Bioinformatics, 32(8), 2016, 1220–1222
doi: 10.1093/bioinformatics/btv710
Advance Access Publication Date: 8 December 2015
Applications Note

Sequence analysis

Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications

Xiaoyu Chen[1], Ole Schulz-Trieglaff[2], Richard Shaw[2], Bret Barnes[1], Felix Schlesinger[1], Morten Källberg[2], Anthony J. Cox[2], Semyon Kruglyak[1] and Christopher T. Saunders[1,*]

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

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on 11 June 2015; revised on 8 October 2015; accepted on 1 December 2015

Abstract

Summary: We describe Manta, a method to discover structural variants and indels from next generation sequencing data. Manta is optimized for rapid germline and somatic analysis, calling structural variants, medium-sized indels and large insertions on standard compute hardware in less than a tenth of the time that comparable methods require to identify only subsets of these variant types: for example NA12878 at 50× genomic coverage is analyzed in less than 20 min. Manta can discover and score variants based on supporting paired and split-read evidence, with scoring models optimized for germline analysis of diploid individuals and somatic analysis of tumor-normal sample pairs. Call quality is similar to or better than comparable methods, as determined by pedigree consistency of germline calls and comparison of somatic calls to COSMIC database variants. Manta consistently assembles a higher fraction of its calls to base-pair resolution, allowing for improved downstream annotation and analysis of clinical significance. We provide Manta as a community resource to facilitate practical and routine structural variant analysis in clinical and research sequencing scenarios.
Availability and implementation: Manta is released under the open-source GPLv3 license. Source code, documentation and Linux binaries are available from https://github.com/Illumina/manta.
Contact: csaunders@illumina.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Whole genome and enrichment sequencing is increasingly used for discovery of inherited and somatic genome variation in clinical contexts, however tools for rapid discovery of structural variants (SVs) and indels in this scenario are limited. We address this gap with Manta, a novel method for accurate discovery and scoring of SVs, medium-sized indels and large insertions in a unified and rapid process. Manta discovers variants from a sequencing assay's paired and split-read mapping information using an efficient parallel workflow. Many advanced structural variant methods are available which focus on research and population genomics (Layer *et al.*, 2014; Rausch *et al.*, 2012; Sindi *et al.*, 2012; Ye *et al.*, 2009). However, none to our knowledge combine as many variant types into a rapid workflow focused on individual or small sets of related samples. Per its focus on clinical pipelines, Manta provides a complete solution for discovery, assembly and scoring using only a reference genome and alignments from any standard read mapper. It provides scoring models for germline analysis of diploid individuals and somatic analysis of tumor-normal sample pairs, with additional applications under development for RNA-Seq, *de novo* variants, and unmatched tumors. We describe

Manta's methods and compare with representative tools to demonstrate high variant call quality with dramatically reduced compute cost.

2 Methods

2.1 Workflow summary

Manta's workflow is designed for high parallelization on individual or small sets of samples. It operates in two phases: first a graph of all breakend associations within the genome is built, then the components of this graph are processed for variant hypothesis generation, assembly, scoring and VCF reporting. The breakend graph contains edges between any genomic regions where evidence of a long range adjacency exists, indel assembly regions are denoted in this scheme as self-edges. The graph does not express specific variant hypotheses so it is very compact, and can be constructed from segments of the genome in parallel. Following graph construction, individual edges (or larger subgraphs) are analyzed for variants in parallel. Each edge is analyzed to find imprecise variant hypotheses, for which variant reads are assembled and aligned back to the genome. Assembly is attempted for all cases, but is not required to report a variant. All paired and split-read evidence is consolidated to a quality score under either a germline or somatic variant model, and filtration metrics complement this quality score to improve call precision. For ease of use, Manta automates estimation of insert size distribution and exclusion of high depth reference compression regions. Details of all workflow components are provided in Supplementary Methods.

2.2 Variant call evaluation

We assess accuracy of germline calls by running variant callers on all members of CEPH pedigree 1463, selecting for calls with pedigree-consistent genotypes and evaluating each caller on one pedigree member (NA12878) against the pedigree-consistent call set. To find pedigree-consistent calls and provide a relative recall comparison for Manta, we select a standard recognized caller in each variant class: Pindel (Ye *et al.*, 2009) for indels and Delly (Rausch *et al.*, 2012) for SVs. Calls from each representative method are used to establish the pedigree-consistent call set together with Manta's. For somatic calls we also use Delly as a standard benchmark and compare calls from both methods for the HCC1954 breast cancer cell line compared to its matched normal cell line (HCC1954BL). These calls are compared to somatic variant entries for HCC1954 in COSMIC v70 (Forbes *et al.*, 2015). Full details of the evaluation procedure are included in Supplementary Methods.

3 Results

We describe NA12878 variant call performance in the top portion of Table 1, comparing the results of each method to pedigree-consistent calls for this sample (see Methods). The first section describes large deletions and duplications, showing that Manta's results are competitive overall and have a somewhat higher recall (or higher rate of pedigree consistency due to correct genotyping). Manta calls consistently show a higher fraction of calls agreeing with the pedigree-consistent set which also have breakends assembled to base-pair resolution. For deletions and insertions smaller than 500 bases, the next section of Table 1 reiterates the large SV pattern of strong performance, with a trend towards higher recall across these smaller indel variant classes.

Somatic call performance for the HCC1954/HCC1954BL tumor/normal sample pair is described in the final portion of Table 1, comparing each method's variant calls to COSMIC variant entries for HCC1954 (see Section 2). In this case, the truth set does not reflect a complete catalog of somatic variants for the cell line, however it does provide a useful relative precision estimate reflecting enrichment for known variants. Here we observe strong performance for Manta calls across all variant types with a trend towards a greater fraction of true calls assembled to base-pair resolution, consistent with germline variant observations.

Table 1. Assessment of variant call accuracy

Variant class	Method	Recall	Prec	Exact%[a]
NA12878 structural variants				
Deletions [500,1k) ($n = 153$)	Manta	**0.941**	**0.929**	**94.1**
	Delly	0.883	0.900	82.1
Deletions [1k,10k) ($n = 479$)	Manta	**0.970**	**0.964**	**95.5**
	Delly	0.873	0.959	91.5
Deletions 10k+ ($n = 33$)	Manta	**0.970**	0.568	**96.8**
	Delly	0.911	**0.688**	93.1
Duplications [500,1k) ($n = 5$)	Manta	**1.000**	**0.333**	**100.0**
	Delly	0.800	0.266	50.0
Duplications [1k,10k) ($n = 17$)	Manta	**1.000**	0.592	**100.0**
	Delly	0.764	**0.722**	76.9
Duplications 10k+ ($n = 5$)	Manta	**1.000**	**0.285**	**50.0**
	Delly	0.600	0.214	33.3
NA12878 indels				
Deletions (50,100) ($n = 417$)	Manta	**0.990**	0.650	–
	Pindel	0.440	**0.708**	–
Deletions [100,500) ($n = 1053$)	Manta	**0.983**	0.799	–
	Pindel	0.710	**0.875**	–
Insertions (50,100) ($n = 276$)	Manta	**1.000**	**0.764**	–
	Pindel	0.342	0.127	–
Insertions [100,500) ($n = 94$)	Manta	**1.000**	**0.531**	–
	Pindel	0.000	0.000	–
HCC1954 somatic structural variants				
Inversions ($n = 100$)	Manta	**0.670**	**0.351**	**97.5**
	Delly	0.660	0.322	90.0
Translocations ($n = 87$)	Manta	**0.839**	**0.271**	**97.3**
	Delly	0.322	0.179	44.4
Duplications 10k+ ($n = 60$)	Manta	0.533	**0.292**	**97.1**
	Delly	**0.550**	0.258	96.9
Deletions 10k+ ($n = 56$)	Manta	0.607	0.256	100.0
	Delly	0.607	**0.268**	100.0
Deletions [1k,10k) ($n = 12$)	Manta	0.417	**0.227**	100.0
	Delly	**0.500**	0.146	100.0

[a]Percent of true positive calls with breakends resolved to base-pair resolution. Bold text is used to highlight the larger value in each comparison.

Table 2. Compute cost evaluation

Sample	Method	Walltime (h)		Memory (Gb)	
		Parallel	Serial	Parallel	Serial
NA12878	Manta	0.327	3.764	2.351	0.233
	Manta-SV[a]	0.102	0.878	1.786	0.125
	Pindel	12.441	124.401	61.840	62.538
	Delly	3.133	6.117	11.188	6.431
HCC1954	Manta	0.852	5.486	3.445	0.244
	Manta-SV[a]	0.544	2.391	2.754	0.186
	Delly	75.911	100.648	11.614	8.540

All tests on dual Xeon E5-2680 v2 server with data on local drive. Parallel tests use all 20 cores, serial tests use 1 core. Memory columns show peak RSS.

[a]By default Manta assembles SVs and indels 8 bases and larger, Manta-SV is a custom SV-only configuration (300 bases and larger).

Table 2 summarizes runtime and memory cost for each variant caller, benchmarked in both parallel and serial modes to show workload distribution and methods efficiency. By either of these runtime or memory metrics we observe that Manta has substantially lower compute cost and turnaround time, while providing coverage of more variant types. We note that Delly is designed to parallelize primarily across, instead of within, samples, so the parallel test reflects a limited use of all server cores. When Manta is restricted to provide variant call coverage similar to Delly (variants 300 bases and larger), observed compute cost is even lower, further highlighting the efficiency of Manta's implementation relative to current methods.

Manta's approach is sufficiently flexible to support several types of sequencing assays. The primary focus for rapid analysis and large-scale SV calling has been whole genome sequencing, but Manta is routinely used to analyze exome and other enrichment-based targeted sequencing assays. The method is not designed for targeted amplicon sequencing but successful results have been reported. We additionally note that Manta has been extensively optimized to handle the shorter fragment lengths and higher chimera rates found in highly degraded FFPE samples as part of an ongoing focus on clinical sequencing workflows.

Conflict of Interest: All authors are employees of Illumina Inc., a public company that develops and markets systems for genetic analysis.

References

Forbes,S.A. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.

Layer,R.M. *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Sindi,S.S. *et al.* (2012) An integrative probabilistic model for identification of structural variation in sequencing data. *Genome Biol.*, **13**, R22.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.