Computational and Structural Biotechnology Journal 16 (2018) 15–24

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data

Chang Xu

Life Science Research and Foundation, Qiagen Sciences, Inc., 6951 Executive Way, Frederick, Maryland 21703, USA

ARTICLE INFO

Article history:
Received 8 September 2017
Received in revised form 20 January 2018
Accepted 28 January 2018
Available online 6 February 2018

Keywords:
Variant calling
Somatic mutation
Unique molecular identifier
Low-frequency mutation
Benchmarking

ABSTRACT

Detection of somatic mutations holds great potential in cancer treatment and has been a very active research field in the past few years, especially since the breakthrough of the next-generation sequencing technology. A collection of variant calling pipelines have been developed with different underlying models, filters, input data requirements, and targeted applications. This review aims to enumerate these unique features of the state-of-the-art variant callers, in the hope to provide a practical guide for selecting the appropriate pipeline for specific applications. We will focus on the detection of somatic single nucleotide variants, ranging from traditional variant callers based on whole genome or exome sequencing of paired tumor-normal samples to recent low-frequency variant callers designed for targeted sequencing protocols with unique molecular identifiers. The variant callers have been extensively benchmarked with inconsistent performances across these studies. We will review the reference materials, datasets, and performance metrics that have been used in the benchmarking studies. In the end, we will discuss emerging trends and future directions of the variant calling algorithms.

Contents

E-mail address: chang.xu@qiagen.com.

https://doi.org/10.1016/j.csbj.2018.01.003

1. Introduction

DNA mutation is the cause of cancer and a major focus of cancer research and treatment. Next-generation sequencing (NGS) is by far the most promising technology for *de novo* mutation detection, thanks to the huge amount of reads that modern sequencers can generate. Theoretically, all mutations regardless of the variant allele frequency (VAF) or genomic region can be *observed* given enough read depth. However, *calling* them with confidence is not trivial due to noise in the reads. Numerous bioinformatics tools have been developed to uncover mutations (variants) from sequencing reads, and such procedures typically consist of three components: read processing, mapping and alignment, and variant calling. First, low quality bases (usually near the 3' end of reads) and exogenous sequences such as sequencing adapters are trimmed with read processing tools such as Cutadapt [1], NGS QC Toolkit [2], and FASTX-Toolkit. Some targeted sequencing protocols use PCR primers or unique molecular identifiers (UMI) during library preparation. In this case, custom-built read processing scripts may be required to trim and extract these oligonucleotides. Second, the cleaned reads are mapped to where they may come from in the reference genome, and then aligned base-by-base. Commonly used mapping and alignment tools include BWA [3], NovoAlign, and TMAP (for Ion Torrent reads) for DNA sequencing, and splice-aware aligners such as TopHat [4] and STAR [5] for RNA sequencing. PCR de-duplication, indel-realignment, and base quality recalibration can be performed in this step as outlined in the Genome Analysis Toolkits (GATK)'s best practice for variant calling [6,7]. The last step, variant calling, is essentially a process of separating real variants from artifacts stemming from library preparation, sample enrichment, sequencing, and mapping/alignment. It has been a very active research field for years and plenty of variant callers have been developed, many freely available. The goal of this article is to review the state-of-the-art variant callers for somatic variants, in the hope to assist practitioners, especially non-bioinformaticians, to select the appropriate variant caller for their own applications.

The underlying assumptions are quite different for germline and somatic variant calling algorithms. Germline variants are expected to have 50 or 100% allele frequencies, therefore germline variant calling is essentially to determine which of the three genotypes, AA, AB, or BB, fit the data best [7–10]. Most artifacts are present in low frequency and unlikely to cause trouble, because homozygous reference would be the most likely genotype in this case. But rejecting these artifacts is not as easy in somatic variant calling, because some real variants could also be present in very low frequencies in cases of impure sample, rare tumor subclone, or circulating DNA. Therefore, the biggest challenge of somatic variant calling is to disambiguate low-frequency variants from artifacts, which requires more sensitive statistical modeling and advanced error correction technology.

Genetic variants can be grouped into three categories by size: single nucleotide variant (SNV), insertion and deletion (indel), and structural variant (SV, including copy number variation, duplication, translocation, etc.). Very few variant callers are versatile enough to call all three because they require very different algorithms. For SNV and short indels (typically $\leq$10bp), the general strategy is to look for non-reference bases from the stack of reads that cover each position. Probabilistic modeling is critical here to infer the underlying genotype or evaluate the odds of variant versus artifacts. For structural variants and long indels, since the reads are too short to span over any variant, the focus is to locate the breakpoints based on the sudden change of read depth or patterns of misalignment with paired end reads. Split-reads and de-novo assembly methods are often used for SV and long indel detection.

In this review, we will focus on somatic SNV calling algorithms. We will review 46 publicly available somatic SNV callers that cover a wide spectrum of applications, in the hope to provide a practical guide for choosing the appropriate software. We will also explain the core algorithm of each variant caller and, if applicable, highlight the strengths and caveats. Germline-only callers, such as GATK UnifiedGenotyper/HaplotypeCaller, inGAP, and MAQ [6,7,11,12] are not included in this review. Although UnifiedGenotyper and HaplotypeCaller have been used for somatic variant calling, their core algorithms are not designed for this task and perform poorly for low-frequency somatic variants, as stated in the GATK documentation and shown by independent studies [13,14]. We will also exclude variant callers that are primarily used for pooled-samples such as CRISP and thunder [15,16].

The article will be structured as follows. We will first describe the general workflow of somatic SNV calling in Section 2. Next, we will explain the core algorithms of individual variant callers and arrange them by the intended application in Sections 3–6. Each dedicated to one type of application. We will then discuss methods of evaluating variant calling performance and review recent progress in benchmarking studies in Section 7. Finally, we will summarize the research field and discuss future directions in Section 8.

2. General workflow of somatic SNV calling

2.1. Pre-processing

In general, variant callers consist of three components: pre-processing, variant evaluation, and post-filtering. The main purpose of pre-processing is to keep low-quality reads from entering the variant evaluation procedure. Read quality is typically measured by average base quality score, mapping quality score, and number of mismatches from the reference genome, etc. If the SNV caller follows a position-based strategy, which basically calls variant at each target position independently and is adopted by most SNV callers, a read can be included at one position and excluded at another, depending on the base quality scores at each individual position. Some variant callers such as Strelka [17] and VarDict [18] implement local indel realignments during pre-processing, resulting in better accuracy around indels. This can also be done using GATK IndelRealigner and BQSR (base quality score recalibration). PCR de-duplication is recommended in whole genome or whole exome sequencing data and can be performed with SAMtools or Picard tools. But it is not recommended in PCR-based amplicon sequencing applications where distinct DNA fragments can share the same genome coordinates. Also included in this step is downsampling during which a subset of reads are randomly selected to proceed to the next steps. Downsampling saves computation time and improves coverage uniformity if done at specific regions, but also makes the results non-deterministic.

2.2. Variant evaluation

Variant evaluation algorithm is the centerpiece of somatic variant callers and hence the focus of this review. Depending on the type of input data and the intended application, the algorithms can be summarized to four categories: matched tumor-normal variant calling, single-sample variant calling, UMI-based variant calling, and RNA-seq variant calling. Individual algorithms will be discussed in detail in Sections 3–6.

2.3. Post-filtering

Sequencing or alignment artifacts may appear to have strong read evidence and trick the statistical model to pass them as real variants. Most variant callers apply a set of filters to identify these artifacts and hence improve the specificity. Strand bias filter, for example, catches artifacts whose reads are only or dominantly observed on one strand, a common error in Illumina reads [19,20]. Strand bias filters rely on the Fisher's exact test to identify imbalanced strand