sciensano

healthy all life long

DEVELOPMENT OF A BIOINFORMATICS PIPELINE FOR ROUTINE ANALYSIS OF WHOLE GENOME SEQUENCING DATA OF *ESCHERICHIA COLI* ISOLATES

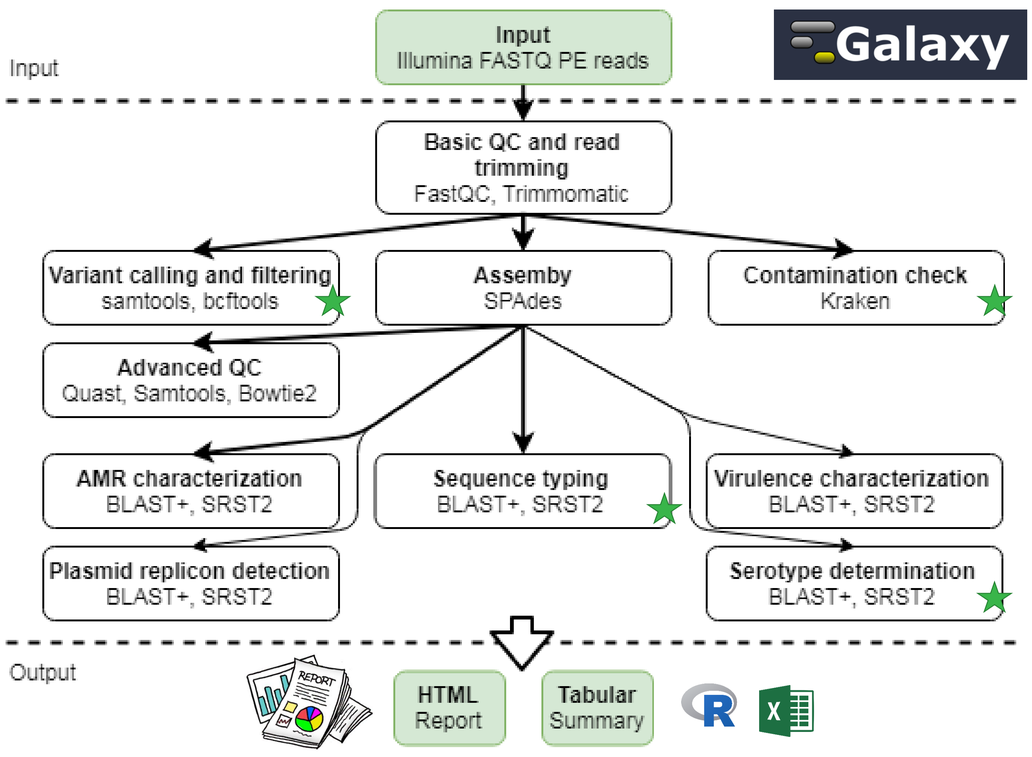
Bert Bogaerts[1,3], Stéphanie Nouws[1,3], Raf Winand[1], Qiang Fu[1], Julien Van Braekel[1], Sarah Denayer[2], Bavo Verhaegen[2], Sigrid C. J. De Keersmaecker[1], Nancy Roosens[1], Kathleen Marchal[3] and Kevin Vanneste[1]

1) Transversal activities in applied genomics, Sciensano, Brussels (1050), Belgium
2) Foodborne pathogens, Sciensano, Brussels (1050), Belgium
3) UGhent – Department of information Technology, IDLab, imec

Abstract

The adaptation of whole genome sequencing (WGS) and bioinformatics for routine molecular typing and pathogen characterization in a public health setting remains problematic, which is partly due to the lack of user-friendly and validated data analysis tools that can be used for routine typing in the National Reference Laboratories (NRLs) and peripheral laboratories. In collaboration with the Belgian NRL for *Escherichia coli*, we developed a pipeline for the routine analysis of *E. coli* isolates that was specifically designed to tackle the aforementioned challenges.

Pipeline architecture



The pipeline uses Illumina WGS data to perform a complete characterization of *Escherichia coli* isolates. The pipeline first generates basic quality reports before trimming the reads. Afterwards, a set of metrics specifically tailored for the pipeline are evaluated to determine if the data quality is sufficient to run the pipeline. The trimmed reads are then used as input for various downstream analyses: contamination check using Kraken, variant calling and filtering against the O157:H7 str. Sakai reference genome, and *de novo* assembly using **SPAdes**. All of the gene detection based assays can be performed by alignment using **blastn** or using a read mapping based approach with **SRST2**. The pipeline output is provided as either a **HTML report** or a tabular summary file which can be further analyses in other software.

★ Detailed on the rest of the poster

A) Sequence typing

The pipeline performs sequence typing using the MLST scheme and cgMLST from **PubMSLT.org**. Databases are automatically updated weekly. The allele detection is based on **BLAST+ alignment** of the de-novo assembly **or** on **read mapping using SRST2** depending on the pipeline setting. The pipeline output is shown below. The detected sequence type is reported at the top, **color codes** are used to indicate the type of hit (see legend on the right).

MLST (Bigsdb, Pasteur);
MLST, cgMLST, wgMLST (EnteroBase, Warwick)

Classic MLST (Pasteur)

ST
563

Locus	Allele	% Identity	HSP/Locus length	Type	Alignment
dinB	11	100.00	450/450	DNA	view
icdA	72	100.00	516/516	DNA	view
pabB	134	100.00	468/468	DNA	view
polB	52	100.00	450/450	DNA	view
putP	25	100.00	456/456	DNA	view
trpA	145	100.00	561/561	DNA	view
trpB	18	100.00	594/594	DNA	view
uidA	2	100.00	600/600	DNA	view

Hit type	Color
Perfect	
Imperfect identity	
Imperfect short	
Multi-hit	
No hit	

Download (TSV)

Last updated: 03-09-2018

B) Serotype determination

Serotype determination is based on the detection of gene variants that determine the **H-type** (*fliC*, *fllA*, *flkA*, *flnA*, *flmA*) and **O-type** (*wzx*, *wzy*, *wzt*, *wzm*). Detection is based on **BLAST+ alignment** of the de-novo assembly **or** on **read mapping using SRST2** depending on the pipeline setting. A set of decision rules is applied to determine the O-type, H-type and the combined serotype. If one the types cannot be determined it is reported as 'ambiguous'. The database sequences are retrieved from the SerotypeFinder tool provided by DTU, and are automatically updated on a weekly basis.

SerotypeFinder - O-type

Locus	Length	Coverage	Mismatches	Uncertainty	Depth	Predicted serotype	Accession
wzx_199	1287	98.76	16holes	edge2.0	6.08	O157	AKMA01000036
wzy_201	1185	99.41	7holes	edge0.0	4.62	O157	JH953200

Download (TSV)

Last updated: 03-09-2018

SerotypeFinder - H-type

Locus	Length	Coverage	Mismatches	Uncertainty	Depth	Predicted serotype	Accession
fliC_15	1758	100.00	0	edge1.0	7.71	H7	AF228487

Download (TSV)

Last updated: 03-09-2018

Detected serotype: **O157:H7**

C) Contamination check

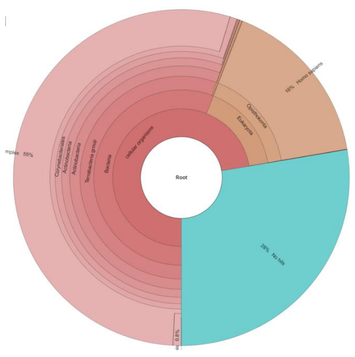
The pipeline uses **Kraken** to check for possible contaminants in the input sample. Warnings and errors are raised for unexpected species that have over 1% and 5% K-mer abundance, respectively. Results are visualized in an interactive **Krona** report.

Species	Percentage
Expected	
Escherichia coli	40.23
Contaminants	
None found	

Percentage	Level	Name
25.58	Unclassified	unclassified
74.42	-	root
74.40	-	cellular organisms
74.39	Domain	Bacteria
74.33	Phylum	Proteobacteria
74.27	Class	Gammaproteobacteria
74.18	Order	Enterobacterales
73.63	Family	Enterobacteriaceae
60.45	Genus	*Escherichia*
40.23	Species	*Escherichia coli*

Krona Report

Last updated: 24-01-2018

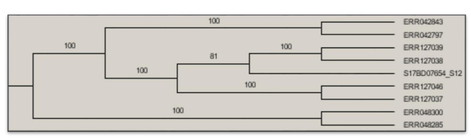


D) Variant calling & filtering

Bowtie2 is used to map the trimmed reads to the O157:H7 str. Sakai reference genome (**NC_002695.2**). Afterwards, *samtools mpileup* followed by *bcftools call* is used to call variants.

Variants are filtered with a set of in-house developed filters based on read depth, SNP quality, mapping quality, distance between variants and the ACTG frequencies at the variant position**.**

The resulting VCF files can be used to construct **phylogenetic trees** to determine the relatedness between isolates.



Reference: *Escherichia coli* O157:H7 (NC_002695.2)

Read mapping

Mapping rate
73.99%

RefSeq

Filtering

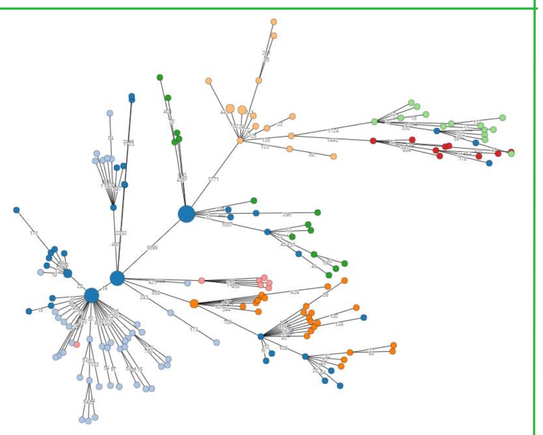
Filter	Variants passed
Depth	64772/67523
SNP quality	63828/64772
Mapping quality	45225/63828
Distance	39216/45225
Z-score	38986/39216

Output files

	Number of variants	VCF file
Unfiltered	67523	Download
Filtered (All positions)	38986	Download

Validation

We will extensively validate the different assays of the pipeline to demonstrate the employed methods are "fit-for-purpose" and provide high-quality results. Our approach will be similar to one employed in a previous pipeline validation for *Neisseria meningitides* (Bogaerts et al, https://doi.org/10.3389/fmicb.2019.00362). We will evaluate repeatability, reproducibility, accuracy, precision, sensitivity, and specificity of the different bioinformatics assays. The validation strategy will be extended to also include the relatedness between isolates based on the sequence typing output and variant calling assays.



Conclusion and future work

We present a pipeline for the complete characterization of *Escherichia coli* isolates using (Illumina) WGS data. The pipeline performance is currently being characterized by means of a set of performance metrics and definitions that were specifically adapted towards bioinformatics assays, and which evaluate repeatability, reproducibility, accuracy, sensitivity, precision, and specificity. Preliminary results on a representative set of samples demonstrate high performance, indicating the feasibility of using WGS in routine public health settings to replace classically employed pathogen typing and characterization techniques. Similar pipelines can be developed for other pathogens and case studies, making bioinformatics analyses less complex and more time-efficient for both expert and non-expert users.

Sciensano • Contact: Kevin Vanneste - T +32 2 642 50 65 - kevin.vanneste@sciensano.be • **www.sciensano.be**

.be