Standardizing unique molecular identifiers in SAM flags would benefit more than RNA-Seq

Unique Molecular Identifiers (UMIs) have been incorporated into RNA-Seq experiments to overcome issues with abundance estimation from samples that may have many PCR amplification cycles. However, the use of UMIs in many different types of sequencing experiments could be beneficial, including amplicon sequencing, ATAC-Seq, and ChIP-Seq. Furthermore, UMIs help to overcome artifacts in high-coverage DNA-Seq, and would enable more accurate RNA-Seq genotyping and allele-specific expression calculation. The main advantage of using UMIs is that identical molecules that are true PCR duplicates can be discerned from unique molecules with identical break points.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2465v1 | CC BY 4.0 Open Access | rec: 20 Sep 2016, publ: 20 Sep 2016

Standardizing unique molecular identifiers in SAM flags would benefit more than RNA-Seq

Elisha Roberson[1,2,*]
[1]Washington University, Department of Medicine, Division of Rheumatology, St. Louis, MO 63110
[2]Washington University, Department of Genetics, St. Louis, MO 63110

[*]Contact: eroberson@wustl.edu; @thatdnaguy

The Sequence Alignment Map (SAM) format for high-throughput sequencing (HTS) data, and other related formats such as CRAM, has evolved over time to allow for more advanced types of analyses. After alignment, DNA sequencing reads in SAM files are frequently tagged to identify potential PCR duplicates. Deduplication is critically important for genotyping accuracy, as PCR over-amplification coupled with PCR-induced sequence errors can lead to spurious variant calls. In this context the possibility of a false-positive duplicate is more acceptable than inflating genotyping errors by tolerating many false-negatives.

Duplicate molecule identification is complicated for quantitative technologies, such as RNA-Seq, or in peak finding applications, such as Chromatin ImmunoPrecipitation (ChIP-Seq) and Assay for Transposase-Accessible Chromatin (ATAC-Seq). RNA-Seq transcript quantification is sensitive to both PCR amplification bias, which is especially apparent for single-cell RNA-Seq, and false-positive duplicate identification. The probability of generating multiple, unique fragments with identical breakpoints is not equal across the length of a transcript; the effect is strongest at the transcript ends where there are few unique breakage sites available. Therefore the strength of effect depends on the transcript size, transcript abundance, and the depth of sequencing. For these reasons it is often better practice to ignore duplicate marking entirely for RNA-Seq. The challenges for ChIP-Seq and ATAC-Seq are similar, but occur from targeting relatively small regions. The width of available chromatin for ATAC-Seq and the size of fragments bound to antibody for ChIP-Seq are relatively small. The probability of generating a unique fragment with identical mapping coordinates is again relatively high because of the small target.

Fortunately, an excellent solution for accurate duplicate molecule marking already exists. By incorporating Unique Molecular Identifiers (UMIs) into each read, a true PCR duplicate can discriminated from a unique molecule with identical mapping. UMIs are

unique barcodes that can be added to one or both ends of a sequencing fragment, or alternatively could be added to the end of the index sequence in the adapter. The number of unique molecules that can be identified for any potentially identical fragments is a function of the barcode length. A 4-base single UMI can discriminate 256 unique molecules, and a 6-base single UMI can discriminate 4096 unique molecules. This is highly advantageous since UMIs should only considered within a set of potential duplicate molecules and does not require uniquely tagging every molecule in the library.

The adoption of UMIs in the process of duplicate marking is not without difficulty, and several steps would need to be implemented to make routine use feasible. First, the FASTQ format should incorporate UMI information into the sequence identifier prior to alignment. The current generally accepted FASTQ format for Illumina data includes additional information in the read identifier beyond the unique ID itself: 1) the pair member number, 2) Y|N for whether the read is filtered, 3) control bits, and 4) index sequence. The format could be expanded to a fifth element that lists the UMI sequence in addition to the index sequence. Second, aligners would need to be made aware of the UMI information stored in the FASTQ read identifiers to include that information in the SAM output. There already exists a SAM barcode tag (BC) and associated barcode quality tag (QT). Aligners can either insert the UMI information in these tags, or into a specific UMI tag. Third, duplicate marking software algorithms would need to take advantage of these flags during duplicate marking. A standard duplicate marking algorithm is to crawl a coordinate sorted SAM file and identify fragments with identical 5' mapping coordinates. The candidate duplicates are scanned, and either randomly or by highest quality one read is selected as primary, while the others are marked as duplicates with the 0x400 bit duplicate flag. Updated crawlers would need an intermediate step, where after identifying potential duplicates the process is further subdivided by UMI. For each unique UMI, only one read is primary and the rest are marked duplicates. Since the duplicate flag is a standard feature in the SAM format, downstream tools would be able to readily take advantage of the new deduplicated data.

The standardization of UMIs in the deduplication process would be beneficial not just for RNA-Seq, as it was designed, but would have a positive effect across the spectrum of HTS technology uses, from alleviating high coverage artifacts from DNA-Seq to increasing the quantification and peak detection accuracy of ChIP-Seq and ATAC-Seq. UMI usage would be particularly beneficial for amplicon-Seq, as marking amplicon-Seq duplicates is impossible without them. Genotyping RNA-Seq is challenging without methodologically sound ways of identifying unique sequence fragments l, and would benefit greatly from advanced duplicate marking. Making changes to deeply embedded formats is not without challenges, and software maintenance is not free. Even with

adequate interest and supportive funding, inertia can be as important a factor. Hopefully as the use of UMIs becomes more routine across experiment types the advantages of updating software UMI awareness will overcome the inertia of resistance to change.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2465v1 | CC BY 4.0 Open Access | rec: 20 Sep 2016, publ: 20 Sep 2016