Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Application note

Canvas SPW: calling *de novo* copy number variants in pedigrees

Sergii Ivakhno[1, *,†], Eric Roller[2, †], Camilla Colombo[1], Philip Tedder[1] and Anthony J. Cox[1]

[1]Illumina Cambridge Ltd., Chesterford Research Park, Little Chesterford, Essex CB10 1XL and [2]Illumina Inc., 5200 Illumina Way, San Diego, CA 92122,

*To whom correspondence should be addressed. [†]These authors contributed equally to the manuscript.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Whole genome sequencing is becoming a diagnostics of choice for the identification of rare inherited and *de novo* copy number variants in families with various pediatric and late-onset genetic diseases. However, joint variant calling in pedigrees is hampered by the complexity of consensus breakpoint alignment across samples within an arbitrary pedigree structure.

Results: We have developed a new tool, Canvas SPW, for the identification of inherited and *de novo* copy number variants from pedigree sequencing data. Canvas SPW supports a number of family structures and provides a wide range of scoring and filtering options to automate and streamline identification of *de novo* variants.

Availability: Canvas SPW is available for download from https://github.com/Illumina/canvas.

Contact: sivakhno@illumina.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The advent of affordable high-throughput sequencing enables the characterization of genes implicated in a wide range of genetic diseases and makes it possible to provide diagnosis in clinical settings as exemplified by Genomics England 100k Genomes Project for the National Health Service (https://www.genomicsengland.co.uk). Whole genome sequencing of families is becoming a standard approach for identifying highly penetrant variants that cause rare disease, such as *de novo* or recessive mutations. Accurate variant and genotype calling is crucial for successful identification of such disease-causing mutations (Acuna-Hidalgo et al. 2016). Unfortunately, false positive and negative results can occur due to technical artifacts or reduced sequencing coverage, which especially impact copy number variants (CNVs) identified through read depth estimation (Teo et al. 2012). CNV calling accuracy in families can be improved over single-sample calling by incorporating pedigree structure into the genotyping model to ensure that copy number genotypes are consistent with Mendelian inheritance and low rates of *de novo* mutation. While a number of tools have been developed for the identification of germline CNVs from sequencing data (Boeva et al., 2011, Abyzov et al., 2011 and Liu et al., 2016), most of them are limited to variant calling in single samples. Even those designed for family-based CNV detection are restricted to only deal with parent-offspring trios (Liu et al., 2016). To expand the number of analyzable family structures and provide explicit and easy-to-interpret *de novo* variant calls, we have developed a new workflow, Canvas SPW (Small Pedigree Workflow), for germline and *de novo* variant calling in pedigrees. In addition to trios, Canvas SPW processes quads and can also perform joint variant calling in medium-size sample batches.

2 METHOD

Outline Canvas SPW comprises five distinct modules designed to (1) process aligned read data and estimate depth in coverage bins, (2) perform outlier removal and normalization of depth estimates, (3) partition bins into segments of uniform copy number, (4) calculate associated allele counts from single nucleotide variants (SNVs) and (5) assign germline and *de novo* copy numbers. While initial multi-sample data processing and normalization steps are extensions of single-sample methods described elsewhere (Roller et al., 2016), segmentation and variant calling steps have been specifically developed for multi-sample pedigree-structured inputs. We briefly describe them below. More information is available in the Supplementary Material, Section 1.

Segmentation The input to the segmentation module is a data matrix produced by processing the aligned read data from all samples. Each row of the input data

matrix represents a single genomic bin with the same genomic coordinates across all samples. In turn, each column represents the normalized depth estimate across all genomic bins for a given sample. A Hidden Markov Model (HMM) with multi-variate negative binomial emission distribution uses this matrix for partitioning. HMM hidden states are initialized to approximately follow copy number states (exact CN assignment is done at the variant calling stage). First, the Expectation Maximization algorithm is used to optimize parameters of the emission and transition distributions. Next, the Viterbi algorithm is used to derive the final partitions.

Variant calling and output For each segment determined by the HMM segmentation module, Canvas SPW uses the distribution of coverage across the various bins along with the allele-specific depths at each SNV within the segment to assign a copy number. As a rule, allele-specific depths are only used when a segment contains enough SNV loci; allele-specific depths are not use in small segments containing only a handful of SNV sites. A probabilistic model is fitted to estimate likelihood (L) of copy number (CN) assignments within a pedigree given coverage data (D) $L(CN_M, CN_F, CN_C / D) \sim P(D_M / CN_M)\, P(D_F / CN_F)\, P(D_C / CN_C) \times P(CN_C / CN_M, CN_F)$, where the last term incorporates both the Mendelian transmission probabilities and the estimated *de novo* rate of CNVs. Likelihood evaluation is done using exhaustive enumeration of all possible *CN* assignments within the pedigree up to the maximal CN threshold. A joint probability table from the model with the maximum likelihood is used to estimate single sample and *de novo* quality scores for variant calls and every variant inconsistent with Mendelian inheritance is assigned a *de novo* flag. A VCF 4.1 compliant file with common and *de novo* CNV calls is produced at the end of each Canvas SPW run.

Implementation and performance Canvas SPW is implemented in the C# programming language and can be run on Linux systems using mono/.NET Core or on Windows systems under the .NET. Using a Linux system with 32 cores and peak RAM consumption of less than 10G, the Canvas SPW runtime on a trio and a quad pedigree with 40X sequencing coverage per-sample was 1.3h and 2.1h respectively.

3 RESULTS

Two key performance characteristics of Canvas SPW were evaluated: (1) ability to accurately call inherited germline variants and (2) correct *de novo* variant detection. To accomplish this we have created a number of pedigree sequencing samples using Platinum Genomes (PG) dataset (Eberle et al., 2016) that include: (1) normal trio, (2) negative control replicates of a single sample, (3) *denovo* enriched trio and quad where parents are derived from the same sample and (4) pedigree simulation through haplotype down-sampling. The latter is an adaptation of the previously described tHapMix simulation framework for somatic variants (Ivakhno et al., 2016) to germline CNVs within a pedigree relationship structure. Truth sets were generated by merging structural variants found in PG data using orthogonal variant calling tools. Further validation of these truth sets was done by checking for full consistency with Mendelian inheritance and comparison with TruSeq synthetic long read data (full details of the assessment methodology and results are available in Supplementary Material, Section 2). Estimation of CNV calling accuracy, precision and recall was performed as previously reported (Roller et al., 2016). We assessed the performance of Canvas SPW against a range of existing CNV calling tools: Control-FREEC (Boeva et al., 2011), CNVnator (Abyzov et al., 2011) and TrioCNV (Liu et al., 2016). All tools were assessed for germline CNV calling accuracy, precision and recall. TrioCNV was also evaluated for its ability to call *de novo* variants. The assessment was done separately four on real synthetic (Table 1) and three haplotype-simulated pedigrees (Table 2).

Table 1. CNV calling performance metrics for real synthetic pedigrees

Method	Germline Variants			*De novo* variants		
	Accu-racy	Preci-sion	Re-call	Accu-racy	Preci-sion	Recall
Canvas SPW	98.91	94.21	94.31	96.12	99.21	96.82
CNVnator	91.22	85.69	85.62	NA	NA	NA
FREEC	45.43	45.43	45.43	NA	NA	NA
TrioCNV	25.87	16.77	18.89	22.17	16.41	14.59

Table 2. CNV calling performance metrics for haplotype simulated pedigrees

Method	Germline Variants			*De novo* variants		
	Accu-racy	Preci-sion	Re-call	Accu-racy	Preci-sion	Recall
Canvas SPW	92.95	92.77	94.35	99.72	96.64	98.58
CNVnator	81.22	78.21	75.12	NA	NA	NA
FREEC	38.19	38.14	73.57	NA	NA	NA
TrioCNV	21.12	10.12	12.51	21.21	13.18	12.22

Canvas SPW showed superior performance in comparison with existing tools for both inherited and *de novo* germline variants. It also outperforms the single-sample germline Canvas workflow (Supplementary Material, Table 2), suggesting that joint CNV calling with pedigree information not only enables *de novo* variants detection, but also improves performance on inherited germline variants. This is particularly true for the recall where the pooling of sequencing coverage from multiple samples amplifies signal, thereby decreasing the number of false negatives. To conclude, Canvas SPW provides fast, accurate and easy to use workflow for the identification of inherited and *de novo* germline CNV variants in pedigrees.

Acknowledgements

We thank Stephen Tanner and Lisa Murray for valuable discussions.
Conflict of Interest: All authors are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis.

References

Acuna-Hidalgo,R. *et al.* (2016) New insights into the generation and role of *de novo* mutations in health and disease. *Genome Biol.*, **17**, 241.

Abyzov,A. *et al.* (2011) CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.*, **21**, 974-984.

Boeva,V. *et al.* (2011) Control-FREEC: a tool for assessing copy number and allelic content using next generation sequencing data. *Bioinformatics*, **28**, 423-425.

Eberle,M. *et al.* (2016) A reference dataset of 5.4 million human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res.*, **27,** 157-164.

Ivakhno,S. *et al.* (2016) tHapMix: simulating tumour samples through haplotype mixtures. *Bioinformatics*, **33**, 280-282.

Liu,Y. *et al.* (2016) Joint detection of copy number variations in parent-offspring trios. *Bioinformatics*, **32**, 1130–1137.

Roller,E. *et al.* (2016) Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*, **32**, 2375-2377.

Teo,SM. *et al.* (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, **28**, 2711-2718.