UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE
BIOINFORMATICS CENTRE

DeepCNV: a deep learning-based method for calling copy number variations by image recognition

Author: *Zuqi Li*

Supervisors: *Alfonso Buil Demur*

Anders Albrechtsen

Submitted on: *15 July 2019*

Contents

Abstract

Copy number variation (CNV) is an important molecular mechanism, leading to many phenotypic changes and even diseases. To better analyze it, various methods have been developed based on different algorithms and data types to call CNVs. Here I implemented a deep learning-based model, called DeepCNV, that takes the total signal intensity and allelic intensity ratio of every 100 SNPs as images to predict CNVs from a subset of iPSYCH study case cohort. The model was tuned in multiple ways and the optimal one reached an accuracy of 87.2% via validation. For generalization, I tested the optimal model on an independent dataset from NeuroDevNet FASD study and it recalled 97.2% deletions and 94.9% duplications stringently labeled by the study. To compare with the 'state-of-the-art' method (PennCNV was chosen), I utilized both of them to call CNVs from 2 datasets separately. The 2 methods have generally similar performance on the iPSYCH dataset and, for the FASD dataset, PennCNV guaranteed higher true positive rate while DeepCNV correctly called more CNVs. The results demonstrate the capacity of DeepCNV to call CNVs from iPSYCH data and its potential across populations, noise levels and SNP array platforms.

1. Introduction

1.1 CNV

Copy number variants (CNVs) are a kind of genomic structural variation which mainly consists of deletion (loss of sequence) and duplication (gain of sequence). With more duplications than deletions, CNVs distribute unevenly over 4.8-9.7% of the human genome, especially in sub-telomeric and pericentromeric regions **(Mehdi et al. 2015)**. Based on 500 base-pair windows of digital comparative genomic hybridization (dCGH) data, human genome was estimated to contain around 64 CNVs and the average length of CNV is roughly 7.4K base pairs **(Sudmant et al. 2015)**.

Even though, in most cases, CNVs won't cause severe phenotypic consequences, they will affect phenotypes and even cause diseases directly by altering gene expression or indirectly by position effect **(Stranger et al. 2007; Feuk et al. 2006)**. Risk of autism spectrum disorders and susceptibility of HIV-1/AIDS, among others, have been reported to be associated with rare CNVs **(Pinto et al. 2010; Gonzalez 2005)**. Another well-documented correlation is between Schizophrenia and the deletion at 1q21.1. George Kirov counted 17 deletions at 1q21.1 in 7918 cases (0.2%) and 11 in 46,502 controls (0.02%) from 3 largest studies in Schizophrenia, indicating a strong correlation (Fisher's Exact Test: P-value = 2.52×10^{-8}) **(Kirov 2010)**. The association between the prevalence of CNVs on chromosome 22q11.2 and risks of psychiatric diseases has been investigated in Danish population **(Olsen et al. 2018)**.

1.2 Current methods

There are various methods for calling CNVs with different data types. ChAMP-CNV and CN450K can be performed with DNA methylation array data **(Morris et al. 2013; Papillon-Cavanagh et al. 2013)**, circular binary segmentation (CBS) and gain and loss analysis of DNA (GLAD) with array CGH data **(Hsu et al. 2011; Hupé et al. 2004)**, PennCNV and QuantiSNP (Illumina only) with SNP array data **(Wang et al. 2007; Colella et al. 2007)**. Apart from that, for genomic sequencing data, Control-FREEC and ExomeCNV are some of the available platforms for CNV calling **(Boeva et al. 2011; Sathirapongsasuti et al. 2011)**.

Among all SNP array-based CNV-calling methods, PennCNV is frequently adopted (cited by over 1,300 times) and thus is used as the 'state-of-the-art' method here for comparison. Based on hidden Markov model (HMM), PennCNV incorporates B allele frequency (BAF) and log R ratio (LRR) of each SNP to detect CNV under the assistance of population-based BAF. More specifically, if a and b represent the normalized signal intensity of the 2 alleles of a single SNP, respectively, the total intensity of both alleles can be measured by $R = a + b$ and the relativity of intensities between the 2 alleles by $\theta = \tan^{-1}(a/b)/(\pi/2)$. BAF is subsequently calculated as the normalized θ and LRR as $\log_2(R_1/R_2)$, where $R_1$is the observed total intensity from target samples while R_2 the expected total intensity from reference samples **(Li and Michael 2012)**. Deletion and duplication can thereby be distinguished from each other and the background through BAF and LRR (**Figure 1**).

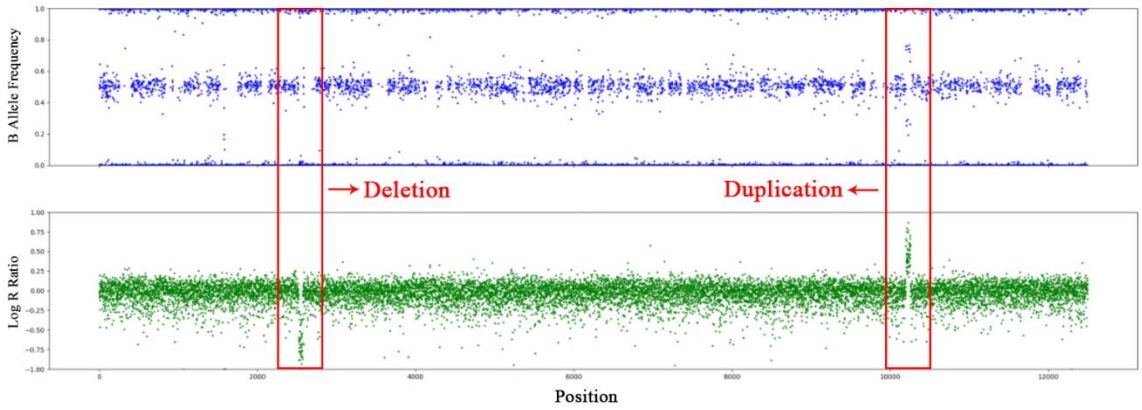

Figure 1: Example of a deletion and a duplication in a 12,500-SNP-long region on chromosome 2. BAF of the 12,500 SNPs are plotted in blue and LRR in green. The 2 red boxes highlight the deletion and duplication, respectively.

1.3 Motivation

Even though various CNV-calling methods are available for SNP arrays, their performances may be limited by the complexity and flexibility of their algorithms. HMM, for example, only takes previous SNPs into consideration to identify copy numbers. PennCNV applies first-order HMM which only looks at the dependency between 2 adjacent SNPs **(Wang et al. 2007)**. Therefore, current methods may be intrinsically deficient on calling large CNVs, especially from noisy data. In addition, it's difficult to further tune HMMs based on new data as well. All the features mentioned lead to deep

learning algorithm, a kind of multi-layer neural network with different architectures. Each layer in neural network is made of nodes, resembling neurons in human brain. In a similar way of perception, the neural network captures patterns from the input by training the parameters representing every node.

In particular, I would like to call CNVs from iPSYCH data whose DNA samples came from phenylketonuria (PKU) test and needed amplification. The noise brought by amplified DNA causes a lot of trouble for normal CNV callers, e.g. PennCNV. To fit the iPSYCH dataset, I developed a deep learning-based method, DeepCNV, that identifies CNVs from BAF and LRR of Illumina PsychArray. Recently, more and more SNP array and CNV data have been cumulated, which provides the foundation to benefit from more complex models, such as deep learning. The success of AlphaFold in the Critical Assessment of protein Structure Prediction (CASP) competition also proves it **(AlQuraishi 2019)**. DeepCNV consists of a convolutional neural network (CNN) and a subsequent recurrent neural network (RNN), which is an architecture having been practiced by other researchers **(Fan et al. 2016**; **Wang et al. 2016)**.