

**Integrated Deep Learning and Bayesian Classification for
Prioritization of Functional Genes in Next-Generation
Sequencing Data**

Chan Khai Ern, Edwin

A thesis submitted to the
Department of Biochemistry
National University of Singapore
in partial fulfilment for the
Degree of Bachelor of Science with
Honours in Life Sciences

Life Sciences Honours Cohort
AY2015/2016 S1

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Chan Khai Ern Edwin

04 April 2017

Acknowledgements

Table of Contents

Contents

1	Introduction	2
1.1	Next Generation Sequencing (NGS) for Clinical Genomics	2
1.2	Variant Calling of NGS Data	2
1.3	Ensemble Methods for Improving the Accuracy of Variant Calling	4
1.4	Deep Learning for Improving the Accuracy of Variant Calling	5
1.5	Prioritisation of Variants with Bayesian Networks	7
1.6	Aims and Approach	9
2	Materials and Methods	10
2.1	Overall Analysis Structure	10
2.2	Programming and Pipelining tools	11
2.2.1	General Programming Language	11
2.2.2	Pipelining	12
2.2.3	Deep Learning	12
2.2.4	Gene Ranking	12
2.3	Simulated Datasets	12
2.4	Alignment and Variant Calling	13
2.5	Feature Engineering	13
2.6	Patient Derived Xenograft Mouse Model Development and Sequencing	15
3	Results	16
3.1	Generation of Simulated Datasets	16
3.2	Feature Engineering	17
3.3	Variant Callers	18
3.4	Network Architecture	19
3.5	Network Tuning and Optimisation	21
3.5.1	Number of Layers	22
3.5.2	Optimiser and Learning Rates	23
3.5.3	Sample Balancing	25
3.6	Benchmarking of Optimised Network with Mason Dataset	26
3.7	Benchmarking of Optimised Network with NA Dataset	28
3.8	Analysis of Gene Importance using Bayesian Ranking systems	30
3.9	Validation of Bayesian Network Ranking on PDX dataset	31

4 Discussion	36
4.1 Comparison of Deep Learning with other Integration Methods	36
4.2 Analysis of Bayesian Network	38
4.3 Future Directions	39
5 Appendixes	40
5.1 Neural Network Learning	40
5.1.1 Feedforward Phase	40
5.1.2 Backpropagation Phase	41
5.1.3 Cost Function in Gradient Descent	42
5.2 Feature Engineering	44
5.2.1 Base Information	44
5.2.2 Sequencing Biases and Errors	45
5.2.3 Calling and Mapping Qualities	46
5.3 Mathematical and Statistical Tools	48
5.3.1 Derivation of F1 Score	48
5.3.2 Principal Components Analysis (PCA)	49
5.3.3 Synthetic Minority Overrepresentation Technique (SMOTE)	50
6 Bibliography	51
7 Relevant Code	62
7.1 generate_matrixes.py	62
7.2 train_network.py	71
7.3 compute_bayesian.py	80

Abstract

The advent of next-generation sequencing technology has enabled large-scale interrogation of the genome to identify variants in patient samples. The accurate identification of functional variants can provide critical insights into the disease process to guide diagnosis and treatment. However, the use of clinical genomics remains limited as (i) the accurate identification of variants remains suboptimal, and (ii) the large number of variants identified may be difficult to interpret without a systematic approach of ranking by functional importance.

Here, we describe the development of a deep learning neural network to improve the accuracy of variant-calling, and a Bayesian classification method for the probabilistic ranking of functionally relevant genes. We show that an optimised neural network can call variants more accurately than single variant callers or concordant callers, with F1 score improvements of 6.5 percent in simulated datasets and 4.5 percent in real datasets over the best concordant methods. Following the identification of high confidence variants, we further demonstrate that a Bayesian classification system can rank functionally relevant genes in a Diffuse Large B-Cell Lymphoma (DLBCL) patient sample.

We propose that the combined use of deep learning and Bayesian network analysis could be extended to build an analytical pipeline for clinical use to augment diagnosis and treatment of diseases by identifying high-confidence variants and ranking them systematically.

1 Introduction

1.1 Next Generation Sequencing (NGS) for Clinical Genomics

There has been a growing interest in using a patient's genome to guide the diagnosis and treatment of diseases (Rehm, 2017; Angrist, 2016), based on the fundamental intuition that variants and mutations in the genome alter gene functions that drive the initiation and progression of the disease. In oncology, for example, identification of the key driver mutations has been shown to be useful in stratifying cancer subtypes (Stratton, Campbell & Futreal, 2009), and identifying mutations for targeted therapy (Janitz, 2011). Furthermore, the development of next generation sequencing (NGS) technologies has dramatically reduced sequencing costs (Metzker, 2010; Mardis, 2008), enabling the adoption of genomic sequencing in clinical labs.

Although clinical genomics holds great promise, there are still two critical issues that limit its use in a clinical setting. Firstly, it is often difficult to obtain high-confidence variant calls from sequencing data, and secondly, the large number of variant calls for patient samples makes interpretation difficult for clinical decision-making.

1.2 Variant Calling of NGS Data

In variant calling, genomic DNA is fragmented, and the short reads are sequenced in a massively parallel manner using next generation sequencing technologies such as sequencing by synthesis. These reads are aligned to the reference genome, and variations in the DNA sequence, such as single nucleotide variants (SNV) and insertions/deletions (indels) are identified by comparing the different reads aligned to the reference genome (Figure 1).

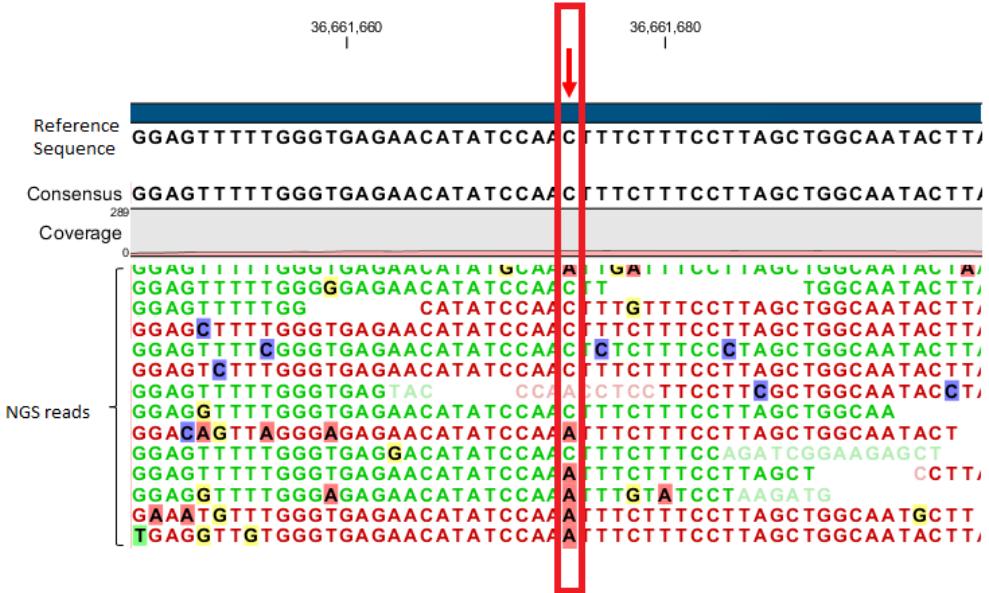


Figure 1: Illustration of Variant Calling Pileup. Due to noise and errors in sequencing and read mapping, it can be difficult to call variants accurately. At the position of interest, there seems to be a possible mutation from the original base of Cytosine to the base Adenine, but not all reads agree with this mutation call. Figure adapted from CLC Genomics Workbench 9.5, Figure 29.8.

Variant calling with NGS data primarily involves the use of various statistical and algorithmic methods to identify variants in the genome (Nielson et al., 2011). These variants represent the deviations and differences between the genome of interest and a reference human genome. This analysis is non-trivial as each variant call requires the integration of multiple sequence reads (e.g. millions of reads) that contain experimental noise and errors (Zook et al., 2014). The calling of variants can be further complicated by errors in mapping the reads to a reference genome.

To account for these errors, variant callers employ a variety of algorithms and statistical models to determine the existence and type of variation/mutation (Zook et al., 2014; Davey et al., 2011). Because of the differences in assumptions and models employed by different variant callers, certain calling algorithms are more sensitive and accurate in calling specific classes of variants but do not perform well in calling other variant types (O’Rawe et al., 2013). To address these problems, ongoing efforts have focused on improving current variant calling algorithms, including optimisation of variant calling for different classes of mutations, as well as the reduction in the number of false positive calls (Mohiyuddin et al., 2015; Gézsi et al., 2015). Despite the variety of approaches used for identifying variants and mutations, the accuracy and

precision of single variant callers remain suboptimal (Cornish and Guda, 2015; O’Rawe et al., 2013). Each variant caller can differ greatly in accuracy depending on the type of sequencing methodology and the statistical algorithm used (Figure 2), making it difficult to identify true high-confidence variant calls.

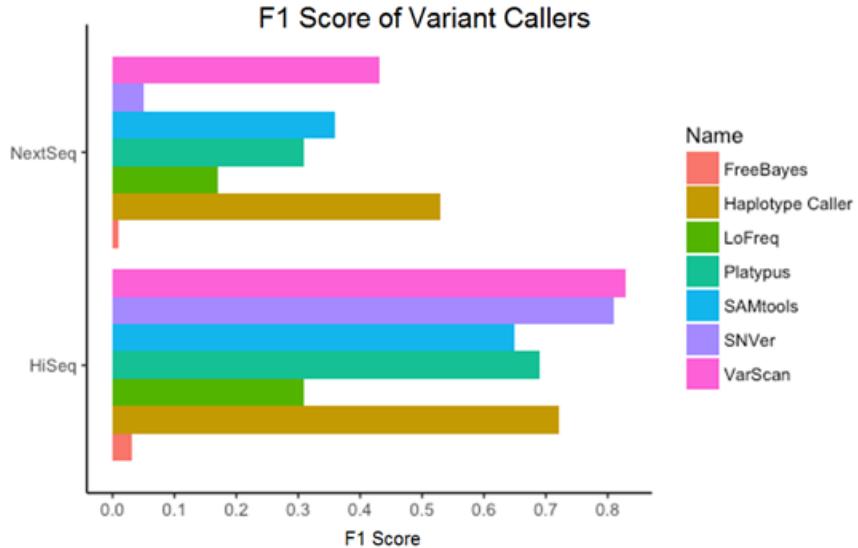


Figure 2: Performance of Variant Calling Tools on Patient Data using Illumina Sequencing Platforms HiSeq and NextSeq. The F1 score indicates how well a caller can predict true positives (See Appendix 5.3 for more details). Notably, the F1 score for the same variant calling tool can differ greatly. Figure adapted from Sandmann et al. (2017)

1.3 Ensemble Methods for Improving the Accuracy of Variant Calling

While it is clear that single variant callers may not perform well across a variety of variant classes, the combination or ensemble of several callers can be used to augment the accuracy of variant callers beyond what can be achieved with a single caller. By aggregating the calls from each different variant caller, the relatively weak prediction calls from each caller can be combined to provide a better aggregate prediction for a variant call. One simple approach to aggregating variant calls is by concordance, where the likelihood of an accurate call depends on multiple variant callers identifying the same variant or mutation (Lam et al., 2012; Wei et al., 2011). While straightforward and intuitive, the recall rates of such a tool is poor with a high number of false negative calls. This is because a high concordance of variant calls will reciprocally decrease the number of true variants calls that are identified by specific variant callers (O’Rawe et al., 2013).

Beyond concordance, supervised machine learning approaches have been used to combine the calls from different variant callers to improve the accuracy. In these approaches, machine learning algorithms have been used predict the accuracy of a variant call by integrating different features of each variant call (e.g. variant frequencies, mapping quality). For example, the support vector machine (SVM) algorithm was used successfully to improve the accuracy of variant calls (Gézsi et al., 2015) over concordance-based methods.

Recent advances in machine learning, in particular, deep learning neural networks, have increased the accuracy of predictions from complex multi-modal data (Ng et al., 2015) beyond traditional algorithms such as SVM and Random Forests. The ability of deep learning networks to learn from complex high-dimensional data suggests that they may be useful in improving the accuracy of high-confidence variant calls derived from the complex features from each variant caller.

1.4 Deep Learning for Improving the Accuracy of Variant Calling

Deep learning is a machine learning approach based on artificial neural networks (LeCun et al., 2015) built on artificial neurons. Each artificial neuron is analogous to a biological neuron where weighted input signals are integrated to produce an output once the signals cross a threshold for activation (Figure 3).

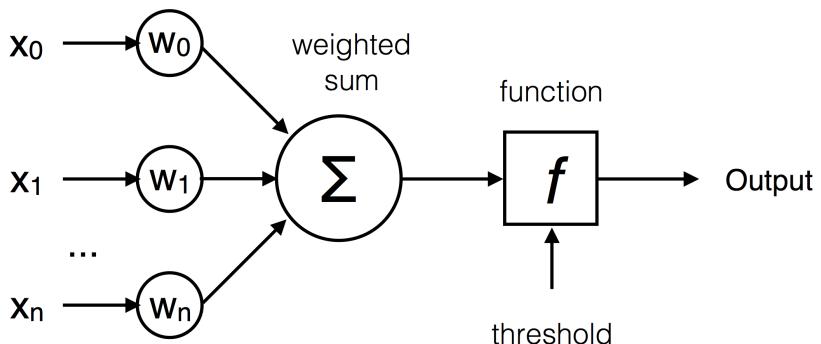
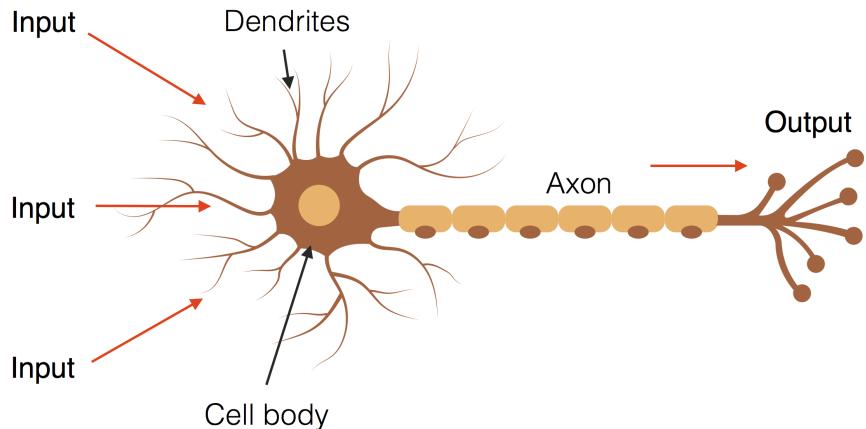


Figure 3: Artificial Neurons as Building Blocks of Neural Networks

Deep learning networks are trained by providing the input data and output data and letting the network learn how to integrate the inputs to create a network of activations that can be used to produce the corresponding output. A more in-depth explanation of their algorithms can be found in Appendix 5.1. The output can be propagated to the next neuron depending on the weighted inputs. By adjusting the weights that determine the propagation of a signal from one neuron to the next, the network can be adjusted to learn from labelled training data so that it can predict the output given new data. Deep learning has been shown to be able to solve complex non-linear decision boundary problems, including drug molecule solubility (Lusci et al., 2013), facial recognition (Sun et al., 2014) and even predicting the best move in the Japanese board game, Go (Silver et al., 2016). Figure 4 shows a typical neural network with five layers, including an input and output layer.

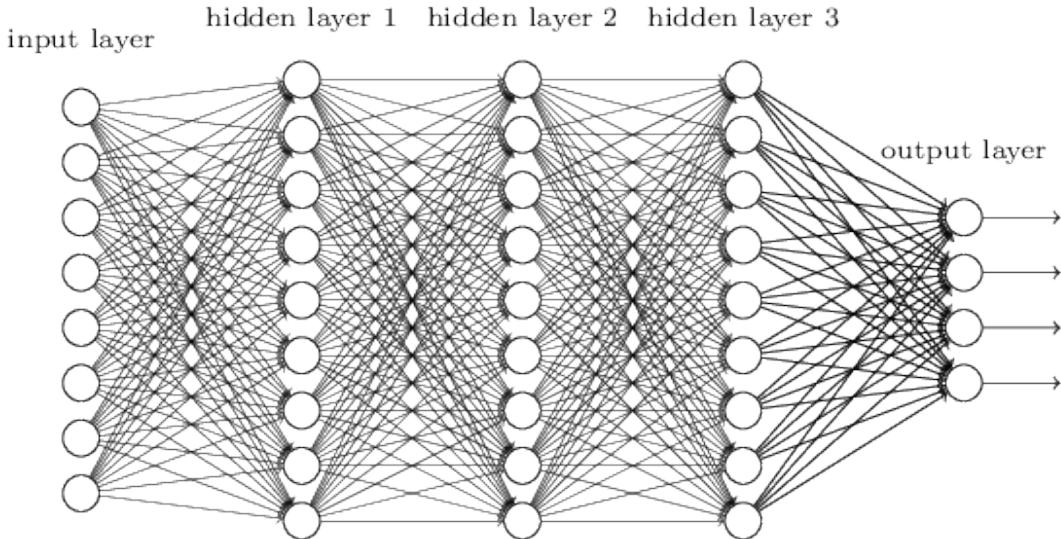


Figure 4: Sample Neural Network with Five Layers, Including Three Hidden Layers. This represents a densely connected neural network, where each node is connected to every node of the preceding and subsequent layers.

In variant calling, we hypothesise that deep learning will allow us to predict based on variant calling features and data whether a variant is valid and exists, or is erroneous. The deep learning network should be able to draw on the diversity of data with different variant callers and learn which patterns will result in a valid call and which patterns are false positives. It will also allow us to tap on the differential sensitivity of different callers, as the neural network can learn which callers work best for which types of mutations. Thus, we propose that deep learning as a combinatorial approach will allow us to improve the accuracy and precision of variant calling.

1.5 Prioritisation of Variants with Bayesian Networks

Once high-confidence variant calls can be established, there remains the second problem of identifying the functional importance of each variant/mutation, given that there are multiple variants in a typical genome (Shen et al., 2013). The ability to systematically prioritise and rank clinically significant variants and mutations would allow clinicians to focus their attention on relevant candidate mutations that can guide decision-making on diagnosis and treatment.

The problem of prioritisation of genetic mutations arises from the multiplicity and complexity of data sources that can be utilised to determine the clinical and functional relevance of a variant or mutation in a gene (Moreau & Tranchevent, 2012). Several approaches include studying previously characterised variants and their phenotypic effects on a person as well as

how the mutation itself will affect protein function through studying the likelihood of amino acid mutation for conserved regions. These functional annotations can be done with the tool ANNOVAR (Wang, Li, & Hakonarson, 2010) but the fundamental problem here is integrating such information in a systematic manner that is clear and understandable to clinicians. Clinicians may not be so familiar with the tools and functional annotation pipelines, and so for them to trust such a ranking system, they have to be able to understand intuitively how it works.

To solve this problem, we can use Bayesian networks to integrate the information from functional annotations as well as the confidence of a call (how likely it is real) to provide a ranking system for how likely the gene is going to be important. Bayesian networks were chosen for this ranking system as it is understandable and yet have proven stable in terms of solving decision-making problems (Pourret et al., 2008; Jensen et al., 1996). Bayesian networks have been applied in medical treatment decision making (Windecker et al., 2014), ecological studies (Johnson et al., 2014) and even predictive epidemiology (Su et al., 2014). A Bayesian network is a network that records the probabilities of events and based on conditional probabilities and observations it updates the final likelihood of an event. This can be seen in Figure 5.

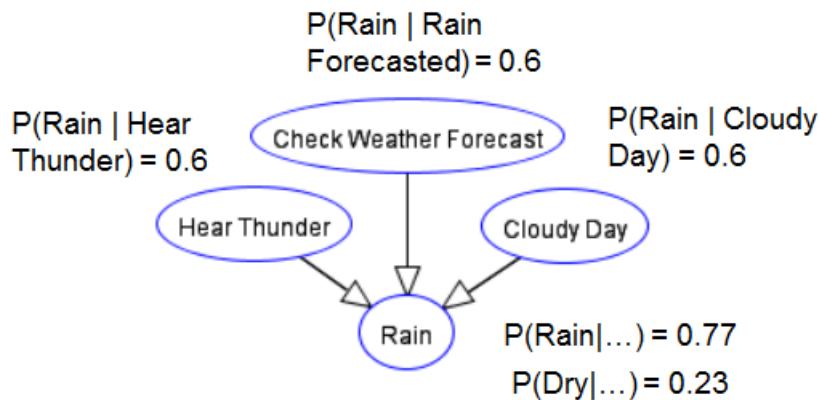


Figure 5: Sample Bayesian Network for Rain Prediction.

Here, we would like to predict how likely it is to rain. Thus, we record observations (we hear thunder, check the weather forecast, notice it is a cloudy day) and updating the likelihood of rain happening based on the conditional probabilities of $P(\text{Rain} \mid \text{Hear Thunder})\dots$ and so on. This model of learning was chosen because a Bayesian Network closely mimics the way humans think - we observe events and form co-relational and causative predictions based on those events. This is advantageous over deep learning as in deep learning we are unable to

interrogate the system to understand intuitively what the network is learning from - it has high predictive power but is essentially a black-box. The Bayesian network allows clinicians and doctors to see what are the components that went into gene ranking and prioritisation, and even change the probabilities, weights or add more observation nodes based on their personal diagnosis and treatment knowledge. Ultimately, this enables the doctors and clinicians to be able to have confidence in the software as they can analyse and understand how it works. This also allows them to be able to explain their methodology and treatment plans clearly to the patient, making the diagnosis and treatment process clear, understandable and transparent.

1.6 Aims and Approach

The overall goal of this thesis is to address the 2 major issues limiting the utility of clinical genomics through the following aims – identification and prioritization of mutations in clinical samples

We describe the development of (i) a deep learning network to identify high-confidence variant calls (focusing on SNVs and short indels) and (ii) a Bayesian network to probabilistically prioritise their functional importance. As a first step, we developed and optimised a deep learning network to identify true variants in both synthetic and real-world datasets. Following the identification of high-confidence variant calls, we will build a Bayesian network based on functional annotations to prioritise mutations and test our network on a cancer dataset.

2 Materials and Methods

2.1 Overall Analysis Structure

To enable the deep learning networks and Bayesian network analysis, we first built two main computational pipelines – the first pipeline is used to train and optimise a neural network, while the second pipeline utilises this trained neural network to perform variant validation and prediction (Figure 6). Two pipelines are required as one pipeline is needed to train the network, and the second pipeline can then use the trained network to perform predictions. The first pipeline involves using a training dataset (either simulated or obtained from real sequencing data), and performing the processing steps of alignment, variant calling and finally deep learning network training. Alignment involves mapping the sequenced reads to the correct location in the genome and variant calling involves using these mapped sequences to call variants. Deep learning training involves preparing a feature vector from this variant call and subsequently using this feature vector to train a deep learning neural network. In all, this pipeline focuses on generating a set of predictions and subsequently validating this set of predictions with a ground truth. The second analysis pipeline is meant for real datasets in which the ground truth is not known, and so it uses a pre-trained neural network from the first pipeline. For this pipeline, we perform similar steps of alignment and variant calling and then use a trained neural network to predict high-confidence variant calls. Finally, we apply the Bayesian classification analysis to determine the important genes in this dataset.

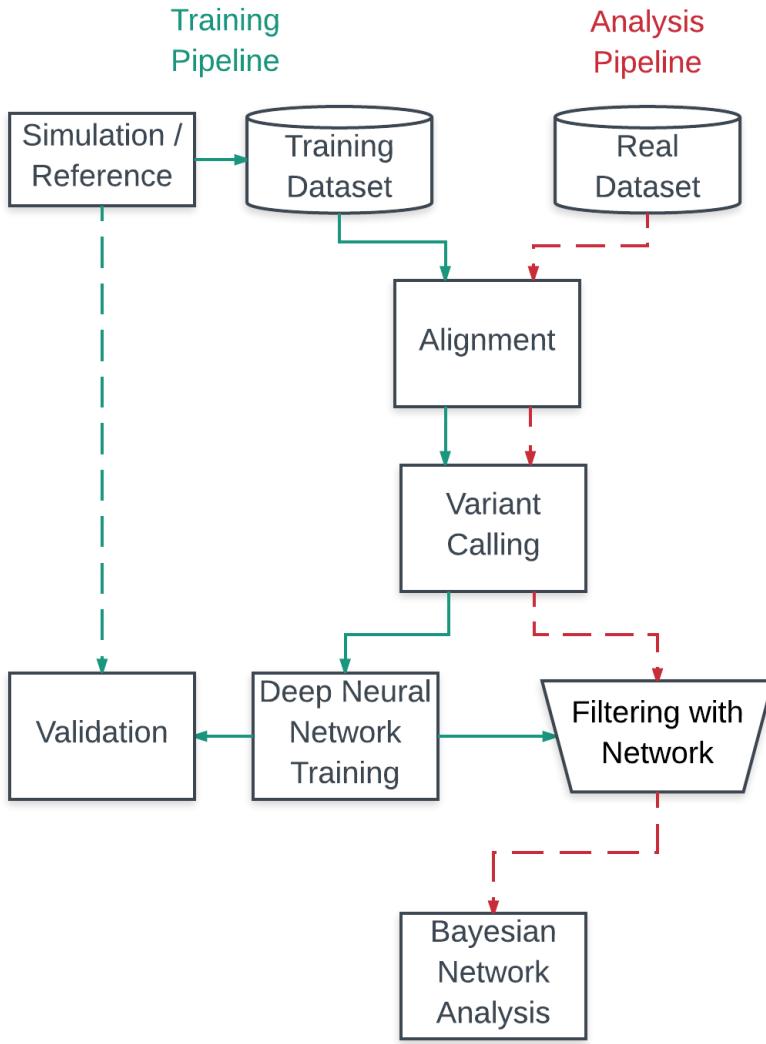


Figure 6: **Overall Structure of Analytical Pipelines.** Pipelines were implemented using the Groovy Domain Specific Language, NextFlow.

2.2 Programming and Pipelining tools

2.2.1 General Programming Language

The general programming language used was Python (v2.7). Python (Van Rossum, 2007) was chosen due to its access to various important libraries, including NumPy, Scikit-learn, Pomegranate and PyVCF. NumPy (v1.11.3) was used to prepare input vectors for deep learning training, Scikit-learn (v0.18.1) was used to perform Principal Component Analysis and Synthetic Minority Oversampling Technique Methods (See Appendix 5.3 for more information). Finally, PyVCF (v0.6.8) was used to parse the VCF files into python objects for easy manipulation. Comparison of variants was also performed using python dictionary lookups.

This method was chosen due to a high number of comparisons required, and with dictionary lookups (which are based on the hash table data-structure), each comparison can be done in $O(1)$ time.

2.2.2 Pipelining

General pipelining and chaining of programmes was done using NextFlow and Bash scripts. NextFlow (v0.21.3.3990) is a Groovy based Domain Specific Language (DSL) that provides easy writing of parallel pipelines with an accessible unix interface (Tommaso et al., 2014). Nextflow was used to run the overall pipelines and control input and output of abstracted core modules, which are in turn either python scripts or Bash shell scripts. This ensures that results are easily replicable and can be later implemented as a single analytic pipeline for clinical use.

2.2.3 Deep Learning

For our deep learning networks, we used the Keras library (v1.1.1) with a TensorFlow backend (v0.11.0). TensorFlow, which was built by Google’s machine learning team (Abadi et al., 2015), was chosen due to its distributed computation and queue management system that enabled better performance in training on a CentOS-7 compute cluster compared to other backends. The code used to generate the feature vectors and train the neural network can be found in Relevant Code – Section 7.1 and 7.2 respectively. For more explanation on the algorithms underpinning deep learning, see Appendix 5.1 for more information.

2.2.4 Gene Ranking

Finally, ANNOVAR (v2015Jun17) was used to generate the functional annotations for the Bayesian probabilistic model (Wang, Li, & Hakonarson, 2010). The protocols used were.snp138, clinvar_20150629 and ljb26_all. Pomegranate (v0.6.1), a Python library, was used to generate and compute the probabilistic model and ranking system for our Bayesian Network. For the probabilistic model, preprocessing was done using Python scripts and subsequently used as input in Pomegranate. The code used to generate the Bayesian network can be found in Relevant Code – Section 7.3.

2.3 Simulated Datasets

Simulated genomes enable the simulation of NGS data with ground truths to test and validate a neural network. For our simulator, we used Mason, a genome mutation software written in

C++ (v2.3.1) to mutate the hg19 reference genome from UCSC (Karolchik et al., 2014). We used indel rates of 0.00002 and SNP rates of 0.00008 to generate sufficient truth variants for analysis, which comprise 229253 SNPs and 57257 indels.

After generating a ground truth model, we simulated sequence reads with error rates and ground truth variants (Figure 7). For error rates, we used published data from Schirmer et al. (2016) as the input to Mason – the general substitution error rate used was 0.0004 per base in the genome, and the insertion and deletion error rate per base were $5 * 10^{-6}$.

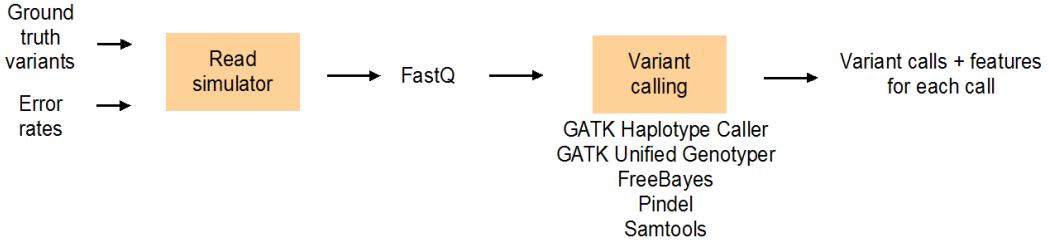


Figure 7: Pipeline for Simulation of Genomes for Analysis

2.4 Alignment and Variant Calling

To perform alignment of simulated and real sequences, the Burrows-Wheeler Aligner (Li, 2013), version 0.7.13, was used. Default settings were used, with the mem option which is known to work well with longer sequences. After alignment, variant calling was performed. Variant callers used were FreeBayes (v1.0.2-16), GATK Haplotype Caller (v3.7-0) and Unified Genotyper (v3.7-0), Samtools (v1.3.1) and finally Pindel (v2.3.0)(Garrison & Marth, 2012; McKenna et al. 2010, DePristo et al. 2011; Li H, et al., 2009; Ye et al., 2009). All callers were used at their default settings.

2.5 Feature Engineering

In order to train a neural network, features in the form of numerical vectors must be used as an input. We subset our features into three broad sets, which are base-specific information, sequencing error and bias information features, and calling and mapping quality. Here we describe the computation of the key features – please see Appendix 5.2 for a more in-depth explanation on their usage and interpretation.

Base Information

Shannon Entropy

Shannon Entropy captures the amount of information contained inside the allele sequences. It is calculated using the equation:

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i) \quad (1)$$

where $P(x_i)$ is the prior probability of finding each base at each position. This prior probability is calculated in two ways – over the entire genome and over a region of space around the allele (10 bases plus the length of the allele in our calculations).

Kullback Leibler Divergence

The Kullback-Leibler Divergence feature is similar to Shannon entropy, but instead, we use this to measure the informational gain from the reference to the allele sequence. The Kullback-Leibler Divergence is calculated as follows:

$$D_{KL}(P||Q) = - \sum_{i=1}^n P(x_i) \log_2 \frac{P(x_i)}{Q(x_i)} \quad (2)$$

where $Q(x_i)$ is the prior probability of finding each base at each position based on the genomic region around the allele, while $P(X_i)$ is the posterior probability of finding a specific base inside the allelic sequence.

Base Quality

Base quality refers to the Phred score probability that the called allele is wrong. It is given by the equation:

$$P = 10^{-\frac{Q}{10}}$$

Where P is the Base Quality, and Q is the probability that the allele called is wrong. This is a number computed by the sequencing machine based on the quality of the base samples provided.

Sequencing Biases and Errors

GC content

This feature comprises the GC content of the reference genome for at least ten bases around the mutation site.

Longest homozygous run

This feature comprises the longest similar string of bases in the reference genome, for at least ten bases around the mutation site.

Allele Count and Allele Balance

This feature is an output from Haplotype Caller and Unified Genotyper, and describes the total number of alleles contributing to a call and the balance between reference and alternate alleles reads.

Calling and Mapping Qualities

Genotype Likelihood

The genotype likelihood score provides the Phred-scaled likelihood scores of how confident the caller is in determining that it is a homozygous or heterozygous call, and is provided by all variant callers.

Read Depth

Mapped read depth refers to the total number of bases sequenced and aligned at a given reference base position. It is provided by all variant callers.

Quality by Depth

Quality by depth is computed by dividing the quality score against allele depth, to obtain an average score of allele quality. This is provided by Haplotype Caller and Unified Genotyper.

Mapping Quality

Mapping quality is a score provided by the alignment method and gives the probability that a read is placed accurately. It is provided by all variant callers except Pindel.

2.6 Patient Derived Xenograft Mouse Model Development and Sequencing

[TO CLARIFY] To test the Bayesian ranking system, we used a patient-derived xenograft mouse model. Athymic mice with the FOX mutation were grown for X number of days, and subsequently, a tumour is grafted onto the mouse's body. Subsequently, the tumour was sequenced on an Illumina MiSeq platform and used for analysis.

3 Results

3.1 Generation of Simulated Datasets

As a first step towards developing a deep learning network, we generated a simulated genome with mutations using the hg19 genome from UCSC (Karolchik et al., 2014) as a reference. The simulated genome contains over 300,000 random mutations spread over the chromosomes as can be seen below in Figures 8 and 9. Simulated genomes are a good method to analyse deep learning networks with, as the ground truth, which are the truth variants inside the genome, are already known. This allows accurate verification of prediction schemes and is a commonly used method to test next generation sequencing related software (Escalona, Rocha & Posada, 2016). This is primarily because it is difficult to obtain complete truth datasets for real genomes as due to the inhibitory cost of checking every variant called via Sanger sequencing. Thus, simulated genomes present a simple way to simulate NGS data with perfectly known ground truth variants to test our validation platform.

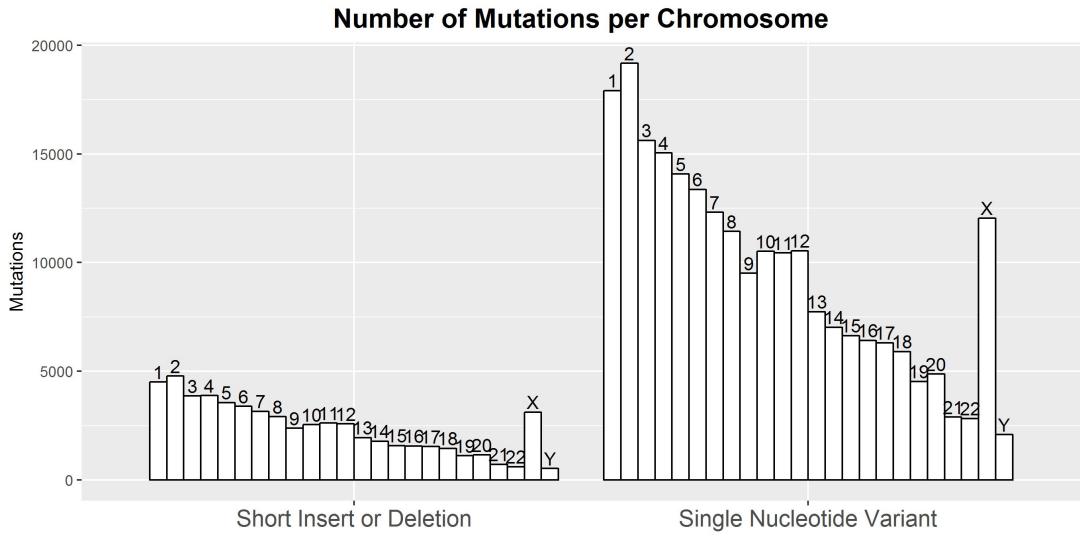


Figure 8: Number of Ground Truth Mutations (Variants) Created in Each Chromosome

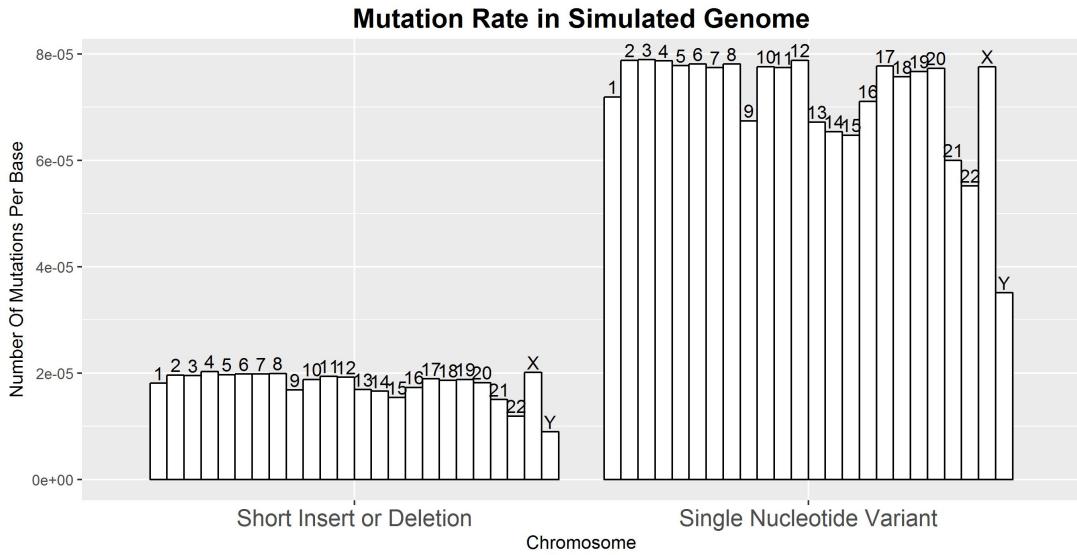


Figure 9: Mutation Rate per Base in Each Chromosome

3.2 Feature Engineering

Deep learning requires features to be vectorised and normalised to be used as input data. We engineered a set of 20 features to use as input data for our deep learning network, using data obtained from the variant callers themselves as well as engineering other features from the dataset. A summary of the features used in training can be found in Table 1, and a description of the full list of features can be found in Appendix 5.2. Features were engineered based on obtaining information on the main aspects of variant calling, which includes the information contained in the sample bases (Base Quality, Entropy, Kullback–Leibler divergence, etc.), the confidence we have in the calling and alignment (Read Depth, Mapping Quality etc) and finally possible biases in the sequencing machine (Allele Balance, Allele Count, GC content).

Table 1: Key Features Engineered from each Caller.

Features	Shannon Entropy (Reference, Alternate and KL-Divergence)	Base Composition (Homopolymer Run, GC content)	Read Depth	Mapping Quality	Base Quality	Allele Balance	Quality by Depth	Allele Count	Genotype Likelihoods
Free Bayes	+	+	+	+	+	+	+		+
Haplotype Caller	+	+	+	+	+	+	+	+	+
Unified Genotyper	+	+	+	+	+	+	+	+	+
Pindel	+	+	+						+
Samtools	+	+	+	+	+	+			+

3.3 Variant Callers

Variant callers were chosen for our deep learning neural network based on their orthogonal calling and reference methodologies – we wanted to optimise the information that the neural network receives (See Table 2). We used two haplotype-based callers, FreeBayes (Garrison & Marth, 2012) and GATK Haplotype Caller (McKenna et al. 2010, DePristo et al. 2011), two position based callers GATK Unified Genotyper and Samtools (Li H, et al., 2009) and finally Pindel, a pattern growth based caller (Ye et al., 2009). When we analysed the concordance rates of the callers on the simulated dataset, we found a high amount call discordance (Figure 10).

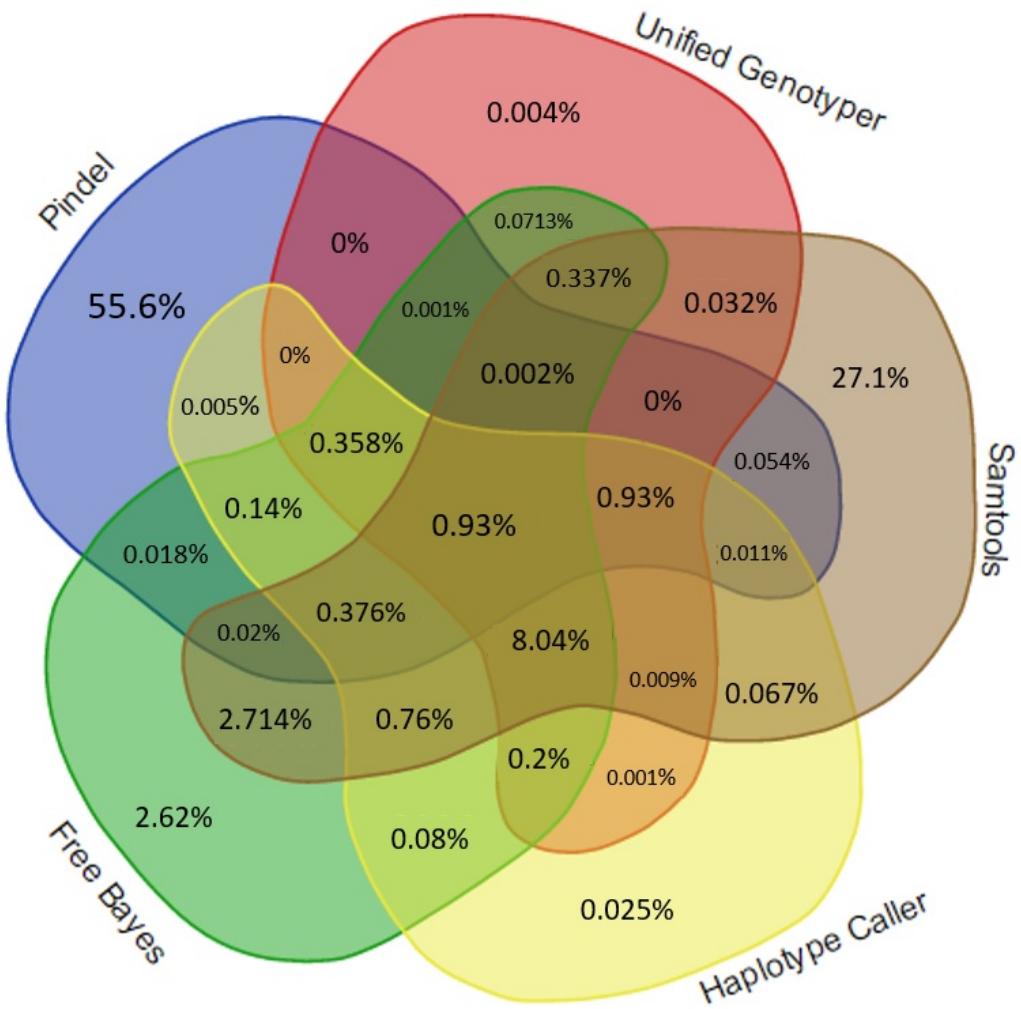


Figure 10: Concordance of Callers on Simulated Dataset

Of all the callers, Pindel was the most discordant caller, with over 1.6 million (55.6%) unique calls that are different from other calls. Samtools was also discordant, with over 800

thousand unique calls (27.1%) that were unique from the other callers, followed by FreeBayes at 80,000 calls. Interestingly, a high amount of calls (about 100,000) was called by exactly two callers. Discordance in the variant callers can be explained by the different methodologies that they use to call variants (Table 2). Due to implementation and design choices, as well as statistical methods, each variant caller has a different calling profile. Discordance in the callers provides a strong argument using deep learning to integrate the information from all the callers in a sophisticated manner.

Table 2: Comparison of Different Methods and Features of Variant Callers.

	GATK Unified Genotyper	Samtools	GATK Haplotype Caller	Free Bayes	Pindel
Calling Method	Uses a list of mapped reads, calling model is probabilistic with increased priors at regions with known SNPs	Uses a list of mapped reads, calling model is probabilistic. Does not assume sequencing errors are independent and has less hard filters compared to Unified Genotyper	Uses Hidden Markov Models to build a likelihood of haplotypes which are then used to call variants	Uses a posteriori probability model to build a set of haplotypes to represent mutations, calling model is probabilistic with population based priors	Locates regions which were mapped with indels or only one end was mapped, and then performs a pattern growth to find inserts and deletions. Shown to be able to identify medium length indels missed by other callers in real samples (Spencer et al., 2013)
Reference and Mapping Method	Position based caller that realigns fragments and analyses each position to call SNPs and indels	Position based caller that uses mapped sequences to call SNPs and indels.	Analyses regions where there is high likelihood of mutation based on activity score, and builds a De Bruijn-like graph that reassembles reads (Haplotypes) in that region	Dynamic sliding window based reference frame, using algorithms to determine window size for analysis. Does not require precise alignment, unlike other callers	Focuses on Unmapped regions, regions known to have insert and deletions or regions with only one end mapped.

3.4 Network Architecture

Next, we tested out various neural network architectures to see which architecture would perform the best for our set of input features. We first explored the flat architecture (Figure 11), which contains stacks of fully connected layers with multiple nodes (initially seven layers, with 80 nodes per layer). This is the simplest architecture, where all the features are loaded onto a single vector, and this entire vector is used as an input to train the neural network. We next explored the PCA + flat architecture which had the same neural network architecture but before the input data was fed into the network, a Principal Components Analysis was done to reduce the dataset to 8 principal components which were then used as input data for the neural network (please see Appendix 5.3.1 for more details of the PCA analysis). Principal components analysis is a dimensionality reduction technique that enables a compressed representation of

data. Each principal component is a linear summation of the original features (X) in the form

$$PC_1 = \beta_{1,1} * X_1 + \beta_{2,1}X_2 + \dots + \beta_{n,1}X_n$$

...

...

$$PC_i = \beta_{1,i} * X_1 + \beta_{2,i}X_2 + \dots + \beta_{n,i}X_n$$

which enables a few principal components to capture a high amount of variance in the dataset. Finally, the last architecture we tested was the merged network this network had a set of layers (initially five layers, 24 nodes per layer) that learns from each caller alone, and then the outputs from each of these layers are subsequently merged and used to make a prediction.

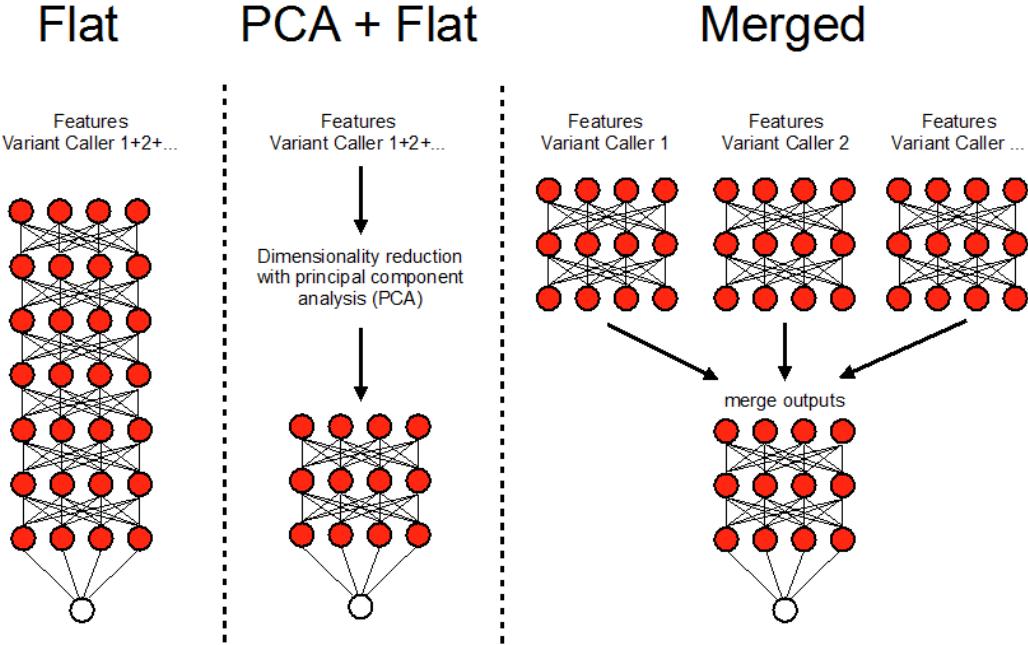


Figure 11: **Different Designs for Neural Network Architecture**

To study how well each architecture is able to perform, we use the metrics of precision, recall and F1 score. Precision measures how many mistakes the predictor makes (the ratio of true positives over false positives and true positives), recall measures what portion of the truth class a predictor can discover (the ratio of true positives over true positives and false negatives) and finally the F1 score is composite function of both precision and recall. The derivations of the metrics can be found in Appendix 5.3.1. Figure 12 shows the precision, recall and F1 score of the three different architectures.

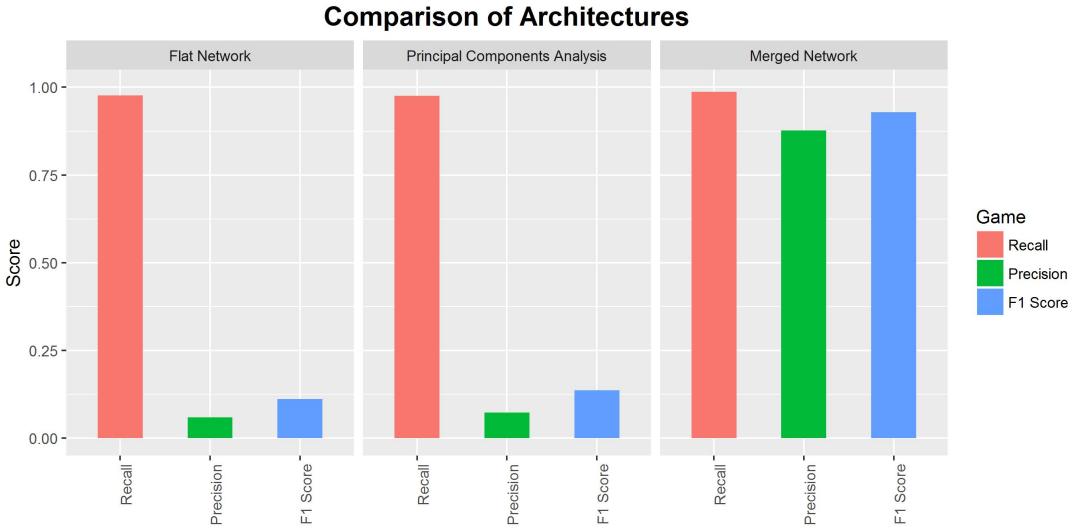


Figure 12: **Analysis of Different Neural Network Architecture**

Initially, with the flat network, the precision rate was very low at 0.059 with an F1 score of 0.112, indicating that the neural network was unable to learn from the input feature set. We suspected that this was due to high dimensionality in the dataset, which led to our second architecture design, the PCA with flat analysis. Principal components analysis has been shown to be able to successfully improve learning in high-dimensionality datasets (Chen et al., 2014; Van Der Maaten, Postma & Van den Herik, 2009). However, the precision and F1 score for the PCA architecture was also low at 0.0735 and 0.137 respectively. Ultimately both failed to learn, indicating to us that perhaps the features from each of the callers had to be analysed separately before being passed into a separate neural network that did the final output integration. With this merged network, we managed to obtain a precision score(0.877) and an F1 score(0.929) that was far better than the previous two architectures. Interestingly, the recall scores for all three architectures were around the same (± 0.01), indicating the main difference for the neural network was in its ability to remove false positive calls.

3.5 Network Tuning and Optimisation

Next, we systematically optimised and tuned the deep learning neural network to maximise its predictive ability. In tuning our network, we also sought to study how the various hyperparameters as well as the data structure affected our network's ability to learn from the data. In particular, we focused on four issues – the number of layers, optimiser choice, learning rate choice and finally sample balancing. These four issues are known to be critical in deep learning networks (Ruder et al., 2016; LeCun, Bengio & Hinton, 2015; Yan et al., 2015; Sutskever et

al., 2013) and would likely be critical to the success of a deep learning neural network.

3.5.1 Number of Layers

Firstly, we studied how many layers should be in the neural network. The number of layers is critical as it determines what kind of information and the representation of data that can be captured by the neural network. Choosing the number of layers is important as sufficient layers are needed to obtain the complex data representation needed for learning, but too many layers might result in the vanishing gradient problem (Sutskever et al., 2013; Bengio et al., 1994). Our initial neural network architecture is shown below (Figure 13), and then we varied the number of layers in at each point.

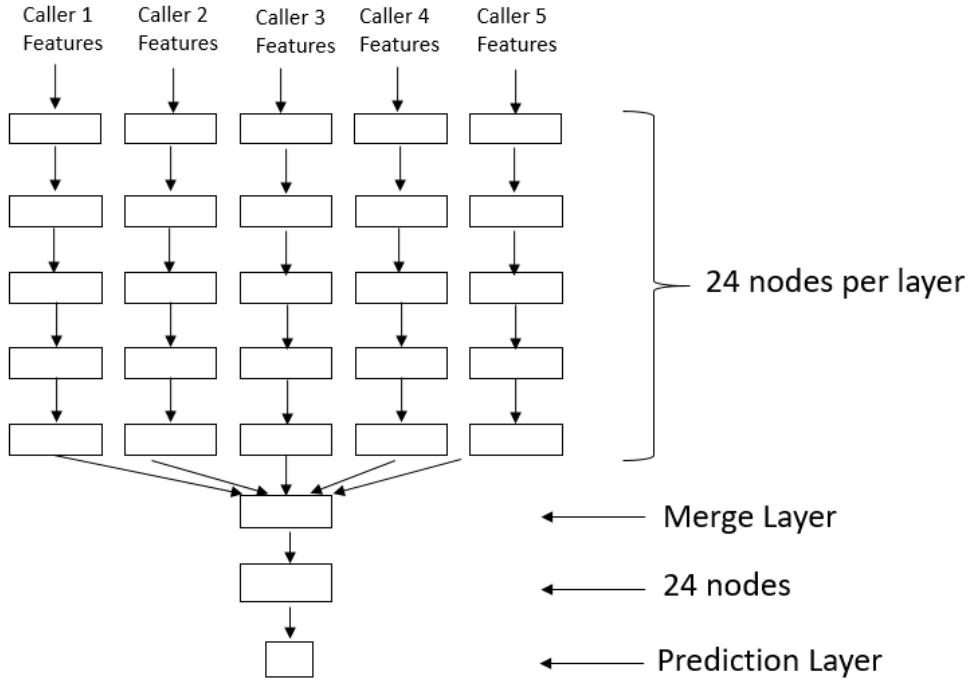


Figure 13: **Basic Merge Network Structure.** Each individual pre-merge layer takes in an input from a single caller. The information is then integrated in a set of merge layers to give a prediction.

For all layers, the LeakyReLU activation function was used. The LeakyReLU is a refinement of the ReLU activation function which minimises the "dying ReLU" problem, and both are well-documented activation functions that have been shown to work well in deep neural networks (Anthimopoulos et al., 2016; LeCun, Bengio & Hinton, 2015; Maas, Hannun & Ng, 2013). We

noticed that changing the number of layers after the merge layer did not significantly vary the output, and so we focused on changing the number of layers before the merge layer. We studied 6 different neural network structures (4 layers to 9 layers). Accuracy was used as the main metric to compare the neural network architectures, and is defined as the fraction of all samples that the neural network is able to correctly predict.

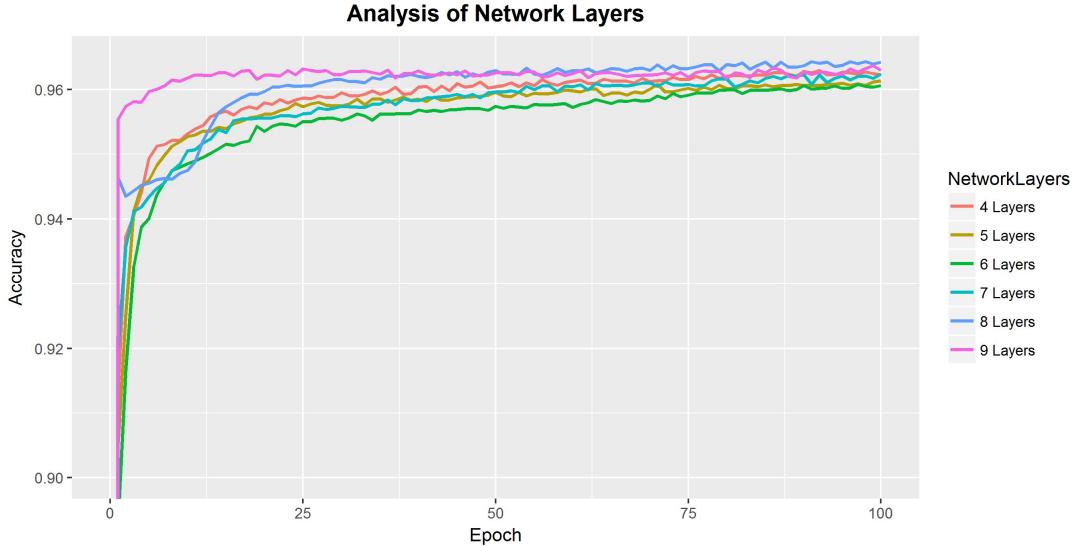


Figure 14: Analysis of Different Number of Layers On Training Accuracy

From Figure 14, we find that the 8 layer neural network seem the best at learning from the input data, with a final accuracy of 0.964 that is about 0.001 higher than other layers. We note that all the layers follow the same rough trend of accuracy, indicating they are all able to learn from the dataset. A final design feature used was to add two dropout filters at the last two layers before merging in order to prevent overfitting in data. Dropout filters have been shown to be an effective in preventing overfitting of data (Srivastava et al., 2014).

3.5.2 Optimiser and Learning Rates

Next, we sought to choose the best optimiser and learning rate for our dataset. Both optimisers and learning rates have been well studied and known to be important in neural network training (Ruder et al., 2016; Sutskever et al., 2013). Optimiser choice is critical as the optimisers determine how the weights and gradients are updated in the network, thus playing an integral part in learning. We studied 3 well-known optimisers for use in our network, ADAM, RMSprop and Stochastic Gradient Descent (SGD). ADAM is an adaptive learning rate optimiser that

is known to be well suited in large dataset and parameter problems(Kingma & Ba, 2014). RMSprop is another adaptive learning rate optimiser that is unpublished, but has been shown to work well for real experimental datasets (Tieleman & Hinton, 2012). SGD is the simplest learning model with no adaptive learning rate but is a useful model because it is the easiest to understand mathematically and has also been shown to solve deep learning problems (Kingma & Ba, 2014). For more information on the mathematical foundations of optimisation and backpropagation, please see Appendix 5.1. For the three optimisers, we ran tests to study the accuracy of the neural network running on each optimiser to predict true variants (Figure 14).

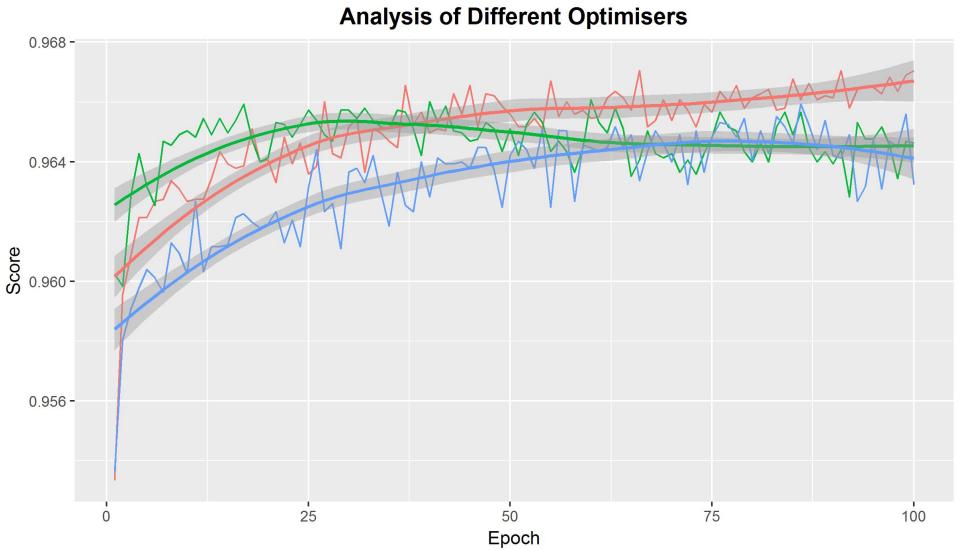


Figure 15: **Optimiser Accuracies for Training at each Epoch.** Due to the noise in accuracies, the overall momentum of the dataset, calculated as a sliding window average is shown. The 95% confidence interval is also shown.

Adam obtained the highest accuracy of 0.9670, while RMSprop and SGD reached maximum accuracies of 0.9660 and 0.9569 respectively. Interestingly the adaptive rate optimisers seemed to have complex learning trajectories, while SGD has a very stable learning rate. This makes sense as adaptive learning rates allow greater gradient descents when the error is high, and decreasing the learning rate at smaller errors (Kingma and Ba, 2014; Zeiler, 2012). This allows Adam and RMSprop to learn at variable rates based on the current gradients. For SGD, it appears that while it takes a while to learn the true minima, it eventually still reaches about the same minima as RMSprop. Ultimately, we chose Adam as our optimiser as the final accuracy discovered by Adam was noted to be higher than RMSprop and SGD, and we note a stable learning curve for Adam, indicating it is able to learn and update the gradients in the neural network to learn from input data at all epochs. Subsequently, we also looked at various initial learning rate for Adam (Figure 16) and found that the most stable learning could be found

at a learning rate of 10^{-5} . This initial learning rate is critical as it determines the first few gradient descents which enable stable adaptive learning throughout the epochs (Sutskever et al., 2013). At any larger learning rates (10^{-4} and below), a very high amount of noise was observed, indicating that the learning rate was too high resulting in minima finding errors. At smaller learning rates (10^{-6} and above), the final accuracy after 100 epochs (0.9639 for 10^{-6}) was lower than the learning rate at 10^{-5} (0.9672). Thus, we chose 10^{-5} to be our learning rate.

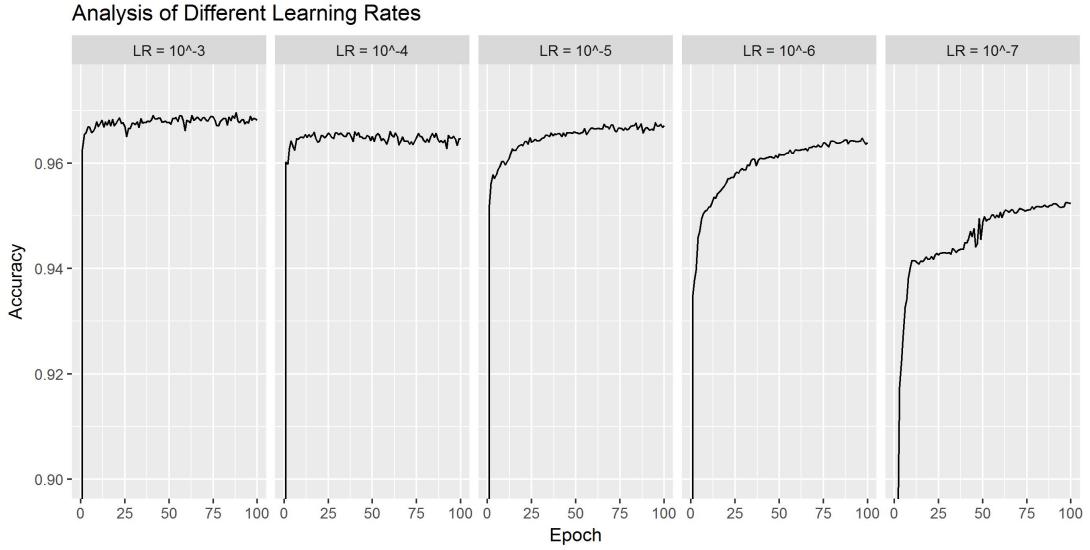


Figure 16: Training Accuracies over Each Epoch for Different Learning Rates

3.5.3 Sample Balancing

Our final concern was sample balancing – the simulated dataset contained an imbalance of positive training examples versus negative training examples. In total, there were 286,510 positive training examples and 4,547,919 negative training examples (about a 15-fold difference). Such a sample imbalance has been known to affect learning adversely (Yan et al., 2015; López et al., 2012). Thus, we sought to implement two methods of sample balancing, undersampling and oversampling. Undersampling was implemented by removing negative training examples until the number of negative training examples was equal to the number of positive training examples. In oversampling, the Synthetic Minority Oversampling Technique(SMOTE) was done, which uses nearest neighbours to create more data points for the positive training example. Specifically, SMOTE looks at two nearby positive class examples, and creates a new synthetic example in the middle of these two examples (see Appendix 5.3 for more details). Figure 17 shows the metrics for each sampling technique.

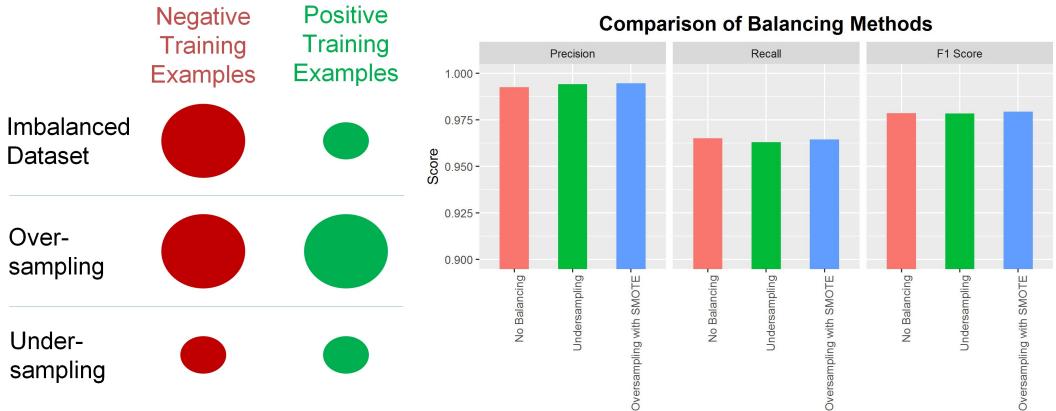


Figure 17: a) Graphical Illustration of Sample Balancing. b) Effect of Sample Balancing Techniques on Prediction Ability.

Interestingly, we note that overall, undersampling, oversampling and no sampling at all had very small effects on precision, recall and the final F1 score. Specifically, all three metrics were within a range of 0.003 for the different techniques. This could be due to clear boundary separation within positive and negative class examples as well as good representative datapoints within the positive training example class. This prevents the imbalanced data from having too much of an effect on variant prediction and classification. Still, we note that oversampling techniques resulted in a marginally higher F1 score (0.001 higher than undersampling and no sampling), and since ensuring that datasets are balanced is a recommended protocol to prevent further bias downstream (Chawla, 2005), we used SMOTE oversampling to produce extra positive training class examples for all analysis pipelines.

3.6 Benchmarking of Optimised Network with Mason Dataset

From optimisation steps, we finalised the network architecture as seen in Figure 9, but with 8 layers before the merge layer. We chose the learning rate to be 10^{-5} , and the optimiser used was Adam. With this network, we benchmarked the neural network against the single variant callers, as well as concordance callers, which are an integration of the outputs of the 5 variant callers. Specifically, the n-concordance variant caller is defined as the set of calls that any n callers agree upon – so 1-concordance includes all the calls made by all callers and 4-concordance includes all the calls made by any 4 callers.

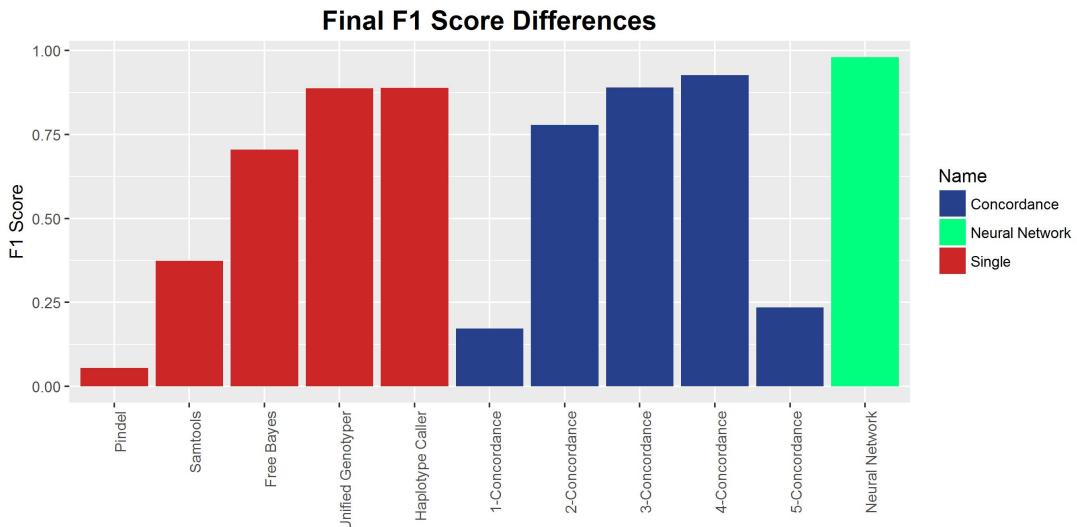


Figure 18: **Overall Comparison of Variant Callers**

In terms of overall F1 score (Figure 18), we see that the neural network was able to outperform single and concordance-based callers. This provides strong evidence that the neural network is able to learn from the input features whether the variant call is real or not, validating its usage in variant calling. The final F1 score obtained by the best single variant caller was Haplotype caller at 0.888, the best concordance caller had an F1 score of 0.927 while the neural network achieved an F1 score of 0.980. To study whether the increase in F1 score is due to improvements in precision or recall, we studied the exact precision, recall and F1 scores of the top 2 variant callers as well as the best single variant caller versus the neural network. We find that the neural network is more precise than both, but the recall is rather similar (Figure 19).

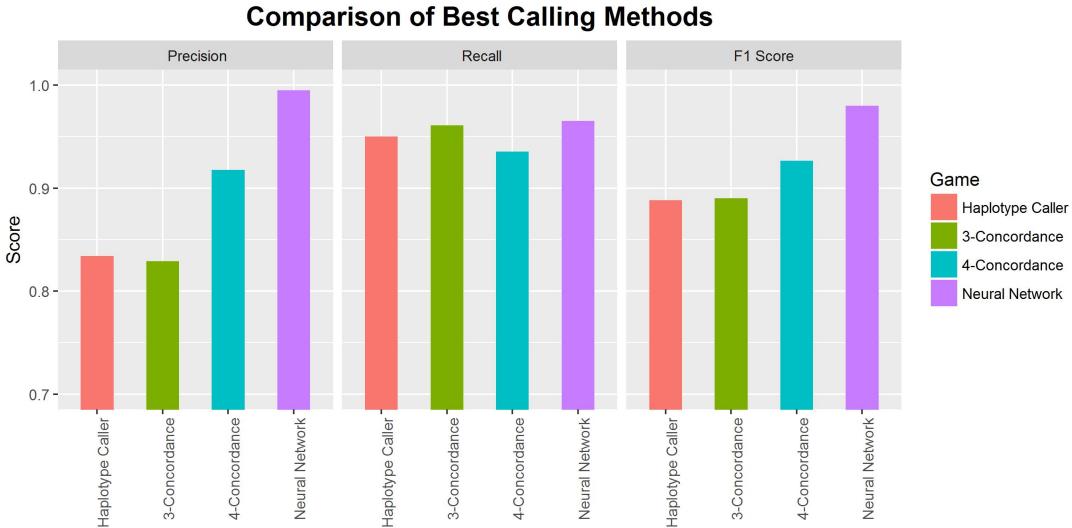


Figure 19: **Comparison of Best Variant Callers in terms of Precision, Recall and F1 Score**

From Figure 18, we see that the neural network is more precise than all four callers, and had the highest precision of 0.995 compared to only 0.917 for 4 concordance. Specifically, this is a 20 fold decrease in the number of false positives or about 23,000 more false positive calls in the 4 concordance network compared to the neural network. Interestingly, the recall of all the callers was high in the range of 0.90 to 0.95, indicating that while all were able to pick out most of the truth class variables, the main errors came from a high number of false positives. Ultimately, the neural network had an F1 score that was 11% above the best single caller and 6% above the best concordance caller. Thus, this provides strong evidence that the neural network is able to sieve out false positives within the dataset and stably predict whether a mutation is true.

3.7 Benchmarking of Optimised Network with NA Dataset

After verification of the optimised neural network on a simulated dataset, we sought to analyse a real dataset to test the validity of the neural network in validating variants. We studied the NA12878 Genome In a Bottle dataset (Zook et al., 2014), which has been used in other variant calling validation pipelines (Talwalkar et al., 2014; Linderman et al., 2014) and contains a set of high-confidence variant calls which we can use as ground truth for training and validation. This set of high-confidence variant calls is obtained from multiple iterations of orthogonal sequencing methods (using Solid, Illumina platforms, Roche 454 sequencing and Ion Torrent technologies). The usage of multiple platforms enables an intersection of variants that can be

considered as the ground truth. We then sought to see if our neural network can predict the ground truth better than single or concordance based variant callers.

We applied the same methodology to the sequences as with the simulated data and then used our neural network to predict the true variants. Validation was done with 47971 high-confidence variant calls in total.

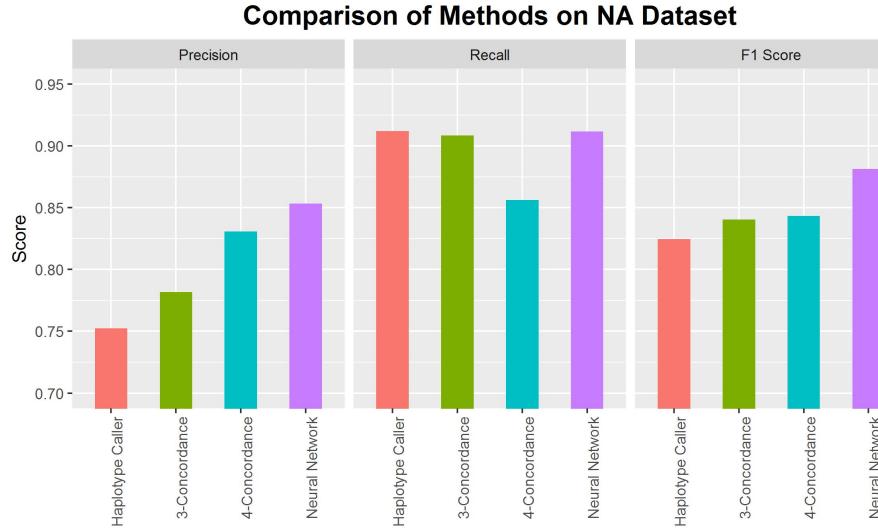


Figure 20: **Comparison of Variant Callers**

As can be seen from Figure 20, the neural network was able to predict with the highest precision (0.859) when compared the best single caller, haplotype caller (0.752) and the 2 best concordance callers, 3-concordance (0.782) and 4-concordance (0.830). In terms of recall, the neural network had a higher recall (0.911) compared to the 4 concordance caller (0.856). Thus, we see that in the NA dataset, the neural network compared with the 4 concordance network is able to call 2650 true variants that were missed by 4 concordance and still had 1228 less false positives. This means the neural network was more aggressive in making calls, yet more of the calls were correct. Compared to the three concordance caller, the neural network had 4253 less false positives. Ultimately when we looked at the F1 score, the neural network was able to outperform concordance variant callers by at least 0.04 and single callers by 0.06. This validates our neural network pipeline in a real genomic dataset and indicates that network is able to learn from the input features.

3.8 Analysis of Gene Importance using Bayesian Ranking systems

After validation of high confidence calls using a deep learning network, we proceeded on to designing a Bayesian network for the clear and understandable ranking of genes. We first build a Bayesian network using known functional annotations from ANNOVAR (Figure 21).

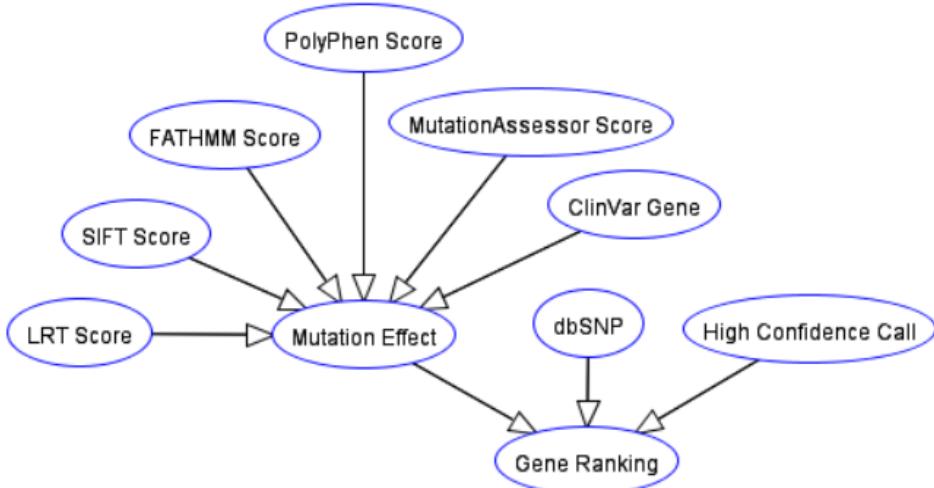


Figure 21: Final Bayesian Network used in Analysis

This network structure was chosen as we wanted to use three different sets of information to update the probability of the gene being important. Firstly, the confidence of the call should matter in how important it is – the more likely a gene is real, the more important it should be. Secondly, the rank should also be determined by how common the variation is, based on studying known SNP polymorphism rates. If it is a common SNP, then the ranking should be downgraded as it is less likely to be a driver mutation (Schork et al., 2009). Finally, we sought to predict the overall effect of mutations via an ensemble of mutation effect predictors. These predictors use different methods to predict the average effect of that mutation – based on statistical methods like position-specific substitution matrixes and Hidden Markov Models to study the effect of a mutation on protein structure and function. We also used the ClinVar database, a curated repository of known Human variants and their resulting phenotypes (Landrum et al., 2014). These scores were then aggregated to update the probability of the mutation effect. To obtain these functional annotations, the informatics tool ANNOVAR (Wang, Li, & Hakonarson, 2010) was used. Table 3 shows the functional annotations obtained from ANNOVAR and how they were computed.

Table 3: Functional Annotations obtained from ANNOVAR

Annotation Name	Information Type	Method	Scoring Method
Likelihood Ratio Test	Deleterious Mutation Score	Likelihood Ratio Test of each amino acid is evolving neutrally to the alternative model of evolution under negative selection	Score normalised to [0,1] and used directly in Bayesian Network
MutationAssessor	Deleterious Mutation Score	Mutation rate of homologous sequence subfamilies	Score normalised to [0,1] and used directly in Bayesian Network
SIFT	Deleterious Mutation Score	Position Specific Scoring Matrixes with conserved Sequences	Score normalised to [0,1] and used directly in Bayesian Network
PolyPhen2	Deleterious Mutation Score	naïve Bayes classifier on various multiple sequence alignments methods of homologous proteins and protein structure-based features	Score normalised to [0,1] and used directly in Bayesian Network
FATHMM	Deleterious Mutation Score	Hidden Markov Model used to score MSA based on protein homologous sequences	Score normalised to [0,1] and used directly in Bayesian Network
ClinVar Genes	Known Pathogenic Genes	Database lookup of curated set of relationship between variant calls and human phenotype	Higher Probability of Importance if known pathogenic variant
dbSNP138	Common Single Nucleotide Polymorphisms	Database lookup of curated set of known Human SNPs	Lower Probability of Importance if known common variant

These were subsequently used to compute the Bayesian probability ranking, which is shown in the equation below. Based on scores provided, we report the update the conditional probabilities using the probabilities chain rule – for the first level; this is given as

$$P(Impt|(Del \cap Uncom \cap High\ Conc)) = P(Impt \cap Del \cap Uncom \cap High\ Conc) * P(Del \cap Uncom \cap High\ Conc) \quad (3)$$

$P(\text{Impt})$ refers to the probability of the gene being important,
 $P(\text{Del})$ refers to the probability of the gene being deleterious,
 $P(\text{Uncom})$ refers to the probability of the gene being uncommon and
 $P(\text{High Conc})$ refers to the probability of the gene being a high confidence call.
Further calculations can be found, and derivations can be found in Appendix A

To compute the final probabilities, the software library Pomegranate was used. This simplifies the node drawing and probabilistic updates of the final ranking scores (see Relevant Code – Section 7.3).

3.9 Validation of Bayesian Network Ranking on PDX dataset

To study the effectiveness of our Bayesian network ranking system, we sequenced and analysed a patient-derived xenograft (PDX) tumour genome. This tumour genome was grafted onto the immunocompromised mouse from a patient with a known cancer subtype – Diffuse Large B-Cell Lymphoma (DLBCL). We chose to analyse lymphoma as it is a well-known and studied disease model with a well-defined disease progression (Knudson et al., 2001; Alizadeh et al., 2000). The patient-derived xenograft model also allows *in vivo* studies of the tumour in its environment and serves as a good model for sequencing and analysis (Tentler et al., 2012). After sequencing the PDX genome, we put it through our full analysis pipeline, which involves

identifying high-confidence mutations using the neural networks and then ranking these genes using the Bayesian network ranking. Figure 22 shows the top 30 genes by probability.

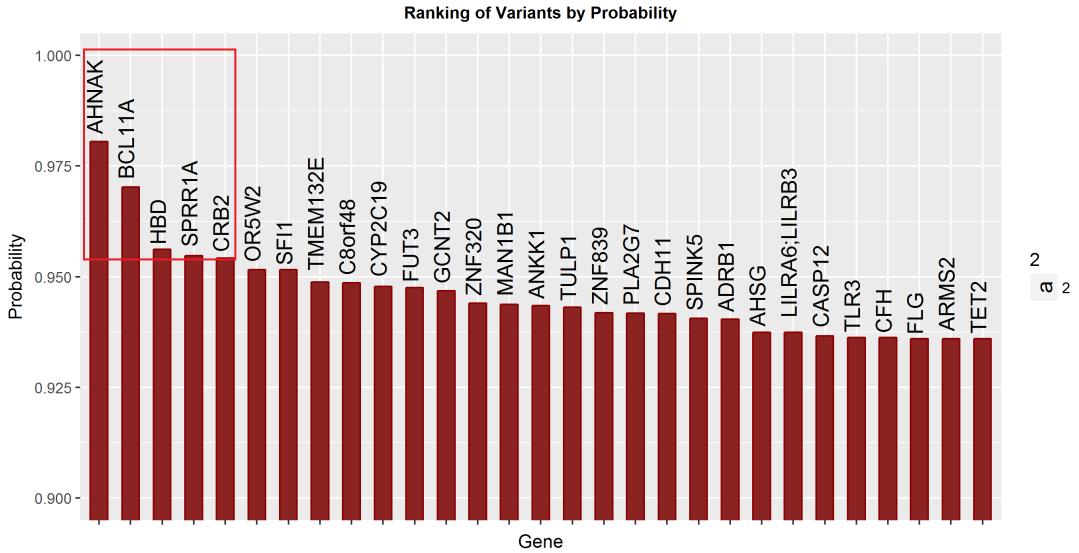


Figure 22: Top 30 genes from Bayesian Ranking Algorithm

Studying the top 5 genes, we found that four of these five genes have been implicated in lymphomas or other cancers (Table 4). AHNAK is a known tumour suppressor and has been known to be downregulated in lines of Burkitt Lymphoma (Lee et al., 2014; Amagai et al., 2004; Shtivelman et al., 1992). BCL11A is a known proto-oncogene in DLBCL and has been found to be overexpressed in 75% of primary mediastinal B-cell Lymphomas, a subset of DLBCL (Weniger et al., 2006; Satterwhite et al., 2001). SPRR1A, the fourth gene ranked in terms of importance, has been shown to be expressed in DLBCL (Zhang et al., 2014) and its expression has been shown to strongly correlate with 5-year survival rate (Figure 23). Finally, development of B-cell lymphoma has been noted in CRB2 related syndrome, which is a bi-allelic mutation of CRB2 (Slavotinek, 2016; Lamont et al., 2016). Interestingly, the last of the high ranked genes was noted to be a subunit of Hemoglobin. While there is no strong evidence for the role of Haemoglobin in DLBCL, it has been shown to be expressed in aggressive glioblastomas lines, indicating a possible previously unknown role in cancer (Emara et al., 2014). This gives us high confidence that the Bayesian ranking method can pick up important and relevant mutations. Without such a ranking system, we would have to look through over 70 thousand genes, without a way to systematically study their likelihood of being important.

Table 4: Highest Ranked Genes from Bayesian Ranking

Gene	Full Name	Known Involvement in Lymphoma or Cancer	Evidence	Mutation Location	Predicted Mutation Type
AHNAK	Neuroblast Differentiation-Associated Protein (Desmoyokin)	<ul style="list-style-type: none"> Known tumour suppressor via modulation of TGFβ/Smad signalling pathway Known to be downregulated in cell lines of Burkitt lymphomas 	Lee et al., 2014; Amagai et al., 2004; Shtivelman et al., 1992	chr11 - 62293433 T -> C	non synonymous SNV
BCL11A	B-Cell CLL/Lymphoma 11A	<ul style="list-style-type: none"> Known proto-oncogene in DLBCL Overexpression of BCL11A was found in 75% of primary mediastinal B-cell lymphomas (a subset of DLBCLs) 	Weniger et al., 2006; Schlegelberger et al., 2001; Satterwhite et al., 2001	chr2 - 60688580 C -> G	non synonymous SNV
HBD	Hemoglobin Subunit Delta	<ul style="list-style-type: none"> Shown to be expressed by aggressive glioblastoma cell lines 	Allalunis-Turner et al., 2013	chr11 - 5255274 G -> A	stop-gain
SPRR1A	Small Proline Rich Protein 1A (Cornifin-A)	<ul style="list-style-type: none"> Known to be expressed in DLBCL and expression has been shown to correlate with 5 year survival rate 	Liu et al., 2014	chr1 - 152957961 G -> C	non synonymous SNV
CRB2	Crumbs 2, Cell Polarity Complex Component	<ul style="list-style-type: none"> Cell polarity and cytoskeletal reorganisation is known to affect B-cell lymphoma migration and invasiveness Development of B-cell lymphoma has also been noted in Crb2-related syndrome (bi-allelic mutation of Crb2) 	Slavotinek, 2015; Gold et al., 2010	chr9 - 126135887 T -> C	non synonymous SNV

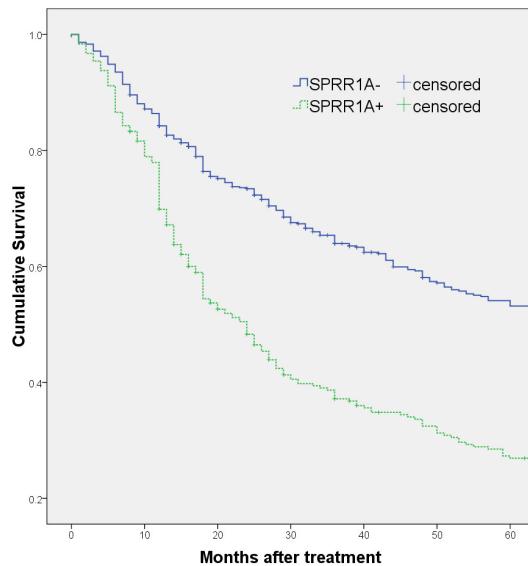


Figure 23: 5 Year Survival Curve of Patients with SPRR1A+ and SPRR1A- DL-BCL. Source : Zhang et al. (2014), Figure 2.

To aggregate the data from our Bayesian Ranking system, we did a Circos plot for the top 300 genes picked up by our gene ranking system (Figure 24). A Circos enables easy visualisation and analysis of large genome datasets, enabling quick understanding and comprehension of results.

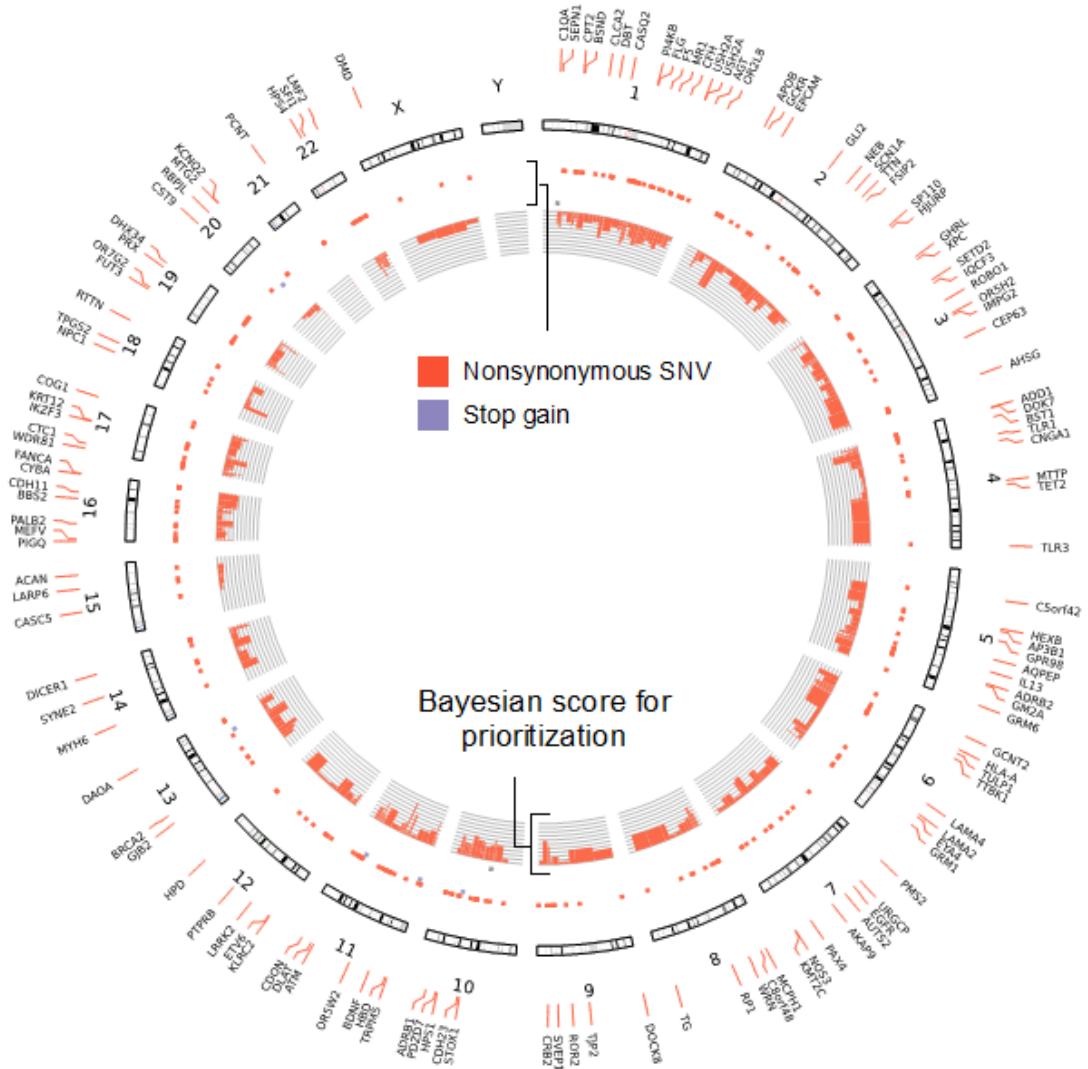


Figure 24: Circos Plot of Top 300 Ranked Genes from Bayesian Network Ranking. In this Circos plot, the outer track indicates the top ranked genes and their positions on the chromosome. The inner track describes the type of mutation that was observed – most mutations were non-synonymous SNVs, with a few stop-gain mutations. The innermost track shows the relative probabilities of each ranked gene.

From the Circos Plot (Figure 24), we find several interesting gene families that might also be relevant in B-Cell Lymphoma. These include several Toll-Like Receptors(TLRs), TLR3 (chr4,rank 26) and TLR1(chr4,rank 77) as well as interleukin receptors IL4R (chr16,rank 37) and IL1 β (chr2,rank 196). TLRs are of significant interest in cancer due to their involvement in the caspase pathway (Kelly et al., 2006), and have been implicated in B-Cell Lymphomas (Marron, Joyce, & Cunningham-Rundles,2012). Interleukins are also important in cancer due to their importance in mediating inflammation and immune response (Balkwill & Mantovani, 2001). Thus, we show that our Bayesian network can be used by clinicians to quickly interrogate the information from functional annotations and database lookups to report important genes.

4 Discussion

We demonstrate the validation of high-confidence variant calls using an optimised deep learning neural network on both real and simulated datasets, and we also show that a Bayesian network can rank and prioritise genes in a systematic way so as to obtain important genes. We show that four of the top five genes had published findings that linked them with B-Cell lymphoma. Looking at the top 300 genes ranked, we also found interesting families of genes that are known to be involved in Lymphoma progression, including the Toll-Like receptor and Interleukin receptors families. To benchmark these results, we compared our variant calling results with other methods like VariantMetaCaller and BAYSIC (Gézsi et al., 2015; Cantarel et al., 2014), and we also looked at other methods of gene prioritisation to see how our ranking system compares.

4.1 Comparison of Deep Learning with other Integration Methods

First, we compare deep learning networks with other method integrating variant call information, including VariantMetaCaller, which uses Support Vector Machines (a decision making machine learning technique), and BAYSIC, a method that uses a Bayesian probabilistic model to integrate variant call information. Both methods were also used to analyse and predict variants for NA12878 data (Gézsi et al., 2015). In terms of the Area Under Prediction Recall Curve (AUPRC) metric, VariantMetaCaller increased AUPRC scores by 0.04 for SNVs and 0.07 for indels when compared to their best single variant caller. BAYSIC also noted a 0.03 increase in AUPRC for SNP prediction and 0.05 increase in AUPRC for indel prediction compared to the best single variant caller. Numerically, this seems comparable to our results of an F1 score increase of 0.06 in both indel and SNP prediction for the NA dataset compared to the best single variant caller. However, since we used the F1 score metric, instead of the AUPRC metric, a direct quantitative comparison is not so simple.

While the AUPRC metric provides evidence of precision and recall improvements at all levels of threshold (Davis & Goadrich, 2006), it does not provide evidence for a predictors performance any specific threshold. The measurement of the F1 score is needed to capture this information, as it looks at the precision and recall at a specific threshold (Abeel, Van de Peer & Saeys, 2009). Instead, what they have shown is that looking at all thresholds, there is an overall increase in precision and recall, but it is still unclear what the improvements are at the optimised thresholds. Fawcett (2006) also mentions this problem, as he notes that 'It is

possible for a high-AUC classifier to perform worse in a specific region of ROC space than a low-AUC classifier'. Here, AUC refers to the Area Under the Receiver Operator Curve (Egan, 1976), which has been shown to be fundamentally the same as the AUPRC (Davis & Goadrich, 2006). Similar bioinformatics studies in predictive classifiers have also used the F1 score instead of the AUPRC as it is a more salient metric when the operating threshold is known (Abeel, Van de Peer & Saeys, 2009). Measurement of the F1 score is more relevant in clinical practice as we are mainly interested in the optimal operating conditions where precision and recall are maximised and all possible conditions. Our results provide direct evidence that at the optimal threshold for all types of callers, we can show a significant F1 score improvement.

Thus, one definite step moving forward is to incorporate VariantMetaCaller and BAYSIC into our pipelines as negative controls, and measure using the same dataset and same processes whether deep learning can outperform these two methods using the F1 score metric. Intuitively, we believe that deep learning will be able to edge out improvements as it can form complex representations of the data to learn from, which Support Vector Machines are unable to do. Furthermore, deep learning networks have been shown to outperform Support Vector Machines in prediction problems (LeCun et al., 2015; Schmidhuber, 2015), and evidence from our flat network architecture shows that putting all the features in a single vector and using that to performing machine learning (in a manner similar to the procedure of a Support Vector Machine) was not effective. Thus, a future research direction is to test this hypothesis by incorporating VariantMetaCaller and BAYSIC in the analytic pipelines.

However, one further complication in the approach of measuring each of the methods against the NA12878 dataset is that the high confidence calls provided in the NA12878 dataset is not the ground truth. Zook et al.(2014) themselves estimate a possible false negative or positive for every 30 million bases in the NA12878 dataset. This is due to variants that are not inside the high confidence dataset because of errors in one sequencing machine, or genomic regions that cause all sequencing machines to have similar biases and noise. Hence, this would result in misclassification and wrongly called false negative and false positive results, thus skewing the classification results. To solve this problem, a lot of effort has to be put in to obtain a set of verified truth variants via gold standard Sanger sequencing (Tsiatis et al., 2014), but this might be prohibitively expensive for a large number of mutations. Still, this would have to be done for us to have a good set of truth variables to test prediction software with before such software can be considered for use in actual treatment and diagnosis.

4.2 Analysis of Bayesian Network

For the Bayesian Network analysis, it is more difficult to numerically benchmark our results for gene prioritisation with current platforms. This is because currently used platforms are qualitative methodologies like gene panels (Olek and Berlin, 2002) or manual literature look-ups of disease-related genes, such as using the ClinVar database. While gene panels work well in a clinical setting, with NGS data it is hoped that as much information about a person's genome as possible can be used in treatment and diagnosis (Meldrum et al., 2011). Using a gene ranking system instead of just looking at a set of implicated genes might allow doctors to find out tease out possible homologs or interacting agents that might be related to the known deleterious genes (perhaps in the gene panel) and integrate that into their treatment and diagnosis.

One weakness of such a network however is its inability to discover new clinically relevant genes. It is biased against unknown but potentially clinically important genes, because it will underestimate their probability of importance. To solve this problem, entirely different approaches have to be taken, which include possible protein and mRNA characterisation methods to discover unknown effectors and drivers, and then subsequently searching for new genes of interest. However, our focus is building a tool for clinical use which requires well-characterised genes. Thus, this systematic under-representation of unknown genes has to be part of the design – the network should filter out uncharacterised genes.

A second weakness in this network is that all the information that goes into it has to be manually curated. Because it taps upon previously curated databases, it is limited in scope to the pre-processed information it already contains. One possible alternative to solve this problem is to develop a data miner to mine the literature for possible relationships between genes and diseases, thus creating a self-updating database. However, this is currently difficult to implement because a machine would have to learn the complex conceptual relationships between genes and phenotype, and this is currently an unsolved problem (Cambria & White, 2014). Furthermore, such a system would require still human intervention to verify the data being mined is valid and appropriate. Ultimately, this should still be a useful tool as it allows for an interface for clinicians and researchers to easily share and obtain information, and a single clinician to tap upon the expertise and data from many different research institutions without having to manually sieve through the literature.

4.3 Future Directions

One interesting extension we would like to move into in the future is to be able to integrate a druggable genome into the network, enabling the prioritisation of genes which have possible candidate drug targets. This method would look up a drug-gene interaction database, for example, DGldb (Griffith et al., 2013), and use the results to inform the importance of a gene. This would enable doctors to notice further possible drug candidates that would work very well on the gene profile of the patient that they might not have considered previously, thus increasing their scope of possible treatment options and augmenting their skills. Other directions in the future include being able to include extra variant callers which will provide it with even more feature data, enabling it to make better predictions. We also hope to build structural variant calling neural networks, as this is a current set of variants that our neural network does not take into account. Finally, we would also like to move everything onto a web interface such that it is accessible for use to perform variant validation and gene prioritisation. This would enable easy access to both the validation and prioritisation pipelines.

Thus, in this paper, we have shown the use of deep learning neural networks to validate variants in both real and simulated datasets successfully. We also show that using a Bayesian network can identify important genes within a lymphoma disease sample. Ultimately, we hope to be able to put these networks to use in a clinical setting to augment treatment and diagnosis of diseases.

5 Appendixes

5.1 Neural Network Learning

Machine learning with deep neural networks is underpinned by two key phases, the feed-forward phase and the backpropagation phase.

5.1.1 Feedforward Phase

The feedforward phase describes the computation of a prediction, and during this phase, the input features are used to compute the final output prediction. For a simple network below:

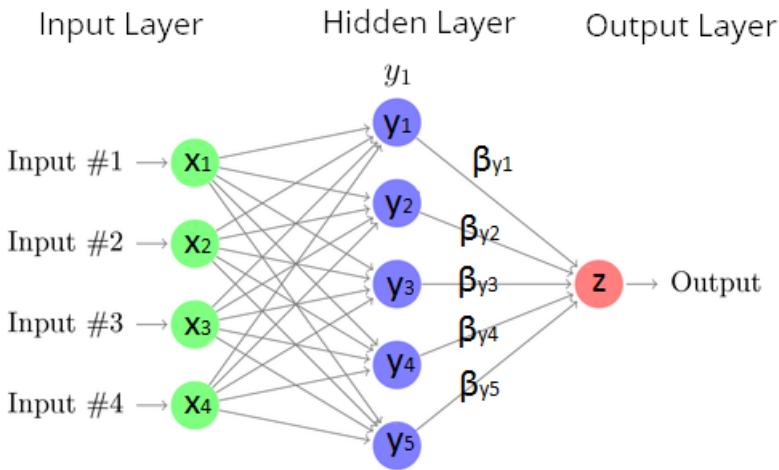


Figure 25: Sample Neural Networks with Labelled Nodes and Weights.

The final prediction, z is computed with the equation:

$$z = \beta_{y1} * y_1 + \beta_{y2} * y_2 + \beta_{y3} * y_3 + \beta_{y4} * y_4 + \beta_{y5} * y_5 \quad (4)$$

Where β indicates, the weights linking each output to the input of z and each of the y_i terms are computed in the same manner from the x_i layer. At each node (x,y,z) , there is also the existence of an activation function that modifies the input of the node to compute an output. Commonly used activation functions include the rectified linear unit (ReLU), sigmoid functions like hyperbolic tangent and logistic function, $S(T) = \frac{1}{1+e^{-T}}$. Thus, the final prediction can be seen as a summation of all weights multiplied by the activation output of each node. In theory, we can expand each of the y_i terms in equation (2) to include the y_i layer activation function as well as rewrite the y_i layer inputs in terms of the sum of outputs and weights from the x_i layers. This complex integration of terms allows for the neural network to form complex

continuous decision boundaries as the neural networks can compute sophisticated non-linear prediction functions despite being a fundamentally linear model.

5.1.2 Backpropagation Phase

After a prediction is made, we then have to check whether it is correct and change our weights if an erroneous prediction was made (Figure 26).

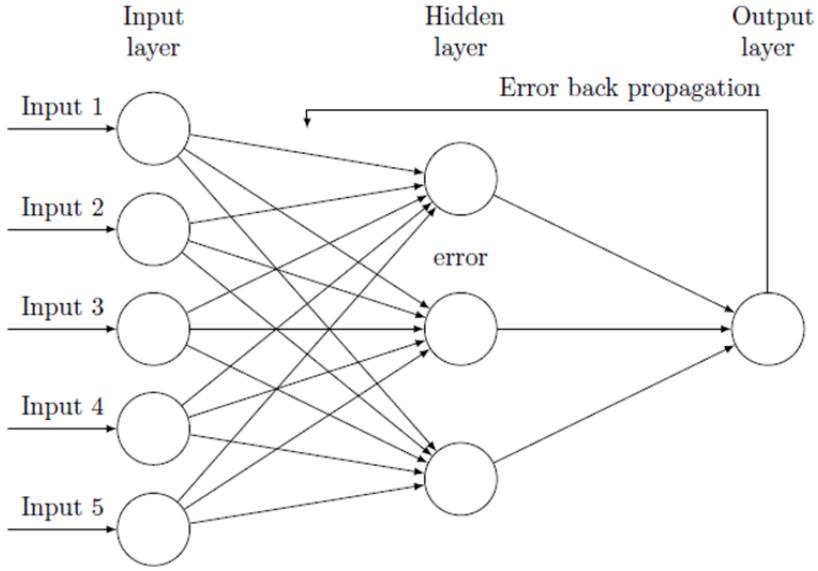


Figure 26: Backpropagation of Error Terms

This is the backpropagation step, which involves backpropagating the error terms from the output layers to the input layer and updating the weights at each node based on the differential relationship between the error and each specific gradient. Specifically, this is governed by the optimiser functions which have been mentioned earlier – one example of such a function is the Stochastic Gradient Descent function, which is

$$\beta_{yi}^n = \beta_{yi}^{n-1} - \alpha \frac{\partial E_n(\beta)}{\partial \beta_i} \quad (5)$$

Here, each β term indicates a gradient, α is a constant for the learning rate and $\frac{\partial E_n(\beta)}{\partial \beta_i}$ is the term used to modify the weight of the gradient based on the cost function $E_n(\beta)$. The idea used in all backpropagation functions is gradient descent, where the contribution of the gradient term to the error is computed, and the gradient is changed by an amount in order to reduce the future contribution of the gradient to that error.

5.1.3 Cost Function in Gradient Descent

Here it is useful to consider what the cost function $E_n(\beta)$ is. It is essentially the error rate when a set of gradients is used to perform predictions, as it measures how many accurate predictions were made and how many wrong predictions were made. For a binary class predictor (which is what we are using, only true and false), this is given by the equation

$$E(\beta) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^2 y_{ij} \log(p_{ij}) \quad (6)$$

where y_{ij} indicates the empirically observed probabilities of each class label while $\log(p_{ij})$ is the theoretical probabilities of each class label. This is also known as binary cross-entropy, which is derived from Shannon's entropy (See Appendix 5.2.1). From this term, we see that if the neural network predicts something with a high probability (y_{ij} is high) and it is false (p_{ij} is low) so then $\log(p_{ij})$ is a big negative number, and so the cost function will very high. On the other hand, if y_{ij} and p_{ij} is high then the entropy will be close to zero, indicating a correct prediction. Since each of the prediction terms can be rewritten in terms of the gradient (rewrite z in terms βy_i and so on), we can theoretically compute the contribution of each gradient to the cost function to see how the cost function changes as the gradient changes. Thus, this is what gradient descent does – it tries to see how the cost function changes as each gradient changes, then attempts to move the gradient in the direction that minimises the error term.

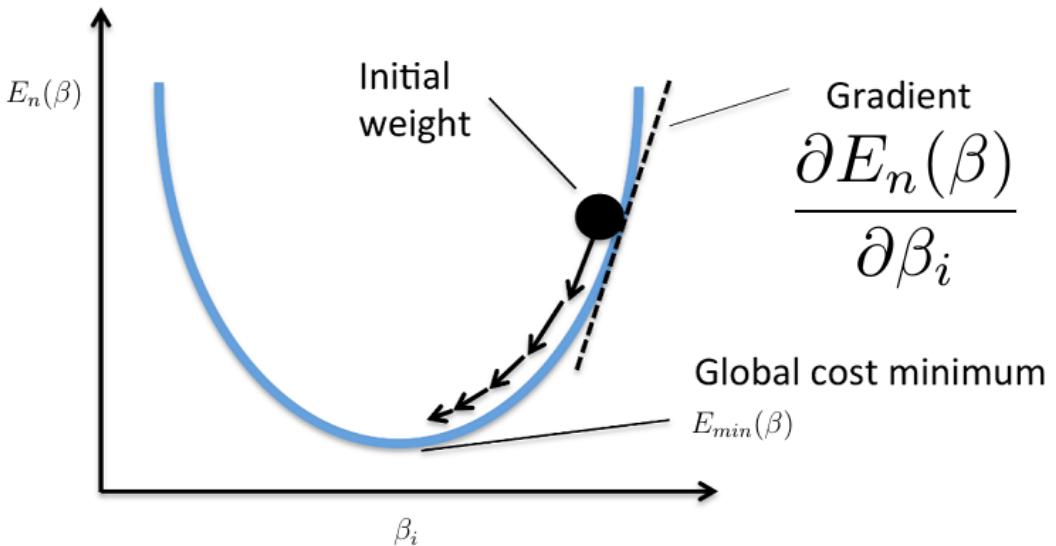


Figure 27: **Graphical Illustration of Gradient Descent.** Gradient descent attempts to find the gradient at which the cost function is minimised (since the cost function depends on the gradient).

This is best seen in Figure 27 above, where the gradient or specifically the partial differentiation of the cost function with regards to each gradient is used to move the gradient to a new position so as to minimise the error term. Thus, machine learning is, in essence, a minimisation problem – we want to find a set of weights that minimises the cost function. Because the cost function describes how many predictions we made correctly, gradient descent essentially trains our network to accurately predict outputs from inputs.

5.2 Feature Engineering

We subset our features into three broad sets, which are base-specific information, sequencing error and bias information features, and calling and mapping quality. Base information tells us base specific properties, including information contained in the base as well as the quality of sequenced bases in the samples. Sequencing error and bias features attempt to tease out potential biases in sequencing, including features such as GC content, longest homopolymer run and as well as allele balances and counts. Finally, calling and mapping quality provides information on the mapping and calling confidence of the variant callers, and includes features such as genotype confidence and mapping quality. In all, these sets of information provide information on the key aspects of variant calling – specifically the properties of the bases in the samples, the characteristics of the sequencing process and finally the variant calling and mapping algorithms.

5.2.1 Base Information

Shannon Entropy

Shannon Entropy captures the amount of information contained in the allele sequences. It is calculated using the equation:

$$H(X) = - \sum_{i=1}^n P(x_i) \log_2 P(x_i) \quad (7)$$

where $P(x_i)$ is the probability of finding each base at each position. Thus, we calculate the entropy by summing up the probabilities/ $\log(\text{probabilities})$ at each position. This prior probability is calculated in two ways, and both are used as features – firstly, the overall genome base probabilities are calculated over the entire genome, and thus the entropy is related to the probability of finding a base at any position in the genome. The second way prior probability is calculated is to take a region of space around the allele (10 bases plus the length of the allele in our calculations) and use those probabilities to calculate the entropy of the allelic sequence. Intuitively, it attempts to find out the amount of information contained within the allelic sequence, and hopefully, the neural network can use the information to determine the validity of a mutation.

Kullback Leibler Divergence

The Kullback-Leibler Divergence feature is similar to Shannon entropy, but instead, we use this to measure the informational change converting from the reference to the allele sequence.

The Kullback-Leibler Divergence is calculated as follows:

$$D_{KL}(P||Q) = - \sum_{i=1}^n P(x_i) \log_2 \frac{P(x_i)}{Q(x_i)} \quad (8)$$

where $Q(x_i)$ is the prior probability of finding each base at each position based on the genomic region around the allele, while $P(X_i)$ is the posterior probability of finding a specific base inside the allelic sequence. Thus, the KL divergence describes the informational gain when the probabilities from Q are used to describe P. Intuitively, since we know the base probabilities of the region, we can then study the probabilities observed in the reference allelic sequence and see how well $Q(X_i)$ probabilities can approximate $P(X_i)$ probabilities.

Base Quality

Base quality refers to the Phred score probability that the called allele is wrong. It is given by the equation:

$$P = 10^{-\frac{Q}{10}}$$

Where P is the Base Quality, and Q is the probability that the allele called is wrong. This is a number computed by the sequencing machine based on the quality of the base samples provided, and tells us how much confidence the sequencing machine has in calling that base.

5.2.2 Sequencing Biases and Errors

GC content

This feature computes the calculated GC content of reference genome, which may affect sequencing results and accuracy as regions with a GC content are known to be more difficult to sequence. This is because of the greater strength of GC bonds, resulting in errors and biases in sequencing (Benjamini & Speed, 2012).

Longest homozygous run

Homopolymer runs (AAAAAAA) are known to cause sequencer errors (Quail et al.,2012),

and might be a factor in determining whether a variant is true. This because long homopolymers provide the same type of signal to the sequencing machine, resulting in a difficult in estimating the magnitude of the signal or rather how many bases are in that homopolymer, resulting in errors and wrongly called variants. The reference sequence region including the allele was checked for homopolymer runs.

Allele Count

Allele count gives the total number of alleles in called phenotypes. This feature provides information of possible biases in the sequencing machine.

Allele Balance

Allele Balance describes the ratio of final allele called over all other alleles called(reference allele for heterozygous calls, or other alleles for homozygous calls). This feature give us information of possible biases in the sequencing machine.

Allele Frequency

Allele Frequency describes the ratio of final allele called over all other alleles called(reference allele for heterozygous calls, or other alleles for homozygous calls). This feature give us information of possible biases in the sequencing machine.

5.2.3 Calling and Mapping Qualities

Genotype Likelihood

The genotype likelihood provides the Phred-scaled likelihood scores of how confident the caller is in determining that it is a homozygous or heterozygous call, and for the homozygous calls whether it is a more likely to be a bi-allelic mutation or no mutation at all. This feature thus gives us the confidence of the caller in determining if one or two alleles have mutated.

Read Depth

Mapped read depth refers to the number of bases sequenced and aligned at a given reference base position. The read depth tells us how many reads contributed to a specific call, and thus provides information on how much evidence there is for the variant call

Read Depth, 4 Variables

This feature is a refinement of Read Depth, and provides 4 variables for the number of forward reference alleles, the number of reverse reference alleles, the number of forward non-ref alleles and finally the number of reverse non-reference alleles.

Allele Depth

This feature is a refinement of Read Depth, and describes the number of alleles supporting the variant call.

Quality by Depth

Quality by Depth is computed by dividing the quality score against allele depth, to obtain an average score of allele quality. This composite feature provides information on the information provided by each read supporting the call

Mapping Quality

Mapping quality is originally a score provided by the alignment method and gives the probability that a read is placed accurately. The variant callers compute an overall mapping quality of the reads that provide evidence for a variant call which is given in this feature. A low mapping quality means that there are multiple positions where the reads contributing to this variant call could have gone, and thus providing evidence that this might not be an accurate call due to poor mapping.

Mapping Quality of Observed Reference Alleles

This feature is a refinement of Mapping Quality, and describes the mapping quality of observed reference alleles.

Mapping Quality of Observed Alternate Alleles

This feature is a refinement of Mapping Quality, and describes the mapping quality of observed alternate alleles.

5.3 Mathematical and Statistical Tools

5.3.1 Derivation of F1 Score

The F1 score is a useful measure as it can measure both the precision as well as the recall of a predictor. For a binary predictor with a binary truth class(Figure 28), we can obtain four types of results – true positives, true negatives, false positives and false negatives.

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive	False Negative
	No	False Positive	True Negative

Figure 28: Confusion Matrix for a Binary Class Predictor.

True positives are positive predictions that are made that are positive class labels, while false positives are positive predictions that are made that have negative class labels. Similarly, true negatives are negative predictions that have negative class labels, while false negatives are negative predictions that are positive class labels. From this, we can define two equations, precision and recall. Precision is defined as (8) while recall is defined as (9).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (9)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (10)$$

Precision tells us how likely a positive prediction made will be true, while recall tells us how much of the truth class positive predictions the predictor can classify successfully . Thus, a predictor can have a high precision but low recall (makes few predictions but are very accurate) or a high recall and low precision(makes many predictions that capture all truth variables, but have a lot of false positives as well). In genomics, both types of errors are not desired – we would want all the predictions to be true (precision), while not losing out on any important mutations (recall). Thus, we use the composite metric, the F1 score, that looks at the overall

precision and recall of a predictor. It is defined as follows:

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (11)$$

5.3.2 Principal Components Analysis (PCA)

Principal Components Analysis (PCA) is a commonly used tool for dimensionality reduction. It was first proposed by Pearson in 1901 (Pearson, 1901) and has been commonplace in many data analytics and signal processing methodologies (Jolliffe, 2002). PCA works by attempting to discover orthogonal principal components (PCs) that are able to represent the original data. Specifically, this means that the PCs can capture variance in the datasets. This is done by finding the Eigenvalues and Eigenvectors of the dataset, with the eigenvectors representing a linear combination of all input variables and the eigenvalues representing the amount of variance that that eigenvector can represent. Ultimately, we select n eigenvectors that can represent a percentage of variance in our dataset. Because each eigenvector is orthogonal, they can capture the variance in the dataset.

For our analysis, we decided to use eight principal components – we took the limit as the last principal component that was able to represent at least 0.5% of the variance in the dataset (Figure 29).

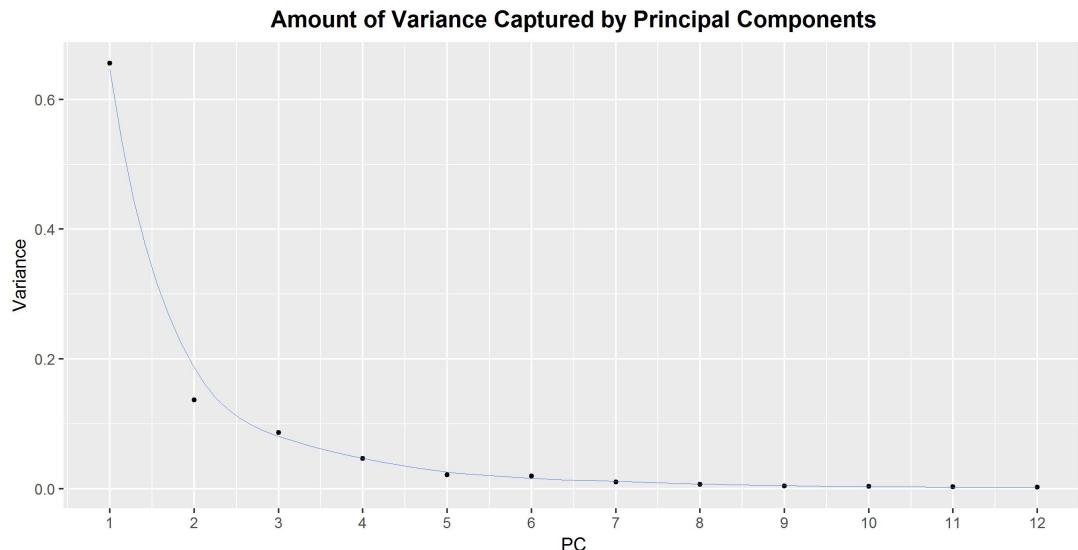


Figure 29: Variance Captured by First 12 Principal Components

To carry out PCA, we used the preprocessing step SciPy to normalise all the input vectors to mean 0 and standard deviation 1. Subsequently, we perform principal components decomposition to obtain the eigenvector transformed representation of the dataset and their corresponding eigenvalues. We then fit 8 of the principal components that explained the largest amount of variance into the neural network to study if it can learn from the compressed representation of the input features.

5.3.3 Synthetic Minority Overrepresentation Technique (SMOTE)

SMOTE is a statistical technique described in by Chawla et al. (2002) to overcome problems with imbalanced datasets that are common in machine learning. SMOTE oversamples the training class with fewer variables in a way that tries not to replicate data points (that makes certain data points over-represented) without creating new invalid training examples. It does this by taking the intersection of two nearest data points of the same training class. This can be seen in Figure 30.

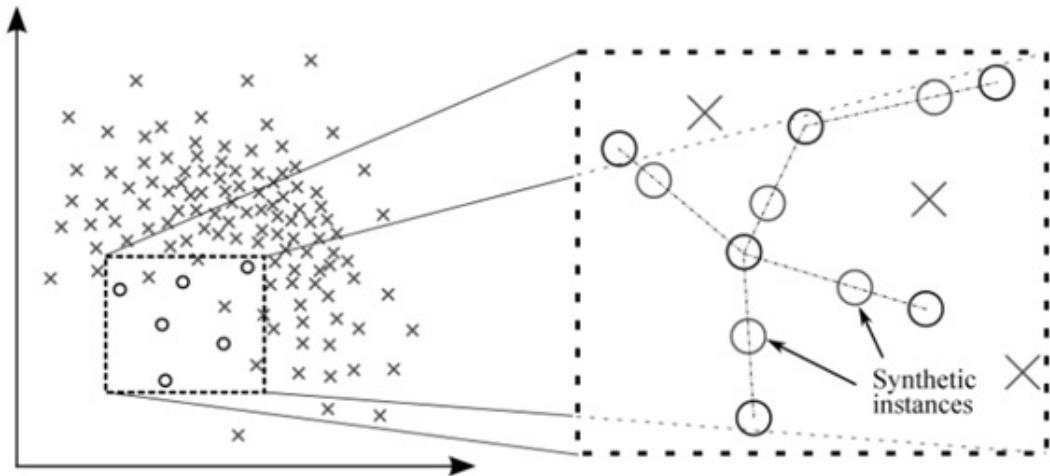


Figure 30: **Illustration of SMOTE Oversampling Algorithm.** Note the use of 2 nearest neighbours to create a new synthetic example. Figure from Chawla et al., 2002

In doing so, it creates a more generalised representation of the sample class with less training examples, without replicating certain datapoints and without creating invalid data. This enables intelligent oversampling of the dataset to balance out the positive and negative feature classes. SMOTE has been shown to be valid for other datasets including sentence boundary detection (Liu et al., 2006) and data mining (Chawla, 2005).

6 Bibilography

Amagai, M. (2004). A mystery of AHNAK/desmoyokin still goes on. *Journal of investigative dermatology* 123, xiv.

Abyzov, A., Li, S., Kim, D., Mohiyuddin, M., Stütz, A., Parrish, N., Mu, X., Clark, W., Chen, K., and Hurles, M. et al. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications* 6, 7256.

Alizadeh, A., Eisen, M., Davis, E., Ma, C., Lossos, I., Rosenwald, A., Boldrick, J., Sabet, H., Tran, T., and Yu, X. (2017). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 403, 503-511.

Angrist, M. (2016). Personal genomics: Where are we now?. *Applied & translational genomics* 8, 1.

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., & Mougiaakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging* 35, 1207-1216.

Balkwill, F., & Mantovani, A. (2001). Inflammation and cancer: back to Virchow?. *The Lancet* 357, 539-545.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 157-166.

Benjamini, Y., & Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research* 40.

Cantarel, B. L., Weaver, D., McNeill, N., Zhang, J., Mackey, A. J., & Reese, J. (2014). BAYSIC: a Bayesian method for combining sets of genome variants with improved

specificity and sensitivity. BMC bioinformatics 15, 104.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. In Data mining and knowledge discovery handbook (pp. 853-867). Springer US.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321-357.

Chen, Y., Lin, Z., Zhao, X., Wang, G., & Gu, Y. (2014). Deep learning-based classification of hyperspectral data. IEEE Journal of Selected topics in applied earth observations and remote sensing 7, 2094-2107.

CLC bio. (2011). Genomics Workbench User Manual 4.8. Finlandsgade 10-12 DK- 8200 Aarhus N, Denmark

Cornish, A., and Guda, C. (2015). A Comparison of Variant Calling Pipelines Using Genome in a Bottle as a Reference. Biomed Research International 2015, 1-11.

Danecek, P., Auton, A., Abecasis, G., Albers, C., Banks, E., DePristo, M., Handsaker, R., Lunter, G., Marth, G., and Sherry, S. et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156-2158.

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M., and Hanna, M. et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics 43, 491-498.

Di Tommaso, P., Chatzou, M., Baraja, P. P., & Notredame, C. (2014). A novel tool for highly scalable computational pipelines.

- Egan, J. P. (1975). Signal detection theory and ROC analysis.
- Emara, M., Turner, A. R., & Allalunis-Turner, J. (2014). Adult, embryonic and fetal hemoglobin are expressed in human glioblastoma cells. International Journal of Oncology 44, 514-520.
- Escalona, M., Rocha, S., & Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. Nature Reviews Genetics 17, 459-469.
- Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters 27, 861-874.
- Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907.
- Gézsi, A., Bolgár, B., Marx, P., Sarkozy, P., Szalai, C., & Antal, P. (2015). VariantMetaCaller: automated fusion of variant calling pipelines for quantitative, precision-based filtering. BMC genomics 16, 1.
- Griffith, M., Griffith, O., Coffman, A., Weible, J., McMichael, J., Spies, N., Koval, J., Das, I., Callaway, M., and Eldred, J. et al. (2013). DGIdb: mining the druggable genome. Nature Methods 10, 1209-1210.
- Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Rajpurkar, P., Migimatsu, T., and Cheng-Yue, R. et al. (2017). An Empirical Evaluation of Deep Learning on Highway Driving. Arxiv 1504.01716.
- Hwang, S., Kim, E., Lee, I., & Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. Scientific reports 5, 17875.
- Janitz, M. (Ed.). (2011). Next-generation genome sequencing: towards personalized medicine. John Wiley & Sons.

Jensen, F. V. (1996). An introduction to Bayesian networks (Vol. 210). London: UCL press.

Johnson, S., Abal, E., Ahern, K., & Hamilton, G. (2014). From science to management: using Bayesian networks to learn about Lyngbya. *Statistical Science* 29, 36-41.

Jolliffe, I. (2002). Principal component analysis. John Wiley & Sons, Ltd.

Karolchik, D., Barber, G., Casper, J., Clawson, H., Cline, M., Diekhans, M., Dreszer, T., Fujita, P., Guruvadoo, L., and Haussler, M. et al. (2017). The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42, D764-D770.

Kelly, M., Alvero, A., Chen, R., Silasi, D., Abrahams, V., Chan, S., Visintin, I., Rutherford, T., and Mor, G. (2006). TLR-4 Signaling Promotes Tumor Growth and Paclitaxel Chemoresistance in Ovarian Cancer. *Cancer Research* 66, 3859-3868.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. *Nature Reviews Cancer* 1, 157-162.

Lam, H., Clark, M., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F., Habegger, L., Ashley, E., and Gerstein, M. et al. (2012). Performance comparison of whole-genome sequencing platforms. *Nature Biotechnology* 30, 562-562.

Lamont, R., Tan, W., Innes, A., Parboosingh, J., Schneidman-Duhovny, D., Rajkovic, A., Pappas, J., Altschwager, P., DeWard, S., and Fulton, A. et al. (2016). Expansion of phenotype and genotypic data in CRB2-related syndrome. *European Journal Of Human Genetics* 24, 1436-1444.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Ma-giott, D. R. (2014). ClinVar: public archive of relationships among sequence variation

and human phenotype. Nucleic acids research 42, D980-D985.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature 521, 436-444.

Lee, I., Sohn, M., Lim, H., Yoon, S., Oh, H., Shin, S., Shin, J., Oh, S., Kim, J., and Lee, D. et al. (2014). Ahnak functions as a tumor suppressor via modulation of TGF β /Smad signaling pathway. Oncogene 33, 4675-4684.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079.

Linderman, M., Brandt, T., Edelmann, L., Jabado, O., Kasai, Y., Kornreich, R., Mahajan, M., Shah, H., Kasarskis, A., and Schadt, E. (2014). Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Medical Genomics 7.

Liu, X., Han, S., Wang, Z., Gelernter, J., & Yang, B. Z. (2013). Variant callers for next-generation sequencing data: a comparison study. PloS one, 8(9), e75619.

Liu, Y., Stolcke, A., Shribberg, E., & Harper, M. (2005, June). Using conditional random fields for sentence boundary detection in speech. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (pp. 451-458). Association for Computational Linguistics.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. Information Sciences 250, 113-141.

Lusci, A., Pollastri, G., & Baldi, P. (2013). Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. Journal of chemical information and modeling 53, 1563.

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013, June). Rectifier nonlinearities improve neural network acoustic models. In Proc. ICML (Vol. 30, No. 1).

Marron, T. U., Joyce, E. Y., & Cunningham-Rundles, C. (2012). Toll-like receptor function in primary B cell defects. *Frontiers in bioscience* (Elite edition), 4, 1853.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., and Daly, M. et al. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297-1303.

Meldrum, C., Doyle, M. A., & Tothill, R. W. (2011). Next-generation sequencing for cancer diagnostics: a practical perspective. *Clin Biochem Rev*, 32(4), 177-195.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nature reviews genetics* 11, 31-46.

Mohiyuddin, M., Mu, J., Li, J., Bani Asadi, N., Gerstein, M., Abyzov, A., Wong, W., and Lam, H. (2015). MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics* 31, 2741-2744.

Moreau, Y., & Tranchevent, L. C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nature Reviews Genetics* 13, 523-536.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443-451.

O'Rawe, J., Jiang, T., Sun, G., Wu, Y., Wang, W., Hu, J., Bodily, P., Tian, L., Hakonarson, H., and Johnson, W. et al. (2013). Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* 5, 28.

Pearson, K. (1901). Principal components analysis. The London, Edinburgh and Dublin Philosophical Magazine and Journal 6, 566.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Weiss, R., Dubourg, V., and Vanderplas, J. et al. (2017). Scikit-learn: Machine learning in Python. Journal Of Machine Learning Research 12, 2825-2830.

Pourret, O., Naïm, P., & Marcot, B. (Eds.). (2008). Bayesian networks: a practical guide to applications (Vol. 73). John Wiley & Sons.

Quail, M., Smith, M., Coupland, P., Otto, T., Harris, S., Connor, T., Bertoni, A., Swerdlow, H., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13, 341.

Rehm, H. L. (2017). Evolving health care through personal genomics. Nature Reviews Genetics.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.

Sandmann, S., de Graaf, A. O., Karimi, M., van der Reijden, B. A., Hellström-Lindberg, E., Jansen, J. H., & Dugas, M. (2017). Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. Scientific Reports 7.

Satterwhite, E., Sonoki, T., Willis, T., Harder, L., Nowak, R., Arriola, E., Liu, H., Price, H., Gesk, S., and Steinemann, D. et al. (2001). The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. Blood 98, 3413-3420.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., & Quince, C. (2016). Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC bioinformatics 17, 125.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks* 61, 85-117.

Schork, N. J., Murray, S. S., Frazer, K. A., & Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development* 19, 212-219.

Shtivelman, E., Cohen, F. E., & Bishop, J. M. (1992). A human gene (AHNAK) encoding an unusually large protein with a 1.2-microns polyionic rod structure. *Proceedings of the National Academy of Sciences* 89, 5472-5476.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, J., Panneershelvam, V., and Lanctot, M. et al. (2017). Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484-489.

Slavotinek, A. M. (2016). The Family of Crumbs Genes and Human Disease. *Molecular Syndromology* 7, 274-281.

Spencer, D., Abel, H., Lockwood, C., Payton, J., Szankasi, P., Kelley, T., Kulkarni, S., Pfeifer, J., and Duncavage, E. (2013). Detection of FLT3 Internal Tandem Duplication in Targeted, Short-Read-Length, Next-Generation Sequencing Data. *The Journal Of Molecular Diagnostics* 15, 81-93.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15, 1929-1958.

Su, C., Borsuk, M. E., Andrew, A., & Karagas, M. (2014, May). Incorporating prior expert knowledge in learning Bayesian networks from genetic epidemiological data. In *Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on* (pp. 1-5). IEEE.

Sun, Y., Wang, X., & Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1891-1898).

Sutskever, I., Martens, J., Dahl, G. E., & Hinton, G. E. (2013). On the importance of initialization and momentum in deep learning. ICML 28, 1139-1147.

Talwalkar, A., Liptrap, J., Newcomb, J., Hartl, C., Terhorst, J., Curtis, K., Bresler, M., Song, Y., Jordan, M., and Patterson, D. (2014). SMASH: a benchmarking toolkit for human genome variant calling. Bioinformatics 30, 2787-2795.

Tentler, J., Tan, A., Weekes, C., Jimeno, A., Leong, S., Pitts, T., Arcaroli, J., Messersmith, W., and Eckhardt, S. (2012). Patient-derived tumour xenografts as models for oncology drug development. Nature Reviews Clinical Oncology 9, 338-350.

Tieleman, T. and Hinton, G. Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning. Technical report, 2012

Tsiatis, A. C., Norris-Kirby, A., Rich, R. G., Hafez, M. J., Gocke, C. D., Eshleman, J. R., & Murphy, K. M. (2010). Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. The Journal of Molecular Diagnostics 12, 425-432.

Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. J Mach Learn Res, 10, 66-71.

Van Rossum, G. (2007, June). Python Programming Language. In USENIX Annual Technical Conference (Vol. 41, p. 36).

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38, e164-e164.

Wei, Z., Wang, W., Hu, P., Lyon, G. J., & Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic acids research* 39, e132-e132.

Weniger, M., Pulford, K., Gesk, S., Ehrlich, S., Banham, A., Lyne, L., Martin-Subero, J., Siebert, R., Dyer, M., and Möller, P. et al. (2006). Gains of the proto-oncogene BCL11A and nuclear accumulation of BCL11AXL protein are frequent in primary mediastinal B-cell lymphoma. *Leukemia* 20, 1880-1882.

Windecker, S., Stortecky, S., Stefanini, G., da Costa, B., Rutjes, A., Di Nisio, M., Silletta, M., Maione, A., Alfonso, F., and Clemmensen, P. et al. (2014). Revascularisation versus medical treatment in patients with stable coronary artery disease: network meta-analysis. *BMJ* 348.

Xie, M., Lu, C., Wang, J., McLellan, M., Johnson, K., Wendl, M., McMichael, J., Schmidt, H., Yellapantula, V., and Miller, C. et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature Medicine* 20, 1472-1478.

Yan, Y., Chen, M., Shyu, M. L., & Chen, S. C. (2015, December). Deep learning for imbalanced multimedia data classification. In *Multimedia (ISM), 2015 IEEE International Symposium on* (pp. 483-488).

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865-2871.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhang, H., Gao, J., Zhao, Z., Li, M., & Liu, C. (2014). Clinical implications of SPRR1A expression in diffuse large B-cell lymphomas: a prospective, observational study. *BMC cancer* 14, 333.

Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., & Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology* 32, 246-251.

7 Relevant Code

3 code segments are provided to clarify the implementation of generating the matrixes for deep learning, deep learning networks and finally bayesian networks. Other code segments not shown include the code base for parsing vcf input into features, concordance generators, NextFlow and Bash code used to simulate and process genomic data as well as to control deep learning and analytic pipelines, and other python helper scripts (e.g. comparing two VCF files, analysis with pre-trained network).

7.1 generate_matrixes.py

```
1 #This python script generates the set of matrixes to be used in deep learning, and then calls the main
2   ↵ method that trains the deep learning network.
3
4 #Input : a directory that contains all the vcf files for processing, as well as a truth file.
5
6 #Output : np.arrays of features from generate_matrixes with accompanying truth labels, feature set
7   ↵ lengths, a dictionary of vcf object records, as well as the list of relevant sample features for
8   ↵ easy reference
9
10 #Notes :
11
12 #Vcf files should have the "vcf" string in their name and truth file should have a "truth" string in its
13   ↵ name.
14
15 #No other file should be present in the folder
16
17 #Overall Strategy :
18
19 #Generate a dictionary of lists, where the keys are mutations, and the value is contains a matrix
20   ↵ containing information of all five callers
21
22 #Secondly, for each mutation label, check if it is inside the truth file or not. The truth is preloaded
23   ↵ into a dictionary
24
25 #Finally, pass the set of features with accompanying truth labels to the neural network
26
27 #The main datastructure used are python dictionaries, which allows O(1) dictionary lookup times
28
29
30 import os
31 import time
32
33 from ANNgeneratorresults import * #this file contains all the main methods for actual neural network
34   ↵ training
35
36 from methods import * #this file contains all the methods for parsing each VCF entry into a numerical
37   ↵ list of features
38
39
40 #declare names of useful files that contains processed data to be saved
41 LIST_OF_INPUTS_NAME = '/ANN/samplelist.p'
42 TRUTH_DICTIONARY_NAME = '/ANN/truthdict.p'
43 CALLER_LENGTH_FILE_NAME = '/ANN/callerlengths.txt'
44 VCF_LIST_FILE_NAME = '/ANN/vcf_list.p'
45 SCORES_NAME = '/ANN/scores.txt'
46 Y_DATA_NAME = '/ANN/myydata.txt'
47 X_DATA_NAME = '/ANN/myXdata.txt'
48
49
50 #Initialise NUMBER_OF_CALLERS
```

```

28     NUMBER_OF_CALLERS = 5
29
30
31     # This method follows the typical input output processing pipeline
32     # It takes in the user input, and loads it into local variables.
33     # It then executes another method, main_analyse_samples_and_truth on the loaded variables
34     # Finally, it then saves files into a directory determined by the final variables, and calls the next
35     # step of the pipeline
36
37     def load_and_save_data(user_input):
38
39         user_input = vars(user_input)
40
41         input_samples, referencepath, output_location = load_references(user_input)    # load user input
42
43         my_x_dataset, my_y_dataset, list_of_samples, truth_dictionary, length_of_caller_outputs, \
44         vcf_record_list = main_analyse_samples_and_truth(input_samples, referencepath)
45
46         save_files(output_location, my_x_dataset, length_of_caller_outputs,
47
48             list_of_samples, truth_dictionary, vcf_record_list, my_y_dataset)
49
50         orig_stdout = sys.stdout  #save print statements into stdout
51
52         f = file(str(output_location) + SCORES_NAME, 'w')
53
54         sys.stdout = f
55
56         main_gather_input_execute_prep_output(length_of_caller_outputs, truth_dictionary, my_x_dataset,
57
58             my_y_dataset, list_of_samples, output_location, vcf_record_list)
59
60
61     # This method first prepares a dictionary of truth to be checked against. It then initialises
62     # a dictionary of samples with all the keys, each key being a variant call, and then fills it up each
63     # key with data from each caller
64
65     # subsequently, it removes dictionary entries that are the wrong size, and then checks whether
66     # each entry in the dictionary is true or not by looking up the truth dictionary
67
68     # subsequently it performs array balancing, and converts the data to np.array, as well as the dictionary
69     # of truth
70
71     # and list of called samples
72
73
74     def main_analyse_samples_and_truth(path, referencepath):
75
76         os.chdir(path)
77
78         truthdict = generate_truth_list(path)
79
80         print "truth dictionary generated at time :", time.time() - start
81
82         callerlengths, list_of_called_samples, vcf_list = generate_input(path, referencepath)
83
84         print "samples generated at time :", time.time() - start
85
86         clean_truth_array, cleaned_sample_array = check_predicted_with_truth(list_of_called_samples,
87
88             truthdict)
89
90         print "samples checked with truth at time :", time.time() - start
91
92         cleaned_sample_array = np.array(cleaned_sample_array, np.float64)
93
94         clean_truth_array = np.array(clean_truth_array)

```

```

66     return cleaned_sample_array, clean_truth_array, list_of_called_samples, truthdict, callerlengths,
67     ↵ vcf_list
68
69 # This method generates the truth dictionary, by iterating through the vcf file, parsing all the vcf
70 # entries and appending them all as keys in the dictionary
71
72 def create_truth_dictionary(generated_truth_dictionary, truth_file):
73     vcf_reader = vcf.Reader(open(truth_file, 'r'))
74     for record in vcf_reader:
75         if "GL" in record.CHROM:      #Ignore non-regular chromosomes in our dataset
76             continue
77         templist = []
78         for item in record.ALT:
79             templist.append(str(item).upper())           #Alternates might be a list, so they have to be
80             ↵ saved as a immutable tuple
81         generated_truth_dictionary[(str(record.CHROM), str(record.POS), str(record.REF).upper())] =
82             ↵ tuple(templist)
83
84 # This method generates the input dictionary, by first initialising the keys of the dictionary by
85 # iterating through the vcf file once, and then
86
87 # Iterating through the vcf file again and parsing all the entries as input vectors
88
89 def generate_input(path, referencepath):
90     reference_dictionary = get_reference_dictionary_for_entropy(referencepath)
91     base_entropy = get_ref_entropy(referencepath)
92     full_dictionary = get_dictionary_keys(path)
93     list_of_called_samples, callerlengths, vcf_list = fill_sample_dictionary(base_entropy,
94     ↵ full_dictionary, path, reference_dictionary)
95     return callerlengths, list_of_called_samples, vcf_list
96
97
98 # This method goes through all the training variant calling files and extracts unique calls as keys in
99 # the sample dictionary
100
101 def get_dictionary_keys(path):
102     sample_dictionary = {}
103     for vcf_file in os.listdir(path):
104         if ignore_file(vcf_file):
105             continue
106         vcf_reader = vcf.Reader(open(vcf_file, 'r'))
107         sample_dictionary = create_dictionary_keys(vcf_reader, sample_dictionary)
108     return sample_dictionary

```

```

102  #This method ensures the feature vector is in the right order - the entries must always be in the order
103  #→ fb, hc, ug, pindel and st.
104
105  def create_list_of_paths(path):
106
107      list_of_paths = [0] * NUMBER_OF_CALLERS
108
109      for vcf_file in os.listdir(path):
110
111          if ignore_file(vcf_file):
112              continue
113
114          if "fb" in vcf_file:
115              list_of_paths[0] = vcf_file
116
117          if "hc" in vcf_file:
118              list_of_paths[1] = vcf_file
119
120          if "ug" in vcf_file:
121              list_of_paths[2] = vcf_file
122
123          if "pind" in vcf_file:
124              list_of_paths[3] = vcf_file
125
126          if "st" in vcf_file:
127              list_of_paths[4] = vcf_file
128
129      return list_of_paths
130
131
132  # This method goes through all the training variant calling files and fills each entry in a sample
133  #→ dictionary
134
135  # with data. If it is empty, it returns an array of length n, where n is the number of variables
136  # that same caller would have provided.
137
138  # Each caller has a different amount of variables because it contains different datasets
139
140
141  def fill_sample_dictionary(base_entropy, sample_dictionary, path, reference_dictionary):
142
143      callerlengths = [0] * number_of_callers
144
145      index = 0
146
147      total_mode_value = 0
148
149      list_of_paths = create_list_of_paths(path)
150
151      for vcf_file in list_of_paths:
152
153          index += 1
154
155          opened_vcf_file = vcf.Reader(open(vcf_file, 'r'))
156
157          removaldict = iterate_over_file_to_extract_data(base_entropy, sample_dictionary,
158
159                                              reference_dictionary, opened_vcf_file,
160                                              → vcf_file)
161
162          mode_value = get_mode_value(removaldict)
163
164          add_length_to_caller_lengths_based_on_file_name(vcf_file, mode_value, callerlengths)
165
166          refill_dictionary_with_zero_arrays_for_each_file(sample_dictionary, index, mode_value)
167
168          total_mode_value += mode_value
169
170          list_of_passed_samples, vcf_list = add_mode_values_into_list_of_samples(sample_dictionary,
171
172          → total_mode_value)
173
174      return list_of_passed_samples, callerlengths, vcf_list

```

```

142
143
144 # this method fills the dictionary with empty arrays with the same length as the ones that were supposed
145 # to be added
146
147 def refill_dictionary_with_zero_arrays_for_each_file(full_dictionary, index, length_of_data_array):
148     empty_set = []
149     for i in range(length_of_data_array):
150         empty_set.append(0)
151     for item in full_dictionary:
152         checksum = len(full_dictionary[item][0])
153         if checksum < index:
154             arbinfo = empty_set
155             full_dictionary[item][0].append(arbinfo)
156
157 # this method iterates through all the files to extract data from each sample. It uses methods from the
158 # methods.py function, which parses each record for data.
159
160 def iterate_over_file_to_extract_data(base_entropy, sample_dictionary, recorddictionary, vcf_reader1,
161                                     vcf_file):
162     removaldict = {}
163     for record in vcf_reader1:
164         if "GL" in str(record.CHROM):
165             continue
166         sample_name = get_sample_name_from_record(record)
167         sample_data = getallvalues(record, recorddictionary, base_entropy, vcf_file)
168         sample_dictionary[sample_name][0].append(sample_data)
169         sample_dictionary[sample_name][1] = record
170         create_removal_dict(sample_data, removaldict)
171
172 # this method counts the mode number of entries in the dictionary. Due to certain vcf files having
173 # multiple possible number of entries for a field, this will create an error
174 # as the size of the input arrays should always be constant. Thus, any sample that does not fit the
175 # array should be removed.
176
177 def create_removal_dict(sample_data, removaldict):
178     count = 0
179     count += len(sample_data)
180     if count not in removaldict:
181         removaldict[count] = 1
182     else:

```

```

182         removaldict[count] += 1
183
184
185     # this method prepares the reference genome dictionary for use in entropy calculations
186
187     def get_reference_dictionary_for_entropy(reference_path):
188         record_dictionary = SeqIO.to_dict(SeqIO.parse(reference_path, "fasta"), key_function=get_chr)
189         return record_dictionary
190
191     # this method ensures that the files inputed are correct
192
193     def ignore_file(vcf_file):
194         if "vcf" not in vcf_file or "truth" in vcf_file:
195             return True
196         return False
197
198     # this method creates the set of keys for the dictionary
199
200     def create_dictionary_keys(vcf_reader, sample_dictionary):
201         for record in vcf_reader:
202             if "GL" in str(record.CHROM):
203                 continue
204             sample_name = get_sample_name_from_record(record)
205             sample_dictionary[sample_name] = [[], []] # fullname has become a key in fulldictionary
206         return sample_dictionary
207
208     # standard method that returns a tuple of the variant call object with the chromosome, position,
209     # reference and tuple of alternates
210
211     def get_sample_name_from_record(record):
212         templist = []
213         for item in record.ALT:
214             templist.append(str(item).upper())
215         sample_name = (str(record.CHROM), str(record.POS), str(record.REF).upper(), tuple(templist))
216         return sample_name
217
218     # this method sets the length of the input neural networks
219
220     def add_length_to_caller_lengths_based_on_file_name(vcf_file, caller_length, callerlengths):
221         if "fb" in vcf_file:
222             callerlengths[0] = caller_length
223         if "hc" in vcf_file:
224             callerlengths[1] = caller_length
225         if "ug" in vcf_file:

```

```

225         callerlengths[2] = caller_length
226
227     if "pind" in vcf_file:
228
229         callerlengths[3] = caller_length
230
231     if "st" in vcf_file:
232
233         callerlengths[4] = caller_length
234
235
236 # this method wraps the create truth dictionary method and is used to checking that the dictionary file
237 # has the correct name
238
239
240
241 def generate_truth_list(path):
242
243     generated_truth_dictionary = {}
244
245     for truth_file in os.listdir(path):
246
247         if "truth" not in truth_file:
248
249             continue
250
251         create_truth_dictionary(generated_truth_dictionary, truth_file)
252
253     return generated_truth_dictionary
254
255
256
257 # this method takes in the mutation (in a tuple) and checks if that mutation exists in the truth
258 # dictionary
259
260 # A mutation exists if the chromosome, reference and position of the variant call is correct, AND one of
261 # the alternate alleles it contains
262
263 # is also an alternate allele in the truth dataset
264
265
266 def check_sample_against_truth_dictionary(tuple_name, final_truth_list, truth_dictionary):
267
268     temp_tuple = (tuple_name[0], tuple_name[1], tuple_name[2])
269
270     if temp_tuple in truth_dictionary:
271
272         for alternate in tuple_name[3]:
273
274             if alternate in truth_dictionary[temp_tuple]:
275
276                 final_truth_list.append(1)
277
278             return
279
280     final_truth_list.append(0)
281
282     return
283
284
285
286 # This method loads the paths of the files into local variables
287
288
289 def load_references(user_input):
290
291     file1 = user_input['input'][0]
292
293     referencepath = user_input['reference']
294
295     output_location = user_input['output']
296
297     return file1, referencepath, output_location
298
299
300
301 # This method saves all the processed data into files that can be used for other purposes later or
302 # loaded natively instead of doing the processing again
303
304

```

```

265  def save_files(output_location, x_array, length_of_caller_outputs, sample_list, truth_dict,
266      ↵ vcf_dictionary_file,
267      ↵ y_array=[]):
268
269      file2 = output_location
270
271      x_data_file_name = str(file2) + str(X_DATA_NAME)
272
273      np.save(x_data_file_name, x_array)
274
275      vcf_file_name = str(file2) + str(VCF_LIST_FILE_NAME)
276
277      caller_length_file_name = str(file2) + str(CALLER_LENGTH_FILE_NAME)
278
279      truth_dictionary_name = str(file2) + str(TRUTH_DICTIONARY_NAME)
280
281      list_of_inputs_name = str(file2) + str(LIST_OF_INPUTS_NAME)
282
283      np.save(caller_length_file_name, length_of_caller_outputs)
284
285      with open(list_of_inputs_name, 'wb') as samplesave1:
286
287          pickle.dump(sample_list, samplesave1)
288
289      with open(truth_dictionary_name, 'wb') as samplesave2:
290
291          pickle.dump(truth_dict, samplesave2)
292
293      with open(vcf_file_name, 'wb') as samplesave3:
294
295          pickle.dump(vcf_dictionary_file, samplesave3)
296
297      if y_array != []:
298
299          y_data_file_name = str(file2) + str(Y_DATA_NAME)
300
301          np.save(y_data_file_name, y_array)
302
303
304      # This method takes in two dictionaries, a dictionary of truth mutations and a dictionary of sample
305      ↵ mutations,
306
307      # checks whether each of the sample variables are inside the truth dictionary
308
309      # and returns 2 arrays, an array of samples and an array of accompanying truth labels
310
311
312  def check_predicted_with_truth(passed_list_of_samples, dictionary_of_truth=[]):
313
314      final_array_of_samples = []
315
316      final_truth_list = []
317
318      for item in passed_list_of_samples:
319
320          if dictionary_of_truth:
321
322              check_sample_against_truth_dictionary(item[0], final_truth_list, dictionary_of_truth)
323
324              temp_array = []
325
326              for row in item[1]:
327
328                  temp_array.extend(row)
329
330              final_array_of_samples.append(temp_array)
331
332          if dictionary_of_truth:
333
334              return final_truth_list, final_array_of_samples
335
336      return final_array_of_samples
337
338
339      # This method ensures that only the variables that have the modal number of features are used
340
341      # in neural network training to ensure all array sizes are the same
342
343
344  def add_mode_values_into_list_of_samples(full_dictionary, mode_value):

```

```

307     list_of_passed_samples = []
308     vcf_list = []
309     for key in full_dictionary:
310         second_count = 0
311         for item in full_dictionary[key][0]:
312             second_count += len(item)
313         if second_count != mode_value:
314             continue
315         list_of_passed_samples.append([key, full_dictionary[key][0]])
316         vcf_list.append(full_dictionary[key][1])
317     return list_of_passed_samples, vcf_list
318
319 # This method gets the modal number of features from a modal dictionary
320
321 def get_mode_value(removaldict):
322     curr = 0
323     mode_value = 0
324     for new_key in removaldict:
325         if removaldict[new_key] > curr:
326             curr = removaldict[new_key]
327             mode_value = new_key
328     return mode_value
329
330 # This method iterates through the dataset to create a modal dictionary which contains a key-value pair
331 # ← of (number of features - number of times seen).
332
333 def iterate_through_dictionary_to_find_mode_size(full_dictionary):
334     removaldict = {}
335     samples = 0
336     for key in full_dictionary:
337         samples += 1
338         if samples == sample_limit:
339             break
340         count = 0
341         for item in full_dictionary[key]:
342             count += len(item)
343         if count not in removaldict:
344             removaldict[count] = 1
345         else:
346             removaldict[count] += 1
347     return removaldict
348
349

```

```
350 if __name__ == "__main__":
351     np.seterr(divide='raise', invalid='raise')
352
353     parser = argparse.ArgumentParser(description="train neural net")
354
355     parser.add_argument('-i', '--input', help="give directories with files", nargs='+')
356     parser.add_argument('-d', '--debug', help="look at matrixes built")
357     parser.add_argument('-r', '--reference', help="")
358     parser.add_argument('-o', '--output', help="")
359
360     paths = parser.parse_args()
361
362     start = time.time()
363
364     load_and_save_data(paths)
```

7.2 train_network.py

```
1 #This script is called by the generate_matrixes.py script and contains the implementation of the neural
2 → network.
3
4 #Input : np.array of features from generate_matrixes with accompanying truth labels, feature set
5 → lengths, a dictionary of vcf object records, as well as the list of sample features
6
7 #Output : A VCF file containing all the filtered entries by the neural network, as well the list of
8 → accompanying scores
9
10 #Overall Strategy :
11
12 #Perform SMOTE oversampling of the input features, and then use the features to train the neural network
13
14 #After training, perform validation on test dataset, and subsequently prepare a vcf file with filtered
15 → entries
16
17
18
19
20
21
22
23
24
25
26
27 #set constants
```

```

28     PCA_COMPONENTS = 8
29     STEP_INCREMENT = 10
30     RECURSION_LIMIT = 0.0002
31     VERBOSE = 1
32     seed = 1337
33
34     # Initialise random seed for reproducibility
35     np.random.seed(seed)
36
37     #Prepare file names for saving
38     vcf_file_name = "/ANN/truevcf.vcf"
39     keras_model_name = "/ANN/model"
40     model_truth_name = "/ANN/modeltruths.txt"
41     model_predictions_name = "/ANN/modelpredictions.txt"
42     original_vcf_reader = "/data/backup/metacaller/stage/data/version6.3a/hc.vcf.normalisedtrain.vcf"
43
44     # this method takes in a path and returns training matrixes for the ANN
45     # The path should contain n caller vcf files and 1 truth file
46     # vcf files should be labelled with vcf and truth file should be labelled with truth
47     # no other file should be present in the folder
48     def main_gather_input_execute_prep_output(array_sizes, dict_of_truth_input, fullmatrix_sample,
        ↪   fullmatrix_truth, list_of_samples_input, save_location, vcf_dictionary):
49         calculated_prediction_actual, calculated_truth_actual = train_neural_net(20, 10, fullmatrix_sample,
        ↪   fullmatrix_truth,
        ↪   save_location, array_sizes)
50         get_all_relevant_scores(calculated_prediction_actual, calculated_truth_actual, dict_of_truth_input,
        ↪   list_of_samples_input, vcf_dictionary, save_location)
51
52     # This method counts the number of false negatives inside the input sample
53
54     def count_false_negative(calculated_prediction_actual, calculated_truth_actual):
55         count_false_negative = 0
56         for i in range(len(calculated_prediction_actual)):
57             if calculated_prediction_actual[i] == 0 and calculated_truth_actual[i] == 1:
58                 count_false_negative += 1
59         return count_false_negative
56
61     # this is the wrapper function for the recursive hill climbing algorithm to get the best f1 score
62     # It starts from a low threshold value, and marginally increases the threshold until it is unable to
        ↪   find
63     # any better F1 scores. It then reports the threshold, F1 score and produces the filtered callset
64
65     def get_all_relevant_scores(calculated_prediction_actual, calculated_truth_actual, dict_of_truth_input,
        ↪   list_of_samples_input, vcf_list, outputpath):
66

```

```

67     print "Here are some predictions", calculated_prediction_actual[:100]
68     print "here are some truths", calculated_prediction_actual[:100]
69     f1_score_left = get_scores(calculated_prediction_actual, calculated_truth_actual, 0.0,
70         ↪ list_of_samples_input, dict_of_truth_input)
71     guess_f1_final_score, guess_f1_final = recursive_best_f1_score(calculated_prediction_actual,
72         ↪ calculated_truth_actual, dict_of_truth_input, list_of_samples_input, 0.0, f1_score_left, 0.2)
73     get_scores(calculated_prediction_actual, calculated_truth_actual, guess_f1_final,
74         ↪ list_of_samples_input, dict_of_truth_input, VERBOSE)
75     produce_vcf_file(calculated_prediction_actual, guess_f1_final, list_of_samples_input, vcf_list,
76         ↪ outputpath)

77
78 # This method produces the vcf file through filtering with the neural network threshold calls
79
80
81 def produce_vcf_file(calculated_prediction_actual, guess_f1_final, list_of_samples_input, vcf_list,
82     ↪ outputpath):
83     prediction = []
84     for item in calculated_prediction_actual:
85         if item > guess_f1_final:
86             prediction.append(1)
87         else:
88             prediction.append(0)
89     list_of_records = []
90     for i in range(len(list_of_samples_input)):
91         if prediction[i] == 1:
92             list_of_records.append(vcf_list[i])
93     vcf_reader = vcf.Reader(filename=original_vcf_reader)
94     vcf_writer = vcf.Writer(open(outputpath + vcf_file_name, 'w'), vcf_reader)
95     for record in list_of_records:
96         vcf_writer.write_record(record)

97
98 # This method is the recursive function that attempts to find the threshold that produces the best f1
99 # score. It does this
100 # by iterating through steps of thresholds (0.2, 0.02 and 0.002) until no better F1 score can be found
101 # for a marginal increase in threshold.
102 # It then returns the best F1 score and the threshold
103
104 def recursive_best_f1_score(calculated_prediction_actual, calculated_truth_actual, dict_of_truth_input,
105     list_of_samples_input, guess, guess_score, step):
106     if step <= RECURSION_LIMIT:
107         return guess_score, guess
108     new_guess = guess + step
109     new_guess_score = get_scores(calculated_prediction_actual, calculated_truth_actual, new_guess,
110         list_of_samples_input, dict_of_truth_input)
111     if new_guess_score > guess_score:

```

```

104     return recursive_best_f1_score(calculated_prediction_actual, calculated_truth_actual,
105                                     ↵ dict_of_truth_input,
106                                     list_of_samples_input, new_guess, new_guess_score, step)
107
108     return recursive_best_f1_score(calculated_prediction_actual, calculated_truth_actual,
109                                     ↵ dict_of_truth_input,
110                                     list_of_samples_input, guess, guess_score, step / STEP_INCREMENT)
111
112
113 # this method uses pre-loaded data to train the neural network. It is optional and only used when this
114 ↵ python script is called natively and not imported
115
116 def load_references(input_paths):
117
118     input_paths = vars(input_paths)
119
120     fullmatrix_sample = np.load(input_paths['input'][0])
121
122     fullmatrix_truth = np.load(input_paths['input'][1])
123
124     with open(input_paths['input'][3], 'rb') as fp1:
125
125         list_of_samples_input = pickle.load(fp1)
126
127     with open(input_paths['input'][4], 'rb') as fp2:
128
128         dict_of_truth_input = pickle.load(fp2)
129
130     array_sizes = np.load(input_paths['input'][5])
131
132     with open(input_paths['input'][6], 'rb') as fp3:
133
133         vcf_dictionary = pickle.load(fp3)
134
135     orig_stdout = sys.stdout
136
136     f = file(str(input_paths['input'][3]) + '.txt', 'w')
137
137     sys.stdout = f
138
138     return array_sizes, dict_of_truth_input, fullmatrix_sample, fullmatrix_truth,
139                                     ↵ list_of_samples_input, input_paths, vcf_dictionary
140
141
142 # this method solves the double false negative problem that is created due to the neural network
143 ↵ prediction scheme
144
145
146 def remove_duplicated_false_negative(prediction_list, truth_list, false_negatives):
147
148     count = 0
149
150     removal_list = []
151
152     for i in range(len(prediction_list) - 1, -1, -1):
153
153         if count == false_negatives:
154
155             break
156
157         if prediction_list[i] == 0 and truth_list[i] == 1:
158
158             removal_list.insert(0, i)
159
160             count += 1
161
162         for index in removal_list:
163
163             prediction_list.pop(index)
164
165             truth_list.pop(index)
166
167     return prediction_list, truth_list

```

```

143 # this method takes in the binary truth and predicted samples and calculates the true positive rate,
144 # false positive rate, recall, precision and f1 score
145
146 def get_scores(actual_predictions, actual_truth, value, sample_list, truth_dictionary, verbose=0):
147     temp_actual_truth = list(actual_truth)
148     prediction = []
149     for item in actual_predictions:
150         if item > value:
151             prediction.append(1)
152         else:
153             prediction.append(0)
154     false_negatives = count_false_negative(actual_predictions, actual_truth)
155     finalpredictionnumbers, finaltruthnumbers = add_negative_data(sample_list, truth_dictionary,
156     ↪ prediction, temp_actual_truth)
157     finalpredictionnumbers, finaltruthnumbers =
158     ↪ remove_duplicated_false_negative(finalpredictionnumbers, finaltruthnumbers, false_negatives)
159     final_f1_score = f1_score(finaltruthnumbers, finalpredictionnumbers)
160     if verbose:
161         print_scores(actual_truth, final_f1_score, finalpredictionnumbers, finaltruthnumbers,
162         ↪ prediction, value)
163     return final_f1_score
164
165
166 # default method for printing all relevant scores
167
168 def print_scores(actual_truth, final_f1_score, finalpredictionnumbers, finaltruthnumbers, prediction,
169     ↪ value):
170     final_false_positive, final_true_negative = perf_measure(finaltruthnumbers, finalpredictionnumbers)
171     print "final false positive rate is :", final_false_positive
172     print "final true negative rate is :", final_true_negative
173     print "final precision score is :", precision_score(finaltruthnumbers, finalpredictionnumbers)
174     print "final recall score is :", recall_score(finaltruthnumbers, finalpredictionnumbers)
175     print "threshold is", value
176     print "final F1 score is : ", final_f1_score
177
178 # This method looks at the set of predicted samples and the set of truths and adds the false negatives
179 # to the predicted sample.
180
181 def add_negative_data(list_of_samples, dict_of_truth, array_of_predicted, array_of_truth):
182     dict_of_samples = generate_sample_dictionary(array_of_predicted, list_of_samples)
183     list_of_truth = generate_list_of_truth(dict_of_truth)
184     new_array_of_predicted = list(array_of_predicted)
185     new_array_of_truth = list(array_of_truth)
186     original_length = len(new_array_of_predicted)
187     for item in list_of_truth:

```

```

181     fillnegative(item, dict_of_samples, new_array_of_predicted, new_array_of_truth)
182     print "number of false data samples are", (len(new_array_of_predicted) - original_length)
183     return new_array_of_predicted, new_array_of_truth
184
185 # This method generates a list of truth variant calls from a dictionary of truth variant calls.
186
187 def generate_list_of_truth(dict_of_truth):
188     list_of_truth = []
189     for key in dict_of_truth:
190         mytuple = dict_of_truth[key]
191         temptuple = []
192         for item in mytuple:
193             temptuple.append(item)
194         list_of_truth.append([key[0], key[1], key[2], temptuple])
195     return list_of_truth
196
197 # This method generates a dictionary of sample variant calls from a list of sample variant calls.
198
199 def generate_sample_dictionary(array_of_predicted, list_of_samples):
200     dict_of_samples = {}
201     for i in range(len(list_of_samples)):
202         item = list_of_samples[i]
203         if array_of_predicted[i] == 0:
204             continue
205         new_key = (item[0][0], item[0][1], item[0][2])
206         new_value = item[0][3]
207         if new_key not in dict_of_samples:
208             dict_of_samples[new_key] = new_value
209         else:
210             dict_of_samples[new_key] = list(dict_of_samples[new_key])
211             dict_of_samples[new_key].extend(new_value)
212             dict_of_samples[new_key] = tuple(dict_of_samples[new_key])
213             # print dict_of_samples[new_key]
214     return dict_of_samples
215
216 # Actual method to calculate false positive, false negative rates
217
218 def perf_measure(y_actual, y_hat):
219     true_positive = 0
220     false_positive = 0
221     false_negative = 0
222     true_negative = 0
223
224     for i in range(len(y_hat)):
```

```

225     if y_actual[i] == 1 and y_hat[i] == 1:
226         true_positive += 1
227     for i in range(len(y_hat)):
228         if y_hat[i] == 1 and y_actual[i] == 0:
229             false_positive += 1
230     for i in range(len(y_hat)):
231         if y_actual[i] == 1 and y_hat[i] == 0:
232             false_negative += 1
233     for i in range(len(y_hat)):
234         if y_hat[i] == 0 and y_actual[i] == 0:
235             true_negative += 1
236
237     print "true positives :", true_positive
238     print "false positives :", false_positive
239     print "false negatives :", false_negative
240     print "true negatives :", true_negative
241
242     true_positive = float(true_positive)
243     false_positive = float(false_positive)
244     false_negative = float(false_negative)
245     if false_positive == 0 and true_positive == 0:
246         false_positive_rate = 0
247     else:
248         false_positive_rate = false_positive / (false_positive + true_positive)
249     if false_negative == 0 and true_positive == 0:
250         true_negative_rate = 0
251     else:
252         true_negative_rate = false_negative / (false_negative + true_positive)
253
254     return false_positive_rate, true_negative_rate
255
256
257 # comparator method that takes a tuple and checks whether it is in the dictionary of samples, if it is
258 # not, then add a false negative call to the dataset
259
260 def fillnegative(tuple1, sampledict, arrayofsamples, arrayoftruths):
261     tuple2 = (tuple1[0], tuple1[1], tuple1[2])
262     if tuple2 in sampledict:
263         for ALT in tuple1[3]:
264             if ALT in sampledict[tuple2]:
265                 return
266             arrayofsamples.append(0)
267             arrayoftruths.append(1)
268

```

```

268 # main method that performs neural network training. This method takes in the sample matrixes, the truth
269 # variables, a save file location, number of epochs,
270 # size of input arrays and the minibatch training size. It first performs SMOTE on the input dataset,
271 # then splits it into training and test dataset. It then
272 # initialises the deep learning layers, compiles the neural network and uses the input data to fit the
273 # network. The best set of weights at any point is saved
274 # to a file and reloaded at the end of the fitting. After training, the neural network is used to
275 # predict the original un-oversampled dataset
276
277 def train_neural_net(mybatch_size, mynb_epoch, myX_train, myy_train, location, arraysize):
278     fb_size, hc_size, ug_size, pindel_size, st_size = get_sizes(array_sizes)
279     X_resampled, y_resampled = do_smote_resampling(myX_train, myy_train)
280     X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled,
281                                                       test_size=0.33, random_state=seed)
282     X_fb, X_hc, X_ug, X_pindel, X_st = prep_input_samples(array_sizes, X_train)
283     X_fb_test, X_hc_test, X_ug_test, X_pindel_test, X_st_test = prep_input_samples(array_sizes, X_test)
284     batch_size = mybatch_size
285     nb_epoch = mynb_epoch
286
287     fb_branch = Sequential()
288     develop_first_layer_matrixes(fb_branch, fb_size)
289
290     hc_branch = Sequential()
291     develop_first_layer_matrixes(hc_branch, hc_size)
292
293     ug_branch = Sequential()
294     develop_first_layer_matrixes(ug_branch, ug_size)
295
296     pindel_branch = Sequential()
297     develop_first_layer_matrixes(pindel_branch, pindel_size)
298
299     st_branch = Sequential()
300     develop_first_layer_matrixes(st_branch, st_size)
301
302     final_model = Sequential()
303     final_model.add(Merge([fb_branch, hc_branch, ug_branch, pindel_branch, st_branch], mode='concat',
304                           concat_axis=1))
305     final_model.add(Dense(24, activation='linear'))
306     final_model.add(LeakyReLU(alpha=0.05))
307     final_model.add(Dense(6, activation='linear'))
308     final_model.add(LeakyReLU(alpha=0.05))
309     final_model.add(Dense(1, activation='linear'))
310     final_model.add(Activation('sigmoid'))
311     print (final_model.summary())

```

```

307     adam = Adam(lr=0.00001, rho=0.9, epsilon=1e-08, decay=0.0)
308     final_model.compile(loss='binary_crossentropy',
309                           optimizer=adam,
310                           metrics=['accuracy'])
311
312     filepath = location + "/best_weights.hdf5"
313     checkpoint = ModelCheckpoint(filepath, monitor='val_acc', verbose=1, save_best_only=True,
314                                  mode='max')
315     callbacks_list = [checkpoint]
316     model_history = final_model.fit([X_train], y_train, batch_size=batch_size, nb_epoch=nb_epoch,
317                                     validation_split=0.2, verbose=2, callbacks=callbacks_list)
317     final_model = load_model(location + "/best_weights.hdf5")
318     print model_history.history['val_acc'], model_history.history['val_acc']
319     print model_history.history['val_loss'], model_history.history['val_loss']
320     np.save(location + "/best_weights.hdf5", model_history.history['val_acc'])
321     np.save(location + "/best_weights.hdf5", model_history.history['val_loss'])
322     scores = final_model.evaluate([X_test], y_test)
323     print scores
324     final_prediction_array_probabilities = final_model.predict([myX_train])
325     final_prediction_array_probabilities = np.squeeze(final_prediction_array_probabilities)
326     save_model_details(final_model, final_prediction_array_probabilities, myy_train, location)
327
328     return final_prediction_array_probabilities, myy_train
329
330 # Method to perform SMOTE oversampling
331
332 def do_smote_resampling(myX_train, myy_train):
333     sm = SMOTE(kind='regular')
334     where_are_NaNs = np.isnan(myX_train)
335     myX_train[where_are_NaNs] = 0
336     X_resampled, y_resampled = sm.fit_sample(myX_train, myy_train)
337     return X_resampled, y_resampled
338
339 # this method saves the details of the neural network
340
341 def save_model_details(final_model, save_model_probabilities, trutharray, location):
342     name1 = location + model_predictions_name
343     name2 = location + model_truth_name
344     name3 = location + keras_model_name
345     np.save(name1, save_model_probabilities)
346     np.save(name2, trutharray)
347     final_model.save(name3)
348
349 # this method gets the array size of the features used

```

```

350
351     def get_sizes(array_sizes):
352         fb_size = array_sizes[0]
353         hc_size = array_sizes[1]
354         ug_size = array_sizes[2]
355         pindel_size = array_sizes[3]
356         st_size = array_sizes[4]
357
358         return fb_size + hc_size + ug_size + pindel_size + st_size
359
360     # this method uses a map function to filter data such that each merge layer gets the correct set of data
361
362     def prep_input_samples(array_sizes, x_training_data):
363         count = 0
364
365         X_fb = np.array(map(lambda x: x[count:array_sizes[0]], x_training_data))
366         count += array_sizes[0]
367
368         X_hc = np.array(map(lambda x: x[count:count + array_sizes[1]], x_training_data))
369         count += array_sizes[1]
370
371         X_ug = np.array(map(lambda x: x[count:count + array_sizes[2]], x_training_data))
372         count += array_sizes[2]
373
374         X_pindel = np.array(map(lambda x: x[count:count + array_sizes[3]], x_training_data))
375         count += array_sizes[3]
376
377         X_st = np.array(map(lambda x: x[count:count + array_sizes[4]], x_training_data))
378         count += array_sizes[4]
379
380         return X_fb, X_hc, X_ug, X_pindel, X_st
381
382
383     if __name__ == "__main__":
384
385         parser = argparse.ArgumentParser(description="train neural net")
386
387         parser.add_argument('-i', '--input', help="give directories with files", nargs='+')
388
389         input_path = parser.parse_args()
390
391         array_sizes, dict_of_truth_input, fullmatrix_sample, fullmatrix_truth, \
392         list_of_samples_input, paths, vcf_dictionary = load_references(input_path)
393
394         main_gather_input_execute_prep_output(array_sizes, dict_of_truth_input, fullmatrix_sample,
395             → fullmatrix_truth, list_of_samples_input, paths, vcf_dictionary)

```

7.3 compute_bayesian.py

```

1  #This script takes in a VCF file with functional annotation already done, and computes the bayesian
2  #network using the annotations. It produces a sorted list of vcf entries in a text file, with
3  #accompanying annotation scores
4
5  #Input : VCF file with functional annotation
6
7  #Output : A sorted list of vcf entries with accompanying annotation scores, redirected from stdout to a
8  #file
9
10 #Overall Strategy :

```

```

5  #First extract all the features from the vcf files and then perform feature-wise normalisation.
6  #Subsequently, prepare the bayesian network by creating edges, nodes, preparing prior distributions
7  #Finally use features to update the bayesian network to obtain final probabilities for importance
8  #Report a list of sorted probabilites for easy ranking

9

10 import matplotlib
11 import vcf
12
13 matplotlib.use('Agg')
14
15 from pomegranate import *
16
17 #main method for loading references into local variables
18
19 def load_reference(paths):
20     paths = vars(paths)
21     input = paths['input']
22     opened_vcf_file = vcf.Reader(open(input, 'r'))
23     name3 = input + "finalscores.txt"
24     # orig_stdout = sys.stdout
25     # f = file(name3 + '.txt', 'w')
26     # sys.stdout = f
27     return opened_vcf_file
28
29 #method for getting functional annotation scores
30
31 def get_scores(record):
32     list_of_important_mutations = [record.INFO['SIFT_score'], record.INFO['LRT_score'],
33                                     record.INFO['MutationAssessor_score'],
34                                     record.INFO['Polyphen2_HVAR_score'], record.INFO['FATHMM_score']]
35     if 'NN_prediction' in record.INFO:
36         NN_prediction = record.INFO['NN_prediction'][0]
37     else:
38         NN_prediction = -1
39     list_of_important_mutations = map(lambda x: x[0], list_of_important_mutations)
40     list_of_important_mutations = map(lambda x: None if x == None else float(x),
41                                     list_of_important_mutations)
41     return NN_prediction, list_of_important_mutations
42
43 #main method that controls I/O - it gets the input, applies the main function and then prepares the
44 #→ output
45
46 def main(paths):
47     vcf_object = load_reference(paths)

```

```

47     full_list_of_scores = analyse_main(vcf_object)
48     prepare_output(full_list_of_scores)
49
50     #this method controls the processes applied to the vcf file - for each record, it extract the list of
51     # scores,
52     # normalises it, compute probabilities, sorts it and then return output
53
54     def analyse_main(vcf_object):
55         full_list_of_scores = extract_list_of_scores(vcf_object)
56         apply_feature_wise_normalisation(full_list_of_scores)
57         compute_network_and_probabilities(full_list_of_scores)
58         full_list_of_scores.sort(key=lambda x: x[4], reverse=True)
59         return full_list_of_scores
60
61     # since print is redirected to stdoutput, print function is used to store output
62
63     def prepare_output(full_list_of_scores):
64         for item in full_list_of_scores:
65             print item[2], item, item[2].INFO['Gene.refGene']
66
67     # wrapper function used to create bayesian network for all records
68
69     def compute_network_and_probabilities(full_list_of_scores):
70         for record in full_list_of_scores:
71             network = create_network_and_compute_probabilities(record)
72             compute_record(network, record)
73
74     # this function applies a featurewise normalisation of all features to a range of 0-1, and flip scores
75     # for certain features
76
77     def apply_feature_wise_normalisation(full_list_of_scores):
78         for i in range(6):
79             min_num = 1000000
80             max_num = -1000000
81             for item in full_list_of_scores:
82                 if item[1][i] != None:
83                     min_num = min(min_num, item[1][i])
84                     max_num = max(max_num, item[1][i])
85             for item in full_list_of_scores:
86                 if item[1][i] != None:
87                     value = ((item[1][i] - min_num) / (max_num - min_num) + 0.2) / 1.3
88                     item[1][i] = value
89                 else:
90                     item[1][i] = 0.5

```

```

90         if i == 0 or i == 5:
91             for item in full_list_of_scores:
92                 if item[1][i] != None:
93                     item[1][i] = -item[1][i]
94
95     # extract list of scores from each record, including all functional annotations, clinvar scores and
96     ↵    dbSNP
97
98     def extract_list_of_scores(vcf_object):
99         count = 0
100        full_list_of_scores = []
101        for record in vcf_object:
102            count += 1
103            nn_prediction, list_of_scores = get_scores(record)
104            if not list(filter(lambda x: x != None, list_of_scores)):
105                continue
106            get_clinvar_scores(list_of_scores, record)
107            snp_present = get_db_snp_scores(record)
108            full_list_of_scores.append([float(nn_prediction), list_of_scores, record, snp_present])
109        return full_list_of_scores
110
111    # Compute the Bayesian Network by assuming observations and attaching mapped probabilities (0,1) to
112    ↵    P(X=True / Y=True)
113
114    def compute_record(network, record):
115        beliefs = network.predict_proba({'Real Gene': 'True', 'ClinVar': 'True', 'PolyPhen': 'True', 'LRT':
116            ↵  'True', 'MutationAssessor': 'True', 'SIFT': 'True', 'FATHMM_gene': 'True', 'rs_gene': 'True'})
117        # print "\n".join("{}\t{}".format(state.name, belief) for state, belief in zip(network.states,
118        ↵  beliefs))
119        # get the probability that the gene is important
120        prob_gene_important = beliefs[2].values()[1]
121        beliefs = map(str, beliefs)
122        record.append(prob_gene_important)
123        record.append(record[2].INFO['snp138'])
124        record.append(record[3])
125
126        # If SNP is present in db-SNP, attach probability of importance to 0.3, else 0.7
127
128    def get_db_snp_scores(record):
129        snp_present = 0.7
130        if record.INFO['snp138'][0] != None:
131            snp_present = 0.3
132        return snp_present

```

```

130  # If snp is present in clinvar, attach probability of importance to 0.7, else 0.3
131
132  def get_clinvar_scores(list_of_scores, record):
133      if record.INFO['clinvar_20150629'][0] != None:
134          list_of_scores.append(0.7)
135      else:
136          list_of_scores.append(0.3)
137
138  # wrapper method to create the bayesian network and compute probabilities
139
140  def create_network_and_compute_probabilities(record):
141      ClinVar_gene, FATHMM_gene, LRT_gene, MutationAssessor_gene, MutationTaster_gene, PolyPhen2_gene,
142      ↪ SIFT_gene, functional_gene, importgene, real_gene, rs_gene = initialise_distributions(
143          record)
144
145  # set up states
146  s1, s10, s11, s2, s3, s4, s5, s6, s7, s8, s9 = generate_states(ClinVar_gene, FATHMM_gene, LRT_gene,
147  ↪ MutationAssessor_gene, MutationTaster_gene, PolyPhen2_gene, SIFT_gene, functional_gene,
148  ↪ importgene, real_gene, rs_gene)
149
150  # set up network
151  network = add_edges_bake_network(s1, s10, s11, s2, s3, s4, s5, s6, s7, s8, s9)
152
153  return network
154
155  # method to create the edges in the network
156
157  def add_edges_bake_network(s1, s10, s11, s2, s3, s4, s5, s6, s7, s8, s9):
158      network = BayesianNetwork("Gene Prediction")
159      network.add_states(s1, s2, s3, s4, s5, s6, s8, s9, s10, s11)
160      network.add_edge(s1, s3)
161      network.add_edge(s2, s3)
162      network.add_edge(s4, s2)
163      network.add_edge(s5, s2)
164      network.add_edge(s6, s2)
165      network.add_edge(s7, s2)
166      network.add_edge(s8, s2)
167      network.add_edge(s9, s2)
168      network.add_edge(s10, s2)
169      network.add_edge(s11, s3)
170      network.bake()
171
172  return network
173
174  # method that generates the nodes in the bayesian network
175
176  def generate_states(ClinVar_gene, FATHMM_gene, LRT_gene, MutationAssessor_gene, MutationTaster_gene,
177  ↪ PolyPhen2_gene, SIFT_gene, functional_gene, importgene, real_gene, rs_gene):

```

```

170     s1 = State(real_gene, name="Real Gene")
171     s2 = State(functional_gene, name="Functional Gene")
172     s3 = State(importgene, name="Important Gene")
173     s4 = State(ClinVar_gene, name="ClinVar")
174     s5 = State(PolyPhen2_gene, name="PolyPhen")
175     s6 = State(LRT_gene, name="LRT")
176     s7 = State(MutationTaster_gene, name="MutationTaster")
177     s8 = State(MutationAssessor_gene, name="MutationAssessor")
178     s9 = State(SIFT_gene, name="SIFT")
179     s10 = State(FATHMM_gene, name="FATHMM_gene")
180     s11 = State(rs_gene, name="rs_gene")
181
182     return s1, s10, s11, s2, s3, s4, s5, s6, s7, s8, s9
183
184
185 #methods to initialise prior distributions in bayesian network
186
187 def initialise_distributions(record):
188
189     ClinVar_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
190
191     PolyPhen2_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
192
193     LRT_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
194
195     MutationTaster_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
196
197     MutationAssessor_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
198
199     SIFT_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
200
201     FATHMM_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
202
203     rs_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
204
205     import_cdp = get_cdp(3, [(record[0] + 0.2) / 1.3, record[3], 0.8])
206
207     functional_cdp = get_cdp(6, record[1])
208
209     functional_gene = ConditionalProbabilityTable(functional_cdp, [ClinVar_gene, PolyPhen2_gene,
210
211         ↳ LRT_gene,
212
213             ↳ MutationAssessor_gene,
214
215             ↳ SIFT_gene, FATHMM_gene])
216
217     real_gene = DiscreteDistribution({'True': 0.5, 'False': 0.5})
218
219     importgene = ConditionalProbabilityTable(import_cdp, [real_gene, rs_gene, functional_gene])
220
221     return ClinVar_gene, FATHMM_gene, LRT_gene, MutationAssessor_gene, MutationTaster_gene,
222
223         ↳ PolyPhen2_gene, SIFT_gene, functional_gene, importgene, real_gene, rs_gene
224
225
226
227 # method that builds the cdp table. n is the number of input variables, probability list gives the
228
229     ↳ probability
230
231     # that the i-th X variable is true P(Xi=True).
232
233
234 def get_cdp(n, prob_list):
235
236     temp_list = create_true_false_matrix(n)
237
238     calculate_probabilities(n, prob_list, temp_list)
239
240     return temp_list

```

```

211
212
213 # Generates a True False matrix using binary counting logic, critical for input in bayesian network
214
215 def create_true_false_matrix(n):
216     temp_list = []
217     for i in range(0, 2 ** n):
218         temp_row = []
219         for j in range(n):
220             number_2 = i // (2 ** (n - j - 1))
221             number_1 = number_2 % 2
222             if number_1 == 0:
223                 temp_row.append('False')
224             else:
225                 temp_row.append('True')
226         temp_list.insert(0, temp_row + ['False'])
227         temp_list.insert(0, temp_row + ['True'])
228     return temp_list
229
230
231 # calculates the probabilities, taking in the true list as well as a list of probabilities. The key here
232     ↪ is
233 # the probability that the mutation is true is related to the scores given by mutation taster etc..
234 # ie  $P(X \text{ is impt} | X \text{ is Clinvar}) = P(X \text{ is Clinvar})$ 
235
236 def calculate_probabilities(n, prob_list, temp_list):
237     for i in range(0, 2 ** (n + 1), 2):
238         true_row = temp_list[i]
239         true_probability = 1
240         false_probability = 1
241         for k in range(0, n, 1):
242             if true_row[k] == 'True':
243                 true_probability *= prob_list[k]
244                 false_probability *= 1 - prob_list[k]    # probability that mutation is false is 1 minus
245                         ↪ mutation is true
246             else:
247                 true_probability *= 1 - prob_list[k]
248                 false_probability *= prob_list[k]
249         final_true_probability = true_probability / (true_probability + false_probability)
250         final_false_probability = false_probability / (true_probability + false_probability)
251         temp_list[i].append(final_true_probability)
252         temp_list[i + 1].append(final_false_probability)
253
254 if __name__ == "__main__":

```

```
253     parser = argparse.ArgumentParser(description="train neural net")
254     parser.add_argument('-i', '--input', help="give directories with files")
255     paths = parser.parse_args()
256     main(paths)
```