During Today's Exercise

1. Do not race ahead in your exercises

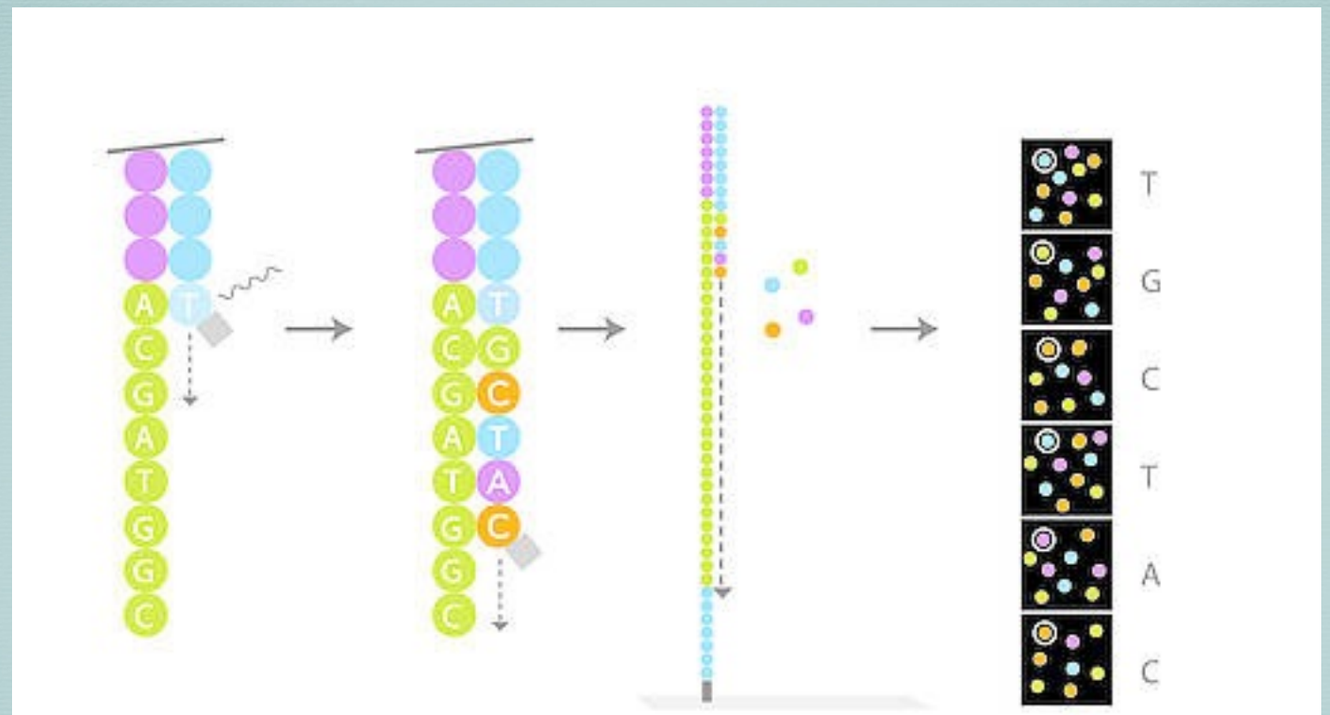
2. You may not leave until the class is over

3. Help your neighbor(s) and ask questions

4. Follow Along the Tutorial

5. Enjoy yourself

What can I do with raw ngs data?



molb-4485/5485

10/25/2016

NGS is just fancy name for high throughput DNA/RNA sequencing platforms

What are some of the data types that rely on ngs?

GBS = Understand genome-wide genetic variation
RADseq = Similar to GBS
RNAseq = Snapshot of genes functional at a given time
WGS = Whole genome sequencing

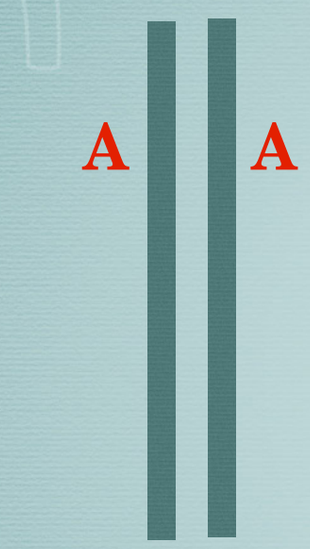
What is a

genome, phenome, transcriptome, proteome ?

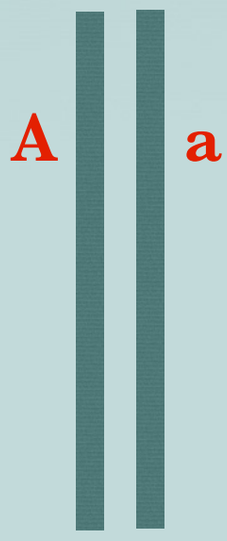
What is genetic variation & why care about it?

A gene or a genetic locus comes in two forms

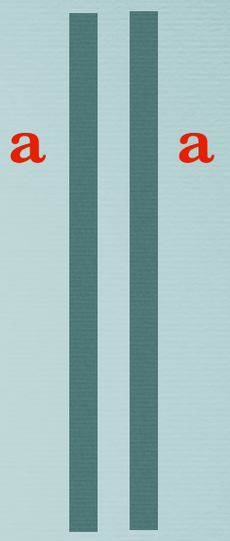
e.g. Gene / Locus A has 2 copies, one on each homologous chromosome



homologous chromosome pair



homologous chromosome pair



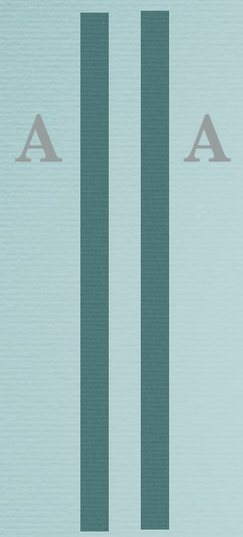
homologous chromosome pair

A and a are two alleles (forms) at this locus

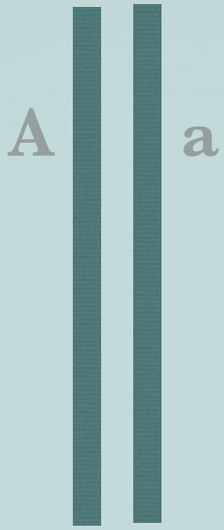
What is genetic variation & why care about it?

A gene or a genetic locus comes in two forms

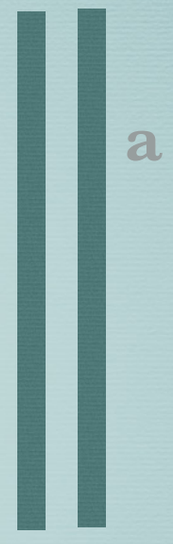
e.g. Gene / Locus A has 2 copies, one on each homologous chromosome



homologous chromosome pair



homologous chromosome pair



homologous chromosome pair

Environments are always in flux (given evolutionary time)

New alleles arise due to mutations and if beneficial in a given environment, spread throughout the gene pool

This is *adaptation*

Why do we care about genetic variation?

Maintaining genetic variation allows organisms to *adapt* to changing environments.

Species that cannot adapt, go *extinct*

Genetic Variation
is raw material for evolution

How to quantify variation?

Single Nucleotide Polymorphism (SNP)

The most prevalent type of variation in the genomes of most organisms


ATGGCAGCTGATATAC


ATGGCAGCTGATATAC


ATGGCAGCTAATATAC

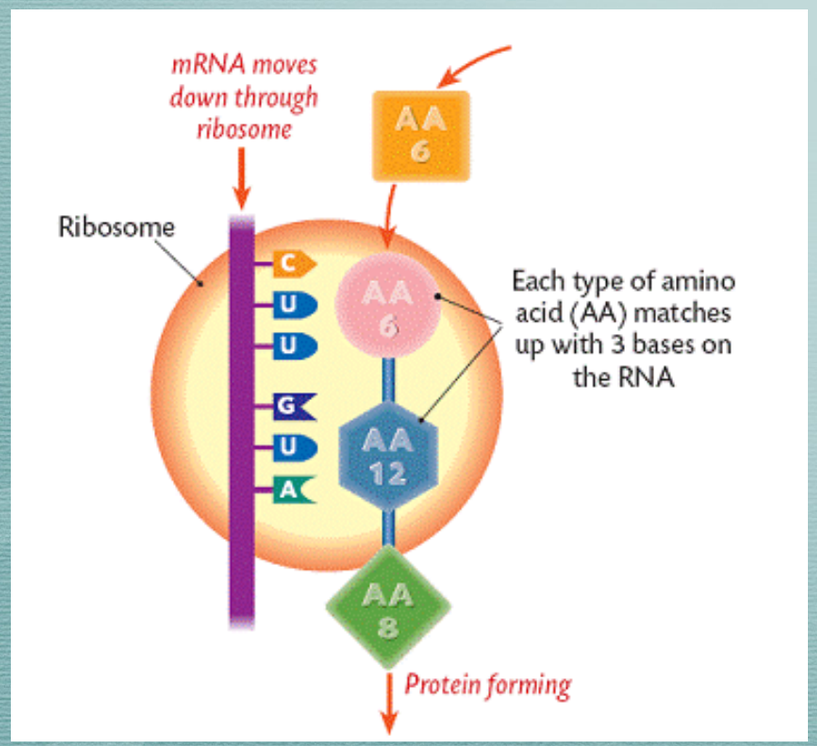
https://www.23andme.com/gen101/snps/

How to quantify variation?

SNPs can be

Synonymous or *Non-Synonymous*

Syn = No changes at protein level
NonSyn = May change protein structure/function



Remember genetic code?

how to go from raw ngs data to SNPs?

raw reads + reference genome = Alignment = Variant Calls

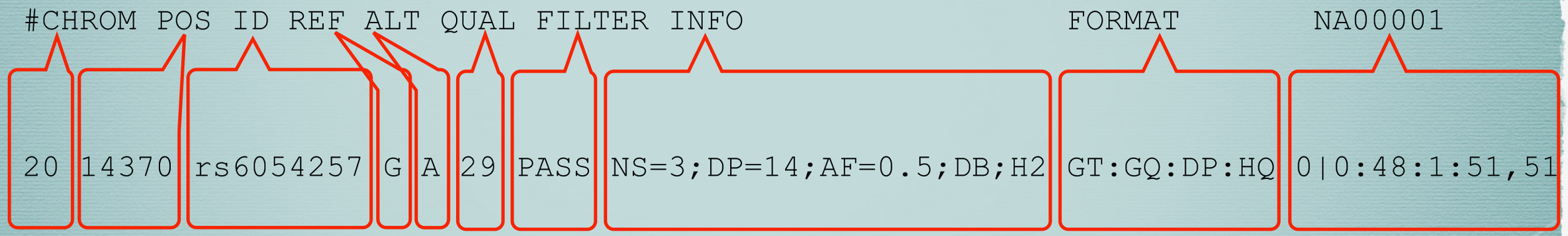
OR

Pool raw reads from many individuals → Assemble reads to form longer contigs → Assemble contigs into even longer scaffolds → *de novo* Reference Genome

Variant Call Format

A highly popular format to store SNP and read quality information

.vcf



http://samtools.github.io

Metrics of Genetic Variation

Allele Frequency = How prevalent is a given 'gene form' (*allele*) in your study population?

Prevalence of different alleles in different geographical areas in a species may be indicative of differential adaptation to the environment.

Population Z

100 individuals
200 chromosomes
400 alleles @ a locus

Allele A: 150 copies
Allele a: 250 copies

Frequencies
$f(A) = 150/400 = 0.375$
$f(a) = 1-0.375 = 0.625$

Population Y

120 individuals
240 chromosomes
480 alleles @ a locus

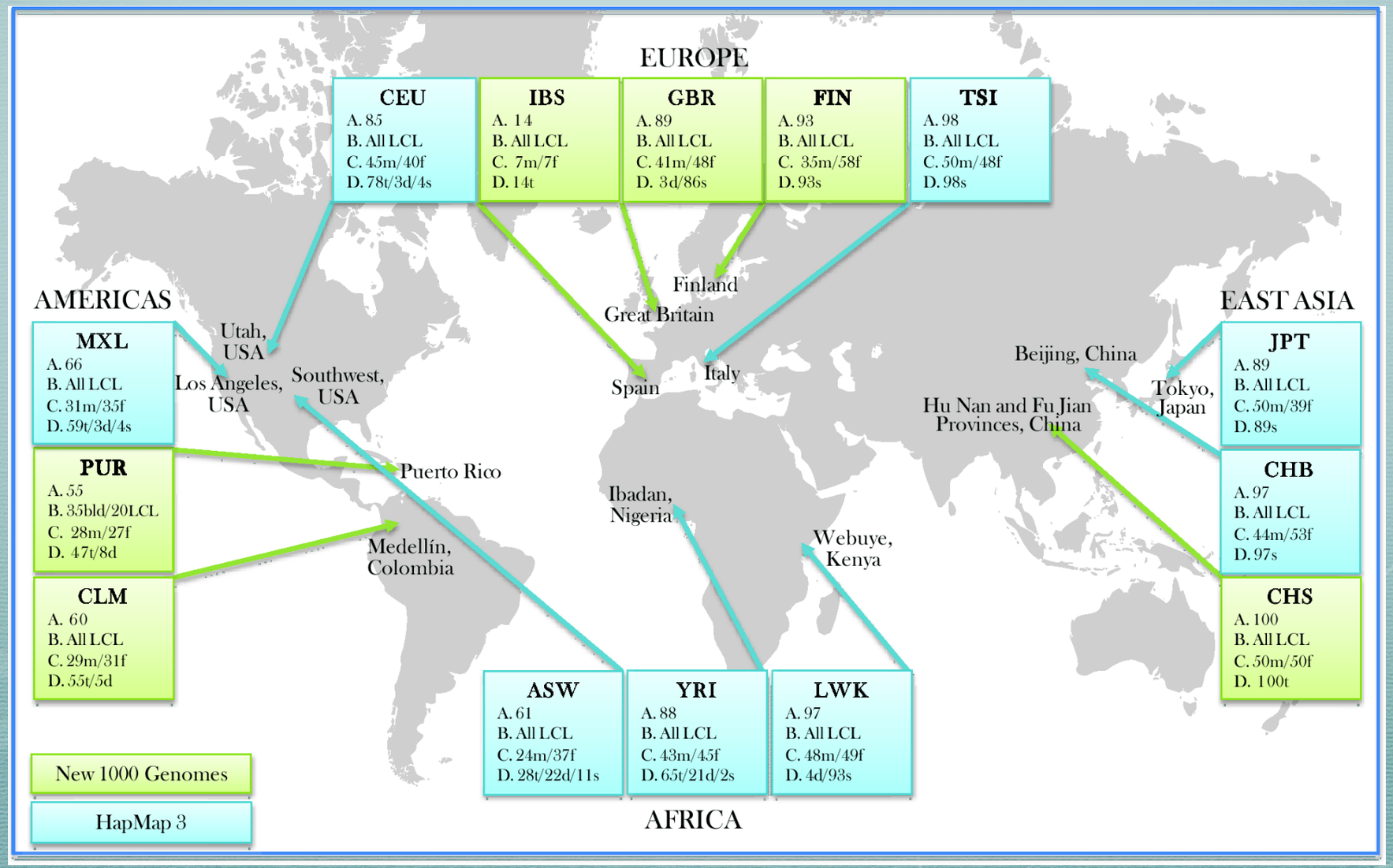
Allele A: 100 copies
Allele a: 380 copies

Frequencies
$f(A) = 100/480 = 0.21$
$f(a) = 1-0.21 = 0.79$

Genetic Diversity in Human Populations

1000 Genomes Project

26 Populations
5 Super Populations
2504 Individuals
4.7 Million SNPs



Genetic Diversity in Human Populations

1000 Genomes Project

26 Populations
5 Super Populations
2504 Individuals
4.7 Million SNPs

We are going to learn about one of these SNPs
in gene LCT
Also called **`rs34100645`**

This SNP plays a small part in Lactose Persistence in Humans and thus may have variable distribution of allele frequencies among global populations

LCT SNP Variation

We will estimate frequencies of the two alleles at this SNP in 26 human populations of various ethnicities

→ Tutorial: **`vcf+R.pdf`**