



README: Genome-Wide Human SNP Array 6.0 Sample Data Set

5/25/2007

HapMap Data Set

This data set contains the 270 samples from the International HapMap Project (www.hapmap.org) run on the Affymetrix Genome-Wide Human SNP Array 6.0. The 270 samples are comprised of 30 CEPH trios, 30 Yoruban trios, 45 unrelated Han Chinese samples and 45 unrelated Japanese samples.

The Genome-Wide Human SNP Array 6.0 consists of 906,600 SNPs (or 931,946 SNPs in the full version of the CDF file). As of HapMap release 21a, a total of about 828,000 SNPs have reference genotypes available for at least 180 of the 270 HapMap samples shared here. These numbers are steadily increasing with each HapMap update.

Chromosome X Titration Data Set

An additional collection of samples are of particular use for exploring copy number variation. The copy number variation dataset consists of five replicates each of five samples. Three of the samples have abnormal copies of the X chromosome, having three, four and five copies respectively. The remaining two are a normal male and a normal female, but are of special interest as the female is the sample studied by fosmid paired-end sequencing in Tuzun *et al.* (2005) and much work has been done on this sample (Redon *et al.*, 2006) relative to the included male HapMap sample NA10851.

It is worth pointing out some of the recent changes in what is stored in the CHP files. The older XDA or GCOS format CHP file is not supported for the Genome-Wide SNP Array 6.0, as it doesn't store the SNP IDs in the file and the SNPs stored in a CHP are not guaranteed to be the same order given the availability of both default and full CDF files. In the AGCC CHP format, a few extra fields are provided in addition to the usual genotype call and confidence fields. These contain useful information, in particular for plotting SNP clusters. The fields provided are:

1. **probeset_id**
2. **genotype_call** - {NoCall,AA,AB,BB} <-> {-1,0,1,2}
3. **call_confidence** - A value in the range [0,1] with lower values corresponding to greater confidence.
4. **Signal value 1** - The x-value in the two-dimensional space recommended for plotting SNPs. For Birdseed, this value is the normalized signalA on the linear scale. For BRLMM-P this value is the transformed contrast which is defined as $f[(A-B)/(A+B)]$. For more details see www.affymetrix.com/support/technical/whitepapers/brlmm_p_whitepaper.pdf.
5. **Signal value 2** - The y-value in the two-dimensional space recommended for plotting SNPs. For Birdseed, this value is the normalized signalB on the linear scale. For BRLMM-P this value is $\log_2(A+B)$.
6. **Forced call** - This value is the genotyping call that would be made if the calling algorithm were tuned to least stringency. Providing this value allows for determining what the calls would have been at a stringency different to that applied to the calls in the second field, without having to rerun the genotyping algorithm.

These samples were subjected to the Genome-Wide Human SNP Nsp/Sty Assay 5.0/6.0 according to the protocol (*Affymetrix[®] Genome-Wide Human SNP Nsp/Sty 6.0 User Guide* [P/N 702504]).

Results

Using the Birdseed algorithm at the default confidence threshold of 0.1, the average call rate for these samples was 99.83 percent and the concordance with HapMap genotypes was 99.84 percent. Call rates and concordance values are averaged over the 270 HapMap experiments. The concordance is based on HapMap Release 21a. The Mendelian inheritance consistency was found to be 99.97 percent. **Figure 1:** Concordance relative to HapMap release 21a. The values are averaged over the 270 experiments.

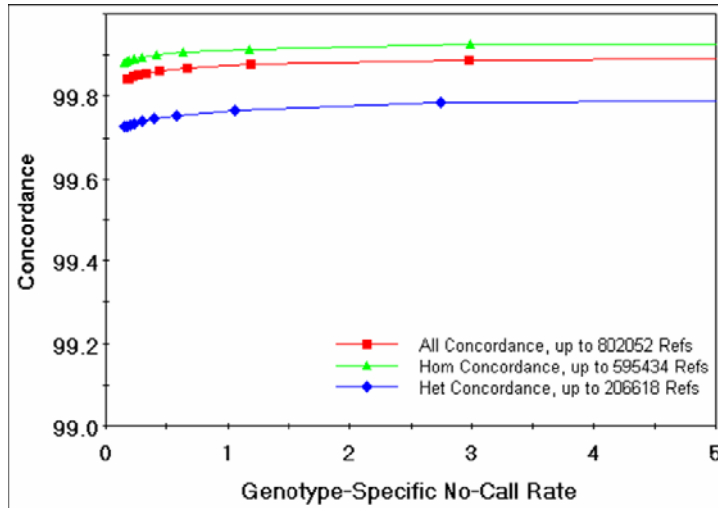


Figure 1: No call rate versus concordance plot

Data Set Distribution Information

The data consists of Affymetrix GeneChip® Command Console™ Software (AGCC) ARR, CEL and/or CHP files. AGCC will be the replacement for GCOS (http://www.affymetrix.com/products/software/specific/command_console_software.affx).

The data are made available in compressed batches for ease of distribution. Each batch is distributed as a bziped tar file. Windows users looking for an application to unzip a bz2 file can find an example at <http://www.bzip.org/downloads.html>. After unzipping, the resulting tar file can be opened by winzip. These files can all be uncompressed into the same location. The data files are grouped into three DVDs in the following manner:

Disc 1:

gw6.cel.ceu.1.tar.bz2 - 30 (10 trios) HapMap Caucasians
gw6.cel.ceu.2.tar.bz2 - 30 (10 trios) HapMap Caucasians
gw6.cel.ceu.3.tar.bz2 - 30 (10 trios) HapMap Caucasians
gw6.cel.chb.1.tar.bz2 - 22 HapMap Chinese
gw6.cel.chb.2.tar.bz2 - 23 HapMap Chinese
gw6.cel.jpt.1.tar.bz2 - 22 HapMap Japanese
gw6.cel.jpt.2.tar.bz2 - 23 HapMap Japanese
gw6.cel.chrX.tar.bz2 - 25 (5X5) 1 - 5 X chromosome titration samples

Each gw6.cel.*.tar.bz2 files contain the ARR files along with the CEL files.

ref.tar.bz2

Supplies pedigree structure of the HapMap samples, along with a large table of reference calls from HapMap21a for the SNPs on the Genome-Wide SNP Arrays 5.0 and 6.0 (please

note that the pedigree information is also available in the ARR files). This is very useful for determination of concordance with HapMap samples. The HapMap data have been translated to use an A/B allele-naming convention consistent with the Genome-Wide SNP Arrays 5.0 and 6.0, and the table format is suitable for comparisons with the output from apt-probeset-genotype (see in particular the program apt-snp-compare).

md5.txt

Checksum file for all data files. Use them to verify the integrity of your local copy of the files. (Windows users looking for an application to generate md5 checksums may find <http://www.fastsum.com/> to be useful).

Disc 2:

gw6.cel.yri.1.tar.bz2 - 30 (10 trios) HapMap Yorubans

gw6.cel.yri.2.tar.bz2 - 30 (10 trios) HapMap Yorubans

gw6.cel.yri.3.tar.bz2 - 30 (10 trios) HapMap Yorubans

Each gw6.cel.*.tar.bz2 files contain the ARR files along with the CEL files.

gw6.results.part1.tar.bz2

gw6.results.part2.tar.bz2

These files contain the text output from the QC and genotyping runs on the 270 HapMap files. Of particular use might be the big text matrix of calls, and the accompanying text matrix of allele A and allele B signals, useful for inspection of SNP clusters.

md5.txt

Checksum file for all data files. Use them to verify the integrity of your local copy of the files.

Disc 3:

gw6.chp.ceu.1.tar.bz2 - 30 (10 trios) HapMap Caucasians

gw6.chp.ceu.2.tar.bz2 - 30 (10 trios) HapMap Caucasians

gw6.chp.ceu.3.tar.bz2 - 30 (10 trios) HapMap Caucasians

gw6.chp.chb.1.tar.bz2 - 22 HapMap Chinese

gw6.chp.chb.2.tar.bz2 - 23 HapMap Chinese

gw6.chp.jpt.1.tar.bz2 - 22 HapMap Japanese

gw6.chp.jpt.2.tar.bz2 - 23 HapMap Japanese

gw6.chp.yri.1.tar.bz2 - 30 (10 trios) HapMap Yorubans

gw6.chp.yri.2.tar.bz2 - 30 (10 trios) HapMap Yorubans

gw6.chp.yri.3.tar.bz2 - 30 (10 trios) HapMap Yorubans

Each gw6.chp.*.tar.bz2 files contain the CHP files.

md5.txt

Checksum file for all data files. Use them to verify the integrity of your local copy of the files.

Additional Information and Support

Please refer to the complete Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0 User Guide for additional information.

To access to the library files for analyzing the CEL files, please visit

<http://www.affymetrix.com/support/technical/libraryfilesmain.affx>.

To access to annotation files, please visit <http://www.affymetrix.com/analysis/index.affx>.

For support information, please contact your FAS or local support team or visit <http://www.affymetrix.com/support/index.affx>.

Affymetrix, Inc.

United States/Canada: 1-888-DNA-CHIP (888-362-2447)

Europe: +44 (0) 1628 552550

Japan: +81 3-5730-8222