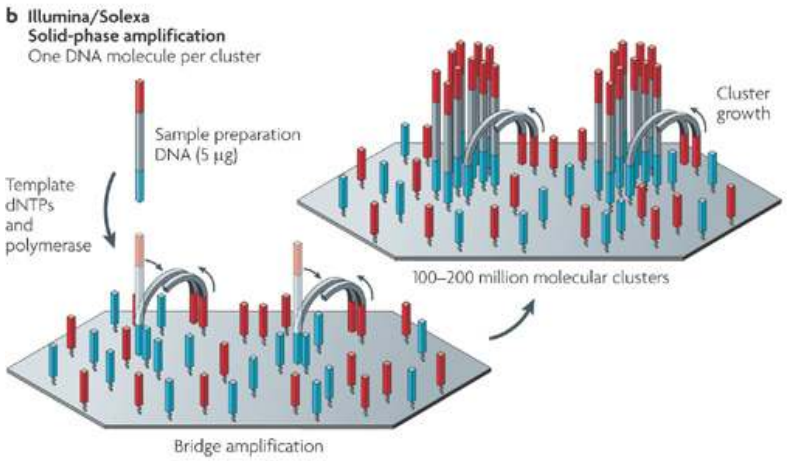
Seq by synthesis
1 nt-type at the time
(monochrome detection)



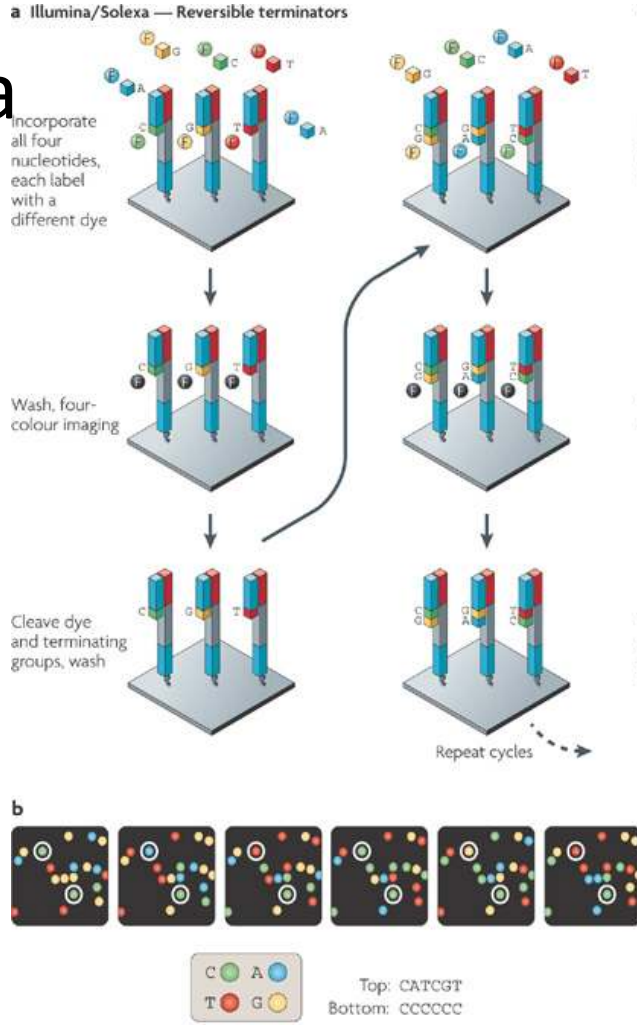
Light intensity proportional to
# of inserted nt:s
(= length of homopolymer)

Homopolymer
problems

Solexa/Illumina



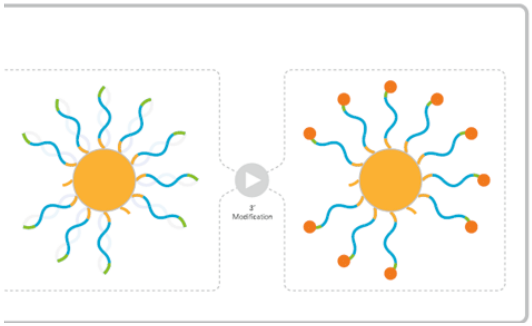
Solid phase “PCR”



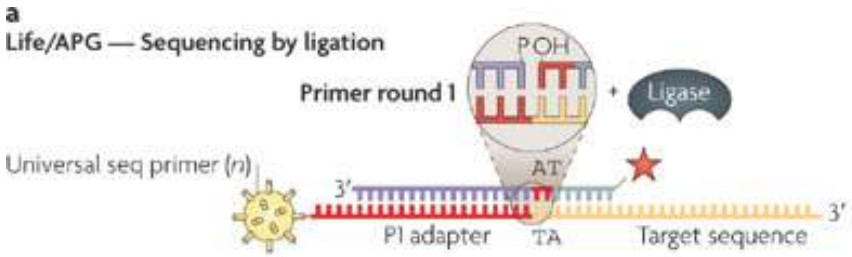
Possible to sequence both ends of one fragment

Seq by synthesis
4 nt-type at the time (4 colours)
Single nt insertions (reversible terminators)

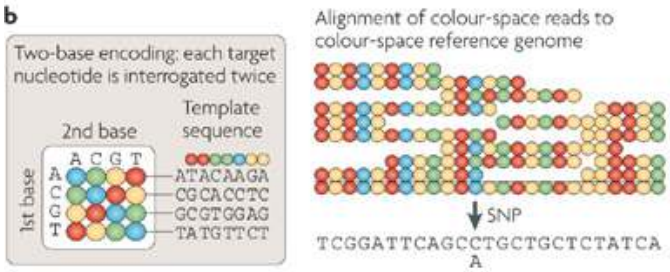
SOLiD



Emulsion PCR
Attach bead to surface



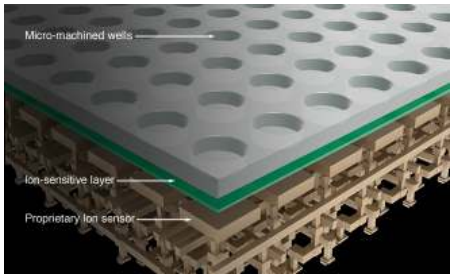
Seq by ligation
4 colours
Multiple rounds



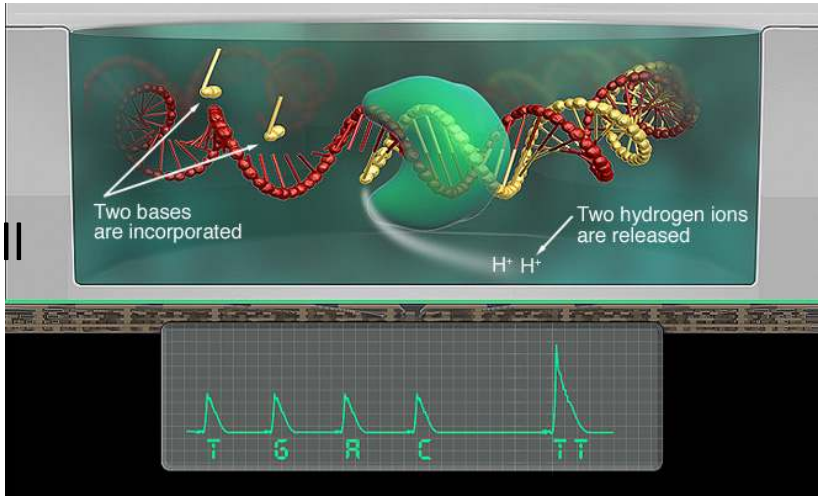
Non-unique colours (!)
Alignment in colourspace

Ion Torrent

"454 without the camera"



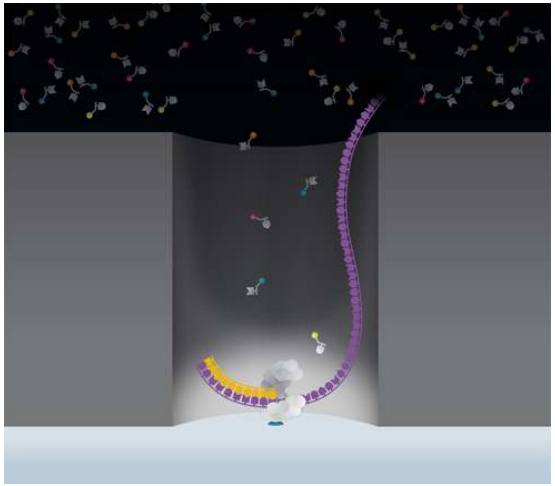
1 amplified template/well



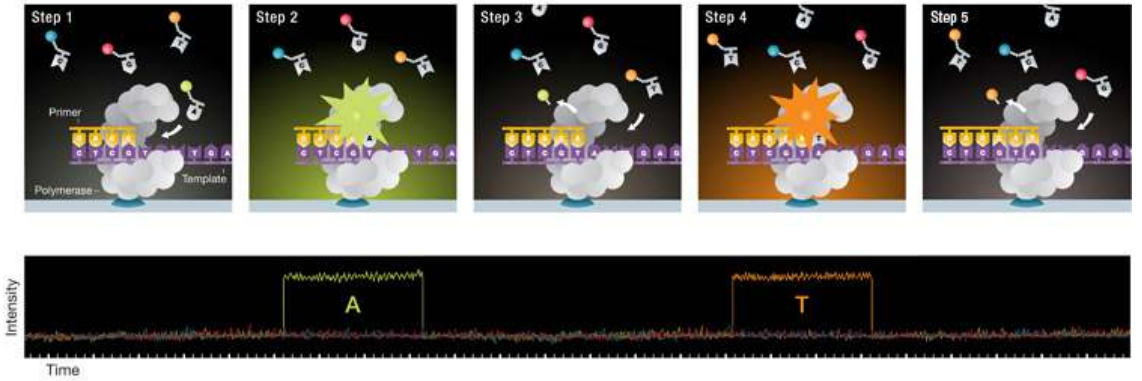
Seq by synthesis with normal nts and pol
1 nt-type at the time (monochrome detection)
Semiconductor technology
Detection by hydrogen ion release (change in pH)

No fluorescence or camera
Template DNA amplification
Homopolymer problems
No data available yet

Pacific Biosciences



Single molecule detection
(No amplification!)
Immobilized DNA
polymerase



Seq by synthesis
Detection in real-time (movie rather than photos)
All nt always available (4 colours)

Aiming for long reads (5-10kb, later 50kb)
“Stroboscope sequencing”

First results obtained with PacBio

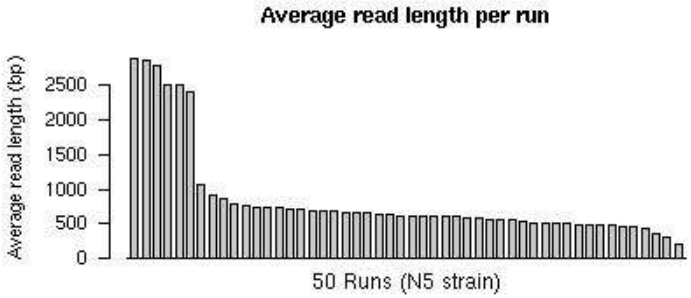
The NEW ENGLAND JOURNAL of MEDICINE

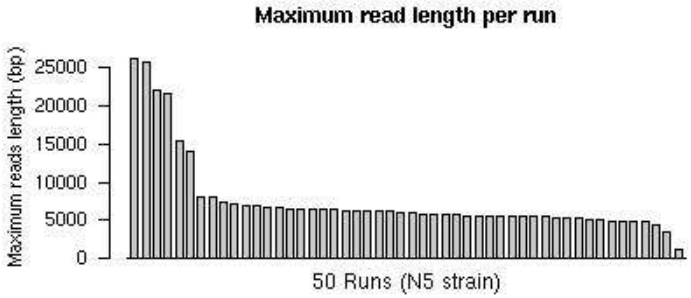
ORIGINAL ARTICLE

The Origin of the Haitian Cholera Outbreak Strain

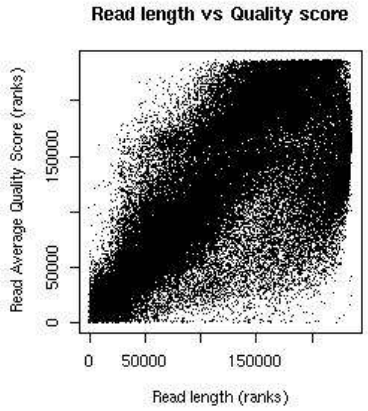
Chen-Shan Chin, Ph.D., Jon Sorenson, Ph.D., Jason B. Harris, M.D., William P. Robins, Ph.D., Richelle C. Charles, M.D., Roger R. Jean-Charles, M.D., James Bullard, Ph.D., Dale R. Webster, Ph.D., Andrew Kasarskis, Ph.D., Paul Peluso, Ph.D., Ellen E. Paxinos, Ph.D., Yoshiharu Yamaichi, Ph.D., Stephen B. Calderwood, M.D., John J. Mekalanos, Ph.D., Eric E. Schadt, Ph.D., and Matthew K. Waldor, M.D., Ph.D.

- Published in Chin et al, NEJM, 2010, Dec. 9. Sequenced 5 *V. cholerae* genomes, among which the one at the source of the cholera outbreak in Haiti





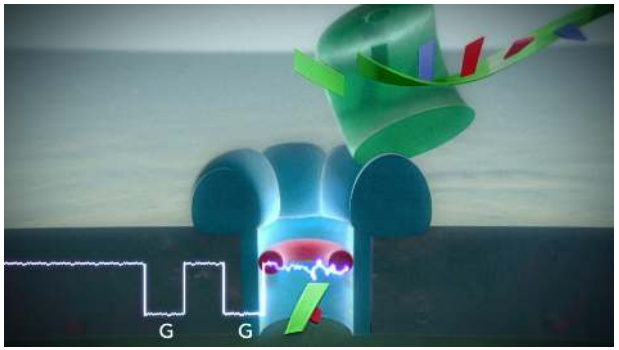
- First results analyzed in Olivier Elemento's blog: http://oelemento.wordpress.com/2011/01/03/a-closer-look-at-the-first-pacbio-sequence-dataset/

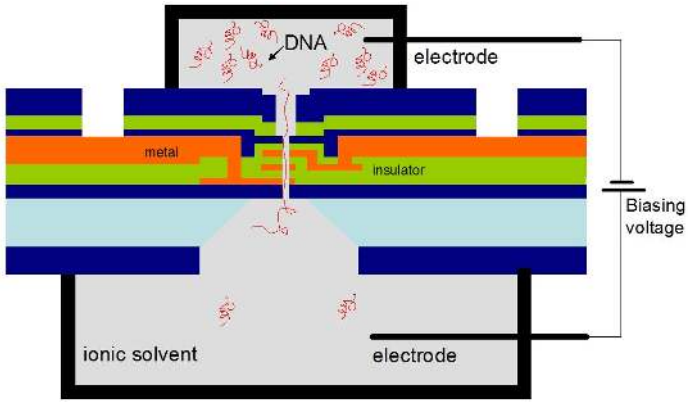


~25% of reads map
~20% error rate compared to ref

Upcoming (or not) technologies

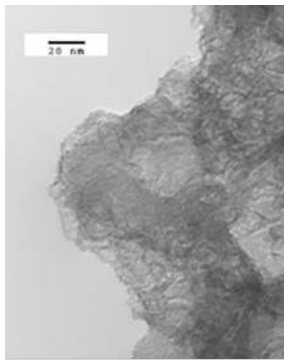
- **Oxford Nanopore**:
  - Restriction enzyme, cut one base at a time.
  - Detection by electrical field change through a nanopore





- **IBM nanopore**:
  - Similar, using voltage biases to control the speed of DNA flow through the pore

- **ZS Genetics**:
  - Using brominated or iodinated probes, visualizing directly DNA with electron microscopy



Platform summary

	Amplification	Sequencing	Read	Gbp/day	Cost $/Mb	Status
Sanger	cloning	synthesis	1 kb	0.006	~500	Standard
454	emPCR	synthesis (multi-nt)	450 bp	0.5	~20	Standard
Solexa	PCR on slide	synthesis (single-nt)	2 x 100 bp	25	~0.5	Standard
SOLiD	emPCR	ligation	2 x 50 bp	10	~0.5	Standard
Ion Torrent	emPCR	semiconductor	100 bp	0.1?	~5?	Market 2011
PacBio	none	synthesis (real-time)	>10 kb	>100?	NA	First results

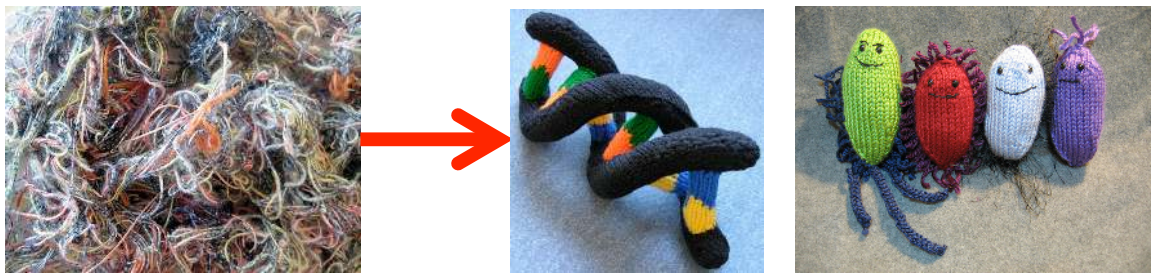
Caution: numbers change every month…

Things to consider…

What is the cost-per base?
What is the cost for tagging / library preparation?
What coverage is needed?
What read length is required?
What kind/frequency of sequencing errors can be tolerated?
What coverage bias can be expected?
Should paired-end reads be used?
What bioinformatics is needed?

What artefacts can be expected in the bioinformatic part?
Can the experiment be designed to increase the signal-to-noise ratio?
Will the final quality/signal be enough to gain new knowledge?

"COST PER INSIGHT"



Comparative cost-per-base, read lengths and error rates

Cost per base:

low PacBio ~ SOLiD ~ Illumina < IonTorrent < 454 < Sanger **high**

Read lengths:

long PacBio >> Sanger > 454 > IonTorrent ~ Illumina ~ SOLiD **short**

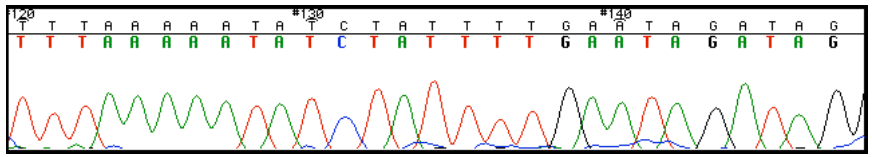
Error rates:

low Sanger ~ SOLiD < Illumina ~ 454* < PacBio **high**

Coverage bias:

low Illumina ~ 454 < SOLiD < Sanger **high**

* non-random errors!



Single-molecule sequencers have higher error rates, but (potentially) lower coverage bias

Sequencing errors

454

Homopolymer errors => artificial indels

Problem remains at high coverage

Hard/impossible for gene-finding software

True biological frameshifts are common at homopolymers (!)

Solexa

Error rates increase rapidly towards read ends

(*100 nt reads barely better than 75 nt reads*)

SOLiD

High accuracy in colorspace mapping

Non-unique color detection causes frameshift problems de-novo assembly

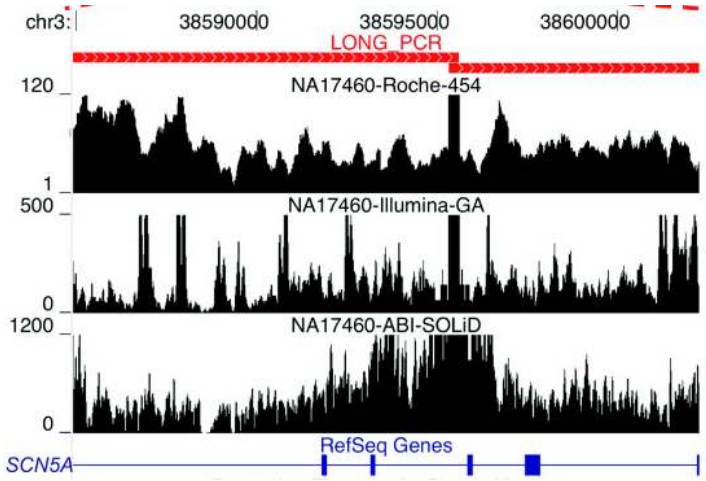
Coverage and coverage bias

Coverage

Depend very much on application/question

De novo: >30X (100X may be too much)

Mapping: 5X sometimes enough, sometimes 100X



Harismendy *et al. Genome Biology* 2009, **10**:R32

Coverage bias

Extremely important for quantitative gene expression (eg RNA-seq)

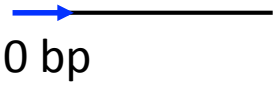
Uneven coverage can be due to biases at several levels

- DNA extraction
- DNA fragment amplification (avoided in single-molecule sequencing!)
- Sequencing

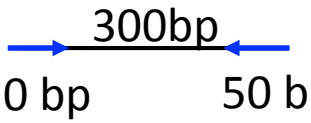
Solexa/454 has less coverage bias than SOLiD

Paired-end reads

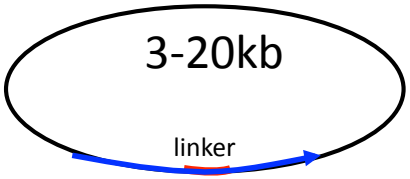
454 **Illumina/SOLiD**



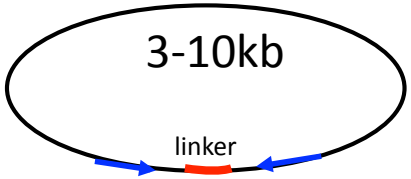
Single end (SE)



Paired-end (PE), short fragment ends



150 + 150 bp
(flip left)



50 + 50 bp
(flip both)

Mate-pair (MP), another type of paired-end, circularization (cloning)

De novo

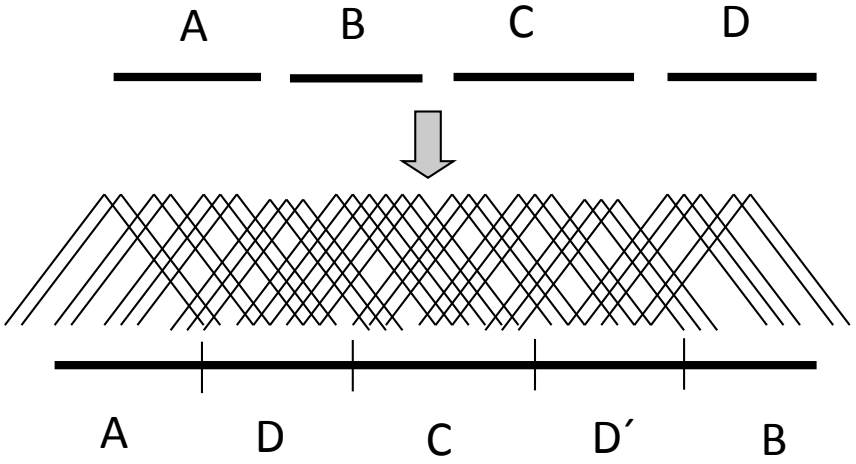
Resolve repeats shorter than the insert

Create scaffolds (contig ordering)

Mapping

Better accuracy

Distinguish repeats



Long vs short insert paired-ends

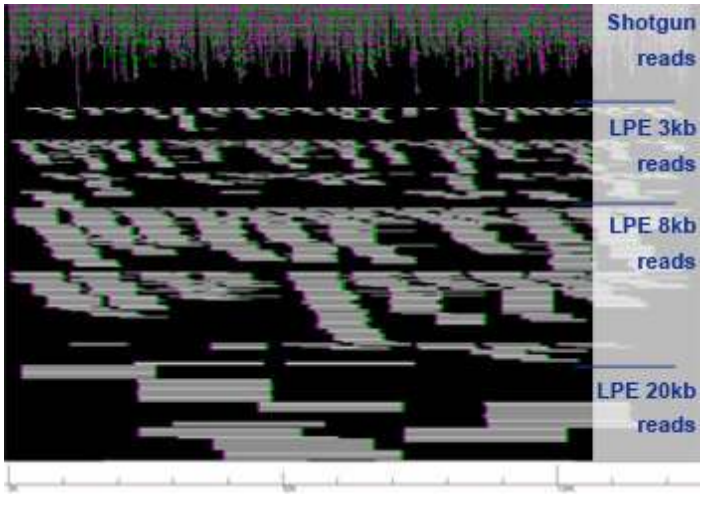
Long inserts: 1-30 kb (high insert length variation)
Good for scaffolding
Detect structural variations (mapping)
Good for metagenomics (better taxonomic resolution)

Short inserts: 100-500 bp (low insert length variation)
Good for short repeat resolving
Enhanced mapping

Illumina 2x100 300bp insert paired-ends ~ 454 450bp reads for de novo sequencing

Length variation: 3-4 kb << 10-20kb

Current de novo sequencing strategies: successive libraries with increasing insert sizes



Bioinformatics

Storage and analysis: increasingly costly compared with sequencing itself

Mapping software are getting pretty good, but might miss crucial differences (repeats…)

De novo assembly still hard to evaluate

Basic programming skills help a lot (Perl, Python, bash…)

Artefacts still very common due to:

- poor quality/short reads
- repeats and indels
- poor assembly of reference genome
- inhomogen samples/complicated biological differences

What sequencing strategy should I take?

To be taken with a pinch of salt…

De novo assembly
454 30X coverage
+ Illumina paired-end (300bp) and (3kb-10 kb) 30X coverage

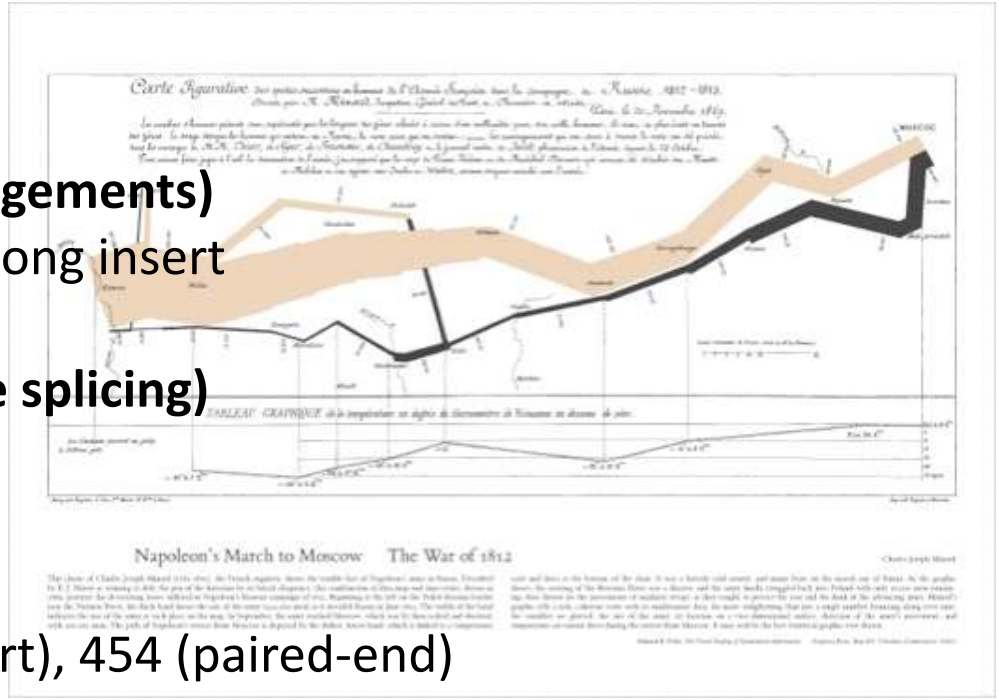
Quantitative transcriptome sequencing
Illumina (or 454). Perhaps 300pb paired-end

SNP detection
Illumina/SOLiD (300bp paired-end?)

Structural variations (indels/rearrangements)
Illumina/SOLiD paired-end, Illumina long insert

Low-expression variants (alternative splicing)
SOLiD

Metagenomics
Illumina (2x100bp, short or long insert), 454 (paired-end)



The future is bright

What can you do with a 10 Tbp run?

The complete 1 Gbp genome for 10 tissue types from 10 individuals from each of your 10 favourite species

…or the complete genomes of 100'000 individual bacteria/archaea cells (i.e. ~100 times as much as all complete bacterial/archaeal genomes in GenBank).

...or the complete genomes of all participants of the workshop and their families

Now, what would you do with that?

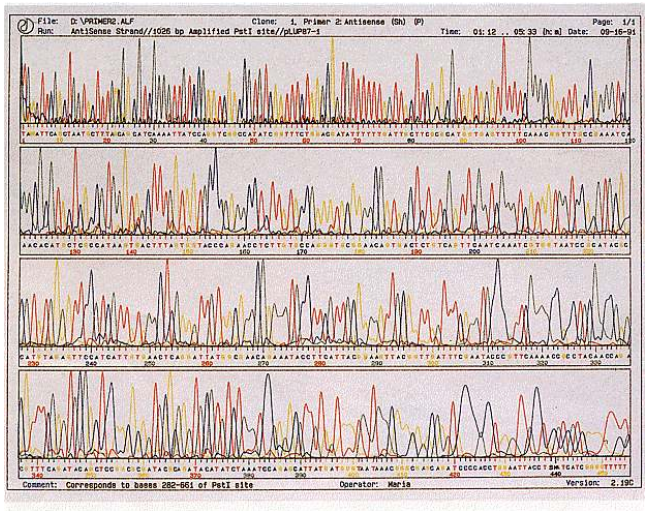
Revolutions to come very soon

Personal diagnostics

Metagenomics

De novo genome assembly

Complete transcriptomics



Bottlenecks

- Whole genome comparisons
- All-against-all comparisons (combinatorial freak-out!)
- Visualization tools
- Standard input/output file formats
- Software benchmarking and post-assembly verifications

Relax…

Keeping up with software/technological development can be stressful.
Things to consider to reduce the stress:

- Choose quality before quantity
- Know your biological system
- Design your studies carefully
- Do small-scale sanity checks at every step
- Combine sequencing with other data
- Share knowledge and talk to your bioinformatics fellows

Thanks