Bioinformatics, 32(15), 2016, 2375–2377
doi: 10.1093/bioinformatics/btw163
Advance Access Publication Date: 24 March 2016
Applications Note

Genome analysis

Canvas: versatile and scalable detection of copy number variants

Eric Roller[1,*,†], Sergii Ivakhno[2,†], Steve Lee[1,†], Thomas Royce[3,†] and Stephen Tanner[1,†]

[1]Illumina Inc, San Diego, CA 92122, USA, [2]Illumina Cambridge Ltd, Chesterford Research Park, Little Chesterford, Essex CB10 1XL, UK and [3]Ashion Analytics, Phoenix, AZ, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, all authors should be regarded as joint First Authors.

Associate Editor: John Hancock

Received on January 5, 2016; revised on March 18, 2016; accepted on March 21, 2016

Abstract

Motivation: Versatile and efficient variant calling tools are needed to analyze large scale sequencing datasets. In particular, identification of copy number changes remains a challenging task due to their complexity, susceptibility to sequencing biases, variation in coverage data and dependence on genome-wide sample properties, such as tumor polyploidy or polyclonality in cancer samples.
Results: We have developed a new tool, Canvas, for identification of copy number changes from diverse sequencing experiments including whole-genome matched tumor-normal and single-sample normal re-sequencing, as well as whole-exome matched and unmatched tumor-normal studies. In addition to variant calling, Canvas infers genome-wide parameters such as cancer ploidy, purity and heterogeneity. It provides fast and easy-to-run workflows that can scale to thousands of samples and can be easily incorporated into variant calling pipelines.
Availability and Implementation: Canvas is distributed under an open source license and can be downloaded from https://github.com/Illumina/canvas.
Contact: eroller@illumina.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The increased throughput of sequencing studies has created high demand for versatile and scalable tools to detect somatic and germline copy number changes. Increasingly complex experimental designs require accurate characterization not only of individual copy number variants (CNVs), but also of global genome and sample properties, such as ploidy, normal contamination and polyclonality (frequently present in cancer samples, Navin *et al.*, 2010). The interaction of these factors creates an array of different somatic genome architectures that confounds optimization of CNV calling algorithms given an often limited availability of training data. While a number of methods for CNV identification have been introduced, most of them harbor shortcomings when it comes to scalability and throughput. First, many tools rely on external segmentation software that complicates workflow management and version control. Second, model parameters to infer global genome and sample properties are often hard-coded and difficult to optimize for individual projects. Finally, many tools require the user to provide pre-specified sample-specific parameter values that might not be known prior to CNV detection. For example, TITAN (Ha *et al.*, 2014) and FREEC (Boeva *et al.*, 2011), while inferring heterogeneity and normal contamination respectively, require genome-wide ploidy values as an input, which incorporates a manual step in the variant calling workflow and complicates automation.

We have developed a new tool for CNV calling, Canvas, to address the aforementioned limitations of existing solutions. It fully implements all steps of the variant calling workflow and requires only aligned sequence data and related reference genome files as input. Canvas offers inference of global tumor genome and sample characteristics, including ploidy, contamination and heterogeneity, as well as loss of heterozygosity. Versatility is attained by offering fast and easy-to-run whole-genome and exome workflows for both somatic and germline variants. This combined functionality makes Canvas a favorable tool for somatic and germline CNV detection in large-scale sequencing studies.

2 Method

2.1 Outline

The Canvas workflow comprises five distinct modules designed to (i) process aligned read data and calculate coverage bins, (ii) perform outlier removal and normalization of coverage estimates, (iii) identify segments of uniform copy number, (iv) calculate minor allele frequencies (MAF) and (v) assign copy number/allelic states and infer genome-wide parameters. Separate workflows exist for somatic, germline and exome sequencing data. The latter workflow can be run either with or without a matched normal control sample and requires a manifest file with locations of targeted regions. The somatic workflow includes logic to cope with normalization of FFPE samples. Detailed step-by-step explanations can be found in Supplementary Methods.

2.2 Implementation and performance

Canvas is implemented in C# programming language and can be run on Linux system using mono or on Windows under the .NET framework. A full per-chromosome parallelization is available for all time-consuming modules. An average Canvas runtime for the tumor/normal workflow on 80×/40× coverage matched sample pair using a Linux node with 32 CPUs is 40 min (70 min when fragment-based GC-content normalization is invoked); with a peak RAM consumption of under 6 GB.

3 Results

Both simulated and real data was used to assess Canvas performance in each workflow category and to enable comparison with alternative CNV calling tools. The selection of third-party methods was based on the principle that they either showed superior performance in the previous benchmarks (as in the case of FREEC, Alkodsi *et al.*, 2015) or that they offered inference of a number of different genome-wide parameters, like THetA (Oesper *et al.*, 2013). We did not aim to perform a comprehensive evaluation of CNV callers, as this was covered elsewhere (Alkodsi *et al.*, 2015; Nam *et al.*, 2015). What follows is an overview of the test data generation strategies and performance results. Supplementary Results provide full details of simulation and evaluation strategies.

3.1 Simulation data

A haplotype mixing workflow was used to generate simulation data for both germline and somatic workflows. Briefly, aligned reads were split into haplotypes inferred from phasing of the Platinum Genomes (PG) family (http://www.platinumgenomes.org). These reads along with manually curated CNV calls from previously sequenced genomes at Illumina were used to create truth sets. The simulation also included parameters to generate tumors of different purity and polyclonality levels.

3.2 Cell lines

For evaluating *somatic* CNV calling, breast carcinoma cell lines HCC2218 and HCC1187 were sequenced to 80× coverage on HiSeq2000 along with matching normal lymphoblastoid cell lines (sequenced to an average of 40×). Exome samples of the same cell lines were prepared using Nextera Rapid Capture Exome reagent kit, and sequenced on a HiSeq 2500. Read titration with normal lymphoblastoid samples was used to approximate different purity levels. A truth set for these cell lines was created by comparison with karyotype data (Newman *et al.*, 2013) and by manual inspection of coverage, allele ratios and read mappings. To benchmark the *germline* workflow, a reference CNV call set for NA12878 individual from the PG family was created by selecting the pedigree-consistent set of deletion calls made using a range of structural variant calling tools (Supplementary Results, Section S1.2).

Table 1. Somatic CNV calling performance metrics (average values across all samples for each workflow type, best preforming metrics in each category are highlighted in bold)

Method	Cell lines			Simulation		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Canvas	**85.32**	**72.80**	67.21	**84.54**	**77.87**	68.04
THetaA	41.22	15.69	25.62	28.89	16.29	28.27
TITAN	73.81	65.62	**76.12**	68.72	59.01	**68.32**
FREEC	65.87	42.71	28.89	71.34	49.78	40.92

Table 2. Germline CNV calling performance metrics (average values across all samples for each workflow type, best preforming metrics in each category are highlighted in bold)

Method	NA12878			Simulation		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Canvas	**78.65**	**96.19**	**78.34**	**93.08**	**98.42**	**92.91**
FREEC	51.86	67.16	51.86	63.67	97.86	63.86
CNVnator	61.91	76.75	62.01	81.22	82.53	80.73

3.3 Evaluation strategy and results

We have focused on exploring concordance between expected and observed copy numbers to calculate accuracy, precision and recall. Tables 1 and 2 show average performance metrics for somatic and germline workflows respectively, while Supplementary Table S4 shows similar performance figures for the exome workflow.

When considering all of the metrics, Canvas attained the highest levels of accuracy among tools used in the comparison for both germline and somatic workflows and across real and simulated datasets. Similar observations were made for the exome workflow, where Canvas outperformed EXCAVATOR (Magi *et al.*, 2013) and ADTEx (Amarasinghe *et al.*, 2014) when considering original HCC cell lines and the purity titration series (Supplementary Results, Tables S6 and S7). Canvas also showed the fastest runtime for whole genome somatic and germline workflows, completing on average 2. 3 times faster than a runner-up tool (Supplementary Results, Table S8). To conclude, Canvas is a versatile tool for CNV identification with superior performance across a range of whole-genome and exome sequencing experiments and fast runtime.

Acknowledgements

We thank Semyon Kruglyak and Anthony J. Cox for valuable discussions.

Conflict of Interest: ER, SI, SL and ST are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis. Sequencing

data for HCC2218 and HCC1187 is being shared in accordance with the terms of a licensing agreement with UT Southwestern, the owners of the cell lines.

References

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.

Alkodsi,A. *et al*. (2015) Comparative analysis of methods for identifying somatic copy number alterations from deep sequencing data. *Brief*. *Bioinf*., **16**, 242–254.

Amarasinghe,K.C. *et al*. (2014) Inferring copy number and genotype in tumour exome data. *BMC Genomics*, **15**, 732.

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.

Ha,G. *et al*. (2014) TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res*., **24**, 1881–1893.

Magi,A. *et al*. (2013) EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol*., **14**, R120.

Nam,J. *et al*. (2015) Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief*. *Bioinf*., **17**, 185–192.

Navin,N.E. (2010) Tracing the tumor lineage. *Mol. Oncol*., **4**, 267–283.

Newman,S. (2013) The relative timing of mutations in a breast cancer genome. *PLoS One*, **8**, e64991.

Oesper,L. *et al*. (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol*., **14**, R80.