Overview of Lecture: Genomes & Phylogenies
Read: Text pgs 534-535, ch 21 & 26

Bullet Points:
- geology, fossils & the history of life on earth
- Linnean systematics vs phylogenetic cladistics
- five kingdoms $\rightarrow$ three domains (or more?) $\rightarrow$ tree of life
- homologous vs analogous (convergent) traits
- orthologous vs paralogous DNA sequences
- retrotransposons: strange paralogs
- the ENCODE project: “junk” is in control!
- copy number variants and human variation
- out groups & shared derived characters (“synapomorphies”)
- mono- para- & polyphyletic groups
  - (goodbye “birds,” watch out “dogs” !)
History of life on Earth

The Earth is a little over 4.5 billion years old, its oldest materials being 4.3 billion-year-old zircon crystals. Its earliest times were geologically violent, and it suffered constant bombardment from meteorites. When this ended, the Earth cooled and its surface solidified to a crust - the first solid rocks. There were no continents as yet, just a global ocean peppered with small islands. Erosion, sedimentation and volcanic activity - possibly assisted by more meteor impacts - eventually created small proto-continentals which grew until they reached roughly their current size 2.5 billion years ago. The continents have since repeatedly collided and been torn apart, so maps of Earth in the distant past are quite different to today's.

The history of life on Earth began about 3.8 billion years ago, initially with single-celled prokaryotic cells, such as bacteria. Multicellular life evolved over a billion years later and it's only in the last 570 million years that the kind of life forms we are familiar with began to evolve, starting with arthropods, followed by fish 530 million years ago (Ma), land plants 475Ma and forests 385Ma. Mammals didn't evolve until 200Ma and our own species, Homo sapiens, only 200,000 years ago. So humans have been around for a mere 0.004% of the Earth's history.

View the Tree of Life

Jump to:
Geological timeline | Geological time periods | Big Five mass extinction events | Mass extinction theories | Ancient Earth habitats
Table 25.1 The Geologic Record

<table>
<thead>
<tr>
<th>Era</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paleozoic</td>
<td></td>
</tr>
<tr>
<td>Permian</td>
<td>Radiation of reptiles; origin of most present-day groups of insects; extinction of many marine and terrestrial organisms at end of period</td>
</tr>
<tr>
<td>Carboniferous</td>
<td>Extensive forests of vascular plants form; first seed plants appear; origin of reptiles; amphibians dominant</td>
</tr>
<tr>
<td>Devonian</td>
<td>Diversification of bony fishes; first tetrapods and insects appear</td>
</tr>
<tr>
<td>Silurian</td>
<td>Diversification of early vascular plants</td>
</tr>
<tr>
<td>Cenozoic</td>
<td></td>
</tr>
<tr>
<td>Eocene</td>
<td>Angiosperm dominance increases; continued radiation of most present-day mammalian orders</td>
</tr>
<tr>
<td>Oligocene</td>
<td>Origins of many primate groups, including apes</td>
</tr>
<tr>
<td>Miocene</td>
<td>Continued radiation of mammals and angiosperms; apelike ancestors of humans appear</td>
</tr>
<tr>
<td>Pliocene</td>
<td>Origin of genus Homo</td>
</tr>
<tr>
<td>Pleistocene</td>
<td>Ice ages; humans appear</td>
</tr>
<tr>
<td>Holocene</td>
<td>Historical time</td>
</tr>
<tr>
<td>Paleogene</td>
<td></td>
</tr>
<tr>
<td>Ediacaran</td>
<td>Diverse algae and soft-bodied invertebrate animals appear</td>
</tr>
<tr>
<td></td>
<td>635</td>
</tr>
<tr>
<td></td>
<td>Oldest fossils of eukaryotic cells appear</td>
</tr>
<tr>
<td></td>
<td>2,100</td>
</tr>
<tr>
<td></td>
<td>2,500</td>
</tr>
<tr>
<td></td>
<td>2,700</td>
</tr>
<tr>
<td></td>
<td>Concentration of atmospheric oxygen begins to increase</td>
</tr>
<tr>
<td></td>
<td>3,500</td>
</tr>
<tr>
<td></td>
<td>Oldest fossils of cells (prokaryotes) appear</td>
</tr>
<tr>
<td></td>
<td>3,800</td>
</tr>
<tr>
<td></td>
<td>Oldest known rocks on Earth’s surface</td>
</tr>
<tr>
<td>Approx. 4,600</td>
<td>Origin of Earth</td>
</tr>
</tbody>
</table>

“Big Bang” ~ 14bya
Linnean systematics is an hierarchical classification scheme. (Domain, Kingdom, Phylum, Class, Order, Family, Genus, Species) It preceded evolutionary theory and is based on phenetics - observed phenotypic similarities {ex: ‘morphospecies’}.

Darwin wrote in The Origin of Species that "our classifications will come to be, as far as they can be so made, genealogies."

We are still in the midst of the an effort to reconcile the existing taxonomy w/ modern phylogenetic cladistics: an attempt to develop classification (taxonomy) consistent with evolutionary history - decent w/ modification.
... phylogenetic trees are hypotheses that best fit the available data. New data often begets new hypotheses.
"The affinities of all the beings of the same class have sometimes been represented by a great tree...
Charles Darwin, 1859
News Feature

The challenge of microbial diversity: Out on a limb

From giant viruses to unexplained marine DNA, there may be more to the tree of life than the three domains currently defined.

*Nucleocytoplasmic large DNA viruses
It is probable that these viruses evolved before the separation of eukaryotes into the extant crown groups.

Convergence from different ancestors to similar descendants confuses inferred ancestry …
In addition to fossil organisms, phylogenetic history can be inferred from certain morphological and molecular similarities among living organisms. Similarities due to shared ancestry are called homologies. The similarity in the number & arrangement of bones in the forelimbs of mammals is due to their descent from a common ancestor with the same bone structure; genes or other DNA sequences are homologous if the nature of their similarity suggests that they are descended from the sequences carried by a common ancestor.

Figure 26.8 Aligning segments of DNA. Systematists use computer software to find and realign similar sequences along DNA segments from two species.
A potential red herring in constructing a phylogeny is similarity due to convergent evolution - called analogy - rather than to shared ancestry (homology). **Convergent evolution** occurs when similar environmental pressures and natural selection produce similar (analogous) adaptations in organisms from different evolutionary lineages.

**Consider 'wings' vs 'forelimbs' of bats & birds.**

Bird and bat wings are analogous - they have separate evolutionary origins, but are superficially similar because they evolved to serve the same function. ... though bird and bat wings are analogous as wings, as **forelimbs they are homologous**. Birds & bats did not inherit wings from a common ancestor, but they did inherit forelimbs from a common ancestor with forelimbs. 

{ are the bones in the forelimbs of birds & bats homologous? }

Convergent evolution in ‘mechanical design of lamnid sharks and tunas.’

*Donley, JM. 2004 Nature 429,61-65*
Blonde hair evolved more than once

Golden locks of dark-skinned Melanesians have different genetic basis to those of Europeans.

About 5–10% of people from Melanesia have naturally blonde hair - the highest prevalence outside Europe.

Yet people from the region have the darkest skin outside Africa.

Now, a study of people from the Solomon Islands shows that they evolved the striking blonde trait independently of people in Europe.

This refutes the possibility that blonde hair was introduced by colonial Europeans … "Blonde hair has clearly evolved twice," says Carlos Bustamante, a co-author.

To investigate the genetic basis underlying the trait, …

compared the genomes of 43 blonde and 42 dark-haired Solomon Islanders, and revealed that the islanders' blonde hair was strongly associated with a single mutation in the TYRP1 gene.

That gene encodes an enzyme that influences pigmentation in mice and humans.

Several genes are known to contribute to blonde hair coloration in Europeans, but TYRP1 is not involved.

By comparing DNA between >900 Solomon Islanders and >900 other individuals from 52 populations around the world, the researchers found that the TYRP1 mutation is probably unique to the Oceanic region …

It is unusual that one particular mutation would explain so much of a population's observable trait, says Bustamante.

{ So, the trait “blond” in Europeans and Melanesians is … what? }
Homologous DNA sequences are **orthologous** if they were separated by a speciation event.

99% of the genes of humans & mice orthologous, 50% of our genes are orthologous with those of yeast.

**Homologous DNA sequences** are **paralogous** if they were separated by a gene duplication event {possibly carried by a retrotransposon} so they are found in more than one copy in the same genome.

The genes encoding myoglobin & hemoglobin are ancient paralogs. {many human hormones come in “families” that are coded by paralogous genes} Humans have huge families of more than 1,000 paralogous olfactory receptor genes. {many are degenerate ‘pseudogenes’}

**Homologous sequences.** Orthologs and Paralogs are two types of homologous sequences. Orthology describes genes in different species that derive from a common ancestor. Orthologous genes may or may not have the same function. Paralogy describes homologous genes within a single species that diverged by gene duplication. http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html
Bitter taste perception is thought to act as a ‘toxin detector’ that warns of the presence of noxious compounds in foods, especially plant toxins such as alkaloids …

Bitter taste responses in humans are mediated by **bitter taste receptors**, a series of ~25 G protein-coupled receptors **encoded by the TAS2R gene family**… allelic variation predicts both the perceived bitterness of saccharin and acesulfame K and functional response in cell-based assays …

Threshold perception concentrations … ranged 64-fold …

Whole-gene sequences obtained from TAS2R30–46 revealed that our {60} subjects collectively harbored 42 single-nucleotide polymorphisms (SNPs) … the frequency of non-synonymous SNPs in our sample, 86%, is high compared with other human genes {“non-synonymous SNPs” means …? }

**Copy-number assays** … targeting TAS2R43 revealed 16 subjects with a copy number of zero, 28 with a copy number of one and 16 with a copy number of two … At TAS2R45, the number of subjects with copy numbers of zero, one and two were 2, 11 and 30 … **Copy-number variation … has the potential to have extreme effects on phenotypes.** {possibly related to autism and other diseases}

… allelic {copy number} variation predicts both
Evolving promiscuously

Of the various types of gene rearrangements that have been found in living organisms, gene duplication-amplification (GDA), a process that alters gene dosage, appears to be especially common and biologically important. GDA is significant from a fundamental evolutionary perspective because it serves as a primary source for genetic innovation (i.e., evolution of new genes) {paralogs diverge into gene families with related but different functions} and also plays an important role in the generation of genomic variability.

Furthermore, recent discoveries in medical genetics show that gene copy number variation in the human genome is an important contributor to many human diseases, phenotypic variability among individuals, and human susceptibility to infectious diseases.

Copy Number Variants: A New Molecular Frontier in Clinical Psychiatry

It is now clear beyond any reasonable doubt that genetic inheritance influences liability to develop almost every major psychiatric disorder. … our growing knowledge regarding the roles of copy number variants in behavioral disorders will soon require revision of standards of evaluation and care for psychiatric patients.

Functional impact of global rare copy number variation in autism spectrum disorders
Pinto et al. 2010. Nature 466, 368–372
Retrotransposons are a strange type of Paralogous “genes”.

L1 Retrotransposons Shape the Mammalian Genome
Haig H. Kazazian Jr. 2000 Science 289, 1152 - 1153
Genomic mobile elements called retrotransposons make up about 40% of the mammalian genome.
... these small pieces of DNA are transcribed into RNA, reverse-transcribed into DNA, and then inserted back into the genome at a new site. {by reverse transcriptase coded by L1, see C&R pg 436}

Which transposable elements are active in the human genome?
~ 35-40 subfamilies of Alu, L1 ... and possibly HERV-K remain actively mobile in the human genome.
... they continue to produce genetic diversity and cause human diseases by integrating into genes.

Estimating the retrotransposition rate of human Alu elements
Cordaux et al. GENE 373: 134-137 MAY 24 2006
We estimate the retrotransposition rate (RR) of Alu elements in humans on the order of one new insertion every ~20 births ...
Retroelements constitute a large portion of our genomes (~45%!!!). One class of these elements, the human endogenous retroviruses (HERVs), is comprised of remnants of ancient exogenous retroviruses that have gained access to the germ line. {inserted in sperm or egg DNA}

... HERV-derived transcripts and proteins have been detected in healthy and diseased human tissues, and HERV-K, the youngest, most conserved family, is able to form virus-like particles. {but has cumulated mutations that prevent release of infectious virus & horizontal transmission by infection}

Although it is generally accepted that the integration of retroelements can cause significant harm by disrupting or disregulating essential genes, the role of HERV expression in ... diseases remains controversial. {MS?}

... HERV proteins might even serve the needs of the host and are therefore under positive selection.

HERV-K: an orthologous “gene” in chimpanzees, bonobos & gorillas

A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans
Evidence from DNA sequencing studies strongly indicated that humans & chimpanzees are more closely related to each other than either is to gorillas.
However, precise details of the lineage leading to humans from those leading to the African great apes have remained uncertain.
The unique insertion sites of endogenous retroviruses, mutations like those of other transposable genetic elements, should be useful for resolving phylogenetic relationships among closely related species.
We identified a human endogenous retrovirus K (HERV-K) provirus that is present at the orthologous position in the gorilla and chimpanzee genomes, but not in the human genome.
Humans contain an intact preintegration site at this locus.
These observations provide very strong evidence that, ... chimpanzees, bonobos, and gorillas are more closely related to each other than they are to humans.
... demonstrate the utility of using the location of HERV-K to trace human evolution.
large, genome-wide studies searching for genetic underpinnings for diseases, such as lung cancer or autism, have pointed to the nether regions of the genome between the protein-producing genes - areas that were often thought to contain “junk” DNA that was not part of the pantheon of known genes.

An international consortium of hundreds of scientists {the ENCODE project} has now deciphered a large portion of the strange language of this junk DNA and found it to be not junk at all. Rather it contains important signals for regulating our genes, determining disease risk, height and many of the other complex aspects of human biology that make each one of us different.

... the group found about 4 million of these so-called switches and can now assign {regulatory} functions to >80% of the entire genome. Compare that to the roughly 2 percent of the genome that is responsible for the protein-coding genes that researchers have been relying on to look for diseases and traits.

These newly catalogued switches not only activate and de-activate genes, but also control how much of each protein gets made and when.

{ note: if ~40% retrotransposons, ~10% HERVs and >80% regulatory, then … }
After systematists have separated homologous from analogous similarities, they must **sort through the homologies to distinguish between shared primitive and shared derived characters.**

For example, **all mammals share the homologous character of a backbone.** However, the presence of a backbone does not distinguish mammals from other vertebrates because nonmammalian vertebrates such as fishes and reptiles also have backbones.

The **backbone** (*vertebral column*) is a homologous structure that predates the branching of the mammalian clade from the other vertebrates; it is a **shared primitive character**, shared beyond the taxon we are trying to define.

In contrast, **hair**, a character shared by all **mammals** but not found in nonmammalian vertebrates, is a **shared derived character**, **{jargon alert: synapomorphy!!!}** an evolutionary novelty unique to a particular clade – in this case, the mammalian clade.

Systematists use **outgroup comparison** to differentiate between **shared derived characters** and **shared primitive characters**.

(b) **Phylogenetic tree.** Analyzing the distribution of these derived characters can provide insight into vertebrate phylogeny.
Patterns of shared characteristics can be depicted in a diagram called a **cladogram**. If the shared characteristics are due to common ancestry (homologous), then the cladogram forms the basis of a phylogenetic tree. Within the tree, a **clade** (from the Greek klados, branch) is defined as a **group of species that includes an ancestral species and all its descendants**.

The cat family represents a clade within a larger clade that also includes the dog family.

But not all groupings of organisms qualify as clades. (a) A valid clade is **monophyletic** (meaning “single tribe”), signifying that it consists of the ancestor species and all its descendants.  

(b) A **paraphyletic** grouping consists of an ancestral species and some, but not all, of the descendants. 

(c) A **polyphyletic** grouping is several species that lack a common ancestor.

(a) **Monophyletic group (clade)**. Group I, consisting of three species (A, B, C) and their common ancestor 1, is a clade, also called a monophyletic group. A monophyletic group consists of an ancestral species and all of its descendants. 

(b) **Paraphyletic group**. Group II is paraphyletic, meaning that it consists of an ancestral species 2 and some of its descendants (species D, E, F) but not all of them (missing species G). 

(c) **Polyphyletic group**. Group III is polyphyletic, meaning that its members have different ancestors. In this case, species B and C share common ancestor 3, but species C has a different ancestor 4.
The group Reptiles does not form a clade, so this is not a valid phylogenetic name. However, if birds are included, Reptilia is a valid phylogenetic name.

http://encyclopedia.laborlawtalk.com/Reptile

In recent years many taxonomists have begun to insist that taxa should be monophyletic, The reptiles would be paraphyletic, since they exclude birds.

A clade is a monophyletic group: most recent common ancestor and all its descendants.

Paraphyletic group: not all descendants of most recent common ancestor.

Polyphyletic group: does not include of most recent common ancestor & all descendants.
Lizard Genome Unveiled
First non-avian reptile sequence helps explain vertebrate evolution
By Lee Sweetlove and Nature magazine | August 31, 2011 | 1

Publication of the genome of the North American green anole lizard has filled a yawning genome-sequence gap in the animal lineage. The paper, which appears today in *Nature*, is the first to sequence the genome of a non-avian reptile. "This fills out a clade that has been completely ignored before," says lead author Jessica Alföldi of the Broad Institute of MIT and Harvard in Cambridge, Massachusetts.
The archaeological record cannot resolve whether domestic dogs originated from a single wolf population or arose from multiple populations at different times.

Mitochondrial DNA sequences were analyzed from 162 wolves at 27 localities worldwide (W#) and from 140 domestic dogs representing 67 breeds (D#).

Wolves (W) were the ancestors of dogs (D).

Clade IV: three haplotypes that were identical or very similar to a wolf haplotype found western Russia, which suggests recent hybridization.

Clade III: a variety of breeds such as the German shepherd, Siberian husky, and Mexican hairless.

Clade II included 2 Scandinavian breeds & 2 wolf haplotypes.

Clade I: many common and ancient breeds such as the dingo, New Guinea singing dog, African basenji, and greyhound.

The coyote {out group} and wolf ... diverged about one million ybp, as estimated from the fossil record.
As available data about DNA sequences increase, the difficulty of building the phylogenetic tree that best describes evolutionary history also grows. What if you are analyzing data for 50 species? There are $3 \times 10^{76}$ different ways to arrange 50 species into a tree! Systematists can never be sure of finding ‘the true phylogeny,’ but they can narrow the possibilities by applying the principles of **maximum parsimony** and **maximum likelihood**. {mathematical methods to search for the simplest plausible phylogeny}

Here are 265 of the phylogeny packages, and 31 free servers, that I know about. This is an attempt to be completely comprehensive. I have not made any attempt to exclude programs that do not meet some standard of quality or