Towards Safer Plant Genetic Resources through improved viral diagnostics (SafePGR). Final report

Claude Pavis, Thierry T. Candresse, Pierre-Yves Teycheney, Philippe Roumagnac

1st call for projects

Final report

*This document is to be filled out by the coordinator in collaboration with the project partners. It must be sent by the coordinator, within the 2 months after the official end of the project with a copy to all the funding bodies of the project. It reports on the activity of **all the project partners**. All the partners must have a copy of the version sent to the JCS.*

SafePGR

A IDENTIFICATION

Project acronym	SafePGR
Project title	Towards Safer Plant Genetic Resources through improved viral diagnostics
Project identification agency numbers if existing (add a line by agency)	Région Guadeloupe: CR/12-16 and CR/12-7 ANR: ANR-11-EBIM-0002-06
Coordinator (company/organization, country/region)	INRA Guadeloupe
Coordinator of the French part of the project (company/organization)	INRA Guadeloupe
Project period (Start date – End date)	1st March 2012 – 31 August 2015
Project website (if applicable)	http://www2.antilles.inra.fr/safepgr/

Author of this report	
Title, first name, surname	Dr Claudie Pavis, with the work package leaders
Telephone	+590 590 25 59 02
E-mail	claudie.pavis@antilles.inra.fr
Date of writing	October 2015

List of partners involved at the end of the project. One line/partner (company/organization, country/region and principal investigator)	INRA Guadeloupe, Dr C. Pavis CIRAD Guadeloupe, Dr P.Y. Teycheney CIRAD Montpellier, Dr Ph. Roumagnac CIRAD Reunion, Dr M. Grisoni INRA Bordeaux, Dr T. Candresse University of Azores, Dr D. Mendonça University of Madeira, Dr M. Carvalho

B CONSOLIDATED PUBLIC SUMMARY

This summary is intended to be disseminated to a large audience in order to promote the project results. Therefore no confidential results should be mentioned. The terminology used should be adapted but without excluding technical words. A funding agency's language and English version of the summary must be provided. It is necessary to follow the instructions detailed below.

B.1 INSTRUCTIONS FOR CONSOLIDATED PUBLIC SUMMARIES

Public summaries must be structured according to the following guidelines:

Catchy title of the project *(about 80 characters including spaces)*
Catchy title, if possible striking and concise, which sums up and explains your project keeping in mind that the target is a large audience: an exhaustive introduction of the project is not necessary, but you should rather insist on the striking aspects of the project.

The two first paragraphs are preceded by a title specific to the project.

Title 1: shows the general objective of the project and the main issues raised *(150 characters max including spaces)*
Paragraph 1: *(about 1200 characters including spaces)*
Paragraph 1 specifies the project objectives and challenges: indicate the context, the general objective, the issues addressed, the solutions explored, the future prospects and outcomes at the technical and /or societal levels.

Title 2: shows the methods or technologies used *(150 characters max including spaces)*
Paragraph 2: *(about 1200 characters including spaces)*
Paragraph 2 indicates how the expected results are obtained according to certain methods and/or technologies. The technologies and/or methods used to overcome bottlenecks are explained (avoid using scientific jargon, acronyms or abbreviations)

Project main results *(about 600 characters spaces including)*
Project highlights to be disseminated to a large audience. Explain the possible applications and/or usages, the original research and/or development tracks, which might have not been foreseen from the beginning.
Also mention any other outcome = international partnerships, new opportunities, new contracts, start-ups, research synergies, competitiveness clusters, etc.

Scientific production and patents since the beginning of the projects *(about 500 characters spaces including)*
Do not only list but add a few comments. You may also indicate the standardization actions.

Illustration
Illustrate with a scheme, graph, picture or photo with a brief legend. The illustration must be clearly readable (size: width: about 6cm / height: about 5 cm). You need sufficient resolution for printing. Only provide illustrations for which you own the rights.

Factual information
Draw up a sentence specifying the type of project (industrial research, fundamental research, experimental development, exploratory, innovation, etc.), the coordinators, the partners, the real starting date and the duration of the project, total and national/regional grants and the project total budget, for instance: "The XXX project is a fundamental research project coordinated by xxx in "Country/region" and by xxx in xxx. It associated xxx, as well as the xxx and xxx laboratories. The project started on … and lasted… months. grant amounted to …€ and xxx grant amounted to xxx€ for a total budget of… €." (give a table with details of grants per funding bodies)

B.2 CONSOLIDATED PUBLIC SUMMARY IN FRENCH, PORTUGUESE OR SPANISH (FUNDING BODIES LANGUAGES)

SafePGR, ou comment limiter les risques de diffusion des maladies virales par échange de ressources génétiques végétales tropicales

Décrypter la diversité virale présente dans les collections de plantes tropicales, et optimiser les outils de diagnostic

Les Centres de Ressources Biologiques (CRB) conservent et distribuent des ressources biologiques végétales à des fins de recherche et développement. Ils fournissent aux programmes d'amélioration variétale des géniteurs, qui sont la clé des adaptations aux changements sociaux et environnementaux actuels. Pour éviter la dispersion ou l'émergence de maladies, les CRB doivent garantir le bon état sanitaire des ressources qu'ils distribuent. Les CRB de Guadeloupe, Madère, les Açores et la Réunion conservent des collections de bananier et plantain, canne à sucre, igname, patate douce, ail tropical et vanille. Ces plantes se multiplient par voie végétative, ce qui induit l'accumulation de virus. L'absence de reproduction sexuée empêche un assainissement naturel, du fait que la plupart des virus ne sont pas transmis par les graines. Par ailleurs, les connaissances sur les virus affectant ces plantes sont très partielles.

Des méthodes d'assainissement existent et permettent de produire des plants indemnes de virus, mais elles nécesssitent de disposer de tests de diagnostic sensibles, polyvalents et fiables. L'objectif général de SafePGR était de renforcer les connaissances sur la diversité des virus affectant les espèces ciblées par les CRB, pour mettre au point des stratégies de diagnostic efficaces. Ceci permettra à terme d'envisager des mouvements sécurisés de ces ressources entre les partenaires, et au-delà.

Combiner les méthodes de biologie moléculaire classique, et les approches de séquençage à haut débit pour explorer la diversité virale

Pour atteindre ses objectifs, le projet a combiné les méthodes de biologie moléculaire classique, et les approches de séquençage à haut débit (NGS), ce qui a conduit à des découvertes de virus sans précédent au sein des plantes étudiées. Les méthodes basées sur les NGS ont l'avantage de ne nécessiter aucune information préalable et ne sont pas ciblées sur des virus particuliers. Elles permettent de mettre en évidence un très large éventail de virus, même dans le cas d'agents viraux inconnus. Les approches NGS, également appelées approches métagénomiques, sont basées sur le dépistage aléatoire des viromes de plantes. Du fait qu'elles permettent de détecter des virus non reconnus par les méthodes classiques utilisant les séquences virales comme cibles, ces approches sont complémentaires des méthodes basées sur la PCR, largement utilisées pour les investigations légales et le diagnostic. Les approches innovantes mises au point dans SafePGR ont rendu possible des découvertes d'espèces virales sans précédent chez les plantes étudiées, à partir desquelles des amorces spécifiques ont ensuite été dessinées. Ceci a conduit *in fine* à un dépistage systématique, optimisé et efficace, des plantes conservées au sein des CRB partenaires.

La partie cachée de l'iceberg : la diversité virale cachée dans les collections de plantes tropicales

Au total, 21 nouvelles espèces virales ont été découvertes et leur diversité moléculaire explorée. Ceci a conduit à la mise au point de méthodes de diagnostic pour ces nouveaux agents, ainsi que l'optimisation du diagnostic pour 10 virus déjà connus. Ainsi, le statut viral des collections est maintenant beaucoup mieux connu. Les équipes ont désormais une base solide pour définir leurs stratégies d'assainissement et de diagnostic, et pour relancer les processus associés. Ces résultats sont d'ores et déjà transférés sur la collection d'ignames en Guadeloupe, et vont contribuer à appuyer un projet sur la mise en place d'une filière de plants de qualité, avec les acteurs économiques.

Production scientifique et brevets issus du projet

- Quatre publications et un chapitre d'ouvrage ont été acceptés dans des revues à comité de lecture. Ils traitent de la découverte de virus et des avancées de la méthode métagénomique. Six autres publications sont en préparation. C'est une contribution importante dans le champ de la virologie et de la métagénomique.
- Les séquences du génome complet de 8 virus identifiés ont été intégrées dans GenBank, dont les séquences de 7 génomes complets du Sugarcane white streak virus et la séquence du génome complet du Yam virus X. De plus, 16 séquences partielles du gène de la RNA-dependent RNA polymerase du Yam virus X ont été déposées dans GenBank.
- Des protocoles nouveaux, ou optimisés, pour le diagnostic sont disponibles sur le site wew du projet http://www2.antilles.inra.fr/safepgr/
- Un pipeline a été adapté pour les analyses bio-informatiques, et peut être transféré aux équipes travaillant sur des jeux de données similaires.

SafePGR: como limitar o risco de dispersão de doenças virais pelas trocas de germoplasma de plantas tropicais

Determinação da diversidade viral em germoplasma de plantas tropicais conservado e aperfeiçoamento das ferramentas de diagnóstico

Os Centros de Recursos Biológicos (CRBs) conservam e distribuem germoplasma vegetal para investigação e desenvolvimento. São eles que providenciam aos programas de melhoramento genético genitores que são essências para a adaptação das culturas às alterações ambientais e sociais em curso. A fim de evitar a disseminação ou emergência de doenças, os CRBs devem garantir o estado sanitário dos recursos que distribuem. Os CRBs de Guadalupe, Madeira, Açores e Reunião conservam germoplasma de banana, inhame (Dioscorea), batata-doce, cana-de-açúcar, alho e baunilha. Estas culturas são propagadas vegetativamente e propensas à acumulação de infeções virais devido à inexistência de reprodução sexuada que normalmente atua como processo de saneamento, já que muitos dos vírus vegetais não são transmitidos por semente. Também, o nosso conhecimento dos vírus que infetam estas culturas é apenas parcial.

Existem métodos de saneamento que permitem a obtenção de plantas isentas de vírus, mas requerem testes de diagnose sensíveis, polivalentes e fiáveis. Os objetivos gerais do projeto SafePGR foram melhorar a conhecimento da diversidade dos vírus que infetam as culturas abordadas pelos CRBs dos parceiros com o fim de desenvolver ou otimizar técnicas de diagnóstico, permitido em última instância a circulação segura de plantas entre os parceiros e para além destes.

Combinação de abordagens de biologia molecular clássica e de sequenciação de última geração para explorar a diversidade viral

Para alcançar os seus objetivos o projeto combinou abordagens de biologia molecular clássica e de sequenciação de última geração (NGS) conduzindo a uma descoberta inédita de vírus nas culturas alvo. A utilização de abordagens baseadas em NGS têm a vantagem de não assumir qualquer informação prévia e, portanto, os genomas virais são obtidos de um modo muito amplo, mesmo no caso de novos agentes. Estas abordagens baseadas em NGS (também designada por abordagem metagenómica) apoiam-se no rastreio aleatório do viroma vegetal. Uma vez que permitem a deteção de vírus que escapariam à deteção por métodos clássicos que têm como alvo sequencias virais conhecidas, estas abordagens são complementares às tradicionais baseadas em PCR que são amplamente utilizadas em ciência forense e diagnóstico. As abordagens inovadoras desenvolvidas no projeto SafePGR permitiram uma descoberta sem precedentes de vírus nas culturas alvo de estudo , a partir das quais primers específicos foram desenvolvidos, conduzindo em suma a um rastreio sistemático, otimizado e eficiente das plantas conservadas nos CRBs parceiros.

Sob a ponta do iceberg: a diversidade viral escondida no germoplasma vegetal tropical conservado

Um total de 21 novas espécies de vírus foram descobertas e analisada a sua diversidade molecular, conduzindo ao estabelecimento de métodos de diagnose para estes novos agentes, bem como à otimização de métodos para 10 vírus já conhecidos. Isto permitiu um melhor conhecimento do estado das coleções de germoplasma relativamente a infeções virais. As equipas têm agora uma boa base para determinar as estratégias de saneamento e diagnostico, e reforçar os processos associados. Os resultados estão a ser transferidos para a coleção de inhame de Guadalupe e a beneficiar um projeto do sector da produção de sementes de qualidade, envolvendo os agentes económicos.

Produção científica e patentes desde o início do projeto

- Cinco artigos aceites em revistas com arbitragem, científica relativos à descoberta de vírus, e outros 6 estão em preparação. Este é um importante contributo no campo da virologia e metagenómica de plantas.
- Sequências completas dos genomas de 8 vírus identificados foram integradas no GeneBank, incluindo 7 sequências inteiras do genoma de Sugarcane white streak virus e uma sequência da totalidade do genoma de Yam virus X. Adicionalmente, 16 sequências parciais do gene da ARN polimerase dependente de ARN do Yam virus X foram depositadas o GenBank. Protocolos de diagnóstico novos e otimizados estão disponíveis na página web do projeto, http://www2.antilles.inra.fr/safepgr/.

SafePGR: how to limit the risk of spreading viral diseases through the exchange of tropical plant germplasm

Deciphering viral diversity in conserved tropical plant germplasm and refining diagnostic tools

Biological Resources Centres (BRCs) conserve and distribute plant germplasm for research and development purposes. They provide breeding programs with genitors that are critical for crop adaptation to ongoing environmental and societal changes. In order to prevent the spread or emergence of diseases, BRCs must guarantee the sanitary status of the resources they distribute. Guadeloupe, Madeira, Azores and Reunion BRCs conserve banana and plantain, sugarcane, yam, sweet potato, garlic and vanilla germplasm. These crops are vegetatively propagated and are prone to the accumulation of viruses, due to the lack of sexual reproduction, which would act as a natural sanitation process since most plant viruses are not seed transmitted. Our knowledge of the viruses infecting these crops is also only partial.

Sanitation methods exist for recovering virus-free plants but they require sensitive, polyvalent and reliable diagnosis tests. The general objective of the SafePGR project was to improve the knowledge of the diversity of viruses infecting the crops addressed by the partner's BRCs, in order to develop or optimize diagnostic techniques, ultimately permitting the safe movement of plants between project partners and beyond.

Combining classical molecular biology and next generation sequencing approaches to explore viral diversity

To reach its objective, the project combines classical molecular biology and next generation sequencing (NGS) approaches, leading to unprecedented virus discovery in the targeted crops. The NGS-based approaches have the advantage of not assuming any prior information and of therefore very broadly targeting viral genomes, even in the case of novel agents. These NGS-based approaches (so-called metagenomics approaches) are based on random screens of the plant viromes. Because they allow detecting viruses that would have escaped classical detection methods that use known viral sequences as targets, these approaches are complementary to the traditional PCR-based approaches that are widely used in forensics and diagnostics. The innovative approaches developed in the SafePGR project have enabled unprecedented virus discovery in the target crops, from which specific primers were further designed, leading in fine to a systematic, optimized and efficient screening of the plants maintained in the partners BRCs.

Beneath the tip of the iceberg: the hidden viral diversity of the conserved tropical plant germplasm

A total of 21 new virus species were discovered and their molecular diversity was explored. This led to the establishment of diagnosis methods for these new agents as well as to the optimization of diagnosis for 10 already known viruses. This allowed improving the knowledge of viral status of the germplasm collections. The teams have now a good basis to determine their sanitation and diagnostic strategies, and to revive the associated processes. The results are now transferred on yam collection in Guadeloupe, and are irrigating a project of a quality seed production sector, with economic actors.

Scientific production and patents since the beginning of the projects (about 500 characters spaces including)

- Four papers and one book chapter have been accepted in peer-review journals, concerning virus discovery, and metagenomics. About 6 other are in preparation. This is an important contribution in the field of virology, and plant metagenomics.
- The complete genome sequences of 8 identified viruses have been integrated in GenBank, including 7 full genome sequences of Sugarcane white streak virus and one full genome sequence of Yam virus X. In addition, 16 partial sequence of Yam virus X RNA-dependent RNA polymerase gene were deposited in GenBank.
- New and optimized diagnostic protocols are available in the project website, http://www2.antilles.inra.fr/safepgr/
- A pipeline has been adapted for bioinformatic analyses, and may be transferred to teams working on similar data sets.

Training for metagenomic methods - Intermediate meeting in Cirad Montpellier, October 2013 © C. Pavis.

Factual information

The SafePGR project is a fundamental research project coordinated by INRA in Guadeloupe. It associated Biological Resources Centres (CRB Plantes tropicales INRA-CIRAD Guadeloupe, Isoplexis University Madeira, CBA-UAc Azores, VATEL INRA La Réunion) as well as virology labs (INRA BFP Bordeaux and CIRAD BGPI Montpellier). The project started on February 2012 and lasted 42 month. Grants amounted to 553 227 € for a total budget of 1 444 121 €.

	Funding agencies (amounts in €)							
PARTNERS	Région Guade-loupe	ANR	Région Réunion	Governo dos Açores	Regiao da Madeira	**Granted by agencies**	Granted by partner	Total budget
ASTRO INRA Guadeloupe	91 000					**91 000**	152 432	243 432
AGAP CIRAD Guadeloupe	53 000					**53 000**	123 120	176 120
BGPI CIRAD Montpellier		180 000				**180 000**	306 749	486 749
PVBMT CIRAD La Réunion			33 000			**33 000**	93 260	126 260
BFP INRA Bordeaux		100 000				**100 000**	102 076	202 076
CBA-UAc Universidad Açores				51 045		**51 045**	38 499	89 544
ISOPlexis Universidad Madeira					45 182	**45 182**	74 761	119 943
Total	144 000	280 000	33 000	51 045	45 182	**553 227**	890 897	1 444 121

C SCIENTIFIC REPORT

***Maximum 5 pages**. Provide here indications on the possible content of the report. This report may be accompanied by more detailed appendices.*

The scientific report covers the whole period of the project. It must produce an auto-sufficient synthesis reminding the objectives, the work achieved and the obtained results against the initial expectations and the state of the art. The format is similar to that of the scientific papers or research monographs. It must reflect the collective character of the effort made by the partners throughout the project. The coordinator prepares this report on the basis of the contributions from all partners. A scientific report reported as confidential will not be disseminated. Briefly justify the reason of the requested confidentiality. Non confidential reports will be likely disseminated by the funding bodies, in particular via the open archives (as an example for ANR: http://hal.archives-ouvertes.fr.)

Confidential scientific report: no

C.1 REPORT SUMMARY

This summary may be taken from the consolidated public summary.

SafePGR: how to limit the risk of spreading viral diseases through the exchange of tropical plant germplasm

Deciphering viral diversity in conserved tropical plant germplasm and refining diagnostic tools

Biological Resources Centres (BRCs) conserve and distribute plant germplasm for research and development purposes. They provide breeding programs with genitors that are critical for crop adaptation to ongoing environmental and societal changes. In order to prevent the spread or emergence of diseases, BRCs must guarantee the sanitary status of the resources they distribute. Guadeloupe, Madeira, Azores and Reunion BRCs conserve banana and plantain, sugarcane, yam, sweet potato, garlic and vanilla germplasm. These crops are vegetatively propagated and are prone to the accumulation of viruses, due to the lack of sexual reproduction, which would act as a natural sanitation process since most plant viruses are not seed transmitted. Our knowledge of the viruses infecting these crops is also only partial.

Sanitation methods exist for recovering virus-free plants but they require sensitive, polyvalent and reliable diagnosis tests. The general objective of the SafePGR project was to improve the knowledge of the diversity of viruses infecting the crops addressed by the partner's BRCs, in order to develop or optimize diagnostic techniques, ultimately permitting the safe movement of plants between project partners and beyond.

Combining classical molecular biology and next generation sequencing approaches to explore viral diversity

To reach its objective, the project combines classical molecular biology and next generation sequencing (NGS) approaches, leading to unprecedented virus discovery in the targeted crops. The NGS-based approaches have the advantage of not assuming any prior information and of therefore very broadly targeting viral genomes, even in the case of novel agents. These NGS-based approaches (so-called metagenomics approaches) are based on random screens of the plant viromes. Because they allow detecting viruses that would have escaped classical detection methods that use known viral sequences as targets, these approaches are complementary to the traditional PCR-based approaches that are widely used in forensics and diagnostics. The innovative approaches developed in the SafePGR project have enabled unprecedented virus discovery in the target crops, from which specific primers were further designed, leading in fine to a systematic, optimized and efficient screening of the plants maintained in the partners BRCs.

Beneath the tip of the iceberg: the hidden viral diversity of the conserved tropical plant germplasm

A total of 21 new virus species were discovered and their molecular diversity was explored. This led to the establishment of diagnosis methods for these new agents as well as to the optimization of diagnosis for 10 already known viruses. This allowed improving the knowledge of viral status of the germplasm collections. The teams have now a good basis to determine their sanitation and diagnostic strategies, and to revive the associated processes. The results are now transferred on yam collection in Guadeloupe, and are irrigating a project of a quality seed production sector, with economic actors.

C.2 CHALLENGES AND ISSUES, STATE OF THE ART

Introduce the initial stakes of the project, the issues addressed by the project, and the state of the art on which the project refers. Outline the possible evolutions throughout the project duration.

It is widely acknowledged that viral diversity is underestimated (1, 2). This is especially true of plant viruses, which have always attracted less attention -and funding- than animal and human viruses, and even more of viruses infecting tropical crops, which are generally less studied than temperate crops. This lack of knowledge limits the efforts made towards the safe movement of germplasm, because many viruses infecting crops have not yet been identified, thus preventing the development of efficient and accurate tools for their detection. Likewise, for known viruses, lack of information on their genetic diversity hampers our ability to detect all isolates of a given virus. This situation has serious consequences on agriculture and food security worldwide, because it is likely responsible for a large fraction of novel threats such as the emergence and/or spread of viral diseases (3).
The situation is particularly critical for vegetatively propagated plants, which do not benefit from the partial sanitation brought by the natural elimination of non-seed-borne viruses through sexual reproduction. As a consequence, viral diseases particularly affect vegetatively propagated crops, which represent the majority of crops cultivated in European outermost regions. Major programmes have been undertaken in European countries to secure the wide genetic diversity of these crops and to promote their safe conservation and distribution through the establishment of Biological Resource and Quarantine or clearance Centres. These efforts are jeopardized by our limited diagnostic capability and by the risk of spreading viruses through the distribution and exchange of infected germplasm.

C.3 SCIENTIFIC AND TECHNICAL APPROACH

The project tackled the issue of the molecular diversity of plant viruses present in the partners BRCs and germplasm collections. To this aim, large-scale studies of this diversity were undertaken in the six crops targeted by the project (banana, sugarcane, sweet potato, garlic, vanilla and yam), using Polymerase Chain Reaction (PCR)-based and Next Generation Sequencing (NGS)-based approaches. Bioinformatic pipelines were developed for the analysis of the large sequence datasets generated by the NGS approaches. The information generated was used to fully or partially characterize the novel viruses identified and develop or optimize, within Work Package 1, diagnostic methods adapted to the diversity of viral populations identified either through classical molecular approaches or through the novel, NGS-based metagenomics approaches.

When the project was initiated (2012), very few viral metagenomics studies had been undertaken and no "golden" metagenomics approach was already adopted by the community of plant and animal virologists (1, 4). The first goal of the work package 2 (WP2_Task 4) of the project was methodological and dealt with the test of several metagenomics protocols that were academically or commercially emerging in the literature at the time. Eight NGS protocols (see annex 3) were used for the detection of plant viruses in all six plant-models studied in the project. Information about the known viruses present within a set of 96 test plants (16 plants for each of the six plant species) grown in the BRCs or quarantine greenhouses was developed. This information, which relied on a range of serological and molecular techniques, was then used to confirm the efficiency of the 8 metagenomics protocols together with data collected in the frame of WP1. Bioinformatics tools (see annex 5) developed in WP3 allowed to begin the comparison of the efficiency of the eight different NGS-based approaches tested. All partners collectively decided during the intermediate meeting to select two metagenomics approaches among the 8 tested protocols, namely the purification of dsRNA using CF11 cellulose chromatography (method developed by INRA Bordeaux, see Protocol #7 in annex 3) and the purification of viral particles (method developed by CIRAD Montpellier, see Protocol #4 in annex 3) for processing more than 1500 plants collected from germplasm collections maintained by the participating BRCs and CIRAD's sugarcane quarantine (WP2_Task 5). This choice was done because results obtained during the initial phase showed these two methods to be complementary in their advantages and detection range (see annex 4): comparison of 38 infected control plants) and enabled making an efficient inventory of plant viruses present in each tested plant.

A training session was organized in Montpellier during the intermediate meeting in order to transfer to all partners the so-called "Protocol #7: Bordeaux dsRNA purification method". This training aimed at transferring to all partners the method in order to perform it in each of the partner's labs (Reunion, Azores, Madeira, Montpellier, Bordeaux and Guadeloupe).

Unprocessed samples (particles purification approach, which was performed in Montpellier) or purified dsRNAs were then centralized in Montpellier to perform the random amplification step needed before the NGS sequencing, which was performed by a commercial company. Results were analysed in Bordeaux and in Montpellier using the locally developed pipelines (see Annex 6).

C.4 ACHIEVED RESULTS

Measure the results in relation to the project deliverables and publications, patents, etc. Revisit the state of the art and the stakes at the end of the project.

The broad inventory of interspecific and intraspecific viral diversity was undertaken in germplasm collections maintained by the participating BRCs and CIRAD's sugarcane quarantine. To this aim, classical approaches using PCR and two innovative approaches based on high throughput NGS were designed and implemented on plant samples from these germplasm collections in an unprecedented survey.

A total of 1526 plants were processed using the viral particle purification at CIRAD Montpellier, including 56 plants from Azores (20 banana, 2 garlic and 34 sweet potato), 740 plants from Guadeloupe (190 banana, 300 sugarcane and 250 yam), 100 from Madeira (36 banana and 64 sweet potato), 357 from Montpellier (305 sugarcane, 2 sweet potato and 50 yam), and 273 from Reunion Island (45 garlic, 19 sweet potato, 199 vanilla and 10 yam). In addition, 716 plants among the set of 1526 plants were processed using the dsRNA CF11 purification method (215 were processed in Reunion Island, 36 in Madeira, 52 in Azores, 25 in Montpellier and 388 in Guadeloupe). Five 454 Roche plates were eventually run.

NGS-based approaches and datamining of expressed sequence tags (ESTs, Task 3.1 of WP3) lead to the complete or partial genome sequencing of 21 new virus species (1 in banana, 3 in garlic, 3 in sweet potato, 4 in sugarcane, 3 in vanilla and 7 in yam) (Annex 1). Based on these data, degenerate primers were designed and used to screen collections maintained in Guadeloupe, Reunion, Madeira and Azores and in the sugarcane quarantine at Montpellier. Nucleic acids were extracted from a total of 1 000 samples and used in a total of 3 939 PCR-based analyses (WP1, Task XXX). Details of the analysed sample counts are provided in Annex 2.

Overall, results show that:

- Very few virus-infected **banana** plants are present in germplasm collections conserved in Guadeloupe, Madeira and the Azores. Concerning the Guadeloupe collection, a possible explanation for this situation is that it underwent sanitation 5 years ago and was replanted from virus-free vitroplants. The same situation is occurring in Madeira, whereas filed collection and plants snow by farmers are produced from vitroplants.

- The **garlic** germplasm maintained in Reunion Island is frequently infected by potyviruses, carlaviruses, allexiviruses and a foveavirus (over 50% prevalence of each virus group). Most accessions are co-infected by two or more viruses.

- In **sweet potato**, the situation differs between Guadeloupe, where a significant proportion of analyzed plants are infected by begomoviruses (although only 18 plants were indexed) and Madeira and the Azores where plants are mostly infected by potyviruses and a few plants are infected by carlaviruses, and Reunion where none of these viruses were detected on the 19 sweet potato accessions tested. All 60 plants analyzed from Guadeloupe and Azores gave a positive signal when indexed for a potential new nucleorhabdovirus, but several lines of evidence suggest that this agent is likely to be an endogenous, integrated viral remnant rather than a replicating infectious agent.

- In **sugarcane**, important prevalence levels were observed in the Guadeloupe collection for a Badnavirus (*Sugarcane bacilliform virus*) and a yet unknown Ampelovirus. In addition, several new geminiviruses and associated satelites were detected in plants from Guadeloupe and from the sugargane quarantine in Montpellier. One of those new geminiviruses is a novel mastrevirus (*Sugarcane white streak virus*; SWSV), which seemingly had escaped detection by all routine quarantine detection assays that were performed so far. Encouragingly, sanitation by meristem-tip culture proved to be efficient to obtain SWSV-free plants.

- In **vanilla**, the prevalence of the new potexvirus discovered by NGS was relatively low in the accessions originating from two vanilla collections of Reunion Island.

- Several new viral species, in the *Macluravirus*, *Ampelovirus*, *Potexvirus* and *Sadwavirus* genera were identified in **yam** in the Guadeloupe collection, with a majority of plants being infected simultaneously by several viruses. In Reunion a Potyvirus (YMMV) and a Badnavirus (DBALV) infect yams.

C.5 EXPLOITATION OF RESULTS

The results obtained during the SafePGR project confirm that vegetatively propagated crops conserved in germplasm collections in the European outermost regions show significant prevalence levels for multiple viruses, including a number of novel viruses discovered and characterized in the frame of the project. Sanitation of the infected accessions is a prerequisite to their movement, either within the regions where they are conserved or beyond. The development or optimization of diagnostic methods for both known and novel agents infecting the six target crops will facilitate these sanitation efforts. Some of these methods have already been successfully implemented in the frame of the yam and sugarcane sanitation programs carried out in Guadeloupe and Montpellier, respectively.

C.6 DISCUSSION

Discuss on the level of execution of the initial objectives, the bottlenecks that are still to be overcome, the breakthroughs, the possible extensions, the prospects offered by the project, the scientific, industrial or societal impact of the results.

The initial objectives were fully executed. More than 800 plants maintained in Guadeloupe, Reunion, Azores, Madeira BRCs and the Montpellier sugarcane quarantine were screened for the presence of the viruses that were already known and characterized at the beginning of the project in 2012. In addition, these plants were screened for the presence of the 21 new virus species that were discovered during this 3-year project. Overall this project that initially aimed at detecting the presence of about 20 known viruses ended up with the diagnosis of 40 viruses, 21 of which are novel agents discovered using the NGS-based approaches and bioinformatics pipelines developed in the frame of the SafePGR project. Noteworthy, this polyphasic approach combining bioinformatics, metagenomics and classical molecular methods not only provided new diagnostics tools for the detection of novel viruses but also supplied optimized diagnosis for 10 already known viruses.

One of the objectives was to screen 1500 plants for the presence of viruses using a metagenomics approach. Our first methodological optimization step highlighted that two methods were efficient for describing a large part of the virome of the tested plants. However, both methods failed to detect some viruses, the dsRNA CF11 purification method failing to detect DNA viruses and the VANA approach failing sometimes to detect some RNA viruses. Consequently, we have decided during the intermediate meeting to use these two complementary methods in an effort to improving our virus detection threshold. A training session was set up during the intermediate meeting for transferring the dsRNA CF11 purification method to all SafePGR partners. Consequently, 800 plants were processed overseas, each partner processing in their own laboratory part of the 800 plants using the dsRNA CF11 purification method. Overall, the results obtained from these overseas experiments were however poor, suggesting that it remains difficult to organize in a short time inter-laboratory metagenomics studies. However, each partner knows now how to use the dsRNA CF11 purification method and can eventually repeat the experiment and remedy the problems encountered.

Our study also shed light on the fact that the metagenomics approaches need further efforts if they are to be used for cataloging the viromes of individual plants from a purely diagnostic perspective (5). The use and application of metagenomics-based diagnostic is still limited by issues of costs, portability and standardization. In particular, standardization of detection sensitivity thresholds is a very important issue since some degree of background sequence contamination is likely unavoidable for such sensitive methods (5).

However, our project with its methodological and bioinformatics developments paves the way toward the use of metagenomics-based diagnostics. We still have effort to make for improving the portability, increasing the degrees of sensitivity to true positives and decreasing the degrees of sensitivity to contamination. We show in this project that combining classical molecular methods and NGS-based approaches still remain the most efficient and reliable approach for cataloging and detecting plant viruses.

C.7 CONCLUSIONS

Besides providing with outstanding academic results, that will be converted into several scientific articles, the integrated and innovative approach set-up in this project that combined bioinformatics, metagenomics and classical molecular methods has also provided new diagnostics tools for the detection of novel viruses but also supplied optimized diagnosis for already known viruses.

Overall, for the six targeted crops, we have more than doubled the number of viruses for which reliable detection assays are available, which will undeniably facilitate sanitation efforts. Noteworthy, some of the sanitation methods have been already successfully implemented in the frame of the yam and sugarcane sanitation programs carried out in Guadeloupe and Montpellier, respectively, which will limit the risk of spreading viruses through the distribution and exchange of infected germplasm.

C.8 REFERENCES

1. **Roossinck MJ, Martin DP, Roumagnac P.** 2015. Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology**:**PHYTO12140356RVW.
2. **Wren JD, Roossinck MJ, Nelson RS, Scheets K, Palmer MW, Melcher U.** 2006. Plant virus biodiversity and ecology. PLoS Biol **4:**e80.
3. **Anderson PK, Cunningham AA, Patel NG, Morales FJ, Epstein PR, Daszak P.** 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. Trends Ecol Evol **19:**535-544.
4. **Roossinck MJ, Saha P, Wiley G, Quan J, White J, Lai H, Chavarria F, Shen G, Roe B.** 2010. Ecogenomics: using massively parallel pyrosequencing to understand virus ecology. Mol Ecol **19:**81-88.
5. **Massart S, Olmos A, Jijakli H, Candresse T.** 2014. Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res **188:**90-96.

D LIST OF DELIVERABLES

When applicable for the project, reproduce the table of deliverables provided at the start of the project. Indicate all the deliverables, including any deliverables deleted or added with respect to the initial list.

Date of supply	**N°**	**Title**	**Nature** (report, software, prototype, date, etc.)	**Partners** (underline the responsible partner)	**Comments**
WP 0 - Coordination					
03/2012	D0-1	A follow-up project website	Website	P1	
TTP	D0-2	Communication supports (talks, web pages, press release)	Pages and files in the website	All P1	
06/2012	D0-3	The consortium agreement	Document	All P1	
07/2013	D0-5	The kick-off meeting report	Report	All P1	
10/2013	D0-6	The mid-report	Report	All P1	
10/2015	D0-7	The final report	Report	All P1	
2013 to 2015	D0-8	Joint scientific and technical publications	Publications	All	
WP 1 - Viral diversity					
Under realisation	D1-1	A catalogue of viruses infecting accessions of BRCs	Files in the website	All P2	
06/2015	D1-2	Sequences comparisons and phylogenetic analyses of identified viruses	Files	P3, P5	
Under realisation	D1-3	Optimized protocols for viral diagnostic	Files in the website	All, P2	
WP 2 - Metagenomics					
10/2013	D2-1	Protocols for obtaining plant viromes	Files on the website	P3	
10/2013	D2-2	Training session in the use of metagenomics	Training session	P3	
04/2015	D2-3	Inventory of viruses present in the BRCs	Files on the website	P2	
WP 1 - Bioinformatics					
10/2012	D3-1	Results of the screening of crops ESTs for viral sequences	Data	P3, P5	
10/2012	D3-2	Sequence information on novel viral agents to WP 1	Data	P3, P5	
09/2013	D3-3	Processing and annotation pipelines for 454 pyrosequencing and Illumina siRNA reads	Data	P5	
10/2013 - 04/2015	D3-4	Training of partners to the use of pipelines	Training session	P5	
10/2015	D3-5	Results of analysis of all NGS data from the validation step of WP2	Data	P5	

TTP: throughout the project

E PROJECT IMPACT

The report gathers all necessary elements for the project assessment. They should permit to measure the general impact of the programme at different levels.

E.1 IMPACT ASSESSMENT INDICATORS

Number of publications and communications (to be detailed in E.2)

List separately single-partner initiatives involving only one partner, and multi-partner initiatives that result from a common work, including the publication in collaboration with at least one of your foreign/ different region or territory partners.

***Important**: avoid artificially increasing the number of publications, mention only those that result directly from the project (after it started, and which mention the support of the different funding bodies and Net-Biome call and the project references).*

		Multi-partner publications	Single-partner publications
International	**Peer-reviewed journals**	2	2
	Books or chapters in books		1
	Communications (conferences)	2	1
France	**Peer-reviewed journals**		
	Books or chapters in books		
	Communications (conferences)	6	2
Azores (foreign partner)	**Peer-reviewed journals**		
	Books or chapters in books		
	Communications (conferences)		1
Madeira	**Publicizing publication**		1
Outreach initiatives	**Please use the specific attached form for outreach**		

Other scientific valorization factors (to be detailed in 0)

In this table, give details of the national and international patents, licences, and other intellectual property elements resulting from the project, the know-how, any other spin-offs from the project, any partnerships, etc. See those announced in the technical appendix in particular.

	Number, years and comments (Actual or likely valorizations)
International patents obtained	0
International patents pending	0
National patents obtained	0
National patents pending	0
Operating licences (obtained / transferred)	0
Company creations or spin-offs	0
New collaborative projects	1, 2016, MALIN (project on infectious diseases - animal, human and vegetal - with scientific teams of Guadeloupe).
Scientific symposiums	0
Others (specify)	1, 2013 (bio-informatic software)

E.2 LIST OF PUBLICATIONS AND COMMUNICATIONS

List here the publications resulting from the project. Follow the categories used in the table of section E.1 using the usual standards for the field. As for the conferences, specify the invited conferences.

Publications in peer reviewed scientific journals

- **Roossinck MJ, Martin DP, Roumagnac P.** (2015). Plant Virus Metagenomics: Advances in Virus Discovery. Phytopathology:PHYTO12140356RVW.
- **Candresse T, Filloux D, Muhire B, Julian C, Galzi S,** et al. (2014). Appearances Can Be Deceptive: Revealing a Hidden Viral Infection with Deep Sequencing in a Plant Quarantine Context. PLoS ONE 9(7): e102945. doi:10.1371/journal.pone.0102945.
- **Massart S, Olmos A, Jijakli H, Candresse T.** (2014). Current impact and future directions of high throughput sequencing in plant virus diagnostics. Virus Res 188:90-96.
- **Acina Mambole I, Bonheur L, Svanella Dumas L, Filloux D, Gomez RM, Faure C, Lange D, Anzala F, Pavis C, Marais A, RoumagnaC P, Candresse T, Teycheney PY** (2014). Molecular characterization of yam virus X, a new potexvirus infecting yams (Dioscorea spp) and evidence for the existence of at least three distinct potexviruses infecting yams. Arch Virol DOI 10.1007/s00705-014-2211-3

Publications in Books or chapters in books

- **Filloux D, Dallot S, Delaunay A, Galzi S, Jacquot E, Roumagnac P.** (2015). Metagenomics approaches based on virion-associated nucleic acids (VANA): an innovative tool for assessing without A priori viral diversity of plants. In : Lacomme Christophe (ed.). Plant pathology: techniques and protocols. New York : Springer [Etats-Unis], Methods in Molecular Biology, 1302: Chapter 18: 249-257.

Oral communications in scientific conferences

- **Grisoni M** (2014) Broad range virus indexing through NGS: the SafePGR case study. International workshop on surveillance and control of cassava diseases, Saint-Pierre de la Réunion.
- **Pavis C** (2014) Development status of the SafePGR project (2014). Net-Biome 1st Joint Call Mid-Term Conference, Guadeloupe.
- **Umber M, Filloux D** (2015) Improved diagnostic tools for the detection badnaviruses in yams unveil the existence of endogenous sequences of extant badnavirus species in yams. 15èmes rencontres de virologie végétale, 18-22 January 2015, Aussois France.
- **Teycheney PY, Filloux D, Contreras S, Julian C, Theil S, Bonheur L, Acina Mambole A, Gomez RM, Bandou E, Lange D, Fernandez E, Pierret A, Rubington M, Faure C, Machado A, Mendonça D, Pinheiro de Carvalho M, Silva E, Daugrois J, Umber M, Pavis C, Grisoni M, Marais A, Roumagnac P, Candresse T** (2015) Viral treasure hunt in European outermost territories: how metagenomics boosts the discovery of novel viral species in tropical and sub-tropical crops germplasm. 15èmes rencontres de virologie végétale, 18-22 January 2015, Aussois (France).
- **Daugrois JH, Darroussat MJ, Fenouillet C, Ferdinand R, Fernandez E, Filloux D, Galzi S, Guinet-Brial I, Julian C, Lubin N, Roques D and Roumagnac P.** (2015). Impact of NGS-based virus discovery on sugarcane quarantine processes. ISSCT XI Pathology and IX Entomology Workshops. Guayaquil-Ecuador

Poster presentations

- **Acina Manbole I, Bonheur L, Anzala F, Gomez RM, Lange D, Faure C, Marais A, Pavis C, Roumagnac P, Filloux D, Candresse T, Teycheney PY.** (2013). Characterization and diagnostic of

Yam virus X (YVX) and Yam necrosis virus (YNV), two novel viruses infecting yams in Guadeloupe. 14èmes rencontres de virologie végétale, 13-17 January 2013, Aussois (France).

- **Julian C, Bernardo P, Fernandez E, Galzi S, Grisoni M, Da Silva Mendonça DM, Pavis C, Roumagnac P, Filloux D.** (2013). Improvement of three nucleic acid isolation protocols for an overall diagnosis of viruses on six vegetative propagated plants. 14èmes rencontres de virologie végétale, 13-17 January 2013, Aussois (France).**Mendonça D, Rocha S, Monjardino P, Luna S, Lopes MS, da Câmara Machado A.** (2013) Projecto SafePGR: desenvolvimento de métodos de diagnóstico de vírus para uma maior seguranca na circulação de plantas. Jornadas "Ciência nos Açores - que futuro?" 7-8 Junho 2013, Ponta Delgada, Açores.
- **Pavis C, Gamiette F, Umber M, Roques D, Boisseau M, Nuissier F, Petro D, Candresse T, Roumagnac P, Teycheney PY** (2013) The Guadeloupe germplasm repository, its role in viral diagnostic & sanitation. First Global Conference on Yam, 3-6 October 2013, Accra, Ghana
- **Fenouillet C, Filloux D, Galzi S, Roumagnac P, Daugrois JH.** (2015). Worldwide genetic diversity of *Sugarcane white streak virus*. 15èmes rencontres de virologie végétale, 18-22 January 2015, Aussois (France).
- **Filloux D, Bonheur L, Umber M, Pavis C, Fernandez E, Galzi S, Julian C, Daugrois JH, Sukal A, Winter S, Teycheney PY, Candresse T, Roumagnac P. (**2015). Metagenomic discovery, worldwide distribution and genetic diversity of novel macluraviruses infecting yams (*Dioscorea* spp.). 15èmes rencontres de virologie végétale, 18-22 January 2015, Aussois (France).

Publicizing publication

- **Da Silva EM, Pinheiro de Carvalho MAA.** (2015). Diagnostico de vírus em culturas de propagação vegetativa, exemplos da Bananeira e Batata-doce, através de técnicas moleculares. DICA (Revista online da DRADR). 123: 1-2. 24-06-2015.
- **Pavis C, Candresse T** (2015) Restitution du projet SafePGR to the stakeholders and the end users, SafePGR final meeting, Guadeloupe.

E.3 LIST OF VALORIZATION FACTORS

List here all valorization factors other than publications provided in the table section E.1, in particular the following:

- *national and international patents, licences, and other elements of intellectual property resulting from the project,*
- *software and any other prototype,*
- *standardization actions,*
- *launching of product or service, new project, contract, etc.,*
- *development of a new partnership,*
- *creation of a platform available to a community,*
- *company creation, spin-off companies, fund-raising,*
- *others (international opening, etc.).*

Indicate the partnerships if any. Present an assessment of the supply of deliverables, if they were specified in the technical appendix.

- VirAnnote - Pipeline for assembly, annotation of metagenomics sequences. Software adaptated from TriAnnote (2013, S. Contreras5 and S. Theil5).
- Sequences in Genebank.
- New and improved diagnostic protocols.
- MALIN - Cooperative project in construction, on infectious diseases - animal, human and vegetal - with scientific teams of Guadeloupe: University, Cirad, INRA, NGO, and governmental agencies. (2015-2017, P2 coordinator).
- EVA-Transfert 2 - Project in construction in Guadeloupe, for the development of a quality seed production in yam sector - with research, extension services and economic actors (2016-2018, P1 and P2 partners).
- Regional project on cassava sanitation in Reunion, Mayotte and Comoros.

For ANR funded labelled projects (competitiveness clusters) only:

Project collaboration with the cluster (s) that labelled it

What collaborations have there been between your project and the competitiveness cluster(s) that labelled it?

Detail the activities carried out by the public laboratories using the complementary funding granted on account of the labelling. Indicate in particular the partners involved and the collaborative work conducted with the cluster(s).

E.4 ASSESSMENT AND FOLLOW-UP OF PERSONNEL RECRUITED ON FIXED-TERM CONTRACTS (EXCLUDING INTERNS): FOR ANR FUNDED PROJECTS ONLY

This table summarizes project recruitment of non-permanent personnel on fix-term contracts or equivalent. Fill out one line per person hired for the project when the hiring has been partially or entirely financed with the ANR grant and the contribution has lasted at least 3 months, all employment contracts combined, subject to the condition that the ANR grant can only represent a portion of the person's remuneration over the duration of his or her participation in the project.

Interns who have an internship agreement with an educational establishment must not be mentioned.

The date collected may be requested to be updated up to 5 years after the end of the project.

Identification				Before recruitment for the project			Recruitment for the project				After the project				
Surname and first name	Sex M/F	E-mail address (1)	Date of last news	Last diploma obtained at time of recruitment	Place of studies (France, EU, outside EU)	Prior professional experience, including post-docs (years)	Partner who hired the person	Position in the project (2)	Duration of missions (months) (3)	End date of mission on project	Professional future (4)	Type of employer (5)	Type of employment (6)	Relation with ANR project (7)	Promotion of professional experience (8)
Lydiane BONHEUR	F	lydiane.bonheur@yahoo.fr	July 2015	MSc	France	1	P1	Technical manager	24	November 2014	Open-ended employment contract	Public research institute	Technician	none	yes
Charlotte JULIAN	F	Charlotte.julian@cirad.fr	September 2015	MSc	France	0	P3	Technical assistant	24	May 2014	Fixed-termed employment contract	Public research institute	Technician	yes	yes
Sandy CONTRERAS	F	Sandy.contreras@laposte.net	August 2015	MSc	France	1	P5	Bioinformatics engineer	24.5	August 2015	Open-ended employment contract	Private company	Engineer in bioinformatics	none	yes
Sara ROCHA	F	sararocha@hotmail.com	August 2015	MSc	Portugal	2	P6	Technical assistant	12	August 2013	Open-ended employment contract	Public research institute	Technician	none	yes
Emanuel SILVA	M	emasil2008@gmail.com	February 2014	Graduation	Portugal	4	7	Technician	17	February 2013	Open-ended employment contract	University	Technician	none	yes

Annex 1 - New viruses identified during the project.

		Virus discovery			Diversity	Diagnostic			Publication	
Host plant	Virus genus	Planned	Achieved WP2	Achieved WP3	Improved know-ledge on diversity	Acronym	Optimized / pre existing	Developped/transferred	Publication status	Expected submission date
Banana	Badnavirus	✓				BA-BAD	✓			
	Potyvirus	✓				BA-PTY	✓			
	Flexivirus	✓				BA-FLE	✓			
	Tobamovirus			✓				✓		
Garlic	Potyvirus	✓			✓	GA-PTY	✓		To be deposited in Genebank	Dec 15
	Carlavirus	✓			✓	GA-CAR		✓	To be deposited in Genebank	Dec 15
	Allexivirus	✓			✓	GA-ALL		✓	To be deposited in Genebank	Dec 15
	Foveavirus		✓	✓	✓	GA-FOV		✓	Paper to write	Dec 15
	Tospovirus	✓				GA-TOS				
	Luteovirus		✓			GA-LUT		✓	Paper to write	Dec 15
	Umbravirus		✓			GA-UMB		✓		
Sweet potato	Begomovirus	✓			✓	SP-BEG	✓			
	Carlavirus	✓				SP-CAR		✓		
	Potyvirus	✓			✓	SP-PTY	✓		New disease report, to write	Jun 16
	Crinivirus	✓				SP-CRI		✓		
	Soymovirus		✓			SP-SOY		✓		
	Mastrevirus		✓			SP-MAS		✓		
	Nucleorhabdovirus		✓			SP-NUC		✓		
Sugarcane	Badnavirus	✓			✓	SC-BAD	✓		Genbank	
	Potyvirus	✓				SC-PTY				
	Mastrevirus	✓	✓		✓	SC-MAS	✓	✓	1 published; new disease report; diversity pubished separately	
	Betaflexivirus	✓		✓		SC-BET		✓		Jun 16
	Closterovirus		✓		✓	SC-CLO		✓	Paper to write	Jun 16
	Umbra/Tombusvirus		✓			SC-UMB		✓	Paper to write	Jun 16
Vanilla	Potexvirus		✓		✓	VA-PTX		✓	Paper to write	Dec 15
	Alphaflexivirus (allexivirus)		✓		✓	VA-ALL		✓	Paper to write	Dec 15
	Luteovirus		✓			VA-LUT			Paper to write	Dec 15
Yam	Badnavirus	✓			✓	YA-BAD	✓		Genbank	
	Closterovirus		✓	✓	✓	YA-CLO		✓	Paper to write	Dec 15
	Macluravirus		✓	✓	✓	YA-MAC		✓	Paper to write	Jun 16
	Potexvirus	✓	✓	✓	✓	YA-PTX	✓	✓	Published	
	Potyvirus	✓	✓		✓	YA-PTY				
	Unassigned *Secoviridae*	✓	✓	✓	✓	YA-SEC		✓	Paper to write	Dec 15
	Flexivirus		✓	✓		YA-FLX				
	Polerovirus			✓		YA-POL				
	Geminviridae	✓	✓	✓		YA-GEM				

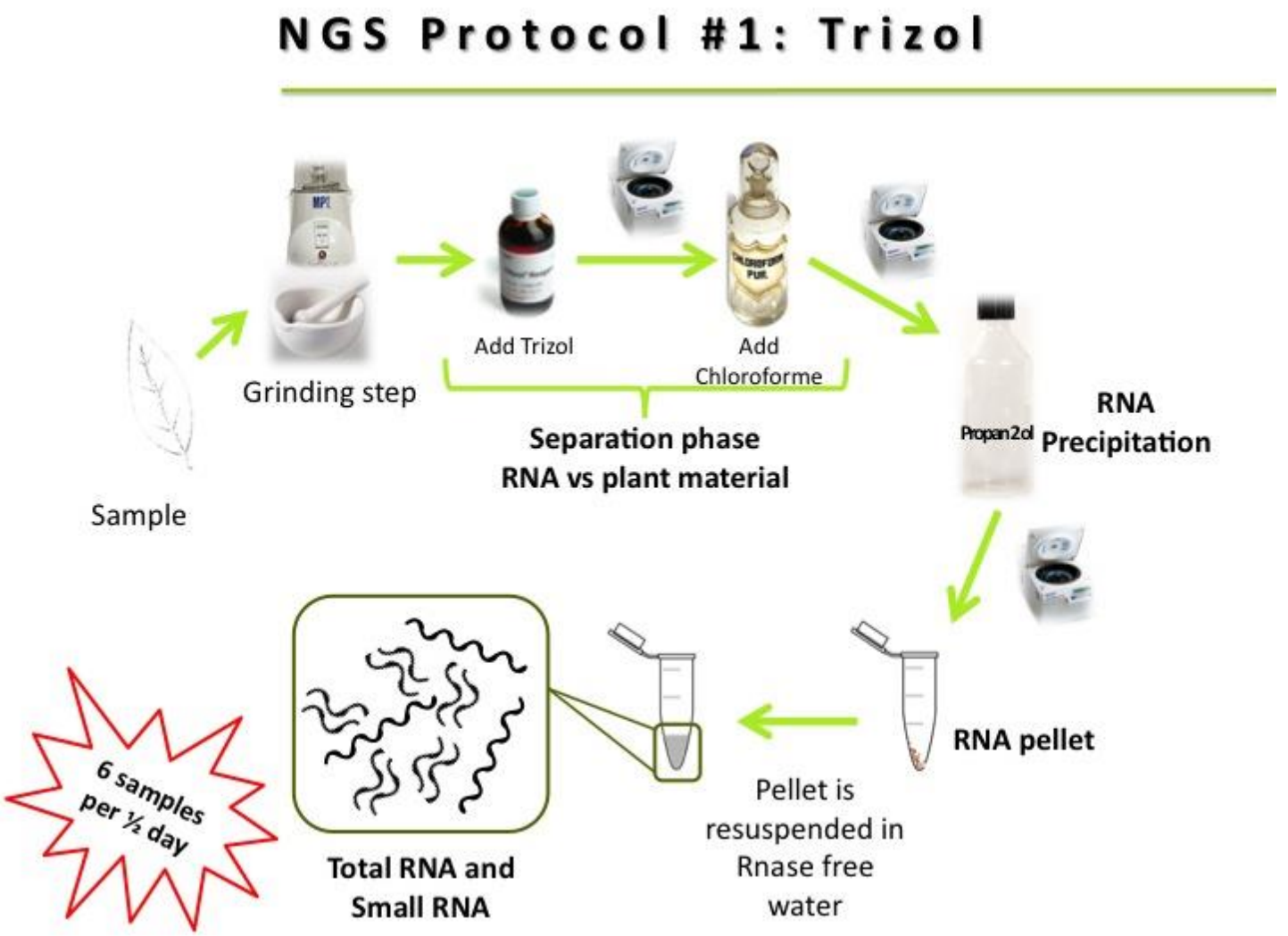
Planned : virus already known

Achieved WP2 and WP3 : virus discovered in the frame of the project.

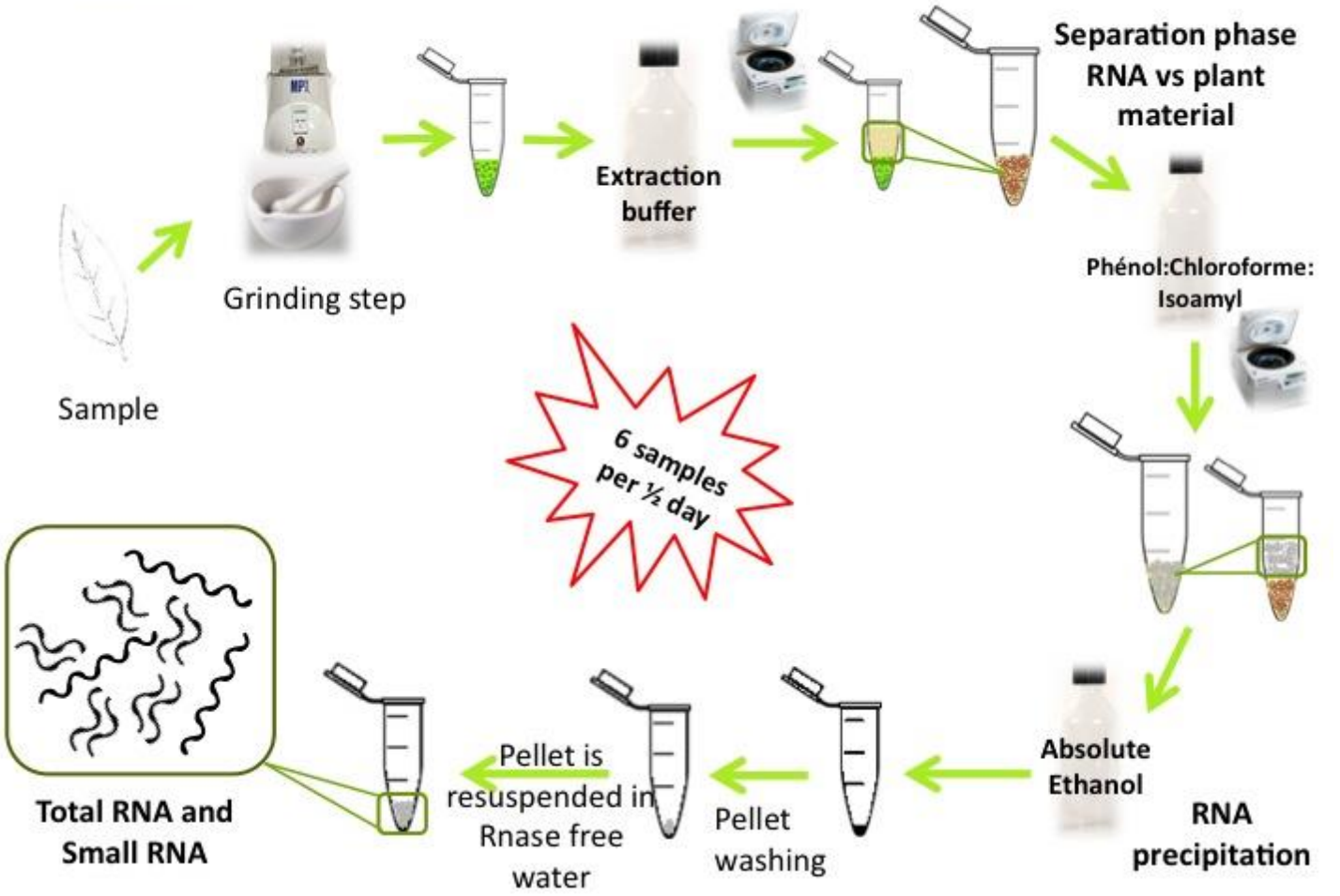
Annex 2 - Detail of the analysed sampled counts.

	Guadeloupe			Réunion			Madeira			Azores			Total	
	sample count (nb accessions)	virus tested	total tests	sample count (nb accessions)	virus tested	total tests	sample count (nb accessions)	virus tested	total tests	sample count (nb accessions)	virus tested	total tests	Total accessions	Total indexings
Banana	230	3	690				36	2	72	56	3	168	322	930
Garlic	0		0	46	6	166							46	166
Sweet potato	18	5	90	89	4	156	64	4	256	42	5	210	213	712
Sugarcane	165	5	825										165	825
Vanilla	0		0	76	1	76							76	76
Yam	170	7	1 190	8	5	40							178	1 230
Total	**583**		**2 795**	**219**		**438**	**100**		**328**	**98**		**378**	**1 000**	**3 939**

Annex 3 - Eight NGS protocols tested for the metagenomic approach.



NGS Protocol #2: Phenol Chloroforme



NGS Protocol #3: KIT mirVana

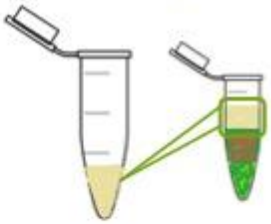
Sample

Grinding step

Lysis / Binding buffer

Acid-Phenol:Chloroform

Separation phase RNA vs plant material



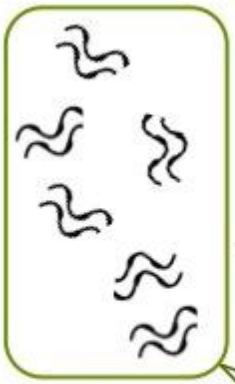
6 échantillons en une matinée

Absolute Ethanol

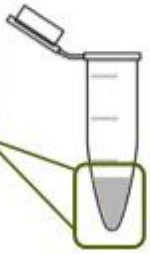
RNA precipitation

Several washing steps

Elution

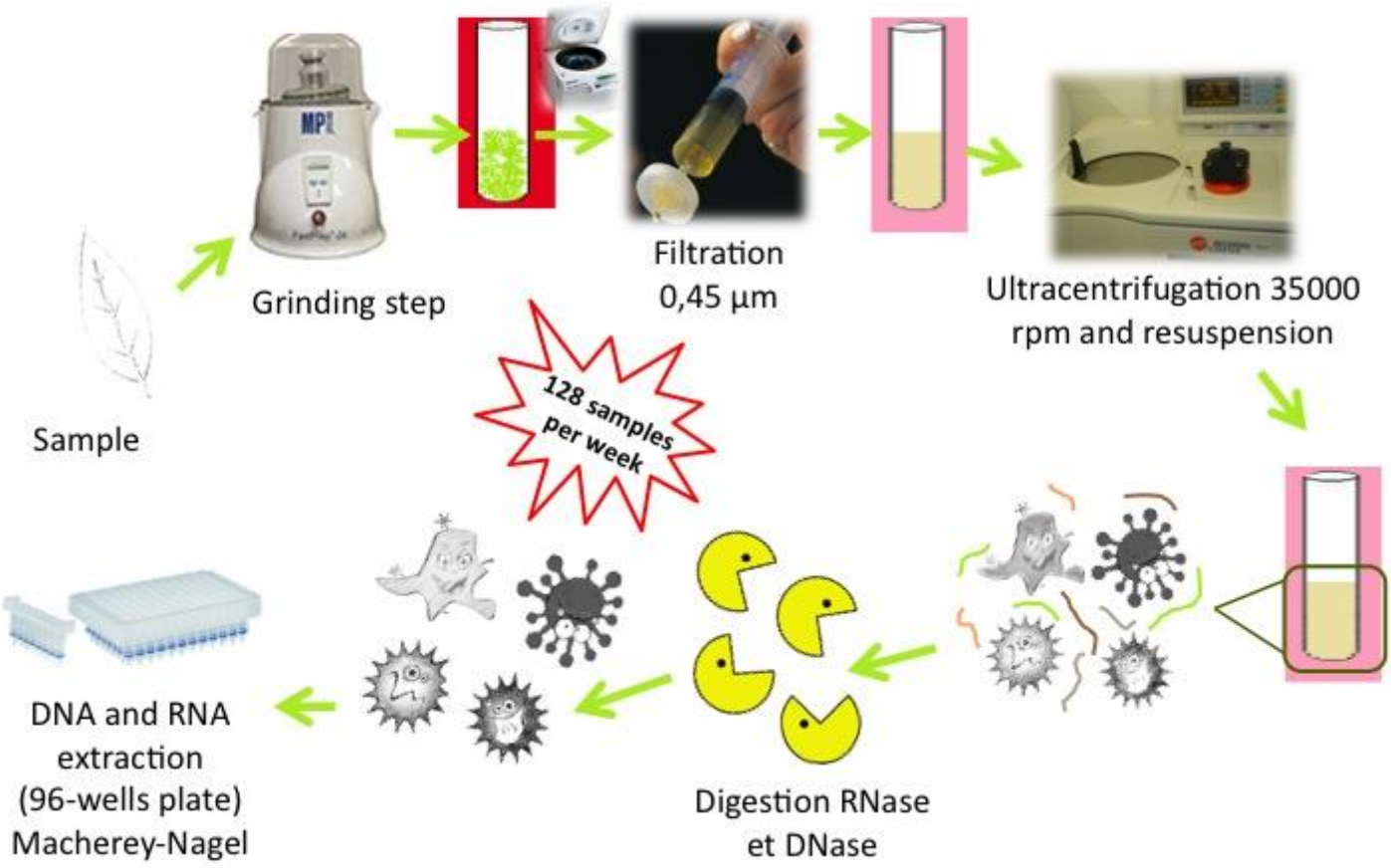


SMALL RNAs



NGS Protocol #4: Viral particle target_1

From Victoria/Blinkova (Victoria, 2009 / Blinkova,2010) modified by Emmanuel Fernandez (UMR BGPI)



NGS Protocol #5: Viral particle target_2

From Filloux et al (2015). Plant Pathology: Techniques and Protocols, 249-257

3 Methods

3.1 Purification of Viral Particles (Modified from ref. [4])

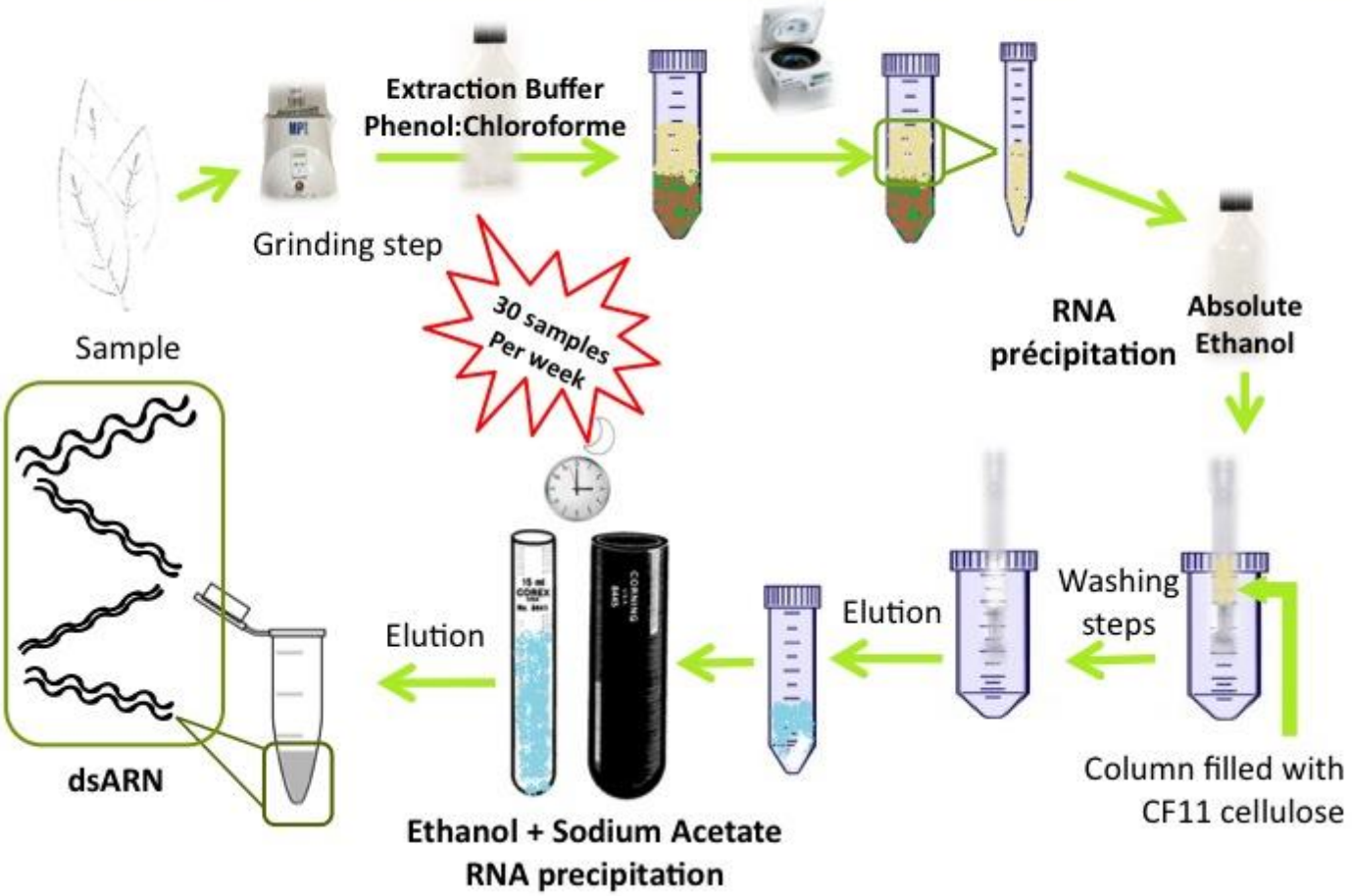
1. Grind 5 g of leaf material in liquid nitrogen in the presence of about 100 mg of carborundum using a pestle and a mortar.
2. Add 15 mL of sodium–potassium phosphate extraction buffer and homogenize.
3. Transfer and filter the homogenized plant extracts in a Miracloth filter placed on a funnel and a 50 mL conical tube.
4. Add 330 µL of Triton® X-100 to the filtrate and mix during 15 min with an orbital shaker.
5. Centrifuge at 14,000 × *g* for 10 min at 4 °C.
6. Transfer the supernatant in a 26.3 mL ultracentrifuge polycarbonate bottles.
7. Add 6 mL of a 30 % sucrose solution diluted in 0.2× sodium–potassium phosphate extraction buffer, at the bottom of the tube with a Pasteur pipette.
8. Centrifuge at 148,000 × *g* for 1 h at 4 °C.
9. Discard the supernatant by pipetting.
10. Wash once the tube carefully with few milliliters of deionized water.
11. Add 150 µL of 1× RQ1 DNase buffer and let the resulting pellet resuspend overnight at 4 °C.
12. Transfer the viral particles suspension by pipetting in a 1.5 mL Eppendorf tube.
13. Add 15 µL of RQ1 DNase (1 U/µL) and 1.5 µL of RNase A (7 U/µL), and incubate at 37 °C for 2 h.

3.2 Viral Nucleic Acid Extraction

1. Extract RNA from the total volume of viral particle suspension obtained at the previous step (approximately 166.5 µL) with the RNeasy Plant Mini Kit according to the manufacturer's protocol.
2. Elute RNA in a 1.5 mL microtube by adding 50 µL of molecular grade water to the center of the RNeasy Mini Spin column and by centrifugation at 8,000 × *g* for 1 min at room temperature.

NGS Protocol #6: CF11_1/ECOGENOMICS

From Roossinck et al (2010)



NGS Protocol #7: CF11_2 / « Bordeaux »

Before starting, prepare:

- STE 10X: NaCl 1M, Tris 500mM, EDTA 10mM, pH 8
- STE 2X: 10 ml STE 10X + 40 ml DEPC-treated water
- STE 1X + 15% alcohol: 4 ml STE 10X + 6 ml absolute alcohol + 30 ml DEPC-treated water (prepare just before use)
- STE 1X: 1 ml STE 10X + 9 ml DEPC-treated water
- SDS 20%
- SDS 2% (diluted in DEPC-treated water)
- Bentonite (40 mg/ml)
- Phénol-TE saturated *(SIGMA 77607)*
- Isopropanol
- Absolute alcohol
- 3M DEPC-treated sodium acetate pH 5.2
- DEPC-treated water *(1ml DEPC + 1 L distilled water – mix well overnight on magnetic agitator – autoclave your DEPC water before using)*
- Cellulose CF11
- 70% Ethanol
- 1M Magnesium Acetate *(VWR A6999.0005)*
- DNAse RQ1 *(EUROMEDEX 1307)*
- 10X SSC *(SIGMA SS6639 – 20X SSC)*
- RNase A *(EUROMEDEX 9707-A)*
- Proteinase K *(EUROMEDEX EU-0090-A)*
- Phenol Chloroform isoamyl alcohol mix *(SIGM A P2069)*
- Chloroform isoamyl alcohol mix *(SIGM A C0549)*
- 1,5 ml tubes with 40 mg CF11/ tube
- 1,5 ml tubes
- Filter tips 1000 µl, 200 µl and 10 µl
- Pipets used only for this manipulation
- Mortar and pestle autoclaved
- Spatula *(VWR 231-0104)*
- Funnels autoclaved *(VWR 221-0154)*

Each solution must be made with DEPC-treated water and autoclaved before use

DAY 1

1. Keep mortar and pestle in liquid nitrogen just before using

Prepare **extraction buffer** in a 15 ml Falcon

- **1 ml STE 2X**
- **70 µl SDS 20%**
- **20 µl Bentonite**
- **1,425 ml Phénol-TE**

Weigh **0.750 g** of sample (or 0.075 g of dried material)
Pulverize the sample in the presence of liquid nitrogen with the precooled mortar and pestle.
Put the funnel in liquid nitrogen container
Grind until obtain a white powder
Transfer the frozen powder to a 15ml Falcon containing extraction buffer with the spatula and funnel
Dispense and thaw the content by vortexing
Homogenize

2. Slow agitation **30 min** about 70 to 85 rpm (Horizontal agitator)
Spin at **3000 g** for **15 min**

3. Transfer the aqueous phase to a new 1,5 ml tube

Spin at **10000 g** for **20 min** at **20°C**
Transfer the aqueous phase to a new 1, 5 ml tube
Add absolute alcohol to obtain a 15 % final concentration
→Absolute alcohol volume to add = **0.176 x phase aqueous volume**
Mix well and transfer to a 1, 5 ml tube containing 40 mg CF11
Mix well
Slow agitation **30 min**
Spin at **5000 g** for **1 min** at**20°C**
Remove the supernatant

4. Wash the silica with **1 ml STE 1X + 15% alcohol**
Mix well and take off the pellet with a tip
Vortex slightly
Slow agitation **5 min**
Spin at **5000 g** for **1 min** at **20°C**
Remove the supernatant
Repeat this step twice or more (until the supernatant became colorless)

At the end of last wash, remove the supernatant and dry cellulose with tip

5. *Elution*
Add **200 µl STE 1X** to the dry CF11
Mix well and take off the pellet with a tip
Vortex slightly
Slow agitation **5 min**
Spin at **5000 g** for **1 min** at **20°C**
Collect (**on ice**) the supernatant
Repeat this step once and collect the supernatant to the same tube (at the last step, be careful to well dry CF11 with your tip)
Spin at **5000 g** for **1 min** at **20°C** to remove cellulose
Retain the supernatant (about **400 µl**)

6. *Nucleic acids precipitation*
Add **1\10 volume 3M sodium acetate pH5.2 (40 µl)**
Add **0.8 volume isopropanol (320 µl)**
Store **overnight** at **-20°C**

DAY 2

7. *Nucleic acids enzymatic treatment*
Spin at **20000 g** for **20 min** at **4°C**
Remove the supernatant
Wash the pellet with **500 µl 70% alcohol**
Spin at **20000 g** for **20 min** at **4°C**
Remove the supernatant
Dry the pellet with speed-vac (about **12 min**)
Dissolve the pellet in **160 µl DEPC-treated water**

DNAse treatment
Add **160µl ARN**
Add **20µl 1M Magnesium Acetate**
Add **10µl DNAse RQ1 1U/µl**
Add **10 µl DEPC-treated water (**to a 200µl total volume)

60 min at **37°C**

RNase A treatment in salt conditions (2X SSC)
Add **60µl 10X SSC**
Add **1µl Rnase A (10 µg/µl)**
Add **39µl DEPC-treated water** (to a 300µl total volume)

30 min at **37°C**

Proteinase K treatment
Add **2.5µl SDS 2%**
Add **8µl Proteinase K (5mg/ml)**

60 min (at least) at **37°C**

8. *Extract once with **Phenol:Chloroform:Isoamylalcohol** mix*
   Add v/v **(300 µl)** of **Phenol:Chloroform:Isoamylalcohol** mix
   Vortex well
   Spin at **10000 g** for **5 min** at **20°C**
   Retain the supernatant
   *Extract the aqueous phase with **Chloroform:isoamyl:alcohol** mix*
   Add v/v **(300 µl) of Phenol:Chloroform:Isoamylalcohol** mix
   Vortex well
   Spin at **10000 g** for **5 min** at **20°C**
   Retain the supernatant

9. *Aqueous phase ethanol precipitation*
   Add **2 volumes absolute alcohol (600 µl)** and **1/10 volume 3M sodium acetate pH 5.2 (30 µl)**
   Store **30 min** at **- 80°C**
   Spin at **20000 g** for **30 min** at **4°C**
   Remove the supernatant
   Wash the pellet with **500 µl 70% alcohol**
   Spin at **20000 g** for **20 min** at **4°C**
   Remove the supernatant
   Dry the pellet with speed-vac (about **12 min**)

10. Dissolve the pellet in **250 µl DEPC-treated water**

DAY 3

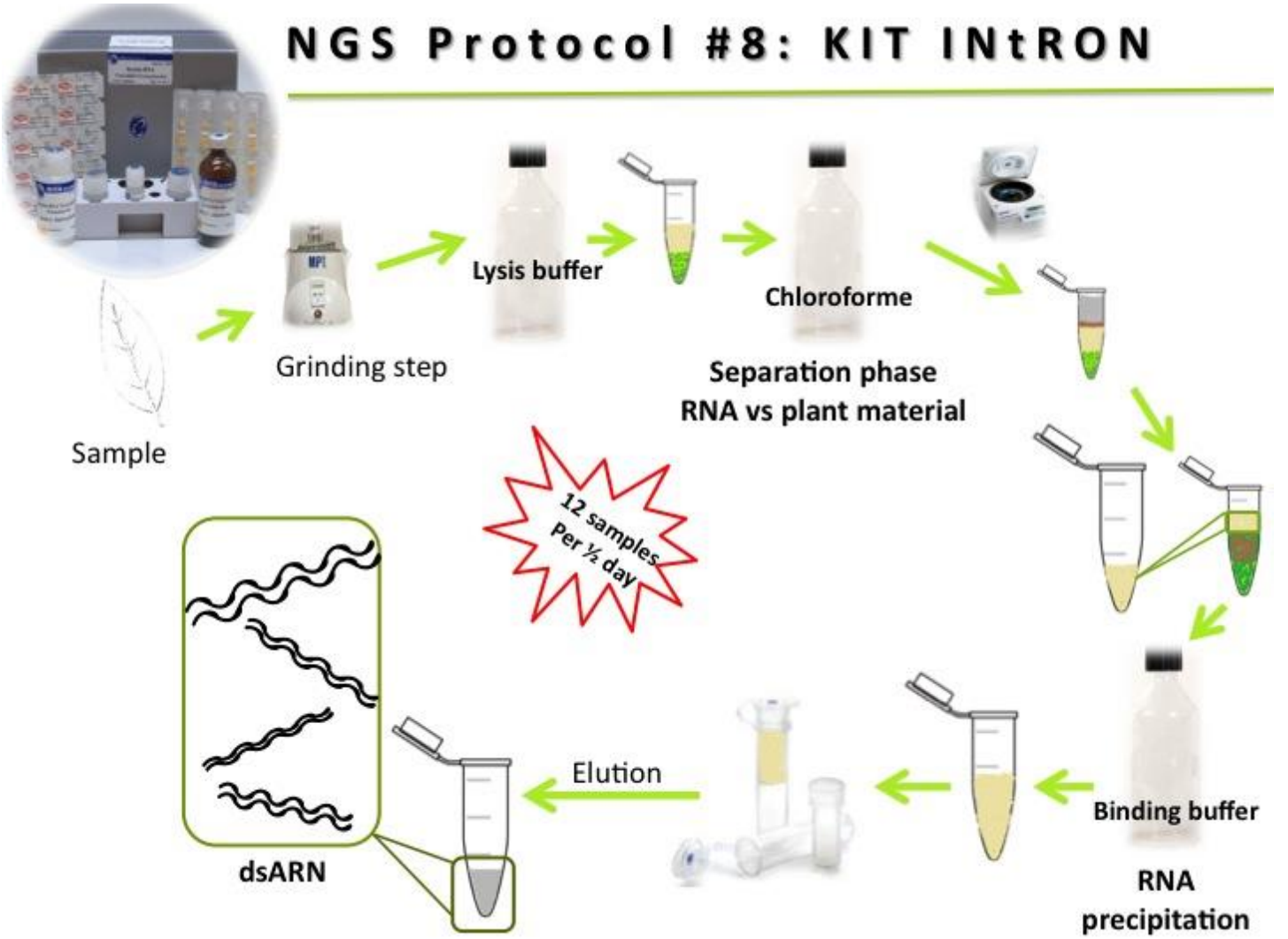
11. Add **44 µl absolute alcohol to 250 µl dsRNA**
    Transfer the mix into a new tube containing 40mg of CF11
    Mix well with a tip if necessary
    Slow agitation **30 min**
    Spin at **10000 g** for **1 min** at **20°C**
    Remove the supernatant

12. Wash the CF11 with **1 ml STE 1X + 15% alcohol**
    Mix well and take off the pellet with a tip
    Vortex slightly
    Slow agitation **5 min**
    Spin at **5000 g** for **1 min** at **20°C**
    Remove the supernatant
    Repeat this step three times
    At the end of last wash, remove the supernatant and dry cellulose with a tip

13. *Elution*
    Add **200 µl STE 1X**
    Mix well and take off the pellet with a tip
    Vortex slightly
    Slow agitation **5 min**
    Spin at **5000 g** for **1 min** at **20°C**
    Collect (**on ice**) the supernatant
    Repeat this step once and collect the supernatant to the same tube
    Spin at **5000 g** for **1 min** at **20°C** to remove cellulose
    Retain the supernatant (about **400 µl**)

14. *Nucleic acids precipitation*
    Add **1\10 volume 3M sodium acetate pH5.2 (40 µl)**
    Add **2 volumes absolute alcohol (800 µl)**
    Store **overnight** at **-20°C**

15. Spin at **20000 g** for **20 min** at **4°C**
    Remove the supernatant
    Wash the pellet with **500 µl 70% alcohol**
    Spin at **20000 g** for **20 min** at **4°C**
    Remove the supernatant
    Dry the pellet with speed-vac (about **12 min**)
    Dissolve the pellet in **20 µl DEPC-treated water**


NGS Protocol #8: KIT INtRON
Sample
Grinding step
Lysis buffer
Chloroforme
Separation phase
RNA vs plant material
12 samples
Per ½ day
Binding buffer
RNA
precipitation
Elution
dsARN

Annex 4 - Results on the comparison of the 38 infected plants.

Comparison VANA / dsRNA on 38 samples (6 plant species)

1/ Interpretation of results is complicated by several factors including

- Probable errors in sample numbering between the two techniques (see below)
- A very prevalent background of low numbers of contaminating reads in the VANA samples. This is much less prevalent in dsRNA samples (but there are nevertheless a few contaminating reads in a few samples) but, with the exception of one Vanilla sample (CRV2148) there are no samples without some contaminating reads coming from viruses infecting other plants in the experiment. Frequently, this involves multiple viruses (for example, between 3 and 8 contaminating viruses in banana samples).
- In some cases, like novel viruses, annotation is complex because different contigs will be annotated as different viruses, so we have to use a "genus level" or "family level" annotation. This also applies in genera with closely related viruses that are in coinfection in a host (Allexiviruses in garlic are a good example) where individual contigs may be annotated as one species or another depending on the genomic region covered. Care has to be used here also.

The problem of contaminating reads is the most critical as it complicates the interpretation of results. If we consider that a virus is positively detected as soon as a single reads is detected for this virus, then we will face evidence of massive contamination. **I think we therefore have to use some kind of threshold to decide that a virus is indeed detected** (we can always go back to individual reads if a virus is not detected using the threshold…). I have tried to use **as a positive detection criterion the identification of a contig** (meaning there are at least two reads for the same region of the viral genome). This removes really a lot of the contaminations and provides a clearer picture, and I provided my analysis of the 38 samples below.

Banana

Sample name	Virus supposed to be present	Viruses detected by dsRNA	Viruses detected by VANA	Comments
Banana-Montpellier-CMV	CMV	CMV	CMV	CMV detected by both techniques 7 contaminating viruses VANA
Banana-Montpellier-Ouro_Mel	BanMMV	BanMMV	BanMMV	BanMMV detected by both techniques 8 contaminating viruses VANA
Banana-Montpellier-Indes_J	BanMMV	BanMMV	BanMMV	BanMMV detected by both techniques 4 contaminating viruses VANA
Banana-Montpellier-BSCav-BSV	BSV	**0 reads BSV**	**no BSV (no contig, 3 reads only)** **Pepino mosaic virus**	BSV status unclear : not detected by dsRNA, marginally seen in VANA as 3 reads but no contig 7 contaminating viruses VANA, PepMV still positive as contig
Banana-Montpellier-BBTV-Nouvelle_Caledonie	BBTV	BBTV	BBTV **Garlic virus C**	BBTV detected by both techniques 7 contaminating viruses VANA, GVC still positive as contig
Banana-Montpellier-ITC-1286	BanMMV	BanMMV BSV (1 contig, 4 reads only)	BanMMV **no BSV (no contig, 2 reads only)**	BanMMV detected by both techniques BSV status unclear, only as reads in VANA, low reads and 1 contig in dsRNA 5 Contaminating viruses VANA 1 contaminating virus dsRNA
Banana-Montpellier-BBrMV	BBrMV	BBrMV	**0 reads BBrMV**	BBrMV not detected by VANA 6 contaminating viruses VANA 1 contaminating virus dsRNA

Synthetic analysis

Both techniques generally OK, good detection of the expected viruses with two exceptions

- **BSV not detected by dsRNA (OK for a DNA virus)**
- **BBrMV not detected by VANA (not logical, Potys well detected in other plant species)**
- **BSV status of plant** Banana-Montpellier-ITC-1286 and Banana-Montpellier-BSCav-BSV **unclear**

Yam

Sample name	Virus supposed to be present	Viruses detected by dsRNA	Viruses detected by VANA	Comments
Yam-Montpellier-NZ366Do	Betaflexi nouveau	Yam latent virus	Yam latent virus	Yam latent (Carla) detected by both techniques 8 contaminating viruses VANA, GVC still positive as contig 3 contaminating viruses dsRNA
Yam-Montpellier-BJ65Ds	DBV	DBV	DBV Iflavirus?	DBV detected by both techniques (surprising for dsRNA….) 11 contaminating viruses VANA, a potential Iflavirus (insect virus) still positive as contig
Yam-Montpellier-VU#206Dn	Betaflexi nouveau?	Potexvirus (Yam virus X)	Potexvirus (Yam virus X)	Potex detected by both techniques 14 contaminating viruses VANA
Yam-Montpellier-VU#333Dn	Potex nouveau	**Cannot be interpreted, low reads number (Potex, 1 read)**	Potex nouveau	Potex detected by VANA 14 contaminating viruses VANA
Yam-Montpellier-VU#331Dn	Macluravirus	**Cannot be interpreted, low reads number (Clostero, 2 reads)**	Maclura Clostero	Macluravirus et Closterovirus detected by VANA 6 contaminating viruses VANA

Synthetic analysis

Problem with dsRNA: low read number obtained from 2 samples so that infection status cannot be determined in a secure way. Results OK in the 3 samples that can be interpreted

Good results obtained with VANA

Garlic

Sample name	Virus supposed to be present	Viruses detected by dsRNA	Viruses detected by VANA	Comments
Garlic-Reunion-CRA0047	Allexi, Luteo, Umbra	Allexi GCLV LYSV OYDV Betaflexiviridae Umbra-like Tombusviridae	Allexi GCLV LYSV OYDV **No Betaflexi** Umbra-like Tombusviridae *(as Tombunodavirus)*	Allexis, GCLV, LYSV, OYDV detected by both techniques No Betaflexi or Umbra-like detected by VANA 3 contaminating viruses VANA
Garlic-Reunion-CRA0017	Allexi, GCLV, Fovea, Luteo	Allexi GCLV LYSV (conta?) Alphaflexi Betaflexi **No OYDV** **No Luteo**	Allexi GCLV **No LYSV** **No Alphaflexi (no contig, 1 read)** Betaflexi OYDV Luteo	Allexis, GCLV and Betaflexi detected by both techniques LYSV and Alphaflexi detected only by dsRNA (reads alphaflex VANA but no contig) OYDV and Luteo detected only by VANA 3 contaminating viruses VANA
Garlic-Reunion-CRA0010	Allexi, GCLV, LYSV, Fovea, Luteo	Allexi Betaflexi STNV **No LYSV** **No OYDV** **No Luteo** **No GCLV** **No GLV**	Allexi Betaflexi STNV LYSV OYDV (conta?) Luteo GCLV GLV (conta?) Potex conta	Allexis, Betaflexi and STNV detected by both techniques LYSV, OYDV, GCLV, GLV and Luteo only detected by VANA but some with very few reads, maybe be contaminations 4 contaminating viruses VANA, 1 Potex conta still positive in contigs
Garlic-Reunion-CRA0038	GCLV-ShLV (ambiguities manip 1)	Allexi Betaflexi STNV GLV (conta?) Iris yellow spot Tospo **No Luteo** **No OYDV**	Allexi Betaflexi **No STNV** **No GLV** **No IYSV** Luteo OYDV	Allexi and Betaflexi detected by both techniques Luteo and OYDV only detected by VANA STNV, GLV and Iris yellow spot Tospovirus only detected by dsRNA 2 contaminating viruses VANA

Garlic-Reunion-CRA0046	LYSV, OYDV (ambiguities manip 1)	Allexi LYSV OYDV	Allexi LYSV OYDV	Allexi, LYSV et OYDV detected by both techniques 4 contaminating viruses VANA
Garlic-Reunion-CRA0008	New Peclu	Allexi Betaflexi LYSV STNV Peclu GCLV (conta?)	Allexi Betaflexi LYSV STNV Peclu **No GCLV (no contig, only 4 reads)**	Allexi, Betaflexi, LYSV, STNV, Peclu detected by both techniques GCLV only detected by dsRNA 2 contaminating viruses VANA

Synthetic analysis

Difficult to interpret because lots of discrepancies between the two techniques, very difficult to know if the samples were not mixed up somehow even if for some of them (CRA0008 with new Peclu) we seem to be OK. There might also be some contamination problems that we cannot clarify in a clean fashion….

Both techniques detect the Allexis, GCLV, LYSV, OYDV, STNV and the new Peclu in one sample or another. There are problems of coherence or of detection of some of the same viruses in other samples. This may results from different competition levels between many viruses.

The Luteo does not seem to be detected efficiently in any sample by dsRNA. This is very surprising given that in the previous experiments of Plates 1-2 it was detected in several plants by dsRNA but was not detected by VANA

IYSV (a Tospovirus, labile particles) is not detected by VANA

We will have to see if these results can be put in a synthesis paper given the broad range of discrepancies between the two techniques.

Sugarcane

Sample name	Virus supposed to be present	Viruses detected by dsRNA	Viruses detected by VANA	Comments
Sugarcane-Guadeloupe-CP52-43-hybrid	Umbravirus, SCYLV, SCBV	**Yam latent virus, cannot be interpreted, contamination with a Yam sample**	SCYLV ScBV	The dsRNA sample contains Yam reads, probable =mix of samples, or contamination, cannot be interpreted SCYLV and ScBV detected by VANA 7 contaminating viruses VANA
Sugarcane-Montpellier-SCYLV_R577_REU	SCYLV, Clostero	SCYLV Clostero	SCYLV Clostero	SCYLV and Clostero detected by both techniques 12 contaminating viruses VANA, one still positive as contig 2 contaminating virus dsRNA
Sugarcane-Montpellier-C13281	Mastre, Clostero	**Cannot be interpreted, low reads number** Clostero (5 reads)	Clostero SCYLV	SCYLV and Clostero detected by VANA Expected Mastre not detected by VANA (initial diagnostic OK ?) 8 contaminating viruses VANA, one still positive as contig
Sugarcane-Montpellier-VARX	Mastre, possibly SCYLV, ScMV, ScBV	Clostero **No Mastre** **No SCYLV**	**No Clostero** Two Mastre (SSEV and SWSV) SCYLV	Clostero not detected by VANA Mastre and SCYLV not detected by dsRNA 3 contaminating viruses VANA 1 contaminating virus dsRNA
Sugarcane-Montpellier-SL7103	SCSMV	SCSMV ScBV	SCSMV	SCSMV detected by both techniques ScBV status dsRNA unclear (integrated ?) 5 contaminating viruses VANA 1 contaminating virus dsRNA
Sugarcane-Montpellier-SRMV_ISOLAT51_USA	SCYLV, SrMV, ScBV	SrMV Clostero ScBV **No SCYLV**	SrMV Clostero ScBV SCYLV	SrMV, Clostero and SCBV detected by both techniques SCYLV not detected by dsRNA 5 contaminating viruses VANA, one still positive as contig 1 contaminating virus dsRNA
Sugarcane-Montpellier-SP716163	SCYLV, Clostero	Clostero **No SCYLV (no contig, only 2 reads)** Chrysovirus, insect virus ?	Clostero SCYLV	Clostero detected by both techniques SCYLV not detected by dsRNA 12 contaminating viruses VANA, one still positive as contig

Synthetic analysis

Two samples cannot be interpreted in dsRNA (low read number for one, Yam contamination for one)

SCYLV not detected in 3 of 4 samples in dsRNA, Clostero not detected in 1 of 5 samples in VANA. Detection SrMV and SCSMV OK with both techniques

Sweet potato

Sample name	Virus supposed to be present	Viruses detected by dsRNA	Viruses detected by VANA	Comments
Sweet_Potato-Azores-Bd0345	SPFMV, SPG/SP2	**Cannot be interpreted, low reads number** (SPFMV, 2 reads, SPV 3 reads)	SPFMV SPVG SPV2	SPFMV, SPVG & SPV2 detected by VANA 18 contaminating viruses VANA
Sweet_Potato-Azores-Bd0346	SPFMV, SPVG/SPV2, Mastre?	SPFMV SPVG SP Mastre	SPFMV **No SPVG** SP Mastre	SPFMV & SP Mastre detected by both techniques SPVG not detected by VANA 16 contaminating viruses
Sweet_Potato-Reunion-CRCo004	Mitovirus, SPSMV1	**No SPFMV** Badna	SPFMV	SPFMV detected by VANA, not by dsRNA **Probable exchange with next sample in the two techniques** 17 contaminating viruses VANA
Sweet_Potato-Reunion-CRCo002	Soymo, Mastre-like_sequences_(CF11)	Soymo SP Mastre SPFMV Badna	**No Soymo** **No SP Mastre** **No SPFMV**	Soymo, Mastre et SPFMV detected by dsRNA SPFMV not detected by VANA **Probable exchange with previous sample in the two techniques** 7 contaminating viruses VANA, 1 still present as contig 1 contaminating virus dsRNA
Sweet_Potato-Azores-Bd2015	Begomovirus ?	SPCSV SPFMV SPVC SPVB3 SP Mastre Insect virus	SPCSV SPFMV SPVC SPVB3 **No SP Mastre**	SPCSV, SPFMV, SPVC & SPVB3 detected by both techniques SP Mastre not detected by VANA ; integrated ? 6 contaminating viruses VANA
Sweet_potato-Montpellier-ZA2Ib	Begomovirus ?	SPCSV SP Mastre Badna Begomo (1 read)	**No SPCSV** **No Mastre** Insect virus Begomo (2 reads)	SPCSV detected by dsRNA, not by VANA SP Mastre not detected by VANA ; integrated ? 5 contaminating viruses VANA

Synthetic analysis

One sample cannot be interpreted in dsRNA (low read number) plus two samples likely exchanged between the two techniques.

Soymo, Mastre (one exception in Bd0346), **Badnas not detected by VANA, all likely integrated (detected but should not be in dsRNA….)**

SPCSV and SPVG detected in only 1 of 2 samples by VANA

Vanilla

Sample name	Virus supposed to be present	Viruses detected by dsRNA	Viruses detected by VANA	Comments
Vanilla-Reunion-CRV0716	New luteo ?	Allexi Betaflexi Potex **No Luteo (no contig, only 1 read)**	Allexi **No Betaflex** **No Potex** Luteo	Allexi detected by both techniques Betaflexi & Potex only by dsRNA, Luteo only by VANA 2 contaminating virus in VANA
Vanilla-Reunion-CRV0853	Possible new Allexi	Allexi Potex **No Betaflex**	Allexi Potex Betaflex **Odontoglossum ringspot virus (no contig, 4 reads as possible conta)**	Allexi & Potex detected by both techniques Betaflex only detected by VANA 3 contaminating viruses VANA
Vanilla-Reunion-CRV2147	Poty	BCMV Allexi Potex	BCMV Allexi Potex	Allexi, Betaflex & Potex detected by both techniques 1 contaminating viruses VANA
Vanilla-Reunion-CRV0064	ORSV, CymMV	ORSV	ORSV	ORSV detected by both techniques Expected CymMV not detected by either technique 2 contaminating viruses VANA
Vanilla-Reunion-CRV0060	ORSV, CymMV	CymMV **No ORSV (no contig, only 1 read)**	CymMV ORSV	CymMV detected by both techniques ORSV detected only by VANA 3 contaminating viruses VANA, 1 still positive by contig
Vanilla-Reunion-CRV1549	BCMV, new potex, new Allexi	CymMV	CymMV	CymMV detected by both techniques **Sample does not contain the expected viruses, including BCMV.** **Problem with numbering?** 2 contaminating viruses VANA
Vanilla-Reunion-CRV2148	Potex, Allexi	BCMV Potex Allexi	BCMV Potex Allexi	BCMV, Potex, Allexi detected by both techniques **Sample contains unexpected BCMV**

Synthetic analysis

Some samples do not contain the expected viruses

Luteo, Betaflex (1 of 2 samples), ORSV (1 of 2 samples) not detected by dsRNA. Betaflex (1 of 2 samples) & Potex (1 of 4) not detected by VANA

Global synthesis

Pooling, number of reads obtained

With VANA all samples generated enough reads and can be interpreted.
With dsRNA 4 samples gave low number of reads (<300) and cannot therefore be easily analyzed. This could suggest inherent difficulties in balancing dsRNA pools, as we have sometimes observed.

Mix-up of samples - contaminations

Occurred in at least two cases, one for dsRNA in sugarcane and one with both techniques in sweet potato. Some interpretation problems may also come from other mix-ups (in particular in Garlic ?) but more difficult to analyze.
VANA consistently gave a background level of low read number contamination with many agents and many samples. A few contaminations are observed with dsRNA but in much lower numbers.

Techniques performance

Generally speaking, both techniques performed well in all tested host plants, allowing the detection of most if not all of the expected viruses.
VANA did not detect Soymo, Badna or Mastre sequences that likely represent integrated viral sequences that are sometimes detected by dsRNA. Conversely and as expected, dsRNA generally did not detected replicating episomal DNA viruses (Mastre, Badna).
Viruses not detected by one technique in a given sample were frequently detected in another sample, or a virus of the same genus could be detected in another host species. This likely indicates that the failure to detect a given agent has more to do with viral concentration and competition with the detection of co-infecting agents than with an inability of a given technique to detect a particular virus or virus genus.
Examples include

VANA:

- No detection of BBrMV in banana or of LYSV in Garlic / detection of other Potys in Sugarcane, Garlic, Sweet Potato and Vanilla
- No detection of Betaflexi in garlic or in Vanilla / detection in other Garlic & Vanilla samples
- No detection of Potex in vanilla / detection in other vanilla samples or detection of other potex (CymMV in vanilla, yam…)
- No detection of Alphaflexi in garlic (detection of Allexis and potex in garlic…
- No detection of clostero in Sugarcane / detection of clostero in other sugarcanes
- No detection of SPCSV in Sweet potato / detection of SPCSV in other SP

A possible counter example concerns IYSV, a Tospovirus, which was not detected in Garlic.

dsRNA

As expected no detection of DNA viruses (but some detections may represent integrated sequences as in SP for example).
Same as VANA, some viruses are not detected in some samples but readily detected in others

- OYDV/LYSV in some Garlic samples not in others
- GCLV in Garlic
- ORSV in Vanilla….

A possible counter-example concerns the detection of Luteoviruses, which generally appears to be poor, since luteovirus contigs were observed in only one of 8 plants (3 Garlic, 4 sugarcane and 1 vanilla) which were positive by VANA. Given that we have readily observed Luteoviral sequences in other situations (BYDV in grasses for example), this may again have more to do with competition with other viruses for detection rather than with an inability of dsRNA analysis to allow the detection of Luteoviruses. **But note that the garlic Luteo had been detected efficiently in 4 plants during the Plates 1-2 experiments, and not detected by VANA. So situation is complex….**

Annex 5 - Development of an annotation pipeline for NGS data analysis and virus identification.

INRA-Bordeaux -

Following the initial steps of demultiplexing and quality trimming of the sequencing reads, the challenge addressed in SafePGR WP3 is to identify viral sequences among the large quantity of NGS data generated. INRA Bordeaux developed a series of tools that allows this task to be performed efficiently and that also allows an easy visualization of the results by the user. Development of such a pipeline was among the deliverables of SafePGR WP3, together with the analysis of the data generated within the project.

The Galaxy interface (https://galaxyproject.org/) was selected to integrate the various tools needed within the pipeline because it provides partners with an user-friendly access better adapted to scientists with no background in bioinformatics. The pipeline is divided in three essential steps, each of which can be implemented using various tools that can be selected depending on the specific needs of the end-user. These three steps are: sequence assembly, annotation and analysis and, finally, representation of results (see Figure).

The first step consist in reads assembly. The pipeline provides the interface for three different bioinformatic tools: Newbler, IDBA UD and MetaVelvet.

- **Newbler** is an assembler developed and provided by Roche and is specialized in treating 454 sequencing reads (which were used throughout SafePGR)(Margulies *et al.*, Nature 2005). Newbler is particularly efficient at dealing with sequencing errors associated with homopolymeric sequence stretches, a type of error frequent in 454 pyrosequencing.
- **IDBA UD** stands for Iterative De Bruijn Assembler for Uneven Depth, and is an assembler specialized in the assembly of paired-end reads generated by Illumina sequencing platforms. Sequencing multiple genomes in one experiment (as in metagenomics analysis such as those performed in SafePGR) creates sequencing depth variations between genomes. Instead of using a simple threshold, IDBA-UD uses a multiple depth-relative threshold to remove errors in both low-depth and high-depth regions and reconstruct longer contigs with higher accuracy (Y. Peng *et al.* Bioinformatics 2012).
- **MetaVelvet** is another metagenomic assembler which has the advantage of handling sequencing data generated using any of the current sequencing platforms. It uses two features, the coverage difference and graph connectivity, for the decomposition of the De Bruijn graph into sub-graphs that may represent single genomes (T. Namiki *et al.*, Nucleic Acids Research 2012).

As the number of reads integrated in each contig is an important consideration for results interpretation, a small tool called **readPerContigs** is also implemented in the pipeline. It is able to extract the read per contigs information from an ACE file (generated by Newbler) or, if another assembler is used, to generate the same information by mapping input reads against contigs. It provides as output the number of reads per contig, the mean depth of each contig, or the RPKM (Read Per Kilobase per Million reads)(A.Mortazavi *et al.*, Nature 2008).

Sequence assembly produces long informative contigs, but information contained in non-assembled reads (singletons) is lost. Since viral infection detection can rely on a single read, we developed a module, **getSingletons**, which is able to retrieve those singletons from either an ACE file or by mapping against contigs when using other assemblers. The retrieved singletons can then be processed, if needed, through the second part (annotation and classification) of the pipeline.

Finally, assembly statistics can be obtained via the **getFastaStats** tool. The tool takes as input multiple fasta files created during the assembly process and generates standard assembly metrics: number of contigs, minimum length, maximum, mean, length standard deviation, sum, N50, L50, N90 and L90.

The second step is the identification and classification of each assembled fragments (contigs annotation, but as explained above, this can also be performed for singletons). This is achieved through similarity search(es) performed by **Blast** against **NCBI non-redundant databanks** (either proteins or nucleotides, using BlastN or BlastX). This is the most time-consuming step because of the high computing requirements. A **customized module** has been developed in order to take advantage of the computing power provided by computer clusters. This module splits the input data, launches multiple Blast analyses in parallel and finally retrieves and collapses all the results in unique output files. This parallelization process is totally transparent for the user.

Taxonomic assignation is performed by the **blast2ecsv** module developed in Bordeaux. It simply links the NCBI gene identifier (gi) given by Blast with the NCBI's taxonomy. The output file is in CSV format with additional information compared to the classical m8 blast format: query length, number of reads for the contig, query and hit coverage for the match, taxonomic identifier, taxonomic string and match description.

Protein functional domain search is performed by **rpstblastn**, which allows to compare contigs against the **CDD database** (Conserved Domain Database, A. Marchler-Bauer *et al.*, Nucleic Acids Research 2005) and/or **Pfam** (Protein families, R.D. Finn *et al.*, Nucleic Acids Research 2014). This module was implemented in the pipeline because it can provide additional clues for the identification of viral fragments, in particular in the case of novel, highly divergent agents, that may be difficult to identify using Blast homology searches.

The third step consists in the representation of results. The module **ecsv2excel** developed in Bordeaux transforms the csv files generated by blast2csv into excel sheets. It can also combine multiple csv file in a single output file as, for example, when merging results from different blast analyses (blastx, blastn, rpsblast...) performed on a single set of contigs. The output excel sheet contains color-coded, highlighted results for easier identification of viral contigs.

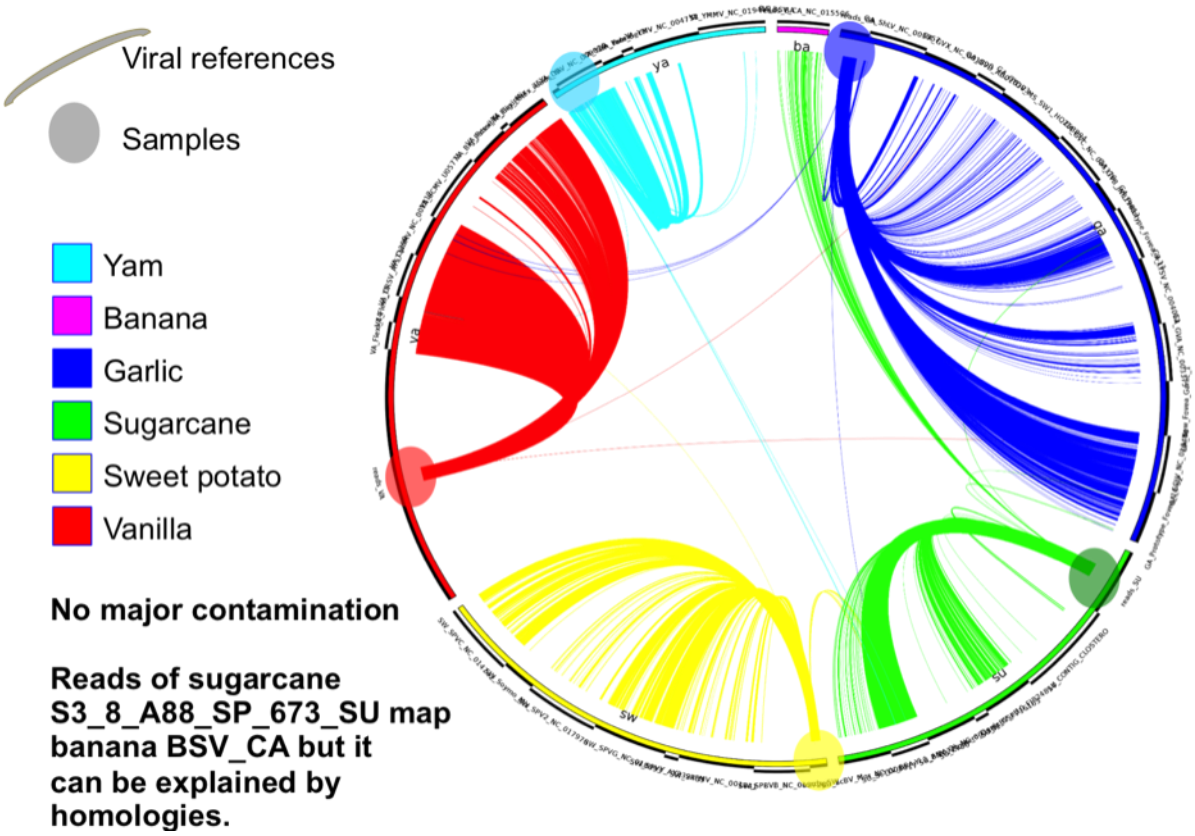
A second integrated module called **ecsv2krona** and is able to produce interactive Krona pie charts (BD. Ondov *et al.*, BMC Bioinformatics 2011) allowing easy hierarchical visualization of taxonomic assignment of all annotated contigs.

Lastly, the module **autoMapper** has been developed to automatically map contigs to a reference assigned by Blast. It takes as input a fasta file containing the contig sequences and the taxonomic annotation file (produced by blast2csv). It first identifies in NCBI the best reference sequence. Since Blast is commonly done on a protein database, the script identifies the best nucleotide reference sequence in a simple manner: if a reference genome for the viral agent considered exists, it is chosen, if not, the longer matching nucleotide sequence is kept. AutoMapper then produces two kinds of outputs: files in fasta format of the contig(s) aligned on the selected reference sequence(s) and html plot files graphically presenting the contig(s) mapped on the reference sequence(s) together with the percentage of identity of the matches. This allows to very rapidly and easily assessing the length, number, position and homology level of viral contigs aligned on the genome of a reference virus.

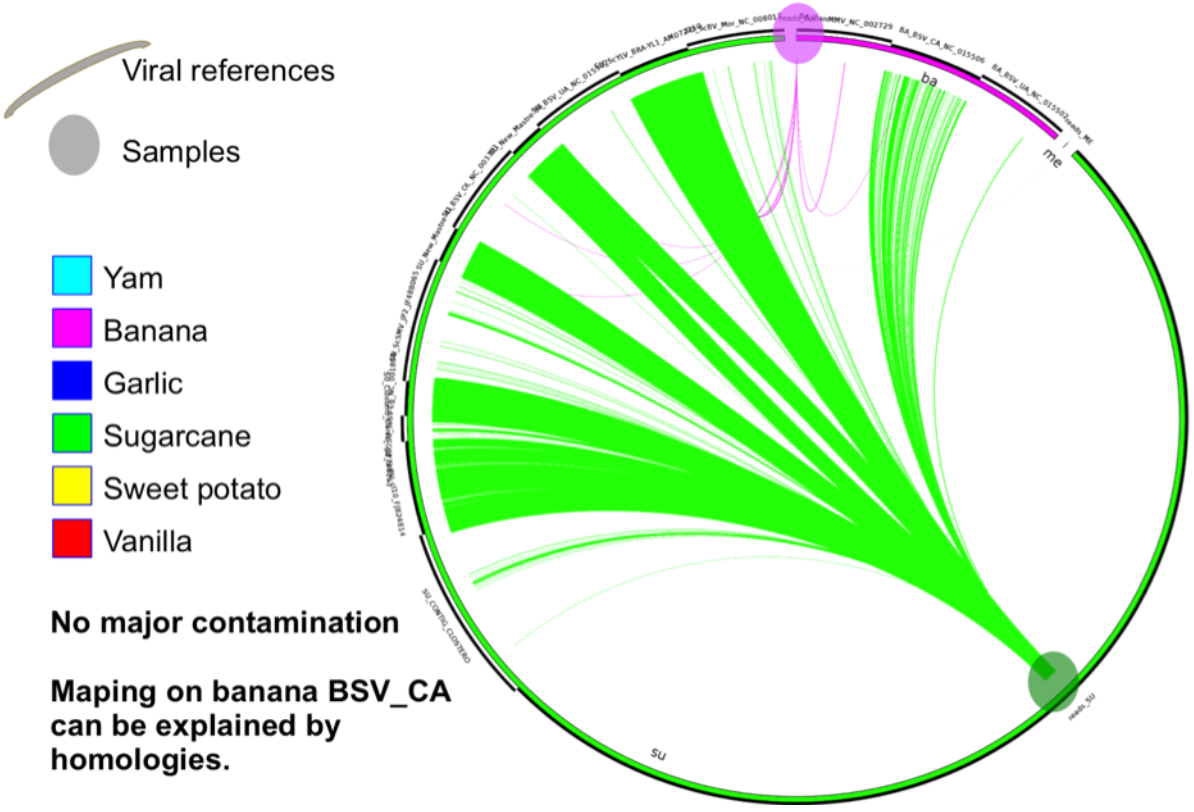
Schematic representation of the Pipeline developed at INRA Bordeaux for the assembly and annotation of NGS data for the identification of viral sequences.

Annex 6 - Results on the bio-informatic analyses.

Mapping against viral references plate 3



Mapping against viral references plate 3



Mapping against viral references plate 3