

A Quantitative Approach in Identifying Natural Selection Signals on Biallelic Single Nucleotide Polymorphisms of BRCA1 Gene in Diverse Populations

Alam Ahmad Hidayat^{1,2*}, *Rudi Nirwantono*^{2,3}, *Joko Pebrianto Trinugroho*³ and *Bens Pardamean*^{3,4}

¹Mathematics Department, School of Computer Science, Bina Nusantara University, 11480 Jakarta, Indonesia

²Bioinformatics and Data Science Research Center, Bina Nusantara University, 11480 Jakarta, Indonesia

³Biotechnology Department, Faculty of Engineering, Bina Nusantara University, 11480 Jakarta, Indonesia

⁴BINUS Graduate Program-Master of Computer Science Program, Bina Nusantara University 11480, Jakarta, Indonesia

Abstract. Population-specific studies reveal that cancer-related mechanisms of BRCA1 gene mutations may vary by ethnicity. The wealth of public genomic data provides insight into the functional roles of BRCA1 in diverse populations. In this study, we performed population differentiation analysis on biallelic SNPs located in the BRCA1 region using variant-calling data from the 1000 Human Genome Project. First, we conducted an Analysis of Molecular Variance (AMOVA) in global populations to infer a differentiation of BRCA1 gene in three population-related hierarchical levels. An F_{ST} -based approach was also conducted for each defined locus in the gene. Moreover, the signals of the natural selection in BRCA1 gene were computed using integrated Haplotype Score (iHS) per locus. The results demonstrated that BRCA1 gene differentiation can be attributed to the continental difference, for example, the genetic difference between Asian and African superpopulations accounts for 25% of the total variance. Imposing the iHS computation, we found that only two East Asian populations that underwent a positive selection, in which only benign variants were observed. In addition, those putative variants are only found in the non-coding regions: intron and 3' UTR. Our study is expected to ignite research interest in cancer-related genes for underrepresented populations.

* Corresponding author: alam.ahmad@binus.ac.id

1 Introduction

BRCA1 is a well-studied gene that encodes for a protein that is responsible for tumour suppression and has an important role in DNA repair pathways [1]. Epidemiological studies agree that germline mutations in BRCA1 (along with BRCA2) predispose to a substantial increase in breast and ovarian cancer risks in most population groups [2]. Therefore, testing for the presence of targeted BRCA1 mutations is a part of the established guideline for breast cancer population screening in most Western countries. Despite being known to contribute to more than 50% probability of developing breast cancer risk in women, the penetrance of BRCA1 mutation shows prominent variation within individuals in different races and ethnicities, familial factors, environmental exposures, etc [3]. The prevalence of BRCA1 mutations and their phenotypic consequences are still largely unknown for richly diverse racial and ethnic groups (i.e., “underrepresented” populations), including most Asian populations for which only a few comprehensive observational studies are available [4]. Hence, population-specific studies in this context provide a pivotal step to inspect accurate clinical implications of BRCA1 mutations, especially for personalized medicine purposes. To account for unique genetic structures among and within populations for accurate genome-wide association studies, we need to explore how the genetic diversity of BRCA1 was shaped by evolution processes as well as natural selections from their respective environment.

The selection footprints in human genomes are expected to be population-specific. Multiple loci located in a certain genomic region are in linkage disequilibrium and segregate as a unit of inheritance that may form a haplotype signature in a population group [5,6]. Therefore, detecting the selective pressure on genes in a particular population during human evolution can be conducted using haplotype-based analyses. The most commonly proposed statistical approaches are performed by testing the presence of a selection in genomic regions against the neutral evolution (null hypothesis) [7]. One of the methods is based on the concept of Extended Haplotype Homozygosity (EHH) to identify regions that undergo the positive selection via selective sweep (i.e., a beneficial allele experiences a “genetic hitchhiking” and reduces the overall genetic variation in a population) [8]. The key idea of EHH follows a method described by Klassman and Gautier [7,9]: If we consider a particular site “focal marker” in a chromosome and the variant at the marker is denoted as “core allele” (an SNP in our case), we then determine the EHH as a region in our sampled sequences that share the core allele (i.e., homozygous in the focal marker). The determined region is measured from both the upstream and downstream directions from the focal marker with a stopping criterion. From that, the integrated statistics were computed for that marker. Three widely used summary statistics in this framework: integrated Haplotype Score (iHS) for a single population, meanwhile XP-EHH, and Rsb are quantities used for comparing an indicative selection between two populations [7].

The EHH-based method has found many applications in predicting the selection signature in human genes. Barreiro et al. utilized the approach in inferring natural selection signals on complete SNPs in the human genome at the time from the Phase II HapMap data [10]. Complementing the F-statistic results, the iHS analysis was used in their work to infer disease-causing genes suspected to undergo a strong selection. The EHH approach was also used by Zhang et al. to inspect the genetic differentiation and natural selection of Zinc Transporter Genes (ZTGs) in humans for various worldwide populations [11]. Their iHS analysis identified some ZTGs with contrasting haplotype properties between Asian and African populations. A study by Esoh et al. was able to identify selective sweep signatures within coding regions of a malaria-related gene in Cameroonian populations using the iHS method [12].

In this work, we focused on applying statistical methods to understand the impact of population differentiation in BRCA1 gene using only markers in the form of biallelic Single Nucleotide Polymorphisms (SNPs). Further, we exploited an EHH analysis to identify putative regions of BRCA1 in each population that are suspected to be under positive selection. Our study is expected to provide a preliminary step to study the evolution of the Human BRCA1 gene, which could also be insightful for genetic testing applications.

2 Materials and Methods

2.1 Data

2.1.1 Data Collection

In this work, we used publicly available high-quality phased genotype data from the recent 1000 Human Genomes (1kGP) with 3202 WGS samples (2,504 unrelated samples and 602 complete trios), according to the high-coverage result described by Byrska-Bishop et al. [13]. The VCF file was retrieved from the International Genome Sample Resource (IGSR) website. The data consists of human subjects from 26 populations (ethnic backgrounds) of five superpopulations (continents). Since the VCF data does not contain information regarding the populations (or hierarchical populations), a separate txt file derived from a BED file in 1000 Human Genomes phase 3 was supplied to provide, for each sample, the following fields: “Population”, and “Superpopulation” for the population stratification.

2.1.2 Data Processing

Using the combination of BCFtools and VCFtools, we only selected the variants in genomic regions corresponding to BRCA1 gene (chromosome 17:43044295-43170245) and filtered only biallelic SNPs by excluding indels and multiallelic SNPs. In total, we obtained 2433 SNPs. Further, a web-based Ensembl Variant Effect Predictor was utilized to determine the location of the variants (coding and non-coding regions, upstream regions, etc), the variant consequences (synonymous and non-synonymous), and the mutational variant scoring in our data [14]. The variant distribution is shown in Fig. 1. The tool predicted that only 80 protein-altering SNPs (synonymous and non-synonymous) exist in our collected BRCA1 biallelic variants.

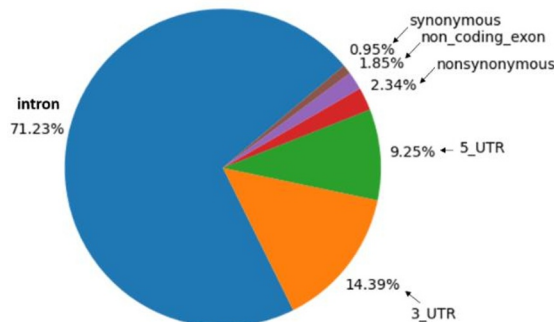


Fig. 1. The distribution of BRCA1 SNPs in terms of variant consequences.

2.2 Analysis of Molecular Variance (AMOVA)

AMOVA aims to infer population differentiations based on a set of genetic markers, taking into account the level of hierarchies in the populations. The method was developed by Excoffier et al. for haploid and diploid markers [15] and has been improved since then for analyzing polyploids [16]. In this work, we utilized package *poppr* to compute the AMOVA with *ade4* implementation in which Euclidean distance was used [17,18]. Technically speaking, we loaded the VCF data into R using package *vcfR* and converted the data into the *genind* object in package *adegenet* [19] using the command *vcfR2genind()*. The object stored information in the form of allele frequency per SNP, which is required for subsequent analysis steps. The population file was loaded into the *genind* object by supplying it into a 'slot' called *strata()* of the object. From that, the *genind* object is ready for the AMOVA implementation in *poppr*. The statistical significance between groups of populations was evaluated via p-value, which was computed based on the permutation test using *randtest()* command with which 1000 bootstrapping samples and a 95% confidence interval were used.

2.3 The Locus-specific F_{ST}

In this work, the degree of population differentiation in the BRCA1 gene was computed via locus-specific Fixation Index (F_{ST}). The F-statistic computation used throughout this work was based on the classical approach by Weir and Cockherham [20]. The F_{ST} value per locus (=per SNP) was computed using package *pegas* in R. In addition, we compared the F_{ST} with a Φ -statistics computation from AMOVA (i.e., hierarchical F_{ST}) in *hierfstat* in R [21] and found that the results were consistent. Further, the F_{ST} between all possible pairs of superpopulations were run using *hierfstat* with command *gene.dist()*. The significance test to obtain the confidence interval of each pairwise F_{ST} value was conducted with 1000 bootstrapping resamples and the confidence interval of 95% using *boot.ppfstat()* command in package *hierfstat*.

2.4 Selection Inference via integrated Haplotype Score (iHS)

The selection signature was identified via EHH analysis by performing iHS-based inference per SNP per population. iHS compared the EHH for an SNP for the ancestral allele with respect to the derived allele [7]. This was implemented in *rehh* package in R [9] by first determining the EHH on each variant with the command *scan_hh()* and then computing the standardized iHS scores for all variants using the command *ihh2ihs()*. The criteria of signal detection were to find a cluster of SNPs in a given window size whose $-\log(\text{p-value})$ values of their iHS score are less than 2. For the averaged computation, we employed a sliding window with a size of 10 kb and an overlapping region of 50%. This was implemented using *calc_candidate_regions()* in *rehh* to infer putative regions. The statistical significance of this test indicates that the distribution of iHS scores is different from the normal distribution which is always assumed to be followed by markers that undergo a neutral evolution.

3 Results

3.1 AMOVA Reveals Hierarchical Differentiations of BRCA1

Table 1. The AMOVA result for all superpopulations.

Comparisons	Variances (%)	Φ -statistic	p-value*
Among superpopulations	16.92	0.1811	0.001
Among populations	1.19	0.0143	0.001
Between Individuals within the population	81.90	0.1692	0.001

From the AMOVA result (Table 1) we found that the genetic diversity of BRCA1 gene was largely attributed to the variations within individuals, explaining 81.89% of the total variance in worldwide populations. All results were statistically significant (p-value < 0.05) according to the permutation test. Further, in specific analysis of the East Asian superpopulation, the same pattern is also replicated for the pairwise comparison between East Asian populations with other superpopulations that occupy more than 70% of the total explained variance in the AMOVA.

Table 2. The AMOVA between East Asians and other superpopulations.

Superpopulation	Variances (%)	Φ -statistic	p-value
EUR	0.68	0.0245	0.072
SAS*	5.89	0.0758	0.001
AMR	0.37	0.0373	0.221
AFR*	26.40	0.2808	0.001

From Table 2, a comparison made between EAS and EUR also seems to indicate the closeness of BRCA1 gene profile between the two ($\Phi < 0.03$). However, the permutation test also demonstrated that there is no evidence (p-value > 0.05) to reject the null hypothesis that the observed variations of BRCA1 found from the two superpopulations are identical. The same conclusion also holds for the comparison between EAS and AMR. Meanwhile, the differentiation between EAS and SAS was supported by the statistical test. Among the pairs of EAS and other superpopulations, the largest percentage of variance explained in the superpopulation level was found between EAS and AFR, with a percentage of 26.40%. Further, the permutation test also corroborated the significant differences (p-value < 0.05) of the Φ value between the two populations for the three hierarchical levels.

3.2 F_{ST} Calculation in Genetic Diversity of BRCA1

The mean F_{ST} value of 0.0170 was found for the whole SNPs and none of the SNPs exceeded an F_{ST} value more than 0.5. In fact, the maximum value of 0.4897 is brought by an intron variant with rsid rs8176287. The overall distribution of locus-specific F_{ST} was a right-skewed distribution. Prior to the analysis, we first removed variants with negative F_{ST} values since the “negative” interpretation is unclear. Next, imposing a 10% cutoff highest F_{ST} values

(i.e., $F_{ST} < 0.045$ in our case), SNPs were classified as “high F_{ST} ” variants or “low F_{ST} ” if otherwise. The distribution of variant consequences based on classified SNPs (according to F_{ST} values) is displayed in Table 3. We only found two highly differentiated missense variants in worldwide populations from our method: rs799917 and rs16941. These two mutations are considered “uncertain-significance to benign” according to ClinVar annotation.

Table 3. The distribution of variants based on F_{ST} values.

F_{ST} Classes	Consequences					
	3' UTR	5' UTR	Intron	Non coding exon	Non synonymous	Synonymous
High	36	20	173	6	2	2
Low	306	200	1511	39	55	21

It is noted that all pairwise F_{ST} values tabulated in Table 4 were found to be statistically significant (p-value < 0.05). In general, all pairwise comparisons involving African superpopulations yielded the highest F_{ST} values (>0.13), indicating that the BRCA1 gene in the African populations poses a strikingly unique genetic signature compared with the rest of the global population. This is certainly consistent with the result from AMOVA. On the other hand, achieving a drastically small pairwise value ($F_{ST} < 0.004$), the low genetic diversity of BRCA1 between European populations and the admixture American populations seems intriguing. In addition, the small extent of differentiation between East Asians and Europeans ($F_{ST} < 0.006$) should be noted as well.

Table 4. Pairwise F_{ST} between superpopulations.

Pairwise F_{ST}	EUR	EAS	AMR	SAS
EAS	0.00557			
AMR	0.00387	0.00607		
SAS	0.03957	0.03271	0.03297	
AFR	0.15383	0.15852	0.13980	0.18340

3.3 Selection Signals from the EHH Analyses

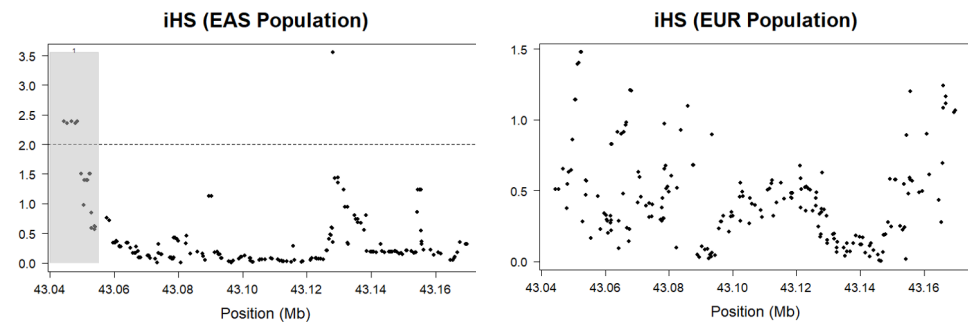


Fig. 2. A comparison between the iHS plots of EAS and EUR.

We identified the putative selective regions in BRCA1 gene using the EHH in the different hierarchical levels. First, we considered the iHS computation in the continent level (combining all ethnicities in a superpopulation). For each superpopulation, SNPs with MAF < 0.05 were removed and ~10-15% of SNPs passing the filter were used for the EHH scan analyses. The results found that only BRCA1 gene in the East Asian superpopulation (out of 5 superpopulations) displayed a positive selection signal (see Fig. 2). The putative selective region is estimated from 43.04 to 43.06 Mb in which 6 candidate SNPs show statistically significant integrated haplotype scores (p-value < 0.01). Considering their narrow genetic distance, we, therefore, concluded that the six loci in the putative region are within linkage with each other. In addition, in Fig. 2 we also provided an example of the plot for the European continent, which revealed that none of the loci in the gene exhibits a positive selection (p-value > 0.05). The European result also follows for other non-East Asian populations.

Table 5. The putative variants found from the iHS analysis in the East Asian populations.

SNP	rsid	Variant Consequence	ClinVar Interpretation
17:43044391:G:A	rs12516	3' UTR	Uncertain Significance to Benign
17:43045257:C:A	rs8176318	3' UTR	Uncertain Significance to Benign
17:43046604:A:G	rs8176314	Intron	Uncertain Significance to Benign
17:43047896:T:C	rs8176310	Intron	Uncertain Significance to Benign
17:43048092:T:C	rs4793190	Intron	Uncertain Significance to Benign
17:43048520:T:C	rs4792972	Intron	Uncertain Significance to Benign

The list of putative SNPs in East Asian populations from the analysis is shown in Table 5. According to the annotation result from Ensembl Variant Effect Prediction, two SNPs are located in a 3' UTR (cis-regulatory) region while four SNPs are in an intron region. Interestingly, these variants are interpreted in ClinVar database as “uncertain-significance to benign” in breast-ovarian cancer with a review status “reviewed by panel expert”.

Further, we also computed the iHS per SNP for 26 different populations. We showed only 5 out of 26 populations that demonstrated the positive signals based on the iHS analysis,

shown in Fig. 3, with the putative regions under selection highlighted in grey. Among the five populations, three populations are from the Asian continent: CDX (Chinese Dai in Xishuanagbanna), CHS (Southern Han Chinese), and KHV (Kinh in Ho Chi Minh City, Vietnam). Despite the potential finding, the signals observed in populations ACB, CDX, and CEU might be not considered “strong enough” compared with the other populations due to only a few numbers significant focal SNPs (only 3 SNPs) located in their putative regions. Meanwhile, populations CHS and KHV contain more consistent results with more than 3 putative SNPs observed in the same genomic region (43.040 Mb - 43.055 Mb). The results in CHS and KHV coincides with the result from the combined East Asian populations in Fig. 2.

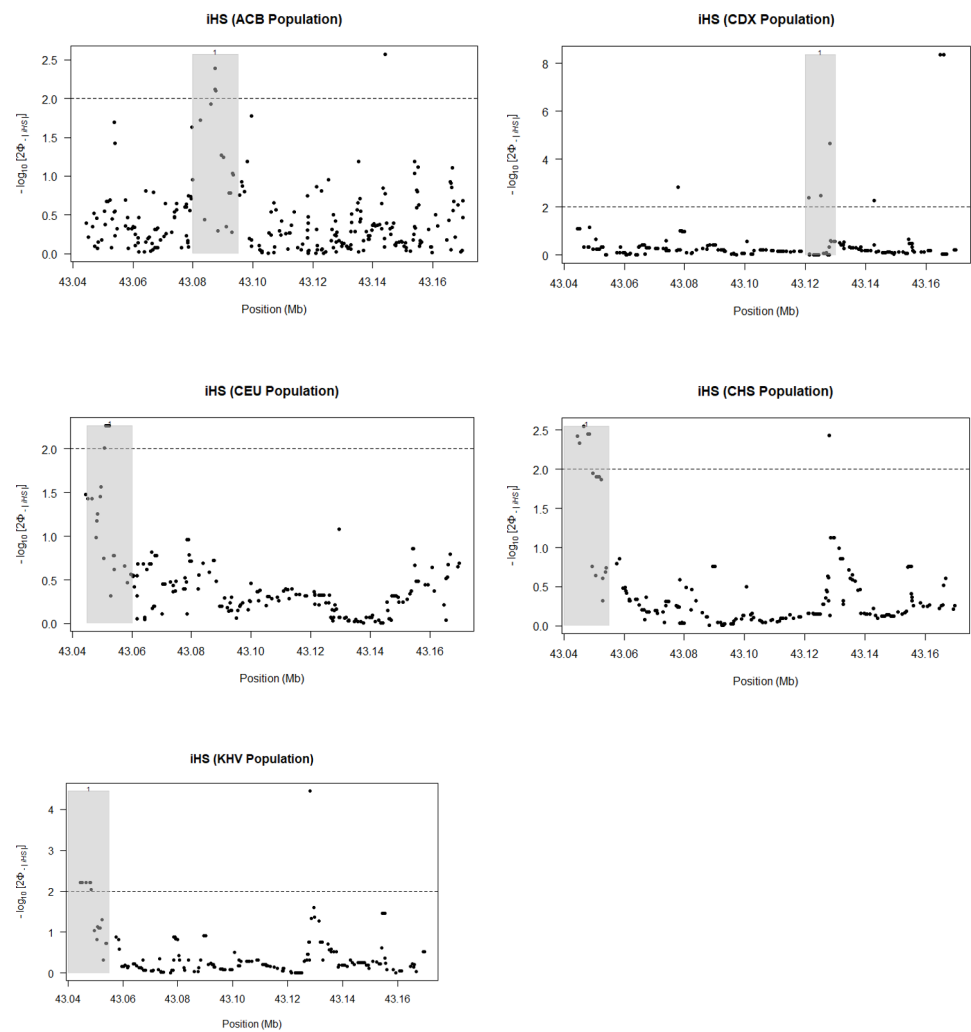


Fig. 3. The iHS plots for populations with statistically positive selection signals.

4 Discussion

The findings from AMOVA and pairwise F_{ST} agreed that the significant divergence of BRCA1 was found in the African population with the rest of the worldwide populations. The result, however, might not be surprising since the large geographical separation between them plays a primary role in the observed BRCA1 heterogeneity. Referring to the “Out of Africa” theory, a long-standing model posits a strong correlation between the genetic differentiation of human populations and the migratory/geographic distance of theirs from an origin in Africa, which was supported by some studies, including Ramachandran et al. [22].

Despite some potential results in our study, the selection signature of BRCA1 in most previous studies is considered weak compared with other disease-causing genes. For example, a genome-wide result using a haplotype test by Barreiro et al. showed that BRCA1 was not included to provide the strongest positive signal [10]. This may explain why the selection footprint on BRCA1 is only noticeable in 2 out of 26 populations in our study.

Interestingly, among the regional populations in the global population, we only found the BRCA1 gene from Kinh people and Southern Han Chinese that were predicted to experience a recent local selective sweep. Considering the linked geographical conditions and the demographic history between Southern China and Northern Vietnam, the identical signature between them might be not surprising. A previous study by He et al., using SNPs from Y chromosome, found that the chromosome lineages of Kinh people are more comparable to the Chinese haplogroup than the South Asian haplogroup [23]. This suggests the possible admixture of Chinese immigrants, who could be mainly from the Southern region, with Kinh people in the northern part of Vietnam. Our analysis may also suggest that an ancient or recent gene flow allows exchanges of disease-related mutations, including BRCA1 mutation carriers, between Southern China and Vietnam regions.

The result observed in the two populations indicates only benign mutations in the 3' UTR and intron regions that underwent selective pressure. Theoretically speaking, non-altering-protein variants with benign consequences (beneficial alleles) in a population are expected to have a larger probability of experiencing a fixation during the population's history. However, the selective signature on the SNPs based on our analysis was only detected in EAS. Hence, there was a local evolutionary force leaving unique footprints on the BRCA1 gene in Kinh and Southern Han Chinese populations, which should be investigated further. Arguably, studies involving the complex interplay of BRCA1 interacting with other genes and environmental factors as well as their phenotypic consequences in the populations must be taken into account to avoid any bias upon drawing the “positive selection” conclusion.

5 Conclusion

In our work, we have provided a preliminary investigation into the population differentiation and the selective sweep signal in BRCA1 gene using the established pipeline from AMOVA, F_{ST} analyses, and the EHH method. Unsurprisingly, African populations demonstrate high differentiation in BRCA1 relative to that of populations from other continents, supporting the “Out of Africa” theory. Meanwhile, the apparent selective pressures in the gene using the iHS approach were only discovered in Kinh and Southern Han Chinese populations. Further, only benign SNPs in non-coding regions were associated with this haplotype signature in the two populations. Our standard framework is expected to be useful for analyzing positive selection in cancer-related genes. This is particularly useful for underrepresented populations in which accurate and effective population screening guidelines for various cancers can be established.

References

1. D.P. Silver and D.M. Livingston, *J. AACR*. **2** (2012)
2. E. Levy-Lahad, A. Lahad and M.-C. King, *JNCI J. Natl. Cancer Inst.* **107** (2015)
3. A.W. Kurian, *Obstet. Gynecol.* **22** (2010)
4. A. Kwong, V.Y. Shin, J.C.W. Ho, E. Kang, S. Nakamura, S.-H. Teo et al., *J. Med. Genet.* **53** (2016)
5. S.S. Lloyd, E.J. Steele and R.L. Dawkins, *Analysis of Haplotype Sequences*, in *Next Generation Sequencing*, J.K. Kulski, ed., IntechOpen, Rijeka, (2016)
6. C.A. Ossa and D. Torres, *Oncologist*. **21** (2016)
7. A. Klassmann and M. Gautier, *PLoS One*. **17** (2022)
8. P.C. Sabeti, D.E. Reich, J.M. Higgins, H.Z.P. Levine, D.J. Richter, S.F. Nature. **419** (2002)
9. M. Gautier and R. Vitalis, *Bioinformatics*. **28** (2012)
10. L.B. Barreiro, G. Laval, H. Quach, E. Patin and L. Quintana-Murci, *Nat. Genet.* **40** (2008)
11. C. Zhang, J. Li, L. Tian, D. Lu, K. Yuan, Y. Yuan et al., *Sci. Rep.* **5** (2015).
12. K.K. Esoh, T.O. Apinjoh, S.G. Nyanjom, A. Wonkam, E.R. Chimusa, L. *Amenga-Etego et al.*, *Sci. Rep.* **11** (2021)
13. M. Byrska-Bishop, U.S. Evani, X. Zhao, A.O. Basile, H.J. Abel, A.A. Regier et al., *Cell*. **185** (2022)
14. W. McLaren, L. Gil, S. E. Hunt, H. Riat, *Genome Biology*. **122** (2016)
15. L. Excoffier, P.E. Smouse and J.M. Quattro, *Genetics*. **131** (1992)
16. P.G. Meirmans and S. Liu, *Front. Ecol. Evol.* **6** (2018)
17. Z.N. Kamvar, J.F. Tabima and N.J. Grünwald, *PeerJ*. **2** (2014)
18. S. Dray and A.-B. Dufour, *J. Stat. Softw.* **22** (2007)
19. T. Jombart, *Bioinformatics*. **24** (2008)
20. B.S. Weir and C.C. Cockerham, *Evolution (N. Y.)*. **38** (1984)
21. J. Goudet, *hierfstat*, *Mol. Ecol. Notes*. **5** (2005)
22. S. Ramachandran, O. Deshpande, C.C. Roseman, N.A. Rosenberg, M.W. Feldman and L.L. Cavalli-Sforza, *Proc. Natl. Acad. Sci.* **102** (2005)
23. J.-D. He, M.-S. Peng, H.H. Quang, K.P. Dang, A.V. Trieu, S.-F. Wu et al., *PLoS One*. **7** (2012)