

CSEP 527  
Computational Biology  
Spring 2016

3: BLAST, Alignment score significance;  
PCR and DNA sequencing

# Outline

Scoring

BLAST

Weekly Bio Interlude: PCR & Sequencing

# Significance of alignment scores



[http://dericbownds.net/uploaded\\_images/god\\_face2.jpg](http://dericbownds.net/uploaded_images/god_face2.jpg)

# Significance of Alignments

Is “42” a good score?

*Compared to what?*

Usual approach: compared to a specific “null model”,  
such as “random sequences”



# Brief Review of Probability

# random variables

---

Discrete random variable: takes values in a finite or countable set, e.g.

$X \in \{1, 2, \dots, 6\}$  with equal probability

$X$  is positive integer  $i$  with probability  $2^{-i}$

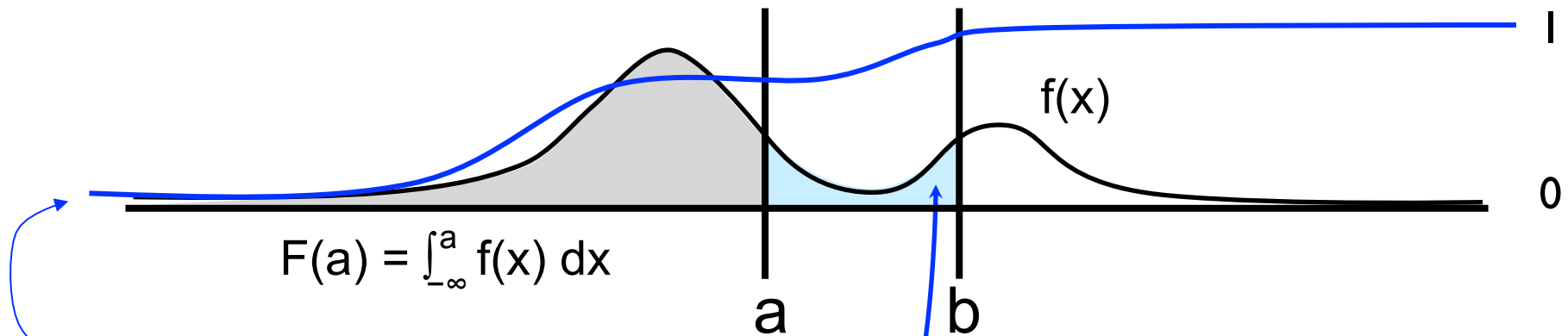
Continuous random variable: takes values in an uncountable set, e.g.

$X$  is the weight of a random person (a real number)

$X$  is a randomly selected point inside a unit square

$X$  is the waiting time until the next packet arrives at the server

$f(x)$  : the *probability density function* (or simply “density”)



$$F(a) = \int_{-\infty}^a f(x) dx$$

$P(X < a) = F(a)$ : the *cumulative distribution function*

$$P(a < X < b) = F(b) - F(a)$$

Need  $f(x) \geq 0$ ,  $\int_{-\infty}^{+\infty} f(x) dx (= F(+\infty)) = 1$

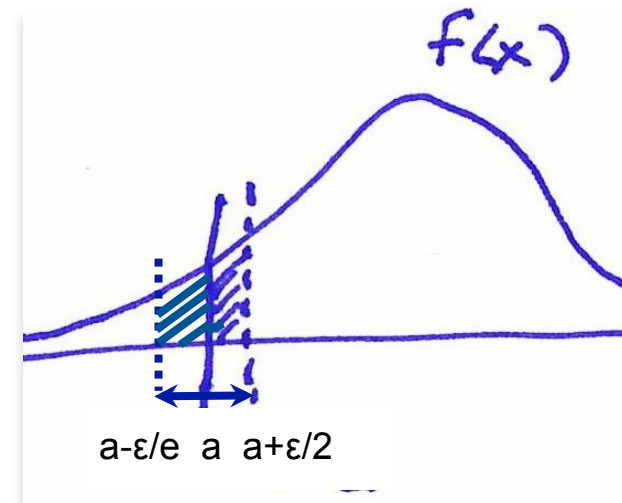
A key relationship:

$$f(x) = \frac{d}{dx} F(x), \text{ since } F(a) = \int_{-\infty}^a f(x) dx,$$

Densities are *not* probabilities; e.g. may be  $> 1$

$$P(X = a) = 0$$

$$\begin{aligned} P(a - \varepsilon/2 \leq X \leq a + \varepsilon/2) &= \\ &F(a + \varepsilon/2) - F(a - \varepsilon/2) \\ &\approx \varepsilon \cdot f(a) \end{aligned}$$



I.e., the probability that a continuous random variable falls *at* a specified point is *zero*

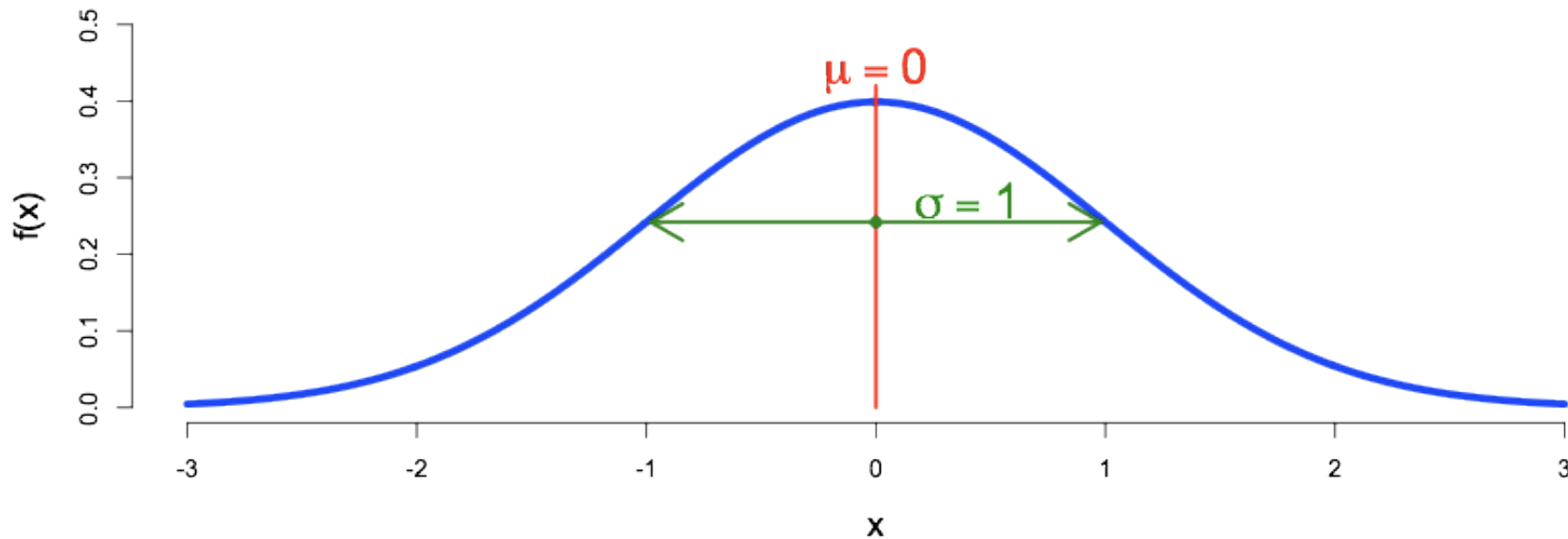
The probability that it falls *near* that point is proportional to the density; in a large random sample, expect more samples where density is higher (hence the name “density”).

X is a normal (aka Gaussian) random variable  $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$$

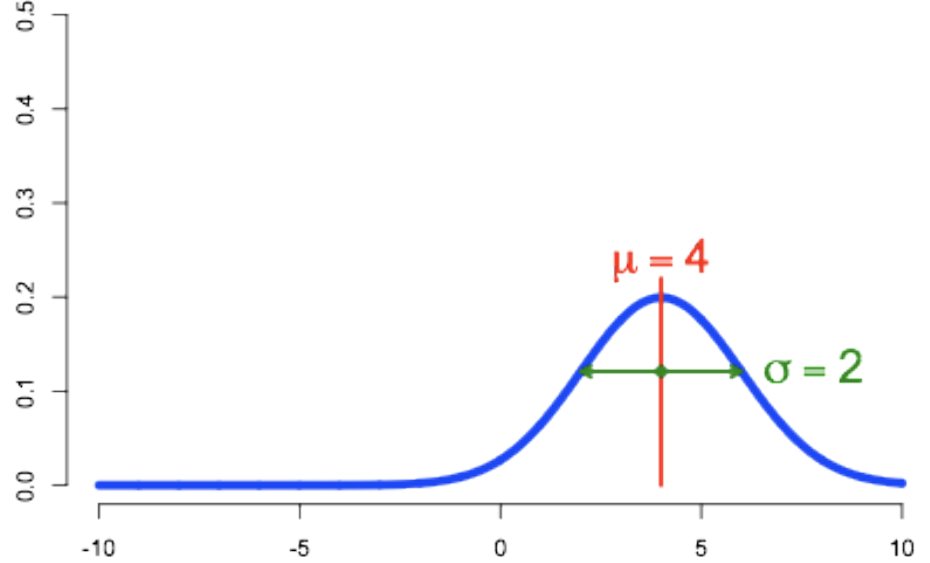
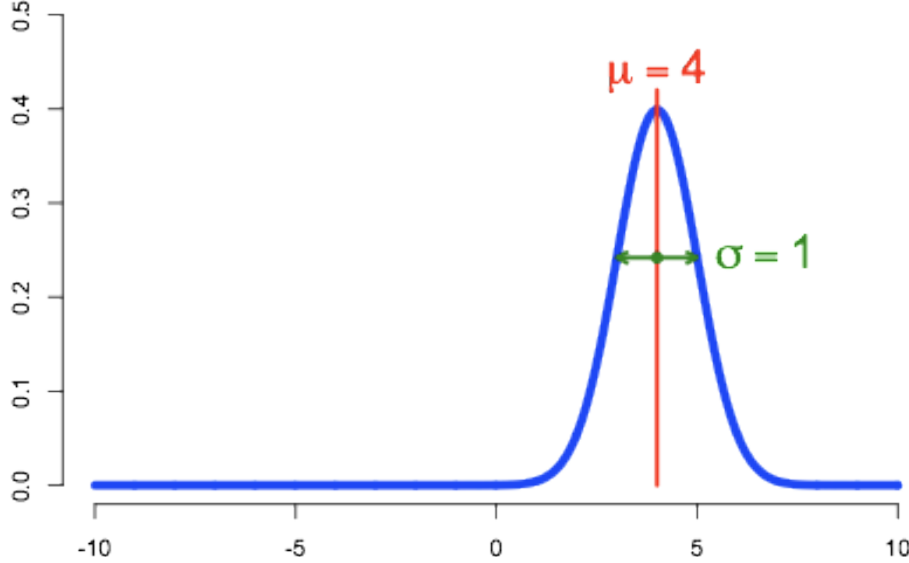
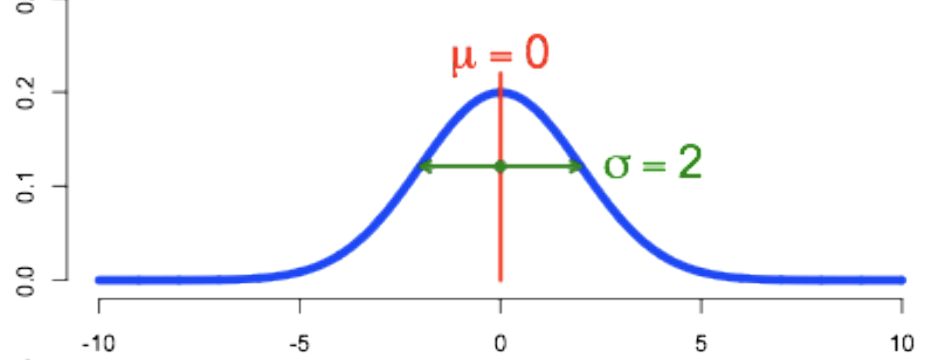
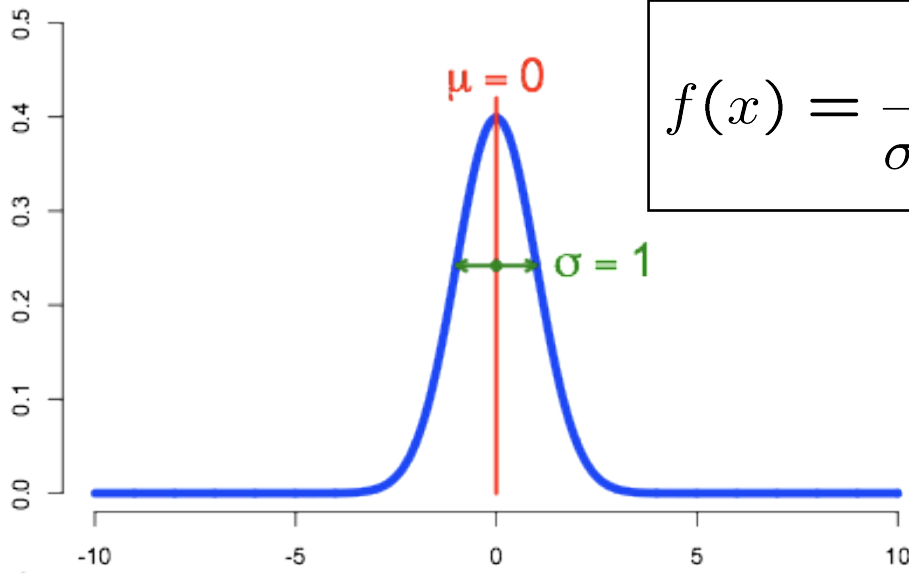
$$E[X] = \mu \quad \text{Var}[X] = \sigma^2$$

### The Standard Normal Density Function



# changing $\mu$ , $\sigma$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$



10

density at  $\mu$  is  $\approx .399/\sigma$

# Z-scores

$$Z = (X - \mu) / \sigma = (X - \text{mean}) / \text{standard deviation}$$

e.g.

$Z = +3$  means “3 standard deviations above the mean”

Applicable to *any* distribution, and gives a rough sense of how usual/unusual the datum is.

If  $X$  is normal( $\mu, \sigma^2$ ) then  $Z$  is normal(0,1), and you can easily calculate (or look up in a table) just *how* unusual

E.g., if normal,  $P(Z\text{-score} \geq +3) \approx 0.001$

# Central Limit Theorem

If a random variable  $X$  is the sum of many independent random variables, then  $X$  will be approximately normally distributed.



# Hypothesis Tests and P-values

# Hypothesis Tests

Competing models might explain some data

E.g., you've flipped a coin 5 times, seeing HHHTH

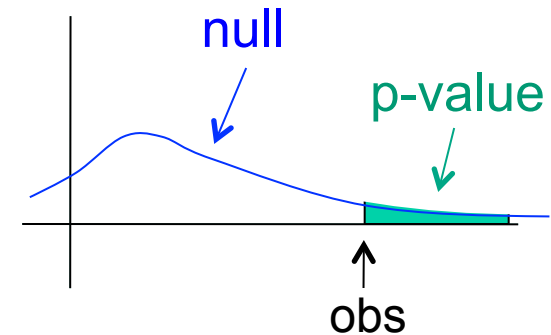
Model 0 (The "null" model):  $P(H) = 1/2$

Model 1 (The "alternate" model):  $P(H) = 2/3$

Which is right?

A possible decision rule: reject the null if you see 4 or more heads in 5 tries

# p-values



The *p-value* of such a test is the probability, assuming that the null model is true, of seeing data as extreme or more extreme than what you actually observed

E.g., we observed 4 heads; p-value is prob of seeing 4 or 5 heads in 5 tosses of a fair coin

Why interesting? It measures *probability that we would be making a mistake in rejecting null*.

Can analytically find p-value for simple problems like coins; often turn to simulation/permutation tests (introduced earlier) or to approximation (coming soon) for more complex situations

Usual scientific convention is to reject null only if p-value is  $< 0.05$ ; sometimes demand  $p \ll 0.05$  (esp. if estimates are inaccurate, and/or big data)

# p-values: controversial

p-values are commonly misused/misinterpreted

Most importantly, it is *not* the probability that the null is true, nor the 1 minus the prob that the alternate is true

*Nevertheless, p-values are very widely used*

Many resources, e.g.:

- <https://en.wikipedia.org/wiki/P-value>
- <http://blog.minitab.com/blog/adventures-in-statistics/how-to-correctly-interpret-p-values>
- <http://www.dummies.com/how-to/content/what-a-pvalue-tells-you-about-statistical-data.html>

# Alignment Scores

## Overall Alignment Significance, I Empirical p-values (via randomization)

You just searched with x, found “good” score for x:y  
Generate N random “y-like” sequences (say  $N = 10^3 - 10^6$ )  
Align x to each & score

If k of them have better score than alignment of x to y,  
then the (empirical) probability of a chance alignment as  
good as your observed x:y alignment is  $(k+1)/(N+1)$

e.g., if 0 of 99 are better, you can say “estimated  $p < .01$ ”

How to generate “random y-like” seqs? Scores depend on:

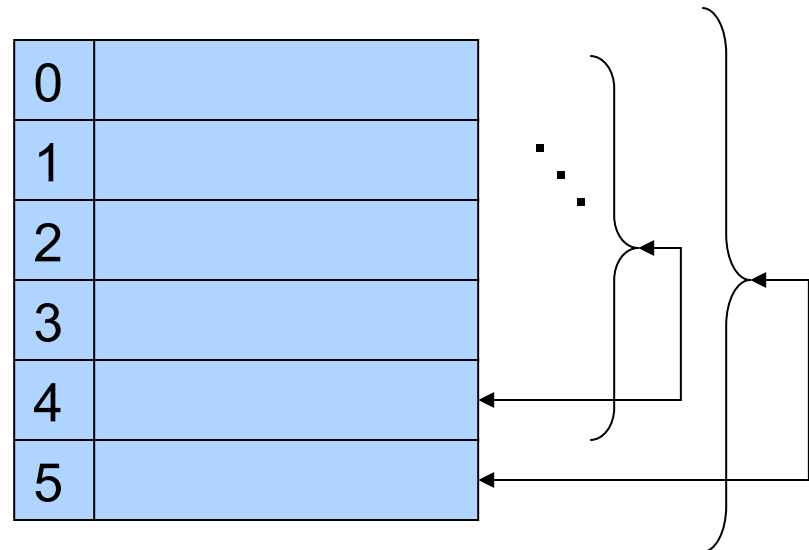
Length, so use same length as y

Sequence composition, so uniform  $1/20$  or  $1/4$  is a bad idea; even  
background  $p_i$  can be dangerous

Better idea: *permute* y N times

# Generating Random Permutations

```
for (i = n-1; i > 0; i--){  
    j = random(0..i);  
    swap X[i] <-> X[j];  
}
```



All  $n!$  permutations of the original data equally likely: A specific element will be last with prob  $1/n$ ; given that, a specific other element will be next-to-last with prob  $1/(n-1)$ , ...; overall:  $1/(n!)$

C.f. [http://en.wikipedia.org/wiki/Fisher–Yates\\_shuffle](http://en.wikipedia.org/wiki/Fisher–Yates_shuffle) and (for subtle way to go wrong) <http://www.codinghorror.com/blog/2007/12/the-danger-of-naivete.html>

# Permutation Pro/Con

## Pro:

Gives empirical p-values for alignments with characteristics like sequence of interest, e.g. residue frequencies

Largely free of modeling assumptions (e.g., ok for gapped...)

## Con:

Can be inaccurate if your method of generating random sequences is un-representative

E.g., probably better to preserve di-, tri-residue statistics and/or other higher-order characteristics, but increasingly hard to know exactly what to model & how

Slow

Especially if you want to assess low-probability p-values



# Theoretical Distribution of Alignment Scores?

A straw man: suppose I want a simple null model for alignment scores of, say MyoD versus random proteins of similar lengths. Consider this: Write letters of MyoD in one row; make a random alignment by filling 2<sup>nd</sup> row with random permutation of the other sequence plus gaps.

MELLSPPLR...

uv---wxyz...

Score for column 1 is a random number from the M row of BLOSUM 62 table, column 2 is random from E row, etc.

By central limit theorem, total score would be approximately normal

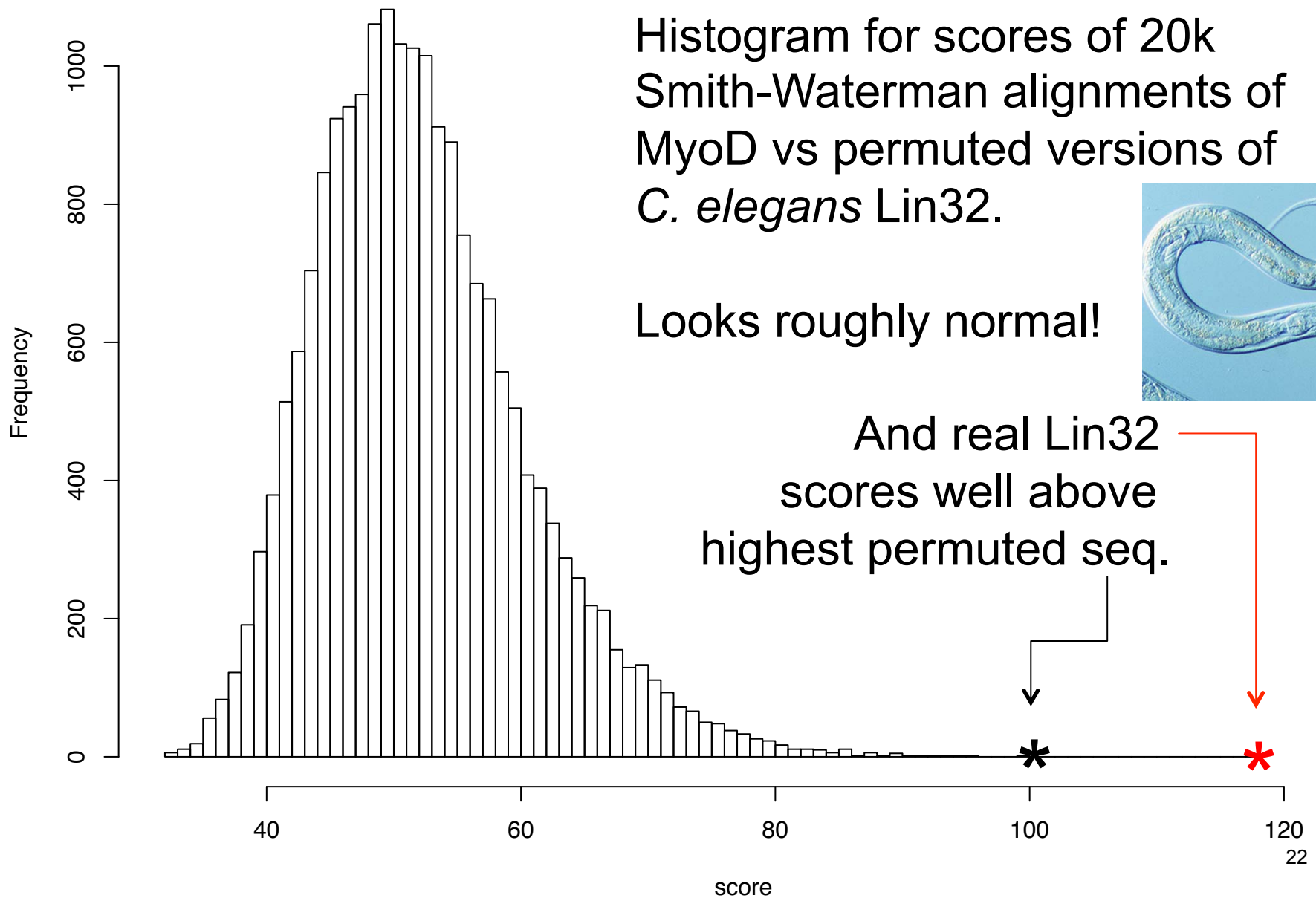
## Permutation Score Histogram vs Gaussian

Histogram for scores of 20k  
Smith-Waterman alignments of  
MyoD vs permuted versions of  
*C. elegans* Lin32.

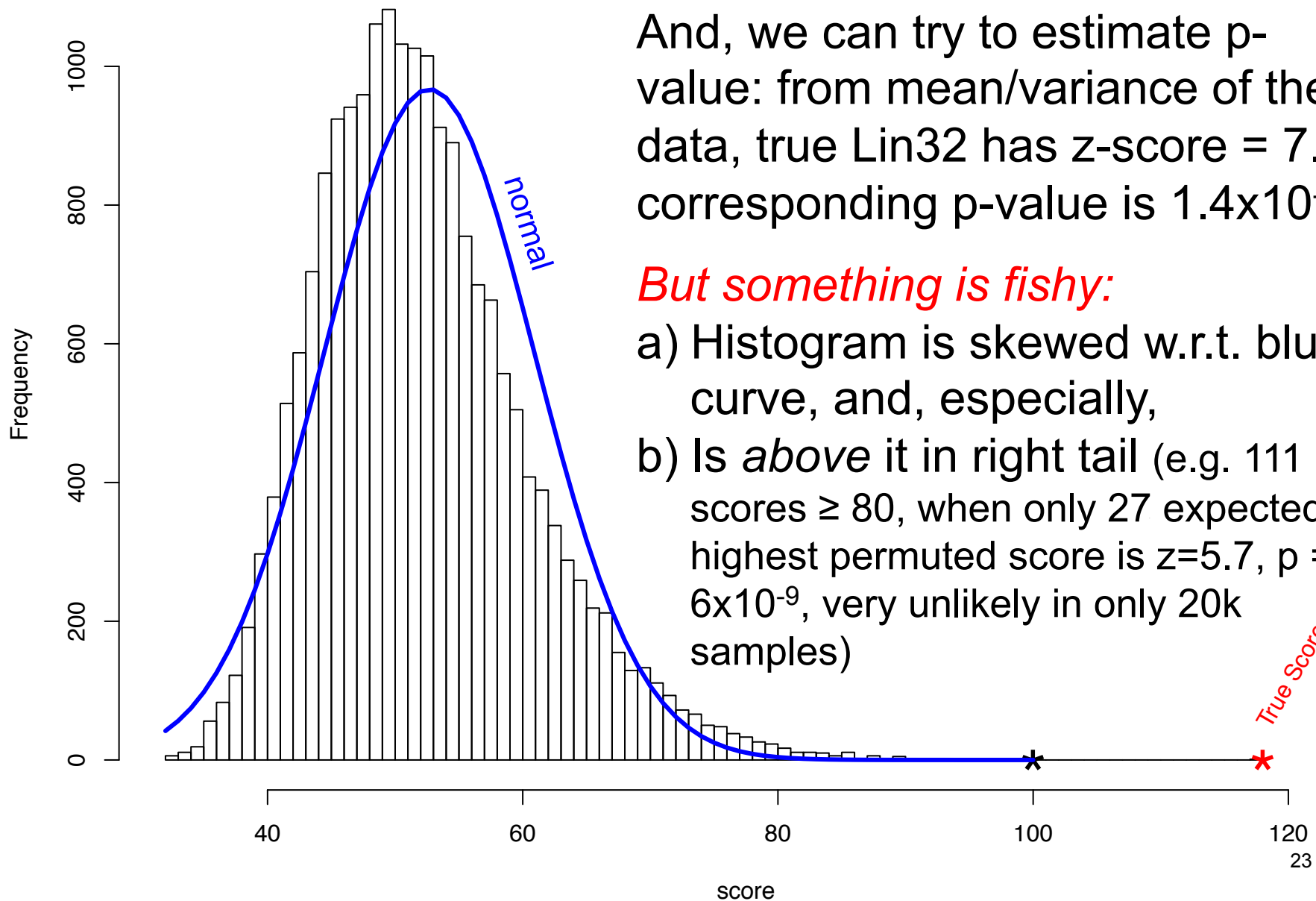
Looks roughly normal!



And real Lin32  
scores well above  
highest permuted seq.



## Permutation Score Histogram vs Gaussian



And, we can try to estimate p-value: from mean/variance of the data, true Lin32 has z-score = 7.9, corresponding p-value is  $1.4 \times 10^{-15}$ .

*But something is fishy:*

- Histogram is skewed w.r.t. blue curve, and, especially,
- Is *above* it in right tail (e.g. 111 scores  $\geq 80$ , when only 27 expected; highest permuted score is  $z=5.7$ ,  $p = 6 \times 10^{-9}$ , very unlikely in only 20k samples)

# Rethinking score distribution

Strawman above is ok: random permutation of letters & gaps *should* give normally distributed scores.

But S-W doesn't stop there; *it then slides the gaps around so as to maximize score, in effect taking the maximum over a huge number of alignments with same sequence but different gap placements, and furthermore trims ends to find the max local score.*

# Overall Alignment Significance, II

## A Theoretical Approach: EVD

Let  $X_i$ ,  $1 \leq i \leq N$ , be indep. random variables drawn from some (non-pathological) distribution

Q. what can you say about distribution of  $Y = \text{sum}\{X_i\}$ ?

A.  $Y$  is approximately *normally* distributed (central limit theorem)

Q. what can you say about distribution of  $Y = \text{max}\{X_i\}$ ?

A. it's approximately an *Extreme Value Distribution (EVD)*

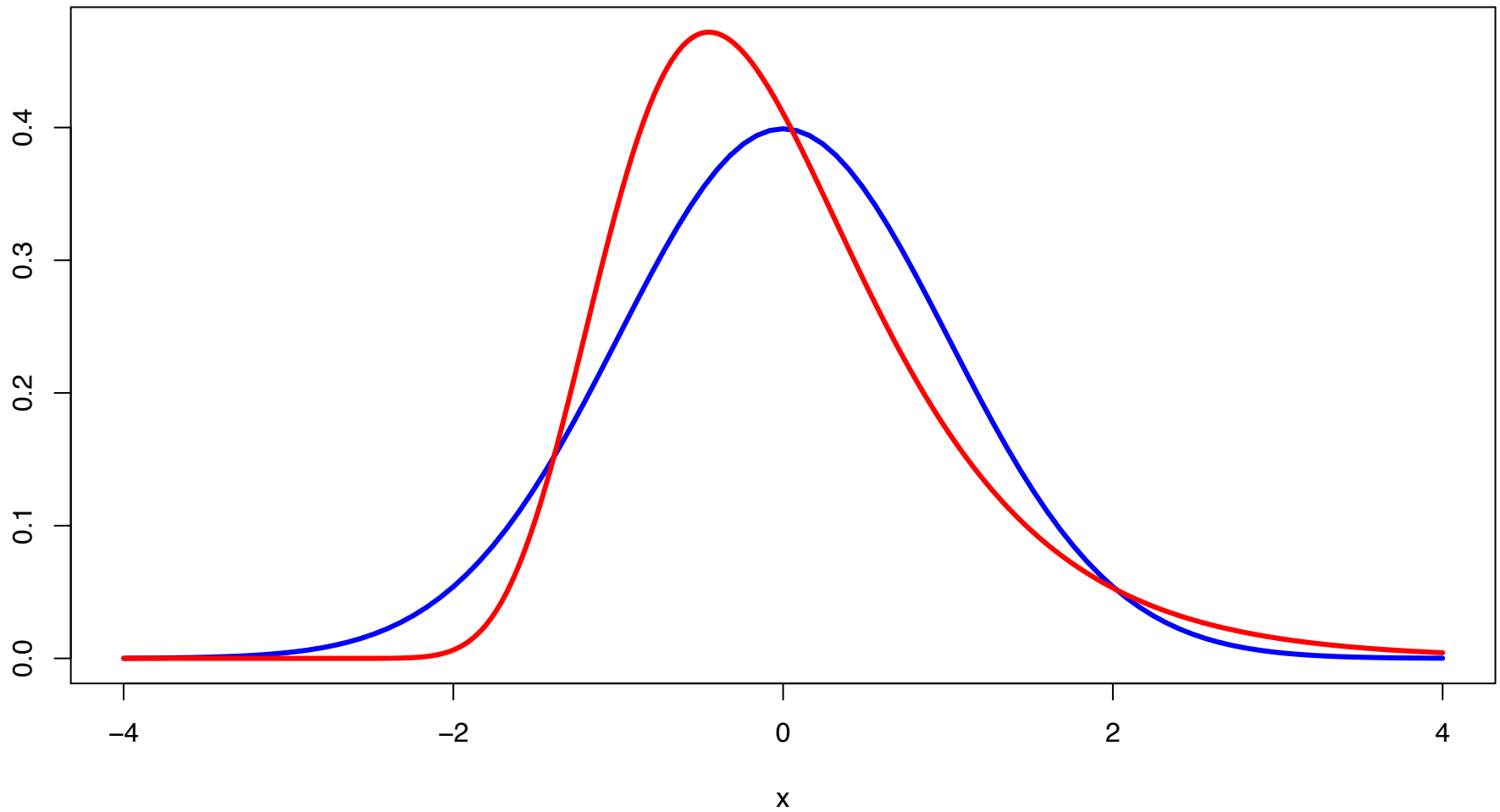
[one of only 3 kinds; for our purposes, the relevant one is:]

$$P(Y \leq z) \approx \exp(-KNe^{-\lambda(z-\mu)}) \quad (*)$$

For ungapped local alignment of seqs  $S, T$ ,  $N \sim |S|^*|T|$

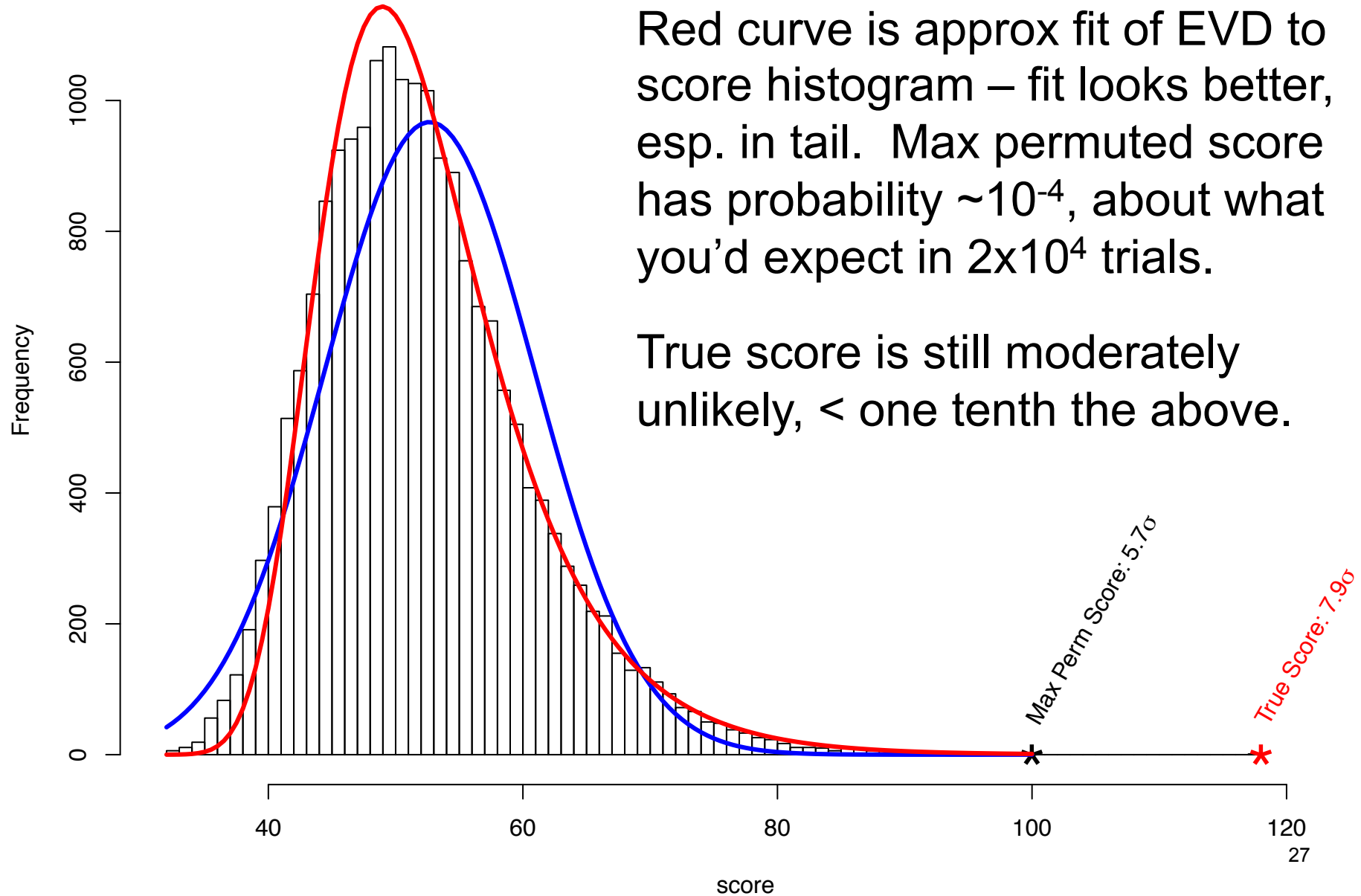
$\lambda, K$  depend on score table, and can be estimated by curve-fitting random scores to (\*), even with gaps. (cf. reading)

Normal (blue) / EVD (red)



Both mean 0, variance 1; EVD skewed & has “fat right tail”

## Permutation Score Histogram vs Gaussian



Red curve is approx fit of EVD to score histogram – fit looks better, esp. in tail. Max permuted score has probability  $\sim 10^{-4}$ , about what you'd expect in  $2 \times 10^4$  trials.

True score is still moderately unlikely,  $<$  one tenth the above.

# EVD Pro/Con

## Pro:

Gives p-values for alignment scores

## Con:

It's only approximate

You must estimate parameters

Theory may not apply. E.g., known to hold for ungapped local alignments (like BLAST seeds). It is NOT proven to hold for gapped alignments, although there is strong empirical support.



# Summary

Assessing statistical significance of alignment scores is crucial to practical applications

Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model (below)

For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments.

Permutation tests are a simple and broadly applicable (but brute force) alternative

# BLAST:

## Basic Local Alignment Search Tool

Altschul, Gish, Miller, Myers, Lipman, J Mol Biol 1990

*The* most widely used comp bio tool

Which is better: long mediocre match or a few nearby, short, strong matches with the same total score?

score-wise, exactly equivalent

biologically, later may be more interesting, & is common

at least, if must miss some, rather miss the former

BLAST is a heuristic emphasizing the later

speed/sensitivity tradeoff: BLAST may miss former, but gains greatly in speed

# BLAST: What

## Input:

A query sequence (say, 300 residues)

A data base to search for other sequences similar to the query  
(say,  $10^6$  -  $10^9$  residues)

A score matrix  $\sigma(r,s)$ , giving cost of substituting r for s (& perhaps gap costs)

Various score thresholds & tuning parameters

## Output:

“All” matches in data base above threshold

“E-value” of each

# Blast: demo

E.g.

<http://expasy.org/sprot>

(or <http://www.ncbi.nlm.nih.gov/blast/> )

look up MyoD

go to blast tab

paste in ID or seq for human MyoD

set params (gapped=yes, blosum62,...)

get top 100 (or 1000) hits

# BLAST: How

*Idea: most interesting parts of the DB have a good ungapped match to some short subword of the query*

Break query into overlapping words  $w_i$  of small fixed length (e.g. 3 aa or 11 nt)

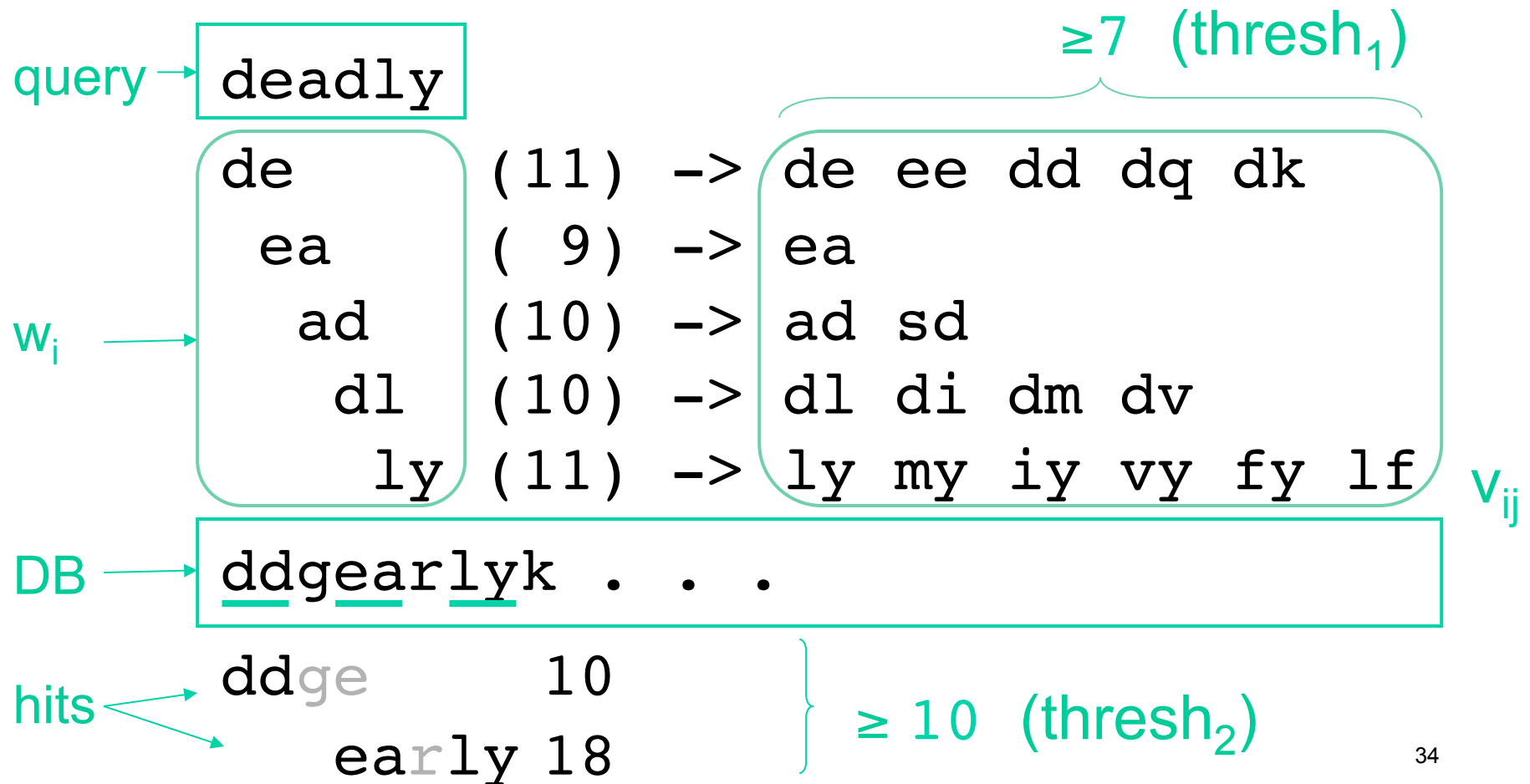
For each  $w_i$ , find (empirically,  $\sim 50$ ) “similar” words  $v_{ij}$  with score  $\sigma(w_i, v_{ij}) > \text{thresh}_1$  (say, 1, 2, ... letters different)

Look up each  $v_{ij}$  in database (via prebuilt index) -- i.e., exact match to short, high-scoring word

Grow each such “seed match” bidirectionally

Report those scoring  $> \text{thresh}_2$ , calculate E-values

# BLAST: Example



# BLOSUM 62 (the “ $\sigma$ ” scores)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	<b>4</b>	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-2	-1	1	0	-3	-2	0	
R	-1	<b>5</b>	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	<b>6</b>	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	<b>6</b>	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	<b>9</b>	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	<b>5</b>	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	<b>5</b>	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	<b>6</b>	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	<b>8</b>	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	<b>4</b>	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	<b>4</b>	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	<b>5</b>	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	<b>5</b>	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	<b>6</b>	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	<b>7</b>	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	<b>4</b>	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	<b>5</b>	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	<b>11</b>	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	<b>7</b>	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	<b>4</b>

# BLAST Refinements

“Two hit heuristic” -- need 2 nearby, nonoverlapping, gapless hits before trying to extend either

“Gapped BLAST” -- run heuristic version of Smith-Waterman, bi-directional from hit, until score drops by fixed amount below max

PSI-BLAST -- For proteins, iterated search, using “weight matrix” (next week?) pattern from initial pass to find weaker matches in subsequent passes

Many others



# Summary

BLAST is a highly successful search/alignment heuristic. It looks for alignments anchored by short, strong, ungapped “seed” alignments

Assessing statistical significance of alignment scores is crucial to practical applications

Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model

For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments

Permutation tests are a simple (but brute force) alternative

# Bio(tech) Interlude

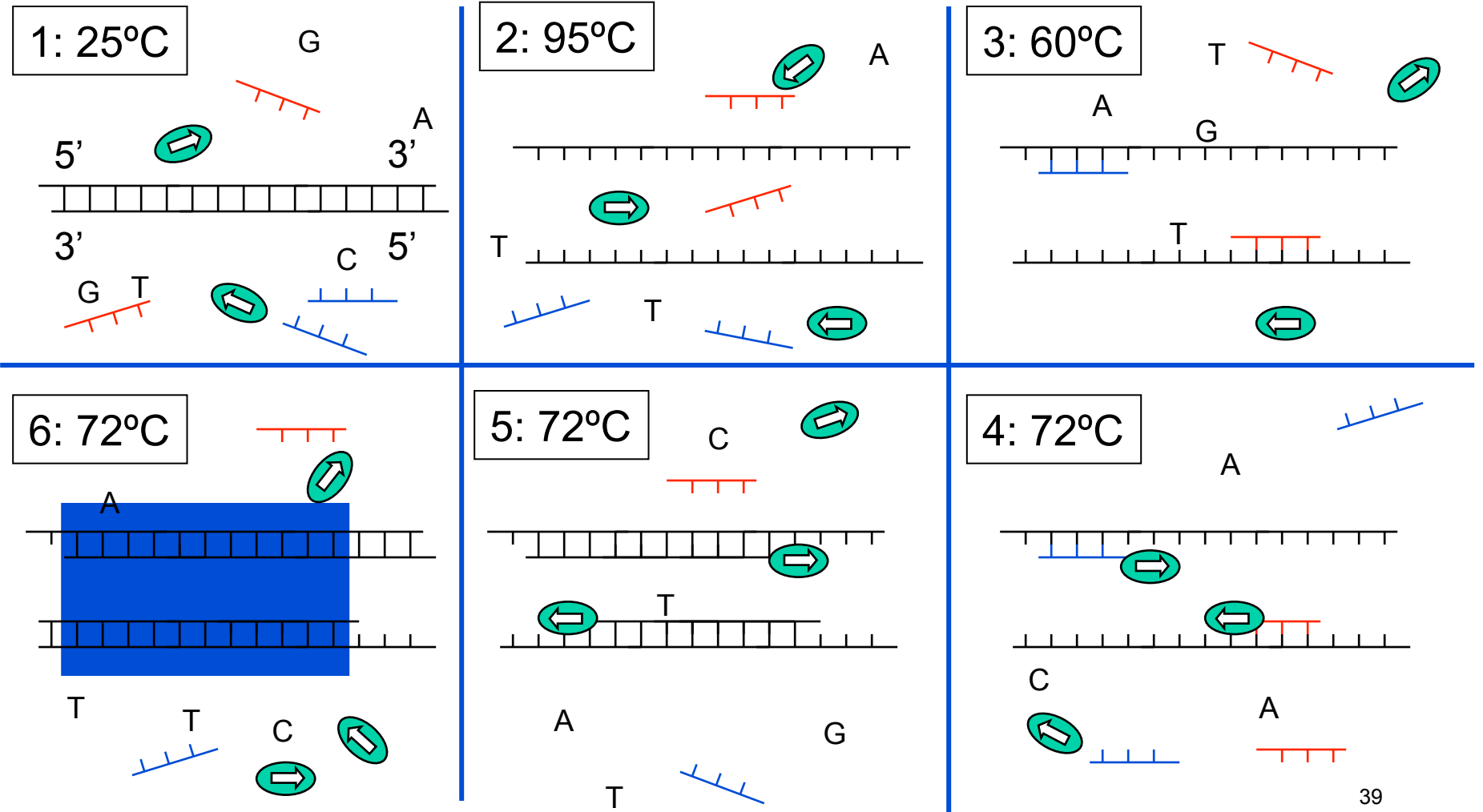
3 Nobel Prizes:

PCR: Kary Mullis, 1993

Electrophoresis: A.W.K. Tiselius, 1948

DNA Sequencing: Frederick Sanger, 1980

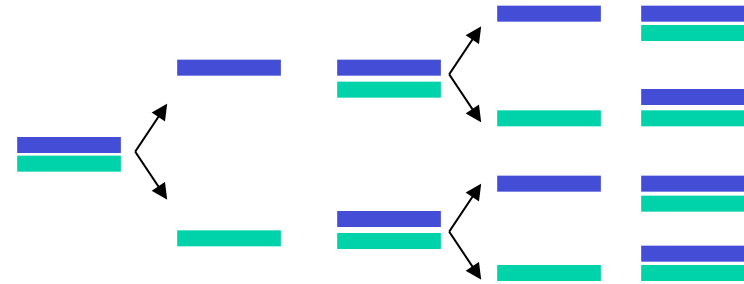
# PCR





Hot spring, near Great Fountain  
Geyser, Yellowstone National Park

# PCR



## Ingredients:

many copies of deoxy nucleotide triphosphates

many copies of two primer sequences (~20 nt each)

readily synthesized

many copies of Taq polymerase (*Thermus aquaticus*),

readily available commercialy

as little as 1 strand of template DNA

a programmable “thermal cycler”

Amplification: million to billion fold

Range: up to 2k bp routinely; 50k with other enzymes & care

# Why PCR?

PCR is important for all the reasons that filters and amplifiers are important in electronics, e.g., sample size is reduced from grams of tissue to a few cells, can pull out small signal amidst “noisy” background

*Very widely used*; forensics, archeology, cloning, sequencing, ...

# DNA Forensics

E.g. FBI “CODIS” (combined DNA indexing system) data base

As of 1/2013, over 10,142,600 offender profiles

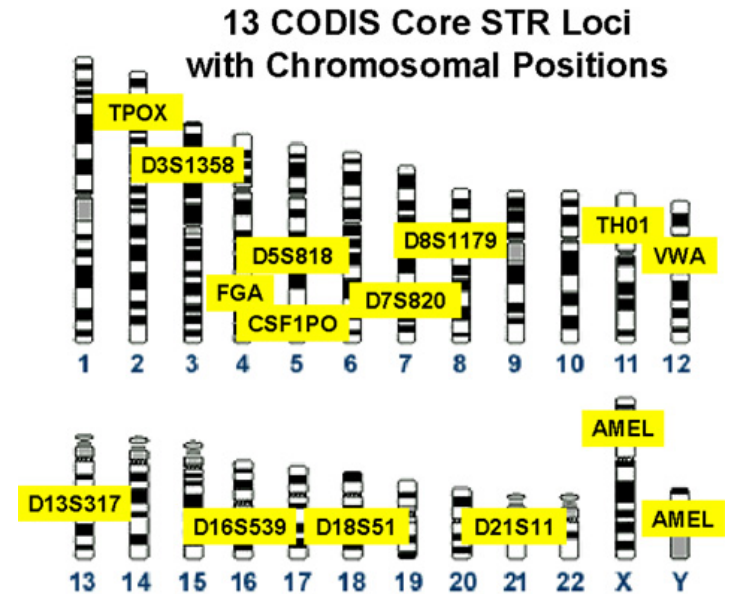
Picked 13 “short tandem repeats”, i.e., variable-length regions of human genome flanked by (essentially) invariant sequences (primer targets), several alleles common at each locus, of which you have 2

Amplify each from, e.g., small spot of dried blood

Measure product lengths (next slides)

<http://www.fbi.gov/about-us/lab/biometric-analysis/codis>

<http://www.dna.gov/solving-crimes/cold-cases/howdatabasesaid/codis/>



# Gel Electrophoresis

DNA/RNA backbone is negatively charged (they're acids)

Molecules moves slowly in gels under an electric field

agarose gels for large molecules

polyacrylamide gels for smaller ones

Smaller molecules move faster

So, you can *separate DNAs & RNAs by size*

Nobel Chem prize, 1948 Arne Wilhelm Kaurin Tiselius

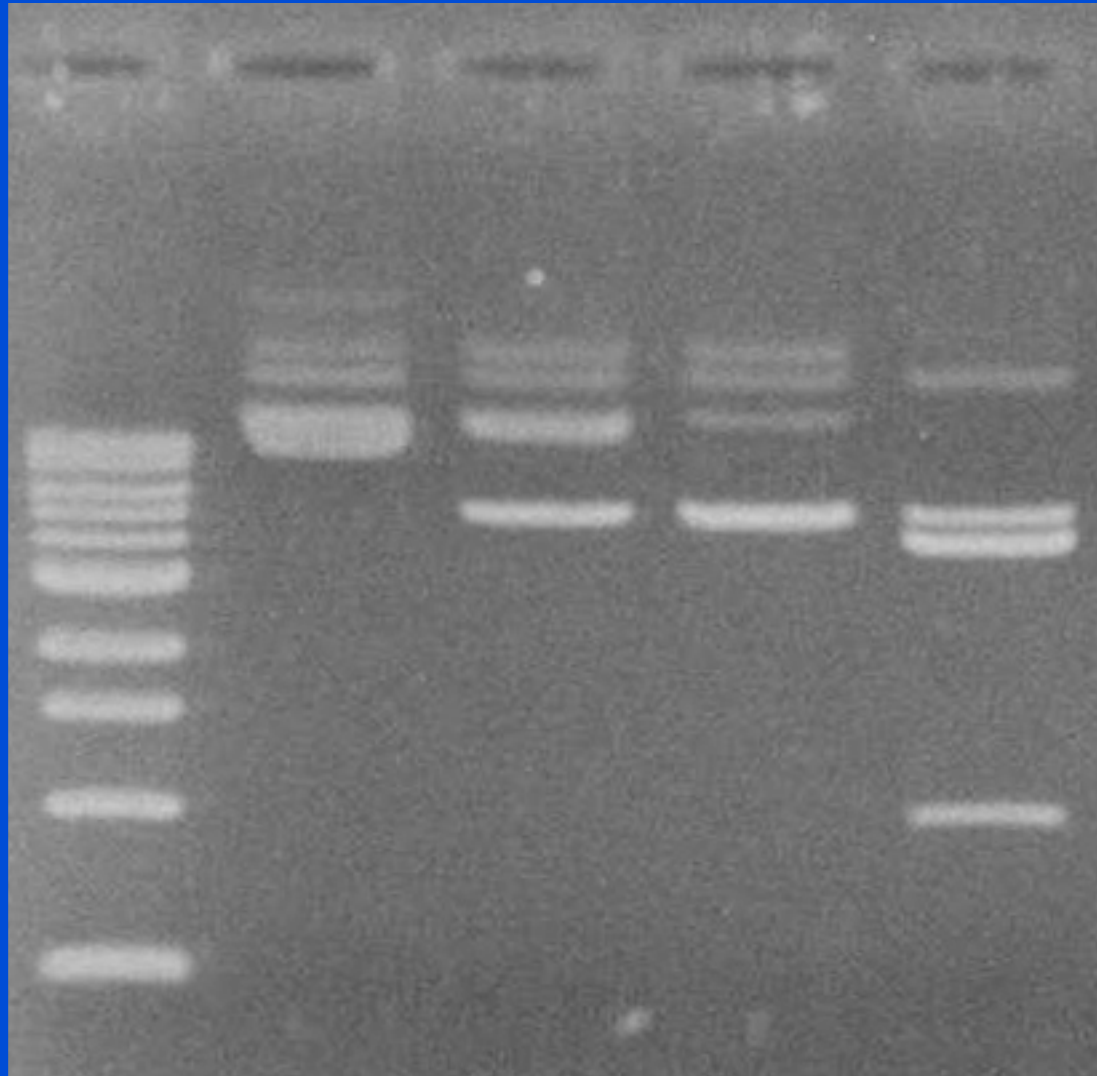


lane 1    lane 2    lane 3    lane 4    lane 5

10,000 bp →

3,000 bp →

500 bp →



-



+

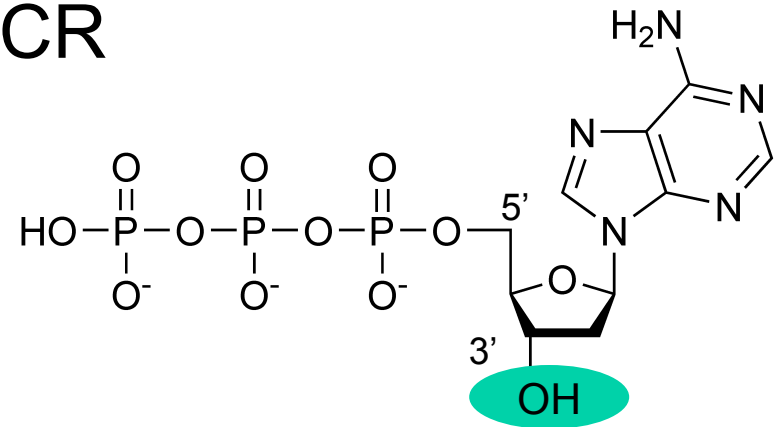
# DNA Sequencing – Sanger Method

Like one-cycle, one-primer PCR

Suppose 0.1% of A's:

are *di*-deoxy adenosine's;  
backbone can't extend

carry a green florescent dye



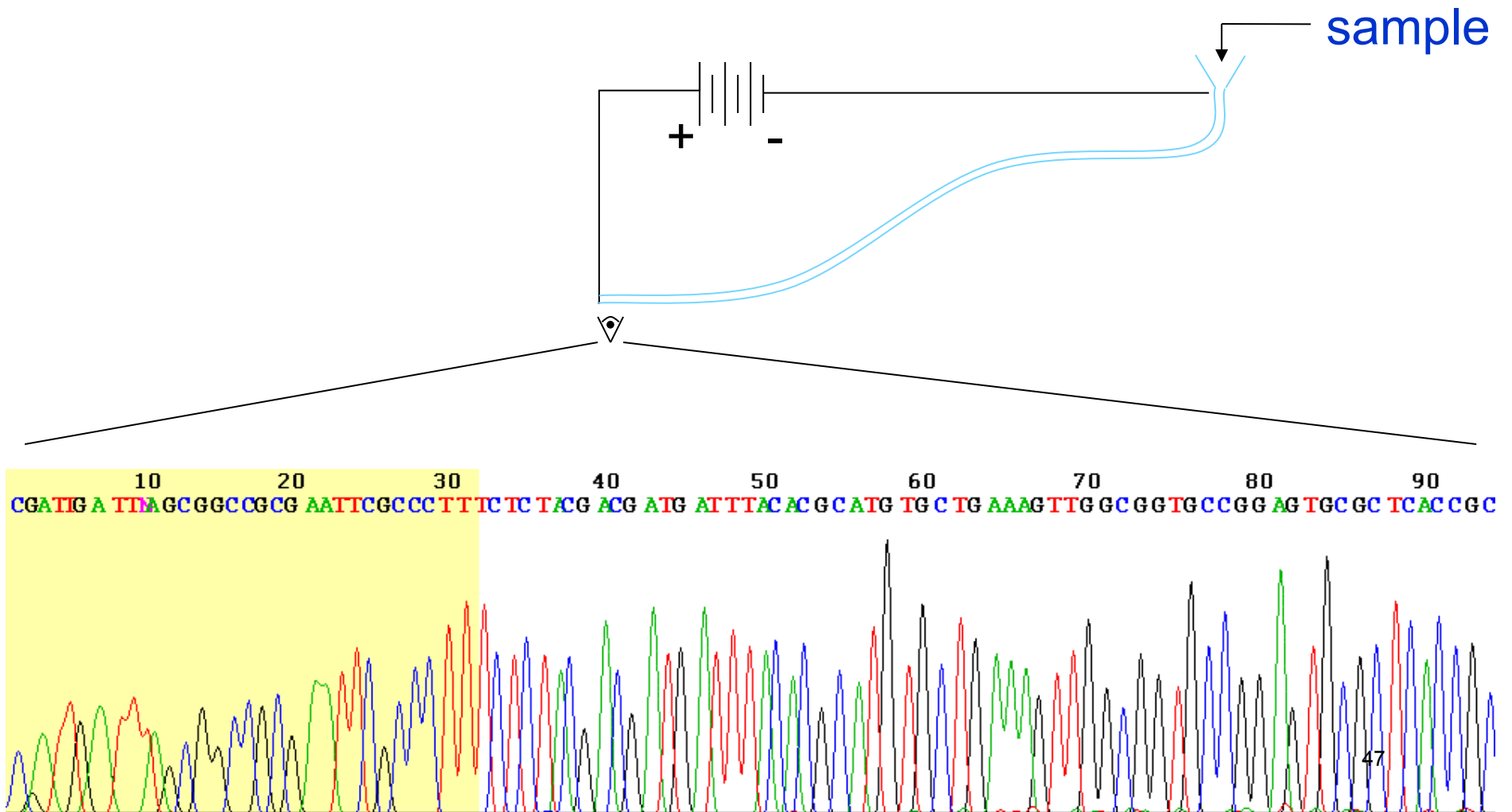
Separate by capillary gel electrophoresis

If frags of length 42, 49, 50, 55 ... glow green,  
those positions are A's

Ditto C's (blue), G's (yellow), T's (red)

# DNA Sequencing

## Sanger with capillary electrophoresis



# Sequencing A Genome

Highly automated

Typical Sanger “read” about 600 nt

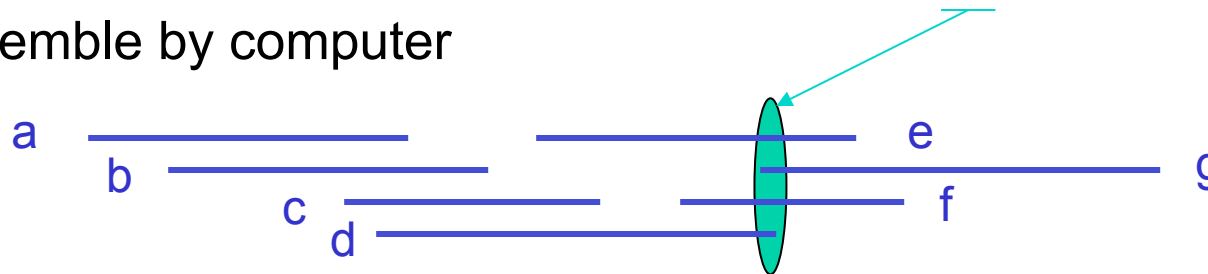
“Whole Genome Shotgun” approach:

randomly fragment (many copies of) genome

sequence many, enough to cover each base 10x or more times

reassemble by computer

E.g., human genome project:  
 $\approx 30\text{Gbases}$  and  
 $\approx 3 \times 10^9 / 600 \times 10$   
 $= 5 \times 10^7$  reads



Complications: repeated region, missed regions, sequencing errors, chimeric DNA fragments, ...

But overall accuracy  $\sim 10^{-4}$ , if careful

# “Next Generation” Sequencing

Many technical improvements to Sanger approach over many years, culminating in highly automated machines used for the HGP

Since then, many innovative new ideas/products:

- Helicos: single molecule fluorescence tethered to flow cell
- Illumina: colony PCR; reversible dye terminator
- Ion Torrent: semiconductor detection of ions released by polymerase
- Roche 454: emulsion PCR; pyro sequencing
- Oxford Nanopore
- Pacific Biosciences: single tethered polymerases in “zero mode waveguide” nano-wells, circularized DNA, “real time”
- ABI SOLiD: emulsion PCR, sequence by ligation, “color-space”
- Complete Genomics: rolling circle replication/DNA nanoballs

Technology is changing rapidly!

# “Next Generation” Sequencing

~1 billion microscopic PCR “colonies” on 1x2” slide

“Read” ~50-150bp of sequence from (1 or 2) ends of each

Ends fluorescently labeled, blocked, chemically cycled

Automated: takes a few days; ~ 100 G bases/day

Costs a few thousand dollars

Generates terabytes of data (mostly images)

I,e., ~ 30x human genome/day (you need 25x-50x to assemble)

Other approaches: long reads, single molecules,...

Technology is changing rapidly!

# Illumina Sequencing

~1 billion microscopic PCR “colonies” on 1x2” slide

“Read” ~50-150bp of sequence from (1 or 2) ends of each

Reversible dye terminators

Automated: takes a few days; ~ 100 G bases/day

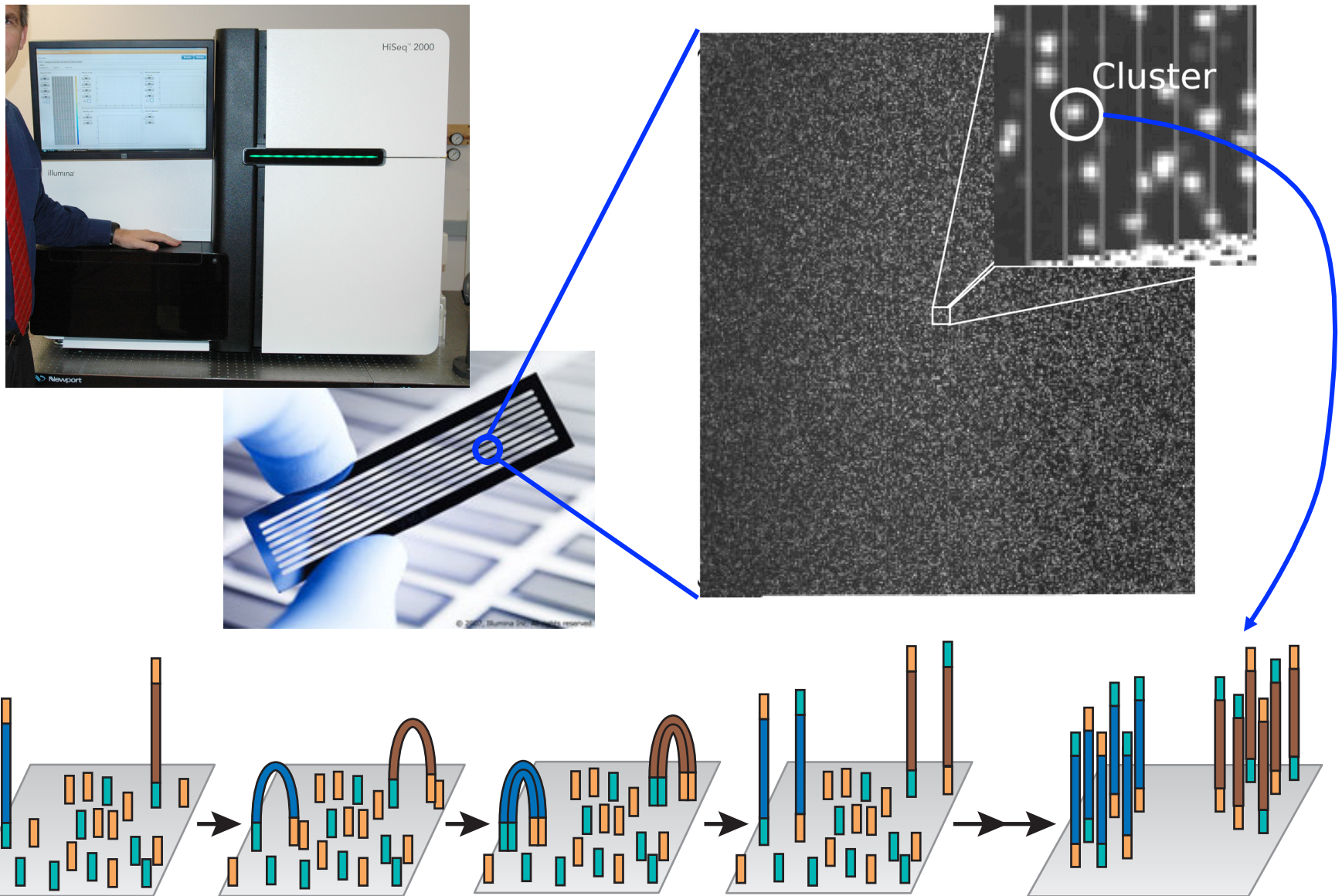
Costs a few thousand dollars

Generates terabytes of data (mostly images)

I,e., ~ 30x human genome/day

(you need 25x-50x to assemble)

(equal to all of pre-2008 Genbank)



[http://www.technologyreview.com/sites/default/files/legacy/pgenome\\_x220.jpg](http://www.technologyreview.com/sites/default/files/legacy/pgenome_x220.jpg)
<http://bioinformatics.oxfordjournals.org/content/25/17/2194/F1.large.jpg>

Fig from: Shendure and Ji 2008. "Next-Generation DNA Sequencing.." *Nature Biotechnol* 26 (10) (October): 1135–1145. doi:10.1038/nbt1486.



# Illumina HiSeq (1500/2500, as of Spring 2013)

## HIGH OUTPUT RUN MODE\*

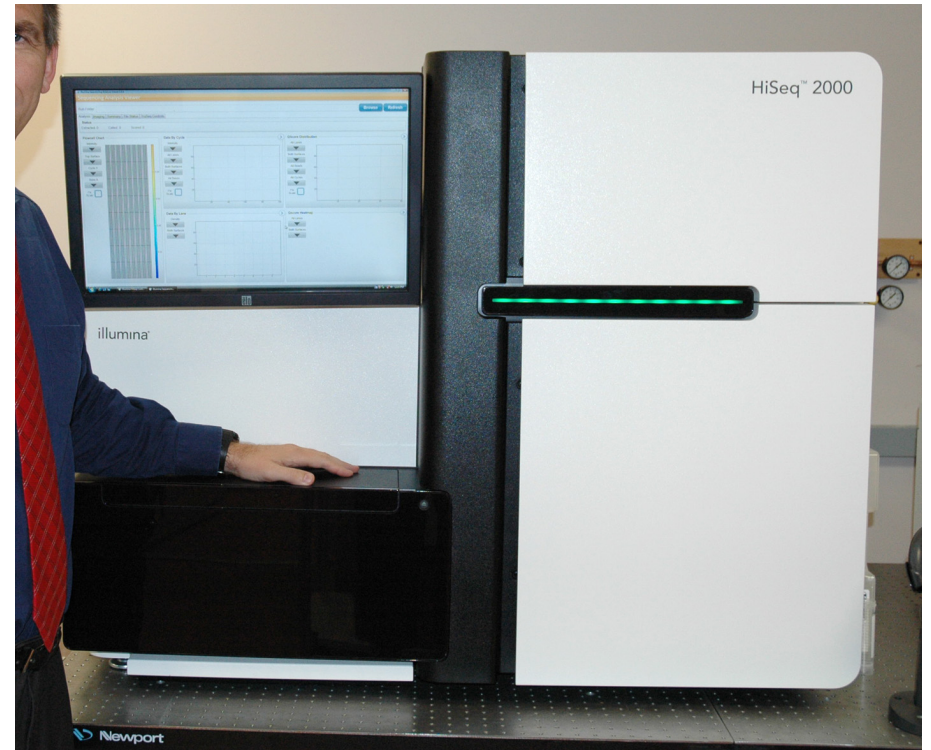
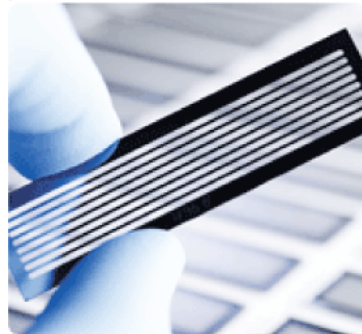
## RAPID RUN MODE\*

Read Length	Dual Flow Cell (2500 only)	Single Flow Cell (1500 or 2500)	Dual Flow Cell Run Time	Dual Flow Cell (2500 only)	Single Flow Cell (1500 or 2500)	Dual Flow Cell Run Time
1 x 36	95-105 Gb	47-52 Gb	2 days	18-22 Gb	9-11 Gb	7 hr
2 x 50	270-300 Gb	135-150 Gb	5.5 days	50-60 Gb	25-30 Gb	16 hr
2 x 100	540-600 Gb	270-300 Gb	11 days	100-120 Gb	50-60 Gb	27 hr
2 x 150	N/A	N/A	N/A	150-180 Gb	75-90 Gb	40 hr
Reads Passing Filter	Up to 3 billion single reads or 6 billion paired-end reads	Up to 1.5 billion single reads or 3 billion paired-end reads		Up to 600 million single reads or 1.2 billion paired-end reads	Up to 300 million single reads or 600 million paired-end reads	
Quality	> 85% of bases above Q30 at 2 x 50 bp > 80% of bases above Q30 at 2 x 100 bp			> 85% of bases above Q30 at 2 x 50 bp > 80% of bases above Q30 at 2 x 100 bp > 75% of bases above Q30 at 2 x 150 bp		

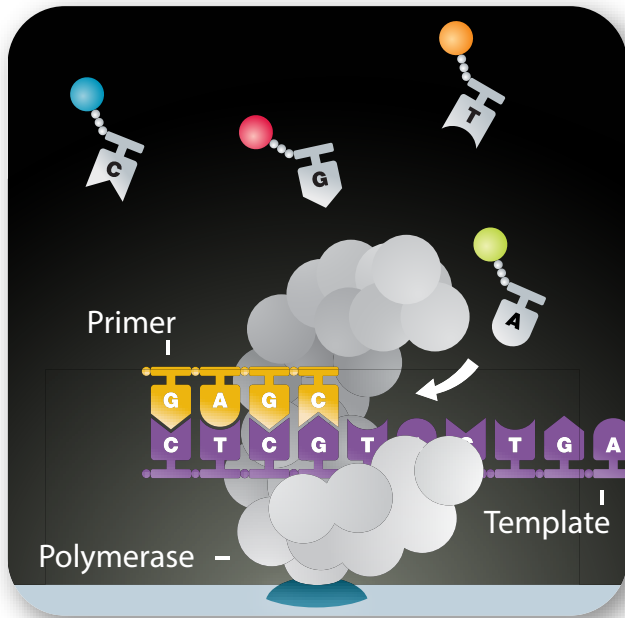
\*Install specifications based on Illumina PhiX control library at supported cluster densities (between 610-678 K clusters/mm<sup>2</sup> passing filter using TruSeq v3 Kits or 700-820 clusters/mm<sup>2</sup> passing filter using TruSeq Rapid Kits. Run times for rapid run mode correspond to on-board cluster generation (1.5 hr) and sequencing; for high output mode, run times correspond to sequencing only. Performance may vary based on sample quality, cluster density, and other experimental factors. Early HiSeq 2000 instruments will run slightly slower when upgraded to a HiSeq 2500.

# Modern DNA Sequencing

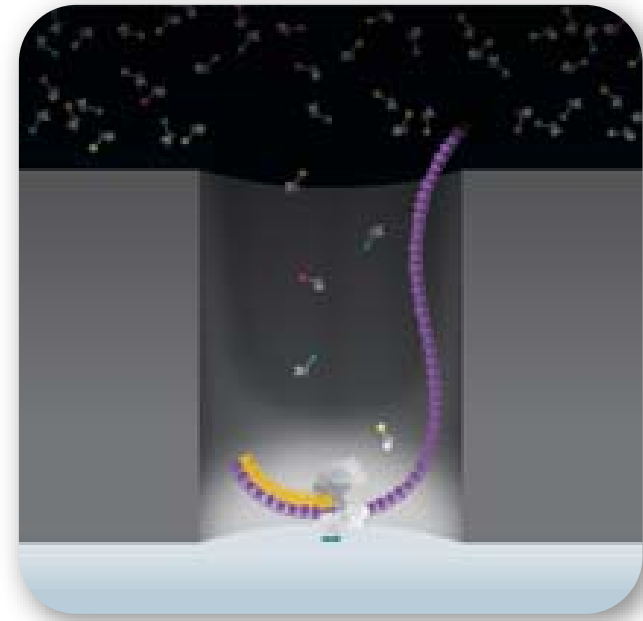
A table-top box the size of your oven (but costs a bit more ... ;- ) can generate ~100 billion BP of DNA seq/day; i.e.  
= 2008 genbank,  
= 30x your genome



# Pacific Biosciences



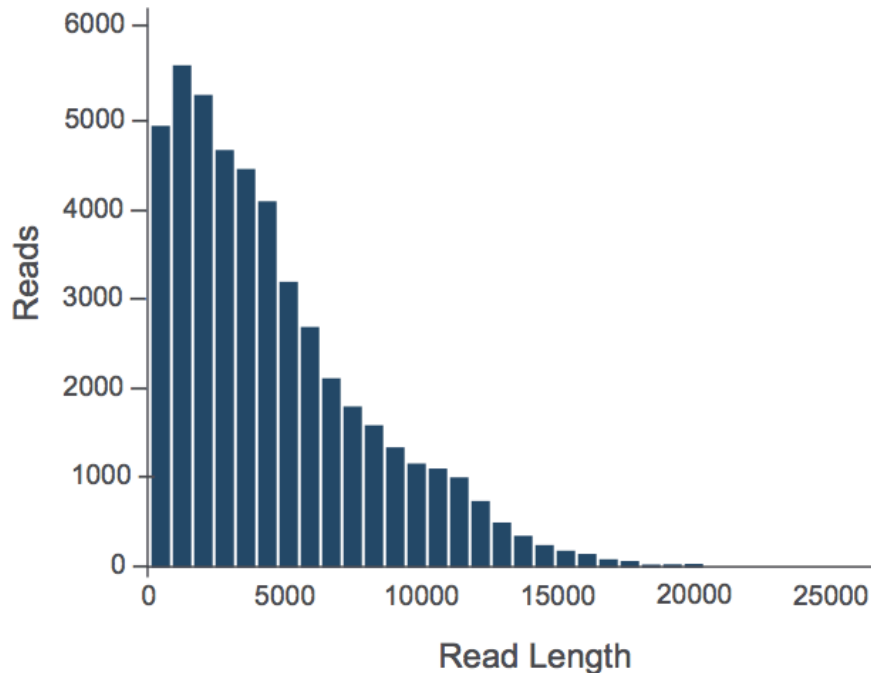
Phospholinked  
nucleotides



Zero-Mode  
Waveguides

# Pacific Biosciences

Read Length Distribution

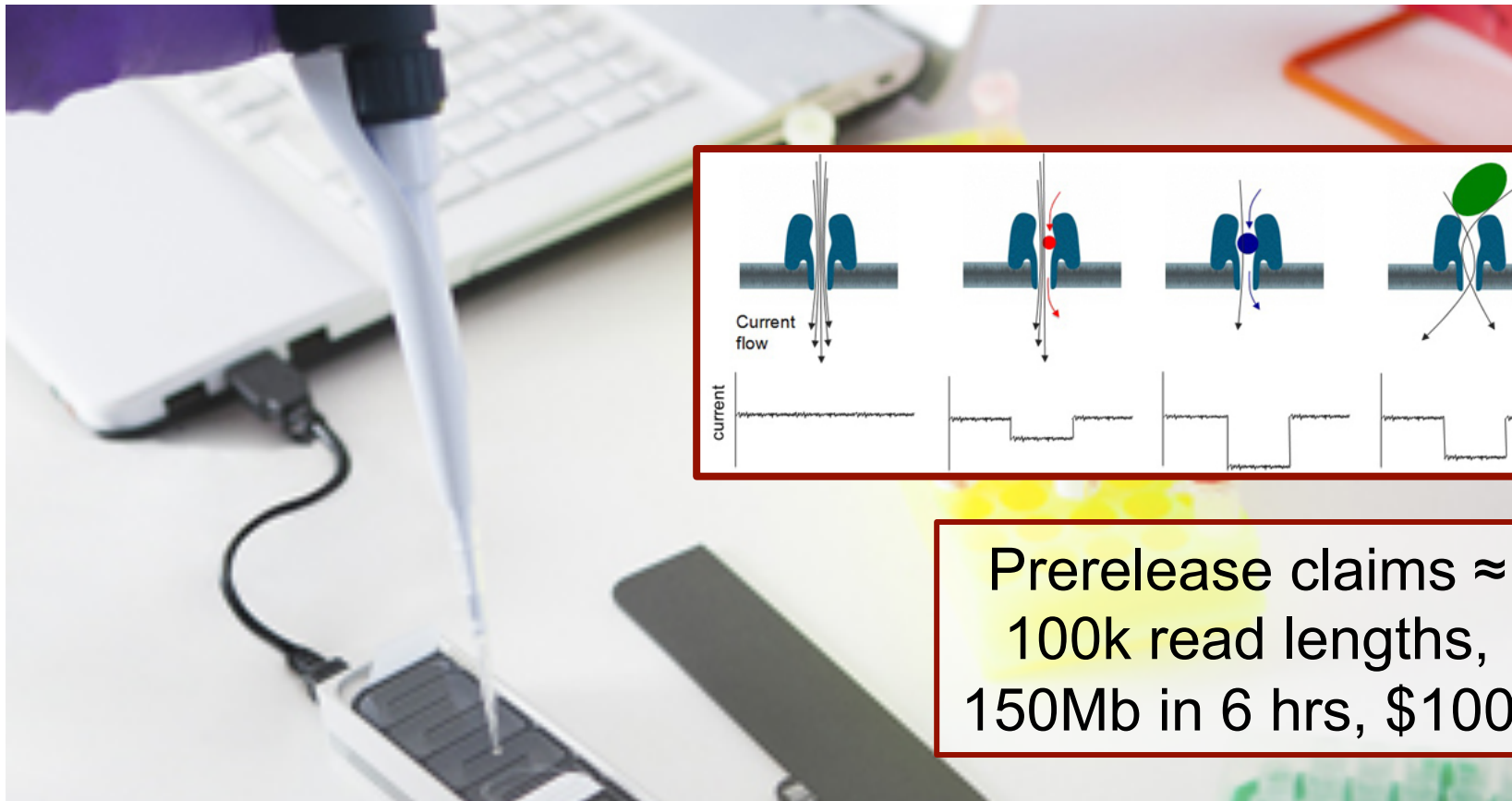


- Advantages:
  - single molecules
  - long reads
  - direct CH<sub>3</sub> detection
- Disadvantages:
  - throughput
  - error rate; (circularize?)

<b>Read Length:</b>	
Average:	4,606 bp
95 <sup>th</sup> Percentile:	11,792 bp
Maximum:	23,297 bp
<b>Throughput per SMRT<sup>®</sup> Cell:</b>	
	216 Mb
	47,197 reads

Based on data from 11 kb plasmid library using a 120 minute movie

# Oxford Nanopore



Prerelease claims  $\approx$   
100k read lengths,  
150Mb in 6 hrs, \$1000

# Personal Genomes

2001: ~\$2.7 billion (Human Genome Project)

2003: ~\$300 million

2007: ~\$1 million

2008: ~\$60 thousand

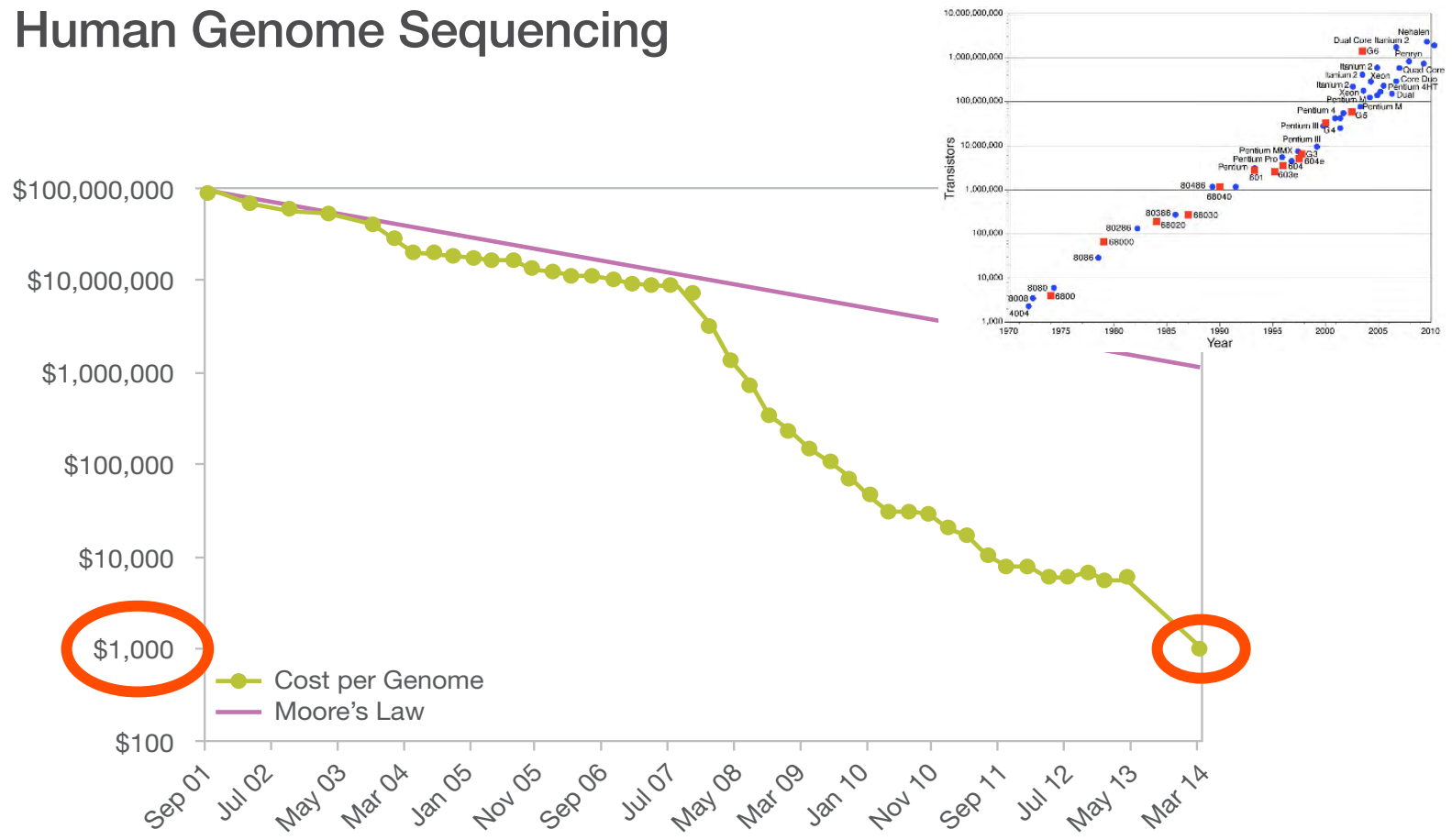
2009: ~\$4400

2015: ~\$1000

*bioinformatics not included...*



# Figure 3: Illumina Sequencing Technology Outpaces Moore's Law for the Price of Whole Human Genome Sequencing





# Summary

PCR allows simple *in vitro* amplification of minute quantities of DNA (having pre-specified boundaries)

Sanger sequencing uses

- a PCR-like setup with modified chemistry to generate varying length prefixes of a DNA template with the last nucleotide of each color-coded

- gel electrophoresis to separate DNA by size, giving sequence

Sequencing random overlapping fragments allows genome sequencing (and many other applications)

“Next Gen” sequencing: many innovations

- throughput up, cost down (lots!)

# More on p-values and hypothesis testing

# P-values & E-values

p-value:  $P(s,n)$  = *probability* of a score more extreme than  $s$  when searching a random target data base of size  $n$

E-value:  $E(s,n)$  = *expected number* of such matches

They Are Related:

$$E(s,n) = \underline{pn} \text{ (where } p = P(s,1) \text{)}$$

$$P(s,n) = 1-(1-p)^n = 1-(1-1/(1/p))^{(1/p)(pn)} \approx 1-\exp(-pn) = \underline{1-\exp(-E(s,n))}$$

E big (say,  $\gg 1$ )  $\Leftrightarrow$  P big ( $\rightarrow 1$ )

$$E = 5 \Leftrightarrow P \approx .993$$

$$E = 10 \Leftrightarrow P \approx .99995$$

E small  $\Leftrightarrow$  P small (both near 0)

$$E = .01 \Leftrightarrow P \approx E - E^2/2 + E^3/3! \dots \approx \underline{E}$$

Both equally valid; E-value is perhaps more intuitively interpretable

# Hypothesis Testing: A Very Simple Example

Given: A coin, either fair ( $p(H)=1/2$ ) or biased ( $p(H)=2/3$ )

Decide: which

How? Flip it 5 times. Suppose outcome  $D = \text{HHHTH}$

Null Model/Null Hypothesis  $M_0$ :  $p(H)=1/2$

Alternative Model/Alt Hypothesis  $M_1$ :  $p(H)=2/3$

Likelihoods:

$$P(D | M_0) = (1/2) (1/2) (1/2) (1/2) (1/2) = 1/32$$

$$P(D | M_1) = (2/3) (2/3) (2/3) (1/3) (2/3) = 16/243$$

$$\text{Likelihood Ratio: } \frac{p(D | M_1)}{p(D | M_0)} = \frac{16/243}{1/32} = \frac{512}{243} \approx 2.1$$

I.e., given data is  $\approx 2.1x$  more likely under alt model than null model

# Hypothesis Testing, II

Log of likelihood ratio is equivalent, often more convenient

add logs instead of multiplying...

“Likelihood Ratio Tests”: reject null if  $LLR > \text{threshold}$

$LLR > 0$  disfavors null, but higher threshold gives stronger evidence against

Neyman-Pearson Theorem: For a given error rate, LRT is as good a test as any (subject to some fine print).

# A Likelihood Ratio

Defn: two proteins are *homologous* if they are alike because of shared ancestry; similarity by descent

Suppose among proteins overall, residue x occurs with frequency  $p_x$   
Then in a random alignment of 2 random proteins, you would expect to find x aligned to y with prob  $p_x p_y$

Suppose among *homologs*, x & y align with prob  $p_{xy}$

Are seqs X & Y homologous? Which is more likely, that the alignment reflects chance or homology? Use a *likelihood ratio test*.

$$\sum_i \log \frac{p_{x_i y_i}}{p_{x_i} p_{y_i}}$$

# Non-*ad hoc* Alignment Scores

Take alignments of homologs and look at frequency of x-y alignments vs freq of x, y overall

## Issues

biased samples  
evolutionary distance

## BLOSUM approach

Large collection of trusted alignments  
(the BLOCKS DB)

Subset by similarity

BLOSUM62  $\Rightarrow$   $\geq$  62% identity

e.g. <http://blocks.fhcrc.org/blocks-bin/getblock.pl?IPB002546>

$$\frac{1}{\lambda} \log_2 \frac{P_{x y}}{p_x p_y}$$

Scores: formula  
above, rounded

# BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	<b>4</b>	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	<b>5</b>	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	<b>6</b>	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	<b>6</b>	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	<b>9</b>	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	<b>5</b>	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	<b>5</b>	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	<b>6</b>	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	<b>8</b>	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	<b>4</b>	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	<b>4</b>	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	<b>5</b>	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	<b>5</b>	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	<b>6</b>	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	<b>7</b>	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	<b>4</b>	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	<b>5</b>	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	<b>11</b>	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	<b>7</b>	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	<b>4</b>



## *ad hoc* Alignment Scores?

Make up any scoring matrix you like

Somewhat surprisingly, under pretty general assumptions<sup>\*\*</sup>, it is *equivalent* to the scores constructed as above from some set of probabilities  $p_{xy}$ , so you might as well understand what they are

NCBI-BLAST: +1/-2    tuned for ~ 95% sequence identity

WU-BLAST: +5/-4    tuned for ~ 66% identity (“twilight zone”)

---

<sup>\*\*</sup> e.g., average scores should be negative, but you probably want that anyway, otherwise local alignments turn into global ones, and some score must be  $> 0$ , else best match is empty

# Summary

Assessing statistical significance of alignment scores is crucial to practical applications

Score matrices derived from “likelihood ratio” test of trusted alignments vs random “null” model

For gapless alignments, Extreme Value Distribution (EVD) is theoretically justified for overall significance of alignment scores; empirically ok in other contexts, too, e.g., for gapped alignments.

Permutation tests are a simple and broadly applicable (but brute force) alternative

Looking at residue substitutions in a large set of “trusted” alignments provides a sound basis for defining the score tables