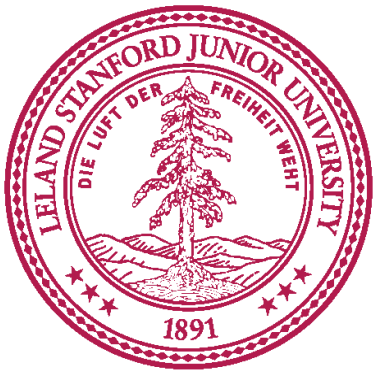
HAPAA Manual



HAPAA is ancestry inference software package that is capable of determining allele specific origin in individual's genome based in dense SNP micro array data. This document represents short introduction to the components of this software package. It is not meant to be a complete description of the software capabilities but should be a good starting point to familiarize yourself with its major part and typical uses.

For more information please contact us using contract information in hapaa.stanford.edu

Binary data file types

Extension	Description
.snpprofile.bin	SNPProfile file
.snpdata.bin	SNPData file
.snpprofiledata.bin	File containing concatenation of a SNPProfile and a SNPData
.diplopop.bin	Set of simulated diploid individuals created from template individuals from a SNPData
.diplopopdata.bin	Ancestral assignments at each position for a collection of diploid individuals
.phases.bin	Phasing information passed between PhaseIncremental and RunHaploModel
.phaseerrors.bin	Phasing information passed between PhaseIncremental and RunHaploModel
.diploposterior.bin	Ancestral posterior probabilities at each position for a collection of diploid individuals

Profile is the information about SNP locations to be analyzed. It lists the labels as well as genetic and genomic (base pair) distance between the adjacent SNP's. Data contains actual SNP values observed.

Program reference

CreateHapMapSNPData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*

Note: Requires 4 GB of RAM to run.

Before running this command you need to make sure that you have the necessary HapMap files in the input_dir directory. If you would like to download relevant HapMap files, you need to go to a directory Data. And execute the following command: "g++ DownloadHapMap.cpp –o DownloadHapMap; ./DownloadHapMap". This will start the download process. For this process to execute you should have 'wget' command installed. If you do not and/or do not wish to use it, the names of all the files to be downloaded are listed in DownloadHapMap.txt file, so that you can acquire them using any other means

Example:

```
CreateHapMapSNPData work_dir=hapmap_dir
```

RestrictIlluminaSites

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input "snpprofiledata.bin" filename
output_file	Output "snpprofiledata.bin" filename
illumina_file	Input text file listing Illumina sites *(default: Hap550v3_561466_SNPs)*

The text files used for "illumina_file" parameter can be downloaded and unzipped from:
http://www.illumina.com/downloads/ILMN_HumanHap550_SNPList.zip
http://www.illumina.com/downloads/ILMN_HumanHap650Y_SNPlist.zip

Example:

```
RestrictIlluminaSites work_dir=hapmap_dir
   input_file=hapmap_chr21.snpprofiledata.bin.gz
   output_file=hapmap550_chr21.snpprofiledata.bin.gz
   illumina_file=Hap550v3_561466_SNPs
```

CreateSNPProfile

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input text filename
output_file	Output "snpprofile.bin" filename

Input text file format:

```
;Lines starting with ";" are ignored. Space/tab is a field delimiter
;chromosome genomic_pos genetic_pos SNP_label(optional)
chr1        1000000     0.0         rs1234
chr1        1001000     0.1         rs1235
chr1        1002000     0.2         rs1236
chr2        5000        0.0         rs2000
chr2        5500        0.001       rs3000
chr2        5900        0.001       rs2900
```

Example:

```
CreateSNPProfile input_file=test_profile.txt
   output_file=test_profile.snpprofile.bin
```

ImportSNPData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input text filename
snpprofile_file	SNPProfile to use for the data, can be either "snpprofile.bin" or "snpprofiledata.bin" file
output_file	Output "snpprofile.bin" filename
switch_error	Haplotype switch error rate *(default 0.5, meaning unphased genotypes)*

Input text file format:

```
;Lines starting with ";" are ignored. Space/tab is a field delimiter
labels      data_label1 data_label2 data_label3 data_label4 (optional row)
pops        pop_label1  pop_label1  pop_label2  pop_label2
rs1234      AC          A-          CC          -
rs1235      AA          AA          TT          T
rs1236      GC          GC          GC          G
rs2000      TT          TA          CT          C
chr2.5500   AT          AA          AC          C
chr2.5900   CT          --          TT          C
```

Example:

```
ImportSNPData snpprofile_file=test_profile.snpprofile.bin
   input_file=test_data.txt output_file=test_data.snpprofiledata.bin.gz
```

ExportSNPData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input data, can be either "snpprofiledata.bin" or "snpdata.bin" file. If "snpdata.bin", then snpprofile_file must also be specified.
snpprofile_file	SNPProfile to use for the data, can be either "snpprofile.bin" or "snpprofiledata.bin" file. Should only be specified if input_file type is "snpdata.bin".
output_file	Output text filename

Example:

```
ExportSNPData input_file=test_data.snpprofiledata.bin.gz
   output_file=exported.txt
```

MergeSNPData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
snpprofile_file	SNPProfile to use for the data, can be either "snpprofile.bin" or "snpprofiledata.bin" file. Should only be used if all the input files are of type "snpdata.bin".
output_file	Output "snpprofiledata" or "snpdata.bin" filename
input_files	Quotes-enclosed, comma-separated list of "snpprofiledata.bin" or "snpdata.bin" files. All data files must use an identical profile.
concat / merge_pop_index / merge_pop_label	Select merge mode. "concat" will concatenate all the data without merging populations. "merge_pop_index" will merge the data while keeping the population index numbers the same. "merge_pop_label" will merge the data while keeping populations with the same name the same.
pop_labels	Quotes-enclosed, comma-separated list of new population names, ordered by index. This is used to relabel the populations. *(optional)*

Example:

```
MergeSNPData input_files="data1.snpprofiledata.bin,data2.snpprofiledata.bin"
   merge_pop_label output_file=combined_data.snpprofiledata.bin
```

RestrictSNPData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
Either:	
input_profile_file	Input "snpprofile.bin" or "snpprofiledata.bin" file
output_profile_file	Output "snpprofile.bin" file
Or:	
input_file	Input "snpprofiledata.bin" or "snpdata.bin" file. If it's a "snpdata.bin" file, also include the parameter "input_profile_file".
output_file	Output "snpprofiledata.bin" or "snpdata.bin" file
More paremeters:	
loci	SNP loci to restrict to, using the integer list spec described below. *(optional)*
chroms	Quotes-enclosed, comma-separated list of chromosome indices to restrict to *(optional)*
pops	Quotes-enclosed, comma-separated list of populations indices to restrict to *(optional)*
pop_labels	Quotes-enclosed, comma-separated list of population names to restrict to *(optional)*
diplos	Diploid individual indices to restrict to, using spec *(optional)*
haplos	Haplotypes indices to restrict to, using spec *(optional)*
pop_diplos	Diploid individual indices within each population to restrict to, using spec *(optional)*
pop_haplos	Haplotype indices within each population to restrict to, using spec *(optional)*
gender	Quotes-enclosed, comma-separated list of genders (0=female, 1=male) to restrict to *(optional)*
map_pops	Quotes-enclosed, comma-separated list of population index remappings. Repeat indices imply a merging of populations. For example, (1,1,0) would remap populations so that populations 0 and 1 become population 1, and population 2 becomes population 0. *(optional)*
restrict_pop_missing	Remove SNP loci for which at least one population is completely unascertained *(optional)*
restrict_monomorphic	Remove SNP loci with only one possible allele *(optional)*

Integer list specification:

List=Elem[,Elem[,Elem…]]
Elem=Single or Range or Count
Single=<integer> ; A single number
Range=<integer>-<integer> ; All integers between the two extremes
Count=(<integer>,<integer>) ; (x, y) indicates integers x, x+1, x+2, x+y-1

Example:

```
RestrictSNPData input_file=combined_data.snpprofiledata.bin
   output_file=pop02.snpprofiledata.bin pops="0,2"
```

SNPInfo

Parameter name	Description
input_dir / work_dir	Input file directory *(optional)*
input_file	Input “snpprofiledata.bin” or “snpprofile.bin” file
show_loci	Output all information of SNP loci as well

Example:

```
SNPInfo input_dir=hapmap_dir input_file=hapmap.snpprofiledata.bin.gz
```

CreateAdmixture

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input "snpprofiledata.bin" file used as ancestral template individuals
output_file	Output "diplopop.bin" file containing all simulated admixed individuals
select	Integer list spec of diploid individuals from each population to use *(optional)*
min_gen	Minimum number of generations of admixture, minimum of 1 *(default: 1)*
max_gen	Maximum number of generations of admixture *(default:20)*
gen_samples	Number of individuals to construct per generation *(default:100)*
gen_samples_list	Quotes-enclosed, comma-separated list of number of individuals to construct per generation, starting with "min_gen" *(optional)*
pop_prob	Quotes-enclosed, comma-separated list of probability of ancestral origin from each population *(default: equal probability for each population)*
force_admixture	Only produce example individuals that have at least 2 populations of origin in their genome *(optional)*
do_gender	Only mate individuals of opposite gender

Example:

```
CreateAdmixture input_file=hapmap550_chr21.snpprofiledata.bin.gz
   output_file=examples.diplopop.bin min_gen=1 max_gen=10 gen_samples=10
   do_gender
```

SimulatePopToSNPData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input "diplopop.bin" file containing admixed individuals
snpprofiledata_file	Input "snpprofiledata.bin" file used as ancestral template individuals
output_file	Output "snpprofiledata.bin" or "snpdata.bin" file
select	Integer list spec of admixed individuals to output
switch_error	Haplotype switch error rate *(default 0.0, meaning perfect haplotypes)*

Example:

```
SimulatePopToSNPData
   snpprofiledata_file=hapmap550_chr21.snpprofiledata.bin.gz
   input_file=tests.diplopop.bin output_file=tests.snpprofiledata.bin
```

SimulatePopToPopData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input "diplopop.bin" file containing admixed individuals
snpprofiledata_file	Input "snpprofiledata.bin" file used as ancestral template individuals
output_file	Output "diplopopdata.bin" or "haplopopdata.bin" file
select	Integer list spec of admixed individuals to output

Outputs the true population ancestry of the admixed individuals.

Example:

```
SimulatePopToPopData
   snpprofiledata_file=hapmap550_chr21.snpprofiledata.bin.gz
   input_file=tests.diplopop.bin output_file=tests.diplopopdata.bin
```

TrainHaploModel

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
model_file	Model parameters "haplomodel.bin" file
default_model	Set initial parameters to a default model
sample_file	Model individuals "snpprofiledata.bin" file
simulate_training_file	Simulated "diplopop.bin" admixed individuals to use for training
diplos	Integer list spec of training examples to use *(optional)*

Note that the model parameters are saved every iteration, so it's okay to interrupt training (ctrl+c) at any time you feel it has converged sufficiently.

Example:

```
TrainHaploModel sample_file=hapmap550_chr21.snpprofiledata.bin.gz
   model_file=model.haplomodel.bin
   simulate_training_file=examples.diplopop.bin
```

PhaseIncremental

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
sample_file	Model phased individuals "snpprofiledata.bin" file
input_file	Unphased target individuals "snpprofiledata.bin" file
output_file	Output phased "snpprofiledata.bin" file, or output "phases.bin" file, for use as input to RunHaploModel
phaseerrors_file	Output "phaseerrors.bin" file, for use as input to RunHaploModel *(optional)*
switch_error_rate	Switch error rate to set for output *(default 0.1)*

Example:

```
PhaseIncremental sample_file=hapmap550_chr21.snpprofiledata.bin.gz
   input_file=tests.snpprofiledata.bin switch_error_rate=0.05
   output_file=tests.phases.bin phaseerrors_file=tests.phaseerrors.bin
```

RunHaploModel

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
model_file	Model parameters "haplomodel.bin" file
default_model	Use a default set of model parameters
sample_file	Model individuals "snpprofiledata.bin" file
testing_file	Test individuals "snpprofiledata.bin" file
phases_file	Phasing information "phases.bin" file from PhaseIncremental *(optional)*
phaseerrors_file	Phasing information "phaseerrors.bin" file from PhaseIncremental *(optional)*
do_haplos	Run inference on individual haplotypes instead of diploid individuals
diplos	Integer list spec of test individuals to use *(optional)*
posterior_file	Output population posterior probabilities "diploposterior.bin" or "haploposterior.bin" file *(optional)*
posterior_popdata_file	Output decoded inferred populations "diplopopdata.bin" or "haplopopdata.bin" *(optional)*
viterbi_popdata_file	Output most-likely sequence decoded populations "diplopopdata.bin" or "haplopopdata.bin" *(optional)*
min_block_len	Minimum block length (in Morgans) for filtering *(optional)*

Example:

```
RunHaploModel sample_file=hapmap550_chr21.snpprofiledata.bin.gz
   model_file=model.haplomodel.bin testing_file=tests.snpprofiledata.bin
   posterior_file=tests_hapaa.diploposterior.bin
   posterior_popdata_file=tests_hapaa.diplopopdata.bin
```

ScoreHaploModelMSE

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
snpprofile_file	Input "snpprofile.bin" or "snpprofiledata.bin" file
true_file	True ancestry "haplopopdata.bin" or "diplopopdata.bin" file
guess_file	Inferred "haplopopdata.bin" or "diplopopdata.bin" file

Example:

```
ScoreHaploModelMSE snpprofile_file=hapmap550_chr21.snpprofiledata.bin.gz
   true_file=tests.diplopopdata.bin guess_file=tests_hapaa.diplopopdata.bin
```

ExportPopData

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_file	Input data, can be either “haplopopdata.bin” or “diplopopdata.bin” file.
snpprofile_file	SNPProfile to use for the data, can be either “snpprofile.bin” or “snpprofiledata.bin” file.
output_file	Output text filename

Example:

```
ExportPopData snpprofile_file=hapmap550_chr21.snpprofiledata.bin.gz
   input_file=tests_hapaa.diplopopdata.bin
   output_file=tests_hapaa.diplopopdata.txt
```

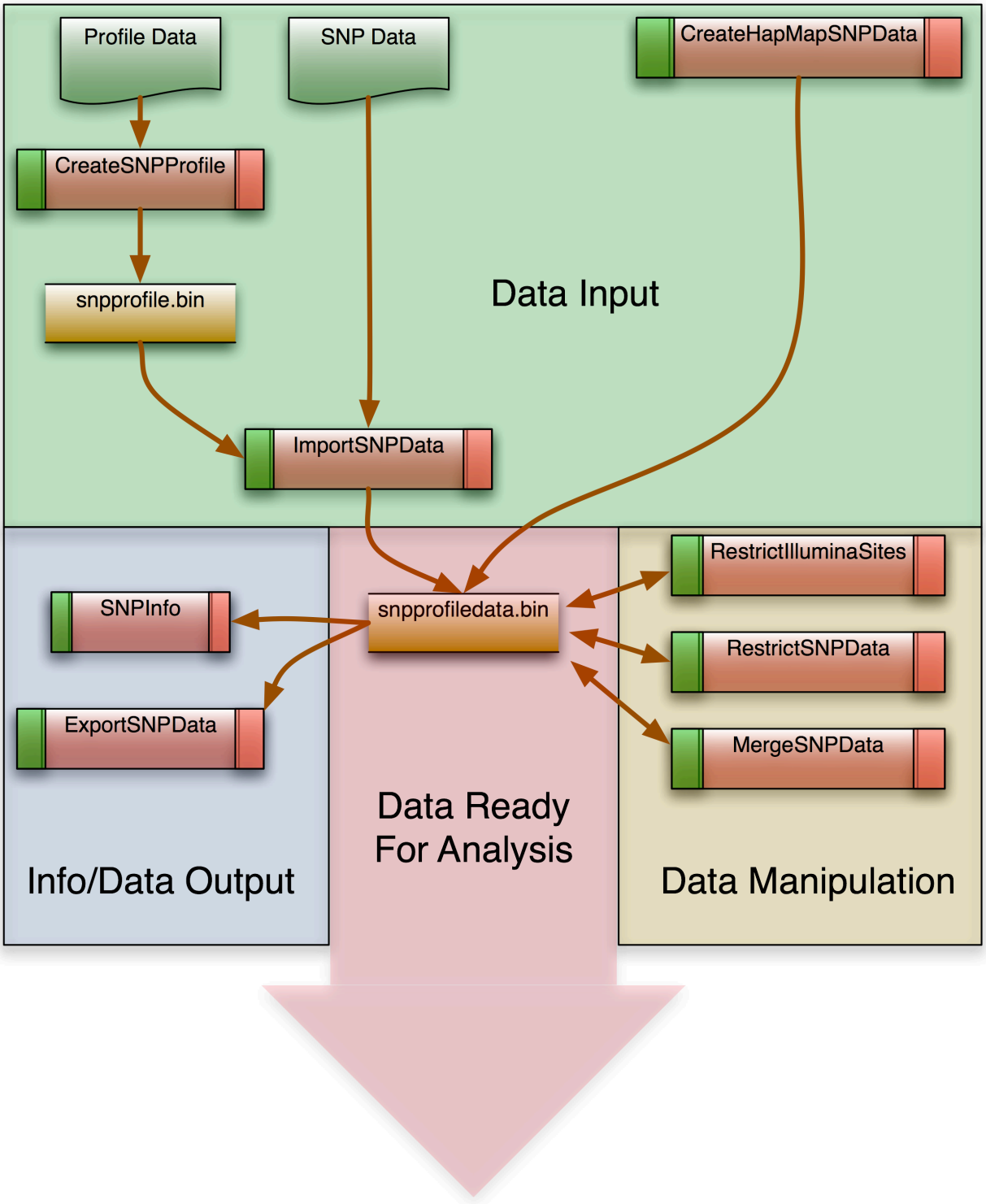
GraphHaploModel

Parameter name	Description
input_dir	Input file directory *(optional)*
output_dir	Output file directory *(optional)*
work_dir	Set both input and output file directories *(optional)*
input_files	Quotes-enclosed, comma-separated list of files to graph. Each one must be one of the types “haplopopdata.bin”, “diplopopdata.bin”, “haploposterior.bin”, “diploposterior.bin”, “snpprofile.bin”, or “snpprofiledata.bin”. One of the files must be a “snpprofile.bin” or “snpprofiledata.bin” file, and only the first one will be used to establish the SNPProfile used for the data. Nothing will be graphed for a “snpprofile.bin” or “snpprofiledata.bin” file.
output_file	Output image file with extension”png”.
image_size_x	Output image x resolution *(default: 1024)*
image_size_y	Output image y resolution *(default: 768)*
select_graphs	Integer list spec of which individuals to graph

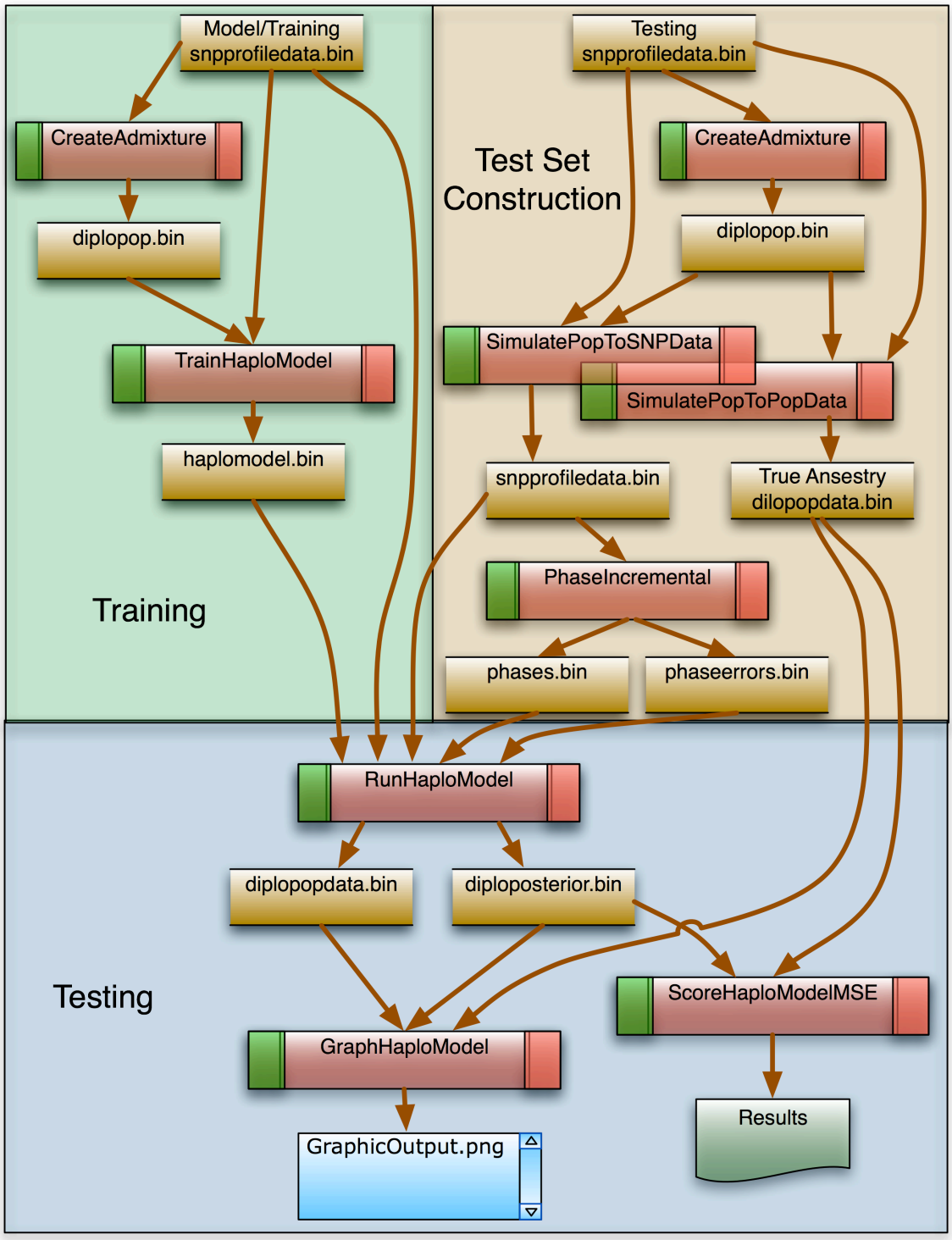
Example:

```
GraphHaploModel
   input_files=”hapmap550_chr21.snpprofiledata.bin.gz,tests.diplopopdata.bin,
   tests_hapaa.diploposterior.bin,tests_hapaa.diplopopdata.bin”
   output_file=tests.png
```

This illustration demonstrates two possible ways to specify input for the analysis. First is based on HapMap data and can be downloaded from their website. The other is using custom data. After snpprofiledata.bin has been created its content can be further restricted to a set of illumina chip locations or any other custom set.



This illustration demonstrates a typical execution flow of HAPAA code after an input file snpprofile.bin has been created. Potrion of the illustration in the left demonstrates how code can be trained. Right portion depicts process of test set creation.



If you are not interested in testing, right portion can be excluded with the exception of PhaseIncremental and snpprofiledata.bin.

Hapaa code can be compiled with a simple ‘make’ command from HAPAA directory after it has been unzipped. The current version was developed in Linux environment. Due to particular installation variations it is possible that you will need to adjust paths, compiler name, compilation flags or some other parameters. All of those values can be found in system.mk file. In case you are interested in porting to a different system you may have to make necessary modifications to System.h and System.cpp files found in src/ directory.

Next page contains a sample code for running a analysis on a single individual. This code is based in HapMap as a model and analyzes HapMap individual.

Notice that there is a naming convention that is enforced in this code. Name of each file consists of 3 parts A=‘identification name’ B=’content type’ C=’extension’. So any file name is A.B.C example: my_file_name.snpprofile.bin or my_file_name.snpprofile.bin.gz. Name such as my_file_name.another_name.bin will be rejected as an invalid input.

```
gunzip HAPAAsrc.tgz

tar –xf HAPAAsrc.tar

cd HAPAA/

make

cd Data

g++ DownloadHapMap.cpp –o DownloadHapMap

./DownloadHapMap

;;long download will follow.

cd ..

./CreateHapMapSNPData work_dir=./Data

;;This process will create very large files containing HapMap information and
;;will take a while to run.

;;download illumina information files listed in the manual.

./RestrictIlluminaSites work_dir=./Data input_file=hapmap.snpprofiledata.bin.gz
output_file=restrict.snpprofiledata.bin.gz illumina_file=Hap550v3_561466_SNPs

./RestrictSNPData input_file=./Data/restrict.snpprofiledata.bin.gz diplos="0-
40" output_file=model.snpprofiledata.bin

./RestrictSNPData input_file=./Data/restrict.snpprofiledata.bin.gz
pop_diplos="41" output_file=analyze.snpprofiledata.bin

./CreateAdmixture input_file=model.snpprofiledata.bin
output_file=model.diplopopdata.bin min_gen=1 max_gen=10 gen_samples=10
do_gender

./TrainHaploModel sample_file=./Data/model.snpprofiledata.bin
model_file=model.haplomodel.bin simulate_training_file=model.diplopopdata.bin

./PhaseIncremental sample_file=./Data/model.snpprofiledata.bin
input_file=analyze.snpprofiledata.bin switch_error_rate=0.05
output_file=model.phases.bin phaseerror_file=model.phaseerrors.bin

./RunHaploModel sample_file=./Data/model.snpprofiledata.bin
model_file=model.haplomodel.bin testing_file=analyze.snpprofiledata.bin
phases_file=model.phases.bin posterior_file=analyze.diploposterior.bin
posterior_popdata_file=model.diplopopdata.bin

./GraphHaploModel input_file="analyze.snpprofiledata.bin,
analyze.diploposterior.bin" output_file=analyze.output.png
```