For deletion discovery, we ran the discovery part of MATE-CLEVER [3], with minor modifications that account for volatilities among library protocols. MATE-CLEVER is an integrated approach. Its major purpose in the frame of the project is to discover deletions of size 30–100 bp (sometimes termed the “twilight zone of NGS indels”). It incorporates CLEVER [2], as an internal segment size based approach that approvedly has state-of-the-art performance rates on indels of size 30–100 bp, and LASER [4], as a split-read aligner. MATE-CLEVER uses CLEVER in a first step, in order to discover deletions of size 30–100 bp at extremely high sensitivity, and uses LASER in a second step, in order to refine the breakpoint annotations made by CLEVER. It also uses several auxiliary tools, as described below. In the following, we provide the full description of details and commands, by which to reproduce MATE-CLEVER’s callset. This pipeline deviates in several details from the description provided in [3] itself, due to the minor modifications mentioned above.

To run the CLEVER-based deletion discovery pipeline, revision 30972 from the git repository at http://clever-sv.googlecode.com has been used. As above-mentioned, it includes CLEVER [2], LASER [4], and several auxiliary tools described below. To run (and parallelize) the below pipeline, we used the Python-based workflow engine Snakemake [1].

1. CLEVER was run on all individuals separately. In detail, that means the following steps were performed.

(a) We used tool `bwabam-to-alignment-priors` to extract prior alignment probabilities. The output was split by chromosome. Commandline:

\[
\text{bwabam-to-alignment-priors -m mean-and-sd.txt ref.fasta input.bam}\\n\text{\quad | split-priors-by-chromosome -z -s input.bam priors}
\]

For each chromosome and each sample in the input, one gzipped file prefixed with `priors` is created. All alternative alignments provided by BWA through XA tags are used. Furthermore, this step generates robust estimates of insert size mean and standard deviation which are stored in the file `mean-and-sd.txt`.

(b) All files with alignment probabilities are sorted by genomic position by means of a standard unix sort (`sort -g -k7`).

(c) The CLEVER core engine is run on each of these files. Commandline:

\[
\text{zcat input.probabilities.gz \pipe clever -c 150 -v > clever.output}
\]

The option `-c 150` instructs CLEVER to limit local coverage to 150 to avoid long runtimes for regions with excess coverage (due to repeats).

(d) For each individual, all CLEVER output files for the chromosomes are concatenated and postprocessed. Commandline:

```
postprocess-predictions -d5 --only-del concatenated-clever-output.txt <mean>
```

where `<mean>` is the mean insert size as estimated in Step 1a. As per option `--only-del`, only deletions are extracted and processed further. Option `-d 5` is used to only retain deletion predictions supported by significant cliques with at least 5 alignment pairs.
2. LASER was run on all regions with putative deletions as identified by CLEVER in Step 1:
   (a) To generate high-quality background distributions for insert size, insertion size, and deletion size from uniquely mapping reads, LASER was used to realign reads from randomly chosen regions. To this end, 5000 regions of length 10000 were sampled uniformly at random. Reads mapped to these regions by BWA were extracted and remapped using LASER as described in [4]. The following command line was used for the LASER core step:

   ```
   join-split-reads -XIS -A14 --anchor_distance 2000 --max_span 2000
   -L insert-size.dist -R insertion-size.dist -D deletion-size.dist
   ref.fasta input.1.fastq.gz input.2.fastq.gz
   ```

   For all individuals/families, the same set of random regions was used.

   (b) For every family, CLEVER deletion predictions for all individuals are pooled and a set of regions of these deletions plus 500 bp up- and downstream of each deletion is created.

   (c) For all individuals, reads aligned to the selected regions by BWA were extracted. When only one read in a pair was successfully mapped by BWA, the unmapped read was also included. The following parameters were used:

   ```
   join-split-reads -XIS -A14 --anchor_distance 2000 --max_span 50000
   -P putative-indels --snp putative-snps ref.fasta
   input.1.fastq.gz input.2.fastq.gz > output.bam
   ```

   In this mode, LASER writes lists of putative SNPs and indels to the given filenames. Each of these comes with an expected support, i.e. the expected number of reads giving evidence for that particular SNP/indel.

   (d) Evidence from all individuals of the considered family is pooled, i.e., the expected support is added over all family members for all indels reported by LASER. Further filtering is done as follows. We retain all deletion candidates that have a total expected support of at least 0.5, are at least 10 bp long, and agree with a deletion prediction made by CLEVER. Here, we deem two putative deletions ”agreeing” if their center points are at most 100 bp apart and the length difference is at most 20 bp. The resulting set of deletion candidates is used for recalibration of alignments produced by LASER in the next step.

   (e) The scores of all alignment pairs are recalibrated as follows. Phred-scaled insertion and deletion probabilities are set according to the empiric distribution obtained in Step 2a. Deletions in the candidate set generated in Step 2d incur a constant (length-independent) alignment cost of 1. This upweights alignments that support deletions that have also been reported by CLEVER in Step 1 and are thus in line with read pair evidence. For each read pair the alignment pair with the highest probability is reported as primary alignment. All secondary alignments are discarded at this stage.

   (f) The number of LASER (primary) alignments supporting each deletion is extracted from the BAM files for further processing.

3. All deletions of CLEVER and LASER are merged. In this step, the whole GoNL population is considered simultaneously. We iterate over all putative deletions as reported by LASER sorted (decreasingly) by total support of primary alignments in the whole population.

   (a) For each individual, the set of CLEVER predictions is searched for deletions with a center distance of at most 100 and a length difference of at most 20 from the currently processed deletion. All CLEVER calls matching the current deletion are marked and ignored in following iterations. This ensures that each CLEVER deletion is assigned to at most one LASER prediction, giving precedence to predictions common in the population.
(b) If a deletion is supported by CLEVER with support at least 5 and by at least on LASER (split) alignment, we report it to be present in the individual in question.

(c) All deletions present in at least one individual are included in the final VCF file.

References


