CIBIO
Centre for Integrative Biology

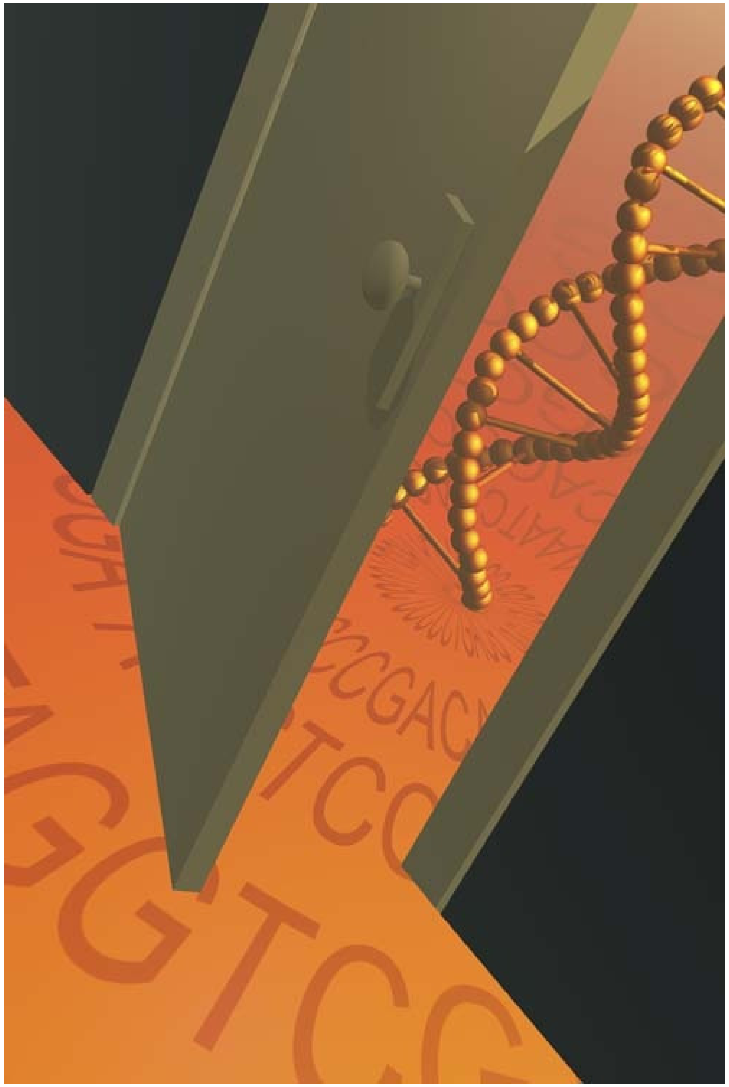
UNIVERSITÀ DEGLI STUDI DI TRENTO

“Next generation sequencing techniques"

Toma Tebaldi
Centre for Integrative Biology
University of Trento

Mattarello
September 28, 2009

Sequencing

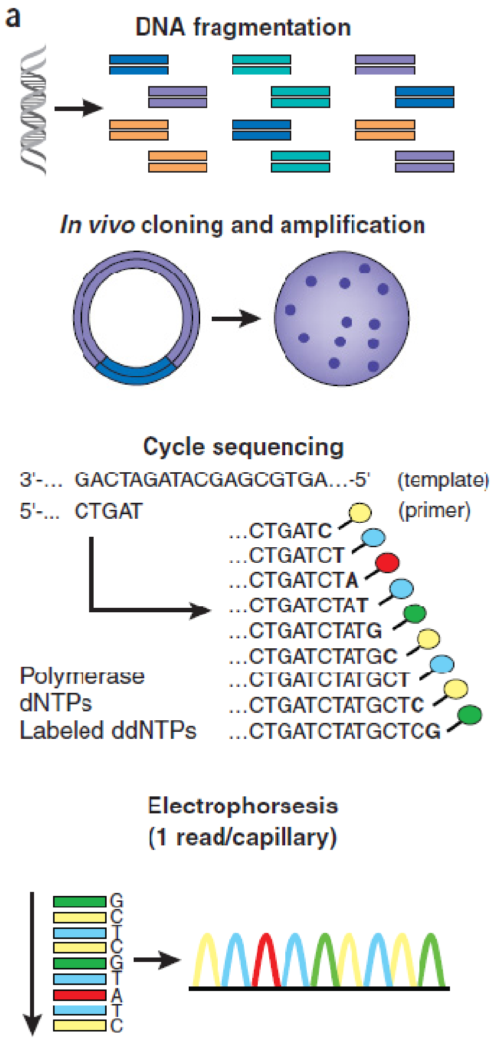


Fundamental task in modern biology

✓ read the information content of biological molecules (DNA, RNA).

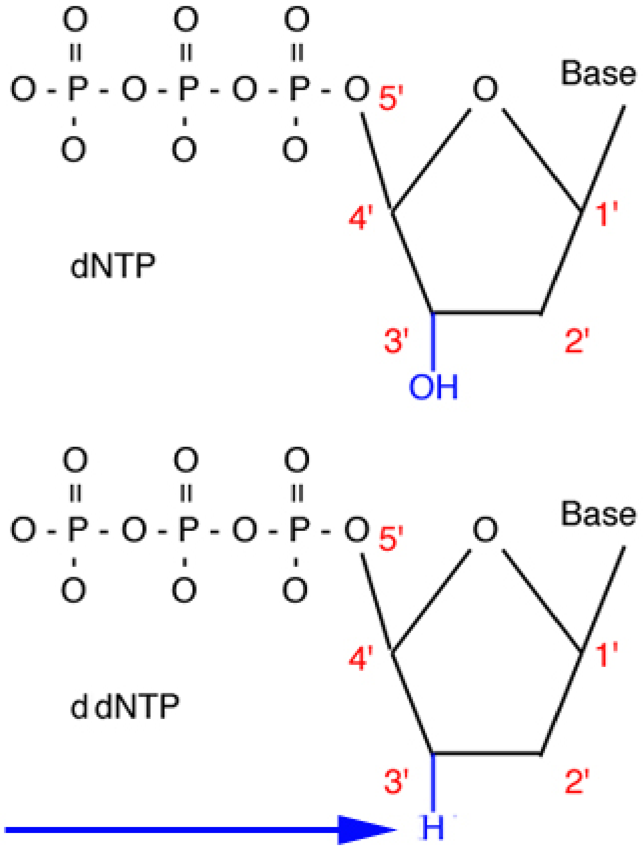
✓ direct and primary access to understand how biological systems function and evolve in time.

First generation sequencing: Sanger



- ✓ DNA is fragmented
- ✓ Cloned to a plasmid vector
- ✓ Cyclic sequencing reaction
- ✓ Separation by electrophoresis
- ✓ Readout with fluorescent tags

Dideoxy nucleoside triphosphates (ddNTPs)



✓ Elongation with a mixture of dNTPs and ddNTPs.

✓ lack an -OH on the 3'-C as well as the 2'-C of the deoxyribose sugar.

✓Each ddNTP is labeled with a different fluorescent dye.

✓Once the ddNTP is incorporated, chain elongation is terminated.

Glossary

✓ **Sequencing depth:** total number of all the sequence reads or base pairs represented in a single sequencing experiment.

✓ **Coverage Depth:** The total number of nucleotides from reads that are mapped to a given position (e.g. 10x).

✓ **Read Length:** length of the sequenced fragments (tags).

✓**Number of sequencing reads**: number of reads (sequence tags) produced in a single experiment.

Second (next) generation sequencing

✓*Greater sequencing throughput*

✓*More economical sequencing technology*

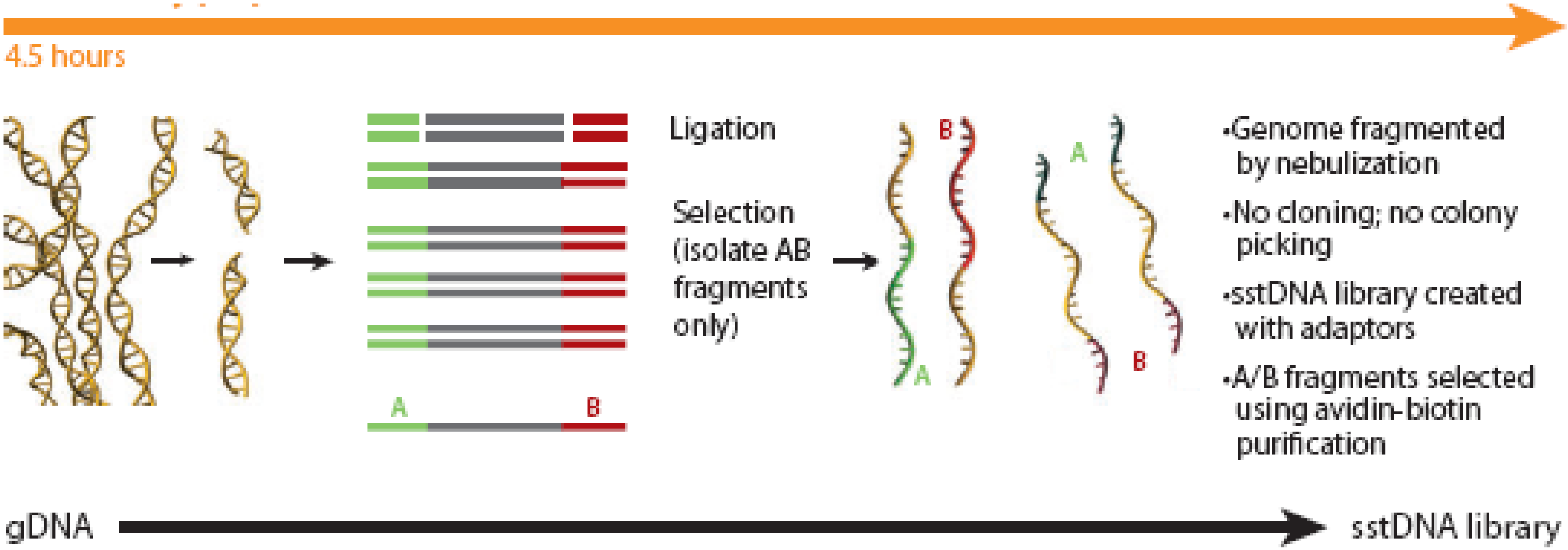
Three leading platforms

✓Roche/454 FLX Pyrosequencer

✓ Illumina/Solexa Genome Analyzer

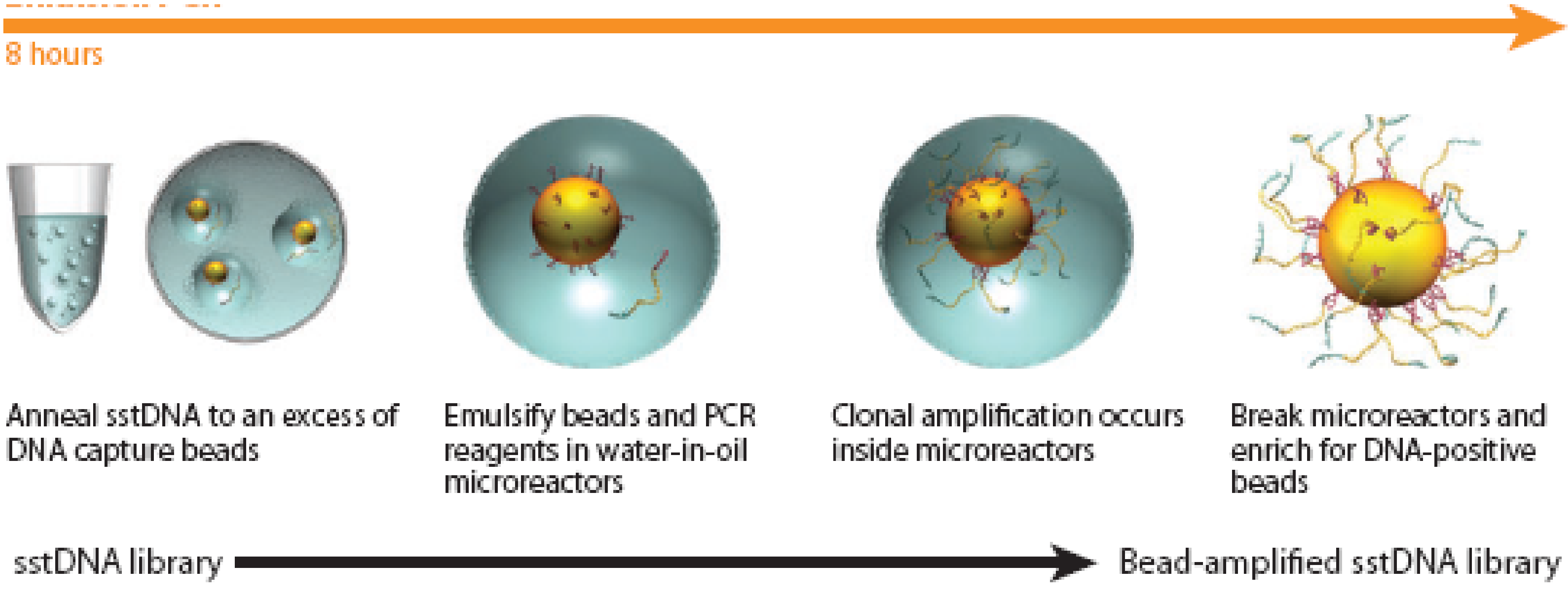
✓Applied Biosystems SOLiD

454 sequencer: DNA library preparation



- Genome fragmented by nebulization
- No cloning; no colony picking
- sstDNA library created with adaptors
- A/B fragments selected using avidin-biotin purification

454 sequencer: Emulsion PCR



8 hours

Anneal sstDNA to an excess of DNA capture beads

Emulsify beads and PCR reagents in water-in-oil microreactors

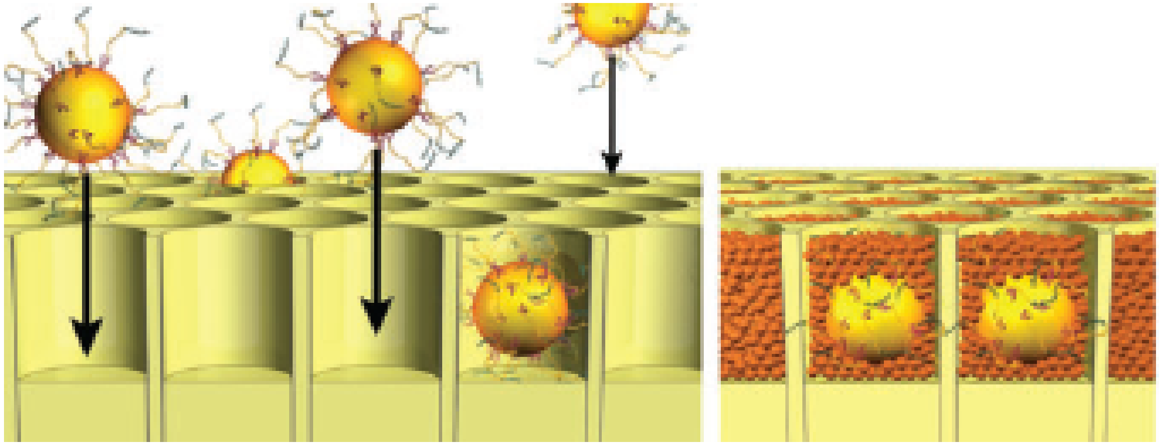
Clonal amplification occurs inside microreactors

Break microreactors and enrich for DNA-positive beads

sstDNA library → Bead-amplified sstDNA library

454 sequencer: Sequencing

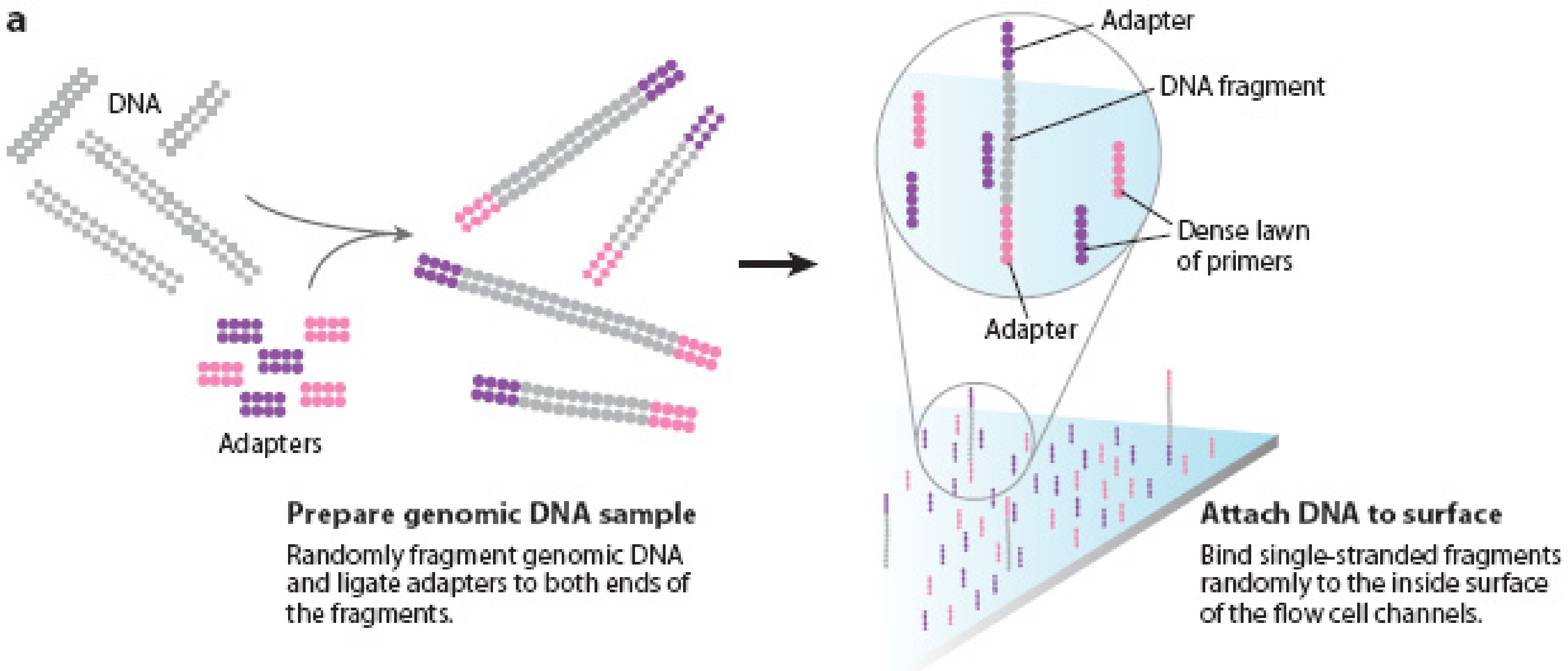
7.5 hours



- •Well diameter: average of 44 µm
- •400,000 reads obtained in parallel
- •A single cloned amplified sstDNA bead is deposited per well

Amplified sstDNA library beads → Quality filtered bases

Illumina: Library Preparation



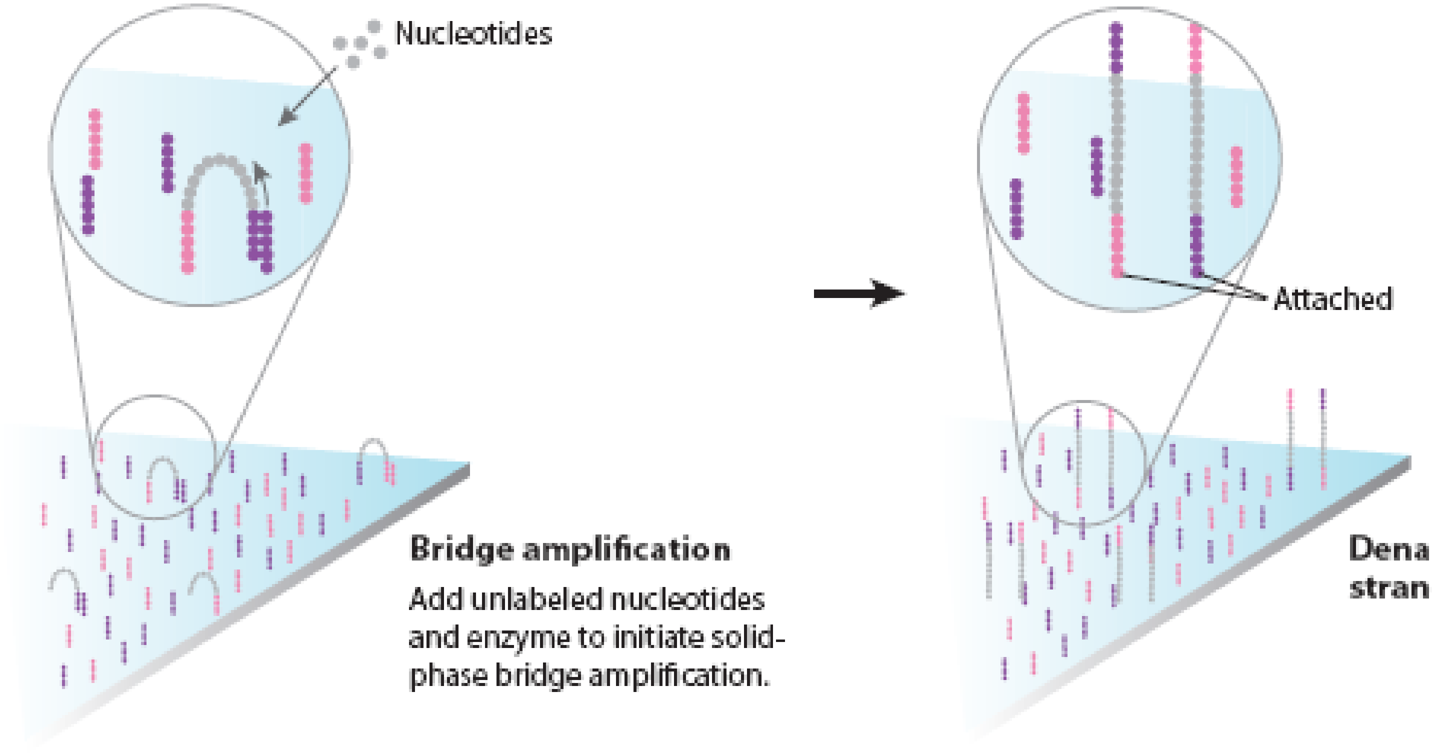
Prepare genomic DNA sample

Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Illumina: Bridge PCR

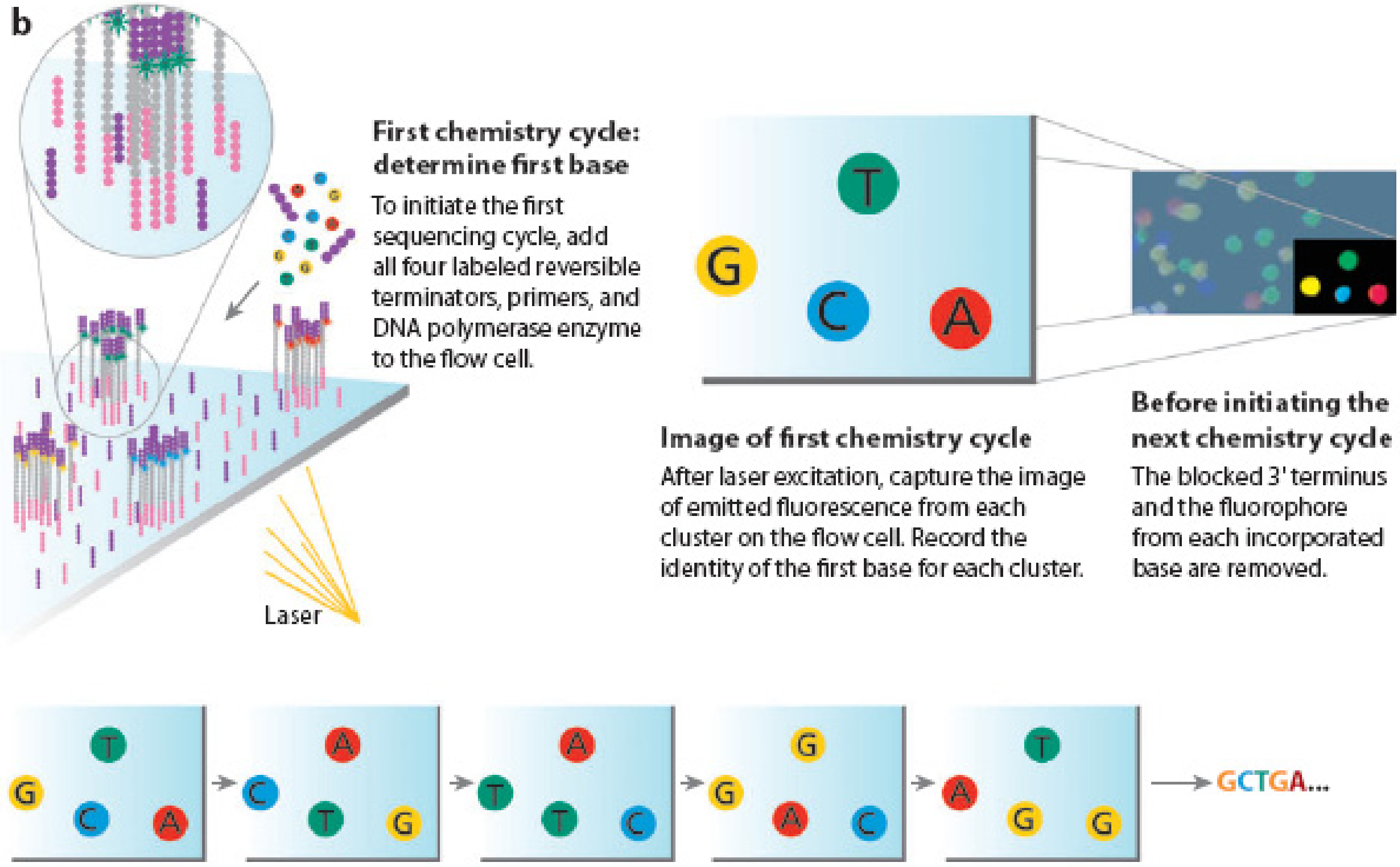


Bridge amplification

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

Illumina: Sequencing by Synthesis



First chemistry cycle: determine first base

To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle

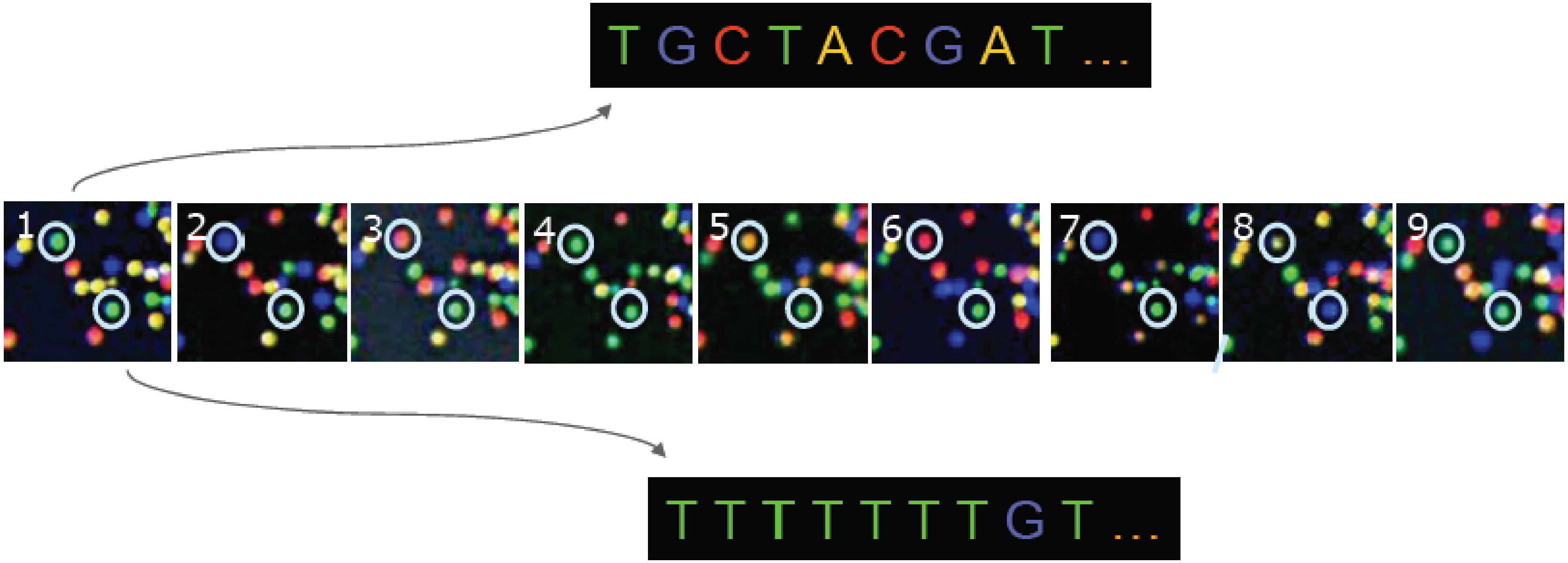
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

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: Base Calling

Base calling from raw data

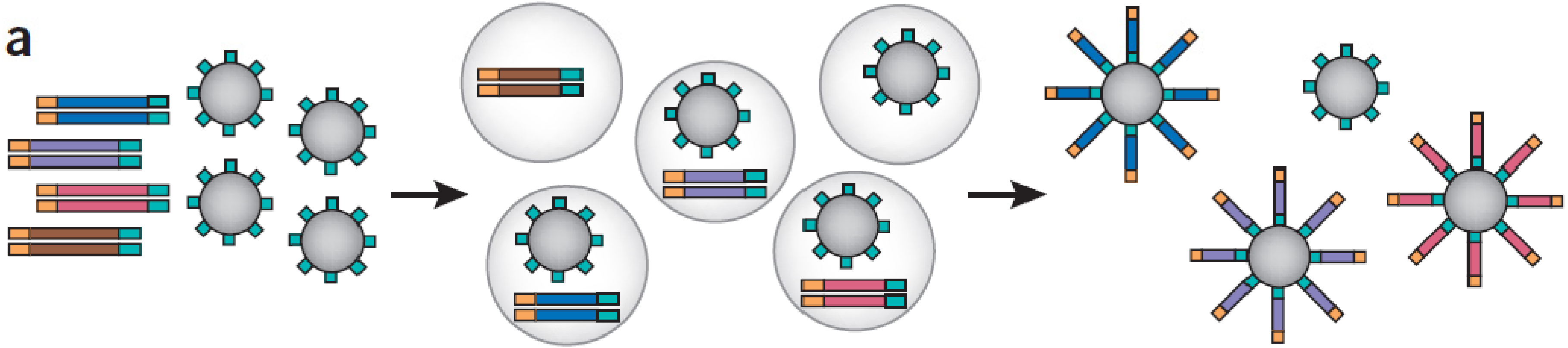


TGCTACGAT...

TTTTTTTGT...

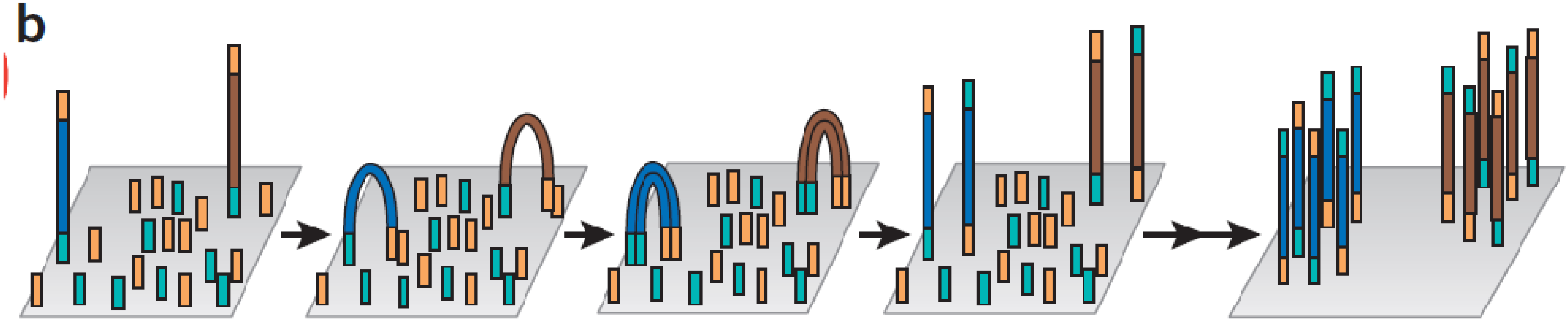
The identity of each base of a cluster is read off from sequential images

Emulsion PCR



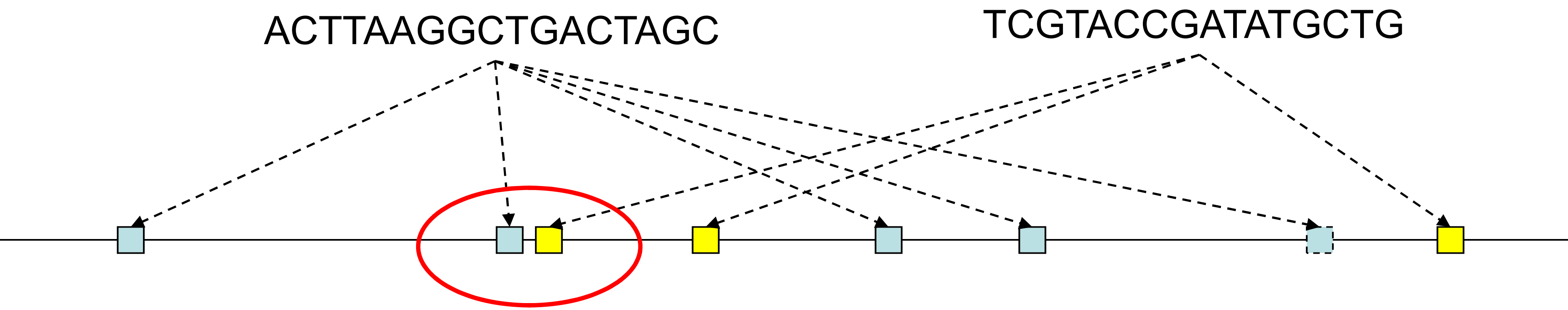
- ✓ Fragments, with adaptors, are PCR amplified within a water drop in oil.
- ✓ One primer is attached to the surface of a bead.
- ✓ Used by 454, Polonator and SOLiD.

Bridge PCR



- ✓ DNA fragments are flanked with adaptors.
- ✓ A flat surface coated with two types of primers, corresponding to the adaptors.
- ✓ Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- ✓ Used by Solexa.

Problems arising with short sequence reads

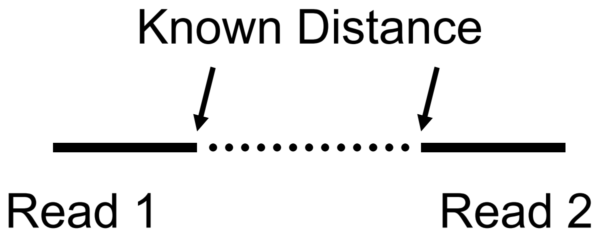


Short sequences do not map uniquely to the genome:

✓ Solution 1: Get longer reads.

✓ Solution 2: Get paired reads

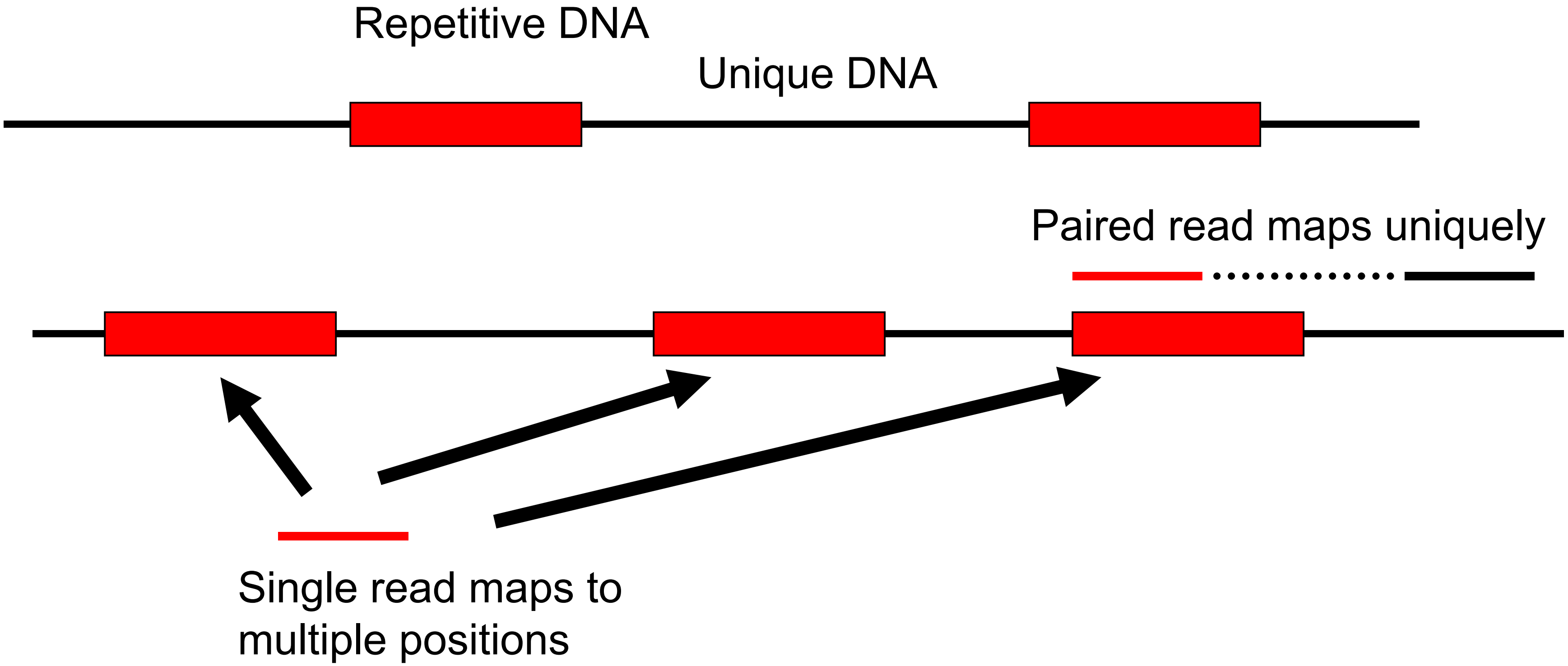
Paired reads are important for mapping



Known Distance

Read 1

Read 2



Repetitive DNA

Unique DNA

Paired read maps uniquely

Single read maps to multiple positions

Platforms comparison

With 3730s, ~60Mb per year

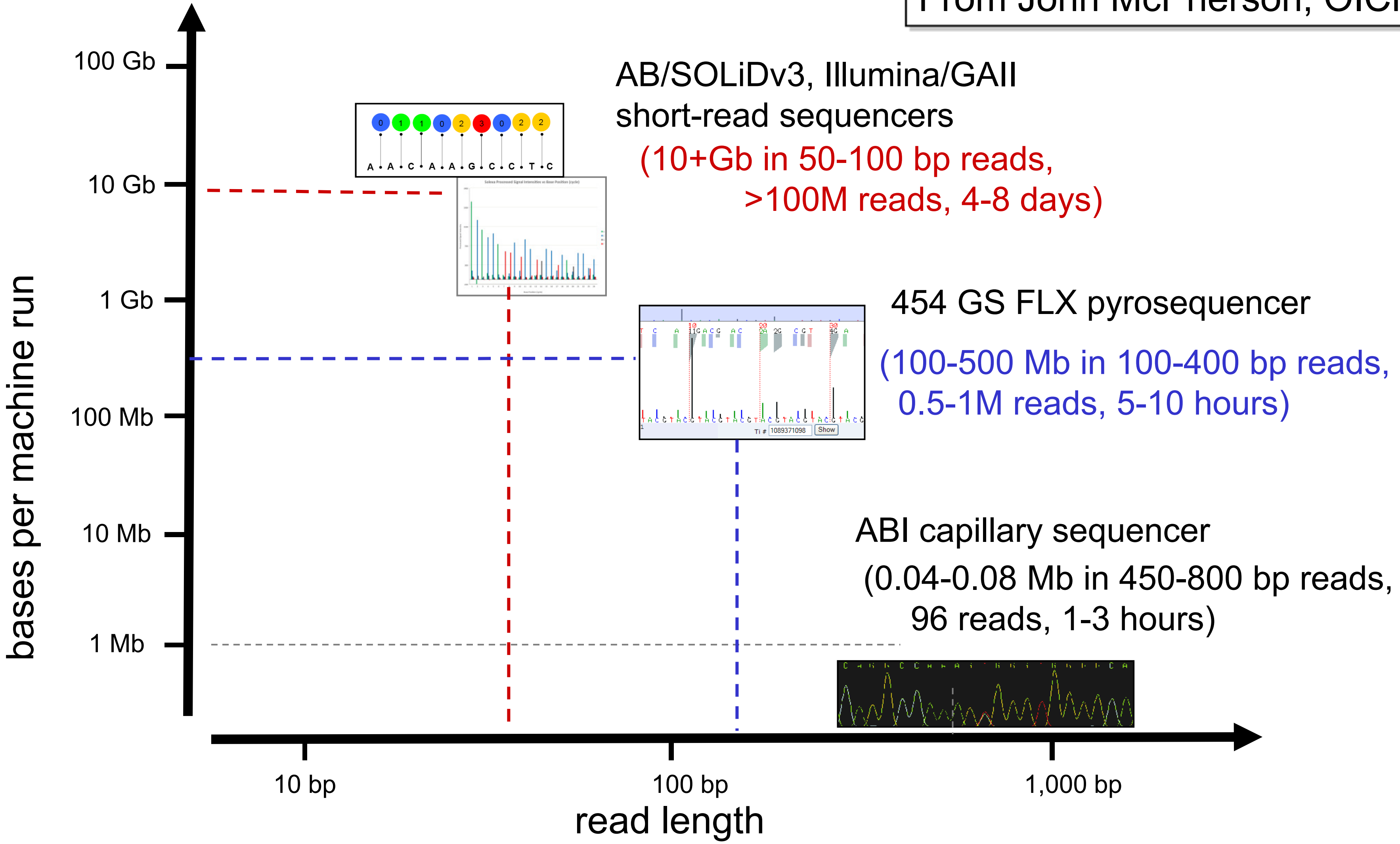
Specifications as of summer 2008

	454	Solexa	SOLiD
Bp per run	400 Mb	2-3 Gb	3-6 Gb
Read length	250-400 bp	35-50 (70-100) bp	35-50 bp
run time	10 hr	2.5 days	5 days
Download	20 min	27 hr (44 min)	~1 day
Analysis	2-5 hr	2 days	2-3 days
Files	20-50 Gb	1T	1 T

- Next-gen sequencing technologies have reduced the cost of sequencing by > 4 orders of magnitude already

Comparisons between methods

From John McPherson, OICR



AB/SOLiDv3, Illumina/GAII
short-read sequencers
(10+Gb in 50-100 bp reads,
>100M reads, 4-8 days)

454 GS FLX pyrosequencer
(100-500 Mb in 100-400 bp reads,
0.5-1M reads, 5-10 hours)

ABI capillary sequencer
(0.04-0.08 Mb in 450-800 bp reads,
96 reads, 1-3 hours)

Computational tasks

✓ Hard to generate clean data: files with quality scores.

✓ Dealing with sequencing errors.

✓ Interpretation of data: need to correctly align sequence tags to a reference genome.

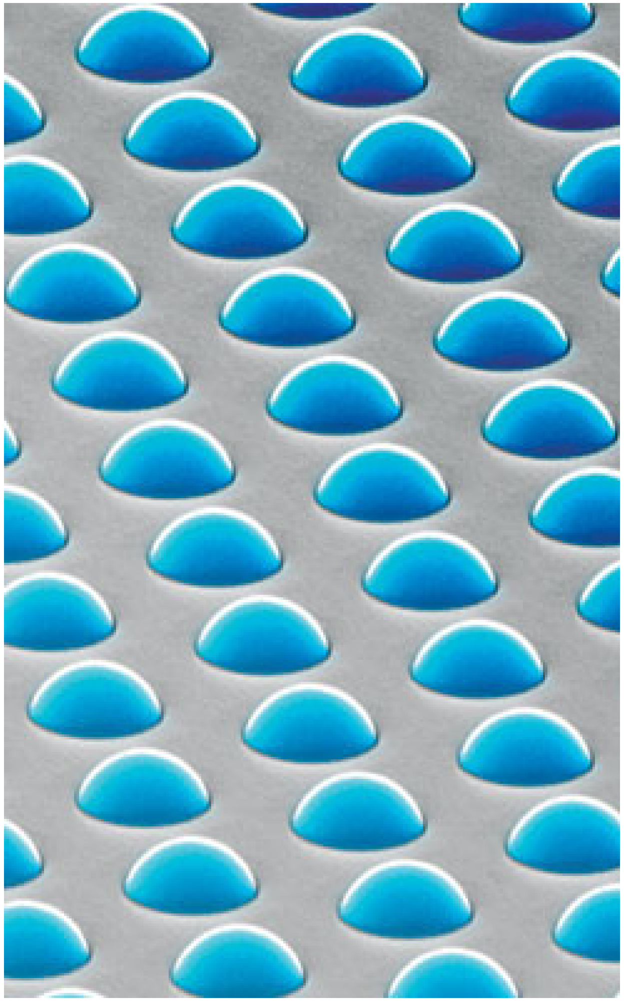
✓ The size of the data will costantly increase.

✓**Analytical bottleneck.**

Applications of Next Generation Sequencing

- Whole-genome sequencing
  - *de novo* genome assembly (much harder with shorter reads)
  - Variant detection (SNPs, indels) and copy number
  - 1000 Genomes Project
- Targeted resequencing (e.g.,exons) using ‘capture and release’ in combination with Agilent or Nimblegen microarrays
- ChIP-seq
  - Protein-DNA binding, histone modifications, nucleosomes
- Expression profiling:
  - RNA-seq – splicing variants
  - Digital expression profiling (DSAGE) – low abundance transcripts
- Small RNA sequencing

Transcriptome profiling: microarray methods



Hybridization-based approaches limitations:

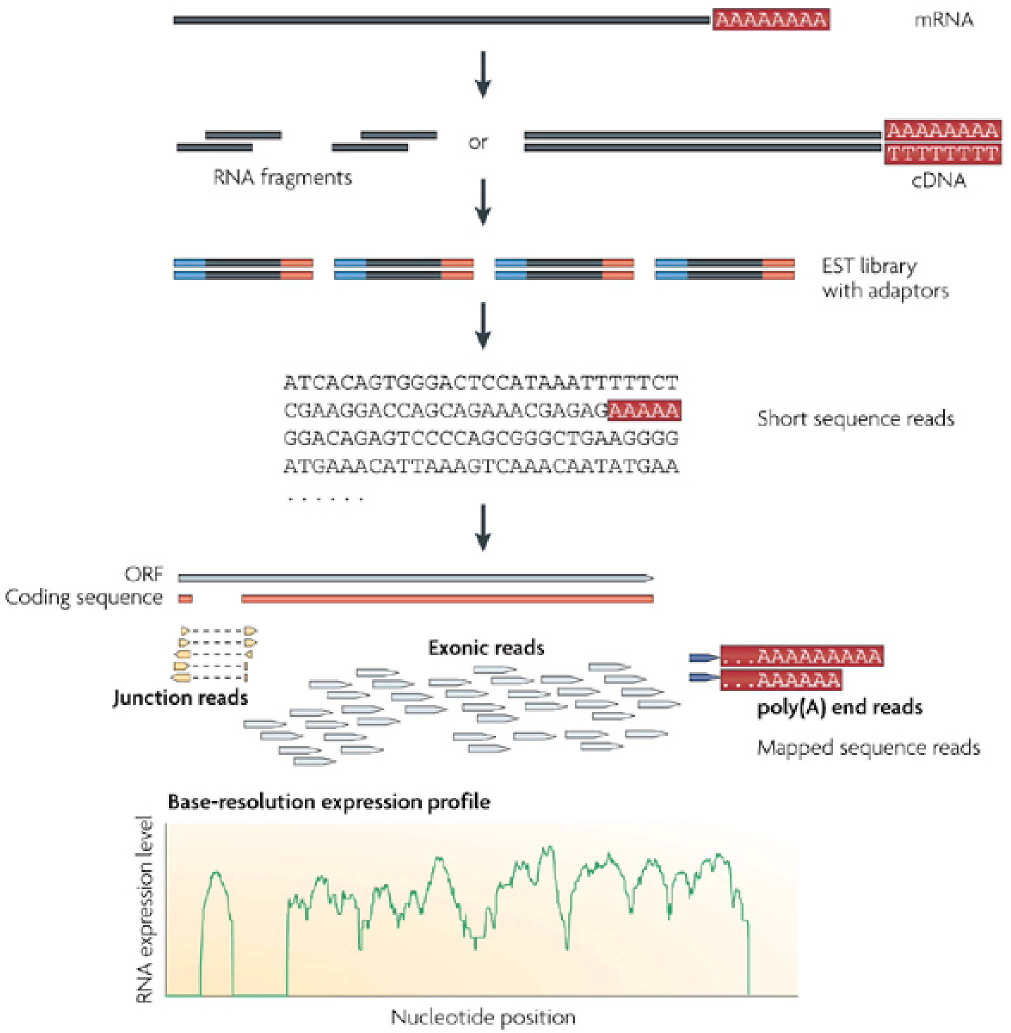
✓ rely upon existing knowledge about genome sequence.

✓ high background levels owing to cross-hybridization.

✓ limited dynamic range of detection due to signal saturation.

✓ normalization methods to compare different experiments.

Transcriptome profiling: sequencing methods



Nature Reviews | Genetics

Serial analysis of gene expression (SAGE): used to produce a snapshot of the messenger RNA population in a sample of interest (CAGE: cap analysis of gene expression).

- ✓ Based on Sanger sequencing

RNA-seq: based on next generation sequencing technologies.

Third (next-next) generation sequencing

Single molecule sequencing

✓Helicos Heliscope

✓ Pacific Biosciences SMRT

✓Nanopore BASE DNA sequencing