Mayo GenomeGPS on iForge:
User Documentation and Standard Operating Procedures

Table of Contents

Pipeline architecture: 3 blocks and 6 run cases

	Alignment block	Realignment/recalibration block	Variant calling block
Case 1	*Perform only alignment* ANALYSIS=ALIGN RESORTBAM=NO	---	---
Case 2	---	*Perform only realignment/recalibration* ANALYSIS=REALIGN_ONLY SKIPVCALL=YES	---
Case 3	---	---	*Perform only variant calling* ANALYSIS=VCALL_ONLY SKIPVCALL=NO
Case 4	*Perform alignment + realignment/recalibration* ANALYSIS=REALIGN SKIPVCALL=YES RESORTBAM=NO		---
Case 5	---	*Perform realignment/recalibration + variant calling* ANALYSIS=REALIGN_ONLY SKIPVCALL=NO	
Case 6	*Invoke the entire pipeline: alignment + realignment/recalibration + variant calling* ANALYSIS=REALIGN SKIPVCALL=NO RESORTBAM=NO		

Runfile options

{OPTION1 | OPTION2} notation means you have a choice between the two options.

i/o

```
INPUTDIR=/full/path/to/folder/with/input/files
SAMPLEDIR=/full/path/to/folder/with/input/files
SAMPLEFILENAMES=/full/path/file.samplenames
OUTPUTDIR=/projects/mayo/GGPSresults/meaningful_output_dir_name
EMAIL=you@email
```

choose the run case

```
ANALYSIS={ ALIGN | REALIGN | REALIGN_ONLY | VCALL_ONLY }
SKIPVCALL={ YES | NO }
RESORTBAM={ YES | NO }    ## NO is always used if ANALYSIS=REALIGN
```

input data

```
PAIRED={ 1 | 0 }          ## 1 for paired-ended reads, 0 for single-ended reads
READLENGTH=100            ## or whatever the read length is
MULTISAMPLE={ YES | NO }
PROVENANCE={ MULTI_SOURCE | SINGLE_SOURCE }
SAMPLEINFORMATION=end2end multisamples
SAMPLENAMES=multisample1 multisample2  ## etcetera, as many words as there are samples
                                       ## these names must match those in samplenames file
SAMPLEID=sample_id_tag
SAMPLELB=hg19
SAMPLEPL=illumina
SAMPLEPU=sample_pu_tag
SAMPLESM=sample_sm_tag
SAMPLECN=Mayo
TYPE={ exome | whole_genome }
DISEASE=cancer
GROUPNAMES=NA
LABINDEX=-:-
LANEINDEX=1:2
```

tools to be used

```
JAVAMODULE=java-1.6
ALIGNER={ NOVOALIGN | BWA }
SORTMERGETOOL={ NOVOSORT | PICARD }
SNV_CALLER=GATK
SOMATIC_CALLER=SOMATICSNIPPER
```

preparatory block

```
BAM2FASTQFLAG={ YES | NO }        ## NO if input if fastq, or cases 2, 3 or 5 are invoked
BAM2FASTQPARMS=INCLUDE_NON_PF_READS=true
REVERTSAM={ 1 | 0 }
FASTQCFLAG={ YES | NO }
FASTQCPARMS=-t 15 -q
```

alignment block parameters

```
BWAPARAMS=-l 32 -t 16
NOVOPARAMS=-g 60 -x 2 -i PE 425,80 -r Random --hdrhd off -v 120 -c 16
BLATPARAMS=-w 50 -m 70 -t 90
```

realign/recalibrate block parameters

```
REALIGNPARMS=
MARKDUP=YES
REMOVE_DUP=NO
REORDERSAM=NO
REALIGNORDER=1
```

regions of interest

```
CHRINDEX=1:2:3:4         ## or whatever regions of the reference genome are of interest
```

variant calling block parameters

```
PEDIGREE=NA
VARIANT_TYPE=BOTH
UNIFIEDGENOTYPERPARMS=-maxAlleles 5
SNVMIX2PARMS=
SNVMIX2FILTER=-p 0.8
```

other parameters - DO NOT EDIT

```
EPILOGUE=/projects/mayo/scripts/epilogue.user
GENOMEBUILD=hg19
EMIT_ALL_SITES=YES
DEPTH_FILTER=0
TARGETTED=NO
```

paths to input output and tools - DO NOT EDIT

```
REFGENOMEDIR=/projects/mayo/reference
REFGENOME=mayo_novo/allchr.fa
DBSNP=mayo_dbsnp/hg19/dbsnp_135.hg19.vcf.gz
KGENOME=kGenome/hg19/kgenome.hg19.vcf
ONTARGET=/projects/mayo/reference/agilentOnTarget
NOVOINDEX=mayo_novo/allchr.nix
BWAINDEX=mayo_novo/allchr.fa
NOVODIR=/projects/mayo/builds/novocraft
BWADIR=/projects/mayo/builds/bwa-0.5.9
PICARDIR=/projects/mayo/builds/picard-tools-1.77
GATKDIR=/projects/mayo/builds/GATK-1.6-9
SAMDIR=/projects/mayo/builds/samtools-0.1.18
FASTQCDIR=/projects/mayo/builds/FastQC
SCRIPTDIR=/projects/mayo/scripts
SNVMIXDIR=/projects/mayo/builds/SNVMix2-0.11.8-r5
DELIVERYFOLDER=delivery
IGVDIR=IGV_BAM
```

pbs resources - DO NOT EDIT

```
PBSPROJECTID=bf0
PBSNODES=8
PBSTHREADS=16
PBSQUEUEEXOME=normal
PBSQUEUEWGEN=long
PBSCPUALIGNWGEN=240:00:00
PBSCPUALIGNEXOME=48:00:00
PBSCPUOTHERWGEN=240:00:00
PBSCPUOTHEREXOME=48:00:00
```

Samplenames file format

Note: the sample names must match those in runfile in all cases.

Single bam input

```
BAM:some_meaningul_samplename=/Full/path/to/inputfilename.bam
```

The .bam extension is obligatory.
Example: `/projects/mayo/scripts/config/BamInput.samplenames`

Single fastq input

```
FASTQ:samplename=/Full/path/to/reads1.fastq /Full/path/to/reads2.fastq
```

File names for left reads and right reads are separated by a space.
Example: `/projects/mayo/scripts/config/FastqInput_SingleSample.samplenames`

Multiple bam inputs

```
BAM:some_meaningul_samplename1=/Full/path/to/inputfilename1.bam
BAM:some_meaningul_samplename2=/Full/path/to/inputfilename2.bam
….  etcetera
```

The sample names field specified here will be used as sampletags during realign/recalibration, unless the user specifies BAM2FASTQFLAG=YES and PROVENANCE=SINGLE_SOURCE. In this case the tag will be whatever already exists in @RG lines of the bam file.
Examples: `/projects/mayo/scripts/config/BamInput_AlignedOnly_Multisamples.samplenames`
`/projects/mayo/scripts/config/BamInput_Multisample.samplenames`

Multiple fastq inputs

```
FASTQ:samplename1=/Full/path/to/reads11.fastq /Full/path/to/reads12.fastq
FASTQ:samplename2=/Full/path/to/reads21.fastq /Full/path/to/reads22.fastq
….  etcetera
```

File names for left reads and right reads are separated by a space. The order matters: specify left reads immediately after the "=" sign, and write the right reads (if any) on the same line.
Example: `/projects/mayo/scripts/config/FastqInput.samplenames`

Step-by-step instructions for setting up the runfile

Case 1: only alignment

This case covers both single sample and multisample inputs, exome and whole-genome data, bam and fastq input file format. The output is one aligned bam file per sample.

```
##i/o
```
- Step 1: copy file /projects/mayo/scripts/config/template.runfile to your home directory.
- Step 2: provide full path to the folder where input files are located by editing fields `INPUTDIR` and `SAMPLEDIR`.
- Step 3: provide full path to the samplenames file by editing field `SAMPLEFILENAMES`.
- Step 4: provide full path to the output folder by editing fields `OUTPUTDIR`.
- Step 5: set your email in the field `EMAIL`.

```
## choose the run case
```
- Step 6: `ANALYSIS=ALIGN`.
- Step 7: `RESORTBAM=NO`.
- Step 8: leave blank `SKIPVCALL=`

```
## input data
```
- Step 9: specify whether data are paired-ended (`PAIRED=1`) or single-ended (`PAIRED=0`).
- Step 10: specify read length.
- Step 11: if supplying multiple input files, are data independent (`MULTISAMPLE=NO`) or samples in the same experiment (`MULTISAMPLE=YES`)?
- Step 12: if the data are multisample, do they come from the same source and have the same sample names among the input files (`PROVENANCE=SINGLE_SOURCE`), or not (`PROVENANCE=MULTI_SOURCE`)? The option `PROVENANCE=SINGLE_SOURCE` in the runfile will cause the pipeline to derive the sample names directly from the @RG tags of the bams, if bam2fastq conversion is performed.
- Step 13: specify `SAMPLEINFORMATION`
- Step 14: list sample names separated by space in the field `SAMPLENAMES`. These must match the names in the samplenames file. If the data are from multiple sources and the sample names do not match among the input files, then just use `multisample1`, etc, like in the template.
- Step 15: edit other fields in this section as is appropriate for the experiment.

```
## tools to be used
```
- Step 16: choose aligner tool and sort/merge tool, as requested by the PI.

```
## preparatory block
```
- Step 17:
  - if the input files are bam, then set `BAM2FASTQFLAG=YES` and choose whether to perform picard revertsam before the conversion (`REVERTSAM=1`) or not (`REVERTSAM=0`)
  - if the input files are fastq, then set `BAM2FASTQFLAG=NO` and leave blank `REVERTSAM=`

Example: `/projects/mayo/scripts/config/FastqInput_AlignOnly.runfile`

Case 2: only realignment /recalibration

##i/o
- Step 1: copy file /projects/mayo/scripts/config/template.runfile to your home directory.
- Step 2: provide full path to the folder where input files are located by editing fields INPUTDIR and SAMPLEDIR
- Step 3: provide full path to the samplenames file by editing field SAMPLEFILENAMES
- Step 4: provide full path to the output folder by editing fields OUTPUTDIR
- Step 5: set your email in the field EMAIL

choose the run case
- Step 6: ANALYSIS=REALIGN_ONLY
- Step 7: SKIPVCALL=YES
- Step 8: set whether to resort the input bam (RESORTBAM=YES) or not (RESORTBAM=NO)

input data
- Step 9: specify whether data are paired-ended (PAIRED=1) or single-ended (PAIRED=0).
- Step 10: specify read length.
- Step 11: if supplying multiple input files, are data independent (MULTISAMPLE=NO) or samples in the same experiment (MULTISAMPLE=YES)?
- Step 12: leave blank PROVENANCE=
- Step 13: specify SAMPLEINFORMATION
- Step 14: list sample names separated by space in the field SAMPLENAMES. These must match the names in the samplenames file. If the data are from multiple sources and the sample names do not match among the input files, then just use multisample1, etc, like in the template.
- Step 15: edit other fields in this section as is appropriate for the experiment.

tools to be used
- Step 16: choose aligner tool and sort/merge tool, as requested by the PI.

preparatory block
- Step 17:
  - leave blank BAM2FASTQFLAG=
  - if the input bam files have already been realigned and recalibrated, then REVERTSAM=1; otherwise, REVERTSAM=0

regions of interest
- Step 18: list the regions of interest (CHRINDEX=1:2:3:4 or whatever).

Example: /projects/mayo/scripts/config/BamInput_RealignOnly_onAlignedOnly_Multisamples.runfile

Case 3: only variant calling

At present, the variant calling block treats all input files as independent, producing .vcf per chromosome, for each input file. This seemed the most meaningful way to separate out the variant calling block. Mayo's original pipeline merges samples after alignment, and from that point on analyzes a single bam file. If we receive multiple bam files, the presumption is that they are

a) either samples from different experiments, and thus should not be variant-called together,

b) or they are post-alignment bams from samples of the same experiment, and they need to be realigned/recalibrated to obtain a single bam.

```
##i/o
```

- Step 1: copy file /projects/mayo/scripts/config/template.runfile to your home directory.
- Step 2: provide full path to the folder where input files are located by editing fields `INPUTDIR` and `SAMPLEDIR`
- Step 3: provide full path to the samplenames file by editing field `SAMPLEFILENAMES`
- Step 4: provide full path to the output folder by editing fields `OUTPUTDIR`
- Step 5: set your email in the field `EMAIL`

```
## choose the run case
```

- Step 6: `ANALYSIS=VCALL_ONLY`
- Step 7: `SKIPVCALL=NO`
- Step 8: leave blank `RESORTBAM=`

```
## input data
```

- Step 9: specify whether data are paired-ended (`PAIRED=1`) or single-ended (`PAIRED=0`).
- Step 10: specify read length.
- Step 11: leave blank `MULTISAMPLE=`
- Step 12: leave blank `PROVENANCE=`
- Step 13: specify `SAMPLEINFORMATION`
- Step 14: list sample names separated by space in the field `SAMPLENAMES`. These must match the names in the samplenames file.
- Step 15: edit other fields in this section as is appropriate for the experiment.

```
## tools to be used
```

- Step 16: leave blank ALIGNER= and SORTMERGETOOL=

```
## preparatory block
```

- Step 17:
  - leave blank `BAM2FASTQFLAG=`
  - leave blank `REVERTSAM=`

```
## regions of interest
```

- Step 18: list the regions for variant calling as appropriate (`CHRINDEX=1:2:3:4` or whatever).

Example: `/projects/mayo/scripts/config/BamInput_VariantcallOnly_Multisample.runfile`

Case 4: alignment + realignment /recalibration

##i/o
- Step 1: copy file /projects/mayo/scripts/config/template.runfile to your home directory.
- Step 2: provide full path to the folder where input files are located by editing fields INPUTDIR and SAMPLEDIR
- Step 3: provide full path to the samplenames file by editing field SAMPLEFILENAMES
- Step 4: provide full path to the output folder by editing fields OUTPUTDIR
- Step 5: set your email in the field EMAIL

choose the run case
- Step 6: ANALYSIS=REALIGN
- Step 7: SKIPVCALL=YES
- Step 8: RESORTBAM=NO

input data
- Step 9: specify whether data are paired-ended (PAIRED=1) or single-ended (PAIRED=0).
- Step 10: specify read length.
- Step 11: if supplying multiple input files, are data independent (MULTISAMPLE=NO) or samples in the same experiment (MULTISAMPLE=YES)?
- Step 12: if the data are multisample, do they come from the same source and have the same sample names among the input files (PROVENANCE=SINGLE_SOURCE), or not (PROVENANCE=MULTI_SOURCE)? The option PROVENANCE=SINGLE_SOURCE in the runfile will cause the pipeline to derive the sample names directly from the @RG tags of the bams, if bam2fastq conversion is performed.
- Step 13: specify SAMPLEINFORMATION
- Step 14: list sample names separated by space in the field SAMPLENAMES. These must match the names in the samplenames file. If the data are from multiple sources and the sample names do not match among the input files, then just use multisample1, etc, like in the template.
- Step 15: edit other fields in this section as is appropriate for the experiment.

tools to be used
- Step 16: choose aligner tool and sort/merge tool, as requested by the PI

preparatory block
- Step 17:
  - if the input files are bam, then set BAM2FASTQFLAG=YES and choose whether to perform picard revertsam before the conversion (REVERTSAM=1) or not (REVERTSAM=0)
  - if the input files are fastq, then set BAM2FASTQFLAG=NO and leave blank REVERTSAM=

regions of interest
- Step 18: list the regions for variant calling as appropriate (CHRINDEX=1:2:3:4 or whatever).

Example: /projects/mayo/scripts/config/FastqInput_AlignRealign.runfile

Case 5: realignment /recalibration + variant calling

##i/o
- Step 1: copy file /projects/mayo/scripts/config/template.runfile to your home directory.
- Step 2: provide full path to the folder where input files are located by editing fields INPUTDIR and SAMPLEDIR
- Step 3: provide full path to the samplenames file by editing field SAMPLEFILENAMES
- Step 4: provide full path to the output folder by editing fields OUTPUTDIR
- Step 5: set your email in the field EMAIL

choose the run case
- Step 6: ANALYSIS=REALIGN_ONLY
- Step 7: SKIPVCALL=NO
- Step 8: set whether to resort the input bam (RESORTBAM=YES) or not (RESORTBAM=NO)

input data
- Step 9: specify whether data are paired-ended (PAIRED=1) or single-ended (PAIRED=0).
- Step 10: specify read length.
- Step 11: if supplying multiple input files, are data independent (MULTISAMPLE=NO) or samples in the same experiment (MULTISAMPLE=YES)?
- Step 12: leave blank PROVENANCE=
- Step 13: specify SAMPLEINFORMATION
- Step 14: list sample names separated by space in the field SAMPLENAMES. These must match the names in the samplenames file. If the data are from multiple sources and the sample names do not match among the input files, then just use multisample1, etc, like in the template.
- Step 15: edit other fields in this section as is appropriate for the experiment.

tools to be used
- Step 16: choose aligner tool and sort/merge tool, as requested by the PI.

preparatory block
- Step 17:
  - leave blank BAM2FASTQFLAG=
  - if the input bam files have already been realigned and recalibrated, then REVERTSAM=1; otherwise, REVERTSAM=0

regions of interest
- Step 18: list the regions for variant calling as appropriate (CHRINDEX=1:2:3:4 or whatever).

Example: /projects/mayo/scripts/config/BamInput_RealignVariantcall.runfile

Case 6: entire pipeline = alignment + realignment /recalibration + variant calling

```
##i/o
```
- Step 1: copy file /projects/mayo/scripts/config/template.runfile to your home directory.
- Step 2: provide full path to the folder where input files are located by editing fields `INPUTDIR` and `SAMPLEDIR`.
- Step 3: provide full path to the samplenames file by editing field `SAMPLEFILENAMES`.
- Step 4: provide full path to the output folder by editing fields `OUTPUTDIR`.
- Step 5: set your email in the field `EMAIL`.

```
## choose the run case
```
- Step 6: `ANALYSIS=REALIGN`.
- Step 7: `RESORTBAM=NO`.
- Step 8: `SKIPVCALL=NO`

```
## input data
```
- Step 9: specify whether data are paired-ended (`PAIRED=1`) or single-ended (`PAIRED=0`).
- Step 10: specify read length.
- Step 11: if supplying multiple input files, are data independent (`MULTISAMPLE=NO`) or samples in the same experiment (`MULTISAMPLE=YES`)?
- Step 12: if the data are multisample, do they come from the same source and have the same sample names among the input files (`PROVENANCE=SINGLE_SOURCE`), or not (`PROVENANCE=MULTI_SOURCE`)? The option `PROVENANCE=SINGLE_SOURCE` in the runfile will cause the pipeline to derive the sample names directly from the @RG tags of the bams, if bam2fastq conversion is performed.
- Step 13: specify `SAMPLEINFORMATION`
- Step 14: list sample names separated by space in the field `SAMPLENAMES`. These must match the names in the samplenames file. If the data are from multiple sources and the sample names do not match among the input files, then just use `multisample1`, etc, like in the template.
- Step 15: edit other fields in this section as is appropriate for the experiment.

```
## tools to be used
```
- Step 16: choose aligner tool and sort/merge tool, as requested by the PI.

```
## preparatory block
```
- Step 17:
  - if the input files are bam, then set `BAM2FASTQFLAG=YES` and choose whether to perform picard revertsam before the conversion (`REVERTSAM=1`) or not (`REVERTSAM=0`)
  - if the input files are fastq, then set `BAM2FASTQFLAG=NO` and leave blank `REVERTSAM=`

```
## variant calling block parameters
```
- Step 18: list the regions for variant calling as appropriate (`CHRINDEX=1:2:3:4` or whatever).

Example: `/projects/mayo/scripts/config/FastqInput_AlignRealignVariantcall.runfile`