# Exome CNV Overlapping (ECO): an integrative copy number variation caller for exome sequencing

Peng Zhang, Hua Ling, Elizabeth Pugh, Kimberly Doheny
Center for Inherited Disease Research (CIDR), Johns Hopkins Genomics, Institute of Genetic Medicine, The Johns Hopkins School of Medicine

## Introduction

Due to the uneven distribution of reads and the sparse nature of target regions for whole exome sequencing (WES) data, calling copy number variations (CNVs) has been a challenge. Most existing programs can only use read counts as inputs and calls often vary between programs. As part of the validation process, we found that some confirmed causal CNVs were called by multiple programs while others were not. In addition, each program often requires different input files and its output format often varies with different breakpoints for the CNV calls, which makes it difficult to compare and summarize results across programs.

We present here a practical pipeline that integrates multiple CNV calling programs and generates one combined VCF-like report with merged calls and annotations. It incorporated three prevalent CNV calling programs (ExomeDepth [Plagnol et al. 2012], CANOES [Backenroth et al. 2014], and CODEX [Jiang et al. 2015]) with the ability to incorporate results from additional programs  such as XHMM [Fromer and Purcell 2014]. In addition, our pipeline: 1) Generates read counts only once, either from BAM or CRAM; 2) Runs the three methods in parallel; 3) Merges calls by a user-defined overlap percentage and a size threshold; 4) Provides annotation such as gene names in the regions and call frequencies.

## Materials and methods

**Sample selection:**
- 1,633 BHCMG samples with WES data.

**Sequencer and reagents:**
- Exome Capture: Agilent SureSelect HumanAllExonV4 or V5 plus clinical content.
- 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.

**Exome CNV calling programs:**
- ExomeDepth, XHMM, CANOES and CODEX.

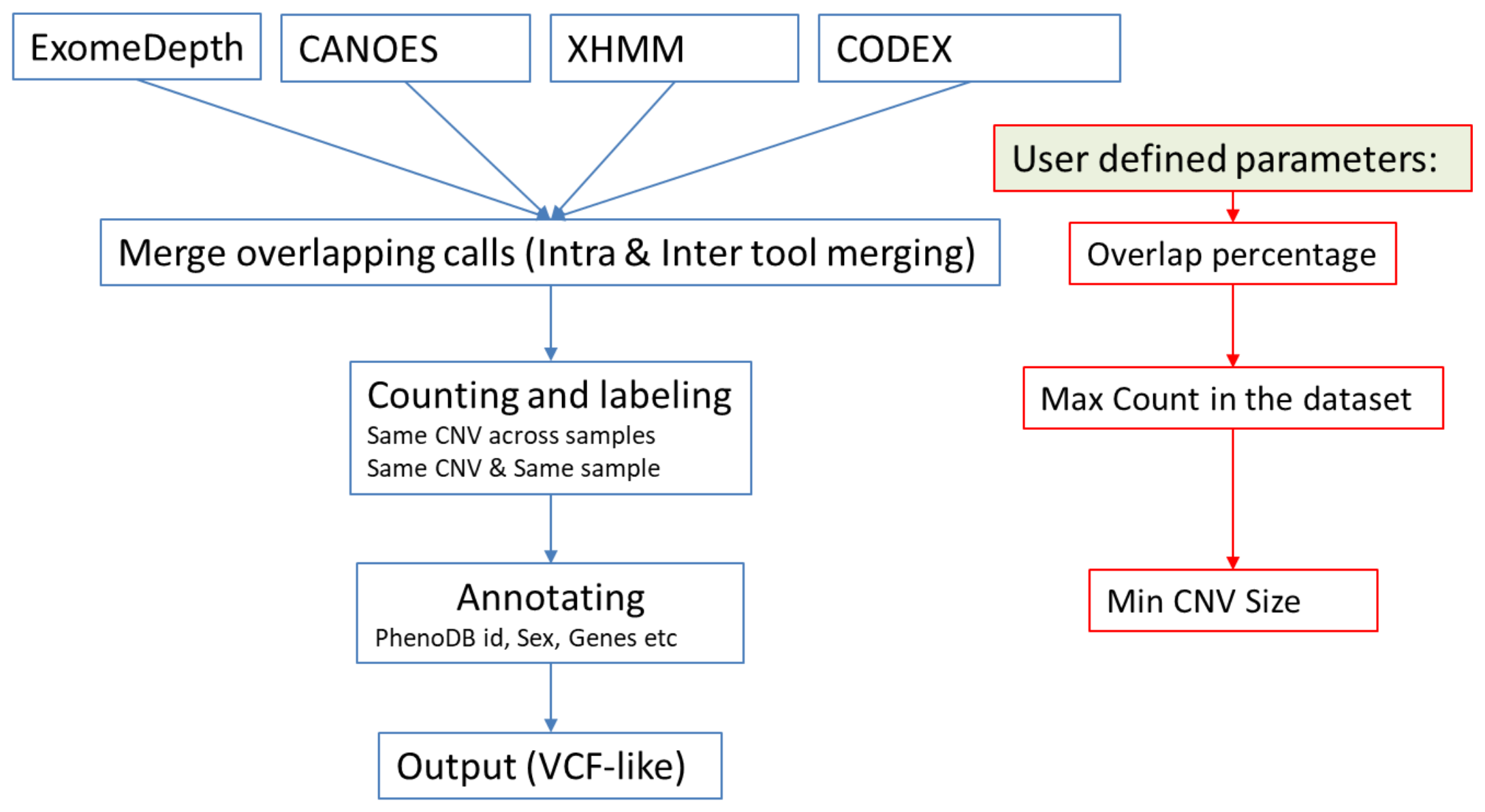
## Results

**The ECO algorithm:**
First pull the results across individuals and programs all together, then separate them by chromosomes and sort the calls by decreasing lengths, then for each chromosome, beginning 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 each call has been assigned with an ID.
**The pipeline:** Python job control scripts adapted from UW-GAC QCpipeline (https://github.com/UW-GAC/QCpipeline)

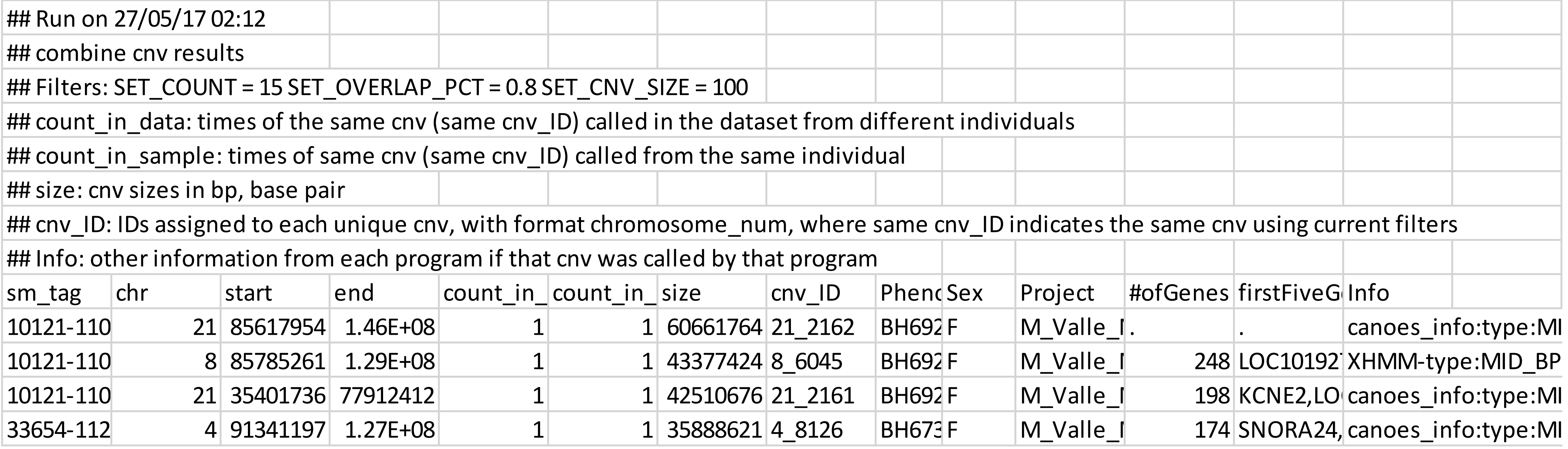
CNV call summary:

| | ExomeDepth | XHMM | CANOES | CODEX |
|---|---|---|---|---|
| Total | 188,703 | 38,952 | 10,607 | 259,146 |
| Median | 85 | 22 | 6 | 43 |
| Range | 30 ~ 1113 | 5 ~ 276 | 1 ~ 97 | 16 ~ 8650 |
| #calledSamples | 1633 | 1563 | 1603 | 1633 |

ECO flow chart:



An example output:


## Run on 27/05/17 02:12
## combine cnv results
## Filters: SET_COUNT = 15 SET_OVERLAP_PCT = 0.8 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_ | count_in_ | size | cnv_ID | Phenc | Sex | Project | #ofGenes | firstFiveG | Info |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 10121-110 | 21 | 85617954 | 1.46E+08 | 1 | 1 | 60661764 | 21_2162 | BH692 | F | M_Valle_I | . | . | canoes_info:type:MI |
| 10121-110 | 8 | 85785261 | 1.29E+08 | 1 | 1 | 43377424 | 8_6045 | BH692 | F | M_Valle_I | 248 | LOC10192 | XHMM-type:MID_BP |
| 10121-110 | 21 | 35401736 | 77912412 | 1 | 1 | 42510676 | 21_2161 | BH692 | F | M_Valle_I | 198 | KCNE2,LO | canoes_info:type:MI |
| 33654-112 | 4 | 91341197 | 1.27E+08 | 1 | 1 | 35888621 | 4_8126 | BH673 | F | M_Valle_I | 174 | SNORA24, | canoes_info:type:MI |

## Summary

- We proposed a pipeline, 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.
- Read counts generated can be used my multiple programs, no need to go back to BAM/CRAM files.
- This framework can be extended to include results from other programs such as and HMZDelFinder [Gambin et al. 2017] and EXCAVATOR [Magi et al. 2013] and whole genome sequencing.
- Future work includes further refinement of calls from each individual program and cross validation across programs.