# Copy Number Variation (CNV) detection in whole exome sequencing data for Mendelian disorders

Peng Zhang, Hua Ling, Elizabeth Pugh, Kurt Hetrick, Dane Witmer, Nara Sobreira, David Valle, Kim Doheny
Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, SOM, Johns Hopkins University, Baltimore, MD 21224

## Introduction

The Centers for Mendelian Genomics (CMG) project uses next-generation sequencing and computational approaches to discover the genes and variants that underlie Mendelian conditions. While SNVs and INDELs explain some Mendelian conditions, many remain unresolved. We are interested to know to what extent unrecognized CNVs would resolve some of these.

Compared to whole genome sequencing (WGS), whole-exome sequencing (WES) is a cost-effective alternative for finding disease genes harboring variants with relatively large effect size. However, CNVs from WES has been a challenge because of the sparseness of the target regions and the non-uniform distribution of reads across genome.

As part of the CMG project, we applied four prevailing CNV calling methods (Conifer, XHMM, ExomeDepth, and EXCAVATOR) on 677 WES samples to search for rare exonic CNVs that might be causal for the disease of interest. We found that ExomeDepth and EXCAVATOR allow users to specify controls in CNV calling, which was appropriate for our family based analyses. We thus focus results from these two methods. In addition, we evaluated how the selection of controls affects CNV calls, taking into consideration disease status, ethnicity, and DNA source.

## Materials and methods

**Sample selection:**
- 677 CMG samples with WES data.

**Sequencer and reagents:**
- Exome Capture: Agilent SureSelect HumanAllExonV4.
- Illumina HiSeq2500 platform (Majority).
- TruSeq Rapid SBS-HS 100 bp Paired Ends (Majority).

**Sequencing data processing:**
- BWA mem 0.7.8 alignment, local alignment and base call quality score recalibration with GATK 3.1-1.

**CNV analysis control selections:**
1) (**Auto**): Run all samples with ExomeDepth, ExomeDepth picked controls based on pair-wise correlation of reads between each test sample and the rest of samples. Then run EXCAVATOR using controls picked by ExomeDepth (selected up to 5).
2) (**Manual**): Selected a subset with 173 samples from three populations. For each population, let ExomeDepth pick controls only from unaffected individuals, then run EXCAVATOR the same as 1).
3) Compare results from 1) and 2), evaluate how the selection of controls affect the CNV calling.

**Program optimizations:**
- ExomeDepth: removed secondary alignment reads in BAM files.
- EXCAVATOR: filtered unmapped reads, as well as PCR duplicates and secondary alignment reads to improve performance (without increasing running time).

## Results

**For the 677 samples:**
- The median number of CNVs called by EXCAVATOR is 10 (min=1, max=250, mean=14); the median number of CNVs called by ExomeDepth is 81 (min=41, max=1356, mean=122).

**For the Subset of 173 samples:**

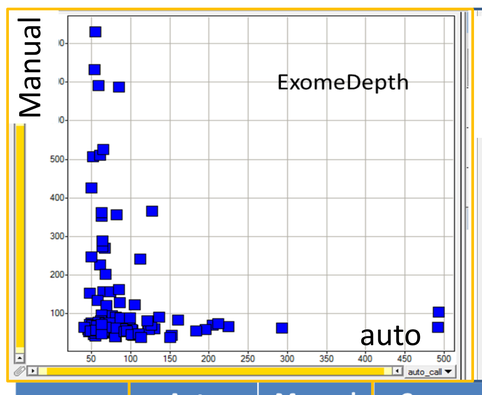



**Figure 1. Number of CNV called for each sample by Auto and Manual control selection for each method:** ExomeDepth is more sensitive to control selection (18.4% of CNVs called by two runs), while EXCAVATOR shows a bigger correlation of number of CNV calls at sample level between the Auto and Manual selection of controls (34.4% concordance).

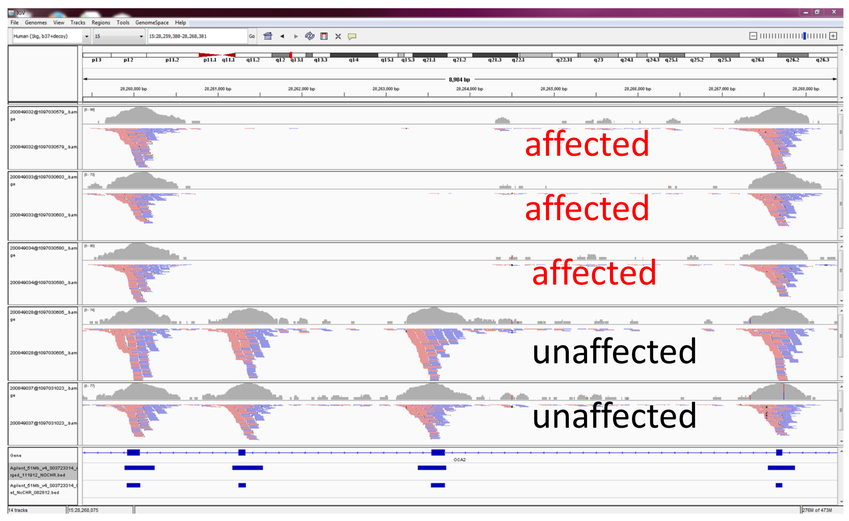
| | Auto | Manual | Concordance(auto vs Manual) | Auto | Manual |
|---|---|---|---|---|---|
| Median | 67 | 65 | ExomeDepth | 11 | 11 |
| Mean | 83 | 107 | 18.4% (3389/18436) | 19 | 13 |
| Min | 41 | 38 | EXCAVATOR | 2 | 5 |
| Max | 493 | 831 | 34.4% (1050/3051) | 251 | 78 |

**Table 1. Comparison between ExomeDepth and EXCAVATOR (173 samples).**
- ExomeDepth calls more aggressively while EXCAVATOR has a higher concordance.

| | ExomeDepth | | | EXCAVATOR | | |
|---|---|---|---|---|---|---|
| | total | both method* | Same CNV# | total | both method* | Same CNV# |
| Auto | 14339 | 44.2% (6342) | 5.3% (763) | 3051 | 67.4% (2055) | 23.9% (728) |
| Manual | 18436 | 33.6% (6193) | 7.2% (1322) | 1984 | 98.6% (1957) | 60.0% (1189) |

*: Same CNVs called by ExomeDepth also called by EXCAVATOR (or vice versa), can from different individuals
#: Same CNVs called by both ExomeDepth and EXCAVATOR from the same individual

**Figure 2. A two-exon deletion detected in affected individuals in one family, shown in IGV.**




This deletion (OCA2 gene, chr15:28,261,251-28263703, ~2.45kb) is called by both ExomeDepth and EXCAVATOR with manual selection of controls, while not called by Auto selection of controls.

## Summary

- We optimized ExomeDepth and EXCAVATOR by adding extra filters to improve their sensitivity and specificity.
- Concordance between different methods may be low. Real CNVs might only be called by one program, so it's important to evaluate the results combined with other information.
- ExomeDepth is more aggressive in calling CNVs and is more sensitive to control selection compared to EXCAVATOR. In comparison, EXCAVATOR is more conservative in CNV calling.
- Control selection affects the number of CNV calls. To identify disease causing CNVs, It's important to match the test sample and controls for the disease status, ethnicity and other factors (batch, calling pipeline etc.).

## References

- http://www.mendelian.org/ (CMG)
- Magi et al. Genome Biology 2013, 14:R120 (EXCAVATOR)
- Plagnol et al. Bioinformatics 2012, 28:2747-2754 (ExomeDepth)