27 March 2019

Integration of genomics in outbreak detection and investigation of foodborne pathogens

Mirko Rossi

Scientific Officer

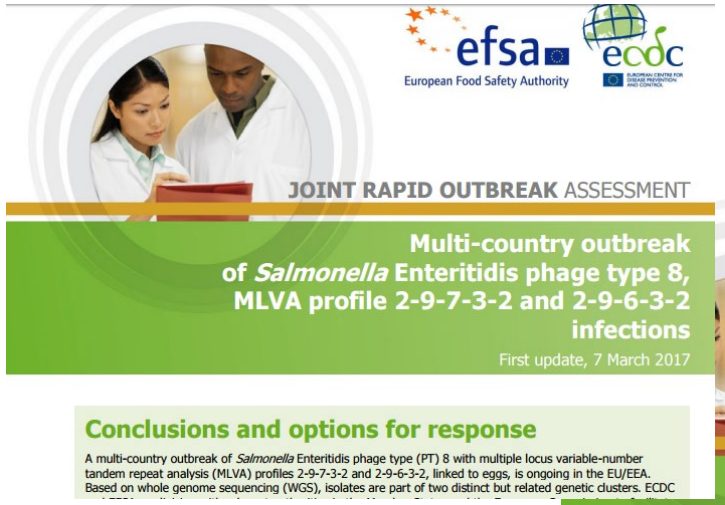
Trusted science for safe food

efsa
European Food Safety Authority

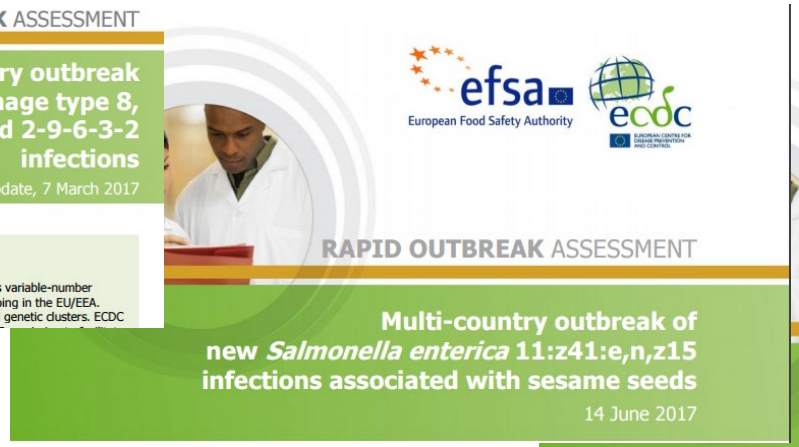
WGS typing in epidemiological investigations

- Generates hypothesis
- Confirms hypothesis quickly
- Adds strength to evidence
- Resolves ambiguous lab-epi data
- Accelerates response due to data sharing
- Enables targeting intervention upstream in the food production chain by identifying the ‘exact’ source
- Gives insights on why is the outbreak happening now, how is it likely to develop and where is it coming from

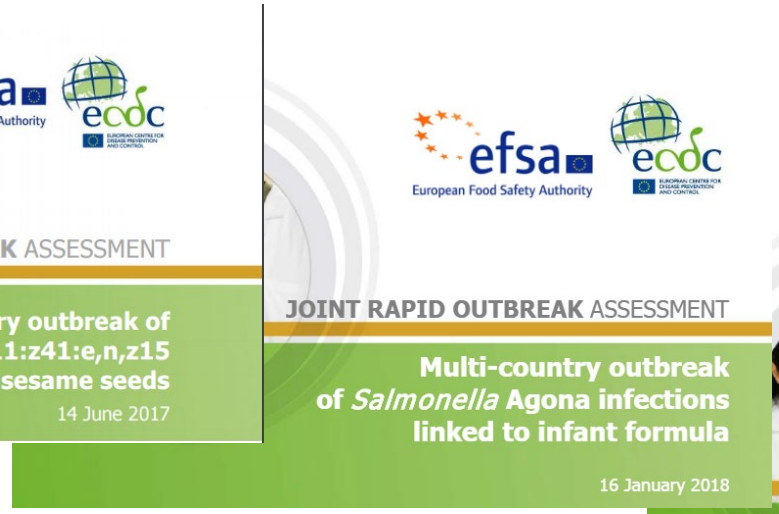
WGS is in used for ECDC-EFSA ROA



PHE SNPs calling
(+ own pipelines)



Ridom Seqsphere,
cgMLST



Bionumerics cgMLST
(+ SNPs + SeqSphere)

- Used daily by several MSs, but different methods in use
- WGS Data are communicated to EFSA/ECDC
- WGS typing included in the case definition in ROA

Courtesy of B. Guerra EFSA

EU WGS DATA Collection

efsa
European Food Safety Authority

"Technical support to collect and analyse whole genome sequencing (WGS) data in the joint ECDC-EFSA molecular typing database"

at least *L. monocytogenes*, *Salmonella*, *E.coli*

- **ToR1**: to analyse **outcome of ECDC and EFSA Surveys on WGS** capacity for foodborne pathogens in MSs (food and PH).
- **ToR2**: **…** to assess the **state of the art of pipelines** for collecting and analysing WGS data…
- **ToR3**: … to assess **needs/requirements** for analysis and comparability; interactions among databases; roles and responsibilities.
- **ToR4**: to prepare a **Technical report**: identification, comparison of potential solutions for a joint EFSA-ECDC

Deadline April 2019

BIOHAZ panel scientific opinion

efsa
European Food Safety Authority

"Self-tasking mandate for scientific opinion on the application and use of next generation sequencing (including whole genome sequencing) for risk assessment of foodborne microorganisms"

- **ToR1.** Evaluate the possible **use of NGS in foodborne outbreak detection/investigation** and hazard identification and underlining the added value for risk assessment.
- **ToR2.** Critically analyse **existing NGS-based methodologies** to assess their ability to complement or replace the **microbiological methods** cited in the current **EU food legislation**

Adoption October 2019

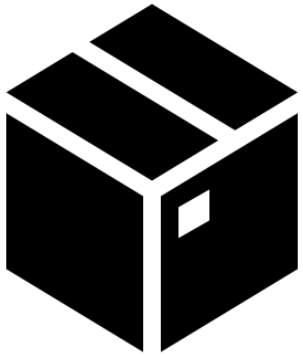
What are the challenges?

- Standardization and harmonization of the process
- Development of a plain language
- Precise communication of the results
- Validation of epidemiological concordance

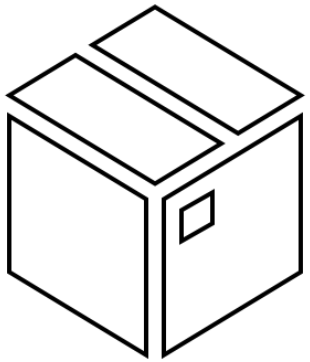
WGS as one-stop-shop for bacterial typing

- Virtually complete characterization at max resolution
- One lab method for all bacteria and all typing
- Sharing of a lot of information in universal format
- Less processing time and personnel workload

Magic box of WGS wonders



vs



Black Box

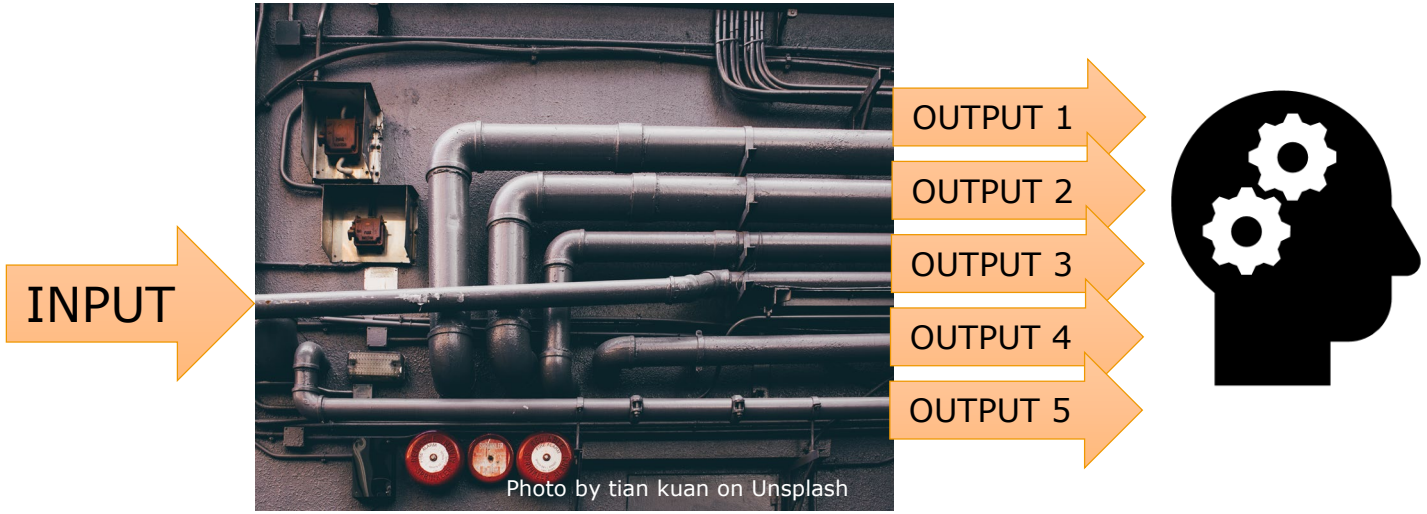
Commercial/Freeware
You get what it gives you
Ready to use
Stealth change
Standalone

See through Box

Freeware
You can "tailor"
Visible changes
Dependencies
"Major" headache

…What do you prefer?

Looking inside the box : pipeline



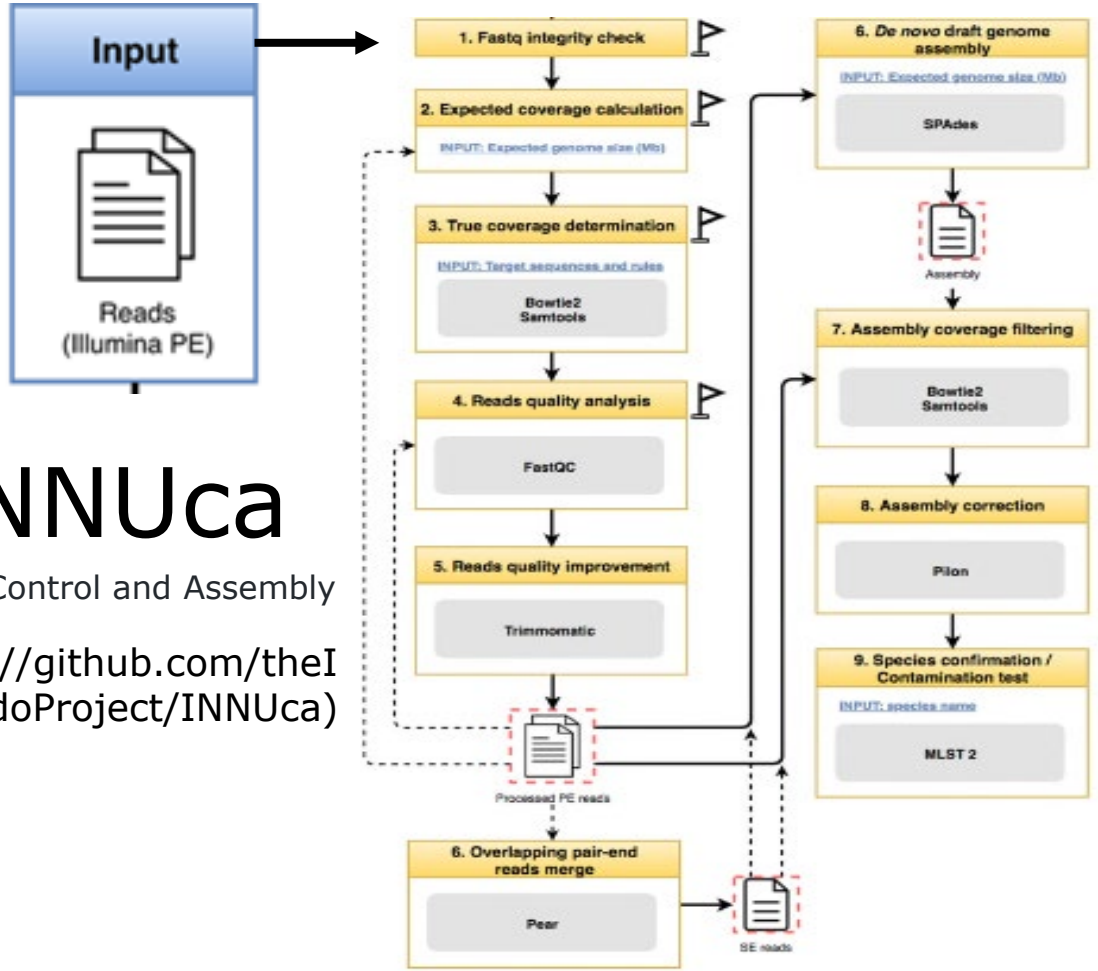
- Chain of processing elements (software or other pipelines)
- The output of each element is the input of the next
- Uses and distils outputs by a lot of software
- Aim for automate analysis; simplify the process

Bioinformatic pipelines

- Comparability: same workflow applied to all the samples
- Accountability: keeping track on the analysis
- Modularity: adding new steps easily
- Reproducibility: same input = the same output
  - NOTE certain software have stochastic steps
- *Software validation: difficult to track bugs*
- *Opacity: difficult to determine which module affect the results; loose track of the assumptions*
- *Software and DB dependencies*

Courtesy of Dr. Carriço UL, Portugal

Example: the assembly pipelines

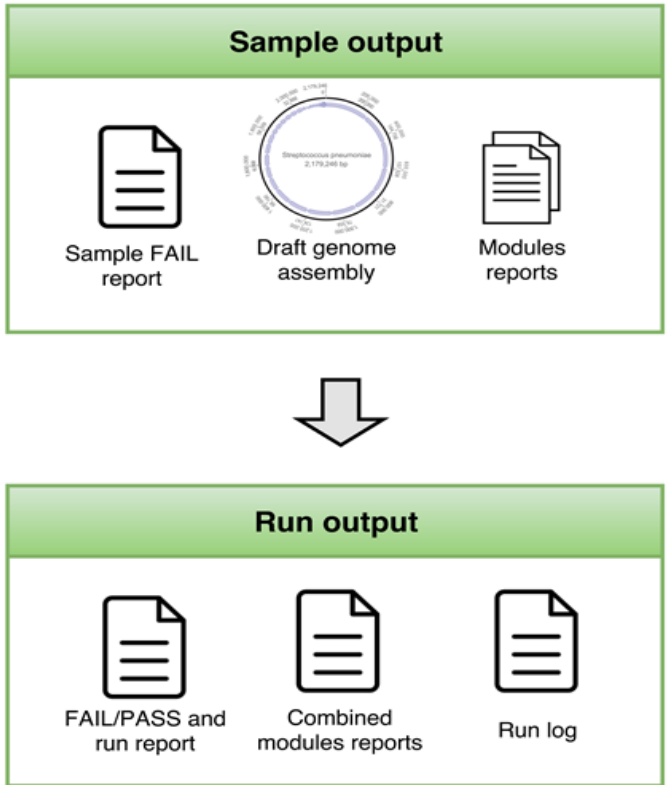


INNUca

Reads Control and Assembly

(https://github.com/theInnuendoProject/INNUca)

```
$ INNUca.py /
        -i read_dir/
        -s 'Salmonella enterica'
        -g 4.7
```



Other assembly pipelines

build passing | License GPL v3 | Language Perl 5

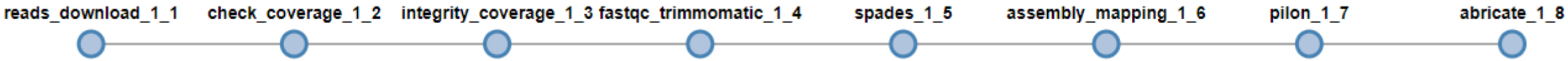
Shovill (https://github.com/tseemann/shovill)

Assemble bacterial isolate genomes from Illumina paired-end reads

Building your pipeline

- https://flowcraft.readthedocs.io/en/latest/index.html
- Overcome the dependencies problem
- Run customized pipeline in any environment
- Maximize the control of each module
- Software run the same exact way, every time

Software **container blocks** → Build pipeline → Execute



Building your pipeline: where to find the blocks

efsa
European Food Safety Authority

FLOWCRAFT

- > 40 images

https://hub.docker.com/u/flowcraft

StaPH-B
State Public Health Bioinformatics

- 23 Docker images

https://hub.docker.com/u/staphb

wellcome sanger institute

… and many others!!

https://hub.docker.com/u/sangerpathogens

Platforms for running pipelines

Galaxy COMMUNITY HUB Use

https://galaxyproject.org/use/

Galaxy / ARIES

EURL VTEC

IRIDA

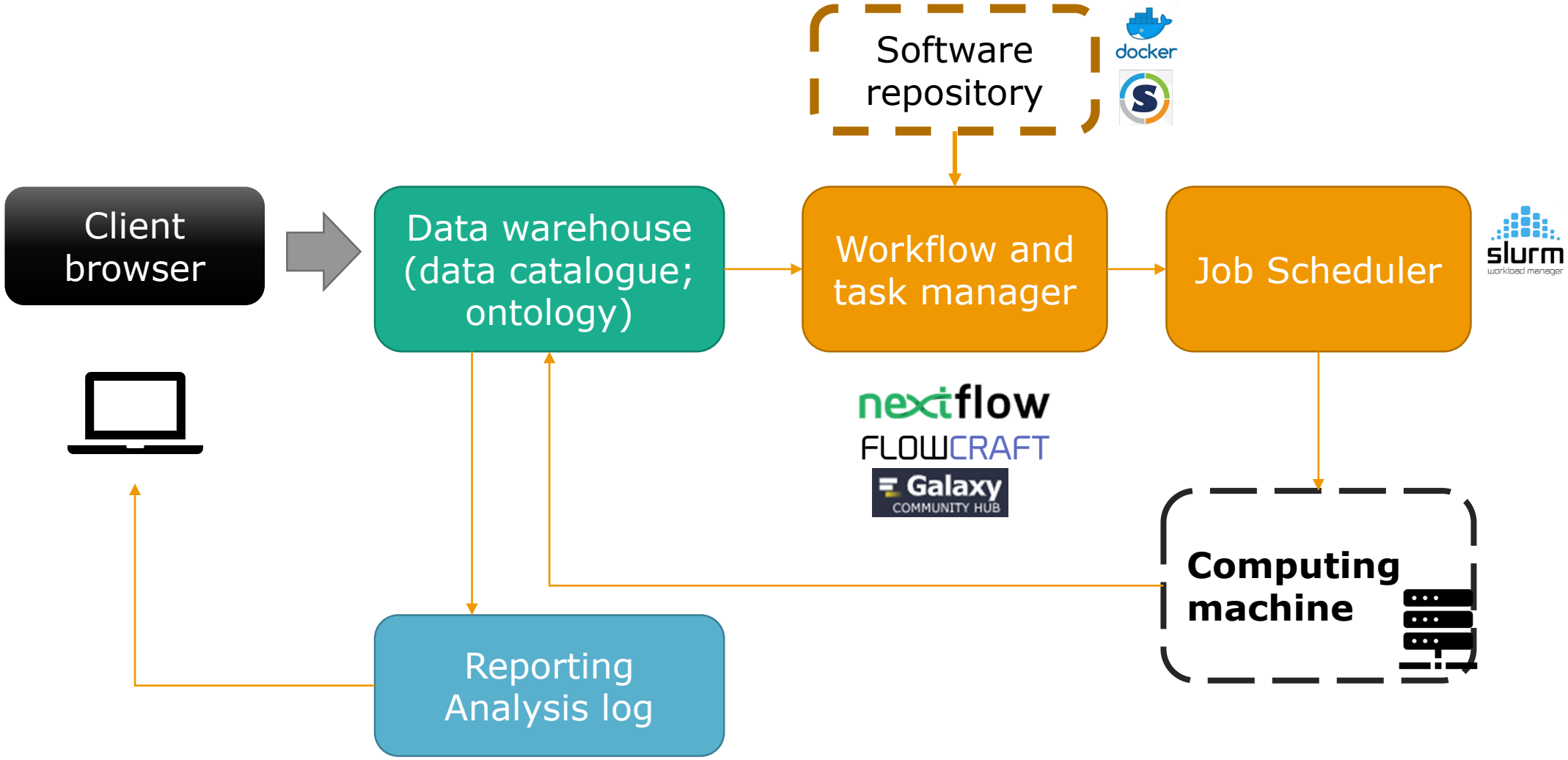
INNUENDO

EnteroBase

BIGSdb

- Facilitate the use of pipelines by non-bioinformaticians
- Facilitate managing and sharing of data
- Local installations or centralized servise-based
- Specialized (i.e. only cgMLST) or generalist (collection of software)

Structure of Bioinformatic platform



INNUENDO

efsa
European Food Safety Authority

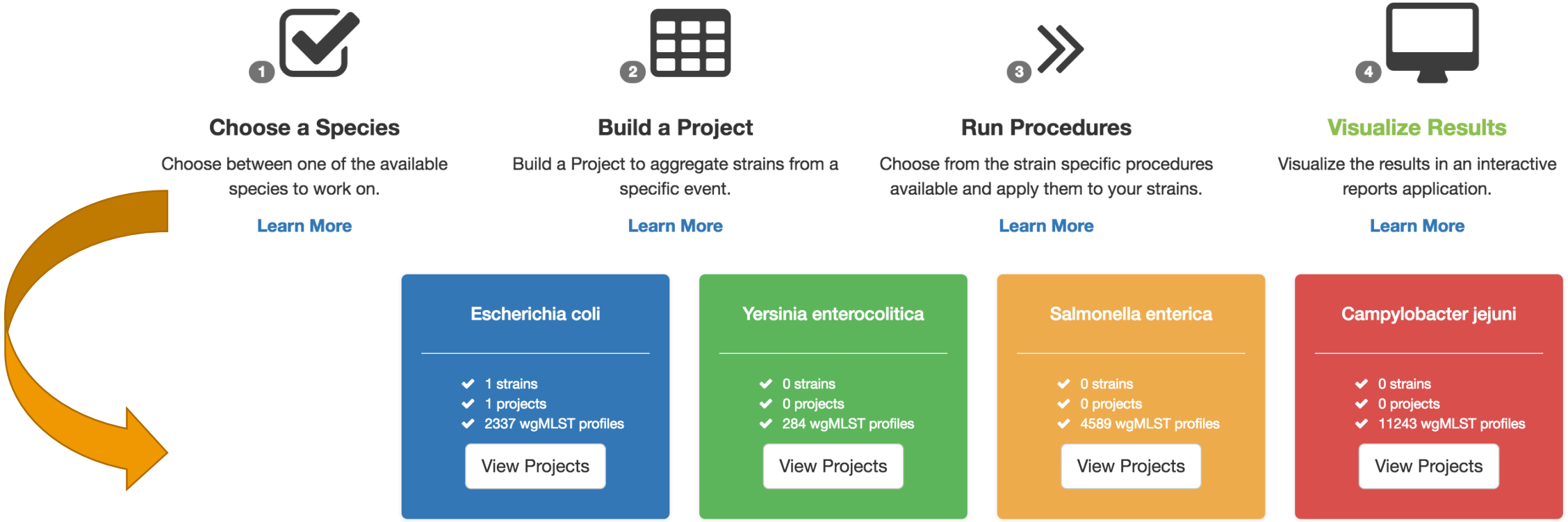
INNUENDO
www.innuendoweb.org

End-to-end management of genomic sequence data and metadata

It is designed for supporting outbreak detection and investigation by matching specific profiles.

Quality verified species-specific genomic databases.

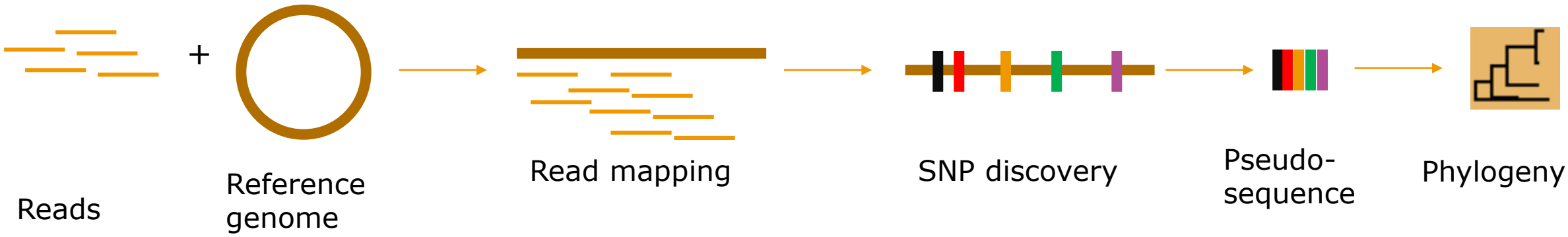
The query is done using allelic profiles



Pipelines for genomic epidemiology

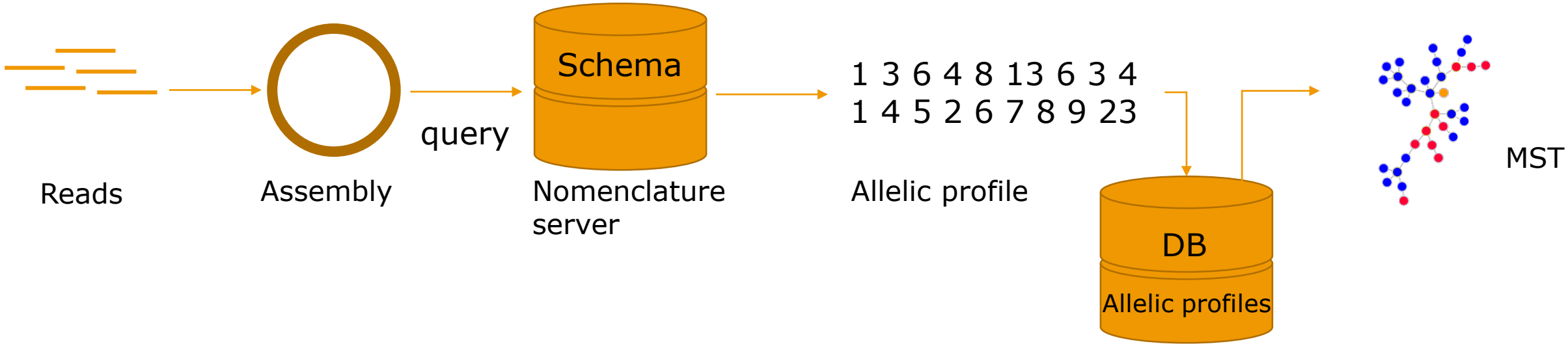
- **The aim**: identify a common ancestor for a set of isolates suspected to be related
- Phylogenetic inferring at higher resolution possible
- It takes times to accurate measure genetic variations
- Pipelines aim to simplify the process (operability)
- Different methodologies (often) reaching similar conclusions
- *Problems: lost in translation and not a complete understanding of the methods from the users*

Genomic epidemiology: SNP calling



- Read Mapping: BWA, Bowtie2, minimap2
- Variant calling: GATK, Freebayes
- Phylogeny: RAxML, IQTREE, FasTree
- **Pipelines: Snippy, Lyve-SET, PHEnix**

Genomic epidemiology: cg/wgMLST



- Expansion of the MLST concept: schemas are species or genus specific
- Mostly assembly based
- Query of the database is often performed using BLAST
- Workflows: **EnteroBase, BIGSdb, chewBBACA + commercials**

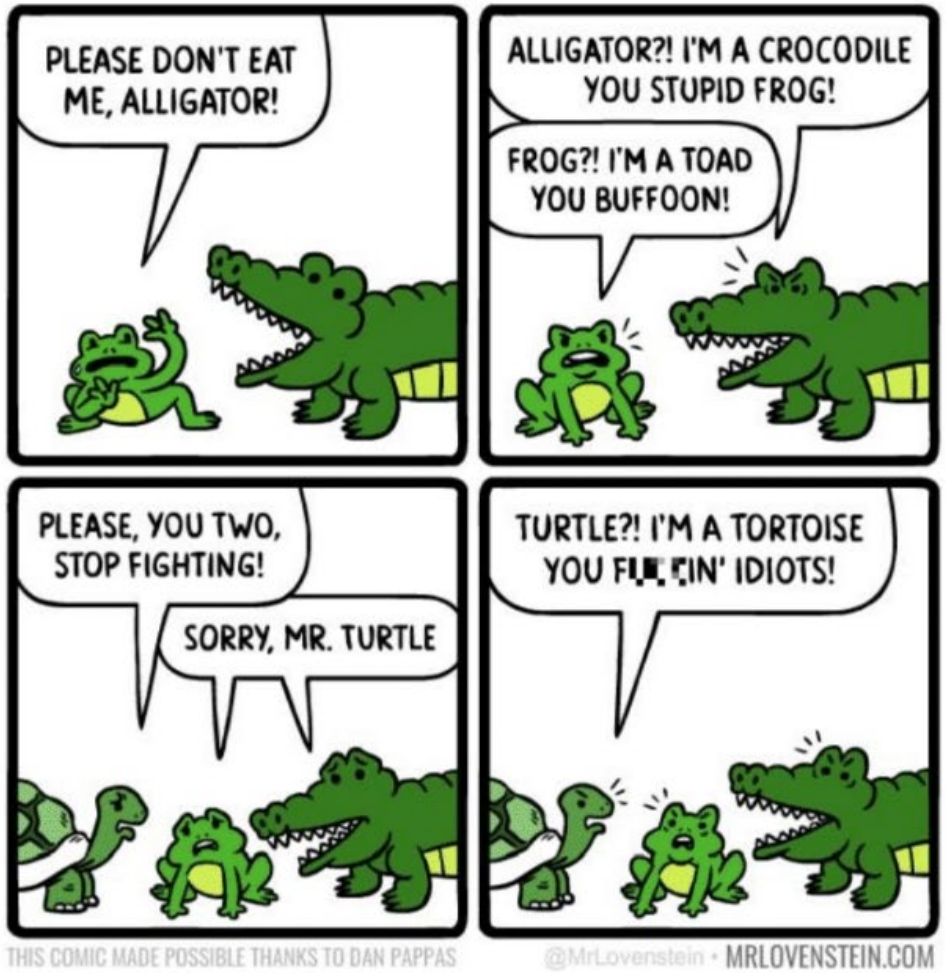
Standardization and harmonization: SNP calling

- The method requires accurate selection of a reference → affect results/resolution/classification
- Multiple aligners and variant calling software & plethora of parameters
- Multi mapping regions/recombination
- Critical values:
  - %coverage of the reference
  - ad-ratio (proportion of reads which support the base call at a specific location)
  - depth of coverage (number of reads covering a base)
- Strain classification

Standardization and harmonization: cg/wgMLST

- Assembly workflow (pre- and post-processing, and assembly algorithm) affects the final result
- The definition of “locus” and “allele”
- Which schema do you use?
- The curation of the schema
  - Not all the loci are “good” to be part of the schema
- Missing loci: particularly relevant in cgMLST
- Strain classification

Lost in translation



Communicate WGS-based typing results

- First step in outbreak investigation is assigning cases to cluster
- Frequently the (only) assumption is that low genetic differences imply recent transmission or common source
- Use of thresholds for communicating microbiological relationship
- Thresholds are source of uncertainties

The problem of the thresholds

- Problems arise in interpreting the relationship at +/- few differences around the cut-offs
- Thresholds for "WGS-based" epidemiological relationships are frequently **ill-defined** and **arbitrary**
- Validation based on biased datasets and biased assumptions → more studies are needed
- Framework frequently misses to consider organism-specific (including lineages) features and to allow for temporal and other epidemiological context

Challenges and factors to be considered when assigning cases to clusters

- Phylogenetic inferring (i.e. tree topology, branch lengths, genetic distance)
- Genetic diversity of population
- Selective pressure
- Mutation rate vs substitution rate
- Vertical inheritance vs horizontal gene transfer
- Role of coverage of mapping (x SNP analyses) or missing loci (x wg/cgMLST)

Possible solutions

- Probabilistic approach for inferring transmission and source attribution
- Building evidences from different data and analyses
  - How would you defend your position in an hypothetical law suit?
- Stress different hypotheses
- Enhance bioinformatics competences

Thanks for your attention

EFSA is committed to:

Excellence, Independency, Responsiveness and Transparency

www.efsa.europa.eu

Contacts:
mirko.rossi@efsa.europa.eu

Stay connected

Subscribe to
www.efsa.europa.eu/en/news/newsletters
www.efsa.europa.eu/en/rss

Engage with careers
www.efsa.europa.eu/en/engage/careers

Follow us on Twitter
@efsa_eu
@plants_efsa
@methods_efsa