Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Software**

HapMap filter 1.0: A tool to preprocess the HapMap genotypic data for association studies

Wei Zhang[1], Shiwei Duan[1] and M. Eileen Dolan[1, 2, 3, *]

[1]Section of Hematology/Oncology, Department of Medicine; [2]Committee on Clinical Pharmacology and Pharmacogenomics; [3]Cancer Research Center, The University of Chicago, IL 60637, USA; M. Eileen Dolan* - E-mail: edolan@medicine.bsd.uchicago.edu; Phone: 773 702 4441; Fax: 773 702 0963; * Corresponding author

received April 29, 2008; accepted May 06, 2008; published May 13, 2008

Abstract:
The International HapMap Project provides a resource of genotypic data on single nucleotide polymorphisms (SNPs), which can be used in various association studies to identify the genetic determinants for phenotypic variations. Prior to the association studies, the HapMap dataset should be preprocessed in order to reduce the computation time and control the multiple testing problem. The less informative SNPs including those with very low genotyping rate and SNPs with rare minor allele frequencies to some extent in one or more population are removed. Some research designs only use SNPs in a subset of HapMap cell lines. Although the HapMap website and other association software packages have provided some basic tools for optimizing these datasets, a fast and user-friendly program to generate the output for filtered genotypic data would be beneficial for association studies. Here, we present a flexible, straightforward bioinformatics program that can be useful in preparing the HapMap genotypic data for association studies by specifying cell lines and two common filtering criteria: minor allele frequencies and genotyping rate. The software was developed for Microsoft Windows and written in C++.

Availability: The Windows executable and source code in Microsoft Visual C++ are available at Google Code (http://hapmap-filter-v1.googlecode.com/) or upon request. Their distribution is subject to GNU General Public License v3.

Keywords: HapMap; genotype; single nucleotide polymorphism; minor allele frequency; genotyping rate

Background:

The International HapMap Project **[1]** provides a resource of genotypic data of more than 3.1 million single nucleotide polymorphisms (SNPs) **[2]** for human lymphoblastoid cell lines (LCLs) derived from the individuals of European (CEU: Caucasians from Utah, USA), African (YRI: Yoruba people from Ibadan, Nigeria) and Asian ancestry (CHB: Han Chinese from Beijing, China and JPT: Japanese from Tokyo, Japan). Association studies using the HapMap genotypic data have generated new insights into the genetic determinants responsible for the risks of common diseases as well as quantitative phenotypes such as gene expression and individual drug response to therapeutic treatments **[3-7]**.

Because of the severity of multiple comparisons due to the large number of SNPs and the running time might be needed for a whole genome association study, the raw genotypic data downloaded from the HapMap website **[8]** requires some degree of preprocessing that includes, but is not limited to, removing uninformative and biased SNPs. The HapMap website **[8]** provides a web-based interface for users to extract genotypic data on each population by setting up parameters including population, minor allele frequency (MAF), SNP location and genomic regions. One limitation of the current website, is the inability to dump filtered genotypic data (eg. data with 80% genotyping rate or data with partial samples in a population). Software packages such as PLINK **[9]** do not provide functions that allow for the selection of certain cell lines, removal of the SNPs with biased genotypic data or do not allow flexibility in defining cutoffs (eg. MAF) **[10]**. Therefore, we wrote a C++ program, HapMap Filter, using Microsoft Visual C++ 6.0 for Windows to generate a high-quality HapMap SNP dataset that are ready to be used in association studies. HapMap Filter v1.0 features two user-specified criteria for filtering the raw HapMap data: MAF and genotyping rate. This tool gives users a flexible way to focus on the common genetic variants and those that have been well-genotyped, thus improving the multiple comparison issue and the running time of the association analyses. The output data can be conveniently used in the following association studies, for example, conducted by a customized script written in R **[11].**

Methodology:

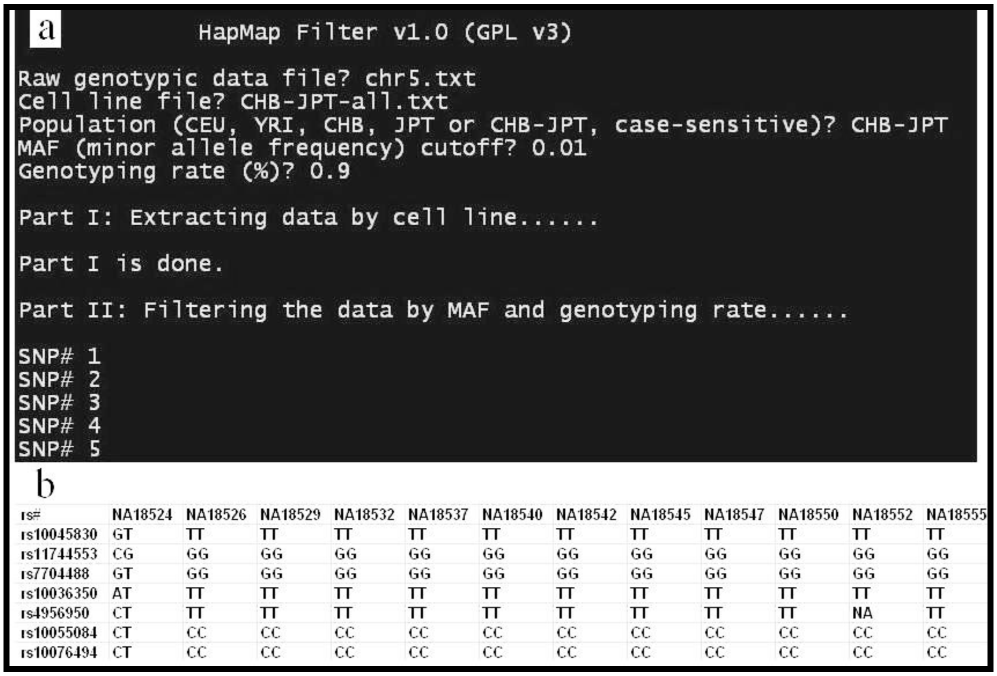
There are two major steps for each run of the software. In step 1, HapMap Filter scans the raw HapMap genotypic data to generate an intermediate output file, which contains data of user-specified cell lines. Users can choose either a subset of cell lines or all cell lines in the HapMap panel using a cell-line file. The format of the cell-line is shown in the on-line supporting materials. For convenience, we have provided cell-line files for the groups of all HapMap cell lines. For example, "CHB-all.txt" contains information for the 45 Han Chinese samples. In step 2, HapMap filters the genotypic data in the intermediate output file from step1 using two criteria: MAF and genotyping rate. If $f(AA)$, $f(Aa)$, and $f(aa)$ are the frequencies of the three genotypes at a locus with two alleles, then the frequency p of the A-allele and the frequency q of the a-allele are obtained by counting alleles. The total frequency p of A-alleles in the population is calculated as in equation 1 (in supplementary material). Similarly, the frequency q of the a-allele is given by equation 2 (under supplementary material). MAF is simply the smaller of these two frequencies. Missing data are not included in the calculation of MAF. The genotyping rate of a particular SNP is defined as $r = 1 - NA(\%)$, where $NA(\%)$ is the proportion of missing data.

Input

HapMap Filter requires the raw genotypic data files downloaded from the HapMap website **[8]**. Besides genotypic data from whole chromosomes, data within a user-specified genomic region (or a particular gene location) saved from the search engine at the HapMap website can also be used as input. The program will prompt for information which includes 1) the name of the HapMap data file (eg. chr1-dat.txt); 2) the name of the cell-line file (eg. CHB-all.txt); 3) population (eg. CHB); 4) the MAF cutoff (eg. 0.01) and; 5) the genotyping rate (eg. 0.90). Figure1a illustrates the screenshot of an example run. For easier input, we suggest the user change the original downloaded file names before running the program.

Output

HapMap Filter has two output files. The first one is "output-temp.txt", which contains all genotypic data of the user-specified cell lines. This file is then used to generate the final output file, "output.txt", which contains the filtered genotypic data by MAF and genotyping rate. Figure1b shows an example of the final output file.



```
          HapMap Filter v1.0 (GPL v3)

Raw genotypic data file? chr5.txt
Cell line file? CHB-JPT-all.txt
Population (CEU, YRI, CHB, JPT or CHB-JPT, case-sensitive)? CHB-JPT
MAF (minor allele frequency) cutoff? 0.01
Genotyping rate (%)? 0.9

Part I: Extracting data by cell line......

Part I is done.

Part II: Filtering the data by MAF and genotyping rate......

SNP# 1
SNP# 2
SNP# 3
SNP# 4
SNP# 5
```

rs#	NA18524	NA18526	NA18529	NA18532	NA18537	NA18540	NA18542	NA18545	NA18547	NA18550	NA18552	NA18555
rs10045830	GT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs11744553	CG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs7704488	GT	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG	GG
rs10036350	AT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT	TT
rs4956950	CT	TT	TT	TT	TT	TT	TT	TT	TT	TT	NA	TT
rs10055084	CT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC
rs10076494	CT	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC	CC

Figure 1: The screenshot of an example run of HapMap Filter 1.0 and its output. (a) The input data include 1) raw genotypic file: chr5.txt; 2) cell line file: CHB-JPT-all.txt; 3) population: CHB-JPT; 4) MAF cutoff: 0.01 (1%) and 5) genotyping rate: 0.9 (90%). (b) The output is a plain text file that has columns for rs numbers and genotypic data for each cell line.

Caveat and future development:

The current v1.0 has been tested for the downloaded data files of the HapMap r22 (March, 2007) and r23 (February, 2008). The file format of the HapMap genotypic data has been consistent for these two most recent releases, though future changes of the HapMap data format will require adjustment of the source code accordingly. Also the source code can be revised easily to accommodate any new populations that are to be included in the International HapMap Project. Future development could include adding options such as filtering data by gender, linkage disequilibrium or SNP location, to name a few.

Acknowledgement:

This Pharmacogenetics of Anticancer Agents Research (PAAR) Group (http://pharmacogenetics.org) study was supported by the

NIH/NIGMS grant U01GM61393. Data generated from PAAR is deposited into PharmGKB (http://www.PharmGKB.org).

References:

[01] International HapMap Consortium, *Nature*, 437: 1299 (2005) [PMID: 16255080]
[02] K. A. Frazer *et al.*, *Nature*, 449: 851 (2007) [PMID: 17943122]
[03] R. S. Huang *et al.*, *Proc Natl Acad Sci USA*, 104: 9758 (2007) [PMID: 17537913]
[04] R. S. Huang *et al.*, *Am J Hum Genet.*, 81: 427 (2007) [PMID: 17701890]
[05] W. Zhang *et al.*, *Bioinform Biol Insights*, 2: 15 (2008) [PMID: 18392109]
[06] W. Zhang *et al.*, *Am J Hum Genet.*, 82: 631 (2008) [PMID: 18313023]
[07] S. Duan *et al.*, *Am J Hum Genet.*, 82: 1101 (2008) [PMID: 18439551]
[08] http://www.hapmap.org
[09] S. Purcell *et al.*, *Am J Hum Genet.*, 81: 559 (2007) [PMID: 17701901]
[10] http://pngu.mgh.harvard.edu/~purcell/plink/res.shtml
[11] http://www.r-project.org

Edited by P. Kangueane
Citation: W. Zhang *et al.*, Bioinformation 2(8): 322-324 (2008)

Supplementary material

Equations

$$p = f(AA) + \frac{1}{2} f(Aa) \quad \rightarrow \quad \textbf{(1)}$$

$$q = f(aa) + \frac{1}{2} f(Aa) \quad \rightarrow \quad \textbf{(2)}$$