WGBS methylation analysis pipeline

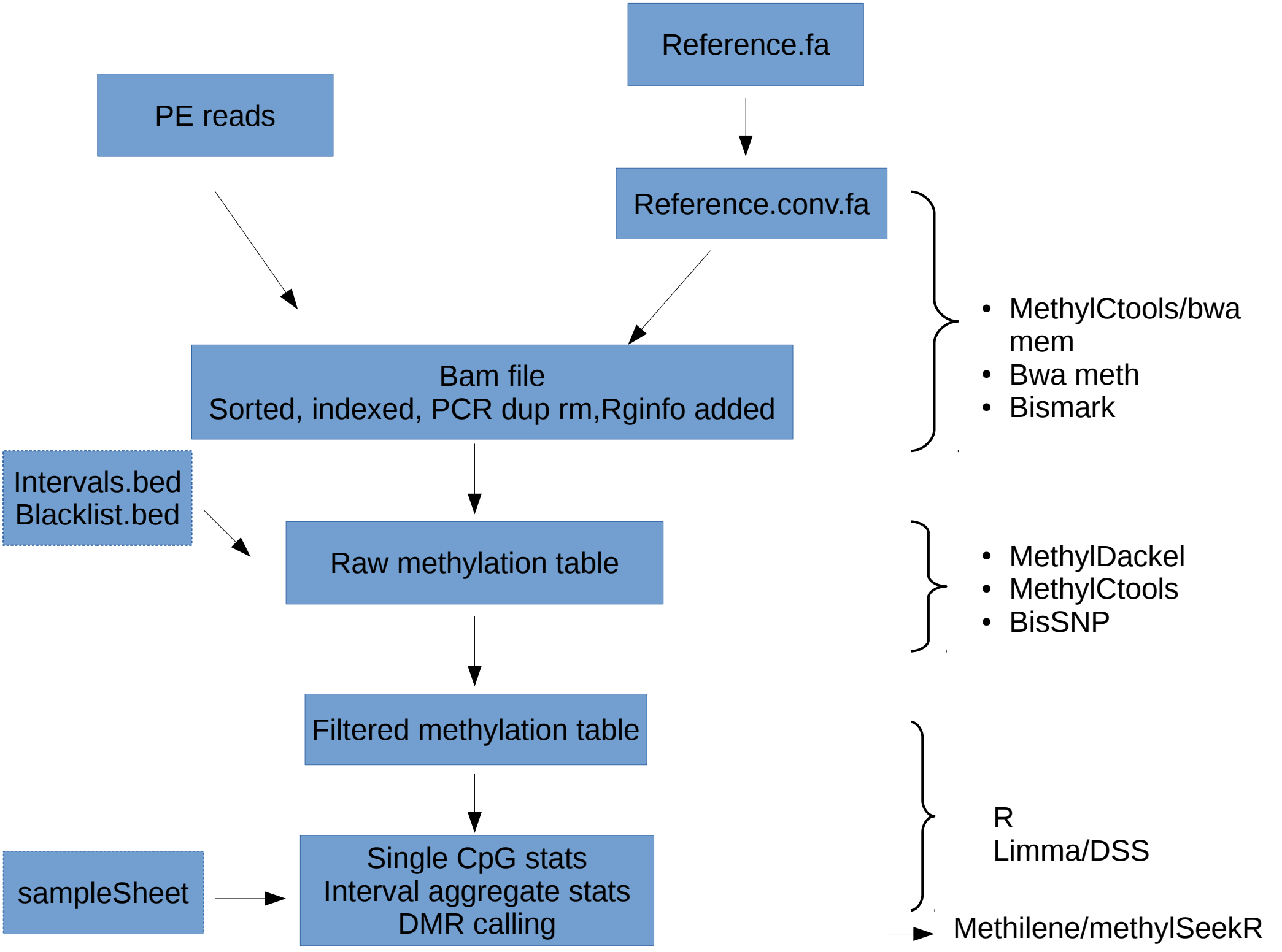
Katarzyna Sikora

V0.0.2

20170712

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Current status and open questions

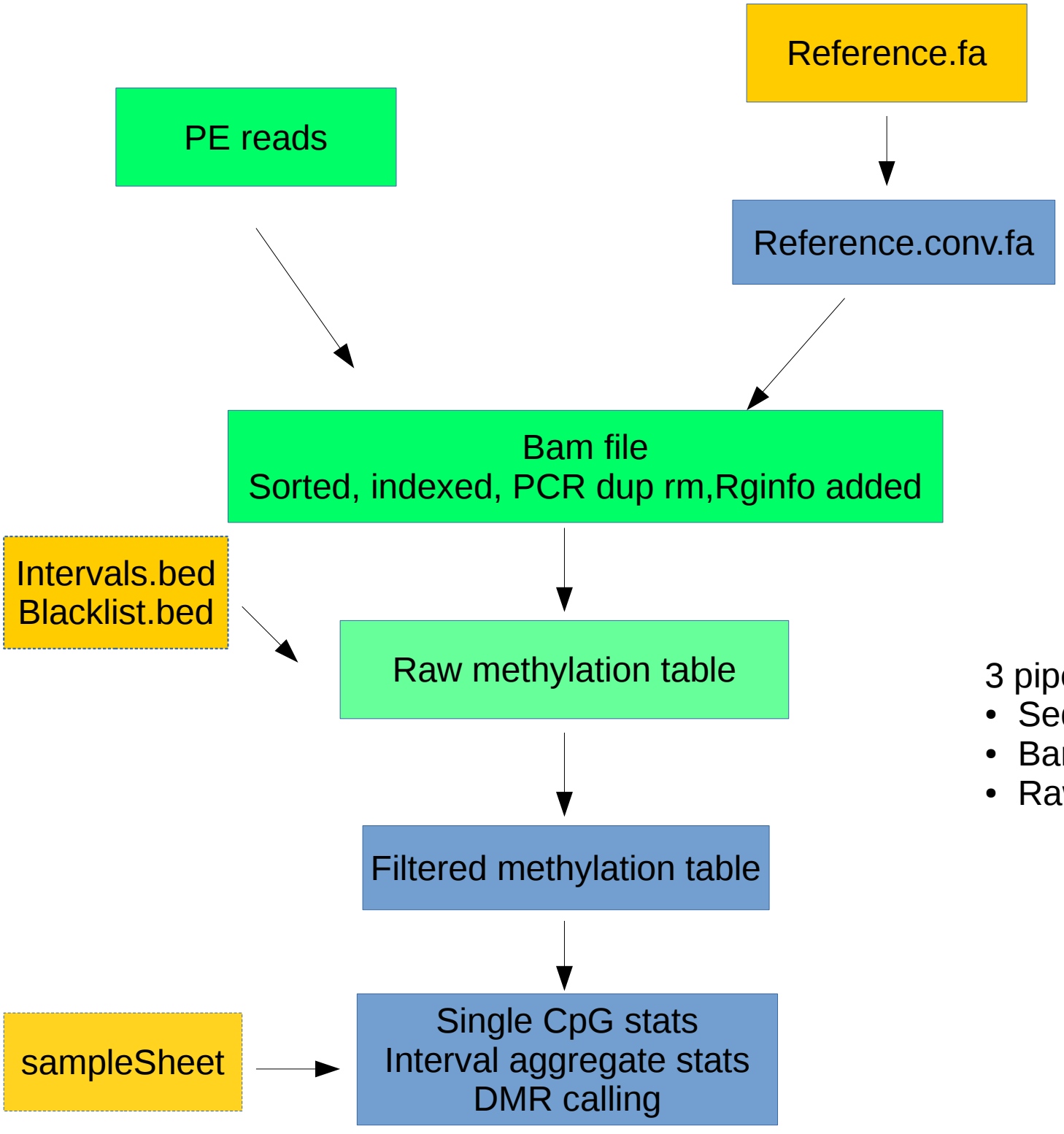

Reference.fa
PE reads
Reference.conv.fa
Bam file
Sorted, indexed, PCR dup rm,Rginfo added
MethylCtools/bwa mem
Bwa meth
Bismark
Intervals.bed
Blacklist.bed
Raw methylation table
MethylDackel
MethylCtools
BisSNP
Filtered methylation table
R
Limma/DSS
sampleSheet
Single CpG stats
Interval aggregate stats
DMR calling
Methilene/methylSeekR

Defaults

- Aligner: bwa-meth v.2016
- Extractor: MethylDackel-0.3.0
- Nthreads: 8
- Batchsize: 10
- R package for differential methylation: limma

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions



3 pipeline entry points

- Sequencing reads
- Bam files
- Raw methylation tables

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions

--readIn	input read folder
--bamIn	input bam folder
--methTabIn	input methylation table folder
--fqcIn	folder with fastqc.zip results for raw reads
--ref	path to indexed reference genome
--cref	Path to converted reference genome
--intList	target interval file(s)
--blackList	SNP black list
--sampleInfo	sample sheet
--wdir	output folder
--batchSize	number of samples to process in parallel
--numThr	number of threads to use per sample
--trimReads	adapter-trim and hardclip the reads
--convRef	BS-convert reference genome
--aligner	mapping software to use
--extractor	methylation extraction software to use
--stats	stats package to use
--DMRpg	DMR calling software to use
--touchOnly	only touch files
--target_tasks	Target tasks for the pipeline
--forcedtorun_tasks	Force up to date tasks

Pipeline entry point: specify one!

Additional input files

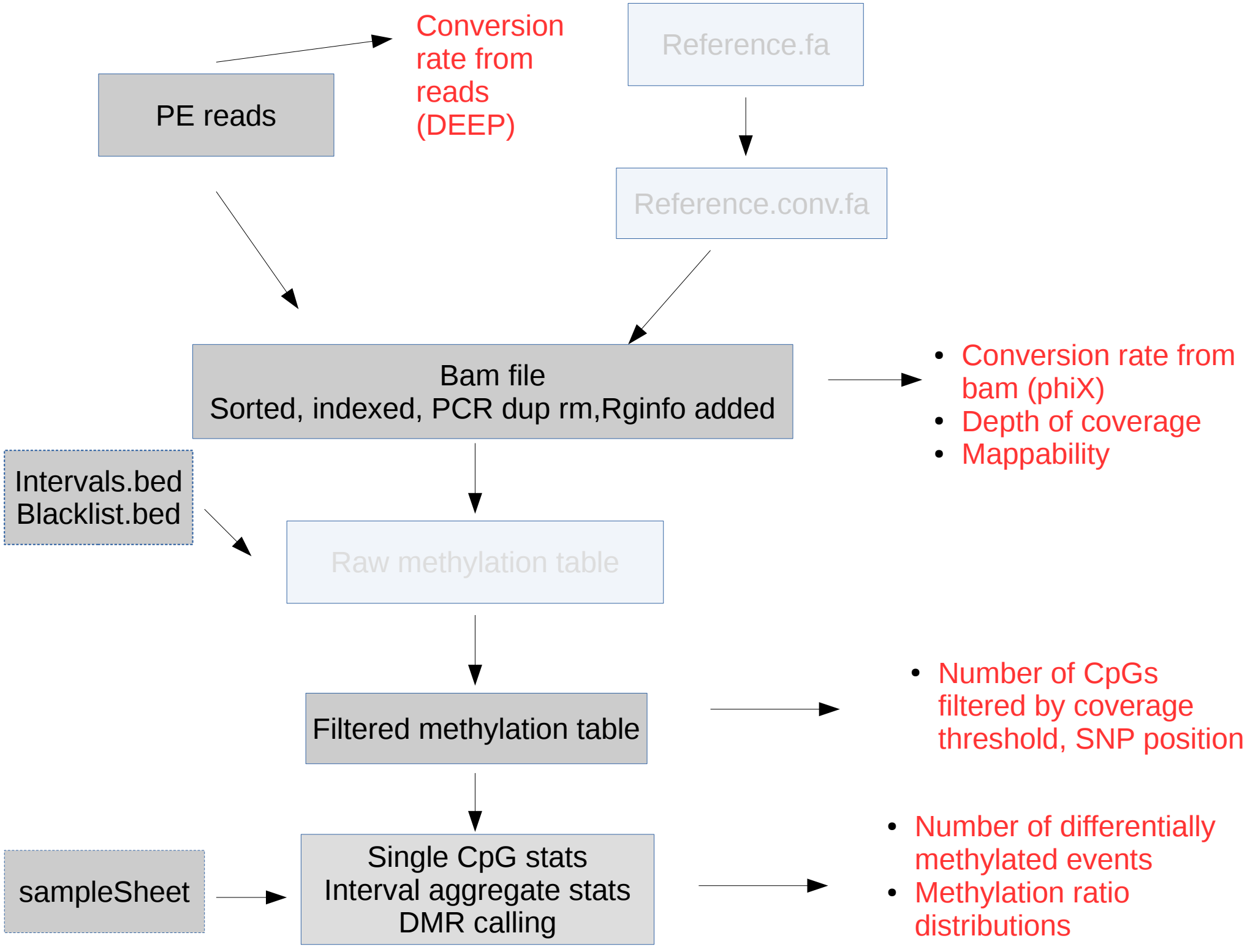
Cluster usage !!!

Software choices

Pipeline control

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions


PE reads
Conversion rate from reads (DEEP)
Reference.fa
Reference.conv.fa
Bam file
Sorted, indexed, PCR dup rm,Rginfo added
Conversion rate from bam (phiX)
Depth of coverage
Mappability
Intervals.bed
Blacklist.bed
Raw methylation table
Filtered methylation table
Number of CpGs filtered by coverage threshold, SNP position
sampleSheet
Single CpG stats
Interval aggregate stats
DMR calling
Number of differentially methylated events
Methylation ratio distributions

QC report (.Rmd, .pdf)

- Conversion rate (>95%, else warning)
- Mapping rate (>80%, else warning)
- Depth of coverage
  - Genome-wide
  - On 1Mln random CpGs in the genome
  - On target intervals in bed file/s, if provided
- Methylation bias – number of nucleotides ignored, mbias plots
- Number/percentage of CpG sites filtered (coverage, SNP)
- * Differential methylation analysis (single CpG, DMR): numbers and plots
-
- * not included in the report pdf, images stored in target folders, numbers in logs
- In main_output_folder/QC_metrics/QC_report.pdf e.g. /data/processing3/WGBS_pipe_example_OUT/QC_metrics/QC_report.pdf

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions

Ruffus framework

Table 1. A classification of modern pipeline frameworks

Syntax	Paradigm	Interaction	Example	Ease of Development	Ease of Use	Performance
Implicit	Convention	CLI	Snakemake, Nextflow, BigDataScript	★★★½	★★★★½	★★★
Explicit	Convention	CLI	Ruffus, bpipe	★★★★	★★★★	★★★
Explicit	Configuration	CLI	Pegasus	★★½	★★★	★★★★
Explicit	Class	CLI	Queue, Toil	★★½	★★★	★★★★★
Implicit	Class	CLI	Luigi	★★★	★★★½	★★★★★
Explicit	Configuration	Open Source Server Workbench	Galaxy, Taverna	★★★	★★★★½	★★★
Explicit	Configuration	Commercial Cloud Workbench	DNAnexus, SevenBridges	★★½	★★★★★	★★★★½
Explicit	Configuration	Open Source Cloud API	Arvados, Agave	★★★	★★★★	★★★★½

Leipzig, Bioinformatics 2016

Ruffus framework: example code

```
if ( args.aligner=='methylCtools' and not args.bamdir and not args.methtabdir ):
    readout=os.path.join(wdir,'conv_reads')
    if args.trimReads:
        @mkdir(readout)
        @transform(trim_reads,suffix('_R1.fastq.gz'),'_R12.conv.fastq.gz',output_dir=readout)
        def convert_reads(input_files,
                          output_file):
            ii1 = input_files[0]
            ii2 = input_files[1]
            oo = output_file
            from BSmapWGBS import methCT_convert_reads
            methCT_convert_reads(ii1,ii2,mCTpath,readout,mySession)
    else:
        @mkdir(readout)
        @transform(IN_files,suffix('_R1.fastq.gz'),'_R12.conv.fastq.gz',output_dir=readout)
        def convert_reads(input_files,
                          output_file):
            ii1 = input_files[0]
            ii2 = input_files[1]
            oo = output_file
            from BSmapWGBS import methCT_convert_reads
            methCT_convert_reads(ii1,ii2,mCTpath,readout,mySession)
```

http://www.ruffus.org.uk/

Modules

- Identify input files
- Trim reads (optional)
- Convert reference (optional)
- Map reads
- Collect QC metrics
- Extract and filter methylation tables
- Run single CpG stats
- Run aggregate stats per interval in bed file (optional)
- Run DMR calling and stats (optional)
- Output QC report

Cluster support: drmaa

```
sys.path.insert(0, "/data/boehm/sikora/tools/ruffus")
from ruffus import *
from ruffus.combinatorics import *
from ruffus.drmaa_wrapper import run_job, run_job_using_drmaa, error_drmaa_job
```

```
#from graphviz import Digraph
from PIL import Image
import string
```

```
#subprocess.check_output('export DRMAA_LIBRARY_PATH=/home/sikora/.local/lib/libdrmaa.so',shell=True)
subprocess.check_output('echo $DRMAA_LIBRARY_PATH',shell=True)
```

```
'/home/sikora/.local/lib/libdrmaa.so\n'
```

```
import drmaa
```

Cluster support: example function call

```
def single_CpG_limma(ii,sampleInfo,outdir,my_session):
    Rstat_cmd='/package/R-3.3.1/bin/Rscript --no-save --no-restore /home/sikora/works/Rscri
+ ' ' + sampleInfo + ' ' + ii
    print(Rstat_cmd)
    with open(os.path.join(outdir,"singleCpG_stats.out" ),'w') as stdoutF, open(os.path.joi
stderrF:
        try:
            stdout_res, stderr_res  = run_job(cmd_str       = Rstat_cmd,
                                              job_name      = 'sCpG',
                                              logger        = logging.getLogger(__name__),
                                              drmaa_session = my_session,
                                              run_locally   = False,
                                              working_directory = os.getcwd(),
                                              job_other_options = '-p bioinfo')
            stdoutF.write("".join(stdout_res))
            stderrF.write("".join(stderr_res))

        # relay all the stdout, stderr, drmaa output to diagnose failures
        except Exception as err:
            print("Single CpG stats error: %s" % err)
            raise
    print('Single CpG stats calculation complete')
    return
```

Overview

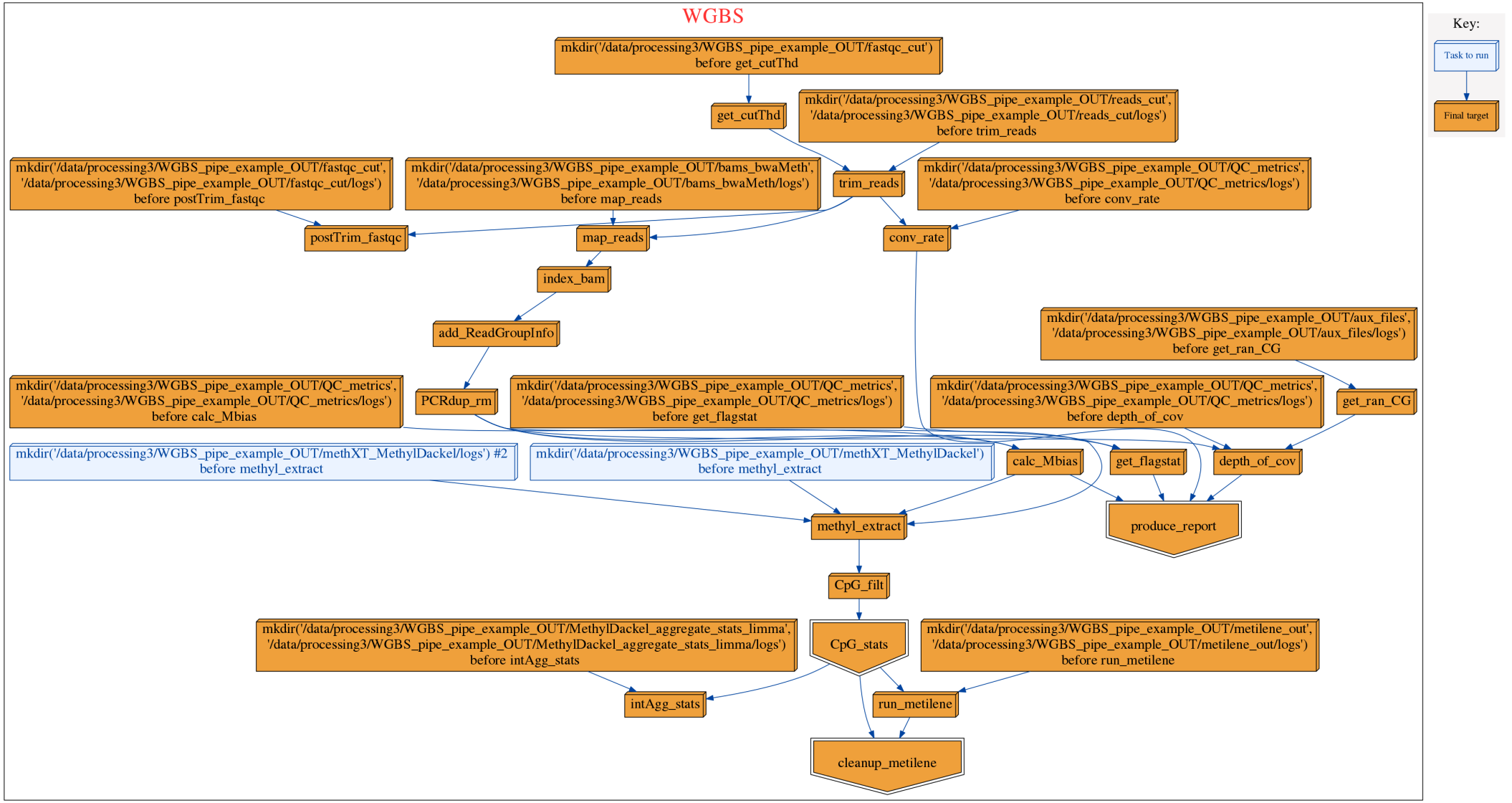
- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions

Starting from sequencing reads

WGBSpipe_example_wrapper.sh

- Exports paths to drmaa, R libraries
- Activates python 2.7 virtual env (conda)
- Python command to run on test data
-
- Make your own copy, edit file paths and source!

Example pipeline graph



Reference genome specification

- Ref genome: folder name of the genome under /data/repository/organisms without the '_ensembl' suffix, e.g. GRCz10, GRCm38
- If bwa-meth (default) or Bismark selected as aligner: cref will be taken from /data/repository/organisms, no need to specify
- If methylCtools as aligner: currently does not work with genomes under /data/repository/organisms; need to provide a genome fasta without spaces in chromosome names! Specified genome will be converted automatically, if cref not provided.

Required arguments

- Input read folder
- Reference genome
- Output folder
- Input folder with fastqc results for input reads

Optional arguments

- Trim reads (adapter removal, hard-trimming on 5' end)
- Interval list: bed file with regions of interest to calculate aggregate methylation values on
  - Specify multiple times if multiple bed files are to be analyzed
- Sample info: limma-style csv file with sample information. See example file: /data/processing3/WGBS_pipe_test_IN/example_sampleSheet.csv
  - If omitted, any differential methylation modules will be skipped

Output folder structure

- See example output: /data/processing3/WGBS_pipe_example_OUT
- Pipeline architecture – under main output folder:
  - pipelineGraph.png → tasks that will be executed
  - pipelinePrint.txt → input and output files for each task to be executed

Output files:

- From alignment and bam post-processing:
  - “bams_”+aligner_choice
  - .PCRrm.bam
    - Sorted, Read Group info added (for compatibility with GATK), PCR duplicates removed, indexed
    - View in IGV “bisulfite mode”; see https://software.broadinstitute.org/software/igv/interpreting_bisulfite_mode
- From methylation extraction: “methXT_”+extractor_choice
  - Raw tables: e.g. “_CpG.bedGraph”
  - Filtered tables (coverage, SNP blacklist etc.): “.CpG.filt2.bed”
- From differential methylation statistics on single CpG:
  - “singleCpG_stats_”+Rpackage_choice+extractor_choice
  - singleCpG.Rdata, limdat.LG.Rdata, PCA and density plots, top table of differentially methylated sites filtered for adjusted p value <0.05 (limdat.LG.CC.tT.FDR5.txt)
- From differential methylation statistics on genomic intervals:
  - extractor_choice+”_aggregate_stats_” + Rpackage_choice
  - bed_file_name+”.aggCpG.Rdata” , PCA and density plots, top table of differentially methylated sites filtered for adjusted p value <0.05 (bed_file_name+“.tT.FDR5.txt”)
- From metilene DMR calling:
  - metilene_out

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions

Logging and error handling

- Currently every sample gets stdout and sterr from every task; stored under 'logs' subfolder in every task folder
- One pipeline log with progress reported:
  - Example /data/processing3/WGBS_pipe_example_OUT/pipeline.log
- Errors from tools are re-raised (pipeline stops) and forwarded to pipeline.log
- Messages from multiple pipeline restarts are appended to the pipeline.log

Overview

- Pipeline design and software choices
- Input files, alternative entry points, arguments
- Quality metrics
- Implementation in Ruffus
- Example usage
- Logging and error handling
- Current status and open questions

Current status and limitations

- Pipeline tested for default software choices on example data
- Task modules currently not executable independently (no main)
- CpG only (not CHG etc.)
- Paired end reads only
- Metilene currently the only implemented DMR caller (future extension: methylSeekR)
- Two-sample comparison only

Questions/comments? - Contact:

- sikora@ie-freiburg.mpg.de
- +49 (0)761-5108-798 (while at the MPI)
- Katarzyna.otylia.sikora@gmail.com
- katarzyna.sikora@alumni.epfl.ch

- https://github.com/katsikora/WGBS_analysis_pipeline