Integrating Bias Correction in Read Depth Analysis of Whole Genome Sequencing Data to Identify Copy Number Gains and Losses

Jin P Szatkiewicz[1], Weibo Wang[2], Wei Sun[1,3], Wei Wang[2], Danyu Lin[3], Jurgen Del-Favero[4], Rolf Adolfsson[5], Patrick F Sullivan[1,6]

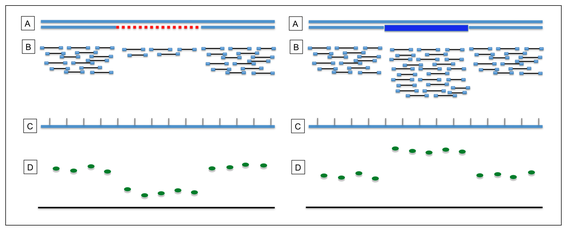
Departments of Genetics[1], Computer Science[2], Biostatistics[3], Psychiatry[6], University of North Carolina, Chapel Hill, North Carolina, USA;
University of Antwerp, Belgium[4]; Umea University, Sweden[5].

Introduction

- CNVs play an important role in the etiology of schizophrenia, autism, developmental delay, brain size, and epilepsy.
- A large amount of high-throughput sequencing (HTS) data is emerging for the study of psychiatric disorders.
- CNV detection from HTS data is challenging. Experimental biases have direct effects on pairing the reads and estimating read depth.
- We present a novel statistical method and software tool for CNV detection from read-depth (RD) analysis of whole genome sequencing data.
- Our method uses a hidden Markov model (HMM) and negative binomial regression framework to identify CNVs and to account for confounders.
- We calibrated our method using 1000 genomes project trio data.
- We then applied our method to detect CNVs using whole genome sequencing data from an individual with schizophrenia.

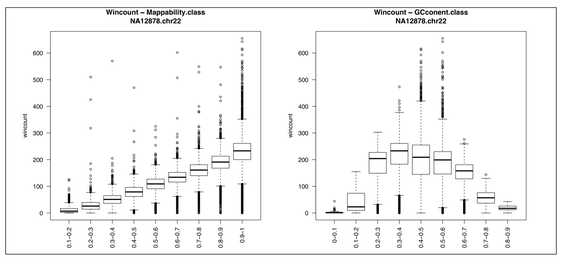
Method Motivation

Ideal: Read-depth analysis in absence of biases



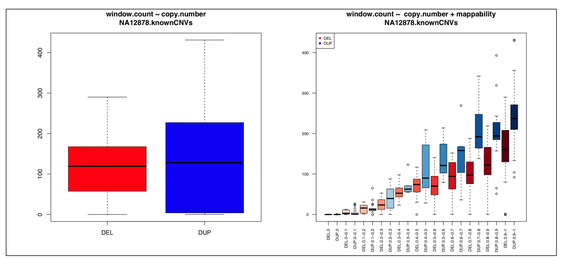
[Left] Deletion. [Right] Duplication. (A) DNA sequence of an individual (B) Uniform sampling of DNA fragments. (C) Reference genome, partitioned to overlapping windows. (D) Read-depth signals used for segmentation.

Reality: Read counts are strongly influenced by biases



[Left] Strong linear relationship with mappability. [Right] Non-linear relationship with GC content

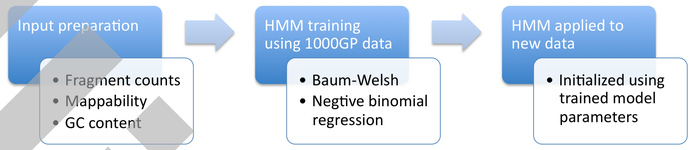
Key idea: Identifying CNVs while simultaneously accounting for the effects of confounders improve detection power.



[Left] Ignoring confounding effects entails loss of power for detecting known CNVs. [Right] Joint estimation improves power for detecting known CNVs.

Method Summary

We use the negative binomial (NB) distribution to model fragment counts, and generalized linear models (GLM) jointly to estimate copy number and confounder effects. Unknown experimental biases are accommodated by the over-dispersion parameter of NB. We use HMM to infer the underlying copy number states.



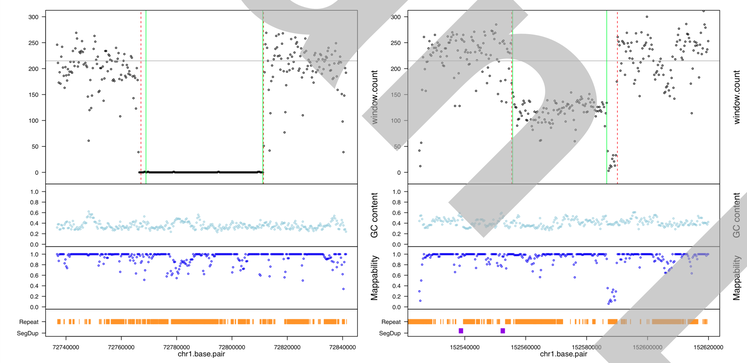
Results: Sensitivity for Detecting Deletions

The sensitivity of our method for detecting deletions was assessed based on the combined gold standard set (Mills et al 2011) derived from individual NA12878. Sensitivity of other methods was reported in the Supplementary Table 8B of Mills et al (2011) using the same approach. We have preliminary evidence that the sensitivity of the HMM-NB method was superior for the detection of deletions and duplications.

Approach	Origin	Method	Sensitivity (DEL, 50% overlap)	Sensitivity (DEL, 1bp overlap)	CPU	Run Time
Read-depth	1000GP-UW	unpublished	0.042	0.097	NA	NA
Read-depth	1000GP-SD	Event-wise testing	0.444	0.556	A linux cluster	>20 h
Read-depth	1000GP- YL	CNVnator	0.634	0.775	2.5GHz Core 2 Duo	A few hours
Read-depth	UNC	HMM-NB	0.813	>0.81	2.5GHz Xeon E5420	6 hour chr1

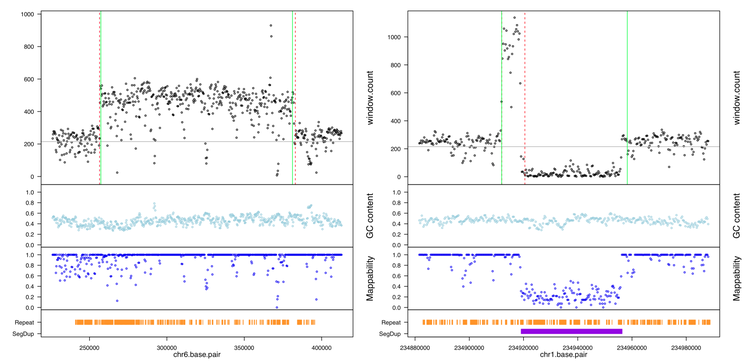
Results: Deletions Detected in NA12878

[Top Panel] Green: CNV boundary from Mills et al (2011) PMID: 21293372; Red: CNV boundary from our method. [Left] A homozygous deletion. [Right] A heterozygous deletion.



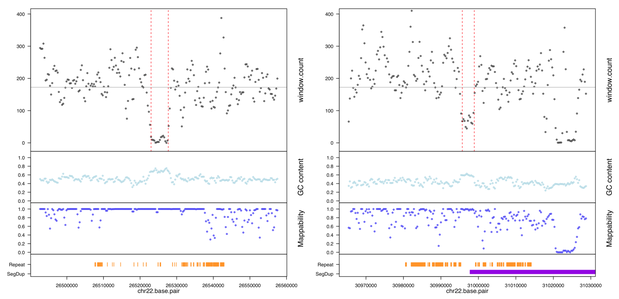
Results: Duplications Detected in NA12878

[Top Panel] Green: CNV boundary from Mills et al (2011); Red: CNV boundary from our method. [Left] A simple duplication. [Right] A complex region overlapping a segmental duplication.



Application to Schizophrenia Dataset

The study sample is an individual with schizophrenia from a large Swedish pedigree rich in schizophrenia and bipolar disorder with three founder pairs dating back approximately twelve generations. Whole genome sequencing data of this sample was obtained at 40X coverage using 33bp paired-end reads. The figures show example deletions detected using our method. [Top Panel] Red: CNV boundary from our method. [Right] A complex heterozygous deletion partially overlapping a segmental duplication.



Conclusions

- Our integrative method outperforms existing read-depth-based method in terms of sensitivity for detecting deletions. The specificity assessment and comparison is ongoing using the 1000GP trio data.
- Our method has been efficiently implemented in C++. We plan to parallelize the alogirthm to further accelerate the computation.
- Integrating bias correction in read-depth analysis improves the power for detecting CNVs and is complementary to read-pair and split-read approaches. A user-friendly implementation of our complete analytic protocol will be freely available.

Acknowledgements

No conflicts reported. Funding for this project was from the US National Institutes of Health who had no role in the design, execution, analysis, and manuscript preparation. This project is funded by K01MH093517 (JPS).

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL