SNP4OrphanSpecies: A bio-informatic pipeline to isolate robust molecular markers for phylogeny, phylogeographical and population genetic studies of genetically orphan species

Benjamin Penaud, Benoît Laurent, Marine Milhes, Camille Noûs, François Ehrenmann, Cyril Dutech

SNP4OrphanSpecies: A bio-informatic pipeline to isolate robust molecular markers for phylogeny, phylogeographical and population genetic studies of genetically orphan species

Penaud, Benjamin[1]; Laurent, Benoît[1] ; Marine Milhes[2] ; Camille Noûs[3] ; François Erhemann[1] ; Dutech, Cyril[1*]

[1] *INRAE, Univ. Bordeaux, UMR BIOGECO 1202, F-33610, Cestas France*
[2] *INRAE, US 1426, GeT-PlaGe, Genotoul, Castanet-Tolosan, France.*
[3] *Cogitamus laboratory, France*

Since several decades, an increase of disease or pest emergence due to introduction or environmental changes has been recorded, causing serious threats to ecosystems. Many of these events are associated to species having poor or no genomic resources (ie. genetically orphan species). This lack of resource is a serious limitation to understand the origin of emergent populations and to predict future consequences on ecosystems. Analysing population genetic diversity is an efficient method to rapidly obtain this information, but required to have available polymorphic genetic markers. We developed a generic bio-informatic pipeline to rapidly isolate these markers in the goal to be applied in numerous different invasive taxa, with a special focus on fungal pathogen and insect pests. This pipeline is based on a quick de-novo assembly genome obtained from a shot-gun whole genome sequencing, using paired-end Illumina technology, and the isolation of single copy genes conserved in the related species of the emergent organisms. Previous studies have shown that intronic regions of these conserved genes generally contain several single nucleotide polymorphisms at the species level. The pipeline was tested on several invasive or expanding pathogen and pest species in Europe (*Armillaria ostoyae, Bursaphelenchus xylophilus, Diplodia sapinea, Erysiphe alphitoides, Thaumetopoea pityocampa*). For each tested species, we successfully isolated several pools of one hundred short gene regions which can be amplified in multiplex. The bio-informatic pipeline is easy to install and to use (i.e. using the concept of container embedding all the computer programs needed for the processing). It also uses little computer resources (i.e. few hundreds of Go, depending of the genome size of the targeted species). We hope that this robust and rapid method of genetic marker isolation will be useful for numerous laboratories involved in the understanding of biological invasions, but with little available resources in bio-informatic.

Keywords : Biological invasion, single-copy genes, SNP, whole-genome sequencing

** Corresponding author : cyril.dutech@inrae.fr*

Introduction

Pest emergences are an important threat for world ecosystems (Fisher et al. 2012) and human well-being (Diagne et al. 2020). A dramatic increase of these events has been recorded for several decades (Santini et al. 2013). In this special context of invasions or recent expansions of pest and disease, there is an urgent need to identify methods for avoiding or, at least, reducing their spread and their

deleterious effects in ecosystems (Filipe et al. 2012, Gonthier et al. 2014). For many of emerging pests, however, taxonomic and biological knowledge may be weak, with no or little information about their geographical origin, routes of colonization, and ability to adapt to environmental changes, as those induced by human activities (e.g. Gross et al. 2021).

Phylogenetic, phylogeographical and demographic inferences may be quickly obtained from population genetic analyses (e.g. Estoup and Guillemaud 2010, Beichman et al. 2018), provided that genetic markers are available for the emerging species or its species complex (i.e. species from the same genus difficult to morphologically identify in the field). Until recent time, microsatellite markers (or simple sequence repeats, SSR) have represented the most efficient genetic markers for such genetic analyses (Selkoe et al. 2006), although they can be sometimes difficult to genotype without ambiguity, and to isolate from genomes of some species (e.g. Dutech et al. 2007, but see for new more efficient methods, as in Lepais et al. 2020). With the emergence of next generation sequencing (NGS) technologies, SNPs (single nucleotide polymorphism) have become the standard markers for population genetic studies. Out of these methods, GBS (genotyping by sequencing), including for example RAD-seq (restricted-site associated DNA sequencing), allows to quickly obtain thousands of markers by using high-throughput DNA sequencing (e.g. Elshire et al. 2011). These methods are powerful because they generate a huge quantity of genetic information, but they may be costly in time and money, since they generally need a good DNA quality for all the samples analysed, high sequencing coverage to identify duplicated genomic regions, and several computational steps for data validation (Ribeiro et al. 2015). All these requirements may be difficult to reach for some species having a genome with numerous repetitive elements, for which DNA extraction of good quality is not easy to obtain (e.g. Dutech et al. 2020), and for research teams without important bio-informatic resources. In addition, the large genetic information generated by these based-NGS methods may largely outperform the simple need of some genetic markers to resolve the first biological questions addressed at the time of the first steps of emergence (e.g. Brodde et al. 2019).

Several previous studies have shown that is generally easy to isolate SNPs within introns of single copy genes conserved in the genus or the family of targeted species (e.g. Feau et al. 2011, Ilves et al. 2014) or in ultraconserved elements (UCEs; e.g. Blaimer et al. 2015). This idea has been successfully applied for several fungal pathogens, and it allowed to investigate population genetic structures and reproductive modes with a small effort of DNA sequencing and SNP genotyping (Dutech et al. 2017, Tsykun et al. 2017, Dutech et al. 2020). Until now, this method to identify the intron candidates have yielded less than fifty unlinked and polymorphic SNPs per study, and it has not been automatized to reduce time of analysis. The objectives of the present study were therefore to develop and to test an automatic bio-informatic pipeline usable for a large number of species, especially focusing on the emerging pest and pathogen species which are frequently genetically orphan species (i.e. with poor or no genomic resources). To validate the method, we expect that this pipeline yields on each tested species, hundreds of short sequences encompassing at least one intronic region, enhancing the probability to isolate SNPs within these sequences. The specifications of this pipeline, hereafter called “SNP4OrphanSpecies” was to be easily installed and used by biologists, and based on a minimum of genomic resources, in order to rapidly provide useful genetic markers for taxonomic identification, resolution of the origin of emergence, and inferences of population dynamics.

The method of SNP isolation is based on an automatic *de-novo* genome assembly without steps of curation, obtained from a paired-end

Illumina sequencing technology. This assembly is quickly validated by checking some parameters (genome size, degree of assembly fragmentation, completeness of the genome, ...). For isolating hundreds of physically unlinked SNPs, we focused on the single copy genes conserved in genomes at a given taxonomic level (*i.e.* genus, family, or order). The focus on these conserved genes allows, first, to control for the taxonomic status of the analysed genome, second, to remove duplicated genes in the genome which can produce possible false positive SNPs, and third, to isolate several SNPs generally present in the introns of these genes. In the final step, SNP4OrphanSpecies yields several pools of pairs of primers for amplification of around 400bp sequences which can be amplified together in one multiplex (100 sequences per pool). By automatisation of these steps, the method decreases the time of genomic analysis, while it selects robust and validated polymorphic markers for the studied species (discarding sequences due to specific mis-identification known in species complex, or laboratory or field DNA contamination; Ballenghien et al. 2017). According to previous studies (Feau et al. 2011, Dutech et al. 2016), we expect that the isolated sequences with intron regions are polymorphic within species, and valuable candidates for future SNPs isolation. The pipeline was validated on five pest or fungal pathogen of tree species which currently cause dramatic disturbances in European forests. Keeping in mind biologist users, we implemented this pipeline with Snakemake (Köster and Rahmann 2012) and Singularity (Kurtzer et al. 2017). This workflow is easy to install, easy to use, reproducible, and can be run on all Linux machines, including high performance cluster. The pipeline, its associated notice, and parameter files can be downloaded on the Portail Data INRAE https://doi.org/10.15454/GWKRKY.

Brief description of the SNP4OrphanSpecies pipeline

The different steps of the analysis are described in Fig. 1. The first step of the pipeline is the whole-genome de novo assembly using paired-end short reads obtained from Illumina technology. Although not tested in this study, a minimum coverage of 10X is recommended for correcting sequencing errors, and probably 20X minimum to obtain a correct *de novo* assembly (e.g. Jiang et al. 2019). This step starts with a quality analysis of the raw data using FastQC (v0.11.9, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads are trimmed by using a sliding window, and filtered using a minimum length with the software trimmomatic (v0.39, Bolger et al. 2014). The parameters for trimming are defined in the parameter file of the pipeline (Snakemake_Config_SNP4Orphan Species.yaml). The de novo assembly is performed using IDBA-UD (v1.0.6, Peng et al. 2010). A basic statistics report (N50, L50, genome size, ….) is then generated on the final assembly using Quast (v5.0.2, Gurevich et al. 2013). The whole-genome de novo assembly is followed by an identification of single copy conserved genes using BUSCO (v4.1.4, Seppey et al. 2019). This step allows to evaluate the completeness of the assembly, and to identify the genes which will be used for the next steps.
The second step of the pipeline (only for Fungi, Bacteria and Viruses in this first version) is the control for the taxonomic assignment of the assembled contigs and the isolated genes. this assignment is performed using Kaiju (v1.7.4, Menzel et al. 2016). Additionally, for fungal species only, isolation of the internal transcribed spacer (ITS) is performed from the *de novo* genome assembly with ITSx (v1.1b, Bengtsson-Palme et al. 2013), followed by a taxonomic assignment of this region by Kaiju. This optional verification requires a big disk space (ideally 125 Go for the nr_euk database available on the Kaiju web server*).*

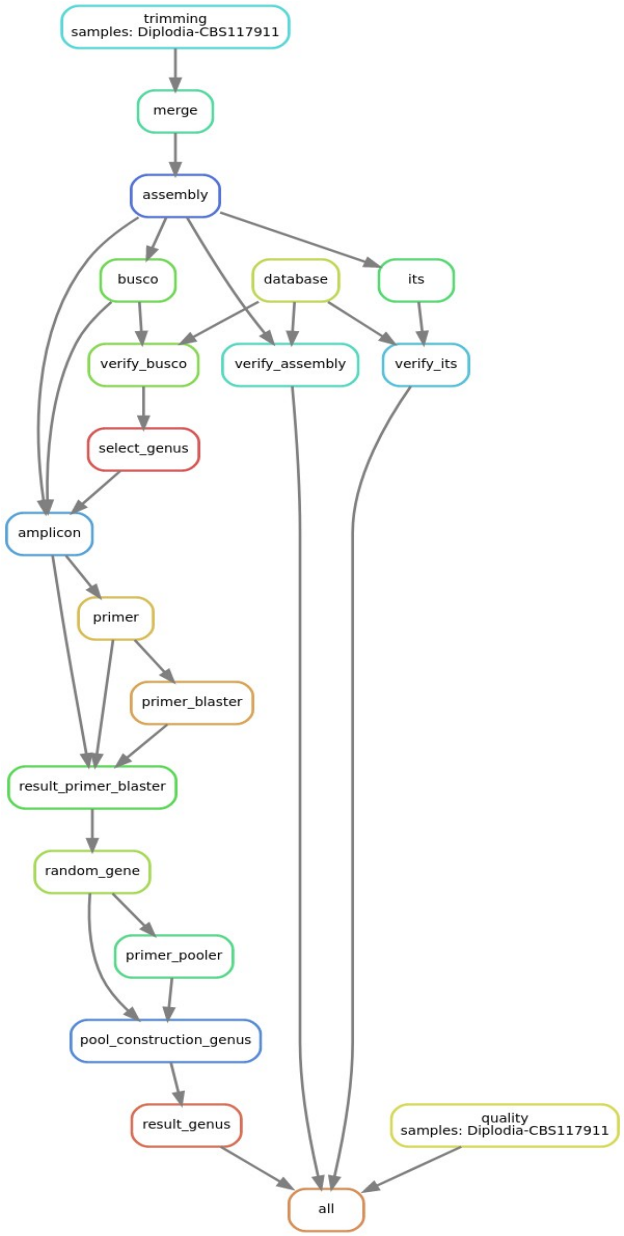

Figure 1: Steps of the SNPs isolation used in SNP4OrphanSpecies

The third step is the isolation of short sequences (400-500 bp hereafter called “amplicon”) to be amplified in pools. For this step, the pipeline selects genes identified by Busco, and with the expected taxonomic assignment (i.e. defined in parameter in the .yaml file). When taxonomy is not defined in the parameter file, genes are selected within the largest taxonomic group given by the automatic gene assignment obtained from BUSCO. Then, the amplicons are chosen to encompass at least one intron in the sequence. For each amplicon, a pair of primers was designed using a home-made Perl script integrating the program Primer3 (v2.5 Koressaar et al. 2007), with stringent parameters favouring the possibility to be amplified jointly in a single multiplex PCR. All the designed primers were blasted against the de-novo genome assembly to test their specificity for the targeted sequences. Each pair of primers for which one of the two primers was found in at least two copies in the genome, was removed. One single pair of primers was finally randomly selected per BUSCO gene to avoid the analysis of physically linked genetic markers. The validated primers are dispatched in several pools for which the primer dimer formation compatibility during a multiplex DNA amplification is tested in silico, using Primer Pooler (v1.71, Brown et al. 2017). A notice available on https://doi.org/10.15454/GWKRKY gives more details about these different steps.

Tests of the pipeline

We tested this pipeline on a new *de novo* genome assembly of the *Diplodia sapinea* isolate CBS117911. *Diplodia sapinea* is a worldwide emergent fungal pathogen infecting a large range of host trees, especially pine species (Brodde et al. 2019). A genomic library was constructed for this isolate using the Illumina TruSeq Nano DNA kit, following the company procedure. A total of 10,544,224 raw 150 bp paired-end sequences was produced on an Illumina HiSeq3000 sequencer at the Get-Plage Genotoul facility (INRAE, France). In addition to this Ascomycete, we also tested this pipeline on four other invasive species from different phyla, and for which a genome assembly has been already published. The first one is another Ascomycete, *Erysiphe alphitoides*, infecting a large range of host plants in the world, and likely introduced in Europe at the beginning of the 20th century (Gross et al. 2021). Genomic resources have recently been published for this obligate

biotroph species (i.e. non-culturable on axenic media), and for which DNA contamination were detected in the genome assembly (Dutech et al. 2020). The second one is a Basidiomycete, *Armillaria ostoyae*, with a large world distribution infecting numerous conifer species (Heinzelmann et al. 2019), and associated with some expanding populations in planted conifer forests, as suggested in south-western France (Labbé et al. 2017). The whole genome sequencing published in Sipos et al. 2017) was used for this test. The third one is a Lepidoptera species, *Thaumetopoea pityocampa*, expanding in Europe due to climatic changes, and causing important defoliation in pine plantations or human health concerns (Battisti et al. 2015). The genome sequencing used to test the pipeline has been published by Gschloessl et al. (2018). Finally, we also used data from the published genome assembly of *Bursaphelenchus xylophilus* (Dayi et al. 2020), a pine wood nematode, infecting several pine species, and introduced in Asia and Europe from North America, where it causes dramatic mortality in forests of the invaded areas (Vicente et al. 2012).

Species	***Diplodia sapinea***	***Erysiphe alphitoides***	***Armillaria ostoyae***	***0,00***	***Bursaphelenchus xylophilus***
class	Dothideomycetes	Leotiomycetes	Agaricomycocetes	0,00	Secermentea
order	Botryosphaeriales	Erysiphales	Agaricales	0,00	Aphelenchida
family	Botryosphaeriaceae	Erysiphaceae	Physalacriaceae	0,00	Parasitaphelenchidae
Reference	This study	Dutech et al. (2019)	Sypos et al. (2017)	0,00	Dayi et al. (2020)
Sequencing	Illumina Hiseq3000	Illumina Hiseq2000	Illumina Hiseq2000	0,00	Illumina Hiseq2000
Strain	CBS117911	MS_42D	C18	0,00	Ka4C1
Number of Reads	10,544,224	369,262,818	116,828,130	0,00	58,326,120
Number of Reads Used to construct the assembly	9,044,726	313,340,218	103,921,206	0,00	55,197,190
Total length	37,650,182	316,911,737	57,720,627	0,00	70,264,222
Nbcontigs>500pb	1,793	131,582	7,119	289,40	10,373
Nbcontigs>1000pb	1,387	79,253	4,666	185,30	7,823
Nbcontigs>50000pb	200	68	215	1,00	76
Largest contig	324,688	102,030	563,590	63,40	148,994
GC(%)	56.71	49,73	48.32	0,00	40.38
N50 (kb)	48,485	3,410	34,291	2,33	15,069
L50 (number)	218	17,657	371	67,37	1,341

Table 1 : Description of the genome assemblies obtained for the five tested species

After the steps of filtering and trimming Illumina reads, new de novo genome assemblies were produced by the pipeline for each species (details of each assembly are given in Table 1). Overall, these genome assemblies were strongly fragmented with small L50 and N50 values, especially compared to their original publication. As expected for a contaminated DNA extraction (Dutech et al. 2020), *E. alphitoides* genome assembly was one of the most fragmented genomes with *T. pityocampa*, and had a surprising estimated genome size for a powdery mildew species (317 Mb *Vs* less than 110 Mb for other published powdery mildew genome assemblies; Frantzeskakis et al. 2019). For fungal species, the identification of the ITS1 sequences using the Kaiju database confirmed that at least a part of the genome assembly may be assigned to the expected genus for each sequenced species (Table S1). Interestingly, ITSx used for this identification, detected several ITS1 in the *E. alphitoides* genome, and congruent with several contigs of the genome assembly which can be assigned to different phyla or fungal families (Figure S1).

Between 98.6 (*D. sapinea*) and 42 % (*T. pityocampa*) of conserved single-copy genes listed in the Busco database were isolated from the genome assemblies (Table 2). For the tested fungal species and using Kaiju, a variable

proportion of genes were actually identified as different from the targeted genus, leading to discard between 70% (*E. alphitoides*) and 0.6 % (*A. ostoyae*) of the initial list of single-copy genes. In the last steps of the analysis, the pipeline defined in each species between 20,991 (*A. ostoyae*) and 1,829 (*E. alphitoides*) short 400bp sequences (i.e. amplicons), encompassing at least one intron region. The design of the pairs of primers for DNA amplification for each amplicon (one single per gene), and the control for their potential duplication in the genomes yielded a final set of pairs of primer between 614 (*E. alphitoides*) and 3,426 (*A. ostoyae*) pairs (Table 2). All these pairs of primers are pooled in five pools for multiple DNA amplification. Depending on the final number of designed pairs of primers, the redundancy rates of pair of primers among the primer pools for each species varied between 19 % (*A. ostoyae*) and 57.4 % (*E. alphitoides*) (Table 2). This redundancy could be manually optimized among pools when rates are too high. Actually, Primer Pooler was not designed to build several pools of primers at the same time, and it may useful to sequentially remove the pairs of primers used in the first pools to build the next pools. Raw sequencing data of *D. sapinea* genome, the output files produced by this pipeline from this new genome assembly, the five pools of the isolated amplicons for each tested species, and their pairs of primers for DNA amplification are available on https://doi.org/10.15454/ GWKRKY.For *D. sapinea*, this bio-informatic analysis was performed on the CBIB Linux cluster (called cortex Centos Linux 7) in 1h02mn25sec using 20 cpus, and it generated 9.1 Go of results files.

Species	*Diplodia sapinea*	*Erysiphe alphitoides*	*Armillaria ostoyae*	*Thaumetopoea pityocampa*	*Bursaphelenchus xylophilus*
Nb of Busco genes	3,786	3,234	3,870	5,286	3,131
Nb of Complete single-copy	3,733	2,353	3,787	2,219	2,068
Nb with the validated genus	3,557	987	3,765	NA	NA
Nb of defined amplicons	6,962	1,829	20,991	3,163	13,256
Nb of genes with amplicons	2,760	685	3,438	887	1,955
Nb of pairs of primers	6,095	1,408	20,617	1,916	10,938
Nb of conserved pairs of primers*	2,570	614	3,426	672	1,928
% gene duplication in pools	20.8	57.4	19	52.8	23

Table 2 : Summary of the genes and primers isolated by SNP4Orphanspecies pipeline for the five tested species

* one single per gene and not duplicated in the genome

Perspectives for population genetic analyses of non-model invasive and emergent species

For all the tested species, the pipeline isolated more than 600 and up to 3,426 short sequences, each one located on a different gene. The different phyla tested in this study suggested that the method can be used for a large number of invasive or emergent species for which genetic markers are looked for. Some limitations may occur for large genomes (i.e. several hundreds Mb), since the assembly, even without any curation steps, requires a minimum of computation resources. For example, the analysis of the *T. pityocampa* genome for which the size was estimated to be more than 500Mb, generated 268 Go (134 Go only for the trimmed fastaQ files), and took more than 18 hrs on a linux cluster using 20 cpus. For such a large genome, it could be interesting to test the method with a reduced whole-genome sequencing (i.e. a lower sequencing coverage, or randomly amplified genome). Based on our results obtained from the highly and

contaminated *E. alphitoides* genome assembly, we speculate that several hundreds of the conserved single-copy genes can be generally isolated by the present method, even from a low-quality or partial genome assembly. Another limitation would be the use of contaminated genome assemblies which may be quite frequent in whole-genome sequencing (Ballenghien et al. 2017), and especially for not easily cultivated micro-organisms. The smallest number of validated sequences was obtained for *E. alphitoides* for which such contamination was assumed (Dutech et al. 2020), and confirmed in this study. Such a contamination, when it involves genetically related species (for example between fungal species), may be difficult to identify and to remove from the genomic data, because of the genetic similarity among the sequences. The use of the Kaiju database to assign the isolated genes is then useful to detect this DNA contamination, and to discard the sequences with the incorrect taxonomic assignment. This contamination likely contributed to remove a large number of amplicon candidates for *E. alphitoides*. First, the similarity among sequences of related taxa may increase the gene duplication rates of the targeted genes. Second, contamination may increase the selection of genes using the BUSCO procedure which are finally discarded after controlling for their taxonomic assignment in fungal species (see Table 2 for example with *E. alphitoides*). In addition, in the final step, for *E. alphitoides*, less primers are validated than for other fungal species, again likely because of potential duplication of these conserved genes within the genome assembly.

Finally, several hundreds, and for three out the five tested species, several thousands of pairs of primers have been designed offering the possibility to develop enough markers for first genetic studies. Several designed primers within species were however automatically integrated in two or more multiplex pools, with duplication rates of primers between 20 and 60% between pools. However, in several cases, this duplication is small (less than 10%) between two pools of primers. For example, for *D. sapinea*, the combination of pool 1 and 3 allows to potentially amplify a total of 193 different amplicons (i.e. only 7 duplicated pairs of primers between these two pools). This example showed that most of the time, the careful selection of the pools of primers should allow to strongly reduce this global redundancy rate among the five automatically designed pools. Two strategies can be then developed after the isolation of these amplicons. The first one would be the sequencing of hundreds samples using one of the designed pools, and next-generation sequencers. The combination of SNPs identified in each amplicon can be treated as microhaplotypes (i.e. multi-SNP loci), that potentially gives more power for population genetic analyses than using bi-allelic SNP loci, and also reduced the need to identify several hundreds of SNPs for a robust genetic analysis (Baetscher et al. 2018). Microhaplotypes defined on short sequences, as those isolated by this pipeline, may be well adapted to identify fine population genetic structures, especially for field samples for which quality of DNA extraction is often poor (e.g. Morin et al. 2021), or for a correct assignment of samples to populations and estimates of population admixture (McKinney et al. 2017). In a preliminary PCR test on 47 samples of *D. sapinea*, most of the 100 pairs of primers tested in multiplex allowed to obtain more than 10 sequencing reads per samples in average. This last result is encouraging for new tests on other pools of primers, and on other genetically orphan species. A second option would be the genotyping of each SNP combined in different pools (i.e. plex) and on several hundreds of samples, using for example a Mass-ARRAY technology (e.g. Chancerel et al. 2013, Dutech et al. 2017). Variations within each amplicon can be first identified by resequencing a ten of genomes or partial genomes, and then, several SNP-plex can be designed for genotyping hundreds of samples (e.g. Dutech et al. 2016).

Although multi-SNP locus information is missed in this case, it allows to genetically characterize more populations from different geographical regions at a limited cost, a central objective when the genetic origin of emergent populations is investigated.

We are aware that SNPs isolated from conserved genes may be under selection. It may seriously affect the inferences of demographic dynamics of populations and should be carefully considered if historical scenarii are tested (Beichman et al. 2018). A first study to identify selection on these conserved genes in *Armillaria sp*., detected 2 out of 20 tested (Dutech et al. 2016). Notwithstanding this potential bias, we argue that for a first estimation of population genetic structures, for phylogeny, or identification of the genetic origin of the emergent populations, the methodology presented here remains efficient. No significant difference on the estimates of genetic structures has been observed when comparing SSR and SNP loci isolated using this method in several European populations of *A. cepistipes* (Tsykun et al. 2017). Basic statistics can also identify loci under selection (e.g. Vitalis et al. 2003). In addition, because these loci are chosen in conserved genes at a given taxonomic level, they may be especially relevant to distinguish between closely related species and their hybrids. Cryptic species which are difficult to identify using morphological criteria, are frequently observed in invasive fungal pathogen species (e.g. in Gross et al. 2021), and they may be identified with these molecular markers. Some loci, such as SSR loci, are sometimes difficult to transfer even within the same genus (e.g. Dutech et al. 2007), and may produce large number of null alleles and missing data. By contrast, the short sequences obtained by this pipeline would be especially relevant for these questions of phylogeny and phylogeography among closely related species, because of the good repeatability and the standardization of genotyping among species, experiments or laboratories, as well as the assessment of the sequence orthology within genomes. They may be an efficient alternative to RAD-seq methods for studying species complex, or for comparing genetic studies produced by different studies or laboratories (see Harvey et al. 2016 for details).

We hope that this pipeline will be used to rapidly improve our knowledge of emerging species in the context of global changes. We designed this method especially in direction of research teams where human and financial resources are limited. We also consider that the time of bio-informatic analyses to isolate and to develop new markers is also seriously reduced thanks to this pipeline, quickly installed on a personal computer, provided that a sufficient computer memory is available. Then, the possibility to rapidly obtain first information on recently emerging populations should help to identify the origin of this emergence, the risks for the ecosystems, and to define the best practices to manage disease or pest species.

Acknowledgment

We thank Pedro Crous (Westerdijk Fungal Biodiversity Institute, Netherlands) for providing us the *Diplodia sapinea* isolate, and G. Sypos and C. Simang (University of Sopron, Hungary) for having sent us the raw sequencing data of the C17 *Armillaria* isolate genome. We also thank O. Lepais for comments on the previous version of this manuscript. The genome of the *Diplodia* isolate was performed in collaboration with the GeT core facility, Toulouse, France (http://get.genotoul.fr). GeT was supported by France Génomique National infrastructure, funded as part of "Investissement d'avenir" program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09). Preliminary tests were performed at the Genome Transcriptome Facility of Bordeaux (grants from the Conseil Regional d'Aquitaine n°20030304002FA and 20040305003FA, the European Union, FEDER n°2003227 and Investissements d'avenir, N°ANR-10-EQPX-16-01). Part of computational resources and infrastructure used in present publication were provided by the Bordeaux Bioinformatics Center (CbiB). This research has benefited of the European HOMED project and received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°771271.

References

Baetscher DS, Clemento AJ, Ng TC, Anderson EC,

Garza JC (2018) Microhaplotypes Provide Increased Power from Short-Read DNA Sequences for Relationship Inference. Molecular Ecology Resources 18 (2): 296–305. https://doi.org/10.1111/1755-0998.12737.

Ballenghien M, Faivre N ,Galtier N (2017) Patterns of Cross-Contamination in a Multispecies Population Genomic Project: Detection, Quantification, Impact, and Solutions. BMC Biology 15 . https://doi.org/10.1186/s12915-017-0366-6

Battisti, A, Larsson S (2015) Climate Change and Insect Pest Distribution Range. In Climate Change and Insect Pests, edited by C Bjorkman and P Niemela, 7:1–15. https://doi.org/10.1079/9781780643786.000

Beichman AC, Huerta-Sanchez E , Lohmueller KE (2018) Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms In Annuel Review of Ecology, Evolution and Systematics 49:433–456. https://doi.org/10.1146/annurev-ecolsys-110617-062431

Bengtsson-Palme J, Ryberg M, Hartmann M, Branco S, Wang Z, Godhe A, De Wit P, et al. (2013) Improved Software Detection and Extraction of ITS1 and ITS2 from Ribosomal ITS Sequences of Fungi and Other Eukaryotes for Analysis of Environmental Sequencing Data. Methods in Ecology and Evolution 4 (10): 914–919. https://doi.org/10.1111/2041-210X.12073

Blaimer BB, Brady SG, Schultz TR, Lloyd MW, Fisher BL, Ward PS (2015) Phylogenomic Methods Outperform Traditional Multi-Locus Approaches in Resolving Deep Evolutionary History: A Case Study of Formicine Ants. BMC Evolutionary Biology 15: 271. https://doi.org/10.1186/s12862-015-0552-5

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. Bioinformatics 30 (15): 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brodde L, Adamson K, Camarero JJ, Castano C, Drenkhan R, Lehtijarvi A, Luchi N, et al. (2019) Diplodia Tip Blight on Its Way to the North: Drivers of Disease Emergence in Northern Europe. Frontiers in Plant Science 9. https://doi.org/10.3389/fpls.2018.01818

Brown SS, Chen Y-W, Wang M, Clipson A, Ochoa E, Du M-Q (2017) PrimerPooler: Automated Primer Pooling to Prepare Library for Targeted Sequencing. Biology Methods and Protocols 2 (1) bpx006. https://doi.org/10.1093/biomethods/bpx006

Chancerel E, Lamy J-B, Lesur I, Noirot C, Klopp C, Ehrenmann F, Boury C, et al. (2013) High-Density Linkage Mapping in a Pine Tree Reveals a Genomic Region Associated with Inbreeding Depression and Provides Clues to the Extent and Distribution of Meiotic Recombination. BMC Biology 11 (1):1–19. https://doi.org/10.1186/1741-7007-11-50

Dayi M, Sun S, Maeda Y, Tanaka R, Yoshida A, Tsai IJ, Kikuchi T (2020) Nearly Complete Genome Assembly of the Pinewood Nematode *Bursaphelenchus Xylophilus* Strain Ka4C1. Microbiology Resource Announcements. 9 (42). https://doi.org/10.1128/MRA.01002-20

Diagne C, Leroy B, Gozlan RE, Vaissiere A-C, Assailly C, Nuninger L, Roiz D, Jourdain F, Jaric I, Courchamp F (2020) InvaCost, a Public Database of the Economic Costs of Biological Invasions Worldwide. Scientific Data 7 (1). https://doi.org/10.1038/s41597-020-00586-z

Dutech C, Enjalbert J, Fournier E, Delmotte F, Barrès B, Carlier J, Tharreau D, Giraud T (2007) Challenges of Microsatellite Isolation in Fungi. Fungal Genetics and Biology 44 (10): 933–949. https://doi.org/10.1016/J.FGB.2007.05.003

Dutech C, Prospero S, Heinzelmann R, Fabreguettes O, Feau N (2016) Rapid Identification of Polymorphic Sequences in Non-Model Fungal Species: The PHYLORPH Method Tested in Armillaria Species. Forest Pathology 46 (4): 298–308. https://doi.org/10.1111/EFP.12256

Dutech C, Labbé F, Capdevielle X, Lung-Escarmant B (2017) Genetic Analysis Reveals Efficient Sexual Spore Dispersal at a Fine Spatial Scale in *Armillaria Ostoyae*, the Causal Agent of Root-Rot Disease in Conifers. Fungal Biolgy 121 (6–7): 550–560. https://doi.org/10.1016/J.FUNBIO.2017.03.001

Dutech C, Feau N, Lesur I, Ehrenmann F, Letellier T, Li B, Mouden C, Guichoux E, Desprez-Loustau M-L, Gross A (2020) An Easy and Robust Method for Isolation and Validation of Single-Nucleotide Polymorphic Markers from a First *Erysiphe Alphitoides* Draft Genome. Mycological Progress 19 (6): 615–628. https://doi.org/10.1007/s11557-020-01580-w

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* 6 (5): 19379.

https://doi.org/10.1371/JOURNAL.PONE.0019379

Estoup A, Guillemaud T (2010) Reconstructing Routes of Invasion Using Genetic Data: Why, How and so What? *Molecular Ecology* 19 (19): 4113–4130. https://doi.org/10.1111/j.1365-294X.2010.04773.x

Feau N, Decourcelle T, Husson C, Desprez-Loustau M-L, Dutech C (2011) Finding Single Copy Genes Out of Sequenced Genomes for Multilocus Phylogenetics in Non-Model Fungi. PLoS ONE 6 (4): e18803. https://doi.org/10.1371/journal.pone.0018803

Filipe JAN, Cobb RC, Meentemeyer RK, Lee CA, Valachovic YS, Cook AR, Rizzo DM, Gilligan CA (2012) Landscape Epidemiology and Control of Pathogens with Cryptic and Long-Distance Dispersal: Sudden Oak Death in Northern Californian Forests. PLos Computational Biology 8 (1). https://doi.org/10.1371/journal.pcbi.1002328

Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ (2012) Emerging Fungal Threats to Animal, Plant and Ecosystem Health. *Nature* 484 (7393):186–194. https://doi.org/10.1038/nature10947

Frantzeskakis L, Németh MZ, Barsoum M, Kusch S, Kiss L, Takamatsu S, and Panstruga R (2019) The *Parauncinula Polyspora* Draft Genome Provides Insights into Patterns of Gene Erosion and Genome Expansion in Powdery Mildew Fungi. MBio 10 (5). https://doi.org/10.1128/MBIO.01692-19

Gonthier P, Anselmi N, Capretti P, Bussotti F, Feducci M, Giordano L, Honorati T, et al. (2014) An Integrated Approach to Control the Introduced Forest Pathogen *Heterobasidion Irregulare* in Europe. Forestry 87 (4): 471–481. https://doi.org/10.1093/forestry/cpu015

Gross, A, Petitcollin C, Dutech C, Ly B, Massot M, Faivre d'Arcier J, Dubois L, Saint-Jean G, Desprez-Loustau M-L (2021) Hidden Invasion and Niche Contraction Revealed by Herbaria Specimens in the Fungal Complex Causing Oak Powdery Mildew in Europe. Biological invasions 23 (3): 885–901. https://doi.org/10.1007/s10530-020-02409-z

Gschloessl B, Dorkeld F, Berges H, Beydon G, Bouchez O, Branco M, Bretaudeau A, et al. (2018) Draft Genome and Reference Transcriptomic Resources for the Urticating Pine Defoliator *Thaumetopoea Pityocampa* (*Lepidoptera: Notodontidae*). Molecular Ecology Resources 18 (3): 602–619. https://doi.org/10.1111/1755-0998.12756

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: Quality Assessment Tool for Genome Assemblies. Bioinformatics 29 (8): 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2016) Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Systematics Biology 65 (5): 910–924. https://doi.org/10.1093/sysbio/syw036

Heinzelmann R, Dutech C, Tsykun T, Labbé F, Soularue J-P, Prospero S (2019) Latest Advances and Future Perspectives in *Armillaria* Research. Canadian Journal of Plant Pathology 41 (1): 1–23. https://doi.org/10.1080/07060661.2018.1558284

Ilves KL, Lopez-Fernandez H (2014) A Targeted Next-Generation Sequencing Toolkit for Exon-Based Cichlid Phylogenomics. Molecular Ecology Resources 14 (4): 802–811. https://doi.org/10.1111/1755-0998.12222

Jiang Y, Jiang Y, Wang S, Zhang Q, Ding X (2019) Optimal Sequencing Depth Design for Whole Genome Re-sequencing in Pigs. *BMC Bioinformatics* 20: 556 https://doi.org/10.1186/s12859-019-3164-z

Koressaar T, Remm M (2007) Enhancements and Modifications of Primer Design Program Primer3. Bioinformatics 23 (10): 1289–1291. https://doi.org/10.1093/bioinformatics/btm091

Köster J, Rahmann S (2012) Snakemake-a Scalable Bioinformatics Workflow Engine. Bioinformatics 28 (19): 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Kurtzer GM, Sochat V, Bauer MW (2017) Singularity: Scientific Containers for Mobility of Compute. PLos One 12 (5). https://doi.org/10.1371/journal.pone.0177459

Lepais O, Chancerel E, Boury C, Salin F, Manicki A, Taillebois L, Dutech C, et al. (2020) Fast Sequence-Based Microsatellite Genotyping Development Workflow. PEERJ 8. https://doi.org/10.7717/peerj.9085

McKinney GJ, Seeb JE, Seeb LW (2017) Managing Mixed-Stock Fisheries: Genotyping Multi-SNP Haplotypes Increases Power for Genetic Stock Identification. Canadian Journal of Fisheries and Aquatic Sciences 74 (4): 429–434. https://doi.org/10.1139/cjfas-2016-0443

Menzel P, Ng KL, Krogh A (2016) Fast and Sensitive Taxonomic Classification for Metagenomics

with Kaiju. Nature Communications 7 (1): 1–9 https://doi.org/10.1038/ncomms11257

Morin PA, Forester BR, Forney KA, Crossman CA, Hancock-Hanser BL, Robertson KM, Barrett-Lennard LG, et al. (2021) Population Structure in a Continuously Distributed Coastal Marine Species, the Harbor Porpoise, Based on Microhaplotypes Derived from Poor-Quality Samples. Molecular Ecology 30 (6): 1457–1476. https://doi.org/10.1111/mec.15827

Peng Y, Leung H C M, Yiu S M, Chin F Y L (2010) IDBA - A Practical Iterative de Bruijn Graph De Novo Assembler. In: Berger B. (eds) Research in Computational Molecular Biology. RECOMB 2010. Lecture Notes in Computer Science, vol 6044. Springer, Berlin, Heidelberg. pp 426-440 https://doi.org/10.1007/978-3-642-12683-3_28

Ribeiro A, Golicz A, Hackett CA, Milne I, Stephen G, Marshall D, Flavell AJ, Bayer M (2015) An Investigation of Causes of False Positive Single Nucleotide Polymorphisms Using Simulated Reads from a Small Eukaryote Genome. BMC Bioinformatics 16. https://doi.org/10.1186/s12859-015-0801-z

Santini A, Ghelardini L, De Pace C, Desprez-Loustau M-L, Capretti P, Chandelier A, Cech T, et al. (2013) Biogeographical Patterns and Determinants of Invasion by Forest Pathogens in Europe. New Phytologist 197 (1): 238–250. https://doi.org/10.1111/j.1469-8137.2012.04364.x

Selkoe KA, Toonen RJ (2006) Microsatellites for Ecologists: A Practical Guide to Using and Evaluating Microsatellite Markers. Ecology Letters 9 (5): 615–629. https://doi.org/10.1111/j.1461-0248.2006.00889.x.

Seppey M, Manni M, Zdobnov EM (2019) BUSCO: Assessing Genome Assembly and Annotation Completeness. In Gene Prediction: Methods and Protocols, edited by M Kollmar, 1962:227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Sipos G, Prasanna AN, Walter MC, O'Connor E, Bálint B, Krizsán K, Kiss B, et al. (2017) Genome Expansion and Lineage-Specific Genetic Innovations in the Forest Pathogenic Fungi Armillaria. Nature Ecology & Evolution 1 (12). Nat Ecol Evol: 1931–1941 https://doi.org/10.1038/S41559-017-0347-8

Tsykun T, Rellstab C, Dutech C, Sipos G, Prospero S (2017) Comparative Assessment of SSR and SNP Markers for Inferring the Population Genetic Structure of the Common Fungus *Armillaria Cepistipes*. Heredity 119 (5): 371–380. https://doi.org/10.1038/hdy.2017.48

Vicente C, Espada M, Vieira P, Mota M (2012) Pine Wilt Disease: A Threat to European Forestry. European Journal of Plant Pathology 133 (1): 89–99. https://doi.org/10.1007/s10658-011-9924-x

Vitalis R, Dawson KD, Boursot P, Belkhir K (2003) DetSel 1.0: A Computer Program to Detect Markers Responding to Selection. Journal of Heredity 94(5):429–431. https://doi.org/10.1093/jhered/esg083