Published online 9 February 2021

doi: 10.1093/nargab/lqab002

Increased yields of duplex sequencing data by a series of quality control tools

Gundula Povysil[1,*,†], Monika Heinzl[1,†], Renato Salazar[1], Nicholas Stoler[2], Anton Nekrutenko[2] and Irene Tiemann-Boege[1,*]

[1]Institute of Biophysics, Johannes Kepler University, 4020 Linz, Austria and [2]Graduate Program in Bioinformatics and Genomics, The Huck Institutes for Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA

Received December 14, 2020; Editorial Decision January 04, 2021; Accepted January 05, 2021

ABSTRACT

Duplex sequencing is currently the most reliable method to identify ultra-low frequency DNA variants by grouping sequence reads derived from the same DNA molecule into families with information on the forward and reverse strand. However, only a small proportion of reads are assembled into duplex consensus sequences (DCS), and reads with potentially valuable information are discarded at different steps of the bioinformatics pipeline, especially reads without a family. We developed a bioinformatics toolset that analyses the tag and family composition with the purpose to understand data loss and implement modifications to maximize the data output for the variant calling. Specifically, our tools show that tags contain polymerase chain reaction and sequencing errors that contribute to data loss and lower DCS yields. Our tools also identified chimeras, which likely reflect barcode collisions. Finally, we also developed a tool that re-examines variant calls from raw reads and provides different summary data that categorizes the confidence level of a variant call by a tier-based system. With this tool, we can include reads without a family and check the reliability of the call, that increases substantially the sequencing depth for variant calling, a particular important advantage for low-input samples or low-coverage regions.

INTRODUCTION

The identification of ultra-rare variants has been relevant in a range of diverse fields, such as cancer research, tumor development and residual disease, somatic mosaicism, evolutionary biology and epidemiology (reviewed in (1)). As such, the last decade has seen an extensive development of technologies for the identification of variants occurring at very low levels (10^{-4}–10^{-9}). Different next generation sequencing protocols based on short paired-end reads (150–300 nucleotides) have been developed for this purpose. To overcome the high error rates (0.1–2%) associated with this sequencing platform (2), different approaches for library preparation have been published that include the addition of tags during library preparation either by a random sequence in the amplification primers (3,4) or the hybridization of indexed molecular inversion probes (5,6). The common strategy of these approaches is that reads are grouped into families, out of which a consensus sequence is built. Real substitutions present in the majority of the reads of a family can be distinguished from polymerase chain reaction (PCR) and sequencing errors (5,3,6,4). Alternatively, family members can be created by the circularization of small DNA fragments followed by rolling circle amplification (7).

This grouping strategy reduces error rates to $<10^{-5}$; however, amplifiable DNA lesions (such as 8-oxoguanine, or deaminated cytosine or 5-methylcytosine) affect the detection limits because they cannot be distinguished from true variants (8). This is resolved in duplex sequencing (DS) (9,10,4), a strategy that tags both strands of the DNA by the ligation of adapters with a random barcode (Figure 1). The paired-end reads are then grouped into families or single strand consensus sequences (SSCS) representing the forward (*ab*-SSCS) or reverse (*ba*-SSCS) strand that are then re-united into the original duplex consensus sequence (DCS). While a substitution is present in both DNA strands, errors due to DNA lesions are only found in one DNA strand and can be distinguished in DS (8), making DS currently the method with the lowest error rate (1,2,4).

However, DS is still quite costly and only a fraction of the input material (1%) results in a DCS (11,2). Changes to the library preparation, such as CRISPR/Cas targeted digestion, reduces the number of amplification steps and increases the number of DCS per input material (6–12%)

[*]To whom correspondence should be addressed. Tel: +43 732 2468 7620; Fax: +43 732 2468 27620; Email: irene.tiemann@jku.at
Correspondence may also be addressed to Gundula Povysil. Tel: +1 212 305 0754; Email: povysil@bioinf.jku.at
[†]The authors wish to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Present address: Gundula Povysil, Institute for Genomic Medicine, Columbia University Irving Medical Center, New York, NY 10032, USA.

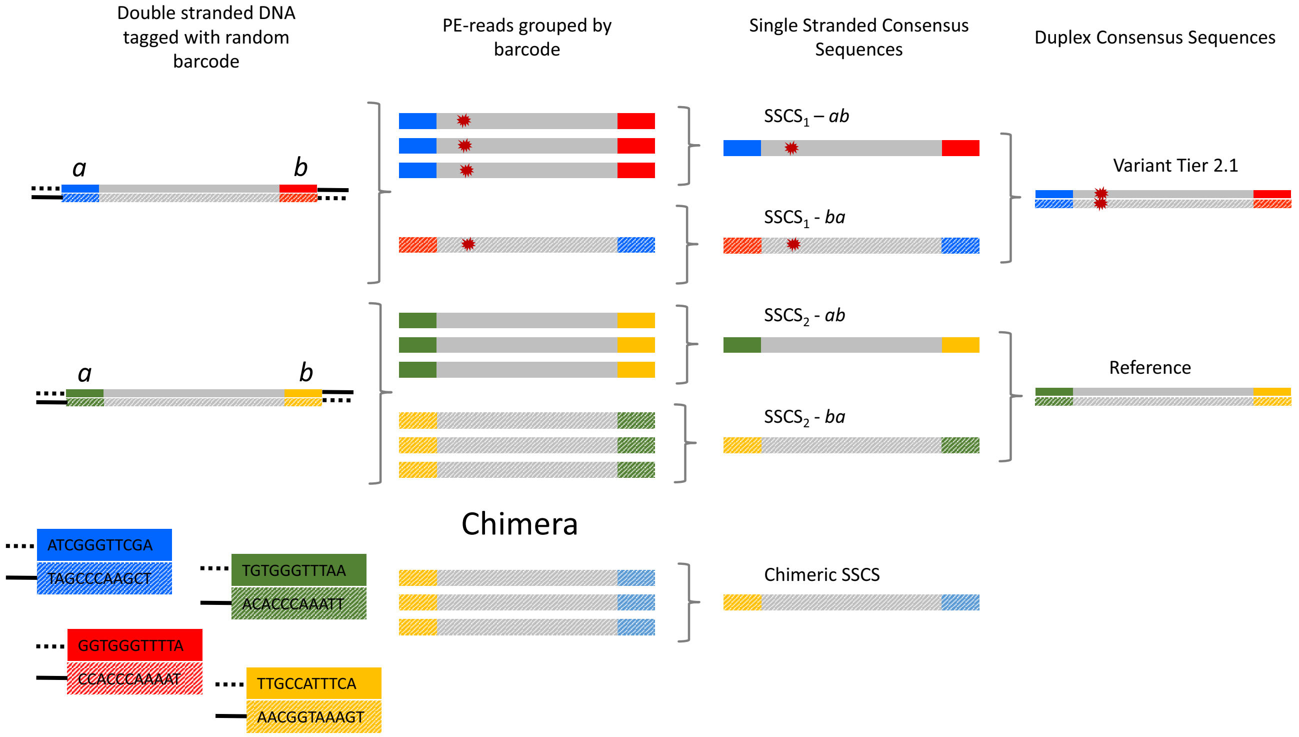

Figure 1. Schematic representation of DS. An adaptor with a unique barcode is ligated to each end (*a* or *b*) of a double-stranded DNA molecule. In this example, adaptors contain 10 random base pairs as exemplified in the red, green, blue, and yellow parts. During library preparation multiple copies are created from the original DNA by PCR that are then sequenced. PE reads with the same tag (combination of the *a* and *b* barcodes) in either *ab* or *ba* orientation are grouped together into a family or SSCS. Complementary SSCS (e.g. SSCS$_1$-*ab* + SSCS$_1$-*ba*) are united into a DCS and substitutions found in both complementary SSCS are classified as true variants. Usually, a minimum of three reads per SSCS are required for consensus building and SSCS with single reads are discarded (e.g. SSCS$_1$-*ba*). In our analysis, we include these and validate the variant calling by a tier-based system. Chimeras form by template switching or by incomplete extensions in PCR and are identified by sharing an identical *a* or *b* end with another SSCS family. These chimeric SSCS will unlikely have a complementary partner and might not form a DCS.

(11). Also changes in the bioinformatic pipeline have improved the data output such as using a reference-free approach, (e.g. Du Novo (12)), which avoids mapping the reads to a reference genome, as done in other pipelines (e.g. UMI-tools, (13)) for the consensus building step. Other tools such as Calib (14) consider not only the barcode of reads, but cluster the reads based also on the sequence of the reads. Moreover, error correction in the barcode helps to re-unite otherwise 'lost' reads into their respective families (15).

Modifications in the library preparation steps, as well as the data analysis could increase the DCS coverage, but this requires a better understanding of the sequencing data produced in DS. Moreover, tools for building DCS require a series of decision taking steps to ensure a low number of false positives. Quality control (QC) tools can enable informed decisions about the parameter settings during consensus building. These include deciding on the smallest number of reads used to build a family or SSCS, the proportion of reads within the family carrying the alternative base, the number of errors allowed in the tag and thresholds used to filter out low-quality bases. Currently, most of these steps are performed automatically using a 'one-fits-all' approach with settings that ensure that a low number of false positives or low-quality consensus reads end up in the DCS. The downside is that with a very conservative setting, more data are lost.

Here, we created a series of QC tools (tag distance (TD), chimera analysis (CA), family size distribution (FSD) and variant analyzer (VAR-A)) that can be implemented at different steps of consensus building and have been tested within Du Novo (12) and compared against the output obtained with the software developed by (11). The purpose of these tools is: (i) Allow for an informed decision as to the best analysis parameters for a particular dataset (currently done by trial and error). (ii) Minimize the number of false positives and false negatives, and at the same time, maximize the number of consensus calling (DCS). (iii) Allow variant calling with more relaxed parameters in the consensus building steps since calls are re-evaluated by a series of summary data validating a variant by a tier-based system that can then be manually examined.

MATERIALS AND METHODS

Family size distribution (FSD)

Here, we define a tag as a combination of the upstream and downstream barcodes of the DNA fragment. Each tag represents a family of paired-end sequences forming SSCS. The FSD analyzes the family size associated with a tag, that is the number of reads per tag, and produces several histograms with the distributed family sizes. We first trimmed the barcodes from all sequencing reads generating a list of e.g. 10 + 10 barcode combinations that were arranged in lexicographic order and then counted the number of times each combination appeared in this list (family size).

Tag distance (TD) analysis

The TD was estimated as reported in (15). We used the same list of tags as already described in the FSD. Since the datasets contained more than one million tags, the comparison of all tags was computationally too demanding. Thus, we parallelized the algorithm and selected 1000 random tags from the dataset and compared them to the whole dataset (189 675, 1 341 763, 138 631 and 1 106 303 tags after barcode correction for the PF1-CRISPR, PF1-Standard, PF2-CRISPR and PF2-Standard datasets, respectively). For the DCS tags, the smaller sizes made it possible to use the complete dataset (15 855, 73 972, 22 456 and 126 714 tags after barcode correction, respectively). We have verified that a sample of 1000 tags (0.1% of the data) is representative for the whole dataset (see Supplementary Figure S2). At each comparison we calculated the number of differences (TD) and reported only the smallest number of differences (minimum TD) observed with any other tag. The distances between tags were calculated using equation 1, where $D_{i,j}$ is the number of sites where X_i and X_j do not match, k is the index of the respective site out of the total number of sites *n*. A detailed guideline for using the TD analysis is described in Supplemental Note 1.

$$Di,\, j = \sum_{k=0}^{n} \left(X_{ik} \neq X_{jk} \right) \tag{1}$$

The output of the tool is a plot of the minimum TD (smallest number of differences) between tags as a frequency histogram categorized after the family sizes.

Chimera analysis (CA)

We have extended the TD tool by the so-called *CA* which allows the identification of chimeric families in the sequencing data. Here, the analysis is described only for one tag in detail, but we repeat this process for 1000 tags (default sample size for TD analysis). First, we split the tag into its upstream and downstream barcode (named *a* and *b*) and compare barcode *a* with all other *a* barcodes of the families in the dataset (∼1 million families). We estimate the sequence distance (TD) among the *a* barcodes and select those tags that have the smallest number of differences ($\mathrm{TD}a_{\min}$) and then calculate from the subset the TD of the *b* barcode. The tags with the largest number of differences are extracted to estimate the maximum TD ($\mathrm{TD}b_{\max}$). The process is repeated starting with the *b* barcode instead and estimates $\mathrm{TD}a_{\max}$ and $\mathrm{TD}b_{\min}$. Next, we calculate the absolute difference between $\mathrm{TD}a_{\min}$ and $\mathrm{TD}b_{\max}$ equal to *delta TD*.

$$delta\ TD = |\mathrm{TD}a_{\min} - \mathrm{TD}b_{\max}|,\quad |TDa_{\max} - \mathrm{TD}b_{\min}| \tag{2}$$

If the same *a* barcode is observed in combination with several different *b* barcodes (as would be expected in a chimera), then *delta TD* will be large (low $\mathrm{TD}a_{\min}$ and high $\mathrm{TD}b_{\max}$) since the TD is contributed only by one of the barcodes (a or b) as the other part is identical (TD = 0). In order to normalize the values between comparisons, we use the relative *delta TD* defined as the ratio of *delta TD* to the sum of the TD of each barcode (TD a+ b).

$$relative\ delta\ TD = \max\left(\frac{|TDa\min - TDb\max|}{TDa\min + TDb\max},\ \frac{|TDa\max - TDb\min|}{TDa\max + TDb\min}\right) \tag{3}$$

For chimeras, the larger of the two *relative delta TD* values is expected to be one, since only one part contributes to the TD. Note that barcode correction was performed before the CA to remove tags with errors from the chimera count. For more information on this analysis see Supplemental Note 2. The *CA* can also be used considering only tags that form DCS. A detailed guideline for using the CA is described in Supplemental Note 1. Note, that a barcode collision event would have the same relative delta TD of one. A barcode collision event happens, when the same upstream barcode is associated with different downstream barcodes because more fragments of a library get sequenced than available different barcodes. A high chimera rate thus may indicate a high rate of barcode collisions and should lead to a re-assessment of the amount of input DNA used.

Variant Analyzer (VAR-A)

Du Novo's analysis was performed with the following settings: barcode correction = 1 and family size = 1 and 3. This was followed by a trimming step (16) of the 3' and 5' ends of the DCS with 10 nucleotides and the alignment to the human genome assembly GRCh38/hg38 using *BWA-MEM* (17) and *BamLeftAlignIndels* (18). Finally, we performed the variant calling with the *FreeBayes* (19) variant caller followed by the tool *VcfAllelicPrimitives* (19). Only variants with a minimum read depth of 100 reads were kept. A detailed workflow of the analysis can be seen in Supplementary Figure S1. Each variant called by Du Novo with its subsequent analysis steps was re-analyzed by VAR-A. First, all tags of the DCS identified to carry a variant were extracted from the bam file obtained in the alignment step. Subsequently all PE reads of these tags were extracted from the data and stored in a fastq file. Further, PE reads were trimmed using *Trimmomatic* (20) with default settings and aligned to the reference using *BWA-MEM*. Finally, VAR-A outputs for each variant, tag, mate and direction (*ab*/*ba*) several statistics: e.g. the number and fraction of reads with reference and alternate allele, as well as the number of unaligned reads and low-quality reads (Phred-scaled base quality score < 20). If both mates overlap a variant, the second mate can either provide additional support for identifying a true variant or help to identify false positive calls. Therefore, confident variant calls become possible even if some of the families have very small sizes (1–2 reads). Furthermore, the output includes the median position of the variant within the reads and information about the number of SSCS carrying the variant, if this variant is a chimera, as well as, all other variants reported for the same tag (variants that are in-phase). To help the user identify high from low confidence calls, we developed a tier system labeling each variant (Table 1 and Supplementary Table S1).

Table 1. Definition of the tier system in the VAR-A tool

Tier 1.1	both *ab* and *ba* SSCS present (>75% of the sites with alternative base) and minimal FS ≥ 3 for both SSCS in at least one mate
Tier 1.2	both *ab* and *ba* SSCS present (>75% of the sites with alt. base) and mate pair validation (min FS = 1) and minimal FS ≥ 3 for at least one of the SSCS
Tier 2.1	both *ab* and *ba* SSCS present (>75% of the sites with alt. base) and minimal FS ≥ 3 for at least one of the SSCS in at least one mate
Tier 2.2	both *ab* and *ba* SSCS present (>75% of the sites with alt. base) and mate pair validation (min FS = 1)
Tier 2.3	both *ab* and *ba* SSCS present (>75% of the sites with alt. base) and minimal FS = 1 for both SSCS in one mate and minimal FS ≥ 3 for at least one of the SSCS in the other mate
Tier 2.4	both ab and ba SSCS present (>75% of the sites with alt. base) and minimal FS = 1 for both SSCS in at least one mate
Tier 3.1	both *ab* and *ba* SSCS present (>50% of the sites with alt. base) and recurring mutation on this position
Tier 3.2	both *ab* and *ba* SSCS present (>50% of the sites with alt. base) and minimal FS ≥ 1 for both SSCS in at least one mate
Tier 4.1	variants at the start or end of the DCS (median position of the variant within the PE reads forming the DCS)
Tier 4.2	variants where the mates contain contradictory information (one mate carries the reference allele whereas the other mate the alternative allele)
Tier 5	other

Tiers 1.2–2.4 include SSCS with small family sizes (<3) and would be discarded by the regular pipeline, yet variants are likely real and are confirmed by both forward (*ab*) and reverse (*ba*) SSCS.

RESULTS AND DISCUSSION

Datasets

In order to test our tools, we used a previously published dataset produced by (11), who sequenced the *TP53* exonic regions with DS using the standard protocol (2) and a targeted genome fragmentation approach based on CRISPR/Cas9 digestion. In particular, we focused on the following four libraries: PF1-CRISPR, PF1-Standard, PF2-CRISPR and PF2-Standard. The main difference between PF1-Standard and PF2-Standard is the starting amount of genomic DNA with ∼10 or 3 μg and the allele frequency of variant chr17:7674230C>T (∼68% versus ∼1% in PF1 and PF2, respectively). Note that the amount of starting material used in the PF1-CRISPR and PF2-CRISPR libraries was 100 ng of DNA. In addition, we produced libraries spiked with known amounts of DNA with the alternative sequence added to wild type at different frequencies to validate the accuracy of our analysis.

Quality control tools

Our QC-tools comprise the following analysis: (i) TD, (ii) TD with the CA, (iii) FSD and (iv) VAR-A. The TD, CA and FSD tools analyze the tag composition extracted from the paired-end reads (PE reads) and are useful for deciding on parameter settings in the bioinformatics pipeline. The tag composition also provides important insights on how different library preparation protocols affect the yields of duplex consensus data. VAR-A provides a comprehensive summary of the called variants by re-analyzing the PE reads, such that the calling of rare variants is verified and borderline cases can be manually inspected. Here we present the results of the tools implemented in the user-friendly Galaxy environment following a general pipeline (Supplementary Figure S1) that can be implemented as part of the DS analysis (see Supplementary Note 1). The workflow presented here uses the Du Novo framework for consensus building, however, our tools are very generic and can also be used with other duplex analysis outputs.

Tag distance (TD)

In DS, each paired-end sequencing read (PE read) is tagged by an upstream (*a*) and downstream (*b*) random barcode resulting in a tag that labels either the forward (*ab*) or the reverse strand (*ba*) of the input DNA (see Figure 1). All the reads containing the same tag (either in the *ab* or *ba* orientation) are grouped into a family or SSCS for the forward (*ab*) or the reverse strand (*ba*). Note, that we refer to tag as the sum of the upstream and downstream barcodes. The DCS is then formed by uniting complementary *ab*+*ba* SSCS. In the library preparation, a large number of unique tag combinations (4^{10+10} or 1.1 × 10^{12}) is used to label the DNA templates (e.g. 10^6 templates). Thus, the probability that two different sequence templates will have the same pair of barcodes is very small; although, as more templates are labeled (e.g. random fragmentation of larger input genomes) this probability will increase (barcode collisions).

The TD tool analyzes the number of nucleotide differences among tags, also known as Hamming distance (21). The tool compares the number of sequence differences of a subset of 1000 tags with the rest of the tags in the dataset and plots the smallest number of differences with another tag (minimum TD) as a histogram stratified by family size (Figure 2). Note that a sample subset of 1000 tags comprise ∼0.1% of the sample size; however, this sample size is representative given that more data (e.g. 1% or 10%) rendered very similar results (see Supplementary Figure S2), but was computationally more time-intensive. For example, for the PF2-CRISPR dataset, the analysis of a subset of 1000 tags took <2.5 min; whereas, the full dataset with 188 354 tags took ∼8 h (see Supplementary Table S2).

Our TD tool is an informative tool to evaluate the optimal tag size for a particular library. This aspect is rather important, since libraries with a very large number of fragments should use longer tags to ensure that the same barcode is not used with more than one fragment; whereas, in smaller libraries longer tags waste sequencing data. This can be nicely illustrated with our tool when comparing the PF2-CRISPR and PF2- and PF1-Standard datasets since the different TD distribution is particularly notorious between these two libraries. These two libraries used slightly different

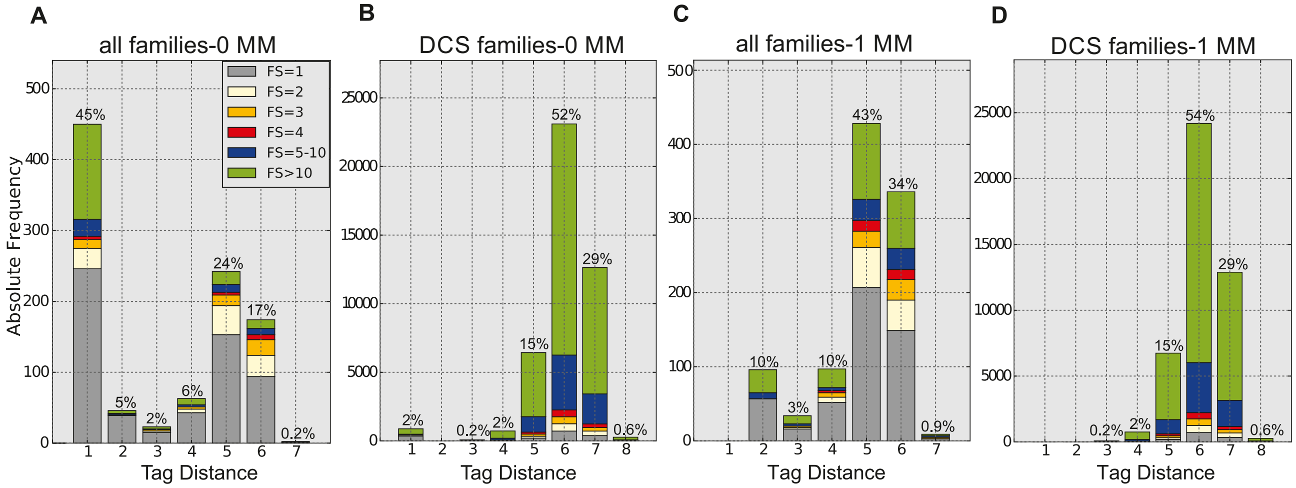

Figure 2. (**A**) TD distribution in the PF2-CRISPR library. The TD was estimated for a subset of 1000 tags, each tag representing a family or SSCS. The legend denotes the family size (FS) e.g. the number of reads with the same tag grouped into a family (**B**) TD distribution using tags that can form a DCS (complementary *ab*+*ba* tags). Note that in this case all tags in the DCS dataset were used for the estimation of the TD. (**C**) TD distribution after barcode correction allowing one mismatch in the tag. TDs smaller or equal to the implemented barcode correction parameter are now reduced to zero. (**D**) TD distribution after barcode correction using tags that form DCS.

library preparation protocols rendering dissimilar numbers of sequenced tags: the PF1-Standard library was prepared using 10 μg of randomly fragmented human genomic DNA and a large sequencing depth (35.7M PE reads); whereas, PF2-Standard started with a third of the fragments (3 μg) and a lower sequencing depth (11.7 M PE reads). Both libraries used the same random barcode of 10 nucleotides on each side of the PE read (10 + 10). Thus, a larger subset of tags was sequenced in the PF1-Standard library. This is also reflected in the TD: in the PF1-Standard library tags differed by less nucleotides (3–5) compared to the other libraries with a TD = 5–7 (compare Supplementary Figures S3A with S4A).

Interestingly, with our TD tool we also observed in all libraries an unexpected number of tags (30–48%) that differed only by one nucleotide (TD = 1). In general, for a bi-modal tag distribution like this, a TD = 1 could be indicative of sequencing or PCR errors in the tags; although, in very large libraries these could be also the result of collision events. In fact, by closer examination of the tag data with our TD tool, we can see that the TD distributions of PF2-Standard had less families with a TD = 1 (34%) compared to PF1-Standard with 48% (Figure 2A and Supplementary Figure S4A, respectively). This difference could be explained by more barcode collisions in PF1-Standard, for which more tags were sequenced. The TD analysis using only tags forming DCS, which should be error-free, provided further information on this aspect and helped distinguish PCR/sequencing mistakes, which are filtered out in DCS from collision events that are still present in the DCS. Figure 2B, Supplementary Figures S3B and S4B show that DCS had almost no tags with a TD = 1 (∼1–2%) in the PF2-CRISPR and PF2-Standard library; whereas, this frequency is increased to 3% in the PF1-Standard library (Supplementary Figure S4B) and are likely collision events.

The shorter the tag or the larger the number of starting fragments, the higher the frequency of tags with one or two differences that are likely barcode collisions that cannot be distinguished from sequencing or PCR mistakes. This is illustrated in Figure 3, in which we computationally shortened the tag from 10 + 10, to 8 + 8 and 6 + 6 nucleotides. As tags get shorter, the proportion of tags with a TD = 4–7 decreases and the proportion of tags with small distances increases (also see Table 2). In short tags (e.g. 6 + 6), errors cannot not be distinguished anymore from real differences between uniquely labeled molecules. Note that a single peak with a very low TD (1–2) observed in a dataset, might not only be the result of too many input fragments to possible barcode combinations (barcode collisions), but could also be the outcome of the sequencing depth being too high leading to errors in every tag. Regardless, an outcome with the majority of the data having a low TD (1–2) means that variant frequencies are unreliable, because there is not a ‘one to one’ mapping between tags and molecules.

Knowing the TD distribution is also quite useful in case a barcode correction is implemented. In barcode correction, tags with sequencing or PCR errors with 1–2 nt differences are re-united to the original family (15). However, before using this tool it is important to know if a low TD is due to errors or mismatches that can be corrected or to barcode collisions, for which this correction would wrongly merge families from different molecules. Here, we showcase how the information of the TD tool helps implementing the barcode correction tool coming back in particular to our example of the PF2-CRISPR and PF2- and PF1-Standard datasets. Given that more than half of the tags have a TD of 5–7 in PF2-Standard, it is appropriate to implement the barcode correction tool (15) allowing for 1–2 mismatches in the tags (one per barcode). In the larger library (PF1-Standard) with 3–5 differences, a barcode correction of one mismatch could already be problematic, since a barcode with one mismatch could translate into two mismatches per tag (Supplementary Figure S4C and D).

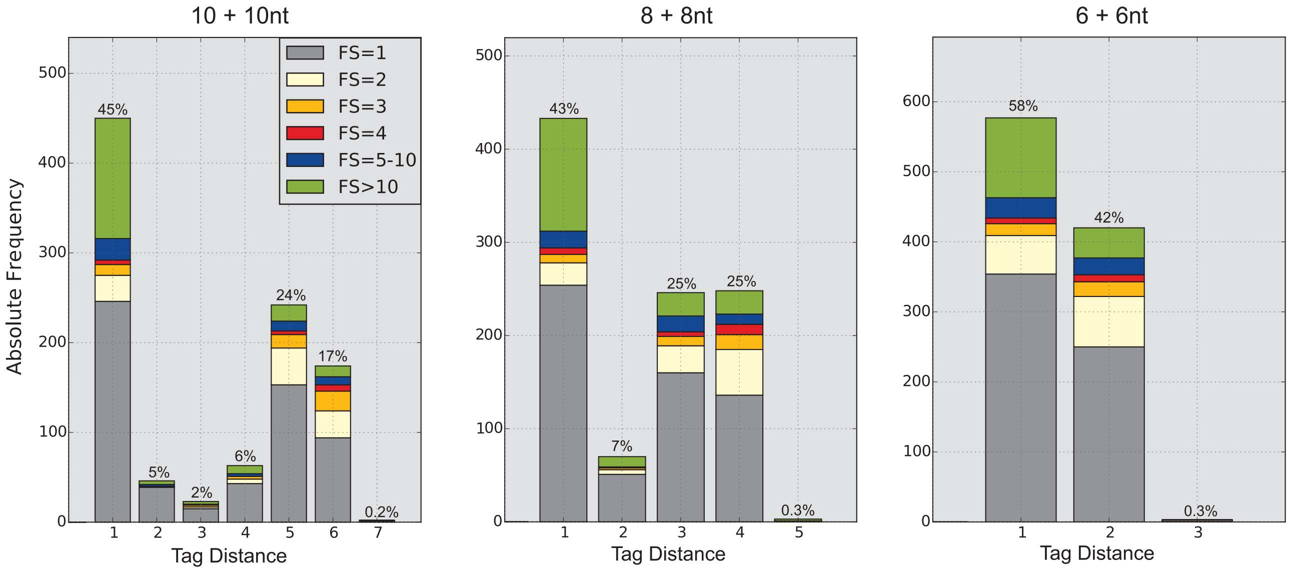

Figure 3. Effect of computationally shortening the length of the original tag in PF2-CRISPR. The original tag (10 + 10) is shortened to 8 + 8 and 6 + 6 nucleotides after splitting the full tag (20 nucleotides) in half and removing 1 and 2 nucleotides at both ends of the half.

Table 2. Proportion of tags with the given TD when computationally shortening the length of the original tag (10 + 10 nt) obtained in the PF2-CRISPR library

tag length (nt)	TD = 1	TD = 2–3	TD = 4–7	Total tags
20 (10 + 10)	45%	6.9%	48.1%	188 354
16 (8 + 8)	43.3%	31.6%	25.1%	
12 (6 + 6)	57.7%	42.3%	0%	

The number of total tags represents all the different families/tags in the dataset.

Allowing mismatches in the barcodes/tags has the advantage that the proportion of families that are partnered into a DCS increases (e.g. for the PF2-CRISPR library the percentage of DCS rises from 23 to 32%) considering a family size of FS $\geq$ 1 (Supplementary Table S3). Note that singleton reads (with no family) are mainly reallocated to larger family sizes (mainly FS = 5–10 and FS > 10). In addition, when using a family size of FS $\geq$ 3, the barcode correction does not influence the proportion of DCS to SSCS as strongly. An extensive description of the error barcode correction tool can be found in (15) and will not be addressed here further.

Chimera analysis (CA)

Chimeric reads are a known problem in PCR-based methods when a primer is not completely extended during PCR and this partial extension acts as a primer on similar templates, or by template switching in which an extended strand jumps from the original template to another one (22,23). During PCR, chimeras form at a frequency of 0.2 or 20% (if the plateau phase is reached in PCR), regardless of the proofreading activity or processivity of the PCR polymerase (24). The formation of chimeras could be a serious problem in DS, because it confounds copies derived from the same template as two independent copies, and potentially leads to a false estimate of variant frequencies.

Our *CA* helps to address this issue. This analysis is based on the examination of differences between the barcodes at both ends of a read (*a*/*b*). Chimeras can be identified by carrying the same barcode at one end combined with multiple different barcodes at the other end of a read (Figure 1). Ideally, given a large excess of random barcodes to input templates, barcode collisions, which occur when two identical barcodes label more than one template, are very unlikely; although, this might quickly change once more templates are tagged and sequenced in the library (see example of PF1-Standard library in the previous section).

Specifically, CA compares tags that are identical at one end and different at the other end of the read. This is done by splitting the tag into the individual barcodes (*a* and *b*) in a subset of 1000 tags (each barcode representing the upstream/downstream end). Tags with the smallest distance in the *a* barcode (TDa_{min}) and the largest distance in *b* (TDb_{max}) in the dataset (and *vice versa*: smallest distance in *b* and largest distance in *a*) are extracted. The tool then analyzes the contribution of each part (*a* + *b*) to the overall TD. In a chimera, it is expected that only one barcode of the tag contributes to the TD of the whole tag. In other words, if the same *a* barcode is observed in combination with several different *b* barcodes (as would be expected in a chimera), then one barcode will have a TD = 0. Thus, the TD difference between the two barcodes (TDa_{min}- TDb_{max}) is the same as the sum of the parts (TDa_{min}+ TDb_{max}) or the ratio of the difference to the sum (*relative delta TD* = TDa_{min}-TDb_{max}/TDa_{min}+ TDb_{max}) will equal to one in chimeric families. Note, that a barcode collision event would have the same relative delta TD of one. A high chimera rate thus may indicate a high rate of barcode collisions and should lead to a re-assessment of the amount of input DNA used in the library or the use of longer barcodes. For experiments, that use random fragmentation, looking at the whole read instead of only the tag for chimera detection would better distinguish chimeras from collision events. However, this approach would not help when using targeted fragmentation (such as CRISPR) since all the targets have the same ends and would increase the computational burden significantly.

In the PF2-CRISPR, PF2-Standard, and PF1-CRISPR libraries ∼44–51% of the tags after barcode correction are formed by chimeric families with a *relative delta TD* value of one (Figure 4A and Supplementary Figure S5A), respectively. Interestingly, the majority (80–86%) of these chimeras just form SSCS, but not DCS (Figure 4B and Supplementary Figure S5B). When analyzing the tags of DCS, the percentage drops to 8% or 4% chimeric DCS, respectively (Figure 4C, Supplementary Figure S5C and Table S4). Note that the DCS chimeras are mainly coming from very large families with more than 10 members (Figure 4D and Supplementary Figure S5D). Interestingly, for the PF1-Standard library (the library with the largest number of sequenced tags), we observed a much larger percentage of chimeric families (97% of all families) and 29% of chimeric DCS (Supplementary Table S4). In the case of the PF1-Standard library, the large number of sequenced tags increased the probability of using the same barcodes resulting in more collision events. This illustrates, how our CA tool helps to assess library preparation quality and differences.

Why is this CA also important? First, it could be an indication of barcode collisions as just discussed. Second, it could be an indicator of data loss with chimeric SSCS that cannot be grouped into DCS given that most chimeric reads form only SSCS and do not have a duplex partner (Supplementary Table S4). However, alternatives are also possible, such as unpaired SSCS being the result of an amplification bias, in which only the forward or the reverse strand gets amplified, as is the case in ∼20–50% of the templates in PCR (see this phenomenon in (8)). Eliminating chimeras could be an important modification in the library preparation protocols to improve DCS coverage and rare variant accuracy. This could be achieved by avoiding amplifying mixtures of very similar DNA, but instead using a single molecule format using for example a water/oil emulsion as described previously (25).

Third, and more importantly, chimeras could influence the accuracy of the variant frequency, in case the same molecule is counted more than once. In order to assess this potential overestimation, we added a ‘chimera count’ to the VAR-A tool (column E and K in sheet ‘Allele frequencies’, Supplementary Tables S5–S8). To distinguish between a variant being a chimera or a collision event, we searched for chimeric tags within the families that share the same variant, since it is very unlikely to see a collision in reads that also carry the same rare variant. In both PF2-CRISPR and PF2-Standard libraries, we detected 37 and 5 chimeras out of 6834 and 7643 variant calls, respectively, for alleles with high absolute counts such as heterozygote variants (e.g. in PF2-CRISPR we observed 14 chimeras for chr17:7674089-A-C; 14 chimeras in chr17:7674109-G-A and 9 chimeras in chr17:7674797-T-C). These chimeras hardly had an effect on the overall variant frequency. We did not detect chimeras in any of the low frequency variants. However, if the chimeric frequency is overall high, the chance of a rare allele being a chimera increases introducing an inflated estimate of the variant frequency.

Family size distribution (FSD)

One of the first steps in the consensus building in DS is grouping together PE reads representing copies of an initial molecule into a family or consensus (in this case SSCS). The more reads within a family carrying the substitution, the more likely the substitution is real and not an artefact. There is a delicate balance around the optimal family size: small families make variant calling less reliable, while larger families reduce the DCS coverage and total yields. Our FSD tool analyzes the family size associated with each tag and renders a graphical and tabular output of the absolute and relative family sizes compared to the total amount of families and total amount of PE reads. This tool can be used to compare the FSD among different libraries or different steps of the bioinformatic pipeline (e.g. barcode correction or sequence trimming).

This tool also analyzes the ratio of SSCS/DCS for each family size. This latter analysis (Figure 5) is particularly useful to decide on the minimal number of PE reads to build a consensus sequence (SSCS). In the early days of DS, an average family size of six to seven members was considered appropriate for reliable variant calling (4), but with time this number has been reduced to three members to increase the data yields and DCS coverage (11). The decision as to the minimal number of PE reads to form a consensus depends on each library, thus having a tool to visualize the FSD is quite useful. In the PF2-CRISPR and PF2-Standard libraries, 32 or 23% of the families are united into DCS (68 or 33% of the total PE reads), respectively, and more than half of the DCS are formed by family sizes between 3 and 20 members (Table 3 and Figure 5). Thus, for these libraries using a family size with a minimal number of three reads is recommended (11); although, with this setting 5–12% of the DCS, formed by smaller family sizes of 1–2 reads, get lost.

Knowing the ratio of SSCS/DCS for each family size also helps to understand the data allocation and DCS yields: for example, Nachmanson and colleagues reported that CRISPR-DS is superior over standard-DS, given the 10-fold higher recovery rate of specific target regions (average DCS coverage per input templates) of the former (11). Our FSD-tool provides a different light on the performance of these two DS protocols: in terms of DCS recovery and allocation of DCS within optimal family sizes (3–20 reads), only one of the four libraries (PF1-Standard) performed suboptimal (Figure 5 and Table 3). In this library, ∼83% of the families (or ∼94% of the total PE reads) had more than 20 reads, which explains the lower DCS coverage of this library. However, the other standard DS library (PF2-Standard), which used a third less of DNA, performed equally well compared to the two CRISPR libraries: all three libraries had 56–62% DCS formed by optimally sized families (3–20). With our tool, it can also be observed that PF1-CRISPR and PF2-Standard formed DCS with small family sizes (FS = 1–2), which would get lost if at least three reads are required in a family. Understanding what factors influence the formation of sub-optimal family sizes for DCS (very small or very large families) during library prepara-

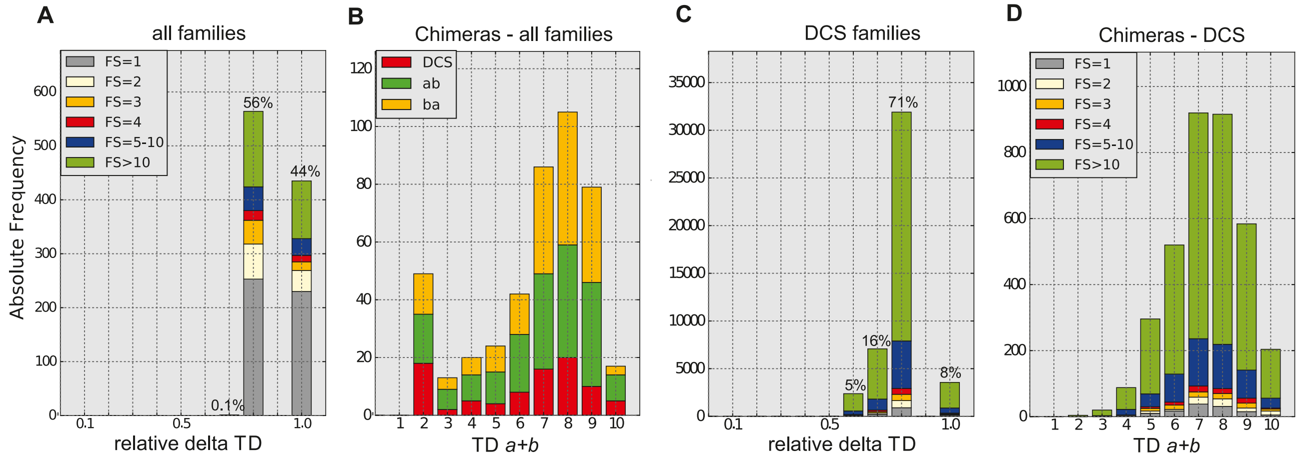

Figure 4. (**A**) CA of a subset of tags (n = 1000) derived from the SSCS of the PF2-CRISPR library after barcode correction (allowing one mismatch). Chimeras have a *relative delta TD* = 1. (**B**) Distribution of TD in chimeric tags (with a *relative delta TD* = 1) stratified into DCS or SSCS. Out of 435 chimeric reads, 181 (42%) belong to SSCS ab; 166 (38%) to SSCS ba and 88 (20%) to DCS. (**C**) CA using the tags of DCS and (**D**) TD of the chimeric DCS. Note that most of the chimeric DCS come from large family sizes (>10). The absolute numbers presented in Figure 4B and Supplementary S5B are listed in Supplementary Table S4. We also added a column in the VAR-A listing the chimeric tag, in case the variant is a chimera (Supplementary Table S5, column AG).

tion is an intensive research focus and our FSD-tool supports a detailed evaluation.

Variant analyzer (VAR-A)

Currently in DS, a family or consensus (SSCS) is built with a minimum of three or more reads (FS≥3). Yet, 5–12% of the DCS in the four tested libraries were formed by families with 1–2 reads. These DCS are discarded; although, they might contain important variant information and their inclusion would increase the coverage, but small families bear a higher risk of false positive calls. We developed the VAR-A to assess the evidence supporting a variant call based on a series of different summary data extracted from the raw PE reads that classifies the confidence level of a variant call by a tier-based system. This allows using more relaxed analysis parameters during consensus building, e.g. small families (including families with only one or two reads) or *ad hoc* stringent trimming parameters.

Only DCS with a variant in the original consensus building output are re-analyzed in VAR-A. The evidence of a variant is then compiled from the raw PE reads, and includes information on the mate (if both mates overlap the position of the variant), the number of high quality PE reads in both forward (*ab*) or reverse (*ba*) SSCS, the proportion of alternate versus reference calls within the family and the median position of the variant within the consensus sequence. The quality of the variant call is re-analyzed, PE reads are trimmed automatically, and positions with low quality (by default PHRED < 20) are removed. Given these additional analysis layers, more data can be used for consensus calling without compromising the reliability of the analysis and false negatives can get potentially recovered. The tabular output of VAR-A also includes relevant information such as the sequence of the tag, if more than one variant is present in a family (multiple consecutive variants in one molecule—in phase), and if the variant is part of a chimeric family.

VAR-A also categorizes variants with a tier-based system (see Table 1 and Supplementary Table S1) that helps the user to distinguish high quality calls from those with lower support. Tier 1 variants have the strongest support with information from multiple reads and mates. However, the inclusion of second-order tiers (1.2–2.4) increases the coverage without seriously compromising the accuracy or reliability of the call that can be removed, if necessary, after manual inspection. These second-order tiers are particularly interesting, because they include small families (1–2 reads) for either the forward or the reverse SSCS, which would be discarded by the regular pipeline. Yet, these variants are likely real since they are present in both forward and reverse SSCS, albeit in one of them at a low number. Table 4 compares the variants identified using different settings for the minimum family size (FS ≥ 3 or FS ≥ 1). When reducing the minimum family size from three to one read, we rescued ∼2200 or ∼25 500 DCS resulting in an ∼10% or 25% increase of DCS coverage in the PF2-CRISPR or PF2-Standard library, respectively (see Table 4).

Since small families make variant calling less reliable, we analyzed the variants with VAR-A using standard (FS ≥ 3) or small minimal family sizes (FS ≥ 1) and compared them with variants identified with the bioinformatics pipeline from University of Washington (11). The full output of the VAR-A analysis for the PF2-CRISPR, PF2-Standard, PF1-CRISPR and PF1-Standard dataset is included as Supplementary Tables S5–S8, respectively.

As expected, heterozygous positions (50%) shown in Table 5 (PF2-CRISPR) or Supplementary Table S9 (PF2-Standard) are detected by all three analyses at very similar frequencies while the coverage in the VAR-A analysis with FS ≥ 1 is higher, demonstrating that reducing FS ≥ 1 renders the same reliable calls as the other two more conservative pipelines, but at a higher coverage. While the increase in coverage might not seem extremely important in this example, it is especially an advantage in low-input samples,

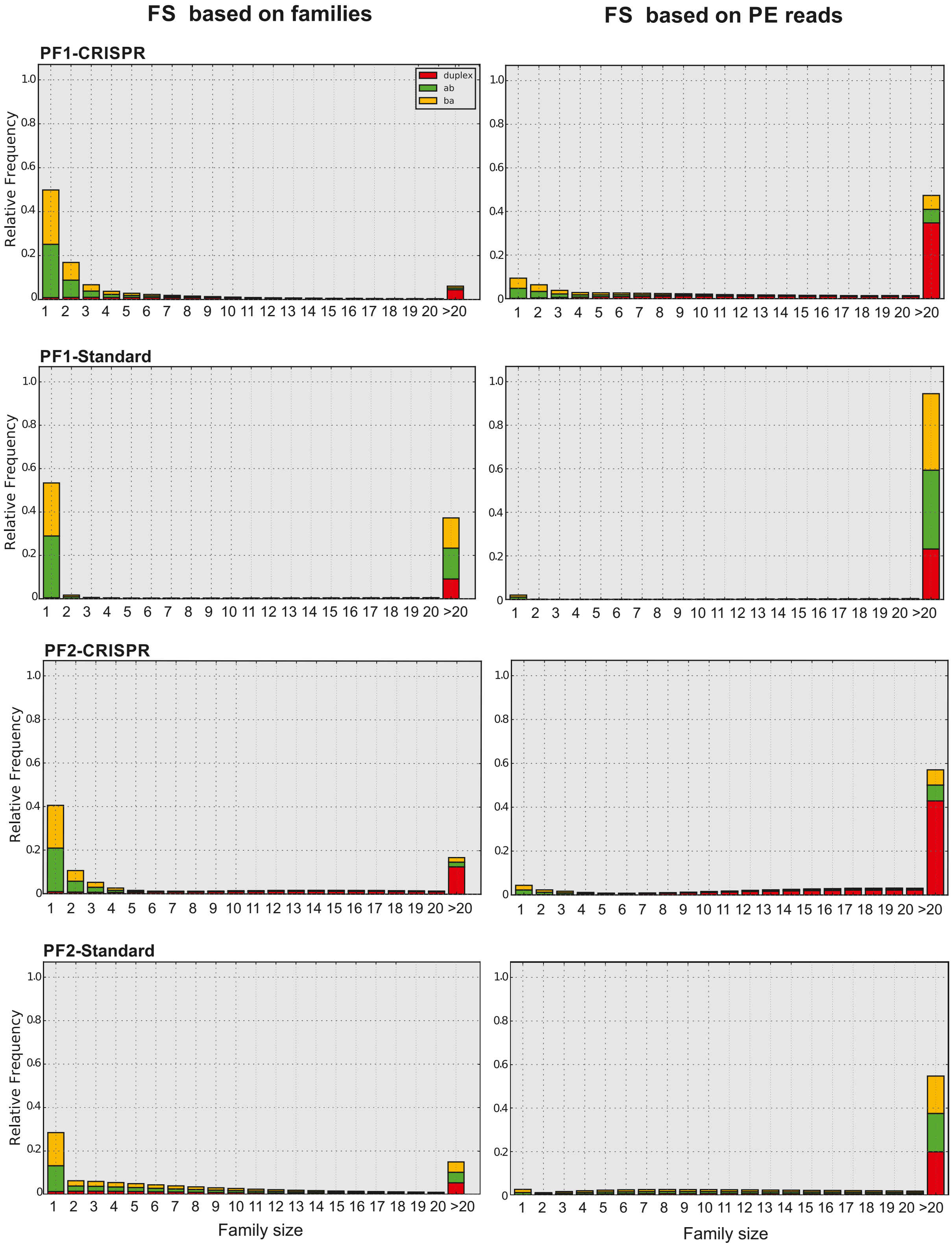

Figure 5. Family Size (FS) distribution in the different libraries. Frequency of reads per family observed in the four analyzed libraries (PF1-CRISPR, PF1-Standard, PF2-CRISPR and PF2-Standard after barcode correction with one mismatch relative to the total number of families or to the total number of PE reads. FS is stratified into SSCS (*ab* or *ba*) or DCS (*duplex*).

Table 3. Fraction of data forming DCS from the total number of families or PE reads obtained in each library (column 1 and 2, respectively) after barcode correction with one mismatch

	DCS families	DCS PE reads	DCS FS1–2	DCS FS3–20	DCS FS>20	mean FS	total families (DCS+SSCS)	Total PE reads
PF1-CRISPR	17% 31 710	53% 538 889	11% 3480	62% 19 787	27% 8443	5	189 675	1 004 941
PF1-Standard	11% 147 944	24% 8 521 521	5% 7281	12% 18 010	83% 122 653	27	1 341 763	35 748 407
PF2-CRISPR	32% 44 912	68% 867 430	5% 2312	56% 25 160	39% 17 440	9	138 631	1 284 504
PF2-Standard	23% 253 428	33% 3 874 141	12% 31 187	64% 162 893	23% 59 348	11	1 106 303	11 717 544

We counted 31 710, 147 944, 44 912 and 253 428 families that can form a DCS for PF1-CRISPR, PF1-Standard, PF2-CRISPR and PF2-Standard, respectively. Column three–five represent the proportion of DCS with a family size (FS) of 1–2 reads, 3–20 reads or more than 20 reads, respectively. Column six reports the average family size (FS) and column seven shows the total number of families counted including SSCS and DCS.

Table 4. shows the coverage of the Du Novo analysis from the PF2-CRISPR and PF2-Standard library (after barcode correction with one mismatch) using different parameters for consensus building for the minimum family size (FS)

PF2-CRISPR	FS ≥ 3	FS ≥ 1	PF2-Standard	**FS ≥ 3**	**FS ≥ 1**
Total reads	1 284 504		**Total reads**	11 717 544	
Total SSCS	138 631		**Total SSCS**	1 106 303	
SSCS FS 1–2	116 175		**SSCS FS 1–2**	979 589	
DCS	20 218	22 456	**DCS**	101 145	126 714
On target	20 217	22 426	**On target**	101 095	125 826

Reducing the FS settings to FS ≥ 1 increased the number of DCS.

where an increase in sequencing depth for variant calling is crucial.

With VAR-A, we also identified important biases or false positives in the DCS of a library. For example, note that in the Standard library (Supplementary Tables S6 and S9) some variants occur with a low tier (mainly 4.1–5). Some of these variants are of high sequence quality and form large family sizes, but co-occur at the end or beginning of the DCS, either as single independent events or as multiple variants within the same family (e.g. chr17:7676340-G-C and chr17:7676341-T-C). This strong positional bias of variants (mainly at the beginning of the DCS, representing the 5′ or 3′ end of the original DNA molecule) makes it highly likely that these are the result of end-polishing during A-tailing before adapter ligation. Variants at the beginning or the end of DCS are labeled by VAR-A as tier 4.1 and can be manually discarded. Another example is variant chr17:7675393-C-T that occurs close to a poly-T homopolymer (chr17:7675394-7675411) reported to be associated with noisy reads (11) and that was tagged by VAR-A as a low tier that can be manually inspected and discarded.

When building consensus sequences with a FS ≥ 1, VAR-A identified two potentially new variants that were missed by the analysis using a FS ≥ 3 (shown in blue in Table 5 and Supplementary Table S9). These occur at an ultra-low frequency (∼10^{-4}), which underlines the power of VAR-A in terms of detecting very rare variants. The identified variant (chr17:7669456-C-A) in the PF2-CRISPR library classified as tier 3.1 is formed by SSCS with 3 and 2 members; in one SSCS, two out of three reads carried the alternative allele and in the other SSCS both reads carried the variant. This call could be considered a borderline case and would require further validation in subsequent libraries with higher coverage. We did not identify this variant in PF2-Standard (DS of the same biological sample) likely because of insufficient coverage in this library at this position. Similarly, in the PF2-Standard library, variant chr17:7675381-T-C was counted twice (with tier 2.1 and 2.4) totaling to an AF of 0.02%. Further inspection of this variant showed that it is formed by high quality PE reads including tier 2.1, 2.4, 4.1 and 5 (the latter two representing the end of a DCS and remaining low quality variants) shown in Supplementary Table S6. This variant also could be a borderline case since it occurs close to the poly-T homopolymer (chr17:7675394–7675411). Note that most of the variants in these positions were filtered out during the different QC steps of VAR-A (tier 5), as were those calls close to another poly-T homopolymer (chr17:7674361–7674376).

We also obtained the same variant allele frequency (1.0 or 1.2% for the PF2-CRISPR or PF2-Std, respectively) for variant chr17:7674230-C-T with the relaxed pipeline as reported by Nachmanson and colleagues (labeled as chr17:7577548-C-T with GRCh37/hg19 as a reference) (11). More interestingly, performing the analysis with FS ≥ 1 increased the coverage of this low frequency variant from 4829 to 5202 DCS or 2003 to 3463 DCS in the PF2-CRISPR or PF2-Standard library, respectively. This example highlights the value of VAR-A to accurately call low variant frequencies at a quite improved coverage.

Moreover, we categorized the variants of two more libraries with the VAR-A: PF1-CRISPR and PF1-Standard (Supplementary Tables S10 and S11). Here again, we obtained very similar variant frequencies with all analyses including the pipeline of the University of Washington (11). Interestingly, for all pipelines we observed in PF1 unusually high allele frequencies (84% for two alleles (chr17:7674089-

A-C and chr17:7674109-G-A). Given the highly similar variant frequency of these two sites, it is likely that these are linked. With our VAR-A tool we could further investigate this, since the VAR-A also provides information if a variant co-occurs with another variant and in what phase in the same DCS. We observed that in the majority of the DCS, the same alleles co-occurred within the same molecule (e.g. 7674089-A and 7674109-G), explaining the highly similar variant frequency of ∼84%. VAR-A further lists the co-occurring variants in column AF and could be used to get a 'haplotype count'. The phase information of different variants is a powerful tool to interrogate if variants could be evolving by clonal expansion within a tumor.

In addition to our analysis of published libraries, we also performed as a proof of principle an orthogonal assay experiment. In this assay, we mixed genomic DNA carrying variants c.742C>T (NA00711), c.746C>G (NA08909), c.749C>G (CD00002) and c.1620C>A (p.540N>K) with wild type DNA at ratios from 1/10 to 1/10,000 in steps of one order of magnitude followed by library preparation as described in (Salazar *et al.*, in preparation). Our analysis shows that our VAR-A is highly reliable and the measured variant fraction correlates with the expected dilutions by a factor R^2 of 0.96 (Supplementary Table S12). Deviations from the expected initial input amounts were due to experimental errors. Also, at the low end we identified two out of four ultra-rare variants diluted to one in 10 000, but given the Poisson distribution of single events it is expected to have missed these rare variants given our achieved coverage. More importantly, with this dataset we clearly illustrate that by relaxing the initial Du Novo analysis parameters followed by our VAR-A, we rescued a higher number of counts with the alternate allele without compromising the accuracy or validity of the call. For the detection of rare-allele this is quite advantageous, since more single counts within one experiment improves the confidence that these variants are real.

We conclude that VAR-A is a powerful tool to increase the coverage without compromising the reliability of the variant calling, which is especially advantageous in low-input samples or low-coverage regions, where an increase in sequencing depth for variant calling is of particular importance.

Table 5. Summary of the variants identified in the PF2-CRISPR dataset

	UW			Du Novo FS ≥ 3			Du Novo FS ≥ 1			Tiers VAR-A with Du Novo FS ≥ 1							
Variant ID	cvrg	mut	Variant allele freq. (%)	cvrg	mut	Variant allele freq. (%)	cvrg	mut	Variant allele freq. (%)	1.1	1.2–2.4	3.1	3.2	4.1	4.2	5	Variant allele freq. in PF2-Std
chr17:7669456-C-A							3983	1	0.03%	0	0	0	1	0	0	0	
chr17:7673663-C-A				2909	1	0.03%	3260	1	0.03%	1	0	0	0	0	0	0	
chr17:7674089-A-C	4696	2318	49.4%	4844	2390	49.3%	5217	2571	49.28%	2385	183	3	0	0	0	0	49.40%
chr17:7674109-G-A	4703	2324	49.4%	4848	2394	49.4%	5224	2575	49.29%	2393	182	0	0	0	0	0	48.30%
chr17:7674230-C-T	4703	48	1.0%	4829	50	1.04%	5202	52	1.00%	50	2	0	0	0	0	0	1.24%
chr17:7674797-T-C	2868	1423	49.6%	2980	1468	49.3%	3292	1631	49.54%	1483	147	1	0	0	0	0	49.60%
chr17:7674926-C-T	2883	1	0.035%	3000	1	0.03%	3310	1	0.03%	1	0	0	0	0	0	0	
chr17:7676111-T-A				3056	1	0.03%	3374	1	0.03%	1	0	0	0	0	0	0	
chr17:7676325-C-Del	2956	1532	51.8%														
chr17:7676705-C-A	2193	1	0.046%	2258	1	0.04%	2571	1	0.04%	1	0	0	0	0	0	0	

We compared the original pipeline of University of Washington (UW) with Du Novo analysis using FS ≥ 3 or FS ≥ 1, the latter combined with VAR-A. The variants identified in VAR-A are classified by the different tiers representing high or lower confidence calls. Details of variants are shown in Table S5. The variant ID is based on the human genome assembly GRCh38/hg38. Variants in blue are only detected by the Du Novo run with relaxed settings such as a family size (FS ≥ 1). Variants in green were detected only by the pipeline of UW. Note that this is due to the fact that VAR-A currently does not support the evaluation of in-dels.

CONCLUSION

Our QC tools are important analysis tools for investigating DS data and can be implemented to identify problems at the experimental level or during the bioinformatic assembly. With our tools we show the number of reads that are part of very large families or are singleton reads (without a family) and contribute to data loss and lower DCS coverage and yields. We also show that chimeras are mainly formed by unpaired SSCS and likely represent collision events. Finally, we demonstrate that reads with small families can be included in the consensus calling since the quality and reliability is validated with our VAR-A tool. The resulting increase in coverage is an advantage especially in low-input samples, where an increase in sequencing depth for variant calling is crucial and allows for an identification of ultra-low frequency variants without blindly increasing the risk of false positive calls.

DATA AVAILABILITY

The tools are written in Python, are open source and readily available through the user-friendly Galaxy platform that can be easily run without any programming experience https://usegalaxy.org/, and GitHub: https://github.com/Single-Molecule-Genetics. All software is freely available under non-restrictive AFL 2.0 license.

The data for the PF1-CRISPR, PF1-Standard, PF2-CRISPR and PF2-Standard library was downloaded from https://www.ncbi.nlm.nih.gov/bioproject/?term = PRJNA412416 on January 23, 2019. The data for the libraries spiked with known mutations can be found in the NCBI Sequence Read Archive (BioProject ID: PRJNA684907).

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We would like to thank Rosana Risques and Jesse Salk for providing the p53 dataset, Barbara Arbeithuber for advice on DS and fruitful comments on the manuscript, and Ulrich Bodenhofer for help in the conceptualization stages of the tools.

Authors' contributions: G.P. and M.H. developed the software and performed the analyses, R.S. participated in the analysis and tested software components and performed the spike in libraries. N.S. improved and trouble-shooted software components. I.T.B. conceived the project and I.T.B. and A.N. provided funding, G.P., M.H., R.S. and I.T.B. wrote the manuscript.

FUNDING

Linz Institute of Technology [LIT213201001 to R.S., M.H., G.P., I.T-B]; Austrian Science Fund [FWFP30867000 to R.S., M.H., G.P., I.T-B].

Conflict of interest statement. None declared.

REFERENCES

1. Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2018) Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat. Rev. Genet.*, **19**, 269–285.
2. Schmitt,M.W., Fox,E.J., Prindle,M.J., Reid-Bayliss,K.S., True,L.D., Radich,J.P. and Loeb,L.A. (2015)Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat. Methods*, **12**, 423–425.
3. Jabara,C.B., Jones,C.D., Roach,J., Anderson,J.A. and Swanstrom,R. (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 20166–20171.
4. Schmitt,M.W., Kennedy,S.R., Salk,J.J., Fox,E.J., Hiatt,J.B. and Loeb,L.A. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **109**, 14508–14513.
5. Hiatt,J.B., Pritchard,C.C., Salipante,S.J., O'Roak,B.J. and Shendure,J. (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.*, **23**, 843–854.
6. O'Roak,B.J., Vives,L., Fu,W., Egertson,J.D., Stanaway,I.B., Phelps,I.G., Carvill,G., Kumar,A., Lee,C., Ankenman,K. *et al.* (2012) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science*, **338**, 1619–1622.
7. Lou,D.I., Hussmann,J.A., McBee,R.M., Acevedo,A., Andino,R., Press,W.H. and Sawyer,S.L. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl Acad. Sci. U.S.A.*, **110**, 19872–19877.
8. Arbeithuber,B., Makova,K.D. and Tiemann-Boege,I. (2016) Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res.*, **23**, 547–559.
9. Ahn,E.H., Hirohata,K., Kohrn,B.F., Fox,E.J., Chang,C.C. and Loeb,L.A. (2015) Detection of ultra-rare mitochondrial mutations in breast stem cells by duplex sequencing. *PLoS One*, **10**, e0136216.
10. Kennedy,S.R., Salk,J.J., Schmitt,M.W. and Loeb,L.A. (2013) Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLos Genet.*, **9**, e1003794.
11. Nachmanson,D., Lian,S., Schmidt,E.K., Hipp,M.J., Baker,K.T., Zhang,Y., Tretiakova,M., Loubet-Senear,K., Kohrn,B.F., Salk,J.J. *et al.* (2018) Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). *Genome Res.*, **28**, 1589–1599.
12. Stoler,N., Arbeithuber,B., Guiblet,W., Makova,K.D. and Nekrutenko,A. (2016) Streamlined analysis of duplex sequencing data with Du Novo. *Genome Biol.*, **17**, 180.
13. Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.
14. Orabi,B., Erhan,E., McConeghy,B., Volik,S.V., Le Bihan,S., Bell,R., Collins,C.C., Chauve,C. and Hach,F. (2019) Alignment-free clustering of UMI tagged DNA molecules. *Bioinformatics*, **35**, 1829–1836.
15. Stoler,N., Arbeithuber,B., Povysil,G., Heinzl,M., Salazar,R., Makova,K.D., Tiemann-Boege,I. and Nekrutenko,A. (2020) Family reunion via error correction: an efficient analysis of duplex sequencing data. *BMC Bioinformatics*, **21**, 96.
16. Blankenberg,D., Gordon,A., Von Kuster,G., Coraor,N., Taylor,J., Nekrutenko,A. and Galaxy,T. (2010)Manipulation of FASTQ data with Galaxy. *Bioinformatics*, **26**, 1783–1785.
17. Li,H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv doi: https://arxiv.org/abs/1303.3997?upload=1, 26 May 2013, preprint: not peer reviewed.
18. Garrison,E. and Marth,G. (2012) Haplotype-based variant detection from short-read sequencing. arXiv doi: https://arxiv.org/abs/1207.3907v2, 20 July 2012, preprint: not peer reviewed.
19. Garrison,E. (2016) Vcflib, a simple C++ library for parsing and manipulating VCF files. https://github.com/vcflib/vcflib.
20. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
21. Hamming,R.W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.*, **29**, 147–160.
22. Kanagawa,T. (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.*, **96**, 317–323.
23. Odelberg,S.J., Weiss,R.B., Hata,A. and White,R. (1995) Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Res.*, **23**, 2049–2057.
24. Boulanger,J., Muresan,L. and Tiemann-Boege,I. (2012) Massively parallel haplotyping on microscopic beads for the high-throughput phase analysis of single molecules. *PLoS One*, **7**, e36064.
25. Palzenberger,E., Reinhardt,R., Muresan,L., Palaoro,B. and Tiemann-Boege,I. (2017) Discovery of rare haplotypes by typing millions of single-molecules with bead emulsion haplotyping (BEH). *Methods Mol. Biol.*, **1551**, 273–305.