"A DNA sequence for the genome of bacteriophage ΦX174 of approximately 5,375 nucleotides has been determined using the rapid and simple 'plus and minus' method. The sequence identifies many of the features responsible for the production of the proteins of the nine known genes of the organism, including initiation and termination sites for the proteins and RNAs. Two pairs of genes are coded by the same region of DNA using different reading frames."
— Frederick Sanger

(Relevant) Trivia

How many base pairs (bp) are there in a human genome?

~3 billion (haploid)

How much did it cost to sequence the first human genome?

~$2.7 billion

How long did it take to sequence the first human genome?

~13 years

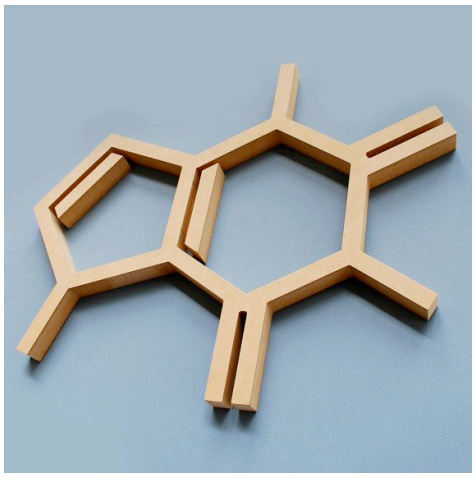
When was the first human genome sequence complete?

2001 (publication date)

Whose genome was it?

Several people's, but actually mostly a dude from Buffalo

So how do we actually get DNA sequence?



Hang on …
there is a lot of chemistry ahead!!

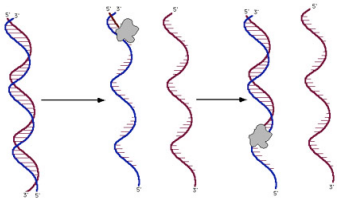
How does sequencing work?

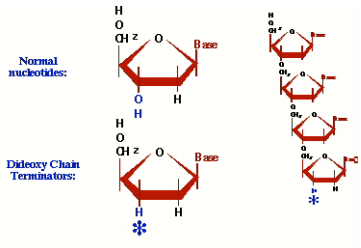
- Sanger Method (dideoxy method)
  1. Amplify DNA
  2. Use small amount dideoxy (ddNTP)
  3. Run synthesis reaction
  4. Use gel to separate fragments by size
  5. Read sequence from gel

Note: Originally only 500-800 bases can be sequenced in one reaction!

How does sequencing work?

1. Similar to PCR – we use a primer to initiate replication of DNA

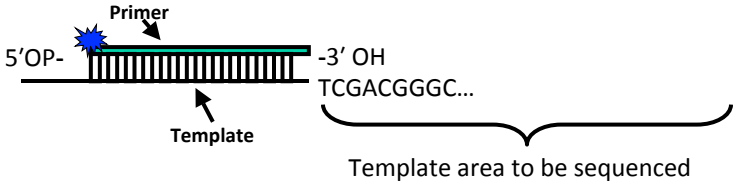




2. Run the reactions in the presence of a small amount of dideoxyribonucleotide

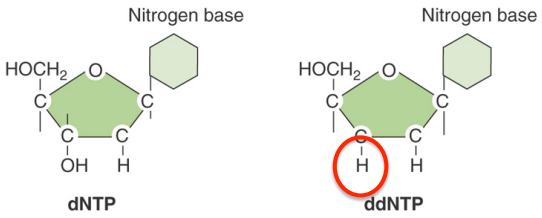
How does sequencing work?

1 A sequencing reaction mix includes primer (to get the reaction started) and template (what you want to sequence).



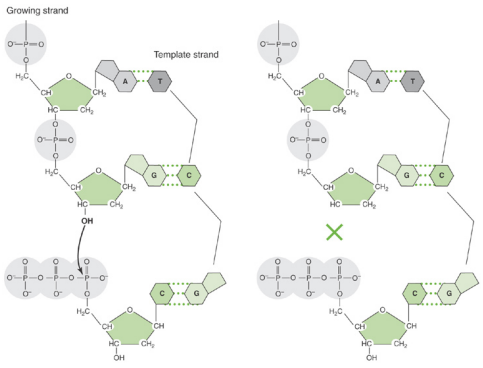
Chain Termination

- 2. Run the reactions in the presence of a small amount of dideoxyribonucleotide



Chain Termination

The 3’-OH group necessary for formation of the phosphodiester bond is missing in ddNTPs.



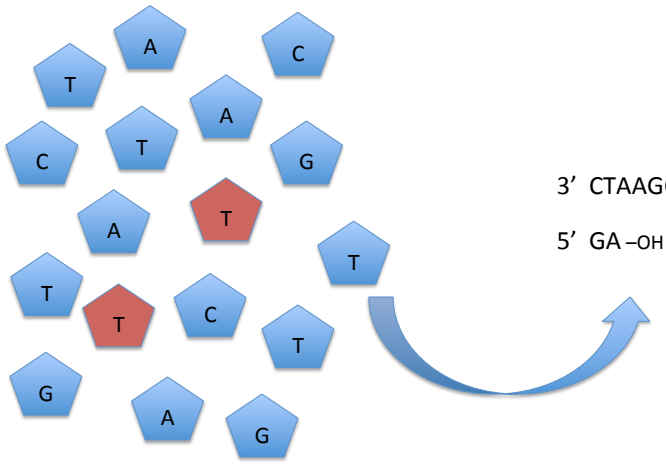
Chain terminates at ddG

3. Replicating a DNA strand in the presence of a specific dideoxyribonucleotide (here dideoxy-T) cause a fraction of copies to stop at this base.

IF 5% of the T nucleotides are actually dideoxy T, then each strand will terminate when it gets a ddT on its growing end:

5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCTACCGAT•
5' - TACGCGGTAACGGTATGTTCGACCGTTTAGCT•
5' - TACGCGGTAACGGTATGTTCGACCGTTT•
5' - TACGCGGTAACGGTATGTTCGACCGTT•
5' - TACGCGGTAACGGTATGTTCGACCGT•
5' - TACGCGGTAACGGTATGTT•
5' - TACGCGGTAACGGTATGT•
5' - TACGCGGTAACGGTAT•
5' - TACGCGGTAACGGT•
5' - TACGCGGT•

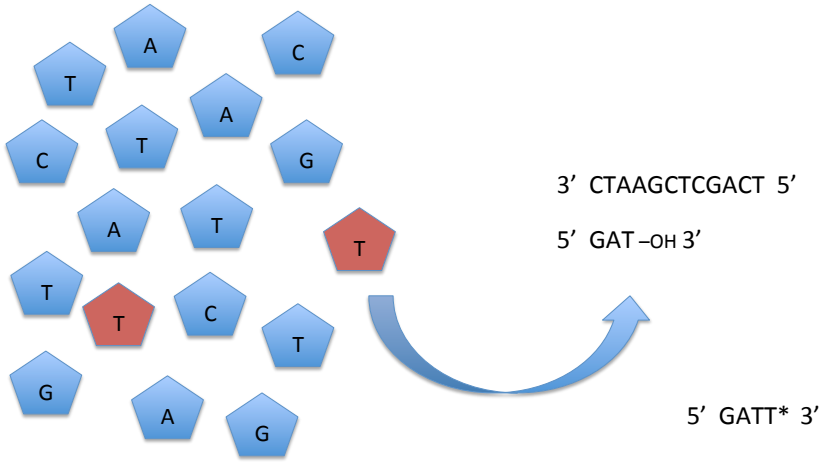
Chain termination is stochastic.

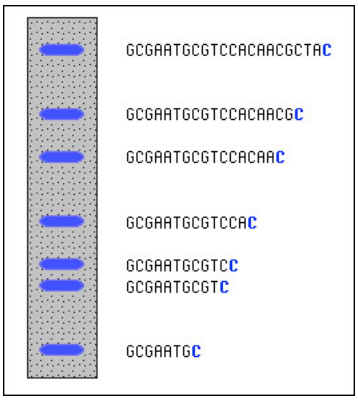


3’ CTAAGCTCGACT 5’

5’ GA –OH 3’

Chain termination is stochastic.



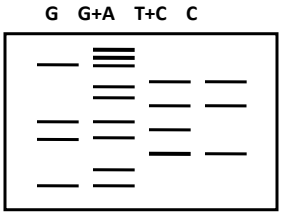


4. Gel electrophoresis can be used to separate the fragments by size and measure them (shown for dideoxy-C).

Running fragments on a gel, separates by size

Longer fragments

Shortest fragments



3′
A
A
G
C
A
A
C
G
T
G
C
A
G
5′

Sequencing gels are read from bottom to top (5′ to 3′).

Need four reactions to have a chance of stopping at every single base.

ddATP + four dNTPs

ddA
dAdGdCdTdGdCdCdCdG

ddCTP + four dNTPs

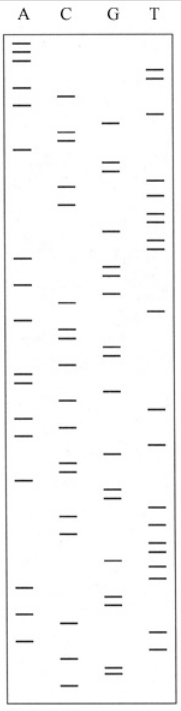
dAdGddC
dAdGdCdTdGddC
dAdGdCdTdGdCddC
dAdGdCdTdGdCdCddC

ddGTP + four dNTPs

dAddG
dAdGdCdTddG
dAdGdCdTdGdCdCdCddG

ddTTP + four dNTPs

dAdGdCddT
dAdGdCdTdGdCdCdCdG



5. Running all four bases possibilities allows one to “read” the sequence off the gel. Remember smaller fragments are at the bottom so we read going UP the gel.

cggctatcaggattgttctctggaccgt…

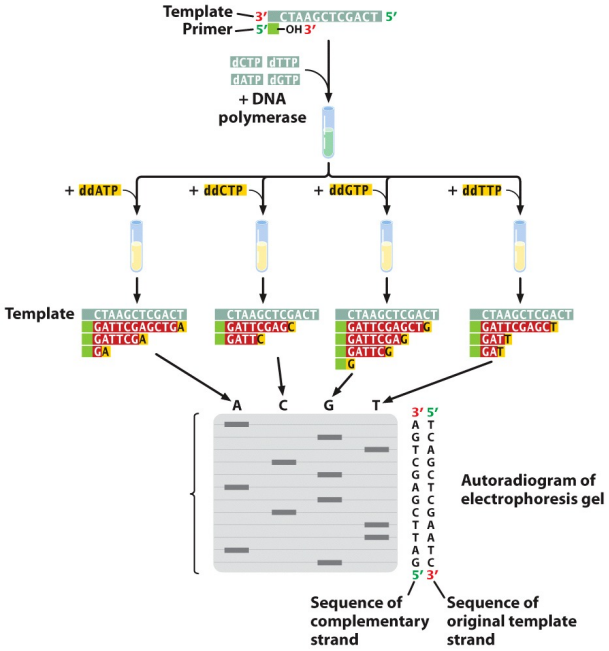
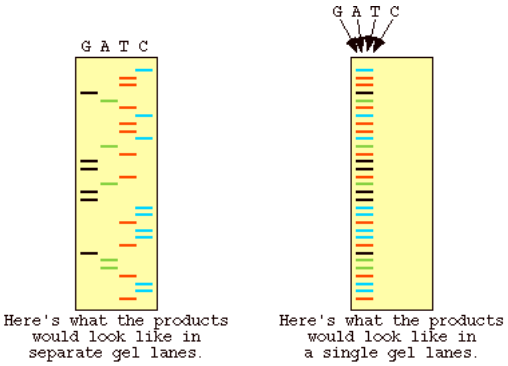


Figure 19.25
Genetics: A Conceptual Approach, Fifth Edition

The Human Genome Project

- In 1988, the NIH established an Office of Human Genome Research, with James Watson as director.
- The human genome project officially began on October 1, 1990.
- Various side projects: genetic diseases, variations between individuals, ethnic variation, comparison to other species.

How do we automate the process?



Innovation #1:
Why not label each nucleotide with a different fluorescent dye?

(note I am using black rather than yellow because it displays easier)

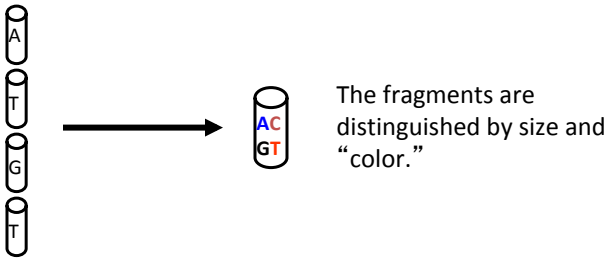
Fluorescent Dyes

- In dye primer sequencing, the primer contains fluorescent dye–conjugated nucleotides, labeling the sequencing ladder at the 5′ ends of the chains.

- In dye terminator sequencing, the fluorescent dye molecules are covalently attached to the dideoxynucleotides, labeling the sequencing ladder at the 3′ ends of the chains.

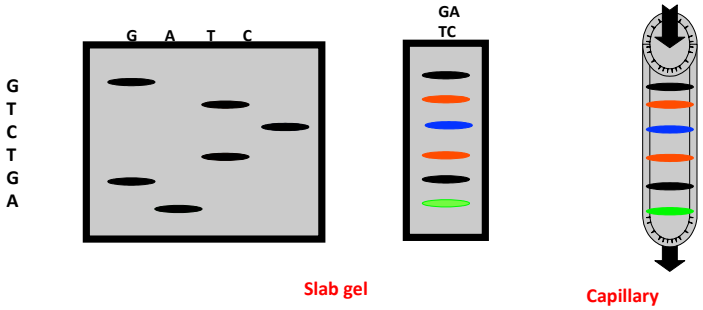
Dye Terminator Sequencing

- A distinct dye or "color" is used for each of the four ddNTP.
- Since the terminating nucleotides can be distinguished by color, all four reactions can be performed in a single tube.

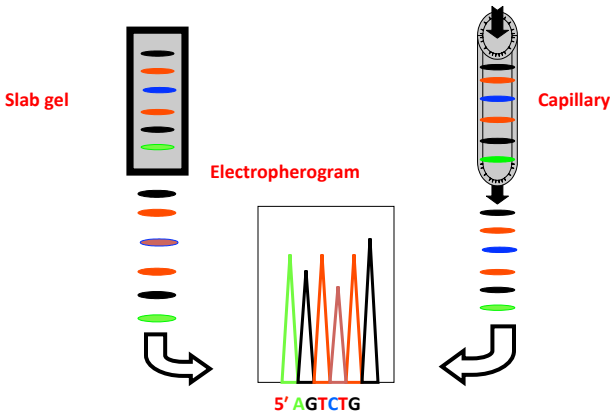


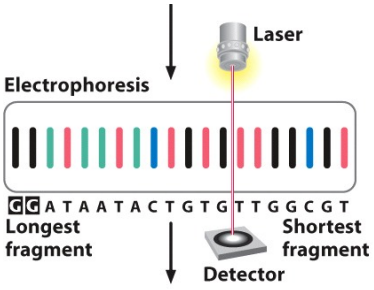
Dye Terminator Sequencing

The DNA ladder is resolved in one gel lane or in a capillary.



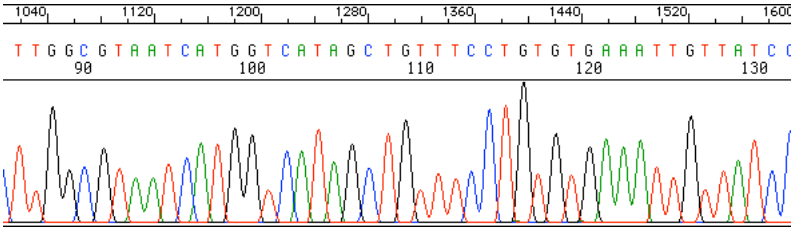
Single lane means it can be made very small.





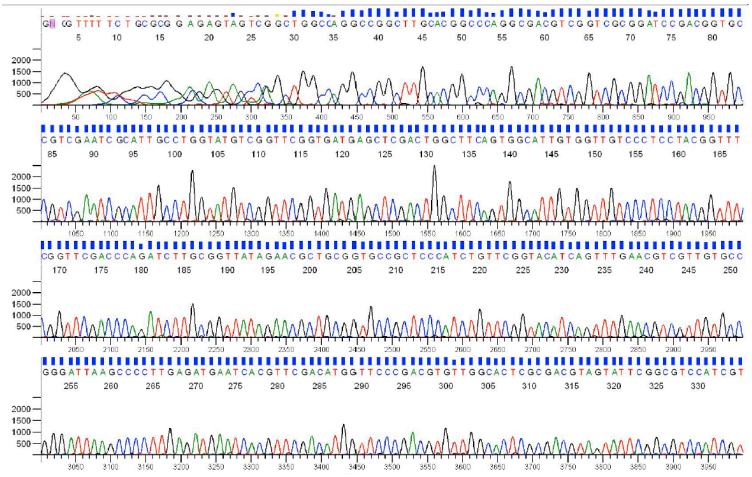
Innovation #2:

Why not let the computer "read" the gel?

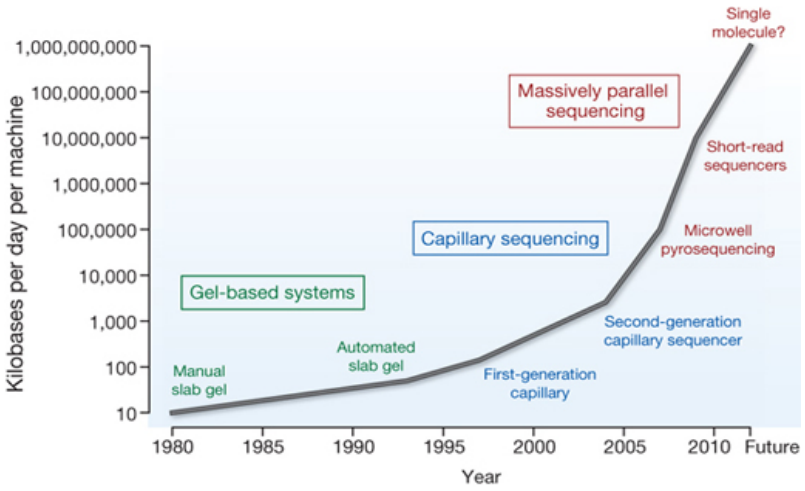


Sanger Sequencing Output

Each sequencing reaction gives us a **chromatogram**, usually ~600-1000 bp:



Astronomical pace of sequencing



http://genome.wellcome.ac.uk/doc_WTX059576.html

Fast moving technology …

Applied Biosystems
ABI 3730XL

Applied Biosystems
SOLiD

Illumina/Solexa
HiSeq

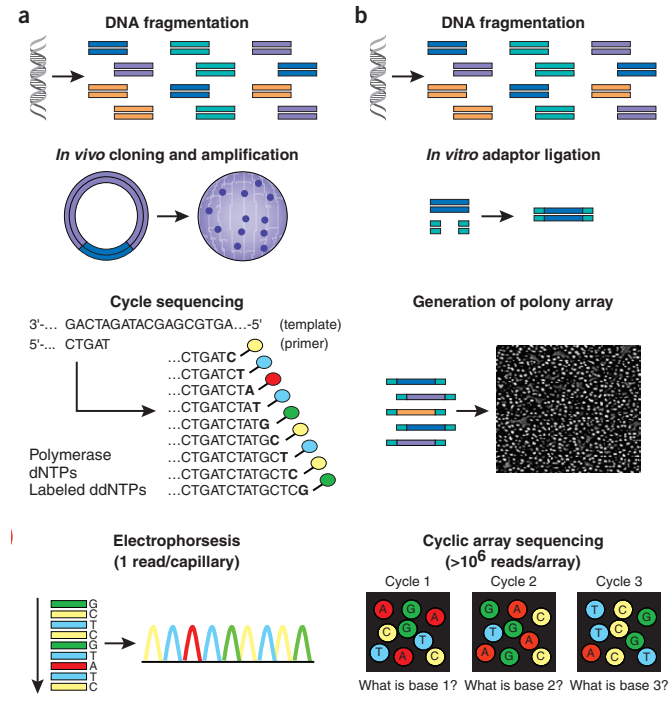
Roche 454
Genome Sequencer FLX

Illumina/Solexa
Genome Analyzer II

Helicos
Heliscope

Innovation #3:

Let's parallelize the whole thing and read MANY sequences at once.



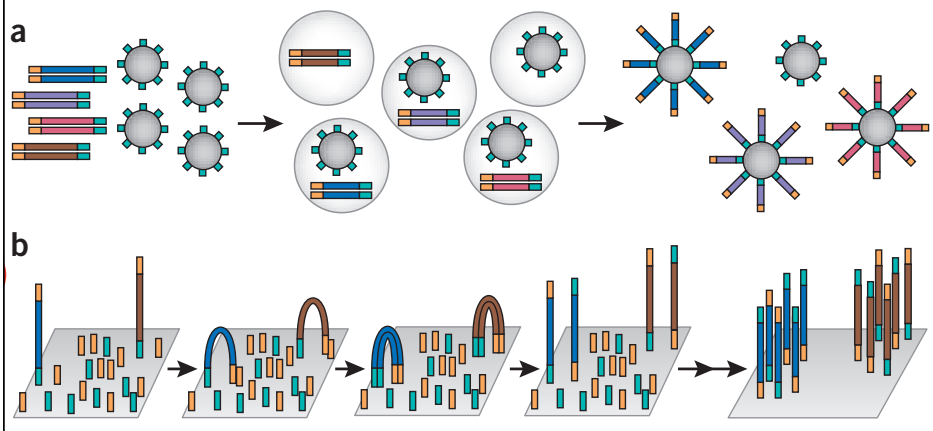
A) Sanger

B) Short read technologies

Note that in modern sequencing you *only* use the terminated nucleotides and depend on chemistry to cleave the blocker for the next round.

Short read sequencing methods

A) Emulsion methods (Ion torrent)
B) Flowcell clusters (Illumina)



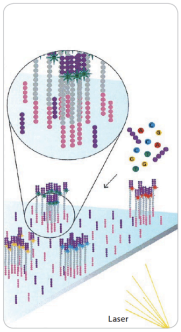
Shendure and Ji (2008) Next-generation DNA sequencing. *Nat. Rev. Biotech*. 26:1135-45

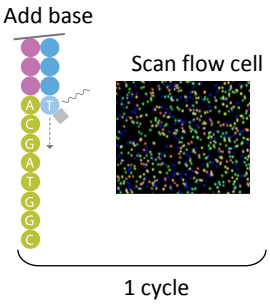
Reuter et al (2015) High Throughput Sequencing Technologies. Mol Cell 58:586-597

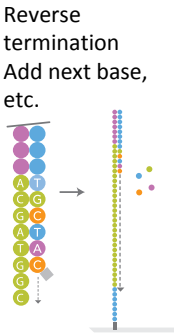
All sequencing strategies have some amount of errors/ mistakes. These can come from humans (user error), chemistry, technology (reading wrong color), etc.

For example: Errors in Illumina sequencing reads

Sequencing by synthesis with reversible dye terminators



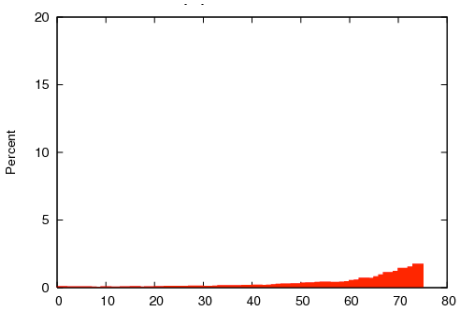


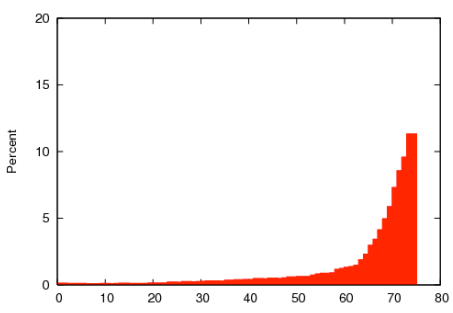


- Wrong "color" called when scanning image
- Wrong nucleotide is added
- Terminator not removed for a cycle
- Multiple bases added (terminator doesn't work for a base)

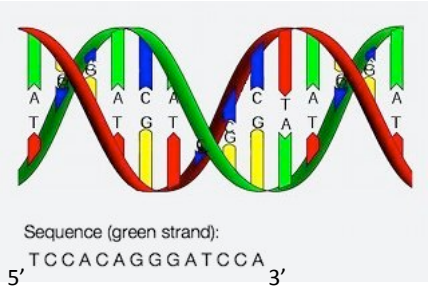
Errors in Illumina sequencing reads

- Error are mainly mismatches (substitutions)
- Error rates increase with increasing cycle number





DNA Sequence: Ambiguity codes



Code	Represents	Complement
A	A (Adenosine)	T
G	G (Guanine)	C
C	C (Cytosine)	G
T	T (Thymidine)	A
U	U (Uracil)	A
M	A or C	K
R	A or G	Y
W	A or T	W
S	C or G	S
Y	C or T	R
K	G or T	M
V	A or C or G	B
H	A or C or T	D
D	A or G or T	H
B	C or G or T	V
N	unknown	N
X	unknown	X

Sanger 1980 Nobel Prize

Basecall quality scores (“Phred” scores)

A quality score (or Q-score) expresses the probability that a basecall is incorrect. Given a basecall, A:

- The estimated probability that A is not correct is P(~A);
- The quality score for A is Q (A) = -10 $\log_{10}$ (P(~A))

A quality score of 10 means a probability of 0.1 that A is the wrong basecall.

P(~A) is platform-specific; Q-scores can be compared across platforms.

Quality scores are logarithmic:

Q-score	Error probability
10	0.1
20	0.01
40	0.0001

Quality Scores

$$Q = -10 \log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$

Phred Quality Scores	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Quality score encoding in FASTQ format

```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
BBBBCCCC?<A?BC?7@@???????DBBA@@@@A@@
```

Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35

Phred* quality score *Q* with base-calling error probability P

$Q = -10 \log_{10} P$

* Name of first program to assign accurate base quality scores. From the Human Genome Project.

Q score	Probability of base error	Base confidence	Sanger-encoded (Q Score + 33) ASCII character
10	0.1	90%	"+"
20	0.01	99%	"5"
30	0.001	99.9%	"?"
40	0.0001	99.99%	"I"

```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
  ..............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
  LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
   |                         |    |        |                              |                     |
 33                        59   64       73                            104                   126

 S - Sanger        Phred+33  range: 0 to 40
 I - Illumina 1.3+ Phred+64  range: 0 to 40
 L - Illumina 1.8+ Phred+33  range: 0 to 41
```

FASTQ encoding variation

- Various flavors:
  - fastq-sanger
  - fastq-illumina
  - fastq-solexa

```
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
*-+*''))**55CCF>>>>>>CCCC
```

Differing in the format of the sequence identifier and in the valid range of quality scores. See:

http://en.wikipedia.org/wiki/FASTQ_format

"…the Sanger version of the FASTQ format has found the broadest acceptance, supported by many assembly and read mapping tools …Therefore, most users will do this conversion very early in their workflows…"

Sequence read lengths remain limiting

Chr1: 249 Mb

249 Mb sequencing read

Current platforms:

- Sanger: A very small number (1-10,000) reads (700-1000 bp) but lowest error rates

- Illumina: A very large number (2 billion) of short reads (75-200 bp) but error rate 0.1-1%
- PacBio: A moderate number (~500,000) of long reads (~10 kb) but error rate as high as 14%

- For most applications reads are **aligned** to a reference genome
- Short reads contain inherently limited information
- *De novo* assembly of short reads is difficult

So now the game is …

If you can convert it to DNA, you can sequence it.