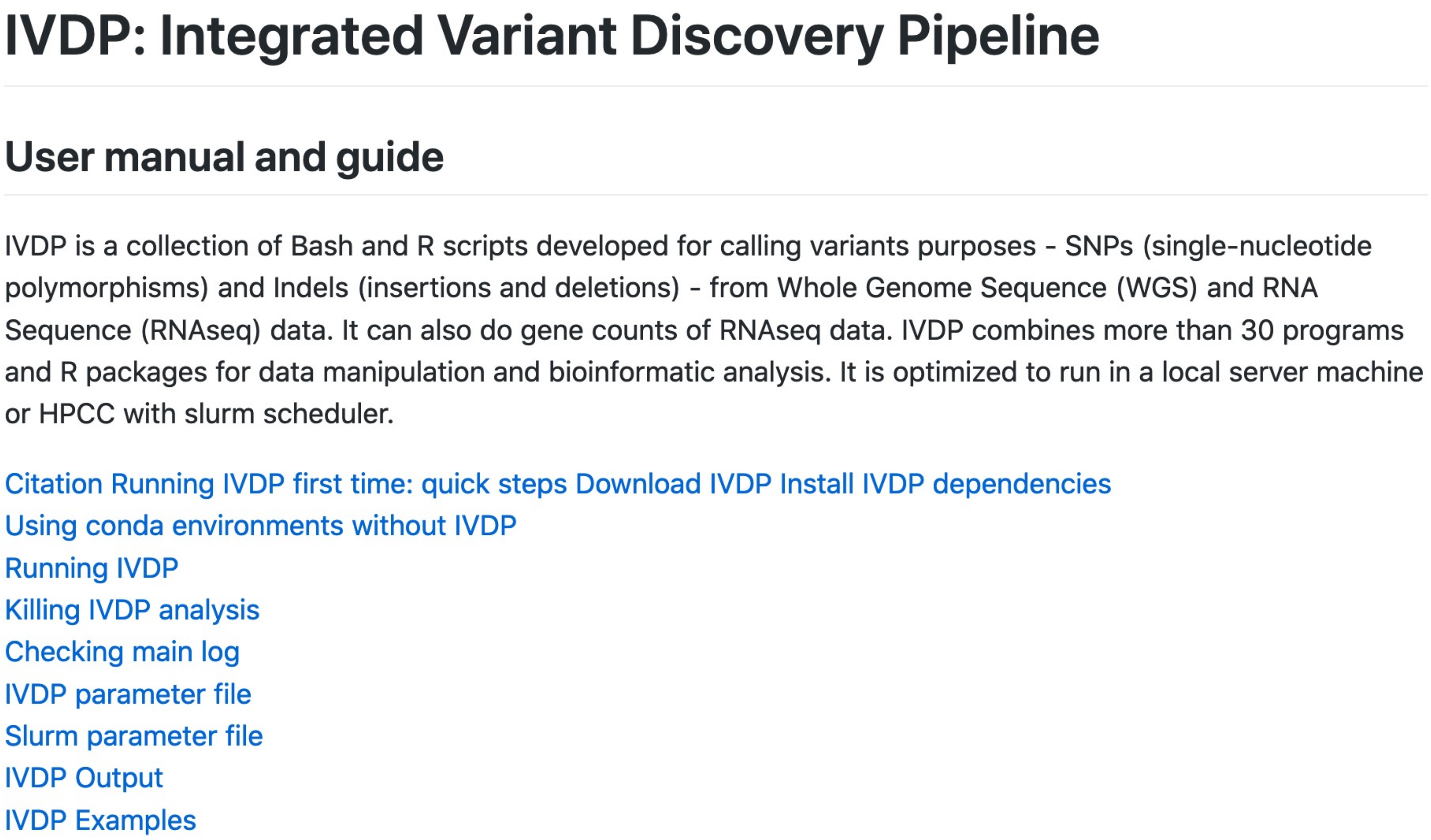
Introduction to Whole Genome Sequencing
IVDP: Integrated Variant Discovery Pipeline

Rodrigo Pelicioni Savegnago
Department of Animal Science

IVDP: Integrated Variant Discovery Pipeline

www.github.com/rodrigopsav/ivdp



IVDP: Integrated Variant Discovery Pipeline

User manual and guide

IVDP is a collection of Bash and R scripts developed for calling variants purposes - SNPs (single-nucleotide polymorphisms) and Indels (insertions and deletions) - from Whole Genome Sequence (WGS) and RNA Sequence (RNAseq) data. It can also do gene counts of RNAseq data. IVDP combines more than 30 programs and R packages for data manipulation and bioinformatic analysis. It is optimized to run in a local server machine or HPCC with slurm scheduler.

Citation Running IVDP first time: quick steps Download IVDP Install IVDP dependencies
Using conda environments without IVDP
Running IVDP
Killing IVDP analysis
Checking main log
IVDP parameter file
Slurm parameter file
IVDP Output
IVDP Examples

IVDP: Integrated Variant Discovery Pipeline

- Collection of Bash and R scripts;
- It integrates more than 30 programs and R packages;
- More than 50 different functions implemented;
- Install dependencies using conda:

```
./install_ivdp_dependencies.sh -f path/to/install/dependencies/programs
```

- Scalable for HPC (High-Performance Computing) that uses SLURM, Torque, HTCondor and other scheduler systems.

IVDP: Integrated Variant Discovery Pipeline

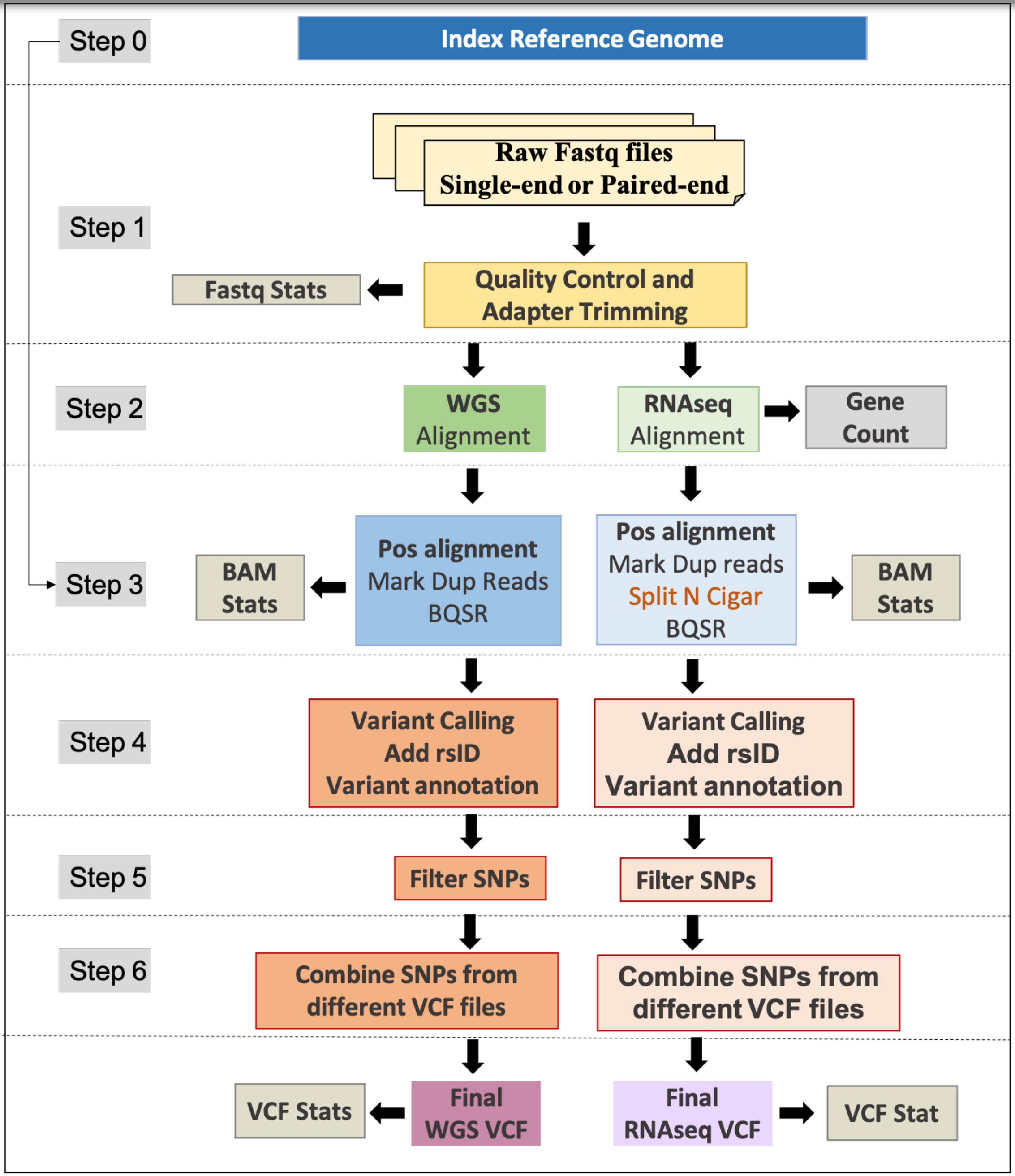
- To run in a local server:

```
./ivdpRun.sh -p parFile.txt
```

- To run on a HPCC with Slurm:

```
./ivdpRun.sh -p parFile.txt -c configSlurm.txt
```

IVDP: workflow



IVDP: parameter file

```
######################## INPUT DATA #####################
|INPUT_DIR=./examples/pe_wgs1
|OUTPUT_DIR=./output
|OUTPUT_NAME=PEex1a
|ANALYSIS=1   # ANALYSIS=1 (for WGS) or ANALYSIS=2 (for RNAseq)
|EXTENSION_PE1=_1.fastq.gz
|EXTENSION_PE2=_2.fastq.gz
####################### INPUT GENOME ####################
|ASSEMBLY=RefGenome
|REFERENCE=./examples/ref_gen/reference.fa
|VARIANTS=./examples/ref_gen/variants.vcf.gz
|ANNOTATION=./examples/ref_gen/annotation.gtf
####################### WORKFLOW STEPS ##################
|STEP_QC_TRIM=yes
|STEP_ALIGNMENT=yes
|STEP_MDUP=yes
|STEP_BQSR=yes
|STEP_VAR_CALL=yes
|STEP_GVCF_TO_VCF=yes
|STEP_VCF_FILTER=yes
#################### CHOOSE THE PROGRAMS ################
|ALIGNER_PROGRAM=bwa2
|CALLER_PROGRAM=gatk4GenomicsDBImport, bcftools
##################### GENERAL PARAMETERS #################
|CALL_BY_CHROM=yes   # Only for local machine
|BP_BY_CHROM=all       # Only for HPCC
|CHROM_SET=2
|MIN_READ_LENGTH=50
|MAX_READ_LENGTH=150
|MIN_DEPTH=3
|MAF=0.01
|MISSING=0.3
|COMBINE_VCF=partial
|THREADS=40        # Only for local machine
|BATCH=5            # Only for local machine
###########################################################
```

Block to define the input files and parameters

Block to define the reference genome / reference variants / reference annotation

Block to activate (yes) or deactivate (no) each step of IVDP

Block to choose aligner and variant callers

Block of general parameters: define trimming, VCF filtering, VCF combining, and parallelism parameters for variant caller

IVDP: parameter file

```
######################## INPUT DATA #####################

|INPUT_DIR=./examples/pe_wgs1

|OUTPUT_DIR=./output

|OUTPUT_NAME=PEex1a

|LIST_SAMPLES=none                    # OPTIONAL

|ANALYSIS=1   # ANALYSIS=1 (for WGS) or ANALYSIS=2 (for RNAseq)

|EXTENSION_PE1=_1.fastq.gz            # PAIRED-END READS 1

|EXTENSION_PE2=_2.fastq.gz            # PAIRED-END READS 2

|EXTENSION_SE=_.fastq.gz              # SINGLE-END READS
```

IVDP: list of samples

FILE NAMES

sample3_S1_L001_R1.fastq.gz
sample3_S1_L001_R2.fastq.gz
sample4_S1_L001_R1.fastq.gz
sample4_S1_L001_R2.fastq.gz
sample5_S1_L001_R1.fastq.gz
sample5_S1_L001_R2.fastq.gz
sample6_S1_L001_R1.fastq.gz
sample6_S1_L001_R2.fastq.gz

IVDP: list of samples

FILE NAMES

sample3_S1_L001_R1.fastq.gz
sample3_S1_L001_R2.fastq.gz
sample4_S1_L001_R1.fastq.gz
sample4_S1_L001_R2.fastq.gz
sample5_S1_L001_R1.fastq.gz
sample5_S1_L001_R2.fastq.gz
sample6_S1_L001_R1.fastq.gz
sample6_S1_L001_R2.fastq.gz

LIST_SAMPLES FILE

```
sample3_S1_L001 ANIMAL1
sample4_S1_L001 ANIMAL1
sample5_S1_L001 ANIMAL2
sample6_S1_L001 ANIMAL3
```

IVDP: parameter file

```
####################### INPUT GENOME ####################
|ASSEMBLY=RefGenome
|REFERENCE=./examples/ref_gen/reference.fa
|VARIANTS=./examples/ref_gen/variants.vcf.gz          # BQSR and rsID
|ANNOTATION=./examples/ref_gen/annotation.gtf   # Annotation,
                                                # Gene count
```

IVDP: parameter file

```
######################## WORKFLOW STEPS ###############
|STEP_QC_TRIM=yes
|STEP_ALIGNMENT=yes
|STEP_GENECOUNT=no        # Only for RNAseq: it requires ANNOTATION
|STEP_MDUP=yes
|STEP_BQSR=yes
|STEP_VAR_CALL=yes
|STEP_GVCF_TO_VCF=yes
|STEP_VCF_FILTER=yes
```

IVDP: parameter file

```
################# CHOOSE THE PROGRAMS ##############

|ALIGNER_PROGRAM=bwa2, bowtie2

(bbmap, bowtie, bowtie2, bwa, bwa2, gsnap, hisat2, star)


|CALLER_PROGRAM=gatk4GenomicsDBImport, bcftools

(bcftools, freebayes, freebayes-parallel, gatk4,

gatk4GenomicsDBImport, gatk4CombineGVCFs, platypus, varscan)


|GENECOUNT_PROGRAM=none      # Only for RNAseq

(none, htseq, featurecounts)
```

IVDP: parameter file

```
################# CHOOSE THE PROGRAMS ##############

|FEATURE_TYPE=exon     # Any feature from 3rd column of gtf file.


|MDUP_PROGRAM=sambamba

(sambamba, markdupspark, or picard)


|BQSR_PROGRAM=bqsrspark  (bqsrspark or bqsr)
```

IVDP: parameter file

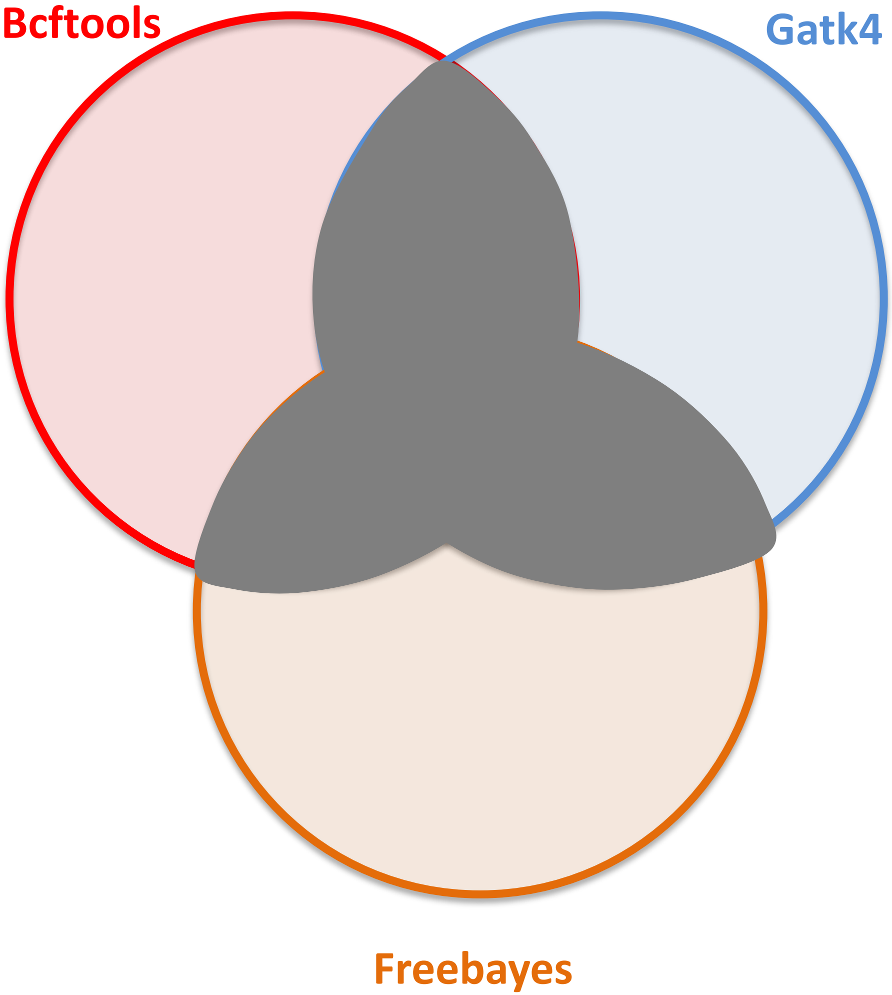
```
################ GENERAL PARAMETERS #############
|CALL_BY_CHROM=yes   # Only for local machine
|BP_BY_CHROM=all       # Only for HPCC
|CHROM_SET=29
|ADAPTER_TRIMMOMATIC=/home/work/rps/softwares/IVDP/program/03.qc_trim/adapters/TruSeq3-PE.fa:2:30:10:8:true
|MIN_READ_LENGTH=50
|MAX_READ_LENGTH=200
```

IVDP: parameter file

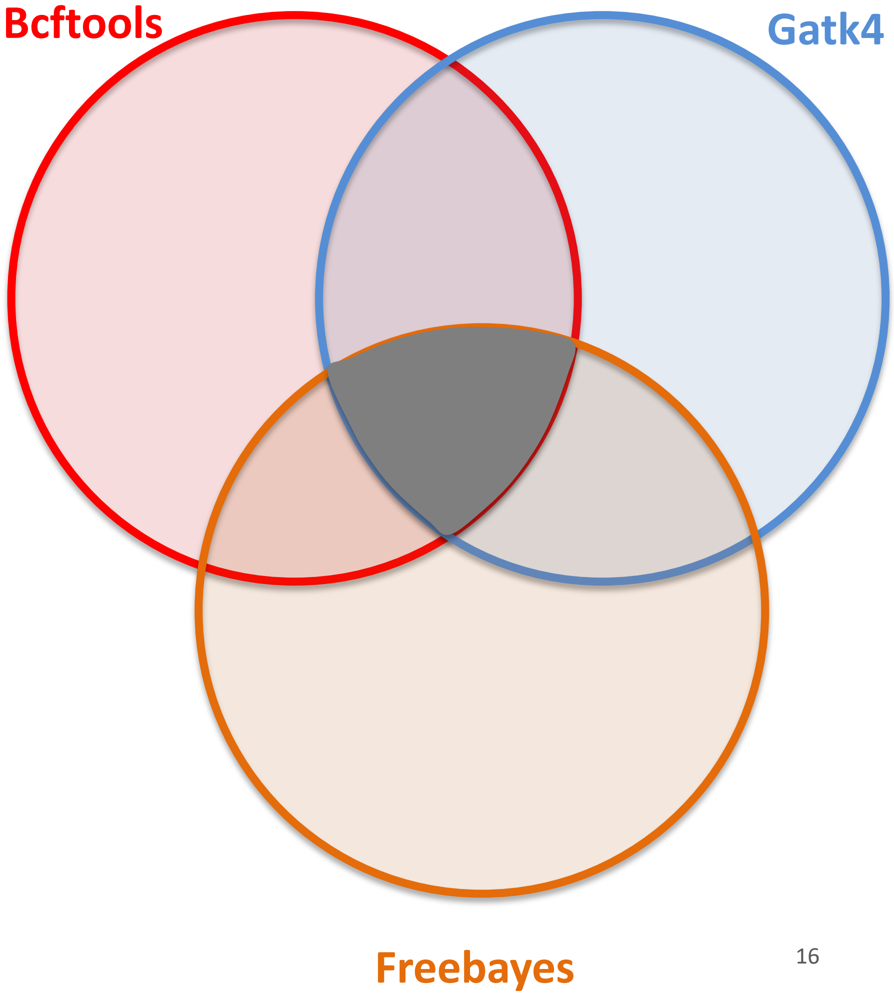
```
################ GENERAL PARAMETERS #############
|MIN_DEPTH=3
|MAF=0.01
|MISSING=0.3
|COMBINE_VCF=none        (none, partial, full)
|THREADS=40        # Only for local machine
|BATCH=5           # Only for local machine
```

IVDP: parameter file

COMBINE_VCF=partial



COMBINE_VCF=full



IVDP: slurm parameter file

```
TRIMMING_TIME=04:00:00
TRIMMING_CPU=16
TRIMMING_MEM=16G
```

Read trim step

```
ALIGNMENT_TIME=04:00:00
ALIGNMENT_CPU=16
ALIGNMENT_MEM=32G
```

Alignment step

```
GENECOUNT_TIME=04:00:00
GENECOUNT_CPU=8
GENECOUNT_MEM=8G
```

Gene count step

```
MDUP_TIME=04:00:00
MDUP_CPU=4
MDUP_MEM=16G
```

Mark duplicated reads step

```
BQSR_TIME=04:00:00
BQSR_CPU=8
BQSR_MEM=16G
```

BQSR step

```
CALLVARIANT_TIME=04:00:00
CALLVARIANT_CPU=4
CALLVARIANT_MEM=16G

GVCF_TO_VCF_TIME=04:00:00
GVCF_TO_VCF_CPU=4
GVCF_TO_VCF_MEM=16G
```

Variant calling step (and gVCF mode)

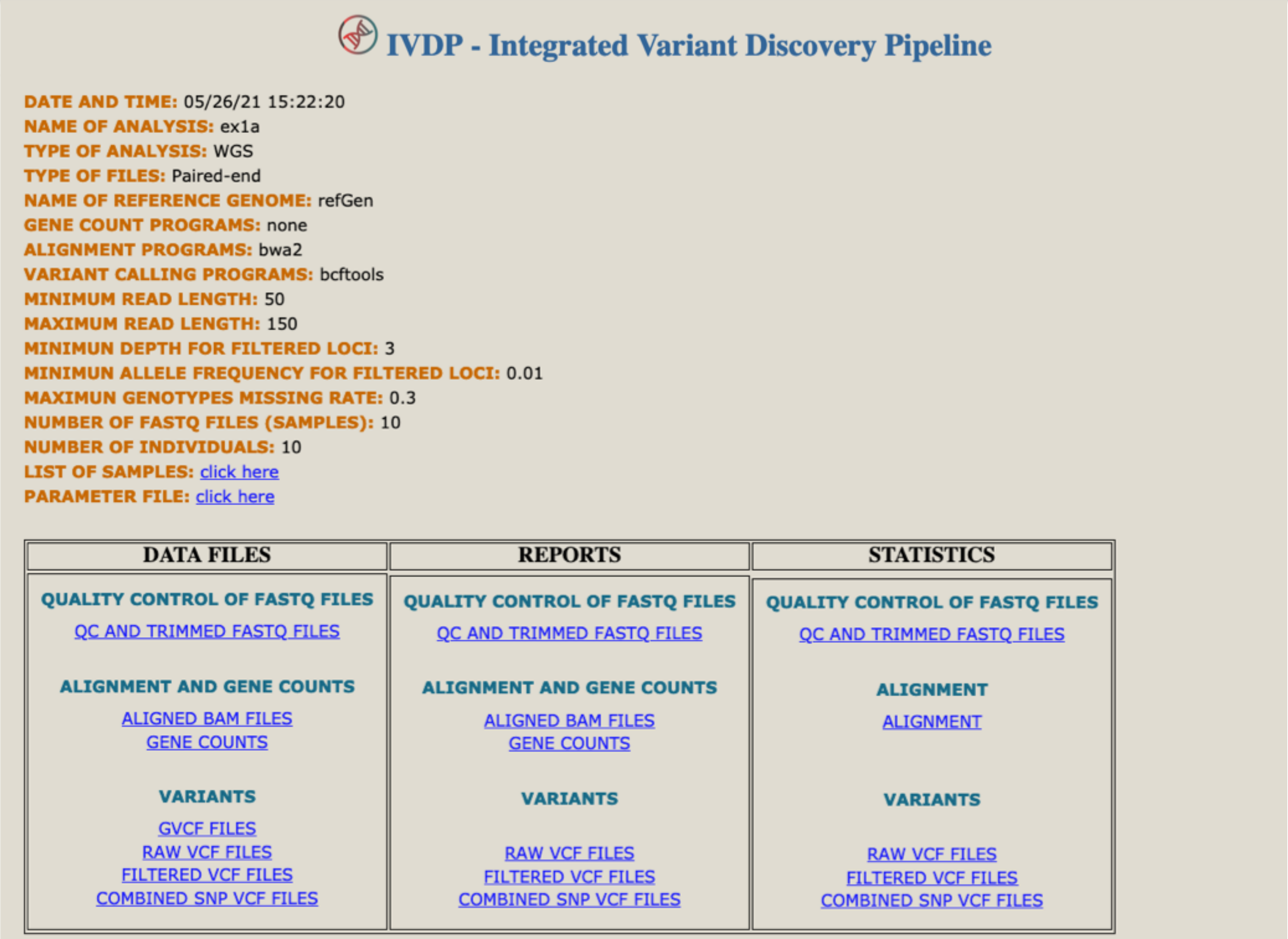
```
FILTERVCF_TIME=04:00:00
FILTERVCF_CPU=4
FILTERVCF_MEM=8G
```

VCF filtering step

IVDP: slurm parameter file

```
TRIMMING_TIME=04:00:00
TRIMMING_CPU=16
TRIMMING_MEM=16G

ALIGNMENT_TIME=04:00:00
ALIGNMENT_CPU=16
ALIGNMENT_MEM=32G

GENECOUNT_TIME=04:00:00
GENECOUNT_CPU=8
GENECOUNT_MEM=8G

MDUP_TIME=04:00:00
MDUP_CPU=4
MDUP_MEM=16G

BQSR_TIME=04:00:00
BQSR_CPU=8
BQSR_MEM=16G

CALLVARIANT_TIME=04:00:00
CALLVARIANT_CPU=4
CALLVARIANT_MEM=16G

GVCF_TO_VCF_TIME=04:00:00
GVCF_TO_VCF_CPU=4
GVCF_TO_VCF_MEM=16G

FILTERVCF_TIME=04:00:00
FILTERVCF_CPU=4
FILTERVCF_MEM=8G
```

IVDP: output



IVDP - Integrated Variant Discovery Pipeline

DATE AND TIME: 05/26/21 15:22:20
NAME OF ANALYSIS: ex1a
TYPE OF ANALYSIS: WGS
TYPE OF FILES: Paired-end
NAME OF REFERENCE GENOME: refGen
GENE COUNT PROGRAMS: none
ALIGNMENT PROGRAMS: bwa2
VARIANT CALLING PROGRAMS: bcftools
MINIMUM READ LENGTH: 50
MAXIMUM READ LENGTH: 150
MINIMUN DEPTH FOR FILTERED LOCI: 3
MINIMUN ALLELE FREQUENCY FOR FILTERED LOCI: 0.01
MAXIMUN GENOTYPES MISSING RATE: 0.3
NUMBER OF FASTQ FILES (SAMPLES): 10
NUMBER OF INDIVIDUALS: 10
LIST OF SAMPLES: click here
PARAMETER FILE: click here

DATA FILES	REPORTS	STATISTICS
QUALITY CONTROL OF FASTQ FILES QC AND TRIMMED FASTQ FILES	**QUALITY CONTROL OF FASTQ FILES** QC AND TRIMMED FASTQ FILES	**QUALITY CONTROL OF FASTQ FILES** QC AND TRIMMED FASTQ FILES
ALIGNMENT AND GENE COUNTS ALIGNED BAM FILES GENE COUNTS	**ALIGNMENT AND GENE COUNTS** ALIGNED BAM FILES GENE COUNTS	**ALIGNMENT** ALIGNMENT
VARIANTS GVCF FILES RAW VCF FILES FILTERED VCF FILES COMBINED SNP VCF FILES	**VARIANTS** RAW VCF FILES FILTERED VCF FILES COMBINED SNP VCF FILES	**VARIANTS** RAW VCF FILES FILTERED VCF FILES COMBINED SNP VCF FILES

Hands on

Working with IVDP

www.github.com/rodrigopsav/ivdp