Recommendations on Read Groups

Release 201911

Sentieon, Inc

Jun 12, 2020

Contents

1 Introduction

This documents describes the recommended usage of the RGID fields to minimize potential problems using the Sentieon Genomics software.

This document should help you determine the best practices for setting the different fields in the RG tags of the bam files used.

2 Detailed description of the RG fields and its usage

2.1 Detailed description of the RG field

The SAM format specification http://samtools.github.io/hts-specs/SAMv1.pdf defines the Read Group as an identifier that groups reads together. The Read Group field in the BAM file can contain the following tags:

- `ID`: IDentifier. A unique identifier for the Read Group. You need to make sure that the RG-ID is unique within the BAM file, and within multiple BAM files that will be used in the same command in a pipeline. This field is required.
- `CN`: Center Name. Name of the sequencing center that sequenced the reads in the Read Group. Typically this tag is not used.

- DS: DeScription. Freeform description of the Read Group. Typically this tag is not used.
- DT: DaTe. Date the run was produced, following ISO8601 date or date/time. Typically this tag is not used.
- FO: Flow Order. The array of nucleotide bases that correspond to the nucleotides used for each flow of each read. Typically this tag is not used.
- KS: Key Sequence. The array of nucleotide bases that correspond to the key sequence of each read. Typically this tag is not used.
- LB: LiBrary. The library used to sequence the reads.
- PG: ProGram. The programs used for processing the read group. Typically the information is included in the PG field of the BAM file, instead of doing this within each Read Group.
- PI: Predicted median Insert size. Typically this tag is not used.
- PL: PLatform. The technology used to sequence the reads. This tag is required if you plan on running BQSR, as it is used to determine the correct error model to apply.
- PM: Platform Model. Freeform text providing further details of the platform/technology used. Typically this tag is not used.
- PU: Platform Unit. Unique identifier for the sequencer unit used to perform the sequencing. This tag is recommended if you plan on running BQSR, as BQSR will model together all reads belonging to the same PU; if the PU is missing, BQSR will model together reads with the same RG-ID.
- SM: Sample name. The sample the reads belong to. This field is required.

2.2 RG field tags and Sentieon

The following are general principles of how RG field tags are used with the Sentieon tools:

- When using multiple input bam files, the ID tags of the bam files need to be unique; there cannot be a RG with the same ID in two different bam input files.
- The tools use the SM tag to identify the reads that belong to the same sample and process them accordingly.
- The Deduplication uses the LB tag to determine which groups may contain duplicates, duplicate reads need to belong to the same library.
- The BQSR model requires the PL tag to determine the error model to apply. If no PL tag is present, the BQSR will not be performed.
- The BQSR modeling will be performed independently on groups of reads identified by the PU tag if it is present; if the PU tag is not present, the BQSR modeling will be performed independently on groups of reads identified by the ID tag.

3 Recommendation on how to fill in the RG fields

Sentieon recommends using the following conventions for the RG field tags:

- ID: sample_name.flowcell.lane.barcode
- SM: sample_name
- PL: technology, i.e. ILLUMINA
- PU: flowcell.lane
- LB: sample_name.library_preparation

The above recommendation makes sure that:

- The read group ID will be unique even across multiple bam files, even for the same sample sequenced in different lanes or using different libraries.
- The BQSR will create a recalibration based on the actual unique sequencing unit, and can be performed on multiple samples if they are sequenced on the same sequencing unit.
- The tumor and normal sample names will be unique for somatic variant calling.

465 Fairchild Drive, Suite 135, Mountain View CA 94043
www.sentieon.com