National Institute for Public Health
and the Environment
Ministry of Health, Welfare and Sport

EURL Salmonella

Verification of the integrity of the raw data files (like md5sum)

Angela van Hoek

FASTQ sequence file format

- Text based
- Stores both raw sequence data and quality scores.
- Have become the standard for storing NGS data
- Compressed; fastq.gz

- Can contain up to millions of entries
- Can be several megabytes or gigabytes in size

- Used as input for a wide variety of secondary data analyses by you and others
- Important to ensure the data integrity of fastq files during and after transfer

MD5 hashes

- Fingerprint of a file
- Is encoded as a 128-bit digest
- The MD5 hash algorithm always produces the same output for the same given input
- Users can compare a hash of the source file with a newly created hash of the destination file to check that it is intact and unmodified.

- MD5 hashes are also commonly used with smaller strings when storing passwords, credit card numbers or other sensitive data in databases

Potential options to determine the MD5 hash with md5sum

- Without having to download anything, see for example;
  - The CertUtil is a pre-installed Windows utility, that can be used to generate hash checksums: *CertUtil -hashfile pathToFileToCheck MD5*
  - The procedure is explained in more detail in for example https://portal.nutanix.com/kb/11848/

- Install md5sum under windows, see for example;
  - https://appuals.com/use-md5sum-windows-command-prompt-environment/
  - https://www.winmd5.com/

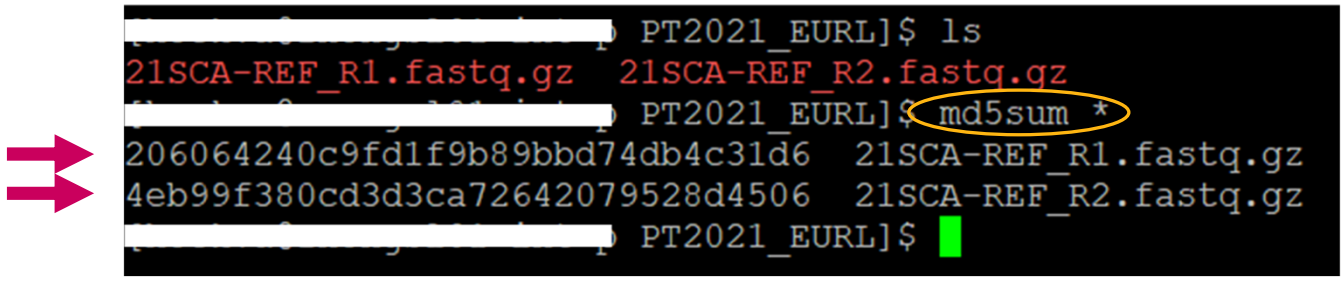
- At EURL-*Salmonella* we routinely use Linux, where we have md5sum installed.

Example; EURL-*Salmonella* PT2021, cluster analysis

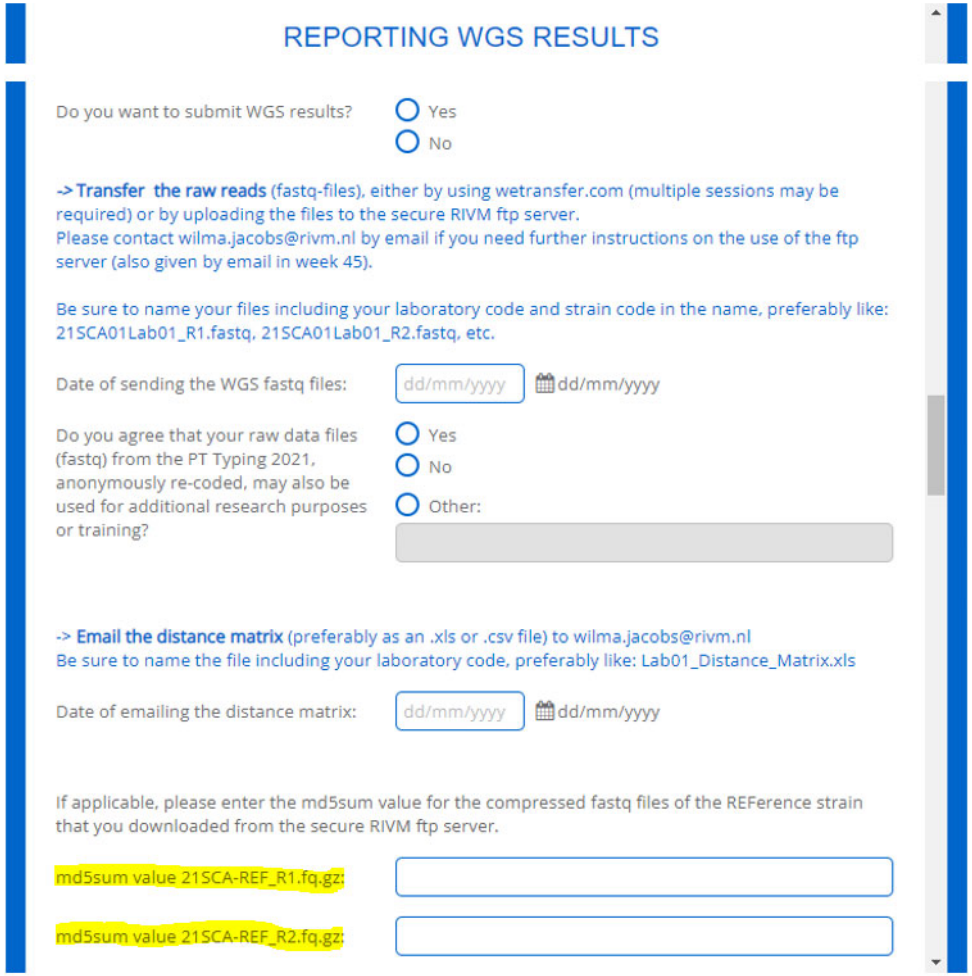
- 10 *S*. Enteritidis strains
  - 9 strains with stable and consistent cgMLST results
  - Technical duplicates: 21SCA06 and 21SCA09

Strain code	Serovar	ST	MLVA-profile	Origin
21SCA01	Enteritidis	11	2-9-9-4-2	Human
21SCA02	Enteritidis	183	2-11-9-3-1	Human
21SCA03	Enteritidis	183	2-11-9-3-1	Human
21SCA04	Enteritidis	11	3-10-4-4-1	Human
21SCA05	Enteritidis	1925	3-10-5-4-1	Human
21SCA06 [a)]	Enteritidis	11	3-10-4-4-1	Human
21SCA07	Enteritidis	3406	2-14-NA-7-NA	Human
21SCA08	Enteritidis	11	3-10-4-4-1	Human
21SCA09 [a)]	Enteritidis	11	3-10-4-4-1	Human
21SCA10	Enteritidis	11	1-10-7-3-2	Human

- Report per strain if a clustering match was found with the reference outbreak strain (21SCA-REF) shared via the RIVM sftp server



- Report form



REPORTING WGS RESULTS

Do you want to submit WGS results? ○ Yes ○ No

-> **Transfer the raw reads** (fastq-files), either by using wetransfer.com (multiple sessions may be required) or by uploading the files to the secure RIVM ftp server.
Please contact wilma.jacobs@rivm.nl by email if you need further instructions on the use of the ftp server (also given by email in week 45).

Be sure to name your files including your laboratory code and strain code in the name, preferably like: 21SCA01Lab01_R1.fastq, 21SCA01Lab01_R2.fastq, etc.

Date of sending the WGS fastq files: dd/mm/yyyy dd/mm/yyyy

Do you agree that your raw data files (fastq) from the PT Typing 2021, anonymously re-coded, may also be used for additional research purposes or training? ○ Yes ○ No ○ Other:

-> **Email the distance matrix** (preferably as an .xls or .csv file) to wilma.jacobs@rivm.nl
Be sure to name the file including your laboratory code, preferably like: Lab01_Distance_Matrix.xls

Date of emailing the distance matrix: dd/mm/yyyy dd/mm/yyyy

If applicable, please enter the md5sum value for the compressed fastq files of the REFerence strain that you downloaded from the secure RIVM ftp server.

md5sum value 21SCA-REF_R1.fq.gz:

md5sum value 21SCA-REF_R2.fq.gz:

Example; EURL-*Salmonella* PT2021, cluster analysis

	md5 checksum	
Labcode	21SCA_REF_R1.fq.gz	21SCA_REF_R2.fq.gz
Expected	**206064240c9fd1f9b89bbd74db4c31d6**	**4eb99f380cd3d3ca72642079528d4506**
1	206064240C9FD1F9B89BBD74DB4C31D6	4EB99F380CD3D3CA72642079528D4506
2		
6	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
7	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
10	e8e2aaff2830d4c2833d7ca203ecba01	4078d467927d6ae2573984a2a8e02a0c
11	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
12	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
14		
16	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
19	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
21	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
22	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
23		
24	NA	NA
26	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
27	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
30	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
34	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
35		

Files on the RIVM sftp server were uncompressed

Example; EURL-*Salmonella* PT2021, cluster analysis

	md5 checksum	
Labcode	21SCA_REF_R1.fq.gz	21SCA_REF_R2.fq.gz
Expected	**206064240c9fd1f9b89bbd74db4c31d6**	**4eb99f380cd3d3ca72642079528d4506**
1	206064240C9FD1F9B89BBD74DB4C31D6	4EB99F380CD3D3CA72642079528D4506
2		
6	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
7	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
10	e8e2aaff2830d4c2833d7ca203ecba01	4078d467927d6ae2573984a2a8e02a0c
11	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
12	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
14		
16	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
19	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
21	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
22	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
23		
24	NA	NA
26	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
27	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
30	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
34	206064240c9fd1f9b89bbd74db4c31d6	4eb99f380cd3d3ca72642079528d4506
35		

md5 checksum	
21SCA_REF_R1.fq	21SCA_REF_R2.fq
e8e2aaff2830d4c2833d7ca203ecba01	**4078d467927d6ae2573984a2a8e02a0c**

Md5 checksum PT 2020: 11/21 (52.4%)
PT 2021: 14/19 (73.7%)

EFSA One Health WGS System