

# **Sweep Documentation**

December 20, 2005

*Note: I will be working to update the documentation in the early new year. Please let me know if there is anything that will make it more useful. Thanks, Pardis  
pardis@broad.mit.edu*

# Table of Contents

<b>DEFINITIONS</b>	<b>4</b>
EXTENDED HAPLOTYPE HOMOZYGOSITY (EHH)	4
RELATIVE EXTENDED HAPLOTYPE HOMOZYGOSITY (REHH)	4
CORE HOMOZYGOSITY	5
MARKER BREAKDOWN	5
<b>INPUTS</b>	<b>6</b>
GENOTYPE DATA FILE	6
SNP DATA FILE	7
.MANY FILE	7
ANCESTRAL.TAB	7
OPENING INPUT FILES	8
<b>CORE SELECTION</b>	<b>10</b>
LOADING CORES	10
OPTIONS FOR CORES	11
<b>LONG-DISTANCE MATCHING</b>	<b>13</b>
TYPES OF DISTANCES TO MATCH	13
<b>MAIN PAGE IMAGES</b>	<b>14</b>
EHH/REHH vs. FREQUENCY SCATTER PLOT	14
GENE TRACK	14
EHH/REHH vs. DISTANCE CHART	15
HAPLOTYPE BIFURCATION DIAGRAMS	15
HAPLOTYPE TABLE	15
ANCESTRAL TREE	16
<b>EXPORTING IMAGES</b>	<b>17</b>
EXPORTING PROCEDURE	17
<b>EXPORTING DATA</b>	<b>18</b>
EHH vs. FREQUENCY DATA	18
EHH/MARKERH CORRELATION DATA FOR ALL CORES	18
EHH/DISTANCE CORRELATION DATA FOR ALL CORES	19

<b>EHH DATA FOR THIS CORE</b>	<b>20</b>
<b>SNP FREQUENCY</b>	<b>20</b>
<b>TOOLS</b>	<b>21</b>
<hr/>	
<b>SHOW EHH vs. FREQUENCY PERCENTILES</b>	<b>21</b>
<b>EXPORT EHH vs. FREQUENCY DENSITY</b>	<b>21</b>
<b>EHH SIGNIFICANCE CALCULATOR</b>	<b>23</b>
<b>ANALYZE CORE H DISTRIBUTION</b>	<b>23</b>
<b>MULTIPLE POPULATION ANALYSIS</b>	<b>24</b>
<b>REFERENCES</b>	<b>25</b>
<hr/>	

## Definitions

### ***Extended Haplotype Homozygosity (EHH)***

The extended haplotype homozygosity (EHH) is defined as the probability that two randomly chosen chromosomes carrying the core haplotype of interest are identical by descent (as assayed by homozygosity at all SNPs) for the entire interval from the core region to a distance  $x$  (Sabeti, Reich et al. 2002). EHH thus detects the transmission of an extended haplotype without recombination. The EHH of a tested core haplotype  $t$  is mathematically calculated as:

$$EHH_t = \frac{\sum_{i=1}^s \binom{e_{ti}}{2}}{\binom{c_t}{2}}$$

where  $c$  is the number of samples of a particular core haplotype,  $e$  is the number of samples of a particular extended haplotype, and  $s$  is the number of unique extended haplotypes.

### ***Relative Extended Haplotype Homozygosity (REHH)***

When we first designed the Long-Range Haplotype Test, there were no good estimates for local recombination rates. In order to correct for local variation in recombination rates, we therefore compare the EHH of the tested core haplotype to that of other core haplotypes present at a locus. We do this using the relative EHH (REHH), the factor by which EHH decays on the tested core haplotype compared to the decay of EHH on all other core haplotypes combined. With new fine-scale recombination rates, we can substitute using REHH and simply look at EHH at carefully matched genetic distances, however REHH is still as useful view (Sabeti, Reich et al. 2002).

To calculate REHH we first calculate the ' $\overline{EHH}$ ', the decay of EHH on all other core haplotypes combined. For this we use to the following equation where  $n$  is the number of different core haplotypes:

$$\overline{EHH} = \frac{\sum_{j=1, j \neq t}^n \left[ \sum_{i=1}^s \binom{e_i}{2} \right]}{\sum_{i=1, i \neq t}^n \binom{c_i}{2}}$$

The relative EHH (REHH) is then simply  $EHH_t / \overline{EHH}$ .

## **Core Homozygosity**

Core homozygosity is a measure of how much variation at a particular core was captured by the SNPs you genotyped. It is determined by the number and characteristics of SNPs genotyped at the core and by the historical haplotype structure of the region. You may want to compare haplotype blocks with the same number of SNPs for their core homozygosity, assessing the distribution and looking for outliers. It may also be useful to match the core homozygosity's across regions you are comparing to make sure that you are comparing analogous data structure.

The core homozygosity is defined as the probability that any 2 randomly chosen core haplotypes from a population will be the same. It is mathematically calculated as:

$$H_{core} = \frac{\sum_{i=1}^s \binom{c_i}{2}}{\binom{n}{2}}$$

where  $n$  is the number of chromosomes,  $c$  is the number of samples of a particular core haplotype, and  $s$  is the number of different core haplotypes.

## **Marker Breakdown**

When comparing EHH/REHH values across regions, it is important to ensure that you are calculating the value at a similar genetic distance. This program now also has cM values from the HapMap fine-scale recombination rates for humans, but for other genomes or for comparison you can match this by the 'marker breakdown,' that is the degree to which each added marker at a further distance causes the extended haplotypes to decay for all core haplotypes (Sabeti, Reich et al. 2002). This gives an evaluation of how much historical recombination (observed recombinants) has occurred over a distance from the core, and therefore what genetic distance you are looking at. This can be calculated as 'all EHH'.

$$allEHH = \frac{\sum_{j=1}^n \left[ \sum_{i=1}^s \binom{e_i}{2} \right]}{\sum_{i=1}^n \binom{c_i}{2}}$$

Where  $n$  is the number of different core haplotypes,  $c$  is the number of samples of a particular core haplotype,  $e$  is the number of samples of a particular extended haplotype, and  $s$  is the number of unique extended haplotypes.

## Inputs

Sweep requires two files as input, a genotype data file and a snp info file. It is helpful but not necessary to give both files the same name with the extensions “.emphase” or “.phase” and “.snp”. You can also load up many files at once using a file with the extension “.many”. We provide sample genotype data, snp data, and .many files for you to test with. There are several options for loading up files either from the File Menu or on the main page. This section goes in depth into the input files and process of loading.

### Sections

1. Genotype data file
2. SNP data file
3. .many file
4. ancestral.tab
5. Opening input files

### Genotype data file

Sweep accepts a standard format of genotype data, fully phased with missing data filled in. It is therefore important to have good quality data with few missing datapoints. We prepare these files using Genehunter to uncover unambiguous phasing using family data. We then use either our own emphase program or PHASE (Stephens, Smith et al. 2001; Stephens and Donnelly 2003) to get complete phased data.

The data format we use looks like what is described below. The columns are tab delimited:

- Column 1: the individual identifier.
- Column 2: the chromosome identifier. For autosomes you should have two chromosomes per individual. We label the two chromosomes T for transmitted and U for untransmitted, (but it can be anything eg. A and B.)
- Columns 3 – N: each column gives the allele for one SNP in the order of its position on the chromosome. The alleles are represented as A=1, C=2, G=3, T=4.

1331-1331FF12	T	1	3	3	2
1331-1331FF12	U	1	1	1	2
1331-1331FM13	T	1	3	3	2
1331-1331FM13	U	1	3	3	4
1331-1331MF14	T	1	1	3	4
1331-1331MF14	U	1	1	1	2
1331-1331MM15	T	1	1	3	4
1331-1331MM15	U	1	1	1	2

The first row therefore represents one chromosome for individual 1331-1331FF12 with the haplotype AGAAT. The second row represents the other chromosome for individual 1331-1331FF12 with the haplotype AAGGC. Etc...

## **SNP Data File**

The SNP data file has 3 tab-delimited columns, which gives information about the markers you genotyped. Be sure to have the three headings snpid, chr, and HG16 (or HG17 if you are using this build). This data is used to display chromosomal positions, match to refgene to bring up genes in the region, and pull chimp alleles for SNPs for which that information has been collected.

- Column 1: The SNP identifier. This can be an rs number or any other name you choose to give.
- Column 2: The chromosome.
- Column 3: The SNP position based on the build identified. HG16 and HG17 are currently recognized.

snpid	chr	HG16
rs267265	3	45548733
rs267262	3	45567119
rs267241	3	45578901
rs2005227	3	45598847
rs267230	3	45619948
rs2012755	3	45633347

## **.many file**

If you have many files you are studying simultaneously, you can load them all at once with a file with extension .many. On each line you give the genotype data file and matching SNP information file separated by a space or tab. The file will have the extension '.many'. You can load this file when given the option to open a genotype data file.

CCR5_ceph.emphase	CCR5_ceph.snp
FY_ceph.emphase	FY_ceph.snp
HBB_ceph.emphase	HBB_ceph.snp
HFE_ceph.emphase	HFE_ceph.snp
LCT_ceph.emphase	LCT_ceph.snp
G6PD_ceph.emphase	G6PD_ceph.snp
CD40L_ceph.emphase	CD40L_ceph.snp

## **ancestral.tab**

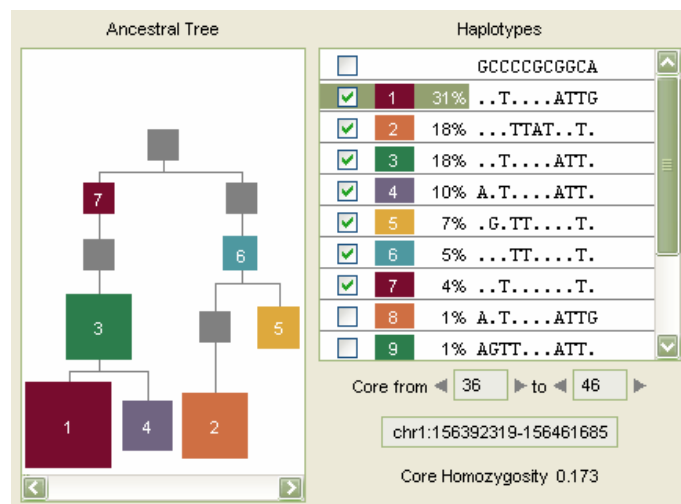
For SNPs where you have outgroup data, you may choose to predict the ancestral allele as the consensus outgroup allele. You can create a file ancestral.tab with two columns,

the name of the SNP and the ancestral allele. The file has no header. The file looks like this:


rs267265	G
rs2856762	T
rs267241	G
rs2005227	C
rs267230	A
rs2012755	N

For all SNPs from HapMap Release 16, the ancestral allele was predicted as the chimp allele and is already stored in the Sweep program.

Where ancestral information is available Sweep will display the predicted ancestral haplotypes above the list of haplotypes observed in your data. The ‘.’s in the observed haplotypes represent alleles that match the ancestral. On the right, the program then creates a phylogenetic tree of haplotypes attempting to root at the ancestral where possible.



## Opening Input Files

Open a specific file by either by going to the File menu > Open or by clicking on the  button at the bottom left hand corner of the screen. The program immediately recognizes .emphase or .phase files and will match it to its corresponding .snp file with the same name. Otherwise you can load the genotype data and snp information file in succession. You can also load a .many file at this point.

In the File Menu there is also an option to “Open recent” which will list the last four files you worked with.

The files that are included are then displayed on the lower left corner of the main page.



Files

Filename
CCR5_ceph.emphase
FY_ceph.emphase
HBB_ceph.emphase
HFE_ceph.emphase
LCT_ceph.emphase
G6PD_ceph.emphase
CD40L_ceph.emphase

+   -   Set Cores

## Core Selection

Setting cores is the way by which you decide what type of core feature you would like to study. Some common ways are looking at a single SNP or looking at a haplotype block as defined by (Gabriel, Schaffner et al. 2002). But there are other options as well. This section goes through how to load a core design and the different options for your core design.

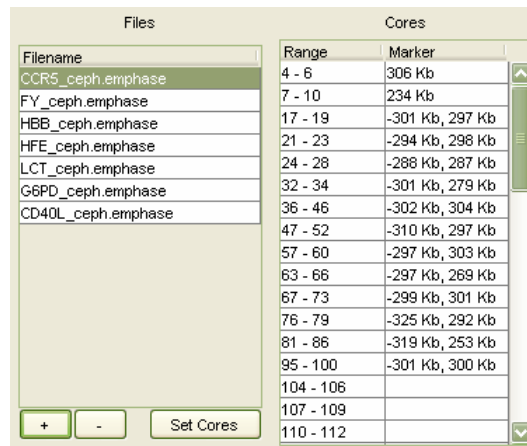
### Sections

1. Loading Cores
2. Options for Core

### Loading Cores

#### Default


When you load up a file, cores will automatically be calculated and presented. The default is to automatically find all haplotype blocks (as defined by (Gabriel, Schaffner et al. 2002)) with between 3 and 20 SNPs. Once cores are identified they are funnelled to the main analysis and presented on the main page under the section ‘Cores’.



The screenshot shows a software interface with two main panels: 'Files' and 'Cores'. The 'Files' panel contains a list of files with 'CCR5\_ceph.emphase' selected. The 'Cores' panel contains a table with columns for 'Range' and 'Marker'. Below the table are buttons for '+', '-', and 'Set Cores'.

Filename	Range	Marker
CCR5_ceph.emphase	4 - 6	306 Kb
FY_ceph.emphase	7 - 10	234 Kb
HBB_ceph.emphase	17 - 19	-301 Kb, 297 Kb
HFE_ceph.emphase	21 - 23	-294 Kb, 298 Kb
LCT_ceph.emphase	24 - 28	-288 Kb, 287 Kb
G6PD_ceph.emphase	32 - 34	-301 Kb, 279 Kb
CD40L_ceph.emphase	36 - 46	-302 Kb, 304 Kb
	47 - 52	-310 Kb, 297 Kb
	57 - 60	-297 Kb, 303 Kb
	63 - 66	-297 Kb, 269 Kb
	67 - 73	-299 Kb, 301 Kb
	76 - 79	-325 Kb, 292 Kb
	81 - 86	-319 Kb, 253 Kb
	95 - 100	-301 Kb, 300 Kb
	104 - 106	
	107 - 109	
	110 - 112	

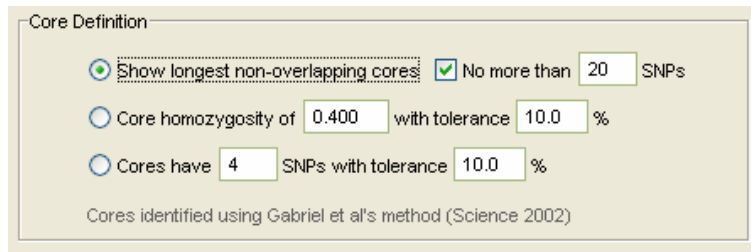
#### Making changes

You can choose a different core paradigm by clicking the  button in the bottom left side of the main page. It is preferable to set your core design before you upload your data, so that the program does not begin mining for cores under the default state.

## Options for Cores

### Core Definition

This image below is from the Set Cores page, showing the three options for defining your cores. The default settings shown below picks the longest non-overlapping cores, limiting cores to no more than 20 SNPs.



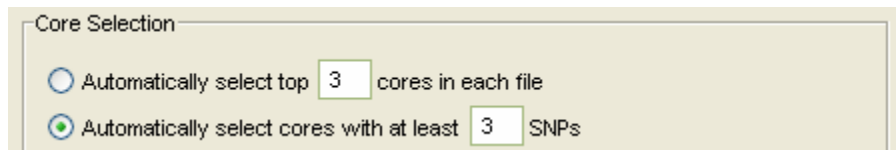
The screenshot shows a 'Core Definition' panel with three radio button options. The first option, 'Show longest non-overlapping cores', is selected and has a checked box next to it with 'No more than 20 SNPs'. The second option is 'Core homozygosity of 0.400 with tolerance 10.0 %'. The third option is 'Cores have 4 SNPs with tolerance 10.0 %'. At the bottom, it says 'Cores identified using Gabriel et al's method (Science 2002)'.

The three options for how to match cores are:

- Longest non-overlapping cores. You can give a maximum number of SNPs to have in a core. Below in the Core Selection section you can give also give a minimum number of SNPs in the block.
- Core homozygosity. You can match haplotype blocks to compare based on the amount of diversity present in the gene so that there are similar numbers and frequencies of haplotypes, with your chosen tolerance. (See definitions for exact calculation.) This option offers very close matching of regions you will compare.
- Number of SNPs. You can set an exact number of SNPs to match to with your chosen tolerance. This option allows you to have comparable information content across regions.

### Automatic Core Selection

This image below is from the Set Cores page, showing the two options for automatic core selection. The default settings shown below are including all cores with at least 3 SNPs.



The screenshot shows a 'Core Selection' panel with two radio button options. The first option is 'Automatically select top 3 cores in each file'. The second option, 'Automatically select cores with at least 3 SNPs', is selected.

The two options for automatic core selection are:

- Top cores. The program randomly includes a set of cores from each file you provide. You can set the exact number.
- All cores. You can include all cores from each file. You can give a minimum number of SNPs you use to define a core.

## Manual Core Selection

The image below shows the options for manual selection of cores. You first choose your criteria for defining a core. The program generates a list of cores that match your criteria. You can then manually select cores of interest by highlighting them and then clicking the Add menu. It will be moved to the Selected Cores section. You can then remove them by again highlighting and clicking remove. You can also restrict the range of SNPs you want presented. When you have multiple files loaded, you go through each at a time by clicking the arrow buttons on the File line.

Manually select cores

File: FY\_ceph.emphase

Range	SNPs	Core H	Haplotypes	Match Error...
36 - 46	11	0.173	13	-56.7%
67 - 73	7	0.219	10	-45.2%
47 - 52	5	0.259	7	-35.3%
81 - 86	6	0.227	9	-43.3%
95 - 100	6	0.453	7	13.3%
113 - 118	6	0.208	11	-47.9%
24 - 28	5	0.571	5	42.8%
7 - 10	4	0.565	5	41.3%

Restrict range to SNPs: 1 to 1 View D' Add >>

Selected Cores

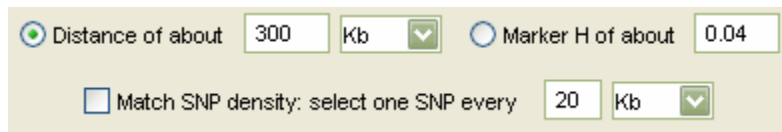
Filename	Range	Haplotypes	Core H
CCR5_ceph.emph...	27 - 32	8	0.196
CCR5_ceph.emph...	63 - 66	6	0.341

<< Remove

## Long-Distance Matching

For each core feature you will analyze, you will be calculating many EHH values at each distance from the core. To compare across many core features, however, you must choose one distance to match. (In the section EHH/Marker Correlation for all cores and EHH/Distance Correlation for all cores, you can see how you can also get a summary value for the cores).


There is a panel on the Main Page, shown below, that gives options for choosing long-distance markers to match.



The screenshot shows a control panel with three main sections. The first section has a radio button selected for "Distance of about" with a text input field containing "300" and a dropdown menu showing "Kb". The second section has a radio button for "Marker H of about" with a text input field containing "0.04". The third section has a checkbox for "Match SNP density: select one SNP every" which is unchecked, followed by a text input field containing "20" and a dropdown menu showing "Kb".

### ***Types of Distances to Match***

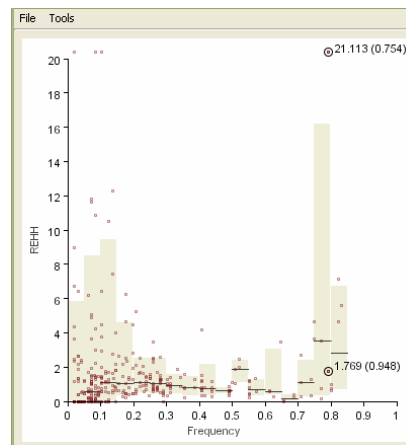
You have several options for matching distances.

- Distance (kb) You can compare across regions at a given physical distance away. For most purposes this is not a useful analysis, because it is well known that the recombination rates vary wildly among regions of the genome. However it is an interesting exploratory option and has utility when studying a specific region of the genome carefully.
- Distance (cM) You can compare across regions at given estimations of genetic distance by clicking on the  button in the panel above and switching to cM. It is ultimately most relevant to compare across the true genetic distances, and better estimates will greatly improve the power of the program. We currently the fine-scale recombination map based on the program LDHat for the HapMap (McVean et al).
- Marker H: We can also match to the observed amount of recombination in the actual data you are loading and testing, as a proxy for genetic distance. This value is the degree to which each added marker at a further distance causes the extended haplotypes to decay for all core haplotypes and can be calculated as 'all EHH' (See Definitions section). A marker H of 0.04 is roughly equivalent to a genetic distance of 0.25cM and is our preferred setting.
- Match SNP Density: This is a useful function when you have different amounts of SNP coverage in different regions of the genome. You can thin out your data to match densities by clicking to select one SNP every specified number of kb.

## Main Page Images

### *EHH/REHH vs. Frequency Scatter Plot*

The scatter plot below gives REHH plotted against haplotype frequency for every core haplotype in your data files, given at a particular long-range distance  $x$  that you designate. If you click on the Y-axis, you can toggle between the view of the EHH and REHH. If you click on the red dots in the scatter plot, it will display the EHH/REHH value for that core haplotype as well for the same core haplotype in the other direction. The other figures on the main page will then display the currently selected haplotype.

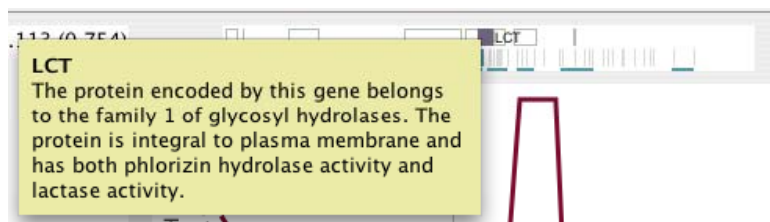


### *Gene Track*

The image below shows the genomic region you have selected to view. The blue-outlined boxes represent the genes in the region from the refgene database. The vertical lines below the genes represent the SNPs in your datafile. SNPs in blue are in the haplotype block you have currently selected. The horizontal blue lines at the bottom represent the haplotype blocks identified in your file.

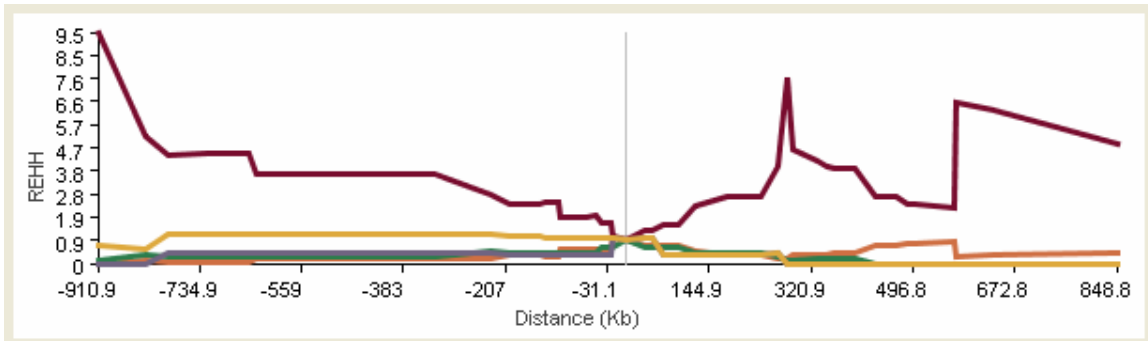


If you put your cursor over the genes, it will display the name of the gene. If you click on a particular gene it will display summary information about the gene.



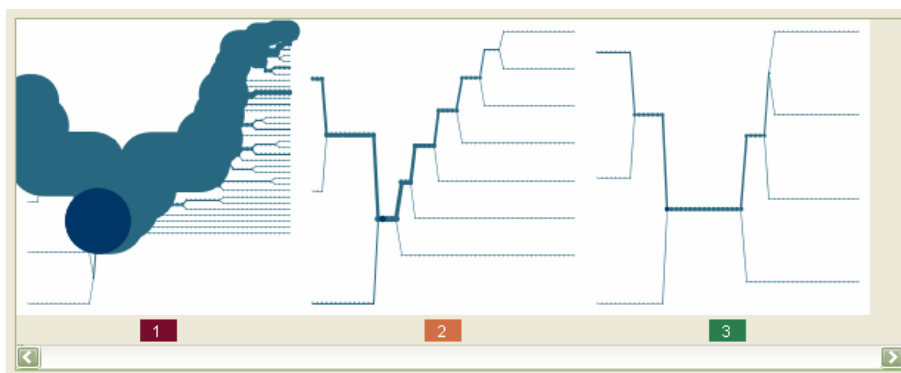
## ***EHH/REHH vs. Distance Chart***

The chart below gives REHH plotted for the selected core haplotype at every long-range distance in both directions from the core region. The different haplotypes are shown together in the plot with the color matching the ‘haplotype table’ view below. If you click on the Y-axis, you can toggle between the view of the EHH and REHH.



## ***Haplotype Bifurcation Diagrams***

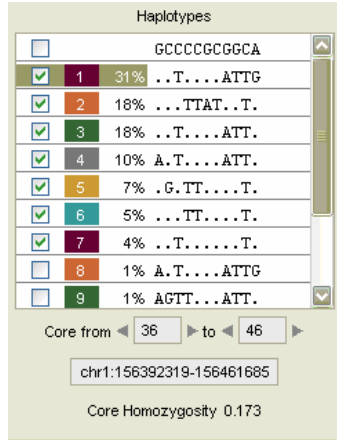
The figure below shows that Haplotype Bifurcation diagram, which visualizes the breakdown of LD at increasing distances from core haplotypes at the selected core region. The root of each diagram is a core haplotype, identified by a dark blue circle. The diagram is bi-directional, portraying both centromere-proximal and centromere-distal LD. Moving in one direction, each marker is an opportunity for a node; the diagram either divides or not based on whether both or only one allele is present. Thus the breakdown of LD on the core haplotype background is portrayed at progressively longer distances. The thickness of the lines corresponds to the number of samples with the indicated long-distance haplotype.



## ***Haplotype Table***

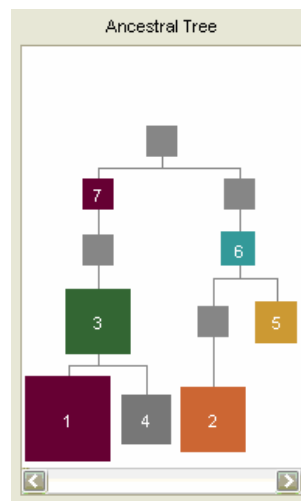
The figure below gives the list of haplotypes at the specified core region. Where ancestral information is available, Sweep will display the predicted ancestral haplotypes above the list of haplotypes observed in your data. Each haplotype in your data is

presented with its sequence and identifier matching the other figures on the main page. The ‘.’s in the observed haplotype sequence represent alleles that match the ancestral. The other allele is given by its nucleotide letter. If there is no ancestral given, both alleles are displayed in gray.



## Ancestral Tree

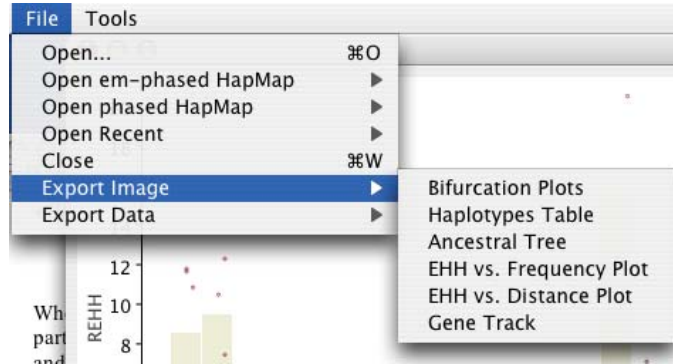
The figure below gives a phylogenetic tree of the haplotypes at your core region. The program attempts to root the tree if the ancestral haplotype is known. Haplotypes closer to the ancestral are at the top of the figure and those many mutational steps away at the bottom. Gray squares represent haplotypes that are not present in your data, but are missing links in the phylogeny. The area of the squares is proportional to the frequency of the haplotype. The program can only create a phylogeny when there are no recombinant haplotypes selected in the ‘haplotype table’. If the program suspects a recombinant it will ask you to deselect potential recombinants. If it can not determine the root of the tree it will report ‘root is ambiguous’.





## Exporting Images

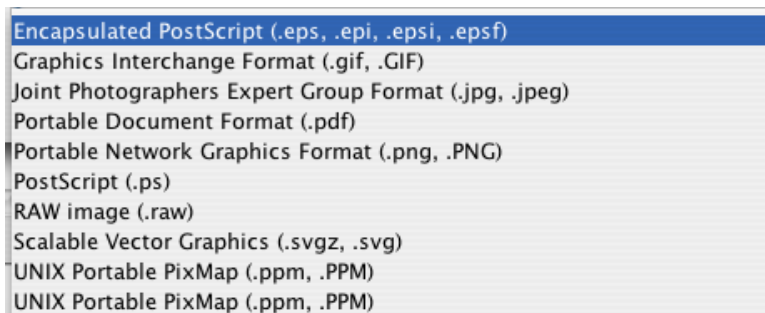
Each of the images on the Main Page can be exported to several file formats including eps, gif, pdf, png, and ps. In the File menu, go to Export Image to display the different images to export. This section reviews the different images for export, and the export procedure.



### Exporting Procedure

The images on the Main Page are exported just as you see them on the screen. So be sure to have the image set just as you want it, including haplotypes of interest, having it on the correct settings etc. The EHH vs. Frequency Plot and EHH vs. Distance Plot will export either the EHH or REHH values, depending on what you have showing on the screen at that point in time.

Once the image looks as you would like on the screen, you can go to the File > Export Images option and click on one of the 6 file-types you would like to export. You will be taken to a menu where you can browse for a file destination and name your file, and you can choose from one of many file type options shown below.



## Exporting Data

You can export your standard analysis in a variety of ways by going to the File Menu > Export Data. The export options are listed below. Each one will be explained in turn.

EHH vs. Frequency Data
EHH/MarkerH Correlation Data for all cores
EHH/Distance Correlation Data for all cores
EHH Data for this Core
List of cores
SNP Frequency Data

### ***EHH vs. Frequency Data***

This exports a series of data points for each core haplotype, in both the centromere-distal and centromere-proximal direction (denoted by the “Dis from Core” being minus or plus). The data points are for the long-distance marker to match as chosen on the main page. The different output columns are explained below.

Header	Explanation
Core Source	Source file from which the haplotype was loaded
Chrom	chromosome
Start SNP	SNP number in the source file for the haplotype start point
End SNP	SNP number in the source file for the haplotype end point
Start Base	chromosomal position for the haplotype start point
End Base	chromosomal position for the haplotype end point
Core H	diversity at the core (explained in Definitions section)
Genes In Region	genes within the window of the haplotype +/- 20kb
Marker #	SNP number in the source file for the extended marker to test
Dist From Core	distance from the core of the the extended marker to test
All EHH	Observed historical recombination (explained in the Definitions section)
Haplotype #	identifier number for the haplotype in the core
Sequence	the allele for each SNP in that haplotype
Hap Freq	frequency of the haplotype
EHH	the EHH for the given core haplotype at the given distance away (explained in the Definitions section)
REHH	the REHH for the given core haplotype at the given distance away (explained in the Definitions section)
Avg EHH	the average EHH for every marker up to the given distance away
Avg REHH	the average REHH for every marker up to the given distance away
Max EHH	the maximum EHH for all markers up to the given distance away
Max REHH	the maximum REHH for all markers up to the given distance away
EHH Percentile	the percentile value for the haplotype's EHH when compared to all other haplotypes in the same 5% frequency bin
REHH Percentile	the percentile value for the haplotype's REHH when compared to all other haplotypes in the same 5% frequency bin
AREHH	a normalized (adjusted) REHH value
AREHH Deviation	the number of standard deviations the REHH value is from other haplotypes in the same 5% frequency bin
Percent Ancestral	the fraction of the chromosomes for the haplotype carrying the putative ancestral long-range haplotype

### ***Distance for fixed EHH for all cores***

This is the exact distance, given your data, in kb and cm (from the fine-scale recombination map) that a core haplotype drops to a specified EHH value. So instead of asking at a given genetic distance, what is the EHH, one can ask given a EHH, what is the genetic distance the haplotype extends. The output columns are explained below.

Header	Explanation
Filename	Source file from which the haplotype was loaded
Chrom	chromosome number
Core Start	SNP number in the source file for the haplotype start point
Core End	SNP number in the source file for the haplotype end point
Start Base	chromosomal position for the haplotype start point
End Base	chromosomal position for the haplotype end point
Genes In Region	genes within the window of the haplotype +/- 20kb)
Haplotype #	identifier number for the haplotype in the core
Sequence	the allele for each SNP in that haplotype
Hap Freq	frequency of the haplotype
EHH to Match	EHH you choose as your decay point to look for
Distance (bases) at said EHH, minus dir	Physical distance in the minus direction the haplotype extends
Distance (cM) at said EHH, minus dir	Genetic distance in the minus direction the haplotype extends
Distance (bases) at said EHH, plus dir	Physical distance in the plus direction the haplotype extends
Distance (cM) at said EHH, plus dir	Genetic distance in the plus direction the haplotype extends

### ***EHH/MarkerH Correlation Data for all cores***

The EHH/MarkerH Correlation gives the genetic distance at which the EHH score falls to a chosen EHH value based on an interpretation of the slope of the decay. The analysis first plots the EHH at increasing genetic distance (measured by observed historical recombination distance) away from the core region. There is high correlation between EHH and genetic distance, giving an associated EHH degradation rate in each direction. Using the degradation rate we then estimate the extended haplotype length (EHL) the genetic distance at which the EHH degrades to a specified value. The output columns are explained below.

Header	Explanation
Filename	Source file from which the haplotype was loaded
Core Start	SNP number in the source file for the haplotype start point
Core End	SNP number in the source file for the haplotype end point
Haplotype #	identifier number for the haplotype in the core
Sequence	the allele for each SNP in that haplotype
Hap Freq	frequency of the haplotype
Direction	minus or plus direction for chromosomal position of extended SNPs
# of markers used	total number of markers to achieve given EHH
$EHH = a \cdot \log(\text{MarkerH}) + b$ : a	the equation for the line representing decay for EHH over distance - value for "a"
$EHH = a \cdot \log(\text{MarkerH}) + b$ : b	the equation for the line representing decay for EHH over distance - value for "b"
R <sup>2</sup> for fit	correlation of each EHH to line
EHH to Match	EHH you choose as your decay point to look for
MarkerH at said EHH	the genetic distance at which EHH decays to your chosen value based on observed historical recombination

### ***EHH/Distance Correlation Data for all cores***

Like the EHH/MarkerH Correlation the EHH/Distance Correlation gives the physical distance at which the EHH score falls to a chosen EHH value. The analysis first plots the EHH at increasing physical distance away from the core region. There is high correlation between EHH and physical distance within the same region of the genome (since local rate of recombination may not vary as greatly), giving an associated EHH degradation rate in each direction. Using the degradation rate we then estimate the extended haplotype

length (EHL) the genetic distance at which the EHH degrades to a specified value. It gives the same output table as EHH/MarkerH Correlation except that the last row is Distance at said EHH.

## ***EHH Data for this core***

This table gives information about EHH/REHH at every distance away from a chosen core. It is the data that underlies the chart on the top right of the Main page. The output columns are given below:

<b>Header</b>	<b>Explanation</b>
core start	SNP number in the source file for the haplotype start point
core end	SNP number in the source file for the haplotype end point
haplo index	identifier number for the haplotype in the core
haplotype	the allele for each SNP in that haplotype
core H	diversity at the core (explained in Definitions section)
haplo freq	frequency of the haplotype
marker index	SNP number in the source file for the extended marker to test
marker H	Observed historical recombination (explained in the Definitions section)
marker EHH	the EHH for the given core haplotype at the given distance away (explained in the Definitions section)
marker not-EHH	the not-EHH for the given core haplotype at the given distance away (explained in the Definitions section)
marker REHH	the REHH for the given core haplotype at the given distance away (explained in the Definitions section)
marker freq(allele 1)	frequency of the extended marker to test
marker dist	distance from the core of the the extended marker to test

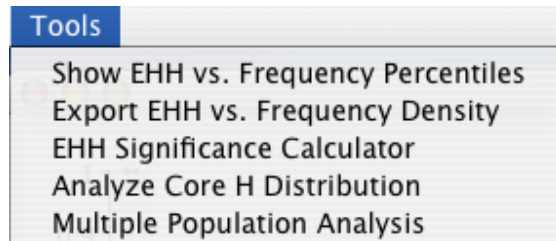
## ***SNP Frequency***

This gives some basic information about the SNPs in your file: allele frequencies, derived allele frequencies, and heterozygosity. It will only give this data for files you highlight in the bottom left corner of the Main page, so highlight all files you want data for. The output columns are explained below.

<b>Header</b>	<b>Explanation</b>
snpid	SNP name (rs number or your given name)
chr	chromosome
hg16	chromosomal position (either HG16 or HG17 named accordingly)
filename	Source file from which the SNP was loaded
major allele	common allele (1=A, 2=C, 3=G, 4=T)
minor allele	rare allele
major frequency	frequency of common allele
minor frequency	frequency of rare allele
ancestral allele	ancestral allele from ancestral.tab or chimp allele for HapMap (if available)
derived allele	derived allele form ancestral.tab or chimp allele for HapMap
derived frequency	frequency of derived allele
heterozygosity	SNP heterozygosity (Nei)

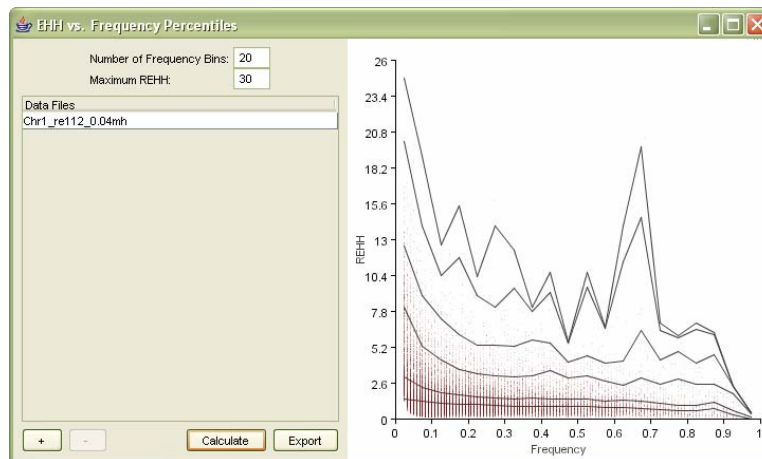
## Tools

In addition to the export data currently uploaded, you can carry out overview analysis of sets of data. The export options are listed below. Each one will be explained in turn.



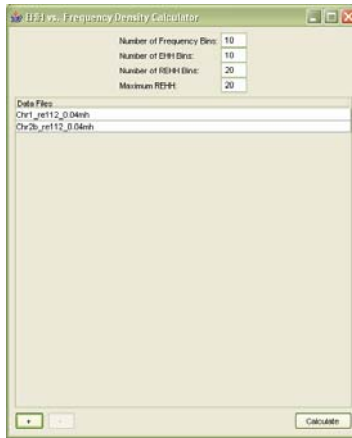
### **Show EHH vs. Frequency Percentiles**

In the percentiles feature, you can load up any number of EHH vs. Frequency data files, and calculate REHH percentiles for different frequency bins. The default is 20 bins created ranges of 0-5%, 5-10% etc... but you can change the number of bins. You can also change the maximum REHH displayed on the Y axis. Then press 'calculate' and the program will calculate the 50<sup>th</sup>, 75<sup>th</sup>, 90<sup>th</sup>, 95<sup>th</sup>, 99<sup>th</sup>, and 99.9<sup>th</sup> percentiles for each frequency bin and display them. You can export the image created using the 'export' button.



### **Export EHH vs. Frequency Density**

The Density feature, gives the number of core haplotypes in each EHH vs. Frequency bin. You load up every file you would like to view at once, and specify the number of bins you would like to make for EHH, REHH, and for frequency.



The feature gives back a table with the total number  $N$  of data points in your input data file. It then bins everything by its corresponding EHH and frequency class and also by its corresponding REHH and frequency class. It gives back two tables, one for EHH and one for REHH, with the number of haplotypes in each bin. An output table is shown below where  $n$  is the number of haplotypes in each bin.

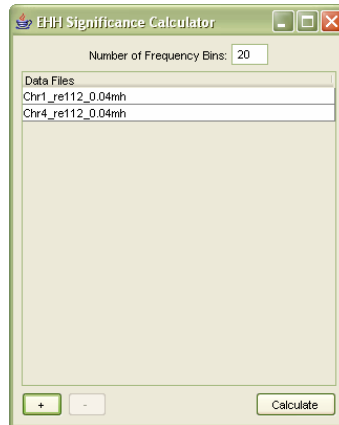
EHH Density ( $N$ data points).		Frequency across, EHH down								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.1	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.2	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.3	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.4	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.5	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.6	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.7	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.8	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
0.9	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$

REHH Density ( $N$ data points).		Frequency across, EHH down								
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
1	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
2	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
3	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
4	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
5	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
6	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
7	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
8	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
9	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
10	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
11	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
12	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
13	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
14	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
15	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
16	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
17	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
18	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
19	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$
Above 20.0	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$	$n$

Input data files  
test

## ***EHH Significance Calculator***

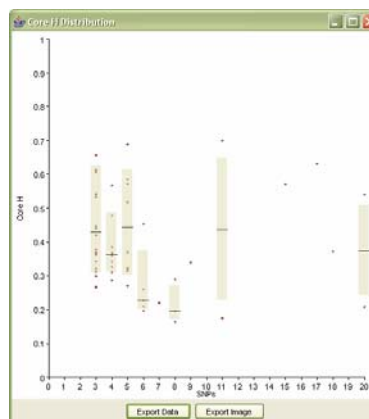
To calculate the significance of EHH or REHH value, you can load up as many “EHH vs. Frequency” data files as you want to compare in the EHH significance calculator. All the haplotypes in the loaded files are then placed into bins based on their frequency. The default is 20 bins creating ranges of 0-5%, 5-10% etc... but you can change the number of bins. P-values are obtained by log-transforming the EHH and REHH in the bin to achieve normality, and calculating the mean and standard deviation.



The output is a concatenated EHH vs. Frequency file with 2 extra columns at the end giving the number of standard deviations that the EHH and REHH for each haplotype are away from the mean in their bin. These can be used to calculate P-values by using  $1 - \text{NORMSDIST}(\text{value})$  in excel.

## ***Analyze Core H Distribution***

You can calculate the core homozygosity (Definitions section) for all the core regions you have selected in the selected core regions. It will display them all grouped by the number of SNPs in the core. Below is the default of cores from 3 to 20 SNPs. You can change the range of SNP numbers in your core on the 'set core' window (See Core Selection section).



## Multiple Population Analysis

You can do some basic comparisons of populations using the Multiple Population Analysis window. It requires data files for multiple populations using the same SNP information file. You first load up the SNP file that matches with all loaded data files. You then load up the data files one by one for each population to compare.



You can either do population analysis for each individual SNP in the file or for all overlapping haplotype blocks in your file by clicking on one of the two buttons at the bottom of the setup window. The program calculates FST, P-excess and contingency chi-squared for every pair of populations, taking as the comparison entity markers or cores common to both populations. Note P-excess is not calculated for cores, since there's no way of taking one population as "ancestral"; biallelic SNPs don't have this problem).

The program will output a table with the data columns presented and explained below.

CCR5Chi71.emphase<->CCR5Yor71.emphase	
Header	Explanation
Core Start	SNP number in the source file for the haplotype start point
Core End	SNP number in the source file for the haplotype end point
CoreH CCR5Chi71.emphase	diversity at the core (explained in Definitions section)
CoreH CCR5Yor71.emphase	diversity at the core (explained in Definitions section)
FST	population differentiation measure (to be explained in Definitions section)
Contingency Chi-squared	differentiation of allele frequencies in each population (to be explained in Definitions section)
X <sup>2</sup> p-value	p-value for chi-square (to be explained in Definitions section)



## Command Lines

To use the command lines, you must go onto unix and cd into the directory where the program lies eg. Sweep-148.

### ***EHH vs Frequency Data***

The most common use of Sweep is to export all haplotype, frequency, EHH, REHH, ancestral data for a large set of regions at a particular matched distance (see the EHH vs Frequency Data section in Exporting Data). Since the task may be needed in bulk, we have created a command line that allows you to specify properties of your core and the long distance region to match.

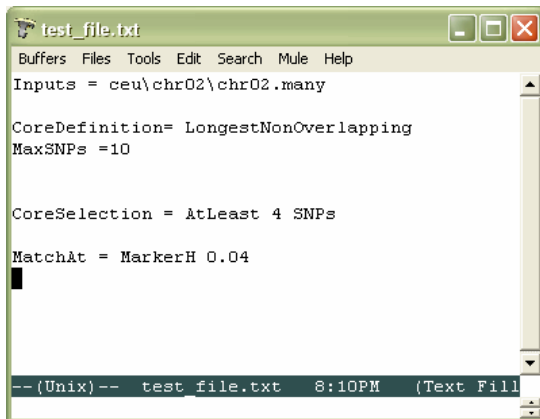
The command line for PC is:

```
./run.bat ExportEHHvsFreqData <input file> <output file>
```

The command line for Mac is:

```
Sweep-.app/run ExportEHHvsFreqData <input file> <output file>
```

The input file is a text file that lists all your parameters. Here is an example input file:



```
test_file.txt
Buffers Files Tools Edit Search Mule Help
Inputs = ceu\chr02\chr02.many

CoreDefinition= LongestNonOverlapping
MaxSNPs =10

CoreSelection = AtLeast 4 SNPs
MatchAt = MarkerH 0.04

-- (Unix)-- test_file.txt 8:10PM (Text File)
```

Inputs: a list of paths to your data files you want to include in the analysis. They can be .phase, .wphase, .emphase or .many files.

CoreDefinition: "LongestNonOverlapping", "ByCoreHomozygosity" or "BySNPCount"  
[default: LongestNonOverlapping]

MaxSNPs: Max # of SNPs in a core *if Core Definition is LongestNonOverlapping*  
[default:20]

CoreH: Value of Core H to match to *if CoreDefinition=ByCoreHomozygosity*  
[default: 0.400]

NumSNPs: Core length to match to *if CoreDefinition=BySNPCount*  
[default: 4]

Tolerance: Percent difference between a core parameter (coreH/SNP count) and specified value for core to be included *if CoreDefinition = ByCoreHomozygosity or BySNPCount*  
[default: 10%]

CoreSelection: "Top <N> Cores" or "AtLeast <N> SNPs"  
[default: AtLeast 3 SNPs]

MatchAt: Long distance criteria to match across regions: "Distance <N> bases/Kb/Mb/cM" or "MarkerH <x>"  
[no default]

SelectOneSNPEvery: SNP Density: "<N> bases/Kb/Mb"  
[default: don't do density matching]

## ***EHH significance***

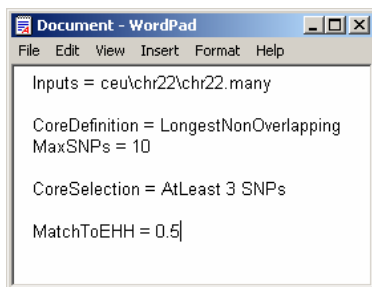
A second important analysis that is often done in bulk is calculating the significance of EHH and REHH for a group of files together (see the EHH Significance in Tools). With this command line you can take in up to 7 EHH vs Frequency Data files, bin haplotypes by frequency, and calculate significance of EHH and REHH.

The command line for PC is:

```
./run.bat CalculateEHHSignificance <number of bins> <input file1> <input file2>  
<input file 3 etc...7> <output file>
```

## ***Distance to EHH***

```
./run.bat EHHCorrelation <input file> <output file>
```



## References

- Gabriel, S. B., S. F. Schaffner, et al. (2002). "The Structure of Haplotype Blocks in the Human Genome." Science **23**: 2225-2229.
- Sabeti, P. C., D. E. Reich, et al. (2002). "Detecting recent positive selection in the human genome from haplotype structure." Nature **419**(6909): 832-7.
- Stephens, M. and P. Donnelly (2003). "A comparison of bayesian methods for haplotype reconstruction from population genotype data." Am J Hum Genet **73**(5): 1162-9.
- Stephens, M., N. J. Smith, et al. (2001). "A new statistical method for haplotype reconstruction from population data." Am J Hum Genet **68**(4): 978-89.