- Illumina Sequencing Technology
- Computational Options
- Applications
- Genome Assembly
    - read mapping
    - reference guided
    - de novo
- Diving In

Computer Exercises

Data-Driven Science vs. Hypothesis Testing

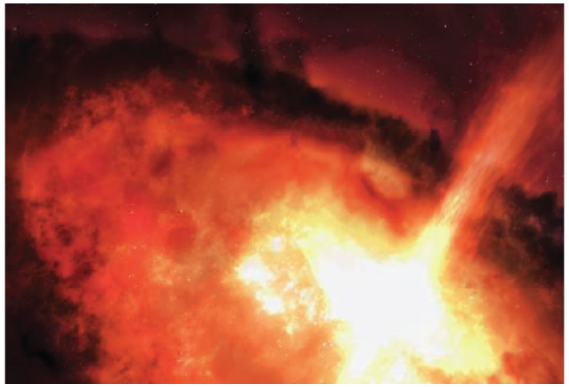
PERSPECTIVES

HISTORY OF SCIENCE

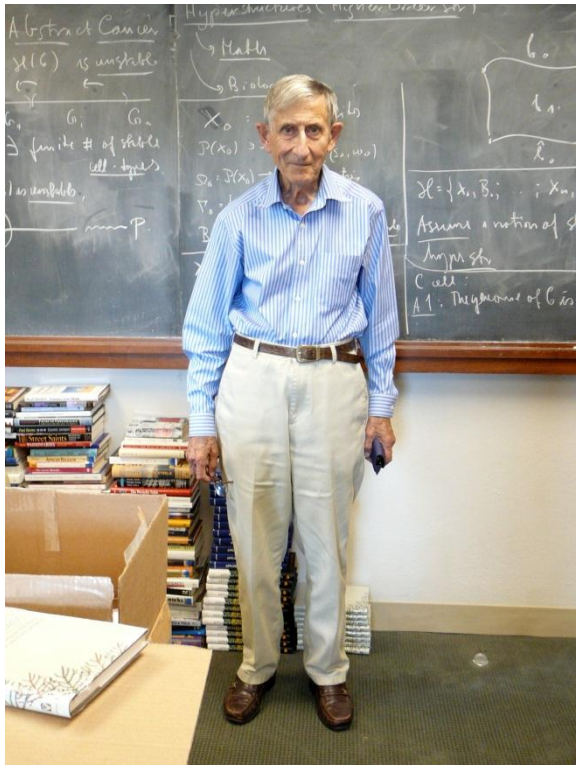
Is Science Mostly Driven by Ideas or by Tools?

Freeman J. Dyson

Thomas Kuhn was a theoretical physicist before he became a historian. He saw the history of science through the eyes of a theorist. He gave us an accurate view of events in the world of ideas. His favorite word, "paradigm," means a system of ideas that dominate the science of a particular place and time. A scientific revolution is a discontinuous shift from one paradigm to another. The shift happens suddenly because new ideas explode with a barrage



The physical sciences have alternated between revolutions driven by new ideas and explorations driven by new tools.



Dyson, F. (2012) Science 338:1426-1427

"It is better to be wrong than to be vague."
F. Dyson. 1999. The Sun, the Genome, and the Internet.

wikipedia.org

Data-Driven Science vs. Hypothesis Testing

Altshuler's path to genomics involved the same attitude that brought about the human genome project, a desire to "get outside of hypothesis-limited science" in which a huge number of scientists pursued the same leads, the same "very small number of reductionist models." Instead there must be a complete catalog of genetic variation, unbiased by educated guesses.

David Altshuler quoted in Victor McElheny (2010) Drawing the Map of Life: Inside the Human Genome Project. Page 198.

LETTER

OPEN

doi:10.1038/nature12309

Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds

Rajinder Singh[1]*, Meilina Ong-Abdullah[1]*, Eng-Ti Leslie I
Leslie Cheng-Li Ooi[1], Siew-Eng Ooi[1], Kuang-Lim Chan[1], N
Nathan Lakey[2], Steven W. Smith[2], Dong He[2], Michael Ho
David Kudrna[4], Jose Luis Goicoechea[4], Rod A. Wing[4], Ric
Robert A. Martienssen[6] & Ravigadevi Sambanthamurthi[1]

are highly expressed in the kernel. We also repor
the South American oil palm *Elaeis oleifera*,
number of chromosomes ($2n = 32$) and produ
hybrids with *E. guineensis*[2] but seems to have
World. Segmental duplications of chromoso
palaeotetraploid origin of palm trees. The oil p
the discovery of genes for important traits a
epigenetic alterations that restrict the use of c
plantings[3], and should therefore help to achi
biofuels and edible oils, reducing the rainfor
tropical plantation crop.

Oil palm genome boosts hopes for tropical forests

Palm oil fruits in a cart. Production of palm oil from the African palm (Elaeis guineesis) is one of the largest industries in Costa Rica. The oil extracted from the palm oil dates is used in many commercial goods including candy, candles and cosmetics.

(AFP) Sequencing of the oil palm, one of the world's most important crops, has pinpointed a gene that should boost yields and ease pressure on tropical rainforests, studies said on Wednesday.

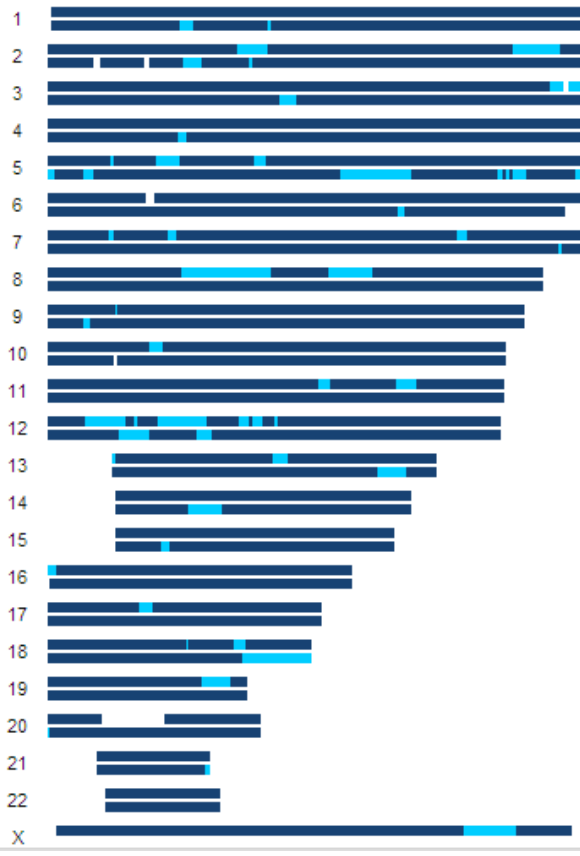
https://www.23andme.com/you/ancestry/composition/

23andMe

Ancestry Composition

Aaron Liston | Conservative Est

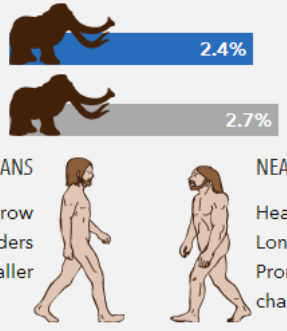
Chromosome View | Regional Resolution



Ancestry Composition tells you what percent of your DNA come each of 22 populations worldwide. The analysis includes DNA received from all of your ancestors, on both sides of your family results reflect where your ancestors lived 500 years ago, before ocean-crossing ships and airplanes came on the scene.

99.4%	European
93.2%	Ashkenazi
0.0%	Southern European
0.0%	Northern European

An estimated 2.4% of your DNA is from Neanderthals.



MODERN HUMANS

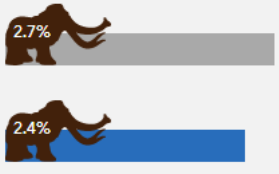
Higher brow
Narrower shoulders
Slightly taller

NEANDERTHALS

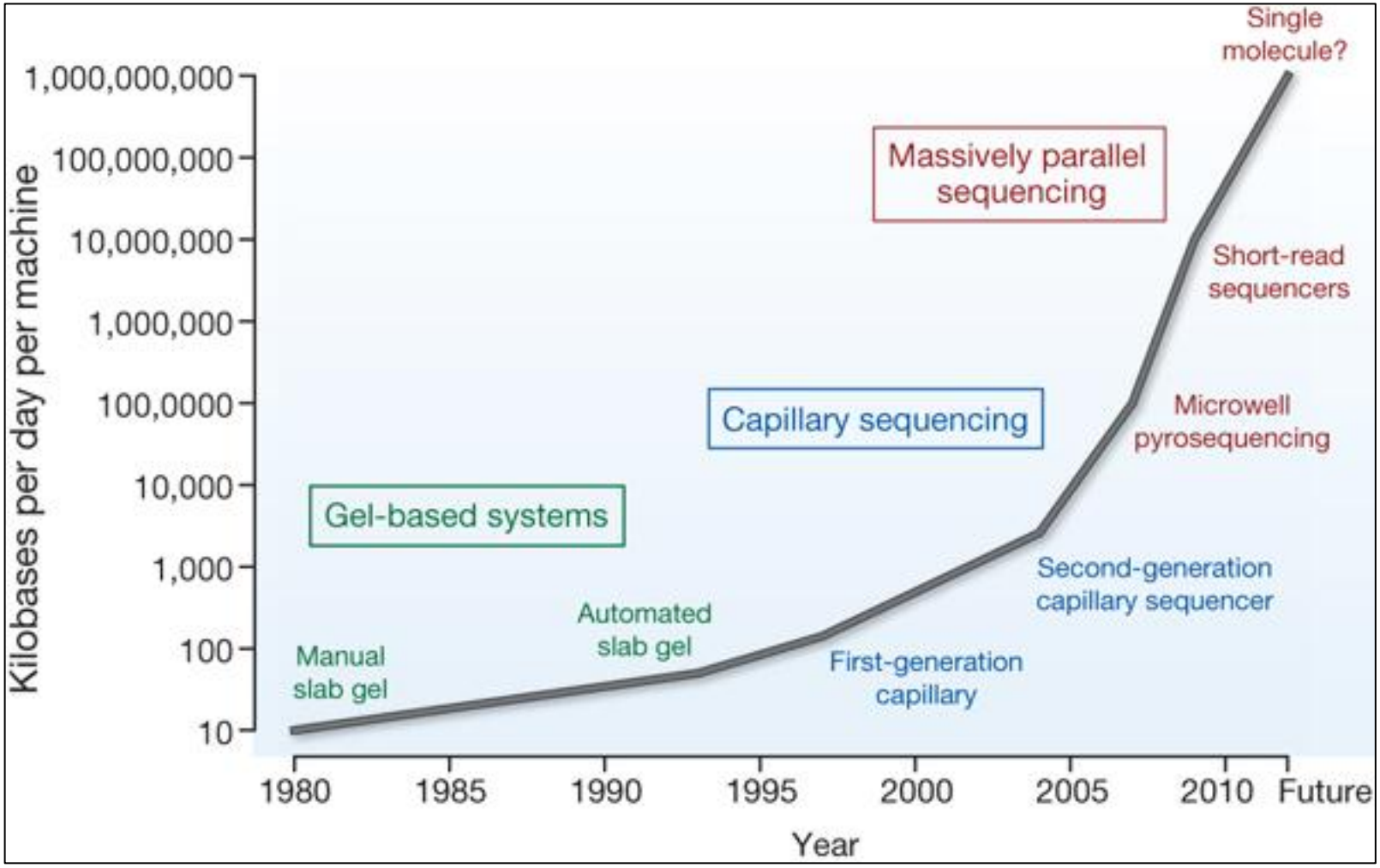
Heavy eyebrow ridge
Long, low, bigger skull
Prominent nose with developed nasal chambers for cold-air protection

Friends & Family

You are ranked 2nd among your friends. Invite more friends

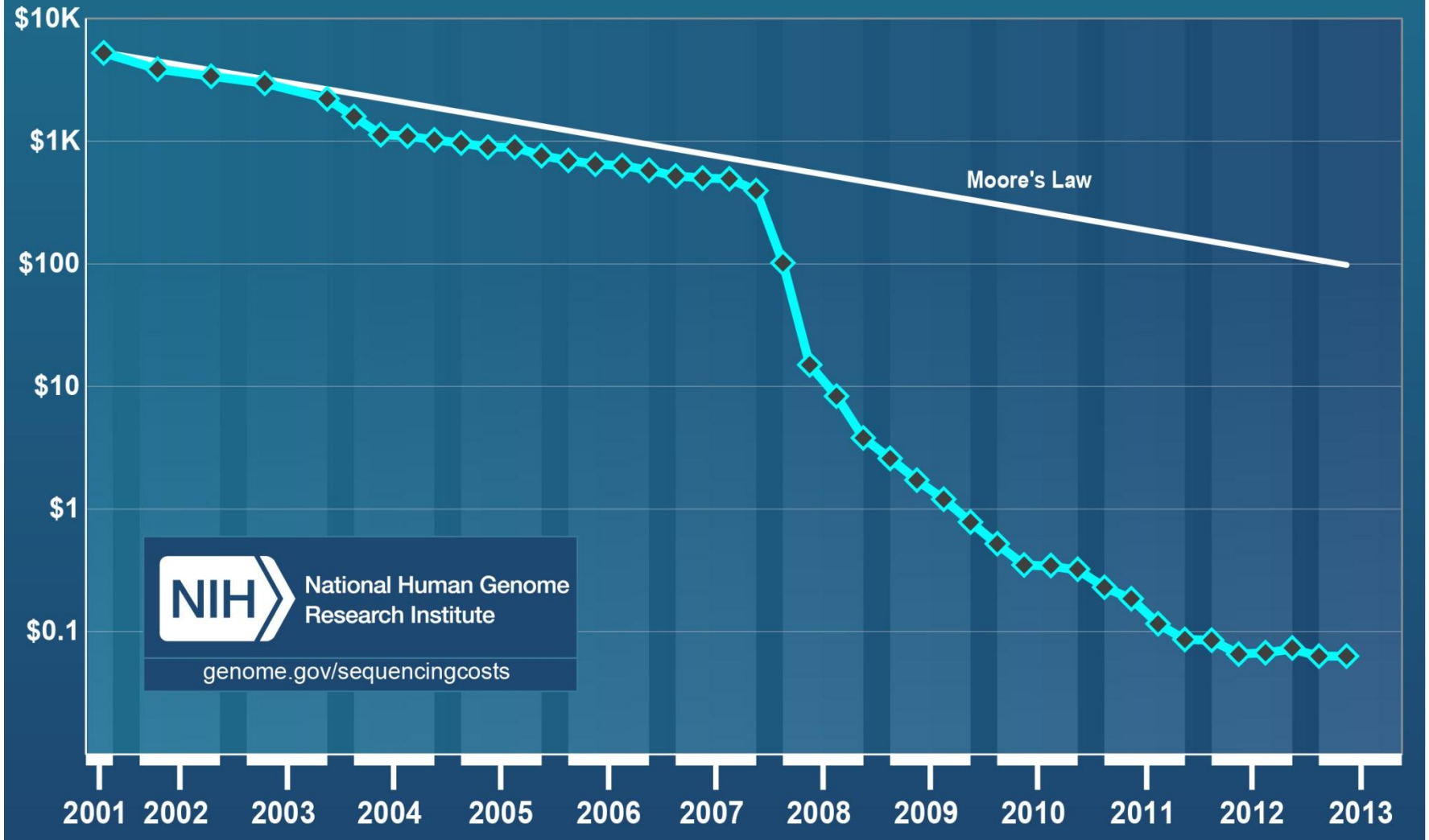


Improvements in the rate of DNA sequencing over the past 30 years



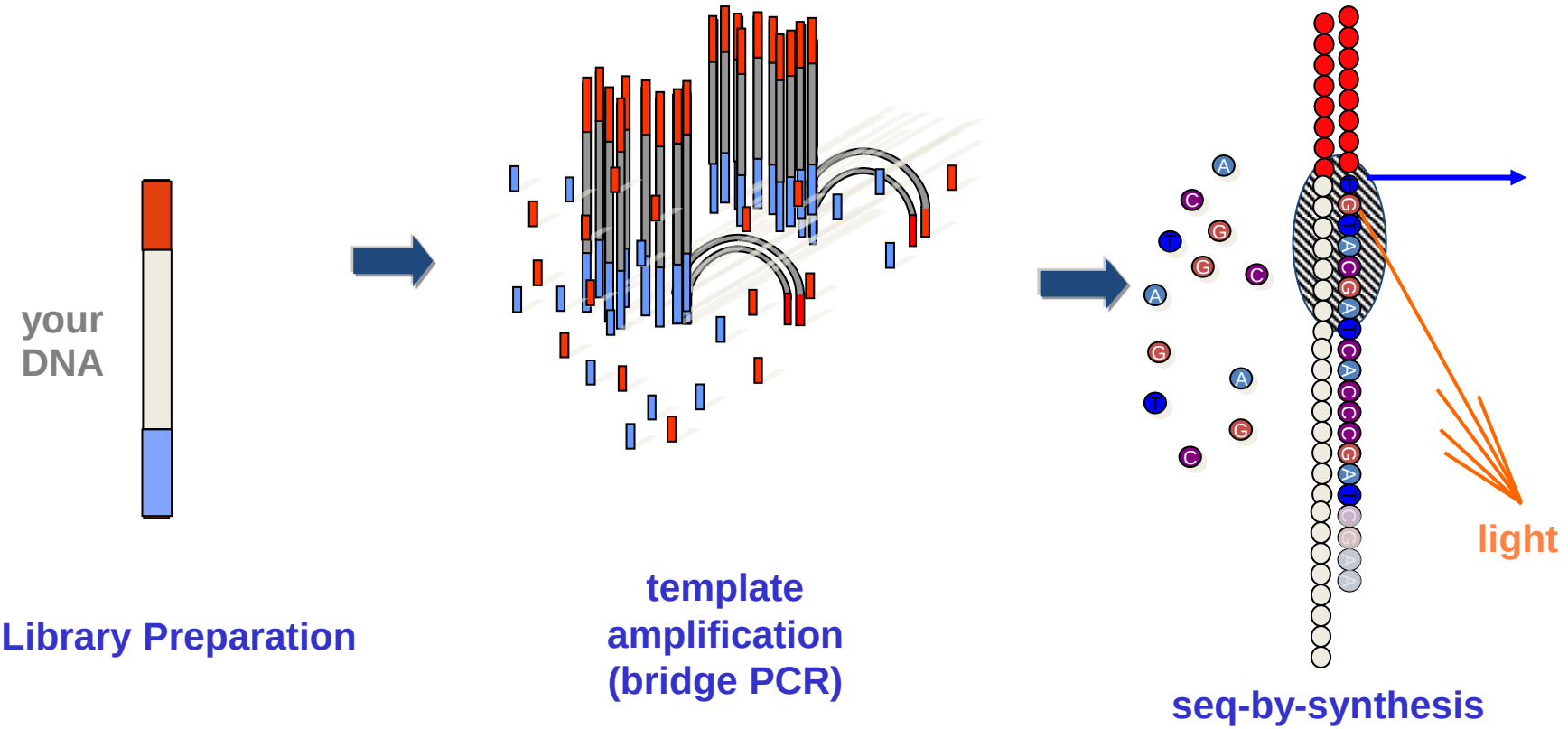
MR Stratton et al. (2009) The cancer genome. Nature **458**, 719-724

Cost per Raw Megabase of DNA Sequence



Illumina Sequencing

Available since 2007 (Solexa purchased by Illumina in 2008)



Library Preparation

template amplification (bridge PCR)

seq-by-synthesis

input:

- **500 ng DNA**
- **1 ug total RNA**

Illumina Library Prep

genomic DNA, RNA
BACs, amplicons ...

DNA fragmentation

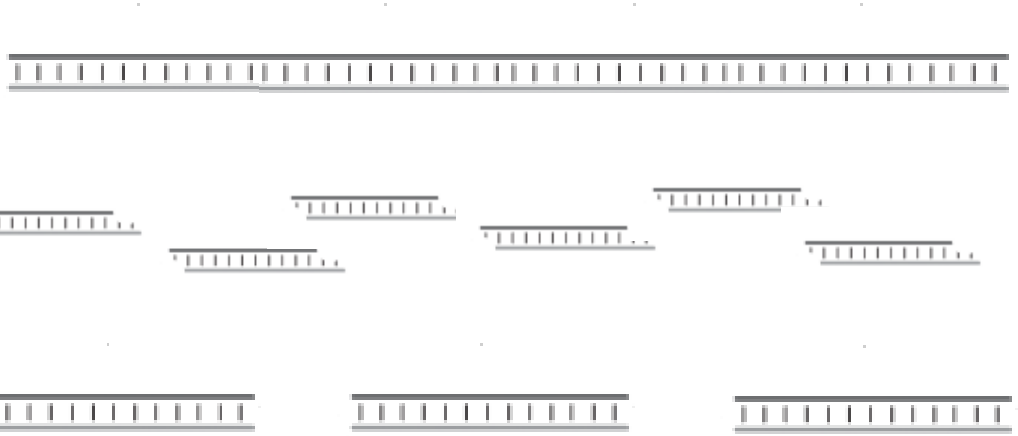
end-repair

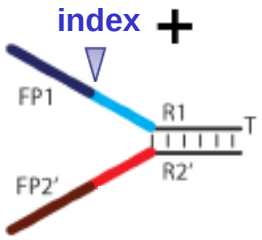
A-tailing

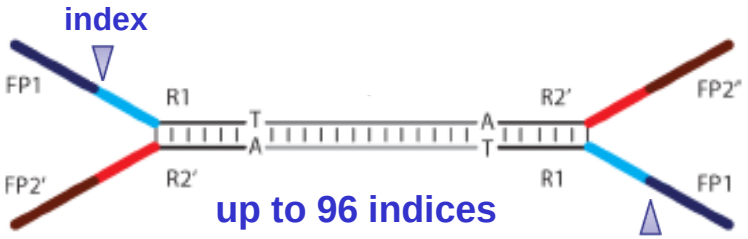
adapter-ligation

size selection

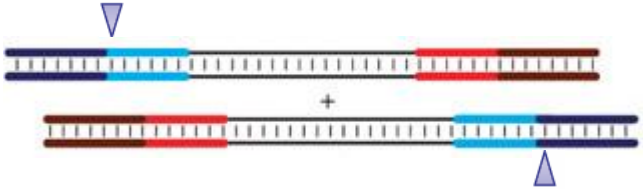
enrichment

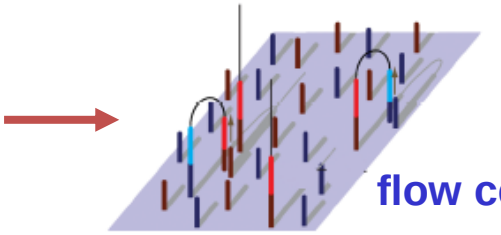






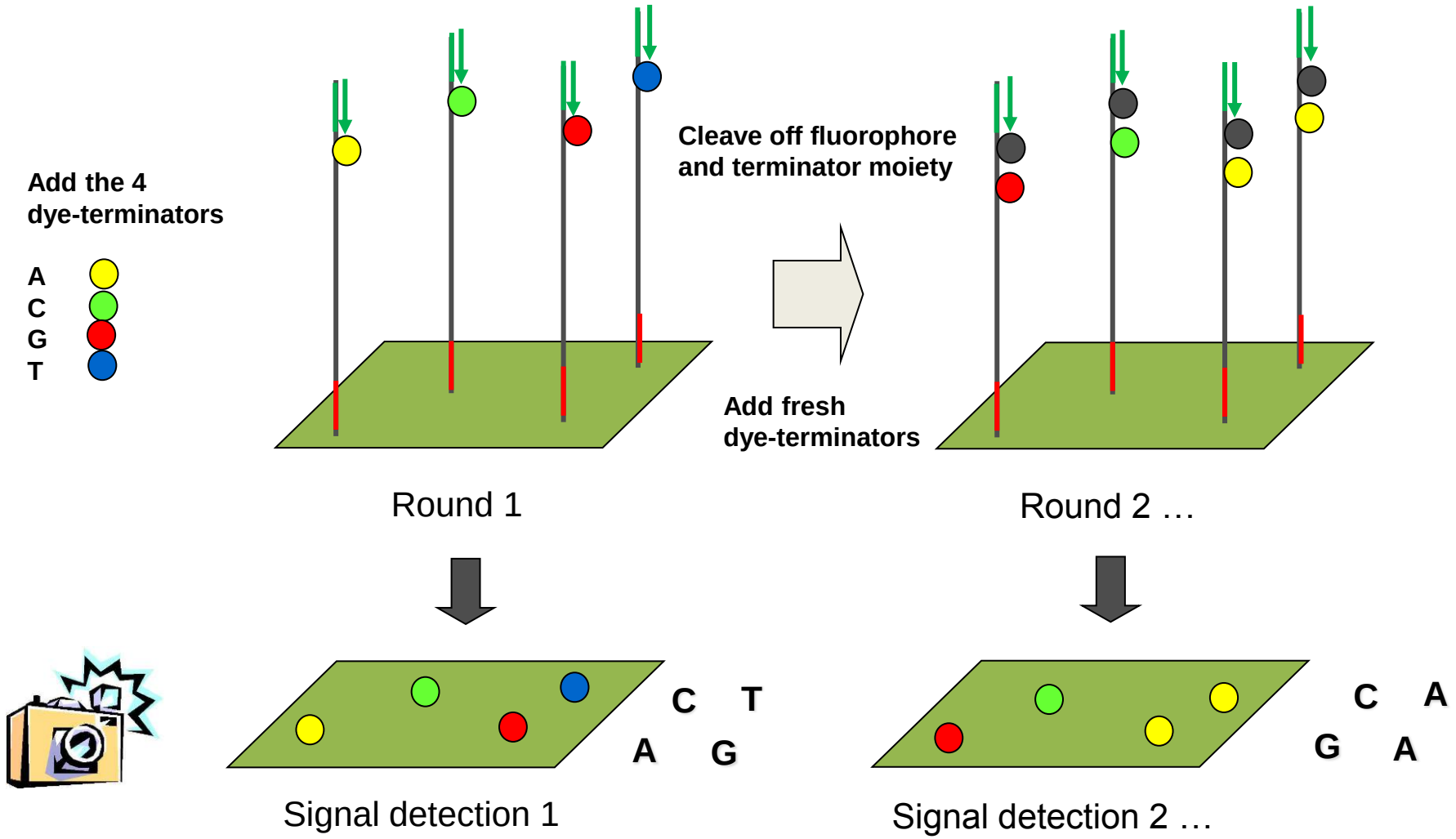
e.g. 200-400 bp





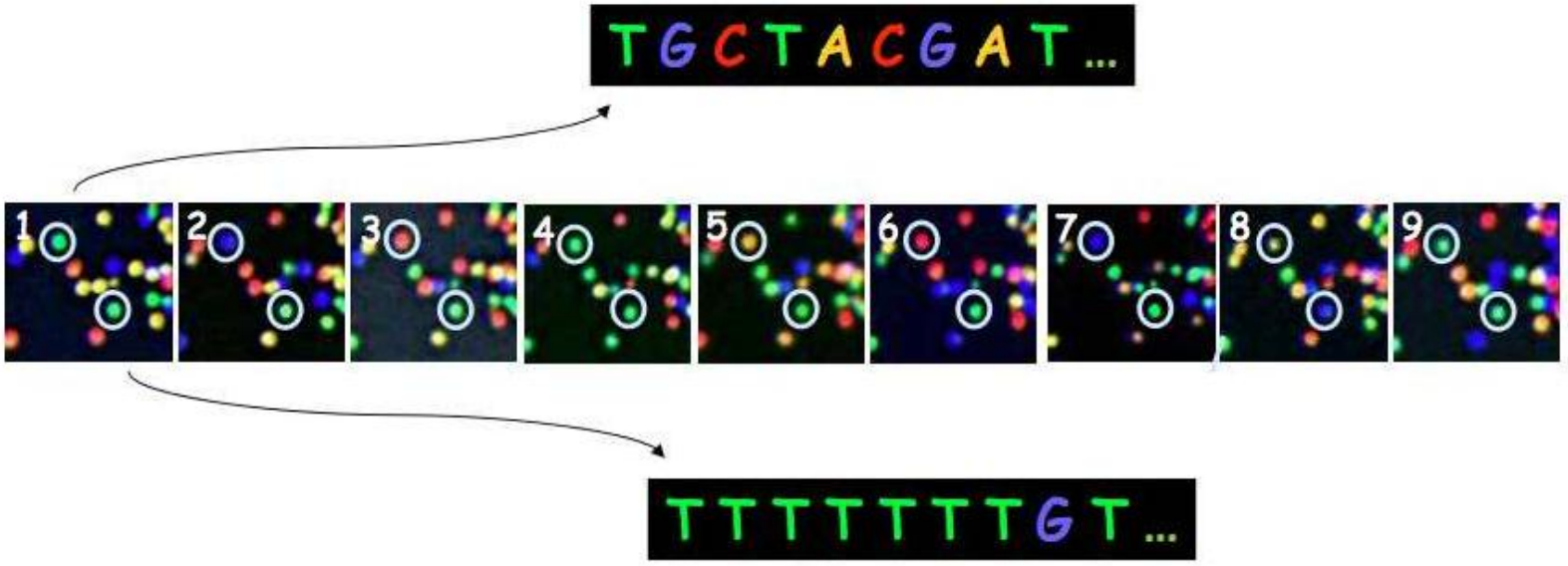
Illumina Sequencing

Sequencing-by-synthesis using reversible dye-terminators



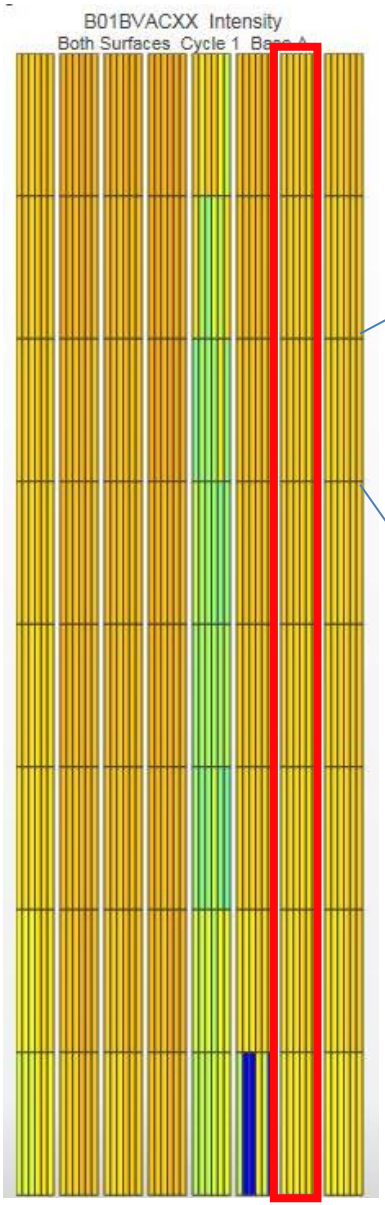
Illumina Sequencing

9 cycles shown, two cluster position marked...

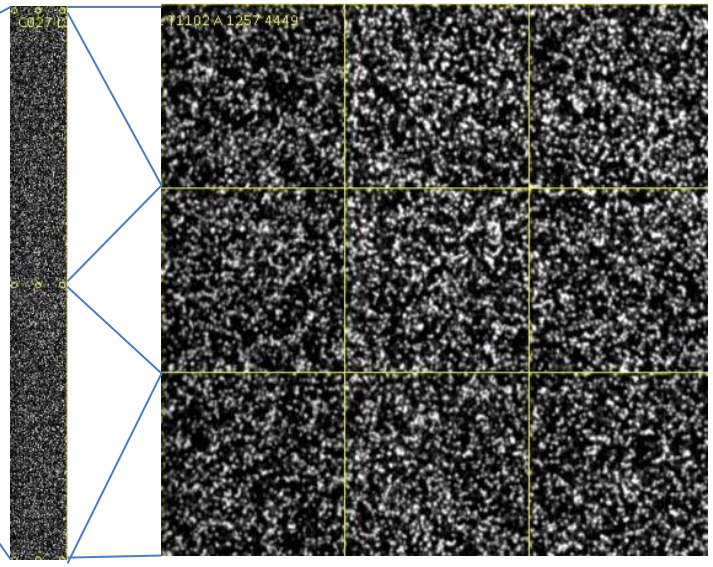


The sequence of each cluster is determined by consecutive images.

Illumina Sequencing



1 lane = on average 200-250 million reads



Illumina Sequencing

HiSeq

Run time	6-12 days
Read length	100+100 bp
Yield/lane	40 Gbp

MiSeq

Run time	1-2 days
Read length	250+250 bp
Yield/lane	5 Gbp

NGS Technology Summary

Platform	Year	Sequencing Method	Amplification	Detection	Features
454	2005	Pyro-sequencing	Emulsion PCR	Light	First NGS
Illumina	2007	Synthesis	Bridge PCR	Light	90% of 2012 Market
SOLiD	2008	Ligation	Emulsion PCR	Light	Lowest Error Rate
Ion Torrent	2010	Synthesis	Emulsion PCR	Hydrogen Ion	Semiconductor Chip
Pacific Biosciences	2010	Synthesis	None = Single Molecule	Light	Longest Reads
Oxford Nanopore	2014?	Nanopore	None = Single Molecule	Electrical Conductivity	"Run Until" Sequencing

Modified from Travis C. Glenn. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759-769

NGS Error Rates

2013 NGS Field Guide. www.molecularecologist.com

Platform	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)	Notes
3730xl (capillary)	Substitution	0.1-1	0.1-1	
454	Indel	1	1	
Illumina	Substitution	~0.1	~0.1	≥ 85% of reads
SOLiD	A-T bias	~5	≤0.1	2x-3x sequencing
Ion Torrent	Indel	~1	~1	0.5-2.4%
PacBio RS	Indel	~13	≤1	consensus of 3 reads
Oxford Nanopore	Deletion	≥4	4	press release only

Deep Sequencing Compensates for Errors

How much will it cost?

2013 NGS Field Guide. www.molecularecologist.com

Instrument	Reagent Cost/run[a]	Reagent Cost/MB	Minimum Unit Cost (% run)
ABI 3730xl (capillary)	$144	$2308	$6 (1%)
454 GS Jr. Titanium	$1100	$22	$1500 (100%)
PacBio RS	$300-1700	$2-17	$500 (100%)
Ion Torrent – 316 chip	$739	$1.20	$1000 (100%)
Illumina MiSeq	$1070	$0.14	$1400 (100%)
Ion Torrent – Proton I	$1050	$0.09	? (100%)
SOLiD – 5500xl	$10,503	<$0.07	$2000 (12%)
Illumina HiSeq 2000	$23,470	≥$0.04	$2400 (6%)

Includes all stages of sample prep. for a single sample (= library prep through sequencing, except capillary = sequencing only).

Multiplexing

Addition of a unique sequence identifier (barcode or index) allowing multiple samples to be run together on a single flow cell lane.

Internal indexes
(Cronn et al 2008)

External indexes
(extra round of sequencing)

- Illumina (12-96)
- Nextflex (48-96)