

- Where does genetic variation come from?
- How can we measure genetic variation?
- How can genetic variation help us ?

Genetic variation I

- **Among individuals within a population**
- Genetic variation among individuals within a population can be identified at a variety of levels. It is possible to identify genetic variation from observations of phenotypic variation in either quantitative traits (traits that vary continuously and are coded for by many genes, e.g., leg length in dogs) or discrete traits (traits that fall into discrete categories and are coded for by one or a few genes (e.g., white, pink, red petal color in certain flowers)).
- Genetic variation can also be identified by examining variation at the level of enzymes using the process of protein electrophoresis. Polymorphic genes have more than one allele at each locus. Half of the genes that code for enzymes in insects and plants may be polymorphic, whereas polymorphisms are less common in vertebrates.
-

Genetic variation II

- Ultimately, genetic variation is caused by variation in the order of bases in the nucleotides in genes. New technology now allows scientists to directly sequence DNA which has identified even more genetic variation than was previously detected by protein electrophoresis. Examination of DNA has shown genetic variation in both coding regions and in the non-coding intron region of genes.
- Genetic variation will result in phenotypic variation if variation in the order of nucleotides in the DNA sequence results in a difference in the order of amino acids in proteins coded by that DNA sequence, and if the resultant differences in amino acid sequence influence the shape, and thus the function of the enzyme.
- **Between populations**
- Geographic variation in genes often occurs among populations living in different locations. Geographic variation may be due to differences in selective pressures or to genetic drift.

Wednesday, September 22, 2010

3

DNA sequence change

- DNA sequences change over time
- substitutions (errors in duplication, repair, environmental damage)
- insertions and deletions (indels)
- multiple substitutions can lead to original base appearing again.
- A -> C -> G -> A
- Interested in rate of change in various portions of genome/genes to learn about mechanism(s) that causes the change.
-

Wednesday, September 22, 2010

4

Genetic variation III

- **Measurement**
- Genetic variation within a population is commonly measured as the percentage of gene loci that are polymorphic or the percentage of gene loci in individuals that are heterozygous.
- **Sources**
- Mutations are the ultimate source of genetic variation because they alter the order of bases in the nucleotides of DNA. Mutations are likely to be rare and most mutations are neutral or deleterious, but in some instances the new alleles can be favored by natural selection.
- Genetic variation can also be produced by the recombination of chromosomes that occurs during sexual reproduction, called independent assortment.
- The crossing over that occurs during meiosis can result in the production of new alleles or new combinations of alleles.

Wednesday, September 22, 2010

5

Genetic variation IV

- **Maintenance in populations**
- A variety of factors maintain genetic variation in populations. Potentially harmful recessive alleles can be hidden from selection in the heterozygous individuals in populations of diploid organisms (recessive alleles are only expressed in the less common homozygous individuals). Natural selection can also maintain genetic variation in balanced polymorphisms. Balanced polymorphisms may occur when heterozygotes are favored or when selection is frequency dependent.
- **Sexual Reproduction**
- Sexually-reproducing, outbreeding population, all individuals are different (except identical twins)
- **Neutral Theory**
- Motoo Kimura's neutral theory of molecular evolution, the majority of mutations in the human genome are neutral,

Wednesday, September 22, 2010

6

The nature of molecular evolutionary changes

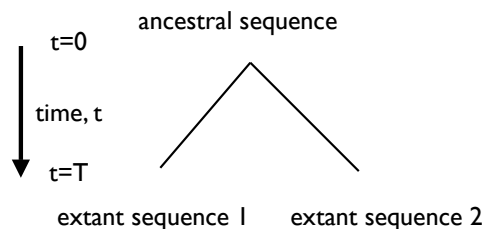
- Neutral divergence/drift compared to selective changes
- Different rates at different sites, gamma distribution describes different rate distributions
- K is the number of substitutions per nucleotide site between two sequences, a very important metric for measuring molecular genetic change
- changes in nucleotides can come in several “flavours”, depending on whether the nucleotide that changed codes for an amino acid (coding) or doesn’t (non-coding)
- To assess how sequences are changing relative to each other, we first need an alignment. Aligned sequences are made up of columns of residues (nucleotides or amino acids). We are making an explicit statement that there is an evolutionary relationship between residues in each column and that these residues are related by descent (homologous).
- Now we can predict the ancestral residue and also look at the nature of any change(s) in a column. We can have a match, mismatch (substitution) or a gap [caused by an indel (insertion or deletion)]. We cannot tell the direction of any of these changes from just two sequences.
- For this we need a multiple (>2) sequence alignment and phylogenetic analysis.

Wednesday, September 22, 2010

7

Rates and patterns of nucleotide substitutions

- If we know the rate of substitution, we can date evolutionary events
- First we need to know whether rates are comparable between groups of species e.g. between rodents and primates. (Rodents are fast-evolving).
- rate of nucleotide substitution, $r = K/2T$



- Different regions of the sequence are subject to different selective pressures and evolve at different rates. These regions are treated independently

Wednesday, September 22, 2010

8

Measurements

- θ_w (Waterson)
- θ_T (Tajima)

$$\theta = \frac{K}{\sum_{i=1}^{n-1} \frac{1}{i}} = 4 N_e \mu$$

where K is the number of segregating sites

Wednesday, September 22, 2010

9

π and θ

- nucleotide diversity, π (Nei 1987, equations 10.5 or 10.6),
- nucleotide polymorphism, $\theta = 4N_e\mu$, where N_e is the effective population size, and μ is the mutation rate per nucleotide (or per sequence) and per generation (Nei 1987, equation 10.3; Tajima 1993, equation 3) and its variance for free and for no recombination (Tajima 1993, equations 4 and 8);

$$\theta = \frac{K}{\sum_{i=1}^{n-1} \frac{1}{i}} = 4 N_e \mu$$

where K is the number of segregating sites

- θ per nucleotide under the finite sites model (Tajima 1996, equations 9-10, 16).
- *** See also Nei & Kumar, 2000

Wednesday, September 22, 2010

10

Types of nucleotide change and their effects on SNPs.

- Synonymous changes encode the same amino acid. Synonymous changes can be 2-fold or 4-fold degenerate
- Non-synonymous changes the amino acid

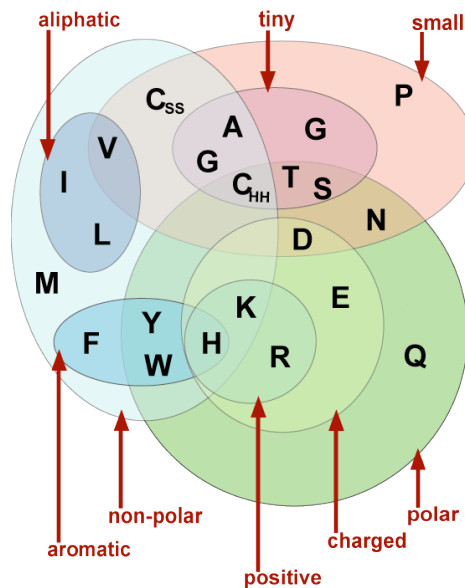
		Second base				
		U	C	A	G	
First base	U	UUU } Phenyl-alanine F UUC } UUA } Leucine L UUG }	UCU } Serine S UCC } UCA } UCG }	UAU } Tyrosine Y UAC } UAA } Stop codon UAG } Stop codon	UGU } Cysteine C UGC } UGA } Stop codon UGG } Tryptophan W	U C A G
	C	CUU } Leucine L CUC } CUA } CUG }	CCU } Proline P CCC } CCA } CCG }	CAU } Histidine H CAC } CAA } Glutamine Q CAG }	CGU } Arginine R CGC } CGA } CGG }	U C A G
	A	AUU } Isoleucine I AUC } AUA } AUG } Methionine start codon M	ACU } Threonine T ACC } ACA } ACG }	AAU } Asparagine N AAC } AAA } Lysine K AAG }	AGU } Serine S AGC } AGA } Arginine R AGG }	U C A G
	G	GUU } Valine V GUC } GUA } GUG }	GCU } Alanine A GCC } GCA } GCG }	GAU } Aspartic acid D GAC } GAA } Glutamic acid E GAG }	GGU } Glycine G GGC } GGA } GGG }	U C A G

Wednesday, September 22, 2010

11

Types of nucleotide change and their effects on SNPs.

- Amino acid changes can make very little difference to the function of the protein (e.g. Lysine (K) to Arginine (R)) or very large differences (Alanine (A) to Phenylalanine (F) or Glutamate (E)).



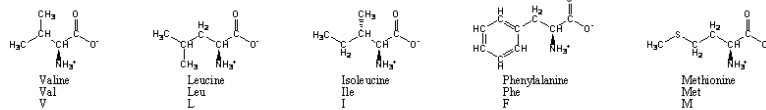
Wednesday, September 22, 2010

12

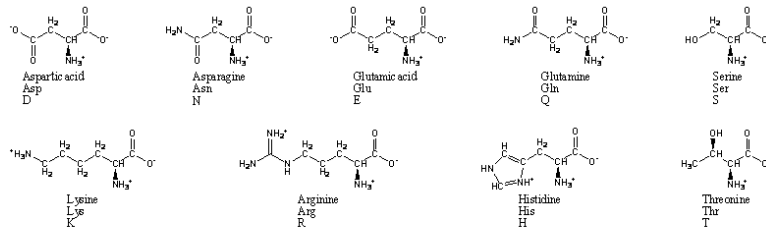
Types of nucleotide change and their effects on SNPs.

- Changes between amino acids with similar chemical groups often do not affect protein function very much.
- The greater the functional change, the rarer it will be fixed if it is deleterious, and the more interesting the SNP that encodes it.

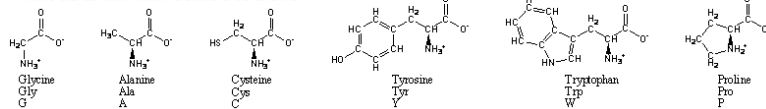
Amino acids with hydrophobic side chains



Amino acids with hydrophilic side chains



Amino acids with intermediate side chains



Wednesday, September 22, 2010

13

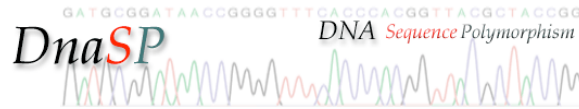
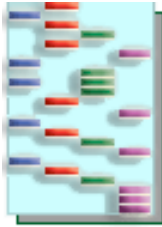
Software for calculating gene frequencies and distances

- Site maintained by J. Felsenstein at <http://evolution.genetics.washington.edu/phylip/software.html>
- Some examples
- DNASP (Windows)
- Phylip (Mac and Windows)
- Popgene
- ycdma
- fstat
- Genepop
- TreeFit

Wednesday, September 22, 2010

14

Software: DNA Sequence Polymorphism DNASP

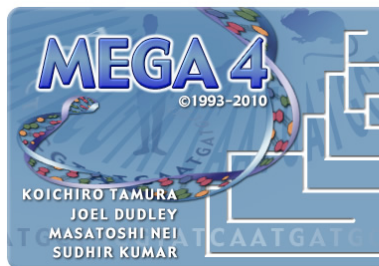


- DnaSP, DNA Sequence Polymorphism, is a software package for the analysis of nucleotide polymorphism from aligned DNA sequence data.
- DnaSP can also carry out several tests of neutrality: Hudson, Kreitman and Aguadé (1987), Tajima (1989), McDonald and Kreitman (1991), Fu and Li (1993), and Fu (1997) tests.
- Additionally, DnaSP can estimate the confidence intervals of some test-statistics by the coalescent. The results of the analyses are displayed on tabular and graphic form.

Wednesday, September 22, 2010

15

Software: MEGA Molecular Evolutionary Genetics Analysis Software



- Molecular Evolutionary Genetics Analysis
- Windows (v4), Mac, (v5. beta) Linux (v4)
- MEGA is an integrated tool for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses.
- Discussed in Phylogenetic Trees Made Easy
- Great tutorial, simple to learn

Wednesday, September 22, 2010

16

Software: Multiple Sequence Alignment

MUSCLE

- muscle is obtained from <http://www.drive5.com/muscle/>



MUSCLE web server
Provided by EBI

<http://www.ebi.ac.uk/Tools/muscle/index.html>

EMBL-EBI Search All Databases Enter Text Here Go Reset Advanced Search Give us feedback

Databases Tools EBI Groups Training Industry About Us Help Site Index

EBI > Tools > Sequence Analysis

MUSCLE

MUSCLE stands for Multiple Sequence Comparison by Log-Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

[Download Software](#)

MUSCLE
multiple sequence alignment with reduced time and space complexity

RESULTS	SEARCH TITLE	YOUR EMAIL
interactive	Sequence	
OUTPUT FORMAT	OUTPUT TREE	OUTPUT ORDER
FASTA	none	aligned

Enter or Paste a set of Sequences in any supported format:

Wednesday, September 22, 2010

17

SNPs and genes

- Location of SNPs in genes is important
- Inferring biological significance of SNPs helps understand function
- Finding SNPs in small datasets (100 sequences)
 - online tools
- Finding SNPs in large datasets (>1,000,000 sequences)
 - large server, cloud computing
- attempt to find a correlation between a SNP present in a gene (allele) and variation in phenotype (function) conferred by that SNP
 - drought, disease resistance, growth characteristics, nutritional qualities of crop
 - accelerate breeding programs
 - sequence to function and related sequences via the genome sequence

Wednesday, September 22, 2010

18