DIVAS
COST Action FA 1407

High Throughput Sequencing of siRNAs and virus diagnostic: do sequence analysis strategies really matter? Results of an international proficiency testing

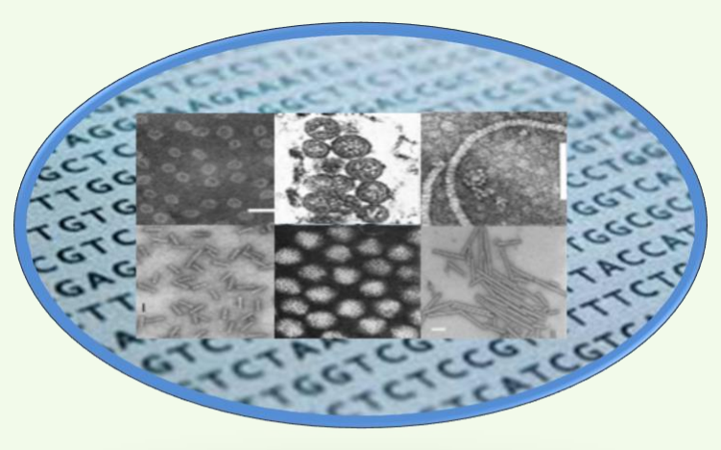
Sebastien Massart[1], Kris De Jonghe[2], Ian Adams, Annalisa Giampetruzzi, Igor Koloniuk, Petr Kominek, Jan Kreuze, Denis Kutnjak, Leonidas Lotos, Hans J. Maree, Thibaut Olivier, Mikhail Pooggin, Ana B. Ruiz-García, Dana Safarova, Pierre H. H. Schneeberger, Noa Sela, Eva Varallyay, Eeva Vainio, Eric Verdin, Marcel Westenberg, Yves Brostaux and Thierry Candresse[a]

[1]: Corresponding author: sebastien.massart@ulg.ac.be, Laboratory of Phytopathology, University of Liège, Gembloux Agro-BioTech, Gembloux, Belgium.
[2]: Presenting author: kris.dejonghe@ilvo.vlaanderen.be, Laboratory of Virology, Plant Sciences Unit, Flanders Research Institute for Agriculture, Fisheries and Food (ILVO)
[a]: the affiliations and addresses of the contributors : http://www.cost.eu/COST_Actions/fa/Actions/FA1407?management

Context

Cost and reliability of NGS technologies

+


Database growth

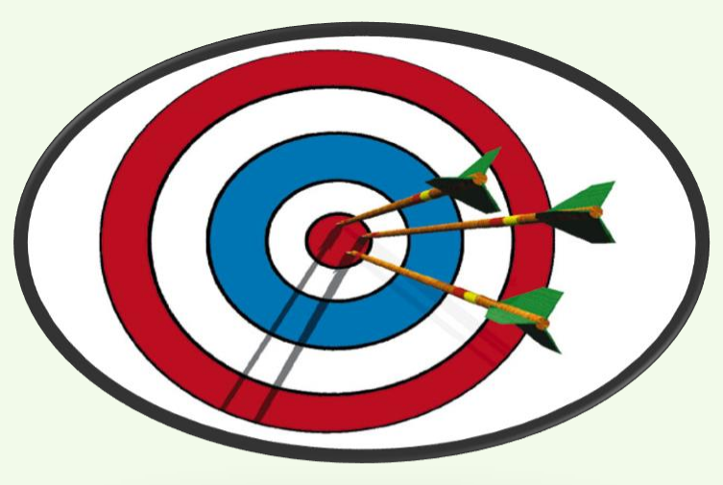
+

Bioinformatic simplification

= Opportunity to implement NGS in diagnostics

Sensitivity/specficty

+


Repeatability

+

Quality assurance

= Criteria for diagnostic use

Does NGS fit criteria for plant virus diagnostic ?

Objective

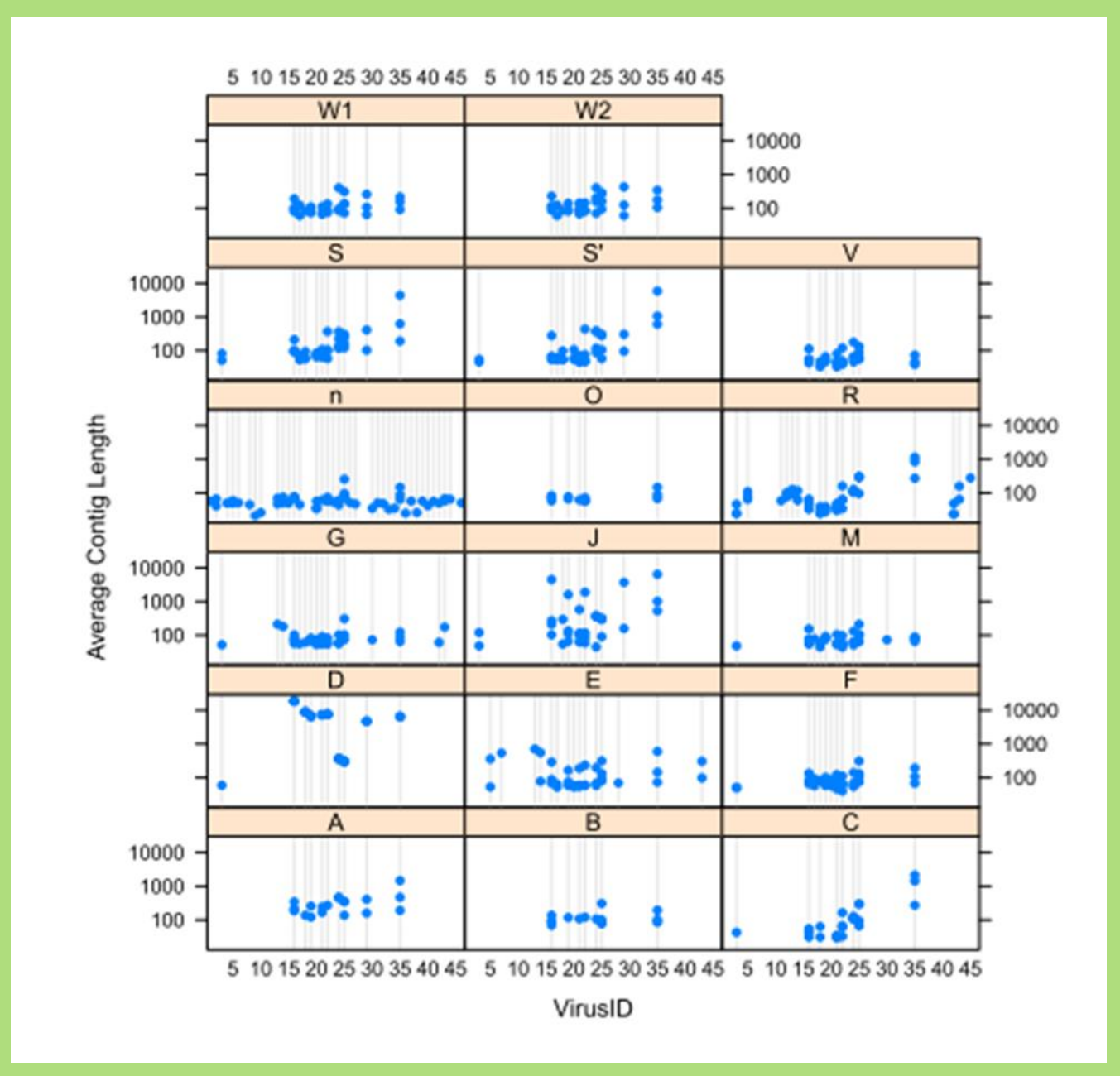
Evaluation of bioinformatic pipelines through proficiency testing on the same dataset of small RNA sequences

Material & Methods

- ✓ small RNA sequencing data from apple (ASGV), potato (PVX, 1 new Nepovirus) & grapevine (GLRa1V, GVA, GVB, HSVd, GYSVd, one marafivirus))
- ✓ Rarefaction at 3 sequencing depths: 50,000 , 250,000 (twice for grapevine) and 2.5 Million (=10 fastq files)
- ✓ **21 participating laboratories** (LabID) applying their own bioinformatic pipelines
- ✓ **One question**: Which viruses do you detected in the 10 fastq files ?

Results

1. Lenght of contigs



Variable distribution of viral contig lenght reflects the huge diversity of pipelines

2. False discovery rate (FDR)

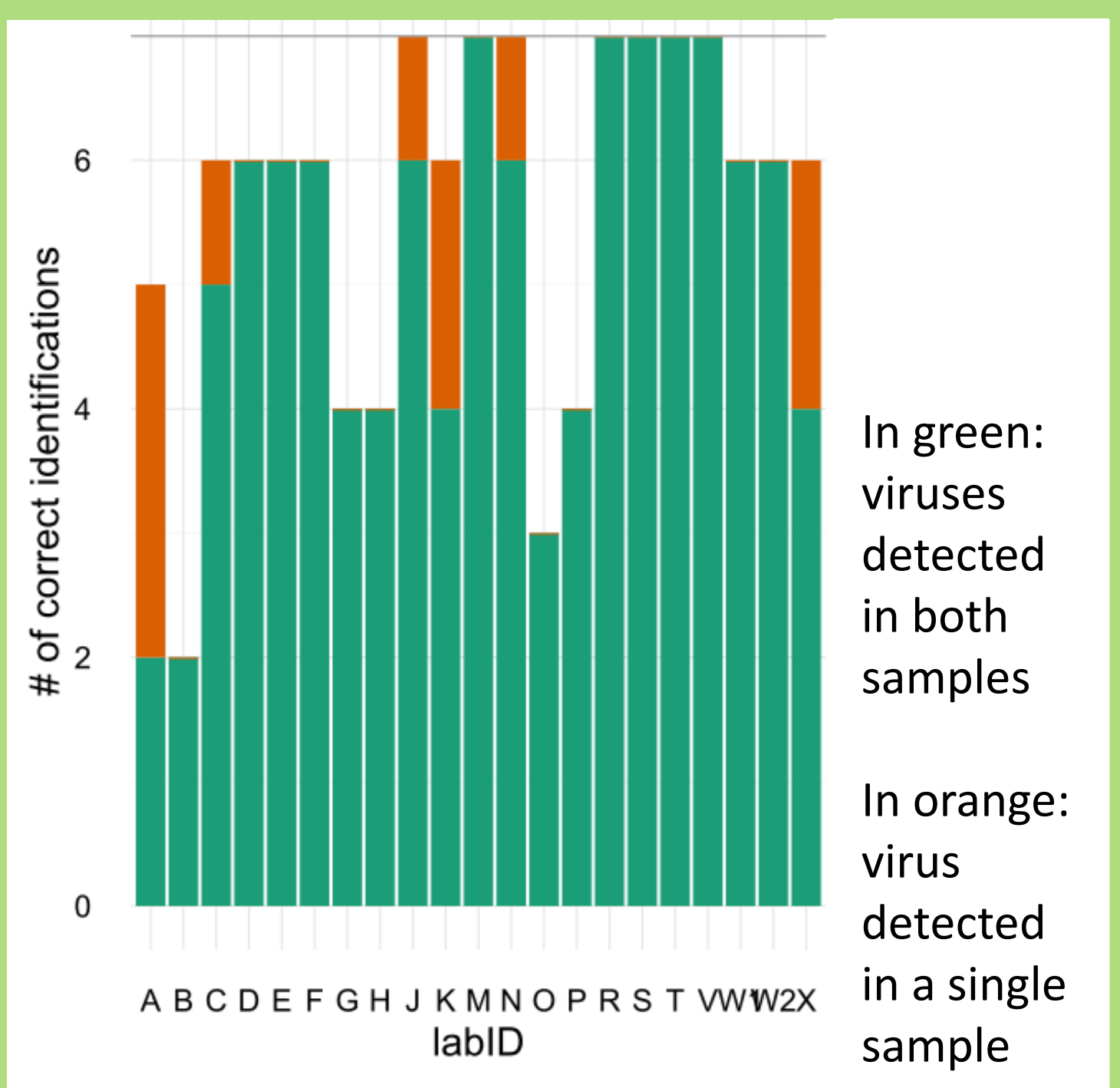
- ❖ Very low (0% for the majority of samples)
- ❖ Expert analysis needed for unknown viruses
- ❖ Report of integrated sequences ?

3. Sensitivity

	Sensitivity		
labID	50	250	2500
A	10%	53%	90%
B	30%	35%	80%
C	60%	71%	80%
D	50%	82%	100%
E	30%	82%	80%
F	80%	88%	100%
G	20%	53%	100%
H	30%	65%	70%
J	70%	94%	100%
K	40%	71%	90%
M	50%	94%	90%
N	30%	82%	90%
O	20%	41%	40%
P	20%	59%	70%
R	100%	100%	100%
S	50%	100%	100%
T	90%	100%	100%
V	60%	88%	80%
W1	40%	82%	90%
W2	60%	82%	90%
X	30%	71%	80%
AVERAGE	46%	75%	86%

- ❖ Decrease with rarefaction
- ❖ Lower for virus with low amount of sequences
- ❖ 1 participant with 100 %
- ❖ 7 participants with 100% at 2.5 M

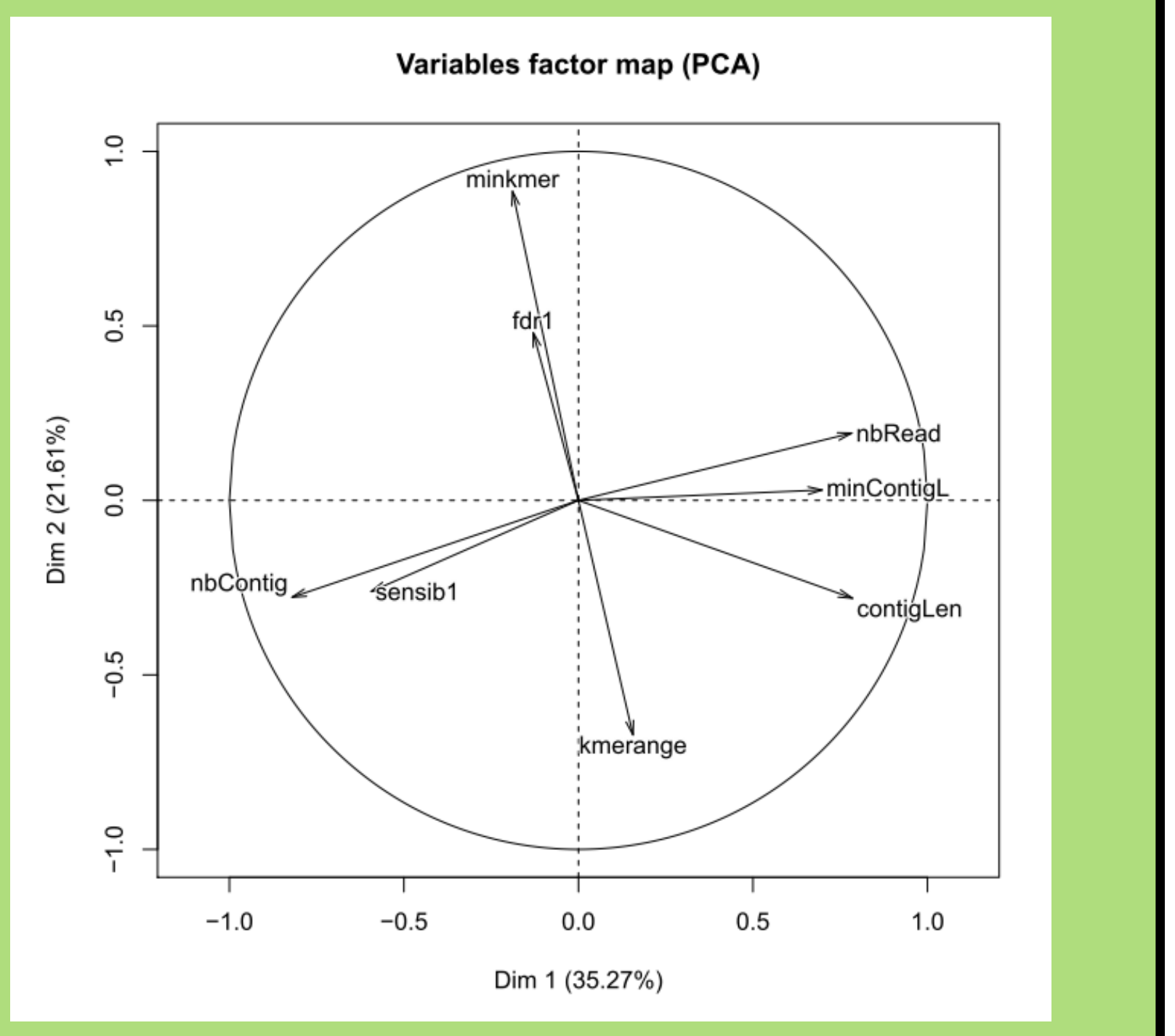
4. Repeatability



Evaluation at 250,000 sequences with 2 files from grapevine

- ❖ Repeatability of 93 %
- ❖ Repeatability of correct virus detection: 74% as some viruses are missed repeatably

5. PCA analysis



Principal Component Analysis:

- ❖ Sensitivity related to high number of viral contigs and small minimal contig lenght
- ❖ Minimum kmer size, kmer range and FDR have little influence on sensitivity
- ❖ Dispersion of participants in the 4 quadrants without clear clustering

Acknowledgement: This publication is based upon work from COST Action FA1407 (DIVAS), supported by COST (European Cooperation in Science and Technology).

cost
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

Conclusions

- ✓ **Huge diversity of pipeline used by participants**
- ✓ **Significant difference in sensitivity and repeatability**
- ✓ **Differences can be explained by the algorithms and their parameters , the used database and the scientist expertise**
- ✓ **An important effort for bioinformatic pipeline standardization is needed**