5-2018

Deep Learning on Abnormal Chromosome Segments: An Intelligent Copy Number Variants Detection System Design

Stephen Shaoyi Liao
City University of Hong Kong, issliao@cityu.edu.hk

Yuan Chen
City of University of Hong Kong, ychen493-c@my.cityu.edu.hk

Xiaobing Ma
Darui Reproductive Technology Co, max@daruisz.com

Puxi Wang
City of University of Hong Kong, puxiwang2-c@my.cityu.edu.hk

Yan Liu
City of University of Hong Kong, yliu627-c@my.cityu.edu.hk

Deep Learning on Abnormal Chromosome Segments: An Intelligent Copy Number Variants Detection System Design

Stephen Shaoyi LIAO
City University of Hong Kong
issliao@cityu.edu.hk

Puxi Wang
City of University of Hong Kong
puxiwang2-c@my.cityu.edu.hk

Yuan Chen
City of University of Hong Kong
ychen493-c@my.cityu.edu.hk

Yan Liu
City of University of Hong Kong
yliu627-c@my.cityu.edu.hk

Xiaobing Ma
Darui Reproductive Technology Co., LTD
max@daruisz.com

Abstract:

Gene testing emerged as a business in the last two decades, and the testing cost has been reduced from 100 million to 1000 dollars for the development of technologies. Preimplantation genetic screening (PGS) is a popular genetic profiling of embryos prior to implantation in gene testing. Copy number variants (CNVs) detection is a key task in PGS which still needs the manual operation and evaluation. At the same time, deep learning technology earns a booming development and wide application in recent years for its strong computing and learning capability. This research redesigns the PGS workflow with the intelligent CNVs detection system, and proposes the corresponding system framework. Deep learning is selected as the proper technology in the system design for CNVs detection, which also fit the task of denoising. The evaluation is conducted on simulation dataset with high accuracy and low time cost, which may achieve the requirements of clinical application and reduce the workload of bioinformatics experts. Moreover, the redesigned process and proposed framework may enlighten the intelligent system design for gene testing in following work, and provide a guidance of deep learning application in AI healthcare.

Keywords:

Copy number variants detection, preimplantation genetic screening, deep learning, AI healthcare

1. Introduction

Human genome has a great significance in an individual's growth and development. Abnormal gene may result in several mutagenic diseases including autism, intellectual

disability, epilepsy, schizophrenia, obesity, and cancer. For example, the deletion of the APOBEC3 gene cluster would increase the probability of breast cancer. Trisomy 21, resulting in one kind of Down's syndrome, may lead to intellectual disability and stunting (Yu Zhenhua, 2016). As for auxiliary reproductive field, embryos with abnormal gene on chromosome aneuploidy which is closely related to genetic disease such as thalassemia should be eliminated in PGS. Furthermore, chromosomal ectopic could lead to abortion and multiple pregnancy. Researchers have always been seeking for superior methods to solve the problem on abnormal gene detection to improve PGS. As for abnormal gene, CNVs are common variants among human genome and it is reported that about 12% of the genome in human populations suffer from copy number change (Zhao, 2013).

Figure 1: PGS workflow

To support prenatal and postnatal care, and further public healthcare, an increasing number of biotechnology corporations participate in the business related with PGS. PGS, is an important means of auxiliary reproductive technology to detect CNVs so as to improve the pregnant rate and reduce abortion rate. CNVs detection is the last task of PGS technical process in the view of the biological information analysis. The first two steps are conducted in biology laboratory, the following two steps take place on sequencing platform and analytical software while the last step is carried out under the guidance of a statistical model which may be likely to be replaced with an artificial intelligence model. Next-generation sequencing (NGS) generate data for CNVs detection which is still a challenging issue owing to the complex process of PGS. Data produced from sequencing platform is inevitably biased (rates from 0.1 to 10%), the original data results closely depend on properties of the instrument, data processing tools and the genome sequence itself [3]. Lab Instrument noise and organism DNA (Deoxyribonucleic Acid) preference variants noise are recognized as two primary noise sources. Based on the noisy data, there are five strategies to detect CNVs, and the one based on read depth is widely employed (Zhao, 2013). Strategy based on read depth aims to handle read counts that is actually signal values detected from a sequencing platform, while many machine learning methods have been applied to this flow while the performance is hard to

match the requirement of clinical application. A bottleneck existed is the troublesome noise of the NGS data.

There are several challenges concerning NGS data noise: a. the unknown distribution of noise; b. the computing capability for Big Data analysis; c. availability of labelling data. Hence, it is an urgent issue to apply an advanced and effective technology to further figure out the problem of CNVs intelligent detection.

An artificial intelligent system may take advantages of its strength and thus make contribution to AI healthcare problem of abnormal gene. For AI healthcare issues of IS research, there are several areas branches: health IT, big data and predictive analytics and design science (Lin, 2017). In this research, we follow the design science paradigm, propose a novel framework to mitigate high-throughput sequencing data noise and solve abnormal chromosome segments detection task for the case of CNVs.

2. Related work

Hidden Markov model (HMM) and circular binary segmentation (CBS) are the top two methods for CNVs detection at this stage, however, as we mentioned, they hardly match the requirements of clinical application. Thus, the practitioners employ these two models in system design only for decision support, the manual audit is still necessary. What's more, statistical models except HMM and CBS like Mean Shift-Based (MSB), Shifting Level Model (SLM) and Expectation Maximization (EM) for CNVs detection task need the knowledge on prior distribution of the read depth. An assumption that a sequence of read segments (gain or loss) along each chromosome could be detected should be subjected to probability distribution functions. Thus, we need to know a good prior probability distribution and in this way their mathematical model may tend to be sensitive to data noise (Zhao, 2013).

One well-investigated bias in read depth-based methods is G+C(G is refered to Guanine and C is refered to Cytosine) content. To eliminate the influence of this noise, an effective approach is GC correction algorithms design such as locally weighted scatterplot smoothing (LOWESS) (Yoon, 2009). The other type of noise that affects CNVs detection is caused by the process of read alignment. And for this problem, baseline correction is widely applied and it aims to capture technical bias of a platform but not the actual CNVs of the samples. To achieve this goal, we must create a robust baseline for GC corrected samples (Li, 2012).

Although these correction approaches have contributed to noise mitigating, the problem cannot be solved thoroughly.

Yao, R., et al. (2017) evaluated three commonly used read-depth based CNVs detection tools based on NGS data but poor concordance of CNVs were observed (Yao, 2017). Although the popular algorithms JointSLM and EWT using high coverage data can achieve a true positive rate >0.8 with a tiny fraction of false positive events, clinic practice still need experts to audit (Magi, 2011).

Moreover, there are many software tools available for CNVs detection like PennCNV, QuantiSNP, and GenoCN. However, these tools have their various limitations and advantages. For instance, the tools adopt CBS or MSB algorithm may perform well when detecting CNVs whose size are over 50bp but perform worse when detecting CNVs whose size are less than 10bp.

Despite machine learning algorithms, researchers have done some work within deep learning technology. Wang, J., et al. made genomic deletion structural variants using Low coverage data based on CNN(Convolutional Neural Network) (Wang, 2017). The union of candidates can be firstly generated by four tools. Then they chose each sample's 49 dimensional features of five aspects to build up a feature matrix for subsequent network input. The proposed deep learning CNNdel model outperforms other CNVs calling tools in accuracy and sensitivity, especially in identifying CNVs whose deletion variants size surpass 1Mb. The other work is proposed by Google team. They present a deep convolutional neural network to call small indel variant. They trained their network with images of read pileups around putative variant sites and ground-truth genotype calls and finally proposed a new approach called DeepVariant which outperformed existing work (Poplin, 2017).

3. Methodology

As what we mentioned above, abnormal gene detection has a great importance on one's health which is a noticeable stream of AI healthcare. High throughout sequencing data and low coverage data may differ in some statistical methods fundamentally while nowadays NGS technology generating data is prevalent. And traditional work of CNVs detection is not satisfying with low sensitivity, high false positive rate and uncertain specificity. Under this circumstance, we carefully analyze weakness of the existed solutions, and find that deep

learning might make a breakthrough.

Deep learning is an innovative technology in artificial intelligence research areas recently, especially for solving some AI healthcare problems (Esses, 2017) (Pratt, 2016) (Paul, 2016). It has been widely used in speech enhancement, natural language processing, audio recognition, social network filtering, machine translation and bioinformatics on account of its strong computing ability to learn complex functions which challenges the typical machine learning methods. By utilizing deep learning, the proposed framework contributes to read depth strategy for seeking the solution of troublesome whole genome sequencing data noise problem and CNVs detection.

Our methodology is dedicated to improve the business process of gene testing from the perspective of data denoising and CNVs detection. The workflow is depicted in figure 2, which consists of five steps. The first step is data generation based on templates construction and noising. The second step is data preprocessing. We select appropriate technology to match the requirement of the task in the third step. The next step focuses on designing the denoising model. In the Final step, we continue to do data training and evaluation.

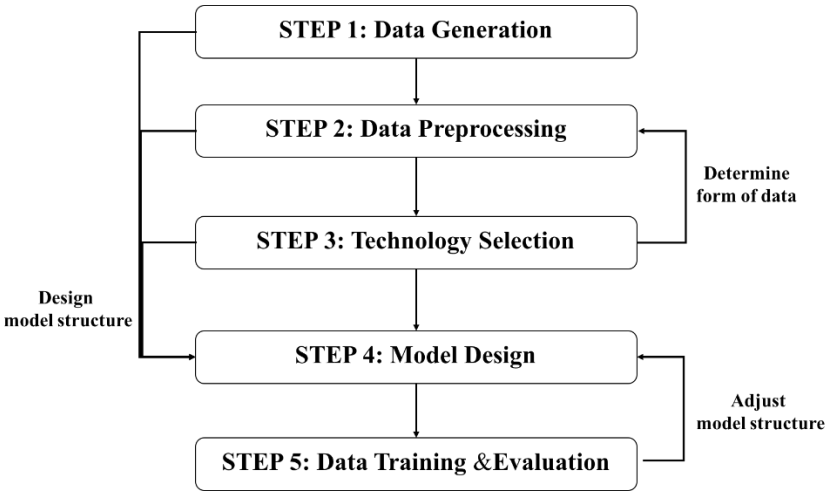

Figure 2: Intelligent CNVs detection system framework

3.1 Data generation

After single cell whole genome amplification and NGS technology, there are tens of thousands of base sequencing reads, then NGS reads are aligned to a human genome reference sequence and mapped on the genome. For CNVs detection, experts regard the count values of reads within windows as the signal value to judge whether the chromosomes are normal, repetitive or absent. We mainly analyze duplicates of 20, i.e., single cell is amplified

20 times. Due to a variety of reasons, these count values of reads are noisy even after GC correction and baseline correction. Data generation is the foundation for following work, our research team has established the cooperation with a biotechnology enterprise. With support of bioinformatics experts, we know what count values of the undetected, haploid, diploid, triploid region, tetraploid after GC correction and baseline correction are, that are actually clean data. Based on the clean data and their extended forms, we generate a set of noisy data. However, the distribution and its characteristics of noise caused by each step of processing are unknown, we choose the plan to add noise to the data randomly.

3.2 Data pre-processing

Generated data based on GC correction and baseline correction is presented as the counts of each bin in gene sequence and difficult to be processed directly by those techniques which are designed for image data processing. Therefore, we transfer the form of the generated data by one-hot approach (Hollaar, 1982). The x direction of images represents 40 bins, in which each bin consists of 0.1Mb, the y direction represents the count values of 40 bins. The simulated data preprocessed is depicted in figure 3, which consists of clean data and noisy data. The above is noisy data and the below is clean data.

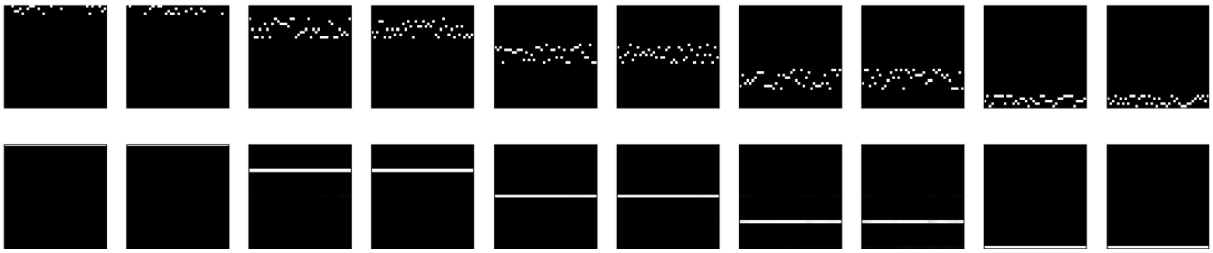

Figure 3: The simulated data preprocessed

3.3 Technology selection

Large noise is the biggest obstacle for CNVs, which seriously influence experts' judgement on chromosome abnormalities. Therefore, denoising after corrections is not only helpful to support experts' decision, but also lay a solid foundation for further intelligent CNVs detection system design.

At present, there are some denoising methods such as smoothing filter and wavelet denoising work on the related issues (Isard, 1998) (Xie, 2002). While the structure or characters of these methods limit the improvement of denosing. For example, Gaussian low pass filter (Carrato, 1996) is a linear smoothing filter with a Gaussian function, which is very effective for

removing the noise of normal distribution, but not for other types of noise. The methods which need the prior knowledge on noise distribution can hardly solve the real world problem directly if the context is complex with diverse noise sources.

The denoising autoencoder (DA) (Vincent, 2008) is one of unsupervised neural networks that is trained to input the corrupted data and attempt to output the original, uncorrupted data. As one of the deep learning model, DA is good at automatically extract some important potential features. The significant advantage is that no human intervention is required, which greatly reduces people's work. Moreover, it has a high robustness and a wide range of applications.

3.4 Model design

Considering the characteristics of NGS data denoising and CNVs detection, we should design a novel model for the workflow reconstruction to gain higher efficiency. Before we design the model, we should review the PGS workflow at first, it can be seen in figure 1. CNVs analysis and detection is the last task in this flow. To better improve the work on this task, in this case, the input should be the NGS data, and by the major IT artifact of our research, the intelligent CNVs detection system, PGS can acquire CNVs as the output automatically. The following figure reflect the reconstructed workflow of PGS.

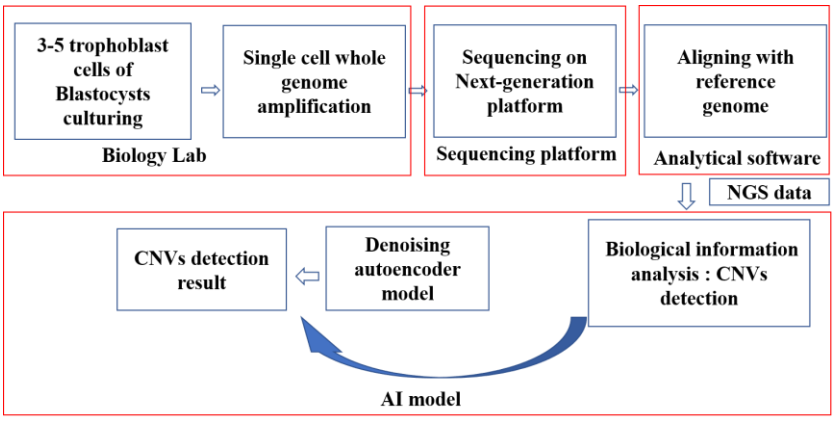

Figure 4: Redesigned workflow of PGS

We use stacked autoencoder as the base structure including input layer, output layer and three hidden layers, which consist of encoder layer and decoder layer. The activation function for all encoder and decoder layers is rectified linear unit (ReLU): $f(x) = \max(0, x)$. And we use the sigmoid function as output layer's activation function: $\sigma(x) = \frac{1}{1+e^{-x}}$. For the learning rate, we not apply a fixed one but variable one through Adadelta optimizer. The activation function, loss function and adaptive learning rate method we used can effectively avoid the problem of

learning slowing down, which accelerates the convergence of stochastic gradient descent (SGD)and shortens our training time. As figure 5 shows, denoising autoencoder is determined after a lot of repeated experimentations, which can provide the best denoising performance.

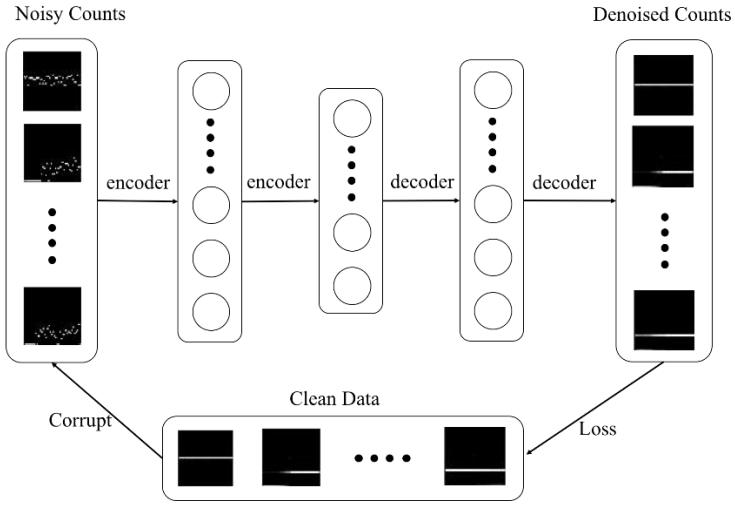

Figure 5: Model Structure

By the design of this model, the PGS workflow is able to achieve the efficiency on three dimensions: service quality, time and flexibility. It can improve the accuracy of CNVs detection more quickly, so as to reduce workload of relevant practioners.

3.5 Data training and Evaluation

We input our preprocessed data into denoising autoencoder model, observe the losses and accuracies of training data and validation data, and adjust super-parameters reasonably according to their variation trend and comparison to avoid overfitting. Finally, we use validation data to examine denoising performance of our denoising autoencoder.

On account of the excellent performance of denoising, the final results we got are clustered around one of the five regions (i.e., Undetected region, Haploid region , Diploid region, Triploid region, Tetraploid region). Moreover, we select 40 bins of the original data for every parent, that is 4Mb base pairs of chromosome. It is within the scope that experts are likely to detect CNVs. Consequently, the denoised results can be directly used to detect CNVs. Based on the evaluation, the accuracy of our intelligent system for testing on each bin (100kb) is higher than 95%, the sensitivity is also higher than 95%, and the testing time cost per patient sample is no more than 10 seconds. Compare with the former works on this problem such as PennCNV which employed the HMM with the sensitivity lower than 90% (Wang. K, 2007), and the one adopted CBS with sensitivity lower than 90% (Zare F, 2017), our intelligent

system design undoubtedly has much more clinical significance.

4. Conclusion and future work

We redesign the workflow of PGS with the intelligent system in this research. By utilizing the deep learning technology, the CNVs detection can be conducted intelligently and automatically, which reduce the workload of bioinformatics experts. PGS is improved by this novel approach on both the dimensions of time cost, service flexibility, service quality. What is more, the proposed framework may give some insights to the following research on AI healthcare system design.

As for the future work, we strive to improve the system by training with more real data, to employ more evaluation index such as ROC, F-score, etc. as the comparison organized in Andrew E. Dellinger' research (Dellinger A E, 2010), and to enrich the theoretical foundations to support the system framework design. Meanwhile, the analysis of characteristics of noise in PGS can better help us to improve its service quality, and better understand each task of PGS which provide more information for technology selection and model design. Furthermore, we hope to contribute to the business of gene testing by research on new topics such as the redesign of preimplantation genetic diagnosis (PGD), and other AI healthcare areas.

Acknowledgements

This work was supported by the development grant from Shen Science ,Technology and Innovation Committee (grant number JCYJ20160229165300897).

References

Carrato, S., et al. (1996). A simple edge-sensitive image interpolation filter. *Image Processing, 1996. Proceedings., International Conference on, IEEE.*

Dellinger A E, Saw S M, Goh L K, et al. (2010) Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic acids research*, 38(9): e105-e105.

Esses, S. J., et al. (2017). Automated image quality evaluation of T2-weighted liver MRI utilizing deep learning architecture. *Journal of Magnetic Resonance Imaging.*

Hollaar, L. A. (1982). Direct implementation of asynchronous control units. *IEEE Transactions on Computers* 12(C-31): 1133-1141.

Isard, M. and A. Blake (1998). A smoothing filter for condensation. *Computer Vision—ECCV*'98: 767-781.

Li, J., et al. (2012). CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 28(10): 1307-1313.

Lin, Y.-K., et al. (2017). Healthcare predictive analytics for risk profiling in chronic care: a Bayesian multitask learning approach. *MIS Quarterly* 41(2).

Magi, A., et al. (2011). Read count approach for DNA copy number variants detection. *Bioinformatics* 28(4): 470-478.

Paul, R., et al. (2016). Combining deep neural network and traditional image features to improve survival prediction accuracy for lung cancer patients from diagnostic CT. *Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on, IEEE.*

Poplin, R., et al. (2017). Creating a universal SNP and small indel variant caller with deep neural networks. *BioRxiv*: 092890.

Pratt, H., et al. (2016). Convolutional neural networks for diabetic retinopathy. *Procedia Computer Science* 90: 200-205.

Vincent, P., et al. (2008). Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th international conference on Machine learning, ACM.*

Wang, J., et al. (2017). CNNdel: Calling Structural Variations on Low Coverage Data Based on Convolutional Neural Networks. *BioMed Research International* 2017.

Wang K, Li M, Hadley D, et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome research*, 17(11): 1665-1674.

Xie, H., et al. (2002). SAR speckle reduction using wavelet denoising and Markov random field modeling. *IEEE Transactions on geoscience and remote sensing* 40(10): 2196-2212.

Yao, R., et al. (2017). Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Molecular Cytogenetics* 10(1): 30.

Yoon, S., et al. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome research* 19(9): 1586-1592.

Yu Zhenhua. (2016). Copy number variant of tumor genome detection algorithm based on

Next-generation sequencing technology. *University of Science and Technology of China*: 2016

Zare F, Ansari S, Najarian K, et al. (2017) Noise cancellation for robust copy number variation detection using next generation sequencing data[C]//Bioinformatics and Biomedicine (BIBM), *2017 IEEE International Conference on. IEEE*: 230-236.

Zhao, M., et al. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14(11): S1.