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

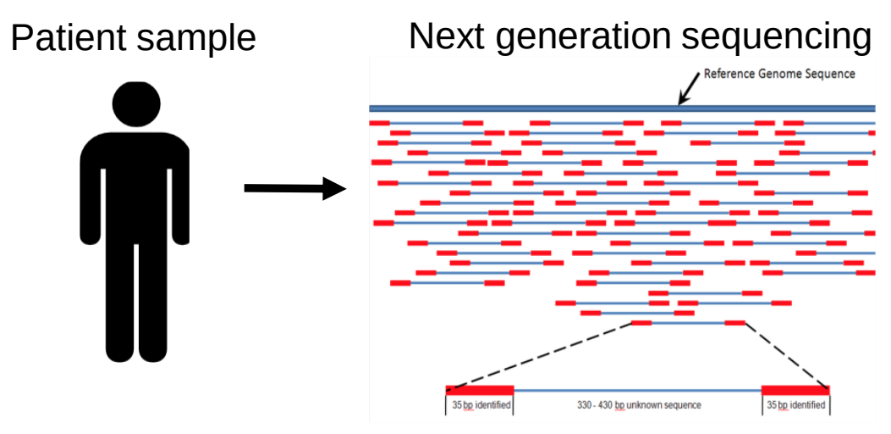
NUS National University of Singapore

Institute of Molecular and Cell Biology A*STAR

Chan Khai Ern Edwin, Manikandan Lakshmanan, Tan Tin Wee & Kenneth Ban

Introduction

Variant calling and gene prioritization are critical steps in NGS analysis



Identifying **high confidence** variants

Annotating and **prioritizing** functional **mutations**

Supporting **medical diagnosis** and **treatment**

- Next generation sequencing (NGS) enables identification of genomic variants in patient samples
- The identification of **high confidence mutations (variant calls)** is critical for downstream analysis
- Given the large number of variant calls from NGS, prioritization is needed to **identify clinically important genes**

Problem 1: Identification of high confidence variant calls

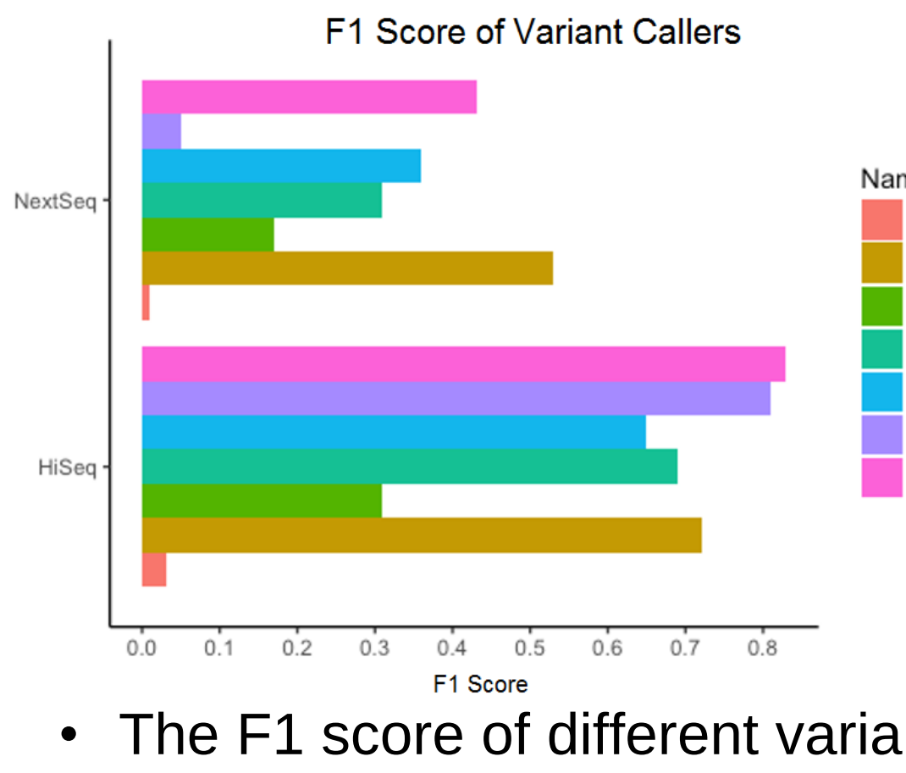
Predicted Class

Actual Class	Yes	No
Yes	TP	FN
No	FP	TN

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

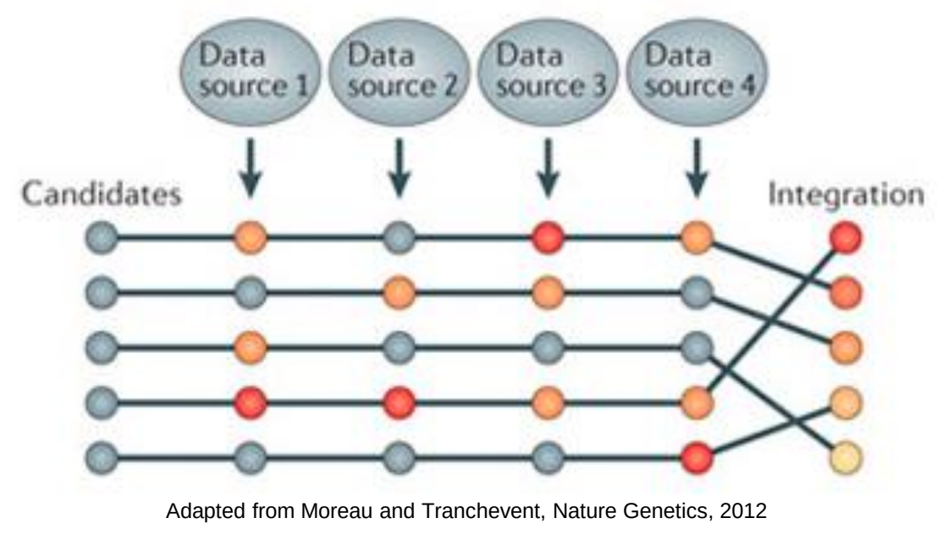
$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$



- The F1 score indicates the precision and recall of each variant caller

- The F1 score of different variant callers can vary
- Each variant caller has its own strengths and weaknesses

Problem 2: Prioritization of functional genes



Adapted from Moreau and Tranchevent, Nature Genetics, 2012

- Multiple candidate mutations can be **difficult to interpret**
- There is a need to integrate different data sources to **rank the importance of mutations**

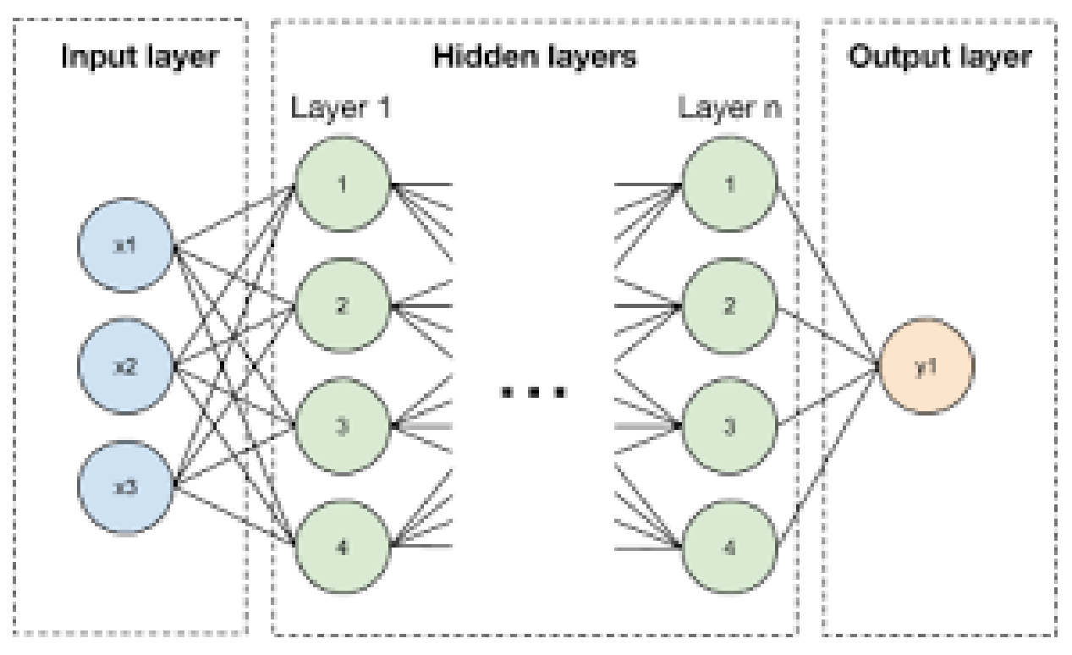
Aims

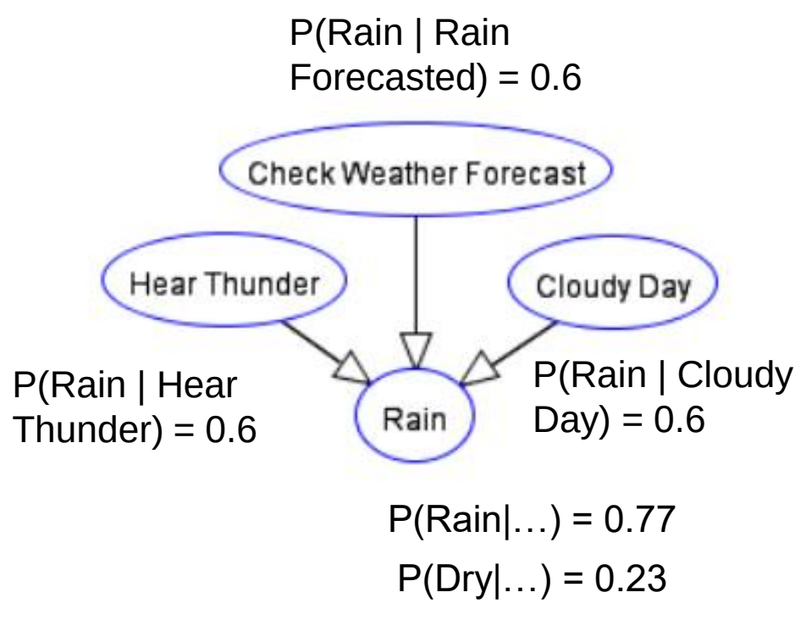
The overall goal is to develop an integrated analytical platform for identifying functionally important mutations via the following methods:

1. To identify high confidence variant calls from an ensemble of variant callers using a **deep learning neural network**
2. To prioritize functional mutations probabilistically using **Bayesian network inference**

Methods

Key Techniques

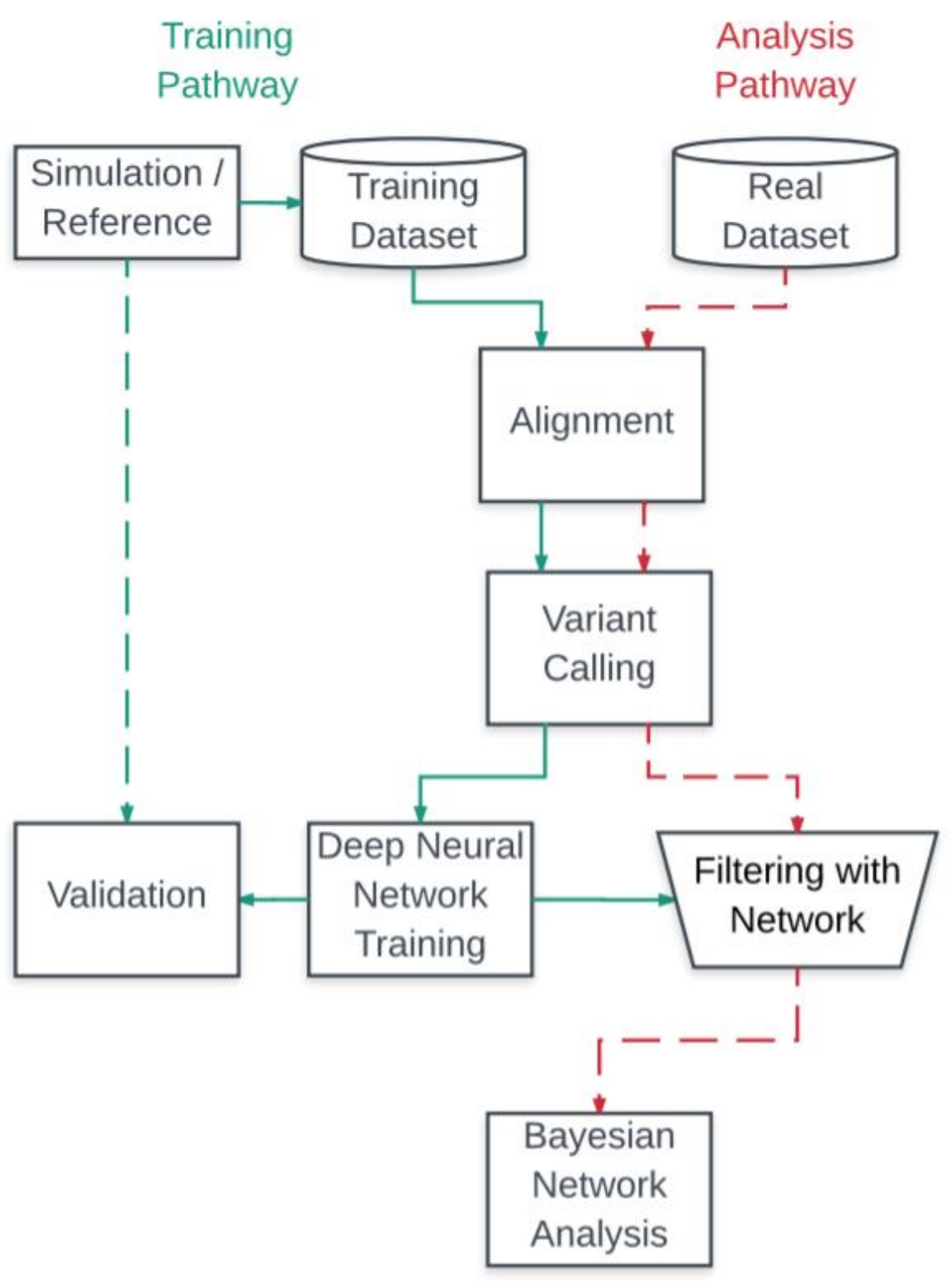



P(Rain | Rain Forecasted) = 0.6
Check Weather Forecast
Hear Thunder
Cloudy Day
P(Rain | Hear Thunder) = 0.6
Rain
P(Rain | Cloudy Day) = 0.6
P(Rain|…) = 0.77
P(Dry|…) = 0.23

- **Deep learning** neural networks comprise a cascade of nonlinear processing units (neurons) that can learn multiple levels of representation from features

- **Bayesian networks** represent probabilistic relationships which can be used to compute the probability of an outcome

Overall Approach



Identification of high confidence variant calls

- Simulate sequence reads with error rates and ground truth variants
- Train deep learning neural network using simulated reads to optimize the network
- Use optimized network to train on reference genome (NA12878) using high confidence calls as ground truth

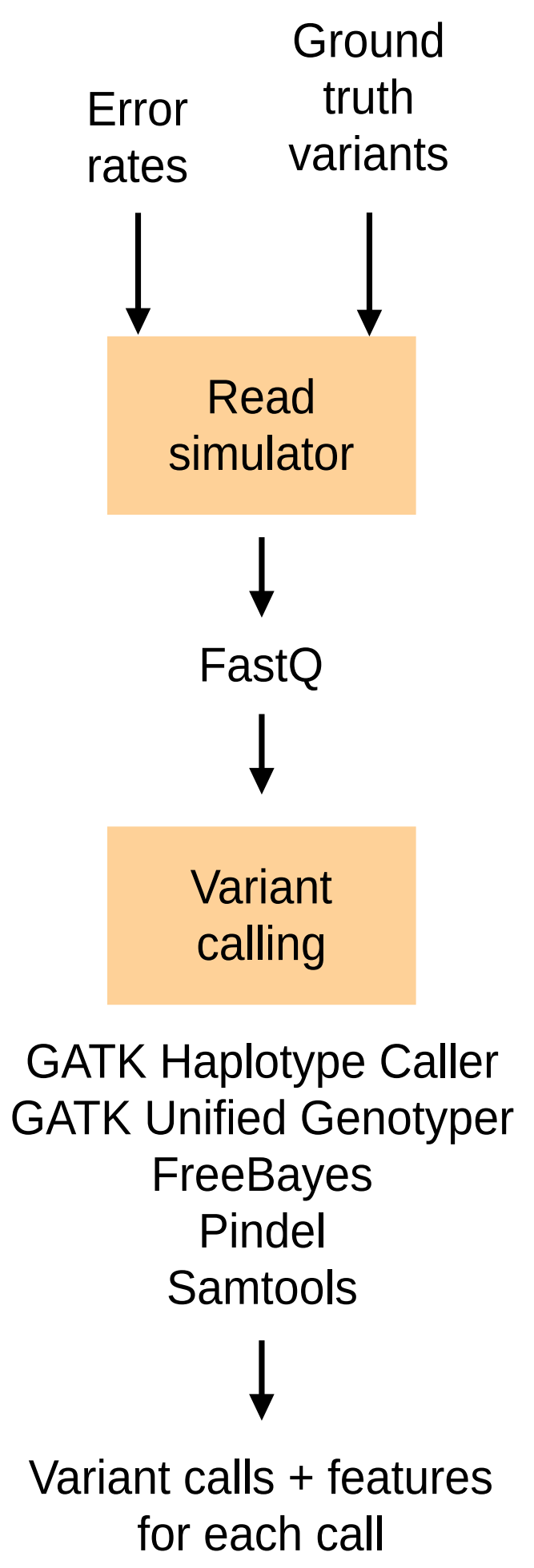
Prioritization of mutations on a cancer dataset

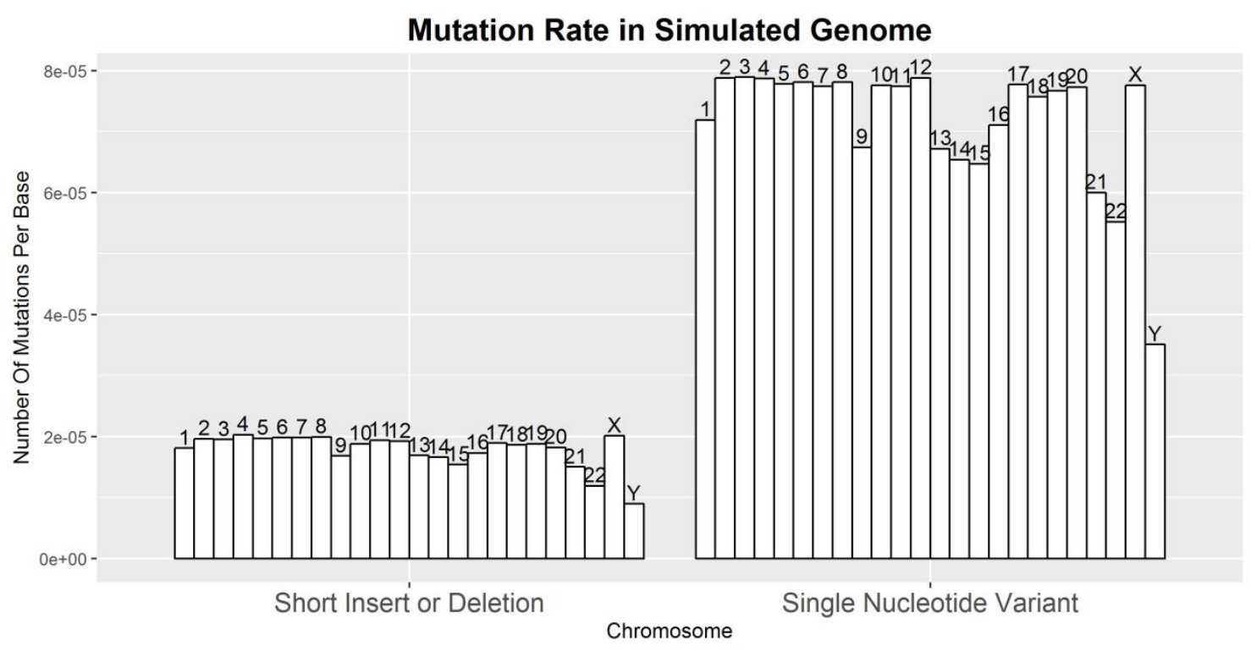
- Build a Bayesian network based on high confidence calls and functional annotations to rank mutations

Results and Discussion

Generation of synthetic sequencing datasets and features for neural network training

- Generation of **synthetic dataset** incorporating sequencer **error rates and profiles** from published data



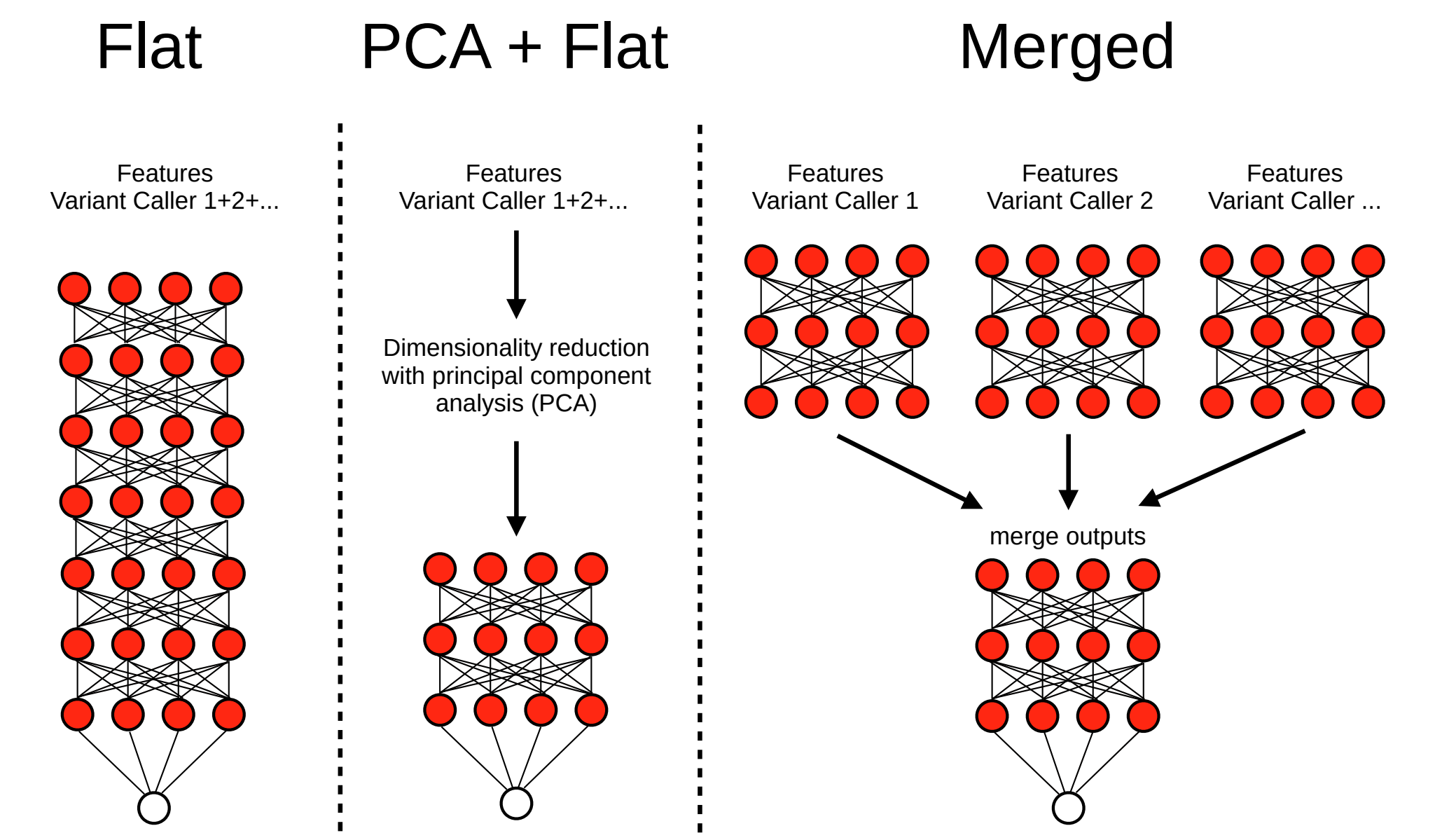


- **Engineering** of **features** for neural network training

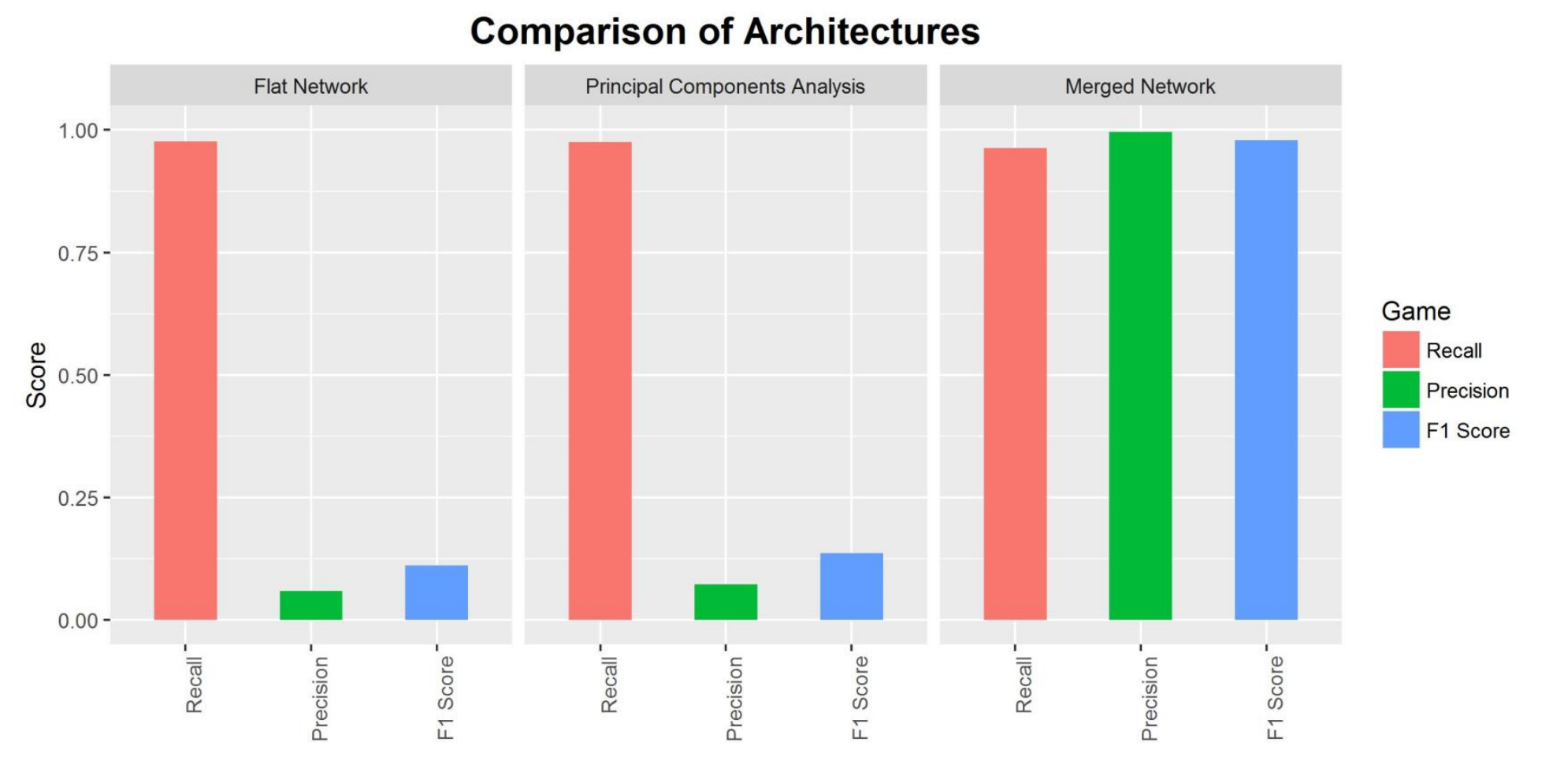
	Read Depth	Allele Count	Mapping Quality	...	Entropy
FreeBayes	+		+		+
GATK HC	+	+	+		+
GATK UG	+	+	+		+
Pindel	+				+
Samtools	+		+		+

Optimization of neural network architecture for variant calling

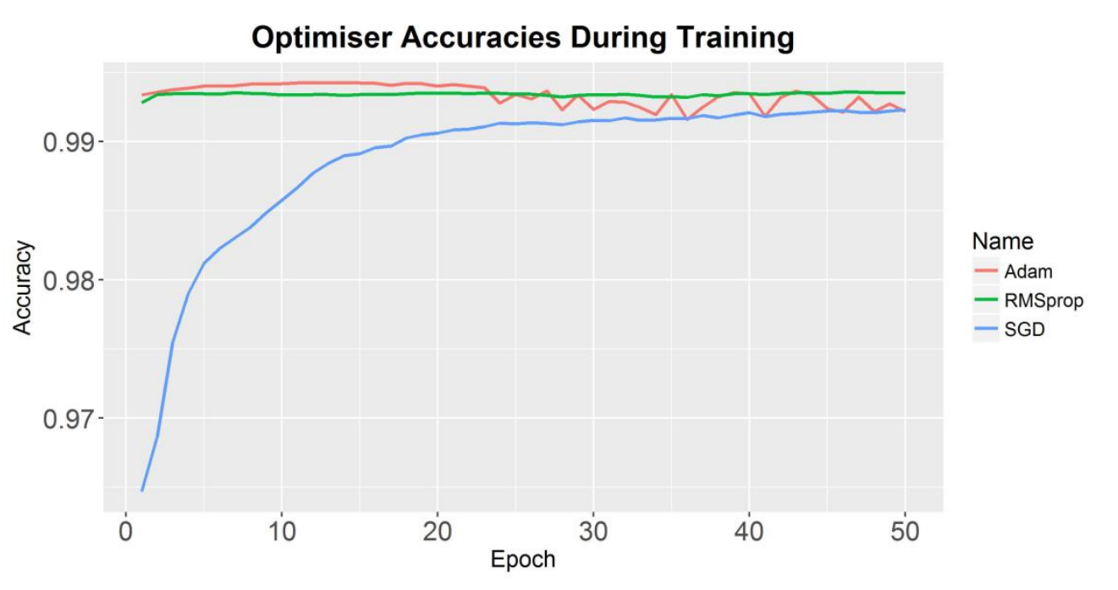
A. Neural network architecture


Flat
PCA + Flat
Merged
Features Variant Caller 1+2+...
Features Variant Caller 1+2+...
Dimensionality reduction with principal component analysis (PCA)
Features Variant Caller 1
Features Variant Caller 2
Features Variant Caller ...
merge outputs

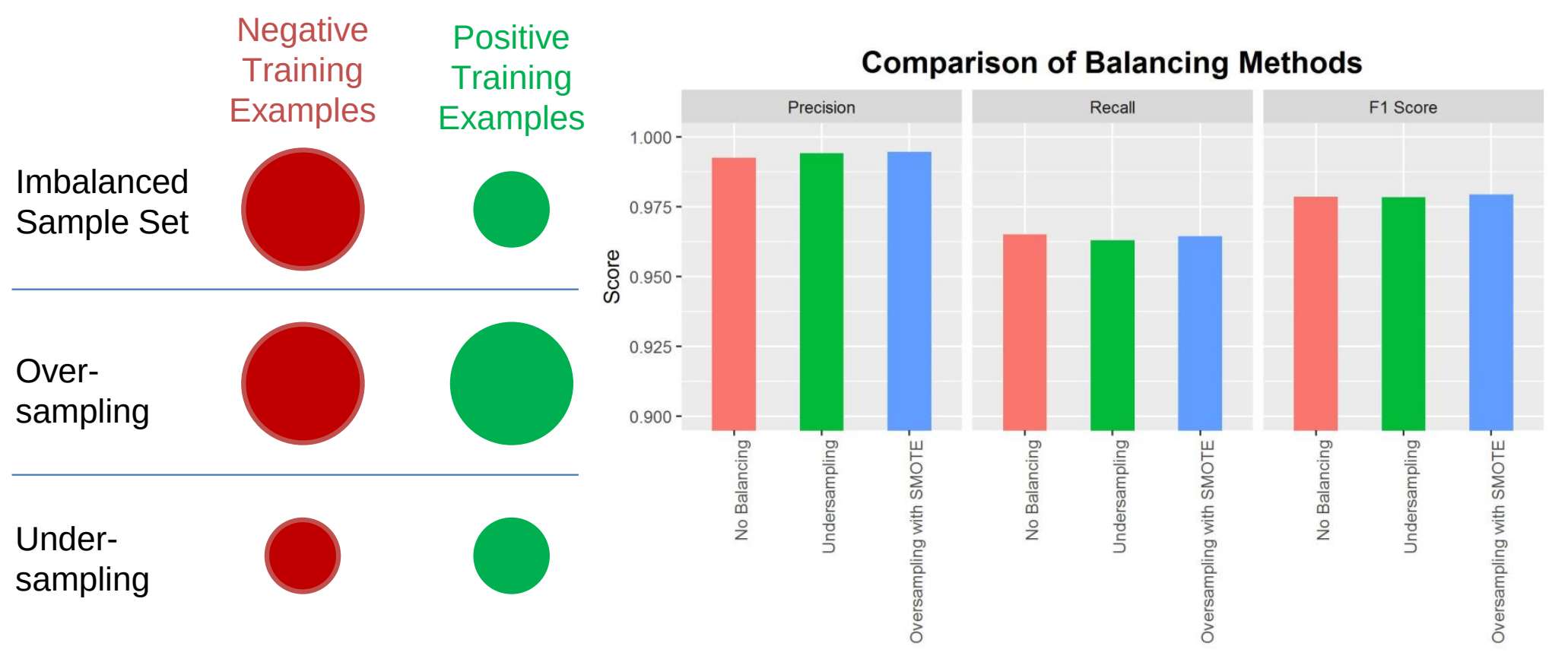
- Three network architectures were assessed for accuracy in identifying high confidence calls



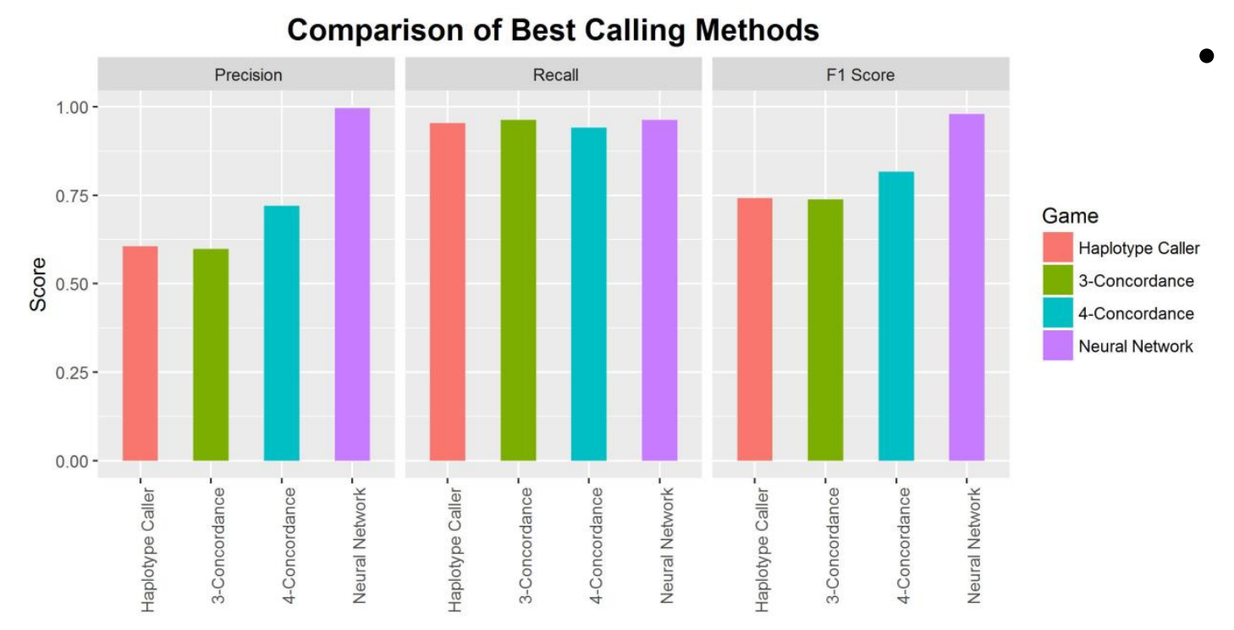
B. Network tuning

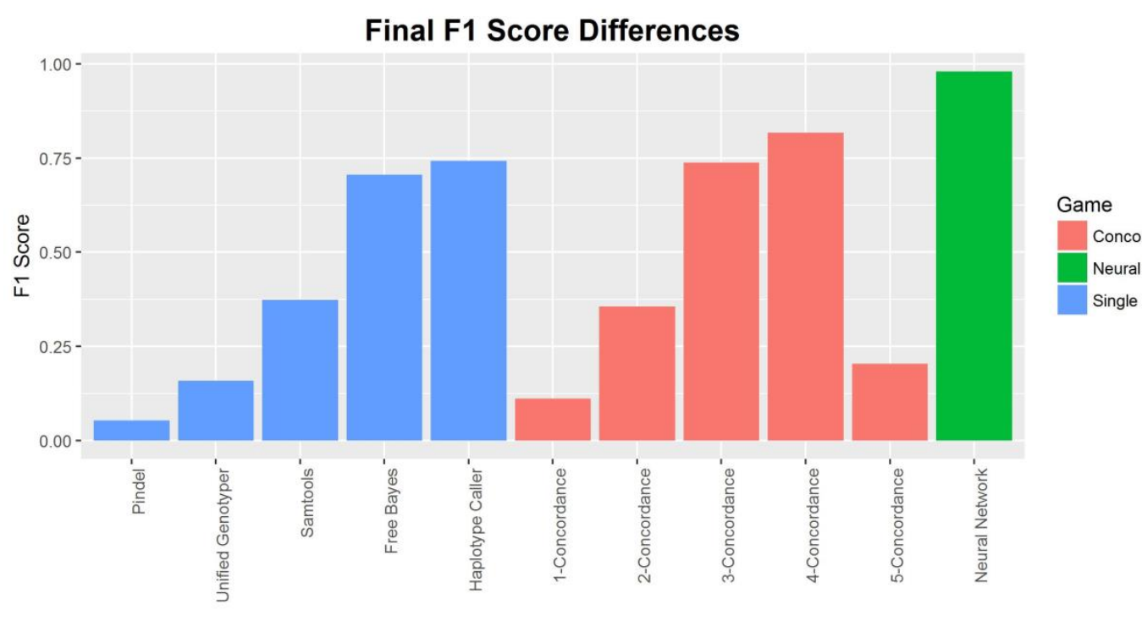


- Tuning of network e.g. (i) testing different **learning algorithms** (ii) **balancing** the number of **positive and negative training samples**



C. Benchmarking



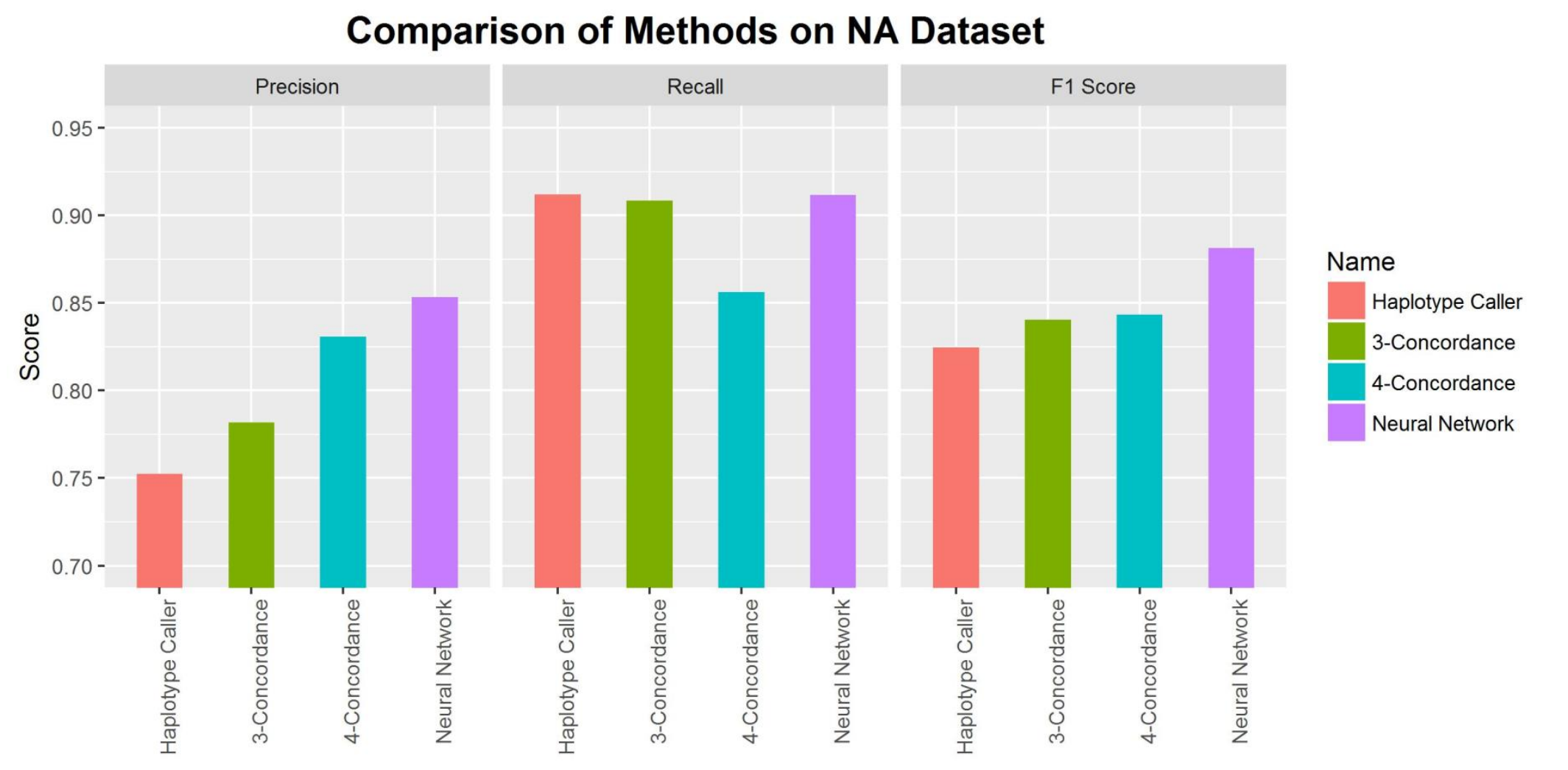


- The neural network outperformed single and concordant based variant callers in terms of precision and overall F1 score

- This indicates that the network is able to learn from multiple features to identify high confidence variants

Validation of deep learning network with benchmark human genome reference

- Deep learning network was validated with the NA12878 reference genome dataset. This dataset contains high confidence calls of SNVs and indels, obtained using orthogonal sequencing methods.

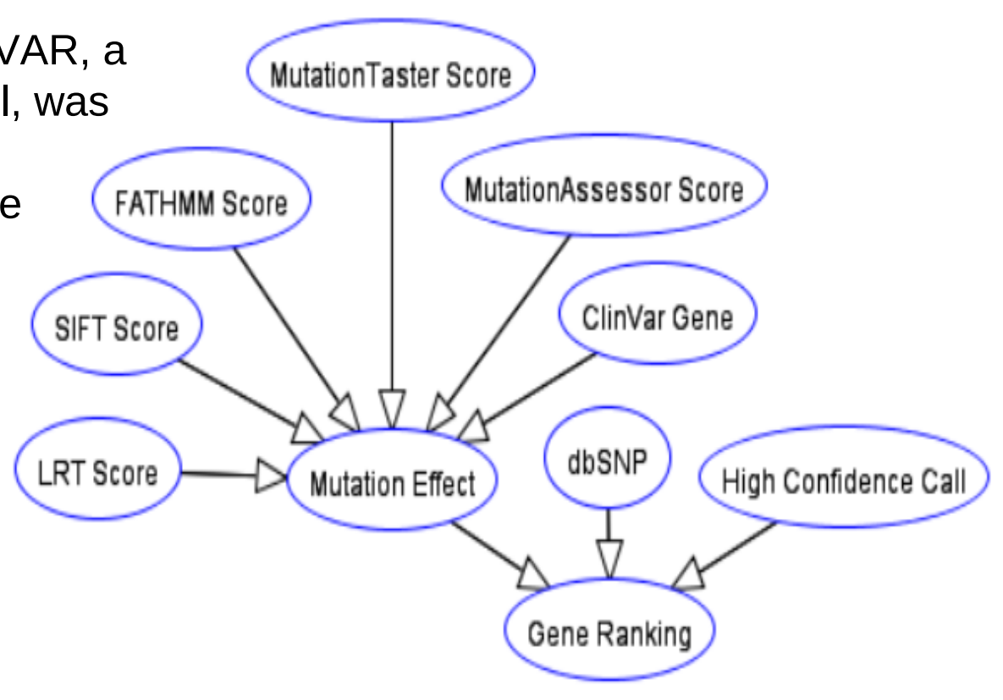


Bayesian network for gene prioritization

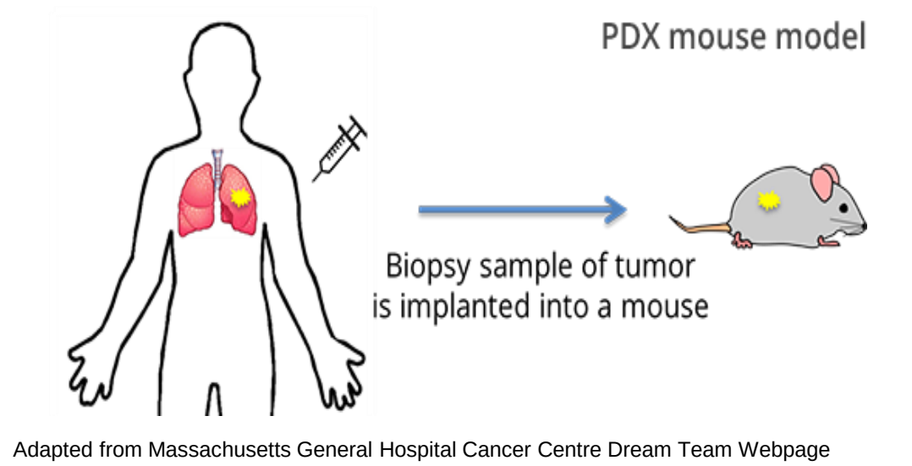
Annotating and **prioritizing** functional **mutations**

Supporting **medical diagnosis** and **treatment**

- Features from ANNOVAR, a variant annotation tool, was used as probabilistic variables to update the Bayesian network



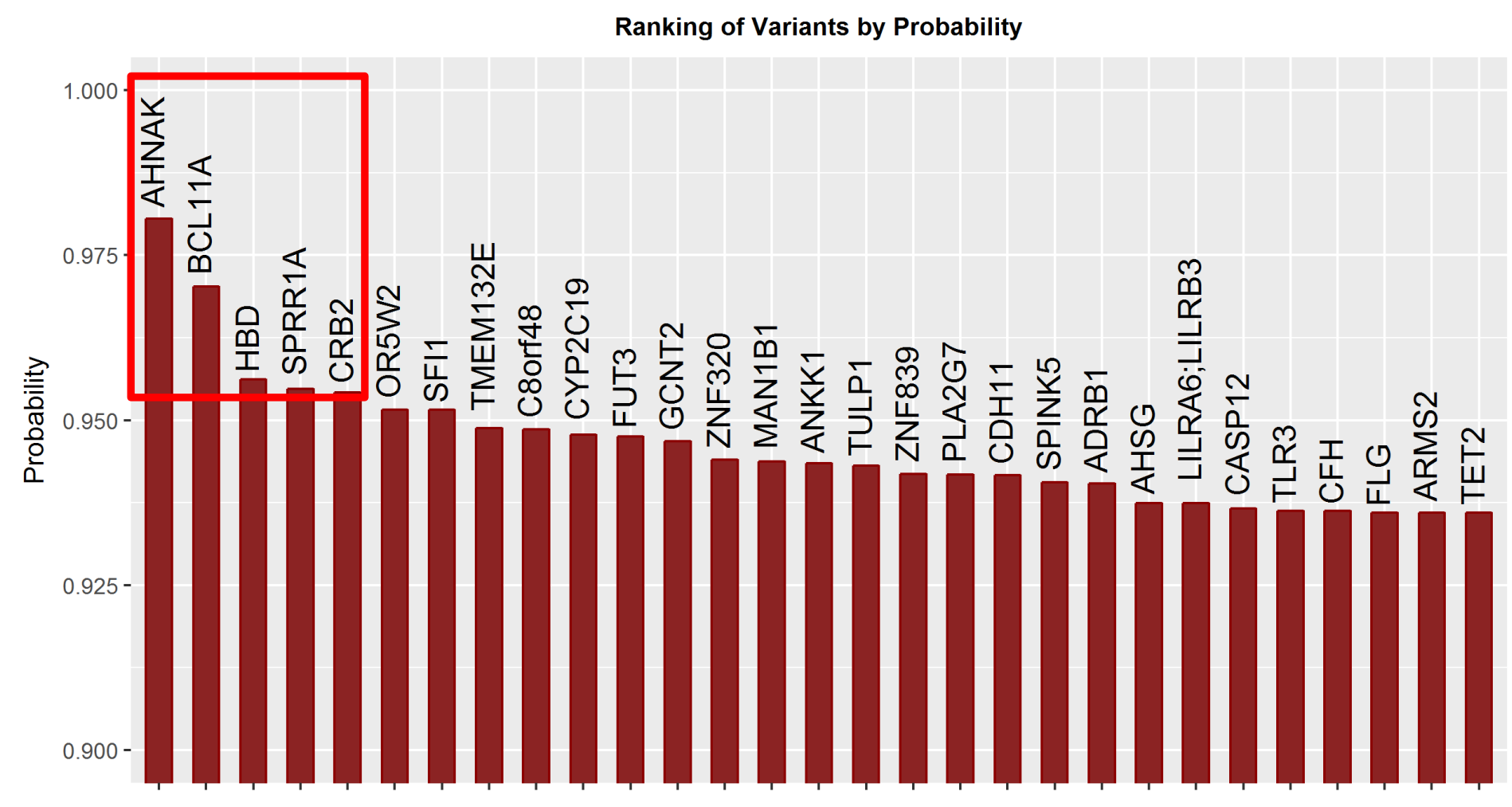
Integrated deep learning and Bayesian analysis of cancer genes in a PDX model of lymphoma



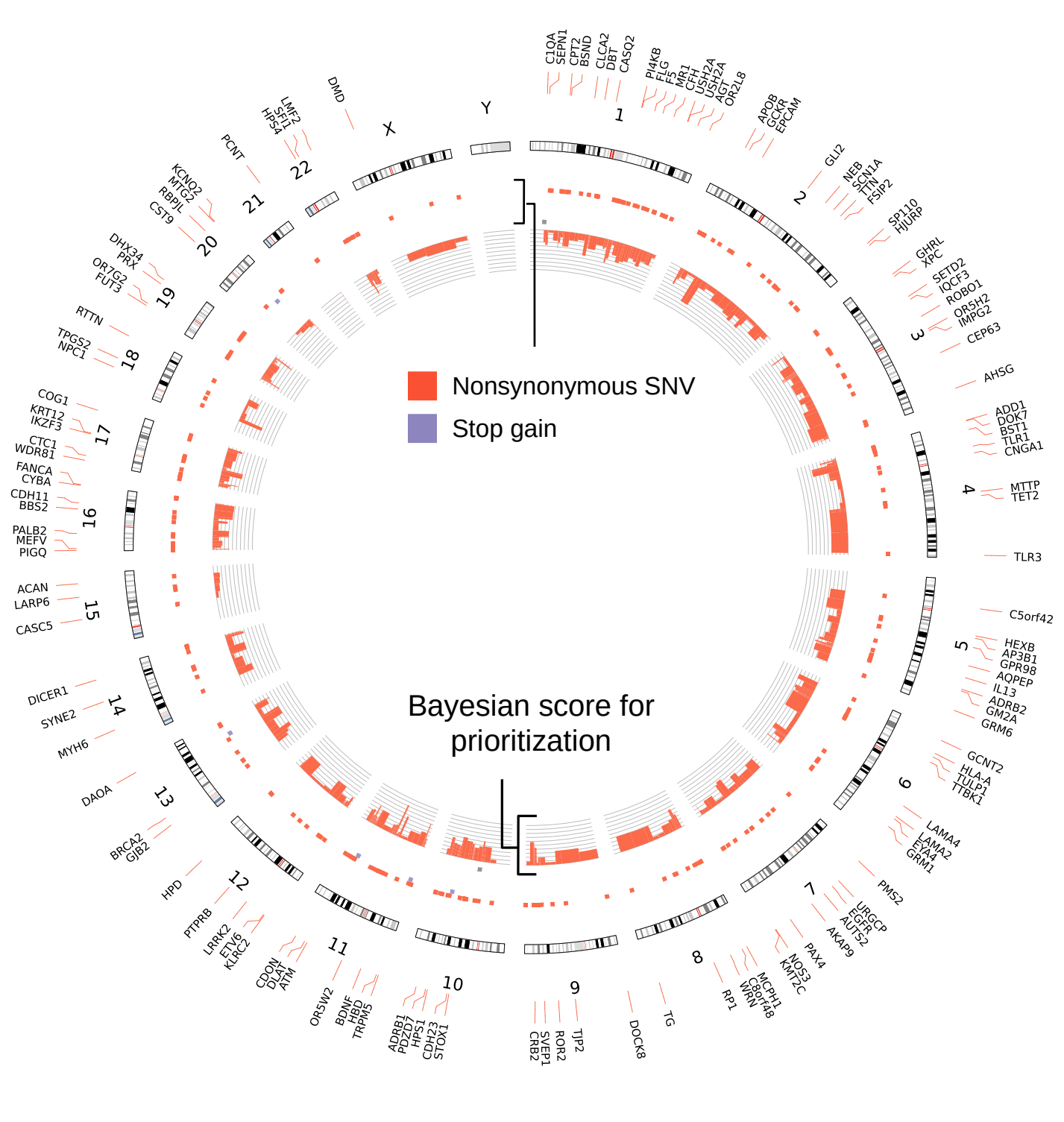
Adapted from Massachusetts General Hospital Cancer Centre Dream Team Webpage
http://www.lungcancerdreamteam.org/progress-in-kras-research/kras-mutant-mouse-models

- In a patient derived xenograft (PDX) model, tumours are implanted into immunocompromised mice and grown as xenografts. This allows the tumours to be studied in their *in vivo* environment

- We sequenced a diffuse large B-cell lymphoma (DLBCL) xenograft and analyzed it using the optimized neural network prior to gene prioritization by Bayesian inference



Gene	Full Name	Known Involvement in Lymphoma or Cancer	Evidence	Mutation Location	Predicted Mutation Type
AHNAK	Neuroblast Differentiation-Associated Protein (Desmoyokin)	• 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	• 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	• 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)	• 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	• **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**



- The Circos plot shows the top 300 genes that are mutated in the DLBCL PDX sample, together with their mutation type and Bayesian score

- The Bayesian method aggregates the probabilistic information from the confidence of variant calls and other functional predictions from gene annotation

- The integrated approach provides a **systematic way of interrogating genes** based on a probabilistic framework

Summary

1. We verified the **usage of deep learning networks** to **predict high confidence variant calls** in both **simulated** and **real datasets**
2. We showed that the Bayesian network is able to identify **highly relevant** genes in diffuse large B-cell lymphoma (DLBCL) that will be useful for clinical analysis

Future Directions

1. Adapting the neural network to **other datasets**, and verification of results with **Sanger sequencing**
2. Extending the Bayesian network to include **druggable datasets** to enable the prioritisation of genes that are druggable to aid in clinical decision making