Integrating copy number variants (CNV) calls from exome sequencing data

Peng Zhang, Hua Ling, Elizabeth Pugh, Kim Doheny

Center for Inherited Disease Research (CIDR), Johns Hopkins Genomics, Institute of Genetic Medicine, The Johns Hopkins School of Medicine

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality Next-Generation Sequencing (NGS), genotyping and statistical genetics consultation to investigators working to discover genes that contribute to disease and mapping a genetic path to better health. Copy number variants (CNVs, > 50 bp) presented in the genome have been shown to contribute to both Mendelian disorders and complex traits. However, calling CNVs from exome sequencing data has been a challenge because of non-continuousness of targets and the unevenness of read depth across the genome. As a result, the CNV calling programs apply various algorithms for normalization and variant calling, resulting in dramatically variable numbers and types of calls between programs. Validating these calls can often be extremely time-consuming as the same CNV may be defined by different programs as having different breakpoints.

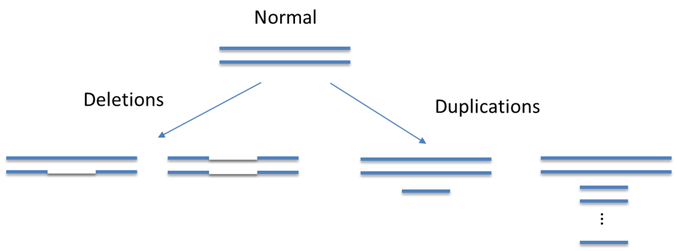


Figure 1. Copy Number Variants (CNVs)

In this project, we propose a new method, ECO (Exome CNV Overlapping), to integrate results from different programs. The program allows the user to merge calls based on the percentage of overlaps, it also filters the results based on the call frequency and size. The goal is to merge the same CNV calls across programs and to filter out as many false positive calls as possible. We applied ECO to the CNV calls from four different programs on 1,576 exome sequencing data that generated at CIDR as part of the Baylor Hopkins Center for Mendelian Genomics (BHCMG) project and summarized results here.

The algorithm and datasets

The ECO algorithm:

- We first pull all the results across individuals and programs together, then separate them by chromosomes and sort the calls by decreasing lengths, then for each chromosome, begining with the largest CNV call (CNV_1), we assign the same unique ID if any of the smaller CNVs (CNV_2) overlap with CNV_1 and length(CNV_2)/length(CNV_1) >= overlap percentage threshold. The process continues until all calls have been assigned with an ID.

Sample selection:

- 1,576 BHCMG 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.
- Samtools 0.1.18 extract mtDNA reads, exclude secondary alignment.

Exome CNV calling programs:

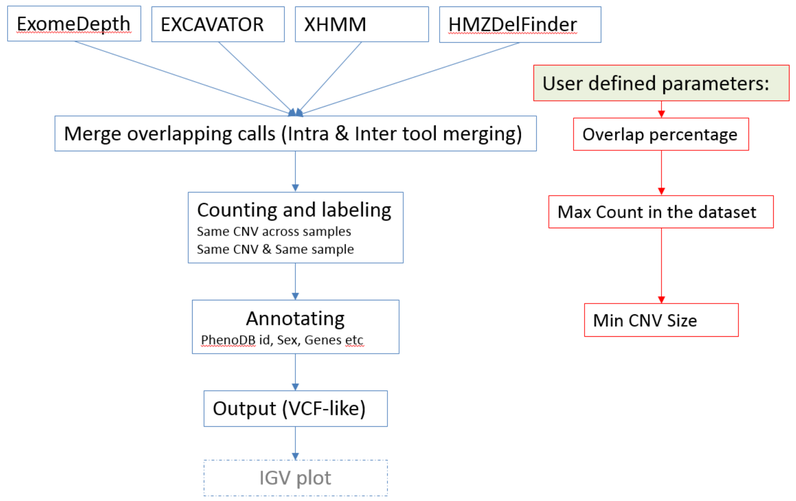
- ExomeDepth, EXCAVATOR, XHMM, and HMZDelFinder.

Results

We ran four exome CNV calling programs on the 1,576 samples sequenced at CIDR, the number of CNV calls were 2,670 (HMZDelFinder), 38,952 (XHMM), 72,229 (EXCAVATOR), and 174,183 (ExomeDepth), respectively.

We ran the ECO program under two overlap percentages 50% and 80%, respectively, with minimum CNV size of 100 bp and maximum count of 15 in the dataset. With the 50% overlap threshold, the combined dataset generated 27,585 unique CNVs from a total of 84, 904 calls, while with 80% overlap threshold, it is 37,235 unique calls and 84,904 overall calls.

Figure 2. A flow chart of ECO algorithm:



An example output:

Run on 02/12/16 10:26
combine cnv results
Filters: SET_COUNT = 15 SET_OVERLAP_PCT = 0.5 SET_CNV_SIZE = 100
count_in_data: times of the same cnv (same cnv_ID) called in the dataset from different individuals
count_in_sample: times of same cnv (same cnv_ID) called from the same individual
size: cnv sizes in bp, base pair
cnv_ID: IDs assigned to each unique cnv, with format chromosome_num, where same cnv_ID indicates the same cnv using current filters
Info: other information from each program if that cnv was called by that program

sm_tag	chr	start	end	count_in _data	count_in _sample	size	cnv_ID	Phe	Sex	Project	gene	Info
10104-110	5	60,368,834	180,374,992	1	1	120,006,158	5_2815	BH2	M	M_Valle_I	JMY,ERAP1,N	EXCAVATO
10057-110	8	47,752,560	146,279,718	2	1	98,527,158	8_1817	BH6	F	M_Valle_I	PPP1R16A,ZN	EXCAVATO
200913368	X	62,519,127	101,581,638	7	1	39,062,511	X_9348	BH2	F	M_Valle_I	.	EXCAVATO
33654-112	12	175,885	9,447,649	1	3	9,271,764	12_2727	BH6	F	M_Valle_I	SLC6A13,IFFO	EXCAVATO
10057-110	8	24,192,779	33,455,186	2	1	9,262,407	8_1799	BH6	F	M_Valle_I	TEX15,SCARA	XHMM-typ
33474-017	1	240,492,159	249,212,798	2	3	8,720,639	1_5636	BH7	F	M_Valle_I	OR2W3,GCSA	EXCAVATO

The info column:

ExomeDepth-type:BF:Exp_reads:Obs_reads:Ratio=:deletion:24.5:117:0:0;XHMM-type:MID_BP:MEAN_RD:MEAN_ORIG_RD =:DEL:28262483:-7.36:0.00;HMZDelFinder-Genes:Mark_num:Exon_num:inAOH_1000=:Ln_53603,Ln_53604:2:2:200849040@1097031048_15:28261175_28263722

Summary

- We proposed a program, ECO, for integrating CNV results from different programs, which allows users to merge with overlap percentage and to filter results based on the CNV size and frequency.
- The program can provide annotations such as the number of genes and the gene names within CNV call region.
- This framework can be extended to include results from other programs and whole genome sequencing.
- Future work: 1) including more programs 2) plotting the CNV in IGV.

References

- http://www.mendelian.org/ (BHCMG)
- Magi et al. Genome Biology 2013, 14:R120 (EXCAVATOR)
- Plagnol et al. Bioinformatics 2012, 28:2747-2754 (ExomeDepth)
- Fromer et al. ASHG 2012, 91(4): 597-607 (XHMM)
- https://github.com/BCM-Lupskilab/HMZDelFinder (HMZDelFinder)