Strand Orientation Bias Detector (SOBDetector) to remove artifacts from sequencing data of formalin fixed samples.

Miklos Diossy[1,2], Zsofia Sztupinszki[2], Marcin Krzystanek[2], Anders Gorm Pedersen[1], Aron C. Eklund[1], István Csabai[3], Zoltan Szallasi[2,4,5,6]

[1] Department of Bio and Health Informatics, Technical University of Denmark, Kemitorvet 208, 2800 Lyngby, Denmark

[2] Danish Cancer Society Research Center, Copenhagen, Denmark.

[3] Department of Physics of Complex Systems, Eötvös Loránd University, Budapest, Hungary

[4] Computational Health Informatics Program, Boston Children's Hospital, USA, Harvard Medical School, Boston, USA

[5] 2nd Department of Pathology, MTA-SE NAP, Brain Metastasis Research Group, Hungarian Academy of Sciences, Semmelweis University, Budapest, Hungary

[6] Correspondence should be sent to:
Zoltan Szallasi, Computational Health Informatics Program (CHIP) Boston Children's Hospital, Harvard Medical School, 300 Longwood Ave., Boston Massachusetts, USA, 02215, e-mail: Zoltan.szallasi@childrens.harvard.edu, +1-617-355-2179

Abstract

Summary: Due to its effectiveness and simplicity, formalin-fixed paraffin-embedded FFPE tissue processing is the most common approach for tissue specimen storage. It is well known however, that formalin reacts with nitrogenous bases (especially with cytosine) and causes DNA lesions, which are sources of sequencing artifacts. This makes the next generation sequencing (NGS) analyses of these specimens challenging.

Here we present SOBDetector, a simple, user-friendly application that can remove the majority of such artifacts from the list of detected variants, by analyzing the orientation of paired end sequences that are spanning the variants' locations. While it was originally created for human somatic mutation filtration, it can be incorporated into any variant calling pipelines of any species.

Availability and Implementation: SOBDetector is implemented in java 1.8, and it is freely available at: www.github.com/mikdio/SOBDetector

Contact: mikdio@bioinformatics.dtu.dk, Zoltan.Szallasi@childrens.harvard.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Formalin fixed paraffin embedded tissue is a valuable resource of translational clinical research. Next generation sequencing of DNA extracted from such samples is often performed for exploratory or validation purposes. It is well-known, however, that the chemical processes fixating the tissue material introduces a significant amount of DNA damage that manifests as misread base pairs during sequencing. The most prominent of those is the deamination of cytosines, which introduces C→T/G→A mutations (Oh *et al.*, 2015).

Since formalin likely affects only one of the strands (e.g. a C|G pair becomes T|G), a paired-end next generation sequencing approach can be utilized for an additional filtering step (Supplementary Figure 1). When counting not just the number of reads that support the alternative alleles, but also their relative orientation (Forward-Reverse:FR or Reverse-Forward:RF), the FFPE artifacts will likely have a strand orientation bias towards one of the directions, while true mutations should have approximately the same amount of FR and RF reads.

This article introduces a tool called Strand Orientation Bias Detector (SOBDetector), which reanalyzes the mutations stored in Variant Call Format (VCF) file, and by using the original binary alignment (BAM) files, evaluates, whether the reads supporting the alternate alleles show a strand orientation bias.

2 SOBDetector

SOBDetector, originally created to remove FFPE-introduced artifacts from a list of human somatic variants, can also be used in any paired-end single-sample variation calling pipelines of any species (Figure 1A). It is implemented in java 1.8, and apart from the presence of a java runtime environment, its only software requirement is samtools (version 1.6 or higher, (Li *et al.*, 2009))] (Supplementary Figure 2).

2.1 Method:

As a measure of the strand orientation, we defined the strand orientation bias (SOB) score, as the following:

$$SOB = \frac{\#ofF1R2alt - \#ofF2R1alt}{\#ofF1R2alt + \#ofF2R1alt},$$

where *#ofF1R2alt* is the number of alternate allele supporting reads that are in the forward 1st, reverse 2nd orientation.

SOB score lies in the [-1,1] interval, and the closer it is to the two extrema, the higher the probability, that the corresponding variant is a sequencing artifact.

SOBDetector takes a variant-containing file (VCF or tab-delimited file) and its related BAM as input. It counts the two orientation quantities, calculates the SOB scores for every variant (online supplementary notes, section 3) and, finally appends the resulting columns to the output VCF.

2.2 Validation:

In order to validate our approach and determine the optimal thresholds on the SOB scores, we used a subset of TCGA patients (http://cancergenome.nih.gov, N = 47), who had both fresh frozen (FF) and FFPE specimens derived from the same tumor (Supplementary figure 3). For the purpose of somatic mutation calling we have selected MuTect 2 v3.8 (Van der Auwera *et al.*, 2013) .

Since most of these donors had multiple FF and/or FFPE tumors, and in some cases even multiple germline samples, we could ensure the robustness of our results by analyzing 235 somatic VCFs.

We found that FFPE samples, unlike FF samples, were dominated by variants that were supported by fewer than 6 alternate allele supporting reads (Supplementary figure 4). In accordance with our assumptions, the great majority of the SNVs with higher than 0.8 SOB scores were coming from FFPE tissues (Supplementary figures 5-7). We also found, that by setting the threshold on the SOB scores at 0.8 and on the number of alternate allele supporting reads at 6, the 2-dimensional distributions of the samples in the SOB score – tumor logarithm of odds (LOD) spaces became more similar (2-dimensional Kolmogorov-Smirnov statistic before filtering: D = 0.368 and after filtering: D = 0.146, Figure 1B). Using this filter, the number of mutations in the breast, uterine, prostate and bladder carcinoma FFPE cases (group A) shrank significantly, while the FF samples were affected negligibly (Supplementary notes, section 4.3). In the colorectal, kidney and lung carcinoma cases (group B) however, the filtration process affected both the FF and FFPE-derived variants at approximately the same level (Supplementary figures 8-13). It is

possible that samples of the latter group might not have been exposed to the effects of formalin for as long as the samples in the former group.

2.3 Concordance:

We have calculated the relative concordance (r) between the number of mutations found in the corresponding FF and FFPE samples before and after the filtration step, and the change in this value (Δr) was analyzed between all the possible sample-type combinations (supplementary notes, section 4.4). In accordance with the previous results, Δr was positive in almost all of the FFPE-FF analytes, and the distribution of Δr was significantly different in both groups (A and B) of the samples (Figure 1C, Supplementary figures 14-18).

3 Conclusions:

In summary, SOBDetector extends the common variant filtering steps by additional, strand orientation bias-specific filters, it can be incorporated into any variant calling pipelines, and its application significantly reduces the number of FFPE-introduced sequencing artifacts from formalin-fixed specimens.

Acknowledgments and funding

This work was supported by the Research and Technology Innovation Fund (KTIA_NAP_13-2014-0021 to Z.S.); Breast Cancer Research Foundation (BCRF-17-156 to Z.S.) and the Novo Nordisk Foundation Interdisciplinary Synergy Programme Grant (NNF15OC0016584 to Z.S.) and by Tesaro Inc.; The Danish Cancer Society grant (R90-A6213 to MK).

Conflict of Interest

We have no conflict of interest to declare.

Figures:

Figure 1.:
A: Standard SOBDetector usage pipeline. The tool uses a BAM file (germline or somatic) and its corresponding variant calling output: a VCF or a tabulated file..

B: Tumor LOD and SOB score distribution of the TCGA FF and FFPE samples, that were used for validation. The D statistic (D), p-value (p) and effective sample size (n) of the 2-dimensional Kolmogorov Smirnov test are presented. After the filtration step, the two histograms became more similar.

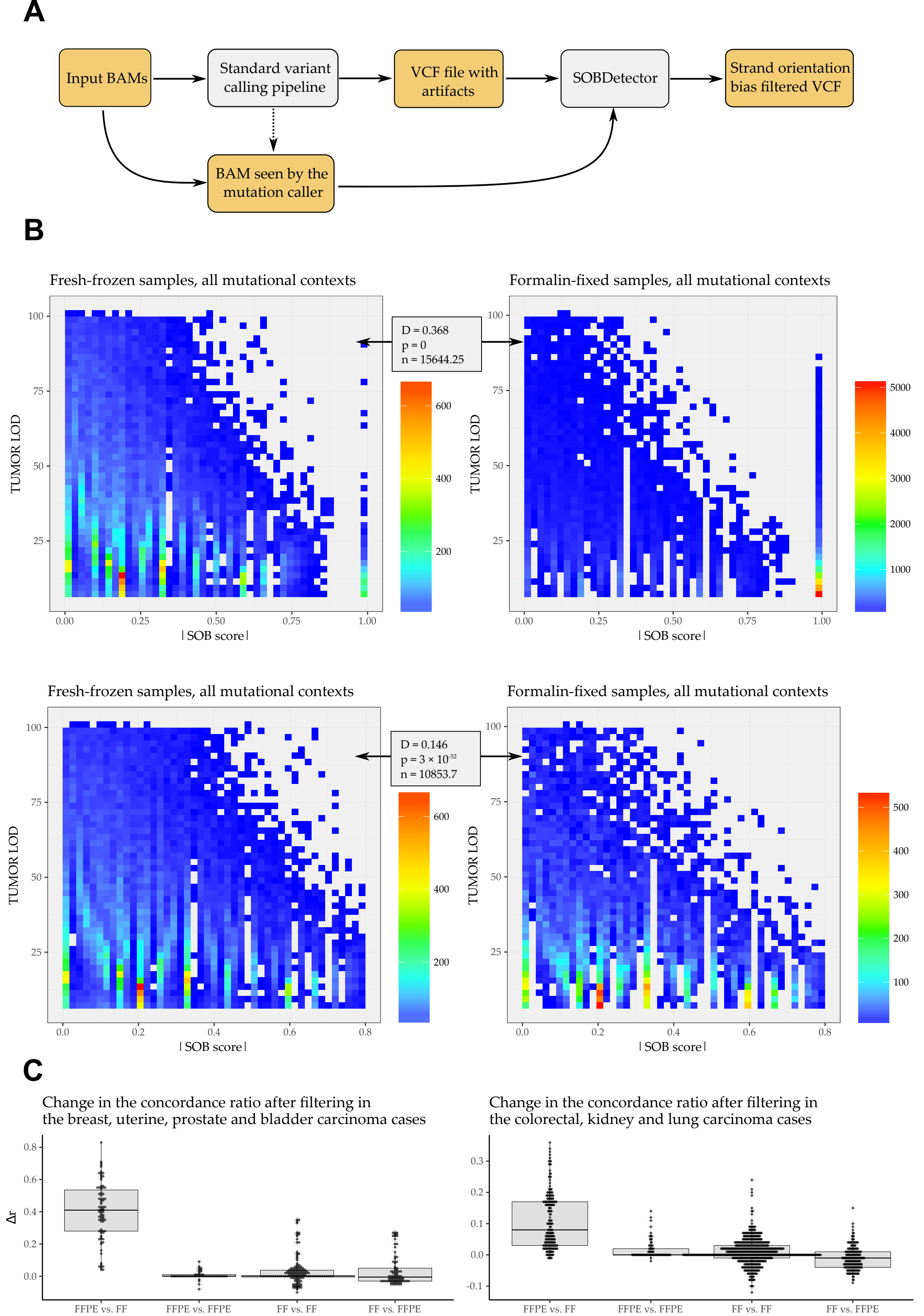
C: Change in the relative concordance ratios between the corresponding samples before and after filtering. Since most of the patients had multiple FFPE/FF specimens, it was possible to compare the change in concordance between FF-FF, and FFPE-FFPE samples as well. FFPE-FF concordance had the highest increase, with the mean value significantly higher than the other 3 classes' ($p < 1E-7$, Tukey ad-hoc test).

References:

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Oh,E. *et al.* (2015) Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. *PLoS ONE*, **10**, e0144162.

Van der Auwera,G.A. *et al*. (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, **43**, 11.10.1–33.


A
Input BAMs
Standard variant calling pipeline
VCF file with artifacts
SOBDetector
Strand orientation bias filtered VCF
BAM seen by the mutation caller
B
Fresh-frozen samples, all mutational contexts
Formalin-fixed samples, all mutational contexts
D = 0.368
p = 0
n = 15644.25
TUMOR LOD
| SOB score |
Fresh-frozen samples, all mutational contexts
Formalin-fixed samples, all mutational contexts
D = 0.146
p = 3 × 10^-52
n = 10853.7
TUMOR LOD
| SOB score |
C
Change in the concordance ratio after filtering in the breast, uterine, prostate and bladder carcinoma cases
Change in the concordance ratio after filtering in the colorectal, kidney and lung carcinoma cases
Δr
FFPE vs. FF
FFPE vs. FFPE
FF vs. FF
FF vs. FFPE