Published online 9 August 2006

Nucleic Acids Research, 2006, Vol. 34, No. 14 ***e101***
doi:10.1093/nar/gkl520

Information-theoretic identification of predictive SNPs and supervised visualization of genome-wide association studies

Kavitha Bhasi, Li Zhang[1], Daniel Brazeau, Aidong Zhang[2] and Murali Ramanathan*

Department of Pharmaceutical Sciences and [1]Department of Computer Science, Eastern Michigan University, Ypsilanti, MI 48197, USA and [2]Department of Computer Science and Engineering, State University of New York, Buffalo, NY 14260, USA

Received March 9, 2006; Revised June 20, 2006; Accepted July 6, 2006

ABSTRACT

The size, dimensionality and the limited range of the data values makes visualization of single nucleotide polymorphism (SNP) datasets challenging. The purpose of this study is to evaluate the usefulness of 3D VizStruct, a novel multi-dimensional data visualization technique for SNP datasets capable of identifying informative SNPs in genome-wide association studies. VizStruct is an interactive visualization technique that reduces multi-dimensional data to three dimensions using a combination of the discrete Fourier transform and the Kullback–Leibler divergence. The performance of 3D VizStruct was challenged with several diverse, biologically relevant published datasets including the human lipoprotein lipase (*LPL*) gene locus, the human Y-chromosome in several populations and a multi-locus genotype dataset of coral samples from four populations. In every case, the SNPs and or polymorphic markers identified by the 3D VizStruct mapping were predictive of the underlying biology.

INTRODUCTION

Technologies capable of simultaneously genotyping thousands of single nucleotide polymorphisms (SNPs) are now widely employed in basic biomedical research for investigating the genetic basis of complex diseases, cancer risk and drug response (1–4). Presently the public SNP database (dbSNP) contains 27 million entries (Build 125, available September 2005), 10 million of which have been identified as unique to the database ('rs' SNPs). Approximately 3 million contain genotype information and >500 000 entries also have frequency data.

Many techniques have been developed to explore these multivariate datasets but one of the key obstacles of exploring genome-wide SNP data is the high dimensionality both in terms of the number of genes involved and the number of polymorphisms within each gene. Additional challenges include the massive size of the datasets (typically data on >10 000–500 000 SNPs can be obtained from a single sample) and the limited range of the data values (the data are typically sequences of ordinal numbers and number values taken by each SNP are very limited: each SNP is typically called as heterozygous or one of two homozygous states). Data analysis is further complicated by the presence of correlated markers delimiting haplotypes. Visualization algorithms can provide effective tools to summarize and interpret datasets, describe the contents and expose features in genome-wide SNP datasets. Although genotyping technologies have advanced considerably and a variety of sequence analysis and alignment algorithms and tools have been developed, analytical visualization of SNP datasets, the primary focus of this research, has not been extensively investigated in the context of SNP data analysis. Fast, efficient, effective and easy-to-use analytical visualization tools are essential for identifying and interpreting patterns in large SNP datasets in order to generate hypotheses and direct subsequent research.

METHODOLOGY AND RESULTS

The VizStruct mapping

At the core of VizStruct is a radial projection that maps the n-dimensional vectors into 2D points while retaining correlation similarity in the original input space (5,6). If the vector $\mathbf{x}[n] = (x[0], x[1], \ldots, x[n-1])$ represents a data item in n-dimensional space, R^n, its mapping to a point $F_1(\mathbf{x}[n])$ in the complex plane C is given by the following equation:

$$F_1(\mathbf{x}[n]) = \sum_{j=0}^{n-1} x[j] e^{-2\pi ij/n}. \quad \mathbf{1}$$

The real and imaginary components of $F_1(\mathbf{x}[n])$ are used for creating the 2D mapping. In Equation 1 $i = \sqrt{-1}$ and the complex exponential has the effect of dividing the circle of

*To whom correspondence should be addressed. Tel: +1 716 645 2842 (ext. 242); Fax: +1 716 645 3693; Email: murali@acsu.buffalo.edu

display into equally spaced sectors. The equation shown represents a substantive reformulation of the usual radial visualization mapping and the use of the complex number notation has significant advantages: it allows easier derivations of the theoretical underpinnings and an intuitive geometric interpretation of the mapping (7–9).

The mapping $F_1(\mathbf{x}[n])$ is equivalent to the first harmonic of the discrete Fourier transform (DFT). The relationship between the DFT and the radial visualization mapping, which was first identified by our group (7–9), allows the computationally efficient fast Fourier transform algorithm [complexity of $O(n \log n)$, where n is the number of dimensions] to be used. It allows a wide range of enhancements, including higher harmonic analysis, that have been described previously (7–9).

For 3D analysis (3D VizStruct), we included the Kullback–Leibler divergence (KLD) as the third dimension or z-coordinate; the complex number corresponding to the first Fourier harmonic is used for the x- and y-axes. The KLD between two probability mass functions $p(x)$ and $q(x)$ is denoted by $D(p\|q)$ and is also known as the relative entropy. The KLD is defined by (10):

$$\mathrm{KLD} = \sum_{x \in X} p(x) \log \left(\frac{p(x)}{q(x)} \right). \qquad \mathbf{2}$$

The base of the logarithm was taken to be 2. The KLD is a measure of the distance between two distributions or equivalently, it is the inefficiency of assuming that the distribution is q when the true distribution is p. The KLD always takes non-negative values, KLD $\geqslant 0$, and is zero only if $p = q$ (11).

As the first step, a contingency table containing the frequencies of the SNP (or polymorphic locus) genotypes in each class was obtained. The frequencies in each cell of the contingency table were normalized using the sample size in the table and these normalized frequencies comprised the probability distribution p. The reference probability distribution, q, for each cell was computed as the product of the corresponding row and column sums of the normalized frequencies table; this is equivalent to using the assumption of independence. The performance of the 3D VizStruct method was measured by calculating the percentage of samples that were misclassified.

Coding of SNP datasets

An ordinal scale was used to code the SNP genotype sequences: the numbers 1, 2 and 3 were used for genotypes that were homozygous in the major allele, heterozygous or homozygous in the minor allele, respectively.

A systematic, sequential approach was used for missing data. Individuals in whom >75% of the SNP genotypes were missing were excluded. SNP locations comprise entirely of a combination of missing data and a single genotype were excluded from the analysis because of the absence of information. In computation of the Fourier dimensions of 3D VizStruct, the remaining missing data points were replaced by the sample mean for that SNP location.

The coding and computations were conducted in Microsoft Excel (Microsoft, Bellevue, WA). The 2D and 3D plots were obtained with Kaleidagraph (Synergy Software, Malvern, PA) and MATLAB (The MathWorks, Natick, MA), respectively.

Evaluation of the VizStruct approach

Analysis of the human lipoprotein lipase genotypes. The human lipoprotein lipase (*LPL*) gene is involved in lipid metabolism and has been characterized in detail for its associations with cardiovascular disease. The *LPL* dataset from http://droog.gs.washington.edu/mdecode/data/lpl/lpl.prettybase.txt.wo_n wherein *LPL* was genotyped at 88 polymorphic sites in 48 individuals (12,13) was analyzed to assess the suitability of the 3D VizStruct approach for supervised visualization of densely characterized candidate gene. The dataset contains genotypes of 24 Americans of African ancestry from Jackson, Mississippi (JMS), who participated in the Family Blood Pressure Program, a hypertension study, and 24 Americans of European ancestry from Rochester, Minnesota (RMN), who participated in the Rochester Family Heart Study. The haplotype phase is available in this dataset, but we intentionally coded each SNP location as being either homozygous in the major allele, heterozygous or homozygous in the minor allele for visualization because haplotype phase information is generally not available in the majority of experimental situations.

Figure 1A shows the VizStruct mapping of the SNPs from LPL dataset; each point in Figure 1A corresponds to a single SNP. The SNPs with the highest values of KLD were identified and their ability to individually classify the JMS and the RMN samples was investigated. The results for the three SNPs with highest KLD values are summarized in Figure 1B–D. Each of the SNPs shown was strongly associated with and informative of the JMS versus RMN class distinction: e.g. at SNP-24 (Figure 1B), 22 of 24 JMS subjects were homozygous for the major allele and all the RMN subjects had the minor allele (i.e. they were heterozygous or homozygous for the minor allele); the percent error was 4.2%. At SNP-40 (Figure 1C), all 19 JMS subjects (genotypes for 5 subjects were missing at this locus) were homozygous for the major allele and 23 of 24 RMN subjects had the minor allele; only 1 RMN subject had the major allele and could be considered as 'misclassified' (1 of 43 or 2.3% error) by the KLD approach. At SNP-79 (Figure 1C), all 24 RMN subjects were homozygous for the major allele and 21 of 24 JMS subjects had the minor allele; 3 JMS subjects had the major allele and could be considered as 'misclassifications' (3 of 48 or 6.3% error).

These results demonstrate that the supervised 3D VizStruct approach is effective for identifying informative SNPs from datasets of densely genotyped candidate genes obtained in 2-class study designs.

The Y-chromosome dataset. Polymorphisms of the Y-chromosome are of practical interest in forensic identification, paternity testing and in the study of human migration since the chromosome is present only in males and transmitted from father to son (14–16). Figure 2A shows the VizStruct mapping of the SNPs from Y-chromosome dataset from the Perlegen Sciences database (http://genome.perlegen.com/browser/download.html) wherein the SNPs in the Y-chromosome was genotyped at 334 polymorphic sites in 33 males differing in race: 11 African-Americans,

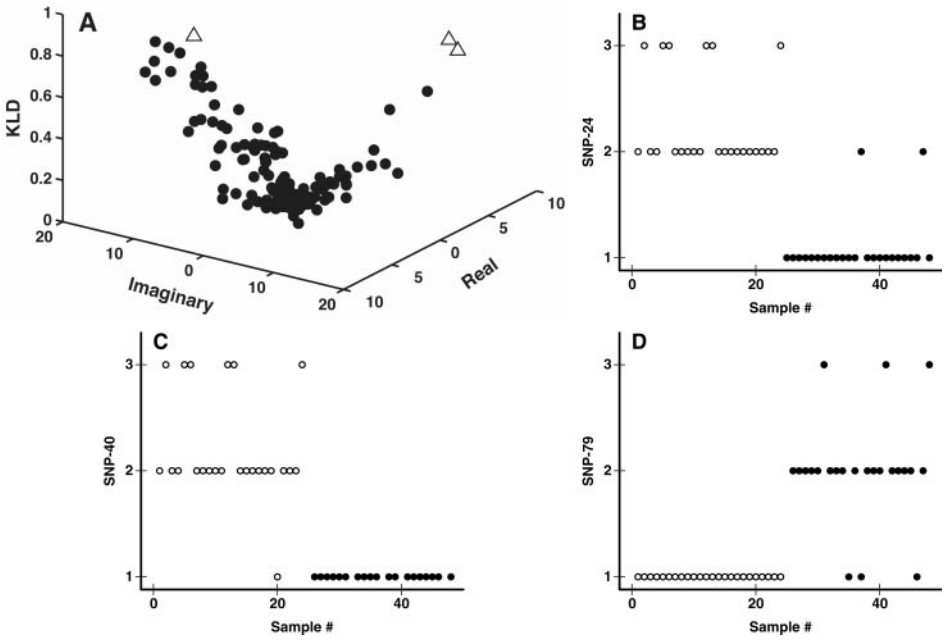

Figure 1. (**A**) (Upper left panel) shows the 3D VizStruct mapping of the *LPL* genotypes. The *x*- and *y*-axes are the real and imaginary components of the first harmonic of the DFT and the *z*-axis is the KLD; each point corresponds to a SNP and the SNPs with the highest KLD values are highlighted with the open triangles. (**B–D**) show the distribution of the genotypes for three SNPs with the highest values of the KLD in the African-American patients from Jackson, MS (closed circles) and Caucasian-American patients from Rochester, MN (open circles). The *x*-axis in (B–D) is the sample number and the *y*-axis are the genotypes with the homozygous genotypes coded as 1 and 3 for the major and minor allele, respectively, and the heterozygous genotype is coded 2.

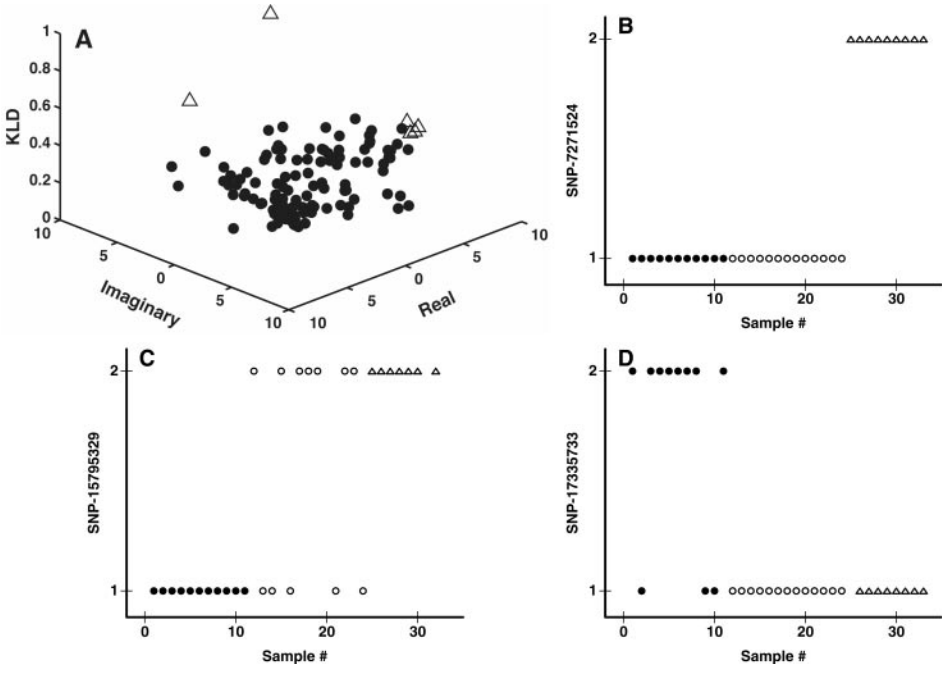

Figure 2. (**A**) (Upper left panel) shows the 3D VizStruct mapping of the Y-chromosome SNPs. The *x*- and *y*-axes are the real and imaginary components of the first harmonic of the DFT and the *z*-axis is the KLD; each point corresponds to a SNP and the SNPs with the highest KLD values are highlighted with the open triangles. (**B–D**) show the distribution of the genotypes for three SNPs with the highest values of the KLD in the African-American (closed circles) and Caucasian-American (open circles) and Han Chinese (open triangles) subjects. The *x*-axis in (**B–D**) is the sample number and the *y*-axes are the genotypes with the homozygous genotypes coded as 1 and 2 for the major and minor allele, respectively.

13 European-Americans and 9 Han Chinese (17). Figure 2B–D summarizes the descriptive capabilities of the three SNPs with the highest KLD values. The alleles for SNP7271524 shown in Figure 2A differentiate between the Han Chinese group versus the European-American and African-American groups. For SNP15795329 (Figure 2B), all the African-American

subjects have the major alleles whereas all the Han Chinese subjects have the minor allele; the European-Americans have both alleles. The pattern for SNP1733733 is also clearly distinctive: all the European-American and Han Chinese subjects have the major allele whereas the majority of African-Americans have the minor allele. This provides a demonstration of the scalability of supervised 3D VizStruct capabilities to a chromosome-wide SNP dataset containing more than two classes.

Analysis of the coral dataset. In the next analysis, we analyzed a dataset obtained by genotyping individual corals from four coral reefs populations (18). These authors used the amplification fragment length polymorphism (AFLP) assay, a multi-locus technique employed for obtaining a genetic fingerprint of organisms with limited available sequence information (19,20). In the AFLP method, a restriction enzyme digest of genomic DNA is annealed with oligonucleotides primers containing short flanking sequences in addition to the adaptor sequences of the restriction enzyme used. The flanking sequences ensure selective PCR amplification of only those restriction fragments that contain the reverse complement of the flanking sequence present. After PCR, the amplified fragments (typically, 50–100 in number) are separated according to size by denaturing gel electrophoresis (19,20). It should be noted that the AFLP method is susceptible to confounding by size homoplasy because the presence of AFLP bands of the same size in different samples is not sufficient to always assure high sequence similarity (21).

The names and locations of the reefs are summarized in Figure 3A: DNA samples from individual coral specimens from geographical locations in the Bahamas (23°28′N, 75°42′W), the Crocker and Conch reefs (two sites separated by 12 km at 24°55′N, 80°31′W near the Key Largo, FL, area) and the Flower Gardens Banks (27°55′N, 93°36′W, 110 km south-southeast of Galveston, TX in the Gulf of Mexico) were analyzed using two separate sets of primers. The number of samples, n, from the Bahamas, Flower Gardens Banks, Crocker and Conch reefs were 22, 28, 17 and 14, respectively. Coral larvae can derive from either local adult populations or immigrate from distant locations. A total of 11 samples from coral larvae (referred to as recruits) from the Flower Gardens Banks reef were also analyzed. The object of the study was to determine the likely source from which the recruits migrated; the authors used discriminant analysis to assign all but one of the recruits to the Flower Gardens banks. The data were nominal variables indicating the presence/absence of PCR products of given lengths generated from two different sets of primers. There were 45 polymorphic markers used in this study. For this dataset, the dimensions were ordered so that the mean across all the samples approximated a cosine-like function; this was achieved by sorting the results for one primer in increasing order and the other primer in decreasing order.

In the VizStruct results shown in Figure 3B, each point corresponds to a single marker. The three markers with the highest values of KLD were examined in detail (Figure 3C–E). Figure 3C demonstrates that the all but one of the samples from the Bahamas (21 of 22) are negative for Marker-8; in contrast, 26 of 28 samples from the Flower Gardens Banks reef, 13 of 17 samples from Crocker reef, 11 of 14 samples from Conch reef and all 11 samples from the recruits were positive for this marker. Figure 3D shows the results for Marker-19, which is associated with a class distinction different from that of Marker-8: the Marker-19 is present in all the Flower Gardens Banks reef and recruits samples, but it is absent in the majority of samples from the Bahamas (absent in 17 of 22 samples), Crocker (absent in 9 of 17 samples) and Conch (absent in 6 of 11 samples) reefs. Figure 3E shows the results for Marker-35, which is associated with yet another class distinction: it is absent in all the Crocker reef and 9 of 11 Conch reef samples; however, Marker-35 is present in the majority of Bahamas (present in 15 of 22 samples), Flower Gardens Banks (present in 22 of 28 samples) and recruits samples (present in 10 of 11 samples). These representative results demonstrate that the KLD is capable of identifying informative genetic polymorphisms when multiple classes are present. The 3D VizStruct analysis also indicates that the recruits are most similar to the samples from the Flower Gardens Banks, which is consistent with the findings of Brazeau *et al*. (18). Discriminant analysis, which was used by Brazeau *et al*. (18), is the conventional 'gold standard' methodology for identifying predictive markers from multi-dimensional datasets. We therefore challenged 3D VizStruct by comparing the three markers with the highest KLD in the 3D VizStruct visualization to the markers present with highest weights in the discriminant function: there was an exact concordance between the top three markers identified by both methods.

These results extend the useful range of 3D VizStruct visualization capabilities to include multi-class data generated by multi-locus genotyping techniques that are used for poorly characterized genomes.

Kullback–Leibler divergence and linkage disequilibrium

The relationship between the KLD and linkage disequilibrium (LD) was analyzed to provide further justification for the use of the KLD in VizStruct for SNP visualization.

A variety of normalized metrics (e.g. R^2 and Lewontin's D'), all of which are based on linkage disequilibrium, D, are widely used in genetic mapping. We therefore first investigated the relationship between the KLD and LD.

The starting point for the definition of the measures of linkage disequilibrium is the standard 2×2 haplotype frequency table shown on in Table 1. Consider two loci, each of which has two alleles. Let A and a denote the major and minor alleles at the first locus and B and b denote the major and minor alleles at the second locus. The proportions of the A, a, B and b alleles are denoted by p_A, p_a, p_B and p_b, respectively. Similarly, denote the proportion of the AB, Ab, aB and ab haplotypes by p_{AB}, p_{Ab}, p_{aB}, p_{ab}, respectively. Linkage disequilbrium D is defined by the following equivalent equations (22):

$$D = p_{AB}p_{ab} - p_{Ab}p_{aB} = p_{AB} - p_Ap_B = p_{ab} - p_ap_b = -p_{Ab} + p_Ap_b = -p_{aB} + p_ap_B. \quad \mathbf{3}$$

Because the reference distribution q in the definition of KLD (Equation 2) is based on the assumption of independence, the KLD of the haplotype frequency table is given by the

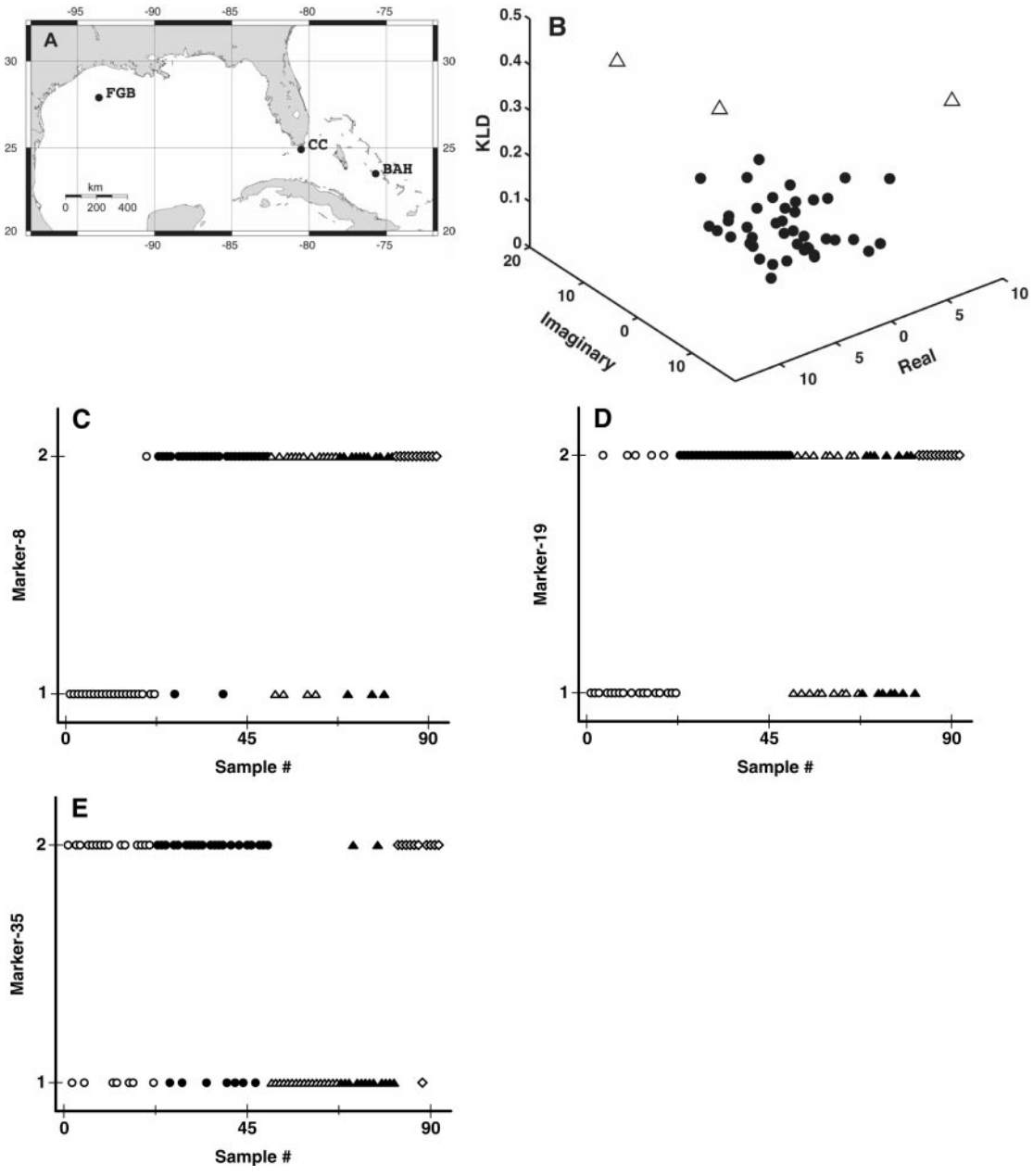

Figure 3. (**A**) (Upper panel) is a map of the southeastern United States (made using M. Weinelt, Online Map Creation: http://www.aquarius.geomar.de/omc/make_map.html) showing the locations of the Bahamas (BAH), Crocker and Conch (CC) and Flower Gardens Banks (FGB) coral reefs from which the samples were derived. The grid on the map indicates latitude north and longitude west. (**B**) Shows the 3D VizStruct mapping of the genotyping results from AFLP analysis of the coral samples. The *x*- and *y*-axes are the real and imaginary components of the first harmonic of the DFT and the *z*-axis is the KLD; each point corresponds to a marker and the markers with the highest KLD values are highlighted with the open triangles. (**C**–**E**) Show the distribution of the genotypes for three amplification fragments with the highest values of the KLD in the samples from the Bahamas (open circles, $n = 22$), the Flower Garden Banks (filled circles, $n = 28$), Crocker (open triangles, $n = 17$), Conch (filled triangles, $n = 14$) and the recruits from the Flower Garden Banks (open diamonds, $n = 11$). The *x*-axis is the sample number and the *y*-axis are the genotypes; the genotype was coded as 1 if the fragment was absent and 2 if the fragment was present.

following equation:

$$\mathrm{KLD} = p_{AB}\log\frac{p_{AB}}{p_A p_B} + p_{Ab}\log\frac{p_{Ab}}{p_A p_b} + p_{aB}\log\frac{p_{aB}}{p_a p_B} + p_{ab}\log\frac{p_{ab}}{p_a p_b}. \quad \mathbf{4}$$

Using the Equation 3, Equation 4 can be re-written in terms of the linkage disequilibrium and the allele frequency terms alone:

$$\mathrm{KLD} = (D + p_A p_B)\log\frac{(D + p_A p_B)}{p_A p_B} + (p_A p_b - D)\log\frac{(p_A p_b - D)}{p_A p_b} + (p_a p_B - D)\log\frac{(p_a p_B - D)}{p_a p_B} + (D + p_a p_b)\log\frac{(D + p_a p_b)}{p_a p_b}. \quad \mathbf{5}$$

Equation 5 formally defines the relationship between linkage disequilibrium and KLD: the KLD depends only on the linkage disequilibrium and on the allele frequencies.

Figure 4 shows the results from numerical experiments designed to investigate the dependence of KLD on allele frequency. The experiment employed 87 simulated datasets wherein the allele frequencies at one locus (locus *B*) were kept constant at ($p_B = 0.9$, $p_b = 0.1$) whereas the allele frequency at the other locus was varied from ($p_A = 0.99$, $p_a = 0.01$) to ($p_A = 0.6$, $p_a = 0.4$). The numerical values of the KLD were calculated for various values of linkage disequilibrium (D). Figure 4 summarizes the relationship between the linkage disequilibrium and KLD with allele frequency as a parameter on linear (Figure 4A) and logarithmic axes (Figure 4B); the curves in Figure 4A appear to 'end' because there are maximum limits to D for a given set of allele frequencies. Figure 4A and B demonstrates that there is direct relationship between the KLD and D despite their disparate underlying formulations—the KLD is based on an information-theoretical framework, whereas D is the determinant of the haplotype probability table. However, over the range examined, for a given allele frequency at one locus, the dependence between KLD and D is approximated by a power-law relationship of the form KLD $\propto D^n$ (Figure 4B); the exponent of the power-law relationship varied between 1.78 for ($p_A = 0.99$, $p_a = 0.01$) to 2.00 for ($p_A = 0.6$, $p_a = 0.4$).

DISCUSSION

The objective of this report was to evaluate 3D VizStruct, a multi-dimensional visualization approach that combines radial visualization with the KLD for visualizing SNP data and for identifying predictive SNPs from large association studies. We analyzed several datasets to demonstrate the usefulness of the 3D VizStruct approach for mining SNP data obtained from studies of large datasets including a densely genotyped candidate gene, *LPL*, the Y-chromosome and samples of reef coral individuals obtained from the wild. We highlighted the effectiveness of identifying predictive SNPs from 3D VizStruct visualization in two-class and multi-class study designs.

Our results demonstrate the 3D VizStruct approach with the KLD in particular is effective for the supervised detection of informative SNPs. The KLD is valuable for identifying class distinctions because it is order-insensitive; however, the availability of the complex Fourier harmonic dimensions for visualization is also important because it spreads the large number of SNPs efficiently in the visualization field, which allows different categories of class distinctions of comparable quality to clearly emerge in multi-class datasets. The KLD is also easy to interpret by users because it represents a 'distance' between two distributions, i.e. two SNP distributions that are similar are placed near each other and those dissimilar are placed distantly from each other. In this context, we are currently conducting theoretical studies to examine the relationships between our supervised KLD formulation and the commonly used measures of linkage disequilibrium, e.g. the correlation coefficient Δ, Lewontin D' (23,24), Yule's Q (25), Kaplan and Weir' proportional difference d (26), the population attributable risk δ (27), because linkage disequilibrium measures remain the fundamental approach used by population geneticists and genetic epidemiologists to identify disease loci (22) and are frequently used for fine mapping by molecular geneticists; identifying these analytical relationships will facilitate acceptance of the 3D VizStruct visualization method. A common feature of linkage disequilibrium measures is that they assess the difference between the observed and expected frequencies of haplotypes between a 'disease' locus and a marker locus of interest (22). Our working hypothesis is that differences between two loci in the KLD dimension are directly related to measures of linkage

Table 1. Haplotype table

	B	*b*	Sum
A	p_{AB}	p_{Ab}	p_A
A	p_{aB}	p_{ab}	p_a
Sum	p_B	p_b	1

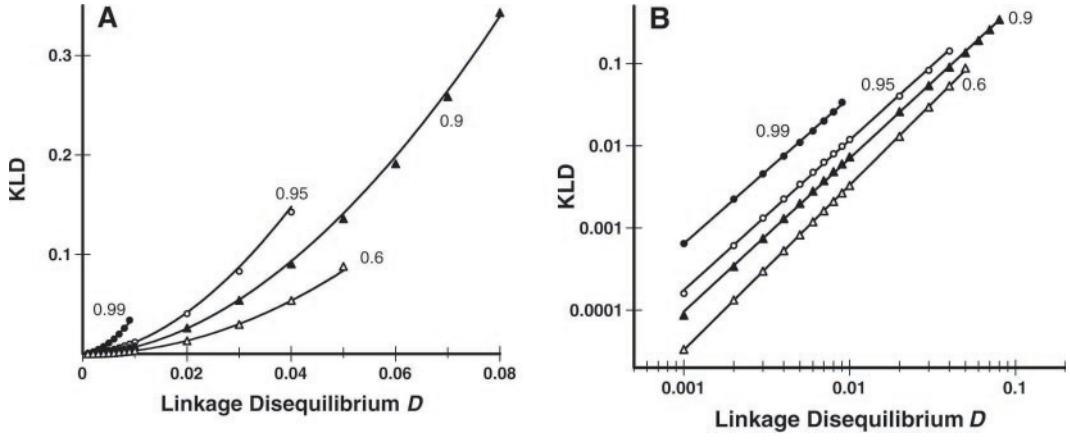

Figure 4. Relationship between the KLD versus the linkage disequilibrium D for a range of allele frequencies. The allele frequencies at one locus were kept constant at 0.9 for the major allele (**A**) and 0.1 for the minor allele (**B**). The major allele frequencies at the other locus were varied as indicated and were 0.99 (filled circles), 0.95 (open circles), 0.9 (filled triangles) or 0.6 (open triangles). The solid lines are a power-law fit to the results. Figure 1A uses linear axes and Figure 1B shows the same data on logarithmic axes.

disequilibrium between the loci. Additional relationships will be systematically derived in the further research but preliminary analyses are promising; e.g. the KLD of two loci is zero in the case the two loci are completely independent of the class distinction and larger for larger values of linkage disequilibrium. One important and useful difference is that the 3D VizStruct approach does not require computation of pairwise linkage disequilibrium values across loci: each SNP vector is individually projected on the visualization field and this reduces the computation needed. An additional advantage with the Fourier and KLD components of 3D VizStruct is that both are generalizable to multiple SNPs and across a wide range of data types. For example, the same approaches could potentially be extended to visualize the associations between SNPs and quantitative traits (e.g. gene expression, clinical and laboratory parameters).

The KLD approach can also be used to identify statistically predictive SNPs because the log-likelihood ratio is equal to the product of the KLD and sample size. For a 2×2 haplotype table with a given sample size, the KLD is distributed according to the χ^2 statistic with one degree of freedom. With these simple relationships and the desired level of significance at hand, predictive genes can be identified easily along the KLD axis.

The 3D VizStruct approach represents a computationally efficient means to visually examine large complex datasets form diverse areas of research. With the ever-expanding numbers of massive datasets there is a real need for visualization tools that can quickly encapsulate the information allowing researchers to more easily interpret data, identify and summarize the most important features in genome-wide SNP datasets.

ACKNOWLEDGEMENTS

This work was supported in part by grants from the Kapoor Foundation, National Science Foundation (Research Grant 0234895) and the National Institutes of Health (P20-GM 067650). Funding to pay the Open Access publication charges for this article was provided by the National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Mir,K.U. and Southern,E.M. (2000) Sequence variation in genes and genomic DNA: methods for large-scale analysis. *Annu. Rev. Genomics Hum. Genet.*, **1**, 329–360.
2. Erichsen,H.C. and Chanock,S.J. (2004) SNPs in cancer research and treatment. *Br. J. Cancer.*, **90**, 747–751.
3. Suh,Y. and Vijg,J. (2005) SNP discovery in associating genetic variation with human disease phenotypes. *Mutat. Res.*, **573**, 41–53.
4. Xu,H., Gregory,S.G., Hauser,E.R., Stenger,J.E., Pericak-Vance,M.A., Vance,J.M., Zuchner,S. and Hauser,M.A. (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. *Bioinformatics*, **21**, 4181–4186.
5. Bhadra,D. and Garg,A. (2001) *An Interactive Visual Framework for Detecting Clusters of a Multidimensional Dataset*. Technical Report 2001–03, State University of New York, Buffalo.
6. Hoffman,P.E., Grinstein,G.G. and Marx,K. (1997) DNA visual and analytic data mining. In *Proceedings of the IEEE Visualization '97*, Phoenix, AZ, pp. 437–441.
7. Zhang,L., Zhang,A. and Ramanathan,M. (2002) Visualized classification of multiple sample types. In *Proceedings of the 2nd Workshop on Data Mining in Bioinformatics (BIOKDD 2002), The ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada.
8. Zhang,L., Zhang,A. and Ramanathan,M. (2003) Enhanced visualization of time series through higher Fourier harmonics. In *Proceedings of the Third ACM SIGKDD Workshop on data mining in bioinformatics (BIOKDD03)*, pp. 49–56.
9. Zhang,L., Zhang,A. and Ramanathan,M. (2004) VizStruct: exploratory visualization for gene expression profiling. *Bioinformatics*, **20**, 85–92.
10. Cover,T.M. and Thomas,J.A. (1991) *Elements of Information Theory*. Wiley, NY.
11. Haykin,S. (1999) *Neural Networks: A Comprehensive Foundation*. College Publishing Co., NY.
12. Nickerson,D.A., Taylor,S.L., Weiss,K.M., Clark,A.G., Hutchinson,R.G., Stengard,J., Salomaa,V., Vartiainen,E., Boerwinkle,E. and Sing,C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet.*, **19**, 233–240.
13. Clark,A.G., Weiss,K.M., Nickerson,D.A., Taylor,S.L., Buchanan,A., Stengard,J., Salomaa,V., Vartiainen,E., Perola,M., Boerwinkle,E. *et al.* (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.*, **63**, 595–612.
14. Butler,J.M. (2005) *Forensic DNA Typing: Biology, Technology and Genetics of STR Markers*, *2nd edn*. Elsevier Press, New York, NY.
15. Hammer,M.F., Karafet,T.M., Redd,A.J., Jarjanazi,H., Santachiara-Benerecetti,S., Soodyall,H. and Zegura,S.L. (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol.*, **18**, 1189–1203.
16. Rosser,Z.H., Zerjal,T., Hurles,M.E., Adojaan,M., Alavantic,D., Amorim,A., Amos,W., Armenteros,M., Arroyo,E., Barbujani,G. *et al.* (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.*, **67**, 1526–1543.
17. Hinds,D.A., Stuve,L.L., Nilsen,G.B., Halperin,E., Eskin,E., Ballinger,D.G., Frazer,K.A. and Cox,D.R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science*, **307**, 1072–1079.
18. Brazeau,D.A., Sammarco,P.W. and Gleason,D.F. (2005) A multi-locus genetic assignment technique to assess sources of *Agaricia agaricites* larvae on coral reefs. *Marine Biol.*, **147**, 1141–1148.
19. Vos,P. (1998) AFLP fingerprinting of *Arabidopsis*. *Methods Mol. Biol.*, **82**, 147–155.
20. Vos,P., Hogers,R., Bleeker,M., Reijans,M., van de Lee,T., Hornes,M., Frijters,A., Pot,J., Peleman,J., Kuiper,M. *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.*, **23**, 4407–4414.
21. Peakall,R., Gilmore,S., Keys,W., Morgante,M. and Rafalski,A. (1998) Cross-species amplification of soybean (Glycine max) simple sequence repeats (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. *Mol. Biol. Evol.*, **15**, 1275–1287.
22. Devlin,B. and Risch,N. (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, **29**, 311–322.
23. Lewontin,R.C. (1988) On measures of gametic disequilibrium. *Genetics*, **120**, 849–852.
24. Lewontin,R.C. and Matsuo,Y. (1963) Interaction of genotypes determining viability in *Drosophila busckii*. *Proc. Natl Acad. Sci. USA*, **49**, 270–278.
25. Yule,G.U. (1900) On the association of attributes in statistics. *Philos. Trans. R Soc. Lond.*, **194**, 257–319.
26. Kaplan,N. and Weir,B.S. (1992) Expected behavior of conditional linkage disequilibrium. *Am. J. Hum. Genet.*, **51**, 333–343.
27. Levin,M.L. and Bertell,R. (1978) RE: 'simple estimation of population attributable risk from case-control studies'. *Am. J. Epidemiol.*, **108**, 78–79.