Untangling bioinformatics bias for the diagnostic of plant viruses by HTS:

Lessons from an international performance testing of sequence analysis strategies

Sébastien Massart, Laboratory of Plant Pathology,
Gembloux Agro-Bio Tech, Liège University, Belgium
sebastien.massart@uliege.be

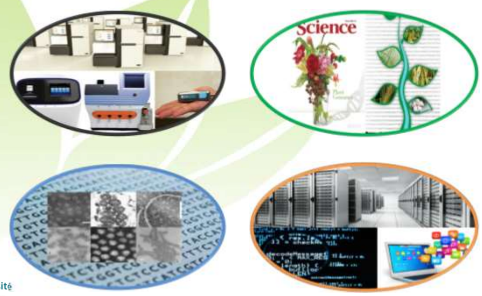
& Michela Chiumenti, Kris De Jonghe, Rachel Glover, Annelies Haegeman, Igor Koloniuk, Petr Komınek, Jan Kreuze, Denis Kutnjak, Leonidas Lotos, Francois Maclot, Varvara Maliogka, Hans J. Maree, Thibaut Olivier, Antonio Olmos, Mikhail M. Pooggin, Jean-Sebastien Reynard, Ana B. Ruiz-Garcıa, Dana Safarova, Pierre H. H. Schneeberger, Noa Sela, Silvia Turco, Eeva J. Vainio, Eva Varallyay, Eric Verdin, Marcel Westenberg, Yves Brostaux, and Thierry Candresse

LIÈGE université Gembloux Agro-Bio Tech

Standardisation of HTS is needed

LIÈGE université Gembloux Agro-Bio Tech

What's happening with HTS ?



LIÈGE université Gembloux Agro-Bio Tech

Objective: evaluation of bio-informatic strategies

Evaluate and compare the performance of bioinformatic pipelines for the analysis of high-throughput siRNA sequencing data

DIVAS COST Action FA 1407

cost EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

LIÈGE université Gembloux Agro-Bio Tech

HTS in Research ?

Starting datasets: FASTQ files from 3 samples

Species	Potato		Apple	Grapevine									
Sequencing technology	Illumina												
Sequencing depth	23M		13M	12M									
Plant	Potato		Apple	Grapevine rootstock Kober 125 aa, Vitis berlandieri x Vitis riparia									
Viruses or viroids detected (one column per virus)	PVX	New Nepovirus	ASGV (2x genetic variants)	Grapevine leafroll-associated virus 1	Grapevine virus A	Grapevine virus B	Grapevine rupestris stem pitting-associated virus	Grapevine rupestris vein feathering virus	Grapevine Syrah virus 1	Grapevine red globe virus	Hop stunt viroid	Grapevine yellow speckle viroid 1
Percentage of reads for each virus/viroid (*)	~23%	~5%	0.12%	5.7%	0.6%	1.2%	0.4%	0.3%			0,5%	0,14%
Confirmation(s) technique(s) for each virus/viroid	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	ELISA, RT-PCR	ELISA, RT-PCR	ELISA, RT-PCR	RT-PCR	not yet verified = not tested	RT-PCR	negative in RT-PCR	RT-PCR	RT-PCR

Difficulties:

1. New unknown virus

LIÈGE université Gembloux Agro-Bio Tech

Starting datasets: FASTQ files from 3 samples

Species	Potato		Apple	Grapevine								
Sequencing technology	Illumina											
Sequencing depth	23M		13M	12M								
Plant	Potato		Apple	Grapevine rootstock Kober 125 aa, Vitis berlandieri x Vitis riparia								
Viruses or viroids detected (one column per virus)	PVX	New Nepovirus	ASGV (2x genetic variants)	Grapevine fabavirus-associated virus 1	Grapevine virus A	Grapevine virus B	Grapevine rupestris stem pitting-associated virus	Grapevine Syrah virus 1	Grapevine red globe virus	Hop stunt viroid	Grapevine yellow speckle viroid 1
Percentage of reads for each virus/viroid (*)	~23%	~5%	0.12%	5.7%	0.6%	1.2%	0.4%	0.3%		0,5%	0,14%
Confirmation(s) technique(s) for each virus/viroid	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	ELISA, RT-PCR	ELISA, RT-PCR	ELISA, RT-PCR	RT-PCR	not yet verified - not tested	RT-PCR	negative in RT-PCR	RT-PCR	RT-PCR

Difficulties:

1. New unknown virus
2. Complex mix of 9 viruses/viroids

LIÈGE université Gembloux Agro-Bio Tech

Methodology for evaluation of bioinformatic pipelines

- sRNA datasets (lenght between 21 and 24 nt)
- Worst case scenario by rarefaction at 3 sequencing depths:
  - 50,000 reads – 3 files
  - 250,000 reads– 4 files (2 grapevine replicates)
  - 2,500,000 reads– 3 files
- 10 fastq files available on a server in double blind
- 21 participants

LIÈGE université Gembloux Agro-Bio Tech

Starting datasets: FASTQ files from 3 samples

Species	Potato		Apple	Grapevine								
Sequencing technology	Illumina											
Sequencing depth	23M		13M	12M								
Plant	Potato		Apple	Grapevine rootstock Kober 125 aa, Vitis berlandieri x Vitis riparia								
Viruses or viroids detected (one column per virus)	PVX	New Nepovirus	ASGV (2x genetic variants)	Grapevine fabavirus-associated virus 1	Grapevine virus A	Grapevine virus B	Grapevine rupestris stem pitting-associated virus	Grapevine Syrah virus 1	Grapevine red globe virus	Hop stunt viroid	Grapevine yellow speckle viroid 1
Percentage of reads for each virus/viroid (*)	~23%	~5%	0.12%	5.7%	0.6%	1.2%	0.4%	0.3%		0,5%	0,14%
Confirmation(s) technique(s) for each virus/viroid	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	ELISA, RT-PCR	ELISA, RT-PCR	ELISA, RT-PCR	RT-PCR	not yet verified - not tested	RT-PCR	negative in RT-PCR	RT-PCR	RT-PCR

Difficulties:

1. New unknown virus
2. Complex mix of 9 viruses/viroids
3. **Very low abundance virus/viroid**

LIÈGE université Gembloux Agro-Bio Tech

Methodology for evaluation of bioinformatic pipelines

Participants free to apply their own bioinformatics strategy to identify the viruses in the 10 datasets (de novo assembly + annotation)

Interpret the data in a diagnostics setting

LIÈGE université Gembloux Agro-Bio Tech

Starting datasets: FASTQ files from 3 samples

Species	Potato		Apple	Grapevine								
Sequencing technology	Illumina											
Sequencing depth	23M		13M	12M								
Plant	Potato		Apple	Grapevine rootstock Kober 125 aa, Vitis berlandieri x Vitis riparia								
Viruses or viroids detected (one column per virus)	PVX	New Nepovirus	ASGV (2x genetic variants)	Grapevine fabavirus-associated virus 1	Grapevine virus A	Grapevine virus B	Grapevine rupestris stem pitting-associated virus	Grapevine Syrah virus 1	Grapevine red globe virus	Hop stunt viroid	Grapevine yellow speckle viroid 1
Percentage of reads for each virus/viroid (*)	~23%	~5%	0.12%	5.7%	0.6%	1.2%	0.4%	0.3%		0,5%	0,14%
Confirmation(s) technique(s) for each virus/viroid	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	RT-PCR & Sanger sequencing	ELISA, RT-PCR	ELISA, RT-PCR	ELISA, RT-PCR	RT-PCR	not yet verified - not tested	RT-PCR	negative in RT-PCR	RT-PCR	RT-PCR

Difficulties:

1. New unknown virus
2. Complex mix of 9 viruses/viroids
3. **Very low abundance virus/viroid**
4. Close virus species with poor genome knowledge

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity

- 70 % sensitivity overall
- Sensitivity from 35 to 100%
- Sensitivity increase with sequencing depth

	Sensitivity			
labID	50,000	250,000	2,500,000	Average
A	10%	53%	90%	51%
B	30%	35%	80%	46%
C	60%	71%	80%	70%
D	50%	82%	100%	78%
E	30%	82%	80%	68%
F	80%	88%	100%	89%
G	20%	53%	100%	57%
H	30%	65%	70%	57%
J	70%	94%	100%	89%
K	40%	71%	90%	68%
M	50%	94%	90%	81%
N	30%	82%	90%	70%
O	20%	41%	40%	35%
P	20%	59%	70%	51%
R	100%	100%	100%	100%
S	50%	100%	100%	86%
T	90%	100%	100%	97%
V	60%	88%	80%	78%
W1	40%	82%	90%	73%
W2	60%	82%	90%	78%
X	30%	71%	80%	62%
AVERAGE	46%	75%	86%	70%

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity

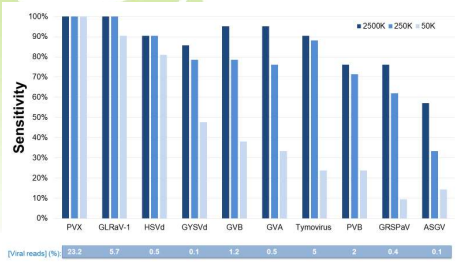
- 70 % sensitivity overall
- Sensitivity from 35 to 100%
- Sensitivity increase with sequencing depth
- Seven strategies have 100% sensitivity at 2.5 M
- Participant R: 100% sensitivity (T close)

-> Both have an additionnal read mapping step

labID	Sensitivity 50,000	250,000	2,500,000	Average
A	10%	53%	90%	51%
B	30%	35%	80%	46%
C	60%	71%	80%	70%
D	50%	82%	100%	78%
E	30%	82%	80%	68%
F	80%	88%	100%	89%
G	20%	53%	100%	57%
H	30%	65%	70%	57%
J	70%	94%	100%	89%
K	40%	71%	90%	68%
M	50%	94%	90%	81%
N	30%	82%	90%	70%
O	20%	41%	40%	35%
P	20%	59%	70%	51%
R	100%	100%	100%	100%
S	50%	100%	100%	86%
T	90%	100%	100%	97%
V	60%	88%	80%	78%
W1	40%	82%	90%	73%
W2	60%	82%	90%	78%
X	30%	71%	80%	62%
AVERAGE	46%	75%	86%	70%

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity per virus



LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity

- Sensitivity improved by lower k-mer lenghts (13-15), decreased by higher lenght (17-21)

-> impact of k-mer lenght

- Sensitivity reduced by setting minimal contig length too high, a value <60nt should be preferred

-> Shorter minimal contig lenght prefered

labID	Sensitivity 50,000	250,000	2,500,000	Average
A	10%	53%	90%	51%
B	30%	35%	80%	46%
C	60%	71%	80%	70%
D	50%	82%	100%	78%
E	30%	82%	80%	68%
F	80%	88%	100%	89%
G	20%	53%	100%	57%
H	30%	65%	70%	57%
J	70%	94%	100%	89%
K	40%	71%	90%	68%
M	50%	94%	90%	81%
N	30%	82%	90%	70%
O	20%	41%	40%	35%
P	20%	59%	70%	51%
R	100%	100%	100%	100%
S	50%	100%	100%	86%
T	90%	100%	100%	97%
V	60%	88%	80%	78%
W1	40%	82%	90%	73%
W2	60%	82%	90%	78%
X	30%	71%	80%	62%
AVERAGE	46%	75%	86%	70%

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity per virus

- Correlation between **percentage of viral reads** in sample and detectability of a given agent
- Other parameters have an influence

1. Small genome size of viroids: better genome coverage and detectability

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity

- Sensitivity improved by contig extension after de novo analysis

-> Contig extension to pass minimal contig lenght threshold

labID	Sensitivity 50,000	250,000	2,500,000	Average
A	10%	53%	90%	51%
B	30%	35%	80%	46%
C	60%	71%	80%	70%
D	50%	82%	100%	78%
E	30%	82%	80%	68%
F	80%	88%	100%	89%
G	20%	53%	100%	57%
H	30%	65%	70%	57%
J	70%	94%	100%	89%
K	40%	71%	90%	68%
M	50%	94%	90%	81%
N	30%	82%	90%	70%
O	20%	41%	40%	35%
P	20%	59%	70%	51%
R	100%	100%	100%	100%
S	50%	100%	100%	86%
T	90%	100%	100%	97%
V	60%	88%	80%	78%
W1	40%	82%	90%	73%
W2	60%	82%	90%	78%
X	30%	71%	80%	62%
AVERAGE	46%	75%	86%	70%

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity per virus

- Correlation between **percentage of viral reads** in sample and detectability of a given agent
- Other parameters have an influence

2. Novel agent absent from database (Nepovirus in potato): only some contig may give significant Blast scores (**avoid blastn only, threshold important**)

LIÈGE université Gembloux Agro-Bio Tech

Results: analytical sensitivity per virus

- Correlation between **percentage of viral reads** in sample and detectability of a given agent
- Other parameters have an influence

3. Minimal contig length must be low to avoid elimination of small viral contigs from low abundance viruses (ASGV for apple)

LIÈGE université Gembloux Agro-Bio Tech

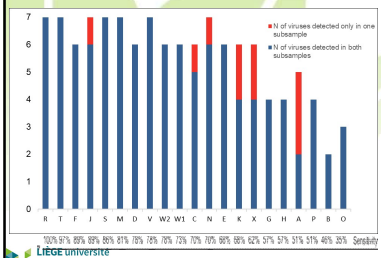
Conclusions: bioinformatics matters !

- ✓ Some rather trivial things to avoid
  - ✓ BlastN only
  - ✓ Stringent Blast cut-off values
  - ✓ MegaBlast (high homology threshold)

LIÈGE université Gembloux Agro-Bio Tech

Results: reproducibility

- Comparing the results reported for the two 250K grapevine pseudo-replicates



Key observations:

- Overall: **91,6% reproducibility**
- 15 strategies: 100% (including FN)
- 5 strategies: 100% reproducibility and analytical sensitivity
- Correlation between sensitivity & reproducibility

LIÈGE université Gembloux Agro-Bio Tech

Conclusions: bioinformatics matters !

- ✓ Some rather trivial things to avoid
  - ✓ BlastN only
  - ✓ High Blast cut-off values
  - ✓ MegaBlast (high homology threshold)
- ✓ Some less obvious lessons
  - ✓ Minimal contig length must be very low (<60nt) for optimal sensitivity with low abundance viruses
  - ✓ For Velvet, broad range of k-mer values including higher ones improve sensitivity
  - ✓ Additional contigs extension and combination of mapping and de novo assembly might improve detection

LIÈGE université Gembloux Agro-Bio Tech

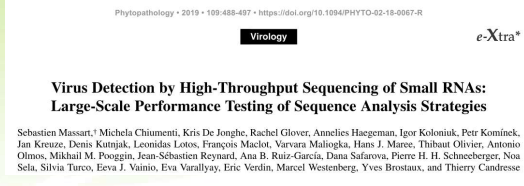
Conclusions: bioinformatics matters !

- ✓ A huge diversity in pipelines and pipeline performance
- ✓ Identification of trivial errors to avoid and other lessons

➢ **siRNA indexing data annotation is not such a straightforward business**

LIÈGE université Gembloux Agro-Bio Tech

For more details: see our publication in Phytopathology


Phytopathology • 2019 • 109:488-497 • https://doi.org/10.1094/PHYTO-02-18-0067-R

Virology

e-Xtra*

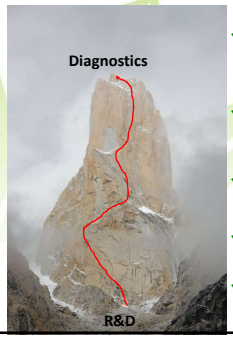
Virus Detection by High-Throughput Sequencing of Small RNAs: Large-Scale Performance Testing of Sequence Analysis Strategies

Sebastien Massart,† Michela Chiumenti, Kris De Jonghe, Rachel Glover, Annelies Haegeman, Igor Koloniuk, Petr Komínek, Jan Kreuze, Denis Kutnjak, Leonidas Lotos, François Maclot, Varvara Maliogka, Hans J. Maree, Thibaut Olivier, Antonio Olmos, Mikhail M. Pooggin, Jean-Sébastien Reynard, Ana B. Ruiz-García, Dana Safarova, Pierre H. H. Schneeberger, Noa Sela, Silvia Turco, Eeva J. Vainio, Eva Varallyay, Eric Verdin, Marcel Westenberg, Yves Brostaux, and Thierry Candresse

LIÈGE université Gembloux Agro-Bio Tech

One step further: applying HTS in plant pest diagnostics

www.valitest.eu



- ✓ Climbing the diagnostic cliff of HTS technologies ?
- ✓ Writing of international guidelines for the use of HTS technologies for plant pest diagnostics
- ✓ 54 scientists & >30 countries
- ✓ For viruses, bacteria, fungi, nematodes and insects
- ✓ 70 pages soon available as scientific paper

LIÈGE université Gembloux Agro-Bio Tech

Thanks for the invitation

Thank you for your attention

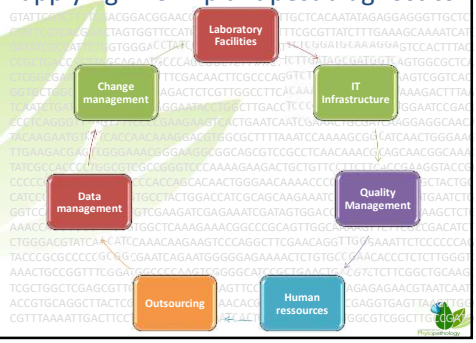
For more information:
sebastien.massart@uliege.be
: @Be_Phytopath

LIÈGE université Gembloux Agro-Bio Tech

One step further: applying HTS in plant pest diagnostics

www.valitest.eu

1. Preparing the laboratory

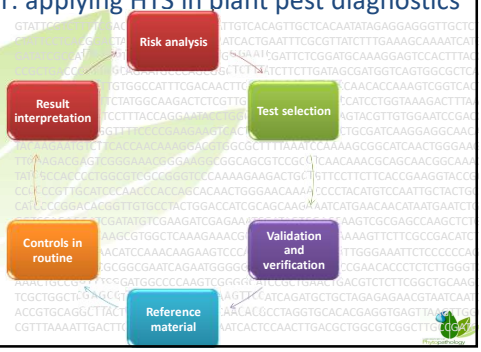


LIÈGE université Gembloux Agro-Bio Tech

One step further: applying HTS in plant pest diagnostics

www.valitest.eu

2. HTS test for diagnostics



LIÈGE université Gembloux Agro-Bio Tech