UNIVERSITY OF CALIFORNIA,

IRVINE

Chromosomal scale length variations as a genetic risk score for predicting complex human diseases in large scale genomic datasets

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Biomedical Engineering

by

Christopher En-Li Toh

Dissertation Committee:

Associate Professor James P. Brody, PhD, Chair

Professor Gregory J. Brewer, PhD

Professor William C. Tang, PhD

Associate Professor Timothy L. Downing, PhD

2021

DEDICATION

To my wife Amy Noreen Toh, my best friend and steadfast helper,

my parents Tein-Su Samuel Toh and Yu Lisa Liu
who sacrificed daily for me,

my brother Andrew En-Le Toh,
who encourages me to find joy,

my friends,
who encourage me towards excellence,

to my aunt and uncle Suzanne and Rongguang Ng,
who were faithful to God even through the cancer Suzanne experienced,

to the patients and their families,
who have suffered from the diseases we have studied

Soli Deo Gloria

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to extend my sincere thanks and appreciation to my committee chair, Dr. James P. Brody, whose support, and guidance have been foundational to this work. He consistently encouraged and listened to my ideas, and urged me to examine my work critically. His expertise and guidance have helped shape and improve my skills as a scientist and researcher. This work would not be possible without him.

I would like to thank my committee member, Dr. William Tang, for the opportunity to serve with him in Vietnam and teaching me the importance to think about global health, while also keeping in mind the unique challenges each individual patient faces. His work has motivated me to broaden the scope of my research and to consider the possibilities of applying engineering techniques beyond their initial purpose.

I would like to thank my committee member, Dr. Timothy Downing for his keen interest in our work and his great advice to approach new ideas with a combined and multi-disciplinary mindset. The collaboration with his lab has been a staunch support to the production of this work.

I would like to thank, Dr. Gregory Brewer, for emphasizing the importance of clear communication in scientific research and the humility he exemplifies in the field.

Financial support was provided by The Henry Samueli School of Engineering at The University of California – Irvine through The Engineering Bridge Fellowship. Additional funding was provided by UC Irvine Graduate Division through the Public Impact Fellowship, Science Communication Fellowship, and Department of Teaching Excellence and Innovation Fellowship.

VITA

Christopher En-Li Toh

2017	B.S. in Bioengineering, University of California, Los Angeles
2017-21	Software Tester, Boeing Digital Solutions & Analytics
2017-19	Graduate Researcher & Teaching Assistant, Department of Biomedical Engineering, University of California Irvine
2019	M.S. in Biomedical Engineering, University of California, Irvine
2019-21	Graduate Researcher & Teaching Assistant, Department of Biomedical Engineering, School of Biological Sciences, University of California Irvine
2021	Ph.D. in Biomedical Engineering, University of California, Irvine

FIELD OF STUDY

Computational Genomics and Machine Learning in Complex Human Diseases

PUBLICATIONS

Chromosomal scale length variation of germline DNA can predict individual cancer risk – Chris Toh & James P. Brody. 2018. https://doi.org/10.1101/303339

Disruption of artificial lipid bilayers in the presence of transition metal oxide and rare earth metal oxide nanoparticles – Acharya et al. *Journal of Physics D: Applied Physics.* 2018. https://doi.org/10.1088/1361-6463/aaeb6e

Improved Measurement of Proteins Using a Solid-State Nanopore Coupled with a Hydrogel – Acharya et al. *ACS Sensors.* 2020. https://doi.org/10.1021/acssensors.9b01928

Evaluation of a Genetic Risk Score for Severity of COVID-19 Using Human Chromosomal-scale Length Variation – Toh, Brody. *Human Genomics BMC.* 2020. https://doi.org/10.1186/s40246-020-00288-y

Applications of Machine Learning in Healthcare – Toh, Brody. Smart Manufacturing - When Artificial Intelligence Meets the Internet of Things. *IntechOpen.* 2021. https://doi.org/10.5772/intechopen.92297

Genetic risk score for ovarian cancer based on chromosomal-scale length variation – Toh, Brody. *BioData Mining.* 2021. https://doi.org/10.1186/s13040-021-00253-y

Genetic risk score for predicting schizophrenia using human chromosomal-scale length variation – Toh, Brody. *In Review.* 2021. https://doi.org/10.21203/rs.3.rs-2685589/v1

Abstract of The Dissertation

Chromosomal scale length variations as a genetic risk score for predicting complex human diseases in large scale genomic datasets

By

Christopher En-Li Toh

Doctor of Philosophy in Biomedical Engineering

University of California, Irvine, 2021

Professor James P. Brody, Chair

Next generation sequencing has created large databases of human genomic information. Utilizing this information to understand disease and genetic risks is a large engineering task. Previous studies have focused primarily on single nucleotide polymorphisms (SNPs) in assessing patient risk for diseases such as cancers and other diseases such as Schizophrenia. These SNP panels do not consider epistatic interactions in the human genome.

Chromosomal scale-length variation (CSLV) is a promising approach for assessing genetic risk scores. CSLV evaluates copy number variations (CNVs), condensing genomic information into a smaller number of parameters. Reducing parameters allows the use of machine learning without the need for millions of patients' data. Machine learning can consider epistatic interactions that might be missed by conventional genome wide association studies (GWAS).

Utilizing machine learning classification algorithms, we assessed prediction of diseases such as ovarian cancer and schizophrenia using CSLV as the sole features for prediction. We have demonstrated the viability of this method in assessing germline inheritance of complex human diseases in The Cancer Genome Atlas (TCGA) and UK Biobank.

We tested 33 different types of cancer from TCGA's 11,000 patients. Glioblastoma multiforme (AUC = 0.87), ovarian cancer (AUC = 0.89), colon adenocarcinoma (AUC = 0.82), and breast invasive carcinoma (AUC = 0.75) could be distinguished greater than chance from cancers.

These results were replicated the UK Biobank using 88 numbers computed from the 22 autosomes for 1,534 women with breast cancer and a control population of 4,391 women without breast cancer and found a classifier with an AUC of 0.83.

1,129 people from the UK Biobank have a diagnosis of schizophrenia. Using a randomized set of 1,129 individuals without schizophrenia we created 150 models using 92 number CSLVs as our feature set. The results provided an average AUC of 0.545 (95% CI 0.539-0.550). Our results indicate that CSLV data can provide an effective genetic risk score for schizophrenia.

In conclusion, CSLV is a promising and novel way to utilize large scale human genetic information in the prediction off complex. Continued improvement of this technique can dramatically improve individualized patient care and can aid physicians in earlier diagnosis.

Chapter 1: Introduction

With the advent of next generation sequencing and rapid acceleration of computer technologies, biological and medical fields can directly benefit from the convergence of multiple disciplines. Utilizing genetic information in order to understand disease and genetic risks is a large engineering task[1]. Previous studies have focused primarily on single nucleotide polymorphisms (SNPs) as the primary feature of interest in assessing patient risk for diseases such as cancers and other possibly heritable diseases such as Schizophrenia[2]. However, these SNP panels do not consider epistatic interactions between various portions of the human genome[3,4].

Next generation sequencing (NGS) technologies have exponentially increased the available data, while also drastically decreasing the cost[1]. However, finding methods to handle massive amounts of data is the next hurdle in engineering solutions which can aid in better clinical outcomes[5]. The objective is to find relevant clinical information which can be used in an actionable manner by physicians and patients alike[6].

In the past, Genome-wide Association Studies (GWAS) have largely focused on single nucleotide polymorphisms (SNPs) and were generally performed on somatic samples derived from patients[7–10]. However, it is difficult to achieve accurate predictive results for diseases such as cancers or schizophrenia. At most, a panel of associated gene mutations or genetic variations are the most common results of these studies[2,8,10,11]. As such risk scores derived from such studies typically do not perform well especially on an individualized basis for most patients.

GWAS derived risk scores for many diseases still struggle to achieve good results. For example, one test for breast cancer achieved an area under the curve (AUC) of 0.68[12]. Other studies that do not report an AUC typically report an odds ratio of 3.36 in the top 1% of women[13]. Current risk scores for schizophrenia perform similarly with an AUC of 0.62[14]. For ovarian cancer one risk score has an odds ratio of 1.77 for the upper quintile[15]. As such we aim to see if a risk score that does not rely on SNP panels can perform just as well or better. Interestingly there is also a recent study which found some genetic correlation between schizophrenia and breast cancer[16].

Chromosomal scale-length variation (CSLV) is a promising new approach for assessing genetic risk scores[17]. This method includes epistatic effects which might be missed by conventional genome wide association studies (GWAS). GWAS typically uses a linear combination of SNP scores to assess genetic risk. CSLV evaluates copy number variations (CNVs) across large sections of the human genome to obtain a comprehensive account of variations which may contribute to inheritance of disease risk.

Utilizing modern machine learning classification algorithms, we assessed prediction of diseases such as ovarian cancer and schizophrenia using CSLV as the sole features for prediction. CSLV measures of a person's genetic variation using copy number variation (CNV) as the basis for the measurement and examines this variation across the large sections of the Chromosomes. This means the test can be done through simple blood samples. Utilizing large databases such as The Cancer Genome Atlas (TCGA) and the UK Biobank, we have demonstrated the viability of this method in assessing germline inheritance of complex human diseases. We utilized h2o, a machine learning framework and assessed the performance of several models, including general linear models (glm), gradient boosted machines (gbm), XGBoost, and stacked ensembles.

Chapter 2: Objectives and Specific Aims

Human genomes contain structural variations composed of repeats and deletions which are often defined as Copy Number Variation (CNV) [18]. In terms of utilizing CNVs, we hypothesize that hereditary cancer disease risk is a result of many different CNVs interacting together in a highly nonlinear fashion. We also hypothesize that other complex diseases may also be affected by this network effect of hereditary CNVs. The hypothesis posits that CNVs create a network effect between each other, contributing to risk, and which we can accurately use to predict risk through machine learning models. This network effect may also include other genomic variations such as SNPs or methylation. However, we aim to demonstrate a significant contribution from CNVs alone.

Objective 1

The primary objective of this study is to develop a method for using germline CNV information from large, public databases to predict diseases. We transform CNVs into Chromosomal Scale Length Variation (CSLV) values as a method of reducing dimensional complexity while attempting to retain the global CNV interactions across the entire genome.

Objective 2

The next objective of this study involves studying complex diseases which may have an inherited or genetic risk component. We aim to determine if such diseases have genetic heritability due to CNVs across an individual person's genomic landscape.

Objective 3

Objective 3 of this study is to develop machine learning techniques for the prediction of different types of diseases using CLSVs. These models would predict whether an individual has a higher risk for specific diseases due to inherited factors. Our models will aim to utilize high dimensional and non-linear relationships between hundreds of structural differences instead of the traditional panels of established, disease associated SNPs. A variety of machine learning algorithms exist, many of which originate in statistical and probabilistic theory [19]. We will explore many different options to achieve the best predictive results and compare between models and techniques.

Objective 4

The next objective of this study aims to use the created models to explain how our models come to make their predictions. This objective will focus primarily on explaining how predictions are made, in order to gain biological insight into what CNV regions are most pertinent to a particular disease.

Objective 5

After creation of our machine learning models, we aim to compare our AUCs with existing genetic risk scores which exist for those specific disease cases.

Objective 6

The final objective of our study is to explore options for improving our risk score. To date, such platforms are in their infancy and are often for specific use cases [20–24]. However, in order to have clinical application, consistent performance is needed to ensure a reliable predictive outcome.

Chapter 3: Background of Artificial Intelligence in Genetic Predictions

Copy Number Variations

One of the more interesting findings resulting from NGS technologies was the discovery of extensive genomic structural variations which include deletions, duplications, triplications, insertions, and translocations of sets of base pairs ranging in size between kilobase pairs (Kbp) to megabase pairs (Mbp) [25]. Significant variations in human genomes occur which can overlap potentially thousands of genes—the full scope of which we do not fully understand due to the difficulty of identifying CNVs 1-50 Kbp in size [26]. Oftentimes CNVs, many of which are hereditary in nature, affect genes implicated in complex human diseases [Appendix 1].

Though regional hotspots of CNV deletions and duplications can arise in chromosomes, studies have shown that occurrence of CNV variations across multiple ethnic backgrounds still confounds studies, as CNVs could indicate evolutionarily ancient mutations and also complicate identification of common disease variations [18]. Considering that CNVs are found in all individuals, experiments concerned with CNV in cancerous diseases have still largely focused on rare single region CNVs [27,28]. Studies looking at germline CNVs also tend to focus on rare single region CNVs and most results are identifying single genes with CNVs associated to a particular cancer [29–34].

CNVs in germline blood samples may differ from somatic tumor CNVs. Though germline CNVs and somatic CNVs both may contain inherited information regarding common pathogenic diseases, drawing conclusions and correlations between such CNVs requires caution [35,36]. Evidence suggests acquisition of CNVs later in life, possibly due to environmental factors [34]. We will distinguish somatic CNVs (sCNV) from immediately inherited CNVs which we will call germline CNVs (gCNV). Determination of such differences is still an active area of research and this study will aim to elucidate further evidence with the hope of confirming the findings in previous literature.

Segment Mean, Log2R Values, and Genomic Addresses

Segment means essentially contain the normalized CNV value across a segment of DNA. TCGA uses and defines segment mean as the value of $\log_2\left(\frac{CN}{2}\right)$, where CN=copy number of a specific segment of the genome. TCGA masks patient information and anonymizes the data by calculating segment means for very large portions of a genome, often for almost an entire chromosome. The data is provided as “Masked CNVs” after being processed with BirdSuite[37,38]. The dataset gives a genomic address to each segment mean, which indicates the chromosome number, the start base pair position, and end base pair position. Chromosome Y is not included for anonymization purposes in TCGA.

These log2r values are also what exists in the UK Biobank flat text files. The UK Biobank has roughly 488,000 individuals while the total number of CNV values is 764,257 across the 22 autosomes, an additional 18,857 CNV values for the X chromosome, and 691 CNV values for the Y Chromosome[39,40]. UK Biobank organizes files as space separated text files for each autosome and chromosome X. We did not utilize chromosome Y in our TCGA studies as TCGA does not provide Y chromosome data for anonymization purposes.

Next Generation Sequencing

Next Generation Sequencing (NGS) platforms have rapidly been improving over the past decade [41]. NGS methods and platforms have become an integral tool in genetic research, providing a wealth of information for researchers regarding the structural landscape of genomes. The biggest improvements to these technologies are the speed and efficiency at which they can now sequence human genomes and as a result the decreasing cost many short-read sequencing techniques have. Some sequencing platforms now fall well below $1,000 USD per gigabase pairs (Gbp) and have runtimes of less than a day [5].

Limitations still exist when sequencing cancer genomes. Cancer genomes are very diverse and complex between cancer types as well as between individuals [42]. Studies have demonstrated that whole genome sequencing tends to have a higher error rate as the depth of coverage increases and unique or rare variants have an even higher rate of error (up to 6 %) [43]. High sequence coverage still has an accuracy of over 95 %, but this accuracy is lower than most of the stated values for the platforms indicating possible systematic errors [44]. In Illumina HiSeq® Platforms, which are the current standard for most sequencing techniques, studies indicate that additional data processing can reduce the errors through quality filtering [45]. Therefore, population wide studies are still the gold standard for genetic studies regarding gene variants and identification of risk associated mutations for hereditary diseases. With large sets of data, researchers can utilize statistical methods to further reduce the error of incorrectly concluding that a mutation is an indicator of a complex human disease.

Affymetrix Arrays

TCGA relies on Affymetrix SNP 6.0 array data to identify genomic regions of copy number variations[46]. TCGA relies on these platforms designed by the company Affymetrix which is owned in part by Thermo Fisher Scientific to harmonize and detect copy number variations based off of GRCh38[47].

UK Biobank Axiom Arrays

The UK Biobank relies on the Axiom Array to probe locations of CNV interest. The resulting files are about 2,300 GBs in size or about 2 TB[48,49]. This information includes, normal SNP genotyping data, calls, confidences, intensities, etc. Downloads are done through UK Biobank's Data Showcase which also works closely with the European Genome Archive (EGA). The exact number of genotypes is 488,377 participants[50].

Machine Learning

Machine learning has its roots and beginnings firmly planted in history. Alan Turing's work in cracking the German Enigma machine became the basis for much of modern computer science. The Turing Test, which aims to see if artificial intelligence (AI) has become indistinguishable from human intelligence, is also named after him [51,52]. Machine learning itself, is a subset of AI and was coined in the late 1950s by Arthur Samuel who published a paper on training computers to play checkers when he worked with IBM [53].

By the late 1960s, researchers were already trying to teach computers to play basic games such as tic-tac-toe[54]. Eventually, the idea of neural networks, which were based on a theoretical model of human neuron connection and communication, was expanded into artificial neural networks (ANNs)[55,56]. These foundational works laid dormant for many years due to the impracticality and poor performance of the systems created. Computing technology had not yet advanced enough to reduce the computational time to a practical level.

The modern computer era led to exponential increases in both computational power and data storage capacity. With the introduction of IBM's Deep Blue and Google's AlphaGo in recent decades, several leaps in AI have shown the capacity of AI to solve real world, complex problems[36,57]. As such, the promise of machine learning has taken hold in almost every sector imaginable.

The widespread adoption of machine learning can be mostly attributed to the availability of extremely large datasets and the improvement of computational techniques, which reduce overfitting and improve the generalization of trained models. These two factors have been the driving force to the rapid popularization and adoption of machine learning in almost every field today. This coupled with the increasing prevalence of interconnected devices or the Internet of Things (IoT) has created a rich infrastructure upon which to build predictive and automated systems.

Machine learning is a primary method of understanding the massive influx of health data today. An infrastructure of systems to complement the increasing IoT infrastructure will undoubtedly rely heavily on these techniques. Many use cases have already shown enormous promise. How do these techniques work and how do they give us insight into seemingly unconnected information?

Experts in the field broadly split machine learning into supervised and unsupervised learning. Algorithms falling under both categories implement mathematical models, with each algorithm aiming to give computers the ability to learn how to perform certain tasks.

Supervised Learning

Supervised learning typically employs data known as training data. Training data has one or more inputs and has a “labeled” output. Models use these labeled results to assess themselves during training, with the goal of improving the prediction of new data (i.e., a set of test data)[58]. Typically, supervised learning models focus on classification and regression algorithms [59]. Classification problems are very common in medicine. For example, diagnosing patient involves a doctor classifying the ailment given a certain set of symptoms. The outcome can be an affirmative diagnosis that the patient has the diseases, or a negative diagnosis that the patient does not have the disease. Regression problems tend to look at predicting numerical results like estimated length of stay in a hospital given a certain set of data like vital signs, medical history, and weight. Another example is estimation of

Common algorithms included in this supervised learning group are random forests (RF), decision trees (DT), Naïve Bayes models, linear and logistic regression, and support vector machines (SVM), though neural networks can also be trained through supervised learning[60].

Unsupervised Learning

Unsupervised learning involves presenting a dataset to learning model with no defined or expected outcome and allowing the model to cluster the data due to latent characteristics found in the data itself[61]. The data is considered unlabeled. Unsupervised learning does not typically have a desired outcome but rather returns groups based on latent differences in the dataset. What the grouping is based on is entirely based on how the learning model approaches the structure of the data. Some examples are k-means or k-medoids[62], hierarchical clustering[63], anomaly detection[64], and certain deep learning algorithms.

Algorithms

In this section, we will focus on the main algorithms in question which we will utilize and study for the prediction of complex diseases using our CSLV techniques.

Generalized Linear Model

Generalized Linear Model (GLM) is a set of models including Gaussian regression, Poisson regression, binomial regression for classification, fractional binomial regression, quasibinomial regression, multinomial classification, gamma regression, ordinal regression, negative binomial regression, and Tweedie distribution. These models can be either classification or regression[65–67]. There are several methods for regularization of GLM models to prevent overfitting including ridge regression and least absolute shrinkage and selection operator (LASSO)[68].

The typical method for finding optimal regularization for GLM models is to perform grid searches over two parameters known as α and λ. The α parameter controls the distribution between LASSO and ridge regression where α of 1.0 represents LASSO and α of 0.0 represents ridge regression. The parameter known as λ controls the amount of applied regularization where λ of 0.0 means no regularization is applied to the GLM model at all[68]. Data does not to be sorted for these models nor does it need to perform any special handling of imbalanced data.

Distributed Random Forest

Random forests are a form of decision trees but are an ensemble set of independently trained decision trees. The resulting predictions of the trees are typically averaged to get a better end result and prediction[63]. Each tree is built by using a random sample of the data with replacement and at each candidate split a random subset of features are also selected. This prevents each learner or tree from focusing too much on apparently predictive features of the training set which may not be predictive on new data. In other words, it increases generalization of the model. Random forests can have hundreds or even thousands of trees and work fairly well on noisy data[69]. The model created from aggregating results from multiple trees trained on the data will give a prediction that can be assessed using test data.

Gradient Boosting Machine and CART Models

We decided to explore Classification and Regression Tree (CART) models—more commonly known as Decision Trees. Decision Trees provide an easily interpretable model after training and have the same predictive power on our disease data, providing some insight into distinguishing features between the populations in the dataset [70]. In many ways ANNs are essentially intertwined decision trees, organized in layers to mimic a biological "neuron" or perceptron as the machine learning field calls them. ANNs still benefit disease genomics research [7,71]. However, we believe interpretability is valuable as the focus of our studies is exploring the differences in inherited genomic landscape between diseases.

The topic of gradient descent and boosting, or gradient boosting, is foundational to modern implementations of CARTs. Gradient Boosting Machines or Gradient Boosted Decision Trees (GBDT) utilize a series of weak classification trees and aggregates their results to form a strong learner. Such models aim to minimize some error function and gradually steps towards best fit, thus descending to the minimum error in mathematical terms and thus tend to handle imbalanced classes better than ANNs [72–75]. Instead of building trees in parallel, the algorithm utilizes the error of prior trees in creating the next tree, correcting the errors of its predecessors. Thus, a given model contains residuals which act as negative gradients of the squared loss function. The drawback of using GBDT models lies in the danger of overfitting to a single dataset. Thus, we will validate models through cross-validation, leave out testing, and other methods such as verifying results on a different dataset. Typically, achieving good results requires some parameter tuning.

XGBoost

XGBoost is a supervised learning algorithm which implements boosting to create an ensemble of parallel trees based on GBM. It is perhaps one of the most useful and best gradient boosting frameworks currently in existence. A significant improvement is that GPU support is available for XGBoost using NVIDIA GPUs provided the system is a Linux system with CUDA 9[76]. XGBoost utilizes two separate modules. The first of which contains binary libraries for each platform and different configurations and tries to load the most powerful library first. This is typically one with GPU and OpenMP support. If it fails it will proceed to attempt the next in the list, with the final fallback being a minimal configuration of single CPU support. The second module contains the XGBoost model and model builder code[77].

Deep Learning

Though there are many different implementations of a deep learning artificial neural net, the most common is a feedforward ANN. This is trained with stochastic gradient descent using back-propagation. The network typically contains many perceptrons organized into many hidden layers. Each perceptron has a *tanh*, rectifier, or some other max-out activation function[78].

It is typically important for deep learning to shuffle training data since each row is fed in sequentially during training. The input layer is also scaled to the number of columns and this is typically an indication of the model's complexity. Backpropagation and loss function assessment occurs after each training sample. An epoch is one pass over the entire dataset. Typically, the default number of epochs is 10.

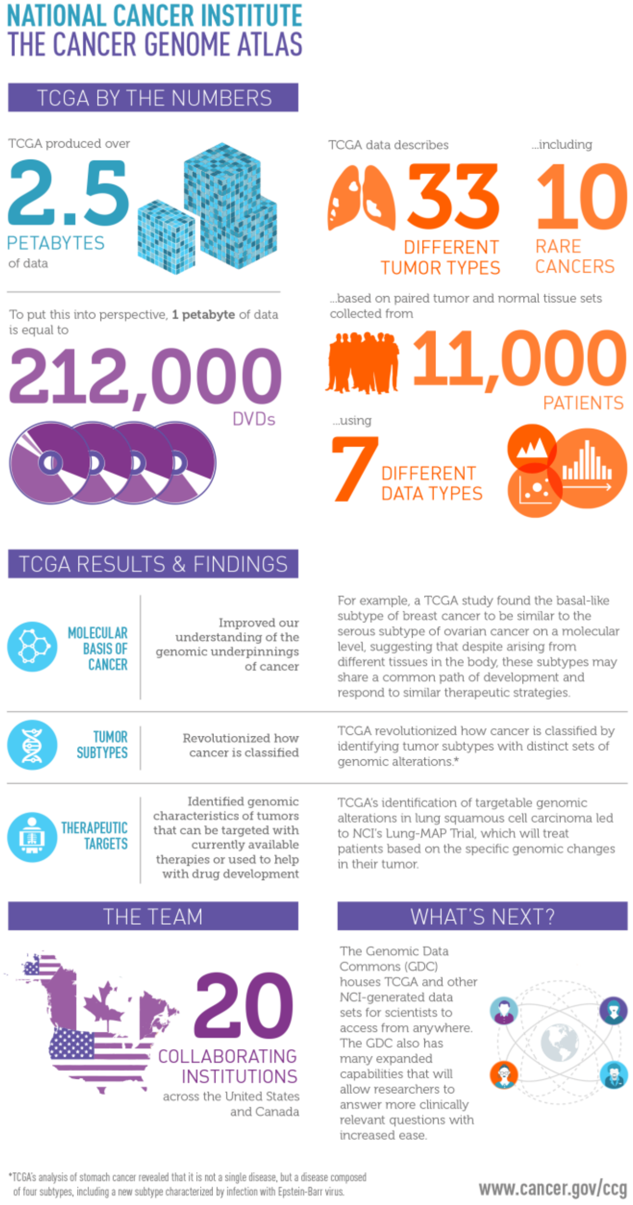
Though many researchers have emphasized ANNs recently, many shortcomings exist when using ANNs on disease data. Specifically, the actual ability to predict disease is only marginally beneficial. The simple ability for a computer to accurately predict outputs from data does not necessarily tell us *how* the models make these predictions. The real insight lies in how models arrive at conclusions using the data. The downside of ANNs lies in the difficulty of interpreting model decisions. Additionally, ANNs do not perform much better than other machine learning models unless the data used is complete (i.e. no data sparsity), unlabeled, and approaches observations in the hundreds of thousands if not more [79–81]. The datasets we utilize contain prelabeled data where physicians have already diagnosed the patients.

Stacked Ensemble

The principle behind stacked ensembles is to use multiple machine learning algorithms to improve the overall predictive performance of the overall model. This technique utilizes stacking, otherwise known as super learning or stacked regression[82,83]. Primarily a "meta-learner" is trained to find the optimal combination of the base algorithms to create a diverse set of learners that work well together. First a list of base algorithms (e.g., GBM, GLM, deep learning, XGBoost, etc.) are chosen, followed by a second-tier algorithm for the meta-learning. The second-tier algorithm can be the same as one of the base algorithms. Then each base algorithm is trained on the training set, followed by a k-fold cross-validation on each of these learners. Then the cross-validated prediction values are taken from each of the base algorithms. This new dataset is then used to train the meta-learning algorithm which forms the "ensemble mode”. To predict on new data we first generate predictions from the base learners, then feed those predictions into the meta-learner which generates the ensemble predictions on the new data[84].

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) was an NIH funded cancer genomic program that focused on the molecular characterization of primary cancer and matched normal samples, spanning 33 different cancer types [85]. The project began in 2006 and was a joint effort between the National Cancer Institute (NCI) and the National Human Genome Research Initiative (NHGRI)[85]. After 12 years, the official project has ended but the publicly available data still contains a large amount of unexplored potential for the field of cancer research.


NATIONAL CANCER INSTITUTE
THE CANCER GENOME ATLAS
TCGA BY THE NUMBERS
TCGA produced over
2.5
PETABYTES
of data
To put this into perspective, 1 petabyte of data is equal to
212,000
DVDs
TCGA data describes
33
DIFFERENT TUMOR TYPES
...including
10
RARE CANCERS
...based on paired tumor and normal tissue sets collected from
11,000
PATIENTS
...using
7
DIFFERENT DATA TYPES
TCGA RESULTS & FINDINGS
MOLECULAR BASIS OF CANCER
Improved our understanding of the genomic underpinnings of cancer

For example, a TCGA study found the basal-like subtype of breast cancer to be similar to the serous subtype of ovarian cancer on a molecular level, suggesting that despite arising from different tissues in the body, these subtypes may share a common path of development and respond to similar therapeutic strategies.

TUMOR SUBTYPES
Revolutionized how cancer is classified

TCGA revolutionized how cancer is classified by identifying tumor subtypes with distinct sets of genomic alterations.*

THERAPEUTIC TARGETS
Identified genomic characteristics of tumors that can be targeted with currently available therapies or used to help with drug development

TCGA's identification of targetable genomic alterations in lung squamous cell carcinoma led to NCI's Lung-MAP Trial, which will treat patients based on the specific genomic changes in their tumor.

THE TEAM
20
COLLABORATING INSTITUTIONS
across the United States and Canada
WHAT'S NEXT?

The Genomic Data Commons (GDC) houses TCGA and other NCI-generated data sets for scientists to access from anywhere. The GDC also has many expanded capabilities that will allow researchers to answer more clinically relevant questions with increased ease.

*TCGA's analysis of stomach cancer revealed that it is not a single disease, but a disease composed of four subtypes, including a new subtype characterized by infection with Epstein-Barr virus.

www.cancer.gov/ccg

Figure 1: Infographic of The Cancer Genome Atlas Project

The project contains over 11,000 patient samples from 33 different cancer types. The database is over 2.5 petabytes and holds both somatic and normal tissue samples along with clinical data associated with the genetic information.

The dataset contains over 11,000 patients' somatic tissue samples (Fig. 1 [86]) from 33 tumor types as well as healthy samples and includes mRNA expression, somatic mutations, DNA methylation, and copy number variation information [85,86]. Additionally, the data set can utilize modern Cloud Server providers such as the Google Cloud Platform™. Cloud servers allow for partial processing in the

Cloud via platforms like Google BigQuery™, making handling of large amounts of data practical and more manageable.

Recently, the project concluded by releasing the Pan-Cancer Atlas and a collection of flagship papers through *Cell Press Journals*, laying the groundwork for classification of molecular differences between the 33 different types of cancers [87]. Most studies utilized an integrated approach to identify subgroups of cancers and genetic differences while others utilized machine learning to identify potential anti-cancer drug targets [88,89]. For our study, we are looking to improve on past studies which examined pathogenic germline variants in cancer as germline copy number variations hold great potential to provide insight regarding the complexity of hereditary risk [30]. Overall, as one of the first large scale cancer genetic datasets, The Cancer Genome Atlas continues to provide a wealth of new insights past its official completion.

The Cancer Genome Atlas Pipeline

The Cancer Genome Atlas was a multi-year, multi-institution program funded by the National Cancer Institute. The project systematically gathered genetic data through a quality-controlled pipeline. TCGA pipeline utilized Affymetrix SNP 6.0 array data for CNVs, sequencing and identifying inferred CNVs. Somatic tumor tissue and blood samples were both taken from patients and processed through the TCGA pipeline via Birdsuite, an open source set of tools created by the Broad Institute [37,90]. The processed SNP microarrays make up much of the copy number information in the TCGA database. Additional information such as clinical information and molecular characterization for biospecimen samples are also available through the Genomic Data Commons (GDC) which is the main storage location for the data [91].

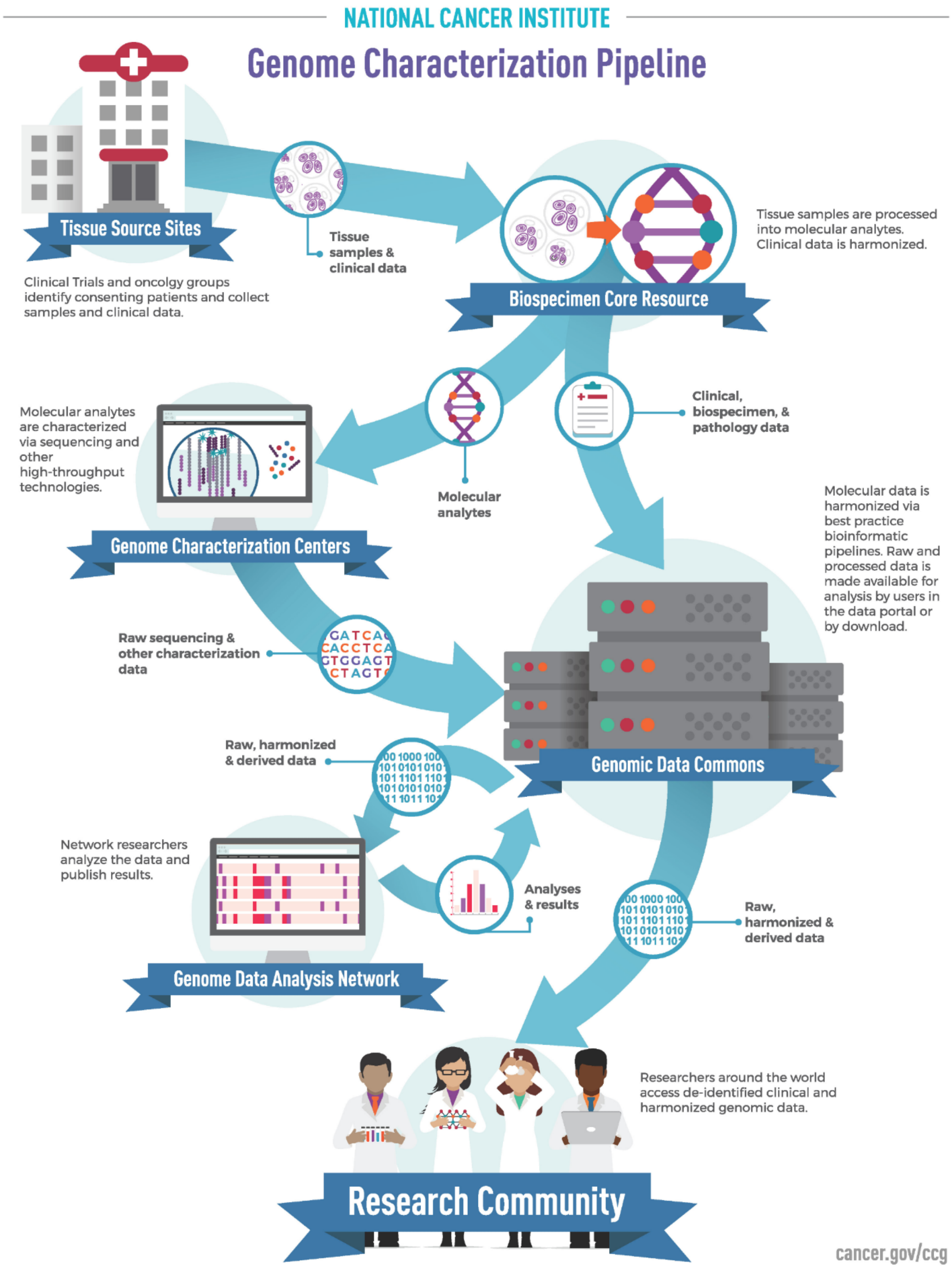

NATIONAL CANCER INSTITUTE
Genome Characterization Pipeline
Tissue Source Sites
Clinical Trials and oncolgy groups identify consenting patients and collect samples and clinical data.
Tissue samples & clinical data
Biospecimen Core Resource
Tissue samples are processed into molecular analytes. Clinical data is harmonized.
Molecular analytes are characterized via sequencing and other high-throughput technologies.
Molecular analytes
Clinical, biospecimen, & pathology data
Genome Characterization Centers
Molecular data is harmonized via best practice bioinformatic pipelines. Raw and processed data is made available for analysis by users in the data portal or by download.
Raw sequencing & other characterization data
Genomic Data Commons
Raw, harmonized & derived data
Network researchers analyze the data and publish results.
Analyses & results
Raw, harmonized & derived data
Genome Data Analysis Network
Researchers around the world access de-identified clinical and harmonized genomic data.
Research Community
cancer.gov/ccg

Figure 2: Infographic of TCGA Pipeline

The TCGA Genome Characterization Pipeline processed all samples, regardless of cancer type, in this manner. The raw genome data resides in the public database along with data processed by the Genome Data Analysis Network.

The NCI's Center for Cancer Genomics (CCG) organized the project, coordinating U.S. and Canadian research teams to produce the clinical datasets that would become TCGA. This standardized workflow began with clinical trials in oncology groups and involved collection of tissue samples, usually from both tumors and blood, taken from patients who chose to participate (Figure 2 [92]). Clinics formalin-fixed samples in paraffin though some tissues were frozen. The CCG's Biospecimen Core Resource (BCR) received the fixed or frozen samples. The Biospecimen Processing Center at Nationwide Children's Hospital is the first component of the BCR and processed all tissues to ensure rigorous quality standards. The Clinical Data Center at Information Management Services, Inc. (IMS), the second component of BCR, oversaw informed consent and anonymized the data to protect patient privacy and clinical data provided with the samples [93].

Next, Genomic Characterization Centers (GCCs) generate data from the DNA, RNA, and proteins they receive from the BCR. There are three GCCs: The Broad Institute, The University of North Carolina, and MD Anderson Cancer Center. The Broad Institute specialized in DNA, whole genome, and whole exome sequencing. The University of North Carolina performed the RNA sequencing. MD Anderson Cancer Center performed reverse phase protein arrays (RPPAs). After sequencing and arrays were performed by the GCCs, the raw sequencing and associated metadata were sent to the GDC, who shared the data with the Genomic Data Analysis Network (GDAN) and research community [93].

GDAN, a team of scientists from 13 institutions, used the raw results and genomic characterization techniques to gain biological insights before publishing results in scientific journals. These Analysis Working Groups (AWGs) produced novel analyses and the GDC harmonized the information and characterization data, making it publicly available for other researches to use across the world [93].

The UK Biobank Project

As one of the most ambitious modern genetic projects, the UK Biobank was established by the Wellcome Trust medical charity, the UK Medical Research Council, the UK Department of Health & Social Care, the Scottish Government, the Northwest Regional Development Agency, with funding from the Welsh Government, British Heart Foundation, Cancer Research UK, and Diabetes UK [94,95]. The UK National Health Service (NHS) is the primary supporter of the project. Beginning in 2006, the UK Biobank recruited 500,000 people between the ages of 40-59 years with the plan to follow them and record health outcomes over their lifetimes [40,95–97]. The database went live in 2017 with no preferential access, meaning any researcher from any institute can pay the fee for the ~12 Petabytes of data and begin doing analysis on it. It contains genetic data and a host of other data such as imaging and exercise data for subsets of the individuals. The UK's National Health Service integrated the project into the healthcare system which aided in performing regular follow ups on a portion of the cohort, providing a large range of health information over time. Thus far, there have been approximately 40,000 incident cancers, ~14,000 deaths, and 1.3 million hospitalizations recorded [98].

UK Biobank Project Genotyping

GlaxoSmithKline and Regeneron are performing exome sequencing on the samples from all 500,000 participants with the first set of data becoming available as of early 2019. Vacutainers at 4 °C store initial blood and urine samples collected from patients. A central processing center collects all the samples storing them in liquid nitrogen. The Cheadle or

Wythenshawe centers processed samples and genotyped them. All samples were then archived in -80 °C[95]. Genomic assays of 820,967 SNPs were conducted on these blood samples with results and data published by 2018 [50,99,100]. Genome-wide genotyping and imputation was performed by the Big Data Institute of Oxford University [101]. About 440,000 were genotyped on the UK Biobank Axiom® Array and 50,000 were genotyped on the Affymetrix UK BiLEVE Axiom® Array with >95% overlap with the previous group [48,99,102]. Additional information can be found here: https://www.ukbiobank.ac.uk/ [103,104].

Chapter 4: Research Design and Methods

Data Acquisition

TCGA Data

The relevant CNV data and associated clinical data are stored on Google BigQuery™, which can return query tables using Structured Query Language (SQL). TCGA is hosted in these Big Query tables publicly. Cloud servers performed the bulk of the original file manipulation, reducing computational time and overhead on the local machines. We used the statistical programming language R to perform the bulk of our own specific data analysis. Using the R package (i.e. “bigrquery”), we downloaded the data tables and completed the remaining manipulation to format the data into an arrangement which can be used in training models [105].

UK Biobank Data

UK Biobank data is provided only to approved researchers through the data showcase: https://biobank.ndph.ox.ac.uk/showcase/. You need to apply to UK Biobank for access. There is about a $500 application fee. Then once the application was approved another 2000 GBP in user fees. The process took several months until we could access data.

When applying for access, we specified what patient data we needed access to. We were looking for cancer data and schizophrenia data, so we specified that. But if we are looking for Alzheimer’s or other diseases, we would choose the appropriate categories listed here: https://biobank.ndph.ox.ac.uk/showcase/browse.cgi.

Once we are granted access, we were sent a key via e-mail. We store the key in our working directory in a file called ".ukbkey" make sure its readable ("chmod 755 .ukbkey"). We can’t download and decrypt data without the key.

They also gave us access to an encrypted file custom made for us called something like ukbXXXXX, where the X's are integers. This file contains *patientIDs* and the variables we chose in our application. So, for us the file has about 500k lines. Each line is *patientID*, sex, and a bunch of information on the patient's such as cancer type.

Now we need to download all the l2r genetic data. Details are here https://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/ukbgene_instruct.html.

First thing we need to get ukbgene. It's the main tool used for downloading. Instructions are here:

https://biobank.ndph.ox.ac.uk/showcase/download.cgi

or just grab it with wget (on Linux)

```
$ wget  -nd  biobank.ndph.ox.ac.uk/showcase/util/ukbgene
```

ukbgene is a Linux executable. We can run it in under windows in a wsl terminal as well (https://docs.microsoft.com/en-us/windows/wsl/install-win10)

Run this command:

```
$ ./ukbgene l2r -c1 -m
```

to download a small (16 MB) file. Rename the file you download as "patientIDSall.fam". This is for chromosome 1 (that's the -c1 argument), but all the chromosomes are the same since we are just pulling patient IDs. So, you only need this one. This patientIDSall.fam file is a list of patientIDs, one per row. The order is important because these are the column headings for the l2r data we'll get next.

Now we can download the data. Use a shell script which executes "`ukbgene l2r -cN`" where N is 1 to 22, the chromosome number to quickly run the command for each chromosome. This will probably take the better part of a month to download. Count on at least two weeks. It downloads about 2.3 terabytes while also performing error checking. The smallest file, chromosome 21, is 34.8 GB. Largest, chromosome 1, is 195 GB. ukbgene occasionally fails and we would have to restart it. We needed to check on it each day. (On a 100 Mbps connection, it should only take a few days to download 2.3 terabytes).

The files are all called something like "ukb_l2r_chrN_v2". Each file is plain text, no headers, just numbers separated by spaces. The file is organized as one column per patient, the patient's ID is given by the patientIDSall.fam file. Each row represents the log2ratio measured at a different SNP location in their array. All data was downloaded in this matter.

Data Processing

TCGA Data

One formatted row in the table holds a data point or observation (i.e., one patient). Each data point contains a column for case-barcode, a unique anonymized identifier for the patient. We unified genomic addresses or molecular locations into a single column. This column contained information about the chromosome locations, including start and end positions (by base pair number) of the gene segment in question. Each column, defined by a molecular address, indicates a chromosome segment, and contains a segment mean. For the masked data, there would be 30 columns as we took the top 30 masked CNVs, while for the unmasked models there would be 100 columns as we took the top 100 unmasked CNVs for those models. The entries for each patient in these columns contain the actual CNV data and could be blank if there is no information found (likely indicating normal CNV). Lastly, a column indicating the TCGA Study the sample is from provides the information for the type of cancer the patient had. Additionally, we can include other columns of information such as gender, age, and ethnicity. For our models we chose to include gender but found little difference in model performance if we excluded gender.

UK Biobank Data

To process the UK Biobank data, you can utilize the script found here: https://github.com/cetoh/brodylab/blob/tohc/ukbiobank/data_handling/condenseSplitCNVs.R to create the CSLV data files. You will have to use your own file paths.

Chromosomal Scale Length Variation

Chromosomal Scale Length Variation is an average of large segments of CNVs across a particular chromosome. As such calculation of these values is based on the desired number of even splits between the values. For example, a chromosome may have something like 11,535 values. In order to calculate “4 splits” an average of 2,884 sequential values would be averaged to form one value for the first 3 splits and then the last split would be the remaining 2,883 sequential values. Since it very difficult to load the entire file of any chromosome into memory, this process is done by reading in each row line by line until reaching a split, averaging, and then clearing the memory. As such, any number of splits can be specified. The mathematical calculation is as follows:

$$\frac{\sum_{i=0}^{n} \log_2 \frac{CN_i}{2}}{n}$$

Where CN represents a copy number value normalized by dividing by 2 for each allele and then taking the base 2 log of that normalization. The $\log_2 \frac{CN_i}{2}$ value represents a single l2r value. These values are then averaged over the split size given a certain n number of l2r values.

Cloud Computing Server Specifications

We created a computing server running Linux Ubuntu 20.04 (64-bit) LTS as the operating system. The server additionally has 2 Intel Xeon E5-2960 2.90 GHz CPUs and one NVIDIA GeForce GT 710 GPU (2GB GDDR3). There is 32GB of RAM available (DDR3 2,133 MHz) with a 10 TB HDD. We also created a 64 GB swap for additional memory on the hard disk.

We also duplicated our findings on our collaborators compute server. We would like to thank Dr. Timothy Downing for allowing us to use their server resources. The specification for this server is Linux Ubuntu 20.04 (64-bit) LTS. It has an AMD Threadripper 3990X CPU (2,200 min MHz, 2,900 max MHz) and an NVIDIA GeForce RTX 3090 GPU (24 GB 0f G6X). There is 64 GB of available RAM (DDR4 3200 MHz). The server also has a 1.6 TB NVME SSD, and two 13 TB HDDs.

R Statistical Programming Language Specifications

The initial work for TCGA was done in R v3.5.1. Subsequent work was performed in R v3.6.3. Instructions for installation are found here: https://cran.r-project.org/bin/linux/ubuntu/README.html.

H2O Machine Learning Specifications

We utilize the leading distributed machine learning platform H2O to train, test, and validate our models. This open-source software is a distributed in-memory machine learning platform (built in Java) which also has a corresponding R package, allowing for seamless model building, data analysis and reproducibility in R while maintain our ability to run SQL queries through Google BigQuery™. Additionally, since the platform utilizes the leading industry algorithms for a variety of machine learning models, it allows us to quickly compare between other algorithms. There are extensive hyper-parameter options which allow us to tune models and training parameters to prevent issues such as overfitting and overcome data sparsity issues [106].

Generalized Linear Model (GLM) with H2O

Following the definitive text by P. McCullagh and J.A. Nelder (1989)[107] on the generalization of linear models to non-linear distributions of the response variable Y, H2O fits GLM models based on the maximum likelihood estimation via iteratively reweighed least squares[108,109].

Let $y_1, \ldots, y_n$ be n observations of the independent, random response variable Y_i.

Assume that the observations are distributed according to a function from the exponential family and have a probability density function of the form:

$$f(y_i) = exp[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i; \phi)]$$

where θ and ϕ are location and scale parameters, and $a_i(\phi)$, $b_i(\theta i)$, and $c_i(y_i; \phi)$ are known functions.

a_i is of the form $a_i = \frac{\phi}{p_i}$ where p_i is a known prior weight.

When Y has a probability distribution function from the exponential family:

$$E(Y_i) = \mu_i = b^{'}var(Y_i) = \sigma_i^2 = b''(\theta_i)a_i(\phi)$$

Let $g(\mu_i) = \eta_i$ be a monotonic, differentiable transformation of the expected value of y_i. The function η_i is the link function and follows a linear model.

$$g(\mu_i) = \eta_i = \boldsymbol{x}'\boldsymbol{i}\beta$$

When inverted: $\mu = g^{-1}(\boldsymbol{x}_i^{'}\beta)$

Maximum Likelihood Estimation

For an initial rough estimate of the parameters $\hat{\beta}$, use the estimate to generate fitted values:

$$\mu_i = g^{-1}(\widehat{\eta}_i)$$

Let z be a working dependent variable such that $z_i = \widehat{\eta}_i + (y_i - \widehat{\mu_i})\left(\frac{d\eta_i}{d\mu_i}\right)$, where $\frac{d\eta_i}{d\mu_i}$ is the derivative of the link function evaluated at the trial estimate.

Calculate the iterative weights: $w_i = \frac{p_i}{\left[b''(\theta_i)\left(\frac{d\eta_i}{d\mu_i}\right)^2\right]}$

where b'' is the second derivative of $b(\theta_i)$ evaluated at the trial estimate.

Assume $a_i(\phi)$ is of the form $\frac{\phi}{p_i}$. The weight w_i is inversely proportional to the variance of the working dependent variable z_i for current parameter estimates and proportionality factor ϕ. Regress z_i on the predictors x_i using the weights w_i to obtain new estimates of β.

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z}$$

where $\boldsymbol{X}$ is the model matrix, $\boldsymbol{W}$ is a diagonal matrix of w_i, and $\boldsymbol{z}$ is a vector of the working response variable z_i.

This process is repeated until the estimates $\hat{\beta}$ change by less than the specified amount.

Distributed Random Forest (DRF) with H2O

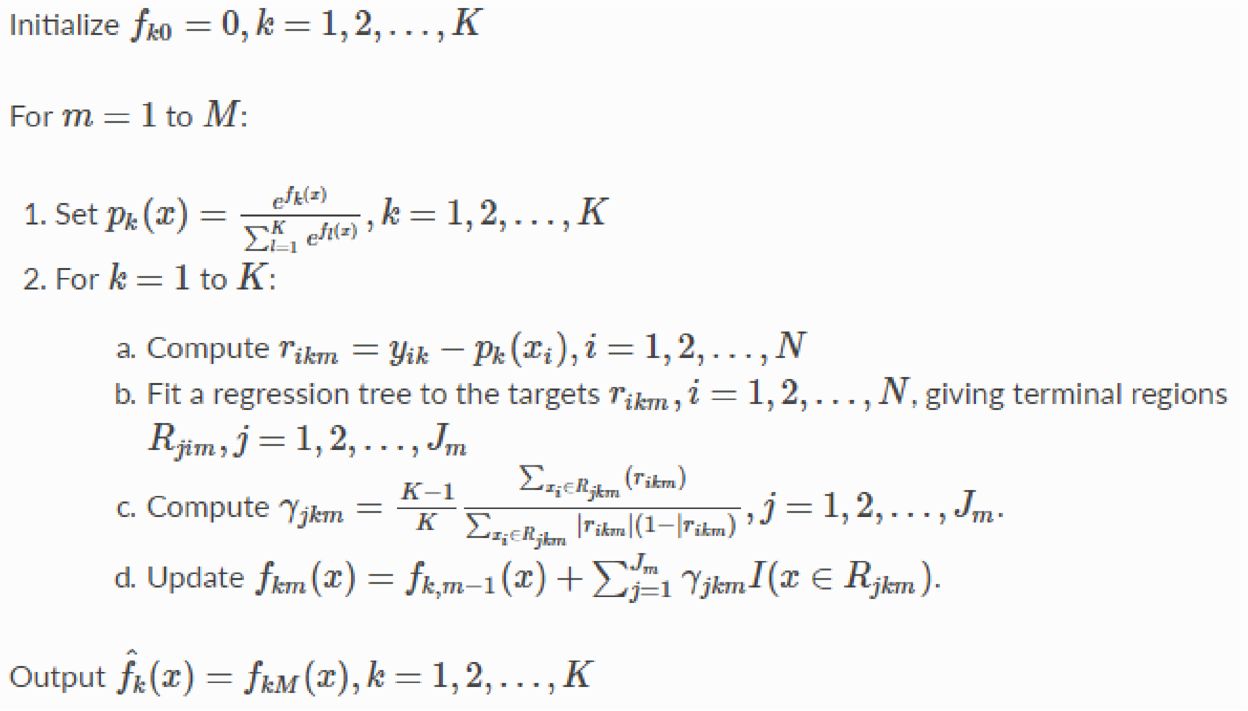
H2O utilizes distributed random forests as a powerful classification and regression tool. This is done by building a set of classifications trees rather than a single classification or regression tree. By increasing the number of trees which are trained as weak learners and taking the average over all the trees' predictions, the model reduces variance. Each node on the computation machine builds a subset of the forest in a parallel manner. Tree building and growth is stopped randomly by several stopping metrics, either tree depth or number of leaves or nodes[110]. The algorithm is like GBM except that the weak learners and trees are built independently without any input from the other trees in the model.

Gradient Boosted Machine (GBM) with H2O

There are extensive hyper-parameter options which allow us to tune GBM models and training parameters to prevent issues such as overfitting and overcome data sparsity issues [106]. GBM is typically one of the best performing algorithms we have utilized in H2O.

Gradient boosting uses these trees as weak learners creating them in an iterative fashion to achieve a single strong learner. From a general sense, the goal of boosting is to teach a model F to predict values of $\hat{y}$ and minimizing some error function such as the mean squared error $\frac{1}{n}\sum_i(\hat{y}_i - y_i)^2$. The algorithm indexes (i) over some training set which is n observations large and compares the squared difference between the actual values of y and predicted values $\hat{y}$. This then provides an imperfect model F_m and to improve on this model there must exist some estimator h which provides a model $F_{m+1}(x) = F_m(x) + h(x)$ which is a better solution. A good h would imply that the model correctly predicted y, thus $F_{m+1}(x) = F_m(x) + h(x) = y$ and in order to find h it follows that $h(x) = y - F_m(x)$ [75]. This provides a residual function which is in fact a negative gradient. From there the algorithm attempts to minimize the loss function for the model which is often denoted as γ, applying the steepest descent step to the minimization problem [74].

H2O builds sequential regression trees following the algorithm specified by Hastie et al. [111]. In the H2O algorithm h is the residuals denoted as r_{ikm} which are in fact gradient values for each of the N bins in the CART model. H2O defines the algorithm as follows:



```
Initialize $f_{k0} = 0, k = 1, 2, \dots, K$

For $m = 1$ to $M$:

  1. Set $p_k(x) = \frac{e^{f_k(x)}}{\sum_{l=1}^{K} e^{f_l(x)}}, k = 1, 2, \dots, K$
  2. For $k = 1$ to $K$:

        a. Compute $r_{ikm} = y_{ik} - p_k(x_i), i = 1, 2, \dots, N$
        b. Fit a regression tree to the targets $r_{ikm}, i = 1, 2, \dots, N$, giving terminal regions
           $R_{jim}, j = 1, 2, \dots, J_m$
        c. Compute $\gamma_{jkm} = \frac{K-1}{K} \frac{\sum_{x_i \in R_{jkm}} (r_{ikm})}{\sum_{x_i \in R_{jkm}} |r_{ikm}|(1-|r_{ikm}|)}, j = 1, 2, \dots, J_m.$
        d. Update $f_{km}(x) = f_{k,m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jkm} I(x \in R_{jkm}).$

Output $\hat{f}_k(x) = f_{kM}(x), k = 1, 2, \dots, K$
```

A conceptual way to view this is that each tree built feeds its result and errors into the building of the next tree so that the newly created tree can benefit from the failings of the previous trees in an iterative fashion [72,73]. This is different than the Distributed Random Forest (DRF) algorithm which also uses decision trees but which instead builds all of them in parallel at the same time, typically much deeper, and keeping the trees in isolation [70].

XGBoost with H2O

The H2O implementation of XGBoost is a supervised learning algorithm which employs boosting as an ensemble technique for improving decision tree models[77]. Each new model attempts to correct deficiencies in the previous model. XGBoost allows for parallel tree boosting that solves many problems in a fast and accurate way.

H2O uses two separated modules, h2o-genmodel-ext-xboost which extends h2ogenmodel and registers an XGBoost-specific Maven plain old Java Object (MOJO)[112]. The module contains multiple libraires for each platform to support a variety of different configurations, including with or without GPU and with or without OMP. H2O will always try to load the most powerful library first, before continuing down the list to the final single CPU minimal configuration.

The second module, h2o-ext-xgboost, contains the XGBoost model and model builder. XGBoost supports multicore implementations and GPU acceleration[76]. Making it much quicker in training the models. In order to use GPU acceleration, you must have an NVIDIA GPU which supports CUDA 9+. XGBoost is also not supported on Windows and OMP and GPU boost is not supported on Windows or Mac OS X.

Deep Learning Neural Networks with H2O

H2O implements a feedforward ANN that is trained stochastic gradient descent using back propagation. It allows for customization of the number and size of hidden layers with a minimum of one hidden layer. Activation of hidden layer perceptrons can be one of several activation functions including tanh, rectifier, and max-out functions. Additionally, H2O trains multiple copies of the same model in parallel through multi-threading in an asynchronous manner[78]. Performance is then periodically averaged across a global network across the entire network copies.

Default settings for deep learning in H2O which typically perform the best have two hidden layers of size 200 each and a stopping metric of log loss for classification. Because training id done in order, it is recommended by H2O to shuffle the training data before training. The input layer will automatically scale to the number of input features or columns for the given dataset as well. As such, if we wish to reduce complexity, we must do so prior to feeding the training data into the neural network.

Stacked Ensembles with H2O

The steps below describe the individual tasks involved in training and testing a Super Learner ensemble. H2O automates most of the steps below so that you can quickly and easily build ensembles of H2O models[113].

1. Set up the ensemble.
    1. Specify a list of L base algorithms (with a specific set of model parameters).
    2. Specify a metalearning algorithm.
2. Train the ensemble.
    1. Train each of the L base algorithms on the training set.
    2. Perform k-fold cross-validation on each of these learners and collect the cross-validated predicted values from each of the L algorithms.
    3. The N cross-validated predicted values from each of the L algorithms can be combined to form a new N x L matrix. This matrix, along with the original response vector, is called the “level-one” data. (N = number of rows in the training set.)

    4. Train the metalearning algorithm on the level-one data. The “ensemble model” consists of the L base learning models and the metalearning model, which can then be used to generate predictions on a test set.

3. Predict on new data.

    1. To generate ensemble predictions, first generate predictions from the base learners.
    2. Feed those predictions into the metalearner to generate the ensemble prediction.

All base models must have the same number of folds if cross-validated. In our case we did 5-fold cross validation for all models. All predictions from the cross-validated predictions must be saved as this data is used train the metalearner. AutoML trains these models via a grid search. Base models must be trained on the same training data. A minimum of two base learners is required.

Additional R Libraries

All R packages are available on CRAN and can be found here https://cran.r-project.org/web/packages/available_packages_by_name.html [114].

ukbtools

This is a set of R tools to visualize primary dataset from UKB file sets (.tab, .r, .html) and query ICD diagnoses, retrieve genetic metadata, read and write standard formats for genetic analyses[115].

tidyverse

The 'tidyverse' is a set of packages that work in harmony because they share common data representations and 'API' design. This package is designed to make it easy to install and load multiple 'tidyverse' packages in a single step. Learn more about the 'tidyverse' at https://tidyverse.org[116].

dplyr

A fast, consistent tool for working with data frame like objects, both in memory and out of memory, this package is built primarily for data manipulation in data frame or table objects[117].

ggplot2

A system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details. Maintained by the same authors as tidyverse at https://ggplot2.tidyverse.org[118].

ggthemes

Some extra themes, geoms, and scales for 'ggplot2'. Provides 'ggplot2' themes and scales that replicate the look of plots by Edward Tufte, Stephen Few, 'Fivethirtyeight', 'The Economist', 'Stata', 'Excel', and 'The Wall Street Journal', among others. Provides 'geoms' for Tufte's box plot and range frame. Found at https://github.com/jrnold/ggthemes [119].

ggsci

A collection of 'ggplot2' color palettes inspired by plots in scientific journals, data visualization libraries, science fiction movies, and TV shows. Found at https://github.com/road2stat/ggsci.

GitHub Repository and Version Control

All code related work is available at the public GitHub Repository: https://github.com/cetoh/brodylab. This code and work will also track all commits and versions of code. Work for this specific dissertation is on the *"tohc"* branch of the repository. The repository is integrated with the cloud computing server and our R Studio Server setup.

Training Gradient Boosting Decision Trees on TCGA Germline Data

To perform model training we utilized data from all TCGA studies and checked for overfitting through ten-fold cross validation for each model. We used genetic data (gCNV) from blood samples only to train our models. We also excluded Acute Myeloid Leukemia (LAML) and Chronic Myelogenous Leukemia (LCML) as these are cancers derived from hematopoietic stem cells and thus blood samples from these groups would skew models for other cancers. For each cancer, we kept them labeled with their cancer short code and set the samples from all other samples as "Normal." Then we used the data to train a GBDT model with 50 trees, did not specify a max depth, and used balanced classes. We performed ten-fold cross validation to achieve a given AUC for the cross-validated results. We created five models in this manner for each cancer type. Finally, we averaged the 5 sets of trials for the 32 types of cancers to get the general performance of a GBDT model created for that cancer type.

Testing TCGA Germline Models on Somatic Samples

We utilized a similar process as before to train a set of GBDT models that would predict on sCNV samples. The main difference in this experiment lies in only using 80% of the patients to train the initial models with gCNV data. We also performed ten-fold cross-validation on these models to verify that that with 80% of the patients the performance was comparable to using 100%. Using the sCNV information of the remaining 20%, we attempted to predict cancer diagnoses using a model trained on the blood samples only. For each of the 5 models per cancer, we acquired an AUC and averaged to get the general performance of GBDT models predicting on somatic samples but trained with blood samples.

Validating TCGA Results on UK Biobank

To check that our results are not due to artifacts within the TCGA data itself, we plan to test our methods and models on a separate database known as the UK Biobank. The methodology will be as follows:

1. Acquire l2r data from UK Biobank for cancers of the same type as ones tested in TCGA
2. Calculate CSLV values for cancer patients and normal non-cancer patients in UK Biobank
3. Train a model to predict between cancer patients and non-cancer patients
4. Assess if resulting model predictions and performance is comparable with the results achieved on TCGA data.

Chapter 5: Results

Chapter 5.1: Prediction and Classification of Cancer Diagnosis between Cancers

Initial results indicate that significant inherited differences exist between cancer types and that predicting cancer from germline copy number variations alone is possible[120]. Using the TCGA public database of 32 different cancer types, we constructed GBDT Models in an iterative fashion. In total, we created 10 models (n=10) for each cancer. We omitted Acute Myeloid Leukemia (LAML) and Chronic Myelogenous Leukemia (LCML) as the germline information utilizes data taken from peripheral blood samples and both cancers are blood related cancers. Somatic samples are samples taken directly from the tumor itself

We utilized masked CNV data and organized the data to indicate a segment mean value for the top 30 CNV segments. This masked data is the segment mean of a segment of DNA. These segment means are normalized averages of very large segments of chromosomes and in some case is the mean for almost the entire chromosome. Certain cancer types, such as Ovarian Cancer (OV) and Glioblastoma Multiforme (GBM), performed very well with an area-under the curve (AUC) of over 0.80. Likewise, most cancers performed better than chance (Table 1). Kidney Chromophobe (KICH), most likely performed poorly due to insufficient sample size (n = 9).

Table 1: Average Performance of GBM Cancer Model in TCGA

Resulting average AUCs of Gradient Boosting Machine models trained on germline CNVs indicates that man cancers can be distinguished between other cancers. for most cancers. Prediction through this method performs better than chance. This table is ordered by AUC. Ovarian Serous Cystadenocarcinoma was the most distinguishable cancer form the rest of the TCGA database. Kidney Chromophobe performed the worst most likely because there are only a few patient samples in TCGA.

Cancer Type	Average AUC	Standard Deviation
Ovarian Serous Cystadenocarcinoma	0.8776165	0.002230274
Glioblastoma Multiforme	0.8181152	0.002663462
Mesothelioma	0.7661127	0.014158982
Pheochromocytoma and Paraganglioma	0.7657664	0.005150056
Kidney Renal Clear Cell Carcinoma	0.7631952	0.009482344
Esophageal Carcinoma	0.7465629	0.010221002
Uveal Melanoma	0.7389811	0.016093650
Colon Adenocarcinoma	0.7361154	0.006013044
Uterine Carpinosarcoma	0.7163798	0.022043596
Thyroid Carcinoma	0.7087212	0.008011903
Adrenocortical Carcinoma	0.7068969	0.013513626
Kidney Renal Papillary Cell Carcinoma	0.7060062	0.006150212
Breast Invasive Carcinoma	0.7016540	0.002804281
Skin Cutaneous Melanoma	0.6967139	0.005545299
Stomach Adenocarcinoma	0.6958837	0.004328998
Uterine Corpus Endometrial Carcinoma	0.6815276	0.006926379
Thymoma	0.6783321	0.017123537
Lung Squamous Cell Carcinoma	0.6783182	0.007591525
Liver Hepatocellular Carcinoma	0.6770429	0.010803151
Pancreatic Adenocarcinoma	0.6685822	0.011634259
Bladder Urothelia Carcinoma	0.6666951	0.004346832
Lung Adenocarcinoma	0.6658928	0.006071668
Sarcoma	0.6590625	0.007575377
Head and Neck Squamous Cell Carcinoma	0.6590391	0.007113240
Testicular Germ Cell Tumors	0.6569525	0.010302327
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	0.6510947	0.036368128
Rectum Adenocarcinoma	0.6494119	0.014650242
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	0.6491048	0.009488570
Brain Lower Grade Glioma	0.6313900	0.005302091
Prostate Adenocarcinoma	0.6299259	0.007695486
Cholangiocarcinoma	0.5190290	0.058815978
Kidney Chromophobe	0.3713367	0.096730788

Additionally, each model was cross-validated 10 times using the original training dataset. We then averaged the AUCs of the receiver-operator curve (ROC) to obtain the resulting performance of the technique for each specific cancer type. We plotted the ROC for individual models, where an AUC of 1 is a perfect model. Models will have a different ROC for the training results and the cross-validated results. We are interested in the cross-validated results as it provides a more accurate depiction of how the model would perform on new data (i.e., indication of model generalization). The resulting performance of our models indicates our technique can differentiate one cancer from other cancers better than chance, using the CSLV data taken as masked CNVs in TCGA (Fig. 3-8). Masked data omits genetic information from the Y chromosome as part of the anonymization process, which is most likely why some cancers such as Prostate Adenocarcinoma (PRAD) do not perform as well.

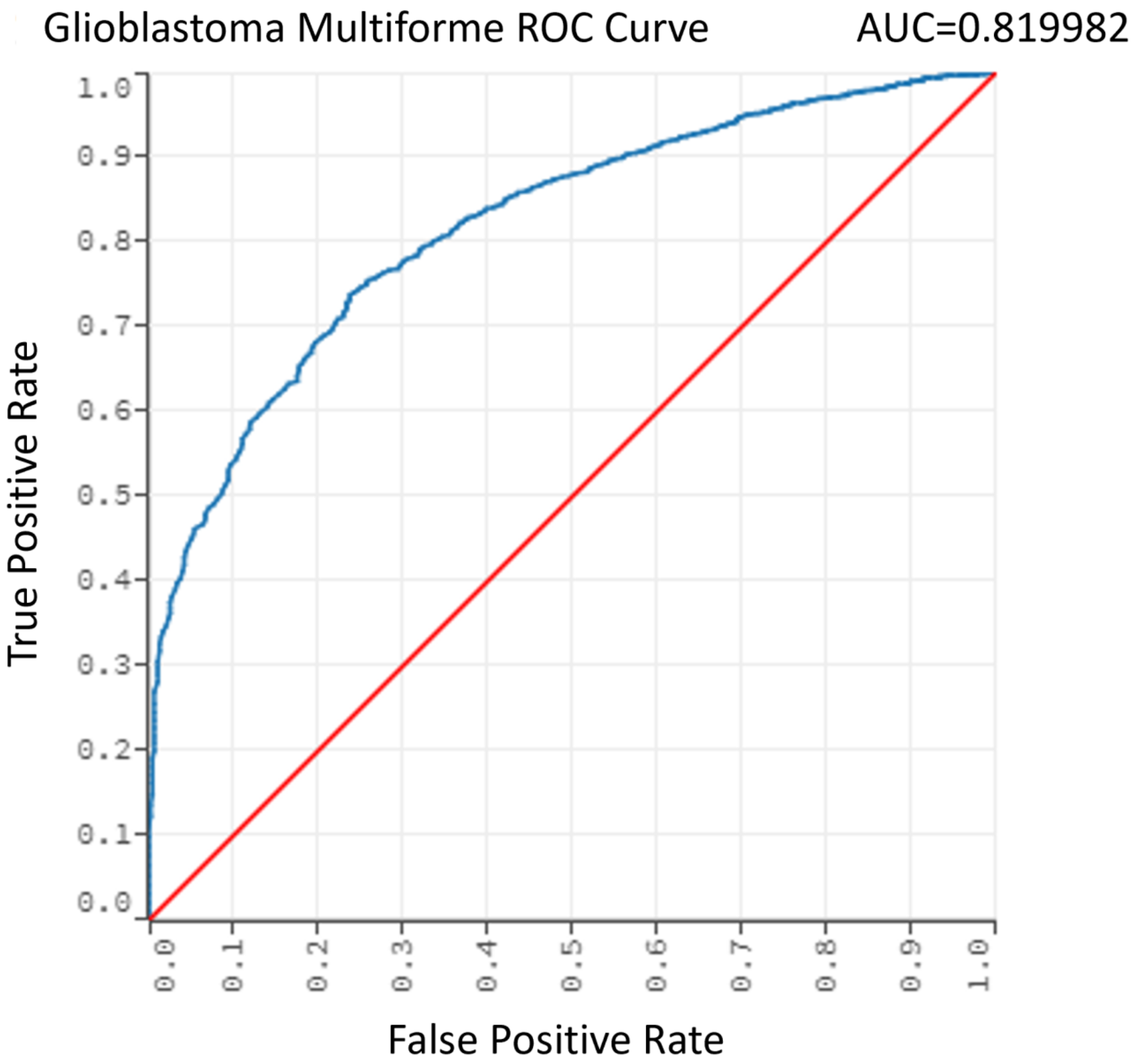


Figure 3: Receiver Operator Curve for Glioblastoma Multiforme

This graph represents the performance of a model trained on TCGA data from masked CNVs from blood samples. The model attempted to classify glioblastoma multiforme from all other cancers in the TCGA dataset. The AUC of this model was ~0.82.

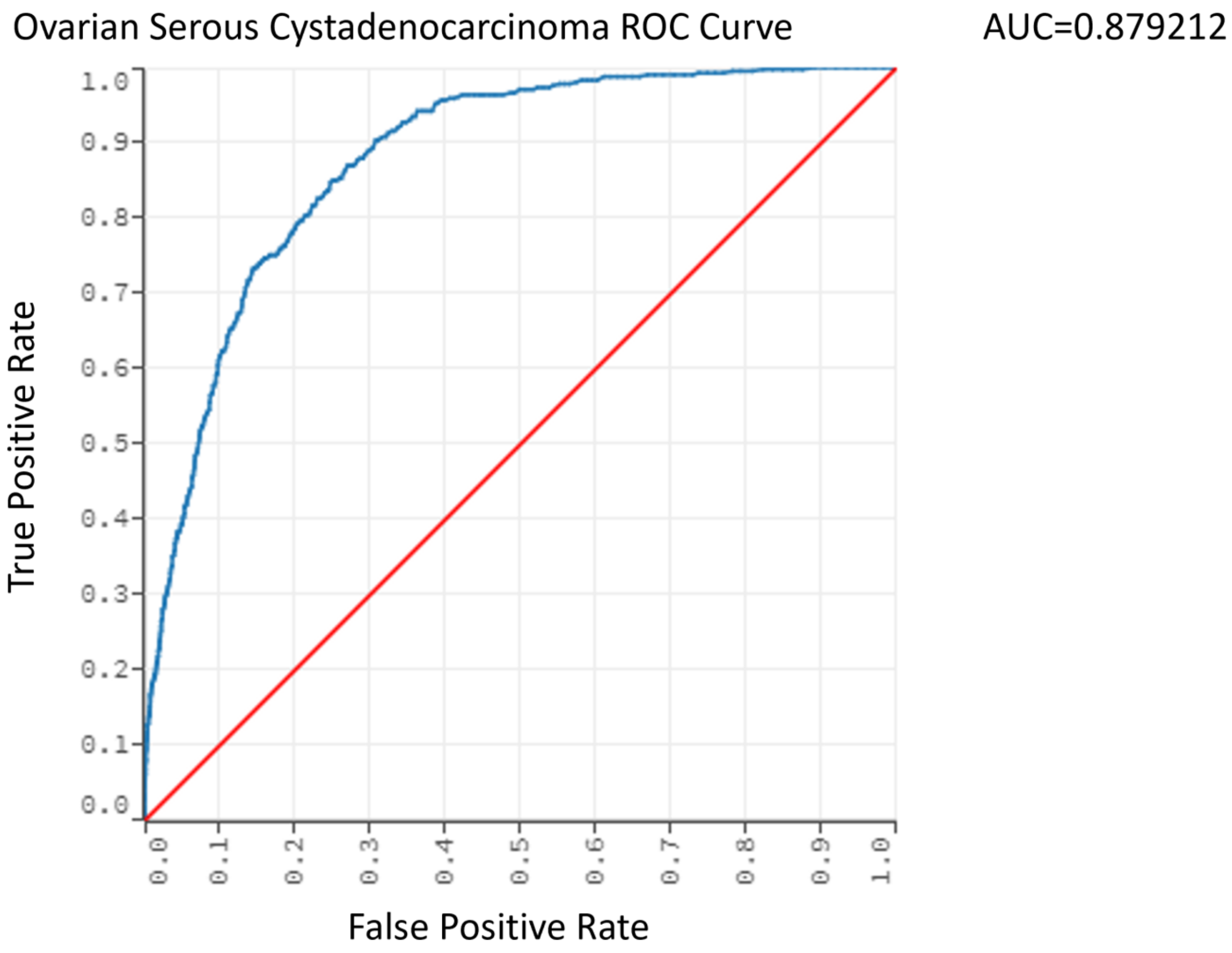


Figure 4: Receiver Operator Curve for Ovarian Serous Cystadenocarcinoma

This graph represents the performance of a model trained on TCGA data from masked CNVs from blood samples. The model attempted to classify ovarian serous cystadenocarcinoma from all other cancers in the TCGA dataset. The AUC of this model was ~0.88.

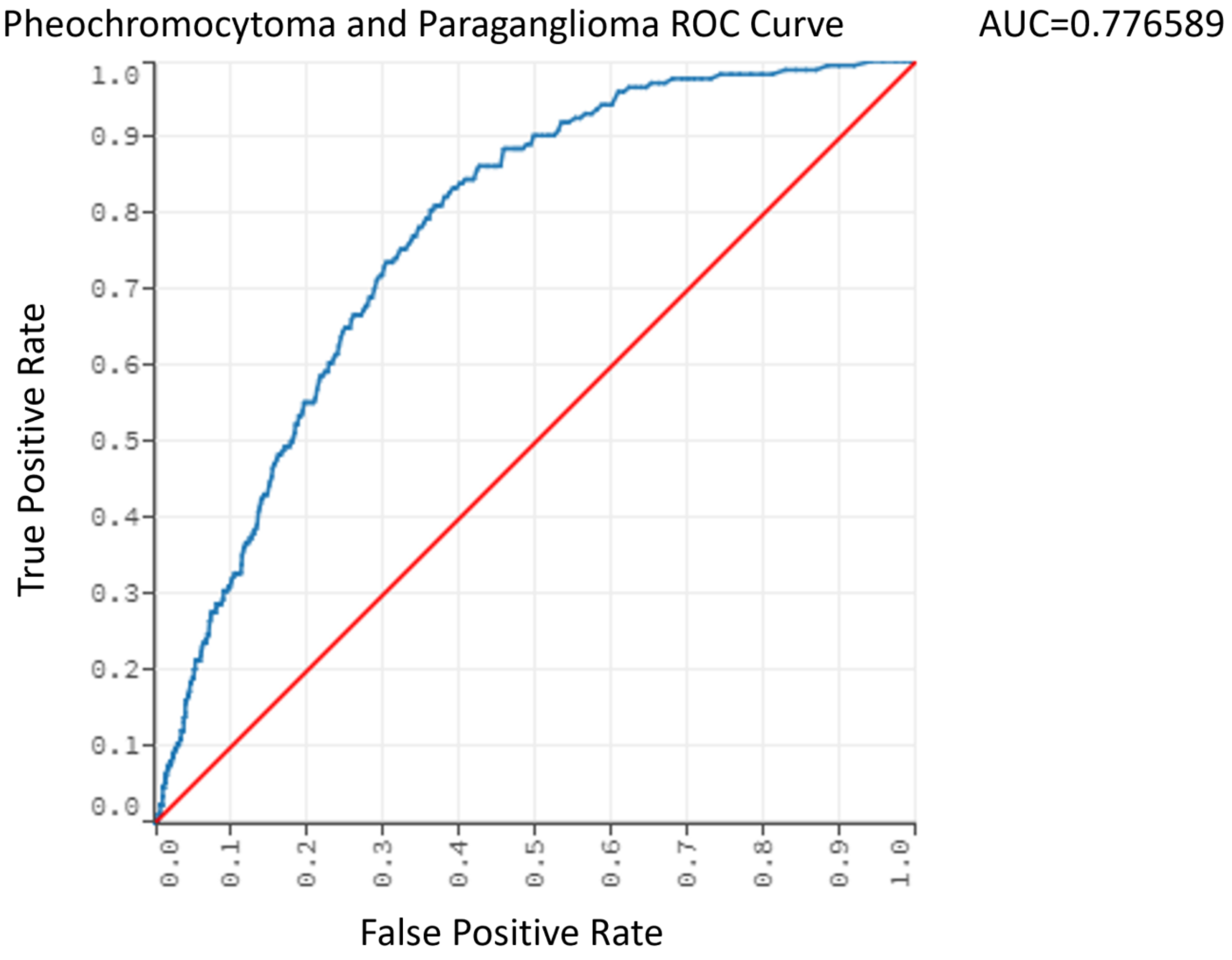


Figure 5: Receiver Operator Curve for Pheochromocytoma and Paraganglioma

This graph represents the performance of a model trained on TCGA data from masked CNVs from blood samples. The model attempted to classify pheochromocytoma and paraganglioma from all other cancers in the TCGA dataset. The AUC of this model was ~0.78.

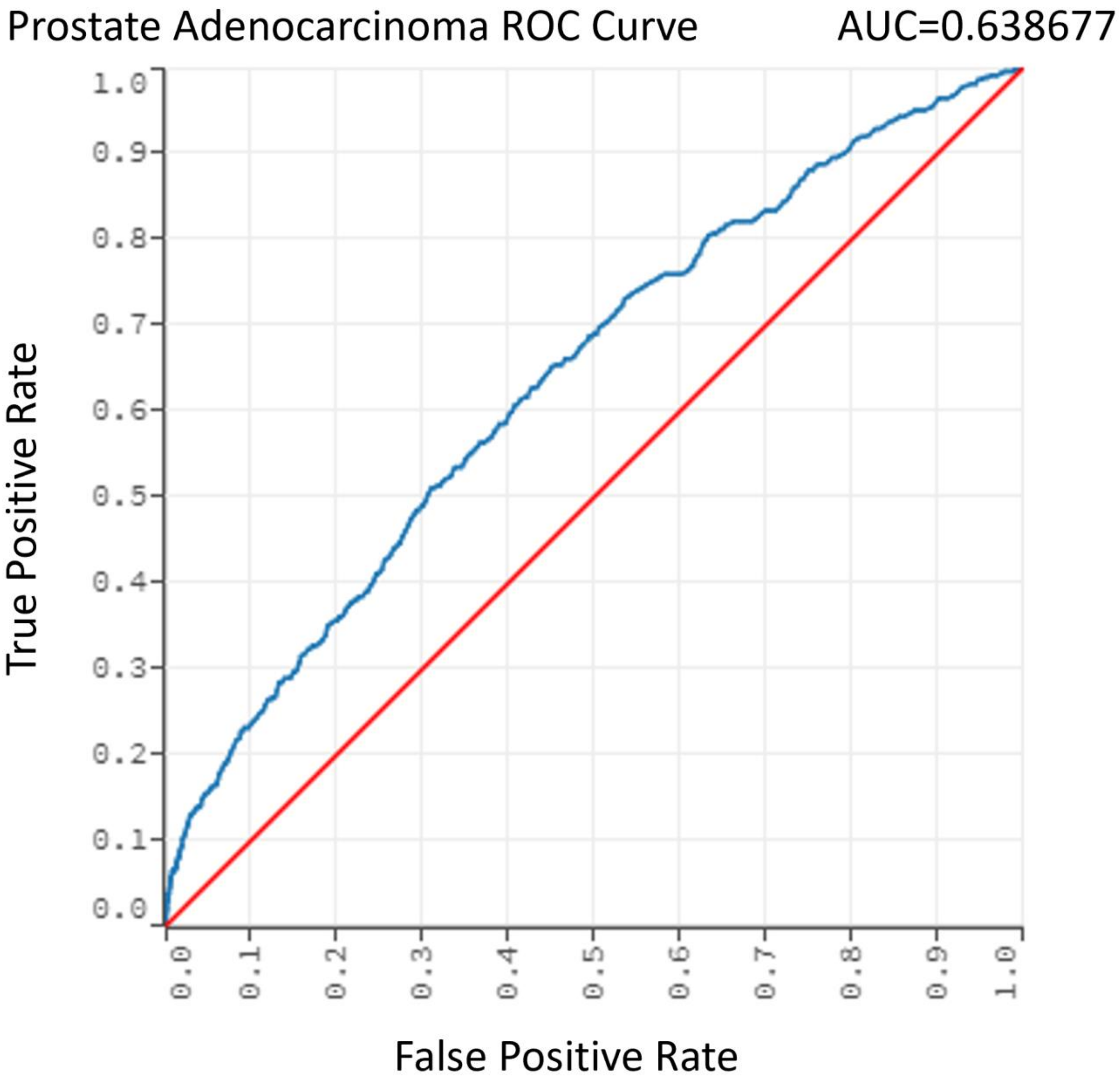


Figure 6: Receiver Operator Curve for Prostate Adenocarcinoma

This graph represents the performance of a model trained on TCGA data from masked CNVs from blood samples. The model attempted to classify prostate adenocarcinoma from all other cancers in the TCGA dataset. The AUC of this model was ~0.64.

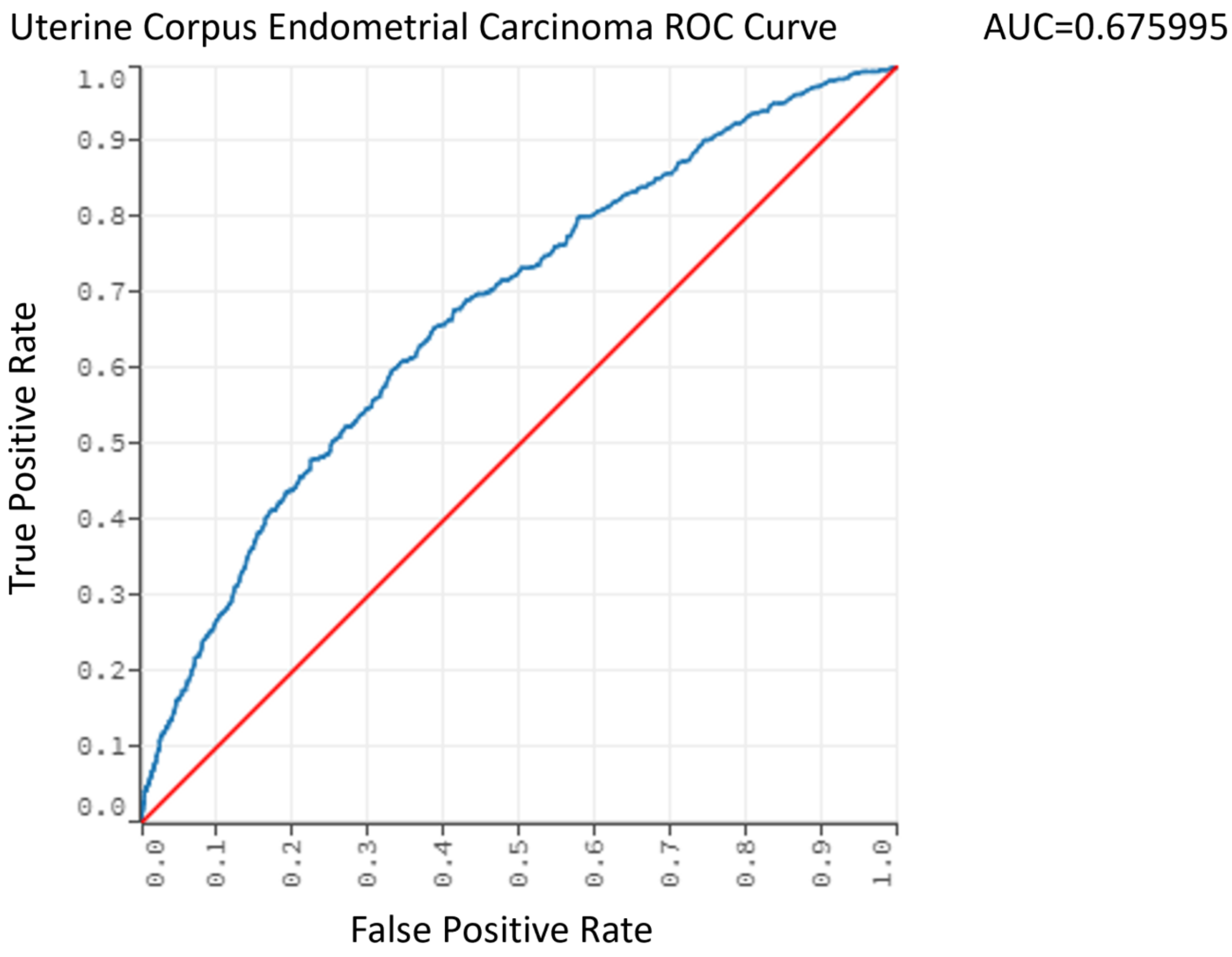


Figure 7: Receiver Operator Curve for Uterine Corpus Endometrial Carcinoma

This graph represents the performance of a model trained on TCGA data from masked CNVs from blood samples. The model attempted to classify uterine corpus endometrial carcinoma from all other cancers in the TCGA dataset. The AUC of this model was ~0.68.

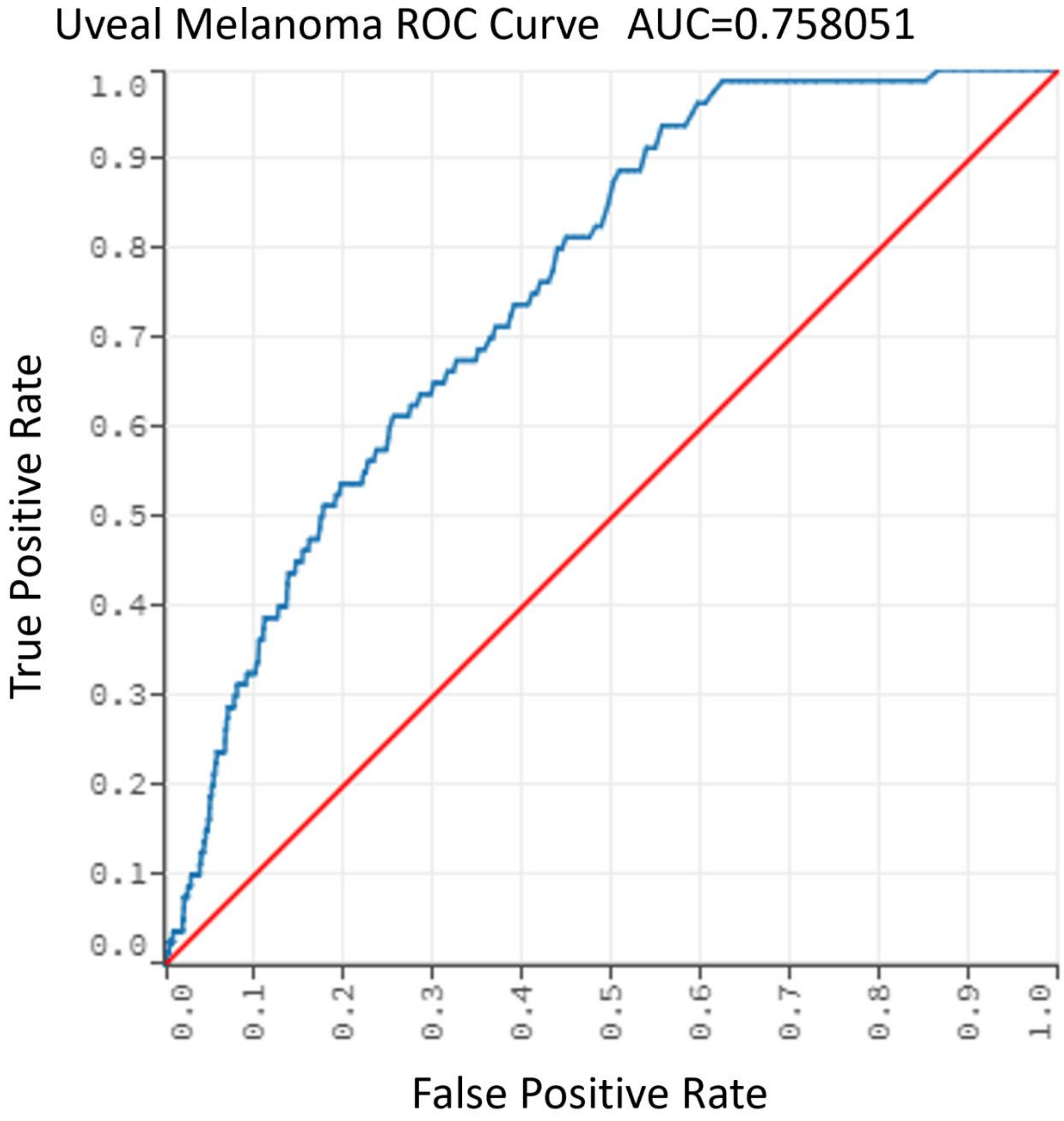


Figure 8: Receiver Operator Curve for Uveal Melanoma

This graph represents the performance of a model trained on TCGA data from masked CNVs from blood samples. The model attempted to classify uveal melanoma from all other cancers in the TCGA dataset. The AUC of this model was ~0.76.

Chapter 5.2: Prediction of Somatic Tissue Samples with Models Trained on Germline Samples

We then sought to examine if models trained on germline peripheral blood samples could perform equally as well when given a set of unknown somatic or tumor samples. To test this, we performed the same process as before but this time only trained our GBDT Models with 80 % of the patients, leaving 20 % aside for later testing. For the training set (containing 80 % of the patient samples) we again utilized germline blood samples and the masked CNV data to iteratively train our models. Each model was cross-validated 10 times. We then took the test group (containing the remaining 20% of samples) and used the masked CNV data from the somatic tumor samples as the set to predict on. The models for certain cancers performed equally well on this new set of somatic tumor data compared to the cross-validation results (Figure 9, 10).

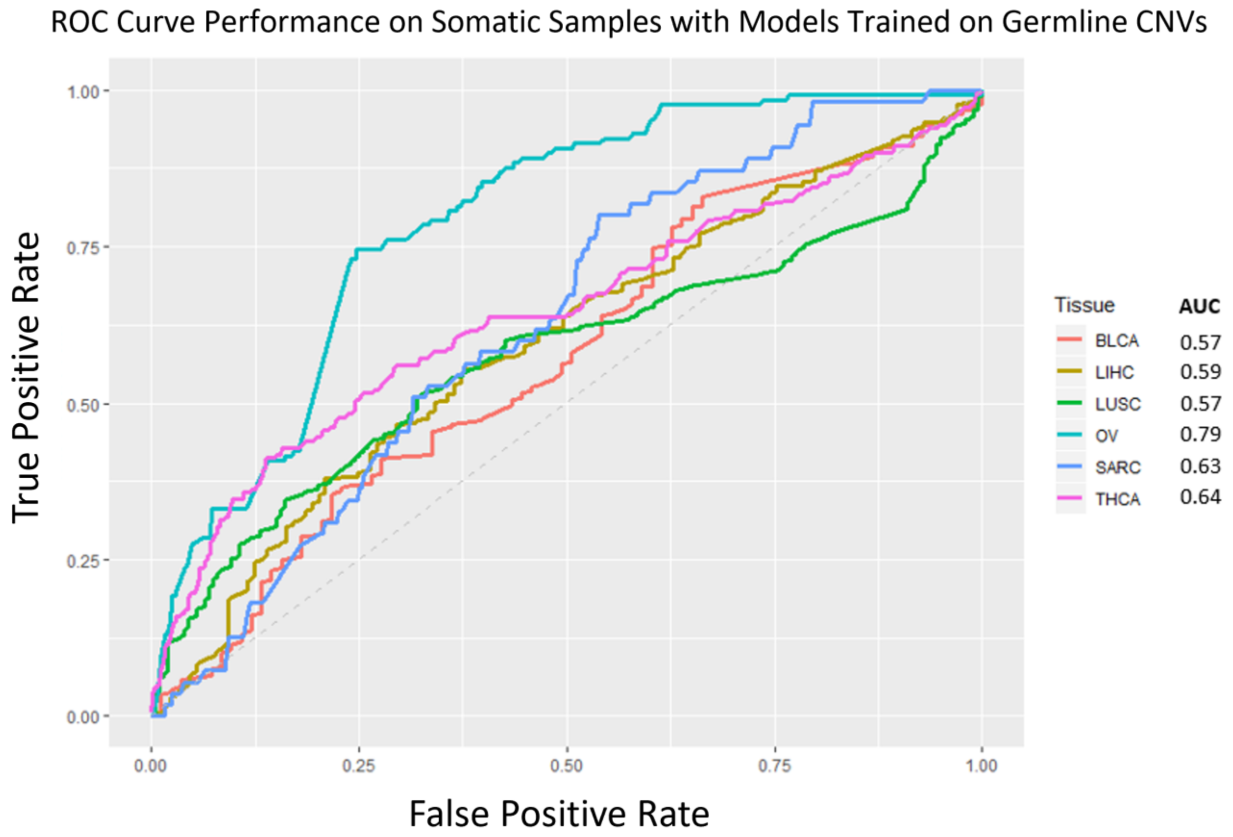


Figure 9 : Prediction on Somatic Samples with models trained on Germline Samples (BLCA, LIHC, LUSC, OV, SARC, THCA)

Performance of models trained with germline CNVs on somatic CNVs. Ovarian Cancer (OV = 0.79) is one of the best performing models even when predicting on somatic CNVs.

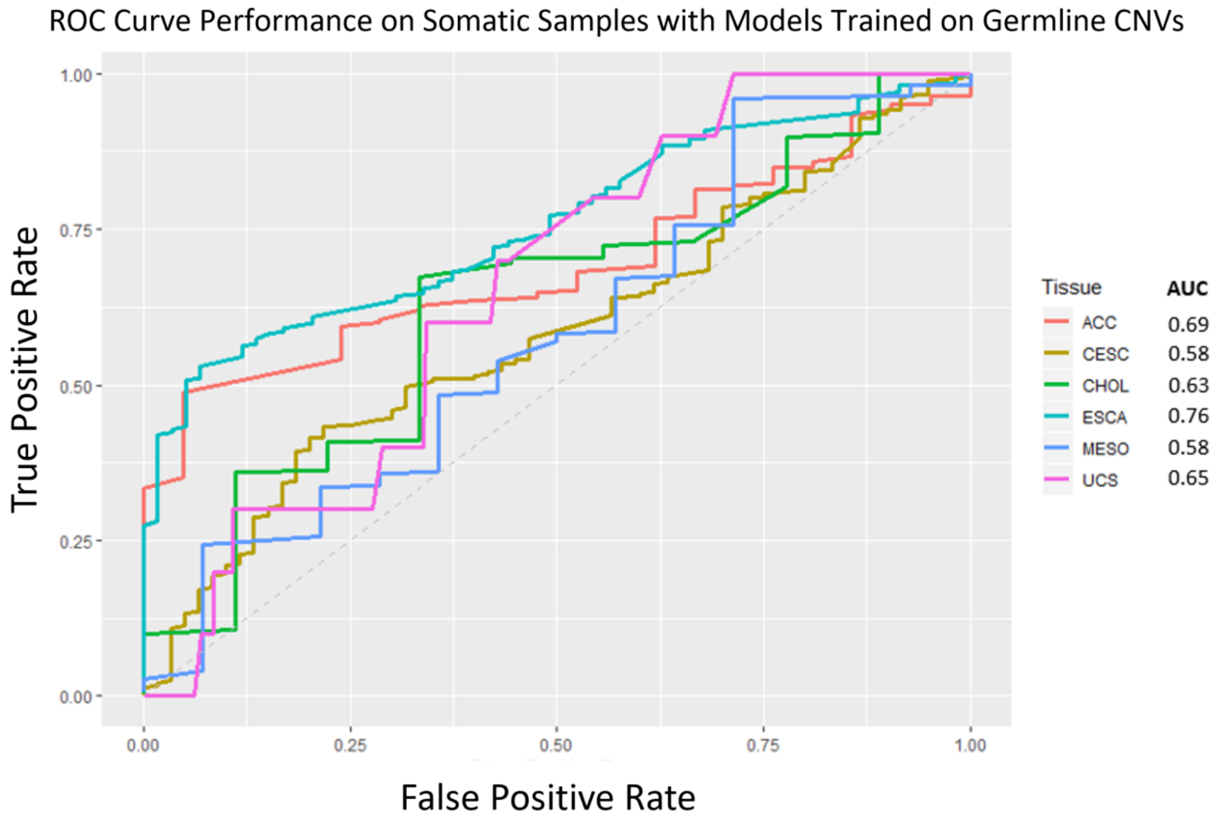


Figure 10: Prediction on Somatic Samples with models trained on Germline Samples (ACC, CESC, CHOL, ESCA, MESO, UCS)

Performance of models trained with germline CNVs on somatic CNVs. Certain cancers still perform better than chance. Of note Esophageal Cancer (ESCA = 0.76) performs rather well indicating possible hereditary factors which could increase risk.

Certain cancer types however performed rather poorly on the somatic samples and were little better than chance (Figure 11, 12, 13). We believe cancers which did not perform better than chance may possibly contain risk factors on the Y chromosome. TCGA removes data from the Y chromsome during the anonymization process. Alternatively, models which did not perform as well on somatic samples may indicate significant environmental factors, which may overshadow the genetic effects in smaller populations. We note also that we trained models using blood samples and predict here on tumor samples, meaning that tumor samples may possess a significantly altered CNV landscape when compared to blood samples.

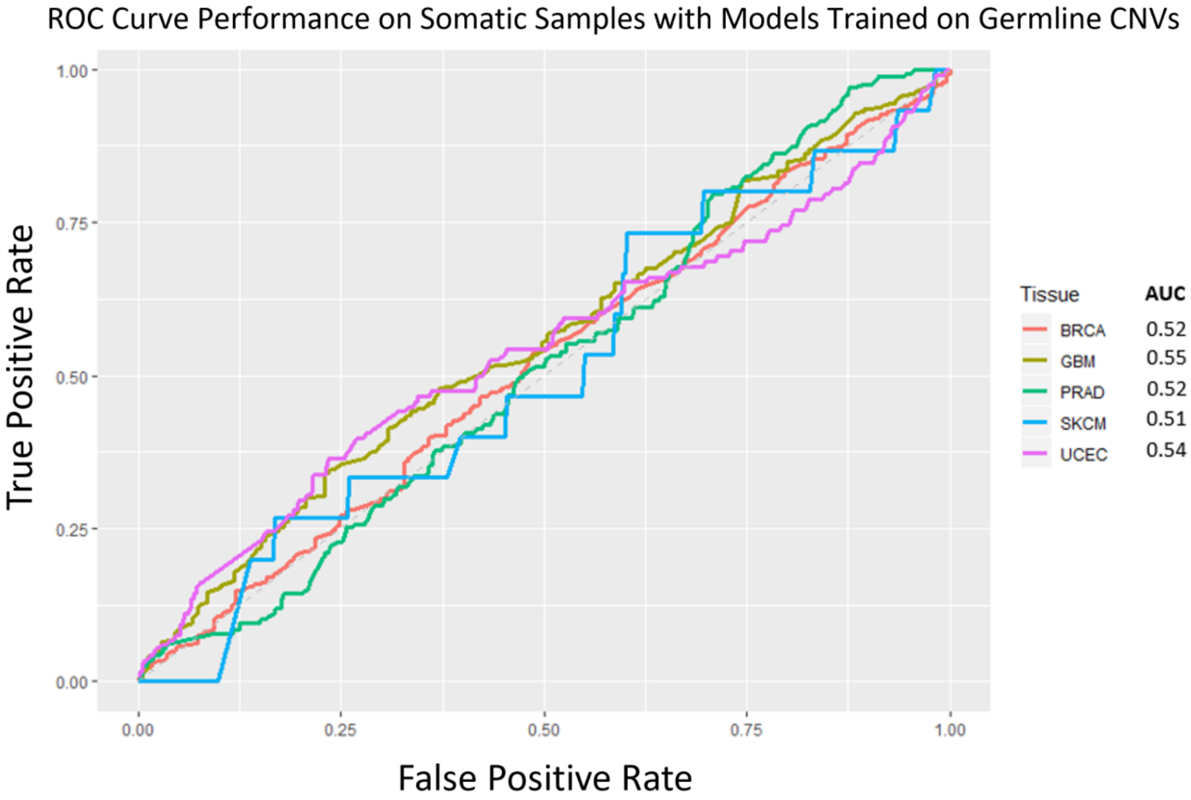


Figure 11: Models indicating possible importance of Y chromosome or environmental factors (BRCA, GBM, PRAD, SKCM, UCEC)

Cancer types where models trained with germline CNVs did not perform much better than chance when predicting on somatic CNVs, indicating that these cancers need more information than the anonymized data in the first 22 chromosomes. Additionally, environmental factors may play a bigger role with smaller population sets. AUCs, BRCA= 0.52, GBM=0.55, PRAD=0.52, SKCM=0.51, UCEC=0.54.

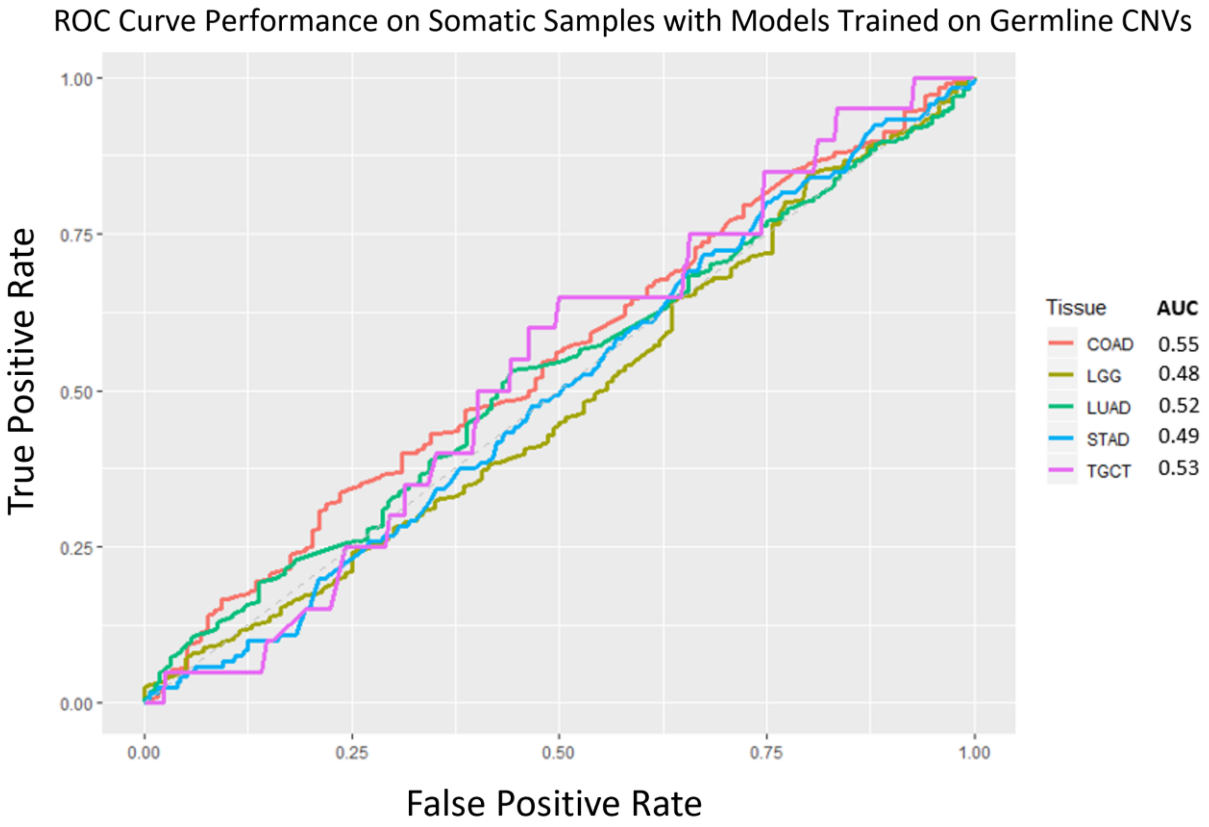


Figure 12: Models indicating possible importance of Y chromosome or environmental factors (COAD, LGG, LUAD, STAD, TGCT)

Cancer types where models trained with germline CNVs did not perform much better than chance when predicting on somatic CNVs, indicating that these cancers need more information than the anonymized data in the first 22 chromosomes. Additionally, environmental factors may play a bigger role with smaller population sets. AUCs, COAD=0.55, LGG=0.48, LUAD=0.52, STAD=0.49, TGCT=0.53.

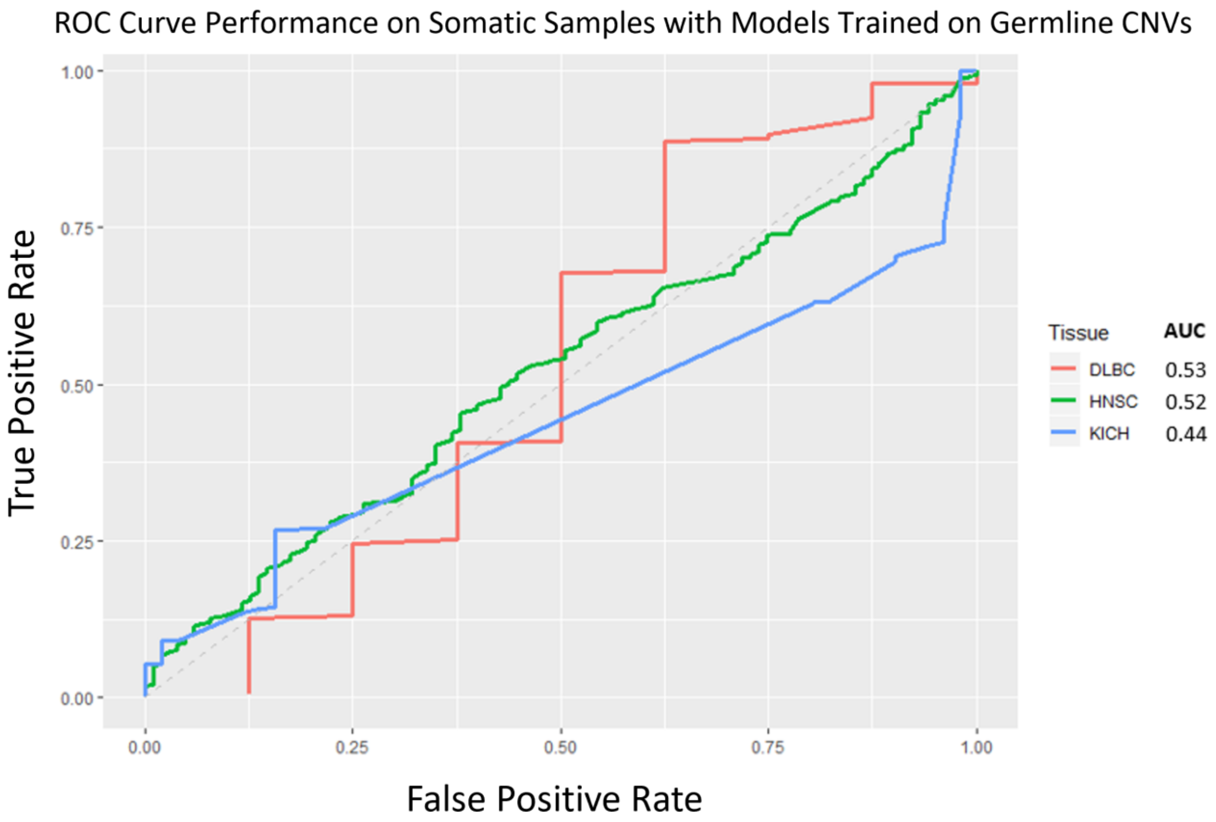


Figure 13: Models indicating possible importance of Y chromosome or environmental factors (DLBC, HNSC, KICH)

Cancer types where models trained with germline CNVs did not perform much better than chance when predicting on somatic CNVs, indicating that these cancers need more information than the anonymized data in the first 22 chromosomes. Additionally, environmental factors may play a bigger role with smaller population sets. AUCs, DLBC=0.53, HNSC=0.52, KICH=0.44.

Chapter 5.3: Comparison of Masked CNV and Unmasked CNV Gradient Boosting Models

Our initial experiments utilized masked CNV data to train our GBDT Models. As such, we wanted to test if our technique performed equally well, or better, using an equal number of the raw copy numbers (i.e., unmasked CNVs). Since the unmasked CNVs are significantly smaller segments than the masked CNVs, we utilized the top 100 raw CNVs. The results indicate that GBDT Models using masked CNVs tend to perform better than unmasked CNVs (Figure 14). This seems to hold true across all cancer types (Figures 15-21).

The unmasked models only use the top 100 CNVs when training the model. Using all the CNVs in the unmasked CNV models may improve the predictive power to a comparable level with the masked CNV models. However, doing so becomes computationally intensive and could also decrease performance due to increasing column variables, creating a dimensionality problem. We also were aiming to demonstrate that because Masked CNVs encompassed larger sections of the chromosome, we could still get good predictive power in a single averaged value.

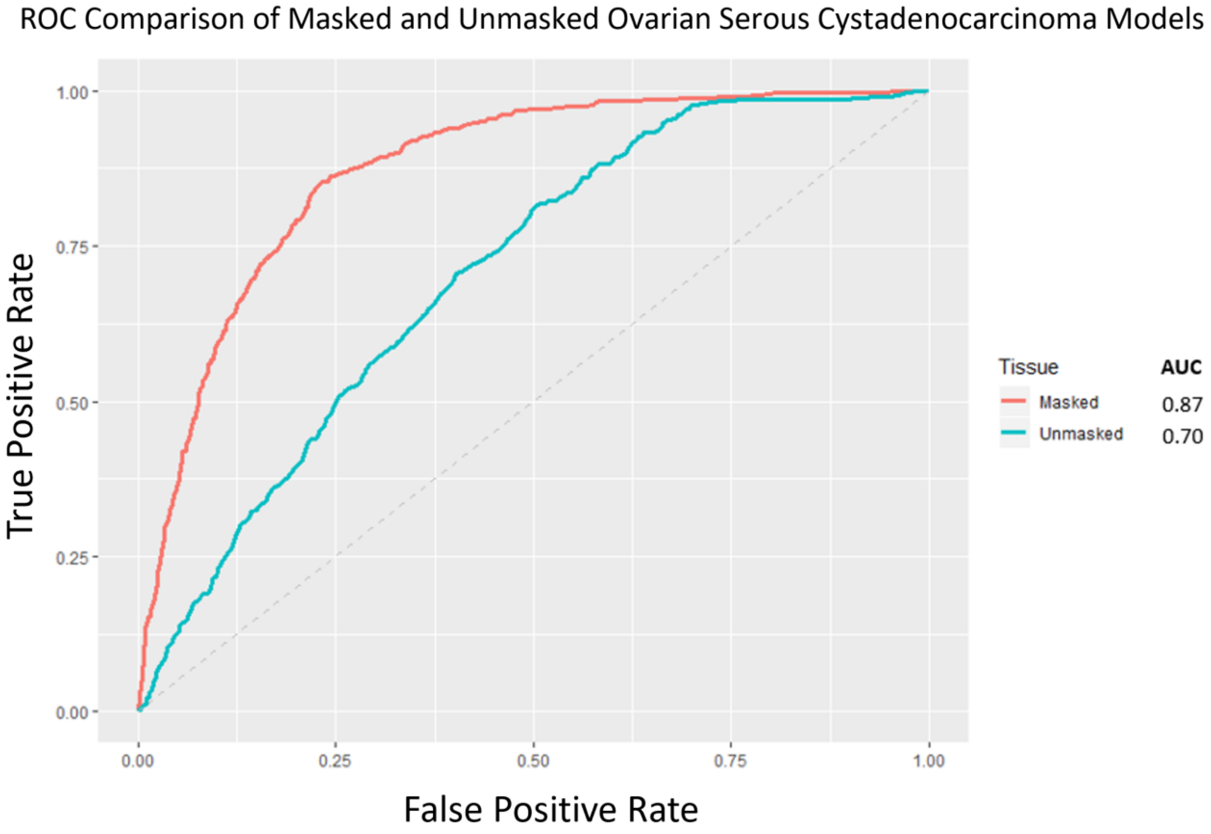


Figure 14: Comparison of Masked and Unmasked Ovarian Serous Cystadenocarcinoma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Ovarian Serous Cystadenocarcinoma, the difference between Masked and Unmasked is 0.87 and 0.70 respectively.

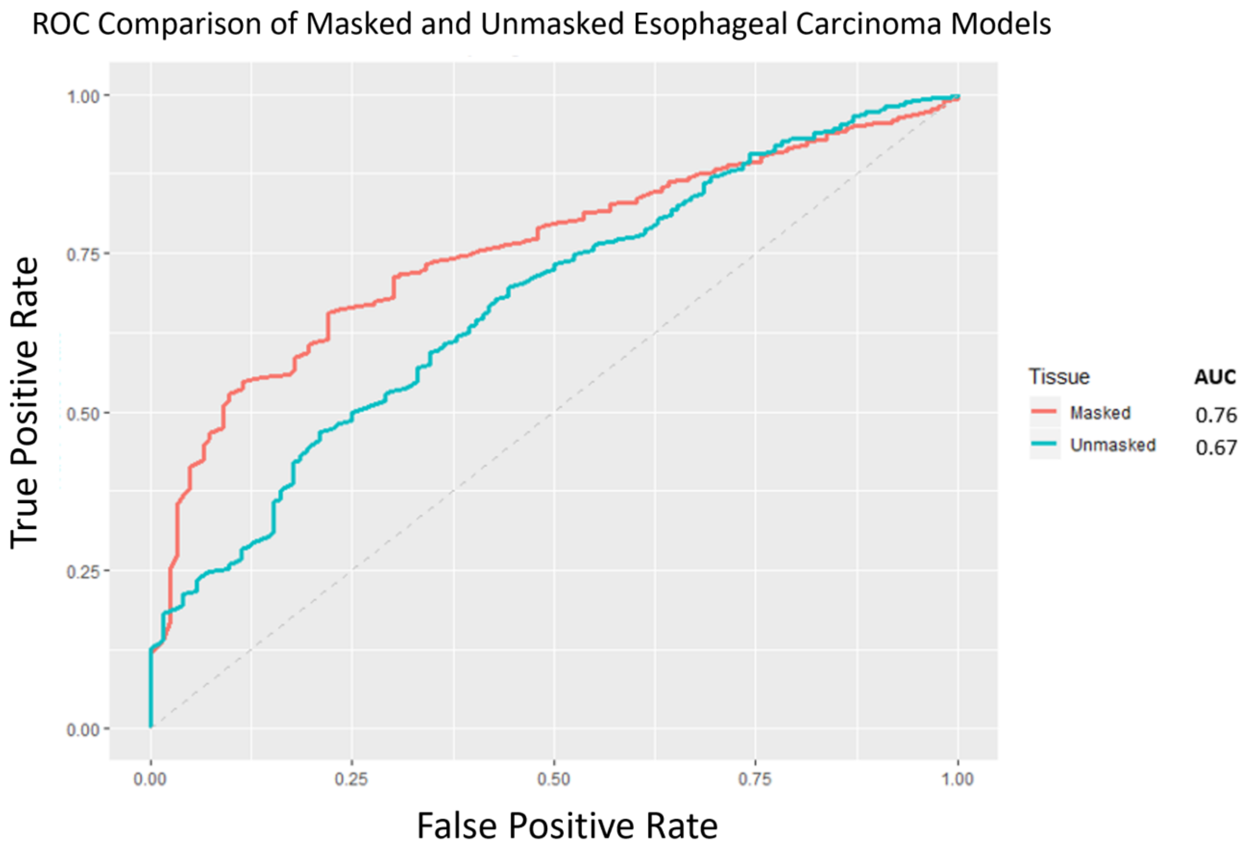


Figure 15: Comparison of Masked and Unmasked Esophageal Carcinoma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Esophageal Carcinoma, the difference between Masked and Unmasked is 0.76 and 0.67 respectively.

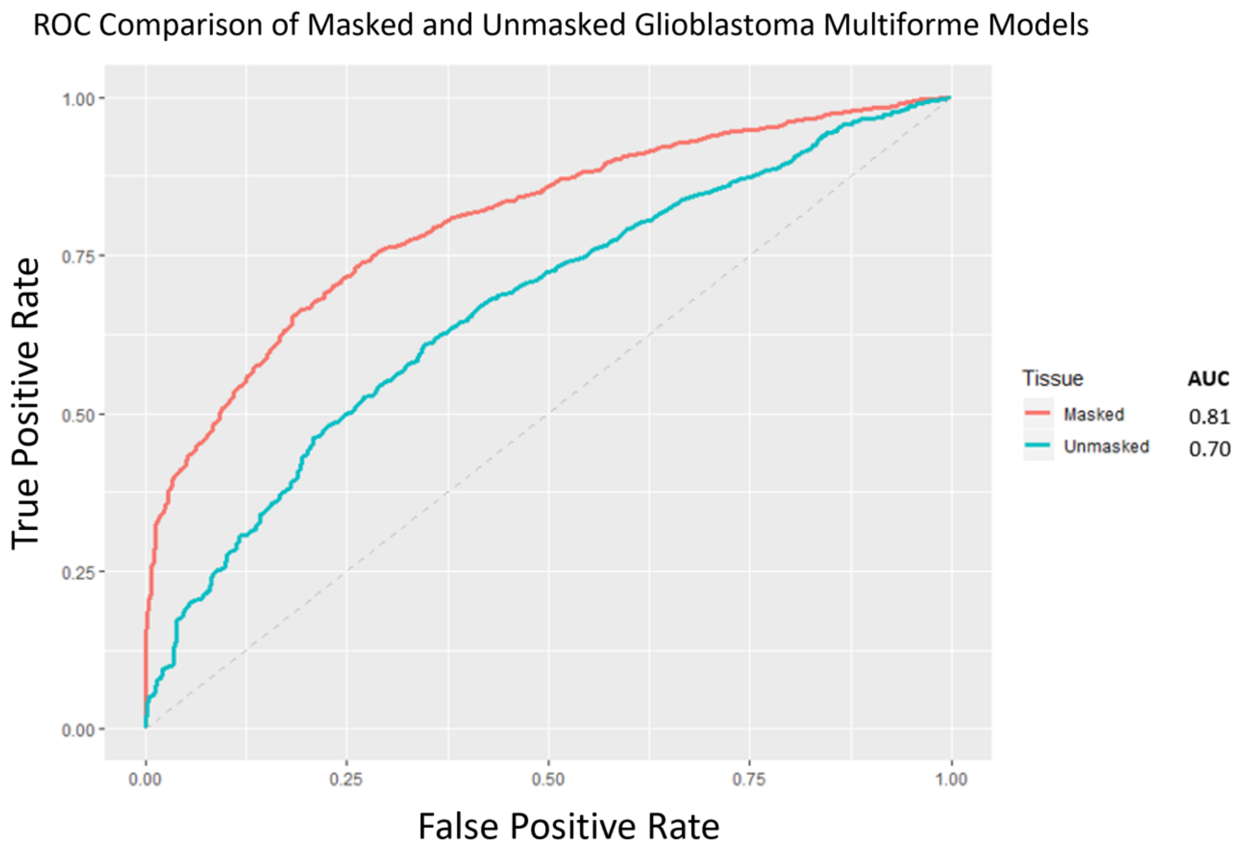


Figure 16: Comparison of Masked and Unmasked Glioblastoma Multiforme Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Glioblastoma Multiforme, the difference between Masked and Unmasked is 0.81 and 0.70 respectively.

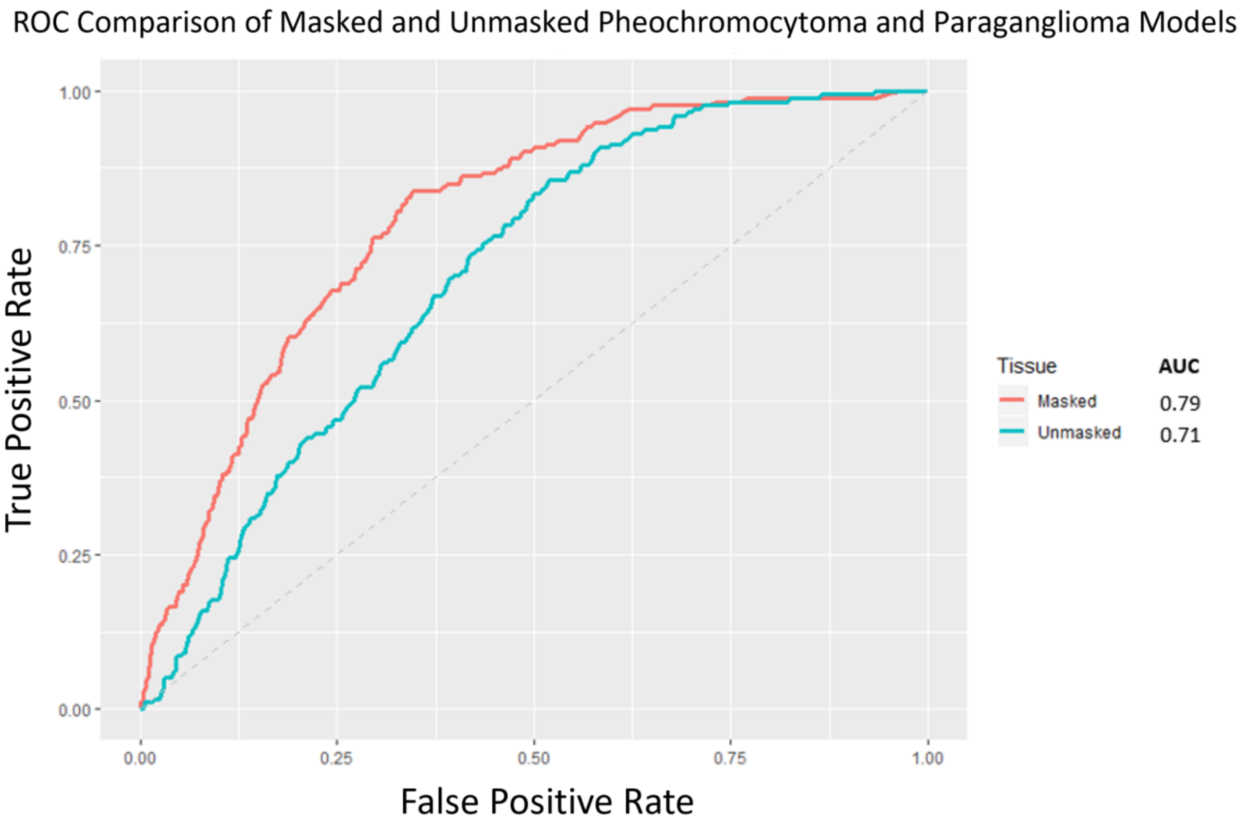


Figure 17: Comparison of Masked and Unmasked Pheochromocytoma and Paraganglioma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Pheochromocytoma and Paraganglioma, the difference between Masked and Unmasked is 0.79 and 0.71 respectively.

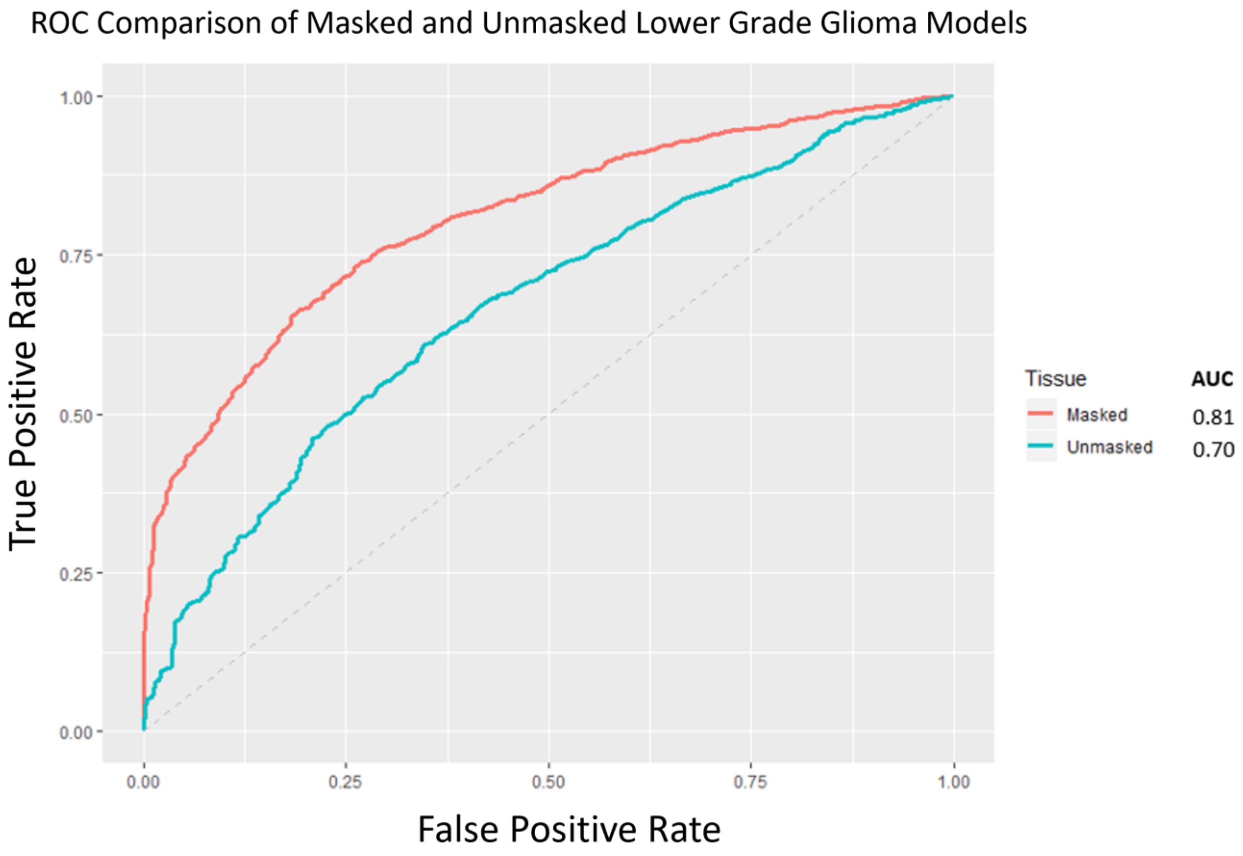


Figure 18: Comparison of Masked and Unmasked Lower Grade Glioma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Lower Grade Glioma, the difference between Masked and Unmasked is 0.81 and 0.70 respectively.

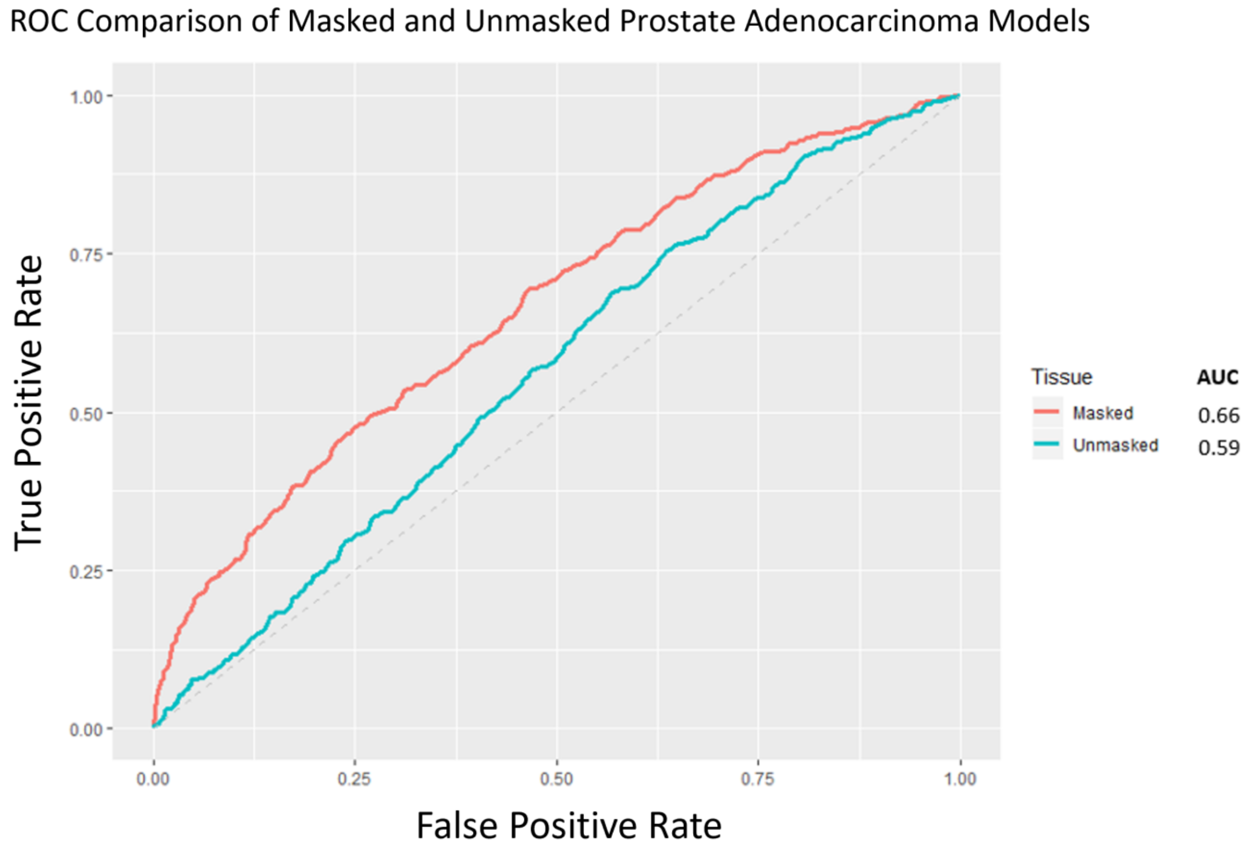


Figure 19: Comparison of Masked and Unmasked Prostate Adenocarcinoma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Prostate Adenocarcinoma, the difference between Masked and Unmasked is 0.66 and 0.59 respectively.

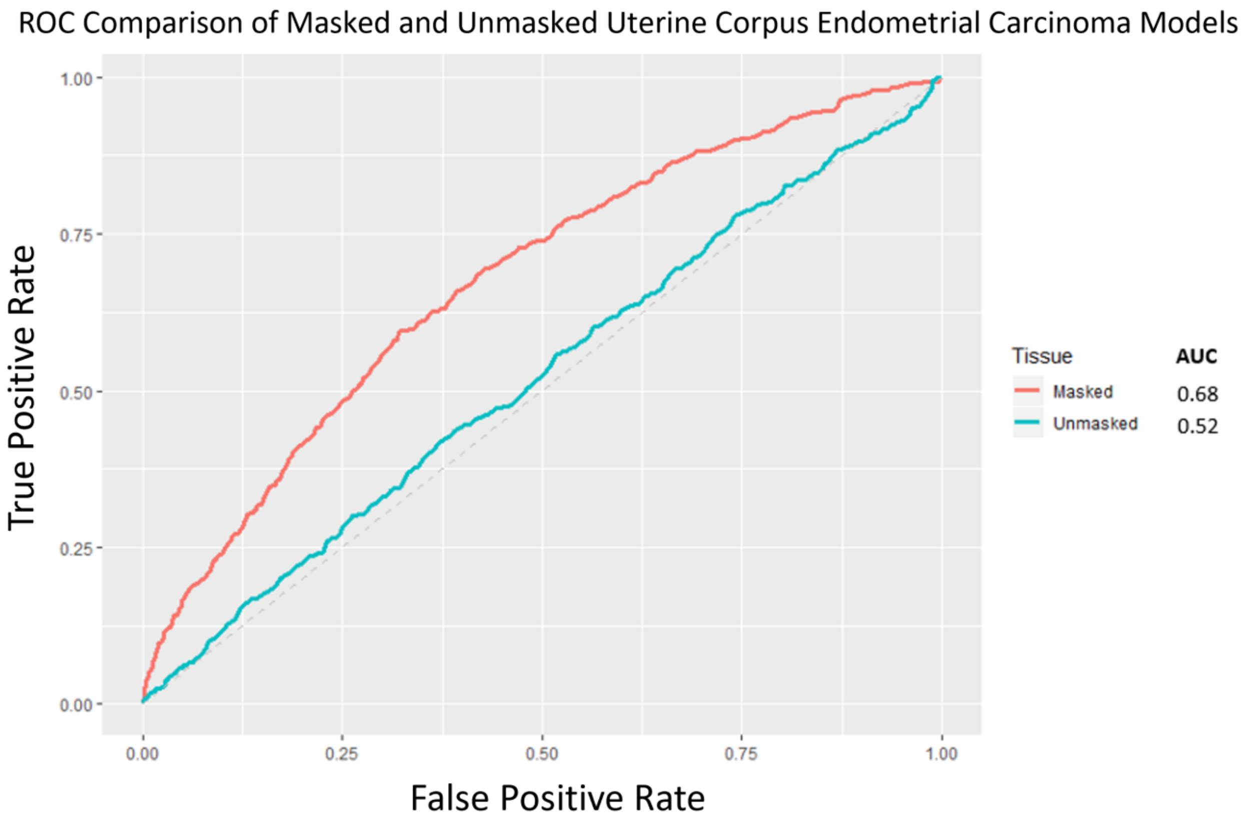


Figure 20: Comparison of Masked and Unmasked Uterine Corpus Endometrial Carcinoma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. For Uterine Corpus Endometrial Carcinoma, the difference between Masked and Unmasked is 0.68 and 0.52 respectively.

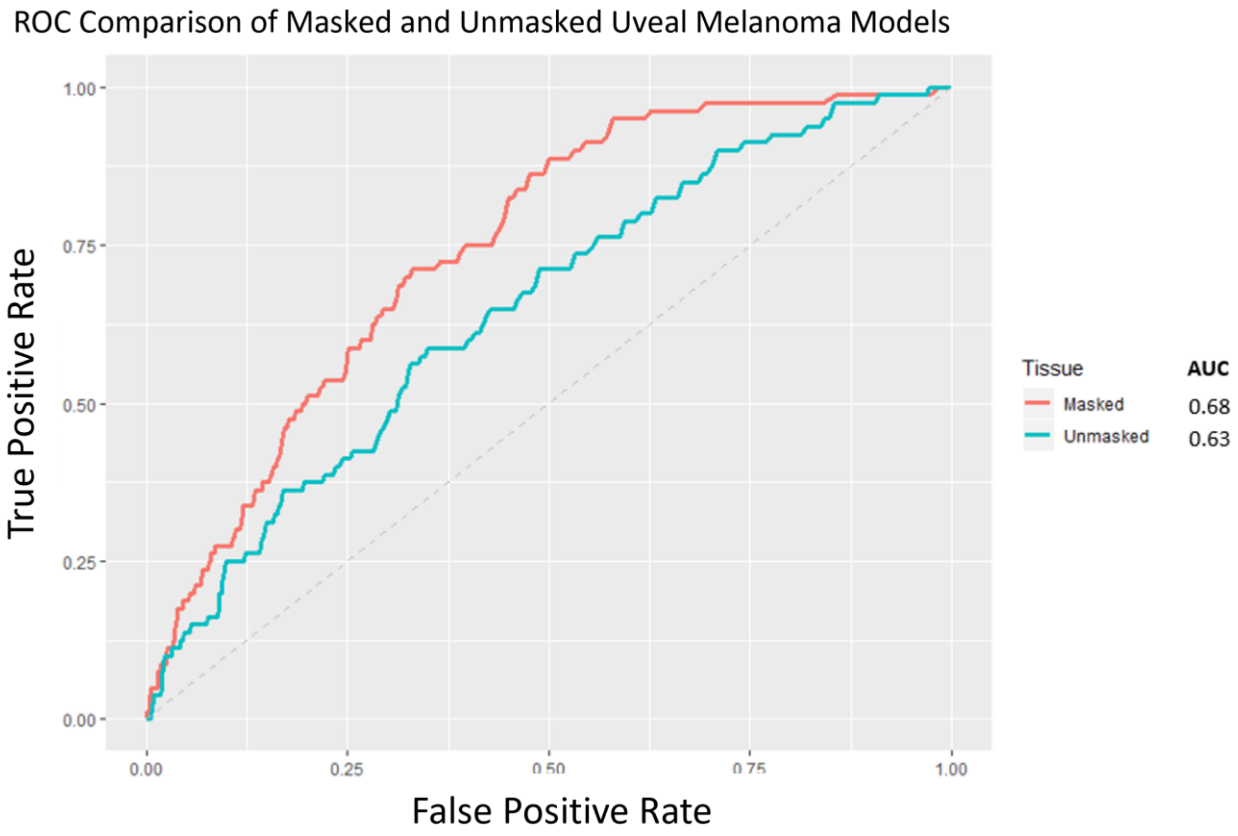


Figure 21: Comparison of Masked and Unmasked Uveal Melanoma Models

Models trained using TCGA Masked CNVs performed better than models trained using Unmasked CNVs. We believe this is mostly because if we take the same number of values from the masked CNV and from the unmasked CNV dataset, the masked dataset encompasses a larger amount of the genome. Uveal Melanoma, the difference between Masked and Unmasked is 0.68 and 0.63 respectively.

Chapter 5.4: Visualization of Decision Trees

After training our GBDT models, we visualized the individual decision trees through various application program interfaces (API) and through H2O's model optimized Java objects (MOJO). We can thus see how the model is mathematically coming to its decisions by looking at the resulting trees (Figure 22 and 23). Each model we built contains 50 trees. During prediction, the model runs each observation through all trees and sums the resulting values to obtain the most likely prediction.

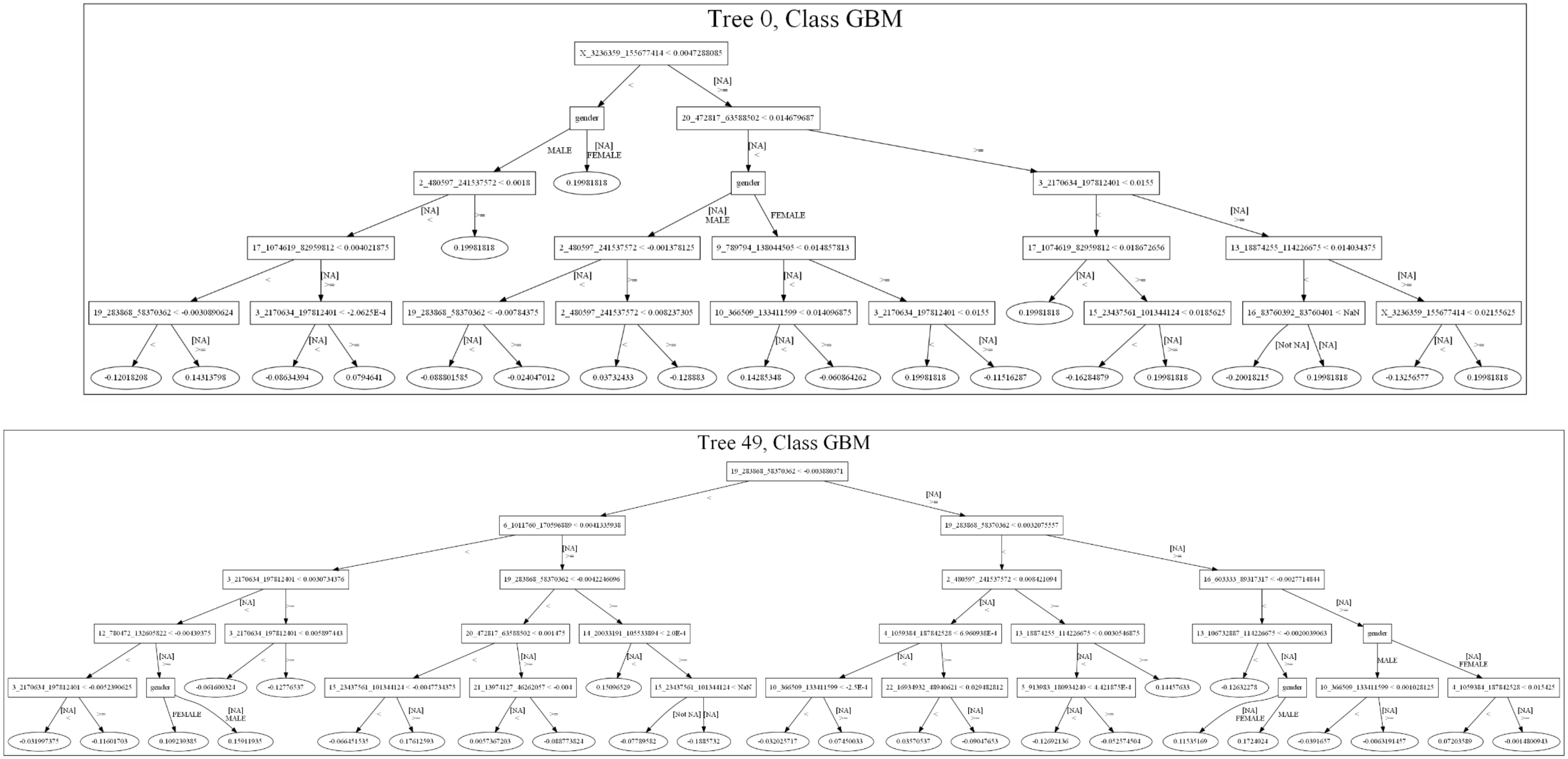


Figure 22: Example Decision Trees for Glioblastoma Multiforme Masked Germline CNV Models

The first and last decision trees are shown above as examples of how Gradient Boosted Decision Trees predict through fitting to the best estimate of the model. Each tree is predicting the residual r and thus when summed will provide a model $f(x) = f_0(x) + f_1(x) + \ldots + f_n(x)$, where each f is the result of a single tree.

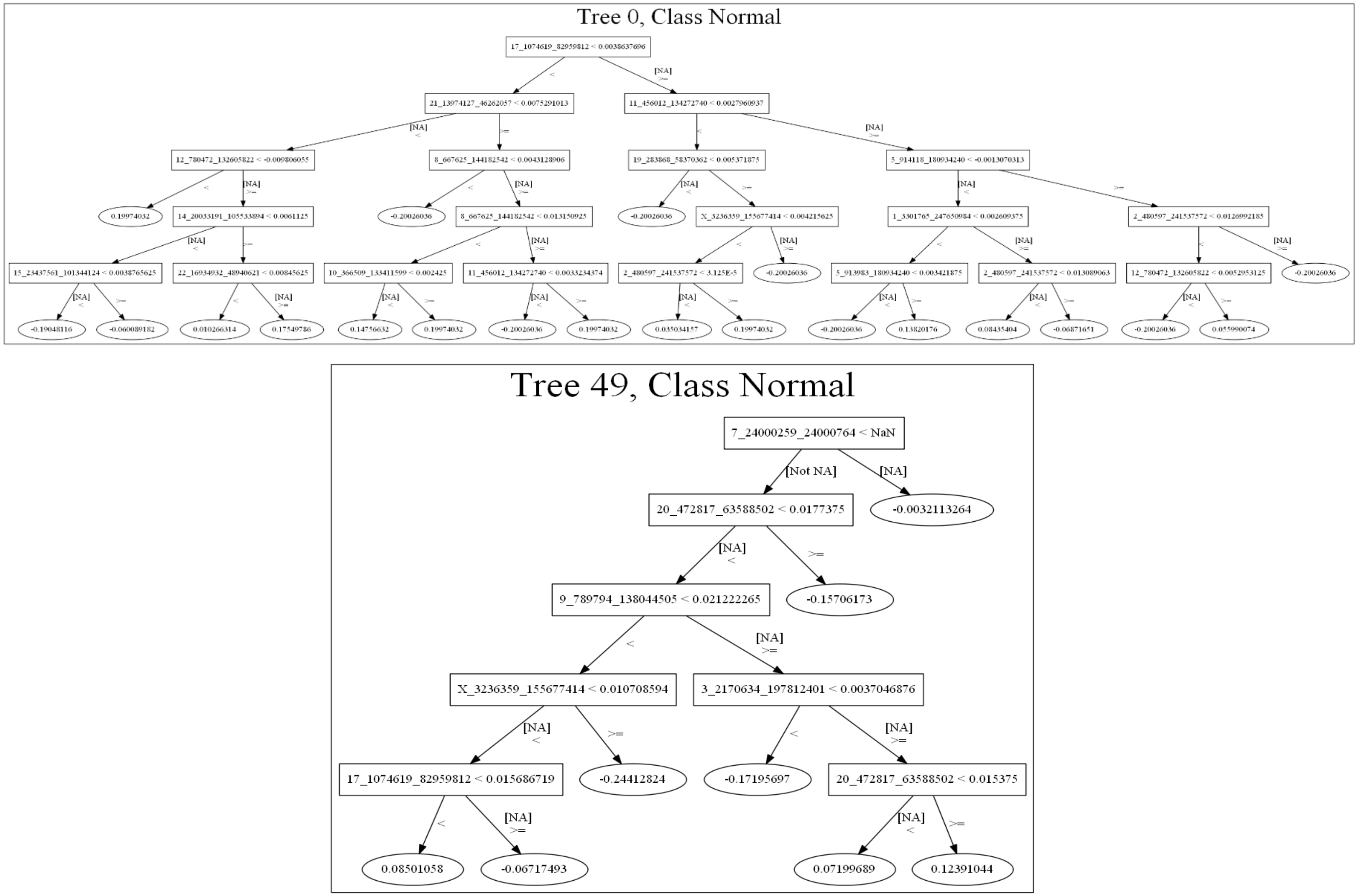


Figure 23: Example Decision Trees for Ovarian Cancer Masked Germline CNV Models

The first and last decision tree in a Gradient Boosted Decision Tree model for Ovarian Cancer. The last decision tree is significantly smaller and has less nodes than the first tree. Since we did not specify max depth or max number of leaves, the algorithm will attempt to account for residuals in the least amount of steps. This ensures adherance to weak classifiers and additive nature of gradient boosting.

Chapter 5.5: Chromosomal Scale Length Variations Indicate Genetic Factors for COVID-19 Severity

The course of COVID-19 varies from asymptomatic to severe (acute respiratory distress, cytokine storms, and death) in patients. The basis for this range in symptoms is unknown. One possibility is that genetic variation is responsible for the highly variable response to infection.

Human genetic variation can affect susceptibility and resistance to viral infections[121]. For instance, variants in the gene IFITM3 affect the severity of seasonal influenza[122]. Patients hospitalized from seasonal influenza had a particular allele of the gene IFITM3 at a higher rate than expected from the general population. Laboratory work determined that this allele can alter the course of the influenza virus infection.

We have previously shown that chromosome-scale length variation is a powerful tool to analyze genome wide associations[123]. This method is particularly appealing for genetic risk scores because it includes epistatic effects that might be missed with conventional genome wide association studies.

We sought to evaluate how well a genetic risk score based on chromosome-scale length variation and machine learning classification algorithms can predict severity of response to SARS-CoV-2 infection. We evaluated this approach on a dataset of 931 patients who had a severe reaction to Covid-19 in 2010. These patients had been previously genotyped as part of the UK Biobank. We also segmented these datasets into three overlapping datasets, shown in Table 2.

Table 2: Segmentation of COVID-19 Datasets

We segmented the dataset into three overlapping subsets. The first, which we called "1930" contained all UK Biobank participants born after 1930 who had a severe reaction to SARS-CoV-2 infection before 27 April 2020. The two subsets contained people born after 1940 and after 1950.

Dataset	Number
1930 (< 90 years of age)	981
1940 (< 80 years of age)	880
1950 (< 70 years of age)	468

The results are presented in Figure 24 and Table 3[17]. We found a significant difference between all three age groupings and their corresponding random controls (Figure 24). This finding indicates germ line genetics of the infected patient, as represented by the set of CSLVs, is correlated with the clinical severity of COVID-19. Additionally, Figure 24 and Table 4 shows that the AUC for the XGBoost classification model was about 0.51, but still significantly greater than 0.50[17].

Table 3: Comparison of AUCs for Overlapping COVID-19 Datasets

We compared the difference in mean AUC values between the various datasets using a t test. The datasets consisting of people born after 1930, 1940, and 1950 all showed significant differences with the corresponding random control. Those three datasets also showed significant differences between the mean AUC and 0.5. The three random controls did not show a significant difference between the mean AUC and 0.5, as expected. An AUC value of 0.5 represents a random classification test, one in which the algorithm is no better than guessing.

Value 1	Value 2	p value of t test
1930 data	1930 random	2×10^{-11}
1940 data	1940 random	1×10^{-9}
1950 data	1950 random	1×10^{-4}

0.5	1930 data	3 x 10^{-14}
0.5	1940 data	4 x 10^{-13}
0.5	1950 data	3 x 10^{-4}
0.5	1930 random	0.1
0.5	1940 random	0.4
0.5	1950 random	0.08

Table 4:Reported AUCs and Standard Deviation of COVID-19 Predictions

The mean and standard deviation of the area under the curve of the receiver operating characteristic curve was recorded after each of the 100 different XGBoost classification models. Each run used a different set of people who did not have a severe reaction to COVID-19. The mean AUC for all three datasets was well described by a normal distribution, as confirmed by a Shapiro normality test.

Dataset	**Mean AUC**	**SD AUC**
1930 data	0.515	0.017
1940 data	0.516	0.019
1950 data	0.511	0.030

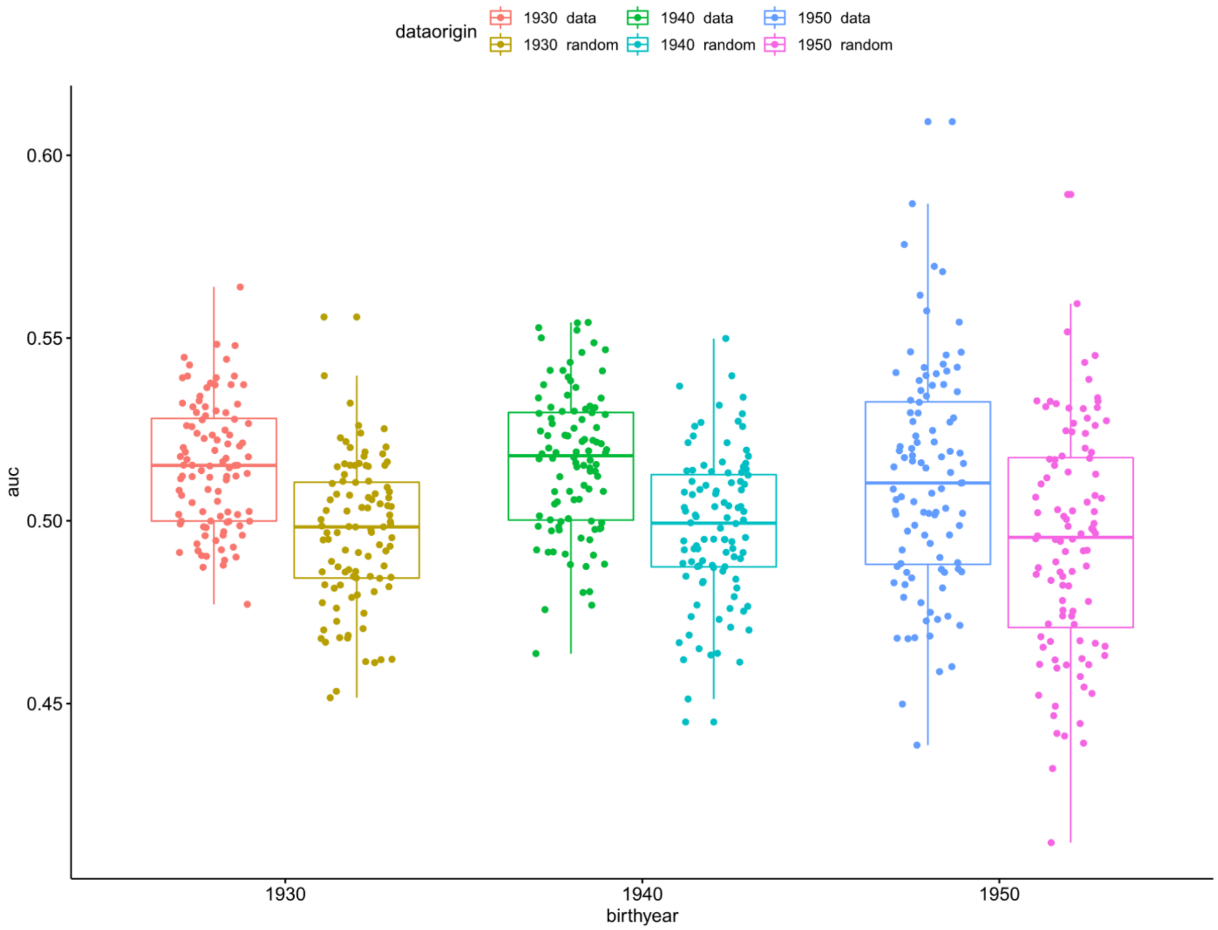


Figure 24: Boxplot of COVID-19 Model AUCs Compared to Random Model AUCs

This boxplot figure presents the results of the machine learning predictions. We created three different datasets, one which includes all patients less than 90 years old, the second includes every patient less than 80 years old, and the third with every patient less than 70 years old. These are indicated as the oldest birthyear "data." Each dataset included an equal number of patients with a "severe reaction" to COVID-19 and an equal number of age-matched people drawn from the general UK Biobank population, "normal." For comparison, we took those three datasets and randomly permuted the status ("severe reaction" or "normal) and repeated the process. This randomly permuted dataset is labeled oldest birthyear "random." For each dataset, we repeated the whole process 100 times, each time with a different set of age-matched people from the general UK Biobank population.

We arrive at two conclusions from these results. First, a genetic difference exists between those who have the most severe diagnoses of COVID-19 and the general population. Second, we were not able to utilize this difference to develop a clinically useful test to distinguish between individuals who experience a severe course of the disease and those who will not.

Though the AUC is low for this test, there are several reasons for this, and which, if addressed, can improve the existing AUC. Primarily the data we had available was a constrained by those listed in the UK Biobank as having as severe course of COVID-19. There are likely a substantial number of people who would also have a sever reaction to COVID-19, who at the time had not been diagnosed or had not contracted the virus. A better approach would be to compare patients with severe reaction, to those with mild or asymptomatic reactions. Since this study was done, it is likely that there is significantly more data and this would improve the overarching model AUC.

Changes in our feature selection and classification algorithm might also improve the AUC. Our feature selection algorithm that transformed “l2r” data into our final chromosomal-scale length variation data took averages over each quarter of a chromosome. We could instead include smaller chromosome segments. Generally, we need the number of features to be much less than the number of observations (patients). So, an increase in the number of observations would allow an increase in the number of features. Also, an alternative machine learning algorithm might improve the AUC. Different algorithms perform differently on different classes of problems and XGBoost generally performs well on tabular data[124]. We did a brief test of different algorithms before choosing XGBoost as the best solution for this problem. But, for instance, a deep learning algorithm might have better performance with proper tuning.

Our results add to the recent work done by others on the link between genetics and severity of COVID-19. For instance, one study from the Netherlands identified four young men from two different families who had severe symptoms of COVID-19 and no preexisting medical conditions. Detailed genetic studies revealed that these four men all had a rare loss of function variant of TLR7, which lies on the X-chromosome[125].

A detailed study of this UK Biobank COVID-19 dataset found that Black and Asian patients were at a significantly higher risk of testing positive compared to white patients[126]. This study also attempted to derive a polygenic risk score. However, when they applied the polygenic risk score to a hold-out group, they found that the mean score was indistinguishable between the group of people who had tested positive and the group that had no positive test. In comparison, our work found that these two groups are distinguishable with a genetic risk score, but only very slightly. We measured the AUC at 0.51. They[126] do not report an AUC, but an indistinguishable test is the equivalent of an AUC of 0.50.

Other more comprehensive metastudies have identified one specific genetic component behind the severity of COVID-19. For instance, one study of COVID-19 patients who experienced respiratory failure at seven hospitals in Italy and Spain found a fairly strong association in a cluster of genes lying on part of chromosome 3 and a borderline association in chromosome 9 encompassing the ABO blood group locus[127]. The “ANA_B2” June 2020 results posted by the COVID-19 Host Genetics Initiative[128,129] also indicate a strong association in chromosome 3 but fail to reproduce the association in chromosome 9. The COVID-19 Host Genetics Initiative “ANA_B2” study compares hospitalized COVID-19 patients to the general population and are mostly derived from patients in Europe and Brazil. Neither study attempted to derive a genetic risk score.

This study has several weaknesses. First, we cannot attribute the severity of COVID-19 to particular genetic variants. This study only finds correlations and does not establish a cause and effect. Second, while it is possible that these correlations relate to underlying biology, it is also possible that the correlations are related to ancestral differences that translate to socio-economic differences. COVID-19 severity is known to be correlated with racial/ethnic background[130,131]. The small effect that we measured might be simply due to the larger complex effect of racial/ethnic disparities in COVID-19 severity.

In conclusion, we found a significant difference exists between the structural genomics of those patients in the UK Biobank who had a severe reaction to the SARS-CoV-2 virus and the general UK Biobank population. However, a test based upon this difference would not be clinically useful in its present state since it had an AUC of 0.51.

Chapter 5.6: Genetic Risk Score for Ovarian Cancer Based on Chromosomal-Scale Length Variation

Ovarian cancer kills about 150,000 women per year worldwide[132]. 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[133]. 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[134–137]. 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[138]. The maximum discriminative accuracy of a genetic risk test is a function of both the heritability and the prevalence of the disease[139,140]. 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[12,141–145]. 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%[146].

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.

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[38,147,148]. 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[46], which included the packages Birdsuite[37] and DNAcopy[149]. 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[150]. 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[46]."

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[151]. 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[152]. 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[73]. For the results reported in Table 5, 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 Figure 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 7 and Figure 26, we estimated 95% confidence intervals for the odds ratios following the method described in[153].

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

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. 25. By convention, these lengths are reported in units of log base 2. A value of "0" represents the consensus, average, chromosome length.

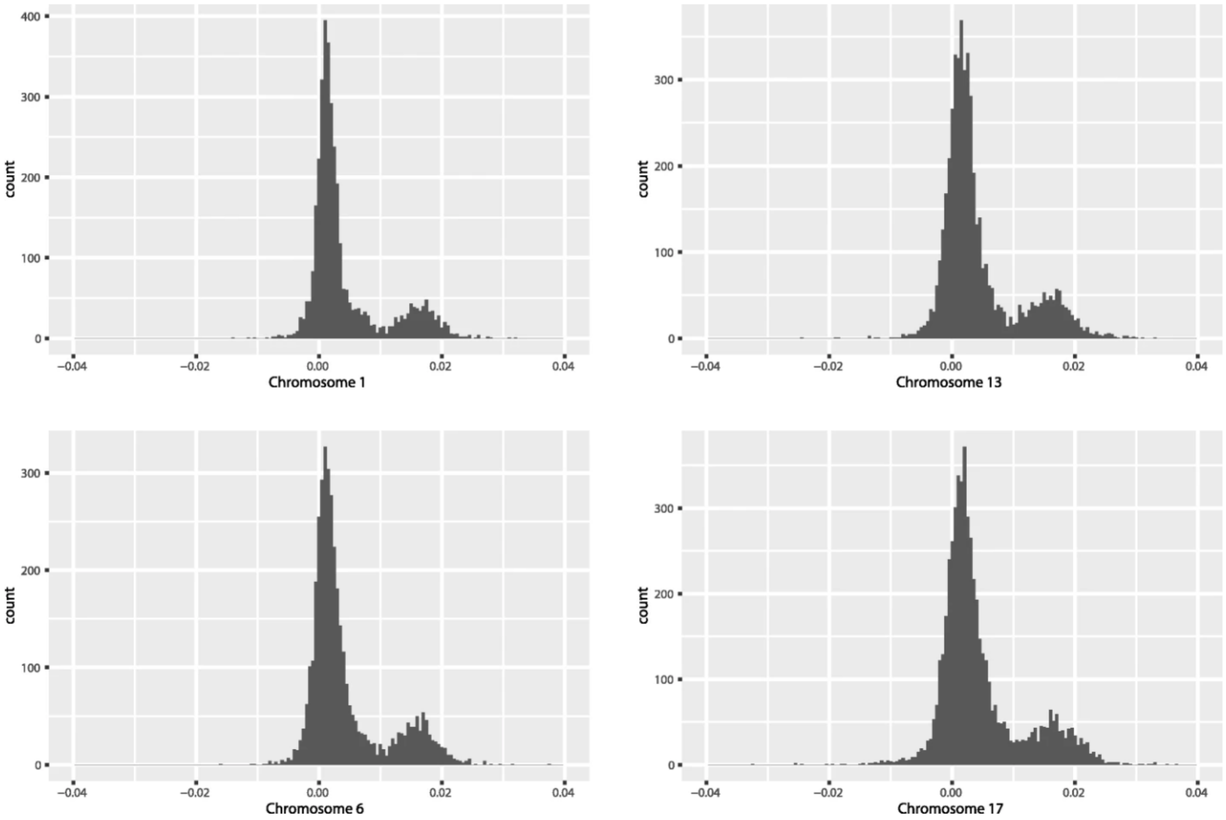


Figure 25: Histogram of CSLV length for Chromosomes 1,6, 7, and 13

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.

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 5.

Table 5: Table TCGA dataset group statistics

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 Serous Carcinoma**
Total	414	4225
Mean age	58.3	59.7
% Black	2/414 = 6%	492/4225 = 12%
% White	352/414 = 85%	3064/4225 = 73%
% Asian	14/414 = 3%	259/4225 = 6%

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 6.

Table 6: Comparison of Machine Learning Algorithms for Ovarian Serous Carcinoma Prediction

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

Based on the results in Table 6, 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.

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.

Table 7: Odds Ratio Quintiles of Ovarian Serous Carcinoma Predictions

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

The comparison is presented in two different forms. Table 7 provides a tabular form of relative risk for the population segmented into five different groups. Figure 26 shows similar information in graphical form, where the population is segmented into 50 groups.

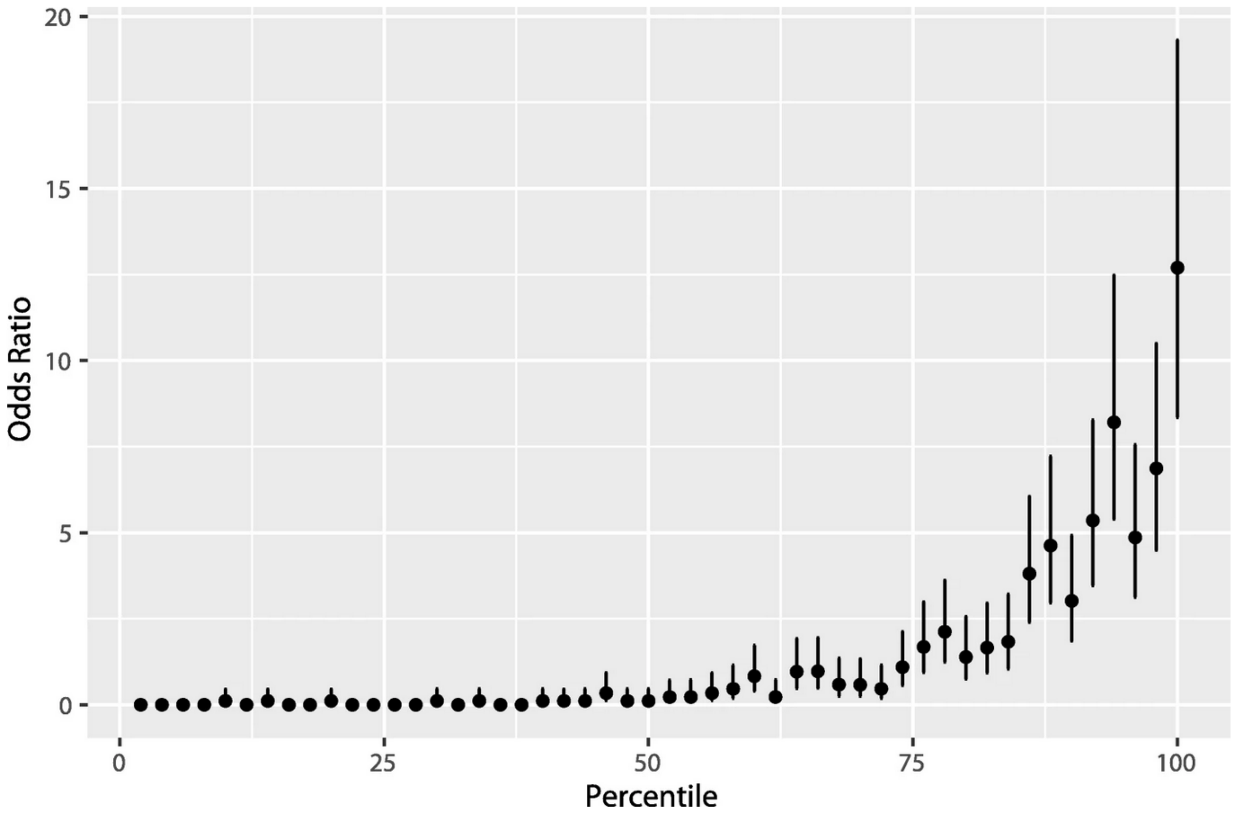


Figure 26: Odds Ratio Plot of 50 equal partitions of Ovarian Serous Carcinoma Model

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.

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 Figure 27, which shows a classic receiver operating characteristic curve of the model's predictions. Figure 28 presents the SHAP contribution plot, which helps explain how the Gradient Boosting Machine model arrives at its result.

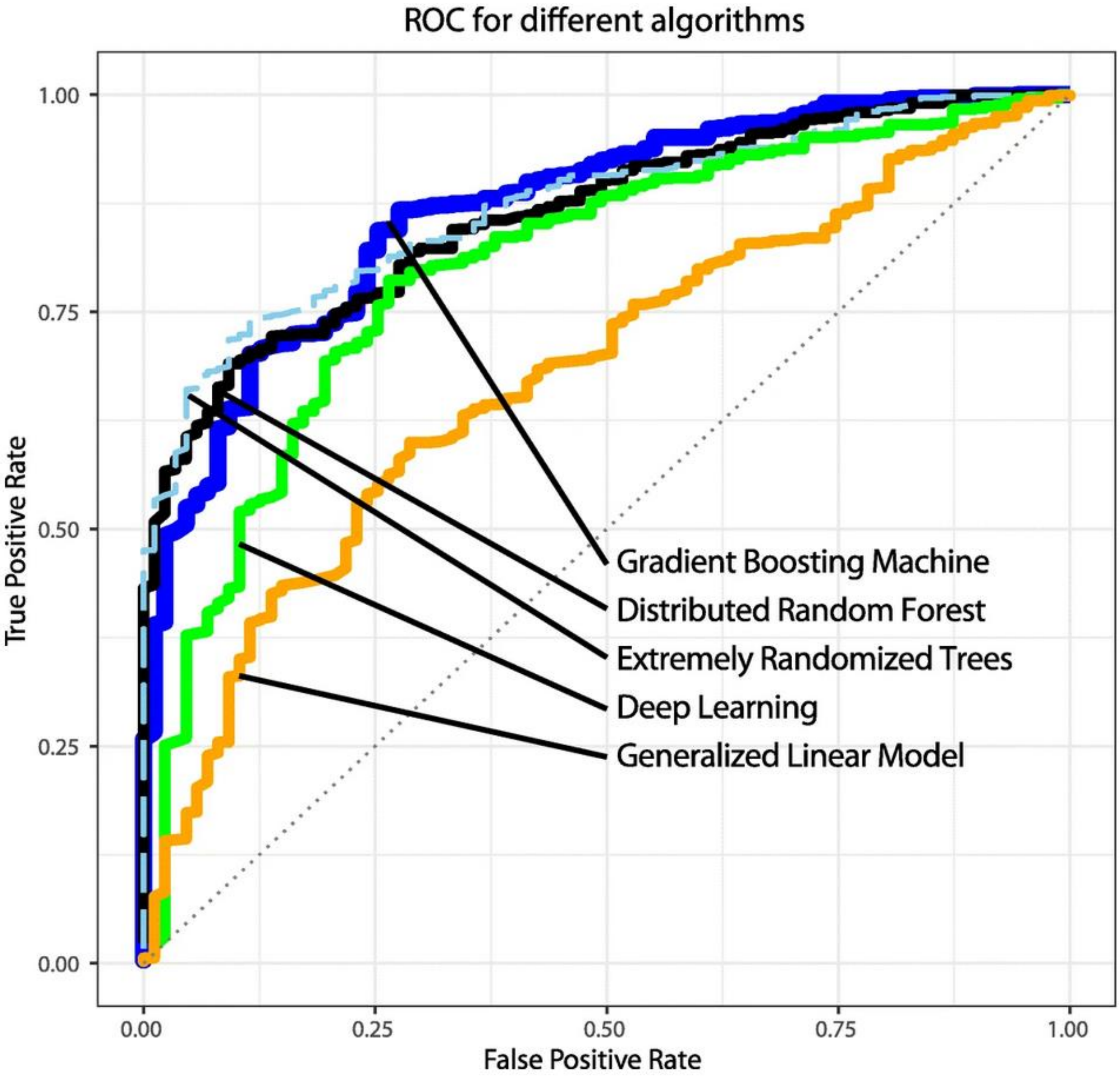


Figure 27: ROC of Algorithms in Predicting Ovarian Serous Carcinoma

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.

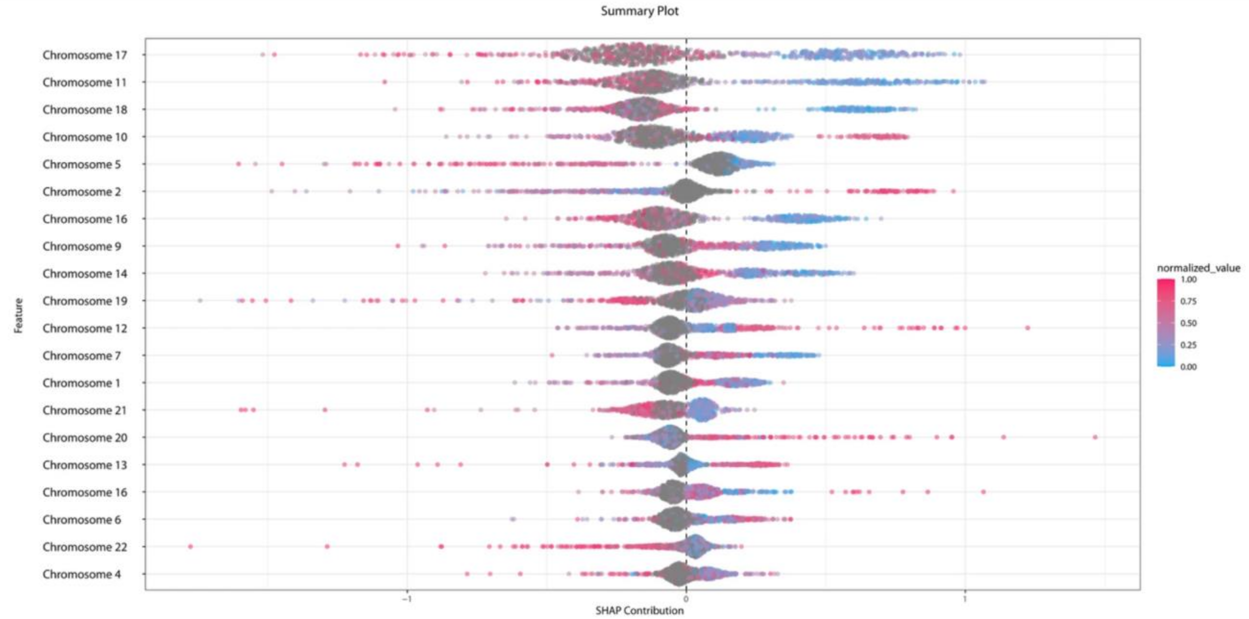


Figure 28: SHAP Contribution Plot of Predictive Model for Ovarian Serous Carcinoma

This SHAP contribution plot ranks the importance to the predictive model for each chromosome[154]*. 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.*

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%[146]. As seen in Table 7, the top 20% in our results had an increase of over 100-fold risk over women who scored in the bottom 20%.

Table 6 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[155]. 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[124]. 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)[156]. Second, most common germline BRCA mutations change the overall length by just a few bases out of the 81 million bases on chromosome 17[157]. 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[158]. 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 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[159]. 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[158].

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.

A genetic risk score based on chromosomal-scale length variation of germ line DNA could provide an effective means of predicting whether a woman will develop ovarian cancer. Several avenues are open to further improve the AUC of this genetic risk score test.

Chapter 5.7: Prediction of Schizophrenia in Individuals from the UK Biobank Indicates Significant X Chromosome Contributions

Schizophrenia is a highly heritable, complex psychiatric disorder[160,161]. Genome wide association studies have identified over one hundred genetic loci that contribute to its heritability[10,11,161–163]. However, these loci still account for less than half of the genetic risk for schizophrenia[11]. Environmental exposure to chemicals appears to play almost no role in the development of schizophrenia, but different forms of trauma experienced during development does appear to be a risk factor[164] . Twin studies have consistently shown a significant genetic contribution to schizophrenia, and many twin studies find that the environmental contribution to schizophrenia exists but that genetic effects provide significant liability to schizophrenia[165].

Genetic risk scores [141,166,167] have been developed for many different forms of disease, including breast cancer[168], coronary artery disease[169], and stroke[170]. Polygenic risk scores based on SNPs clearly can predict schizophrenia. One study measured an odds ratio of about 8 (95% CI 4-14) for the highest decile compared to the lowest decile[171]. A second study found that polygenic risk scores for schizophrenia (and bipolar disorder) are also associated with creativity[172]. A review of polygenic risk scores for schizophrenia highlighted the difficulty these studies had finding a consistent diagnosis of schizophrenia[173].

Copy number variations (CNVs) and copy-neutral loss of heterozygosity (CN-LOH) have been implicated in significant clonal selection[174]. We have previously shown that chromosome-scale length variation is a powerful tool to predict phenotypes from a person's genome [123]. This method is particularly appealing for genetic risk scores because it includes epistatic effects that might be missed with conventional genome wide association studies, which use logistic regression—a linear combination of SNP scores.

We aimed to evaluate how well a genetic risk score based on chromosome-scale length variation and machine learning classification algorithms can predict schizophrenia in individuals. We evaluated this approach on a dataset of 1129 patients who had schizophrenia in the UK Biobank dataset. These patients were previously genotyped as part of the UK Biobank project.

Figure 29 presents results showing the performance of different machine learning algorithms. We found that the stacked ensemble models consistently performed best. As Figure 30 shows, we found a slight difference between algorithms and their performance. But all algorithms could predict schizophrenia significantly better than chance (AUC=0.50). This finding indicates that germ line genetics of the patient, as represented by the set of chromosome-scale length variation numbers, demonstrates predictability of schizophrenia.

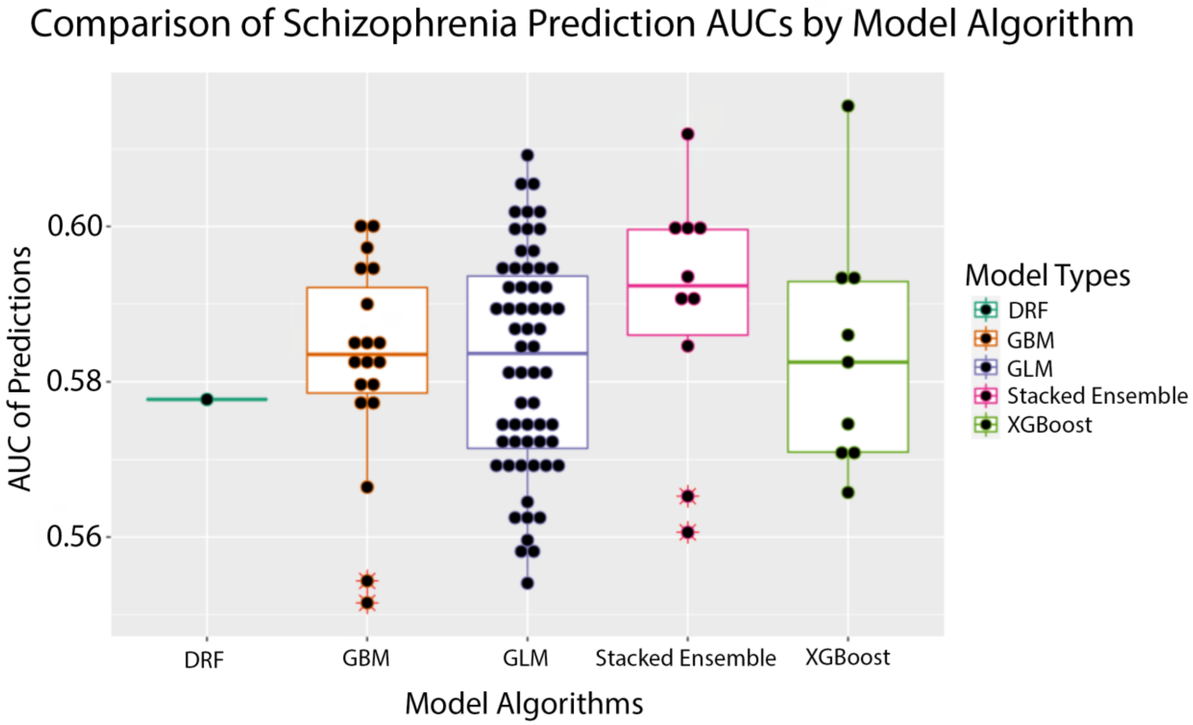


Figure 29: Comparison of Schizophrenia Prediction AUCs by Model Algorithms

This boxplot figure presents the results of the machine learning predictions. We created 100 different datasets. For each dataset, we used the same set of schizophrenia patients with a distinct set of age matched people from the general UK Biobank population as controls. Then H2O was used to perform a grid-search of possible best algorithms. The best performing algorithm was then reported with an AUC. The differences between algorithms are reported here. The machine learning algorithms tested were distributed random forests (drf), gradient boosting machine (gbm), general linear model (GLM), stacked ensemble (a combination of the other four algorithms) and XGBoost (XGBoost).

The AUC (area under the curve of the receiver operating characteristic curve) for the machine learning classification models was 0.583 (standard deviation 0.014, 95% confidence interval of 0.581-0.586). A classification model with an AUC of 0.50 is equivalent to random guessing. The measured AUC differs from 0.50 with $p<0.00001$.

We also tested how well each model could predict schizophrenia on a holdout set of validation data. The holdout set was 30% of the original test data and was not included in the training of the models. The AUC of the holdout set was 0.5734 with a 95% confidence interval of 0.569-0.578.

We also tested whether increasing the number of splits improves model performance. We constructed three overlapping datasets with 1 split, 4 splits, and 8 splits. The phrase “1 split” represents the average l2r value measured across an entire chromosome for all 23 chromosomes giving a total of 23 numbers, “4 splits” represents the average of each quarter of the 23 chromosomes l2r values for a total of 92 numbers, and “8 splits” represent the average of each eighth of the 23 chromosomes’ l2r values for a total of 184 numbers.

Figure 30 shows how models compare on the 3 different split datasets. Overall, a stacked ensemble had the best performance, however a general linear model (glm) was most often the best candidate model.

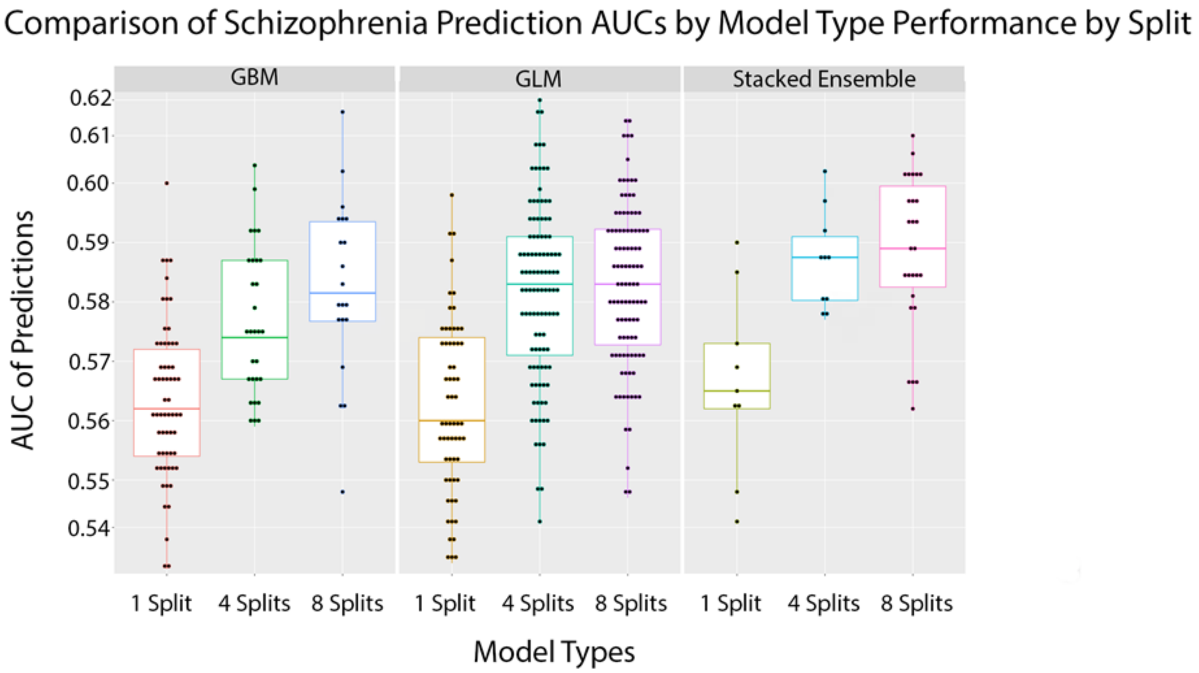


Figure 30: Comparison of Schizophrenia Prediction AUCs by Model Type Performance by Splits

We tested whether finer splits of each chromosome lead to better predictability. We split each chromosome into either one, four, or eight subsections. We computed the chromosome scale length variation for each of these subsections for each person. This set of numbers was used to predict whether patients had schizophrenia. The quality of this prediction was characterized by the AUC. This plot demonstrates how the quality of these predictions increase with finer information on chromosome length variation. The Stacked Ensemble algorithm performs the best across all split variations.

In all models, increasing splits improves model performance for the same runtime. Figure 3 demonstrates the difference of all models for 1 split, 4 splits, and 8 splits datasets. We tested whether finer splits of the dataset provided significantly improved AUCs. As shown in Table 8, the p-value of the 4 splits model compared to the 1 split model is $p = 1 \times 10^{-24}$. Comparing the mean AUC for the 8 splits model to the 1 split model gave a p-value of $p = 3 \times 10^{-30}$ indicating that finer splits significantly improved the predictive ability of the models.

The 4 splits and 8 splits models performed better than the 1 split models by a significant amount.

Table 8: Table of Comparing AUC by CSLV Splits.

The mean and standard deviation of the cross-validated AUCs of 1split, 4 splits, and 8 splits datasets of 150 models for each.

Dataset	Mean AUC	Standard Deviation	P-value vs 1 split
1 split	0.5614	0.0148	
4 splits	0.5807	0.0146	1×10^{-24}
8 splits	0.5838	0.0141	3×10^{-30}

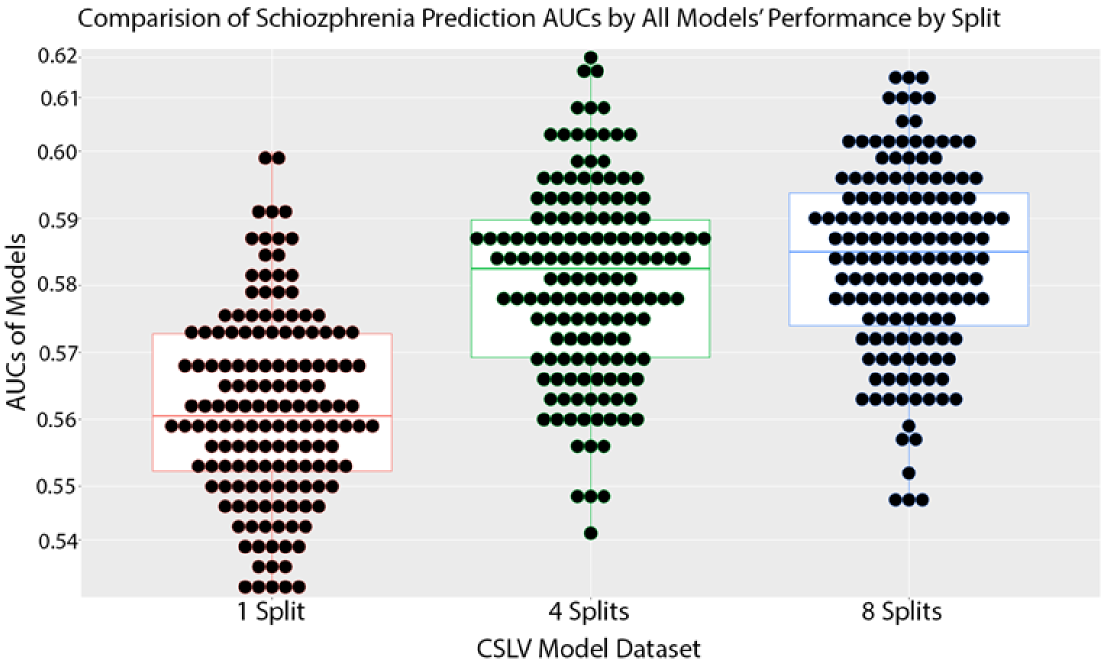


Figure 31: Comparison of Schizophrenia Prediction AUCs of All Models by CSLV Splits

This plot represents the average performance of 150 models for each type of CSLV split for a total of 450 models.

We then calculated the odds ratio (OR) of our predictions drawn from the cross-validated model. Table 9 shows that a patient in the upper quintile is approximately twice as likely to have schizophrenia when compared to the lower quintile. The odds ratio of the upper is 1.3 compared to the lower which is 0.67 thus giving us $\frac{1.3}{0.67} = 1.94$.

Table 9: Odds Ratio of Schizophrenia by Quintiles

This table represents the odds ratio between the quintiles of predicted results from our cross-validated results. The result indicates that the top quintile is twice as likely to have an accurate prediction for Schizophrenia as the bottom quintile.

Quintile	Normal	Schizophrenia	Odds Ratio	Count	95% CI
1	185	123	0.67	308	0.51-0.85
2	156	152	0.97	308	0.76-1.24
3	153	155	1.0	308	0.79-1.3
4	142	165	1.2	307	0.91-1.5
5	133	174	1.3	307	1.0-1.7

In order to understand, how our models came to their conclusions we created several plots to explain them from H2O's "explainability" framework. The first is a variable importance heatmap across the generated models which is shown in Figure 32. Our analysis here indicated that chromosome X was one of the highest contributing variables in predicting Schizophrenia, especially in tree models such as GBM and XGBoost. We then confirmed this with a Shapley Additive exPlanation or SHAP plot in Figure 33. This plot also indicates that chromosome X was the leading factor in our leading model for predicting schizophrenia.

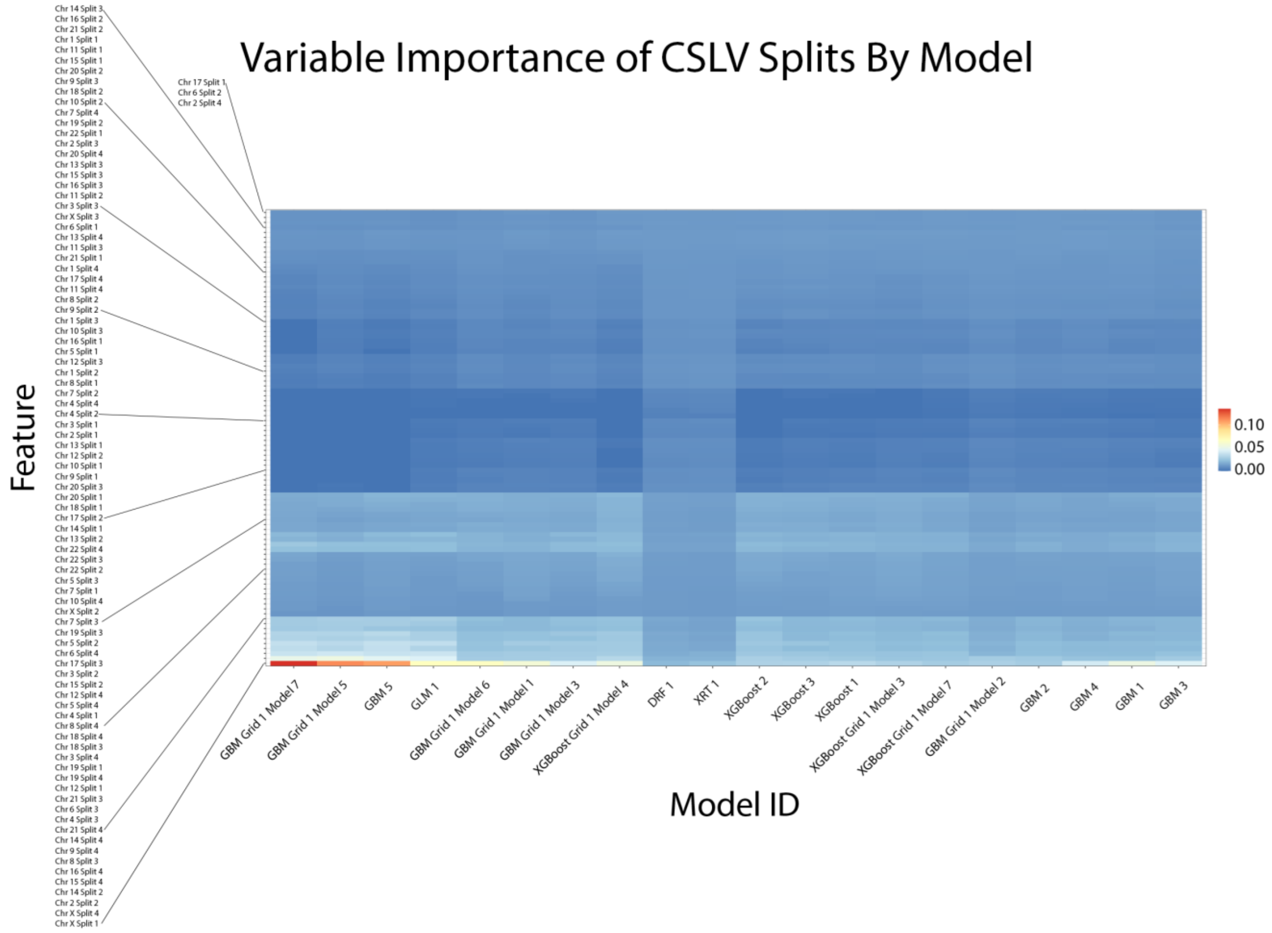


Figure 32: Variable Importance Heatmap of 4 Split Schizophrenia CSLV Models

This variable importance heatmap shows the variables which most affected the performance and outcome of decisions made by the specified model. A value closer to 1.0 indicates higher importance of that variable. In most tree-based models the CSLV values for chromosome X have the highest importance.

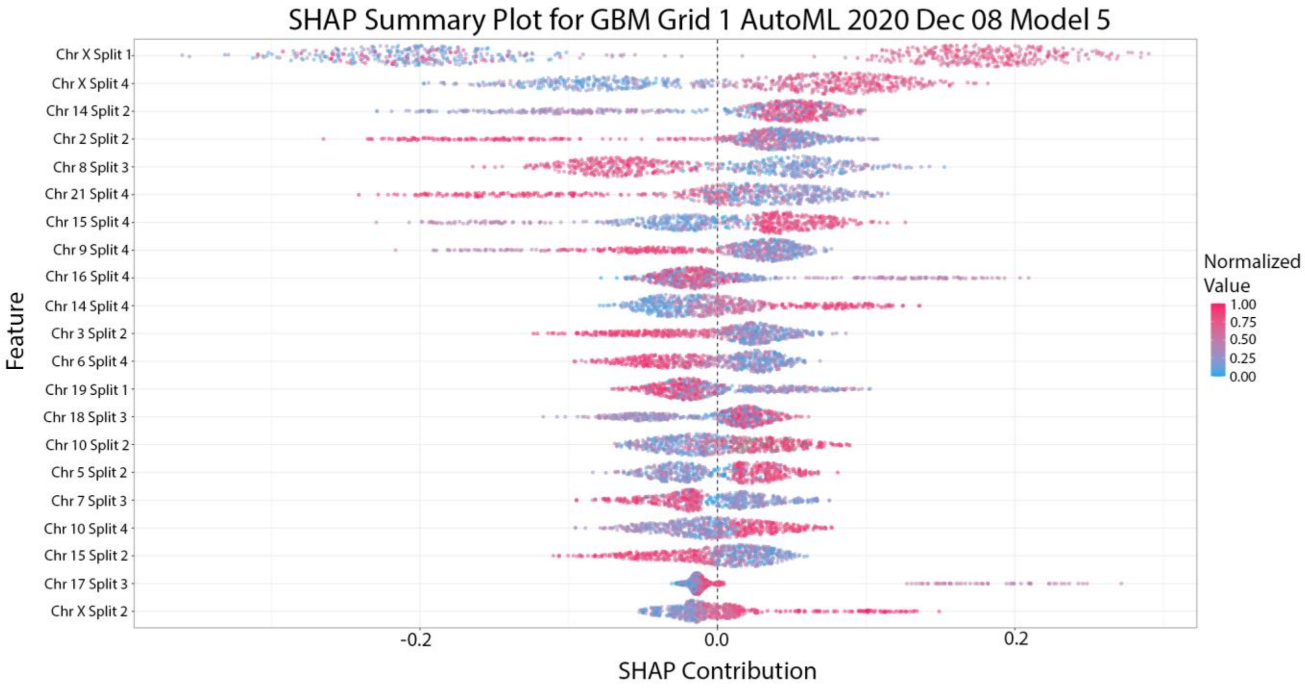


Figure 33: SHAP Plot of Leading 4 Split Schizophrenia CSLV Model

This SHAP plot indicates that the leading model for our 4-splits model relied heavily on the first quarter and last quarter value of chromosome X with some contribution from other regions and the second quarter of chromosome X.

Utilizing our findings above, we then proceeded to train models using only CSLV values from chromosome X but with 64 CSLV splits. This model did not contain any information from the 22 autosomes but instead relied solely on CNVs in the X chromosome and our aim was to see if the model would be comparable to our previous 4-split and 8-split models. We found that on average these models had a comparable performance of about 0.58 with the highest being around 0.627 as shown in Figure 34.

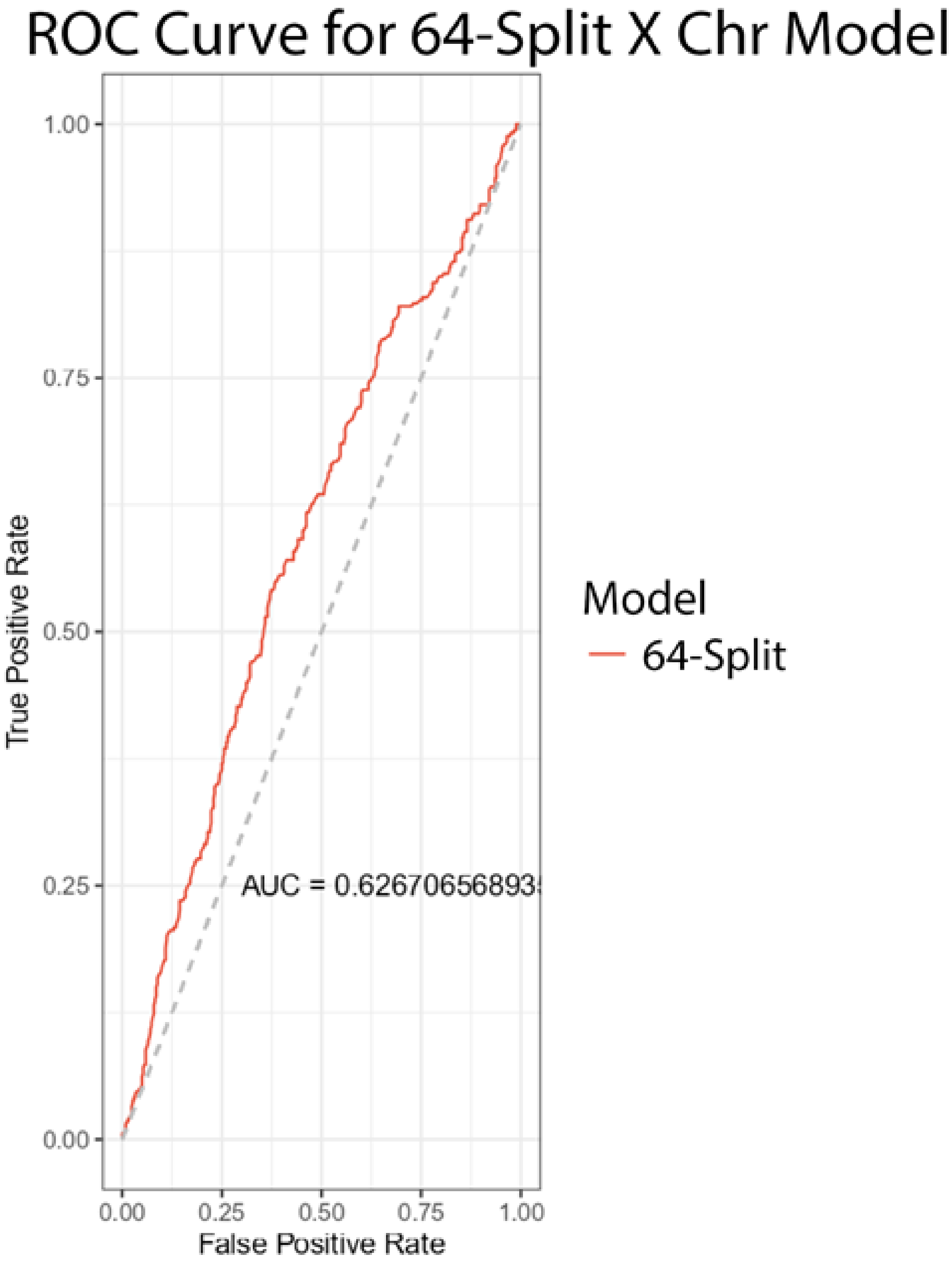


Figure 34: ROC Curve for 64-Split X Chromosome Schizophrenia Model

This ROC Curve for a schizophrenia prediction model utilizing 64-splits or 64 CSLVs of chromosome X only. The reported AUC is 0.627.

We then again performed a variable importance heatmap analysis to get greater granularity of our understanding of the contributing CSLVs in chromosome X. We found that this was again consistent with the previous findings from the 4-split model. Figure 35 indicates that the top features of variable importance are again being found in the first and last regions of chromosome X. As such it appears that much of the predictive power of any model trained with CSLV and when predicting schizophrenia in an individual is a result of CNVs on chromosome X. We also report corresponding estimates of hg38 coordinates in Table 10.

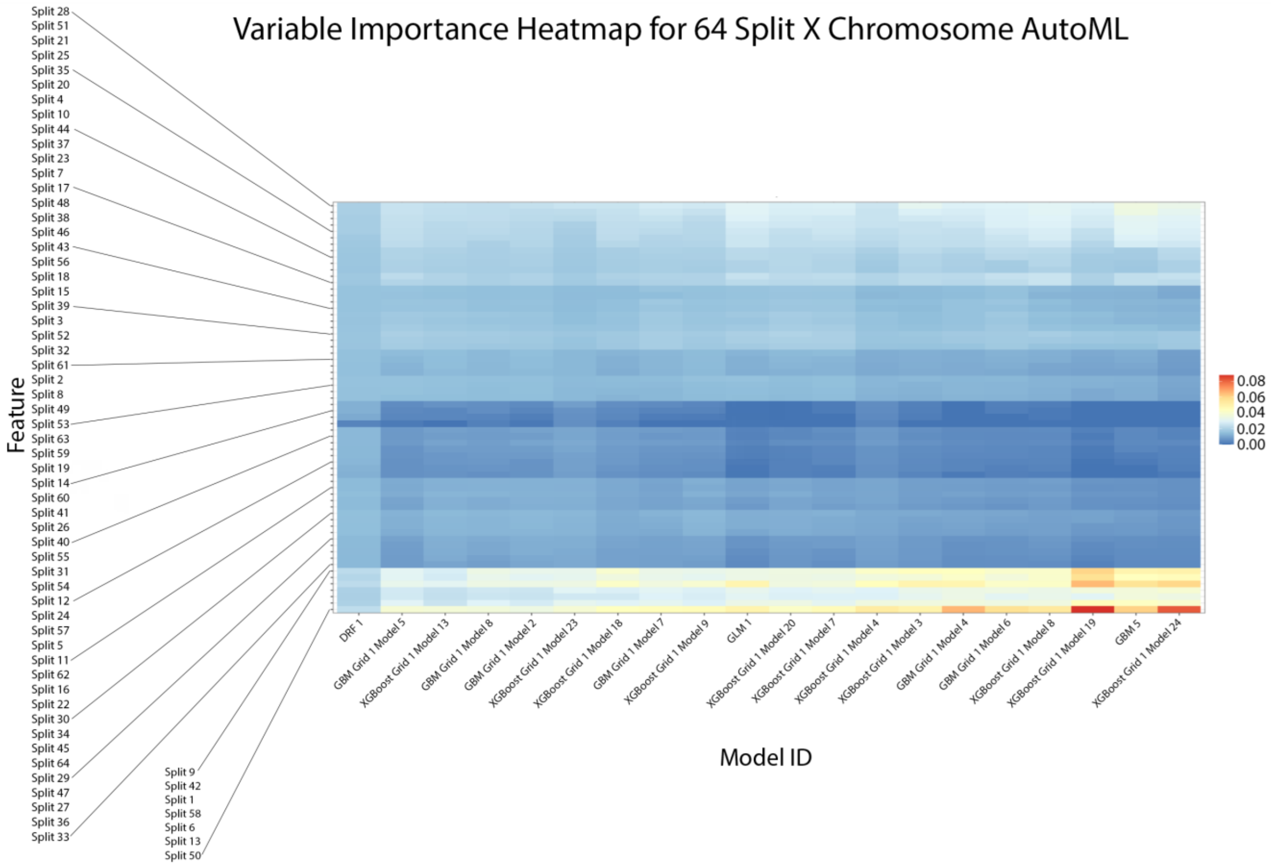


Figure 35: Variable Importance Heatmap of 64 Split X Chromosome Schizophrenia Models

This variable importance heatmap shows the variables which most affected the performance and outcome of decisions made by the specified model. A value closer to 1.0 indicates higher importance of that variable. In most of the models we find that the CSLV values were mostly centered around split 50, 1, 9, 42, 13, 58, and 6. This is consistent with Figure 12.

Table 10: Estimated hg38 Coordinates for High Variable Importance CSLVs in Schizophrenia Prediction

This table shows the estimated hg38 coordinates for the corresponding CSLV splits with high variable importance as shown in Figure 35.

CSLV Split	Estimated hg38 Coordinates
1	chrX:60425-634774
6	chrX:5651118-7792613
9	chrX:11426091-13234434
13	chrX:20912585-22990332
42	chrX:107331058-110669244
50	chrX:128031497-130523635
58	chrX:145709120-147908169

We wanted to ensure these results were not due to inherent sex differences. We trained 50 models using the 64 split chromosome X dataset which were not only age-matched with the controls but also sex-matched. 25 of the AutoML models were trained with the actual data with correctly labeled disease states. The other 25 AutoML models were trained with the schizophrenia diagnosis randomly shuffled. The results are shown in Table 11. Here we can see that a portion of the previous performance is most likely due to CSLV differences inherent between males and females (Figure 36). However, a portion of the prediction is statistically still better than random guessing.

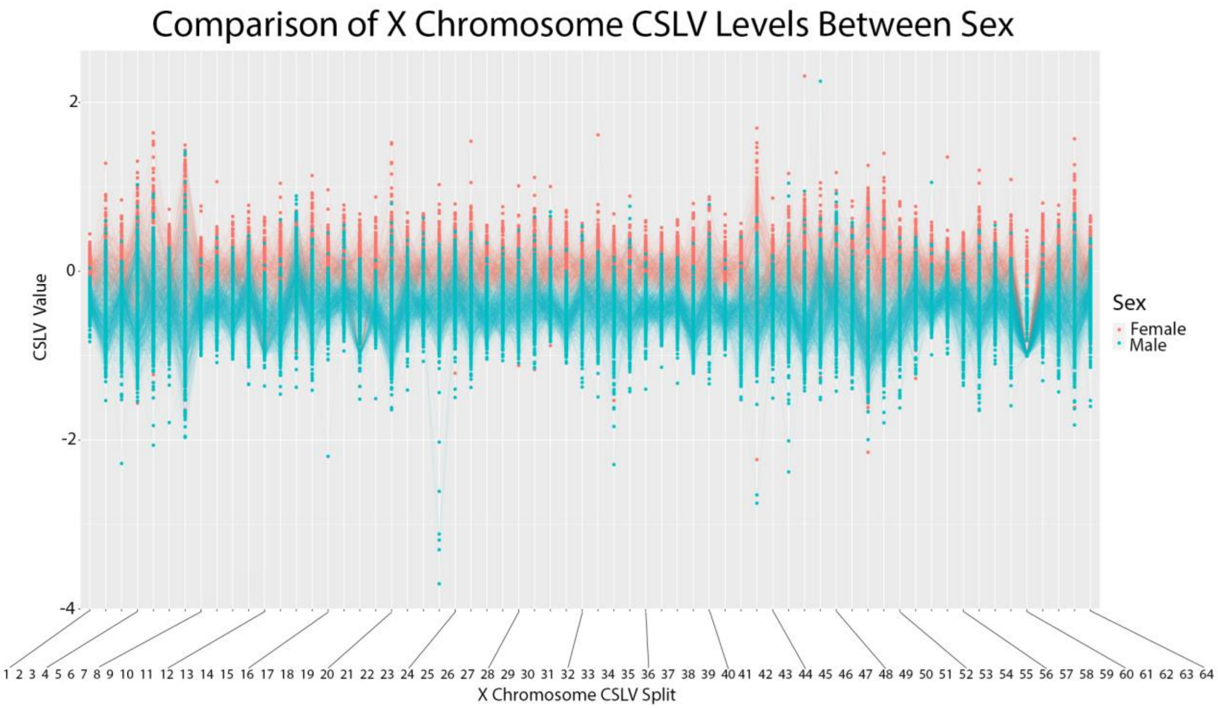


Figure 36:Comparison of X Chromosome CSLV Levels Between Sex

This plot compares the X Chromosome CSLV values of schizophrenia patients by sex. It contains all patients from the UK Biobank with a diagnosis of schizophrenia. In general females have higher average CSLV values than males. This is consistent with the fact that females have two X chromosomes while males only have one X chromosome.

Table 11: Comparison of Age & Sex Matched 64 Split Chromosome X Schizophrenia Models with Randomized Models

This table shows a comparison of the age and sex matched models using 64 Split chromosome X data. The reported mean AUCs demonstrates that a portion of the previous performance is attributed to differences between male and females in X Chromosome CSLV le

Dataset	**Mean AUC**	**Standard Deviation**
64 Split X Normal	0.545	0.01373103
64 Split X Random	0.525	0.01363745

Welch Two Sample t-test Between Normal and Random	T = -5.0111
	df = 47.998
		p-value = 7.763e-06

These results indicate that germline genetic variation contributes at least to some degree to the onset of schizophrenia in individuals. Our results indicate that genetic structural variation across the global chromosomal scope is sufficient to predict, better than guessing, whether an individual will have schizophrenia. The patients were an equal number of patients by gender between the control and disease group and the ages of patients in the control group also were matched to the ages of patients in the disease group. Further analysis revealed that length variation in a handful of regions of the X chromosome was sufficient to reproduce the predictive model. Recently, there has been revived discussion of copy number variations as a large contributing factor to several neurological ailments including schizophrenia [175]. Additionally, hypotheses about sex chromosome links to schizophrenia inheritance have been discussed for several decades and our findings lend support to this idea [176].

On average, a stacked ensemble is the best approach to creating a predictive model for the prediction of schizophrenia. However, all models that were tested still created models with predictive power better than chance (Appendix 3-5). Since H2O's AutoML performs a grid-search of all the possible datasets and each trial we ran included the same disease group but with different control groups, we can see in Figure 9 that a general linear model (GLM) oftentimes provided the best immediate performance. Gradient Boosted Machines (GBM) and XGBoost also typically performed the same as GLM. When a Stacked Ensemble did work well, it was significantly better than all other algorithms.

Utilizing a more granularized dataset by splitting the autosomes into quarters and eighths performs significantly better than using a CSLV averaged across an entire chromosome. This observation suggests we can increase performance by increasing splits. In the future, we plan on exploring the tradeoff in run time and computational resources required by increasing splits.

The CSLV values are averages of copy number variation (CNV) measured at each SNP location. Simply using every single CNV value introduces a dimensionality problem as our dataset only has roughly 488,000 individuals while the total number of CNV values is 764,257 across the 22 autosomes and an additional 18,857 CNV values for the X Chromosome. This means there is likely diminishing returns for using more splits unless it can be offset with increased data.

This approach has several limitations. First, CSLV is an averaged measure of copy-number variations across a large section of the entire chromosome. We used SHAP values to highlight the regions that seem to be more important, but this does not provide a mechanistic explanation. Second, the dataset lacks diversity. The UK Biobank population is primarily Caucasian individuals in the United Kingdom (although not exclusively). Finally, the diagnosis of schizophrenia in an individual is difficult to quantify and the disease might consist of a heterogeneous group of underlying biological processes.

We were able to create machine learning models for prediction of schizophrenia in patients. These models perform better than chance with an average AUC of 0.545. Prediction was performed with only chromosomal scale length variation measurements as the input variables. Further analysis of the SHAP values suggests that the length variation of several regions of the X chromosome are sufficient to reproduce this predictive value.

Chapter 6: Significance of Work

The totality of how hereditary factors contribute to complex diseases like cancers and schizophrenia remains unclear. Our understanding of these diseases has grown tremendously in the past decade; however, we do not understand a large portion of the specific mechanisms for disease onset and risk. Challenges exist in separating the genetic information that contains mutations occurring due to environmental factors from the genetic information that contains mutations passed from generation to generation.

Current GWAS methods often focus on somatic SNPs without considering epistatic interactions greater than two SNPs. Whether these SNPs are the main cause of the genetic process causing cancers, or are simply the largest contributors from a greater group of SNPs, remains undetermined. Germline CNVs hold a wealth of information that we have utilized to understand how genes interact in conjunction with a broader network of variants to affect disease ontology. This study of germline CNVs provides insight into both the epistatic interactions between genes and highlights the degree hereditary factors contribute to specific cancers. We have also demonstrated that schizophrenia has risk factors in latent CNVs which could be determined and exploited to predict a person's risk for the disease.

Global CNVs implicate higher risk for certain complex diseases

This study has demonstrated that CNVs likely act within an individual person's genome in a networked fashion. Globally, CNV levels can be used to determine risk of diseases. Since CNVs are an inherited feature of the structural genomics, the possibility of utilizing this knowledge to better assess and diagnose diseases such as cancer and schizophrenia is promising. In terms of cancer, we have demonstrated that CSLVs can be used to distinguish between different types of cancer. This is beneficial for tests that rely on blood or serum-based diagnostics. This could make diagnosing specific types of cancer from blood tests alone a reality soon. Additionally, support for the idea of the epigenome playing a large role in cancer onset is further bolstered by our findings here.

Schizophrenia is most likely inherited to some degree by global X Chromosome CNVs. As a neurological and psychiatric disease, studying the exact cause of this disease is always difficult. Clinical and ethical issues with understanding brain related diseases have always been a concern and rightly so. There have been hypotheses of inherited risk for schizophrenia for some time. Our findings show that there is indeed a sex chromosome link to schizophrenia when it comes to CSLVs on the X chromosome. This will aid in more targeted studies which could help us understand the underlying genetic mechanisms for schizophrenia.

Chromosomal Scale Length Variation is an effective method to utilize CNVs for study of complex diseases

CSLV demonstrates that a global variation of CNVs is potentially an important factor for inheritance of disease. CSLV utilizes many CNVs in an efficient way, while still maintaining the relevant information. Using machine learning methods, we can perform this analysis with CSLVs to gain useful biological insight in a reasonable amount of time and computational resources.

CSLV demonstrates risk factors of CNVs across the global genetic landscape. This means that isolated CNVs are unlikely to be the single contributor to a disease. CNVs are likely working in concert with each other in a highly connected and dependent network of interactions. Unraveling and understanding this network will most likely demonstrate a multifactorial problem. This problem has a complexity of an order much higher than our previous SNP approaches to disease understanding.

A comparison of the AUC for risk scores vs. current published risk scores.

In order to understand the relevance of our findings it is best to compare with other reported genetic risk scores. Risk scores are typically reported as one of two metrics: 1. The AUC of the receiver operator curve, or 2. Odds Ratio between upper and lower groups. As shown in Table 12, our studies have demonstrated that using CSLV as a genetic risk score for complex disease performs comparably to previously reported risk scores. As such, we believe that CSLV is a promising feature in studying complex diseases which may have inherited risk from germ line genetics.

Table 12: Comparison of Genetic Risk Scores

A comparison of risk scores based off CSLVs and other reported risk scores from literature for the same disease. The odds-ratio (OR) is for the upper quintile unless otherwise indicated. The AUCs are for also given for the predictive effectiveness of the models. Values are for 95% confidence interval unless otherwise indicated.

**This reported Odds Ratio only applied to patients with bloody type B and did not work with other blood types.*

Disease	CLSV AUC	CLSV OR	Other AUC	Other OR
Breast Cancer (BRCA)	0.73[177]	1.98	0.63[168]	1.61[168]
Ovarian Cancer (OV)	0.88[178]	4.76	0.64[179]	3.4[146]
Glioblastoma Multiforme (GBM)	0.81[180]	3.78	0.719[181]	6.91*[182], 2.12
Schizophrenia	0.62[183]	1.94	0.56[184]	1.93[185]

Machine Learning applications can help provide insight into future targeted CNV research

There are many variations of machine learning techniques that can still be explored when utilizing CSLVs. We have studied a large variety of the most common and best performing machine learning algorithms. However, this is by no means an exhaustive study. There are still many variations and iterations which can still be examined. Some may improve our results, especially in the area of more specific hyperparameter tuning.

In terms of CSLVs, there are several areas which may be of interest. These include better feature selection in terms of the length and size of the CSLVs. Specifying CSLV areas may also improve our results. Another possibility is the calculation of the CSLV value itself. Currently, we are calculating a simple value average, however this average could be calculated as an average based on genomic location rather than number of CNV values. Or standard deviation or some other metric could be used to calculate the CSLVs.

Finally, increasing data is the most likely to improve our results. Even with TCGA and UK Biobanks, the disease set is still well below 5,000 individuals. This means we have a large control set, but the actual number of patients with the diseases in question is still quite low when compared to the data in a particular genetic database. These databases will likely only increase in size and scope and as other countries begin their own sample collection, augmenting results from one database with another will greatly improve results.

Chapter 7: Summary

Genetic medical research is currently in an exciting era. It is our opinion that the collision of informatics, biology, engineering, chemistry, and computer science will rapidly accelerate our knowledge of both hereditary and environmental factors contributing to the onset of later life and age dependent diseases. This study shows the potential of utilizing copy number variations in the form of chromosomal scale length variations in the prediction of complex diseases along with the promise of utilizing machine learning techniques to create an interpretable method of understanding how the genomic landscape interlinks across genes to contribute to inherited disease risk.

The Cancer Genome Atlas and UK Biobank are invaluable resources, providing high statistical power to our analysis. As other large-scale population data projects near completion in the coming decade, the methods laid on the foundation of The Cancer Genome Atlas and UK Biobank will continue to benefit and improve as sample sizes easily begin to move into the regime of millions of patients. Examining populations around the world will truly aid in the goal of precision medicine.

We believe we can apply our methods to other complex diseases such as Alzheimer's disease and dementia, asthma, and autoimmune diseases. Information about how human genetic variation can contribute to individual susceptibility allows patients and doctors to make early lifestyle changes in a preventative manner. Likewise, it can inform physicians which types of prognostics and diagnostics would be the most relevant for a specific patient, saving both time and money, while improving patient outcomes in the long term.

References

1. Schatz MC, Langmead B. The DNA data deluge. *IEEE Spectr*. 2013;50(7):28-33. doi:10.1109/MSPEC.2013.6545119

2. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*. 2017;101(1):5-22. doi:10.1016/j.ajhg.2017.06.005

3. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109(4):1193-1198. doi:10.1073/pnas.1119675109

4. Crawford L, Zeng P, Mukherjee S, Zhou X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet*. 2017;13(7). doi:10.1371/journal.pgen.1006869

5. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet*. 2016;17(6):333-351. doi:10.1038/nrg.2016.49

6. Toh C, Brody JP. *Chapter Applications of Machine Learning in Healthcare*. www.intechopen.com.

7. Yuan Y, Shi Y, Li C, et al. Deepgene: An advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics*. 2016;17. doi:10.1186/s12859-016-1334-9

8. Stadler ZK, Thom P, Robson ME, et al. Genome-wide association studies of cancer. *J Clin Oncol*. 2010;28(27):4255-4267. doi:10.1200/JCO.2009.25.7816

9. Galvan A, Ioannidis JPA, Dragani TA. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends Genet*. 2010;26(3):132-141. doi:10.1016/j.tig.2009.12.008

10. Ripke S, O'Dushlaine C, Chambert K, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet*. 2013. doi:10.1038/ng.2742

11. Lee SH, Decandia TR, Ripke S, et al. Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat Genet*. 2012. doi:10.1038/ng.1108

12. Khera A V., Chaffin M, Aragam KG, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018;50(9):1219-1224. doi:10.1038/s41588-018-0183-z

13. Mavaddat N, Pharoah PDP, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. *J Natl Cancer Inst*. 2015;107(5). doi:10.1093/jnci/djv036

14. Jonas KG, Lencz T, Li K, et al. Schizophrenia polygenic risk score and 20-year course of illness in psychotic disorders. *Transl Psychiatry*. 2019;9(1). doi:10.1038/s41398-019-0612-5

15. Bardenet R, Brendel M, Kégl B, Sebag M. Collaborative hyperparameter tuning. In: *30th International Conference on Machine Learning, ICML 2013*. ; 2013.

16. Lu D, Song J, Lu Y, et al. A shared genetic contribution to breast cancer and schizophrenia. *Nat Commun*. 2020;11(1). doi:10.1038/s41467-020-18492-8

17. Toh C, Brody JP. Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. *Hum Genomics*. 2020;14(1):36. doi:10.1186/s40246-020-00288-y

18. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet*. 2005;77(1):78-88. doi:10.1086/431652

19. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature*. 2015;521(7553):452-459. doi:10.1038/nature14541

20. Zhu Y, Xu Y, Helseth DL, et al. Zodiac: A Comprehensive Depiction of Genetic Interactions in Cancer by Integrating TCGA Data. *J Natl Cancer Inst*. 2015;107(8). doi:10.1093/jnci/djv129

21. Leiserson MDM, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet*. 2015. doi:10.1038/ng.3168

22. Cline MS, Craft B, Swatloski T, et al. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci Rep*. 2013;3. doi:10.1038/srep02652

23. Dauchel H, Lecroq T. Findings from the Section on Bioinformatics and Translational Informatics. *Yearb Med Inform*. 2017;26(01):188-192. doi:10.15265/iy-2016-050

24. Krasnov GS, Dmitriev AA, Melnikova N V., et al. CrossHub: A tool for multi-way analysis of the Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms. *Nucleic Acids Res*. 2016. doi:10.1093/nar/gkv1478

25. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437-455. doi:10.1146/annurev-med-100708-204735

26. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004;36(9):949-951. doi:10.1038/ng1416

27. Sismani C, Koufaris C, Voskarides K. Copy number variation in human health, disease and evolution. In: *Genomic Elements in Health, Disease and Evolution: Junk DNA*. ; 2015:129-154. doi:10.1007/978-1-4939-3070-8_6

28. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*. 2013;45(10):1134-1140. doi:10.1038/ng.2760

29. Kuusisto KM, Akinrinade O, Vihinen M, Kankuri-Tammilehto M, Laasanen SL, Schleutker J. Copy Number Variation Analysis in Familial BRCA1/2-Negative Finnish Breast and Ovarian Cancer. *PLoS One*. 2013;8(8). doi:10.1371/journal.pone.0071802

30. Huang K lin, Mashl RJ, Wu Y, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell*. 2018;173(2):355-370.e14. doi:10.1016/j.cell.2018.03.039

31. Krepischi ACV, Pearson PL, Rosenberg C. Germline copy number variations and cancer predisposition. *Futur Oncol*. 2012;8(4):441-450. doi:10.2217/fon.12.34

32. Kuiper RP, Ligtenberg MJL, Hoogerbrugge N, Geurts van Kessel A. Germline copy number variation and cancer risk. *Curr Opin Genet Dev*. 2010;20(3):282-289. doi:10.1016/j.gde.2010.03.005

33. Krepischi ACV, Achatz MIW, Santos EMM, et al. Germline DNA copy number variation in

familial and early-onset breast cancer. *Breast Cancer Res*. 2012;14(1). doi:10.1186/bcr3109

34. Park RW, Kim TM, Kasif S, Park PJ. Identification of rare germline copy number variations over-represented in five human cancer types. *Mol Cancer*. 2015;14(1). doi:10.1186/s12943-015-0292-6

35. Bruder CEG, Piotrowski A, Gijsbers AACJ, et al. Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. *Am J Hum Genet*. 2008;82(3):763-771. doi:10.1016/j.ajhg.2007.12.011

36. Fanciulli M, Petretto E, Aitman TJ. Gene copy number variation and common human disease. *Clin Genet*. 2010;77(3):201-213. doi:10.1111/j.1399-0004.2009.01342.x

37. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 2008. doi:10.1038/ng.237

38. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet*. 2013;45(10):1113-1120. doi:10.1038/ng.2764

39. UK Biobank Coordinating Centre. *UK Biobank-Genotyping and Imputation Data Release*.; 2018. http://www.ukbiobank.ac.uk/wp-content/uploads/2018/03/UKB-Genotyping-and-Imputation-Data-Release-FAQ-v3-2.pdf. Accessed December 3, 2019.

40. Allen NE, Sudlow C, Peakman T, Collins R. UK biobank data: Come and get it. *Sci Transl Med*. 2014;6(224). doi:10.1126/scitranslmed.3008601

41. Mardis ER. Next-Generation Sequencing Platforms. *Annu Rev Anal Chem*. 2013. doi:10.1146/annurev-anchem-062012-092628

42. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*. 2010;11(10):685-696. doi:10.1038/nrg2841

43. Wall JD, Tang LF, Zerbe B, et al. Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Res*. 2014;24(11):1734-1739. doi:10.1101/gr.168393.113

44. Harismendy O, Ng PC, Strausberg RL, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol*. 2009;10(3). doi:10.1186/gb-2009-10-3-r32

45. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol*. 2011;12(11). doi:10.1186/gb-2011-12-11-r112

46. NIH - National Cancer Institute G. Copy Number Variation Analysis Pipeline. https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/. Published 2021. Accessed January 28, 2021.

47. NIH - National Cancer Institute G. Affymetrix SNP 6.0. https://docs.gdc.cancer.gov/Encyclopedia/pages/Affymetrix_SNP_6.0/. Published 2021. Accessed January 29, 2021.

48. UK Biobank, Affymetrix. *UK Biobank Axiom Array - Content Summary*.; 2014. http://tools.thermofisher.com/content/sfs/brochures/uk_axiom_biobank_contentsummary_brochure.pdf?cid=2014070005. Accessed April 12, 2019.

49. UK Biobank Coordinating Centre. UK Biobank - Gentotype and Imputation Data Release. https://www.ukbiobank.ac.uk/media/cffi4mx5/ukb-genotyping-and-imputation-data-release-faq-v3-2-1.pdf. Published 2018. Accessed January 29, 2021.

50. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

51. Copeland J. The turing test. *Minds Mach*. 2000;10(4):519-539. doi:10.1023/A:1011285919106

52. French RM. The turing test: The first 50 years. *Trends Cogn Sci*. 2000;4(3):115-122. doi:10.1016/S1364-6613(00)01453-4

53. A.~L.~Samuel. Some Studies in Machine Learning Using the Game of Checkers. *IBM J Res Dev*. 1959;3(3):210-229.

54. Samuel AL. Programming Computers to Play Games. *Adv Comput*. 1960;1(C):165-192. doi:10.1016/S0065-2458(08)60608-7

55. Fukushima K. Cognitron: A self-organizing multilayered neural network. *Biol Cybern*. 1975;20(3-4):121-136. doi:10.1007/BF00342633

56. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol Cybern*. 1980;36(4):193-202.

doi:10.1007/BF00344251

57. David S, Demis H. AlphaGo: Mastering the ancient game of Go with Machine Learning. *Google Res Blog*. 2016. https://research.googleblog.com/2016/01/alphago-mastering-ancient-game-of-go.html.

58. Stephens CD. *Artificial Intelligence.* Vol 180.; 1996. doi:10.1038/sj.bdj.4809057

59. Pentakalos O. *Introduction to Machine Learning*.; 2019. doi:10.4018/978-1-7998-0414-7.ch003

60. Kotsiantis SB. Supervised machine learning: A review of classification techniques. *Inform*. 2007;31(3):249-268. doi:10.31449/inf.v31i3.148

61. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321-332. doi:10.1038/nrg3920

62. Garbade M. Understanding K-means Clustering in Machine Learning. Medium - Towards Data Science.

63. Hastie T, Tibshirani R, Friedman J. *Springer Series in Statistics*. Vol 27.; 2009. doi:10.1007/b94608

64. Liu L, Özsu MT, eds. *Encyclopedia of Database Systems*. New York, NY: Springer New York; 2020. doi:10.1007/978-1-4899-7993-3

65. H2O.ai. Generalized Linear Model. doi:10.1201/9781420060386.ch5

66. Lee Y, Nelder JA. Hierarchical Generalized Linear Models. *J R Stat Soc Ser B*. 1996.

doi:10.1111/j.2517-6161.1996.tb02105.x

67. Lee Y, Nelder JA, Pawitan Y. *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*.; 2006. doi:10.1111/j.1467-985x.2007.00485_4.x

68. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010. doi:10.18637/jss.v033.i01

69. Ho TK. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Anal Appl*. 2002;5(2):102-112. doi:10.1007/s100440200009

70. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*.; 2017. doi:10.1201/9781315139470

71. Oustimov A, Vu V. Artificial neural networks in the cancer genomics frontier. *Transl Cancer Res*. 2014;3(3):191-201. doi:10.3978/j.issn.2218-676X.2014.05.01

72. Ye J, Chow JH, Chen J, Zheng Z. Stochastic gradient boosted distributed decision trees. In: *International Conference on Information and Knowledge Management, Proceedings*. ; 2009:2061-2064. doi:10.1145/1645953.1646301

73. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. 2002;38(4):367-378. doi:10.1016/S0167-9473(01)00065-2

74. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232. doi:10.1214/aos/1013203451

75. Widrow B, Lehr MA. 30 Years of Adaptive Neural Networks: Perceptron, Madaline, and Backpropagation. *Proc IEEE*. 1990;78(9):1415-1442. doi:10.1109/5.58323

76. Mitchell R, Frank E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput Sci*. 2017;2017(7). doi:10.7717/peerj-cs.127

77. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Vol 13-17-Augu. ; 2016:785-794. doi:10.1145/2939672.2939785

78. Candel, Arno, Viraj Parmar, Erin LeDell AA. Deep learning with H2O. *H2O ai Inc*. 2016;(October):1-21.

79. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *J Mach Learn Res*. 2014;15:3133-3181. doi:10.1117/1.JRS.11.015020

80. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. In: *ACM International Conference Proceeding Series*. Vol 148. ; 2006:161-168. doi:10.1145/1143844.1143865

81. Caruana R, Karampatziakis N, Yessenalina A. An empirical evaluation of supervised learning in high dimensions. In: *Proceedings of the 25th International Conference on Machine Learning*. ; 2008:96-103. doi:10.1145/1390156.1390169

82. Breiman L. Stacked regressions. *Mach Learn*. 1996;24(1):49-64. doi:10.1007/bf00117832

83. Van Der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6(1). doi:10.2202/1544-6115.1309

84. Wolpert DH. Stacked generalization. *Neural Networks*. 1992;5(2):241-259.

doi:10.1016/S0893-6080(05)80023-1

85. The Cancer Genome Atlas Program - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Accessed March 14, 2019.

86. tcga-infographic-enlarge.__v100169753.png (1400×2580). https://www.cancer.gov/PublishedContent/Images/images/nci/organization/tcga/tcga-infographic-enlarge.__v100169753.png. Accessed March 14, 2019.

87. Welcome to the Pan-Cancer Atlas. https://www.cell.com/pb-assets/consortium/pancanceratlas/pancani3/index.html. Accessed March 14, 2019.

88. Hoadley KA, Yau C, Hinoue T, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*. 2018;173(2):291-304.e6. doi:10.1016/j.cell.2018.03.022

89. Malta TM, Sokolov A, Gentles AJ, et al. Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell*. 2018. doi:10.1016/j.cell.2018.03.034

90. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*. 2008;40(10):1166-1174. doi:10.1038/ng.238

91. The Cancer Genome Atlas Reports Brain Tumors Study - National Cancer Institute. http://www.cancer.gov/newscenter/newsfromnci/2008/tcgaglioblastoma. Accessed April 9, 2019.

92. Genome-Characterization-Pipeline-infographic-2019 - Enlarge.__v2001486135.png (1924×2508). https://www.cancer.gov/PublishedContent/Images/about-nci/organization/ccg/research/genomic-pipeline/Genome-Characterization-Pipeline-infographic-2019 - Enlarge.__v2001486135.png. Accessed April 9, 2019.

93. Genome Characterization Pipeline - Center for Cancer Genomics - National Cancer Institute. https://www.cancer.gov/about-nci/organization/ccg/research/genomic-pipeline#collection-processing. Accessed April 9, 2019.

94. Biobank. Biobank: About UK Biobank. https://www.ukbiobank.ac.uk/about-biobank-uk/. Published 2016. Accessed April 12, 2019.

95. UK Biobank. UK Biobank: Protocol for a large-scale prospective epidemiological resource. *UKBB-PROT-09-06 (Main Phase)*. 2007;06(March):1-112. https://www.ukbiobank.ac.uk/wp-content/uploads/2011/11/UK-Biobank-Protocol.pdf.

96. Ollier W, Sprosen T, Peakman T. UK Biobank: From concept to reality. *Pharmacogenomics*. 2005. doi:10.2217/14622416.6.6.639

97. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med*. 2015;12(3). doi:10.1371/journal.pmed.1001779

98. Littlejohns TJ, Sudlow C, Allen NE, Collins R. UK Biobank: opportunities for cardiovascular research. *Eur Heart J*. 2017;44:1158-1166. doi:10.1093/eurheartj/ehx254

99. Welsh S, Peakman T, Sheard S, Almond R. Comparison of DNA quantification

methodology used in the DNA extraction protocol for the UK Biobank cohort. *BMC Genomics*. 2017. doi:10.1186/s12864-016-3391-x

100. Elliott LT, Sharp K, Alfaro-Almagro F, et al. Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*. 2018. doi:10.1038/s41586-018-0571-7

101. UK Biobank — Oxford Big Data Institute. Big Data Institute. https://www.bdi.ox.ac.uk/research/uk-biobank. Published 2019. Accessed April 12, 2019.

102. Wain L V, Shrine N, Miller S, et al. Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med*. 2015;3(10):769-781. doi:10.1016/S2213-2600(15)00283-0

103. Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol*. 2008;37(2):234-244. doi:10.1093/ije/dym276

104. Peakman TC, Elliott P. The UK Biobank sample handling and storage validation studies. *Int J Epidemiol*. 2008;37(SUPPL. 1). doi:10.1093/ije/dyn019

105. An Interface to Google's "BigQuery" "API" • bigrquery. https://bigrquery.r-dbi.org/. Accessed April 9, 2019.

106. Overview — H2O 3.24.0.1 documentation. http://docs.h2o.ai/h2o/latest-stable/h2o-docs/index.html. Accessed April 9, 2019.

107. McCullagh P, Nelder JA. Generalized Linear Models, Second Edition (Monographs on

Statistics and Applied Probability). *Lavoisierfr*. 1989.

108. H2O.ai. Generalized Linear Model (GLM). http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/glm.html. Published 2021. Accessed February 12, 2021.

109. Hastie TJ, Pregibon D. Generalized linear models. In: *Statistical Models in S*. ; 2017. doi:10.1201/9780203738535

110. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42. doi:10.1007/s10994-006-6226-1

111. Crosby MH. Interest in teaching skills development: a survey of Virginia nurses. *J Contin Educ Nurs*. 1977;8(4):35-36. doi:10.3928/0022-0124-19770701-11

112. H2O.ai. XGBoost. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/xgboost.html. Accessed February 19, 2021.

113. H2O.ai. Stacked Ensembles. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html. Published 2021. Accessed February 19, 2021.

114. 3.6.1. RDCT. A Language and Environment for Statistical Computing. *R Found Stat Comput*. 2019;2:https://www.R--project.org. http://www.r-project.org.

115. Hanscombe KB, Coleman JRI, Traylor M, Lewis CM. UKBTools: An R package to manage and query UK Biobank data. *PLoS One*. 2019;14(5). doi:10.1371/journal.pone.0214311

116. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686. doi:10.21105/joss.01686

117. Wickham H. Francois R. dplyr: A Grammar of Data Manipulation. R package version 0.4. 3. 2015. *Media*. 2018.

118. Wickham H. *Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York*.; 2009.

119. Wickham H, Kolaczyk ED, Csárdi G, et al. *R-Bloggers*.; 2018.

120. Toh C, Brody JP. Chromosomal scale length variation of germline DNA can predict individual cancer risk. *bioRxiv*. 2018. doi:10.1101/303339

121. Kenney AD, Dowdle JA, Bozzacco L, et al. Human Genetic Determinants of Viral Diseases. *Annu Rev Genet*. 2017;51:241-263. doi:10.1146/annurev-genet-120116-023425

122. Everitt AR, Clare S, Pertel T, et al. IFITM3 restricts the morbidity and mortality associated with influenza. *Nature*. 2012;484(7395):519-523. doi:10.1038/nature10921

123. Toh C, Brody JP. Chromosomal scale length variation of germline DNA can predict individual cancer risk. *bioRxiv*. 2018:303339. doi:10.1101/303339

124. Olson RS, La Cava W, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. In: *Pacific Symposium on Biocomputing*. Vol 0. ; 2018:192-203. doi:10.1142/9789813235533_0018

125. Van Der Made CI, Simons A, Schuurs-Hoeijmakers J, et al. Presence of Genetic Variants among Young Men with Severe COVID-19. *JAMA - J Am Med Assoc*. 2020;324(7):663-673. doi:10.1001/jama.2020.13719

126. Kolin DA, Kulm S, Elemento O. Clinical and Genetic Characteristics of Covid-19 Patients

from UK Biobank. *medRxiv*. 2020. doi:10.1101/2020.05.05.20075507

127. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J Med*. 2020;383(16):1522-1534. doi:10.1056/nejmoa2020283

128. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet*. 2020. doi:10.1038/s41431-020-0636-6

129. Covid-19 Host Genetics Initiative Results. https://www.covid19hg.org/results/. Accessed June 29, 2020.

130. Webb Hooper M, Napoles AM P-SE. COVID-19 and Racial/Ethnic Disparities. JAMA. https://jamanetwork.com/journals/jama/fullarticle/2766098. Published 2020. Accessed June 14, 2020.

131. Garg S, Kim L, Whitaker M, O’Halloran A, Cummings C, Holstein R et al. Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019 — COVID-NET, 14 States, March 1–30, 2020. MMWR Morbidity and Mortality Weekly Report. http://www.cdc.gov/mmwr/volumes/69/wr/mm6915e3.htm?s_cid=mm6915e3_w. Published 2020. Accessed June 14, 2020.

132. 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(6). doi:10.3322/caac.21492

133. Torre LA, Trabert B, DeSantis CE, et al. Ovarian cancer statistics, 2018. *CA Cancer J Clin*. 2018;68(4). doi:10.3322/caac.21456

134. Bast RC. Status of Tumor Markers in Ovarian Cancer Screening. *J Clin Oncol*. 2003;21(90100). doi:10.1200/JCO.2003.01.068

135. Andrews L, Mutch DG. Hereditary Ovarian Cancer and Risk Reduction. *Best Pract Res Clin Obstet Gynaecol*. 2017;41. doi:10.1016/j.bpobgyn.2016.10.017

136. Grossman DC, Curry SJ, Owens DK, et al. Screening for Ovarian Cancer. *JAMA*. 2018;319(6). doi:10.1001/jama.2017.21926

137. Trimbos JB. Surgical treatment of early-stage ovarian cancer. *Best Pract Res Clin Obstet Gynaecol*. 2017;41. doi:10.1016/j.bpobgyn.2016.10.001

138. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. 2016;315(1). doi:10.1001/jama.2015.17703

139. Janssens ACJW, van Duijn CM. Genome-based prediction of common diseases: advances and prospects. *Hum Mol Genet*. 2008;17(R2). doi:10.1093/hmg/ddn250

140. 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(7). doi:10.1097/01.gim.0000229689.18263.f4

141. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018. doi:10.1038/s41576-018-0018-x

142. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. *Hum*

Mol Genet. 2019;28(R2). doi:10.1093/hmg/ddz187

143. Pharoah PDP, Tsai Y-Y, Ramus SJ, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet*. 2013;45(4). doi:10.1038/ng.2564

144. Kuchenbaecker KB, Ramus SJ, Tyrer J, et al. Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nat Genet*. 2015;47(2). doi:10.1038/ng.3185

145. Lewis CM, Vassos E. Polygenic risk scores: from research tools to clinical instruments. *Genome Med*. 2020;12(1). doi:10.1186/s13073-020-00742-5

146. Goode EL, Chenevix-Trench G, Song H, et al. A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat Genet*. 2010;42(10). doi:10.1038/ng.668

147. Cancer Genome Atlas Research Network T, Bell D, Berchuck A, et al. Integrated genomic analyses of ovarian carcinoma. *Nature*. 2011. doi:10.1038/nature10166

148. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*. 2018;173(2). doi:10.1016/j.cell.2018.03.042

149. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004. doi:10.1093/biostatistics/kxh008

150. National Cancer Institute Genomic Data Commons. https://gdc.cancer.gov/. Accessed March 26, 2021.

151. Reynolds SM, Miller M, Lee P, et al. The ISB Cancer Genomics Cloud: A Flexible Cloud-Based Platform for Cancer Genomics Research. *Cancer Res*. 2017;77(21). doi:10.1158/0008-5472.CAN-17-0617

152. LeDell E, Gill N, Aiello S, et al. *R Interface for "H2O."*; 2019.

153. Tenny S, Hoffman MR. Odds Ratio. *StatPearls*. July 2020.

154. Lundberg SM, Allen PG, Lee S-I. *A Unified Approach to Interpreting Model Predictions*. https://github.com/slundberg/shap.

155. Zhang B. Colorectal cancer predictive test using germ-line DNA data and multiple machine learning methods. 2019. https://escholarship.org/uc/item/44f3f487.

156. Cancer T, Atlas G, Network TCGA, et al. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012. doi:10.1038/nature11412

157. Abeliovich D, Kaduri L, Lerer I, 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;60(3):505-514. https://pubmed.ncbi.nlm.nih.gov/9042909.

158. Efron B. Prediction, Estimation, and Attribution. *J Am Stat Assoc*. 2020;115(530). doi:10.1080/01621459.2020.1762613

159. 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;38(8). doi:10.1038/ng1847

160. Flint J, Munafò M. Genesis of a complex disease. *Nature*. 2014;511(7510):412-413. doi:10.1038/nature13645

161. Ripke S, Neale BM, Corvin A, et al. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014. doi:10.1038/nature13595

162. Ripke S, Sanders AR, Kendler KS, et al. Genome-wide association study identifies five new schizophrenia loci. *Nat Genet*. 2011. doi:10.1038/ng.940

163. Farrell MS, Werge T, Sklar P, et al. Evaluating historical candidate genes for schizophrenia. *Mol Psychiatry*. 2015;20(5):555-562. doi:10.1038/mp.2015.16

164. Van Os J, Kenis G, Rutten BPF. The environment and schizophrenia. *Nature*. 2010. doi:10.1038/nature09563

165. Sullivan PF, Kendler KS, Neale MC. Schizophrenia as a Complex Trait: Evidence from a Meta-analysis of Twin Studies. *Arch Gen Psychiatry*. 2003. doi:10.1001/archpsyc.60.12.1187

166. Sugrue LP, Desikan RS. What Are Polygenic Scores and Why Are They Important? *JAMA*. 2019;321(18):1820. doi:10.1001/jama.2019.3893

167. Lello L, Raben TG, Yong SY, Tellier LCAM, Hsu SDH. Genomic Prediction of 16 Complex Disease Risks Including Heart Attack, Diabetes, Breast and Prostate Cancer. *Sci Rep*. 2019. doi:10.1038/s41598-019-51258-x

168. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet*. 2019.

doi:10.1016/j.ajhg.2018.11.002

169. Fahed AC, Wang M, Homburger JR, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. *Nat Commun*. 2020. doi:10.1038/s41467-020-17374-3

170. Abraham G, Malik R, Yonova-Doing E, et al. Genomic risk score offers predictive performance comparable to clinical risk factors for ischaemic stroke. *Nat Commun*. 2019. doi:10.1038/s41467-019-13848-1

171. Agerbo E, Sullivan PF, Vilhjálmsson BJ, et al. Polygenic risk score, parental socioeconomic status, family history of psychiatric disorders, and the risk for schizophrenia: A Danish population-based study and meta-analysis. *JAMA Psychiatry*. 2015. doi:10.1001/jamapsychiatry.2015.0346

172. Power RA, Steinberg S, Bjornsdottir G, et al. Polygenic risk scores for schizophrenia and bipolar disorder predict creativity. *Nat Neurosci*. 2015. doi:10.1038/nn.4040

173. Mistry S, Harrison JR, Smith DJ, Escott-Price V, Zammit S. The use of polygenic risk scores to identify phenotypes associated with genetic risk of schizophrenia: Systematic review. *Schizophr Res*. 2018. doi:10.1016/j.schres.2017.10.037

174. Loh P-R, Genovese G, McCarroll SA. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature*. 2020;584(7819). doi:10.1038/s41586-020-2430-6

175. Zarrei M, Burton CL, Engchuan W, et al. A large data resource of genomic copy number

variation across neurodevelopmental disorders. *npj Genomic Med*. 2019;4(1). doi:10.1038/s41525-019-0098-3

176. Bache WK, DeLisi LE. The Sex Chromosome Hypothesis of Schizophrenia: Alive, Dead, or Forgotten? A Commentary and Review. *Mol Neuropsychiatry*. 2018;4(2):83-89. doi:10.1159/000491489

177. Ko C, Toh C, Brody JP. Abstract PR-09: Genetic risk scores for breast cancer based on machine learning analysis of chromosomal-scale length variation. In: *Oral Presentations - Proffered Abstracts*. American Association for Cancer Research; 2021. doi:10.1158/1557-3265.ADI21-PR-09

178. Toh C, Brody JP. Genetic risk score for ovarian cancer based on chromosomal-scale length variation. *BioData Min*. 2021;14(1). doi:10.1186/s13040-021-00253-y

179. Li K, Hüsing A, Fortner RT, et al. An epidemiologic risk prediction model for ovarian cancer in Europe: The EPIC study. *Br J Cancer*. 2015;112(7). doi:10.1038/bjc.2015.22

180. Ko C, Brody JP. A genetic risk score for glioblastoma multiforme based on copy number variations. *Cancer Treat Res Commun*. 2021;27. doi:10.1016/j.ctarc.2021.100352

181. Hu N, Cheng H, Zhang K, Jensen R. Evaluating the Prognostic Accuracy of Biomarkers for Glioblastoma Multiforme Using The Cancer Genome Atlas Data. *Cancer Inform*. 2017;16. doi:10.1177/1176935117734844

182. Heenkenda MK, Malmström A, Lysiak M, et al. Assessment of genetic and non-genetic risk factors for venous thromboembolism in glioblastoma – The predictive significance of

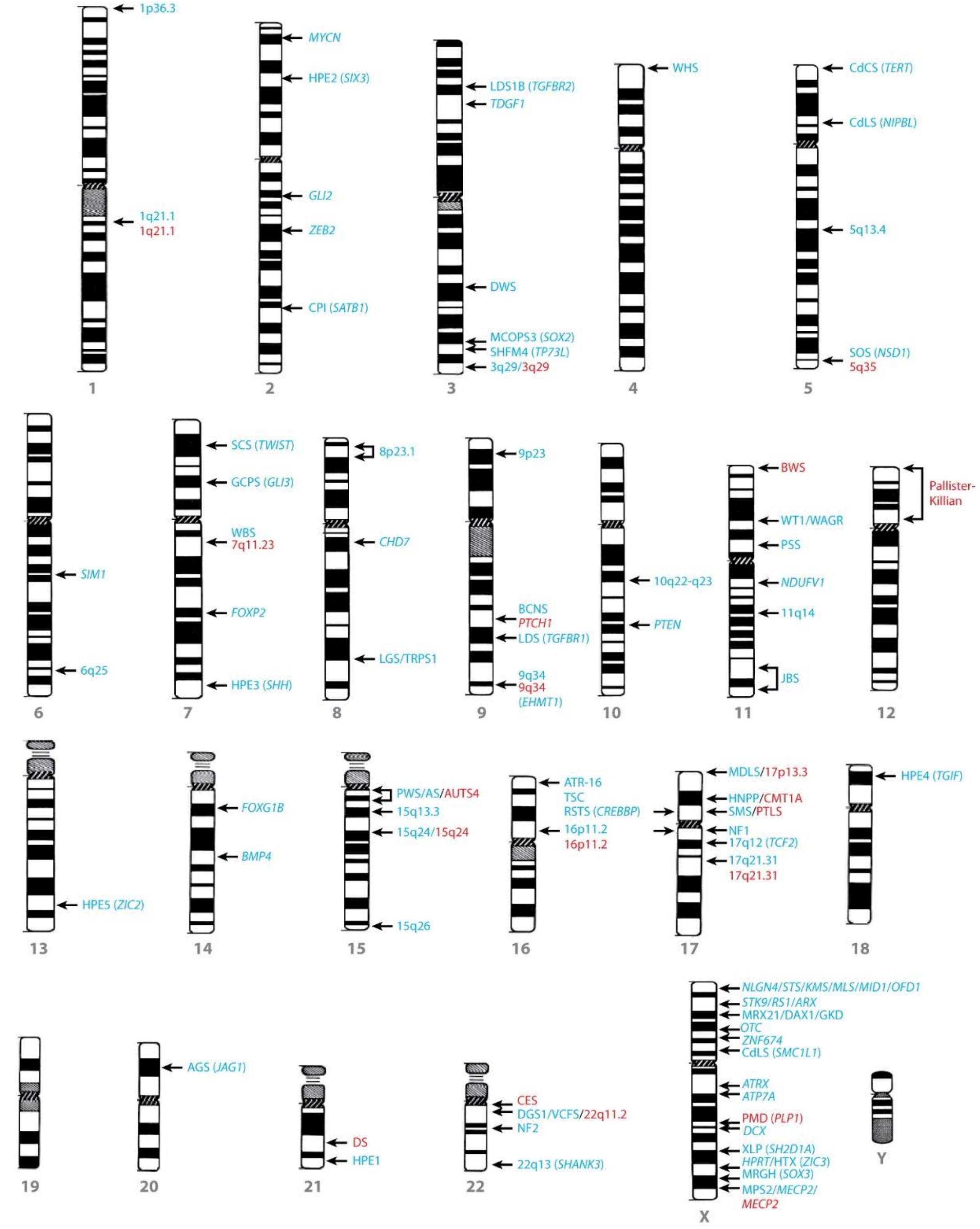
B blood group. *Thromb Res*. 2019;183. doi:10.1016/j.thromres.2019.10.009

183. Toh C, Brody J. Genetic Risk Score for Predicting Schizophrenia Using Human Chromosomal-Scale Length Variation. March 2021.

184. Bracher-Smith M, Menzies G, Kendall K, et al. F29INVESTIGATING SUPERVISED MACHINE LEARNING METHODS FOR PREDICTION OF SCHIZOPHRENIA IN UK BIOBANK. *Eur Neuropsychopharmacol*. 2019;29. doi:10.1016/j.euroneuro.2018.08.109

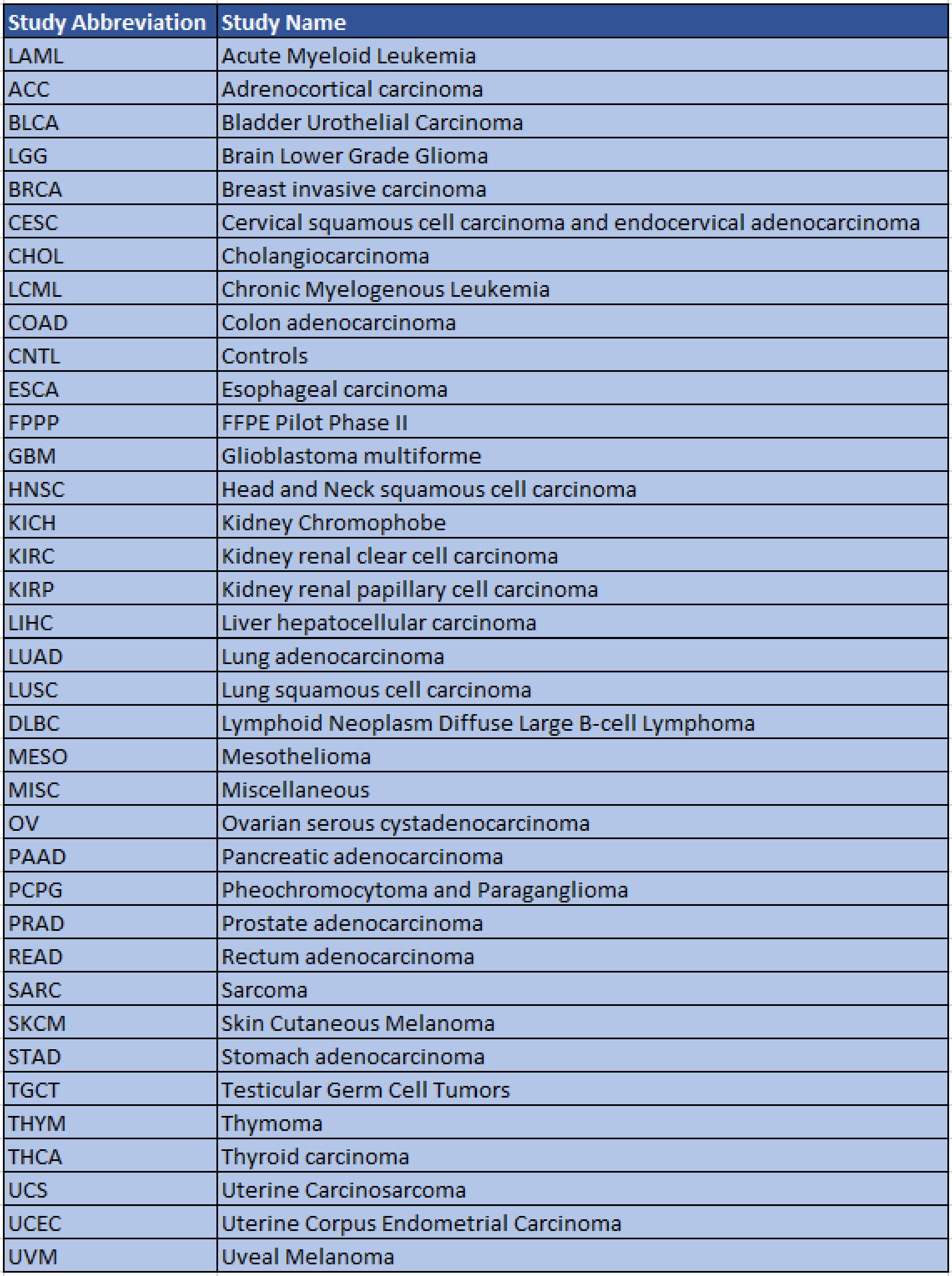
185. González-Peñas J, Amigo J, Santomé L, et al. Targeted resequencing of regulatory regions at schizophrenia risk loci: Role of rare functional variants at chromatin repressive states. *Schizophr Res*. 2016;174(1-3). doi:10.1016/j.schres.2016.03.029

Appendix



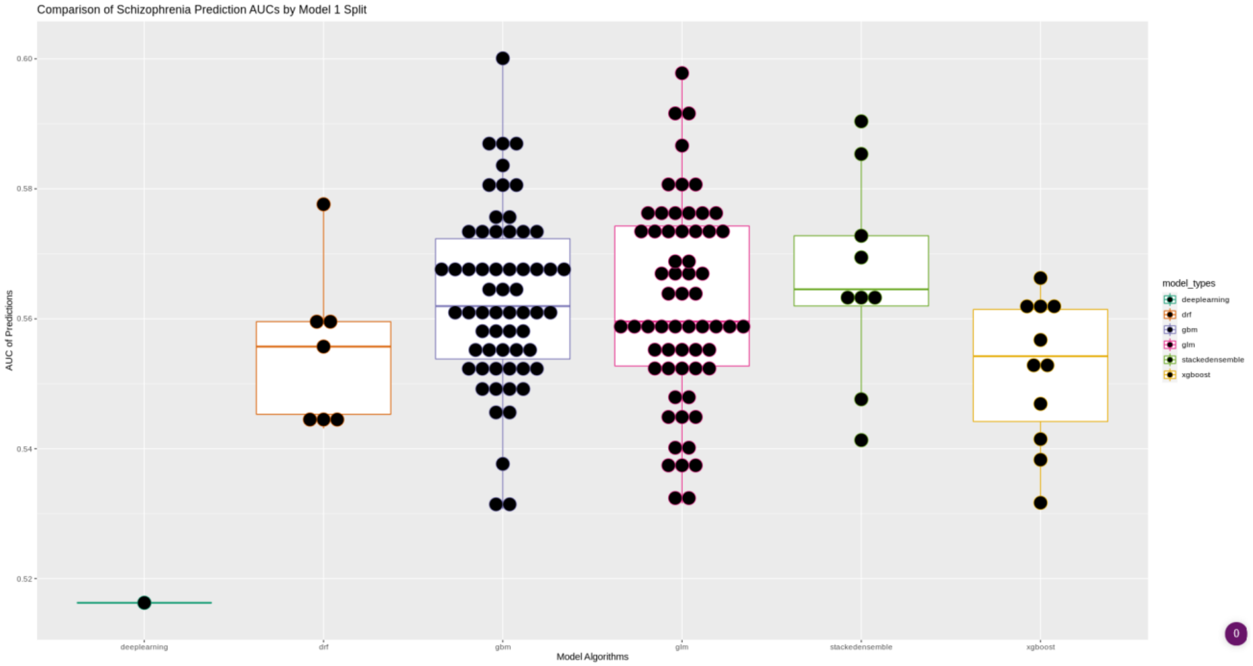
Stankiewicz P, Lupski JR. 2010.
Annu. Rev. Med. 61:437–55

Appendix 1:Examples of Known CNV Locations Associated with Complex Human Diseases. Stankiewicz et al.

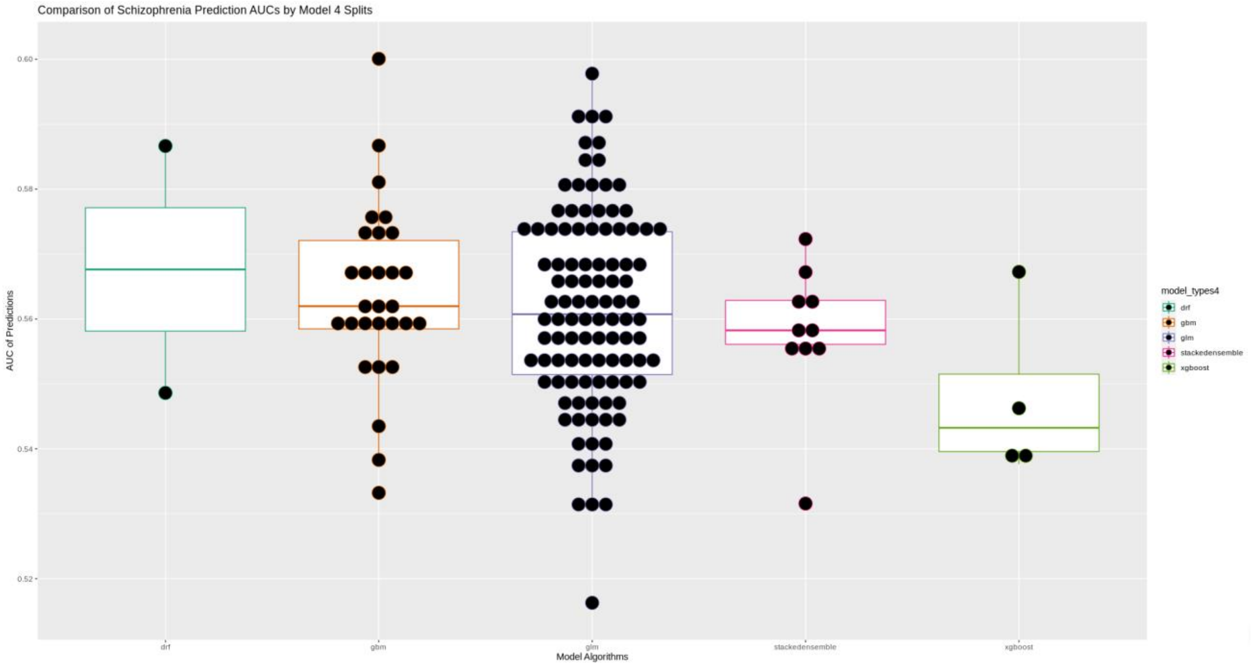


Study Abbreviation	Study Name
LAML	Acute Myeloid Leukemia
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
LGG	Brain Lower Grade Glioma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
CHOL	Cholangiocarcinoma
LCML	Chronic Myelogenous Leukemia
COAD	Colon adenocarcinoma
CNTL	Controls
ESCA	Esophageal carcinoma
FPPP	FFPE Pilot Phase II
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
MESO	Mesothelioma
MISC	Miscellaneous
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular Germ Cell Tumors
THYM	Thymoma
THCA	Thyroid carcinoma
UCS	Uterine Carcinosarcoma
UCEC	Uterine Corpus Endometrial Carcinoma
UVM	Uveal Melanoma

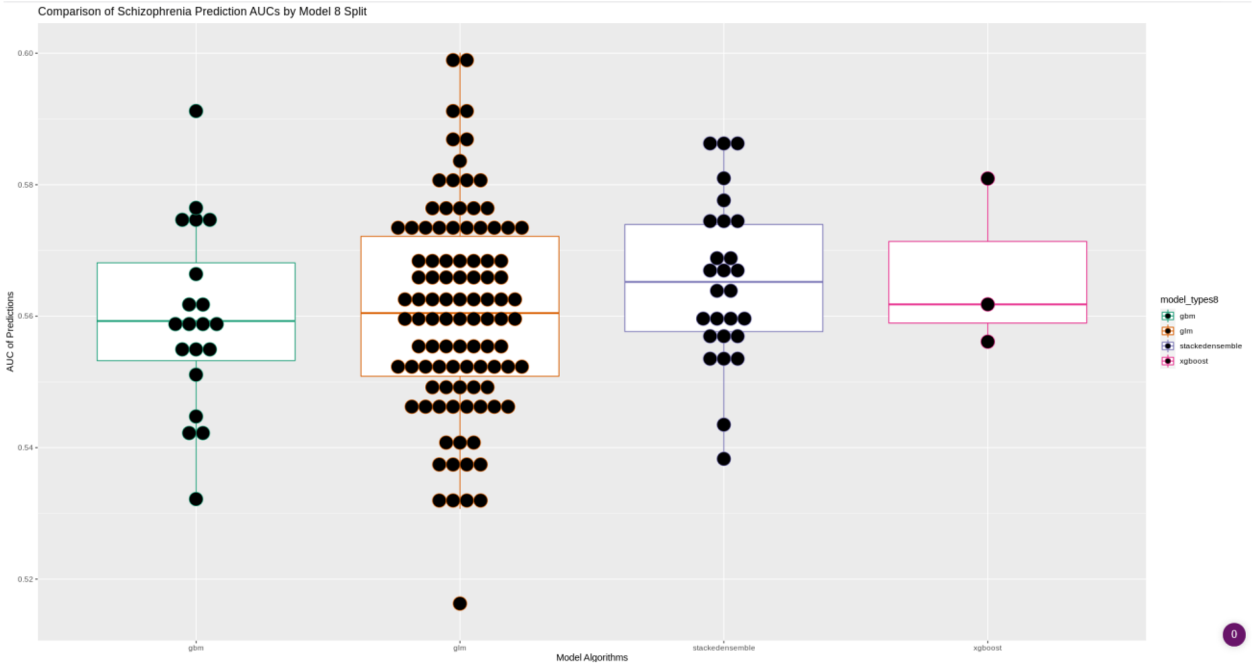
Appendix 2: Table of TCGA Study Codes or Abbreviations



Appendix 3: Comparison of Schizophrenia Prediction AUCs by Model for 1 Split Dataset



Appendix 4: Comparison of Schizophrenia Prediction AUCs by Model for 4 Split Dataset



Appendix 5: Comparison of Schizophrenia Prediction AUCs by Model for 1 Split Dataset