D4.1 Repository of defined requirements

WP4: Enhancement & Consolidation of Digitalisation of Workflows

Author(s): Daniel Polzer

This project has received funding from the European Union's Health and Digital Executive Agency (HaDEA) under Grant Agreement No 101102440.

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

Copyright message

Document Information

Grant Agreement Number	101102440	Acronym	HERA 2
Full Title	Consolidation of WGS/RT-PCR and infrastructure processes in surveillance and outbreak investigation activities		
Call	EU4H-2022-DGA-MS-IBA-1		
Topic	EU4H-2022-DGA-MS-IBA-01-02		
Type of Action	EU4H-PJG		
Start Date	01.10.2022	**Duration (in months)**	42
HaDEA Project Officer	Marie de Looz-Corswarem		
Task	T4.1		
Work Package	4		
Date of Delivery	**Contractual** M11/August 2023	**Actual**	M11/August 2023
Nature	R — Document, report	**Dissemination Level**	PU — Public
Lead Beneficiary	AGES		
Lead Author	Daniel Polzer	**Organization**	AGES
Other authors	Adriana Cabal Rosel		
Reviewer(s)	KR, JS		

Document History

Version	Issue Date	Stage	Changes	Contributor
0.1	04.05.2023	First template generated in Steering Workshop		AGES, CIPH, EODY, NNK
0.2	01.08.23	Work in progress	Input from survey added	AGES
0.3	17.08.23	Work in progress	Executive summary; Conclusions and Graphs added	AGES
0.4	18.08.23	Ready for internal circulation	Added Table of content; list of abbreviations; next steps	AGES

0.5	21.08.23	Feedback integration	Integrate feedback from AGES members	AGES
1.0	29.08.2023	Quality-checked version		AGES

Table of Content

1. Executive Summary

In the course of the first project round, the focus on national infrastructure and capacity up-scaling enhancing WGS and/or RT-PCR to respond to the COVID-19 pandemic and future health threats set a solid but challenging basis for the enhancement of processes in national health agencies and public analytic facilities. The priorities are now to consolidate these new processes. The consolidation of WGS and RT-PCR activities aims to ensure the sustainable use and integration of enhanced infrastructure into routine surveillance and outbreak investigation activities, in synergy with relevant ongoing work at the international level is the key priority.

The COVID-19 pandemic aggravated the existing shortcomings that have to be tackled as soon as possible to improve the WGS workflow in terms of speed, efficiency, costs, and high-throughput analysis in general but also in terms of preparedness for future epidemics and pandemics.

This deliverable D4.1. "Repository of defined requirements" is based on a survey that the Consortium conducted by distributing it between partners and relevant stakeholders to evaluate different national needs and possible gaps in regards to digitization in WGS workflows.

2. List of abbreviations

EU	European Union
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
RT-PCR	Reverse Transcription Polymerase Chain Reaction
MLST	Multi-Locus-Sequence-Typing
CGE	Centre for Genomic Epidemiology
WGS	Whole Genome Sequencing
NCBI	National Centre for Biotechnology Information
ENA	European Nucleotide Archive
EURL	European Reference Laboratory
WHO	World Health Organization
ECDC	European Centre for Disease Prevention and Control
EFSA	European Food Safety Authority
RKI	Robert Koch Institute

3. Background of the project

The goal of this 42-month-long project is to establish stable WGS and RT-PCR processes in four relevant national institutions under the One Health approach with the aim of a more effective response to all current and future challenges during the outbreak, monitoring, and resolution of epidemics and other crises in public health. The cooperation of these professional public health institutions will enable an interdisciplinary approach to global and one health, aimed at mutual encouragement and the establishment of new highly specialized work processes.

4. Objectives and content

One of the objectives of work package 4 is to evaluate the different national needs and possible gaps regarding digitization in WGS workflows through a systematic and objective assessment of the relevance, efficiency, effectiveness, and sustainability of the existing systems. This deliverable is based on a survey completed by stakeholders in the four consortium countries.

5. Survey results on digitization

5.1. Digitisation in Austria

8 surveys were filled out from Austria. One survey was from an organisation with 21-50 members, another one from an organisation with 51-100 members, one from an organisation with 101-200 members. The other 5 were returned by members of an organisation with more than 500 members.

All survey participants from Austria report, that whole genome shotgun sequencing is performed in their organisation. One survey from Austria reported the use of amplicon-based sequencing.7 participants reported that they work with bacteria in their institute and 6 reported to work on viruses.

Three of the surveys report a sample volume regarding sequencing in the range of 1-10 samples per week, two in the range of 51-100 samples per week and another two reported a sample volume of over 200 samples per week. The final survey reported that they do not know the sample volume.

Only 1 survey reported the use of only long read methods, 4 surveys reported the use of only short read methods and 3 the use of both long- and short-reads methods.

3 surveys report to perform sequencing with the goal of obtaining consensus genomes, 7 surveys reported the aim of phylogenetic analysis, 7 survey participants perform sequencing to screen for certain mutations. 5 survey participants perform sequencing to detect antimicrobial resistances, three surveys report surveillance of antimicrobial resistances as a goal in sequencing experiments, 5 surveys report using sequencing for outbreak detection, 4 surveys report using sequencing for cluster allocation. One survey also reports using sequencing for genotyping.

7 of the surveys reported, that 1 to 3 people are working in bioinformatics. The 8th survey responded with "I do not know". Three surveys reported that only dedicated bioinformaticians in the organisation perform bioinformatic, one survey reported that only wet lab members perform bioinformatic analysis and four surveys report that bioinformatics is performed by both bioinformaticians and wet lab members.

Regarding bioinformatics the survey participants were asked about tools used for quality control and trimming. 5 participants reported the use of FastQC (a tool for quality control of short reads), 4 surveys reported the use of Trimmomatic for trimming of short reads.

5 surveys reported on quality criteria that decide whether the quality of the reads allows further analysis. Two of these surveys reported the Phred score as a criterium one reported the warning messages of the tool FastQC as a parameter, while the other two reported criteria for assembled genomes as a criterion. These two surveys listed the coverage of the contigs, the number of contigs as well as the N50 value and the estimated genome size. Additionally for samples which are analysed by core-genome Multi-locus-sequence-typing (MLST) the percentage of good targets and the percentage of missing values were reported.

Participants were asked about the tools they use for genome assembly and annotation. 2 participants reported BWA, 1 participant bowtie, 4 participants SPAdes, 2 participants Unicycler, 1 participant CGE (Centre for Genomic Epidemiology), 1 participant KRAKEN, 1 participant Guppy, 1 participant SKESA, 1 participant Flye, and 1 participant Geneious Prime. 2 participants did not know which tools were used for these steps.

Regarding Data visualization 6 participants reported the use of SeqSphere+, 1 participant mentioned figtree and the Biopython package for Python, one participant mentioned Geneious prime.

Participants noted that one advantage of SeqSphere and Geneious Prime are their graphical user interface (GUI) and as a result the ease to use. SPAdes on the other hand was reported to be fast.

SeqSphere+ users use it to visualize phylogenetic trees and relatedness of samples. Geneious Prime users report using it for visualisation of phylogenetic trees, assemblies, alignments, and variants.

The survey participants were asked which database they use to obtain reference sequences. The NCBI database was reported 3 times, Gisaid was reported twice, BIGSDB Pasteur, CGE tools, GenBank and SeqSphere+ were mentioned once each.

5 participants reported to upload genomes to GISAID, 3 to the NCBI databases, and one participant reported to upload to the ENA.1 participant reported to upload to no database. The participants reported, that they share data with European reference laboratories (EURLs) (2 out of 8), ECDC,EFSA,RKI and the Centre of Excellence FoodHub (1 out of 8 each). 4 Participants reported to share geographical data, 5 to share the collection date, 4 share the type of organism and one participant share the sample matrix with these external organisations. 1 participant reported sharing less than 10% of their data, 2 participants between 10 and 50%, 1 participant between 51 and 90% with external institutes or databases.

4 participants reported to share genomes in the FASTA format, 3 in the Fastq format, and 1 participant reported to also share phylogenetic data in the Newick format as well as coverageplots internally.

Participants were asked if they use any bioinformatic workflow management tools. 1 participant reported the usage of Bash scripts, Nextflow and the pyrpipe package for python. Another participant reported the use of the Galaxy web tool.

5.2. Digitisation in Croatia

6 surveys were filled out from Croatia. One survey was from an organisation with 21-50 members, another one from an organisation with 101-200 members, 3 from an organisation with 201-500 members. The last survey was returned by a member of an organisation with more than 500 members.

4 survey participants from Croatia report, that whole genome shotgun sequencing is performed in their organisation. The other 2 reported that they do not perform sequencing but that it was planned to be implemented within the next three years.3 participants reported that they work with bacteria in their institute, 1 reported to work on parasites and 4 reported to work on viruses.

1 of the surveys report a sample volume regarding sequencing in the range of 1-10 samples per week, 1 in the range of 51-100 samples per week , 1 reported a range of 101-200 samples per week. The last responder that performed sequencing in their lab, did not know how many samples they are sequencing.

Only 1 survey reported the use of only long read methods, 2 surveys reported the use of only short read methods and 1 the use of both long- and short-reads methods.

2 surveys report to perform sequencing with the goal of obtaining consensus genomes, 5 surveys reported the aim of phylogenetic analysis, 1 survey participants perform sequencing to screen for certain mutations. 3 survey participants perform sequencing to detect antimicrobial resistances, 2 surveys report surveillance of antimicrobial resistances as a goal in sequencing experiments, 4 surveys report using sequencing for outbreak detection, 2 surveys report using sequencing for cluster allocation.

1 of the surveys reported, that nobody in their institute is working in bioinformatics, the other 5 report 1-3 people working in this field .2 surveys reported that only dedicated bioinformaticians in the organisation perform bioinformatics, and 4 surveys reported that only wet lab members perform bioinformatic analysis in their institute.

Regarding bioinformatics the survey participants were asked about tools used for quality control and trimming. 2 participants reported the use of FastQC (a tool for quality control of short reads), 2 surveys reported the use of Trimmomatic for trimming of short reads.1 survey reported cutapt, porechop, SeqTK, Trimmgalore and MEGA11

2 surveys reported on quality criteria that decide whether the quality of the reads allows further analysis. This surveys reported the Phred score as a criterium. The other responder also takes the read length into account.

Participants were asked about the tools they use for genome assembly and annotation. 2 participants reported BWA, bowtie, SPAdes, Unicycler, Kraken and PROKKA respectively.

1 participant reported the usage of ABySS, Canu, velvet CARD, SKESA, Flex, MEGA 11, participant MTBseq, TB profiler and PhyResSe respectively.

Regarding Data visualization 1 participants reported the use of R Studio, Freyja, Geneious prime, Bionumerics and Chromas respectively.

Speed, reliability, and ease of use were reported as advantages for this software.

The survey participants were asked which database they use to obtain reference sequences. The NCBI database was reported 3 times, GISAID, ResFinder, CARD, Galaxy, Geneious, CGE DTU and ENA were reported once each.

3 participants reported to upload genomes to GISAID, 2 to the NCBI databases, and 1 participant reported to upload to the ENA.2 participant reported to upload to no databases.

The participants reported that they share data with European reference laboratories (EURLs) universities and local public health institutes.

3 Participants reported to share geographical data, 3 to share the collection date, 3 share the type of organism and 2 participants share the sample matrix with these external organisations.

1 participant reported sharing all of their sequencing data with external institutes or databases.

3 participants reported to share genomes in the FASTA format. No other format was reported.

Participants were asked if they use any bioinformatic workflow management tools. 1 participant reported the usage of Bash scripts, Nextflow and the Galaxy web tool.

5.3. Digitisation in Hungary

7 surveys were filled out from Hungary. 2 surveys were from an organisation with 1 to 10 members, 3 were from an organisation with 11-20 members, 1 from an organisation with 21-50 members, The last survey was returned by a member of an organisation with 101-200 members.

4 survey participants from Hungary report, that shotgun sequencing is performed in their organisation 3 surveys report the use of amplicon sequencing.3 surveys report to perform no sequencing.

7 participants reported that they work with bacteria in their institute, 4 reported to work on parasites and 5 reported to work on viruses.

1 of the surveys report a sample volume regarding sequencing in the range of 1-10 samples per week, 2 in the range of 51-100 samples per week, 1 reported a range of 101-200 samples per week.

Only 1 survey reported the use of only long read methods 3 surveys reported the use of both long- and short-reads methods.

3 surveys report to perform sequencing with the goal of obtaining consensus genomes, 3 surveys reported the aim of phylogenetic analyses, 3 survey participants perform sequencing to screen for certain mutations. 5 survey participants perform sequencing to detect antimicrobial resistances, 2 surveys report surveillance of antimicrobial resistances as a goal in sequencing experiments, 2 surveys report using sequencing for outbreak detection, 3 surveys report using sequencing for cluster allocation.

4 of the surveys reported, that nobody in their institute is working in bioinformatics, 1-3, 4-9 and 10-19 members working in bioinformatics was reported once respectively.

2 surveys reported that only dedicated bioinformaticians in the organisation perform bioinformatics, 2 surveys reported that wet lab members in the organisation perform bioinformatics and 1 survey reported that only wet lab members perform bioinformatic analysis in their institute.

Regarding bioinformatics the survey participants were asked about tools used for quality control and trimming.4 participants reported the use of FastQC, 3, participants the use of porecho,2

participants reported the use of Cutapt, Trimgalore, Trimmomatic and NanoFilt. Filtlong, the QUAST pipeline, SeqTK, fastp, Bloocoo and KMA were reported in one survey each.

2 surveys reported on quality criteria that decide whether the quality of the reads allows further analysis. This surveys reported the Phred score, adapter content, insert size and the throughput as a criterium.

Participants were asked about the tools they use for genome assembly and annotation.

4 participants reported the use of Kraken and Guppy. 3 participants reported the use of Bowtie, SPAdes, Unicycler, CARD and PROKKA. 2 participants reported the use of BWA, MEGA6 and Flye respectively.1 participant reported the usage RAST, Enterobase, canSNPer, Canu, CGE, MEGAN, Kaiju, Velvet, VBF SKESA, Medaka, megahit, gam-ngs, VFDB, platon and Bakta respectively.

Regarding Data visualization 4 participants reported the use of R Studio. Freyja, Microreact and Seqsphere+, as well as the python libraries plotly, matplotlib and GGplot were mentioned once each.

The survey participants were asked which database they use to obtain reference sequences. The NCBI database and CARD were reported 2 times. VFDB, PLSDB and the ENA were reported once each.

2 participants reported to upload genomes to GISAID, 4 to the NCBI databases, and 1 participant reported to upload to the ENA. 3 participants reported to share genomes in the FASTA format, and 2 surveys reported to upload fastq files. 3 Participants reported to upload geographical data, 3 to upload the collection date, 3 upload the type of organism and 1 participant uploads the sample matrix to these databases. 1 participant reported uploading less than 10% another reported uploading 10-50% and 1 participant reported to upload more than 90% of their data.

The participant reported to not share data with external institutes. Data sensitivity and legal obligations were reported as a reason.

Participants were asked if they use any bioinformatic workflow management tools. 3 participants reported the usage of Bash scripts,2 participants the usage of Nextflow and R, 1 participant reported the usage of Galaxy.

5.4. Digitisation in Greece

4 surveys were filled out from Greece. 2 surveys were from an organisation with 1 to 10 members, 1 from an organisation with 21-50 members, The last survey was returned by a member of an organisation with 51-100 members.

1 survey participants from Greece report, that shotgun sequencing is performed in their organisation 2 surveys report the use of amplicon sequencing.1 survey report to perform no sequencing.

4 participants reported that they work with bacteria in their institute and 2 reported to work on viruses.

1 of the surveys report a sample volume regarding sequencing in the range of 11-50 samples per week, 1 in the range of 51-100 samples per week and 1 reported a range of 101-200 samples per week.

2 surveys report the use of short read sequencing methods and 1 survey reported the use of both long- and short-reads methods.

3 surveys report to perform sequencing with the goal of obtaining consensus genomes, 3 surveys reported the aim of phylogenetic analysis, 3 survey participants perform sequencing to screen for certain mutations. 2 survey participants perform sequencing to detect antimicrobial resistances, 2 surveys report surveillance of antimicrobial resistances as a goal in sequencing experiments, 2 surveys report using sequencing for outbreak detection, 2 surveys report using sequencing for cluster allocation.

1 of the surveys reported, that nobody in their institute is working in bioinformatics, 3 report 1-3 people working in this field.

1 survey reported that only wet lab members in the organisation perform bioinformatics and 2 surveys reported that wet lab members and dedicated bioinformaticians perform bioinformatic analysis in their institute.

Regarding bioinformatics the survey participants were asked about tools used for quality control and trimming.3 participants reported the use of FastQC and Trimmomatic, 1 survey user utilizes the QUAST pipeline, uparse and custom-made scripts.

3 surveys reported on quality criteria that decide whether the quality of the reads allows further analysis. These surveys reported the Phred score as a criterium for short read methods and the number of pores as a criterium for long read sequencing.

Participants were asked about the tools they use for genome assembly and annotation.

3 participants reported the use of Bowtie, BWA, and SPAdes. 2 participants reported the use of PROKKA. And 1 participant reported the usage CGE, KRAKEN and star respectively.

Regarding Data visualization 4 participants reported the use of R Studio. Freyja, Microreact and Seqsphere+, as well as the python libraries plotly, matplotlib and GGplot were mentioned once each.

The survey participants were asked which database they use to obtain reference sequences. The NCBI database was reported 2 times. USCS genome browser, ensembl, GISAID and Usher were reported once each.

3 participant reported to share data with external institutes. 2 participants reported to upload genomes to GISAID. 3 participants reported to share genomes in the FASTA format and 1 survey reported to upload or share fastq files and bam files. 2 Participants reported to upload geographical data, 2 to upload the collection date, 3 upload the type of organism and 3 participants uploads the sample matrix to these databases.

1 participant reported uploading less than 10% and 2 participants reported to upload more than 90% of their data.

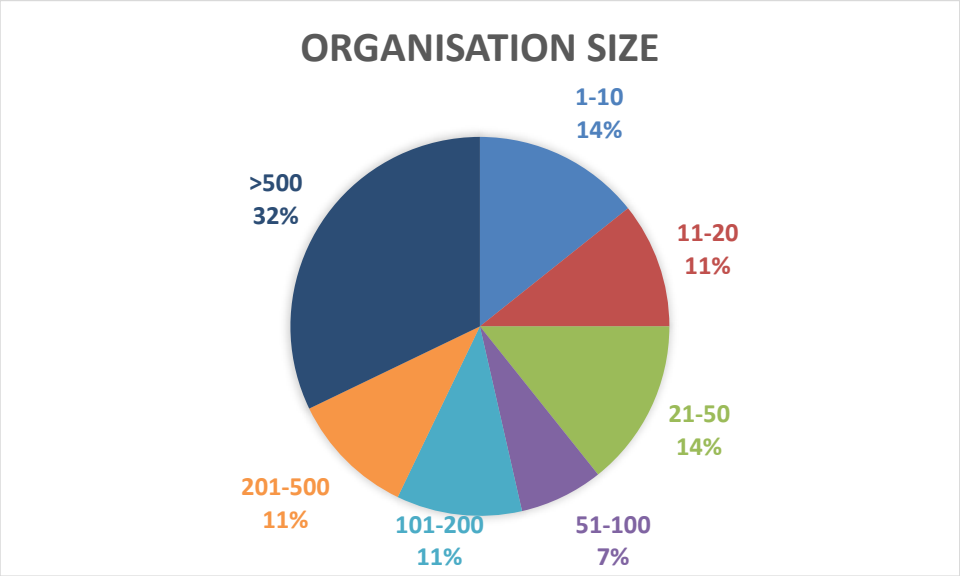
Participants were asked if they use any bioinformatic workflow management tools. 2 participants reported the usage of Bash scripts, R and Galaxy. 1 participant reported the usage of make and snakemake.

6. Results

28 surveys were filled out and reported in the consortium countries in total. Here we want to highlight some findings of this survey.

6.1. Organization size

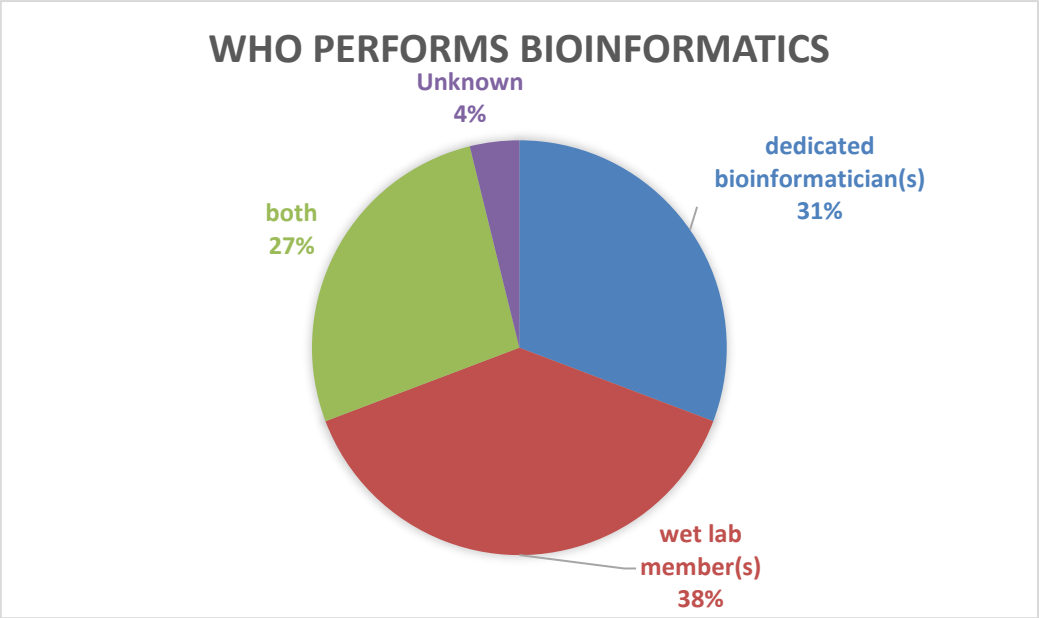
The following graph represents the organization's size from which the survey was reported:



Roughly a third of the respondents came from an organization with more than 500 members. Roughly a half of the responses came from organizations with less than 100 members.

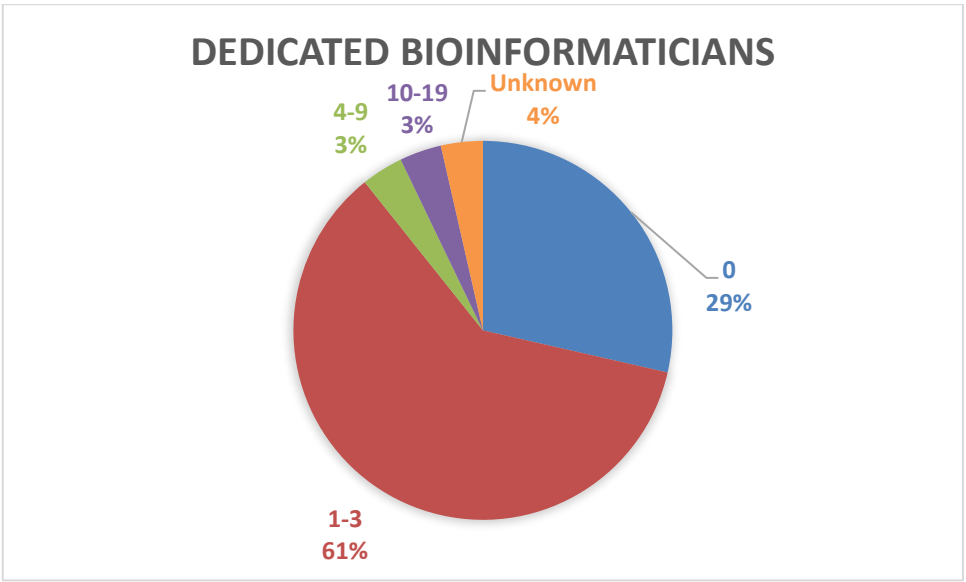
6.2. Bioinformatic personal

The following graph highlights who is performing bioinformatic analyses in the organizations:



In about two thirds of the responses, it was reported that scientists working in the wet lab also have to perform bioinformatics. This could proof as a bottleneck in case of major outbreaks since these scientists would be burdened with the wet lab analysis as well as the bioinformatics.

The following graph highlights the number of dedicated bioinformaticians in the organizations:

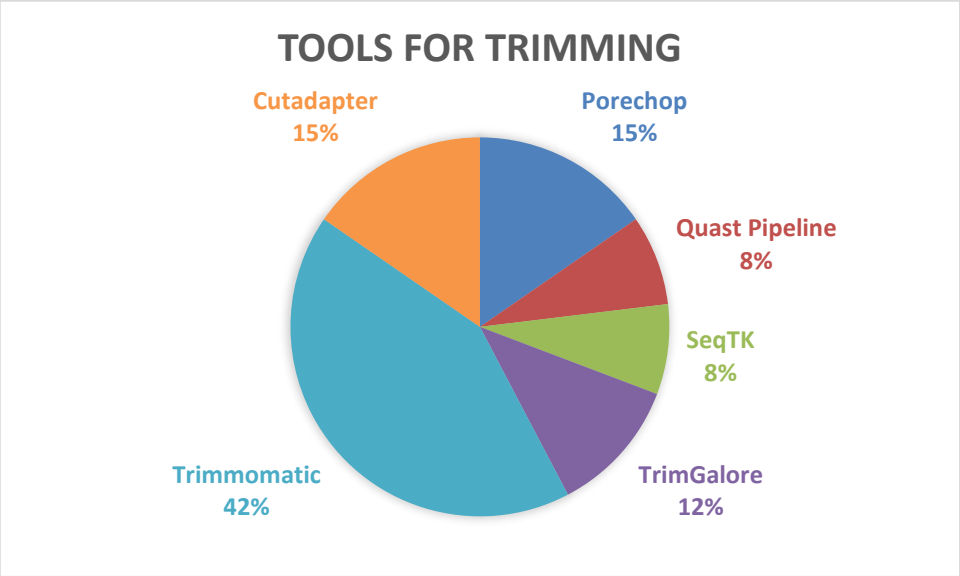


Roughly 90 percent of the responders reported less than 4 bioinformaticians in their organization. 29 percent report having zero dedicated bioinformaticians. This can explain why wet lab scientists also have to perform bioinformatics in two thirds of the organizations.

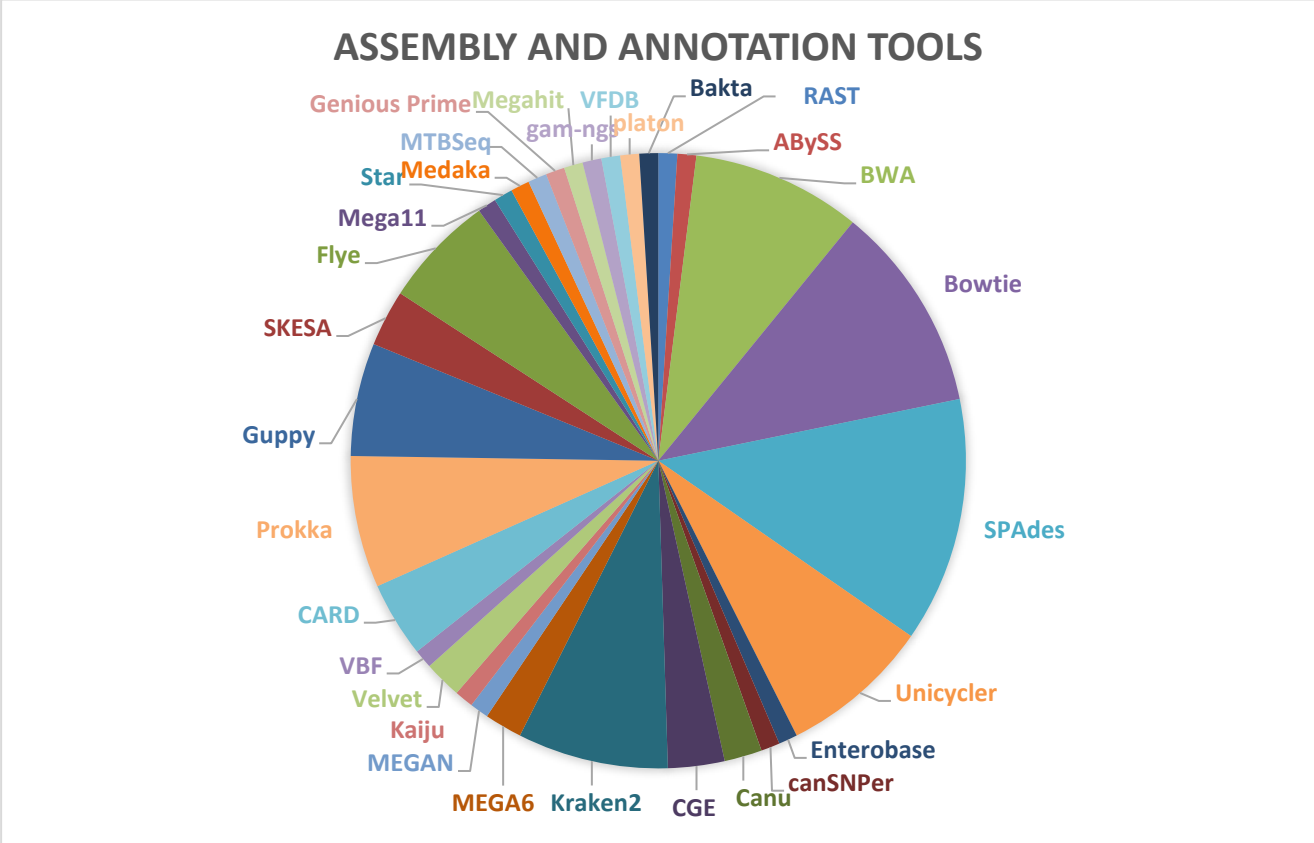
The consortium´s members are in agreement, that more bioinformaticians are necessary in their respective organizations to be able to adequately and quickly respond to potential future outbreak scenarios.

6.3. Bioinformatic Tools

The following graphic represents the reported tools used for trimming short or long reads. The usage of a variety of tools was reported. Still Trimmomatic is clearly one of the more used trimming tools.



Looking at Assembly and Annotation tools an even bigger variety is used. This is represented in the following graph.

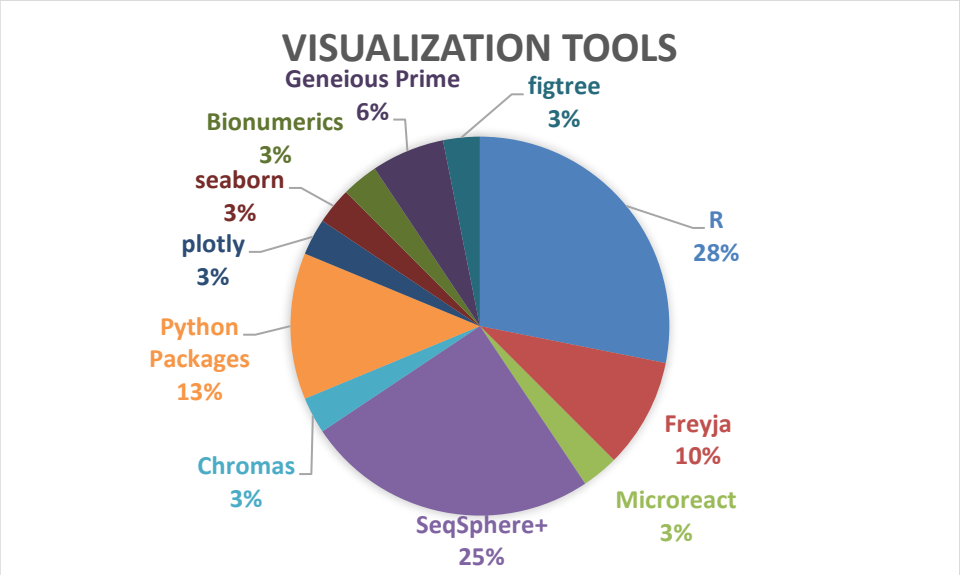


Listed were tools used for assembly of long reads and tools for short reads. Some of these tools run only on the Linux operative system, some also on Microsoft Windows. Some are command line interface only, others come with a graphical user interface. Some are broad in scope, other specialised on certain pathogens. This highlights the diversity in bioinformatic analyses currently in use. Since all of these tools have their unique advantages and disadvantages it can be assumed, that the survey participants use these tools purposefully. This also highlights an already high level of know-how in the survey participants and a flexibility regarding data analysis.

It is not the goal of this report to prioritize one tool over the other. Therefore it is required that each organization uses adequat methods for a given task and given computational resources. However it shoul be noted that new tools are released continiously. Additionaly many already released tools are updated regularly as well. As a result the used tools should be reviewed regularly, to determine if they still are the best suited to fulfill the specific needs.

6.4. Visualisation

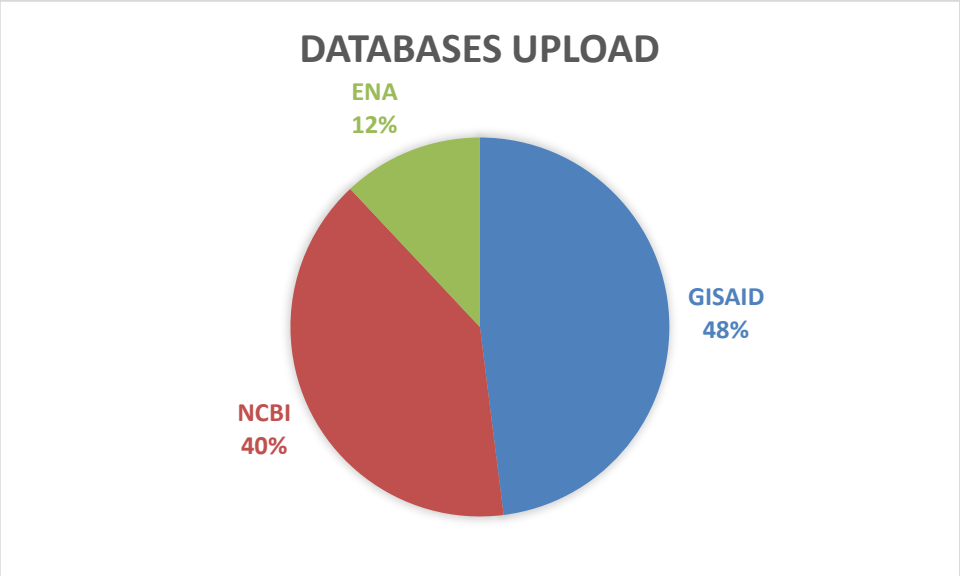
Participants were asked to report the software they use to visualize data obtained by next generation sequencing.



Interestingly R and Python packages make up over a third of the total reports. Meaning that for data vizulisation some scripting knowledge is present in the survey participants. It is not the goal of this report to prioritize one tool over the other. Therefore it is required that each organization uses adequat methods for a given task and given computational resources.

6.5. Databases

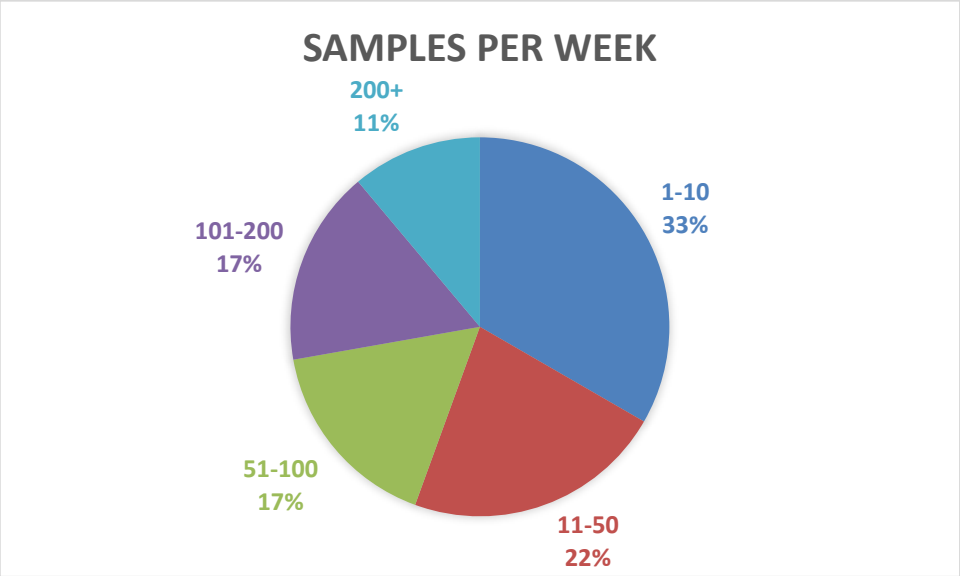
Survey participants upload their generated data into three databases:



Notably GISAD and NCBI except files in FASTA format whereas the ENA excepts the reads of a sequencing experiment in Fastq format. If sequenced samples are to be uploaded into one of these databases it is necessary, that all required metadata is also collected and saved.

6.6. Data storage

Generated data also has to be stored to be used for future analyses. To estimate the required capacity survey participants were asked to report how many samples the are currently sequencing per week. This graph shows the high diversity in sample throughput in the survey participants. Therefor a general recommendation for required storage capacity cannot be made.



There are however a number of factors that have to be taken into account when estimating the required storage space:

- Sample throughput
  - The more samples are sequenced the more data is generated.
- Size of target sequences
  - The larger the sequences the more data is generated.
- Target coverage
  - The higher the target coverage the more data is generated.
- Duration of storage
  - The longer data need to be stored the more storage space is needed.
- Ability to upload to public databases.
  - Uploading to public databases allows to use external storage capacity while still maintaining ability to access one's data.

7. Next steps

This report gives an overview of the current status of digitization in regard to NGS in Austria, Croatia, Greece and Hungary. It is not possible to define concrete requirements that are necessary for each of the consortium members and stakeholders in the countries. However, this report includes information that can be used by each consortium member to better identify their personal needs and as a result requirements.

The mayor databases that are uploaded to were identified as NCBI, GISAID and ENA. A roadmap to harmonize the country specific databases to these databases is planned in the deliverable 4.2.