Evaluation of Copy Number Variation (CNV) detection methods in whole exome sequencing (WES) data

Peng Zhang[1], Hua Ling[1], Elizabeth Pugh[1], Kurt Hetrick[1], Dane Witmer[1], Nara Sobreira[2], David Valle[2], Kimberly Doheny[1]

[1]Center for Inherited Disease Research, Institute of Genetic Medicine, The Johns Hopkins School of Medicine

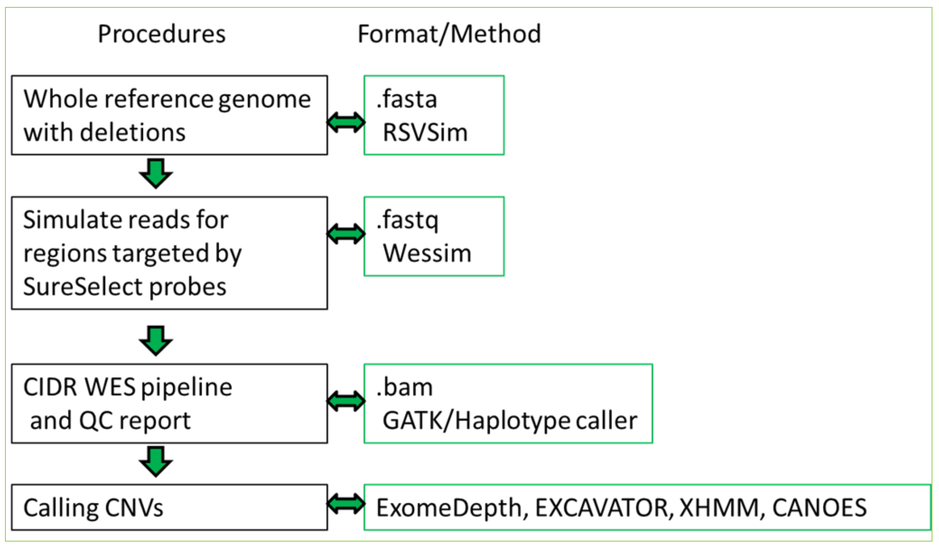
[2]Institute of Genetic Medicine, Johns Hopkins University School of Medicine

Introduction

As part of the Baylor Hopkins Center for Mendelian Genomics (BHCMG) (http://www.mendelian.org/), CIDR has performed whole exome sequencing (WES) for over 1000 samples to discover the genetic basis for Mendelian conditions. However, calling Copy Number Variants (CNVs) from WES is a challenge because of the sparseness of the target regions and the non-uniform distribution of reads across genome. As a result, the concordance across different methods is usually low, and it's often hard to evaluate the CNV calls from real sequencing data. We present here a simulation framework for evaluating the performance of ExomeDepth, EXCAVATOR, XHMM and CANOES. We simulated WES data with CNVs of various sizes and frequencies. We then evaluate the sensitivity and specificity of each CNV calling programs under different scenarios.

Materials and methods

Figure 1: Software and procedures in simulating WES data.



Simulation design:

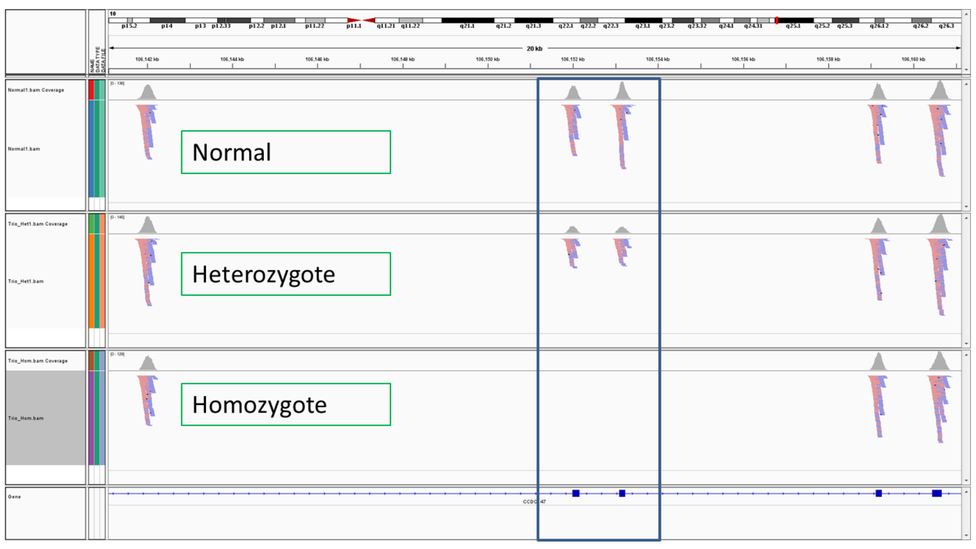
We simulated a total of 50 samples that carry each of the five sets of deletions in the whole genome reference sequence including:

1): **Singleton**, 1 sample with this set of homozygous deletions. 2): **Doubleton**, 2 samples, both are homozygous deletions. 3): **Trio**, 1 sample with homozygous deletions, 2 with heterozygous deletions. 4): **Pct10**, 1 sample with homozygous deletions, 8 samples with heterozygous deletions. 5): **Pct20**, 2 samples with homozygous deletions, 16 samples with heterozygous deletions. The rest of 17 samples are normal samples with no deletions.

Size: 50 bp to 50 Kb

Mean target coverage: 110X

Figure 2: Example of simulated WES data with two-exon deletion shown in IGV (Chr10: 106151892-106157746).



Running the CNV calling methods:

ExomeDepth can select reference control from a set of sample based on pair-wise correlation of read counts. We ran EXCAVATOR using the first five samples ExomeDepth picked for controls. For XHMM, we added 250 real WES samples from BHCMG to reach an appropriate sample size. For CANOES, we ran the 50 simulated samples alone.

Results

Table 1: Detection of homozygous deletions.

		Singleton (n = 1066)	Doubleton (n = 1193)	Trio (n = 1143)	Pct10 (n = 1123)	Pct20 (n = 1105)
ExomeDepth	sensitivity	**0.977** 1041/1066	**0.873** 1042/1193	**0.899** 1027/1143	**0.963** 1081/1123	**0.896** 990/1105
	specificity	0.974 931/956	0.951 962/1012	0.993 943/950	0.982 998/1016	0.996 938/942
EXCAVATOR	sensitivity	0.374 399/1066	0.372 444/1193	0.368 421/1143	0.391 439/1123	0.381 421/1105
	specificity	**0.993** 289/291	**0.982** 320/326	**0.997** 311/312	**1.0** 298/298	**1.0** 292/292
XHMM	sensitivity	0.443 472/1066	0.425 507/1193	0.430 492/1143	0.295 331/1123	0.02 23/1105
	specificity	0.930 413/444	0.937 417/445	0.936 409/437	0.894 270/302	0.351 20/57
CANOES	sensitivity	0.90 959/1066	0	0.524 599/1143	0.902 1013/1123	0.033 36/1105
	specificity	0.952 893/938	0	0.987 520/527	0.976 950/973	0.939 31/33

Table 2: Detection of heterozygous deletions.

		Trio (n = 1143)	Pct10 (n = 1123)	Pct20 (n = 1105)
ExomeDepth	sensitivity	0.381 435/1143	0.001 1/1123	0 0/1105
	specificity	0.995 412/414	0.50 1/2	0 0/1
EXCAVATOR	sensitivity	0.01 12/1143	0	0
	specificity	0.833 10/12	0	0
XHMM	sensitivity	0.226 258/1143	0.034 38/1123	0
	specificity	0.936 225/243	0.944 34/36	0
CANOES	sensitivity	0	0	0
	specificity	0	0/3	0/17

		Trio_Het (n = 1143)	Pct10_Het (n = 1123)	Pct20 (n = 1105)
ExomeDepth	sensitivity	0.787 899/1143	0.816 916/1123	0.807 892/1105
	specificity	0.964 832/863	0.959 861/898	0.963 839/871
EXCAVATOR	sensitivity	0.267 305/1143	0.265 298/1123	0.257 284/1105
	specificity	0.968 210/217	0.976 201/206	0.990 197/199

Table 3: Better sensitivity and specificity for calling heterozygous deletions by selecting normal samples as controls.

Summary

For homozygous deletions:

- ExomeDepth has the best sensitivity and higher specificity in general; EXCAVATOR has the best specificity but not high sensitivity.
- XHMM have high specificity only for less common deletions.
- CANOES can have high sensitivity and specificity for rare deletions, but may not be stable.

For heterozygous deletions:

- All the methods have lower sensitivity for detecting heterozygous deletions, while ExomeDepth generated the best results.
- High specificity was only observed for rare heterozygous deletions.

Discussions:

- ExomeDepth produced the best performance overall in this simulated dataset but we are aware the simulation setting is in favor of ExomeDepth, as there are a lot of small deletions and EXCAVATOR is designed to detect larger CNVs. In addition, EXCAVATOR can differentiate between heterozygous and homozygous deletions. And we added 250 real WES samples in order to run XHMM, which may introduce noise to its CNV calling.
- XHMM and CANOES are more affected by the frequency of deletions in the dataset than ExomeDepth and EXCAVATOR.
- **Reference selection:** Using normal samples as reference significantly increase the sensitivity and specificity calling heterozygous deletions, but not for homozygous deletions.
- **Still need to check how the methods affected by CNV sizes.**