Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Database**

FstSNP-HapMap3: a database of SNPs with high population differentiation for HapMap3

Shiwei Duan[1], Wei Zhang[1], Nancy Jean Cox[1, 2] and Mary Eileen Dolan[1, 3, 4, *]

[1]Section of Hematology/Oncology, Department of Medicine; [2]Department of Human Genetics; [3]Committee on Clinical Pharmacology and Pharmacogenomics; [4]Cancer Research Center, The University of Chicago, IL 60637, USA; M. Eileen Dolan* - Email: edolan@medicine.bsd.uchicago.edu; Phone: 773 702 4441; Fax: 773 702 0963; * Corresponding author

received September 16, 2008; revised October 10; accepted October 15, 2008; published November 09, 2008

Abstract:
The International HapMap Project has recently made available genotypes and frequency data for phase 3 (NCBI build 36, dbSNP b129) of the HapMap providing an enriched genotype dataset for approximately 1.6 million single nucleotide polymorphisms (SNPs) from 1,115 individuals with ancestry from parts of Africa, Asia, Europe, North America and Mexico. In the present study, we aim to facilitate pharmacogenetics studies by providing a database of SNPs with high population differentiation through a genomewide test on allele frequency variation among 11 HapMap3 samples. Common SNPs with minor allele frequency greater than 5% from each of 11 HapMap3 samples were included in the present analysis. The population differentiation is measured in terms of fixation index (Fst), and the SNPs with Fst values over 0.5 were defined as highly differentiated SNPs. Our tests were carried out between all pairs of the 11 HapMap3 samples or among subgroups with the same continental ancestries. Altogether we carried out 64 genomewide Fst tests and identified 28,215 highly differentiated SNPs for 49 different combinations of HapMap3 samples in the current database.

Availability: http://FstSNP-hapmap3.googlecode.com/

Keywords: database; HapMap3; SNP; population differentiation; Fst; human genome

Background:
With a public dataset of both genotypes and inferred haplotypes for millions of SNPs in ethnically diverse samples, the International HapMap Project [**1**] has provided a landscape of the human genome that enables genetic scientists to compare their genetic variant results with that of a reference. The vast information from the International HapMap Project has significantly driven the development of more efficient statistical tools for high-throughput analysis of large genetic data sets **[2]** and enhanced our knowledge of population genetics **[3-4]**. Given that the HapMap samples are Epstein-Barr virus (EBV)-transformed lymphoblastoid cell lines (LCLs) that can be purchased **[5]**, researchers can perform integrated analysis that includes genetic variation, gene expression **[6-9]**, gene transcript isoforms **[10]** and sensitivity to drugs **[11-15]**.

HapMap3 enhances the initial HapMap samples with 1,115 individual samples collected from 11 populations around the world, thus providing more diversity. The HapMap3 includes samples from individuals of African, Asian, European and Mexican descent residing in various locations. Therefore, it has become an ideal resource to study genotype-phenotype relationships with cellular phenotypes.

In this analysis, we compare the genotypic frequencies of over 1 million SNPs among the samples. We hypothesize that the list of candidate SNPs, which have significantly different allele frequency distributions among the diverse samples, will facilitate studies attempting to correlate genotype with cellular phenotypes such as pharmacogenomics, and alert investigators to SNPs that might contribute disproportionately to substructure within more heterogeneous samples.

Methodology:
HapMap3 genotypic data and SNP function classes
The genotypic data of 1,115 individuals from 11 population samples were downloaded from the International HapMap Project website (Phase III, release 1). Two samples NA18955, NA18962 in the CHB group actually belong to the JPT sample set, and thus were dropped out of the analysis. In the current study, the tested population panel focused only on the 931 unrelated individuals comprised of Gujarati Indians in Houston, Texas (GIH, n = 83), individuals of Mexican ancestry in Los Angeles, California (MEX, n = 47) and individuals of African ancestry from the Southwest USA (ASW, n = 47), Luhya in Webuye, Kenya (LWK, n = 83), Maasai in Kinyawa, Kenya (MKK, n = 143), Yoruba in Ibadan, Nigeria (YRI, n = 108), Han Chinese from Beijing,

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(3): 139-141 (2008)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Database**

China (CHB, n = 80), Chinese from metropolitan Denver, Colorado (CHD, n = 70), Japanese from Tokyo, Japan (JPT, n= 82); Utah residents with Northern and Western European ancestry from the CEPH collection (CEU, n = 111), and Tuscans from Italy (TSI, n = 77). Of note, two samples (NA18955, NA18962) included in the CHB group belong to JPT. Five groups of samples (ASW, CEU, MEX, MKK and YRI) are familial samples; and the rest are unrelated samples. A total of 1,614,792 polymorphic SNPs were genotyped in the HapMap3 Project. The number of shared SNPs ranges from 1,047,055 to 1,487,361 for the pair-wise combinations of the 11 HapMap3 population samples (Fig.1). Using the same reference allele of a SNP, we calculated the allele frequencies across all the populations.

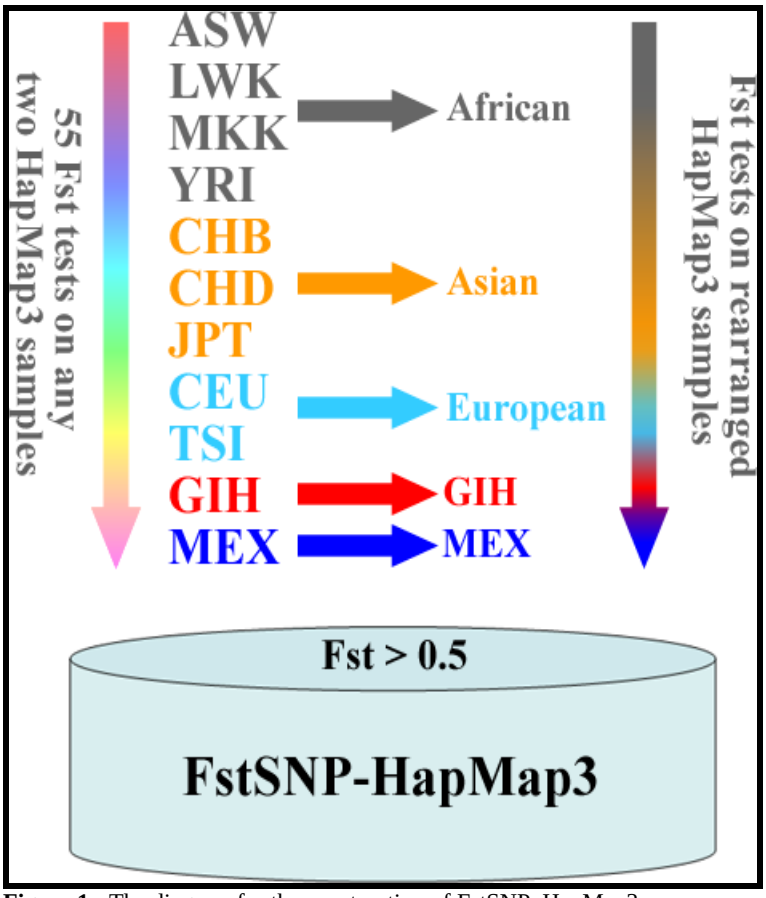

Figure 1: The diagram for the construction of FstSNP_HapMap3.

Fst calculation

Fst, a metric representation of the effect of population subdivision, was estimated according to Wright's approximate formula, $Fst = (H_T - H_S) / H_T$ where H_T represents expected heterozygosity per locus of the total population and H_S represents expected heterozygosity of a subpopulation [**16**]. An Fst value was calculated for each SNP of interest with allele frequencies estimated from the unrelated individuals in each population. We calculated Fst values for only SNPs with minor allele frequencies greater than 5% in each of the HapMap3 samples.

Dataset

The dataset [**17**] contains all SNPs with Fst value over 0.5 from the tests between any two of the 11 HapMap3 samples. There are 28,215 common SNPs with high population differentiation observed for at least 1 of 49 different combinations of HapMap3 population samples in the databases.

Development

The genotypes of approximately 1.6 million SNPs were downloaded from the International HapMap Project (HapMap3, release 1). Fst was evaluated for over 1 million SNPs for each of the 55 paired combinations from the 11 HapMap3 samples, followed by rearrangement into five distinct groups defined by historical geographic ancestry (African, Asian, European, MEX and GIH) (Fig.1). The Fst test was evaluated again among the combinations of the five geographical defined groups. Altogether there are 64

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(3): 139-141 (2008)

genomewide Fst tests for the HapMap3 samples. SNPs with high population differentiation (Fst > 0.5) were categorized into two major classes: genic and nongenic SNPs. The genic SNPs were further divided into six function classes including intron, UTR (utr-3 or utr-5), locus region (near-gene-3 or near-gene-5), splice site, coding synonymous or nonsynonymous based on the annotation in the NCBI dbSNP129 database. The SNP annotation was added as a separate column to facilitate additional applications in the future.

Database content

This database **[17]** provides 115, 212 entries of 28,215 SNPs with high differentiation among 48 different combinations of the HapMap3 samples. There are six columns for the database as follows: SNP, SNP_CHR (SNP Chromosome), SNP_POSI (SNP position in dbSNP129), Fst value, Tested_Populations and DbSNP129_Class. The number of the entries in each group ranges from 3 for GIH_TSI group to 10,416 for JPT_YRI group.

Database usage

The users can download the FstSNP_HapMap3 **[17]** and then query the database by the conditions of genomic regions or a list of SNPs or SNP Class.

Caveats:

The Fst test was evaluated in the HapMap3 population samples, some of which have relatively small sample size (e.g. 47 individuals in ASW and MEX). Furthermore, we note that SNPs included in HapMap3 are subject to a number of ascertainment biases; thus, these SNPs cannot be considered as representative of Fst values that might be calculated for all common variants.

Acknowledgment:

This Pharmacogenetics of Anticancer Agents Research (PAAR) Group http://pharmacogenetics.org) study was supported by the NIH/NIGMS grant U01GM61393.

References:

[01] http://www.hapmap.org

[02] J. Li *et al.*, *Am J Hum Genet.* 79: 628 (2006) [PMID: 16960799]

[03] P. C. Sabeti *et al.*, *Nature*, 449: 913 (2007) [PMID:1794313]

[04] L. B. Barreiro *et al.*, *Nat Genet.*, 40: 340 (2008) [PMID: 18246066]

[05] http://www.coriell.org

[06] B. E. Stranger *et al.*, *Nat Genet.*, 39: 1217 (2007) [PMID: 17873874]

[07] W. Zhang *et al.*, *Am J Hum Genet.*, 82: 631 (2008) [PMID: 18313023]

[08] S. Duan *et al.*, *Am J Hum Genet.*, 82: 1101 (2008) [PMID: 18439551]

[09] R. S. Huang *et al.*, *Pharmacogenet Genomics*, 18: 545 (2008) [PMID: 18496134]

[10] T. Kwan *et al.*, *Nat Genet.*, 40: 225 (2008) [PMID: 18193047]

[11] S. Duan *et al.*, *Cancer Res.*, 67: 5425 (2007) [PMID: 17545624]

[12] R. S. Huang *et al.*, *Proc Natl Acad Sci U S A.*, 104: 9758 (2007) [PMID: 17537913]

[13] R. S. Huang *et al.*, *Am J Hum Genet.*, 81: 427 (2007) [PMID: 17701890]

[14] S. J. Shukla *et al.*, *Pharmacogenet Genomics*, 18: 253 (2008) [PMID: 18300947]

[15] R. S. Huang *et al.*, *Cancer Res.*, 68: 3161 (2008) [PMID: 18451141]

[16] S. Wright, *Nature*, 166: 247 (1950) [PMID: 15439261]

[17] http://FstSNP-hapmap3.googlecode.com

Edited by P. Kangueane

Citation: Duan *et al.*, Bioinformation 3(3): 139-141 (2008)

ISSN 0973-2063 (online) 0973-2063 (print)
Bioinformation 3(3): 139-141 (2008)