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

Mutation
Selection & genetic drift
Sampled genomes

Mutations happen--

- Nucleotide base substitution, eg C --> T
- Insertion or deletion of one or few bases, eg ATCCGAT --> 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
  - 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. catgag -> catgta)
- Common
- Primary focus of molecular evolution & population genetics in past
  - Has been relatively easy to obtain single gene sequence
- Phenotypic effect depends on position in genome (e.g., silent/replacement, coding/ncoding, gene/non-gene, exon/intron)
- Many underlying physical mechanisms
  - Mismatches due to errors in replication
  - Failure of proofreading enzymes that act after replication
  - U 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 frameshift in protein-coding sequences
    - Replicating 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 generate a null allele
    - Associated with human genetic disease
    - E.g., T:CAT-3' expansion repeat -- poly glutamine tract, too many copies in HD locus result in Huntington's disease

Mutation: role of DNA repair

- Point mutation during replication -- E. coli case
  - Error rate for DNA synthesis (after replication & proofreading) is ~1 in 10^7
  - 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 replace it
  - 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 females because more replication events, differential effect on sex chromosomes in outcrosses
  - 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, e.g. --

- 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)
- Error repair (larger regions of damage)
- Larger scale homologous recombination (recA & recF)
DNA repair genes in some Proteobacteria

<table>
<thead>
<tr>
<th>Genome Size (Mb)</th>
<th>E. coli</th>
<th>R. prowazekii</th>
<th>A. aphidicola (Ap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mismatch repair</td>
<td>mutL</td>
<td>mutS</td>
<td>mutT</td>
</tr>
<tr>
<td>uvr repair</td>
<td>uvrA</td>
<td>uvrB</td>
<td>uvrC uvrD</td>
</tr>
<tr>
<td>Recombination pathways</td>
<td>recA</td>
<td>recB</td>
<td>recC recD</td>
</tr>
</tbody>
</table>

Example

How do particular repair enzyme functions affect the mutational spectrum?

Study using metabolic revertants:


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

Counts of mutations in 150 bp region of yeast strains with different repair gene inactivations.

Greene & Jinks-Robertson 2001

Effect of repair gene function on relative frequencies of mutations in single base runs of different lengths in yeast.

Greene & Jinks-Robertson 2001

Length of initial run

Example -- study of mutational patterns

Based on mapping of substitutions on phylogeny for 11 strains of E. coli

P. Francino et al 1996. Science

Strand bias ((G-C)/(G+C)) along fully sequenced bacterial chromosomes.

Position on chromosome

Some effect on leading and lagging strands

Arrows indicate locations of the origin of replication (when known).

From Z. Medzhit & S. Karlin. 1999. PNAS 95, 3720-3725
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
     - Compare 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
     - Compare 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 diseases (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
  - E.g., human dominant autosomal diseases such as achondroplastic dwarfism with per gene mutation rate of ~10^-9

Mutation rates estimated from specific loci in higher eukaryotes

<table>
<thead>
<tr>
<th>Organism</th>
<th>G (Gb)</th>
<th>Ge (Gb)</th>
<th>µ(b)</th>
<th>µ(g)</th>
<th>µ(eg)</th>
<th>µ(egs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. elegans</td>
<td>8.0 x 10^9</td>
<td>1.8 x 10^9</td>
<td>2.3 x 10^-10</td>
<td>0.018</td>
<td>0.004</td>
<td>0.036</td>
</tr>
<tr>
<td>Drosophila</td>
<td>1.7 x 10^9</td>
<td>1.6 x 10^9</td>
<td>3.4 x 10^-10</td>
<td>0.056</td>
<td>0.005</td>
<td>0.14</td>
</tr>
<tr>
<td>Mouse</td>
<td>2.7 x 10^9</td>
<td>8.0 x 10^9</td>
<td>1.8 x 10^-10</td>
<td>0.49</td>
<td>0.014</td>
<td>0.9</td>
</tr>
<tr>
<td>Human</td>
<td>3.2 x 10^9</td>
<td>8.0 x 10^9</td>
<td>5.0 x 10^-10</td>
<td>0.16</td>
<td>0.004</td>
<td>1.6</td>
</tr>
</tbody>
</table>

G: Genome size in bases or base pairs (haploid unless otherwise indicated)
Ge: Effective genome size (portion in which most mutations are deleterious)
µ(b): Mutation rate per base pair per replication
µ(g): Mutation rate per genome per replication
µ(eg): Mutation rate per effective genome per replication
µ(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

---

AGCTGATTG

K = 1 change/10 sites = 0.1

AGCAATTG

Reconstructing substitutions on phylogenetic tree
Estimating sequence divergence (K)

ATGCTCTAGAGTCTAGTCACTGACCGATGCAATAG
ATGCACTAGAGATCTAGCTAGCAAGATCGAGGCGATGCAATAG

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?

New Mutations

Substitutions in lineage

“Selective Sieve”

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 & Ne, 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)

Using divergence of 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

Basic for “molecular clocks” using molecular divergence to date past divergence events (More next week)

50 Million years ago,

AGCTGATTG K = 0.1
AGCAGATTG r = 0.1 subst/100My
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


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

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.

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

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
- 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

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

- Neutral alleles have slow fixation time ($t=(2/s)\ln(2N)$)
- There are more neutral mutations

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

- Neutral alleles have slow fixation time ($t=(2/s)\ln(2N)$)
- There are more neutral mutations

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 Ne on rate of change; larger populations give proportionately 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 Ne
- 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 Ne
- Genetic Drift is main factor for Ne*s << 1
- Selection is important for Ne*s >> 1
- Affected by linkage of mutant allele to other chromosomal regions under selection

Review of effective population size (Ne)
- Based on number of breeding individuals in the population over time
- Strongly influenced by bottlenecks in population size—temporary reductions in numbers
  - For humans, Ne ~ 20,000, due to migration, population expansions in recent history
  - For Drosophila melanogaster, Ne ~ 300,000
  - For E. coli ~ 10^7
- Can be estimated from polymorphism levels: more polymorphism with large Ne because alleles are not lost as quickly through sampling

Distribution of selective coefficients of new mutations
- High frequency of slightly deleterious mutations
- Completely neutral mutations probably common
- Beneficial mutations rare

Allele frequency
- More effective with large N and large |s|
- Less effective with small N or small |s|

New Mutations

Substitutions in lineage

"Selective Sieve"
Neutral sites evolve faster than selected sites--but by varying amounts among organisms and genes.

Each point is a related pair of bacterial species or strains. Ochman et al 2001 PNAS

Neutral sites evolve faster than selected sites--but by varying amounts among organisms and genes.

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 sometime, 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 (E. Sharp on bacteria 1980's & H. Akashi and others on Drosophila, 1990's)
  - Preferred codons generally correspond to more abundant tRNAs
- Codon bias (genes with preference for certain codons) greater in large population organisms
  - E. coli, Bacillus, many other bacteria
  - Yeast
  - Drosophila species
  - Human: coding, non-coding
  - Genes with strong codon bias likely to be in small population genes
  - Bacteria that are symbiotic or chronically pathogenic

Effects of selection and genetic drift on codon use

<table>
<thead>
<tr>
<th>Codon</th>
<th>E. coli</th>
<th>Drosophila melanogaster</th>
<th>Human</th>
<th>Gene type</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUU</td>
<td>1.38</td>
<td>1.28</td>
<td>1.29</td>
<td>G+C-rich</td>
</tr>
<tr>
<td>AUA</td>
<td>1.12</td>
<td>1.29</td>
<td>1.05</td>
<td>A+T-rich</td>
</tr>
<tr>
<td>AUC</td>
<td>0.01</td>
<td>0.90</td>
<td>0.66</td>
<td>A+T-rich</td>
</tr>
</tbody>
</table>

Value of 1 indicates match to random expectation

Small Ne=number of sites affected by drift
### Some major points about sequence evolution

<table>
<thead>
<tr>
<th>Mutation</th>
<th>Fixation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varies among sites and regions of genomes</td>
<td>Independent of population size for neutral mutations</td>
</tr>
<tr>
<td>Varies among genomes</td>
<td>Ns=0: Selection governs outcome</td>
</tr>
<tr>
<td>Can be replication-dependent, time-dependent or both</td>
<td>Ns=0: Genetic drift governs outcome</td>
</tr>
<tr>
<td>Profile of selection coefficients of new mutations critical, but hard to measure</td>
<td>Neutral View: Genetic drift most important: $n=0$ for many mutations</td>
</tr>
<tr>
<td>Mutation rate determines rate of evolution of neutral sites</td>
<td>Nearly Neutral View: Genetic drift most important: $N^*e=1$</td>
</tr>
</tbody>
</table>