7.36 / 20.390 / 6.802

7.91 / 20.490 / 6.874 / HST.506

Lecture #2

C. Burge

Feb. 6, 2014

Local Alignment (BLAST) and Statistics

Topic 1 Info

- CB office hours
  - after lectures (Tues/Thurs 2:30-3:00) - 68-271A (except today)
  - or by request
- Slides will generally be posted (PDF) by 12:15 pm on day of lecture*

- Overview slide has blue background - readings for upcoming lectures are listed at bottom of overview slide

- Review slides will have purple background

- PS1 is posted

- PS2 will be posted soon. Look at the programming problem

The two Python tutorials will be:

Friday, Feb. 7 3:00 – 4:00 PM
Monday, Feb. 10 4:00 – 5:00 PM

* If printing, to save paper, can print multiple slides per page using Acrobat Reader. Under “Page scaling:” choose “Multiple pages per sheet”

For those reg'd for grad versions of course

- Please email by Tuesday Feb 11th:

Name
Email
G/U Program_name
Background (1 sentence)
Comp Bio Interests (1 sentence or a few keywords)

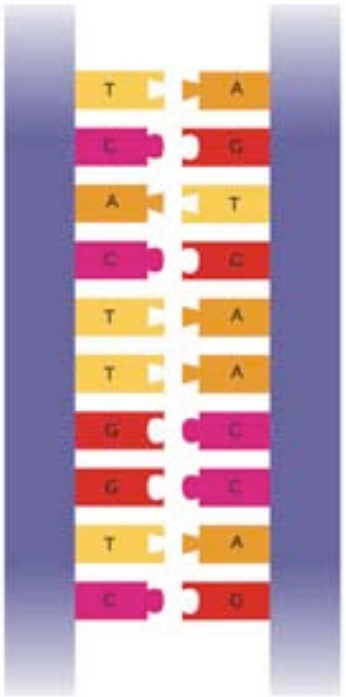
for posting

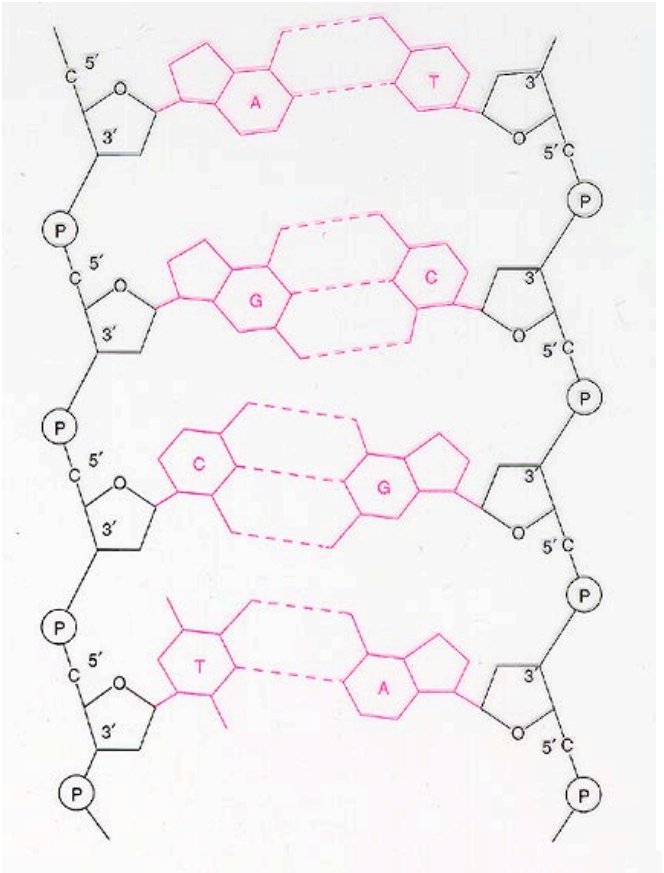
Local Alignment (BLAST) and Statistics

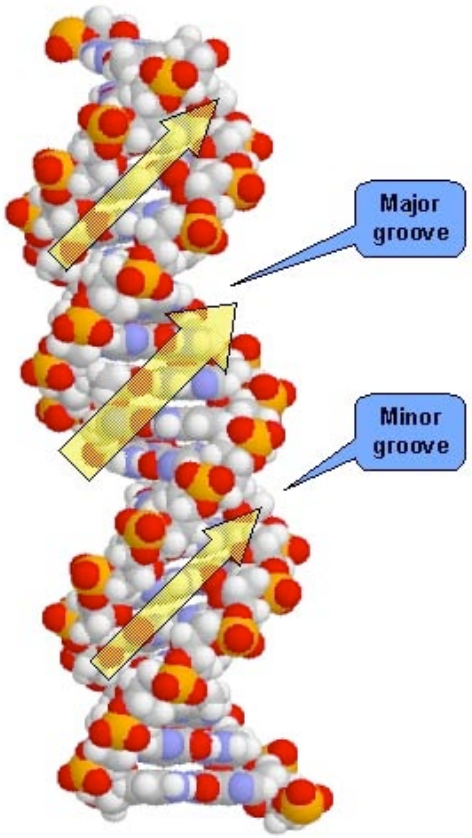
- Sequencing
  - Conventional
  - 2nd generation
- Local Alignment:
  - a simple BLAST-like algorithm
  - Statistics of matching
  - Target frequencies and mismatch penalties for nucleotide alignments

Background for next two lectures: Z&B Ch. 4 & 5

1D, 2D and 3D Representations of DNA

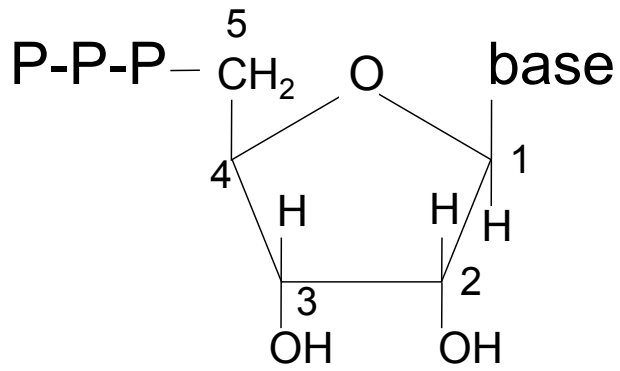


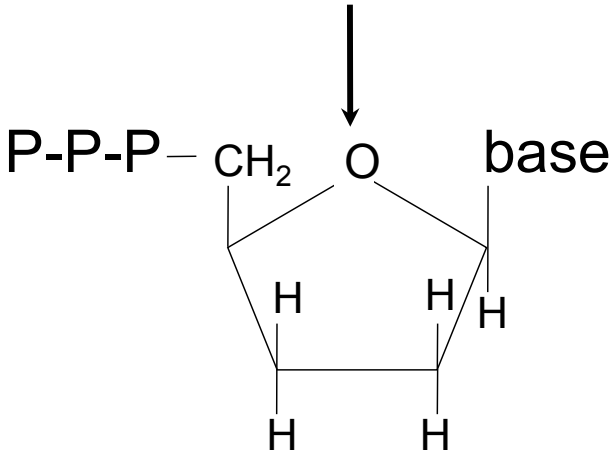


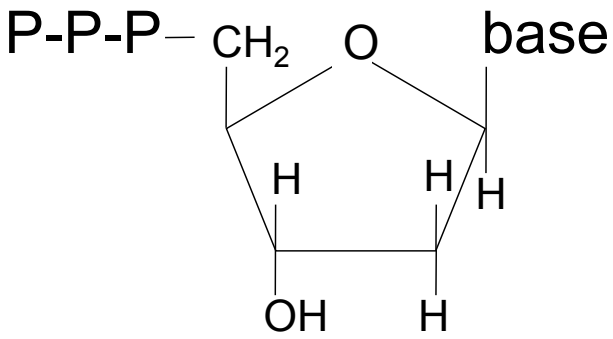


Types of Nucleotides

- ribonucleotide
- deoxyribonucleotide
- dideoxyribonucleotide

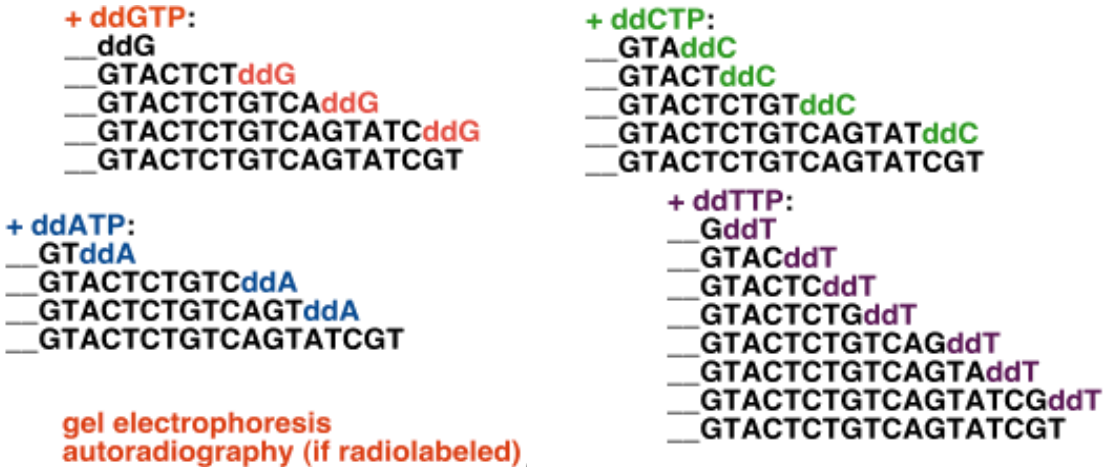






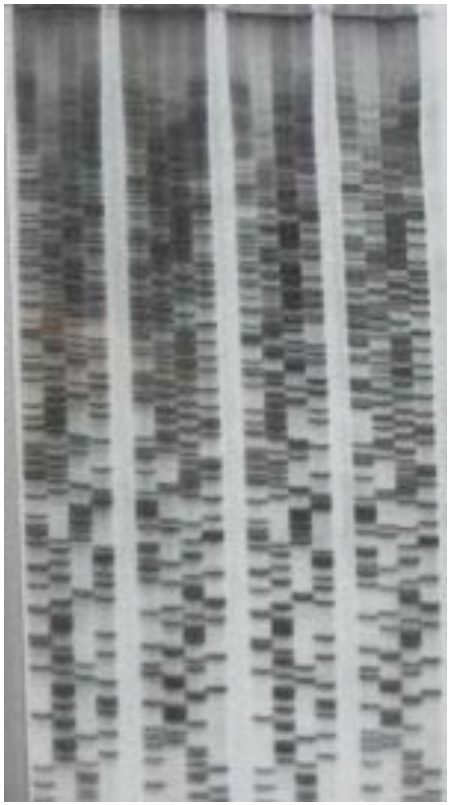
Primer
5' NNN
3' NNNCATGAGACAGTC…
Template

Sanger sequencing method



Evolution of Sequencing Technologies

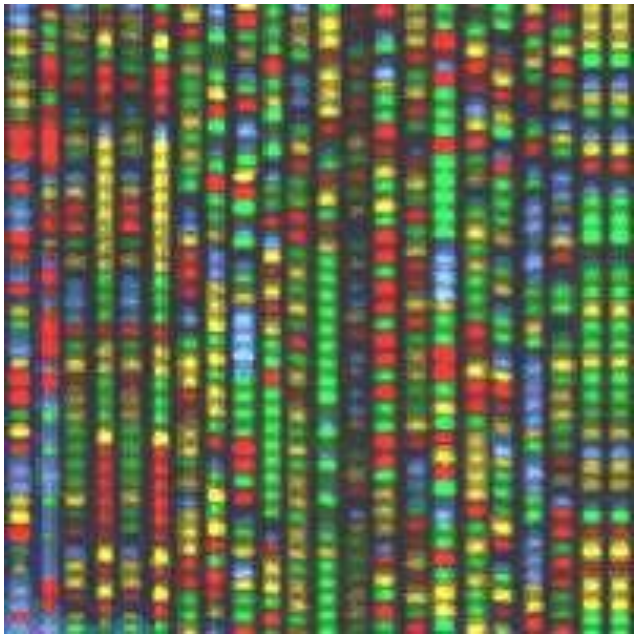
- Traditional Sanger / chain termination sequencing (70s, 80s, 90s)



A T C G ...

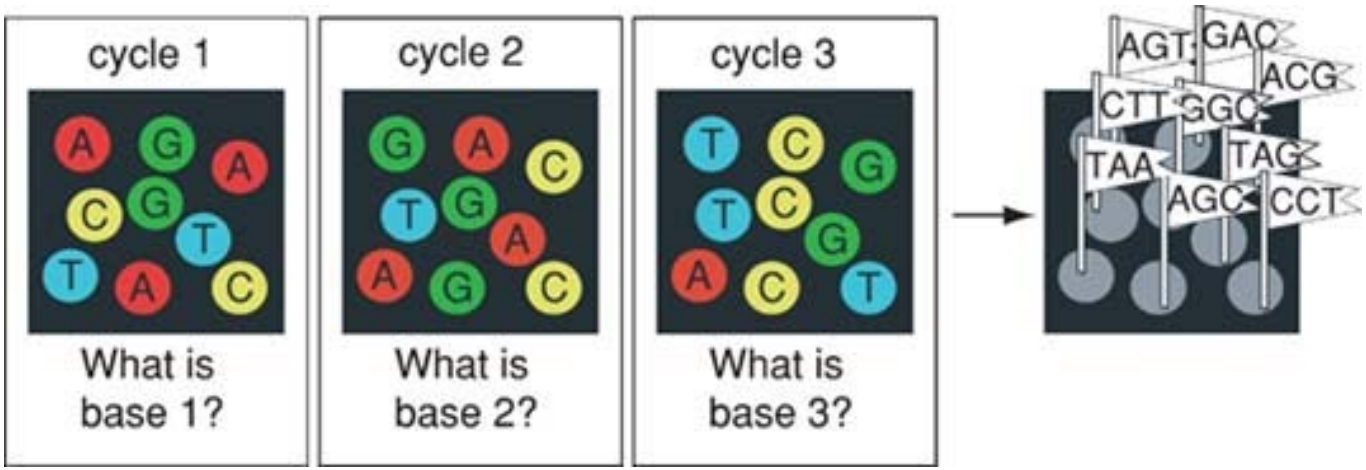
- Large polyacrylamide gels, radiolabeled DNA, 4 lanes per read

- Fluorescent-based / dye terminator sequencing (90s - present)



- Capillary electrophoresis, fluorescent tags for each base, 1 lane per read

‘Next Generation’ Sequencing Technologies



Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Shendure, Jay, and Hanlee Ji. "Next-generation DNA Sequencing." *Nature Biotechnology* 26, no. 10 (2008): 1135-45.

A variety of technologies. Differ in aspects of:

- DNA template
- Modified nucleotides used
- Imaging / image analysis

Metzker NRG 2010

Comparison of Platforms

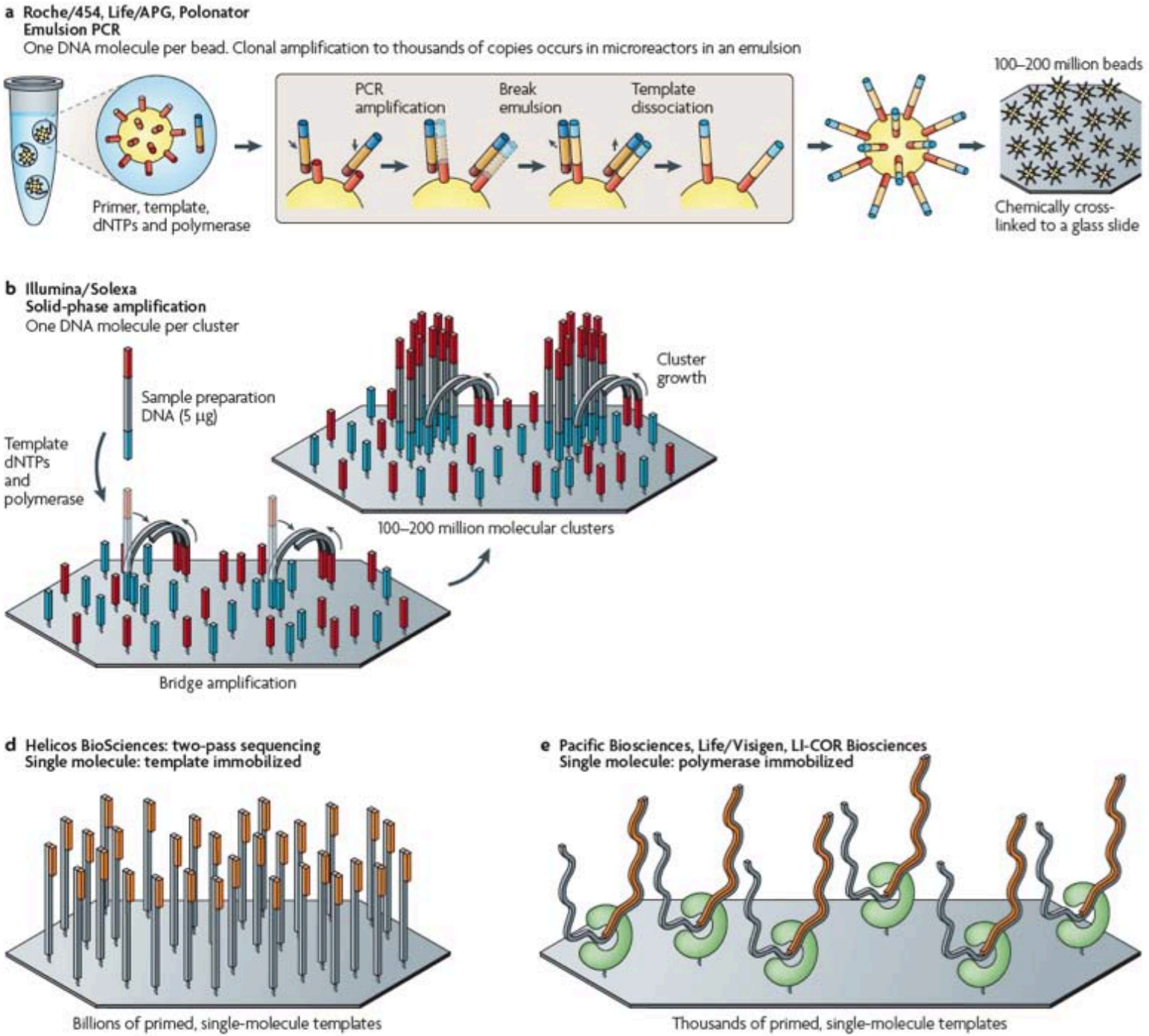
Table 1 | **Comparison of next-generation sequencing platforms**

Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Metzker NRG 2010

Next-gen Sequencing: Templates

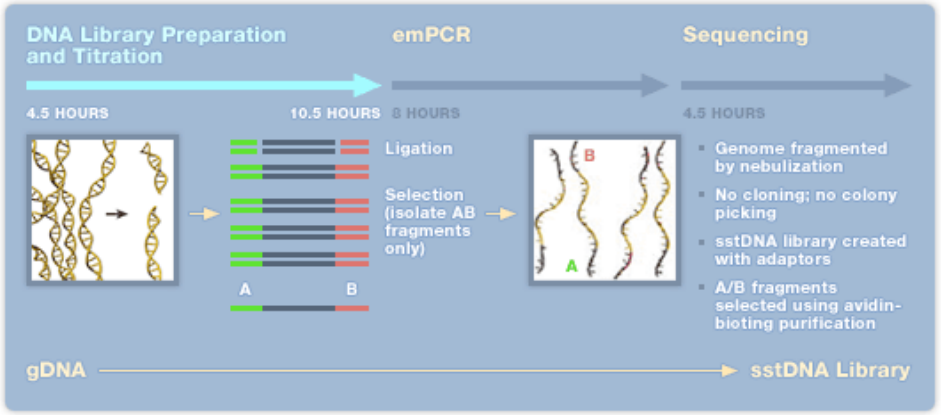


Metzker NRG 2010

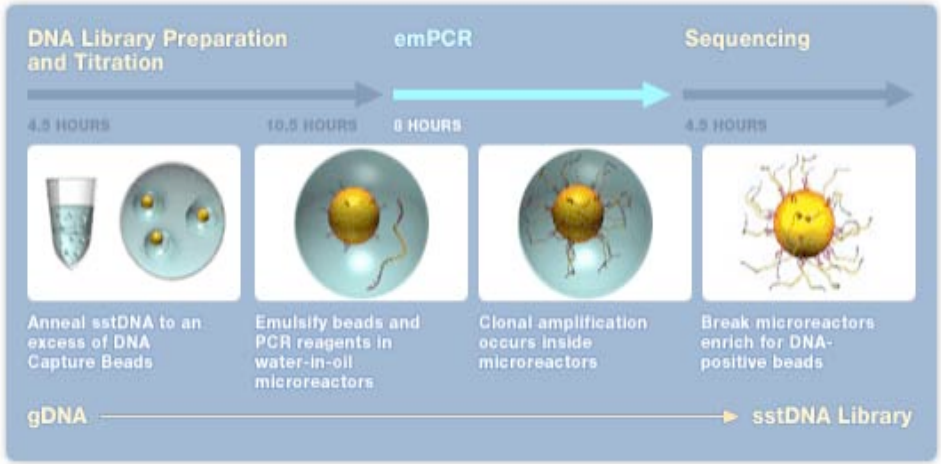
Source: Metzker, Michael L. "Sequencing Technologies—The Next Generation." *Nature Reviews Genetics* 11, no. 1 (2009): 31-46.

Example: bead-based pyrosequencing 1

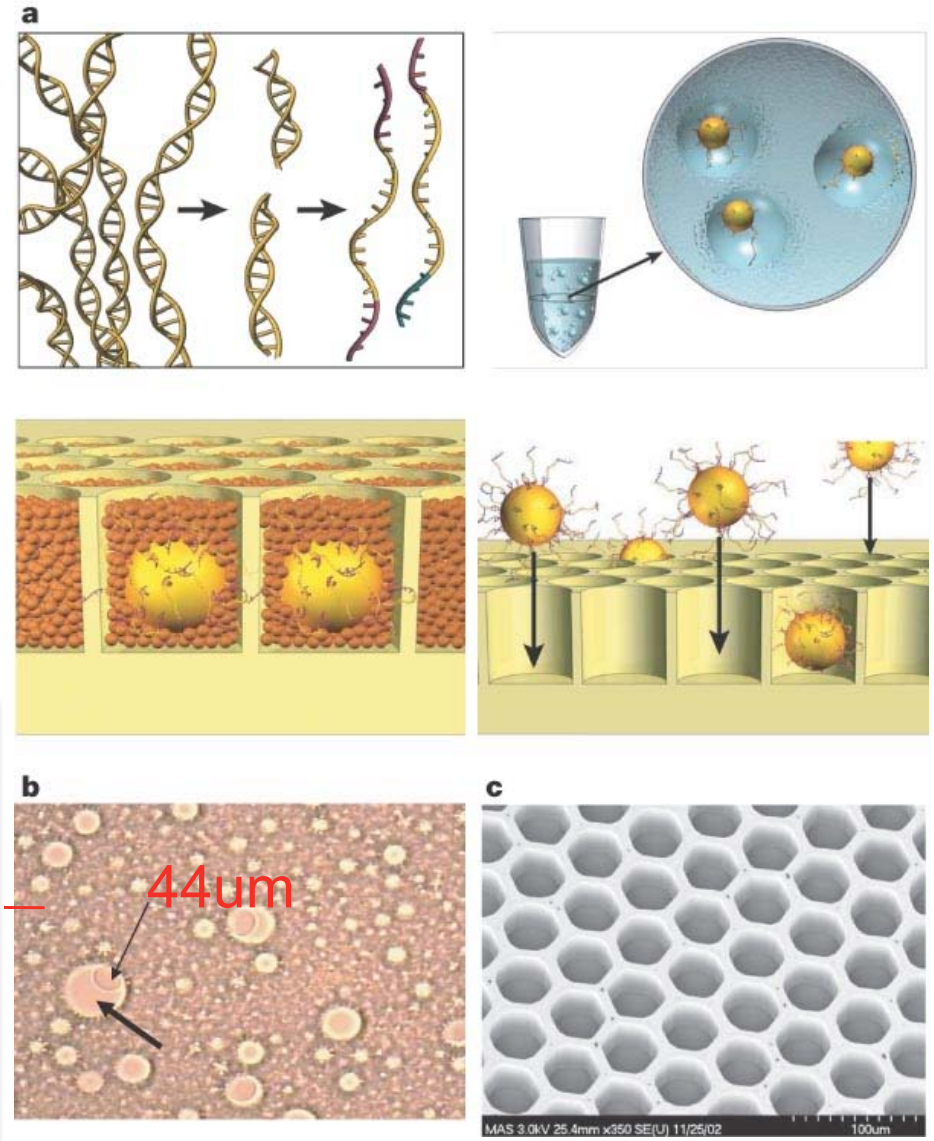
Step 1. DNA Library Preparation



Step 2. PCR



Courtesy of 454 Life Sciences, A Roche Company. Used with permission.



Margulies et al. *Nature* 2005

Courtesy of Macmillan Publishers Limited. Used with permission. Source: Margulies, Marcel, Michael Egholm, et al. "Genome Sequencing in Microfabricated High-density Picolitre Reactors." *Nature* 437, no. 7057 (2005): 376-80.

Bead-based pyrosequencing 2

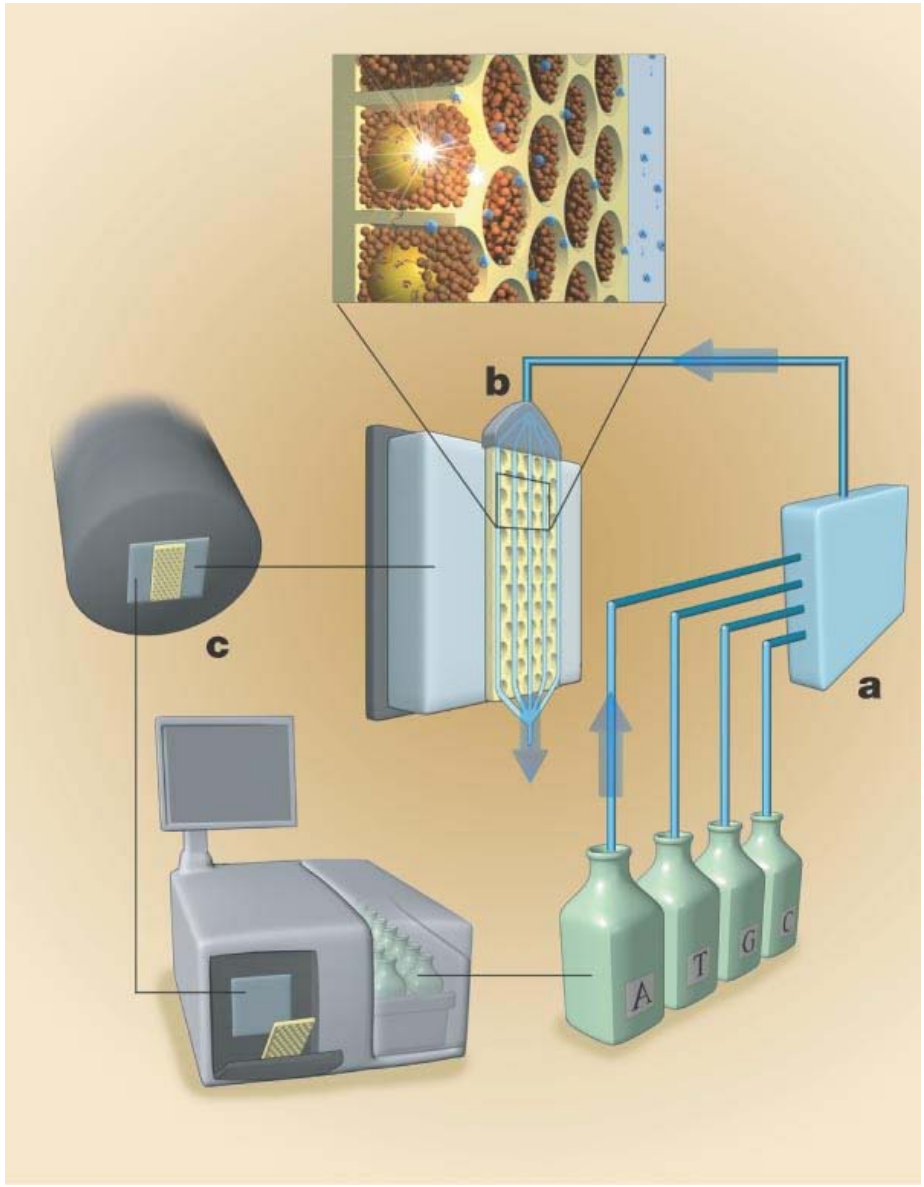
Generates ~400+ nt per well

x 1,000,000 wells with single bead

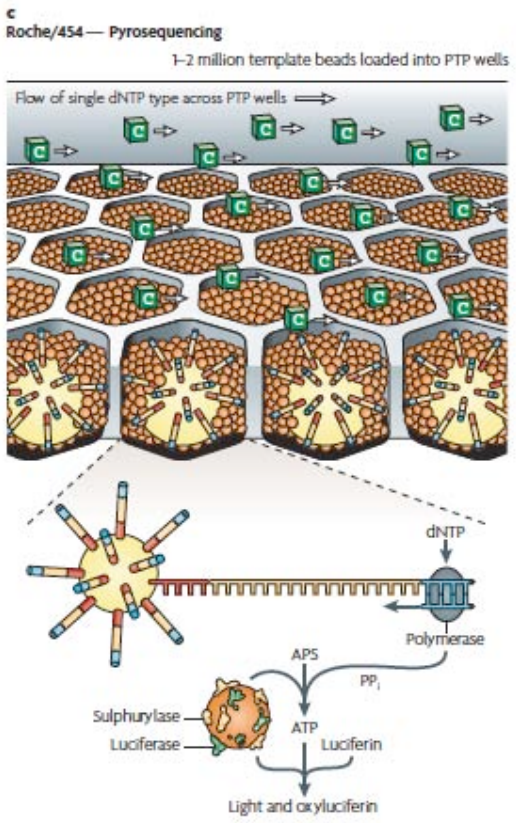
= ~400 Mbp per run

(10 hours, several $K)

(stats updated since publication)



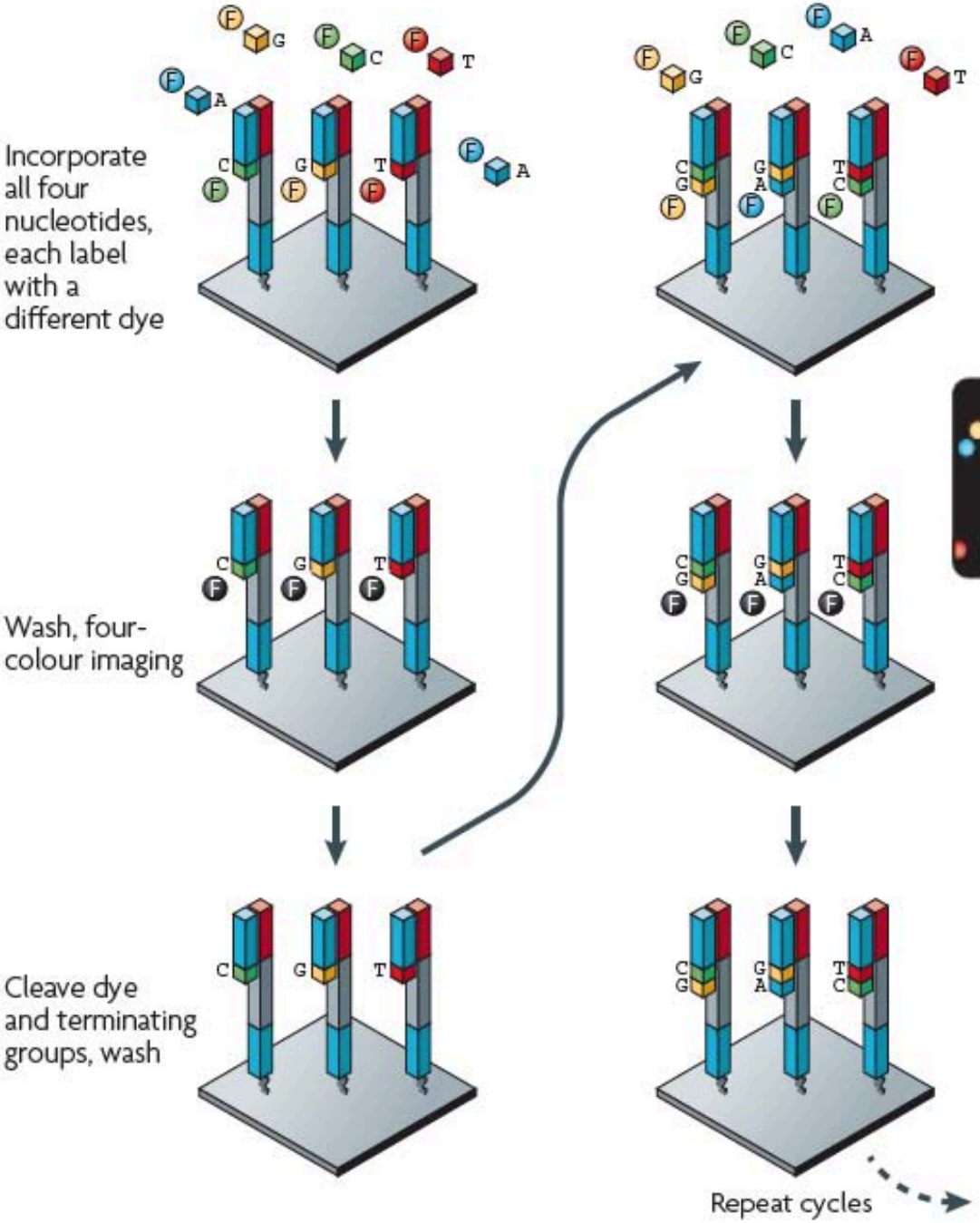
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Margulies, Marcel, Michael Egholm, et al. "Genome Sequencing in Microfabricated High-density Picolitre Reactors." *Nature* 437, no. 7057 (2005): 376-80.

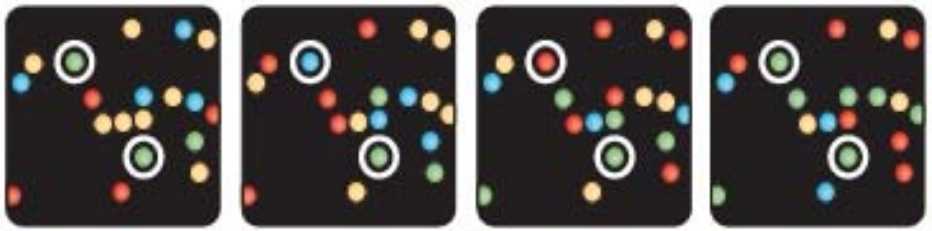


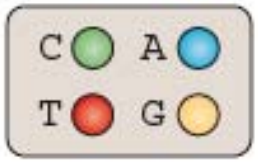
Margulies et al. *Nature* 2005

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Metzker, Michael L. "Sequencing Technologies—The Next Generation." *Nature Reviews Genetics* 11, no. 1 (2009): 31-46.

Illumina/Solexa sequencing





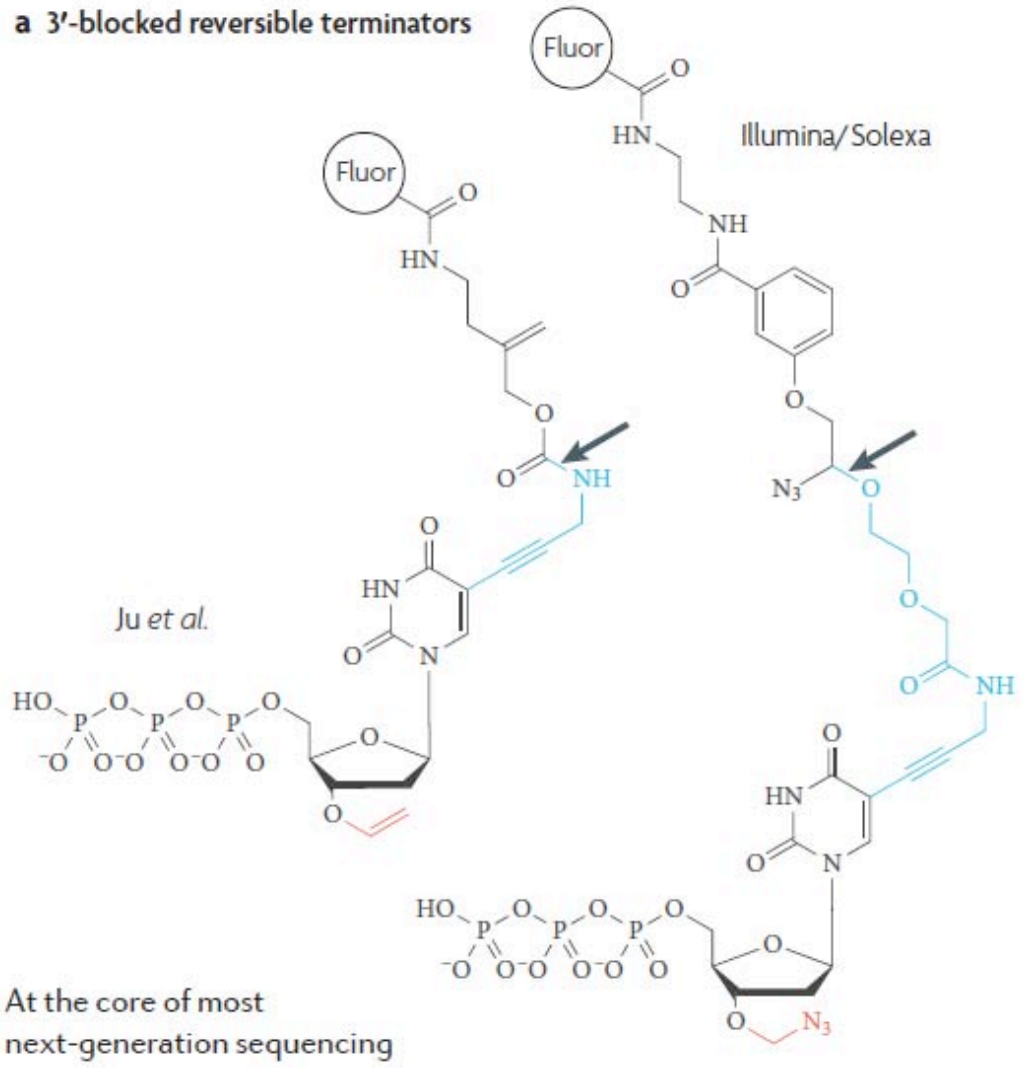


Top: CATC
Bottom: CCCC

Metzker NRG 2010

Source: Metzker, Michael L. "Sequencing Technologies—The Next Generation." *Nature Reviews Genetics* 11, no. 1 (2009): 31-46.

Illumina/Solexa sequencing



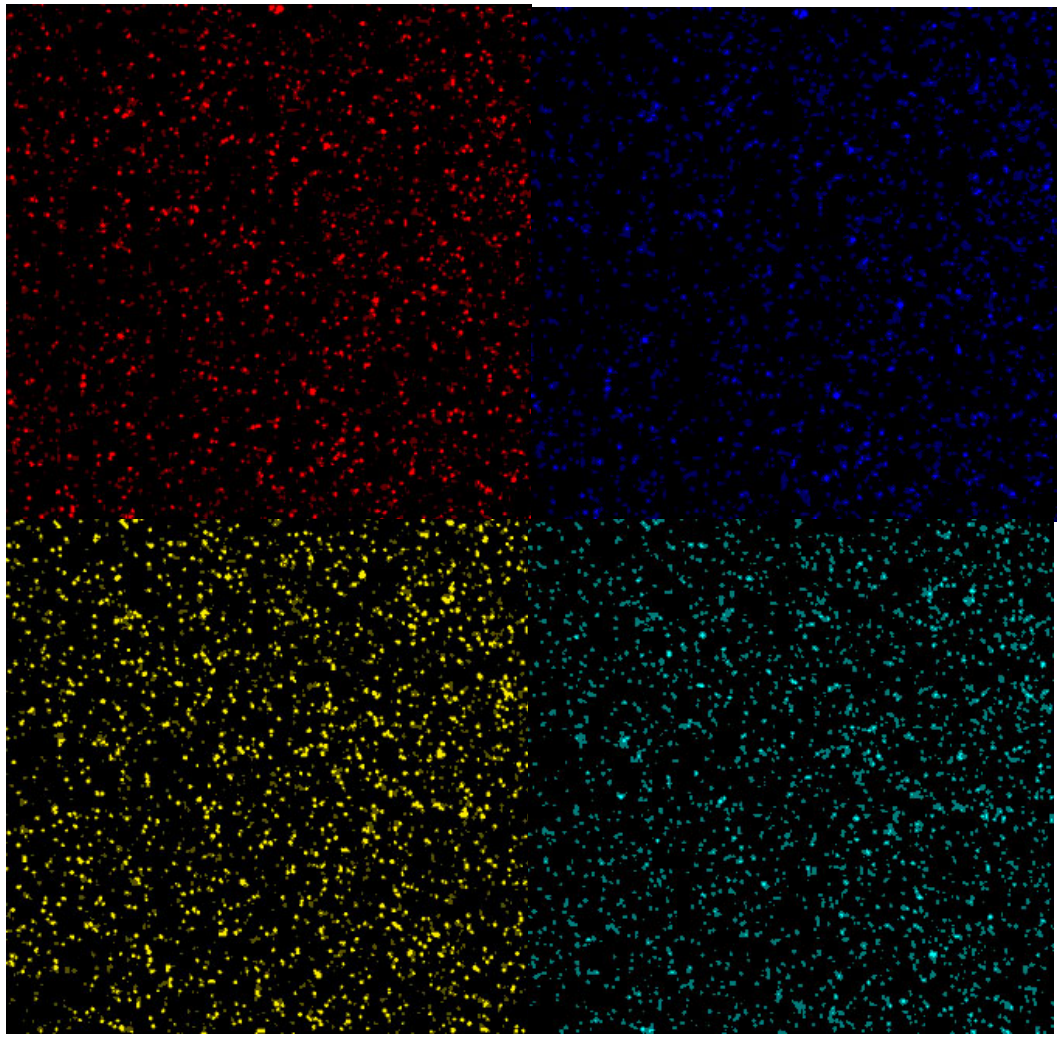
Metzker NRG 2010

Source: Metzker, Michael L. "Sequencing Technologies—The Next Generation." *Nature Reviews Genetics* 11, no. 1 (2009): 31-46.

Illumina Sequencing Images

A channel

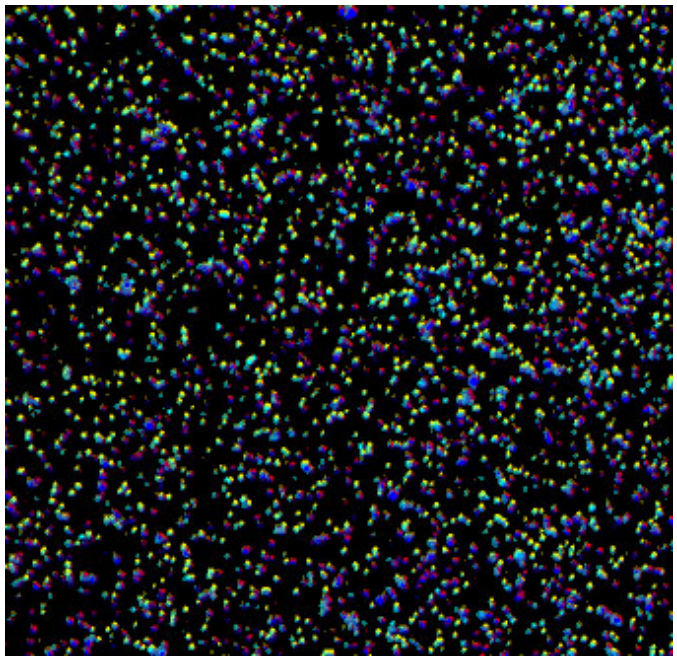
C channel



G channel

T channel

Merge



1/4 of one tile (0.03% of a flow cell, GA2)

Example 2:

Illumina cluster-based sequencing

Current throughput (HiSeq 2000 instrument)
one flow cell = 8 lanes, several days, ~$20K in reagents

8 lanes x $2x10^8$ reads/lane x 100 bp / read = ~160×10^9 bp

Can double throughput by:
- PE sequencing
- Sequencing 2 flow cells at once

Why Align Sequences?

Which alignments are significant?

Local alignment:

find shorter stretches of high similarity
don’t require alignment of whole sequence

DNA Sequence Alignment I: Motivation

You are studying a recently discovered human non-coding RNA.

You search it against the mouse genome using BLASTN (N for nucleotide) and obtain the following alignment:

```
Q: 1   ttgacctagatgagatgtcgttcacttttactcaggtacagaaaa 45
       |||| |||||||||||| | |||||||||||| || |||||||||
S: 403 ttgatctagatgagatgccattcacttttactgagctacagaaaa 447
```

Is this alignment significant?
Is this likely to represent a homologous RNA?

How to find alignments?

DNA Sequence Alignment II

Determining significance of nucleotide local alignments

```
Q: 1   ttgacctagatgagatgtcgttcacttttactcaggtacagaaaa 45
       |||| |||||||||||| | |||||||||||| || |||||||||
S: 403 ttgatctagatgagatgccattcacttttactgagctacagaaaa 447
```

Identify high scoring segments whose score S exceeds a cutoff x using a **local alignment** algorithm (e.g., BLAST)

Scores follow an extreme value (aka Gumbel) distribution:

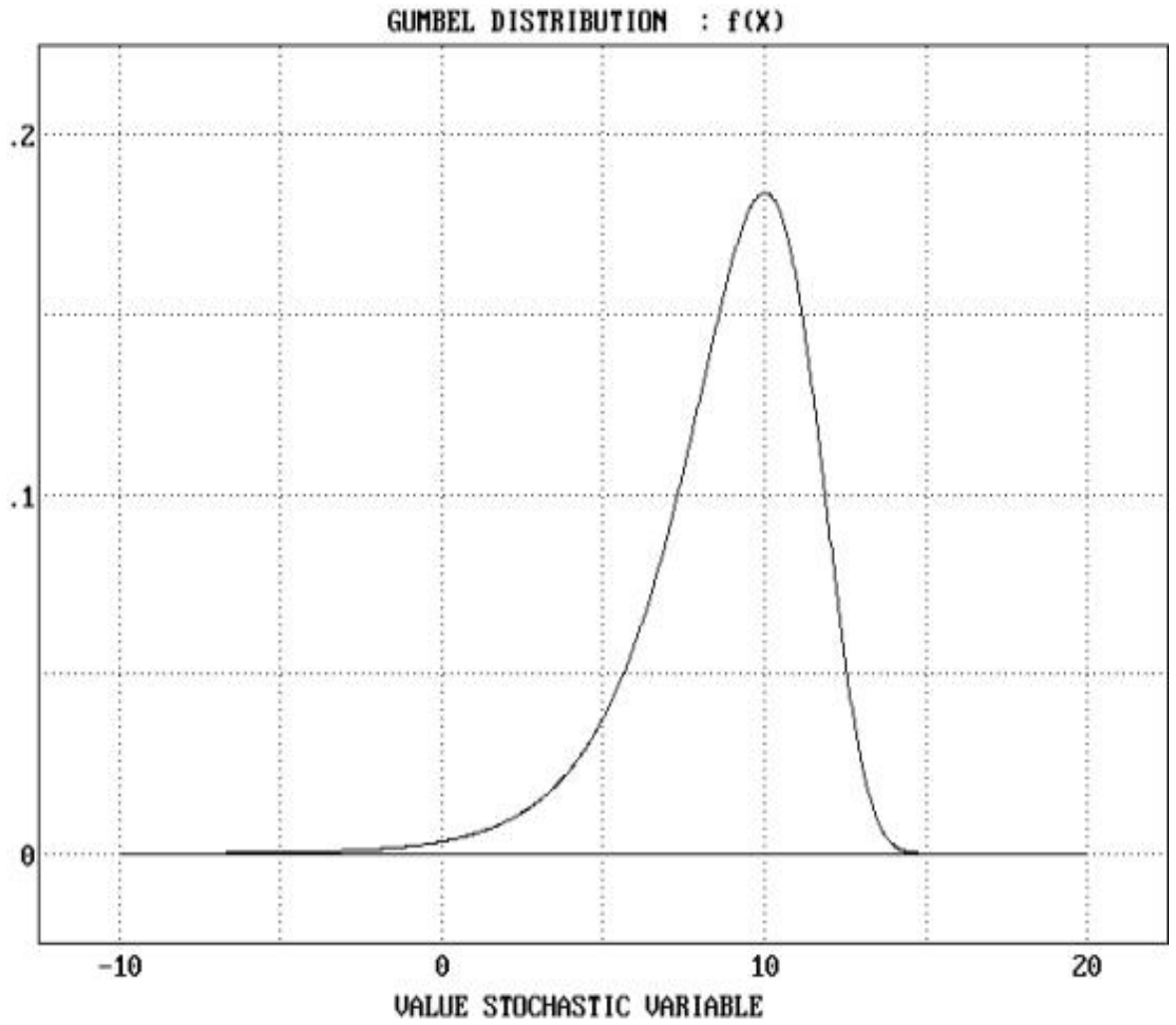
$$P(S > x) = 1 - \exp[-KMN\, e^{-\lambda x}]$$

For sequences/databases of length M, N where K, λ are positive parameters that depend on the score matrix and the composition of the sequences being compared

Conditions: expected score is negative, but positive scores possible

Karlin & Altschul 1990

Extreme Value (Gumbel) Distribution



DNA Sequence Alignment III

How is λ related to the score matrix?

λ is the unique positive solution to the equation*:

$$\sum_{i,j} p_i r_j e^{\lambda s_{ij}} = 1$$

p_i = freq. of nt i in query, r_j = freq. of nt j in subject

s_{ij} = score for aligning an i,j pair

What kind of an equation is this? (transcendental)

What would happen to λ if we doubled all the scores? (reduced by half)

What does this tell us about the nature of λ? (scaling factor)

*Karlin & Altschul, 1990

DNA Sequence Alignment IV

What scoring matrix to use for DNA?

Usually use simple match-mismatch matrices:

$s_{i,j}$:

i j:	A	C	G	T
A	1	m	m	m
C	m	1	m	m
G	m	m	1	m
T	m	m	m	1

m = "mismatch penalty" (must be negative)

When would you use a mismatch penalty of: -1 -3 -5 ?

DNA Sequence Alignment V

Figuring out how to choose the mismatch penalty …

"Target frequencies"* : $q_{ij} = p_i p_j e^{\lambda s_{ij}}$ $\Rightarrow$ $\boxed{s_{ij} = \ln(q_{ij} / p_i p_j)/\lambda}$

q_{ij} are nt pair frequencies expected in high-scoring matches

If you want to find regions with R% identities:

$r = R/100$ $q_{ii} = r/4$ $q_{ij} = (1-r)/12$ $(i \neq j)$ Set $s_{ii} = 1$

Then $m = s_{ij} = s_{ij}/s_{ii} = (\ln(q_{ij} / p_i p_j)/\lambda) / (\ln(q_{ii} / p_i p_i)/\lambda)$ $(i \neq j)$

$\Rightarrow$ $\boxed{m = \ln(4(1-r)/3)/\ln(4r)}$ (Assuming all $p_i, p_j = 1/4$, $1/4 < r < 1$)

*Karlin & Altschul, 1990

DNA Sequence Alignment VI

Optimal mismatch penalty m for given target identity fraction r

$$m = \ln(4(1-r)/3)/\ln(4r)$$

Examples:

r	0.75	0.95	0.99
m	-1	-2	-3

r = expected fraction of identities in high-scoring BLAST hits

7.91J / 20.490J / 20.390J / 7.36J / 6.802J / 6.874J / HST.506J Foundations of Computational and Systems Biology
Spring 2014