Bioinformatics, 35(9), 2019, 1579–1581
doi: 10.1093/bioinformatics/bty849
Advance Access Publication Date: 9 October 2018
Applications Note

OXFORD

Sequence analysis

Pisces: an accurate and versatile variant caller for somatic and germline next-generation sequencing data

Tamsen Dunn[1,*], Gwenn Berry[1], Dorothea Emig-Agius[1], Yu Jiang[1], Serena Lei[1], Anita Iyer[1], Nitin Udar[1], Han-Yu Chuang[1], Jeff Hegarty[1], Michael Dickover[1], Brandy Klotzle[1], Justin Robbins[1], Marina Bibikova[1], Marc Peeters[2] and Michael Strömberg[1]

[1]Departments of Bioinformatics and Clinical Genomics, Illumina Inc., San Diego, CA 92122, USA and [2]Department of Oncology, Antwerp University Hospital, 2650 Edegem, Belgium

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on April 16, 2018; revised on September 28, 2018; editorial decision on October 1, 2018; accepted on October 8, 2018

Abstract

Motivation: Next-generation sequencing technology is transitioning quickly from research labs to clinical settings. The diagnosis and treatment selection for many acquired and autosomal conditions necessitate a method for accurately detecting somatic and germline variants.

Results: We have developed Pisces, a rapid, versatile and accurate small-variant calling suite designed for somatic and germline amplicon sequencing applications. Accuracy is achieved by four distinct modules, each incorporating a number of novel algorithmic strategies.

Availability and implementation: Pisces is distributed under an open source license and can be downloaded from https://github.com/Illumina/Pisces. Pisces is available on the BaseSpace™ SequenceHub. It is distributed on Illumina sequencing platforms such as the MiSeq™ and is included in the Praxis™ Extended RAS Panel test which was recently approved by the FDA.

Contact: pisces@illumina.com

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

1 Introduction

The diagnosis and treatment for many oncological conditions necessitate a method for accurately detecting somatic and germline variants (Dietel *et al.*, 2015; Dong *et al.*, 2015). Many algorithms have been developed for somatic single nucleotide variant (SNV) detection in matched tumor-normal DNA sequencing, and many algorithms have been developed for detecting germline variants, GATK being the most well-known (McKenna *et al.*, 2010). However, there is no single front runner, and different callers dominate in different situations. Particularly in the context of amplicon workflows, the standardization of variant calling pipelines remains elusive (Betge *et al.*, 2015; Horak *et al.*, 2016).

Pisces is unique primarily because it excels in the difficult and common situation where no matched normal sample exists for a given tumor sample. Pisces also performs well on germline samples. Pisces requires only aligned sequence data (BAM files) and a reference genome, and it returns a variant call file with SNVs and small indels. We present an overview of the Pisces algorithms, and compare the results to alternative small-variant calling tools.

2 Materials and methods

Pisces comprises four modules, each with a novel algorithmic strategy:

1. Pisces read stitcher: reduces noise by stitching paired reads into consensus reads.
2. Pisces Variant Caller: calls small variants, includes a collapsing algorithm to rescue variants broken up by read boundaries.

Table 1. Accuracy metrics by Variant Caller

Work flow	Dataset	Tool	SNV recall	SNV precision	Indel recall	Indel precision	#Truth Var	*F*1
Somatic	Titr	Pste	99.9	99.1	97.9	91.3	2100	97.0
	Titr	Pvc	99.9	98.4	97.9	87.2	2100	95.7
	Titr	LoFreq	99.2	91.0	99.8	72.8	2100	89.9
	Titr	VarDict	96.8	75.6	82.2	85.0	2100	84.7
	RAS	Pste	98.1	84.1	NA	NA	638	90.5
	RAS	Pvc	98.3	78.4	NA	NA	638	87.2
	RAS	LoFreq	98.3	66.7	NA	NA	638	79.5
	RAS	VarDict	98.0	66.8	NA	NA	638	79.4
Germline	VP	Pste	100.0	100.0	98.9	100.0	3376	99.7
	VP	Pvc	100.0	100.0	100.0	100.0	3376	100.0
	VP	GATK	79.2	97.0	91.0	97.1	3376	90.7
	VP	VarScan	94.3	94.5	97.8	87.7	3376	93.5
	Myl	Pste	93.6	94.8	91.4	98.8	749	94.6
	Myl	Pvc	93.6	94.8	92.6	99.0	749	94.9
	Myl	GATK	90.0	94.0	63.6	38.9	749	71.2
	Myl	VarScan	84.4	93.9	95.7	58.0	749	82.4

3. Pisces variant quality recalibrator: in the event that the variant calls overwhelmingly follow a pattern associated with thermal damage or formalin-fixed paraffin-embedded (FFPE) deamination, this step will recalculate the variant QScore given the signature of the detected noise.
4. Pisces variant phaser (Scylla): uses a read-backed greedy clustering method to assemble small variants into complex alleles.

Runtime for the Pisces Variant Caller on a 470 MB BAM (8 million reads) is 85 s. Runtime for a 2 GB BAM (60 million reads) is about 4 min. All were run with 20 threads on 2.60 GHz processors.

2.1 Testing methodology

We compared Pisces performance with the following alternative small-variant calling tools: the GATK HaplotypeCaller, LoFreq, VarDict and VarScan (Koboldt *et al.*, 2013; Lai *et al.*, 2016; Wilm *et al.*, 2012). The selection of third-party tools was based on the principle that they showed a superior performance in previous benchmarking studies (Dietel *et al.*, 2015; Horak *et al.*, 2016). Each tool chosen offers a different variant calling strategy and might be optimal in other situations. A comprehensive comparison of tools is given elsewhere (Sandmann *et al.*, 2017).

For our testing, we generated BAMs from four amplicon datasets, using the Illumina amplicon aligner, and then processed the BAMs through the variant callers. The results were assessed using the Hap.py accuracy assessment tool (https://github.com/Illumina/hap.py). The datasets were selected to include both well-characterized samples and realistic cancer samples.

2.2 Datasets

All germline testing was done using established cell line samples from individuals NA12878 and NA12877 from the Coriell Institute. High-confidence variant calls are available for these individuals via Platinum Genomes build 2016-1.0 (Eberle *et al.*, 2017). These samples were run on two different panels to produce two distinct datasets. The Variant Panel was designed to target known variants in the NA12878 and NA12877 samples, specifically for the purpose of assessing the accuracy of sequencing applications. The Myeloid Panel is a commercial panel which targets genes frequently mutated in blood cancer disorders.

The somatic datasets are as follows: the Titration dataset is a mixture of the NA12878 and NA12877 cell line samples, serially diluted to present a range of variant frequencies, observed down to 1%. The titrated samples were run with the Variant Panel, and cover the same high-confidence variants. The RAS Panel dataset was generated from a set of colorectal cancer tissue blocks which were FFPE treated and extracted 8–9 years later. Those samples were evaluated by alternate methods (Sanger sequencing and therascreen KRAS test by Qiagen) to provide a gold standard.

3 Results

In Table 1, we show average accuracy metrics by variant caller across all samples for each dataset. The *F*-score given is the average of the *F*1 for SNVs and the *F*1 for indels. Pste means the full Pisces Suite was used and Pvc means only the Pisces Variant Caller was used. In each of the four datasets, Pisces attained the highest number of best-performing metrics. For germline calling, Pisces Variant Caller alone does slightly better than the more complex pipeline. In the somatic case, best results are achieved with the full Pisces Suite. Pisces' success with respect to indel calling is due to its variant collapsing algorithm, while the stitching algorithm enabled higher accuracy for low frequency datasets. We give more discussion in the Supplementary Results section. To conclude, Pisces is an accurate tool for small-variant detection.

Acknowledgements

GATK3 was made available for use by Illumina through the generosity of the Broad Institute.

Conflict of Interest: T.D., G.B., D.E.-A., Y.J., S.L., A.I., N.U., H.Y.C., J.H., M.D., B.K., J.R., M.B. and M.S. are employees of and own stock in Illumina Inc., a public company that develops and markets systems for genetic analysis.

References

Betge,J. *et al.* (2015) Amplicon sequencing of colorectal cancer: variant calling in frozen and formalin-fixed samples. *PLoS One*, **10**, e0127146.

Dietel,M. *et al.* (2015) A 2015 update on predictive molecular pathology and its role in targeted cancer therapy: a review focusing on clinical relevance. *Cancer Gene Ther.*, **22**, 417–430.

Dong,L. *et al.* (2015) Clinical next generation sequencing for precision medicine in cancer. *Curr. Genomics*, **16**, 253–263.

Eberle,M.A. *et al*. (2017) A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res*., **27**, 157–164.

Horak,P. *et al*. (2016) Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls. *ESMO Open*, **1**, e000094.

Koboldt,D.C. *et al*. (2013) Using VarScan 2 for germline variant calling and somatic mutation detection. *Curr. Protoc. Bioinformatics*, **44**, 1541–1517.

Lai,Z. *et al*. (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*., **44**, e108.

McKenna,A. *et al*. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*., **20**, 1297–1303.

Sandmann,S. *et al*. (2017) Evaluating variant calling tools for non-matched next-generation sequencing data. *Sci. Rep*., **7**, 43169.

Wilm,A. *et al*. (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res*., **40**, 11189–11201.