Hindawi
BioMed Research International
Volume 2021, Article ID 6667201, 7 pages
https://doi.org/10.1155/2021/6667201

Research Article

Predicting Ovarian/Breast Cancer Pathogenic Risks of Human BRCA1 Gene Variants of Unknown Significance

Hui-Heng Lin,[1] **Hongyan Xu**,[2] **Hongbo Hu**,[2] **Zhanzhong Ma**,[3] **Jie Zhou**,[2] **and Qingyun Liang**[2]

[1]*Yuebei People's Hospital, Shantou University Medical College, Shaoguan 512025, China*
[2]*Department of Gynecology, Yuebei People's Hospital, Shantou University Medical College, Shaoguan 512025, China*
[3]*Clinical Laboratory, Yuebei People's Hospital, Shantou University Medical College, Shaoguan 512025, China*

Correspondence should be addressed to Hongyan Xu; medicalres@protonmail.com

Received 15 October 2020; Revised 17 March 2021; Accepted 7 April 2021; Published 15 April 2021

Academic Editor: Luis Morales-Quintana

High-throughput sequencing is gaining popularity in clinical diagnoses, but more and more novel gene variants with unknown clinical significance are being found, giving difficulties to interpretations of people's genetic data, precise disease diagnoses, and the making of therapeutic strategies and decisions. In order to solve these issues, it is of critical importance to figure out ways to analyze and interpret such variants. In this work, BRCA1 gene variants with unknown clinical significance were identified from clinical sequencing data, and then, we developed machine learning models so as to predict the pathogenicity for variants with unknown clinical significance. Through performance benchmarking, we found that the optimized random forest model scored 0.85 in area under receiver operating characteristic curve, which outperformed other models. Finally, we applied the best random forest model to predict the pathogenicity of 6321 BRCA1 variants from both sequencing data and ClinVar database. As a result, we obtained the predictive pathogenic risks of BRCA1 variants of unknown significance.

1. Introduction

For diagnosis of ovarian cancer, the symptom-based diagnostic approaches tend to be less precise because they usually do not display obvious and specific symptoms in early-stage patients. Therefore, unfortunately, the usual cases are that, when confirmed, the cancer is already developed to a late stage. The difficulties in detecting specific symptoms in early-stage ovarian cancer have affected precise diagnosis, and it is one of the important reasons causing the high mortality rate of ovarian cancer [1–3].

Since ovarian cancer is a multigenic disease, molecular genetic diagnosis is better for ovarian cancer diagnosis, compared with symptom-based methods, especially in cases of early-stage cancer. According to investigations, several genes are associated with the pathogenesis of ovarian cancer, and amongst them, two genes—BRCA1 and BRCA2—are well-known ones and found to have significant associations with ovarian cancer [4]. "BRCA1" stands for "BReast CAncer type 1 protein." It is a tumor suppressor and found to be associated with familial breast cancer [5]. Since the discovery of BRCA1, scientists have been researching the molecular structure and functions of it and its products [6–9]. Thanks to their efforts, parts of its functions and roles in biological processes have been elucidated. For instance, BRCA1 is known to participate in the processes of DNA repairing [10, 11], and its mutations/variants are known to have association with the onset of breast cancer and ovarian cancer [12–14]. Moreover, germline mutations of BRCA1 have also been discovered, and it is reported that, for patients of familial ovarian cancer, over 80% of patients carry BRCA1 (or BRCA2) mutation [4, 15, 16]. Because of the high susceptibility of BRCA1 to ovarian cancer, for molecular and genetic tests of ovarian cancer, BRCA1 is one of the most indispensable genes for probing.

Nowadays, modern sequencing technologies are not only applied to researches but also clinical diagnoses. Sequencing detects genotypes via reading nucleotide sequences in the human body, while challenges exist. A significant one is the gene Variants of Unknown Significance (VUSs). Sequencing keeps generating large sets of novel gene variants. However, except for their physical/structural variation information, such as the mutation position, and nucleic acids' changes, nothing else is known. This gives great difficulties in data interpreting and clinical diagnoses. For example, real cases happened that novel VUSs of BRCA1 were detected in people who seem to be healthy (at least no symptom was detected), while doctors still had no way to analyze the genetic pathogenic risks for these people. Therefore, it is urgent and necessary to find ways to analyze and interpret VUSs of genes.

Theoretically, biomedical characteristics and molecular functions of a VUS can be explored via biochemical approaches; e.g., immunoprecipitation and immunoblotting can probe the potential binding partner molecules with the protein of VUSs. Crystallization and 3D structural analyses could reveal the structure-function relationships of the protein of VUSs. While these methods are highly costly in all sorts of aspects including time, labor forces, and budget, in fact, large amounts of such VUSs are seen in clinical data and doctors usually face more than one VUS. Thus, wet lab methods are impractical for the purpose of quickly and efficiently characterizing large amounts of VUSs. The more realistic and practical way is computational analysis.

In the present study, BRCA1 VUSs were identified from clinical sequencing data. (Note that in this study, we also consider "likely pathogenic" and "likely benign" variants as the VUSs, since these variants lack solid evidence that can demonstrate their pathogenic risks.) In order to interpret the clinical significances of these data, we analyzed the data and developed multiple machine learning models to predict the pathogenic risks of these VUSs. After benchmarking, the optimized random forest model was found to have the best performance and it was chosen to predict BRCA1 VUSs from both our sequencing data and ClinVar database [17]. As a result, predictive pathogenicity of total 6322 VUSs was obtained. Amongst them, 1593 VUSs were predicted to be pathogenic and 4729 VUSs were predicted to be benign.

2. Materials and Methods

Initially, we identified and processed BRCA1 variant data from RNA-seq. Since the year 2017, we have started to ask hospital visitors if they are willing to donate their samples for our cancer genetic research purposes. For those who agreed, we sampled their bloods and made a third-party contractor research organization (Beijing Genome Institute, Shenzhen, China) to perform RNA-seq to these samples. After bioinformatic analyses, we further confirmed, analyzed, and identified the BRCA1 variant data via queries against databases (Ensembl [18], dbSNP [19], and ClinVar [17], as of 15 April 2020), information parsing, and data cleaning. Accordingly, these variation data were parsed and converted into DNA sequences and they were prepared for loading into predictive model predicting risks of ovarian/breast cancer.

Next, we prepared the dataset for training and benchmarking predictive models. We extracted BRCA1's variant data from the ClinVar, dbSNP, and Ensembl databases following such criteria. (1) Only retrieve BRCA1 variant data that are labeled "benign" and "pathogenic" for ovarian/breast cancer. (2) Choose variants that only have single nucleotide base substitution. (3) Choose variants that are reviewed by expert panel. As a result, 499 pathogenic BRCA1 variants and 585 benign BRCA1 variants were obtained. These variant data were further transformed into DNA sequences, and subsequently, through different types of molecular descriptors, DNA sequences were converted into feature vectors. Totally, more than 100 kinds of feature combinations were tested. Lastly, we found that the combination of 117 features gave the best performance for machine learning. The adopted feature set included vectors of DNA 3-mer [20], genomic location of variants, and di- (tri-) nucleotide-based autocross covariance [21, 22].

In order to obtain good predictive results, we selected multiple predictors and benchmarked their performance. Through benchmarking, we chose the most well-performed model to predict the pathogenic risk of BRCA1 VUSs so as to obtain the better and more precise results.

After preparation of datasets, we initially chose 5 predictors, i.e., the naïve Bayes [23], the support vector machine [24], the random forest [25], the PolyPhen program [26], and the SIFT program [27]. Amongst them, the former 3 are classic machine learning models, and we trained, generated, and tested the models by ourselves. While the latter 2 were programs developed by other researchers, their performances were used as the references to our models. Primary performance tests (used 8 : 2 ratio to randomly split dataset into training set and testing set) on 5 predictors identified the strongly biased performance of the naïve Bayes classifier against our datasets; thus, naïve Bayes was excluded from further analyses.

A series of methods were used for machine learning model optimizations; e.g., we adopted the oversampling method for balancing the positive and negative training set. We applied 10-fold cross-validation strategy to models. We tried to select a feature set giving the best performance to models. We also tried to perform standardization, normalization, principal component analysis, etc., to our dataset. In particular, for random forest and support vector machine, model-specific optimizations were conducted. For the support vector machine, the cost coefficient, gamma, kernel, etc., were tuned. For random forest, specific parameter tunings including the number of trees and number of features were carried out.

Predictive performances of all predictors were visualized as ROC plot, and the values of AUC were calculated for quantitative comparison. Accordingly, we also computed optimized random forest's true-positive rate, true-negative rate, false-positive rate, false-negative rate, positive predictive values, accuracy, balanced accuracy, and F-measure, so as to examine its extra performance scores in different perspectives.

Upon identification of the machine learning model with the best performance in benchmarking, both sets of BRCA1 VUSs, including 6 VUSs identified from sequencing data and 6315 VUSs identified from ClinVar database, were loaded into the optimized random forest for prediction. The obtained predictive results were statistically analyzed and then compared with their original pathogenicity annotations in the database.

The aforementioned data processing and computational analytic tasks including data parsing, data cleaning, sequence manipulations, format conversions, statistical analyses, and numeric computations were done in R computing environment and Rstudio [28, 29]. Besides self-scripted analytic pipelines, other used R packages include BioMedR [30], Bioconductor [31], Biostrings [32], e1071 [33], ROCR [34], and RandomForest [25].

3. Results

3.1. BRCA1 VUS Data. Through sequencing data analyses and database queries, we totally identified 7 *BRCA1* VUSs that have substitution of single nucleotide base. The relevant data and information are shown in Table 1. Of the 7 *BRCA1* variants listed in Table 1, 4 of which have unknown clinical significance. Here, the term "unknown clinical significance" indicates that whether or not these VUSs will increase the risk of having relevant diseases remains unclear. For the rest of the other 3 *BRCA1* variants, they are annotated as "likely benign" in databases. This indicates that so far these variants lack sufficient or solid evidence supporting their associations with pathogenic risk levels, and these uncertainties give difficulties to diagnosis, as well. Hence, the word "likely" is used and it is necessary to further analyze these variants. Specifically, one of our 7 VUSs, the "c.1348A>T," could not be found in the results of database queries or searching engines. It is indicating that the *BRCA1* "c.1348A>T" variant is a new variant identified from our sequencing data. Regarding this novel *BRCA1* variant and its molecular functions, though nothing is known except its sequence variation information, through our computational analyses and machine learning prediction, its pathogenicity was characterized.

Table 1: Seven VUSs of *BRCA1* identified from sequencing data and databases.

ID	Variation	refSNP ID (a.k.a. rs number)	Clinical significance in database
1	c.1255G>C	rs876658873	Unknown
2	c.824G>A	rs397509327	Unknown
3	c.3448C>T	rs80357272	Unknown
4*	c.1348A>T	Not available	Unknown
5	c.2566T>C	rs80356892	Likely benign
6	c.3748G>A	rs28897686	Likely benign
7	c.571G>A	rs80357090	Likely benign

*This *BRCA1* variant was not found through querying databases.

3.2. Performances of Predictive Models. As described in Materials and Methods, initially, 5 predictors were chosen for the present study. Because of the obviously biased performance of naïve Bayes in primary tests [23], it was excluded from further analysis and only 4 other predictors were used. Upon benchmarking with datasets, each model's receiver operating characteristic (ROC) curve was plotted and the relevant value of area under the curve (AUC) was computed accordingly. Eventually, the performances of 4 included models are as shown in Figure 1.

The 4 tested models, i.e., the support vector machine [24], the random forest [25], the PolyPhen [26, 35], and the SIFT [27], have the AUC values of 0.74, 0.78, 0.74, and 0.78, respectively. For our own models, the random forest outperformed the support vector machine. For comparison between our own models and others' models, the support vector machine had similar overall performance with PolyPhen, and the random forest had similar performance with SIFT as well.

Since our own models had similar performances with the two reference models developed by other researchers, in order to obtain the better model, we subsequently tried to optimize our own support vector machine and random forest models, hoping to see improvements in our models' predictive performances. The optimization works are described in Materials and Methods, and the performances of optimized models are as seen in Figure 2.

For both optimized models, the AUC of ROC indicates that the optimized random forest model had better overall performance than the optimized support vector machine. While comparing with our own models developed before optimization works, the support vector machine seemed to have little improvement (merely very slight improvement from 0.74 to 0.75, only 0.01 difference in AUC value). And for the random forest model, the optimization helped the random forest model improve its AUC value from 0.78 to 0.853 (the best values). Specifically, in order to detect the robustness of the model's performance, we further carried out 10 times 10-fold cross-validation to test the optimized random forest model. As a result, the mean value and standard deviation of AUCs are 0.85 and 0.006, respectively. This result showed that the optimized random forest had stable performance. For the best model of random forest, in Table 2, we listed its other performance indicators based on confusion matrix computing. Overall, the optimized random forest model had good performance.

3.3. Prediction of Pathogenic Risks of BRCA1 VUSs. Upon identification of the model with the best predictive performance, i.e., the optimized random forest, it has been applied to predict the *BRCA1* VUS pathogenic risks for ovarian/breast cancer. We initially predicted the *BRCA1* VUSs identified from our sequencing dataset. The results of the machine learning predictive analysis are shown in Table 3. For the first 3 VUSs, our best model predicted the first VUS to be pathogenic while the other 2 are benign (Table 3). And for the rest 4 VUSs, according to our predictive analysis, their prefix word "likely" was removed from their original clinical significances. In other words, our model had given further

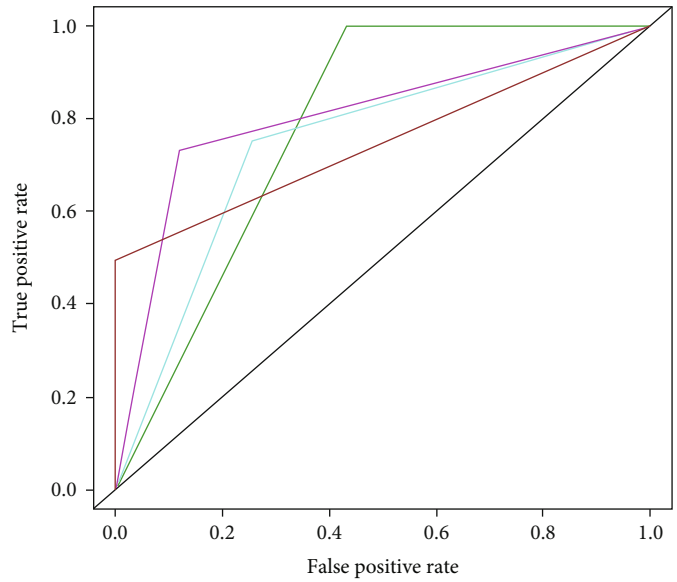

Figure 1: ROCs of 4 kinds of predictive models. ROCs indicating varied performances of different models were plotted. And the relevant AUCs were also computed to indicate models' overall performance. For support vector machine (the light blue curve), random forest (the purple curve), PolyPhen (the red curve), and SIFT (the green curve), their AUC values were 0.74, 0.78, 0. 74, and 0.78, respectively.

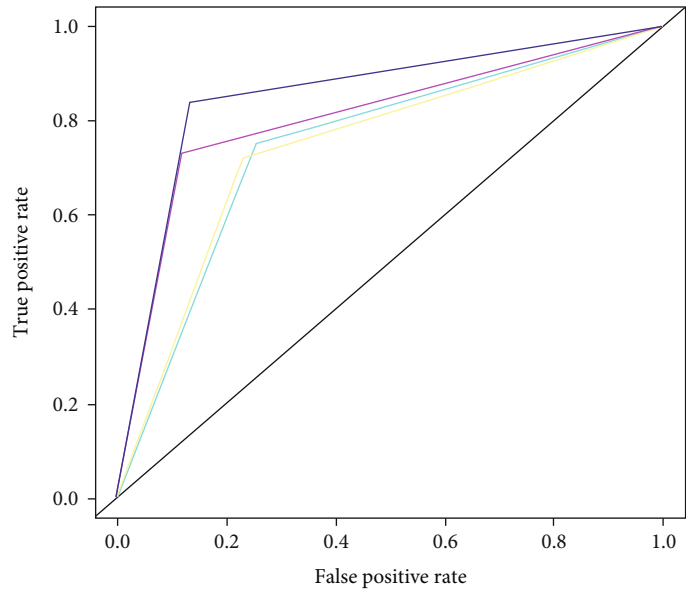

Figure 2: The overall performance of optimized support vector machine, optimized random forest model, original (not optimized) support vector machine, and original random forest. The optimized random forest (dark blue) had an obviously larger AUC than the optimized support vector machine (light blue) and the original random forest (purple), while no significant increase of AUC was observed between the original support vector machine (yellow) and the optimized one (light blue). The quantified AUC values of optimized random forest and optimized support vector are 0.85 and 0.75, respectively, indicating that the random forest model had better performance after optimization, while the support vector machine did not.

confidence to the likelihood of database's original pathogenic annotations of them.

Additionally, we found that large amounts of VUSs exist in the ClinVar database, too. Many gene variants were discovered and submitted to ClinVar. However, as mentioned before, parts of variants do not have any function annotation and hence they remain to be the VUSs. (Note that, here for variants in ClinVar, we consider those variants of "likely benign," "likely pathogenic," "not provided," and "conflicting interpretations of pathogenicity" the same as those variants

Table 2: Other performance indicators of the best random forest model. Tp, Tn, Fp, and Fn stand for the number of true-positive, true-negative, false-positive, and false-negative instance in the machine learning confusion matrix, respectively.

ID	Indicator	Value	Calculation
1	True-positive rate (a.k.a. sensitivity or recall)	0.84	Tp/(Tp + Fn)
2	True-negative rate (a.k.a. specificity)	0.86	Tn/(Tn + Fp)
3	False-positive rate	0.13	Fp/(Fp + Tn)
4	False-negative rate	0.16	Fn/(Fn + Tp)
5	Positive predictive value (a.k.a. precision)	0.77	Tp/(Tp + Fp)
6	Accuracy	0.85	(Tp + Tn)/(Tp + Tn + Fp + Fn)
7	Balanced accuracy	0.85	(True-positive rate + true-negative rate)/2
8	*F*-measure (a.k.a. *F*1 score)	0.80	2Tp/(2Tp + Fp + Fn)

Table 3: Predictive pathogenic risks for 7 VUSs of *BRCA1* identified from our sequencing data.

ID	Variation	(Original) clinical significance	Predictive pathogenic risk
1	c.1255G>C	Unknown	Pathogenic
2	c.824G>A	Unknown	Benign
3	c.3448C>T	Unknown	Benign
4[a]	c.1348A>T	Unknown	Pathogenic
5	c.2566T>C	Likely benign	Benign
6	c.3748G>A	Likely benign	Benign
7	c.571G>A	Likely benign	Benign

[a]Except this variant, the rest of the variants can be found in the ClinVar database.

of "uncertain significance," as the VUSs. And hence, all of them were the targets for our predictive analysis.) For example (as of 16 May 2020), for *BRCA1*, the number of variants of "likely benign," "likely pathogenic," "not provided," "conflicting interpretations of pathogenicity," and "uncertain significance" is 937, 82, 2797, 264, and 2235, respectively (Table 4). In order to provide more insights for these *BRCA1* VUSs which are not covered by our datasets, we also used the optimized random forest model to carry out pathogenic risk prediction of these *BRCA1* VUSs. Notice that 6 of VUSs identified from sequencing data were also found in ClinVar's VUSs. For a total of 6315 predicted variants, in which 6 VUSs identified from sequencing were excluded, 4724 (74.81%) were predicted to be benign and 1591 (25.19%) were predicted to be pathogenic (Table 4). Interestingly, we found that the percentages of predicted pathogenic variants in different subclasses are close. These percentages range from 69.51% to 77.80%. For variants of "likely pathogenic," "uncertain significance," and "conflicting interpretations of pathogenicity," the predicted pathogenic ratios of variants are 69.51%, 71.19%, and 71.97%, respectively. These values fluctuate around 70%, while the two percentages, 77.80% for "not provided" subgroup of *BRCA1* VUSs and 75.78% for "likely benign," are slightly higher than the former 3 percentages (Table 4). The full predictive results can be found in supplementary file S1.

4. Discussion

We believe that the present study of us is of high significance. Through analyzing potential pathogenicity of BRCA1 VUSs via machine learning approaches, we succeeded in carrying out such translational research that can help clinicians interpret the clinical significances of VUSs. Our work not only facilitates the precise clinical diagnosis but also provides references to clinical therapeutic decision-making.

Before we had the performance scores of the optimized machine learning models, we expected the AUC value of our best model could reach around 0.9. However, it did not despite our efforts and multiple trials, while we believe there exist ways for improvement of predictive models, such as using other advanced machine learning models or other optimization methods. For example, through feature engineering approaches, more advanced feature sets could be generated and tested.

In the present work, our analyses only covered single base substitutions of BRCA1 VUSs, while the forms of BRCA1 VUSs are diverse. Besides single base substitution mentioned in this work, there are deletion, insertion, and other kinds of structural variations as well. These variations and variants, though more complex than single nucleotide base substitution, are of equally high importance for medical researches, clinical data interpretation, and clinical diagnosis. In the future, we may take this challenge and try to analyze these more complex forms of BRCA1 VUSs, so as to acquire more insights between BRCA1 VUSs and cancers.

Variants' precise functional annotations and associated information are prerequisite for gene-disease relationship analysis and thus play a vital and indispensable role in precise diagnosis. On the one hand, we know that the precise functional annotations of gene variants come from the labor-, time-, and resource-consuming biochemical assays. On the other hand, more and more novel VUSs are being discovered. And the discovery of VUSs is in faster speed than the carry-out speed of biochemical experiments for VUS characterization. But for the purposes of clinical diagnosis, there is no doubt that it is necessary to efficiently interpret the VUSs. Hence, before databases can accumulate and disclose sufficient biochemical assay-based annotation data for gene variants, computational methods for variant analysis would still play important roles in the long run.

Table 4: Overview of predictive results for 6321 *BRCA1* VUS pathogenic risks.

Class	Clinical significance (ClinVar database)	Variant amount	Number of predictive pathogenic variants (%)	Number of predictive benign variants (%)
1	Likely benign	937	227 (24.22%)	710 (75.78%)
2	Likely pathogenic	82	25 (30.49%)	57 (69.51%)
3	Not provided	2797	621 (22.20%)	2176 (77.80%)
4	Uncertain significance	2235	644 (28.81%)	1591 (71.19%)
5	Conflicting interpretations of pathogenicity	264	74 (28.03%)	190 (71.97%)
6	Total	6315	1591 (25.19%)	4724 (74.81%)

5. Conclusions

In this work, we identified BRCA1 VUSs from sequencing data, and subsequently, we developed machine learning predictive models and benchmarked the performance of predictive models. The best predictive model scored an AUC value 0.85, namely, the optimized random forest, which was used to predict the pathogenic risk of BRCA1 VUSs, including those in the ClinVar database. In total, 6322 variants of unknown clinical significance were predicted. Amongst them, one variant "c.1348A>T" identified from the sequencing data has not been found in databases, and hence, we considered it as a novel VUS. And it was predicted to be pathogenic. For the other 6 VUSs identified from sequencing data, "c.1255G>C" was predicted to be pathogenic, as well, while the remaining 5 were predicted to be benign. For VUSs in ClinVar, our model predicted 4724 benign variants and 1591 pathogenic ones.

We believe that the present study of us is of high significance. Through analyzing potential pathogenicity of BRCA1 VUSs via machine learning approaches, we succeeded in carrying out such translational research that can help clinicians interpret the clinical significances of VUSs. Our work not only facilitates the precise clinical diagnosis but also provides references to clinical therapeutic decision-making.

Data Availability

The dataset of this study would be available upon reasonable request.

Disclosure

This study was submitted to medRxiv preprint previously. The URL is https://www.medrxiv.org/content/10.1101/2020.06.04.20120055v1.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Shaoguan Science and Technology Bureau (grant numbers 200812114531428 owned by Hui-Heng Lin and 2017cx/006 owned by Hongyan Xu) and Science and Technology Department of Guangdong Province (grant number 201803011 owned by Zhanzhong Ma).

Supplementary Materials

Supplementary Data S1 is available. The predictive results of pathogenic risks of BRCA1 VUSs. *(Supplementary Materials)*

References

[1] D. Chi, A. Berchuck, D. S. Dizon, and C. M. Yashar, *Principles and Practice of Gynecologic Oncology*, Lippincott Williams & Wilkins, Wolters Kluwer, Philadelphia, 2017.

[2] J. L. Bendedet, H. Bender, and H. Jones III, "FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers," *International Journal of Gynecology & Obstetrics*, vol. 20, no. 70, pp. 209–262, 2000.

[3] B. Fervers, J. Hardy, and T. Philip, "Standards, options and recommendations (clinical practice guidelines for cancer care from the French National Federation of Cancer (FNCLCC))," *british journal of cancer*, 2011.

[4] T. S. Frank, A. M. Deffenbaugh, J. E. Reid et al., "Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals," *Journal of Clinical Oncology*, vol. 20, no. 6, pp. 1480–1490, 2002.

[5] J. M. Hall, M. K. Lee, B. Newman et al., "Linkage of early-onset familial breast cancer to chromosome 17q21," *Science*, vol. 250, no. 4988, pp. 1684–1689, 1990.

[6] R. Baer, "With the ends in sight: images from the BRCA1 tumor suppressor," *Nature Structural Biology*, vol. 8, no. 10, pp. 822–824, 2001.

[7] W. S. Joo, P. D. Jeffrey, S. B. Cantor, M. S. Finnin, D. M. Livingston, and N. P. Pavletich, "Structure of the 53BP1 BRCT region bound to p53 and its comparison to the Brca1 BRCT structure," *Genes & Development*, vol. 16, no. 5, pp. 583–593, 2002.

[8] S. L. Clark, A. M. Rodriguez, R. R. Snyder, G. D. Hankins, and D. Boehning, "Structure-function of the tumor suppressor BRCA1," *Computational and structural biotechnology journal*, vol. 1, no. 1, article e201204005, 2012.

[9] J. M. Beckta, S. M. Dever, N. Gnawali et al., "Correction: mutation of the BRCA1 SQ-cluster results in aberrant mitosis, reduced homologous recombination, and a compensatory increase in non-homologous end joining," *Oncotarget*, vol. 7, no. 36, article 58716, 2016.

[10] T. T. Paull, D. Cortez, B. Bowers, S. J. Elledge, and M. Gellert, "Direct DNA binding by Brca1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 11, pp. 6086–6091, 2001.

[11] S. T. Durant and J. A. Nickoloff, "Good timing in the cell cycle for precise DNA repair by BRCA1," *Cell Cycle*, vol. 4, no. 9, pp. 1216–1222, 2005.

[12] C. A. Wilson, L. Ramos, M. R. Villaseñor et al., "Localization of human BRCA1 and its loss in high-grade, non-inherited breast carcinomas," *Nature Genetics*, vol. 21, no. 2, pp. 236–240, 1999.

[13] D. D. Bowtell, "The genesis and evolution of high-grade serous ovarian cancer," *Nature Reviews. Cancer*, vol. 10, no. 11, pp. 803–808, 2010.

[14] C. Sun, N. Li, Z. Yang et al., "miR-9 regulation of BRCA1 and ovarian cancer sensitivity to cisplatin and PARP inhibition," *Journal of the National Cancer Institute*, vol. 105, no. 22, pp. 1750–1758, 2013.

[15] B. Modan, P. Hartge, G. Hirsh-Yechezkel et al., "Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation," *The New England Journal of Medicine*, vol. 345, no. 4, pp. 235–240, 2001.

[16] S. A. Narod, P. Sun, P. Ghadirian et al., "Tubal ligation and risk of ovarian cancer in carriers of BRCA1 or BRCA2 mutations: a case-control study," *Lancet*, vol. 357, no. 9267, pp. 1467–1470, 2001.

[17] M. J. Landrum, J. M. Lee, M. Benson et al., "ClinVar: public archive of interpretations of clinically relevant variants," *Nucleic Acids Research*, vol. 44, no. D1, pp. D862–D868, 2016.

[18] A. D. Yates, P. Achuthan, W. Akanni et al., "Ensembl 2020," *Nucleic Acids Research*, vol. 48, no. D1, pp. D682–D688, 2020.

[19] S. T. Sherry, M. H. Ward, M. Kholodov et al., "dbSNP: the NCBI database of genetic variation," *Nucleic Acids Research*, vol. 29, no. 1, pp. 308–311, 2001.

[20] B. Chor, D. Horn, N. Goldman, Y. Levy, and T. Massingham, "Genomic DNA k-mer spectra: models and modalities," *Genome Biology*, vol. 10, no. 10, p. R108, 2009.

[21] B. Liu, Y. Liu, X. Jin, X. Wang, and B. Liu, "iRSpot-DACC: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance," *Scientific Reports*, vol. 6, no. 1, article 33483, 2016.

[22] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. C. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. W1, pp. W65–W71, 2015.

[23] D. D. Lewis, "Naive (Bayes) at forty: the independence assumption in information retrieval," in *Machine Learning: ECML-98*, vol. 1398, pp. 4–15, Springer, Berlin, Heidelberg, 1998.

[24] B. Schölkopf, P. Simard, A. J. Smola, and V. Vapnik, "Prior knowledge in support vector kernels," in *Advances in Neural Information Processing Systems*, pp. 640–646, Denver, Colorado, USA, 1997.

[25] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[26] W. McLaren, L. Gil, S. E. Hunt et al., "The Ensembl variant effect predictor," *Genome Biology*, vol. 17, no. 1, p. 122, 2016.

[27] P. C. Ng and S. Henikoff, "SIFT: predicting amino acid changes that affect protein function," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3812–3814, 2003.

[28] R Core Team, "R: a language and environment for statistical computing," 2013.

[29] Rstudio Team, "RStudio: integrated development for R," *R Studio*, vol. 42, p. 14, 2015.

[30] J. Dong, M. F. Zhu, Y. H. Yun, A. P. Lu, T. J. Hou, and D. S. Cao, "BioMedR: an R/CRAN package for integrated data analysis pipeline in biomedical study," *Briefings in bioinformatics*, vol. 22, no. 1, pp. 474–484, 2021.

[31] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[32] H. Pages, P. Aboyoun, R. Gentleman, S. DebRoy, and M. H. Pages, "Biostrings: efficient manipulation of biological strings," *R package version*, 2017.

[33] D. Meyer, E. Dimitriadou, K. Hornik et al., "Package 'e1071'," *The R Journal*, 2019.

[34] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R," *Bioinformatics*, vol. 21, no. 20, pp. 3940-3941, 2005.

[35] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting functional effect of human missense mutations using PolyPhen-2," *Current Protocols in Human Genetics*, vol. 76, no. 1, pp. 7–20, 2013.