View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Chapman University Digital Commons

e-Research: A Journal of Undergraduate Work

Volume 1
Number 2 *Vol 1, No 2 (2010)*

Article 5

September 2014

Haplotype Variety Analysis of Human Populations: an Application to HapMap Data

Michelle Creek

Cyril Rakovski

Follow this and additional works at: http://digitalcommons.chapman.edu/e-Research

Part of the Genetics Commons

Recommended Citation

Creek, Michelle and Rakovski, Cyril (2014) "Haplotype Variety Analysis of Human Populations: an Application to HapMap Data," *e-Research: A Journal of Undergraduate Work*: Vol. 1: No. 2, Article 5.
Available at: http://digitalcommons.chapman.edu/e-Research/vol1/iss2/5

This Article is brought to you for free and open access by Chapman University Digital Commons. It has been accepted for inclusion in e-Research: A Journal of Undergraduate Work by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.

e-Research: A Journal of Undergraduate Work, Vol 1, No 2 (2010)

HOME ABOUT USER HOME SEARCH CURRENT ARCHIVES

Home > Vol 1, No 2 (2010) > Creek

Haplotype Variety Analysis of Human Populations: an Application to HapMap Data

Michelle Creek, Cyril Rakovski

Abstract

We undertake a study to investigate the haplotype variety of distinct human populations. We use a natural measure of haplotype variety, the total number of haplotypes (TNH) present that reflects the number of haplotypes with nonzero frequencies estimated from the data at hand for each selection of multiple loci. For the analysis of real human populations, we use the haplotype data of the Denver Chinese, Tuscan Italians, Luhya Kenyans, and Gujarati Indians from release III of the HapMap database. Moreover, we show that the TNH statistic is biased in small sample data scenarios such as the HapMap and implement a nested simulation study to estimate and remove such bias. We perform a preliminary analysis of means and variances of the population allele frequencies in the four populations. Lastly, we implement a generalized linear model to detect and quantify the differences in haplotype structures of these populations. Our results show that all populations possess significantly different adjusted average TNH values. Our findings extend previous results based on alternative statistical approaches and demonstrate the existence of pronounced differences in the haplotype variety of the analyzed populations even after controlling for haplotype span as well as all allele frequencies and their two-way interactions.

Keywords: Multimarker Disequilibrium Analysis, Generalized Linear Models, Single Nucleotide Polymorphism

Introduction

Differences in the genetic structure of human populations have been studied and assessed from various viewpoints [Gu, et al. 2008; Joubert, et al. ; Lundmark, et al. 2008; Marvelle, et al. 2007; Sved, et al. 2008]. In this work we analyze the haplotype variety in distinct human populations via a naturally arising measure, the total number of haplotypes present that reflects the number of haplotypes with nonzero frequencies estimated from the data at hand for each selection of multiple loci. Thus, we define a counting measure, as an alternative to existing approaches such as the haplotype entropy [Nothnagel, et al. 2002] and haplotype diversity [Clayton 2001], that can be inherently modeled by generalized linear models or generalized estimating equations. We employ the TNH statistic to perform analysis of real human populations using haplotype data of the Denver Chinese (CHD), Tuscan Italians (TSI), Luhya Kenyans (LWK), and Gujarati Indians (GIH) from release III of the HapMap database [Tanaka 2009]. It is well known that the HapMap data possesses intrinsic limitations [Biswas, et al. 2007; Check

Copyright © 2010 M. Creek, C. Rakovski

2007; Zhang and Dolan 2008]. Further, we show that the TNH statistic is biased in small sample data scenarios such as the HapMap and implement a nested simulation study to estimate and remove such bias. We perform a preliminary analysis of means and variances of the population allele frequencies in the four populations. Further, we determine the differences in haplotype structures of these populations via an appropriate generalized linear model. As a secondary note, we estimate the increase of haplotype variety as a function of haplotype span which can be viewed as a multivariate generalization of previous results on the decay of linkage disequilibrium (LD) as a function of physical distance between pairs of single nucleotide polymorphisms (SNPs) [Bosch, et al. 2009].

Methods

Assume that we consider $\boldsymbol{n}$ biallelic markers $\boldsymbol{M_1, M_2, \ldots, M_n}$ with alleles at each locus coded as 1 or 2. Let $\boldsymbol{i_1 i_2 \ldots i_n}$ be a haplotype based on these markers and $\mathbf{H}^n = \{\boldsymbol{i_1 i_2 \ldots i_n} \mid \boldsymbol{i_k} \in \{1,2\}\}$ denote the set of all 2^n such haplotypes with 2 representing the rarer allele. A natural measure of haplotype variety is the total number of haplotypes present (TNH) defined in the following fashion:

(1)
$$TNH(M_1, M_2, \ldots, M_n) = \sum_{i_1 i_2 \ldots i_n \in H^n} I\{P(i_1 i_2 \ldots i_n) > 0\},$$

where $\boldsymbol{P(i_1 i_2 \ldots i_\mathrm{n})}$ denote the population frequency of haplotype $\boldsymbol{i_1 i_2 \ldots i_n}$.

Clearly, the values of the TNH counting measure are bounded by one and 2^n with one representing the scenarios with the fewest number of haplotypes and all possible haplotypes present respectively. We are interested in comparing the haplotype variety of different populations through the THN statistic after adjusting for potential confounders. However, the true population haplotype frequencies that appear in definition are unknown and we need to replace them with their maximum likelihood estimates (MLE) based on the data at hand,

(2)
$$T\hat{N}H(M_1, M_2, \ldots, M_n) = \sum_{i_1 i_2 \ldots i_n \in H^n} I\{\hat{P}(i_1 i_2 \ldots i_n) > 0\}.$$

Clearly, by using equation we add bias to the TNH measure of haplotype variety. For instance, under independent allelic transmissions (i.e. by ignoring the reduction of haplotype variety due to complex multilocus structure), all haplotypes should be present for each selection of loci. For instance, in the simple case of haplotypes based on four markers with minor allele frequencies 0.05 (the lowest considered in this work), the population frequency of the rarest haplotype will be $P(2222) = 0.05^4 = 10^{-6}$ (6.25) and therefore highly unlikely to be observed in small sample data such as the HapMap.

For each selection of markers $\boldsymbol{M_1, M_2, \ldots M_n}$ from a particular population and chromosome of the HapMap data, we can estimate the allele frequencies

$$\hat{P}(i_1), \hat{P}(i_2), \ldots, \hat{P}(i_n)$$

through the sample proportions which are the MLEs. We can estimate the above-mentioned bias analytically by noticing that under independent allelic transmissions, the joint distribution of all haplotypes is multinomial,

(3)
$$P(h_1, h_2, \ldots, h_{2^n}) \sim Mult\left[\prod_{k=1}^{n} \hat{P}(i_k^1), \prod_{k=1}^{n} \hat{P}(i_k^2), \ldots, \prod_{k=1}^{n} \hat{P}(i_k^{2^n})\right],$$

where $\boldsymbol{h_k = i_1^k i_2^k \ldots i_n^k, k = 1, 2, \ldots 2^n}$ denote all elements of $\boldsymbol{H^n}$. However, we proceed with an empirical approach by undertaking a nested simulation study to estimate and consequently remove this bias from the TNH statistic. Similarly, for each selection of selection of markers $\boldsymbol{M_1, M_2, \ldots M_n}$ we estimate the allele frequencies

$$\hat{P}(i_1), \hat{P}(i_2), \ldots, \hat{P}(i_n)$$

and simulate $\boldsymbol{K}$ datasets, $\boldsymbol{D_1, D_2, \ldots D_K}$ consisting of 170 haplotypes (this is the number of independent haplotypes for the selected populations in the HapMap data) under the condition of independent allelic transmissions i.e.

$$\hat{P}(i_1 i_2 \ldots i_n) = \prod_{j=1}^{n} \hat{P}(i_j)$$
.

Then, the bias can be empirically estimated by the mean of the sample-specific bias values,

(4) $$\hat{B}ias[T\hat{N}H(M_1, M_2, \ldots, M_n)] = 2^n - \sum_{j=1}^{K} \left\{ \sum_{i_1 i_2 \ldots i_n \in H^n} I\{\hat{P}_j(i_1 i_2 \ldots i_n) > 0\} \right\} / K,$$

where

$$\hat{P}_j(i_1 i_2 \ldots i_n)$$

denotes the estimated population frequency of haplotype $\boldsymbol{i_1 i_2 \ldots i_n}$ based on simulated dataset $\boldsymbol{D_j}$.

Next, the comparison of haplotype structures of distinct isolated populations is performed through a linear regression analysis with outcome variable being the bias-corrected TNH. We also investigate the effect of haplotype span, allele frequencies and their two-way interactions,

(5) $$\log[T\hat{N}H_i - \hat{B}ias(T\hat{N}H_i)] = X_i \beta + \varepsilon_i, \ \varepsilon_i \square N(0, \sigma^2).$$

The logarithmic transformation the TNH measure decreases is a standard approach to attain better fit in count data settings.

Lastly, as a preliminary data analysis step, we compare the variances and means of the allele frequencies of the four populations using classical tools such as F-tests and t-tests. These analytical methods are robust to model misspecification which is likely the case with high density SNP data where the independence assumptions are violated due to presence of linkage disequilibrium (LD).

Data

For the study on isolated populations we chose the Chinese in Denver, Colorado, the Toscani in Italy, the Luhya in Webuye, Kenya, and the Gujarati Indians in Houston, Texas. We use the haplotypes from 22 autosomal chromosomes of the corresponding datasets (CHD, TSI, LWK, and GIH) on unrelated subjects from release III of the HapMap database. Details on these dataset characteristics are shown in Table 1.

Table 1. HapMap datasets characteristics.

Population	Total number of unrelated samples	Total number of haplotypes	Total number of SNPs	Total number of common SNPs*
CHD	85	170	1,387,466	966,507
GIH	88	176	1,387,466	1,071,872
LWK	90	180	1,387,486	1,141,860
TSI	88	176	1,387,486	1,069,127

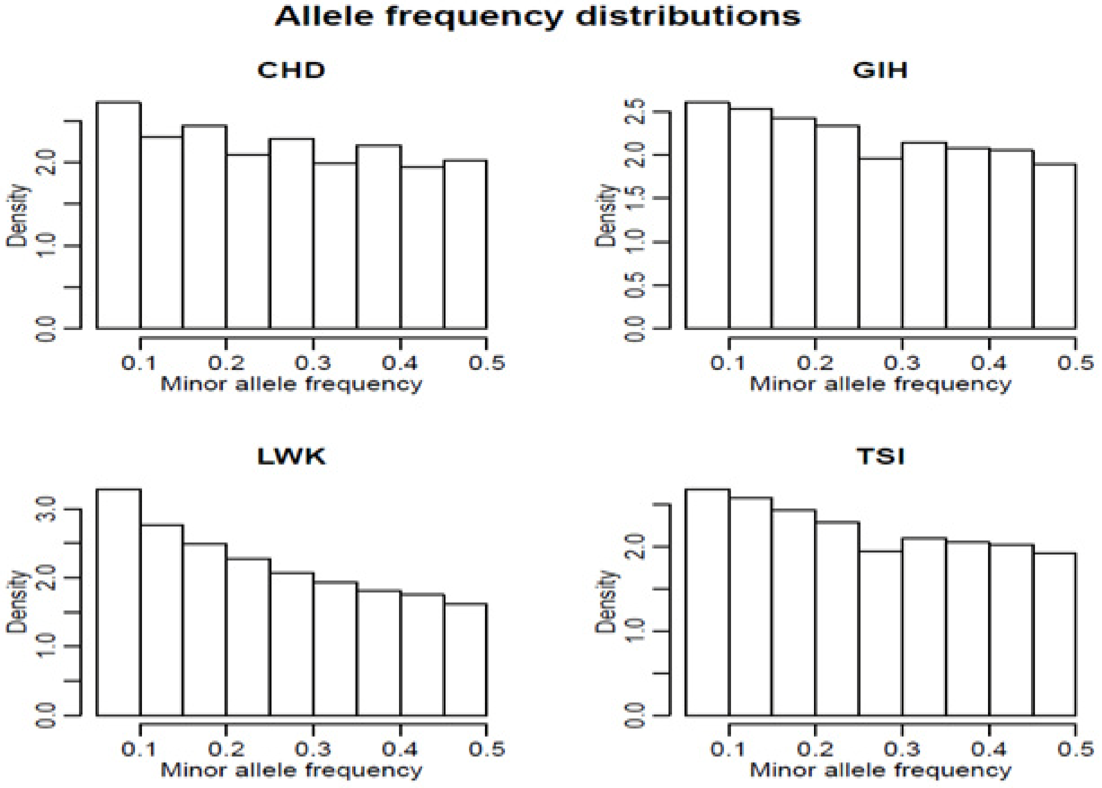
*SNPs with minor allele frequency greater than 0.05.

We removed the SNPs with minor allele frequencies smaller than 0.05 (reduction of 30, 23, 18 and 23% from each dataset respectively). Further, for each population and autosomal chromosome pair we randomly selected 1000 quadruples of loci (a total of 88,000) and in the subsequent analysis, we investigated the diversity of the haplotypes based on these markers via the TNH measure.

Results

We performed the complete analysis for haplotypes based on three, four and five markers. We present results for the case of four-locus haplotypes. The minor allele frequency distributions of the four studied populations reveal noteworthy dissimilarities. The corresponding histograms are shown in Figure 1.

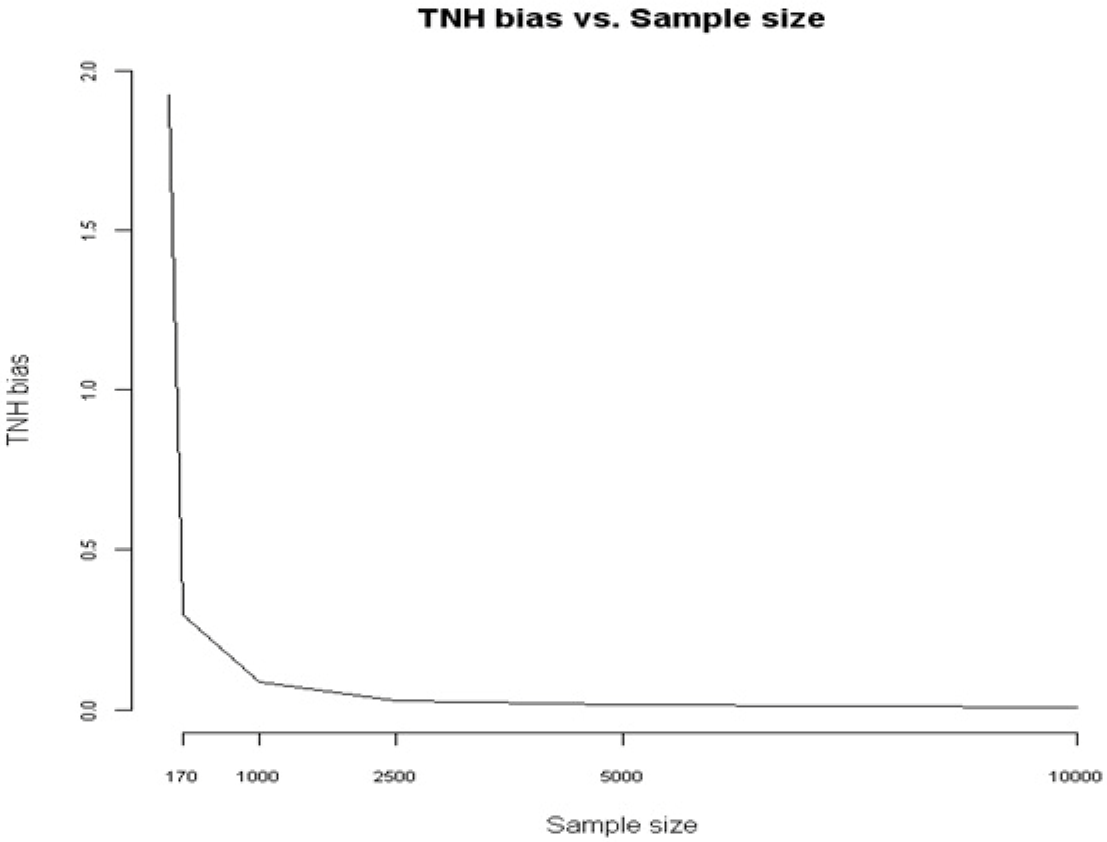
Figure 1: Minor allele frequency distributions for the four populations.



Preliminary analysis via two-sample t-tests confirms that the population average minor allele frequencies differ significantly between all possible pairs of populations with p-values smaller than 3.10^{-14}. Interestingly, the only population variance that differs significantly from the rest is that of the Tuscan Italians with all F-test p-values smaller than 0.002.

Next, we investigate the magnitude of the TNH bias as a function of the sample size. Figure 2 shows the TNH bias averaged over 10,000 randomly selected haplotypes for four different sample sizes, 170, 100, 2500, 5000, and 10000 respectively.

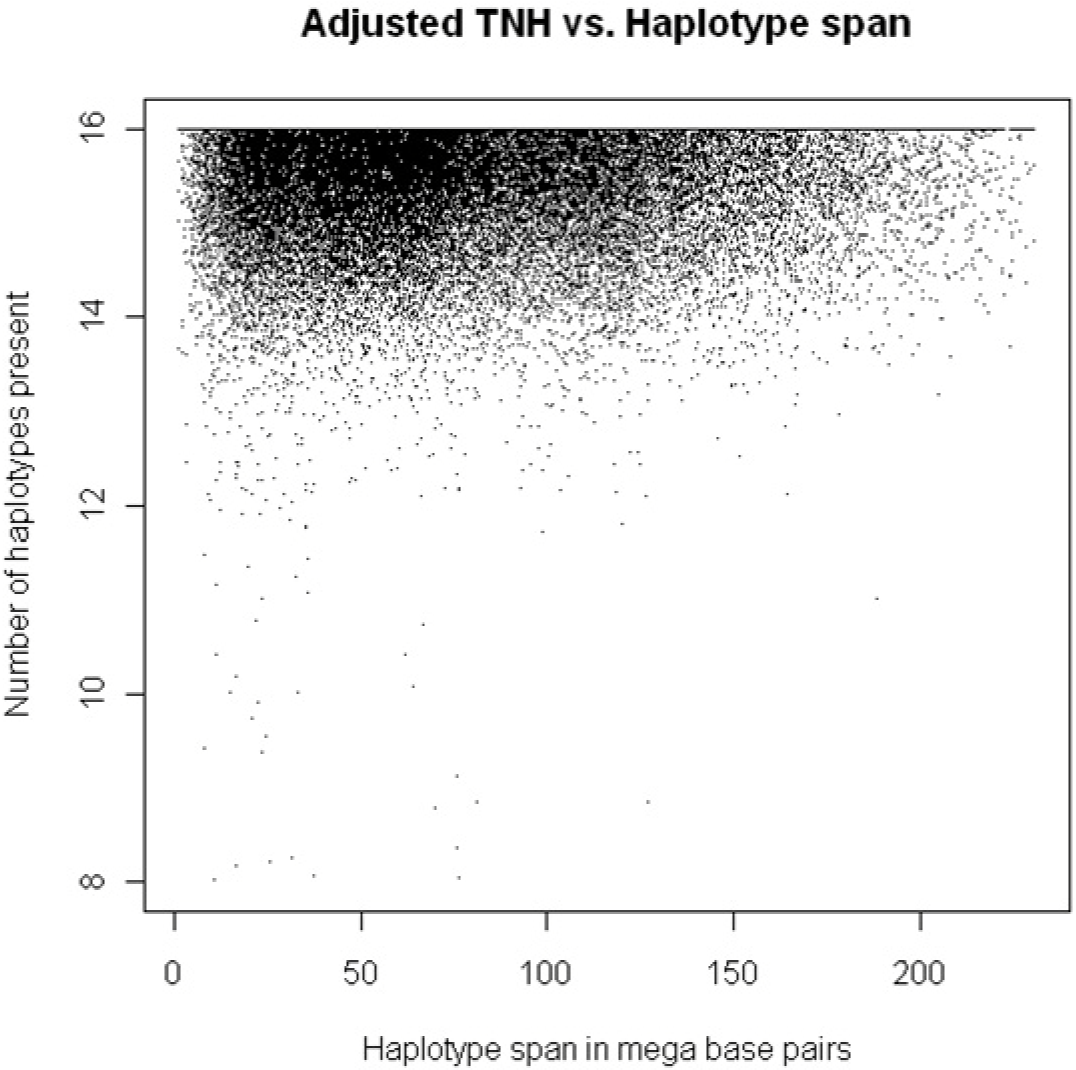
Figure 2: Average TNH bias versus sample size.



The average bias for the HapMap dataset sizes is approximately 0.3. In the subsequent analysis, for each random selection of four loci, we estimate and remove this selection-specific bias using expression .

The decrease of linkage disequilibrium as the distance between pairs of markers increase is well-known. Figure 3 shows a multilocus version of these phenomena, the increase of THN as the haplotype span grows.

Figure 3: Bias-adjusted TNH versus Haplotype span in mega base pairs.



As expected, the number of haplotypes approaches the maximum possible number of haplotypes as the physical distance spanned by the flanking markers increases. However, the rate of increase seems much slower than anticipated. The magnitude of the adjusted haplotype span effect is estimated in the subsequent regression analysis.

Lastly, the linear regression results are presented in Table 2. We arrived at this model through standard model fitting techniques such as forward selection and backward elimination of potential predictors as well as residual diagnostic and goodness of fit approaches.

Table 2. Regression analysis results.

Variable	Estimate	SE	t-value	p-value
Intercept	2.773	2.069e-03	1339.883	<2e-16
Haplotype span	1.855e-11	4.271e-12	4.343	1.41e-05
GIH	3.171e-03	5.456e-04	5.811	6.24e-09
LWK	5.646e-03	5.479e-04	10.305	< 2e-16
TSI	3.066e-03	5.457e-04	5.617	1.95e-08
AlFreq1	1.016e-02	5.259e-03	1.932	0.053368
AlFreq2	1.421e-02	5.236e-03	2.714	0.006653
AlFreq3	5.669e-03	5.258e-03	1.078	0.280979
AlFreq4	1.430e-02	5.247e-03	2.725	0.006424
AlFreq1:AlFreq2	2.796e-02	1.121e-02	-2.495	0.012594
AlFreq1:AlFreq3	-1.449e-02	1.121e-02	-1.293	0.196074
AlFreq1:AlFreq4	-2.450e-02	1.120e-02	-2.187	0.028763
AlFreq2:AlFreq3	-1.698e-02	1.122e-02	-1.513	0.130375
AlFreq2:AlFreq4	-3.725e-02	1.124e-02	-3.315	0.000917
AlFreq3:AlFreq4	-2.005e-02	1.124e-02	-1.785	0.074316

The important results from this analysis is that the haplotype varieties differ significantly among the four populations even after adjusting for haplotype span, allele frequencies and their two-way interaction. Controlling for these covariates in the model shows the depth and intricacies of the haplotype structure distinctions. Moreover, based on the adjusted average TNH values, the Luhya Kenyans possess the highest haplotype variety followed by the Gujarati Indians, the Toscani Italians and the Chinese in Denver respectively. The haplotype span is measured kilo base pairs and surprisingly, the adjusted effect of haplotype span is extremely small, 10^{-11} with p-value of 2.10^{-5}.

Discussion

The difference among human populations studied in this work are pronounced and complex. Not only are all allele frequency means and some of the variances significantly different but also the haplotype varieties even after we controlling for span, allele frequencies and their two way interactions. Moreover, the increase of haplotype variety as the distance between the flanking markers increases is dramatically smaller than anticipated given the rate of decay of LD between pairs of SNPs. However, these conclusions are difficult to compare to existing results due to the unique modeling approach.

Having obtained a degree of insight into the intricacies of the haplotypes structures of human populations we can perform a large-scale analysis using more precise measures of haplotype variety. We can explore the joint distributional structure of the all possible haplotypes under independence given in through a Monte Carlo test procedure [Hope 1968] that circumvents the minimum expected count requirements characteristic to classical large-sample chi-square contingency tables goodness-of-fit tests. Alternatively, we can improve the TNH statistic by measuring not only the number of haplotypes present but also the extent of departure of the observed haplotypes frequencies from the expected frequencies under independent allelic transmissions.

References

Biswas NK, Dey B, Majumder PP. 2007. Using HapMap data: a cautionary note. Eur J Hum Genet 15(2):246-9.

Bosch E, Laayouni H, Morcillo-Suarez C, Casals F, Moreno-Estrada A, Ferrer-Admetlla A, Gardner M, Rosa A, Navarro A, Comas D and others. 2009. Decay of linkage disequilibrium within genes across HGDP-CEPH human samples: most population isolates do not show increased LD. BMC Genomics 10:338.

Check E. 2007. Time runs short for HapMap. Nature 447(7142):242-3.

Clayton DG. 2001. Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. Nature.

Gu CC, Yu K, Rao DC. 2008. Characterization of LD structures and the utility of HapMap in genetic association studies. Adv Genet 60:407-35.

Hope ACA. 1968. A Simplified Monte Carlo Significance Test Procedure. Journal of the Royal Statistical Society Series B-Statistical Methodology 30(3):582-&.

Joubert BR, North KE, Wang Y, Mwapasa V, Franceschini N, Meshnick SR, Lange EM. Comparison of genome-wide variation between Malawians and African ancestry HapMap populations. J Hum Genet.

Lundmark PE, Liljedahl U, Boomsma DI, Mannila H, Martin NG, Palotie A, Peltonen L, Perola M, Spector TD, Syvanen AC. 2008. Evaluation of HapMap data in six populations of European descent. Eur J Hum Genet 16(9):1142-50.

Marvelle AF, Lange LA, Qin L, Wang Y, Lange EM, Adair LS, Mohlke KL. 2007. Comparison of ENCODE region SNPs between Cebu Filipino and Asian HapMap samples. J Hum Genet 52(9):729-37.

Nothnagel M, Furst R, Rohde K. 2002. Entropy as a measure for linkage disequilibrium over multilocus haplotype blocks. Hum Hered 54(4):186-98.

Sved JA, McRae AF, Visscher PM. 2008. Divergence between human populations estimated from linkage disequilibrium. Am J Hum Genet 83(6):737-43.

Tanaka T. 2009. [HapMap project]. Nippon Rinsho 67(6):1068-71.

Zhang W, Dolan ME. 2008. On the challenges of the HapMap resource. Bioinformation 2(6):238-9.