WETENSCHAPPELIJK INSTITUUT
VOLKSGEZONDHEID
INSTITUT SCIENTIFIQUE
DE SANTÉ PUBLIQUE

Kevin Vanneste, PhD

Bioinformatics Platform

Platform Biotechnology and Molecular Biology

Department Expertise, Service Provision and Customer Relations

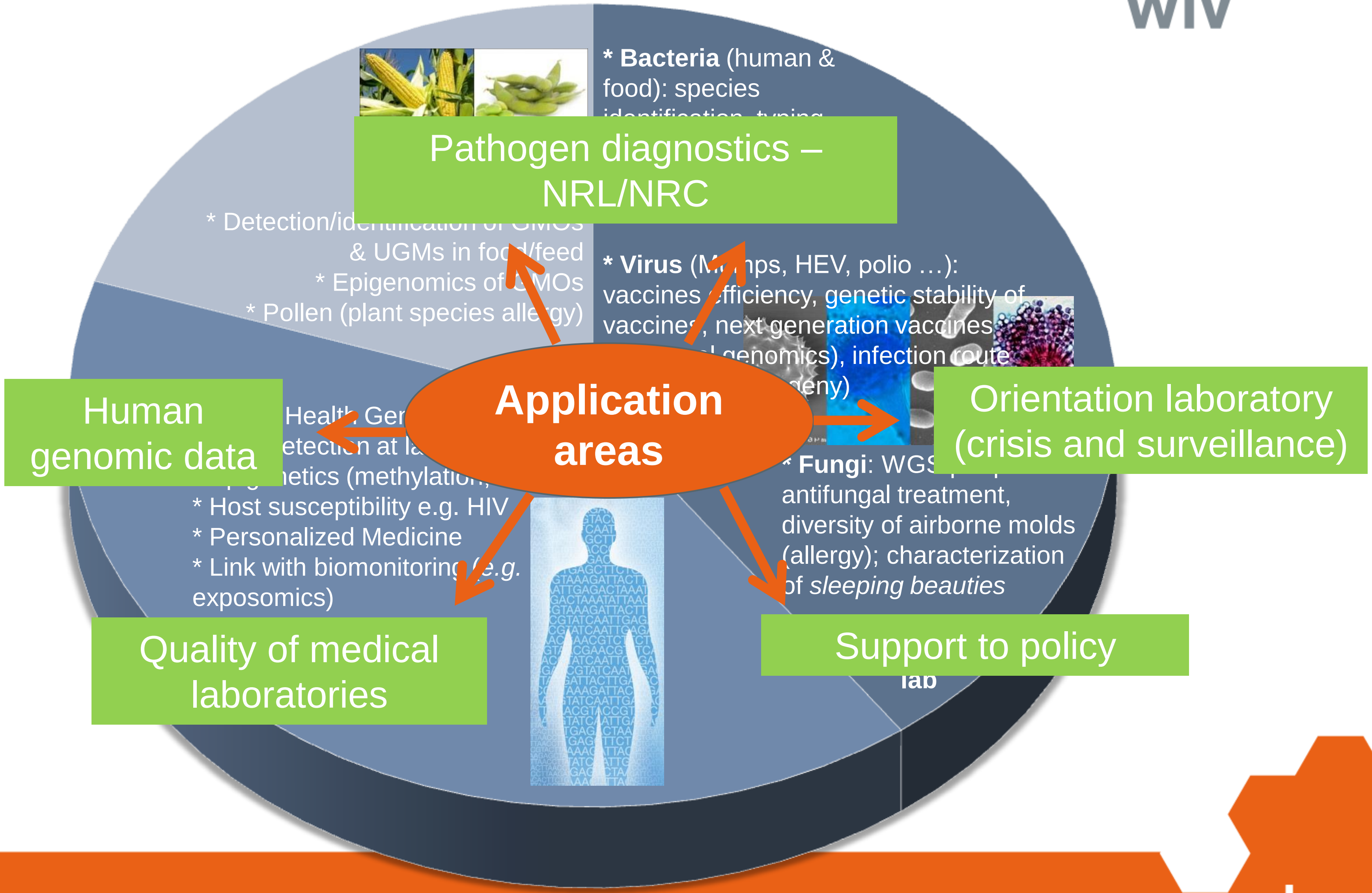
Collaboration between the EURL-VTEC and the Platform for Biotechnology and Molecular Biology (WIV-ISP, Belgium) for the development of a bioinformatics pipeline for routine analysis of whole genome sequencing data for typing of STEC/*E. coli* using Galaxy

12th Annual Workshop of the National Reference Laboratories for *E. coli* in the EU (12/10/2017-13/10/2017)

Rue Juliette Wytsmanstraat 14 | 1050 Brussels | Belgium
T +32 2 642 50 65 | F +32 2 642 52 92 | email: kevin.vanneste@wiv-isp.be | www.wiv-isp.be

.be

Use of NGS & bioinformatics @WIV-ISP?



Application areas

Pathogen diagnostics – NRL/NRC

Orientation laboratory (crisis and surveillance)

Support to policy

Quality of medical laboratories

Human genomic data

*** Bacteria** (human & food): species identification, typing

* Detection/identification of GMOs & UGMs in food/feed
* Epigenomics of GMOs
* Pollen (plant species allergy)

*** Virus** (M nps, HEV, polio …): vaccines efficiency, genetic stability of vaccines, next generation vaccines l genomics), infection route geny)

*** Fungi**: WGS antifungal treatment, diversity of airborne molds (allergy); characterization of *sleeping beauties*

lab

Health Ge etection at la netics (methylation,
* Host susceptibility e.g. HIV
* Personalized Medicine
* Link with biomonitoring (*e.g.* exposomics)

The Next-Generation Sequencing revolution

NGS as a well established research tool…

- Universal method
- Single nucleotide resolution
- High-throughput, ultimate multiplex tool

…but many challenges remain regarding data analysis and interpretation for routine applications in a public health setting!

- Computational requirements
- Validation of standardized & optimized pipelines
- User-friendly access for non-experts
- Trade-off between quality and speed of analysis
- Traceability (databases, runs)

NGS & bioinformatics platform

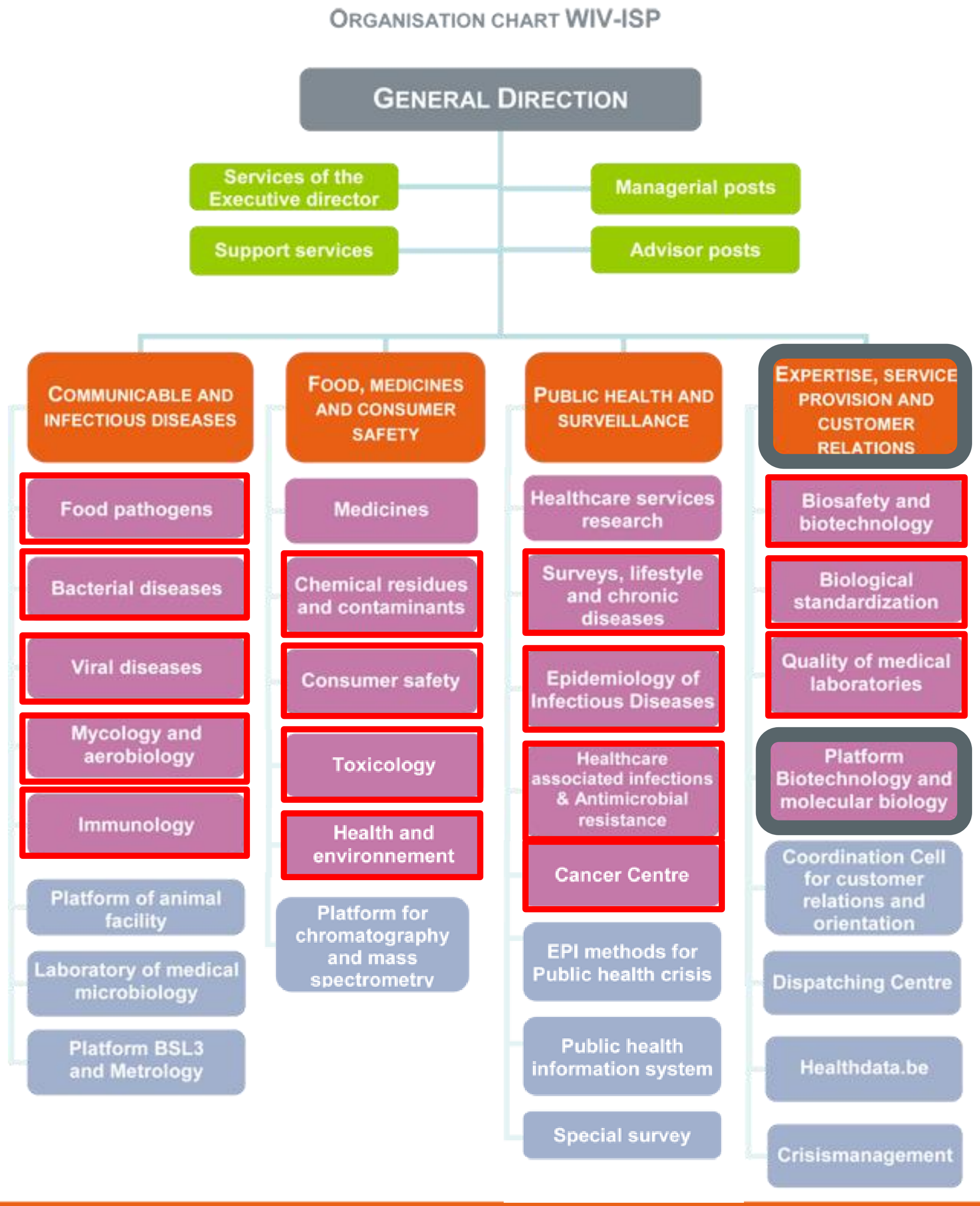
Mission: Utilizing NGS & bioinformatics

- for the diagnosis, surveillance, control and characterisation of potentially harmful organisms
- to promote public health genomics by the effective integration into clinical use and public health policy

Objectives: Develop and implement solutions and provide data acquisition and analysis tools to

- complement the WIV-ISP laboratories services
- integrate the knowledge of genomics into public health policy

NGS & bioinformatics platform @WIV-ISP



ORGANISATION CHART WIV-ISP

GENERAL DIRECTION

Services of the Executive director

Managerial posts

Support services

Advisor posts

COMMUNICABLE AND INFECTIOUS DISEASES
- Food pathogens
- Bacterial diseases
- Viral diseases
- Mycology and aerobiology
- Immunology
- Platform of animal facility
- Laboratory of medical microbiology
- Platform BSL3 and Metrology

FOOD, MEDICINES AND CONSUMER SAFETY
- Medicines
- Chemical residues and contaminants
- Consumer safety
- Toxicology
- Health and environnement
- Platform for chromatography and mass spectrometry

PUBLIC HEALTH AND SURVEILLANCE
- Healthcare services research
- Surveys, lifestyle and chronic diseases
- Epidemiology of Infectious Diseases
- Healthcare associated infections & Antimicrobial resistance
- Cancer Centre
- EPI methods for Public health crisis
- Public health information system
- Special survey

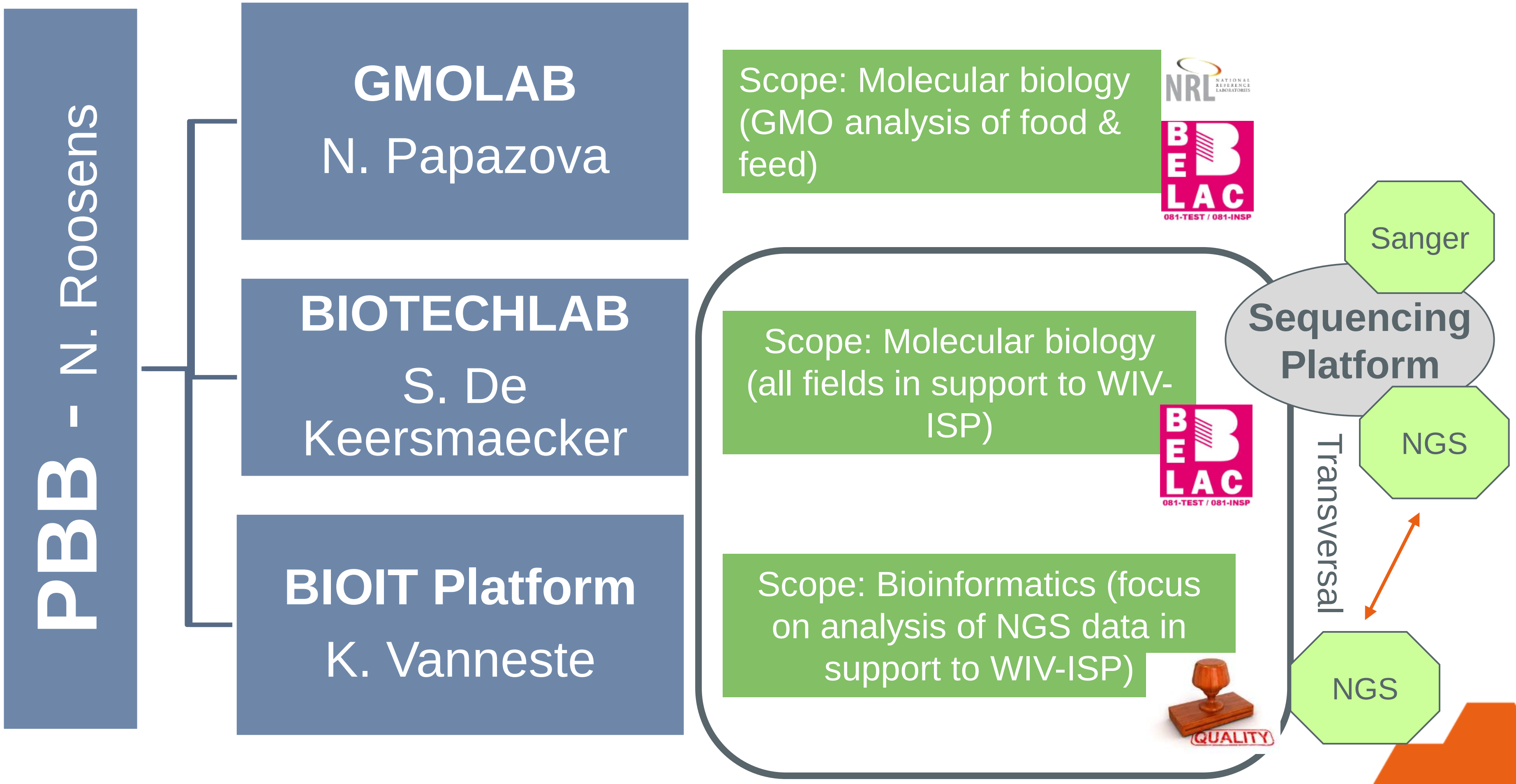
EXPERTISE, SERVICE PROVISION AND CUSTOMER RELATIONS
- Biosafety and biotechnology
- Biological standardization
- Quality of medical laboratories
- Platform Biotechnology and molecular biology
- Coordination Cell for customer relations and orientation
- Dispatching Centre
- Healthdata.be
- Crisismanagement

Mission

PBB is a transversal scientific service using **molecular biology & bioinformatics** to conduct routine analysis as well as scientific and technologic research.

By building internal and external partnership, it generates new knowledge & customized tools to anticipate present and futures challenges affecting public health

Platform Biotechnology and Molecular Biology



PBB - N. Roosens

GMOLAB
N. Papazova

Scope: Molecular biology (GMO analysis of food & feed)

BIOTECHLAB
S. De Keersmaecker

Scope: Molecular biology (all fields in support to WIV-ISP)

BIOIT Platform
K. Vanneste

Scope: Bioinformatics (focus on analysis of NGS data in support to WIV-ISP)

Sequencing Platform

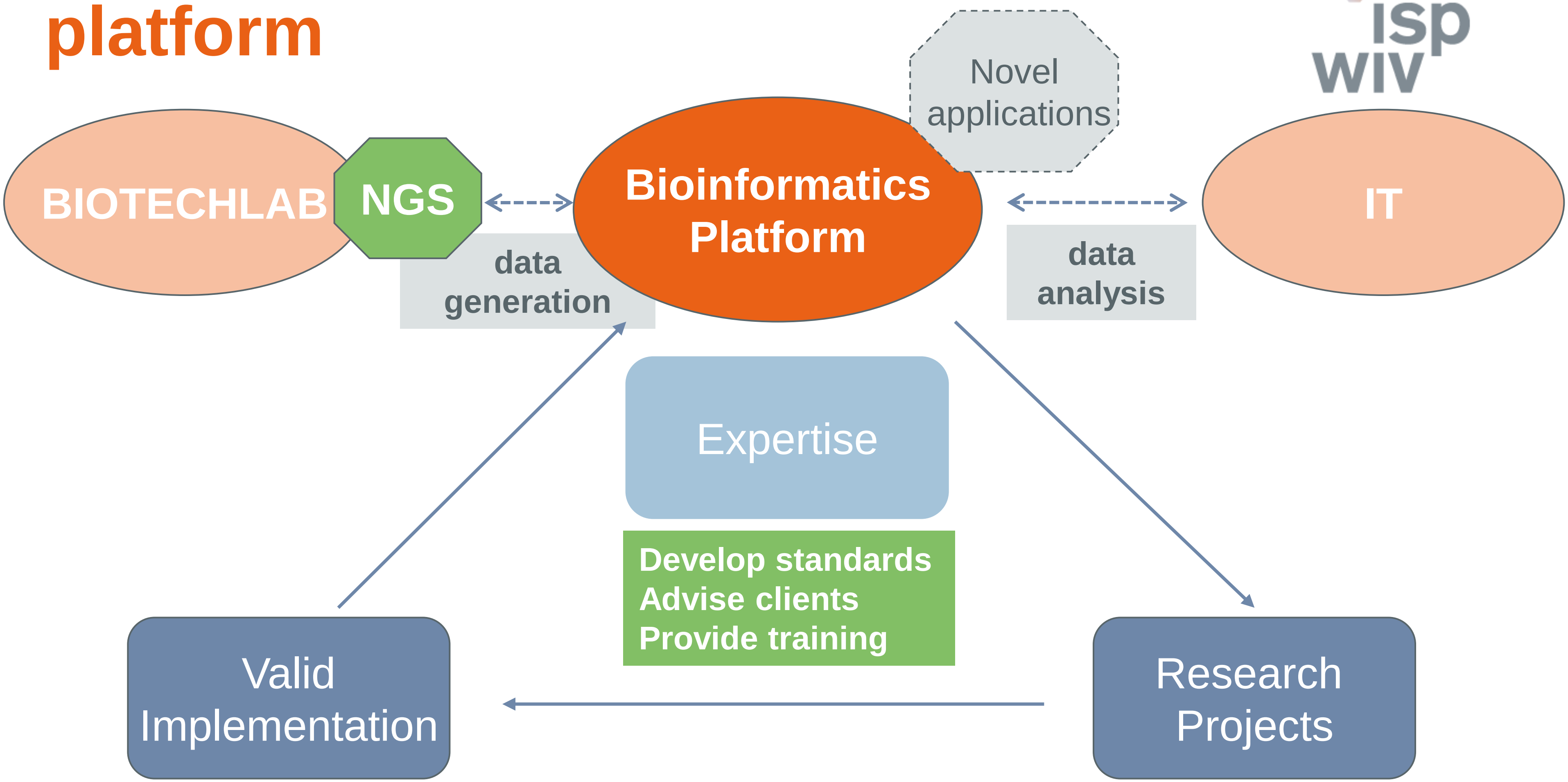
Sanger

NGS

NGS

Transversal

Activities of the bioinformatics platform



Develop standards
Advise clients
Provide training

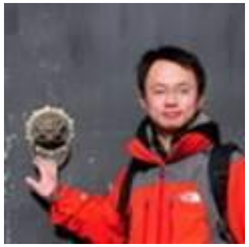
Centralized bioinformatics platform
User-friendly workflows
Obtain accreditation

Develop state-of-the-art pipelines
Collaborate (inter)nationally
Identify upcoming needs

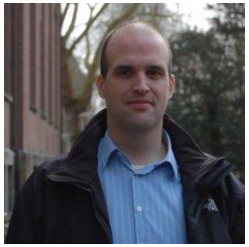
Develop and maintain centralized bioinformatics platform

Bioinformatics platform

Bioinformatician
(lead)



Bioinformatician



Bioinformatician

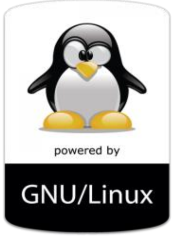
Bioinformatician

Bioinformatician

Software engineer

Development and implementation of user-friendly bioinformatics tools, pipelines and databases

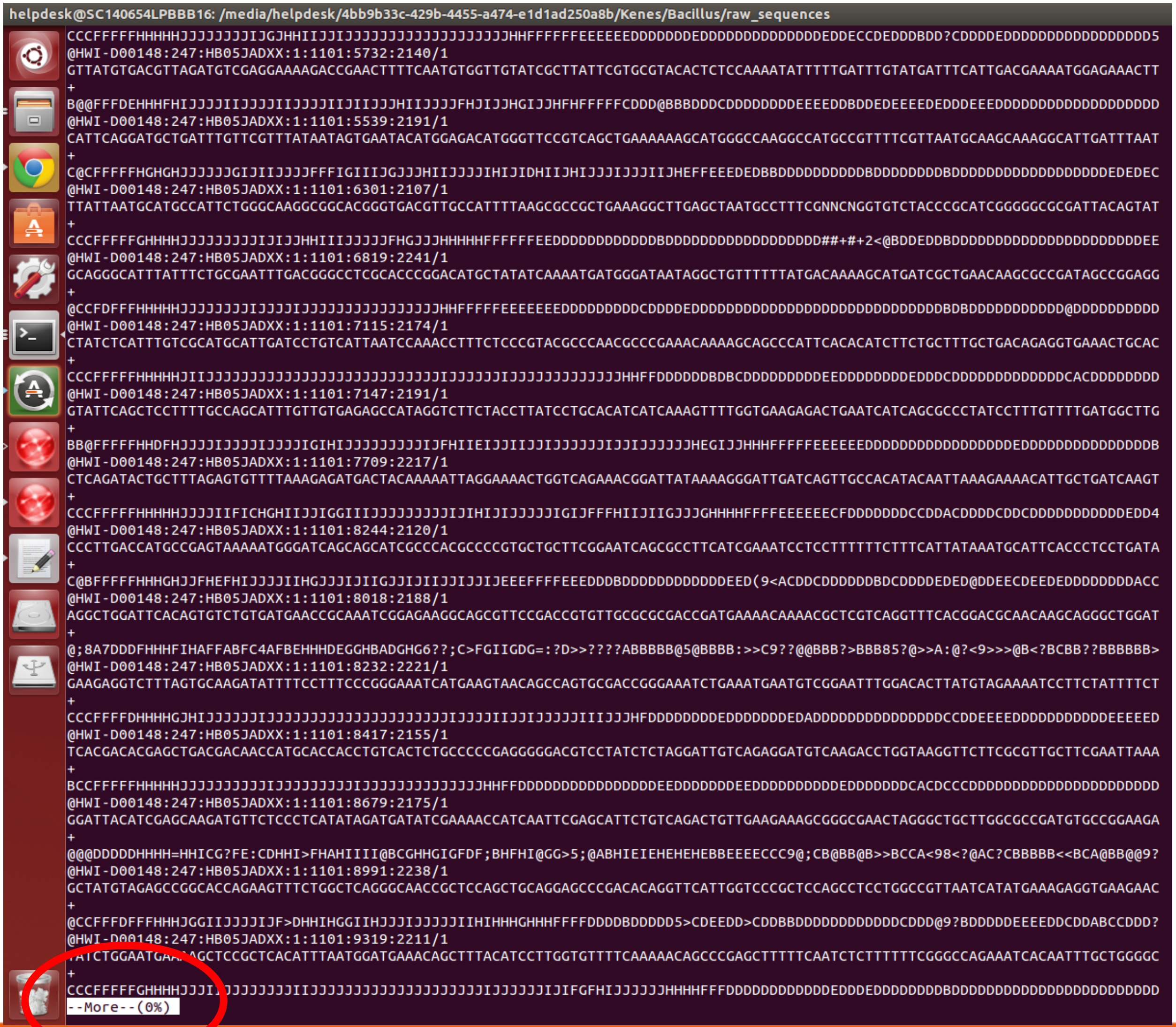
Commercial solutions



In-house developed solutions

Microbial isolates
Mixed samples
Human samples

The data analysis bottleneck



helpdesk@SC140654LPBBB16: /media/helpdesk/4bb9b33c-429b-4455-a474-e1d1ad250a8b/Kenes/Bacillus/raw_sequences

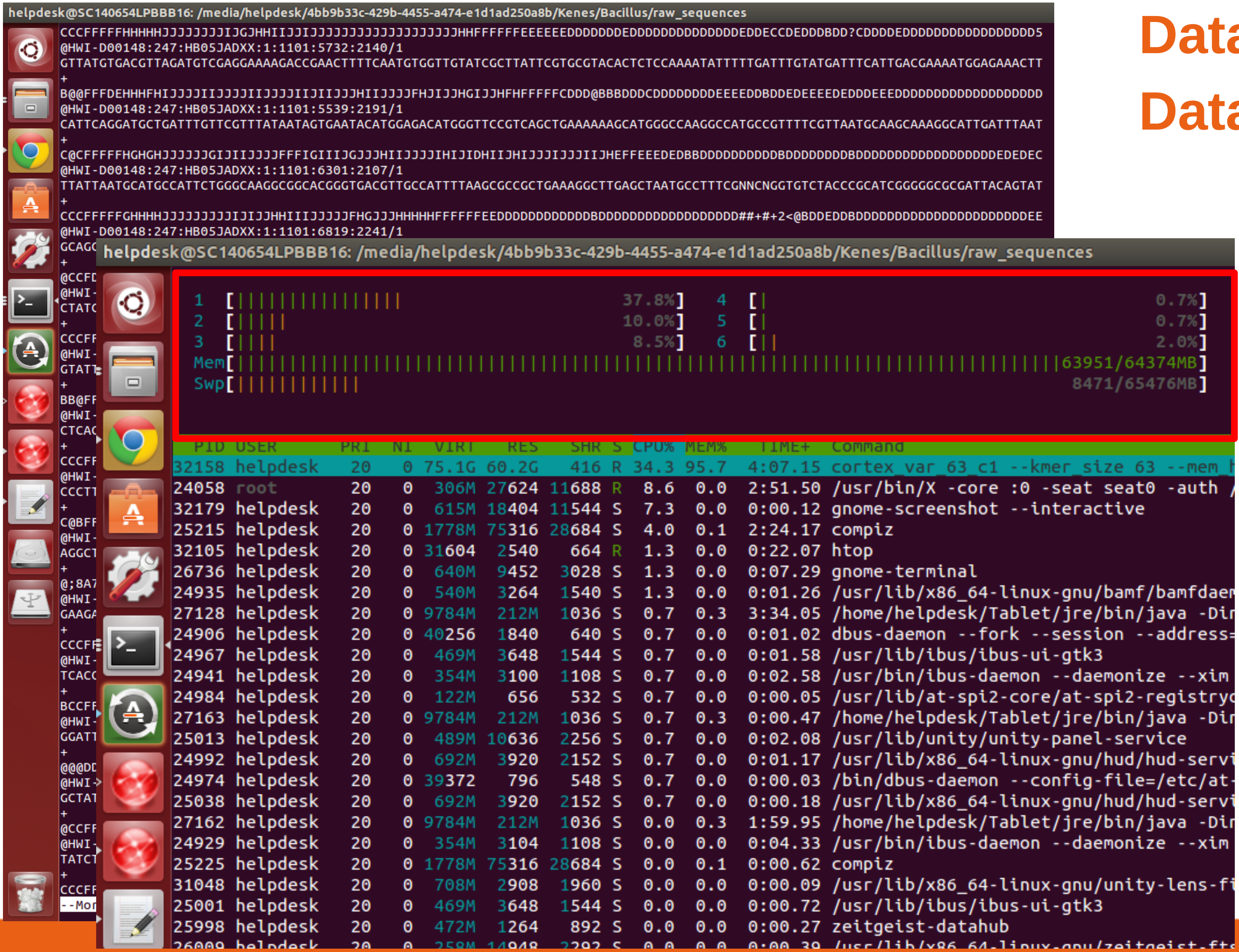
```
CCCFFFFFHHHHHJJJJJJJJIJGJHHIIJJIJJJJJJJJJJJJJJJJJJJJJJHHFFFFFFEEEEEEDDDDDDDEDDDDDDDDDDDDDDDEDDECCDEDDDBDD?CDDDDEDDDDDDDDDDDDDDDD5
@HWI-D00148:247:HB05JADXX:1:1101:5732:2140/1
GTTATGTGACGTTAGATGTCGAGGAAAAGACCGAACTTTTCAATGTGGTTGTATCGCTTATTCGTGCGTACACTCTCCAAAATATTTTTGATTTGTATGATTTCATTGACGAAAATGGAGAAACTT
+
B@@FFFDEHHHFHIJJJJIIJJJJIIJJJJIIJIIJJJHIIJJJJFHJIJJHGIJJHFHFFFFFCDDD@BBBDDDCDDDDDDDDEEEEDDBDDEDEEEEDEDDDEEEDDDDDDDDDDDDDDDDDDD
@HWI-D00148:247:HB05JADXX:1:1101:5539:2191/1
CATTCAGGATGCTGATTTGTTCGTTTATAATAGTGAATACATGGAGACATGGGTTCCGTCAGCTGAAAAAAGCATGGGCCAAGGCCATGCCGTTTTCGTTAATGCAAGCAAAGGCATTGATTTAAT
+
C@CFFFFFHGHGHJJJJJJGIJIIJJJJFFFIGIIIJGJJJHIIJJJJIHIJIDHIIJHIJJJIJJJIIJHEFFEEEDEDBBDDDDDDDDDDBDDDDDDDDBDDDDDDDDDDDDDDDDDDEDEDEC
@HWI-D00148:247:HB05JADXX:1:1101:6301:2107/1
TTATTAATGCATGCCATTCTGGGCAAGGCGGCACGGGTGACGTTGCCATTTTAAGCGCCGCTGAAAGGCTTGAGCTAATGCCTTTCGNNCNGGTGTCTACCCGCATCGGGGGCGCGATTACAGTAT
+
CCCFFFFFGHHHHJJJJJJJJJIJIJJHHIIIJJJJJFHGJJJHHHHHFFFFFFEEDDDDDDDDDDDDBDDDDDDDDDDDDDDDDDDD##+#+2<@BDDEDDBDDDDDDDDDDDDDDDDDDDDDDDDEE
@HWI-D00148:247:HB05JADXX:1:1101:6819:2241/1
GCAGGGCATTTATTTCTGCGAATTTGACGGGCCTCGCACCCGGACATGCTATATCAAAATGATGGGATAATAGGCTGTTTTTTATGACAAAAGCATGATCGCTGAACAAGCGCCGATAGCCGGAGG
+
@CCFDFFFHHHHHJJJJJJJJIJJJJIJJJJJJJJJJJJJJJJHHFFFFFEEEEEEEDDDDDDDDDCDDDDEDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDBDBDDDDDDDDDDD@DDDDDDDDDD
@HWI-D00148:247:HB05JADXX:1:1101:7115:2174/1
CTATCTCATTTGTCGCATGCATTGATCCTGTCATTAATCCAAACCTTTCTCCCGTACGCCCAACGCCCGAAACAAAAGCAGCCCATTCACACATCTTCTGCTTTGCTGACAGAGGTGAAACTGCAC
+
CCCFFFFFHHHHHJIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJJJJIJJJJJJJJJJJJJHHFFDDDDDDBDBCDDDDDDDDDEEDDDDDDDDEDDDCDDDDDDDDDDDDDCACDDDDDDDD
@HWI-D00148:247:HB05JADXX:1:1101:7147:2191/1
GTATTCAGCTCCTTTTGCCAGCATTTGTTGTGAGAGCCATAGGTCTTCTACCTTATCCTGCACATCATCAAAGTTTTGGTGAAGAGACTGAATCATCAGCGCCCTATCCTTTGTTTTGATGGCTTG
+
BB@FFFFFHHDFHJJJJIJJJJIJJJJIGIHIJJJJJJJJJIJFHIIEIJJIIJJIJJJJJJIJJIJJJJJJHEGIJJHHHFFFFFEEEEEEDDDDDDDDDDDDDDDDDEDDDDDDDDDDDDDDDDB
@HWI-D00148:247:HB05JADXX:1:1101:7709:2217/1
CTCAGATACTGCTTTAGAGTGTTTTAAAGAGATGACTACAAAAATTAGGAAAACTGGTCAGAAACGGATTATAAAAGGGATTGATCAGTTGCCACATACAATTAAAGAAAACATTGCTGATCAAGT
+
CCCFFFFFHHHHHJJJJIIFICHGHIIJJIGGIIIJJJJJJJJJIJIHIJIJJJJJIGIJFFFHIIJIIGJJJGHHHHFFFFEEEEEECFDDDDDDDCCDDACDDDDCDDCDDDDDDDDDDDDEDD4
@HWI-D00148:247:HB05JADXX:1:1101:8244:2120/1
CCCTTGACCATGCCGAGTAAAAATGGGATCAGCAGCATCGCCCACGCCCCGTGCTGCTTCGGAATCAGCGCCTTCATCGAAATCCTCCTTTTTTCTTTCATTATAAATGCATTCACCCTCCTGATA
+
C@BFFFFFHHHGHJJFHEFHIJJJJIIHGJJJIJIIGJJIJIIJJIJJIJEEEFFFFEEEDDDBDDDDDDDDDDDDDEED(9<ACDDCDDDDDDBDCDDDDEDED@DDEECDEEDEDDDDDDDDACC
@HWI-D00148:247:HB05JADXX:1:1101:8018:2188/1
AGGCTGGATTCACAGTGTCTGTGATGAACCGCAAATCGGAGAAGGCAGCGTTCCGACCGTGTTGCGCGCGACCGATGAAAACAAAACGCTCGTCAGGTTTCACGGACGCAACAAGCAGGGCTGGAT
+
@;8A7DDDFHHHFIHAFFABFC4AFBEHHHDEGGHBADGHG6??;C>FGIIGDG=:?D>>????ABBBBB@5@BBBB:>>C9??@@BBB?>BBB85?@>>A:@?<9>>>@B<?BCBB??BBBBBB>
@HWI-D00148:247:HB05JADXX:1:1101:8232:2221/1
GAAGAGGTCTTTAGTGCAAGATATTTTCCTTTCCCGGGAAATCATGAAGTAACAGCCAGTGCGACCGGGAAATCTGAAATGAATGTCGGAATTTGGACACTTATGTAGAAAATCCTTCTATTTTCT
+
CCCFFFFFDHHHHGJHIJJJJJJIJJJJJJJJJJJJJJJJJJJJJJJJIJJJJIIJJIJJJJJJIIIJJJHFDDDDDDDDEDDDDDDDEDADDDDDDDDDDDDDDDDCCDDEEEEDDDDDDDDDDDEEEEED
@HWI-D00148:247:HB05JADXX:1:1101:8417:2155/1
TCACGACACGAGCTGACGACAACCATGCACCACCTGTCACTCTGCCCCCGAGGGGGACGTCCTATCTCTAGGATTGTCAGAGGATGTCAAGACCTGGTAAGGTTCTTCGCGTTGCTTCGAATTAAA
+
BCCFFFFFHHHHHJJJJJJJJJJIJJJJJJJJJJIJJJJJJJJJJJJJJJHHFFDDDDDDDDDDDDDDDDDDEEDDDDDDDEEDDDDDDDDDDEDDDDDDDCACDCCCDDDDDDDDDDDDDDDDDDDDDDD
@HWI-D00148:247:HB05JADXX:1:1101:8679:2175/1
GGATTACATCGAGCAAGATGTTCTCCCTCATATAGATGATATCGAAAACCATCAATTCGAGCATTCTGTCAGACTGTTGAAGAAAGCGGGCGAACTAGGGCTGCTTGGCGCCGATGTGCCGGAAGA
+
@@@DDDDDHHHH=HHICG?FE:CDHHI>FHAHIIII@BCGHHGIGFDF;BHFHI@GG>5;@ABHIEIEHEHEHEBBEEEECCC9@;CB@BB@B>>BCCA<98<?@AC?CBBBBB<<BCA@BB@@9?
@HWI-D00148:247:HB05JADXX:1:1101:8991:2238/1
GCTATGTAGAGCCGGCACCAGAAGTTTCTGGCTCAGGGCAACCGCTCCAGCTGCAGGAGCCCGACACAGGTTCATTGGTCCCGCTCCAGCCTCCTGGCCGTTAATCATATGAAAGAGGTGAAGAAC
+
@CCFFFDFFFHHHJGGIIJJJJIJF>DHHIHGGIIHJJJIJJJJJIIHIHHHGHHHFFFFDDDDBDDDDD5>CDEEDD>CDDBBDDDDDDDDDDDDDCDDD@9?BDDDDDEEEEDDCDDABCCDDD?
@HWI-D00148:247:HB05JADXX:1:1101:9319:2211/1
TATCTGGAATGAAAAGCTCCGCTCACATTTAATGGATGAAACAGCTTTACATCCTTGGTGTTTTCAAAAACAGCCCGAGCTTTTTCAATCTCTTTTTTCGGGCCAGAAATCACAATTTGCTGGGGC
+
CCCFFFFFGHHHHJJJIJJJJJJJJIIJJJJJJJJJJJJJJJJJJJJJJIJJJJJJIJIFGFHIJJJJJHHHHFFFDDDDDDDDDDDEDDDEDDDDDDDDBDDDDDDDDDDDDDDDDDDDDDDDDDD
--More--(0%)
```

Data storage?

~10s GB raw data (microbial)

~100s GB raw data (human)

The data analysis bottleneck



```
helpdesk@SC140654LPBBB16: /media/helpdesk/4bb9b33c-429b-4455-a474-e1d1ad250a8b/Kenes/Bacillus/raw_sequences
CCCFFFFFHHHHHJJJJJJJJIJGJHHIIJJIJJJJJJJJJJJJJJJJJJJJJHHFFFFFFEEEEEEDDDDDDDEDDDDDDDDDDDDDDDEDDECCDEDDDBDD?CDDDDEDDDDDDDDDDDDDDDDDD5
@HWI-D00148:247:HB05JADXX:1:1101:5732:2140/1
GTTATGTGACGTTAGATGTCGAGGAAAAGACCGAACTTTTCAATGTGGTTGTATCGCTTATTCGTGCGTACACTCTCCAAAATATTTTTGATTTGTATGATTTCATTGACGAAAATGGAGAAACTT
+
B@@FFFDEHHHFHIJJJJIIJJJJIIJJJJIIJIIJJJHIIJJJJFHJIJJHGIJJHFHFFFFFCDDD@BBBDDDCDDDDDDDDEEEEDDBDDEDEEEEDEDDDEEEDDDDDDDDDDDDDDDDDDD
@HWI-D00148:247:HB05JADXX:1:1101:5539:2191/1
CATTCAGGATGCTGATTTGTTCGTTTATAATAGTGAATACATGGAGACATGGGTTCCGTCAGCTGAAAAAAGCATGGGCCAAGGCCATGCCGTTTTCGTTAATGCAAGCAAAGGCATTGATTTAAT
+
C@CFFFFFHGHGHJJJJJJGIJIIJJJJFFFIGIIIJGJJJHIIJJJJIHIJIDHIIJHIJJJIJJJIIJHEFFEEEDEDBBDDDDDDDDDDDBDDDDDDDDBDDDDDDDDDDDDDDDDDDEDEDEC
@HWI-D00148:247:HB05JADXX:1:1101:6301:2107/1
TTATTAATGCATGCCATTCTGGGCAAGGCGGCACGGGTGACGTTGCCATTTTAAGCGCCGCTGAAAGGCTTGAGCTAATGCCTTTCGNNCNGGTGTCTACCCGCATCGGGGGCGCGATTACAGTAT
+
CCCFFFFFGHHHHJJJJJJJJJIJIJJHHIIIJJJJJFHGJJJHHHHHHFFFFFFEEDDDDDDDDDDDDDBDDDDDDDDDDDDDDDDDDDD##+#+2<@BDDEDDBDDDDDDDDDDDDDDDDDDDDDEE
@HWI-D00148:247:HB05JADXX:1:1101:6819:2241/1
GCAGC
+
@CCFD
@HWI-
CTATC
+
CCCFF
@HWI-
GTATT
+
BB@FF
@HWI-
CTCAC
+
CCCFF
@HWI-
CCCTT
+
C@BFF
@HWI-
AGGCT
+
@;8A7
@HWI-
GAAGA
+
CCCFF
@HWI-
TCACC
+
BCCFF
@HWI-
GGATT
+
@@@DD
@HWI->
GCTAT
+
@CCFF
@HWI-
TATCT
+
CCCFF
--Mor
```

```
helpdesk@SC140654LPBBB16: /media/helpdesk/4bb9b33c-429b-4455-a474-e1d1ad250a8b/Kenes/Bacillus/raw_sequences

  1  [||||||||||||||||                          37.8%]    4  [|                                        0.7%]
  2  [|||||                                     10.0%]    5  [|                                        0.7%]
  3  [||||                                       8.5%]    6  [||                                       2.0%]
  Mem[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||63951/64374MB]
  Swp[||||||||||||                                                                        8471/65476MB]

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
32158 helpdesk   20   0 75.1G 60.2G   416 R 34.3 95.7  4:07.15 cortex_var_63_c1 --kmer_size 63 --mem_h
24058 root       20   0  306M 27624 11688 R  8.6  0.0  2:51.50 /usr/bin/X -core :0 -seat seat0 -auth /
32179 helpdesk   20   0  615M 18404 11544 S  7.3  0.0  0:00.12 gnome-screenshot --interactive
25215 helpdesk   20   0 1778M 75316 28684 S  4.0  0.1  2:24.17 compiz
32105 helpdesk   20   0 31604  2540   664 R  1.3  0.0  0:22.07 htop
26736 helpdesk   20   0  640M  9452  3028 S  1.3  0.0  0:07.29 gnome-terminal
24935 helpdesk   20   0  540M  3264  1540 S  1.3  0.0  0:01.26 /usr/lib/x86_64-linux-gnu/bamf/bamfdaem
27128 helpdesk   20   0 9784M  212M  1036 S  0.7  0.3  3:34.05 /home/helpdesk/Tablet/jre/bin/java -Dir
24906 helpdesk   20   0 40256  1840   640 S  0.7  0.0  0:01.02 dbus-daemon --fork --session --address=
24967 helpdesk   20   0  469M  3648  1544 S  0.7  0.0  0:01.58 /usr/lib/ibus/ibus-ui-gtk3
24941 helpdesk   20   0  354M  3100  1108 S  0.7  0.0  0:02.58 /usr/bin/ibus-daemon --daemonize --xim
24984 helpdesk   20   0  122M   656   532 S  0.7  0.0  0:00.05 /usr/lib/at-spi2-core/at-spi2-registryd
27163 helpdesk   20   0 9784M  212M  1036 S  0.7  0.3  0:00.47 /home/helpdesk/Tablet/jre/bin/java -Dir
25013 helpdesk   20   0  489M 10636  2256 S  0.7  0.0  0:02.08 /usr/lib/unity/unity-panel-service
24992 helpdesk   20   0  692M  3920  2152 S  0.7  0.0  0:01.17 /usr/lib/x86_64-linux-gnu/hud/hud-servi
24974 helpdesk   20   0 39372   796   548 S  0.7  0.0  0:00.03 /bin/dbus-daemon --config-file=/etc/at-
25038 helpdesk   20   0  692M  3920  2152 S  0.7  0.0  0:00.18 /usr/lib/x86_64-linux-gnu/hud/hud-servi
27162 helpdesk   20   0 9784M  212M  1036 S  0.0  0.3  1:59.95 /home/helpdesk/Tablet/jre/bin/java -Dir
24929 helpdesk   20   0  354M  3104  1108 S  0.0  0.0  0:04.33 /usr/bin/ibus-daemon --daemonize --xim
25225 helpdesk   20   0 1778M 75316 28684 S  0.0  0.1  0:00.62 compiz
31048 helpdesk   20   0  708M  2908  1960 S  0.0  0.0  0:00.09 /usr/lib/x86_64-linux-gnu/unity-lens-fi
25001 helpdesk   20   0  469M  3648  1544 S  0.0  0.0  0:00.72 /usr/lib/ibus/ibus-ui-gtk3
25998 helpdesk   20   0  472M  1264   892 S  0.0  0.0  0:00.27 zeitgeist-datahub
26009 helpdesk   20   0  258M 14948  2292 S  0.0  0.0  0:00.39 /usr/lib/x86_64-linux-gnu/zeitgeist-fts
```

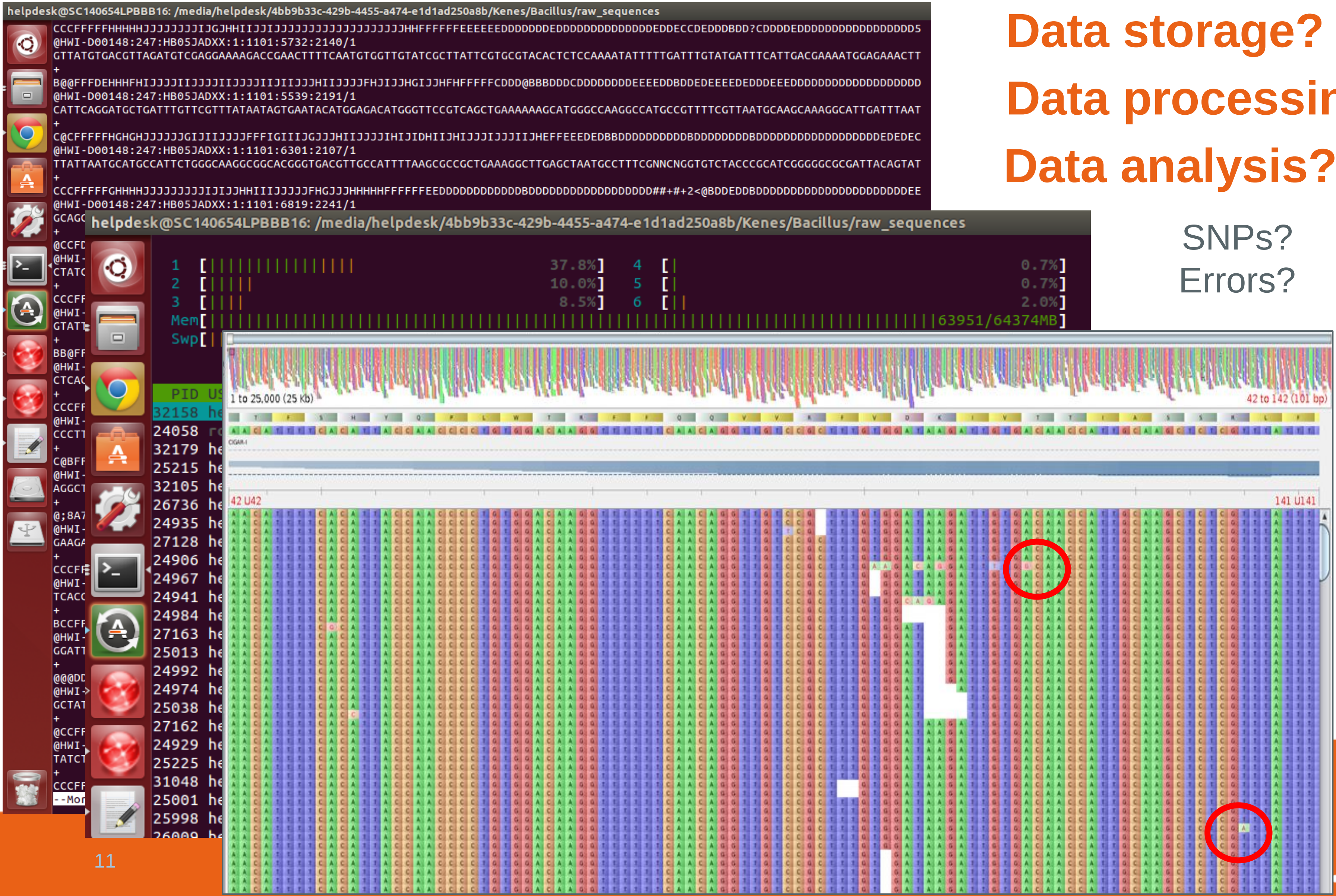
Data storage?

Data processing?

~10s GB RAM (microbial)

~100s GB RAM (human)

The data analysis bottleneck



helpdesk@SC140654LPBBB16: /media/helpdesk/4bb9b33c-429b-4455-a474-e1d1ad250a8b/Kenes/Bacillus/raw_sequences

```
CCCFFFFFHHHHHJJJJJJJJIJGJHHIIJJIJJJJJJJJJJJJJJJJJJJJHHFFFFFFEEEEEEDDDDDDDEDDDDDDDDDDDDDDDEDDECCDEDDDBDD?CDDDDEDDDDDDDDDDDDDDDDDD5
@HWI-D00148:247:HB05JADXX:1:1101:5732:2140/1
GTTATGTGACGTTAGATGTCGAGGAAAAGACCGAACTTTTCAATGTGGTTGTATCGCTTATTCGTGCGTACACTCTCCAAAATATTTTTGATTTGTATGATTTCATTGACGAAAATGGAGAAACTT
+
B@@FFFDEHHHFHIJJJJIIJJJJIIJJJJIIJIIJJJHIIJJJJFHJIJJHGIJJHFHFFFFFCDDD@BBBDDDCDDDDDDDDEEEEDDBDDEDEEEEDEDDDEEEDDDDDDDDDDDDDDDDDDD
@HWI-D00148:247:HB05JADXX:1:1101:5539:2191/1
CATTCAGGATGCTGATTTGTTCGTTTATAATAGTGAATACATGGAGACATGGGTTCCGTCAGCTGAAAAAAGCATGGGCCAAGGCCATGCCGTTTTCGTTAATGCAAGCAAAGGCATTGATTTAAT
+
C@CFFFFFHGHGHJJJJJGIJIIJJJJFFFIGIIIJGJJJHIIJJJJIHIJIDHIIJHIJJJIJJJIIJHEFFEEEDEDBBDDDDDDDDDDDBDDDDDDDDBDDDDDDDDDDDDDDDDDDDEDEDEC
@HWI-D00148:247:HB05JADXX:1:1101:6301:2107/1
TTATTAATGCATGCCATTCTGGGCAAGGCGGCACGGGTGACGTTGCCATTTTAAGCGCCGCTGAAAGGCTTGAGCTAATGCCTTTCGNNCNGGTGTCTACCCGCATCGGGGGCGCGATTACAGTAT
+
CCCFFFFFGHHHHJJJJJJJJJIJIJJHHIIIJJJJJFHGJJJHHHHHFFFFFFEEDDDDDDDDDDDDDBDDDDDDDDDDDDDDDDDDD##+#+2<@BDDEDDBDDDDDDDDDDDDDDDDDDDDDDDEE
@HWI-D00148:247:HB05JADXX:1:1101:6819:2241/1
GCAGC
+
@CCFD
@HWI-
CTATC
+
CCCFF
@HWI-
GTATT
+
BB@FF
@HWI-
CTCAC
+
CCCFF
@HWI-
CCCTT
+
C@BFF
@HWI-
AGGCT
+
@;8A7
@HWI-
GAAGA
+
CCCFF
@HWI-
TCACC
+
BCCFF
@HWI-
GGATT
+
@@@DD
@HWI-
GCTAT
+
@CCFF
@HWI-
TATCT
+
CCCFF
--Mor
```

helpdesk@SC140654LPBBB16: /media/helpdesk/4bb9b33c-429b-4455-a474-e1d1ad250a8b/Kenes/Bacillus/raw_sequences

```
  1  [|||||||||||||||||                                      37.8%]     4  [|                                                      0.7%]
  2  [|||||                                                  10.0%]     5  [|                                                      0.7%]
  3  [||||                                                    8.5%]     6  [||                                                     2.0%]
  Mem[|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||63951/64374MB]
  Swp[|

  PID US
32158 he
24058 r
32179 he
25215 he
32105 he
26736 he
24935 he
27128 he
24906 he
24967 he
24941 he
24984 he
27163 he
25013 he
24992 he
24974 he
25038 he
27162 he
24929 he
25225 he
31048 he
25001 he
25998 he
26009 he
```

Data storage?

Data processing?

Data analysis?

SNPs?
Errors?

Bring data analysis into routine

Tools

Staple bioinformatics tools used and adopted by the scientific community

Tools can be combined to make pipelines

'Push-on-the-button' pipelines

Pipelines engineered by the BIOT platform

Case studies tackled according to priorities defined by direction committee

E.g. *Neisseria meningitidis* pipeline

Computational requirements

Tools and pipelines integrated directly into high-performance computational infrastructure WIV-ISP

Validation of standardized & optimized pipelines

Pipelines use validated parameters

User-friendly access for non-experts

Galaxy Workflow Management System for access to non-bioinformaticians

Trade-off between quality and speed of analysis

Different modes of analysis (e.g. surveillance versus outbreak)

Traceability

Automatically updated databases, logging of all parameters and runs

Offer a high-quality service platform

Providing a high-quality service....

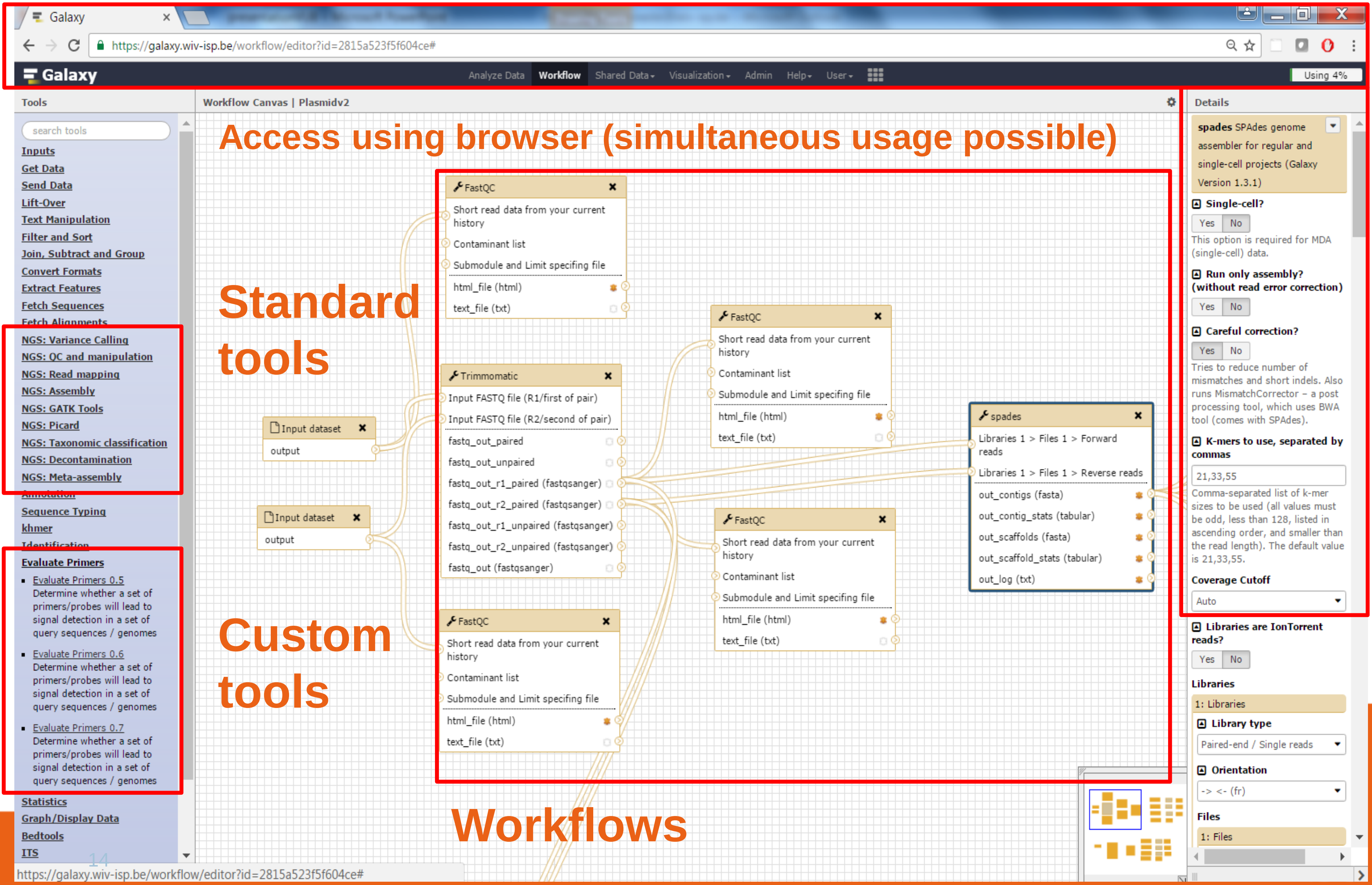
Version control
Code review
Basic testing
Technical documentation (bioit Wiki)
User documentation (bioit Wiki)
Several in-take meetings with client to define needs
DTAP principle (**D**evelopment -> **T**esting -> **A**cceptance -> **P**roduction)

Building up the quality system…

2017: Benchmarking to consolidate internal quality system
2018: Obtain certification / accreditation

Combine tools to make pipelines

Tweak parameters



Access using browser (simultaneous usage possible)

Standard tools

Custom tools

Workflows

Galaxy

https://galaxy.wiv-isp.be/workflow/editor?id=2815a523f5f604ce#

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 4%

Tools

search tools

- Inputs
- Get Data
- Send Data
- Lift-Over
- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Convert Formats
- Extract Features
- Fetch Sequences
- Fetch Alignments
- NGS: Variance Calling
- NGS: QC and manipulation
- NGS: Read mapping
- NGS: Assembly
- NGS: GATK Tools
- NGS: Picard
- NGS: Taxonomic classification
- NGS: Decontamination
- NGS: Meta-assembly
- Annotation
- Sequence Typing
- khmer
- Identification
- Evaluate Primers
  - Evaluate Primers 0.5 Determine whether a set of primers/probes will lead to signal detection in a set of query sequences / genomes
  - Evaluate Primers 0.6 Determine whether a set of primers/probes will lead to signal detection in a set of query sequences / genomes
  - Evaluate Primers 0.7 Determine whether a set of primers/probes will lead to signal detection in a set of query sequences / genomes
- Statistics
- Graph/Display Data
- Bedtools
- ITS

Workflow Canvas | Plasmidv2

Input dataset: output

Input dataset: output

FastQC: Short read data from your current history; Contaminant list; Submodule and Limit specifing file; html_file (html); text_file (txt)

Trimmomatic: Input FASTQ file (R1/first of pair); Input FASTQ file (R2/second of pair); fastq_out_paired; fastq_out_unpaired; fastq_out_r1_paired (fastqsanger); fastq_out_r2_paired (fastqsanger); fastq_out_r1_unpaired (fastqsanger); fastq_out_r2_unpaired (fastqsanger); fastq_out (fastqsanger)

FastQC: Short read data from your current history; Contaminant list; Submodule and Limit specifing file; html_file (html); text_file (txt)

FastQC: Short read data from your current history; Contaminant list; Submodule and Limit specifing file; html_file (html); text_file (txt)

FastQC: Short read data from your current history; Contaminant list; Submodule and Limit specifing file; html_file (html); text_file (txt)

spades: Libraries 1 > Files 1 > Forward reads; Libraries 1 > Files 1 > Reverse reads; out_contigs (fasta); out_contig_stats (tabular); out_scaffolds (fasta); out_scaffold_stats (tabular); out_log (txt)

Details

spades SPAdes genome assembler for regular and single-cell projects (Galaxy Version 1.3.1)

Single-cell? Yes No
This option is required for MDA (single-cell) data.

Run only assembly? (without read error correction) Yes No

Careful correction? Yes No
Tries to reduce number of mismatches and short indels. Also runs MismatchCorrector – a post processing tool, which uses BWA tool (comes with SPAdes).

K-mers to use, separated by commas
21,33,55
Comma-separated list of k-mer sizes to be used (all values must be odd, less than 128, listed in ascending order, and smaller than the read length). The default value is 21,33,55.

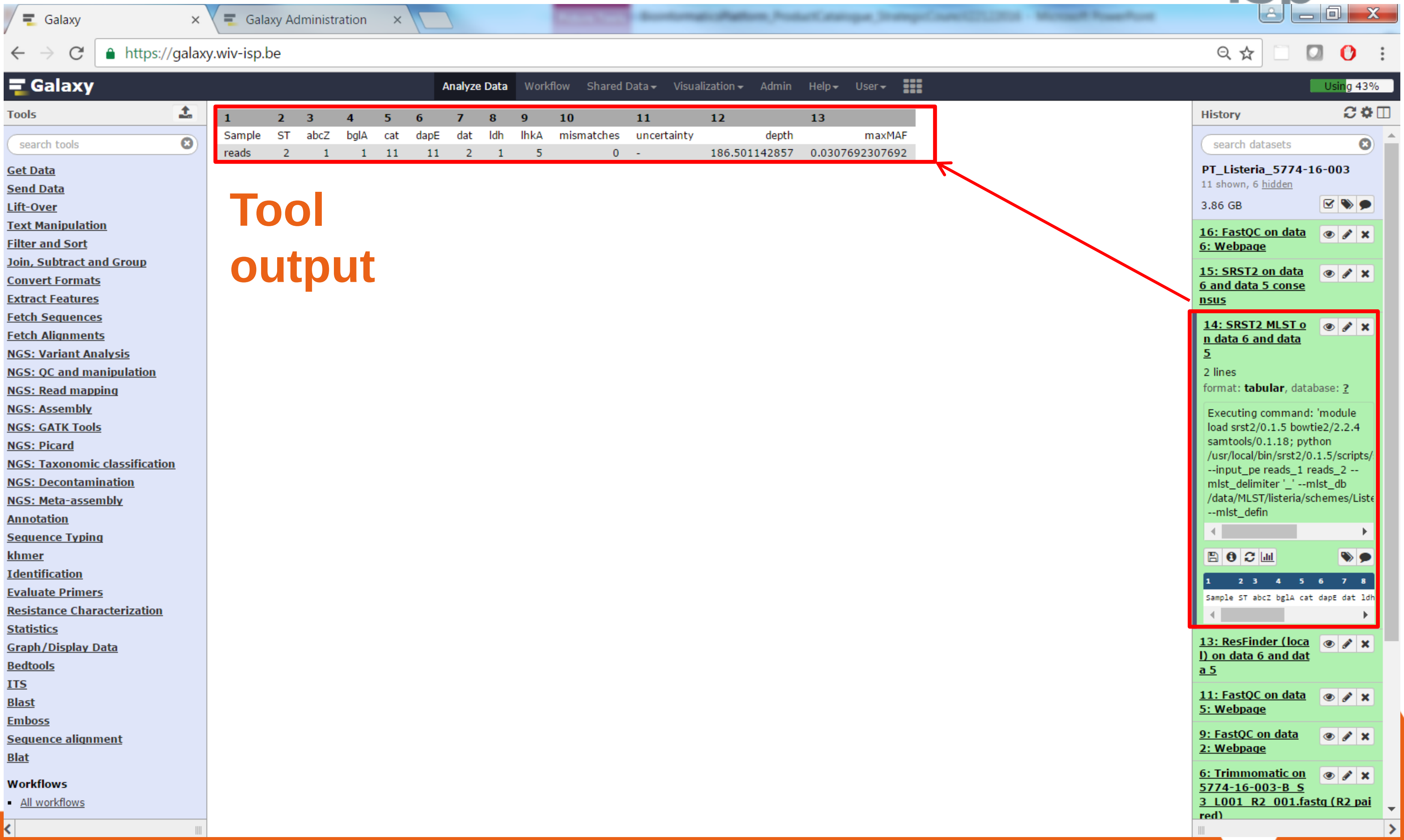
Coverage Cutoff
Auto

Libraries are IonTorrent reads? Yes No

Libraries
1: Libraries
Library type
Paired-end / Single reads
Orientation
-> <- (fr)
Files
1: Files

https://galaxy.wiv-isp.be/workflow/editor?id=2815a523f5f604ce#

Combine tools to make pipelines



1	2	3	4	5	6	7	8	9	10	11	12	13
Sample	ST	abcZ	bglA	cat	dapE	dat	ldh	lhkA	mismatches	uncertainty	depth	maxMAF
reads	2	1	1	11	11	2	1	5	0	-	186.501142857	0.0307692307692

Tool output

14: SRST2 MLST on data 6 and data 5

2 lines

format: **tabular**, database: ?

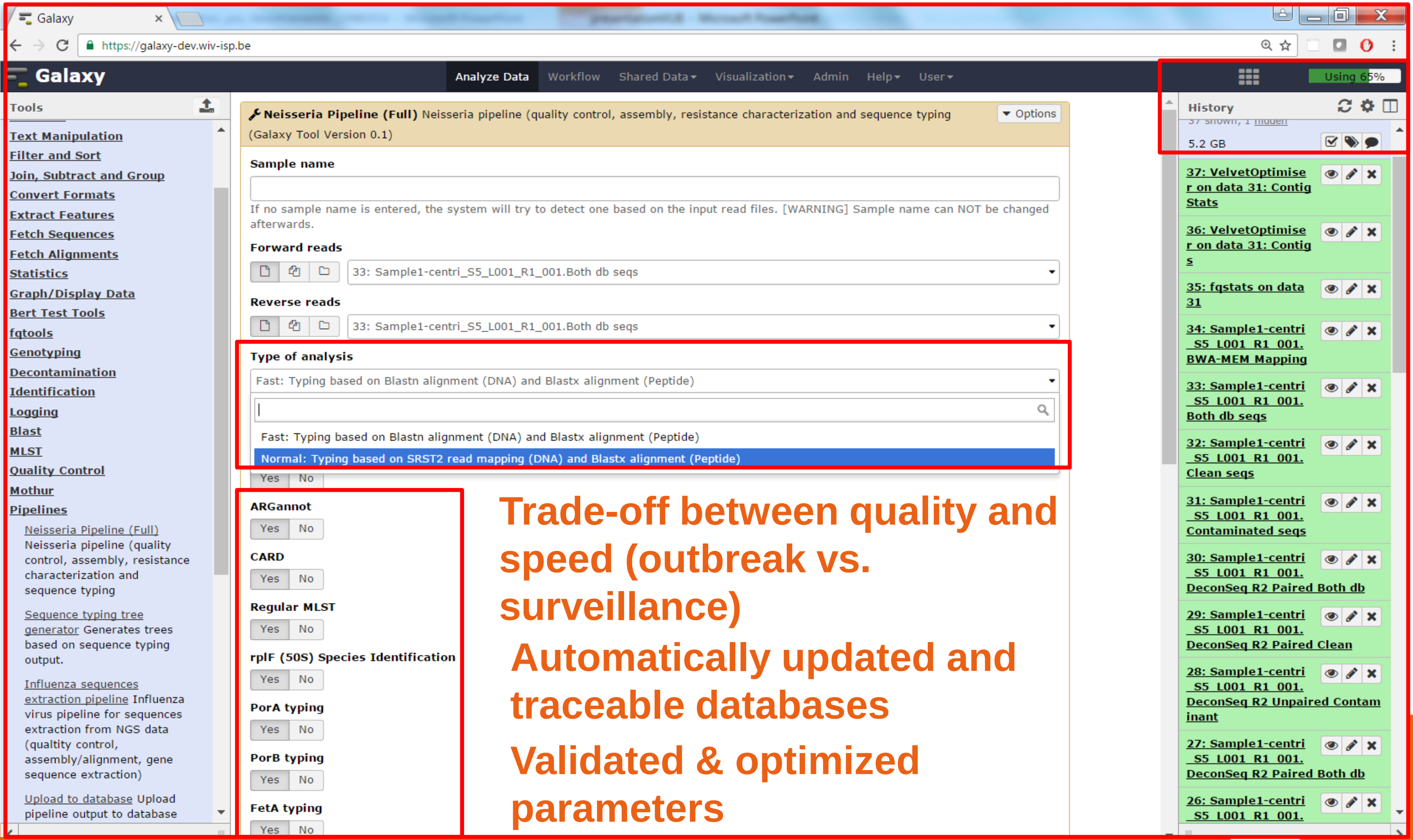
```
Executing command: 'module
load srst2/0.1.5 bowtie2/2.2.4
samtools/0.1.18; python
/usr/local/bin/srst2/0.1.5/scripts/
--input_pe reads_1 reads_2 --
mlst_delimiter '_' --mlst_db
/data/MLST/listeria/schemes/Liste
--mlst_defin
```

1	2	3	4	5	6	7	8
Sample	ST	abcZ	bglA	cat	dapE	dat	ldh

Optimized ‘push-on-the-button’ pipelines

User-friendly access

Centralized computational infrastructure



Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Statistics
Graph/Display Data
Bert Test Tools
fqtools
Genotyping
Decontamination
Identification
Logging
Blast
MLST
Quality Control
Mothur
Pipelines

Neisseria Pipeline (Full) Neisseria pipeline (quality control, assembly, resistance characterization and sequence typing

Sequence typing tree generator Generates trees based on sequence typing output.

Influenza sequences extraction pipeline Influenza virus pipeline for sequences extraction from NGS data (qualtity control, assembly/alignment, gene sequence extraction)

Upload to database Upload pipeline output to database

Neisseria Pipeline (Full) Neisseria pipeline (quality control, assembly, resistance characterization and sequence typing (Galaxy Tool Version 0.1) ▾ Options

Sample name

If no sample name is entered, the system will try to detect one based on the input read files. [WARNING] Sample name can NOT be changed afterwards.

Forward reads

33: Sample1-centri_S5_L001_R1_001.Both db seqs

Reverse reads

33: Sample1-centri_S5_L001_R1_001.Both db seqs

Type of analysis

Fast: Typing based on Blastn alignment (DNA) and Blastx alignment (Peptide)

Fast: Typing based on Blastn alignment (DNA) and Blastx alignment (Peptide)
Normal: Typing based on SRST2 read mapping (DNA) and Blastx alignment (Peptide)

Yes No

ARGannot
Yes No

CARD
Yes No

Regular MLST
Yes No

rplF (50S) Species Identification
Yes No

PorA typing
Yes No

PorB typing
Yes No

FetA typing
Yes No

History

37 shown, 1 hidden

5.2 GB

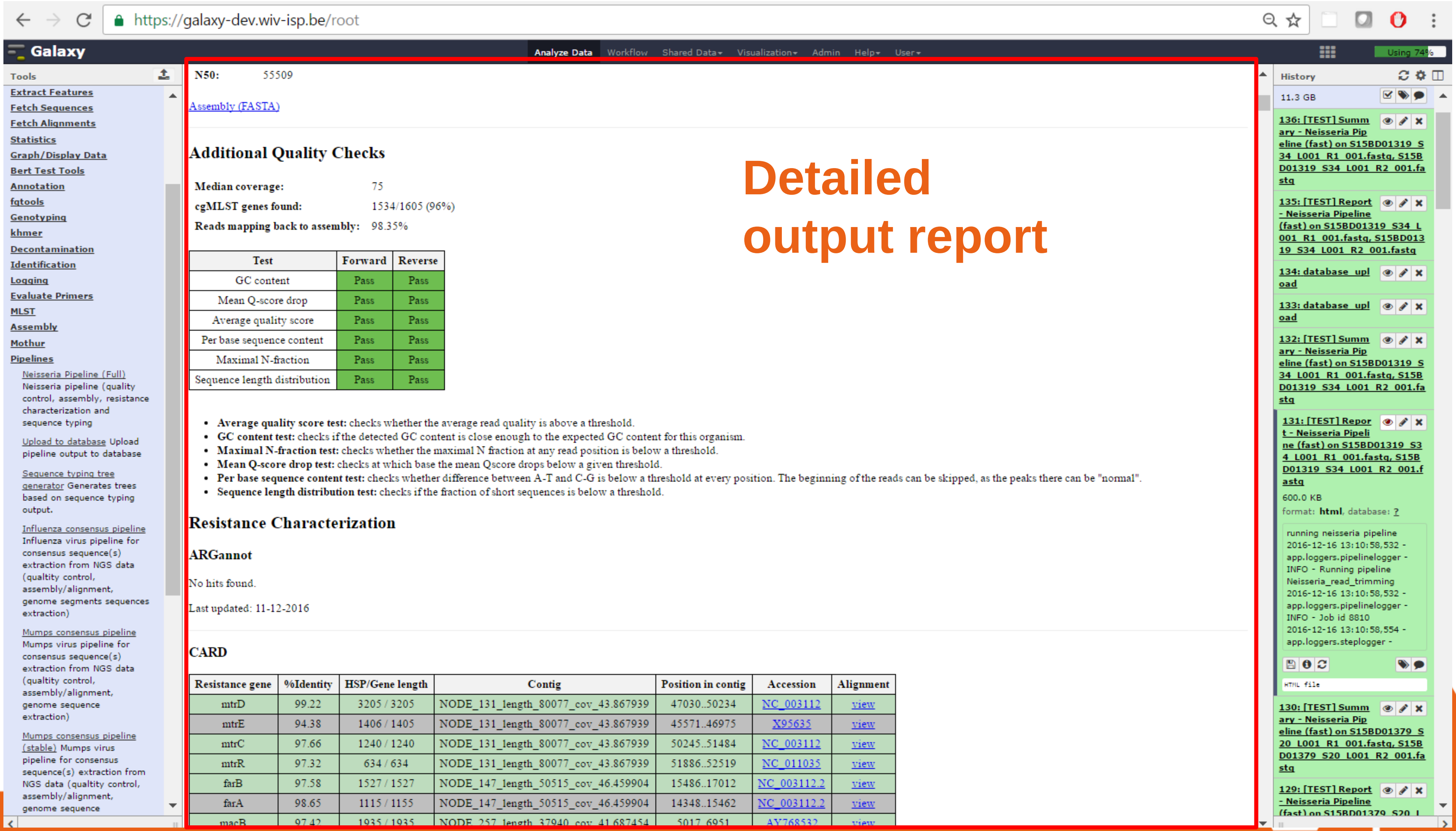
37: VelvetOptimiser on data 31: Contig Stats
36: VelvetOptimiser on data 31: Contigs
35: fqstats on data 31
34: Sample1-centri_S5_L001_R1_001. BWA-MEM Mapping
33: Sample1-centri_S5_L001_R1_001. Both db seqs
32: Sample1-centri_S5_L001_R1_001. Clean db seqs
31: Sample1-centri_S5_L001_R1_001. Contaminated seqs
30: Sample1-centri_S5_L001_R1_001. DeconSeq R2 Paired Both db
29: Sample1-centri_S5_L001_R1_001. DeconSeq R2 Paired Clean
28: Sample1-centri_S5_L001_R1_001. DeconSeq R2 Unpaired Contaminant
27: Sample1-centri_S5_L001_R1_001. DeconSeq R2 Paired Both db
26: Sample1-centri_S5_L001_R1_001.

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

Automatically updated and traceable databases

Validated & optimized parameters

Optimized 'push-on-the-button' pipelines



Neisseria Pipeline (Full) Neisseria pipeline (quality control, assembly, resistance characterization and sequence typing

Upload to database Upload pipeline output to database

Sequence typing tree generator Generates trees based on sequence typing output.

Influenza consensus pipeline Influenza virus pipeline for consensus sequence(s) extraction from NGS data (qualtity control, assembly/alignment, genome segments sequences extraction)

Mumps consensus pipeline Mumps virus pipeline for consensus sequence(s) extraction from NGS data (qualtity control, assembly/alignment, genome sequence extraction)

Mumps consensus pipeline (stable) Mumps virus pipeline for consensus sequence(s) extraction from NGS data (qualtity control, assembly/alignment, genome sequence

N50: 55509

Assembly (FASTA)

Additional Quality Checks

Median coverage:	75
cgMLST genes found:	1534/1605 (96%)
Reads mapping back to assembly:	98.35%

Test	Forward	Reverse
GC content	Pass	Pass
Mean Q-score drop	Pass	Pass
Average quality score	Pass	Pass
Per base sequence content	Pass	Pass
Maximal N-fraction	Pass	Pass
Sequence length distribution	Pass	Pass

- **Average quality score test:** checks whether the average read quality is above a threshold.
- **GC content test:** checks if the detected GC content is close enough to the expected GC content for this organism.
- **Maximal N-fraction test:** checks whether the maximal N fraction at any read position is below a threshold.
- **Mean Q-score drop test:** checks at which base the mean Qscore drops below a given threshold.
- **Per base sequence content test:** checks whether difference between A-T and C-G is below a threshold at every position. The beginning of the reads can be skipped, as the peaks there can be "normal".
- **Sequence length distribution test:** checks if the fraction of short sequences is below a threshold.

Resistance Characterization

ARGannot

No hits found.

Last updated: 11-12-2016

CARD

Resistance gene	%Identity	HSP/Gene length	Contig	Position in contig	Accession	Alignment
mtrD	99.22	3205 / 3205	NODE_131_length_80077_cov_43.867939	47030..50234	NC_003112	view
mtrE	94.38	1406 / 1405	NODE_131_length_80077_cov_43.867939	45571..46975	X95635	view
mtrC	97.66	1240 / 1240	NODE_131_length_80077_cov_43.867939	50245..51484	NC_003112	view
mtrR	97.32	634 / 634	NODE_131_length_80077_cov_43.867939	51886..52519	NC_011035	view
farB	97.58	1527 / 1527	NODE_147_length_50515_cov_46.459904	15486..17012	NC_003112.2	view
farA	98.65	1115 / 1155	NODE_147_length_50515_cov_46.459904	14348..15462	NC_003112.2	view
macB	97.42	1935 / 1935	NODE_257_length_37940_cov_41.687454	5017..6951	AY768532	view

136: [TEST] Summary - Neisseria Pipeline (fast) on S15BD01319_S34_L001_R1_001.fastq, S15BD01319_S34_L001_R2_001.fastq

135: [TEST] Report - Neisseria Pipeline (fast) on S15BD01319_S34_L001_R1_001.fastq, S15BD01319_S34_L001_R2_001.fastq

134: database_upload

133: database_upload

132: [TEST] Summary - Neisseria Pipeline (fast) on S15BD01319_S34_L001_R1_001.fastq, S15BD01319_S34_L001_R2_001.fastq

131: [TEST] Report - Neisseria Pipeline (fast) on S15BD01319_S34_L001_R1_001.fastq, S15BD01319_S34_L001_R2_001.fastq

600.0 KB

format: **html**, database: ?

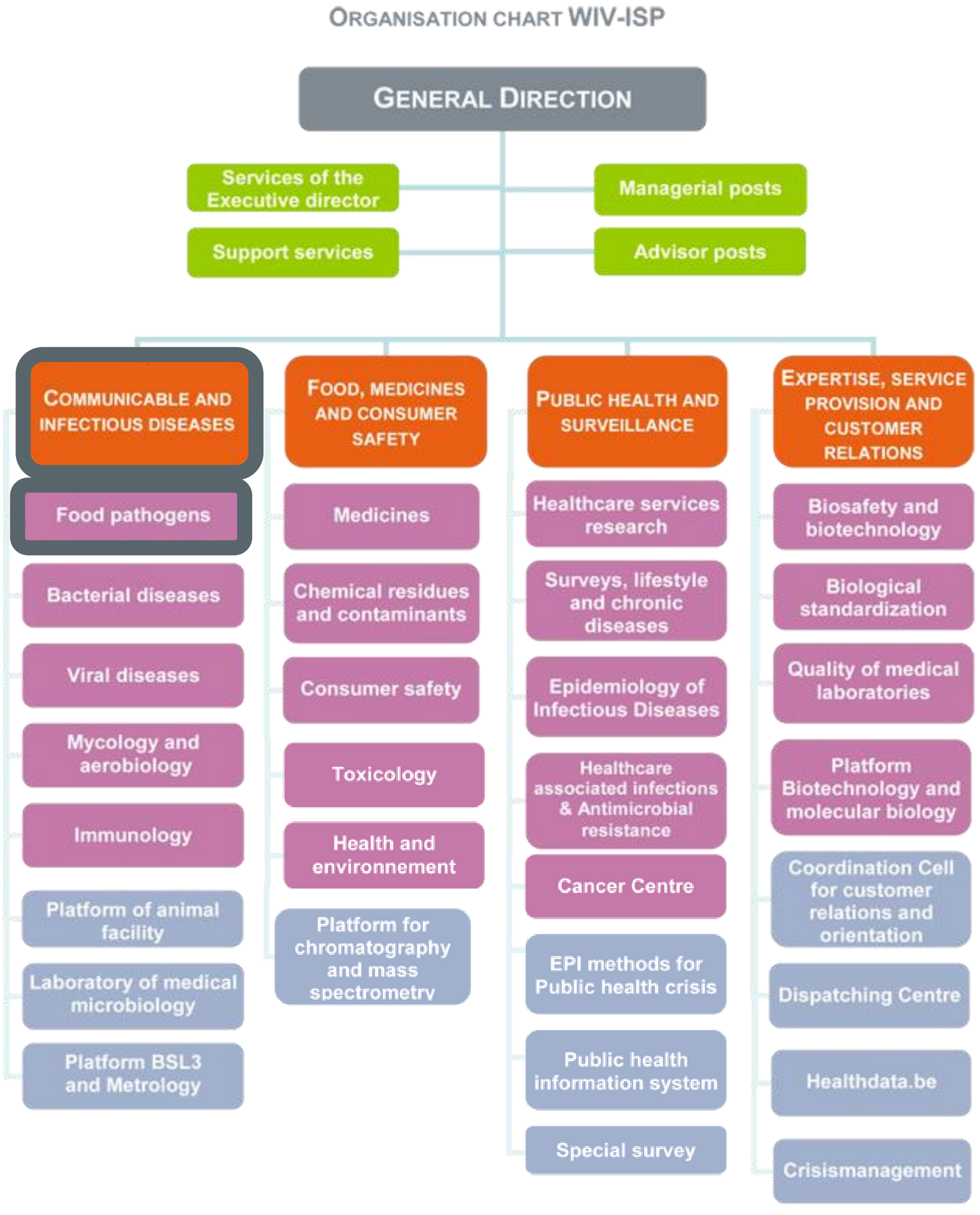
```
running neisseria pipeline
2016-12-16 13:10:58,532 -
app.loggers.pipelinelogger -
INFO - Running pipeline
Neisseria_read_trimming
2016-12-16 13:10:58,532 -
app.loggers.pipelinelogger -
INFO - Job id 8810
2016-12-16 13:10:58,554 -
app.loggers.steplogger -
```

HTML file

130: [TEST] Summary - Neisseria Pipeline (fast) on S15BD01379_S20_L001_R1_001.fastq, S15BD01379_S20_L001_R2_001.fastq

129: [TEST] Report - Neisseria Pipeline (fast) on S15BD01379_S20_L

Detailed output report


ORGANISATION CHART WIV-ISP

GENERAL DIRECTION

Services of the Executive director

Managerial posts

Support services

Advisor posts

COMMUNICABLE AND INFECTIOUS DISEASES
- Food pathogens
- Bacterial diseases
- Viral diseases
- Mycology and aerobiology
- Immunology
- Platform of animal facility
- Laboratory of medical microbiology
- Platform BSL3 and Metrology

FOOD, MEDICINES AND CONSUMER SAFETY
- Medicines
- Chemical residues and contaminants
- Consumer safety
- Toxicology
- Health and environnement
- Platform for chromatography and mass spectrometry

PUBLIC HEALTH AND SURVEILLANCE
- Healthcare services research
- Surveys, lifestyle and chronic diseases
- Epidemiology of Infectious Diseases
- Healthcare associated infections & Antimicrobial resistance
- Cancer Centre
- EPI methods for Public health crisis
- Public health information system
- Special survey

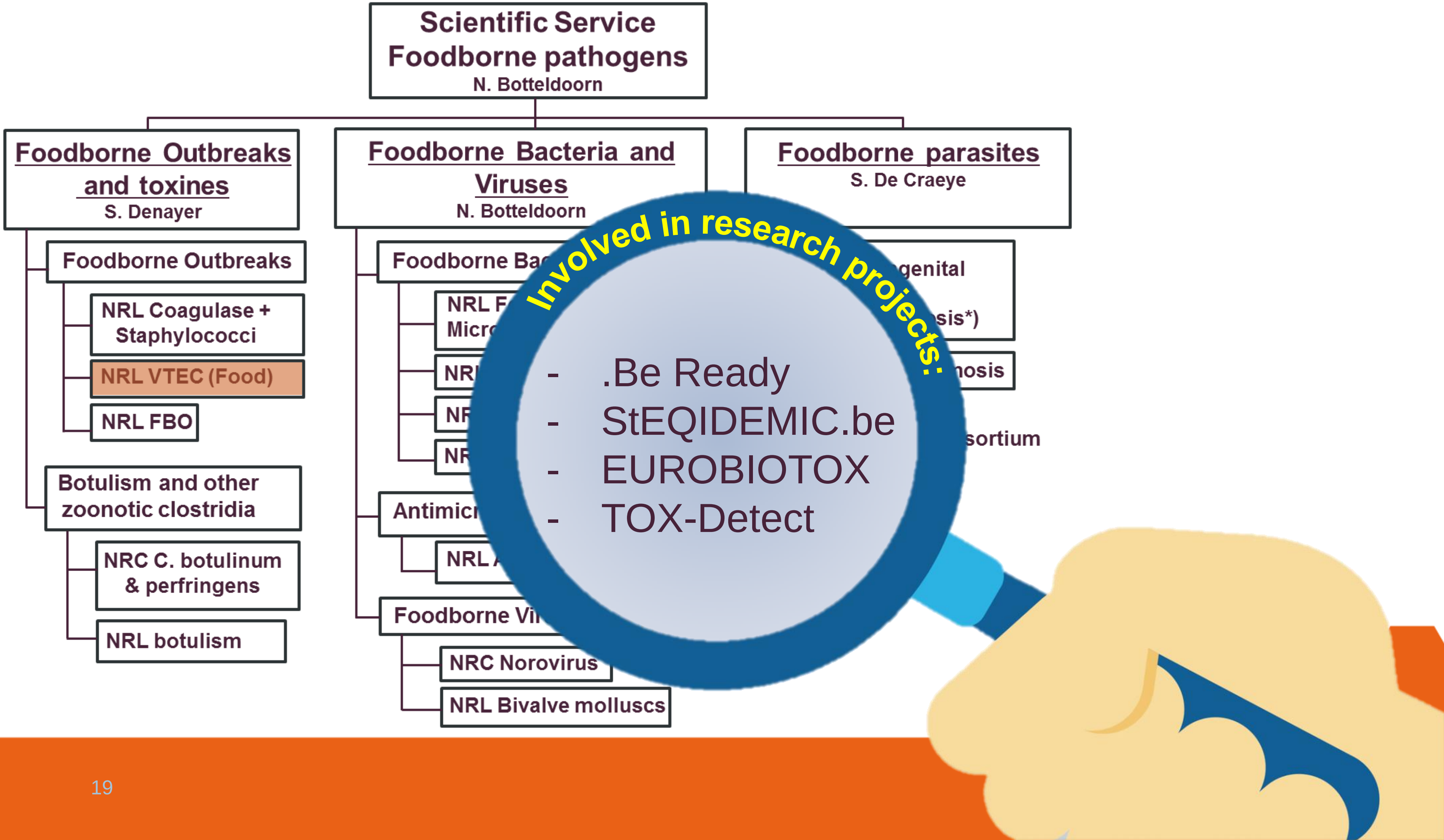
EXPERTISE, SERVICE PROVISION AND CUSTOMER RELATIONS
- Biosafety and biotechnology
- Biological standardization
- Quality of medical laboratories
- Platform Biotechnology and molecular biology
- Coordination Cell for customer relations and orientation
- Dispatching Centre
- Healthdata.be
- Crisismanagement

NRL STEC - Belgium

Department communicable and infectious diseases

Food pathogens

NRL STEC - Belgium



Scientific Service Foodborne pathogens
N. Botteldoorn

Foodborne Outbreaks and toxines
S. Denayer

Foodborne Outbreaks

NRL Coagulase + Staphylococci

NRL VTEC (Food)

NRL FBO

Botulism and other zoonotic clostridia

NRC C. botulinum & perfringens

NRL botulism

Foodborne Bacteria and Viruses
N. Botteldoorn

Foodborne Ba

NRL F
Micr

NRI

NF

NF

Antimicr

NRL A

Foodborne Vi

NRC Norovirus

NRL Bivalve molluscs

Foodborne parasites
S. De Craeye

genital
sis*)

nosis

sortium

Involved in research projects:

- .Be Ready
- StEQIDEMIC.be
- EUROBIOTOX
- TOX-Detect

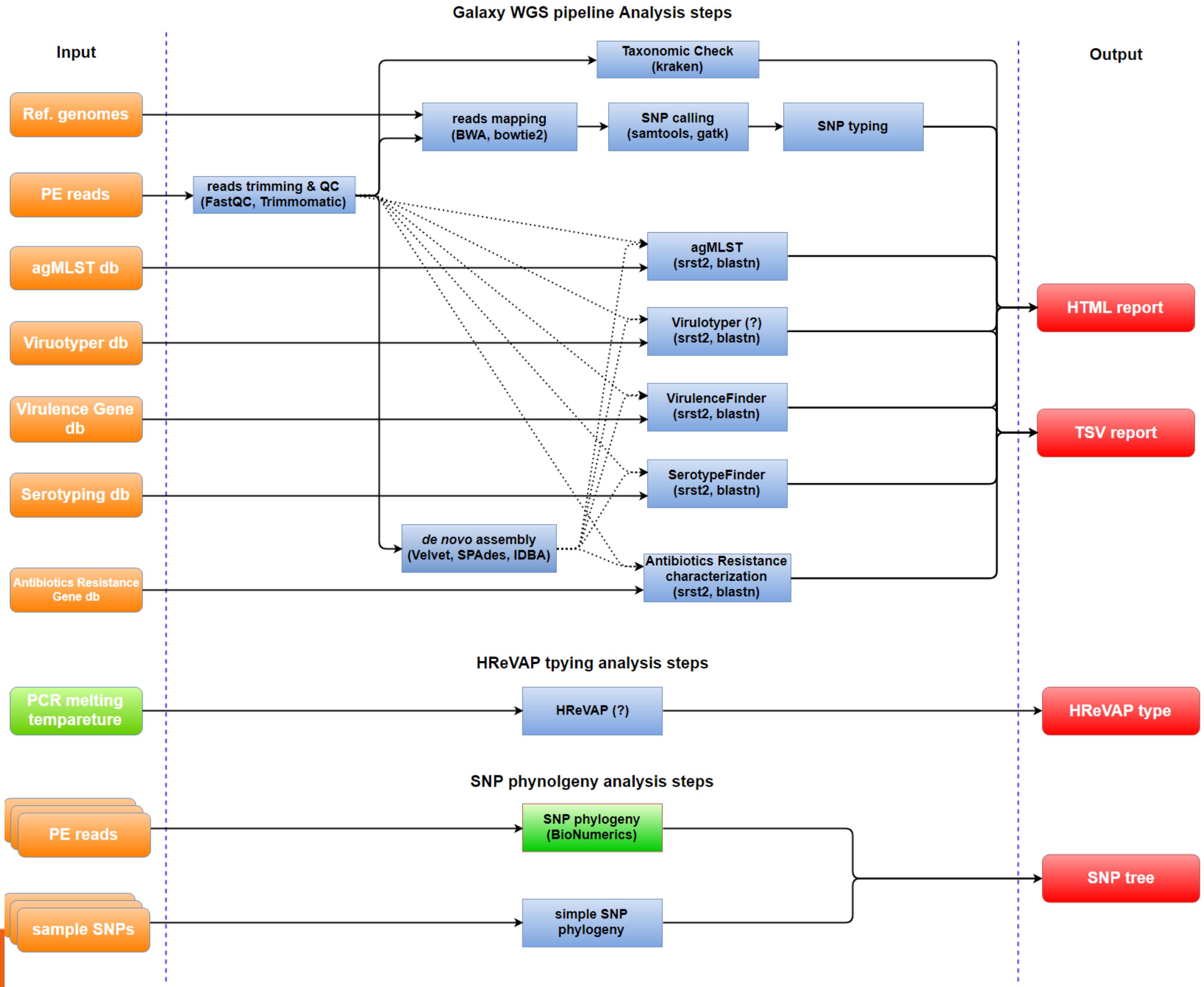
NRL STEC - Belgium

E-mail: nrlvti-lnrtia@wiv-isp.be

Intake BIOIT needs NRL STEC

- Requirement for a routine pipeline for the analysis of WGS data generated on VTEC and other *E. coli* isolates
- Pipeline should be ‘push-on-the-button’
- All functionality should also be available as stand-alone tools
- **Pipeline should be streamlined with the functionality made available by the EURL-VTEC in their ARIES platform**

Analysis BIOIT needs NRL STEC


Galaxy WGS pipeline Analysis steps
Input
Output
Ref. genomes
PE reads
agMLST db
Viruotyper db
Virulence Gene db
Serotyping db
Antibiotics Resistance Gene db
reads trimming & QC (FastQC, Trimmomatic)
Taxonomic Check (kraken)
reads mapping (BWA, bowtie2)
SNP calling (samtools, gatk)
SNP typing
agMLST (srst2, blastn)
Virulotyper (?) (srst2, blastn)
VirulenceFinder (srst2, blastn)
SerotypeFinder (srst2, blastn)
Antibiotics Resistance characterization (srst2, blastn)
de novo assembly (Velvet, SPAdes, IDBA)
HTML report
TSV report
HReVAP tpying analysis steps
PCR melting tempareture
HReVAP (?)
HReVAP type
SNP phynolgeny analysis steps
PE reads
sample SNPs
SNP phylogeny (BioNumerics)
simple SNP phylogeny
SNP tree

Data processing

- Quality control:
  - FastQC for generation of raw read reports
  - Trimmomatic for trimming of raw reads
  - FastQC for generation of trimmed read reports
- Contamination check
  - Kraken for identification/confirmation of species based on *kmer* counting, using the entire NCBI RefSeq Microbial database (updated automatically)
- *De novo assembly*
  - SPAdes or Velvet(Optimiser) for assembly of trimmed reads
  - QUAST for quality control of assembly
- ‘Advanced’ quality control
  - A series of custom checks to ensure adequate quality for functional interpretation (%cgMLST genes found, median coverage, N-content…)
  - Provide 3 outcomes: ‘pass’ (all checks passed), ‘warning’ (questionable quality but OK for interpretation), and ‘fail’ (sample should be re-sequenced)

Sequence typing

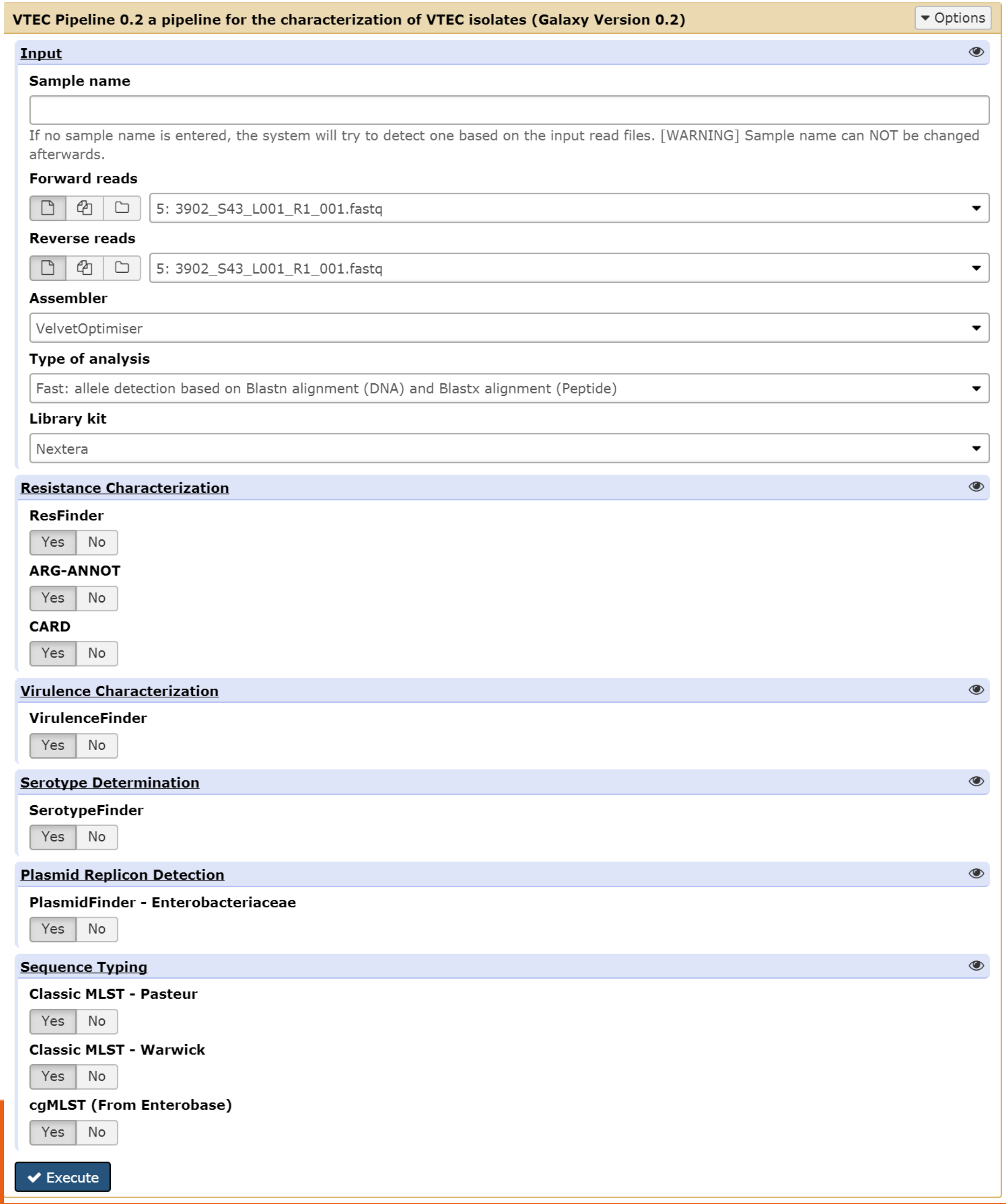
- MLST, cgMLST, wgMLST, agMLST
  - SRST2 for direct read mapping or BLAST+ for checking assembly
  - MLST (Pasteur/Warwick), cg/wgMLST (Enterobase), agMLST (ARIES) - updated automatically
- Serotyping
  - SRST2 for direct read mapping or BLAST+ for checking assembly
  - SerotypeFinder (DTU/CGE) - updated automatically
- Virulence typing
  - SRST2 for direct read mapping or BLAST+ for checking assembly
  - VirulenceFinder (DTU/CGE) - updated automatically
- Antibiotics resistance
  - SRST2 for direct read mapping or BLAST+ for checking assembly
  - ResFinder (DTU/CGE), CARD (McMaster University), ARG-annot (University of Marseille) - updated automatically
- Plasmid typing
  - SRST2 for direct read mapping or BLAST+ for checking assembly
  - PlasmidFinder(DTU/CGE) - updated automatically

SNP typing/phylogeny

- SNP typing
  - SnapperDB for SNP typing
  - Italian/UK/Belgian database(s) for typing (?)
- SNP phylogeny
  - PHEnix pipeline, CFSAN pipeline, in-house implementation using Samtools for determining phylogeny based on SNPs
  - Output that consists out of a newick tree file and basic visualization of the resulting phylogenetic tree
  - Due to the nature of the these tools, they cannot be integrated in the ‘push-on-the-button’ pipeline

Others

- HReVAP (ARIES)
- Community requirements?



VTEC Pipeline 0.2 a pipeline for the characterization of VTEC isolates (Galaxy Version 0.2)

Input

Sample name

If no sample name is entered, the system will try to detect one based on the input read files. [WARNING] Sample name can NOT be changed afterwards.

Forward reads

5: 3902_S43_L001_R1_001.fastq

Reverse reads

5: 3902_S43_L001_R1_001.fastq

Assembler

VelvetOptimiser

Type of analysis

Fast: allele detection based on Blastn alignment (DNA) and Blastx alignment (Peptide)

Library kit

Nextera

Resistance Characterization

ResFinder

[✓ Yes] No

ARG-ANNOT

[✓ Yes] No

CARD

[✓ Yes] No

Virulence Characterization

VirulenceFinder

[✓ Yes] No

Serotype Determination

SerotypeFinder

[✓ Yes] No

Plasmid Replicon Detection

PlasmidFinder - Enterobacteriaceae

[✓ Yes] No

Sequence Typing

Classic MLST - Pasteur

[✓ Yes] No

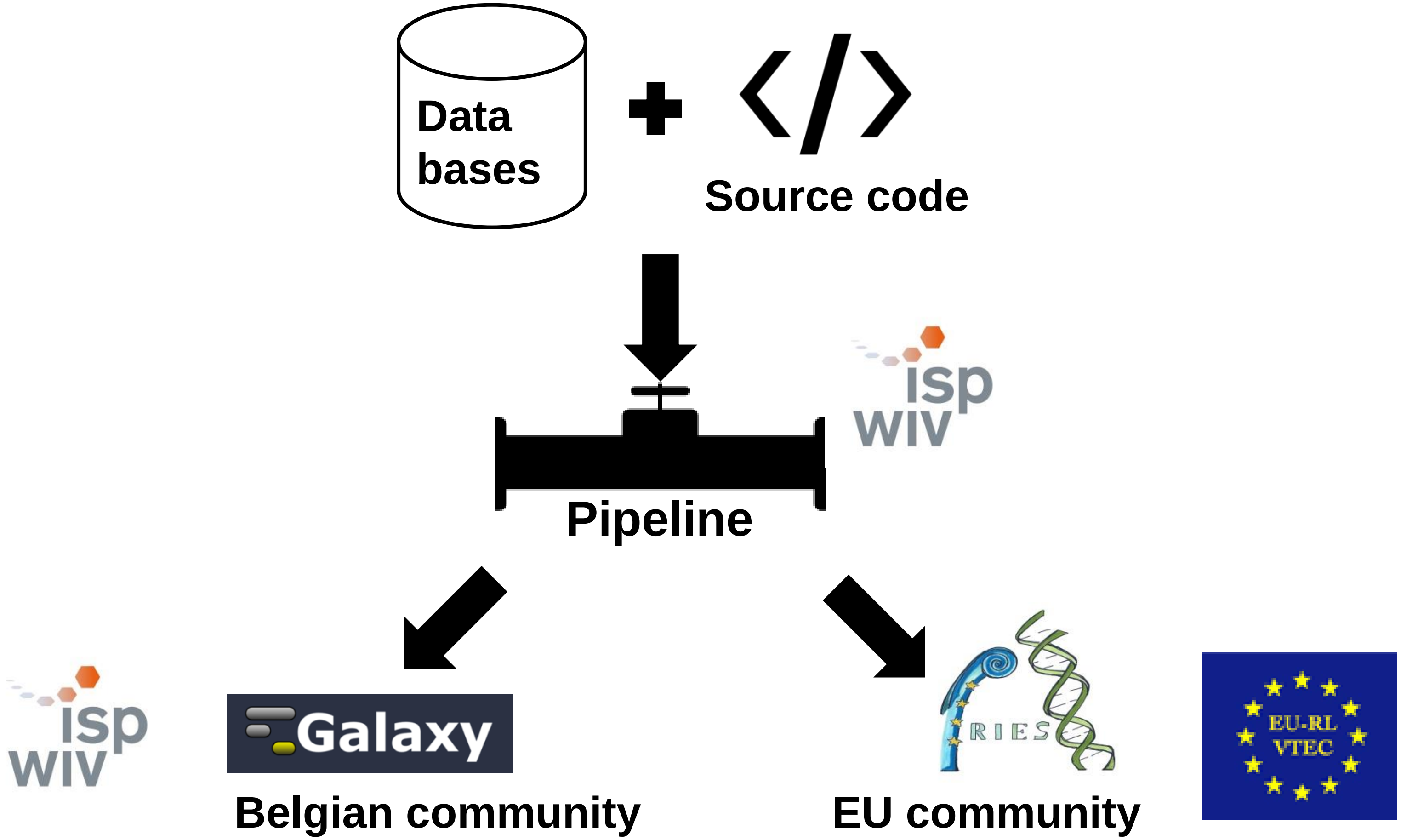
Classic MLST - Warwick

[✓ Yes] No

cgMLST (From Enterobase)

[✓ Yes] No

Collaboration ISS and WIV-ISP



Data bases

Source code

Pipeline

Belgian community

EU community

Thank you for your attention!
Questions?