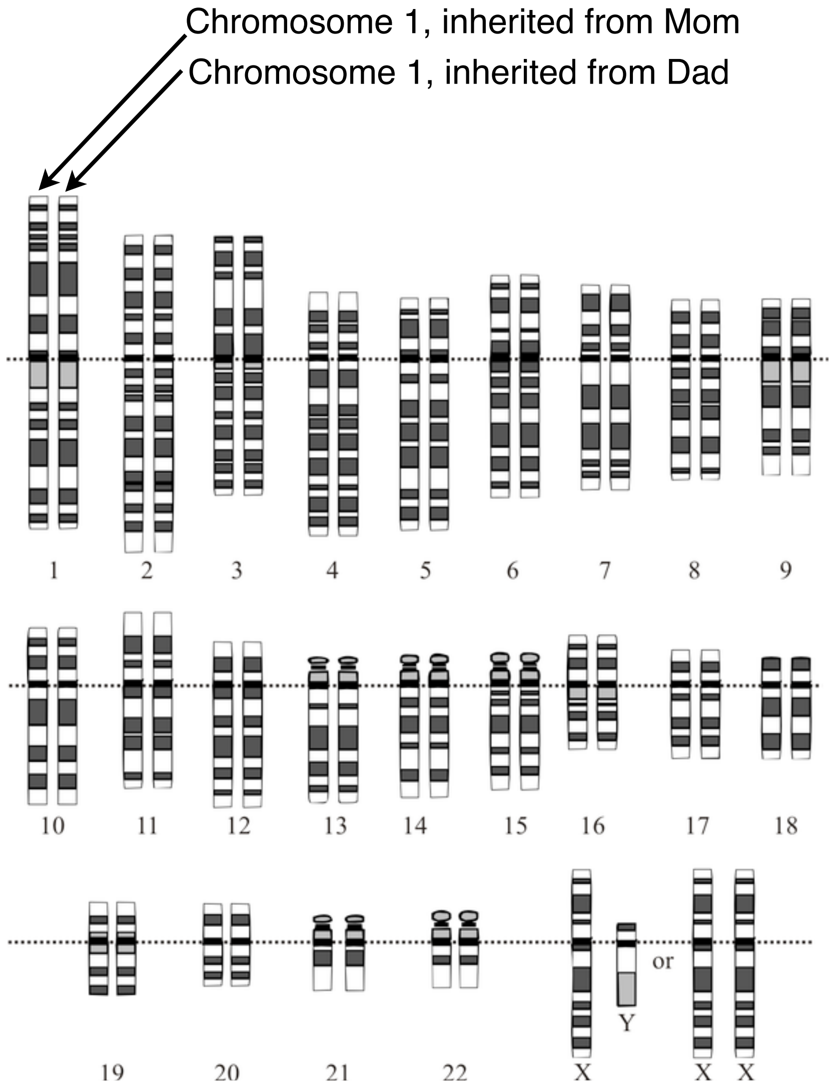
Basics of DNA & Sequencing by Synthesis

Lecture 2

1/11/18

Most slides courtesy of Ben Langmead at John Hopkins

The genome: where genotypes live


Chromosome 1, inherited from Mom
Chromosome 1, inherited from Dad
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
X
or
Y
X
X

Human chromosomes

23 pairs, 46 total
22 pairs are “autosomes”
1 pair are “sex chromosomes”

Genome is the entire DNA sequence of an individual; all chromosomes

Human genome is 3 billion nt long

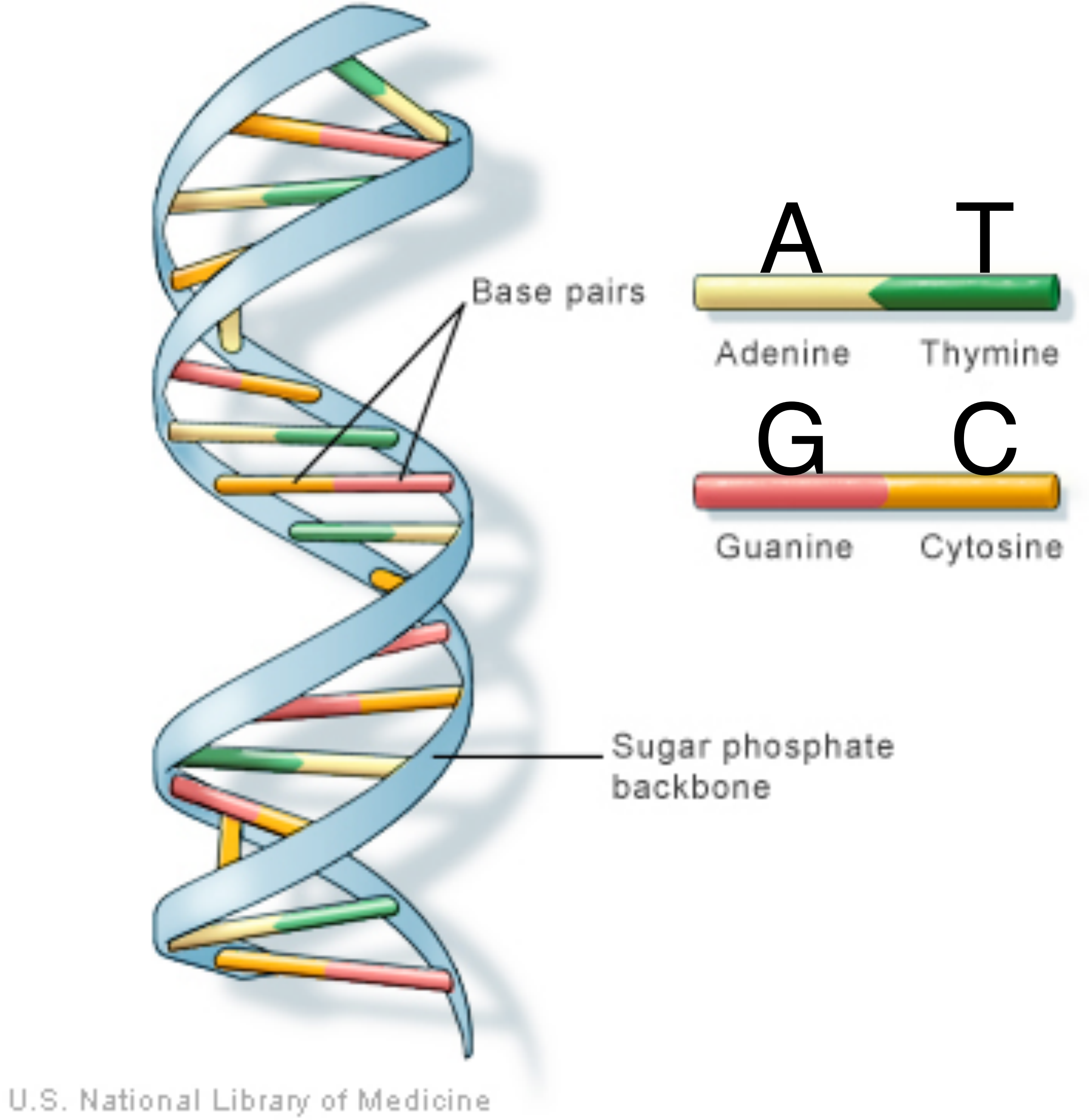
“nt” = nucleotides
similarly: “bp”

Most bacterial genomes are a few million nt. Most viral genomes are tens of thousands of nt. This plant’s genome is about 150 billion nt.

Paris japonica

Pictures: http://en.wikipedia.org/wiki/Chromosome, http://en.wikipedia.org/wiki/Paris_japonica

DNA: the genome's molecule



Picture: http://ghr.nlm.nih.gov/handbook/basics/dna

Deoxyribonucleic acid

"Rungs" of DNA double-helix are base pairs. Pair combines two complementary

Complementary pairings: A-T, C-G

Single base also called a "nucleotide"

DNA fragment lengths are measured in
"base pairs" (abbreviated bp),
"bases" (b) or "nucleotides" (nt)

Stringizing DNA

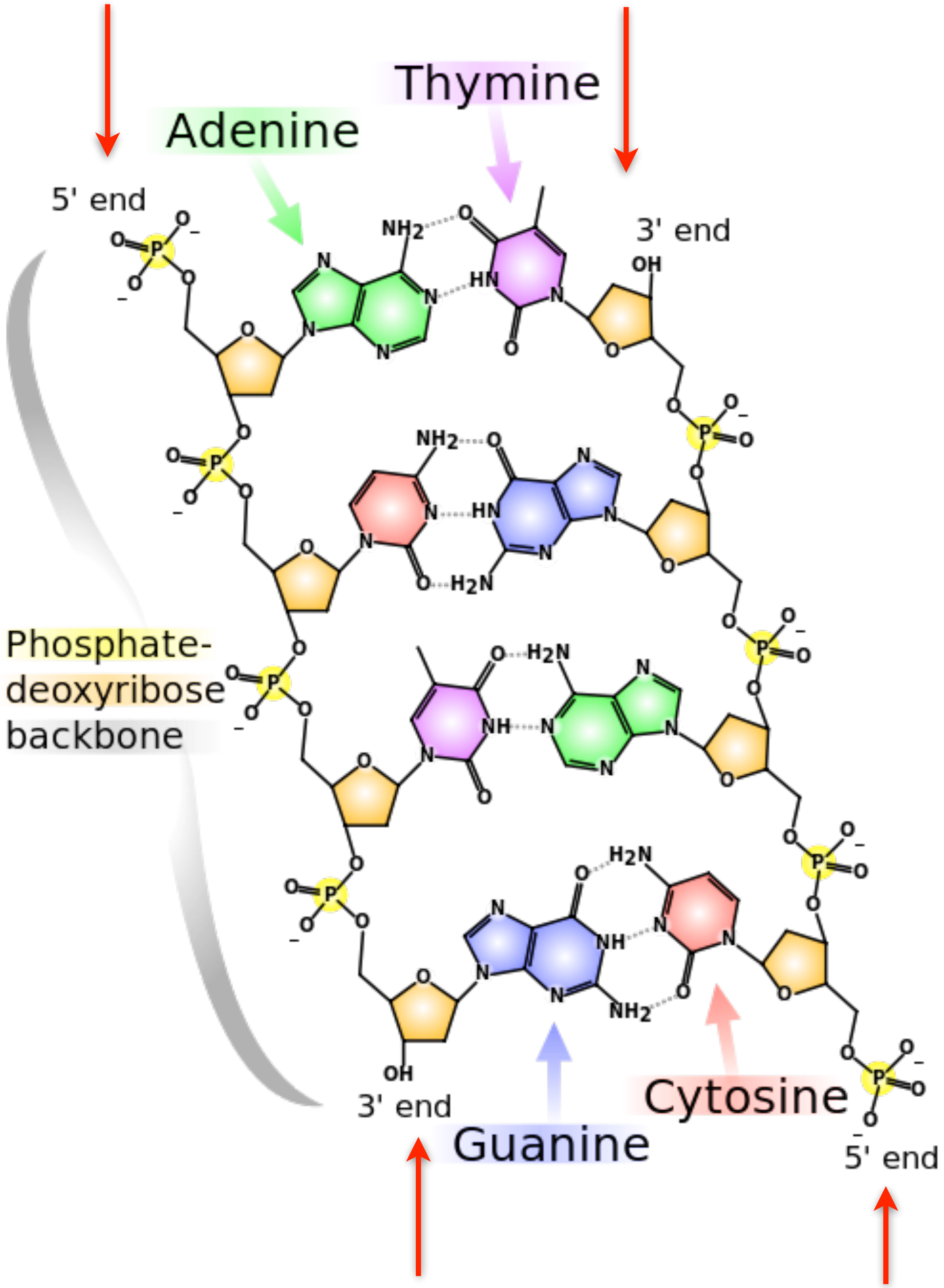
DNA has *direction* (a *5' head* and a *3' tail*). When we write a DNA *string*, we follow this convention.

When we write a DNA string, we write just one strand. The other strand is its *reverse complement*.

To get reverse complement, reverse then complement nucleotides
(i.e. interchange A/T and C/G)

5' end (head)

3' end (tail)



3' end (tail)

5' end (head)

Picture: http://en.wikipedia.org/wiki/DNA

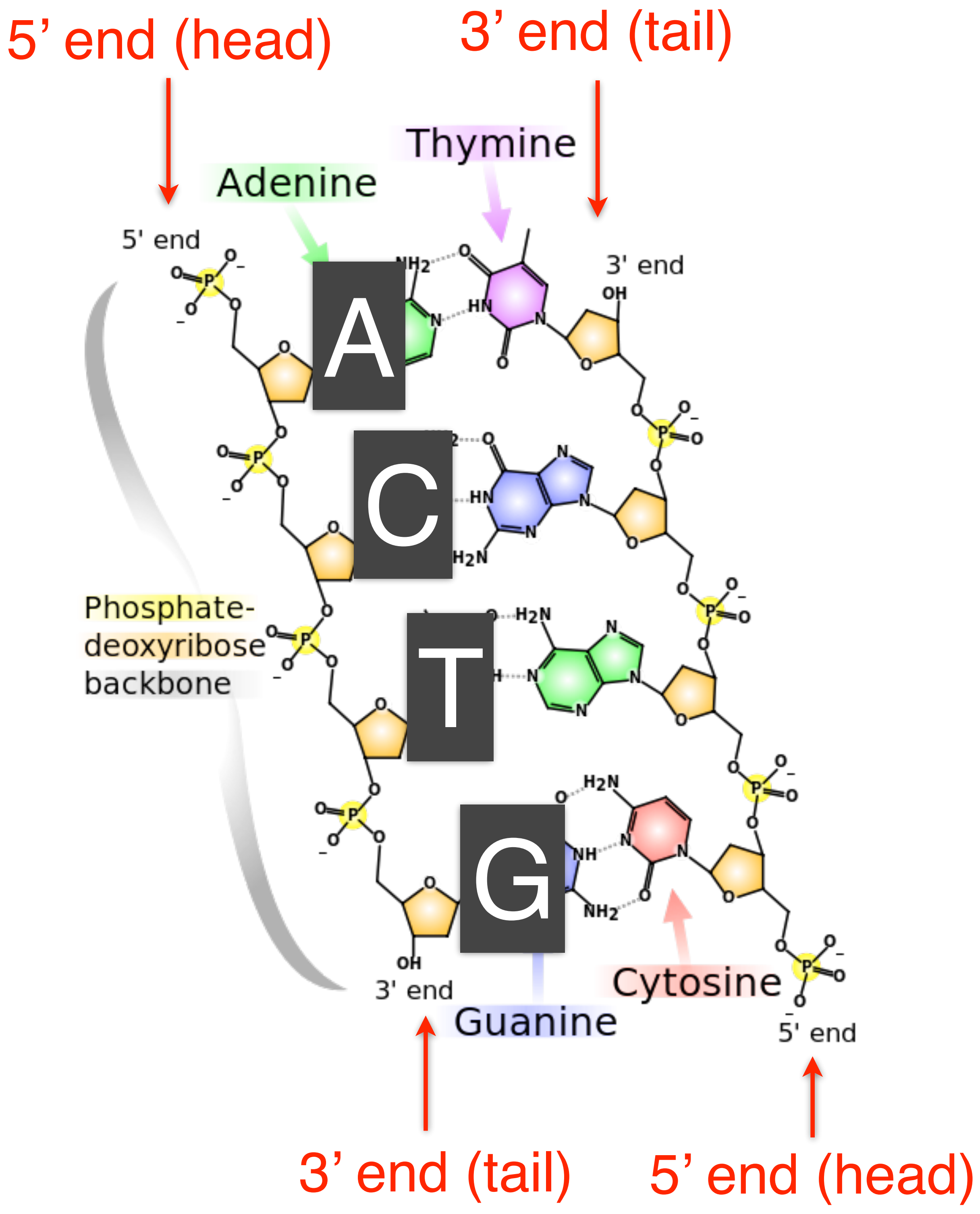
Stringizing DNA

DNA has *direction* (a *5' head* and a *3' tail*). When we write a DNA *string*, we follow this convention.

When we write a DNA string, we write just one strand. The other strand is its *reverse complement*.

To get reverse complement, reverse then complement nucleotides
(i.e. interchange A/T and C/G)

5' end A C T G *3' end*



Picture: http://en.wikipedia.org/wiki/DNA

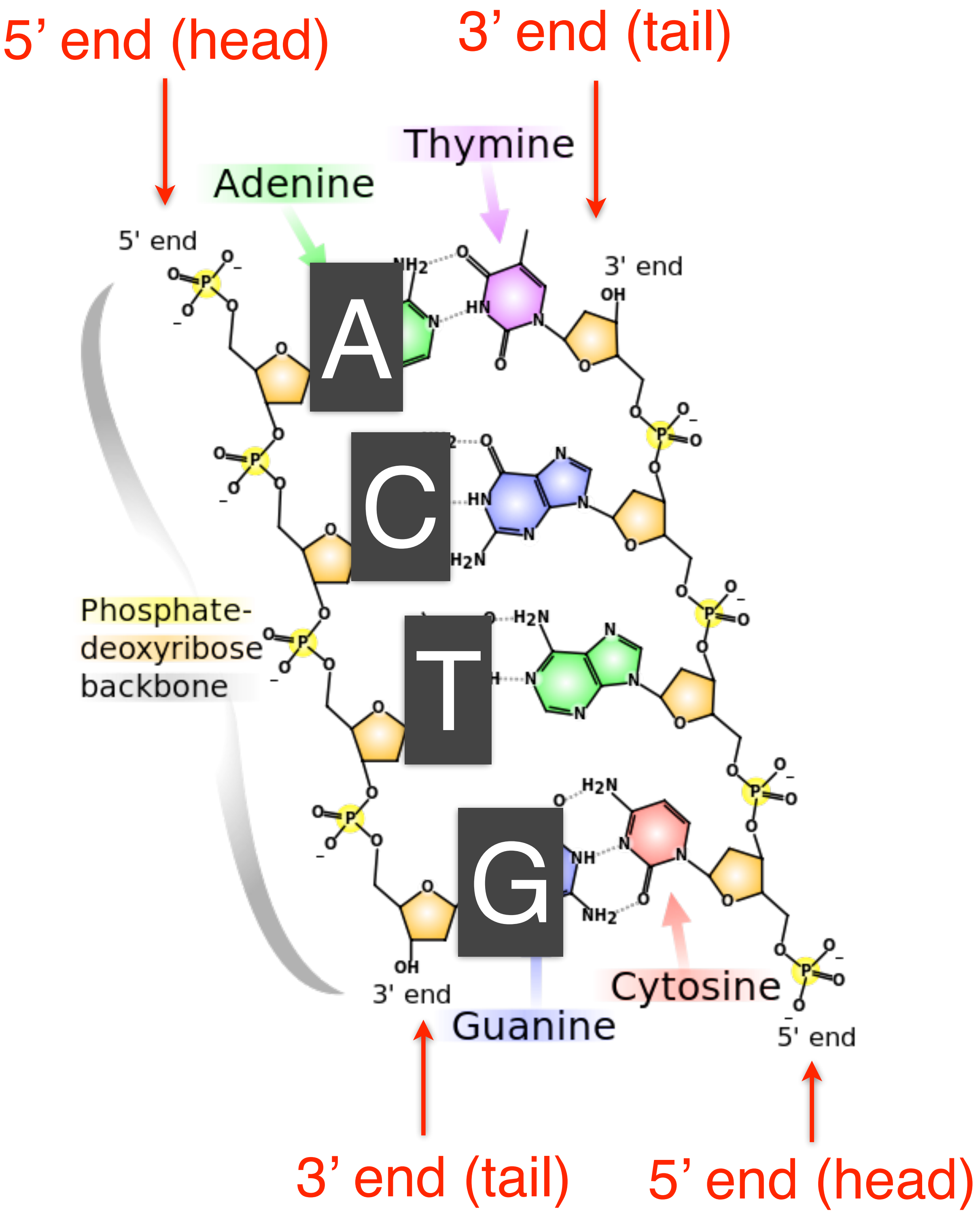
Stringizing DNA

DNA has *direction* (a *5’ head* and a *3’ tail*). When we write a DNA *string*, we follow this convention.

When we write a DNA string, we write just one strand. The other strand is its *reverse complement*.

To get reverse complement, reverse then complement nucleotides (i.e. interchange A/T and C/G)

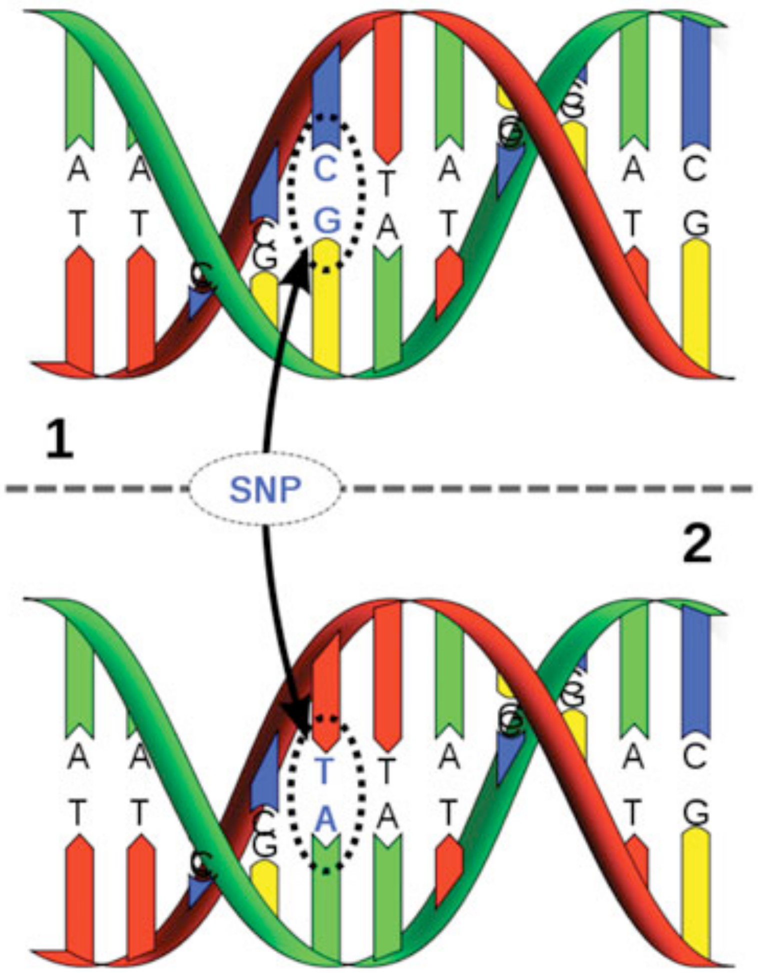
5’ end A C T G *3’ end*
↑
reverse complement
↓
5’ end C A G T *3’ end*



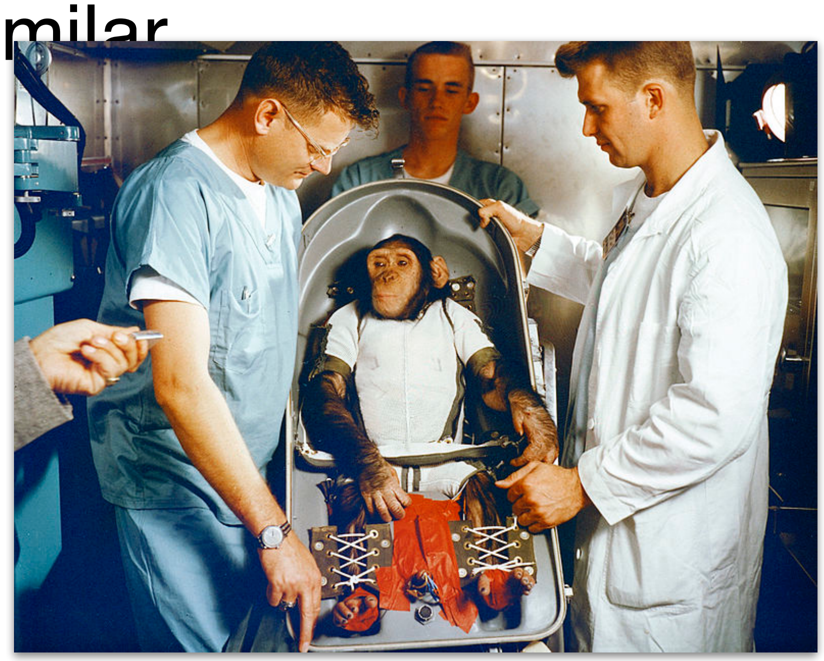
Picture: http://en.wikipedia.org/wiki/DNA

The genome: variation

Two unrelated humans have genomes that are ~99.8% similar by sequence. There are about 3-4 million differences. Most are small, e.g. Single Nucleotide Polymorphisms



Human and chimpanzee genomes are about 96% similar



Pictures: http://www.dana.org/news/publications/detail.aspx?id=24536, http://en.wikipedia.org/wiki/Chimpanzee

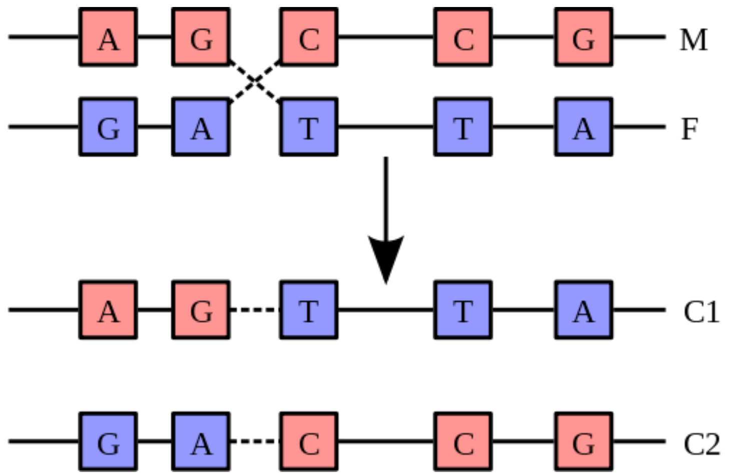
Evolution: why *these* genotypes?

Organisms reproduce, offspring *inherit* genotype from parents

Random *mutation* changes genotypes and *recombination* shuffles chunks of genotypes together in new combinations

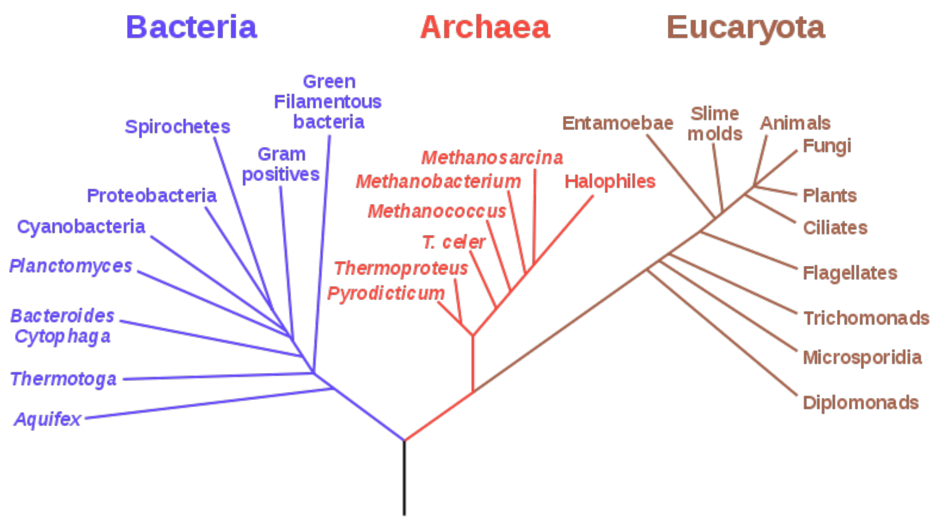
Natural *selection* favors phenotypes that reproduce more

Over time, this yields the variety of life on Earth. Incredibly, all organisms share a common ancestor.



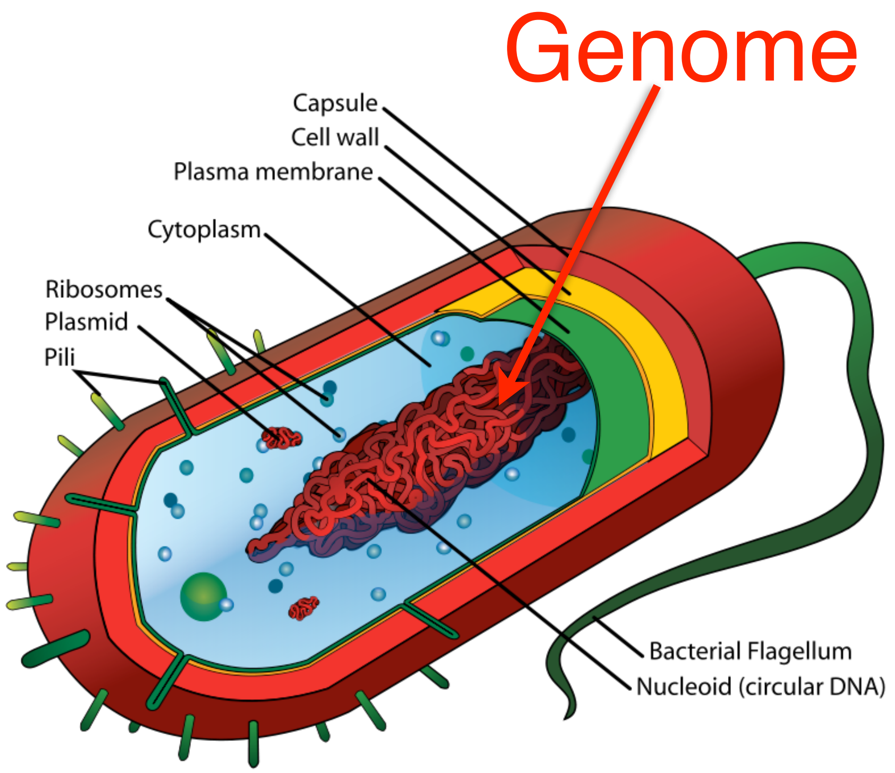
http://en.wikipedia.org/wiki/Genetic_recombination

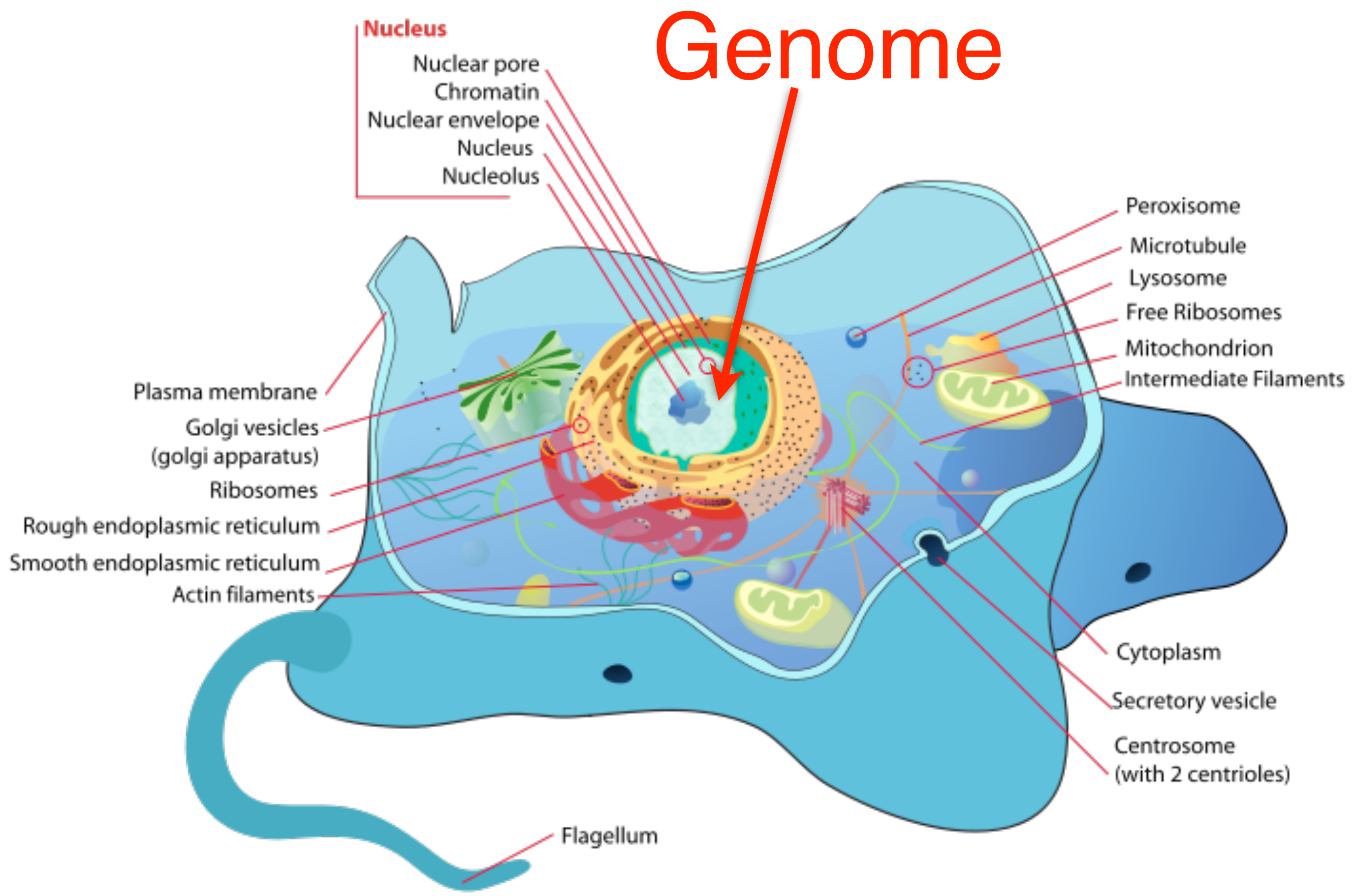
Phylogenetic Tree of Life



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

Cells: where genomes live


Genome


Genome

Prokaryotic cell

A bacterium consists of a single prokaryotic cell

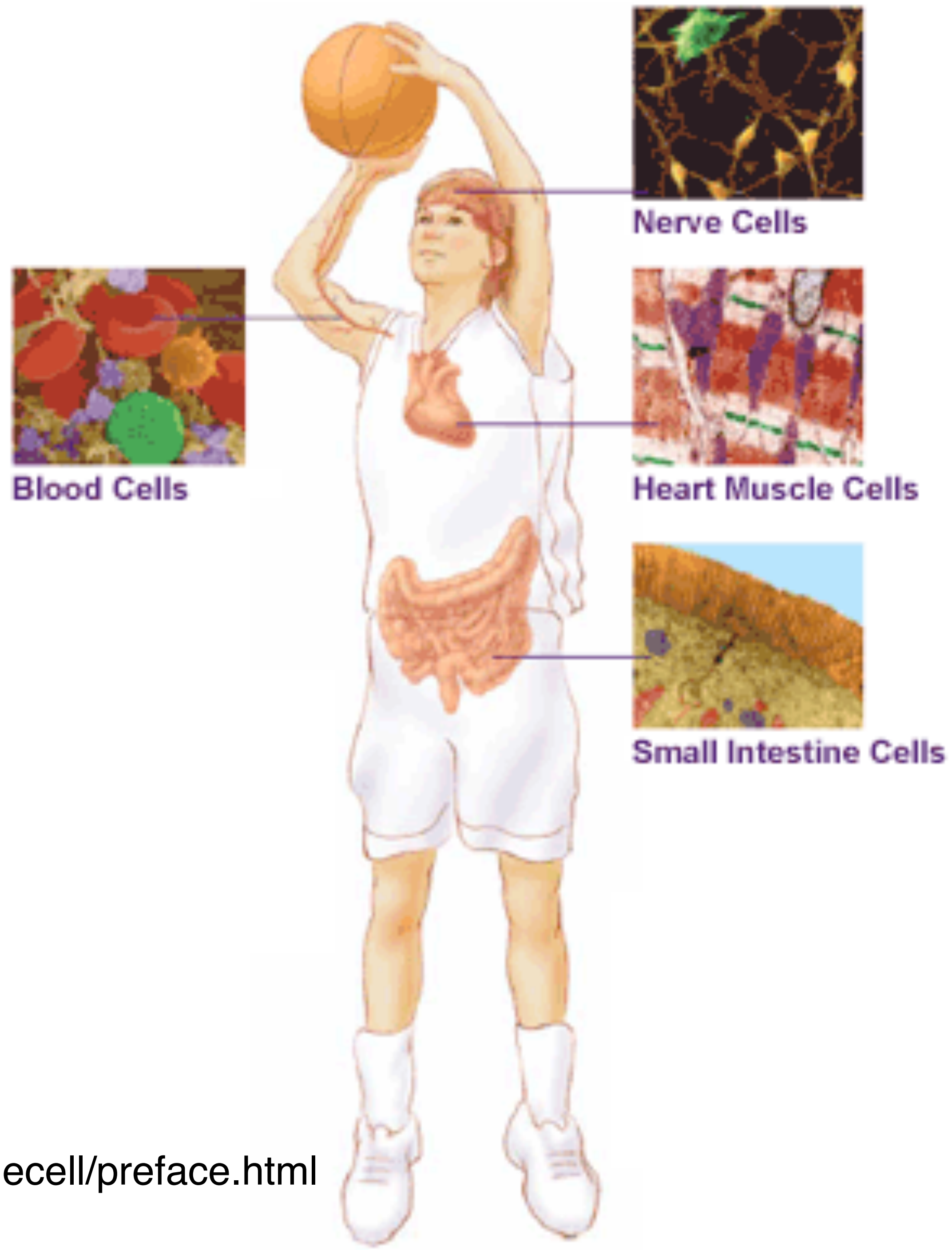
Eukaryotic cell
(pictured: animal cell)

Make up animals, plants, fungi, other eukaryotes

Pictures: http://en.wikipedia.org/wiki/Cell_(biology)

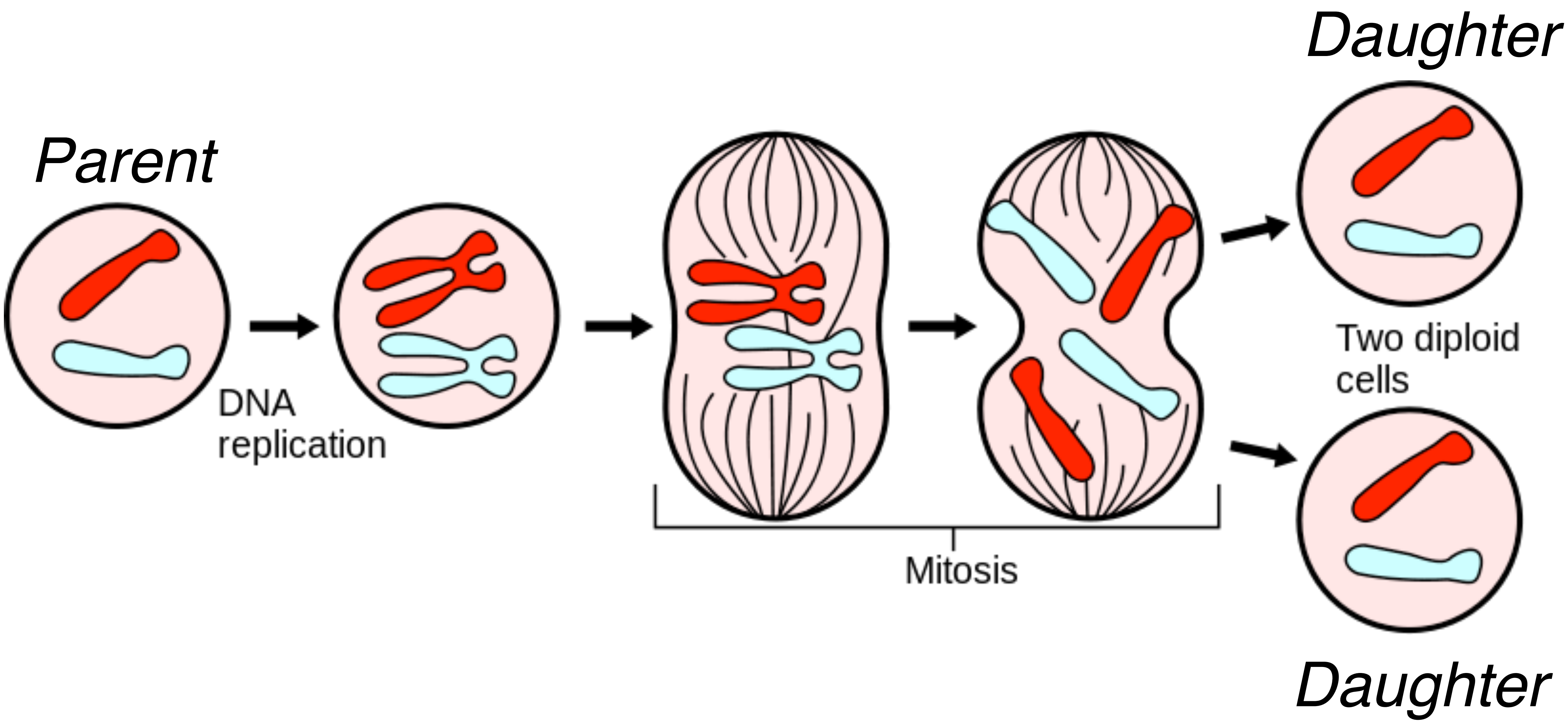
Cells: where genomes live

All the trillions of cells in a person have same genomic DNA in the nucleus



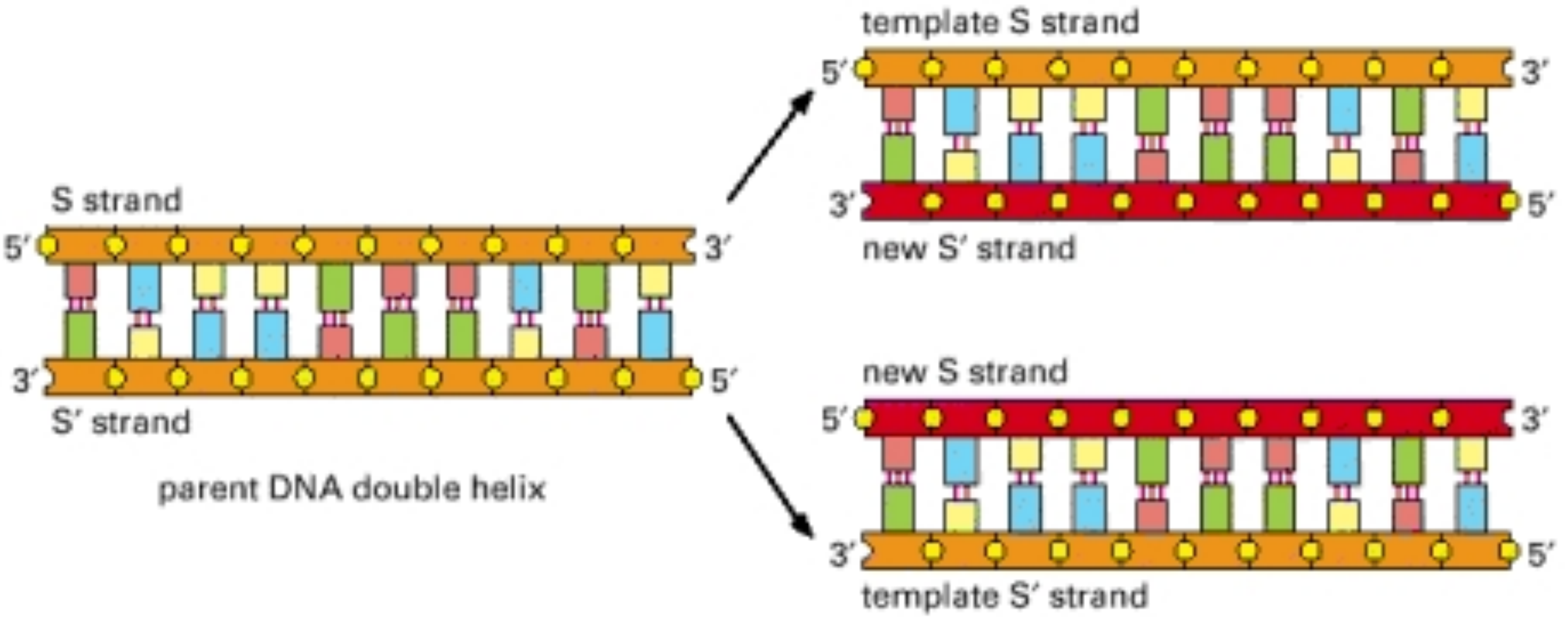
Picture: http://publications.nigms.nih.gov/insidethecell/preface.html

Cells: division



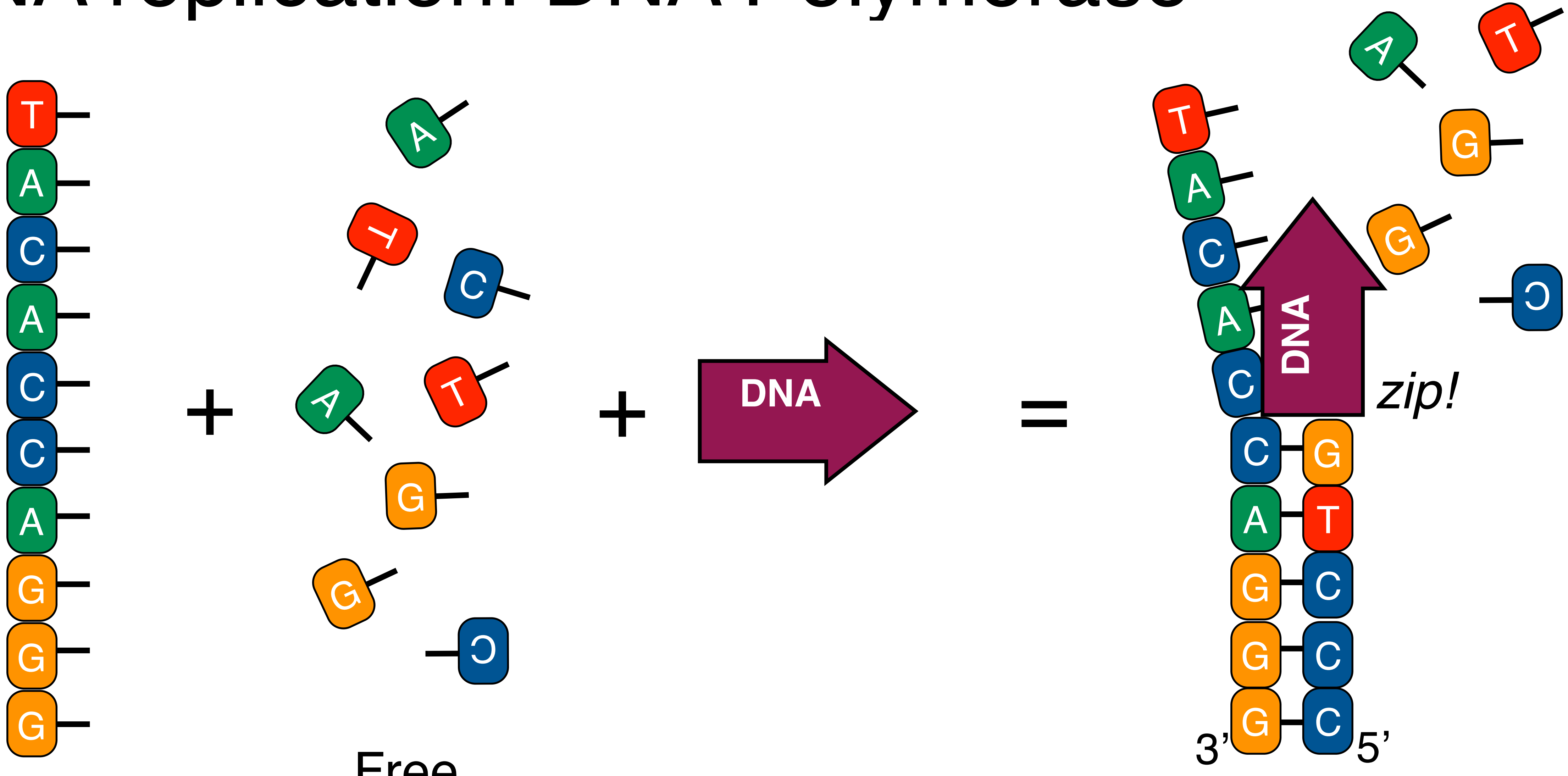
During cell division (*mitosis)*, the genome is copied

Picture: http://en.wikipedia.org/wiki/Mitosis



Each strand becomes a template for replication.

DNA replication: DNA Polymerase



Single-stranded DNA *template*

Free nucleotides dNTPs

DNA polymerase

Strand synthesis

DNA polymerase moves along the template in one direction, integrating complementary nucleotides as it goes.

A short RNA primer starts the replication process.

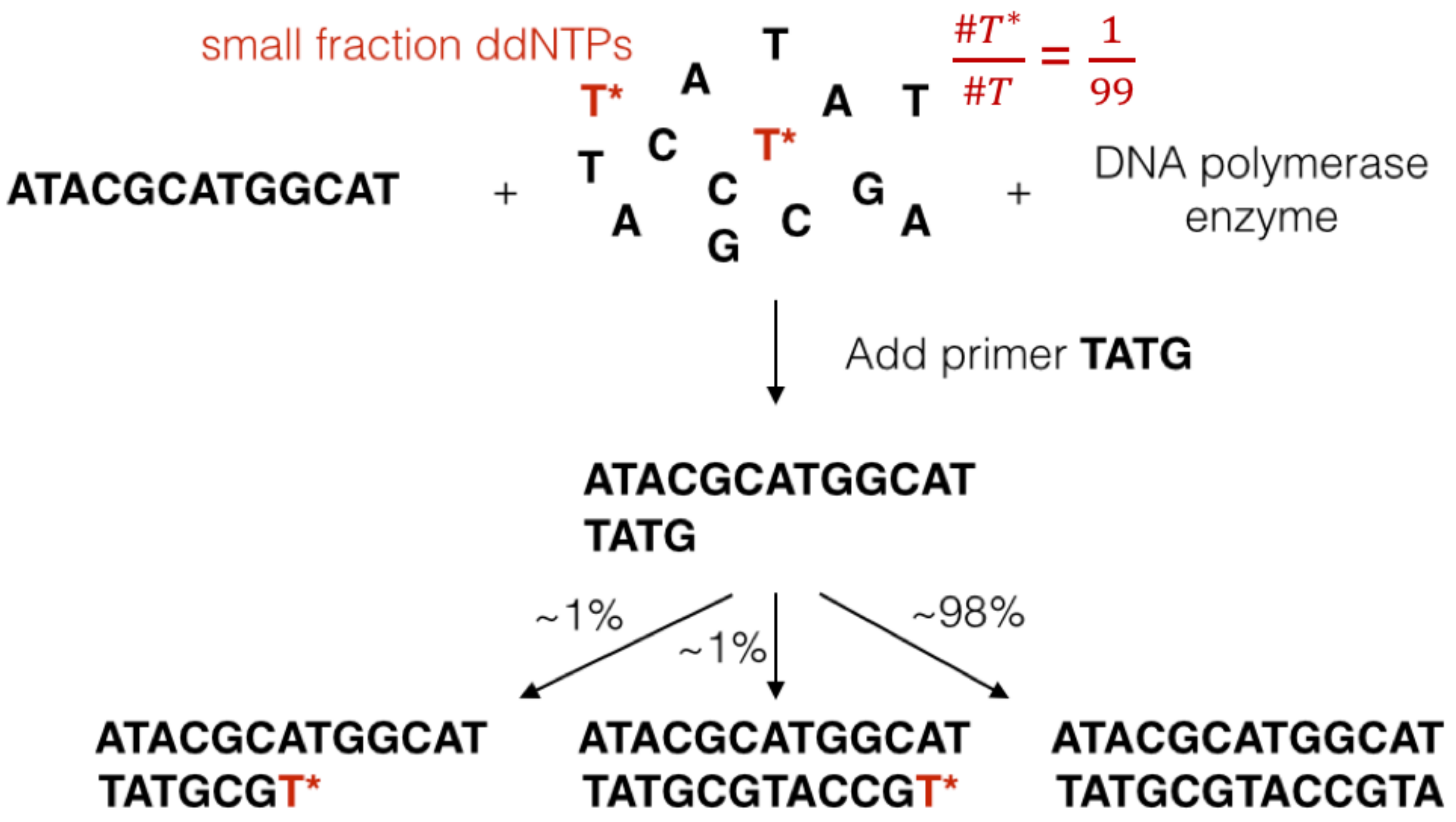
Sanger Sequencing

1. Replicate sequence using PCR (polymerase chain reaction).

2. Break the sequences into many fragments.

3. Break apart the two strands of each fragment by heating.

4. “Simulate” DNA replication to read each fragment.



small fraction ddNTPs

$$\frac{\#T^*}{\#T} = \frac{1}{99}$$

ATACGCATGGCAT + T A T T* A T C T* T C G A C G A G + DNA polymerase enzyme

Add primer **TATG**

ATACGCATGGCAT
TATG

ATACGCATGGCAT
TATGCGT*

ATACGCATGGCAT
TATGCGTACCGT*

ATACGCATGGCAT
TATGCGTACCGTA

Measure length of each strand with gel electrophoresis to determine the position of **A** in each template strand

TATGCGT*   TATGCGTACCGT*   TATGCGTACCGTA

Repeat above process using **A***, **C***, and **G*** ddNTPs in parallel lanes

An Example

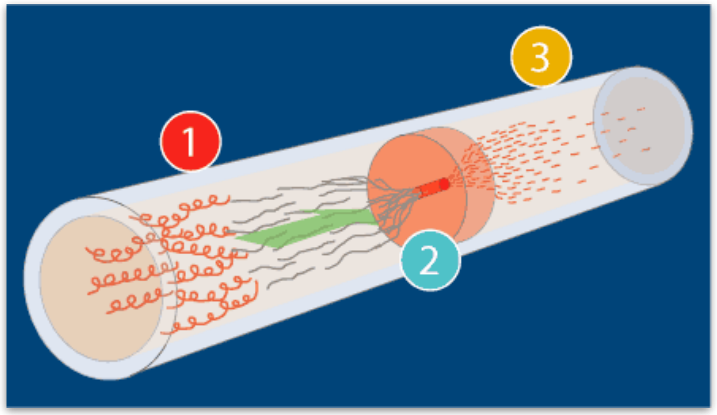
A	C	G	T
30.0	48.2	56.7	86.3
61.3	99.3		
74.4			

30.0 - A
48.2 - C
56.7 - G
61.3 - A
74.4 - A
86.3 - T
99.3 - C

Sequencing by synthesis: second gen

1. Take DNA sample, which includes many copies of the genome, and chop it into single-stranded fragments (“templates”)

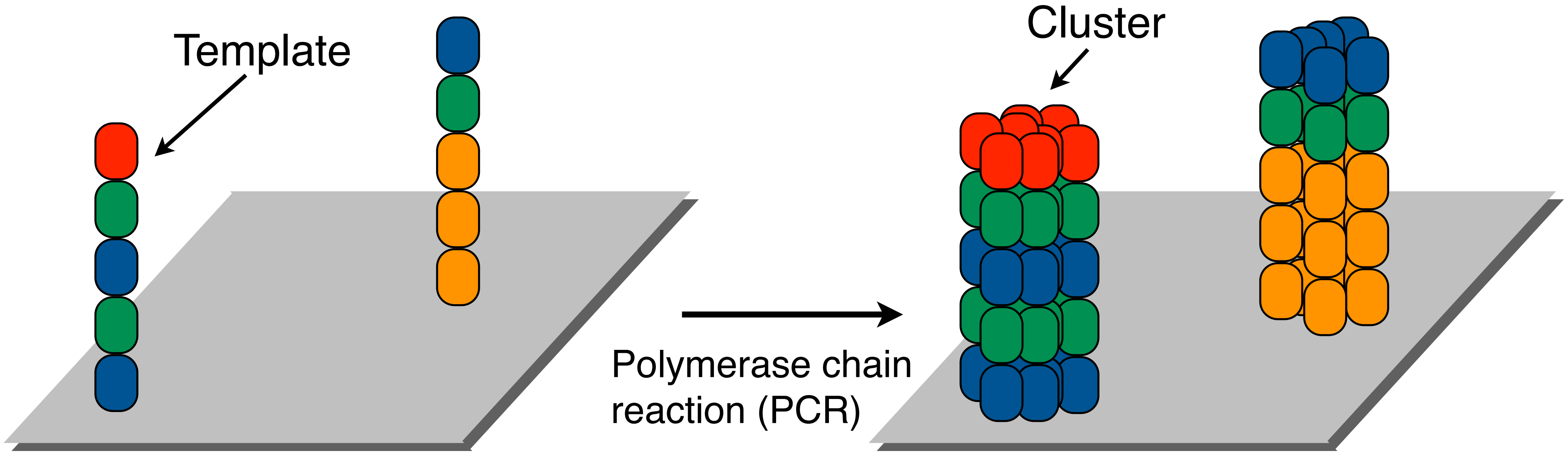
   E.g. with ultrasound waves, water-jet shearing (pictured), divalent cations



Picture: http://www.jgi.doe.gov/sequencing/education/how/how_1.html

2. Attach templates to a surface

3. Make copies so that each template becomes a “cluster” of clones



Template

Cluster

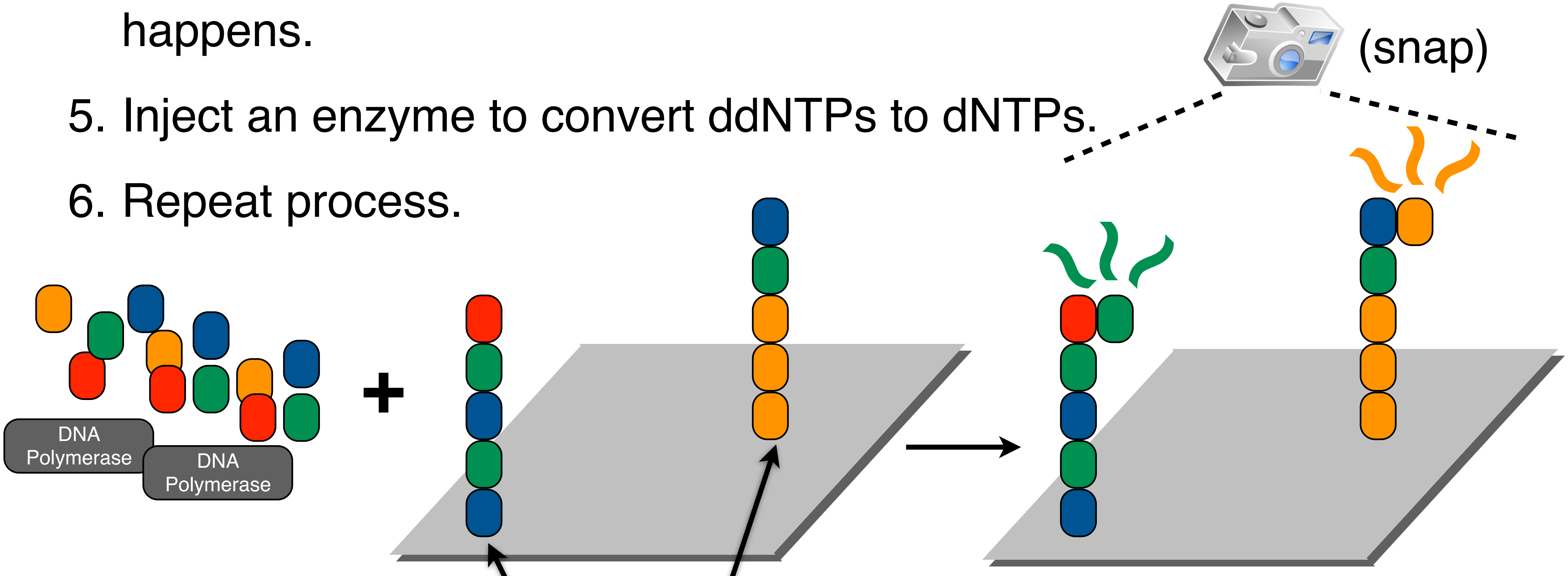
Polymerase chain reaction (PCR)

Sequencing by synthesis

4. Inject mixture of *fluorescence-tagged* ddATP, ddCTP, ddGTP and ddTTP’s and DNA polymerase. When a complementary nucleotide is added to a cluster, the corresponding color of light is emitted. Capture images of this as it happens.

5. Inject an enzyme to convert ddNTPs to dNTPs.

6. Repeat process.

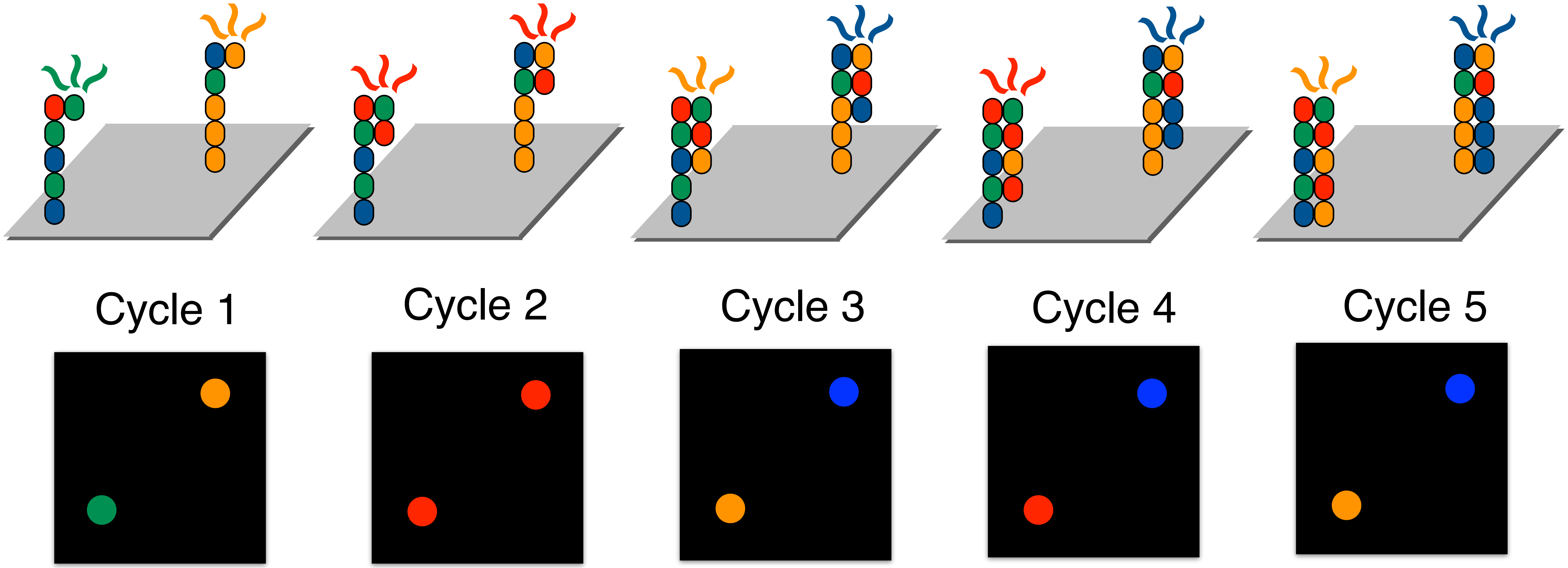


Pretend these are clusters

Shown here is just the first *sequencing cycle*

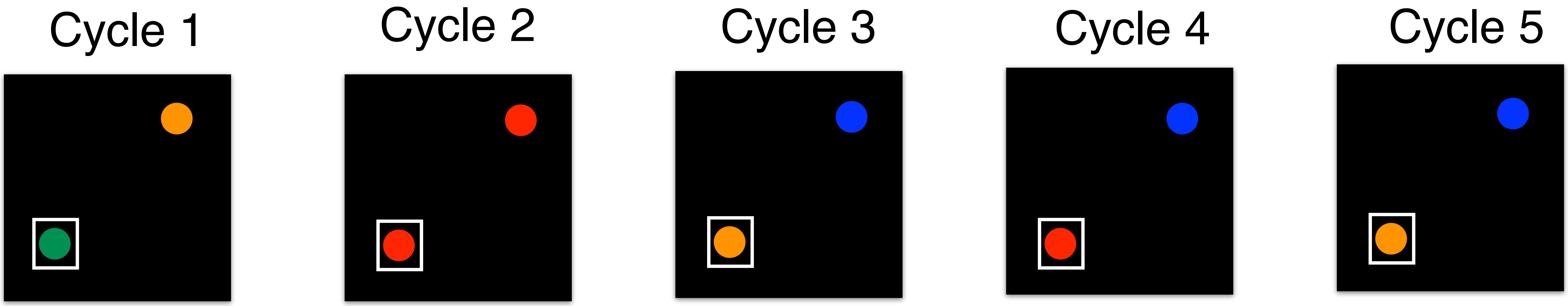
Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides

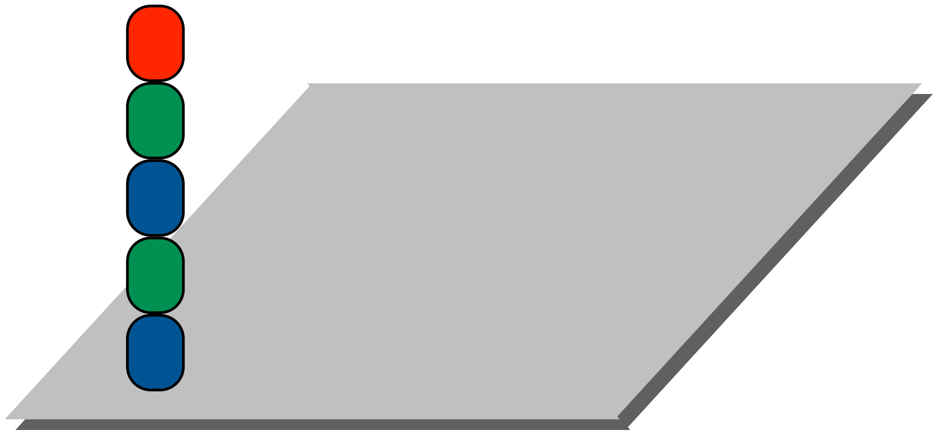


Sequencing by synthesis

5. Line up images and, for each cluster, turn the series of light signals into corresponding series of nucleotides



"Base caller" software looks at this cluster across all images and "calls" the complementary nucleotides: TACAC, corresponding to the template sequence



TACAC is a "sequence read," or "read." Actual reads are usually 100 or more nucleotides long.

Sequencing by synthesis

A modern sequencing-by-synthesis instrument such as the HiSeq sequences *billions* of clusters simultanously

A single “run” takes about 10 days to generate about 600 billion nucleotides of data

Cost of the reagents is $5-10K per run; multiplexing (sequencing many samples per run) further reduces cost per genome