Selection of Minimum Subsets of Single Nucleotide Polymorphisms to Capture Haplotype Block Diversity

H.I. Avi-Itzhak, X. Su, F.M. De La Vega

Pacific Symposium on Biocomputing 8:466-477(2003)

SELECTION OF MINIMUM SUBSETS OF SINGLE NUCLEOTIDE POLYMORPHISMS TO CAPTURE HAPLOTYPE BLOCK DIVERSITY

HADAR I. AVI-ITZHAK[1,2], XIAOPING SU[1], FRANCISCO M. DE LA VEGA[1]

[1]Applied Biosystems, Bioinformatics R&D, 850 Lincoln Center Drive, Foster City CA 94404, and [2]Imagenix Corporation, P.O. Box 735, Los Altos CA 94023

We present a simple numerical algorithm to select the minimal subset of SNPs required to capture the diversity of haplotype blocks or other genetic loci. This algorithm can be used to quickly select the minimum SNP subset with no loss of haplotype information. In addition, the method can be used in a more aggressive mode to further reduce the original SNP set, with minimal loss of information. We demonstrate the algorithm performance with data from over 11,000 SNPs with average spacing of 6 to 11 Kb, across all the genes of chromosomes 6, 21, and 22, genotyped on DNA samples of 45 unrelated African-Americans and 45 Caucasians from the Coriell Human Diversity Collection. With no loss of information, we reduced the number of SNPs required to capture the haplotype block diversity by 25% for the African-American and 36% for the Caucasian populations. With a maximum loss of 10% of haplotype distribution information, the SNP reduction was 38% and 49% respectively for the two populations. All computations were performed in less than 1 minute for the entire dataset used.

1 Introduction

Single nucleotide polymorphisms (SNPs) are the most abundant form of genetic variation. These single nucleotide changes are found approximately every 500 base pairs (bp) in the human genome. Almost all SNPs are bi-allelic (i.e., only two different alleles exist), and typically one allele is present in the majority of the chromosomes of a population, and the alternative variant (i.e., the minor allele) is present with less frequency.

SNPs are promising tools for mapping susceptibility mutations that contribute to complex diseases. Although most SNPs are neutral (i.e., do not affect phenotype), they can be used as surrogate markers for positional cloning of genetic loci, because of the allelic association, known as linkage disequilibrium (LD), that can be shared by groups of adjacent SNPs. LD is eroded by gene conversion [3] and recombination [12], and the amount of LD depends on the age of the mutations and on the demographic history of the population. The extent of LD across a genomic region dictates the density of SNP markers necessary to ensure association between a marker and the causative allele sought.

Early attempts to model the extent of LD on theoretical grounds predicted very short regions of LD, extending only a few kilobases (Kb) [13]. However, empirical surveys reported average LD distances between 5 Kb and 60 Kb, with the upper range extending up to hundreds of Kb [2,4,14]. More recently, studies have reported a discontinuous structure in the patterns of LD across specific genomic regions,

where long stretches of strong LD are punctuated by recombination hot-spots [4,8,9]. These LD "blocks" show little evidence of historical recombination. These results suggest that a reduced set of contiguous chromosomal segments, or haplotypes, exist in specific populations. For example, for a block spanning tens of Kbs for which 10 SNPs exist, instead of the 2^{10} theoretically possible haplotypes, it has been found that 95% of the haplotype diversity is made up of only 4 to 6 so-called common haplotypes [8].

It is noteworthy that these LD block patterns change depending on the population sampled because of historical differences; for example, populations that have experienced bottlenecks (e.g., Caucasians) show longer LD blocks and less evidence of historical recombination events than other populations [8]. The haplotype diversity in a given population is typically constant in a given block irrespective of the number of SNPs sampled [8]; therefore typing an arbitrarily large number of SNPs within a LD block is unnecessary. Selecting the minimum subset of SNPs within LD blocks, or any other discrete genetic loci, that enable discrimination of the common haplotypes present in a block without loss of information is the optimum approach.

Most experimental methods for typing the specific SNP alleles present in a DNA sample produce unphased genotypes (i.e., the alleles detected cannot be assigned to either the maternal or the paternal chromosome). Although cumbersome methods exist to directly determine haplotypes, algorithms are widely used to infer the haplotypes from genotypes using maximum-likelihood or Bayesian principles. Family relationships of the DNA donors, if available, can be used to increase haplotype inference accuracy. Even in absence of family information, the Expectation-Maximization (EM) algorithm introduced by Excoffier and Slatkin [7] is quite accurate in most realistic situations, especially in regions of low diversity [16]. The analysis of haplotype distributions has been reported to provide more power for finding associations in genetic studies undertaken to find susceptibility mutations in case-control populations [6].

2 The Algorithm for Determining the Minimum Set

Given a block containing N SNPs and M haplotypes, define P, a probability vector of length M, where P_i is the relative frequency of the i^{th} haplotype. Also define A, a haplotype/SNP allele state matrix of N columns and M rows, where A_{ij}, (the i^{th} row of the j^{th} column of this matrix) indicates the allele state ('1' or '2') of the j^{th} SNP for the i^{th} haplotype. The algorithm will eliminate as many columns of A, while preserving as much of the information in P as possible. For the purpose of quantifying the information in P, we used the information-theoretic quantity known as Shannon Entropy [15] as the measure of haplotype diversity information within each LD block [11]:

$$H = -\sum_{i=1}^{M} P_i \log_2(P_i)$$

However, it is quite useful in many cases to use the algorithm in *lossless* mode, in which case it is irrelevant which information measure is used for the haplotype distribution. In addition, the algorithm can use any other measure of information preferred by the user.

2.1 Lossless Mode

The algorithm consists of two sequentially performed phases. Phase I is *elimination by columns*. Phase II is *elimination by rows*. Initially, we will define these operations in *lossless* mode only:

Phase I: **Any column that is identical to another column, or is the exact opposite of another column, can be eliminated.**

A column that is identical to another column represents a SNP that behaves identically to another SNP for all tested samples. Thus, under the assumption that we have enough DNA samples to infer the major haplotypes, this redundant SNP will not provide any useful information. Similarly, a column that is the exact opposite of another column is a SNP whose behavior can always be predicted from the behavior of another SNP simply by inverting it; therefore this SNP will not provide new information. Note that merely selecting a set of basis vectors from the matrix A would often miss elimination of columns that are an exact opposite of other columns. The 2x2 identity matrix illustrates this.

Assume that after phase I, the N columns of matrix A have been reduced to N' unique columns where $N' \leq N$.

Phase II: **Any column whose elimination does not reduce the number of unique rows should be eliminated.**

Each row represents the allelic states of the SNPs for a specific haplotype. Removing a "useful" SNP would eliminate the ability to detect at least one haplotype. In such a case, two or more haplotypes would register the same allelic states at the remaining SNPs, thereby reducing the number of unique rows. Therefore, if the elimination of a column does not reduce the number of unique rows, it can be omitted.

Note that phase II actually subsumes phase I, in the sense that if phase I were skipped, phase II would eliminate the SNPs that phase I would have eliminated. However, without the "pruning" that takes place in phase I, phase II can quickly become computationally unmanageable, especially in *lossy* mode.

The following example illustrates the method. Table 1 shows four SNPs and the four haplotypes they participate on a hypothetical LD block.

Table 1. Haplotype/SNP Allele State Matrix

	SNP_1	SNP_2	SNP_3	SNP_4
Haplotype 1	1	1	1	2
Haplotype 2	2	2	1	1
Haplotype 3	2	2	2	1
Haplotype 4	1	2	2	2

The fourth column is the exact opposite of the first column. This implies that either SNP_4 or SNP_1 is redundant. If SNP_4 is removed from the SNP set, no information is lost. When SNP_1 registers allele "1", the state of SNP_4 is known as allele "2", and conversely, when SNP_1 registers allele "2", the state of SNP_4 is known as allele "1". Removing SNP_4 leaves the matrix in Table 2.

Table 2. Haplotype/SNP Allele State Matrix after Phase I

	SNP_1	SNP_2	SNP_3
Haplotype 1	1	1	1
Haplotype 2	2	2	1
Haplotype 3	2	2	2
Haplotype 4	1	2	2

The three columns are unique (including accounting for opposites), thus phase I is complete. $N = 4$ has been reduced to $N' = 3$, as phase II is entered.

Table 3 depicts the three remaining matrices, following the removal of SNP_1, SNP_2, or SNP_3, respectively. The first and the third matrices only have three unique rows, whereas the second matrix has four unique rows. Thus, if the haplotype list is exhaustive, we can eliminate SNP_2 with no loss of haplotype detection.

Table 3. Three Possible Haplotype/SNP Allele State Matrices

	SNP_2	SNP_3
Haplotype 1	1	1
Haplotype 2	2	1
Haplotype 3	2	2
Haplotype 4	2	2

	SNP_1	SNP_3
	1	1
	2	1
	2	2
	1	2

	SNP_1	SNP_2
	1	1
	2	2
	2	2
	1	2

We have shown that the set $\{SNP_1, SNP_3\}$ provides the same haplotype detection ability as the full set $\{SNP_1, SNP_2, SNP_3, SNP_4\}$. In the above example, each phase caused the elimination of exactly one SNP. However, in general, each phase could result in the elimination of multiple SNPs, or possibly none.

2.2 Lossy Mode

An alternative version of the algorithm can be used to further reduce the retained SNP set while minimizing the loss of haplotype detection. Phase I will

remain unchanged, however phase II will now select the optimal SNPs to eliminate by performing an exhaustive search.

Phase II: **For the $\binom{N'}{k}$ possible selections of k SNPs, compute the entropy *H* for the resulting *P*. Choose the selection with the highest *H* as the best selection.**

When only k out of N' SNPs are selected, $N' - k$ columns are eliminated. The resulting matrix (with k columns) could have fewer unique rows than the full matrix (with N' columns). When a row is repeated more than once, it implies that several "minor" haplotypes will now be measured as a single "major" haplotype. This is because with fewer SNPs, we have lost the ability to make the finer distinction between them. The relative frequency (probability) of this "major" haplotype is equal to the sum of the frequencies of the "minor" haplotypes. Thus, when elimination of columns results in repeating rows, the repeating rows can be combined into a single row, and their respective probabilities summed to form a new probability. The vector P will be shorter and have larger numbers. This will always reduce the value of the entropy, H. The combination with the smallest reduction of entropy is deemed the optimal selection. Obviously, if all the rows are unique after elimination of $N' - k$ columns, the entropy is not reduced, and k SNPs can be used with no loss of information, as in the *lossless* case.

3 Implementation and Test Data

The algorithm was implemented in MATLAB v6.1 (The MathWorks Inc., Natick, MA, USA) without further optimization. The computations were completed on a 700MHz PC for all the 2,874 blocks in our test dataset in less than 1 minute. To validate and assess the utility of the algorithm on a realistic dataset, we used genotyping data from 11,160 SNPs distributed in a gene-centric fashion across chromosomes 6, 21, and 22 (see Table 7 below). The SNPs were scored with 5'nuclease assays with TaqMan®-MGB probes from Applied Biosystems' Assays-on-Demand™ SNP Genotyping Products (Foster City, CA, USA)[5]. The samples typed included 45 African-American and 45 Caucasian DNAs from the Coriell Human Diversity Collection (Coriell Institute for Medical Research, Camden, NJ, USA). LD blocks and haplotypes were computed independently for each population using methods described elsewhere [1,8]. Only blocks of 3 or more SNPs were considered in this study. Therefore, 4,864 SNPs were used for the African-American population and 7,347 SNPs were used for the Caucasian population, which is known, in general, to have more and longer LD blocks.

4 Experimental Results

To exemplify the usefulness of the algorithm, we applied it in *lossy* mode to an LD block that we discovered using the Caucasian population panel, in chromosome 6, overlapping the human gene *TTK* (RefSeq ID NM_003318, Celera ID hCG401205). The block consists of 17 SNPs, and the EM algorithm inferred 8 haplotypes, including two major haplotypes: haplotype 2 and haplotype 7 with frequencies of approximately 43% and 33% respectively. The remaining 24% of the diversity is spread among the remaining 6 haplotypes. Table 4 below summarizes the allelic states of the 17 SNPs, as well as the respective probability, for each of the 8 haplotypes.

Table 4. Original Haplotype/SNP Allele State Matrix

Haplotype		**SNP No.**																
Number	P	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.1136	1	1	1	1	1	2	1	2	1	2	1	1	2	2	2	1	2
2	0.4318	1	1	1	1	1	2	1	2	1	2	1	1	2	2	2	2	2
3	0.0114	1	1	1	2	2	1	2	1	2	1	2	2	1	1	1	1	1
4	0.0454	1	2	2	1	1	2	1	2	1	2	1	1	2	2	2	1	1
5	0.0454	2	1	1	1	1	2	1	2	1	2	1	1	2	2	2	1	2
6	0.0118	2	2	2	2	2	1	2	1	2	1	2	1	1	1	1	1	1
7	0.3292	2	2	2	2	2	1	2	1	2	1	2	2	1	1	1	1	1
8	0.0114	2	2	2	2	2	1	2	1	2	2	1	2	1	1	1	1	1

After running phase I of the algorithm, the number of SNPs is reduced to 7, with the remaining SNP set {SNP1, SNP_2, SNP_4, SNP_{10}, SNP_{12}, SNP_{16}, SNP_{17}}. All the haplotype information is preserved, including their distribution, as shown in Table 5 below. The entropy of the original distribution of haplotypes is $H(P)$ = 2.0351 bits.

Table 5. Haplotype/SNP Allele State Matrix After Phase I

Haplotype Number	P	SNP_1	SNP_2	SNP_4	SNP_{10}	SNP_{12}	SNP_{16}	SNP_{17}
1	0.1136	1	1	1	2	1	1	2
2	0.4318	1	1	1	2	1	2	2
3	0.0114	1	1	2	1	2	1	1
4	0.0454	1	2	1	2	1	1	1
5	0.0454	2	1	1	2	1	1	2
6	0.0118	2	2	2	1	1	1	1

Table 5. Haplotype/SNP Allele State Matrix After Phase I (continued)

7	0.3292	2	2	2	1	2	1	1
8	0.0114	2	2	2	2	2	1	1

After running phase II of the algorithm in *lossless* mode, it is apparent that the number of SNPs in the set can be reduced to 5 without any loss of information (Table 6). Reducing the set to 4 SNPs causes haplotype 8, the rarest haplotype, to merge into haplotype 7 because we have lost the ability of distinguish between the two. As a result 3.5% of the entropy is lost in the resultant haplotype distribution. The least amount of entropy that can be lost by reducing the SNP set to 3 is 9.2% when haplotypes 3 and 4 are merged, and haplotypes 6, 7, and 8 are merged.. Interestingly, the optimal single SNP is SNP_{16}. With SNP_{16} , we can detect only "haplotype 2" or "other". Haplotype 2 is the most common haplotype, with 43.2% of the frequency. Therefore it seems intuitively obvious that if we were only allowed to choose a single SNP, SNP_{16} would be our most useful choice.

Table 6. Lossy Min. SNP Set Example

No. of SNPs (k)	No. of Combinations $\binom{7}{k}$	Optimal Set of k SNPs	Haplotype Distribution Resulting from the Optimal SNP Set	Resulting Entropy (H) (bits)
7	1	{ SNP_1, SNP_2, SNP_4, SNP_{10}, SNP_{12}, SNP_{16}, SNP_{17}}	(0.114, 0.432, 0.011, 0.045, 0.045, 0.012, 0.329, 0.011)	2.0351
6	7	{ SNP_1, SNP_4, SNP_{10}, SNP_{12}, SNP_{16}, SNP_{17}}	(0.114, 0.432, 0.011, 0.045, 0.045, 0.012, 0.329, 0.011)	2.0351
5	21	{ SNP_1, SNP_{10}, SNP_{12}, SNP_{16}, SNP_{17}}	(0.114, 0.432, 0.011, 0.045, 0.045, 0.012, 0.329, 0.011)	2.0351
4	35	{ SNP_1, SNP_{12}, SNP_{16}, SNP_{17}}	(0.114, 0.432, 0.011, 0.045, 0.045, 0.012, 0.341)	1.9631
3	35	{ SNP_1, SNP_{16}, SNP_{17}}	(0.114, 0.432, 0.057, 0.045, 0.352)	1.8475
2	21	{ SNP_{12}, SNP_{16}}	(0.216, 0.432, 0.352)	1.5311
1	7	{SNP_{16} }	(0.5682 , 0.4318)	0.9865

Table 7 below summarizes the results after applying the algorithm to the full dataset of haplotype blocks detected for chromosomes 6, 21, and 22. The African-American population panel is denoted by 'A' and the Caucasian population panel is

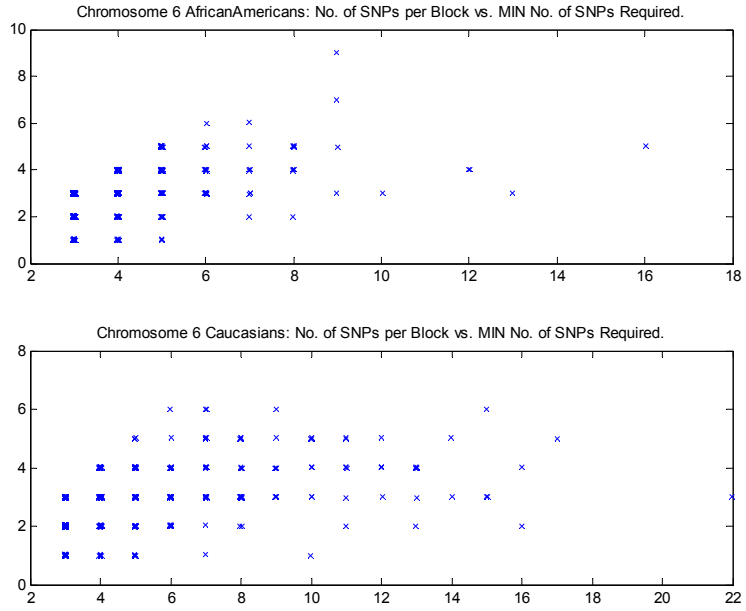
denoted by 'C'. The results show that the mean number of SNPs per block is reduced by on average 36% for the Caucasian population, whereas the reduction for the African-American panel is 25%, using the *lossless* mode of the algorithm. Using the *lossy* mode and accepting <10% information loss, the algorithm accomplishes an additional 17% reduction in the average number of SNPs per block for the African-American panel, versus 19% for the Caucasian samples.

Table 7. Experimental Results Summary

Chr.	Pop.	Total No. of SNPs	Mean Intragenic Spacing Between SNPs (bp)	No. of Haplotype Blocks	Mean Block Size (bp)	Mean No. of SNPs per Block	Mean No. of Min. SNP per block	
							Lossless	<10% Loss
6	A	2,504	10,840	646	23,000	3.88	2.94	2.44
	C	4,009	10,630	883	34,000	4.54	2.86	2.33
21	A	955	7,382	242	14,933	3.95	2.92	2.39
	C	1,555	7,031	336	21,032	4.63	2.88	2.32
22	A	1,405	6,035	350	13,714	4.01	2.99	2.47
	C	1,783	7,760	417	17,505	4.28	2.81	2.27

Figure 1 graphically illustrates the relationship between the original number of SNPs in an LD block (horizontal axis) and the minimum number of SNPs required to genotype the LD block with no loss of information (vertical axis). The thickness of the 'x' corresponds to the number of different blocks found in chromosome 6 with the same properties. This figure shows that irrespective of the original number of SNPs per LD block (up to 18), the maximum number of minimum SNPs levels off at about 6 for the Caucasian population, and is not too different for the African-American panel which has a few outlier blocks with 7 and 9 minimum SNPs.

Figure 1. SNP per block vs. minimum informative SNP subset



5 Discussion

Frequently, when SNPs are initially selected for typing, not much is known about the existence or location of LD blocks, nor about the number and relative frequencies of haplotypes within the blocks. It is therefore typical to "over-sample" the chromosomal region, (i.e., select SNPs as densely as one's budget permits). However, using large numbers of SNPs on a study adds significant cost; therefore the algorithm introduced in this paper can be useful for reducing the set of SNPs to the minimum number required for adequate coverage with no loss of haplotype information. Furthermore, the method can be used to eliminate additional SNPs while minimizing loss of haplotype information.

Previous work attempting to find the minimum SNP sub-set [10,11] have generally focused on complete genes or randomly selected loci, as opposed to LD blocks. In such cases, the number of haplotypes was expected to be higher, and more importantly, the amount of information in the haplotype distribution was expected to be much greater. As a result, these studies were challenged by numerical complexity, and solutions were optimized for relatively small regions and

specific local conditions. Our method computes the global optimum, whether in *lossless* or *lossy* mode, making it more broadly applicable to a variety of purposes.

The method described by Judson *et al* [11] is essentially equivalent to phase II of the *lossy* version of our method. However, it is limited to $k \leq 11$. This is expected, because without the efficient pruning of SNPs performed by phase I of our method, the exponential nature of phase II can require practically infinite execution time. For example, the largest block we encountered consisted of 22 SNPs. With Judson's method, comparing sub-sets of 22 SNPs requires examining over almost 4.2 million combinations. Even using only half of the SNPs, sub-sets 1 to 11, would require examining over 2.4 million combinations, after which it would not be clear whether the solution is optimal, since it could represent a local optimum. Our method uses phase I to quickly reduce the 22 SNPs to a subset of 4 SNPs. As a result, phase II can find the global optimum (3 SNPs in lossless mode or 2 SNPs with less than 10% loss) by examining only 15.

The method described in the on-line supplement to the paper by Johnson et al. [10] appears to strive to compromise the maximization of the information detected by the SNP set with other considerations (e.g., maximization of the individual SNPs' properties). However, there is little explanation of the method details. One example asserts that any haplotype matrix with full rank cannot be pruned. However, we show in Table 1 a haplotype matrix of full rank can be pruned with no loss of information. The on-line supplement also provides executable programs, but the maximum subset size is limited to $k \leq 5$, which would lead to suboptimal results, because we have found the global optimum to be greater than 5 in some blocks.

Previous reports have suggested that the Caucasian population has longer LD blocks than the African-American population [8]. As Table 7 shows, the Caucasian population initially yielded more blocks, and more SNPs per block. However, after "compression" of the SNP sets of each block by our algorithm into the minimum required to represent the information (with no loss), the set size is almost the same for the two populations. Further compression with our *lossy* method, produced a very similar reduction for both populations. We believe that this reflects the arbitrary nature of the criteria to define LD blocks, which was applied uniformly to both populations.

It is important to note that the premise of the algorithm, as well as any other that attempts to reduce the size of the SNP set, is that the DNA sample size for each population is large enough so that the inferred haplotypes adequately represent reality. The accuracy of the computational haplotype inference is dependent on sample size, among other factors. There is always a risk that a SNP whose behavior is identical to another SNP (and thus deemed worthless in terms of new information) for the sample size used, could differentiate an additional haplotype inferred in a larger sample size. Although this risk is low for common haplotypes, it is possible that a rare haplotype can harbor the causative mutation sought and be present in higher frequency in affected individuals. Another factor that must be considered is the possibility of experimental errors that can eliminate data points

and thus render suboptimal the minimum SNP subset. Therefore, it will always be necessary to supplement the minimum SNP subset with additional SNPs to enhance robustness. An optimal procedure for choosing additional SNPs that increase robustness, as well as the ability to bias the selection of the minimum SNP subset as a function of cost or availability of specific SNP assays, is currently under investigation.

6 Conclusions

We have described a simple algorithm that can select a minimum subset of SNPs without loss of haplotype information, or an even smaller subset with some acceptable loss of information. In a practical example encompassing 3 human chromosomes, we were able to reduce the SNP set by 25% for the African American population and by 36% for the Caucasian population with no loss of haplotype distribution information. It is clear that for an arbitrarily large genomic segment with numerous SNPs and haplotypes, the number of computations required by our algorithm would grow exponentially even after our phase I heuristic, and become unpractical. This situation would require different approaches to reduce the combinatorial complexity of the problem. However, for most practical situations our algorithm shall suffice and produce optimal results in a reasonable time, even allowing for the real-time calculation of minimum SNP subsets for haplotype blocks.

7 Acknowledgments

We are indebted to Charles Scafe, Heinz Hemken, Yu Wang, Marion Webster, Lewis Wogan, Lily Xu, Xiaoqing You, and Janet Ziegle for their role in the selection of SNPs, design of assays, genotyping of DNA samples, and management of the experimental results used in this work. Thanks are also due to Eugene Spier, Sorin Istrail, and Andrew Clark for many helpful discussions, and Mignon Fogarty for assistance with the manuscript.

References

1. Abecasis, G.R., Cherny, S.S., Cookson, W.O. and Cardon, L.R. *Nat Genet* 30:97-101 (2002).
2. Abecasis, G.R., Noguchi, E., Heinzmann, A., Traherne, J.A., Bhattacharyya, S., Leaves, N.I., Anderson, G.G., Zhang, Y., Lench, N.J., Carey, A., Cardon, L.R., Moffatt, M.F. and Cookson, W.O. *Am J Hum Genet* 68:191-197 (2001).
3. Ardlie, K., Liu-Cordero, S.N., Eberle, M.A., Daly, M., Barrett, J., Winchester, E., Lander, E.S. and Kruglyak, L. *Am J Hum Genet* 69:582-589 (2001).
4. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. *Nat Genet* 29:229-232 (2001).
5. De La Vega, F.M., Dailey, D., Ziegle, J., Williams, J., Madden, D. and Gilbert, D.A. *Biotechniques* Suppl:48-50, 52, 54 (2002).
6. Drysdale, C.M., McGraw, D.W., Stack, C.B., Stephens, J.C., Judson, R.S., Nandabalan, K., Arnold, K., Ruano, G. and Liggett, S.B. *Proc Natl Acad Sci U S A* 97:10483-10488 (2000).
7. Excoffier, L. and Slatkin, M. *Mol Biol Evol* 12:921-927 (1995).
8. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. *Science* 296:2225-2229 (2002).
9. Jeffreys, A.J., Kauppi, L. and Neumann, R. *Nat Genet* 29:217-222 (2001).
10. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F., Twells, R.C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S.C., Clayton, D.G. and Todd, J.A. *Nat Genet* 29:233-237 (2001).
11. Judson, R., Salisbury, B., Schneider, J., Windemuth, A. and Stephens, J.C. *Pharmacogenomics* 3:379 (2002).
12. Ke, X., Tapper, W. and Collins, A. *Bioinformatics* 17:581-586 (2001).
13. Kruglyak, L. *Nat Genet* 22:139-144 (1999).
14. Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R. and Lander, E.S. *Nature* 411:199-204 (2001).
15. Shannon, C.E. *Bell System Technical J.* 27:379-423 (1948).
16. Xu, C.F., Lewis, K., Cantone, K.L., Khan, P., Donnelly, C., White, N., Crocker, N., Boyd, P.R., Zaykin, D.V. and Purvis, I.J. *Hum Genet* 110:148-156 (2002).