From cluster to computer: the progressive matching of diagnosticians with bioinformatics

Prof. Sébastien Massart – Liège University – Gembloux Agro-Bio Tech
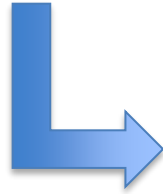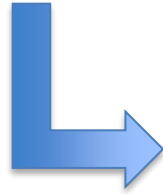
Us... 20 years ago

Us... 10 years ago

Us… from now

In 2009…

➢ Microbiome analysis by amplicon sequencing

Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa

Carlotta De Filippo[a], Duccio Cavalieri[a], Monica Di Paola[b], Matteo Ramazzotti[c], Jean Baptiste Poullet[d], Sebastien Massart[d], Silvia Collini[b], Giuseppe Pieraccini[e], and Paolo Lionetti[b,1]

[a]Department of Preclinical and Clinical Pharmacology, University of Florence, 50139 Firenze, Italy; [b]Department of Pediatrics, Meyer Children Hospital, University of Florence, 50139 Firenze, Italy; [c]Department of Biochemical Sciences, University of Florence, 50134 Firenze, Italy; [d]DNA Vision Agrifood S.A., B-4000 Liège, Belgium; and [e]Centro Interdipartimentale di Spettrometria di Massa, University of Florence, 50139 Firenze, Italy

➢ One month for informatician (& me) on a super calculation cluster of 500 k€

In 2020…

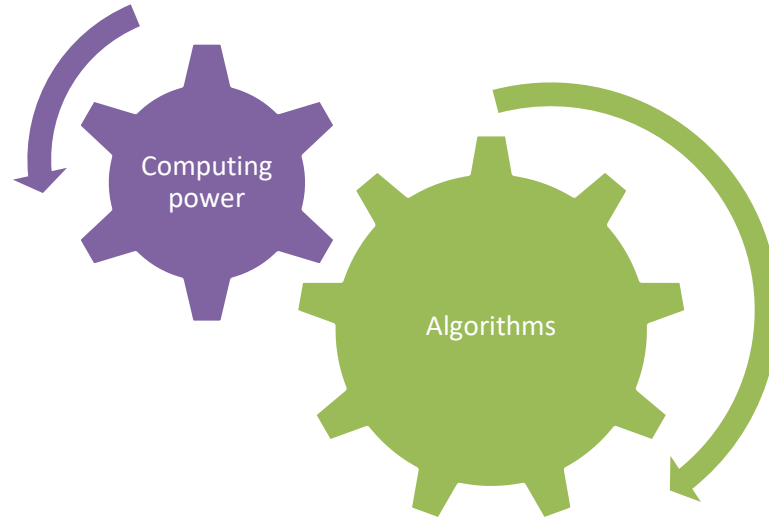➢ Microbiome analysis by amplicon sequencing

Techniques

*e-X*tra*

Comparison of qPCR and Metabarcoding Methods as Tools for the Detection of Airborne Inoculum of Forest Fungal Pathogens

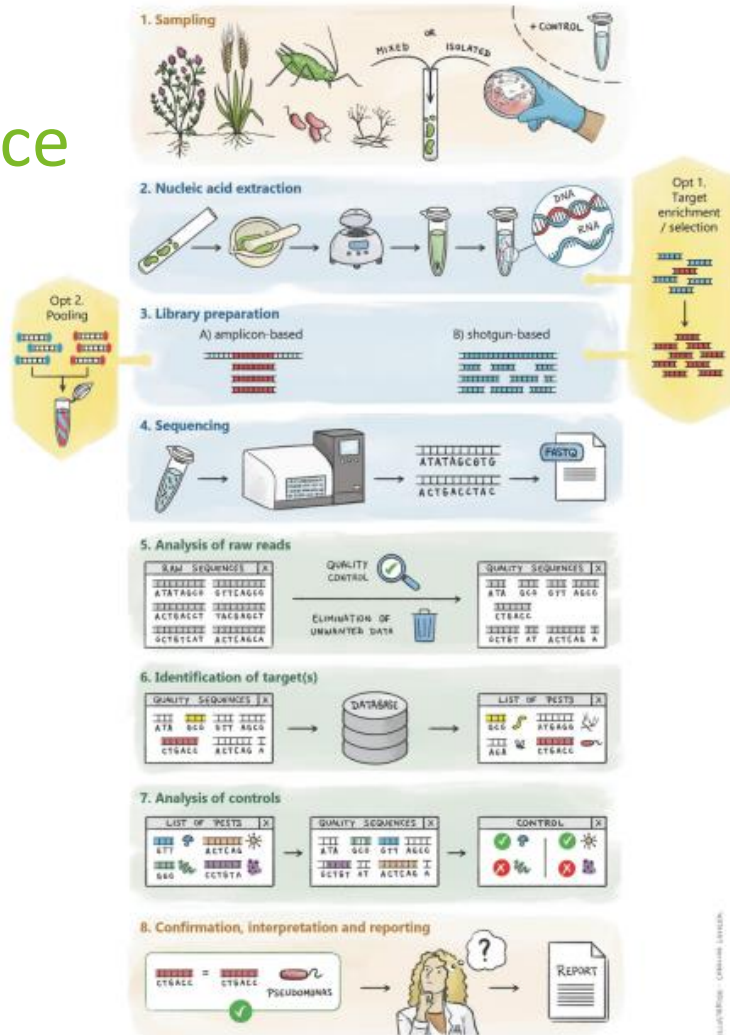Anne Chandelier,[1,†] Julie Hulin,[2] Gilles San Martin,[1] Frédéric Debode,[1] and Sébastien Massart[3]

➢ A scientist on a personal computer !!

Progressive matching ?

Computing power

Algorithms

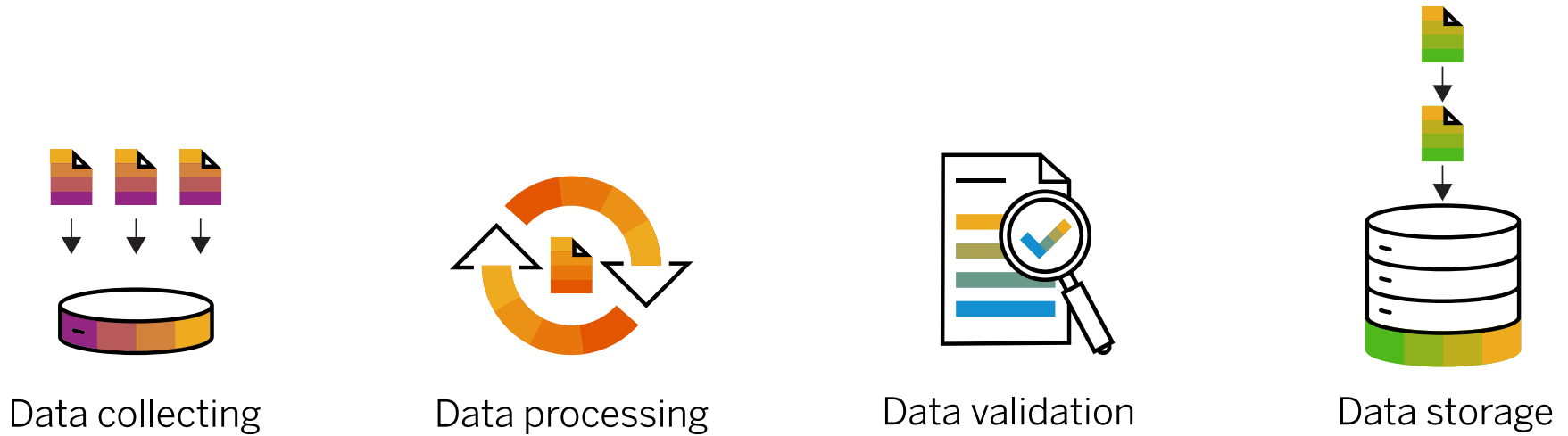# Laboratory & bioinformatics balance

Data management: overview

Data collecting

Data processing

Data validation

Data storage

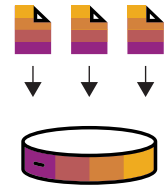5. Analysis of raw reads

6. Identification of target(s)

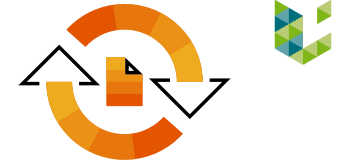7. Analysis of controls

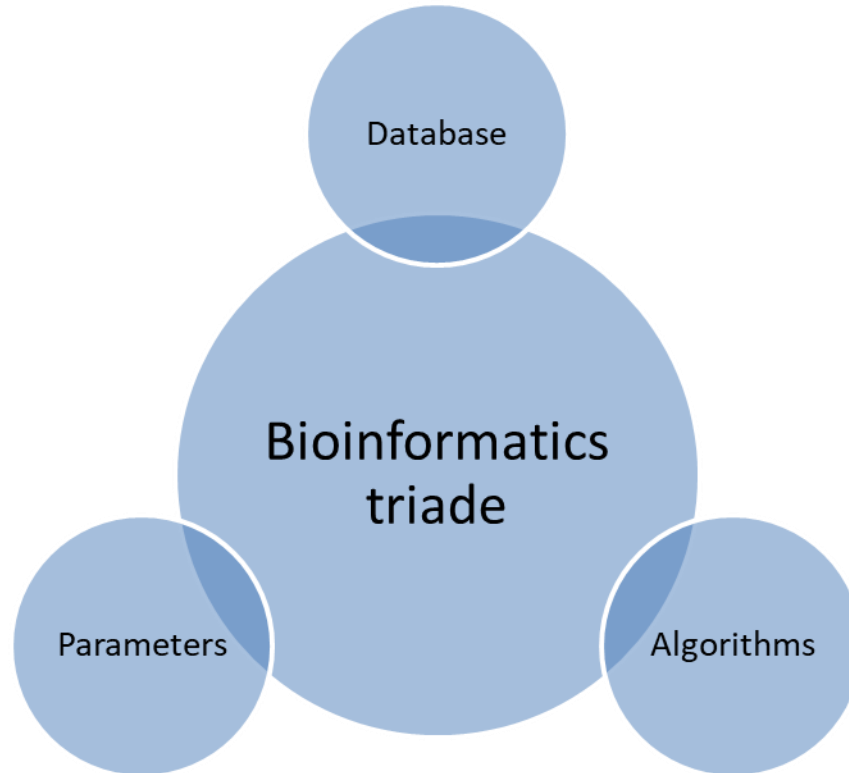# Data management: collection of data

Data collecting

➢ Results: amplicon (Mb) or shotgun (Gb)
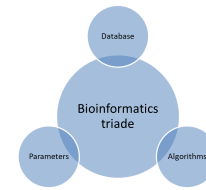
➢ Transfer and storage

➢ Reference database

Data management: data processing

Data processing

Bioinformatics triade

Database

Parameters

Algorithms

Data management: data processing

> # Bioinformatic Triade :

Data processing

**Virology**

*e-**X**tra**

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

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%**

Data management: data processing
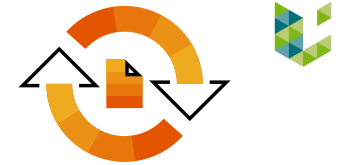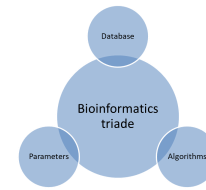
➤ Bioinformatic Triade

➤ At each step of data processing :

➤ Step 1: analysis of raw reads
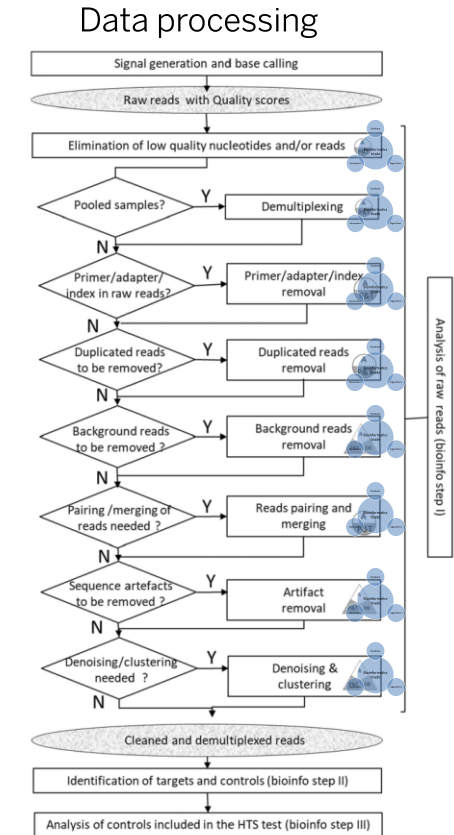
Data processing

Data management: data processing
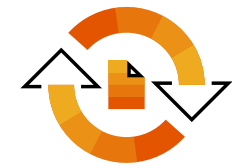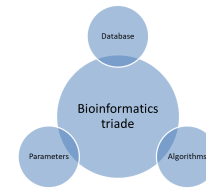
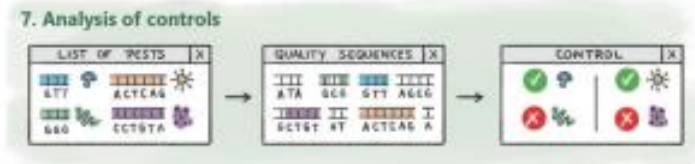Data processing

- ➢ Bioinformatic Triade

- ➢ At each step of data processing

- ➢ Step 2: identification of targets
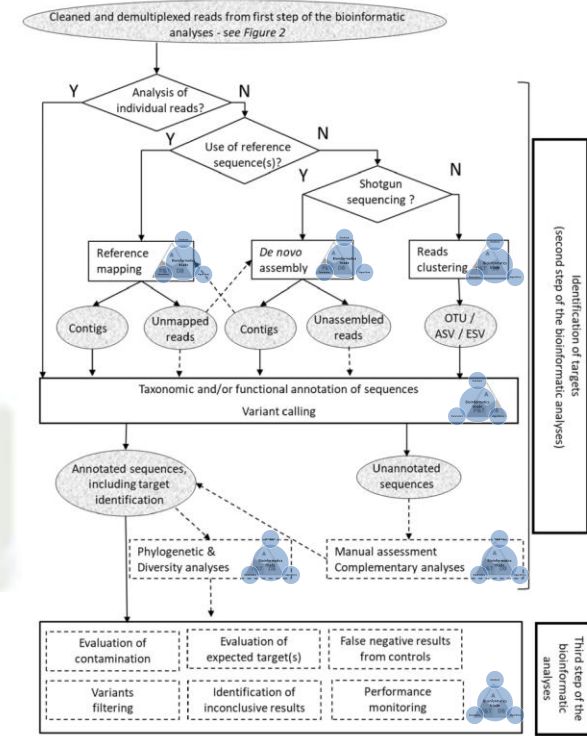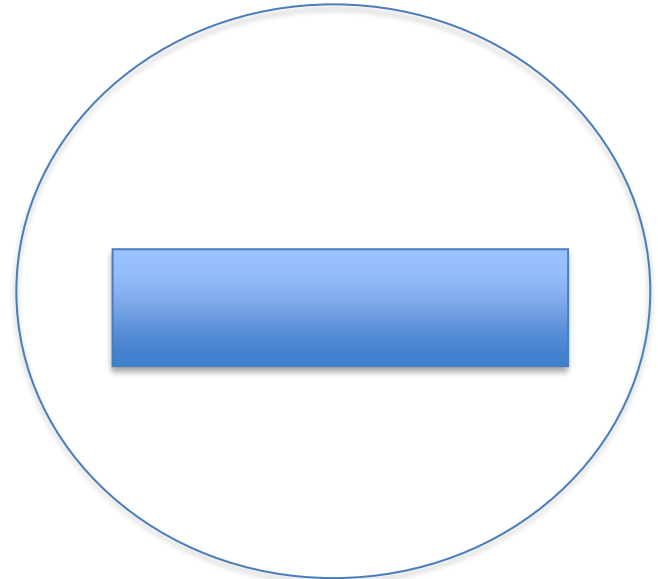
Data management: data processing

> ## About controls: before HTS

&

Data management: data processing

7. Analysis of controls

Data validation

The alien control

> With ... s also a
> nega...

Data management: data processing

7. Analysis of controls

Data validation

> Bioinformatic Triade & scientist expertise to gain collectively

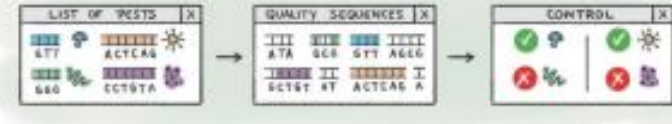| Library name | Batch | Number of reads mapped to RefSeq without validation threshold | | | | |
		BanMMV	BBrMV	BBTV	CMV	BYDV
Alien control	3		1			125616
ITC0146	3				2	7
ITC0148	3					9
ITC1498	3				845	6
ITC1565	3	7	2		686	11
ITC1607	3				2	3
ITC1847	3				1	4
ITC1859 - Old Kit	3	950				288
ITC1859	3	1840	4			5
ITC1861	3	1903				2
ITC1863	3	816			1	3
ITC1867 - Old Kit	3	558	1		1	9
ITC1867	3	1167	5		1	1
Negative Control	3				1	184
Positive Control - Old Kit	3	512	7205	52	22320	11
Positive Control	3	807	16787	66	32672	6
Sample 2	3		4		2	6
Sample 3	3		5		3	8
Sample 4	3		3			2
Sample 5	3		1		1	4
Sample J	3	3247	288675			8
Sample EM4	3	903	114		1884	1
BBrMV2 No.208	3		3099			5
BBrMV3 No.58	3		1519		2	40
BBrMV4 No.15	3		4715		2	22
BSOLV1 No.10	3		1		2	10
BBTV No.9.58.2	3		8	715	1	44

# Data management: data processing

7. Analysis of controls

Data validation

➤ **Bioinformatic Triade & scientist expertise to gain collectively**

| Library name	Batch	Number of reads mapped to RefSeq without detection threshold				
		BanMMV	BBrMV	BBTV	CMV	BYDV
Alien control	1		3		1	22525
Alien control	1	1			2	25366
Alien control	1	1			11	8405
Alien control	1				6	30928
ITC1543	1		2		4	
ITC1543	1					
ITC1543	1	1	10		1	
ITC1543	1				3	
ITC1543	1					7
ITC1723	1				1	5
ITC1783	1	325			4	
ITC1799	1		3			
ITC1817	1					
ITC1827	1				4	
ITC1831	1	485				
ITC1833	1				4	
ITC1840	1		2			
ITC1843	1	2			2	
ITC1843	1					
ITC1845	1					
ITC1845	1					
ITC1852	1	749				
ITC1855	1	1557				
Positive control Mix	1	22	330	3	1988	
Positive control Mix	1	522	894	69	6271	
Positive control Mix	1	338	25592	30	79496	
Positive control Mix	1	437	28971	34	835	

Can I store my data ?
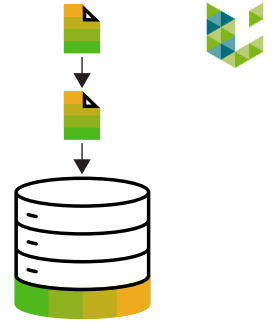
Data storage

- 2400–3000 Gbytes for 16-20 billions reads (up to 3,000 billions bases)

- Virus: 10M reads -> 2-3 Gbytes for one sample

- Which intermediate results to keep ?

Progressive matching
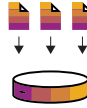
Data collecting

Data processing

Data validation

Data storage

Storage capacity

Computing power

Algorithms

Progressive matching: user friendly interfaces

1. Linux B&W screens

Progressive matching: use friendly interfaces

2. User-friendly interface (€)

Progressive matching: use friendly interfaces

VirusDetect
An automated pipeline for efficient virus discovery using deep sequencing of small RNAs

virtool/**virtool**

Viral infection diagnostics using next-generation sequencing

A 16 Contributors   ⊙ 8 Issues   💬 10 Discussions   ☆ 37 Stars   ⑂ 22 Forks

3. Virus-oriented (linux or windows)

Progressive matching: use friendly interfaces

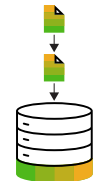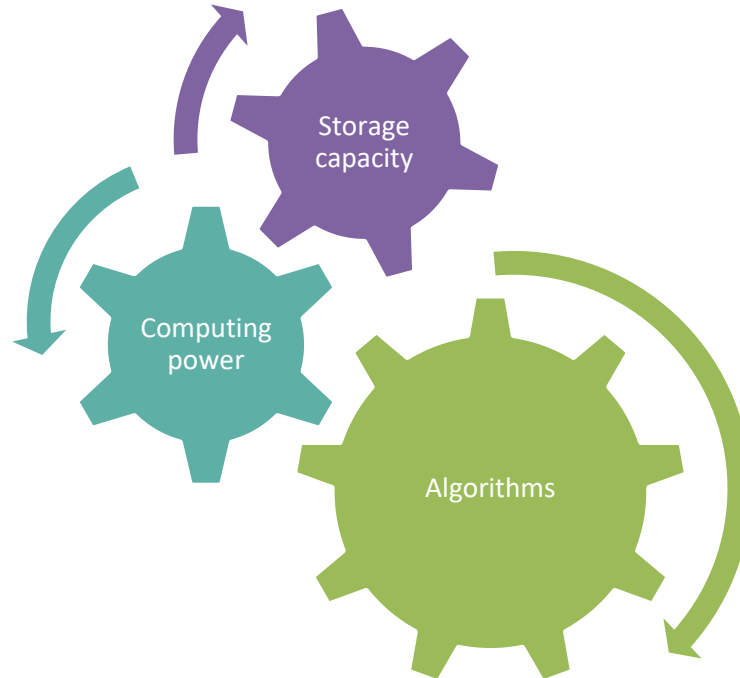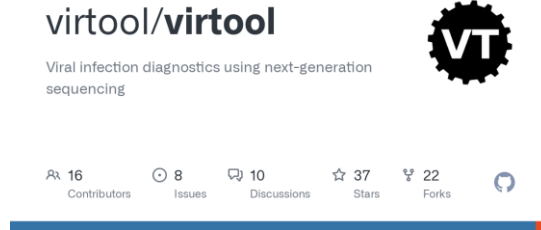# 4. HTML-based

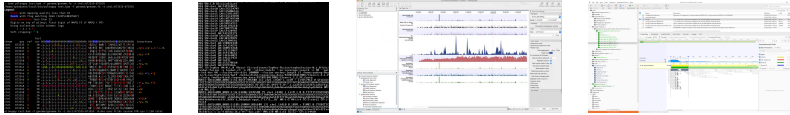# Integrity, safety & hacking
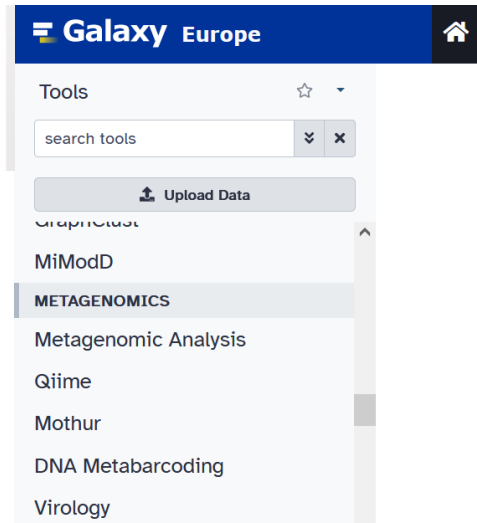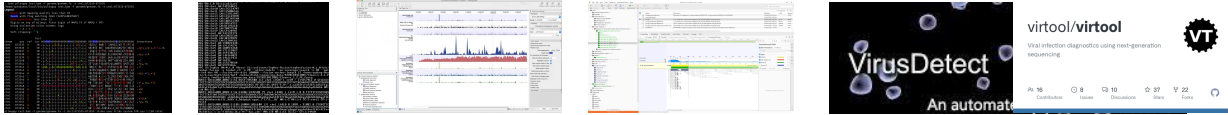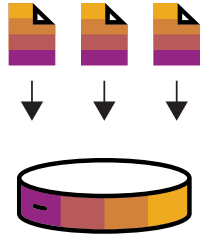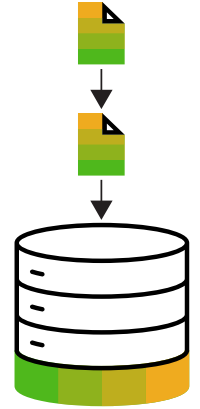

Data collecting

Data processing

Data validation

Data storage

Guidelines published last year

Peer Community Journal

Section: Infections

RESEARCH ARTICLE

Published
2022-10-25

Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests

Sebastien Massart[1], Ian Adams[2], Maher Al Rwahnih[3], Steve Baeyen[4], Guillaume J. Bilodeau[5], Arnaud G. Blouin[1], Neil Boonham[6], Thierry Candresse[7], Anne Chandellier[8], Kris De Jonghe[4], Adrian Fox[2], Yahya Z.A. Gaafar[9], Pascal Gentit[10], Annelies Haegeman[4], Wellcome Ho[11], Oscar Hurtado-Gonzales[12], Wilfried Jonkers[13], Jan Kreuze[14], Denis Kutjnak[15], Blanca B. Landa[16], Mingxin Liu[17], François Maclot[1], Marta Malapi-Wight[18], Hans J. Maree[19,20], Francesco Martoni[21], Natasa Mehle[15,22], Angelantonio Minafra[23], Dimitre Mollov[24], Adriana G. Moreira[25], Mark Nakhla[26], Françoise Petter[27], Alexander M. Piper[21], Julien P. Ponchart[1,28], Robbie Rae[29], Benoit Remenant[10], Yazmin Rivera, Brendan Rodoni[21], Marleen Botermans[30], J.W. Roenhorst[30], Johan Rollin[1], Pasquale Saldarelli[23], Johanna Santala[31], Rose Souza-Richards[32], Davide Spadaro[33], David J. Studholme[34], Stefanie Sultmanis[5], René van der Vlugt[35], Lucie Tamisier[1], Charlotte Trontin[27], Ines Vazquez-Iglesias[2], Claudia S.L. Vicente[36], Bart T.L.H. van de Vossenberg[30], Marcel Westenberg[30], Thierry Wetzel[37], Heiko Ziebell[9], and Benedicte S. M. Lebas[1]

Volume 2 (2022), article e62

https://doi.org/10.24072/pcjournal.181

Cite as
Sebastien Massart, Ian Adams, Maher Al Rwahnih, Steve Baeyen, Guillaume J. Bilodeau, Arnaud G. Blouin, Neil Boonham, Thierry Candresse, Anne Chandellier, Kris De Jonghe, Adrian Fox, Yahya Z.A. Gaafar, Pascal Gentit, Annelies Haegeman, Wellcome Ho, Oscar Hurtado-Gonzales, Wilfried Jonkers, Jan Kreuze, Denis Kutjnak, Blanca B. Landa, Mingxin Liu, François Maclot, Marta Malapi-Wight, Hans J. Maree, Francesco Martoni, Natasa Mehle, Angelantonio Minafra, Dimitre Mollov, Adriana G. Moreira, Mark Nakhla, Françoise Petter, Alexander M. Piper, Julien P. Ponchart, Robbie Rae, Benoit Remenant, Yazmin Rivera, Brendan Rodoni, Marleen Botermans, J.W. Roenhorst, Johan Rollin, Pasquale Saldarelli, Johanna Santala, Rose Souza-Richards, Davide Spadaro, David J. Studholme, Stefanie Sultmanis, René van der Vlugt, Lucie Tamisier, Charlotte Trontin, Ines Vazquez-Iglesias, Claudia S.L. Vicente, Bart T.L.H. van de Vossenberg, Marcel Westenberg, Thierry Wetzel, Heiko Ziebell and Benedicte S. M. Lebas (2022) *Guidelines for the reliable use of high throughput sequencing technologies to detect plant pathogens and pests*, Peer Community Journal, 2: e62.

Correspondence
sebastien.massart@uliege.be

Peer-review

DOI: 10.1111/epp.12863

ORIGINAL ARTICLE

Facilitating the adoption of high-throughput sequencing technologies as a plant pest diagnostic test in laboratories: A step-by-step description

Benedicte Lebas[1] | Ian Adams[2] | Maher Al Rwahnih[3] | Steve Baeyen[4] | Guillaume J. Bilodeau[5] | Arnaud G. Blouin[1] | Neil Boonham[6] | Thierry Candresse[7] | Anne Chandelier[8] | Kris De Jonghe[4] | Adrian Fox[2] | Yahya Z. A. Gaafar[9] | Pascal Gentit[10] | Annelies Haegeman[4] | Wellcome Ho[11] | Oscar Hurtado-Gonzales[12] | Wilfried Jonkers[13] | Jan Kreuze[14] | Denis Kutjnak[15] | Blanca Landa[16] | Mingxin Liu[17] | François Maclot[1] | Martha Malapi-Wight[18] | Hano J. Maree[19,20] | Francesco Martoni[21] | Natasha Mehle[15] | Angelantonio Minafra[22] | Dimitre Mollov[23] | Adriana Moreira[24] | Mark Nakhla[25] | Françoise Petter[26] | Alexander M. Piper[21] | Julien Ponchart[1] | Robbie Rae[27] | Benoit Remenant[10] | Yazmin Rivera[25] | Brendan Rodoni[21] | Johanna W. Roenhorst[28] | Johan Rollin[1] | Pasquale Saldarelli[22] | Johanna Santala[29] | Rose Souza-Richards[30] | Davide Spadaro[31] | David J. Studholme[32] | Stefanie Sultmanis[5] | René van der Vlugt[33] | Lucie Tamisier[1] | Charlotte Trontin[26] | Ines Vazquez-Iglesias[2] | Claudia S. L. Vicente[34] | Bart T. L. H. Vossenberg[28] | Thierry Wetzel[35] | Heiko Ziebell[9] | Sebastien Massart[1]
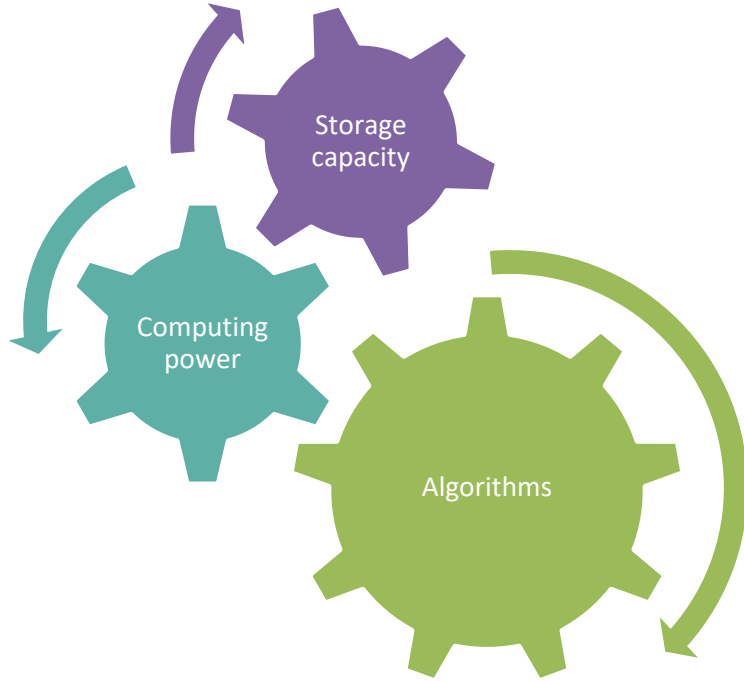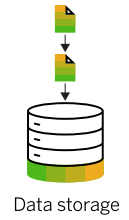
EPPO standard on diagnostics

EPPO STANDARD ON DIAGNOSTICS

PM 7/151 (1) Considerations for the use of high throughput sequencing in plant health diagnostics[1]

Conclusion

Data collecting

Data processing

Data validation

Data storage

Storage capacity

Computing power

Algorithms

Workshop this afternoon: from survey

➢ Hardware capacity and upgrade : storage ?

➢ Hardware capacity and upgrade: CPU ?

➢ IT security and policies compatible ?

➢ Which level of bioinformatic knowledge ?

Special thanks to hundreds of plant pathologists collaborating on disseminating HTS and making it more reliable

Thank you for your attention