Population Genetic Variation

Overview

- Polymorphisms
- Genetic Linkage
- Application: Natural Selection

Polymorphisms

- Types
- Allele Frequencies/Hardy Weinberg
- Ancestral State
- Population differences

Types of Polymorphisms

- Single Nucleotide Polymorphisms
- Variable number tandem repeats
- Insertion/Deletions
- Large-scale polymorphisms (inversions, copy number polymorphisms)

Types of Polymorphisms

- Single nucleotide polymorphisms (SNPs)

Sickle Cell Anemia

> 30

Huntington's Disease
Types of Polymorphisms

- Insertion/Deletions

Cystic Fibrosis

Allele and Genotype Frequencies

2 alleles: A (16) and T (4)

- P(A): p = 0.8
- P(T): q = 0.2

3 genotypes: AA (6), TT (0), AT (4)

- P(AA): p^2 = 0.64
- P(AT): 2pq = 0.32
- P(TT): q^2 = 0.04

Hardy Weinberg Principle

Both allele and genotype frequencies in a population are in equilibrium from generation to generation in a population with the following assumptions:

- Random mating (eg. no inbreeding or assortative mating)
- No Mutation
- No Migration
- No Selection
- Large Population

Hardy Weinberg Principle

3 genotypes: Homozygotes AA (6) and TT (0), Heterozygotes AT (4)

2 alleles: A (16) and T (4)

- p + q = 1
- \( p^2 \) + 2pq + q^2 = 1

Observed vs Expected

- P(AA) = p^2 = 0.64
- P(AT) = 2pq = 0.32
- P(TT) = q^2 = 0.04

Genetic Drift

Generation
We determine the ancestral state based on comparison to closely related species (outgroup).

The allele that is present in the outgroup is likely the ancestral allele.
Natural History of a Polymorphism

- TGAGG - 5.7 mya

Chimpanzee
- TGAGG

Africa
- TGTTG

Europe

Human

Natural History of a Polymorphism

- TGAGG - 5.7 mya

Chimpanzee
- TGAGG

Africa
- TGTTG

Europe

Human

Natural History of a Polymorphism

- TGAGG - 5.7 mya

Chimpanzee
- TGAGG

Africa
- TGTTG

Europe

Human

Natural History of a Polymorphism

- TGAGG - 5.7 mya

Chimpanzee
- TGAGG

Africa
- TGTTG

Europe

Human

Population Differences

- We can measure the differences in allele frequency in populations by the statistic Fst

- Fst estimates the reduction in heterozygosity expected when 2 different populations are erroneously grouped

Measuring Derived Allele Frequency

\[
\text{chimp CATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTG}
\]

Non-ancestral (derived) allele is T, it is at 20% in the population

2 isolated populations

Subpopulation 1
P(AA) = 1
P(A) = 1

Subpopulation 2
P(TT) = 1
P(T) = 1
Subpopulation 1
P(A) = 1
P(T) = 0

Subpopulation 2
P(A) = 0
P(T) = 1

Total
P(A) = 0.5
P(T) = 0.5

Fst = \frac{\text{Heterozygosity (total) - Heterozygosity (subpopulations)}}{\text{Heterozygosity (total)}}

Fst = \frac{0.5 - 0}{0.5} = 1

Subpopulation 1
H = 0.18

Subpopulation 2
H = 0.42

Total
H = 0.32

Fst = \frac{0.32 - 0.30}{0.32} = 0.0625

Population 1: AA (4) and TT (0), AT (1)
Population 2: AA (2) and TT (0), AT (3)

Linkage disequilibrium
Allele correlations

Law of independent assortment

Not always independent
Locus A
P(1) = 0.8
P(2) = 0.2

Locus B
P(1) = 0.5
P(2) = 0.5

Loci AB
P(11) = 0.4
P(12) = 0.4
P(21) = 0.1
P(22) = 0.1

Expected
Loci AB
P(11) = 0.3
P(12) = 0.5
P(21) = 0.2
P(22) = 0

Observed
A B
1 1
1 2
2 1
2 2

Not always independent

LD Measurement: D

D = \text{ObsP}(11)\cdot\text{ObsP}(22) - \text{ObsP}(12)\cdot\text{ObsP}(21)

if in complete equilibrium, then this just equals

\[ D = \frac{\text{ObsP}(11)\cdot\text{ObsP}(22) - \text{ObsP}(12)\cdot\text{ObsP}(21)}{\text{ExpP}(12)\cdot\text{ExpP}(21)} \]

which equals 0

LD Measurement: D'

D = \text{ObsP}(11)\cdot\text{ObsP}(22) - \text{ObsP}(12)\cdot\text{ObsP}(21)

D' = \frac{D}{D_{\text{max}}} = \frac{0.1}{0.1} = 1

LD Measurement: D'

D = \text{ExpP}(12)\cdot\text{ExpP}(21)

D' = \frac{D}{D_{\text{max}}} = \frac{0.25}{0.25} = 1

LD Measurement: D'

D = \text{ObsP}(12)\cdot\text{ObsP}(21)

D' = \frac{D}{D_{\text{max}}} = \frac{0.09}{0.09} = 1

LD Measurement: D'

D = \text{ExpP}(12)\cdot\text{ExpP}(21)

D' = \frac{D}{D_{\text{max}}} = \frac{0.25}{0.25} = 1

LD Measurement: D'

D = \text{ObsP}(11)\cdot\text{ObsP}(22) - \text{ObsP}(12)\cdot\text{ObsP}(21)

D' = \frac{D}{D_{\text{max}}} = \frac{0.09}{0.09} = 1
Decay of LD

Complete LD but alleles not correlated

Decay of LD a function of the recombination rate \( r \) and time

History also matters

• Application to the human genome

Natural Selection

• Methods for detecting positive natural selection

• Localizing to individual polymorphisms

Forces Shaping Human Evolution

Genomics Signals of Natural Selection

1950s: HBB Resistance to Malaria

1990s: LCT Lactose tolerance

2000s: Scans of the Human genome

1858 1871 1948

1948 - J.B.S. Haldane's "Malaria Hypothesis"
The geographical distribution of P. falciparum malaria is correlated with common erythrocyte diseases suggests that carriers might have greater resistance to malarial infection.

The Study of Natural Selection in Humans

Sickle cell anemia
Thalassemias
G6PD
Ovalocytosis

1858

1871

1948

1950s: HBB Resistance to Malaria

prevalence

generations
Genomics Signals of Natural Selection

before

after

prevalence

generations

Tests:
1) Long-range correlations
   • iHS, XP-EHH

Long:

Broken Haplotype
Long Correlations

Genomics Signals of Natural Selection

before

after

Tests:
1) Long-range correlations
   • iHS, XP-EHH
2) High frequency derived

Genomics Signals of Natural Selection

before

after

Tests:
1) Long-range correlations
   • iHS, XP-EHH
2) High frequency derived
3) High differentiation
   • Fst

Genome-wide searches for selected loci now possible

The Genomic Era

New resources
- Complete genomes: human, chimpanzee, mouse, gorilla, etc...
- Large database of known genetic variants
- Large datasets of genotype data

Scanning the Genome for Selection

Nodes
~300 regions detected in genome surveys

~5 regions with known functional adaptations

Combining scores can localize regions

1. Long Haplotype

2. Derived allele

3. Differentiated

Simulations to investigate selected regions

Simulation mimics 1000 Genomes and HapMa
1500 replicates, 1 Mb region, ~10,000 SNPs
Sweeps to 20, 40, 60, 80, and 100%

S. Schaffner et al., Gen. Res. 15:1576-1583.(2005)

Properties of signals in selected regions

Absent selection, signals are nearly independent
Composite Likelihood

\[ CL(x) = \sum \log P(\text{causal} \mid \text{stat}(x)) \]

\[ CL = \log(0.001) + \log(0.005) + \log(0.023) + \log(0.005) \]

Properties of signals in selected regions

Applied to Known Pigmentation Gene: MATP

Applying localization to regions

Chr15: Northern Europe, Skin pigmentation

SLC24A5 is a determinant of pigment in zebrafish.
Threonine to alanine mutation associated with skin pigment differences in humans.

Chr2: Asia, Hair traits

In putative binding site of EDAR's death domain.
Linked to hair diameter in Japanese population sample.
Hair traits identified

In putative binding site of EDAR's death domain

Linked to hair diameter in Japanese population sample

Overexpression of EDAR creates East Asian hair form
C Mou et al., Human Mutation, epub June 17 (2008)

Hairst and Sweat in Human Evolution

What are different about these families?

Chr10: Asia, Trait Unknown

Variant in highly conserved cadherin domain

Potential to influence hearing, vision or balance

Plays key role in development and function of sensory hair cells

Mutations in PCDH15 cause Usher syndrome, a disorder marked deafness, blindness, and severe balance problems

Residue in calcium binding pocket of cadherin repeat
B. Nagar et al., Nature 380:360-64 (1996)

Chr12: Africa, Trait Unknown

Peak signal in intron of the gene PAWR
Involved in T cell proliferation and regulation
Mice homozygous for disruptions in the gene display increased proliferations of both B and T cells, and enhanced T cell activation in response to stimulation
Potential to regulate gene expression

- Measured gene expression levels for in all HapMap individuals’ cell lines
- Correlated expression levels of each gene with all genetic variants within 1 Mb
- Variants near PAWR are associated with significant differences in expression of the gene in African populations

Future Directions

- Signal of Selection
- Target variant
- Normal Function
- Functional Variation
- Other Signals Evolutionary Adaptation

- Apply to new datasets
  - HapMap3
  - 1000 genomes

- Follow up candidates
  - Additional genotyping & sequencing
  - Functional validation
  - Model organisms

Applied to Unknown Regions

- Coding mutations
  - 20 high-scoring coding mutations

Regulatory changes

- Pigmentation in non-African populations
  - 40 regions that hit eQTLs (p<10^-5)
  - 31 regions contain lincRNAs
  - 6 peaks contain splicing QTLs

40 regions that hit eQTLs (p<10^-5)
31 regions contain lincRNAs
6 peaks contain splicing QTLs
Metabolism in all populations

Response to infection in populations with disease

Natural Selection Shaped Human Genome

What have we found