getools Enables Ultra-fast and Accurate Quantification of Gene-editing Outcomes with Both Bulk and Single-cell Data

Linjun Wu, **Yongjian Zhang**, Pengfei Yuan

EdiGene Inc., Beijing, China

yjzhang@edigene.com

ASGCT 2022

W-88

Abstract # 962

BACKGROUND

The gene-editing systems as represented by CRISPR/Cas and their derivates enable genome-wide targeted editing in humans and a wide variety of other organisms by either DNA or RNA level modulation. Quantifying the on- and off-target editing outcomes and profiling the mixed complex mutational spectrum *in vitro* and *in vivo*, in bulk or even at the single-cell level, at high accuracy and sensitivity becomes an urgent need for both industry and clinical studies. However, the vast diverse mutations with variable frequency plus multiple-target gene-editing makes systematical enumerating the editome and cataloging rare but relevant co-occurring events and monitoring the single-cell level clonal dynamics very challenging in the real world, especially for the interpretation of the noisy amplicon deep sequencing data. Despite many efforts, the existing pipelines and analytic tools are either defective in processing high-depth multiplexed amplicon data or do not support single-cell data analysis. To address these challenging issues and attempt to standardize the gene-editing data analysis, we have developed getools, a native integrated gene-editing data analysis software toolkit for determining editing efficiency and cataloging outcomes at the 'per-read' level from next-generation sequencing data.

HIGHLIGHTED FEATURES

- *getools* recognizes and splits typically pooled libraries and single-cell libraries by both primer match and sequence alignment, and enables fast and error-tolerant in-place processing of pooled data from thousands of samples each with hundreds of targets;
- *getools* improves in pair-end data merge, using kmers as well as overlap strategy with a precomputed optimal minimal overlap region length for each amplicon, to prohibit false-positive insertion brought by greedy sliding in repeat regions;
- Amplicon data are firstly deduplicated to minimize the memory consumption and downstream computation;
- Optimized Smith-Waterman local alignment algorithm with the customized scoring strategy to make cleavage site aware;
- Comprehensive filters to remove amplification and sequencing artifacts with an unsupervised clustering strategy using the alignment score and variant count, while hard filters could also be applied according to the sequence count and allele frequency threshold;
- *getools* calculates mutation frequency and outcomes for all types of edit events, including snv, insertion, deletion, deletion-insertions (delins), separately or by any combination, in single sample mode or paired with a supplied control sample;
- Implemented as a high-performance standalone software for localized deployment without any data security concerns, and provides industry-level bam, bcf, json, tsv output plus vivid HTML report with processing details and summary metrics for whole cycle data tracing.

ALGORITHMS & IMPROVEMENTS

Overall computation framework of getools

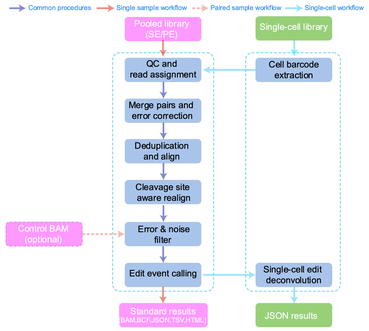

Figure 1. Overall computation framework of *getools*. *getools* support single-end (SE) and paired-end (PE) libraries from the common batch libraries or single-cell libraries. There is a module that supports fast parsing single-cell libraries based on cell barcodes. These cell barcodes will be extracted and kept as extra information for parsing cell-level edit events later.

Versatile split of multiplexed libraries

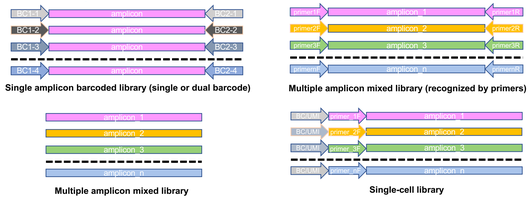

Figure 2. *getools* support various types of multiplexed libraries either by barcodes, primers, or barely amplicon sequences. For the typical barcoded library, either single- or dual-barcode at fixed or nonfixed length could be recognized. The amplicon alignment strategy is complementary to primer match and is critical to dealing with amplification by universal sequences (adapter) or possible disruption in the primer region. *getools* can also process complex libraries such as single-cell libraries.

Improved paired-end reads merge algorithm

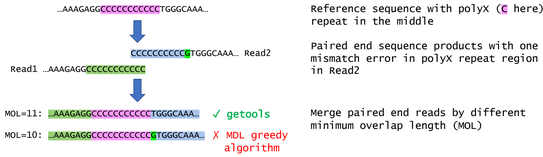

Figure 3. *getools* dynamically precompute the desired minimum overlap length needed for each template and prohibit false positive insertion from greedy overlap by minimum mismatch rate strategy.

Cas-aware cleavage site re-alignment

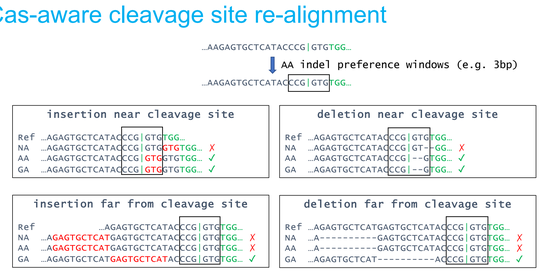

Figure 4. Naïve alignment tools (NA) without any indel preference window limitation might lead to ambiguous indels, advanced alignment tools (AA) with predefined indel preference window might not make all indel detectable, *getools* (GA) realign indels in the whole range of reference to make them closer to cleavage site and make more indel detectable.

Comprehensive event filters

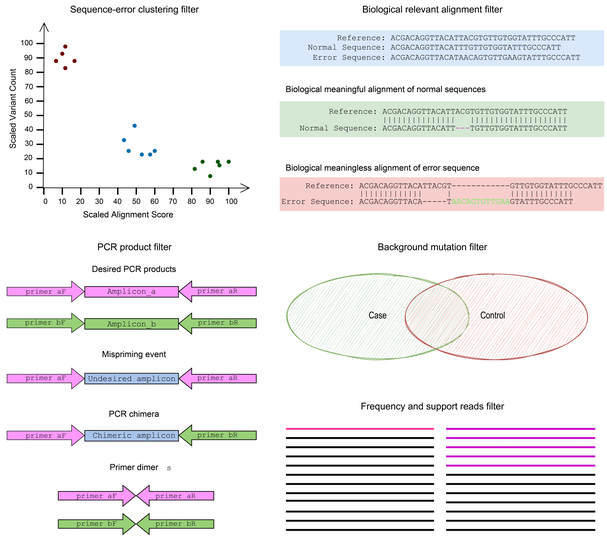

Figure 5. *getools* support various false-positive events filters. Such as an accelerated k-means cluster filter to filter out events with relative low alignment scores and high variant counts; a biological meaning filter to filter out events that might not come from real biological procedures; various PCR artifacts filters to filter out PCR mispriming, chimeras, and dimers; case/control paired background events filter, frequency and support reads as well.

BENCHMARKING RESULTS

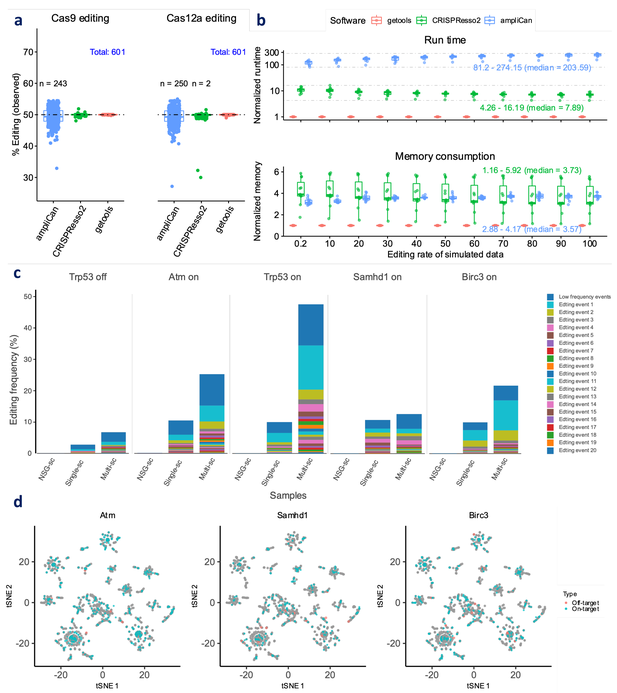

Figure 6. Benchmarking of getools with current pipelines on bulk data (a-b) and mouse single-cell data (c-d). **a**) *getools* outperforms all the existing open-source software/pipeline in accuracy with the synthetic test set constructed by Kurgan et al,.(2021), numbers denote targets with larger than 2% error in editing efficiency determine; **b**) *getools* runs much faster and exhaust much less memory under different edit efficiency libraries; Mouse single-cell data were analyzed in bulk for editing patterns (**c**) or in single-cell mode for co-occurrence (**d**).

CONCLUSION

Designed as a versatile gene editing data analysis toolkit, getools is able to decode the complex editome from gene editing experiments, and outperforms other known pipelines in accuracy and speed, demonstrating the power of getools as a robust and efficient tool in analyzing gene editing data.

getools is now under continuously development and intensively internal test.

For more information

Scan the QR code or contact us at: ngs@edigene.com