Toh and Brody *BioData Mining* (2021) 14:18
https://doi.org/10.1186/s13040-021-00253-y

BioData Mining

RESEARCH Open Access

Genetic risk score for ovarian cancer based on chromosomal-scale length variation

Christopher Toh and James P. Brody*

* Correspondence: jpbrody@uci.edu
Department of Biomedical Engineering, University of California, Irvine, USA

Abstract

Introduction: Twin studies indicate that a substantial fraction of ovarian cancers should be predictable from genetic testing. Genetic risk scores can stratify women into different classes of risk. Higher risk women can be treated or screened for ovarian cancer, which should reduce ovarian cancer death rates. However, current ovarian cancer genetic risk scores do not work that well. We developed a genetic risk score based on variations in the length of chromosomes.

Methods: We evaluated this genetic risk score using data collected by The Cancer Genome Atlas. We synthesized a dataset of 414 women who had ovarian serous carcinoma and 4225 women who had no form of ovarian cancer. We characterized each woman by 22 numbers, representing the length of each chromosome in their germ line DNA. We used a gradient boosting machine to build a classifier that can predict whether a woman had been diagnosed with ovarian cancer.

Results: The genetic risk score based on chromosomal-scale length variation could stratify women such that the highest 20% had a 160x risk (95% confidence interval 50x-450x) compared to the lowest 20%. The genetic risk score we developed had an area under the curve of the receiver operating characteristic curve of 0.88 (95% confidence interval 0.86–0.91).

Conclusion: A genetic risk score based on chromosomal-scale length variation of germ line DNA provides an effective means of predicting whether or not a woman will develop ovarian cancer.

Keywords: Genetic risk score, Ovarian cancer, Polygenic risk score, Risk prediction, Copy number variation, TCGA the Cancer genome atlas

Introduction

Ovarian cancer kills about 150,000 women per year worldwide [1]. The most common form of ovarian cancer, ovarian serous carcinoma is often diagnosed late (stage III (51%) or IV (29%)) and has a relatively bleak 5-year survival rate [2]. If women with an elevated risk of developing ovarian cancers could be identified, interventions could be taken that would reduce the number of women who die from ovarian cancer. These interventions include prophylactic oophorectomies, which would completely avoid ovarian cancer, and more targeted screening, which could identify ovarian cancers in

earlier stages, where surgery is an effective cure [3–7]. These interventions could both increase 5-year survival times and reduce the overall number of deaths due to ovarian cancer.

A substantial fraction of ovarian cancers should be predictable by genetic testing. The heritability of ovarian cancer has been measured at about 40% (95% confidence interval 23–55%) by the Nordic Twin Study [8]. The maximum discriminative accuracy of a genetic risk test is a function of both the heritability and the prevalence of the disease [9, 10]. Based on the measured heritability (about 40%) and prevalence (about 0.1%) of ovarian cancer, the maximum accuracy, measured by the area under the receiver operating characteristic curve (AUC), should be greater than 0.95, where 1.0 indicates a perfect test. Current genetic risk scores do not approach that level of accuracy.

Most current genetic risk scores are derived from single nucleotide polymorphisms (SNPs) identified by genome wide association studies [11–16]. These tests, called polygenic risk scores, construct a score based on a linear combination of the value of a collection of SNPs. This strategy has been moderately successful with ovarian cancer. One study followed this strategy to construct a polygenic risk score where women who scored in the top 20% had a 3.4-fold increased risk compared to women who scored in the bottom 20% [17].

We developed an alternative strategy to compute genetic risk scores. Our strategy is based on structural variation rather than SNPs and uses machine learning algorithms, which include non-linear effects, rather than linear combinations.

Methods

We tested this strategy with data from the Cancer Genome Atlas (TCGA) project. TCGA was a project sponsored by the National Cancer Institute to characterize the molecular differences in 33 different human cancers [18–20]. The project collected samples from about 11,000 different patients, all of whom were being treated for one of 33 different types of tumors. The samples collected usually included tissue samples of the tumor, tissue samples of normal tissue adjacent to the tumor and normal blood samples. (Normal blood samples were not available from patients diagnosed with leukemias.)

Most of the patient normal blood samples were processed to extract and characterize germline DNA. All germline DNA samples were processed by a single laboratory, the Biospecimen Core Resource at Nationwide Children's Hospital. Single nucleotide polymorphisms (SNPs) were measured from the patient samples with an Affymetrix SNP 6.0 array. This SNP data was then processed (by the TCGA project) through a bioinformatics pipeline [21], which included the packages Birdsuite [22] and DNAcopy [23]. The result of this pipeline is, for each sample, a listing of a chromosomal region (characterized by the chromosome number, a starting location, and an ending location) and the associated value given as the "segmented mean value." The segmented mean value is defined as the logarithm, base 2 of one-half the copy number. A normal diploid region with two copies will have a segmented mean value of zero.

The Affymetrix SNP 6.0 array provides intensity measurements indicating whether or not specific probes on the array bind to specific sequences in the sample. These intensity measurements are usually interpreted in a binary fashion, indicating whether a

specific sequence is absent or present in the sample. This process provides the genotype for a sample, quantified by the presence or absence of single nucleotide polymorphisms (SNPs). If these intensity measurements are instead interpreted in an analog fashion, one can discern whether specific sequences are absent, present with a single copy, two copies, three copies, etc. Thus providing a relative copy number value at each SNP location. By collecting these values across the chromosome scale, we compute a number that we call the chromosome-scale length.

NCI has provided most of the TCGA data on the Genomic Data Commons [24]. The copy number variation data available is called the masked copy number variation on the Genomic Data Commons. The masking process removes "Y chromosome and probe sets that were previously indicated to have frequent germline copy-number variation." [21].

This research uses de-identified coded datasets produced by TCGA. Therefore it is not considered human subjects research.

We accessed the TCGA data through Google's BigQuery, a cloud-based database. This resource is hosted and maintained by the Institute of Systems Biology [25]. We used the copy number segment (masked) table extracted from the Genomic Data Commons in February 2017. We also used information from the Biospecimen (extracted April 2017) and Clinical (extracted June 2018) tables. The copy number table contained all the information for the chromosome scale length variation data. The Biospecimen table was used to identify which samples were from normal blood (representing germ line DNA). The Clinical table provided information on the individual patient's gender, race, and ovarian cancer status. Information in the different tables was tied together by the sample barcode parameter.

All patients in the TCGA ovarian cancer sample had a well characterized form of ovarian cancer. TCGA only included those who were newly diagnosed with ovarian serous adenocarcinoma. The tumor had was confirmed to be serous by a board-certified pathologist after examining histological samples of the tumor. Mucinous, endometrioid and other types of ovarian tumors were excluded.

The final dataset consisted of a dataset with 4639 rows, each representing a different patient. Each row started with a label, "ovarian cancer" or "normal", and then 22 numbers. The mean age at diagnosis of the patients with ovarian cancer was 59.7 years, while the mean age for the "normal" sample was 58.6 years. Each number represented a measure of the length for one of the chromosomes. These length measurements were reported by the TCGA bioinformatics pipeline as extremely long copy number variations, usually greater than 90% of the length of the chromosome. We obtained these numbers from the TCGA dataset stored on Google's BigQuery. The TCGA bioinformatics pipeline did not report any copy number values for many specific genomic regions, presumably that indicates the copy number value is normal, with two copies. However, we coded these as not available, or "N/A" in our dataset. This dataset was used for the machine learning analysis.

We used the statistical computer language R to query the BigQuery database, collect the data and manipulate it into different forms. We took extensive care to avoid typical problems that lead to falsely high AUCs in machine learning. For instance, we ensured that no data leakage occurred, which can lead to deceivingly high AUCs when copies of a sample appear in both the training and test sets.

We used the H2O machine learning package in R to create machine learning models. H2O takes care of setting many of the proper default values, depending on whether the goal of the model is classification or regression. For the gradient boosting machine (GBM) models, H2O performs preprocessing, randomization, encoding categorical variables, and other data processing steps appropriate for the chosen model.

H2O has an automated machine learning algorithm, named AutoML [26]. Given a spreadsheet like- dataset, AutoML will run through four different machine learning algorithms and evaluate which provides the best models for the given problem. For each of the machine learning algorithms, it will evaluate several different hyperparameters. The process is limited by the amount of time devoted to it. After the allotted time, AutoML reports a scoreboard ranking the best algorithms. For the gradient boosting machine algorithm, we started with the default H2O settings. These default settings build trees to a maximum depth of five trees with a sample rate of 1 [27]. For the results reported in Table 2, we used an allotted time of one hour. In tests, we found that the results do not change substantially with times up to 10 h.

We used 5-fold cross validation with the GBM algorithm to produce Table 3 and Fig. 2. Cross validation uses repeated model runs with non-overlapping data. This approach allows one to use of all samples in the limited dataset. For Table 3 and Fig. 2, we estimated 95% confidence intervals for the odds ratios following the method described in [28].

Figure 3 was produced with a single model run by splitting the dataset into a training set holding 80% of the data and a test set containing 20% of the data.

Code is available to reproduce this work at: https://github.com/jpbrody/cancer-prediction-cnv/blob/master/ovarian-TCGA.R

Results

Using the TCGA dataset, we identified a measure that we call *chromosome-scale length variation*. Taken together, structural variations like insertions, deletions, translocations and copy number variations slightly alter the overall length of an individual's chromosome. Thus, the lengths of the set of chromosomes can be used to characterize a person. A histogram showing the distribution of relative chromosome lengths taken from germ line DNA samples in the TCGA dataset is shown in Fig. 1. By convention, these lengths are reported in units of log base 2. A value of "0" represents the consensus, average, chromosome length.

From the TCGA dataset, we synthesized a case-control study to test whether chromosome-scale length variation data can construct a genetic risk score. We identified 4225 women who had not been diagnosed with any form of ovarian cancer and 414 women who had been diagnosed with ovarian serous carcinoma. Statistical descriptions of the two populations are shown in Table 1.

Next, we evaluated the effectiveness of several different machine learning algorithms. We measured how well these algorithms could classify a woman, based solely on the set of 23 chromosome-scale length variation measurements, into either the class with ovarian cancer or without. The measurement of success we used was the area under the curve (AUC) of the receiver operating characteristic curve. The results of these measurements are shown in Table 2.

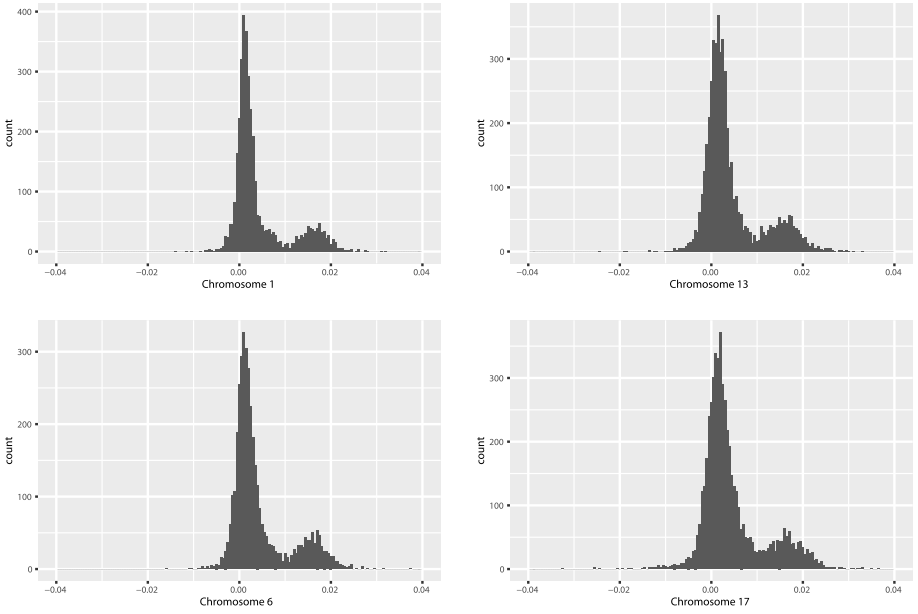

Fig. 1 This figure shows a histogram of chromosome scale length variation for most of chromosomes 1,6,13, and 17. For most patients in the TCGA dataset, a normal blood sample was taken, genomic DNA was extracted from that sample and analyzed with an Affymetrix SNP 6.0 array. The data from this array was processed by the TCGA project through a bioinformatic pipeline that resulted in a segment mean value, which is a number equal to the log base two of one half the copy number value. This histogram indicates that most people have a nominal value of 0, indicating exactly two copies of the diploid chromosome. A value of 0.02 would indicate the person has on average 2.028 copies of the chromosome, or about 1.4% longer than the average length of the chromosome

Based on the results in Table 2, we used the Gradient Boosting Machine algorithm throughout the rest of this manuscript. In the next step, we sought to classify the 4669 women in the dataset. We used a k-fold cross validation procedure, with $k = 5$. The dataset was randomly partitioned into five equal groups. The first group was held out (to be the test set), while the other four groups were used to train a model to distinguish the two classes (women with ovarian cancer and women without ovarian cancer). The trained model assigned a numerical score to each of the women in the first group (test set) quantifying how likely that woman was a member of the ovarian cancer class. The process was repeated 5 times, with a different group held out each time. The result is a numerical score for each of the 4669 women.

Table 1 From the TCGA dataset, we constructed two groups, both solely composed of women. The first group, containing 414 women, all had been diagnosed with ovarian serous carcinoma. None of the second group, with 4225 women, had been diagnosed with any form of ovarian cancer. This table compares the two populations

	Diagnosed with Ovarian Serous Carcinoma	Not diagnosed with Ovarian cancer
Total	414	4225
Mean age	58.3 years	59.7 years
% Black	25/414 = 6%	492/4225 = 12%
% White	352/414 = 85%	3064/4225 = 73%
% Asian	14/414 = 3%	259/4225 6%

Table 2 This table lists five different machine learning algorithms we evaluated for predicting ovarian cancer from chromosome-scale length variation data using the H2O package in R. The algorithms are ranked by the best AUC it achieved using 5-fold cross validation

Algorithm	AUC
Gradient Boosting Machine	0.88
Distributed Random Forest	0.87
Extremely Randomized Trees	0.86
Deep learning	0.82
Generalized Linear Model	0.68

The predictions were compared to the known ovarian cancer status of each of the 4669 women. First, all 4669 women were ranked by their score, representing the likelihood that they were from the ovarian cancer class. By comparing this ranking with their known ovarian cancer status, we can evaluate how well the model classified the women.

The comparison is presented in two different forms. Table 3 provides a tabular form of relative risk for the population segmented into five different groups. Figure 2 shows similar information in graphical form, where the population is segmented into 50 groups.

Finally, we took the dataset of 4669 women and split it into a training set (80%) and a test set (20%). Using H2O, we trained a Gradient Boosting Machine model to predict whether a woman was in the group with ovarian cancer, or not. The results are presented in Fig. 3, which shows a classic receiver operating characteristic curve of the model's predictions. Figure 4 presents the SHAP contribution plot, which helps explain how the Gradient Boosting Machine model arrives at its result.

Discussion

The results presented here compare favorably to other genetic risk scores for ovarian cancer. For instance, a previous study found that a polygenic risk score in the top 20% conferred a 3.4-fold risk increase compared to women in the bottom 20% [17]. As seen in Table 3, the top 20% in our results had an increase of over 100-fold risk over women who scored in the bottom 20%.

Table 3 Using 5-fold cross validation, each woman in the dataset received a score from the model built to predict ovarian cancer. The women were ranked by score from lowest to highest and then partitioned into five quintiles. This table presents the number of women with and without ovarian cancer in each quintile along with the odds ratio (relative to the entire group) and the 95% confidence interval for the odds ratio

Quintile	Number of women without ovarian cancer	Number of women with ovarian cancer	Total number of women	Odds ratio	95% confidence interval
1	925	3	928	0.03	0.01--0.09
2	925	3	928	0.03	0.01--0.09
3	901	27	928	0.30	0.21--0.45
4	842	86	928	1.04	0.82--1.33
5	632	295	927	4.76	4.01--5.65

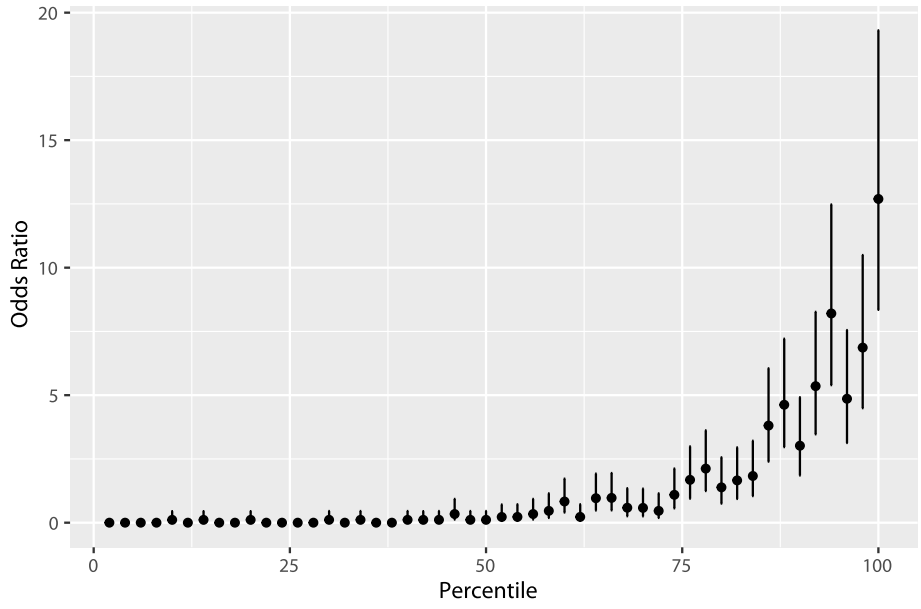

Fig. 2 This figure shows that women ranked higher by the predictive model are significantly more likely to have ovarian cancer. The predictive model ranked all 4669 women in the dataset based on their likelihood of having ovarian cancer, based solely on germ line DNA data. This ranking was then split into 50 equal partitions, each with about 93 women. This plot shows the odds ratio (relative to 414 ovarian cases out of 4669 total) of each of the 50 equal partitions along with the 95% confidence intervals

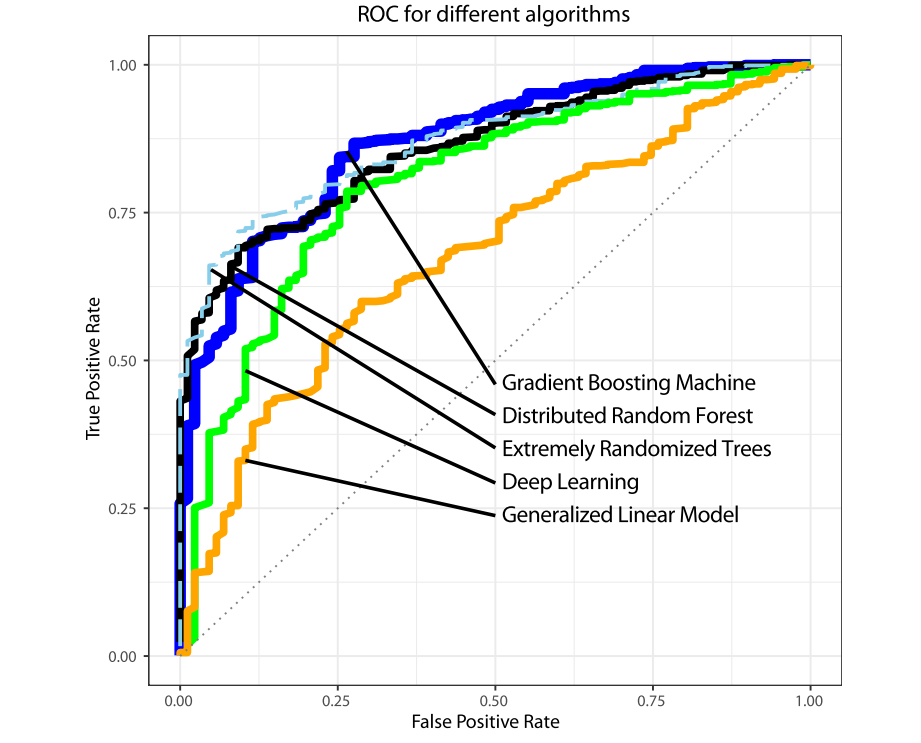

Fig. 3 The receiver operating characteristic curves for different model predictions. The area under the curve for the Gradient Boosting Machine model was 0.89. An actual predictive test for ovarian cancer would require choosing a threshold. Depending on the threshold, the true positive rate and false positive rate (or equivalently the sensitivity and specificity) will vary. This graph demonstrates how these two factors will vary

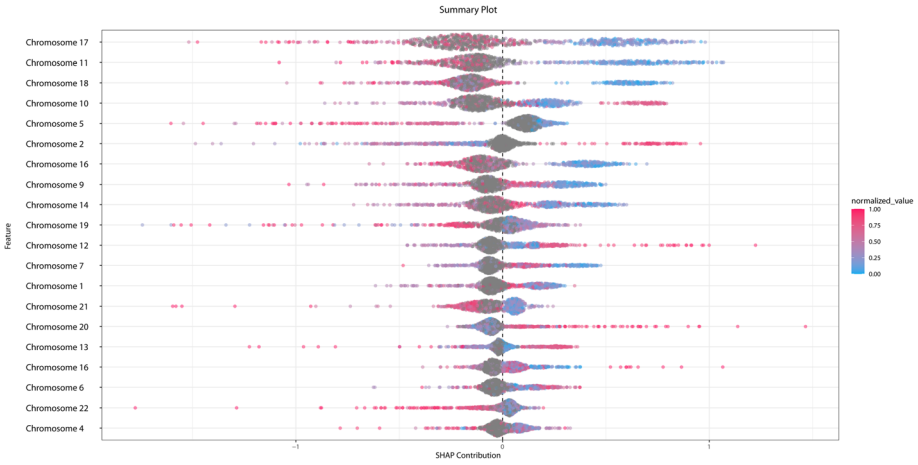

Fig. 4 This SHAP contribution plot ranks the importance to the predictive model for each chromosome [29]. Each person is represented by a dot. The color of the dot represents the normalized chromosome length. The position of the dot on the x-axis represents the impact of that chromosome on the model's prediction for that person. The figure indicates that Chromosome 17's length is more important than Chromosome 4's length in predicting ovarian cancer

Table 2 quantifies different algorithms applied to this problem. These results are illustrative, but not conclusive. Tuning machine learning models is an art, and it might be possible, for instance, to tune a deep learning network to obtain superior results. In similar work on TCGA colon cancer data, we found that a pairwise neuron network algorithm performs equal to a gradient boosting machine [30]. The gradient boosting machine generally runs faster and is easier to tune. Others have evaluated different machine learning algorithms for different bioinformatic problems and found that no one algorithm is superior [31]. They also found that a gradient boosting machine algorithm does perform well on many different types of datasets, consistent with these findings.

Germline mutations in the genes BRCA1 and BRCA2 are known to predispose women to ovarian and breast cancers. We considered whether these mutations had a significant effect on our results. First, 22 women in the TCGA ovarian cancer category had BRCA1 or BRCA2 germline mutations, while another 27 in the control group had BRCA1 or BRCA2 mutations (these were breast cancer patients, included here as controls because they were non-ovarian cancer women patients) [32]. Second, most common germline BRCA mutations change the overall length by just a few bases out of the 81 million bases on chromosome 17 [33]. This change would be imperceptible in our data, which focuses on large scale variations. Based on these two factors, we do not believe that BRCA1/2 mutations are responsible for the predictive ability presented here.

A disadvantage of this approach, compared to more conventional SNP-based genetic risk scores, is that the results are difficult to understand and extract biological meaning. A fundamental difference exists between statistical methods for prediction and those for attribution [34]. The method presented here is optimized for prediction, SNP-based genetic risk scores grew out of genome wide association studies, which were designed for attribution, identifying specific genes responsible for cancer.

The Gradient Boosting Machine computational model is complex, consisting of dozens of decision trees. Furthermore, the data that is used to traverse the decision tree is also complex. The data consists of chromosome scale length variation, which is the

result of many different insertions, deletions, translocations, and other structural changes. Polygenic risk scores based on SNPs are easy to interpret. One can identify how much each SNP contributes to the score and one can locate this SNP in the genome and understand the function of nearby genes that might change. Although this approach is lacking in explanatory power, its ultimate goal is predictive power.

We considered whether the results were due to two common problems faced by genome wide association studies: batch effects or population stratification. We found it unlikely that our model is identifying batch effects rather than real effects. First, all samples were collected from the same tissue, blood. This eliminates one common source of batch effects, since the DNA extraction process is the same for each sample. Second, all samples were processed by the same laboratory, the Nationwide Children's Hospital Biospecimen Core Resource, with the same type of instrument. This laboratory followed the same protocol throughout their processing phase. Finally, we looked up the batch history of each sample. The 424 ovarian cancer samples were processed in 15 separate batches. The non-ovarian samples were processed in several hundred different batches. For these reasons, we do not believe the results are due to batch effects.

Population stratification occurs in case/control studies when the cases and controls contain substantially different proportions of genetically discernable subclasses. Most TCGA samples were collected in the United States from a racially diverse group. For instance, over half the ovarian cancer samples were collected at five locations in the United States: Memorial Sloan Kettering, Washington University, University of Pittsburgh, Duke, and Mayo Clinic- Rochester. Table 1 lists demographic information about the two populations. Although the table does indicate slightly different proportions by race in the case and control groups, it does not seem to be different enough to account for the AUC observed. We cannot rule out population effects, but do not believe they would be responsible for such a large effect.

We could not use the typical process to correct for population stratification, because it is specific to logistic regression. The typical process uses the algorithm EIGENSTRAT to identify a number of (typically ten) principal components of the population [35]. Then, these principal components are fed into the logistic regression analysis to "correct for" or "adjust for" population stratification. This process of "adjusting for" a factor is unique to linear/logistic regression, it cannot be done in the same way with the non-linear machine learning algorithms. Again, the statistical algorithms for prediction are fundamentally different than those used for attribution [34].

This study has several weaknesses. First, the control population in this analysis is not randomly drawn from the general population, but instead consists of women who were part of the study because they were diagnosed with another form of cancer. This may lead to confound effects of the conclusions. Second, the results rely on a single dataset. The general applicability of this method would be better established if we were able to show that a model trained on one dataset would perform well on a second dataset that was collected independently. Demonstrating that a model is transferrable is a longer-term goal of ours.

Future work could refine this method to improve the predictive ability of this method. The AUC might be improved through several strategies, including feature engineering, for instance using sub-chromosomes rather than complete chromosomes, data augmentation strategies, and the inclusion of SNP data. Further work can also

establish how robust the model is: can a model trained with the TCGA data be successfully applied to a person not in the TCGA dataset.

Conclusion

A genetic risk score based on chromosomal-scale length variation of germ line DNA could provide an effective means of predicting whether or not a woman will develop ovarian cancer. Several avenues are open to further improve the AUC of this genetic risk score test.

Abbreviations

AUC: Area under the curve of the receiver operating characteristic curve; DNA: Deoxyribonucleic Acid; GBM: Gradient boosting machine; SNP: Single nucleotide polymorphism; TCGA: The Cancer Genome Atlas

Acknowledgements

The results published here are in whole or part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Availability of materials

The datasets analyzed during the current study are available from the NIH Data Commons at https://portal.gdc.cancer.gov/

Authors' contributions

JB conceived of the study. JB/CT analyzed the data. JB/CT wrote and edited the manuscript. The author(s) read and approved the final manuscript.

Declarations

Ethics approval and consent to participate

Ethics approval for the initial data collection was obtained by the TCGA project. The work presented here is based on anonymized data and is thus not considered human subjects research.

Consent for publication

Not Applicable.

Competing interests

The authors' do not have any competing interests.

Received: 3 November 2020 Accepted: 28 February 2021
Published online: 09 March 2021

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394–424. https://doi.org/10.3322/caac.21492.
2. Torre LA, Trabert B, DeSantis CE, Miller KD, Samimi G, Runowicz CD, et al. Ovarian cancer statistics, 2018. CA Cancer J Clin. 2018;68:284–96. https://doi.org/10.3322/caac.21456.
3. Bast RC. Status of Tumor Markers in Ovarian Cancer Screening. Journal of Clinical Oncology. 2003;21: 200s–2205. doi: https://doi.org/10.1200/JCO.2003.01.068
4. Andrews L, Mutch DG. Hereditary ovarian Cancer and risk reduction. Best Pract Res Clin Obstet Gynaecol. 2017;41:31–48. https://doi.org/10.1016/J.BPOBGYN.2016.10.017.
5. Grossman DC, Curry SJ, Owens DK, Barry MJ, Davidson KW, Doubeni CA, et al. Screening for ovarian cancer US preventive services task force recommendation statement. JAMA. 2018. https://doi.org/10.1001/jama.2017.21926.
6. Trimbos JB. Surgical treatment of early-stage ovarian cancer. Best Pract Res. 2017. https://doi.org/10.1016/j.bpobgyn.2016.10.001.
7. Bast RC, Lu Z, Han CY, Lu KH, Anderson KS, Drescher CW, et al. Biomarkers and Strategies for Early Detection of Ovarian Cancer. Cancer Epidemiology Biomarkers & Prevention. 2020; cebp.1057.2020. doi:https://doi.org/10.1158/1055-9965.EPI-20-1057
8. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial risk and heritability of Cancer among twins in Nordic countries. JAMA. 2016;315:68–76. https://doi.org/10.1001/jama.2015.17703.
9. Janssens ACJW, Aulchenko YS, Elefante S, Borsboom GJJM, Steyerberg EW, van Duijn CM. Predictive testing for complex diseases using multiple genes: fact or fiction? Genet Med. 2006;8:395–400. https://doi.org/10.1097/01.gim.0000229689.18263.f4.
10. Janssens ACJW, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. Hum Mol Genet. 2008;17:R166–73. https://doi.org/10.1093/hmg/ddn250.
11. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018. https://doi.org/10.1038/s41576-018-0018-x.
12. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019. https://doi.org/10.1093/hmg/ddz187.

13. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat Genet. 2018;50:1219–24. https://doi.org/10.1038/s41588-018-0183-z.
14. Pharoah PDP, Tsai Y-Y, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat Genet. 2013;45:362–70. https://doi.org/10.1038/ng.2564.
15. Kuchenbaecker KB, Ramus SJ, Tyrer J, Lee A, Shen HC, Beesley J, et al. Identification of six new susceptibility loci for invasive epithelial ovarian cancer. Nat Genet. 2015;47:164–71. https://doi.org/10.1038/ng.3185.
16. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. Genome Med. 2020;12:44. https://doi.org/10.1186/s13073-020-00742-5.
17. Goode EL, Chenevix-Trench G, Song H, Ramus SJ, Notaridou M, Lawrenson K, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nat Genet. 2010. https://doi.org/10.1038/ng.668.
18. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer genome atlas pan-Cancer analysis project. Nat Genet. 2013;45:1113–20. https://doi.org/10.1038/ng.2764.
19. Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, Dao F, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15. https://doi.org/10.1038/nature10166.
20. Hutter C, Zenklusen JC. The Cancer genome atlas: creating lasting value beyond its data. Cell. 2018;173:283–5. https://doi.org/10.1016/j.cell.2018.03.042.
21. Copy Number Variation Analysis Pipeline. [cited 18 Jan 2018]. Available: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/
22. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008;40:1253–60. https://doi.org/10.1038/ng.237.
23. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics. 2004;5:557–72. https://doi.org/10.1093/biostatistics/kxh008.
24. National Cancer Institute Genomic Data Commons. [cited 18 Jan 2018]. Available: https://gdc.cancer.gov/
25. Reynolds SM, Miller M, Lee P, Leinonen K, Paquette SM, Rodebaugh Z, et al. The ISB Cancer genomics cloud: a flexible cloud-based platform for Cancer genomics research. Cancer Res. 2017;77:e7–e10. https://doi.org/10.1158/0008-5472.CAN-17-0617.
26. Gijsbers P, LeDell E, Thomas J, Poirier S, Bischl B, Vanschoren J. An Open Source AutoML Benchmark. 6th ICML Workshop on Automated Machine Learning. 2019. Available: https://arxiv.org/pdf/1907.00909.pdf
27. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002;38:367–78. https://doi.org/10.1016/S0167-9473(01)00065-2.
28. Tenny S, Hoffman MR. Odds ratio (OR). StatPearls. StatPearls publishing; 2020. Available: http://www.ncbi.nlm.nih.gov/pubmed/28613750.
29. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Proces Syst. 2017.
30. Zhang B. Colorectal cancer predictive test using germ-line DNA data and multiple machine learning methods. 2019. Available: https://escholarship.org/uc/item/44f3f487
31. Olson RS, Cava W, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. Pacific symposium on Biocomputing Pacific symposium on Biocomputing. 2018;23: 192–203. Available: http://www.ncbi.nlm.nih.gov/pubmed/29218881.
32. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490:61–70. https://doi.org/10.1038/nature11412.
33. Abeliovich D, Kaduri L, Lerer I, Weinberg N, Amir G, Sagi M, et al. The founder mutations 185delAG and 5382insC in BRCA1 and 6174delT in BRCA2 appear in 60% of ovarian cancer and 30% of early-onset breast cancer patients among Ashkenazi women. Am J Hum Genet. 1997.
34. Efron B. Prediction, estimation, and attribution. J Am Stat Assoc. 2020. https://doi.org/10.1080/01621459.2020.1762613.
35. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006. https://doi.org/10.1038/ng1847.

Publisher's Note