deepSNV
Jason Dean

Introduction

Mendelian disorders and somatic mosaicism form the genomic foundation of diseases ranging from cystic fibrosis to cancer (Tomasetti, Li, & Vogelstein, 2017). Identifying single nucleotide variants (SNVs) is one of the first steps in discovering and understanding genetic links between clinical outcomes related to disease, progression, and drug resistance. SNVs, or single nucleotide 'mismatches' in the genome of an individual, can be identified in a high throughput manner by next generation sequencing (NGS) of genomic DNA. From a technical perspective, SNV calling is confounded by two primary factors:

1) The error rate of NGS ranges from 0.1-10%
2) Somatic SNVs can be present at levels as low as 0.1%

Thus, the main challenge in SNV assignment is separating the signal from a true SNV from the inherent technical noise in the measurement assay. Tremendous progress has been made in this area with the development of complicated, hand-tuned statistical models (do Valle et al., 2016). These models, although certainly impressive in their performance, are laborious to develop and are still error prone. In this work I chose to approach the SNV calling problem from a deep learning perspective. Deep learning algorithms have demonstrated near human performance on tasks like image recognition (Krizhevsky, Sutskever, & Geoffrey E., 2012), however applications in the context of genomics are still being developed and exciting progress has been made in the areas of SNV calling (Poplin et al., 2016) and automated analysis of microscopy images (Kraus et al., 2017). Here I adapt a convoluted neural network (CNN), an algorithm traditionally used in image classification, to SNV calling.

Methods

Access to accurately labeled data is critical for machine learning model building, and this study required human genome NGS data with accurately annotated SNVs. Here I chose to use sample NA12878 (50X coverage, 2X100bp reads, PCR free preparation) from the Illumina Platinum Genome data set (Eberle et al., 2017). This dataset contains over 4 million curated SNVs that were identified and confirmed via both deep sequencing of a large family pedigree and multiple publically available SNV calling pipelines.

Sample NA12878 aligned to reference genome hg19 was downloaded as a bam file from: https://www.illumina.com/platinumgenomes.html and indexed using samtools (Li et al., 2009). A VCF file containing 4,049,513 verified SNVs was used for identification of positive genomic sites for training. Data was formatted for input to a (CNN) as shown in Figure 1.

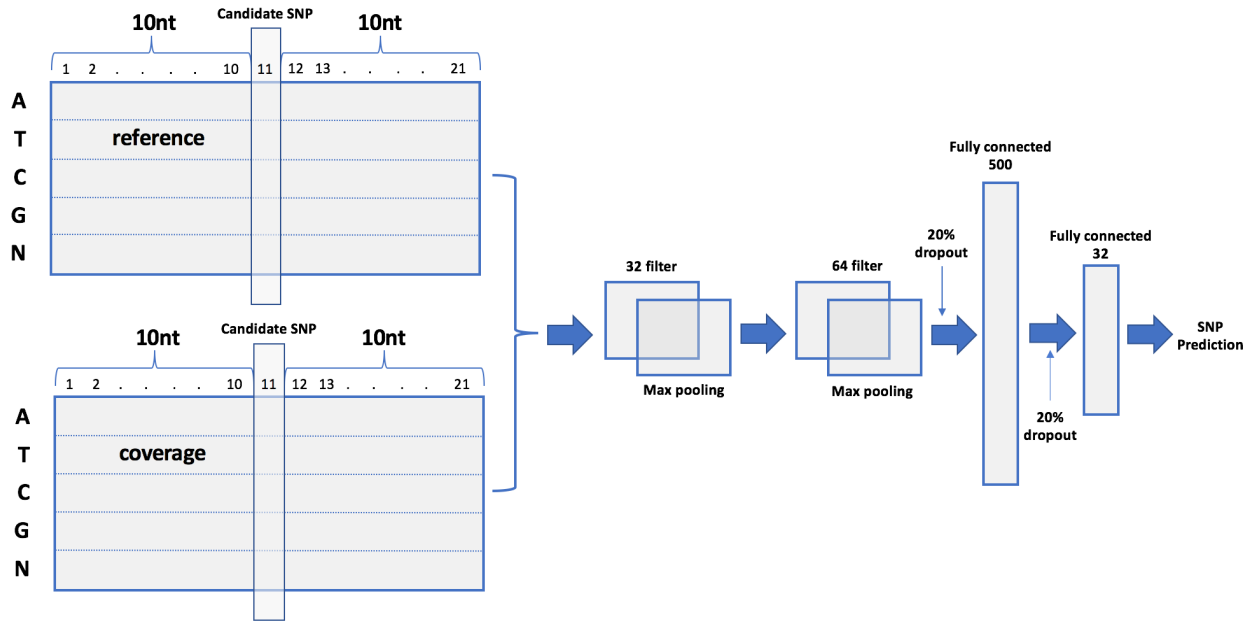

Figure 1. Overview of data format and CNN model structure.

CNN architectures are typically employed for image recognition, and here we investigated if genomic data formatted into the multiple channel matrix format of images is amenable to machine learning. First, 100,000 positive SNV and 100,000 non-SNV sites were selected and combined for model development. 70% of the data was randomly chosen for a training set and 30% was set aside for testing. Each genomic coordinate in the data set was formatted in a dual channel format as shown in Figure 1. One channel contains information pertaining to the reference genome sequence at the coordinate of interest along with 10bp up and downstream. Each element of the reference 5X21 matrix is set to zero with the exception of the row/column coordinate corresponding to the nucleotide at a given coordinate, which is set to one. The second channel contains information related to the coverage of the NGS data around the same genomic region as channel one. While channel one contains data describing the reference sequence, channel two contains the read count in at each genomic coordinate for each nucleotide. Outputs were one hot encoded as either one for a SNV or zero for a non-SNV site.

Once the data was formatted as described in the strategy above it was fed to the CNN architecture shown in Figure 1 for training. In particular, two CNN layers of 32 and 64 filters, each followed by dimension reduction via max pooling, were used to extract features from the two channels of genomic information. The features from the max pooling following the 64 filter layer were then fed to a two layer fully connected network consisting of 500 and 32 nodes, respectively, with 20% dropout between each layer. Model weights were updated by back propagation via minimization of the cross-entropy loss function with a batch size of 16 and 50 epochs.

Results

As shown in Figure 2, I found that the cross-entropy loss and accuracy reached steady values after approximatively 15 epochs of training, suggesting that the 50 epochs used for training may have resulted in overfitting.

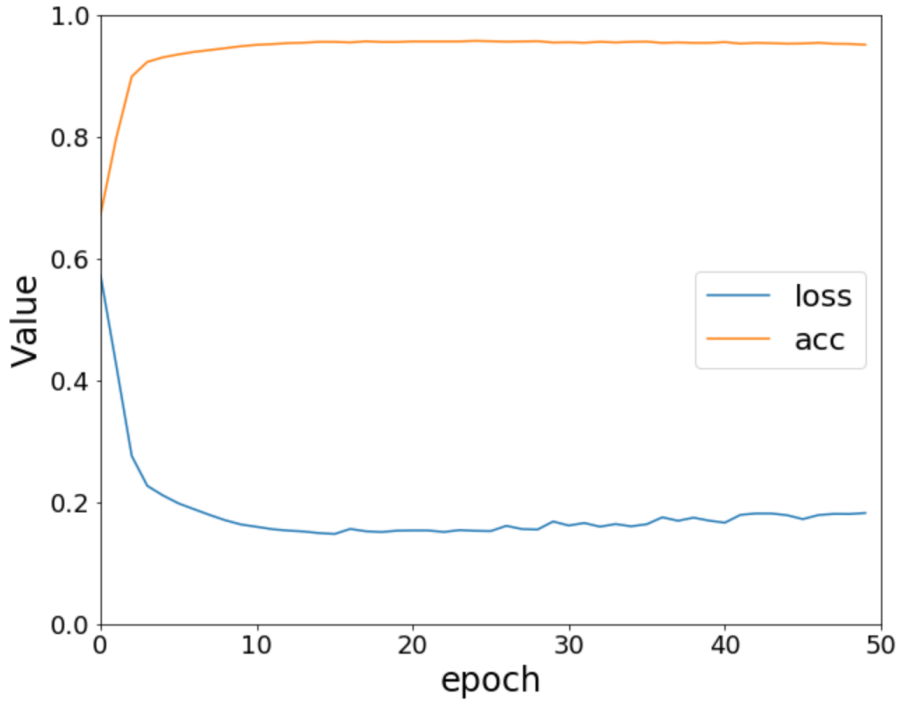

Figure 2. Cross-entropy loss and accuracy over 50 epochs.

The trained model was then used to make predictions on the 60,000 held out test samples. Remarkably, I found that the model achieved 95.3% accuracy on the test data along with 99.6% precision, and 90.9 % recall (Figure 3).

	Predicted Non-SNV	Predicted SNV
Actual Non-SNV	29718	90
Actual SNV	2727	27465

Figure 3. Confusion matrix for model predictions on test data.

Conclusions

In conclusion, I found that the CNN model performed strikingly well with regards to multiple performance metrics despite minimal model tuning and iterative updating. I found that NGS data, when formatted into the multi-channel matrix format of digital images, is amenable to deep learning architectures such as CNN. Since the model captures local feature information

surrounding each putative SNV coordinate this model may extend to NGS data generated from other organisms. Finally, this project demonstrates the utility of using deep learning to extract information from genomic NGS data.

References

do Valle, Í. F., Giampieri, E., Simonetti, G., Padella, A., Manfrini, M., Ferrari, A., … Castellani, G. (2016). Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinformatics*, *17*. https://doi.org/10.1186/s12859-016-1190-7

Eberle, M. A., Fritzilas, E., Krusche, P., Källberg, M., Moore, B. L., Bekritsky, M. A., … Bentley, D. R. (2017). A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Research*, *27*(1), 157–164. https://doi.org/10.1101/gr.210500.116

Kraus, O. Z., Grys, B. T., Ba, J., Chong, Y., Frey, B. J., Boone, C., & Andrews, B. J. (2017). Automated analysis of high-content microscopy data with deep learning. *Molecular Systems Biology*, *13*(4), 924. https://doi.org/10.15252/msb.20177551

Krizhevsky, A., Sutskever, I., & Geoffrey E., H. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems 25 (NIPS2012)*, 1–9. https://doi.org/10.1109/5.726791

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Poplin, R., Newburger, D., Dijamco, J., Nguyen, N., Loy, D., Gross, S. S., … DePristo, M. A. (2016). Creating a universal SNP and small indel variant caller with deep neural networks. *bioRxiv*, 92890. https://doi.org/10.1101/092890

Tomasetti, C., Li, L., & Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science*, *355*(6331), 1330–1334. https://doi.org/10.1126/science.aaf9011