Ana Rita Rebelo

anrire@food.dtu.dk

Recommendations and proposals to develop harmonised protocols for CRE/CCRE surveillance and outbreak detection

Funded by the European Union

Rationale – Why WGS?

Advantages

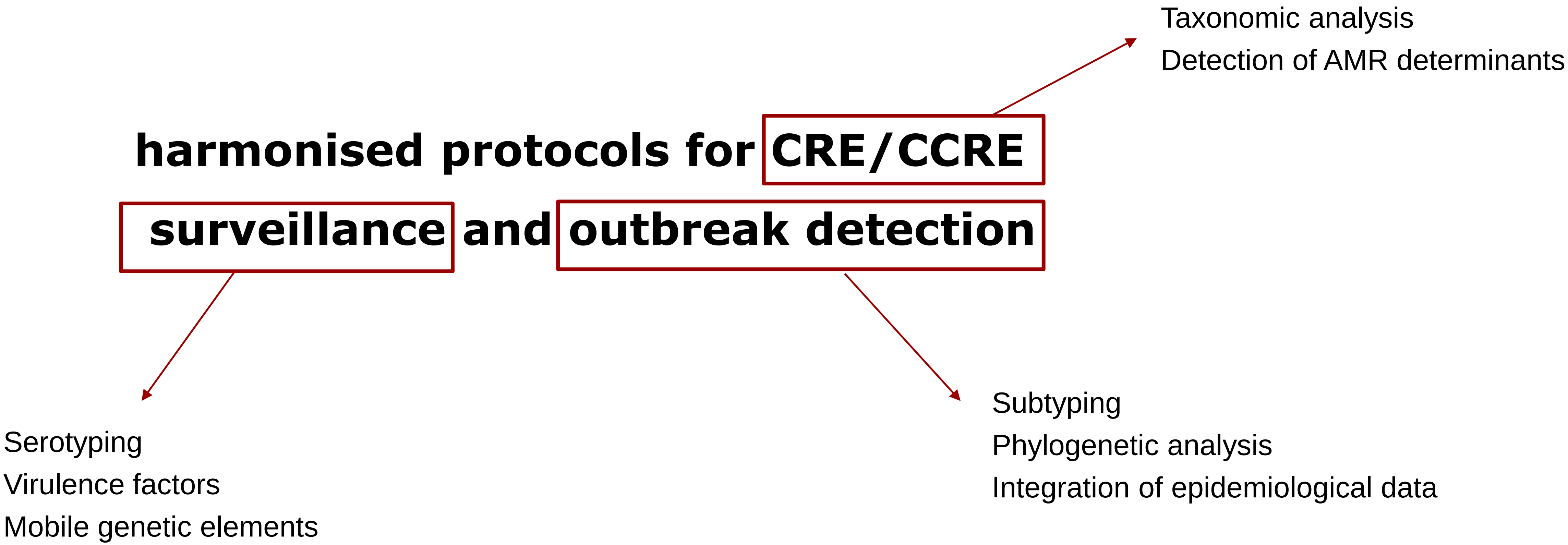
- Only one protocol
- Very large amount of data
- Higher discriminatory power
- Harmonised and automatic analysis
- Direct comparison
- Ease of storage
- Retroactive screening

Why now?

- Increase in sequencing accuracy
- Decrease in cost
- Coordinated efforts throughout Europe

Funded by the European Union

Rationale – What should we be able to do?



Taxonomic analysis
Detection of AMR determinants

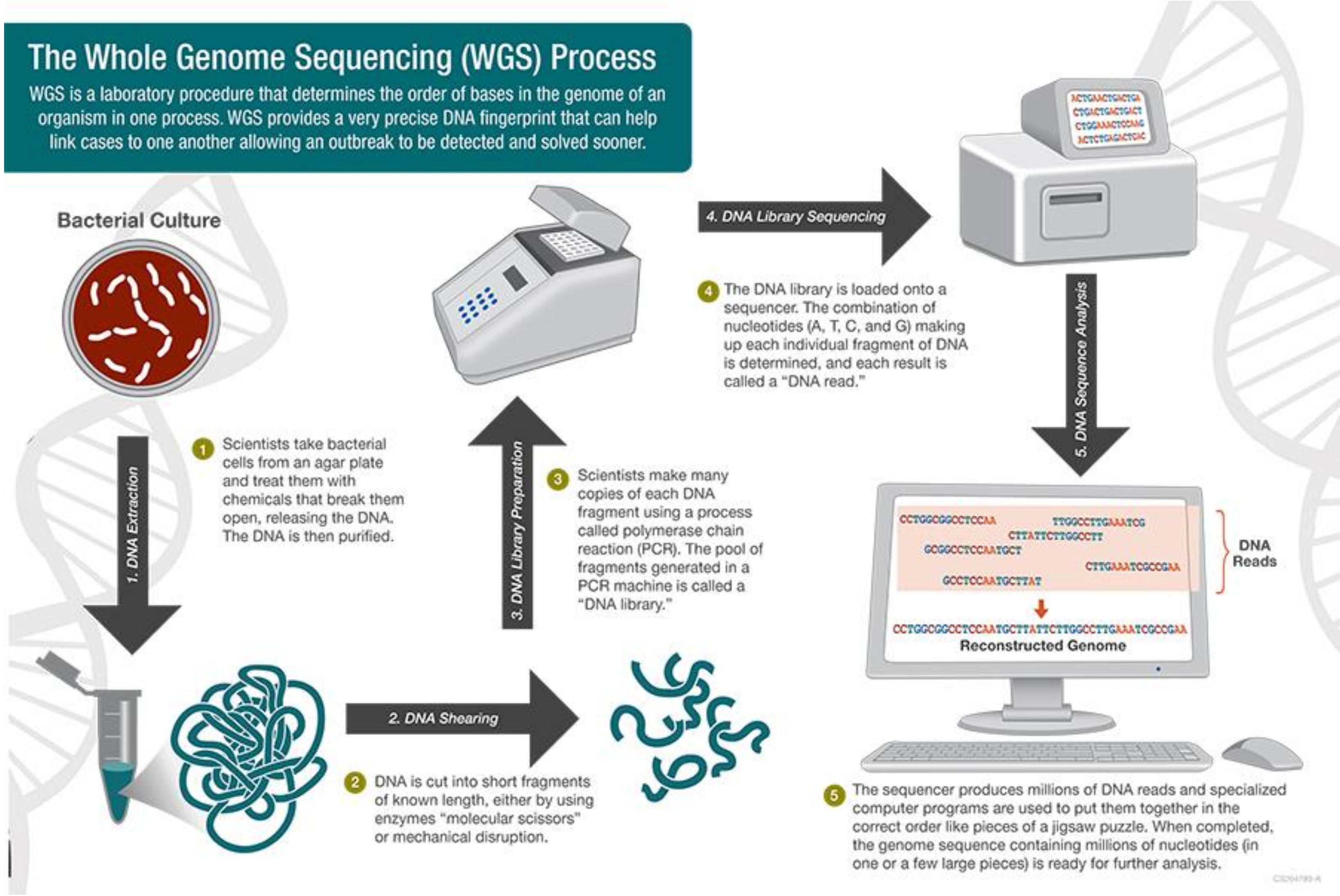
harmonised protocols for CRE/CCRE surveillance and outbreak detection

Serotyping
Virulence factors
Mobile genetic elements

Subtyping
Phylogenetic analysis
Integration of epidemiological data

Funded by the European Union

WGS-BASED ANALYSIS OF BACTERIA - OVERVIEW



The Whole Genome Sequencing (WGS) Process

WGS is a laboratory procedure that determines the order of bases in the genome of an organism in one process. WGS provides a very precise DNA fingerprint that can help link cases to one another allowing an outbreak to be detected and solved sooner.

1 Scientists take bacterial cells from an agar plate and treat them with chemicals that break them open, releasing the DNA. The DNA is then purified.

2 DNA is cut into short fragments of known length, either by using enzymes "molecular scissors" or mechanical disruption.

3 Scientists make many copies of each DNA fragment using a process called polymerase chain reaction (PCR). The pool of fragments generated in a PCR machine is called a "DNA library."

4 The DNA library is loaded onto a sequencer. The combination of nucleotides (A, T, C, and G) making up each individual fragment of DNA is determined, and each result is called a "DNA read."

5 The sequencer produces millions of DNA reads and specialized computer programs are used to put them together in the correct order like pieces of a jigsaw puzzle. When completed, the genome sequence containing millions of nucleotides (in one or a few large pieces) is ready for further analysis.

https://www.cdc.gov/pulsenet/pathogens/protocol-images.html#wgs

Funded by the European Union

WGS-BASED ANALYSIS OF BACTERIA - REQUIREMENTS

- Expertise on DNA extraction methods
- Expertise on library preparation methods

Not too technically demanding

Ideally a dedicated room

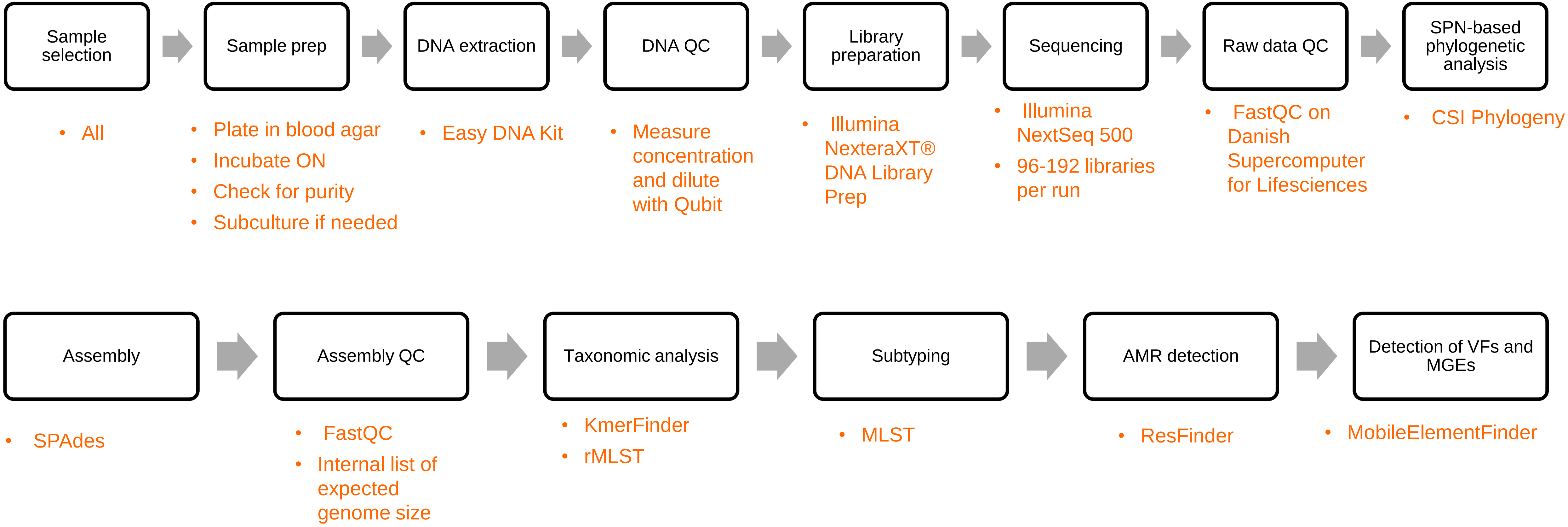
- Access to sequencing platform
- Access and expertise on bioinformatics tools

Main challenges: cost, implementation

- Data management infrastructure

Main challenges: cost, compatibility

EXAMPLE: A COMPLETE WGS WORKFLOW



Sample selection
- All

Sample prep
- Plate in blood agar
- Incubate ON
- Check for purity
- Subculture if needed

DNA extraction
- Easy DNA Kit

DNA QC
- Measure concentration and dilute with Qubit

Library preparation
- Illumina NexteraXT® DNA Library Prep

Sequencing
- Illumina NextSeq 500
- 96-192 libraries per run

Raw data QC
- FastQC on Danish Supercomputer for Lifesciences

SPN-based phylogenetic analysis
- CSI Phylogeny

Assembly
- SPAdes

Assembly QC
- FastQC
- Internal list of expected genome size

Taxonomic analysis
- KmerFinder
- rMLST

Subtyping
- MLST

AMR detection
- ResFinder

Detection of VFs and MGEs
- MobileElementFinder

+++ integrating metadata

Funded by the European Union

NOT THAT SIMPLE – TOO MANY OPTIONS

Tool	Reference database	Description and output
Tools for taxonomic analysis and typing		
KmerFinder [150,151]	KmerFinder	Provides hits of the query genome against whole reference genomes, the respective % of identity and % of coverage
SILVA [152]	SILVA	Collection of 16S rRNA genes, also possible to perform phylogenetic analysis and obtain phylogenetic trees
MLST [153]	PubMLST	MLST schemes, provides the sequence type
rMLST [126]	rMLST	rMLST schemes, provides the predicted species and respective allelic support metric
SerotypeFinder [154]	SerotypeFinder	Serotype, specific for E. coli
SeqSero [		
PneumoC		
Tools for		
cgMLST [		be used exclusively for typing but also clustering
CSIPhylog		through FastTree
Evergreen		
Tools for detection of antimicrobial resistance determinants		
ResFinder [129]	ResFinder, PointFinder	Provides hits against reference ARGs and PMs and the respective % of identity and % of coverage, position in genome and predicted phenotype
KmerResistance [150,151]	KmerResistance	Provides hits of the query genome against reference genomes, as well as the detected ARGs and respective % of identity and % of coverage
CARD/RGI [158]	CARD	Provides hits against reference ARGs and respective % of identity and % of coverage. Other options are possible and the service is highly focused on ontology and standardization
AMRFinder [159]; AMRFinderPlus [160]	NCBI RefSeq	Provides hits against reference ARGs and respective % of identity and % of coverage
ARIBA [161]	CARD, ResFinder, NCBI Bacterial AMR Reference Gene DB, ARG-ANNOT, MEGARes, PubMLST, others defined by user	Provides hits against reference ARGs and respective % of identity and % of coverage
Tools for detection of virulence factors		
VirulenceFinder [162]	VirulenceFinder	
Victors [131]	Victors	
Tools for detection and analysis of mobile genetic elements		
PlasmidFinder [163]	PlasmidFinder	
Platon [164]	Platon	Provides hits against reference plasmids and respective % of identity and % of coverage, as well as relevant genes
pMLST [153]	PubMLST	Plasmid typing schemes
MobileElementFinder [135]	MobileElementFinder	Provides type and reference sequences of MGEs, respective % of identity and % of coverage, as well as associated ARGs and VFs
Pipelines for extensive analyses		
NCBI Pathogen Detection	NCBI DBs	Detects ARGs and VFs, provides SNP-based phylogenetic analysis
Pathogenwatch [165]	Pathogenwatch, tools' DBs	Performs taxonomic analysis, determines MLST and cgMLST and provides cgMLST-based phylogenetic clustering
BIGSdb [166]	PubMLST BIGSdb	Performs annotation and taxonomic analysis, detects ARGs and plasmids, determines MLST, rMLST and cgMLST, provides phylogenetic and spatio-phylogenetic analysis
PATRIC [167]	PATRIC, but also includes others such as CARD, NDARO and VFDV	Performs assemblies, quality control, annotation and taxonomic analysis, detects ARGs and performs phenotype prediction, detects VFs and MGEs, provides phylogenetic analysis, variation analysis and genome alignments

This is just a subset….

….and just from tools for analysis .

RECOMMENDATIONS FOR HARMONIZATION

i. Accounting for future priority pathogens and priority AMR profiles for surveillance.

ii. Choose well defined AMR genotypes for validation of the WGS approach.

iii. Establish the control parameters to be used.

iv. Establish the thresholds for the control parameters.

v. Define a set of bioinformatics tools and databases as potential candidates to be included in the harmonised approach.

vi. Consider the results of other tasks incurred in the context of this project.

Funded by the European Union

RECOMMENDATION 1

Accounting for future priority pathogens and priority AMR profiles for surveillance.

Funded by the European Union

RECOMMENDATION 1

MAIN OBJECTIVE: the harmonised WGS approach is streamlined for surveillance and analyses of CRE/CCRE bacteria

BUT: should already account for the future integration of two additional epidemic-prone healthcare-associated antimicrobial-resistant bacterial pathogens of public health priority

BECAUSE: current European guidelines and Commission Implementing Decisions might at any time be reviewed if new epidemiological situations become established, changing the scope of organisms under surveillance

AND: we should avoid working in silos

EXAMPLES:

○ selection of ARG databases that contain ARGs and point mutations for other microorganisms besides CRE/CCRE

○ defining QC parameters at are not exclusive to *Enterobacterales*

establishing the acceptable genome size deviation as a percentage of the expected genome size, and not as a numerical deviation from the expected 5 million base-pairs.

Funded by the European Union

RECOMMENDATION 2

Choose well defined AMR genotypes for validation of the WGS approach.

Funded by the European Union

Recommendation 2

MAIN OBJECTIVE: Predict AMR phenotypes (which are often regulated by a combination of several genotypic determinants)

BUT: many of the genetic mechanisms are currently not well defined and prove difficult to be detected through WGS approaches

SOLUTION: Defining a subset of well-studied ARGs and point mutations allows for comparison of sampling methods, laboratory protocols and bioinformatics workflows

EXAMPLE: Adequately dectecting *mcr*-genes in colistin-resistant *Enterobacterales* could correspond to the benchmark indicating an adequate approach, even knowing that other mechanisms of resistance (e.g. PMs) exist

Funded by the European Union

RECOMMENDATION 3 AND RECOMMENDATION 4

Establish the control parameters to be used.

Establish the thresholds for the control parameters.

Funded by the European Union

RECOMMENDATION 3 AND RECOMMENDATION 4

Many different:

- DNA extraction kits
- Sequencing platforms
- Bioinformatics approaches
- Bioinformatics tools

Is harmonization feasible?

Well defined set of QC parameters

- For the raw data

  E.g. nr. and length of raw reads, depth of coverage

- For the assembled genomes

  E.g. N50, nr. of contigs, genome size

- For the performance of the tools

  E.g. accurately detect PMs and ARGs in sets of benchmarking data

Recommendation 3 and Recommendation 4

We propose two sets of QC parameters that will allow the users to maintain flexibility in their choice of WGS platform and bioinformatics tools:

- Sequencing QC parameters
- Data management QC parameters

Funded by the European Union

RECOMMENDATION 3 AND RECOMMENDATION 4

Sequencing QC parameters:

any sequencing platform, protocol and bioinformatics tool

↓

raw data and assembled genomes with quality equal or above limited thresholds for defined control parameters

Current general consensus:

- ○ "number and length of raw reads"
- ○ "depth of coverage"
- ○ "number of contigs in the assembled genome"
- ○ "N50 and deviation from expected genome size"

Funded by the European Union

RECOMMENDATION 3 AND RECOMMENDATION 4

Sequencing QC parameters:

- the depth of coverage (both of the raw data and also of the assembled genome) should be at least of 30 times (30X)

  15X has proven sufficient to an adequate detection of ARGs and point mutations in E. coli

- the number of contigs in the assembled genome must be lower than 1,000, and ideally lower than 500

  short-read technologies

- the size of the assembled genome is dependent on the target species. To account for genome plasticity and mobile genetic elements we suggest that a maximum threshold of 10% of variation in the number of base-pairs (BPs) should be adopted.

  CRE/CCRE: genome size 5 million BPs = a variation of plus/minus 0.5 million BPs would be acceptable

Funded by the European Union

Recommendation 3 and Recommendation 4

Data management QC parameters:

Ensure that all workflows respect the same data management directions and there is:

-traceability of data and methods

-compatibility of data types and formats between different bioinformatics approaches

-comparability of results between settings

Funded by the European Union

RECOMMENDATION 3 AND RECOMMENDATION 4

Data management QC parameters:

Parameters:

?

ISO standard will soon become available and help guide us

Examples from ISO draft:

- defining the minimum metadata requirements and respective adequate descriptors
- describing the proper registry methods for the DNA extraction and sequencing protocols, WGS platforms and bioinformatics analyses.

Funded by the European Union

BENCHMARKING DATASETS

Data management QC parameters:

--secret--

BENCHMARKING DATASETS

European Commission's Joint Research Centre

2018: “The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies”

2021: “A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing”

- Benchmarking approaches to validate sequencing and bioinformatics workflows

- Ensure that different pipelines can be used while at the same time adhering to the same minimum standards of performance

Funded by the European Union

BENCHMARKING DATASETS

HOW?

- creating platform-specific validation datasets
- using simulated data that complies with specific certifications (not biases towards the creating platform)
- accepting the *fastq* format as the standardized input for analysing the performance of bioinformatics tools
- **accuracy should be dependent on the ability of the workflow to identify the correct AMR determinants that are introduced in the dataset (and not on agreement with phenotypic results)**
- these AMR determinants should include species and mechanisms which are selected based on international priority lists

Funded by the European Union

RECOMMENDATION 5

Define a set of bioinformatics tools and databases as potential candidates to be included in the harmonised approach.

Funded by the European Union

BIOINFORMATICS TOOLS - OVERVIEW

Purpose

Quality control
Assembly
Taxonomic analysis
Phylogeny
Serotyping/Subtyping
Detection of AMR determinants
Detection of other determinants

Accessibility

Web-based
Command line
Local vs. server

Data

Raw data as input
Assemblies as input
Integration of metadata

Maintenance

Benchmarked
Curated

Cost

Open access
Subscription

Funded by the European Union

Example: Bioinformatics tools for prediction of AMR

Tool	Target species	Reference database	Output	Comments
SRST2	All	CARD, PubMLST, or others defined by user	Reference sequences and respective % of coverage, depth	Also taxonomy, phylogeny, VFs, plasmids, other, depending on provided databases
ARIBA	All	CARD, ResFinder, ARG-ANNOT, MEGARes, NCBI Bacterial AMR Reference Gene Database, PubMLST, or others defined by user	Reference sequences and respective % of identity, % of coverage	Also phylogeny, VFs and plasmids, depending on provided databases (such as plasmidfinder, VFDB, VirulenceFinder)
KmerResistance	All	Own	Reference genomes, ARGs and respective % of identity, % of coverage	Also taxonomy
ResFinder	All	Own	ARGs and respective % of identity, % of coverage, position in genome, predicted phenotype	NA
PointFinder	Limited	Own	Mutated gene, protein translation, predicted phenotype	Included in ResFinder but can be used locally by itself. Currently under development for Klebsiella spp.
RGI	All	CARD	Reference sequences and respective % of identity, % of coverage, other options	Integrated in the Galaxy server; allows proteome analysis
AMRFinder; AMRFinderPlus	All (Limited PMs)	NCBI RefSeq	Reference sequences and respective % of identity, % of coverage	Included in NCBI Pathogen Detection
SSTAR	All	Own (created by merging ResFinder and ARG-ANNOT)	Reference sequences and respective % of coverage, depth	Can be used with other reference databases
ABRicate	All	CARD, ResFinder, ARG-ANNOT, MEGARES, NCBI AMRFinderPlus, or others defined by user	Reference sequences and respective % of identity, % of coverage	Also VFs and plasmids, depending on provided databases Also VFs and plasmids, depending on provided databases (such as plasmidfinder and VFDB)
CARD	All	Own	ARGs and point mutations, respective prevalence and predicted phenotype	Highly focused on ontology and standardization. VFs and mobile genetic elements currently being added

Funded by the European Union

EXAMPLE: BIOINFORMATICS PIPELINES FOR PREDICTION OF AMR

Tool	Target species	Reference database	Output
Pathogenwatch	Limited	Own, tools' databases	Taxonomy, MLST, cgMLST and clustering. Other functionalities for Klebsiella spp. derived from Kleborate
Enterobase	Limited	Tools' databases	Genome assembly and annotation, serotyping, MLST, cgMLST, rMLST, phylogenetic analysis
BIGSdb	All	PubMLST BIGSdb	Annotation, taxonomy, ARGs, plasmids, MLST, rMLST, cgMLST, phylogenetic and spatio-phylogenetic analysis, comparative genomics
NCBI Pathogen Detection	Limited	Own	ARGs, VFs, SNP-based phylogenetic analysis
PATRIC	All	Own, but also includes others such as CARD, NDARO and VFDV	Assemblies, QC, annotation, taxonomy, ARGs, phenotype prediction, VFs, mobile elements, phylogenetic analysis, variation analysis, genome alignments, comparative genomics, other options
Ridom SeqSphere+ €	All	Own, tools' databases; Includes NCBI AMRFinder and VFDB	Assemblies, QC, taxonomy, ARGs, VFs, MLST, cgMLST, phylogenetic analysis
Bionumerics €	All (wgMLST schemes available for limited species)	Own, tools' databases, others provided by user	Assemblies, QC, annotation, taxonomy, ARGs, plasmids, MLST, rMLST, cgMLST, wgMLST, phylogenetic analysis, comparative genomics, other options for E. coli (ARGs, PMs, VFs, plasmids, serotypes)

Funded by the European Union

RECOMMENDATION 5

- should be **open-access**, in order to respect probable budget limitations of certain users

- should be available as **online interfaces** (or be part of other interfaces) to avoid the need for expensive computing resources and specific professionals

- should also be **downloadable for local** usage for users with the resources and interest in doing so

- should be **benchmarked**, and should be transparently and continuously **curated**

- The owners should be available for **collaboration** with the European Commission, ECDC, and this provider to facilitate implementation and testing, and to potentially coordinate events such as suspension of updates during External Quality Assessments, and to provide scientific and technical support to the users

Funded by the European Union

RECOMMENDATION 6

Consider the results of other tasks incurred in the context of this project.

- ○ WGS approaches currently used by the NRLs
- ○ minimum conflict as possible with the ones currently used

Funded by the European Union

SOME CONSIDERATIONS

There’s no one-size-fits-all approach

There’s no perfect bioinformatics tool

Essential requirements:

QC

Benchmarking datasets

→

Ensuring accuracy of your approach

Knowing which database supports each tool

Understanding how the tool works

→

Knowing what data to provide

Knowing the limitations of your results

DISCUSSION AND QUESTIONS

Input from NRLs with established WGS workflows

Positive and negative feedback regarding their specific approaches

Challenges so far

Input from NRLs with no-established systems

What part of the whole WGS workflow seems more demanding?

Specific fears (e.g. command-line seems too difficult)

To be discussed in break-out session → Very important comments now

Funded by the European Union

anrire@food.dtu.dk

Thank you!

Funded by the European Union