

Nancy Moran readings

April 1 & 3
Chapter 4 in Graur and Li

April 8 & 10
Paper to be announced

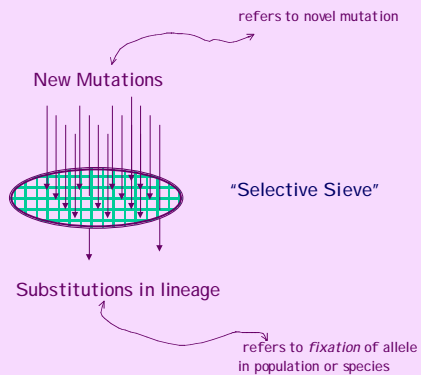
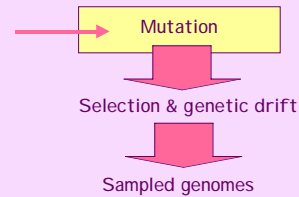
ALSO
Chapter 7 in Futuyma
The History of Life

Molecular Evolution

- Why?
- Basic genetic system universal to life as we know it
- Genomes retain a record of the past population structure and evolutionary forces affecting organisms
 - Understanding the basis of phenotypic evolution
 - Single gene level (evolution of alcohol dehydrogenase, etc)
 - Whole genome level (origins of new genes and new complexity or metabolic capabilities, loss of genes and ecological consequences)
 - Reconstructing the history of populations, species, higher clades
 - Phylogenetic reconstruction
 - Past history of population bottlenecks and expansions
 - Timing of evolutionary events/Molecular clocks

What is the basis of gene and genome change?

- Frequency and basis of mutation events <--first
- Persistence once mutation occurs <--second
 - Determined by population processes
 - Natural selection
 - Genetic drift
- Using molecular data to reconstruct the history of life <--next week



Mutations happen--

- Nucleotide base substitution, eg. C --> T
- Insertion or deletion of one or few bases, eg. ATCCGTAT --> ATCC-TAT
- Deletion of large fragment containing genes
- Duplication of a gene or a large region
- Acquisition of one or more genes or genetic elements from "foreign" source
- Rearrangements on small or large scale (inversions, translocations)
- Polyploidization

These events are ultimately the source of all variation within and among species.

Mutations are not random--

- Depends on organism
 - Some DNA polymerases more error-prone than others
 - Differences in repair genes
 - RNA viruses have extremely high rates
- Depends on position in genome
 - Regions with repeats prone to replication error, "slippage" in long runs of single bases
 - Differences between transcribed vs. untranscribed region
- Depends on environment
 - Mutagens
- Direction often biased: deletion v. insertion, nucleotide base changes
- "mutation is random" means that "mutation is random with respect to effect on fitness of the organism" -- which is, at least mostly, true.

Mutation: Single base substitutions...

- "Point" mutations in nucleotide sequence (e.g. catgcca --> catgta)
- Common
- Primary focus of molecular evolution & population genetics in past
 - Has been relatively easy to obtain single gene sequences
- Phenotypic effect depends on position in genome (e.g., silent/replacement, coding/noncoding, gene/non-gene, exon/intron)
- Many underlying physical mechanisms
 - Mismatches due to errors in replication
 - Failure of proofreading enzymes that act after replication
 - uv or other damage not corrected by repair (C->U deamination)
 - during transcription, non-coding strand is vulnerable to damage

Mutation: insertion or deletion ("indels")

- Replication slippage
 - Template and copy shift relative positions, leading to a section being skipped or copied twice
 - Can lead to frameshifts in protein-coding sequences
 - Remaining sequence has completely different amino acid sequence, may have premature stop codon
 - Usually destroys protein function
 - Deletion or insertion of repeated short sequence or single base
 - Trinucleotide repeat expansions
 - Can give amino acid repeats
 - Associated with human genetic diseases
 - E.g., 5'-CAG-3' expansion repeat --poly glutamine tract, too many copies in HD locus result in Huntingtons disease

...Mutation: role of DNA repair

- Point mutation during replication -- *E. coli* case
 - Error rate for DNA synthesis (after replication & proofreading) is ~ 1 in 10^{10}
 - Actual rate of error is 1 in 10^{10} or 10^{11} per chromosome replication
 - ~ 1 error in the genome for every 1000 chromosome replications (~ 5 Mb genome)
 - Improvement is due to DNA repair system
- Mutation can also result from damage occurring between rounds of replication
 - Results in incorporation of wrong base during next round of replication
 - Specific repair enzymes may recognize damaged base or region and remove/replace
 - DNA glycosylases, can recognize uracil, others, present in all organisms
 - urABC recognition of damage, removal of damaged strand
 - Many DNA repair genes are distributed across all of the domains of life

...Mutational rates: time units

- Time units depend on mutational process
- Can be replication-dependent
 - Will scale to generation time or number of replication events
 - "Male-driven evolution": more mutations in male lineages than female because more replication events, differential effect on sex chromosomes v autosomes
- Can be dependent on mutagens that act between rounds of replication, then result in mutation when replication occurs
 - Will scale to absolute time
- Usually both-- so time scale may be complex
 - Not strictly linear with number of replication events or time

Mutation frequency and spectrum vary among organisms due to differences in replication and repair machinery, eg --

- Bacterial genomes vary in possession of pathways affecting frequency of specific mutations
 - Cytosine deamination to uracil (C->U) can be corrected by uracil glycosylase (recognizes U in DNA, replaces with C) but this enzyme is inactivated in some bacteria
 - Can affect base composition of genome
 - mismatch repair (single nt mismatches),
 - uvr repair (larger regions of damage),
 - larger scale homologous recombination (*recA* & *recF*)

DNA repair genes in some *Proteobacteria*

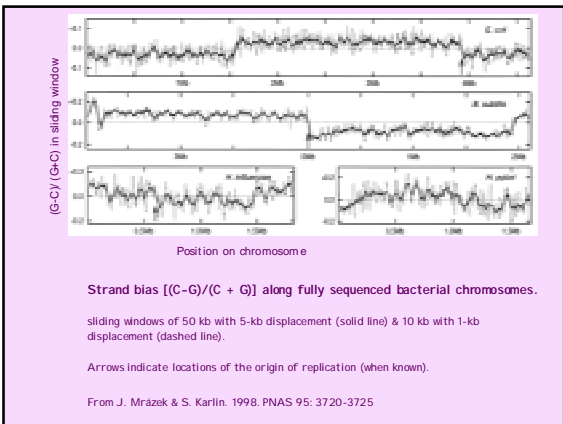
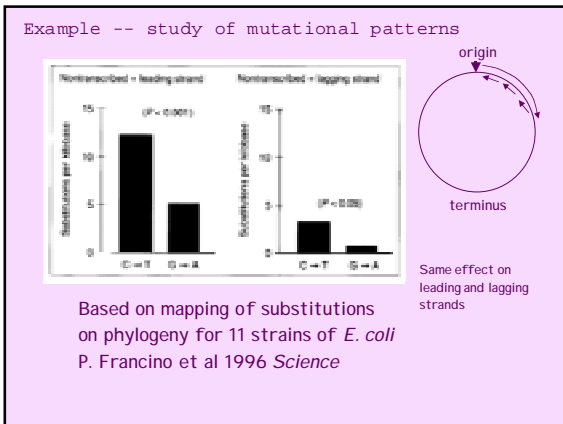
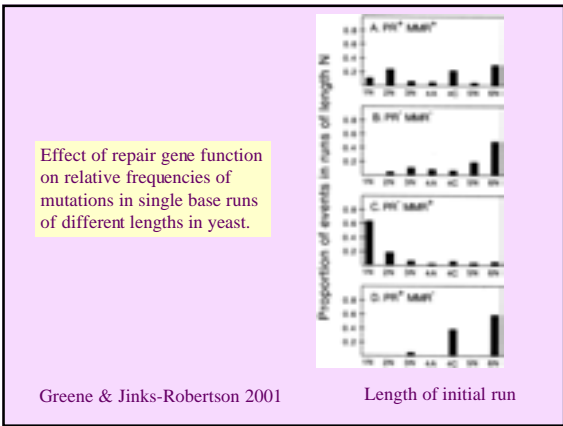
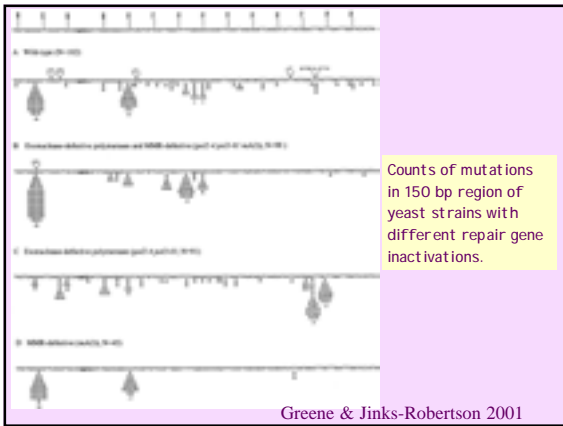
genome size(Mb):		<i>Escherichia coli</i> 4.6	<i>Rickettsia prowazekii</i> 1.1	<i>Buchnera aphidicola</i> (Ap) 0.64
Mismatch repair	<i>mutL</i>	█	█	█
	<i>mutS</i>	█	█	█
	<i>mutT</i>	█	█	█
uvr repair	<i>uvrA</i>	█	█	█
	<i>uvrB</i>	█	█	█
	<i>uvrC</i>	█	█	█
	<i>uvrD</i>	█	█	█
Recombinase pathways	<i>recA</i>	█	█	█
	<i>recB</i>	█	█	█
	<i>recC</i>	█	█	█
	<i>recD</i>	█	█	█
	<i>recF</i>	█	█	█
	<i>recJ</i>	█	█	█
	<i>recN</i>	█	█	█
	<i>recO</i>	█	█	█
	<i>recR</i>	█	█	█

Example
How do particular repair enzyme functions affect the mutational spectrum?

Study using metabolic revertants:

CN Greene & S Jinks-Robertson. 2001. Spontaneous frameshift mutations in *Saccharomyces cerevisiae*: Accumulation during DNA replication and removal by proof-reading and mismatch repair activities. *Genetics* 159, 65-75

1. Make genetic constructs having inactivated repair genes
2. Detect revertants for different strains.
3. Sequence the relevant region to characterize each mutant.



Measuring mutation rates and patterns

- Mutations are rare, on the order of 1 mutation per nucleotide site per 10^{10} replication events
- How to screen enough to calculate a rate?
- 1. Direct approach: Screen large numbers of individuals for new mutations
 - Compares many genomes to allow sufficient numbers of mutations to be identified
 - Mutations happen in the time frame of the study
- 2. Indirect approach: Compare divergent genotypes to estimate number of mutations since they diverged
 - Compares over very long time periods to allow sufficient number of replications
 - Mutations happened in the past

Measuring mutation rates directly

- Direct approach-mutation is rare, so problem is how to screen enough individuals
 - Screen for reversion mutations: revertant regains a metabolic function that is easily screened, allows detection of rare mutational events by screening large numbers of colonies.
 - Dominant genetic disease (medical profession screens human populations)

Estimating current mutations--

- Experimental studies: screening for mutations with known effects, such as reestablishment of metabolic ability, appearance of phenotype corresponding to known mutation...
 - Can examine frequency of a single base substitution in a particular gene
 - *Drosophila*, *C. elegans*, *E. coli*, yeast, mouse
 - Example above for screening mutations in yeast with inactivated repair functions
- Genealogical studies for spontaneous appearance of dominant mutation
 - Eg, human dominant autosomal diseases such as achondroplastic dwarfism with per gene mutation rate of $\sim 10^{-5}$

Mutation rates estimated from specific loci in higher eukaryotes

Organism	G	Ge	$\mu(b)$	$\mu(g)$	$\mu(eg)$	$\mu(egs)$
<i>C. elegans</i>	8.0×10^7	1.8×10^7	2.3×10^{-10}	0.018	0.004	0.036
<i>Drosophila</i>	1.7×10^8	1.6×10^7	3.4×10^{-10}	0.058	0.005	0.14
Mouse	2.7×10^9	8.0×10^7	1.8×10^{-10}	0.49	0.014	0.9
Human	3.2×10^9	8.0×10^7	5.0×10^{-11}	0.16	0.004	1.6

G Genome size in bases or base pairs (haploid unless otherwise indicated)
 Ge Effective genome size (portion in which most mutations are deleterious)

$\mu(b)$ Mutation rate per base pair per replication
 $\mu(g)$ Mutation rate per genome per replication
 $\mu(eg)$ Mutation rate per effective genome per replication
 $\mu(egs)$ Mutation rate per effective genome per sexual generation

(*Effective genome* = portion important to function, susceptible to deleterious mutation)

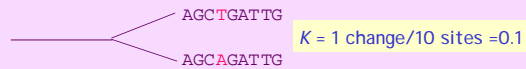
From compilation by J Drake et al 1998 *Genetics* 149:1667

Measuring mutation rates indirectly

- Estimate number of mutations in diverging lineages since time of ancestor
 - Estimates mutations in the past, accumulating over many generations
 - Can be based on pairwise comparison or on mapping of mutations onto phylogeny
 - Main issue is to eliminate effects of selection

--Estimating past substitutions (=fixed mutations)

- Examining divergence of homologous sequences



- Reconstructing substitutions on phylogenetic tree



Estimating sequence divergence (K)

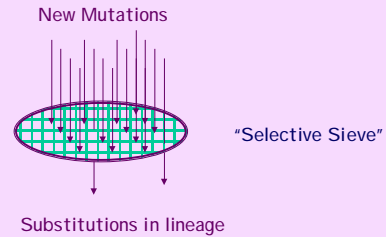
ATGCGCTAGAGGTCCTAGCTAGCATGATCGACGCGATGCAATAG
 ↓ ?
 ATGCATTAGAGATCTAGCTAGCAAGATCGAGGCGATGCGATAG

Need to adjust for multiple "hits" and reversions unless divergence is very low (when #diff = # events)

Estimate of K is based on model of substitution:
 Jukes-Cantor (model: all nt changes equally likely)
 Kimura 2-parameter (different transition and transversion rates)
 Other distributions of nt changes, especially different frequencies of mutations at different sites
 Maximum likelihood estimation of substitution model from data
 --can take base composition or other features into account

Use of sequences to infer mutational patterns/rate brings us to

How is sequence evolution governed by population level processes: selection and genetic drift?



Estimating past mutations--

- MUST use neutral sites to estimate the rates and patterns of spontaneous mutations from sequence data
- At sites under selection, fixation is dependent on selection and population size--which we generally do not know.

--Estimating past mutations

- Why do fixation events at neutral sites reflect the rate & pattern of mutation?
 - By definition every neutral mutation has equal chance of fixation, so profile of original mutation categories is the same as that for fixation events.
- Why does fixation at neutral sites reflect the rate of mutation?
 - Rate of change (r) = mutation rate for population * probability of fixation
 - For selected sites, probability of fixation depends on s & N_e , usually unknown
 - For neutral sites, $s = 0$.
 - Mutation rate per generation = μ (per individual site, gene, or genome)
 - So, chance of a neutral mutation being fixed in a haploid population = ?
 - You already know this... (REVIEW from Birky lectures, see Grauer and LI, Chapter 2)

--Estimating past mutations--

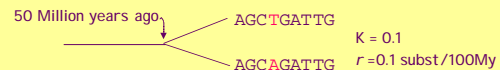
- Chance of a neutral mutation being fixed in a haploid population = $1/N$
- Mutation rate per generation per individual copy = μ
- Mutation rate per generation in haploid population = $N \mu$
- Rate of incorporation of neutral mutations, $r = N \mu * 1/N = \mu$
- *The mutation rate equals the rate of change for neutral sites.*
- In other words:
 - More mutations in a large population
 - Smaller chance of fixation for each
 - No effect of population size on fixation
 - For the most part

Using divergence at neutral sites to estimate mutation rate--

Advantages:
 easy (just sequence, or better yet, use a database...)
 more likely to give actual mutation pattern in nature, no lab artifacts

Can sometimes be calibrated with respect to absolute time
 dating ancestor can come from fossil-derived dates
 sometimes extended using phylogenetic information

Basis for "molecular clocks": using molecular divergence to date past divergence events
 (More next week)



Assigning directionality to mutations inferred from sequence comparisons

From pairwise divergence (K):
Can estimate number and rate of changes, but not direction

Assigning directionality to substitutions

Using rooted phylogenies to reconstruct ancestral states

Can count numbers of each type of change on branches based on parsimony

Example of using divergences to study mutation rates

S Kumar & S Subramanian. 2002. Mutation rates in mammalian genomes. *PNAS* 99: 803-808

Using divergences to examine mutation rates for different genes and genome regions and for different mammalian lineages.

Estimating mutation rates among genes in mammals (mouse-human)
similar mutation rates among genes as evidenced by similar silent divergences

Neutral evolutionary distances estimated by using 4X-degenerate sites of 2,019 human and mouse genes. The curve is based on observed mean and expected variance under the hypothesis of uniform neutral mutation rate among genes.

The distribution of neutral evolutionary distances estimated from genes of varying sequence lengths.

Molecular clock dating Fossil dating

Accumulation of neutral substitutions over time in diverse mammalian species pairs. (Kumar & Subramanian 2002)

Similar mutation rates in different lineages (rodents, primates, bats...)

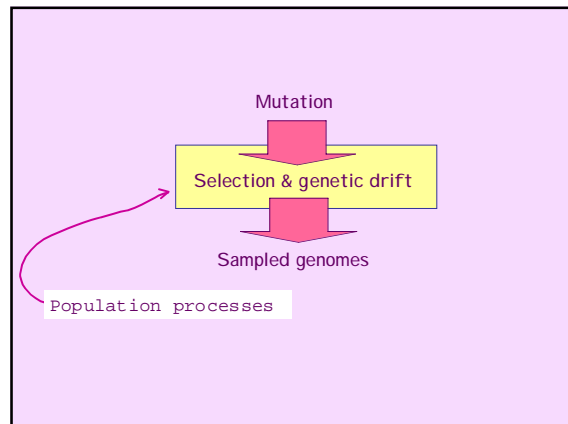
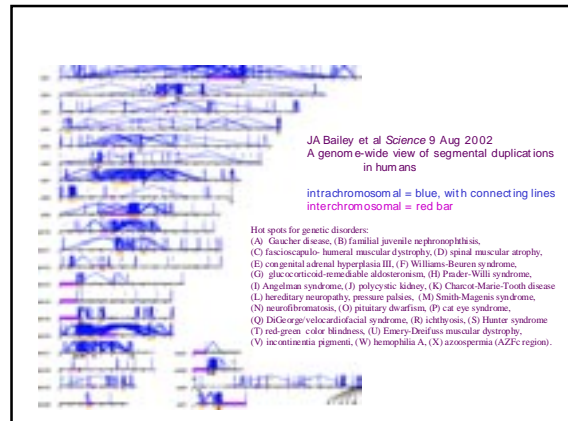
Indicates that mutation scales to absolute time rather than to generation time--ie is not mostly replication-dependent.

Estimating mutations-- contradictions

- Experimental (direct) methods for determining rate or pattern of mutation often do not agree with (indirect) estimates from sequence comparisons
 - Lab estimates can be higher or lower
- Why?
 - Sequence comparisons may not have eliminated effects of selection
 - Limits of detection in lab may be weak (based on few sites)
 - Lab environments may facilitate/prevent certain mutations
 - Eg, Mutational spectrum in bacteria can differ under anaerobic environment which is usual part of life cycle

Larger scale mutations

- Duplications of genes and larger regions, deletions of larger regions, rearrangements of genes and chromosome regions
- Genome-level data are showing that these are important (plants, animals, bacteria)
- Little studied by evolutionary biologists so far... extensive data only recently available
- Horizontal gene transfer in bacteria is the most developed area
- Less likely to be neutral so difficult to estimate rates
- "Hot spots" for gene duplications or rearrangements within eukaryotic genomes
 - Linked to presence of short sequence repeats
 - Associated with genetic diseases in humans



Once a mutation occurs, likelihood of persistence depends on population processes

- Natural selection
 - Selection between individual organisms
 - Segregation distortion (differential transmission)
- Genetic drift
 - Increase or decrease by chance sampling

Molecular view of evolution reveals variation within populations and divergence between species. Major goal is to explain:

- (1) the cause and maintenance of molecular variation in populations.
- (2) the forces producing molecular divergence between species.

Since the 1960s, two conflicting general explanations for both:

selectionist view: natural selection acting on advantageous mutations is the dominant force

neutralist view: genetic drift acting on neutral alleles is the major basis of molecular change

selectionist & neutralist views extended earlier schools of thought (arising before molecular basis of genetic variation was known):

Classical (Muller): natural selection is mostly a purifying force that removes variation from populations

Balancing (Dobzhansky): complex forms of overdominant (balancing) selection maintain variation in populations

1960's Lewontin and Hubby used allozyme electrophoresis to assess levels of genetic variation in populations of flies:

polymorphism and heterozygosity levels were surprisingly high under balancing view this created the problem of "genetic load" - differential fitness must continually result in a lot of death or sterility to permit the greater success of the more fit individuals. i.e. mean fitness must be far below the highest fitness also segregational load due to constant regeneration of homozygotes

Neutral view: another way to explain the high levels of variation

Neutral Theory: Mutations with $s=0$ account for much of the genetic divergence between lineages & most genetic polymorphism within populations

- Motoo Kimura *Nature* 1968
- King and Jukes *Science* 1969 - similar ("Non-Darwinian evolution"), emphasized between-population changes

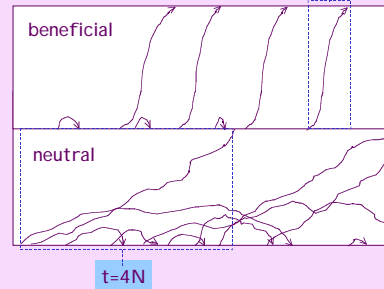


Motoo Kimura

Neutral alleles contribute more to polymorphism within populations, due to slow fixation time

- Neutral alleles have slow fixation time (t)
- There are more neutral mutations

$$t = (2/s) \ln(2N)$$



Neutral view: another way to explain within population variation: extended period of polymorphism expected for neutral variants

Other observations contributing to neutral view:

1. molecular divergence between species higher than expected if most changes are driven by selection
2. change is clock-like (rates similar across lineages) as expected if no selection and mutation

Based on early protein sequences from mammal species.

Kimura 1968 estimated 1 aa replacement per 2 years in mammals, which required an implausibly high genetic load if driven by positive selection.

Both selectionist & neutralist views allow that: positive selection for beneficial alleles is important in evolution selection against deleterious mutations is ubiquitous Positions have converged as data have become available

Main contribution of neutral theory: introduced testable null expectation into population genetics and molecular evolutionary biology

Can observed patterns arise by chance?

Tomoko Ohta 1973
 ("Nearly Neutral theory"):
 much genetic variation and
 many fixed differences
 are slightly deleterious
 rather than entirely neutral



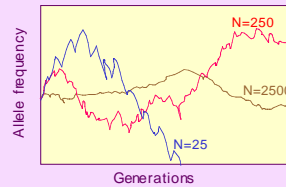
- Neutral Theory: most genetic variation and divergence detected at the molecular level is neutral
 - no effect of N_e on rate of change: larger populations give proportionally more mutations but correspondingly smaller chance that each mutation will be fixed
- Nearly Neutral Theory: much genetic variation and many fixed differences are slightly deleterious
 - Persistence and frequency are affected by N_e
 - Smaller populations will evolve faster at nearly neutral sites, because fixation is through genetic drift.

What determines the relative roles of selection and genetic drift?

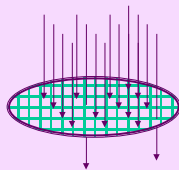
- s and N_e
- Genetic Drift is main factor for $N_e * s \ll 1$;
- Selection is important for $N_e * s \gg 1$
- Affected by linkage of mutant allele to other chromosomal regions under selection

Review of effective population size ($=N_e$)

- based on number of breeding individuals in the population over time
- strongly influenced by bottlenecks in population size--temporary reductions in numbers
 - For humans, $N_e < 20,000$, due to migration, population expansions in recent history
 - For *Drosophila melanogaster*, $N_e \sim 300,000$,
 - For *E. coli* $\sim 10^8$
- Can be estimated from polymorphism levels: more polymorphism with large N_e because alleles are not lost as quickly through sampling



New Mutations



Substitutions in lineage

"Selective Sieve"

More effective with large N and large $|s|$
 Less effective with small N or small $|s|$

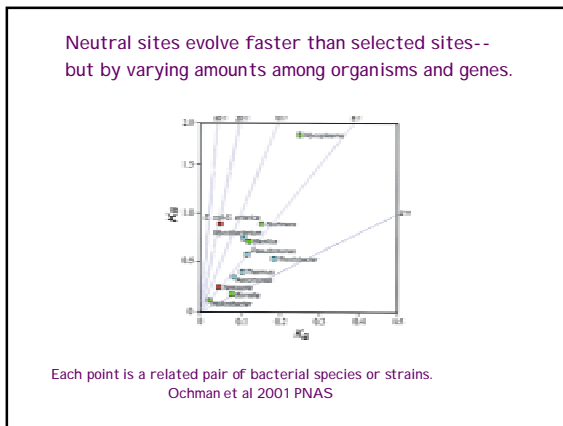
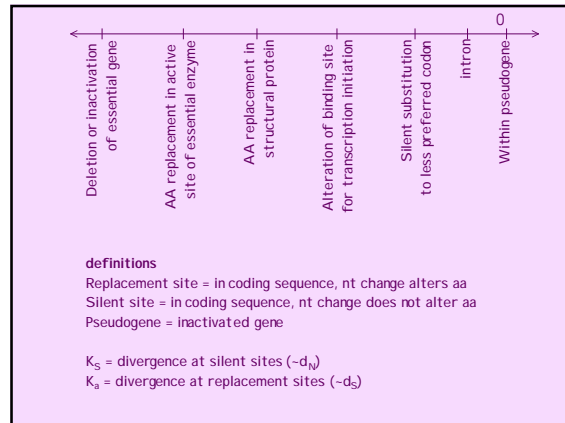
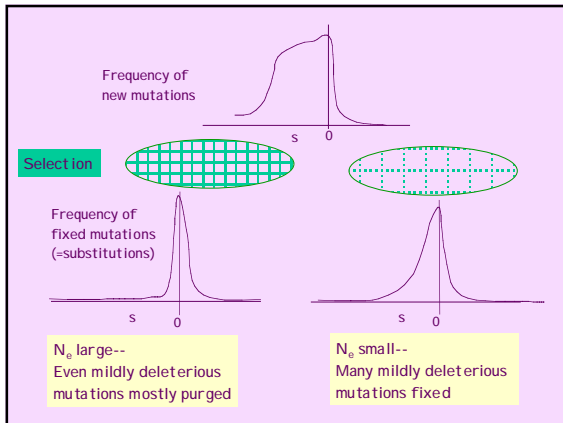
Distribution of selective coefficients of new mutations --- ?

High frequency of slightly deleterious mutations

Completely neutral mutations probably common

Beneficial mutations rare

Deleterious-----neutral-----Beneficial
 (s)



- ### What organisms are most affected by Genetic Drift?
- Many large organisms (often have fewer individuals)
 - Island populations compared to mainland (birds)
 - Microbes that live only in large organisms and do not persist outside their hosts
 - Species in which bottlenecks occur due to ecology or life cycle
 - May have large numbers of individuals sometimes, but boom-bust cycles
 - Especially relevant to human populations
 - Evidence for species-wide population bottleneck during evolution of modern humans

- ### Extent of genetic drift is reflected in sites under weak purifying selection
- "silent" sites in ORFs: Codon bias
 - Codon choice less important than amino acid choice in determining fitness
 - But-Codon choice can affect efficiency of translation
 - Highly expressed genes show strong codon bias (P. Sharp on bacteria 1980's; H. Akashi and others on Drosophila, 1990's)
 - Favored codons generally correspond to more abundant tRNAs
 - Codon bias (genome wide preference for certain codons) greater in large population organisms
 - E. coli, Bacillus, many other bacteria
 - Yeast
 - Drosophila species
 - Varies among genomes, even among Drosophila species
 - Codon preference less apparent in small populations
 - Bacteria that are symbiotic or chronically pathogenic
 - mammals

Effects of selection and genetic drift on codon use

Large N_e : selection due to translational efficiency bias affects favors certain codons in high expression genes

codon	<i>E. coli</i>		<i>Drosophila melanogaster</i>		human	
	High expr gene	Low expr gene	High expr gene	Low expr gene	Gene in G+C rich region	Gene in A+T rich region
AUU	.48	1.38	1.26	1.29	0.45	1.60
AUC	2.51	1.12	1.29	0.66	2.43	0.76
AUA	0.01	0.50	0.0	1.05	0.12	0.64

Value of 1 indicates match to random expectation

Small N_e : mutational bias affects silent sites in Isochores

Some major points about sequence evolution

Mutation

Varies among sites and regions of genomes

Varies among genomes

Can be replication-dependent, time-dependent or both

Profile of selection coefficients of new mutations critical, but hard to measure

Mutation rate determines rate of evolution at neutral sites.

Fixation

Independent of population size for neutral mutations

$Ns \gg 1$:
Selection governs outcome

$Ns \ll 1$:
Genetic Drift governs outcome

Neutral View: Genetic drift most important: $s \sim 0$ for many mutations

Nearly Neutral View: Genetic drift most important: $N^*s \ll 1$