A MAP OF THE RECENT POSITIVE SELECTION IN THE HUMAN GENOME

Benjamin F. Voight, Sridhar Kudaravalli, Xiaoquan Wen, Jonathan K. Pritchard

Presented by:
Debolina Ganguly
Overview:

- Identifies signals of recent positive selection, knowledge of which provides information about the adaptation of modern humans to local conditions.
- Dramatic changes in the local environment resulted in powerful selection pressures on new genotypes that are better suited for the new environments. Examples: Response to Malaria, Lactase gene in response to dairy farming etc.
Best examples of recent selection, until now are from studies of candidate genes.

Thus it is not known as of:

- how widespread these signals are
- whether these are the same genes that were important in the earlier evolution of the human lineage
- whether they are geographically restricted
Aim of the study

- Find loci where there is strong, very recent selection in favor of alleles that have not yet reached fixation.
- Detect signals of selective sweeps in progress.
- Creation of selection maps on ongoing sweeps.
Results

- Analyzed genome wide SNP data from phase 1 of the Hap Map project. Had 800,000 polymorphic SNP’s in a total of 309 unrelated individuals.
- There were 3 distinct population samples of unrelated individuals:
  - 89 Japanese and Han Chinese individuals from Tokyo and Beijing respectively referred as the East Asians.
  - 60 individuals from northern and western Europe.
  - 60 Yoruba from Ibadan, Nigeria.
- Unless mentioned, study focused on autosomes only.
- They derived genome wide high resolution LD based recombination maps separately for all 3 samples.
- The goal of the study was the identification of loci where strong selection had driven new alleles up to intermediate frequencies.
The data consisted of pre ascertained SNP’s for the genome wide scale and a new test statistic the integrated haplotype score (iHS) was determined.

- The test began with extended haplotype homozygosity (EHH). EHH measures the decay of identity, as a function of distance of haplotypes that carry a particular core allele at one end. The haplotype homozygosity for each allele starts from 1 and decays to 0 as its distance increases from the core site.
- As an allele increases in frequency due to strong selection, it tends to have high levels of haplotype homozygosity extending further away than what is expected under the neutral model. The integrated EHH (iHH) is denoted as $IHH_A$ or $IHH_D$. 
Figure 1. Decay of EHH in Simulated Data for an Allele at Frequency 0.5
(A) Decay of haplotypes in a single region in which a new selected allele (red, center column) is sweeping to fixation, replacing the ancestral allele (blue). Horizontal lines are haplotypes; SNP positions are marked below the haplotype plot using blue for SNPs with intermediate allele frequencies (minor allele .0.2), and red otherwise. For a given SNP, adjacent haplotypes with the same color carry identical genotypes everywhere between that SNP and the central (selected) site. The left- and right-hand sides are sorted separately. Haplotypes are no longer plotted beyond the points at which they become unique.
(B) Decay of haplotype homozygosity for ten replicate simulations. When the core SNP is neutral ($s=0$; left side) the haplotype homozygosity decays at similar rates for both ancestral and derived alleles. When the derived alleles are favored ($s=2N\alpha=250$; right side), the haplotype homozygosity decays much slower for the derived alleles than for the ancestral alleles. The discrepancy in the overall areas spanned by these two curves forms the basis of our text for selection (iHS).
Figure 3. Plots of Chromosome 2 SNPs with Extreme iHS Values Indicate Discrete Clusters of Signals
SNPs with $|\text{iHS}| > 2.5$ (top 1%) are plotted. The bottom plot combines signals for all three populations, plotting only SNPs with derived frequency $>0.5$ and $\text{iHS} < -2.5$. Such SNPs correspond to high-frequency-derived SNPs in the range for which our test is most powerful. The short vertical bars below each plot indicate 100-kb windows whose signals are in the top 1% of windows genome-wide.
Figure 6. Signals of Selection for Three Candidate Selection Regions Discussed in the Text

The columns show (left) scatter plots of negative iHS scores, (center) haplotype plots, and (right) decay of haplotype homozygosity. In each case the Core SNP for the center and right-hand plots was chosen as a SNP with high negative iHS score (starred in the scatter plots); the allele marked in red is derived. For each signal, values are listed for the derived allele frequency ($p_d$) and the local deCode recombination rate estimate.
For every gene, the number of SNP’s with high iHS value in a 50 SNP window centered on the gene was determined. Genes in the top 10% are considered targets of selection.

- Chemosensory perception, olfaction as well as gametogenesis and fertilization. Might be due to sexual competition and defense against pathogens.
- Genes related to the metabolism of carbohydrates, lipids, phosphates and vitamin C.
- Genes in skeletal development and hair formation and patterning in Yoruba.
- Alcohol dehydrogenase cluster in east Asians, carbohydrate metabolism genes like mannose in Yoruba and sucrose and lactose in the Europeans.
- 2 microcephaly genes namely, CDK5RAP2 in Yoruba and CENP in Europeans and east Asians.
- Electron transport genes in Europeans (CYP genes).
  Detailed report in Table 2 of the paper.
Figure 7. Sharing of iHS Signals between Populations
The numbers listed inside circles represent the numbers of 100-kb windows that are in the top 1% of the empirical distributions in at least one population. The numbers in the intersection regions are in the top 1% for one population, and the top 5% for one or both of the other populations. The counts that would be expected if signals were independent across populations are shown in parentheses. The number of windows not in any circle is reported in the upper-left corner.
Discussions Points:

- **How effective are the study statistics?**
- **According to the paper the candidate sweep regions tend to be narrower in Yoruba than in the non African populations, indicating that sweep events are younger in the non African populations. Why?**