BioCompute Object for Regulatory Review

BCO Title: GATK Best Practice Data Pre-processing 4.1.0.0

BCO Generation Date: February 16, 2023

BCO Specification Version: v1.3.0

BCO Generator: Seven Bridges

Contents

1 BioCompute Object Domain Entries

1.1 Top Level Fields

```
["https://w3id.org/biocompute/1.4.2/",
"https://biocompute.sbgenomics.com/bco/58218981-5b14-4883-90c2-48c188be74d8",
"57f437f0e2f1162ca3e1b3690860f80cac325996bb89e4965179241d224b9beb"]
```

1.2 Provenance Domain

```
{
"name": "GATK Best Practice Data Pre-processing 4.1.0.0",
"version": "1.0.0",
"review": [],
"derived_from":
"https://cgc-api.sbgenomics.com/v2/apps/phil_webster/bco-cwl-examples/broad-best-practice-data-
"obsolete_after": "2023-02-16T00:00:00+0000",
"embargo": ["2023-02-16T00:00:00+0000",
"2023-02-16T00:00:00+0000"],
"created": "2023-02-16T00:00:00+0000",
"modified": "2023-02-16T00:00:00+0000",
"contributors": [],
"license": "https://spdx.org/licenses/CC-BY-4.0.html"
}
```

1.3 Usability Domain

```
"**Note:** This version of the GATK Best Practice Data
Pre-processing 4.1.0.0 workflow was created for testing
purposes regarding github actions and CI/CD only. Changes
vs the public tool are purely to run tests and should't
affect functionality, but this version is not supported by
SBG in production.\n\n**BROAD Best Practice Data
Pre-processing Workflow 4.1.0.0** is used to prepare data
for variant calling analysis. \n\nIt can be divided into
```

```
two major segments: alignment to reference genome and data
cleanup operations that correct technical biases [1].\n\n*A
list of all inputs and parameters with corresponding
descriptions can be found at the bottom of this
page.*\n\n***Please note that any cloud infrastructure
costs resulting from app and pipeline executions, including
the use of public apps, are the sole responsibility of you
as a user. To avoid excessive costs, please read the app
description carefully and set the app parameters and
execution settings accordingly.***\n\n### Common Use
Cases\n\n* **BROAD Best Practice Data Pre-processing
Workflow 4.1.0.0** is designed to operate on individual
samples.\n* Resulting BAM files are ready for variant
calling analysis and can be further processed by other
BROAD best practice pipelines, like **Generic germline
short variant per-sample calling workflow** [2], **Somatic
CNVs workflow** [3] and **Somatic SNVs+Indel workflow**
[4].\n\n\n### Changes Introduced by Seven Bridges\n\nThis
pipeline represents the CWL implementation of BROADs
[original WDL
file](https://github.com/gatk-workflows/gatk4-data-processing/pull/14)
available on github. Minor differences are introduced in
order to successfully adapt to the Seven Bridges Platform.
These differences are listed below:\n*
**SamToFastqAndBwaMem** step is divided into elementary
steps: **SamToFastq** - converting unaligned BAM file to
interleaved FASTQ file, **BWA Mem** - performing alignment
and **Samtools View** - used for converting SAM file to
BAM.\n* A boolean parameter **Ignore default RG ID** is
added to **BWA MEM Bundle** tool. When used, this parameter
ensures that **BWA MEM Bundle** does not add read group
information (RG) in the BAM file. Instead, RG ID
information obtained from uBAM is added by **GATK
MergeBamAlignment** afterwards. \n* **SortAndFixTags** is
divided into elementary steps: **SortSam** and
**SetNmMdAndUqTags**\n* Added **SBG Lines to Interval
List**: this tool is used to adapt results obtained with
**CreateSequenceGroupingTSV** for platform execution, more
precisely for scattering.\n\n\n\n### Common Issues and
Important Notes\n\n* **BROAD Best Practice Data
Pre-processing Workflow 4.1.0.0** expects unmapped BAM
(uBAM) file format as the main input. One or more read
groups, one per uBAM file, all belonging to a single sample
(SM).\n* **Input Alignments** (`--in_alignments`) -
provided uBAM file should be in query-sorted order and all
reads must have RG tags. Also, input uBAM files must pass
validation by **ValidateSamFile**.\n* For each tool in the
workflow, equivalent parameter settings to the one listed
in the corresponding WDL file are set as defaults. \n\n###
Performance Benchmarking\nSince this CWL implementation is
meant to be equivalent to GATKs original WDL, there are no
additional optimisation steps beside instance and storage
definition. \nThe c5.9xlarge AWS instance hint is used for
WGS inputs and attached storage is set to 1.5TB.\nIn the
table given below one can find results of test runs for WGS
and WES samples. All calculations are performed with
reference files corresponding to assembly 38.\n\n*Cost can
be significantly reduced by spot instance usage. Visit the
[knowledge
center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n| Input Size | Experimental Strategy
| Coverage| Duration | Cost (spot) | AWS Instance Type |\n|
--- | --- | --- | --- | --- | --- | \n| 6.6 GiB | WES | 70
|1h 19min | $2.61 | c5.9 |\n|3.4 GiB | WES | 40 | 42min |
$1.40 | c5.9 |\n| 111.3 GiB| WGS | 30 |22h 41min | $43.86 |
c5.9 |\n| 37.2 GiB | WGS | 10 | 4h 21min | $14.21 | c5.9
|\n\n\n\n### API Python Implementation\nThe app's draft
task can also be submitted via the **API**. In order to
learn how to get your **Authentication token** and **API
endpoint** for corresponding platform visit our
[documentation](https://github.com/sbg/sevenbridges-python#authentication-and-configuration).\
Initialize the SBG Python API\nfrom sevenbridges import
Api\napi = Api(token=\"enter_your_token\",
url=\"enter_api_endpoint\")\n# Get project_id/app_id from
your address bar. Example:
https://igor.sbgenomics.com/u/your_username/project/app\nproject_id
= \"your_username/project\"\napp_id =
\"your_username/project/app\"\n# Replace inputs with
appropriate values\ninputs = {\n\t\"in_alignments\":
list(api.files.query(project=project_id,
names=[\"<unaligned_bam>\"])), \n\t\"reference_index_tar\":
api.files.query(project=project_id,
names=[\"Homo_sapiens_assembly38.fasta.tar\"])[0],
\n\t\"in_reference\": api.files.query(project=project_id,
names=[\"Homo_sapiens_assembly38.fasta\"])[0],
\n\t\"ref_dict\": api.files.query(project=project_id,
names=[\"Homo_sapiens_assembly38.dict\"])[0],\n\t\"known_snps\":
api.files.query(project=project_id,
names=[\"Homo_sapiens_assembly38.dbsnp.vcf\"])[0],\n
\"known_sites\": list(api.files.query(project=project_id,
names=[\"Homo_sapiens_assembly38.known_indels.vcf\",
"Mills_and_1000G_gold_standard.indels.hg38.vcf",
"Homo_sapiens_assembly38.dbsnp.vcf"\n]))}\n# Creates draft
task\ntask = api.tasks.create(name=\"BROAD Best Practice
Data Pre-processing Workflow 4.1.0.0 - API Run\",
project=project_id, app=app_id, inputs=inputs,
run=False)\n```\n\nInstructions for installing and
configuring the API Python client, are provided on
[github](https://github.com/sbg/sevenbridges-python#installation).
For more information about using the API Python client,
consult [the client
documentation](http://sevenbridges-python.readthedocs.io/en/latest/).
**More examples** are available
[here](https://github.com/sbg/okAPI).\n\nAdditionally, [API
R](https://github.com/sbg/sevenbridges-r) and [API
Java](https://github.com/sbg/sevenbridges-java) clients are
available. To learn more about using these API clients
please refer to the [API R client
documentation](https://sbg.github.io/sevenbridges-r/), and
[API Java client
documentation](https://docs.sevenbridges.com/docs/java-library-quickstart).\n\n\n###
References\n\n[1] [Data
Pre-processing](https://software.broadinstitute.org/gatk/best-practices/workflow?id=11165)\n[2
[Generic germline short variant per-sample
calling](https://software.broadinstitute.org/gatk/best-practices/workflow?id=11145)\n[3]
[Somatic
CNVs](https://software.broadinstitute.org/gatk/best-practices/workflow?id=11147)\n[4]
[Somatic SNVs+Indel pipeline
](https://software.broadinstitute.org/gatk/best-practices/workflow?id=11146)"
```

1.4 Extension Domain

```
{
  "fhir_extension": {
    "fhir_endpoint": "",
    "fhir_version": "",
    "fhir_resources": {}
  },
  "scm_extension": {
```

```
    "scm_repository": "",
    "scm_type": "git",
    "scm_commit": "",
    "scm_path": "",
    "scm_preview": ""
  }
}
```

1.5 Description Domain

```
{
"keywords": [],
"xref": [],
"platform": [
"Seven Bridges Platform"
],
"pipeline_steps": [
{
"step_number": "1",
"name": "gatk_markduplicates_4_1_0_0",
"description": "The **GATK MarkDuplicates** tool identifies
duplicate reads in a BAM or SAM file.\n\nThis tool locates
and tags duplicate reads in a BAM or SAM file, where
duplicate reads are defined as originating from a single
fragment of DNA. Duplicates can arise during sample
preparation e.g. library construction using PCR. Duplicate
reads can also result from a single amplification cluster,
incorrectly detected as multiple clusters by the optical
sensor of the sequencing instrument. These duplication
artifacts are referred to as optical duplicates [1].\n\nThe
MarkDuplicates tool works by comparing sequences in the 5
prime positions of both reads and read-pairs in the SAM/BAM
file. The **Barcode tag** (`--BARCODE_TAG`) option is
```

```
available to facilitate duplicate marking using molecular
barcodes. After duplicate reads are collected, the tool
differentiates the primary and duplicate reads using an
algorithm that ranks reads by the sums of their
base-quality scores (default method).\n\n\n###Common Use
Cases\n\n* The **GATK MarkDuplicates** tool requires the
BAM or SAM file on its **Input BAM/SAM file** (`--INPUT`)
input. The tool generates a new SAM or BAM file on its
**Output BAM/SAM** output, in which duplicates have been
identified in the SAM flags field for each read. Duplicates
are marked with the hexadecimal value of 0x0400, which
corresponds to a decimal value of 1024. If you are not
familiar with this type of annotation, please see the
following [blog
post](https://software.broadinstitute.org/gatk/blog?id=7019)
for additional information. **MarkDuplicates** also
produces a metrics file on its **Output metrics file**
output, indicating the numbers of duplicates for both
single and paired end reads.\n\n* The program can take
either coordinate-sorted or query-sorted inputs, however
the behavior is slightly different. When the input is
coordinate-sorted, unmapped mates of mapped records and
supplementary/secondary alignments are not marked as
duplicates. However, when the input is query-sorted
(actually query-grouped), then unmapped mates and
secondary/supplementary reads are not excluded from the
duplication test and can be marked as duplicate reads.\n\n*
If desired, duplicates can be removed using the **Remove
duplicates** (`--REMOVE_DUPLICATES`) and **Remove
sequencing duplicates** ( `--REMOVE_SEQUENCING_DUPLICATES`)
options.\n\n* Although the bitwise flag annotation
indicates whether a read was marked as a duplicate, it does
```

````
not identify the type of duplicate. To do this, a new tag
called the duplicate type (DT) tag was recently added as an
optional output of a SAM/BAM file. Invoking the **Tagging
policy** ( `--TAGGING_POLICY`) option, you can instruct the
program to mark all the duplicates (All), only the optical
duplicates (OpticalOnly), or no duplicates (DontTag). The
records within the output SAM/BAM file will have values for
the 'DT' tag (depending on the invoked **TAGGING_POLICY**
option), as either library/PCR-generated duplicates (LB),
or sequencing-platform artifact duplicates (SQ). \n\n* This
tool uses the **Read name regex** (`--READ_NAME_REGEX`) and
the **Optical duplicate pixel distance**
(`--OPTICAL_DUPLICATE_PIXEL_DISTANCE`) options as the
primary methods to identify and differentiate duplicate
types. Set **READ_NAME_REGEX** to null to skip optical
duplicate detection, e.g. for RNA-seq or other data where
duplicate sets are extremely large and estimating library
complexity is not an aim. Note that without optical
duplicate counts, library size estimation will be
inaccurate.\n\n* Usage example:\n\n```\ngatk MarkDuplicates
\\\n --INPUT input.bam \\\n --OUTPUT marked_duplicates.bam
\\\n --METRICS_FILE
marked_dup_metrics.txt\n```\n\n###Changes Introduced by
Seven Bridges\n\n* All output files will be prefixed using
the **Output prefix** parameter. In case **Output prefix**
is not provided, output prefix will be the same as the
Sample ID metadata from the **Input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, output prefix will be
inferred from the **Input SAM/BAM** filename. This way,
having identical names of the output files between runs is
avoided. Moreover, **dedupped** will be added before the
extension of the output file name. \n\n* The user has a
````

```
possibility to specify the output file format using the
**Output file format** option. Otherwise, the output file
format will be the same as the format of the input
file.\n\n###Common Issues and Important Notes\n\n*
None\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK
MarkDuplicates** for a couple of different samples,
executed on the AWS cloud instances:\n\n| Experiment type |
Input size | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 1.8 GB | 3min | ~0.02$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 5.3 GB | 9min | ~0.06$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 8.8 GB | 16min | ~0.11$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 17 GB | 30min | ~0.20$ | c4.2xlarge
(8 CPUs) |\n\n*Cost can be significantly reduced by using
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
MarkDuplicates](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "2",
"name": "bwa_mem_bundle_0_7_15",
"description": "BWA-MEM is an algorithm designed for
aligning sequence reads onto a large reference genome.
BWA-MEM is implemented as a component of BWA. The algorithm
can automatically choose between performing end-to-end and
local alignments. BWA-MEM is capable of outputting multiple
```

alignments, and finding chimeric reads. It can be applied to a wide range of read lengths, from 70 bp to several megabases. \n\n*A list of **all inputs and parameters** with corresponding descriptions can be found at the bottom of the page.*\n\n\n## Common Use Cases\nIn order to obtain possibilities for additional fast processing of aligned reads, **Biobambam2 sortmadup** (2.0.87) tool is embedded together into the same package with BWA-MEM (0.7.15).\n\nIn order to obtain possibilities for additional fast processing of aligned reads, **Biobambam2** (2.0.87) is embedded together with the BWA 0.7.15 toolkit into the **BWA-MEM Bundle 0.7.15 CWL1.0**. Two tools are used (**bamsort** and **bamsormadup**) to allow the selection of three output formats (SAM, BAM, or CRAM), different modes of sorting (Quarryname/Coordinate sorting), and Marking/Removing duplicates that can arise during sample preparation e.g. library construction using PCR. This is done by setting the **Output format** and **PCR duplicate detection** parameters.\n- Additional notes:\n - The default **Output format** is coordinate sorted BAM (option **BAM**).\n - SAM and BAM options are query name sorted, while CRAM format is not advisable for data sorted by query name.\n - Coordinate Sorted BAM file in all options and CRAM Coordinate sorted output with Marked Duplicates come with the accompanying index file. The generated index name will be the same as the output alignments file, with the extension BAM.BAI or CRAM.CRAI. However, when selecting the CRAM Coordinate sorted and CRAM Coordinate sorted output with Removed Duplicates, the generated files will not have the index file generated. This is a result of the usage of different Biobambam2 tools - **bamsort** does not have the ability to write CRAI files (only supports outputting BAI

```
index files), while **bamsormadup** can write CRAI files.\n
- Passing data from BWA-MEM to Biobambam2 tools has been
done through the Linux piping which saves processing times
(up to an hour of the execution time for whole-genome
sample) of reading and writing of aligned reads into the
hard drive. \n - **BWA-MEM Bundle 0.7.15 CWL1** first needs
to construct the FM-index (Full-text index in Minute space)
for the reference genome using the **BWA INDEX 0.7.17
CWL1.0** tool. The two BWA versions are compatible.\n\n###
Changes Introduced by Seven Bridges\n\n- **Aligned
SAM/BAM/CRAM** file will be prefixed using the **Output
SAM/BAM/CRAM file name** parameter. In case **Output
SAM/BAM/CRAM file name** is not provided, the output prefix
will be the same as the **Sample ID** metadata field from
the file if the **Sample ID** metadata field exists.
Otherwise, the output prefix will be inferred from the
**Input reads** file names.\n- The **Platform** metadata
field for the output alignments will be automatically set
to \"Illumina\" unless it is present in **Input reads**
metadata, or given through **Read group header** or
**Platform** input parameters. This will prevent possible
errors in downstream analysis using the GATK toolkit.\n- If
the **Read group ID** parameter is not defined, by default
it will be set to ‘1’. If the tool is scattered within a
workflow it will assign the **Read Group ID** according to
the order of the scattered folders. This ensures a unique
**Read Group ID** when processing multi-read group input
data from one sample.\n\n### Common Issues and Important
Notes \n \n- For input reads FASTQ files of total size less
than 10 GB we suggest using the default setting for
parameter **Total memory** of 15GB, for larger files we
suggest using 58 GB of memory and 32 CPU cores.\n- When the
```

```
desired output is a CRAM file without deduplication of the
PCR duplicates, it is necessary to provide the FASTA Index
file (FAI) as input.\n- Human reference genome version 38
comes with ALT contigs, a collection of diverged alleles
present in some humans but not the others. Making effective
use of these contigs will help to reduce mapping artifacts,
however, to facilitate mapping these ALT contigs to the
primary assembly, GRC decided to add to each contig long
flanking sequences almost identical to the primary
assembly. As a result, a naive mapping against GRCh38+ALT
will lead to many mapQ-zero mappings in these flanking
regions. Please use post-processing steps to fix these
alignments or implement
[steps](https://sourceforge.net/p/bio-bwa/mailman/message/32845712/)
described by the author of the BWA toolkit.  \n- Inputs
**Read group header** and **Insert string to header** need
to be given in the correct format - under single-quotes.\n-
BWA-MEM is not a splice aware aligner, so it is not the
appropriate tool for mapping RNAseq to the genome. For
RNAseq reads **Bowtie2 Aligner** and **STAR** are
recommended tools. \n- Input paired reads need to have the
identical read names - if not, the tool will throw a
``[mem_sam_pe] paired reads have different names``
error.\n- This wrapper was tested and is fully compatible
with cwltool v3.0.\n\n### Performance Benchmarking\n\nBelow
is a table describing the runtimes and task costs on
on-demand instances for a set of samples with different
file sizes :\n\n| Input reads | Size [GB] | Output format |
Instance (AWS) | Duration | Cost | Threads
|\n|-------------------|-----------|---------------|---------------------------|-----------|---
HG001-NA12878-30x | 2 x 23.8 | SAM | c5.9xlarge (36CPU,
72GB) | 5h 12min | $7.82 | 36 |\n| HG001-NA12878-30x | 2 x
```

```
23.8 | BAM | c5.9xlarge (36CPU, 72GB) | 5h 16min | $8.06 |
36 |\n| HG002-NA24385-50x | 2 x 66.4 | SAM | c5.9xlarge
(36CPU, 72GB) | 8h 33min | $13.08 | 36 |\n\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*",
"version": "0.7.15",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "3",
"name": "gatk_mergebamalignment_4_1_0_0",
"description": "The **GATK MergeBamAlignment** tool is used
for merging BAM/SAM alignment info from a third-party
aligner with the data in an unmapped BAM file, producing a
third BAM file that has alignment data (from the aligner)
and all the remaining data from the unmapped BAM.\n\nMany
alignment tools still require FASTQ format input. The
unmapped BAM may contain useful information that will be
lost in the conversion to FASTQ (meta-data like sample
alias, library, barcodes, etc... as well as read-level
tags.) This tool takes an unaligned BAM with meta-data, and
the aligned BAM produced by calling
[SamToFastq](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_sam
and then passing the result to an aligner. It produces a
new SAM file that includes all aligned and unaligned reads
and also carries forward additional read attributes from
the unmapped BAM (attributes that are otherwise lost in the
process of converting to FASTQ). The resulting file will be
```

```
valid for use by Picard and GATK tools. The output may be
coordinate-sorted, in which case the tags, NM, MD, and UQ
will be calculated and populated, or query-name sorted, in
which case the tags will not be calculated or populated
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n###Common Use Cases\n\n* The **GATK
MergeBamAlignment** tool requires a SAM or BAM file on its
**Aligned BAM/SAM file** (`--ALIGNED_BAM`) input, original
SAM or BAM file of unmapped reads, which must be in
queryname order on its **Unmapped BAM/SAM file**
(`--UNMAPPED_BAM`) input and a reference sequence on its
**Reference** (`--REFERENCE_SEQUENCE`) input. The tool
generates a single BAM/SAM file on its **Output merged
BAM/SAM file** output.\n\n* Usage example:\n\n```\ngatk
MergeBamAlignment \\\\n --ALIGNED_BAM aligned.bam \\\\n
--UNMAPPED_BAM unmapped.bam \\\\n --OUTPUT merged.bam
\\\\n --REFERENCE_SEQUENCE
reference_sequence.fasta\n```\n\n###Changes Introduced by
Seven Bridges\n\n* The output file name will be prefixed
using the **Output prefix** parameter. In case **Output
prefix** is not provided, output prefix will be the same as
the Sample ID metadata from **Input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, output prefix will be
inferred from the **Input SAM/BAM file** filename. This
way, having identical names of the output files between
runs is avoided. Moreover, **merged** will be added before
the extension of the output file name. \n\n* The user has a
possibility to specify the output file format using the
**Output file format** argument. Otherwise, the output file
format will be the same as the format of the input aligned
file.\n\n###Common Issues and Important Notes\n\n* Note:
```

```
This is not a tool for taking multiple BAM/SAM files and
creating a bigger file by merging them. For that use-case,
see
[MergeSamFiles](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_
Benchmarking\n\nBelow is a table describing runtimes and
task costs of **GATK MergeBamAlignment** for a couple of
different samples, executed on the AWS cloud
instances:\n\n| Experiment type | Aligned BAM/SAM size |
Unmapped BAM/SAM size | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
RNA-Seq | 1.4 GB | 1.9 GB | 9min | ~0.06$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 4.0 GB | 5.7 GB | 20min | ~0.13$ |
c4.2xlarge (8 CPUs) | \n| RNA-Seq | 6.6 GB | 9.5 GB | 32min
| ~0.21$ | c4.2xlarge (8 CPUs) | \n| RNA-Seq | 13 GB | 19
GB | 1h 4min | ~0.42$ | c4.2xlarge (8 CPUs) |\n\n*Cost can
be significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
MergeBamAlignment](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/pica
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "4",
"name": "gatk_samtofastq_4_1_0_0",
"description": "The **GATK SamToFastq** tool converts a SAM
or BAM file to FASTQ.\n\nThis tool extracts read sequences
and qualities from the input SAM/BAM file and writes them
into the output file in Sanger FASTQ format.\n\nIn the RC
```

```
mode (default is True), if the read is aligned and the
alignment is to the reverse strand on the genome, the read
sequence from input SAM file will be reverse-complemented
prior to writing it to FASTQ in order to correctly restore
the original read sequence as it was generated by the
sequencer [1].\n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n###Common Use Cases\n\n* The **GATK
SamToFastq** tool requires a BAM/SAM file on its **Input
BAM/SAM file** (`--INPUT`) input. The tool generates a
single-end FASTQ file on its **Output FASTQ file(s)**
output if the input BAM/SAM file is single end. In case the
input file is paired end, the tool outputs the first end of
the pair FASTQ and the second end of the pair FASTQ on its
**Output FASTQ file(s)** output, except when the
**Interleave** (`--INTERLEAVE`) option is set to True. If
the output is an interleaved FASTQ file, if paired, each
line will have /1 or /2 to describe which end it came
from.\n\n* The **GATK SamToFastq** tool supports an
optional parameter **Output by readgroup**
(`--OUTPUT_BY_READGROUP`) which, when true, outputs a FASTQ
file per read group (two FASTQ files per read group if the
group is paired).\n\n* Usage example (input BAM file is
single-end):\n\n```\ngatk SamToFastq \n --INPUT input.bam\n
--FASTQ output.fastq\n```\n\n\n\n\n\n* Usage example (input
BAM file is paired-end):\n\n```\ngatk SamToFastq \n --INPUT
input.bam\n --FASTQ output.pe_1.fastq\n --SECOND_END_FASTQ
output.pe_2.fastq\n --UNPAIRED_FASTQ
unpaired.fastq\n\n```\n\n###Changes Introduced by Seven
Bridges\n\n* The GATK SamToFastq tool is implemented to
check if the input alignments file contains single-end or
paired-end data and according to that generates different
```

```
command lines for these two modes and thus produces
appropriate output files on its **Output FASTQ file(s)**
output (one FASTQ file in single-end mode and two FASTQ
files if the input alignment file contains paired-end
data). \n\n* All output files will be prefixed using the
**Output prefix** parameter. In case the **Output prefix**
is not provided, the output prefix will be the same as the
Sample ID metadata from the **input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, the output prefix
will be inferred from the **Input SAM/BAM** filename. This
way, having identical names of the output files between
runs is avoided.\n\n* For paired-end read files, in order
to successfully run alignment with STAR, this tool adds the
appropriate **paired-end** metadata field in the output
FASTQ files.\n\n###Common Issues and Important Notes\n\n*
None\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK SamToFastq**
for a couple of different samples, executed on the AWS
cloud instances:\n\n| Experiment type | Input size |
Paired-end | # of reads | Read length | Duration | Cost |
Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
RNA-Seq | 1.9 GB | Yes | 16M | 101 | 4min | ~0.03$ |
c4.2xlarge (8 CPUs) | \n| RNA-Seq | 5.7 GB | Yes | 50M |
101 | 7min | ~0.04$ | c4.2xlarge (8 CPUs) | \n| RNA-Seq |
9.5 GB | Yes | 82M | 101 | 10min | ~0.07$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 19 GB | Yes | 164M | 101 | 20min |
~0.13$ | c4.2xlarge (8 CPUs) |\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n###References\n\n[1] [GATK
```

```
SamToFastq](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.12.0/picard_sar
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "5",
"name": "gatk_sortsam_4_1_0_0",
"description": "The **GATK SortSam** tool sorts the input
SAM or BAM file by coordinate, queryname (QNAME), or some
other property of the SAM record.\n\nThe **GATK SortOrder**
of a SAM/BAM file is found in the SAM file header tag @HD
in the field labeled SO.  For a coordinate\nsorted SAM/BAM
file, read alignments are sorted first by the reference
sequence name (RNAME) field using the reference\nsequence
dictionary (@SQ tag).  Alignments within these subgroups
are secondarily sorted using the left-most
mapping\nposition of the read (POS).  Subsequent to this
sorting scheme, alignments are listed
arbitrarily.<\/p><p>For\nqueryname-sorted alignments, all
alignments are grouped using the queryname field but the
alignments are not necessarily\nsorted within these groups.
Reads having the same queryname are derived from the same
template\n\n\n###Common Use Cases\n\nThe **GATK SortSam**
tool requires a BAM/SAM file on its **Input SAM/BAM file**
(`--INPUT`) input. The tool sorts input file in the order
defined by (`--SORT_ORDER`) parameter. Available sort order
options are `queryname`, `coordinate` and `duplicate`.
\n\n* Usage example:\n\n```\njava -jar picard.jar SortSam\n
--INPUT=input.bam \n
--SORT_ORDER=coordinate\n```\n\n\n###Changes Introduced by
```

```
Seven Bridges\n\n* Prefix of the output file is defined
with the optional parameter **Output prefix**. If **Output
prefix** is not provided, name of the sorted file is
obtained from **Sample ID** metadata from the **Input
SAM/BAM file**, if the **Sample ID** metadata exists.
Otherwise, the output prefix will be inferred form the
**Input SAM/BAM file** filename. \n\n\n###Common Issues and
Important Notes\n\n* None\n\n\n###Performance
Benchmarking\nBelow is a table describing runtimes and task
costs of **GATK SortSam** for a couple of different
samples, executed on the AWS cloud instances:\n\n|
Experiment type | Input size | Paired-end | # of reads |
Read length | Duration | Cost | Instance (AWS) |
\n|:-------------:|:-----------:|:-------:|:-------:|:---------:|:----------:|:------:|:---
WGS | | Yes | 16M | 101 | 4min | ~0.03$ | c4.2xlarge (8
CPUs) | \n| WGS | | Yes | 50M | 101 | 7min | ~0.04$ |
c4.2xlarge (8 CPUs) | \n| WGS | | Yes | 82M | 101 | 10min |
~0.07$ | c4.2xlarge (8 CPUs) | \n| WES | | Yes | 164M | 101
| 20min | ~0.13$ | c4.2xlarge (8 CPUs) |\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n\n###References\n[1] [GATK SortSam
home
page](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.12.0/picard_sam_SortS
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "6",
"name": "gatk_setnmmdanduqtags_4_1_0_0",
"description": "The **GATK SetNmMdAndUqTags** tool takes in
a coordinate-sorted SAM or BAM and calculatesthe NM, MD,
and UQ tags by comparing it with the reference. \n\nThe
**GATK SetNmMdAndUqTags** may be needed when **GATK
MergeBamAlignment** was run with **SORT_ORDER** other than
`coordinate` and thus could not fix these tags.
\n\n\n###Common Use Cases\nThe **GATK SetNmMdAndUqTags**
tool fixes NM, MD and UQ tags in SAM/BAM file **Input
SAM/BAM file** (`--INPUT`) input. This tool takes in a
coordinate-sorted SAM or BAM file and calculates the NM,
MD, and UQ tags by comparing with the reference **Reference
sequence** (`--REFERENCE_SEQUENCE`).\n\n* Usage
example:\n\n```\njava -jar picard.jar SetNmMdAndUqTags\n
--REFERENCE_SEQUENCE=reference_sequence.fasta\n
--INPUT=sorted.bam\n```\n\n\n###Changes Introduced by Seven
Bridges\n\n* Prefix of the output file is defined with the
optional parameter **Output prefix**. If **Output prefix**
is not provided, name of the sorted file is obtained from
**Sample ID** metadata form the **Input SAM/BAM file**, if
the **Sample ID** metadata exists. Otherwise, the output
prefix will be inferred form the **Input SAM/BAM file**
filename. \n\n\n\n###Common Issues and Important Notes\n\n*
The **Input SAM/BAM file** must be coordinate sorted in
order to run **GATK SetNmMdAndUqTags**. \n* If specified,
the MD and NM tags can be ignored and only the UQ tag be
set. \n\n\n###References\n[1] [GATK SetNmMdAndUqTags home
page](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard_sam_SetNm
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
```

```
},
{
"step_number": "7",
"name": "gatk_baserecalibrator_4_1_0_0",
"description": "**GATK BaseRecalibrator** generates a
recalibration table based on various covariates for input
mapped read data [1]. It performs the first pass of the
Base Quality Score Recalibration (BQSR) by assessing base
quality scores of the input data.\n\n*A list of **all
inputs and parameters** with corresponding descriptions can
be found at the bottom of the page.*\n\n###Common Use
Cases\n\n* The **GATK BaseRecalibrator** tool requires the
input mapped read data whose quality scores need to be
assessed on its **Input alignments** (`--input`) input, the
database of known polymorphic sites to skip over on its
**Known sites** (`--known-sites`) input and a reference
file on its **Reference** (`--reference`) input. On its
**Output recalibration report** output, the tool generates
a GATK report file with many tables: the list of arguments,
the quantized qualities table, the recalibration table by
read group, the recalibration table by quality score,\nthe
recalibration table for all the optional covariates
[1].\n\n* Usage example:\n\n```\ngatk --java-options
\"-Xmx2048M\" BaseRecalibrator \\\n --input my_reads.bam
\\\n --reference reference.fasta \\\n --known-sites
sites_of_variation.vcf \\\n --known-sites
another/optional/setOfSitesToMask.vcf \\\n --output
recal_data.table\n\n```\n\n###Changes Introduced by Seven
Bridges\n\n* The output file will be prefixed using the
**Output name prefix** parameter. If this value is not set,
the output name will be generated based on the **Sample
ID** metadata value from the input alignment file. If the
```

Sample ID value is not set, the name will be inherited from the input alignment file name. In case there are multiple files on the **Input alignments** input, the files will be sorted by name and output file name will be generated based on the first file in the sorted file list, following the rules defined in the previous case. Moreover, **recal_data** will be added before the extension of the output file name which is **CSV** by default.\n\n* **Include intervals** (`--intervals`) option is divided into **Include intervals string** and **Include intervals file** options.\n\n* **Exclude intervals** (`--exclude-intervals`) option is divided into **Exclude intervals string** and **Exclude intervals file** options.\n\n* The following GATK parameters were excluded from the tool wrapper: `--add-output-sam-program-record`, `--add-output-vcf-command-line`, `--arguments_file`, `--cloud-index-prefetch-buffer`, `--cloud-prefetch-buffer`, `--create-output-bam-index`, `--create-output-bam-md5`, `--create-output-variant-index`, `--create-output-variant-md5`, `--gatk-config-file`, `--gcs-max-retries`, `--gcs-project-for-requester-pays`, `--help`, `--lenient`, `--QUIET`, `--sites-only-vcf-output`, `--showHidden`, `--tmp-dir`, `--use-jdk-deflater`, `--use-jdk-inflater`, `--verbosity`, `--version`\n\n\n\n###Common Issues and Important Notes\n\n* **Memory per job** (`mem_per_job`) input allows a user to set the desired memory requirement when running a tool or adding it to a workflow. This input should be defined in MB. It is propagated to the Memory requirements part and "-Xmx" parameter of the tool. The default value is 2048MB.\n* **Memory overhead per job** (`mem_overhead_per_job`) input allows a user to set the

```
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n* Note: GATK tools
that take in mapped read data expect a BAM file as the
primary format [2]. More on GATK requirements for mapped
sequence data formats can be found
[here](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapped
Note: **Known sites**, **Input alignments** should have
corresponding index files in the same folder. \n* Note:
**Reference** FASTA file should have corresponding .fai
(FASTA index) and .dict (FASTA dictionary) files in the
same folder. \n* Note: These **Read Filters**
(`--read-filter`) are automatically applied to the data by
the Engine before processing by **BaseRecalibrator** [1]:
**NotSecondaryAlignmentReadFilter**,
**PassesVendorQualityCheckReadFilter**,
**MappedReadFilter**,
**MappingQualityAvailableReadFilter**,
**NotDuplicateReadFilter**,
**MappingQualityNotZeroReadFilter**,
**WellformedReadFilter**\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"LibraryReadFilter\",
the **Library** (`--library`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"PlatformReadFilter\", the **Platform
filter name** (`--platform-filter-name`) option must be set
to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set
to\"PlatformUnitReadFilter\", the **Black listed lanes**
(`--black-listed-lanes`) option must be set to some value.
```

```
\n* Note: If the **Read filter** (`--read-filter`) option
is set to \"ReadGroupBlackListReadFilter\", the **Read
group black list** (`--read-group-black-list`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"ReadGroupReadFilter\",
the **Keep read group** (`--keep-read-group`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to
\"ReadLengthReadFilter\", the **Max read length**
(`--max-read-length`) option must be set to some value.\n*
Note: If the **Read filter** (`--read-filter`) option is
set to \"ReadNameReadFilter\", the **Read name**
(`--read-name`) option must be set to some value.\n* Note:
If the **Read filter** (`--read-filter`) option is set to
\"ReadStrandFilter\", the **Keep reverse strand only**
(`--keep-reverse-strand-only`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"SampleReadFilter\", the **Sample**
(`--sample`) option must be set to some value.\n* Note: The
following options are valid only if the appropriate **Read
filter** (`--read-filter`) is specified: **Ambig filter
bases** (`--ambig-filter-bases`), **Ambig filter frac**
(`--ambig-filter-frac`), **Max fragment length**
(`--max-fragment-length`), **Maximum mapping quality**
(`--maximum-mapping-quality`), **Minimum mapping quality**
(`--minimum-mapping-quality`), **Do not require soft
clips** (`--dont-require-soft-clips-both-ends`), **Filter
too short** (`--filter-too-short`), **Min read length**
(`--min-read-length`). See the description of each
parameter for information on the associated **Read
filter**.\n* Note: The wrapper has not been tested for the
SAM file type on the **Input alignments** input port, nor
```

```
for the BCF file type on the **Known sites** input
port.\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK
BaseRecalibrator** for a couple of different samples,
executed on AWS cloud instances:\n\n| Experiment type |
Input size | Duration | Cost (on-demand) | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 2.2 GB | 9min | ~0.08$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 6.6 GB | 19min | ~0.17$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 11 GB | 27min | ~0.24$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 22 GB | 46min | ~0.41$ | c4.2xlarge
(8 CPUs) |\n\n*Cost can be significantly reduced by using
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
BaseRecalibrator](https://gatk.broadinstitute.org/hc/en-us/articles/360036726891-BaseRecalibra
[GATK Mapped sequence data
formats](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapp
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "8",
"name": "gatk_createsequencegroupingtsv_4_1_0_0",
"description": "**CreateSequenceGroupingTSV** tool generate
sets of intervals for scatter-gathering over
chromosomes.\n\nIt takes **Reference dictionary** file
(`--ref_dict`) as an input and creates files which contain
chromosome names grouped based on their
sizes.\n\n\n###**Common Use Cases**\n\nThe tool has only
```

one input (`--ref_dict`) which is required and has no
additional arguments. **CreateSequenceGroupingTSV** tool
results are **Sequence Grouping** file which is a text file
containing chromosome groups, and **Sequence Grouping with
Unmapped**, a text file which has the same content as
Sequence Grouping with additional line containing
\"unmapped\" string.\n\n\n* Usage example\n\n\n```\npython
CreateSequenceGroupingTSV.py \n --ref_dict
example_reference.dict\n\n```\n\n\n\n###**Changes
Introduced by Seven Bridges**\n\nPython code provided
within WGS Germline WDL was adjusted to be called as a
script (`CreateSequenceGroupingTSV.py`).\n\n\n###**Common
Issues and Important Notes**\n\nNone.\n\n\n###
Reference\n[1]
[CreateSequenceGroupingTSV](https://github.com/gatk-workflows/broad-prod-wgs-germline-snps-inde
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "9",
"name": "gatk_gatherbqsrreports_4_1_0_0",
"description": "**GATK GatherBQSRReports** gathers
scattered BQSR recalibration reports into a single file
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n\n### Common Use Cases \n\n* This tool is
intended to be used to combine recalibration tables from
runs of **GATK BaseRecalibrator** parallelized
per-interval.\n\n* Usage example:\n```\n gatk
--java-options \"-Xmx2048M\" GatherBQSRReports \\\n --input

````
input1.csv \\\n --input input2.csv \\\n --output
output.csv\n\n```\n\n\n###Changes Introduced by Seven
Bridges\n\n* The output file will be prefixed using the
**Output name prefix** parameter. If this value is not set,
the output name will be generated based on the **Sample
ID** metadata value from **Input BQSR reports**. If the
**Sample ID** value is not set, the name will be inherited
from the **Input BQSR reports** file name. In case there
are multiple files on the **Input BQSR reports** input, the
files will be sorted by name and output file name will be
generated based on the first file in the sorted file list,
following the rules defined in the previous case. Moreover,
**.recal_data** will be added before the extension of the
output file name.\n\n* The following GATK parameters were
excluded from the tool wrapper: `--arguments_file`,
`--gatk-config-file`, `--gcs-max-retries`,
`--gcs-project-for-requester-pays`, `--help`, `--QUIET`,
`--showHidden`, `--tmp-dir`, `--use-jdk-deflater`,
`--use-jdk-inflater`, `--verbosity`,
`--version`\n\n\n###Common Issues and Important Notes\n\n*
**Memory per job** (`mem_per_job`) input allows a user to
set the desired memory requirement when running a tool or
adding it to a workflow. This input should be defined in
MB. It is propagated to the Memory requirements part and
"-Xmx" parameter of the tool. The default value is
2048MB.\n\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n\n\n###Performance
````

```
Benchmarking\n\nThis tool is fast, with a running time of a
few minutes. The experiment task was performed on the
default AWS on-demand c4.2xlarge instance on 50 CSV files
(size of each ~350KB) and took 2 minutes to finish
($0.02).\n\n*Cost can be significantly reduced by using
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n###References\n\n[1] [GATK
GatherBQSRReports](https://gatk.broadinstitute.org/hc/en-us/articles/360036359192-GatherBQSRRe
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "10",
"name": "gatk_applybqsr_4_1_0_0",
"description": "The **GATK ApplyBQSR** tool recalibrates
the base quality scores of an input BAM or CRAM file
containing reads.\n\nThis tool performs the second pass in
a two-stage process called Base Quality Score Recalibration
(BQSR). Specifically, it recalibrates the base qualities of
the input reads based on the recalibration table produced
by the **GATK BaseRecalibrator** tool. The goal of this
procedure is to correct systematic biases that affect the
assignment of base quality scores by the sequencer. The
first pass consists of calculating the error empirically
and finding patterns in how the error varies with basecall
features over all bases. The relevant observations are
written to the recalibration table. The second pass
consists of applying numerical corrections to each
individual basecall, based on the patterns identified in
```

the first step (recorded in the recalibration table), and writing out the recalibrated data to a new BAM or CRAM file [1].\n\n*A list of **all inputs and parameters** with corresponding descriptions can be found at the bottom of the page.*\n\n###Common Use Cases\n\n* The **GATK ApplyBQSR** tool requires a BAM or CRAM file on its **Input alignments** (`--input`) input and the covariates table (= recalibration file) generated by the **BaseRecalibrator** tool on its **BQSR recal file** input (`--bqsr-recal-file`). If the input alignments file is in the CRAM format, the reference sequence is required on the **Reference** (`--reference`) input of the tool. The tool generates a new alignments file which contains recalibrated read data on its **Output recalibrated alignments** output.\n\n* Usage example\n\n```\n gatk --java-options \"-Xmx2048M\" ApplyBQSR \\\n --reference reference.fasta \\\n --input input.bam \\\n --bqsr-recal-file recalibration.table \\\n --output output.bam\n\n```\n\n* Original qualities can be retained in the output file under the \"OQ\" tag if desired. See the **Emit original quals** (`--emit-original-quals`) argument for details [1].\n\n###Changes Introduced by Seven Bridges\n\n* The output file will be prefixed using the **Output name prefix** parameter. If this value is not set, the output name will be generated based on the **Sample ID** metadata value from the input alignments file. If the **Sample ID** value is not set, the name will be inherited from the input alignments file name. In case there are multiple files on the **Input alignments** input, the files will be sorted by name and output file name will be generated based on the first file in the sorted file list, following the rules defined in the previous case. Moreover, **recalibrated**

```
will be added before the extension of the output file
name.\n\n* The user has a possibility to specify the output
file format using the **Output file format** argument.
Otherwise, the output file format will be the same as the
format of the input file.\n\n* **Include intervals**
(`--intervals`) option is divided into **Include intervals
string** and **Include intervals file** options.\n\n*
**Exclude intervals** (`--exclude-intervals`) option is
divided into **Exclude intervals string** and **Exclude
intervals file** options.\n\n* The following GATK
parameters were excluded from the tool wrapper:
`--add-output-vcf-command-line`, `--arguments_file`,
`--cloud-index-prefetch-buffer`, `--cloud-prefetch-buffer`,
`--create-output-bam-md5`, `--create-output-variant-index`,
`--create-output-variant-md5`, `--gatk-config-file`,
`--gcs-max-retries`, `--gcs-project-for-requester-pays`,
`--help`, `--lenient`, `--QUIET`,
`--sites-only-vcf-output`, `--showHidden`, `--tmp-dir`,
`--use-jdk-deflater`, `--use-jdk-inflater`, `--verbosity`,
`--version`\n\n###Common Issues and Important Notes\n\n*
**Memory per job** (`mem_per_job`) input allows a user to
set the desired memory requirement when running a tool or
adding it to a workflow. This input should be defined in
MB. It is propagated to the Memory requirements part and
"-Xmx" parameter of the tool. The default value is
2048MB.\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n* Note: GATK tools
```

that take in mapped read data expect a BAM file as the primary format [2]. More on GATK requirements for mapped sequence data formats can be found [here](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mappe
Note: **Input alignments** should have corresponding index files in the same folder. \n* Note: **Reference** FASTA file should have corresponding .fai (FASTA index) and .dict (FASTA dictionary) files in the same folder. \n* Note: This tool replaces the use of PrintReads for the application of base quality score recalibration as practiced in earlier versions of GATK (2.x and 3.x) [1].\n* Note: You should only run **ApplyBQSR** with the covariates table created from the input BAM or CRAM file [1].\n* Note: This **Read Filter** (`--read-filter`) is automatically applied to the data by the Engine before processing by **ApplyBQSR** [1]: **WellformedReadFilter**\n* Note: If the **Read filter** (`--read-filter`) option is set to \"LibraryReadFilter\", the **Library** (`--library`) option must be set to some value.\n* Note: If the **Read filter** (`--read-filter`) option is set to \"PlatformReadFilter\", the **Platform filter name** (`--platform-filter-name`) option must be set to some value.\n* Note: If the **Read filter** (`--read-filter`) option is set to\"PlatformUnitReadFilter\", the **Black listed lanes** (`--black-listed-lanes`) option must be set to some value. \n* Note: If the **Read filter** (`--read-filter`) option is set to \"ReadGroupBlackListReadFilter\", the **Read group black list** (`--read-group-black-list`) option must be set to some value.\n* Note: If the **Read filter** (`--read-filter`) option is set to \"ReadGroupReadFilter\", the **Keep read group** (`--keep-read-group`) option must be set to some value.\n* Note: If the **Read filter**

```
(`--read-filter`) option is set to
\"ReadLengthReadFilter\", the **Max read length**
(`--max-read-length`) option must be set to some value.\n*
Note: If the **Read filter** (`--read-filter`) option is
set to \"ReadNameReadFilter\", the **Read name**
(`--read-name`) option must be set to some value.\n* Note:
If the **Read filter** (`--read-filter`) option is set to
\"ReadStrandFilter\", the **Keep reverse strand only**
(`--keep-reverse-strand-only`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"SampleReadFilter\", the **Sample**
(`--sample`) option must be set to some value.\n* Note: The
following options are valid only if an appropriate **Read
filter** (`--read-filter`) is specified: **Ambig filter
bases** (`--ambig-filter-bases`), **Ambig filter frac**
(`--ambig-filter-frac`), **Max fragment length**
(`--max-fragment-length`), **Maximum mapping quality**
(`--maximum-mapping-quality`), **Minimum mapping quality**
(`--minimum-mapping-quality`), **Do not require soft
clips** (`--dont-require-soft-clips-both-ends`), **Filter
too short** (`--filter-too-short`), **Min read length**
(`--min-read-length`). See the description of each
parameter for information on the associated **Read
filter**.\n* Note: The wrapper has not been tested for the
SAM file type on the **Input alignments** input
port.\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK ApplyBQSR**
for a couple of different samples, executed on the AWS
cloud instances:\n\n| Experiment type | Input size |
Duration | Cost (on-demand) | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 2.2 GB | 8min | ~0.07$ | c4.2xlarge (8 CPUs) |
```

\n| RNA-Seq | 6.6 GB | 23min | ~0.21$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 11 GB | 37min | ~0.33$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 22 GB | 1h 16min | ~0.68$ |
c4.2xlarge (8 CPUs) |\n\n*Cost can be significantly reduced
by using **spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
ApplyBQSR](https://gatk.broadinstitute.org/hc/en-us/articles/360036725911-ApplyBQSR)\n\n[2]
[GATK Mapped sequence data
formats](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapp
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "11",
"name": "gatk_gatherbamfiles_4_1_0_0",
"description": "**GATK GatherBamFiles** concatenates one or
more BAM files resulted form scattered paralel anaysis.
\n\n\n### Common Use Cases \n\n* **GATK GatherBamFiles**
tool performs a rapid \"gather\" or concatenation on BAM
files into single BAM file. This is often needed in
operations that have been run in parallel across genomics
regions by scattering their execution across computing
nodes and cores thus resulting in smaller BAM files.\n*
Usage example:\n```\n\njava -jar picard.jar
GatherBamFiles\n --INPUT=input1.bam\n
--INPUT=input2.bam\n```\n\n### Common Issues and Important
Notes\n* **GATK GatherBamFiles** assumes that the list of
BAM files provided as input are in the order that they
should be concatenated and simply links the bodies of the

```
BAM files while retaining the header from the first file.
\n* Operates by copying the gzip blocks directly for speed
but also supports the generation of an MD5 in the output
file and the indexing of the output BAM file.\n* This tool
only support BAM files. It does not support SAM
files.\n\n###Changes Intorduced by Seven Bridges\n*
Generated output BAM file will be prefixed using the
**Output prefix** parameter. In case the **Output prefix**
is not provided, the output prefix will be the same as the
**Sample ID** metadata from the **Input alignments**, if
the **Sample ID** metadata exists. Otherwise, the output
prefix will be inferred from the **Input alignments**
filename. This way, having identical names of the output
files between runs is avoided.",
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "12",
"name": "samtools_view_1_9_cwl1_0",
"description": "**SAMtools View** tool prints all
alignments from a SAM, BAM, or CRAM file to an output file
in SAM format (headerless). You may specify one or more
space-separated region specifications to restrict output to
only those alignments which overlap the specified
region(s). Use of region specifications requires a
coordinate-sorted and indexed input file (in BAM or CRAM
format) [1].\n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n####Regions\n\nRegions can be specified
```

```
as: RNAME[:STARTPOS[-ENDPOS]] and all position coordinates
are 1-based. \n\n**Important note:** when multiple regions
are given, some alignments may be output multiple times if
they overlap more than one of the specified
regions.\n\nExamples of region specifications:\n\n-
**chr1** - Output all alignments mapped to the reference
sequence named `chr1' (i.e. @SQ SN:chr1).\n\n-
**chr2:1000000** - The region on chr2 beginning at base
position 1,000,000 and ending at the end of the
chromosome.\n\n- **chr3:1000-2000** - The 1001bp region on
chr3 beginning at base position 1,000 and ending at base
position 2,000 (including both end positions).\n\n-
**'\\*'** - Output the unmapped reads at the end of the
file. (This does not include any unmapped reads placed on a
reference sequence alongside their mapped mates.)\n\n-
**.** - Output all alignments. (Mostly unnecessary as not
specifying a region at all has the same effect.)
[1]\n\n###Common Use Cases\n\nThis tool can be used for:
\n\n- Filtering BAM/SAM/CRAM files - options set by the
following parameters and input files: **Include reads with
all of these flags** (`-f`), **Exclude reads with any of
these flags** (`-F`), **Exclude reads with all of these
flags** (`-G`), **Read group** (`-r`), **Minimum mapping
quality** (`-q`), **Only include alignments in library**
(`-l`), **Minimum number of CIGAR bases consuming query
sequence** (`-m`), **Subsample fraction** (`-s`), **Read
group list** (`-R`), **BED region file** (`-L`)\n- Format
conversion between SAM/BAM/CRAM formats - set by the
following parameters: **Output format**
(`--output-fmt/-O`), **Fast bam compression** (`-1`),
**Output uncompressed BAM** (`-u`)\n- Modification of the
data which is contained in each alignment - set by the
```

```
following parameters: **Collapse the backward CIGAR
operation** (`-B`), **Read tags to strip** (`-x`)\n-
Counting number of alignments in SAM/BAM/CRAM file - set by
parameter **Output only count of matching records**
(`-c`)\n\n###Changes Introduced by Seven Bridges\n\n-
Parameters **Output BAM** (`-b`) and **Output CRAM** (`-C`)
were excluded from the wrapper since they are redundant
with parameter **Output format** (`--output-fmt/-O`).\n-
Parameter **Input format** (`-S`) was excluded from wrapper
since it is ignored by the tool (input format is
auto-detected).\n- Input file **Index file** was added to
the wrapper to enable operations that require an index file
for BAM/CRAM files.\n- Parameter **Number of threads**
(`--threads/-@`) specifies the total number of threads
instead of additional threads. Command line argument
(`--threads/-@`) will be reduced by 1 to set the number of
additional threads.\n\n###Common Issues and Important
Notes\n\n- When multiple regions are given, some alignments
may be output multiple times if they overlap more than one
of the specified regions [1].\n- Use of region
specifications requires a coordinate-sorted and indexed
input file (in BAM or CRAM format) [1].\n- Option **Output
uncompressed BAM** (`-u`) saves time spent on
compression/decompression and is thus preferred when the
output is piped to another SAMtools command
[1].\n\n###Performance Benchmarking\n\nMultithreading can
be enabled by setting parameter **Number of threads**
(`--threads/-@`). In the following table you can find
estimates of **SAMtools View** running time and cost.
\n\n*Cost can be significantly reduced by using **spot
instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
```

```
for more details.* \n\n| Input type | Input size | # of
reads | Read length | Output format | # of threads |
Duration | Cost | Instance
(AWS)|\n|---------------|--------------|-----------------|---------------|------------------|-
BAM | 5.26 GB | 71.5M | 76 | BAM | 1 | 13min. | \\$0.12 |
c4.2xlarge |\n| BAM | 11.86 GB | 161.2M | 101 | BAM | 1 |
33min. | \\$0.30 | c4.2xlarge |\n| BAM | 18.36 GB | 179M |
76 | BAM | 1 | 60min. | \\$0.54 | c4.2xlarge |\n| BAM |
58.61 GB | 845.6M | 150 | BAM | 1 | 3h 25min. | \\$1.84 |
c4.2xlarge |\n| BAM | 5.26 GB | 71.5M | 76 | BAM | 8 |
5min. | \\$0.04 | c4.2xlarge |\n| BAM | 11.86 GB | 161.2M |
101 | BAM | 8 | 11min. | \\$0.10 | c4.2xlarge |\n| BAM |
18.36 GB | 179M | 76 | BAM | 8 | 19min. | \\$0.17 |
c4.2xlarge |\n| BAM | 58.61 GB | 845.6M | 150 | BAM | 8 |
61min. | \\$0.55 | c4.2xlarge |\n| BAM | 5.26 GB | 71.5M |
76 | SAM | 8 | 14min. | \\$0.13 | c4.2xlarge |\n| BAM |
11.86 GB | 161.2M | 101 | SAM | 8 | 23min. | \\$0.21 |
c4.2xlarge |\n| BAM | 18.36 GB | 179M | 76 | SAM | 8 |
35min. | \\$0.31 | c4.2xlarge |\n| BAM | 58.61 GB | 845.6M
| 150 | SAM | 8 | 2h 29min. | \\$1.34 | c4.2xlarge
|\n\n###References\n\n[1] [SAMtools
documentation](http://www.htslib.org/doc/samtools-1.9.html)",
"version": "1.9",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "13",
"name": "sbg_lines_to_interval_list_abr",
"description": "This tools is used for splitting GATK
sequence grouping file into subgroups.\n\n### Common Use
```

```
Cases\n\nEach subgroup file contains intervals defined on
single line in grouping file. Grouping file is output of
GATKs **CreateSequenceGroupingTSV** script which is used in
best practice workflows sush as **GATK Best Practice
Germline Workflow**.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "14",
"name": "sbg_lines_to_interval_list_br",
"description": "This tools is used for splitting GATK
sequence grouping file into subgroups.\n\n### Common Use
Cases\n\nEach subgroup file contains intervals defined on
single line in grouping file. Grouping file is output of
GATKs **CreateSequenceGroupingTSV** script which is used in
best practice workflows sush as **GATK Best Practice
Germline Workflow**.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
}
]
}
```

1.6 Execution Domain

```
{
"keywords": [],
"xref": [],
"platform": [
"Seven Bridges Platform"
],
"pipeline_steps": [
{
"step_number": "1",
"name": "gatk_markduplicates_4_1_0_0",
"description": "The **GATK MarkDuplicates** tool identifies
duplicate reads in a BAM or SAM file.\n\nThis tool locates
and tags duplicate reads in a BAM or SAM file, where
duplicate reads are defined as originating from a single
fragment of DNA. Duplicates can arise during sample
preparation e.g. library construction using PCR. Duplicate
reads can also result from a single amplification cluster,
incorrectly detected as multiple clusters by the optical
sensor of the sequencing instrument. These duplication
artifacts are referred to as optical duplicates [1].\n\nThe
MarkDuplicates tool works by comparing sequences in the 5
prime positions of both reads and read-pairs in the SAM/BAM
file. The **Barcode tag** (`--BARCODE_TAG`) option is
available to facilitate duplicate marking using molecular
barcodes. After duplicate reads are collected, the tool
differentiates the primary and duplicate reads using an
algorithm that ranks reads by the sums of their
base-quality scores (default method).\n\n\n###Common Use
Cases\n\n* The **GATK MarkDuplicates** tool requires the
BAM or SAM file on its **Input BAM/SAM file** (`--INPUT`)
input. The tool generates a new SAM or BAM file on its
**Output BAM/SAM** output, in which duplicates have been
identified in the SAM flags field for each read. Duplicates
are marked with the hexadecimal value of 0x0400, which
corresponds to a decimal value of 1024. If you are not
```

familiar with this type of annotation, please see the following [blog post](https://software.broadinstitute.org/gatk/blog?id=7019) for additional information. **MarkDuplicates** also produces a metrics file on its **Output metrics file** output, indicating the numbers of duplicates for both single and paired end reads.\n\n* The program can take either coordinate-sorted or query-sorted inputs, however the behavior is slightly different. When the input is coordinate-sorted, unmapped mates of mapped records and supplementary/secondary alignments are not marked as duplicates. However, when the input is query-sorted (actually query-grouped), then unmapped mates and secondary/supplementary reads are not excluded from the duplication test and can be marked as duplicate reads.\n\n* If desired, duplicates can be removed using the **Remove duplicates** (`--REMOVE_DUPLICATES`) and **Remove sequencing duplicates** (`--REMOVE_SEQUENCING_DUPLICATES`) options.\n\n* Although the bitwise flag annotation indicates whether a read was marked as a duplicate, it does not identify the type of duplicate. To do this, a new tag called the duplicate type (DT) tag was recently added as an optional output of a SAM/BAM file. Invoking the **Tagging policy** (`--TAGGING_POLICY`) option, you can instruct the program to mark all the duplicates (All), only the optical duplicates (OpticalOnly), or no duplicates (DontTag). The records within the output SAM/BAM file will have values for the 'DT' tag (depending on the invoked **TAGGING_POLICY** option), as either library/PCR-generated duplicates (LB), or sequencing-platform artifact duplicates (SQ). \n\n* This tool uses the **Read name regex** (`--READ_NAME_REGEX`) and the **Optical duplicate pixel distance**

```
(`--OPTICAL_DUPLICATE_PIXEL_DISTANCE`) options as the
primary methods to identify and differentiate duplicate
types. Set **READ_NAME_REGEX** to null to skip optical
duplicate detection, e.g. for RNA-seq or other data where
duplicate sets are extremely large and estimating library
complexity is not an aim. Note that without optical
duplicate counts, library size estimation will be
inaccurate.\n\n* Usage example:\n\n```\ngatk MarkDuplicates
\\\n --INPUT input.bam \\\n --OUTPUT marked_duplicates.bam
\\\n --METRICS_FILE
marked_dup_metrics.txt\n```\n\n###Changes Introduced by
Seven Bridges\n\n* All output files will be prefixed using
the **Output prefix** parameter. In case **Output prefix**
is not provided, output prefix will be the same as the
Sample ID metadata from the **Input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, output prefix will be
inferred from the **Input SAM/BAM** filename. This way,
having identical names of the output files between runs is
avoided. Moreover, **dedupped** will be added before the
extension of the output file name. \n\n* The user has a
possibility to specify the output file format using the
**Output file format** option. Otherwise, the output file
format will be the same as the format of the input
file.\n\n###Common Issues and Important Notes\n\n*
None\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK
MarkDuplicates** for a couple of different samples,
executed on the AWS cloud instances:\n\n| Experiment type |
Input size | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 1.8 GB | 3min | ~0.02$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 5.3 GB | 9min | ~0.06$ | c4.2xlarge (8 CPUs)
```

```
| \n| RNA-Seq | 8.8 GB | 16min | ~0.11$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 17 GB | 30min | ~0.20$ | c4.2xlarge
(8 CPUs) |\n\n*Cost can be significantly reduced by using
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
MarkDuplicates](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "2",
"name": "bwa_mem_bundle_0_7_15",
"description": "BWA-MEM is an algorithm designed for
aligning sequence reads onto a large reference genome.
BWA-MEM is implemented as a component of BWA. The algorithm
can automatically choose between performing end-to-end and
local alignments. BWA-MEM is capable of outputting multiple
alignments, and finding chimeric reads. It can be applied
to a wide range of read lengths, from 70 bp to several
megabases. \n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n\n## Common Use Cases\nIn order to obtain
possibilities for additional fast processing of aligned
reads, **Biobambam2 sortmadup** (2.0.87) tool is embedded
together into the same package with BWA-MEM (0.7.15).\n\nIn
order to obtain possibilities for additional fast
processing of aligned reads, **Biobambam2** (2.0.87) is
embedded together with the BWA 0.7.15 toolkit into the
**BWA-MEM Bundle 0.7.15 CWL1.0**.  Two tools are used
```

(**bamsort** and **bamsormadup**) to allow the selection of three output formats (SAM, BAM, or CRAM), different modes of sorting (Quarryname/Coordinate sorting), and Marking/Removing duplicates that can arise during sample preparation e.g. library construction using PCR. This is done by setting the **Output format** and **PCR duplicate detection** parameters.\n- Additional notes:\n - The default **Output format** is coordinate sorted BAM (option **BAM**).\n - SAM and BAM options are query name sorted, while CRAM format is not advisable for data sorted by query name.\n - Coordinate Sorted BAM file in all options and CRAM Coordinate sorted output with Marked Duplicates come with the accompanying index file. The generated index name will be the same as the output alignments file, with the extension BAM.BAI or CRAM.CRAI. However, when selecting the CRAM Coordinate sorted and CRAM Coordinate sorted output with Removed Duplicates, the generated files will not have the index file generated. This is a result of the usage of different Biobambam2 tools - **bamsort** does not have the ability to write CRAI files (only supports outputting BAI index files), while **bamsormadup** can write CRAI files.\n - Passing data from BWA-MEM to Biobambam2 tools has been done through the Linux piping which saves processing times (up to an hour of the execution time for whole-genome sample) of reading and writing of aligned reads into the hard drive. \n - **BWA-MEM Bundle 0.7.15 CWL1** first needs to construct the FM-index (Full-text index in Minute space) for the reference genome using the **BWA INDEX 0.7.17 CWL1.0** tool. The two BWA versions are compatible.\n\n### Changes Introduced by Seven Bridges\n\n- **Aligned SAM/BAM/CRAM** file will be prefixed using the **Output SAM/BAM/CRAM file name** parameter. In case **Output

```
SAM/BAM/CRAM file name** is not provided, the output prefix
will be the same as the **Sample ID** metadata field from
the file if the **Sample ID** metadata field exists.
Otherwise, the output prefix will be inferred from the
**Input reads** file names.\n- The **Platform** metadata
field for the output alignments will be automatically set
to \"Illumina\" unless it is present in **Input reads**
metadata, or given through **Read group header** or
**Platform** input parameters. This will prevent possible
errors in downstream analysis using the GATK toolkit.\n- If
the **Read group ID** parameter is not defined, by default
it will be set to ‘1’. If the tool is scattered within a
workflow it will assign the **Read Group ID** according to
the order of the scattered folders. This ensures a unique
**Read Group ID** when processing multi-read group input
data from one sample.\n\n### Common Issues and Important
Notes \n \n- For input reads FASTQ files of total size less
than 10 GB we suggest using the default setting for
parameter **Total memory** of 15GB, for larger files we
suggest using 58 GB of memory and 32 CPU cores.\n- When the
desired output is a CRAM file without deduplication of the
PCR duplicates, it is necessary to provide the FASTA Index
file (FAI) as input.\n- Human reference genome version 38
comes with ALT contigs, a collection of diverged alleles
present in some humans but not the others. Making effective
use of these contigs will help to reduce mapping artifacts,
however, to facilitate mapping these ALT contigs to the
primary assembly, GRC decided to add to each contig long
flanking sequences almost identical to the primary
assembly. As a result, a naive mapping against GRCh38+ALT
will lead to many mapQ-zero mappings in these flanking
regions. Please use post-processing steps to fix these
```

```
alignments or implement
[steps](https://sourceforge.net/p/bio-bwa/mailman/message/32845712/)
described by the author of the BWA toolkit.  \n- Inputs
**Read group header** and **Insert string to header** need
to be given in the correct format - under single-quotes.\n-
BWA-MEM is not a splice aware aligner, so it is not the
appropriate tool for mapping RNAseq to the genome. For
RNAseq reads **Bowtie2 Aligner** and **STAR** are
recommended tools. \n- Input paired reads need to have the
identical read names - if not, the tool will throw a
``[mem_sam_pe] paired reads have different names``
error.\n- This wrapper was tested and is fully compatible
with cwltool v3.0.\n\n### Performance Benchmarking\n\nBelow
is a table describing the runtimes and task costs on
on-demand instances for a set of samples with different
file sizes :\n\n| Input reads | Size [GB] | Output format |
Instance (AWS) | Duration | Cost | Threads
|\n|-------------------|-----------|---------------|--------------------------|-----------|---
HG001-NA12878-30x | 2 x 23.8 | SAM | c5.9xlarge (36CPU,
72GB) | 5h 12min | $7.82 | 36 |\n| HG001-NA12878-30x | 2 x
23.8 | BAM | c5.9xlarge (36CPU, 72GB) | 5h 16min | $8.06 |
36 |\n| HG002-NA24385-50x | 2 x 66.4 | SAM | c5.9xlarge
(36CPU, 72GB) | 8h 33min | $13.08 | 36 |\n\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*",
"version": "0.7.15",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "3",
"name": "gatk_mergebamalignment_4_1_0_0",
"description": "The **GATK MergeBamAlignment** tool is used
for merging BAM/SAM alignment info from a third-party
aligner with the data in an unmapped BAM file, producing a
third BAM file that has alignment data (from the aligner)
and all the remaining data from the unmapped BAM.\n\nMany
alignment tools still require FASTQ format input. The
unmapped BAM may contain useful information that will be
lost in the conversion to FASTQ (meta-data like sample
alias, library, barcodes, etc... as well as read-level
tags.) This tool takes an unaligned BAM with meta-data, and
the aligned BAM produced by calling
[SamToFastq](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_sam
and then passing the result to an aligner. It produces a
new SAM file that includes all aligned and unaligned reads
and also carries forward additional read attributes from
the unmapped BAM (attributes that are otherwise lost in the
process of converting to FASTQ). The resulting file will be
valid for use by Picard and GATK tools. The output may be
coordinate-sorted, in which case the tags, NM, MD, and UQ
will be calculated and populated, or query-name sorted, in
which case the tags will not be calculated or populated
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n###Common Use Cases\n\n* The **GATK
MergeBamAlignment** tool requires a SAM or BAM file on its
**Aligned BAM/SAM file** (`--ALIGNED_BAM`) input, original
SAM or BAM file of unmapped reads, which must be in
queryname order on its **Unmapped BAM/SAM file**
(`--UNMAPPED_BAM`) input and a reference sequence on its
```

```
**Reference** (`--REFERENCE_SEQUENCE`) input. The tool
generates a single BAM/SAM file on its **Output merged
BAM/SAM file** output.\n\n* Usage example:\n\n```\ngatk
MergeBamAlignment \\\\\n --ALIGNED_BAM aligned.bam \\\\\n
--UNMAPPED_BAM unmapped.bam \\\\\n --OUTPUT merged.bam
\\\\\n --REFERENCE_SEQUENCE
reference_sequence.fasta\n```\n\n###Changes Introduced by
Seven Bridges\n\n* The output file name will be prefixed
using the **Output prefix** parameter. In case **Output
prefix** is not provided, output prefix will be the same as
the Sample ID metadata from **Input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, output prefix will be
inferred from the **Input SAM/BAM file** filename. This
way, having identical names of the output files between
runs is avoided. Moreover, **merged** will be added before
the extension of the output file name. \n\n* The user has a
possibility to specify the output file format using the
**Output file format** argument. Otherwise, the output file
format will be the same as the format of the input aligned
file.\n\n###Common Issues and Important Notes\n\n* Note:
This is not a tool for taking multiple BAM/SAM files and
creating a bigger file by merging them. For that use-case,
see
[MergeSamFiles](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_
Benchmarking\n\nBelow is a table describing runtimes and
task costs of **GATK MergeBamAlignment** for a couple of
different samples, executed on the AWS cloud
instances:\n\n| Experiment type | Aligned BAM/SAM size |
Unmapped BAM/SAM size | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
RNA-Seq | 1.4 GB | 1.9 GB | 9min | ~0.06$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 4.0 GB | 5.7 GB | 20min | ~0.13$ |
```

c4.2xlarge (8 CPUs) | \n| RNA-Seq | 6.6 GB | 9.5 GB | 32min
| ~0.21$ | c4.2xlarge (8 CPUs) | \n| RNA-Seq | 13 GB | 19
GB | 1h 4min | ~0.42$ | c4.2xlarge (8 CPUs) |\n\n*Cost can
be significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
MergeBamAlignment](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/pica
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "4",
"name": "gatk_samtofastq_4_1_0_0",
"description": "The **GATK SamToFastq** tool converts a SAM
or BAM file to FASTQ.\n\nThis tool extracts read sequences
and qualities from the input SAM/BAM file and writes them
into the output file in Sanger FASTQ format.\n\nIn the RC
mode (default is True), if the read is aligned and the
alignment is to the reverse strand on the genome, the read
sequence from input SAM file will be reverse-complemented
prior to writing it to FASTQ in order to correctly restore
the original read sequence as it was generated by the
sequencer [1].\n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n###Common Use Cases\n\n* The **GATK
SamToFastq** tool requires a BAM/SAM file on its **Input
BAM/SAM file** (`--INPUT`) input. The tool generates a
single-end FASTQ file on its **Output FASTQ file(s)**
output if the input BAM/SAM file is single end. In case the

```
input file is paired end, the tool outputs the first end of
the pair FASTQ and the second end of the pair FASTQ on its
**Output FASTQ file(s)** output, except when the
**Interleave** (`--INTERLEAVE`) option is set to True. If
the output is an interleaved FASTQ file, if paired, each
line will have /1 or /2 to describe which end it came
from.\n\n* The **GATK SamToFastq** tool supports an
optional parameter **Output by readgroup**
(`--OUTPUT_BY_READGROUP`) which, when true, outputs a FASTQ
file per read group (two FASTQ files per read group if the
group is paired).\n\n* Usage example (input BAM file is
single-end):\n\n```\ngatk SamToFastq \n --INPUT input.bam\n
--FASTQ output.fastq\n```\n\n\n\n\n\n* Usage example (input
BAM file is paired-end):\n\n```\ngatk SamToFastq \n --INPUT
input.bam\n --FASTQ output.pe_1.fastq\n --SECOND_END_FASTQ
output.pe_2.fastq\n --UNPAIRED_FASTQ
unpaired.fastq\n\n```\n\n###Changes Introduced by Seven
Bridges\n\n* The GATK SamToFastq tool is implemented to
check if the input alignments file contains single-end or
paired-end data and according to that generates different
command lines for these two modes and thus produces
appropriate output files on its **Output FASTQ file(s)**
output (one FASTQ file in single-end mode and two FASTQ
files if the input alignment file contains paired-end
data). \n\n* All output files will be prefixed using the
**Output prefix** parameter. In case the **Output prefix**
is not provided, the output prefix will be the same as the
Sample ID metadata from the **input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, the output prefix
will be inferred from the **Input SAM/BAM** filename. This
way, having identical names of the output files between
runs is avoided.\n\n* For paired-end read files, in order
```

```
to successfully run alignment with STAR, this tool adds the
appropriate **paired-end** metadata field in the output
FASTQ files.\n\n###Common Issues and Important Notes\n\n*
None\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK SamToFastq**
for a couple of different samples, executed on the AWS
cloud instances:\n\n| Experiment type | Input size |
Paired-end | # of reads | Read length | Duration | Cost |
Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
RNA-Seq | 1.9 GB | Yes | 16M | 101 | 4min | ~0.03$ |
c4.2xlarge (8 CPUs) | \n| RNA-Seq | 5.7 GB | Yes | 50M |
101 | 7min | ~0.04$ | c4.2xlarge (8 CPUs) | \n| RNA-Seq |
9.5 GB | Yes | 82M | 101 | 10min | ~0.07$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 19 GB | Yes | 164M | 101 | 20min |
~0.13$ | c4.2xlarge (8 CPUs) |\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n###References\n\n[1] [GATK
SamToFastq](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.12.0/picard_sa
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "5",
"name": "gatk_sortsam_4_1_0_0",
"description": "The **GATK SortSam** tool sorts the input
SAM or BAM file by coordinate, queryname (QNAME), or some
other property of the SAM record.\n\nThe **GATK SortOrder**
```

```
of a SAM/BAM file is found in the SAM file header tag @HD
in the field labeled SO.  For a coordinate\nsorted SAM/BAM
file, read alignments are sorted first by the reference
sequence name (RNAME) field using the reference\nsequence
dictionary (@SQ tag).  Alignments within these subgroups
are secondarily sorted using the left-most
mapping\nposition of the read (POS).  Subsequent to this
sorting scheme, alignments are listed
arbitrarily.<\/p><p>For\nqueryname-sorted alignments, all
alignments are grouped using the queryname field but the
alignments are not necessarily\nsorted within these groups.
Reads having the same queryname are derived from the same
template\n\n\n###Common Use Cases\n\nThe **GATK SortSam**
tool requires a BAM/SAM file on its **Input SAM/BAM file**
(`--INPUT`) input. The tool sorts input file in the order
defined by (`--SORT_ORDER`) parameter. Available sort order
options are `queryname`, `coordinate` and `duplicate`.
\n\n* Usage example:\n\n```\njava -jar picard.jar SortSam\n
--INPUT=input.bam \n
--SORT_ORDER=coordinate\n```\n\n\n###Changes Introduced by
Seven Bridges\n\n* Prefix of the output file is defined
with the optional parameter **Output prefix**. If **Output
prefix** is not provided, name of the sorted file is
obtained from **Sample ID** metadata from the **Input
SAM/BAM file**, if the **Sample ID** metadata exists.
Otherwise, the output prefix will be inferred form the
**Input SAM/BAM file** filename. \n\n\n###Common Issues and
Important Notes\n\n* None\n\n\n###Performance
Benchmarking\nBelow is a table describing runtimes and task
costs of **GATK SortSam** for a couple of different
samples, executed on the AWS cloud instances:\n\n|
Experiment type | Input size | Paired-end | # of reads |
```

Read length | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
WGS | | Yes | 16M | 101 | 4min | ~0.03$ | c4.2xlarge (8
CPUs) | \n| WGS | | Yes | 50M | 101 | 7min | ~0.04$ |
c4.2xlarge (8 CPUs) | \n| WGS | | Yes | 82M | 101 | 10min |
~0.07$ | c4.2xlarge (8 CPUs) | \n| WES | | Yes | 164M | 101
| 20min | ~0.13$ | c4.2xlarge (8 CPUs) |\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n\n###References\n[1] [GATK SortSam
home
page](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.12.0/picard_sam_SortS
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "6",
"name": "gatk_setnmmdanduqtags_4_1_0_0",
"description": "The **GATK SetNmMdAndUqTags** tool takes in
a coordinate-sorted SAM or BAM and calculatesthe NM, MD,
and UQ tags by comparing it with the reference. \n\nThe
GATK SetNmMdAndUqTags may be needed when **GATK
MergeBamAlignment** was run with **SORT_ORDER** other than
`coordinate` and thus could not fix these tags.
\n\n\n###Common Use Cases\nThe **GATK SetNmMdAndUqTags**
tool fixes NM, MD and UQ tags in SAM/BAM file **Input
SAM/BAM file** (`--INPUT`) input. This tool takes in a
coordinate-sorted SAM or BAM file and calculates the NM,
MD, and UQ tags by comparing with the reference **Reference

```
sequence** (`--REFERENCE_SEQUENCE`).\n\n* Usage
example:\n\n```\njava -jar picard.jar SetNmMdAndUqTags\n
--REFERENCE_SEQUENCE=reference_sequence.fasta\n
--INPUT=sorted.bam\n```\n\n\n###Changes Introduced by Seven
Bridges\n\n* Prefix of the output file is defined with the
optional parameter **Output prefix**. If **Output prefix**
is not provided, name of the sorted file is obtained from
**Sample ID** metadata form the **Input SAM/BAM file**, if
the **Sample ID** metadata exists. Otherwise, the output
prefix will be inferred form the **Input SAM/BAM file**
filename. \n\n\n\n###Common Issues and Important Notes\n\n*
The **Input SAM/BAM file** must be coordinate sorted in
order to run **GATK SetNmMdAndUqTags**. \n* If specified,
the MD and NM tags can be ignored and only the UQ tag be
set. \n\n\n###References\n[1] [GATK SetNmMdAndUqTags home
page](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard_sam_SetNmI
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "7",
"name": "gatk_baserecalibrator_4_1_0_0",
"description": "**GATK BaseRecalibrator** generates a
recalibration table based on various covariates for input
mapped read data [1]. It performs the first pass of the
Base Quality Score Recalibration (BQSR) by assessing base
quality scores of the input data.\n\n*A list of **all
inputs and parameters** with corresponding descriptions can
be found at the bottom of the page.*\n\n###Common Use
Cases\n\n* The **GATK BaseRecalibrator** tool requires the
```

```
input mapped read data whose quality scores need to be
assessed on its **Input alignments** (`--input`) input, the
database of known polymorphic sites to skip over on its
**Known sites** (`--known-sites`) input and a reference
file on its **Reference** (`--reference`) input. On its
**Output recalibration report** output, the tool generates
a GATK report file with many tables: the list of arguments,
the quantized qualities table, the recalibration table by
read group, the recalibration table by quality score,\nthe
recalibration table for all the optional covariates
[1].\n\n* Usage example:\n\n```\ngatk --java-options
\"-Xmx2048M\" BaseRecalibrator \\\n --input my_reads.bam
\\\n --reference reference.fasta \\\n --known-sites
sites_of_variation.vcf \\\n --known-sites
another/optional/setOfSitesToMask.vcf \\\n --output
recal_data.table\n\n```\n\n###Changes Introduced by Seven
Bridges\n\n* The output file will be prefixed using the
**Output name prefix** parameter. If this value is not set,
the output name will be generated based on the **Sample
ID** metadata value from the input alignment file. If the
**Sample ID** value is not set, the name will be inherited
from the input alignment file name. In case there are
multiple files on the **Input alignments** input, the files
will be sorted by name and output file name will be
generated based on the first file in the sorted file list,
following the rules defined in the previous case. Moreover,
**recal_data** will be added before the extension of the
output file name which is **CSV** by default.\n\n*
**Include intervals** (`--intervals`) option is divided
into **Include intervals string** and **Include intervals
file** options.\n\n* **Exclude intervals**
(`--exclude-intervals`) option is divided into **Exclude
```

```
intervals string** and **Exclude intervals file**
options.\n\n* The following GATK parameters were excluded
from the tool wrapper: `--add-output-sam-program-record`,
`--add-output-vcf-command-line`, `--arguments_file`,
`--cloud-index-prefetch-buffer`, `--cloud-prefetch-buffer`,
`--create-output-bam-index`, `--create-output-bam-md5`,
`--create-output-variant-index`,
`--create-output-variant-md5`, `--gatk-config-file`,
`--gcs-max-retries`, `--gcs-project-for-requester-pays`,
`--help`, `--lenient`, `--QUIET`,
`--sites-only-vcf-output`, `--showHidden`, `--tmp-dir`,
`--use-jdk-deflater`, `--use-jdk-inflater`, `--verbosity`,
`--version`\n\n\n\n###Common Issues and Important
Notes\n\n* **Memory per job** (`mem_per_job`) input allows
a user to set the desired memory requirement when running a
tool or adding it to a workflow. This input should be
defined in MB. It is propagated to the Memory requirements
part and "-Xmx" parameter of the tool. The default value is
2048MB.\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n* Note: GATK tools
that take in mapped read data expect a BAM file as the
primary format [2]. More on GATK requirements for mapped
sequence data formats can be found
[here](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mappe
Note: **Known sites**, **Input alignments** should have
corresponding index files in the same folder. \n* Note:
**Reference** FASTA file should have corresponding .fai
```

```
(FASTA index) and .dict (FASTA dictionary) files in the
same folder. \n* Note: These **Read Filters**
(`--read-filter`) are automatically applied to the data by
the Engine before processing by **BaseRecalibrator** [1]:
**NotSecondaryAlignmentReadFilter**,
**PassesVendorQualityCheckReadFilter**,
**MappedReadFilter**,
**MappingQualityAvailableReadFilter**,
**NotDuplicateReadFilter**,
**MappingQualityNotZeroReadFilter**,
**WellformedReadFilter**\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"LibraryReadFilter\",
the **Library** (`--library`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"PlatformReadFilter\", the **Platform
filter name** (`--platform-filter-name`) option must be set
to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set
to\"PlatformUnitReadFilter\", the **Black listed lanes**
(`--black-listed-lanes`) option must be set to some value.
\n* Note: If the **Read filter** (`--read-filter`) option
is set to \"ReadGroupBlackListReadFilter\", the **Read
group black list** (`--read-group-black-list`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"ReadGroupReadFilter\",
the **Keep read group** (`--keep-read-group`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to
\"ReadLengthReadFilter\", the **Max read length**
(`--max-read-length`) option must be set to some value.\n*
Note: If the **Read filter** (`--read-filter`) option is
set to \"ReadNameReadFilter\", the **Read name**
```

```
(`--read-name`) option must be set to some value.\n* Note:
If the **Read filter** (`--read-filter`) option is set to
\"ReadStrandFilter\", the **Keep reverse strand only**
(`--keep-reverse-strand-only`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"SampleReadFilter\", the **Sample**
(`--sample`) option must be set to some value.\n* Note: The
following options are valid only if the appropriate **Read
filter** (`--read-filter`) is specified: **Ambig filter
bases** (`--ambig-filter-bases`), **Ambig filter frac**
(`--ambig-filter-frac`), **Max fragment length**
(`--max-fragment-length`), **Maximum mapping quality**
(`--maximum-mapping-quality`), **Minimum mapping quality**
(`--minimum-mapping-quality`), **Do not require soft
clips** (`--dont-require-soft-clips-both-ends`), **Filter
too short** (`--filter-too-short`), **Min read length**
(`--min-read-length`). See the description of each
parameter for information on the associated **Read
filter**.\n* Note: The wrapper has not been tested for the
SAM file type on the **Input alignments** input port, nor
for the BCF file type on the **Known sites** input
port.\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK
BaseRecalibrator** for a couple of different samples,
executed on AWS cloud instances:\n\n| Experiment type |
Input size | Duration | Cost (on-demand) | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 2.2 GB | 9min | ~0.08$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 6.6 GB | 19min | ~0.17$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 11 GB | 27min | ~0.24$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 22 GB | 46min | ~0.41$ | c4.2xlarge
(8 CPUs) |\n\n*Cost can be significantly reduced by using
```

```
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
BaseRecalibrator](https://gatk.broadinstitute.org/hc/en-us/articles/360036726891-BaseRecalibra
[GATK Mapped sequence data
formats](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Map
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "8",
"name": "gatk_createsequencegroupingtsv_4_1_0_0",
"description": "**CreateSequenceGroupingTSV** tool generate
sets of intervals for scatter-gathering over
chromosomes.\n\nIt takes **Reference dictionary** file
(`--ref_dict`) as an input and creates files which contain
chromosome names grouped based on their
sizes.\n\n\n####**Common Use Cases**\n\nThe tool has only
one input (`--ref_dict`) which is required and has no
additional arguments. **CreateSequenceGroupingTSV** tool
results are **Sequence Grouping** file which is a text file
containing chromosome groups, and **Sequence Grouping with
Unmapped**, a text file which has the same content as
**Sequence Grouping** with additional line containing
\"unmapped\" string.\n\n\n* Usage example\n\n\n```\npython
CreateSequenceGroupingTSV.py \n --ref_dict
example_reference.dict\n\n```\n\n\n\n####**Changes
Introduced by Seven Bridges**\n\nPython code provided
within WGS Germline WDL was adjusted to be called as a
script (`CreateSequenceGroupingTSV.py`).\n\n\n###**Common
```

Issues and Important Notes**\n\nNone.\n\n\n###
Reference\n[1]
[CreateSequenceGroupingTSV](https://github.com/gatk-workflows/broad-prod-wgs-germline-snps-inde
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "9",
"name": "gatk_gatherbqsrreports_4_1_0_0",
"description": "**GATK GatherBQSRReports** gathers
scattered BQSR recalibration reports into a single file
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n\n### Common Use Cases \n\n* This tool is
intended to be used to combine recalibration tables from
runs of **GATK BaseRecalibrator** parallelized
per-interval.\n\n* Usage example:\n```\n gatk
--java-options \"-Xmx2048M\" GatherBQSRReports \\\n --input
input1.csv \\\n --input input2.csv \\\n --output
output.csv\n\n```\n\n\n###Changes Introduced by Seven
Bridges\n\n* The output file will be prefixed using the
Output name prefix parameter. If this value is not set,
the output name will be generated based on the **Sample
ID** metadata value from **Input BQSR reports**. If the
Sample ID value is not set, the name will be inherited
from the **Input BQSR reports** file name. In case there
are multiple files on the **Input BQSR reports** input, the
files will be sorted by name and output file name will be
generated based on the first file in the sorted file list,
following the rules defined in the previous case. Moreover,

.recal_data will be added before the extension of the
output file name.\n\n* The following GATK parameters were
excluded from the tool wrapper: `--arguments_file`,
`--gatk-config-file`, `--gcs-max-retries`,
`--gcs-project-for-requester-pays`, `--help`, `--QUIET`,
`--showHidden`, `--tmp-dir`, `--use-jdk-deflater`,
`--use-jdk-inflater`, `--verbosity`,
`--version`\n\n\n###Common Issues and Important Notes\n\n*
Memory per job (`mem_per_job`) input allows a user to
set the desired memory requirement when running a tool or
adding it to a workflow. This input should be defined in
MB. It is propagated to the Memory requirements part and
"-Xmx" parameter of the tool. The default value is
2048MB.\n\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n\n\n###Performance
Benchmarking\n\nThis tool is fast, with a running time of a
few minutes. The experiment task was performed on the
default AWS on-demand c4.2xlarge instance on 50 CSV files
(size of each ~350KB) and took 2 minutes to finish
($0.02).\n\n*Cost can be significantly reduced by using
spot instances. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n###References\n\n[1] [GATK
GatherBQSRReports](https://gatk.broadinstitute.org/hc/en-us/articles/360036359192-GatherBQSRRep
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],

```
"output_list": []
},
{
"step_number": "10",
"name": "gatk_applybqsr_4_1_0_0",
"description": "The **GATK ApplyBQSR** tool recalibrates
the base quality scores of an input BAM or CRAM file
containing reads.\n\nThis tool performs the second pass in
a two-stage process called Base Quality Score Recalibration
(BQSR). Specifically, it recalibrates the base qualities of
the input reads based on the recalibration table produced
by the **GATK BaseRecalibrator** tool. The goal of this
procedure is to correct systematic biases that affect the
assignment of base quality scores by the sequencer. The
first pass consists of calculating the error empirically
and finding patterns in how the error varies with basecall
features over all bases. The relevant observations are
written to the recalibration table. The second pass
consists of applying numerical corrections to each
individual basecall, based on the patterns identified in
the first step (recorded in the recalibration table), and
writing out the recalibrated data to a new BAM or CRAM file
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n###Common Use Cases\n\n* The **GATK
ApplyBQSR** tool requires a BAM or CRAM file on its **Input
alignments** (`--input`) input and the covariates table (=
recalibration file) generated by the **BaseRecalibrator**
tool on its **BQSR recal file** input
(`--bqsr-recal-file`). If the input alignments file is in
the CRAM format, the reference sequence is required on the
**Reference** (`--reference`) input of the tool. The tool
```

```
generates a new alignments file which contains recalibrated
read data on its **Output recalibrated alignments**
output.\n\n* Usage example\n\n```\n gatk --java-options
\"-Xmx2048M\" ApplyBQSR \\\n --reference reference.fasta
\\\n --input input.bam \\\n --bqsr-recal-file
recalibration.table \\\n --output output.bam\n\n```\n\n*
Original qualities can be retained in the output file under
the \"OQ\" tag if desired. See the **Emit original quals**
(`--emit-original-quals`) argument for details
[1].\n\n###Changes Introduced by Seven Bridges\n\n* The
output file will be prefixed using the **Output name
prefix** parameter. If this value is not set, the output
name will be generated based on the **Sample ID** metadata
value from the input alignments file. If the **Sample ID**
value is not set, the name will be inherited from the input
alignments file name. In case there are multiple files on
the **Input alignments** input, the files will be sorted by
name and output file name will be generated based on the
first file in the sorted file list, following the rules
defined in the previous case. Moreover, **recalibrated**
will be added before the extension of the output file
name.\n\n* The user has a possibility to specify the output
file format using the **Output file format** argument.
Otherwise, the output file format will be the same as the
format of the input file.\n\n* **Include intervals**
(`--intervals`) option is divided into **Include intervals
string** and **Include intervals file** options.\n\n*
**Exclude intervals** (`--exclude-intervals`) option is
divided into **Exclude intervals string** and **Exclude
intervals file** options.\n\n* The following GATK
parameters were excluded from the tool wrapper:
`--add-output-vcf-command-line`, `--arguments_file`,
```

```
`--cloud-index-prefetch-buffer`, `--cloud-prefetch-buffer`,
`--create-output-bam-md5`, `--create-output-variant-index`,
`--create-output-variant-md5`, `--gatk-config-file`,
`--gcs-max-retries`, `--gcs-project-for-requester-pays`,
`--help`, `--lenient`, `--QUIET`,
`--sites-only-vcf-output`, `--showHidden`, `--tmp-dir`,
`--use-jdk-deflater`, `--use-jdk-inflater`, `--verbosity`,
`--version`\n\n###Common Issues and Important Notes\n\n*
**Memory per job** (`mem_per_job`) input allows a user to
set the desired memory requirement when running a tool or
adding it to a workflow. This input should be defined in
MB. It is propagated to the Memory requirements part and
"-Xmx" parameter of the tool. The default value is
2048MB.\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n* Note: GATK tools
that take in mapped read data expect a BAM file as the
primary format [2]. More on GATK requirements for mapped
sequence data formats can be found
[here](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mappe
Note: **Input alignments** should have corresponding index
files in the same folder. \n* Note: **Reference** FASTA
file should have corresponding .fai (FASTA index) and .dict
(FASTA dictionary) files in the same folder. \n* Note: This
tool replaces the use of PrintReads for the application of
base quality score recalibration as practiced in earlier
versions of GATK (2.x and 3.x) [1].\n* Note: You should
only run **ApplyBQSR** with the covariates table created
```

```
from the input BAM or CRAM file [1].\n* Note: This **Read
Filter** (`--read-filter`) is automatically applied to the
data by the Engine before processing by **ApplyBQSR** [1]:
**WellformedReadFilter**\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"LibraryReadFilter\",
the **Library** (`--library`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"PlatformReadFilter\", the **Platform
filter name** (`--platform-filter-name`) option must be set
to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set
to\"PlatformUnitReadFilter\", the **Black listed lanes**
(`--black-listed-lanes`) option must be set to some value.
\n* Note: If the **Read filter** (`--read-filter`) option
is set to \"ReadGroupBlackListReadFilter\", the **Read
group black list** (`--read-group-black-list`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"ReadGroupReadFilter\",
the **Keep read group** (`--keep-read-group`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to
\"ReadLengthReadFilter\", the **Max read length**
(`--max-read-length`) option must be set to some value.\n*
Note: If the **Read filter** (`--read-filter`) option is
set to \"ReadNameReadFilter\", the **Read name**
(`--read-name`) option must be set to some value.\n* Note:
If the **Read filter** (`--read-filter`) option is set to
\"ReadStrandFilter\", the **Keep reverse strand only**
(`--keep-reverse-strand-only`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"SampleReadFilter\", the **Sample**
(`--sample`) option must be set to some value.\n* Note: The
```

following options are valid only if an appropriate **Read
filter** (`--read-filter`) is specified: **Ambig filter
bases** (`--ambig-filter-bases`), **Ambig filter frac**
(`--ambig-filter-frac`), **Max fragment length**
(`--max-fragment-length`), **Maximum mapping quality**
(`--maximum-mapping-quality`), **Minimum mapping quality**
(`--minimum-mapping-quality`), **Do not require soft
clips** (`--dont-require-soft-clips-both-ends`), **Filter
too short** (`--filter-too-short`), **Min read length**
(`--min-read-length`). See the description of each
parameter for information on the associated **Read
filter**.\n* Note: The wrapper has not been tested for the
SAM file type on the **Input alignments** input
port.\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK ApplyBQSR**
for a couple of different samples, executed on the AWS
cloud instances:\n\n| Experiment type | Input size |
Duration | Cost (on-demand) | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 2.2 GB | 8min | ~0.07$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 6.6 GB | 23min | ~0.21$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 11 GB | 37min | ~0.33$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 22 GB | 1h 16min | ~0.68$ |
c4.2xlarge (8 CPUs) |\n\n*Cost can be significantly reduced
by using **spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
ApplyBQSR](https://gatk.broadinstitute.org/hc/en-us/articles/360036725911-ApplyBQSR)\n\n[2]
[GATK Mapped sequence data
formats](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapp
"version": "4.1.0.0",
"prerequisite": [],

```
"input_list": [],
"output_list": []
},
{
"step_number": "11",
"name": "gatk_gatherbamfiles_4_1_0_0",
"description": "**GATK GatherBamFiles** concatenates one or
more BAM files resulted form scattered paralel anaysis.
\n\n\n### Common Use Cases \n\n* **GATK GatherBamFiles**
tool performs a rapid \"gather\" or concatenation on BAM
files into single BAM file. This is often needed in
operations that have been run in parallel across genomics
regions by scattering their execution across computing
nodes and cores thus resulting in smaller BAM files.\n*
Usage example:\n```\n\njava -jar picard.jar
GatherBamFiles\n --INPUT=input1.bam\n
--INPUT=input2.bam\n```\n\n### Common Issues and Important
Notes\n* **GATK GatherBamFiles** assumes that the list of
BAM files provided as input are in the order that they
should be concatenated and simply links the bodies of the
BAM files while retaining the header from the first file.
\n* Operates by copying the gzip blocks directly for speed
but also supports the generation of an MD5 in the output
file and the indexing of the output BAM file.\n* This tool
only support BAM files. It does not support SAM
files.\n\n###Changes Intorduced by Seven Bridges\n*
Generated output BAM file will be prefixed using the
**Output prefix** parameter. In case the **Output prefix**
is not provided, the output prefix will be the same as the
**Sample ID** metadata from the **Input alignments**, if
the **Sample ID** metadata exists. Otherwise, the output
prefix will be inferred from the **Input alignments**
```

```
filename. This way, having identical names of the output
files between runs is avoided.",
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "12",
"name": "samtools_view_1_9_cwl1_0",
"description": "**SAMtools View** tool prints all
alignments from a SAM, BAM, or CRAM file to an output file
in SAM format (headerless). You may specify one or more
space-separated region specifications to restrict output to
only those alignments which overlap the specified
region(s). Use of region specifications requires a
coordinate-sorted and indexed input file (in BAM or CRAM
format) [1].\n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n####Regions\n\nRegions can be specified
as: RNAME[:STARTPOS[-ENDPOS]] and all position coordinates
are 1-based. \n\n**Important note:** when multiple regions
are given, some alignments may be output multiple times if
they overlap more than one of the specified
regions.\n\nExamples of region specifications:\n\n-
**chr1** - Output all alignments mapped to the reference
sequence named `chr1' (i.e. @SQ SN:chr1).\n\n-
**chr2:1000000** - The region on chr2 beginning at base
position 1,000,000 and ending at the end of the
chromosome.\n\n- **chr3:1000-2000** - The 1001bp region on
chr3 beginning at base position 1,000 and ending at base
position 2,000 (including both end positions).\n\n-
```

```
**'\\*'** - Output the unmapped reads at the end of the
file. (This does not include any unmapped reads placed on a
reference sequence alongside their mapped mates.)\n\n-
**.** - Output all alignments. (Mostly unnecessary as not
specifying a region at all has the same effect.)
[1]\n\n###Common Use Cases\n\nThis tool can be used for:
\n\n- Filtering BAM/SAM/CRAM files - options set by the
following parameters and input files: **Include reads with
all of these flags** (`-f`), **Exclude reads with any of
these flags** (`-F`), **Exclude reads with all of these
flags** (`-G`), **Read group** (`-r`), **Minimum mapping
quality** (`-q`), **Only include alignments in library**
(`-l`), **Minimum number of CIGAR bases consuming query
sequence** (`-m`), **Subsample fraction** (`-s`), **Read
group list** (`-R`), **BED region file** (`-L`)\n- Format
conversion between SAM/BAM/CRAM formats - set by the
following parameters: **Output format**
(`--output-fmt/-O`), **Fast bam compression** (`-1`),
**Output uncompressed BAM** (`-u`)\n- Modification of the
data which is contained in each alignment - set by the
following parameters: **Collapse the backward CIGAR
operation** (`-B`), **Read tags to strip** (`-x`)\n-
Counting number of alignments in SAM/BAM/CRAM file - set by
parameter **Output only count of matching records**
(`-c`)\n\n###Changes Introduced by Seven Bridges\n\n-
Parameters **Output BAM** (`-b`) and **Output CRAM** (`-C`)
were excluded from the wrapper since they are redundant
with parameter **Output format** (`--output-fmt/-O`).\n-
Parameter **Input format** (`-S`) was excluded from wrapper
since it is ignored by the tool (input format is
auto-detected).\n- Input file **Index file** was added to
the wrapper to enable operations that require an index file
```

```
for BAM/CRAM files.\n- Parameter **Number of threads**
(`--threads/-@`) specifies the total number of threads
instead of additional threads. Command line argument
(`--threads/-@`) will be reduced by 1 to set the number of
additional threads.\n\n###Common Issues and Important
Notes\n\n- When multiple regions are given, some alignments
may be output multiple times if they overlap more than one
of the specified regions [1].\n- Use of region
specifications requires a coordinate-sorted and indexed
input file (in BAM or CRAM format) [1].\n- Option **Output
uncompressed BAM** (`-u`) saves time spent on
compression/decompression and is thus preferred when the
output is piped to another SAMtools command
[1].\n\n###Performance Benchmarking\n\nMultithreading can
be enabled by setting parameter **Number of threads**
(`--threads/-@`). In the following table you can find
estimates of **SAMtools View** running time and cost.
\n\n*Cost can be significantly reduced by using **spot
instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.* \n\n| Input type | Input size | # of
reads | Read length | Output format | # of threads |
Duration | Cost | Instance
(AWS)|\n|---------------|-------------|----------------|---------------|-----------------|-
BAM | 5.26 GB | 71.5M | 76 | BAM | 1 | 13min. | \\$0.12 |
c4.2xlarge |\n| BAM | 11.86 GB | 161.2M | 101 | BAM | 1 |
33min. | \\$0.30 | c4.2xlarge |\n| BAM | 18.36 GB | 179M |
76 | BAM | 1 | 60min. | \\$0.54 | c4.2xlarge |\n| BAM |
58.61 GB | 845.6M | 150 | BAM | 1 | 3h 25min. | \\$1.84 |
c4.2xlarge |\n| BAM | 5.26 GB | 71.5M | 76 | BAM | 8 |
5min. | \\$0.04 | c4.2xlarge |\n| BAM | 11.86 GB | 161.2M |
101 | BAM | 8 | 11min. | \\$0.10 | c4.2xlarge |\n| BAM |
18.36 GB | 179M | 76 | BAM | 8 | 19min. | \\$0.17 |
c4.2xlarge |\n| BAM | 58.61 GB | 845.6M | 150 | BAM | 8 |
61min. | \\$0.55 | c4.2xlarge |\n| BAM | 5.26 GB | 71.5M |
76 | SAM | 8 | 14min. | \\$0.13 | c4.2xlarge |\n| BAM |
11.86 GB | 161.2M | 101 | SAM | 8 | 23min. | \\$0.21 |
c4.2xlarge |\n| BAM | 18.36 GB | 179M | 76 | SAM | 8 |
35min. | \\$0.31 | c4.2xlarge |\n| BAM | 58.61 GB | 845.6M
| 150 | SAM | 8 | 2h 29min. | \\$1.34 | c4.2xlarge
|\n\n###References\n\n[1] [SAMtools
documentation](http://www.htslib.org/doc/samtools-1.9.html)",
"version": "1.9",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "13",
"name": "sbg_lines_to_interval_list_abr",
"description": "This tools is used for splitting GATK
sequence grouping file into subgroups.\n\n### Common Use
Cases\n\nEach subgroup file contains intervals defined on
single line in grouping file. Grouping file is output of
GATKs **CreateSequenceGroupingTSV** script which is used in
best practice workflows sush as **GATK Best Practice
Germline Workflow**.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "14",
```

"name": "sbg_lines_to_interval_list_br",
"description": "This tools is used for splitting GATK sequence grouping file into subgroups.\n\n### Common Use Cases\n\nEach subgroup file contains intervals defined on single line in grouping file. Grouping file is output of GATKs **CreateSequenceGroupingTSV** script which is used in best practice workflows sush as **GATK Best Practice Germline Workflow**.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
}
]
}

1.7 Parametric Domain

{
"keywords": [],
"xref": [],
"platform": [
"Seven Bridges Platform"
],
"pipeline_steps": [
{
"step_number": "1",
"name": "gatk_markduplicates_4_1_0_0",
"description": "The **GATK MarkDuplicates** tool identifies duplicate reads in a BAM or SAM file.\n\nThis tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample

```
preparation e.g. library construction using PCR. Duplicate
reads can also result from a single amplification cluster,
incorrectly detected as multiple clusters by the optical
sensor of the sequencing instrument. These duplication
artifacts are referred to as optical duplicates [1].\n\nThe
MarkDuplicates tool works by comparing sequences in the 5
prime positions of both reads and read-pairs in the SAM/BAM
file. The **Barcode tag** (`--BARCODE_TAG`) option is
available to facilitate duplicate marking using molecular
barcodes. After duplicate reads are collected, the tool
differentiates the primary and duplicate reads using an
algorithm that ranks reads by the sums of their
base-quality scores (default method).\n\n\n###Common Use
Cases\n\n* The **GATK MarkDuplicates** tool requires the
BAM or SAM file on its **Input BAM/SAM file** (`--INPUT`)
input. The tool generates a new SAM or BAM file on its
**Output BAM/SAM** output, in which duplicates have been
identified in the SAM flags field for each read. Duplicates
are marked with the hexadecimal value of 0x0400, which
corresponds to a decimal value of 1024. If you are not
familiar with this type of annotation, please see the
following [blog
post](https://software.broadinstitute.org/gatk/blog?id=7019)
for additional information. **MarkDuplicates** also
produces a metrics file on its **Output metrics file**
output, indicating the numbers of duplicates for both
single and paired end reads.\n\n* The program can take
either coordinate-sorted or query-sorted inputs, however
the behavior is slightly different. When the input is
coordinate-sorted, unmapped mates of mapped records and
supplementary/secondary alignments are not marked as
duplicates. However, when the input is query-sorted
```

```
(actually query-grouped), then unmapped mates and
secondary/supplementary reads are not excluded from the
duplication test and can be marked as duplicate reads.\n\n*
If desired, duplicates can be removed using the **Remove
duplicates** (`--REMOVE_DUPLICATES`) and **Remove
sequencing duplicates** ( `--REMOVE_SEQUENCING_DUPLICATES`)
options.\n\n* Although the bitwise flag annotation
indicates whether a read was marked as a duplicate, it does
not identify the type of duplicate. To do this, a new tag
called the duplicate type (DT) tag was recently added as an
optional output of a SAM/BAM file. Invoking the **Tagging
policy** ( `--TAGGING_POLICY`) option, you can instruct the
program to mark all the duplicates (All), only the optical
duplicates (OpticalOnly), or no duplicates (DontTag). The
records within the output SAM/BAM file will have values for
the 'DT' tag (depending on the invoked **TAGGING_POLICY**
option), as either library/PCR-generated duplicates (LB),
or sequencing-platform artifact duplicates (SQ). \n\n* This
tool uses the **Read name regex** (`--READ_NAME_REGEX`) and
the **Optical duplicate pixel distance**
(`--OPTICAL_DUPLICATE_PIXEL_DISTANCE`) options as the
primary methods to identify and differentiate duplicate
types. Set **READ_NAME_REGEX** to null to skip optical
duplicate detection, e.g. for RNA-seq or other data where
duplicate sets are extremely large and estimating library
complexity is not an aim. Note that without optical
duplicate counts, library size estimation will be
inaccurate.\n\n* Usage example:\n\n```\ngatk MarkDuplicates
\\\n --INPUT input.bam \\\n --OUTPUT marked_duplicates.bam
\\\n --METRICS_FILE
marked_dup_metrics.txt\n```\n\n###Changes Introduced by
Seven Bridges\n\n* All output files will be prefixed using
```

```
the **Output prefix** parameter. In case **Output prefix**
is not provided, output prefix will be the same as the
Sample ID metadata from the **Input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, output prefix will be
inferred from the **Input SAM/BAM** filename. This way,
having identical names of the output files between runs is
avoided. Moreover, **dedupped** will be added before the
extension of the output file name. \n\n* The user has a
possibility to specify the output file format using the
**Output file format** option. Otherwise, the output file
format will be the same as the format of the input
file.\n\n###Common Issues and Important Notes\n\n*
None\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK
MarkDuplicates** for a couple of different samples,
executed on the AWS cloud instances:\n\n| Experiment type |
Input size | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 1.8 GB | 3min | ~0.02$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 5.3 GB | 9min | ~0.06$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 8.8 GB | 16min | ~0.11$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 17 GB | 30min | ~0.20$ | c4.2xlarge
(8 CPUs) |\n\n*Cost can be significantly reduced by using
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
MarkDuplicates](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
```

{
"step_number": "2",
"name": "bwa_mem_bundle_0_7_15",
"description": "BWA-MEM is an algorithm designed for
aligning sequence reads onto a large reference genome.
BWA-MEM is implemented as a component of BWA. The algorithm
can automatically choose between performing end-to-end and
local alignments. BWA-MEM is capable of outputting multiple
alignments, and finding chimeric reads. It can be applied
to a wide range of read lengths, from 70 bp to several
megabases. \n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n\n## Common Use Cases\nIn order to obtain
possibilities for additional fast processing of aligned
reads, **Biobambam2 sortmadup** (2.0.87) tool is embedded
together into the same package with BWA-MEM (0.7.15).\n\nIn
order to obtain possibilities for additional fast
processing of aligned reads, **Biobambam2** (2.0.87) is
embedded together with the BWA 0.7.15 toolkit into the
BWA-MEM Bundle 0.7.15 CWL1.0. Two tools are used
(**bamsort** and **bamsormadup**) to allow the selection of
three output formats (SAM, BAM, or CRAM), different modes
of sorting (Quarryname/Coordinate sorting), and
Marking/Removing duplicates that can arise during sample
preparation e.g. library construction using PCR. This is
done by setting the **Output format** and **PCR duplicate
detection** parameters.\n- Additional notes:\n - The
default **Output format** is coordinate sorted BAM (option
BAM).\n - SAM and BAM options are query name sorted,
while CRAM format is not advisable for data sorted by query
name.\n - Coordinate Sorted BAM file in all options and
CRAM Coordinate sorted output with Marked Duplicates come

with the accompanying index file. The generated index name will be the same as the output alignments file, with the extension BAM.BAI or CRAM.CRAI. However, when selecting the CRAM Coordinate sorted and CRAM Coordinate sorted output with Removed Duplicates, the generated files will not have the index file generated. This is a result of the usage of different Biobambam2 tools - **bamsort** does not have the ability to write CRAI files (only supports outputting BAI index files), while **bamsormadup** can write CRAI files.\n - Passing data from BWA-MEM to Biobambam2 tools has been done through the Linux piping which saves processing times (up to an hour of the execution time for whole-genome sample) of reading and writing of aligned reads into the hard drive. \n - **BWA-MEM Bundle 0.7.15 CWL1** first needs to construct the FM-index (Full-text index in Minute space) for the reference genome using the **BWA INDEX 0.7.17 CWL1.0** tool. The two BWA versions are compatible.\n\n### Changes Introduced by Seven Bridges\n\n- **Aligned SAM/BAM/CRAM** file will be prefixed using the **Output SAM/BAM/CRAM file name** parameter. In case **Output SAM/BAM/CRAM file name** is not provided, the output prefix will be the same as the **Sample ID** metadata field from the file if the **Sample ID** metadata field exists. Otherwise, the output prefix will be inferred from the **Input reads** file names.\n- The **Platform** metadata field for the output alignments will be automatically set to \"Illumina\" unless it is present in **Input reads** metadata, or given through **Read group header** or **Platform** input parameters. This will prevent possible errors in downstream analysis using the GATK toolkit.\n- If the **Read group ID** parameter is not defined, by default it will be set to ‘1’. If the tool is scattered within a

```
workflow it will assign the **Read Group ID** according to
the order of the scattered folders. This ensures a unique
**Read Group ID** when processing multi-read group input
data from one sample.\n\n### Common Issues and Important
Notes \n \n- For input reads FASTQ files of total size less
than 10 GB we suggest using the default setting for
parameter **Total memory** of 15GB, for larger files we
suggest using 58 GB of memory and 32 CPU cores.\n- When the
desired output is a CRAM file without deduplication of the
PCR duplicates, it is necessary to provide the FASTA Index
file (FAI) as input.\n- Human reference genome version 38
comes with ALT contigs, a collection of diverged alleles
present in some humans but not the others. Making effective
use of these contigs will help to reduce mapping artifacts,
however, to facilitate mapping these ALT contigs to the
primary assembly, GRC decided to add to each contig long
flanking sequences almost identical to the primary
assembly. As a result, a naive mapping against GRCh38+ALT
will lead to many mapQ-zero mappings in these flanking
regions. Please use post-processing steps to fix these
alignments or implement
[steps](https://sourceforge.net/p/bio-bwa/mailman/message/32845712/)
described by the author of the BWA toolkit.  \n- Inputs
**Read group header** and **Insert string to header** need
to be given in the correct format - under single-quotes.\n-
BWA-MEM is not a splice aware aligner, so it is not the
appropriate tool for mapping RNAseq to the genome. For
RNAseq reads **Bowtie2 Aligner** and **STAR** are
recommended tools. \n- Input paired reads need to have the
identical read names - if not, the tool will throw a
``[mem_sam_pe] paired reads have different names``
error.\n- This wrapper was tested and is fully compatible
```

```
with cwltool v3.0.\n\n### Performance Benchmarking\n\nBelow
is a table describing the runtimes and task costs on
on-demand instances for a set of samples with different
file sizes :\n\n| Input reads | Size [GB] | Output format |
Instance (AWS) | Duration | Cost | Threads
|\n|-------------------|-----------|---------------|--------------------------|-----------|---
HG001-NA12878-30x | 2 x 23.8 | SAM | c5.9xlarge (36CPU,
72GB) | 5h 12min | $7.82 | 36 |\n| HG001-NA12878-30x | 2 x
23.8 | BAM | c5.9xlarge (36CPU, 72GB) | 5h 16min | $8.06 |
36 |\n| HG002-NA24385-50x | 2 x 66.4 | SAM | c5.9xlarge
(36CPU, 72GB) | 8h 33min | $13.08 | 36 |\n\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*",
"version": "0.7.15",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "3",
"name": "gatk_mergebamalignment_4_1_0_0",
"description": "The **GATK MergeBamAlignment** tool is used
for merging BAM/SAM alignment info from a third-party
aligner with the data in an unmapped BAM file, producing a
third BAM file that has alignment data (from the aligner)
and all the remaining data from the unmapped BAM.\n\nMany
alignment tools still require FASTQ format input. The
unmapped BAM may contain useful information that will be
lost in the conversion to FASTQ (meta-data like sample
alias, library, barcodes, etc... as well as read-level
```

```
tags.) This tool takes an unaligned BAM with meta-data, and
the aligned BAM produced by calling
[SamToFastq](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_sa
and then passing the result to an aligner. It produces a
new SAM file that includes all aligned and unaligned reads
and also carries forward additional read attributes from
the unmapped BAM (attributes that are otherwise lost in the
process of converting to FASTQ). The resulting file will be
valid for use by Picard and GATK tools. The output may be
coordinate-sorted, in which case the tags, NM, MD, and UQ
will be calculated and populated, or query-name sorted, in
which case the tags will not be calculated or populated
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n###Common Use Cases\n\n* The **GATK
MergeBamAlignment** tool requires a SAM or BAM file on its
**Aligned BAM/SAM file** (`--ALIGNED_BAM`) input, original
SAM or BAM file of unmapped reads, which must be in
queryname order on its **Unmapped BAM/SAM file**
(`--UNMAPPED_BAM`) input and a reference sequence on its
**Reference** (`--REFERENCE_SEQUENCE`) input. The tool
generates a single BAM/SAM file on its **Output merged
BAM/SAM file** output.\n\n* Usage example:\n\n```\ngatk
MergeBamAlignment \\\\n --ALIGNED_BAM aligned.bam \\\\n
--UNMAPPED_BAM unmapped.bam \\\\n --OUTPUT merged.bam
\\\\n --REFERENCE_SEQUENCE
reference_sequence.fasta\n```\n\n###Changes Introduced by
Seven Bridges\n\n* The output file name will be prefixed
using the **Output prefix** parameter. In case **Output
prefix** is not provided, output prefix will be the same as
the Sample ID metadata from **Input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, output prefix will be
```

```
inferred from the **Input SAM/BAM file** filename. This
way, having identical names of the output files between
runs is avoided. Moreover, **merged** will be added before
the extension of the output file name. \n\n* The user has a
possibility to specify the output file format using the
**Output file format** argument. Otherwise, the output file
format will be the same as the format of the input aligned
file.\n\n###Common Issues and Important Notes\n\n* Note:
This is not a tool for taking multiple BAM/SAM files and
creating a bigger file by merging them. For that use-case,
see
[MergeSamFiles](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/picard_
Benchmarking\n\nBelow is a table describing runtimes and
task costs of **GATK MergeBamAlignment** for a couple of
different samples, executed on the AWS cloud
instances:\n\n| Experiment type | Aligned BAM/SAM size |
Unmapped BAM/SAM size | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
RNA-Seq | 1.4 GB | 1.9 GB | 9min | ~0.06$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 4.0 GB | 5.7 GB | 20min | ~0.13$ |
c4.2xlarge (8 CPUs) | \n| RNA-Seq | 6.6 GB | 9.5 GB | 32min
| ~0.21$ | c4.2xlarge (8 CPUs) | \n| RNA-Seq | 13 GB | 19
GB | 1h 4min | ~0.42$ | c4.2xlarge (8 CPUs) |\n\n*Cost can
be significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
MergeBamAlignment](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.1.0.0/pica
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
```

```
},
{
"step_number": "4",
"name": "gatk_samtofastq_4_1_0_0",
"description": "The **GATK SamToFastq** tool converts a SAM
or BAM file to FASTQ.\n\nThis tool extracts read sequences
and qualities from the input SAM/BAM file and writes them
into the output file in Sanger FASTQ format.\n\nIn the RC
mode (default is True), if the read is aligned and the
alignment is to the reverse strand on the genome, the read
sequence from input SAM file will be reverse-complemented
prior to writing it to FASTQ in order to correctly restore
the original read sequence as it was generated by the
sequencer [1].\n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n###Common Use Cases\n\n* The **GATK
SamToFastq** tool requires a BAM/SAM file on its **Input
BAM/SAM file** (`--INPUT`) input. The tool generates a
single-end FASTQ file on its **Output FASTQ file(s)**
output if the input BAM/SAM file is single end. In case the
input file is paired end, the tool outputs the first end of
the pair FASTQ and the second end of the pair FASTQ on its
**Output FASTQ file(s)** output, except when the
**Interleave** (`--INTERLEAVE`) option is set to True. If
the output is an interleaved FASTQ file, if paired, each
line will have /1 or /2 to describe which end it came
from.\n\n* The **GATK SamToFastq** tool supports an
optional parameter **Output by readgroup**
(`--OUTPUT_BY_READGROUP`) which, when true, outputs a FASTQ
file per read group (two FASTQ files per read group if the
group is paired).\n\n* Usage example (input BAM file is
single-end):\n\n```\ngatk SamToFastq \n --INPUT input.bam\n
```

```
--FASTQ output.fastq\n```\n\n\n\n\n\n* Usage example (input
BAM file is paired-end):\n\n```\ngatk SamToFastq \n --INPUT
input.bam\n --FASTQ output.pe_1.fastq\n --SECOND_END_FASTQ
output.pe_2.fastq\n --UNPAIRED_FASTQ
unpaired.fastq\n\n```\n\n###Changes Introduced by Seven
Bridges\n\n* The GATK SamToFastq tool is implemented to
check if the input alignments file contains single-end or
paired-end data and according to that generates different
command lines for these two modes and thus produces
appropriate output files on its **Output FASTQ file(s)**
output (one FASTQ file in single-end mode and two FASTQ
files if the input alignment file contains paired-end
data). \n\n* All output files will be prefixed using the
**Output prefix** parameter. In case the **Output prefix**
is not provided, the output prefix will be the same as the
Sample ID metadata from the **input SAM/BAM file**, if the
Sample ID metadata exists. Otherwise, the output prefix
will be inferred from the **Input SAM/BAM** filename. This
way, having identical names of the output files between
runs is avoided.\n\n* For paired-end read files, in order
to successfully run alignment with STAR, this tool adds the
appropriate **paired-end** metadata field in the output
FASTQ files.\n\n###Common Issues and Important Notes\n\n*
None\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK SamToFastq**
for a couple of different samples, executed on the AWS
cloud instances:\n\n| Experiment type | Input size |
Paired-end | # of reads | Read length | Duration | Cost |
Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
RNA-Seq | 1.9 GB | Yes | 16M | 101 | 4min | ~0.03$ |
c4.2xlarge (8 CPUs) | \n| RNA-Seq | 5.7 GB | Yes | 50M |
```

101 | 7min | ~0.04$ | c4.2xlarge (8 CPUs) | \n| RNA-Seq |
9.5 GB | Yes | 82M | 101 | 10min | ~0.07$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 19 GB | Yes | 164M | 101 | 20min |
~0.13$ | c4.2xlarge (8 CPUs) |\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n###References\n\n[1] [GATK
SamToFastq](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.12.0/picard_sa
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "5",
"name": "gatk_sortsam_4_1_0_0",
"description": "The **GATK SortSam** tool sorts the input
SAM or BAM file by coordinate, queryname (QNAME), or some
other property of the SAM record.\n\nThe **GATK SortOrder**
of a SAM/BAM file is found in the SAM file header tag @HD
in the field labeled SO. For a coordinate\nsorted SAM/BAM
file, read alignments are sorted first by the reference
sequence name (RNAME) field using the reference\nsequence
dictionary (@SQ tag). Alignments within these subgroups
are secondarily sorted using the left-most
mapping\nposition of the read (POS). Subsequent to this
sorting scheme, alignments are listed
arbitrarily.<\/p><p>For\nqueryname-sorted alignments, all
alignments are grouped using the queryname field but the
alignments are not necessarily\nsorted within these groups.
Reads having the same queryname are derived from the same

```
template\n\n\n###Common Use Cases\n\nThe **GATK SortSam**
tool requires a BAM/SAM file on its **Input SAM/BAM file**
(`--INPUT`) input. The tool sorts input file in the order
defined by (`--SORT_ORDER`) parameter. Available sort order
options are `queryname`, `coordinate` and `duplicate`.
\n\n* Usage example:\n\n```\njava -jar picard.jar SortSam\n
--INPUT=input.bam \n
--SORT_ORDER=coordinate\n```\n\n\n###Changes Introduced by
Seven Bridges\n\n* Prefix of the output file is defined
with the optional parameter **Output prefix**. If **Output
prefix** is not provided, name of the sorted file is
obtained from **Sample ID** metadata from the **Input
SAM/BAM file**, if the **Sample ID** metadata exists.
Otherwise, the output prefix will be inferred form the
**Input SAM/BAM file** filename. \n\n\n###Common Issues and
Important Notes\n\n* None\n\n\n###Performance
Benchmarking\nBelow is a table describing runtimes and task
costs of **GATK SortSam** for a couple of different
samples, executed on the AWS cloud instances:\n\n|
Experiment type | Input size | Paired-end | # of reads |
Read length | Duration | Cost | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|:----------:|:------:|:---
WGS | | Yes | 16M | 101 | 4min | ~0.03$ | c4.2xlarge (8
CPUs) | \n| WGS | | Yes | 50M | 101 | 7min | ~0.04$ |
c4.2xlarge (8 CPUs) | \n| WGS | | Yes | 82M | 101 | 10min |
~0.07$ | c4.2xlarge (8 CPUs) | \n| WES | | Yes | 164M | 101
| 20min | ~0.13$ | c4.2xlarge (8 CPUs) |\n\n*Cost can be
significantly reduced by using **spot instances**. Visit
the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n\n###References\n[1] [GATK SortSam
home
```

```
page](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.12.0/picard_sam_SortS
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "6",
"name": "gatk_setnmmdanduqtags_4_1_0_0",
"description": "The **GATK SetNmMdAndUqTags** tool takes in
a coordinate-sorted SAM or BAM and calculatesthe NM, MD,
and UQ tags by comparing it with the reference. \n\nThe
**GATK SetNmMdAndUqTags** may be needed when **GATK
MergeBamAlignment** was run with **SORT_ORDER** other than
`coordinate` and thus could not fix these tags.
\n\n\n###Common Use Cases\nThe **GATK SetNmMdAndUqTags**
tool fixes NM, MD and UQ tags in SAM/BAM file **Input
SAM/BAM file** (`--INPUT`) input. This tool takes in a
coordinate-sorted SAM or BAM file and calculates the NM,
MD, and UQ tags by comparing with the reference **Reference
sequence** (`--REFERENCE_SEQUENCE`).\n\n* Usage
example:\n\n```\njava -jar picard.jar SetNmMdAndUqTags\n
--REFERENCE_SEQUENCE=reference_sequence.fasta\n
--INPUT=sorted.bam\n```\n\n\n###Changes Introduced by Seven
Bridges\n\n* Prefix of the output file is defined with the
optional parameter **Output prefix**. If **Output prefix**
is not provided, name of the sorted file is obtained from
**Sample ID** metadata form the **Input SAM/BAM file**, if
the **Sample ID** metadata exists. Otherwise, the output
prefix will be inferred form the **Input SAM/BAM file**
filename. \n\n\n\n###Common Issues and Important Notes\n\n*
The **Input SAM/BAM file** must be coordinate sorted in
```

```
order to run **GATK SetNmMdAndUqTags**. \n* If specified,
the MD and NM tags can be ignored and only the UQ tag be
set. \n\n\n###References\n[1] [GATK SetNmMdAndUqTags home
page](https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard_sam_SetNm
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "7",
"name": "gatk_baserecalibrator_4_1_0_0",
"description": "**GATK BaseRecalibrator** generates a
recalibration table based on various covariates for input
mapped read data [1]. It performs the first pass of the
Base Quality Score Recalibration (BQSR) by assessing base
quality scores of the input data.\n\n*A list of **all
inputs and parameters** with corresponding descriptions can
be found at the bottom of the page.*\n\n###Common Use
Cases\n\n* The **GATK BaseRecalibrator** tool requires the
input mapped read data whose quality scores need to be
assessed on its **Input alignments** (`--input`) input, the
database of known polymorphic sites to skip over on its
**Known sites** (`--known-sites`) input and a reference
file on its **Reference** (`--reference`) input. On its
**Output recalibration report** output, the tool generates
a GATK report file with many tables: the list of arguments,
the quantized qualities table, the recalibration table by
read group, the recalibration table by quality score,\nthe
recalibration table for all the optional covariates
[1].\n\n* Usage example:\n\n```\ngatk --java-options
\"-Xmx2048M\" BaseRecalibrator \\\n --input my_reads.bam
```

```
\\\n --reference reference.fasta \\\n --known-sites
sites_of_variation.vcf \\\n --known-sites
another/optional/setOfSitesToMask.vcf \\\n --output
recal_data.table\n\n```\n\n###Changes Introduced by Seven
Bridges\n\n* The output file will be prefixed using the
**Output name prefix** parameter. If this value is not set,
the output name will be generated based on the **Sample
ID** metadata value from the input alignment file. If the
**Sample ID** value is not set, the name will be inherited
from the input alignment file name. In case there are
multiple files on the **Input alignments** input, the files
will be sorted by name and output file name will be
generated based on the first file in the sorted file list,
following the rules defined in the previous case. Moreover,
**recal_data** will be added before the extension of the
output file name which is **CSV** by default.\n\n*
**Include intervals** (`--intervals`) option is divided
into **Include intervals string** and **Include intervals
file** options.\n\n* **Exclude intervals**
(`--exclude-intervals`) option is divided into **Exclude
intervals string** and **Exclude intervals file**
options.\n\n* The following GATK parameters were excluded
from the tool wrapper: `--add-output-sam-program-record`,
`--add-output-vcf-command-line`, `--arguments_file`,
`--cloud-index-prefetch-buffer`, `--cloud-prefetch-buffer`,
`--create-output-bam-index`, `--create-output-bam-md5`,
`--create-output-variant-index`,
`--create-output-variant-md5`, `--gatk-config-file`,
`--gcs-max-retries`, `--gcs-project-for-requester-pays`,
`--help`, `--lenient`, `--QUIET`,
`--sites-only-vcf-output`, `--showHidden`, `--tmp-dir`,
`--use-jdk-deflater`, `--use-jdk-inflater`, `--verbosity`,
```

```
`--version`\n\n\n\n###Common Issues and Important
Notes\n\n* **Memory per job** (`mem_per_job`) input allows
a user to set the desired memory requirement when running a
tool or adding it to a workflow. This input should be
defined in MB. It is propagated to the Memory requirements
part and "-Xmx" parameter of the tool. The default value is
2048MB.\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n* Note: GATK tools
that take in mapped read data expect a BAM file as the
primary format [2]. More on GATK requirements for mapped
sequence data formats can be found
[here](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mappe
Note: **Known sites**, **Input alignments** should have
corresponding index files in the same folder. \n* Note:
**Reference** FASTA file should have corresponding .fai
(FASTA index) and .dict (FASTA dictionary) files in the
same folder. \n* Note: These **Read Filters**
(`--read-filter`) are automatically applied to the data by
the Engine before processing by **BaseRecalibrator** [1]:
**NotSecondaryAlignmentReadFilter**,
**PassesVendorQualityCheckReadFilter**,
**MappedReadFilter**,
**MappingQualityAvailableReadFilter**,
**NotDuplicateReadFilter**,
**MappingQualityNotZeroReadFilter**,
**WellformedReadFilter**\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"LibraryReadFilter\",
the **Library** (`--library`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"PlatformReadFilter\", the **Platform
filter name** (`--platform-filter-name`) option must be set
to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set
to\"PlatformUnitReadFilter\", the **Black listed lanes**
(`--black-listed-lanes`) option must be set to some value.
\n* Note: If the **Read filter** (`--read-filter`) option
is set to \"ReadGroupBlackListReadFilter\", the **Read
group black list** (`--read-group-black-list`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"ReadGroupReadFilter\",
the **Keep read group** (`--keep-read-group`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to
\"ReadLengthReadFilter\", the **Max read length**
(`--max-read-length`) option must be set to some value.\n*
Note: If the **Read filter** (`--read-filter`) option is
set to \"ReadNameReadFilter\", the **Read name**
(`--read-name`) option must be set to some value.\n* Note:
If the **Read filter** (`--read-filter`) option is set to
\"ReadStrandFilter\", the **Keep reverse strand only**
(`--keep-reverse-strand-only`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"SampleReadFilter\", the **Sample**
(`--sample`) option must be set to some value.\n* Note: The
following options are valid only if the appropriate **Read
filter** (`--read-filter`) is specified: **Ambig filter
bases** (`--ambig-filter-bases`), **Ambig filter frac**
(`--ambig-filter-frac`), **Max fragment length**
(`--max-fragment-length`), **Maximum mapping quality**
```

(`--maximum-mapping-quality`), **Minimum mapping quality**
(`--minimum-mapping-quality`), **Do not require soft
clips** (`--dont-require-soft-clips-both-ends`), **Filter
too short** (`--filter-too-short`), **Min read length**
(`--min-read-length`). See the description of each
parameter for information on the associated **Read
filter**.\n* Note: The wrapper has not been tested for the
SAM file type on the **Input alignments** input port, nor
for the BCF file type on the **Known sites** input
port.\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK
BaseRecalibrator** for a couple of different samples,
executed on AWS cloud instances:\n\n| Experiment type |
Input size | Duration | Cost (on-demand) | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 2.2 GB | 9min | ~0.08$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 6.6 GB | 19min | ~0.17$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 11 GB | 27min | ~0.24$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 22 GB | 46min | ~0.41$ | c4.2xlarge
(8 CPUs) |\n\n*Cost can be significantly reduced by using
spot instances. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
BaseRecalibrator](https://gatk.broadinstitute.org/hc/en-us/articles/360036726891-BaseRecalibra
[GATK Mapped sequence data
formats](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Map
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{

```
"step_number": "8",
"name": "gatk_createsequencegroupingtsv_4_1_0_0",
"description": "**CreateSequenceGroupingTSV** tool generate
sets of intervals for scatter-gathering over
chromosomes.\n\nIt takes **Reference dictionary** file
(`--ref_dict`) as an input and creates files which contain
chromosome names grouped based on their
sizes.\n\n\n####**Common Use Cases**\n\nThe tool has only
one input (`--ref_dict`) which is required and has no
additional arguments. **CreateSequenceGroupingTSV** tool
results are **Sequence Grouping** file which is a text file
containing chromosome groups, and **Sequence Grouping with
Unmapped**, a text file which has the same content as
**Sequence Grouping** with additional line containing
\"unmapped\" string.\n\n\n* Usage example\n\n\n```\npython
CreateSequenceGroupingTSV.py \n --ref_dict
example_reference.dict\n\n```\n\n\n\n####**Changes
Introduced by Seven Bridges**\n\nPython code provided
within WGS Germline WDL was adjusted to be called as a
script (`CreateSequenceGroupingTSV.py`).\n\n\n###**Common
Issues and Important Notes**\n\nNone.\n\n\n###
Reference\n[1]
[CreateSequenceGroupingTSV](https://github.com/gatk-workflows/broad-prod-wgs-germline-snps-inde
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "9",
"name": "gatk_gatherbqsrreports_4_1_0_0",
"description": "**GATK GatherBQSRReports** gathers
```

~~~~
scattered BQSR recalibration reports into a single file
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n\n### Common Use Cases \n\n* This tool is
intended to be used to combine recalibration tables from
runs of **GATK BaseRecalibrator** parallelized
per-interval.\n\n* Usage example:\n```\n gatk
--java-options \"-Xmx2048M\" GatherBQSRReports \\\n --input
input1.csv \\\n --input input2.csv \\\n --output
output.csv\n\n```\n\n\n###Changes Introduced by Seven
Bridges\n\n* The output file will be prefixed using the
**Output name prefix** parameter. If this value is not set,
the output name will be generated based on the **Sample
ID** metadata value from **Input BQSR reports**. If the
**Sample ID** value is not set, the name will be inherited
from the **Input BQSR reports** file name. In case there
are multiple files on the **Input BQSR reports** input, the
files will be sorted by name and output file name will be
generated based on the first file in the sorted file list,
following the rules defined in the previous case. Moreover,
**.recal_data** will be added before the extension of the
output file name.\n\n* The following GATK parameters were
excluded from the tool wrapper: `--arguments_file`,
`--gatk-config-file`, `--gcs-max-retries`,
`--gcs-project-for-requester-pays`, `--help`, `--QUIET`,
`--showHidden`, `--tmp-dir`, `--use-jdk-deflater`,
`--use-jdk-inflater`, `--verbosity`,
`--version`\n\n\n###Common Issues and Important Notes\n\n*
**Memory per job** (`mem_per_job`) input allows a user to
set the desired memory requirement when running a tool or
adding it to a workflow. This input should be defined in
MB. It is propagated to the Memory requirements part and
~~~~

```
"-Xmx" parameter of the tool. The default value is
2048MB.\n\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n\n\n###Performance
Benchmarking\n\nThis tool is fast, with a running time of a
few minutes. The experiment task was performed on the
default AWS on-demand c4.2xlarge instance on 50 CSV files
(size of each ~350KB) and took 2 minutes to finish
($0.02).\n\n*Cost can be significantly reduced by using
**spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n\n###References\n\n[1] [GATK
GatherBQSRReports](https://gatk.broadinstitute.org/hc/en-us/articles/360036359192-GatherBQSRRep
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "10",
"name": "gatk_applybqsr_4_1_0_0",
"description": "The **GATK ApplyBQSR** tool recalibrates
the base quality scores of an input BAM or CRAM file
containing reads.\n\nThis tool performs the second pass in
a two-stage process called Base Quality Score Recalibration
(BQSR). Specifically, it recalibrates the base qualities of
the input reads based on the recalibration table produced
by the **GATK BaseRecalibrator** tool. The goal of this
```

```
procedure is to correct systematic biases that affect the
assignment of base quality scores by the sequencer. The
first pass consists of calculating the error empirically
and finding patterns in how the error varies with basecall
features over all bases. The relevant observations are
written to the recalibration table. The second pass
consists of applying numerical corrections to each
individual basecall, based on the patterns identified in
the first step (recorded in the recalibration table), and
writing out the recalibrated data to a new BAM or CRAM file
[1].\n\n*A list of **all inputs and parameters** with
corresponding descriptions can be found at the bottom of
the page.*\n\n###Common Use Cases\n\n* The **GATK
ApplyBQSR** tool requires a BAM or CRAM file on its **Input
alignments** (`--input`) input and the covariates table (=
recalibration file) generated by the **BaseRecalibrator**
tool on its **BQSR recal file** input
(`--bqsr-recal-file`). If the input alignments file is in
the CRAM format, the reference sequence is required on the
**Reference** (`--reference`) input of the tool. The tool
generates a new alignments file which contains recalibrated
read data on its **Output recalibrated alignments**
output.\n\n* Usage example\n\n```\n gatk --java-options
\"-Xmx2048M\" ApplyBQSR \\\n --reference reference.fasta
\\\n --input input.bam \\\n --bqsr-recal-file
recalibration.table \\\n --output output.bam\n\n```\n\n*
Original qualities can be retained in the output file under
the \"OQ\" tag if desired. See the **Emit original quals**
(`--emit-original-quals`) argument for details
[1].\n\n###Changes Introduced by Seven Bridges\n\n* The
output file will be prefixed using the **Output name
prefix** parameter. If this value is not set, the output
```

```
name will be generated based on the **Sample ID** metadata
value from the input alignments file. If the **Sample ID**
value is not set, the name will be inherited from the input
alignments file name. In case there are multiple files on
the **Input alignments** input, the files will be sorted by
name and output file name will be generated based on the
first file in the sorted file list, following the rules
defined in the previous case. Moreover, **recalibrated**
will be added before the extension of the output file
name.\n\n* The user has a possibility to specify the output
file format using the **Output file format** argument.
Otherwise, the output file format will be the same as the
format of the input file.\n\n* **Include intervals**
(`--intervals`) option is divided into **Include intervals
string** and **Include intervals file** options.\n\n*
**Exclude intervals** (`--exclude-intervals`) option is
divided into **Exclude intervals string** and **Exclude
intervals file** options.\n\n* The following GATK
parameters were excluded from the tool wrapper:
`--add-output-vcf-command-line`, `--arguments_file`,
`--cloud-index-prefetch-buffer`, `--cloud-prefetch-buffer`,
`--create-output-bam-md5`, `--create-output-variant-index`,
`--create-output-variant-md5`, `--gatk-config-file`,
`--gcs-max-retries`, `--gcs-project-for-requester-pays`,
`--help`, `--lenient`, `--QUIET`,
`--sites-only-vcf-output`, `--showHidden`, `--tmp-dir`,
`--use-jdk-deflater`, `--use-jdk-inflater`, `--verbosity`,
`--version`\n\n###Common Issues and Important Notes\n\n*
**Memory per job** (`mem_per_job`) input allows a user to
set the desired memory requirement when running a tool or
adding it to a workflow. This input should be defined in
MB. It is propagated to the Memory requirements part and
```

```
"-Xmx" parameter of the tool. The default value is
2048MB.\n* **Memory overhead per job**
(`mem_overhead_per_job`) input allows a user to set the
desired overhead memory when running a tool or adding it to
a workflow. This input should be defined in MB. This amount
will be added to the Memory per job in the Memory
requirements section but it will not be added to the "-Xmx"
parameter. The default value is 100MB. \n* Note: GATK tools
that take in mapped read data expect a BAM file as the
primary format [2]. More on GATK requirements for mapped
sequence data formats can be found
[here](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mappe
Note: **Input alignments** should have corresponding index
files in the same folder. \n* Note: **Reference** FASTA
file should have corresponding .fai (FASTA index) and .dict
(FASTA dictionary) files in the same folder. \n* Note: This
tool replaces the use of PrintReads for the application of
base quality score recalibration as practiced in earlier
versions of GATK (2.x and 3.x) [1].\n* Note: You should
only run **ApplyBQSR** with the covariates table created
from the input BAM or CRAM file [1].\n* Note: This **Read
Filter** (`--read-filter`) is automatically applied to the
data by the Engine before processing by **ApplyBQSR** [1]:
**WellformedReadFilter**\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"LibraryReadFilter\",
the **Library** (`--library`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"PlatformReadFilter\", the **Platform
filter name** (`--platform-filter-name`) option must be set
to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set
to\"PlatformUnitReadFilter\", the **Black listed lanes**
```

```
(`--black-listed-lanes`) option must be set to some value.
\n* Note: If the **Read filter** (`--read-filter`) option
is set to \"ReadGroupBlackListReadFilter\", the **Read
group black list** (`--read-group-black-list`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to \"ReadGroupReadFilter\",
the **Keep read group** (`--keep-read-group`) option must
be set to some value.\n* Note: If the **Read filter**
(`--read-filter`) option is set to
\"ReadLengthReadFilter\", the **Max read length**
(`--max-read-length`) option must be set to some value.\n*
Note: If the **Read filter** (`--read-filter`) option is
set to \"ReadNameReadFilter\", the **Read name**
(`--read-name`) option must be set to some value.\n* Note:
If the **Read filter** (`--read-filter`) option is set to
\"ReadStrandFilter\", the **Keep reverse strand only**
(`--keep-reverse-strand-only`) option must be set to some
value.\n* Note: If the **Read filter** (`--read-filter`)
option is set to \"SampleReadFilter\", the **Sample**
(`--sample`) option must be set to some value.\n* Note: The
following options are valid only if an appropriate **Read
filter** (`--read-filter`) is specified: **Ambig filter
bases** (`--ambig-filter-bases`), **Ambig filter frac**
(`--ambig-filter-frac`), **Max fragment length**
(`--max-fragment-length`), **Maximum mapping quality**
(`--maximum-mapping-quality`), **Minimum mapping quality**
(`--minimum-mapping-quality`), **Do not require soft
clips** (`--dont-require-soft-clips-both-ends`), **Filter
too short** (`--filter-too-short`), **Min read length**
(`--min-read-length`). See the description of each
parameter for information on the associated **Read
filter**.\n* Note: The wrapper has not been tested for the
```

```
SAM file type on the **Input alignments** input
port.\n\n###Performance Benchmarking\n\nBelow is a table
describing runtimes and task costs of **GATK ApplyBQSR**
for a couple of different samples, executed on the AWS
cloud instances:\n\n| Experiment type | Input size |
Duration | Cost (on-demand) | Instance (AWS) |
\n|:--------------:|:------------:|:--------:|:-------:|:---------:|\n|
RNA-Seq | 2.2 GB | 8min | ~0.07$ | c4.2xlarge (8 CPUs) |
\n| RNA-Seq | 6.6 GB | 23min | ~0.21$ | c4.2xlarge (8 CPUs)
| \n| RNA-Seq | 11 GB | 37min | ~0.33$ | c4.2xlarge (8
CPUs) | \n| RNA-Seq | 22 GB | 1h 16min | ~0.68$ |
c4.2xlarge (8 CPUs) |\n\n*Cost can be significantly reduced
by using **spot instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.*\n\n###References\n\n[1] [GATK
ApplyBQSR](https://gatk.broadinstitute.org/hc/en-us/articles/360036725911-ApplyBQSR)\n\n[2]
[GATK Mapped sequence data
formats](https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapp
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "11",
"name": "gatk_gatherbamfiles_4_1_0_0",
"description": "**GATK GatherBamFiles** concatenates one or
more BAM files resulted form scattered paralel anaysis.
\n\n\n### Common Use Cases \n\n* **GATK GatherBamFiles**
tool performs a rapid \"gather\" or concatenation on BAM
files into single BAM file. This is often needed in
operations that have been run in parallel across genomics
```

regions by scattering their execution across computing
nodes and cores thus resulting in smaller BAM files.\n*
Usage example:\n```\n\njava -jar picard.jar
GatherBamFiles\n --INPUT=input1.bam\n
--INPUT=input2.bam\n```\n\n### Common Issues and Important
Notes\n* **GATK GatherBamFiles** assumes that the list of
BAM files provided as input are in the order that they
should be concatenated and simply links the bodies of the
BAM files while retaining the header from the first file.
\n* Operates by copying the gzip blocks directly for speed
but also supports the generation of an MD5 in the output
file and the indexing of the output BAM file.\n* This tool
only support BAM files. It does not support SAM
files.\n\n###Changes Intorduced by Seven Bridges\n*
Generated output BAM file will be prefixed using the
Output prefix parameter. In case the **Output prefix**
is not provided, the output prefix will be the same as the
Sample ID metadata from the **Input alignments**, if
the **Sample ID** metadata exists. Otherwise, the output
prefix will be inferred from the **Input alignments**
filename. This way, having identical names of the output
files between runs is avoided.",
"version": "4.1.0.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "12",
"name": "samtools_view_1_9_cwl1_0",
"description": "**SAMtools View** tool prints all
alignments from a SAM, BAM, or CRAM file to an output file

```
in SAM format (headerless). You may specify one or more
space-separated region specifications to restrict output to
only those alignments which overlap the specified
region(s). Use of region specifications requires a
coordinate-sorted and indexed input file (in BAM or CRAM
format) [1].\n\n*A list of **all inputs and parameters**
with corresponding descriptions can be found at the bottom
of the page.*\n\n####Regions\n\nRegions can be specified
as: RNAME[:STARTPOS[-ENDPOS]] and all position coordinates
are 1-based. \n\n**Important note:** when multiple regions
are given, some alignments may be output multiple times if
they overlap more than one of the specified
regions.\n\nExamples of region specifications:\n\n-
**chr1** - Output all alignments mapped to the reference
sequence named `chr1' (i.e. @SQ SN:chr1).\n\n-
**chr2:1000000** - The region on chr2 beginning at base
position 1,000,000 and ending at the end of the
chromosome.\n\n- **chr3:1000-2000** - The 1001bp region on
chr3 beginning at base position 1,000 and ending at base
position 2,000 (including both end positions).\n\n-
**'\\*'** - Output the unmapped reads at the end of the
file. (This does not include any unmapped reads placed on a
reference sequence alongside their mapped mates.)\n\n-
**.** - Output all alignments. (Mostly unnecessary as not
specifying a region at all has the same effect.)
[1]\n\n###Common Use Cases\n\nThis tool can be used for:
\n\n- Filtering BAM/SAM/CRAM files - options set by the
following parameters and input files: **Include reads with
all of these flags** (`-f`), **Exclude reads with any of
these flags** (`-F`), **Exclude reads with all of these
flags** (`-G`), **Read group** (`-r`), **Minimum mapping
quality** (`-q`), **Only include alignments in library**
```

(`-l`), **Minimum number of CIGAR bases consuming query sequence** (`-m`), **Subsample fraction** (`-s`), **Read group list** (`-R`), **BED region file** (`-L`)\n- Format conversion between SAM/BAM/CRAM formats - set by the following parameters: **Output format** (`--output-fmt/-O`), **Fast bam compression** (`-1`), **Output uncompressed BAM** (`-u`)\n- Modification of the data which is contained in each alignment - set by the following parameters: **Collapse the backward CIGAR operation** (`-B`), **Read tags to strip** (`-x`)\n- Counting number of alignments in SAM/BAM/CRAM file - set by parameter **Output only count of matching records** (`-c`)\n\n###Changes Introduced by Seven Bridges\n\n- Parameters **Output BAM** (`-b`) and **Output CRAM** (`-C`) were excluded from the wrapper since they are redundant with parameter **Output format** (`--output-fmt/-O`).\n- Parameter **Input format** (`-S`) was excluded from wrapper since it is ignored by the tool (input format is auto-detected).\n- Input file **Index file** was added to the wrapper to enable operations that require an index file for BAM/CRAM files.\n- Parameter **Number of threads** (`--threads/-@`) specifies the total number of threads instead of additional threads. Command line argument (`--threads/-@`) will be reduced by 1 to set the number of additional threads.\n\n###Common Issues and Important Notes\n\n- When multiple regions are given, some alignments may be output multiple times if they overlap more than one of the specified regions [1].\n- Use of region specifications requires a coordinate-sorted and indexed input file (in BAM or CRAM format) [1].\n- Option **Output uncompressed BAM** (`-u`) saves time spent on compression/decompression and is thus preferred when the

```
output is piped to another SAMtools command
[1].\n\n###Performance Benchmarking\n\nMultithreading can
be enabled by setting parameter **Number of threads**
(`--threads/-@`). In the following table you can find
estimates of **SAMtools View** running time and cost.
\n\n*Cost can be significantly reduced by using **spot
instances**. Visit the [Knowledge
Center](https://docs.sevenbridges.com/docs/about-spot-instances)
for more details.* \n\n| Input type | Input size | # of
reads | Read length | Output format | # of threads |
Duration | Cost | Instance
(AWS)|\n|---------------|--------------|-----------------|---------------|------------------|-
BAM | 5.26 GB | 71.5M | 76 | BAM | 1 | 13min. | \\$0.12 |
c4.2xlarge |\n| BAM | 11.86 GB | 161.2M | 101 | BAM | 1 |
33min. | \\$0.30 | c4.2xlarge |\n| BAM | 18.36 GB | 179M |
76 | BAM | 1 | 60min. | \\$0.54 | c4.2xlarge |\n| BAM |
58.61 GB | 845.6M | 150 | BAM | 1 | 3h 25min. | \\$1.84 |
c4.2xlarge |\n| BAM | 5.26 GB | 71.5M | 76 | BAM | 8 |
5min. | \\$0.04 | c4.2xlarge |\n| BAM | 11.86 GB | 161.2M |
101 | BAM | 8 | 11min. | \\$0.10 | c4.2xlarge |\n| BAM |
18.36 GB | 179M | 76 | BAM | 8 | 19min. | \\$0.17 |
c4.2xlarge |\n| BAM | 58.61 GB | 845.6M | 150 | BAM | 8 |
61min. | \\$0.55 | c4.2xlarge |\n| BAM | 5.26 GB | 71.5M |
76 | SAM | 8 | 14min. | \\$0.13 | c4.2xlarge |\n| BAM |
11.86 GB | 161.2M | 101 | SAM | 8 | 23min. | \\$0.21 |
c4.2xlarge |\n| BAM | 18.36 GB | 179M | 76 | SAM | 8 |
35min. | \\$0.31 | c4.2xlarge |\n| BAM | 58.61 GB | 845.6M
| 150 | SAM | 8 | 2h 29min. | \\$1.34 | c4.2xlarge
|\n\n###References\n\n[1] [SAMtools
documentation](http://www.htslib.org/doc/samtools-1.9.html)",
"version": "1.9",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "13",
"name": "sbg_lines_to_interval_list_abr",
"description": "This tools is used for splitting GATK
sequence grouping file into subgroups.\n\n### Common Use
Cases\n\nEach subgroup file contains intervals defined on
single line in grouping file. Grouping file is output of
GATKs **CreateSequenceGroupingTSV** script which is used in
best practice workflows sush as **GATK Best Practice
Germline Workflow**.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
},
{
"step_number": "14",
"name": "sbg_lines_to_interval_list_br",
"description": "This tools is used for splitting GATK
sequence grouping file into subgroups.\n\n### Common Use
Cases\n\nEach subgroup file contains intervals defined on
single line in grouping file. Grouping file is output of
GATKs **CreateSequenceGroupingTSV** script which is used in
best practice workflows sush as **GATK Best Practice
Germline Workflow**.",
"version": "1.0",
"prerequisite": [],
"input_list": [],
"output_list": []
```

```
}
]
}
```

1.8 Input/Output Domain

```
{
  "input_subdomain": [
    {
      "uri": [
        {
          "filename": "",
          "uri": "",
          "access_time": ""
        }
      ]
    }
  ],
  "output_subdomain": [
    {
      "mediatype": "",
      "uri": [
        {
          "uri": "",
          "access_time": ""
        }
      ]
    }
  ]
}
```

1.9 Error Domain

```
{
  "empirical_error": [],
  "algorithmic_error": []
}
```

2 Funding

The Seven Bridges Cancer Genomics Cloud has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

3 References

Lau et al (2017) The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized—A New Paradigm in Large-Scale Computational Research. Cancer Res. 77(21):e3-e6. doi: 10.1158/0008-5472.CAN-17-0387.

4 Appendix 1: BioCompute Object Specification v1.3.0

Name	ID	Description
Top Level Fields		
BioCompute Object Indetifier	BCO_id	Unique identifier that should be apllied to each BCO instance. Assigned by a BCO database engine, like URL. It never be reused.
Type	type	As any object of the type, it has its own fields.
Digital signature	digital_signature	A string-type, read-only generated and stored by a BCO database, protecting the object from internal or external alterations without proper validation. It can be used for validation, downloading, and transferring BCOs.
BCO version	bco_spec_version	The version of the BCO specification used to define this document.
Provenance Domain		
Name	name	Name of the BCO.
Structured name	structured_name	Computable text field designed to represent a BCO instance name in visible interfaces
Version	version	Records the versioning of this BCO instance object. A change in the BCO affecting the outcome of the computation should be deposited as a new BCO, not as a new version.
Review	review	Describes the status of an object in the review process. Status flags: unreviewed, in-review, approved, suspended, rejected.
Inheritance/derivation	derived_from	If the object is derived from another, this field will specify the parent object, in the form of the objectid. It is null, if inherits only from the base BioCompute Object or a type definition.
Obsolescence	obsolete	If the object has an expiration date this field will specify that using the datetime type.
Embargo	embargo	If the object has a period of time that it is not public, that range can be specified using these fields. Using the datetime type a start and end time are specified for the embargo.
Created	created	Using the datetime type the time of initial creation of the BCO is recorded.
Modification	modified	Using the datetime type the time of most recent modification of the BCO is recorded.
Contributors	contributors	List to hold contributor identifiers and a description of their type of contribution, including a field for ORCIDs to record author information, as they allow for the author to curate their information after submission.
License	license	A space for Creative commons licence or other licence information. The default or recommended licence can be Attribution 4.0 International.
Usability Domain		
Usability Domain	usability_domain	Provides a space for the author to define the usability domain of the BCO. It is an array of free text values. This field is to aid in search-ability and provide a specific description of the object. It helps determine when and how the BCO can be used.
Extension Domain		

(continued)

Name	ID	Description
Extension Domain	extension_domain	For a user to add more structured information that is defined in the type definition. This section is not evaluated by checks for BCO validity or computational correctness.
Extension to External References: SMART on FHIR Genomics	Extension to External References: SMART on FHIR Genomics	SMART on FHIR Genomics provides a framework for HER-based apps to built on FHIR that integrate clinical and genomics information.
Extension to External References: GitHub	Extension to External References: GitHub	Include an extension to GitHub repositories where HTS computational analysis pipelins, workflows, protocols, and tool or software source code can be stored, deposited, downloaded.
Description Domain		
Description Domain	description_domain	Structured field for description of external references, the pipeline steps, and the relationship of IO objects. Information in this domain is not used for computation. Capture information that is currently being provided in FDA submission in journal format.
Keywords	keywords	List of key map fields to hold a list of keywords to aid in search-ability and description of the object.
External References	xref	It contains a list of the databases and/or ontology IDs that are cross-referenced in the BCO. It provides more specificity in the information related to BCO entries.
Pipeline tools	pipeline_steps	For recording the specifics of a pipeline. Each individual tool is represented as step, at the discretion of the author. Step Number (step_number), Name (name), Tool Description (description), Tool Version (version), Tool Prerequisites (prerequisite), Input List (input_list), Output List (output_list).
Execution Domain		
Execution Domain	execution_domain	The filelds required for execution of the BCO have been encapsulated together in order to clearly separate information needed for deployment, software configuration, and running applications in a dependent enviroment.
Script Access Type	script_access_type	This field indicates whether the code of the script to execute the BioCompute Object is access as an external file via HTTP or in-line text in the script field.
Script	script	Points to an internal or external reference to a script object that was used to perform computations for this BCO instance. This may be reference to Galaxy Project or Seven Bridges Genomics pipeline, a Common Workflow Language (CWL) object in GitHub, HIVE computational service or any other type of script.
Pipeline Version	pipeline_version	This field records the version of the pipeline implementation.
Platform/Environment	platform	The multi-value reference to a particular deployment of an existing platform where this BCO can be reproduced (Galaxy or HIVE or CASAVA).
Script Driver	script_driver	The reference to an executable that can be launched in order to perform a sequence of commands described in the script. For example if the pipeline is driven by a HIVE script, the script driver is the hive execution engine. For CWL based scripts specify cwl-runner. Another very general commonly used in Linux based operating systems is shell.

(continued)

Name	ID	Description
Algorithmic tools and Software Prerequisites	software_prerequisites	Field listing the minimal necessary prerequisites, library, tool versions needed to successfully run the script to produce BCO.
Domain Prerequisites	domain_prerequisites	Listing the minimal necessary domain specific external data source access in order to successfully run the script to produce BCO.
Enviromental parameters	env_parameters	Multi-value additional key value pairs useful to configure the execution environment on the target platform, like compute cores, available memory use of the script.
Parametric Domain		
Parametric Domain	parametric_domain	List of parameters customizing the computational flow which can affect the output of the calculations. These fields are custom to each type of analysis and are tied to a particular pipeline implementation.
Input and Output Domain		
Input and output Domain	io_domain	This represents the list of global input and output files created by the computational workflow, excluding the intermediate files.
Input Subdomain	input_subdomain	This field records the references and input files for the entire pipeline. Each type of input file is listed under a key for that type.
Output Subdomain	output_subdomain	This field records the outputs for the entire pipeline .
Error Domain, acceptable range of variability		
Error Domain, acceptable range of variability	error_domain	Consists of two subdomains: empirical and algorithmic._The empirical subdomain contains the limits of_detectability_ fps, fns, statistical confidence of outcomes, etc. The algorithmic subdomain is descriptive of errors that originated by fuzziness of the algorithms, driven by stochastic processes, in dynamically parallelized multi-threaded executions, or in machine learning methodologies where the state of the machine can affect the outcomeConsists of two subdomains: empirical and algorithmic. The empirical subdomain contains the limits of detectability FPs, FNs, statistical confidence of outcomes, etc. The algorithmic subdomain is descriptive of errors that originated by fuzziness of the algorithms, driven by stochastic processes, in dynamically parallelized multi-threaded executions, or in machine learning methodologies where the state of the machine can affect the outcome.

5 Appendix 2: The Complete BioCompute Object

```
{
  "spec_version": "https://w3id.org/biocompute/1.4.2/",
  "object_id": "https://biocompute.sbgenomics.com/bco/58218981-5b14-4883-90c2-48c188be74d8",
  "etag": "57f437f0e2f1162ca3e1b3690860f80cac325996bb89e4965179241d224b9beb",
  "provenance_domain": {
    "name": "GATK Best Practice Data Pre-processing 4.1.0.0",
    "version": "1.0.0",
    "review": [],
    "derived_from": "https://cgc-api.sbgenomics.com/v2/apps/phil_webster/bco-cwl-examples/broa
    "obsolete_after": "2023-02-16T00:00:00+0000",
    "embargo": ["2023-02-16T00:00:00+0000", "2023-02-16T00:00:00+0000"],
    "created": "2023-02-16T00:00:00+0000",
    "modified": "2023-02-16T00:00:00+0000",
    "contributors": [],
    "license": "https://spdx.org/licenses/CC-BY-4.0.html"
  },
  "usability_domain": "**Note:** This version of the GATK Best Practice Data Pre-processing 4.
  "extension_domain": {
    "fhir_extension": {
      "fhir_endpoint": "",
      "fhir_version": "",
      "fhir_resources": {}
    },
    "scm_extension": {
      "scm_repository": "",
      "scm_type": "git",
      "scm_commit": "",
      "scm_path": "",
      "scm_preview": ""
    }
  },
  "description_domain": {
    "keywords": [],
    "xref": [],
    "platform": [
      "Seven Bridges Platform"
    ],
    "pipeline_steps": [
      {
        "step_number": "1",
        "name": "gatk_markduplicates_4_1_0_0",
        "description": "The **GATK  MarkDuplicates** tool identifies duplicate reads in a BAM
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "2",
        "name": "bwa_mem_bundle_0_7_15",
        "description": "BWA-MEM is an algorithm designed for aligning sequence reads onto a la
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "4",
        "name": "gatk_samtofastq_4_1_0_0",
        "description": "The **GATK SamToFastq** tool converts a SAM or BAM file to FASTQ.\n\nT
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "5",
        "name": "gatk_sortsam_4_1_0_0",
        "description": "The **GATK SortSam** tool sorts the input SAM or BAM file by coordinate
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "6",
        "name": "gatk_setnmmdanduqtags_4_1_0_0",
        "description": "The **GATK SetNmMdAndUqTags** tool takes in a coordinate-sorted SAM or
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "7",
        "name": "gatk_baserecalibrator_4_1_0_0",
        "description": "**GATK BaseRecalibrator** generates a recalibration table based on var
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "8",
        "name": "gatk_createsequencegroupingtsv_4_1_0_0",
        "description": "**CreateSequenceGroupingTSV** tool generate sets of intervals for scat
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "9",
        "name": "gatk_gatherbqsrreports_4_1_0_0",
        "description": "**GATK GatherBQSRReports** gathers scattered BQSR recalibration report
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "10",
        "name": "gatk_applybqsr_4_1_0_0",
        "description": "The **GATK ApplyBQSR** tool recalibrates the base quality scores of an
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "11",
        "name": "gatk_gatherbamfiles_4_1_0_0",
        "description": "**GATK GatherBamFiles** concatenates one or more BAM files resulted fo
        "version": "4.1.0.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "12",
        "name": "samtools_view_1_9_cwl1_0",
        "description": "**SAMtools View** tool prints all alignments from a SAM, BAM, or CRAM
        "version": "1.9",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "13",
        "name": "sbg_lines_to_interval_list_abr",
        "description": "This tools is used for splitting GATK sequence grouping file into subg
        "version": "1.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      },
      {
        "step_number": "14",
        "name": "sbg_lines_to_interval_list_br",
        "description": "This tools is used for splitting GATK sequence grouping file into subg
        "version": "1.0",
        "prerequisite": [],
        "input_list": [],
        "output_list": []
      }
    ]
  },
  "execution_domain": {
    "script": [
      "https://cgc-api.sbgenomics.com/v2/apps/phil_webster/bco-cwl-examples/broad-best-practic
    ],
    "script_driver": "Seven Bridges Common Workflow Language Executor",
    "software_prerequisites": [],
    "external_data_endpoints": [],
    "environment_variables": []
  },
  "parametric_domain": [],
  "io_domain": {
    "input_subdomain": [
      {
        "uri": [
          {
            "filename": "",
            "uri": "",
            "access_time": ""
          }
        ]
      }
    ],
    "output_subdomain": [
      {
        "mediatype": "",
        "uri": [
          {
            "uri": "",
            "access_time": ""
          }
        ]
      }
    ]
  },
  "error_domain": {
    "empirical_error": [],
    "algorithmic_error": []
  }
}
```