Reconstruction of high read-depth signals from low-depth whole genome sequencing data using deep learning

Yao-zhong Zhang
Institute of Medical Science, the University of Tokyo
Tokyo, Japan
yaozhong@ims.u-tokyo.ac.jp

Seiya Imoto
Institute of Medical Science, the University of Tokyo
Tokyo, Japan
imoto@ims.u-tokyo.ac.jp

Satoru Miyano
Institute of Medical Science, the University of Tokyo
Tokyo, Japan
miyano@ims.u-tokyo.ac.jp

Rui Yamaguchi
Institute of Medical Science, the University of Tokyo
Tokyo, Japan
ruiy@ims.u-tokyo.ac.jp

***Abstract*—Motivation: Next-generation sequencing (NGS) technologies using DNA, RNA, or methylation sequencing are prevailing tools used in modern genome research. For DNA sequencing, whole genome sequencing (WGS) and whole exome sequencing (WES) are two typical applications with a different preference on the trade-off between sequencing depth and base coverage. Although sequencing costs have been greatly reduced, the sequence depth used in WGS is relatively lower than WES (e.g., ∼35x vs. 100x∼). In addition, biases and batch effects may exist in different stages of a NGS experiment. Using low-depth and biased WGS data for downstream analyses is more sensitive to the bias problem and makes it even more difficult to uncover real biological signals in the data. In this work, we focused on reconstructing high read-depth signals from low-depth WGS data. We make use of a pair of WGS data with different read-depth for the same sample and learn a mapping from low-depth signals to high-depth in the given platform.**

Results: We explored three different reconstruction models from shallow to deep. Our experimental results show that by only using the read depth information, deeper models do not perform far better than a linear regression model. Through incorporating additional information, such as GC-content, mappability and nucleotide sequence information, the performance of convolutional neural network (CNN) models can be further improved. We made use of the reconstructed read-depth signals in downstream analysis to identify copy number variation segments for single sample. The experiment results show that segments that are not detected using low-depth data, can be detected with the reconstructed signals by the CNN model using extra biological information.

Availability: The source code will be available at https://github.com/yaozhong/DLRec

***Index Terms*—Whole genome sequencing, sequencing bias, deep learning**

I. Introduction

With more than a decade development, second-generation sequencing (also known as next-generation sequencing (NGS)) technologies have been widely used in genomic research and there is an explosive growth of NGS data in both scale and sample diversity. For DNA sequencing, whole-genome sequencing (WGS) and whole-exome sequencing (WES, or targeted sequencing in general) are the two most common usages. WGS sequences whole genome with a wide breadth and relatively low read-depth [1](10x∼35x), while WES focuses on target regions in a genome and sequences with high read-depths (60x∼). Although cost per Megabase has been greatly reduced in a NGS run, high-depth WGS is still magnitude more expensive than WES. In this study, we focused on the reconstruction of high read-depth signals from low-depth WGS data. We make use of a pair of WGS data with different read depths for the same sample and learn a mapping from low-depth data to high-depth data in the given platform(s). We assume the reconstruction method can be used for doing high-depth WGS for large-scale samples of the same specie, like 1000 genomes project (1000GP) [11]. After learning the mapping function from few high-depth WGS data, low-depth WGS is then conducted and reconstructed as high-depth WGS signals for downstream analysis, which provides a cheaper solution to do high-depth WGS for large-scale samples.

In WGS data, biases and batch effects commonly exist and can be incorporated in different stages of a NGS experiment. This makes learning the mapping from low-depth signals to high-depth a non-trivial task. For example, in the step of conventional library construction, amplification through polymerase chain reaction (PCR) has sequence-dependence efficiency, which makes the read coverage not as uniform as expected in practical situations. GC-content bias also affects read coverage in WGS data. These biases interact with the others, which makes the effect difficult to be described explicitly. Recently, deep learning methods have achieved impressive state-of-the-art results on many challenging tasks, such as image recognition and natural language understanding. In the field of bioinformatics, deep learning-based methods are being explored for many problems, such as predicting sequence specificity of protein binding [2], [13]. Koh et al. [5] proposed using convolutional neural network (CNN) for

Identify applicable funding agency here. If none, delete this.

[1]In this paper, we refer read-depth to the depth of per-base coverage for short. Per-base coverage is the average number of times a base of a genome is sequenced.

denoising ChIP-seq data. In their work, they utilized a two-layer CNN model to recover clean signal tracks and peak calls for each histone marker based on all available histone marker data. We take inspiration from this work and explored using deep learning models to learn the mapping from low-depth WGS data to high-depth WGS data. As we know the both data is from the same sample, but biases and batch effects may be incorporated separately. It is intuitive to make use of deep learning to model such a sophisticated correlation. Different from their work for the ChIP-seq data, we make use of the read depth information and extra biological knowledge, such as GC-content, mappability and sequence information in deep learning models.

Our experimental results show that by only using the read depth information, deeper models such as multiple layer perceptron (MLP) and CNN do not perform far better than a linear regression model. But CNN can take the advantage of incorporating additional biology information to further improve the reconstruction performance. We also use the reconstructed read-depth signals for identification of copy number variation segments for single sample (without any case-control sample). In the experiment, additional segments can be detected with reconstructed signals by CNN using extra biological knowledge. Those segments are detected in the high-depth data, while are not detected in the low-depth data.

II. Methods

The first step to build a reconstruction model is to collect a pair of WGS data with different read depths of the same sample for target NGS platform(s). The pair of WGS data can be generated from different platforms or the same platform[2] using different experiment settings. Here, we focused on the case of the same platform using different settings. We made use of the data from 1000 Genomes Project. The pair of WGS data is first aligned to a reference genome and segmented into non-overlapping bins across the genome, as shown in Figure 1(b). ld_i and hd_i are the read depth vectors of the i^{th} bin in low-depth data and high-depth data, respectively. $D = \{(ld_i, hd_i)|i = 1 \dots n\}$ is used as training data.

We formalize the reconstruction of high read-depth signals from low-depth data as a regression task. The high read-depth signals are predicted based on the low read-depth signals inside a bin. We applied three types of models from shallow to deep, which are linear regression model (LM), multiple-layer perceptron (MLP) and convolutional neural network (CNN). Compared with the LM, the MLP has one hidden layer with the number of hidden units the same as the input. The CNN uses a typical architecture of a convolutional network. It contains three convolutional layers using 32/64/128 filters with 7/3/3 bp length (stride=1, padding=0) for the first/second/third convolutional layer. Each convolutional layer is followed by a ReLU and a 3 bp length max-pooling operation with stride 3 for downsampling. Besides the read-depth information, we also explore the usage of additional biological knowledge, such as GC-content, mappability (or uniqueness) and nucleotide sequence information, which are incorporated into the CNN as extra input channels. It is reported that GC-content affects fragment count (read coverage) in Illumina sequencing data [3] and mappability has a major influence on the averaged mapped depth [6], [9]. For concise, we treat those additional information as a part of the model. We then have, $rd_i = Reconstruction(model, ld_i)$, where rd_i is the reconstructed read depth for the i^{th} bin.

The loss function used for model optimizing is the mean absolute value (MAE), which is conceptually simpler and more interpretable.

$$loss(D) = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} |rd_{i,j} - hc_{i,j}|,$$

where n is the number of bins and m is the bin length. $rd_{i,j}$ and $hc_{i,j}$ represent the read-depth in the j^{th} position of i^{th} bin. The averaged consine similarity between rd_i and hc_i is also used as an evaluation metric.

We use the trained reconstruction model for the augmentation of low-depth WGS data generated in the same environment as in the training data pair. It can be used for a large scale of populational low-depth WGS data, such as 1000GP.

III. Results

A. Experiment setup

The sample used in this study is the NA12878 CEPH individual from the 1000GP. It is one of the most comprehensively characterized samples with a large amount of genomic data and is widely used in various benchmark studies, such as Genome in a Bottle (GIAB) [14]. We chose this sample for the reason that both high-depth and low-depth WGS data is available. The high-depth WGS data has approximately 50-fold read-depth of coverage (50x). It was sequenced in a Illumina Hiseq platform using a PCR-free protocol and generated 250 bp paired-end reads. The PCR-free protocol is used in the high-depth data to reduce coverage bias. The low-depth WGS data has a lower read-depth of coverage around 7x and consists of 100 bp paired-end reads. PCR amplification was used for the low-depth data.

Both WGS data is mapped to the reference genome GRCh37/hg19. In this study, we directly downloaded the reference file (hs37d5.fa) and aligned bam files from the 1000GP FTP[3]. We filtered out sequencing inaccessible genome regions [10] which contain more than 1000 contiguous 'N' characters. The size of each sequencing-accessible region is characterized by the number of nucleotide bases. We randomly selected accessible regions from chromosome 1 to 22, and the total size of selected regions in each chromosome is in proportion to its sequencing-accessible chromosome size. We selected 100M bp covering regions for training and 10M bp covering regions for testing. For training, the dataset is further randomly split into

[2]The platform series, such as Illumina Hiseq series, are treated as the same platform.

[3]ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/

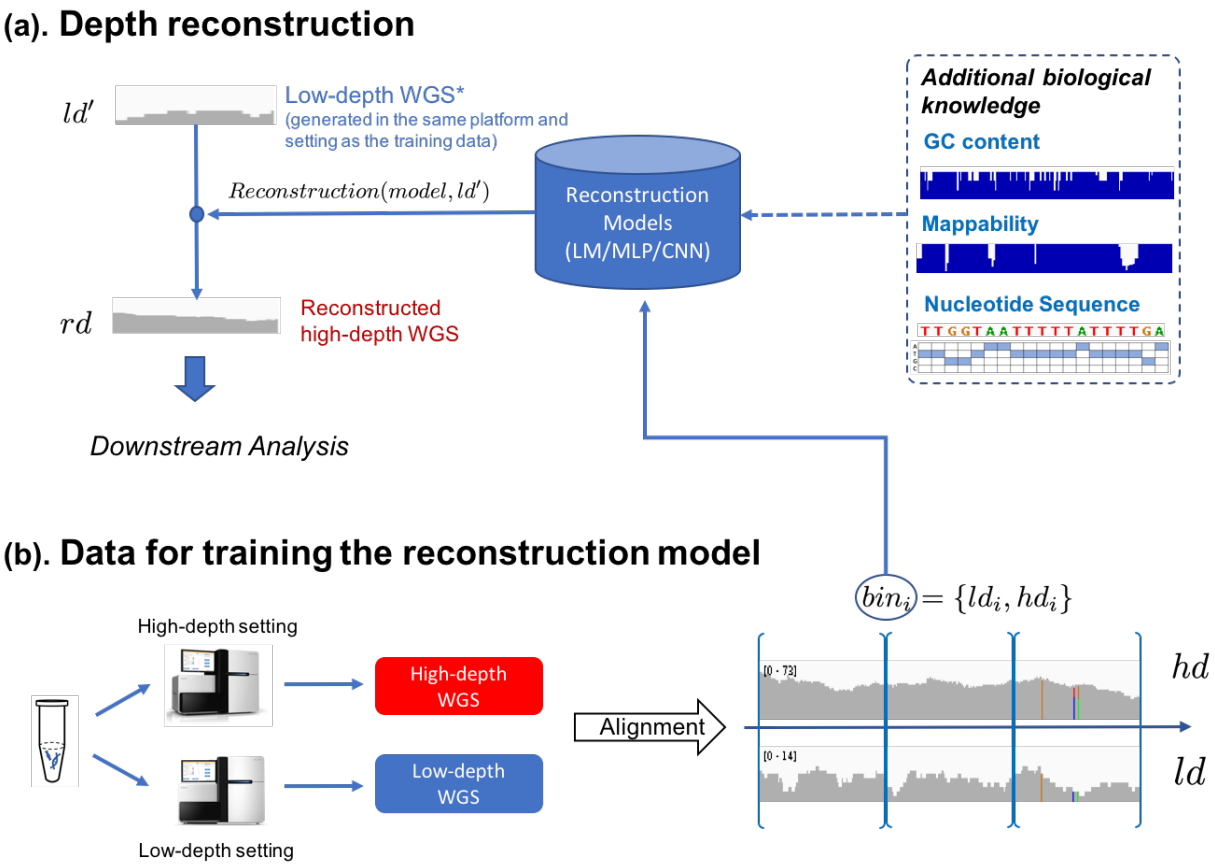


Figure 1. Reconstruction of high read-depth signals from low-depth WGS data. (a) shows a general pipeline of read-depth reconstruction. It is platform-specific that the WGS data to be reconstructed is supposed to be generated in the same platform as the training data. (b) shows the training data generation process.

training set and validation set (around 10% of the total training dataset). The GC-content and mappability profile are acquired from UCSC Genome Browser. The GC-content profile has a span size of 5 bases and the mappability profile uses the 100-mer mapping uniqueness for each coordinate.

Our reconstruction models are implemented with Chainer 2.0 library and are trained with Adam [4] adaptive learning rate algorithm and stochastic gradient descent. The minibatch-size used for training is 128. The training curves are monitored for the pre-defined number of epochs (100 epochs used here). The models are chosen with the lowest validation loss-score in all run epochs. MAE and averaged cosine-similarity on the test set are used as evaluation metrics. The training time of 100-epoch for 1000 bp bin-size training data, on a Intel machine with 3.5Ghz CPU and NVidia GTX1080 GPU acceleration is 239.8 s/259.0 s/601.9 s for LM/MLP/CNN, respectively.

B. Bias effect

We first show the difficulty of learning a base-wise mapping function from low-depth to high-depth signals, when the coverage bias exists. As illustrated in the above subsection, the high-depth data was generated with a PCR-free protocol, while the low-depth data used PCR amplification for library preparation. The low-depth data is supposed to contain a coverage bias [1]. The averaged base-wise ratio r between high-depth and low-depth data can be treated as the multiplication factor of the simplest mapping function that $hd_i = r \times ld_i$. We use the standard deviation of base-wise ratio of sampled coordinates as the indicator. Higher variance of the ratio indicates more difficulty in using the base-wise mapping to describe the relationship between low-depth and high-depth data. We down-sampled the high-depth data (to 50%) using *Samtools* (v1.3.1) [7] to synthesize low-depth data in PCR-free setting. From Table I, we can observe that the base-wise mapping from the low-depth data with coverage bias (non-PCR-free) has a higher variance than the down-sampled case (approximately 31 times). This indicates finding a mapping from low-depth data to high-depth data is not a trivial task, when biases and batch effects exist in the data.

C. Bin size effect on reconstruction models

Instead of only using base-wise information, we make use of the surrounding read-depth information in the bin of low-depth data to predict high-depth signals. We investigated 7 different bin sizes of 50 bp, 100 bp, 250 bp, 500 bp, 1000 bp, 2500 bp and 5000 bp.

Larger bin size indicates more surrounding read-depth information is used in the prediction. For LM, all input units are

Table I
THE MEAN AND STANDARD DEVIATION OF RATIOS BETWEEN THE HIGH-DEPTH AND DIFFERENT LOW-DEPTH SIGNALS OF SAMPLED COORDINATES. THE NUMBER OF BASE PAIRS USED FOR THE CALCULATION IS 2×10^7.

Mapping data pair	mean ratio	std
50% down-sampled HD $\rightarrow$ HD	2.057	0.39
LD $\rightarrow$ HD	15.43	12.28

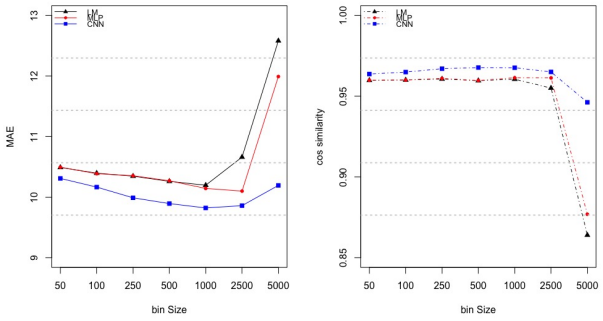


Figure 2. Performance of LM, MLP and CNN with different bin sizes. The left figure is the curve of MAE and the right figure is the cosine-similarity curve.

directly connected to output units using linear combinations. MLP uses one additional hidden layer to connect the input and the output. CNN uses local filters scanning the input bin and follows with rectification and max-pooling. From Figure 2, we can observe MAEs are gradually reduced and cosine-similarities are slightly increased as the bin size increased up to 1000 bp for the three models. When bin size exceeds 1000 bp, performances become worse as the bin size is increased. When comparing the performances in the bin size of 1000 bp and 5000 bp, a performance gap is observed for both MAE and cosine-similarity curve. CNN is less sensitive to the bin size change over 1000 bp, when compared with LM and MLP. For bin sizes less than 1000 bp, deeper models (CNN $>_{deeper\ than}$ MLP $>_{deeper\ than}$ LM) do not show an overwhelming advantage over shallow models, when the read-depth information is only used.

D. Additional biological knowledge used for the reconstruction

We make use of additional biological knowledge for the reconstruction task. We fixed the bin size as 1000 bp for the following experiments. We explored using GC-content (G), mappability (M) and nucleotide sequence (S) information in CNN. Compared with LM and MLP, it is more convenient to incorporate data of different scales as different kernels into CNN. We evaluated CNN with different extra features and feature combinations.

By incorporating GC-content, mappability and nucleotide sequence information separately, we can observe a performance improvement over basic CNN (read-depth only) on both MAE and cosine-similarity, shown in Table II. Among those three features, the largest gain is acquired by using the sequence information, which reduce the MAE from 9.837 to 9.06 (7.9% error reduction) and the Consine-similarity is also increased from 0.967 to 0.972 (0.5% improvement). CNN+M performs better than CNN+G on both MAE and cosine-similarity. We then investigated the combination of the three features. The GC-content and mappability combination (CNN+GM) generates a better performance than using those two separately. The GC-content and sequence feature combination (CNN+GS) gives a better performance than using G feature only, but worse than using S feature only. Interestingly, CNN using the mappability and sequence feature combination (CNN+MS) achieve the best results (MAE: 8.77 and cosine-similarity: 0.974), and even better than CNN with all three combination features (CNN+GMS). In our implementation, the nucleotide sequence information is represented in one-hot encoding for each coordinate and is incorporated through four different channels. GC-content and mappability are calculated based on nucleotide sequences. GC-content uses on 5-base window filter for sequences and calculates the Guanine and Cytosine percentage for in the window. Mappability uses 100-mer sequences and is calculated through mapping 100-mers back to the reference genome, which is a more global information. We think this might be a reason that the combination of more global mappability information and local nucleotide sequence information gives a better performance of CNN+MS. The nucleotide sequence feature overlaps with the GC-content feature to some extend and the GC-content feature may confound the usage of the nucleotide sequence feature (CNN+GS $<$ GNN+S and CNN+GMS $<$ CNN+MS).

Table II
THE PERFORMANCE OF DIFFERENT BIOLOGICAL INFORMATION USED AS THE INPUT FOR CNN MODELS. BIN SIZE IS 1000 BP. G, M AND S REPRESENT GC-CONTENT, MAPPABILITY AND NUCLEOTIDE SEQUENCE INFORMATION.

Input Features	MAE	Cosine-similarity
RD only	9.837	0.967
+ G(GC)	9.617	0.969
+ M(Mappability)	9.51	0.970
+ S(Sequence)	9.06	0.972
+ GM	9.25	0.971
+ GS	9.32	0.970
+ MS	8.77	0.974
+ GMS	9.06	0.973

E. CNVs detection using reconstructed high-depth signals

We further evaluated the reconstructed high-depth signals in a downstream analysis. We detected copy number variation (CNV) segments based on the WGS read-depth signals for single sample without any case-control sample. We used R-package of *DNAcopy* [8], which applies circular binary segmentation (CBS) algorithm [12] to detect the genomic region that has copy number different from its adjacent regions. For computational efficiency and noise reduction, we averaged the signals in each bin. For example, we calculated the mean read-

depth of the 1000 bp coordinates for each bin. We used the undo split option in *DNAcopy* for the segment in the range of three standard deviation of read-depth. We compared CNV segments detected with original signals (low-depth and high-depth) and reconstructed signals (LM, MLP, CNN, CNN+MS and CNN+GMS) on the test dataset.

Since the reconstructed high-depth data is to mimic the original high-depth data, we use the segments detected with the original high-depth data as the golden standard. We draw the heatmap of sensitivity and false discovery rate for all test regions shown in Figure 3(a) and 3(b). The most left column for Figure 3(a) and Figure 3(a) is the number of segments detected with high-depth signals, which is used as the ground truth. Based on the heatmaps, the segment results can be categorized into two groups. The first group contains the left two columns in the heatmaps, which are CNN+MS and CNN+GMS (CNN with extra biological features). The second group covers the right four columns in heatmaps, which are LD, LM, MLP and CNN. From Figure 3(a), we can observe the first group detects extra segments, which are not detected in the second group. In the first group, the results of CNN+MS and CNN+GMS are not totally overlapped. CNN+MS works a bit better based on the sensitivity and FDR. For example, Figure 4 shows a positive example on $chr6 : 61880166 - 62125166$ that one extra segment is detected with CNN+(G)MS reconstructed signals. The segment is detected in the original high-depth data and is not detected in the low-depth data and reconstruction models without extra biological features. CNN+MS detect the exactly break point at the relative position of 239, while CNN+GMS has 1000 bp shift that the break point is at the relative position of 238. The above results also show the effectiveness of using the extra biological knowledge in the high-depth reconstruction process.

IV. Discussion

In this study, we focused on the task of reconstructing high read-depth signals from low-depth WGS data. Our general goal is to reduce the cost of high-depth WGS through low-depth WGS with data reconstruction. We made use of paired WGS data generated from target NGS platform(s) with different read depths for the same sample to train a reconstruction model. The high-depth signals can be reconstructed using the model for the low-depth WGS data generated in the same NGS setting as the low-depth training data. We incorporated extra biological knowledge into a CNN model to further enhance the reconstruction performance. In this paper, we only use the data from the same platform with different sequencing settings. We would like to apply this method for WGS data from different platforms, such as Nanopore WGS data, for future work.

Our work is motivated by [5], who proposed to use CNN for denoising ChIP-seq data. In the work, they utilized a two-convolutional-layer CNN model without max-pooling to recover clean signal tracks from noisy ChIP-seq data based on the input of 5 histone markers. Different from their work, we applied the reconstruction task for whole genome sequence data on the original read-depth data without averaging for bins and used the max-pooling layer in the CNN. There are two reasons for using the max-pooling layer here. First, we don't use the overlapping bins as input for prediction, which can avoid repetitive computational cost in CNN. Second, max-pooling is used to smooth the original base-wise signals. Our experimental trails indicate CNN with max-pooling gives better results than CNN without max-pooling and the three-layer model gives better performance than the two-layer model on the dataset.

V. Conclusions

In this paper, we proposed an approach to reconstruct high read-depth signals from low-depth WGS data using deep learning models. Besides the original read-depth information inside a bin, we explored the integration of additional biological knowledge into the CNN model. Our experiment results show that those additional biological information, especially the combination of the mappability and nucleotide sequences, can further enhance the reconstruction performances. Further more, we evaluated the reconstructed signals in a CNV segmentation task for single sample and the CNN-based reconstruction model with extra biological information shows a potential to enhance low-depth signals for detecting CNV segments.

Ackownledgement

This research was supported by MEXT as Priority Issue on Post-K computer hp170227) and MEXT Innovative Area (15H05912).

References

[1] D. Aird, M. G. Ross, W.-S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke, "Analyzing and minimizing pcr amplification bias in illumina sequencing libraries," *Genome biology*, vol. 12, no. 2, p. R18, 2011.

[2] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of dna-and rna-binding proteins by deep learning," *Nature biotechnology*, vol. 33, no. 8, pp. 831–838, 2015.

[3] Y. Benjamini and T. P. Speed, "Summarizing and correcting the gc content bias in high-throughput sequencing," *Nucleic acids research*, vol. 40, no. 10, pp. e72–e72, 2012.

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization." *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14

[5] P. W. Koh, E. Pierson, and A. Kundaje, "Denoising genome-wide histone chip-seq with convolutional neural networks," *Bioinformatics*, vol. 33, no. 14, pp. i225–i233, 2017.

[6] H. Lee and M. C. Schatz, "Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score," *Bioinformatics*, vol. 28, no. 16, pp. 2097–2105, 2012.

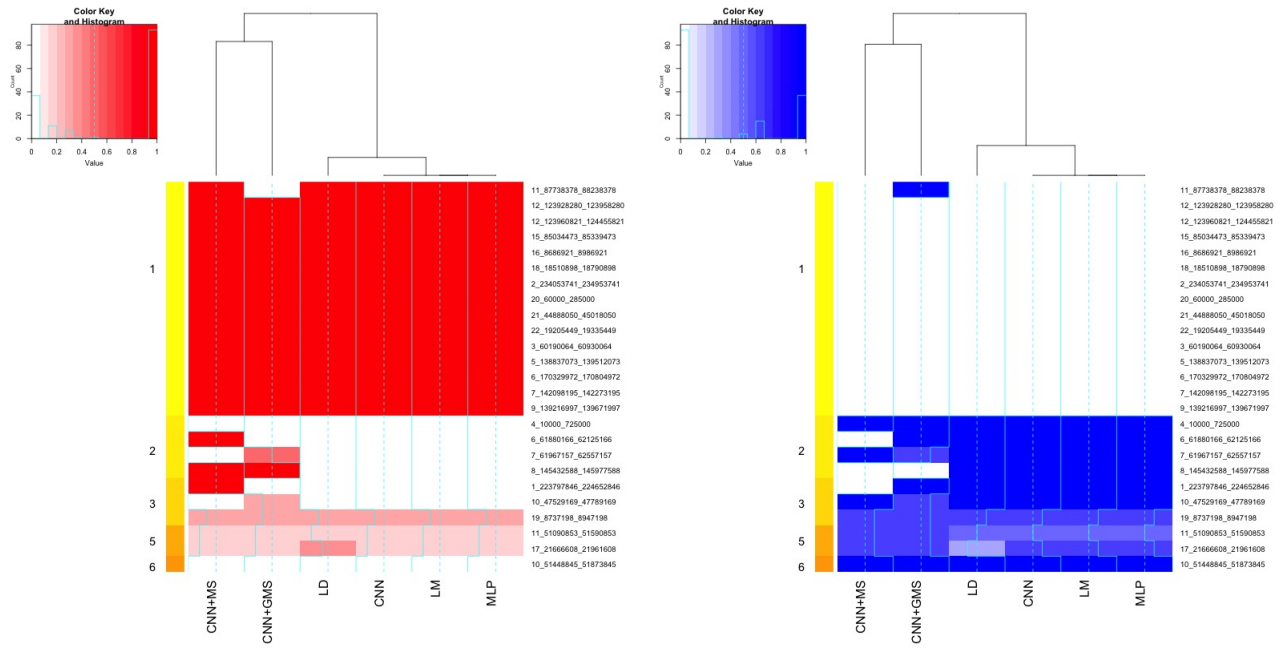
[7] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[8] V. E. Seshan and A. Olshen, "Dnacopy: Dna copy number data analysis," *R package version*, vol. 1, no. 1, 2016.

[9] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, 2014.

[10] E. Talevich, A. H. Shain, T. Botton, and B. C. Bastian, "Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing," *PLoS computational biology*, vol. 12, no. 4, p. e1004873, 2016.

[11] The 1000 Genomes Project Consortium, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, p. 68, 2015.



(a) Sensitivity of each test sample

(b) False discovery rate of each test sample

Figure 3. The heatmap of sensitivity and FDR for all test regions with low-depth signals and reconstructed signals. Segments detected with high-depth (HD) data is used as the golden standard. The most left "yellow-to-orange" column refers to the number of segments for the regions. If there is no breakpoint in the region, the number of segment for the region is counted as one.

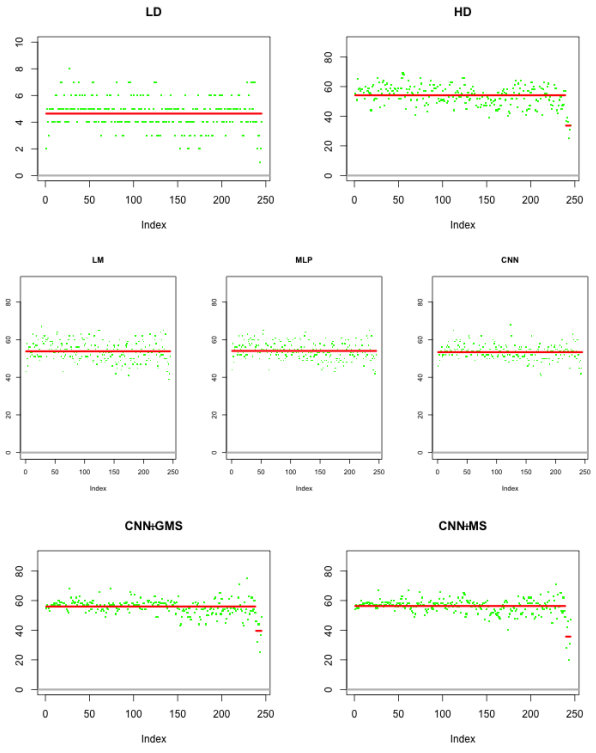


Figure 4. CNV segmentation results with different read-depth signals for the region chr6:61880166-62125166. The X-axis is in 1k bp scale and Y-axis is the read-depth. The high-depth (HD) data has a breakpoint at the relative position of 239.

[12] E. Venkatraman and A. B. Olshen, "A faster circular binary segmentation algorithm for the analysis of array cgh data," *Bioinformatics*, vol. 23, no. 6, pp. 657–663, 2007.

[13] J. Zhou and O. G. Troyanskaya, "Predicting effects of noncoding variants with deep learning–based sequence model," *Nature methods*, vol. 12, no. 10, p. 931, 2015.

[14] J. M. Zook *et al.*, "Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls," *Nat Biotechnol*, vol. 32, pp. 246–51, 2014.