Methods 129 (2017) 3–7

Contents lists available at ScienceDirect

Methods

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

Toolkit for automated and rapid discovery of structural variants

Arda Soylev[a], Can Kockan[a,1], Fereydoun Hormozdiari[b,*], Can Alkan[a,*]

[a] Department of Computer Engineering, Bilkent University, Ankara, Turkey
[b] Department of Biochemistry and Molecular Medicine, MIND Institute and UC-Davis Genome Center, University of California, Davis, CA, United States

ARTICLE INFO

Article history:
Received 16 February 2017
Received in revised form 20 May 2017
Accepted 30 May 2017
Available online 2 June 2017

Keywords:
Structural variation
High throughput sequencing
Combinatorial algorithms

ABSTRACT

Structural variations (SV) are broadly defined as genomic alterations that affect >50 bp of DNA, which are shown to have significant effect on evolution and disease. The advent of high throughput sequencing (HTS) technologies and the ability to perform whole genome sequencing (WGS), makes it feasible to study these variants in depth. However, discovery of all forms of SV using WGS has proven to be challenging as the short reads produced by the predominant HTS platforms (<200 bp for current technologies) and the fact that most genomes include large amounts of repeats make it very difficult to unambiguously map and accurately characterize such variants. Furthermore, existing tools for SV discovery are primarily developed for only a few of the SV types, which may have conflicting sequence signatures (i.e. read pairs, read depth, split reads) with other, untargeted SV classes. Here we are introduce a new framework, TARDIS, which combines multiple read signatures into a single package to characterize most SV types simultaneously, while preventing such conflicts. TARDIS also has a modular structure that makes it easy to extend for the discovery of additional forms of SV.

1. Introduction

Genome structural variations (SVs), defined as genomic alterations >50 bp [1,2], play major roles in both genome evolution [3] and pathogenesis of diseases of genomic origin such as schizophrenia, epilepsy, and autism [4]. Although *-by count-* less number of SVs are found in each human genome with respect to the reference than single nucleotide polymorphisms (SNPs), the total number of affected basepairs by SVs far exceed those affected by SNPs [2]. It is, therefore, of utmost importance to accurately and comprehensively characterize all forms of SVs, including copy number variants (CNVs, i.e. deletions, insertions and duplications), mobile element insertions, and balanced rearrangements (inversions and translocations).

Algorithm development for structural variation discovery and genotyping using high throughput sequencing (HTS) data was accelerated during the 1000 Genomes Project [2,5,6]. Briefly, all algorithms use one or several of four basic read mapping *signatures*: read pair, split read, read depth, and assembly [1]. The detection accuracy of using each sequence signature differs depending on the type, size, and the underlying sequence properties of genomic location of the SV. Therefore, although the first few SV discovery algorithms focused on using a single sequence signature [7–14], more recent SV callers use multiple signatures [15–19]. However, most SV calling algorithms aim to characterize one or a few types of SV, and they do not try to resolve conflicting SV within the same locations, or sequence signature that signal more than one type of SV.

Here we introduce TARDIS, a toolkit for automated and rapid discovery of SVs. TARDIS integrates read pair, read depth, and split read (using soft clipped mappings) sequence signatures to discover several types of SV, while resolving ambiguities among different putative SVs: 1) at the same locations signaled by different sequence signatures, and 2) in different locations signaled by the same mapping information. TARDIS is fully automated and requires no user intervention. Additionally, it is suitable for cloud use as the memory footprint is low. The current version is capable of characterizing deletions, small novel insertions, tandem duplications, inversions, and mobile element retrotransposition.

TARDIS is implemented in C using HTSLib (http://www.htslib.org), and it is freely available at https://github.com/BilkentCompGen/tardis.

2. Methods

We have previously developed some of the first tools to discover various types of SV that also incorporate multi-mapping of

* Corresponding authors.
E-mail addresses: fhormozd@ucdavis.edu (F. Hormozdiari), calkan@cs.bilkent.edu.tr (C. Alkan).

[1] Current address: School of Informatics and Computing, Indiana University, Bloomington, IN, United States.

http://dx.doi.org/10.1016/j.ymeth.2017.05.030

reads, such as mrCaNaVaR/mrFAST [20], VariationHunter [8], VariationHunter-CR [13], NovelSeq [21], Pamir [22], and CommonLAW [23]. All of these tools use a similar objective function for SV discovery although they are developed to discover different types of SV under different conditions (e.g. single vs. multi-sample) using different sequence signatures [1,12]. We now further improve our algorithms for SV detection and integrate them into a single package (TARDIS) that can simultaneously characterize different forms of SVs using read pairs, read depth, and split reads. TARDIS is a user-friendly single executable with a potential to be easily extended for discovering additional forms of complex SV (e.g. translocations) and for supporting different sequencing technologies such as linked read sequencing [24] and long read sequencing (i.e., PacBio, nanopore). However, the current version of TARDIS is developed only for whole genome sequencing (WGS) data generated with the Illumina platform, and in the remainder of the paper we assume the input is Illumina WGS. Below we first define the terminology and then provide problem formulation and our solution.

We first define some of the terms that we use in this paper below.

- *fragment size:* the Illumina WGS protocol generates paired-end reads from both ends of longer fragments. The lengths of these fragments are assumed to be sampled from a normal distribution. Therefore, in the absence of structural variants, mapping locations of the paired ends *span* within an interval $[\delta_{min}, \delta_{max}]$. Most (>90%) of paired-end reads are sampled from no-SV regions, therefore the fragment size distribution can be learned empirically for each WGS data set separately.
- *concordant reads:* a read pair is called *concordant* if they can be mapped to the reference genome as "expected": (a) mapped to opposing strands where the upstream read is mapped to the forward strand and the downstream read is mapped to the reverse strand,[2] (b) the distance between ends is between the minimum and maximum expected fragment size.
- *discordant reads:* briefly, any non-concordant read pair is considered *discordant*. Note that, by definition, the discordant read pairs signal potential SVs. The sequence signature produced by these type of reads is known as read-pair signature [1,12].
- *split reads:* a read that can only be mapped to the reference genome by breaking into two sub-reads is called a *split-read*. These types of reads also indicate a potential SV or a short insertion or deletion (indel).
- *read depth:* number of reads that map within a region of the genome. Overall genome-wide read depth is also referred to as *depth of coverage*. It is expected that the number of reads that "cover" each base-pair to follow a Poisson distribution. Therefore, if the read depth over a certain region deviates significantly from this distribution, it signals for a potential copy number variation (CNV) [1,20,12].

2.1. Problem formulation

One of the main drawbacks of high-throughput sequencing technologies is that reads are usually very short (<200 bp). This results in mapping ambiguity as some reads may map to more than one location equally likely due to genomic repeats and segmental duplications [25]. Similar to our previous work [8,13,23], TARDIS uses the signatures explained above and it also considers all map locations of multi-mapping reads. However, TARDIS also has a *quick* mode, which considers only the best map location provided in the input BAM file. We formulate our problem formulation under the assumption of *maximum parsimony*.

As in VariationHunter [8] the objective function that TARDIS tries to optimize is also based on maximum parsimony. Briefly, TARDIS aims to minimize the total number of structural variation inferred from all discordant read pairs and split reads. We have previously showed that maximum parsimony SV discovery problem is NP-Complete [8] by reduction from the SET-COVER problem [26]. Additionally we provided a greedy algorithm with an approximation factor of $O(\log\ n)$ using only the read pair signature.

In addition to the read pair signature, TARDIS also uses read depth and split read signatures for SV discovery. Briefly, after clustering discordant read pairs (Section 2.2), we can assign weights to the clusters based on the GC%-normalized read depth within the inferred cluster coordinates (Section 2.3). Note that, since the read depth weights are calculated for each cluster once, and they mainly represent a score, the approximation ratio of the greedy algorithm does not change.

2.2. Maximal valid clusters of read pairs

We define a set of discordant read pairs that signal the same SV (i.e. same type and size) as a *valid cluster*. Similarly, we define a *maximal valid cluster* as a valid cluster where no additional discordant read pairs can be added without violating its validity. Valid clusters for some of the SV types are previously defined in [27,8,28].

2.3. Read-depth signature

We use read depth signature to score and eliminate likely false positive CNV calls (deletions). We model read depth distribution as Poisson, and we calculate the read depth of each putative SV as the summation of read depths for each base pair within the SV breakpoints. Other discrete binomial distributions have been suggested for modeling read depth such as the *negative binomial* distribution [29]. Calculation of the distribution function is implemented as a module in TARDIS, thus it can be replaced in upcoming versions.

Note that the summation of two Poisson distributions is also a Poisson distribution. Additionally, we use a statistical smoothing method (i.e. LOESS transformation) to normalize read depth values based on the GC% content as previously described elsewhere [20,30].

Next, we calculate the probability $P(RD|CN = i)$[3] for each putative deletion within breakpoint intervals (B_l, B_r) as follows. We first calculate the *expected read depth* (denoted as E_{RD}) within the deletion breakpoints normalized with respect to its GC% content using a sliding window of size 100 bp. Here, the expected read depth refers to "normal" read depth (i.e. no CNV).

We then calculate for every region the copy number corrected (i.e. $CN = i$) expected read depth as

$$E_i = \frac{E_{RD} \times i}{2}$$

We also denote observed read-depth as O. Thus assuming Poisson distribution we calculate the probability $P(RD|CN = i)$ as:

$$P(RD|CN = i) = \frac{{E_i}^{O} \times e^{-E_i}}{O!}$$

We consider a deletion prediction to be correct if the likelihood of the observed read depth is significantly higher for a copy number that supports a deletion (i.e. CN = 0 or CN = 1) compared to that of CN> 1. More formally, we calculate the deletion likelihood assuming the copy number is bounded by 10.

[2] This is correct for most Illumina WGS data sets, however, there are alternative library preparation protocols with different strand rules.

[3] RD: read depth, CN: copy number, and i denotes an integer for copy number.