Abstract

SAMBLASTER [1] marks duplicates in one pass over a read-id grouped SAM file. This affords it much of its time and space advantage over other duplicate marking programs such as PICARD *MarkDuplicates* (http://picard.sourceforge.net/) and SAMBAMBA *markdup* (https://github.com/lomereiter/sambamba) that require two passes over a position sorted input file. However, two-pass algorithms have the advantage that they can choose to keep from amongst a set of duplicates, the read pair with the "best" score using some metric that differentiates sequence and/or alignment quality. In contrast, SAMBLASTER's one pass approach can only keep the first pair from a set of duplicate pairs. Given the high quality of Illumina paired-end sequencing, we find this makes little difference in practice. We support this claim with statistics from a well-studied dataset.

Results

Previously, we compared the runtime performance of SAMBLASTER to PICARD and SAMBAMBA in marking duplicates in the ~50X-coverage whole genome sequence data for NA12878 from the Illumina Platinum Genomes (ENA Accession: ERP001960). We now extend this analysis to include the quantity and quality of reads marked as duplicates by SAMBLASTER and PICARD, the generally agreed upon gold standard for quality duplicate marking. In particular, we count the number of reads that each tools marks as duplicate that fall into various categories, note the percent of the total represented by that category, and report the mean alignment score (MAS), mean number of alignment mismatches (MNM) and the mean base sequencing quality scores (MBQS) as measures of alignment and/or sequence quality. The results are summarized in Table 1. Although there are some notable differences in the number of duplicates in various read categories, PICARD and SAMBLASTER find almost the same total number of duplicates. The resulting non-duplicate reads have almost identical MAS, MNM, and MBQS statistics.

Among the duplicates, by far the largest and the most interesting group are the doubly mapped (DM) pairs in which an alignment is found for both reads in the pair. Because PICARD and SAMBLASTER use the same calculation to locate the 5' end of reads used to identify duplicates, both find the identical number of duplicate DM pairs. In addition, they agree on which of the DM pairs to mark as duplicate ~80% of the time. Assuming that PICARD has a metric that distinguishes all of these DM pairs, we would expect that SAMBLASTER could choose to keep the better scoring pair by chance only 50% the time. Therefore, to explain this 80% concordance, two things must be true. First, at least 60% of the duplicate DM pairs must be considered by Picard to be of the same quality, and therefore it has no way to choose between them. We call these "don't cares". Second, PICARD and SAMBLASTER must pick the same DM pair to mark as duplicate for these "don't cares" a disproportionate percentage of the time. This latter condition is likely caused by the fact that the input file to PICARD was position sorted using Novosort (http://www.novocraft.com/Novosort), which is known to use a stable sort algorithm. Therefore, many of the reads with the same nominal genomic position as reported in the SAM file will be in the same read-id order in the input to PICARD as they were in the SAM file used as input to SAMBLASTER. The high percentage of agreement on "don't care" pairs

Faust, G.G. and Hall, I.M., *SAMBLASTER:* fast duplicate marking and structural variant read extraction. *Bioinformatics* Sept. 2014; **30**(17): 2503-2505.

can then be explained if PICARD chooses to keep the first of these "don't care" cases as the non-duplicate pair, thereby picking the same one as SAMBLASTER. For the remaining 20% of the DM pairs in which PICARD and SAMBLASTER disagree (0.43% of the total reads), the duplicates marked by SAMBLASTER have slightly better MAS, MNM, and MBQS statistics than those marked by PICARD, with the concomitant result that the corresponding non-duplicate pairs kept by SAMBLASTER have worse scores for these pairs.

For the remainder of the read categories, it is clear that PICARD and SAMBLASTER are using different strategies to identify duplicates. Read pairs in which one read is mapped and the other unmapped are called "orphans". SAMBLASTER compares orphans only to other orphans to find duplicates, and always marks both reads in an orphan pair as either duplicate or not duplicate. PICARD marks many more mapped reads in orphans as duplicate than SAMBLASTER, and marks no unmapped reads in orphans as duplicates. One possible explanation for this large number of mapped orphan duplicates is that Picard compares the mapped orphan reads to all mapped reads to determine if it is a duplicate. This could also account for the better MAS, MNM, and MBQS scores for PICARD mapped orphan duplicates when compared to either the DM pairs, or SAMBLASTER orphan statistics. Finally, SAMBLASTER marks as duplicate any secondary alignments associated with primary duplicates, while PICARD currently does not. By definition, these are the result of a split mapping of the read, are therefore shorter alignments, and have correspondingly much lower MAS and MNM statistics. The lower MAS and MNM scores for the SAMBLASTER duplicate secondary alignments and mapped orphans are partially compensating for the better scores for duplicate mismatched DM pairs, leading to a final total non-duplicate MAS and MNM for SAMBLASTER that is very close to that of PICARD.

Table 1. Statistics for the number and quality of duplicates and non-duplicate reads for PICARD and SAMBLASTER runs on NA12878. Statistics include the mean alignment score (MAS), mean number of mismatches (MNM), and mean base quality scores (MBQS). The MAS and MNM numbers exclude unaligned reads.

	PICARD					SAMBLASTER				
Read Category	Reads	% Total	MAS	MNM	MBQS	Reads	% Total	MAS	MNM	MBQS
Total Reads	1,578,585,456	100.00	96.954	0.560	36.355	1,578,585,456	100.00	96.954	0.560	36.355
Total Non-duplicates	1,542,595,943	97.72	97.520	0.474	36.353	1,543,282,023	97.76	97.511	0.476	36.341
Total Duplicates	35,989,513	2.28	72.848	4.195	36.439	35,303,433	2.24	72.630	4.234	36.941
Total DM Pairs	34,690,470	2.20	72.600	4.280	36.600	34,690,470	2.20	72.980	4.240	37.050
DM Matching	27,877,300	1.77	72.570	4.310	36.720	27,877,300	1.77	72.570	4.310	36.720
DM Mismatching	6,813,170	0.43	72.740	4.160	36.100	6,813,170	0.43	74.660	3.990	38.420
Orphans, mapped	1,299,043	0.08	79.337	1.964	32.152	190,924	0.01	54.329	3.631	29.703
Orphans, unmapped		0.00				190,924	0.01	NA	NA	24.439
Secondary Alignments		0.00				231,115	0.01	34.761	3.158	36.456

Bibliography

1. Faust, G.G. and Hall, I.M., *SAMBLASTER:* fast duplicate marking and structural variant read extraction*. Bioinformatics* Sept. 2014; **30**(17): 2503-2505.