



INSTITUTE
FOR RESEARCH
IN BIOMEDICINE

The `bam-utils` Software Package

Daniel Ortiz

5 Nov 2018

Table of Contents

1. Introduction
2. Package Overview
3. Main Tools and File Formats
4. Whole Pipeline Example

Introduction

Introduction

- Executing bam file pipelines is a tedious task
 - Downloading of very large files
 - Combination of tools with different input requirements
 - Existence of dependencies between tools
 - Tools may need to be added or removed
 - Each tool has specific computational requirements
 - Pipeline may need to be executed for hundreds of files
 - Parallelism should be exploited when possible
 - ...
- bam-utils has been created as a highly portable, configurable and extensible solution

Package Overview

Package Dependencies

- Shell Bash
- Python
- Conda
- Database download clients
- Slurm Workload Manager (optional)

Package Installation

- Obtain the package using git:

```
git clone gitlab@fsupeksvr.irbbarcelona.pcb.ub.es:dortiz/bam-utils.git
```

- Change to the directory with the package's source code and type:

```
./reconf  
./configure  
make  
make install
```

NOTE: use `--prefix` option of `configure` to install the package in a custom directory

Additional configure Options

- `--with-icgcstor=DIR`: sets location of ICGC's storage client
- `--with-aspera=DIR`: enables Aspera Connect download client
- `--with-egadecrypt=DIR`: location of EGA decryptor tool

Functionality

- Automate execution of pipelines for pairs of normal-tumor bam files in a highly configurable manner
- Automate processing of all of the normal-tumor samples of a dataset
- Handle file downloading as part of pipeline execution
- Keep track of which analysis steps of a pipeline for a pair of bam files have been completely executed and which haven't
- Specification of computational resources for each step

Execution Model

- `bam-utils` follows a simple execution model based on a file containing a list of analysis steps to be executed
- All of the steps defined in the list are by default executed simultaneously unless dependencies between them are specified
- This model requires the availability of a workload manager
- Otherwise, steps are executed sequentially

Supported Databases and Download Clients

- Databases
 - EGA
 - ICGC
- Download clients
 - aspc
 - icgc-storage-client
 - Amazon cloud
 - Collaboratory cloud
 - pyega3

Implemented Analysis Steps

- **bam file downloading:**
 - `download_ega_{norm|tum}_bam`
 - `download_ega_asp_{norm|tum}_bam`
 - `download_aws_{norm|tum}_bam`
 - `download_collab_{norm|tum}_bam`
- **bam file manipulation:**
 - `sort_{norm|tum}_bam`
 - `index_{norm|tum}_bam`
 - `delete_bam_files`

Implemented Analysis Steps

- **Small Indels and Single Nucleotide Variant Callers:**
 - `manta_somatic`
 - `platypus_germline`
 - `strelka_somatic`
- **Copy Number Variant Callers:**
 - `ascatngs`
 - `cnvkit`
 - `facets`
 - `wisecondorx`
- **MSI Analyzers:**
 - `msisensor`

Main Tools and File Formats

- `submit_bam_analysis`
- `query_ega_metadata`
- `query_icgc_metadata`
- `analyze_dataset`

- Automates analysis of a normal-tumor sample pair
- Main input parameters:
 - -r <string>: file with reference genome
 - -n <string>: file with normal bam file
 - -t <string>: file with tumor bam file
 - -extn <string>: external database id of normal bam file
 - -extt <string>: external database id of tumor bam file
 - -a <string>: file with analysis steps to be performed
 - -o <string>: output directory

- Content of output directory:
 - data: directory containing the normal-tumor bam files
 - scripts: directory containing the scripts used for each analysis step
 - <analysis_step_name>: directory containing the results of the analysis step of the same name

- Extracts information from EGA metadata
- Main input parameters:
 - -s <string>: file with sample information
 - -a <string>: file with analysis information
 - -t <string>: file with study information
 - -p <string>: file listing Aspera box content
 - -f <int>: output format

- Extracts information from ICGC metadata
- Main input parameters:
 - -d <string>: file with donor information
 - -a <string>: file with aws manifest
 - -t <string>: table file in json format
 - -f <int>: output format:

- Uses metadata information to automate analysis of a whole dataset
- Main input parameters:
 - -r <string>: file with reference genome
 - -m <string>: file with metadata, one entry per line
 - -a <string>: file with analysis steps to be performed
 - -p: only print the commands executing the analysis

The `bam_utils_lib.sh` Library

- Shell library containing functions used by the previously described tools
- Functions can be classified as follows:
 - Implementation of the package execution model
 - Automated creation of scripts executing analysis steps
 - Implementation of analysis steps

Additional Tools

- Reference genome operations:
 - `filter_contig_from_genref`
 - `gen_bed_for_genome`
- Data preparation for analysis steps:
 - `convert_snppos_to_snpgcc`
 - `create_snv_pos_ascat`
 - `gen_wisecondorx_ref`
- Reporting tools:
 - `get_analysis_status`

- **Analysis file:** file describing all of the analysis steps to be carried out when processing a normal-tumor sample
- **EGA/ICGC metadata:** information regarding a whole dataset that is typically spread out in a set of files
- **Analysis metadata:** file providing all the information of a given dataset that is relevant to automate its analysis
- **Analysis automation script:** file with a sequence of commands automating the analysis of a dataset

- **Entry format** (one entry per line):

Step name, Slurm account, Slurm partition, CPUs, Memory limit, Time limit, Dependencies

- **Dependency types:** none, after, afterok, afternotok, afterany

```
download_ega_norm_bam dortiz normal_prio 1 2048 10:00:00 jobdeps=none
download_ega_tum_bam dortiz normal_prio 1 2048 10:00:00 jobdeps=none
index_norm_bam dortiz normal_prio 1 1G 4:00:00 jobdeps=afterok:download_ega_norm_bam
index_tum_bam dortiz normal_prio 1 1G 4:00:00 jobdeps=afterok:download_ega_tum_bam
manta_somatic dortiz normal_prio 8 3G 6:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam
strelka_somatic dortiz normal_prio 8 6G 6:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam,
    afterok:manta_somatic
delete_bam_files dortiz normal_prio 1 1G 0:10:00 jobdeps=afterok:manta_somatic,afterok:strelka_somatic,
    afterok:msisensor,afterok:cnvkit,afterok:facets,afterok:ascatngs,afterok:platypus_germline
```


- Sample information (`Sample_File.map`)
 - contains file name info
- Analysis information (`Analysis_Sample_meta_info.map`)
 - contains donor and phenotype information
- Study information (`Study_analysis_sample.map`)
 - contains EGA id information
- Aspera box content (`dbbox_content`)

- Donor information (`donor.<study_name>.tsv`)
 - contains gender information
- AWS manifest (`manifest.aws-virginia.<code>.tsv`)
 - contains object id, file name and donor id information
- JSON table file (`icgc_table.json`)
 - contains phenotype information

Analysis Metadata (EGA)

- Created with the `query_ega_metadata` tool
- Example entries:

```
EGAF00001664282 phenotype=Blood|Normal_blood gender=male ; EGAF00001664327 phenotype=Skin|  
  Tumour_metastasis_to_local_lymph_node gender=male  
  
EGAF00001670586 phenotype=Blood|Normal_blood gender=male ; EGAF00001664289 phenotype=Skin|  
  Tumour_metastasis_to_local_lymph_node gender=male  
  
EGAF00001664356 phenotype=Skin|Tumour_metastasis_to_distant_location gender=male ; EGAF00001670533  
  phenotype=Blood|Normal_blood gender=male  
  
EGAF00001661882 phenotype=Blood|Normal_blood gender=male ; EGAF00001661538 phenotype=Skin|  
  Tumour_metastasis_to_local_lymph_node gender=male  
...
```

Analysis Metadata (EGA Aspera)

- Created with the `query_ega_metadata` tool
- Example entries:

```
EGAD00001003388/PART_2/EGAZ00001300436_20170516_AWS_MELA_3c3ed66c-1505-4614-ac9d-575a6713b06a.bam.crypt
  phenotype=Blood|Normal_blood gender=male ; EGAD00001003388/PART_3/
EGAZ00001300354_20170516_AWS_MELA_daf1ffd8-0a0f-4869-abc8-5be0b4fc1a21.bam.crypt phenotype=Skin|
Tumour_metastasis_to_local_lymph_node gender=male

EGAD00001003388/PART_3/EGAZ00001303407_20170516_AWS_MELA_a197619e-f3e2-41f6-aef7-d1fadb3c1f5b.bam.crypt
  phenotype=Blood|Normal_blood gender=male ; EGAD00001003388/PART_2/
EGAZ00001300389_20170516_AWS_MELA_3a9bf676-1a7b-4718-8396-fb36cc89b688.bam.crypt phenotype=Skin|
Tumour_metastasis_to_local_lymph_node gender=male

EGAD00001003388/PART_3/EGAZ00001300416_20170516_AWS_MELA_f64eba46-d8a1-46f2-ba66-1b509e16c946.bam.crypt
  phenotype=Skin|Tumour_metastasis_to_distant_location gender=male ; EGAD00001003388/PART_3/
EGAZ00001303394_20170516_AWS_MELA_7bb66858-7533-4f96-9cd4-41aae2fe18b2.bam.crypt phenotype=Blood|
Normal_blood gender=male

...
```

Analysis Metadata (ICGC)

- Created with the `query_icgc_metadata` tool
- Example entries:

```
34fa2369-424f-5886-9d23-6d19f8f15278 tumor female ; d759d07f-330c-5d0c-bd28-af72147dfb17 normal female
284f1424-d250-59cf-b105-da277b061e4a normal female ; e7e69d23-fb0d-5d3d-9027-ebf355053dbf tumor female
c42fffad-4ffd-59ba-93f1-2c573547369c normal female ; 3a33ef20-dfd0-50b0-afc2-38de9a5baa32 tumor female
37f076d6-fa64-5b5d-a0d0-b5cd7428d4a2 normal female ; 2c34270b-98d2-54b9-bdd3-068c6a9d858f tumor female
...
```

Analysis Automation Script

- Created with the `analyze_dataset` tool (`-p` option)
- At each entry (one per line), `submit_bam_analysis` tool is used to analyze a normal-tumor bam file pair
- Entry example:

```
/home/dortiz/bio/software/bam-utils/bin/submit_bam_analysis -r /home/dortiz/bio/data/genome_references/
refseq_hg19_filt.fa -extn d759d07f-330c-5d0c-bd28-af72147dfb17 -extt 34fa2369-424f-5886-9d23-6
d19f8f15278 -a /home/dortiz/bio/software/bam-utils/share/bam-utils/examples/basic_analysis_test.
csv -g XX -o /mnt/raid/dortiz/bio/tasks/bam_analysis_testing_pipeline/d759d07f-330c-5d0c-bd28-
af72147dfb17_34fa2369-424f-5886-9d23-6d19f8f15278 -cr /home/dortiz/bio/data/genome_references/
refseq_hg19_filt.fa.bed -sv /home/dortiz/bio/data/facets_info/00-common_all.vcf -sg /home/dortiz/
bio/data/ascatngs_info/r93/SnpGcCorrections_GRCh37_1000g.tsv -mc chrY -egastr 50 -egacred /home/
dortiz/bio/software/ega-download-client-python/dortiz_cred.json
```

Extending Package Functionality

- Creating a new pipeline step requires only two actions:
 - Add a function in `bam_utils_lib.sh` defining the step
 - Define a function in `submit_bam_analysis.sh` passing the input parameters
- Once the step is created, it can be executed within a pipeline by adding it into an analysis file

Extending Package Functionality: bam_utils_lib.sh

```
execute_cnvkit()
{
    # Initialize variables
    local ref=$1
    local normalbam=$2
    local tumorbam=$3
    local step_outd=$4
    local cpus=$5

    # Activate conda environment
    conda activate cnvkit > ${step_outd}/conda_activate.log 2>&1 || exit 1

    # Run cnvkit
    cnvkit.py batch ${tumorbam} -n ${normalbam} -m wgs -f ${ref} -d ${step_outd} -p ${cpus} > ${step_outd}/cnvkit.log 2>&1 || exit 1

    # Deactivate conda environment
    conda deactivate > ${step_outd}/conda_deactivate.log 2>&1

    # Create file indicating that execution was finished
    touch ${step_outd}/finished
}
```


Extending Package Functionality: submit_bam_analysis.sh

```
get_pars_cnvkit()
{
    echo "$ref_$normalbam_$tumorbam_${step_outd}_$cpus"
```

Whole Pipeline Example

Analysis File

```
download_ega_norm_bam dortiz normal_prio 1 2048 10:00:00 jobdeps=none
download_ega_tum_bam dortiz normal_prio 1 2048 10:00:00 jobdeps=none
sort_norm_bam dortiz normal_prio 1 4G 10:00:00 jobdeps=afterok:download_ega_norm_bam
sort_tum_bam dortiz normal_prio 1 4G 10:00:00 jobdeps=afterok:download_ega_tum_bam
index_norm_bam dortiz normal_prio 1 1G 4:00:00 jobdeps=afterok:sort_norm_bam
index_tum_bam dortiz normal_prio 1 1G 4:00:00 jobdeps=afterok:sort_tum_bam
manta_somatic dortiz normal_prio 8 3G 6:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam
strelka_somatic dortiz normal_prio 8 6G 6:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam,
    afterok:manta_somatic
msisensor dortiz normal_prio 8 6G 5:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam
facets dortiz normal_prio 1 20G 4:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam
cnvkit dortiz normal_prio 8 8G 10:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam
ascatngs dortiz normal_prio 8 25G 12:00:00 jobdeps=afterok:index_norm_bam,afterok:index_tum_bam
platypus_germline dortiz normal_prio 1 4G 5:00:00 jobdeps=afterok:index_norm_bam
delete_bam_files dortiz normal_prio 1 1G 0:10:00 jobdeps=afterok:manta_somatic,afterok:strelka_somatic,
    afterok:msisensor,afterok:cnvkit,afterok:facets,afterok:ascatngs,afterok:platypus_germline
```

Pipeline

