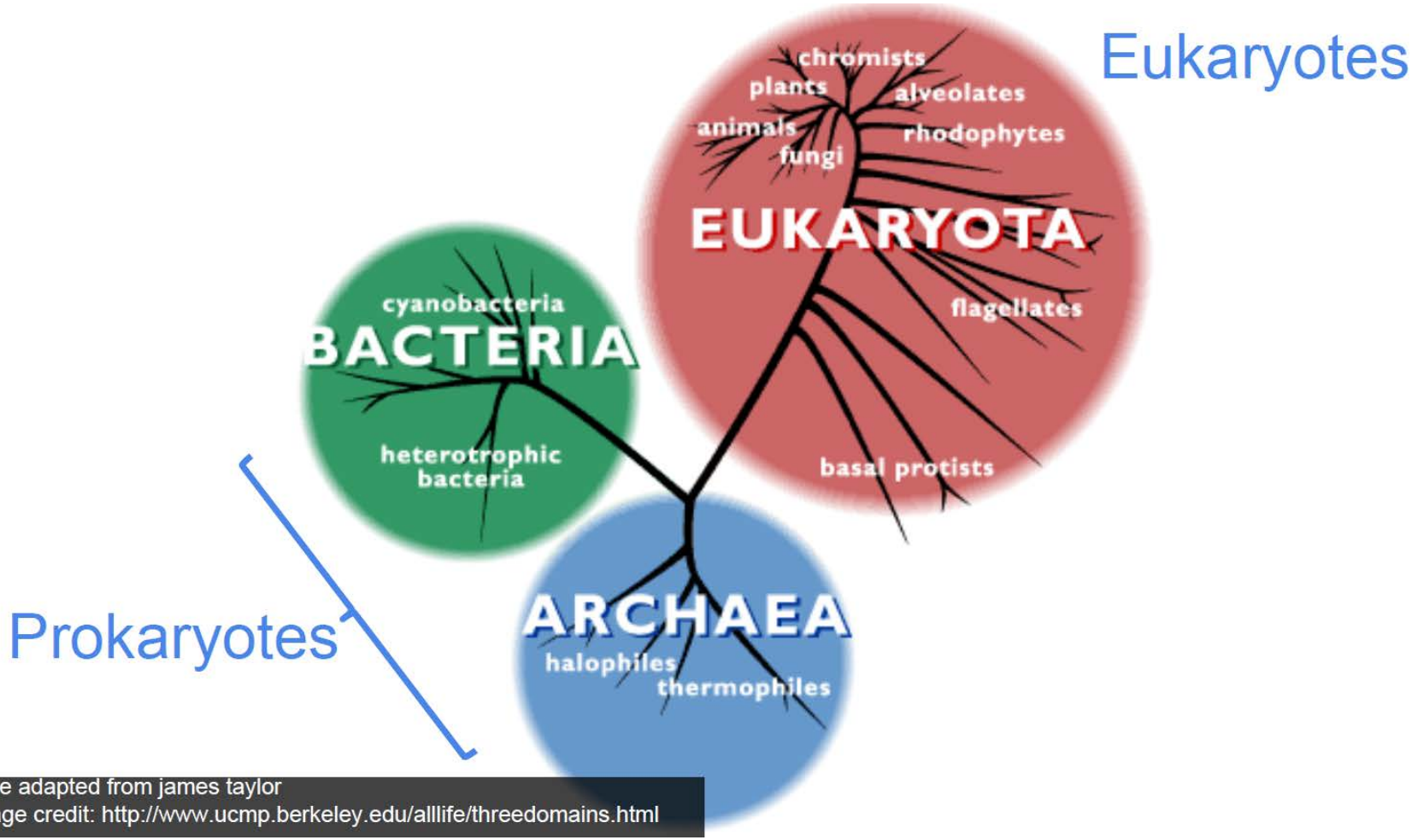
DNA/RNA sequencing

Computational Biology II

Credit to Coursera, Genomic data science specialization

Basic domains of life



slide adapted from james taylor
image credit: http://www.ucmp.berkeley.edu/alllife/threedomains.html

Basic types of cells

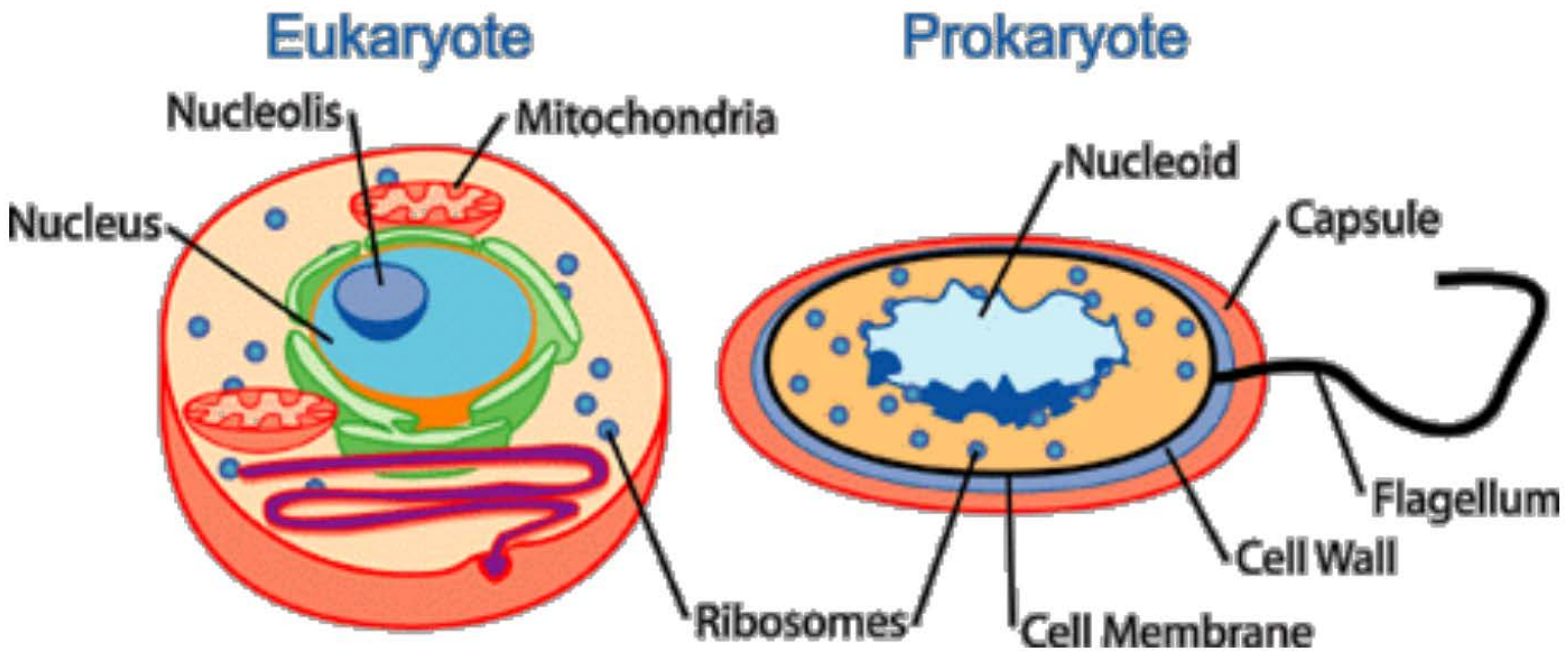


image credit: http://commons.wikimedia.org/wiki/File:Celltypes.png
slide adapted from james taylor

Eukaryotic cells

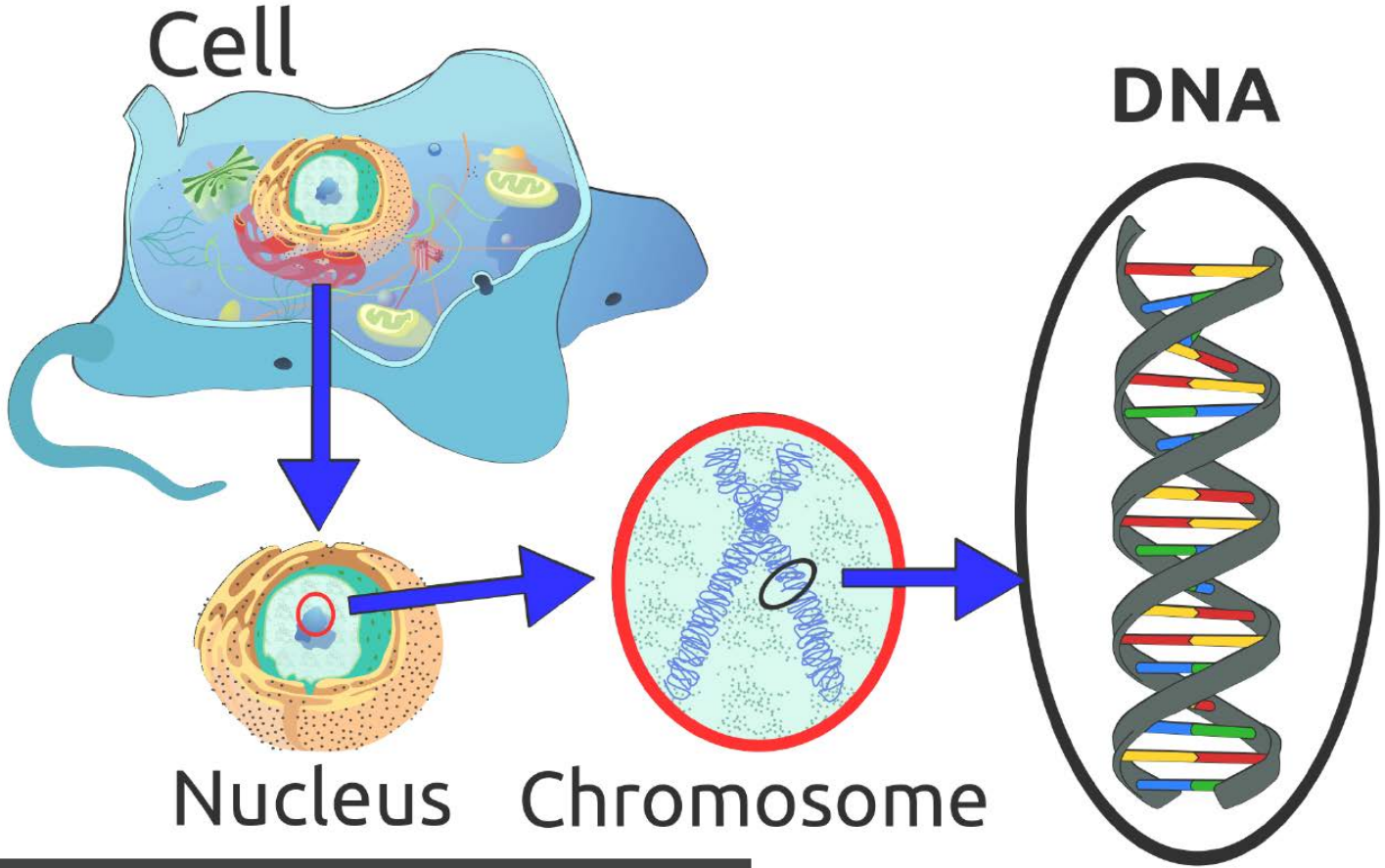


image credit: http://en.wikipedia.org/wiki/DNA

DNA nucleotides

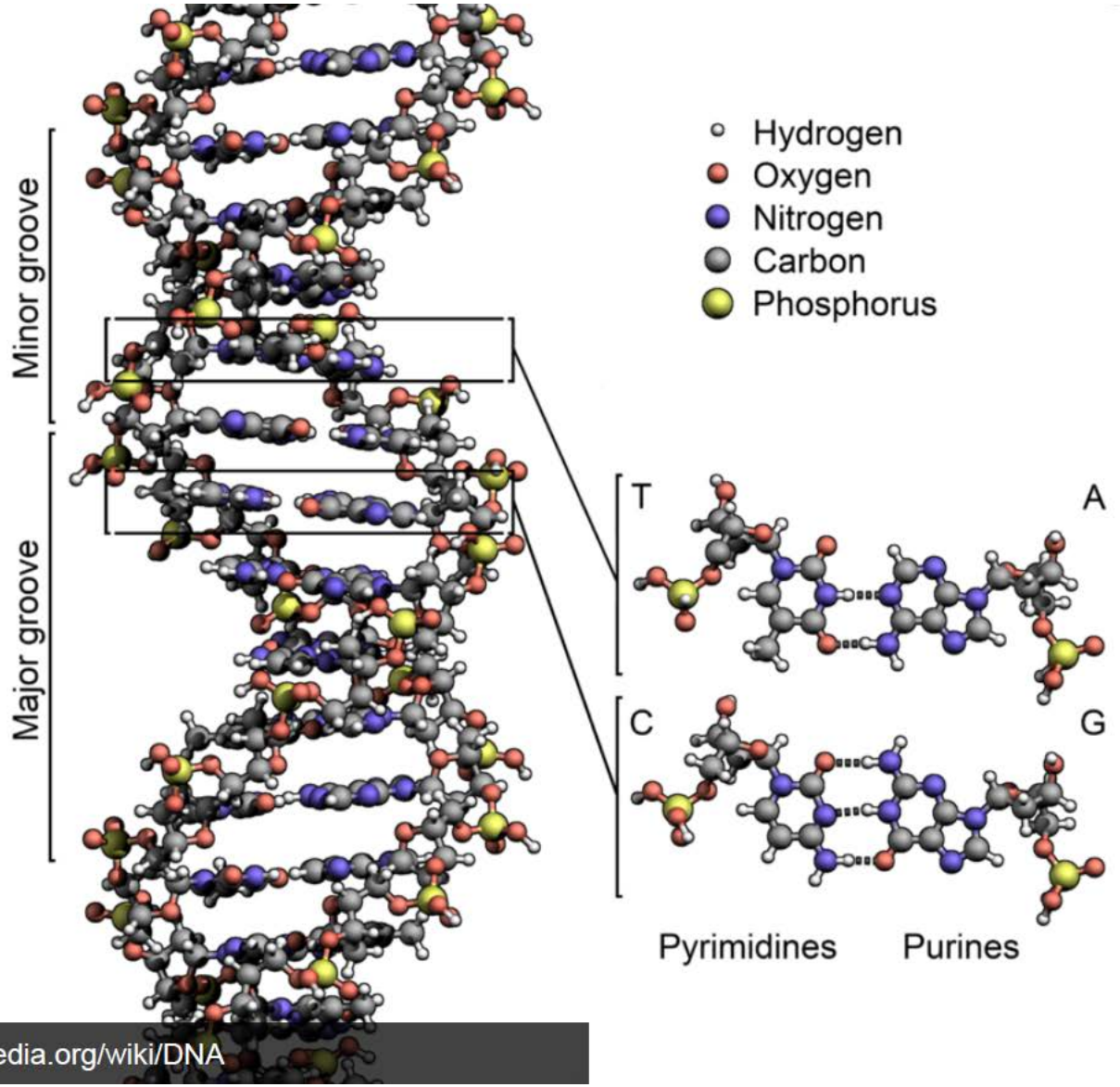


image credit: http://en.wikipedia.org/wiki/DNA

DNA structure

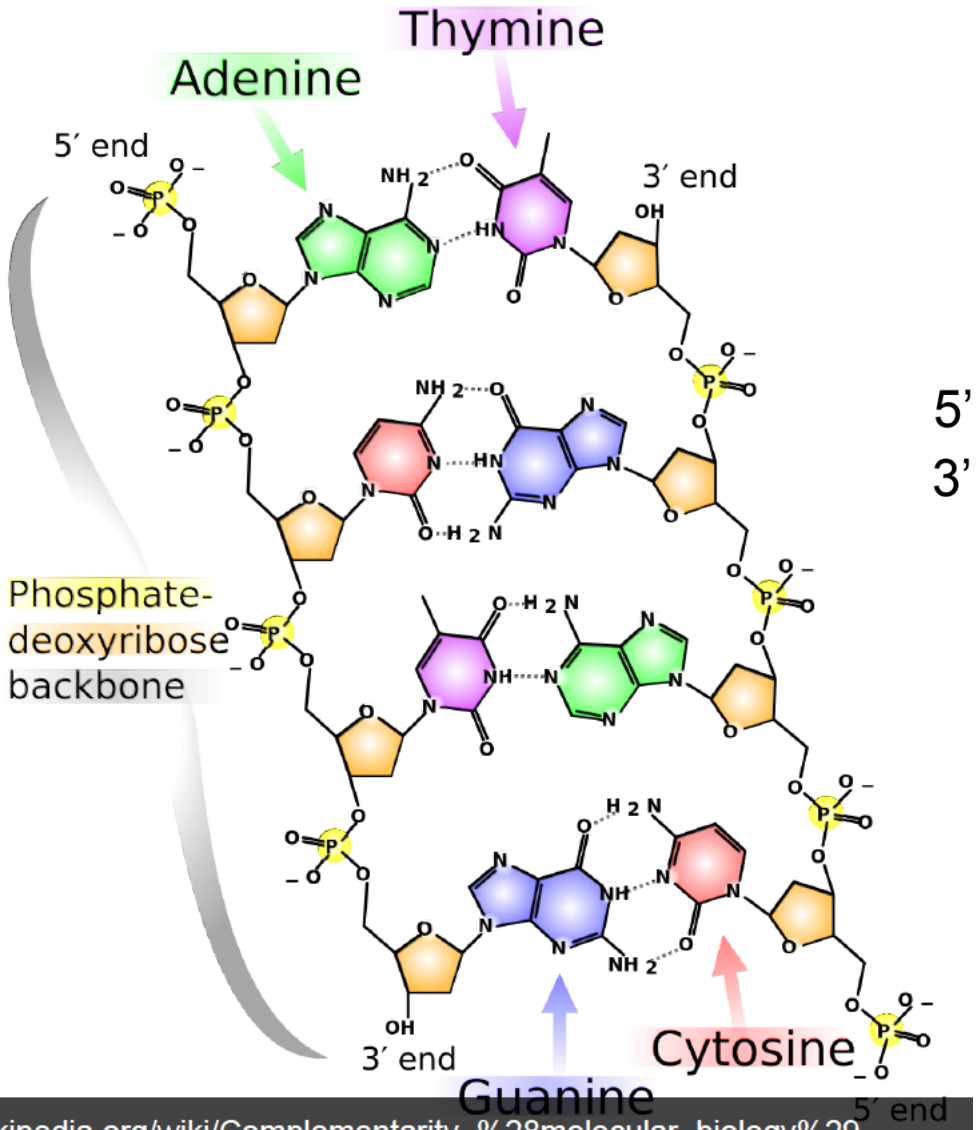


image credit: http://en.wikipedia.org/wiki/Complementarity_%28molecular_biology%29

"+" strand

5' - ACACCGGTT - 3'

3' - TGTGGCCAA - 5'

"-" strand

RNA structure

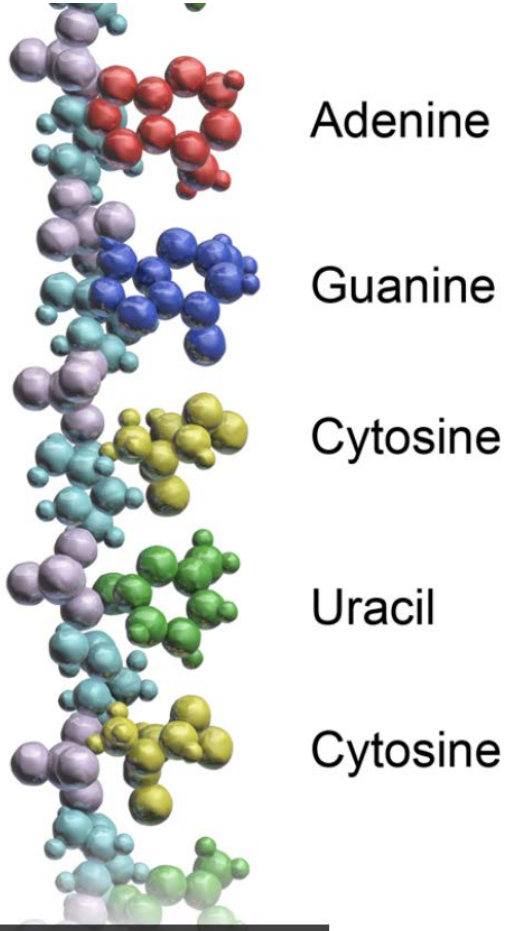
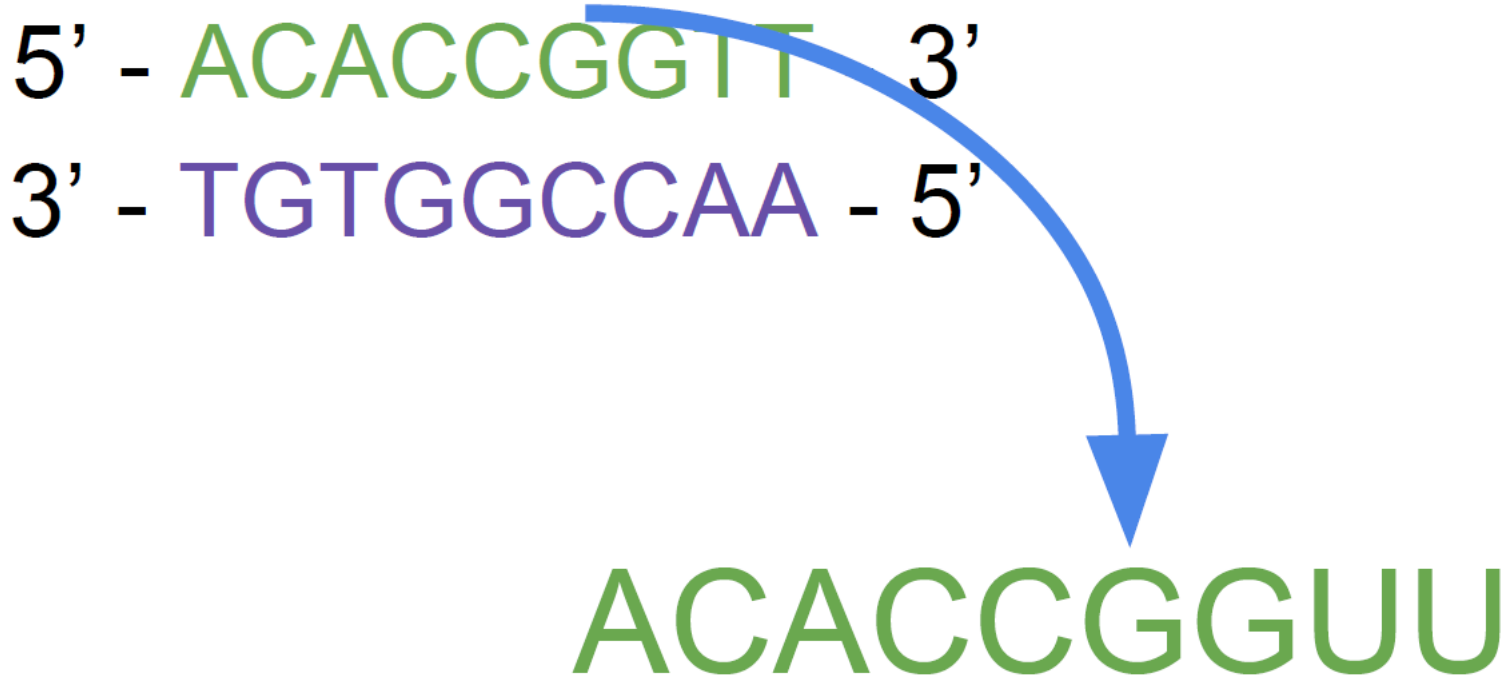


image credit: http://en.wikipedia.org/wiki/RNA

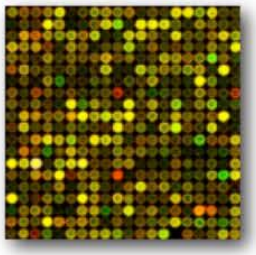
Single stranded nucleic acid

RNA sequence



DNA sequencing

Sanger DNA sequencing 1977-1990s



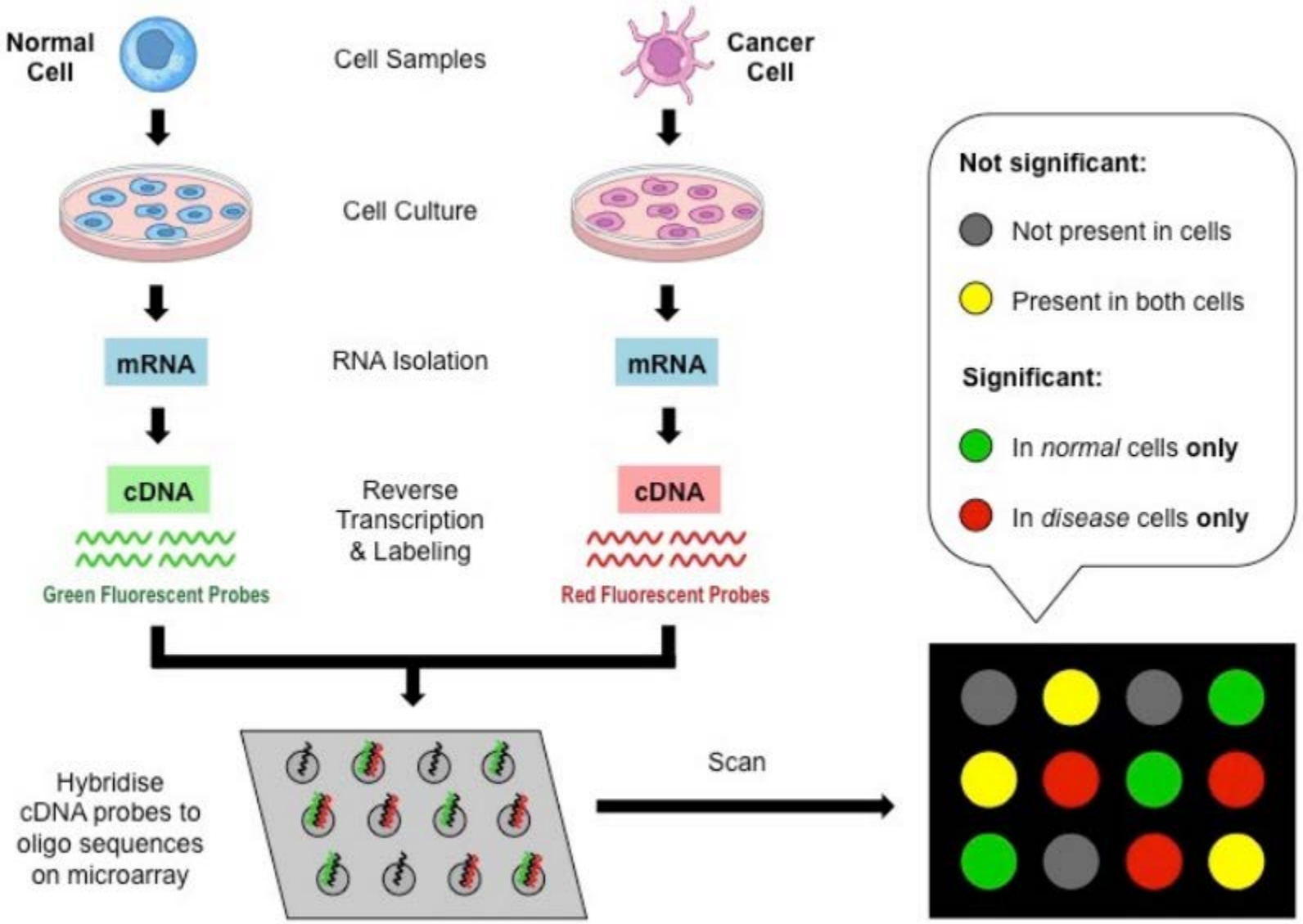
DNA Microarrays Since mid-1990s

2nd Generation DNA Sequencing Since ~2007

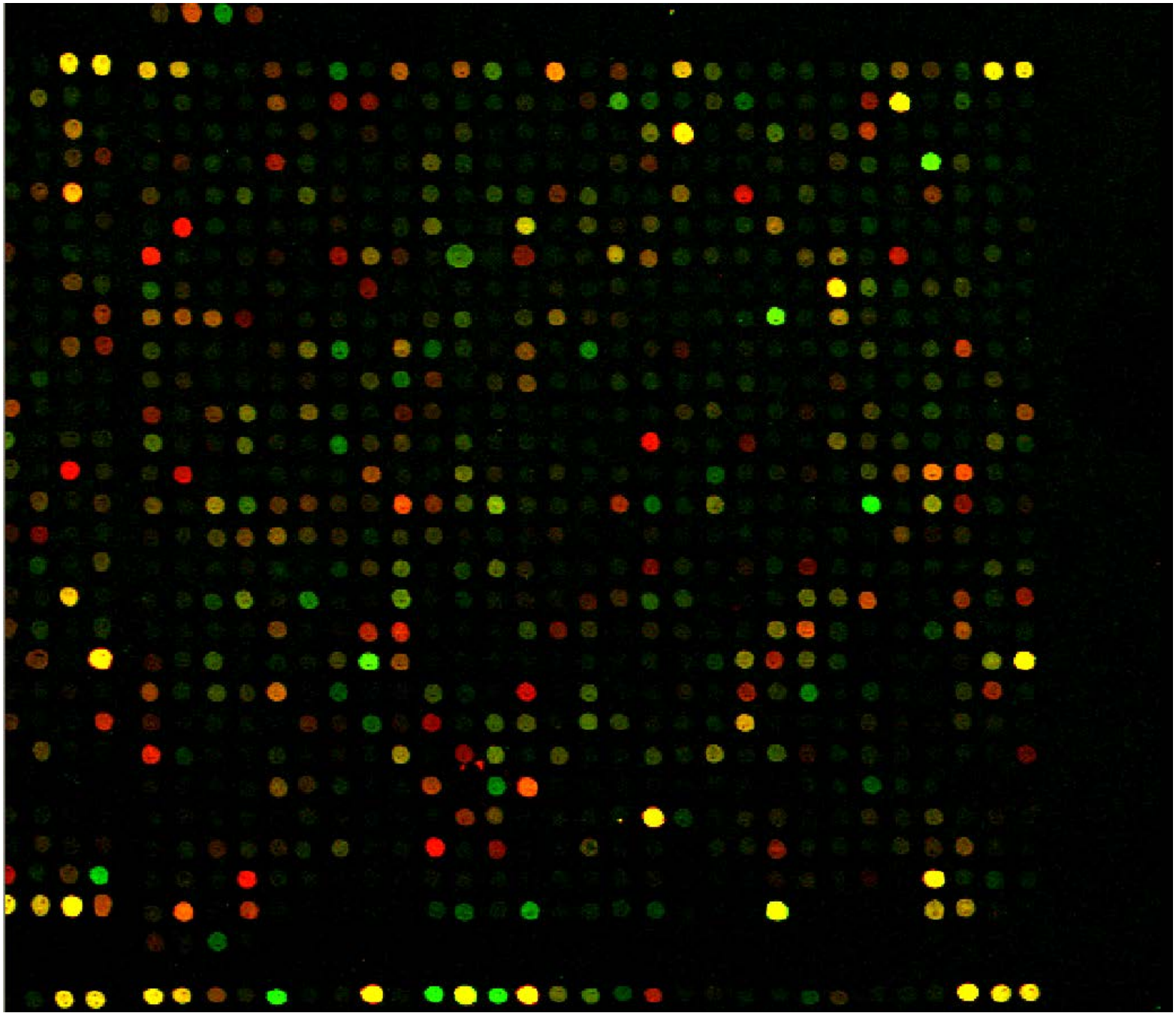
3rd Generation & single molecule Sequencing Since ~ 2010

Slide adapted from Ben Langmead

Microarrays

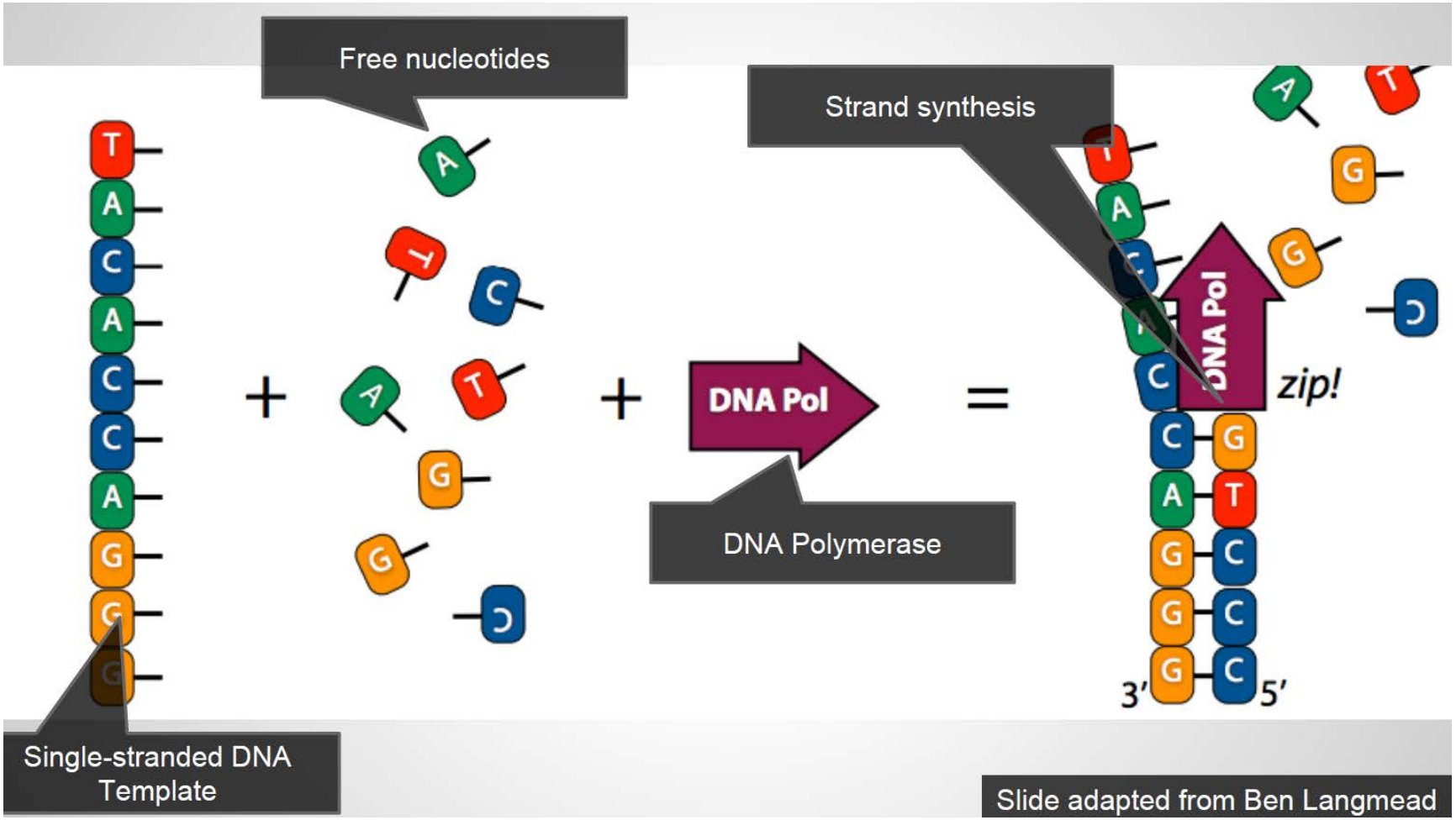


BioNinja

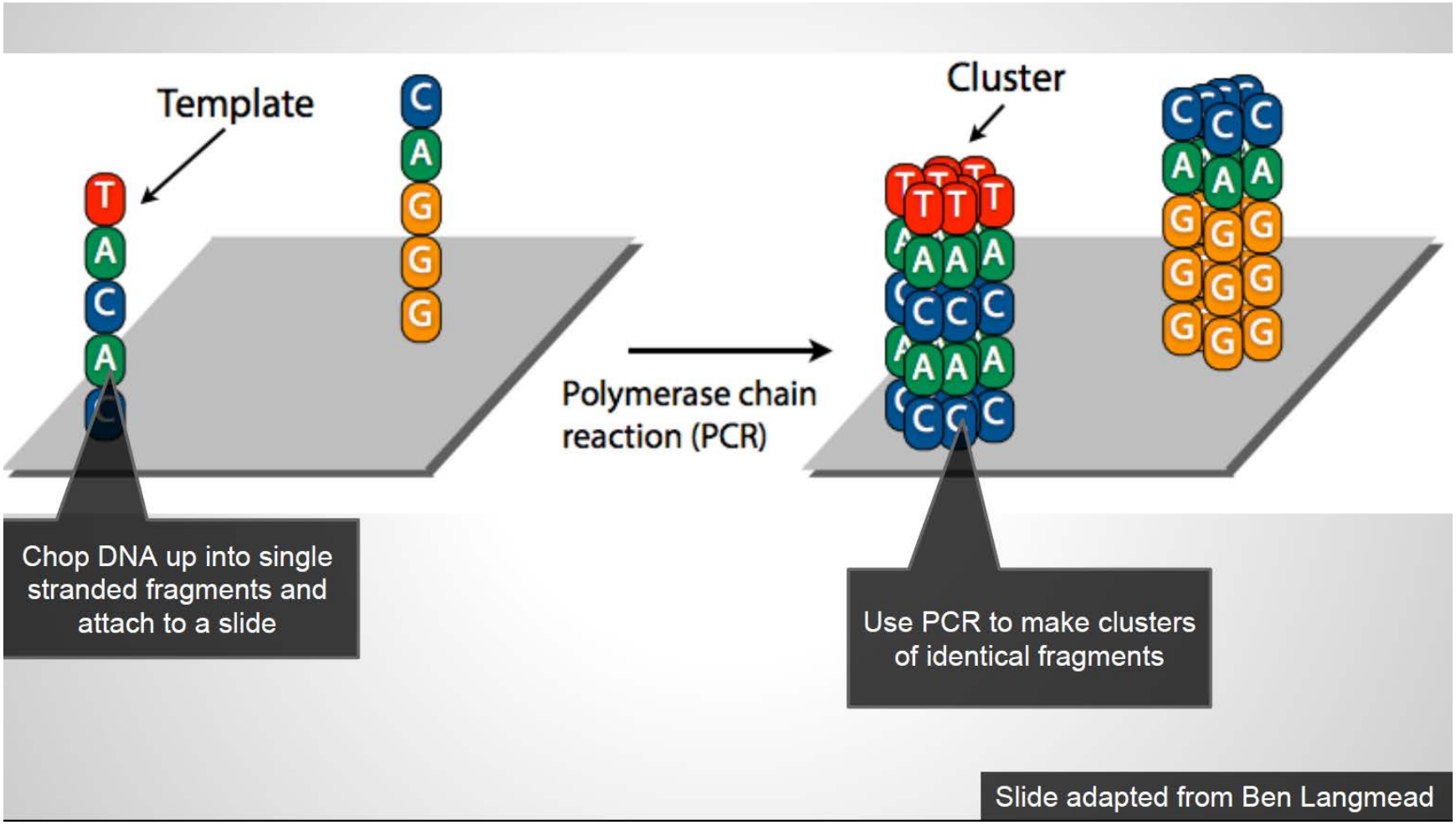


www.microarray.org

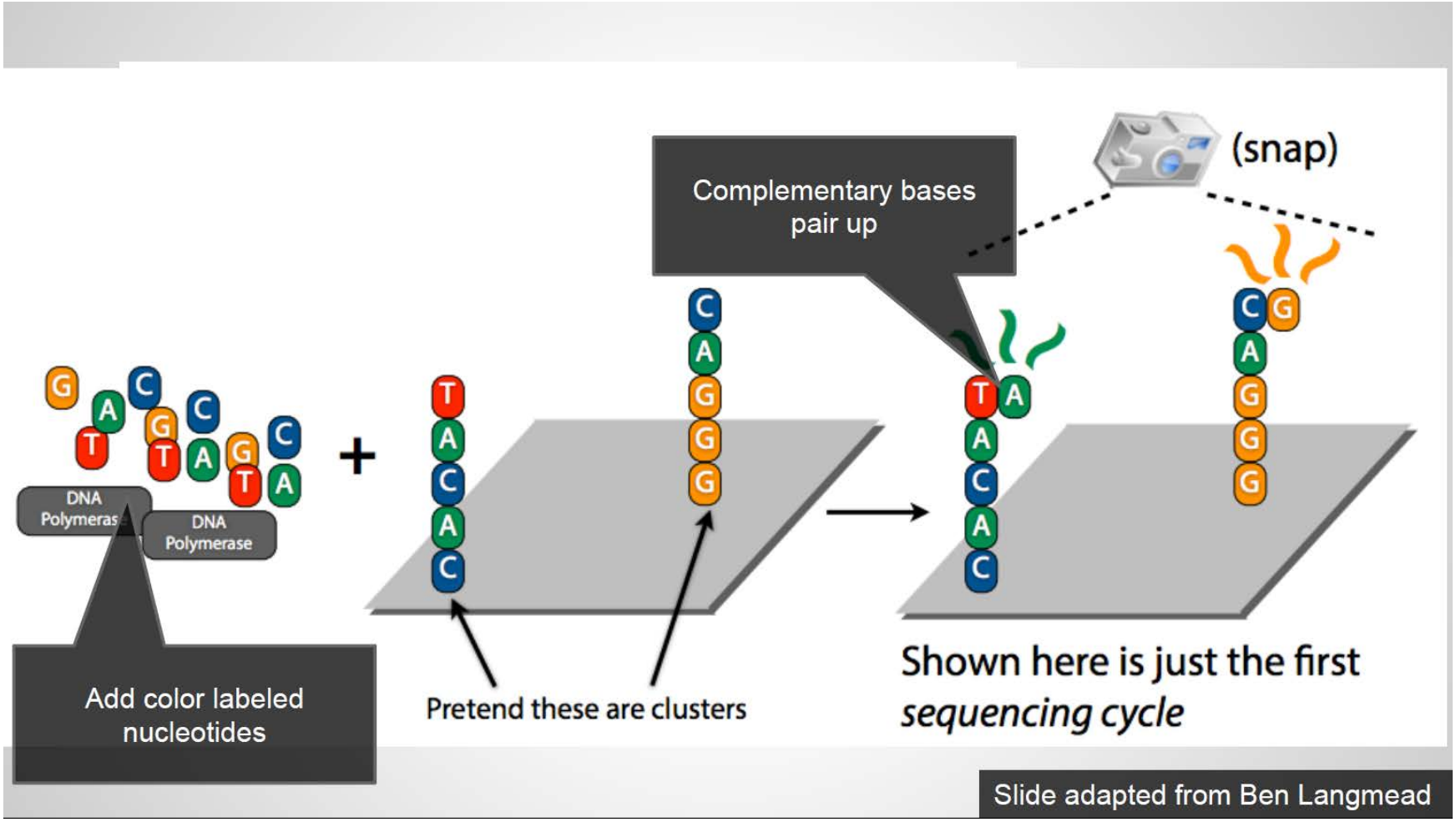
Polymerase chain reaction (PCR)



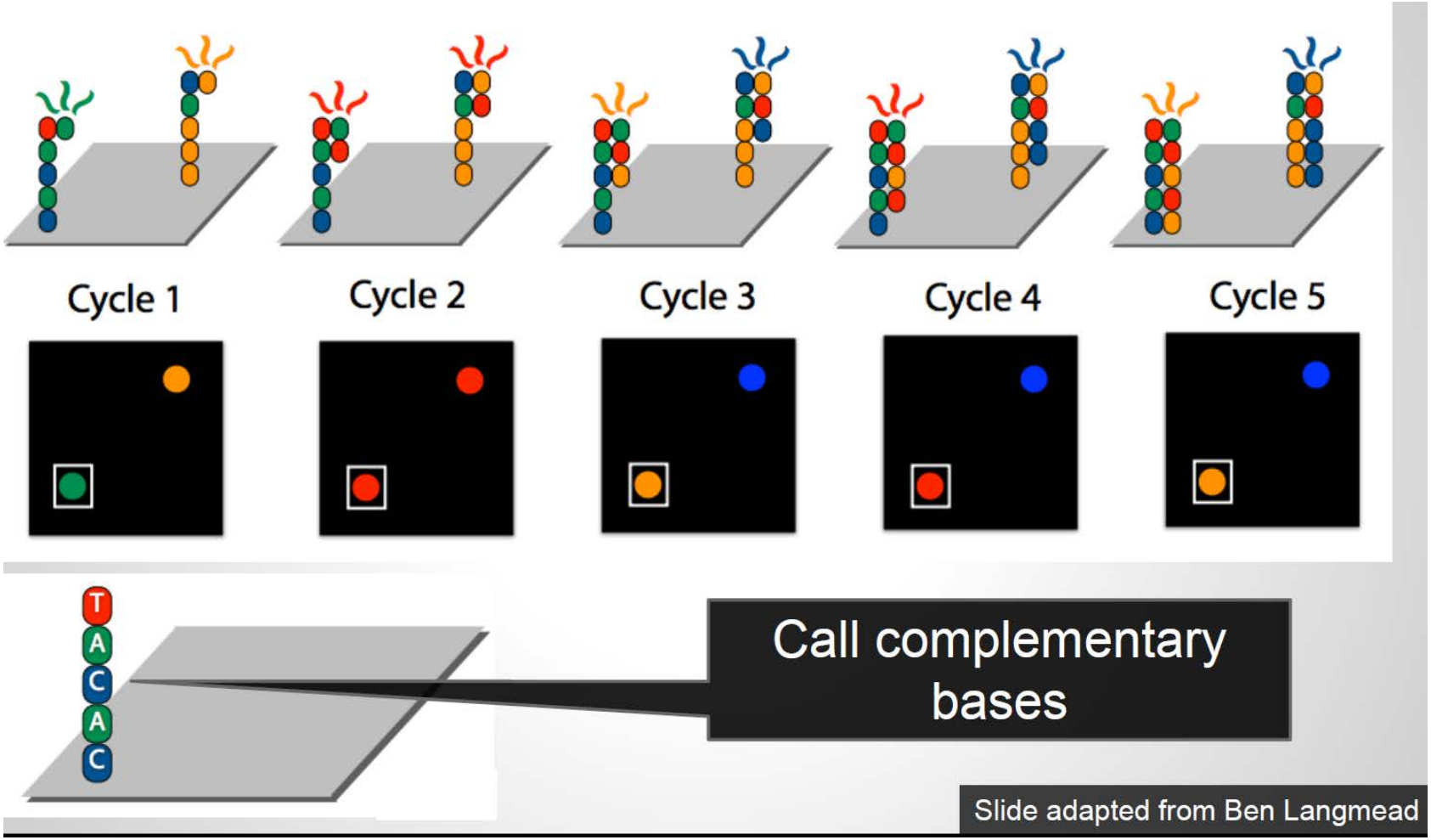
Polymerase chain reaction (PCR) II



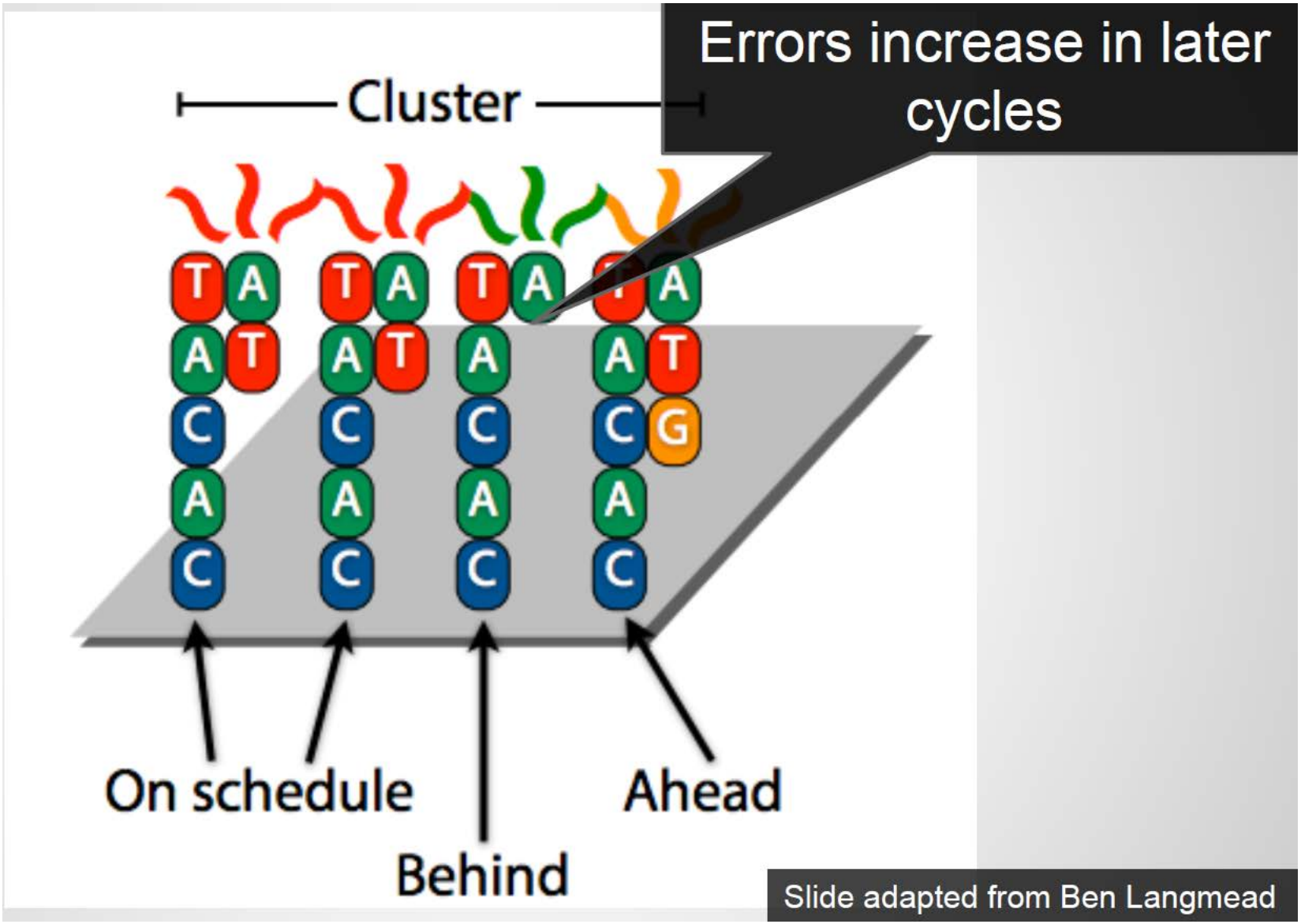
Next generation sequencing



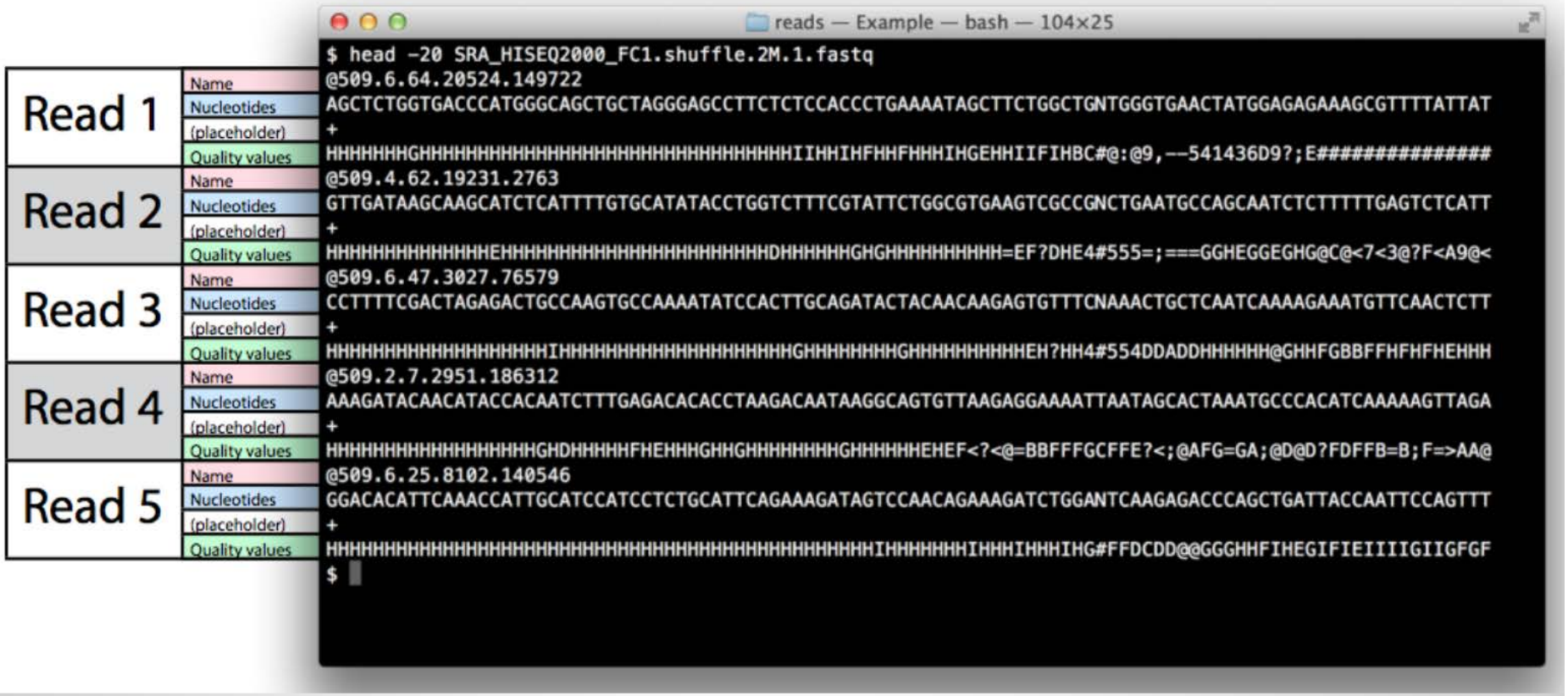
Sequencing cycles



Sequencing errors



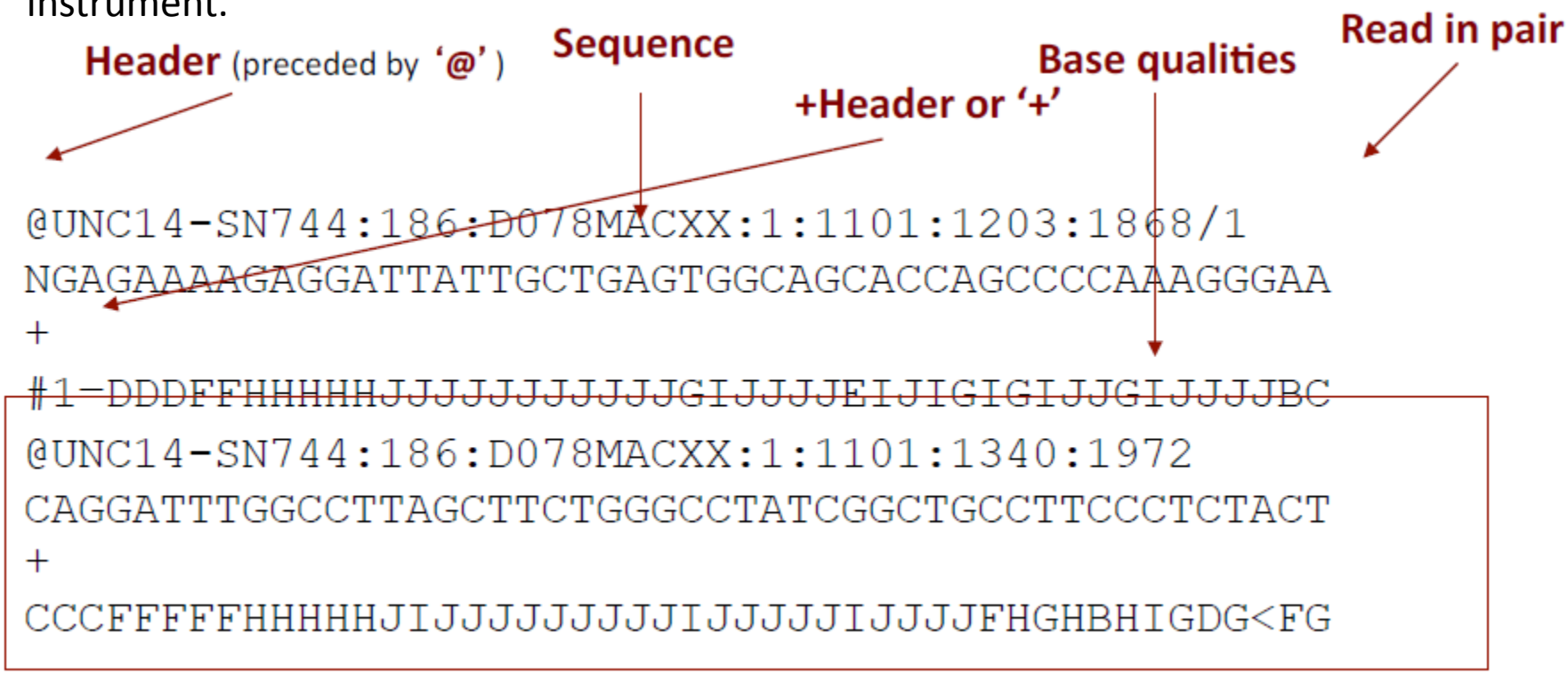
Output FASTQ file



Slide adapted from Ben Langmead

Output FASTQ file II

Usually these identifiers give information about the precise location and geometry of the cluster within the sequencing instrument.



Base quality score

- Let p_b = probability that the call at base *b* is incorrect
- Quality value: $Q_{sanger} = -10 \log_{10} p_b$ (integer)
- Sanger (Phred quality scores): 0..93 (ASCII characters 33..126)

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`
abcdefghijklmnopqrstuvwxyz{|}~
```

- In practice, the maximal quality value is ~40.
- Quality values below 20 are typically considered low.

Representation: Single sequences, Fasta/Pearson format

- Fasta/Pearson – the most common format for Sanger sequences, and for genomic and gene sequences

Header (1 line, starts with '>')

```
>gi|23238195|ref|NM_000878.2| Homo sapiens interleukin 2 receptor, beta
(IL2RB), mRNA
GCAGCCAGAGCTCAGCAGGGCCCTGGAGAGATGGCCACGGTCCCAGCACCGGGGAGGACTGGAGAGCGCG
CGCTGCCACCGCCCCATGTCTCAGCCAGGGCTTCCTTCCTCGGCTCCACCCTGTGGATGTAATGGCGGCC
CCTGCTCTGTCCTGGCGTCTGCCCCTCCTCATCCTCCTCCTGCCCCTGGCTACCTCTTGGGCATCTGCAG
CGGTGAATGGCACTTCCCAGTTCACATGCTTCTACAACTCGAGAGCCAACATCTCCTGTGTCTGGAGCCA
AGATGGGGCTCTGCAGGACACTTCCTGCCAAGTCCATGCCTGGCCGGACAGACGGCGGTGGAACCAAACC
TGTGAGCTGCTCCCCGTGAGTCAAGCATCCTGGGCCTGCAACCTGATCCTCGGAGCCCCAGATTCTCAGA
AACTGACCACAGTTGACATCGTCACCCTGAGGGTGCTGTGCCGTGAGGGGGTGCGATGGAGGGTGATGGC
CATCCAGGACTTCAAGCCCTTTGAGAACCTTCGCCTGATGGCCCCCATCTCCCTCCAAGTTGTCCACGTG
GAGACCCACAGATGCAACATAAGCTGGGAAATCTCCCAAGCCTCCCACTACTTTGAAAGACACCTGGAGT
TCGAGGCCCGGACGCTGTCCCCAGGCCACACCTGGGAGGAGGCCCCCCTGCTGACTCTCAAGCAGAAGCA
GGAATGGATCTGCCTGGAGACGCTCACCCCAGACACCCAGTATGAGTTTCAGGTGCGGGTCAAGCCTCTG
CAAGGCGAGTTCACGACCTGGAGCCCCTGGAGCCAGCCCCTGGCCTTCAGGACAAAGC
```

Sequence (nucleotide or protein) (1 or several lines)

Representation: Multiple sequences, Multi-Fasta format

>gi|28178860|ref|NM_000586.2| Homo sapiens interleukin 2 (IL2), mRNA

CGAATTCCCCTATCACCTAAGTGTGGGCTAATGTAACAAAGAGGGATTTCACCTACATCCATTCAGTCAG
TCTTTGGGGGTTTAAAGAAATTCCAAAGAGTCATCAGAAGAGGAAAAATGAAGGTAATGTTTTTTCAGAC
AGGTAAAGTC. . . . . ATAAAAAAAAAAAA

>gi|31982837|ref|NM_008366.2| Mus musculus interleukin 2 (Il2), mRNA

ATCACCCTTGCTAATCACTCCTCACAGTGACCTCAAGTCCTGCAGGCATGTACAGCATGCAGCTCGCATC
CTGTGTCACATTGACACTTGTGCTCCTTGTCAACAGCGCACCCACTTCAAGCTCCACTTCAAGCTCTACA
GCGGAAGCACAGCAGCAGCAGCAGCAGCAGCAG . . . . . AGCTCTCCTCT

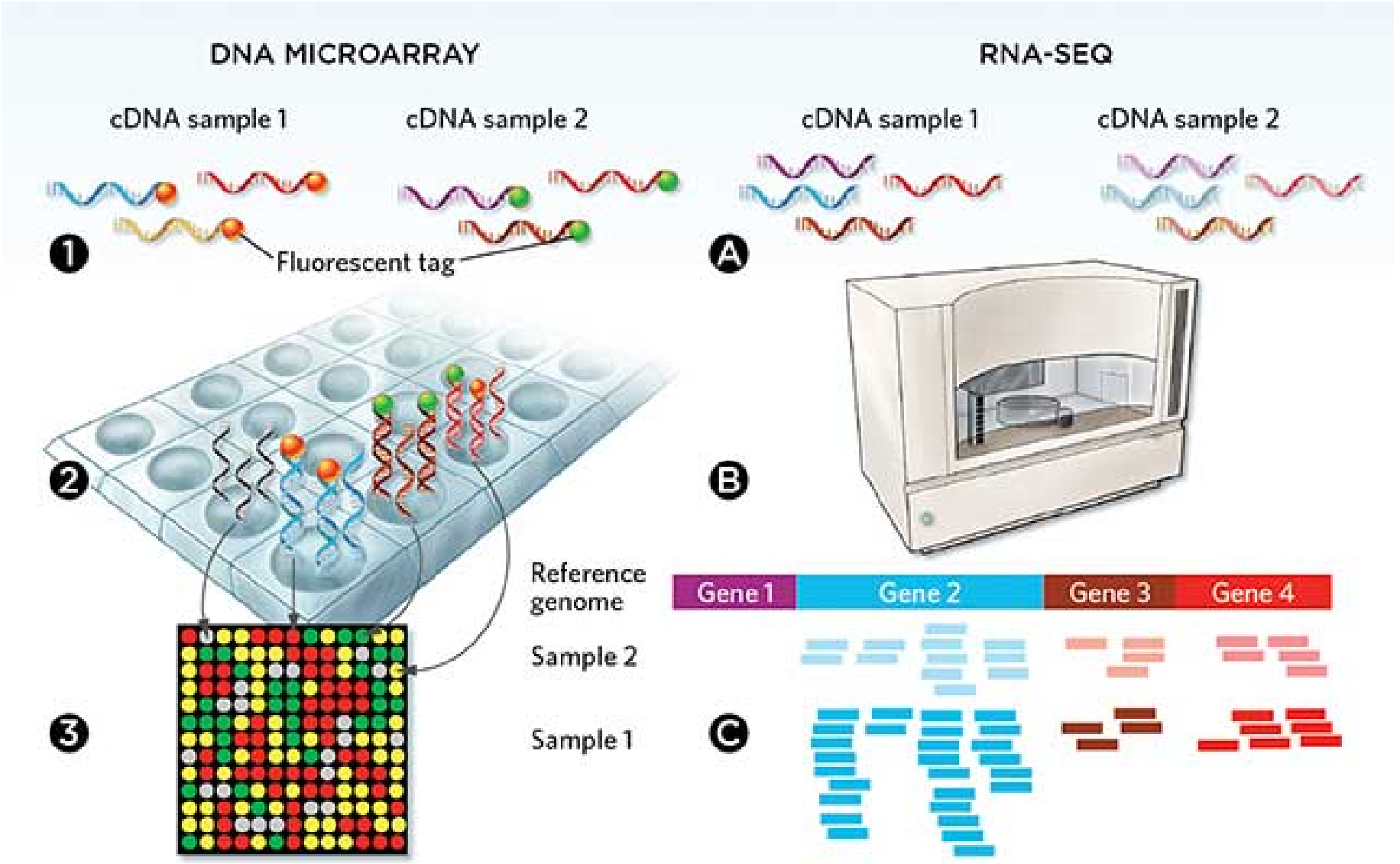
>gi|16758691|ref|NM_053836.1| Rattus norvegicus interleukin 2 (Il2), mRNA

GAAGTCCTGCAAGCATGTACAGCATGCAGCTCGCATCCTGTGTTGCACTGACGCTTGTCCTCCTTGTCAA
CAGCGCACCCACTTCAAGCCCTGCAAAGGAAACACAGCAGCACCTGGAGCAGCTGTTGCTGGACTTACAG
GTGCTCCTGAGAGGGATC . . . . . AGTATTTAGAAGAGTCGATGAA

>gi|30794289|ref|NM_180997.1| Bos taurus interleukin 2 (IL2), mRNA

GGGCTATCTGTTCGGTCGTTCATGTCAGCAATGTACAAGATACAACTCTTGTCTTGCATTGCACTAACTC
TTGCACTCGTTGCAAACGGTGCACCTACTTCAAGCTCTACGGGGAACACAATGAAAGAAGTGAAGTCATT
GCTGCTGGATTTACAGTTGCTTTTGGAGAAAGTTAAAAATCCTGAGAACCTCAAGCTCTCCAGGATGCAT
. . . . ATTTAATAAAGTTGATGAATAAAAAC

Summary



The Scientist