Introduction of the webinar series

The use and validation of High Throughput sequencing (HTS) tests for diagnostics of plant pests.

Webinar 1	What is High Throughput Sequencing (HTS)?	Friday 30th April, 2 pm-3:30 pm
Webinar 2	How to prepare your laboratory to conduct HTS tests?	Monday 3rd May, 2 pm-3 pm
Webinar 3	How to develop, validate and routinely use HTS for diagnostic purpose?	Tuesday 4th May, 2 pm-3:30 pm
Practical training activity	How to apply the guidelines to your laboratory?	Wednesday 5th May, 2 pm to 4:30 pm. Friday 7th May, 2 pm to 4:30 pm.

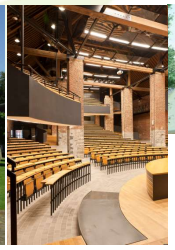
VALITEST Webinar series

What is High Throughput Sequencing (HTS)?

Sebastien Massart

Gembloux AgroBio Tech – Liège University

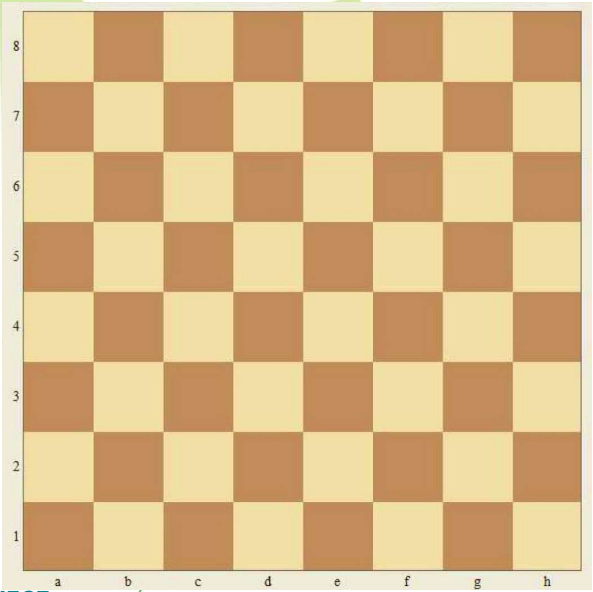
30/04/2021



1. Introduction : History

LIÈGE université Valitest

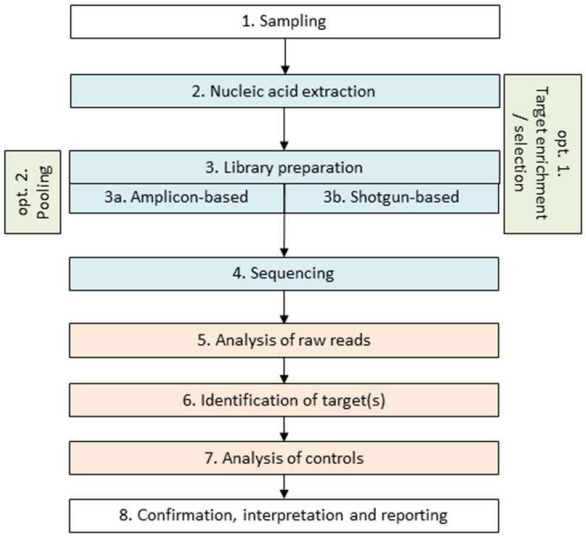
1. Introduction



www.wooclap.com/OCIPBU

LIEGE université Valitest

1. Introduction: HTS steps

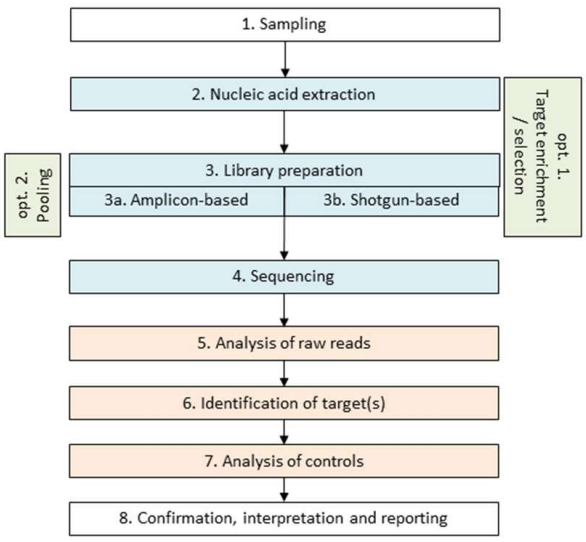


2. HTS Laboratory protocols

2.1 Sampling

- Representative of sample status -> like any other test
- Matrix:
  1. Multiple organisms (plant, environmental samples, spore traps, insects with microbiota…)
  2. Isolated organisms (ufc)

Steps of HTS technologies



2.2 Nucleic Acids extraction

- Huge diversity of protocols used
- DNA, RNA, smallRNA, dsRNA…
- Look into the M&m in the most recent publication you have read on THS (virus, fungi, bacteria…) – (5 minutes)
- What are the protocol used for N.A. extraction ?
- Are there specificities ?
- Send in the chat your observation

2.2 Nucleic acids extraction

- ✓ Protocols used for NA extraction ?

www.wooclap.com/OCIPBU

2.2 Nucleic Acids extraction

- Huge diversity of protocols used
- DNA, RNA, smallRNA, dsRNA…
- Not a best fit-for-all kit -> depending on your plant, tissue and experience
- Like (RT)-PCR but higher quality needed -> CONTROL !
- Quantity of RNA needed is going down
  5-10 µg -> 1 µg

2.3 RNA quality control

How the RNA Integrity Number (RIN) Works

Step 1:
Researchers deposit their total RNA sample into an RNA Nano LabChip.

Step 2:
They insert the LabChip into the Agilent bioanalyzer and let the analysis run, generating a digital electropherogram.

Source: www.ag.unr.edu

2.3 RNA quality control

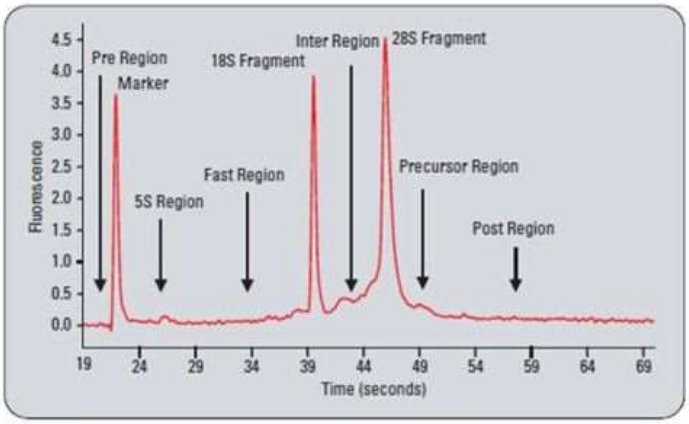
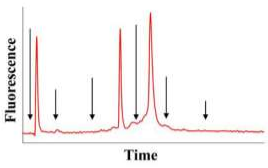


Figure 3
Electropherogram detailing the regions that are indicative of RNA quality.

Source: http://www.openwetware.org/wiki/BioMicroCenter:RIN

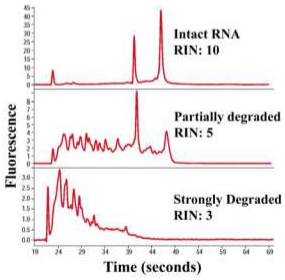
2.3 RNA quality control

Step 3:
The new RIN algorithm then analyzes the entire electrophoretic trace of the RNA sample, including the presence or absence of degradation products, to determine sample integrity. Important elements of the electropherogram that are indicative of RNA quality are shown in the figure at right.



Step 4:
The algorithm assigns a 1 to 10 RIN score, where level 10 RNA is completely intact. Because interpretation of the electropherogram is automatic and not subject to individual interpretation, universal and unbiased comparison of samples is enabled and repeatability of experiments is improved.

The RIN algorithm was developed using neural networks and adaptive learning in conjunction with a large database of eukaryote total RNA samples, which were obtained mainly from human, rat, and mouse tissues. The RIN score is largely independent of the amount of RNA used and the origin of the sample.



Source: www.ag.unr.edu

2.3 RNA quality control

What the RIN **can** do:

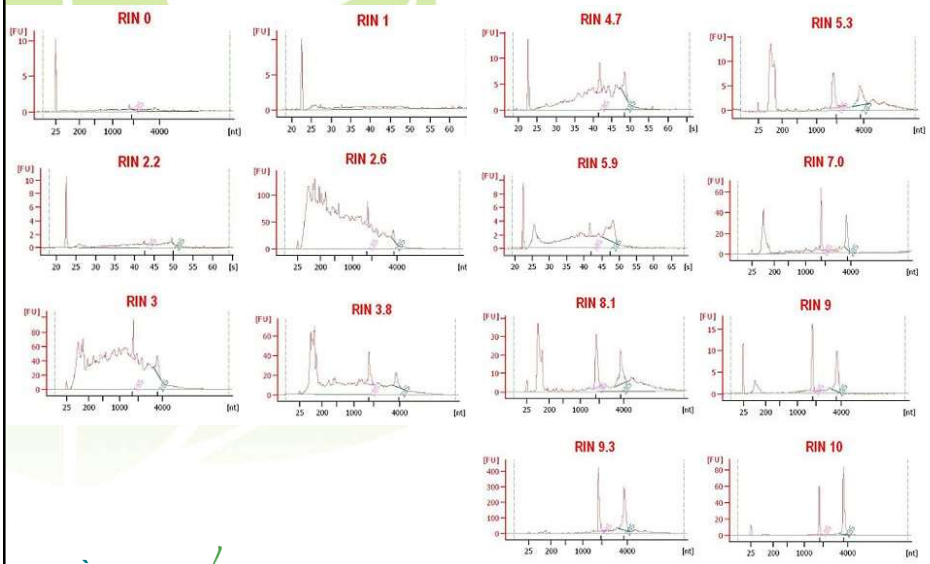
- ✓ Obtain a numerical assessment of the integrity of RNA.
- ✓ Directly compare RNA samples, e.g. before and after archival,
- ✓ compare integrity of same tissue across different labs.
- ✓ Ensure repeatability of experiments, e.g. if RIN shows a given value and is suitable for microarray experiments, then the RIN of the same value can always be used for similar experiments given that the same organism/tissue/extraction method is used.

What the RIN **cannot** do:

Tell a scientist ahead of time whether an experiment will work or not if no prior validation was done (e.g. RIN of 5 might not work for microarray experiments, but might work well for an appropriate RT-PCR experiment. Also, a RIN that might be good for a 3' amplification might not work for a 5' amplification).

Minimal requested RIN = 7 or 8… complicated for plants !!!

2.3 RNA quality control

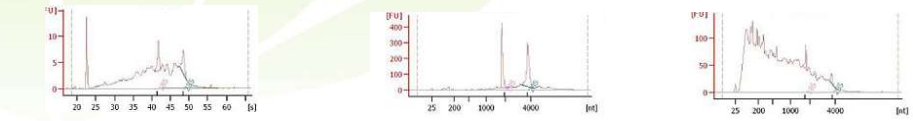


2.3 RNA quality control

✓ Evaluation of RIN:

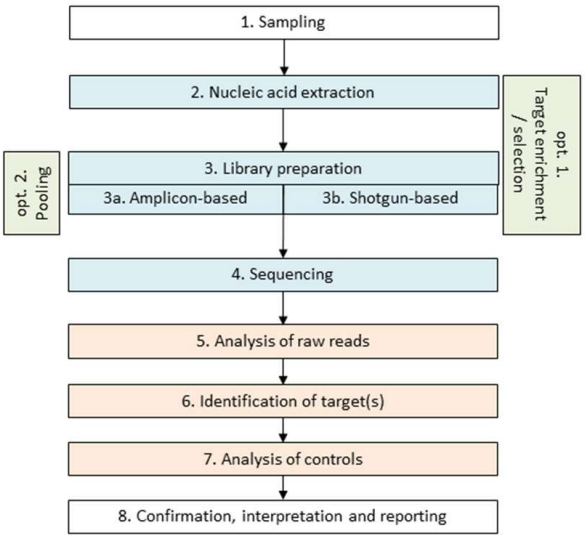
www.wooclap.com/OCIPBU

Fair | Excellent | Bad-very bad



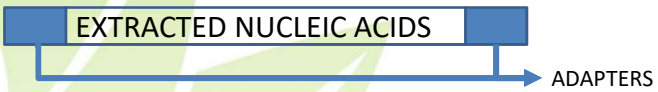
Yet, viruses can still be identified from samples with low RIN quality

Steps of HTS technologies



2.4 Library preparation

- Format of nucleic acids **compatible** with the sequencing platform in sufficient quantity of the appropriate size



- Two groups of protocols:

1. AMPLICON SEQUENCING – PCR based

2. SHOTGUN SEQUENCING

2.4 Library preparation - conclusion

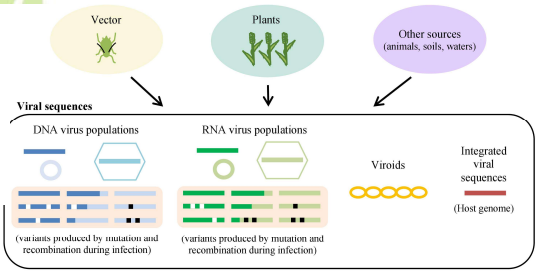
It is only about classical molecular biology: shearing, digesting, ligating, amplifying

2.3 Used protocols for virus discovery

REVIEW ARTICLE
Front. Microbiol., 16 October 2020 | https://doi.org/10.3389/fmicb.2020.578064

Illuminating an Ecological Blackbox: Using High Throughput Sequencing to Characterize the Plant Virome Across Scales

François Maclot[1*], Thierry Candresse[2], Denis Filloux[3,4], Carolyn M. Malmstrom[5], Philippe Roumagnac[3,4], René van der Vlugt[6] and Sébastien Massart[1*]

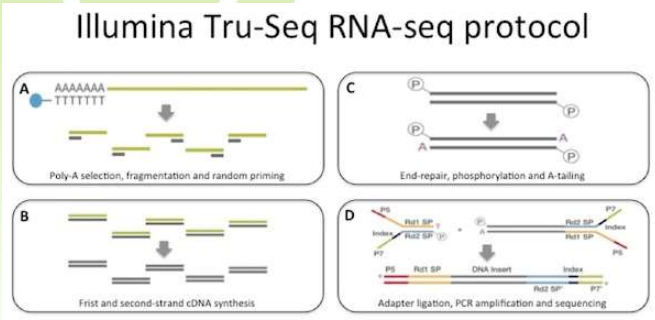


HTS

Nucleic acids	Total RNA	dsRNA	small RNA (21, 22, 24 nt)	VANA
Potentially detectable viral sequences				
Ex. of platforms used in ecological studies	Illumina™ / PacBio™	Illumina™ / PacBio™	Illumina™	Illumina™ / PacBio™

2.4 Library preparation – RNA example

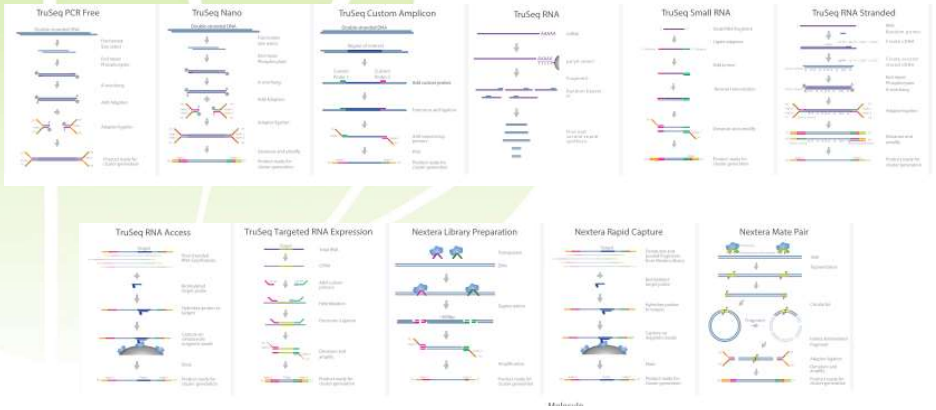
✓ Total RNA Truseq protocol after ribodepletion :



Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

2.4 Library preparation – shotgun RNA

- For Illumina: 11 kits existing

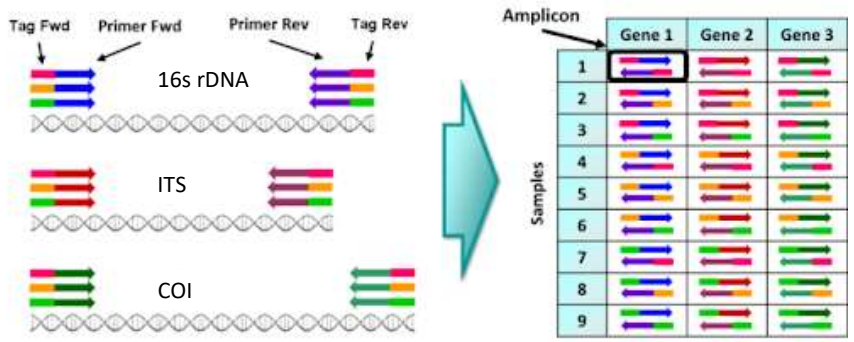


2.4 Library preparation – shotgun RNA

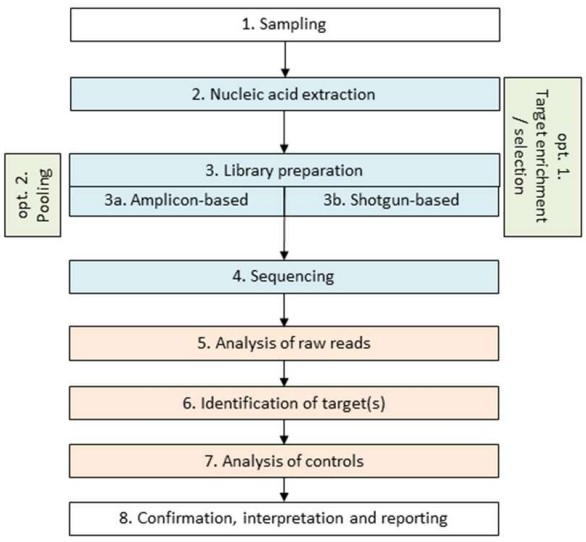
- For Illumina: >100 protocols

2.4 Library prep. – amplicon example

- ✓ Analysing bacterial, fungal and insect composition of a composite sample:



Steps of HTS technologies



2.5 Available sequencing technologies

- ✓ Illumina
- ✓ Ion Torrent
- ✓ Pacific biosciences
- ✓ Oxford Nanopore technologies

2.5 Available sequencing technologies

Roche FLX, Helicos, Polonator, Complete Genomics, Solid,…

2.5 Available sequencing technologies

- ✓ For all the sequencing technologies, specific protocols of library preparation have been developped
- ✓ Based on the same steps but some specificities make them incompatible

2.5 Available sequencing technologies

- ✓ Four videos will be presented.
- ✓ Compare the technologies based on the following criteria
  - ✓ Detection method
  - ✓ Nucleotide used
  - ✓ Enzyme location
  - ✓ Amplification
  - ✓ Lenght of sequence
  - ✓ Fixation of DNA on support

Illumina - Ion torrent - Pacific Bioscience - Oxford Nanopore

2.5 Available sequencing technologies

	Illumina	Ion torrent	PacBio	Minion
Detection method	Fluorescence	Electric signal	Fluorescence	Electric signal
Nucleotides	Mix	One by one	Mix	n.a.
Enzyme	Reagent	Reagent	Fixed	Fixed
Amplification	Bridge	Beads	No	No
Lenght	Max 300 nt	Max 300 nt	Up to 10 kb	Up to 10 kb
Fixation of DNA	Oligo on cell	Oligo on beads	Enzyme	Enzyme + nanopore

2.5 Available sequencing technologies

Technology	Read length (bp)	Throughput	Reads	Runtime	Error profile	Instrument cost (US$)	Cost per Gb (US$, approx.)
Sequencing by synthesis: CRT							
Illumina MiniSeq Mid output	150 (SE)	2.1–2.4 Gb	14–16 M	17 h	<1%, substitution	$50	$200–300
Illumina NextSeq 500/550 Mid output	75 (PE)	16–20 Gb	Up to 260 M (PE)	15 h	<1%, substitution	$250	$42
	150 (PE)	32–40 Gb		26 h			$40
Illumina HiSeq X	150 (PE)	800–900 Gb per flow cell	2.6–3 B (PE)	<3 d	0.1%, substitution	$1,000	$7.0
Sequencing by synthesis: SNA							
Ion Proton	Up to 200 (SE)	Up to 10 Gb	60–80 M	2–4 h	1%, indel	$224	$80
Ion S5 540	200 (SE)	10–15 Gb	60–80 M	2.5 h	1%, indel	$65	$300
Single-molecule real-time long reads							
Pacific BioSciences RS II	~20 Kb	500 Mb–1 Gb	~55,000	4 h	13% single pass, ≤1% circular consensus read, indel	$695	$1,000
Pacific Biosciences Sequel	8–12 Kb	3.5–7 Gb	~350,000	0.5–6 h	?	$350	?
Oxford Nanopore MK 1 MinION	Up to 200 Kb	Up to 1.5 Gb	>100,000	Up to 48 h	~12%, indel	$1	$750

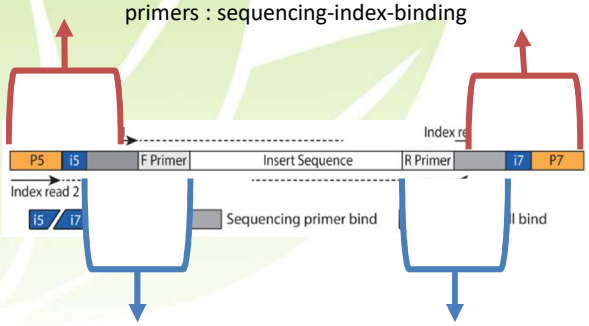
Goodwin et al., 2016. Nature Review genetics

Frequency of use in virus discovery and/or characterisation

2.6 Library preparation & sequencing

✓ Adapters at both ends for Illumina sequencing: amplicon

For amplicon sequencing: second round of PCR with primers : sequencing-index-binding

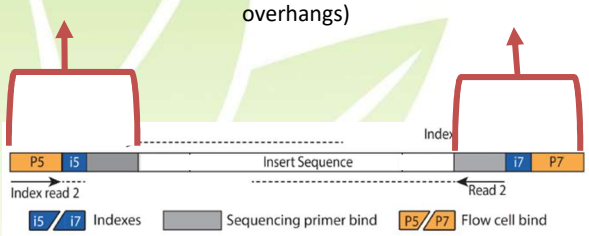


For amplicon sequencing: composite primers (target region in 3' & sequencing primer in 5')

2.6 Library preparation & sequencing

✓ Adapters at both ends for Illumina sequencing: shotgun

For shotgun sequencing: ligation of adapter (A overhangs)



3. HTS bioinformatics protocols

LIÈGE université

3.1 Introduction

- Balance between laboratory and bioinformatic

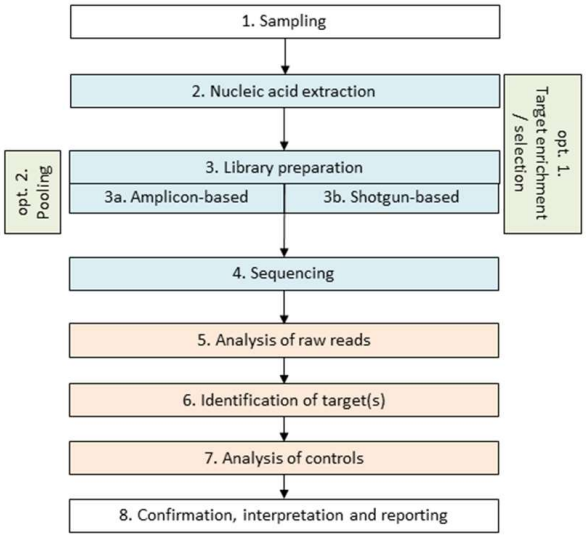
3.1 Introduction

- Importance of bio-informatic in generation of bias:

3.1 Introduction

- Parameter setting & result interpretation :

Steps of HTS technologies



3.2 Analysis of raw reads

- From light to fastq format:

@M00991:160:000000000-BC4TY:1:1101:17783:1494 1:N:0:CTGAAGCT+CAGGACGT — NAME

CTCTAAGAAGCTGGCTGCGGAGGCTCACCTTCACATAGCTAGTTAGCAGGCTGAGGTCTCGTT
CGTTAACGGAATTAACCAGACAAATCGCTCCACCAACTAAGAACGGCCATGCACCACCACCCA
TAGAATCAAGAAAGA — Sequence

+

3AA3FFFFFFFGCE4FGCEAEEEHFCEGGBGDHFHFBGFEFDHGFHEFAACG1BFHHHDFFGHF
GHAFHGG3CFFH4FFHGHFFHBGHHFGHHGHFFEEEHGGGHHHGGEEEGHH1DGFFFGHGG
GGHHGGFFHGFHHGH — Quality of each base

@M00991:160:000000000-BC4TY:1:1101:16780:1498 1:N:0:CTGAAGCT+CAGGACGT — NAME

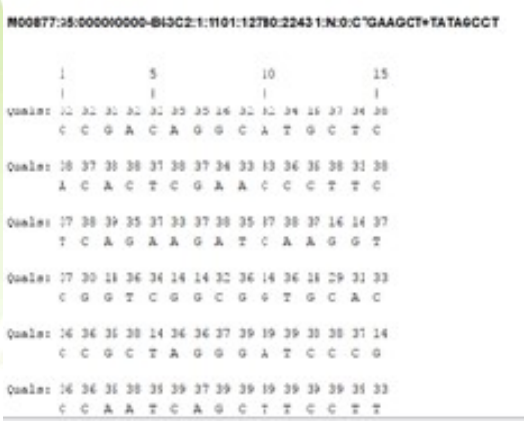
CCCGGTATTGTTATTTATTGTCACTACCTCCCCGTGTCAGGATAGATCGGAAGAGCACACGTCT
GAACTCCAGTCACCTGAAGCTATCTCGTATGC — Sequence

+

>ABABDDAFFDDDGGGFGGGGGBGFHFGHHHGGGEGHHHGFHGGHHGFBGHHGEGHHF
GGGHGFHHGHHH0BGHHGGE////////0/?1// — Quality of each base

6.2 Analysis of raw reads

- From light to fastq format:



6.2 Analysis of raw reads

- Quality control of the sequences thanks to the quality score of each single base

  -> elimination of reads
  -> elimination of nucleotides

- Importance ?

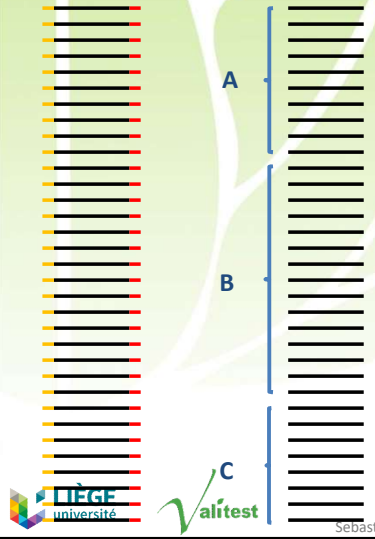
  www.wooclap.com/OCIPBU

Stringency of quality control depends on the purpose of the HTS test

6.2 Analysis of raw reads

1. Raw data with index

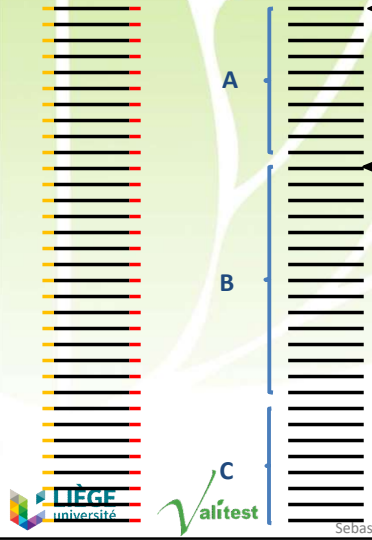
2. Raw data by sample



6.2 Analysis of raw reads

1. Raw data with index

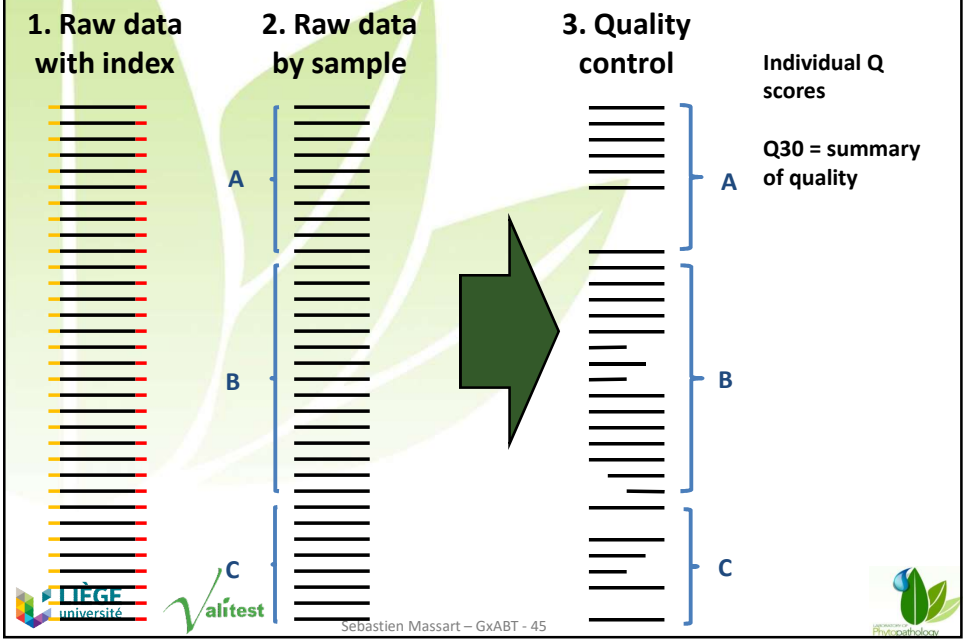
2. Raw data by sample



@Sequence 1
ATCCGACTAGATCAGATCAGAGAGGGAAA
+
40 40 40 40 30 20 10 10 10 40 38 10 20 24

@Sequence 10
CGTTTTAGAGAAACGACAGACTAGACAGAT
+
36 38 40 20 40 40 20 28 32 32 30 10 10 40

6.2 Analysis of raw reads



6.2 Analysis of raw reads

- ✓ **Removal of redundant reads**

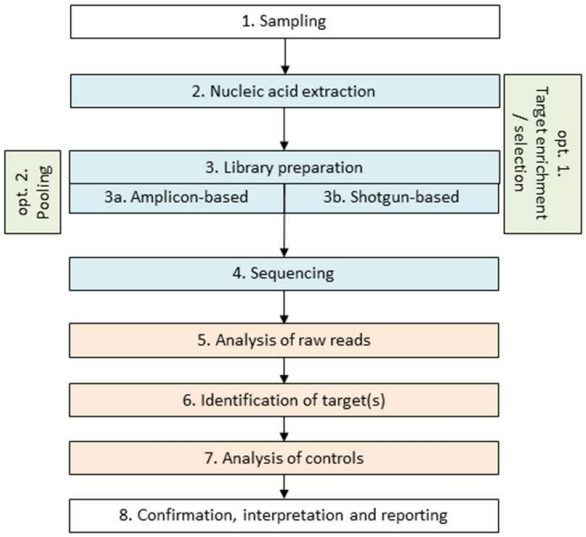
  Elimination of reads that are 100% identical to another (full sequence, including first and last position)

- ✓ **Pairing of the reads**

- ✓ **Merging the reads**

  May improve the assembly

Steps of HTS technologies

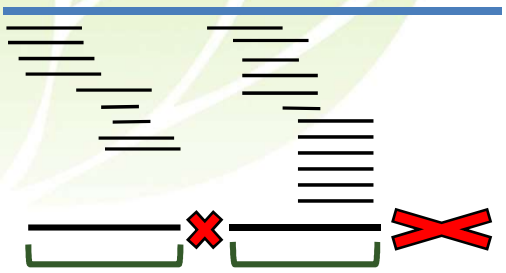


6.3 Identification of targets

From reads to contigs

- ✓ A contig (from contiguous) is a set of overlapping DNA segments that together represent a consensus region of DNA
- ✓ Two options:

1. **Assembly *vs.* reference genome**



6.3 Identification of targets

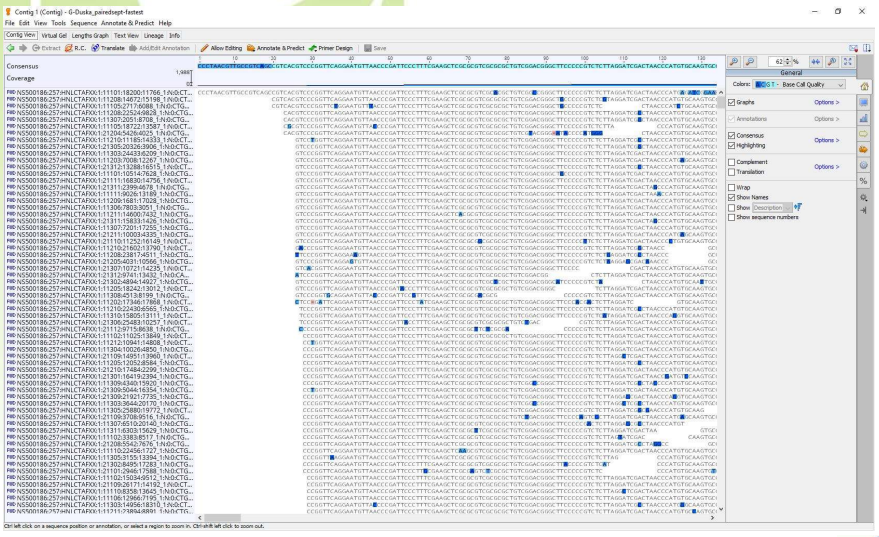
From reads to contigs

✓ Two options:

1. **Assembly *vs.* reference genome**
2. **De novo assembly**

6.3 Identification of targets

From reads to contigs



6.3 Identification of targets

From reads to contigs

De novo assembly	Mapping on genome
+ Reflect real genomes in the sample	+ Quicker (n <-> 1)
- Computing power (n <-> n)	- Bias between strains/cultivars and reference sequences

There is no opposition between both methodologies

6.3 Identification of targets

Grouping the reads together

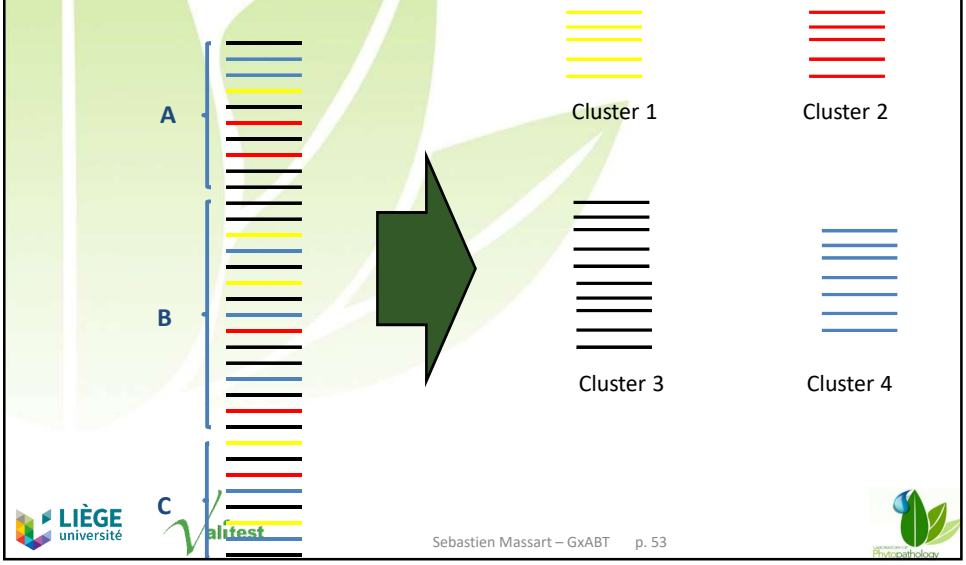
✓ For **amplicon sequencing** : grouping the sequences together based on their identity: example with the sequencing of 4 bacterial species in a mixed population:

- Species 1
- Species 2
- Species 3
- Species 4

6.3 Identification of targets

Grouping the reads together

- ✓ For amplicon sequencing :



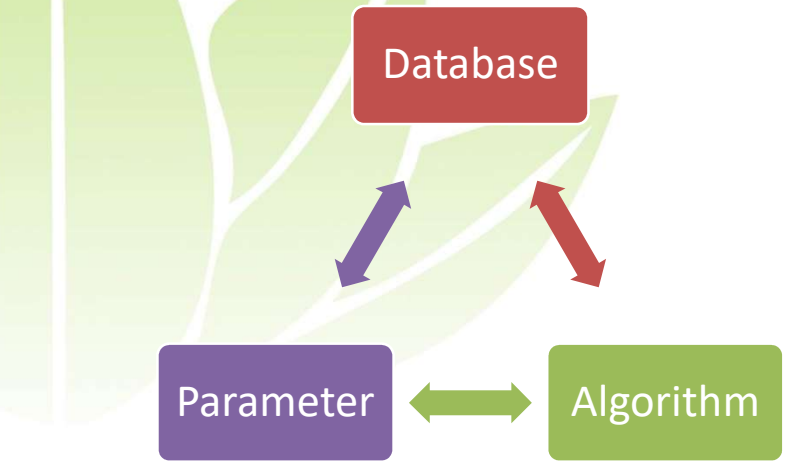
6.3 Identification of targets

Analysing each individual read

- ✓ Computing power improvement
- ✓ Algorithm optimisation
- ✓ Each individual read is identified (Kraken, Bracken, Kodoja…)

6.3 Identification of targets

Comparing with databases



6.3 Identification of targets

Blast algorithm

- Hows does it work ?

1. BLAST scans database for 'words' of a predetermined length (a 'hit') with some minimum threshold parameter, T.
2. BLAST then extends the hit until the score falls below the maximum score yet attained minus some value X.
3. From the 'hit', BLAST will extend the length of homology calculating score of homology for each additional nucleotide or aa and stop when the score is below a threshold
4. BLAST compute the probability that the homology is due to chance (depending on the lenght of the homology, the percentage of the homology and the size of the database)

6.3 Identification of targets

Blast algorithm

Query sequence:
ACTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATAG

Division in words of 5 letters

CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA

In total : 40 words of 5 letters (nt or aa) for the sequence

By default, word lenght is 28 letters (nucleotides)

6.3 Identification of targets

Blast algorithm

- Millions sequences in database: same process -> cut sequences in words of same lenght
- Search for 100% homology

Database sequence:
>SARS-Cov-2 genome sequence for E protein
ACTTGGTCATGTCAG TGTGT ATAATGTCAT

A hit is found !!!

CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA
CTGGCTGTCGTGTGTAAAGCACTAGATAGCTAGATCGATA

It is a first anchoring for a further extended comparison (longer sequence)

6.3 Identification of targets

Blast algorithm

SARS-Cov-2: ACTTGGTCTGTCGTGTGTATA A TGTCAT

Query sequence CTGGCTGTCGTGTGTAAAGCACTAG

High Scoring Segment Pair (HSP) for each additional residue

Calcul of probability score of having by chance this alignment

Stopping when score is going under threshold

6.3 Identification of targets

Blast algorithm

SARS-Cov-2: CTTGGTCTGTCAGTGTGTATA A TGTCAT

Query sequence CTGGCTGTCGTGTGTAAAGCACTAG

Very bad score ?

6.3 Identification of targets

Blast algorithm

SARS-Cov-2: CTTGGTCTGTCAGTGTGTATA A TGTCAT

Query sequence CTGGCTGTC+GTGTGTAAAGCACTAG

Very bad score ?

No ! Insertion/deletion are taken into account

Stopping when score is going under threshold

Key Parameter = e-value

6.3 Identification of targets

Comparing with databases by blast

- Comparison of contigs (reads) *vs*. database

BLAST program	Query	Database
Nucleotide blast (blastn)	Nucleotide	Nucleotide
Protein blast (blastp)	Protein	Protein
Blastx	Translated nucleotide	Protein
tblastn	Protein	Translated nucleotide
tblastx	Translated nucleotide	Translated nucleotide

What are the pros and cons of translation of DNA and comparison with protein databases ?

6.3 Identification of targets

Comparing with databases by blast

For each contig: 110's of hits with databases

Usual presentation of Blast results: best hit for each contig -> OK but…..

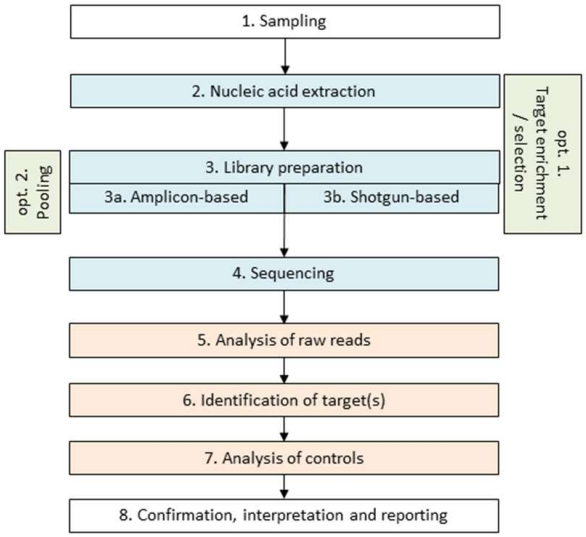
Look behind the tree hiding the forest !!!!

6.3 Identification of targets

Comparing with databases by blast

- Some elements related to Blast :

Steps of HTS technologies



6.4 Biological interpretation

frontiers in Microbiology

PERSPECTIVE
published: 24 January 2017
doi: 10.3389/fmicb.2017.00045

A Framework for the Evaluation of Biosecurity, Commercial, Regulatory, and Scientific Impacts of Plant Viruses and Viroids Identified by NGS Technologies

Sebastien Massart[1], Thierry Candresse[2], José Gil[3], Christophe Lacomme[4], Lukas Predajna[5], Maja Ravnikar[6], Jean-Sébastien Reynard[7], Artemis Rumbou[8], Pasquale Saldarelli[9], Dijana Škorić[10], Eeva J. Vainio[11], Jari P. T. Valkonen[12], Hervé Vanderschuren[13], Christina Varveri[14] and Thierry Wetzel[15]*

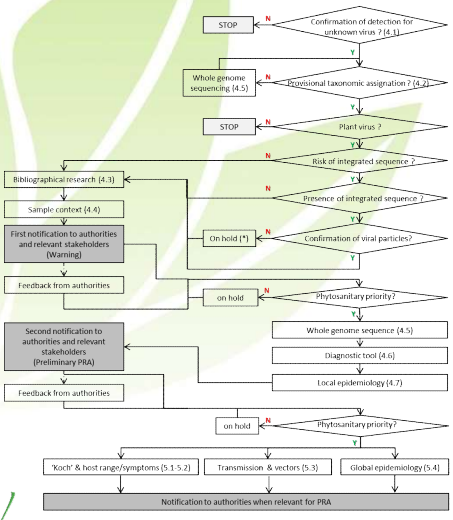
[1] Plant Pathology Laboratory, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium, [2] Institut National de la Recherche Agronomique (INRA), University of Bordeaux, CS20032 UMR 1332 BFP, Villenave d'Ornon, France, [4] Plant Biology, Linnean Centre for Plant Biology, Uppsala BioCentre, Swedish University of Agricultural Sciences, Uppsala, Sweden, [4] Virology and Zoology, Science and Advice for Scottish Agriculture, Edinbourgh, UK, [5] Department of Plant Virology, Institute of Virology, Biomedical Research Center, Slovak Academy of Science (SAS), Bratislava, Slovakia, [6] Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia, [7] Virology, Agroscope, Nyon, Switzerland, [8] Division Phytomedicine Lentzeallee, Faculty of Life Sciences, Albrecht Daniel Thaer-Institute of Agricultural and Horticultural Sciences, Humboldt-University of Berlin, Berlin, Germany, [9] National Research Council Institute for Sustainable

OPEN ACCESS

Edited by:
David Gilmer,
University of Strasbourg, France

6.4 Biological interpretation

Figure 2: Framework proposal following discovery of new viruses. Y means positive response and N negative response. The numbers between brackets correspond to the chapters in the text. (*): the absence of viral particle should be recorded and further experiments carried out if new cases are discovered



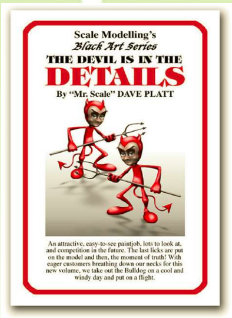
LIÈGE université

Valitest

7. Conclusions

LIÈGE université

1. Reviewing all the steps at both lab and bioIT levels
2. Diversity of protocols causing all bias



You can miss a pest in the sample because of library preparation method, sequencing depth, bioinformatic software and parameters !!

1. Reviewing all the steps at both lab and bioIT levels
2. Diversity of protocols causing all bias
3. Standardisation of protocols
   - Library preparation (Illumina kits) :

1. Reviewing all the steps at both lab and bioIT levels
2. Diversity of protocols causing all bias
3. Standardisation of protocols
   - Library preparation (Illumina kits)
   - Bioinformatic (smallRNA):

METHODOLOGY ARTICLE | OPEN ACCESS

An internet-based bioinformatics toolkit for plant biosecurity diagnosis and surveillance of viruses and viroids

Roberto A. Barrero†, Kathryn R. Napier†, James Cunnington, Lia Liefting, Sandi Keenan, Rebekah A. Frampton, Tamas Szabo, Simon Bulman, Adam Hunter, Lisa Ward, Mark Whattam and Matthew I. Bellgard
† Contributed equally

BMC Bioinformatics BMC series – open, inclusive and trusted 2017 **18**:26 | https://doi.org/10.1186/s12859-016-1428-4

Received: 12 July 2016 | **Accepted:** 15 December 2016 | **Published:** 11 January 2017

3. Standardisation of protocols:

Thank you for your attention!

E-mail:

sebastien.massart@uliege.be

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 773139