

# **MICADo - Looking for mutations in PacBio cancer data: an alignment-free method**

Journal:	<i>Bioinformatics</i>
Manuscript ID	BIOINF-2016-0485
Category:	Original Paper
Date Submitted by the Author:	01-Apr-2016
Complete List of Authors:	Rudewicz, Justine; Université de Bordeaux, Computer Science Soueidan, Hayssam; NKI-AVL, Molecular Carcinogenesis Uricaru, Raluca; Université de Bordeaux, Computer Science Bonnefoi, Hervé; Centre Régional de Lutte Contre le Cancer Institut Bergonié Iggo, Richard; Centre Régional de Lutte Contre le Cancer Institut Bergonié Bergh, Jonas Nikolski, Macha; LaBRI / CNRS, Computer Science
Keywords:	HiTSeq, PacBio, Cancer Genomics, Next-generation sequencing, Mutation Calling, de Bruijn graphs

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



Subject Section

MICADo - Looking for mutations in PacBio cancer data: an alignment-free method

Justine Rudewicz<sup>1,2,3\*</sup>, Hayssam Soueidan<sup>1,2</sup>, Raluca Uricaru<sup>1,2</sup>, Hervé Bonnefoi<sup>3</sup>, Richard Iggo<sup>3</sup>, Jonas Bergh<sup>4</sup> and Macha Nikolski<sup>1,2\*</sup>

<sup>1</sup>Univ. Bordeaux, CBiB, F-33000 Bordeaux, France and <sup>2</sup>Univ. Bordeaux, CNRS/LaBRI, F-33405 Talence, France and <sup>3</sup>INSERM U916, Bergonié Cancer Institute, Univ. Bordeaux, F-33076 Bordeaux, France and <sup>4</sup>Karolinska Institute and University Hospital, 171 76 Stockholm, Sweden

\*To whom correspondence should be addressed.  
Associate Editor: XXXXXXXX  
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

**Motivation:** Targeted sequencing is commonly used in clinical application of NGS technology since it enables generation of sufficient sequencing depth in the targeted genes of interest and thus ensures the best possible downstream analysis. This notwithstanding, the accurate discovery and annotation of disease causing mutations remains a challenging problem even in such favorable context. The difficulty is particularly salient in the case of third generation sequencing technology, such as PacBio.

**Results:** We present MICADo, a de Bruijn graph based method that makes possible to distinguish between patient specific mutations and other alterations for targeted sequencing of a cohort of patients. MICADo analyses NGS reads for each sample within the context of the data of the whole cohort in order to capture the differences between specificities of the sample with respect to the cohort. MICADo is particularly suitable for sequencing data from highly heterogeneous samples, especially when it involves high rates of non-uniform sequencing errors. It was validated on PacBio sequencing datasets from several cohorts of patients.

**Contact:** justinerudewicz@gmail.com  
**Implementation:** The source code is available at <http://github.com/cbib/MICADo>.

1 Introduction

Capturing known cancer genes by next generation sequencing, approach known as "gene panel" or targeted sequencing, is commonly used for tumor genotyping. Such studies enable the discovery of point mutations, insertions/deletions (indels), copy number variations and gene rearrangements. Second-generation sequencing platforms, like Illumina, 454 Life Sciences (Roche) and Life Technologies Ion Torrent (van Dijk et al., 2014), are currently widely used for diagnostic applications in cancer. Comparatively, third-generation sequencing platforms, like Pacific Biosciences (PacBio), offer many theoretical advantages such as longer reads, reduced cost and greater speed (Eid et al., 2009). However, PacBio technology is known to have a high rate of sequencing errors and thus has not yet entered the mainstream diagnostic use. Moreover, the absence of PCR bias in PacBio is to be counterbalanced by the presence of PCR in the case of targeted sequencing (Robasky et al., 2014). Error rate as

high as 17.9% has been reported (Chin et al., 2011), the majority of errors being indels according to (Eid et al., 2009). Notwithstanding the high error rate, the expected positive characteristic of PacBio technology is supposedly the absence of context-specific errors that affect other short-read sequencing platforms (Carneiro et al., 2012). Indeed, this technology is alleged to generate errors randomly and in the following proportions: 10-12% of insertions, 2-5% of deletions and about 1% of mismatches (Ross et al., 2013; Carneiro et al., 2012). If the PacBio error model is random, the consequence is that given a sufficient sequencing coverage, high consensus accuracy can be obtained from the reads thus opening the avenue for diagnostic applications of this technology.

The goal of variant calling is to identify positions where at least one of the bases differs from the reference genome, the main types of such short differences being Single Nucleotide Polymorphisms (SNPs) and indels. SNPs are variants that affect a single nucleotide, are present in a population and most often do not have any impact on health. On the other hand, a somatic point mutation – called somatic Single Nucleotide Variant (somatic SNV) – occurs in a tumor sample, the control allele being the wild

type. Usually, somatic SNVs are identified by comparing a tumor sample with a corresponding normal sample. A large number of tools for variant calling has been developed in recent years (Pabinger et al., 2014; Leggett and MacLean, 2014), including methods specific for the identification of somatic mutations in cancer (see (Cancer Genome Consortium, 2013) and (Wang et al., 2013) for review). A number of comparative studies have shown that variant callers available within the GATK pipeline (McKenna et al., 2010) show the best performance (see e.g. (Pirooznia et al., 2014; Yi et al., 2014)). For cancer data, VarScan (Koboldt et al., 2012) is a popular choice. A recent study (Warden et al., 2014) has compared the performance of these two tools and has shown a significant overlap of high-quality GATK and VarScan variant calls.

Despite the existence of these numerous computational solutions, calling somatic mutations remains challenging due to a number of factors like technical artifacts, sequencing errors, biases of alignment algorithms, DNA contamination (control samples contaminated with tumor DNA and vice versa) and tumor heterogeneity. This issue is even more salient for the third generation sequencing data, such as PacBio. Indeed, since very high read depths are required for achieving sequence accuracy close to that of Illumina and Ion Torrent (Quail et al., 2012), variant calling potentially suffers from high false positive and negative rates. Moreover, variant callers are notorious for encountering difficulties with indels (Zook et al., 2014), which is the most common error type in PacBio sequencing data. These issues can be circumvented through the use of *de Bruijn graphs* (DBG) structure. Indeed, some new approaches based on the DBG structure have recently emerged (Iqbal et al., 2012; Uricaru et al., 2014; Rizk et al., 2014) but either remain specific to the detection of one variation type or are not appropriate for cancer data. Therefore, only few variant callers are recommended for PacBio data, such as the aforementioned GATK pipeline and the PacBio's Quiver algorithm.

In the clinical setting, distinguishing between true mutations and artifacts introduced by sequencing errors and variant calling tools, is essential. Indeed, systematic molecular characterization of relevant genes within patient cohorts is the key for understanding cancer initiation and progression. When sequencing data obtained from a patients' cohort is analysed for the presence of SNVs, identification of genetic variants is performed for each patient's sample individually. These individual results can be further jointly analyzed in the downstream statistical analysis. Such standard analysis workflows do not take advantage of the whole tumor cohort data. Concomitantly, multiple studies have shown the existence of specific error profiles in second generation sequencing. For example, GC rich regions and the extremities of sequence reads exhibit higher error rates (Allhoff et al., 2013), and both Roche 454 and Ion Torrent encounter difficulties in the homopolymer regions (Ross et al., 2013). Based on these observations, Bansal (Bansal, 2010) has proposed a statistical method that identifies rare and common variants in DNA pools of diploid individuals. However the diploidy requirement prevents its application to cancer data where samples are highly heterogeneous.

In this manuscript we pursue the idea that analyzing sequencing data produced for a cohort of patients as a whole should make it possible to distinguish between real patient mutations and other alterations, including sequencing errors. Indeed, one sample can be analyzed in the context of the whole dataset, serving itself as statistical basis for filtering out systematic alterations. This idea is even more pertinent in the case of sequencing data carrying non-uniform sequencing errors as, shown in this paper, it is the case for PacBio data. We introduce a method called *MICADO* that distinguishes patient-specific from cohort-specific alterations and show that it efficiently performs even on samples carrying SNVs in the context of contamination by germline DNA and in the presence of sequencing errors. *MICADO* is based on the well-known representation of NGS sequencing reads, DBG, which provides the double advantage of circumventing a time-consuming alignment step required by most SNV callers and avoiding

additional biases due to the alignment itself. In theory, *MICADO* can be used on any kind of sequencing data, however its advantage with respect to other variant calling methods should be lesser in the case of short read datasets with low error ratios. Therefore, *MICADO* was evaluated on PacBio sequencing datasets: (i) a novel sequencing of TP53 of a breast cancer cohort, (ii) a publicly available dataset of FLT3 sequencing of an acute myeloid leukemia cohort and (iii) a synthetic dataset based on the TP53 data.

## 2 Approach

Our approach, called *MICADO*, takes as input the reference sequences for a gene of interest and several read sets corresponding to the gene panel sequencing of a cohort of patients. This method allows for accurate detection of patient mutations in such targeted sequencing data and their distinction from other alterations. *MICADO* is built of three major steps.

First, the reference and the cohort sequencing data are efficiently represented with de Bruijn graphs. A DBG, widely used in NGS data processing methods (Pevzner et al., 2001; Compeau et al., 2011) is a directed graph that depicts the redundancy encountered in a read set by representing all  $k$  length words contained in the data, i.e.  $k$ -mers, as vertices in the graph, and all  $k - 1$  overlaps between  $k$ -mers as directed edges in the graph. Here we use an extension of the classical DBG, meaning a *colored* DBG (Iqbal et al., 2012). A colored DBG comprises several read sets and has multi-colored vertices meant to keep track of the datasets to which every  $k$ -mer belongs. We build a colored DBG *reference graph* for the gene reference sequences, including splicing variants and known SNPs. The vertices of this graph are colored according to their original reference sequences. We then build a colored DBG *sample graph* per patient dataset. The vertices of the sample graph are colored according to the patient dataset, thus with only one color.

Second, we compare the reference graph to each sample graph in order to highlight paths in the sample graph that are not present in the reference graph, i.e. *alternative paths*, see Figure 1 (A).

Finally, for each alternative path, we verify whether it is specific to the sample or common within the cohort. This is achieved by generating random samples populated by reads from the whole cohort and checking the presence of each alternative path previously identified within these random sets, see Figure 1 (B).

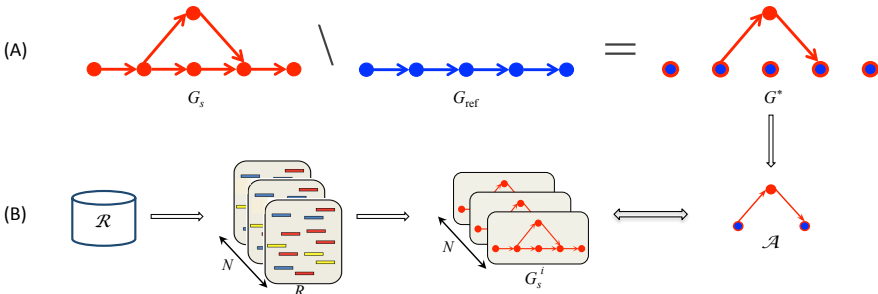
## 3 Methods

### 3.1 Datasets

*MICADO* was tested on three datasets: TP53 sequencing PacBio dataset, FLT3 PacBio sequencing dataset and a synthetic dataset.

**TP53 sequencing data** The p53 protein encodes for the tumor suppressor gene TP53, the most frequently mutated gene in human cancers (Olivier et al., 2010), given its major implication in response to cellular stress. Most of mutations are found in its DNA Binding Domain used to activate genes involved in apoptosis, growth arrest, or senescence of highly damaged cells (Brosh and Rotter, 2009).

We have sequenced 48 samples of TP53 transcripts extracted from breast tumor biopsies (Bonnefoi et al., 2011) by PacBio circular sequencing (PacBio RSII; P4-C2 chemistry). The PacBio sequencing data is available from the NCBI SRA database under the accession number SRP064161 BioProject PRJNA290142. The main interest of this dataset for testing the accuracy of SNV detection is the existence of SNV calling results obtained from 454 Roche sequencing data generated for the 48 samples (available from NCBI SRA database under the accession number SRP020456, BioProject PRJNA193388; see (Iggo et al., 2013) for details). Moreover, for these 48 samples there exists a classification into three categories (based on the percentage of red colonies in the yeast assay, see (Bonnefoi et al., 2011) and (Iggo et al., 2013)). These categories are:



**Fig. 1.** Summary of MICADo.  $G_{\text{ref}}$  and  $G_s$  are respectively constructed from multiple reference sequences and the sample read set. After capturing differences between  $G_s$  and  $G_{\text{ref}}$  in  $G^*$ , the goal is to decide whether they are sample specific by checking their presence in random samples  $R_i$  constructed from the whole readset of the cohort  $\mathcal{R}$ .

- *negative control*: 12 samples considered to be wild type;
- *positive control*: 18 samples that are mutated with high rate of altered reads;
- *difficult group*: 18 samples that are mutated with low rate or complex mutations.

Sequencing was centered on the DNA Binding Domain of the protein. TP53 transcripts were split into two distinct fragments, 482 nucleotides long N fragment and 454 nucleotides long C fragment. After demultiplexing and primer cutting (see Supplementary section 2 for more details), the number of reads varies from 132 to 8372 for the C fragment and from 89 to 8771 for the N fragment with respective averages of 974 and 826.

**FLT3 sequencing data** The receptor tyrosine kinase FLT3 is highly involved in the development of stem cells and in the immune system (Gilliland and Griffin, 2002). Mutations of FLT3 have been detected in about 20% of patients with acute myeloid leukemia (AML) and are associated with poor prognosis (Smith et al., 2012).

This dataset is composed of 19 samples obtained by PacBio circular sequencing of FLT3 transcripts for 8 AML patients before treatment and the same patients after relapse, and 3 healthy individuals with no cancer history. The data is available from NCBI SRA database under the accession number SRP011010 Bioproject PRJNA85103. All patients present an internal tandem duplication (ITD) that is associated after relapse with new point mutations; a complete description of mutations carried by each patient can be found in (Smith et al., 2012). Sequencing was done on the FLT3 kinase domain which is 1346 nucleotides long, with an average of 1348 reads per sample (the number of reads varies from 59 to 4856 reads).

**Synthetic data** Synthetic samples with artificially inserted alterations were generated from the pooled TP53 dataset as follows. Approximately 20k reads from all the TP53 samples from the negative control group were first pooled in a single file and mapped with GMAP against the TP53 transcript variant 1 (NCBI Reference Sequence: NM\_000546.5). To build a synthetic sample, we first sample  $n$  reads ( $n$  ranging from 150 to 1000) from all the mapped reads (chosen with uniform probability), we then sample  $i$  positions ( $i$  ranging from 1 to 3) to alter between the minimal and the maximal mapped positions. For each of the  $i$  alterations, we choose with uniform probabilities between a mismatch, an insertion and a deletion, as well as an alteration length  $l$  ( $l$  is 1 for mismatches and ranges from 1 to 5 for insertions or deletions). In the case of insertions, we randomly generate a  $l$ -nt sequence. Finally, we generate a fastq file with  $n$  reads, out of which a fraction  $\varphi$  ( $\varphi$  ranging from 3% to 80%) harbors the same randomly generated alterations. The base quality of each alteration was fixed to the highest base quality.

3.2 Construction of de Bruijn graphs

A DBG is a directed graph whose vertices are words of a given length  $k$ , called  $k$ -mers, and whose edges connect words overlapping by  $k - 1$  letters. Two definitions for *de Bruijn* graphs are encountered in the literature,

depending on the way edges are dealt with: the implicit definition that infers the edges from the vertices, *i.e.* every  $k - 1$  overlap is accounted for whether the corresponding  $k$ -mers are consecutive or not within the reads (Uricaru et al., 2014), and the explicit formulation that specifically represents the edges in addition to the vertices, in which case only  $k - 1$  overlaps between consecutive  $k$ -mers present in the reads can be considered (Iqbal et al., 2012). In this work we employ the explicit formulation. Also, it is commonly accepted that a vertex in a DBG represents both the  $k$ -mer and its reverse complement. Here, as reverse reads are reverse complemented prior to building the graph (see Supplementary section 2), we solely represent  $k$ -mers (and not their reverse complements).

Moreover, we use an extended version of the DBG, a *colored* DBG (Iqbal et al., 2012), best suited to the multi-sample case. A colored DBG takes several read sets corresponding to multiple samples, and aggregates them in a union DBG, while coloring the vertices with respect to the samples they belong to. A more formal definition follows below.

**Definition 1.** For  $n$  given sets of sequences  $S_1, \dots, S_n$  and  $k \geq 2$ , the corresponding colored de Bruijn graph  $G = \langle V, E, l \rangle$  is such that its set of vertices  $V$  is composed of all  $k$ -mers of  $S_1, \dots, S_n$ , its edge set is

$$E = \{(v, w) : v, w \in V \text{ and } v_2 \dots v_k = w_1 \dots w_{k-1} \text{ and } \exists i \in [1, n] \text{ s.t. } \exists s \in S_i, v_1 v_2 \dots v_k w_k \subseteq s\},$$

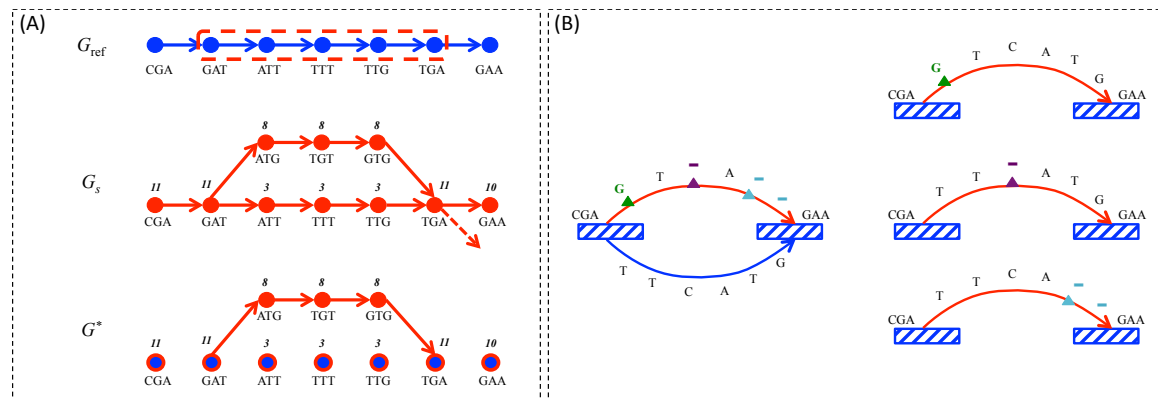
and the labeling function  $l : V \rightarrow \mathcal{P}(L)$  provides colors for the vertices with  $L$  being the set of  $n$  labels (colors) corresponding to the  $n$  sets of sequences.

When it is clear from the context, we will use a shortcut notation  $v = w$  for the equality of  $k$ -mers that are encoded by the corresponding vertices and omit the notation of the labeling function  $l$ .

For a gene of interest, its *reference graph*, noted  $G_{\text{ref}} = \langle V_{\text{ref}}, E_{\text{ref}}, l \rangle$ , is a  $n$ -colored DBG constructed from the  $k$ -mer decomposition of the  $n$  sequences corresponding to its splicing variants and known SNPs. A *sample graph*, noted  $G_s = \langle V_s, E_s \rangle$ , is a 1-colored DBG, thus a simple DBG, constructed from the sequencing reads of one sample. For the sample graph we define the *read support* of a vertex  $v$ , denoted as  $r(v)$ , as the number of reads the corresponding  $k$ -mer appears in. In order to discard indisputable sequencing errors, vertices having a read support below a fixed threshold are removed from  $G_s$  (see Supplementary section 7).

3.3 Alternative path search

Given the reference sequences and the sample reads modeled by  $G_{\text{ref}}$  and  $G_s$ , detecting sequence alterations (insertions, deletions and mismatches) comes to capturing the differences between these graphs. This means that we have to identify paths, *i.e.* sequences of vertices, present in  $G_s$  but absent in  $G_{\text{ref}}$ , that we call *alternative paths*. Basically, an alternative path corresponds to a sequence of  $k$ -mers (vertices) within the sample that differ from the reference sequence except for the two anchoring  $k$ -mers which are common to both the sample and the reference (see Figure 2 (A)).



**Fig. 2.** (A)  $G_{\text{ref}}$  encodes the sequence CGATTGAA for  $k = 3$ . The sequencing set corresponding to the sample is composed of 8 reads containing CGATTGAA sequence, 2 reads with CGATTGAA and 1 read containing CGATTGGA sequence followed by something else than A;  $G_s$  represents this read set for  $k = 3$ . The difference graph  $G^*$  contains the alternative path  $p_a$  encoding the subsequence ATGTG with its read support  $r(p_a) = 8$ , as well as the singleton vertices (blue dots circled in red) for shared  $k$ -mers between  $G_{\text{ref}}$  and  $G_s$ , but with no alternative path between them. The reference path corresponding to  $p_r$  is circled in red in  $G_{\text{ref}}$ . (B) An alternative path  $p_a$  encoding the sequence CGAGTAGAA is identified with the corresponding  $p_r$  encoding CGATTGATGAA.  $k$ -mers common to  $p_a$  and  $p_r$  are depicted by striped boxes.  $p_a$  carries 3 alterations, corresponding to the following edit operations:  $\delta_1$  - substitution of T by a G (green triangle),  $\delta_2$  - deletion of C (lilac triangle) and  $\delta_3$  - deletion of TG (blue triangle).  $p_a$  is thus split in three atomic alternative paths, each carrying only one alteration.

**Definition 2.** We define the difference graph  $G^* = \langle V^*, E^* \rangle$  to be such that  $V^* = V_s \cup V_{\text{ref}}$  and  $E^* = E_s \setminus E_{\text{ref}}$ . A path  $p_a = (v_1^* \dots v_n^*)$  in  $G^*$  is said to be an alternative path if there exists a path  $p_r = (v_1 \dots v_m)$  in  $G_{\text{ref}}$  such that  $v_1^* = v_1$  and  $v_n^* = v_m$ .

We generalize the notion of *read support*, originally defined on vertices, to paths in the sample graph: the *read support* of a path  $p$  is the number  $r(p)$  of reads from the sample in which the corresponding succession of  $k$ -mers composing  $p$  appears. We say that a read supports a path if and only if it supports all  $k$ -mers composing the path.

An alternative path  $p_a$  can carry  $n \geq 1$  alterations with respect to its reference path  $p_r$ . We use the Levenshtein distance  $\text{lev}(p_a, p_r)$  between the sequences defined by  $p_a$  and  $p_r$ , to determine the minimal set  $\delta_i$  (with  $1 \leq i \leq n$ ) of edit operations (deletions, insertions and substitutions) that transform  $p_a$  into  $p_r$ . This defines the set of *atomic alternative paths* corresponding to  $\{\delta_i(p_a)\}$  (see Figure 2 (B)). The read support of these atomic alternative paths is set to be equal to  $r(p_a)$ .

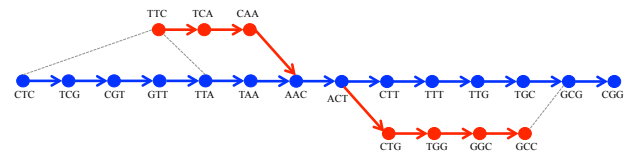
In the following, the set of tuples corresponding to the alterations computed for a given sample is denoted by  $\mathcal{A} = \{ \langle p_a, p_r, c \rangle \}$ , where  $p_a$  is an atomic alternative path in  $G^*$ ,  $p_r$  its corresponding reference path in  $G_{\text{ref}}$ , and  $c(p_a)$  its *count ratio* with  $c(p_a) = r(p_a) / (r(p_a) + r(p_r))$ , where  $r(p_r)$  is computed in the sample graph. For example, in figure 2 (A) the count ratio for the alternative path  $p_a$  is computed as  $c(p_a) = 8 / (8 + 3) = 0.72$ . Notice that  $p_a$  is atomic as it carries only one alteration (G vs. T mismatch).

Moreover, an alternative path has to be *simple*, that is composed by at most two vertices from  $V_{\text{ref}}$ , meaning that composite paths are not considered. The existence of multiple reference sequences can engender a combinatorial explosion of reference paths for a given alternative path. Consequently, for  $p_r$  we retain the reference path that maximises the intersection, in terms of supporting reads, between the alternative path and the reference paths.

To construct  $\mathcal{A}$ , we apply the alternative path search algorithm (**Algorithm 1**) presented below.

Notice that certain  $v \in V_{\text{start}}$  may not belong to  $V_{\text{ref}}$ . We call such start or end vertices *tips*. Tips may appear in  $G^*$  for three reasons: (i) the presence of reads that do not start or end at the beginning, respectively at the end of the region of interest, (ii) the presence of reads that carry alterations at the very beginning or end of their sequence, and (iii) because of the removal from  $G_s$  of vertices having a read support below a fixed threshold. For the algorithm to work in all cases, we have to define an appropriate reference path for a  $p_a$  starting or ending with a tip. This is

done by a heuristic lookup for the most plausible reference path as shown in Figure 3. To do this, we look for *anchors* of these tips in  $G_{\text{ref}}$ . These anchors are defined as vertices that belong to paths between the start node in  $G_{\text{ref}}$  and the corresponding  $v_e$  (symmetrically,  $v_s$ ) having the smallest Levenshtein distance between the associated  $k$ -mers. If there is more than one such vertex, we choose the one defining the  $p_r$  that is the closest in length to  $p_a$ .



**Fig. 3.** Given the  $G_{\text{ref}}$  encoding the sequence CTCGTTAACTTTGCGG (in blue) and  $G_s$  encoding TTCAACTGGCC (in red), there are two tips: TTC and GCC. Two alternative paths are  $p_a^1$  corresponding to TTCAAC and  $p_a^2$  corresponding to ACTGGCC. For TTC the closest  $k$ -mers among those on the path from CTC to AAC in terms of Levenshtein distance are CTC and TTA (depicted by gray dotted lines). We choose TTA since the paths  $|CTC \dots AAC| = 6$  and  $|TTA \dots AAC| = 2$  - the latter path being the closest in length to  $p_a^1$ . In the same manner we choose GCG as anchor for GCC.

### 3.4 Alternative path specificity and variant calling

Once the set of atomic alternative paths  $\mathcal{A}$  is computed, MICADO partitions it into sample-specific alterations (that is, mutations) and other paths corresponding either to non-specific alterations (recurrent within the cohort) or to sequencing errors. To this end, we use a permutation test based on rearranging reads and sample assignments. This approach is justified by the following two hypothesis.

1. Mutations are assumed to be independent across samples and non-recurrent across the cohort. This is particularly true for loss of function mutations (such as those leading to the inactivation of TP53).
2. Sequencing errors are supposed to be distributed non-uniformly and are recurrent across samples. This is known to be true for the base call errors within homopolymers by second generation sequencing technologies and we suppose it to hold for PacBio targeted data.

Based on these hypothesis, we can determine whether an atomic alternative path  $p_a$  is specifically associated with a given sample by capitalizing on the background information contained in the cohort. More specifically, an atomic alternative path  $p_a$  is specific to a sample graph  $G_s$  if it is rare in the whole set of reads for the cohort, denoted  $\mathcal{R}$ . Mutations are Levenshtein edit operations between sample-specific atomic alternative paths and the corresponding reference paths.



Algorithm 1: Alternative path search

```
1  $\mathcal{A} = \emptyset$ 
2  $V_{\text{start}} = \{v \in V^* \mid \deg^+(v) > 0\}$  // start vertices
3  $V_{\text{end}} = \{v \in V^* \mid \deg^-(v) > 0\}$  // end vertices
4 for each  $v_s$  in  $V_{\text{start}}$  do
5   for each  $v_e$  in  $V_{\text{end}}$  do
6     compute  $A$ , the set of alternative paths  $p_a$  between  $v_s$  and  $v_e$ 
       in  $G^*$ 
7     if  $A \neq \emptyset$  then
8       retrieve the best reference path  $p_r$  between  $v_s$  and  $v_e$  in
          $G_{\text{ref}}$ 
9       for each  $p_a \in A$  do
10        compute  $A' = \{\delta_i(p_a) \mid p_a \in A\}$  the set of atomic
          alternative paths
11        for each  $p'_a \in A'$  do
12           $c = r(p'_a)/(r(p'_a) + r(p_r))$ 
13           $\mathcal{A} = \mathcal{A} \cup \langle p'_a, p_r, c \rangle$ 
14 return  $\mathcal{A}$ 
```

We measure the probability of this association between  $p_a$  and  $G_s$  by computing the count ratio  $c(p_a)$  statistics under the null hypothesis of no association between reads and samples. For an observed  $c(p_a)$ , we generate  $N$  random samples  $R_i$  by sampling without replacement reads from  $\mathcal{R}$ . For each of the  $R_i$ , we compute the corresponding ratio  $c_i(p_a)$ , that is the count ratio of  $p_a$  observed in the graph  $G_i^s$  that encodes  $R_i$ . We iterate this step to obtain a distribution of  $c_i(p_a)$  under the null hypothesis, and therefore deduce a  $p$ -value by comparing this distribution and the sample read counts for each alteration. Once we have performed  $N$  resamplings, we can reject the null hypothesis of no association between the sample and the path if less than  $\alpha$ -percent of all  $c_i(p_a)$  are smaller than  $c(p_a)$ , where  $\alpha$  is the required significance level (0.01 in our experiments). In such cases, we identify the altered nucleotides by identifying edit operations (insertions, replacements or deletions) between  $p_r$  and  $p_a$  and the associated  $p$ -value  $p = \frac{|\{c_i(p_a) > c(p_a)\}|}{N}$  for this alteration. Finally, following the observation that most  $c_i(p_a)$  were normally distributed during our simulations and executions, we also determine a – standardized –  $z$ -score for  $c(p_a)$  and only consider alterations for which the ratio  $c(p_a)$  is at least 10 standard deviation away from the mean  $z(c(p_a)) \geq 10$ .

4 Results

Three pipelines based on GATK, VarScan and MICADo (see Supplementary material section 3 for details on how these pipelines were executed) were evaluated on both synthetic and real data.

4.1 Evaluation on synthetic data

We evaluated the ability of the three pipelines to recover artificially inserted somatic mutations. We developed to this end a simple reads simulator based on sampling reads from the TP53 dataset. Although numerous synthetic read simulators for NGS data are available, we opted to use a simple non-parametric sampling-altering scheme based on existing samples in order to preserve as much as possible sequencing biases present within the data. We generated 9245 synthetic datasets by sampling and altering reads from the pool of negative control samples belonging to the TP53 dataset (see section 3.1 and Supplementary section 6 for details). Moreover, any identified variants corresponding to known SNPs (described in Supplementary section 2) were excluded from this analysis.

The three pipelines were evaluated by computing, for each sample, the number of true positives (number of correctly identified mutations), false positives (number of wrongly identified mutations) and false negatives (number of missed mutations). The results are reported in Figure 4. In terms

of precision, GATK and MICADo offer similar performances when more than 10% of the reads are altered. VarScan on the other hand, systematically generates false positives (from 1 to 8 per sample), which is not the case for GATK and MICADo.

When samples contain a limited fraction of altered reads (less than 5%), we observe that the vast majority of mutations are not identified by either GATK or VarScan (the first four classes with “FN+” label depicted in Figure 4 indicate that at least one mutation was missed). For VarScan, this is expected, since we set the threshold of the minimum variant allele frequency to 5%; if not, the results are riddled with false positives. On the contrary, MICADo was run with a lower threshold of 3% minimum variant allele frequency and still adequately identifies both with high precision and recall the majority of mutations even when as few as 3.5% of the total number of reads contain these mutations.

Note that synthetic samples were generated by sampling from the negative control group of the TP53 dataset that were qualified as being “wild-type” by an independent sequencing run as well as by a yeast colony assay test (see section 3.1 for details). Therefore, any variant found in the negative control group is either a known SNP or is considered to be a false positive. Consequently, only variants that were inserted by our read simulator were considered as true positives.

We then evaluated in more details the ability of MICADo to identify single nucleotide variants and small indels (less than 5nt). We report in figure 5 calling results for each of the three possible classes of variants (substitutions denoted by “X”, deletions by “D” and insertions by “I”). We observe that MICADo adequately detects most deletions and insertions, task considered to be difficult for variant calling.

4.2 TP53 targeted data

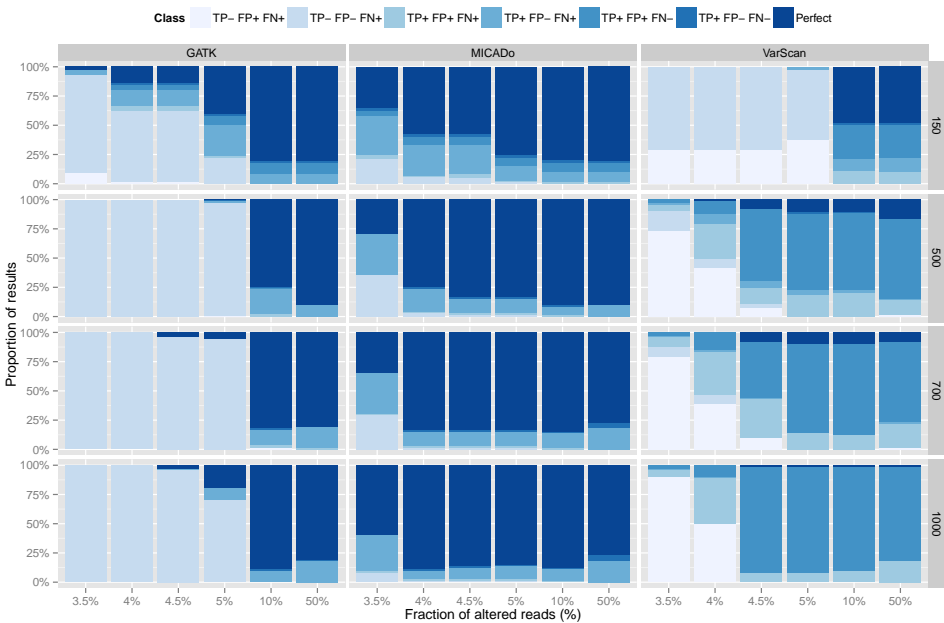
The three pipelines based on MICADo, VarScan and GATK, were used to characterize the mutations of the newly sequenced 48 TP53 samples. See section 3.1 for details on the samples and Supplementary material section 3 for how these pipelines were run. The results are depicted in figure 6 and details of the identified alterations are reported in Supplementary table 1. Note that known polymorphisms affecting the sequencing region were filtered out, either by incorporating known SNPs in the reference graph for MICADo; or by post-processing calls from GATK and VarScan.

We previously reported (Iggo et al., 2013) by independent 454 sequencing that for samples of the difficult group and positive control, we expect one or two significant alterations (mutations) per sample, while we expect no mutations for the negative control group. The only sample that carries two mutations is 276\_1, belonging to the difficult group, where MICADo correctly call one of them (the other one is a false negative).

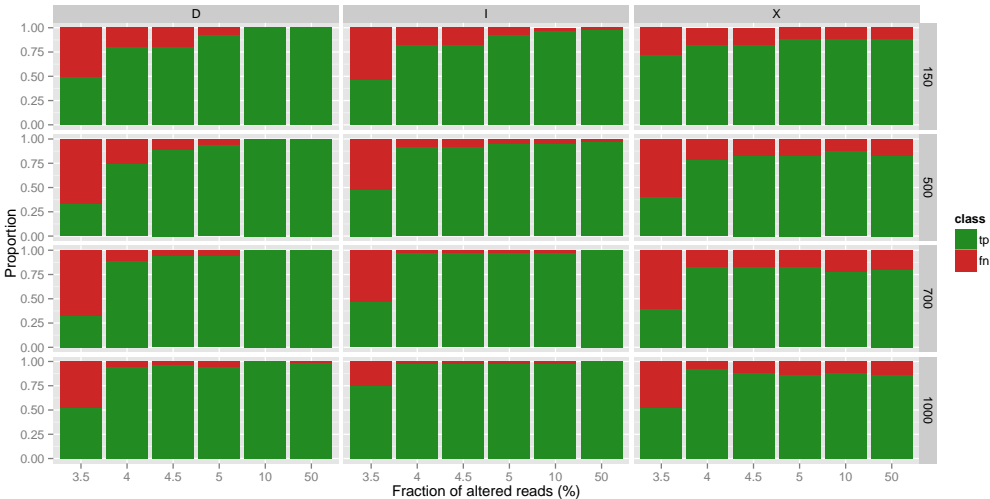
We observe in figure 6 and Supplementary table 1 that both GATK and VarScan exhibit high false positive rate, and that the number of identified alterations seems to be independent of the difficulty category. On the contrary, we observe that MICADo adequately limits the false positive rate and that only one sample from the negative control category is identified as being mutated. Indeed, the sample 256\_1 exhibits a significant alteration while being in the negative control group. This sample is known to contain a missense variant of doubtful functional significance (Iggo et al., 2013).

Closers inspection (see Supplementary table 1) shows that alterations identified as significant for 83\_1 and 83\_2 are in fact false positives (these samples are mutated, but not at the positions identified by MICADo).

We then determined if alterations identified by GATK and VarScan were clustered around hotspots. In our case, we expect loss of function mutations to be uniformly distributed over the transcript. Therefore, if we observe hotspots, these hotspots would likely be due to sequencing errors rather than to alteration recurrence. We thus aggregated the results of the whole cohort and determined weather some alterations were more frequent than others. We detail in table 7 the top 25 most frequently altered locations, as well as the nucleotidic context before and after the putative alteration.



**Fig. 4.** Comparison of three pipelines on synthetic data. Three pipelines, GATK, MICADo and VarScan (vertical panels) were run on synthetic data containing from 1 to 3 inserted alterations, harbored by a varying fraction of altered reads (x-axis) among a total of 150, 500, 700 and 1000 reads (horizontal panels). TP- label indicates no True Positives and TP+ indicates that at least one True Positive was identified. The same notation applies for False Positives (FP) and False Negatives (FN). We identify the results where all the true positives were retrieved, and only them, as “Perfect”. Each result thus belongs to one out of 7 classes (sequential colors) and we report the proportion of results for each class (y-axis).



**Fig. 5.** Proportion of correctly discovered alterations (y-axis) function of the fraction of altered reads (x-axis). Vertical panels report substitutions (“X”), deletions (“D”) and insertions (“I”). Horizontal panels report results for various sample sizes ((150, 500, 700 and 1000 reads). We see that MICADo identifies indels with high precision.

Note that MICADo is absent from this table as by definition it reports no recurrent alterations. We immediately observe in this table that alterations identified by VarScan and GATK are (i) recurrent, with as much as 34 independent samples exhibiting the same deletion, (ii) frequently found in the context of a homopolymer and (iii) that the deleted nucleotide is one of the homopolymer’s nucleotides. This confirms that the sequencing exhibits a higher rate of error in homo-polymeric regions, and that these errors are recurrent across samples.

Note that both VarScan and GATK were run with stringent minimal base quality threshold (cf. Supplementary section 3 for details), and that we observed that quality for bases in these homo-polymers very frequently received the highest possible quality score output by PacBio sequencers. This implies that callers relying on quality score to filter out false positives would not be able to avoid these errors.

In summary, these results demonstrate that several alterations that are significant at the individual level (such as those reported by GATK and

VarScan) are actually resulting from systematic sequencing biases. With the MICADo algorithm, we measure the specificity of each alteration by contrasting it with background samples generated by resampling, thus accounting for the systematic biases.

4.3 FLT3 targeted data

To further test MICADo, we applied it to the publicly available PacBio sequencing targeted on the FLT3 kinase domain. For each patient, we check for absence of mutations in samples before relapse in treatment and we search for at least one mutation in samples after relapse in accordance with (Smith et al., 2012). Normal samples (Normal Control no. 1, no. 2 and no. 3) are not supposed to carry any alterations which was confirmed by MICADo (data not shown). Alteration results obtained by MICADo on the remaining samples were compared with those reported in the original paper, results are shown in table 1.



Fig. 6. Comparison of caller's results for TP53 samples. Three pipelines (vertical panels) were applied on the TP53 samples (x-axis) to identify alterations in the three groups (horizontal panels). We report here the number of identified significant alterations (y-axis) per category (vertical panels) and per sample.

Pos	Before	R	A	Type	After	Count
412	AGGCTGCT	TC	T	D	CCCCCGGT	34
1099	ACGAGCTG	GC	G	D	CCCCCAGG	34
652	ATTCCACA	AC	A	D	CCCCCGCC	29
464	TGCACCAG	GC	G	D	CCCCCTCC	28
581	CACGTACT	TC	T	D	CCCCTGCC	16
729	AGGCGCTG	GC	G	D	CCCCCACC	16
1078	GCAAGAAA	AG	A	D	GGGGAGCC	13
767	TGGTCTGG	GC	G	D	CCCCTCCT	12
1147	GCTCCTCT	TC	T	D	CCCCAGCC	11
1201	AGATCCGT	TG	T	D	GGGCGTGA	11
475	CCTCCTGG	GC	G	D	CCCCTGTC	6
652	ATTCCACA	A	AC	I	CCCCCGCC	6
729	AGGCGCTG	G	GC	I	CCCCCACC	6
502	CTTCCAG	GA	G	D	AAAACCTA	5
1126	GAGCACTG	GC	G	D	CCCAACAA	5
996	AATCTACT	TG	T	D	GGGACGGA	4

Pos	Before	R	A	Type	After	Count
452	ACCGGCGG	GC	G	D	CCCCTGCA	26
652	ATTCCACA	AC	A	D	CCCCCGCC	13
412	AGGCTGCT	TC	T	D	CCCCCGGT	11
422	CCCCGTGG	GC	G	D	CCCCTGCA	10
729	AGGCGCTG	GC	G	D	CCCCCACC	10
1099	ACGAGCTG	GC	G	D	CCCCCAGG	9
767	TGGTCTGG	GC	G	D	CCCCTCCT	7
464	TGCACCAG	GC	G	D	CCCCCTCC	5

Fig. 7. Top 25 “alteration hotspots” along the DNA Binding Domain of TP53 for VarScan (left panel) and GATK (right panel). For each base (column Pos, reference transcript's coordinate), we report in the column Count the number of significant alterations reported by each caller, as well as the context Before, the context After, the Reference sequence and the Alternative sequence. Column Type reports the alteration type: insertions (I) and deletions (D).

In the original paper, authors have manually analyzed a small region restricted to only 4 codons downstream from the ITD region. We have processed the entirety of the targeted sequences, but focused our analysis on the region after the ITD in the same way as in the original paper. The pre-treatment samples all had the percent of altered reads below 0.43%. We have consequently set the number of expected alterations to be 0 (exp. # column in table 1). For the relapse samples we have counted only 1 alteration for each altered position, thus taking into the account only the majority clone in our counting of expected alterations.

Subject #	Exp. pr. #	MICADo pr.	Exp. pr. #	MICADo rel.
1009-003	0	0/0	1	0/0
1011-006	0	0/0	1	1/1
1011-007	0	0/0	2	1/1
1005-004	0	0/1	1	1/2
1005-006	0	0/0	1	1/1
1005-007	0	0/0	1	1/1
1005-009	0	0/0	1	1/1
1005-010	0	0/0	1	1/1

Table 1. Results of MICADo variant calls for FLT3 PacBio sequencing data. For each sample we provide its expected number of mutations for pretreatment (Exp. pr. #) and relapse states (Exp. rel. #) and the number of correct calls (True Positives) (c) over total calls (t) for pretreatment and relapse states obtained by MICADo (MICADo pr. and MICADo rel., respectively) c/t.

As can be seen in table 1 MICADo identified mutations with high precision. Indeed, there are 2 mutation calls that were not previously reported (identified as false positives in the table 1) and 1 false negative. The second mutation in the subject 1011-007 (relapse) is identified by MICADo, but is filtered out by the z-score threshold. The potential false positives in pretreatment and relapse subjects 1005-004 are identical and correspond to a mismatch in a region that has not been analysed in the original paper. For subjects 1009-003 (pretreatment and relapse), 1011-007 (pretreatment and relapse) and 1005-007 MICADo identified large insertions corresponding to the ITD regions with  $k = 30$ . In other subjects the duplication can be identified, given  $k > 30$ .

5 Conclusion

We describe here a novel method for calling mutations in sequencing data produced for a cohort of patients that makes it possible to distinguish between real mutations and other alterations, including sequencing errors. The originality of MICADo is three-fold. First, it is based on de Bruijn graphs and does not require reads to be aligned on a reference sequence, a step known to be particularly error-prone in the case of indels. Second, MICADo uses cohort information and an exact test to correct for systematic and recurrent alterations caused by technical biases. This has the advantage



of allowing correct identification of loss-of-function mutations, which are known to be non-recurrent at the nucleotide level over a cohort, but recurrent in a given sample. Third, by construction, MICADo allows known alterations and multiple reference sequences to be jointly analyzed in a single run.

Our method is especially relevant in the case of PacBio sequencing data that has been previously considered as unsuitable for accurately predicting mutations in clinical applications. MICADo was able to achieve significant improvement over widely used available methods (GATK and VarScan) by controlling both the low false positive and false negative rates.

Our analysis also suggests that PacBio targeted sequencing harbors recurrent errors in homopolymer regions – observation that goes counter to the often admitted hypothesis of uniform distribution of sequencing errors in PacBio reads.

We believe that our method can be a useful addition to the presently available mutation calling tools and could be effectively used in targeted sequencing of heterogeneous samples from cohorts of patients.

## Acknowledgements

The authors would like to thank Eric Rivals for valuable discussions. The authors also acknowledge support from EORTC, the Science for Life Laboratory, the National Genomics Infrastructure funded by the Swedish Research Council, and Uppsala Multidisciplinary Center for Advanced Computational Science for assistance with massively parallel sequencing.

## Funding

This work was supported in part by the SIRIC BRIO. Jonas Bergh's research group was supported by grants from the Swedish Cancer Society, Knut and Alice Wallenberg's fund, the research funds at Radiumhemmet, Karolinska Institutet and Stockholm County Council, BRECT and ALF.

## References

- Allhoff, M., Schönhuth, A., Martin, M., Costa, I. G., Rahmann, S., and Marschall, T. (2013). Discovering motifs that induce sequencing errors. *BMC Bioinformatics*, 14(Suppl 5):S1.
- Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, 26(12):i318–i324.
- Bonnefoi, H., Piccart, M., Bogaerts, J., Mauriac, L., Fumoleau, P., Brain, E., Petit, T., Rouanet, P., Jassem, J., Blot, E., Zaman, K., Cufer, T., Lortholary, A., Lidbrink, E., André, S., Litière, S., Lago, L. D., Becette, V., Cameron, D. A., Bergh, J., and Iggo, R. (2011). {TP53} status for prediction of sensitivity to taxane versus non-taxane neoadjuvant chemotherapy in breast cancer (eortc 10994/big 1-00): a randomised phase 3 trial. *The Lancet Oncology*, 12(6):527 – 539.
- Brosh, R. and Rotter, V. (2009). When mutants gain new powers: news from the mutant p53 field. *Nature Reviews Cancer*, 9(10):701–713.
- Carneiro, M. O., Russ, C., Ross, M. G., Gabriel, S. B., Nusbaum, C., and DePristo, M. A. (2012). Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics*, 13(1):375.
- Chelala, C., Khan, A., and Lemoine, N. R. (2009). SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, 25(5):655–661.
- Chin, C.-S., Sorenson, J., Harris, J. B., Robins, W. P., Charles, R. C., Jean-Charles, R. R., Bullard, J., Webster, D. R., Kasarskis, A., Peluso, P., et al. (2011). The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine*, 364(1):33–42.
- Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991.
- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498.
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138.
- Gilliland, D. G. and Griffin, J. D. (2002). The roles of FLT3 in hematopoiesis and leukemia. *Blood*, 100(5):1532–1542.
- Iggo, R., Rudewicz, J., Monceau, E., Sevenet, N., Bergh, J., Sjoblom, T., and Bonnefoi, H. (2013). Validation of a yeast functional assay for p53 mutations using clonal sequencing. *The Journal of Pathology*, 231(4):441–448.
- Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature genetics*, 44(2):226–232.
- Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research*, 22(3):568–576.
- Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.
- Leggett, R. M. and MacLean, D. (2014). Reference-free SNP detection: dealing with the data deluge. *BMC genomics*, 15(Suppl 4):S10.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*, 17(1):pp. 10–12.
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernysky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303.
- Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harbor perspectives in biology*, 2(1):a001008.
- Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J., and Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics*, 15(2):256–278.
- Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753.
- Pirooznia, M., Kramer, M., Parla, J., Goes, F., Potash, J., McCombie, W. R., and Zandi, P. (2014). Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8(1):14.
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., and Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1):341.
- Guillaume Rizk, Anaïs Gouin, Rayan Chikhi, and Claire Lemaitre. Mindthegap: integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24):3451–3457, 2014.
- Robasky, K., Lewis, N. E., and Church, G. M. (2014). The role of replicates for error mitigation in next-generation sequencing. *Nature Reviews Genetics*, 15(1):56–62.
- Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, 14(5):R51.
- Schult, D. A. and Swart, P. (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pages 11–16.
- Smith, C. C., Wang, Q., Chin, C.-S., Salerno, S., Damon, L. E., Levis, M. J., Perl, A. E., Travers, K. J., Wang, S., Hunt, J. P., et al. (2012). Validation of ITD mutations in FLT3 as a therapeutic target in human acute myeloid leukaemia. *Nature*, 485(7397):260–263.
- the International Cancer Genome Consortium Mutation Pathways and of the Bioinformatics Analyses Working Group, C. S. (2013). Computational approaches to identify functional genetic variants in cancer genomes. *Nature Methods*, pages 723–729.
- Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., Lemaitre, C., and Peterlongo, P. (2014). Reference-free detection of isolated SNPs. *Nucleic Acids Research*.
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426.
- Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., Dahlman, K. B., Pao, W., Zhao, Z., et al. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med*, 5(10):91.
- Warden, C. D., Adamson, A. W., Neuhausen, S. L., and Wu, X. (2014). Detailed comparison of two popular variant calling packages for exome and targeted exon studies. *PeerJ*, 2:e600.
- Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875.
- Yi, M., Zhao, Y., Jia, L., He, M., Kebebew, E., and Stephens, R. M. (2014). Performance comparison of SNP detection tools with Illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic acids research*, page gku392.
- Zook, J. M., Chapman, B., Wang, J., Mittelman, D., Hofmann, O., Hide, W., and Salit, M. (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature biotechnology*, 32(3):246–251.