Phylogenomic and population history inference using ancient DNA

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

SERGIO MAURICIO LATORRE OCHOA

aus Bogotá / Kolumbien

Tübingen

2020

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 18.09.2020

Stellvertretender Dekan:	Prof. Dr. József Fortágh
1. Berichterstatter:	Prof. Dr. Hernán A. Burbano
2. Berichterstatter:	Prof. Dr. Oliver Bossdorf

Zusammenfassung

Evolutionsstudien haben sich die enorme Menge genomischer Daten zunutze gemacht, die nach der Revolution der Next-Generation-Sequencing Plattformen generiert wurden. Dieses Momentum ging einher mit bedeutenden Entwicklungen bei der DNA-Isolation, Protokollen zur Herstellung von Sequenzier-Bibliotheken, sowie einem stetigen Anstieg der Rechenkapazität für die Verarbeitung großer Datensätze. Da Evolution die Untersuchung der Akkumulation genetischer Variation über die Zeit ist, kann der direkte Zugang zu Momentaufnahmen genetischer Variation in verschiedenen historischen Perioden die Genauigkeit und das Verständnis evolutionärer Rückschlüsse bemerkenswert erhöhen. Dies wurde durch die Sequenzier-Revolution ermöglicht, da nun der technischen Zugang zu Genomen alter und historischer Proben, z.B. subfossiler Aufzeichnungen, Museumspräparate und Herbariumsbelege, erschlossen wurde.

Durch die Studie sogenannter 'ancient DNA' (alter DNA) aus verschiedenen Quellen habe ich die phylogenomische Verwandtschaft und die demographische Geschichte in zwei verschiedenen Systemen untersucht. Zum einen nutzte ich Museumspräparate ausgestorbener *Oryctes*-Skarabäuskäfer aus den Maskarenen um in Kombination mit frischen Präparaten und den daraus ermittelten hochwertigen Genom-Assemblierungen deren phylogenomische Verwandtschaft zu untersuchen. Dies erlaubte mir erstmalig, die evolutionären Beziehungen zwischen diesen Käfern mit Hilfe der molekularen Phylogenetik sowie D-Statistiken zu beschreiben. Darüber hinaus entdeckte ich auf der Grundlage dieser Rekonstruktion zwei vermutlich unabhängige Kolonisationen der Insel Reunión durch diese Käfergattung, platzierte eine zuvor nicht zur Gruppe gehörende Art innerhalb der Gattung, und terminierte den Verlust ihrer Flugfähigkeit und ihres Zwergwuchses auf einen Zeitraum nach der Kolonisierung. Diese Ergebnisse sind wegweisend für die Verwendung von 'ancient DNA' aus Museumspräparaten ausgestorbener Arten bei der Untersuchung der Phylogenomik von Insekten.

Darüber hinaus habe ich die Populationsstruktur und die demographische Geschichte des Reisbohrpilzes *Magnaporthe oryzae* studiert. Dazu nutze ich eine zeitliche Serie von herbarisierten infizierten Reisblättern und die daraus

entschlüsselten Genome zusammen mit einem globalen Datensatz. Ich konnte zeigen, dass die aktuelle Populationsstruktur des Pilzes in drei pandemische klonale Linien aufgeteilt ist, sowie in ein Cluster rekombinierender Individuen nahe des postulierten Ursprungszentrums. Anhand der Herbariumsbelege konnte ich die Entstehung der pandemischen klonalen Linien auf vor weniger als 400 Jahre abschätzen. Durch die Untersuchung des Repertoires von Effektorgenen in den untersuchten Genomen konnte ich zeigen, dass verschiedene Muster ihres Verlusts und Gewinns mit den verschiedenen klonalen Linien assoziiert sind. Diese Effektorgene können nun funktionell untersucht werden bezüglich ihrer Auswirkung auf die Fitness des Pathogens.

Abstract

Evolutionary studies have taken advantage of the massive amount of genomic data produced after the revolution of the next generation sequencing platforms. This momentum has been accompanied by major developments in DNA isolation methods, library preparation protocols and a steady increase in computational capacity to handle big datasets. Because evolution is the study of the accumulation of genetic variation through time, the direct access to snapshots of genetic variation through different historical periods can remarkably increase the accuracy and understanding of evolutionary inferences. Ancient DNA research has profited from the sequencing revolution by enabling the technical access to the genomes of ancient and historical sources e.g., subfossil records, museum specimens and herbarium samples.

Making use of ancient DNA sources, I investigated the phylogenomic relationships and demographic histories in two different systems. First, with the combined use of extinct pinned insect museum specimens and high-quality sequenced reads and assemblies derived from present-day specimens, I sought to investigate the phylogenomic relations between the *Oryctes* scarab beetles in the Mascarene islands. For the first time, I ascertained the evolutionary relationships among those beetles using molecular phylogenetics and site patterns-based *D*-statistics. Moreover, and based on this reconstruction, I discovered two likely independent colonization events in the Reunión island and, a post-colonization case of dwarfism and loss of flight ability in one of the species, before classified as an outgroup. These results pave the way on the use of ancient DNA and extinct museum specimens in the study of insect phylogenomics as one of the most biologically important yet underrepresented taxa.

Additionally, I analyzed the population structure and demographic history of the rice blast fungus *Magnaporthe oryzae*. By combining the temporal resolution of genomes from infected rice leaf herbarium samples together with a worldwide dataset, I showed that the current population structure of the pathogen is grouped in three pandemic clonal lineages and a cluster of recombining individuals mainly grouped in their likely centre or origin. With the use of the herbarium samples, I estimated the time of origin of the

pandemic clonal lineages to less than 400 years ago. Finally, by inspecting the repertoire of effector genes of the isolates, I showed different patterns of loss and gains of effector genes associated with the different clonal lineages. This work opens the way to functional tests about the role of effector gains and loss through time and their effect on the overall pathogen fitness.

Acknowledgements

First, I want to thank Hernán A. Burbano for his mentorship and constant supervision during my entire PhD. Working with you has greatly improved my scientific knowledge and has inspired me to always keep an attitude of humbleness, rigurosity and permanent questioning. Many thanks for that. Thanks also for setting up a great environment for constant discussion and help in your group.

I also want to thank Detlef Weigel for creating such a great working atmosphere in which I always felt welcomed. Many thanks for your ample support during my entire doctorate and for your input during seminars and TAC meetings. I would also like to thank Oliver Bossdorf and Claude Becker for agreeing to be part of my committee and for their valuable feedback during my TAC meetings. Many thanks to Marja Timmermans for agreeing to be part of my thesis defense committee as well as Oliver Bossdorf and Detlef Weigel.

I want to specially thank Rebecca Schwab for being a great support in many aspects. Certainly my PhD was bearable with your constant advice and help.

I also want to thank the many collaborators and people who contributed to my projects. In respect to the research on museum phylogenomics of extinct beetles, I would like to thank Rafal Gutaker for laboratory assistance, including the implementation of the museum DNA extraction protocol; E. Reiter for access to clean-room facilities and technical support; Vrinda Venu for technical support during linked-read library preparations; Felicity Jones for useful input during linked-read the library preparations and genome assembly; Kay Pruefer, Aida Andres and members of the Burbano and Sommer laboratories for useful discussions and input on data analysis; Detlef Weigel for supporting my visit to London during the final stages of the project; and A. Andres, R. Gutaker, Talia Karasov and Michael Werner for comments on the manuscript. We are also indebted to the Parc national de la Réunion and the Office National des Forets on Réunion for long-term support and permits; the CIRAD at St. Pierre de la Réunion for housing the Max Planck laboratory; Jaques Rochat and Micropoda for entomological expertise and logistic support; and Max Barclay from the Natural History Museum London and Giulio Cuccodoro from the Natural

History Museum Geneva for providing historical museum samples. In respect to the research project on the genetic history of the rice blast, I want to thank Adeline Harant for the DNA isolation on the contemporary italian *M. oryzae* samples; Sonja Kersten for the construction of the Nextera libraries on the same isolates; E. Reiter for access to clean-room facilities and technical support; Michael Dannemann, Aida Andrés and members of the Kamoun and Burbano laboratories for useful discussions and input on data analysis; Talia Karasov and Francois Balloux for comments on the manuscript; and Detlef Weigel for supporting my visit to London during the final stages of the project.

I am grateful to all my great colleagues at the Max Planck Institute. Specially to Hülya Wicher for her support in many matters.

Working in Tübingen allowed me to meet friends that became part of my family: Julián R., Sebastian G., Clemens W., Moisés E. Patricia L., Cristina B., Jorge K., Adrián C., Tess R., Thanvi S., Gülüm A., Talia K., Mike W., Rafal G., Cris Z., Effie S., Rewati T., Aseem B., Piotrek W., Annia G., Alejandra D., Jacobo O., Francesco A., Prateek K., Diana A., George D., Sofia D., Monika Z., Leonardo M., Andrea B., Brigit WV., Marko WV., Sara F., Mohannad D., Lamia A., Vahid S., Albane R., Isabella C., Zeliha B., Elena J., Alba G., Edu L., Ainhoa I. Xixi C., Simone L., Ezgi D., Emre B.

To Mariaelena for much love.

And, of course: Pá, Má, Hugo y Juli.

General Remarks

- In accordance with the standard scientific protocol, the personal pronoun ‘we’ will be used to indicate the reader and the writer, or my scientific collaborators and myself.

Contents

Prologue

This thesis is organized as follows. Initially, I will introduce the theoretical background relevant to the work presented here, including particular aspects of ancient DNA research in the context of phylogenomics and population genetics In particular, I will briefly present a historical overview about ancient DNA (aDNA) research. This perspective will allow me to explain the unique contribution of aDNA research to evolutionary studies, the technical challenges derived from the degraded nature of aDNA, as well as state-of-the-art molecular biology and computational methods to sequence and analyze aDNA. Subsequently, I will introduce specific phylogenomic developments relevant for the work presented here. I will finalize the introduction, by describing the significance of different population genomic analyses to quantify genetic diversity and infer the different evolutionary forces that gave rise to the observed patterns. The introduction is followed by two chapters, each of which will have its own introductory section relevant for its scope and biological studied system. Consequently, each chapter will have its own Materials and Methods sections, as well as its results, discussions and particular conclusions. I will conclude this thesis by discussing altogether, the most relevant conclusions and the major outlook of my dissertation in a broader context.

1. Introduction

The basis for evolutionary studies is the accumulation of genetic variation through time. However, as in many fields, the discrepancy between the real phenomena and the measurements heavily relies on theoretical developments, as well as on the advances of methods and technologies.

Traditionally, analyses with high organisms used morphological differences as well as enzymatic profiles and few genetic marker genes to build evolutionary and phylogenetic relations among them. The reason behind the study of morphological and biochemical properties of proteins is that they were easily quantifiable and, at the same time, are a good proxy for the underlying molecular processes that were difficult to measure at that time.

With the rapid development of the DNA isolation protocols, as well as with the development of DNA sequencing through, at least, three major technological leaps, the use of genomic data as the primary source of information has become a gold standard to understand and study evolution. Nowadays, decreasing costs of Whole Genome Sequences (WGS), have made possible to harness and improve existing methods while opening the possibility to explore new niches for genomic information e.g., subfossil records, museum specimens and herbaria samples. In fact, taking advantage of those developments, ancient DNA (aDNA) studies are complementing evolutionary studies and shedding light into a diverse set of evolutionary questions.

1.1. Ancient DNA

Advances in high throughput sequencing technologies have allowed the sequencing of whole-genome sequences from multiple individuals from different taxa. Along with the increasing amount of available data, evolutionary studies have shifted from few phenotypic and molecular markers to genome-wide markers and genomes, which improve the accuracy of evolutionary inferences. Nevertheless, modern genomic information provides only indirect evidence about the past and the underlying historical processes that gave rise to present-day diversity (Hofreiter, Serre, et al., 2001). Hence, the

possibility to sample and sequence DNA retrieved from historical and ancient samples increases the confidence of evolutionary inference.

The aDNA-based framework has recently emerged as a real solution because it opens the potential of a great variety of historical sources from which it is possible to retrieve DNA, e.g: subfossil records, historical museum specimens and herbaria samples. Through these precious samples, it is possible to examine directly the genetic diversity at a given specific time in the past. There is however a major tradeoff by working with aDNA: greater temporal signal in the data at the expense of reduced DNA quality due to the degraded nature of the DNA retrieved from old sources. Therefore, some experimental cautions and methodological corrections should be considered when working with aDNA.

Once an organism dies, its organic molecules start a decay phase and in the specific case of nucleic acid molecules, different processes contribute to DNA degradation. Three main types of biochemical reactions take place when DNA is exposed to both environmental conditions and the action of endonucleases: i) oxidative damage (Höss et al., 1996), ii) depurination and, iii) hydrolytic damage (Hofreiter, Serre, et al., 2001). Those reactions leave specific and distinctive footprints on the molecules of aDNA.

Particularly, as a consequence of the combined action of the hydrolytic depurination, aDNA molecules suffer a process of β'-elimination, generating patterns of single-strand breaks and fragmentation of the DNA molecules (Lindahl, 1993). Moreover, both oxidative damage of the nucleotides (Höss et al., 1996) and cross-links between DNA and other macromolecules like proteins (Poinar et al., 1998) can cause blocking lesions that prevent polymerase-mediated DNA synthesis. However, blocking lesions have been found to be less frequent than originally assumed (Heyn et al., 2010). Finally, one of the most important signatures that is a specific type of hydrolytic damage is cytosine deamination: a process in which cytosine residues (C) become uracil residues (U) due to the loss of the amino group. As a consequence, adenine residues (A) are incorporated during DNA replication leading to apparent C to T substitutions when DNA molecules are sequenced (Dabney, Meyer, et al., 2013; Hofreiter, Serre, et al., 2001) (Figure 1).

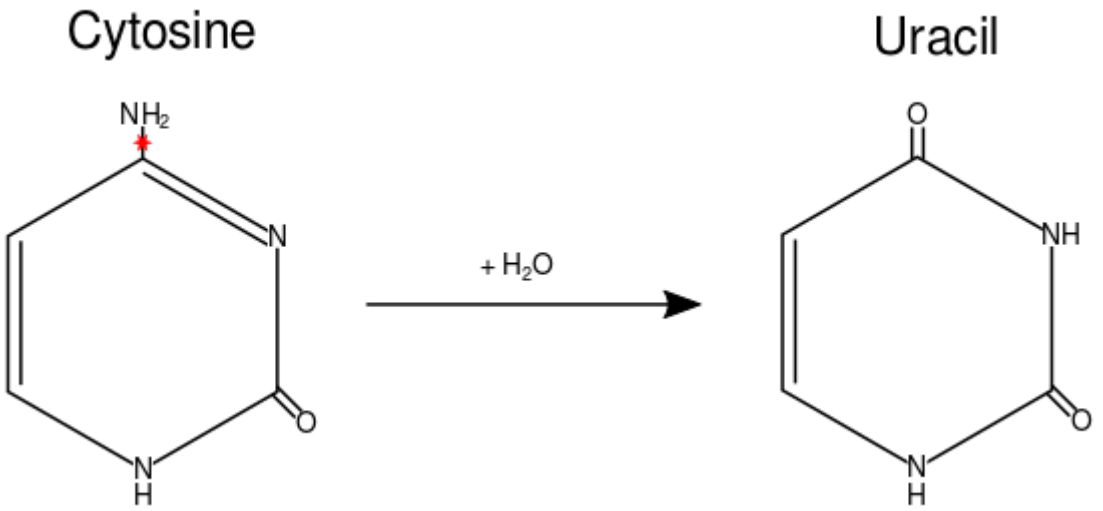


Figure 1. Hydrolytic deamination of cytosine residues. The red star indicates the amino group where the hydrolytic reaction takes place.

The described biochemical properties of aDNA molecules have direct consequences in the way DNA extractions, genomic libraries and sequencing protocols are implemented. In particular, the short fragment size caused mainly by the action of depurination has led to the implementation of protocols that retrieve ultra-short DNA fragments, and index-based library preparations suitable to be tailored with second-generation sequencing technologies like Illumina (Dabney, Knapp, et al., 2013; Gutaker et al., 2017; Kircher et al., 2012; M. Meyer & Kircher, 2010). An extra advantage, particularly for the analysis of short DNA molecules, is that pair-ended libraries will increase base calling confidence since each molecule is likely to be covered twice during sequencing due to their short length (Kircher, 2012).

Once molecules have been sequenced and the genomic raw data is available, another type of aDNA characteristic becomes more relevant. Indeed, cytosine deamination represents the most important source of error for aDNA genomic data. As mentioned before, after the sequencing of aDNA molecules, an excess of C-to-T transitions is observed because U's coming from deaminated C's residues are sequenced as apparent T's (Briggs et al., 2007). This characteristic has been shown to be enriched towards the extremes of DNA molecules since single-stranded DNA hangouts are more likely to suffer deamination (Hofreiter,

Jaenicke, et al., 2001; Lindahl et al., 1977; Pääbo et al., 2004). Distinguishing between deamination-caused transitions from true substitutions poses a challenge which can be addressed either using molecular biology protocols or bioinformatic approaches.

When the effect caused by deamination is either pervasive or undesired in the downstream analyses, biochemical enzymatic steps can be incorporated during the library construction process. The use of Uracil DNA Glycosylase (UDG) (Lindahl et al., 1977), can remove the uracil residues caused by deamination. The use of UDG reduces the majority of apparent C-to-T transitions while removing the main source of systematic error for downstream genomic analyses (Briggs et al., 2010)..

Authentication and computational tools for aDNA

Although the unique features of aDNA molecules generate challenges for genomic analysis, such aDNA-specific features can be used to authenticate historical samples and differentiate their genomic information from external sources of contamination: microorganisms that have colonized the sample *post-mortem*; technical contamination that arises during sample or handling stages; as well as crossed contamination during the DNA isolation and library preparation (Cooper & Poinar, 2000).

By taking advantage on the temporal footprints of DNA decay, aDNA genomic data can be authenticated by inspecting: i) depurination signatures , ii) small fragment sizes that lead to DNA fragmentation and ii) presence of enriched cytosine deamination patterns manifested as C-to-T substitutions (Figure 2).

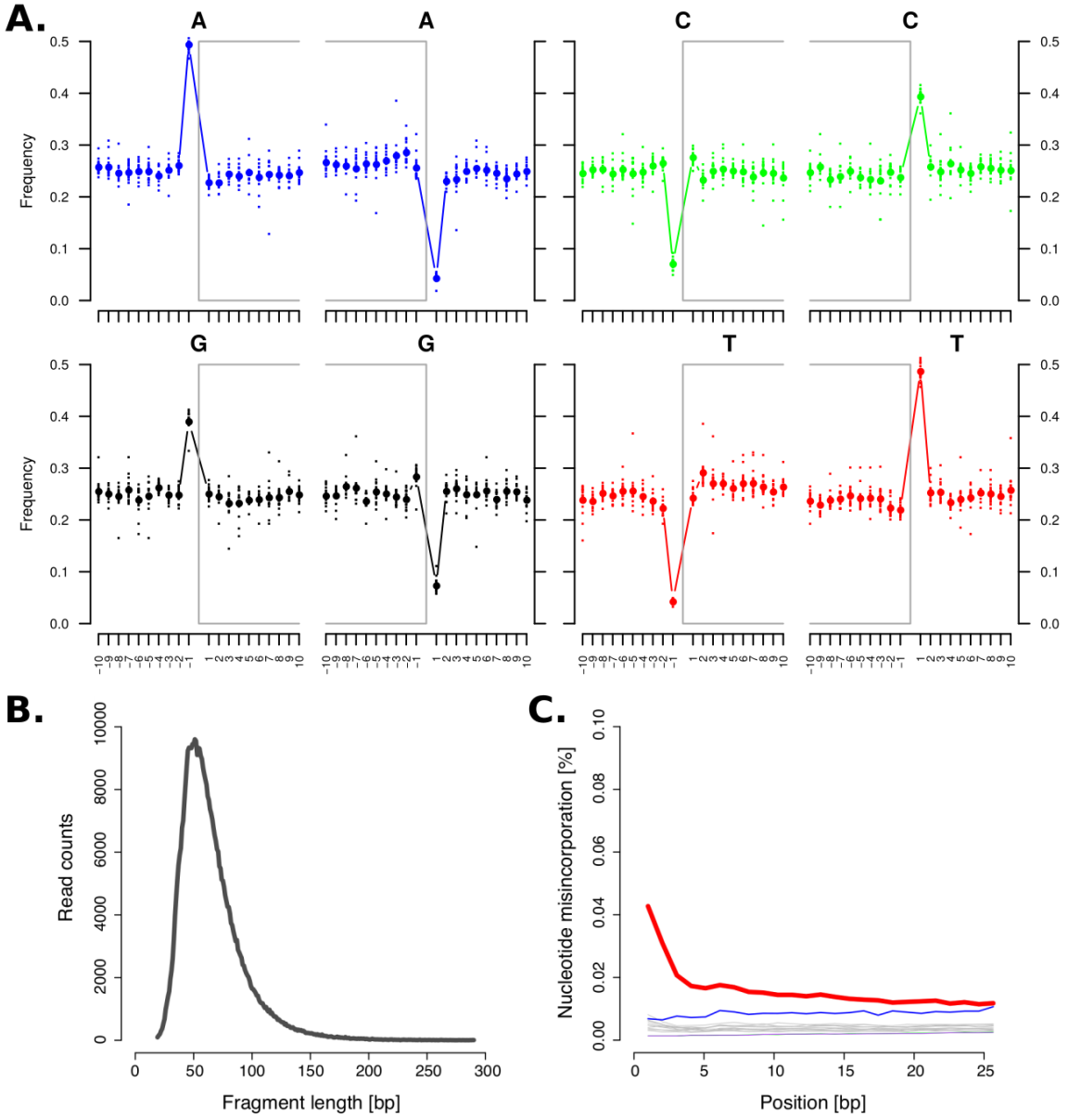


Figure 2. Unique aDNA characteristics used to authenticate historic samples. (A) Purines (adenines and guanines) are enriched at DNA breakpoints in the reference genome used for mapping.. The figure depicts typical breaking points in aDNA molecules (plot produced with MapDamage 2.0 (Jónsson et al., 2013) using a historic sample of *Magnaporthe oryzae* collected in 1891). **(B)** In consequence, distribution of aDNA fragment lengths are short. **(C)** The red line depicts cytosine to thymine substitutions when fragments are aligned to a reference genome. Blue lines depict guanine to adenine substitutions, and grey lines represent other substitutions. (**B. and C.** adapted from Gutaker and Burbano (Gutaker & Burbano, 2016)).

The assessment of the before mentioned signatures, require the existence of a relatively close reference genome to which the data should be aligned and therefore computational mapping methods suitable to work with the intrinsic aDNA characteristics (Prüfer et al., 2010; Schubert et al., 2012).

Most of the commonly used reference mapping software and pipelines are heuristic methods. Such methods rely on the presence of a small-sized exact match (seed) from which the mapping extension will begin (Schubert et al., 2012). Due to the fact that in aDNA molecules, most of the deamination-driven C-to-T substitutions are expected to be enriched towards the extremes of the reads, the possibility of finding exact matches will decrease in proportion to the level of deamination. Hence, the use of mapping algorithms that relax the stringency in the seed or positional dependence substitution matrices are suitable to handle aDNA genome mapping and aligning (Kircher, 2012; Schubert et al., 2012).

Moreover, since the basis of genomic evolutionary studies heavily relies on a robust assessment of nucleotide differences between ortholog regions, it is fundamental to reduce as possible the main source of genetic false-positive polymorphic signal originated from deamination. The use of simulation of deamination and decay process on modern genomic data implemented in *gargammel* (Renaud et al., 2016), the recalibration by downgrading the mapping quality scores in the extremes of the reads where deamination is more prone to happen performed by *mapDamage 2.0* (Jónsson et al., 2013), the use of optimized variant callers for aDNA data like *snapAD* (Prüfer, 2018), as well as the ascertainment of variable positions with curated datasets, are part of a useful toolbox warranted to be utilized in aDNA research. Several software suites and pipelines aimed to handle aDNA-type genomic information have been developed in the last decade (Hanghøj et al., 2016; Jónsson et al., 2013; Peltzer et al., 2016; Schubert et al., 2014; Skoglund et al., 2014).

1.2. Ancient DNA and phylogenomics

Understanding the relationship and common histories between lineages and species is one of the major and most important challenges for evolutionary studies. The way different lineages adapt to a variety of pressures and

environmental conditions with physiological changes but also with the underlying dynamics in their genomic-scale changes are a cornerstone of the evolutionary analysis.

Thanks to the new advances in genomic sequencing technologies, the availability of full species genomes has opened the possibility to increase the level of resolution in the inference of phylogenetic relationships between lineages and species (Eisen, 1998; Philippe et al., 2005). The fact that all exonic and intronic regions can be included in a phylogenetic framework, reduces the sampling error associated with the selection of few genes for phylogenetic reconstruction. The potential consequences of reconstructing erroneous phylogenetic relationships due to the high variance of mutation rates along the genome can get significantly reduced by increasing the amount of data input, e.g. whole genome sequences (Philippe et al., 2005). For instance, the resolution of the true phylogenetic relationship between gorillas, chimpanzees and humans, was only possible when a longer genomic region of non-coding DNA (10.8 kb) was utilized (Miyamoto et al., 1988), after unsuccessful attempts by inspecting single genes (Koop et al., 1986). During the last decade, empirical studies have studied the probability of calculating incongruent phylogenetic topologies as a result of genomic sampling (Rokas et al., 2003). Moreover, because the evolutionary course can be expressed as a function of time, a better understanding of the main evolutionary forces enables a better inference of temporal events, like speciation or lineage coalescence. Thus, more information coming from bigger genomic regions encode more and a better temporal signal.

The reasoning behind the presence of a sequence-embedded temporal signal is that mutations accumulate at different rates in different lineages. The fidelity in which DNA is replicated and repaired varies across different taxa, enabling some lineages to accumulate mutations at a faster pace than their counterparts in the tree of life (Lynch, 2010). While some small fraction of those mutations can be beneficial, it has been shown that most of them have either neutral or slightly deleterious effects (Drake, 2006; T. Ohta, 1973; Tomoko Ohta, 2002). Nevertheless, because mutations are the ultimate source of all variation, organisms with limited survival strategies e.g. prokaryotes with small genome

sizes and high effective population sizes, display, in general, high mutation rates.

Because mutations are the ultimate source of all variation (Lynch, 2010; Lynch & Conery, 2003), phylogenetic reconstructions are heavily impacted by their accumulation. Therefore, phylogenetic analyses can be utilized as estimators of relative time units between lineages: specifically, in an evolutionary neutral scenario, the length of the branches are expected to be a good estimator of the time separating different leaves and nodes in a given topology. Indeed, the idea that substitutions were accumulated in a clockwise manner, was first highlighted by analyzing aminoacid substitutions in both albumin and hemoglobin data collected from primates (Wilson & Sarich, 1969).

Dating a phylogenetic reconstruction

Since phylogenies encode a measure of a relative temporal signal on their topology, the conversion from relative time units to absolute time, can be performed by three different ways: *i*) with the addition of prior information on the substitution rate, *ii*) by informing about divergence times on the internal nodes of the tree generally by the use of dated fossil records and, *iii*) through the incorporation of heterochronous data or time-stamped samples (Heled & Drummond, 2012).

In a scenario in which individuals from the same species are analyzed or in which the phylogenetic distance among the lineages is close enough, it is fair to assume that mutation rate is stable through the whole tree in a way that the above mentioned *a posteriori* dating strategies will effectively convert the relative time encoded in branch lengths into absolute time. The use of both mutation rates and information about dates in fossil records (Heled & Drummond, 2012; Ho et al., 2011) can be modelled as prior distributions in Bayesian frameworks and their power will be heavily dependent on the accuracy of the estimations. For example, the discordance between displayed morphological characters in fossil records and evolutionary information from genomic sequences can lead to incorrect temporal inferences (Ronquist et al., 2016; C. Zhang & Wang, 2019).

With the progressive accumulation of time-series data in genomic databases, it was possible to use individuals from the same species or genetic group yet

sampled in different years. It was shown, that especially in organisms in which their mutation rate is fast, like viruses, a few years of sampling difference was enough to make it possible to calibrate trees and therefore estimate mutation rates (Rambaut, 2000). Because the collected information was assigned to calibrate the tips of a given topology with the precision of the sampling date, in contrast to an estimated range of time from fossil records, this strategy is becoming a powerful tool to estimate absolute time events (Rieux & Balloux, 2016). Briefly, if the sampling time difference between two lineages is significant enough relative to their relation to an outgroup, it is possible to use the observed genetic distance of these two lineages to estimate an evolutionary rate (Figure 3).

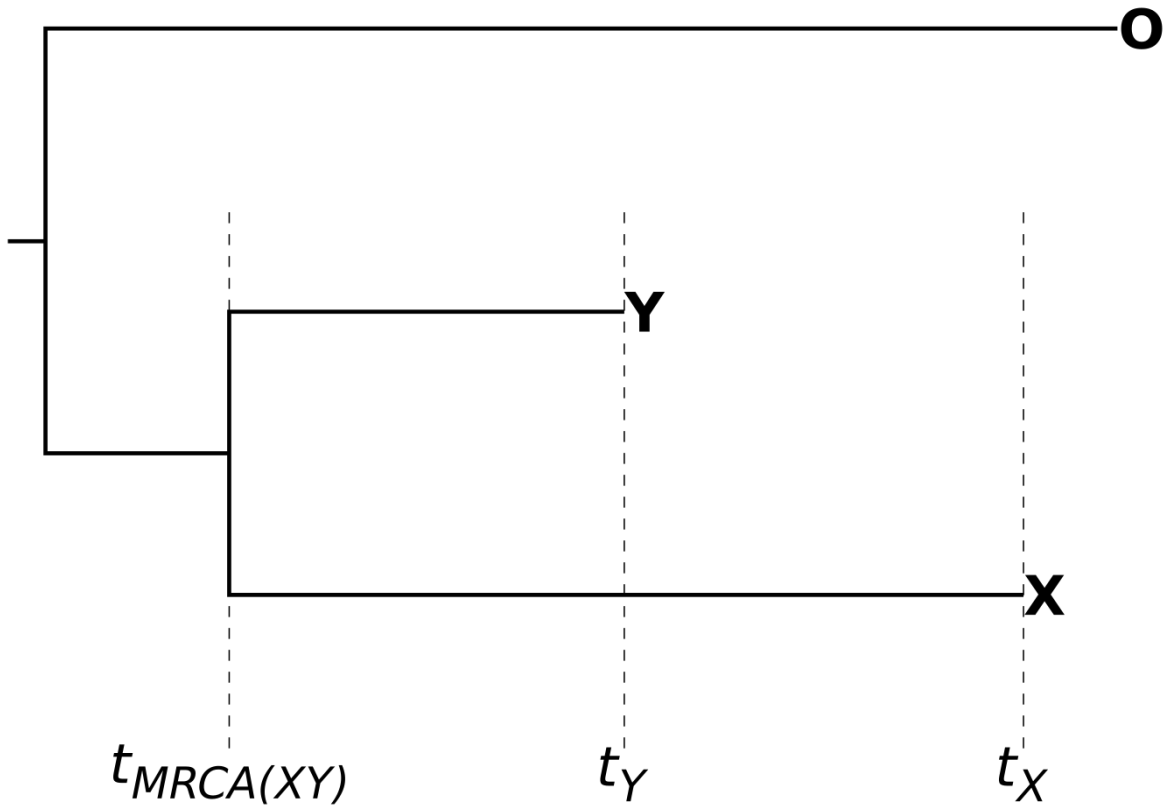


Figure 3. Schematics for tip-dating calibration. In the given phylogenetic tree, the evolutionary rate can be inferred from the relative genetic distance from an **O**utgroup to both **X** and **Y** divided by the sampling differential time between **X** and **Y**: $\mu = (d(OX) - d(OY)) / (t_X - t_Y)$. Adapted from (Rieux & Balloux, 2016).

Because the uniformity of the evolutionary rate between lineages is an important assumption, the addition of more distant taxa to this phylogenetic approach, can both underestimate and overestimate branch lengths for some lineages in which their mutation rates are slower or faster. Thus, adaptations and extensions that allow different parts of the topology to have different molecular clock models (relaxed clock models), have better estimation accuracy (Alexei J. Drummond et al., 2006; Heled & Drummond, 2012). Furthermore, even when analyzing single species, it has been shown that rates of evolution can have extraordinary variability, as is the case of *Yersinia pestis (Cui et al., 2013)*.

Finally, because the proposed tip-dating strategy was limited to the historical records in databases and especially to organisms with fast mutation rates, it is possible to use the power of aDNA to increase the resolution and power of detection of temporal parameters by directly sampling the observed diversity in the past (Green & Shapiro, 2013; Gutaker & Burbano, 2016; Yoshida et al., 2014). For example, with the use of a mitochondrial aDNA dataset with a temporal resolution spanning 40.000 years, it was possible to both improve the estimation of mitochondrial mutation rate in humans and calculate divergence times in human populations with high accuracy (Q. Fu et al., 2013). Moreover, the use of this framework enables the reconstruction of historical demographic histories on different lineages. For instance, the relation between the drastic environmental changes associated with the last glacial cycle and the population decimation of megafauna has been possible to be reconstructed due to the use of aDNA samples of extinct Bisons (Beth Shapiro et al., 2004) and Cave Bears (Stiller et al., 2010). The reconstruction of the dynamics of Effective Population Sizes gains accuracy with the temporal information embedded in aDNA samples. Finally, by taking advantage of the vast collections represented in worldwide herbarium samples and insect museum collections (Tegelberg et al., 2014), it is possible to expand the reconstruction of aDNA evolutionary histories of a diverse set of species (Lang et al., 2019).

1.3. Population genomics

Population assessment

Analysis of genetic populations conceptually assumes the existence of formed populations. It also suggests a historical origin and a process of differentiation into the extant populations which are the target of the analyses.

In order to both understand the origins of populations and analyse their evolutionary course, Fisher (R. A. Fisher, 1931) and Wright (S. Wright, 1931) proposed a model in which population dynamics are studied from a theoretical standpoint. This model assumes a population of constant size in which individuals are hermaphrodites and no mating restrictions or biases are present. This ideal panmictic population and its characteristics can be used to estimate the expected change of allele frequencies in a scenario where no selection takes part, i.e. just the action of the arrival of random mutations and their shuffling via recombination through generations. This is, in fact, a null model from which deviations driven by different types of evolutionary processes can be analyzed.

Furthermore, because of the action of demographic events that generate population subdivision, once individuals are isolated from its main population, the stochastic accumulation of mutations in incipient subpopulations together with a random sampling of reproducing individuals - better known as *genetic drift* - will ultimately give rise to new populations. The genomic manifestation of these processes can be quantified as a reduction or deficit of the frequency of heterozygous alleles, an observation known as the Wahlund effect (Wahlund, 1928).

Since both, the gaining of new mutations and reduction of frequency of heterozygous alleles are continuous and progressive processes, any attempt to identify and separate individuals in populations will resort, in some degree, to *ad hoc* thresholds. Thus, different methods that can establish thresholds to classify individuals into genetic groups or populations have been developed. These methods take advantage of the increased power of genome-wide genetic variation by either using dimensionality reduction multivariate techniques or analyzing patterns of allele sharing.

As part of the dimensionality reduction methods, Principal Component Analysis (PCA) is suitable for coarse-grained population analysis in large and complex datasets (Jolliffe, 2002). Particularly, whole-genome SNP datasets can

be considered as data matrices that reflect the genomic relatedness among individuals. Hence, a PCA can be used to infer any type of underlying structure in an unsupervised manner (Patterson et al., 2006). The robustness and quality of the identified clusters can be assessed *a posteriori* by methods based on the properties of cohesion and separation between clusters like Silhouette Scores (Lovmar et al., 2005; Rousseeuw, 1987) or k-means based algorithms like Discriminant Analysis of Principal Components (Jombart et al., 2010). Moreover, the identification of genetic clusters or populations can be carried out by using the shared evolutionary history between individuals. Specifically, f3-outgroup statistics measure the amount of shared derived alleles of two individuals since they diverged from a common outgroup and therefore f3-outgroup statistics can be interpreted as a measure of pairwise genetic distance between individuals relative to an outgroup species (Raghavan et al., 2014). Finally, another possibility is to infer populations by identifying individuals from which their allele frequencies are consistent with Hardy–Weinberg equilibrium, as this theoretical stage, assumes no intervention from non-neutral evolutionary forces and therefore it can be used as a proxy to identify the genetic structure (Pritchard et al., 2000; Rosenberg et al., 2002; Tang et al., 2005).

Population genomics metrics

Even though the definition of populations and subpopulations can be artificial to some degree, there are also common characteristics shared by the members of a given genetic cluster. Their relative distances within and among clusters, the composition of their variation and the distinctive footprint of different evolutionary processes, are some of the characteristics that give uniqueness to each population.

A major informative metric is the degree of genetic diversity within each population. In fact, the amount of time and generations that a population has experienced and the presence of any disruptive process like selection or different types of demographic histories, have consequences in both the amount and the type of exhibited variation within each population. The parameter *theta* (**θ**) describes and summarizes the diversity of a population

assuming neutrality and proxies of θ can be inferred directly from genomic data.

Two approximations to estimate θ from genomic data can take advantage of the sum of pairwise polymorphic sites in a given set of sequences ((π or θ_π) (Nei & Li, 1979)) or the total number of segregating sites ((θ_W) (Watterson, 1975)). Under a model of neutral-equilibrium, both estimators should be equal and be themselves a good estimator of the theoretical θ. Different implementations of these estimators have been recently developed, tailoring their calculations with the actual data sizes from new sequencing platforms while taking into account missing data (Begun et al., 2007; Ferretti et al., 2012). Moreover, another way of comparing diversity between different populations is by measuring the amount of Wahlund effect, that is, the proportion of deficit of heterozygotes from a neutral-equilibrium expectation between two populations. In other words the correlations between alleles within subpopulations relative to the whole population (Holsinger & Weir, 2009). The Fixation Index (F_{ST}) (Sewall Wright, 1949), was developed with such purpose and has become a standard tool in the analysis of population differentiation. Genome-wide outlier values of F_{ST} can indicate regions under positive selection.

Although the above-mentioned estimators can give an idea of the amount of diversity within and between populations, other types of statistics and estimators can give more information regarding the nature and characteristics of the observed diversity based on the Allele Frequency Spectrum or the Site Frequency Spectrum (SFS) (Braverman et al., 1995). By computing the frequencies of minor or derived alleles in a population, it is possible to infer whether the population have the expected frequencies under neutrality, i.e., only affected by mutation and drift (Evans et al., 2007), or whether the frequencies deviate from expectations, suggesting the indirect effect of selection - linked selection - on allele frequencies. Moreover, summary statistics derived from the SFS have been commonly used to describe such deviations. Using the difference between different estimators of θ, the Tajima's *D* genetic test is effective in detecting deviations from expected allele frequencies in the presence of selection or non-equilibrium demographic histories (Tajima, 1989).

A common characteristic of the above-mentioned tests is that all them detect changes in the frequency of alleles or in the overall genetic diversity. Such

changes are driven by evolutionary processes that affect the parameter **θ**. There is, however, an important biological process that can have the same apparent manifestation, although its underlying cause is the creation of new haplotype diversity while maintaining allele frequencies. Indeed, in the presence of recombination, the gamete sorting process will generate a rearrangement of haplotypes that are only possible to differentiate from other evolutionary forces because of the display of four gamete sets. Assuming two loci with two alleles each, the only possibility to generate four gametes is under recombination (or recurrent mutations). Based on this observation, Hudson and Kaplan (Hudson & Kaplan, 1985) and later Myers and Griffiths (S. R. Myers & Griffiths, 2003) developed the basis of the four-gamete test by estimating the minimum number of recombination events that must have occurred in a sample, so four gamete sets can be observed in the genetic sequences.

Signatures of population genomic measures in clonal and recombinant populations

Analysis of genome-wide variation between and among populations permit to describe among others, the preferred strategy of reproduction of different types of organisms. Whether individuals grouped in a population are experiencing sexual recombination or instead asexuality or clonality; whether those populations are experiencing reduction or expansion of their population size. Those are relevant questions that the before explained analyses and theoretical body can help to understand.

Specifically, sexual recombination has been extensively studied and its footprint on different genomic analysis are well described (Felsenstein, 1974; Griffiths, 1981; Hudson & Kaplan, 1985). One of the most appreciable signatures of recombination is the breakage of physically linked regions across chromosomes. As a consequence, specific polymorphic sites in which four-gamete sets are present are more likely to happen in populations where there is sexual recombination, hence there is an expected surplus of the number of violations of the four-gamete test. Conversely, clonal reproducing populations will have in contrast a lower number of violations detected by the four-gamete test, being the captured ones, most likely, a consequence of recurrent mutations or sequencing errors.

In addition, the combined analysis of nucleotide diversity and SFS-based analyses like Tajima's *D* can tell about both demographic and selection events. For instance, assuming a constant mutation rate between populations, reduced nucleotide diversity in companion with a prevalence of acquisition of rare alleles suggested by negative Tajima's *D* values, are typical signatures of populations experiencing a growth phase after either a drastic population bottleneck or the existence of a founder effect. Alternatively, negative Tajima's *D* observations can also be caused by selective sweeps. Conversely, positive Tajima's *D* is a consequence of intermediate frequency alleles in the populations and thus, its interpretations can be associated either with abrupt population contractions or the presence of a balancing selection, which maintains the diversity levels.

Introgression between populations

Until now, I have discussed the acquisition of diversity within populations mainly through the way that mutations and drift act at the population level. However, because of migration and/or absence of geographical barriers, populations can experience the introduction of new allele diversity in the form of gene flow or introgression from external populations, regardless of their overall consequence will have neutral, negative or positive fitness effects in the recipient population. In practice, introgression has been shown to be an important source of genetic variation and diversity, with great impact in environmental adaptation, e.g. toll-like receptors which enhanced modern humans defence against pathogenic bacteria, fungi and parasites and were acquired through introgression from Denisovan and Neardenthal-like populations (Dannemann et al., 2016); colour pattern loci adaptive introgression in *Heliconius* butterflies between recently diverged species (Heliconius Genome Consortium, 2012; Pardo-Diaz et al., 2012; W. Zhang et al., 2016); and the contribution of domestication traits via introgression from wild relatives into sunflower modern cultivars (Baute et al., 2015).

Detection of gene flow or introgression can be however confounded with any other signal that creates discordance between the species and gene phylogenies, being Incomplete Lineage Sorting (ILS) a big source of discordance (Degnan & Rosenberg, 2009; Maddison, 1997). Indeed, ILS is the phenomenon in which the lineages from a given population do not coalesce among them prior to their

coalesce into an ancestral population (Avise et al., 1983). In other words, discordances due to ILS are a consequence of the lack of time for the lineages to sort all their alleles as a population. For this reason, when divergence times between lineages/species are short, a significant proportion of alleles or genes are expected to be discordant, obscuring the detection of introgression events. However, ILS and introgression differ both in the arrangement and length of their genomic footprints. ILS is characterized by randomly distributed events along the genome (Pease & Hahn, 2015), while introgression is manifested by longer genomic stretches.

Following this rationale, Green et al (Green et al., 2010), created a test that can differentiate whether the presence of a genome-wide discordant signal is likely explained by introgression rather than the manifestation of ILS, which is, instead, used as a null hypothesis. Given a four-taxon conformation (((X,Y),T),O), where *O* is a true outgroup and therefore defines the ancestrality (A), the test will inspect for specific site patterns where the lineage *T* has a derived allele (B) (Figure 4). In the absence of discordance events, the site pattern expectations will follow the lineage genealogy, so lineages *X* and *Y* will always have the same allele (Figure 4A,B). In contrast, in the presence of discordant events, lineages *X* and *Y* will have different site patterns, and the overall site pattern will not follow the genealogy (Figure 4C,D). Moreover, in the presence of ILS, the proportions of the discordant site patterns are expected to be equal, and therefore the statistic will not be significant (*Patterson's D* = 0). However, if the proportions of ABBA sites are higher (Figure 4C), there will be genomic-wide evidence of recombination between lineages *T* and *Y* (*Patterson's D* > 0). Conversely, if the proportions of BABA sites are higher (Figure 4D), there will be genomic-wide evidence of introgression between lineages *T* and *X* (*Patterson's D* < 0).

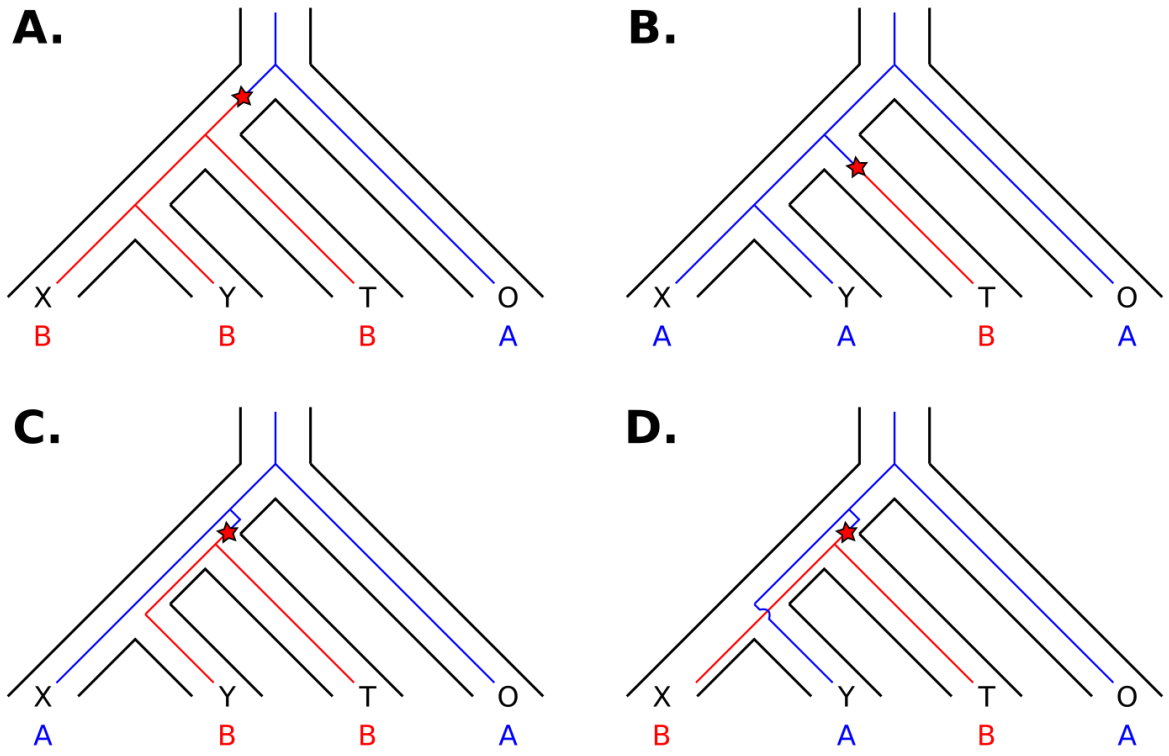


Figure 4. The rationale for Patterson's *D* statistic calculation. The true genealogy of a four-taxon group with an (O)utgroup is depicted by the black trees (((X,Y) ,T),O). Allele trees and site patterns are depicted as coloured lines where blue corresponds to the ancestral state (A) and red corresponds to the derivative state (B). Mutation occurrence is represented by a start. There are two possible concordant site patterns (**A.** and **B.**). In the presence of Incomplete Lineage Sorting (ILS), site patterns with the configuration ABBA (**C.**) and BABA (**D.**) are expected to occur with the same proportion. In a scenario of introgression, ABBA sites (**C.**) are more likely to happen if there is gene flow between T and Y, while BABA sites (**D.**) are expected to happen if there is gene flow between T and X.

1.4. Objectives of this work

As highlighted in this introductory section, I will make use of ancient DNA as a powerful source of information to inform and complement phylogenomics and population genomics approaches. By studying two different eukaryotic biological systems, a fungus and insect, I will describe the population history, evolutionary relationships and evolutionary forces that have shaped present historical and present-day diversity.

The first objective will be to describe and resolve the phylogenetic relationships among scarab beetles from the Mascarene Islands, some of which are most

likely extinct. To address this goal, I will combine whole genomic data from museum pinned-specimens together with high quality reads and assemblies from two extant beetle species. I will ascertain the evolutionary relationships by building a molecular phylogeny and by using site patterns based on *D*-statistics. Furthemore, I will estimate divergence times and analyze them in the light of the Mascarene Plateau geological events. Finally, I will use morphological standard observations to jointly analyze the estimated molecular-based phylogeny. For the first time I will propose a robust phylogenetic relationship of this important insect taxa by using whole genome markers instead of the until now used gene markers and morphological data.

The second objective will be to describe the present genetic history of the rice blast fungus, *Magnaporthe oryzae*. To address this goal, I will combine different datasets: I will use the two recent and largest genomic datasets publicly available, which have a worldwide geographical coverage. I also will generate new genomic data from present-day isolates coming from Europe, in order to fill this underrepresented region in the mentioned studies. And finally, I will make use of historic herbarium infected rice leaves which have been collected from 1891 to 1948 and therefore enhance the dataset with a temporal resolution of more than one century. I will define the current population structure of the pathogen by utilizing different methods that exploit different genomic features like allele sharing patterns, as well as genetic distances. Subsequently, I will describe the characteristics associated with the previously identified genetic groups. At this stage, I will use the information to build the phylogenetic relations and reconstruct the demographic history by using again different yet complementary phylogenetic approaches. Finally, by inspecting the repertoire of effectors as the most important proteins involved in the disease recognition by the host, I will jointly analyze the patterns of presence and absence in the light of the previously described genetic structure. By accomplishing the described objectives, this work will hopefully shed light and new hypotheses about candidate functional units involved in the adaptive process of the pathogen to different hosts, geographical regions and timescales.

2. Museum phylogenomics of extinct *Oryctes* beetles from the Mascarene Islands

Contributions

This chapter consists in the manuscript entitled: "Museum phylogenomics of extinct *Oryctes* beetles reveal independent colonization of a young volcanic island" available as a *bioRxiv* preprint (doi: 10.1101/2020.02.19.954339). The following people have contributed to the work presented in this chapter: Hernán A. Burbano (HAB), Ralf J. Sommer (RJS) and myself conceived and designed the project with input from Matthias Herrmann (MH) and Christian Rödelsperger (CR). I carried out DNA extractions and library preparations of historical museum specimens. Andreea Dréau (AD) and Waltraud Röseler performed linked-read library preparations. AD and CR carried out *de novo* assembly of extant genomes. MH and M.J. Paulsen performed morphological analyses and contributed entomological expertise. I performed the phylogenomic and population history analyses, developed the software application, analyzed the data and prepared all the figures with input from HAB. RJS, HAB and myself wrote the manuscript with input from all authors.

Abstract

The sheer size of insect museum collections and the possibility to sequence genomes from historical insect specimens present an unique opportunity to address phylogenetic and population history questions in light of increasing biodiversity loss. On islands, for instance, extreme morphological diversity, high endemism and high extinction rates make evolutionary inference difficult. Although it is evident that historical specimens could fill in for recently extinct taxa, they have not yet been widely harnessed for phylogenomic reconstructions. Here, we generated genomes from historical museum specimens from the Mascarene Islands to reconstruct the evolutionary relationships of three extinct species of the rhinoceros beetle genus *Oryctes*. We compared their genomes with those of an extant *Oryctes* species from the island of Réunion, as well as a flightless Réunion-based species previously placed into the supposedly unrelated genus *Marronus*. We found that contrary to the

previously proposed phylogeny based on morphological characters, *Marronus borbonicus* belongs instead to the genus *Oryctes*. Moreover, the two Réunion-based species (*O. borbonicus* and *M. borbonicus*) are not sister taxa, suggesting two independent colonizations. The divergence time between them (<3Myr) overlaps with the volcanic formation of Réunion, likely indicating that *M. borbonicus* became flightless in situ. Our study showcases the power of combining present-day and historical genomes to inform phylogenetic inference and paves the way for the inclusion of insect historical genomes in short-term longitudinal studies.

2.1. Introduction

Insects are the most diverse taxon on Earth and, thus, provide essential services to ecosystems. These services are currently under threat due to anthropogenic activities that have led to insect population declines and extinctions (Cardoso et al., 2020). In spite of their relevance, and given their diversity, insects are to date underrepresented in global change studies (Kharouba et al., 2018). To quantify the scale of biodiversity loss, ascertain its causes and formulate conservation strategies, it is fundamental to have a short-term record of diversity change spanning the last century. Fortunately, the sheer size of insect historical collections curated in natural history museums provide a window into the past to study, for instance, geographical ranges as well as phenotypic and phenological changes of both extant and extinct insect species (Kharouba et al., 2018). Due to recent developments in the retrieval and sequencing of ancient DNA (aDNA) (B. Shapiro & Hofreiter, 2014), it is now possible to sequence complete genomes of historical and ancient specimens. However, these approaches, with very few exceptions (Cridland et al., 2018; Mikheyev et al., 2015), are not yet widely spread in insect evolutionary genomics. Here, we applied state-of-the-art aDNA techniques to address phylogenomic and population history questions harnessing whole-genomes retrieved from historical specimens of extinct beetle species.

Coleoptera (beetles) are the most diverse order in metazoans with almost 400,000 described species (Hammond, 1992). Some lineages have given rise to spectacular forms that have fascinated humans for millennia. For example, illustrations featuring rhinoceros beetles have been found in Crete from the

Minoan period (2000-1600 BC) (Paulian, 1985). Large and comprehensive museum collections of insect specimens exist throughout the world and the overall beetle phylogeny has been studied in great detail using both morphological and molecular tools based on a handful of loci (Hunt et al., 2007; Jin et al., 2016). The rhinoceros beetle genus *Oryctes* includes some of the largest beetles, such as *O. gigas* (Dechambre & Lachaume, 2001), a species well known for its impressive horns (Hu et al., 2019). In total, Oryctes contains 42 valid species distributed in Africa, Southeast Asia and the Indian Ocean (Dechambre & Lachaume, 2001). The genus displays extremely high levels of endemism on islands in the Indian Ocean, which are considered major terrestrial biodiversity hotspots (N. Myers et al., 2000). For instance, Madagascar and the Mascarene Islands (Réunion, Mauritius and Rodrigues) alone harbor 16 species, some of which have gone extinct due to extreme habitat loss, i.e. *O. tarandus* and *O. chevrolatii* on Mauritius and *O. minor* on Rodrigues (Figure 2.1A). While several systematic classifications have been suggested based on morphological characters, no molecular phylogeny is available because traditionally used PCR-based methods are not suitable for highly-degraded DNA present in historical museum specimens, and state-of-the-art library-based methods have not been employed for phylogenomic studies in extinct insect species. Given the distribution and diversity of rhinoceros beetles across the Mascarene Islands, establishing their evolutionary relationships is fundamental to understanding how geological processes, such as landmass emergence and submergence, have shaped the Mascarene's endemic biodiversity.

Réunion, the youngest of the Mascarene islands, harbors *O. borbonicus,* and an additional rhinoceros beetle that was placed in the monotypic genus *Marronus* based on morphological analyses (Endrődi & Others, 1985). This genus has been included in the tribe Pentodontini, and presumed to be distantly related to *Oryctes* in the tribe Oryctini. Like many island beetles, *M. borbonicus* (hereafter referred to as *Marronus* to avoid confusion with *Oryctes borbonicus*) is a flightless species that has undergone dwarfism. The creation of the monotypic genus *Marronus* based only on morphology is problematic since morphological features are often lost, and frequently convergent, in dwarf species. Moreover, the high prevalence of character displacement in islands can obscure morphological synapomorphies. Both *O. borbonicus* and *Marronus* are hosts of the nematode *Pristioncus pacificus*, a well-studied model organism for integrative

evolutionary biology (Sommer & McGaughran, 2013). Thus, understanding the phylogenetic relationship between these two sympatric species might also shed light on the evolutionary history of the association between Réunion-based beetles and *P. pacificus*.

To investigate the phylogenetic relationships among *Oryctes* species from the Mascarene Islands and between them and *Marronus*, we used Illumina sequencing and 10X Genomics to refine the draft genome of *O. borbonicus*, and generated a new draft genome of *Marronus*. Furthermore, we used minute amounts of tissue from pinned beetle museum specimens to generate for the first time genome-wide data from extinct insect species from Mauritius and Rodrigues. The combined analysis of extant and extinct genomes permitted us to infer phylogenetic relationships, divergence times, and the colonization history of *Oryctes* beetles of the Mascarene Islands.

2.2. Results

2.2.1. *De novo* assembly of Reunion's *Oryctes borbonicus* and *Marronus borbonicus*

Despite the large number of beetle species (~400,000), genome sequences of only 20 species have been reported (D. D. McKenna, 2018; D. D. McKenna et al., 2019). Available genomic data includes a draft assembly of *O. borbonicus* (J. M. Meyer et al., 2016). Here, we sequenced DNA from two specimens of *O. borbonicus* and *Marronus*, the two extant endemic beetles from Réunion, on the 10X Genomics platform to improve the draft genome of *O. borbonicus*, and to generate a draft genome of *Marronus*. Both libraries were individually sequenced on the Illumina HiSeq 3000 platform yielding 330 million paired end reads (2x150bp), which translates into roughly 120X coverage per genome. These data were assembled into 411 Mb for *O. borbonicus* and 413 Mb for *Marronus* (Supplementary Table 2.1). In comparison to the previously published draft genome of *O. borbonicus* (J. M. Meyer et al., 2016), this led to a huge reduction in contig number from over 150,000 to 9,526 and an 80-fold increased contiguity, i.e. N50 was raised from 105kb to 8.4Mb (Supplementary Table 2.1). Coverage analysis of the previous *O. borbonicus* assembly, revealed a high abundance of genomic regions with half the expected coverage possibly

pointing towards a problem of allelism, i.e. two divergent haplotypes were assembled separately (Barrière et al., 2009; J. M. Meyer et al., 2016). Indeed, sequencing of linked reads allowed the new *O. borbonicus* assembly to be resolved into pseudohaplotypes (Weisenfeld et al., 2017). Consequently, the coverage profile of the new assembly is substantially shifted towards higher coverage, most likely autosomal regions (Supplementary Figure 2.1) indicating that the increased size of the previous assembly was largely due to separate assembly of divergent haplotypes. In total, 17,736 and 14,738 protein coding genes were annotated in the assemblies of *O. borbonicus* and *Marronus*, respectively. Comparison of protein sequence identity between 9,656 orthologous pairs revealed a median percentage identity of 98%, indicating that *O. borbonicus* and *Marronus* are closely related. Indeed, further phylogenomic analysis in the context of 13 phylogenetically broad Coleopteran genomes indicated that *O. borbonicus* and *Marronus* are closely related, similar in phylogenetic distance to that between *Hycleus phaleratus* and *Hycleus cichorii*, two members of the same genus (Supplementary Figure 2.2).

2.2.2. Sampling and sequencing of historical beetle genomes

As three other endemic *Oryctes* species from both Mauritius (*O. tarandus* and *O. chevrolatii*) and Rodrigues (*O. minor*) are extinct, the only way to robustly reconstruct their evolutionary relationships is by retrieving genomes from historical museum specimens. To this purpose we obtained pinned museum specimens from these three extinct species, and also from *O. mayottensis* from Mayotte Island (as an outgroup for the Mascarene species), and an *O. borbonicus* museum specimen for comparison (Figure 2.1 and Supplementary Table 2.2). The age range of the museum specimens was between 53 and 99 years old. Although museum specimens from insects have been used before for phylogenetic analyses, this was mainly by PCR-based methods that aimed at amplifying one or a handful of loci (Goldstein & Desalle, 2003; Thomsen et al., 2009). In general, PCR amplifications have failed when using samples older than 50 years (Watts et al., 2007). In contrast, whole-genome data derived from beetle museum specimens have not been used for phylogenetic analyses. To minimize the degree of sample destruction, we extracted DNA from one leg from each museum specimen amounting to ~7-23 mg of tissue (Supplementary Figure 2.3A-E). The DNA extractions (Gutaker et al., 2017) and library

preparations were carried out in a clean room facility to avoid contamination from exogenous DNA. Shallow sequencing of the libraries showed damage patterns and length distributions typical of ancient DNA (aDNA) (Supplementary Figure 2.3F-G) (Briggs et al., 2007), and endogenous beetle DNA percentage varied from 5 to 85% (Supplementary Figure 2.3H). The variation in the percentage of endogenous DNA did not correlate with the distance of each species to the draft genomes of either *O. borbonicus* or *Marronus* (Supplementary Figure 2.3I). The same DNA extracts were used to generate chemically repaired libraries with significantly reduced ancient DNA-associated damage (Briggs et al., 2010) (Supplementary Figure 2.4), which were sequenced using the Illumina platform achieving on average 1X coverage (Supplementary Table 2.3). Only sequencing data derived from these chemical repaired libraries were used for subsequent analysis.

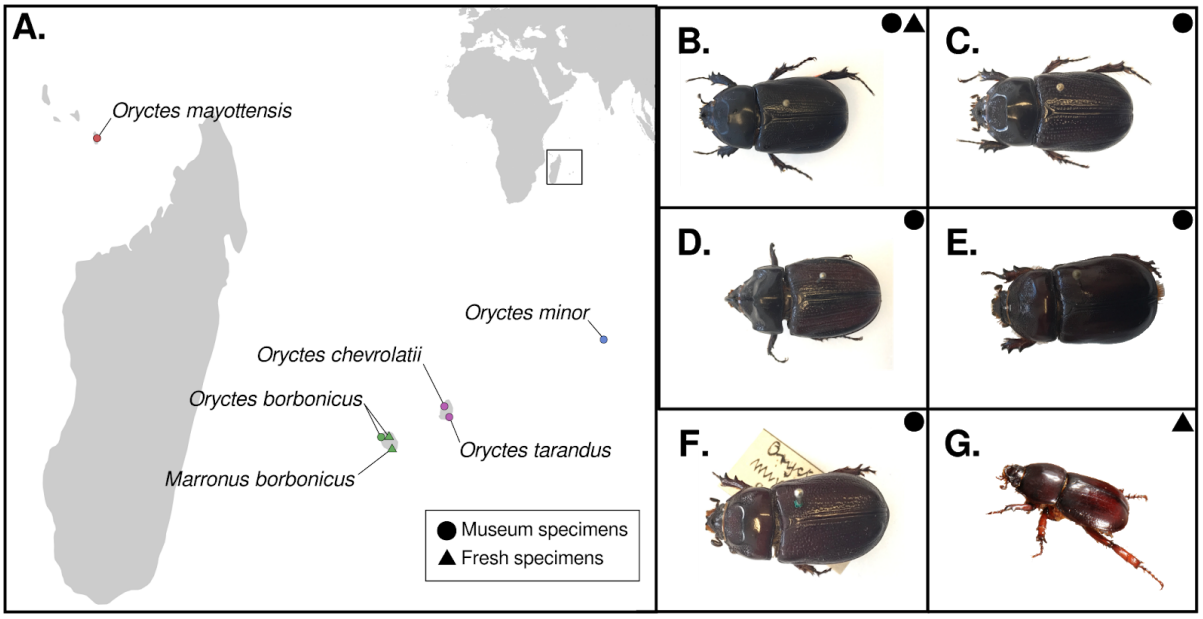


Figure 2.1. Provenance of fresh and museum rhinoceros beetle specimens. (A) The map displays beetle species endemic to different islands in the Indian Ocean on the coast of East Africa. The largest island is Madagascar. From west to east the sampled species included *Oryctes mayottensis* in Mayotte, *Oryctes borbonicus* and *Marronus borbonicus* in Réunion, *Oryctes chevrolatii* and *Oryctes tarandus* in Mauritius, and *Oryctes minor* in Rodrigues. Réunion, Mauritius and Rodrigues comprise the Mascarene Archipelago. *O. chevrolatii*, *O. tarandus* and *O. minor* have not been observed in decades and are presumably extinct. **(B-F)** Beetle museum specimens used in this study. **(B)** *O. borbonicus* **(C)** *O. tarandus* **(D)** *O. chevrolatii* **(E)** *O. minor* **(F)** *O. mayottensis* **(G)** Fresh specimen of *M. borbonicus*.

2.2.3 Evolutionary relationships and divergence times of *Oryctes* spp. and *Marronus borbonicus*

To establish the evolutionary relationship among all beetle species, we first mapped the sequenced reads to the updated *O. borbonicus* draft genome, which provides a common coordinate system for all beetle species. To ascertain single nucleotide polymorphisms (SNPs) for each sample, we randomly selected one base at each segregating site. This method, also known as "pseudo-haploidization", is commonly applied to low-coverage aDNA datasets, allowing the estimation of genetic relatedness from low-coverage data (Green et al., 2010) (See Materials and Methods). Implementing this methodology, we identified a total of 1,541,675 SNPs.

Initially, we summarized the genetic variation among beetle species using principal components analysis (PCA), which showed that PC1 does not separate species by genera but, instead, clusters together *Oryctes* and *Marronus* beetles from Réunion and Mauritius, and separates the species from Mayotte and Rodrigues (*O. mayottensis* and *O. minor*) from each other, and from the species from Réunion and Mauritius (Supplementary Figure 2.5A). This separation suggests that the vast majority of SNPs might have occurred in the lineages leading to *O. mayottensis* and *O. minor*. To test this hypothesis, we implemented a Minor Allele Frequency (MAF) filter that requires the SNPs to be segregating in at least two out of the seven samples (2/7). This filtering step reduced both the number of SNPs to 304,417, and the percentage of variance explained by PC2 in 30%. Indirectly, this reduction along the PC2 axis increased the separation among the beetles from Réunion and Mauritius (Figure 2.2A and Supplementary Figure 2.5). These observations support the hypothesis that the variance explained by PC2 was driven by variation that is private to either *O. mayottensis* or *O. minor* (Supplementary Figure 2.5A). Thus, after implementing the MAF filtering, the PC2 further separated beetle species. Importantly, PC2 also grouped together *O. borbonicus* from fresh and museum samples, which demonstrates that the use of chemically-repaired libraries and the appropriate identification of SNPs permit the accurate clustering of species, independent of

their present-day or historical origin (Figure 2.2A and Supplementary Figure 2.6).

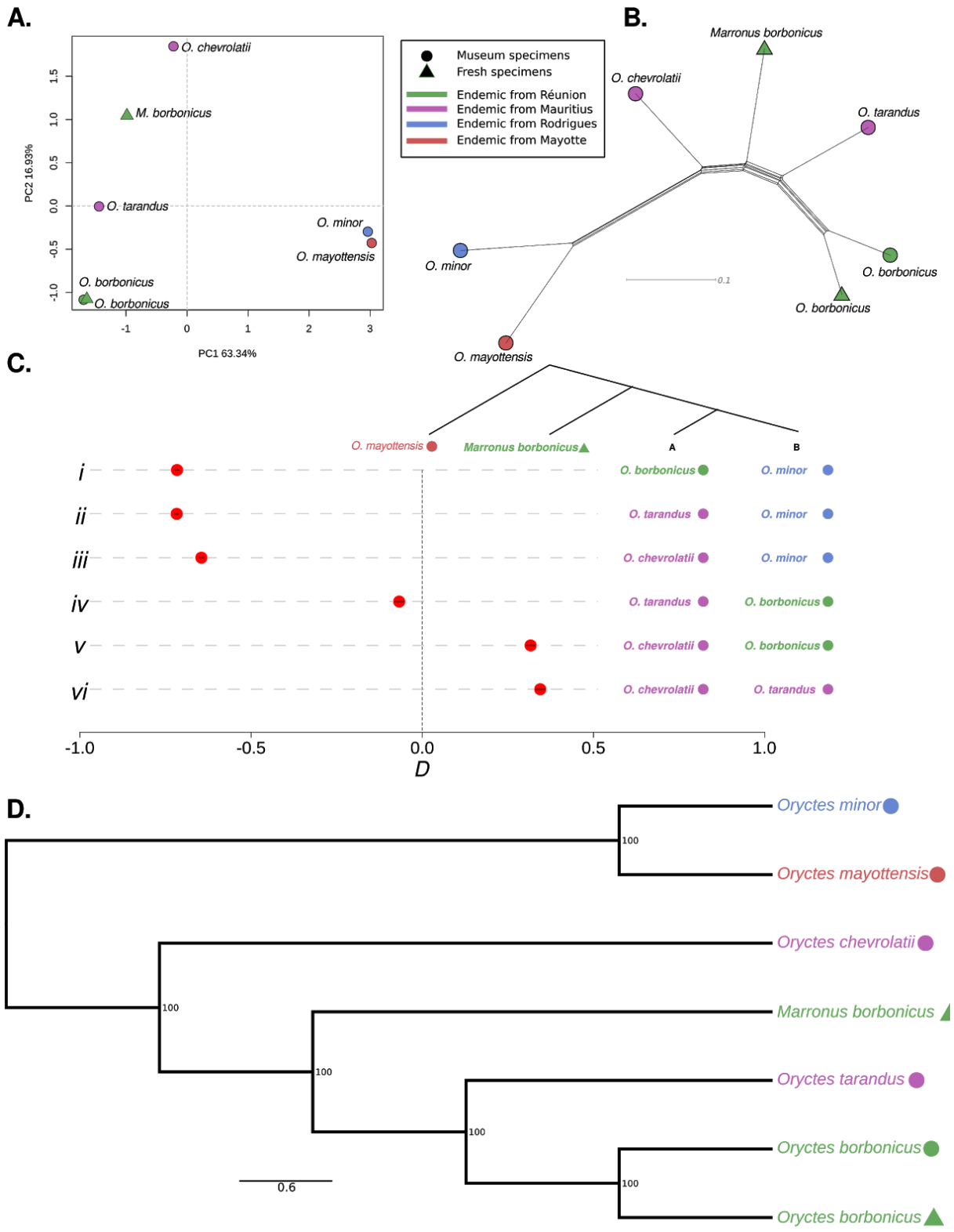


Figure 2.2. Evolutionary relations among rhinoceros beetles. (A) Principal component analysis plot based on 304,417 SNPs. Genetic distances between beetle samples are projected onto the first two PCs. Axis labels indicate the fraction of total variation explained by each PC. **(B)** Phylogenetic network based on 304,417 SNPs using the neighbor-net method. **(C)** Testing the robustness of phylogenetic relations among scarab beetle species using *D*-statistics of the type *D*(B,A; *Marronus borbonicus*, outgroup), as depicted in the phylogenetic tree. *O. mayottensis* was used as an outgroup. Each row (*i-vi*) shows a different *D*-statistic configuration. A negative *D*-statistic indicates that *M. borbonicus* is closer to species A, whereas a positive *D*-statistic indicates that *M. borbonicus* is closer to species B. The points depict the result of each *D*-statistic test and the lines their respective 95% confidence intervals. Rows *i-iii* show that *M. borbonicus* is closer to the *Oryctes* spp. from Réunion and Mauritius. Rows *v-vi* show that *M. borbonicus* is closer to both *O. borbonicus* and *O. tarandus* than to *O. chevrolatii*. Finally, row *iv* shows the closest *D*-statistic to zero, which indicates that *M. borbonicus* is slightly closer to *O. tarandus* than to *O. borbonicus* and **(D)** SVDquartets species tree. Numbers at nodes indicate bootstrap support (1000 replicates).

To further refine the evolutionary relationships among these Mascarene beetles, we built phylogenetic networks using either of the two SNP sets, i.e. with and without MAF filtering. As expected from the PCA analysis, the use of the MAF filtering reduced the branch lengths of *O. mayottensis* and *O. minor*, while preserving the network topology (Figure 2.2B and Supplementary Figure 2.6A-B). To focus on the evolutionary relationships among Réunion and Mauritius beetles, from here on, we carried out all analyses implementing the MAF 2/7 filtering. The phylogenetic networks revealed that (i) sympatric species pairs from Réunion and Mauritius do not cluster together but, instead, *O. borbonicus* and *O. tarandus* appear as sister groups, (ii) *Marronus* falls within the *Oryctes* genus, and (iii) *O. mayottensis* and *O. minor*, as suggested by the PCA analysis, are outgroups for beetles from Réunion and Mauritius. The absence of pervasive reticulation in the phylogenetic network might suggest that introgression is not substantial between Mascarene beetles.

In order to test the phylogenetic relationships (the "treeness") suggested by the phylogenetic network, we used *D*-statistics (Durand et al., 2011; Green et al., 2010). We employed *D*-statistics of the following form: *D*(Outgroup, *Marronus*; species A, species B), using *O. mayottensis* as an outgroup and different configurations of the four *Oryctes* species from Réunion and Mauritius as species A and B (Figure 2.2C). The first three rows (*i-iii*) of Figure 2.2C indicated that *Marronus* is closer to *Oryctes* species from Reunión and Mauritius than to *O.*

minor from Rodrigues, whereas the last two rows (*v-vi*) showed that *Marronus* is closer to both *O. borbonicus* and *O. tarandus* than to *O. chevrolatii*. The extreme negative and positive *D*-statistics of rows *i-ii* and *v-vi* indicated that the tested phylogenetic hypotheses are likely incorrect. Finally, row *iv*, where the *D*-statistic is the closest to zero, showed that *Marronus* is slightly closer to *O. tarandus* than to *O. borbonicus*, which could be explained by post-speciation introgression between *Marronus* and *O. tarandus*. To evaluate the influence of genome reference bias in our phylogenomic inferences, we repeated the *D*-statistic analyses but instead of mapping the reads to the *O. borbonicus* draft genome, we mapped the reads to the *Marronus* genome. *D*-statistics values for configurations presented in all rows but row *iv* showed qualitatively similar results, very negative for rows *i-iii* and very positive for rows *v-vi* (Figure 2.2C and Supplementary Figure 2.8C). Also consistent with previous analysis, row *iv* had the *D*-statistics closest to zero but this time with a positive value, which indicates a closer relationship between *Marronus* and *O. borbonicus* that again could be caused by post-speciation introgression, this time between *Marronus* and *O. borbonicus*. The fact that the sign of the *D*-statistic in row *iv* switched between negative and positive depending on the reference genome used very likely indicates that *Marronus* is equally distant to both *O. borbonicus* and *O. tarandus* and that the true value would overlap zero. A *D*-statistic not different from zero likely suggests that negligible or no post-speciation introgression took place between *Marronus* and either *O. borbonicus* and *O. tarandus*, thus any segment of the genome showing a discordant phylogeny is most likely the result of incomplete lineage sorting (Green et al., 2010). The lack of pervasive introgression suggested by both the lack of reticulations in the phylogenetic network and the *D*-statistics prompted us to carry out a phylogenetic analysis using genome-wide SNPs. First, we generated a species tree using a quartet-based method that analyzes sequencing data on a SNP-by-SNP basis (Chifman & Kubatko, 2014) (Figure 2.2D). Second, we carried out phylogenetic inferences using concatenated SNPs under both Maximum Likelihood and Bayesian frameworks (Supplementary Figure 2.7). Both the SNP-by-SNP method and the concatenation approach confirmed that *Marronus* indeed belongs to the genus *Oryctes* and, thus, do not warrant a monotypic genus. Additionally, the fact that *O. borbonicus* and *O. tarandus*, although endemic to different islands, are sister groups, and that *Marronus* is an outgroup to both of

them, likely indicates that Réunion was colonized independently by *O. borbonicus* and *Marronus*. All our results (Figure 2.2) proved to be robust to the choice of reference genome (Supplementary Figure 2.8).

To relate the speciation events with the geological processes that gave rise to the Mascarene Islands, we set out to calculate the divergence times between Mascarene's beetles. Initially, to investigate the variation of DNA sequence divergence along the genome, we analyzed the nucleotide divergence of all species relative to the *O. borbonicus* draft genome using 100 kb windows (Figure 2.3A). As suggested by their phylogenetic relationship as outgroups for Réunion and Mauritius beetles, *O. minor* and *O. mayottensis* displayed the highest nucleotide divergence relative *O. borbonicus*, with medians of 3 and 3.8%, respectively. The use of pairwise genetic differences reflected the true genetic distance between *O. borbonicus* and both *O. mayottensis* and *O. minor* as shown in the non-MAF filtered genetic network (Supplementary Figure 2.6A). The distributions of nucleotide divergences between *O. borbonicus* and Réunion (*Marronus*) and Mauritius (*O. tarandus* and *O. chevrolatii*) beetles showed a high degree of overlap (Figure 2.3A), reflecting very close divergence times, as expected given their close phylogenetic relationships. Although, we found that mean pairwise nucleotide divergence does not perfectly correlate with the inferred phylogenetic relationships between beetle species, we expect that mean sequence divergence will provide a rough estimate of sequence divergence times. To translate mean sequence divergence between all sequenced beetle species and *O. borbonicus*, we used the most commonly used insect mutation rate: 1.15% (1.1%-1.2%) per million years (Brower, 1994). Although this rate was calculated using mitochondrial DNA (mtDNA) from a different insect order (Lepidoptera) (Brower, 1994), a recent attempt to calibrate the insect molecular clock using beetle mtDNA estimated a rate very similar to the Lepidoptera-based one (Papadopoulou et al., 2010). Moreover, it has been suggested that mtDNA and nuclear mutation rates are very similar in insects (Papadopoulou et al., 2010; Sharp & Li, 1989).

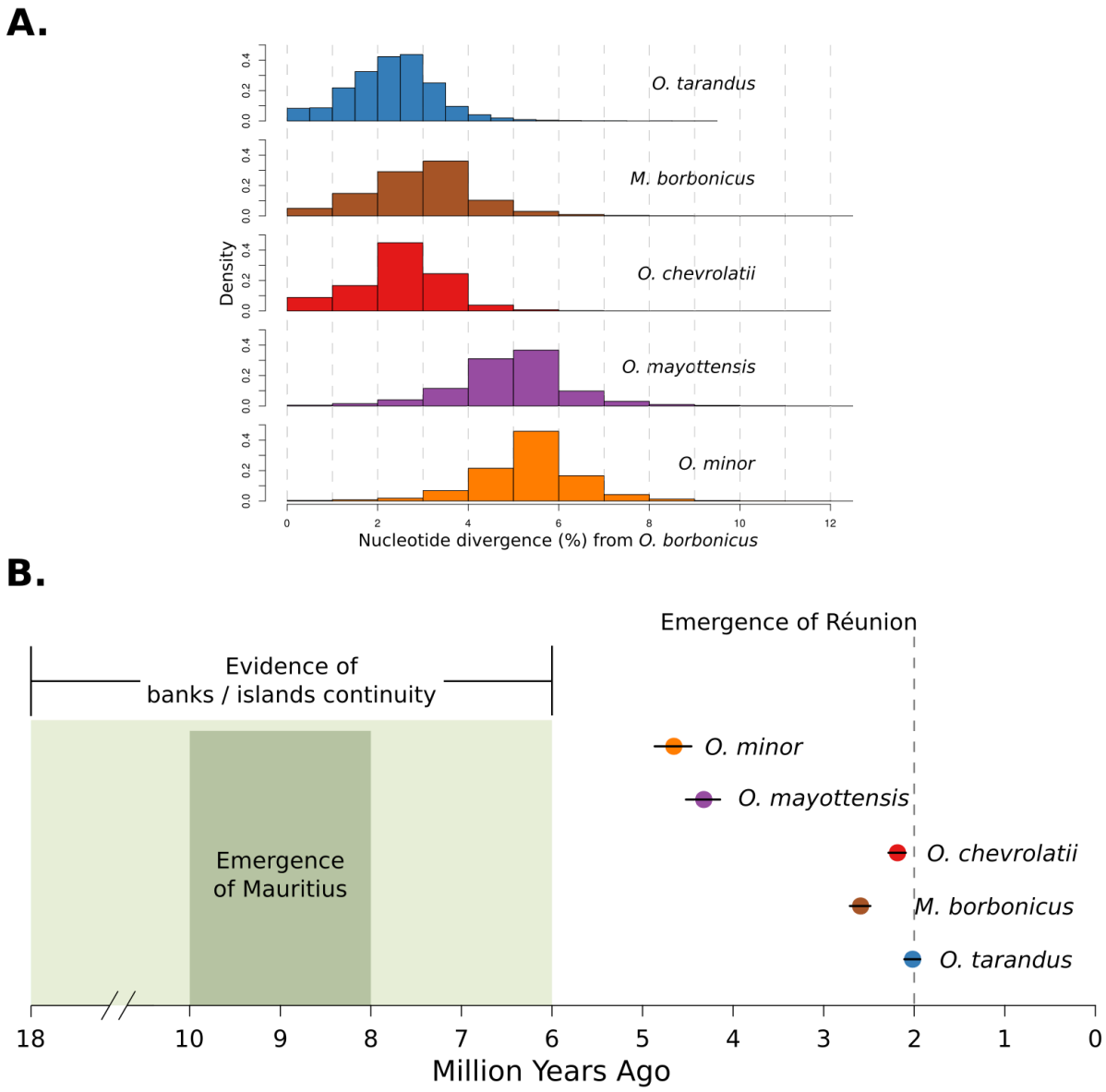


Figure 2.3. Nucleotide divergence of rhinoceros beetles. (A) Distribution of nucleotide divergence from the *Oryctes borbonicus* genome among genomic segments of 100 kb (N=1,553). **(B)** Divergence times of rhinoceros beetles relative to the *O. borbonicus* genome. The distributions of nucleotide divergences from (A) were converted into divergence times in million years using a constant substitution rate of 1.15 (1.1 - 1.2) % pairwise sequence divergence per million years per lineage (Brower, 1994). The horizontal bars indicate upper and lower bounds for the divergence times based on the confidence intervals of the substitution rate. The x-axis shows the timing of major geological events in the Mascarene plateau, including the emergence of present-day Mascarene Islands.

The translation of mean sequence divergence resulted in sequence divergence times (in million years) of 1.70 (1.63-1.68) for *O. chevrolatii*, 1.90 (1.82-1.99) for *O.*

tarandus and 1.79 (1.77-1.87) for *Marronus* (Figure 2.3B). Thus, all sequence divergence times between Réunion and Mauritius beetles relative to *O. borbonicus* are younger than 2 million years, and close in time to each other, showing overlapping confidence intervals in the case of *Marronus* and *O. tarandus*. Note that our approach to calculate sequence divergence times overestimates population divergence times, since sequence divergence times include the coalescence within the common ancestor of two species, while the population divergence - the point at which species stop exchanging genes - takes place much later in time. However, our estimated divergence times provide approximate estimates that can be overlaid with the geology of the Mascarene Islands. This overlay suggests that *Marronus* diverged from the common ancestor of *O. borbonicus* and *O. tarandus* at a time posterior to the emergence of Réunion (Figure 2.3B). These results imply that *Marronus* became flightless *in situ*, since it is less likely for a flightless species to be able to colonize an island, as has been proposed for the flightless Rodrigues solitaire (Beth Shapiro et al., 2002).

2.2.4. Morphological analysis of *Marronus borbonicus* and its relation to *Oryctes* spp.

In light of the phylogenomic analyses presented here that locate *Marronus* within the genus *Oryctes*, we revisited the morphological evidence that placed *M. borbonicus* in its own monotypic genus. *Marronus* has previously been classified in the tribe Pentodontini, separated from the Oryctini by the form of the apex of the metatibiae (truncate vs. toothed, respectively) and by the more pronounced sexual dimorphism of oryctines (Endrödi, 1969). However, these characters doubtlessly have evolved numerous times independently and are not taxonomically robust, leaving the monophyly of the tribes in doubt (Ratcliffe et al., 2013). Truncate metatibiae are a frequent accompaniment to the suite of characters that are related to flightlessness, e.g. dwarfism, atrophied wings, reduced eyes, and thickened legs. Other island flightless rhinoceros beetles purported to be genera of Pentodontini such as *Neoryctes* Arrow (Galapagos Islands) and *Mellissius* Wollaston (St. Helena) display truncate metatibiae, as do numerous other flightless scarabaeoid beetles (Scholtz, 2000). Most importantly, *Marronus* displays the same sexually dimorphic characters found in *Oryctes* species, just in a reduced form (Figure 2.4). These consist of a frontal

horn and distinct anterior pronotal fovea in males and a frontal tubercle and smaller pronotal fovea in females. The pronotal fovea of male *Oryctes* is often bordered behind by a bifurcate process and laterally by coarse punctures, and these structures can also be found in males of *Marronus* in reduced form. Thus, the morphology of *Marronus* is consistent with that of *Oryctes* species, agreeing with our phylogenomic analyses, and therefore, *Marronus* should be transferred to *Oryctes*. As two species of the same genus cannot retain identical names, a replacement name will be designated elsewhere.

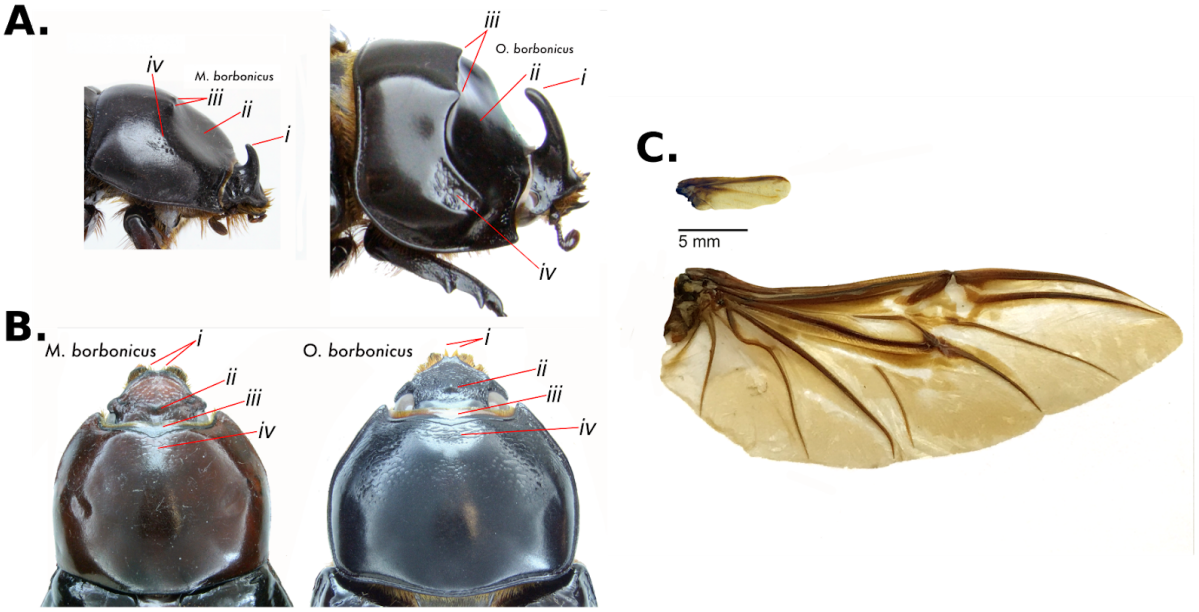


Figure 2.4. Morphological analysis of *Oryctes borbonicus* and *Marronus borbonicus*. (A) Similarities in male specimens of *M. borbonicus* (left) and *O. borbonicus* (right). Roman numbers and red lines point to the presence of analogous morphological characters: *i)* horn, *ii)* pronotal fovea, *iii)* bifurcate process, and *iv)* coarse punctures. **(B)** Similarities in female specimens of *M. borbonicus* (left) and *O. borbonicus* (right). Roman numbers and red lines point the presence of analogous morphological characters: *i*) bidentate clypeus, *ii)* central tubercle, *iii)* anterior margin of pronotum thickened medially, and *iv*) flattened area anteromedially with transverse rugae. **(C)** Wings of *M. borbonicus* (upper) and *O. borbonicus* (lower). While the *O. borbonicus* wing shows venation, size and shape typical of a functional beetle wing, *M. borbonicus* wing size and venation are highly reduced.

2.3. Discussion

Current studies of biodiversity are challenged by rapid environmental change driven by factors such as increased pollution, widespread habitat loss and fragmentation, and unprecedented climate change, which has resulted in increasing extinction rates (De Vos et al., 2015). Island ecosystems are particularly vulnerable to these threats due to their isolation and ecological uniqueness, which is characterized by high levels of endemism. It is estimated that 80% of species extinctions have taken place in islands and 45% of Red List endangered species also inhabit islands. Thus, while characterizing biodiversity and the evolutionary relationships among different groups of organisms is today more relevant than ever, species are vanishing from Earth at a higher rate than our capacity to study them. Hence, it is becoming necessary to take a retrospective approach to the study of biodiversity that resides in the global archive provided by natural history collections (Kharouba et al., 2018).

Here, we took such a retrospective approach to study the phylogenetic relationships among extant and extinct beetles from the Mascarene Islands. In contrast to most aDNA studies, which usually investigate species that went extinct thousands (e.g. mammoths (Pečnerová et al., 2017)) or hundreds of years ago (e.g. moas (Allentoft et al., 2014)), we studied species that went extinct only a few decades ago. Comparisons between recent extinctions and closely-related extant relatives will shed light on the genetic and ecological factors driving either extinction or persistence of species. To generate genome-wide datasets from historical pinned insect museum specimens, we used state-of-the-art ancient genomics methods including the use of preventive measures to avoid contamination by exogenous DNA, an extraction protocol that enriches for short DNA fragments (Gutaker et al., 2017), and library preparation methods tailored for aDNA (Briggs et al., 2010; M. Meyer & Kircher, 2010). Methodologically, we go beyond previous studies, since the use of these methods permits responsible analysis of precious historical samples with minimal tissue destruction and to present positive evidence of the historical nature of the sequenced DNA (Weiß et al., 2015).

Our phylogenomic analyses support two independent colonizations of rhinoceros beetles to Réunion island and suggest that *Marronus* became

flightless *in situ*. Importantly, estimated sequence divergence times are in agreement with the well accepted geological age of Réunion. Réunion is believed to have emerged about 3 mya and several biological reasons suggest that it has been colonized after Mauritius and Rodrigues, i.e. the absence of flightless birds only on Réunion. In contrast to Réunion, our estimated sequence divergence time for *O. minor* from Rodrigues disagrees with the generally accepted geological age of this island. Specifically, our data suggest that *O. minor* emerged around 5 mya, a finding that is consistent with data on other biota on the island (Cheke & Hume, 2010). Importantly, previous geological studies have recently been contradicted by discoveries of older relicts of lava that support an age of Rodrigues similar to that of Mauritius (Cheke & Hume, 2010). These recent findings are also more in agreement with our estimated divergence times. Thus, phylogenomics combining extant and extinct species can provide important biological support for the geology and colonization of islands with limited or contrasting geological data.

Our study showcases an integrative taxonomic approach that combines traditional morphological analyses with genome-wide variation from extant and extinct species. In light of current global environmental challenges, and given the vast number of plant and animal collections curated in natural history collections, the widespread use of this approach will be fundamental to catalogue Earth's biodiversity through space and at different timescales. Furthermore, we envision that the approach presented here will be also used in more applied fields such as the study of insect pests and insect invasive species.

2.4. Materials and Methods

2.4.1. Materials and laboratory methods

Biological material of fresh specimens

Samples from *Marronus* and *Oryctes borbonicus* were collected in Réunion.

DNA extraction of fresh specimens

In order to recover DNA from beetle specimens, two legs plus one thoracic muscle were manually extracted by tweezers from one male specimen of an *O. borbonicus* beetle and three legs plus one thoracic muscle for *Marronus*. The

beetle material was ground in liquid nitrogen and processed further according to the QIAGEN Genomic DNA Kit using the 100/G Genomic-tips. DNA was precipitated with 7µl of GlycoBlue (Invitrogen, 15mg/ml) and resuspended in 35µl of EB Buffer (Qiagen). DNA quality was checked by Nanodrop, Qubit and Pulse Field.

Library preparation and sequencing of fresh specimens

The preparation of the linked-read sequencing library was done as described previously (Dréau et al., 2019). Raw sequences were assembled into draft assemblies with the help of *SuperNova (Weisenfeld et al., 2018)* (v. 2.0.1). *SuperNova* was run on the full sequencing data set as well as multiple downsampled read sets. In terms of assembly contiguity, the best results were obtained using around 170 million single reads, which translates into roughly 60X coverage per genome. To guide gene annotation, we generated a transcriptome of a male *Marronus* individual by grinding several legs in liquid nitrogen and used the Zymo Direct-zol RNA Miniprep Kit according to the manufacturer's instructions to extract RNA, which was then eluted in 25µl dH_20. We followed previously described methods (Rödelsperger et al., 2018) to prepare an RNA-seq library and sequenced it on a multiplexed run on a Illumina HiSeq 3000 resulting in 24 million paired end reads (2 x 150bp).

Biological material of museum specimens

Museum sample provenances are provided in Supplementary Table 2.2.

DNA Extraction of museum specimens

To prevent contamination by exogenous DNA, museum samples were handled using standard ancient DNA precaution measures, i.e. sterilization with UV light of all equipment, surfaces and hoods after each extraction round, and the use of different hoods for handling of samples, reagents and DNA extracts, and of protective gear by researchers. DNA extractions were carried out in the clean-room facility at the Institute of Archeological Sciences at the University of Tübingen. The tissue (one leg per specimen) was ground inside a microtube with a stainless steel pestle until finely powdered and a N-phenacylthiazolium bromide (PTB) and Qiagen Plant DNEasy® Mini Kit (Qiagen)-based protocol

was used to isolate the DNA (Gutaker et al., 2017). A microtube without tissue was used as negative DNA extraction control.

Library preparation and sequencing of museum specimens

Genomic libraries for all museum specimens were prepared in a DNA clean-room facility taking the same preventive measures described in the DNA extraction section.

Non-UDG treated library preparation of museum specimens

We used a protocol that permits the preparation of indexed sequencing ancient DNA libraries (M. Meyer & Kircher, 2010). After adapter ligation in the clear-room facility, the indexing and PCR amplification of the libraries were performed in a different laboratory, located in a separate building. Briefly, the libraries were indexed using two barcoded primers (Kircher et al., 2012) during 10 cycles of PCR amplification using AccuPrime™ Pfx polymerase (Thermo Fisher Scientific). The MinElute PCR Purification Kit (Qiagen) was used to clean PCR residues and samples were pooled in equimolar concentrations. The samples were sequenced at the Genome Center facility located at the Max Planck Institute for Developmental Biology, with the Illumina MiSeq Platform using MiSeq Reagent Kit v2, 300 cycles (Illumina). Together with the prepared libraries, both aDNA extraction and a library preparation negative controls were sequenced. All non-UDG library sequences were used only for authentication purposes and not included in any further analyses.

UDG treated library preparation of museum specimens

In spite of the fact that cytosine to thymine (C-to-T) substitutions are useful for the authentication of the samples, they are not desirable in the phylogenomic inferences (Briggs et al., 2007; Hofreiter, Serre, et al., 2001). Thus, in order to reduce the effect of C-to-T substitutions, we prepared new DNAlibraries adding uracil-DNA glycosylase (USER™ enzyme (New England Biolabs)) during the blunting step (Briggs et al., 2010). The rest of the steps were done as described in the non-UDG treated library section. Finally, after measuring the final DNA concentration per sample and assessing the DNA endogenous concentration (Supplementary Figure 2.3H), we prepared a pool with a calculated equimolar content of endogenous molecules. Sequencing was done using the Illumina

HiSeq 3000 platform (Illumina) located at the Genome Center facility at the Max Planck Institute for Developmental Biology.

2.4.2. Bioinformatic analyses

Assembly of fresh specimens reads

Assembly quality was assessed using an approach based on benchmarking of universal single copy orthologs (BUSCO (Simão et al., 2015)). The raw genome assemblies as well as a data set of 1658 orthologous genes from 42 insect species (insecta_odb9) were taken as input of the program *run_BUSCO.py* (v. 3.0.1). To analyse the coverage profile of the previous and current *O. borbonicus* assemblies, we aligned raw reads to both assemblies with the help of the *mem* program of the *BWA* software suite (v. 0.7.17, default options) (Li & Durbin, 2009). Coverage profiles were calculated by the *samtools depth* program (v. 0.1.18, default options) (Li et al., 2009).

Gene annotation and comparative genomic analysis of fresh specimens

Transcriptomic data for *Marronus* and reads from a transcriptomic library of *O. borbonicus* (J. M. Meyer et al., 2016) were assembled by the software *Trinity* (v. 2.2.0, default settings) (Grabherr et al., 2011). Full or partial open reading frames were called as described previously (Rödelsperger et al., 2018). In cases where *Trinity* annotated multiple isoforms, the isoform with the longest ORF was chosen as a representative isoform for subsequent analysis. The resulting ORFs and protein sequences were mapped against their reference assemblies by the *protein2genome* program of the *exonerate* package (version 2.2.0, --bestn 1 --dnawordlen 20 --maxintron 20000) (Slater & Birney, 2005). Among all gene annotations which share an identical exon, the annotation resulting in the longest protein product was taken as representative annotation. Pairwise BLASTP (v. 2.6.0, e-value 0.00001) searches were done between proteins of *Marronus* and *O. borbonicus* and best reciprocal hits were extracted to estimate the median protein sequence identity between orthologous gene pairs. For further comparative genomic analyses, we obtained protein sequences for *Nicrophorus vespilloides* and *Aethina tumida* from the i5k website (https://i5k.nal.usda.gov/, accessed July, 12th 2019) (Poelchau et al., 2015), *Anoplophora glabripennis*, *Dendroctonus ponderosae*, *Tribolium castaneum* from

Ensembl Metazoa (release 44), *Protaetia brevitarsis*, *Pyrocoelia pectoralis*, *Hycleus cichorii* and *Hycleus phaleratus* from GigaDB (X. Fu et al., 2017; Wang et al., 2019; Wu et al., 2018), *Onthophagus taurus* from the U.S. Department of Agriculture website (https://data.nal.usda.gov, accessed July, 12th 2019), *Agrilus planipennis* and *Leptinotarsa decemlineata (Schoville et al., 2018)* from the ftp server of the Human Genome Sequencing Center of the Baylor College of Medicine (ftp://ftp.hgsc.bcm.tmc.edu/I5K-pilot/, accessed July, 12th 2019), and *Hypothenemus hampei* from the website of the NYU Center for Health Informatics and Bioinformatics (https://genome.med.nyu.edu/) (Vega et al., 2015). In cases of multiple isoforms per gene, the sequence corresponding to the longest protein was taken for further analysis. Assignment of protein sequences into orthologous clusters, concatenation, and phylogenetic reconstruction were performed as described previously (Rödelsperger et al., 2018).

Pre-processing of reads of museum specimens

Raw reads were trimmed and merged with *AdapterRemoval* (v. 2.2.2) (Schubert et al., 2016). Merged reads were aligned to the *O. borbonicus* and *M. borbonicus* assemblies presented here using *bwa mem* (v. 0.7.12) (Li & Durbin, 2009). Subsequently, PCR duplicates were removed with *picard tools* (v. 2.8.1).

Museum specimens authentication

Ancient DNA (aDNA) has multiple signatures which can be used to its authentication: i) C-to-T substitutions are expected to be enriched at the ends of reads with a decay towards the inner part of the molecule, ii) aDNA is expected to have shorter fragments than fresh DNA molecules and, iii) aDNA is expected to be a mixture between endogenous DNA and non-endogenous DNA (Dabney, Meyer, et al., 2013; Gutaker & Burbano, 2016). We used *MapDamage* (v. 2.0) (Jónsson et al., 2013) to analyse both the C-to-T misincorporation patterns and fragment lengths. We also calculated the endogenous DNA content by dividing the number of mapped reads by the total number of reads. All museum samples displayed the expected signatures of museum / historical samples (Supplementary Figure 2.3).

Identification of segregating positions

To achieve a comparable coverage between historical and fresh samples, four million raw reads from both fresh *O. borbonicus* and *Marronus* were subsampled with the program *samtools view* (v. 1.4) (Li et al., 2009). Together with the museum mapped reads, we created a single bam file with the program *samtools merge* (v. 1.4) (Li et al., 2009) and used it as the input for the discovery of variant positions using the program *bsh-denovo* (Weiß, 2019). Only positions with full information were considered (flag *-m = 1*) and the alternative allele was randomly sampled (flag *-a = 0.001*). To account for the effect of the Minimum Allele Frequency (MAF) in the evolutionary relationships, we filtered the positions with both 1/7 and 2/7. Since a MAF of 1/7 favours variant sites privative to the more genetically distinct individuals, we used a MAF of 2/7 for all downstream analyses (Supplementary Figures 2.5 and 2.6).

Phylogenetic analysis

A matrix of pairwise Hamming distances between the individuals was calculated using *Plink* (v. 1.9) (Purcell et al., 2007) and PCA was computed with the function *prcomp* from the *R* package *stats* V. 3.4.4 (R Core Team, 2018). Neighbor-net analysis and NJ tree calculations were performed using *SplitsTree* (v. 4.14.6) (Huson & Bryant, 2006). To test the phylogenetic relations among the species, we used *D*-statistics (Durand et al., 2011; Green et al., 2010). Based on the results of the previous analysis, we fixed *Oryctes mayottensis* as outgroup species while testing for the relations between *Marronus* and the rest of the beetles denoted as A and B in the following configuration: *D*(((A, B), *Marronus*), *O. mayottensis*). *D*-statistics were calculated using *popstats (Skoglund et al., 2015)*. Finally, we formally assessed the phylogeny of the scarab beetles by generating a species tree using *SVDquartets* (Chifman & Kubatko, 2014, 2015), a SNP-by-SNP quartet-based method. We used *SVDquartets* implementation in *PAUP V4.0a* (Swofford, 2002). We examined all possible quartets (N = 35) and assessed the node support by performing a bootstrap of 1000 replicates. Finally we generated species trees using the concatenated variant sites for implementing both a Maximum Likelihood and Bayesian-based methods. *RAxML-NG* (v. 0.9.0) (Kozlov et al., 2019) was used for the reconstruction of a ML-based phylogeny. We chose a GTR+G4 substitution model (Abadi et al., 2019) and performed 200 bootstrap replicates. We also performed a Markov-Chain-Monte-Carlo-based phylogenetic reconstruction using *BEAST*

(v. 2.6) (Bouckaert et al., 2019). To reduce the complexity of the model we chose a Strict Clock and to reduce the effect of demographic history assumptions we chose a Coalescent Extended Bayesian Skyline approach (A. J. Drummond et al., 2005). Both logs and trees from four independent MCMC chains of 10 million each, with ESS values over 200, were merged using *LogCombiner*. Finally, the Maximum Clade Credibility Tree was computed using *TreeAnnotator*.

In order to discard biases due to the effect of the chosen reference genome, we used the mapped reads to the *Marronus* reference genome (here presented) and repeated all the analysis following the same methodologies we described here (Supplementary Figure 2.8).

Divergence and time estimation

We calculated nucleotide divergence from each individual against the *O. borbonicus* reference genome. The values were calculated within non-overlapping windows of 100kb as Number of variant positions / Total number of positions. Only bases with a sequence support of >= 3X were selected. Nucleotide divergences were converted into time estimates using the reported estimates for arthropods (Brower, 1994).

Data and materials availability:

Both sequencing raw data and genome assemblies for *O. borbonicus* and *M. borbonicus* have been uploaded to the European Nucleotide Archive (ENA) under the study accession number PRJEB34604. Museum specimens raw data for both UDG and non-UDG treated libraries, have been uploaded under the study accession number PRJEB36751 (Supplementary Table 2.2). Pipelines and scripts are available at: https://gitlab.com/smlatorreo/museum_phylogenomics_extinct_oryctes_beetles

3. Genetic history if the rice blast fungus

Contributions

Parts of the content of this chapter have also been published in the article "Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus" *BMC Biology* 2020 doi:10.1186/s12915-020-00818-z. This chapter also includes newly sequenced present-day and herbarium samples. The results presented in the paper have been updated accordingly.

Sophien Kamoun (SK), Hernán A. Burbano (HAB) and myself conceived the study. Claudia Sarai Reyes-Avila, Angus Malmgren (AM) and Joe Win generated the effector database and called the presence and absence of the effectors in most of the isolates. AM and myself analyzed the profiles of effector presence and absence in the isolates. I performed the population history, developed the software application, analyzed the data and designed all the figures with input from SK and HAB. SK, HAB and myself wrote the paper. All authors read and approved the final manuscript.

Abstract

Background: Understanding the mechanisms and timescales of plant pathogen outbreaks requires a detailed genome-scale analysis of their population history. The fungus *Magnaporthe* (Syn. *Pyricularia*) *oryzae* —the causal agent of blast disease of cereals— is among the most destructive plant pathogens to world agriculture and a major threat to the production of rice, wheat and other cereals. Although *M. oryzae* is a multihost pathogen that infects more than 50 species of cereals and grasses, all rice-infecting isolates belong to a single genetically defined lineage. Here, we combined the two largest published genomic datasets together with newly sequenced herbarium and contemporary isolates to reconstruct the genetic history of the rice-infecting lineage of *M. oryzae* based on 149 isolates from 23 countries.

Results: The global population of the rice blast fungus consists mainly of three well-defined genetic groups and a diverse set of individuals. Multiple

population genetic tests revealed that the rice-infecting lineage of the blast fungus probably originated from a recombining diverse group in South East Asia followed by three independent clonal expansions that started ~400 years ago. Analysis of historical herbarium samples showed that one of the clonal lineages was already present in Europe during the XIX century, revealing genetic continuity of the rice blast fungus population in Europe for at least 124 years. Moreover, patterns of allele sharing identified a subpopulation from the recombining diverse group that introgressed with one of the clonal lineages before its global expansion. Remarkably, the four genetic lineages of the rice blast fungus vary in the number and patterns of presence and absence of candidate effector genes. These genes encode secreted proteins that modulate plant defense and allow pathogen colonization. In particular, two clonal lineages carry a reduced repertoire of effector genes compared with the diverse group, and specific combinations of presence and absence of effector genes define each of the pandemic clonal lineages.

Conclusions: Our analyses reconstruct the genetic history of the rice-infecting lineage of *M. oryzae* revealing three clonal lineages associated with rice blast pandemics. The inclusion of historical herbarium isolates revealed the genetic continuity of one of the clonal lineages in Europe for at least the last 124 years. Each of these clonal lineages displays a specific pattern of presence and absence of effector genes that may have shaped their adaptation to the rice host and their evolutionary history.

3.1. Introduction

Plant diseases are a persistent threat to food production due to a notable increase in the emergence and spread of new pathogens (M. C. Fisher et al., 2012; Savary et al., 2019). Understanding the mechanisms and timescales associated with new epidemics is essential for both basic studies and the implementation of effective response measures (Carvajal-Yepes et al., 2019). A fundamental component of this knowledge is a detailed genome-scale understanding of the population structure and dynamics of global plant pathogen populations (Croll & Laine, 2016; Grandaubert et al., 2019; Terauchi & Yoshida, 2010). Population genetic information drives the selection of isolates for activities as diverse as basic mechanistic research and plant germplasm

screening for disease resistance. It also helps to pinpoint the origin of pandemic strains and the evolutionary potential of different pathogen populations (Cooke et al., 2012; Goss et al., 2009; Hubbard et al., 2015; Islam et al., 2016; Radhakrishnan et al., 2019; Saunders et al., 2019). A thorough understanding of the global population structure is essential for any surveillance program that aims at rapidly detecting pathogen incursions into new geographical areas. In addition, the recent knowledge gained in the biology of pathogen effectors —secreted molecules that modulate host responses— brings yet another dimension to the population genetics framework, as it enables the reconstruction of the evolutionary history of virulence traits and helps guide the deployment of disease resistant cultivars (Cooke et al., 2012; Mohd-Assaad et al., 2019; Rietman et al., 2012; Vleeshouwers et al., 2008; Vleeshouwers & Oliver, 2014).

Fungal plant pathogens account for ~10-80% of crop losses in agriculture and are viewed as a major threat to global food security (Bebber et al., 2014; Bebber & Gurr, 2015; M. C. Fisher et al., 2012; Savary et al., 2019). Cereal crops like rice, oat, millet, barley and wheat have provided the foundation of modern agriculture and the success of humankind. Today's agriculture is facing the challenge of ensuring global food security for an ever-expanding world population, which is estimated to exceed 9 billion within the next 30 years (Nations & United Nations, 2019). The ascomycete fungus *Magnaporthe* (Syn. *Pyricularia*) *oryzae*, the causal agent of blast disease of cereals, is often ranked as the most destructive fungal pathogen, causing losses in rice production that, if mitigated, could feed several hundred million people (R. Dean et al., 2012; M. C. Fisher et al., 2012). Despite its Linnean name, *M. oryzae* is a multihost pathogen that can also cause the blast disease on other cereal crops, notably on wheat where it has recently spread from South America to Bangladesh resulting in destructive outbreaks (Inoue et al., 2017; Islam et al., 2016, 2019). *M. oryzae* reproduces mainly asexually and field isolates of *M. oryzae* are haploid. Asexual reproduction is the predominant mode of reproduction in almost all rice-growing regions, however, population genetics evidence has identified sexually reproducing populations in Southeast Asia, indicating that *M. oryzae* likely lost sexual reproduction outside of its center of origin (Saleh et al., 2012).

Comparative genomics analyses provided insights into the population structure and host-specialization of *M. oryzae* (Chiapello et al., 2015; Gladieux, Condon, et al., 2018; Yoshida et al., 2016). This pathogen consists of a complex assemblage of genetically distinct lineages that tend to be associated with particular host genera (Gladieux, Condon, et al., 2018). Remarkably, all rice-infecting isolates belong to a single genetic lineage that is thought to have originated from isolates infecting foxtail millet (*Setaria italica* and *Setaria viridis*). *M. oryzae* host-specific lineages exhibit limited gene flow but recurrent gene gain/loss particularly in regions of the genome linked to transposable elements (Chiapello et al., 2015; Yoshida et al., 2016). As in many other plant pathogens, effector genes exhibit a high degree of presence and absence polymorphisms and signatures of adaptive evolution (e.g. higher rate of non-synonymous over synonymous mutations) (Yoshida et al., 2016). Loss of so-called AVR effector genes —activators of host immunoresponses— can dramatically impact the fitness of the blast fungus by enabling virulence on resistant host genotypes (Białas et al., 2018; Inoue et al., 2017; Yoshida et al., 2009).

Although the genome sequence of the *M. oryzae* strain 70-15 was at the time of its publication the first fungal plant pathogen genome to be described (R. A. Dean et al., 2005), it took about a decade before comparative genomics analyses of this pathogen started to be reported (Chiapello et al., 2015; Xue et al., 2012; Yoshida et al., 2016). Until recently, understanding of the population genomics structure of the rice blast fungus has remained limited. In 2018, two studies reported whole genome sequences from non-overlapping sets of globally distributed rice-infecting *M. oryzae* isolates (Gladieux, Ravel, et al., 2018; Zhong et al., 2018). Both studies suggested the presence of a diverse Southeast Asian population and two major clonal groups. However, due to sampling or analytical limitations the two studies reached different conclusions about the composition of worldwide populations, i.e. the number of genetic groups, their temporal distribution and the historical processes that gave rise to them.

Here, we performed a combined analysis that builds on the studies of Gladieux *et al.* (Gladieux, Ravel, et al., 2018) and Zhong *et al.* (Zhong et al., 2018) and incorporates a new dataset composed by individuals isolated from rice infected herbarium leaves and modern individuals from the same geographic region. Our analyses permitted us to reconcile the two published datasets and increase

the number of examined *M. oryzae* individuals to 149 isolates from 23 countries. This has enabled us to assess the global genetic structure of rice-infecting *M. oryzae* more comprehensively than the prior separate analyses of the two datasets. Moreover, the use of time-stamped herbaria samples permitted an accurate estimate of *M. orzyae*'s mutation rate and the recalculation of divergence times among major lineages. We discovered that the global population of the rice blast fungus consists mainly of three well-defined genetic groups and a diverse set of individuals. Multiple population genetic tests revealed that the rice blast fungus probably originated from a recombining population in South East Asia followed by three independent clonal expansions that took place over the last ~200-400 years. The assignment of all the European historical samples to one of the clonal lineages, revealed the genetic continuity of the rice blast fungal population for at least 124 years. Patterns of allele sharing identified a subpopulation from the recombining group that introgressed with one of the clonal lineages before its global expansion. Remarkably, the genetic lineages of the rice blast fungus vary in the number and patterns of presence and absence of secreted protein predicted as effectors. In particular, two clonal lineages are defined by specific sets of effectors that may have shaped their adaptation to the rice host and their evolutionary history.

3.2. Results and discussion

3.2.1. The global population structure of rice-infecting *Magnaporthe oryzae* consists of three well defined genetic groups and a diverse set of individuals

To complement the published genomic datasets presented below, we generated new genome-wide genetic information from both present-day diversity and historical herbarium samples with lesions compatible with rice blast infection. Harnessing minute amounts of infected herbarium leaves, we used state-of-art protocols for DNA isolation and library construction to generate historical rice blast fungus genomes (Materials and Methods and Supplementary Figure 3.1).

To assess the global population structure of rice-infecting *M. oryzae*, we combined the new generated data from present-day samples (N=9) and

herbarium samples (N=9) with genome sequences from Gladieux *et al.* (N=43) (Gladieux, Ravel, et al., 2018) and Zhong *et al.* (N=88) (Zhong et al., 2018) for a total dataset of 149 genome sequences (Supplementary Table 3.1). The combined use of samples from different geographical locations and time spans increases not only the number of *M. oryzae* samples but also their geographical spread and temporal distribution (Figure 3.2B-C and Supplementary Table 3.1). We identified a total of 48,484 Single Nucleotide Polymorphism (SNPs) (see the "Methods" section). For subsequent analyses, we only used SNPs ascertained in all samples ("full information") (N=8,379 SNPs).

We first sought to investigate the number of distinct genetic groups in our global sample of *M. oryzae* given previous discrepancies in the number of clades or lineages identified in the two studies. We identified three well-defined groups and a diverse set of individuals based on two lines of evidence. First, we used f3-outgroup statistics (Raghavan et al., 2014) to evaluate the pairwise relatedness between *M. oryzae* samples relative to an outgroup. The f3-outgroup statistics measure the amount of shared evolutionary history between samples, which can be interpreted as shared genetic drift (always relative to an outgroup). We summarized the results of all tests by performing hierarchical clustering based on pairwise shared genetic drift comparisons, i.e. z-scores derived from f3-outgroup statistic tests (Figure 3.1). Additionally, we calculated pairwise Hamming genetic distances between all samples and summarized the information using Principal Component Analysis (PCA). The samples clustered again in three distinct groups and one diverse set of individuals using PC1 and 2, which together explained ~90% of the variance (Supplementary Figure 3.2A). We assessed the robustness of these clusters using Silhouette scores, which indicate how similar an individual is to its own cluster compared to other clusters (Lovmar et al., 2005). We found that the best mean Silhouette scores were obtained when the dataset was divided into four clusters (Supplementary Figure 3.2B).

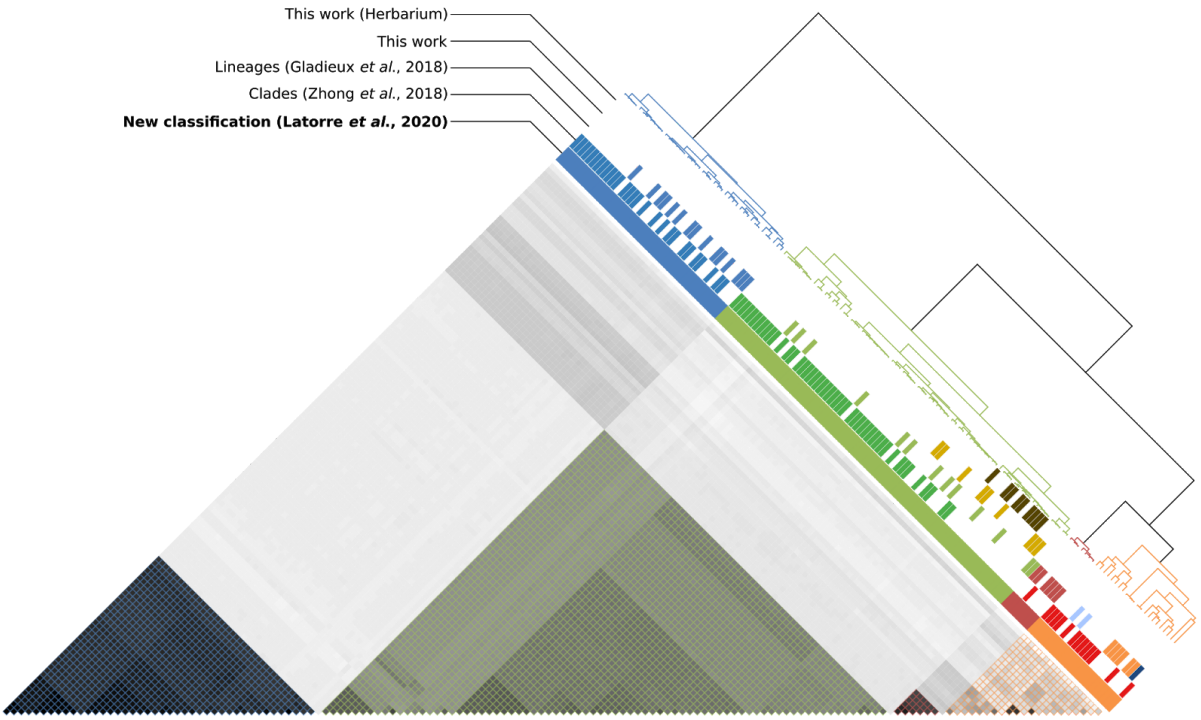


Figure 3.1. Genetic clustering of *Magnaporthe oryzae* reveals three defined groups and a diverse set of individuals. The pairwise relatedness between *M. oryzae* samples (X and Y) was estimated using f3-outgroup statistics of the form *f3*(X, Y; Outgroup), which measures the amount of shared genetic history (genetic drift) between X and Y after the divergence from an outgroup (*M. oryzae* strain from *Setaria*). The hierarchical clustering is based on *f3-scores* resulting from f3-outgroup statistic calculations. Darker colors indicate more shared drift.

Since our two approaches consistently revealed the presence of four groups, we named them group I, II, III and IV. Whereas group II and III are geographically widespread, Group I is mainly located in South-East Asia and group IV in the Indian subcontinent (Figure 2). The correspondence between our classification and previously described nomenclatures can be found in Supplementary Table 3.2. Our grouping very likely recapitulates the four lineages proposed by Saleh *et al.* based on microsatellite data (Saleh et al., 2014). Although it is not possible to directly link the microsatellite data with our analysis, we linked the correspondence between groups indirectly through the analysis presented in Gladieux *et al.*, the same group that previously performed the microsatellite analysis. Zhong *et al.* (Zhong et al., 2018) divided their dataset in three groups (I-III) but did not identify group IV, since their dataset only included one individual from this group. In addition to groups I-IV, Gladieux *et al.* (Gladieux,

Ravel, et al., 2018) identified two additional lineages based on a set of phylogenetic analyses. The combined analysis presented here showed that these additional lineages from Gladieux *et al*. are within the genetic diversity of group I, thus splitting of group I is not warranted. Remarkably, the fact that all historical and present-day samples from Europe belong to the genetic group II, suggests a process of genetic continuity in the European rice blast fungal populations for at least 124 years.

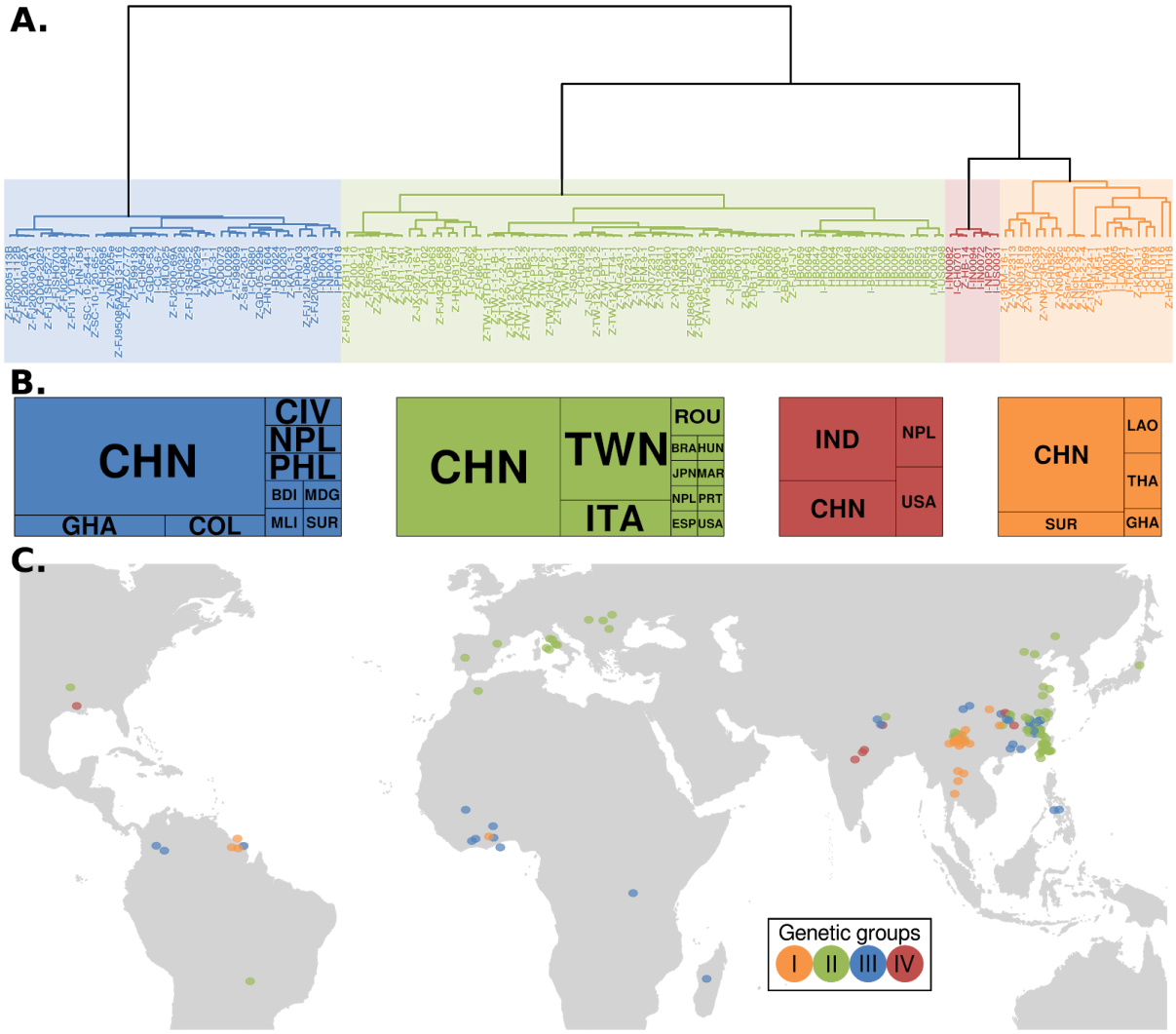


Figure 3.2. Geographic location of *Magnaporthe oryzae* isolates shows global distribution of **defined genetic groups (II-III) and a preferential South-East Asian location for the diverse group (I). (A)** Dendrogram showing the hierarchical clustering based on pairwise f3 values (same as Fig. 1). Prefixes of the isolate names correspond to the database source: G = Gladieux *et al.*, 2018 (Gladieux, Ravel, et al., 2018); Z = Zhong *et al.*, 2018 (Zhong et al., 2018); HB = This work. **(B)** Country of origin for *M. oryzae* isolates. The overall size of the boxes represents the total number

of samples within each genetic group. The size of each internal box is proportional to the number of isolates per country. Countries are represented as three-letter codes (ISO 3166-1 alpha-3): BDI=Burundi, BRA=Brazil, CHN=China, CIV=Côte d'Ivoire, COL=Colombia, ESP=Spain, GHA=Ghana, HUN=Hungary, IND=India, ITA=Italy, JPN=Japan, LAO=Lao People's Democratic Republic, MAR=Morocco, MDG=Madagascar, MLI=Mali, NPL=Nepal, PHL=Philippines, PRT=Portugal, ROU=Romania, SUR=Suriname, THA=Thailand, TWN=Taiwan, Province of China, USA=United States of America. **(C)** Geographical origin of samples used in this study. A random jitter was used on the coordinates of geographical-close samples for better visualization.

3.2.2. Global population of rice-infecting *Magnaporthe oryzae* probably arose from a recombining South East Asian population followed by clonal expansions

To determine the evolutionary origin of the four *M. oryzae* groups identified in this study, we used a set of statistics that evaluate genetic diversity, recombination and population differentiation. Initially, we visualized the relationships among samples using a phylogenetic network, which are more appropriate for visualizing reticulate evolution (Figure 3.3A) (Huson & Bryant, 2006). We found that group I exhibited a high degree of reticulation. In contrast, the phylogenetic network showed long internal branches with terminal star-shape phylogenetic configurations almost devoid of reticulations for the well-defined groups II, III and IV (Figure 3.3A). Such configurations are typical of expanding populations after genetic bottlenecks, driven, for instance, by clonal expansions (Exposito-Alonso et al., 2018). We, therefore, queried whether genetic diversity levels and recombination rates support clonality in groups II, III, and IV. Two lines of evidence support clonality in these groups compared with the diverse group I: i) reduced nucleotide diversity measured as pi (π) (Nei & Li, 1979) (Figure 3.3B); ii) lower detectable recombination events calculated using the four-gamete test (Hudson & Kaplan, 1985) (Figure 3.3C). The reduced levels of diversity in groups II, III, and IV in conjunction with their star-like phylogenies are tell-tale signs of populations that have experienced a strong reduction of diversity followed by a population expansion. Reductions in diversity followed by population expansion are typical of both demographic bottlenecks or founder effects (i.e., the establishment of a new population from a reduced number of individuals). Independent of the exact nature of the

demographic processes and evolutionary forces that gave rise to the changes in population size, the diversity and phylogenetic patterns that we observed are mostly driven by the population expansion phase. To calculate a proxy for recombination we used the four-gamete test, which puts a bound to the minimum number of recombination events in a sample (Hudson & Kaplan, 1985). Although it is known that this test underestimates recombination events, it is a simple and useful proxy for differences in recombination between populations. Our results showed that groups II, III and IV have on average ~5-fold less recombination events than the diverse group I. In agreement with these analyses, we characterized *M. oryzae*'s mating types based on the breadth of coverage of the *MAT-1* and *MAT-2* genes (Materials and Methods) and found the presence of only one mating type in the groups II, III and IV, whereas two mating types were segregating in the diverse group I (Supplementary Figure 3.3 and Supplementary Table 3.1). In sum, we conclude that group II, III and IV are likely clonal lineages, while group I consists of genetically diverse and recombining individuals (Figure 3.3A-C). The original microsatellite-based study by Saleh *et al.* (Saleh et al., 2014) reported a high level of genetic variability in group IV, however both our analyses and the ones carried out by Gladieux *et al.* (Gladieux, Ravel, et al., 2018) supported the clonal nature of this group.

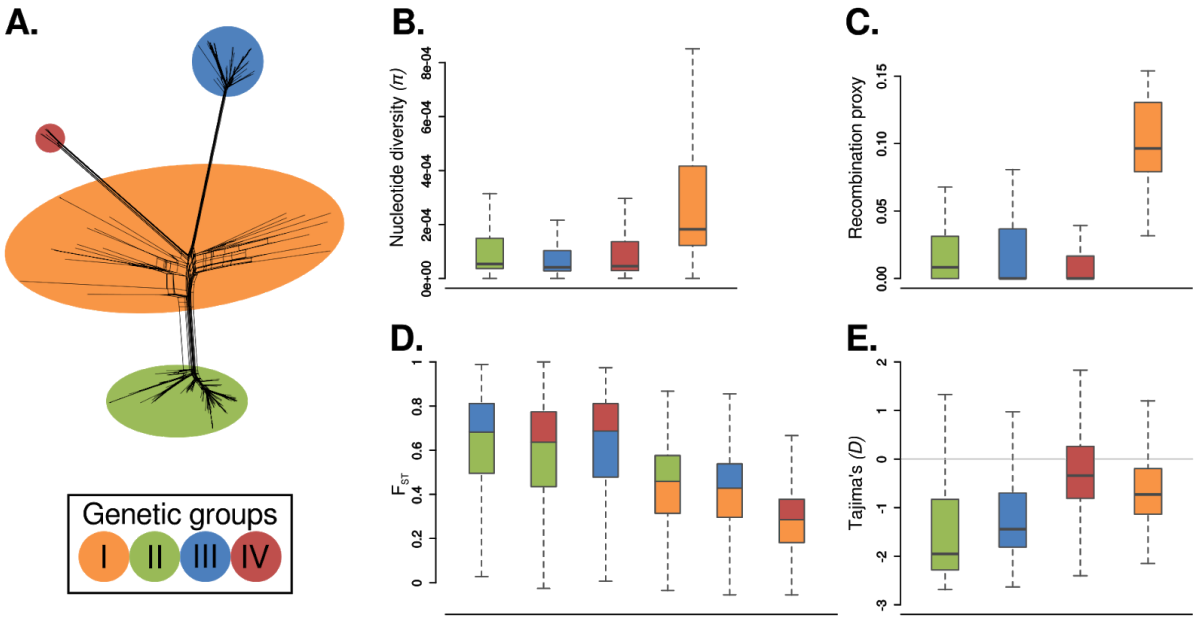


Figure 3.3. *Magnaporthe oryzae* population structure is driven by recombination and global clonal expansions. **(A)** Phylogenetic network showing the three well-defined groups (green, blue and red) and the diverse set of individuals (orange) from figure 1. **(B)** Within-population comparisons of nucleotide diversity measured as π. **(C)** Recombination proxy calculated by dividing the number of violations of the Four-gamete Test by the total number of SNPs. **(D)** Genetic distances between groups measured as Fixation Indices (F_{st}). The box colors depict the pairwise comparisons between groups. **(E)** Tajima's *D*.

To further investigate the relationships and demographic history of *M. oryzae* groups, we measured population differentiation among groups and leveraged the site frequency spectrum (SFS) for each group individually. To measure population differentiation we used F_{ST} (Weir & Cockerham, 1984) and found that when clonal groups II, III, and IV are compared among them, their F_{ST} distances were the highest. Although a fraction of the allele frequency differences that resulted in high F_{ST} values could have been driven by selection, the fact that on average F_{ST} values are much higher among clonal groups likely reflects a long history of independent drift. In contrast, whenever the diverse group I is compared with any of the clonal groups, the F_{ST} distances decreased, suggesting that group I is a common source of genetic diversity for all clonal lineages (Figure 3.3D). Subsequently, for every group we investigated their corresponding SFS using Tajima's *D* (Tajima, 1989), as this statistic records changes in allele frequencies driven, for instance, by variation in population sizes. We found that Tajima's *D* values for all clonal lineages were negative (Figure 3.3E). A demographic interpretation of negative Tajima's *D* values is consistent with population bottlenecks or founder effects followed by population expansions and a concurrent accumulation of rare alleles. Negative Tajima's *D* values are consistent with star-like phylogenies, as new mutations that occurred during the expansion phase accumulate in terminal branches lowering Tajima's *D* values. The inspection of the SFS also revealed an excess of high-frequency derived alleles, a feature of the SFS found mostly in rapidly adaptive populations, and that is particularly strong in asexual organisms or in organisms where meiotic recombination happen infrequently (Neher, 2013) (Supplementary Figure 3.4). By using multiple outgroups, we discarded that our observation is caused by misassignemnt of the ancestral allele. We believe, instead, that the excess of high-frequency derived alleles might be driven by a process dubbed genetic draft, i.e. the random association of alleles with genetic

backgrounds of different fitness (Gillespie, 2000). Thus, although the SFS is mainly driven by genetic drift during the population expansion phase - as manifested by the negative Tajima's *D* -, linked selection via genetic draft contributes to the fate of neutral alleles. Further theoretical work is needed to quantify the role of genetic draft in clonal populations of *M. oryzae*.

Overall, our results are consistent with a model where south-east Asia is a likely centre of origin of rice-infecting *M. oryzae* and in which three distinct clonal lineages arose from this ancestral population. These findings are consistent with previous models of the evolution of the rice lineage of *M. oryzae* (Saleh et al., 2014).

3.2.3. The expansion of *Magnaporthe oryzae* rice-infecting clonal lineages started 400 years ago

To estimate the divergence time of the clonal expansions of *M. oryzae*, we first used a Bayesian phylogenetic analysis leveraging the sampling dates of the combined dataset of herbarium and modern samples, which span 124 years. Such dates were used for tip-calibration (Alexei J. Drummond & Bouckaert, 2015; Heled & Drummond, 2012). To carry out the analysis, we first removed the diverse group I and used only the three clonal lineages, as the recombining group violates the assumptions of phylogenetic reconstruction. We used a concatenation approach including SNPs in the input pseudo-alignment. We also codified the amount of invariant sites in the configuration file, since the exclusion of invariant sites will lead to deeper divergence times. We estimated an evolutionary rate of 2.27 e-8 substitutions/site/year (1.93 e-8 - 2.63 e-8 HPD 95%), which was similar and contains in its HPD 95% a previously calculated rate (1.98e-8 substitutions/site/year) (Gladieux, Ravel, et al., 2018). Our approach of including only the clonal lineages permitted the reconstruction of a robust phylogeny and a more accurate estimation of divergence times, as reflected in the high posterior probabilities supporting the nodes and the narrow HPD 95% confidence intervals of node ages (Figure 3.4). This contrasts with previous studies that included individuals from the diverse recombining group I in the phylogenetic analysis and produced broader HPD 95% confidence intervals (Gladieux et al., 2018: Fig. 5 (Gladieux, Ravel, et al., 2018)).

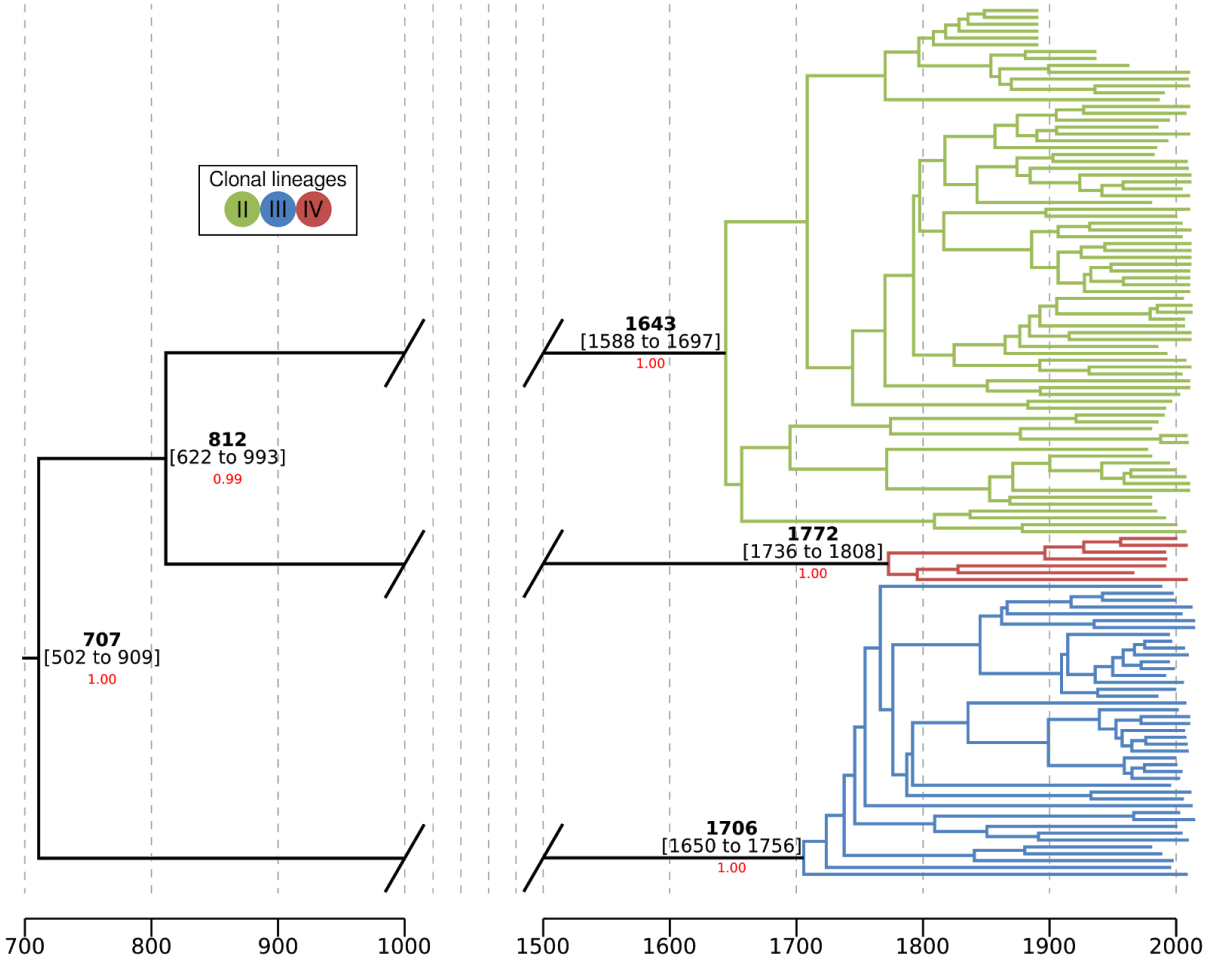


Figure 3.4. Clonal expansions of *Magnaporthe oryzae* took place in the last 400 years. Bayesian tip calibrated phylogenetic tree using individuals belonging to clonal lineages. Average, and HPD 95% confidence intervals are shown in calendar years. The Bayesian posterior probability is shown in red for nodes leading to the clonal lineage expansions.

The phylogenetic reconstructions revealed that all three clonal expansions occurred relatively recently over the last 400 years (243-372) (Fig. 3.4). These expansions happened concomitant with an increase of the effective population size of all clonal lineages (Supplementary Figure 3.5).

To assess the robustness of the phylogenetic reconstruction we carried out a coalescent-based method for multi-locus unlinked data, that infers the quartet trees for all subsets of isolates, and then combines the quartets in a single tree

(Chifman & Kubatko, 2014, 2015) (Supplementary Figure 3.7). This analysis confirmed the monophyly of each clonal group.

3.2.4. Patterns of allele frequency sharing identify introgression between a subpopulation of the diverse group I and clonal lineage II

Since the identification of admixture between populations facilitates the reconstruction of the evolutionary history of populations, we investigated the admixture history of *M. oryzae* using *D-statistics* (Durand et al., 2011; Green et al., 2010). This test employs counts of site patterns, which are patterns of alternative alleles at a given genomic position and evaluates whether these site patterns support one of two alternative discordant topologies. The *D*-statistics will return a value of zero if the two discordant phylogenies are supported equally, whereas positive or negative values indicate asymmetric support and, therefore, introgression. We test the three possible configurations of the following form: *D*(Outgroup, Diverse Group I; Clonal lineage X, Clonal lineage Y) (tree insets in Figure 3.5A-C). Whilst for clonal lineages II, III, and VI we used a strain representative for each clonal lineage, we performed a test for every one of the 22 members of the diverse group I. The test will retrieve positive values when the diverse group I is closer to clonal lineage Y and negative values when the diverse group I is closer to clonal lineage X. We found that group II has drifted farther apart from the diverse group I than the two other clonal lineages, as manifested from positive *D*-statistics when group II was included (as clonal lineage X) in the comparisons (Figure 3.5B-C). This accumulation of genetic drift is consistent with the fact that group II was the clonal lineage that diverged earliest from the recombining diverse group (Figure 3.4). We retrieved positive *D*-statistics in tests including almost all individuals of the diverse group I, with the exception of two individuals collected in China —*CH1016* (Gladieux, Ravel, et al., 2018) and *HB-LTH18* (Zhong et al., 2018)— that showed strong signals of genetic introgression with the clonal lineage II, as manifested by negative *D*-statistic values (Figure 3.5B-C). Since we detected introgression between these two Chinese samples and all members of group II, including herbarium isolates, regardless of their geographic origin (Supplementary Figure 3.8), we inferred that the admixture should have taken place before the clonal expansion that

gave rise to group II about 372 years ago (Figure 3.4). Previous attempts to detect interlineage recombination were not statistically robust and plagued with false positives (Gladieux, Ravel, et al., 2018). In contrast, *D*-statistics provide a statistically robust framework that reliably permits distinguishing between introgression and incomplete lineage sorting using genome-wide SNPs (Durand et al., 2011; Green et al., 2010).

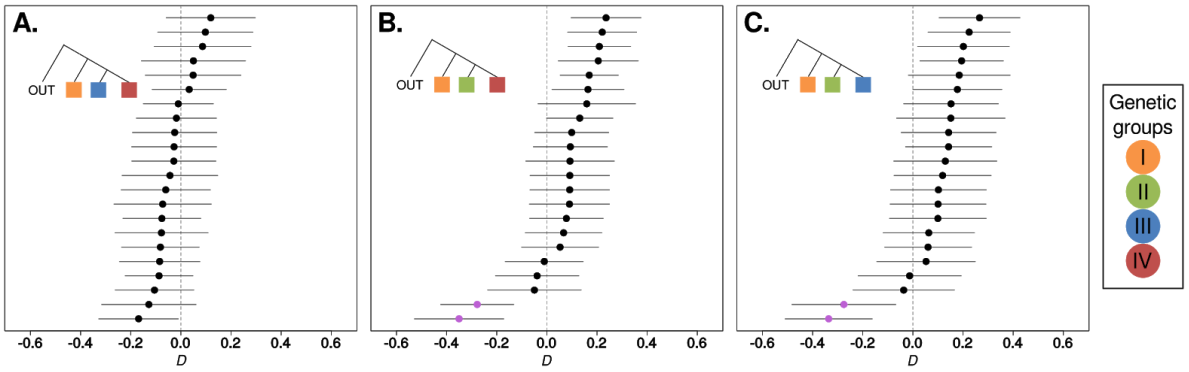


Figure 3.5. Patterns of allele frequency sharing identify introgression between a Chinese *Magnaporthe oryzae* subpopulation and clonal lineage II. *D*-statistics using three different phylogenetic configurations (depicted as colored inset trees). **(A)** *D*(Outgroup, Orange; Blue, Red). **(B)** *D*(Outgroup, Orange; Green, Red). **(C)** *D*(Outgroup, Orange; Green, Blue). In all cases, a *M. oryzae* strain from wheat was used as an outgroup and a fixed individual was selected as representative from each clonal lineage (Blue, Orange, Red). Points represent *D*-statistic tests for each of the 22 individuals assigned to the diverse clade (Orange), and lines depict 95% confidence intervals. Purple dots in **B.** and **C.** correspond to Chinese individuals *CH1016* and *HB-LTH18*, which are the closest individuals to the Green clonal lineage.

To further investigate the extent and location of the introgression between group II and the two Chinese Group I individuals (*CH1016/HB-LTH18*), we segmented the genomes of the two Chinese individuals based on their similarity at segregating sites to either Group I or Group II (Supplementary Figure 3.9B). This analysis revealed that the genome-wide percentage of Group II-like fragments in the Chinese individuals is 44.58%, including a ~4 Mb region in chromosome 3 (Supplementary Figure 3.9B). To test whether those fragments are a good proxy for the percentage of introgression we carried out two additional tests. First, we repeated the *D*-statistic test presented in figure 3.5B

and supplementary figure 3.9A, but this time removing the candidate introgressed fragments. In contrast to the outcome of the test with whole-genome data, this time the test did not indicate introgression, i.e., it was not different from zero (Supplementary Figure 3.9C). Second, we estimated the proportion of introgression by using a f4-ratio test (Patterson et al., 2012) with the following setup: (Group III, Group II, Group I (without introgressed Chinese individuals), Outgroup) / (Group III, Group II, Chinese introgressed individuals, Outgroup). This test estimated the mixture proportion to be ~31.68%, a lower but similar value to the overall percentage of identified Group II-like fragments in the Chinese individuals.

3.2.5. Lineages of *Magnaporthe oryzae* show distinct patterns of presence and absence of effector genes

In *M. oryzae*, regions of the genome containing effector genes exhibit a high rate of structural variation as illustrated by presence and absence polymorphisms (Yoshida et al., 2016). We investigated the distribution of known and predicted effector genes within the population structure framework we defined for the rice lineage of *M. oryzae*. We mapped the genome sequences of the 149 isolates to the sequences of 178 known and candidate effectors predicted from the genomes of *M. oryzae* from hosts as diverse as rice, wheat, finger millet, foxtail millet, oat and *Digitaria* spp. (Petit-Houdenot et al., 2019). This pan-effectorome set enabled us to capture as much effector gene diversity as possible. In total, 134 effectors were identified in the 149 isolates (Supplementary Table 3.3). Remarkably, the number of effectors per isolate varied from 111 to 128 with clonal lineages III and IV carrying a reduced repertoire of effector genes compared with the diverse genetic group (Figure 3.6A-B). In the particular case of clonal lineage II, there is also evidence of a substantial loss of effectors, yet counterbalanced by the presence of effectors at high population frequencies. Such effectors are either absent or segregate at low frequencies in the rest genetic groups. This indicates that clonal-expansion-driven bottlenecks not only reduced the overall genetic diversity of all pandemic clonal lineages but are associated with a less diverse repertoire of dispensable genes such as effectors. In pathogenic bacteria, a reduction in the effectiveness of purifying selection has been associated with an increase in gene loss (Hershberg et al., 2007). Moreover, gene loss is particularly prevalent in clonal pathogenic

bacteria and has been postulated as a source of phenotypic variation in these otherwise genetically similar species (Bolotin & Hershberg, 2015). The association between gene loss and reduced purifying selection in bacteria is a consequence of their strong deletional bias, i.e. bacteria with reduced effective population size experience genome reduction (Mira et al., 2001). In contrast, eukaryotes with small effective population sizes have larger genomes (Lynch & Conery, 2003) and filamentous plant pathogens are notorious for having repeat-driven genome expansions associated that display a “two-speed” architecture (Dong et al., 2015; Raffaele & Kamoun, 2012). This relation is, however, more complex in the rice blast fungal phylum Ascomycota, where both genome expansions and reductions have been observed (Kelkar & Ochman, 2012). It remains to be tested if the concurrent loss of genetic diversity and dispensable/non-core genes is a widespread consequence of clonality-driven bottlenecks in eukaryotic pathogens that undergo clonal expansions, or if (adaptive) phenotypic novelty resulting from gene loss drives clonal expansions.

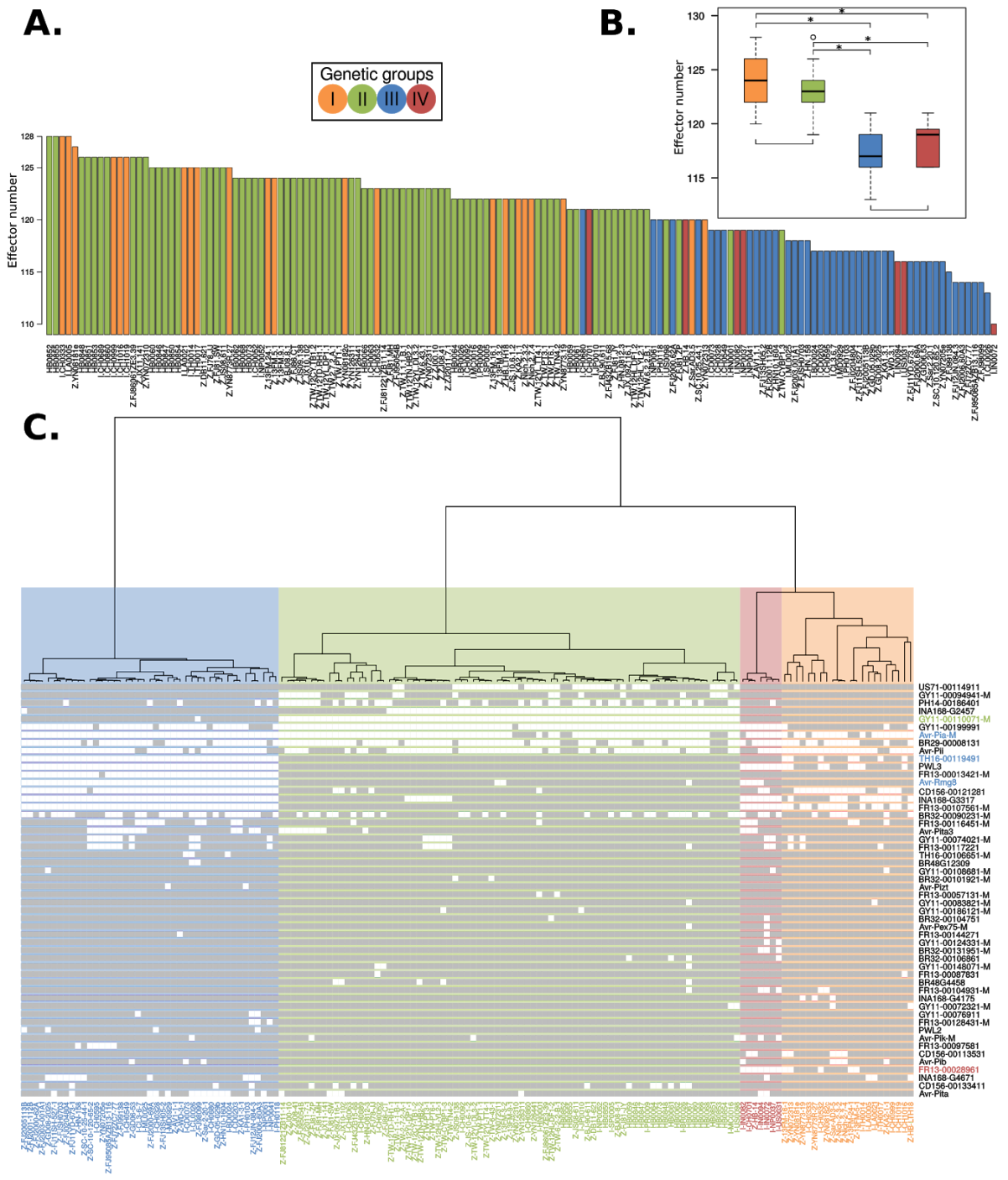


Figure 3.6. Rice blast genetic lineages vary in the number and patterns of presence and absence of candidate effector genes. **(A)** Clonal lineages carry a reduced repertoire of effector genes compared with the diverse group I. **(B)** The box-and-whisker plots show the distribution of effector number per isolate for each genetic group. Asterisks represent a *p-value* < 0.01 for a one-tailed Wilcoxon non-parametric test. **(C)** The dendrogram shows the clustering based on f3-outgroup statistic (as in Figure 3.1). Light and dark colors on the rows, represent absence and presence of effectors, respectively. Rows were grouped using a hierarchical clustering algorithm. Labels in green and blue font denote effectors missing in Clonal Group II and III, respectively.

We next mapped the distribution of the subset of 69 effectors that display presence and absence polymorphisms across all strains (Figure 3.6C). The resulting matrix clearly shows that there are distinct patterns of presence and absence of effectors across the genetically defined groups. For example, a set of three effectors (Avr-Pia-M, Avr-Rmg8 and TH16-00119491) are absent in Group III. Likewise, PWL3, INA168-3317 and FR13-00107561-M are absent in Groups III and IV; GY11-00110071-M is absent in Group II and FR13-00028961 is absent in Group IV (Figure 3.6C, Supplementary Table 3.4).

To determine which effectors have the strongest association with the defined genetic structure, we conducted two separate analyses based on the presence and absence effector repertoire per isolate. First, a PCA and effector loadings analysis revealed a set of 13 effectors that explained 90% of the variance of both PC1 and PC2 (Supplementary Figure 3.10A-B). Similarly, by using extremely randomized trees (a classification machine learning technique), we identified a set of 16 effectors that explained 90% of the variance (Supplementary Figure 3.1012C). Although the two methods produced different rankings of the impact of each effector gene, we found an overlap of 92.3% between the top 13 effectors found in the two subsets. In both cases the top effectors reproduced the separation of the isolates in the described genetic clusters (Supplementary Figure 3.10D-E). A close inspection of this group of top effectors, which were selected in an unbiased way, revealed that they are differentially (almost) present or (almost) absent in the four *M. oryzae* genetic groups (Supplementary Table 3.4). Thus, this group of effectors might have played an important role in the initial adaptation of *M. oryzae* clonal expansions to different rice subspecies and varieties.

The matrix in Figure 3.6C indicates that patterns of presence and absence of effector genes reflect different timescales in the evolution of the clonal lineages of *M. oryzae*. AVR effectors, such as AVR-Pita and AVR-Pii, show a patchy distribution within the clonal lineages. Their recurrent deletion in *M. oryzae* populations has generated virulent races (Yoshida et al., 2009). This may reflect the fact that their matching resistance genes have been repeatedly bred and deployed into rice cultivars. Other candidate effectors that display a similar

patchy distribution may be candidate AVR effectors that are detected by one of the dozens of blast resistance genes that have been bred into rice cultivars.

We also compare the presence and absence distribution between herbaria isolates and modern samples from the same geographical region (Italy and Romania) (Figure 3.7). Notwithstanding the relatively short amount of time between historical and modern isolates (~90 years), we found striking changes in their effector genes repertoire. In particular, the effector BR29-00008131 was completely absent in historical samples and present in all present-day samples. This pattern either supports gain of this effector in present-day Italian samples, or a loss of it in historical samples. The latter will require the existence of an unsampled population, where BR29-00008131 was still segregating. Moreover, the presence and absence pattern of other effector genes might indicate that they are on its way to fixation or deletion in the European subpopulations of the clonal lineage II (e.g. Avr-Pia-M, Avr-Pii, US71-00114911, GY11-00094941-M) (Figure 3.7).

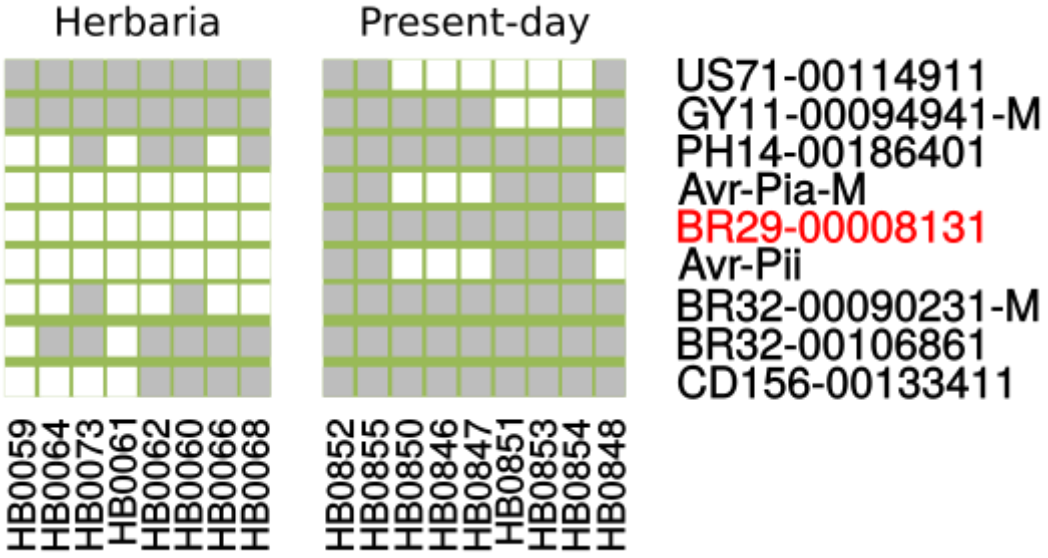


Figure 3.7. Differential presence and absence of effectors in herbarium and modern isolates. A subset of the matrix displayed in Figure 3.6C shows the different repertoire of polymorphic effector genes between herbaria and modern isolates from the same geographical region (Italy and Romania). Sample HB0067 was removed due to low coverage.

Our finding that the clonal lineages of rice-infecting *M. oryzae* display distinct repertoires of effectors raises a number of interesting questions. It is possible that this reflects the distinct genotype of the founding individual of the given clone. It is also possible that the absence of a given AVR effector(s) has facilitated the spread of the clonal lineage to otherwise resistant host genotypes as previously noted in *M. oryzae* (Huang et al., 2014; Inoue et al., 2017; Xue et al., 2012; Yoshida et al., 2016). In the future, it would be interesting to test the extent to which effectors that define the clonal lineages are detected by particular resistance genes. For example, AVR-Rmg8, which is known in wheat blast isolates to mediate avirulence on Rmg8 containing wheat varieties, may also be detected by a rice resistance gene. Future experiments will tease out the degree to which the distinct effector repertoires of the clonal lineages of *M. oryzae* reflect their adaptation to the rice host and their evolutionary history. Such analyses will require new genomic resources that permit a more accurate identification of effectors in canonical chromosomes and mini-chromosomes (Langner et al., 2020; Peng et al., 2019). To this aim, it will be fundamental to generate multiple reference genomes sequenced with long-read technologies in conjunction with a detailed characterization of structural variation and genomic rearrangements, which will include a per isolate inventory of mini-chromosome repertoires.

3.3. Conclusion

Our analyses reconstruct the genetic history of the rice-infecting lineage of *M. oryzae* revealing three clonal lineages that have emerged over the last ~400 years and have been associated with rice blast pandemics. These lineages display differential loss of effector genes that may have shaped their adaptation to the rice host and their evolutionary history. These findings provide a framework for further comparative analyses of the genomes of rice-infecting *M. oryzae*. One particular interesting research avenue will be to establish the degree to which structural variation, notably mini-chromosomes, have impacted the evolution of this lineage. Moreover, the inclusion of historical herbarium samples demonstrated the genetic continuity of the clonal lineage II in Europe for at least 124 years. The temporal resolution provided by the herbarium samples, also permitted to formulate testable hypotheses based on the

time-stamped differential patterns of presence and absence of effector genes. Thus, this framework opens the possibility to functionally test the role of gains and loss of effectors in planta in order to understand in greater detail the coevolution between rice and the rice blast fungus.

3.4. Materials and Methods

Herbarium samples DNA extraction

To prevent contamination by exogenous DNA, all the herbarium samples were handled using standard ancient DNA precaution measures, i.e. sterilization with UV light of all equipment, surfaces and hoods after each extraction round, and the use of different hoods for handling of samples, reagents and DNA extracts, and of protective gear by researchers. DNA extractions were carried out in the clean-room facility at the Institute of Archeological Sciences at the University of Tübingen. We used leaves with clear signs of tissue infection or lesions, previously described as *M. oryzae*. The tissue was ground inside a microtube with a stainless steel pestle until finely powdered and a N-phenacylthiazolium bromide (PTB) and Qiagen Plant DNEasy® Mini Kit (Qiagen)-based protocol was used to isolate the DNA (Gutaker et al., 2017). A microtube without tissue was used as negative DNA extraction control.

Library preparation and sequencing of herbarium samples

Genomic libraries for all herbarium samples were prepared in a DNA clean-room facility taking the same preventive measures described in the DNA extraction section.

Non-UDG treated library preparation of herbarium samples

We used a protocol that permits the preparation of indexed sequencing ancient DNA libraries (M. Meyer & Kircher, 2010). After adapter ligation in the clear-room facility, the indexing and PCR amplification of the libraries were performed in a different laboratory, located in a separate building. Briefly, the libraries were indexed using two barcoded primers (Kircher et al., 2012) during 10 cycles of PCR amplification using AccuPrime™ Pfx polymerase (Thermo Fisher Scientific). The MinElute PCR Purification Kit (Qiagen) was used to clean PCR residues and samples were pooled in equimolar concentrations. The

samples were sequenced at the Genome Center facility located at the Max Planck Institute for Developmental Biology, with the Illumina MiSeq Platform using MiSeq Reagent Kit v2, 300 cycles (Illumina). Together with the prepared libraries, both aDNA extraction and a library preparation negative controls were sequenced. All non-UDG library sequences were used only for authentication purposes (Supplementary Figure 3.1) and not included in any further analyses.

UDG treated library preparation of herbarium samples

Cytosine to thymine (C-to-T) substitutions are useful for the authentication of the samples but they can mislead evolutionary inferences (Briggs et al., 2007). Thus, in order to reduce the effect of C-to-T substitutions, we prepared new DNAlibraries adding uracil-DNA glycosylase (USER™ enzyme (New England Biolabs)) during the blunting step (Briggs et al., 2010). The rest of the steps were done as described in the non-UDG treated library section. Finally, after measuring the final DNA concentration per sample and assessing the DNA endogenous concentration, we prepared a pool with a calculated equimolar content of endogenous molecules. Sequencing was done using the Illumina HiSeq 3000 platform (Illumina) located at the Genome Center facility at the Max Planck Institute for Developmental Biology.

Herbarium specimens authentication

Ancient DNA has multiple signatures which can be used to its authentication: i) C-to-T substitutions are expected to be enriched at the ends of reads with a decay towards the inner part of the molecule and, ii) aDNA is expected to have shorter fragments than their contemporary counterparts due to depurination and further breakage of the DNA molecule (Dabney, Meyer, et al., 2013; Gutaker & Burbano, 2016). This depurination-driven breakage leaves signatures that can be analyzed after mapping reads to a reference genome (Briggs et al., 2007). We used MapDamage (v. 2.0) (Jónsson et al., 2013)to analyse both the C-to-T misincorporation patterns and fragment lengths. We also calculated the endogenous DNA content by dividing the number of mapped reads by the total number of reads. All herbarium samples displayed the expected signatures of historical samples (Supplementary Figure 3.1).

Library preparation of modern samples

Genomic high molecular weight DNA was used as input for a modified (Karasov et al., 2018) Nextera protocol (Caruccio, 2011). Sequencing was done using the Illumina HiSeq 3000 platform (Illumina) located at the Genome Center facility at the Max Planck Institute for Developmental Biology.

Pre-processing of reads

Raw reads were trimmed with AdapterRemoval (v. 2.2.2) (Schubert et al., 2016). In addition, reads from herbarium samples were also merged with the same software. Trimmed and / or merged reads were aligned to the *M. oryzae* reference genome (GUY-11 PacBio assembly) (Bao et al., 2017) using bwa-mem V.0.7.12 (Li and Durbin, 2009) with default parameters.

Datasets and mapping

We used *M. oryzae* Illumina reads from two recent resequencing studies (43 samples from Gladieux *et al.* (Gladieux, Ravel, et al., 2018), and 88 samples from Zhong *et al.* (Zhong et al., 2018) (Supplementary Table 3.1)). Raw sequencing reads were downloaded and mapped to the *M. oryzae* reference genome (*GUY-11* PacBio assembly (Bao et al., 2017)) using *bwa-mem* V.0.7.12 (Li & Durbin, 2009) with default parameters.

Variant identification and filtering

De novo variants were identified using *GATK* V.3.8.0 (A. McKenna et al., 2010). The following set of filters were applied: QD < 5.0; QUAL < 5000.0; MQ < 20.0; -2.0 < ReadPosRankSum < 2.0; -2.0 < MQRankSum < 2.0; -2.0 < BaseQRankSum < 2.0. In all subsequent analyses we used only biallelic SNPs present in all samples ("full information").

Characterization of *Magnaporte oryzae*'s mating types

Raw reads were mapped to a reference genome built from sequences of both mating type complete CDSs (MAT1-1 Genbank: AB080668.2 and MAT1-2 Genbank: AB080669.2) using *bwa-mem* V.0.7.12 (Li & Durbin, 2009) with default parameters. Coverage in both contigs was estimated using *samtools depth* V1.6 (Li et al., 2009)mating types were assessed accordingly (Supplementary Table 3.2).

Population Structure Analyses

To assess the global population structure of *M. oryzae* we first determined patterns of allele sharing using f3-outgroup statistics (Raghavan et al., 2014). We performed the test using the program *qp3Pop* from the *AdmixTools* package (Patterson et al., 2012). The test was used to establish the pairwise relatedness between *M. oryzae* samples (X and Y) after divergence from an outgroup: *f3*(X, Y; Outgroup). We used a deeply diverged *Setaria*-infecting *M. oryzae* strain HB0075 as an outgroup. We calculated z-scores for every possible pairwise sample comparison included in the f3-statistics test (N=20,458). Subsequently, we carried out hierarchical clustering using the function *hclust* from the *R* package *stats* (R Core Team, 2018). As input we used a distance matrix generated from the f3-statistics-derived z-scores (Figure 3.1A).

Additionally, we determined the level of population structure using genetic distances coupled with dimensionality reduction methods. We calculated pairwise Hamming distances using *Plink V.1.9* (Purcell et al., 2007). Such distances were used as input for Principal Component Analysis (PCA) using the function *prcomp* from the *R* package *stats* (R Core Team, 2018) (Supplementary Figure 3.2A). To assess the robustness of the clusters, PCA coordinates were used to compute silhouette scores using the function *silhouette* from the *R* package *cluster (Maechler et al., 2012)*. We calculated mean silhouette scores for different number of clusters (K=2-6) and found that the highest mean silhouette scores were obtained when K=4. We also used Discriminant Analysis of Principal Components (DAPC) (Jombart et al., 2010), implemented in the *adegenet R* package. The analysis was carried out by capturing the variance in the 10 first PC's. The Bayesian Information Criterion (BIC) indicated that the best number of groups was K=4 (Supplementary Figure 3.2C-D). We used the grouping of individuals in four clusters for subsequent analyses.

Population genetics analyses

We constructed a neighbor network using the program *SplitsTree V.4.14.6* (Huson & Bryant, 2006). As a proxy for recombination within each of the clusters, we used the four-gamete test (Hudson & Kaplan, 1985) as implemented in *RminCutter* (Ross-Ibarra, 2009). To this aim, we created consensus *fasta* sequences from the contigs 1 to 7 using the filtered vcf file with *bcftools* V. 1.3.1 (*Bcftools by samtools*, n.d.). The summary statistic was calculated by dividing the total number of violation events of the four-gamete test by the total number of

SNPs. Nucleotide diversity (π), Fixation Indices F_{ST} and Tajima's *D* values were calculated using *vcftools V.0.5.15* (Danecek et al., 2011). We calculated the unfolded Site Spectrum (SFS) for each genetic group using custom scripts. Ancestral alleles were ascertained requiring concordance between a *Setaria*- and a wheat-infecting outgroup strain (SA05-144 (Yoshida et al., 2016) and BTJP-4(12) (Soanes et al., 2017)).

We computed *D*-statistic values (Green et al., 2010) as:

$$D(O,T;X,Y) = \frac{(p_O - p_T)(p_X - p_Y)}{(p_O + p_T - 2p_O p_T)(p_X + p_Y - 2p_X p_Y)}$$

where P_O, P_T, P_X, and P_Y are frequencies of randomly selected alleles in populations (O)utgroup, (T)est, X, and Y at each locus. The reported 95% confidence intervals were calculated as $D \pm (SE \times 1.96)$ where the Standard Error was computed using a jackknife weighted by the number of SNPs for each 5 Mb block in the genome (Reich et al., 2009). We performed the calculations using *popstats* (Skoglund et al., 2015).

Genomic segmentation analysis

Based on the *D*-statistic results, two isolates from the diverse group I (CH1016 and HB-LTH18) showed genome-wide introgression evidence with the clonal lineage II. In order to identify which regions of the genomes of CH1016 and HB-LTH18 show higher nucleotide similarity to clonal lineage II than to members of the diverse group I, we performed a window-based similarity analysis. These regions, especially if they overlap between CH1016 and HB-LTH18, will be strong candidates for being introgressed from the clonal lineage II. Consequently, we performed window-based pairwise nucleotide similarity comparisons between an example isolate of clonal lineage II (TW-PT3-1) and the two Chinese individuals (CH1016 and HB-LTH18). To this end we divided the seven chromosomes in 400 windows, each of which had the same number of SNPs. To ascertain the basal level of similarity among clonal lineage II individuals we compared our example clonal lineage II isolate TW-PT3-1 with another clonal lineage II isolate (BR0026). Finally, to ascertain the nucleotide similarity between clonal lineage II and non-introgressed individuals from the diverse group I, we compared our example clonal lineage II isolate TW-PT3-1 with diverse group I isolates CH0532 and CH0333.

Phylogenetic analysis

We first carried out a Bayesian tip-dated phylogenetic analysis. To perform this analysis, we first removed individuals from the diverse group I, as these recombining group of individuals do not comply with the assumptions of any phylogenetic analysis (Figure 3.2C). We kept only biallelic variant positions to perform a Markov-Chain-Monte-Carlo-based phylogenetic reconstruction using *BEAST V.2.4.8* (Bouckaert et al., 2014). We used the isolates' collection dates (Supplementary Table 3.1) as prior information for the estimation of divergence times. We used *ModelTest-NG* (Darriba et al., n.d.) to assess the best suitable substitution model. Based on the lowest Akaike Information Criterion (AIC), we selected the General Time Reversible model. Since the calculation was performed with non-recombining individuals from the same species, we used a strict clock rate with a prior value of 1.98 e-8 substitutions/site/year, which was the rate ascertained in Gladieux *et al*. (Gladieux, Ravel, et al., 2018). To test the hypothesis of a non-clocklike data, we estimated the coefficient of variation in a model relaxed clock log normal model to be 0.0042, suggesting strong evidence for a clock-like data (Alexei J. Drummond et al., 2006). In order to reduce the effect of demographic history assumptions, and to calculate the dynamics of the population size through time, we also chose a Coalescent Extended Bayesian Skyline approach (A. J. Drummond et al., 2005). Invariant sites were explicitly considered in the model by adding a "*constantSiteWeights*" tag in the XML configuration file. We combined the output of four independent MCMC chains. Each chain had a length of 10 million iterations and was logged every 1000 iterations. We only used chains with overall ESS values above 200 and summarized a maximum clade credibility tree with *TreeAnnotator*. We summarized Effective Population Size through time using an Extended Bayesian Skyline Plot (Supplementary Figure 3.5). Configuration and log files are provided in our repository (see code and data availability).

To assess the robustness of the phylogenetic reconstruction to different sources of tree dicordances, we carried out several additional analyses. First, to illustrate the effect of recombination in the phylogeny, we included all the individuals from the genetic group I, who displayed signatures of sexual recombination (Figure 3.2C) and computed a new phylogeny following the same approach employed for the clonal groups (Supplementary Figure 3.6). As an alternative to

our described concatenation approach, we implemented a full phylogenetic method that takes into account Incomplete Lineage Sorting and assumes free recombination between variant sites. We used SVDquartets (Chifman & Kubatko, 2014, 2015) with its implementation in PAUP V4.0a. (Swofford, 2002). We selected a multispecies coalescent tree with an exhaustive examination of all possible quartets (N = 10,334,625). To assess node support, a bootstrap with 100 replicates was implemented (Supplementary Figure 3.7).

Effector genes repertoire

To determine the effector gene repertoire for each of 149 *M. oryzae* isolates described in Supplementary Table 3.1, we mapped the publicly available genomic short read sequences from these isolates to a reference set of diverse effector candidate sequences. We used the recently reported database from Petit-Houdenot and colleagues (Petit-Houdenot et al., 2019), which is composed by 195 candidates with similarity to both AVR and MAX effectors from isolates infecting a wide variety of hosts (e.g. rice, wheat, finger millet, foxtail millet, oat and other *Digitaria* species). We reduced the redundancy of the reference by removing highly similar sequences (≥ 90% identity). The final reference set included 178 coding DNA sequences for candidate effectors (Supplementary Table 3.5) from different *M. oryzae* lineages infecting hosts such as rice, wheat, oat, millet, and wild grasses. The coordinates of the reference effector genes corresponding to *M. oryzae* PacBio genome GUY-11 (GenBank accession GCA_002368485.1) are stored in the Gitlab repository (see availability od data and materials). We used elongation factor 2 mRNA sequence (GenBank accession XM_003714691.1) from *M. oryzae* as a positive control for presence of a gene, and a secreted protein gene *CoMC69* from the fungus *Colletotrichum orbiculare* as a negative control for absence of a gene in the reference for short read mapping.

Mapping was performed with *bwa-mem* V.0.7.15 (Li, 2013). An effector was deemed present if more than 80% of its sequence was recovered with a minimum depth of 3x, using *SAMtools* V1.6 (Li et al., 2009).

To summarize effector content per isolate, we built a presence and absence matrix indicating presence and absence of effector genes with 1 and 0, respectively (Figure 3.6A-B and Supplementary Table 6). For subsequent

analyses we excluded effector genes that were either present or absent in all lineages, as they are uninformative for clustering algorithms. This filtering resulted in a presence and absence matrix that contains a set of 69 informative effectors. We organized the columns of this matrix according to the dendrogram of genetic groups (Figure 3.1), while the rows were sorted using hierarchical clustering with the function *hclust* from *R stats* package (R Core Team, 2018).

To determine which effectors have the strongest association with the defined genetic structure of *M. oryzae*, we conducted PCA and loading analysis using the presence and absence matrix per isolate as input. Analyses were carried out with the *princomp* function of the *R stats* package (R Core Team, 2018). Then, by multiplying the absolute value of X and Y coordinates of each loading vector in PC1 and PC2, we assessed the strength of importance per effector (Supplementary Figure 3.10A). We selected a subset of effectors that contains the 13 most important effectors (i.e. loading vectors with the highest magnitudes), as they together explained 90% of the variance (Supplementary Figure 3.10C). We recalculated the PCA using only this subset of 13 effectors, which resulted in an increase from 44.8% to 73.8% (total increase of 29%) of the variance explained by PC1 and PC2 together.

A similar analysis was also carried out using the Extremely Randomized Trees algorithm implemented in the *Python scikit-learn* module (Pedregosa et al., 2012), with 100 trees per forest, and trained using all the effector presence and absence data. The feature importances were extracted from the trained model. This process was repeated 2,500 times to ensure consistency, and the mean effector importance for reconstructing the population structure was calculated in order to rank the effectors. Using this method, 90% of the variance was explained by 16 effectors (Supplementary Figure 3.10C).

Availability of data and materials

The datasets and scripts generated during and/or analysed during the current study are available in the Gitlab repository, https://gitlab.com/smlatorreo/genetic_history_of_rice-infecting_magnaporthe_oryzae (Latorre et al., 2020).

4. Conclusions and Outlook

During the last decade, evolutionary studies have been transitioning to whole genomic information as the primary source for the evolutionary inferences. This major change has only been possible due to simultaneous technological developments and decreasing costs of DNA isolation methods (Gutaker et al., 2017; Kistler, 2012; Rogers & Bendich, 1985; Rohland & Hofreiter, 2007), library preparation protocols (Caruccio, 2011; Gansauge et al., 2017; Gansauge & Meyer, 2013; M. Meyer & Kircher, 2010), WGS technologies (Kircher & Kelso, 2010; Metzker, 2005, 2010; Jun Zhang et al., 2011) and increasing computational capacity to analyze big datasets (Stein, 2011). Taking advantage of this major leap, ancient DNA-based studies have led to the incorporation of temporal information into evolutionary studies. While evolutionary studies have been interpreting past events by inferring them from contemporary genetic variance, aDNA studies can directly use the temporal information associated with subfossil records, museum specimens and herbarium samples. These sources provide timepoint anchors from which past events can be calibrated and translated into absolute time (Heled & Drummond, 2012; Ho et al., 2011; Rambaut, 2000; Rieux & Balloux, 2016).

During my doctorate, I used aDNA to understand phylogenomic relations between species, as well as describing the demographic history of different populations from the same species. The combined use of historical samples (pinned insect museum specimens and herbarium leaves), together with present-day datasets, enabled me to better study the colonization of the Mascarene islands by scarab beetles, and describe the population structure as well as the demographic history of global populations of the the rice blast fungus.

In particular, with the incorporation of museum specimens from extinct rhinoceros beetles from the genus *Oryctes*, I ascertained their evolutionary relationships by building for the first time, a molecular phylogeny using site patterns-based *D*-statistics. Previous efforts were hindered by the recent extinctions within the genus, and by the implicit problems of morphological-based phylogenies. Moreover, I showed that the unrelated

monotypic genus *Marronus* (species *M. borbonicus*) falls instead inside the observed variation of the genus *Oryctes*, requiring a taxonomic reclassification, which will be presented elsewhere. Furthermore, by analyzing their speciation history in the light of the geological events that took place in the the Mascarene islands, I inferred: i) two independent colonization events that took place in the Réunion island and, ii) a case of loss of flight ability and dwarfism after the colonization of Réunion by *M. borbonicus*.

As shown with scarab beetles from the Mascarene islands, the inclusion of recently extinct taxa by sequencing whole-genome data from museum specimens can significantly improve the reconstruction of the tree of life. In the particular case of insects, which are the most diverse taxon on earth, most of its taxonomic and phylogenetic studies have been ascertained using morphological characters and/or a handful of molecular markers (Hunt et al., 2007; Jin et al., 2016). As I demonstrated here, morphological characters can be obscured or hidden, especially when adaptive radiations take place (Brower, 1994; D. D. McKenna et al., 2019; Pease et al., 2016; Jing Zhang et al., 2019), while phylogenetic trees based on few genetic markers might not reflect the relationships between lineages or species (Philippe et al., 2005). The inspection of whole genome datasets will encourage a vast revision of the tree of life. Furthermore, because of the rapid extinction rate of species, particularly in islands, retrospective approaches are necessary to gather a full picture from extinct, yet informative lineages. Indeed, fundamental knowledge has been gained by analyzing the genetic contributions from extinct species and populations to the extant lineages, specifically in the adaptive and non-adaptive roles played by introgression (Dannemann et al., 2016; Speidel et al., 2019; Vattathil & Akey, 2015; Vernot & Akey, 2014).

In this work, I also used a combined dataset of historical samples and contemporary diversity to unveil the population structure and demographic history of the rice blast pathogen *Magnaporthe oryzae*. I concluded that the current diversity of the pathogen is organized in three main clonal lineages and a set of recombining individuals, which are the likely source of genetic diversity. I described a process by which, after experiencing a bottleneck, the three clonal lineages acquired population-specific genetic diversity in the recent past (~ 400 years ago). These population-specific differences seem to be

partially correlated to the genetic make-up of the host organism (*Oryza sativa*). Moreover, because effector secreted proteins are a fundamental part of the coevolutionary process between host and pathogen - since they participate in protein-protein interactions between them -, I analyzed the presence and absence of effector secreted proteins among the different *M. oryzae* isolates. Their different patterns suggest that the arms race between plants and their pathogens has shaped the effector repertoire of *M. oryzae*'s clonal lineages. Furthermore, the incorporation of historical samples permitted a better estimation of the above-mentioned characteristics, as well as enabled to ascertain the divergence times of different lineages in the demographic history of the pathogen, and to formulate hypotheses that warrant to be tested with functional assays.

It is still unknown whether genetic drift is the predominant evolutionary force in clonally expanding *M. oryzae* populations, or whether different events like introgression, structural variation and retention of extrachromosomal material like mini-chromosomes, are important sources of diversity and functional adaptation among different lineages. Further studies integrating short- and long-read technologies will be fundamental to quantify the role played by these sources of variation.

Finally, the anthropic footprint and its associated consequences on climate change and other drivers of environmental change, are endangering entire ecosystems and biodiversity hotspots worldwide, which directly translates into an increase of the already high rate of species extinction (De Vos et al., 2015; McLaughlin et al., 2002; Veron et al., 2019). A full comprehension of past extinctions of lineages should not just analyze the environmental triggers and causes, but also the consequences at genomic levels over the extinct and extant lineages. Evolutionary adaptation processes in extreme and harsh environmental scenarios, can shed light into predicting, preventing and therefore contributing to informing conservation efforts of biodiversity hotspots and their associated species (Carlson et al., 2014; Exposito-Alonso et al., 2020; Lewin et al., 2018; Tigano & Friesen, 2016; Waldvogel et al., 2020).

With a study case, I have demonstrated the power of combining phylogenomics and aDNA to understand colonization and phylogenetic relationships between different species with shared histories in the Mascarene islands, one of the most

endangered biodiversity hotspots. This study paves the way for harnessing the rich and valuable information present/found in natural museum collections around the world, to study the genetic footprints of extinction and adaptation (Tegelberg et al., 2014) in insects, since they are an ecologically important, yet understudied taxa (Warren et al., 2018).

As the temperature increases different populations are predicted to migrate laditunally. Moreover, important anthropic activities such as agriculture exploitations are expected to follow the same trajectory, expanding the spectra of migrations to crop associated organisms, particularly pests and pathogens (Bebber et al., 2013; Lindner et al., 2010). Consequently, plant pathogen pandemics are expected to increase together with the expected migration trajectories (Thomas et al., 2004; Warren et al., 2018). In a case study, I have utilized historical and present-day samples from the most important rice pathogen. As concluded, similar frameworks can differentially analyze the genetic diversity with special effort on those geographical regions where big population sizes are expected to arise. Understanding already established populations in those regions, can help to forecast and elucidate different mechanisms used by pests and pathogens to adapt to new environments.

References

Abadi, S., Azouri, D., Pupko, T., & Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, *10*(1), 934. https://doi.org/10.1038/s41467-019-08822-w

Allentoft, M. E., Heller, R., Oskam, C. L., Lorenzen, E. D., Hale, M. L., Gilbert, M. T. P., Jacomb, C., Holdaway, R. N., & Bunce, M. (2014). Extinct New Zealand megafauna were not in decline before human colonization. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13), 4922–4927. https://doi.org/10.1073/pnas.1314972111

Avise, J. C., Shapira, J. F., Daniel, S. W., Aquadro, C. F., & Lansman, R. A. (1983). Mitochondrial DNA differentiation during the speciation process in Peromyscus. *Molecular Biology and Evolution*, *1*(1), 38–56. https://doi.org/10.1093/oxfordjournals.molbev.a040301

Bao, J., Chen, M., Zhong, Z., Tang, W., Lin, L., Zhang, X., Jiang, H., Zhang, D., Miao, C., Tang, H., Zhang, J., Lu, G., Ming, R., Norvienyeku, J., Wang, B., & Wang, Z. (2017). PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe oryzae. *Molecular Plant*, *10*(11), 1465–1468. https://doi.org/10.1016/j.molp.2017.08.008

Barrière, A., Yang, S.-P., Pekarek, E., Thomas, C. G., Haag, E. S., & Ruvinsky, I. (2009). Detecting heterozygosity in shotgun genome assemblies: Lessons from obligately outcrossing nematodes. *Genome Research*, *19*(3), 470–480. https://doi.org/10.1101/gr.081851.108

Baute, G. J., Kane, N. C., Grassa, C. J., Lai, Z., & Rieseberg, L. H. (2015). Genome scans reveal candidate domestication and improvement genes in cultivated sunflower, as well as post-domestication introgression with wild relatives. *The New Phytologist*, *206*(2), 830–838. https://doi.org/10.1111/nph.13255

Bcftools by samtools. (n.d.). Retrieved August 30, 2018, from http://samtools.github.io/bcftools/

Bebber, D. P., & Gurr, S. J. (2015). Crop-destroying fungal and oomycete

pathogens challenge food security. *Fungal Genetics and Biology: FG & B*, *74*, 62–64. https://doi.org/10.1016/j.fgb.2014.10.012

Bebber, D. P., Holmes, T., & Gurr, S. J. (2014). The global spread of crop pests and pathogens. *Global Ecology and Biogeography: A Journal of Macroecology*, *23*(12), 1398–1407. https://doi.org/10.1111/geb.12214

Bebber, D. P., Ramotowski, M. A. T., & Gurr, S. J. (2013). Crop pests and pathogens move polewards in a warming world. *Nature Climate Change*, *3*(11), 985–988. https://doi.org/10.1038/nclimate1990

Begun, D. J., Holloway, A. K., Stevens, K., Hillier, L. W., Poh, Y.-P., Hahn, M. W., Nista, P. M., Jones, C. D., Kern, A. D., Dewey, C. N., Pachter, L., Myers, E., & Langley, C. H. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. *PLoS Biology*, *5*(11), e310. https://doi.org/10.1371/journal.pbio.0050310

Białas, A., Zess, E. K., De la Concepcion, J. C., Franceschetti, M., Pennington, H. G., Yoshida, K., Upson, J. L., Chanclud, E., Wu, C.-H., Langner, T., Maqbool, A., Varden, F. A., Derevnina, L., Belhaj, K., Fujisaki, K., Saitoh, H., Terauchi, R., Banfield, M. J., & Kamoun, S. (2018). Lessons in Effector and NLR Biology of Plant-Microbe Systems. *Molecular Plant-Microbe Interactions: MPMI*, *31*(1), 34–45. https://doi.org/10.1094/MPMI-08-17-0196-FI

Bolotin, E., & Hershberg, R. (2015). Gene Loss Dominates As a Source of Genetic Variation within Clonal Pathogenic Bacterial Species. *Genome Biology and Evolution*, *7*(8), 2173–2187. https://doi.org/10.1093/gbe/evv135

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., Suchard, M. A., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *10*(4), e1003537. https://doi.org/10.1371/journal.pcbi.1003537

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Müller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., … Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*(4), e1006650.

https://doi.org/10.1371/journal.pcbi.1006650

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., & Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics*, *140*(2), 783–796. https://www.ncbi.nlm.nih.gov/pubmed/7498754

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., & Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(37), 14616–14621. https://doi.org/10.1073/pnas.0704665104

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., & Pääbo, S. (2010). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Research*, *38*(6), e87. https://doi.org/10.1093/nar/gkp1163

Brower, A. V. (1994). Rapid morphological radiation and convergence among races of the butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *91*(14), 6491–6495. https://doi.org/10.1073/pnas.91.14.6491

Cardoso, P., Barton, P. S., Birkhofer, K., Chichorro, F., Deacon, C., Fartmann, T., Fukushima, C. S., Gaigher, R., Habel, J. C., Hallmann, C. A., Hill, M. J., Hochkirch, A., Kwak, M. L., Mammola, S., Ari Noriega, J., Orfinger, A. B., Pedraza, F., Pryke, J. S., Roque, F. O., … Samways, M. J. (2020). Scientists' warning to humanity on insect extinctions. *Biological Conservation*, *242*, 108426. https://doi.org/10.1016/j.biocon.2020.108426

Carlson, S. M., Cunningham, C. J., & Westley, P. A. H. (2014). Evolutionary rescue in a changing world. *Trends in Ecology & Evolution*, *29*(9), 521–530. https://doi.org/10.1016/j.tree.2014.06.005

Caruccio, N. (2011). Preparation of Next-Generation Sequencing Libraries Using Nextera™ Technology: Simultaneous DNA Fragmentation and Adaptor Tagging by In Vitro Transposition. In Y. M. Kwon & S. C. Ricke (Eds.), *High-Throughput Next Generation Sequencing: Methods and Applications* (pp.

241–255). Humana Press. https://doi.org/10.1007/978-1-61779-089-8_17

Carvajal-Yepes, M., Cardwell, K., Nelson, A., Garrett, K. A., Giovani, B., Saunders, D. G. O., Kamoun, S., Legg, J. P., Verdier, V., Lessel, J., Neher, R. A., Day, R., Pardey, P., Gullino, M. L., Records, A. R., Bextine, B., Leach, J. E., Staiger, S., & Tohme, J. (2019). A global surveillance system for crop diseases. *Science*, *364*(6447), 1237–1239. https://doi.org/10.1126/science.aaw1572

Cheke, A., & Hume, J. P. (2010). *Lost Land of the Dodo: The Ecological History of Mauritius, Réunion and Rodrigues*. Bloomsbury Publishing. https://play.google.com/store/books/details?id=8xXSBAAAQBAJ

Chiapello, H., Mallet, L., Guérin, C., Aguileta, G., Amselem, J., Kroj, T., Ortega-Abboud, E., Lebrun, M.-H., Henrissat, B., Gendrault, A., Rodolphe, F., Tharreau, D., & Fournier, E. (2015). Deciphering Genome Content and Evolutionary Relationships of Isolates from the Fungus Magnaporthe oryzae Attacking Different Host Plants. *Genome Biology and Evolution*, *7*(10), 2896–2912. https://doi.org/10.1093/gbe/evv187

Chifman, J., & Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics* , *30*(23), 3317–3324. https://doi.org/10.1093/bioinformatics/btu530

Chifman, J., & Kubatko, L. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, *374*, 35–47. https://doi.org/10.1016/j.jtbi.2015.03.006

Cooke, D. E. L., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J., Deahl, K. L., Farrer, R. A., Gilroy, E. M., Goss, E. M., Grünwald, N. J., Hein, I., MacLean, D., McNicol, J. W., Randall, E., Oliva, R. F., Pel, M. A., Shaw, D. S., Squires, J. N., … Kamoun, S. (2012). Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS Pathogens*, *8*(10), e1002940. https://doi.org/10.1371/journal.ppat.1002940

Cooper, A., & Poinar, H. N. (2000). Ancient DNA: do it right or not at all [Review of *Ancient DNA: do it right or not at all*]. *Science*, *289*(5482), 1139. https://doi.org/10.1126/science.289.5482.1139b

Cridland, J. M., Ramirez, S. R., Dean, C. A., Sciligo, A., & Tsutsui, N. D. (2018). Genome Sequencing of Museum Specimens Reveals Rapid Changes in the Genetic Composition of Honey Bees in California. *Genome Biology and Evolution*, *10*(2), 458–472. https://doi.org/10.1093/gbe/evy007

Croll, D., & Laine, A.-L. (2016). What the population genetic structures of host and pathogen tell us about disease evolution [Review of *What the population genetic structures of host and pathogen tell us about disease evolution*]. *The New Phytologist*, *212*(3), 537–539. https://doi.org/10.1111/nph.14203

Cui, Y., Yu, C., Yan, Y., Li, D., Li, Y., Jombart, T., Weinert, L. A., Wang, Z., Guo, Z., Xu, L., Zhang, Y., Zheng, H., Qin, N., Xiao, X., Wu, M., Wang, X., Zhou, D., Qi, Z., Du, Z., … Yang, R. (2013). Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(2), 577–582. https://doi.org/10.1073/pnas.1205750110

Dabney, J., Knapp, M., Glocke, I., Gansauge, M.-T., Weihmann, A., Nickel, B., Valdiosera, C., García, N., Pääbo, S., Arsuaga, J.-L., & Meyer, M. (2013). Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(39), 15758–15763. https://doi.org/10.1073/pnas.1314445110

Dabney, J., Meyer, M., & Pääbo, S. (2013). Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*, *5*(7). https://doi.org/10.1101/cshperspect.a012567

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics* , *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Dannemann, M., Andrés, A. M., & Kelso, J. (2016). Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *American Journal of Human Genetics*, *98*(1), 22–33. https://doi.org/10.1016/j.ajhg.2015.11.015

Darriba, D., Posada, D., & Stamatakis, A. (n.d.). *ModelTest-NG* (Version 0.1.5) [Computer software]. Github. Retrieved August 1, 2019, from https://github.com/ddarriba/modeltest

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., Thon, M., Kulkarni, R., Xu, J.-R., Pan, H., Read, N. D., Lee, Y.-H., Carbone, I., Brown, D., Oh, Y. Y., Donofrio, N., Jeong, J. S., Soanes, D. M., Djonovic, S., … Birren, B. W. (2005). The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature*, *434*(7036), 980–986. https://doi.org/10.1038/nature03449

Dean, R., Van Kan, J. A. L., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., Rudd, J. J., Dickman, M., Kahmann, R., Ellis, J., & Foster, G. D. (2012). The Top 10 fungal pathogens in molecular plant pathology. *Molecular Plant Pathology*, *13*(4), 414–430. https://doi.org/10.1111/j.1364-3703.2011.00783.x

Dechambre, R. P., & Lachaume, G. (2001). *Dynastidae. Le genre Oryctes. Les Coléoptères du Monde*. Hillside Books, Canterbury, United Kingdom.

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, *24*(6), 332–340. https://doi.org/10.1016/j.tree.2009.01.009

De Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R., & Pimm, S. L. (2015). Estimating the normal background rate of species extinction: Background Rate of Extinction. *Conservation Biology: The Journal of the Society for Conservation Biology*, *29*(2), 452–462. https://doi.org/10.1111/cobi.12380

Dong, S., Raffaele, S., & Kamoun, S. (2015). The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Genetics & Development*, *35*, 57–65. https://doi.org/10.1016/j.gde.2015.09.001

Drake, J. W. (2006). Chaos and order in spontaneous mutation. *Genetics*, *173*(1), 1–8. https://www.ncbi.nlm.nih.gov/pubmed/16723419

Dréau, A., Venu, V., Avdievich, E., Gaspar, L., & Jones, F. C. (2019). Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nature Communications*, *10*(1), 4309.

https://doi.org/10.1038/s41467-019-12210-9

Drummond, A. J., & Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*. Cambridge University Press. https://play.google.com/store/books/details?id=Nn25CgAAQBAJ

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, *4*(5), e88. https://doi.org/10.1371/journal.pbio.0040088

Drummond, A. J., Rambaut, A., Shapiro, B., & Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, *22*(5), 1185–1192. https://doi.org/10.1093/molbev/msi103

Durand, E. Y., Patterson, N., Reich, D., & Slatkin, M. (2011). Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), 2239–2252. https://doi.org/10.1093/molbev/msr048

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Research*, *8*(3), 163–167. https://doi.org/10.1101/gr.8.3.163

Endrödi, S. (1969). Monographie der Dynastinae 4. Tribus : Pentodontini (Coleoptera, Lamellicornia). *Entomologische Abhandlungen*, *87*, 1–145. https://ci.nii.ac.jp/naid/20001034599/

Endrődi, S., & Others. (1985). *The Dynastinae of the world*. Dr. W. Junk. https://www.cabdirect.org/cabdirect/abstract/19860536087

Evans, S. N., Shvets, Y., & Slatkin, M. (2007). Non-equilibrium theory of the allele frequency spectrum. *Theoretical Population Biology*, *71*(1), 109–119. https://doi.org/10.1016/j.tpb.2006.06.005

Exposito-Alonso, M., Becker, C., Schuenemann, V. J., Reiter, E., Setzer, C., Slovak, R., Brachi, B., Hagmann, J., Grimm, D. G., Chen, J., Busch, W., Bergelson, J., Ness, R. W., Krause, J., Burbano, H. A., & Weigel, D. (2018). The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genetics*, *14*(2), e1007155. https://doi.org/10.1371/journal.pgen.1007155

Exposito-Alonso, M., Drost, H.-G., Burbano, H. A., & Weigel, D. (2020). The Earth BioGenome project: opportunities and challenges for plant genomics and conservation. *The Plant Journal: For Cell and Molecular Biology*, *102*(2), 222–229. https://doi.org/10.1111/tpj.14631

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, *78*(2), 737–756. https://www.ncbi.nlm.nih.gov/pubmed/4448362

Ferretti, L., Raineri, E., & Ramos-Onsins, S. (2012). Neutrality tests for sequences with missing data. *Genetics*, *191*(4), 1397–1401. https://doi.org/10.1534/genetics.112.139949

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., & Gurr, S. J. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature*, *484*(7393), 186–194. https://doi.org/10.1038/nature10947

Fisher, R. A. (1931). XVII.—The Distribution of Gene Ratios for Rare Mutations. *Proceedings of the Royal Society of Edinburgh*, *50*, 204–219. https://doi.org/10.1017/S0370164600044886

Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., Ronchitelli, A. M., Martini, F., Cremonesi, R. G., Svoboda, J., Bauer, P., Caramelli, D., Castellano, S., Reich, D., Pääbo, S., & Krause, J. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes. *Current Biology: CB*, *23*(7), 553–559. https://doi.org/10.1016/j.cub.2013.02.044

Fu, X., Li, J., Tian, Y., Quan, W., Zhang, S., Liu, Q., Liang, F., Zhu, X., Zhang, L., Wang, D., & Hu, J. (2017). Long-read sequence assembly of the firefly Pyrocoelia pectoralis genome. *GigaScience*, *6*(12), 1–7. https://doi.org/10.1093/gigascience/gix112

Gansauge, M.-T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A., & Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Research*, *45*(10), e79. https://doi.org/10.1093/nar/gkx033

Gansauge, M.-T., & Meyer, M. (2013). Single-stranded DNA library preparation

for the sequencing of ancient or damaged DNA. *Nature Protocols*, *8*(4), 737–748. https://doi.org/10.1038/nprot.2013.038

Gillespie, J. H. (2000). Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics*, *155*(2), 909–919. https://www.ncbi.nlm.nih.gov/pubmed/10835409

Gladieux, P., Condon, B., Ravel, S., Soanes, D., Maciel, J. L. N., Nhani, A., Jr, Chen, L., Terauchi, R., Lebrun, M.-H., Tharreau, D., Mitchell, T., Pedley, K. F., Valent, B., Talbot, N. J., Farman, M., & Fournier, E. (2018). Gene Flow between Divergent Cereal- and Grass-Specific Lineages of the Rice Blast Fungus Magnaporthe oryzae. *mBio*, *9*(1). https://doi.org/10.1128/mBio.01219-17

Gladieux, P., Ravel, S., Rieux, A., Cros-Arteil, S., Adreit, H., Milazzo, J., Thierry, M., Fournier, E., Terauchi, R., & Tharreau, D. (2018). Coexistence of Multiple Endemic and Pandemic Lineages of the Rice Blast Pathogen. *mBio*, *9*(2). https://doi.org/10.1128/mBio.01806-17

Goldstein, P. Z., & Desalle, R. (2003). Calibrating phylogenetic species formation in a threatened insect using DNA from historical specimens. *Molecular Ecology*, *12*(7), 1993–1998. https://doi.org/10.1046/j.1365-294x.2003.01860.x

Goss, E. M., Larsen, M., Chastagner, G. A., Givens, D. R., & Grünwald, N. J. (2009). Population genetic analysis infers migration pathways of Phytophthora ramorum in US nurseries. *PLoS Pathogens*, *5*(9), e1000583. https://doi.org/10.1371/journal.ppat.1000583

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Grandaubert, J., Dutheil, J. Y., & Stukenbrock, E. H. (2019). The genomic determinants of adaptive evolution in a fungal pathogen. *Evolution Letters*,

3(3), 299–312. https://doi.org/10.1002/evl3.117

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., … Pääbo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722. https://doi.org/10.1126/science.1188021

Green, R. E., & Shapiro, B. (2013). Human evolution: turning back the clock. *Current Biology: CB*, *23*(7), R286–R288. https://doi.org/10.1016/j.cub.2013.02.050

Griffiths, R. C. (1981). Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, *19*(2), 169–186. https://doi.org/10.1016/0040-5809(81)90016-2

Gutaker, R. M., & Burbano, H. A. (2016). Reinforcing plant evolutionary genomics using ancient DNA. *Current Opinion in Plant Biology*. Submitted.

Gutaker, R. M., Reiter, E., Furtwängler, A., Schuenemann, V. J., & Burbano, H. A. (2017). Extraction of ultrashort DNA molecules from herbarium specimens. *BioTechniques*, *62*(2), 76–79. https://doi.org/10.2144/000114517

Hammond, P. M. (1992). *Species inventory. In "Global Biodiversity: Status of the Earth"s Living Resources. A Report Compiled by the World Conservation Monitoring Centre'.(Ed. B. Groombridge.) pp. 17--39*. Chapman & Hall: London.

Hanghøj, K., Seguin, A., Schubert, M., Madsen, T., Pedersen, J. S., Willerslev, E., & Orlando, L. (2016). Fast, accurate and automatic ancient nucleosome and methylation maps with epiPALEOMIX. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msw184

Heled, J., & Drummond, A. J. (2012). Calibrated tree priors for relaxed phylogenetics and divergence time estimation. *Systematic Biology*, *61*(1), 138–149. https://doi.org/10.1093/sysbio/syr087

Heliconius Genome Consortium. (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, *487*(7405), 94–98.

https://doi.org/10.1038/nature11041

Hershberg, R., Tang, H., & Petrov, D. A. (2007). Reduced selection leads to accelerated gene loss in Shigella. *Genome Biology*, *8*(8), R164. https://doi.org/10.1186/gb-2007-8-8-r164

Heyn, P., Stenzel, U., Briggs, A. W., Kircher, M., Hofreiter, M., & Meyer, M. (2010). Road blocks on paleogenomes--polymerase extension profiling reveals the frequency of blocking lesions in ancient DNA. *Nucleic Acids Research*, *38*(16), e161. https://doi.org/10.1093/nar/gkq572

Hofreiter, M., Jaenicke, V., Serre, D., von Haeseler, A., & Pääbo, S. (2001). DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. *Nucleic Acids Research*, *29*(23), 4793–4799. https://doi.org/10.1093/nar/29.23.4793

Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M., & Pääbo, S. (2001). Ancient DNA. *Nature Reviews. Genetics*, *2*(5), 353–359. https://doi.org/10.1038/35072071

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature Reviews. Genetics*, *10*(9), 639–650. https://doi.org/10.1038/nrg2611

Höss, M., Jaruga, P., Zastawny, T. H., Dizdaroglu, M., & Pääbo, S. (1996). DNA damage and DNA sequence retrieval from ancient tissues. *Nucleic Acids Research*, *24*(7), 1304–1307. https://doi.org/10.1093/nar/24.7.1304

Ho, S. Y. W., Lanfear, R., Bromham, L., Phillips, M. J., Soubrier, J., Rodrigo, A. G., & Cooper, A. (2011). Time-dependent rates of molecular evolution. *Molecular Ecology*, *20*(15), 3087–3101. https://doi.org/10.1111/j.1365-294X.2011.05178.x

Huang, J., Si, W., Deng, Q., Li, P., & Yang, S. (2014). Rapid evolution of avirulence genes in rice blast fungus Magnaporthe oryzae. *BMC Genetics*, *15*, 45. https://doi.org/10.1186/1471-2156-15-45

Hubbard, A., Lewis, C. M., Yoshida, K., Ramirez-Gonzalez, R. H., de Vallavieille-Pope, C., Thomas, J., Kamoun, S., Bayles, R., Uauy, C., &

Saunders, D. G. O. (2015). Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology*, *16*, 23. https://doi.org/10.1186/s13059-015-0590-8

Hudson, R. R., & Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, *111*(1), 147–164. https://www.ncbi.nlm.nih.gov/pubmed/4029609

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O. S., Wild, R., Hammond, P. M., Ahrens, D., Balke, M., Caterino, M. S., Gómez-Zurita, J., Ribera, I., Barraclough, T. G., Bocakova, M., Bocak, L., & Vogler, A. P. (2007). A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science*, *318*(5858), 1913–1916. https://doi.org/10.1126/science.1146954

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), 254–267. https://doi.org/10.1093/molbev/msj030

Hu, Y., Linz, D. M., & Moczek, A. P. (2019). Beetle horns evolved from wing serial homologs. *Science*, *366*(6468), 1004–1007. https://doi.org/10.1126/science.aaw2980

Inoue, Y., Vy, T. T. P., Yoshida, K., Asano, H., Mitsuoka, C., Asuke, S., Anh, V. L., Cumagun, C. J. R., Chuma, I., Terauchi, R., Kato, K., Mitchell, T., Valent, B., Farman, M., & Tosa, Y. (2017). Evolution of the wheat blast fungus through functional losses in a host specificity determinant. *Science*, *357*(6346), 80–83. https://doi.org/10.1126/science.aam9654

Islam, M. T., Croll, D., Gladieux, P., Soanes, D. M., Persoons, A., Bhattacharjee, P., Hossain, M. S., Gupta, D. R., Rahman, M. M., Mahboob, M. G., Cook, N., Salam, M. U., Surovy, M. Z., Sancho, V. B., Maciel, J. L. N., NhaniJúnior, A., Castroagudín, V. L., Reges, J. T. de A., Ceresini, P. C., … Kamoun, S. (2016). Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. *BMC Biology*, *14*(1), 84. https://doi.org/10.1186/s12915-016-0309-7

Islam, M. T., Kim, K.-H., & Choi, J. (2019). Wheat Blast in Bangladesh: The Current Situation and Future Impacts. *The Plant Pathology Journal / the*

Korean Society of Plant Pathology, *35*(1), 1–10. https://doi.org/10.5423/PPJ.RW.08.2018.0168

Jin, H., Yonezawa, T., Zhong, Y., Kishino, H., & Hasegawa, M. (2016). Cretaceous origin of giant rhinoceros beetles (Dynastini; Coleoptera) and correlation of their evolution with the Pangean breakup. *Genes & Genetic Systems*, *91*(4), 209–215. https://doi.org/10.1266/ggs.16-00003

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, New York, NY. https://doi.org/10.1007/b98835

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94. https://doi.org/10.1186/1471-2156-11-94

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* , *29*(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Karasov, T. L., Almario, J., Friedemann, C., Ding, W., Giolai, M., Heavens, D., Kersten, S., Lundberg, D. S., Neumann, M., Regalado, J., Neher, R. A., Kemen, E., & Weigel, D. (2018). Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host & Microbe*, *24*(1), 168–179.e4. https://doi.org/10.1016/j.chom.2018.06.011

Kelkar, Y. D., & Ochman, H. (2012). Causes and consequences of genome expansion in fungi. *Genome Biology and Evolution*, *4*(1), 13–23. https://doi.org/10.1093/gbe/evr124

Kharouba, H. M., Lewthwaite, J. M. M., Guralnick, R., Kerr, J. T., & Vellend, M. (2018). Using insect natural history collections to study global change impacts: challenges and opportunities. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *374*(1763). https://doi.org/10.1098/rstb.2017.0405

Kircher, M. (2012). Analysis of High-Throughput Ancient DNA Sequencing Data. In B. Shapiro & M. Hofreiter (Eds.), *Ancient DNA* (pp. 197–228).

Humana Press. https://doi.org/10.1007/978-1-61779-516-9_23

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536. https://doi.org/10.1002/bies.200900181

Kircher, M., Sawyer, S., & Meyer, M. (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, *40*(1), e3. https://doi.org/10.1093/nar/gkr771

Kistler, L. (2012). Ancient DNA extraction from plants. *Methods in Molecular Biology* , *840*, 71–79. https://doi.org/10.1007/978-1-61779-516-9_10

Koop, B. F., Goodman, M., Xu, P., Chan, K., & Slightom, J. L. (1986). Primate eta-globin DNA sequences and man's place among the great apes. *Nature*, *319*(6050), 234–238. https://doi.org/10.1038/319234a0

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btz305

Langner, T., Harant, A., Gomez-Luciano, L. B., Shrestha, R. K., Win, J., & Kamoun, S. (2020). Genomic rearrangements generate hypervariable mini-chromosomes in host-specific lineages of the blast fungus. In *bioRxiv*.

Lang, P. L. M., Willems, F. M., Scheepens, J. F., Burbano, H. A., & Bossdorf, O. (2019). Using herbaria to study global environmental change. *The New Phytologist*, *221*(1), 110–122. https://doi.org/10.1111/nph.15401

Latorre, S. M., Reyes-Avila, C. S., Malmgren, A., Win, J., Kamoun, S., & Burbano, H. A. (2020). *Dataset and Scripts for: Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus*. https://doi.org/10.5281/zenodo.3893626

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E.,

Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), 4325–4333. https://doi.org/10.1073/pnas.1720115115

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. In *arXiv [q-bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, *362*(6422), 709–715. https://doi.org/10.1038/362709a0

Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B., & Sperens, B. (1977). DNA N-glycosidases: properties of uracil-DNA glycosidase from Escherichia coli. *The Journal of Biological Chemistry*, *252*(10), 3286–3294. https://www.ncbi.nlm.nih.gov/pubmed/324994

Lindner, M., Maroschek, M., Netherer, S., Kremer, A., Barbati, A., Garcia-Gonzalo, J., Seidl, R., Delzon, S., Corona, P., Kolström, M., Lexer, M. J., & Marchetti, M. (2010). Climate change impacts, adaptive capacity, and vulnerability of European forest ecosystems. *Forest Ecology and Management*, *259*(4), 698–709. https://doi.org/10.1016/j.foreco.2009.09.023

Lovmar, L., Ahlford, A., Jonsson, M., & Syvänen, A.-C. (2005). Silhouette scores for assessment of SNP genotype clusters. *BMC Genomics*, *6*, 35. https://doi.org/10.1186/1471-2164-6-35

Lynch, M. (2010). Evolution of the mutation rate. *Trends in Genetics: TIG*, *26*(8), 345–352. https://doi.org/10.1016/j.tig.2010.05.003

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science*,

302(5649), 1401–1404. https://doi.org/10.1126/science.1089370

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, *46*(3), 523–536. https://doi.org/10.1093/sysbio/46.3.523

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2012). *Cluster: Cluster Analysis Basics and Extensions*. *1*. http://dx.doi.org/

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKenna, D. D. (2018). Beetle genomes in the 21st century: prospects, progress and priorities. *Current Opinion in Insect Science*, *25*, 76–82. https://doi.org/10.1016/j.cois.2017.12.002

McKenna, D. D., Shin, S., Ahrens, D., Balke, M., Beza-Beza, C., Clarke, D. J., Donath, A., Escalona, H. E., Friedrich, F., Letsch, H., Liu, S., Maddison, D., Mayer, C., Misof, B., Murin, P. J., Niehuis, O., Peters, R. S., Podsiadlowski, L., Pohl, H., … Beutel, R. G. (2019). The evolution and genomic basis of beetle diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(49), 24729–24737. https://doi.org/10.1073/pnas.1909655116

McLaughlin, J. F., Hellmann, J. J., Boggs, C. L., & Ehrlich, P. R. (2002). Climate change hastens population extinctions. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(9), 6070–6074. https://doi.org/10.1073/pnas.052131199

Metzker, M. L. (2005). Emerging technologies in DNA sequencing. *Genome Research*, *15*(12), 1767–1776. https://doi.org/10.1101/gr.3770505

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Meyer, J. M., Markov, G. V., Baskaran, P., Herrmann, M., Sommer, R. J., & Rödelsperger, C. (2016). Draft Genome of the Scarab Beetle Oryctes borbonicus on La Réunion Island. *Genome Biology and Evolution*, *8*(7),

2093–2105. https://doi.org/10.1093/gbe/evw133

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), db.prot5448. https://doi.org/10.1101/pdb.prot5448

Mikheyev, A. S., Tin, M. M. Y., Arora, J., & Seeley, T. D. (2015). Museum samples reveal rapid evolution by wild honey bees exposed to a novel parasite. *Nature Communications*, *6*, 7991. https://doi.org/10.1038/ncomms8991

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics: TIG*, *17*(10), 589–596. https://doi.org/10.1016/s0168-9525(01)02447-7

Miyamoto, M. M., Koop, B. F., Slightom, J. L., Goodman, M., & Tennant, M. R. (1988). Molecular systematics of higher primates: genealogical relations and classification. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(20), 7627–7631. https://doi.org/10.1073/pnas.85.20.7627

Mohd-Assaad, N., McDonald, B. A., & Croll, D. (2019). The emergence of the multi-species NIP1 effector in Rhynchosporium was accompanied by high rates of gene duplications and losses. *Environmental Microbiology*, *21*(8), 2677–2695. https://doi.org/10.1111/1462-2920.14583

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, *403*(6772), 853–858. https://doi.org/10.1038/35002501

Myers, S. R., & Griffiths, R. C. (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics*, *163*(1), 375–394. https://www.ncbi.nlm.nih.gov/pubmed/12586723

Nations, U., & United Nations. (2019). World Population Prospects 2019: Highlights. In *Statistical Papers - United Nations (Ser. A), Population and Vital Statistics Report*. https://doi.org/10.18356/13bf5476-en

Neher, R. A. (2013). Genetic Draft, Selective Interference, and Population Genetics of Rapid Adaptation. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 195–215.

https://doi.org/10.1146/annurev-ecolsys-110512-135920

Nei, M., & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273. https://doi.org/10.1073/pnas.76.10.5269

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature*, *246*(5428), 96–98. https://doi.org/10.1038/246096a0

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(25), 16134–16137. https://doi.org/10.1073/pnas.252626899

Pääbo, S., Poinar, H., Serre, D., Jaenicke-Despres, V., Hebler, J., Rohland, N., Kuch, M., Krause, J., Vigilant, L., & Hofreiter, M. (2004). Genetic analyses from ancient DNA. *Annual Review of Genetics*, *38*, 645–679. https://doi.org/10.1146/annurev.genet.37.110801.143214

Papadopoulou, A., Anastasiou, I., & Vogler, A. P. (2010). Revisiting the insect mitochondrial molecular clock: the mid-Aegean trench calibration. *Molecular Biology and Evolution*, *27*(7), 1659–1672. https://doi.org/10.1093/molbev/msq051

Pardo-Diaz, C., Salazar, C., Baxter, S. W., Merot, C., Figueiredo-Ready, W., Joron, M., McMillan, W. O., & Jiggins, C. D. (2012). Adaptive introgression across species boundaries in Heliconius butterflies. *PLoS Genetics*, *8*(6), e1002752. https://doi.org/10.1371/journal.pgen.1002752

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Paulian, R. (1985). Les coléoptères Scarabaeidae canthonines de

Nouvelle-Guinée. *Annales de La Société Entomologique de France*, *21*, 219–238.

Pease, J. B., Haak, D. C., Hahn, M. W., & Moyle, L. C. (2016). Phylogenomics Reveals Three Sources of Adaptive Variation during a Rapid Radiation. *PLoS Biology*, *14*(2), e1002379. https://doi.org/10.1371/journal.pbio.1002379

Pease, J. B., & Hahn, M. W. (2015). Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, *64*(4), 651–662. https://doi.org/10.1093/sysbio/syv023

Pečnerová, P., Díez-Del-Molino, D., Dussex, N., Feuerborn, T., von Seth, J., van der Plicht, J., Nikolskiy, P., Tikhonov, A., Vartanyan, S., & Dalén, L. (2017). Genome-Based Sexing Provides Clues about Behavior and Social Structure in the Woolly Mammoth. *Current Biology: CB*, *27*(22), 3505–3510.e3. https://doi.org/10.1016/j.cub.2017.09.064

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1201.0490

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., & Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. *Genome Biology*, *17*, 60. https://doi.org/10.1186/s13059-016-0918-z

Peng, Z., Oliveira-Garcia, E., Lin, G., Hu, Y., Dalby, M., Migeon, P., Tang, H., Farman, M., Cook, D., White, F. F., Valent, B., & Liu, S. (2019). Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. *PLoS Genetics*, *15*(9), e1008272. https://doi.org/10.1371/journal.pgen.1008272

Petit-Houdenot, Y., Langner, T., Harant, A., Win, J., & Kamoun, S. (2019). *A clone resource of Magnaporthe oryzae effectors that share sequence and structural similarities across host-specific lineages*. https://doi.org/10.5281/zenodo.3268775

Philippe, H., Delsuc, F., Brinkmann, H., & Lartillot, N. (2005). Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 541–562.

https://doi.org/10.1146/annurev.ecolsys.35.112202.130205

Poelchau, M., Childers, C., Moore, G., Tsavatapalli, V., Evans, J., Lee, C.-Y., Lin, H., Lin, J.-W., & Hackett, K. (2015). The i5k Workspace@NAL--enabling genomic data access, visualization and curation of arthropod genomes. *Nucleic Acids Research*, *43*(Database issue), D714–D719. https://doi.org/10.1093/nar/gku983

Poinar, H. N., Hofreiter, M., Spaulding, W. G., Martin, P. S., Stankiewicz, B. A., Bland, H., Evershed, R. P., Possnert, G., & Pääbo, S. (1998). Molecular coproscopy: dung and diet of the extinct ground sloth Nothrotheriops shastensis. *Science*, *281*(5375), 402–406. https://doi.org/10.1126/science.281.5375.402

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://www.ncbi.nlm.nih.gov/pubmed/10835412

Prüfer, K. (2018). snpAD: an ancient DNA genotype caller. *Bioinformatics* , *34*(24), 4165–4171. https://doi.org/10.1093/bioinformatics/bty507

Prüfer, K., Stenzel, U., Hofreiter, M., Pääbo, S., Kelso, J., & Green, R. E. (2010). Computational challenges in the analysis of ancient DNA. *Genome Biology*, *11*(5), R47. https://doi.org/10.1186/gb-2010-11-5-r47

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Radhakrishnan, G. V., Cook, N. M., Bueno-Sancho, V., Lewis, C. M., Persoons, A., Mitiku, A. D., Heaton, M., Davey, P. E., Abeyo, B., Alemayehu, Y., Badebo, A., Barnett, M., Bryant, R., Chatelain, J., Chen, X., Dong, S., Henriksson, T., Holdgate, S., Justesen, A. F., … Saunders, D. G. O. (2019). MARPLE, a point-of-care, strain-level disease diagnostics and surveillance tool for complex fungal pathogens. *BMC Biology*, *17*(1), 65. https://doi.org/10.1186/s12915-019-0684-y

Raffaele, S., & Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews. Microbiology*, *10*(6), 417–430. https://doi.org/10.1038/nrmicro2790

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., Rasmussen, S., Stafford, T. W., Jr, Orlando, L., Metspalu, E., Karmin, M., Tambets, K., Rootsi, S., Mägi, R., Campos, P. F., Balanovska, E., Balanovsky, O., Khusnutdinova, E., Litvinov, S., … Willerslev, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, *505*(7481), 87–91. https://doi.org/10.1038/nature12736

Rambaut, A. (2000). Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics* , *16*(4), 395–399. https://doi.org/10.1093/bioinformatics/16.4.395

Ratcliffe, B. C., Cave, R. D., & Cano, E. B. (2013). *The dynastine scarab beetles of Mexico, Guatemala, and Belize (Coleoptera: Scarabaeidae: Dynastinae)*. http://www.entomologica.es/index.php?d=secciones&sec=14&n=33

R Core Team. (2018). *R: A language and environment for statistical computing*. https://www.R-project.org/

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–494. https://doi.org/10.1038/nature08365

Renaud, G., Hanghøj, K., Willeslev, E., & Orlando, L. (2016). gargammel: a sequence simulator for ancient DNA. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btw670

Rietman, H., Bijsterbosch, G., Cano, L. M., Lee, H.-R., Vossen, J. H., Jacobsen, E., Visser, R. G. F., Kamoun, S., & Vleeshouwers, V. G. A. A. (2012). Qualitative and quantitative late blight resistance in the potato cultivar Sarpo Mira is determined by the perception of five distinct RXLR effectors. *Molecular Plant-Microbe Interactions: MPMI*, *25*(7), 910–919. https://doi.org/10.1094/MPMI-01-12-0010-R

Rieux, A., & Balloux, F. (2016). Inferences from tip-calibrated phylogenies: a

review and a practical guide. *Molecular Ecology*, *25*(9), 1911–1924. https://doi.org/10.1111/mec.13586

Rödelsperger, C., Röseler, W., Prabh, N., Yoshida, K., Weiler, C., Herrmann, M., & Sommer, R. J. (2018). Phylotranscriptomics of Pristionchus Nematodes Reveals Parallel Gene Loss in Six Hermaphroditic Lineages. *Current Biology: CB*, *28*(19), 3123–3127.e5. https://doi.org/10.1016/j.cub.2018.07.041

Rogers, S. O., & Bendich, A. J. (1985). Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Molecular Biology*, *5*(2), 69–76. https://doi.org/10.1007/BF00020088

Rohland, N., & Hofreiter, M. (2007). Comparison and optimization of ancient DNA extraction. *BioTechniques*, *42*(3), 343–352. http://www.ncbi.nlm.nih.gov/pubmed/17390541

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*(6960), 798–804. https://doi.org/10.1038/nature02053

Ronquist, F., Lartillot, N., & Phillips, M. J. (2016). Closing the gap between rocks and clocks using total-evidence dating. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *371*(1699). https://doi.org/10.1098/rstb.2015.0136

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*(5602), 2381–2385. https://doi.org/10.1126/science.1078311

Ross-Ibarra, J. (2009). *RminCutter* (Version 1.05) [Computer software]. https://www.plantsciences.ucdavis.edu/faculty/ross-ibarra/code/files/rmin.html

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Saleh, D., Milazzo, J., Adreit, H., Fournier, E., & Tharreau, D. (2014). South-East

Asia is the center of origin, diversity and dispersion of the rice blast fungus, Magnaporthe oryzae. *The New Phytologist*, *201*(4), 1440–1456. https://doi.org/10.1111/nph.12627

Saleh, D., Xu, P., Shen, Y., Li, C., Adreit, H., Milazzo, J., Ravigné, V., Bazin, E., Nottéghem, J.-L., Fournier, E., & Tharreau, D. (2012). Sex at the origin: an Asian population of the rice blast fungus Magnaporthe oryzae reproduces sexually. *Molecular Ecology*, *21*(6), 1330–1344. https://doi.org/10.1111/j.1365-294X.2012.05469.x

Saunders, D. G. O., Pretorius, Z. A., & Hovmøller, M. S. (2019). Tackling the re-emergence of wheat stem rust in Western Europe. *Communications Biology*, *2*, 51. https://doi.org/10.1038/s42003-019-0294-9

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, *3*(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Scholtz, C. H. (2000). Evolution of flightlessness in Scarabaeoidea (Insecta, Coleoptera). *Deutsche Entomologische Zeitschrift*, *47*(1), 5–28. https://doi.org/10.1002/dez.200000003

Schoville, S. D., Chen, Y. H., Andersson, M. N., Benoit, J. B., Bhandari, A., Bowsher, J. H., Brevik, K., Cappelle, K., Chen, M.-J. M., Childers, A. K., Childers, C., Christiaens, O., Clements, J., Didion, E. M., Elpidina, E. N., Engsontia, P., Friedrich, M., García-Robles, I., Gibbs, R. A., … Richards, S. (2018). A model species for agricultural pest genomics: the genome of the Colorado potato beetle, Leptinotarsa decemlineata (Coleoptera: Chrysomelidae). *Scientific Reports*, *8*(1), 1931. https://doi.org/10.1038/s41598-018-20154-1

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., & Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols*, *9*(5), 1056–1082. https://doi.org/10.1038/nprot.2014.063

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., Krogh, A., & Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, *13*, 178. https://doi.org/10.1186/1471-2164-13-178

Schubert, M., Lindgreen, S., & Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes*, *9*, 88. https://doi.org/10.1186/s13104-016-1900-2

Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., Pybus, O. G., Gilbert, M. T. P., Barnes, I., Binladen, J., Willerslev, E., Hansen, A. J., Baryshnikov, G. F., Burns, J. A., Davydov, S., Driver, J. C., Froese, D. G., Harington, C. R., Keddie, G., … Cooper, A. (2004). Rise and fall of the Beringian steppe bison. *Science*, *306*(5701), 1561–1565. https://doi.org/10.1126/science.1101074

Shapiro, B., & Hofreiter, M. (2014). A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, *343*(6169), 1236573. https://doi.org/10.1126/science.1236573

Shapiro, B., Sibthorpe, D., Rambaut, A., Austin, J., Wragg, G. M., Bininda-Emonds, O. R. P., Lee, P. L. M., & Cooper, A. (2002). Flight of the dodo. *Science*, *295*(5560), 1683. https://doi.org/10.1126/science.295.5560.1683

Sharp, P. M., & Li, W. H. (1989). On the rate of DNA sequence evolution in Drosophila. *Journal of Molecular Evolution*, *28*(5), 398–402. https://doi.org/10.1007/bf02603075

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N., & Reich, D. (2015). Genetic evidence for two founding populations of the Americas. *Nature*, *525*(7567), 104–108. https://doi.org/10.1038/nature14895

Skoglund, P., Northoff, B. H., Shunkov, M. V., Derevianko, A. P., Pääbo, S.,

Krause, J., & Jakobsson, M. (2014). Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(6), 2229–2234. https://doi.org/10.1073/pnas.1318934111

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. https://doi.org/10.1186/1471-2105-6-31

Soanes, D., Ryder, L. S., Islam, M. T., & Talbot, N. J. (2017). *Genome assemblies of Magnaporthe oryzae isolated from Bangladesh in 2016 and 2017*. https://doi.org/10.6084/m9.figshare.5236381.v1

Sommer, R. J., & McGaughran, A. (2013). The nematode Pristionchus pacificus as a model system for integrative studies in evolutionary biology. *Molecular Ecology*, *22*(9), 2380–2393. https://doi.org/10.1111/mec.12286

Speidel, L., Forest, M., Shi, S., & Myers, S. R. (2019). A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, *51*(9), 1321–1329. https://doi.org/10.1038/s41588-019-0484-x

Stein, L. D. (2011). An introduction to the informatics of "next-generation" sequencing. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *Chapter 11*, Unit 11.1. https://doi.org/10.1002/0471250953.bil1101s36

Stiller, M., Baryshnikov, G., Bocherens, H., Grandal d'Anglade, A., Hilpert, B., Münzel, S. C., Pinhasi, R., Rabeder, G., Rosendahl, W., Trinkaus, E., Hofreiter, M., & Knapp, M. (2010). Withering away--25,000 years of genetic decline preceded cave bear extinction. *Molecular Biology and Evolution*, *27*(5), 975–978. https://doi.org/10.1093/molbev/msq083

Swofford, D. L. (2002). *PAUP: phylogenetic analysis using parsimony, version 4.0 b10*. Sinauer Associates, Sunderland, MA.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595. https://www.ncbi.nlm.nih.gov/pubmed/2513255

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, *28*(4), 289–301. https://doi.org/10.1002/gepi.20064

Tegelberg, R., Mononen, T., & Saarenmaa, H. (2014). High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. *Taxon*, *63*(6), 1307–1313. https://onlinelibrary.wiley.com/doi/abs/10.12705/636.13

Terauchi, R., & Yoshida, K. (2010). Towards population genomics of effector-effector target interactions: Research review. *The New Phytologist*, *187*(4), 929–939. https://doi.org/10.1111/j.1469-8137.2010.03408.x

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., De Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Peterson, A. T., Phillips, O. L., & Williams, S. E. (2004). Extinction risk from climate change. *Nature*, *427*(6970), 145–148. https://doi.org/10.1038/nature02121

Thomsen, P. F., Elias, S., Gilbert, M. T. P., Haile, J., Munch, K., Kuzmina, S., Froese, D. G., Sher, A., Holdaway, R. N., & Willerslev, E. (2009). Non-destructive sampling of ancient insect DNA. *PloS One*, *4*(4), e5048. https://doi.org/10.1371/journal.pone.0005048

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. *Molecular Ecology*, *25*(10), 2144–2164. https://doi.org/10.1111/mec.13606

Vattathil, S., & Akey, J. M. (2015). Small Amounts of Archaic Admixture Provide Big Insights into Human History. *Cell*, *163*(2), 281–284. https://doi.org/10.1016/j.cell.2015.09.042

Vega, F. E., Brown, S. M., Chen, H., Shen, E., Nair, M. B., Ceja-Navarro, J. A., Brodie, E. L., Infante, F., Dowd, P. F., & Pain, A. (2015). Draft genome of the most devastating insect pest of coffee worldwide: the coffee berry borer, Hypothenemus hampei. *Scientific Reports*, *5*, 12525. https://doi.org/10.1038/srep12525

Vernot, B., & Akey, J. M. (2014). Resurrecting surviving Neandertal lineages

from modern human genomes. *Science*, *343*(6174), 1017–1021. https://doi.org/10.1126/science.1245938

Veron, S., Mouchet, M., Govaerts, R., Haevermans, T., & Pellens, R. (2019). Vulnerability to climate change of islands worldwide and its impact on the tree of life. *Scientific Reports*, *9*(1), 14471. https://doi.org/10.1038/s41598-019-51107-x

Vleeshouwers, V. G. A. A., & Oliver, R. P. (2014). Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. *Molecular Plant-Microbe Interactions: MPMI*, *27*(3), 196–206. https://doi.org/10.1094/MPMI-10-13-0313-IA

Vleeshouwers, V. G. A. A., Rietman, H., Krenek, P., Champouret, N., Young, C., Oh, S.-K., Wang, M., Bouwmeester, K., Vosman, B., Visser, R. G. F., Jacobsen, E., Govers, F., Kamoun, S., & Van der Vossen, E. A. G. (2008). Effector genomics accelerates discovery and functional profiling of potato disease resistance and phytophthora infestans avirulence genes. *PloS One*, *3*(8), e2875. https://doi.org/10.1371/journal.pone.0002875

Wahlund, S. (1928). Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. *Hereditas*, *11*(1), 65–106. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1601-5223.1928.tb02483.x

Waldvogel, A.-M., Feldmeyer, B., Rolshausen, G., Exposito-Alonso, M., Rellstab, C., Kofler, R., Mock, T., Schmid, K., Schmitt, I., Bataillon, T., Savolainen, O., Bergland, A., Flatt, T., Guillaume, F., & Pfenninger, M. (2020). Evolutionary genomics can improve prediction of species' responses to climate change. *Evolution Letters*, *4*(1), 4–18. https://doi.org/10.1002/evl3.154

Wang, K., Li, P., Gao, Y., Liu, C., Wang, Q., Yin, J., Zhang, J., Geng, L., & Shu, C. (2019). De novo genome assembly of the white-spotted flower chafer (Protaetia brevitarsis). *GigaScience*, *8*(4). https://doi.org/10.1093/gigascience/giz019

Warren, R., Price, J., Graham, E., Forstenhaeusler, N., & VanDerWal, J. (2018). The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5°C rather than 2°C. *Science*, *360*(6390), 791–795.

https://doi.org/10.1126/science.aar3646

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276. https://doi.org/10.1016/0040-5809(75)90020-9

Watts, P. C., Thompson, D. J., Allen, K. A., & Kemp, S. J. (2007). How useful is DNA extracted from the legs of archived insects for microsatellite-based population genetic analyses? *Journal of Insect Conservation*, *11*(2), 195–198. https://doi.org/10.1007/s10841-006-9024-y

Weir, B. S., & Cockerham, C. C. (1984). ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution; International Journal of Organic Evolution*, *38*(6), 1358–1370. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767. https://doi.org/10.1101/gr.214874.116

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2018). Corrigendum: Direct determination of diploid genome sequences. *Genome Research*, *28*(4), 606.1. https://doi.org/10.1101/gr.235812.118

Weiß, C. L. (2019). *bsh-denovo*. Github. https://github.com/clwgg/bsh-denovo

Weiß, C. L., Dannemann, M., Prüfer, K., & Burbano, H. A. (2015). Contesting the presence of wheat in the British Isles 8,000 years ago by assessing ancient DNA authenticity from low-coverage data. *eLife*, *4*. https://doi.org/10.7554/eLife.10005

Wilson, A. C., & Sarich, V. M. (1969). A molecular time scale for human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *63*(4), 1088–1093. https://doi.org/10.1073/pnas.63.4.1088

Wright, S. (1931). Evolution in Mendelian Populations. *Genetics*, *16*(2), 97–159. https://www.ncbi.nlm.nih.gov/pubmed/17246615

Wright, S. (1949). THE GENETICAL STRUCTURE OF POPULATIONS. *Annals*

of Eugenics, *15*(1), 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Wu, Y.-M., Li, J., & Chen, X.-S. (2018). Draft genomes of two blister beetles Hycleus cichorii and Hycleus phaleratus. *GigaScience*, *7*(3), 1–7. https://doi.org/10.1093/gigascience/giy006

Xue, M., Yang, J., Li, Z., Hu, S., Yao, N., Dean, R. A., Zhao, W., Shen, M., Zhang, H., Li, C., Liu, L., Cao, L., Xu, X., Xing, Y., Hsiang, T., Zhang, Z., Xu, J.-R., & Peng, Y.-L. (2012). Comparative analysis of the genomes of two field isolates of the rice blast fungus Magnaporthe oryzae. *PLoS Genetics*, *8*(8), e1002869. https://doi.org/10.1371/journal.pgen.1002869

Yoshida, K., Burbano, H. A., Krause, J., Thines, M., Weigel, D., & Kamoun, S. (2014). Mining herbaria for plant pathogen genomes: back to the future. *PLoS Pathogens*, *10*(4), e1004028. https://doi.org/10.1371/journal.ppat.1004028

Yoshida, K., Saitoh, H., Fujisawa, S., Kanzaki, H., Matsumura, H., Yoshida, K., Tosa, Y., Chuma, I., Takano, Y., Win, J., Kamoun, S., & Terauchi, R. (2009). Association genetics reveals three novel avirulence genes from the rice blast fungal pathogen Magnaporthe oryzae. *The Plant Cell*, *21*(5), 1573–1591. https://doi.org/10.1105/tpc.109.066324

Yoshida, K., Saunders, D. G. O., Mitsuoka, C., Natsume, S., Kosugi, S., Saitoh, H., Inoue, Y., Chuma, I., Tosa, Y., Cano, L. M., Kamoun, S., & Terauchi, R. (2016). Host specialization of the blast fungus Magnaporthe oryzae is associated with dynamic gain and loss of genes linked to transposable elements. *BMC Genomics*, *17*, 370. https://doi.org/10.1186/s12864-016-2690-6

Zhang, C., & Wang, M. (2019). Bayesian tip dating reveals heterogeneous morphological clocks in Mesozoic birds. *Royal Society Open Science*, *6*(7), 182062. https://doi.org/10.1098/rsos.182062

Zhang, J., Chiodini, R., Badr, A., & Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics = Yi Chuan Xue Bao*, *38*(3), 95–109. https://doi.org/10.1016/j.jgg.2011.02.003

Zhang, J., Cong, Q., Shen, J., Opler, P. A., & Grishin, N. V. (2019). Genomics of a complete butterfly continent. In *bioRxiv* (p. 829887).

https://doi.org/10.1101/829887

Zhang, W., Dasmahapatra, K. K., Mallet, J., Moreira, G. R. P., & Kronforst, M. R. (2016). Genome-wide introgression among distantly related Heliconius butterfly species. *Genome Biology*, *17*, 25. https://doi.org/10.1186/s13059-016-0889-0

Zhong, Z., Chen, M., Lin, L., Han, Y., Bao, J., Tang, W., Lin, L., Lin, Y., Somai, R., Lu, L., Zhang, W., Chen, J., Hong, Y., Chen, X., Wang, B., Shen, W.-C., Lu, G., Norvienyeku, J., Ebbole, D. J., & Wang, Z. (2018). Population genomic analysis of the rice blast fungus reveals specific events associated with expansion of three main clades. *The ISME Journal*. https://doi.org/10.1038/s41396-018-0100-6

Publication List

- P.M.M. Lang, C.L. Weiß, S. Kersten, **S.M. Latorre**, S. Nagel, B. Nickel, M. Meyer, H.A. Burbano. (2020). Hybridization ddRAD-sequencing for population genomics of nonmodel plants using highly degraded historical specimen DNA. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13168

- **S.M. Latorre**, C.S. Reyes-Avila, A. Malmgren, J. Win, S. Kamoun & H.A. Burbano. (2020). Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus. *BMC Biology*. https://doi.org/10.1186/s12915-020-00818-z

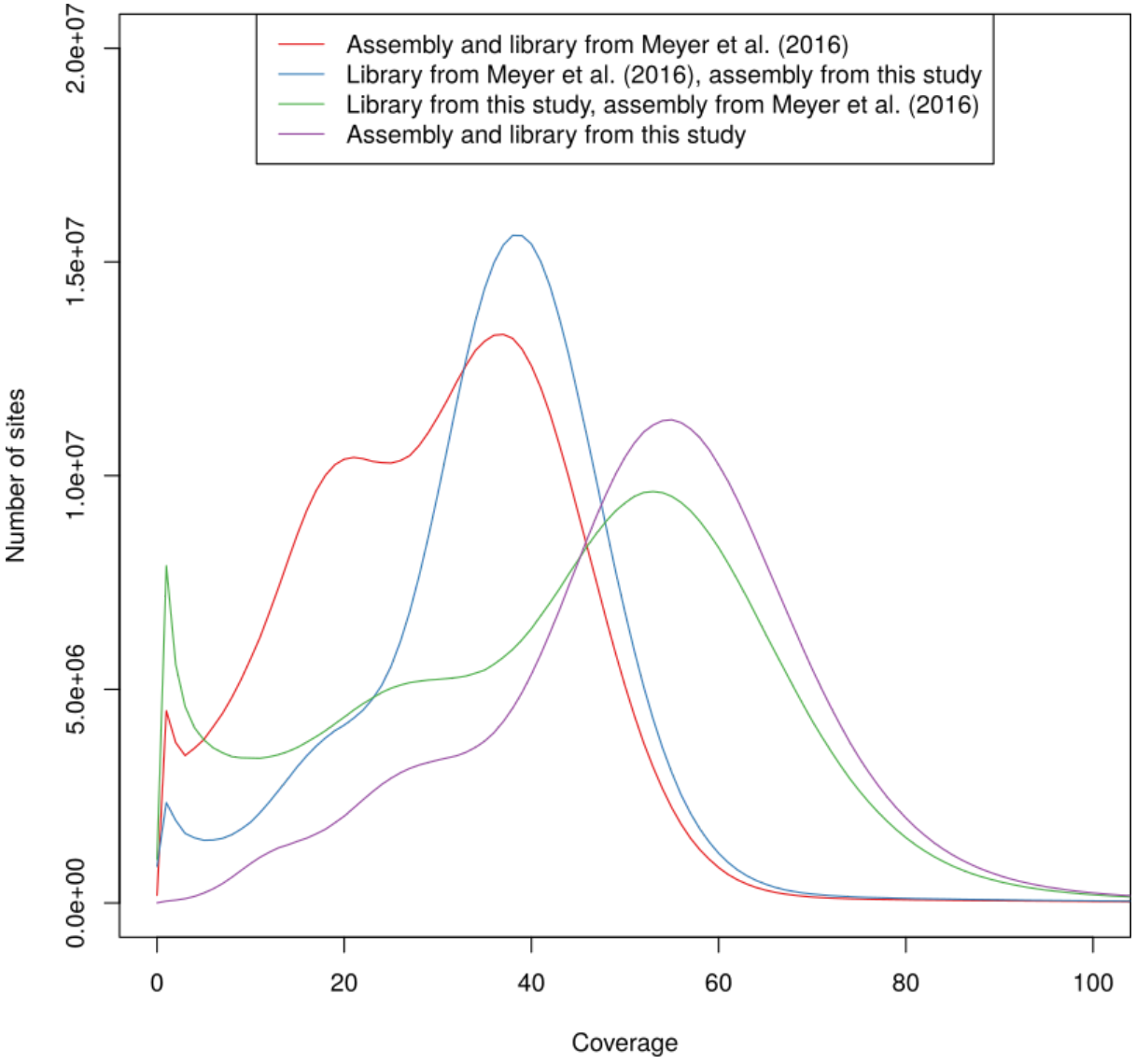
- **S.M. Latorre**, M. Herrmann, M.J. Paulsen, C. Rödelsperger, A. Dréau, W. Röseler, R.J. Sommer & H.A. Burbano. (2020). Museum phylogenomics of extinct *Oryctes* beetles from the Mascarene Islands. *bioRxiv*. https://doi.org/10.1101/2020.02.19.954339

Abbreviations

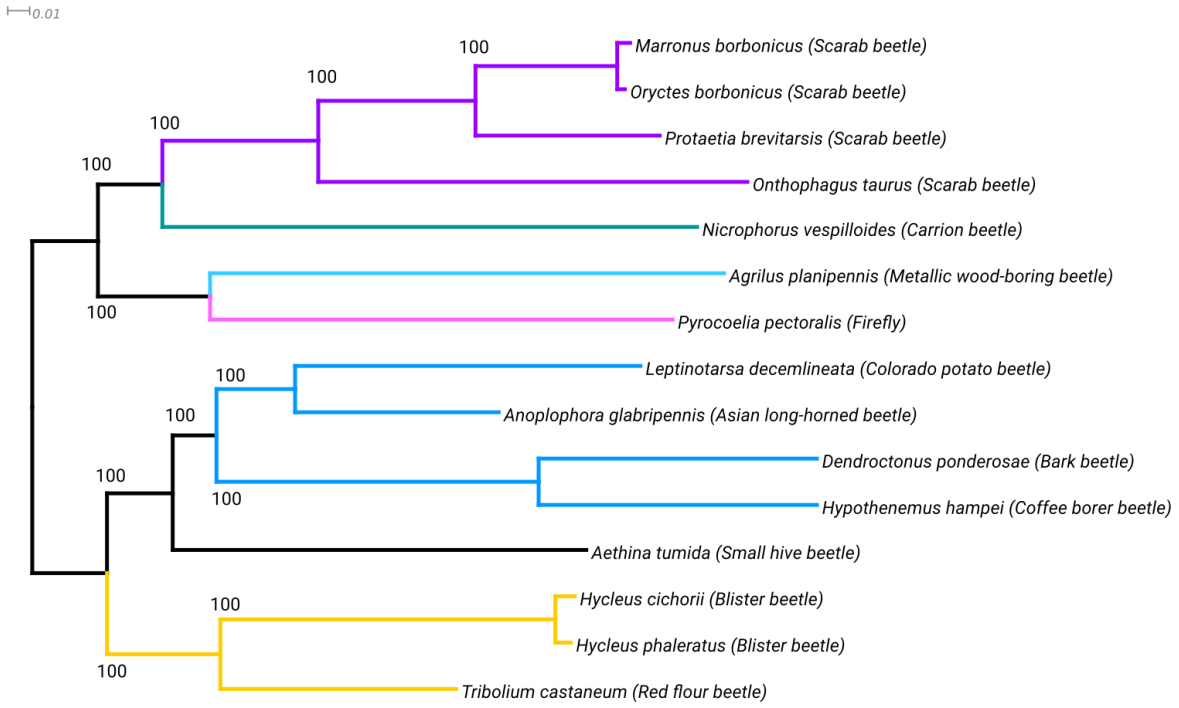
A, G, T, C, U	Adenine, Guanine, Cytosine, Uracil
DNA	Deoxyribonucleic acid
aDNA	ancient DNA
UDG	Uracil-DNA glycosylase
PTB	N-phenacylthiazolium bromide
PCR	Polymerase Chain Reaction
WGS	Whole Genome Sequencing
ENA	European Nucleotide Archive
SAM	Sequence Alignment Map
BAM	Binary Alignment Map
VCF	Variant Call Format
SNP	Single Nucleotide Polymorphism
MAF	Minimum Allele Frequency
PCA	Principal Component Analysis
MDS	Multidimensional Scaling
F_{ST}	Fixation Index
SFS	Site Frequency Spectrum
ILS	Incomplete Lineage Sorting
HPD	Highest Posterior Density

Supplementary Material

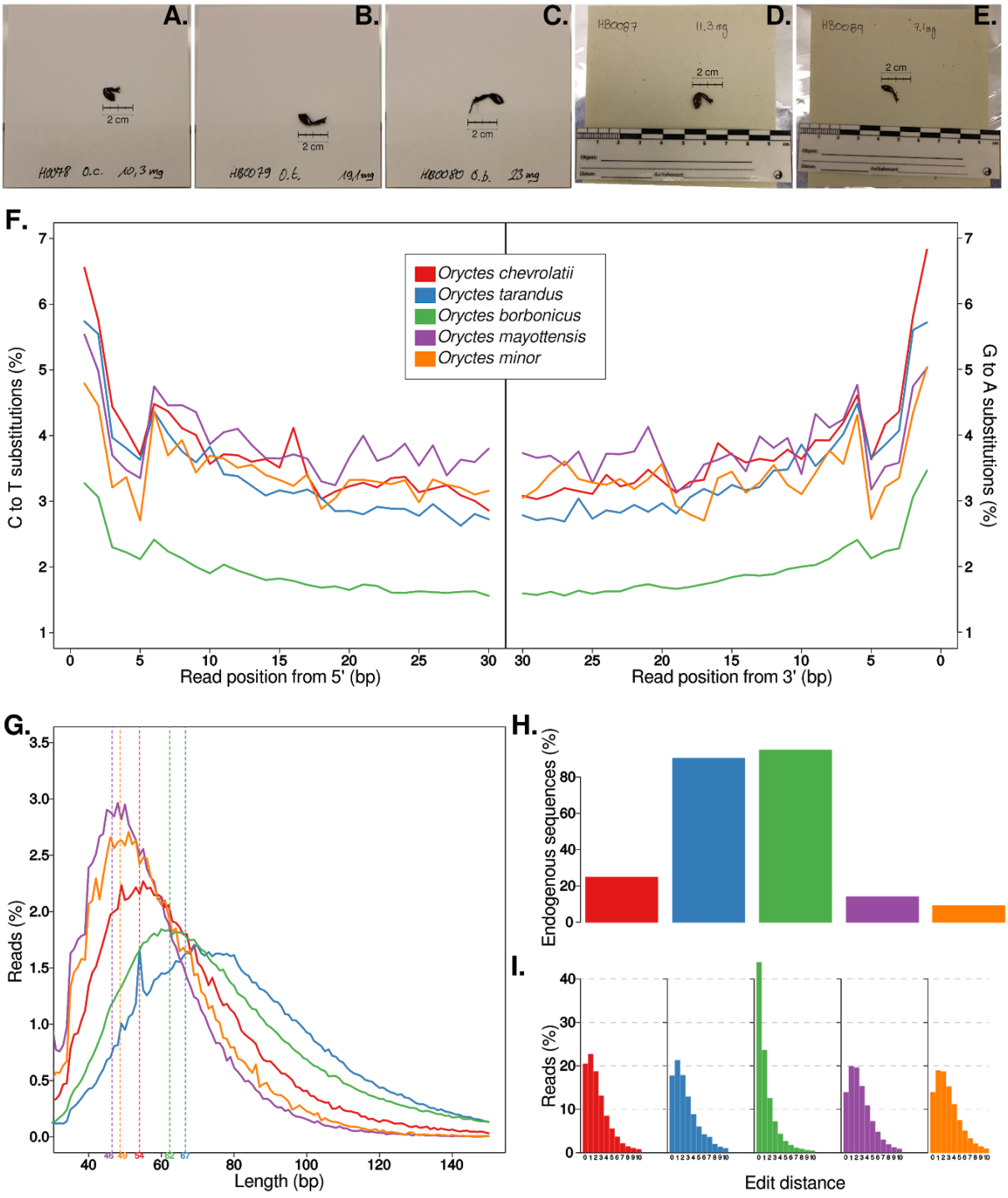
Supplementary Material for Chapter 2



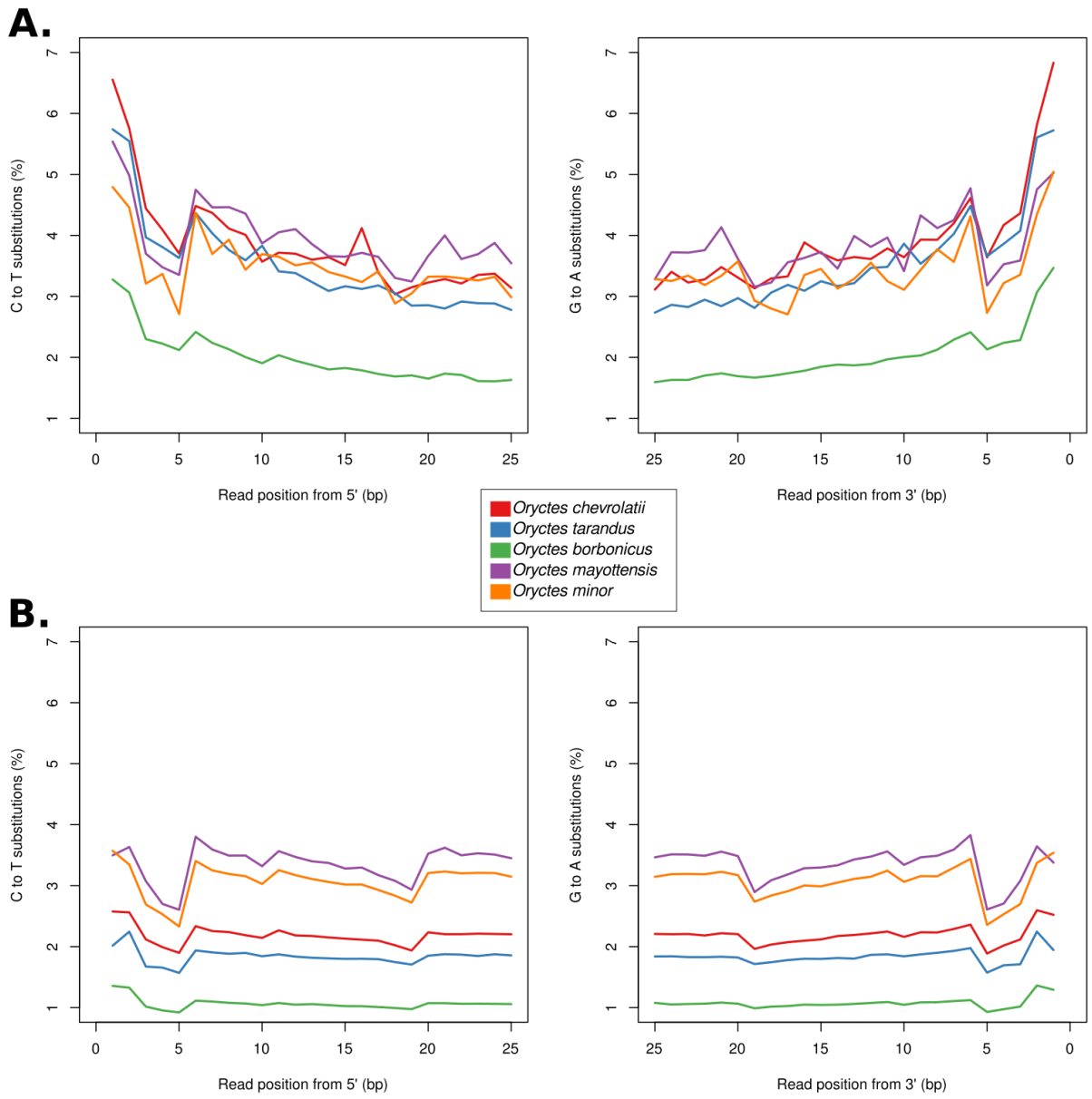
Supplementary Figure 2.1. Coverage analysis of different *Oryctes borbonicus* assemblies. Raw reads from two different whole genome sequencing libraries were aligned to the current and the previously published *O. borbonicus* assembly (J. M. Meyer et al., 2016). For both data sets, the previously published assembly shows a more pronounced peak at half of the expected coverage pointing at the potential problem of allelism in the old assembly.



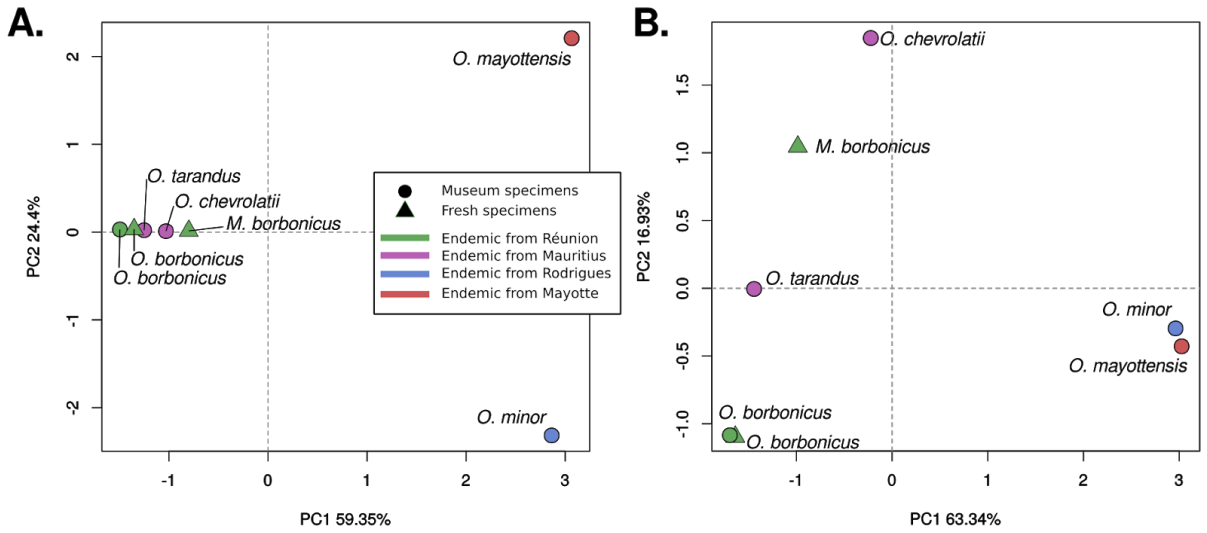
Supplementary Figure 2.2. Phylogenetic tree based on genome data from 15 beetle species. Protein sequences from 363 orthologous genes were concatenated. A Maximum likelihood tree was calculated based on the resulting alignment of 107,398 amino acids (100 bootstrap pseudoreplicates). Subtree coloring was chosen for easier comparison with the phylogeny by McKenna et al. (D. D. McKenna et al., 2019). The two focal species, *Oryctes borbonicus* and *Marronus borbonicus*, display similar levels of divergence as two species of the same genera, *Hycleus cichorii* and *H. phaleratus*.



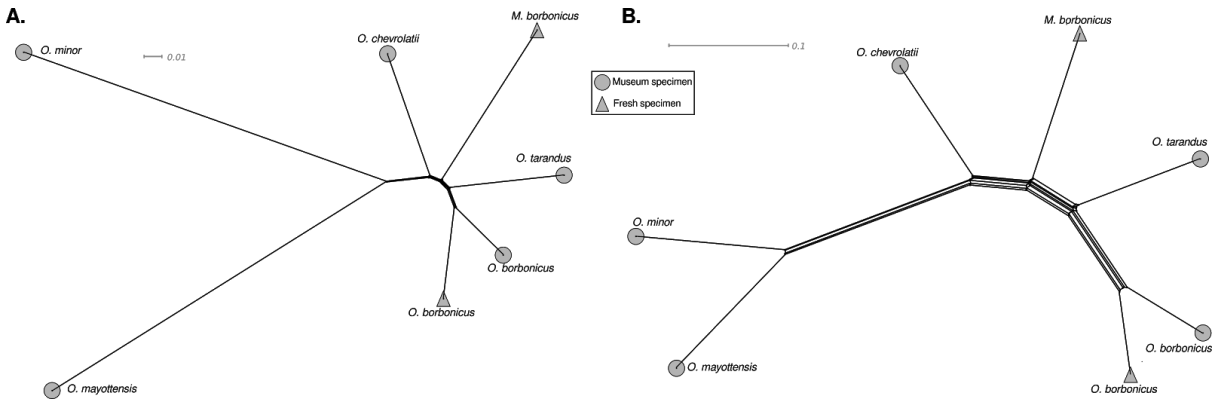
Supplementary Figure 2.3. Ancient DNA characteristics and edit distances of museum specimens. **(A-E)** Pictures of samples used for the DNA extraction: **(A)** *Oryctes chevrolatii*, **(B)** *O. tarandus*, **(C)** *O. borbonicus*, **(D)** *O. mayottensis*, **(E)** *O. minor*. **(F)** Cytosine to Thymine and Guanine to Adenine substitutions at the 5'- and 3'-end, respectively. **(G)** Distribution of fragment lengths of merged reads. Dotted lines indicate the mean value of each distribution. **(H)** Percentage of merged reads that mapped to the *O. borbonicus* genome. **(I)** Distribution of edit distances of reads mapped to the *O. borbonicus* genome.



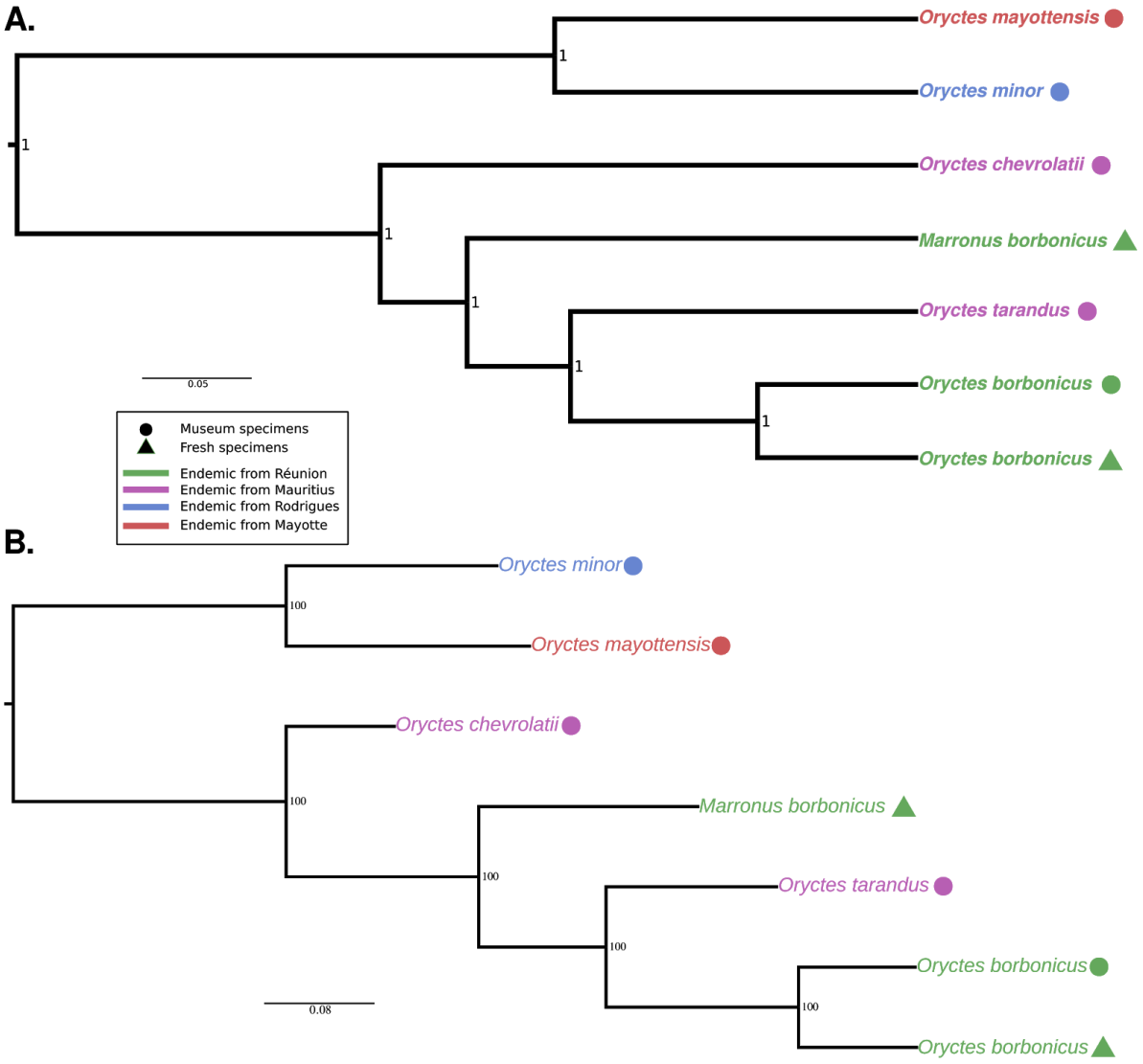
Supplementary Figure 2.4. Cytosine to Thymine and Guanine to Adenine substitutions at the 5'- and 3'-end, before and after uracil enzymatic library repair. (A) Described substitutions present in museum specimens before enzymatic repair (same as Supplementary Figure 2.3F). **(B)** Described Substitution after enzymatic library reparation.



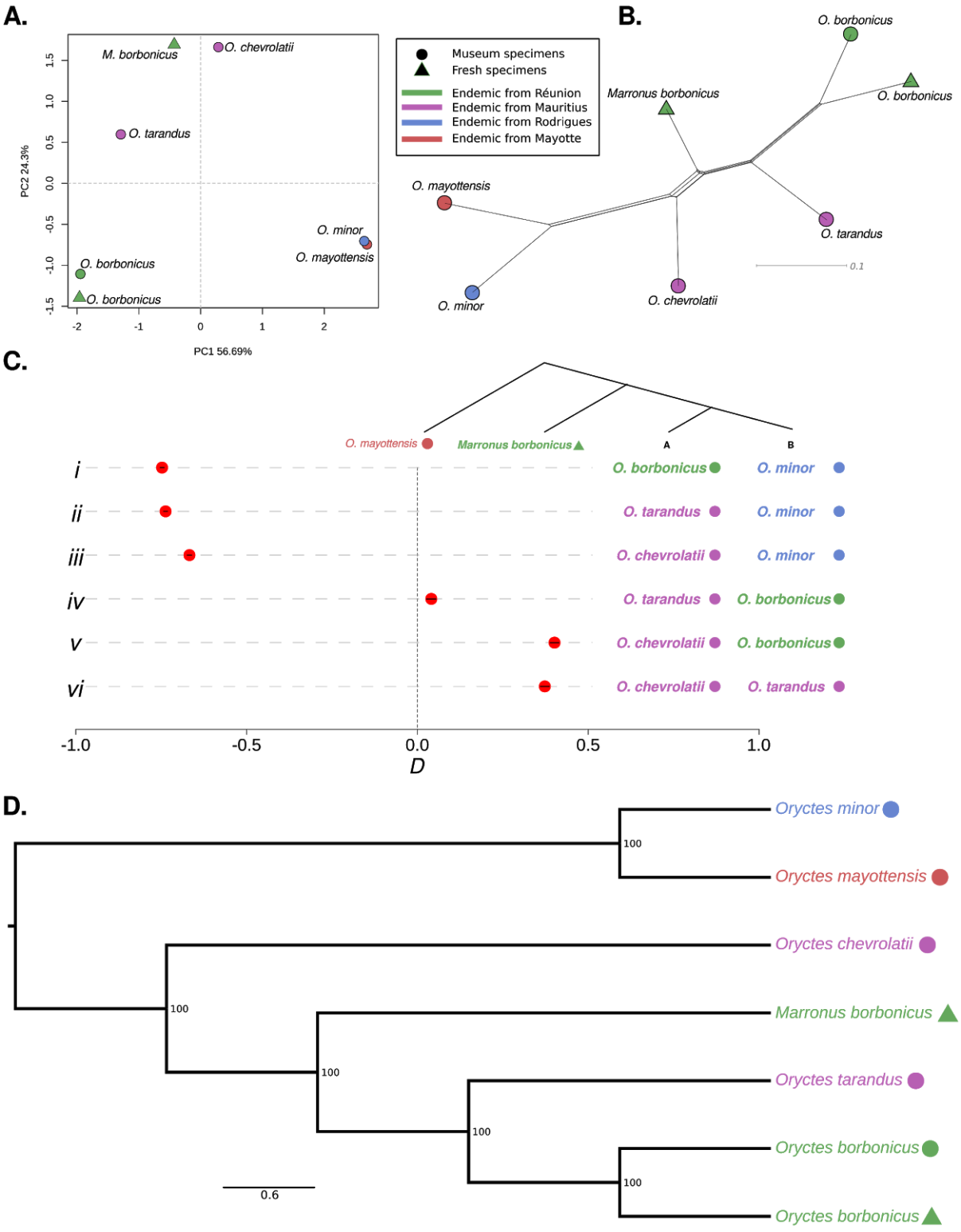
Supplementary Figure 2.5. Effect of Minimum Allele Frequency (MAF) filter on the genetic distances. Hamming distances-based PCAs show the effect of the MAF on the separation of the endemic beetles from Réunion and Mauritius. **(A)** MAF of 1/7 with 2,144,289 SNPs and **(B)** MAF of 2/7 with 304,417 SNPs.



Supplementary Figure 2.6. Effect of Minimum Allele Frequency (MAF) filter on the evolutionary relationships and branch lengths. The phylogenetic networks show a similar topology and relations between samples but different branch lengths. **(A)** MAF of 1/7 with 2,144,289 SNPs and **(B)** MAF of 2/7 with 304,417 SNPs.



Supplementary Figure 2.7. Concatenated Bayesian and Maximum Likelihood phylogenetic trees. (A) Bayesian phylogenetic reconstruction. Maximum clade credibility tree from 36,000 trees. Numbers at the nodes indicate posterior probability support. **(B)** Concatenated Maximum Likelihood phylogenetic tree. Numbers at nodes indicate bootstrap support (200 replicates). In both **A** and **B**, only sites with complete information were included, leaving 304,417 sites in the final dataset.



Supplementary Figure 2.8. Robustness of evolutionary relations among scarab beetles. The figures represent the same analyses shown in Figure 2.2. In order to discard the possible effect of ascertainment bias due to the selection of *Oryctes borbonicus* as reference genome, we repeated the analysis by mapping the raw reads to a *Marronus borbonicus* assembly. **(A)** Principal component analysis plot based on 330,529 SNPs. Genetic distances between beetle samples are projected onto

the first two PCs. Axis labels indicate the fraction of total variation explained by each PC. **(B)** Phylogenetic network based on 330,529 SNPs using the neighbor-net method. **(C)** Testing the robustness of phylogenetic relations among scarab beetle species using *D*-statistics of the type *D*(B,A; *Marronus borbonicus*, outgroup), as depicted in the phylogenetic tree. *O. mayottensis* was used as an outgroup. Each row (*i-vi*) shows a different *D*-statistic configuration. A negative *D*-statistic indicates that *M. borbonicus* is closer to species B, whereas a positive *D*-statistic indicates that *M. borbonicus* is closer to species A. The points depict the result of each *D*-statistic test and the lines their respective 95% confidence intervals. Rows *i-iii* show that *M. borbonicus* is closer to the *Oryctes* spp. from Réunion and Mauritius. Rows *v-vi* show that *M. borbonicus* is closer to both *O. borbonicus* and *O. tarandus* than to *O. chevrolatii*. Finally, row *iv* shows the closest *D*-statistic to zero, which indicates that *M. borbonicus* is slightly closer to *O. borbonicus* than to *O. tarandus*. **(D)** SVDquartets species tree. Numbers at nodes indicate bootstrap support (1000 replicates).

Supplementary Table 2.1. Overview of genome sequencing and assembly of extant beetle specimens.

Species	Ref	Number of Scaffolds	Genome size (assembled) [Mb]	N50 [Mb]	BUSCO Complete single copy [%]
O. borbonicus	(Meyer et al. 2016)	150,243	494.4	0.1	95.3
O. borbonicus	This study	9,526	411.2	8.4	95.9
M. borbonicus	This study	9,046	412.6	12.4	95.8

Supplementary Table 2.2. Samples information

ID	(M)useum / (F)resh	Species	Collection Year	Collection[1]	Collector	Non-UDG treated libraries[2]	UDG treated libraries[2]
HB0078	M	*Oryctes chevrolatii*	1963	NHMG	Rouillard	ERR3932721	ERR3932716
HB0079	M	*Oryctes tarandus*	1966	NHMG	Yves Gomy	ERR3932722	ERR3932717
HB0080	M	*Oryctes borbonicus*	1966	NHMG	Yves Gomy	ERR3932723	ERR3932718
HB0087	M	*Oryctes mayottensis*	2010	MPI	NA	ERR3932724	ERR3932719
HB0089	M	*Oryctes minor*	1918	NHML	Snell & Thomasset	ERR3932725	ERR3932720
-	F	*Oryctes borbonicus*	2017	MPI	Matthias Herrmann	ERR3685131 - ERR3685134	NA
-	F	*Marronus borbonicus*	2017	MPI	Matthias Herrmann	ERR3685135 - ERR3685138	NA

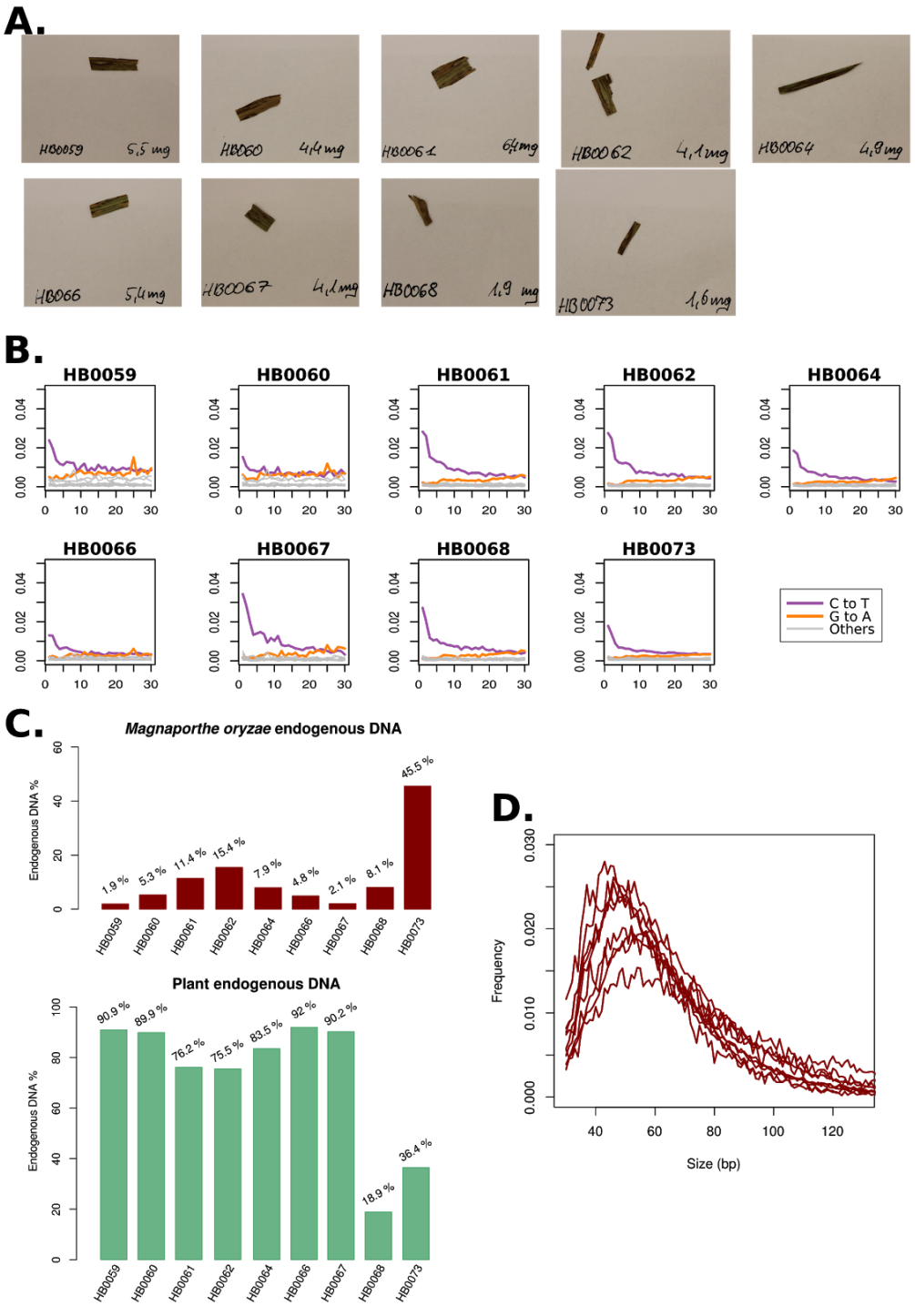
[1]NHMG (Natural History Museum Geneva); NHML (Natural History Museum London); MPI (Max Planck Institute for Developmental Biology, Tuebingen).

[2] ENA accession numbers.

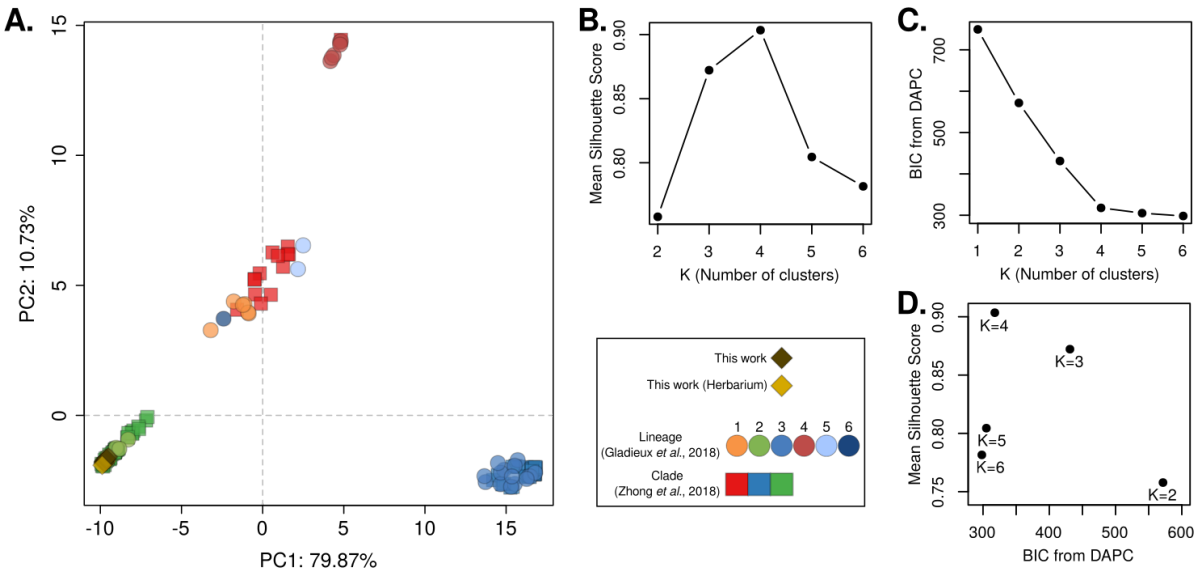
Supplementary Table 2.3. Average depth in museum specimens mapped to the *Oryctes borbonicus* draft genome

Sample	Average depth (X)
Oryctes borbonicus	1.16
Oryctes tarandus	1
Oryctes chevrolatii	0.79
Oryctes mayottensis	0.8
Oryctes minor	1.38

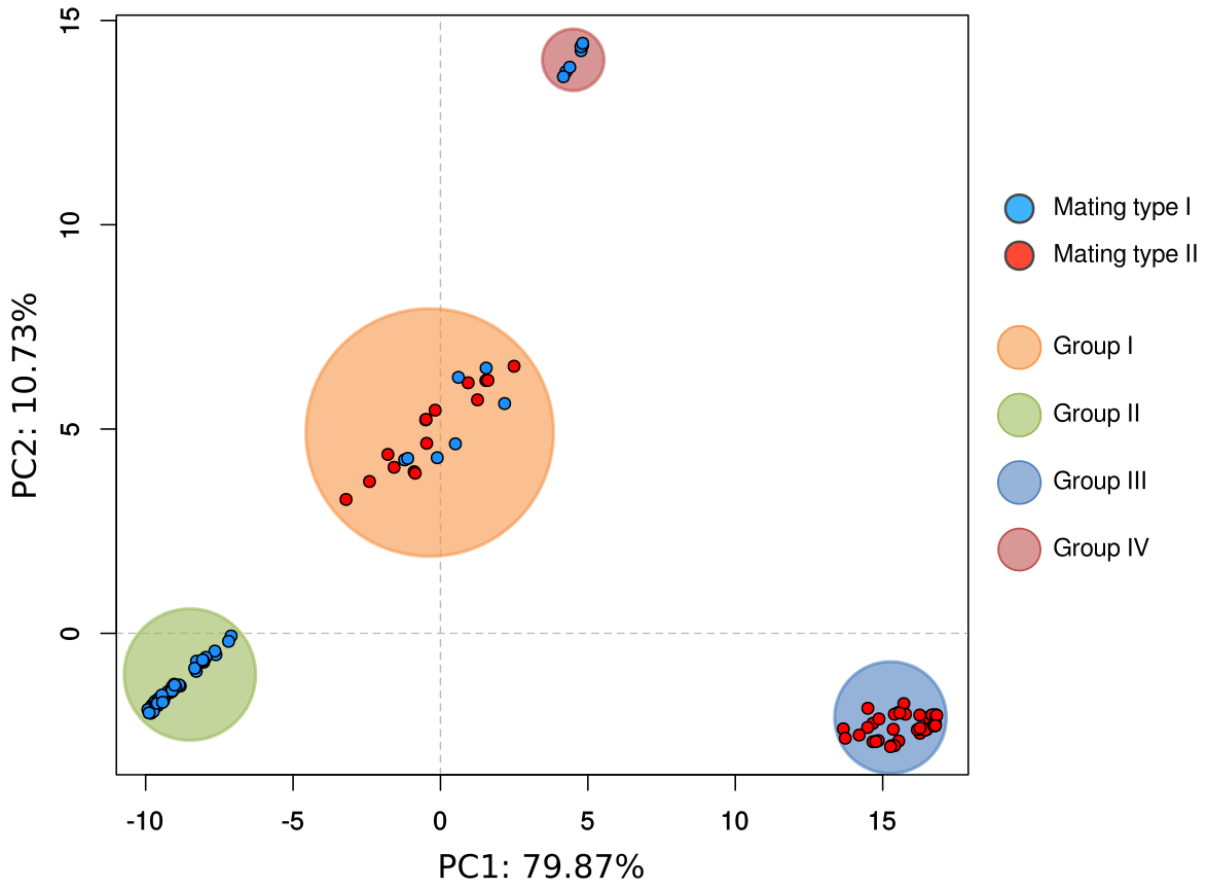
Supplementary Material for Chapter 3



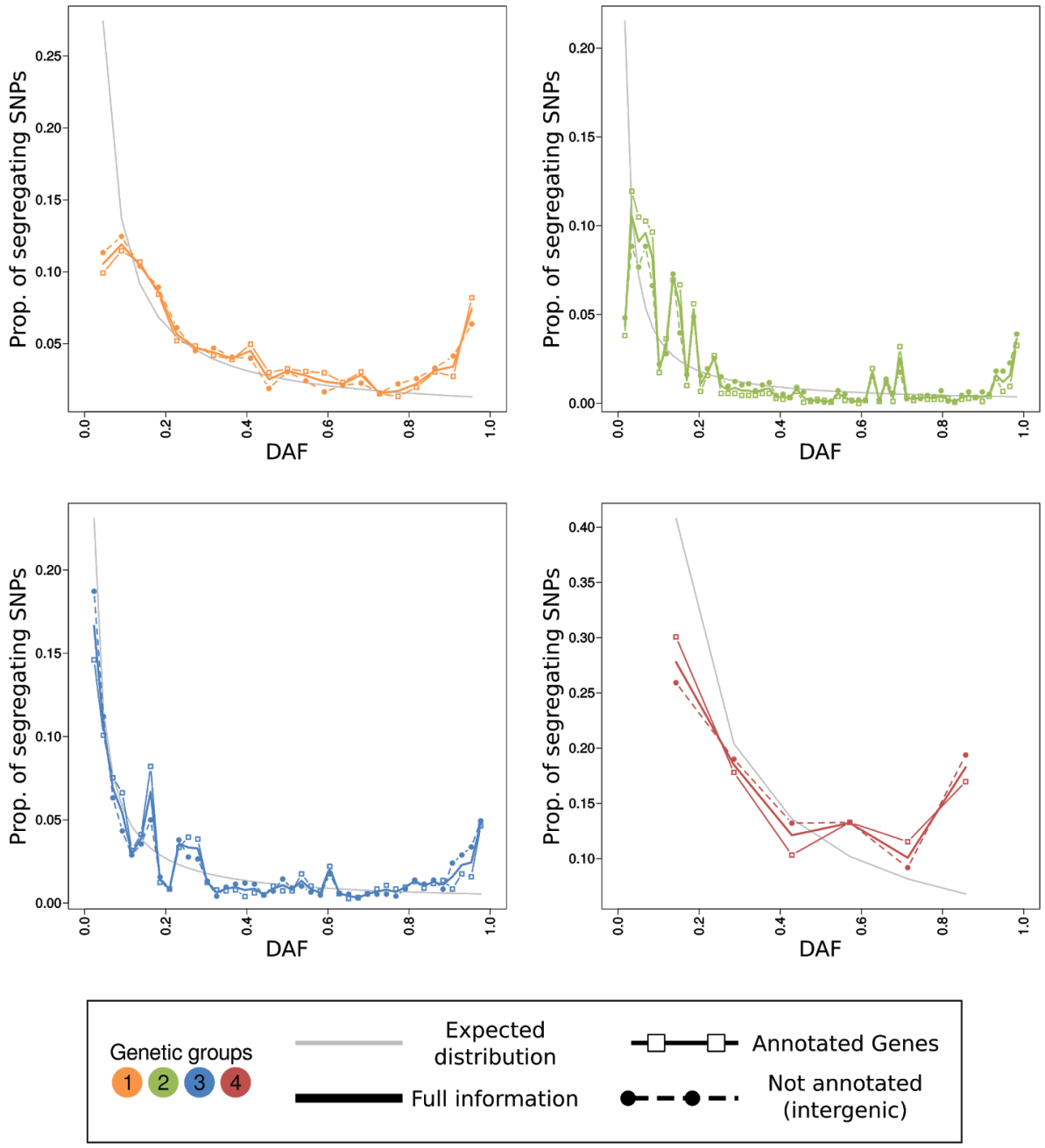
Supplementary Figure 3.1. Ancient DNA characteristics of herbarium isolates. **(A)** Pictures of samples used for the DNA extraction. **(B)** Cytosine to Thymine,Guanine to Adenine and other substitutions at the 5' end of the reads. **(C)** Percentage of reads aligned to either the host plant *Oryza sativa* or the rice blast *Magaporthe oryzae* reference genomes. **(D)** Distribution of fragment lengths of merged reads.



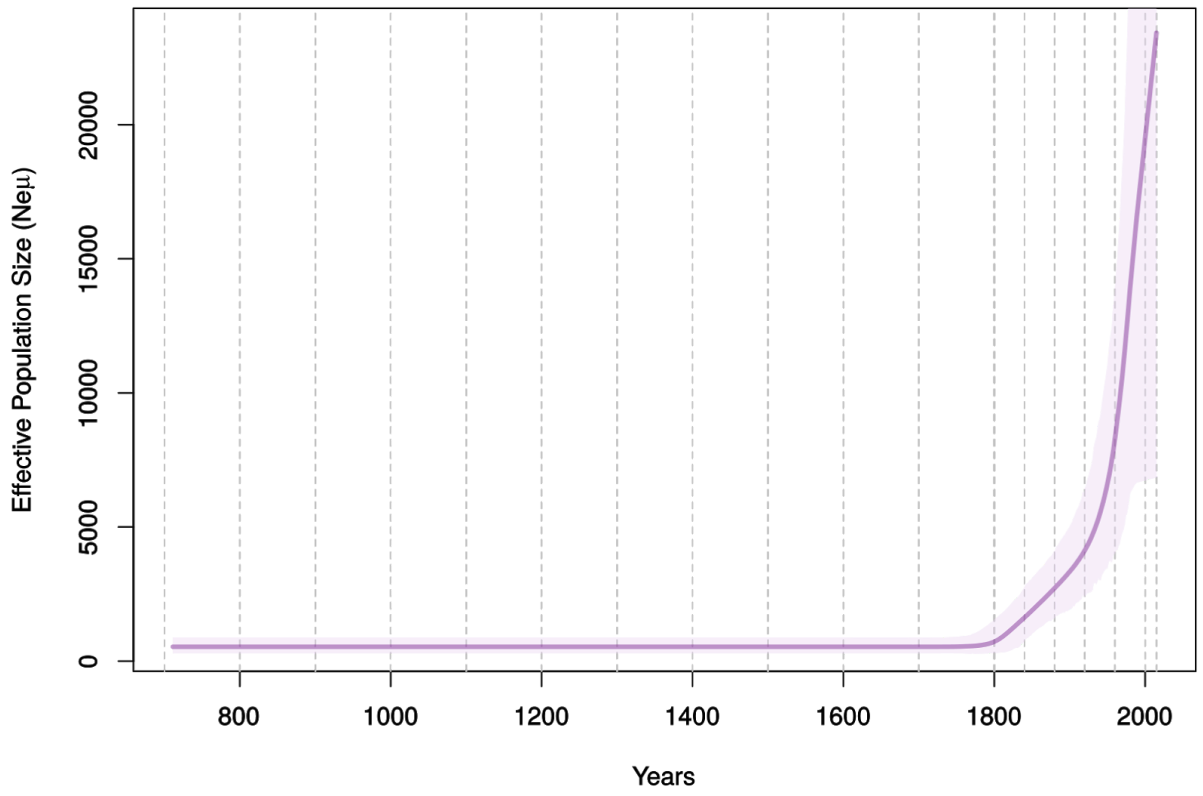
Supplementary Figure 3.2. Principal component analysis (PCA) reveals four defined groups. (A) PCA based on pairwise Hamming distances. **(B)** Silhouette score analysis shows best averages per group scores when *K*=4. **(C)** Discriminant Analysis of Principal Components (DAPC) shows stabilization of the Bayesian Information Criterion (BIC) when *K*=4. **(D)** BIC from DAPC versus mean silhouette scores shows an optimal number of groups when *K*=4.



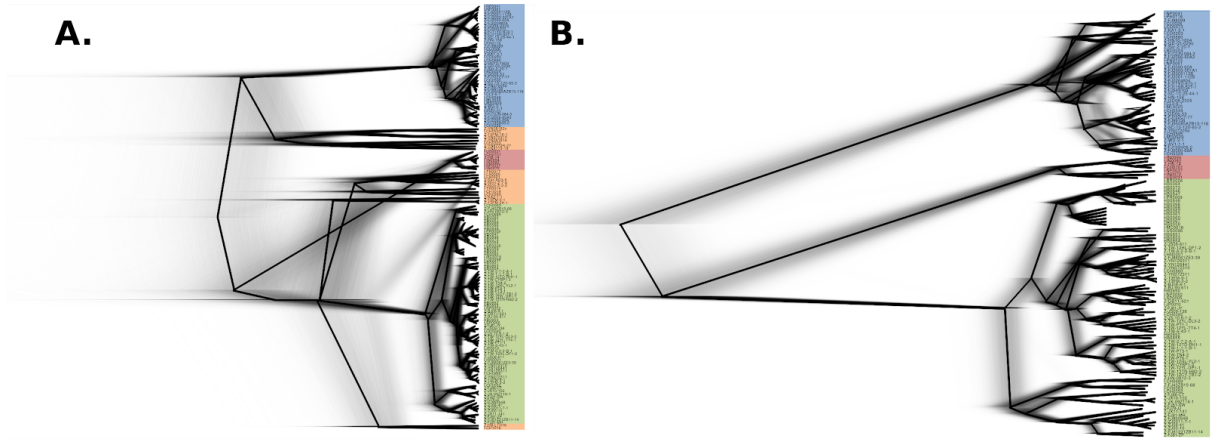
Supplementary Figure 3.3. Relation between genetic groups and sample mating type. PCA based on pairwise Hamming distances (same coordinates as Supp. Fig. 3.2). The color scheme codifies the assessed mating type.



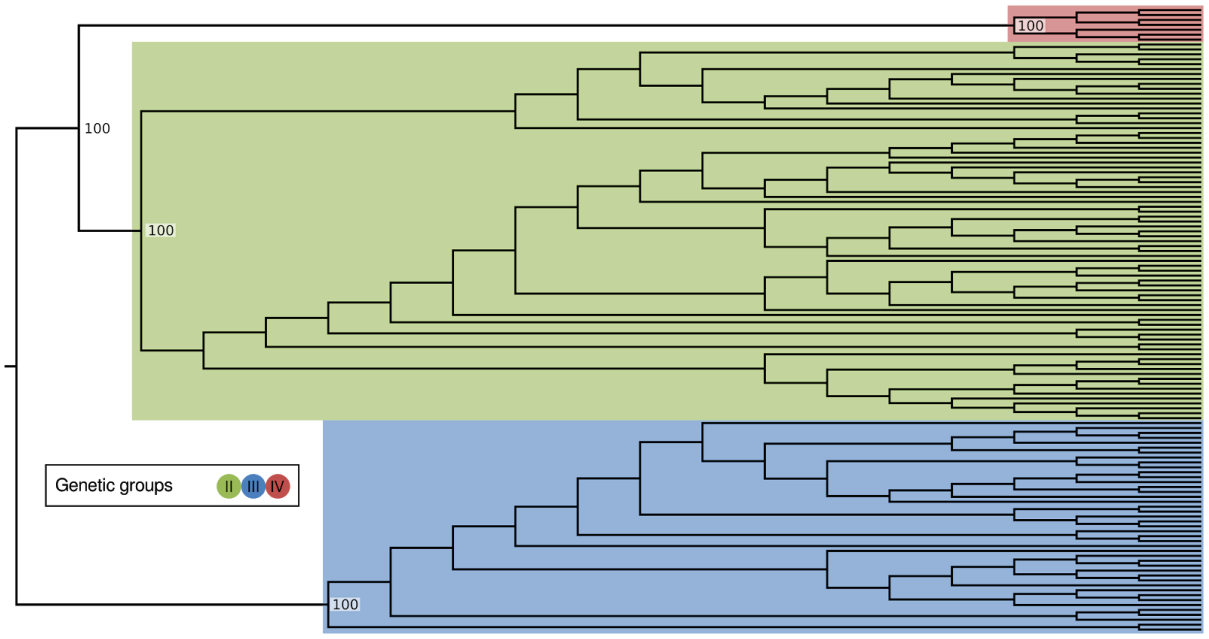
Supplementary Figure 3.4. The unfolded Site Frequency Spectrum (SFS). The SFS was calculated independently for each genetic group using genome-wide segregating sites, or only segregating sites within annotated genes or intergenic regions (see inset). The ancestral allele was ascertained using two different outgroups (see 3.4. Materials and Methods). The gray line shows the expected distribution of the SFS assuming no linked selection.



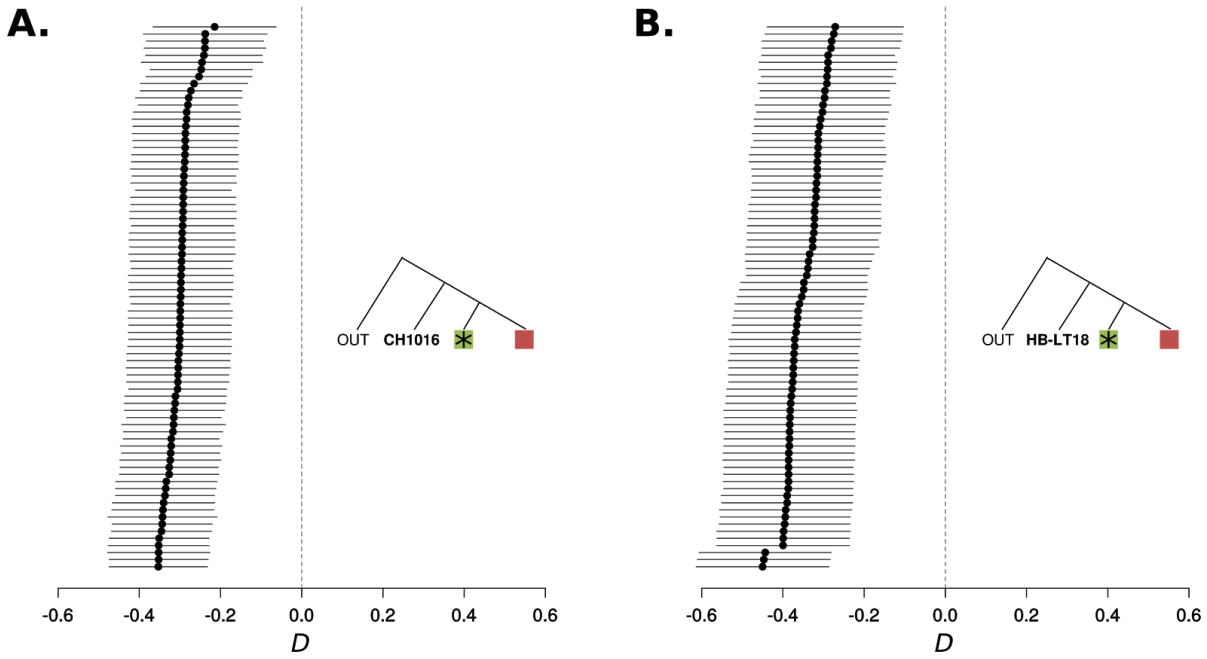
Supplementary Figure 3.5. Recent increase of population size in clonal lineages of *Magnaporthe oryzae*. An Extended Bayesian Skyline Plot was calculated from the Bayesian tip calibrated phylogeny (Figure 3.4). The thick line depicts the median Effective Population Size and the light colored silhouette represents its 95 HPD interval.



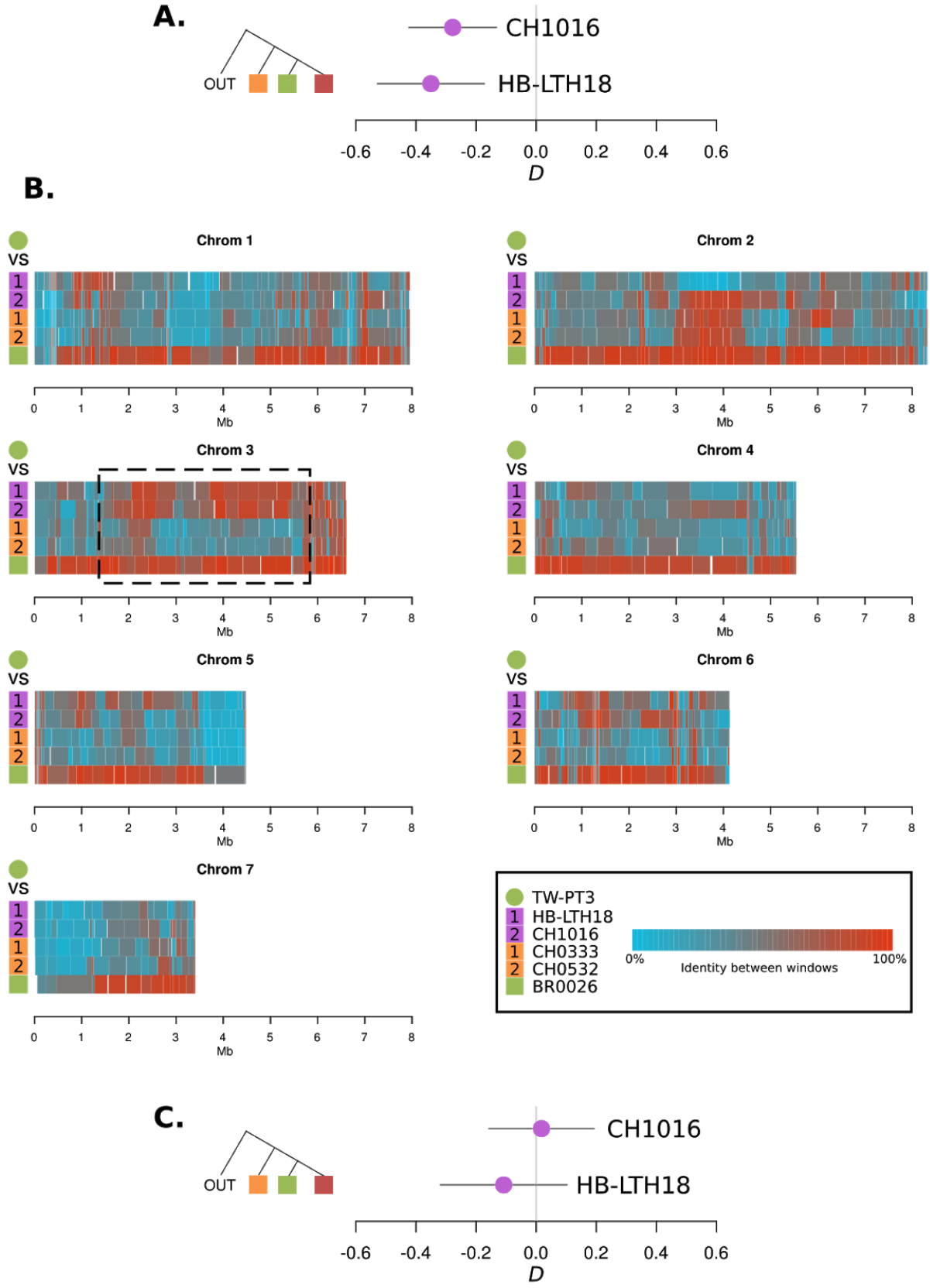
Supplementary Figure 3.6. Effect of recombination on the *Magnaporthe oryzae* phylogeny construction. Comparison between phylogenies with **(A)** and without **(B)** the diverse recombining group. Grey diffused lines depict all calculated trees, whereas the black lines represent the maximum clade credibility tree.



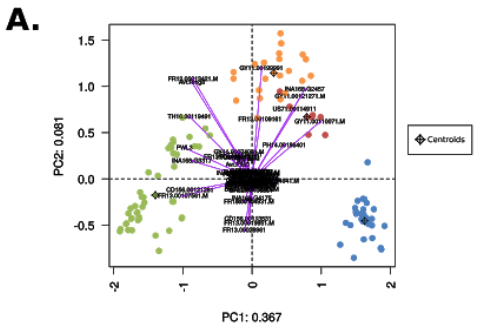
Supplementary Figure 3.7. Phylogenetic inference using SVDquartets. The combination of the sets of all possible quartets of isolates is represented in a single phylogenetic tree. The values on the nodes indicate the bootstrap support after 100 replications.

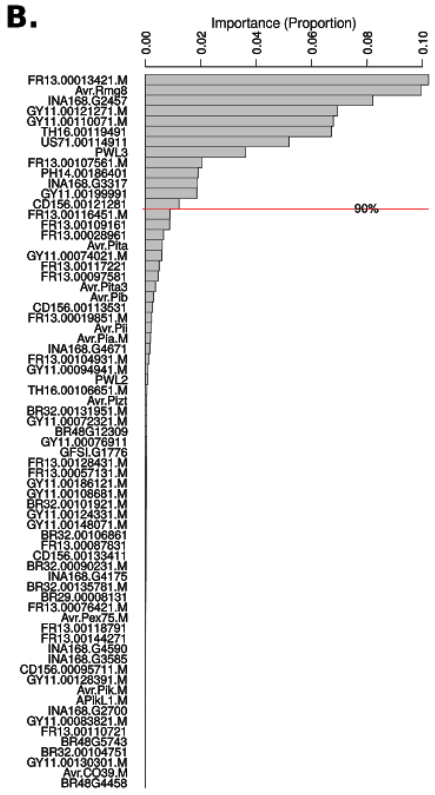


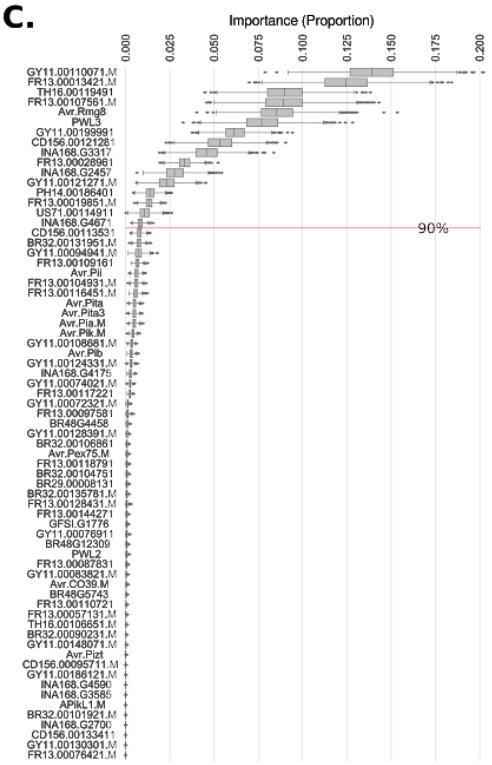
Supplementary Figure 3.8. Two Chinese individuals display consistent introgression with the clonal lineage II. *D*-statistics using different phylogenetic configurations depicted as colored inset trees. The green box with the asterisk represents a position in which all individuals from the clonal lineage II were placed in an iterative way. Red boxes represent a fixed individual from the clonal lineage IV. **(A)** *D*(Outgroup, CH1016, Lineage II individual, Lineage IV individual). **(B)** *D*(Outgroup, HB-LT18, Lineage II individual, Lineage IV individual). Points represent *D*-statistic tests, and lines depict 95% confidence intervals.

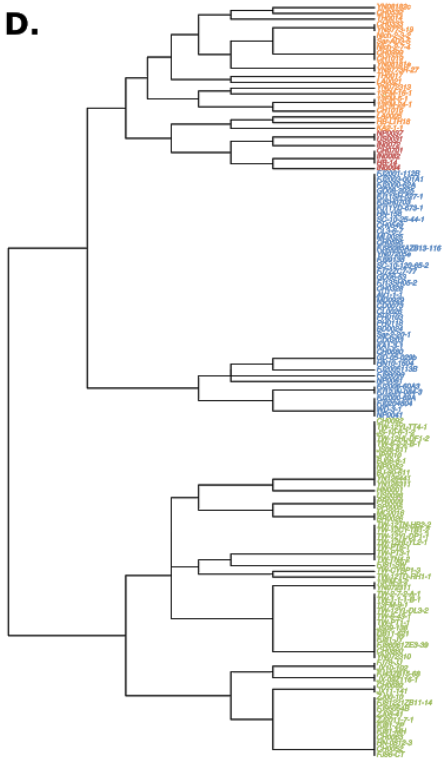

A.
OUT
CH1016
HB-LTH18
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6
D
B.
Chrom 1
Chrom 2
Chrom 3
Chrom 4
Chrom 5
Chrom 6
Chrom 7
VS
1
2
1
2
0 1 2 3 4 5 6 7 8
Mb
TW-PT3
HB-LTH18
CH1016
CH0333
CH0532
BR0026
0%
100%
Identity between windows
C.
OUT
CH1016
HB-LTH18
-0.6 -0.4 -0.2 0.0 0.2 0.4 0.6
D

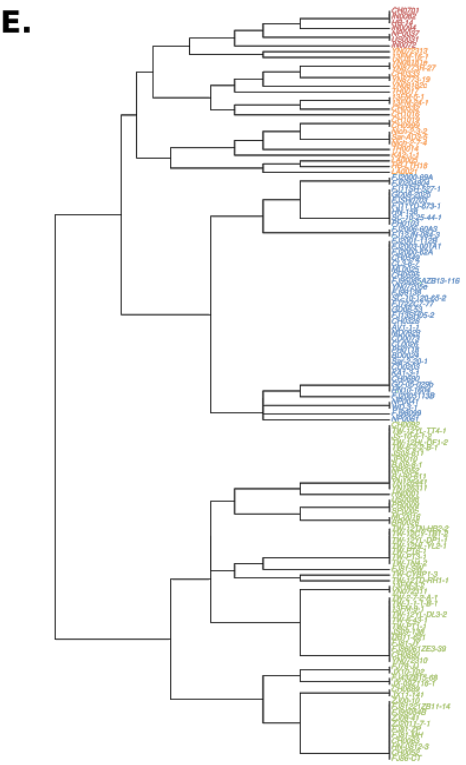
Supplementary Figure 3.9. Ancestry-based genomic segmentation of Chinese individuals *CH1016* and *HB-LTH18* reveals a 4 Mb putative introgressed region on chromosome 3. (A) The inset tree shows the *D*-statistics configuration, *D*(Outgroup, Orange; Green, Red), used to detect introgression between clonal lineage II and two individuals (*CH1016* and *HB-LTH18*) from the diverse group I (as in Fig 3.5B). Introgression is inferred based on the significant negative *D*-statistics. **(B)** Each panel shows homologous chromosomes from *CH1016* and *HB-LTH18* in addition to the control individuals from the diverse group I (CH033 and CH0532) and the clonal lineage II (BR0026), segmented according to their ancestry. The color coding represents the level of SNP similarity between each individual and the chosen clonal lineage II individual (TW-PT3) for that particular segment. Chromosome 3 shows a 4 Mb segment inferred to be introgressed between the clonal lineage II and both *CH1016* and *HB-LTH18* (boxed area). **(C)** The same *D*-statistic test as in A. was carried out after all putative introgressed fragments (in red) with a percentage similarity value of >= 60 were removed. The test was not significant, i.e., no different from zero.


A.
PC2: 0.081
PC1: 0.367
Centroids


B.
Importance (Proportion)
90%


C.
Importance (Proportion)
90%


D.


E.

Supplementary Figure 3.10. Effector loadings reveal major effector loss in clonal lineage III. **(A)** Biplot based on the presence/ and absence effector matrix (Figure 3.6C). Dots represent isolates color-coded by their genetic group. Vectors and labels correspond to the effector loadings for PC1 and PC2. All vectors were scaled by 3X for better representation. **(B)** The barplot represents the absolute value of the product of the PC1 and PC2 coordinates for each loading vector. The horizontal red line represents the cumulative sum of 90% of the data. **(C)** The set of boxplots summarize the distribution of the importance of each effector as decision factors for the genetic group assignment under 2,500 iterations of the extremely randomized trees algorithm. The horizontal red line represents the cumulative 90% of the data based on the mean values. **(D)** Hierarchical cluster-based dendrogram built by subsetting the 13 first effectors showed in B. **(E)** Hierarchical cluster-based dendrogram built by subsetting the 16 first effectors showed in **C.**

Supplementary Table 3.1. Samples information.

* Samples HB0075 and HB0076 correspond to *Setaria*-infecting individuals

Study: (H)erbaria This work ; (M)odern This work ; (G)ladieux *et al.*, 2018 ; (Z)hong *et al.*, 2018

Study	ID	Year	Country	Loc1	Loc2	Asserted MAT Type
H	HB0059	NA	ITA	NA	NA	1
H	HB0060	NA	ITA	NA	NA	1
H	HB0061	1891	ITA	NA	NA	1
H	HB0062	1891	ITA	NA	NA	1
H	HB0064	1937	ROU	NA	NA	1
H	HB0066	1963	ROU	NA	NA	1
H	HB0067	1891	ITA	NA	NA	1
H	HB0068	1891	ITA	NA	NA	1
H	HB0073	1937	ROU	NA	NA	1
H	HB0075	1948	ROU	NA	NA	NA
H	HB0076	1948	ROU	NA	NA	NA
M	HB0846	2010	ITA	Sozzago_No	NA	1
M	HB0847	2011	ITA	Vigevano_Pv	NA	1
M	HB0848	2011	ITA	Ferrara	NA	1
M	HB0850	2011	ITA	Olcenengo_Vc	NA	1
M	HB0851	2011	ITA	Verceli_Vc	NA	1
M	HB0852	2011	ITA	Siziano_Pv	NA	1
M	HB0853	NA	ITA	Oristano	NA	1
M	HB0854	2003	ITA	Dossena	NA	1
M	HB0855	2001	ITA	NA	NA	1
G	I-BD0024	1989	BDI	Burundi	Noebe	2
G	I-BR0026	1987	BRA	Mato_Grosso_Do_Sul	Fazenda_Progresso	1
G	I-CD0073	1981	CIV	Côte_d'Ivoire	Bouake	2
G	I-CD0203	2003	CIV	Côte_d'Ivoire	Tiassalé	2
G	I-CH0052	1991	CHN	Hunan	NA	1

G	I-CH0063	1985	CHN	Hunan	NA	1
G	I-CH0092	1983	CHN	Zhejiang	NA	1
G	I-CH0328	1998	CHN	Hunan	Anhua_Yanxi	2
G	I-CH0333	1998	CHN	Hunan	Yanxi	1
G	I-CH0532	1995	CHN	Yunnan	Baoshan	2
G	I-CH0549	1995	CHN	Yunnan	Baoshan	2
G	I-CH0595	1997	CHN	Yunnan	Yiliang	2
G	I-CH0680	2001	CHN	Hunan	Anhua_Yanxi	2
G	I-CH0689	2001	CHN	Hunan	Anhua_Yanxi	1
G	I-CH0701	2001	CHN	Hunan	Anhua_Yanxi	1
G	I-CH0860	2006	CHN	Yunnan	Wenshan_Dehou	1
G	I-CH0999	2008	CHN	Yunnan	Yule	1
G	I-CH1016	2008	CHN	Yunnan	Yule	2
G	I-CH1019	2008	CHN	Yunnan	Yule	2
G	I-CL0026	1989	COL	Meta	Villavicencio	2
G	I-CL3-6-7	2000	COL	Meta	Villavicencio	2
G	I-HN0001	1993	HUN	Hungary	Szarvas	1
G	I-IN0072	1992	IND	NA	NA	1
G	I-IN0082	1992	IND	NA	NA	1
G	I-IN0094	1993	IND	NA	NA	1
G	I-JP0010	1986	JPN	NA	Fukushima	1
G	I-LA0005	2009	LAO	Luang prabang	Silalek	1
G	I-LA0021	2009	LAO	Luang prabang	Silalek	2
G	I-MC0016	1997	MAR	NA	Tazi	1
G	I-MD0929	2005	MDG	NA	Andranomanelatra	2
G	I-ML0025	1986	MLI	NA	Niema	2
G	I-NP0037	2009	NPL	Kaski	Begnas	1
G	I-NP0041	2009	NPL	Kaski	Begnas	2
G	I-NP0052	2008	NPL	Kathmandu	Sangle_VDC9	1

G	I-NP0061	2008	NPL	Kathmandu	Sangle_VDC9	2
G	I-PH0103	NA	PHL	NA	NA	2
G	I-PH0118	NA	PHL	NA	NA	2
G	I-PR0009	1991	PRT	NA	Sado_Torre_Do_Cleri	1
G	I-SP0005	1994	ESP	Catalonia	Amposta_Salats	1
G	I-TH0014	1987	THA	NA	NA	2
G	I-TH0017	1987	THA	NA	NA	2
G	I-US0031	1967	USA	Louisiana	NA	1
G	I-US0098	1992	USA	Arkansas	Lonoke	1
Z	Z-13FM-16-1	2013	CHN	Yunnan	NA	2
Z	Z-13FM-24-1	2013	CHN	Yunnan	NA	2
Z	Z-13FM-3-2	2013	CHN	Yunnan	NA	1
Z	Z-13FM-5-1	2013	CHN	Yunnan	NA	2
Z	Z-13FM-9-1	2013	CHN	Yunnan	NA	1
Z	Z-AV1-1-1	2015	GHA	Aveyime	NA	2
Z	Z-BJ08-8-1	1980s	CHN	Beijing	NA	1
Z	Z-BJ-90-611	1990s	CHN	Beijing	NA	1
Z	Z-DB11-621	2011	CHN	Liaoning	NA	1
Z	Z-FJ0204804	2002	CHN	Fujian	NA	2
Z	Z-FJ11SH-527-1	2011	CHN	Fujian	Shanghang	2
Z	Z-FJ11YD-673-1	2011	CHN	Fujian	Yongding	2
Z	Z-FJ12JN-084-3	2012	CHN	Fujian	Jianning	2
Z	Z-FJ13SH05-2	2013	CHN	Fujian	Shanghang	2
Z	Z-FJ2000-62A	2000	CHN	Fujian	NA	2
Z	Z-FJ2000-69A	2000	CHN	Fujian	NA	2
Z	Z-FJ2001-112B	2001	CHN	Fujian	NA	2
Z	Z-FJ2003-001A1	2003	CHN	Fujian	NA	2
Z	Z-FJ2005113B	2005	CHN	Fujian	NA	2
Z	Z-FJ2006-60A3	2006	CHN	Fujian	NA	2

Z	Z-FJ43ZB15-68	1992	CHN	Fujian	Youxi	1
Z	Z-FJ72ZC7-77	1992	CHN	Fujian	Jianyang	2
Z	Z-FJ78-JJ	1978	CHN	Fujian	Jinjiang	1
Z	Z-FJ81221ZB11-14	1981	CHN	Fujian	Fuzhou	1
Z	Z-FJ81-JY	1981	CHN	Fujian	Jianyang	1
Z	Z-FJ81-MH	1981	CHN	Fujian	Minhou	1
Z	Z-FJ81-SW	1981	CHN	Fujian	Shaowu	1
Z	Z-FJ81-ZP	1981	CHN	Fujian	Zhangpu	1
Z	Z-FJ86061ZE3-39	1986	CHN	Fujian	Shaowu	1
Z	Z-FJ86-CT	1986	CHN	Fujian	Changting	1
Z	Z-FJ95054B	1995	CHN	Fujian	Changting	1
Z	Z-FJ95085AZB13-116	1995	CHN	Fujian	Longyan	2
Z	Z-FJ98099	1998	CHN	Fujian	Changting	2
Z	Z-FJ99138	1999	CHN	Fujian	NA	2
Z	Z-FJSH0703	2007	CHN	Fujian	Shanghang	2
Z	Z-GD-05-029b	2005	CHN	Guandong	Conghua	2
Z	Z-GD06-53	2006	CHN	Guandong	Jiexi	2
Z	Z-GD08-2025	2008	CHN	Guandong	Wengyuan	2
Z	Z-HB-14	2009	CHN	Hubei	Xianfeng	1
Z	Z-HB-LTH18	2009	CHN	Hubei	Xianfeng	1
Z	Z-HN-0812-3	2008	CHN	Hunan	Taojiang	1
Z	Z-HN10-1604	2010	CHN	Hunan	Taojiang	2
Z	Z-HN-158	2009	CHN	Hunan	Taojiang	2
Z	Z-JS08-611	2008	CHN	Jiangsu	NA	1
Z	Z-JS09-138	2009	CHN	Jiangsu	NA	1
Z	Z-JS-10-6-1-2	2010	CHN	Jiangsu	NA	1
Z	Z-JX-09Z116-1	2009	CHN	Jiangxi	Wanan	1
Z	Z-JX10-102	2010	CHN	Jiangxi	Fengcheng	1
Z	Z-JX11-141	2011	CHN	Jiangxi	Jiangxi	1

Z	Z-KA1-3-1	2015	GHA	Ghana	Kade-abaam	2
Z	Z-KA2-1-1	2015	GHA	Ghana	Kade-abaam	2
Z	Z-Nich-2-3-2	2014	SUR	Suriname	Nickerie	2
Z	Z-Nich-2-7-4	2014	SUR	Suriname	Nickerie	2
Z	Z-Sar-2-20-1	2013	SUR	Suriname	Saramacca	2
Z	Z-Sar-AD3-5	2013	SUR	Suriname	Nickerie	2
Z	Z-SC-10-120-65-2	2010	CHN	Sichuan	Nanbu	2
Z	Z-SC-10-25-44-1	2010	CHN	Sichuan	Nanbu	2
Z	Z-TW-1-1-1-B-1	2005	TWN	Kaohsiung	NA	1
Z	Z-TW-12CY-TB1-2	2012	TWN	Chiayi	NA	1
Z	Z-TW-12HL-DF1-2	2012	TWN	Hualien	NA	1
Z	Z-TW-12HL-YL2-1	2012	TWN	Hualien	NA	1
Z	Z-TW-12TD-RH1-1	2012	TWN	Taitung	NA	1
Z	Z-TW-12TN-HB2-2	2012	TWN	Tainan	NA	1
Z	Z-TW-12YL-DL3-2	2012	TWN	Yunlin	NA	1
Z	Z-TW-12YL-DP1-1	2012	TWN	Yunlin	NA	1
Z	Z-TW-12YL-TT4-1	2012	TWN	Yunlin	NA	1
Z	Z-TW-2-7-2-A-1	2005	TWN	Pingtung	NA	1
Z	Z-TW-6-2-2-B-1	2005	TWN	Chiayi	NA	1
Z	Z-TW-6-43-1	2005	TWN	Chiayi	NA	1
Z	Z-TW-CYBP1-3	2011	TWN	Chiayi	NA	1
Z	Z-TW-PT1-1	2011	TWN	Pingtung	NA	1
Z	Z-TW-PT3-1	2011	TWN	Pingtung	NA	1
Z	Z-TW-PT6-1	2011	TWN	Pingtung	NA	1
Z	Z-TW-TN4-2	2011	TWN	Tainan	NA	1
Z	Z-WD-3-1	2015	GHA	Ghana	Weed-dabala	2
Z	Z-YN07205e	2007	CHN	Yunnan	NA	2
Z	Z-YN072310	2007	CHN	Yunnan	NA	1
Z	Z-YN072311	2007	CHN	Yunnan	NA	1
Z	Z-YN072313	2007	CHN	Yunnan	NA	1

Z	Z-YN08181e	2008	CHN	Yunnan	NA	1
Z	Z-YN08182c	2008	CHN	Yunnan	NA	2
Z	Z-YN126311	2012	CHN	Yunnan	NA	1
Z	Z-YN126441	2012	CHN	Yunnan	NA	1
Z	Z-YN8773-19	1987	CHN	Yunnan	NA	2
Z	Z-YN8773R-27	1987	CHN	Yunnan	NA	1
Z	Z-ZJ00-10	2000	CHN	Zhejiang	NA	1
Z	Z-ZJ08-41	2008	CHN	Zhejiang	NA	1
Z	Z-ZJ2011-7-1	2011	CHN	Zhejiang	NA	1

Supplementary Table 3.2. New classification assessed in this study.

Study: (H)erbaria This work ; (M)odern This work ; (G)ladieux *et al.*, 2018 ; (Z)hong *et al.*, 2018

Dataset	ID	Lineage (Gladieux *et al.*, 2018)	Clade (Zhong *et al.*, 2018)	Classification
G	CH0333	*5*	NA	I
G	CH0532	*5*	NA	I
G	CH0999	1	NA	I
G	CH1016	6	NA	I
G	CH1019	1	NA	I
G	LA0005	1	NA	I
G	LA0021	1	NA	I
G	TH0014	1	NA	I
G	TH0017	1	NA	I
Z	13FM-16-1	NA	1	I
Z	13FM-24-1	NA	1	I
Z	13FM-5-1	NA	1	I
Z	HB-LTH18	NA	1	I
Z	KA2-1-1	NA	1	I
Z	Nich-2-3-2	NA	1	I
Z	Nich-2-7-4	NA	1	I
Z	Sar-AD3-5	NA	1	I
Z	YN072313	NA	1	I
Z	YN08181e	NA	1	I
Z	YN08182c	NA	1	I
Z	YN8773-19	NA	1	I
Z	YN8773R-27	NA	1	I
G	BR0026	2	NA	II
G	CH0052	2	NA	II
G	CH0063	2	NA	II
G	CH0092	2	NA	II

G	CH0689	2	NA	II
G	CH0860	2	NA	II
G	HN0001	2	NA	II
G	JP0010	2	NA	II
G	MC0016	2	NA	II
G	NP0052	2	NA	II
G	PR0009	2	NA	II
G	SP0005	2	NA	II
G	US0098	2	NA	II
H	HB0059	NA	NA	II
H	HB0060	NA	NA	II
H	HB0061	NA	NA	II
H	HB0062	NA	NA	II
H	HB0064	NA	NA	II
H	HB0066	NA	NA	II
H	HB0067	NA	NA	II
H	HB0068	NA	NA	II
H	HB0073	NA	NA	II
M	HB0846	NA	NA	II
M	HB0847	NA	NA	II
M	HB0848	NA	NA	II
M	HB0850	NA	NA	II
M	HB0851	NA	NA	II
M	HB0852	NA	NA	II
M	HB0853	NA	NA	II
M	HB0854	NA	NA	II
M	HB0855	NA	NA	II
Z	13FM-3-2	NA	3	II
Z	13FM-9-1	NA	3	II
Z	BJ08-8-1	NA	3	II

Z	BJ-90-611	NA	3	II
Z	DB11-621	NA	3	II
Z	FJ43ZB15-68	NA	3	II
Z	FJ78-JJ	NA	3	II
Z	FJ81221ZB11-14	NA	3	II
Z	FJ81-JY	NA	3	II
Z	FJ81-MH	NA	3	II
Z	FJ81-SW	NA	3	II
Z	FJ81-ZP	NA	3	II
Z	FJ86061ZE3-39	NA	3	II
Z	FJ86-CT	NA	3	II
Z	FJ95054B	NA	3	II
Z	HN-0812-3	NA	3	II
Z	JS08-611	NA	3	II
Z	JS09-138	NA	3	II
Z	JS-10-6-1-2	NA	3	II
Z	JX-09Z116-1	NA	3	II
Z	JX10-102	NA	3	II
Z	JX11-141	NA	3	II
Z	TW-1-1-1-B-1	NA	3	II
Z	TW-12CY-TB1-2	NA	3	II
Z	TW-12HL-DF1-2	NA	3	II
Z	TW-12HL-YL2-1	NA	3	II
Z	TW-12TD-RH1-1	NA	3	II
Z	TW-12TN-HB2-2	NA	3	II
Z	TW-12YL-DL3-2	NA	3	II
Z	TW-12YL-DP1-1	NA	3	II
Z	TW-12YL-TT4-1	NA	3	II
Z	TW-2-7-2-A-1	NA	3	II
Z	TW-6-2-2-B-1	NA	3	II

Z	TW-6-43-1	NA	3	II
Z	TW-CYBP1-3	NA	3	II
Z	TW-PT1-1	NA	3	II
Z	TW-PT3-1	NA	3	II
Z	TW-PT6-1	NA	3	II
Z	TW-TN4-2	NA	3	II
Z	YN072310	NA	3	II
Z	YN072311	NA	3	II
Z	YN126311	NA	3	II
Z	YN126441	NA	3	II
Z	ZJ00-10	NA	3	II
Z	ZJ08-41	NA	3	II
Z	ZJ2011-7-1	NA	3	II
G	BD0024	3	NA	III
G	CD0073	3	NA	III
G	CD0203	3	NA	III
G	CH0328	3	NA	III
G	CH0549	3	NA	III
G	CH0595	3	NA	III
G	CH0680	3	NA	III
G	CL0026	3	NA	III
G	CL3-6-7	3	NA	III
G	MD0929	3	NA	III
G	ML0025	3	NA	III
G	NP0041	3	NA	III
G	NP0061	3	NA	III
G	PH0103	3	NA	III
G	PH0118	3	NA	III
Z	AV1-1-1	NA	2	III
Z	FJ0204804	NA	2	III

Z	FJ11SH-527-1	NA	2	III
Z	FJ11YD-673-1	NA	2	III
Z	FJ12JN-084-3	NA	2	III
Z	FJ13SH05-2	NA	2	III
Z	FJ2000-62A	NA	2	III
Z	FJ2000-69A	NA	2	III
Z	FJ2001-112B	NA	2	III
Z	FJ2003-001A1	NA	2	III
Z	FJ2005113B	NA	2	III
Z	FJ2006-60A3	NA	2	III
Z	FJ72ZC7-77	NA	2	III
Z	FJ95085AZB13-116	NA	2	III
Z	FJ98099	NA	2	III
Z	FJ99138	NA	2	III
Z	FJSH0703	NA	2	III
Z	GD-05-029b	NA	2	III
Z	GD06-53	NA	2	III
Z	GD08-2025	NA	2	III
Z	HN10-1604	NA	2	III
Z	HN-158	NA	2	III
Z	KA1-3-1	NA	2	III
Z	Sar-2-20-1	NA	2	III
Z	SC-10-120-65-2	NA	2	III
Z	SC-10-25-44-1	NA	2	III
Z	WD-3-1	NA	2	III
Z	YN07205e	NA	2	III
G	CH0701	4	NA	IV
G	IN0072	4	NA	IV
G	IN0082	4	NA	IV
G	IN0094	4	NA	IV

G	NP0037	4	NA	IV
G	US0031	4	NA	IV
Z	HB-14	NA	1	IV

Supplementary Table 3.3. Members of the pan-effectorome that are present or absent in all 131 *Magnaporthe oryzae* isolates used in this study along with the effectors showing presence and absence polymorphism.

Present in all isolates (69 effectors)		**Absent in all isolates (40 effectors)**	**Effectors showing presence and Absence polymorphism (70 effectors)**	
Avr-Pi54	FR13-00089161-M	APikL3-M	APikL1-M	FR13-00118791
Avr-Pi9	FR13-00093031-M	BR29-00004921-M	Avr-CO39-M	FR13-00128431-M
BR32-00003601-M	FR13-00094721-M	BR29-00026641-M	Avr-Pex75-M	FR13-00144271
BR32-00040381	FR13-00094761	BR29-00027131-M	Avr-Pia-M	GY11-00072321-M
BR32-00066181-M	FR13-00099751-M	BR29-00043011-M	Avr-Pib	GY11-00074021-M
BR32-00081411-M	FR13-00101061	BR29-00052161-M	Avr-Pii	GY11-00076911
BR32-00101991-M	FR13-00107801-M	BR29-00081821-M	Avr-Pik-M	GY11-00083821-M
BR32-00117431-M	FR13-00108711-M	BR29-00089541	Avr-Pita3	GY11-00094941-M
BR32-00131211-M	FR13-00114581-M	BR29-00091361-M	Avr-Pita	GY11-00096851
BR32-00132551	FR13-00122001-M	BR29-00091381-M	Avr-Pizt	GY11-00108681-M
BR32-00133701-M	FR13-00143751-M	BR29-00091681-M	Avr-Rmg8	GY11-00110071-M
BR32-00141021-M	GY11-00054591-M	BR29-00092011	BR29-00008131	GY11-00124331-M
BR32-00141031-M	GY11-00060011	BR29-00094821-M	BR32-00090231-M	GY11-00128391-M
BR32-00141181-M	GY11-00067931	BR29-00096171-M	BR32-00090371-M	GY11-00128491-M
BR48G12130	GY11-00069551-M	BR29-00096401	BR32-00101921-M	GY11-00148071-M
BR48G630	GY11-00121271-M	BR29-00096771-M	BR32-00101931-M	GY11-00186121-M
BR48G9741	GY11-00130301-M	BR29-00097421	BR32-00101951-M	GY11-00199991
BR58-G4698	GY11-00137731-M	BR29-00105021-M	BR32-00104751	INA168-G2457
CD156-00095711-M	GY11-00138681-M	BR29-00105921	BR32-00106861	INA168-G2700
CD156-00118721	INA168-G4590	BR29-00106461-M	BR32-00128081-M	INA168-G3317
CD156-00122701-M	MGDIG41	BR29-00107161	BR32-00131951-M	INA168-G3585
CD156-00131711-M	MOBR58	BR29-00107301	BR32-00135781-M	INA168-G4175
CD156-00138051-M	TH12-00085091-M	BR29-00112111-M	BR32-00143481	INA168-G4671
FR13-00001281-M	TH12-00121641-M	BR29-00114891-M	BR48G12309	PH14-00186401
FR13-00003821-M	TH16-00079081-M	BR29-00115871	BR48G4458	PWL2
FR13-00004061-M	TH16-00100711-1-M	BR29-00115881	BR48G5743	PWL3

Present in all isolates (69 effectors)		**Absent in all isolates (40 effectors)**	**Effectors showing presence and Absence polymorphism (70 effectors)**	
FR13-00019851-M	TH16-00118381-M	BR29-00117591-M	BR58-G2338	TH16-00106651-M
FR13-00020841-M	US71-00065681-M	BR29-00118801-M	CD156-00113531	TH16-00119491
FR13-00035971	US71-00116921-M	BR29-00119471	CD156-00121281	US71-00000121-M
FR13-00041321-M		BR29-00119491-M	CD156-00133411	US71-00114911
FR13-00057121-M		BR29-00119511-M	FR13-00013421-M	
FR13-00057131-M		BR29-00121041-M	FR13-00028961	
FR13-00066391-M		BR29-00121721	FR13-00076421-M	
FR13-00067231		BR29-00125251-M	FR13-00087831	
FR13-00068741-M		BR32-00102581	FR13-00097581	
FR13-00069491		BR32-00120141	FR13-00104931-M	
FR13-00070821-M		DIG41-G5801	FR13-00107561-M	
FR13-00080861-M		DIG41-G7456	FR13-00110721	
FR13-00081261-M		FR13-00109161	FR13-00116451-M	
FR13-00088361-M		GFSI-G1776	FR13-00117221	

Supplementary Table 3.4. Patterns of effector presence and absence polymorphism in *Magnaporthe oryzae* isolates

Pattern	Group showing presence	Group showing absence	Effector
Lineage/group specific	I, II, IV	III	TH16.00119491* FR13.00013421.M* Avr-Rmg8* CD156.00121281*
	I, II	III, IV	PWL3* INA168.G3317* FR13-00107561.M*
	I, II, III	IV	FR13.00028961*
	I, III, IV	II	GY11.00110071.M*
	I	II, III, IV (with a few exceptions)	GY11.00199991*
Patchy distribution	Varies	Varies	Avr-Pia.M Avr-Pii
Partial gene loss within a group	I, III, IV	II	PH14.00186401* GY11.00094941.M GY11.00121271.M* INA168.G2457* US71.00114911*
Near-complete gene loss across all groups	Varies	Varies	FR13.00109161 APikL1 CD156.00133411 Avr-CO39.M BR32.00090231.M BR29.00008131 GFSI.G1776