PB3470: Screening Copy Number Variation with an Autoencoder

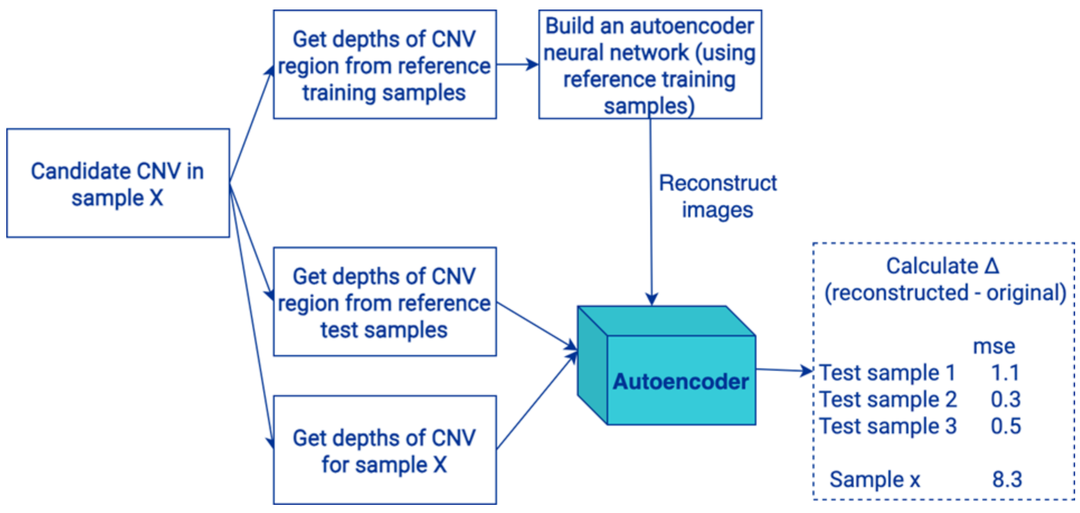
Authors: Kate Im, Pauline Ng, Premal Shah, Akash Kumar

BACKGROUND

- Copy number variants (CNVs) are known causes of Mendelian diseases such as 22q11.2 microdeletion syndrome and have also been associated with more complex phenotypes such as autism
- CNV calling from short-read whole genome sequencing (WGS) has higher resolution than other methodologies, but can suffer from reference assembly gaps and repetitive sequence context
- Visual inspection of many samples can help identify false positives, but is time-consuming
- An autoencoder can be used for anomaly detection to screen out false positives and save time

METHODS

Figure 1. Schematic of how the autoencoder is built for each CNV



- Depth profiles for regions spanning a candidate CNV were obtained from > 20 reference samples known not to have large CNVs
- An autoencoder was trained on reference samples
- We tested whether the autoencoder's reconstruction error from a clinical sample was similar to reference samples (no CNV) or significantly different (CNV present)
- We evaluated the autoencoder on clinical samples with known large pathogenic CNVs (≥100kb)
- We extended evaluation to smaller (≥1kb), rare (frequency <0.01) CNVs from Genome in a Bottle's HG002

An autoencoder is a highly sensitive, cost-saving method for filtering large CNVs

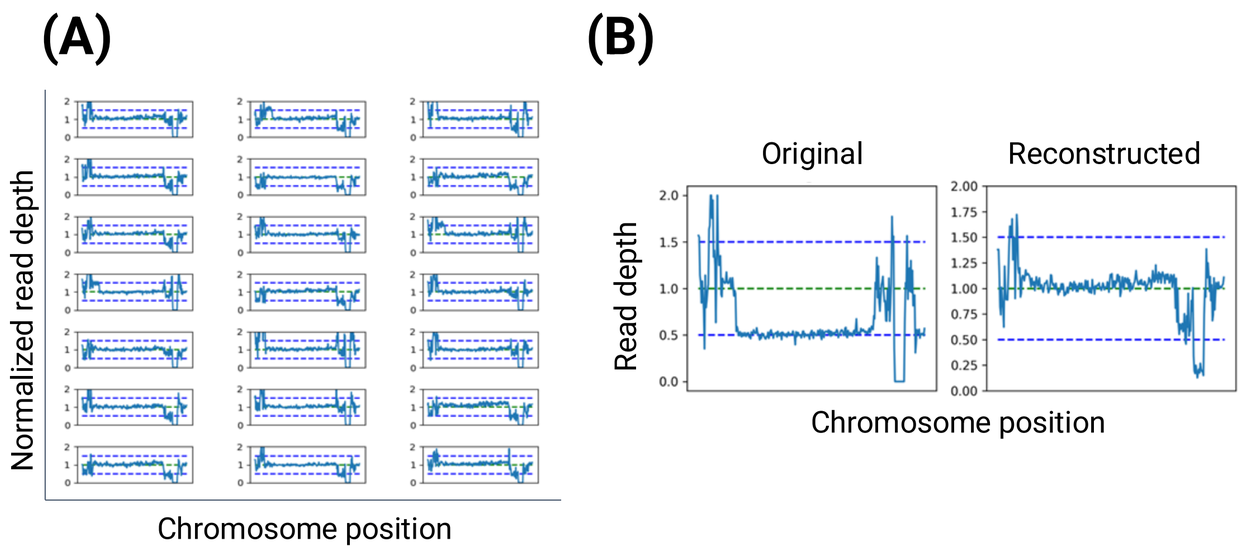
RESULTS

Table 1. Sensitivity and specificity across two methods of CNV classification

Method	Sensitivity of Clinical Samples	Sensitivity of Rare CNVs	Specificity of True Negative CNVs
Autoencoder	100% (22/22)	99.85% (667/668)	87.5% (385/440)
Z-score	100% (22/22)	99.4% (664/668)	88.2% (388/440)

- We observed improved sensitivity with minimal impact on specificity with the autoencoder compared to using z-scores based on average read depth
- The autoencoder showed sensitivity of **100%** (22/22) and **99.85%** (667/668) on clinical samples and HG002, respectively
- Specificity for the autoencoder and z-score approach is 87.5% (385/440) and 88.2% (388/440), respectively

Figure 2. Example of reconstructed image on chromosome 22. (A) Depth plots of reference samples. (B) Autoencoder reconstruction for test sample



CONCLUSIONS AND FUTURE DIRECTIONS

- This neural network approach is a sensitive method to automate an otherwise time intensive component of CNV reporting, reducing the cost of interpretation and confirmation
- The autoencoder has been applied to deletions and duplications. We need to extend to other structural variants such as inversions and translocations

Correspondence: Kate Im, kate@myome.com