- Project presentations and papers due:
  Thursday, MARCH 15, 2007,
  1030-1220

Adaptation in the Human Genome

A genome-wide scan for signatures of adaptive evolution using SNP data

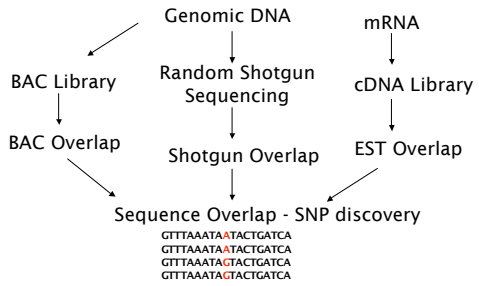
Joanna Kelley
GE414
20 Feb 07

Single Nucleotide Polymorphism (SNP)

A nucleotide difference at a given location in the genome

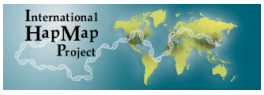
GTAAGCCTAC
GTACGCCTAC

Discovering SNPs in the Human Genome



~ 10 Million SNPs Available http://www.ncbi.nlm.gov/SNP/

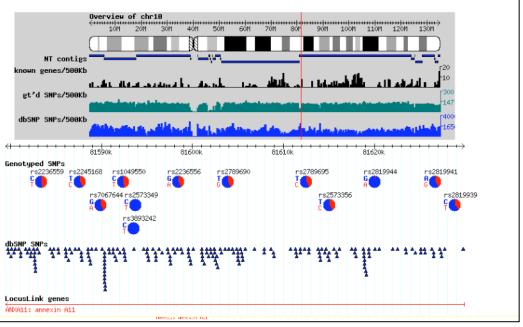
Debbie Nickerso



HapMap

- Genetic resource for developing association maps
  - Compare genotype patterns between individuals and populations
- Populations
  - Nigerian
  - Japanese
  - Chinese
  - European (Individuals from Utah - CEPH)
- Total number of genotyped SNPs released (Jan 07): 3,904,218
- Approximately 1 SNP per 1000 base pairs

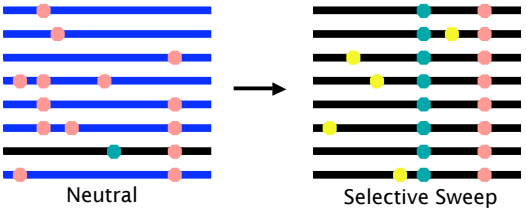
Screen Capture of HapMap Gene: annexin A11



The HapMap is a Resource for Population Genetic Studies

- SNP data can be used to identify natural selection
- Genome-wide scan for regions of adaptive evolution
  - Selective sweeps
  - Balancing selection

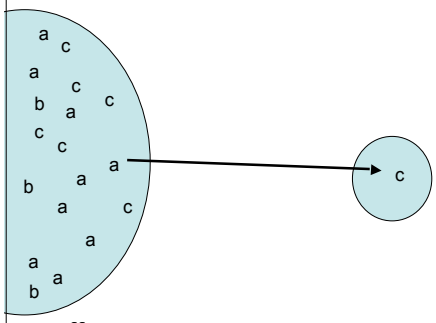
Studying natural selection



Look for an excess of rare alleles using Tajima's D

- Nucleotide polymorphism, θ
- Nucleotide diversity, π

Population demographics



Affects entire genome, not just one locus

Nucleotide Polymorphism, θ

number of polymorphic nucleotides, normalized for sample size

$$\theta = S / (\sum_{i=1}^{n-1}(1/i))$$

```
tcagaaccatgctgccattcatgcatctgctcaaatcaatcacttctgcaatgccatcat
tcagaaccatgctgcgattcatgcatcagctcaaatcaatcacttctgcaatgccatcat
tcagaaccatgctgcgattcatgcatcagctcaaatcaatcacttcttcaatgccatcat
tcagaaccatgctgcgattcatgcatctgctcaaatcaatcacttctgcaatgccatcat
```

3 Segregating (polymorphic) Sites

n = 4 (number of sequences)
$\theta = 0.027$

Note: θ does not depend on nucleotide frequency

Nucleotide diversity, π

proportion of nucleotides that differ between two random sequences in a sample

$$\pi = (^{n}/_{(n-1)}) \sum_{ij} x_i x_j p_{ij}$$

x_i, x_j - frequencies of i^{th} and j^{th} sequences

```
tcagaaccatgctgccattcatgcatctgctcaaatcaatcacttctgcaatgccatcat
tcagaaccatgctgcgattcatgcatcagctcaaatcaatcacttctgcaatgccatcat
tcagaaccatgctgcgattcatgcatcagctcaaatcaatcacttcttcaatgccatcat
tcagaaccatgctgcgattcatgcatctgctcaaatcaatcacttctgcaatgccatcat
```

p_{ij} - proportion of different nucleotides between two sequences

	Seq1	Seq2	Seq3	Seq4
Seq1	0	2	3	1
Seq2	2	0	1	1
Seq3	3	1	0	2
Seq4	1	1	2	0

n = 4 (number of sequences)
$\pi = 0.0267$

Note: π depends on nucleotide frequencies

Tajima's D: Relation between θ and π

- Under neutrality, expect polymorphism and diversity to be equal ($\theta = \pi$)
- Tajima's D is a test statistic that tests for deviations from this expectation
  - Tajima's D > 0, excess of high frequency (common) alleles
    - Heterozygote advantage, balancing selection
    - Population bottleneck
    - Population structure
  - Tajima's D < 0, excess of low frequency (rare) alleles
    - Directional selection
    - Population expansion
    - Population admixture

Methods

Download SNP information from SNP database (dbSNP)

↓

Create a list of SNPs that map to gene regions

↓

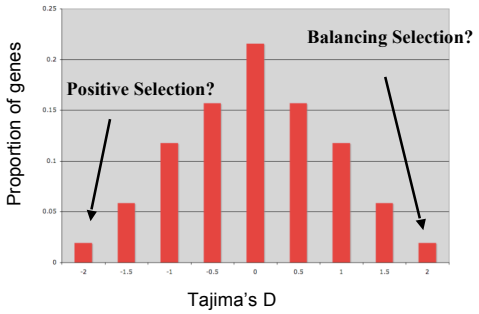
Download genotype data from HapMap

↓

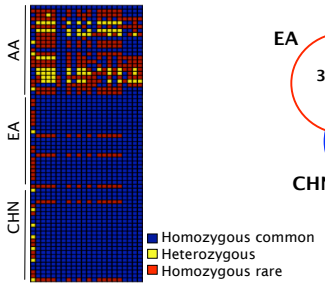
Analyze the data

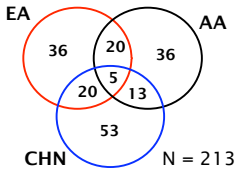
Calculate #SNPs (sites) per gene, Tajima's D, etc.

Distribution of Genes by Tajima's D



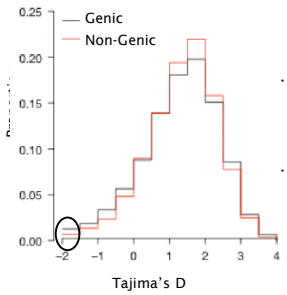
Genes show dramatic differences in allele frequencies between populations





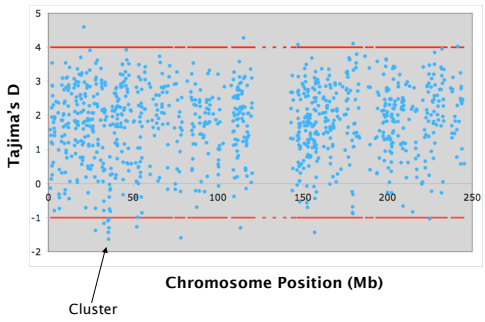
What do you expect Tajima's D to look like in non-gene regions?

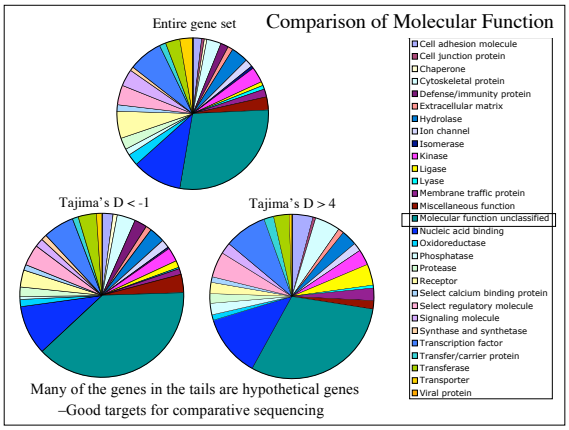
Distribution of genes by Tajima's D



- Gene regions have similar demographic history and ascertainment as non-genic regions
- Comparison of genic and simulated non-genic regions show excess of genes with highly negative Tajima's D
- Identified 385 genes showing evidence of positive selection

Tajima's D across Chromosome 1





Verification Using Sequence Data

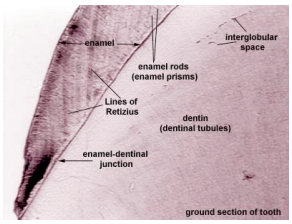
- Several of the genes have been verified by re-sequencing projects (SeattleSNPs and EGP)
- Targeted sequencing of genes
  - Genes with:
    - Tajima's D < -1.0
    - Tajima's D > 4.0
  - 1.5 kb each, targeted to HapMap polymorphic sites
    - ENAM - dental enamelin
    - PKD1-like - polycystic kidney disease 1-like
    - POLL - Polymerase lamda
    - RAGE - Renal tumor antigen

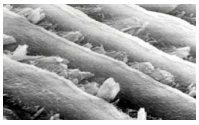
An in-depth study: *enamelin*

- Extreme population differences
- Ecological relevance
- Potential for phenotypic correlation

What is the evolutionary history of *enamelin*?

Tooth enamel properties





- Hardest, most mineralized tissue in human body
- Made up of small mineralized crystallites
- 85% mineral by volume

Tooth enamel thickness

- Tooth enamel thickness used in characterizing hominoid fossils
- Thickness varies from tooth to tooth, individual to individual and species to species
  - Enamel thickness is heritable (Hlusko *et al.* 2004)
  - Documented human population specific differences in enamel thickness (Harris *et al.* 2001)
- Diet correlated to enamel thickness
  - Carnivores / hard-object feeders = thick enamel
  - Herbivores / soft-object feeders = thin enamel

Enamel thickness examples

Herbivore

Carnivore

enamelin and tooth enamel formation

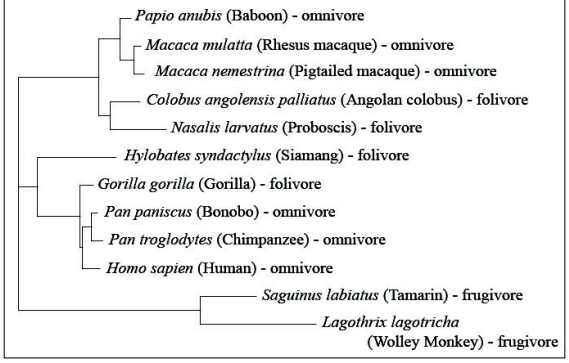
- Encodes secretory protein
  - Largest protein in enamel matrix
  - Comprises 5% of total enamel matrix protein
- Function
  - Involved in determining enamel crystallite growth and length
  - Critical for proper enamel formation
    - Mutations lead to thin, malformed enamel

⇨**Enamel thickness**

Evolutionary history of *enamelin*

- Identified by polymorphism study
  - Tajima's D
  - Confounded by population demographics or ascertainment biases
- Need to confirm
  - Direct sequencing
  - Other methods: d_N/d_S

enamelin primate species tree



Evolutionary selection between species

- Synonymous nucleotide substitution (d_S) – amino acid remains unchanged
- Non-synonymous substitution (d_N) – encoded amino acid changes

No selection (Neutrality): $d_N / d_S = 1$

Purifying selection: $d_N / d_S < 1$

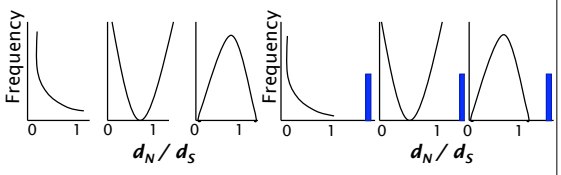
Positive selection: $d_N / d_S > 1$

d_N / d_S - measure of selective pressure

d_N/d_S Model Comparisons

Compare likelihood of neutral vs. selection models

Neutral model	Selection model
Assume beta distribution estimated of d_N / d_S in interval (0,1) (all site classes fall within (0,1)	Beta distribution, d_N / d_S (one unrestricted site class)

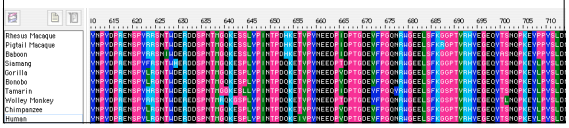


Methods of Nielsen and Yang (1998), Yang et al. (2000)

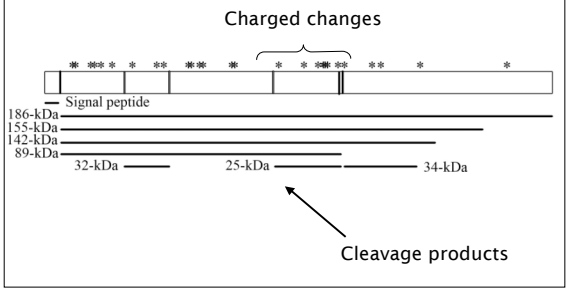
d_N/d_S sites model analysis

Models Compared	-2 Δ lnL	d_N/d_S estimates	% sites under selection
Neutral (M1) vs. Selection (M2)	35.04** (df = 2)	6.76	4.3
One-ratio (M0) vs. Discrete (M3)	63.22** (df = 4)	6.76	4.3
Beta (M7) vs. Beta & ω (M8)	35.06** (df = 2)	6.80	4.3

** p<0.01



Predicted sites under selection



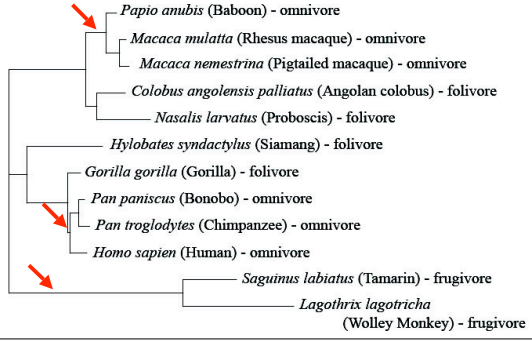
Have specific lineages been subject to positive selection?

- Correlations exist between diet and enamel thickness in primates
- Are specific lineages with dietary changes correlated to bursts of adaptive evolution?

Method to test lineage specific selection

- Reconstruct ancestral diet
- Identify lineages with dietary shift
- Hypothesis: selection on lineages with dietary shift
- Test for selection
  - Neutral model branches w = 1
  - Selection model branches w = free

Bursts of adaptive evolution are correlated to lineages with diet change

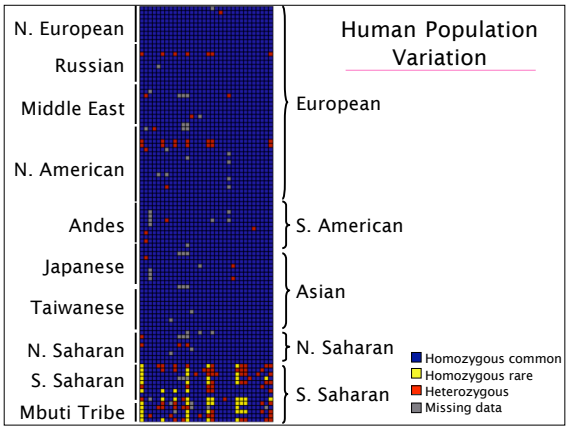


Existing evidence for molecular change tracking diet change

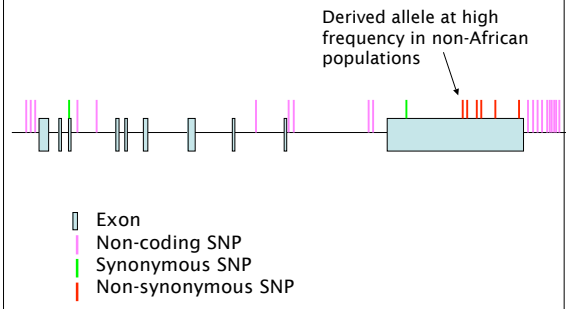
- Lysozyme and pancreatic RNAse
  (Messier & Stewart 1997, Yang 1998, Zhang 2003, Zhang 2006)
  - Specific activity in foregut fermenting species (ruminants and colobine monkeys)
  - Adaptive bursts coinciding with dietary changes

enamelin primate evolution

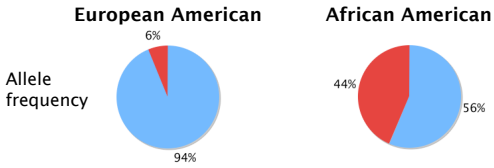
- Primate d_N/d_S analysis indicates adaptive evolution
- *enamelin* molecular change tracks with dietary change
- What is the pattern of adaptive evolution within humans?



Distribution of SNPs in *enamelin*



Non-synonymous derived allele and enamel thickness



Documented population specific differences in tooth enamel thickness (Harris *et al.* 2001)

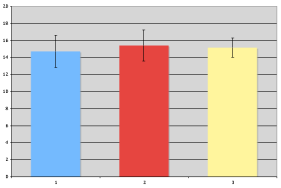
Mean enamel thickness	**5.04 mm**	**5.56 mm**

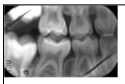
Association study:
non-synonymous SNP and enamel thickness

- 220 African and African American patients
- Bite-wing radiographs for enamel thickness measurements
- Cheek swabs for DNA sequencing
- Statistical analysis for association study

Ilona Khosh, dental student, taking a cheek swab

Preliminary Results





Conclusions

- Primate analysis indicates adaptive evolution
- Human polymorphism data provide evidence for human population specific adaptation
- Association study between genotype and enamel thickness phenotype

Questions

- What two mechanisms can reduce nucleotide variation?
- How can we potentially tell the difference between population demographics and positive selection?
- What does an association study look for a correlation between?
- What does Tajima's D test statistic measure?
- Describe a genome-wide scan for adaptive evolution.