sciensano

healthy all life long

Pathogen profiling in a clinical and public health context

The use of whole genome sequencing (WGS) for the typing of bacterial pathogenic isolates

Capita selecta in Bioinformatics
17/02/2020

Dr. ir. Sigrid De Keersmaecker – Transversal activities in Applied Genomics

.be

Overview

- Sciensano, the Belgian Scientific Institute of Public Health
- Context pathogen profiling
- Case study: The use of whole genome sequencing (WGS) for the typing of bacterial pathogenic isolates
- Bioinformatics @ Sciensano
- Constraints, including working in a quality system (ISO)

sciensano

.be

Who is Sciensano?

- Belgian federal research centre (established April 2018)
- Merger between ISP WIV (Scientific Institute of Public Health - Belgium) & CODA - CERVA (Veterinary counterpart)
- >1 century of scientific expertise in human and animal health
  - 700 staff members
- Support to policy, professionals and citizens through innovative research, analyses, monitoring activities and expert advice

sciensano .be

Our core: One health and beyond

Core: "One health" concept

Holistic view on health

Combining different research perspectives and disciplines to prevent, evaluate and mitigate health problems.

Unique contribution to health

sciensano

Sciensano : fields of expertise

Sciensano's scientists explore 6 fields of action

- Animal health
- Effectiveness and safety of vaccines, medicines and health products - Quality of medical laboratories
- Food consumption and food safety
- Health and disease monitoring
- Health and environment
- Quality of healthcare

- emerging diseases and diseases
- Quality control of medicines (incl.
- Safety of the food chain
- Status of health of population
- Monitoring circulating diseases & their causes
- Biomonitoring (e.g. Wetteren)
- National health survey
- Registers
- Cancer center
- Rare diseases

Sciensano (01/04/2018)

sciensano .be

Transversal activities in Applied Genomics

Objective? Transversal unit developing and implementing molecular biology and bioinformatics to perform both routine analyses, and scientific and technological research.

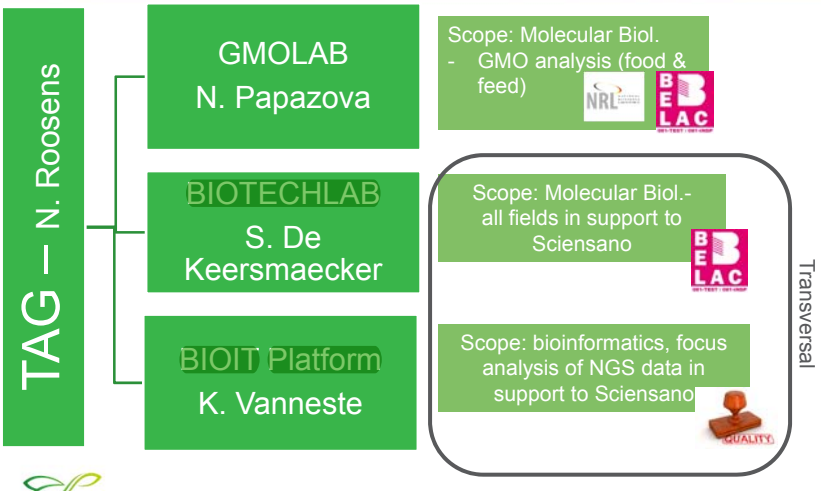
How? By setting up internal and external partnerships, we generate new knowledge and tools tailored to face and anticipate the current and future challenges affecting human and animal health.

Activities?

- in the domain of molecular detection, identification and characterisation of (pathogenic) micro-organisms (bacteria, viruses, and fungi), biotech organisms (GMO) and human nucleic acid biomarkers.
- high-tech equipment & expertise (e.g. real time & digital PCR, Sanger and NGS sequencers).
- bioinformatic tools & expertise

sciensano .be

Organigram



sciensano .be

Sciensano in figures

700 staff members

10 patents
in the field of public health

110 parliamentary questions
processed annually

260 peer-reviewed articles
published annually

More than 130 000 sample analyses
conducted annually

30 surveys, studies and registries
to monitor the evolution of public health

135 scientific projects
financed by Belgian and international partners

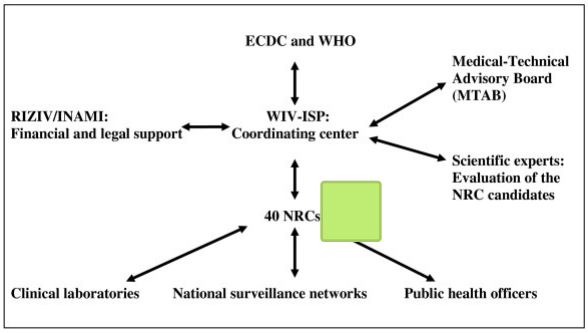
More than 80 reference centres and labs

sciensano .be

National reference centers & laboratories

Tasks: service, research and expertise for the Belgian authority

National reference centers (human) (NRC): RIZIV/INAMI – ECDC - WHO



National reference laboratories (food) (NRL): FAVV/AFSCA - EFSA

sciensano .be

Context pathogen profiling

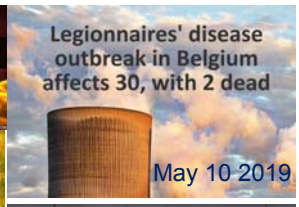
Table 1 | **Classes of determinants for pathogen profiling**

Class of determinant	Data type	Uses
Pathogen identification	Presence of pathogen, genus and species-specific gene	Confirmation of identity of a pathoger
Virulence	Presence or absence of individual genes or mutants associated with virulence	Primary risk assessment or outcome prediction*
Transmissibility	Presence or absence of individual genes associated with transmissibility	Secondary risk assessment or outcom prediction*
Antimicrobial resistance	Presence or absence of individual genes or mutations associated with resistant phenotype	Treatment response prediction
Clonality	Genotypes and epidemiological data	Confirmation of epidemiological links or generation of hypotheses about relationships in the absence of epidemiological data[1]; Tracking geographical and temporal spread of pathogens of public health importance
Clinical information	Patient's demographics and location, laboratory number	Unique identifier, temporal and geo-positioning

Sintchenko et al., 2007 Nature Rev Microbiol

sciensano .be

Outbreak



VRT NWS

Sept 10th, 2019

WESTVLAANDEREN

150 leerlingen en leerkrachten Spermalie Brugge ziek na besmetting met salmonella

Tartaarsaus oorzaak salmonella-uitbraak in hotelschool Spermalie

Sept 23rd, 2019

Verse tartaarsaus is de bron van de salmonella-uitbraak in de Brugse hotelschool Spermalie begin september.

Op vrijdag 6 september en in de daaropvolgende dagen zijn in de hotel- en toerismeschool spermalie in brugge een tweehonderdtal leerlingen en leerkrachten ziek geworden. Uit laboratoriumonderzoek van stoelgangstalen bleek al snel dat spermalie getroffen werd door een uitbraak van salmonella. Daarop werden de stalen genalyseerd van de maaltijden die in het schoolrestaurant werden geserveerd. Ook werd er een online enquête bij de leerlingen en leerkrachten opgestart om na te gaan wie in welk restaurant wat gegeten had.

.be

Case study:

The use of whole genome sequencing (WGS) for the typing of bacterial pathogenic isolates

SCIENTIFIC OPINION

EFSA Journal

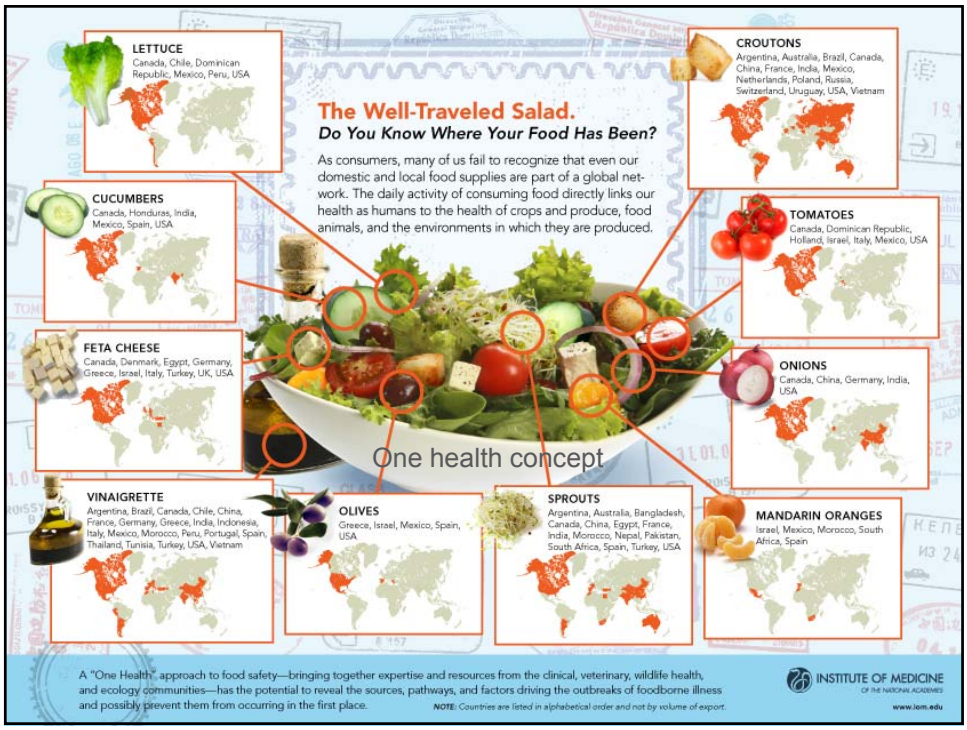
ADOPTED: 23 October 2019

doi: 10.2903/j.efsa.2019.5898

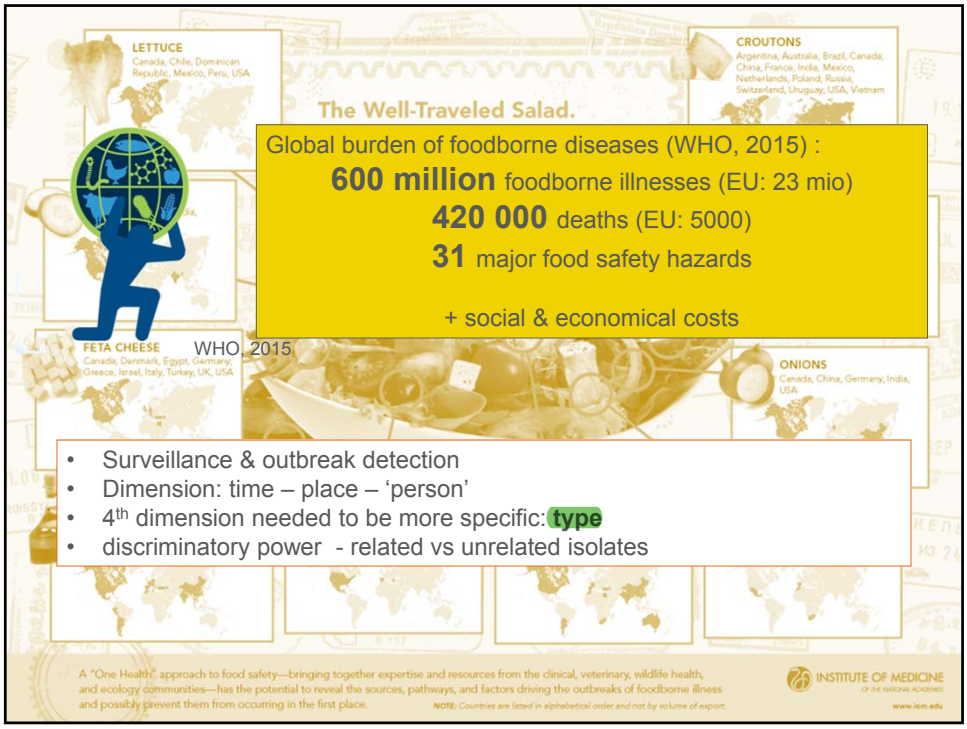
Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms

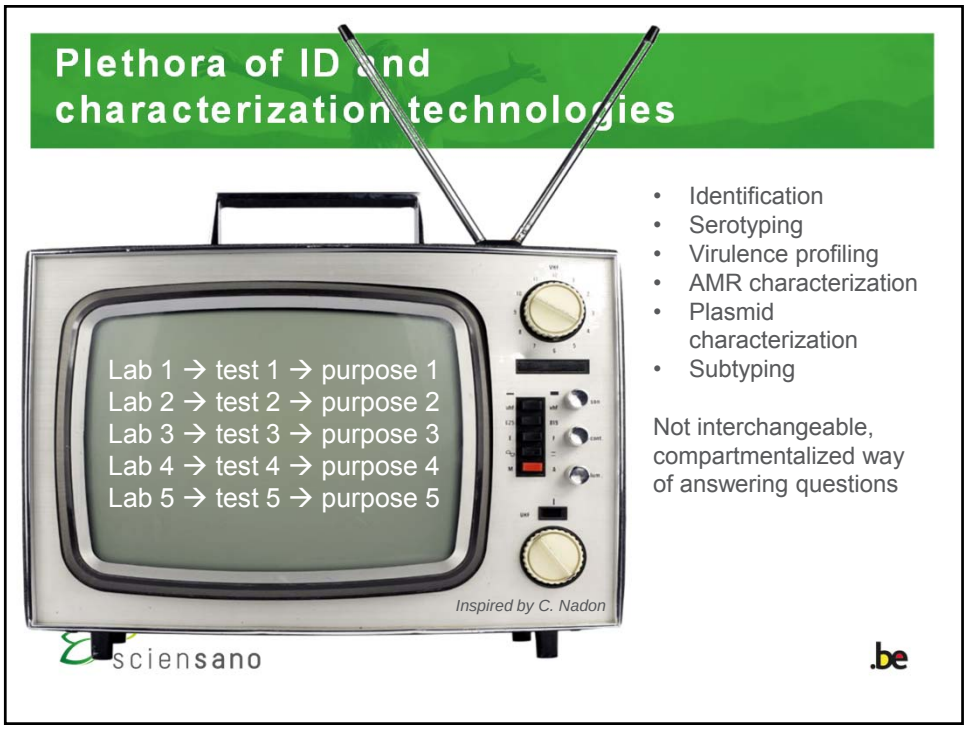
EFSA Panel on Biological Hazards (EFSA BIOHAZ Panel), Kostas Koutsoumanis, Ana Allende, Avelino Alvarez-Ordóñez, Declan Bolton, Sara Bover-Cid, Marianne Chemaly, Robert Davies, Alessandra De Cesare, Friederike Hilbert, Roland Lindqvist, Maarten Nauta, Luisa Peixe, Giuseppe Ru, Marion Simmons, Panagiotis Skandamis, Elisabetta Suffredini, Claire Jenkins, Burkhard Malorny, Ana Sofia Ribeiro Duarte, Mia Torpdahl, Maria Teresa da Silva Felicio, Beatriz Guerra, Mirko Rossi and Lieve Herman

sciensano .be

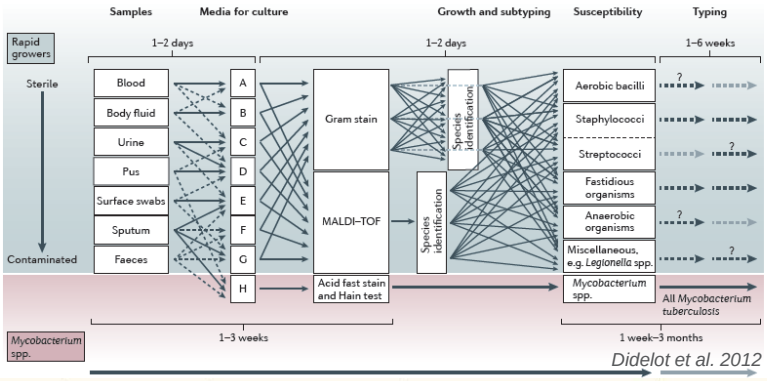


One health concept

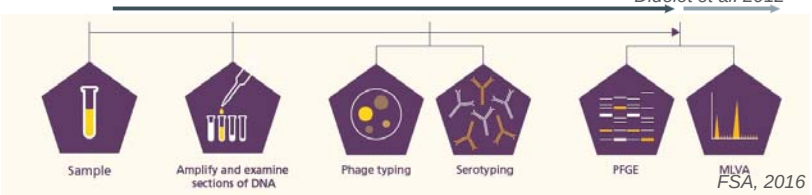

The Well-Traveled Salad.
LETTUCE
Canada, Chile, Dominican Republic, Mexico, Peru, USA
CROUTONS
Argentina, Australia, Brazil, Canada, China, France, India, Mexico, Netherlands, Poland, Russia, Switzerland, Uruguay, USA, Vietnam
FETA CHEESE
Canada, Denmark, Egypt, Germany, Greece, Israel, Italy, Turkey, UK, USA
ONIONS
Canada, China, Germany, India, USA
Global burden of foodborne diseases (WHO, 2015) :
600 million foodborne illnesses (EU: 23 mio)
420 000 deaths (EU: 5000)
31 major food safety hazards
+ social & economical costs
WHO, 2015
• Surveillance & outbreak detection
• Dimension: time – place – 'person'
• 4th dimension needed to be more specific: type
• discriminatory power - related vs unrelated isolates
A "One Health" approach to food safety—bringing together expertise and resources from the clinical, veterinary, wildlife health, and ecology communities—has the potential to reveal the sources, pathways, and factors driving the outbreaks of foodborne illness and possibly prevent them from occurring in the first place.
NOTE: Countries are listed in alphabetical order and not by volume of export.
INSTITUTE OF MEDICINE
www.iom.edu


Plethora of ID and characterization technologies
Lab 1 → test 1 → purpose 1
Lab 2 → test 2 → purpose 2
Lab 3 → test 3 → purpose 3
Lab 4 → test 4 → purpose 4
Lab 5 → test 5 → purpose 5
Inspired by C. Nadon
• Identification
• Serotyping
• Virulence profiling
• AMR characterization
• Plasmid characterization
• Subtyping
Not interchangeable, compartmentalized way of answering questions
sciensano
.be

Timescale & total cost traditional methods



Didelot et al. 2012



FSA, 2016

Few weeks to few months

Each test adding to the total cost

.be

Subtyping of pathogens

What?

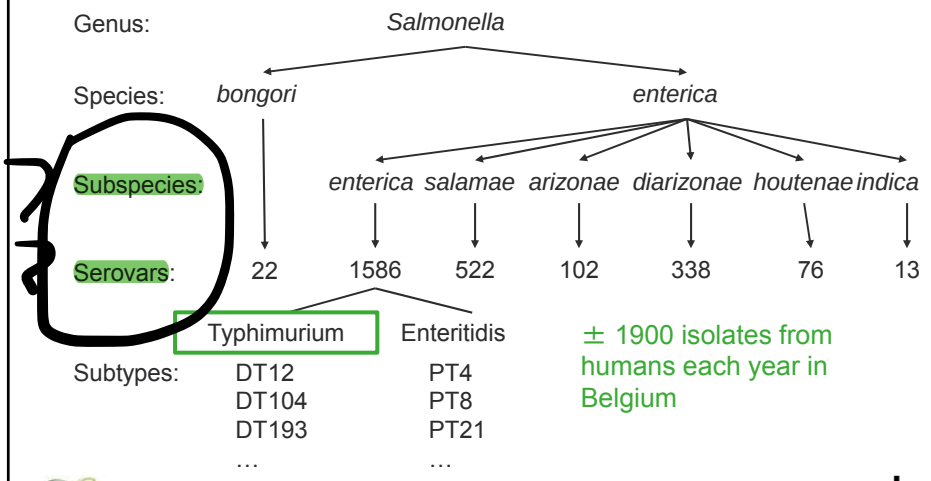
Characterisation of bacterial isolates beyond the species and subspecies level

→ Case study: *Salmonella* subtyping
important foodborne pathogen

sciensano

.be

Salmonella subtypes

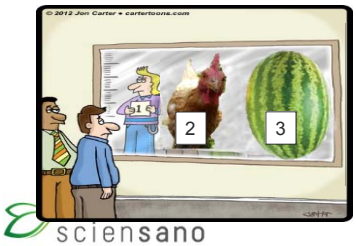


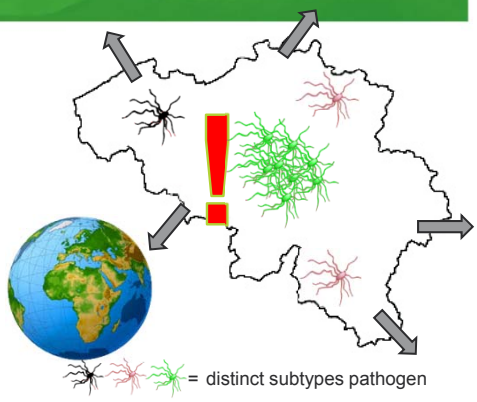
± 1900 isolates from humans each year in Belgium

sciensano .be

Subtyping: why needed?

- Surveillance: dissemination and incidence of pathogens
- Outbreak detection
- Trace the source of an outbreak
  → confinement of outbreak
  (link human and food isolate)





National Reference Centres (NRC)
National Reference Laboratories (NRL)

→ international network (ECDC, EFSA and WHO)

sciensano .be

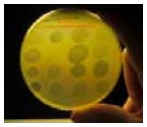
Ideal global public health molecular tool

- Robust
- Easily implemented and standardised
- Sufficiently discriminative
- Objective data → appropriate interpretation and transfer between laboratories, reproducibility & portability
- Rapid
- Inexpensive
- Universally applicable

sciensano .be

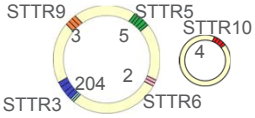
Classical subtyping methods for *Salmonella*

Phage typing



+ inexpensive
+ discriminative
- high level of expertise required
- interpretation of lysis patterns is subjective
- NT and RDNC strains

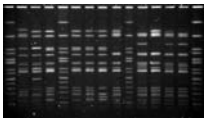
MLVA



Profile: 3-5-2-4-204

+ rapid
+ profiles easily compared between laboratories
- relatively expensive
- only available for small number of pathogens
- data analysis with commercial software
- too discriminative for *Salmonella* Typhimurium

PFGE



+ medium discriminative
+ public database available
- relatively slow
- limited reproducibility
- data analysis with commercial software

sciensano

+ AMR, (q)PCR virulence genes….

.be

MLST: multilocus sequence typing

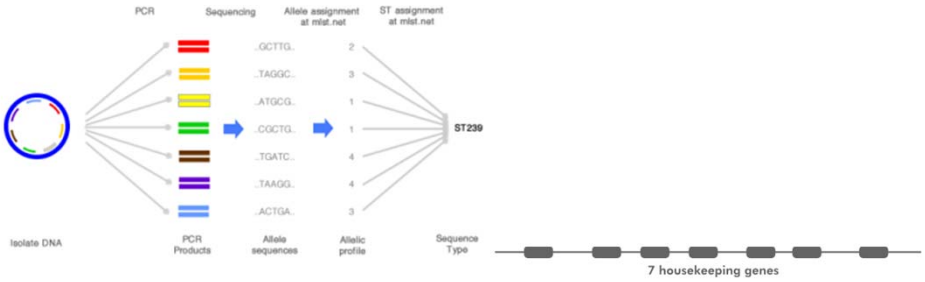
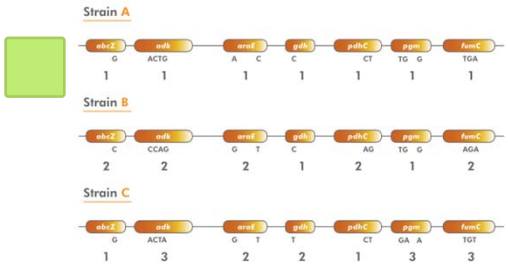
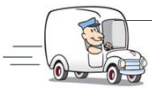


Fig. MLST.net



sciensano

A day in the life of a reference lab…



Salmonella (5,000 / year) | ***Shigella*** (400 / year) | ***Listeria*** (80 / year) | ***N. meningitidis*** (100 / year)

Species confirmation

- Biochemistry
- MALDI-TOF / PCR (neg. result)

Serotyping

SELECTION (1,000 / y)

Subtyping

- Drug Susceptibility Testing
- PCR/Seq-based identification of resistance genes (QRDR, ESBLs)
- Genotyping using PFGE and MLST/MLVA

Why not immediately replacing all this by WGS?

Slide from Dr. Ir. W. Mattheus, Unit Bacterial Diseases– Sciensano

Whole genome sequencing (WGS)

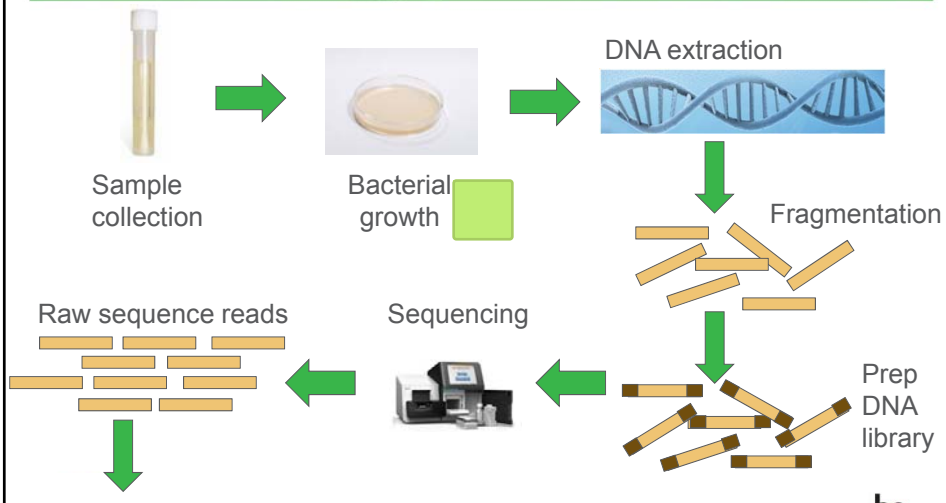
- fast
- conceptual simple workflow
- ONE universal method (i.e. species independent)
- high resolution (maximal strain discrimination)
- multiplex
- no *a priori* knowledge
- more and more affordable
- 1 test to answer >1 question at a time:
  - Identification
  - Define lineage & surveillance
  - Predict resistance
  - Serotyping
  - Virulence profiling
  - AMR characterization
  - Plasmid characterization
  - Subtyping – outbreak investigation

« One method to rule them all… »

sciensano

.be

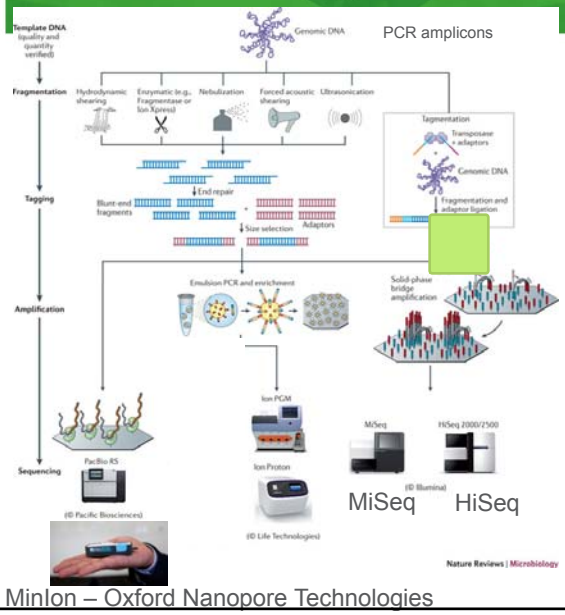
Workflow WGS



DATA analysis (identification, subtyping, virulence marker detection, AMR predictions …)

.be

WGS workflow – data generation



MiSeq HiSeq

MinIon – Oxford Nanopore Technologies

Short reads: MiSeq (~56h), HiSeq
Long reads: MinION (real-time, 15 min-24h)

.be

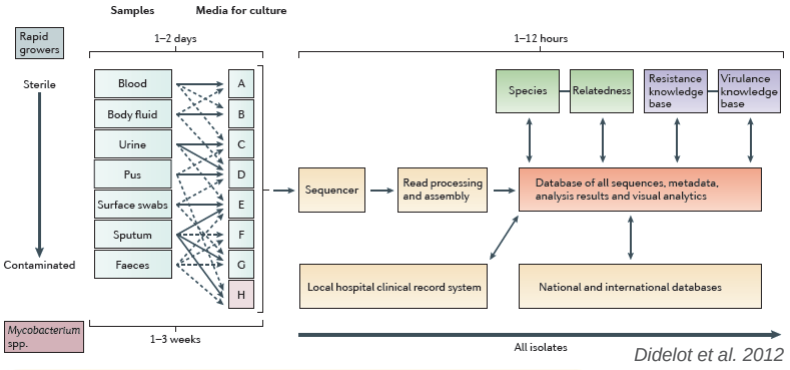
In silico typing
Backward comparability

'Classical' technique	Via WGS?	Comment
PCR-based	**yes**	multiplex
Antimicrobial susceptibility testing	**yes**	Likely to be predicted from WGS with sufficient accuracy, depending on mechanism
Serotyping	**yes**	Likely to be predicted from WGS with sufficient accuracy
Multiple-locus VNTR analysis (MLVA)	no	Cannot yet be derived from routine WGS data, potentially in future
Multilocus sequence typing (MLST)	**yes**	Can be derived unambiguously from WGS data
Pulse-field gel electrophoresis (PFGE)	no	Cannot be derived from routine WGS, to be determined if possible in the future
WGS for subtyping	**yes**	SNP-based for high resolution

sciensano

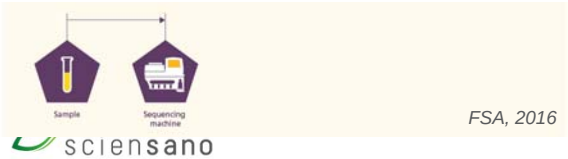
.be

Timescale & total cost WGS



Few days (expected: 1 day)

Didelot et al. 2012



FSA, 2016

sciensano

.be

Cost and timeliness typing methods

Pathogen	Typing method	Weight	Countries	Samples per year	Consumables per isolate (EUR)	Operator time per isolate (h)	Total time per isolate (d)
Salmonella	Isolation, confirmation and pure culture	100%	5–7	18315–18465	1–**4**–15	0.1–**0.5**–1.5	1–**1.5**–4
	Serotyping agglutination[4]	100%	6–8	21430–21580	5–**25**–55	0.08–**0.5**–1	<1–**2**–17
	MLVA Typhimurium and Enteritidis	67%	6–8	2395–2430	7–**17.5**–30	0.67–**1**–3	1–**1.5**–2
	PFGE XbaI	33%	6–8	1165–1435	20–**25**–42	0.36–**0.9**–1	2–**3.5**–6
	Resistance testing	0%	5–6	4630–4730	1.5–**9**–20	0.17–**0.5**–1	1–**2**–3
	Total[5]	**N/A**	**N/A**	**N/A**	**17–49–104**	**0.7–2–4.8**	**3.3–5.7–24.3**
Listeria monocytogenes	Isolation, confirmation and pure culture	100%	4–5	700–775	1.5–**3**–20	0.1–**0.69**–3	1–**2**–6
	Serotyping agglutination	33%	4–5	143–148	10–**14**–40	0.25–**0.49**–1	<1–**1**–2
	Serotyping PCR	67%	5–6	290–360	1.2–**6.5**–35	0.33–**1**–3.5	0.09–**1**–2
	PFGE ApaI+AscI	100%	7–8	440–515	24–**30**–60	1–**2**–2.8	3–**3.5**–6
	Total[5]	**N/A**	**N/A**	**N/A**	**30–42–117**	**1.4–3.5–8.5**	**4.4–6.5–14**
Campylobacter	Isolation, confirmation and pure culture	100%	4–5	1015–1165	1.5–**3**–20	0.1–**0.69**–3	1–**2**–6
	Species typing	0%	4–6	2000–2150	1–**15**–30	0.07–**0.5**–2	<1–**1**–1
	MLST	100%	3–4	650	25–**61.6**–250	3–**3.5**–5	2–**3**–5
	Total	**N/A**	**N/A**	**N/A**	**27–65–270**	**3.1–4.2–8**	**3–5–11**
E. coli	Isolation, confirmation and pure culture	100%	6–7	2980–3150	1–**6**–35	0.1–**0.75**–3	1–**2**–3
	vtx1+vtx2+eae gene detection (not subtyping)	100%	7–8	4105–4275	7–**20**–40	0.7–**1**–2	0.5–**1**–2
	O-group typing agglutination	100%	6–8	2285–2375	5–**10**–85	0.2–**1**–2.5	1–**2**–7
	PFGE XbaI	100%	7–8	765–805	20–**25**–50	0.36–**0.9**–1.27	3–**4**–12
	ESBL production	0%	2–4	303–403	10–**13.5**–15	0.25–**0.5**–2	1–**1.5**–2
	Total[5]	**N/A**	**N/A**	**N/A**	**33–61–210**	**1.4–3.7–8.8**	**5.5–9–24**
WGS[5]	Isolation, confirmation and pure culture	100%	19–24	N/A	1–**6**–35	0.1–**0.5**–3	1–**2**–3
	DNA Extraction	100%	10–19	N/A	1.5–**5**–10	0.5–**0.67**–3	0.25–**1**–2
	Library preparation and sequencing	100%	6–8	N/A	48.3–**80**–151	0.09–**0.48**–1	2–**3**–12
	Data analysis	100%	-	-	-	-	-
	Total	**N/A**	**N/A**	**N/A**	**51–91–196**	**0.7–1.7–7**	**3.25–6–17**

Price: assumption that full batch
WGS: no standard yet
Equipment costs not included (min 17 euro per isolate for WGS)
Data storage costs not included (min 2.7 euro/isolate)

E. coli & *Campylobacter*: less costly
Listeria: cost idem
Salmonella: higher cost, depending on throughput

ECDC – Expert opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA, 2015

sciensano

.be

Economic impact of using WGS for foodborne outbreaks

An economic analysis of salmonella detection in fresh produce, poultry, and eggs using whole genome sequencing technology in Canada

Sonali Jain[a], Kakali Mukhopadhyay[a,b,*], Paul J. Thomassin[a]

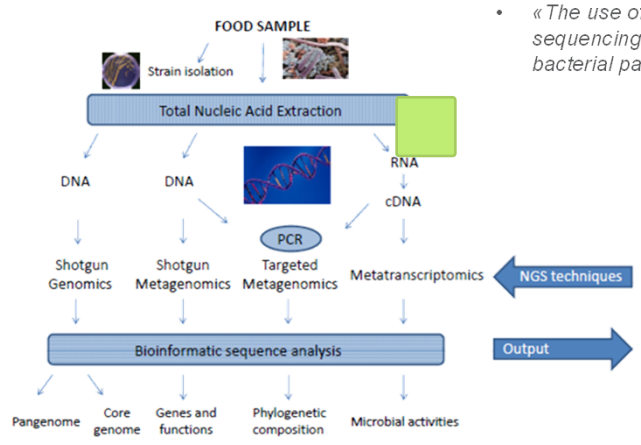
[a] Department of Agricultural Economics, McGill University, Canada
[b] Gokhale Institute of Politics and Economics, Pune 411004, India

	Canada
Incidence of illness	47082 per year
Direct healthcare costs (hospitalization, physician, traditional lab costs PFGE, medication)	$6.43 mio ($98.7 mio value of statistical life)
Indirect costs (productivity loss)	$21.13 mio
Federal costs	$161.44 mio
Total illness cost traditional methods	$287.78 mio
Total net benefit use WGS	**$90.25 mio**

Belgium: 24.81 mio euro ? (Salmonellosis only)

sciensano .be

NGS for food microbiology



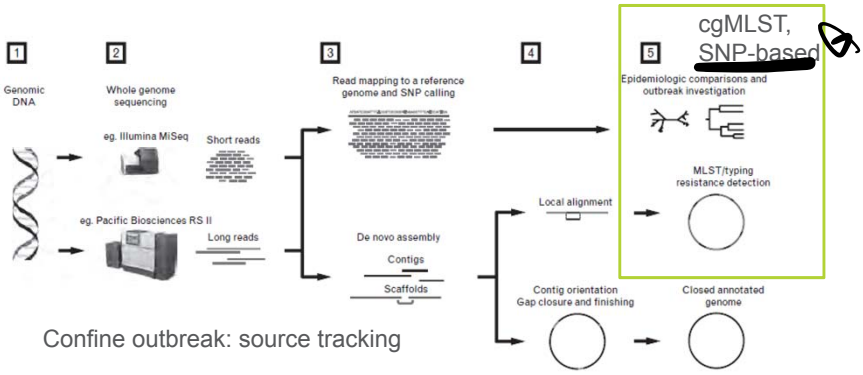
- *« The use of whole genome sequencing (WGS) for the typing of bacterial pathogenic isolates »*

- WGS vs NGS
- Isolate vs population
- Bacterial vs virus
- Pathogenic

Mayo et al. 2014

sciensano .be

WGS workflow – data analysis



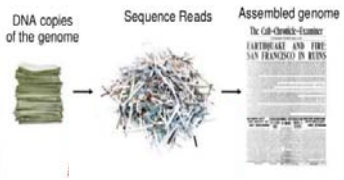
Confine outbreak: source tracking

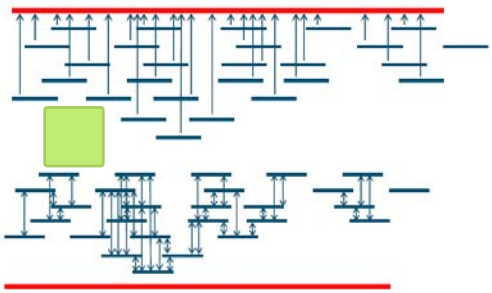
Kwong et al. 2015

Human isolate Food isolate

Some terms

- **Read mapping**: map the reads against a reference genome (e.g. BWA or Bowtie2)
- **Assembly**: assembly the reads into *de novo* genomes without knowing *a priori* which species is present (e.g. Velvet and SPAdes)





Some terms (2)

- **Coverage**: average number of times the data is covered in the genome
  - **C = N.L/G**
  - N = number of reads
  - L = read length
  - G = genome size (target or assembly)

Example:
N = 5 million
L = 100 bp
G = 5 Mb
C = 5*100/5 = 100x
On average, 100 reads covers each position in the genome

- **Depth**: number of reads that co[illegible]a particular nucleotide in each position in the genome
  - Reads/site = depth

sciensano .be

Deriving phylogeny
Some considerations

- Assess whether 2 or more isolates originate from an epidemiologically relevant **common ancestor** (**source**) (dendograms)
- Entire genome available for analysis: not straightforward analysis & **interpretation** for epidemiologists
- Amount of diversity generated in one unit of time depends on the **rate of reproduction and the evolutionary pressure** that organisms are under (environment specific) -> highly unlikely that the number of SNPs difference found between 2 isolates can be reliably related to the number of days that have passed since they diverged from a common ancestor
- Distantly related strains may still be epidemiologically linked (several contamination foci along the same food chain)

sciensano .be

Deriving phylogeny
Two approaches

SNP-based typing/relatedness

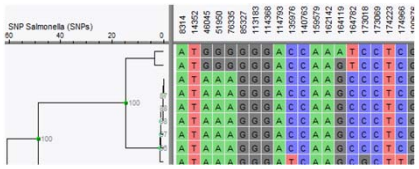
- Best estimate/approximation for the true evolutionary distance -> very high accuracy at the highest possible resolution (SNP)
- Novel technique (databases, standards schemes etc. lacking)

(wg)MLST-based typing/relatedness

- Lower resolution but very robust approach
- MLST established technique for which WGS can be used -> expand to wgMLST to include core and/or pan genome (databases, standards schemes etc. however lacking – under development)

sciensano .be

Phylogeny based on SNPs

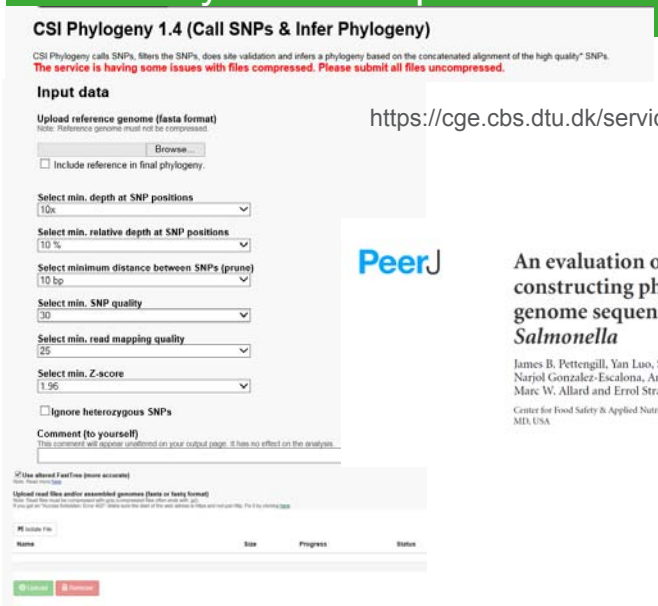


Bionumerics

- Occur with a relatively well understood frequency
- Derived from alignment with reference genome, also reference-free methods available (kmer-based), but less performant
- Difference with sequencing error/mapping error -> filtering out false SNPs (e.g. consecutive SNPs, SNPs at end of read)
- Computationally demanding
- Most rewarding for closely related isolates (low proportion of false SNPs): highly accurate, highest possible resolution (e.g. after cgMLST)
- Different algorithms/pipelines possible – different outcome?

sciensano .be

SNP analysis - examples



https://cge.cbs.dtu.dk/services/CSIPhylogeny/

PeerJ

An evaluation of alternative methods for constructing phylogenies from whole genome sequence data: a case study with *Salmonella*

James B. Pettengill, Yan Luo, Steven Davis, Yi Chen, Narjol Gonzalez-Escalona, Andrea Ottesen, Hugh Rand, Marc W. Allard and Errol Strain

Center for Food Safety & Applied Nutrition, U.S. Food & Drug Administration, College Park, MD, USA

.be

SNP - interpretation

- **SNP Distance**

  How close are the isolates? No single threshold for all species/types - need to incorporate epidemiological data

  Rough, conservative guides @ FDA

  - Inclusion: <=20 SNPs match, virtually identical (examples outbreaks – at most 1 SNP apart)
  - Inconclusive: 20-100 SNPs
  - Exclusion: > 100 SNPs exclude

  remark: Outbreak – 2 isolates 100 SNPs different, might not be considered related if other clustered isolates were differing by 10 SNPs. However if 90 of the 100 SNPs co-located within a small genomic segment suggestive of a recombination event, still represent evidence of transmission

- **Bootstrapping**

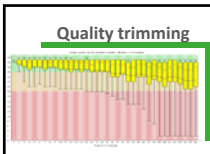
  Do the isolates form a unique cluster w/ >= 95% support?

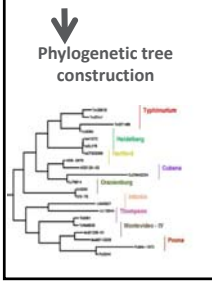
  Is the cluster distinct from other isolates in the tree?

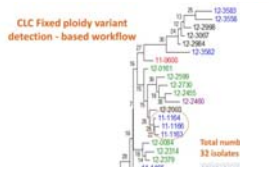
sciensano

.be

SNP – different workflows, different results







PLOS ONE | https://doi.org/10.137

not be possible to define a single cut-off value for delineation of an outbreak
Threshold = workflow dependent; variability of observed SNP distances depending on the analysis settings need to be investigated, taken into account in the decision-making process.

PhD
Assia Saltykova

PLOS ONE

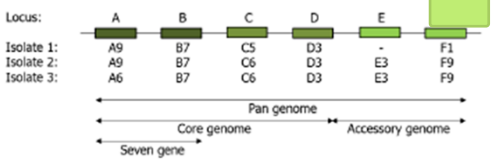
RESEARCH ARTICLE

Comparison of SNP-based subtyping workflows for bacterial isolates using WGS data, applied to *Salmonella enterica* serotype Typhimurium and serotype 1,4,[5],12:i:-

Assia Saltykova[1,2], Véronique Wuyts[1], Wesley Mattheus[3], Sophie Bertrand[3], Nancy H. C. Roosens[1], Kathleen Marchal[2,4,5]*, Sigrid C. J. De Keersmaecker[1]*

Phylogeny based on allele differences
(wgMLST - cgMLST)

a) Assignment of unique allele identifiers



b) Distance between isolates according to different schemes

Isolate pair	Seven gene	Core genome	Pan genome
1-2	0	1	3
1-3	1	2	4
2-3	1	1	1

c) Assignment of sequence types

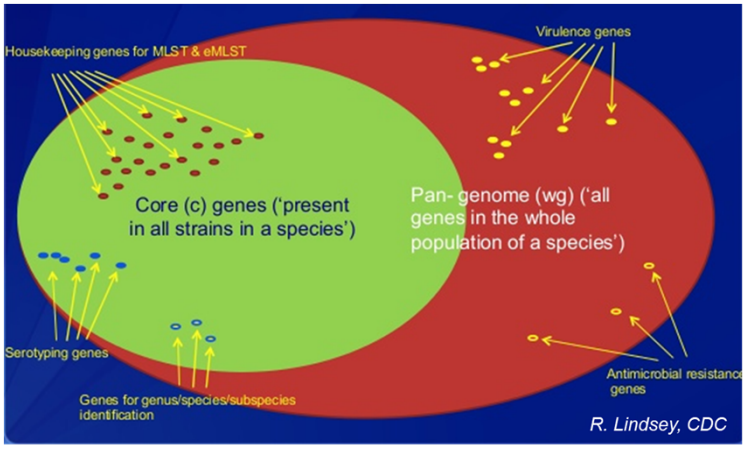
Isolate	Seven gene	Core genome	With sublevels (AB / CD)
1	23 (A9-B7)	115 (A9-B7-C5-D3)	23-9 (A9-B7 / C5-D3)
2	23 (A9-B7)	267 (A9-B7-C6-D3)	23-15 (A9-B7 / C6-D3)
3	47 (A6-B7)	349 (A6-B7-C6-D3)	47-93 (A6-B7 / C6-D3)

- Based on read mapping (better results) or assembly
- Only coding regions (excludes a large amount of potentially informative genomic information, such as phages, insertion sequences, and other mobile genetic elements that may indicate direct transmission)
- Allele sequences identical or different, regardless of the number of actual evolutionary events
- Missing loci if bad quality WGS data
- Reduced resolution, but more robust
- Database needed
- Pan/core/accessory genome

ECDC – Expert opini… introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA, 2015

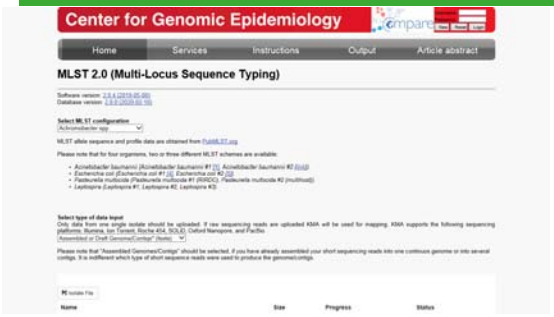
sciensano .be

MLST, cgMLST, wgMLST



sciensano .be

MLST analysis - examples



Inouye et al. Genome Medicine 2014, 6:90
http://genomemedicine.com/content/6/11/90

Genome Medicine

SOFTWARE Open Access

SRST2: Rapid genomic surveillance for public health and hospital microbiology labs

Michael Inouye[1,2], Harriet Dashnow[3,4], Lesley-Ann Raven[1], Mark B Schultz[3], Bernard J Pope[4,5], Takehiro Tomita[2,6], Justin Zobel[5] and Kathryn E Holt[1*]

sciensano .be

Nomenclature

- Short, yet still informative human readable code for isolate
- Same code = share properties
- Causal relationship between vehicle and human cases
- Efficient (international) communication between stakeholders
- Stability of system

sciensano .be

wgMLST nomenclature

c) Assignment of sequence types

Isolate	Seven gene	Core genome	With sublevels (AB / CD)
1	23 (A9-B7)	115 (A9-B7-C5-D3)	23-9 (A9-B7 / C5-D3)
2	23 (A9-B7)	267 (A9-B7-C6-D3)	23-15 (A9-B7 / C6-D3)
3	47 (A6-B7)	349 (A6-B7-C6-D3)	47-93 (A6-B7 / C6-D3)

ECDC , 2015

- Fixed set of loci for a particular species or genus, agreed upon by all parties
- Core genome (cgMLST), pan genome (wgMLST) – different for each species
- Unique alleles for these loci -> unique identifier in global nomenclature database accessible to all (allele nomenclature)
- Unique identifier to each combination of alleles observed (schemes) (sequence type – 7-gene MLST; sub-sequence types, covering the next N most stable loci per top level ST) (strain nomenclature)
- Allele calling algorithm (extract allele sequences from a new sequence for all defined loci, match against known allele sequences) & strain classification algorithm (determine type based on allele identifiers)
- No phylogenetic relationships (e.g. 23-25 and 47-93 with 1 SNP in 1 of the 7 HKG -> other place in tree)

MLST databases

a) wgMLST schemes

Scheme	Loci
PanGenome	A-B-C-D-E-F
CoreGenome	A-B-C-D
SevenGene	A-B
Sublevels	A-B / C-D

wgMLST locus definitions

RefGenome	Locus	Location
Ref1	A	50001-51123
Ref1	B	250001-251320
Ref1	C	300001-301566
Ref1	D	321285-322284

wgMLST allele sequences

Locus	Allele	Sequence
A	A1	ATGCATTAT…
A	A2	ATGCAATAG…
B	B1	ATGGGCCTG…
B	B2	ATGGGCTTG…

wgMLST sequence-type nomenclature

Scheme	SeqType	Alleles
CoreGenome	115	A9-B7-C5-D3
CoreGenome	267	A9-B7-C6-D3
SevenGene	23	A9-B7
Sublevels	23-15	A9-B7 / C6-D3

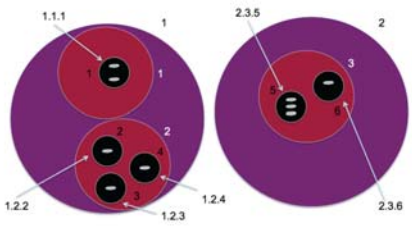
e.g. PubMLST

ECDC , 2015

sciensano .be

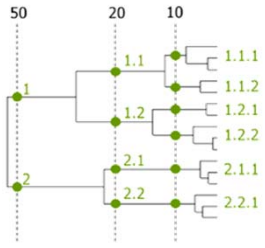
SNP nomenclature (taxonomical)

- Phylogenetic relationships
- Hierarchical description of where a particular isolate fits in a dendrogam that spans all isolates
- SNP address (PHE)

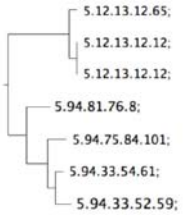


Ashton … Dallman et al, 2015, PHE

ECDC , 2015



Height cut-offs of 50, 20 and 10 correspond to the average number of SNPs or allele differences



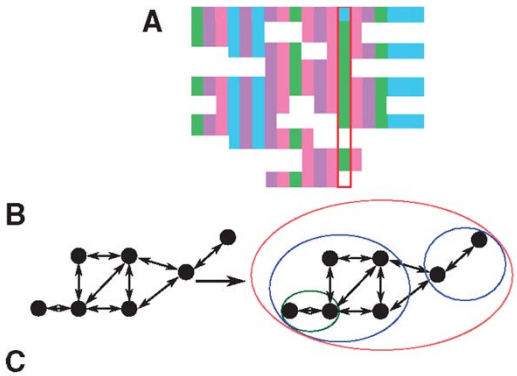
Specified index
e.g. 50:25:10:5:0

Each number represents the cluster membership at each descending SNP distance threshold. The resultant 'SNP address' provides an isolate level nomenclature where two isolates with the same SNP addresses have 0 SNP differences.

.be

Snapper DB (PHE)

(**A**) Illustration of SNP difference between a reference sequence (top row) and a set of isolate sequences (remaining row). (**B**) Single linkage clustering of SNP differences into 0, 5 and 10 SNP levels for a set of isolates. (**C**) Examples of SNP addresses based on the seven descending SNP thresholds; 250, 100, 50, 25, 10, 5, 0



C

SNP threshold	250	100	50	25	10	5	0
Isolate 1	1	1	1	158	199	222	243
Isolate 2	1	1	1	158	199	222	256
Isolate 3	1	2	2	35	60	125	160

Bioinformatics, Volume 34, Issue 17, 01 September 2018, Pages 3028–3029, https://doi.org/10.1093/bioinformatics/bty212
The content of this slide may be subject to copyright: please see the slide notes for details.

OXFORD UNIVERSITY PRESS

Bioinformatics, 34(17), 2018, 3028–3029
doi: 10.1093/bioinformatics/bty212
Advance Access Publication Date: 5 April 2018
Applications Note

OXFORD

Genome analysis

SnapperDB: a database solution for routine sequencing analysis of bacterial isolates

Timothy Dallman[1,*], Philip Ashton[1,2], Ulf Schafer[3], Aleksey Jironkin[3], Anais Painset[1], Sharif Shaaban[4], Hassan Hartman[1], Richard Myers[3], Anthony Underwood[3], Claire Jenkins[1] and Kathie Grant[1]

[1]Gastrointestinal Bacteria Reference Unit, National Infections Service, Public Health England, London, NW9 5EQ, UK, [2]Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam, [3]Infectious Disease Informatics, National Infections Service, Public Health England, London, NW9 5EQ, UK and [4]Roslin Institute, University of Edinburgh, Scotland, EH25 9RG, UK

sciensano

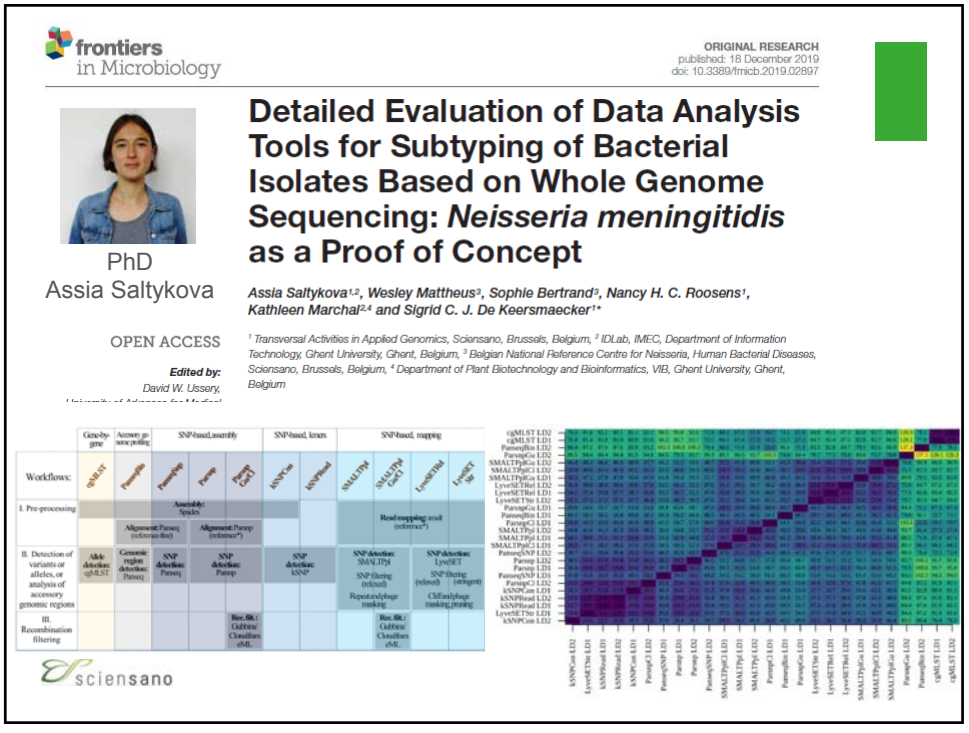
.be

MLST vs SNP

	SNP	MLST
Epidemiological concordance	High	High
Stable nomenclature	(No)	Yes
Reference characterization: identification, serotyping, virulence & resistance markers	No	Yes
Speed	Slow SNP calling, slow analysis	Slow allele calling, fast analysis
Local computing requirements	Medium-High	Low
Local bioinformatics expertise	Yes	No
Reference used to perform analysis	Most often sequence of closely related annotated strain (reference-free also possible)	Allele database
Requires curation	No	(Yes)

P. Gerner-Smidt, PulseNet

MLST = primary approach for public health (cfr. PulseNet); SNP used if more detail is needed or MLST fails
wgMLST & hqSNP = equally discriminatory, trees concordant



frontiers in Microbiology

ORIGINAL RESEARCH
published: 18 December 2019
doi: 10.3389/fmicb.2019.02897

PhD
Assia Saltykova

Detailed Evaluation of Data Analysis Tools for Subtyping of Bacterial Isolates Based on Whole Genome Sequencing: *Neisseria meningitidis* as a Proof of Concept

*Assia Saltykova[1,2], Wesley Mattheus[3], Sophie Bertrand[3], Nancy H. C. Roosens[1], Kathleen Marchal[2,4] and Sigrid C. J. De Keersmaecker[1]**

[1] Transversal Activities in Applied Genomics, Sciensano, Brussels, Belgium, [2] IDLab, IMEC, Department of Information Technology, Ghent University, Ghent, Belgium, [3] Belgian National Reference Centre for Neisseria, Human Bacterial Diseases, Sciensano, Brussels, Belgium, [4] Department of Plant Biotechnology and Bioinformatics, VIB, Ghent University, Ghent, Belgium*

OPEN ACCESS

Edited by:
David W. Ussery,

sciensano

Phenotype prediction

- Antimicrobial resistance
- Virulence
- Serotype

- Complexity of prediction dependent on cellular process involved (single gene for antimicrobial resistance vs. Multiple virulence genes in *E. coli* or AMR through mutations or *Salmonella* (…en)
- Associated phenotypic data r…d to 'train' predictive engine

sciensano .be

Mining WGS for AMR

Repositories

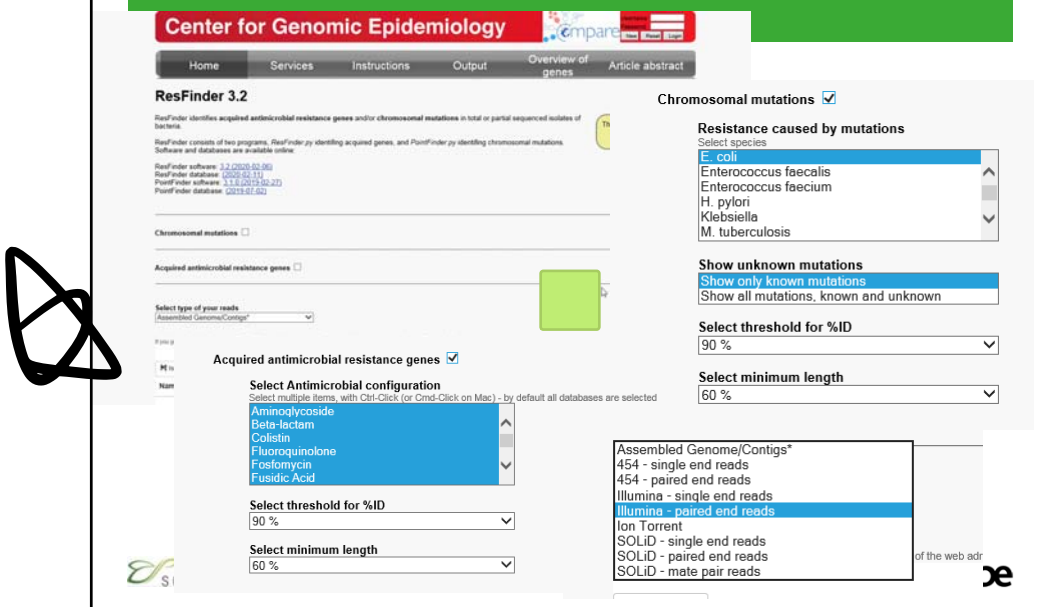
- Webserver
  - ResFinder
  - Comprehensive Antibiotic Resistance Database (CARD)
- Offline
  - ARG-ANNOT database
- Combination (to be installed locally), e.g. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor

AMR efforts e.g. at NCBI

- Database of sequenced isolates with standardized AMR metadata (i.e. accept antibiograms)
  http://www.ncbi.nlm.nih.gov/biosample/?term=antibiogram[filter]
- Stable, up-to-date database of AMR genes with standardized nomenclature
- Implement & validate tools for identifying AMR genes in new isolates

sciensano .be

Webserver - ResFinder



ResFinder results

ResFinder-2.1 Server - Results

Aminoglycoside

Resistance gene	%Identity	Query/HSP length	Contig	Position in contig	Predicted phenotype	Accession number
aadD	99.74	771 / 771	scaffold28_size1986	574..1344	Aminoglycoside resistance Alternate name; ant(4')-Ia and aadD2	AF181950

Beta-lactam

Resistance gene	%Identity	Query/HSP length	Contig	Position in contig	Predicted phenotype	Accession number
blaTEM-116	100.00	861 / 861	scaffold25_size2701	605..1465	Beta-lactam resistance	AY425988

Fluoroquinolone

No resistance genes found.

Fosfomycin

No resistance genes found.

Fusidic Acid

No resistance genes found.

ResFinder – database updates

Database Updates (Acquired antimicrobial resistance)

The ResFinder database download site

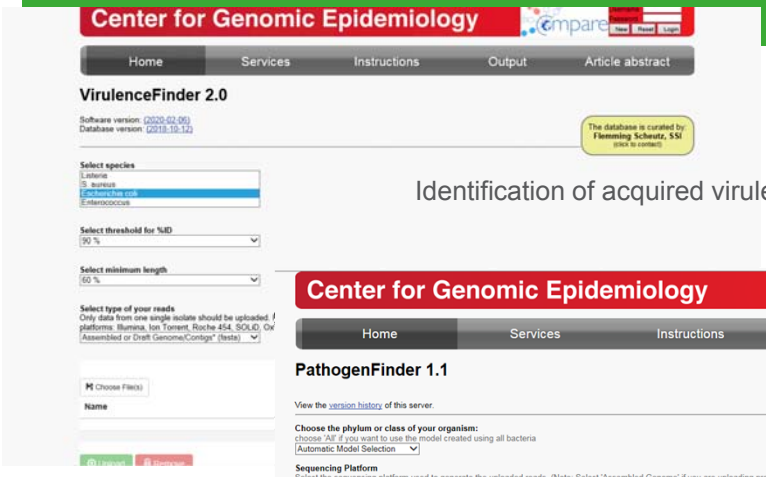
- **01-Oct-2019** Add tet(X3) and tet(X4) to tetracycline db
- **28-Apr-2019** Add mcr-9.1 to colistin db
- **22-Feb-2019** Updates of fusidic acid db
- **20-Feb-2019** Updates of blaCARB and blaFRI genes in beta-lactam db
- **28-Nov-2018** Updates and a few format changes in several db
- **10-Sep-2018** colistin db was updated to mirror the official mcr nomenclature (JAC doi:10.1093/jac/dky262)
- **10-Aug-2018** blaIMP-64 sequence was corrected in beta-lactam db
- **03-Aug-2018** Genes encoding NDM, VIM, KPC, IMP, TEM, SHV, CTX-M, CMY, ACC, DHA in beta-lactam db were revised and updated
- **04-Jul-2018** mec genes in beta-lactam db were revised and updated
- **04-Jul-2018** New genes added in quinolone, macrolide, oxazolidinone, phenicol and sulphonamide db
- **04-Jul-2018** Glycopeptide db was restructured
- **04-Jul-2018** All databases were formatted and corrected for duplicates
- **21-Mar-2018** Major updates in tetracycline db
- **21-Mar-2018** mcr-4.2_MG026621, mcr-5.2_ MG384740 and mcr-7.1_ MG267386 added to colistin db
- **19-Feb-2018** Major updates and corrections in quinolone and colistin database
- **30-Nov-2017** General updates and corrections
- **14-Sep-2017** Fosfomycin database was updated with 20 genes
- **25-Aug-2017** Colistin database was updated with the genes mcr-4 and mcr-5
- **25-Aug-2017** Beta-lactam database was updated with the gene blaOXA-427
- **25-Aug-2017** Beta-lactam database entries for blaSHV-5, blaSHV-12 and blaSHV-129 were corrected
- **25-Aug-2017** Sulphonamide database entries for sul3 were corrected
- **04-Jul-2017** Colistin database was updated with the gene mcr-3
- **03-Jul-2017** ResFinder database updated
- **17-Feb-2017** 62 new genes in beta-lactam, oxazolidinone and quinolone databases. Removed duplicates in phenicol and oxazolidinone data
- **02-Nov-2016** Beta-lactam database was updated with the gene blaBRO-1 beta-lactamase
- **02-Nov-2016** Beta-lactam database was updated with the gene blaBRO-2 beta-lactamase
- **02-Nov-2016** Beta-lactam database was updated with the gene blaBKC-1 beta-lactamase
- **01-Sep-2016** Colistin database was updated with the gene mcr-1.2, colistin
- **08-Jul-2016** Colistin database was updated with the gene mcr-2, colistin
- **21-Jun-2016** Default threshold (%ID) changed to 90%
- **02-Jun-2016** blaVCC-1 was added to beta-lactamase and 13 new variant of optrA to oxazolidinone and phenicol

Limitations of using WGS for predicting resistance

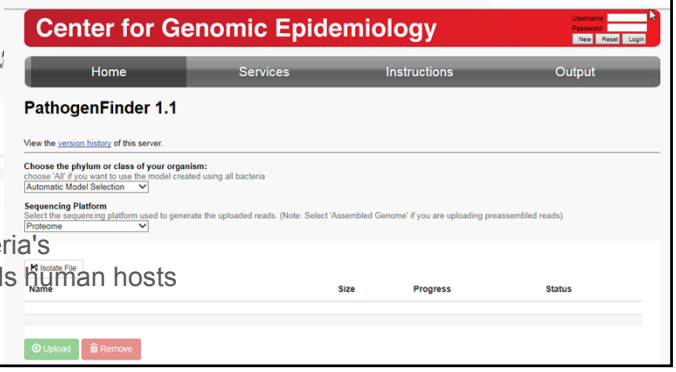
- **Technical**
  - Difficulty in closing plasmids
  - Exclusion of repeat regions with short reads
  - Missing IS that might interrupt regulatory loci
  - Flawed phenotyping
  - Incorrect breakpoint
- **Scientific (incomplete understanding of resistance)**
  - Chromosomal mutations that alter gene expression leading to:
    - Modified expression of OMPs
    - Modified expression of efflux pumps
    - Changes to the LPS
  - Possibility to predict susceptibility in the presence of an undiscovered determinant
  - Genetic instability (e.g. loss of plasmids)
  - Heteroresistance (e.g. Staph, MTB)
  - Resistance from loss of function in pro-drug activators (any nt possible)

sciensano .be

Virulence



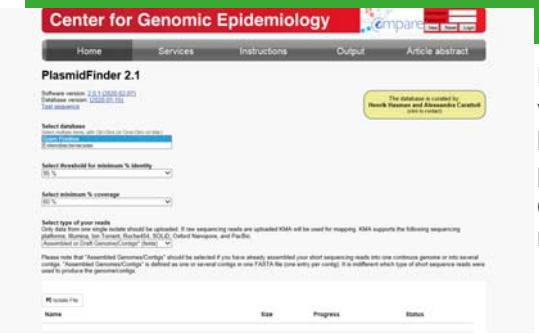
Identification of acquired virulence genes



Prediction of a bacteria's pathogenicity towards human hosts

sciensano

Plasmids



Detecting & classifying variants of known plasmids based on similarity with plasmids present in plasmid database – not able to find new plasmids

plasmidSPAdes: Assembling Plasmids from Whole Genome Sequencing Data

Extended abstract

Dmitry Antipov [1], Nolan Hartwick [2], Max Shen [3], Mikhail Raiko [2], Alla Lapidus [1] and Pavel A. Pevzner [1,2]

[1] Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia
[2] Department of Computer Science and Engineering, University of California, San Diego, USA
[3] Bioinformatics and Systems Biology Program, Massachusetts Institute of Technology, USA

Assembling plasmids from WGS

- http://spades.bioinf.spbau.ru/plasmidSPAd

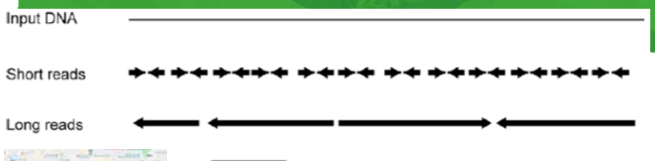
sciensano

.be

Long read sequencing

sciensano .be

Short read versus long read sequencing



Adapted from Peona et al 2018

Distance Brussels-Midi station – Sciensano (STEC genome)
5.500.000 bp

Autobus (long reads)
15,000 bp

Smartphone (short reads)
150 bp - 300 bp

Putting things in perspective

Taken from W. Timp

Building a genome is like putting together a puzzle

- Larger (blurry) pieces = easier puzzle
- Small pieces = hard to put together, can't figure out « blue sky »
- Small + large pieces = clearly and easier to put together

.be

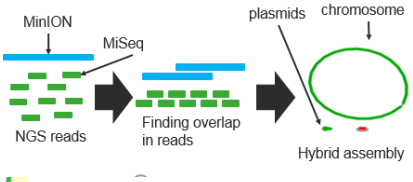
Hybrid assembly

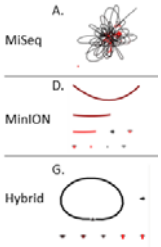
NGS strategy for plasmid reconstruction: combining short and long reads

For the reconstruction of plasmids with NGS the Illumina (MiSeq) and Nanopore (MinION) technologies will be used:

MiSeq:
- Short reads (25-250 bp)
+ High accuracy (99.9%)
+ Affordable
+ Standardized protocols

MinION:
+ Long reads (>8000 bp)
- Low accuracy (75-90%)
- Expensive (~1000 euro for 1 flowcell)
- Protocols constantly evolving and improving
- Stringent requirements for input DNA





PhD
Bas Berbers

Title

Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified *Bacillus*

Authors

Bas Berbers[1,2], Assia Saltykova[1,2], Cristina Garcia-Graells[3], Patrick Philipp[4], Fabrice Arella[4], Kathleen Marchal[2,5], Raf Winand[1], Kevin Vanneste[1], Nancy H. C. Roosens[1,6], Sigrid C.J. De Keersmaecker[1,*]

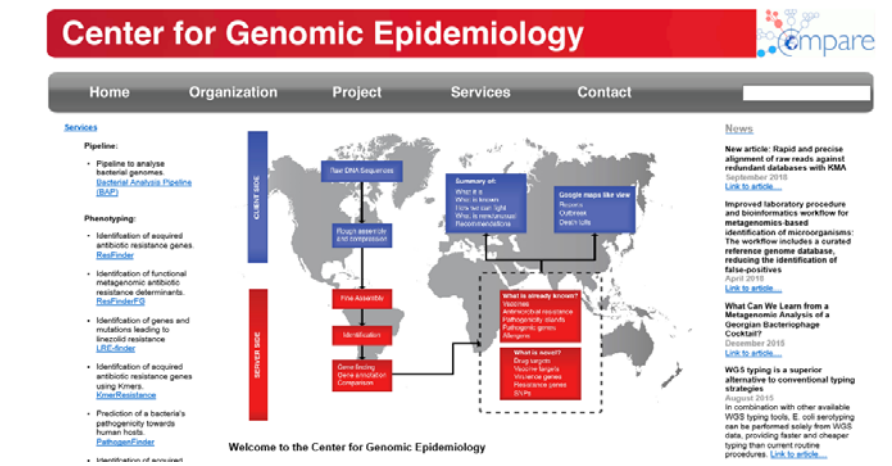
[1]Transversal activities in Applied Genomics, Sciensano, Brussels, Belgium; [2]Department of Information Technology, IDLab, Ghent University, IMEC, Ghent, Belgium; [3]Foodborne Pathogens, Sciensano, Brussels, Belgium, [4]Service Commun des Laboratoires, Illkirch-Graffenstaden, France, [5]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium.

(User-friendly) bioinformatics (1/3)

- Commercial packages
  - Bionumerics
  - CLC Genomics Workbench
  - SeqSphere
  - …

    - **Advantages**
      - Very powerful
      - Graphical user interface -> more user-friendly
    - **Disadvantages**
      - Licenses -> expensive
      - Black box
      - Limitations in applications
      - Might be complicated to learn how to use

sciensano .be



sciensano .be

(User-friendly) bioinformatics (2/3)

- Webservices
  - e.g. Resfinder
  - Integrated pipeline (BAP): https://cge.cbs.dtu.dk/services/CGEpipeline-1.1/
  - PubMLST

    - **Advantages**
      - Very powerful
      - Graphical user interface -> more user-friendly
      - Easy to use
    - **Disadvantages**
      - Black box
      - Limitations in applications
      - Security? Privacy?
      - Accessibility/Availability at all times?

sciensano .be

(User-friendly) bioinformatics (3/3)

- « in-house » pipelines/databases, to be installed/implemented locally
  - Open source tools (e.g. SRST) (less expensive)
  - Linux -> use Galaxy as interface to use Linux-based tools in a command-line free way, centralized resource (browser/network (nothing locally), multiple simultaneous users, scalable)
  - Push on the button pipelines, customized to pathogen
    - **Advantages in-house**
      - Both entire system and individual processes known and controlled (regulated)
      - Ability to provide a quick response (accessibility)
      - No need to send data to external servers (privacy, security)
      - In-house controlled and tracked databases and code (traceability)
      - No dependencies on external providers (availability)
      - Can be tailored to specific needs (flexibility)
    - **Disadvantages**
      - Requires more development & maintenance compared to commercial applications

sciensano .be

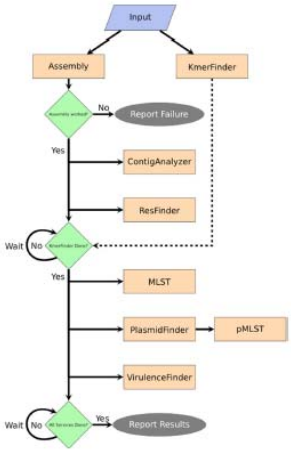
PLOS ONE

PLOS ONE | DOI:10.1371/journal.pone.0157718 June 21, 2016

A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance

Martin Christen Frølund Thomsen[1]*, Johanne Ahrenfeldt[1], Jose Luis Bellod Cisneros[1], Vanessa Jurtz[1], Mette Voldby Larsen[1], Henrik Hasman[2], Frank Møller Aarestrup[2], Ole Lund[1]

1 Department of Systems Biology, Technical University of Denmark, Kemitorvet Building 208, 2800 Kgs. Lyngby, Denmark, 2 National Food Institute, Technical University of Denmark, Søltofts Plads Building 221, 2800 Kgs. Lyngby, Denmark

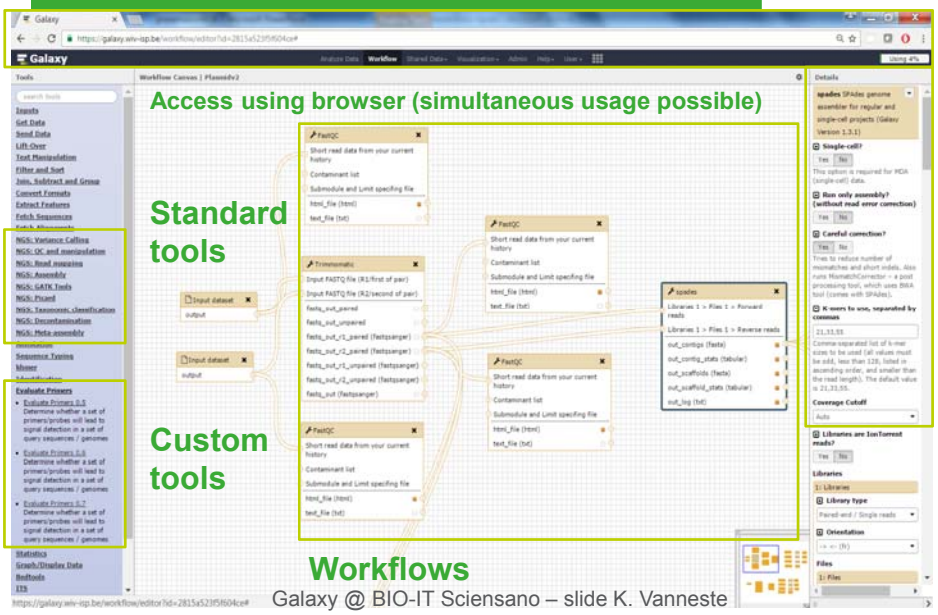


https://cge.cbs.dtu.dk/services/CGEpipeline-1.1/

sciensano .be

Combine tools to make pipelines

Tweak parameters

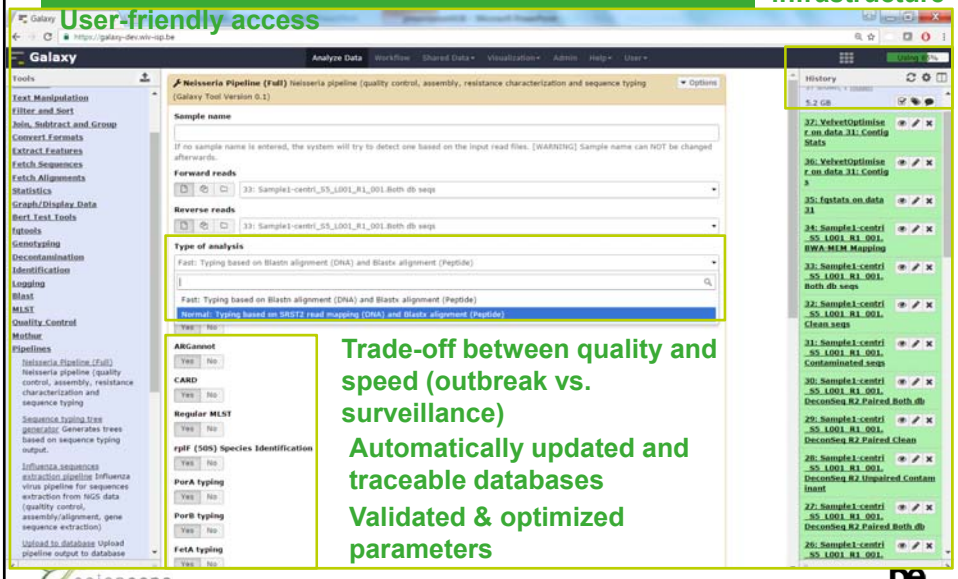


Galaxy @ BIO-IT Sciensano – slide K. Vanneste

Optimized 'push-on-button' pipelines

Centralized computational infrastructure

User-friendly access



Trade-off between quality and speed (outbreak vs. surveillance)

Automatically updated and traceable databases

Validated & optimized parameters

Galaxy @ BIO-IT Sciensano – slide K. Vanneste

Added value of WGS – routine NRL/NRC *Salmonella* Enteritidis outbreak

9/12/2015 Whole Genome Sequence Analysis of Salmonella Enteritidis PT4 Outbreaks from a National Reference Laboratory's Viewpoint – PLOS Currents O...

PLOS | CURRENTS OUTBREAKS

Whole Genome Sequence Analysis of Salmonella Enteritidis PT4 Outbreaks from a National Reference Laboratory's Viewpoint

September 11, 2015 · Research

Citation

Wuyts V, Denayer S, Roosens NH, Mattheus W, Bertrand S, Marchal K, Dierick K, De Keersmaecker SC. Whole Genome Sequence Analysis of Salmonella Enteritidis PT4 Outbreaks from a National Reference Laboratory's Viewpoint. PLOS Currents Outbreaks. 2015 Sep 11 . Edition 1. doi: 10.1371/currents.outbreaks.aa5372d90826e6cb0136ff66bb7a62fc.

Tweet

PhD Véronique Wuyts - SalMolType

sciensano .be

April – May 2014: 2 foodborne outbreaks

Same source or different sources for both outbreaks?

Social event: ± 220 guests
Food: catering service
Disease onset: April 23rd, 2014
No. cases: 45
No. hospitalised cases: 5

Social event: ± 300 guests
Food: bbq by volunteers
Disease onset: May 1st, 2014
No. cases: ± 40
Hospitalised cases: yes

Human cases linked to egg-containing food samples

Non-commercial eggs

.be

Microbiological investigation

Outbreak Flanders:

- NRL-FBO: 2 chocolate mousse samples tested positive for *Salmonella*
  → prepared with non-commercial eggs
- NRCSS: 11 human isolates received
  → 2 isolates randomly selected for further testing

Outbreak Wallonia:

- NRL-FBO: raw egg tested positive for *Salmonella*
  → non-commercial eggs used to prepare tiramisu
- NRCSS: 1 human isolate received

All outbreak isolates:

serotyped as *Salmonella enterica* subsp. *enterica* serovar Enteritidis

sciensano .be

Microbiological investigation

Outbreak	Isolate	Origin	Phage type	MLVA	Antimicrobial resistance
Flanders	S14FP01640	Chocolate mousse [a]	PT4	3-10-5-4-1	Colistin
Flanders	S14FP01642	Chocolate mousse [a]	PT4	3-10-5-4-1	Colistin
Flanders	S14BD01605	Human	PT4	3-10-5-4-1	Colistin
Flanders	S14BD01672	Human	PT4a	3-10-5-4-1	Colistin – ampicillin
Wallonia	S14FP01877	Raw egg [b]	PT4	3-10-5-4-1	Colistin
Wallonia	S14BD01753	Human	PT4	3-10-5-4-1	Colistin

[a] Prepared with non-commercial eggs from private laying hens
[b] Non-commercial egg from private laying hen, used to prepare tiramisu

sciensano .be

WGS analysis

WGS:

- Illumina HiSeq 2000
- 100 bp paired-end reads
- 40-plex

→ Raw FASTQ reads

WGS data analysis on a Windows 7 platform

- CLC Genomics Workbench:
  - quality trim of raw FASTQ reads
  - *de novo* assembly → contigs of chromosome and plasmids
  - read mapping on *S.* Enteritidis P125109 chromosome (NC_011294)
    + *de novo* assembly of unmapped reads → contigs of plasmids
- Tools on server of the Center for Genomic Epidemiology (DTU):
  MLST – ResFinder – PlasmidFinder
- BRIG (BLAST Ring Image Generator)

sciensano .be

Center for Genomic Epidemiology: MLST – ResFinder - PlasmidFinder

Outbreak	Isolate	Origin	MLST [a]	ResFinder	PlasmidFinder
Flanders	S14FP01460	Chocolate mousse	ST-11	-	pSLA5
Flanders	S14FP01642	Chocolate mousse	ST-11	-	pSLA5
Flanders	S14BD01605	Human	ST-11	-	pSLA5
Flanders	S14BD01672	Human	ST-11	*bla*$_{TEM-1B}$	pSLA5 – pSD107
Wallonia	S14FP01877	Raw egg	ST-11	-	pSLA5
Wallonia	S14BD01753	Human	ST-11	-	pSLA5

[a] *aroC-dnaN-hemD-hisD-purE-sucA-thrA*

1 analysis → similar results as traditional methods

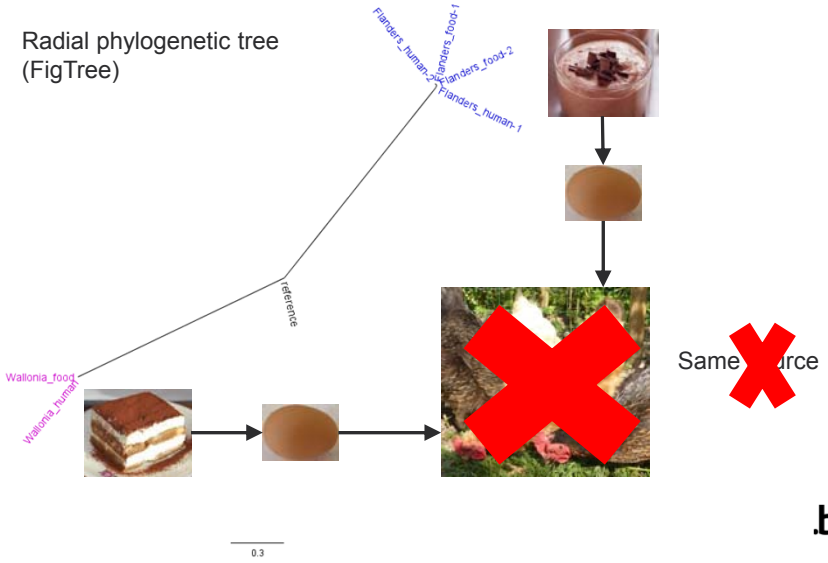
sciensano .be

Center for Genomic Epidemiology: CSI Phylogeny

		Flanders				Wallonia		Reference P125109
Flanders		0	0	1	1	52	52	46
		0	0	1	1	52	52	46
		1	1	0	2	53	53	47
		1	1	2	0	51	51	45
Wallonia		52	52	53	51	0	0	44
		52	52	53	51	0	0	44
Reference P125109		46	46	47	45	44	44	0

Center for Genomic Epidemiology: CSI Phylogeny

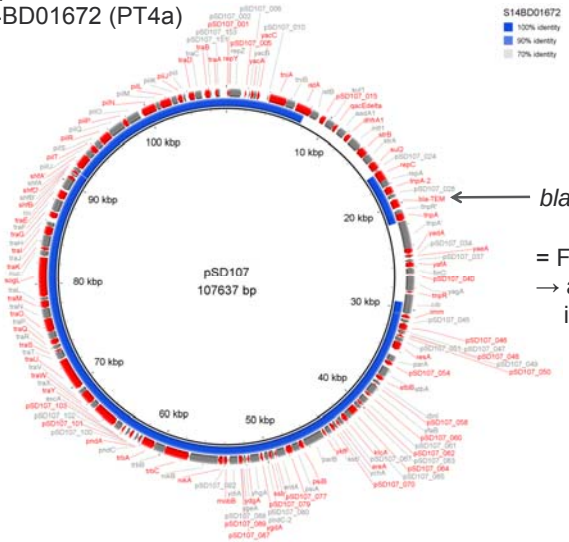
Radial phylogenetic tree (FigTree)



Same source

BRIG: plasmid pSD107

Contigs of plasmids of human isolate S14BD01672 (PT4a)



bla_{TEM}

= Fine-tuning
→ acquired during infection?

Added value WGS for outbreaks

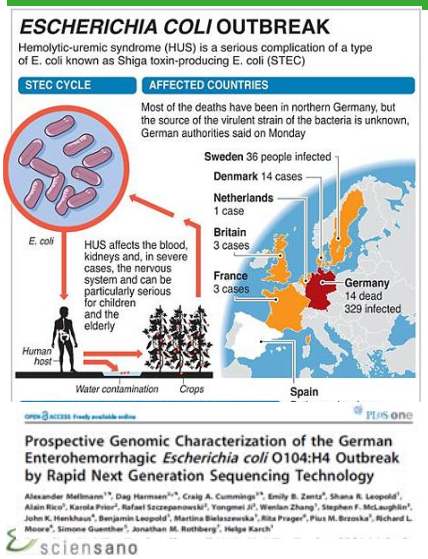
- WGS confirmed link between human isolates and isolates from food samples
  - = initial epidemiological and microbiological investigation
- WGS could additionally discriminate between each outbreak based on SNP analysis
  - > initial epidemiological and microbiological investigation
- Exploration of plasmids for fine-tuning of the SNP analysis results

Added value of WGS for outbreak investigation

WGS is feasible within the setting of an average NRL or NRC with user-friendly tools

Need for database of circulating strains

Added value of WGS
German 2011 *E. coli* outbreak



Traditional methods do not offer insight

↓

NGS allows detailed characterization of the genomic events that led to extreme virulence

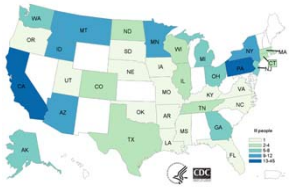
↓

Multiple antibiotic resistance, acid tolerance, enteroaggregative capacity, and shigatoxin production

.be

Example: human pathogenic Shiga toxin producing *Escherichia coli* (STEC)

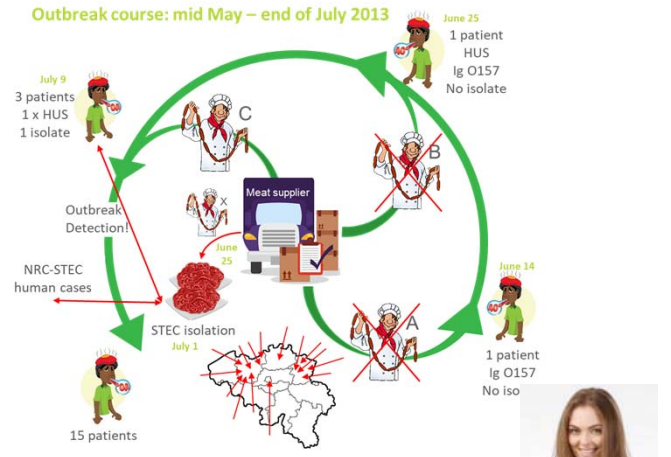
The romaine lettuce US multistate outbreak (E0157:H7), 2018

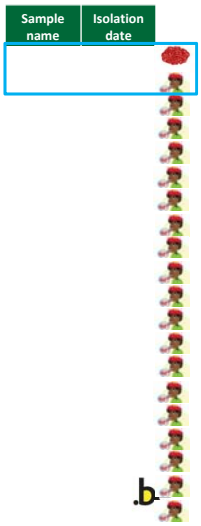


Limburg 2012 outbreak

Link human & food isolate to confine outbreak

STEC outbreak





PhD Stéphanie Nouws

Added value of WGS

More accurate and reliable phylogeny

- Accelerates cluster completion
- Database of background isolates
  → Importance of good surveillance
  → Not convenient with PFGE

Generation of large amount of information about pathogen

- Increased insight → Stimulation further research

WGS – faster workflow

- Limited effect
  - ✓ Real-time data analysis
  - ✓ Metagenomics: no isolation
- Time until outbreak detection
  - ✓ Stimulate rapid report by clinicians
  - ✓ Database with background strains

Added value of WGS – routine NRL/NRC
Neisseria meningitidis

Méningite sur le campus de Dijon, une campagne de vaccination à la rentrée

Après la mort de deux étudiantes, victimes d'une méningite à l'université de Dijon, une campagne de vaccination devrait être lancée dès le 3 janvier.

LE MONDE | 27.12.2016 à 11h45 • Mis à jour le 27.12.2016 à 12h03

- 2 cases in France (deaths)
- Student of Univerisity of Dijon died in Saint Pierre hospital in Brussels because of meningitis (20/12/2016)
- Isolate received @ NRC Sciensano (S. Bertrand) (23/12/2016)
- Classical testing: invasive strain W (or W 135) cfr. cases in France (23/12/2016)
- WGS started 03/01/2017 @ TAG
- WGS data available 9/01/2017
- Bioinformatic pipeline '*Neisseria*': same cluster W:P1-5,2:F1-1:cc11 (09/01/2017)
- Phylogenetic analysis by French authorities: confirmation isolate belongs to same clone UK/South America, as the other isolates in Dijon (10/01/2017)

sciensano .be

WGS data storage requirements

Pathogen	Genome size (Mb)	Average coverage	Raw read FASTQ file size, uncompressed (MB)	Fully assembled genome FASTA file size, uncompressed (MB)
Salmonella spp.	5.1	50	510	5.1
Listeria monocytogenes	2.9	50	290	2.9
E. coli	4.6–5.4	50	460–540	4.6–5.4
Campylobacter spp.	1.6	50	160	1.6

RawReadsFileSize (MB, uncompressed) = GenomSize (Mb) x Coverage x 2
Size file fully assembled genome (MB) = size of genome in Mb

- Zipping of FASTQ files (2-3x reduction)
- Need to keep raw reads for accreditation purposes, litigation, reanalysis
- Back-up of data
- Sharing of data – bandwidth (secure FTP sites, file sharing services)

ECDC – Expert opinion on the introduction of next-generation typing methods for food- and waterborne diseases in the EU and EEA, 2015

sciensano .be

Global data sharing

Benefits

- Improves food safety sciences
- Efficient use of resources (prevent duplication)
- Enables trend analysis of pathogen evolution & rapid response
- Mitigates health, social, economic impacts

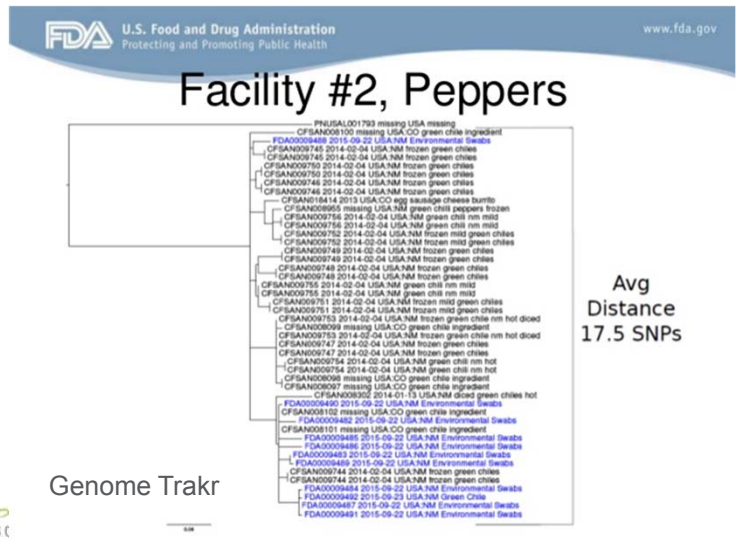
Challenges

- ? Inequities in capacity & resources
- ? Data ownership & metadata access
- ? Accountability & transparency
- ? Validation, standards & quality (ISO working group ongoing)
- ? Fear of being 'scooped'
- ? Technical trade barriers
- ? Privacy law

FAO & WHO Technical background paper 2016: Applications of whole genome sequencing in food safety management

sciensano .be

Added value open data sharing



.be

WGS resource considerations

« WGS is too advanced, too difficult and too expensive to implement »

PERCEPTION

« WGS is the perfect solution to all our problems »

sciensano .be

WGS – benefits & drawbacks

Performance (specificity/sensitivity)

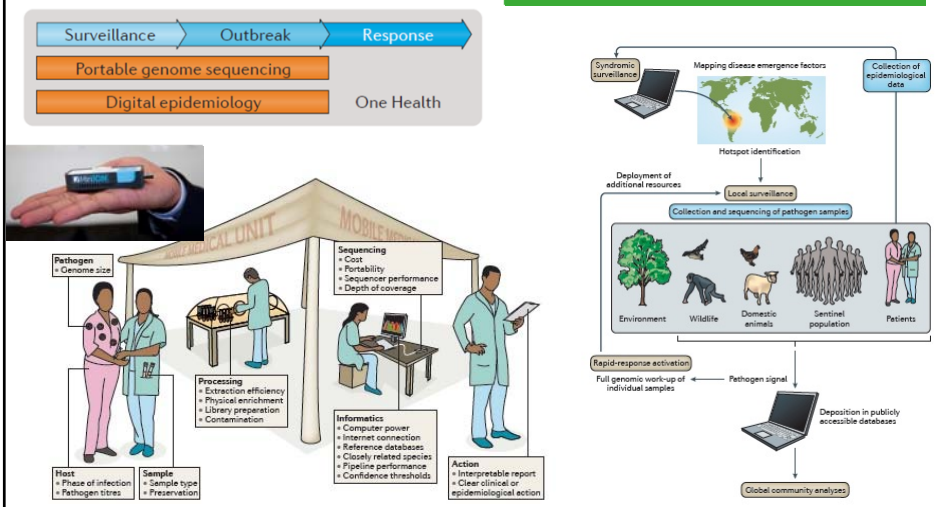
+ Confidence in clusters and links isolates
+ Variety of outputs based on ONE « simple » laboratory procedure
+ Flexibility/universality – not pathogen specific
+ Speed, amenable to reanalysis
+ Ease of sharing (basic common language, data repositories)
- No phenotype
- Education & experience needed for analysis (bioinformatics) and interpretation
- Willingness to (global) data sharing
- Standardization needed

Cost

+ Cost effective alternative to classical typing if replacing several classical methods (not complementary) & adequate volumes of isolates; economic cost
+ Flexibility – same person can analyse different pathogens
- Expensive to establish (equipment, education, computational) – but can be shared across institutes

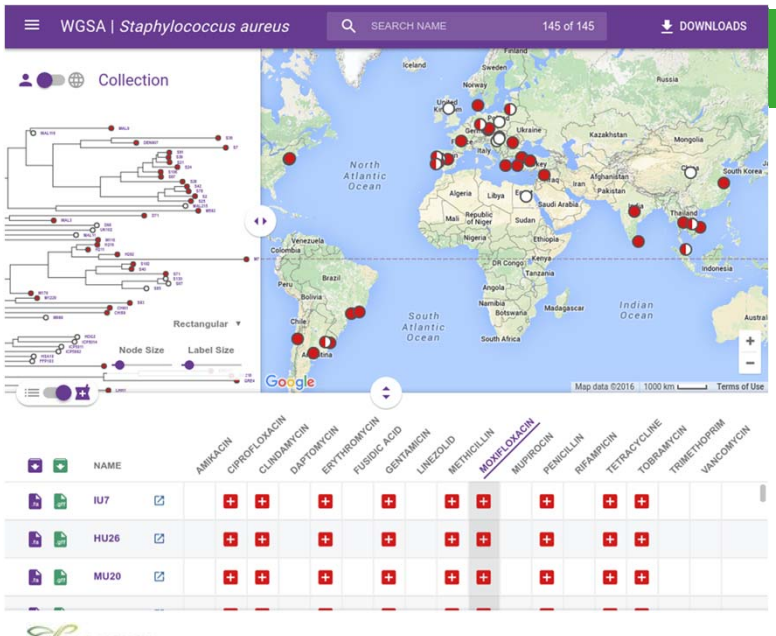
sciensano .be

Future?: Towards a genomics-informed, real-time, global pathogen surveillance system



sciensano .be

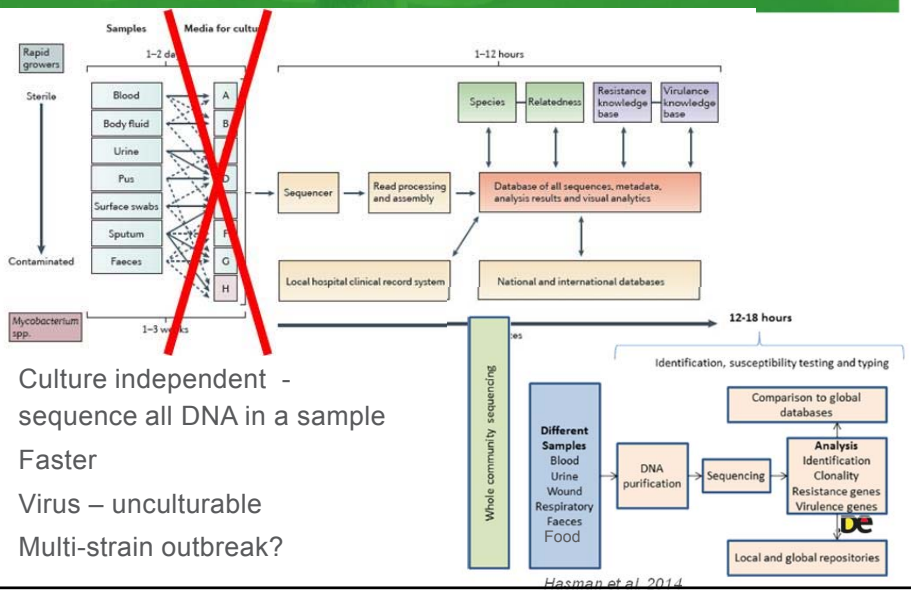
Gardy & Loman Nat Rev Genetics 2017



Wgsa.net (microreact)

sciensano .be

Isolates (WGS) vs. metagenomics



Hasman et al. 2014

Culture independent - sequence all DNA in a sample

Faster

Virus – unculturable

Multi-strain outbreak?

Isolates (WGS) vs. Metagenomics (2)

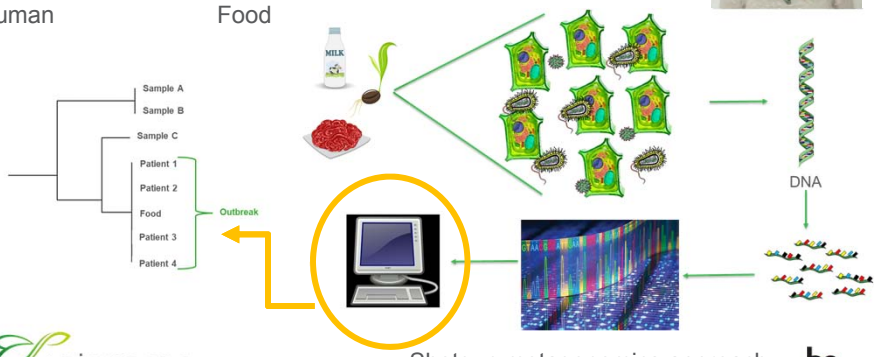
- **16S rRNA PCR & sequencing**
  - Taxonomically and phylogenetically informative marker
  - Only to the genus/species level
  - Bias PCR (primers not universal), artificial amplicons after assembly
  - Low sensitivity if insufficient pathogen DNA
  - Affected by presence of contaminating DNA from other bacterial species
  - No functional/biological information
- **Shotgun metagenomics**
  - Taxonomical & biological information
  - Filtering out unwanted DNA in post-sequencing analysis
  - Higher sensitivity as organisms can be identified from different segments of DNA
- **Considerations**
  - Dead & alive organisms
  - Massive sequence data linked to e.g. food processing facility (pathogen?)
  - Human reads? Privacy law

sciensano .be

Future: metagenomics?
Towards isolation-free outbreak investigation

Metagenomics: no isolation…

PhD Florence Buytaers

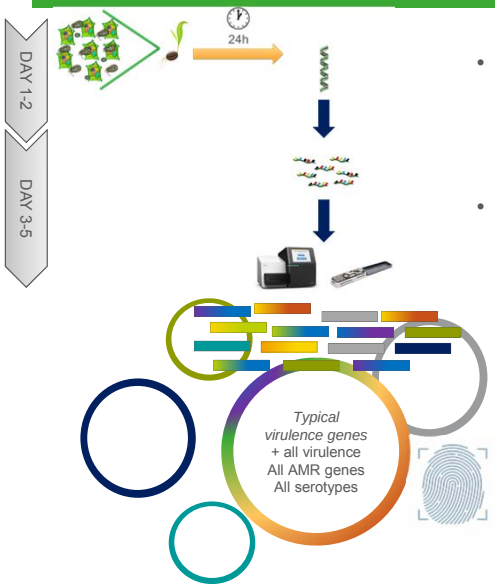


sciensano

Shotgun metagenomics approach

.be

Metagenomics-based outbreak investigation



- **Information**:
  - All species present
  - No isolation needed -> shorter analysis time
- **Complexity**:
  - all DNA present in sample sequenced
  - (Short) reads: big puzzle: what belongs to which genome?
  - Reconstruct genomes present
  - Phylogeny of reconstructed genomes (SNP-level): link human & food isolate

.be

Metagenomics & food safety

AEM

Application of Metagenomic Sequencing to Food Safety: Detection of Shiga Toxin-Producing *Escherichia coli* on Fresh Bagged Spinach

Susan R. Leonard, Mark K. Mammel, David W. Lacher, Christopher A. Elkins
Division of Molecular Biology, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, Laurel, Maryland, USA

December 2015 Volume 81 Number 23 Applied and Environmental Microbiology aem.asm.org 8183

TABLE 1 Results of spiking spinach with STEC Sakai[a]

Sampling time and spike (CFU)	*E. coli* phylogroup E (%)[b]	Mapped reads (%)[c]	Coverage (fold)[d]
Preenrichment			
10	0.003	0.01	0.04
1,000	0.003	0.05	0.11
100,000	0.119	0.03	0.10
1,000,000	0.356	0.13	0.41
10,000,000[e]	4.251	0.84	3.16
5-h enrichment			
10[f]	0.008	0.08	0.24
1,000[f]	0.066	0.12	0.39
10,000[e]	7.418	3.30	13.3
100,000	30.449	17.09	70.9
23-h enrichment			
10[f]	66.954	49.73	184
1,000[f]	84.330	61.77	252

[a] Samples of 100 g spinach were spiked with STEC Sakai.
[b] Percentage of the bacterial population belonging to *E. coli* phylogroup E.
[c] Percentage of total reads mapping to the STEC Sakai genome.
[d] Average chromosome coverage. The values were normalized to 10,000,000 total reads per sample. The coverage for the two plasmids was greater than that of the chromosome for all samples.
[e] Average of triplicate biological samples.
[f] Average of duplicate biological samples.

TABLE 2 Number of reads mapping to virulence genes and serotyping loci for spinach samples spiked with STEC Sakai

Sampling time and spike (CFU)	No. of reads mapping to the following virulence gene or serotyping locus[a]: *stx*$_{1a}$	*stx*$_{2a}$	*γ-eae*	*ehxA*	O157 *wzx*	O157 *wzy*	H7 *fliC*
Preenrichment							
10	0	0	0	0	0	0	0
1,000	0	0	0	0	0	0	0
100,000	1	1	1	1	0	0	0
1,000,000	2	0	7	4	3	0	2
10,000,000[b]	21	17	79	44	15	12	12
5-h enrichment							
10	0	0	0	0	0	1	0
1,000	0	2	4	5	0	0	0
10,000[b]	57	65	329	231	36	22	43
100,000	373	275	1,676	1,382	194	126	232
23-h enrichment							
10[c]	1,225	1,361	3,467	4,013	988	850	978
1,000	1,715	1,647	4,773	4,529	1,419	1,110	1,298

[a] Number of reads mapping to virulence genes and serotyping loci for 100-g spinach samples spiked with STEC Sakai. For comparison, the numbers are normalized to 10,000,000 total reads per sample.
[b] Average of triplicate biological samples.
[c] Average of duplicate biological samples.

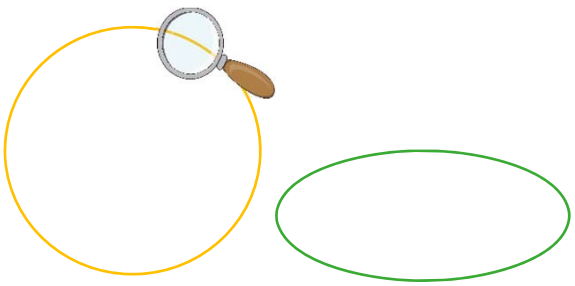
Shotgun metagenomics

LOD: 10 CFU/100g if enriched 8h

Virulence genes & assembled genome

.be

Strain-level metagenomic analysis of bacterial pathogens in enriched food samples



Outbreak

Future work: regulation to be changed – still isolate needed…

sciensano

PhD Assia Saltykova

PhD Florence Buytaers

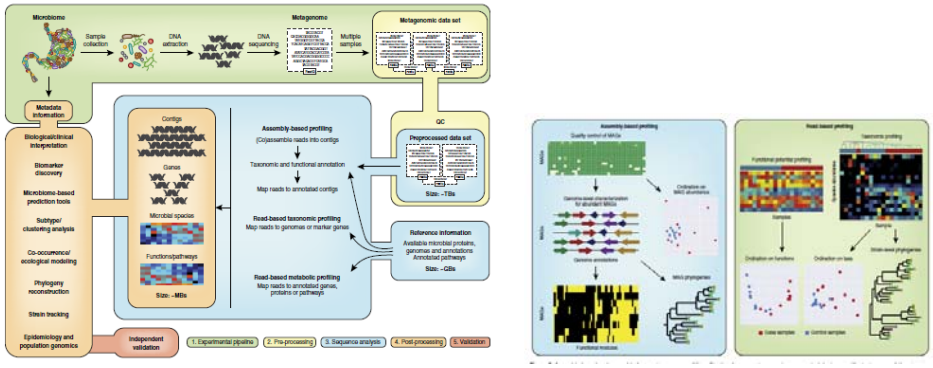
.be

nature biotechnology

REVIEW

Shotgun metagenomics, from sampling to analysis

Christopher Quince[1,7], Alan W Walker[2,7], Jared T Simpson[3,4], Nicholas J Loman[5] & Nicola Segata[6]



WGS will impact foodsafety worldwide

WGS will become standard methodology in food safety

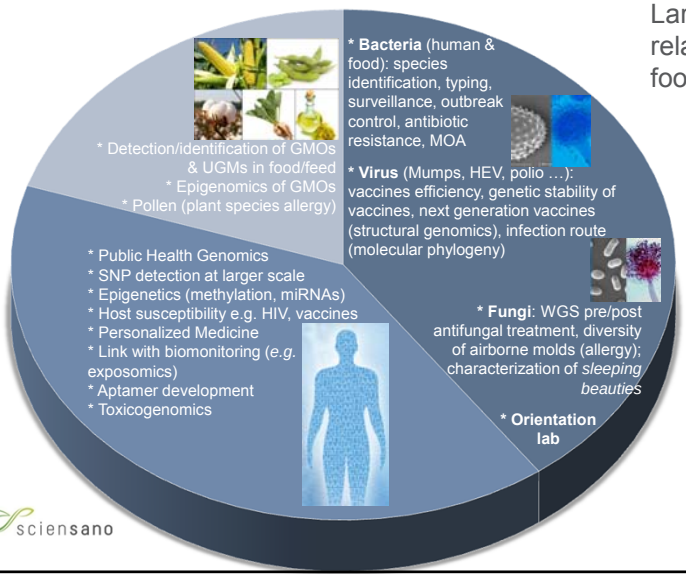
- Foodborne disease surveillance
- Food inspection (testing) and monitoring
- Outbreak detection and investigation (higher resolution, more confident definition of clusters, linking sporadic cases…)
- Food source attribution studies
- Food technology developments

- + human & animal disease surveillance & outbreak response
- = thé One Health technology
- Several European countries already WGS implemented/ing in routine, supported by the competent authorities – Belgium?

sciensano .be

NGS/WGS @ Sciensano

Needs identified in all 4 DSs of Sciensano



Large part related to food safety

"*nice*" to use NGS

Research

* To address specific biological questions

"*need*" to use NGS

Routine analysis

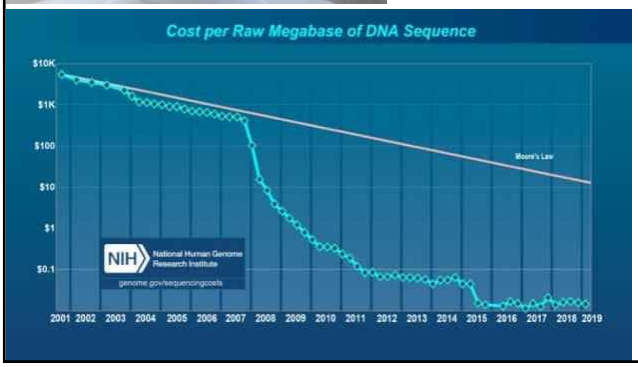
* Demand of stakeholders (ECDC, INAMI…)

International branding & project funding

sciensano .be

The Next-Generation Sequencing revolution

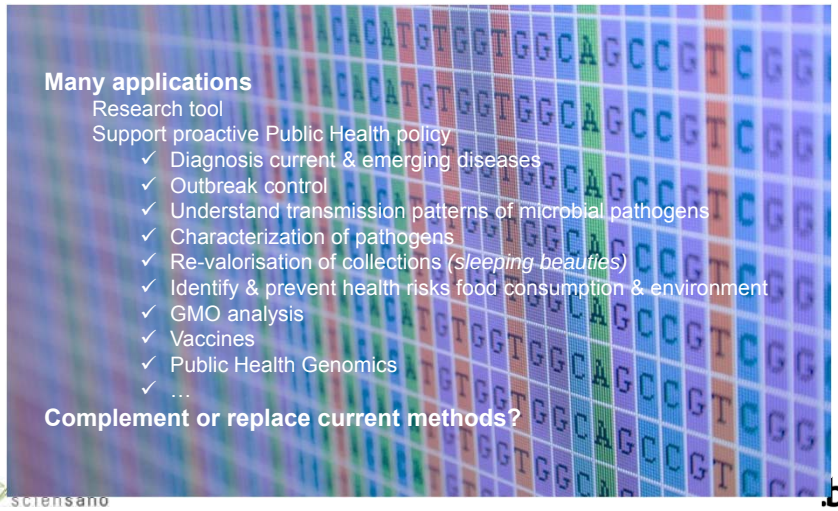
NGS
Massive parallel sequencing



2016
Bacterial genome < $100
Human genome < $1000

.be

NGS: Next-generation revolution



Many applications

Research tool

Support proactive Public Health policy

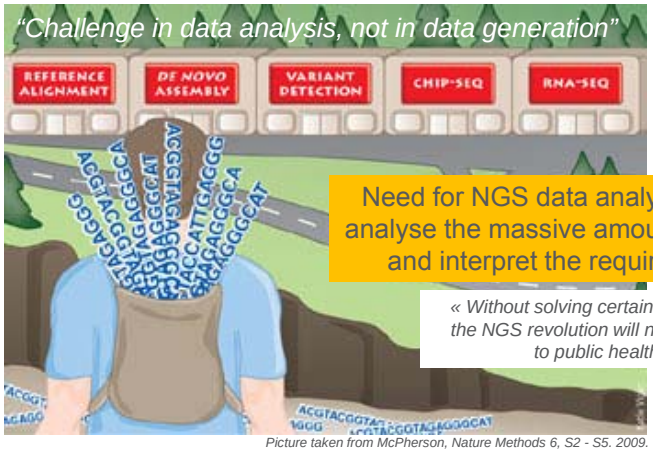
- ✓ Diagnosis current & emerging diseases
- ✓ Outbreak control
- ✓ Understand transmission patterns of microbial pathogens
- ✓ Characterization of pathogens
- ✓ Re-valorisation of collections (*sleeping beauties*)
- ✓ Identify & prevent health risks food consumption & environment
- ✓ GMO analysis
- ✓ Vaccines
- ✓ Public Health Genomics
- ✓ …

Complement or replace current methods?

sciensano .be

NGS: Next-generation gap

"Challenge in data analysis, not in data generation"



Picture taken from McPherson, Nature Methods 6, S2 - S5. 2009.

Need for NGS data analysis tools to rapidly analyse the massive amount of data to extract and interpret the required info correctly

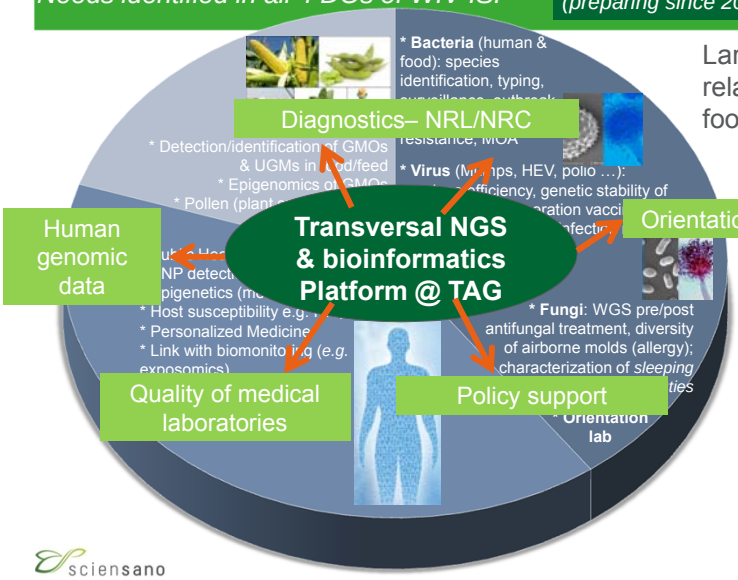
« Without solving certain bioinformatics challenges, the NGS revolution will not be extensively available to public health professionals »

sciensano .be

NGS/WGS @ Sciensano

Needs identified in all 4 DOs of WIV-ISP

'Official' start 'NGS & bioinformatics platform': 2015 *(preparing since 2011)*



Large part related to food safety

In-house MiSEQ (outbreak, accreditation envisaged)

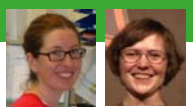
Bioinformatics team (~5 people)

sciensano .be

MiSeq instrument for routine

Dr Ir senior MoI biologist (lead)

NGS expert technician



+ 2 other technicians trained for NGS

- NGS-instrument: in-house
  - Time– crisis (<-> outsourcing: 6-8 weeks)
  - Technology: ~ stable
  - Strategic importance: central points for BE stakeholders
- Benchtop model: Illumina MiSeq
  - Short analysis time – acceptable amount of data
  - User-friendly
  - 'gold standard' for microbial NGS + cancer panels
- ISO17025 (ISO15189) accreditation (member of ISO working group (ISO/TC 34/SC 9/WG25): Whole genome sequencing for genomic typing and characterization)

Other technologies: R&D

- MinION
- PacBio (outsourcing)
- HiSeq (outsourcing)

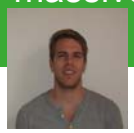
.be

Bioinformatics Platform for massive data analysis


K. Vanneste
Bioinformatician
(lead)


Bioinformatician


Bioinformatician


Bioinformatician


Bioinformatician

Software engineer

NGS data analysis

Development and implementation of user-friendly bioinformatics pipelines & databases

Commercial solutions

In-house developed solutions

Microbial isolates
Mixed samples
Human samples

Computational high-performance infrastructure

- Scientific storage
- In-house cluster
- Several (web)servers
- Highly controlled (versioning, logging of parameters/pipelines, stress tests…)

sciensano .be

R&D versus routine applications in a public health setting, many challenges

- Sample preparation

sciensano .be

More emphasis on pre-sequencing steps

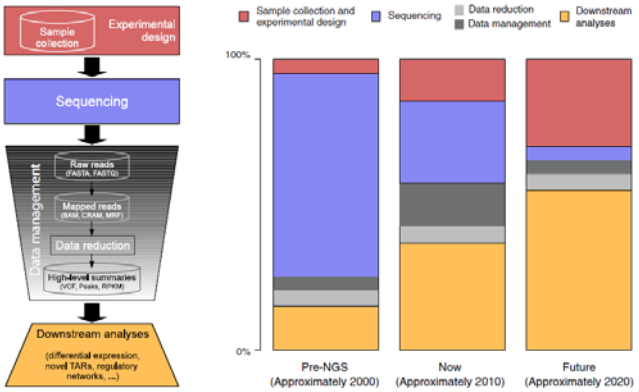


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

Study design (e.g. which platform to use?)

Sample preparation

Sboner A., Mu X. J. *et al.* 2011. The real cost of sequencing: higher than you think! **Genome Biology** 12:125.

.be

Example AmpliVi.x project: RNA viruses in clinical samples

Level of sample preparation

- Viral diagnostics, discovery and molecular epidemiology
- Pathogen-specific PCR:
  - Narrow range of detection
- Multiplex assays
  - Design extremely laborious
  - To be continuously updated with new targets according to the description of new species and/or variants
  - No novel virus detection
- NGS
  - No need to design and update specific primers to amplify target sequences for each specific virus
  - Cultivation = bottleneck
  - W/O cultivation: enrichment
    - With specific primers -> cfr. pathogen-specific PCR
    - Metagenomics approach -> problem of low abundance virus, human/bacterial contaminants
  - **-> physical enrichment of virus before library prep**

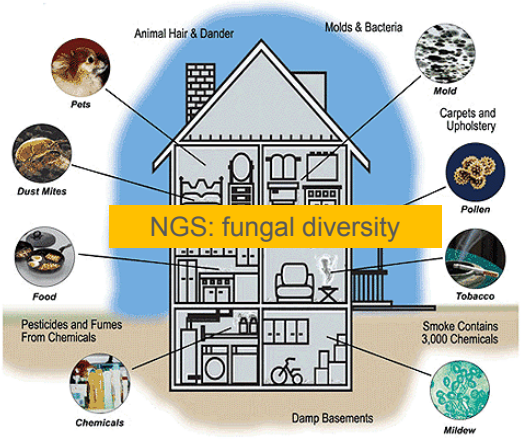
sciensano

Research A. Marichal

Example MycoMOLAIR project: Indoor air quality

PhD Xavier Libert

Level of sample & library preparation



Indoor time: 75-90%

Mold exposure: 18-50%

-> Severe public health problems (e.g. asthma) – causal relationship?

-> consequences

Moisture – energy saving houses?

Accurate identification needed

Problems with classical (culture & microscopy based) analysis

Dead/uncultivable fraction: allergenic?-> fungal diversity?

Need for molecular methods

sciensano .be

Metagenomics approach – ITS

Sampling in offices
E. jeanselmei
detection in indoor air

Sampling of residences
Fungal diversity

300 L/min during 5 minutes
Collection liquid :
Water + Tween 20 (0.01 %)

Classical analysis
Culture and microscopic identification

gDNA extraction
ITS1*, ITS2* amplicon sequencing

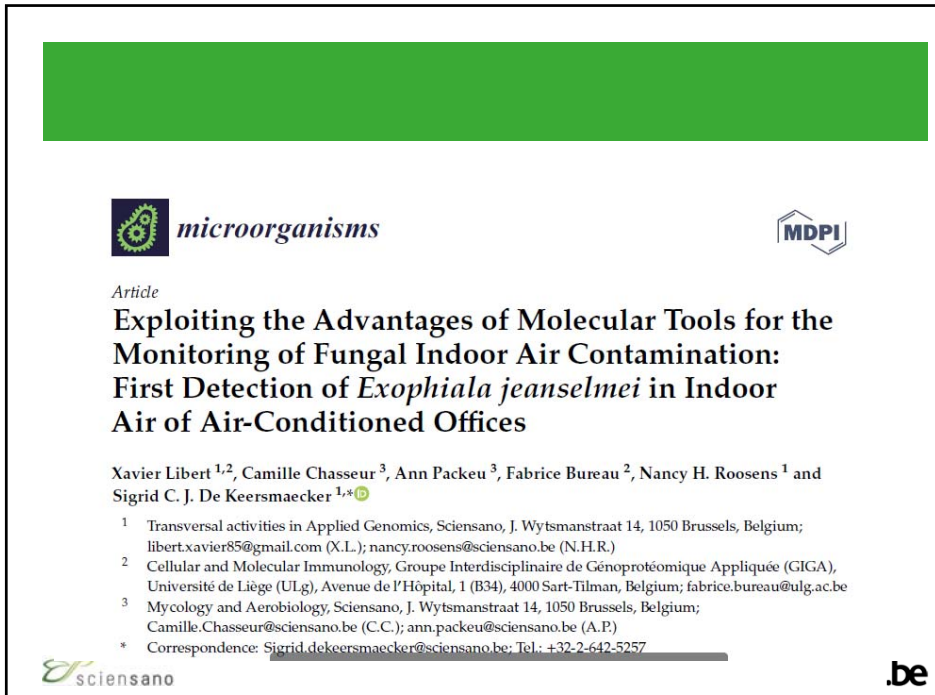
ITSoneDB

unite

NGS data analysis – ITS databases

sciensano .be

**ITS1 is located between 18S and 5.8S rRNA genes, while ITS2 is between 5.8S and 28S rRNA genes*



microorganisms

MDPI

Article

Exploiting the Advantages of Molecular Tools for the Monitoring of Fungal Indoor Air Contamination: First Detection of *Exophiala jeanselmei* in Indoor Air of Air-Conditioned Offices

Xavier Libert [1,2], Camille Chasseur [3], Ann Packeu [3], Fabrice Bureau [2], Nancy H. Roosens [1] and Sigrid C. J. De Keersmaecker [1,*]

1 Transversal activities in Applied Genomics, Sciensano, J. Wytsmanstraat 14, 1050 Brussels, Belgium; libert.xavier85@gmail.com (X.L.); nancy.roosens@sciensano.be (N.H.R.)
2 Cellular and Molecular Immunology, Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA), Université de Liège (ULg), Avenue de l'Hôpital, 1 (B34), 4000 Sart-Tilman, Belgium; fabrice.bureau@ulg.ac.be
3 Mycology and Aerobiology, Sciensano, J. Wytsmanstraat 14, 1050 Brussels, Belgium; Camille.Chasseur@sciensano.be (C.C.); ann.packeu@sciensano.be (A.P.)
* Correspondence: Sigrid.dekeersmaecker@sciensano.be; Tel.: +32-2-642-5257

sciensano .be

R&D versus routine applications in a public health setting, many challenges

- Sample preparation
- Standardized, harmonized generation of NGS data for different applications

- Computational requirements
- Validation of standardized & optimized pipelines, for different applications in public health (isolates, clinical samples, metagenomics…)
- User-friendly access for non-experts (application in NRL/NRC)
- Trade-off between quality and speed of analysis (crisis…)
- Traceability (databases, runs)
- Integration with other high-throughput technologies

sciensano .be

Accreditation: ISO 17025

- General requirements for the competence of testing and calibration laboratories

General considerations

• management system ensuring adequate qualification and training of staff;
• metrology system ensuring periodical calibration of equipment;
• method verification demonstrating that laboratory is meeting the performance characteristics

Goal	Tools
High quality service to the customer	• Management review • Documents (protocols and traceability) • Quality: proficiency tests • Internal & external audits • NC (internal and external complaints) • Maintenance of instruments • Staff training

Data generation & data analysis!
ISO working group

Accreditation ISO15189: 2012 specifies requirements for quality and competence in medical laboratories (Additional requirements regarding ethics, contingency plan, risk assessment, referral laboratories, organization structure (medical staff),LIS – reporting)

.be

Bioinformatics & ISO

- No ISO standard yet
- Best practices:
  - Version control (codebase + databases) (GIT)
  - Code review
  - DTAP approach (Development, Testing, Acceptance, Production)
  - Technical documentation (Wiki)
  - User documentation (Wiki + SOP)
  - Traceability: automatically updated databases + logging of all parameters (including version of databases used) and runs
    - All relevant information to repeat an analysis with a pipeline is stored:
      - Input files (checksums), Parameters, Output files (checksums), Database versions, Username, Date, …
  - Extensive validation of pipelines
    - Repeatability
    - Reproducibility
    - Part of complete workflow within pathogen profiling

More information: Lambert et al. 2017. Journal of AOAC International Vol. 100, No.3 DOI: 10.5740/jaoacint.16-0269

Validation of bioinformatics pipelines

LAMBERT ET AL.: JOURNAL OF AOAC INTERNATIONAL VOL. 100, NO. 3, 2017 721

FOOD BIOLOGICAL CONTAMINANTS

Baseline Practices for the Application of Genomic Data Supporting Regulatory Food Safety

DOMINIC LAMBERT
Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Ottawa, ON, Canada
ARTHUR PIGHETTI
U.S. Food and Drug Administration, Center for Food Safety and Applied Nutrition, College Park, MD
EMMA GRIFFITHS
Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC, Canada
GARY VAN DOMSELAAR
Public Health Agency of Canada, National Microbiology Laboratory, Winnipeg, MB, Canada
PETER EVANS
U.S. Department of Agriculture, Food Safety and Inspection Service, Washington, DC
SHARON BERTHELET
Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Ottawa, ON, Canada
DUNCAN CRAIG
Food Standards Australia New Zealand, Canberra, Australia
P. SCOTT CHANDRY
Commonwealth Scientific and Industrial Research Organisation, Melbourne, Australia
ROBERT STONES
Food and Environment Research Agency, Sand Hutton, York, United Kingdom
FIONA BRINKMAN
Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC, Canada
ALEXANDRE ANGERS-LOUSTAU and JOACHIM KRUYSSA
European Commission, Joint Research Center, Ispra, Italy
WEIDA TONG
U.S. Food and Drug Administration, National Center for Toxicological Research, Little Rock, AR
BURTON BLAIS[1]
Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Bldg 22, 960 Carling Ave, Central Experimental Farm, Ottawa, ON, Canada K1A 0Y9

sciensano .be

Neisseria case study

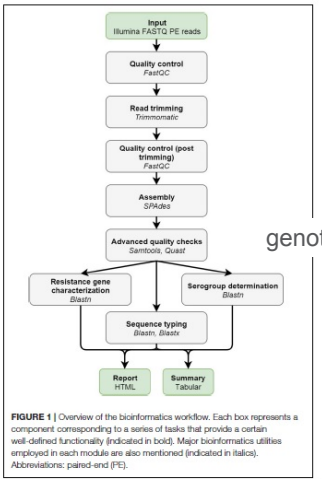


FIGURE 1 | Overview of the bioinformatics workflow. Each box represents a component corresponding to a series of tasks that provide a certain well-defined functionality (indicated in bold). Major bioinformatics utilities employed in each module are also mentioned (indicated in italics). Abbreviations: paired-end (PE).

genotyping

frontiers in Microbiology

PhD Bert Bogaerts

ORIGINAL RESEARCH
published: 06 March 2019
doi: 10.3389/fmicb.2019.00362

Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: *Neisseria meningitidis* as a Proof-of-Concept

*Bert Bogaerts[1], Raf Winand[1], Qiang Fu[1], Julien Van Braekel[1], Pieter-Jan Ceyssens[2], Wesley Matthouss[2], Sophie Bertrand[2], Sigrid C. J. De Keersmaecker[1], Nancy H. C. Roosens[1] and Kevin Vanneste[1]**

[1] Transversal Activities in Applied Genomics, Sciensano, Brussels, Belgium, [2] Bacterial Diseases, Sciensano, Brussels, Belgium

OPEN ACCESS

Edited by:
Rustam Aminov,
University of Aberdeen,
United Kingdom

Reviewed by:
Maria de Toro,

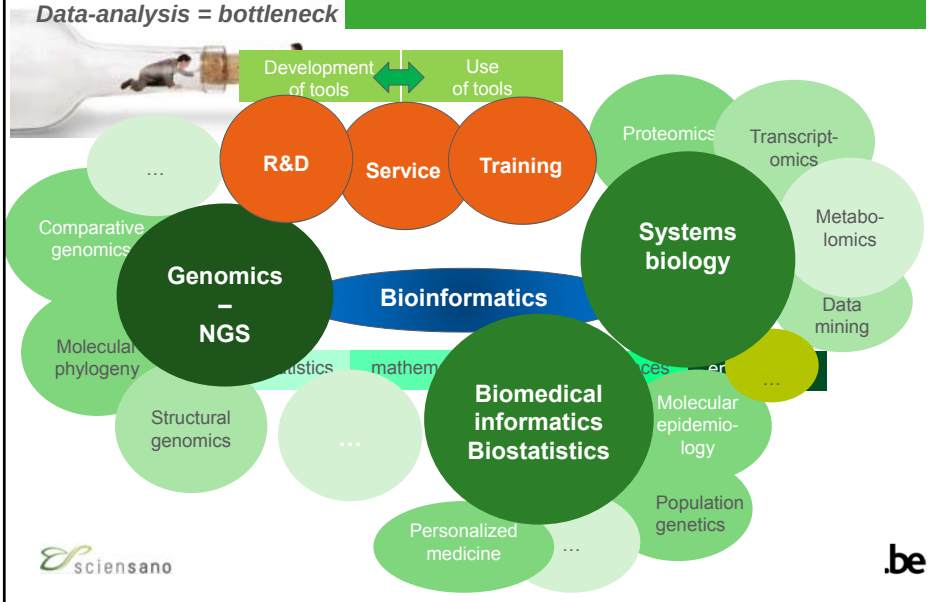
sciensano

Push-button pipeline = user-friendly, for NRL/NRC Report

.be

Need for bioinformatics @ Sciensano

Data-analysis = bottleneck



Acknowledgements

- Sciensano
  - Transversal activities in Applied Genomics: Nancy Roosens, Kevin Vanneste, Véronique Wuyts, Assia Saltykova, Stefan Hoffman, Maud Delvoye, Els Vandermassen, Dirk Van Geel, Loïc Lefèvre, Bas Berbers, Stéphanie Nouws, Florence Buytaers, Qiang Fu, Bert Bogaerts, Raf Winand, Thomas Delcourt, Julien Van Braekel, Axel Marichal, Xavier Libert
  - Food Pathogens: Katelijne Dierick, Nadine Botteldoorn, Sarah Denayer, Bavo Verhaegen, Cristina Garcia-Graells
  - Bacterial Diseases: Sophie Bertrand, Wesley Mattheus, Vanessa Mathys, Pieter-Jan Ceyssens
  - QSE: Anne-Marie Vanherle, Patricia Cliquet
- Ughent: Kathleen Marchal
- ISO working group TC 34/ SC 9/ WG 25 Whole-genome sequencing for typing and genomic characterization

sciensano .be

sciensano

healthy all life long

Thank you for your attention!
Questions?

Contact

Sigrid.dekeersmaecker@sciensano.be

Sciensano • Rue Juliette Wytsmanstraat 14 • 1050 Brussels • Belgium
T +32 2 642 51 11 • T Press +32 2 642 54 20 • info@sciensano.be • www.sciensano.be

.be