Natural Selection – and the distribution of IBD in the human genome

Rasmus Nielsen
Departments of Integrative Biology and Statistics, UC-Berkeley
and
Department of Biology, University of Copenhagen

Identity By Descent (IBD) and Selection

- Previous methods for identifying selection in the human genome have indirectly used IBD.
- During a selective sweep, IBD in the population increases.

Calculation of Probabilities of IBD

\[ F = \sum_{i=1}^{k} p_i^2 F_i \]

- \( F \) = probability of identity by descent
- \( k \) = number of alleles segregating in the population
- \( p_i \) = frequency of allele \( i \)
- \( F_i \) = \( F \) within allele \( i \)

Selective Sweeps

New advantageous mutation

Change of \( F \) in one generation with selection

\[ F(t+1) = \sum_{i=1}^{k} p_i(t+1)^2 \frac{1}{p_i(t)N} \]

\[ p_i(t+1) = \frac{\omega_i p_i(t)}{\bar{\omega}}, \quad \bar{\omega} = \sum_{i=1}^{k} \omega_i p_i(t) \]

\[ F(t+1) = \sum_{i=1}^{k} \left( \frac{\omega_i p_i(t)}{\bar{\omega}} \right)^2 \frac{1}{N p_i(t)} = \frac{\omega^2}{\bar{\omega}^2} \frac{1}{N} \]
Increase in IBD (1)

In a previously outbred population the relative increase in IBD due to selection is

\[ \frac{W^2(t)}{W^2(0)} \]

This result is quite general and holds for any type of selection.

Change of $F$ in a selective sweep

Change in allele frequency:

\[ \frac{\partial p_i(t)}{\partial t} = s_i (1 - p_i(t)) p_i(t) \]

With solution:

\[ p_i(t) = \frac{e^{s_i t} p_i(0)}{1 - p_i(0) + e^{s_i t} p_i(0)} \]

Re-writing the result from before:

\[ F_i(t + 1) - F_i(t) = \frac{1 - F_i(t)}{N p_i(t)} \]

In continuous time:

\[ \frac{\partial F_i(t)}{\partial t} = \frac{1 - F_i(t)}{N p_i(t)} \]

With boundary condition $F_i(0) = 0$, we get:

Increase in IBD (2)

The IBD of selected allele $i$ during a selective sweep is, if the population is initially outbred, is

\[ F_i(t) = 1 - \exp \left( -\frac{e^{s_i t} p(0) - p(0) + s_i t - s_i p(0) t}{s_i (1 - p(0)) N} \right) \]

$p(0)$: initial frequency of selected allele

$t$: generations after sweep began

$N$: chromosomal population size

$s_i$: selection coefficient acting on allele $i$

Selection on standing variation

![IBD vs Generations](chart)

Initial allele frequency: 1%, $s = 0.1$, $N = 100,000$

Inference of IBD tracts

- Assume a Hidden Markov Model (HMM) of pairwise IBD relationship.

- Use dense SNP genotype maps.


Time (distance) dependent transition probabilities can then be found analytically by exponentiation of $Q$.

If two individuals are only related through one cycle in the pedigree then $k_{ij}$ is a simple function of the number of meiosis $m$ and the recombination rate $\phi$.

$$\alpha = -m \log(1 - \phi)$$

The number of meiosis can be calculated from the relatedness estimates

$$m = m_a + m_b, m_i = 1 - \log(x_i)/\log(2),$$

where

$$x_a = \frac{k_1 + 2k_2 + \sqrt{(k_1 + 2k_2)^2 - 4k_2}}{2},$$

$$x_b = \frac{k_2}{x_a}$$

Genotyping errors are also incorporated in the emission probabilities but are not shown here.
Original motivation: relatedness mapping

- 7 cancer patients
- 60 controls

Population based IBD mapping

- Combining evidence of relatedness to locate a locus shared by IBD.
- If disease mutation was introduced recently, affected individuals will share more IBD in the region around the mutation.
- Does not require inference of specific haplotypes – or choice on how to test based on haplotypes.
- Deals with the multiple testing problem.
- Has the potential to combine information from pedigrees and outbred populations directly.
- May have very high mapping power – especially when the mutation is recent and rare.

Selection on standing variation (IBD)

- Initial frequency 10%.
- $p=0.1$
- $N=10,000$
Selection on standing variation (Tajima’s D)

- Initial frequency 10%.
- $S=0.1$
- $N=10,000$

Human Genome - CEPH

Overdominance

HLA Region

Human Genome - CEPH

HLA Region

HLA (MHC)

- Some of the most important immune and defense related genes.
- Bind antigens and present them on the outside of the cell to T-cells.
- Stimulate anti-body producing B-cells or attract killer T cells that destroy cells.
- Also important in auto-immune diseases (incl. type I diabetes) and in cancer.
- The first set of genes to be demonstrated to be under selection in humans due to high dN/dS ratios and generally extremely high levels of variability, presumably due to balancing selection.
MHC residues predicted to be under positive selection are located in the antigen recognition site.

Acknowledgments

- Anders Albrechtsen, University of Copenhagen
- Ida Moltke, University of Copenhagen