Available online at www.medicinescience.org

ORIGINAL RESEARCH

Medicine Science International Medical Journal

Medicine Science 2018;7(4):852-6

Bioinformatics; The comparison of softwares based on genetics data analysis

Tahsin Ertas[1], Celal Guven[2], Onur Ozturk[3]

[1]*Istanbul University School of Medicine, Department of Biophysics, Istanbul, Turkey*
[2]*Nigde Omer Halisdemir University School of Medicine, Department of Biophysics, Nigde, Turkey*
[3]*Inonu University School of Medicine, Department of Biophysics, Malatya, Turkey*

Received 02 June 2018; Accepted 16 June 2018
Available online 11.10.2018 with doi:10.5455/medscience.2018.07.8903

Abstract

The comparative analysis with relational statistics programs Arlequin 3.5 and Power Marker 3.25 is performed, using the published polymorphic loci on the β-globin gene to investigate possible associations of these loci with the Hb D-Los Angeles mutation. It is envisaged that the results obtained by the processing of genetic data by different software will provide source data in terms of reliability of the analysis of the genetic data, compatibility of the programs, discrepancies, if any, and reasons for these differences and the researchers guide the program selection and benchmarking for the study purposes. Under this point of view; the main purpose of this study is to compare the possible differences or common results of analysis of gene data on beta globin gene family and beta globin gene in Hb D-Los Angeles [β121 (GH4) Glu→Gln] model with two statistical software. Arlequin and Power Marker software calculated the haplotype frequencies associated with the Hb D-Los Angeles mutation in both populations with equal frequency values. Considering the molecular diversity and mismatch distribution parameters, Arlequin software can provide important advantages in determining the historical development processes of populations. For each locus, parameters such as allele frequency calculations, allele pair frequency calculations, haplotype tendency regression, and difference test for both populations are presented as specific tests of Power Marker software. Findings from two different bioinformatics analyze and software advantages and disadvantages compared to each other are present.

Keywords: Bioinformatics, genetic data analysis, arlequin, power marker, population genetics

Introduction

At the point reached by modern technology today, researches on hereditary clinical problems or those that do not show any clinical symptoms but can be transferred by heredity have significant contributions to the development of the literature databases [1]. The rapid increase in data generated from studies involving genes and proteins is hopeful in understanding the diseases that cause health problems and in developing possible treatment methods. Naturally, the rapid increase in interest in genetic research and the generation of complex data clusters cause genetic technology and computer technology to find a common working area. For this purpose genetic database software is very useful in the basic issues such as storage, analysis and sharing of produced data. In this context, it is assumed that within the molecular biology there must be a basic research theme, such as collecting, processing, storing and using information in the living system model [2]. For many years scientists have been trying to get the complete genome sequence of many organisms [3]. The greatest of these has been the 'human genome project, which was launched in the 1980s and officially started in October 1990. Bioinformatics, which enables the co-operation of computer technology and medical science, enables the storage, analysis, processing, and use of results of DNA, RNA and protein data obtained from this major project and other genomic projects. The main aim of the bioinformatics science is; to collect, manage and distribute the rapidly growing and increasing knowledge, to ensure that the information is reached in the fastest and easiest way, to try to define the works of biological systems which are very complex [4]. The evaluation of these data may produce valuable information to the anthropological, paleoclimatic, archaeological and phylogeographical approaches to the process of biological development of the daily human being from the past [5]. Under this point of view; the main purpose of this study is to compare the possible differences or common results of analysis of gene data on beta globin gene family and beta globin gene in Hb D-Los Angeles model with two statistical software (Arlequin ver. 3.5 and Power Marker ver 3.25) [6-8].

***Coresponding Author:** Onur Ozturk, Inonu University School of Medicine, Department of Biophysics, Malatya, Turkey
E-mail: onurphysics@gmail.com

Materials and Methods

Sample definition

We analyzed 40 patients with Hb D-Los Angeles and 59 normal DNA data. These DNA samples that we have analyzed have been obtained from previous studies and have been provided using published data in the literature [9-12].

Statistical analysis

Firstly, we performed the data work of both populations with Arlequin 3.5 software, which uses an "unknown gametic phase method" and haplotype analysis [6,13], Hardy–Weinberg equilibrium tests [14,15] perform of genetic variation and group differentiation parameters, analysis of molecular variance (AMOVA) using F-statistics (FST, FIT, FIS) [16-19], Linkage disequilibrium (LD), historical-demographic analyses (Tajima's Fu's tests) [20,21], mismatch distribution analysis, analyses of tau (τ) and initial theta, SSD, the Harpending's raggedness index (Hri) and p-values of SSD [22-28] as previously reported [9]. Historic demographic expansions were also investigated by the analyzed of frequency distributions of pairwise differences between sequences (mismatch distribution), which is based on three parameters: θ0, θ1 (θ number of individuals before and after) and τ (time since expansion expressed in the unit of mutational time). According to Rogers (1995), to re-express τ years must be divided by 2u (twice the mutation rate) and multiply by the length of a generation, say, 25 years. Unfortunately, the mutation rate is not known with great accuracy. The rate of human mitochondrial nucleotide divergence has been variously estimated at 2% and 4% per million years, but the confidence intervals around these estimates are unknown. The two estimates u to be 7.5 x 10-4 and 1.5 x 10-3, respectively [25,26].

In the course of our work, we analyzed two populations statistically using the tests included in the PowerMarker software [7,8]. PowerMarker is a statistical analysis program with a user-friendly interface, including other genetic analyzes supported by statistical models, which can perform SNP analyzes primarily. Genetic markers are highly advantageous in determining historical parameters such as past genetic research, allele links, gene diversity, population-migration relationship, linkage disequilibrium (LD) [7]. PowerMarker interface allows more than 50 statistical analyzes to be easily viewed and analysis data edited. These analysis methods include Hardy-Weinberg and linkage test, population genetics parameter tests, $\chi 2$ tests, likelihood ratio tests and exact tests. LD, D and r2, a commonly used analysis group, can be performed by PowerMarker. In addition, the obtained LD results can be displayed in graphical and matrix form with its 2D viewer. Traditionally, genetic relationships and differences between comparable groups are calculated using F-statistics. In the PowerMarker software, these relationships are tested using four different F-statistic analyzes [7,8].

Results

Table 1 and Table 2 show the haplotype diversity and related frequencies obtained using the two software Arlequin 3.5 and PowerMarker. These tables summarize the top five highest frequencies.

Table 1. β-globin gene cluster haplotypes for the seven loci in association with the Hb D-Los Angeles population calculated with Power Marker and Arlequin

No.	Hb D-Los Angeles population Haplotype diversity	Power Marker % frequency	Arlequin % frequency
1.	+−−−−++	0.34694	0.34690
2.	−++−+++	0.29860	0.29863
3.	+−−−−−+	0.12496	0.12500
4.	+−−−−+−	0.06559	0.06559
5.	−+−−+++	0.02500	0.02500

Maximum-likelihood haplotype frequencies generated by Arlequin 3.5 software. Sum of the 14 listed frequencies and generated by Power Marker software. Sum of the 25 listed frequencies

Table 2. β-globin gene cluster haplotypes for the seven loci in association with the Normal population calculated with Power Marker and Arlequin

No.	Hb D-Los Angeles population Haplotype diversity	Power Marker % frequency	Arlequin % frequency
1.	+−−−−++	0.24751	0.26665
2.	−+−++++	0.16464	0.14668
3.	+−−−−+−	0.12952	0.12587
4.	−++−+++	0.07326	0.06264
5.	+−−−−−+	0.08545	0.05145
6.	+−−−−−−	0.01111	0.05438

Maximum-likelihood haplotype frequencies generated by Arlequin 3.5 software. Sum of the 14 listed frequencies and generated by Power Marker software. Sum of the 25 listed frequencies

We tested the genetic differentiation of both populations using the analysis of molecular variance (AMOVA) [28] implemented in Arlequin 3.5 and PowerMarker software [7] (Table 3).

Between population diversity indices, the haplotype diversity (h), nucleotide diversity (π) and average number of pairwise nucleotide differences (k) with their standard deviations computed with Arlequin and are showed at Table 4.

Table 3. (AMOVA) F-statistics (FST) calculated with Power Marker and Arlequin

AMOVA	Power Marker %	Arlequin %
Genetics differentiation of among populations	4.37	4.27

Exact test of sample differentiation based on haplotype frequencies Global test of differentiation among populations. Analysis of molecular variance (AMOVA) performed using the program ARLEQUIN ver. 3.5 and Power Marker ver. 3.25.

Harpending's raggedness index (Hri) and P values of SSD computed with Arlequin and showed at Table 5 [23].

The linkage disequilibrium (LD) P value is calculated by the chikare test and the permutation of its P value. In addition, the relation is shown in the table as indicated by the '+' and '-' signs (Table 6).

According to our comparative analysis results, HWE (P> 0.05) was shown in both groups calculated with Arlequin 3.5 and PowerMarker software (Table 7).

Table 4. Summary of molecular diversity performed with Arlequin

							Tajima's D		Fu's Fs		Mismatch distribution		
Populations	**n**	**No. of haplo.**	**k (95%CI)**	**θ_S**	**h**	**π**	**D**	**P**	**F_S**	**P**	**τ (95%CI)**	**θ_0**	**θ_1**
Normal	59	24	3.03 (2.19-4.55)	1.30 ±0.56	0.93 ±0.02	0.43±1.76	3.00	0.99	-16.88	0.00	3.46 (4.71-2.07)	0.01	25.82
Hb-D Los Angeles	40	14	2.56 (1.87-4.25)	1.41 ±0.62	0.87±0.02	0.36±1.54	1.97	0.95	-6.37	0.00	3.16 (5.08-1.58)	0.00	12.64

Number of individuals (n), number of haplotype, average pairwise differences among individuals (k), number of segregating sites (S), haplotype diversity (h ± standard deviation), nucleotide diversity (π ± standard deviation) for each populations. Tajima's D and Fu's Fs, corresponding P-value, and mismatch distribution parameter estimates for each population. D Tajima's D estimate population expansion, Fs Fu's Fs estimate population expansion. Values for τ, θ_0, and θ_1 are the age of the expansion, the population size before the expansion, and the population size after expansion, respectively, all expressed in units of mutation time. Insignificant P > 0.05, significant P ≤ 0.05. Tajima's D and Fu's Fs, corresponding P values, mismatch distribution parameter estimates and error estimates for populations are ±standard deviation as calculated by Arlequin.

Table 5. Values of the mismatch distribution test statistics SSD and rg performed with Arlequin

	Goodness-of-fit tests			
Populations	**SSD**	**SSD-P value**	**rg**	**P value**
Normal	0.00352	0.290	0.02279	0.540
Hb-D Los Angeles	0.00304	0.570	0.02137	0.800

P (SSD) is the probability of observing by chance a less than good fit between the observed and mismatch distribution for a demographic history of the population defined by the estimated parameters τ, θ0, and θ1. SSD: sum of squared deviations / rg: Harpending's raggedness

Table 6. Linkage disequilibrium (LD) calculated with Power Marker and Arlequin (significance level=0.05, P < 0.05; (+), P > 0.05; (−))

Hb D-Los Angeles population	**Power Marker**		**Arlequin**	
Pairs of loci	**Chi-square p value**	**Significant LD diagram**	**Chi-square p value**	**Significant LD diagram**
locus 1 - locus 2	0.0001	+	0.7548	−
locus 2 - locus 3	0.0001	+	0.0001	+
locus 3 - locus 4	0.5377	−	0.1751	−
locus 4 - locus 5	0.0029	+	0.0029	+
locus 5 - locus 6	0.0001	+	0.0009	+
locus 6 - locus 7	0.1222	−	0.7537	−
Normal population	**Power Marker**		**Arlequin**	
Pairs of loci	**Chi-square p value**	**Significant LD diagram**	**Chi-square p value**	**Significant LD diagram**
locus 1 - locus 2	0.0001	+	0.5931	−
locus 2 - locus 3	0.0001	+	0.0001	+
locus 3 - locus 4	0.0015	+	0.9113	−
locus 4 - locus 5	0.0001	+	0.0001	+
locus 5 - locus 6	0.1179	−	0.0004	+
locus 6 - locus 7	0.0207	+	0.0012	+

There is a listing of the log-likelihoods under the null and alternative hypotheses, a p-value determined by permutation, the $\chi 2$ test statistic and its corresponding (asymptotic) p-value. Table is provided, in which a '+' sign denotes nominal evidence of there are linkage disequilibrium (LD) between the alleles at two loci.

Table 7. Hardy-Weinberg equilibrium (HWE) test for all Loci calculated with Power Marker and Arlequin

No.	Normal population HWE P-value - Power Marker	Normal population HWE P-value - Arlequin
1.	0.4378	0.4352
2.	0.3052	0.4345
3.	0.6044	0.6050
4.	0.1303	0.2487
5.	0.2944	0.3038
6.	0.1059	0.1916
7.	0.3483	0.5478
Locus	**Hb D-Los Angeles population HWE P-value - Power Marker**	**Hb D-Los Angeles population HWE P-value - Arlequin**
1.	1.0000	1.0000
2.	1.0000	1.0000
3.	0.4860	0.4895
4.	1.0000	1.0000
5.	0.0940	0.1055
6.	0.0750	0.0820
7.	1.0000	1.0000

Tests for HWE for each locus within each population used an HWE test analogous to Fisher's exact test. P values were obtained using power marker 3.25 and arlequin 3.5 software. Insignificant $P > 0.05$, significant $P \leq 0.05$.

Discussion

Our study aimed to analyze the frequency of alleles determined by haplotype analysis of populations and haplotype types using Arlequin and PowerMarker software and to compare the results. The data obtained for both normal and Hb D-Los Angeles populations is in Hardy–Weinberg equilibrium ($P > 0.05$) for each of the seven polymorphic loci in both software as shown in Table 7. In both populations, the 6th locus p values are calculated to be close to the limit values in both Arlequin and PowerMarker software results (Table 7). The nearly equal p values shown in Table 7 indicate that both software offer reliable results for the HWE test. The PowerMarker software offers an advantage for researchers in comparing HWE results with multiple statistical tests. In both softwares haplotype I [+ – – – – + +] is the most frequent haplotype block (Table 1 and Table 2). Differently, the PowerMarker software provides results that include a large number of (25 types) haplotypes with lower percentages. Arlequin software lists 14 haplotype types. This can be considered as an advantage and suggests that it would be useful to determine the minimum frequency of genetic variability within the group. Arlequin software combines low-frequency and similar haplotypes. However, if the purpose of the research is to identify the low-frequency haplotypes in the population, this computation of Arlequin software may be considered to be disadvantageous.

Molecular diversity, historical gene flow parameters, Hri and SSD P values, population development and population age calculations provided by Arlequin can not be calculated by the PowerMarker software (Table 4 and Table 5). When mismatch distribution parameters are considered, Arlequin software can provide important advantages in determining the historical development processes of populations. Based on these data, we calculated the (AMOVA) fixation index (FST) to measure the degree of genetic differentiation between these populations with two softwares (Table 3). When the results of genetic differentiation of the two software are compared, it seems that there is a low difference. This low difference is caused by PowerMarker software calculation of a greater number of haplotypes.

Linkage disequilibrium (LD) means that the alleles in closest loci have similar frequencies in future generations. Given the genetic events, mutation and recombination may have quite pronounced effects on LD calculations, but other factors should not be ignored. Many of these factors affect the demographic characteristics of the population and cause to decrease the link between LD and loci [29]. According to the results of LD analysis, it was determined that there is a difference between the two software (Table 6). PowerMarker software only tests neighboring locus connections according to LD analysis method. However, Arlequin software performs LD analysis by comparing each locus, neighboring locus and non-neighboring locus. When evaluated in terms of LD test statistics, it appears that Arlequin software is more useful because it provides both the test parameters used and the LD results in an easy way to understand the format.

Conclusion

As a result, it is thought that findings presented in our study may provide valuable contributions for researchers to make appropriate software preferences in studies using bioinformatics and genetic data.

Acknowledgments
This study reflects the results of the MSc Thesis done by Tahsin ERTAŞ in İnönü University Graduate School in Health Sciences entitled as "Bioinformatics; The Comparison of Softwares Based on Genetics Data Analysis," June 2017.

Competing interests
The authors declare that they have no competing interest

Financial Disclosure
The financial support for this study was provided by the investigators themselves.

Author Contributions
TE, provided support in the data analysis of the laboratory results. OÖ is thesis director and supervised study, involved in the data interpretation and manuscript preparation. CG is support as co-director in MSc Thesis.

Reference

1. Giardine B, Borg J, Higgs DR, et al. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. Nat Genet. 2011;20;43:295–301.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, et al. GenBank. Nucleic Acids Res. 2002;1;30:17-20.
3. Kuonen D. Challanges in Bioinformatics for Statistical Data Miners, Bulletin of the Swiss Statistical Society. 2003;46:10-7.
4. Kumaresan V, Bhatt P, Palanisamy R, et al. Bioinformatics characterization, gene expression and proteolytic activity. Biologia. 2014;69:395-406.
5. Oppenheimer S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Philos Trans R Soc Lond B Biol Sci. 2012;367:770–84.
6. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online. 2007; 23;1:47-50.
7. Kejun Liu, Spencer V. Muse. PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics. 2005;21:2128–9.
8. Liu K. PowerMarker: New Genetic Data Analysis Software, Version 3.0. 2003; Free program distributed by the author over Internet at http://www.powermarker.net
9. Ozturk O, Arikan S, Atalay A, et al. Analysis of the population genetic structure of Hb D-Los Angeles [β121 (GH4) Glu→Gln GAA→CAA] in

Denizli, Turkey; genetic diversity, historical demography and estimation of the mutation rates based on haplotype variation. Am J Hum Biol. 2016;28:476-83.

10. Ozturk O. The beta globin gene cluster haplotypes associated with Hb D-Los Angeles in Denizli Province. M.Sc. Thesis, 41 p. Pamukkale University Graduate School of Health Sciences Denizli, Turkey, 2007.

11. Öztürk O, Atalay A, Köseler A, et al. Beta globin gene cluster haplotypes of abnormal hemoglobins observed in Turkey. Turk J Haematol. 2007;24:146-54.

12. Atalay EO, Atalay A, Ustel E, et al. Genetic origin of Hb D-Los Angeles according to beta globin gene cluster haplotypes. Hemoglobin. 2007;31:387–391.

13. Excoffier L and Heckel G. Computer programs for population genetics data analysis: a survival guide. Nat Rev Genet. 2006;7:745–58.

14. Excoffier L, Lischer H. E. L. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour. 2010;10:564–7.

15. Excoffier L, Laval G, Schneider S. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinform Online. 2005;1:47-50.

16. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967;27:209–20.

17. Schneider S, Roessli D, Excoffier L. Arlequin: A software for population genetics data analysis, version 2.000. Genetics and Biometry Laboratory, Department of Anthropology, University of Geneva, Switzerland. 2000; Retrieved from http://www.cmpg.uni be.ch/software/arlequin/archive/website/software/2.000/manual/Arlequin.pdf

18. Slatkin M. A measure of population subdivision based on microsatellite allele frequencies. Genetics. 1995;139:457–62.

19. Wright S. The interpretation of population structure by F-statistic with special regard to system of mating. Evolution. 1965;19:395–420.

20. Fu Y. Statistical tests of neutrality of mutations against population growth, hitchhiking and backgroud selection. Genetics. 1997;147:915-25.

21. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989a;123:585–95.

22. Excoffier L. Patterns of DNA sequence diversity and genetic structure after a range expansion: Lessons from the infinite-island model. Molr Ecol. 2004;13:853–864.

23. Harpending H. C. Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Human Biol. 1994;66:591–600.

24. Ray N, Curratand M, Excoffier L. Intra-deme molecular diveraity in spatially expanding populations. Mol Biol Evol. 2003;20:76–86.

25. Rogers A.R. Genetic evidence for a Pleistocene population explosion. Evolution. 1995;49:608–615.

26. Rogers A R, Harpending H. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol. 1992;9:552–69.

27. Schneider S, Excoffier L. Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. Genetics. 1999;152:1079–108.

28. Excoffier L, Smouse P, Quattro J. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics. 1992;131:479–91.

29. Ardlie KG, Kruglyak L, Seielstad M. Patterns of Linkage disequilibrium in the human genome. Genetics. 2002;3:299–309.