Bioinformatics, 37(14), 2021, 2055–2057
doi: 10.1093/bioinformatics/btab199
Advance Access Publication Date: 24 March 2021
Applications Note

Genome analysis

muCNV: genotyping structural variants for population-level sequencing

Goo Jun [1,*], Fritz Sedlazeck[2], Qihui Zhu[3], Adam English[2], Ginger Metcalf[2], Hyun Min Kang[4], Human Genome Structural Variation Consortium (HGSVC)[2], Charles Lee[3], Richard Gibbs[2] and Eric Boerwinkle[1]

[1]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA, [3]Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA and [4]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.
Associate Editor: Peter Robinson

Received on July 10, 2020; revised on January 31, 2021; editorial decision on XXXX X, 2021; accepted on March 13, 2021

Abstract

Motivation: There are high demands for joint genotyping of structural variations with short-read sequencing, but efficient and accurate genotyping in population scale is a challenging task.
Results: We developed muCNV that aggregates per-sample summary pileups for joint genotyping of >100 000 samples. Pilot results show very low Mendelian inconsistencies. Applications to large-scale projects in cloud show the computational efficiencies of muCNV genotyping pipeline.
Availabilityand implementation: muCNV is publicly available for download at: https://github.com/gjun/muCNV.
Contact: goo.jun@uth.tmc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent advances in high-throughput sequencing had enabled population-scale analysis of whole genome sequencing data, but most of analyses have been focused on small variants. Structural variations (SVs) potentially have significant functional implications (Chiang *et al.*, 2017; Conrad *et al.*, 2010), but genome-wide landscape and genetic contributions of such variants are still largely unknown. There are still challenges in efficient and accurate genotyping of SVs from short-read sequencing data, especially at large scale. Challenges are mainly two-folds: controlling false discoveries and managing computational efficiencies.

A plethora of SV detection methods has been developed utilizing read depth, discordant read pairs, split reads and soft-clipped read information of short-read sequencing data. Each algorithm captures unique characteristics of SVs; hence has its merits in specific SV types and size regimes. While there is not a single gold-standard tool for SV discovery, it has been reported that merging outputs from multiple tools provide better sensitivity (Zarate et al., 2020). Many SV discovery and genotyping tools run on a single sample, so SV genotyping typically involves two steps of merging processes: merging across multiple callers on a single sample and then merging across multiple samples to construct a project-wide SV discovery set. These approaches greatly improve overall sensitivity of SV discovery, but at the same time it increases the risk of false discoveries. This is especially challenging for large-scale studies because true SVs would share the same breakpoints across different callers and samples, but we cannot expect the same for false discoveries. Even if we manage false discovery rates of a SV caller at a very low level, say 10 unique false events per sample and per caller, it could result in millions of false discoveries when merged across several different callers and across 100 000 samples, order-of-magnitude more than the number of true SVs. A jointly genotyped set is crucial for the study of genetic associations of SVs in large scale.

Joint genotyping of a variant requires accessing sequencing data across all samples; hence computationally challenging. Simply accessing sequencing data across large sample size is expensive; 100 000 whole genome CRAM files at 30× sequencing takes several petabytes of storage. These are typically stored in the cloud and a concurrent access of all samples is not practical. To address this issue, GATK, a popular short variant caller, adopted gVCF format that summarizes possible variant information with summary of reference alleles (Poplin *et al.*, 2018), similar to pileup data of GotCloud (Jun *et al.*, 2015). These summary files are useful for short variants but do not have enough information for SV genotyping.

We present muCNV, a software for joint genotyping of SVs in large-scale sequencing studies. Our approach utilizes both read depth and read mapping information and improves genotyping accuracy by utilizing multi-sample statistics from population-level data. It also provides an efficient workflow that can be easily parallelized in the clouds by generating sample-level pileup data and using the pileup for joint genotyping. The strength of our approach is separating SV discovery

from genotyping to incorporate multiple external SV discovery tools that have complementary strengths to each other.

2 Materials and methods

Genotyping by muCNV consists of two major steps: (i) generating summary pileups with read depth and read pair information and (ii) joint genotyping of candidate events across all samples from the pileups. Currently muCNV supports genotyping of deletions, duplications, multi-allelic CNVs and inversions in autosomes. The overall workflow of muCNV pipeline is shown in Supplementary Figure S1.

The pileup step takes list of candidate SV events generated by external callers as input and generate three files to summarize SV-related features from each sample's CRAM file in a single scan: pileup, var and idx files. The pileup file contains summary of sequencing statistics (depth, insert size, GC-curve), strand and position of discordant read pairs, split reads and soft clips, and average sequencing depth per each 100-bp genomic region for the entire genome. The var file contains average sequencing depth for each candidate SV event and the idx file contains index information for the pileup file to enable random access using genomic coordinates. These pileup files can be merged across samples in arbitrary numbers to prevent concurrent handling of tens of thousands of files in the genotyping step. The merging step is optional but highly recommended for projects with more than 10 000 samples.

The genotyping step reads all (merged) pileups and determine whether each candidate SV is polymorphic or not. For each candidate SV, muCNV tries to refine breakpoints first from the consensus of split read information across all samples and collect counts of split reads, discordant read pairs and soft clips around the refined breakpoints. Collected counts are then used together with the read depth distribution to fit two-dimensional mixture of Gaussian distributions, where one dimension is normalized depth and the other dimension is fraction of reads with evidences (discordant read pairs, split reads and soft clips). The number of components in the mixture model is determined by Bayesian information criterion (BIC) and overlap between mixture components. The mixture model is then used for genotype (and copy number) assignments. When there are not enough number of split reads for breakpoint refinements, muCNV utilizes soft clips and then reported breakpoints from the candidate SV event. If the two-dimensional Gaussian model fails to call genotypes, muCNV also tries to fit Gaussian mixtures using depth information only, followed by call rate and Hardy-Weinberg equilibrium (HWE) based filtering for common variants. For inversions, muCNV applies one-dimensional Gaussian mixture model with fraction of reads with evidences only.

The entire muCNV pipeline is highly parallelizable. The pileup step is sample-by-sample process and can be distributed across many computing nodes. The run-time of pileup step is dependent on the number of candidate SVs, but typically it takes 1–2 single-thread CPU hours to process a single 30× whole genome CRAM file when ran on a Google cluster node with a CRAM file stored in a bucket storage. The genotyping step is parallelized per genomic regions. Overall, the total computing cost for muCNV is slightly more than two CPU hours per sample for large-scale projects. The muCNV pipeline is currently being actively used to generate SV genotypes in conjunction with Parliament2 multi-discovery pipeline (Zarate *et al.*, 2020) on population-scale (>100 000) sequencing projects including the NHLBI TOPMed project, where initially tens of millions of candidate SV calls were generated but only a small fraction of these candidate SVs could be jointly genotyped and used for genotype–phenotype analyses.

3 Results

To assess genotyping accuracies with publicly available data, we ran a benchmark analysis by running muCNV on Illumina 30× whole genome sequences from the 1000 Genomes Project (1000G). We used 168 samples from the Yoruba (YRI) trios, selected from the union of the 2504 unrelated individuals from 1000G Phase 3 release (Sudmant *et al.*, 2015) and 698 additional individuals that are family (trio) members of the 2504. The candidate SV set was generated by merging SVs from 14 different callers for each sample by FusorSV (Becker *et al.*, 2018) and then merging across 2504 individuals using SURVIVOR (Jeffares *et al.*, 2017) with 200-bp merging interval. We used muCNV version 1.0.0 version.

Table 1. Summary of SVs with Mendelian error (ME) rates genotyped from 168 YRI trio samples

	No. of SV		No. of per sample		ME rate (%)	
	All	1KG new	All	1KG new	All	1KG new
Deletion	16 883	6156	3294	976	0.39	0.65
Duplication	2138	1866	194	180	2.5	2.5
Inversion	438	180	66	26	0.58	0.81
CNV	400	373	146	135	–	–

Note: 1KG new means SVs not overlapping 1000G SVs.

In total, muCNV genotyped 19 859 SVs in 168 individuals. On average, an individual had 3294 deletions, 194 bi-allelic duplications, 146 CNVs and 66 inversions (Table 1). We assessed genotyping accuracies by Mendelian inconsistencies. Deletions had the lowest non-reference error rate at 0.39%, followed by inversions at 0.58% and bi-allelic duplications at 2.5% (Supplementary Tables S1–S3). Compared to reported error rate of 2–3% for 'high confidence' calls of svtools (Larson *et al.*, 2019) and 3.8% of gnomAD-SV (Collins *et al.*, 2020), our result show much lower overall error rate. We evaluated false discovery rate by measuring fraction of *de novo* heterozygous genotypes in children, which was very low at 0.6%. More details on the comparison, additional experiments and evaluation are in Supplementary Text.

We also compared our result to 1000G Phase 3 SVs and we identified 8576 additional SVs that do not overlap with the 1000G data similar Mendelian inconsistency rates to the overall data while maintaining 80% sensitivity (Supplementary Fig. S2), meaning that most of these 1000G-novel SVs are likely to be true variants. The results show that muCNV provides an efficient and accurate genotyping pipeline for multi-sample SV analyses.

Funding

This work was supported by the National Institutes of Health [1R01DK118631, 1R03HD098552, 5UM1HG008898-04, 1OT2OD00 2751-01 and HHSN26817HV00002R]. The sequencing data were generated at the New York Genome Center with funds provided by NHGRI [3UM1HG008901-03S1 and 3UM1HG008901-04S2].

Conflict of Interest: none declared.

Data availability

The sequencing data used in this paper is available from the International Genome Sample Resource.

References

Becker,T. *et al.* (2018) FusorSV: an algorithm for optimally combining data from multiple structural variation detection methods. *Genome Biol.*, **19**, 38.

Chiang,C. *et al.*; GTEx Consortium. (2017) The impact of structural variation on human gene expression. *Nat. Genet.*, **49**, 692–699.

Collins,R.L. *et al.*; Genome Aggregation Database Consortium. (2020) A structural variation reference for medical and population genetics. *Nature*, **581**, 444–451.

Conrad,D.F. *et al.*; Wellcome Trust Case Control Consortium. (2010) Origins and functional impact of copy number variation in the human genome. *Nature*, **464**, 704–712.

Jeffares,D.C. *et al.* (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.*, **8**, 14061.

Jun,G. *et al.* (2015) An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.*, **25**, 918–925.

Larson,D.E. *et al.* (2019) svtools: population-scale analysis of structural variation. *Bioinformatics*, **35**, 4782–4787.

Poplin,R. *et al.* (2018) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv, 201178.

Sudmant,P.H. *et al.*; 1000 Genomes Project Consortium. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.

Zarate,S. *et al.* (2020) Parliament2: fast structural variant calling using optimized combinations of callers. *GigaScience*, **9**, 1–9.