Send Orders for Reprints to reprints@benthamscience.net

The Open Bioinformatics Journal, 2013, *7*, 19-24

Open Access

Haplotype Classification Using Copy Number Variation and Principal Components Analysis

Kevin Blighe*

Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield, S10 2TH, United Kingdom

Abstract: Elaborate downstream methods are required to analyze large microarray data-sets. At times, where the end goal is to look for relationships between (or patterns within) different subgroups or even just individual samples, large data-sets must first be filtered using statistical thresholds in order to reduce their overall volume. As an example, in anthropological microarray studies, such 'dimension reduction' techniques are essential to elucidate any links between polymorphisms and phenotypes for given populations. In such large data-sets, a subset can first be taken to represent the larger data-set. For example, polling results taken during elections are used to infer the opinions of the population at large. However, what is the best and easiest method of capturing a sub-set of variation in a data-set that can represent the overall portrait of variation?

In this article, principal components analysis (PCA) is discussed in detail, including its history, the mathematics behind the process, and in which ways it can be applied to modern large-scale biological datasets. New methods of analysis using PCA are also suggested, with tentative results outlined.

Keywords: Principal components analysis, multivariate data analysis, haplotype-tagging, copy number variation.

INTRODUCTION

Principal components analysis and other multivariate tools are used to analyze large volumes of data in order to tease out the differences/relationships between the logical entities being analyzed (for example, a data-set consisting of a large number of samples, each with their own data points/varia-bles) [1]. It extracts the fundamental structure of the data without the need to build any model to represent it [2]. This 'summary' of the data is arrived at through a process of reduction that transforms the large number of variables into a lesser number that are uncorrelated (i.e. the 'principal' components), whilst at the same time being capable of easy interpretation on the original data [3, 4].

Principal components analysis has broad applications and is used in a wide range of areas. Examples include craniofacial recognition [5], analysis of water quality[3], and to derive a set of highly confident genes [6] or single nucleotide polymorphisms (SNPs) [7, 8] for classification purposes. It has also been used in subject areas such as climatology, geology, meteorology, psychology, quality control [4], forensics and population genetics (particularly in relation to SNPs), medical genetics [2], and bacteriology [9]. It can also help in the identification of subgroups within samples by visually scanning the resulting bi-plot created to represent the data [10]. There has also been notable success of applying PCA to protein datasets. Du [11] successfully adapted and applied PCA to protein data in the form of Amino Acid PCA (AAPCA), where the aim was to classify proteins into structural classes; meanwhile, Li [12] combined PCA with continuous wavelet transform (CWT) to also successfully predict protein structural classes; Zhao [13] also used PCA to help predict protein-protein interaction (PPI) networks by using this method to first derive an optimised subset and then using this subset as input to a support vector machine (SVM). Chou [14] also outlines 'pseudo-amino acid composition' as a means of managing and using the large amount of protein sequence that is currently held in public repositories. With pseudo-amino acid composition and PCA, patterns in protein sequences can be found, which can then be used to infer the cellular attributes of the corresponding proteins. Pseudo-amino acid composition and PCA has also been employed by Liu [15].

In PCA experiments, two different approaches can be taken: 1) looking at relationships between variables; and 2) looking at relationships between samples. If only two variables are involved, then a simple linear correlation analysis can be employed. However, having numerous variables prevents this [3].

Post-PCA Analysis

After carrying out PCA and deriving the principal components, there is no standard way of choosing how many components to include or exclude in end-analysis, a fact that is probably related to the broad spectrum of analyses on which PCA is carried-out. The end goal of the study is of critical importance: if you wanted to determine the variables that defined differences between samples, then you would observe the first few principal components (or even just the first); if you wanted to determine the variables that were common across samples, then you would observe the last few. If it was the former, then choosing a certain number of components whose combined percentage of variance accumulated to a pre-determined level (generally ≥70%)

*Address correspondence to this author at the Sheffield Children's NHS Foundation Trust, Western Bank, Sheffield, S10 2TH, United Kingdom; Tel: +44 7500 190333; E-mail: kevinblighe@gmail.com

1875-0362/13

could be employed [4, 16]. They could also be chosen using 'Kaiser's rule [17]', which states that all principal components with an eigenvalue greater than 1 should be retained, or by invoking and analyzing a Scree plot [18], which shows a line-graph of the eigenvalues of each principal component (see Appendix for detailed information on how principal components are derived).

After deriving the principal components, three methods for arriving at a subset of variables/markers for end-analysis can be pursued: 1. observe the resulting PCA bi-plot, remove a pre-determined number of markers, and then re-create the bi-plot to see if the original structure and variance is still visually/graphically represented. If the same structure is present, then remove further markers and repeat until a manageable subset of markers has been chosen. In essence, this method involves comparing the principal components derived from subsets of markers to those of the full set, but key markers could be lost in the process. 2. Only choose markers that have high correlation coefficients to each of the generated eigenvectors of the chosen principal components and then search for overlap between them [4]. There is no standard cut-off value for the correlation coefficients, but Mahloch [19] indicates that a value larger than 0.6 (r^2) is sufficient, while a coefficient larger than 0.8 is regarded as good. Correlation coefficients close to 0 indicate that the marker is not significantly-contributing to variance and is common across samples [2, 20].

A third method of selecting the markers to include in end-analysis is to first perform an orthogonal rotation of the derived principal components. There are a number of ways to do this, including varimax, quartimax, equamax, and promax rotation. The rotation of principal components is performed to increase the accuracy of the relationship/correlation of the original markers to the newly-derived eigenvectors (principal components). As a result, this also serves to maximize the differences between each eigenvector [3,7], in respects resembling the complete linkage method in hierarchical clustering.

Limitations of Principal Components Analysis

There are some limitations to employ just PCA as the sole analytical tool for large data-sets. For example, PCA is a linear transformation; thus any data-set that is non-linear will not be represented sufficiently after data-reduction. In addition, PCA assumes that the directions with the largest variances are of most interest [21], which might not necessarily be true. It also follows the assumption that the original data variables are correlated - if they are not, then PCA cannot reduce the data [21, 22].

Case Study: Deriving a Genetic Signature

A method used to minimally characterize individual samples from large data-sets was previously employed by finding haplotype-tagging SNPs (htSNPs) - i.e. a reduced list of SNP markers that captured the majority of haplotype diversity in a population [8] - whilst Horne and Camp [23] devised a separate method that aimed to find SNPs in linkage disequilibrium (LD), known as group-tagging SNPs (gtSNPs). The original htSNP method, which was devised by Meng [7], was only applied to analyze certain regions of the genome. Additionally, redundancy could still exist using this approach if, for example, 2 or more of the derived htSNPs mapped to regions affected by linkage disequilibrium (LD) [2, 24], which can encompass segments ranging in size from 1-100Kb on the same chromosome and can also be used for classification purposes [25]. If two alleles are in LD with each other, measuring the value of one can reveal with a certain level of confidence that of the other due to their high correlation [24]. It can arise for different reasons, such as selection, favorable mutations, population mixture and migration, et cetera; if measured around known genes, it can help in the elucidation of those that are involved in disease [25, 26]. It is more common around genes that are 'rare' or recently evolved as these would have had less time for recombination to break the disequilibrium in question [25].

Lin & Altman [8] found that using this method by Meng [7] could produce a list of htSNPs that achieved a 90% reconstruction precision of each observed haplotype. The process involved the generation of eigenvectors (the synonymous term 'eigenSNPs' was coined by Meng [7]) and then reverse-mapping these to original SNPs to arrive at a minimal set that could define maximum diversity/variance. Single nucleotide polymorphisms are useful in this regard because they are regarded as the most common type of genetic variation in the human genome [27].

The derivation of eigen SNPs has also been employed in population and anthropological studies. The present-day genetic variation in humans around the world exists and has been strengthened by the history of migration patterns [2,28]. What contributed to this included mutation, genetic drift, and natural selection that were each driven by an interaction with new climates, pathogens, etc. [2]. This prompted Paschou [2] to attempt to build a scoring system to assign an individual of unknown origin to a population group based on PCA and SNPs - some success was achieved. However, their system was based on diverse population groups between whom was exhibited already-known differences/separation, both genetically and geographically. Thus, their method of assignation to a particular clade/population was made easier. Also, they only had small sample sizes from each observed group.

The method has yet to be applied to other polymorphisms in the human genome. However, through the use of a high-density microarray that can scan the entire genome, applying the htSNP method could generate a genetic signature for a particular population group or even disease state, if the study was such. In the latter sense, it could potentially provide a minimum set of diagnostic and prognostic markers and assist in disease-type classification.

Thus, the aim of this work was to derive a set of haplotype-tagging copy number variants (htCNVs) amongst 128 female samples from the International HapMap that could be used for assignation and characterization purposes to deduce the origin of unknown samples in the future.

METHODS

In total, 270 (128 female and 172 male) Genome-Wide Human SNP Array 6.0 (Affymetrix SNP 6.0) CEL files from the International HapMap (build 270 na30 r1 a5) (International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/) were processed, incorporating the following sub-

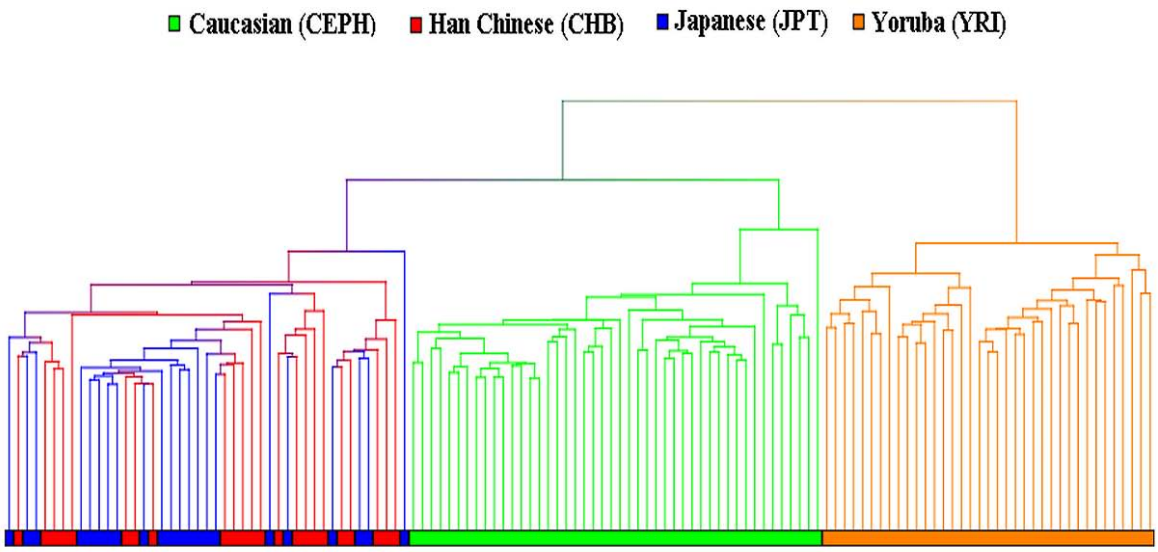

Fig. (1). Principal components analysis incorporating the htCNV pipeline reveals a reduced set of 4,594 copy number markers that can distinguish the International HapMap sub-population groups' female samples in hierarchical clustering.

population groups: CEPH (Utah, USA residents with ancestry from northern and western Europe); CHB (Han Chinese in Beijing, China); JPT (Japanese in Tokyo, Japan); and YRI (Yoruba in Ibadan, Nigeria).

Principal Components Analysis

Principal components analysis (PCA) was performed as follows: In brief, principal components were determined using a covariance matrix method (used for mean-centred data) with normalized eigenvector scaling (see Supplementary Methods for detailed information on how principal components were derived). A Bonferroni-corrected $p< 0.0001$ for multiple comparisons was used to filter-out markers of insignificance before determining principal components. False discovery rate was not used as it was considered too lenient and unsuitable for the amount of comparisons involved. The component loadings for each of the derived principal components were rotated using varimax rotation with Kaiser normalization [31] on a UNIX-based system using custom R [32] scripts.

Haplotype-tagging CNV

For deriving the genetic signature of variation amongst the HapMap samples, the method according to Meng [7] was applied as follows: Using absolute values on all component loadings, the mean correlation coefficient for each marker to the first few principal components whose total variance accumulated to ≥70% was obtained. Then, the corresponding mean correlation coefficient for the marker was calculated for all remaining principal components. If the mean of a marker for the first set of components was greater than its corresponding mean for the remaining components, then the marker was included.

Pathway analysis and keyword/term-enrichment for genes was performed using DAVID [33, 34].

RESULTS

On the Affymetrix SNP 6.0, 762,463 markers target known genes that are listed in the RefSeq gene database [29]. After pre-filtering for markers of difference amongst the 128 healthy female HapMap samples through a Bonferroni-corrected ANOVA, the number of markers for PCA was reduced to 5,896. The data generated through PCA was then channelled through the htCNV pipeline, which was capable of reducing them further to 4,594. This reduced number of markers covered a total of 2,866 genes that were significantly driving differences based on copy number between the four HapMap sub-population groups analyzed. A total of 1,893 of these genes were enriched for the UniProt [30] key-term 'sequence variant', whilst 1,838 were enriched for the keyword 'polymorphism'. In addition, 1,412 were genes that are expressed in the brain, while 456 are expressed in the epithelium.

Hierarchical clustering using this severely reduced number of 4,594 markers was capable of distinguishing the different sub-populations (Fig. **1**). However, similarities were revealed between the CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan). The dendrogram also revealed that the YRI (Yoruba in Ibadan, Nigeria) were distinct from all other groups. The remaining samples from the International HapMap, which comprised 172 healthy male samples, were then added and clustered using the same panel of markers to again reveal sub-groups based on both ethnicity and sex (Fig. **2**).

DISCUSSION AND CONCLUSION

Thus, the htCNV method -through PCA- is capable of defining a reduced number of markers for characterization purposes in a large sample cohort. Moreover, it is then robust in the sense that new samples can be prospectively added to the cohort using these markers with correct classification based on both population group and sex. It is reasonable to suggest that this method could be applied to other larger data-sets and used to derive panels of markers for diagnostic purposes. For example, in metabolomic studies, it could be used to drastically reduce the large -and sometimes incoherent- number of variables to a select few that had much meaning between a healthy and disease state.

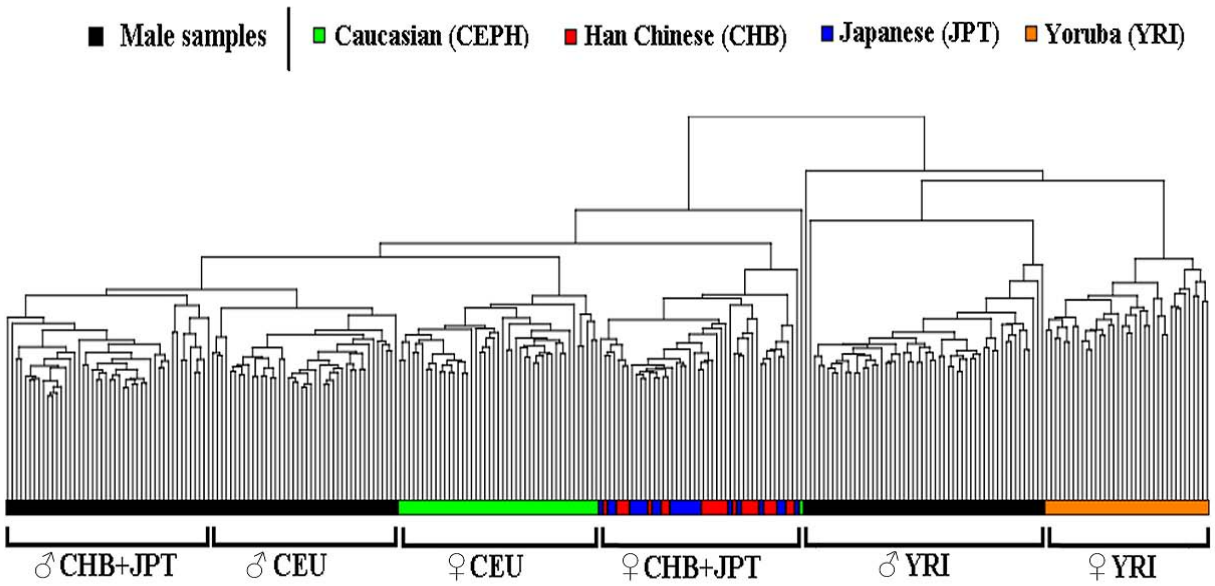

Fig. (2). By adding 172 male samples and using the same markers as per Fig. 1, the htCNV pipeline is also shown to be capable of distinguishing between both sex and sub-population group within the Internationall HapMap.

Although the derived, reduced set of markers cannot provide 100% confidence that the genomic loci to which they target are indeed capable of classifying the population groups studied -and that further work in the wet-laboratory would be required to prove this- it is my belief that the results herewith are evidence of the sound computational methodology employed. Indeed, such a method had not previously been applied to copy number markers in the human genome; thus, the results show how copy number loci can equally provide for haplotype classification along with other polymorphism-types in the human genome.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENT

I would like to thank Bentham Science, for their great support during the peer-review and publication process.

APPENDIX

Principal Components Analysis

The process of reducing data using PCA involves the following steps: 1. subtract mean; 2. build covariance matrix; 3. calculate eigenvectors and eigenvalues; 4. transform original data; 5. end-analysis.

The first step in PCA is to subtract the mean of each data-set from each data-point in the set. For example, if we had data-sets X, Y, and Z, then from each value in X, Y, and Z, we would subtract the mean of each data-set, respectively. This would result in the data-sets having means of 0. The covariance matrix is then built using the mean-subtracted data-sets (note: if the variables were measured in different units, then a correlation matrix should be employed, whereas if the variables were in the same units and are mean-centred, then a covariance matrix should be employed). Thirdly, eigenvectors (a) and eigenvalues (e) for the built covariance matrix are determined. Eigenvectors characterize the data using straight, orthogonal lines and each is scaled with an eigenvalue. Eigenvalues indicate the amount of variance that a component contains (similar to the percentage of variance). The eigenvector is the direction cosine of the axis of the principal component, such that:

$$\sum_{v=1}^{n} z_v = a_v x$$

Where: n=Number of original variables; Z_v=v^{th} principal component; a_v=v^{th} eigenvector; x=vector of the original variables

As mentioned, it also holds true that the variance of a principal component is equivalent to its corresponding eigenvalue:

$$\text{var}(z_v) = e_v$$

Where: Z_v=v^{th} principal component; e_v=v^{th} eigenvalue

The eigenvectors are then ordered by their eigenvalue (highest to lowest), indicating the level of significance to the data-set.

Transformation of the original data then occurs by multiplying a matrix containing the derived eigenvectors by one containing the mean-subtracted original data. The result is a matrix with the same dimensions as that of the mean-subtracted data but whose values have been transformed. The original data can then be said to be expressed with regard to the patterns within it. Once the data has been transformed, end-analysis can be performed, which can involve the selection of significant data-points and eigenvectors - it can also involve viewing the data on bi-plots.

The transformed data contains the principal components and each is assigned a percentage that corresponds to the amount of total variance in the data towards which the component contributes. The relationship of the original variables to the new components is represented by correlation (r) values that are scaled between -1 and +1. If there are variables having high correlation to one or more of the derived components, then most of the variation will be accounted for by these components and such variables will be the ones that are accounting for the differences amongst samples. The last few, in such a case, will account for little variation as they will define constant or near-constant linear relationships amongst the variables (i.e. variables whose values were common across samples).

Eigenvectors and Eigenvalues

Common statistical measures include the mean ($\overline{X}$) and standard deviation (σ). However, variance (var), which is merely the square of the standard deviation (σ2), and covariance (cov) can also be used. To derive the eigenvectors and eigenvalues for a set of data, the covariance matrix must first be derived.

The standard deviation and variance are represented as follows:

$$\sigma = \sqrt{\frac{\sum_{i=i}^{n}\left(X_i - \overline{X}\right)^2}{(n-1)}}$$

$$\text{var} = \sigma^2 = \frac{\sum_{i=i}^{n}\left(X_i - \overline{X}\right)^2}{(n-1)}$$

Where: n=Number of values in the data-set; $\overline{X}$ =Mean of the data-set; X_i=i^{th} element of the data-set

These measure the spread of the data. However, whereas the variance is one-dimensional, covariance is used on two-dimensional data (for example, measuring the relationships between the height of a person and their body-weight, or between hours studied and exam results). Covariance retains the same formula as variance, except for a minor difference:

$$\text{cov}(XY) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1)}$$

Where: n=Number of values in the data-set; $\overline{X}$ & $\overline{Y}$ =Means of data-sets X and Y; X_i & Y_i=i^{th} elements of data-sets X and Y

For n-dimensional data-sets, a covariance matrix can be constructed that represents the covariance between each possible combination of data-sets. For example, a covariance matrix for a three-dimensional data-set (X, Y, Z) would look like the following:

$$C = \begin{bmatrix} \text{cov}_{XX} & \text{cov}_{XY} & \text{cov}_{XZ} \\ \text{cov}_{YX} & \text{cov}_{YY} & \text{cov}_{YZ} \\ \text{cov}_{ZX} & \text{cov}_{ZY} & \text{cov}_{ZZ} \end{bmatrix}$$

The matrix is symmetrical across the main diagonal. Also, the covariance values on this diagonal are equivalent to simply finding the covariance of each particular data-set to itself. If we multiplied two matrices together (one a transformation matrix and the other a vector), the result would be a matrix that is either an integer-multiple of the original vector or not. Integer-multiples are eigenvectors, with the scaling value being the eigenvalue. For example, observe the following:

$$A\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix} B\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 20 \\ 20 \end{bmatrix} = 10\begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Where: A=Transformation matrix; B=Vector; 10=Eigenvalue; $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$=Eigenvector

Eigenvectors are orthogonal (i.e. - perpendicular) to each other. For any nXn matrix, only n possible eigenvectors can be found, and they are scaled to have a length of 1. Visually, the orthogonal nature of eigenvectors means that at most 3 eigenvectors can be displayed on a three-dimensional plot.

REFERENCES

[1] L. I. Smith, "A tutorial on Principal Components Analysis", February 2002. Available from: http://www.cs.otago.ac.nz/cosc4-53/student_tutorials/principal_components.pdf

[2] P. Paschou, J. Lewis, A. Javed, and P. Drineas, "Ancestry informative markers for fine-scale individual assignment to worldwide populations", *J. Med. Genet.*, vol. 47, pp. 835-847, 2010.

[3] N. Mazlum, A. Özer, and S. Mazlum, "Interpretation of Water Quality Data by Principal Components", *Turk. J. Eng. Environ. Sci.*, vol. 23, pp. 19-26, 1999.

[4] I.T. Joliffe Ed., *Principal Components Analysis,* 2nd *Ed*. Springer: New York, 2002.

[5] M.M. Chakravarty, R. Aleong, G. Leonard, M. Perron, G.B. Pike, L. Richer, S. Veillette, Z. Pausova, and T. Paus, "Automated analysis of craniofacial morphology using magnetic resonance images", *PLoS One*, vol. 6, no. 5, e20241v, 2011. [Online] Available: http://www.plosone.org/

[6] S. Bicciato, A. Luchini, and C. Di Bello, "PCA disjoint models for multiclass cancer analysis using gene expression data", *Bioinformatics*, vol. 19, pp. 571-578, 2003.

[7] Z. Meng, D.V. Zaykin, C.F. Xu, M. Wagner, and M.G. Ehm, "Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes", *Am. J. Hum. Genet.*, vol. 73, pp. 115-130, 2003.

[8] Z. Lin and R.B. Altman, "Finding haplotype tagging SNPs by use of principal components analysis", *Am. J. Hum. Genet.*, vol. 75, pp. 850-861, 2004.

[9] Z.H. Qi, J.M. Wang, and X.Q. Qi, "Classification analysis of dual nucleotides using dimension reduction", *J. Theor. Biol.*, vol. 260, pp. 104-9, 2009.

[10] M.A. Van de Wiel, F. Picard, W.N. van Wieringen, and B. Ylstra, "Preprocessing and downstream analysis of microarray DNA copy number profiles", *Brief. Bioinform.*, vol. 12, pp. 10-21, 2011.

[11] Q.S. Du, Z.Q. Jiang, W.Z. He, D.P. Li, and K.C. Chou KC, "Amino Acid Principal Component Analysis (AAPCA) and its applications in protein structural class prediction", *J. Biomol. Struct. Anal.*, vol. 23, pp. 635-40, 2006.

[12] Z.C. Li, X.B. Zhou, Z. Dai, and X.Y. Zou, "Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis", *Amino Acids*, vol. 37, pp. 415-25, 2009.

[13] X.W. Zhao, Z.Q. Ma, and M.H. Yin, "Predicting protein-protein interactions by combing various sequence- derived features into the general form of Chou's Pseudo amino acid composition", *Protein Pept. Lett.*, vol. 19, pp. 492-500, 2012.

[14] K.C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition", *Proteins*, vol. 43, pp. 246-55, 2001.

[15] X.L. Liu, J.L. Lu, and X.H. Hu, "Predicting thermophilic proteins with pseudo amino acid composition:approached from chaos game representation and principal component analysis", *Protein Pept. Lett.*, vol. 18, pp. 1244-50, 2011.

[16] C. Chatfield and A.J. Collins, *Introduction to Multivariate Analysis*, New York: Chapman and Hall, 1980.

[17] H.F. Kaiser, "The application of electronic computers to factor analysis", *Educ. Psychol. Meas.*, vol. 20, pp. 141-151, 1960.

[18] R.B. Cattell, "The scree test for the number of factors", *Multivariate Behav. Res.*, vol. 1, pp. 629-637, 1966.

[19] J.L. Mahloch, "Multivariate Techniques for Water Quality Analysis", *J. Environ. Eng. Div.*, vol. 100, pp. 1119-1132, 1974.

[20] T. Madsen, "Multivariate data analysis with PCA, CA and MS", 2007. Available at: *http://www.archaeoinfo.dk/PDF%20files/Multivariate%20data%20analysis.pdf* [Accessed on 25th June 2010]

[21] M. Maathuis, "Principal component Analysis (PCA)", 2008. Available at: *http://stat.ethz.ch/~maathuis/teaching/fall08/Notes3.pdf* [Accessed 15th August 2012]

[22] K.J. Parsons, W.J. Cooper, and R.C. Albertson, "Limits of principal components analysis for producing a common trait space: implications for inferring selection, contingency, and chance in evolution", *PLoS One*, vol. 4, pp. 7957, 2009.

[23] B.D. Horne and N.J. Camp, "Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation", *Genet. Epidemiol.*, vol. 26, pp. 11-21, 2004.

[24] T. LaFramboise, "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances", *Nucleic Acids Res.*, vol. 37, pp. 4181-4193, 2009.

[25] D.E. Reich, M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, T. Lavery, R. Kouyoumjian, S.F. Farhadian, R. Ward, and E.S. Lander ES, "Linkage disequilibrium in the human genome", *Nature*, vol. 411, pp. 199-204, 2001.

[26] C. Sabatti and N. Risch, "Homozygosity and linkage disequilibrium", *Genetics*, vol. 160, pp. 1707-1719, 2002.

[27] L. Kruglyak and D.A. Nickerson, "Variation is the spice of life", *Nat. Genet.*, vol. 27, pp. 234-236, 2001.

[28] P. Paschou, E. Ziv, E.G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M.W. Mahoney, and P. Drineas, "PCA-correlated SNPs for structure identification in worldwide human populations", *PLoS Genet.*, vol. 3, pp. 1672-86, 2007.

[29] K.D. Pruitt, T. Tatusova, and D.R. Maglott, "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins", *Nucleic. Acids. Res.*, vol. 1, D501-4, 2005.

[30] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)", *Nucleic. Acid. Res.*, vol. 40, D71-75, 2012.

[31] H.F. Kaiser HF, "The varimax criterion for analytic rotation in factor-analysis", *Psychometrika*, vol. 23, pp. 187-200 1958.

[32] R Development Core Team, *R: a language and environment for statistical computing*, Vienna: R Foundation for Statistical Computing, 2008.

[33] D.W. Huang, B.T. Sherman, and R.A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists", *Nucleic Acids Res*., vol. 37, pp. 1-13, 2009.

[34] D.W. Huang, B.T. Sherman, and R.A. Lempicki RA, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources", *Nat. Protoc.*, vol. 4, pp. 44-57, 2009.

Received: August 30, 2013 Revised: October 19, 2013 Accepted: October 29, 2013