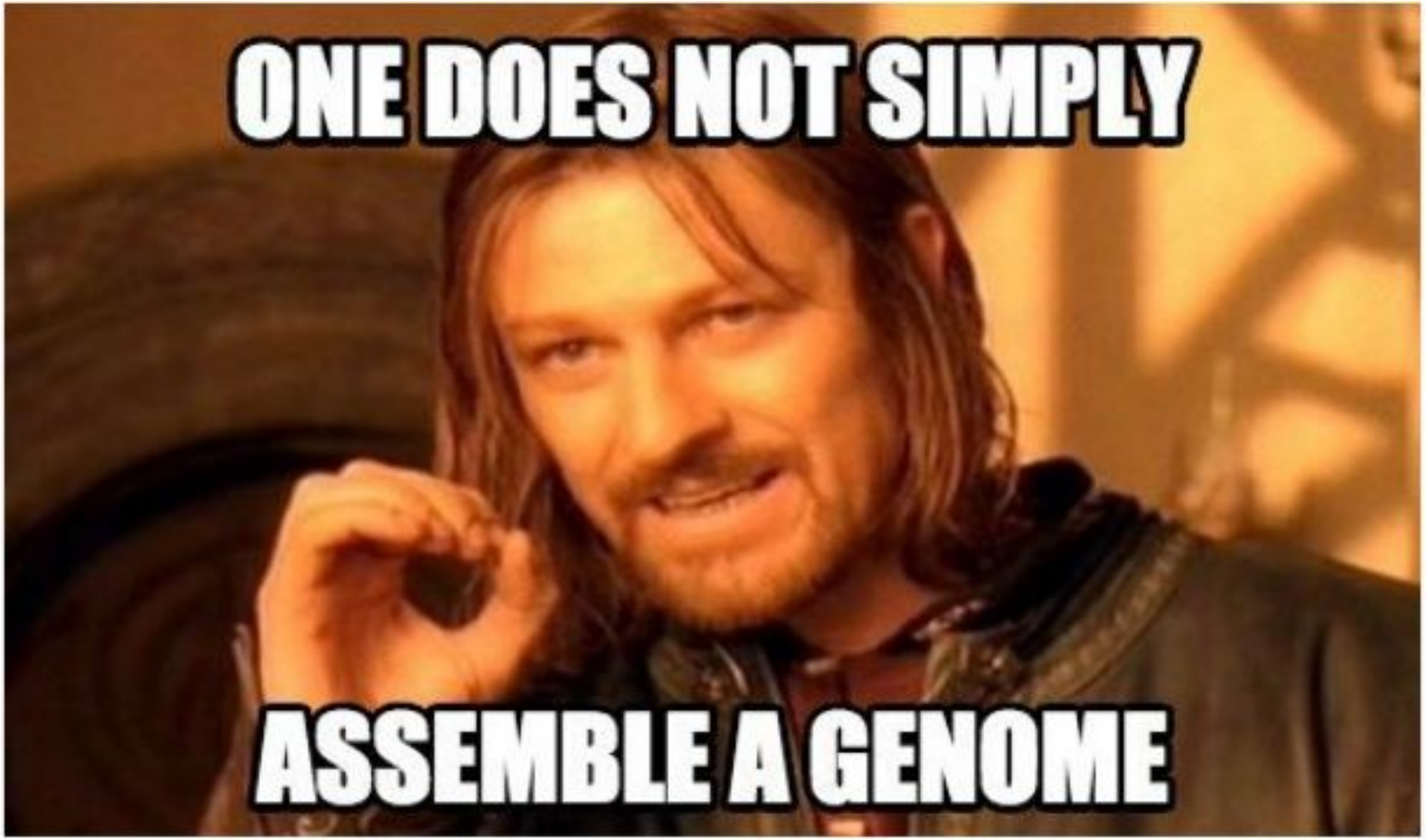

ONE DOES NOT SIMPLY
ASSEMBLE A GENOME

Week 2: The Human Genome Project (HGP)

Looking Forward:

- **Week 3: Data we'll be using**
- Week 4: Technology behind the data in week 3
- **Week 5: Analysis we'll be replicating**

Notes:

- Office Hours: Wed 4-5PM; Th 10:30-11:30
- Your responses to weekly Discussion Q's added to each week's README.
- On your name tag circled in red: # correct / # attempted (18 Qs)

1990

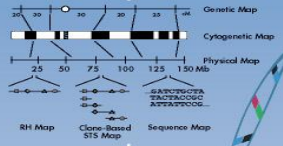
Human Genome Project (HGP) launched in the U.S.

Ethical, Legal, and Social Implications (ELSI) programs founded at NIH and DOE

First gene for breast cancer (BRCA1) mapped

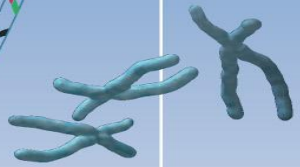
1991

First U.S. Genome Centers established



1992

Second-generation human genetic map developed



Rapid data release guidelines established by NIH and DOE

1993

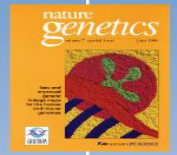
New five-year plan for the HGP in the U.S. published

Sanger Centre founded (later renamed Wellcome Trust Sanger Institute)

The Wellcome Trust

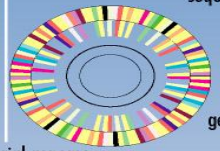
1994

HGP's human genetic mapping goal achieved



1995

HGP's human physical mapping goal achieved



First bacterial genome (*H. influenzae*) sequenced

U.S. Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

1996

First human gene map established

Pilot projects for human genome sequencing begin in U.S.

First archael genome sequenced

Yeast (*S. cerevisiae*) genome sequenced

HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established

1997

DOE forms Joint Genome Institute

JGI DOE JOINT GENOME INSTITUTE US DEPARTMENT OF ENERGY OFFICE OF SCIENCE

NCHGR becomes NHGRI

National Human Genome Research Institute

E. coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded

1998

Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the U.S. published

RIKEN Genomic Sciences Center (Japan) established

Roundworm (*C. elegans*) genome sequenced

SNP initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers (in Beijing and Shanghai) established

1999

Full-scale human sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000

Draft version of human genome sequence completed

President Clinton and Prime Minister Blair support free access to genome information

Fruit fly (*D. melanogaster*) genome sequenced

Mustard cress (*A. thaliana*) genome sequenced

Executive order bans genetic discrimination in U.S. federal workplace

2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002

Draft version of mouse genome sequence completed and published

Draft version of rat genome sequence completed

Draft version of rice genome sequence completed and published

2003

Finished version of human genome sequence completed

HGP ends with all goals achieved

to be continued..

How BIG is the Human Genome?

If the human genome were compiled in books:

- 200 volumes, 1000 pages each
- read 10 bases/second = 315,360,000 bases/year
- 9.5 years to read out loud (without stopping)

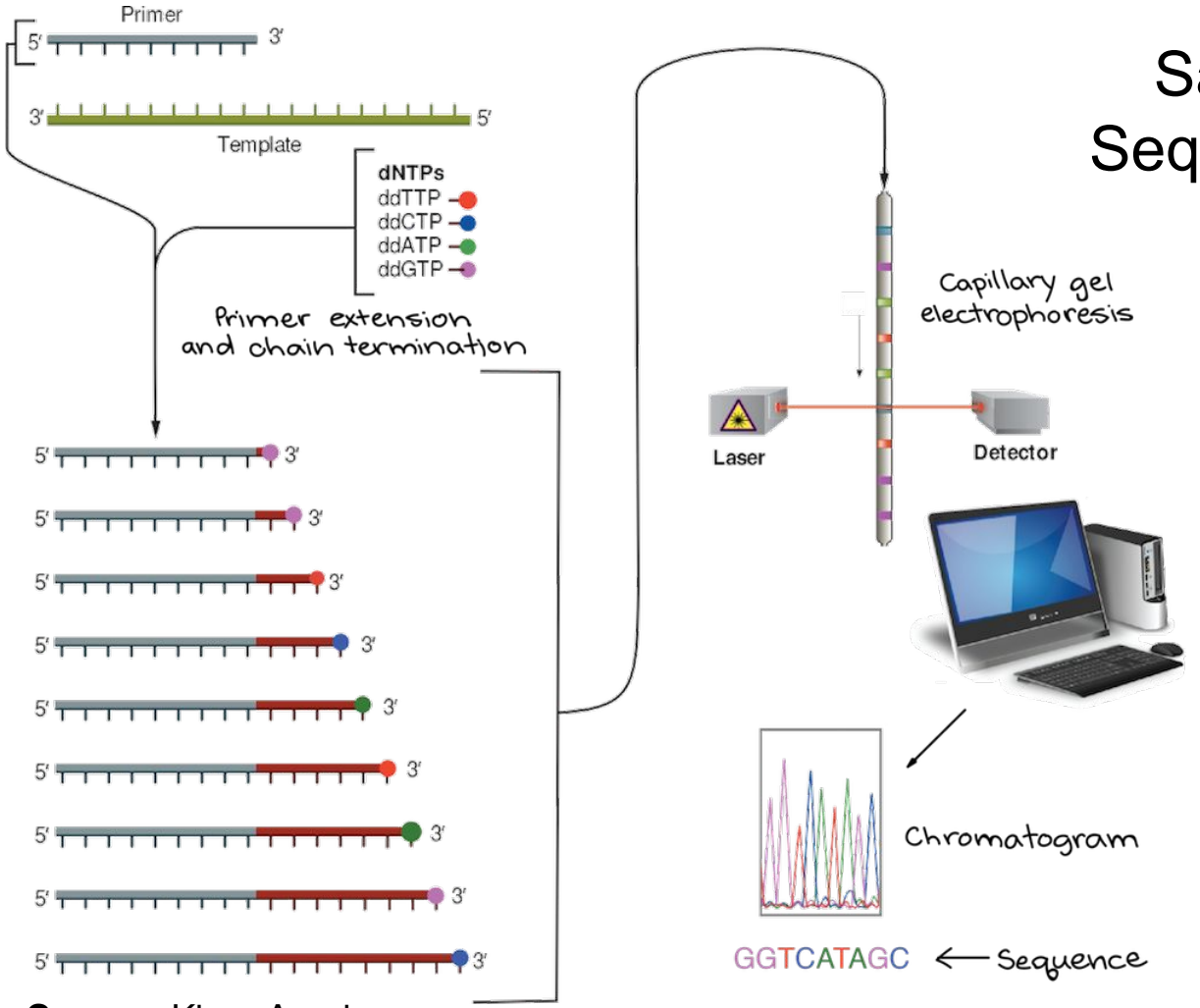
How do you sequence a genome?

- Determine order of bases on all 23 (24) chromosomes
- Can only read 30 to 700 bases at a time
- Cannot sequence a genome in one run
- “Whole Genome Shotgun” sequencing

Discussion Questions

- What is the human reference genome?
- Why did scientists set out to determine the reference genome?
- What are three (3) important things we learned from the human genome project (HGP)?
- What is a SNP? Why do we care about them?
- **What is 'genome assembly'?**

Sanger Sequencing



Source: Khan Academy

Sanger is slow & limited in how much it can sequence.

Two options: primer walk vs shotgun

primer walk - one piece at a time

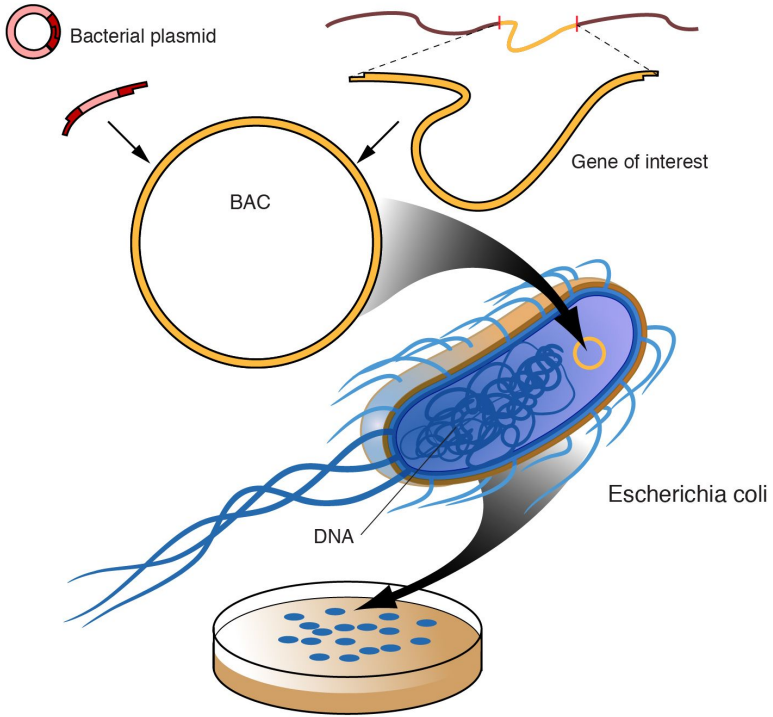
1. Sequence a short piece
2. Design a primer to short piece
3. Repeat and walk

shotgun - blast genome to pieces and assemble later

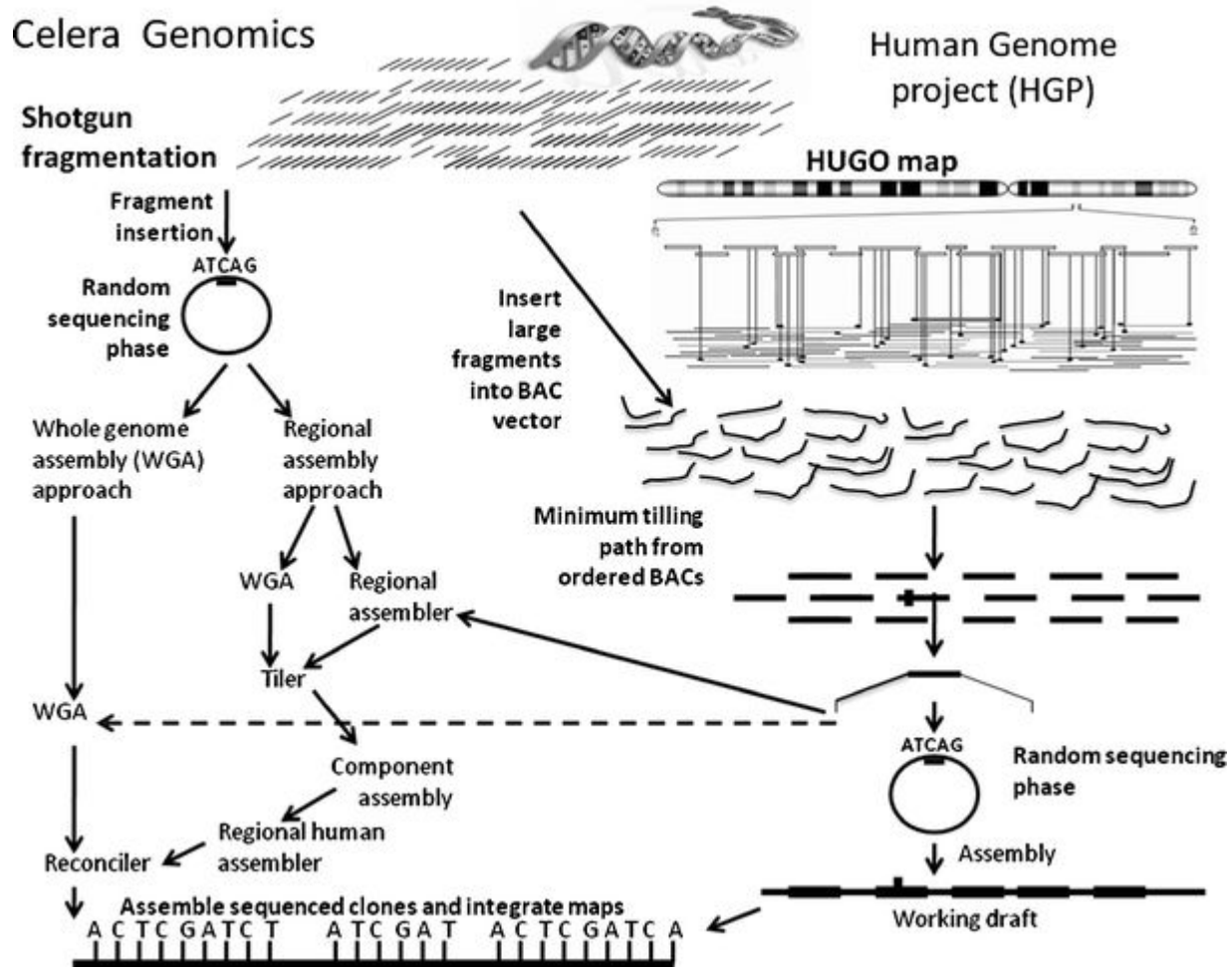
1. Split up DNA “randomly” into small pieces
2. Sequence each “read” (small piece)
3. Use overlapping reads to figure out how they fit together

Human Genome Project

1. Physical Map is known
2. Multiple genomes are sheared
3. Pieces are cloned into BACs
4. Clones contain copies of small DNA pieces (BAC library)
5. BACs are sequenced as needed
6. Overlaps used for assembly



Source: NHGRI



Pareek, Chandra & Smoczyński, Rafał & Tretyn, Andrzej. (2011). Sequencing technologies and genome sequencing. Journal of applied genetics. 52. 413-35. 10.1007/s13353-011-0057-x.

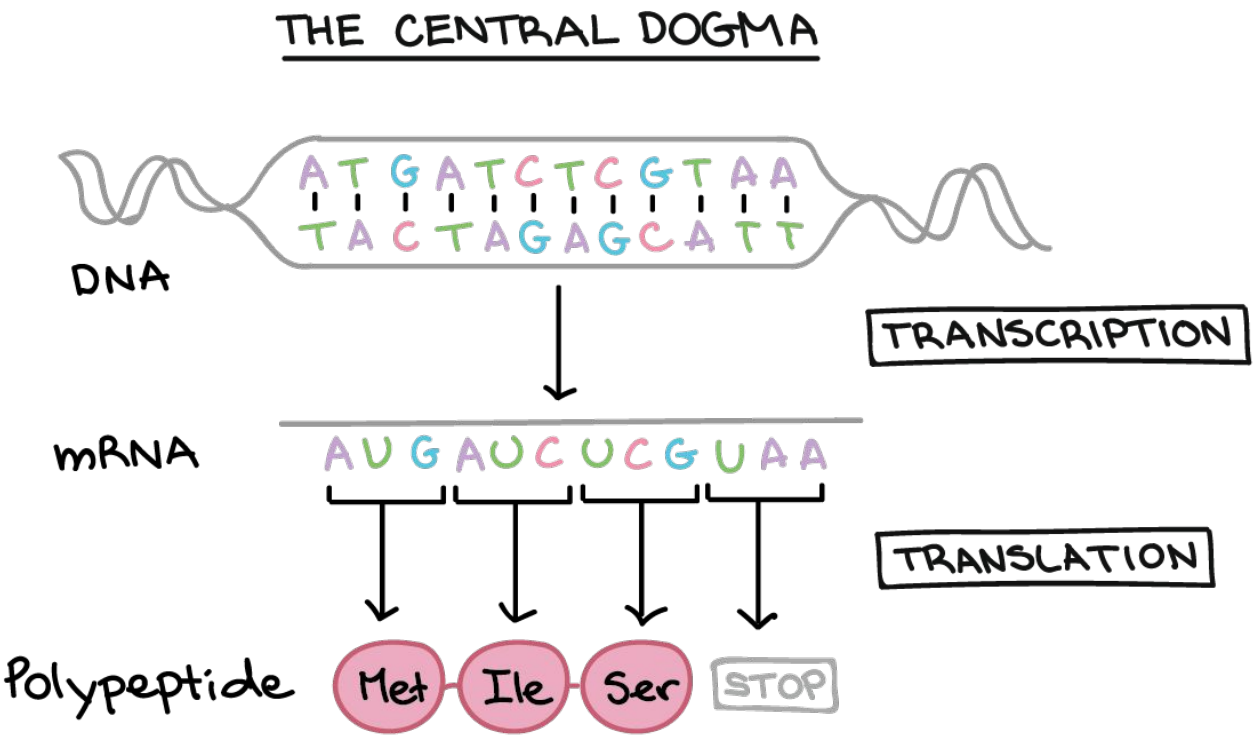
To help in assembly

STSs (Sequence Tag Sites) - short (~500bp) pieces of DNA of known sequence and chromosomal location + PCR to screen for their existence

Discussion Questions

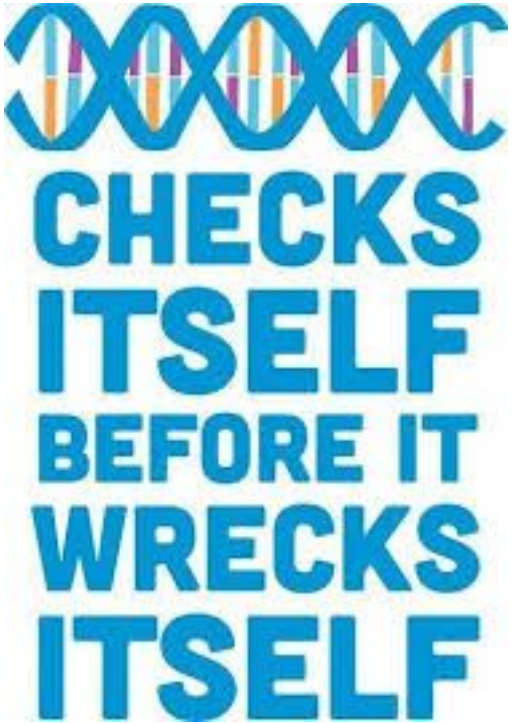
- What is the human reference genome?
- Why did scientists set out to determine the reference genome?
- What are three (3) important things we learned from the human genome project (HGP)?
- **What is a SNP? Why do we care about them?**
- What is 'genome assembly'?

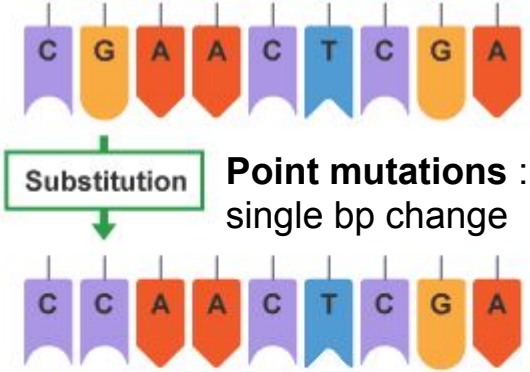
The Central Dogma of Genetics



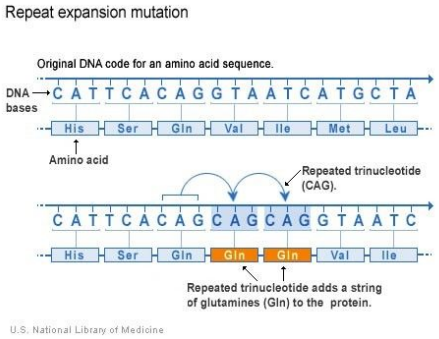
Source: Khan Academy

Mutations

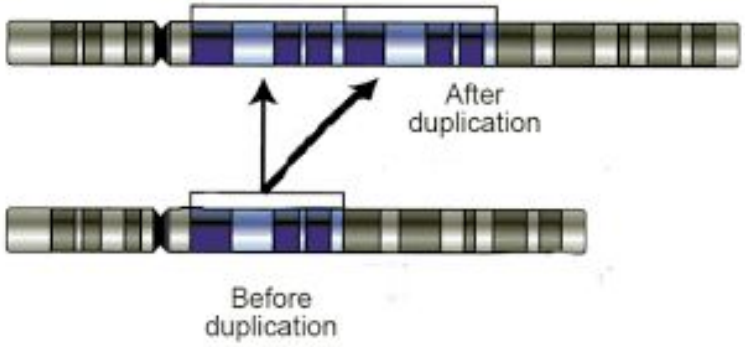




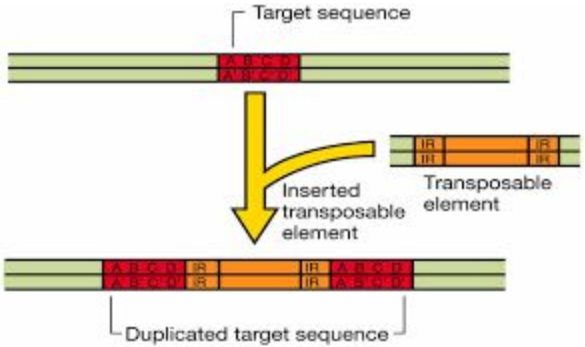
Point mutations : single bp change



Repeat Expansions - Short DNA sequences that are repeated a number of times in a row.



Duplications - A piece of DNA that is abnormally copied one or more times.



Transpositions : the ability of genes to change position on chromosomes (“mobile DNA”)

The genome is bonkers.

Insertions/Deletions (Indels) - The addition or deletion of one or more bases from the genome.

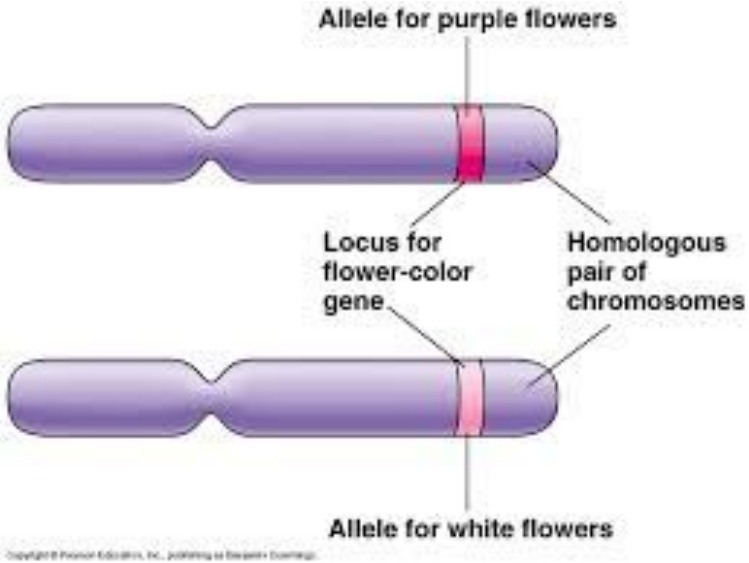
Genetics Terminology

Gene - a unit of heredity; a section of DNA sequence encoding a single protein

Locus - location on a strand of DNA

Alleles - possible variations of a single gene

Genome - the entire set of genes in an organism



Causes of mutations

Replication errors <- these happen all the time!

Crossing over <- happens every time a zygote forms

Radiation

Chemicals

DNA intercalating agents

mutations == human variation

- Human genome: 3B base pairs across 22 autosomes + 2 sex chromosomes
- Any two people differ at a position along their genome ~ every 100 bp
  - Most of these do not happen in genes (benign) or don't affect protein (silent)
  - Mutations that happen in important places -> disease
- There are **80M+** variants in the human genome
- Genetically similar populations have more similar genomes