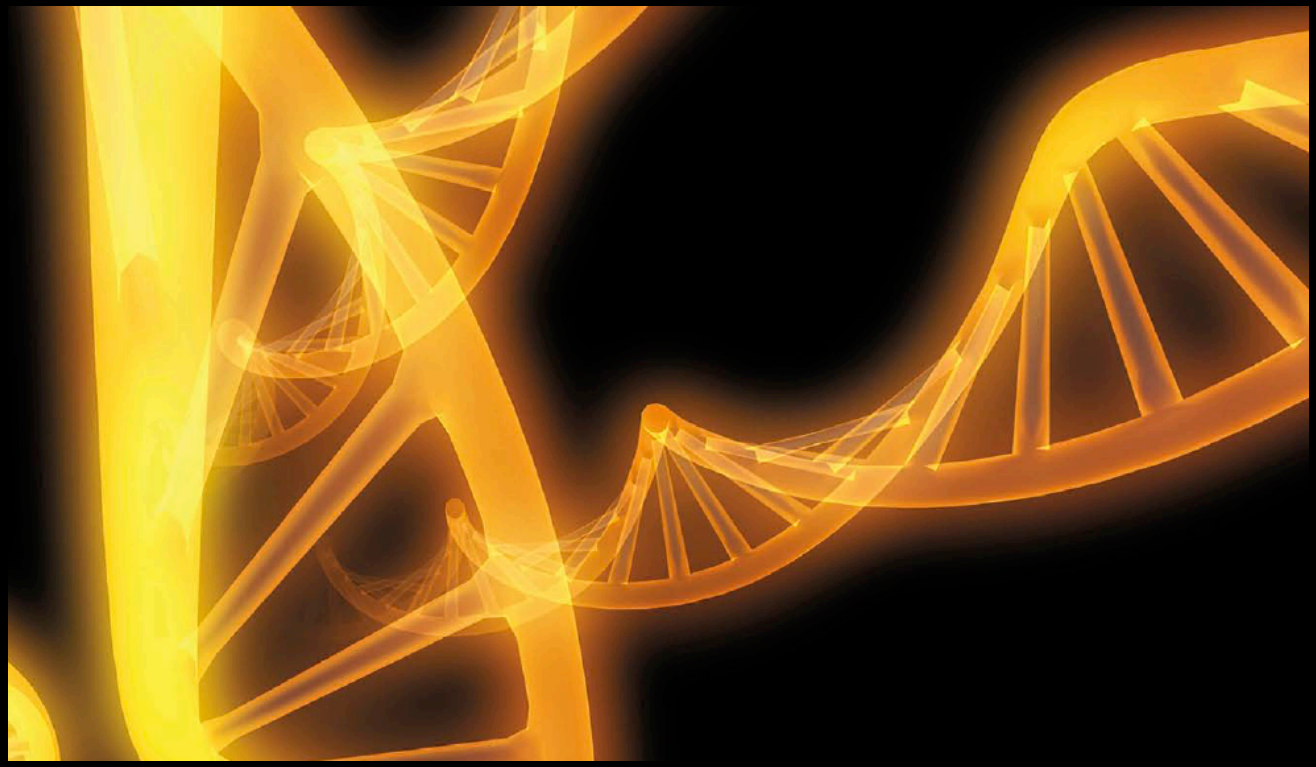
Next Generation Sequencing:

Technologies and Applications



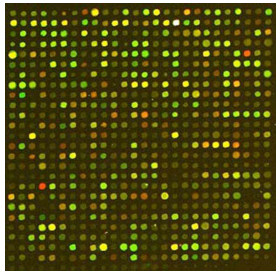
Jim Noonan
Department of Genetics
james.noonan@yale.edu
www.yale.edu/noonanlab

Sequence as the readout for biological processes

Determining the biological state of cells, tissues and organisms requires the quantification of sequence information

- Gene expression
- Protein-DNA interactions (ChIP)
- DNA-DNA interactions (3C/4C/5C)
- Chromatin state
- DNA methylation
- Genetic variation (SNPs/CNVs)

Indirect measures



microarrays, PCR, etc.

Direct measures

CTATGATCAGTC...
TCAATCTGATCTG...
GGACTTCGAGATC...
AAGTCGCTGACGT...

Sequencing

Outline

First-generation sequencing technology

- Sanger sequencing
- Parallelization in human genome project

Current massively parallel sequencing platforms

- 454
- Illumina
- SOLiD
- Helicos

Applications, advantages, issues

Third-generation sequencing

- Pacific Biosciences
- Nanopore sequencing

Metrics for evaluating sequencing methods

Throughput

- Number of high quality bases per unit time
- Number of independent samples run in parallel - multiplexing
- Difficulty of sample prep

Yield

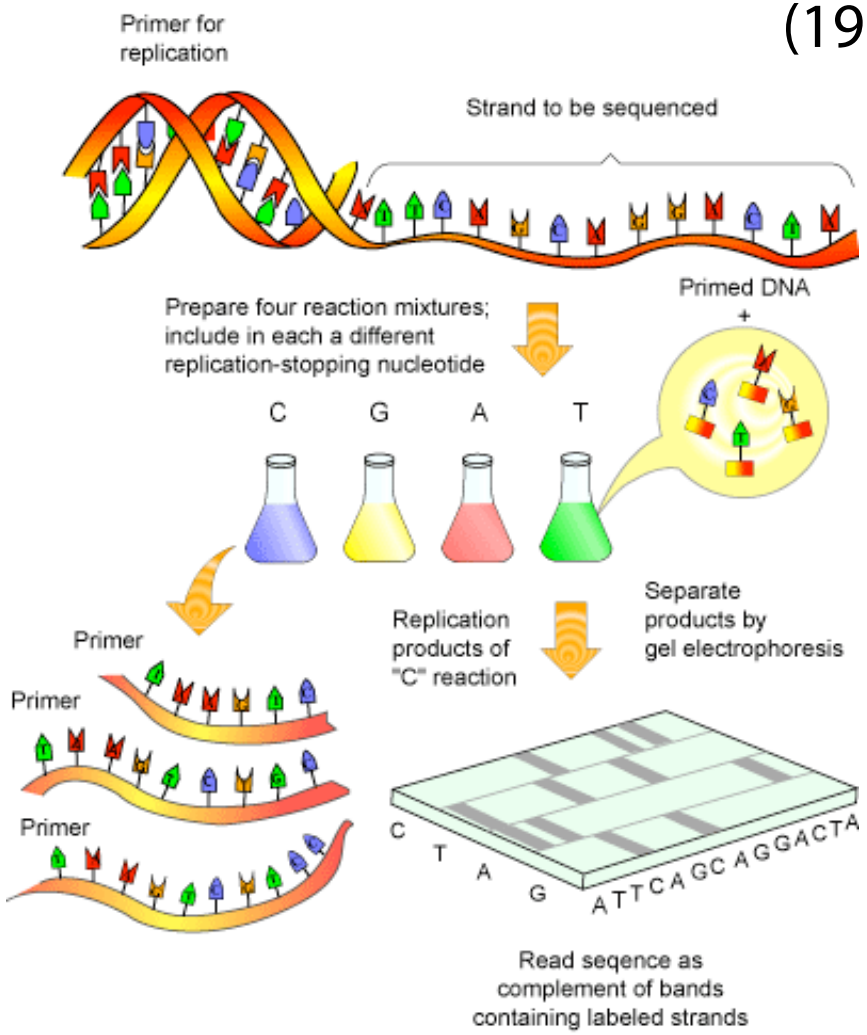
- Number of useful/mappable reads per sample
- Read length

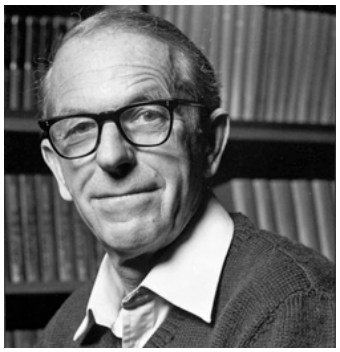
Cost

- Per run and per base
- Equipment
- Reagents
- Infrastructure
- Labor
- Analysis

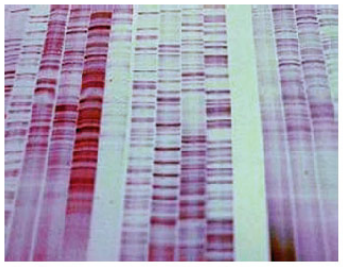
The goal of all new sequencing technologies is to increase throughput and yield while reducing cost

Sanger sequencing (1975-1977)

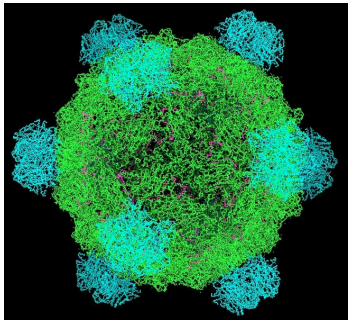




1980 Nobel Prize in chemistry



gels read by hand

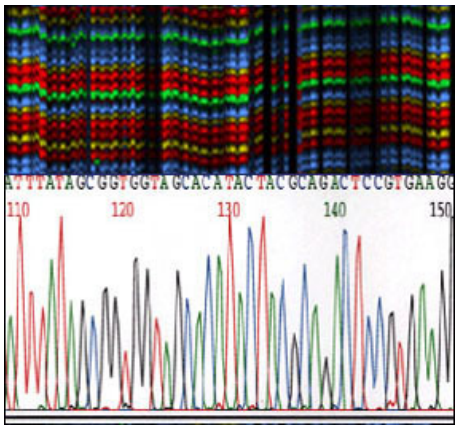


phi X 174
~5300 bp

- radiolabeled dideoxyNTPs
- one lane per nucleotide
- 800 bp reads
- low throughput (several kb/gel)

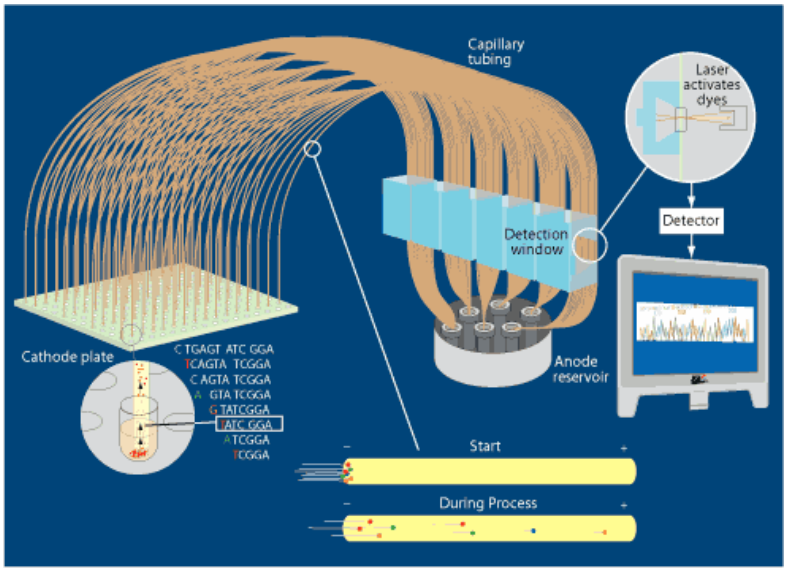
Parallelization of Sanger sequencing: Technology

Semi-automated gel electrophoresis



- four color ddNTP labeling
- 800 bp reads
- 96 samples/gel
- 70,000 bp/gel
- automated readout
- metrics for basecalling and quality scoring

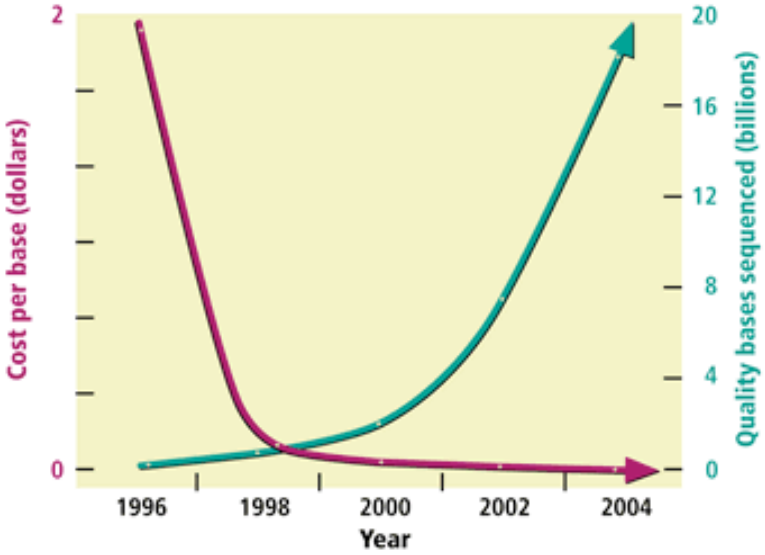
Capillary electrophoresis



- 800-1000 bp reads
- 384 samples/cap
- 300,000 bp/cap

Parallelization of Sanger sequencing: Infrastructure

•enormous increase in sequencing production capacity throughout the HGP



•industrialization of Sanger sequencing, library construction, sample preparation, analysis, etc.

•$3 billion total cost

•1 billion bp/month at largest centers (2005)

National Human Genome Research Institute

HGSC HUMAN GENOME SEQUENCING CENTER

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

wellcome trust sanger institute

Second-generation sequencing

"Democratizing" sequencing production

- Massive parallelization
- Reduction in per-base cost
- Eliminate need for huge infrastructure
- Millions of reads - >1Gb sequence per run

Novel sequencing applications

- RNA-seq
- ChIP-seq
- Methyl-seq

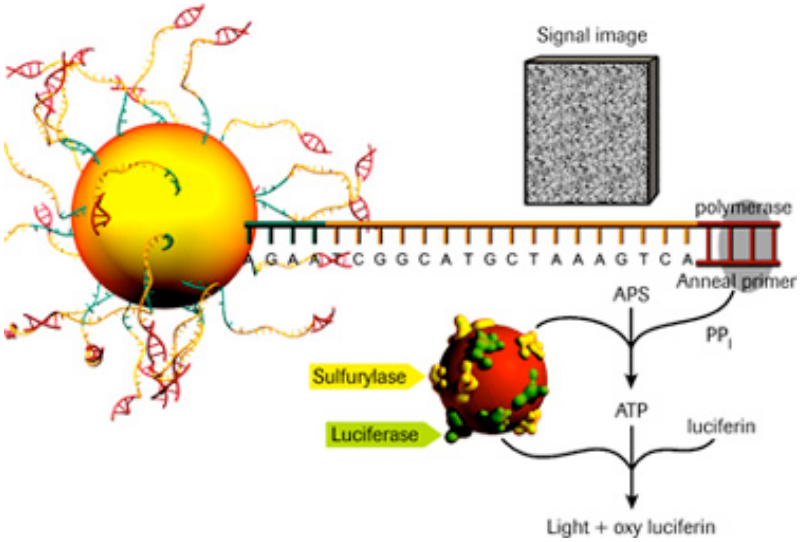
Counting applications

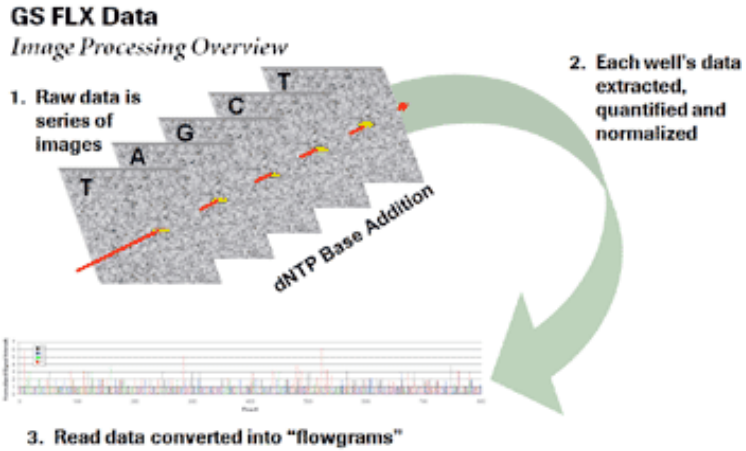
- Whole-genome and targeted resequencing

Challenges

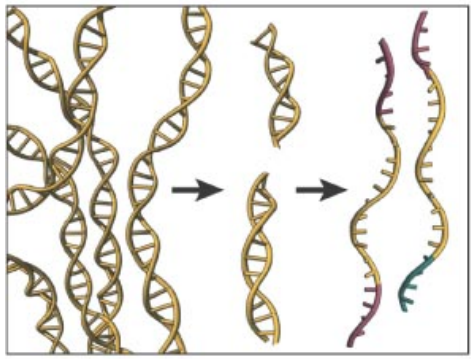
- Read length
- Quality
- Data analysis

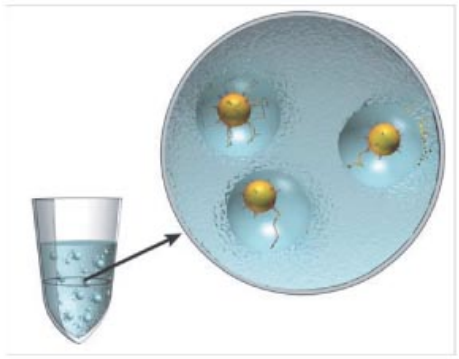
454 pyrosequencing

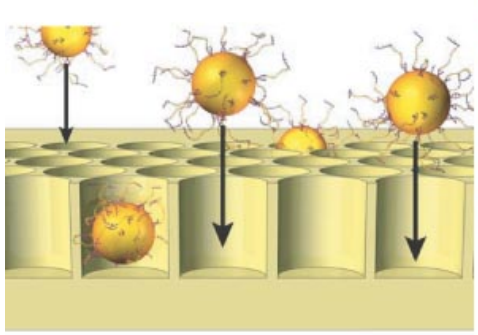


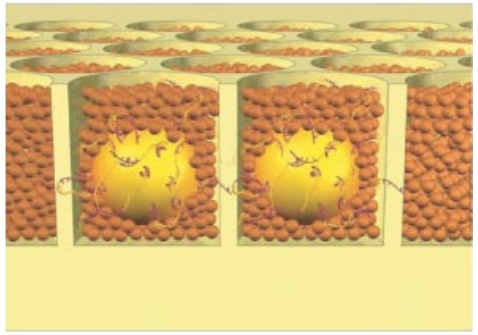


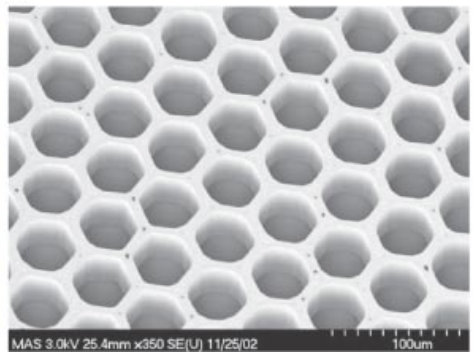
1 cycle: T-A-G-C flowed in sequence across plate
Intensity of signal determines how many nt (i.e. A vs. AAAAA) are incorporated











454 pyrosequencing

Throughput & Yield

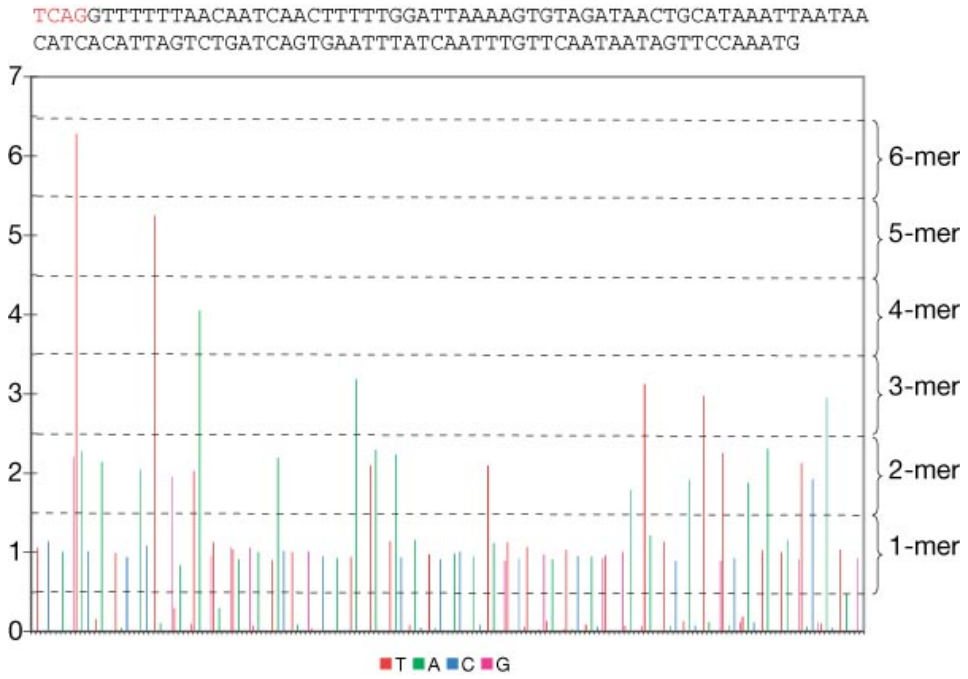
- 1 million 400 bp reads/10 hour run
- >8 samples/run (more with barcoding)

Cost

- Machine: $500k; reagents ~$8000k/run

Issues

- High indel rate in homopolymers
- Longer reads but fewer than other systems



454 sequencing applications

The complete genome of an individual by massively parallel DNA sequencing

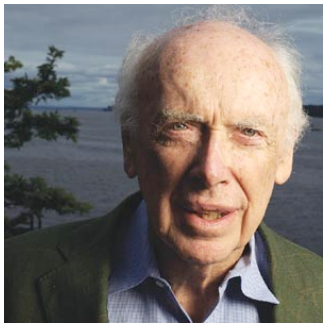


Table 1 | Single nucleotide variation in 454 reads

Subject	Filter*	Total variation	Known†	Novel
Watson	Raw	14,829,087	3,283,273	11,545,814
	1	4,427,488	2,815,322	1,612,166
	2	3,971,513	2,752,991	1,218,522
	3	3,322,093	2,715,296	606,797
Venter‡	4	3,470,669	2,822,902	647,767

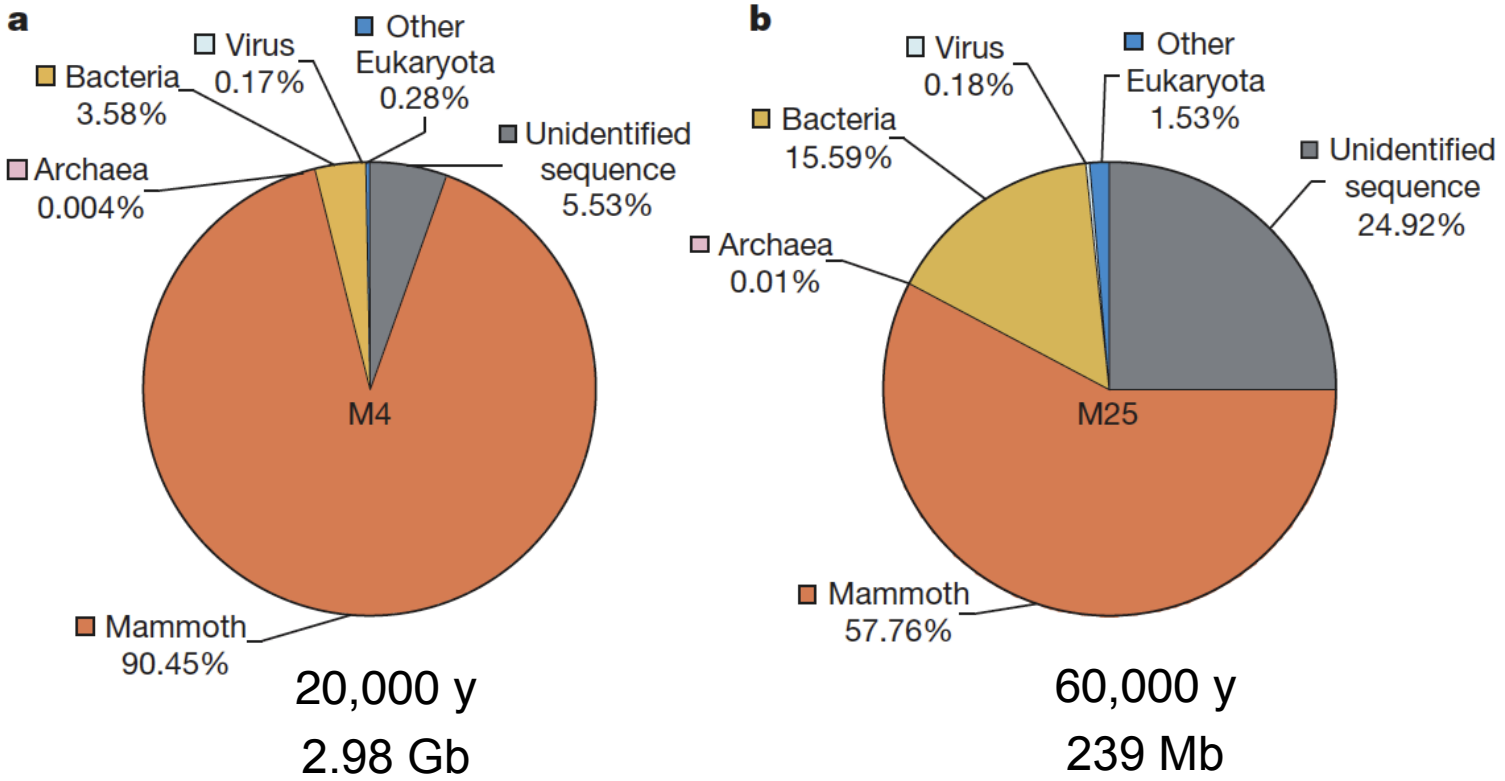
- 106.5 million reads
- 24.5 Gb seq
- 7.4 x coverage

- >10,000 amino acid replacements
- CNVs up to 1.5 Mb

Nature 452:872 (2008)

454 sequencing applications

Sequencing the nuclear genome of the extinct woolly mammoth



20,000 y
2.98 Gb

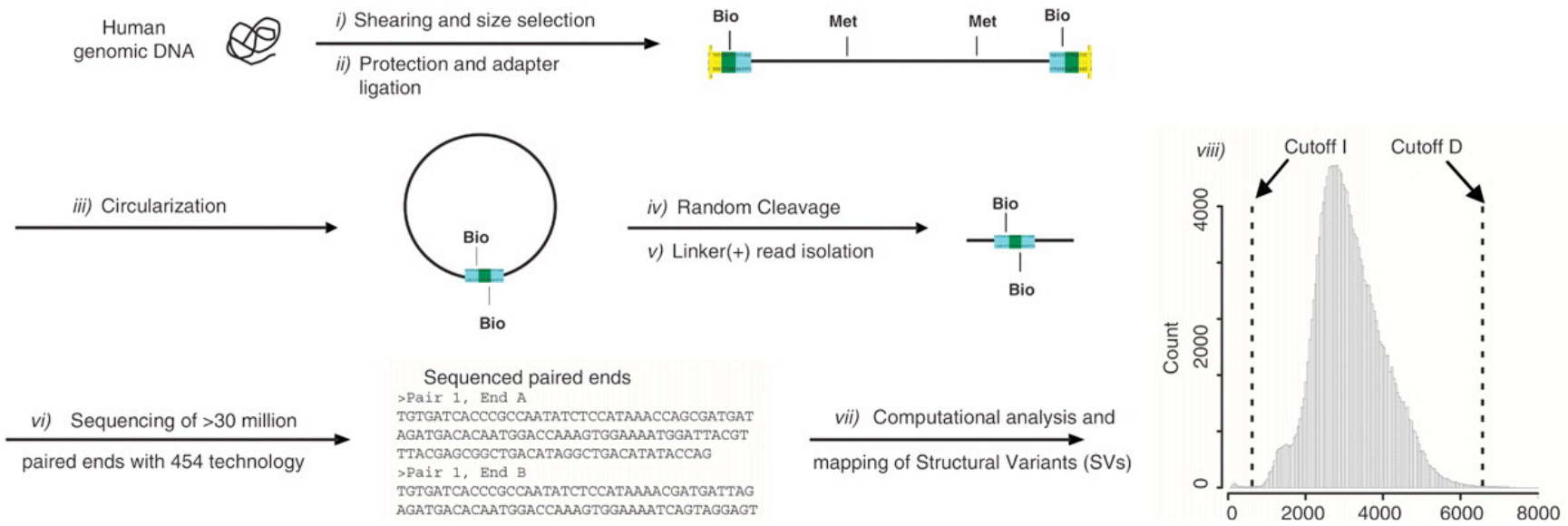
60,000 y
239 Mb

Analysis of one million base pairs of Neanderthal DNA

Richard E. Green[1], Johannes Krause[1], Susan E. Ptak[1], Adrian W. Briggs[1], Michael T. Ronan[2], Jan F. Simons[2], Lei Du[2], Michael Egholm[2], Jonathan M. Rothberg[2], Maja Paunovic[3]‡ & Svante Pääbo[1]

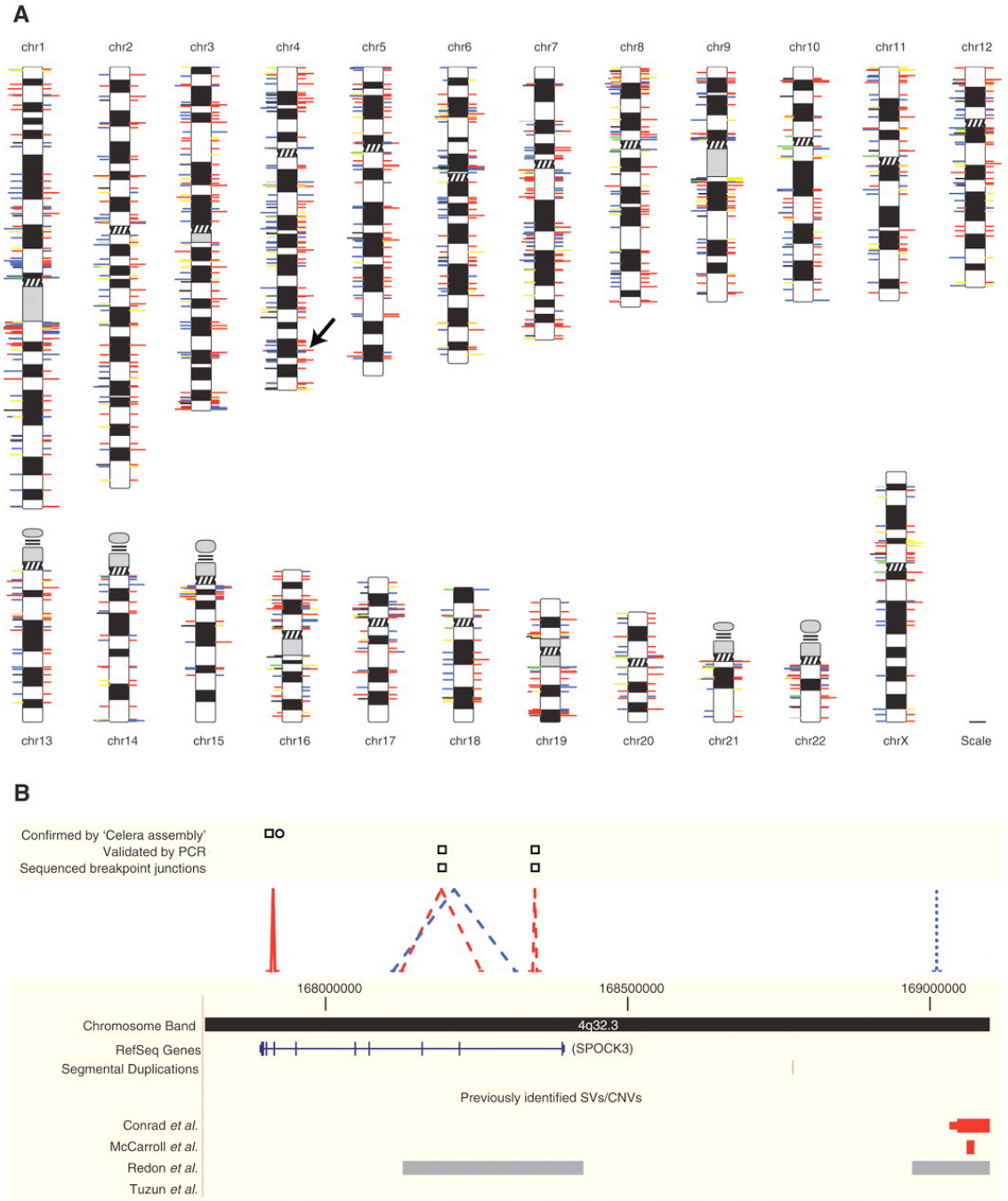
454 sequencing applications

CNV detection by paired-end sequencing



Korbel et al. Science 318:420 (2007)

CNV detection by 454



Korbel et al. Science 318:420 (2007)

Short read technologies

Illumina

- Sequencing by synthesis
- 100 million 36-75 bp reads/run
- $6500 in reagent cost/run
- 3-6 day run time

SOLiD

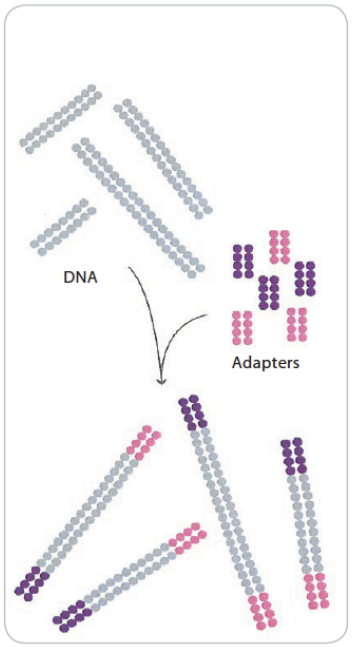
- Sequencing by ligation
- ~400 million 35-50 bp reads/run
- ~$5000 in reagent cost/run
- 3-6 day run time

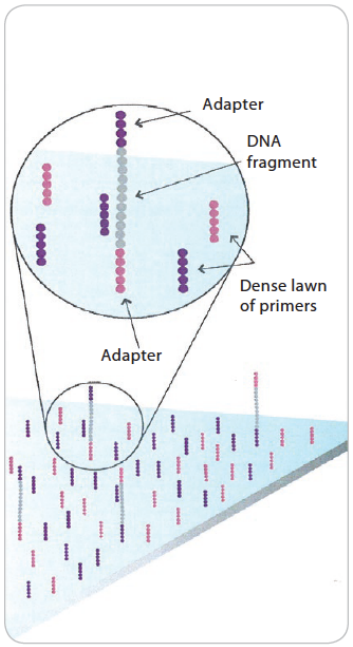
Helicos

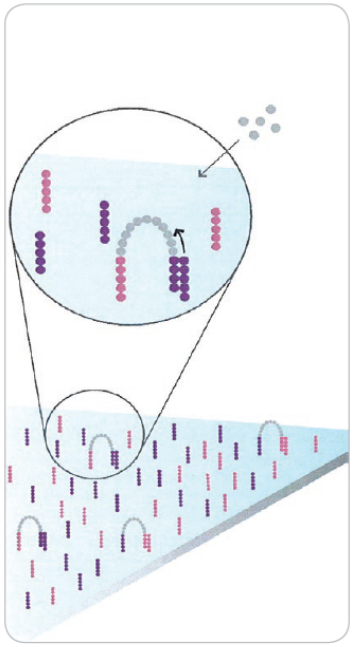
- Sequencing by synthesis
- No amplification
- 750 million reads/run
- $18k run cost
- 8 day run time

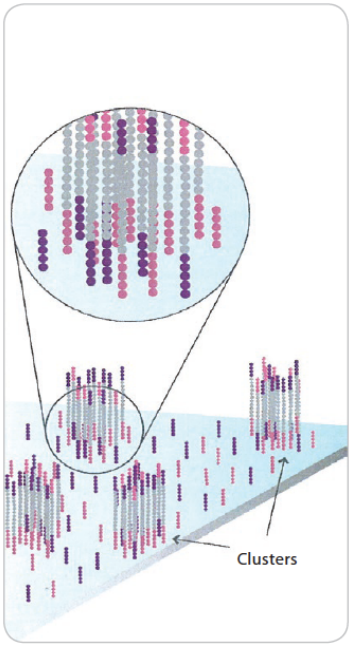
Illumina

Cluster PCR
on flowcell
(8 lanes)

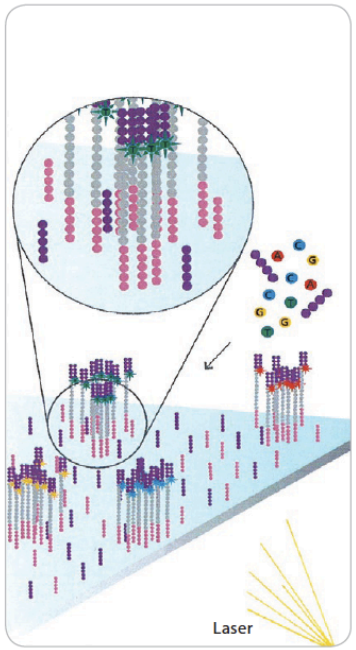


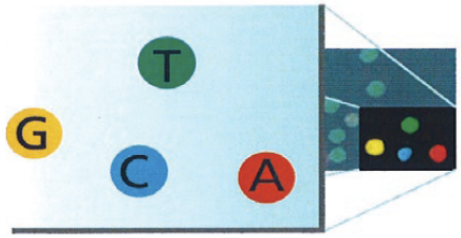


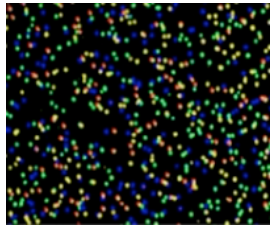




Sequencing
by synthesis
with reversible
dye terminators



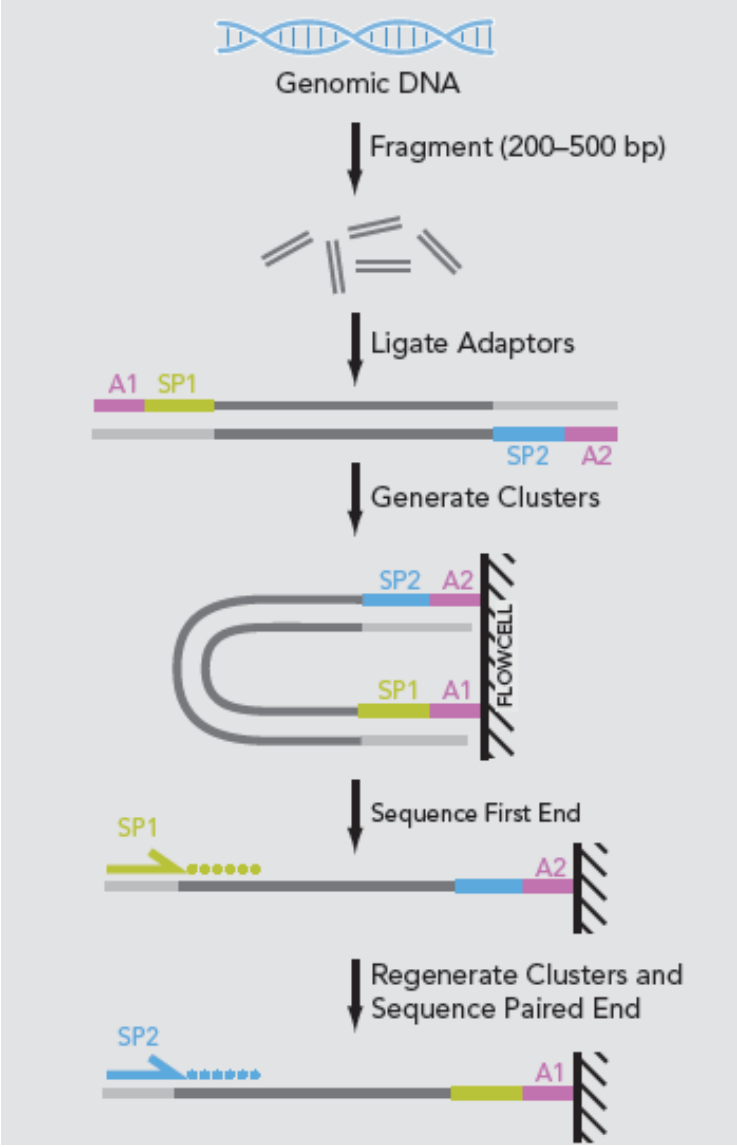




Scan flowcell
Reverse termination
Add next base

1 cycle

Paired end sequencing



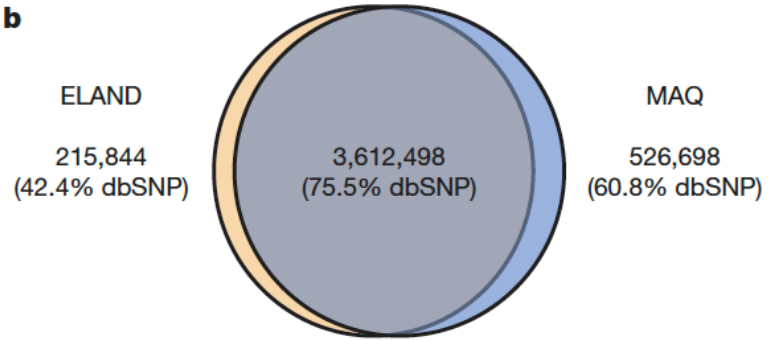
Accurate whole human genome sequencing using reversible terminator chemistry

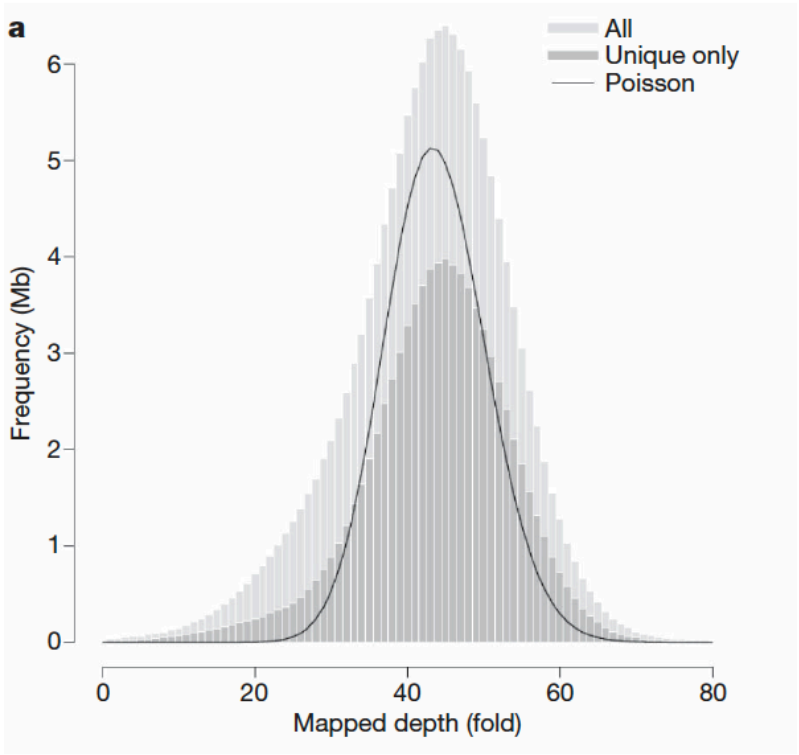
Yoruba male from Nigeria

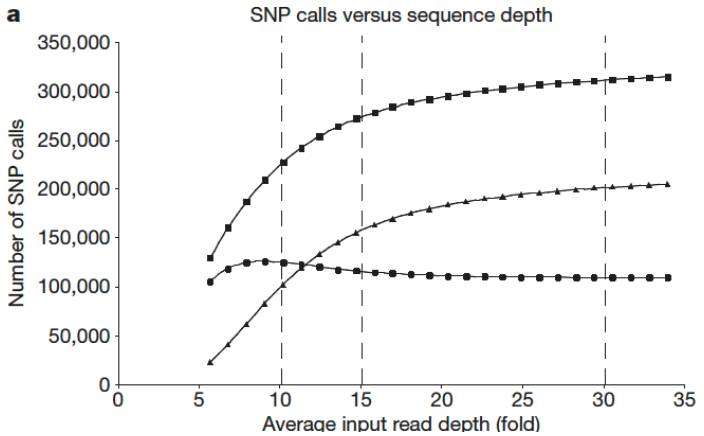
a

Call	ELAND SNPs (*n*)	ELAND In dbSNP (%)	MAQ SNPs (*n*)	MAQ In dbSNP (%)
Homozygote	1,417,320	90.1	1,503,420	90.8
Heterozygote	2,411,022	63.9	2,635,776	63.8
All	3,828,342	73.6	4,139,196	73.6

b

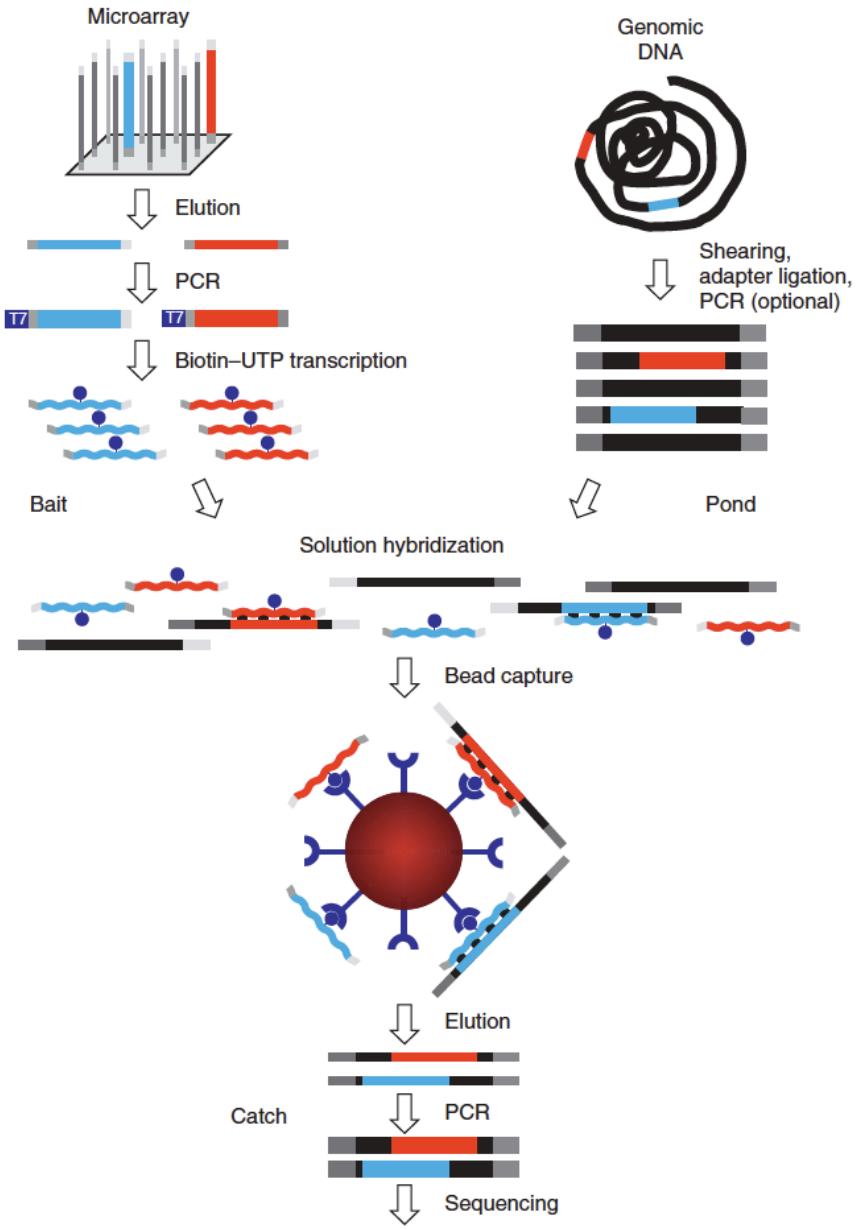






Nature 456:53 (2008)

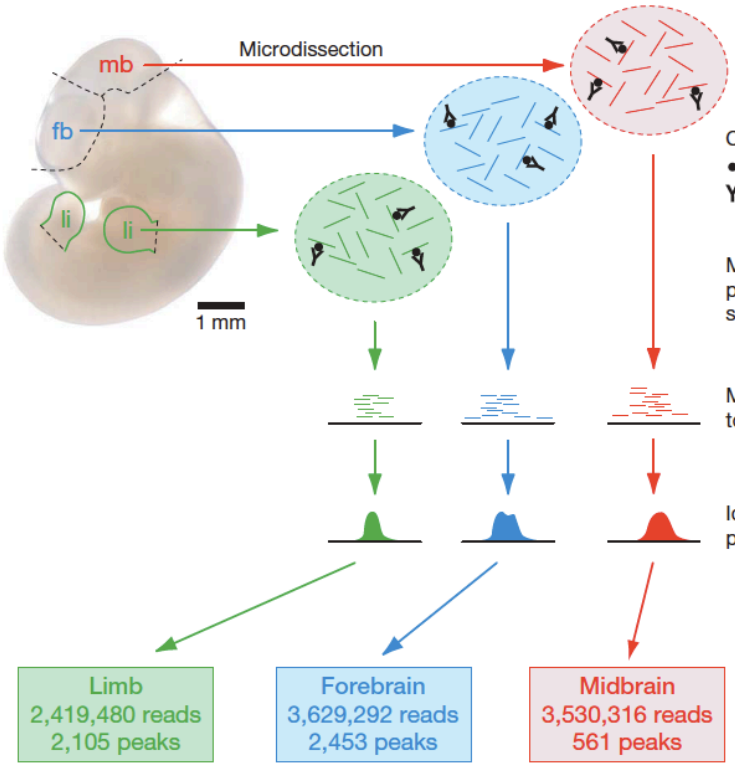
Targeted resequencing by sequence capture

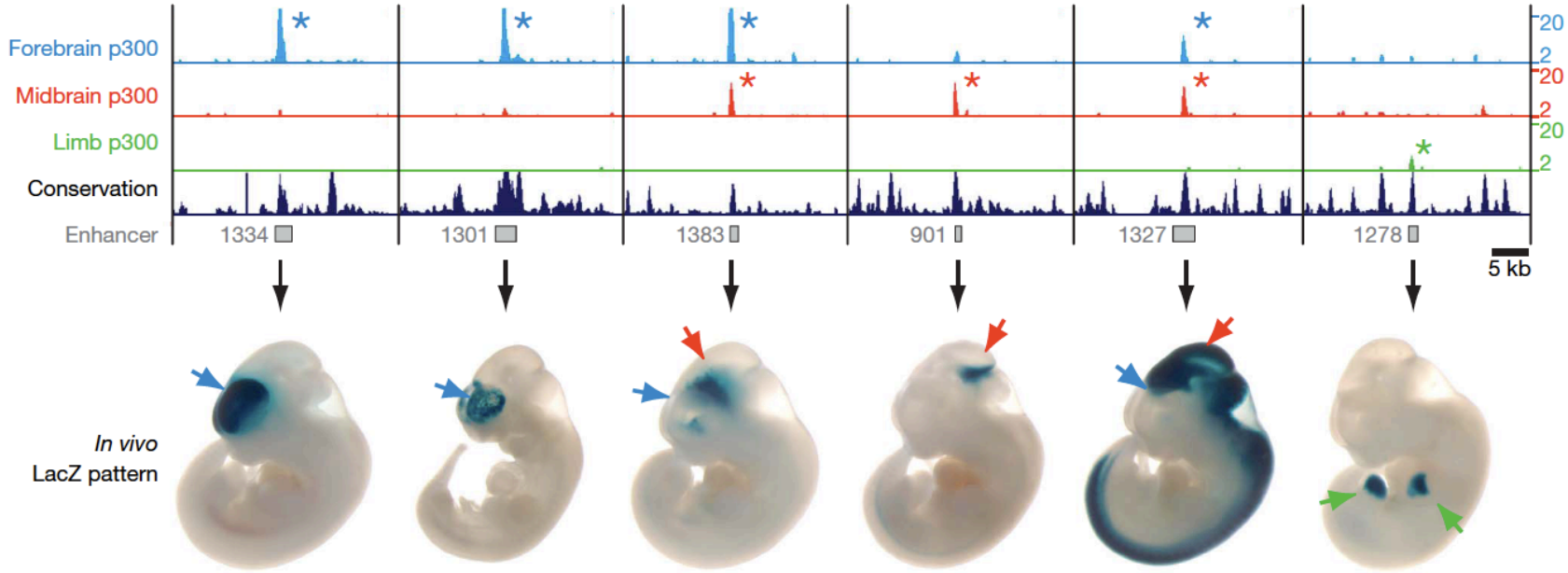


ChIP-seq: enhancer identification *in vivo*

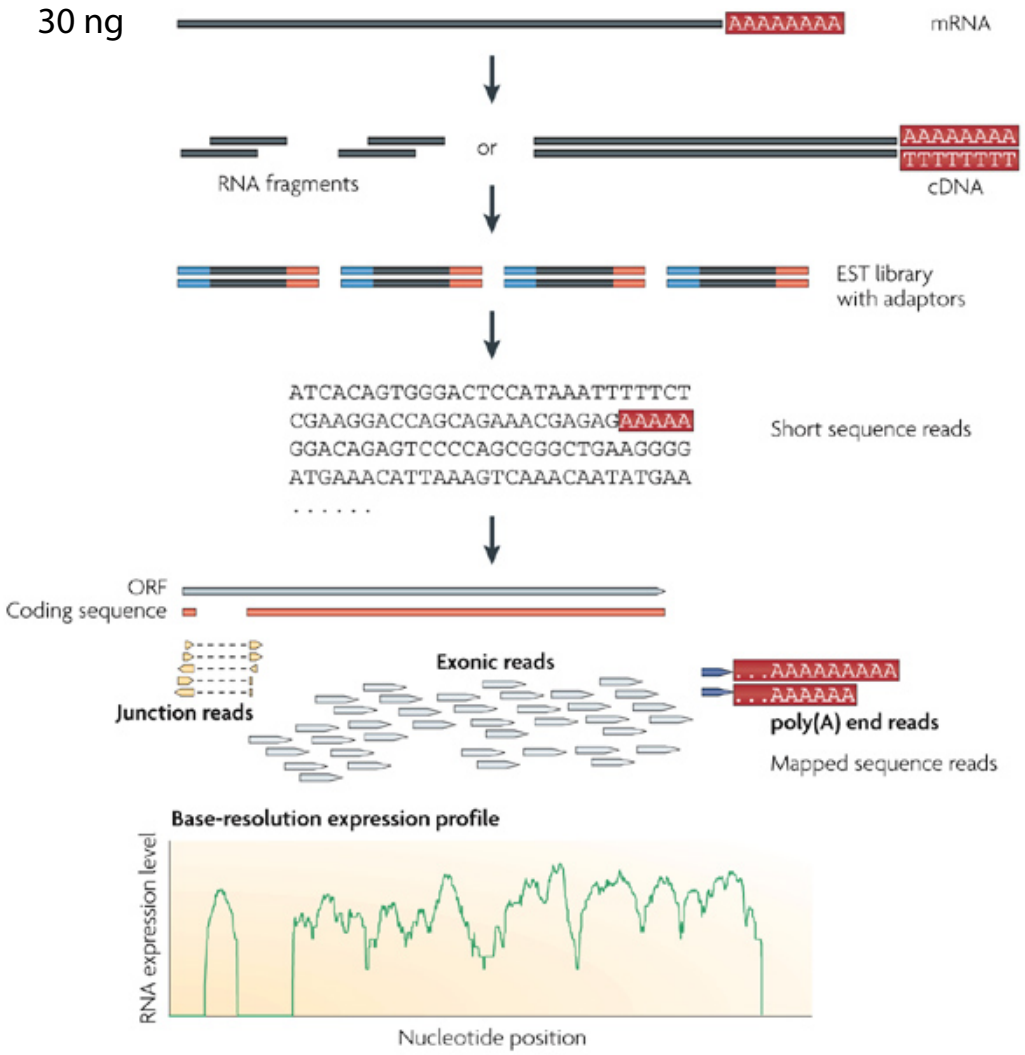
Visel et al. Nature 457:854 (2009)

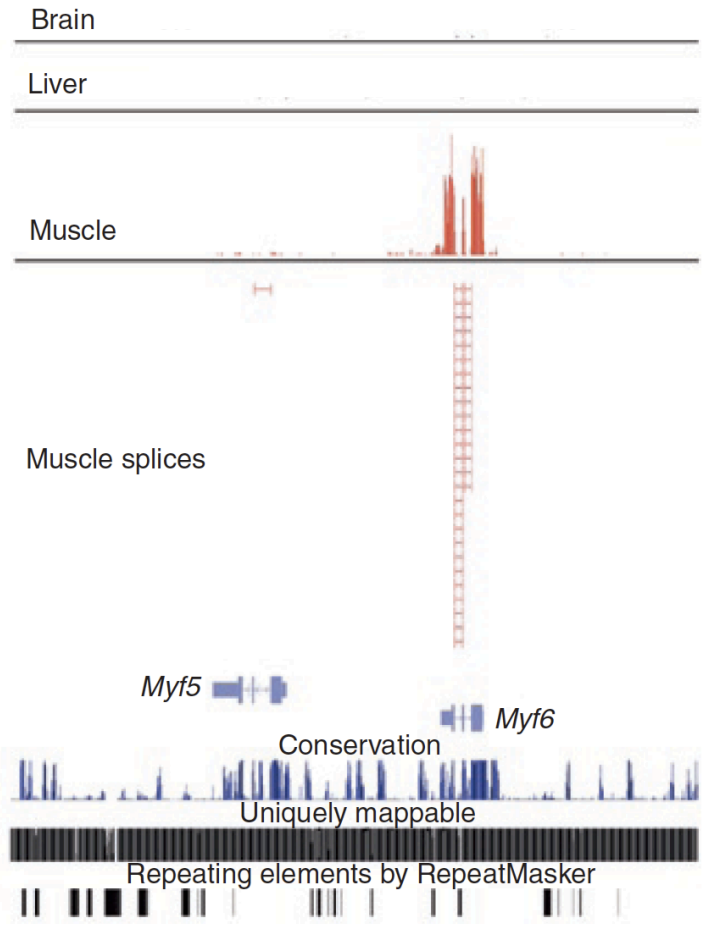
- p300 = enhancer-associated factor
- p300 binding = ~90% predictive of enhancer activity





Gene expression profiling by massively parallel RNA sequencing (RNA-seq)



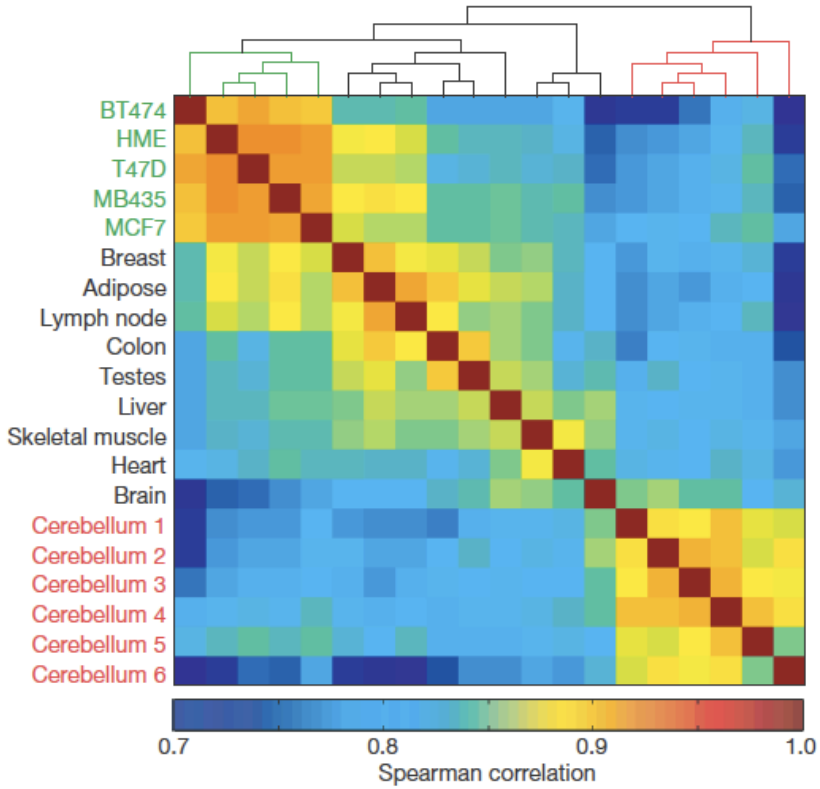


Analysis of alternative splicing by RNA sequencing

Alternative transcript events	Total events (×10³)	Number detected (×10³)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon	37	35	10,436	6,822	65	72
Retained intron	1	1	167	96	57	71
Alternative 5′ splice site (A5SS)	15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)	17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)	4	4	167	95	57	66
Alternative first exon (AFE)	14	13	10,281	5,311	52	63
Alternative last exon (ALE)	9	8	5,246	2,491	47	52
Tandem 3′ UTRs	7	7	5,136	3,801	74	80
Total	105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

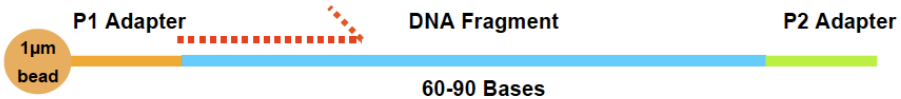


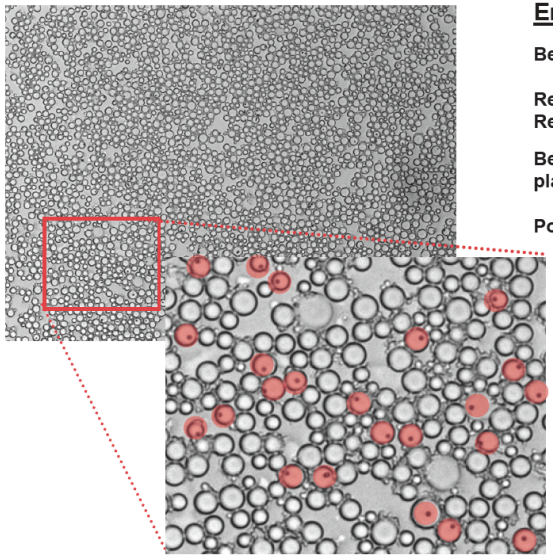
>92% of human genes undergo alternative splicing

Splicing varies more among tissues than among individuals

Wang et al. Nature 456:470 (2008)

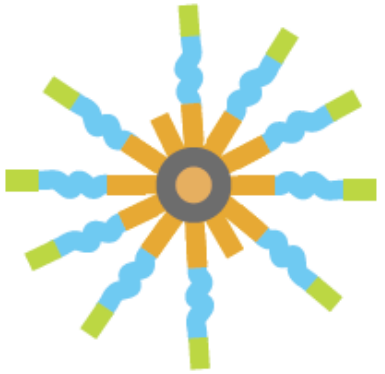
SOLiD

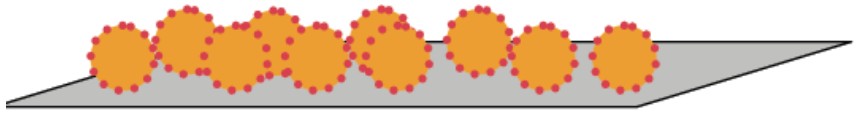




Emulsion Metrics

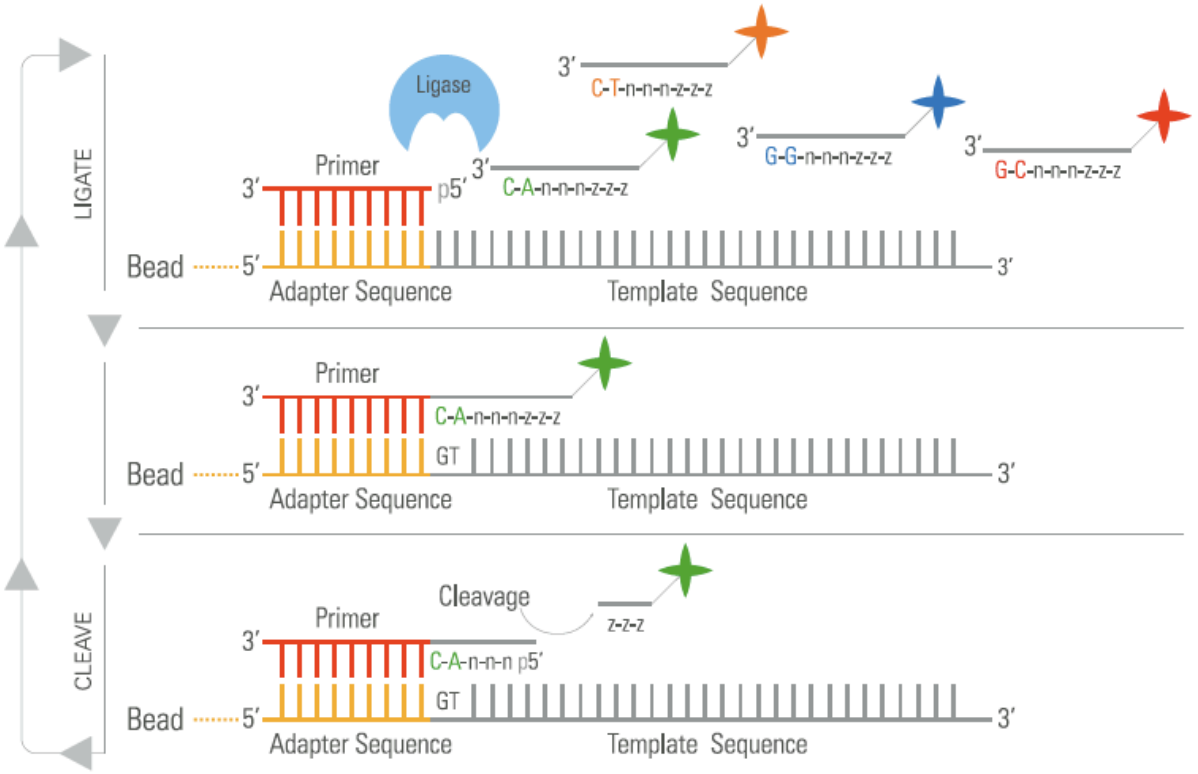
Bead size:	1 µm
Reactor size:	4 µm
Reactor volume:	34 fL
Beads / emulsion plate (96-well):	2-4 x 10^9
Post Enrichment:	~500M / plate

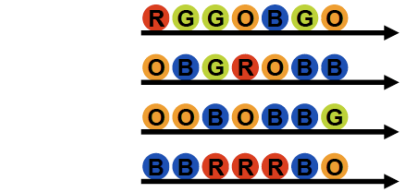


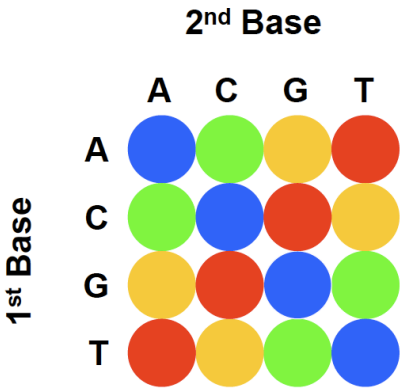


Beads attached to glass surface in a **random** array

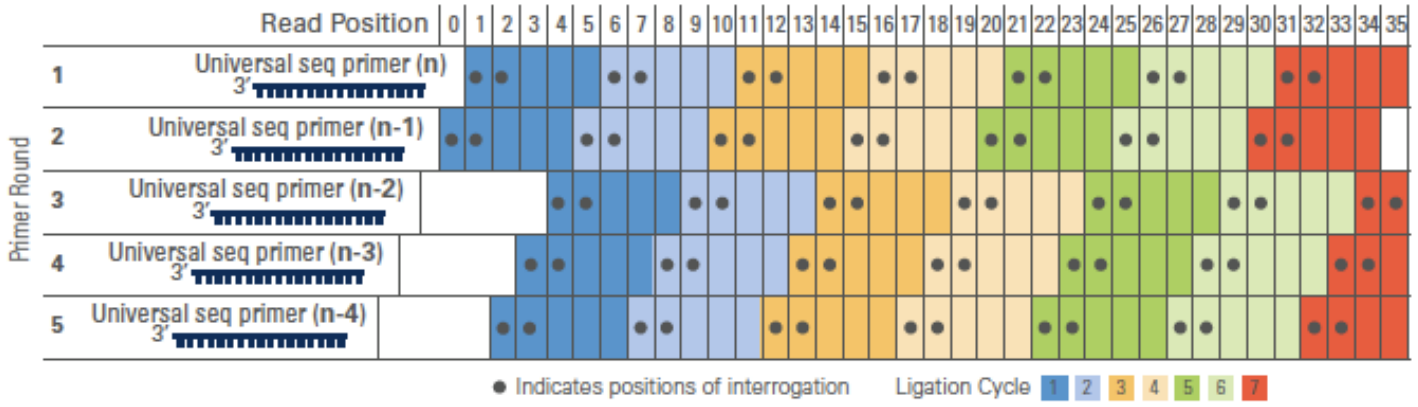
SOLiD



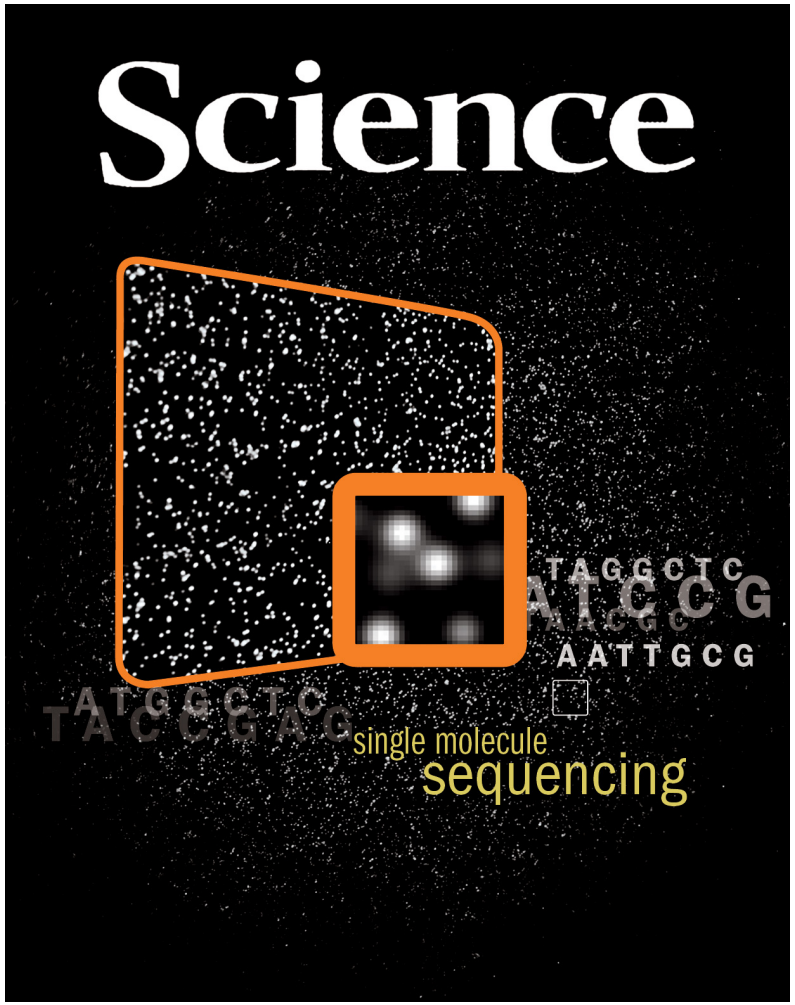




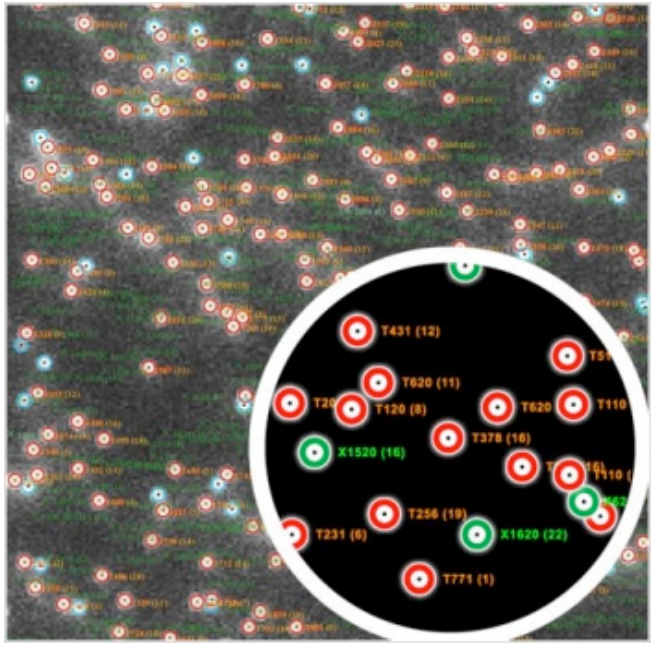
Two base encoding in color space



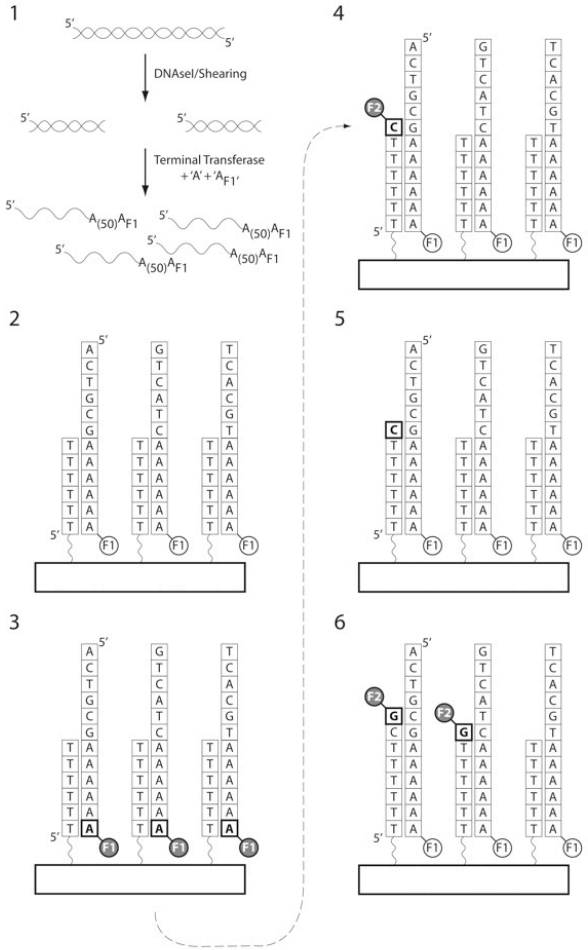
Single molecule sequencing: Helicos

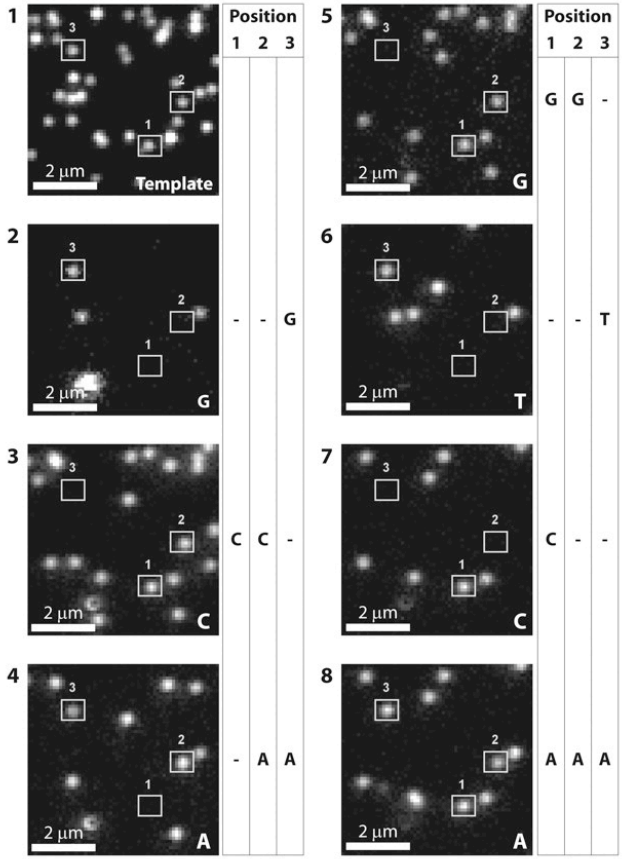


Helicos
BioSciences Corporation



Single molecule sequencing: Helicos





Base #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Oligo #																				
1	C	G	C	A	G	A	T	A	T	C	G	C	A	T	C	A				
2			C	A	G	A	T	A	T	C	G	C	A	T	C	A	G	T		
3					G	A	T	A	T	C	G	C	A	T	C	A	G	T	C	A

- No PCR
- 800 million reads, 50 samples
- high indel rate (3%)

Third-generation sequencing

Extremely high-throughput sequencing at very low cost

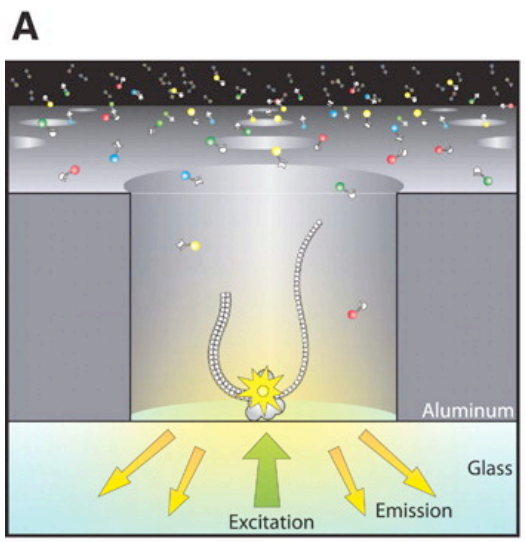
Pacific Biosciences

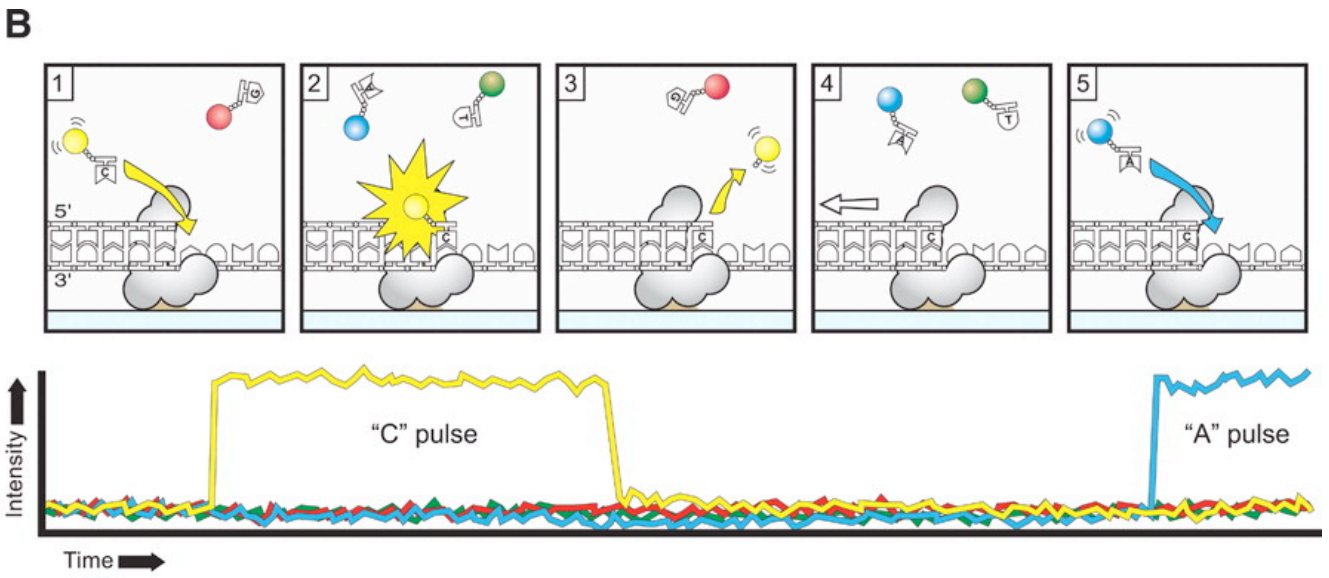
- Sequence in real time with fluorescent NTPs
- Rate limited by processivity of polymerase
- Very long reads (>10 kb)
- Not well parallelized (few reads)

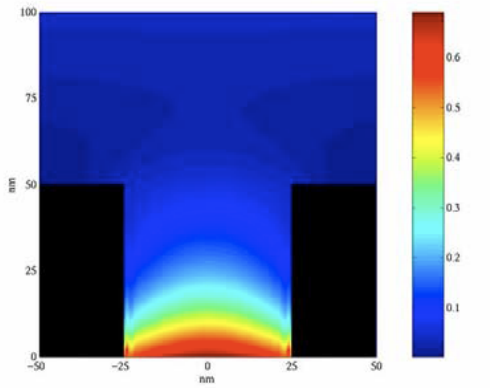
Nanopore sequencing

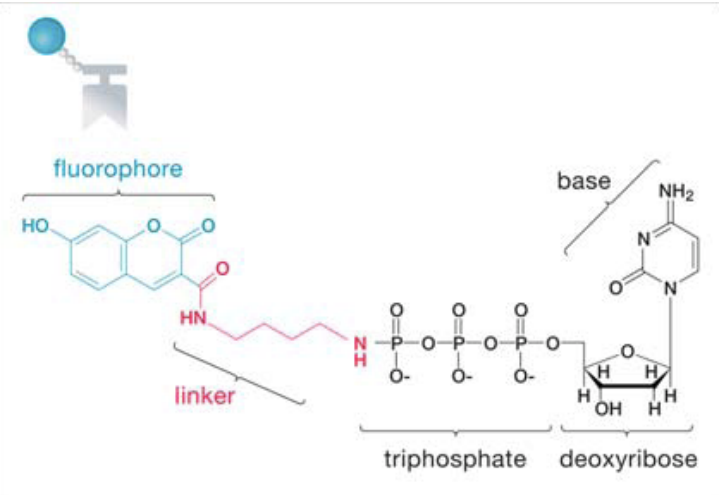
- Sequencing by exonuclease cleavage of native DNA
- Bases are read as they pass through a modified nanopore - base-specific change in current

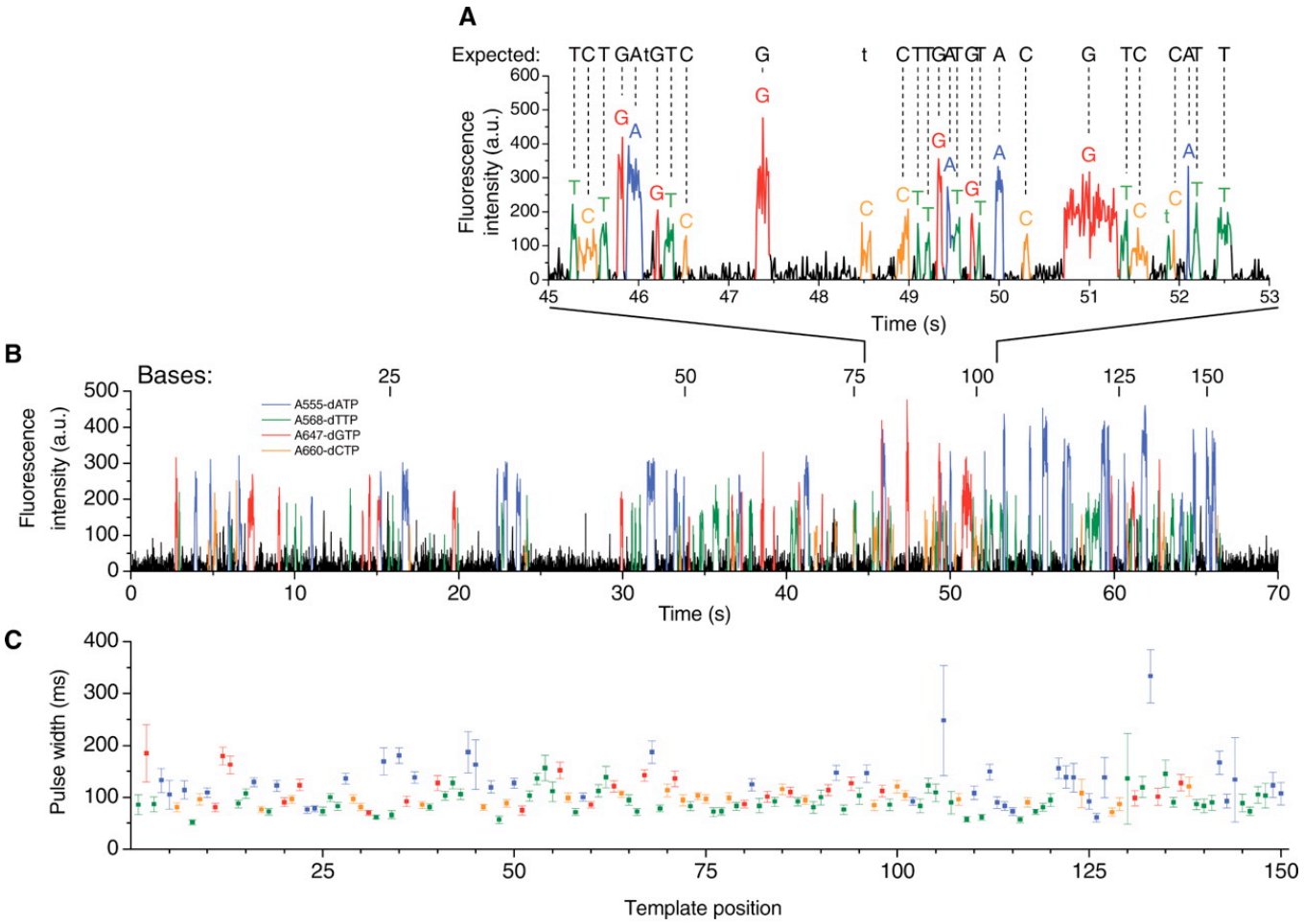
Sequencing in real time: Pacific Biosciences



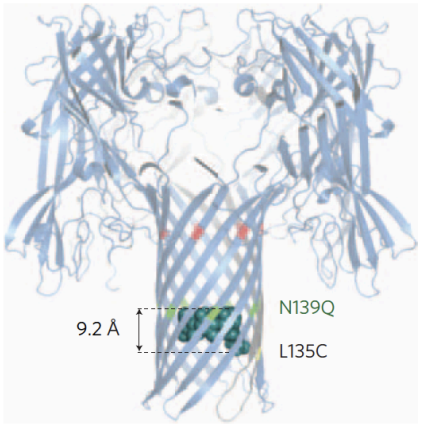


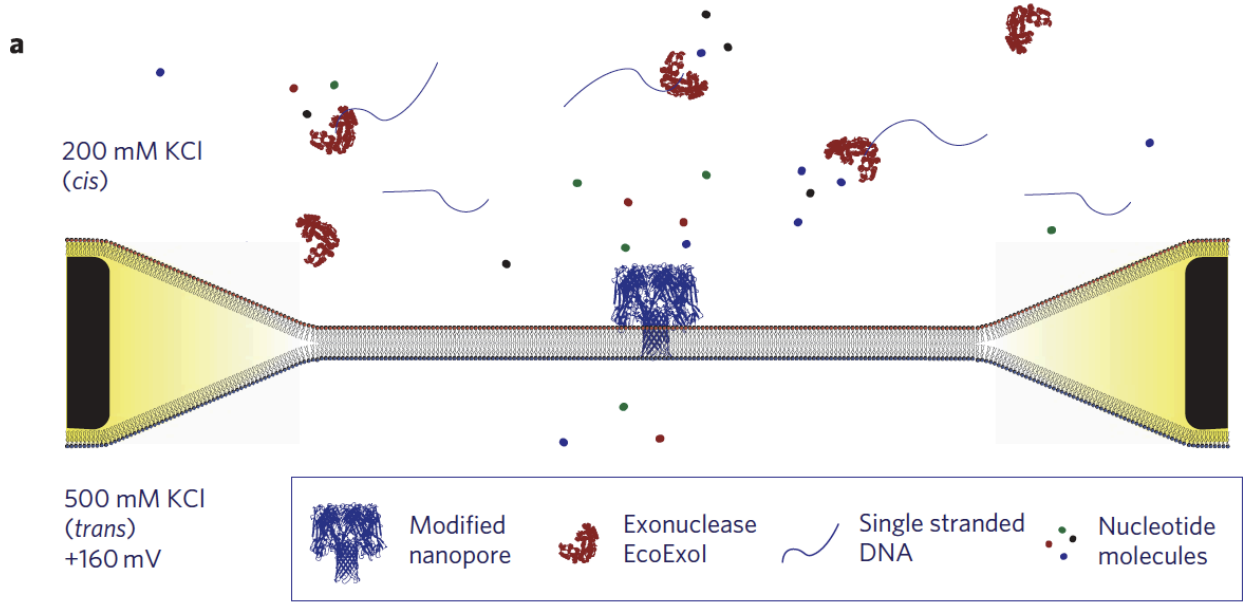


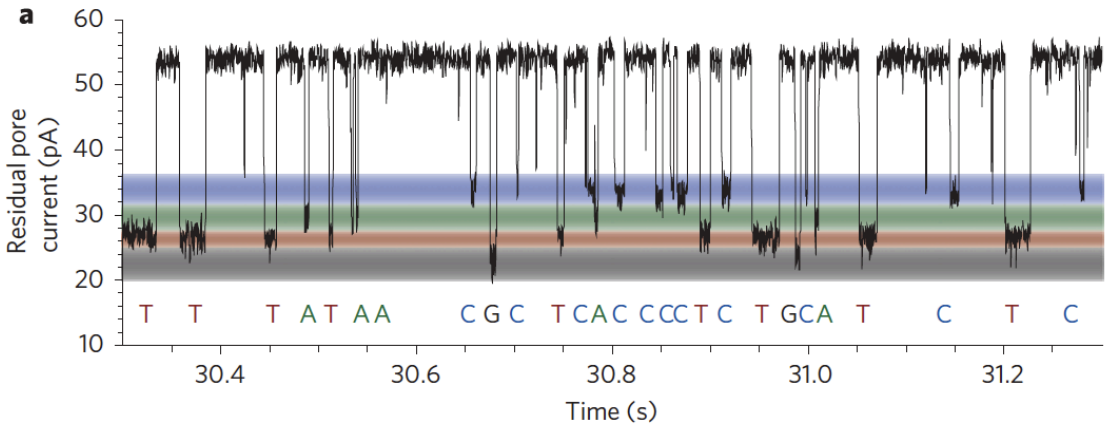


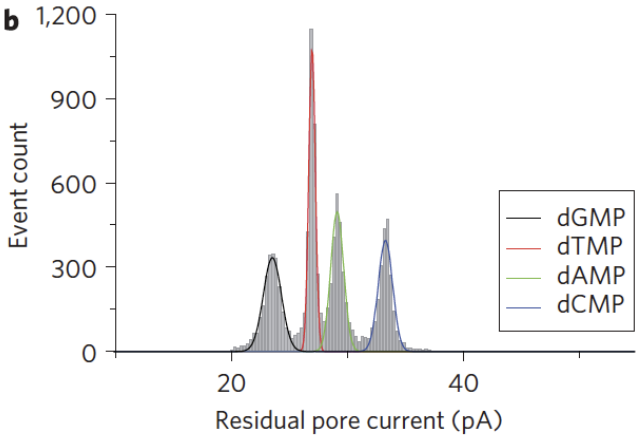

A
Expected: TC T GAtGT C G t CTTGATGT A C G TC CAT T
Fluorescence intensity (a.u.)
Time (s)
B
Bases:
A555-dATP
A568-dTTP
A647-dGTP
A660-dCTP
Fluorescence intensity (a.u.)
Time (s)
C
Pulse width (ms)
Template position

Nanopore sequencing









Conclusions

- High-throughput sequencing has become democratized - moved out of industrial-scale genome centers
- Sequence is no longer limiting - next generation of sequencers will make sequencing very inexpensive
- Earlier methods for counting / resequencing applications are largely obsolete
- Current challenge: how do we handle all the data?