LoLoPicker: Detecting Low Allelic-Fraction Variants in Low-Quality Cancer Samples

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. Compared with other methods, we showed superior performance of LoLoPicker with significantly improved specificity. The algorithm of LoLoPicker is particularly useful for calling variants from 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.

Introduction

The detection of somatic mutations in tumors 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 clonal heterogeneity. The difficulty of identifying those variants is magnified by the fact that sequencing technologies are still imperfect and error-prone (Flickinger *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 *et al*., 2014, Williams *et al*., 1999).

Next generation sequencing 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 to enable them filtering out false-positive calls that are specific to library preparation, protocols, instruments, environmental factors or analytical pipeline. Such databases provide an opportunity to precisely estimate the site-specific error rates that give the advantage to increase the sensitivity of calling single nucleotide variants (SNVs) on sites with lower error rates, and reduce false positives on sites with high error rates. This idea has been successfully implemented for targeted re-sequencing experiments (Gerstung *et al*., 2013).

However, there are no software able to perform low allelic-fraction SNV calling with high specificity on the genome scale. Here, we present LoLoPicker. The program requires users to provide a control panel consisting of normal samples processed using similar procedures as the test sample (tumor), and uses this control to estimate site-specific error rates. Then, a binominal test is performed to determinate whether the ratio of altered reads of the tumor variant exceeds the background error rate (Figure 1). Detailed description of the 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 *et al*., 2013, Koboldt *et al*., 2012, Wilm *et al*., 2012). Somatic mutations validated by Sanger in an ovarian tumor were used as a set of true positives. BAM files obtained from WES of the tumor sample and its matched blood sample were mixed 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. Moreover, 500 unrelated germ-line samples from non-cancer patients were used in our control panel. As the results, LoLoPicker showed highest sensitivity and specificity, even when we reduced the coverage of variants (Figure 1).

Applying LoLoPicker to Real Data

High-quality tumor samples

We then applied LoLoPicker, MuTect and VarScan2 to analyze a real cancer sample with matched blood sample from a glioblastoma (GBM) patient (GBM_9). Known GBM driving mutations were identified, including mutations in *TP53*, *H3F3A*, *ATRX*, and *PIK3CA*. Only LoLoPicker successfully identified all of them. MuTect discarded the *TP53* mutation because it found three reads supporting the variant in the normal sample. In LoLoPicker, the mutation was retained because we count overlapping read-pair covering same variant as one (Figure S3). 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 allelic-fraction SNVs. Moreover, 14 SNVs in GBM_9 were selected for validation. We found that all variants called by both LoLoPicker and MuTect were validated as true positives, whereas the ones that LoLoPicker rejected were not validated. These included four variants with higher coverage (>=5X) supporting the altered bases (Table S3). Our results suggested the high specificity of LoLoPicker without rejecting true positives as trade-off.

FFPE tumor 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 S6). In previously published work on small cell carcinoma of the ovary, hypercalcemic type (SCCOHT), we showed that only one gene – *SMARCA4* – was recurrently mutated, and no other recurrent mutations were observed (Witkowski *et al*., 2014). We therefore tested LoLoPicker on an FFPE-SCCOHT sample (UN5). Although very few somatic mutations other than *SMARCA4* 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 92 SNVs. When we switched our controls to 35 FFPE-normal tissues, only 18 variants were called. Most of the LoLoPicker rejected calls were errors (C to T and G to A transitions) known to be induced by the FFPE protocol (Spencer *et al*., 2013), suggesting the necessity of providing a control cohort to further reduce false positive calls related to batch effects, especially FFPE-specific artifacts (Figure 1).

Discussions

LoLoPickers is designed to detect somatic SNVs, particularly tailored for low allelic-fraction SNVs. LoLoPicker maintains highest sensitivity among other programs. More importantly, the specificity of LoLoPicker is dramatically improved, thus highlighting the importance of precisely measuring site-specific error rate from additional control samples, rather than from a matched normal sample solely. Although we expect that LoLoPicker will handles 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 reducing FFPE-specific artifacts. Compared to simply filtering out recurrent calls from normal samples, 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, although we showed that 35 FFPE controls are able to reject most of the false positives. As FFPE is 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 for his help with targeted re-sequencing. JM is the recipient of a Canada Research Chair in Genomics.

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.

Spencer, D.H. et al. (2013) Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *The Journal of Molecular Diagnostics,* **15**, 623-633.

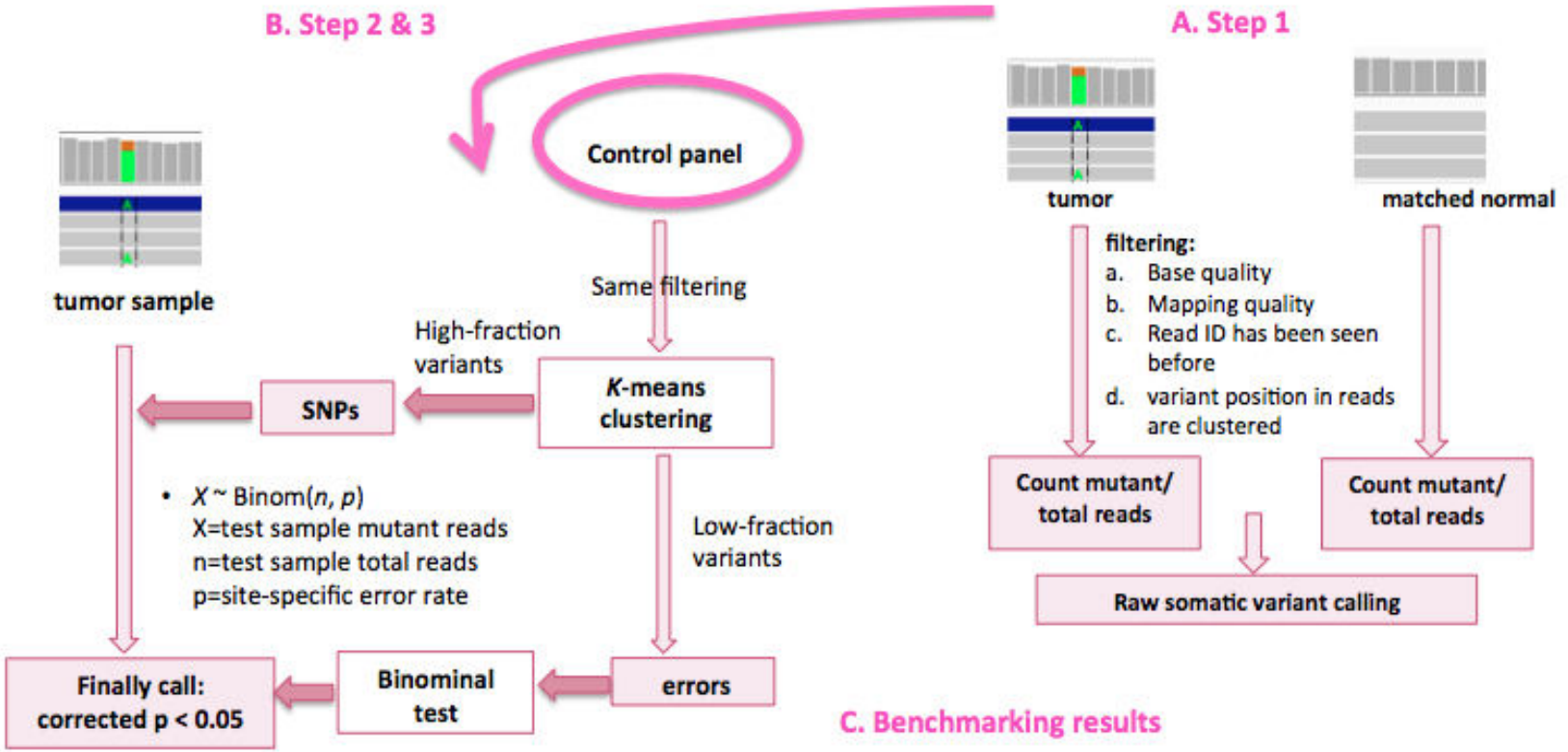
Van Allen,E,M. et al. (2014) Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature Method*, **20**, 682-688.

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.

Witkowski,L. et al. (2014) Germline and somatic *SMARCA4* mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nature genetics,* **46**, 438-443.

Figure 1: **Workflow and performance of LoLoPicker.** A. LoLoPicker first performs raw variant calling using tumor and matched normal sample. B. LoLoPicker then performs its core statistical framework using a user-provided control cohort**.** C. Number of true positives and false positives called by LoLoPicker, MuTect, VarScan and LoFreq from benchmarked samples, suggesting high sensitivity and specificity of LoLoPicker. D. C to T and G to A transitions, which are known FFPE-induced artifacts, are most frequently observed in LoLoPicker-rejected variants, and MuTect-called variants; whereas these transitions are barely observed among variants called by LoLoPicker.



C. Benchmarking results

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

D. FFPE-related errors are rejected by LoLoPicker