Manual SIG-BIO
WGS outbreak proficiency test

Contact *jordy.coolen@gmail.com* & *casperjamin@gmail.com*

Introduction

Thank you for subscribing for the Whole-genome sequencing outbreak proficiency test 2019 initiated by SIG Bioinformatics in Medical Microbiology. With this document we would like to inform you on how to obtain the data, report the data, and stay up-to-date to future updates.

The goal of this proficiency test is to obtain an overview of current bioinformatics workflows used within Dutch centers for detecting outbreaks by using Whole-genome sequencing. Furthermore, the proficiency test will make it possible to compare your results to others and could be served as a first result for validation and accreditation.

For people who participated at the 5th SIG Bioinformatics in MM meeting. We would like to advert you for the fact that some of the templates presented during the meeting have changed due to practical reasons. Therefore, we would advise you all to follow this manual carefully.

Inhoud

Introduction 1
Case definition 3
Timeline 3
Checklist 3
1.0 Obtaining the data 4
1.1 Check md5sum (recommended but optional) 4
2.0 Description of pipeline_template.xlsx 5
2.1 the 01_Start_sheet 5
2.2 the 02_QC_rejection_parameters 5
2.3 the 03_cluster cutoffs 6
2.4 the 04_pipeline sheet 7
3.0 Description of KP_Results_sheet.xlsx and VRE_Results_sheet.xlsx 7
3.1 the 01_Start_sheet 7
3.2 the 02_SampleInfo 8
3.3 the 03_Sample_to_Sample 9
3.4 the 04_AMR_reporting 10
4.0 Data Handling 10
5.0 Report results 11
6.0 Frequently asked questions: 12
7.0 Contact: 12

Case definition

To outline the metadata involved in this proficiency test, we formulated the following case study:

"On Monday 07-10-19, the hospital infection prevention (IPC) team is informed regarding an increase in carriage of Vancomycin resistant Enterocci (VRE) on ward X. Simultaneously on ward Y, a high prevalence of multi-drug resistant *Klebsiella pneumoniae* (KP) was observed. The IPC requests to perform whole genome sequencing (WGS) for all isolates and would like to know if there is/are outbreaks of the same bacterium on these wards. The WGS data of 40 KP and 40 VRE samples have just arrived. The IPC would like you to report potential outbreak clusters based on the generated WGS results. Furthermore, the IPC would like to obtain a report on the antimicrobial resistance genes carried by these bacteria."

Timeline

To set a time window, the benchmark will start at 16th of October 2019 and we would like to receive the results by 25th of November 2019. We anticipate that this would give everyone enough time to be able to obtain, analyse, and return the results.

Checklist

- ☐ check MD5 sum to confirm file integrity (optional)
- ☐ Download datasets
- ☐ Analyse KP dataset
- ☐ Analyse VRE dataset
- ☐ fill in pipeline sheet
- ☐ fill in KP results sheet
- ☐ fill in VRE results sheets
- ☐ sign Consent_from
- **Request upload link from Jordy Coolen (jordy.coolen@gmail.nl)**
- ☐ upload FASTA files to surfdrive
- ☐ upload pipeline sheet to surfdrive
- ☐ upload Klebsiella result sheet to surfdrive
- ☐ upload VRE result sheet to surfdrive
- ☐ upload Consent_form

1.0 Obtaining the data

The data used for this proficiency test is pre-deposit data on the SRA. The samples provided are all anonymized and labelled according to species.

The 40 WGS *Klebsiella pneumoniae* (KP) samples of and WGS 40 *Enterococcus faecium* (VRE) samples are available via following surfdrive link:

https://surfdrive.surf.nl/files/index.php/s/b9HwlT0ESijcaEO

Also, the needed documents (pipeline_template.xlsx, KP_Results_sheet.xlsx, VRE_Results_sheet.xlsx) are available using the above link.

1.1 Check md5sum (recommended but optional)

To ensure, that the download is not corrupt, we provided a md5sum of the zipped KP and the VRE datasets. Furthermore, inside the .zip files there are also md5sum showing the independent md5sum of the independent .gz files.

What is md5 (https://en.m.wikipedia.org/wiki/Md5sum)?

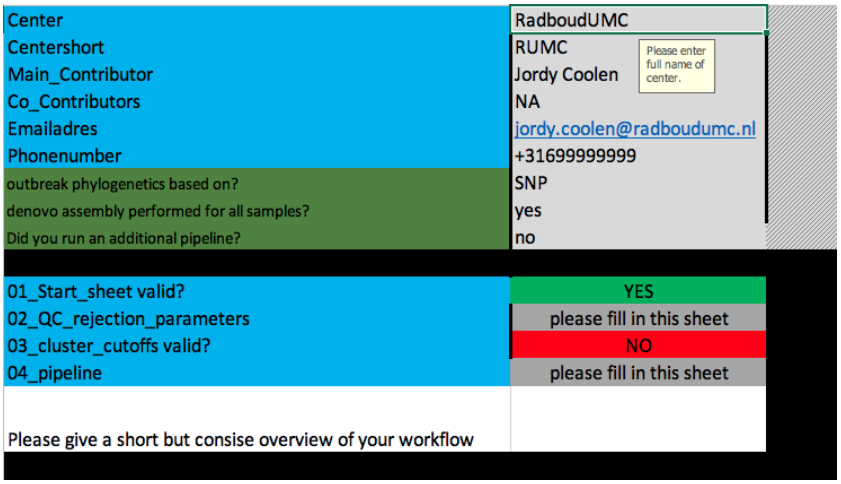
- MACOS users:
  - `md5 -r [file.ext]`
- Unix users:
  - md5sum [file.ext]
- Windows users:
  - Download this software tool (http://www.pc-tools.net/win32/md5sums/)
  - Drag and drop the file on the downloaded program
  - It will return the md5sum

2.0 Description of pipeline_template.xlsx

2.1 the 01_Start_sheet

This is the first sheet asking information from the participant to fill in. This information will mainly be used to contact participants and in case of problems to be able to correctly match results from centers. Please fill in with care.

Example:



2.2 the 02_QC_rejection_parameters

We would like to compile a list of generally used QC rejection criteria. Therefore, we would like you to fill in the QC rejection parameters sheet. One can use various parameters to determine insufficient data quality of a dataset. Please specify what kind of QC your center performs. Although we may have deemed the datasets as sufficient quality for this proficiency test, you might disagree. Still we would like you to use all datasets for this study and do not disregard any of the provided 2x40 samples. Furthermore, we assume that QC parameters are applicable for both species. However, if you would use different values as QC parameters per species, specify this in the SPECIES column.

Example:

	A	B	C	D	E
1	description or toolname	rejection criteria	SPECIES	comments	
2	fastqc	read quality lower (Q30) less than 70%		on average, more than 70% of the bases per read have to have a Q score of 30 or higl	example
3	denovo read coverage	coverage less than 30 fold		after denovo assembly, the assembly depth must be bigger than 30 fold on average	example
4	wgMLST alleles	less than 2000 alleles called	Enterococcus	for Enterococci, if less than 2000 alleles were found in a sample, this sample is disreg	example
5	N50 metric	lower than 15000			example

2.3 the 03_cluster cutoffs

The goal of the cluster cutoff sheet is to specify what cut-offs were used in this outbreak analysis. This may vary between species. Please be precise in which cut-offs were used and what the unit of measurement is. This can be based on SNPs, alleles different but also a normalized measure such as “alleles different / number of alleles compared”.

PLEASE NOTE: These values are interpreted as “up to and including” meaning if you have a cut-off of 15 SNPs and you observe 15 SNPs between 2 samples, these are considered to be clonally related.

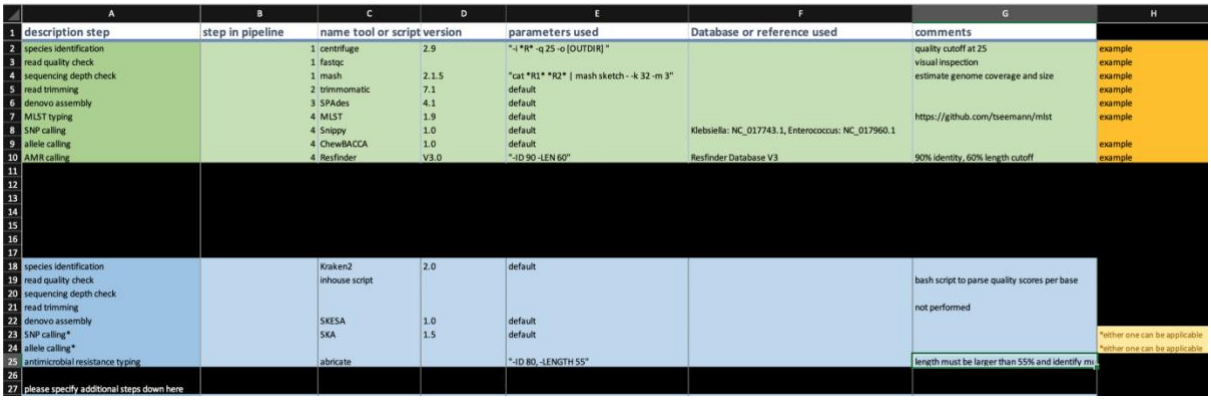
Example:

	A	B	C	D
1	cutoff	rejection criteria	unit	
2	hard cluster cutoff	5	SNPs	example
3	soft cluster cutoff	10	SNPs	example
4				
5	hard cluster cutoff	10	alleles different	example
6	soft cluster cutoff	15	alleles different	example
7				
8	hard cluster cutoff	0,001	alleles / number of alleles compared	example
9	soft cluster cutoff	0,005	alleles / number of alleles compared	example
10				
11				
12				
13	***Klebsiella pneumoniae***			
14	hard cluster cutoff	3	alleles different	
15	soft cluster cutoff	8	alleles different	
16				
17	***Enterococcus faecium***			
18	hard cluster cutoff	12	SNPs	
19	soft cluster cutoff	19	SNPs	
20				
21				
22	is this sheet sheet valid?	valid		

2.4 the 04_pipeline sheet

The goal of the pipeline sheet is to obtain all information on the bioinformatic workflow used per centre. Please specify the entire workflow used in this project. if additional methods were used, just specify this as well in this sheet. We consider the use of additional pipelines as a single big pipeline.

Example:



	A	B	C	D	E	F	G	H
1	description step	step in pipeline	name tool or script	version	parameters used	Database or reference used	comments	
2	species identification	1	centrifuge	2.9	"-i *R* -q 25 -o [OUTDIR] "		quality cutoff at 25	example
3	read quality check	1	fastqc				visual inspection	example
4	sequencing depth check	1	mash	2.1.5	"cat *R1* *R2* | mash sketch - -k 32 -m 3"		estimate genome coverage and size	example
5	read trimming	2	trimmomatic	7.1	default			example
6	denovo assembly	3	SPAdes	4.1	default			example
7	MLST typing	4	MLST	1.9	default		https://github.com/tseemann/mlst	example
8	SNP calling	4	Snippy	1.0	default	Klebsiella: NC_017743.1, Enterococcus: NC_017960.1		
9	allele calling	4	ChewBACCA	1.0	default			example
10	AMR calling	4	Resfinder	V3.0	"-ID 90 -LEN 60"	Resfinder Database V3	90% identity, 60% length cutoff	example
11								
12								
13								
14								
15								
16								
17								
18	species identification		Kraken2	2.0	default			
19	read quality check		inhouse script				bash script to parse quality scores per base	
20	sequencing depth check							
21	read trimming						not performed	
22	denovo assembly		SKESA	1.0	default			
23	SNP calling*		SKA	1.5	default			*either one can be applicable
24	allele calling*							*either one can be applicable
25	antimicrobial resistance typing		abricate		"-ID 80, -LENGTH 55"		length must be larger than 55% and identify mu	
26								
27	please specify additional steps down here							

3.0 Description of KP_Results_sheet.xlsx and VRE_Results_sheet.xlsx

These two sheets are intended to report the results of all samples. Each excelsheet consists of 4 worksheets; 01_Start_sheet, 02_SampleInfo, 03_Sample_to_Sample, and 04_AMR_reporting.

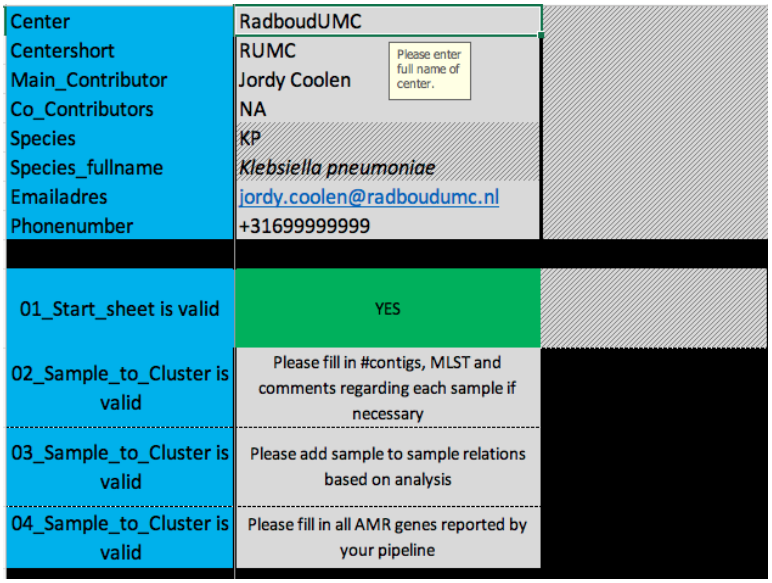
- ***KP_Results_sheet.xlsx*:** For reporting sample results of the KP (*Klebsiella pneumoniae*) dataset.
- ***VRE_Results_sheet.xlsx***: For reporting sample results of the VRE (*Enterococcus feacium*) dataset.

3.1 the 01_Start_sheet

This is the first sheet asking information of the participant to fill in. This information will mainly be used to contact participants and in case of problems to be able to correctly match results from centers. Please fill in with care.

Row 11 will check if this sheet is filled in correctly.
Rows 12-14 gives information about the other worksheets.

Example:



After completion please proceed to sheet 02_SampleInfo.

3.2 the 02_SampleInfo

This sheet provides the participants with limited info per sample.

Colum F: Assembly_name: Is the name of the result fasta that we would like to receive.

Please fill in Coverage and MLST types of each sample in corresponding columns (for both only numbers allowed).
If there are remarks on any of the samples please add a note in the comment column.

Example:

SampleName	R1	R2	Sequence platform	Assembly_name	Coverage	MLST	Comments
KP01	KP01_R1.fastq.gz	KP01_R2.fastq.gz	Illumina HiSeq 2500	KP01_RUMC.fasta	91	6	
KP02	KP02_R1.fastq.gz	KP02_R2.fastq.gz	Illumina HiSeq 2500	KP02_RUMC.fasta	79	5	
KP03	KP03_R1.fastq.gz	KP03_R2.fastq.gz	Illumina HiSeq 2500	KP03_RUMC.fasta	35	67	
KP04	KP04_R1.fastq.gz	KP04_R2.fastq.gz	Illumina HiSeq 2500	KP04_RUMC.fasta	78	91	
KP05	KP05_R1.fastq.gz	KP05_R2.fastq.gz	Illumina HiSeq 2500	KP05_RUMC.fasta	84	28	
KP06	KP06_R1.fastq.gz	KP06_R2.fastq.gz	Illumina HiSeq 2500	KP06_RUMC.fasta	31	15	
KP07	KP07_R1.fastq.gz	KP07_R2.fastq.gz	Illumina HiSeq 2500	KP07_RUMC.fasta	99	93	
KP08	KP08_R1.fastq.gz	KP08_R2.fastq.gz	Illumina HiSeq 2500	KP08_RUMC.fasta	94	71	
KP09	KP09_R1.fastq.gz	KP09_R2.fastq.gz	Illumina HiSeq 2500	KP09_RUMC.fasta	41	98	
KP10	KP10_R1.fastq.gz	KP10_R2.fastq.gz	Illumina HiSeq 2500	KP10_RUMC.fasta	21	61	
KP11	KP11_R1.fastq.gz	KP11_R2.fastq.gz	Illumina HiSeq 2500	KP11_RUMC.fasta	10	66	Very low covera

If completely filled please proceed to sheet 03_Sample_to_Sample.

3.3 the 03_Sample_to_Sample

This sheet has as goal to fill in all the sample to sample relations resulting from your outbreak analysis. Please report all the sample to sample relations that you will report to the Infection Prevention team (Please define relation based on the cut-offs that you provided in *2.3* 03_cluster cutoffs). The sheet contains 40 rows and 40 columns corresponding to all 40 samples per set. Only the grey part with 0 in the cells have to be filled. Each cell contains a dropdown menu with; 1 related, 0.5 maybe related, 0 not related.

The definition of these cut-offs are as follows:

Related: Infection prevention measures must be taken

Maybe related: Infection prevention measures must be taken

Not related: **No** infection prevention measures must be taken

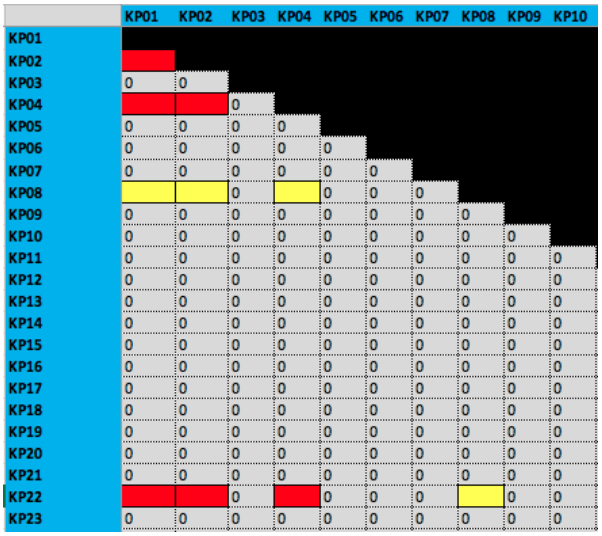
The filling process can take some time depending on the number of clusters found in the dataset but will afterwards make comparisons between centers more comprehensive. Please take your time to fill in the sheet.

Example:

We have a cluster of sample 1,2,4, and 22.

Sample 8 is closely/maybe related to this cluster.
Relations are: 1,2 1,4 1,22 2,4 2,22 4,22 with value 1 related.

Maybe relations are: 1,8 2,8 4,8 22,8 with value 0.5 maybe related.



	KP01	KP02	KP03	KP04	KP05	KP06	KP07	KP08	KP09	KP10
KP01										
KP02										
KP03	0	0								
KP04			0							
KP05	0	0	0	0						
KP06	0	0	0	0	0					
KP07	0	0	0	0	0	0				
KP08			0		0	0	0			
KP09	0	0	0	0	0	0	0	0		
KP10	0	0	0	0	0	0	0	0	0	
KP11	0	0	0	0	0	0	0	0	0	0
KP12	0	0	0	0	0	0	0	0	0	0
KP13	0	0	0	0	0	0	0	0	0	0
KP14	0	0	0	0	0	0	0	0	0	0
KP15	0	0	0	0	0	0	0	0	0	0
KP16	0	0	0	0	0	0	0	0	0	0
KP17	0	0	0	0	0	0	0	0	0	0
KP18	0	0	0	0	0	0	0	0	0	0
KP19	0	0	0	0	0	0	0	0	0	0
KP20	0	0	0	0	0	0	0	0	0	0
KP21	0	0	0	0	0	0	0	0	0	0
KP22			0		0	0	0		0	0
KP23	0	0	0	0	0	0	0	0	0	0

We enforce all to all relations, meaning that sample 1 in above example is in the same cluster as 2, 4, and 22. Therefore, all are related to each other.

This also holds for sample 8 in this example which is a 0.5 maybe related sample.

After completion please proceed to sheet 04_AMR_reporting.

3.4 the 04_AMR_reporting

This open field sheet lets you fill in AMR found resistance genes. We are aware that notation of genes may differ from center to center depending on the database and tools used. Still we ask you to fill in the notation as reported by your tool and database(s) (the tools and databases you used and noted in *2.4* 04_pipeline) Please use the cut-off that you applied and noted in "pipeline_template" excelsheet. Please report all AMR genes that you will report to your Infection Prevention team.

Example:

SampleName	KP01	KP02	KP03	KP04	KP05	KP06
Gene01	blaKPC-1	blaKPC-43		blaCTX-M-4		
Gene02	blaCMY-4			blaKPC-43		
Gene03				qnrA		
Gene04						
Gene05						
Gene06						
Gene07						
Gene08						
Gene09						
Gene10						
Gene11						

After completion of this sheet you can continue filling in the other species Results_sheet if you haven't filled it in yet.

4.0 Data Handling

All results reported by participants will only be communicated within the SIG bio members. For publication results will be anonymized to a random center number. Upon publication all centers will be mentioned as participating members of the study. (Please fill in Consent_form)

5.0 Report results

When you have completed all the analysis and completed the sheets it is time to upload the results to the surfdrive.

Please email to jordy.coolen@gmail.com and communicate your institute or center in order to receive your private upload link.

Files to upload:

1. Pipeline_template.xlsx
2. KP_Results_Sheet.xlsx
3. VRE_Results_Sheet.xlsx
4. 40 fasta files, *denovo* assembly results of KP (to see how to name the files see Results_sheet 02_SampleInfo Column Assembly_name)
5. 40 fasta files, *denovo* assembly results of VRE (to see how to name the files see Results_sheet 02_SampleInfo Column Assembly_name)
6. Signed Consent_form

Uploading the files is not difficult.

- open de private link
- drag and drop all files

6.0 Frequently asked questions:

- "I have a couple of samples that may be clonally related to the outbreak, how do I report this?"

  For samples that are maybe related you can use option 0.5 maybe related in the sheet. Please fill in all relations to all samples of cluster.

- "I would normally reject some samples included in this study; how do I proceed?"

  Add to the comment section of Result_sheet 02_SampleInfo that you normally would reject this sample. But for this study please include the sample in the complete outbreak analysis.

- "Sometimes I perform additional analyses next to our default pipeline, how do I report on this?"

  Please specify in the pipeline_template (see 2.4 04_pipeline), the complete workflow performed in this study. on the first page of this sheet you can specify if additional analyses were performed.

7.0 Contact:

jordy.coolen@gmail.com

casperjamin@gmail.com

SIGnl Bioinformatics in Medical Microbiology