bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

LoLoPicker: Detecting Low-Fraction Variants in Low-Quality Cancer Samples from Whole-exome Sequencing Data

Jian Carrot-Zhang[1,2] and Jacek Majewski[1,2]

1. Department of Human Genetics, McGill University, Montreal, Quebec, Canada; 2. McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada.

Abstract

Summary: We developed an efficient tool dedicated to call somatic variants from whole-exome sequencing (WES) data using tumor and its matched normal tissue, plus a user-defined control panel of non-cancer samples. We showed superior performance of LoLoPicker with significantly improved specificity, especially for low-quality cancer samples such as formalin-fixed and paraffin-embedded (FFPE) samples.

Implementation and Availability: The main scripts are implemented in Python 2.7.8 and the package is released at https://github.com/jcarrotzhang/LoLoPicker.

bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

The detection of tumor-only mutations remains challenging. One of the major complexities is that variants with low allelic-fraction are commonly observed in tumor samples, owing to normal tissue contamination, local copy number change and cancer heterogeneity. The difficulty of identifying those low allelic-fraction variants is magnified by the fact that sequencing technologies are imperfect and produce errors (Flickinger,M. *et al*., 2015). Moreover, technical artifacts may arise from the formalin fixation process, and therefore decrease the accuracy of calling variants from FFPE samples (Van Allen,E.M. *et al*., 2014, Williams,C *et al*., 1999).

WES has emerged as a promising tool to discover disease-causing genes. For many basic research or clinical laboratories, the number of samples being sequenced has increased dramatically. Some laboratories build their in-house database of WES data to enable them to filter out false-positive calls that are specific to library preparation, protocols, instruments, environmental factors or analytical pipeline. Such database also provides an opportunity to 1) rule out polymorphisms not reported by public database, and to 2) precisely estimate the site-specific error rates using control samples. Accurate site-specific error rate gives the advantage to increase the sensitivity of calling low-fraction, single nucleotide variants (SNVs) on sites with lower error rates, and reduce false

bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

positives on sites with high error rates. This idea has been successfully implemented for targeted re-sequencing experiments (Gerstung,M. *et al*., 2013).

However, to the best of our knowledge, there are no software able to perform low-fraction SNV calling on the WES scale. Here, we present LoLoPicker that allows users to provide a control panel, which contains normal samples underwent similar procedures as the test sample (tumor), and uses this control to estimate site-specific error across the exomes. Then, a binominal test followed by Bonferroni correction are performed to determinate whether the ratio of altered reads of the tumor variant exceeds the background error rate obtained from the control samples. Detailed description of this algorithm is provided in the Supplementary Information file.

Benchmarking Analysis

To access the performance of LoLoPicker in comparison to other variant callers, we benchmarked LoLoPicker, MuTect, VarScan2 and LoFreq against two datasets (Cibulskis,K. *et al*., 2013, Koboldt,D.C. *et al*., 2012, Wilm,A. *et al*., 2012). Somatic mutations validated by Sanger in an ovarian tumor were used as true positives. The tumor sample was mixed with its matched blood to ensure that variants were present in low allelic-fraction. For specificity, a sample that underwent WES twice in two different batches was used, and all variants called between the two batches were considered as false positives. As the results,

bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

LoLoPicker showed much better specificity, while maintained the highest sensitivity (Table 1). When reducing the coverage of variants, the sensitivity of all callers were dropped, but LoLoPicker and MuTect showed highest sensitivity.

Applying LoLoPicker to Read Data

High-quality tumor samples

Because LoLoPicker, MuTect and VarScan2 showed better performances in calling low-fraction SNVs, we then applied them on a real cancer sample with matched blood sample from a glioblastoma (GBM) patient (GBM_9). About 500 germ-line samples were used as controls. Known GBM driving mutations were identified, including mutations in *TP53*, *H3F3A*, *ATRX*, and *PIK3CA*. LoLoPicker successfully identified all of them. MuTect filtered out the *TP53* mutation because it found three reads supporting the variant in the normal sample. In LoLoPicker, the mutation was retained becasue overlapping read-pair covering same variant, meaning that they sequence variant from same DNA fragment, are counted once (Figure S4). VarScan2 did not call *PIK3CA* mutation as a high-confidence variant. In particular, the *PIK3CA* mutation showed low allelic-fraction at 6%. Again, this demonstrates that LoLoPicker has a high sensitivity of calling low-fraction SNVs. Moreover, 14 low-fraction SNVs in GBM_9 were selected for targeted re-sequencing validation. All the variants called by both LoLoPicker and MuTect were validated as true positives, whereas the ones that LoLoPicker rejected were not validated. Those included four variants with higher coverage

bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(>=5X) supporting the altered bases (Table S3). This result suggested that the specificity of LoLoPicker was improved without rejecting true positives as trade-off.

FFPE samples

Error rates across different sites vary. Site-specific error rates in low-quality samples, such as FFPE samples are much higher than high-quality samples (Figure S7). In previously published work, we showed that no recurrent mutations, other than *SMARCA4* mutations were observed in small cell carcinoma of the ovary, hypercalcemic type (SCCOHT). We therefore, tested LoLoPicker on an FFPE-SCCOHT sample. Although few somatic mutations were expected, both MuTect and VarScan2 called a large number of SNVs (502 and 143, respectively). When using germ-line samples as controls, LoLoPicker called 113 SNVs. When we switched our controls to 35 FFPE-normal tissues, only 60 variants were called. Most of the LoLoPicker rejected calls were known FFPE-induced C to T or G to A, known to be induced by the FFPE protocol, suggesting the necessity of providing a control cohort to further reduce false positive calls related to batch effects, especially FFPE-specific artifacts (Figure S8).

Discussions

LoLoPickers is a new algorithm designed to detect somatic SNVs, particularly tailored for low frequency SNVs. While LoLoPicker maintains highest sensitivity

bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of calling low-fraction variants among other programs, the specificity of LoLoPicker is dramatically improved, thus highlighting the importance of precisely measuring site-specific error rate from a larger number of control samples, rather than from a matched normal sample solely. Samples provided as additional controls are essential in estimating the background error rate. Although we expect that LoLoPicker will well handles WES data from any sequencing platforms and alignment methods, we suggest that samples processed in similar experimental protocols should be used. For example, having a panel of FFPE samples helped in filtering FFPE-specific artifacts. Compared to simply filtering out recurrent calls from the control panel, LoLoPicker's statistical framework retains sites with low-level artifacts, allowing high sensitivity. Finally, the LoLoPicker algorithm can be easily parallelized to allow the analysis against a lager number of control samples in a reasonable time. As FFPE are commonly used in clinical laboratories, our method will provide unprecedented information for analyzing FFPE samples and pave the way to apply WES into cancer clinical testing.

Acknowledgements

We thank Hamid Nikbakht, Xiaojian Shao and Rui Li for helpful discussions. Pierre Lepage from McGill University and Genome Quebec Innovation Centre for his help with targeted re-sequencing. JM is the recipient of a Canada Research Chair in Genomics.

bioRxiv preprint doi: https://doi.org/10.1101/043612; this version posted March 15, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Number of true positives and false positives called by LoLoPicker, MuTect, VarScan and LoFreq from benchmarked samples.

Tools	True Positives		False Positives
	High Coverage	Low Coverage	
LoLoPicker	18/18	9/13	3
MuTect	18/18	9/13	25
VarScan2	18/18	8/13	21
LoFreq	18/18	7/13	53

Reference

Cibulskis,K. et al. (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, **31**, 213-219.

Flickinger,M. et al. (2015) Correcting for Sample Contamination in Genotype Calling of DNA Sequence Data. *The American Journal of Human Genetics*, **97**, 284-290.

Gerstung,M. et al. (2013) Subclonal variant calling with multiple samples and prior knowledge. *Bioinformatics*, **30**, 1198-1204.

Koboldt,D.C. et al. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**, 568-576.

Van Allen,E,M. et al. (2014) Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature Method*, **20**, 682-688.

Williams,C. et al. (1999) A High Frequency of Sequence Alterations Is Due to Formalin Fixation of Archival Specimens. *The American Journal of Pathology*, **155**, 1467-1471.

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 Research*, **40**, 11189-11201.