Haplotter Guide

March 8, 2006

Introduction

Haplotter is a web tool that has been developed to display the results of a scan for positive selection in the human genome using the HapMap data (www.hapmap.org). The publication associated with this tool is Voight et al 2006.

iHS (Integrated Haplotype Score) is a statistic that has been developed to detect evidence of recent positive selection at a locus. It is based on the differential levels of linkage disequilibrium(LD) surrounding a positively selected allele compared to the background allele at the same position. An extreme positive iHS score ($iHS > 2$) means that haplotypes on the ancestral allele background are longer compared to derived allele background. An extreme negative iHS score ($iHS < -2$) means that the haplotypes on the derived allele background are longer compared to the haplotypes associated with the ancestral allele. Under our model we expect selected derived alleles to harbour excessive LD relative to the background. However, an extreme positive iHS score is also considered as a candidate because the ancetral allele may be hitchhiking along with the selected allele, or the ancestral allele itself may be the target of selection. iHS has the highest power to detect selection when the selected allele has swept to intermediate frequencies. The power of iHS also depends on many other factors such as local SNP density, SNP ascertainment, availability of haplotype phase information and intensity of selection among others. For more detailed information on this statistic and results of a whole genome scan, please refer to our associated publication (Voight et al, 2006).

Fay and Wu's H (Fay and Wu, 2000) and Tajima's D (Tajima, 1989) are based on the frequencies of the polymorphims segregating in the region of interest. Episodes of selection (positive or negative) tend to skew the frequencies in the different directions compared to neutral model. Positive selection results in an excess of high frequency derived alleles compared to neutral expectations when the selected allele has swept to high frequencies. Fay and Wu's H is sensitive to this signature in the frequency spectrum. H has the greatest power to detect selection when the selected allele is very close to being fixed or has just fixed in the population. Positive selection also results in an excess of low frequency polymorphisms, especially when the selected allele is close to fixation or right after fixation. This signature is used to detect positive selection by Tajima's D. Statistics that normally summarize the frequency spectrum, like the H and D, normally require reseqeunced data. However, we calculated these statistics for the HapMap SNP data treating the data as if it were resequenced data. This has the effect of biasing the statistics, since the data in the hapmap represents a preacertained subset of the actual sequence variation which underlies the data. As a result, p-values are based on an empirical rather than from a population genetics based model, although the underlying intuition is the same, i.e. signals of selective sweeps will result in high negative D and H . We showed a correlation between these measures and our statistic under a partial seletive sweep model of selection, but not under a neutral model. (Voight et al, 2006), which demonstrates the utility of both aspects of the data. Finally, F_{st} is a measure that is used to measure the degree of population differentiation. In situations where selection is

restricted to certain populations or geographical locations, the allele frequencies at the locus that are is undergoing selection may vary significantly between different populations. F_{st} can be used to measure the level differentiation between populations at a locus.

Features

Haplotter can be used as a resource to examine various population genetic measures in a genomic region. Measures that are currently displayed include iHS, ascertainment biased versions of Fay and Wu's H, Tajima's D and F_{st}. For information on how these statistics were generated please refer to the associated publication (Voight et al, 2006).

Usage

There are three ways in which the data can be viewed:

- Gene-centered approach
- Region-centered aproach
- Single SNP query.

These options are available in the left panel of the webpage.

There are four graphic panels displayed for each gene or region queried. They represent iHS, ascertainment versions of Fay and Wu's H, Tajima's D and F_{st}.

Of the four display panels, the first three are displayed for all three of our study populations (CEPH, Yoruba and East Asians), while the F_{st} plot consists of the three pairwise comparisons of the three populations. Each point on the y-axis for these plots represents the negative log of the rank of the observed statistic for a given SNP divided by the total number of SNPs. The statistic that is ranked is obtained independently for each of the four statistics separately for each population. For iHS, for each SNP, 25 SNPs on either side of the SNP are scanned for $|iHS| > 2$. The proportion of SNPs in this 51 SNP window with $|iHS| > 2$ is computed. For H and D, the estimated value of H and D (see Voight et al, 2006 for details on how these were estimated) were used for ranking. For F_{st}, the statistic to be ranked is obtained in a similarmanner as that for iHS except for each population comparison, the thresholds for defining a significant F_{st} is based on the top 5% cutoff for each population comparion. The different thresholds used for F_{st} are (CEPH-Yoruba: 0.2976, CEPH-East Asians: 0.2055, Yoruba-East Asians: 0.3374). In addition to these, the SNPs with very high F_{st} (in the top 1% within each population comparion) are plotted as points with their F_{st} value represented on the right hand side of the plot.

The information displayed for the gene and region display methods varies slightly. The horizontal bars of varying sizes displayed under each panel of the graphic display represent

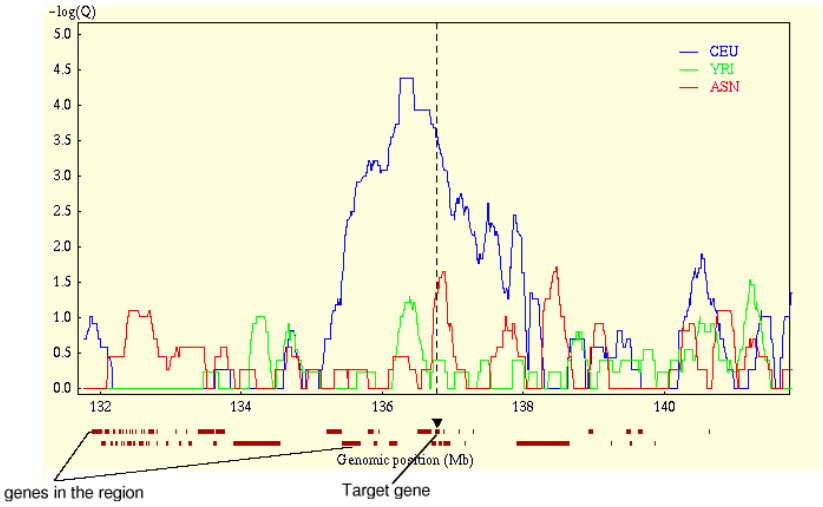

Figure 1: A sample iHS plot output by querying a gene

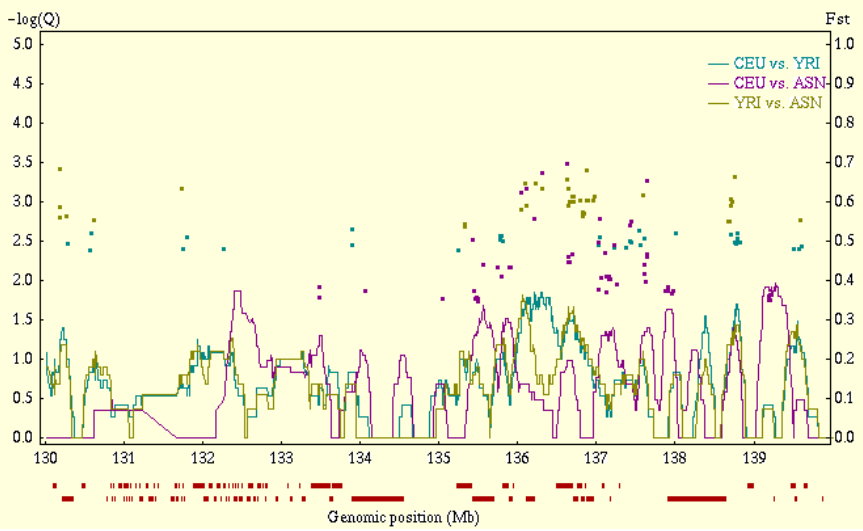

Figure 2: A sample F_{st} plot output by querying a genomic region

genes present the region. If a specific gene is queried then the center of that gene is marked by ablack arrow. A vertical dashed line is used as a reference for comparisons within a graphic panel and across the graph panels.

In addition to the four graph panels, there are two tables displayed at the bottom of the page. The first table is a list of all the genes that are present in the region and have data available in at least one of the populations. The values presented for each gene and for each population represent an empirical p-value. Please refer to the associated publication (Voight et al, 2006) for details on how these were estimated. Within the table if a cell is considered to show significant evidence for selection, then it is colored light blue. If a gene was queried, the entire row representing the gene is colored yellow.

401013	FLJ34870	133893259 - 134536800	0.999955	0.304367	0.548327
4249	MGAT5	135219937 - 135417237	0.093768	0.999955	0.542952
81615	DKFZP566N034	135424329 - 135687339	0.003801	0.999955	0.541331
130013	ACMSD	135806957 - 135870373	0.000679	0.607928	0.999954
905	CCNT2	135887162 - 135925356	0.000543	0.999955	0.999954
80122	YSK4	135947981 - 135949991	0.000679	0.999955	0.999954
84083	ZRANB3	136105255 - 136207200	0.000226	0.607928	0.359281
23518	R3HDM1	136499852 - 136693608	0.000000	0.408690	0.539616
23190	UBXD2	136710107 - 136753395	0.000091	0.408690	0.060467
3938	LCT	136756184 - 136805519	0.000226	0.408690	0.043647
4175	MCM6	136807965 - 136844780	0.000407	0.408690	0.026874
391448	LOC391448	136866288 - 136867465	0.000679	0.408690	0.023723
1615	DARS	136875023 - 136953885	0.000860	0.999955	0.060467
7852	CXCR4	137082676 - 137086487	-	-	0.539616
389053	LOC389053	137168110 - 137168949	0.001946	0.607928	0.539616
80731	KIAA1679	137904047 - 138646056	0.191067	0.618052	0.153693
3176	HNMT	138932577 - 138983005	0.541838	0.999955	0.359281

Genes with "significant" results

Gene queried

Figure 3: A sample text output by querying a gene

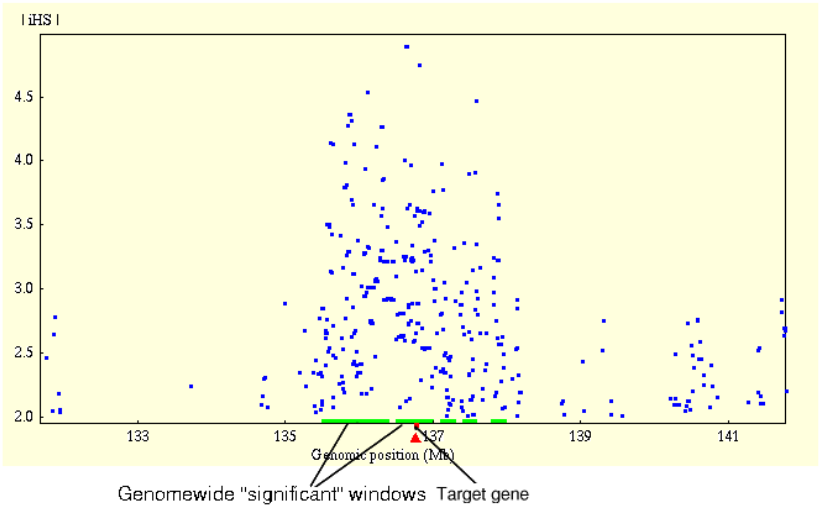

Figure 4: A sample iHS scatter plot

Clicking on the gene would result in a similar set of graphs being displayed for that gene. In addition to this, clicking on the p-values would display a scatter plot of iHS values in and around the gene. The red horizontal bar displayed in this plot represents the gene coresponding to the p-value that was clicked, while green horizontal bars represent 100Kb windows that lie in the top 1% of our results for each population i.e. areas marked in green show especially strong evidence for selection. Below this plot is a table with the top 20 iHS scores in the displayed region.

The second and last table lists regions that show significant evidence for selection, but lie in nongenic regions of the genome.

In addition to the region and gene centered views of the data, one can query a single SNP. This will display two graphs. The first graph consists of an ordered display of haplotypes at

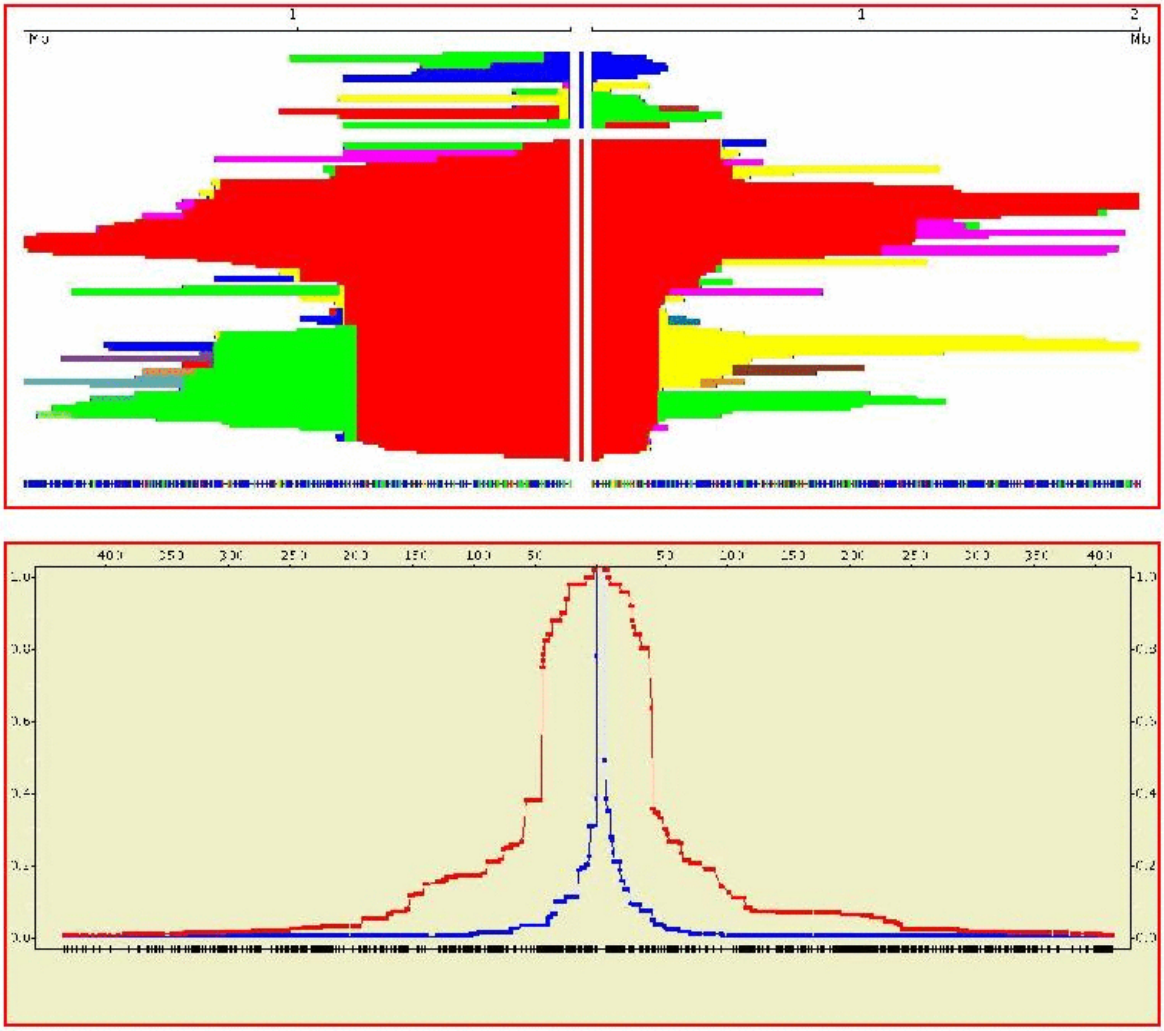

Figure 5: A sample output by querying a SNP

different distances from the queried SNP (which is at the center of the display). At the center of this display is a vertical line in two colors: blue color represents the ancestral state and the derived state is represented in red. The distances over which the haplotypes are spread is displayed at the top of the graph. A continous block of the same color represents a haplotype block that is shared among many chromosomes. When a chromosome switches to a new color at some distance away from the core SNP, it means that that particular chromosome has a different allele relative to the remaining chromosomes that shared a common haplotype with it until that distance. In effect, an origin of a new color represent a new haplotype from that point on. Haplotypes are no longer plotted if they become unique in the sample. The second graph displays the decay of Extended Haplotype Homozygosity (EHH) at different distances from the queried SNP (Sabeti et al, 2002). The table below this figure displays the iHS, Fay and Wu's H, Tajima's D, derived allele frequency and F_{st} (between the chosen population and the remaining two populations).

References

1. Voight B.F, Kudaravalli S., Wen X., Pritchard J.K. (2006). A map of recent positive selection in the human genome. PLoS Biology 4(3): e72

2. Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.

3. Tajima, F., (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595

4. Fay J.C., Wu C.I. (2000) Hitchhiking under positive Darwinian selection. Genetics Jul;155(3):1405-13