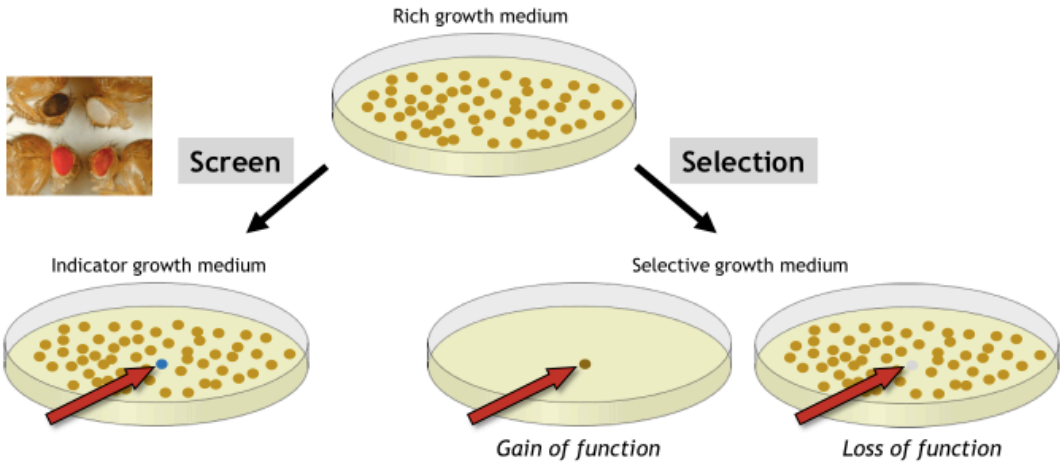
Bacterial Genetics

1: Understanding & working with bacterial genes

- Life
  - 3 Domains: Bacteria, Archaea (these two are prokaryotes w/ no organelles) and Eucarya (membrane enclosed structures with organelles)
  - Length of tree line corresponds with how much it diverged from the ancestor
  - We are closer to fungi than we are to eukaryotes
  - Viruses are outside the three domains of life
    - Marine viruses; 4 x 10^{30} viruses exist in ocean waters
- Why Study Bacteria?
  - Regulation of biogeochemical cycles:
    - Sulfur Cycle for proteins; Nitrogen Cycle to fix N2 in soil for plants
  - Vast economic, industrial, medical and agricultural importance
    - Bacteria to make snow; Squash, maize, & beans w/ N2 fixing bacteria; Plants make ethanol for fuel
  - They account for >60% of the Earth's biomass, can survive at high altitudes, temperatures, and depths —> possible because they don't live as isolated species but in a **microbiome**
- Microbiome
  - **Assemblage of microbes that reside in a specific niche; microbial community of a particular habitat/ environment; *human microbiome has niche as human body***
  - Human Gut Microbiome
    - 500-1000 bacterial species; 10x more gut bacterial cells than human cells in ENTIRE HUMAN BODY; 100x more genes in microbiome than human genes
    - Essential for health: regulation of metabolism +aiding digestion + training immune system
- Bacterial Genetics
  - How we can use it: vast analytical power for identifying and manipulating genetic variation
    - Huge population size (1 billion cells in 1 mL); Extensive genetic variation (very fast genome-wide mutation rate); Fast generation time;
    - Easy mutagenesis; Phenotypes found by screens (gof and lof) and selections; potential for mutant enrichment
  - Genetic Dark Matter: genes don't exist until we can isolate a mutant with a phenotype

- Huge proportion of genes exist which we don't know the function of (because we haven't screened them under the right selection pressures)

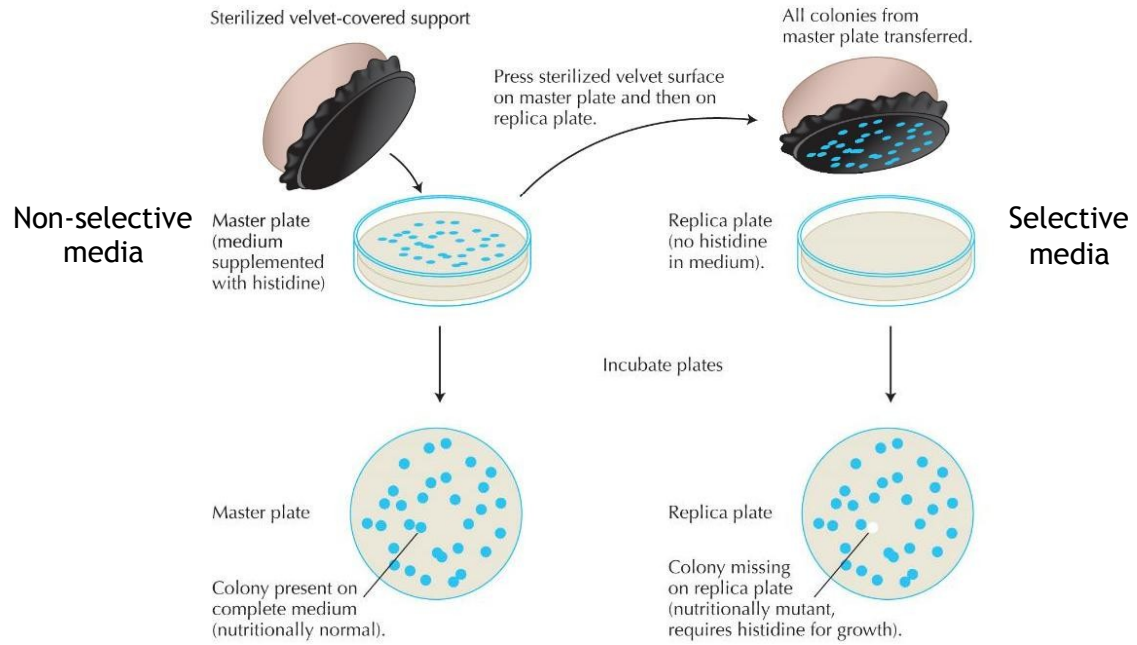
- Identifying Mutants:
  - We can keep diluting test tube and reduce the # of E.coli (or any bacteria) cells in each test tube when we transfer them over; then we spread 0.05 mL (50 microL) or about 1 drop over the surface of a solidified agar containing nutrient broth in a Petri dish —> incubate —> have ~50 colonies per dish and each colony has ~10^6 cells
  - **Genetic Screen:** *Identification of genotype of interest based on differential phenotypes*
  - **Genetic Selection:** *Identification of genotype of interest based on differential survival*
  - We use **Genetic Selection** to establish conditions in which only the desired mutant will grow. Example: Select for StrR mutants on streptomycin. We use **Genetic Screen** to examine each colony in a population for its phenotype. Example: screening for auxotrophs



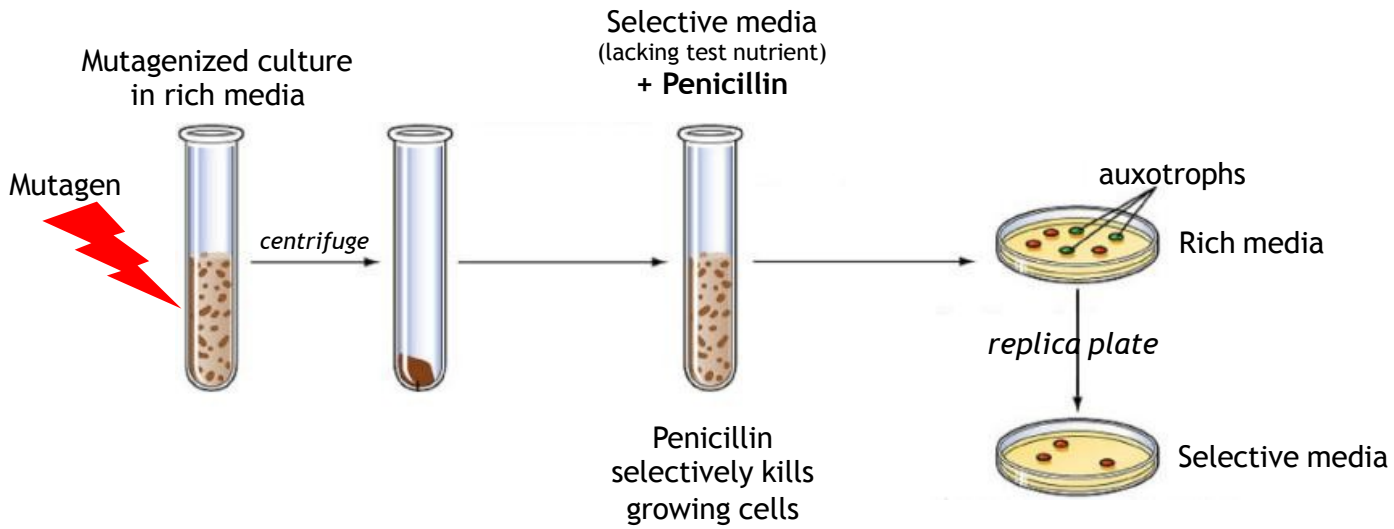
Rich growth media=agar with normal nutrient broth
Indicator growth media=has chemical indicator which changes colour of colony of interest (and in turn identify bacterial cells of interest)
Selective grown media=gof shows colonies carrying the mutations to survive the selection (e.g. antibiotic like clindamycin) and lof shows colonies without mutation to survive

- Replica Plating
  - Selecting for metabolic traits
  - **Prototrophs:** *microbes w/ same nutritional capabilities as wild type*
    - i.e. if we have a plate -nutrient and a colony grows, this colony is nut+ because it can metabolize the nutrient without it being present in the media; GOF mutation

- **Auxotrophs:** *mutants that lost the ability to synthesize a nutrient/substance required for growth*
  - i.e if we have a plate -nutrient and a colony does not grow, then it is nut- because it cannot grow when the nutrient isn't present; LOF mutation
  - We screen for LOF mutants via **replica plating**
- **Replica Plating by Ether Lederberg**
  - **SCREENING** for LOF mutants (auxotrophs)



- **Penicillin Enrichment for Auxotrophs**
  - **SELECTING** for LOF mutants
  - Penicillin kills by inhibiting proteins which cross-link peptidoglycan in the cell wall; when bacteria divides in the presence of penicillin, it cannot fill the holes left in the cell wall —> leaky cell wall allows influx of H2O and apoptosis occurs



- The penicillin only affects growing cells (kills 99% of prototrophs) and not auxotrophs (because the auxotrophs dont grow in -nutrient selective media) and the resulting plate has the 1% of prototrophs and all the auxotrophs; then we once again use a selective -nutrient media to select for auxotrophs

- Understanding the Nature of Mutations

- When we develop bacteria R to antibiotics, is it a random mutation or is it selective pressure (Lamarck=physiological response Darwin=random)
- Luria and Delbruck used bacteriophages to make two hypotheses:

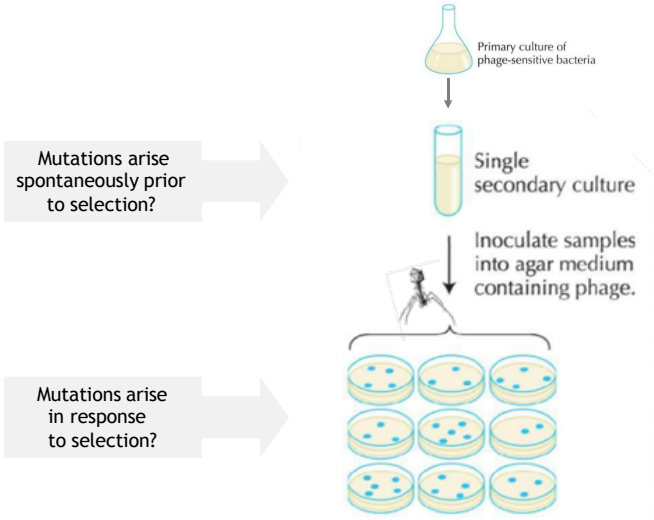
1. Mutations arise randomly: finite probability for any b to mutate during lifetime from S to R and every offspring will be R unless reverse mutation occurs
2. Mutations arise due to a specific physiological response to selective pressure: small finite probability for any b to survive attack by virus, and immunity is conferred to offspring. Note that clones of the original survivor b are not guaranteed or likely to also survive, we cannot infer this

Bacterial Genetics

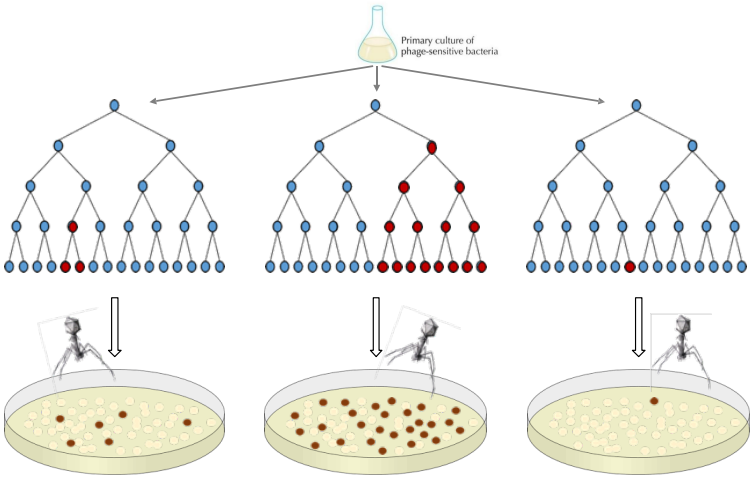
2: The nature of mutations

- Luria Delbruk Flucutation Test

  - We want to understand the motive force underlying the generation of genetic diversity and how selection acts on genetic diversity
  - **Mutation Theory**: *Mutations are due to random genetic processes and are present prior to application of selection; Darwinian Inheritance and survival of most fit variants*
  - **Adaptation Theory:** *Mutants occur as specific physiological response to a selection pressure and mutants are not present prior to application of selection; Lamarckian Inheritance and heritability of acquired characteristics*
  - They collected mean and variance data and found that it was poisson distributed, meaning that the mean=variance (shows they got consistent results), for **resistant bacteria in different samples from the same culture**
    - If mutations arise in response to phage selection then the single subculture split into 3 samples —> similar



Replicate subsamples from a single secondary culture give similar numbers of phage-resistant colonies



between the replicate cultures

numbers of resistant colonies between replicate cultures (Poisson Distribution)

•Data found that the number of resistant bacteria in a series of similar cultures actually have high variance and are not poisson distributed

•If mutations arise spontaneously during growth, then we expect mutations to occur at different times in each subculture and these will be passed to the offspring; there is high variance in the number of resistant colonies

- Conclusion: Genetic mutations arise spontaneously and not in response to selection
  - Random mutations occur with a constant probability and these mutations are passed down to offspring
  - This confirms that b (and all other life) acquire genetic variation **spontaneously and randomly**, and this is the variation that natural selection acts on
    - *The underlying fuel of evolution, is derived through random chance processes*
  - It isn't possible to induce specific mutations, but it is possible to **select** for mutations

Bacterial Genomes

- Clonal Reproduction
  - Offspring are copies of their parents (**vertical gene transfer**)
  - Propagation doesn't require genetic mixing of individuals (**reproduction unlinked to genetic exchange**)
- Genetic Exchange
  - Not required, but occurs in nearly all species; rate of GE varies among species
  - Can occur within same species (recombination) and between species (horizontal/lateral gene transfer)
  - Transformation, Transduction, Conjugation (Lecture 3)
- Extremely Dynamic Genomes
  - Much larger degree of variation in gene content for b than there is for eukaryotes
  - Core Genome=shared by all strains of species and is 20-40% of all genes
  - Flexible/Accessory Genome=genes variably resent among strains of species; **mobile genetic elements**
  - Pan Genome=All genes; Core + Flexible

Mobile Genetic Elements

- Genetic elements capable of **moving to different genomic locations within a strain; transferring between different strains**
- Play central role in **evolution of microbes**, and **spread** of virulence and resistance factors
- Classes:

A. Insertion Sequence (IS) Elements

B. Transposons

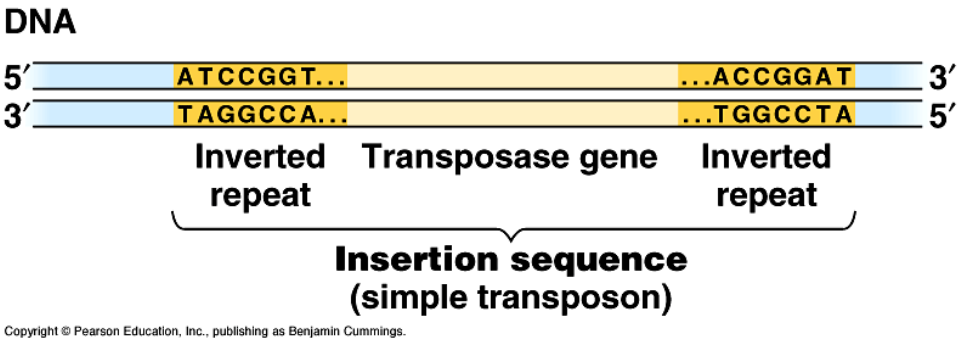
C. Plasmids

D. Bacteriophage

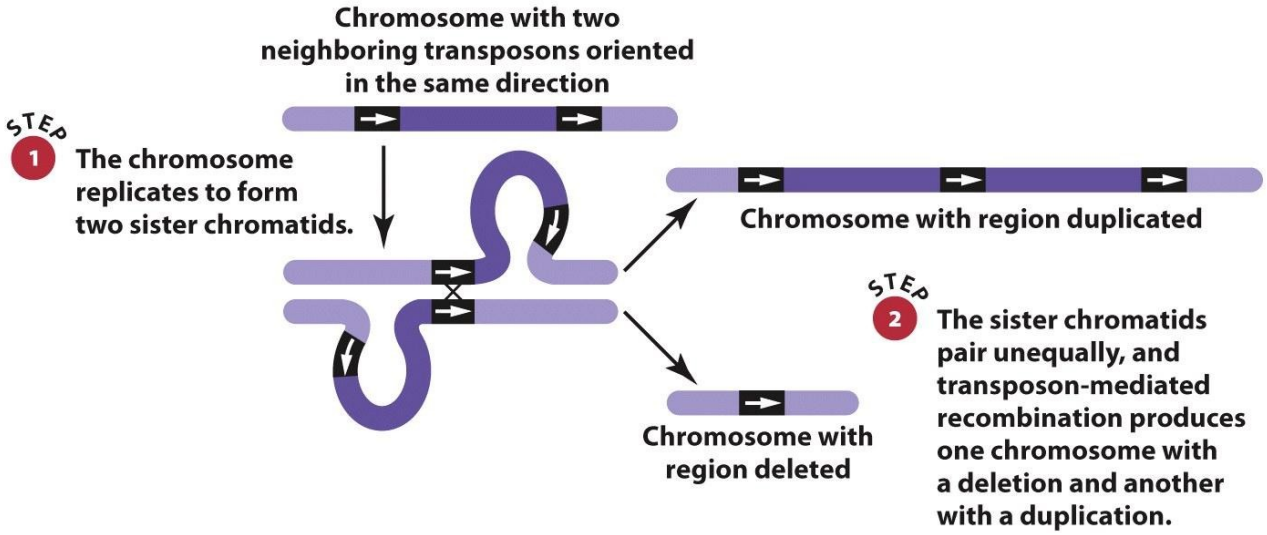
E. Genomic Island

A) Insertion Sequence (IS) Elements

- Do not cause phenotype; induce cleavage at inverted repeats; encode only a transposes gene
- Impact the expression of other genes and induce genomic rearrangements
- **Recombination between trans-IS elements —> gene duplication**

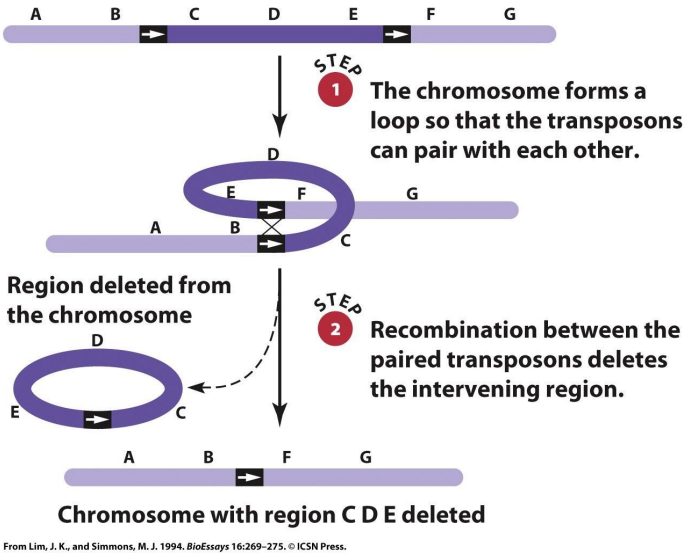


- First, a chromosome replicates to form 2 sister chromatids. Then, the sister chromatids pair **unequally** and transposon mediated RC produces one chromosome with a deletion and another with a duplication



From Lim, J. K., and Simmons, M. J. 1994. *BioEssays* 16:269–275. © ICSN Press.

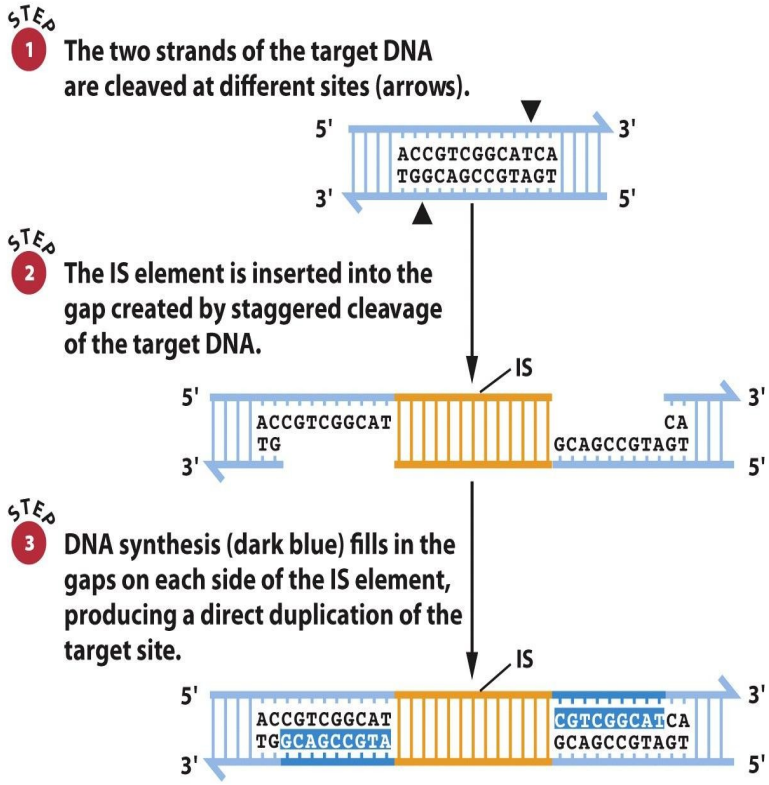
- **Recombination between cis-IS elements —> deletion (through cross over event in RC)**
  - First, the chromosome forms a loop, so that the transposons can pair with each other. Then, RC between the paired transposons deletes the intervening region
  - Deleted region CDE is a **plasmids**



From Lim, J. K., and Simmons, M. J. 1994. *BioEssays* 16:269–275. © ICSN Press.

B) Transposons

- These are composite elements with a flanking pair of IS elements
- They have cargo genes encoding adaptive proteins like antibiotic resistance or toxins
- A transposon is a sequence of DNA that can move to new positions within the genome of a single cell. Transposition can create significant mutations and alter the cell's genome size.
- First, the two strands of the target DNA are cleaved at different sites, then the IS element is inserted into the gap created by staggered cleavage of the target DNA. Finally, DNA synthesis fills in the gaps on each side of the IS element producing a direct duplication of the target site.



C) Plasmids

- Self-replicating elements **independent** of the organisms chromosome; encode for genes encoding adaptive proteins such as antibiotic resistance or toxins
  - Some benefit b, some neutral, some only benefit the genetic element and not the host b

D) Bacteriophages

- Bacterial viruses (non-cellular organisms); Used to infect b with their genetic material

E) Genomic Islands

- Discrete regions of the genome carrying genes involved in **adaptation** to specific **environments** (pathogenesis, symbiosis, etc.); may or may not be mobile
  - Type III Secretion System aids in pathogenesis; effectors enter the eukaryotic cell with needles then pierce the cell wall
  - EPEC (enteropathogenic) and EHEC (enterohemmorhagic) E.coli form pedestals which flatten the intestine
  - The pathogens adhere to intestinal cells (attaching) and promote microvilli destruction (effacement) and flattening
  - Upon attachment, EPEC and EHEC recruit host cytoskeletal proteins to form actin-filled membranous protrusions/pedestals

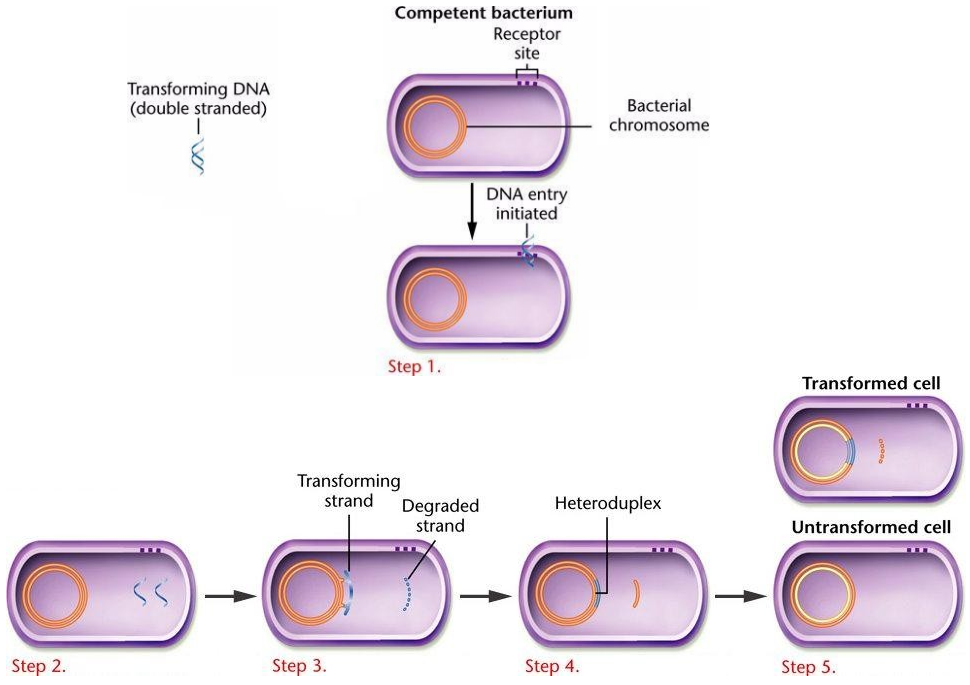
Bacterial Genetics

3-Bacterial Sex

- Three mechanisms of genetic exchange:
  A. **Transformation** (*uptake of naked environmental DNA*)
  B. **Transduction** (*phage-mediated transfer*)
  C. **Conjugation** (*plasmid-mediated transfer*) ***Lecture 4***

A) Transformation

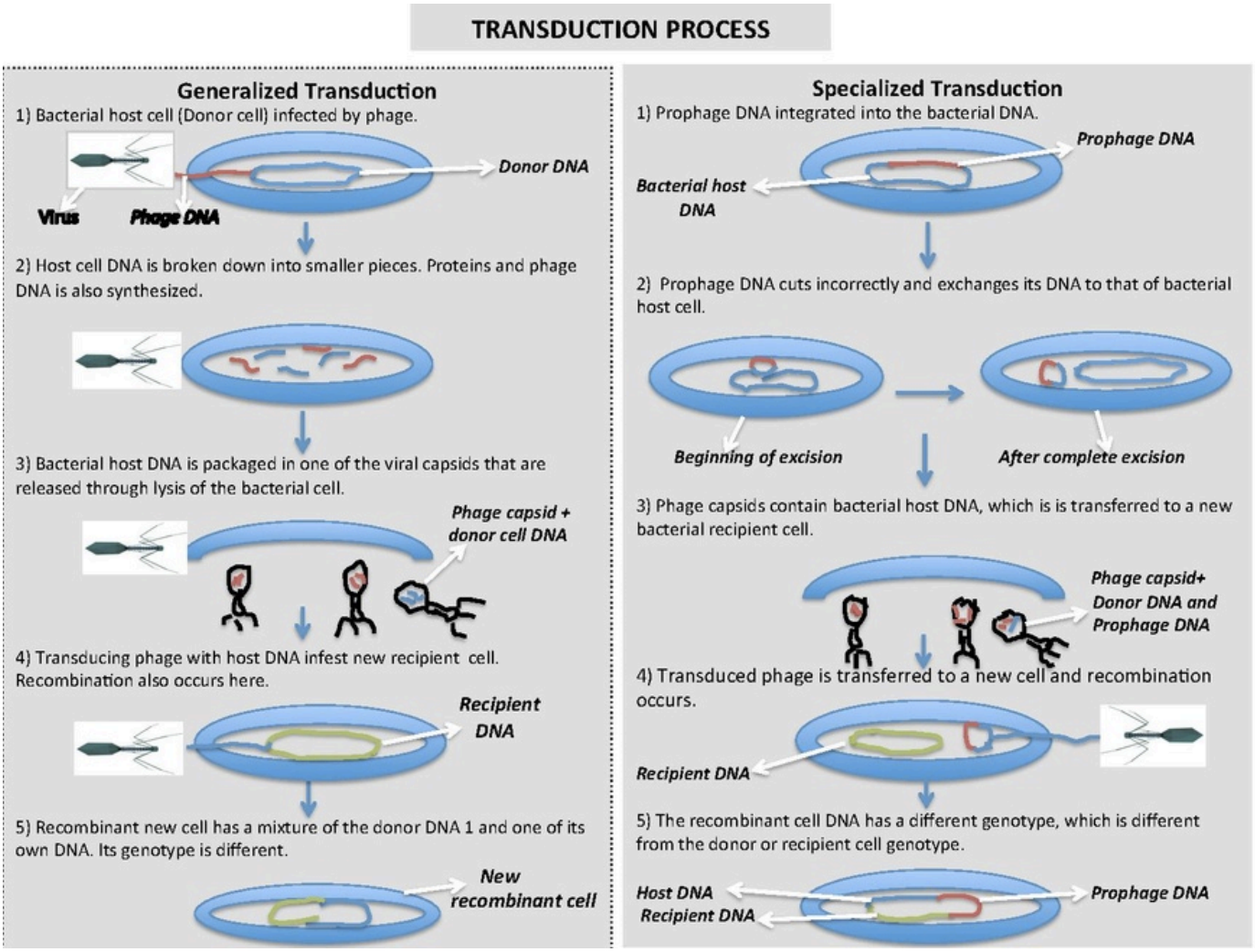
- **Frederick Griffiths** experimented with two types of pneumococcal b, S was virulent and R was non-virulent. When he injected heat killed **lysed** S cells with living R cells, the **mouse still died** (even though neither strain can kill the mouse). Griffith was also able to isolate both live R and live S strains of pneumococcus from the blood of these dead mice. Griffith concluded that the R had been "**transformed**" into the lethal S strain by a "transforming principle" that was somehow part of the dead S strain bacteria.
- Avery, Macleod and McCarty narrowed down that from the four possible transforming molecules of **proteins, polysaccharides, RNA & DNA**, only isolated, naked DNA in the environment could transform R cells into S cells (which was surprising at the time because they though that protein would do this)
- The purpose of transformation is:
  - **Genetic Exchange**; Adaptability vs Viability; Inserting a random piece of DNA into the genome is almost never beneficial; it is unlikely to evolve a mechanism that is almost always neutral or deleterious



  - **Mutation Repair**; graphs show that DNA uptake rapidly increases after b is damaged by e.g. Mitomycin or UV Irradiation
  - **Nutrition;** as [nucleotide] decreases, transformation frequency increases

B) Transduction

- As an example, the emergence of *Vibrio cholera* depends on the **sequential** acquisition of two bacteriophage (BP)
  - Toxin Co-regulated Pilus (TCP) Phage produces the **fimbria** essential for the colonization of humans and virulence; also creates receptor for CTX Phage
  - CTX Phage produces **cholera** toxin
- Lytic and Lysogenic Cycles occur when BP binds to b; **Lytic: virulent** where phage replicates and lyses the host cell; **Lysogenic: temperate** where phage DNA is incorporated into the host genome (form **prophage**), and is passed onto subsequent generations; Induction means that the prophage is induced and becomes lytic
- Phage can only integrate into b genome at **specific sites**
- **Specialized Transduction**
  - Transfer of only a few specific genes from 1 b cell to another; Phage DNA enters b and integrates into the phase integration site; errors in phage particle excision from the host DNA results in b host DNA being incorporated into the phage; only DNA that flanks the phage integration site can be incorporated (specialized)
  - The phage is now defective, but it can still inject its DNA into another b cell, and the original host b will have its DNA transduced; transduction occurs during the transition from lysogenic to lytic phase
- **Generalized Transduction**
  - Transfer of any genes from host b cell to another; a BP injects its nucleic acids/phage DNA into the host b cell; a phage enzyme is produced which breaks down the host DNA into smaller fragments; the phage DNA will replicate and phage coat proteins are produced; during production of phage heads, some may surround the host b DNA; the BP is released from the cell and it injects another b cell, leading to the incorporation of the host b DNA into the new b
  - Occurs during lytic phase, an example is if the host b DNA has an “a+” gene, allowing the b to survive as a prototroph in media without nutrient “a” present; this “a+” gene can be transduced into a new bacteria, which is “a-“, or it CANNOT survive in media without nutrient “a” present; the transduction will allow this bacteria to survive, along with its offspring.
  - “only 47% of thy+ can grow on media lacking lys —>P(lys being incorporated with thy is relatively high compared to other numbers=0.47”
    - We can map this to see how close the genes are and if they were cotransduced (two genes transduced at the same time into the b)
  - thyA+=“can grow on media lacking thyA”; “0% of lys+ b cotransduced with cys+, and we infer that cysC must lie on the other side of thyA, because thy+ cotransduced 2% with cys+
    - This means lysA is furthest from cysC, and thyA is closer, but both thy and lys are far from cys (0.47 and 0.50 chance)
    - Now we know that the order is lysA, short space, thyA, long space, CysC



(a) Donor: *thyA*$^+$ *lysA*$^+$ *cysC*$^+$

↓ make P1 lysate; infect recipient

Recipient: *thyA*$^-$ *lysA*$^-$ *cysC*$^-$

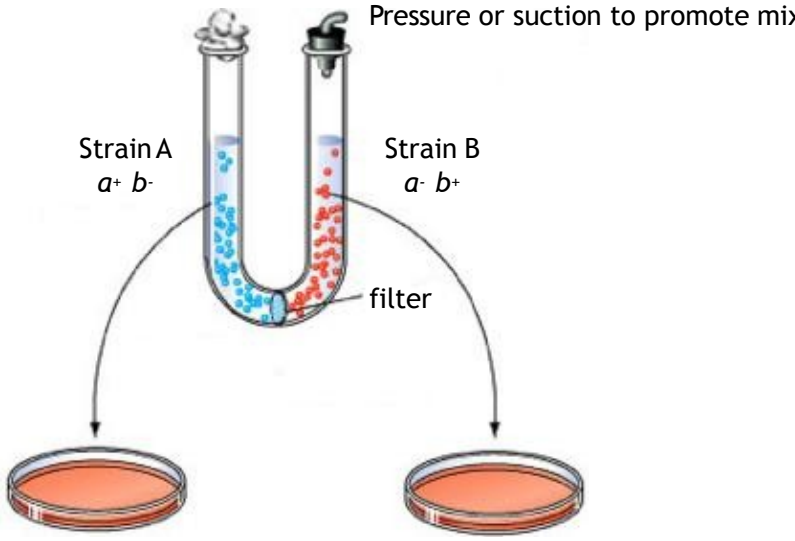
Selected marker	Unselected marker
thy$^+$	47% *lys*$^+$; 2% *cys*$^+$
lys$^+$	50% *thy*$^+$; 0% *cys*$^+$

(b)

Bacterial Genetics

4-Conjugation and the F Factor

- Plasmids and Antibiotic Resistance
  - Sex pilus=bridge to transfer DNA; Plasmids=self-replicating and independent; plasmids have adaptive genes;
  - They have a replication site to allow independent replication; they have partitioning system that allows them to separate into individual b rather than full in one b and none in another b
  - An example of adaptive gene is the gene which codes for B lactamase, makes b resistant to penicillin; there are heavy metal resistors which will not resist our heavy metal medical treatments
  - Graph shows that # of antibiotic resistance genes exponentially increasing over time; another example is with methicillin resistance, where 60% of incidences were met with resistant b; we can also use antibiotics for agriculture (e.g. cows and poultry)
- Conjugation
  - Lederberg and Tatum; They placed Strain A (a+ is wild type and b-) and B (a- mutant type and b+) into two different minimal media and found that there were no prototrophs (no GOF mutants); then they combined A and B into 1 test tube; then they found that the combination TT formed prototrophs in the minimal media as Strain AB with a+ b+
  - Possibilities: Mutation (no, bc there are no colonies of prototrophs showing that no mutants appear and no mutations occur), Genetic Exchange, or Cross Feeding; Davis later did an experiment with a filter design
  - The filter pore size only allows media and no b to pass; pressure applied to move liquid back and forth
    - When the filter was absent, prototrophs appeared
    - Filter present however, no prototrophs appeared and no growth occurred in minimal media
  - Therefore, **appearance of prototrophs requires cell-cell (physical) contact**

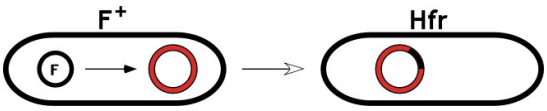


- F Factor and Conjugation
  - Lederberg discovered the lambda phage
  - Discovered F (Fertility) plasmid and it is the plasmid that moves through the filter
  - F+ is donor and F- is recipient (sexually transmitted) and is easily lost; the F Factor has genes on it that prevents infection from the recipient cell

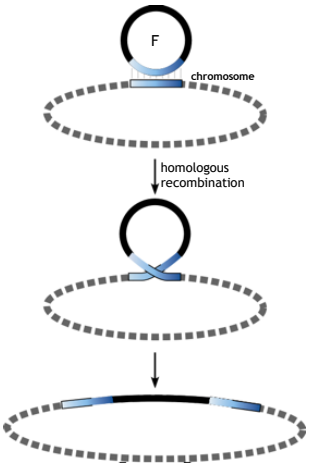
- The F Factor maintains a copy of itself in the donor strain and sends another copy (tDNA guided by relaxase) into the recipient

Isolation of Hfr strains

- Hfr=high frequency RC b with F Factor plasmid; 10^6 rates of mutation; Hfr strains readily transfer chromosomal encoded genetic markers (not just plasmid being transferred); the F Factor can integrate into the host chromosome



- F Factor and b host have circular chromosomes and **homologous RC** b/w the host chromosome and the F Factor to create a single larger circular chromosome
- RC usually occurs at an insertion element (1st mobile genetic element)

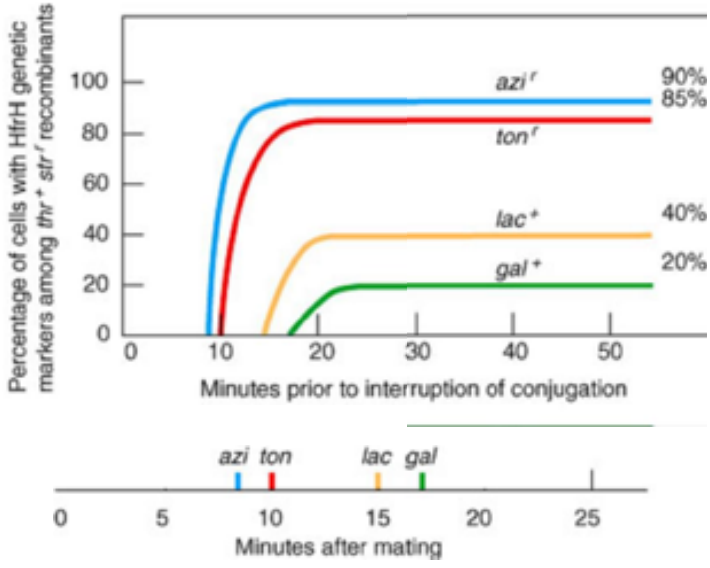


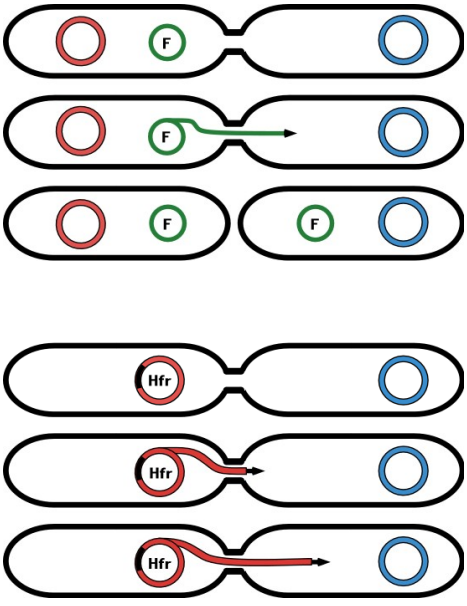
Hfr Strains show Gradient of Transfer

- F Factor has orientation
  - origin of replication=beginning of conjugal transfer
  - Terminus=end of conjugal transfer and required to be a function F Factor
  - DNA polymerase recruited to F Factor at tip of arrow until reaching Terminus
  - There is a long way to go, so it never actually makes it to the Terminus
    - This takes **time**
- We can measure genetic distance and gene order with the time aspect

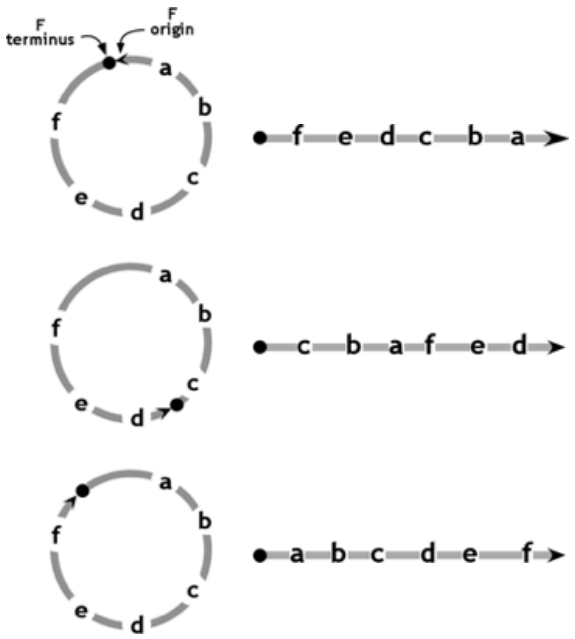
Interrupted Mating

- Hfr is strS and F- is strR; Break mating at specific time intervals (using blender to break mating bride) then plate them onto plate with streptomycin (F- is strR) and we screen only for F- that has acquired genetic material from the Hfr strain; screen for each gene with this method and you see that the gene markers appear at specific times, specific sequence and markers have different probabilities of being transferred





- E.coli example, the genome size is 100 minutes;



- Different Hfr strains have different gradients of transfer
  - F Factor can insert in different chromosomal locations and in different orientations
  - Integration size and orientation determine order of the marker transfer (replication occurs in different orientations)
    - We can measure this by looking at different Hfr strains
    - Genes further away from the origin of replication have lower probability of replication, so we can looking at multiple Hfr strains and figure out the order
- Hfr Strains Don't Convert Recipients
  - Because recipient conversion requires transfer of the terminus
  - No transfer of the Terminus
- How does the exchanged DNA get incorporated into the recipient?
  - Integration of markers into recipient requires a double RC event

TABLE 8.1

Distinguishing between the Three Parasexual Processes in Bacteria

Recombination Process	Criterion	
	Cell Contact Required?	Sensitive to DNase?
Transformation	no	yes
Conjugation	yes	no
Transduction	no	no

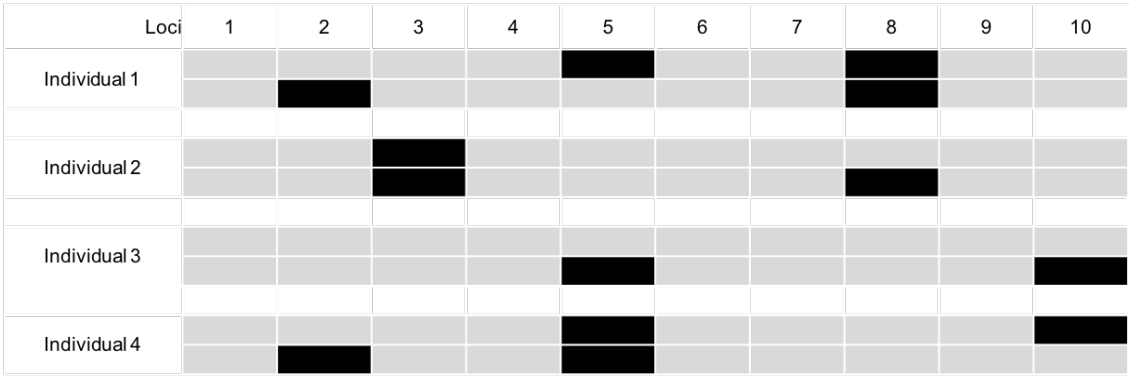
Population Genetics

1-Measuring Genetic Variation & HW

- What is Population Genetics?
  - Study of genetic variation among individuals in a population; Population=group of interbreeding individuals;
  - Gene Pool= collection of genes shared by a population of individuals
  - Some Questions:
    - What is the structure of the gene pool and how does it change over time?
    - What **biological characteristics** impact the structure of the gene pool?
      - Population Structure (variation within the population); Breeding System (haploid or diploid); Age Structure; Fecundity (over what period of time can individuals reproduce)
    - What are the **evolutionary forces** that change the gene pool?
      - **M**utation; **M**igration; **N**atural Selection; **G**enetic Drift; **R**C

	Allele Frequency (%)				
Locus	1	2	3	4	Heterozygosity
1	100				0.000
2	99	1			0.020
3	90	10			0.180
4	80	20			0.320
5	70	30			0.420
6	60	40			0.480
7	50	50			0.500
8	50	40	10		0.580
9	50	30	20		0.620
10	50	25	25		0.625
11	50	25	12.5	12.5	0.656
Total Heterozygosity					0.400

- Genetic Diversity
  - SNPs (Single Nucleotide Polymorphisms): mutations that change 1 letter that can change or might not
  - Insertions/Deletions (InDels): entire region of genome gets cut out or even inserted
  - Copy Number Variation (CNV): two copies of the gene
  - Structural Variation (SV):



•Locus: specific location in the genome

-Monomorphic loci: locations with no variation e.g. 1, 4; Polymorphic loci:

locations with multiple variants e.g. 2,3

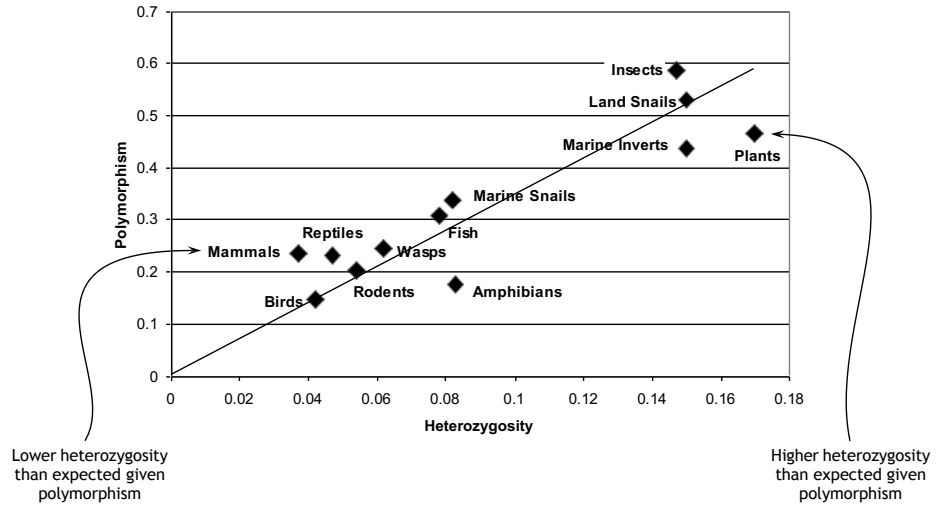
- Allele: different form of the same locus
  - Major Allele: Allele at the highest frequency e.g. grey; Minor Allele: low frequency allele e.g. black

- Heterozygosity
  - Polymorphism: co-occurence of 2 or more alleles at a locus within a population e.g. 2,3
  - Heterozygosity: probability/fraction of individuals expected to be heterozygous at a particular locus given the allele frequencies at that same locus (**HW**); population statistic NOT a statement of genotype (bc we're looking at frequency in a population)
    - A locus can be polymorphic, have no heterozygous individuals, and still have heterozygosity
    - Heterozygosity NOT EQUAL TO Heterozygote
  - **IMPORANT EQUATIONS:**

$$H = 1 - \sum_{i=1}^{k} p_i^2 \qquad H = 1 - \frac{1}{m}\sum_{l=1}^{m}\sum_{i=1}^{k} p_i^2$$

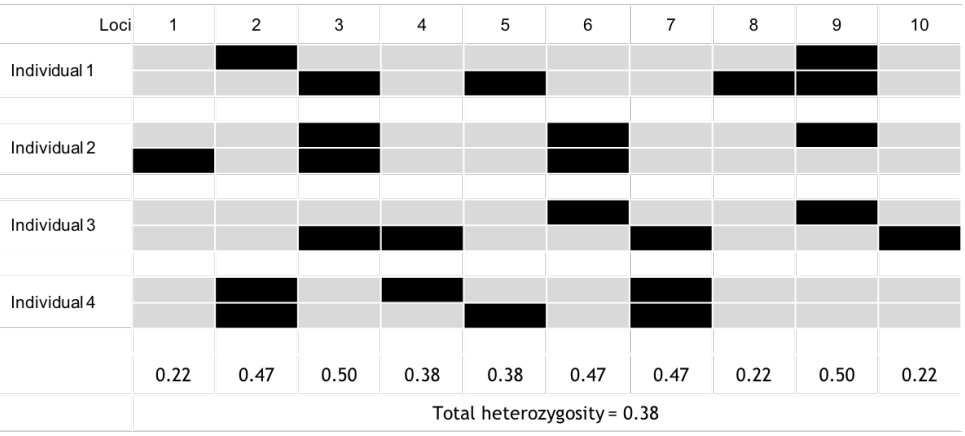
$H = 1 - (0.99^2 + 0.01^2)$
$H = 1 - (0.9801 + 0.001)$
$H = 1 - 0.9802$
$H = 0.0198$

- where p_i is the frequency of the i^{th} of k alleles
- where m is the number of loci



- How can haploid organisms be heterozygous?
  - They use effective heterozygosity: probability of sampling two different alleles from a haploid population

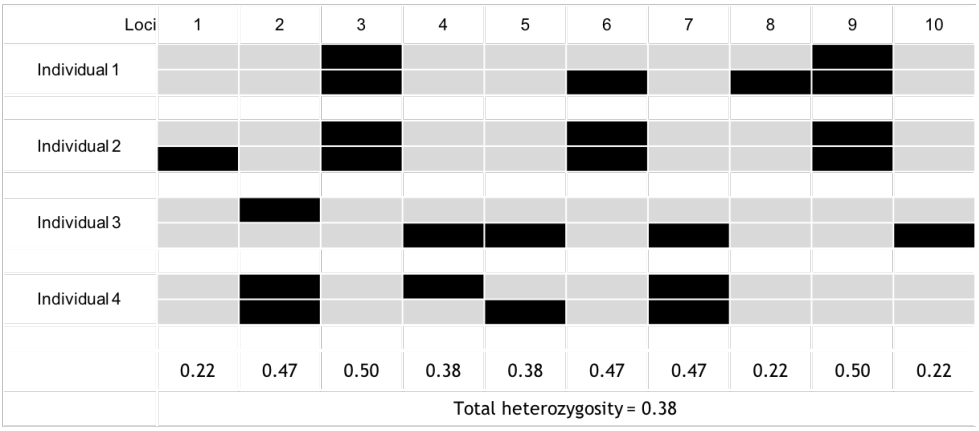
- Heterozygosity Visualized



•High Heterozygosity shared among all individuals with a **single interbreeding population** e.g.

•Same total heterozygosity, but subdivided into 2 populations is a **population substructure** e.g. 3,6 and 9 show individuals 1, 2 separate

•Same total heterozygosity, but no heterozygosity within individuals is a **high inbreeding/selfing mating system** e.g.



-Hardy Weinberg

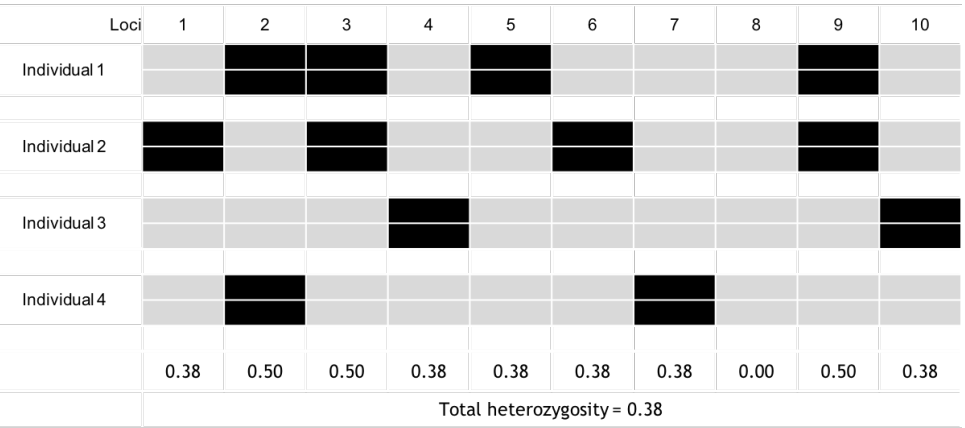
•Model for explaining how Mendelian principles influence the distribution and fate of genetic variation over time

•Strong Assumptions:

-**∞ Population Size**

-**Random Mating**

-**No Mutation, Selection, or Migration**



•Population that fulfills these assumptions will have allele frequencies that don't change over time (non-evolving population)

-HWE (Equilibrium); the null hypothesis is that it is a non-evolving population

•A population NOT in HWE must have violated an assumption; ∴ we can use the HW framework to identify what evolutionary forces are acting on a population

•Punnet Square gives us probability of getting each allele after crossing a heterozygote mother and father

•p=freq(A) and q=freq(a); p+q=1

Popul ation		Female Gametes	
		A	*a*
Male Gametes	*A*	*AA*	*Aa*
	a	*aA*	*aa*

• Population Genetics

2-HW cont., Mutation, Genetic Drift

- Hardy Weinberg

	Observed Numbers				Observed Genotype Freq.			Allele Freq.		Expected Genotype Freq.			Expected Numbers		
pop.	*AA*	*Aa*	*aa*	*n*	*AA*	*Aa*	*aa*	*p*	*q*	*AA*	*Aa*	*aa*	*AA*	*Aa*	*aa*
I	30	0	70	100	0.3	0.0	0.7	0.3	0.7	0.09	0.42	0.49	9	42	49
II	6	6	18	30	0.2	0.2	0.6	0.3	0.7	0.09	0.42	0.49	2.7	12.6	14.7
III	58	232	290	580	0.1	0.4	0.5	0.3	0.7	0.09	0.42	0.49	52.2	243.6	284.2

- Genotype Frequencies: $p^2 + 2pq + q^2 = 1$ (HWE Distribution); Example calculations below
- p= freq (AA) + 0.5 freq (Aa); q=freq (aa) + 0.5 freq (Aa) (or 1-p)
- Expected AA frequency=p^2; Expected Aa frequency=2pq; Expected Aa frequency= q^2
  - Is the **observed genotype freq** is different from the **expected genotype freq**?
  - Yes: the assumptions hold true and no evolutionary forces are acting on the population **HWE**
  - No: an assumption has been violated and some evolutionary force is in play **MMN**s**G**d**R**

- Implications of HW:
  - No change in allele freq over time (no loss of genetic variation); are the frequencies the same in t and t+1 generations?
  - HWE is obtained after just one generation irrespective of the genotype frequencies in the Parental generation IF THE HW assumptions are true!
  - Mechanism of Medelian genetics (law of independent segregation) maintains genetic variation
  - Assumptions are true because they only apply to the locus under study
- The Forces of Evolution
  - **Mutation, GD, Migration (gene flow), Selection**

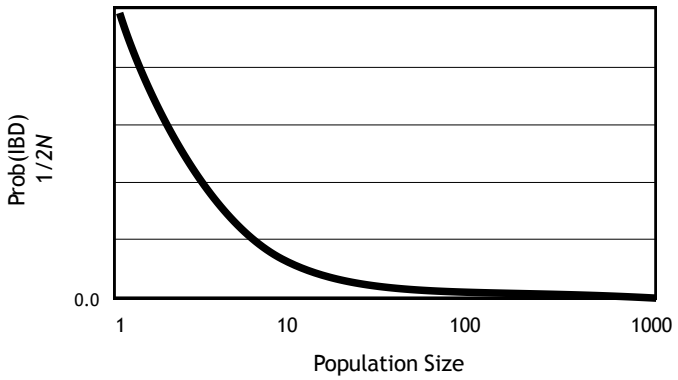
1) Mutation:

- Ultimate source of all genetic variability; occurs with very high variance
- Directional Selection
- 2 alleles: A & a; freq(A) = p0; Rate of mutational change from A —> a = μ; Freq(A) after mutation=p1
- p1=initial frequency (p0) - frequency of mutated alleles (μ*p0); $\mathbf{p_t = p_0*(1-\mu)^t}$
- **Since 1-μ < 1,** as **t —> ∞, p_t —> 0**

Population Genetics

3-Genetic Drift

- What is GD?
  - Change in allele freq due to random sampling variation between generations (stochastic sampling process; different from the deterministic process in HW)
  - Only significant in finite populations; we can relax the infinite population size assumption; magnitude of GD is inversely related to population size (bigger population means less GD)
  - Importance: it stochastically (randomly) changes allele frequencies; These changes occur without respect to the fitness of alleles or individuals;
  - The decreased heterozygosity = Increased homozygosity; Increased likelihood of exposing deleterious recessive alleles; May lower fitness of population
- Genetic Drift
  - We are looking for things that happen randomly in the population (e.g. tornado, boat crash); drastic decrease in population size (caused by bottleneck); larger population means less variation of frequency of allele
- Details
  - N (number) of diploid individuals with 2N haploid gametes (alleles) and the freq of each gamete is 1/(2N)
  - Fixation Index (F): probability that 2 alleles chosen at random from the population are ***Identical by Descent*** (IBD)
  - F=increase in homozygosity, or fixation, that results from inbreeding
    - Homozygosity=F, Heterozygosity=1-F; F changes from generation to generation under the influence of GD
    - **$F_t = 1/2N + (1 - 1/2N) F_{t-1}$**
    - Heterozygosity= 1-F=Het_t = (1-1/2N)*Het_{t-1}
      - Strongly influenced by population size, and with higher populations we see Het_t closer to 1
      - Over time, we will always lose heterozygosity due to GD



- Effective Population Size
  - Number of individuals in an ideal population that would lose Heterozygosity at the same rate as the actual population

- Number of individuals in a population contributing offspring to the next generation if **all were mating randomly**, the **sex ratio** was **equal** and **all matings** have rise to the **same number of offspring**
- **Ratio: N_e/N e.g. N=2000 (census population); Ratio=0.04; N_e=85 (much fewer individuals available to produce offspring than the entire population**

- Causes of GD reducing Heterozygosity

A. Unequal contributions to the next generation due to:

1. Non-random (assortive) mating/population structure

2. Unequal sex ratio

3. Unequal number of offspring produced from parents

B. Fluctuations in population size

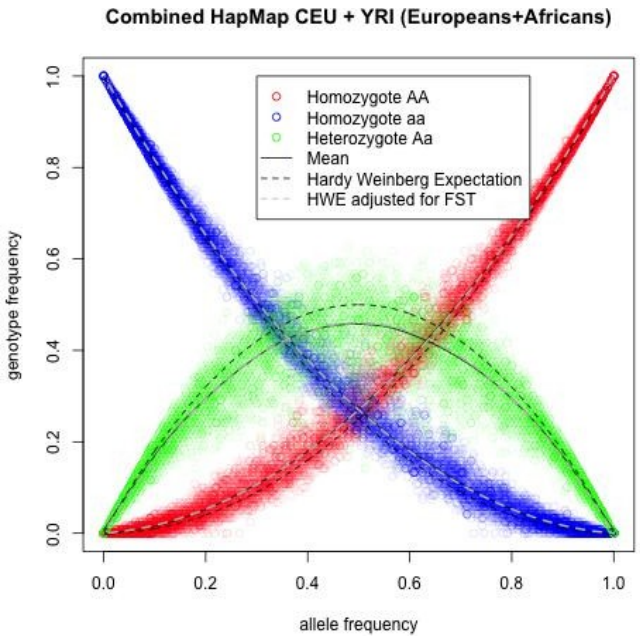
C. Inbreeding

D. Population Bottlenecks

E. Founder Events

A) Unequal contributions to the next generation: *assortative mating*

- Graph is basically saying that the mean expected does not match the observed mean because the populations are **isolated from each other (not at HWE)**



A) Unequal contributions to the next generation: *unequal sex ratio + unequal number of offspring*

- Table shows that N_e=(4xFxM)/ (F+M)

	Equal	Unequal
Females	20	20
Males (breeding males)	20 (20)	20 (1)
Census Pop Size	40	40
N Breeding Individuals	40	21
Effective Pop Size	40	3.8

B) Fluctuations in population size

- $1/N_e = (1/t)\ (1/N_0 + 1/N_1 + \ldots + 1/N_t)$
- e.g. as crop matures, more plant pathogens attacks and once you harvest, the pathogen disappears
- Breeding population can be estimated from harmonic mean of actual population (formula)

C) Inbreeding

- Mating between relatives increases IBD (increases Fixation Index) and gametes came from similar individuals
- Inbreeding Coefficient: F_{IS} (expected % of homozygosity arising from a given system of breeding/probability of IBD)

- **$F_{IS} = s/(2-s) = 1/(2N)\ (1-\text{Heterozygosity})$**

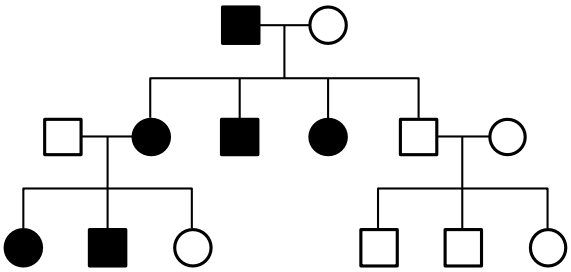
- s = selfing rate
  - completely outbred population: s=0 —> $F_{IS} = 0$
  - completely selfing population: s=1 —> $F_{IS} = 1$

- **$N_e = N / (1+F_{IS})$**
  - s=0 —> $N_e=N$
  - s=1 —> $N_e=N/2$
- Inbreeding can lead to problems in dog breeds due to lose of heterozygosity

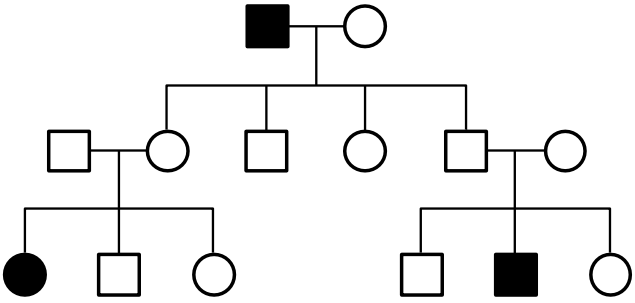
Population Genetics

4- Gene Flow, Natural Selection

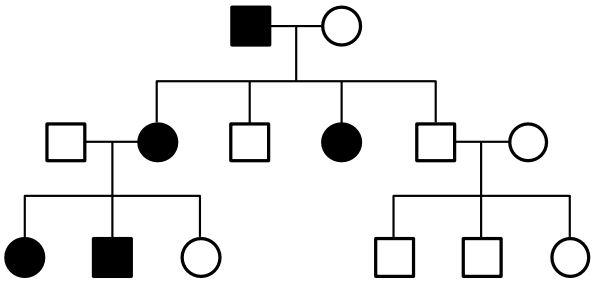
- Pedigrees
  - Autosomal Dominant Trait
    - Affects both sexes equally
    - Does not skip generations
    - Both sexes transmit the trait to their offspring
    - Affected offspring must have an affected parent
    - Unaffected parents do not transmit the trait



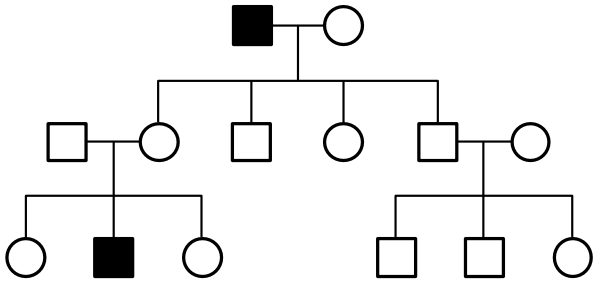
  - Autosomal Recessive Trait
    - Affects both sexes equally
    - Tends to skip generations
    - Affected offspring are usually born to unaffected parents



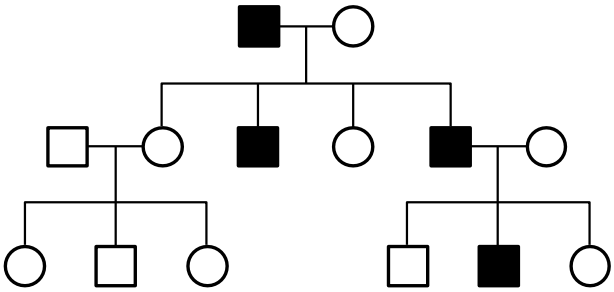
  - X-Linked Dominant Trait
    - Affects both sexes equally
    - Does not skip generations
    - Affected sons must have an affected mother
    - Affected daughters must have one affected parent
    - Affected fathers will pass the trait on to all their daughters
    - Affected fathers will not pass the trait to sons



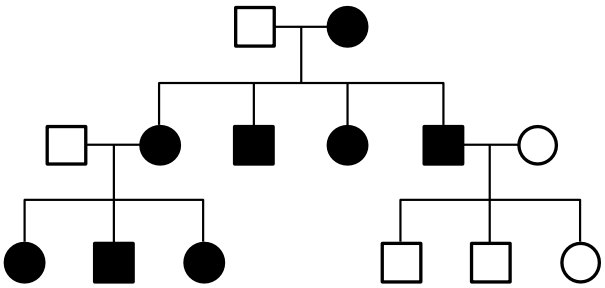
- X.Linked Recessive Trait
  - Affects males much more than females
  - Usually skips generations
  - **Affected sons are usually born to unaffected mothers**
  - Never passed from father to son
  - **Affected fathers will pass the trait (as carriers) to all daughters**



- Y-Linked Trait
  - Affects only males
  - Does not skip generations
  - Affected fathers pass it to all sons
  - Only transmitted and expressed in males

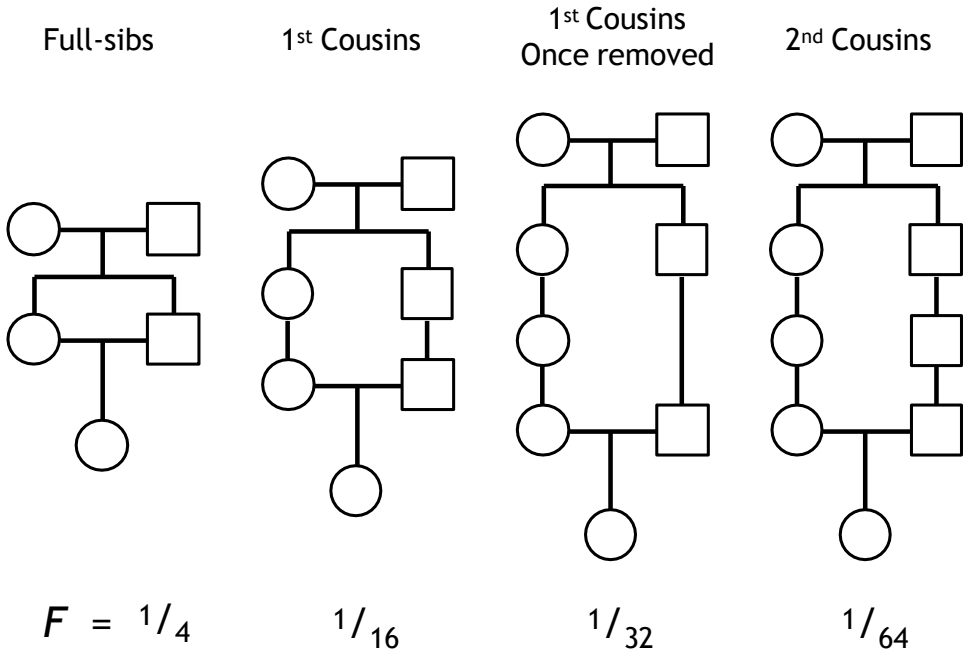


- Mitochondrial Trait
  - Affects both sexes equally
  - Does not skip generations
  - Only transmitted by females
  - Affected mother will pass to all offspring



Trait	Affects Both Sexes	Skips Generations	Transmitted From	Transmission Details
Autosomal Dominant	Yes	No	Both parents	Affected offspring must have an affected parent
Autosomal Recessive	Yes	Yes	Both parents	Affected offspring are usually born to unaffected parents
X-Linked Dominant	Yes	No	Both parents	Affected sons must have affected mother Affected daughters must have one affected parent
X-Linked Recessive	Mostly males	Yes	Both parents	Never passed from father to son
Y-Linked	Only males	No	Father	Affected fathers pass it to all sons
Mitochondrial	Yes	No	Mother	Affected mother will pass to all offspring

- Genetic Drift
  - Inbreeding is source of GD bc you get over representative of 1 allele
  - To calculate F: ([1/2)^(# of arrows)] x (# of individuals involved in inbreeding)



- F = (½)^7 x 4 = 0.03125

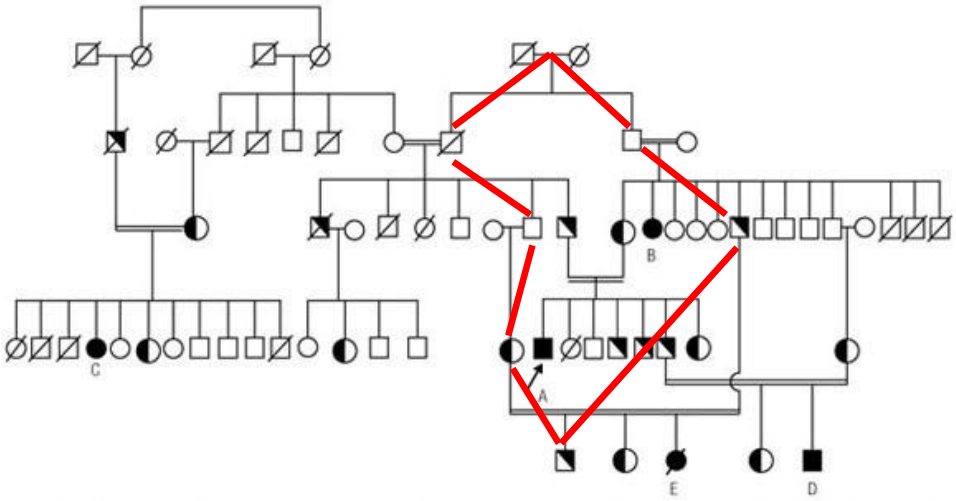


Figure 2. Pedigree chart of the Turkish cohort. Patient A: 34-year-old male; patient B: 44-year-old female; patient C: 49-year-old female; patient D: 10-year- old boy (actual ages); patient E: 9-year-old deceased girl.

D) and E) Population Bottlenecks and Founder Events

- Dramatically reduce population size due to some unfavourable events (e.g. disease)
- Founder Event: establishment of new population by small number of individuals
- e.g. of 5-alpha reductase converting testosterone into DHT
  - 5-ARD genetic males are born physical female and at puberty, the test. Masculines their bodies and genitalia
  - Occurs in Domincon Republic and Santa Dominigo, 1/90 boys born physically F
- e.g. Huntington's disease in Venezuela
  - **Autosomal dominant disease**
  - Late onset (after reproductive age); prevalence of 0.005% to 0.001% in Venezuela

- Migration (Force of Evolution)
  - Movement of alleles among partially isolated populations; m=proportion of Pop Y who migrated FROM Pop X
  - p_x is allele frequency in Population X and p_y is allele frequency for SAME ALLELE in Population Y
  - $\mathbf{p_y}$ after t+1 is $\mathbf{p'_y}$; $\mathbf{p'_y = m*p_x + (1 - m)p_y}$
    - $\mathbf{\Delta p = p'_y - p_y}$
    - $\mathbf{\Delta p = m*p_x + (1 - m)p_y - p_y}$; $\mathbf{\Delta p = m*(p_x - p_y)}$
- F Statistics and Population Structure
  - F-Statistics are fixation indices and they are a measure of homozygosity and genetic population structure
  - Three Levels of Structure: Individual (I); Subpopulation (S); Total Population (T)
  - Three Levels of Comparison:
    - If you are looking at individual fixation index compared to total population: $\mathbf{F_{IT}}$ **=Total homo = 1-Het**
    - Identity by Descent: how likely it is that you inherited an enriched allele, and therefore you have a locus that is identical by descent —> $\mathbf{F_{IS}}$ **= Inbreeding coefficient (identity by descent)**
      - Correlation of uniting gametes relative to gametes drawn at random from within a subpopulation (Individual within the Subpopulation)
    - F_{ST} = Population substructure; Correlation of gametes within subpopulations relative to gametes drawn at random from the entire population (Subpopulation within the Total population)
  - Looking at gametes because we can track gametes: $\mathbf{(1–F_{IT}) = (1-F_{IS})(1-F_{ST})}$

- FST is the primary heterozygosity stat used to measure population differentiation; it compares the level of genetic variation within 2 or more sub-populations relative to all the sub-populations combined (Total P)
  - **FST = 1 – (observed(freq(Aa)) / expected(freq(Aa)))**
  - If FST=0 —> if subpopulations are reflective of total population
  - If FST=1 —> if subpopulations are completely different from each other
  - FST=1-(0.5/0.5) = 0

	Observed							Allele Freq.		Expected						
pop.	*N(AA)*	*N(Aa)*	*N(aa)*	*N*	*AA*	*Aa*	*aa*	*p*	*q*	*AA*	*Aa*	*aa*	*N(AA)*	*N(Aa)*	*N(aa)*	*Chi-sq P*
A	150	300	150	600	0.25	0.50	0.25	0.50	0.50	0.25	0.50	0.25	150	300	150	1.0
B	150	300	150	600	0.25	0.50	0.25	0.50	0.50	0.25	0.50	0.25	150	300	150	1.0
AB	300	600	300	1200	0.25	0.50	0.25	0.50	0.50	0.25	0.50	0.25	300	600	300	1.0

  - FST = 1-(0/0.5) = 1

	Observed							Allele Freq.		Expected						
pop.	*N(AA)*	*N(Aa)*	*N(aa)*	*N*	*AA*	*Aa*	*aa*	*p*	*q*	*AA*	*Aa*	*aa*	*N(AA)*	*N(Aa)*	*N(aa)*	*Chi-sq P*
A	10000	0	0	10000	1.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	10000	0	0	1.0
B	0	0	10000	10000	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0	0	10000	1.0
AB	10000	0	10000	20000	0.5	0.0	0.5	0.5	0.5	0.25	0.5	0.25	5000	10000	5000	0.0

  - FST = 1- (0.46/0.5) = 0.07

	Observed							Allele Freq.		Expected						
pop.	*N(AA)*	*N(Aa)*	*N(aa)*	*N*	*AA*	*Aa*	*aa*	*p*	*q*	*AA*	*Aa*	*aa*	*N(AA)*	*N(Aa)*	*N(aa)*	*Chi-sq P*
A	390	426	120	936	0.42	0.46	0.13	0.64	0.36	0.42	0.46	0.13	388	429	118	0.98
B	150	400	300	850	0.18	0.47	0.35	0.41	0.59	0.17	0.48	0.35	144	412	294	0.71
AB	540	826	420	1786	0.30	0.46	0.24	0.53	0.47	0.28	0.50	0.22	509	889	389	0.01

Population Genetics

5-Natural Selection

- **Differential** survival and **reproduction** of genotypes THIS IS NATURAL SELECTION
  - **Deterministic** force that **promotes** adaptation
  - Dependent upon **fitness differences** between genotypes
  - Darwinian Fitness
    - *Relative **reproductive ability** of a genotype AND the **probability of survival** + **rate of reproduction** of an **average individual** of a **specified** genotype*
  - Environment-Dependent: the same genotype will have different fitness in different environments
- Fitness: can be modelled based on a selection coefficient
  - Relative Fitness (to whole population); Fitness of one genotype relative to a reference genotype; Reference genotype, W = 1; When Aa has W=1, we make 12 offspring, but with lower W, there are less offspring
- Directional Selection:
  - Selection for 1 allele
  - Positive: selecting for the beneficial allele (A) (A becomes fixed over time)
  - Negative: selecting against a detrimental allele (a)

	Genotype		
	AA	*Aa*	*aa*
Initial frequency	p^2	$2pq$	q^2
Relative fitness	W_{AA}	W_{Aa}	W_{aa}
Relative fitness (*s*)	1	1	1-*s*

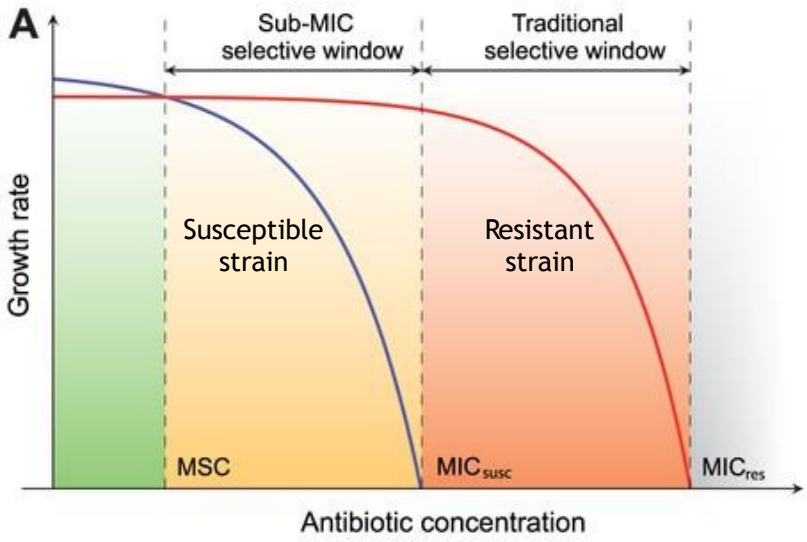
- Equations
  - s=selection coefficient = 1-W (+ selection=+s); relative intensity of selection for or against a genotype
  - p2 + 2pq + q2(1-s) —>1- sq^2
  - Find frequency after selection with modified equations like
    - AA: $p^2/(1-sq^2)$
    - Aa: $2pq/(1-sq^2)$
    - aa: $q^2(1-s)/(1-sq^2)$

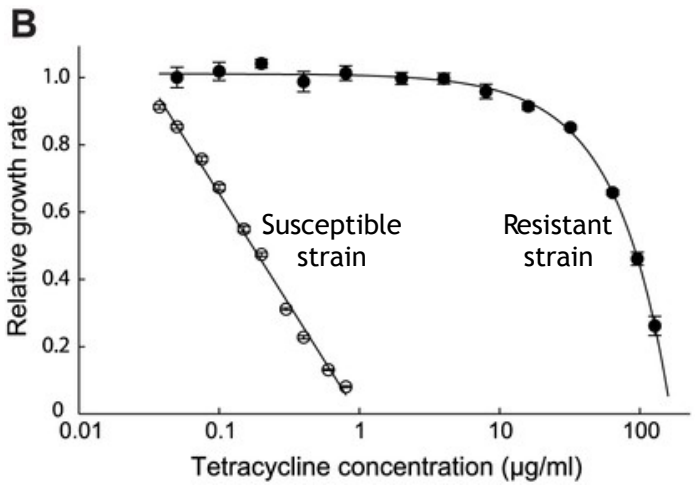
	Genotype			
	AA	*Aa*	*aa*	
Initial frequency	p^2	$2pq$	q^2	
Relative fitness	W_{AA}	W_{Aa}	W_{aa}	
Relative fitness (*s*)	1	1	1-*s*	
Freq after selection (*s*)	p^2	$2pq$	$q^2(1-s)$	$= 1-sq^2$
Norm. freq. after selection	$p^2/(1-sq^2)$	$2pq/(1-sq^2)$	$q^2(1-s)/(1-sq^2)$	= 1

Initial Frequency	0.09	0.42	0.49	= 1.0
Relative Fitness	1	1	0.2	**s = 0.8**
Frequencies after selection	0.15	0.69	0.16	= 1.0

  - In full dominance situation, we can never get rid of Aa (2pq) OR aa (q), even with s=1 and 100 generations the q will be present at 0.0001 but is non-zero

- Evolution of Antibiotic Resistance
  - It is known that resistant b are selected against at high concentrations of antibiotics —>what happens at much lower antibiotic concentration?
  - MIC: minimal inhibitory concentration
    - Lowest [ ] of an antimicrobial that will inhibit the visible growth of a micro organism
  - MSC: minimal selective concentration
    - Fitness cost of resistance = antibiotic conferred selection for the resistant mutant
    - Selection coefficient goes from negative to positive as the environment changes (as the antibiotic concentration increases)





Population Genetics

6-Natural Selection (cont.)

- Why is NS so ineffective in removing deleterious recessive alleles from the population?
  - Because so many generations required to fix allele A with selection against a
    - The rate that allele a decreases is correlated with its frequency
    - As allele **a** becomes **rare**, it spends **more time** in the **heterozygous** state
    - Selection can only act on a recessive allele when homozygous
  - ***Therefore, recessive alleles are "protected" when heterozygous***
  - When p changes from p=0 (no dominant allele) to p=0.5 to p=0.9 then p=0.99 we see Aa changing from 0% to 67% to 94.7% to 99.5%
    - Protecting recessive allele
- Why Eugenics is scientifically stupid:
  - Because it will take way too many generations to select against a recessive allele
- Balancing Selection
  - Selection that maintains multiple alleles in a population; **Heterozygote advantage** and negative frequency dependent selection

	Genotype			
	AA	Aa	aa	total
Initial freq	p^2	$2pq$	q^2	1
Relative Fitness	$1-s$	1	$1-t$	
Freq after sel.	$p^2(1-s)$	$2pq$	$q^2(1-t)$	$1-sp^2-tq^2$
Norm. Freq.	$p^2(1-s)\ /(1-sp^2-tq^2)$	$2pq\ /(1-sp^2-tq^2)$	$q^2(1-t)\ /(1-sp^2-tq^2)$	1

    - Heterozygote advantage over both AA and aa; the complicated equations are normalized frequencies to =1
  - **Equilibrium Frequency: p hat=t/(s+t); even if p(0) is different the final p hat is the same**
    - **EF is frequency that maximizes the mean population fitness**
- Heterozygote Advantage with Sickle Cell Anemia
  - HbS homozygous (HbS/HbS) = fatal with anemia and W=0.2; HbA/HbS is resistant to malaria with sickle cell anemia and W=1.0; HbA/HbA has W=0.9
  - We calculate each row, until relative fitness row, here we want the heterozygote to be 1.0, so 1.12/1.12 is 1.0; use 1.12 to divide other **Observed/Expected Ratio**s as well

- P hat = t/(s+t)= 0.877

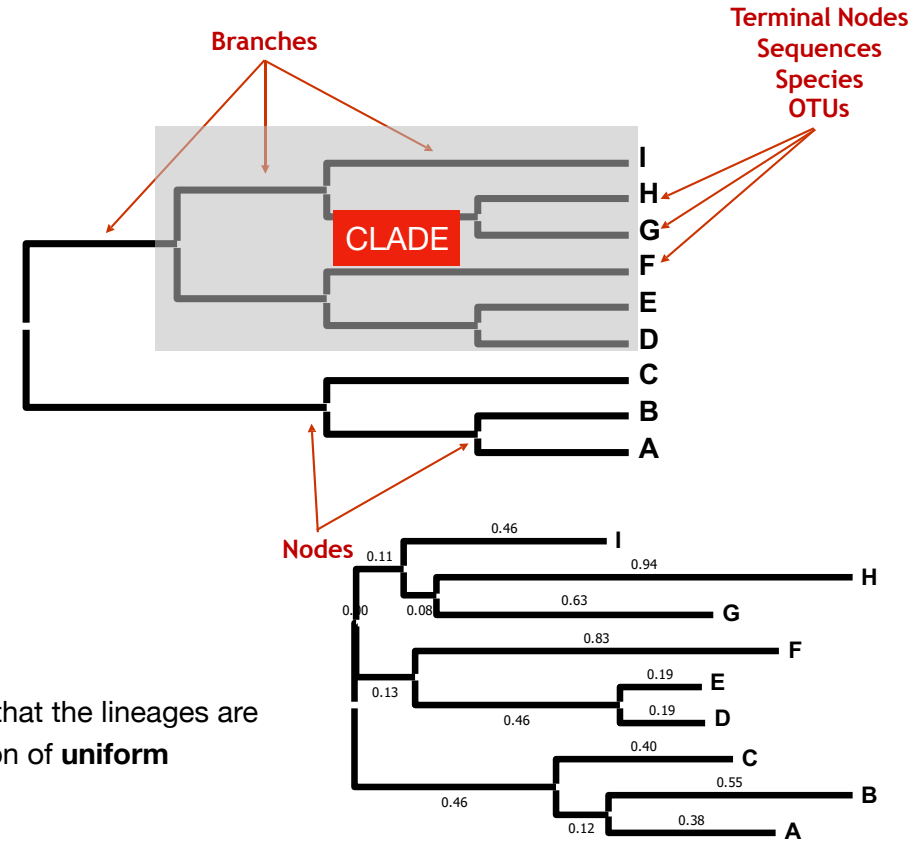
Yoruba tribe	HbA / HbA	HbA / HbS	HbS / HbS	total
Obs #	9,365	2,993	29	12,387
Obs freq	0.756	0.242	0.0023	1
Allele freq	p = 0.877		q = 0.123	1
Exp freq	p^2 = 0.769	$2pq$ = 0.216	q^2 = 0.015	1
Exp #	9,525.6	2,675.6	185.8	12,387
Obs'd/Exp'd	0.98	1.12	0.16	
Relative W	0.88	1.0	0.14	
Selection coef.	s = 0.12		t = 0.86	

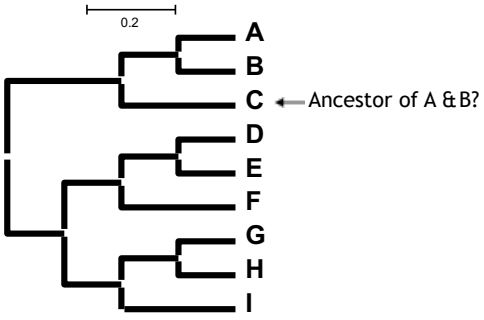
- Negative Frequency-Dependent Selection
  - Selection for an allele only when it is rare
  - e.g. As the receptor evolves to abandon its deleterious shape which binds the pathogen the pathogen also evolves it's shape to fit into the new receptor; they go back and forth where one's allele expression is dependent on the others So there is a relative connection to selection It also ties in with the red queen hypothesis where not only should the receptor adapt for it's own pace but also do it twice as fast as the pathogen might catch up
  - Red Queen Hypothesis: Organisms must constantly adapt in order to survive and reproduce since they are interacting with other ever-adapting organisms and a constantly changing environment.
  - Maintenance of multiple alleles in the population, and there are oscillations in the frequency of alleles
  - Another example is with beards vs clean-shaven: Attractiveness depends not only on the individual phenotype, but also the distribution of rivals' phenotypes
    - Clean shaven is more attractive when clean shaven is rare; same with full beard when full beard is rare

Phylogenetics

Study of Evolutionary Relationships

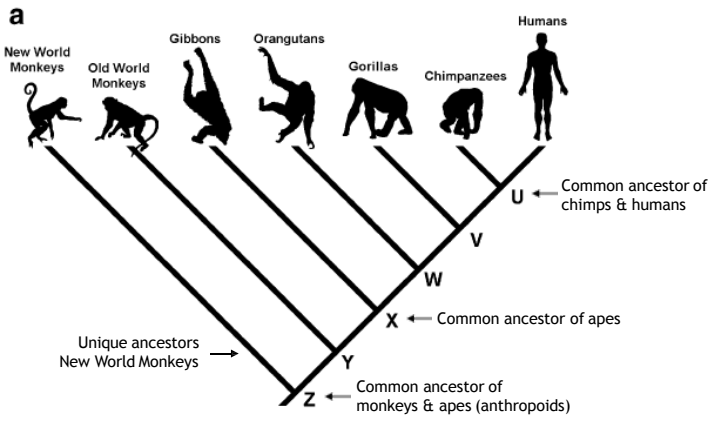
- Trees and Their Components
  - **Clade:** ***defines a monophyletic group; consists of an ancestral node and ALL its descendants***
  - **Cladogram:** ***only shows branching order***
  - **Phylogram:** ***branch lengths are proportional to evolutionary divergence***
    - From diagram, from I to H it is 1.48 (0.46+0.94+0.08)
  - Rooted Trees
    - Root provides orientation/polarity for the tree; Methods:
      - **Midpoint Rooting:** present the tree so that the lineages are balanced with the underlying assumption of **uniform evolutionary rates** among lineages
      - **Outgroup Rooting**: inclusion of 1 or more sequences that are **known** to lie **outside the diversity** of the sequences of interest
- Siblings vs. Ancestors





•Did A and B evolve from C?
NOOOOOOO

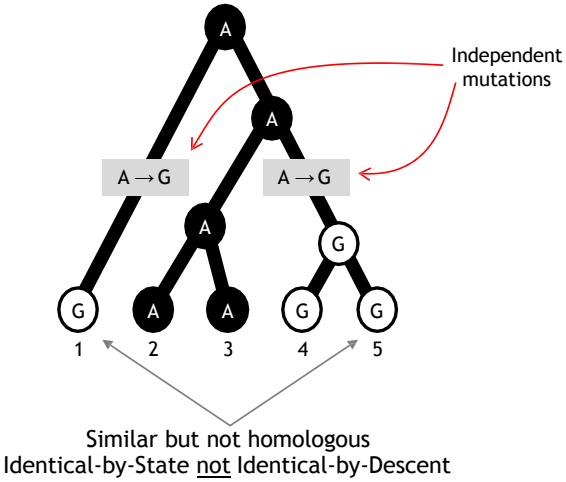
-Because all sequences are **extant**, therefore none can be an ancestor



•**Extant:** ***a species still in existence;*** **Ancestor:** ***a progenitor species giving rise to a descendant species***

•Humans **did not descend** from monkeys, but humans did **share a common ancestor** with monkeys*; all living organisms are equally evolved*

- Ancestral and Derived States
  - Ancestral State: character state (e.g. A,T,G,C) shared with common ancestor
  - Derived State: character state that has changed since the common ancestor
  - Diagram: The top A is the Ancestral State, and 2,3 are ancestral states while 1, 4 and 5 are Derived States
  - **Homology: *Descent from a common ancestor; Similarity is not the same as Homology;*** 1 is similar to 4,5 and 4 and 5 are homologous to each other
  - MRCA= Most Recent Common Ancestor



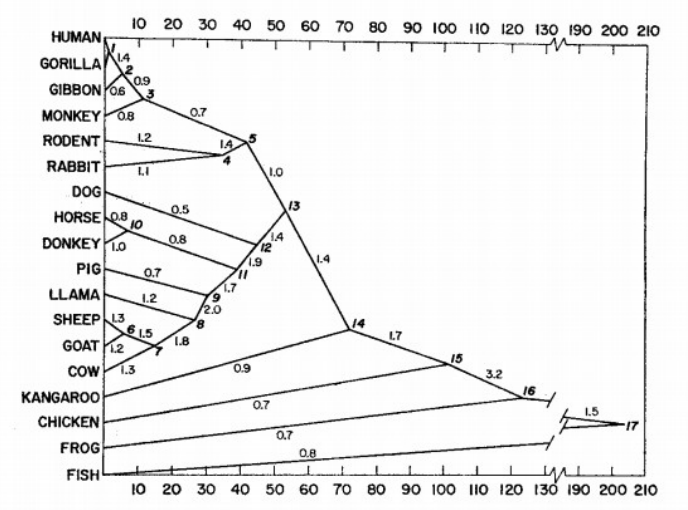
Molecular Population Genetics

Neutral Theory

- Classical vs. Balanced Schools of Population Genetics
- **Classical School —> promotes homozygosity**
  - They say that polymorphisms are **rare and transient (remain for short time)**
  - They say NS primarily **removes deleterious alleles via purifying selection**
- **Balanced School —> promotes heterozygosity**
  - They say that polymorphisms are **common and long-lived**
  - They say NS primarily **maintains polymorphisms via balancing selection**
- Apparent evidence for Balanced School:
  - Protein variants=allozymes; **Protein electrophoresis** studies of enzyme polymorphisms consistently showed very **large amounts of variation**
  - Did not find what the classical school suggested that polymorphism and heterozygosity is rare
- Observations Countering Balanced School
  - These observations appeared despite the observed level of polymorphisms
- **Segregation Genetic Load**
  - Loss in fitness caused when a population is segregated less-fit homozygotes due to a heterozygote fitness advantage
  - EQN: **Segregational Load ~ (s*t)/(s+t)**
  - e.g. SL of 0.11 —> 11% of the population dies each generation due to polymorphism (not advantageous)
  - A population cannot survive if variation is maintained at many loci via heterozygote advantage; if heterozygote advantage doesn't maintain variation … the only viable alternative is that the variants are neutral
- **Molecular Clock**
  - Individual proteins show a constant rate of AA divergence over time, and the mutations occur with clock-like regularity
  - Proteins with different functional constraints will have different proportions of neutral sites and therefore, are **free to change without affecting protein function** OR **organism fitness**
    - Critical functions and highly constrained: histones (evolve very slowly) under lots of selection
    - Moderate/Low functional constraints: firbinopeptides (evolve more quickly) under less selection

- No functional constraints: pseudogenes (evolve very fast) under no selection

- If we assume a constant mutation rate, we can estimate time since divergence based on the level of sequence divergence
- There is statistical proportionality between the time elapsed since the last common ancestor of two homologous sequences and the number of differences between the sequences (nucleotide substitution shows us that humans diverged from cows much time ago)
- Hawaiian Honeycreepers and Drosophila
  - Honeycreepers different species have different beaks developed for different types of food
  - Most recent divergence observed corresponding with most recent island formed



- In our bodies, some cells can be cancerous and unstable and accumulate more mutations; tumours from older patient has more mutations
- Molecular clock used by classical school and in summary: the clock depends only on the mutation rate and proportion of sites that are neutral
  - It does not hold true if there are varying selective pressures on a sequence

- Asking the right question:
  - How much genetic variation exists in the natural world? **A lot, but it cannot be explained due to the action of balancing selection due to:**
    - ***Extensive heterozygote advantage leading to lethal levels of segregation load***
    - ***The molecular clock indicates the most genetic variation is neutral***
  - The right question: ***how much of the genetic variation in natural populations is adaptive???***
- Neutral Theory of Molecular Evolution
  - Theory says that mutations are going to be neutral and there is no selection against them (makes no difference), but they appear after stochastic changes through GD; in addition to beneficial and positive selection and detrimental and negative selection
  - Most genetic variation doesn't affect fitness (proven by molecular clock) —> but this doesn't imply non-functionality
  - The fate of most neutral genetic variation is controlled by genetic drift

- Distribution of Fitness Effects
  - Deleterious mutations are common but are rapidly purged from the population by NS (negative or purifying)
  - Advantageous mutations are rare; Most genetic variation found within and between species is neutral
  - Molecular Clock supports Neutral Theory (proportional over time that nucleotide substitutions increase with time; the rate of increase is linear and doesn't change)

- What DOESNT Neutral Theory Say:
  - DOESN'T SAY: that organisms aren't adapted to their environments
  - DOESN'T SAY: that ALL morphological variation is neutral
  - DOESN'T SAY: that ALL genetic variation is neutral
  - DOESN'T SAY: that NS is UNIMPORTANT in shaping genomes/organisms

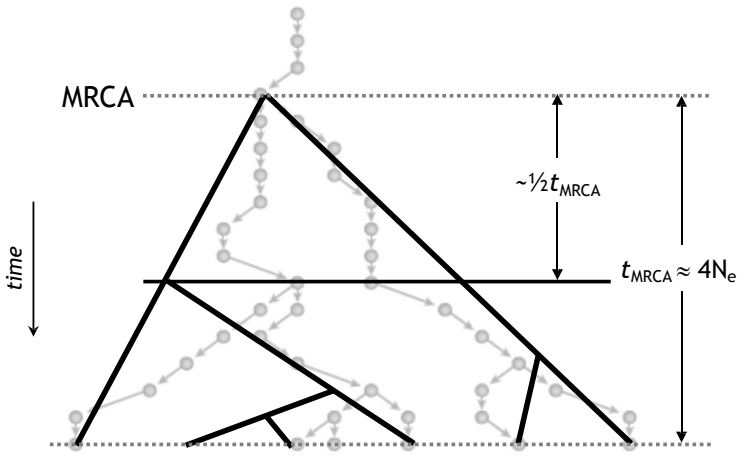
Phenotypic changes are likely to be dominated by NS and the phenotypic evolution may not always be clocklike

We have many phenotypic changes from chimpanzees

Molecular Population Genetics

Coalescent Theory + Linkage Disequilibrium

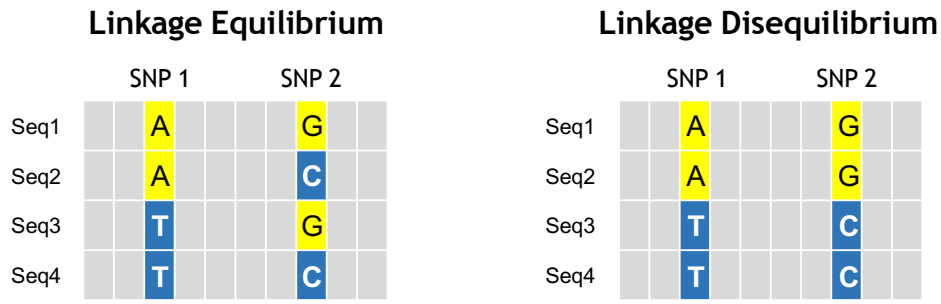
- Coalescent Theory—> going back in time and lines coalesce to ancestor
  - We want to look at lots of individuals, sample some, find out what is the relationship between the individuals, what is the shape of the evolutionary tree, what is the TMRCA, etc
  - **What past evolutionary forces have acted to give us the scope and distribution of genetic variation that we see today?**
  - **Retrospective model** (looking @past)—> classical population genetics looks at predictive models and see how evolutionary forces affect the future change in allelic and genotypic frequencies
  - The Evolutionary Processes: M M Ns Gd R
  - Demographic Factors: Population Structure, Population Expansion OR Bottlenecks, Breeding Systems
  - Biological Logic: Parents produce variable # of offspring —> variance in offspring # results in GD —> repeated GD results in eventual elimination of all genetic variation
  - Coalescent Modelling: model based on mutation and GD; the null model is based on expectation of how neutral mutations are generated and move through time
  - How it Works:
    - Sampled lineages randomly 'pick' parent alleles as go back in time—> Some parental alleles are over-sampled, while others are under-sampled—>Lineages coalesce when the same parental allele is picked multiple times—> Eventually, all lineages coalesce into a single most recent common ancestor (MRCA) and we have **key findings**
  - Key Findings:
    - TMRCA proportions to the N_e of the population IRRESPECTIVE of the # of alleles sampled (more time to MRCA means larger effective population)
    - ~0.5 TMRCA is the time for the last two alleles to coalesce, IRRESPECTIVE of the # of alleles sampled



  - The rate at which lineages coalesce depends on:
    - Size of the population: more parents, slower the rate AND
    - Number of sampled lineages: more lineages, faster the rate

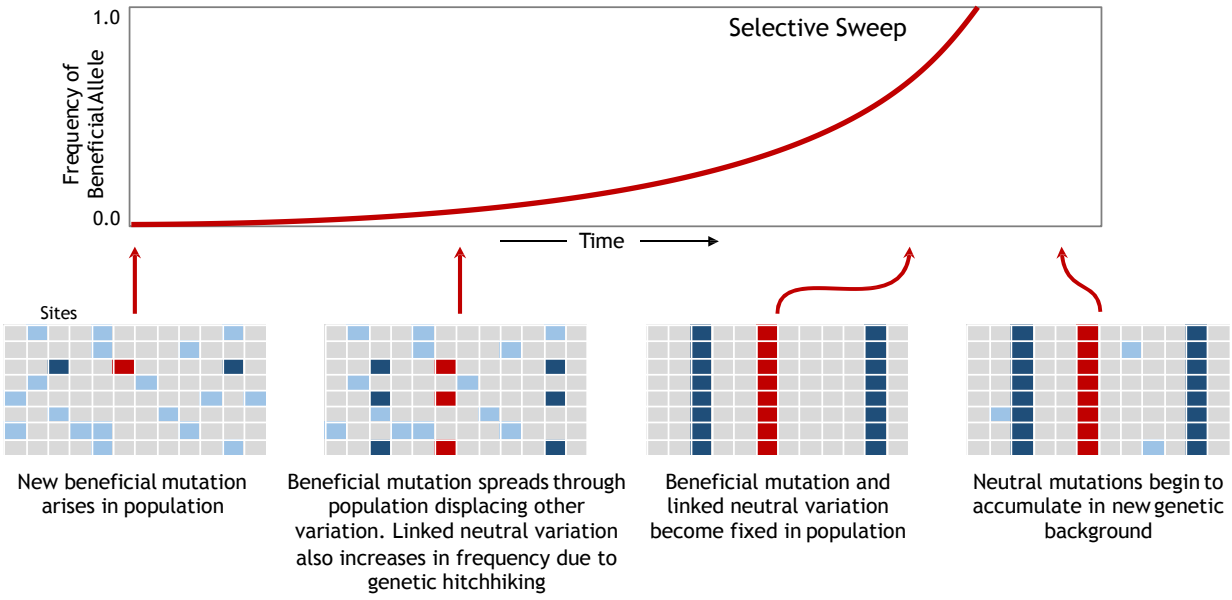
- Most of the genetic variation in a population is relatively young
- Simulations assume: a constant effective population size; different number of sampled alleles

- Linkage Disequilibrium: **non-random association between alleles at multiple loci**
  - With LD, there is an occurrence of a combo of alleles in a population more or less often than the expected combo from a random formation based on their frequencies
  - Shuffling of chromosomal sequences through recombination randomly across the genome (more likely to occur in larger genome than small genome) —>break LD



  - Why it Occurs:
    - All mutations occur in a context of a pre-existing genetic background
    - Any new mutation will be in LD with the genetic background that they arise in
  - Recombination: it is the only mechanism where mutations can be moved out of a genetic background AND a genetic background can be reshuffled —>**RC=only mechanism that can break down LD**
    - Shuffling of chromosomal sequences through recombination randomly across the genome (more likely to occur in larger genome than small genome) —>break LD and the linked genes wont be linked anymore

- Example of beneficial mutation occurring in genome w/o RC and LD occurring:



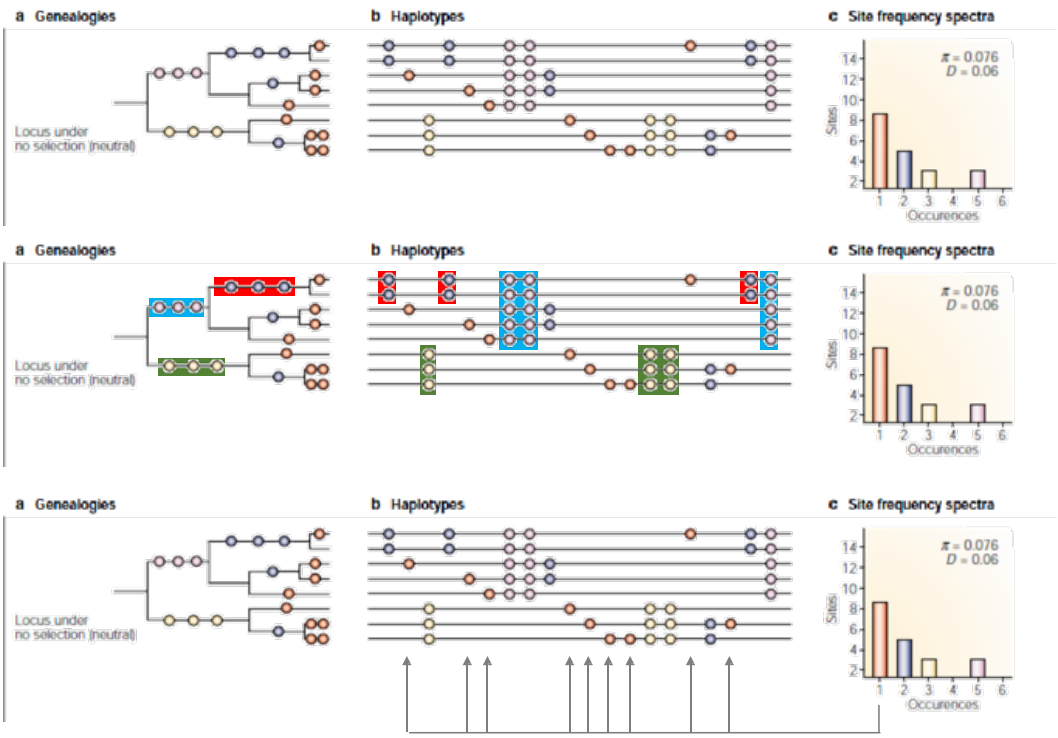
- Selection for beneficial mutation (RED)—> increases beneficial mutation's frequency in the population; linked genetic variation that is neutral (BLUE) will also increase its frequency because of **genetic hitchhiking —>** selective sweep occurs and the beneficial mutation + linked variation becomes fixed in population
  - Allele changes frequency not because it itself is under natural selection, but because it is near another gene that is undergoing a selective sweep and that is on the same DNA chain
- If there is RC at this site, there is no longer linked neutral variation and LD is broken
- Given a rate of RC, the probability of RC between any two sites increases with: Distance between sites, the time that passes (how many times mitosis occurs, generation number), and recombination rate

- Haplotypes
  - Combination/**group of polymorphism** that **segregate together due to LD**
  - Correlated genetic variation that **travels in blocks** and is inherited as a unit (when you see one allele, you always see the other one too)
  - All **new mutants** enter the population in a pre-existing (**ancestral**) **haplotype**
  - Linkage Disequilibrium among the polymorphism in a haplotype is **broken down** by **RC**
- Haplogroups
  - Group of **similar haplotypes** that share a common ancestor
  - The **common ancestor** is defined by unique mutational events
- Haplotype-Tagged SNPs
  - Polymorphism(s) that uniquely identify specific haplotypes
- Over time, haplogroups are seen as being maintained over time even though the species are branching off from each other with different ancestors

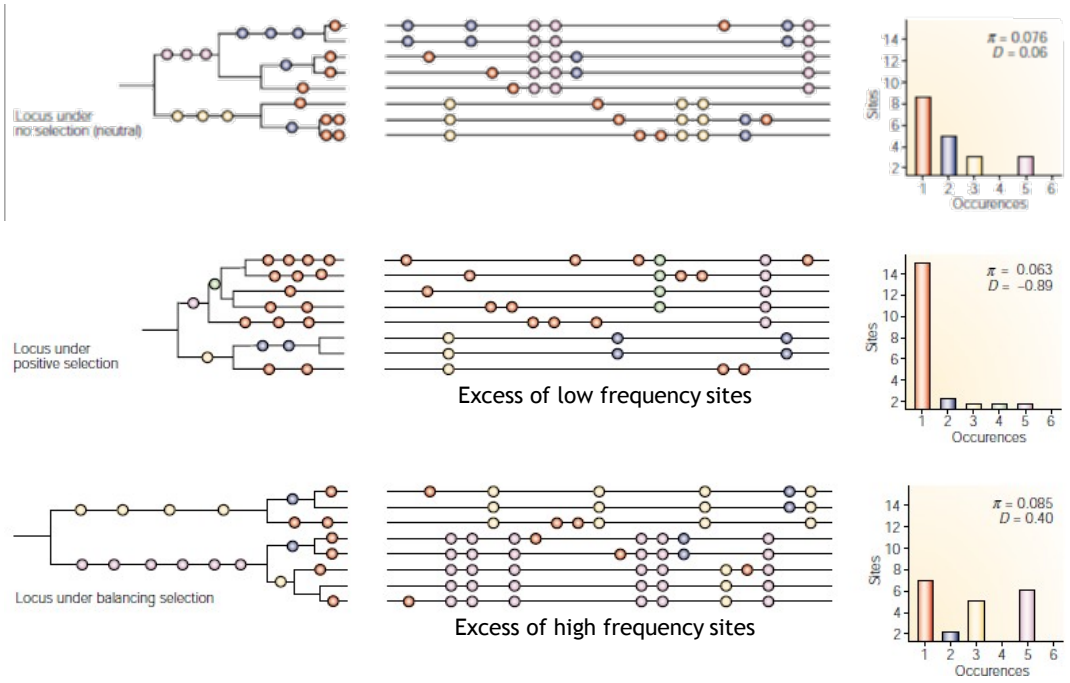
Evolutionary Genetics

Linkage Disequilibrium, Haplotypes, Selection

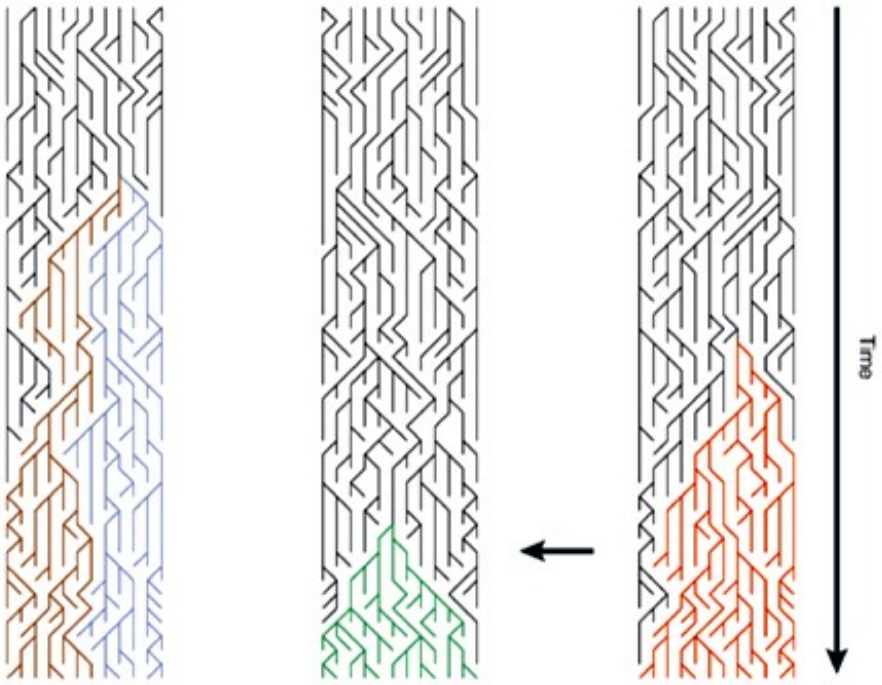
- Taking advantage of LD
  - **Data Reduction:** *allow scoring of relationships using small number of haplotype-tagged SNPs*
  - **Test Reduction:** *we only need to identify associations with haplotypes rather than with SNPs*
  - **Population Structure:** *we can quickly determine if there is an underlying substructure in the sample*
  - **Ancestry Reconstruction:** *we can infer evolutionary history based on non-recombining regions of the genome*
- Non-Recombining Genetic Material:
  - Y Chromosome to track paternal lineage AND mtDNA to track maternal lineage (autosomal/nuclear DNA recombines)
- What we can do with evolutionary genetics so far:
  - Phylogenetics —> understand how to reconstruct evolutionary history
  - Coalescent Theory —> understand how to model transmission of alleles through time
  - Neutrality —> understanding the functional impact of mutations
  - Linkage Disequilibrium —> understanding how alleles at different sites influence each other
- How to Detect Natural Selection
  - Identify genes and variants that don't meet neutral expectations
  - We look at distribution of polymorphism in the sample with 3 steps:
    - Gene Genealogies: evolutionary relationship of samples based on shared derived mutations
    - Haplotypes: polymorphisms that tend to be inherited together
    - Site Frequency Spectrum: number of polymorphic sites that occur



- Positive selection: excess of low frequency sites because of selective sweep which selection of beneficial mutation increases and other mutations arise and D=- value; coalescent theory shows it goes back to MCRA very fast (TMRCA is small and small branches
- Balancing selection: much more of the high frequency sites and coalescent theory shows that is takes longer to get to MRCA



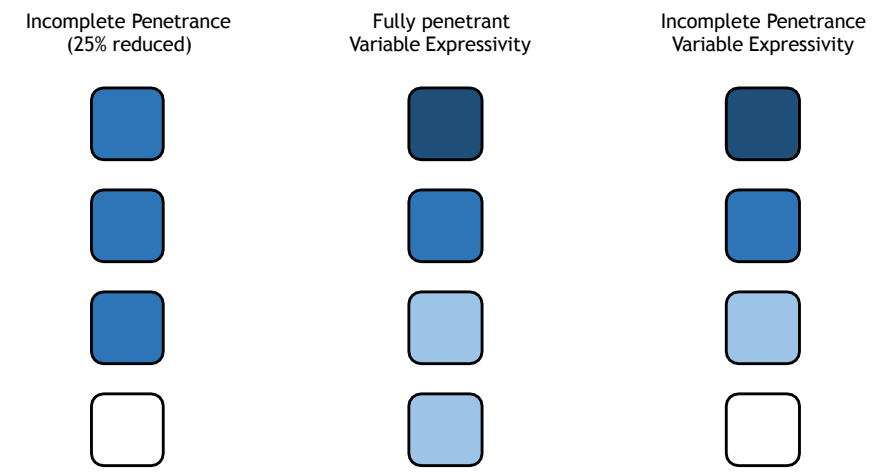
- Also detect using nucleotide base changes
  - Non-Synonymous Substitution—> AA changing —> d_N
  - Synonymous Substitution—> silent and no AA change —> d_S
  - d_N / d_S —> greater than 1 means + selection and selecting for and closer to 0 means selection against non-synonymous mutations alleles and - selection
  - FOXP2=locus with mutations that allow for speech in humans but not apes, monkeys, etc
    - From ancestors, a non-synonymous substitution led to humans talking and not chimps



Quantitative Genetics

Genetics of Complex Traits, Epigenetics, Heritability

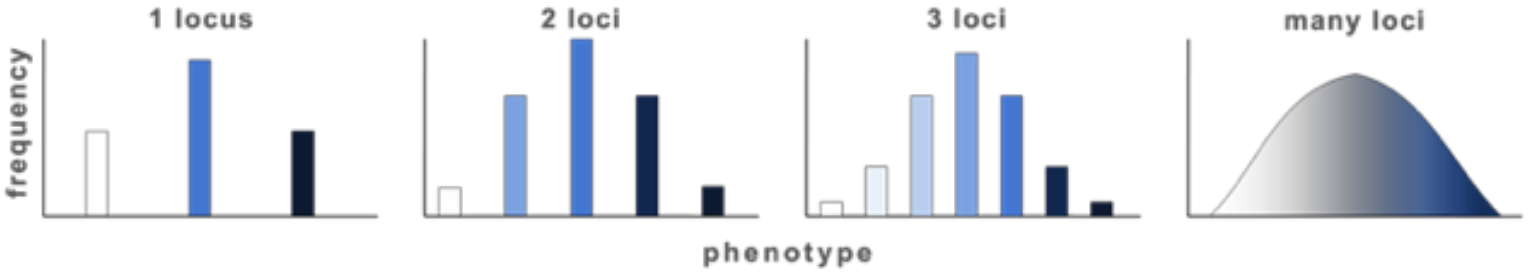
- Looking at many loci and is not discrete but continuous (like for height)
- Artificial selection has increased size of fruits like peaches, watermelons, corn; we also selected for more meaty chickens (more breast meat and are taller)
  - Chickens got taller due to selection but also better nutrition over time
- Genetics of Complex Traits
  - **Penetrance: *likelihood of showing*** *the trait given that the gene/allele is present*
    - *This is a **qualitative** measure, of whether or not a phenotype is expressed and 1 genotype may or may not display the associated trait*
    - *Example: BRCA1 and BRCA2 mutation in gene that encodes tumour suppressor proteins (that repair DNA damage) only accounts for 25% of breast cancer (low penetrance)*



  - **Expressivity: *1 genotype*** *produces* ***different degrees*** *of the phenotype*
    - *This is a **quantitative** measure of the degree to which the phenotype is expressed and the expressivity can be influenced by **genetic background, environment, or age***
    - *Example: mutation in the R-spondin 4 results either in absence of fingers/toes (anonychia) OR hypoplasia*
  - **Phenocopy: *phenotype*** *produced by an* ***environmental factor*** *rather than a genetic factor*
    - *The phenotype is **not inherited***
    - *Example: Bithorax mutation in fruit fly when the embryo is exposed to ether and the result is two sets of wings*
  - **Genocopy: *phenotype*** *produced by* ***multiple different genotypes***
    - *Phenotype **is inherited***
    - *Example: Huntington's Disease, the CAG repeats have multiple genes causing Huntington's*
  - **Pleiotropy: *multiple phenotypes/functions*** *caused by 1 genotype (Sickle cell anemia)*

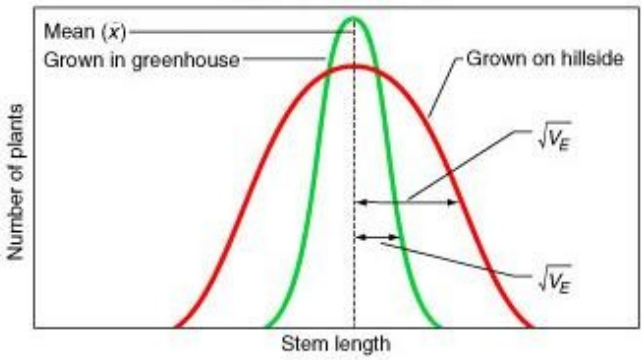
- **Antagonistic Pleiotropy:** *gene that **changes the sign of its selection coefficient** (going from beneficial to detrimental) over the course of the organisms life cycle*
  - *Initially, Huntington's Disease patients have high fertility early in life but later they have symptoms of the disease*
  - *Theory of Aging: selective value of a gene depends on how it affects the total reproductive probability and also selection is weakest of genes expressed when organisms have no reproductive potential (Huntington's Disease comes back at old age when you may no longer be reproducing)*
  - *Cancer-Aging Hypothesis: cancers and aging are related in that they both involve the success or failure of tumour suppressor mechanisms that control cellular senescence (organ breakdown)*
    - *Tumour suppressor gene p53 is induced by stress signals (like DNA damage) and results in repair of damage AND suppression of abnormal cell proliferation (cell growth and replication) leading to accelerated aging*

- Polygenic Traits
  - Quantitative=height
  - Polygenic Traits=quantitative traits controlled by alleles at multiple loci
    - Complex quantitative traits can be explained by Mendelian inheritance (discrete categories) if several genes affect the trait; example is skin colour

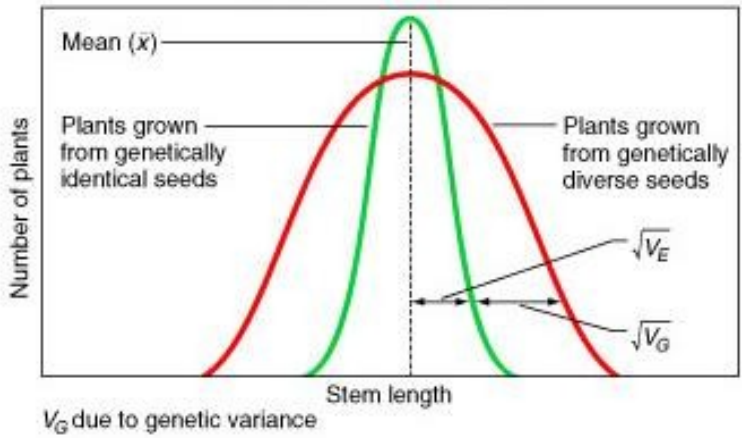


    - THEREFORE: Quantitative traits can be polygenic traits!

- Genetically Identical Plants (grown in greenhouse and hill)
  - Plants from natural environment have V_E as being higher than those from the identical and uniform environmental conditions with $V_E=0$



V_E is smaller among greenhouse-grown plants.



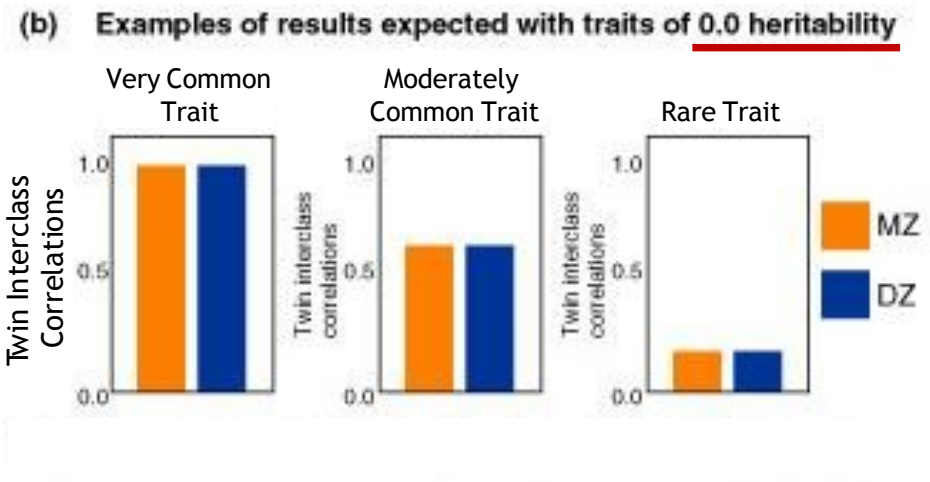
V_G due to genetic variance

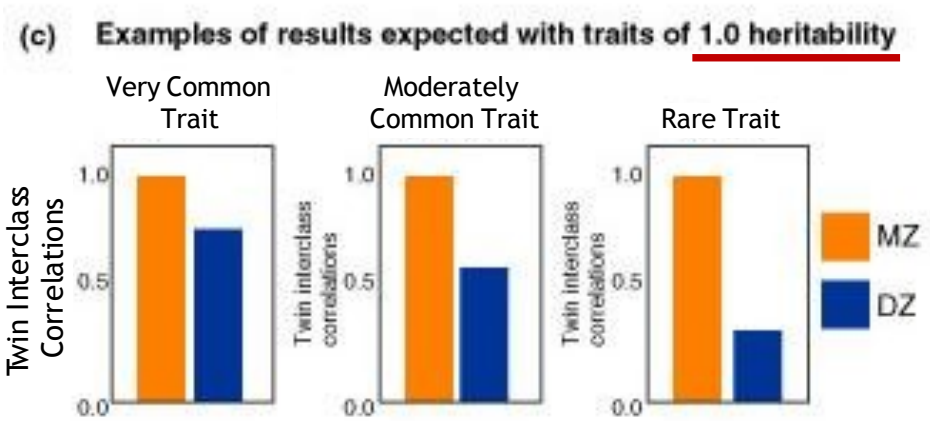
•Phenotypic variance = $V_P = V_E + V_G$

- Heritability:
  - EQUATION:
  - Proportion of the phenotypic variance attributable to genetic variance
  - 0 means no correlation between parent and offspring and 1 means high correlation between p and o

$$h^2 = \frac{V_G}{V_G + V_E} = \frac{V_G}{V_P}$$

- Heritability is not deterministic
  - Monozygotic and Dizygotic Twins Study
  - MZ=genetically identical and DZ aren't
  - MZ highly correlated with 1.0 heritability but DZ less correlated
  - MZ twins: if one gets autism, it is more likely that next twin will develop autism BUT from DZ twins, if one twin gets autism, it is less likely that second twin will develop autism





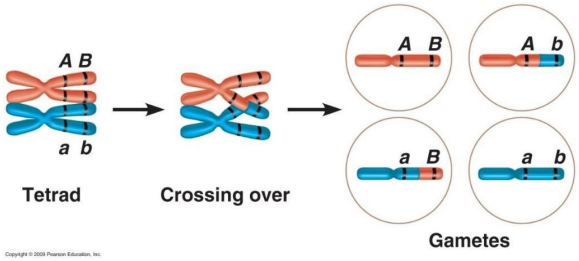
Genetic Mapping

Linkage, GWAS, EWAS, QTL

- Heritability and Potential for Evolution
  - h^2=heritability; proportion of phenotypic variance attributable to genetic variance
  - S=selection differential; strength of selection and is the difference b/w trait value of parents and trait value of total population
  - R=response to selection; amount of evolution; **$R=h^2xS$**
- Practice:
  - Hen produces on average 200 eggs/year; Heritability for egg production, h^2 = 0.5; How many eggs will be produced if you select hens with higher egg production
    - Pick hens that produce 250 eggs/yr (you pick this number, can be anything)
    - S = 250-200 = 50; R = h^2S = (0.5)(50) = 25 extra eggs/yr
    - Therefore, total egg production averages to 225 eggs/yr
  - Flour beetle length; Body length shows a continuous distribution with a mean = 6 mm; A group of males and females with body lengths = 9 mm are removed and interbred; The body lengths of their offspring average 7.2 mm
  - Calculate the heritability of body length in this population
    - Selection differential S = 9 - 6 = 3 mm; Response to selection R = 7.2 - 6 = 1.2 mm; Heritability h^2 = 1.2/3 = 0.4
  - Bean Weight
  - INBRED F1 population, the variance in bean weight = 1.5; The F1 is selfed to produce an F2 population, the variance in bean weight = 6.1; Estimate the heritability of bean weight in the F2 population •
    - All F1 variance must be environmental because all individuals are the same genotype
    - F2 variance is combination of environmental and genetic because all genes that are heterozygous in F1 will segregate in the F2 pop to give an array of genotypes V_E = 1.5 V_G +V_E =6.1=V_P therefore, V_G = 4.6 h^2 = 4.6/6.1 = 0.75
- **Important caveats**
  - **Heritability=property of populations, not individuals; Heritability says nothing about whether a gene influences a trait, only the extent to which genetic variation contributes to phenotypic variation; An estimate of heritability only applies to the environment in which it was measured in**

- Epigenetics: Developmental Genetics
  - Genes interact with an epigenetic landscape to produce a phenotype; goes from undifferentiated to differentiated
  - Originally applied to development of the organism, but has current relevance for stem cell research
- Epigenetics: Heritable Effects
  - Dynamic remodelling of the genome by DNA methylation
  - Methylation of DNA has silencing effects on genes, induces a switch in gene activity, and takes part in cell fertilization
- Epigenetics: Molecular Genetics
  - Chromatin based epigenetics; alterations to DNA like methylation, acetylation, phosphorylation, and sumoylation

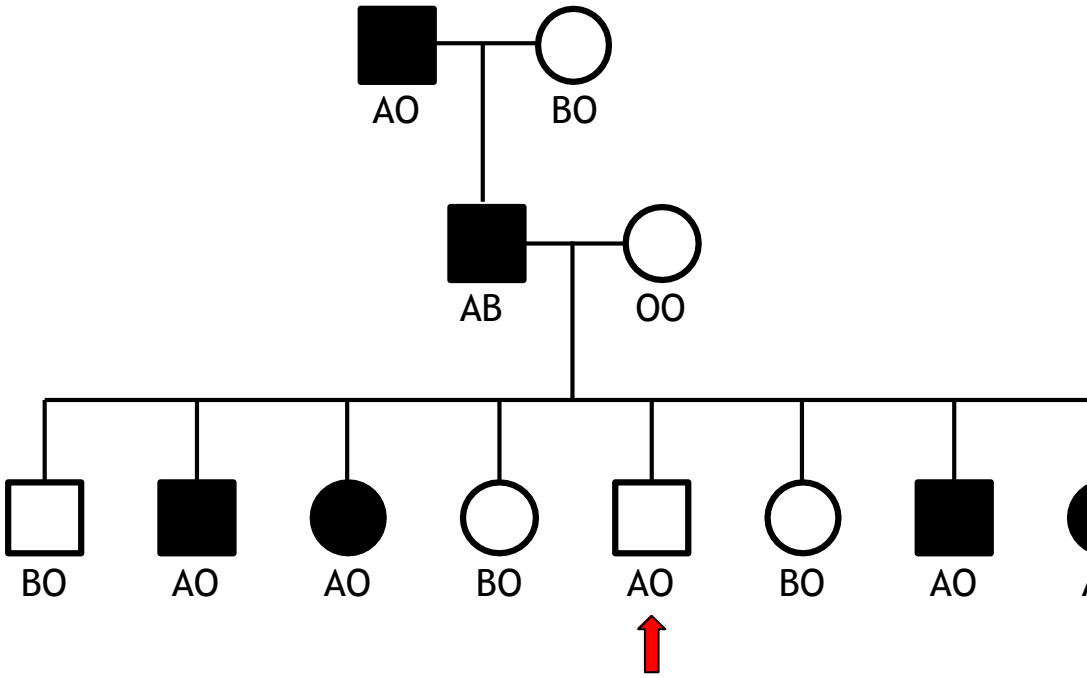
- Genetic Mapping
  - How do we find the small number of selected variants in a sea of neutral variants?
    - Neutral variation is linked to selected variation through LD; we can infer selection even when we dont know the specific selected mutation…by following associated neutral variation
- Linkage Mapping
  - A family-based (pedigree) method to identify variants underlying a trait by demonstrating ***co-segregation*** of the trait with genetic markers
    - ***Transmission of two or more linked genes on a chromosome to the same daughter cell leading to the inheritance by the offspring of these genes together***
  - Goal: identify markers that track the progression of a trait through a family
  - Theory: is that the genotype is in multiple generations of a family, and we **associate the inheritance** of the trait of interest with **genetic markers**
  - We use the RC Factor (r): ratio of recombined gametes to the total gametes produced
    - If the marker and the gene of interest are unlinked? *There is an **equal** proportion of **parental gametes** and **recombinant gametes** and **r=0.5***
    - **If the marker and the gene of interest are linked?** *There is **only** a parental gamete found and **r=0 and crossing over is less likely to separate them***



Parental genotypes: AB / ab

Recombinant genotypes: aB / Ab

- Here, the marker is on the A gene and the trait is not shown on AO, meaning that the r=1/8 (8 kids)



Recombination between ABO marker and trait

Recombination fraction, r = 1/8 = 0.125

- Logarithm of Odds Score: measure of statistical association between marker and trait
  - LOD3 means they're highly likely to be linked but LOD 0 means highly unlikely
  - With same diagram, the RC factor=0.125, the probability of a parental genotype= 1-0.125=0.875
  - Probability of observed:
    - Observe 7 parental types and 1 recombinant type: $(0.875)^7 \times (0.125)^1 = 0.049$
  - Probability of expected: $(0.5)^8 = 0.0039$
  - LOD Score = log (Prob Obsv/Prob Exp)=log(12.56) = 1.1
    - ***Trait is 12.56x more likely to be linked than unlinked***

Evolutionary Genetics

Homology

- Limitations of Linkage Mapping
  - Need to find appropriate families
  - Low Resolution
    - Works with fairly small number of markers
    - RC is fairly rare within this time frame so LD breaks down slowly
    - Information provided by each marker covers large regions of the genome
    - Identifying specific gene is a tedious and very long process
  - Strongest linkage signals are from **dominant** and highly penetrant (and thus generally rare) diseases like Huntington's Disease

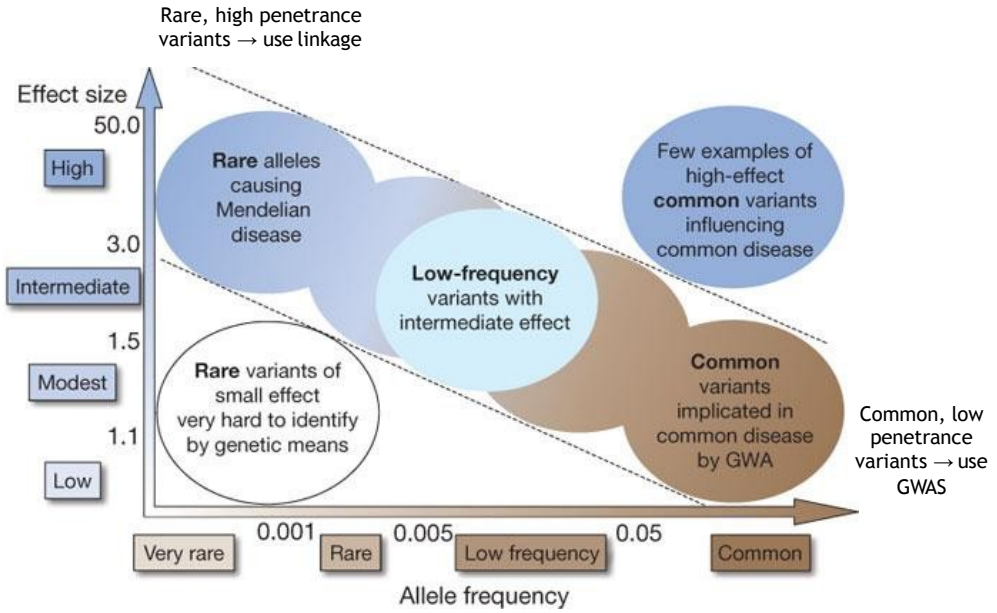
Genome Wide Association Studies (GWAS)

- Common Disease Common Variant Hypothesis
  - Many high frequency but low effect variants contribute to disease risk
    - Each variant will only have a small additive or multiplicative effect on the disease phenotype
  - May be due to:
    - Small effective population size
    - Some alleles that are now disease-predisposing might have been advantageous in the past
    - Selection pressure is weak for late-onset diseases and on variants that contribute only a small risk
- Logic of GWAS
  - A mutation associated with a trait should be more common among affected individuals (cases) than in random controls
  - Neutral markers near the mutation of interest will be associated with the trait due to LD
  - Association b/w linked neutral markers and mutation of interest should drop off with distance from mutation
- Requirements
  - Populations of unrelated individuals with cases (expressing trait of interest) + controls

  - You also need genome wide genetic variation bc SNP data from high variation genome is put through sequencing
- Linkage mapping uses related siblings but GWAS uses unrelated cases and controls
- GWAS typically finds that there is an excess of the mutant allele among the cases, relative to the controls, as well as an excess of alleles that are tightly linked to it
- How to get your results:
  - Use a Chi-Squared Test to assess association between markers and traits; then correct results for the large number of tests (multiple test correction)
  - $p<10^{-8}$ are considered significant

- HWE:
  - Critical test for all polymorphic sites in an association study
  - Deviation from HWE due to:
    - Genotyping (data collection/analysis) Error:
      - **Sporadic deviations scattered randomly around genome and NO positional bias**
    - Population Structure
      - **Deviations across large regions of the genome/associated with a haplotype and HAS haplotype bias**
    - True Association
      - **Deviations cluster in a local block due to LD and HAS positional bias**

- **Linkage Mapping vs. GWAS:**
  - **Effect Size=magnitude of difference between 2 groups (e.g. case and controls)**



Quantitative Trait Locus (QTL) Mapping

Now look at molecular markers in the phenotype of interest

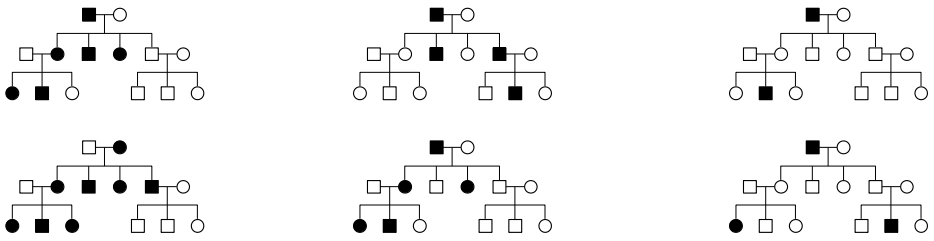
- Uncover the genetic basis of quantitative phenotypic variation
- Phenotype of interest must be variable within the mapping population
- Mapping Population:
  - Typically derived from crossing parental lines that vary in phenotype of interest
    - Analogous to linkage mapping
  - Can also use natural diversity
    - Analogous to GWAS
- Method:
  - it's basically a a statistical method that links phenotypic data and genotypic data in an attempt to uncover the genetic basis of variation in complex traits
  - Actually don't need too much: you need two individuals of whatever species you're interested in who differ in the phenotype of interest (so for example if you're studying the color of fish you would want one fish perhaps that is yellow you would want another fish perhaps that is blue)
  - Next we need knowledge of the molecular markers in the species of interest (molecular markers are what we are going to use to see what genetic locus is in this species and have a relationship with the phenotype that we are interested in)
  - One type of genetic molecular markers that we could use are SNPs but in order to do QTL analysis we're going to need some knowledge of what the common SNPs are in this species we're studying
  - Parental Generation with 2 different phenotypes —> F1 generation with 1 type of tree —> take self-cross to make F2 —>huge range of phenotypes
  - Use a probability score vs position graph and find the marker with highest probability associated with phenotype of interest; Investigate locus for genes controlling for phenotype
  - Step 1: Crosses Step 2: Genotyping Step 3: Relationship between a marker and the trait
- QTL Mapping of Floral differences found that after crossing the two species, the 2 mutants had adaptation to different pollinators via divergence in flower colour; the reproductive isolation was maintained between flower populations; this leads to sexual isolation and eventually speciation

	Linkage	GWAS	QTL
Population	Families	Unrelated cases & controls	Inbred Crosses
Trait	**Categorical**	**Categorical**	Continuous
Recombination	From family	Historical	From crosses
Statistics	Recombination fraction (Linkage)	Association	Regression
Range of Detection	Long (Mb range)	**Short (Kb range)**	**Short (Kb range)**
Number of Markers	Moderate (100s – 1000s)	**Large (>100,000)**	**Large (>100,000)**
Best Application	Rare, dominant traits	**Common traits**	**Common traits**
Types of mutations	Coding changes	**Coding & expression changes**	**Coding & expression changes**

All Equations + Key Points from Reading

Practice Questions

1. What are the advantages of using bacteria for genetics?
2. Draw an example of a GOF mutant:
3. What protein structure is required for infection by the Type III Secretion System?
4. Outline all 5 mobile genetic elements:
5. Describe Avery, MacLeod and McCarty's experiment:
6. What mode of genetic exchange is prevented by DNA-ase?
7. What is the allelic frequency p for the A allele: 40 homozygous AA; 280 homozygous aa; 80 Aa
8. What are the causes of deviation from HWE?
9. Outline the factors that lead to genetic drift:
10. Define natural selection:
11. Find equilibrium frequency, p hat, in balancing selection with heterozygote advantage where the fitness of the different genotypes is WAA= 0.88, WAa=1, Waa=0.72:
12. Pedigree ID: Explain your answer
13. Explain why natural selection is so ineffective in removing deleterious recessive alleles:



14. Which class of proteins have very slow rates of AA divergence over time relative to other proteins, outline all the proteins as well:
15. What are some observations that provide support the balanced school of population genetics? Provide criticism?
16. What is observed in linkage disequilibrium? Explain.
17. What is the best marker to choose when studying ancestry and why?
18. What is assortative mating?
19. Give an example of a highly penetrate genotype and explain:
20. What observations would supports the hypothesis that alleles common enough to be of medical significance are unlikely to be confined to one population and will have similar effects?