SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS IN APPLICATION TO FINE GENE MAPPING

by

Manish Sampat Pungliya

A Thesis submitted to the Faculty of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

In

Biology

By

_____

May 2001

APPROVED BY:

_____	_____
Dr. Julia Krushkal , Major Advisor	Dr. Elizabeth Ryder, Advisor on Record
_____	_____
Dr. Carolina Ruiz, Thesis Committee	Dr. David Adams, Thesis Committee

_____

Dr. Ronald Cheetham, Head of the Department

ABSTRACT

Single nucleotide polymorphisms (SNPs) are single base variations among groups of individuals. In order to study their properties in fine gene mapping, I considered their occurrence as transitions and transversions. The aim of the study was to classify each polymorphism depending upon whether it was a transition or transversion and to calculate the proportions of transitions and transversions in the SNP data from the public databases. This ratio was found to be 2.35 for data from the Whitehead Institute for Genome Research database, 2.003 from the Genome Database, and 2.086 from the SNP Consortium database. These results indicate that the ratio of the numbers of transitions to transversions was very different than the expected ratio of 0.5. To study the effect of different transition to transversion ratios in fine gene mapping, a simulation study was performed to generate nucleotide sequence data. The study investigated the effect of different transition to transversion ratios on linkage disequilibrium parameter (LD), which is frequently used in association analysis to identify functional mutations. My results showed no considerable effect of different transition to transversion ratios on LD. I also studied the distribution of allele frequencies of biallelic SNPs from the Genome Database. My results showed that the most common SNPs are normally distributed with mean allele frequency of 0.7520 and standard deviation of 0.1272. These results can be useful in future studies for simulating SNP behavior. I also studied the simulated data provided by the Genetic Analysis Workshop 12 to identify functional SNPs in candidate genes by using the genotype-specific linkage disequilibrium method.

ACKNOWLEDGMENTS

I am very thankful to many people who were directly and indirectly involved in my completion of my thesis. This is the only way to express my sincere gratitude towards them. First of all I would like to thank my advisor Dr. Julia Krushkal for giving me the unique opportunity of working in the exciting field of Bioinformatics. Her continuous encouragement and support made this thesis as one of the most important things to be cherished throughout my life.

I extend my special gratitude to Dr. Elizabeth Ryder, Dr. David Adams and Dr. Carolina Ruiz for their sincere suggestions and helpful discussions throughout my years of graduate study. Special thanks goes to Dr. Matthew Ward for providing me with the program 'Scansort'. I am also thankful to Christopher Shoemaker and Michael Sao Pedro who did the data conversion and provided me with the formatted data for Genetic Analysis Workshop 12 (GAW12). I would also like to make a mention of my friend Raju Subramanian for guiding me in writing algorithms in C++. Analysis of the GAW12 data presented in this thesis was funded by the grant "Computational algorithms for analysis of genomic data" from the Research Development Council of Worcester Polytechnic Institute. I am also thankful to the National Institutes of Health grant GM31575 to GAW12.

Finally, I express my gratitude to my parents for their unconditional love and moral support throughout my graduate study. And last but not the least I would like to thank all my friends at WPI who have made my stay here a memorable thing in my life.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

Page no.

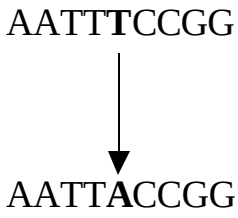
INTRODUCTION

At the present time, we are at a stage where we can read nearly the entire genetic code of the human genome, as recently a rough draft of the human genome sequence has been determined (International Human Genome Sequencing Consortium 2001; Venter et. al 2001). It represents the sequence of A, G, C, and T letters that are symbols for nucleotides. The specific sequence of these nucleotides constitutes all our genes that have specific characteristics and expression in the human being. These genes are responsible for various physical, physiological and pharmacological activities in the body. Since mutations in these genes are often the cause of many heritable diseases, it is sometimes necessary to find specific genetic mutations responsible for a particular disease. This is one of the current aims of the Human Genome Project.

Until now, genome analysts have concentrated their efforts on finding the similarities between different individuals, and they have found that 99.9 percent of anyone's genes perfectly match those of another individual (Brown 2000). But the remaining 0.1 percent of the genes varies among individuals. It is these nucleotide variations that are of interest to many researchers, as these variations might change the properties of that particular gene. Even a simple single nucleotide polymorphism (SNP) in a gene sequence can change the property of that gene and cause a disease, as differences in DNA of that gene could change the phenotype. Below I describe the properties of SNPs in more detail.

Single Nucleotide Polymorphism

SNPs are the most common single base variations in the human population. A SNP is a variation where two alternative bases occur at appreciable frequency (the frequency of each base is above 1% in a population) (Lander et. al 1998). Most of these variants are neutral, but some are functional. One of the important goals of genetic analysis is to identify those SNPs and SNP variants (alleles), which are associated with a disease. For example, consider the following nucleotide sequence,



The change from T to A is considered a SNP provided both alleles are present in more than 1 percent in the population. SNPs are often binary, i.e. they most often have only two alleles. They are less susceptible to mutations than microsatellite repeat markers. A microsatellite is a short sequence of repeated nucleotides in a genome e.g.: AATGAATGAATG-----, where AATG is repeated a variable number of times. Due to their stability, SNPs are very useful for studying human evolutionary history.

Importance of SNPs

The occurrence of SNPs is approximately one in every 1000-2000 base pairs, and the total number of SNPs in the human genome estimated by November 2000 is 1,433,393 (The International SNP Map Working Group 2001) and 2,104,820 (Venter

et. al 2001). These polymorphisms are present in coding as well non-coding regions. Less than 1% of all the SNPs are present in the protein-coding regions of the genome (Venter et. al 2001). This suggests that a very small proportion of SNPs may be responsible for phenotypic variation.

SNPs in association analysis

Association analysis and linkage analysis are methods for gene mapping of complex human disorders. In the case of association analysis, one tries to find an association between a marker locus (could be a SNP) and a disease locus (may be or may not be a SNP) from the genetic data of the human population (Risch and Merikangas 1996). Such analysis tests for association of loci at short distances apart and attempts to identify individual mutations. SNPs are therefore very useful in association analysis, as SNPs are so frequent and close to each other that the loci stay associated even after much recombination. In contrast, in linkage analysis, one tests for genetic linkage between a disease locus and a marker locus which is generally a microsatellite marker (Kruglyak et al. 1996). The distances detected in linkage analysis are generally much larger than those in association analysis. In linkage studies, one needs to have a family with a certain proportion of individuals with the disease. Linkage analysis takes into account all the available information from a pedigree of a chromosomal region and a trait that cosegregate. Such sample data are often difficult to obtain. This disadvantage is overcome by using association analysis, as it does not require the pedigree information for a trait.

Association analysis does not involve the creation of a model based on the pedigree information. In fact, it is based on the association that exists between two phenotypes or loci when they occur together in a group of individuals more often than expected by chance. Association may or may not be due to linkage, as in linkage one tries to find an association between two loci in a pedigree.

Association analysis is a very useful tool in mapping genes for complex phenotypes. An etiology of a complex phenotype can be associated with factors such as marker locus, disease gene, other disease genes, environment, and cultural factors. All of these factors interact with each other in a very complex way, resulting in the expression of that phenotype. The main purpose of an association study to map disease genes and find causative mutations is to minimize the effect of other genetic, environmental, and cultural factors, and in turn increase the correlations of marker and disease gene with the complex phenotype. By increasing the number of marker loci, one can improve the efficacy of association studies in finding functional mutations, as there is always an increased probability of finding at least one marker locus associated with the disease gene.

The disadvantage of a linkage study compared to an association study lies in its basic requirement of number of affected individuals necessary to detect the linkage to a complex trait (Risch and Merikangas 1996). This number is very large when the proportion of affected individuals in a population is small. In contrast, the requirement of such number is vastly less in association methods as one tries to study association between a single locus or multiple loci together with the disease locus in affected/unaffected individuals. One would have better understanding of the

association of a disease locus with other loci that lie on the same chromosome or on different chromosomes when genome-wide association tests are performed.

Linkage disequilibrium

To investigate the association between the two loci/multiple loci, one of the ways of evaluating association is to use a parameter called linkage disequilibrium. Linkage disequilibrium mapping is a frequently used analytical approach involving SNPs. Linkage disequilibrium (LD) is defined as a nonrandom association between SNPs in proximity to each other (Terwilliger and Weiss 1998). So if thousands of SNPs are mapped over the entire genome, then through LD, associations could be established between any of the susceptibility regions of the gene and a particular SNP marker. Two sites are said to be in linkage disequilibrium if the presence of one marker locus at one site enhances the predictability of another locus on the same chromosome or different chromosome. This indicates that these two sites are associated, and it helps in mapping the disease gene, as one of the loci could be the disease locus, or might be associated with the disease locus so that it is transmitted in a family.

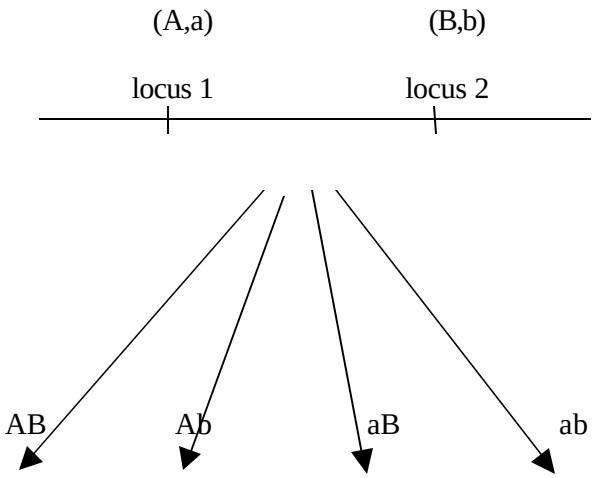


Figure 1: Linkage disequilibrium phenomena (see text for explanation)

Suppose there are two loci, 1 and 2, that are close to each other (Figure 1). Loci 1 and 2 may be on the same or on different chromosomes with two alleles each: A,a and B,b respectively. As a result, there are four haplotypes possible: AB, Ab, aB, and ab. If allele A has a frequency of ρ_A in the population and allele B has a frequency of ρ_B, then haplotype AB would have frequency $\rho_A\rho_B$ in the absence of association, as they would occur independent of each other in the population. If alleles A and B are associated, then the frequency of haplotype AB would be $\rho_A\rho_B + D$, where D is the measure of the strength of LD between the two loci 1 and 2.

$$D = \rho_{AB} - \rho_A\rho_B$$

If allele B at locus 2 is a disease causing locus, then the frequency of allele A would be much higher in affected individuals than in unaffected individuals because of its association with B. This method of association analysis is used in case-control studies. In case-control studies, there are two samples, one of cases (with disease) and one of control (no disease) individuals. The two groups are further classified according to one marker on the basis of its presence or absence. By performing chi-square analysis, one can estimate the significance of the association of the marker with the disease. This method can be extrapolated to markers having more than one allele (reviewed by Elston 1998).

Thus if one has a large map of such marker loci over the entire genome, one can test them to find the linkage disequilibrium with the disease locus of interest. The markers are generally polymorphisms like microsatellites, or single nucleotide polymorphisms within or outside the gene, but lying close to it so that they could be associated with the disease locus. This is the basis for linkage disequilibrium mapping

of common disease genes. Linkage disequilibrium mapping is considered as the indirect strategy (Kruglyak 1999) of association studies, as it involves testing for associations between the disease locus and nearby polymorphism which is not physically linked to the disease gene. It is done by using a dense map of polymorphic markers across the genome which can be tested for association with the disease locus. Biallelic single nucleotide polymorphisms are the markers of choice for this kind of analysis because of their abundance in the genome and low mutation rates compared to microsatellites. This is particularly important for whole-genome association studies for identification of complex disease genes (Schafer and Hawkins, 1998). Lai et al (1998) have proved the feasibility and importance of creating such SNP maps for identification of genes of interest. The importance of creating a dense map of SNPs is that it is useful in linkage disequilibrium mapping so that a susceptibility allele and a marker lie within the range of linkage disequilibrium (McCarthy and Hilfiker 2000). In the second half of this thesis, I have presented the application of association analysis to a simulated data set.

Recently, it has been shown that the average extent of linkage disequilibrium in the general human population is approximately 3kb; thus roughly 500,000 SNPs (as markers) may be needed for systematic whole-genome linkage disequilibrium studies (Kruglyak 1999) so that one would have good placing of SNP markers (in the range of 3kb) over the entire genome. LD tends to decrease when the distance between the markers is in the range of 10-100kb, as there is a high probability of recombination and genetic drift with increased distance (McCarthy and Hilfiker 2000).

Application of SNPs in Pharmacogenomics

Linkage disequilibrium mapping has also been important in pharmacogenomic studies (McCarthy and Hilfiker 2000). The use of LD mapping using SNPs has gained a lot of importance, as it provides the necessary information about the drug response in a genetically heterogeneous population used in clinical trials. It will help to uncover the secret of why some drugs are effective in certain people and not in others. These small variations may cause differences in drug response among patients as they alter gene expression. Some drugs may have a beneficial effect on some patients, but might prove harmful to others. So in the future a simple genetic test may determine whether an individual can be treated effectively by a given drug. Such application of genomics to pharmaceuticals is categorized as Pharmacogenomics, which is the branch of genomics addressing molecular pharmacology and toxicology. Hopefully this new branch will help reduce the cost of drug development (presently it is in the range of $400-500 million), at the same time increasing the speed of the development process (Rothberg, Ramesh and Burgess 2000). Pharmacogenomics involves detecting and cataloguing SNPs in genes responsible for drug response. This will uncover the variability of individual drug responses, and facilitate appropriate patient selection during clinical trials and the aftermarket of a particular drug.

Other applications

SNPs are present in any part of the human genome. If they are present in the coding or regulatory part of the gene, a SNP variant might alter gene function, and

thus may be related to the disease progression. However, if a SNP is not functional it is still important in mapping studies, as it might be present very close to the disease gene. This is an important basis of linkage disequilibrium analysis as one can study the presence of that particular SNP in a population to find out its association with the disease gene in the population.

Presence of SNPs

The presence of SNPs is related to nucleotide substitutions in DNA sequences. To study the process of nucleotide substitutions, several mathematical models based on the probability of nucleotide substitutions have been proposed in the literature (reviewed by Li 1997). The two simplest and most frequently used models are

1) Jukes and Cantor's (1969) one-parameter model and
2) Kimura's (1980) two-parameter model

1) *Jukes and Cantor's one-parameter model*

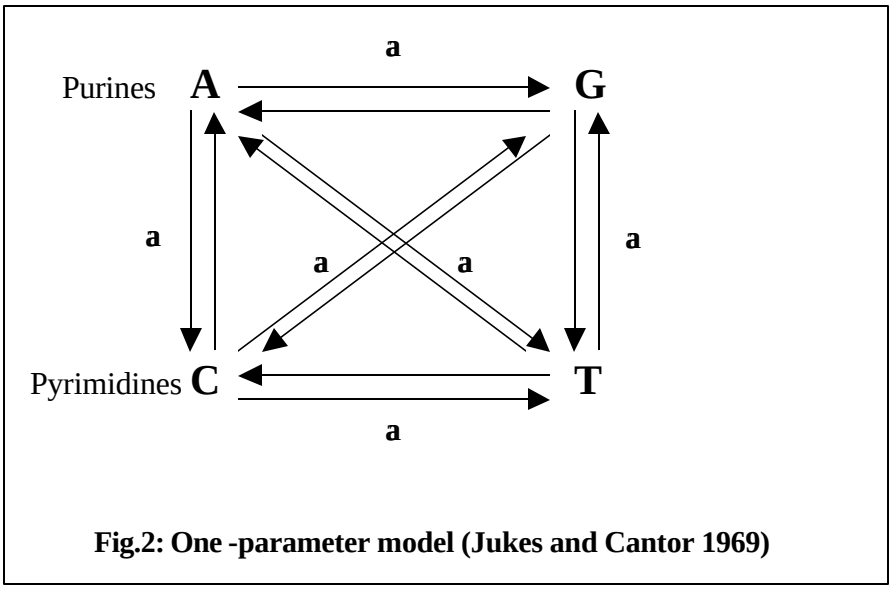


Fig.2: One -parameter model (Jukes and Cantor 1969)

This model assumes the substitution of nucleotides in DNA sequences is a random process with all possible changes occurring with equal probabilities. For example, nucleotide A can change to nucleotides T or C or G with equal probability. Let α be the rate of substitution per time in each of the three possible directions of nucleotide A. In this model, the rate of substitution for each nucleotide is 3α. Because only one parameter (α) is involved in this model, it is called a one-parameter model.

The basic assumption of the one-parameter model (that all nucleotide substitutions occur randomly) is unrealistic in most cases. For example, *transitions* (changes between A and G, or between C and T) are generally more frequent than *transversions* (changes between A and C or T, and between G and C or T, and vice versa). To take this fact into account, a two-parameter model was proposed by Kimura (1980).

2) *Kimura's (1980) two-parameter model*

This model attempts to account for different frequencies of transitions and transversions (Figure 3). In this model, there are two parameters involved. One is α, which is the rate of transition (changes within purines or pyrimidines) and the other is β, which is the rate of transversion (changes between purines and pyrimidines). If transitions and transversions occur with equal rates ($\alpha = \beta$), the expected ratio between all possible transitions and all possible transversions from the two-parameter model would be,

$$N_{transitions}/N_{transversions} = \alpha/2\beta = 0.5$$

as there are four possibilities of transitions and eight possibilities of transversions.

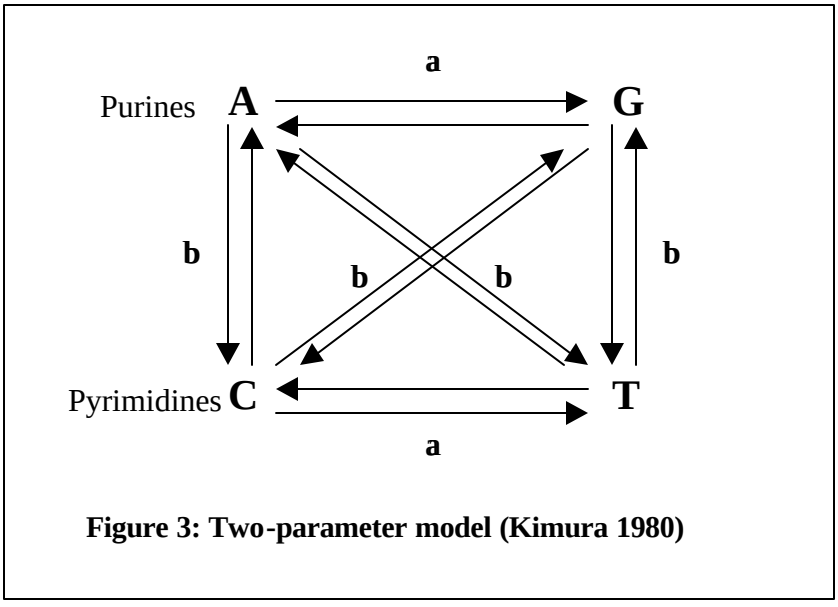


Figure 3: Two-parameter model (Kimura 1980)

One- and two-parameter models of nucleotide substitutions may not apply to a relatively short evolutionary time and also the possibility of occurring four types of transitions (or eight types of transversions) with the same rate may not be true. There

are many other substitution models, some of which were described by Li (1997). These models are more complicated mathematically and take into account complex nucleotide substitution matrix, as some of these higher models are based on six-parameters or nine-parameters substitution matrix (as these models consider different rates of substitutions within transitions or transversions). Felsenstein's (1981) suggested the equal-input model, which is based on the equal rate of substitution of one nucleotide with the other three nucleotides (similar to the one-parameter model). Hasegawa et al. (1985) suggested a newer model (Model HKY85) with the addition of additional substitution parameters and base frequencies. All these models are mathematically complicated to compute. Felsenstein's, Jukes and Cantor's and Kimura's models are special cases of Hasegawa's model. Felsenstein's model takes into account the ratio of transition to transversion equal to 1.0. Jukes and Cantor's model (one-parameter model) considers equal nucleotide frequencies and transition to transversion ratio to 1.0, while Kimura's model (two-parameter model) considers only equal nucleotide frequencies.

THESIS OBJECTIVES

In the first part of my work, I investigated the exact ratio of transition to transversion in SNPs from the publicly available databases and compared it with the expected ratio of 0.5 ($\alpha = \beta$). I also studied the distribution of the most common SNPs from one of the data sets explained in the next section.

In previous studies, most association methods that used SNPs did not take into account the possible differences between the rates of transitions and transversions. These studies considered the transition and transversion occurring with equal rates as in the one-parameter model. To investigate the validity of this approach, after finding the actual proportion of transitions/transversions in the public databases, I studied the effect of different transition to transversion ratios (including observed and expected) on fine gene mapping using computer simulations.

In the second part of this thesis, I describe how association methods are applied for fine gene mapping using simulated SNP data from Genetic Analysis Workshop 12 (GAW 12).

Part I

Investigation of the effect of the transition to transversion ratio

METHODS

DATA COLLECTION

To find out the ratio of transition to transversion in the SNP data, the SNPs from human chromosome 6 were collected. I selected human chromosome 6 specifically because it has a number of genes associated with human diseases such as breast cancer, hypertension (Krushkal et al. 1999), maple syrup urine disease (Nobukuni et al. 1991), diabetes mellitus (Davies et al. 1994), psoriasis (Balendran et al. 1999), and schizophrenia (Cao et al. 1997)[1]. I used three publicly available databases to collect the data:

1) Whitehead Institute of Biomedical Research/MIT Center for Genome Research[2]
2) The Genome Database [3]
3) The SNP Consortium Ltd [4]

1) Whitehead Institute of Biomedical Research/MIT Center for Genome Research

This center started work on SNPs with the intention of developing a dense map of SNPs (about 100,000 in number) (Collins, Guyer, and Chakravarti 1997). This database has an anonymous list of SNPs from both coding and noncoding regions of the genome. The SNPs are listed on a

[1] http://www.ncbi.nlm.nih.gov:80/htbin-post/Omim/getmap?d2417
[2] http://www-genome.wi.mit.edu
[3] http://www.gdb.org
[4] http://snp.cshl.org

genetic distance map (cM) with corresponding sequence tagged sites (STSs) (Lander et al. 1998). The output is shown in Appendix 1.

2) *The Genome Database (GDB)*

This database is a result of an international collaboration in support of the Human Genome Project. It has the list of genes for each chromosome. In this database, the list of SNPs, labeled as point variations, is provided for individual genes. Another advantage of the GDB is that it also provides the allele frequencies for many polymorphisms, which is useful in finding the allele frequency distribution for SNPs in human population. The data collected from this database are in Appendix 2.

3) *The SNP Consortium Ltd*

This is the most comprehensive SNP database. The SNP Consortium Ltd. is a non-profit foundation organized for the purpose of providing public genomic data. This database was one of the goals of Genome II, the next phase of the Human Genome Project (Brower 1998). Its mission was to develop up to 100,000 SNPs distributed evenly throughout the human genome and to make the information related to these SNPs available to the public without intellectual property restrictions. SNP screening is performed at the three major genomic centers (Washington University at St. Louis, The Whitehead Institute for Biomedical Research and The Sanger Center), by using a panel of unrelated, anonymous individuals. The fifth release (April

2000) of this database consisted of 296,990 SNPs from all the chromosomes. In the present thesis, I collected data from the fifth release of this database. An example of this data is shown in Appendix 3.

ALLELE FREQUENCY DISTRIBUTION

SNPs are generally present in biallelic form, but sometimes they are present in more than two forms. I investigated the distribution of the frequencies of the most common alleles in the SNP data obtained from The Genome Database (GDB). I only considered biallelic SNPs in this data set. The Genome Database has given the allele frequency for biallelic polymorphisms. I considered those allele frequencies that are more than 0.5 (More than 0.5 indicates frequent occurrence of that allele).

To check the symmetry and normality of the data, I performed a kurtosis plot analysis to test if the data are normally distributed and also to verify that the distribution is not skewed to the left or right of the mean. By using symmetry and kurtosis measures, the normality of the allele frequency data was assessed. The null hypothesis of population normality was tested by using the test statistic,

$$K^2 = Z^2_{g1} + Z^2_{g2}$$

where Z^2_{g1} and Z^2_{g2} are the parameters for symmetry and kurtosis. K^2 has the same distribution as χ^2 and it is the parameter for assessing normality using symmetry (Z_{g2}) and kurtosis (Z_{g2}) measures. Z_{g1} is the parameter for testing the population's symmetry. But not all symmetrical distributions are normal and to check the normality, kurtosis measure is used. Here, Z_{g2} is the population kurtosis parameter. The calculations of Z_{g2} and Z_{g2} are as follows.

$$Z_{g1} = E \ln(F + \sqrt{F^2 + 1})$$

$$F = A / \sqrt{2 \div C - 1}$$

$$E = 1 / \sqrt{\ln D}$$

$$D = \sqrt{C}$$

$$C = \sqrt{2(B - 1)} - 1$$

$$B = \frac{3(n^2 + 27n - 70)(n + 1)(n + 3)}{(n - 2)(n + 5)(n + 7)(n + 9)}$$

$$A = \sqrt{b_1} \sqrt{(n + 1)(n + 3) / 6(n - 2)}$$

where, n = number of SNPs considered and $\sqrt{b_1}$ = beta measure of symmetry.

$$Z_{g2} = \frac{1 - \frac{2}{9K} - \sqrt[3]{L}}{\sqrt{\frac{2}{9K}}}$$

$$L = \frac{1 - \frac{2}{K}}{1 + H\sqrt{\frac{2}{K - 4}}}$$

$$K = 6 + \frac{8}{J}\left[\frac{2}{J} + \sqrt{1 + \frac{4}{J^2}}\right]$$

$$J = \frac{6(n^2 - 5n + 2)}{(n + 7)(n + 9)} \sqrt{\frac{6(n + 3)(n + 5)}{n(n - 2)(n - 3)}}$$

$$H = \frac{(n - 2)(n - 3) |g_2|}{(n + 1)(n - 1)\sqrt{G}}$$

$$G = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

g_2 is the sample statistic for the measure of kurtosis and for further calculations of g_2 refer (Zar 1999 (b)).

SIMULATION STUDY

To study the effect of different ratio of transitions to transversions, I generated sequence data by using a computer simulation program called TREEVOLVE (Grassly and Rambaut, unpublished) [1].

Principle of TREEVOLVE

TREEVOLVE simulates DNA sequences based on Kingman's (1982a,b,c) coalescent approach (cited in Kingman 2000). The coalescent model simulates the evolutionary history of a gene backwards in time until it reaches a point of most recent common ancestor (MRCA) for that gene. This generates a tree for a particular gene in the population sample. The tracing backwards in time is based on a Markov chain (Kingman 2000). An example shown in the following figure (figure 4),

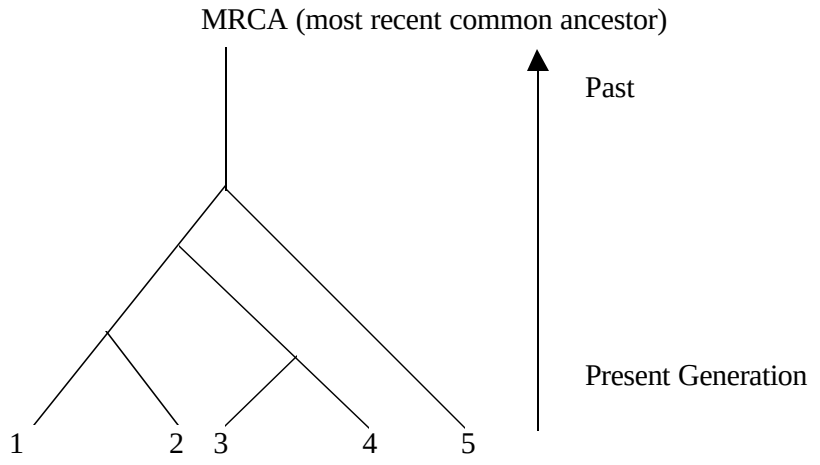


Figure 4: Simulation of population history by TREEVOLVE.

[1] http://evolve.zoo.ox.ac.uk

According to this figure, if there are n members of a particular generation, then TREEVOLVE traces their family tree backward through time until it coalesces (i.e. the lineages find a common ancestor). The number n is reduced by one each time coalescence occurs so that next time (n-1) lines (members) will be traced back until they coalesce and so on, until the number of lines is reduced to one (MRCA). While simulating the tree, the program does not take into account any sequence information from the present-day generation individuals. The lineages are joined randomly back in time. Depending on the number of replicates specified in the input parameter file, a new random tree is generated in each replication. Each tree corresponds to a separate replicate. After generating each tree, the molecular sequences are simulated down the genealogy under the substitution model specified by a user.

Substitution model

For each tree, TREEVOLVE generates present-day sequences. The number of sequences depends on the input parameter file shown in figure 6. The sequences differ completely between the replicates. After generating each coalescent tree, the DNA sequences are generated according to the genealogy down the line with time (forward direction) starting at the most recent common ancestor (MRCA) as shown in figure 5. TREEVOLVE has an option of using different substitution models based on the type of analyses and the choice of variables in the sequence generation. Two of the models are F84 (Felsenstein and Churchill 1996) and HKY85 (Hasegawa et al. 1985) that takes into account the transition to transversion ratio and also the base frequencies as parameters, but differing in their values. I used the one- and two-

parameter models which can be presented as a partial case of F84 model with equal base frequencies in all simulations. In the case of the one-parameter model which assumes equal rate of transitions and transversions, I used the transition to transversion ratio ($\alpha/2\beta$) of 0.5. When I used the two-parameter models, the ratio of $\alpha/2\beta$ was varied between 1, 2.35, 3 and 5. First of all, for a given set of simulation parameters, TREEVOLVE generates 200 trees (200 replicates). For each tree, it generates present-day sequences depending on the selected substitution model. Each tree corresponds to a separate replicate and these sequences differ completely between the replicates. While simulating these sequences TREEVOLVE also takes into account the mutation rate and recombination rate. As a result, some polymorphic sites within each replicate are produced. An example of an input parameter file for TREEVOLVE is shown in figure 6, and the complete list of parameters for TREEVOLVE used in this study is shown in table 1. The simulations were carried out under no recombination, and also with recombination rates varied between 10^{-8}, 3×10^{-5} and 10^{-3} as suggested by other studies (Zollner and Haeseler 2000) keeping all other parameters constant. I used no population subdivision (m = 0 in figure 6), no migration, and no exponential growth (e = 0.0 in figure 6) to minimize the effect of admixture, allelic heterogeneity and environment. The mutation rate was 10^{-7}, which was constant in all simulations, as suggested by other studies (Zollner and Haeseler 2000). I used a sequence length of 1000bp with 200 sequences (same as sample size) for each file. The substitution model used was either a one- or two-parameter model as described earlier. The simulations were performed as shown in table 2. A sample run of TREEVOLVE is shown in figure 7.

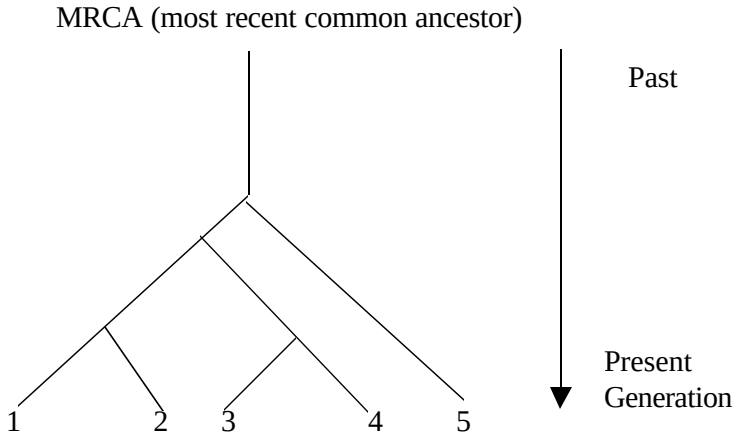


Figure 5: Se quence simulation by TREEVOLVE after the population history has been simulated.

```
BEGIN TVBLOCK
[sequence length] l1000
[sample size] s200
[mutation rate] u0.0000001
[number of replicates] n1
[substitutionmodel] vF84 t0.5 [f0.25,0.25,0.25,0.25] [r0.1667,0.1667,0.1667,0.1667,
0.1667, 0.1667]
[output coalescent times] oCoal.Times
[diploid]
[generation time/variance in offspring number] b1.0

*PERIOD 1
   [length of period] t100000.0
   [population size] n100000 e0.0
   [subdivision] d1 m0
   [recombination] r0.0
*END
```

Figure 6: An example of input parameter file for TREEVOLVE.

Table 1: Input parameter values for TREEVOLVE used in this study.

Parameters	Value
Sequence length, l	1000 bp
Sample size, s	200 sequences
Mutation rate, u	0.0000001
Number of replicates, n	200
Substitution model, m	One- and two-parameter model
Transition/transversion ratio, t	Varies (0.5 to 5.0)
Base frequencies, f	Equal (0.25)
Rate heterogeneity	None
Ploidy	Diploid
Generation time/var. in offspring no., b	1.0
Run time for population dynamic model,t	100000.0
Effective population size, n	100000
Exponential growth rate, e	0.0
Number of demes, d	1
Migration rate, m	0.0
Recombination rate	Varies from 0.0 to 10^{-3}

Sequence1 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG
Sequence2 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG
Sequence3 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG
Sequence4 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG
Sequence5 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG
Sequence6 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG
Sequence7 TCAGGAACAACAGCTAATGAGCTTATATTTTCATGACATAACG

Figure 7: A sample output of TREEVOLVE for sample size of 7 with sequence length of 43. There are no SNPs in nucleotide sequences shown because only few initial bases are shown for each sequence for illustration.

Table 2: The table indicating different simulations performed for different parameters.

Transition / Transversion	Recombination rate			
	0.0	10^{-8}	3×10^{-5}	10^{-3}
0.5	√	√	√	√
1.0	√	√	√	√
2.35	√	√	√	√
3.0	√	√	√	√
5.0	√	√	√	√

LINKAGE DISEQUILIBRIUM ANALYSIS OF SIMULATED DATA FROM TREEVOLVE

Different types of software are used in order to compute values of LD (linkage disequilibrium) using the sequence data. In the present study, program Arlequin[1] (Schneider et al. 1997) was used for calculating the LD in the data simulated by TREEVOLVE. It is versatile software available for analysis of genetic data of various different types, including RFLPs, DNA sequence, and microsatellite data. It allows one to perform a number of statistical tests using population data, including linkage disequilibrium analysis. Arlequin has a graphical interface that is user-friendly. Its other advantage is that one can run a number of input files simultaneously with the same or different parameter lists by creating a batch file, thus reducing the analysis time of the user. An example of an input file of Arlequin used in this study is shown in figure 8.

During analysis by Arlequin, there were 200 input files (replicates from TREEVOLVE) for each set of parameters. All these input files were analyzed together by creating a batch file (figure 9).

[1] http://anthro.unige.ch/arlequin/

```
[Profile]
Title="SNPAnalysis Ratio5 r0.0"
NbSamples=1

GenotypicData=0
 DataType=DNA
 LocusSeparator=NONE
 MissingData='?'

[Data]

[[Samples]]

SampleName="Replicate1"
SampleSize=10
SampleData= {
sequence1 1 ACAACAGCTAATGAGCTTATATTTTCATGACATAACGGGAAC
sequence2 1 ACAACAGCTAATGAGCTTATATTTTCATGACATAACGGGAAC
sequence3 1 ACAACAGCTAATGAGCTTATATTTTCATGACATAACGGGAAC
sequence4 1 ACAACAGCTAATGAGCTTATATTTTCATGACATAACGGGAAC
sequence5 1 ACAACAGCTAATGAGCTTATATTTTCATGACATAACGGGAAC
}
```

Figure 8: Sequence.arp – An input file for Arlequin for the calculation of LD. The field under SampleData is the output of treevolve with transition to transversion ratio of 0.5 and recombination rate of 0.0 (Only first five sequences are shown with only forty nucleotides in each sequence. The data simulated by TREEVOLVE had 200 such sequences with 1000 nucleotides in each sequence for each set of parameters).

replicate1.arp
replicate 2.arp
replicate 3.arp
replicate 4.arp
replicate 5.arp
replicate 6.arp
replicate 7.arp
replicate 8.arp
replicate 9.arp
replicate 10.arp

Figure 9: Batch.arb – An example of a Batch file for Arlequin. During analysis, 100 to 200 replicates were analyzed in each batch file.

Arlequin calculates LD by doing pair-wise comparisons between different SNPs within each replicate. The LD is calculated by using the formula,

$$D_{ij} = \rho_{ij} - \rho_i \rho_j,$$

where ρ_{ij} is the frequency of the haplotype having allele i at the first locus and allele j at the second locus, and ρ_i and ρ_j are the frequencies of alleles i and j in the replicate respectively.

After obtaining the linkage disequilibrium (D) values for all the input parameter files, I calculated the mean D and standard deviation (SD) values for all the combined 200 replicates for each parameter file (i.e. for different transition to transversion ratios at different recombination rates).

RESULTS

SUMMARY OF DATA COLLECTED

In order to test whether the ratio of transitions to transversions is 0.5 ($\alpha/2\beta$) as predicted, I collected SNP data from three databases. The summary of the SNP data collected is given in table 3. The results indicate that the transition to transversion ratio in the real data is at least four times higher than the expected ratio of 0.5.

Table 3: Data Summary from all the three databases

Dataset	Number of SNPs	$\alpha/2\beta$
The Whitehead Institute of Genome Research	146	2.35
The Genome Database	36	2.003
The SNP Consortium	5102	2.086

1) *The Whitehead Institute of Genome research*

The SNP data is shown in appendix 1. The distribution of SNPs in this data is shown below in figure 10. The observed ratio of $\alpha/2\beta$ is 2.35, which is approximately five times higher than the expected ratio of 0.5. The total number of biallelic SNPs in this dataset was 146.

Proportion of SNPs:

			Proportion
A/G, G/A:	25+24 =	49	0.336
A/C, C/A:	10+4 =	14	0.096
A/T, T/A:	3+5 =	08	0.055
G/C, C/G:	6+7 =	13	0.089
G/T, T/G:	5+3 =	08	0.055
C/T, T/C:	24+30 =	54	0.370

		Total = 146	

Proportion of transitions and transversions:

	Proportion
Transitions-	
A/G + G/A + C/T + T/C = 103	0.705
Transversions-	
A/T + T/A + A/C + C/A + G/T + T/G + G/C + C/G = 43	0.300

Figure 10: SNP distribution in the chromosome 6 data from The Whitehead Institute of Genome Research.

2) *The Genome database*

The SNP data from this database are shown in appendix 2. The distribution of SNPs in this database is shown in figure 11. The observed ratio of $\alpha/2\beta$ is 2.003, which is four times higher than the expected ratio of 0.5. There were 36 biallelic SNPs in this dataset.

Proportion of SNPs:

			Proportion
A/G, G/A:	06+07 =	13	0.361
A/C, C/A:	01+01 =	02	0.055
A/T, T/A:	01+01 =	02	0.055
G/C, C/G:	04+01 =	05	0.139
G/T, T/G:	02+01 =	03	0.083
C/T, T/C:	10+01 =	11	0.305

	Total =	36	

Proportion of transitions and transversions:

	Proportion
Transitions -	
A/G + G/A + C/T + T/C = 24	0.667
Transversions -	
A/T + T/A + A/C + C/A + G/T + T/G + G/C + C/G = 12	0.333

Figure 11: SNP distribution in the chromosome 6 data from The Genome Database.

3) *The SNP Consortium Database*

An example of the SNP data from the fourth release of the SNP Consortium Database is shown in appendix 3. The distribution of SNPs in this dataset is given in figure 12. The total number of biallelic SNPs in this data set was 5102. The observed ratio of $\alpha/2\beta$ is 2.09 which is four times higher than the expected ratio of 0.5. The total number of biallelic SNPs in this data set is 5102. This number is much higher than that for the first two data sets.

Proportion of SNPs:

		Proportion
A/G, G/A:	1750	0.343
A/C, C/A:	425	0.083
A/T, T/A:	342	0.067
G/C, C/G:	432	0.085
G/T, T/G:	454	0.089
C/T, T/C:	1699	0.333

Total = 5102		

Proportion of transitions and transversions:

	Proportion
Transitions -	
A/G + G/A + C/T + T/C = 3449	0.676
Transversions -	
A/T + T/A + A/C + C/A + G/T + T/G + G/C + C/G = 1653	0.324

Figure 12: SNP distribution in the chromosome 6 data from The SNP Consortium.

The results showing the number of SNPs are summarized in table 3. This table indicates that transitions are much more common than transversions, and therefore a two-parameter model may better describe the SNP properties than the one-parameter model. The overall results from figures 10, 11, and 12 are summarized in table 4.

Table 4: Summarized results from figures 10, 11, and 12.

Database	Transitions	Transversions	Transitions/transversions
1	103	43	0.705/0.300 = 2.35
2	24	12	0.667/0.333 = 2.003
3	3449	1653	0.676/0.324 = 2.086

ALLELE FREQUENCY DISTRIBUTION

The main aim was to understand the distribution of SNP variants in human population. This knowledge may be helpful in the future when generating simulated SNP data.

To see the distribution of the allele frequencies in the human population of the SNPs, I analyzed the allele frequency data collected for the SNPs from the Genome Database. I selected those alleles which were common (i.e. they have a frequency of more than 0.5). The distribution of allele frequencies is shown in figure 13. This distribution suggests that the allele frequency of most common alleles in SNP data follows a normal distribution with mean allele frequency of 0.7520 and standard deviation of 0.1272.

From the Kurtosis plot analysis performed to see the symmetry and normality of the allele frequency data,

$K^2 = 2.016915$ ($0.25 < P < 0.50$) (From the Chi-squared distribution table with two degrees of freedom)

Thus, the result of K^2 indicates that the data are normally distributed. These results will help in the future for simulations of SNP data while considering the allele frequencies.

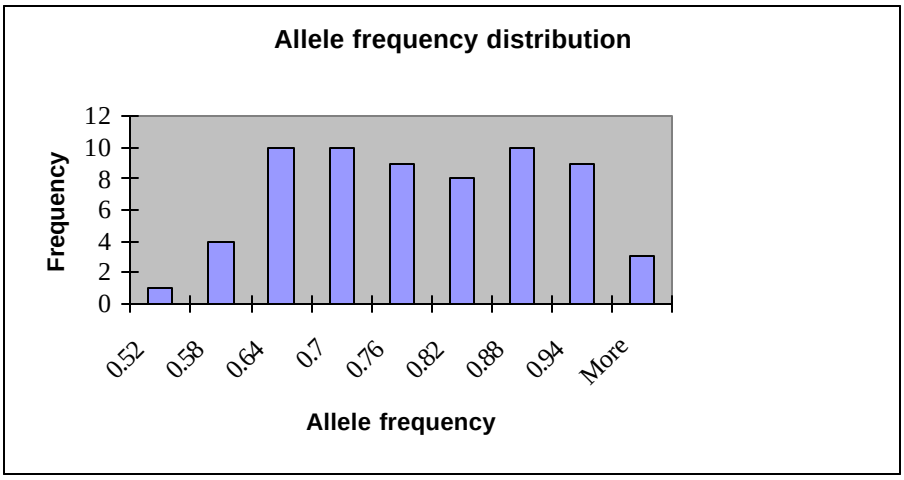


Figure 13: Allele frequency distribution for all common alleles of SNPs from human chromosome 6.

EFFECT OF THE RATIO OF TRANSITIONS TO TRANSVERSIONS ON LINKAGE DISEQUILIBRIUM

In order to study the effect of different ratios of transition to transversion on linkage disequilibrium, which is a frequently used parameter in association analysis, a simulation study was performed. After running the simulations for each set of parameters and analyzing the simulated data by Arlequin, I calculated the mean linkage disequilibrium values for each set of parameters. The results of this analysis are shown in the table 5.

Table 5: Mean linkage disequilibrium values ± standard deviation for different values of the transition to transversion ratios at different recombination rates.

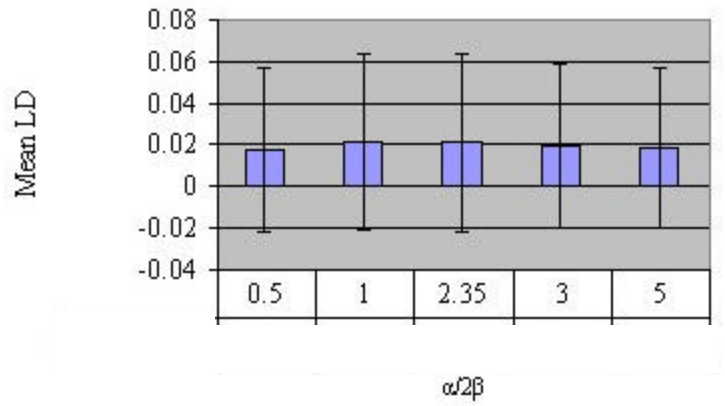
Transition/transversion (a/2b)	**Recombination rate**		
	0.0	**10^{-8}**	**3×10^{-5}**
0.5	0.0178 ±0.0393	0.0221 ±0.0454	0.0001 ±0.0006
1.0	0.0214 ±0.0420	0.0236 ±0.0495	0.00014 ±0.0007
2.35	0.0211 ±0.0426	0.0175 ±0.0423	0.00011 ±0.0006
3.0	0.0193 ±0.0394	0.0205±0.0475	0.00008 ±0.0006
5.0	0.0183 ±0.0386	0.0158 ±0.0391	0.00012 ±0.0007

The relationships between different transition to transversion ratios and mean LD at three recombination rates of 0.0, 10^{-8} and 3×10^{-5} are shown in figures 14, 15 and 16, respectively, with error bars indicating standard deviation (SD). No variable loci were observed for data simulated under the highest recombination rate (10^{-3}). I also observed a variation in the number of polymorphic loci for different

recombination rates, with the number of SNP loci decreasing as the recombination rate increased.

Figure 14: Plot of Mean LD Vs. a /2b at recombination rate (r) equal to 0.0.

Error bars indicate the standard deviations (SD)



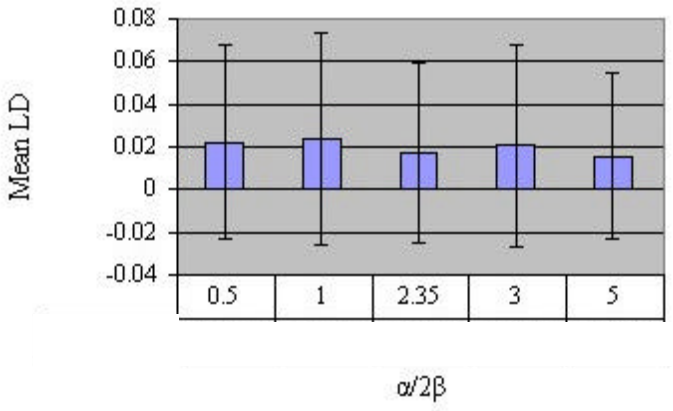


Figure 15: Plot of Mean LD Vs. a /2b at recombination rate (r) equal to 10^{-8}.

Error bars indicate the standard deviations (SD)

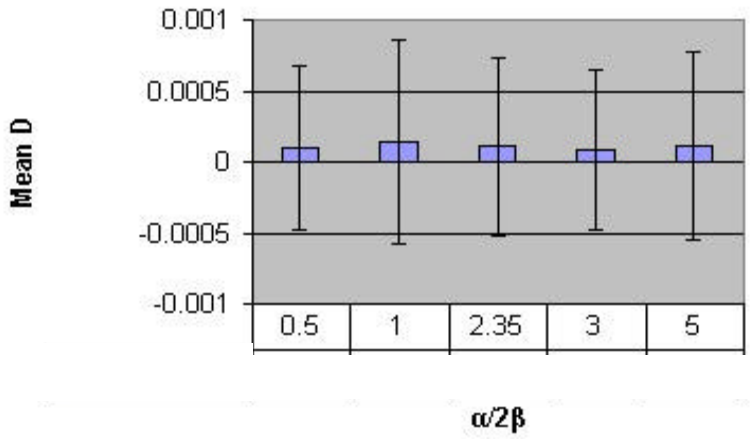


Figure 16: Plot of Mean LD Vs. a /2b at recombination rate (r) equal to $3´10^{-5}$

Error bars indicate the standard deviations (SD)

DISCUSSION

According to the one-parameter model, the expected ratio of transitions to transversions is 0.5 (since the transition rate is expected to be equal to the rate of the transversion). To see the actual proportion of transitions to transversions in the general human population, I collected data from publicly available databases. The ratio observed from the SNP data from the public databases showed that transitions are more frequent than transversions (transitions are approximately four times higher than transversions). Therefore, the two-parameter model (where transition and transversion rates differ) may better approximate the SNP behavior. My analyses showed that the proportion of transitions to transversions was very different (70% to 30%, 66% to 33%, and 68% to 32% in the Whitehead Institute of Genome Research, the Genome Database and the SNP Consortium databases respectively) than that expected (33% to 66%) under the scenario of no differences in mutation rates between different nucleotide changes.

To study the distribution of the frequencies of the most common alleles in the SNP data obtained from The Genome Database (GDB), I considered all biallelic SNPs in this data set. The allele frequency distribution of the most common alleles turned out to be normal with mean allele frequency of 0.7520 and standard deviation of 0.1272. After performing the Kurtosis plot analysis on this allele frequency data, it showed that the distribution of allele frequency is symmetric as well as normal (K^2 = 2.016915, corresponding to $0.25 < P < 0.50$). This analysis will help in the future for simulating SNP behavior, because allele frequencies play an important role in detecting strong association between a marker SNP and a susceptibility loci. If

marker allele frequencies are substantially different from susceptibility allele frequencies, then one needs a large sample size or a large number of markers or both to have a strong association (McCarthy and Hilfiker 2000). Therefore, the allele frequency distribution of common SNPs would help for future simulation studies involving SNP behavior in application to association studies involving allele frequencies.

I analyzed the effect of different transition to transversion ratios ($\alpha/2\beta$) on linkage disequilibrium. Results from the linkage disequilibrium analysis of simulated population showed that the linkage disequilibrium remained approximately same for different transition to transversion ratios for the parameters used in the simulations. The results obtained were for a sequence lengths of 1000 bp. As the recombination rate increased, the mean LD value decreased. The increase in recombination rate also resulted in reduction of number of polymorphic loci, thus reducing the strength of LD. The reduction in number of loci could be related to the small sequence length (1000 bp) in simulations and high recombination rate. An average extent of useful levels of LD in the general human population is approximately 3kb as shown by Kruglyak (1999). It is possible that the $\alpha/2\beta$ ratio might have an effect on LD when longer sequence length is used. Longer sequences can be investigated in future studies. One can also analyze the transitions and transversions separately from the SNP data obtained from the simulation studies. In the present study, the LD was evaluated between the polymorphic loci without considering whether each SNP was a transition or transversion.

No polymorphic sites were observed in the sequences simulated by TREEVOLVE for recombination rate of 10^{-3}. This result was observed with short sequence length and high recombination rate, which reduces the strength of LD considerably, because the chances of a polymorphism being fixed in a population are considerably less at a higher recombination rate while simulating the sequences. This phenomenon is because of the way the TREEVOLVE simulates the data. The program tries to simulate a polymorphism in the population while generating the sequences and not while simulating the population tree. When the recombination rate is much higher than the mutation rate, TREEVOLVE simulates a population that does not have polymorphism. These results support those previously obtained by Kruglyak (1999), in an independent analysis, in which he found no LD at recombination rate of 3×10^{-4} or higher (corresponding to physical distance of approximately 30kb). At such a high rate of recombination, there is higher separation between polymorphic loci. Such a high separation is probably unable to be accommodated in a sequence length of 1000 bp, resulting in the absence of any polymorphic loci in the population.

Part II

Application of SNPs in fine gene mapping

INTRODUCTION

As described in the earlier part of this thesis, association analysis has been of considerable importance in various fields ranging from population genetics, pharmacogenomics, population genetic epidemiology and toxicology. An association is said to exist between two phenotypes (of which one is resulting into a disease) if they occur in the same individual more often than expected by chance. To investigate whether the two phenotypes are associated or not, one collects the two groups of individuals, one with the affected individuals and one for controls (unaffected individuals). Then by counting the proportion of individuals having disease and the other phenotype and individuals with disease and not having the other phenotype, one can perform a standard 2×2 chi-square allelic association analysis or a 3×2 chi-square genotypic association analysis to examine the significance of the association.

A combination of alleles at a specific gene characterizes a genotype of an individual. Alleles and genotypes play a very important role in association analysis. If a gene is a disease pre-disposing gene, it is possible that only one of the alleles of that gene is actually responsible for the disease predisposition. Therefore, an allelic association was a common way of fine gene mapping until recently, although allelic heterogeneity may complicate this method (Terwilliger and Weiss 1998). Thus, genotypic association analysis can be used in place of allelic association. In allelic association, alleles are used to identify an association with disease. In contrast, in genotypic analysis one looks at association of a disease with genotypes (similar to genotypic linkage disequilibrium described by Weir 1996). In genotypic association

analysis one studies differences in genotype frequencies in healthy and affected individuals to find if any genotype is associated with the disease.

Objective

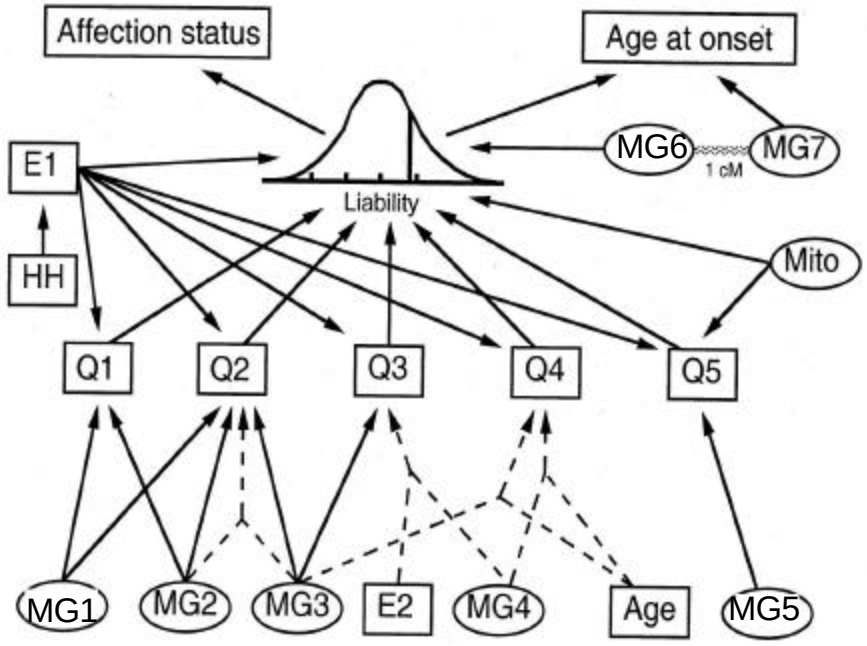
In this part, I describe an approach for fine gene mapping using SNPs from a general population by using genotype-specific disequilibrium analysis. The approach is different from the allelic disequilibrium analysis, because it considers the frequency of a genotype and not of individual alleles.

Description of the simulated data

A committee of Genetic Analysis Workshop 12 (GAW12) organized by Southwest Foundation for Biomedical Research, San Antonio, Texas, USA, simulated the data for GAW12 2000. The data used in this thesis were simulated for a large general population. The disease prevalence in the population was about 25%. The data were provided for 23 extended pedigrees with 1497 total individuals (1000 living) in the population. The disease was more prevalent in females than in males. Five quantitative risk factors (Q1 to Q5) and two environmental factors (E1 and E2) were also associated with the disease. Seven major genes (MG1 to MG7) influence these five quantitative risk factors as shown in the figure 17. A major gene is a gene related to the disease risk. The overall summary of the generating model is shown in figure 17. This model was not known during the analysis. In the data available to me, each living individual had information on affectation status, age at last exam, age at onset of disease if affected, five quantitative risk factors and two environmental

factors as well as genotypic data for SNPs present in 7 candidate genes. These genes were named from 1 through 7. These seven candidate genes were the genes, which were potential candidates that might affect the disease. The data for these seven candidate genes were provided by the GAW12 for analysis. These candidate genes were present in the major genes. The goal of the study described in this thesis was to test whether any of these candidate genes were contributing to the disease, and to identify any functional SNPs that could be related to the disease risk. The data were provided for 50 such replicates (50000 living individuals). There were 165 founders in each replicate. There was a total of 9515 original SNPs in the population of 50,000 individuals. The organizers also provided us with the information about which was the best replicate in the data. The best replicate was replicate 42 that had the data with contributions from all the factors discussed above. It represented the best sample population among all 50 replicates, with the mean simulated parameters being the closest to the parameters originally used I simulations.

Figure 17: The phenotypic model of the data simulated by GAW12. Boxes indicate items provided in the GAW12 data set, including quantitative traits, affectation status, age at onset, environmental factors (E1 and E2) and household membership (HH). Circles indicate genetic factors including seven major genes and a mitochondrial (Mito) component.



METHODS

Individuals used in the study

The study was performed on pedigree founders as well as all the living pedigree members. A founder is an individual in a family tree who does not have any living ancestors and is not related to anyone in his or her generation (Figure 18).

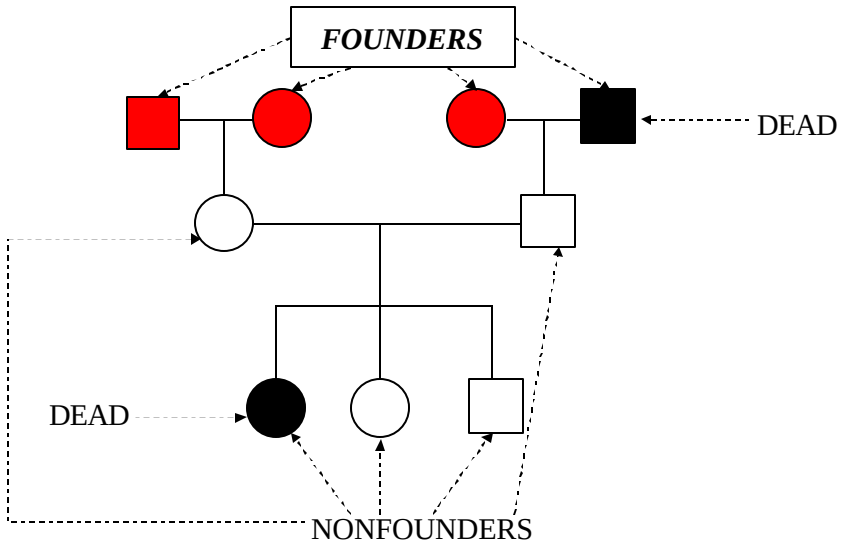


Figure 18: Founders in a pedigree.

I considered founders initially, because they were unrelated. Such an approach minimizes the correlations among family members in the SNP association analysis.

Several data sets were used in this study.

1) 8250 pedigree founders from all 50 replicates studying all candidate genes 1 to 7.

2) 8250 pedigree founders from all 50 replicates, analyzing only for genes 1 and 2 separately.
3) 165 pedigree founders from the best replicate 42, using only genes 1 and 2 separately.
4) 1000 living individuals from the best replicate 42 using only genes 1 and 2 separately.

Genotypic and phenotypic data

I considered the genotypic data for biallelic SNPs from 7 candidate genes for each individual along with his or her affectation status. The genotypic data were for 715 candidate SNPs selected after the data reduction. The data for each SNP genotype were provided by the GAW12 in binary format, i.e. 11, 12, or 22 instead of the nucleotides. Therefore, the transition or transversion nature of the polymorphisms was not taken into account.

Data reduction

The genotypic data set was very large. Considering all the 50 replicates, there were 50000 individuals. To minimize the dimensionality of the data set, only those SNPs were considered that were present in the pedigree founders in each of the 50 replicates by using a program called DATACONVERT (SaoPedro, unpublished). As a result, 715 SNPs were obtained from the total of 9515 SNPs after the data reduction. I considered the SNPs from the both coding and non-coding regions in the analysis.

Sorting technique

The first step in the analysis was to sort the data. I considered affection status as the variable of interest. I used program Scansort (Ward 2000, unpublished), which sorts the data according to the sum of the absolute differences between frequencies of healthy and affected subjects with the three SNP genotypes (11, 12, or 22, where 1 is a wild type, or ancestral, allele and 2 is a mutated allele). The output of the program is an eleven-dimensional data set, which contains one record for each SNP position in the following order:

a. SNP index (original order)

b. Frequency of healthy individuals with genotype 11 (f_{11h})

c. Frequency of healthy individuals with genotype 12 (f_{12h})

d. Frequency of healthy individuals with genotype 22 (f_{22h})

e. Frequency of affected (sick) individuals with genotype 11 (f_{11s})

f. Frequency of affected (sick) individuals with genotype 12 (f_{12s})

g. Frequency of affected (sick) individuals with genotype 22 (f_{22s})

h. Sum of all the absolute differences (d)

i. Absolute difference between c and f

j. Absolute difference between d and g

k. Absolute difference between e and h

Table 6 represents this information in tabular form.

Table 6: Calculation of maximum difference by Scansort program.

Individual	Genotype		
	11	12	22
Healthy	f_{11h}	f_{12h}	f_{22h}
Sick (Affected)	f_{11s}	f_{12s}	f_{22s}

Scansort calculates d, the sum of the absolute differences, by using the following formula.

$$d = (| f_{11h} - f_{11s}| + |f_{12h}-f_{12s}| + |f_{22h}-f_{22s}|) / \text{Number of individuals in the data set}$$

Scansort sorts the SNP data according to the d value. Therefore, Scansort orders the SNP data in a useful way where SNPs that have very different proportions in healthy and affected individuals are present at the top of the list. As a result, the output of the Scansort was used for the conventional 3×2 chi-square analysis.

Comparative statistical analysis of genotype frequencies

I used the chi-square statistic to find out the significant SNPs associated with the disease. The Bonferroni correction was applied to correct the significance levels for multiple testing. Statistical analysis was performed using Microsoft Excel.

Chi-square statistic is calculated by using the following formula,

$$\chi^2 = \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

The chi-square analysis for individual SNP position was performed by considering a 3×2 contingency table (Table 7).

Table 7: Chi-square analysis for an individual SNP position.

Individual	Genotype			
	11	12	22	
Healthy	f_{11h}	f_{12h}	f_{22h}	$R_1 = f_{11h} + f_{12h} + f_{22h}$
Sick (Affected)	f_{11s}	f_{12s}	f_{22s}	$R_2 = f_{11s} + f_{12s} + f_{22s}$
	$C_1 = f_{11h} + f_{11s}$	$C_2 = f_{12h} + f_{12s}$	$C_3 = f_{22h} + f_{22s}$	$A = C_1+C_2+C_3 = R_1+R_2$

Let $A = C_1+C_2+C_3 = R_1+R_2$ (total individuals in a population), then

Expected frequency for f_{11h}, $f_{11he} = C1 \times R1/A$

$$\chi^2 = \sum \frac{(\text{Observed frequency} - \text{expected frequency})^2}{(\text{expected frequency})}$$

The chi-square distribution is calculated by using appropriate degrees of freedom. The degree of freedom is calculated by using following formula (Zar 1999),

$$\text{Degrees of freedom} = (\text{columns} - 1) \times (\text{rows} - 1)$$

The Bonferroni correction was used to adjust the significance level, as there are a large number of tests to be performed because of the high number of loci (Weir 1996). α was calculated by using following formula.

$$\alpha = 1 - (1 - \alpha')^{1/L}$$

where $\alpha' = 0.05$, and L is the number of SNPs analyzed.

RESULTS

To find the significant SNPs associated with the disease I performed chi-square analysis initially on a data set containing the founders from all the 50 replicates. Each of the 715 SNPs in genes 1 through 7 were analyzed for the differences in their genotype frequencies between healthy and affected subjects for the data set with all the pedigree founders from the 50 replicates. These SNPs were then sorted by the Scansort program for the sum of their absolute differences between their genotype frequencies. The top twenty SNPs from this analysis are shown in table 8.

Table 8: Scansort output for data set 1 (8250 pedigree founders) for top 20 SNPs sorted by the value of d. All the 20 SNPs are from gene 1.

SNP ID	11h	12h	22h	11s	12s	22s	d (diff)
557	93	1657	4103	368	1281	748	0.777903
76	4113	1647	93	754	1278	365	0.776313
2619	4219	1547	87	864	1235	298	0.720753
1553	4236	1536	81	871	1229	297	0.720721
3573	4197	1565	91	863	1236	298	0.71407
3835	4189	1572	92	860	1238	299	0.713839
3853	92	1575	4186	302	1236	859	0.713648
3742	92	1576	4185	302	1236	859	0.713307
3456	4238	1530	85	889	1224	284	0.706386
5757	4122	1620	111	842	1240	315	0.705964
7281	4081	1654	118	838	1237	322	0.695291
2942	139	1738	3976	330	1264	803	0.688615
2923	137	1766	3950	336	1258	803	0.679731
11180	4007	1710	136	837	1229	331	0.670839
1478	172	1845	3836	356	1257	784	0.65663
189	268	2154	3431	468	1303	626	0.650071
596	3433	2151	269	627	1309	461	0.64992
4471	200	1913	3740	360	1272	765	0.639679
4752	203	1922	3728	361	1273	763	0.637248
3534	262	2079	3512	396	1274	727	0.593477

After having obtained the ordered SNP data for 715 SNPs from all the genes by Scansort, I performed chi-square analysis on the frequency data obtained from the Scansort output. The results for the most significant SNPs according to the p-values in genes 1, 2, and 6 are shown in tables 8, 9, and 10, respectively. I did not found any significant SNP in any other genes except 1,2, and 6. The resulting Bonferroni correction for each SNP was 7.17363×10^{-5} when the initial significance level was set to 0.05.

Table 9: The 20 most significant SNPs in gene 1. The data set analyzed was the 8250 pedigree founders from all the 50 replicates.

SNP ID	11h	12h	22h	11s	12s	22s	Chi Square	p-value
557	93	1657	4103	368	1281	748	1315.64227	$2.0507*10^{-286}$
76	4113	1647	93	754	1278	365	1308.13783	$8.7394*10^{-285}$
1553	4236	1536	81	871	1229	297	1124.2512	$7.4465*10^{-245}$
2619	4219	1547	87	864	1235	298	1112.56292	$2.5706*10^{-242}$
3853	92	1575	4186	302	1236	859	1090.46042	$1.62*10^{-237}$
3742	92	1576	4185	302	1236	859	1089.66349	$2.4131*10^{-237}$
3573	4197	1565	91	863	1236	298	1088.87717	$3.5754*10^{-237}$
3835	4189	1572	92	860	1238	299	1087.27638	$7.9603*10^{-237}$
3456	4238	1530	85	889	1224	284	1068.70713	$8.5742*10^{-233}$
5757	4122	1620	111	842	1240	315	1052.39334	$2.9901*10^{-229}$
7281	4081	1654	118	838	1237	322	1024.87688	$2.8236*10^{-223}$
2942	139	1738	3976	330	1264	803	984.310527	$1.8183*10^{-214}$
2923	137	1766	3950	336	1258	803	976.29031	$1.0028*10^{-212}$
11180	4007	1710	136	837	1229	331	954.378676	$5.7451*10^{-208}$
189	268	2154	3431	468	1303	626	916.243316	$1.0972*10^{-199}$
596	3433	2151	269	627	1309	461	905.951689	$1.884*10^{-197}$
1478	172	1845	3836	356	1257	784	902.347863	$1.1419*10^{-196}$
4471	200	1913	3740	360	1272	765	838.786422	$7.2417*10^{-183}$
4752	203	1922	3728	361	1273	763	831.839741	$2.335*10^{-181}$
12185	3543	2059	251	747	1237	413	750.82136	$9.1456*10^{-164}$

Table 10: The 15 most significant SNPs in gene 2. The data set analyzed was the 8250 pedigree founders from all the 50 replicates.

SNP ID	11h	12h	22h	11s	12s	22s	Chi Square	p-value
4894	680	2643	2530	170	873	1354	127.7856	$1.7853*10^{-28}$
4977	3605	2003	245	1782	565	50	125.28594	$6.2302*10^{-28}$
4766	672	2634	2547	172	873	1352	120.04499	$8.5617*10^{-27}$
3185	614	2594	2645	158	848	1391	117.5252	$3.0180*10^{-26}$
861	569	2556	2728	145	828	1424	116.38561	$5.3356*10^{-26}$
1495	567	2555	2731	144	829	1424	115.68125	$7.5881*10^{-26}$
715	2730	2554	569	1423	830	144	115.47719	$8.4032*10^{-26}$
4030	627	2597	2629	158	865	1374	112.08614	$4.5793*10^{-25}$
4538	633	2604	2616	167	859	1371	111.29881	$6.7884*10^{-25}$
5961	3826	1834	193	1844	518	35	110.24763	$1.1482*10^{-24}$
2540	570	2539	2744	152	856	1389	88.29132	$6.7264*10^{-20}$
3155	2914	2454	485	1454	814	129	84.504551	$4.4675*10^{-19}$
5219	957	2869	2027	281	1059	1057	73.380148	$1.1633*10^{-16}$
2805	738	2684	2431	216	948	1233	71.968459	$2.356*10^{16}$
5499	978	2877	1998	287	1073	1037	70.221367	$5.6444*10^{-16}$

Table 11: The 12 most significant SNPs in gene 6. The data set analyzed was the 8250 pedigree founders from all the 50 replicates.

SNP ID	11h	12h	22h	11s	12s	22s	Chi Square	p-value
6805	4948	873	32	1725	625	47	185.364929	$5.6042*10^{-41}$
7332	4948	873	32	1725	625	47	185.364929	$5.6042*10^{-41}$
8067	4951	870	32	1727	623	47	184.886869	$7.1174*10^{-41}$
5782	4947	874	32	1725	625	47	184.845233	7.26721E-41
7073	4946	874	33	1724	626	47	184.492956	$8.6668*10^{-41}$
5007	4942	879	32	1724	626	47	183.275133	$1.5933*10^{-40}$
8226	35	893	4925	47	633	1717	179.169005	$1.2414*10^{-39}$
1987	202	1793	3858	146	879	1372	67.4235199	$2.2864*10^{-15}$
1748	3855	1797	201	1372	879	146	67.1851486	$2.5759*10^{-15}$
993	3854	1796	203	1373	877	147	66.4377271	$3.7430*10^{-15}$
11782	513	2420	2920	155	909	1333	26.8361224	$1.4880*10^{-06}$
13869	514	2415	2924	158	911	1328	24.2681663	$5.3732*10^{-06}$

The range of significant SNPs for each gene is reported in table 12. According to the significance level of 7.17363×10^{-5} (Bonferroni cut-off correction), only 172 SNPs out of 715 SNPs were significant, and thus could be associated with the disease.

Table 12. Number of significant sequence polymorphisms, the range of their significance values, and the total number of polymorphisms, by gene, determined from the 8250 pedigree founders from all the 50 replicates.

Gene	# of significant SNPs	Range of significant p-values	Total # of SNPs
1	107	$2.0507\text{x}10^{-286}$ to $6.21184\text{x}10^{-5}$	157
2	52	$1.78529\text{x}10^{-28}$ to $4.29522\text{x}10^{-5}$	90
6	13	$5.60425\text{x}10^{-41}$ to $5.37322\text{x}10^{-6}$	34

I considered only founders initially as they were responsible to transmit the SNPs if any in the next generation. Hence I analyzed the SNP data considering all the founders from the 50 replicates in the hope to find all the SNPs significant with respect to the disease causing polymorphisms. Analysis of all the pedigree founders showed that most of the SNPs in genes 1 and 2 might be associated with the disease. Therefore, to narrow down the SNPs of most interest, I analyzed 8250 pedigree founders for SNPs only in genes 1 and 2 separately following the same procedure as in 1. The results of this analysis are shown in table 13.

Table 13: Number of significant SNPs in genes 1 and 2 analyzed separately for 8250 pedigree founders. (cut-off p-values from Bonferroni correction are 0.000327 for gene 1 and 0.000570 for gene 2)

Gene	# of significant SNPs	Range of p-values	Total # of SNPs
1	114	2.0507×10^{-286} to 0.000287	157
2	61	1.78529×10^{-28} to 0.000398	90

I also analyzed the SNP data for genes 1 and 2 separately for the founders in best replicate 42 in an attempt to identify the most significant SNPs associated with the disease in the best replicate and then to compare it with the significant SNPs obtained after analyzing 8250 pedigree founders. These results are shown in table 14.

Table 14: Number of significant SNPs in genes 1 and 2 analyzed separately for 165 pedigree founders in best replicate 42. (cut-off p-values from Bonferroni correction are 0.000327 for gene 1 and 0.000570 for gene 2)

Gene	# of significant SNPs	Range of p-values	Total # of SNPs
1	35	9.54×10^{-8} to 0.000325	157
2	0	--	90

I also analyzed the SNP data from all individuals in replicate 42 for genes 1 and 2 separately and then compare it with the results obtained for the same analysis performed with 8250 pedigree founders and 165 founders from replicate 42

separately to see the effect of sample size on the number of significant SNPs. The results are shown in table 15.

Table 15: Number of significant SNPs in genes 1 and 2 analyzed separately for 1000 individuals in best replicate 42. (cut-off p-values from Bonferroni correction are 0.000327 for gene 1 and 0.000570 for gene 2)

Gene	# of significant SNPs	Range of p-values	Total # of SNPs
1	55	2.18×10^{-35} to 0.000125	157
2	1	6.01×10^{-6}	90

DISCUSSION

The analysis was performed to find the most significant SNPs associated with the disease in the simulated GAW12 data set. I used a genotype disequilibrium approach in which I considered the difference in genotype frequencies of the SNPs. According to the results obtained, I found possible mutations in genes 1, 2, and 6 that were associated with the disease. This analysis was performed without the GAW 12 answers, which were made available to us after the analysis was completed.

The chi-square analysis on data set of 8250 pedigree founders identified 107 significant SNPs in gene 1 (out of a total of 157 SNPs in that gene), 52 SNPs in gene 2 (out of a total of 90 SNPs in gene 2) and 13 SNPs in gene 6 (out of a total of 36 SNPs in gene 6). Because genes 1, 2, and 6 contain multiple polymorphisms which are associated with the disease, it is difficult to identify causative mutations for the disease. Hence, one hypothesis that I pursued was that the sequence polymorphisms showing the lowest *p*-values were the most likely candidates for affecting the disease state. In this regard, I identified SNPs at nucleotide position 557 in gene 1 (lowest *p*-value of 2.05×10^{-286}), nucleotide positions 6805 and 7332 in gene 6 (*p*-value = 5.60×10^{-41}) and nucleotide position 4894 in gene 2 ($p = 1.79 \times 10^{-28}$) that were most likely associated with the disease. Many of the SNPs located near these mutations also have very low p-values, most likely because of the linkage disequilibrium between SNP variants. For example, in gene 1 the SNP at position 557 has the lowest *p*-value of 2.05×10^{-286}. This is the sixth polymorphism occurring in this gene, counting from the start of the gene. The second polymorphism from the start of this gene, at position 76, has the next lowest *p*-value of 8.74×10^{-285}. The third

polymorphism from the start of the gene, at nucleotide position 189, has a *p*-value of 1.10 $x10^{-199}$, and the seventh polymorphism, at position 596, has *p*-value = 1.88 $x10^{-197}$. The fourth, eighth, ninth and tenth polymorphisms at nucleotide positions 286, 610, 730 and 885 respectively, also show significant association with the disease (with *p*-value<0.000327). This observation suggests a likely functional role of the 5' end of gene 1 in the disease.

There was generally a good correlation between the SNPs identified by the chi-square test and the polymorphisms that had the top scores identified by program Scansort. The majority of SNPs that showed significant association with the disease by the chi-square test were located in the top portion of the arrays generated by Scansort. The most significant SNPs from genes 1, 2, and 6 (sequence positions 557 in gene 1, 4894 in gene 2 and 6805 and 7332 in gene 6, respectively) were ranked as number 1, 61, 47 and 48 when sorted by their chi-square values in data set of 8250 pedigree founders. After sorting by Scansort in the same data set for the sum of the absolute differences in genotype frequencies in healthy and affected individuals, their positions were 1, 48, 56 and 57, respectively, in the sorted data set. When the 172 top SNP positions that were significant in data set containing all pedigree founders from all the 50 replicates according to the chi-square method were compared to the top 172 SNP positions identified by the Scansort, all but 24 SNPs were found to be shared between the two lists, with the proportion of shared SNPs equal to 87%.

The sensitivity of our tests seemed to increase with the increase of the number of individuals in the data set. This was the case whether relatives or non-relatives were added to the data set. I found 25 SNPs in gene 1 which were significant in data

set 1 with founders from all 50 replicates, but not significant in either data set 3 or 4. Analysis of the data set containing all the 715 SNPs with all the founders from 50 replicates revealed 107 significant SNPs in gene 1, while that number was 114 in data set containing the SNPs for gene 1 only with all the founders from 50 replicates, 35 in data set containing just founders from replicate 42, and 55 in data set containing all the individuals in replicate 42. In case of gene 2, I identified 52 significant SNPs in data set containing the 715 SNPs with all the founders from the 50 replicates, 61 significant SNPs in data set containing all the founders from 50 replicates for gene 2 only, 0 in data set containing just founders from replicate 42 and 1 in data set which contained all the individuals in replicate 42. The p-values were also smaller in the larger data sets. Therefore, it seems beneficial to add individuals to the data set even if they are related to one another.

The answers to this problem provided by the GAW12 organizers (GAW12 Abstracts) were very close to the answers obtained in this analysis. The SNPs associated with the disease were in genes 1, 2, and 6 according to the GAW12 answers (Figure 17) where candidate gene 1 corresponds to the major gene 6, candidate gene 2 corresponds to the major gene 5 and candidate gene 6 corresponds to major gene 1. My analysis also showed association of these three genes with the disease. The exact positions of the SNPs in these genes according to the answers were at nucleotide position 557 in gene 1, and position 5782 in gene 6. In gene 2, there were number of multi-allelic functional variants; in the regulatory elements or in the first or second bp of a codon; leading to amino acid substitutions in gene 2 and were associated with the disease. The effect of these genes is shown in figure 17. The

analysis presented here also showed the SNP at 557 position in gene 1 as the most likely candidate; however, the SNPs in gene 6 identified at positions 6805 and 7332 were not correct when compared to the GAW12 answers. Despite discrepancies related to individual SNPs, the analysis presented here correctly identified the importance of all the three genes1, 2, and 6 in disease, and these results show importance of genotypic linkage disequilibrium methods in fine gene mapping.

BIBLIOGRAPHY

1) Balendran, N. et al. Characterization of the major susceptibility region for psoriasis at chromosome 6p21.3. J. Invest. Derm. 113: 322-328 (1999).
2) Brower, V. Genome II: The next frontier. Nature Biotechnology 16(11): 1004-1007 (1998).
3) Brown, K. The human genome business today. Scientific American: 50-55 (July 2000).
4) Cao, Q., Martinez, M., Zhang, J., Sanders, A. R., Badner, J. A., Cravchik, A., Markey, C. J., Beshah, E., Guroff, J. J., Maxwell, M. E., Kazuba, D. M., Whiten, R., Goldin, L. R., Gershon, E. S., Gejman, P. V. Suggestive evidence for a schizophrenia susceptibility locus on chromosome 6q and a confirmation in an independent series of pedigrees. Genomics 43: 1-8 (1997).
5) Collins, F. S., Guyer, M. S., and Chakravarti, A. Variations on a theme: Cataloging human DNA sequence variations. Science 278: 1580-1581 (1997).
6) Davies, J. L. A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371: 130-136 (1994).
7) Elston, R. Linkage and association. Genetic Epidemiology 15: 565-576 (1998).
8) Felsenstein, J. and Churchill, G. A. A hidden markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. 13: 93-104 (1996).

9) Grassly and Rambaut, TREEVOLVE v. 1.3 (http://evolve.zoo.ox.ac.uk), unpublished.

10) Hasegawa, M., Kishino, H. and Yano, T. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22: 160-174 (1985).

11) International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409: 860-921 (2001).

12) Kingman, J. F. C. Origins of the coalescent. Genetics 156: 1461-1463 (2000).

13) Kruglyak, L. et al. Parametric and nonparametric linkage analysis: A unified approach. American Journal of Human Genetics 58: 1347-1363 (1996).

14) Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nature Genetics 22: 139-144 (1999).

15) Krushkal, J., et al. Genome-wide linkage analysis of systolic blood pressure using highly discordant siblings. Circulation 11: 1407-1410 (1999).

16) Lai, E. et al. A 4-Mb high-density single nucleotide polymorphism-based map around human APOE. Genomics 54: 31-38 (1998).

17) Lander, E.S. et al. Large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. Science 280: 1077-1082 (1998).

18) Li, W. Evolutionary change in nucleotide sequences. In: Li, W. editor. Molecular Evolution. Sunderland: Sinauer Associates, Inc. p. 53 (1997).

19) McCarthy, J. J., and Hilfiker, R. The use of single-nucleotide polymorphism maps in Pharmacogenomics. Nat. Biotech. 18: 505-508 (2000).

20) Nobukuni, Y. et al. Maple syrup urine disease: complete defect of the E1-beta subunit of the branched chain alpha-ketoacid dehydrogenase complex due to a deletion of an 11-bp repeat sequence which encodes a mitochondrial targeting leader peptide in a family with the disease. J. Clin. Invest. 87: 1862-1866 (1991).

21) Risch, N. and Merikangas, K. The future of genetic studies of complex human diseases. Science 273: 1516-1517 (1996).

22) Rothberg, B. E. G., Ramesh, T. M. and Burgess, C. E. Integrating expression-based drug response and SNP-based pharmacogenetic strategies into a single comprehensive pharmacogenomics program. Drug Development Research 49: 54-64 (2000).

23) Schafer, A. J., and Hawkins, J. R. DNA variation and future of human genetics. Nature Biotechnology. 16: 33-39 (1998).

24) Schneider, S. et al. Arlequin: a software for population genetics data analysis user manual version 2.0. Genetics and Biometry Lab. Department of Anthropology, University of Geneva. (http://anthro.unige.ch/arlequin/).

25) Spielman, R. S., McGinnis, R. E. and Ewen, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J of Hum Genet. 52(3): 506-516 (1993).

26) Terwilliger, J. D., and Weiss, K. M. Linkage disequilibrium mapping of complex diseases: fantasy or reality? Current opinion in Biotechnology 9: 578-594 (1998).

27) The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928-933 (2001).

28) Venter, C. et al. The sequence of the human genome. Science 291(5507): 1304-1351 (2001).

29) Weir, B. Disequilibrium. In: Weir, B. S. editor. Genetic data analysis II - methods for discrete population genetic data. Sunderland, MA: Sinauer Associates. Inc. p. 91-139 (1996).

30) Zar, J. H. Testing for goodness of fit. In: Zar, J. H. editor. Biostatistical analysis. New Jersey: Prentice Hall. p. 461-485 (1999) (a).

Zar, J. H. The normal distribution. In: Zar, J. H. editor. Biostatistical analysis. New Jersey: Prentice Hall. p. 65-90 (1999) (b).

31) Zollner, S. and Haeseler, A. A Coalescent approach to study linkage disequilibrium between single-nucleotide polymorphisms. Am. J. Hum. Genet. 66: 615-628 (2000).

APPENDICES

Appendix 1: An example of the SNP data for human chromosome 6 from the Whitehead Institute of Genome Research. In all there were 146 SNPs. (cR corresponds to the radiation hybrid distance). VNTR (variable number of tandem repeats) are the top and bottom markers for that particular SNP position.

SNP NAME	TYPE	GENETIC DISTANCE (cM)	TOP VNTR	BOTTOM VNTR
WIAF 1583	A/G	1.4	D6S344	D6S344
WIAF 857	C/T	1.37(cR)	D6S344	D6S344
WIAF 1034	C/T	0.00(cR)	D6S344	D6S344
WIAF 2096	G/A	6.4	D6S1617	D6S1617
WIAF 131	T/C	41.75(cR)	D6S1617	D6S1617
WIAF 132	T/C	41.75(cR)	D6S1617	D6S1617
WIAF 950	A/C	51.37(cR)	D6S296	D6S470
WIAF 549	C/T	9.0	D6S296	D6S470
WIAF 890	G/C	59.80(cR)	D6S1674	D6S1674
WIAF 1939	C/G	54.83(cR)	D6S1674	D6S1674
WIAF 1020	A/G	17.7	D6S470	D6S470
WIAF 881	G/A	67.56(cR)	D6S470	D6S1578
WIAF 1567	G/A	20.5	D6S470	D6S1578
WIAF 967	T/C	66.11(cR)	D6S470	D6S1578
WIAF 1891	C/T	81.31(cR)	D6S469	D6S288
WIAF 116	G/A	85.33(cR)	D6S469	D6S288
WIAF 1541	C/G	34.2	D6S1688	D6S1688
WIAF 1966	G/A	111.66(cR)	D6S1688	D6S422
WIAF 709	T/A	34.0	D6S1688	D6S422
WIAF 105	C/T	118.48(cR)	D6S422	D6S1686
WIAF 106	C/A	118.48(cR)	D6S422	D6S1686
WIAF 415	T/C	121.60(cR)	D6S1686	D6S1686
WIAF 1896	A/T	123.20(cR)	D6S1686	D6S1691
WIAF 613	G/T	40.0	D6S1686	D6S1691
WIAF 614	A/G	40.0	D6S1686	D6S1691
WIAF 1959	T/C	128.64(cR)	D6S1691	D6S464
WIAF 1574	T/C	130.56(cR)	D6S464	D6S464
WIAF 1460	A/G	46	D6S276	D6S439
WIAF 1461	T/A	46	D6S276	D6S439
WIAF 1462	G/T	46	D6S276	D6S439
WIAF 2020	C/G	161.68(cR)	D6S276	D6S439
WIAF 2021	T/C	161.68(cR)	D6S276	D6S439
WIAF 139	A/C	165.13(cR)	D6S276	D6S439
WIAF 1551	T/C	46.6	D6S276	D6S439
WIAF 1552	G/A	46.6	D6S276	D6S439
WIAF 1553	G/A	46.6	D6S276	D6S439
WIAF 1554	T/C	46.6	D6S276	D6S439
WIAF 1722	A/T	46.7	D6S276	D6S439

WIAF 110	C/T	179.84(cR)	D6S276	D6S439
WIAF 556	C/T	177.90(cR)	D6S276	D6S439
WIAF 557	C/T	177.90(cR)	D6S276	D6S439
WIAF 558	T/C	177.90(cR)	D6S276	D6S439
WIAF 257	C/A	178.21(cR)	D6S276	D6S439
WIAF 258	A/C	178.21(cR)	D6S276	D6S439
WIAF 1185	C/G	46.9	D6S276	D6S439
WIAF 1453	A/G	46.9	D6S276	D6S439
WIAF 1935	C/T	47.0	D6S276	D6S439
WIAF 897	C/T	178.21(cR)	D6S276	D6S439
WIAF 898	A/G	178.21(cR)	D6S276	D6S439
WIAF 1696	C/G	47.0	D6S276	D6S439
WIAF 643	A/G	Unassigned	D6S276	D6S439
WIAF 1306	A/G	47.1	D6S276	D6S439
WIAF 1084	C/T	47.2	D6S276	D6S439
WIAF 1009	T/C	47.2	D6S276	D6S439
WIAF 1335	G/C	179.74(cR)	D6S276	D6S439
WIAF 1336	A/G	179.74(cR)	D6S276	D6S439
WIAF 1337	C/A	179.74(cR)	D6S276	D6S439
WIAF 1338	A/G	179.74(cR)	D6S276	D6S439
WIAF 1339	G/A	179.74(cR)	D6S276	D6S439
WIAF 1340	T/C	179.74(cR)	D6S276	D6S439
WIAF 1341	G/A	179.74(cR)	D6S276	D6S439
WIAF 1342	T/C	179.74(cR)	D6S276	D6S439
WIAF 2177	G/C	47.7	D6S276	D6S439
WIAF 2106	C/T	47.7	D6S276	D6S439
WIAF 2103	A/G	47.8	D6S276	D6S439
WIAF 1006	T/C	188.58(cR)	D6S276	D6S439
WIAF 1746	A/G	194.89(cR)	D6S276	D6S439
WIAF 916	T/C	193.76(cR)	D6S276	D6S439
WIAF 1586	G/A	52.8	D6S291	D6S1610
WIAF 1587	T/C	52.8	D6S291	D6S1610
WIAF 2075	G/A	65.5	D6S426	D6271
WIAF 1780	A/C	252.98(cR)	D6S426	D6271
WIAF 1608	A/G	246.37(cR)	D6271	D6271
WIAF 1609	A/G	246.37(cR)	D6271	D6271
WIAF 1060	G/A	258.00(cR)	D6271	D6271
WIAF 98	T/C	251.95(cR)	D6271	D6S459
WIAF 1359	T/C	69.0	D6271	D6S459
WIAF 1360	T/C	69.0	D6271	D6S459
WIAF 538	C/T	69.0 TO 72.0	D6S459	D6S438
WIAF 539	C/A	69.0 TO 72.0	D6S459	D6S438
WIAF 997	G/T	271.85(cR)	D6S459	D6S438
WIAF 520	C/T	77.0	D6S436	D6S466
WIAF 805	C/T	307.81	D6S466	D6S466
WIAF 2212	T/C	79.9	D6S257	D6S257
WIAF 589	T/A	462.24(cR)	D6S1589	D6S1589

WIAF 922	A/C	533.70(cR)	D6S1589	D6S1589
WIAF 1486	A/G	557.69(cR)	D6S1601	D6S1570
WIAF 1669	G/A	557.21(cR)	D6S1601	D6S1570
WIAF 1670	G/A	557.21(cR)	D6S1601	D6S1570
WIAF 220	G/T	539.89(cR)	D6S1601	D6S1570
WIAF 1654	A/C	102.4	D6S417	D6S424
WIAF 78	C/T	626.11(cR)	D6S417	D6S424
WIAF 1427	G/C	618.65(cR)	D6S417	D6S424
WIAF 1476	T/C	629.40(cR)	D6S1716	D6S468
WIAF 207	G/A	116.2	D6S278	D6S278
WIAF 844	A/C	682.24	D6S278	D6S278
WIAF 845	A/C	682.24	D6S278	D6S278
WIAF 1875	C/T	116.6	D6S278	D6S302
WIAF 1876	G/A	116.6	D6S278	D6S302
WIAF 1877	G/C	116.6	D6S278	D6S302
WIAF 1878	T/C	116.6	D6S278	D6S302
WIAF 1879	C/T	116.6	D6S278	D6S302
WIAF 1880	A/G	116.6	D6S278	D6S302
WIAF 1881	C/T	116.6	D6S278	D6S302
WIAF 2219	T/G	116.6	D6S278	D6S302
WIAF 2126	G/A	116.9	D6S278	D6S302
WIAF 584	T/C	707.96(cR)	D6S423	D6S423
WIAF 726	C/T	121	D6S423	D6S423
WIAF 985	T/A	705.43(cR)	D6S423	D6S1712
WIAF 1203	G/T	121.0	D6S423	D6S1712
WIAF 2171	G/A	124.2	D6S423	D6S1712
WIAF 224	C/T	122.0(cR)	D6S423	D6S1712
WIAF 225	T/C	122.0(cR)	D6S423	D6S1712
WIAF 736	C/G	731.63(cR)	D6S423	D6S1712
WIAF 737	G/A	731.63(cR)	D6S423	D6S1712
WIAF 492	A/T	733.67(cR)	D6S407	D6S262
WIAF 1674	T/C	736.71(cR)	D6S407	D6S262
WIAF 738	A/G	129.0	D6S407	D6S262
WIAF 1366	G/A	744.83(cR)	D6S407	D6S262
WIAF 642	T/C	138.0	D6S292	D6S1699
WIAF 1604	G/C	144.5	D6S453	D6S453
WIAF 466	A/G	748.63(cR)	D6S453	D6S308
WIAF 1873	G/A	788.32(cR)	D6S453	D6S308
WIAF 3	T/G	790.34(cR)	D6S453	D6S308
WIAF 2034	T/C	145.6	D6S308	D6S308
WIAF 1704	T/G	150.4	D6S311	D6S1687
WIAF 1623	A/G	150.4	D6S311	D6S1687
WIAF 592	A/G	811.66(cR)	D6S311	D6S1687
WIAF 982	C/T	814.38(cR)	D6S1687	D6S1687
WIAF 563	A/C	155.0	D6S1687	D6S448
WIAF 762	C/G	837.32(cR)	D6S419	D6S1579
WIAF 1954	T/C	165.7	D6S419	D6S1579
WIAF 1955	G/A	165.7	D6S419	D6S1579
WIAF 1493	T/A	165.7	D6S419	D6S1579

WIAF 1494	A/G	165.7	D6S419	D6S1579
WIAF 888	T/C	165.7	D6S419	D6S1579
WIAF 915	G/A	175.9	D6S1579	D6S1719
WIAF 97	G/A	833.43(cR)	D6S1579	D6S1719
WIAF 397	C/T	843.63	D6S1719	D6S1719
WIAF 1690	A/C	178.4	D6S1719	D6S264
WIAF 30	C/T	846.82(cR)	D6S1719	D6S264
WIAF 12	T/C	167.0	D6S1719	D6S264
WIAF 13	A/G	167.0	D6S1719	D6S264
WIAF 1739	A/G	184.8	D6S1585	D6S446
WIAF 1740	A/G	184.8	D6S1585	D6S446
WIAF 2186	A/G	189.0	D6S446	D6S1693

Appendix 2: The SNP data of human chromosome 6 from the Genome Database (PV = point variation)

GENE NAME	SNP	TYPE/ LOCATION	CHANGE IN AMINO ACID	SNP
ABCB2 (TAP1)	PV	Codon 333	ILE. to VAL.(ATC to GTC)	A/G
	PV	Codon 370	ALA to VAL (GCT to GTT)	C/T
	PV	Codon 458	VAL to LEU (GTG to TTG)	G/T
	PV	Codon 637	ASP to GLY(GAC to GGC)	A/G
	PV	Codon 648	ARG to GLN (CGA to CAA)	G/A
ABCB3 (TAP2)	PV	Codon 163	VAL to VAL (GTC to GTT)	C/T
	PV	Codon 379	ILE to VAL (ATA to GTA)	A/G
	PV	Codon 386	GLY to GLY (GGG to GGT)	G/T
	PV	Codon 387	VAL to VAL (GTG to GTA)	G/A
	PV	Codon 436	ASN to ASN (AAC to AAT)	C/T
	PV	Codon 565	ALA to THR (GCT to ACT)	G/A
	PV	Codon 651	ARG to CYS (CGT to TGT)	C/T
	PV	Codon 665	ALA to THR (GCA to ACA)	G/A
	PV	Codon 687	GLN to STOP (CAG to TAG)	C/T
	PV	Codon 697	VAL to VAL (GTT to GTG)	T/G
C4A 4B		Codon 1101	C to T	C/T
		Codon 1186	G to C	G/C
	RFLP	Exon 40-> nt 5095	CTG to CTA (no aa change)	G/A
CDKN1A	PV	Codon31	SER to ARG (C to A)	C/A
	PV	Codon 91	A to T	A/T
COL10A1	PV		GLY to ARG	
	PV2	exon 2 position 608	GTG to GTC	G/C

CYP21A1P/ A2	PV	exon 7 codon 269	SER to THR (G to C)	G/C
ESR1	PV	End of Intron 1 and before exon2 (nt 257)	ALA to VAL (C to T)	C/T
	PV	Nt 261	G to C (Silent)	G/C
GMPR	PV	Nt 766 from chain initiat. Codon	PHE to ILE (T to A)	T/A
	PV		C to T (silent)	C/T
HSPA1B	RFLP	Nt 1267	A To G	A/G
LPA	PV	Codon 4168	THR to MET (C to T)	C/T
LTA/ TNFB (ii)	RFLP	Intron 1 position 26	ASN to THR (AAC to ACC)	A/C
MEP1A	PV	Codon 176	A To G	A/G
		D6S282 To D6S272	A to G	A/G
MLN	PV	Nt115 of gene Position-11of signal peptide	VAL to ALA (T to C)	T/C
	PV	Nt 184 of 1st Exon	C To G	C/G
RDS	PV	Position 558 of gene	C To T (no aa change)	C/T
TNF (TNFA)	PV	Nt –308 of 5' of the gene	G To A	G/A
	PV	Nt –238	G To A	G/A

Appendix 3: An example of the SNP data of human chromosome 6 from fifth release of the SNP Consortium.

SNP	dbSNP #	Chrom	Band	GenBank Version	Links	Type of Change
TSC0089213	119642	Chr6	6p35.2	AL080251.3		a/g
TSC0089214	119643	Chr6	6p35.2	AL080251.3		c/t
TSC0016149	74583	Chr6	6q22.22	AL135839.2		a/g
TSC0148520	110255	Chr6	6q22.22	AL135839.2		a/g
TSC0089004	119501	Chr6	6q22.22	AL135839.2		a/g
TSC0010080	54655	Chr6	6q22.22	AL135839.2		a/g
TSC0026679	80938	Chr6	6q22.22	AL135839.2		a/g
TSC0097341	125463	Chr6	6p35.2	AL080251.3		g/t
TSC0119139	93589	Chr6	6q22.22	AL135839.2		c/g
TSC0102942	129620	Chr6		AC004842.2	NT_002179	g/t
TSC0110225	133601	Chr6	6p25	AL021328.1	NT_000213	a/c
TSC0110226	133602	Chr6	6p25	AL021328.1	NT_000213	a/c
TSC0090977	120849	Chr6	6p25	AL035696.14	NT_002179	a/g
TSC0067822		Chr6	6p25	AL035696.14	NT_002179	c/t
TSC0110227	133603	Chr6	6p25	AL021328.1	NT_000213	a/g
TSC0043505	59435	Chr6	6p25	AL035696.14	NT_002179	a/g
TSC0001995	25023	Chr6	6p25	AL035696.14	NT_002179	g/t
TSC0067821		Chr6	6p25	AL035696.14	NT_002179	a/g
TSC0102721	129487	Chr6		AC004842.2	NT_002179	a/g
TSC0069979		Chr6	6p25	AL021328.1	NT_000213	c/t

Note: This example represents 20 SNPs out of 5102 SNPs collected from the SNP Consortium database for this project.