Analysis of copy number variation from germline DNA can predict individual cancer risk

Authors: Chris Toh[1] and James P. Brody[1]*

Affiliations:

[1]Department of Biomedical Engineering, University of California, Irvine.

*Correspondence to: jpbrody@uci.edu.

Abstract:

BACKGROUND: Twin studies suggest that inherited factors are responsible for a substantial portion of many different forms of cancer. However, most of these inherited factors have yet to be identified. Most analyses of genome wide association data focus on the additive effects of single nucleotide polymorphisms found in germline DNA. The objective of this study is to test whether germline DNA copy number variation data can be used to predict whether a person will develop cancer.

METHODS: We tested this objective using data compiled by The Cancer Genome Atlas (TCGA) project. The TCGA data consisted of information on DNA copy number variation extracted from the peripheral blood of 8,821 different patients, each of whom had one of 32 different types of cancer. We used a machine-learning algorithm to test how well we could classify whether or not a patient had a certain cancer based solely on this data.

RESULTS: We found that all 32 different types of cancer tested could be predicted better than random chance using copy number variation data. Specifically, we measured the area under the receiver operator curve, known as the AUC, for ovarian cancer in women (0.89), glioblastoma multiforme (0.86), breast cancer in women (0.75), and colon adenocarcinoma (0.79), with a 95% confidence interval width of less than 0.01 in each case.

CONCLUSIONS: All cancers have a substantial inherited component. This method should be applicable to developing predictive genetic tests for conditions known to have a substantial genetic component.

Introduction:

Different lines of evidence suggest that a large fraction of common cancers is due to inherited factors. Cancer rates in monozygotic and dizygotic twins differ(*1*). Similarly, age-specific incidence studies suggest that most cancers occur in small subset of the population, probably determined by inherited genetics (*2*).

Genome wide association studies (GWAS) can identify inherited genetic factors that influence phenotypes(*3*). GWAS has been very successful at identifying single DNA alterations that influence traits, but if a trait requires two or more independent DNA alterations, known as epistatic interactions, it is generally invisible to GWAS using the typical statistical analysis(*4*, *5*).

Machine learning can take many weak predictors and create a strong ensemble classifier, combining these predictors in non-linear fashion; these algorithms should be able to identify epistatic interactions. However, machine-learning algorithms require more observations than predictors do. Hence, one cannot effectively apply machine learning algorithms to GWAS data when the number of observations, or cases and controls (typically less than 10,000), is less than the number of predictors, or single nucleotide polymorphisms, observed (typically half a million or more).

The best current genetic predictive tests use a polygenic risk score computed by adding the risk from 70-80 different single nucleotide polymorphisms (SNPs). In 2015, a large collaboration published a study demonstrating that a polygenic risk score based on 77 SNPs could stratify breast cancer risk in women. They found that women who scored in the top 1% had a threefold increase in risk compared to a woman who scored in the middle quintile (*6*). More recently, a group from Myriad Genetics developed a risk score based upon 82 different SNPs and applied it to a group of women who have a family history of breast cancer, but lack monogenic mutations in well-known breast cancer genes. They found a significant increase in breast cancer among those women with the highest risk scores(*7*).

To identify inherited disease caused by epistatic interactions one needs to identify a genetic measurement that condenses the information contained within the genome into fewer numbers, while maintaining sufficient information about genetic differences that characterize differences among humans. Here we test whether computing average copy number variations across large portions of a chromosome fulfills this need.

Results:

Figures 1 and Table 1 present the overall results. Table 1 shows the cancer type, number of batches, area under the receiver operator curve (AUC), the number of cases in the dataset, and the number of controls in the dataset. The average of five measurement is reported. With one exception, the standard deviation of the five measurements was less than 0.03, usually less than 0.01. The standard deviation of kidney chromophobe's five measurements was higher, 0.10, probably because the dataset only contains nine cases of kidney chromophobe. Figure 1 presents the receiver operator curves for four of the tumor types. These four were selected for different reasons. Gliobastoma multiforme (*15*) and ovarian serous cystadenocarcinoma (*16*) had the best AUC values, while colon adenocarcinoma (*17*) and breast invasive carcinoma (*18*) were chosen because they are the most common types of cancer.

Figure 2 presents the results of a study to determine how many CNV sites were contributing to the accuracy of the model. We studied models containing up to 10,000 different sites,

but found that most of the gains in AUC occurred with just the most common 25 CNV sites. Examination showed that these sites were usually hundreds of megabases long regions of different chromosomes.

Table 2 presents the results of an example test to predict glioblastoma multiforme in patients. A test dataset consisting of 20% of the data was set aside. A machine-learning model was trained using the other 80% of the data. The model was then applied to the test data. We found that we could predict about 10% of these brain cancers without any false positives.

We also tested the effect of race on these predictions. The racial category of patients in the TCGA data is listed as either "white" (6,406 patients), "black or African American" (776 patients), or "Asian" (603 patients). A very small number are listed as "American Indian or Alaska native" (20 patients) or "native Hawaiian or other Pacific Islander" (11 patients). "Unknown or Not available" was listed for 1,010 of the 8,826 patients. We constructed a training set consisting of 80% of the patients with a racial category listed as "White." We then built models to predict ovarian cancer, breast cancer, colon cancer and glioblastoma multiforme using data from only white patients. We tested the predictive ability of these models on three different datasets of patients identified as either white, black or African American, and Asian. The breakdown for these datasets is shown in Table 3. The results of these predictive tests are shown in Table 4.

Discussion:

Several different polygenic risk scores have been developed for different forms of cancer(*19*, *20*). Breast cancer was the first target of genetic risk scores with the discovery of BRCA1. It probably has the most advanced polygenic risk scores available today. The 2015 study that computed a polygenic risk score based on 77 SNPs found that women who scored in the top 1% had a three-fold increase in risk compared to a woman who scored in the middle quintile (*6*). For comparison, in the ovarian cancer prediction presented here 24 of 33 in the top 3% had ovarian cancer, while 2 of 33 in the middle 3% had ovarian cancer, for an approximate tenfold increase in risk. All 11 of the women scoring in the top 1% had developed ovarian cancer. Our test for ovarian cancer performs substantially better than the best existing predictive test for cancer.

Machine learning techniques have been used to build polygenic risk scores to predict other complex traits(*21–23*). For instance, diabetes can be predicted from SNP data with an area under the receiver operator curve (AUC) of 0.602(*23*) using a gradient boosted regression tree.

Heritability measures how well one can predict a phenotype of an offspring based on the two parents(*24*). This definition is often referred to as "broad-sense heritability." This heritability can be estimated from studies of monozygotic and dizygotic twins(*1*, *25*). A large gap exists between heritability measured from twin and family studies and the extent to which we can currently predict a phenotype from genetic data. This gap is called the missing heritability or phantom heritability(*5*, *26*). Twin studies assume that monozygotic twins are genetically identical. This assumption, however, does not hold for copy number variation(*27*, *28*). Substantial difference in copy number variation occurs after conception. Thus, predictions based on copy number variation measured in peripheral DNA could be even higher than estimated by heritability measured in twin studies.

Interpreting the predictions made here is difficult, but not necessary. Although risk prediction and association studies share common methods, the end goals differ. Association studies

often try to identify alterations in specific genes that can be mechanistically tied to specific diseases. Risk prediction, however, is only concerned with maximizing the predictive power(*20*).

Our predictions are generated using gradient boosting machines, which are derived from decision trees(*23*, *29*). Decisions trees are easily interpretable. However, the useful variables, a form of copy number variations, are not. For instance, one point in the decision tree asks whether a certain measured copy number variation, expressed as a segmented mean value of a 133 megabase region of chromosome 11 (specifically between bases 456,012 and 134,272,740 in hg38 coordinates) is greater than or less than 0.0090125. Copy number variations are typically integral, but this measurement is not. It measures the average copy number variation over 99% of chromosome 11, and the segmented mean value of 0.0090125 represents an excess of about 1% in that region. This is one of thousands of different decisions the algorithm examines to compute a prediction. Identifying the mechanisms behind these predictions is not simple, but something we plan on working towards.

Interpreting the area under the receiver operator curve (AUC), which is also known a C-index is straightforward. Classification tests are a tradeoff between the true positive rate, or sensitivity, and the false positive rate, or 1-selectivity. These tests provide a numerical value. In practical use, a cutoff value must be selected, where all tests above the cutoff fall into one class and all below fall into another. The cutoff value is arbitrary; different cutoff values will give different sensitivity and selectivity values. The receiver operator curve summarizes the sensitivity and selectivity values for different potential cutoff values. The integral of this curve, known as the area under the curve (AUC) summarizes the performance of the test independent of the arbitrary cutoff value. An AUC value of 1.0 is perfection. The test classifies each case correctly. An AUC value of 0.5 is useless; it represents a random test.

Some example area under the receiver operator curve (AUC)s for diagnostic tests are: the BNP test for congestive heart failure AUC=0.91 (*30*), the PSA test for prostate cancer AUC=0.68 (*31*), and more recently the Cancer SEEK test, a “blood biopsy” to determine whether one of six types of cancer are present AUC=0.91 (*32*). Predictive tests generally have lower AUCs than diagnostic tests. A predictive test for coronary artery disease using genetic information has an AUC of 0.62, but when combined with conventional risk factors it can be increased to 0.68(*33*). The best predictive breast cancer tests incorporating multiple SNPs and family history have an AUC in the range 0.7 to 0.8 (*6*, *7*). For comparison, 7 of the 31 cancer tests shown here have an AUC greater than or equal to 0.8 and 15 of the 31 tests have an AUC greater than or equal to 0.75 suggesting that most of these tests will be at least as useful as the current predictive tests for breast cancer.

These test results should be considered lower limits. The results could improve in several ways. The machine-learning algorithm could be improved. Deep learning networks can outperform gradient boosting machines with sufficient tuning and more data. The results might improve with unmasked copy number variation data. Figure 3 shows that most of the increase in AUC occurs in the first 25 variables. Examination of the number of patients who had each copy number variation indicated a sharp drop off after the first 25, presumably an artifact of the masking process. Many of the copy number variation measurements overlap and this information is only contained in the model in a rudimentary way. We saw a substantial improvement in the AUC by combining all copy number variants that shared the same endpoint.

We considered whether the results were due to two common problems faced by GWAS studies: batch effects or population stratification. This analysis might be identifying batch effects rather than real effects (*34*), but it seems unlikely. 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. Most of the samples were processed in multiple batches (the number of batches is indicated in Table 2). Samples from five different types of tumor patients were processed in individual batches, but these were not outliers in the results.

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 3 outlines the breakdown. This table does indicate slightly different proportions, by race, in the case and control groups. The ultimate test of a predictive model is whether one derived from one group can applied to a different group. Table 4 demonstrates that the predictive model derived from a strictly white population works equally well, if not better, in an Asian population. It also performs good, but substantially worse in a black/African American population. This observation is consistent with our understanding of how genetic variation correlates with race(*35*, *36*).

Cancer is the result of a complex interaction between genetics and the environment. In some cases, for instance, lung cancer and smoking or mesothelioma and asbestos exposure, the required environmental exposure is significant and well known. In other cases, the required environmental exposure is probably minor and not well known. The genetic signature identified here is a necessary, but not sufficient factor in developing the cancer. Since this is a retrospective study of people who already developed cancer, sufficient environmental exposure has already occurred. A prospective study would need to be performed to determine the effect of environmental exposure on how effective these predictions are.

Conclusions:

In conclusion, we found that most cancers have significant inherited components. Machine-learning techniques applied to germline DNA copy number variation can effectively predict whether a person will develop many of these tumors. This method should be generally applicable to predict any cancers or other complex genetic conditions including conditions as diverse as heart disease and schizophrenia.

Methods:

The Cancer Genome Atlas (TCGA) was a project sponsored by the National Cancer Institute to characterize the molecular differences in 33 different human cancers (*8*, *9*). 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 acute myeloid leukemia.)

Most of the patient blood samples were processed to extract 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 (*10*), which included the packages Birdsuite (*11*) and DNAcopy (*12*). 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.

NCI has provided most of the TCGA data on the Genomic Data Commons (*13*). The copy number variation is called the masked copy number variation. The masking process removes "Y chromosome and probe sets that were previously indicated to have frequent germline copy-number variation." (*10*)

The final data set that we used is the masked copy number variation data. It originates from normal blood samples extracted from 8,826 different patients: 4,692 female and 4,134 male. The patient's age ranged from 10 to 90 years old.

In this dataset, about 695,000 different genomic regions characterized by the chromosome number, a start position, and an end position have additions or deletions, known as copy number variations, in at least one patient.

We initially selected the top 10,000 of these 695,000 regions, ranked by the number of different patient samples that contained that CNV, to create a dataset with 8,826 rows and 10,000 columns. After some work on this dataset, we noted that many overlapping CNVs exist. As an improved second version, we used an alternate grouping characterized simply by the chromosome number and the end location. This groups together all CNVs that share the same end location. This second version gave better results than the first version for most tumors. We used the second version throughout the manuscript.

For each type of tumor, we created a case/control study. Cases, shown in Table 2, were the number of "normal blood" samples from patients with those diagnoses. This quantity is the number of patients who also have measurements of copy number variation from DNA derived from normal blood in the database.

Controls, shown in Table 2, were the number of "normal blood" samples from patients without that specific diagnosis. These normal blood samples included patients with other diagnosis. A few forms of cancer were specific to women (cervical squamous cell carcinoma, ovarian serous cystadenocarcinoma, uterine corpus endometrial carcinoma, and breast invasive carcinoma) so the controls only consisted of female controls. The male specific forms of cancer (prostate adenocarcinoma and testicular germ cell tumors) had only male controls. Nine cases of male breast cancer are included in the dataset, so breast cancer was analyzed both as female specific and for either sex. Each cancer test was set up as a binary classification supervised learning task that was trying to distinguish between patients diagnosed with one cancer (breast cancer, for example) and those not diagnosed with that cancer.

We performed the analysis using the statistical language R. The TCGA copy-number variation data was acquired through a query to Google's BigQuery, which hosts a copy of the

TCGA data as part of the Institute of Systems Biology Cancer Genomics Cloud(*14*). The data was reformatted for analysis, and then was fed into the machine-learning algorithm. The data included the patient's sex, whether or not they had the particular cancer, and measurements of copy number variation at different locations derived from the patient's peripheral blood sample. The data fed into the machine-learning algorithm did not include age, since germline DNA should not depend upon age. The results are independent of the patient's age.

The machine-learning algorithm used the h2o platform, provided by the h2o library in R (http://www.h2o.ai). We initially used a deep learning network for the analysis and had good results without tuning any hyperparameters. We used the default values provided by h2o (two hidden layers, each with 200 neurons, the default learning rate, etc.). We tested several other algorithms. We generally obtained the best results using the gradient boosting machine algorithm, again without tuning any of the hyperparameters. The gradient boosting machine algorithm was considerably faster, so we present our results here with the gradient boosting machine algorithm.

We used 10-x cross-validation for the analysis. The result of the machine learning analysis is a computational model that provides a "score," indicating a likelihood of a patient having a cancer, given a list of copy number variations. The computational model can be characterized by the area under the curve of the receiver operator curve, commonly referred to as the AUC of the ROC. We recorded the AUC of the model and repeated the process five times. This provided a measure of the variance of the process.

An independent validation strategy gave similar results to the cross validation analysis. We split the data randomly into two sections, an 80% training set, and a 20% validation set. The training set was used to build the model without any input from the validation set. Once the model was built, it was applied to the validation set. The performance metrics for this strategy were similar to those observed in the cross validation study.

We also tested a random control. One third of the patients in the dataset were randomly labelled as cancer, while the other two thirds were labelled as healthy. Following the same independent validation procedure, we attempted to build a model that could predict to which category a patient should belong. For this random control, we found that the area under the receiver operator curve (AUC) was 0.50 with a standard deviation of 0.01. The classification was equivalent to random guessing.

References and Notes:

1. P. Lichtenstein *et al.*, Environmental and Heritable Factors in the Causation of Cancer — Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).

2. L. Soto-Ortiz, J. P. Brody, A theory of the cancer age-specific incidence data based on extreme value distributions. *AIP Adv.* **2** (2012), doi:10.1063/1.3699050.

3. P. M. Visscher *et al.*, 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

4. L. Crawford, P. Zeng, S. Mukherjee, X. Zhou, Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genet.* **13**, e1006869 (2017).

5. O. Zuk, E. Hechter, S. R. Sunyaev, E. S. Lander, The mystery of missing heritability:

Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1193–8 (2012).

6. N. Mavaddat *et al.*, Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. *JNCI J. Natl. Cancer Inst.* **107** (2015), doi:10.1093/jnci/djv036.

7. E. Hughes *et al.*, Development and validation of a residual risk score to predict breast cancer risk in unaffected women negative for mutations on a multi-gene hereditary cancer panel. *J. Clin. Oncol.* **35**, 1579–1579 (2017).

8. K. Chang *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

9. D. M. Muzny *et al.*, Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. **487**, 330–337 (2012).

10. Copy Number Variation Analysis Pipeline, (available at https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/).

11. J. M. Korn *et al.*, Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).

12. A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. **5**, 557–572 (2004).

13. National Cancer Institute Genomic Data Commons, (available at https://gdc.cancer.gov/).

14. S. M. Reynolds *et al.*, The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research. *Cancer Res.* **77**, e7–e10 (2017).

15. C. W. Brennan *et al.*, The somatic genomic landscape of glioblastoma. *Cell*. **155**, 462–77 (2013).

16. D. Bell *et al.*, Integrated genomic analyses of ovarian carcinoma. *Nature*. **474**, 609–615 (2011).

17. T. C. G. A. Network *et al.*, Comprehensive molecular characterization of human colon and rectal cancer. *Nat. 2012 4877407*. **487**, 330 (2012).

18. D. C. Koboldt *et al.*, Comprehensive molecular portraits of human breast tumours. *Nature*. **490**, 61–70 (2012).

19. H. Li *et al.*, Breast cancer risk prediction using a polygenic risk score in the familial setting: a prospective study from the Breast Cancer Family Registry and kConFab. *Genet. Med.* **19**, 30–35 (2017).

20. G. Abraham, M. Inouye, Genomic risk prediction of complex human disease and its clinical application. *Curr. Opin. Genet. Dev.* **33**, 10–16 (2015).

21. S. Okser *et al.*, Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLoS Genet.* **10**, e1004754 (2014).

22. J. Kruppa, A. Ziegler, I. R. König, Risk estimation and risk prediction using machine-learning methods. *Hum. Genet.* **131**, 1639–54 (2012).

23. G. Paré, S. Mao, W. Q. Deng, A machine-learning heuristic to improve gene score prediction of polygenic traits. *Sci. Rep.* **7**, 12665 (2017).

24. P. M. Visscher, W. G. Hill, N. R. Wray, Heritability in the genomics era — concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

25. L. A. Mucci *et al.*, Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. **315**, 68 (2016).

26. B. Maher, Personal genomes: The case of the missing heritability. *Nature*. **456**, 18–21 (2008).

27. C. E. G. Bruder *et al.*, Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82**, 763–71 (2008).

28. A. Abdellaoui *et al.*, CNV Concordance in 1,097 MZ Twin Pairs. *Twin Res. Hum. Genet.* **18**, 1–12 (2015).

29. J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **29**, pp. 1189–1232.

30. C. M. Florkowski, Sensitivity, specificity, receiver-operating characteristic (ROC) curves and likelihood ratios: communicating the performance of diagnostic tests. *Clin. Biochem. Rev.* **29 Suppl 1**, S83-7 (2008).

31. I. M. Thompson *et al.*, Operating Characteristics of Prostate-Specific Antigen in Men With an Initial PSA Level of 3.0 ng/mL or Lower. *JAMA*. **294**, 66 (2005).

32. J. D. Cohen *et al.*, *Science (80-.).*, in press, doi:10.1126/SCIENCE.AAR3247.

33. M. Inouye *et al.*, Genomic risk prediction of coronary artery disease in nearly 500,000 adults: implications for early screening and primary prevention. *bioRxiv*, 250712 (2018).

34. J. T. Leek *et al.*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

35. L. B. Jorde, S. P. Wooding, Genetic variation, classification and "race." *Nat. Genet.* **36**, S28–S33 (2004).

36. K. L. Hunley, G. S. Cabana, J. C. Long, The apportionment of human diversity revisited. *Am. J. Phys. Anthropol.* **160**, 561–569 (2016).

Acknowledgments:

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

Fig. 1. The receiver operator curves for four different cancer types. The receiver operator curves characterize how the test's true positive rate varies with its false positive rate. The overall curve can be summarized by the area under the curve (AUC). An AUC of 0.50 is indicative of a random guess. An AUC of 1.0 is a perfect test; all cancer cases and controls are correctly classified.

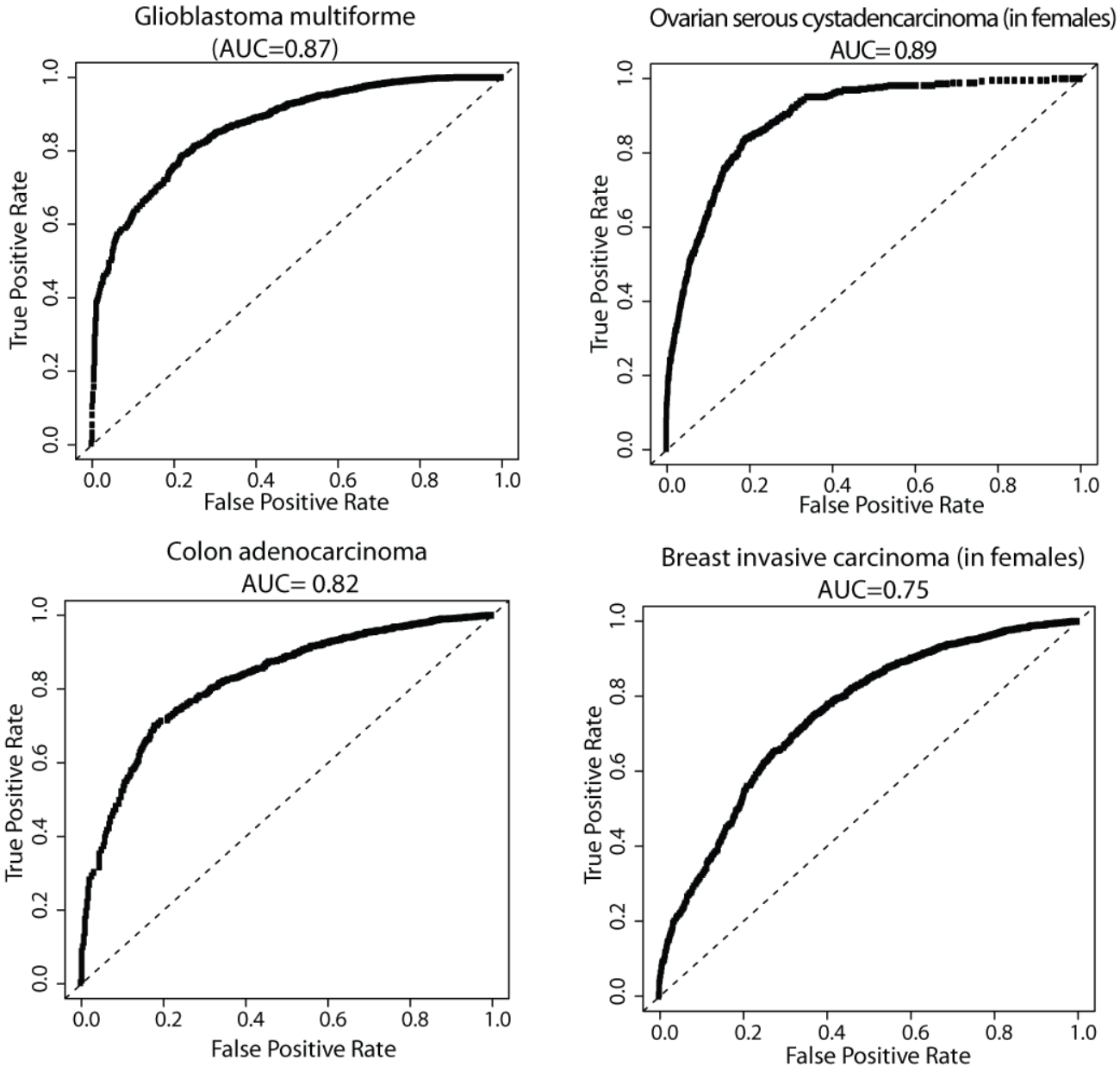

Fig. 2. This figure presents the AUC as a function of the number of CNV sites included in the model for four types of cancer: ovarian (OV), glioblastoma multiforme (GBM), colon (COAD), and breast (BRCA). We measured this up to 10,000 sites, but found that most of the information was contained in the top 25 sites.

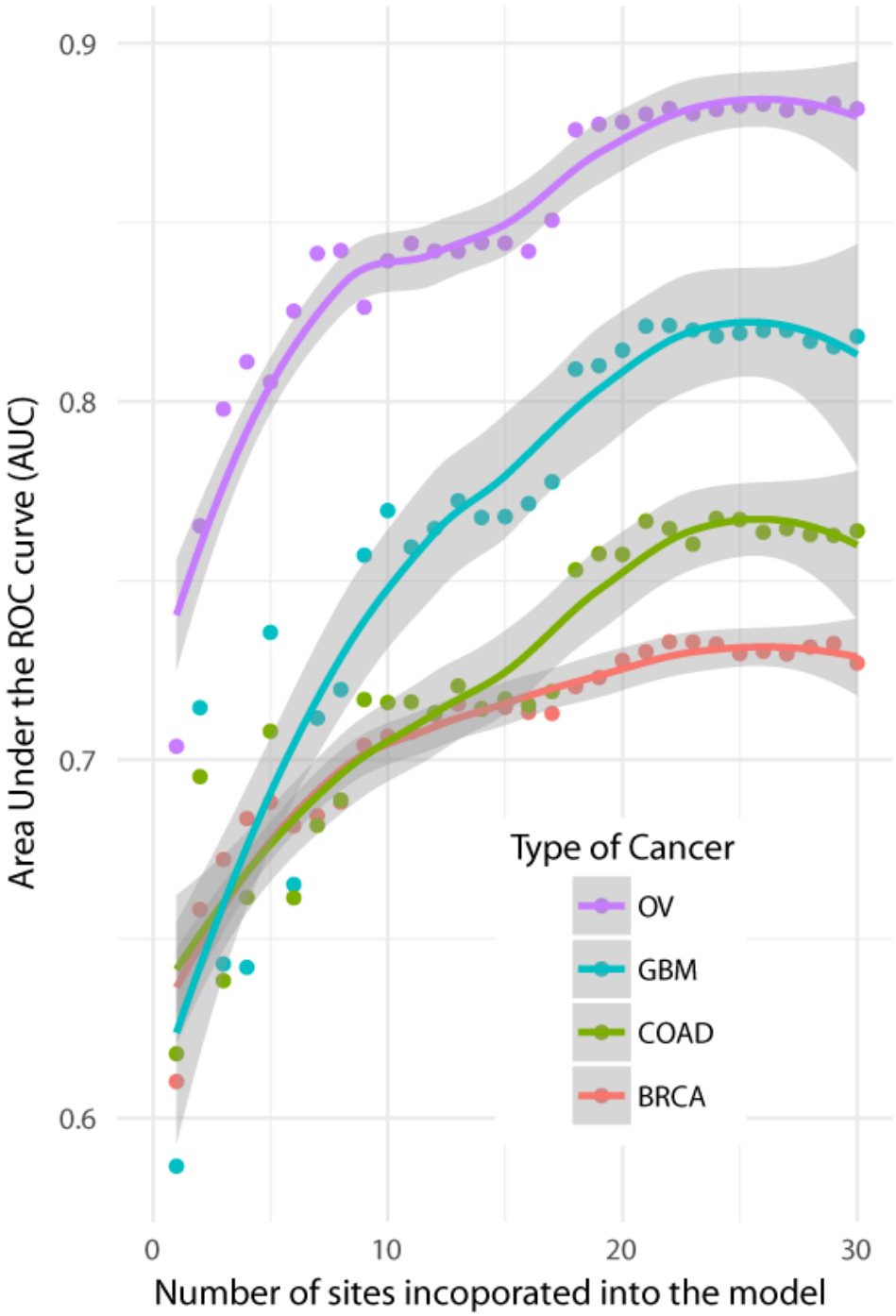

Table 1. This table shows the cancer type, number of batches, area under the receiver operator curve (AUC), the number of cases in the dataset, and the number of controls in the dataset. The average of five measurements is reported. With one exception, the standard deviation of the five measurements was less than 0.03, usually less than 0.01. The standard deviation of kidney chromophobe's five measurements was higher, 0.10, probably because there were only nine cases of kidney chromophobe in the dataset.

Type of Cancer	Batches	Cases	Controls	AUC
Adrenocortical carcinoma	2	85	9713	0.77
Bladder Urothelial Carcinoma	23	361	9437	0.70
Breast invasive carcinoma (in women)	37	968	3715	0.73
Breast invasive carcinoma (in men and women)	37	977	8821	0.81
Cervical squamous cell carcinoma	19	281	4411	0.71
Cholangiocarcinoma	2	34	9764	0.71
Colon adenocarcinoma	20	374	9424	0.79
Lymphoid/ Diffuse Large B-cell Lymphoma	6	42	9756	0.76
Esophageal carcinoma	10	124	9674	0.78
Glioblastoma multiforme	19	484	9314	0.86
Head and Neck squamous cell carcinoma	18	488	9310	0.66
Kidney Chromophobe	3	9	9789	0.71
Kidney renal clear cell carcinoma	11	110	9688	0.82
Kidney renal papillary cell carcinoma	16	219	9579	0.75
Brain Lower Grade Glioma	16	485	9313	0.66
Liver hepatocellular carcinoma	19	306	9492	0.69
Lung adenocarcinoma	20	402	9396	0.67
Lung squamous cell carcinoma	21	296	9502	0.72
Mesothelioma	2	85	9713	0.82
Ovarian serous cystadenocarcinoma	15	424	4268	0.89
Pancreatic adenocarcinoma	11	146	9652	0.70
Pheochromocytoma and Paraganglioma	1	175	9623	0.84
Prostate adenocarcinoma	18	423	3711	0.66
Rectum adenocarcinoma	10	132	9666	0.75
Sarcoma	15	233	9565	0.72
Skin Cutaneous Melanoma	14	465	9333	0.73
Stomach adenocarcinoma	17	369	9429	0.74
Testicular Germ Cell Tumors	1	132	4002	0.73
Thyroid carcinoma	17	417	9381	0.73
Thymoma	1	111	9687	0.78
Uterine Corpus Endometrial Carcinoma	25	504	4188	0.71
Uterine Carcinosarcoma	1	48	4644	0.78
Uveal Melanoma	1	80	9718	0.80

Table 2. A machine-learning ranking for glioblastoma multiforme (GBM). The dataset was split into a training/validation set (80%) and a test set (20%). The model was built with the training/validation set and then applied to the test set. The test set contained DNA copy number variation information from 1704 patients. Of these, 101 patients had been diagnosed with glioblastoma multiforme and 1603 had not. The 1603 were designated as controls. The model computes a GBM-score based on the patient's gender and DNA copy number variation. The patient's age is not used in the model, but is provided here. This table lists the top 25 patients, ranked by the model's GBM-score. The probability that the top 12 were all drawn from the GBM class by chance is about one in 10^{14}.

Rank	GBM-score	Class	Gender	Age (years)
1.	0.998461	GBM	MALE	50
2.	0.993502	GBM	MALE	69
3.	0.993179	GBM	MALE	71
4.	0.987589	GBM	MALE	34
5.	0.98755	GBM	MALE	54
6.	0.984182	GBM	MALE	54
7.	0.98272	GBM	MALE	70
8.	0.973697	GBM	FEMALE	65
9.	0.971379	GBM	FEMALE	75
10.	0.955155	GBM	FEMALE	63
11.	0.949079	GBM	FEMALE	54
12.	0.921413	GBM	FEMALE	51
13.	0.812008	Control	FEMALE	35
14.	0.793035	Control	FEMALE	54
15.	0.788271	Control	FEMALE	52
16.	0.748749	GBM	MALE	64
17.	0.728986	GBM	FEMALE	58
18.	0.700573	Control	FEMALE	58
19.	0.6934	GBM	FEMALE	79
20.	0.685881	GBM	FEMALE	78
21.	0.665199	Control	FEMALE	61
22.	0.662652	GBM	FEMALE	59
23.	0.640001	GBM	FEMALE	65
24.	0.595295	GBM	MALE	24
25.	0.56173	Control	FEMALE	52

Table 3. The racial composition of the cases and controls for four types of cancer in the dataset.

	Ovarian		Gliobastoma Multi-forme		Breast Cancer		Colon Cancer	
	Case	**Control**	**Case**	**Control**	**Case**	**Control**	**Case**	**Control**
White	359	3098	422	5984	652	2805	196	6210
Black or Afr. Am.	25	496	45	731	163	358	57	719
Asian	16	261	11	592	59	218	11	603

Table 4. We trained a model with data exclusively from patients labelled as "white." A portion (20%) of the white patients was withheld from the training set. The model was then applied to three different validation datasets. The first contained the withheld 20% of the white patients, the second contained all patients labelled Black or African American and the third contained all patients labelled Asian. The model trained on white patients consistently performed better on Asian patients than on Black/African American patients.

Validation	Ovarian	Gliobastoma Multiforme	Breast Cancer	Colon Cancer
	AUC	**AUC**	**AUC**	**AUC**
White	0.93	0.82	0.74	0.81
Black or Afr. Am.	0.70	0.73	0.56	0.57
Asian	0.98	0.88	0.83	0.81