Is Mutational Meltdown a threat to the mega-diverse genus *Begonia*?

Thibauld Michel

Doctoral Thesis
Royal Botanic Gardens Edinburgh
University of Edinburgh
2023

Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree. Except where stated otherwise by reference or acknowledgement, the work presented is entirely my own.

Thibauld Michel
01/02/2023

Acknowledgements

This project is the product of researches that have been undertaken mainly in the Royal Botanic Garden of Edinburgh, in coordination with multiple institutes. There are many collaborators and colleagues that have contributed to this work over the years and I am grateful to have made their acquaintance.

In the RBGE, I have to thank first My principal supervisor Dr. Catherine Kidner, who is the best supervisor possible and has developed my taste for plant biology, genetics, and Oolong tea. My thanks to my supervisor assistant Dr. Mark Hughes for his indefectible help, the Dr. Lucía Campos for answering my daily questions, the Dr Flávia Pezzini for bioinformatic help, Dr. Michelle Hart and Dr. Laura Forrest for molecular biology assistance, Dr. Peter Moonlight for his insights on *Begonia* taxonomy, Dr. Giada Ferrari for aDNA advisory, Dr. Anthony Miller an Dr. Alan Forrest for specimens and documentation about the flora of Socotra, Dr. David Harris for herbarium references, Dr. Michael Möller for advices on inbreeding estimation, and Dr. Markus Ruhsam for his critical thinking. Thanks to Dr. Yu-Hsin Tseng from the National Chung Hsing University for her inspiring advices. Among the horticulture team I wish to thank especially Andy Ensoll for his help and information about the Socotran begonia, and the rest of the team for maintaining the collection, and sharing their botanical knowledge.

I wish to thanks my fellow PhD students with whom I have shared many coffee through the years: Gustavo, Surhabi, Ellen, Anya, Cynthia, Charlie, Ludwig, Natalia, Darth Madhavi, Katie, Pakkapol, Nadia, Jess, Hazel, Matt, Ding, Briggitthe, Isuru, and Andres. A special thanks to the Msc. and Bsc. students that have worked conjointly on this project, Philippa Stone, and Zitian Tang.

In the University of Edinburgh I wish to thank among others, Dr. Alex Twyford for his help and insights, Dr. Richard Milne for his stimulating ideas and support, and Dr. Andrew Hudson for welcoming me in his laboratory.

In the NTNU University Museum in Trondheim, my thanks go to Dr. Michael Martin who has welcomed me in the Holomuseomics research group, and to Dr. Vanessa Bieker who has supervised me during this fantastic experience.

I must thank all the partners and students who have collaborated in the context of the European project Plant.ID. They are too numerous to be all quoted here, but among my fellow Early Stage Researchers, I wish to thanks Dr. Yannick "BOT" Wousta and Dr. Anne-Sophie Quatela for their stimulating botanical enthusiasms and their friendship.

I thank Saray Garcia who has supported me through this project, and put her considerable talent to illustrate the Paleogenomics chapter. Eventually, thanks to my French family including P'pa, M'man, Carl, Louise, and Sylvain, as well as my Scottish family including Baklava, Edeline, Andy, Rebecca, Sandy, Martha, and the 15-year-old kitten Mothie.

Abstract

Begonia is one of the most species-rich angiosperm genera, studied for its rapid species radiation in tropical regions, and high morphological diversity. Typical populations are isolated and many display characteristics of narrow endemism. Endemic populations are prone to inbreeding and vulnerable to anthropogenic disturbance, while being isolated and difficult to access for population size estimation. For these rare species, herbarium specimens are the most accessible material available, even though the number of specimens collected for a single population is few.

We have developed a pipeline to use genomic data recovered from a single herbarium specimen to estimate the degree of inbreeding and the demographic history of the population. This pipeline has been designed to process low-coverage ancient DNA datasets from non-model organisms and assess the inbreeding coefficient using several genomic homozygosity estimators.

The pipeline integrate several tools to manage ancient DNA (aDNA) damage patterns, duplicated genes, problematic baits, and to determine homozygosity patterns in fresh and historical specimens.

The pipeline includes mapDamage, a tool to quantify nucleotides substitution A to G or C to T in the set of data, and recalibrate the quality score of the alignment files, minimizing the bias due to aDNA patterns of damages.

Target capture baits matching multiple regions of the genome have been identified, characterised, and removed from the analysis as well to prevent subsequent incorrect variant call.

Many paralogous genes are found in *Begonia* genomes due to an early whole genome duplication event in the history of the genus. As this can introduce a bias in the variant calling step of the pipeline, we have implemented a step to detect baits capturing sequences from paralogous genes in our analysis. Three methods have been considered for this: deviation of the genotype frequencies expected in a mapping population, detection of a unexpected level of heterozygosity (HDplot tool), or segregating multiple contigs aligning to the same bait (pipeline HybPiper). This analysis used genome skims from a mapping population to test the approaches. The study showed low overlap between the baits detected as capturing paralogs between the three methods with only 73 detected in all of them.

Herbarium historical specimens from a single population are scarce, and at one time point considered we can expect to find a reduced number of specimens available for analysis. In a lot of cases, only a unique specimen is available and represent the whole population. Therefore, rather than using inbreeding coefficients based on alleles frequencies, we are using Runs of Homozygosity (ROH) to estimate inbreeding and need only a single sample to be measured. To be able to measure ROH with Hyb-Seq data, we needed to know what part of the genome the Begonia baits are capturing with contiguous baits. The length of genome captured by the bait set has been calculated for the four most complete *Begonia* genomes available to determine the length of syntenic regions which can be captured.

This was a key point to establish the last part of the pipeline to calculate the size of ROH. We used PLINK to detect and quantify ROHs from VCF files produced by variant calling. The estimators derived are the total length of ROH in the dataset (SROH), the total number of ROH in the dataset (NROH), and the frequency of ROH for each sample (FROH). The confrontation of the SROH and NROH scores on a scatter plot provide an estimation of the relative size of the population, and give clues about an admixture with another population, a bottleneck event, or consanguinity are provided by this plot. The FROH estimator is less informative but follows linearly the size of the population estimated by the NROH/SROH plot. It has been used to study the biogeography of the specimens and mapped to their phylogenetic reconstruction to investigate the patterns of homozygosity.

We have analysed two sets of target-capture data with the pipeline, one with Arabian *Begonia*, and the second with *Begonia* from Papua New Guinea.

The first set is composed of 43 specimens of Arabian *Begonia* specimens from the Socotran archipelago including the species *B. socotrana* and *B. samhaensis* and with silica-dried and herbarium-dried historical specimens. Examination of the Hyb-Seq Socotran dataset revealed uneven coverage across the baits. This capture has been used to show the limitation of the pipeline, as phylogenetic reconstruction has not been successful beyond species level, and the ROH estimations were not significant.

The second set of target capture data included 160 samples from the New Guinea Highlands, from silica-dried and herbarium-dried historical specimens. As output of the pipeline, 10 specimens showed high homozygosity levels indicating a bottleneck in their demographic history, 3 outliers were suspected to be inbred, 60 were found to be from a large population or showing introgression, and 87 did not display homozygosity patterns significant enough and were filtered out by the pipeline. Mapping FROH metrics to the phylogeny shows a group within section *Petermannia* with consistently high homozygosity levels. Biogeographical analysis of the distribution of the samples did not reveal any clear relation between patterns of homozygosity and geographic location of the populations sampled. The data analysis has revealed a higher genetic diversity than expected in the Papua New Guinea *Begonia* collected and has given clues about the origin of the homozygosity patterns observed which seem more related to phylogenetic relationship rather than microevolution at population level.

Lay summary

Begonia is one of the largest group of flowering plant, with more than 2,000 species growing mostly in subtropical and tropical climates. They are usually studied for their diversification into a large array of species with different morphologies, notably a vast diversity of leaf shapes. *Begonia* usually display distant populations with a low number of individuals, and due to poor pollen dispersal, most populations are isolated from each other. It might be the cause of the high number of species in the *Begonia* group and might promote genetic isolation as well.

Isolation of population increases loss of genetic diversity, which can reduce fitness of the plants and eventually lead to inbreeding. This is a serious threat for this group of plants, as it can prevent their adaptation to fast-evolving environmental changes or other ecological pressure.

As many tropical plants share the same population isolation than *Begonia* species and might be under threat, we decided to set up a tool to detect if a population is inbred, and fragilized at a genetic level by its isolation. This pipeline is designed to analyse freshly collected plants and historical museum specimens, mostly herbarium sheets specimens. Using plants genomic records collected at different date, we can visualize the evolution of their inbreeding level over time and assess if the population should be prioritized for conservation.

However using herbarium specimens has several constrains. First the DNA contained in the plant material is scarce and degraded, we had to use a particular method of extraction and specific tools to analyse the data. Another issue with herbarium dried historical material is that a very few plants are available for one population. Often, only one single sample is available for a population at a given time. Therefore, we selected a method to estimate the inbreeding level requiring only a single specimen.

The informatics pipeline that has been set up in the frame of this project has been built around these constrains, and incorporate tools to analyse single herbarium specimens and estimate their inbreeding level. Two batches of data have been analysed with the inbreeding detection tools: one of Arabian *Begonia*, and another one of *Begonia* of Papua New Guinea. Preliminary results suggest that a very low number of our specimens are from consanguineous populations, and they might not be representative of their populations as there is clues they are product of self-pollination. Further inquiries should be made on the Arabian and Papua New Guinea specimens, with re-sampling and a different sequencing method to validate the results of the present study. Meanwhile, the analysis pipeline itself is operating well, and will be used on other groups of tropical herbaceous plants to test their genetic health and help to assess their conservation status.

Contents

Chapter 1

Introduction to museomics

1.1 Herbaria as plant libraries

Herbaria

Museum collections of plants specimens are maintained in herbaria all across the world (Fig.1.1). World-wide there are 3,522 active herbaria in 2021, containing more than 397 millions archived specimens (Thiers, 2022). These plant collections have been used for decades as reference for taxonomical identification, and more recently as sampling material for genetic, ecological, and environmental studies. They are records of global biodiversity encompassing several centuries of patient plant sampling (Rønsted et al., 2020). This database of natural variation can be used to answer many research questions related to evolutionary history, taxonomy, demographic changes, ecology, conservation, climatic changes and archaeology (Larridon et al., 2020, Nic Lughadha et al., 2019, Brown, 1999, Lang et al., 2019, Gutaker et al., 2019). The global and historical nature of herbaria collections provides a unique perceptive on geographic and temporal patterns in plant diversity which is key to solving large-scale ecological problems, such as the present ecological crisis. Furthermore, global herbaria collections continue to grow and provide a record for the future to help solve questions we are not aware of yet.

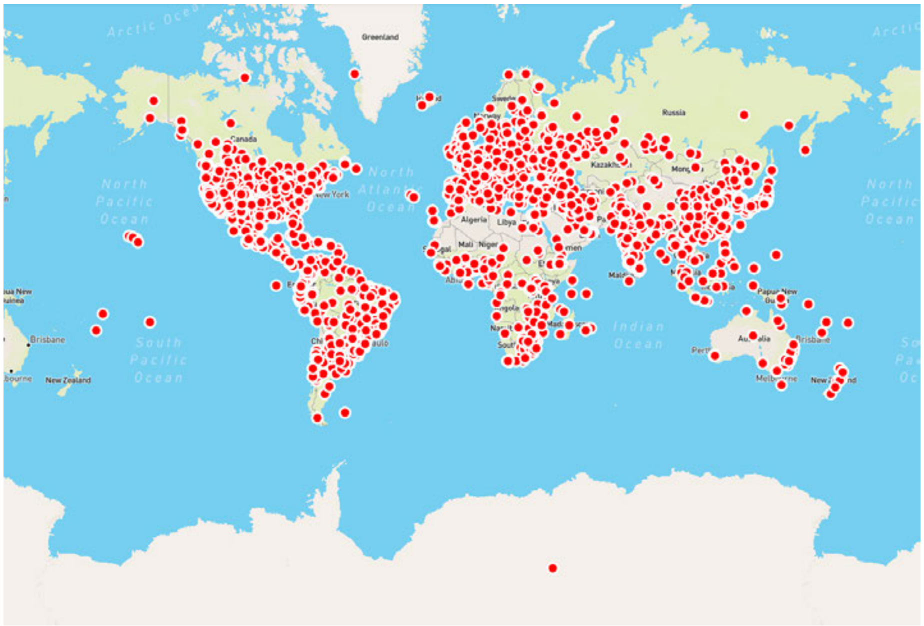


Figure 1.1: Location of herbaria worldwide (NYBG steere herbarium website)

Herbarium specimens

Herbarium specimens themselves are pressed dried plants presented to best display structures enabling morphological identification of the organism: leaves, stems, flowers, roots, or fruits (Fig.1.2b). But vouchers can also be represented by other types of specimens, the woody part of a tree preserved in a Xylarium, or ethnobiological specimens including plants parts as trade goods (fruit, seeds, bark), clothes, woven basket, paper (Albani Rocchetti et al., 2021, Salick et al., 2014).

Although the process of drying plant specimen with sheets of absorbent paper has remained unchanged since the beginning of plant collecting, drying specimens in the humid conditions of a tropical collection site is complicated. The specimens can be moist and become easily contaminated with invasive fungi or bacteria. Apply heat to speed drying and alcohol soaking and pickling to prevent fungal growth have been widely used.

Production of good herbarium vouchers that maintain most of their morphological features, particularly those that are used for taxonomic purposes can present some difficulties, such as succulent plants which are difficult to dry out, large plants which do not fit into presses, delicate tissues which lose morphological detail on pressing, and colour/texture which changes on drying. A variety of methods have been developed to deal with difficult material (Eggli et al., 1996). Once prepared, the vouchers must also be careful preserved in dark and dry conditions, and are usually treated to deter insects and fungi by freezing, or soaking in alcohol or more dangerous solutions, such as mercury (Fig.1.2b).

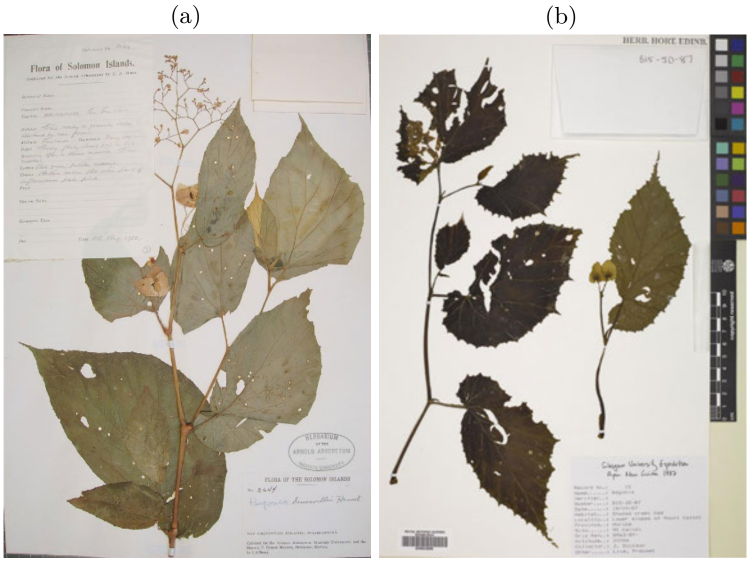


Figure 1.2: Herbarium specimens from RBGE's Papua New Guinean collection. (a) Damaged leaves in a *B. somervillei* specimen from Solomon Island (1932, RBGE herbarium). The specimen might have been damaged by incorrect handling, or insects. Residue of insect repellent (mercury or other chemical) might be found on the specimen or the capsule containing non-mounted plant material. (b) Partially annotated specimen of *B. sect. Petermannia* not including GPS coordinates (1987, RBGE herbarium).

Voucher specimens

The preserved plants are mounted for archiving as voucher specimens, a herbarium specimen representative of a species and usually collected to support a research project (Culley, 2013).

The information in the specimen itself is of little value without detail on its origin and identity. All herbarium vouchers are mounted with labels covering key metadata. The basic label includes the scientific name of the plant species, its vernacular name, the collector's name, the date of collection, the locality or geographical location of sampling, and notes about the habitat and ecological conditions. However, given different collecting practises, and the historical nature of collections the information in the label can vary widely (Fig.1.2b).

Extra information is often included on labels, such as phenotypical features, ecological conditions, distribution of populations in the landscape, and other information lost during the pressing process, colour, or smell. Some now have links to silica samples or sequence data (Bridson, 2000).

The voucher used to define a species is referred to as a type specimen, and becomes the reference against which all samples are compared for identification. Herbarium specimens are therefore the key to plant specimen identification. The type collections in herbaria are international resources that plant scientists all around the world use for the fundamental task of identifying species (Funk, 2003).

Temporal sampling with herbarium specimens

Records of plants preserved dried and described start as early as the 16th century, with the specimens collected by Brunfelds and Fuchs in Italy to support their botanical drawings (Bellorini, 2016). Even though those dried plant specimens did not survive to this day, other specimens do survive from the 17th century. The oldest herbarium specimen found yet in RBGE is a Cape Myrtle (*Myrsine africana L.*) collected by Alexander Brown, a ship's surgeon, picked on the Cape of Good Hope in 1697 (RBGE, 2020, Fig.1.3).

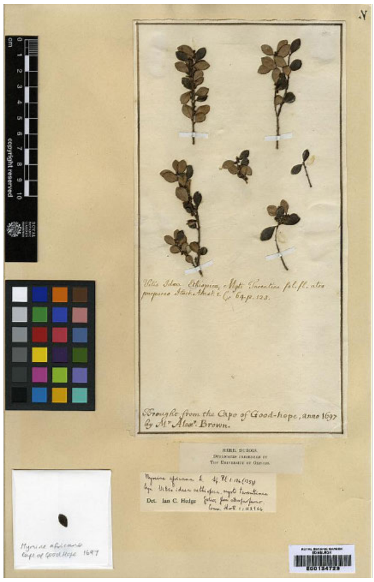


Figure 1.3: Herbarium specimen of Cape Myrtle *Myrsine africana L.* from 1697 (RBGE herbarium).

Over the next 350 years thousands of people worldwide contributed to building up today's herbarium collections. As well as scientists, this included local naturalists and collectors, herbalists and foresters, administrators, and commercial collectors, as well as collections made as mementos of travel and for artistic purposes. The contributions of different people to the generation of herbarium resources has not always been acknowledged, but the scope and richness of the collections produced by their effects is now becoming an open resource to improve everyone's life (Nordling, 2022). The very different reasons behind individual collections provides a patchy and eclectic sampling of any one species. Early samples may be primarily from close to population centres, trade routes or popular holiday destinations. There are also many gaps from times and places where collections was impossible (Fig.1.4), such as during war. Even with current professional, focussed, collecting trips, a full survey of flora across a regions is expensive, time consuming and rare. With these caveats, herbarium collections often offer the best possible overview of variation in a species across time and space.

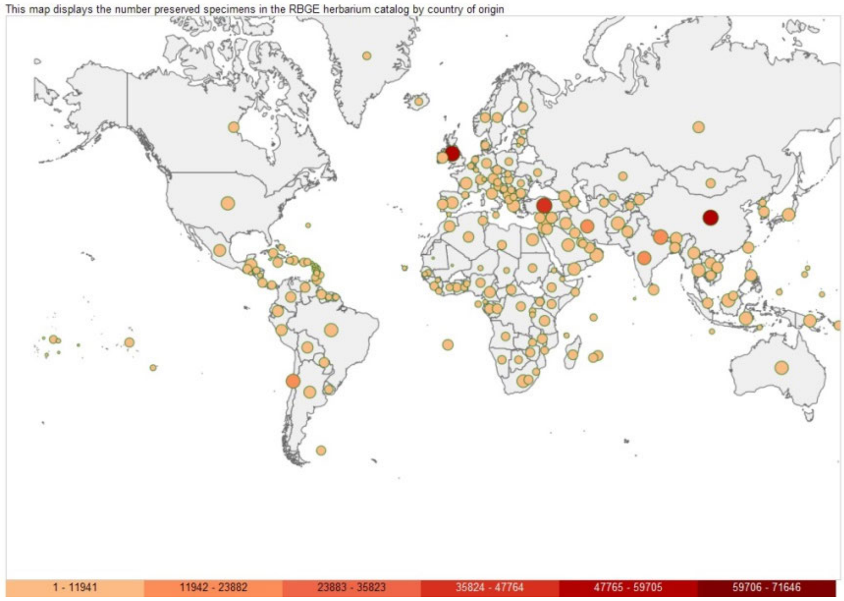

This map displays the number preserved specimens in the RBGE herbarium catalog by country of origin
1 - 11941
11942 - 23882
23883 - 35823
35824 - 47764
47765 - 59705
59706 - 71646

Figure 1.4: Location of samples digitised from the RBGE Herbarium collection

Availability of specimens and modern archiving

Access to herbaria specimens was formerly restricted to physical visits, limiting the consultation of collections to those who could travel and had permissions. Digitization of the collections has made them accessible to the larger community, and in most cases fully open access via aggregated portals like JSTOR Global Plants (*Global Plants on JSTOR* 2023), the Global Biodiversity Information Facility (*GBIF* 2023), iDigBio (*iDigBio Home | iDigBio* 2023), and more local initiative as the Reflora Virtual Herbarium (*Reflora | Kew* 2023) (Rønsted et al., 2020, Walker et al., 2022). One limitation on the digitisation process is the increasing number of specimens. The collections were already large and continue to grow at a rate that could outpace digitisation efforts. However, automated digitalization processes with Artificial Intelligence (AI) may contribute to overcoming this issue (Sweeney et al., 2018, Carranza-Rojas et al., 2017).

The digitization process include capture of images of the specimens and images of the herbarium sheet labels. Digitization workflow had previously been focused mainly on capturing specimens labels information , newer technologies as deep learning may contribute to specimens identification through the training of a neural network to extract patterns from a training set (Walker et al., 2022).

1.2 Herbarium specimens pest infestation, biocides, and contamination

Although digitisation opens up the images and the metadata to all, there is still much value in the physical specimens. Digitisation cannot capture all the detail which may be needed and advances in analysis may allow us to garner more information from a specimen than a jpeg can replace. However, keeping the specimens in good condition is a difficult task.

Physical conservation treatments and humidity levels

Environmental control is one of the key factors for specimen conservation in herbaria. Dry, cool conditions are essential to prevent the growth of moulds, but also prevents degradation of biomolecules in the specimens. Conditions of preservation have been shown to have more impact on the biomolecules in plant material than the age of the specimen (Forrest et al., 2019, Bakker, 2022). The environmental controls include control of air flow into the herbarium, temperature and humidity regulation and monitoring, and using desiccant dust on the herbarium specimens themselves (Hall, 1988).

Animal threats to herbaria specimens

Pests dwelling in herbarium facilities can potentially damage plants, voucher and other documents and contaminate the plant material (Fig.1.2a). The most destructive feed directly on the preserved plant material, and include beetles feeding on cellulose. The tobacco beetle *Lasioderma serricorne* is the most common pest in tropical herbaria, while the 'Herbarium' beetle *Stegobium paniceum* fits the same ecological niche and is resilient to drugs and spices toxic to other species. Xylariums and wooden museum specimens can be damaged by specialist wood pests such as drywood termites *Cyptotermes spp*, *Kalotermes spp* and Powder-post beetles *Lyctus spp.*

The generalists: Silverfishes *Liepisma saccharina*, booklices *Liposcelis*, spider beetles *Ptinus tectus*, and cockroaches *Blattela germanica* have all been noted as serious herbarium pests (Hall, 1988, Bridson, 2000).

Deep freezing decontamination

Deep freezing is the best general method of decontamination for herbarium specimens entering the herbarium (Bridson, 2000). The incubation time needed to kill insects and their egg is defined as Operating time, composed of the Chill-time the time necessary for the samples to reach the adequate temperature, and the Kill-time the delay needed to effectively kill the pests. The temperature needed to kill the insects is -18°C with an average of 17 hours for the Chill-time and between 3 and 9 hours required for the Kill-time, depending of the species of insect targetted (Hall, 1988). Eggs of several species might survive a temperature of -18°; for those a temperature of -30°has been recommanded for decontamination (O'Connor, 1991).

Chemical treatments

Chemical methods to repel herbivores include poisoning the herbarium specimens themselves, using insect repellents or insecticides in the herbarium facilities to present invasion, anti-fungal reagents in the case of long exposure to high humidity, or fumigation of the facilities and regular use of decontamination cabinets (Bridson, 2000). Whilst protecting the herbaria specimens from pests and fungi, the chemicals used to coat specimens directly may complicate DNA extraction and genomic library preparation. Poisons applied on the specimens as long-term solutions to repel herbivores, commonly included *mercuric chloride*, a very toxic reagent used since the 19th century and still in use until the end of the 20th century (Clark, 1986). Specimens could be dipped (a process called kyanizing for wood specimens) or painted with the reagent that did not alter the physical appearance of a specimen. However, it can blacken paper and leave crystalline deposits like hairs on the specimens. It is corrosive to human skin and sublimates producing a vapour which can destroy mucous membranes (Hall, 1988, Briggs et al., 1983, Havermans et al., 2015, Cabassi et al., 2020). A more recent solution used in herbarium conservation is *Lauryl pentachlorophenate* (LPCP). It is assumed as a safer to handle than *mercuric chloride*, but requires the specimen to be totally immersed in the solution, which can cause blurring of labels, and is not as long-lasting contrary as*mercuric chloride.* The presence of these chemicals might affect subsequent molecular work, especially the preparation of genomic libraries preparation as the DNA polymerase α, involved is inhibited by the action of a mercury salt (mercuric acetate) (Williams et al., 1987).

1.3 Museomics, and biomolecules contained in herbarium specimens

Conservation of aDNA in herbarium specimens

The conditions of preservation of herbaria specimens allow good preservation of biomolecules contained in the plant material. DNA, lipids, and proteins are still available in herbarium specimens and can be used to trace the plant's genotype, evolutionary history, and even pathogen infections (Cappellini et al., 2018, Phillipson, 1982).

DNA does degrade in herbarium specimens. The long threads of DNA breakdown into fragments of hundreds of base pairs, or shorter, and become contaminated with DNA from bacteria and fungi growing on the decaying plant material. It also sustains damage comparable to ancient DNA (aDNA) found in coprolites, archaeological remains, or sediments. It has been estimated that the DNA from plant material contained in herbarium specimens sustain six times more degradation than in ancient bones. This is likely due to the drying and preserving process rather than breakdown during herbarium storage (Weiß et al., 2016). The molecular and bioinformatic tools and protocols developed for aDNA analysis are appropriate for use on herbarium specimens as well (Malaspinas, 2016).

Methods of aDNA sampling

Even though the presence of DNA in ancient plant seeds was suggested as early as 1973 (Hallam et al., 1972), one of the first attempts to isolate genetic material from dried preserved plant tissue dates from the 80s, with the isolation of rRNA fragments from Egyptian cress seeds (*lepidium sativum L.*) recovered from an archaeological site in Thebes and dating from 1400 years B.C (Rollo, 1985). This study used RNA radio-marking and hybridization with cress sequences. The recovery of rRNA is more likely than good DNA for ancient specimens as its secondary structure and association with protein make is a more stable than genomic DNA (Livio et al., 1990). This attempt was followed by the recovery of aDNA fragments from pre-Columbian maize seeds using PCR to amplify the tiny amount of genetic material recovered (Rollo et al., 1991). Subsequently, ancient DNA extraction and analysis has been widely applied to herbarium specimens, first to detect a limited number of genetic markers, and as the discipline progressed, as genome skims and target capture to recover large portions of the plant genomes (Fig.1.5).

1.4 Target capture method on Herbarium specimens DNA

Target capture used biotinylated RNA baits to select matching sequences from a genomic DNA library. The capture sequences are then amplified and sequenced. The reduction in complexity from a heterogeneous DNA prep to sequences corresponding to hundreds of loci allows high depth coverage and confident assembly of consensus sequences. The hybridisation affinity of the bait is usually enough to capture sequences from a wide range of related species, and bait sets which capture across extensive lineages have been produced and are now widely used (Lemmon et al., 2013, Barrett et al., 2016, Brewer et al., 2019).

As this method takes libraries produced from fragments a few hundred base pairs long as input it can work on both fresh material and the fragmented DNA extractable from museum material, enabling recovery of useful sequences even with reduced material available (Villaverde et al., 2018, Kates et al., 2021, Michel et al., 2022). It can be used in combination with genome skim derived chloroplast sequences for phylogenetic reconstruction (Weitemier et al., 2014) with high-coverage. But as stand-alone it has proved efficient to capture nuclear sequences from old and degraded DNA sufficient for phylogenetic analysis or even population genetics (Hart et al., 2016, Kates et al., 2021, Vatanparast et al., 2018, Brewer et al., 2019, Couvreur et al., 2019, Stull et al., 2020). Analysis of the factors affecting capture efficiency in helps in design of sampling, and advances in preparation of libraries from ever-smaller amounts of DNA continue to make this a feasible approach for even unpromising samples (Brewer et al., 2019, Hart et al., 2016, Forrest et al., 2019).

Microbiology of herbaria specimens

The total DNA of a herbarium specimen can be seen as more of an environmental sample than a sample from a single individual. Some of the first studies of herbarium specimens were focussed not on the the plants but on the pathogens they carried. Pathogen DNA (or RNA in the the case of some virus) is present in high copy number in infected material, and this greatly simplifies the analysis compared to analysis of nuclear DNA of the host plant (Meineke et al., 2019; Rønsted et al., 2020, Bieker et al., 2020).

The microbiome of herbarium specimens can also provide data on the plant's demography and ecological context. Fewer pathogens have been detected in herbarium specimens compared to silica-dried modern specimens, probably due to the shorted DNA fragments of historical specimens, which complicate the identification of pathogens on the specimen (Bieker et al., 2022). One difficulty with this approach is that pathogens identified in an historical specimens could originate from the herbarium facilities during storage of the specimen rather than the original ecological microbiome of the plant (Bieker et al., 2020).

1.5 Ancient DNA analysis and evolutionary history

The use of herbaria specimens to trace evolutionary, demographic, or domestication events is increasing thanks to technical progress that enables better investigation of ancient genetic material (Fig.1.5).

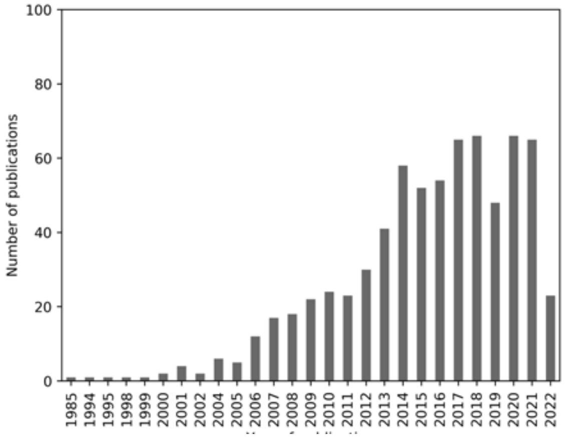


Figure 1.5: Plant museomics publications over time.

Phylogenetic inference versus temporal sampling

Usual methods of investigating evolutionary processes are based on inference from patterns of genetic variation. Thanks to recent developments in aDNA recovery, the alleles frequency of historical specimens can be directly estimated and their evolution over time observed if a time series of specimen is accessible (Gutaker et al., 2017). This field of research allows reconstruction of the evolutionary history of extant populations descended from populations sampled as herbarium specimens. It uses a temporal distribution of samples, which provides information about the evolutionary history and time scales associated to a plant population (Zedane et al., 2015). Although limited to the very recent past, and constrained by the sample bias of herbarium collections, there is still the potential to gather valuable information from such analyses.

Value of including museum samples in studies

Museomics studies are the only route to studying genetic information from extinct populations (Paer et al., 2016; Zedane et al., 2015; Humphreys et al., 2019), and make it much easier to sample populations difficult to collect in the wild (Zedane et al., 2015), or from rare species (Silva et al., 2017). This allows genetic studies a much more complete view of variation across a lineage and is particularly valuable in population studies examining issues of rarity (Gutaker et al., 2017). Although focussed on recent dates this is ideal for analysis of domestication, and the impacts of the Anthropocene (Ash et al., 2017, VanAndel et al., 2022).

Population genetic studies are mostly based on a coalescent reconstruction of ancestral state from a single point in time. The assumptions of ancient evolutionary change based on contemporary specimens can be biased, as recent events such as gene flow or changes in population size are difficult to separate from signatures of selection (Dehasque et al., 2020). Museum collections specimens can provide a better estimation of a diversity within a species or within a population, as well as showing the progressive selective sweeps resulting from a domestication process and to quantify them. A series of soft selective sweep have been detected in Andean beans domestication from 2500 years to 600 years ago, followed by a sudden loss of diversity in modern cultivar (Trucchi et al., 2021) Similar signatures of selection have also been detected as well in maize (Beissinger et al., 2016). Use of only modern cultivars would have missed the presence of and patterns in earlier variation. With the capacity to include present and historical samples we can observe the range expansion, different speeds of radiations, and introgression of invasive plants in a new biome. This has been done for many years using plant record to track an invasion, but now we can also examine the evolution of genomic architecture through introgression and selection during this process. A study on *Ambrosia artemisiifolia* (common ragweed) that has been introduced to Europe in the late 19$^{\text{th}}$ century, has shown a signature of introgression from European natives *Ambrosia* species closely related along with signature in defence against plants pathogens genes, possibly granting the invader with an immunity against microbial threats (Bieker et al., 2022).

1.6 The *Begonia* genus

With more than 2000 species, *Begonia* is one of the largest plant genera. The group has a pan-tropical distribution, being found in Central and South America, Africa, and Asia, being absent only from Australian tropical forests (Neale et al., 2006). Phylogenetic studies indicate that the group is originated from the African continent, counting fewer species but displaying larger morphological diversity (Forrest, 2000, Plana, 2003).

The *Begonia* populations are usually small with low pollen dispersal and restricted gene flow between populations (Matolweni et al., 2000, Hughes et al., 2002c). It might be one of the factors responsible for the high speciation rate within the genus and the cause of the high degree of endemism observed in the typical *Begonia* population (Hughes et al., 2003, BPG et al., 2022, Brennan et al., 2012). *Begonia* evolutionary history, population structure, and endemism patterns are discussed further in Chapter 3.

Many researches on *Begonia* have been driven by horticultural interest, as a lot of species within the genus hybridise readily, with short generation time and high seeds production (Neale et al., 2006). Notable examples are the crosses made between Arabian and South American species, yielding the very successfully *Hiemalis* and *Cheimantha* groups of hybrids, nicknamed Winter-flowering *Begonia*. The group has been intensely researched as well in reason of their large array of different morphological shapes (Fan, 2023, Li et al., 2022). Combination of easy hybridization and very different morphologies makes *Begonia* a good candidate for genetic studies supported by segregation genetics

(Twyford et al., 2014b).

While studies have focused on *Begonia* high level of speciation or evolutionary history, the question of genetic isolation and impact of inbreeding depression on populations is still mostly unexplored. Using the segregation patterns of alleles observed in a mapping population from South American *Begonia* (Twyford et al., 2014b) and Hyb-Seq datasets from Papua New Guinea (Wilson et al., 2020) and Arabia, we will analyse the genome of individual plants, assess their recent demographic history, and detect markers of inbreeding depression.

1.7 Aims of this thesis

This thesis aims to build on advances in herbarium genomics to set up a pipeline allowing estimation of the genetic health of species from single specimen.

In this document we are referring to genetic health as the biological fitness of a population, maintaining a genetic variation or gene pool sufficient to avoid inbreeding depression, the detrimental effects of inbreeding.

Often single specimens are all that is available for rare species, and information on size of populations, range and tolerance are difficult to obtain. By using the genetic data from that single specimen levels and patterns of heterozygosity can be obtained which will allow estimation of levels of inbreeding and the likely resilience of the species to change. We will use this data to answer a question about how robust many of the 2,000+ *Begonia* species are.

Hypothesis: Mutational meltdown is a driver of extinction in Begonia.

- **Test**: Species with smaller ranges, highly endemic species, species now extinct or rare in the wild are highly homozygous and this homozygosity appears recent - uniform across chromosomes.
- **Aim**: Establish a pipeline to call population genetic parameters from herbarium samples of Begonia.
- **Experiment 1**: Hybrid capture in mapping population (fresh, silica and herbarium samples) to confirm reliability of metrics called.
- **Experiment 2**: Hybrid capture in wild populations and herbarium samples of *B. socotrana* and *B. samhaensis* to confirm metric agree with previous work using microsatellites.
- **Experiment 3**: Use of Hybrid capture from a range of species to establish frequency of different patterns across *Begonia* species, and determine what proportion appear at risk.

Chapter 2

Palaeobotany

Thibauld Michel[1], **Michael D. Martin**[2], **Catherine Kidner**[1]

1 Royal Botanic Garden of Edinburgh, University of Edinburgh, Edinburgh, United Kingdom
2 NTNU University Museum, Norwegian University of Science and Technology, Trondheim, Norway

Thibauld Michel __________________
Michael D. Martin
Catherine Kidner

Citation: Michel T, Martin MD, Kidner C (2022) Chapter 21. Palaeobotany. In: de Boer H, Rydmark MO, Verstraete B, Gravendeel B (Eds) Molecular identification of plants: from sequence to species. Advanced Books. https://doi.org/10.3897/ab.e98875

2.1 Identification

2.1.1 The evolution of ancient plant DNA analysis across time

The study of ancient plant remains was historically limited to morphological studies, palaeontology being the primary field of study of past organisms. However, since the 1980s, genetic analysis of biological matter within fossils has become increasingly informative, thanks to the development of new molecular methods such as polymerase chain reaction (PCR) and a revolution in sequencing methods that has enabled high-throughput sequencing (HTS). Since the first identification of aDNA from extinct species in 1984 (Higuchi et al., 1984), it is now possible to identify many organisms, including plant taxa from only microscopic plant remnants or even short fragments of DNA bound to a substrate using modern molecular tools. These new methods allow us to explore their recent evolutionary history, study the ecology of palaeoenvironments, and understand population relationships, migration, and domestication processes. While the domain of palaeogenetics is limited to the study of a few genetic markers, the establishment of new DNA isolation techniques, the generalization of HTS analysis, and more robust computational methods have enabled the analysis of longer DNA fragments and in the last couple of decades shifted the domain towards the field of palaeogenomics with the analysis of full plant genomes (Mitchell et al., 2021).

2.1.2 Ancient plant DNA in historical remains

In the context of palaeogenetics, ancient DNA (aDNA) is DNA from long-deceased tissues preserved by conditions allowing DNA survival. Despite appropriate preservation conditions, aDNA is usually degraded by biotic or abiotic processes. Though often damaged, it can carry valuable historical information (Schlumbaum et al., 2008). Ancient DNA from wild plants can be used to reconstruct the evolutionary and demographic histories of populations to trace ecological and climatic changes (Hofreiter et al., 2001). Ancient plant DNA from anthropogenic sources can be used for studying the processes of plant domestication, generating insights into past plant usage, agricultural techniques, and the migration patterns of ancient human societies (Kistler et al., 2014; Trucchi et al., 2021). The sequences used for most plant aDNA studies are derived from the nuclear and organellar genomes, and are quite often the same markers typically used for plant identification or studies of evolutionary history.

Markers from plastids, in particular from chloroplast DNA (cpDNA) are usually favoured as they are markers inherited uniparentally (usually maternally in angiosperms and paternally in gymnosperms) and therefore represent a single haplotype transmitted from a single parent to the offspring, and facilitate the identification of lineages. Furthermore, their relative high copy number facilitate the recovery of cpDNA of small or damaged specimens (*Molecular Ecology, 3rd Edition | Wiley* 2020). Genes transfer from the plastome to nuclear and mitochondrial genomes complicates the analysis as the mutational rate differs between in the nucleus and in other organelles (Kistler et al., 2014; Wales et al., 2016). Compared to chloroplasts markers, mitochondrial markers are considered less informative, and have rarely been used in plant aDNA due to their relative low level of copies and to an higher rate of rearrangement (Schlumbaum et al., 2008, *Molecular Ecology, 3rd Edition | Wiley* 2020). However, they are not the only target to consider as traces of ancient RNA have also been amplified in cress seeds by hybridization and later sequenced in maize (Fordyce et al., 2013; Gnirke et al., 2009; Rollo, 1985). Other targets previously detected include epigenetics patterns such as methylation in response to pathogen infection, and small plant RNA (miRNA) as a response to environmental stress (Smith et al., 2017, Smith et al., 2014) in barley. Bacterial and viral DNA can also be amplified from ancient plant material (Bieker et al., 2020; Kistler et al., 2014).

2.1.3 Historical and archaeological aDNA challenges

Several difficulties are inherent to working with aDNA from plant specimens: the complexity and variability of the genome, aDNA damages, and potential contamination increase downstream analytical difficulties. The combination of often very low aDNA concentrations with the amplification power of PCR dramatically increases the probability of amplifying contaminating modern DNA. Specialised methods and laboratory procedures have been established to reduce the risk of contamination. These include: the use of positively pressurised clean laboratory facilities dedicated to aDNA work, the replication of experimental works in different institutions, and the use of biomarkers for prediction of DNA survival such as mitochondrial DNA (mtDNA) detection, aDNA damage patterns, and detection of associated remains (Capo et al., 2021; Cooper, 2000). Another complication in the analysis of plant DNA is its variability. The presence of different organelle genomes (plastid and mitochondrial) as well as the interspecific differences in ploidy level and chromosome size can complicate the alignment of sequencing reads to a reference sequence (Kapusta et al., 2017; Kistler et al., 2020). Target capture is one strategy that can be used to reduce this complexity, even in sequences that are heavily degraded and/or contaminated (Parducci et al., 2019).

2.1.4 Sources of plant DNA

Macrofossils

Macrofossils are defined as fossils that are observable without magnification, and in the case of plant-based studies, they are ancient preserved tissues found in archaeological or sedimentological contexts. aDNA can be extracted from macroscopic plant remains such as leaves, needles, bud scales, wood, or seeds. However, individually-based approaches on plant macrofossils are scarce and most of the studies focusing on plant DNA are based on metabarcoding using sedimentary DNA (sedaDNA) material (Jaenicke-Despres, 2003; Rollo et al., 2002; Schwörer et al., 2022). The scarcity of macrofossils in plant studies can be explained by the difficulty of aDNA recovery in preserved plants, which can be due to the low-level of endogenous DNA in plant remains, high amounts of contaminant microbial DNA, and aDNA specific damages (Eleanor Green et al., 2017). Regardless of this limitation, macrofossil DNA studies have some advantages: they can be directly dated without the use of proxies, they represent local species in contrast to pollen studies where pollen grains could be dispersed over large distances, and DNA from a single analysis can be authenticated from its aDNA specific damage patterns (Schwörer et al., 2022).

Charred and desiccated

A very common archaeological plant material is charred remains. One example is superficially burnt seeds in hearth remains found in ancient settlements. Molecular identification of even lightly charred remains is however challenging since the DNA is often very fragmented and contaminated (Palmer et al., 2012). Target enrichment has not yet been able to overcome this issue (Nistelberger et al., 2016). Thus, charred plant remains are primarily identified using morphological analysis. In contrast, dessicated samples are often suitable for molecular analysis. Dessicated samples are typically found in dry environments such as caves, shelters formed by rock features (well suited for long-term food storage), or deserts. Desiccation can limit DNA degradation, and plastid and mitochondrial DNA from sunflower seeds as old as 3,100 years old has been successfully recovered (Kistler et al., 2011; Mascher et al., 2016; Swarts et al., 2017; Wales et al., 2019).

Waterlogged

Biological remains preserved under waterlogged anaerobic conditions may also contain sufficient aDNA for molecular identification. Lakes and marine sediments can provide sedimentary DNA (sedaDNA) from plant remains and pollen grains found in different strata of core samples. They can be used to reconstruct past ecological diversity. Microorganism communities can as well be a source of aDNA. For example, diatoms are commonly used bioindicators for assessing the biological composition (trophic state) of a lake since their morphology is highly sensitive to the surrounding environment (Ibrahim et al., 2021). The taxonomic diversity of diatoms found in the sediments of glacial and thermokarst lakes has for instance been linked to lake type and age, environmental changes, and surrounding vegetation (Huang et al., 2020). Cyanobacteria, which are sensitive to temperature, can be used as a biomarker for detecting the effects of climate change by studying their population diversity. The microbial communities of Lake Constance (Central Europe) for instance, including microbial eukaryotes, diatoms, and cyanobacteria, have been used as bioindicators for both biotic and abiotic changes due to warming by studying the phylogenetic distance of microbial communities, and their geographic and temporal change of diversity (Monchamp et al., 2019). Waterlogged remains can be found in the context of archaeological studies. Wells, latrines, ditches, and pits can result in anaerobic conditions. DNA from grape seeds from the Iron Age have been sequenced successfully with Hyb-Seq, and it was shown that the grapes are related to present-day West European cultivars, which provides evidence that there has been 900 years of uninterrupted vegetative propagation of the crop (Ramos-Madrigal et al., 2019). Gourd rinds, squash seeds, and oak wood thousands of years old have provided high-quality aDNA using target-capture methods, or using plastid or mitochondrial DNA (Wagner et al., 2018). This has led to a correction on the view of how gourd domestication happened by showing that the pre-Columbian bottle gourds originated from Africa and reached Latin America via the Atlantic by ocean drift (Kistler et al., 2014). Other studies have shown a link between the Holocene megafauna extinction and the decline of wild *Cucurbita*, while domestic lineages thrived because of cultivation (Kistler et al., 2015).

Mineralized and embedded

Mineralized samples or those embedded in resin or fossilised in amber are both potential sources for aDNA, though the high probability of contamination, extreme fragmentation of the material, and non-reproducibility of the results have led some authors to strongly discourage aDNA analysis from amber-preserved fossils (Modi et al., 2021). However, partially mineralized remains (subfossils) less than 10,000 years old can still contain biological material and are potentially a source of biomolecules including DNA (Wagner et al., 2018). Recently developed methodologies for specimen extraction from amber that reduces contamination have enabled the recovery of insect DNA up to 3,900 years old from copal, a precursor to amber (Peris et al., 2020). This leads to the possibility that these sample types may be sources for plant aDNA in the future.

Microfossils

Microfossils can be found in any environment, including in humid conditions and tropical zones where macrofossil preservation is rare. These include pollen, starch grains, and phytoliths. Plastid aDNA obtained from pollen grains is very often endogenous, and its amplification has previously established the first genetic link between extant and fossilised Scots Pine specimens from post glacial lake sediments in Sweden (Parducci et al., 2005). Phytoliths enable radiocarbon dating, even though no aDNA has been isolated from them so far (Elbaum et al., 2009). Yet, they are hypothesised to be a potential source of aDNA (Grass et al., 2015).

Sedimentary DNA

Sediments found in lakes, temperate caves, permafrost, and ice cores can retain plant aDNA for thousands, and in some cases, millions of years (Kirkpatrick et al., 2016). Sedimentary DNA may be used as a proxy for the reconstruction of the palaeoenvironment, even though other plant structures have been destroyed (Willerslev, 2003). Metabarcoding to amplify short amplicons of cpDNA is by far the most commonly used approach (Capo et al., 2021; Parducci et al., 2017; Rijal et al., 2021). Shotgun sequencing has only been used sparsely because of the lack of reference libraries (Slon et al., 2017), but as full genome reference databases are being built, this method could improve the ability to investigate lake sedaDNA (Parducci et al., 2019). More recently, shotgun metagenomics was used for retrieval of whole plant genomes from archaeological settlements and marine deposits (Parducci et al., 2019; Pedersen et al., 2013; Slon et al., 2017). However, sedaDNA taphonomy for sedimentary material is still a subject to explore, as the conditions that lead to its preservation are not yet clear (Kistler et al., 2020). SedaDNA provides a broad understanding of the past environment, climate, and ecology of the palaeosol studied. It can also provide insights on the movement and cultivation of plants by Neolithic populations and their social network in absence of other archaeological evidence (Brown et al., 2021; Smith et al., 2015). sedaDNA from lake sediments has been used to reconstruct ancient plant vegetation and to assess the impact of anthropogenic activities on the palaeoenvironment. For example, the impact of cattle grazing on deforestation dynamics during the Late Iron Age and Roman period has been demonstrated by using a metabarcoding approach on sediment samples from a subalpine lake (Giguet-Covex et al., 2014). sedaDNA can also be used to study the impact of climatic changes on plant biodiversity and help prioritise conservation management. A research project using metabarcoding of lake sediments was able to show that an heterogeneous mountain landscape served as a refugium for arctic-alpine plants in a warm climate (Clarke et al., 2019). Another study on Arctic Canada lake sediments gave clues about the effect of the rise in temperature during the Last Interglacial period (LIG) on plant population dynamics. Previous attempts to reconstruct the LIG palaeoclimate with climate modelling based on the simulation of atmosphere, sea, and ice circulation have yielded inconsistent results (Otto-Bliesner et al., 2013). Comparison of the model results with sedaDNA vegetation reconstruction suggests that models underestimated the magnitude of Arctic warming during the LIG. This discrepancy could be due to the lack of vegetation-related feedback such as arctic greening in the models, but are observable in sedaDNA records (Crump et al., 2021). We can improve modelling of future climate change effects on plant diversity based on these studies that inform how plant richness has evolved in reaction to previous episodes of climate warming. Several environmental changes that might have been overlooked such as arctic amplification or arctic greening can be studied with sedaDNA (Clarke et al., 2019; Crump et al., 2021; Liu et al., 2021). The impact of sea ice on plant colonisation of Iceland during different periods of the Holocene suggests that the melting of the ice sheet due to future warming might limit plant distribution rather than favour it (Alsos et al., 2021). SedaDNA studies are furthermore more robust than pollen-based methods for detecting plant richness, and deliver taxa diversity with more resolution (Crump et al., 2021). As an example, a study based on multiple-sites lake sedaDNA analysis and pollen records shows the steep increase of plant richness in the early Holocene in northern Fennoscandia (Rijal et al., 2021). The causes of this increase are the higher level of available soil nutrients and the lower level competition just after deglaciation. However, the pollen records did not match the sedaDNA findings that taxonomic richness has continued to increase even after climate stabilisation. These discrepancies are due to problems affecting pollen records such as overabundance of a few taxa and under-representation of others (swamping). In contrast, sedaDNA provides higher taxonomic resolution, lower swamping effect, and represents local plant groups. The same observations can be done using sedaDNA extracted from permafrost, as presented in a study encompassing 50,000 years of megafauna diet and arctic vegetation history from samples collected across the Arctic. While pollen-based reconstruction showed a majority of graminoids in unglaciated Arctic during the Late Glacial Maximum, the metabarcoding approach has revealed a forb-dominated vegetation (Willerslev et al., 2014).

Palaeofaeces

Ancient faeces, though relatively uncommon, are a rich source of biomolecules and palaeodietary information that can be related to demographic, ecological, and climatic changes in the locations in which they are found (Eleanor Green et al., 2017). Genetic identification from plastome barcoding can also provide evidence missing in classic macroscopic morphological analysis (Poinar, 1998; Gilbert et al., 2008; Rollo et al., 2002). Recent approaches using shotgun metagenomic methods provided identification of plants in ancient faeces as well as information on the gut microbiome, parasitic worms, and the actual identification of the defecator (Boast et al., 2018; Wood et al., 2016).

2.1.5 Bioinformatic tools and challenges

The analysis of an aDNA dataset is complicated by postmortem DNA degradation that leads to short fragments, specific nucleotide substitution patterns, and overall low DNA yields (Briggs et al., 2007). These difficulties will affect subsequent evolutionary inferences and population genetics studies. Consequently, numerous tools have been developed to detect and quantify nucleotide substitution, deletion, and DNA fragmentation. The initial alignment step with a reference genome during bioinformatic analyses is already affected by aDNA pattern of damages, which can increase the apparent error rate and lower the alignment accuracy. Subsequent steps in variant calling of genetic markers can be complicated by the high mapping error rate and low coverage (Bilinski et al., 2018). Strategies have been developed to prevent bias resulting from low coverage. This can include random sampling of a single read at each locus of interest (Bakker et al., 2016; Kistler et al., 2018) and genotype likelihood estimation (Korneliussen et al., 2014). More specific tools have also been designed to solve the issue of identifying the ancestry of unknown samples with a low coverage dataset using multidimensional scaling (MDS) methods (Malaspinas et al., 2014; Ramos-Madrigal et al., 2019). Issues related to aDNA specific damage patterns can be prevented using strategies such as only considering transversion polymorphisms, using statistical algorithms to rescale the base quality scores before variant calling (Jónsson et al., 2013), or soft-clipping fragment ends to avoid deamination sites

(Kistler et al., 2018). Tools for rescaling base quality scores have also been implemented into bioinformatic pipelines that are dedicated to aDNA alignment (Schubert et al., 2014).

2.2 Applications

2.2.1 Evolutionary studies

The evolutionary history of a species or a population can be established based on genomic inference from modern samples, providing clues about the evolutionary processes that form the basis for present genomic variation. However, allelic patterns in contemporary specimens are shaped by a range of demographic events, including changes in population size, gene flow, and hybridization events. These may be due to very recent events, and do not necessarily represent the lineage's deeper evolutionary history. A time series of samples can provide greater resolution in a genomic analysis and resolve phylogenetic questions. It can also detect recent demographic events such as population bottlenecks and provide chronological estimates for these events without using a molecular clock. Allele frequencies can be directly estimated for each time point and used to estimate the strength of selection pressure during that period (Malaspinas, 2016). This approach can be used to distinguish between different selection processes and to establish their tempo across time (Dehasque et al., 2020). The Dramatic global warming and extinction events that occurred during the later Anthropocene coincided with the active collection of specimens for museums and herbaria (Bieker et al., 2018). Genetic analysis of collections provides a detailed understanding on how human activity has shaped the evolutionary fate of many organisms. Modern techniques also allow us to recover information on extinct species. One example is the genus *Hesperelaea* from the *Oleaceae* family, which was collected once 140 years ago in Mexico, and is now extinct. A genomic analysis of this *H. palmeri* specimen traced its American lineage, the date of its divergence, and helped to characterise its endemism (Zedane et al., 2016). Positive selection can also be detected in contemporary specimens using statistical tools such as coalescence, population differentiation (Fst), and linkage disequilibrium. Selection pressure, however, can be conflated with demographic change or background selection. Specific methods have been developed to detect positive selection on a polygenic trait using an admixture graph to represent the admixture events relating different populations through time (Racimo et al., 2018). Purifying selection or negative selection can be detected in present-day specimens as signals of reduced genetic diversity. However, similar signals can be caused by demographic events such as population bottlenecks or background selection (Henn et al., 2015). Again, using a sample time series, these signatures can be disentangled by considering regions with lower recombination rate where selection has more impact (Murray et al., 2017). Therefore, loci located in regions with low recombination rates and lower genetic diversity are likely to be a signature for selection rather than past demographic events. Genetic maps and good understanding of genome dynamics in the target species is thus key to accurate inference of selection. Balancing selection is more difficult to detect since it affects narrow genomic regions on a short timescale. This can be mistaken for positive selection, demographic events, or introgression (Fijarczyk et al., 2015). For these reasons, methods using contemporary specimens have low statistical power. A time series of samples can help detect alleles under balancing selection as their frequencies are maintained over time at frequencies higher than expected by the Hardy-Weinberg equilibrium. Although this method is limited by the number of specimens available to calculate genome frequencies.

2.2.2 Tracing domestication

All current crops are the products of single or repeated domestication events starting less than 12,000 years ago from the ancestral wild species (Kistler et al., 2015; Larson et al., 2014). Understanding the geographical origin and the ancestral lineages of domesticated species during the Holocene and the subsequent spread of the cultivars are central questions for different domains such as archaeology, anthropology, and ecology. Archaeobotanical remains can be arranged in a time series to study the evolution of domestication over time and space. They can indicate the number of times that domestication events occurred and their location, the pace and stringency of anthropogenic selection, introgression with wild relatives and between different cultivars, and be used to determine the date of these events (Brown, 1999). Molecular methods have made an increasingly large contribution to the field of archaeobotany. Starting with simple genetic analysis for taxonomic identification to supplement morphological examination, the field has rapidly progressed following advances in high-throughput technologies in archaeogenomics. Methods such as shotgun sequencing have enabled genome-wide studies, exploring in detail the genome of domesticated plants and analysing the genome-wide rearrangements that occurred during this process (Palmer et al., 2012). As both a key crop and a genetic model organism deeply studied for over 100 years, a wealth of domestication studies have been conducted on maize, revealing a detailed picture of evolution. Molecular analysis of palaeobotanical remains continues to provide new information on maize evolution, and PCR-based studies have identified the likely geographic region of its original domestication in Mexico, and traced its dispersal across Central America and South America (Kistler et al., 2018). The target capture method, or Hyb-Seq has been used to confirm and refine models for maize domestication over time mediated with progressive introgression from wild relatives (daFonseca et al., 2015). A recent study on maize domestication and diversification in South-America based on the genomes of present-day and ancient American maize cobs has shown that maize had a stratified mode of domestication that started with a large Mesoamerican gene pool that was partially domesticated. This was followed by dispersal to different locations in which the sub populations become reproductively isolated by different selection pressures (Kistler et al., 2018). Wheat domestication has not been studied as extensively as maize, but modern genome-wide studies on emmer wheat chaff found shared haplotypes between 3,000-year-old Egyptian emmer wheat from museum collection and modern emmer wheat, including domestication loci as two Quantitative Traits Loci (QTLs) related to grain size and seed dormancy. Although several haplotypes present in historical specimens are absent from modern emmer, similarities between museum specimens and Arabian and Indian emmer landraces suggest an early South-Eastern dispersal of ancient Egyptian emmer (Scott et al., 2019). Bottlenecks are a common feature in the domestication process and have also been revealed from ancient plant material in beans. One of the symptoms of a bottleneck event in the demographic history of a lineage is genetic erosion, the loss of allele diversity in a population due to genetic drift and inbreeding caused by the bottleneck event. This effect was found in the case of the Andean bean domestication, which was likely triggered by stringent varietal selection (Trucchi et al., 2021). In this study, ancient bean genomes dated between 600 and 2,500 years ago showed ten times more heterozygosity than modern genomes, despite that the set of genes that characterise the domestication had already been selected. It is likely that initial improvements in common beans occurred via soft sweeps rather than under strong selection pressure, while selection strategies in recent centuries produced further improvement at the cost of genetic erosion (Trucchi et al., 2021).

2.2.3 Phylogeography

Climatic and environmental changes can be responsible for major shifts in species' geographic distributions. For example, the glaciation cycles over the past 2.4 million years have restricted some species in separate refugia, often resulting in a loss of allelic variation that persists after the species' expansion out of the refugium. Phylogeography allows studying the history of geographic distribution of genealogical lineages using population genetic tools to detect the changes in genetic variation caused by historical events such as migration and dispersal (Cruzan et al., 2000). In contrast to studies of modern populations using selection inference from a single time point, aDNA studies including multiple time points can show the shift of alleles before and after periods of environmental or demographic change, providing information about the selection coefficient of the event (Bank et al., 2014). Early plant phylogeography studies were based on plastid DNA (pDNA) sequencing methods, as a study of the distribution and circumpolar migration of saxifrage, suggesting the possibility that plant refugia were located in the Arctic (Abbott et al., 2000). Later studies used DNA fingerprinting, such as amplified fragment length polymorphism (AFLP), in addition to pDNA to disentangle signatures of hybridization due to isolation in a refugia and postglacial migration for two species of Birches (Eidesen et al., 2015). More recently, a target capture method has been used on lake sediments to recover the complete *larix* chloroplast genome and study its dynamics at population level (Schulte et al., 2021). Another recent study has used shotgun sequencing to analyse Ice Age algal populations from lake sediments. It has enabled the mapping of chloroplast and mitochondrial genomes to reconstruct the genomic variation of the lake populations (Lammers et al., 2021).

2.2.4 Palaeoecology

Ancient DNA studies can unravel the ecological past and temporally explore the adaptation mechanism and interactions between organisms. This can include processes such as convergent evolution of different species in a similar environment, present plant adaptations due to standing or de novo mutation in the evolutionary history of a species, or metagenomics of a aDNA specimen to reveal the dynamics of plant pathogens (Bieker et al., 2020; Kistler et al., 2020). Innovations in shotgun metagenomics have increased the possibilities for using sedDNA analysis for reconstruction of past vegetation with higher taxonomic resolution than with pollen DNA barcoding (Bjune et al., 2021; Clarke et al., 2020), and they can detect more taxa in a single sample than macrofossils (Alsos et al., 2016). Provided that an appropriate reference library is available, minimal sampling can enable the identification of hundreds of different taxa in a few samples, giving an estimation of species diversity. This information allows reconstruction of the palaeoenvironment and its biodiversity change over time (Anderson-Carpenter et al., 2011). Some limitations do however remain. SedaDNA is preserved in lake environments since the stable temperature conditions can conserve DNA. However, sampling can be challenging in these areas. There are also major challenges in detecting species that are rare or have a low biomass. Additionally, the taxonomic resolution provided by sedaDNA is variable in function of the method used. While metabarcoding sedaDNA almost always provides higher resolution than direct pollen analysis (Clarke et al., 2020; Sønstebø et al., 2010), the reference library to match the dataset must match the method used and the flora of the region (Parducci et al., 2019)).

2.2.5 Conservation archaeogenomics

The Anthropocene presents major global challenges, including climate change, loss of biodiversity through extinction, and emerging zoonotic infectious diseases. An understanding of previous human interactions with the environment can guide conservation management during this era of massive environmental change and rapid loss of biodiversity. The field of conservation archaeogenomics involves analysing aDNA with the goal of guiding present-day biological conservation (Hofman et al., 2015). Genomic archaeological data can also reveal details about the time and potential reasons for local or global extinction events, and help to understand the resulting consequences on ecosystems and human societies. Studies that use these data may also contribute to better understanding how human activities and behaviours may have contributed to past extinction events. Studying the distribution of species and how they colonise new areas can also help us to anticipate how ecosystems may respond to future climate change (Alsos et al., 2021). A theoretical application of the recent progress in molecular biology and sequencing techniques follows from the concept of "de-extinction" or "species revivalism". The possibility of de-extinction is controversial and still debated on both technical and ethical levels, as it is difficult to justify the ecological need for reviving extinct species rather than supporting current conservation efforts for endangered species (Orlando et al., 2014).

2.2.6 Future perspectives on plant aDNA analysis

Over the last several decades, palaeogenetics has made substantial contributions towards our understanding of ancient plant science, ecology, and archaeology. In contrast, palaeogenomic is just in its infancy and sequencing and analysis techniques are constantly improving. The study of full genome datasets has allowed to accurately characterise taxonomic diversity (Wagner et al., 2018), to study changes in distribution and demography over time, including changes in population size and measurement of genetic diversity on a population scale (Schwörer et al., 2022; Zimmermann et al., 2017), to investigate the origin of ancient domesticated plant cultivars with high resolution (Ramos-Madrigal et al., 2019; Scott et al., 2019; Trucchi et al., 2021), and to reconstruct entire palaeoenvironments (Capo et al., 2021). The race to understand biological diversity before it is lost is, to some degree, mitigated by the presence of valuable genomic information in archaeological and natural history collections that include extinct and endangered species. As this field of research provides information about common species and their ecological background, it provides a framework in which to study and understand how the past 200 years of human activity have impacted patterns of genetic diversity in the natural world. It is essential that we use insights from the study of ancient plant genomics to help us reduce biodiversity loss over the next 200 years.

2.3 Questions

1. Human faecal material recovered from the latrines of an ancient settlement were analysed with a shotgun sequencing approach, yielding puzzling results. The plants identified from this archaeological site were not domesticated at the time of its occupation, and are not supposed to be present at this location. How can you explain this discrepancy? What protocols can be used to verify this result?

2. A study of the Holocene glacial retreat will be designed to assess the time and zone affected by deglaciation using plant aDNA as a proxy. What aDNA specimens can be used to assess the changes in plant diversity over time at each sampling point, and identify the species involved?

2.4 Glossary

Amber Fossilised tree resin, may contain animal or plant material as inclusion.

Palaeogenetics The study of the past using genetic material from ancient specimens.

Palaeogenomics Genome-scale sequencing studies of genetic material from ancient specimens.

Balancing selection Different selective processes which maintain genetic diversity at a frequency superior to that expected under neutral genetic drift.

Coprolite (or coprolith) Fossilised human or animal faeces. Contrary to palaeofaeces, most of their original composition has been replaced by mineral deposit.

cpDNA Chloroplast DNA, or plastome.

De-extinction Theoretical possibility to rebuild extinct species using aDNA sequences.

Ice core Long cylinder of ice recovered by drilling through ice sheets or glaciers.

mtDNA Mitochondrial DNA.

Palaeoecology (or paleoecology) The study of interactions between organisms and their environment across geologic timescales.

Palaeofaeces (or paleofeces) Ancient animal or human faeces. Contrary to coprolites, they retain some parts of their original biological composition, although in practice the terms are used interchangeably.

Permafrost Ground continuously frozen (below 0°C) for two or more years.

Phytoliths Silica microstructures found in some plant tissues.

Plant domestication Human selection of desirable traits in plants that has taken place in the last 12,000 years.

Positive selection (or directional selection) Process by which one phenotype is selected preferentially to others, causing allele frequency to shift over time towards this phenotype.

Purifying selection (or negative selection) The removal of deleterious alleles from a population genome.

SedDNA Sedimentary DNA, younger and better preserved sedimentary DNA.

SedaDNA Sedimentary ancient DNA, older, more poorly preserved.

Subfossil Organism partially fossilised still containing biological matter such as bone, skin, or faecal deposit, while a fossil is completely mineralized.

Taphonomy Study of how organic remains pass from the biosphere to the lithosphere, including processes affecting remains from the time of death of an organism through decomposition, burial, and preservation as mineralized fossils or other stable biomaterials.

2.5 Answers

1. Several biases specific to aDNA analysis lead to an incorrect identification of the specimen species. The low quantity of aDNA in historical specimens can increase the effect of cross-contamination between samples and differential amplification of the DNA fragments during the PCR process of making genomic libraries. Different replicates of the samples can be analysed in separate facilities to test reproducibility of the results, and a negative control devoid of DNA can be used to check contamination. Ancient DNA damages such as substitution or deletion can affect the DNA sequence itself and lead to incorrect identification. Software assessing aDNA damage and recalibrating the alignment file can be used to minimise this bias. Another source of error can be the incompleteness of the plastid reference database used to match the sequencing reads. If many species are missing from the reference database, the detection might occur at genus level instead of species level. For more information, this question is based on a study that characterised the diet and intestinal parasites of ancient communities in Northern Europe and Middle East from latrines remains aDNA (Søe et al., 2018).

2. To study the evolution of plant richness over time, we can use a time series of samples to reconstruct the evolution of vegetation diversity at the sampling point. A range of datasets from several sampling points can be used to model the Holocene glacial retreat over time. Lake sedaDNA can be an adequate source of aDNA to study climatic change via taxonomic plant diversity detection. SedaDNA is extracted from lake sediment cores, each sediment layer of the core corresponds to a different era. This kind of sampling might provide a measure of plant vegetation richness before and after deglaciation, and might be used to confirm models of the Holocene glacial retreat. For more information about lake sedaDNA cores used to reconstruct changes in plant diversity over time and geographically, have a look at a study using sedaDNA to characterise the emergence of vascular plants after glaciation in Greenland (Epp et al., 2015) or another study reconstructing the post-glacial plant colonisation of Iceland (Alsos et al., 2021; Epp et al., 2015).

Chapter 3

Development of an herbarium specimen Hyb-Seq analysis pipeline for *Begonia* specimens

3.1 Introduction

3.1.1 Endemism and rare species extinction

In the history of conservation genetics, the relationship between species extinction risk and genetic factors has been controversial, several studies claiming that species were driven to extinction by external threats before genetic features as inbreeding could play a role in their demise while others pointed out the relationship between inbreeding and extinction rates (Frankham, 1998, Elgar et al., 2001, Frankham, 2001). Insular population suffering from significantly higher inbreeding rates compared to mainland populations were more likely to be associated with elevated extinction rates (Frankham, 1998). Other studies questioned this point of view, suggesting that the findings were biased by phylogenetic differences between taxonomic groups (Elgar et al., 2001). The ecological context of the populations and species considered was though to be more relevant to management strategies than genetic factors. Reasons underlying extinction events include external factors as habitat loss, overexploitation, chemical pollution and major ecological or climatic change. These threats act in historical time, sometimes in years or months and are therefore thought to affect populations before any genetic factors can play a role, this hypothesis is called the 'No Genetic Impact' hypothesis (Spielman et al., 2004). There have been questions about comparing threatened and non-threatened species with different levels of homozygosity in a meta-analysis on IUCN-listed species

(Greeff et al., 2003, Spielman et al., 2004). However, regardless of the speed with which external factors can impact populations, inbreeding and loss of genetic diversity can lower population fitness and population size, affecting capacity to adapt to ecological and environmental changes. For example, *Primula scotica* is endemic to a small area in Caithness, Surtherland and Orkney. It is vulnerable to changes in grazing management in a way that would not be the case if it had a larger population. It has a mixed breeding system but genetic analysis shows very little variation between individuals, suggesting little capacity to respond to environmental change and poor adaptative potential for re-introduction (Glover et al., 1995).

3.1.2 The *Begonia* genus

Begonia is a pan-tropical genus of herbs, shrubs, and lianas belonging to the family *Cucurbitaceae* within the order *Cucurbitales* along with six other families (*Datiscaceae*, *Tetramelaceae*, *Anisophylleaceae*, *Coriariaceae*, and *Corynocarpaceae*) (Schwarzbach et al., 2000, APG, 2003, Schaefer et al., 2011). *Cucurbitales* families are morphologically distinct, and appearance can change drastically from one family to the other despite common traits, as these groups include few woody species and several genera are monotypic or include a few number of species.

The genus is well-known for its diversity of leaf shapes, patterns, and textures (Li et al., 2022). The group has been intensively studied for its high morphological diversity and high speciation rate (BPG et al., 2022, Moonlight et al., 2018).

Begonia is indeed one of the ten largest genera of plants with more than 2000 species known, and the species number recorded has grown faster than for any genera in the last decade (Frodin, 2004, Moonlight et al., 2018), and more than 200 have been published in the last five years (BPG et al., 2022). As the genus include a large array of different species closely related, it is used as a model for evolutionary history studies focused on the speciation process and its causes (Frodin, 2004, Dewitte et al., 2011, Brennan et al., 2012).

Furthermore, there is only one other sister genus in the family *Begoniaceae*, *Hillebrandia*, with a single species endemic from the Hawaiian Islands, phylogenetically and morphologically distinct from *Begonia* (Clement et al., 2004). Studies have suggested that *Hillebrandia* is a surviving relict endemic genus on the Hawaiian Islands, possibly due to the relatively stable maritime climate of the archipelago. It was suggested as well that separation from its sister group *Begonia* is a relatively ancient event in the history of the *Begoniaceae* and pre-date the formation of the current Hawaiian Island (Clement et al., 2004).

Several explanations have been proposed to explain diversity in the genus, for example high levels of outcrossing due to protandrous pollination (Agren et al., 1991, Ågren et al., 1993, Neale et al., 2006).

Most of *Begonia* populations are small and endemic, while genetic studies show strong population structure and strict population delimitation (Hughes et al., 2003, BPG et al., 2022, Brennan et al., 2012). They are considered as narrow endemic and occupy micro-habitats, with limited gene flow between populations (Li et al., 2022). They are mostly distributed in tropical countries with a higher number of species in South America and Asia compared to Africa, where the genus is probably originated (Neale et al., 2006).

A low gene flow between population is a possible reason for speciation over small geographical scales (Hughes et al., 2008, Twyford et al., 2013). This genetic isolation might be a factor promoting lower genetic diversity, thus poor environmental adaptation, reduced fitness, and vulnerability at a genus scale.

For example, in South Africa, *Begonia sutherlandii* populations have F_{ST}=0.485 (Hughes et al., 2002a), *Begonia dregei*, F_{ST}=0.882 and *Begonia homonyma*, F_{ST}=0.937 (Matolweni et al., 2000). These are levels expected between species. In more widespread species in Mexico lower but still substantial F_{ST} were recorded (*B. heracleifolia*, F_{ST}=0.364; *B. nelumbiifolia*, F_{ST}=0.277) (Twyford et al., 2014a, Twyford et al., 2014b).

The possible effects of such genetic isolation are seen in *B. samhaensis*, an Arabian *Begonia* restricted to the highest point of the island of Samha. No genetic variation has been seen in this population, making it extremely vulnerable to any ecological change (Hughes et al., 2002b). Several species are affected with narrow endemism and not available in the wild any more. The species *B .monicae* has been described from a single Holotype collected in North-East of Madagascar by Aymonin and Bosser in 1920 (Bosser, 1983), and has never been identified in the wild since. This specimen is the only record we have of the entire species. Similarly, *B. antaisaka* records are six specimens collected in 1947 by Humbert in the South-West of Madagascar, and no other specimen has been discovered since (Humbert, 1972). Same situation with *B. bekopakensis* with three specimens collected in 1962 ("Begonia Bekopakensis" 1983).

Anthropogenic activities as urban expansion or farming restrict the area dedicated to native tropical plants, and especially tropical forest. Much has already been lost (Zheng et al., 2021) and much is under great threat (Corlett, 2016). It has been demonstrated that old-growth forests are necessary to preserve endemic herbaceous tropical plants as *Begonia* (Raveloaritiana et al., 2021).

3.1.3 Inbreeding depression and mutational meltdown

Inbreeding can be defined as the mating of close relatives, and the result of this is an increased homozygosity level in a population, which leads to an increased frequency of genetic diseases. The resulting effects are called inbreeding depression (Silvertown, 2001).

A particular case of inbreeding is caused by genetic erosion, a process that affects many endangered species by loss of habitat. Genetic erosion is the result of small population having a limited gene pool, loosing particular alleles as individual with unique genes dies without breeding, and triggering a diminution of available alleles within population (Silvertown, 2001, Gillespie, 2004). Fewer individuals also maximise the impact of any harmful mutations fixed permanently in the population, and the accumulation of of deleterious mutation is eventually ending in mutational meltdown (Lavrentovich et al., 2016).

Mutational meltdown, or error catastrophe, is an accumulation of deleterious mutations in a small population, reducing its fitness and size. Mutations occur in a wild-type population and accumulate with low effect of selection. The deleterious alleles frequency increase with the number of mutant-type individuals, until they are eventually fixed in the population, reaching a frequency of 100%. Under low selective force, the genetic drift has more effect than the purging selection, and drive the population into a downward spiral to extinction by continuously accumulating deleterious mutation (Lynch et al., 1993). First characterized in small, asexual populations, mutational meltdown has been identified as a threat that can affect small, random-mating monoecious populations, as recombination does not affect this process (Lynch et al., 1993; Lynch et al., 1995). The size of the population is a key factor of this particular case of genetic drift. The mean time of extinction of a population is directly linked to its size, with other factors as background of recurrent deleterious mutations and environment or demography fluctuations (Gralka et al., 2016).

Endemic *Begonia* species are likely candidate to mutational meltdown in reason of their repartition and relative isolation.

3.1.4 Using aDNA from historical specimens to test genetic health

Herbaria specimens are a unique resource, providing an opportunity to explore the genome of species not in cultivation, difficult to collect or extinct, as well as the ability to monitor the evolution of allelic frequencies over time (Chapter 1). DNA sampling from herbaria is destructive, and this can limit its use on precious samples such as types. Specific extraction methods have been developed to optimize the DNA recovery from this precious material.Tests have established which specimens are likeliest to give usable data and therefore are worth the destructive sampling (Kates et al., 2021, Forrest et al., 2019).Non-destructive methods have also been developed, such as using a Staedtler “Mars Plastic” eraser to collect sufficient DNA molecules from the surfaces of herbarium samples to allow PCR amplification (Shepherd, 2017). Herbaria specimens have been increasingly used over the past ten years for DNA-based research, but despite the optimisation of extraction methods low, fragmented and contaminated yields are the general rule. Target capture overcomes issues of low yield and of contamination and is now well developed as a tool for generating genetic data from herbarium samples (Gutaker et al., 2019, Bieker et al., 2018, Gutaker et al., 2019, Bieker et al., 2020, Kates et al., 2021, Bieker et al., 2022). This has driven interest in developing specific pipelines to manage herbarium-captured DNA (Fig.3.1).

3.1.5 Target capture

Target capture is one method of reduced-representation sequencing, using molecular probes to target specific loci and isolate them from the rest of the genomic sequences (Fig.3.1). The ability to collect just the matching sequences from a heterogeneous mixture of highly fragmented DNA makes targeted capture ideal for dealing with the problems of herbarium samples (Cronn et al., 2012). Several sets of baits are available for plants including very wide-ranging ones, designed from conserved single copy elements to amplify across angiosperms (Johnson et al., 2019), and many sets designed for specific clades (Woudstra et al., 2021, Cowman et al., 2020, Eserman et al., 2021, Michel et al., 2022)).

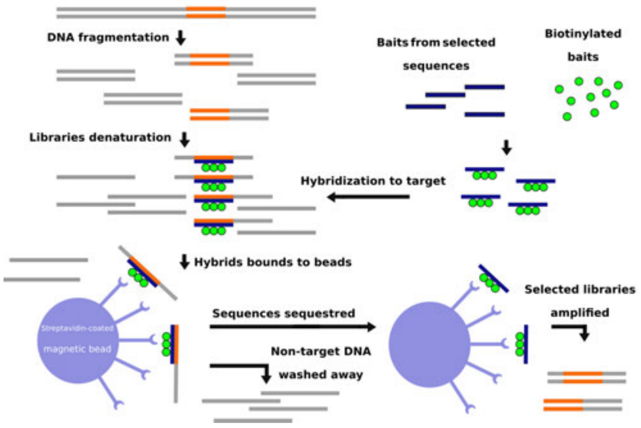


Figure 3.1: Mechanism of Target Capture.

Pipelines developed for NGS datasets

Sequence data from hybrid capture is usually generated with Illumina, often as short paired-end sequences. Bioinformatically hybrid capture datasets have the advantage of a known reference (the bait set) and, hopefully, a fairly simple structure with high coverage on distinct non-repetitive sequences. Pipelines use a combination of de-novo assembly, mapping reads or contigs to target sequences and filtering to remove chimeric sequencing or paralogs to generate a set of sequences per target per sample for further analysis (Fig.3.1). The focus may be on deriving single and easy to analyse consensus sequence for each target locus, or on harvesting as much data as possible, including non-bait sequence from introns or up and downstream regions, and multiple paralogs for a very rich as dataset. In this study five assembly pipelines have been considered for hybrid capture data analysis. In parallel to my researches, a trial has been set up to compare HybPiper, HybPhyloMaker, SECAPR, PALEOMIX, and a home-made tool called the BASIC pipeline. The parameters of the comparison and specificities of the pipelines are described in Michel et al., 2022. Eventually, the PALEOMIX pipeline has been selected for most steps of this study due to the volume of data produced, the size of the contigs aligned, the integration of an aDNA analysis tool, and its modularity allowing easy modifications of the pipeline (Michel et al., 2022). Furthermore, a reference-based pipeline was necessary in downstream analysis, notably the Runs Of Homozygosity (ROH) measurements that requested long scaffold genomes references (Ceballos et al.,

Pipeline	Data	Purpose	aDNA	Ploidy management	Modules	References
Phylucce	Hyb-Seq	Alignment phylogenetic analysis	No	Yes	Paralog detection	Faircloth, 2015
HybPiper	Hyb-Seq	Exon and flanking intron recovery	No	No	Paralog detection	Johnson et al., 2016
HybPhyloMaker	Hyb-Seq	Data analysis and species tree reconstruction	No	No		Fér et al., 2018
Secapr	Hyb-Seq	Non-model organism	No	No		Andermann et al., 2018
Paleomix	Hyb-Seq	aDNA analysis	No	Yes		Schubert et al., 2014

Table 3.1: NGS pipelines for target enrichment dataset analysis.

2018), available for the *Begonia* group since 2021 (Li et al., 2022). HybPiper has been used to detect paralogous sequences for a comparative study described further in this Chapter.

Phyluce

Phyluce was designed to focuses on UCE (Ultra Conserved Elements) markers for phylogenetic analysis and comparison specifically in non-model species (Faircloth et al., 2012). The software uses de-novo assembly to generate contigs from the reads then aligns contigs to target loci using LASTZ. Detection of the paralogous sequences is done by comparing the contigs matching probes addressed to different loci or single locus. It aligns the remaining orthologous sequences for phylogenetic inference (Faircloth, 2015).

HybPiper

HybPiper was designed to work with the Angiosperm 353 bait set, and to extend the sequence recovery beyond the coding sequence into the more variable intron regions, valuable for phylogenetic analysis (Johnson et al., 2016). Sequencing reads are aligned to the target sequence using Bowtie2, then the set of reads mapping to each target are individually de-novo assembled into contigs and coding sequence and introns identified. Either exons, intron or all the sequence (supercontigs) can be extracted for further use in the next steps. Targets with many paralogs are identified and paralogs can be analysed using additional modules (Jackson et al., 2021).

HybPhylomaker

HybPhylonaker has been developed specifically for target enrichment analysis from raw reads to supertree- and multispecies coalescent-based species tree reconstruction (Fér et al., 2018). Similarly to HybPiper, HybPhyloMaker build a pseudoreference sequence with a de-novo assembly of the raw reads which allows to recover the intronic part at the fringe of the exon. This method seems to be efficient to isolate orthologous sequences, as they are more abundant due to paralogs to have little sequence dissimilarity with the baits (Fér et al., 2018). HybPhyloMaker includes as well a whole module of gene tree estimation using either FastTree or RaxML, and the last part of the pipeline is a species tree reconstruction step using ASTRAL (Zhang et al., 2018) or ASTRID (Vachaspati et al., 2015) (coalescent method), MRL (Nguyen et al., 2012) (supertree method), and FastTree (Price et al., 2009) and ExaML (Kozlov et al., 2015) (concatenation method).

SECAPR

The Sequence Capture Processor pipeline (SECAPR) allows either de-novo assembly, to generate contigs from captured data without a reference, or reference-guided assembly, producing phased or unphased consensus sequences (Andermann et al., 2018). Contigs produced from a de-novo assembly can be used as reference in subsequent steps. De-novo assembly uses Abyss and the resultant contigs are mapped to targets using LASTZ (Harris, n.d.), producing alignments for each target for all loci which will include off-target sequence such as introns and flanking sequence. The reference based assembly uses BWA to map the reads and remove duplicates. Consensus sequences are derived from the BAM files using SAMtools (Danecek et al., 2021), with possible phasing to derive both alleles.

PALEOMIX

PALEOMIX is a modular pipeline processing raw sequencing reads through several analytical stages before a step of alignment against one or more reference genomes (Schubert et al., 2014). The pipeline integrate the tool mapDamage2.0 (Jónsson et al., 2013) that evaluate the presence of postmortem DNA damage signatures in the reads alignments to authenticate aDNA data set and recalculate base quality scores in aligned reads in order to reduce the noise in downstream analysis.

3.1.6 Ancient DNA damages pattern in herbarium specimens

The aDNA contained in herbarium specimens is usually highly fragmented and can include degenerated nucleotides that could impair variant calling.

Two types of DNA degradation are common in herbaria material. DNA fragmentation is the most common, observed at various degree in silica-dried and herbarium material (Forrest et al., 2019). Due to processes of depurination of the DNA sugar phosphate backbone and β-elimination, it occurs in slowly-dried specimens, specimens exposed to heat, or exposed to moisture during storage (Dabney et al., 2013). On the contrary, substitution C to T and G to A, caused by nucleotides deamination and located at extremities of DNA fragments occur during amplification of aDNA, and is a more reliable marker of historical aDNA. Modern contaminants can be integrated in a batch of samples and be over-amplified during library preparation and mistakenly analysed with the whole dataset of specimens (Kistler et al., 2020). These issues can be addressed during preparation of the samples, using clean laboratories facilities dedicated to aDNA extractions and DNA repair kits (Mouttham et al., 2015). However, bioinformatic methods have also provided means to identify and quantify aDNA patterns, allowing us to authenticate the part of the dataset representing historical specimens sequences and lower their impact on variant calling (Ginolhac et al., 2011, Peyrégne et al., 2020, Neukamm et al., 2021). In this study we used two pipelines for identifying and managing DNA damage.

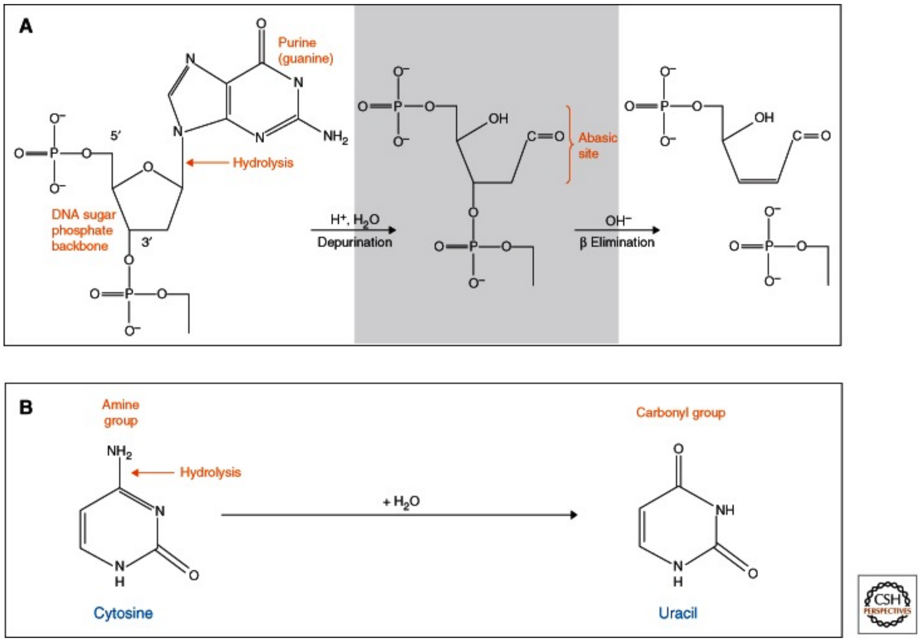


Figure 3.2: Fragmentation and deamination process (Dabney et al., 2013).

Antcaller is a python based programme which is based on the GATK approach and uses Bayesian analysis to calculate the probability of each genotype, including estimates of probability of damage for C to T and G to A (Zhou et al., 2017).

MapDamage 2.0 is a software using Python and R scripts, it calculates misincorporatated nucleotides and fragmentation with respect to a reference (Jónsson et al., 2013). It allows identification of ancient DNA in a mixed sample. It has been incorporated into the pipeline Paleomix to allow more accurate SNP calling on damaged DNA (Schubert et al., 2014).

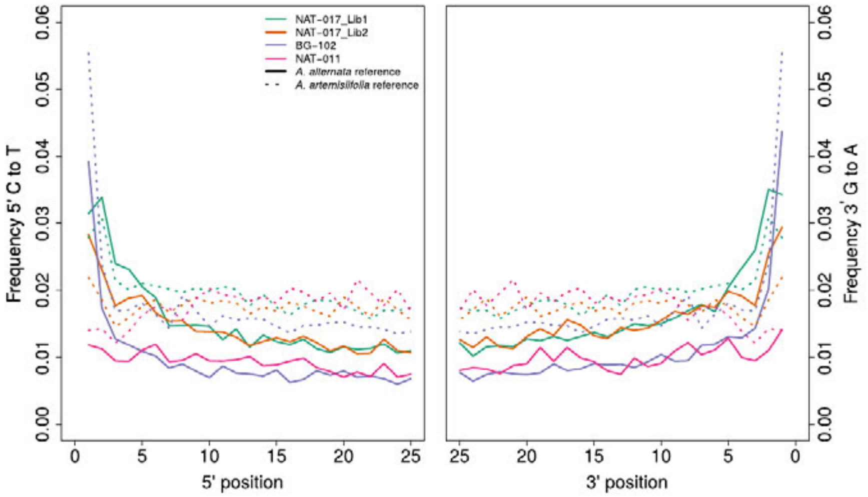


Figure 3.3: Misincorporation rates for C to T and G to A in ancient DNA (aDNA) (Bieker et al., 2018).

3.1.7 A pipeline to detect homozygosity patterns and inbreeding depression

In this chapter we aim to compare the success of the different tools and pipelines described for accurate detection of damaged DNA, paralogs, SNP calling and runs of homozygosity using hybrid capture from two sample sets.

3.2 Methods

3.2.1 Begonia target capture baits

An hybrid target capture set of baits has been set up by Dr. Catherine Kidner for phylogenetic analysis, population genetic analysis, and functional studies. The purpose is to recover nuclear genes for phylogenomics analysis, targeting Ultraconserved Elements (UCEs) as slow-evolving regions that can be used on a large phylogenetic diversity of organisms. In contrast protein-coding genes can be used to design the probes in order to capture low-copies exonic genes for phylogenetic and molecular evolution analysis (Zhang et al., 2012). Additionally, flanking intronic regions can inform more recent relationships. Then to achieve a high coverage on thousand of markers selected for evolutionary history (McKain et al., 2018). The target capture bait set used for this study includes a total of 1,239 loci from *B. conchifolia* and *B. luzhaiensis* transcriptomes, designed to capture conserved sequences for phylogenetic and population genetic studies and functional regions including developmental genes (Michel et al., 2022). The functional genes included in the set of baits include annotated genes involved in shade adaptation. While Universal probe sets of baits has been developed for cross-species analysis among the angiosperm clade targeting conserved regions (Johnson et al., 2019), claiming no phylogenetic bias caused by difference of relatedness between the sample analysed and the design of the bait set. The set has been designed as well to avoid high polymorphism sites related to incorrect sequence assembly (assembly errors) and gappy regions. A preliminary study has been made to check where are the loci targetted in different *Begonia* genomes, if they are contiguous, and what are the linkage group involved (Brennan et al., 2012, Fig.3.4).

Transcriptome from leaves and flower buds of *Begonia luzhaiensis* T.C.Ku (Tseng et al., 2017). BLASTN transcriptome on its own sequence to identify sequences over 100bp with single match to identify single-copy genes with 98% identity. BLASTN *B. luzhaiensis* to annotated genes from *B. conchifolia* with 90% identity (Campos-Dominguez, 2020). The output were filtered using the genome assembly method used by Yang et al., 2012 to match cucumber (*Cucumis sativus* L. cv.) genes with matches above 90% identity.

The recovery of sequence data from silica-dried and historical herbarium specimens with this set of baits had been demonstrated (Michel et al., 2022).

3.2.2 Target capture set of samples

The specimens considered for setting up and testing our analysis pipeline are *B. conchifolia*, *B. plebeja* and F1 hybrids and backcrosses. They are a mapping population set up for a previous work related to the production of a *Begonia* genetic map (Brennan et al., 2012, Twyford et al., 2014b). The *B. conchifolia* parent has a sequenced genome, *B. plebeja* a sequenced transcriptome, and the known parentage of the F1s and backcrosses allows prediction of expected alleles and level of heterozygosity (Table .3.2). We included in the batch of samples herbarium-dried and and fresh silica-dried samples from the same individuals to study the pattern of damage occurring during the herbarium-drying process. This mapping population has been selected to build and test the pipeline as the *B. conchifolia* parent has a sequenced genome, *B. plebeja* a sequenced transcriptome, and the known parentage of the F1s and backcrosses allows prediction of expected alleles and level of heterozygosity. The expected segregation pattern in parents, F1, and backcrosses will be used to discriminate the orthologous sequences from the paralogous duplicated sequences and to exclude paralogs from the analysis.

Leaf tissue has been collected for each accession and silica-dried. DNA extraction followed the standard Qiagen DNeasy Plant Mini Kit protocol. DNA was quantified using a Qbit 4 Fluorometer with the dsDNA HS chemistry kit, and a quality check performed on an Agilent TapeStation. All samples were normalized to 2 ng/uL before fragmentation step. Fresh and recent historical specimens were fragmented to 350bp using a Covaris M220 Focused-ultrasonicator. Library preparation followed the protocol of the NebNext Ultra II DNA Library Prep for Illumina Kit. Seramag Sample Purification beads were used for size selection of samples above 50 ng, and clean up for less concentrated samples. An Agilent Tapestation with High Sensitivity kit was used for libraries quality check.

Table 3.2: Mapping populations specimens collected to build and test the pipeline.

Specimen ID	Species	Condition	Living collection ID
Plebeja_1_silica	*B. plebeja*	silica dried	20051406
Plebeja_2_silica	*B. plebeja*	silica dried	20051406
Conchifolia_1_silica	*B. conchifolia*	silica dried	20042082
Conchifolia_2_silica	*B. conchifolia*	silica dried	20042082
F1_CKB137_1_silica	hybrid	silica dried	CKB137.1
F1_CKB137_2_silica	hybrid	silica dried	CKB137.2
F1_CKB137_6_silica	hybrid	silica dried	CKB137.6
F1_CKB137_9_silica	hybrid	silica dried	CKB137.9
BC_ARB312_5_silica	hybrid	silica dried	BC ARB312.5
BC_ARB312_71_silica	hybrid	silica dried	BC ARB312.71
BC_ARB312_76_silica	hybrid	silica dried	BC ARB312.76
BC_ARB312_117_silica	hybrid	silica dried	BC ARB312.117
Plebeja_1_herbarium	*B. plebeja*	herbarium dried	20051406
Plebeja_2_herbarium	*B. plebeja*	herbarium dried	20051406
Conchifolia_1_herbarium	*B. conchifolia*	herbarium dried	20042082
Conchifolia_2_herbarium	*B. conchifolia*	herbarium dried	20042082
F1_CKB137_1_herbarium	hybrid	herbarium dried	CKB137.1
F1_CKB137_2_herbarium	hybrid	herbarium dried	CKB137.2

Subsequently, libraries have been normalised to 10 nM, then pooled according to fragment size and quality. Three pools of 10, 14, and 19 libraries were made. The hybridization step followed the MyBaits Hybridization Capture for Targeted NGS Manual version 4.01. According to the guidelines of the manual relating to degraded or contaminated DNA libraries, the hybridization time was extended to 24 hours with a temperature of 62°C. 16 post-amplification cycles were performed on all the samples. Pools were sequenced by Edinburgh Genomics on a single lane of NovaSeq6000 SP with 250bp paired end Illumina reads.

Unfortunately the loci recovery of the target capture protocol has been uneven across loci, with random variable read depth, improper to develop our pipeline. We then shifted our test group for a set of specimens processed via genome skimming by the Dr. Cynthia Fan (Fan, 2023).

Genome skimming set of samples

As the target capture protocol did not achieve the results expected, another batch of samples has been added subsequently to the study in order to build

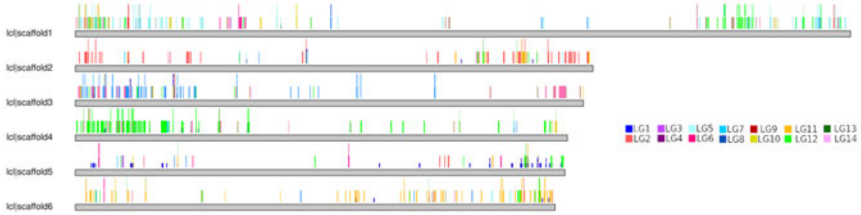


Figure 3.4: Baits location and linkage group associated on six pseudo-chromosomes of the *B. masoniana* genome.

the pipeline. This set of data has been produced by Dr. Cynthia Fan in the context of her PhD thesis, and consist of genome skimming of the same mapping population previously described for the target capture set (Fan, 2023). A total of 82 samples from the mapping population have been included in this dataset: one Parent plant *B. plebeja*, one F1 plant *Begonia* hybrid identified as CKB137.8, and 80 F2 plants *Begonia* hybrid backcrosses of *B.conchifolia* identified as B.08 Sprouting leaves were flash-frozen in liquid nitrogen. We have proceeded to DNA extraction using the protocol described in Nishii et al., 2022, excluding the nuclei isolation step and with several editions to the method. QIAGEN Genomic-tip 20G kits were used for High-molecular weight DNA extraction, loaded with 200mg of ground material to prevent overloading the column. After leaves were ground, the tissues cells were lysed with 4ml of QIAGEN G2 buffer. We resuspended the DNA extracted in 1ml of 0.1 x Tris-EDTA buffer. DNA quality check has been done with a Denovix Qubit. The samples have been sequenced with 150bp paired end Illumina reads, providing an average of 10x coverage.

3.2.3 Alignment pipeline

After sequencing, the raw reads were processed through the PALEOMIX pipeline. The makefile of the pipeline was set up with QualityOffset of 33 accordingly to the choice of Illumina platform. The AdapterRemoval option for trimming was set up according to the following parameters: –adapter1: AGATCGGAAGAG-CACACGTCTGAACTCCAGTCA, –adapter2: AGATCGGAAGAGCGTCGT-GTAGGGAAAGAGTGT, –mm: 3, –minlength: 25. We used BWA as aligner with the following parameters: Algorithm: mem, MinQuality: 20, FilterUnmappedReads: no, UseSeed: yes. PCR duplicates have been filtered out with Picard MarkDuplicates (*Picard Toolkit* 2019). We used mapDamage with default settings, a downsample of the input to 100,000 hits (Jónsson et al., 2013). After adapter removal, we excluded from the alignment single-end reads, non-collapsed paired-end reads, paired-end reads for which the mate was discarded, and overlapping paired-ended reads collapsed into a single sequence. The BAM files were indel-realigned using GATK Indel realigner (O'Connor BD, 2020).

3.2.4 SNP calling

We have used Genome Analysis ToolKit (GATK) with gatk HaplotypeCaller to join call variants (McKenna et al., 2010). Jointly calling all samples in a cohort has been described as a gold standard strategy for analysing rare and under-represented SNPs, especially for low-coverage data (Chen et al., 2020, Nho et al., 2014). To filter them, and have a proper SNPs reliability, the method recommanded is VQSR. Since we do not have the option of an adaptative filtering, as it require a well-curated database of markers, hard filtering has been chosen as an alternative. The filters were calibrated according to the distribution of the different variants annotations of the VCF files INFO field, following the recommendation given by the GATK guidelines (O'Leary et al., 2018, Caetano-Anolles, 2022) (Fig.3.5). This method has been proved to reduce the number of unfiltered false positives (De Summa et al., 2017), we choose this conservative approach at the risk to filter out true variants from the analysis.

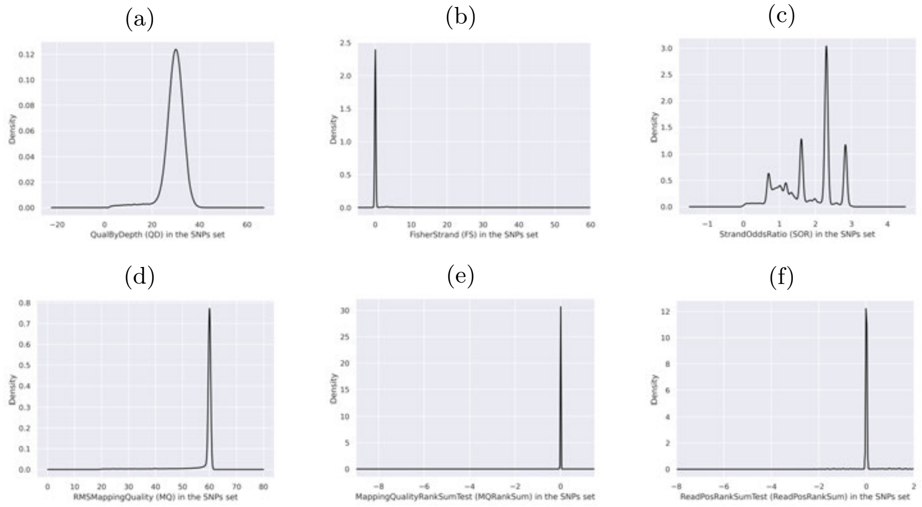


Figure 3.5: Density of the annotations values of the INFO field of the VCF file. (a) QualByDepth (DP), (b) FisherStrand (FS), (c) StrandOddsRatio (SOR), (d) RMSMappingQuality (MQ), (e) MappingQualityRankSumTest (MQRankSum), (f) ReadPosRankSumTest (ReadPosRankSum).

3.2.5 Paralogous genes detection

The presence of duplicated genes in *Begonia* genomes can be detected in the evolutionary history of the genus. A recent Whole Genome Duplication event (WGD) occurred early in Begoniaceae evolution, at the origin of the family some time before before 22 MYA (Brennan et al., 2012). Several genome duplications have occurred since, in different lineages (Campos-Dominguez et al., 2022) and segmental and tandem duplication are common (Li et al., 2022). This

Table 3.3: Reference genomes used to map the reads.

Reference genome	Section	Size (Mbp)	Assembly	Source
B. loranthoides	*Tetraphila*	671	chromosome-scale	Li et al., 2022
B. masoniana	*Coelocentrum*	799	chromosome-scale	Li et al., 2022
B. darthvaderiana	*Petermannia*	785	chromosome-scale	Li et al., 2022
B. peltatifolia	Unassigned	309	chromosome-scale	Li et al., 2022
B. bipinnatifida	*Petermannia*	1,099	draft genome	Lucia Campos-Dominguez
B. conchifolia	*Gireoudia*	564	draft genome	Lucia Campos-Dominguez
B. dregei	*Augustia*	546	draft genome	Lucia Campos-Dominguez
B. fuchsioides	*Lepsia*	322	draft genome	Lucia Campos-Dominguez
B. johnstonii	*Rostrobegonia*	291	draft genome	Lucia Campos-Dominguez
B. luxurians	*Pritzelia*	254	draft genome	Lucia Campos-Dominguez
B. socotrana	*Peltaugustia*	346	draft genome	Lucia Campos-Dominguez
Begonia baits set	-	1.9	bait set	Michel et al., 2022

kind of event is common in angiosperm lineages and are possibly a vector of diversification. It is very difficult to distinguish close paralogs from alleles, so WGD and other duplications can lead to errors in demographic analysis. The *Begonia* bait set had been designed to exclude multi-copy genes, but was based on a single genome and single transcriptome. Copy number variability certainly would exist across the genus.

We used the recently published set of 4 *Begonia* genomes including *B. loranthoides*, *B. masoniana*, *B. darthvaderiana*, and *B. peltatifolia* (Li et al., 2022). We used as well 7 draft genomes (Lucia Campos-Dominguez, being published) to examine variation in copy number of the baits across the genus. The 7 draft genomes include *B. bipinnatifida*, *B. conchifolia*, *B. dregei*, *B. fuchsioides*, *B. johnstonii*, *B. luxurians*, *B. socotrana* (Table. 3.3).

3.2.6 Measurements of Runs Of Homozygosity

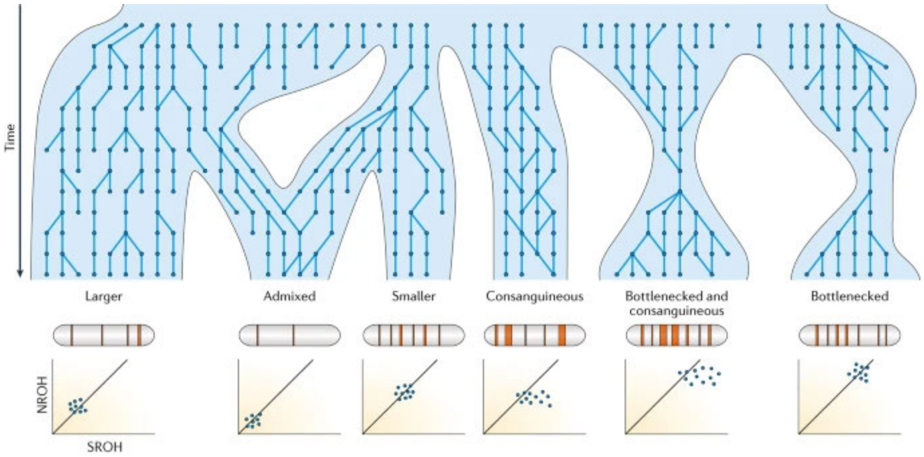
The purpose of our study is to detect inbreeding in the populations sampled. Even though the research community agree on a simple definition of this concept 'inbreeding is the mating of related parents' (Charlesworth et al., 1987, Curik et al., 2014) many disagreements have emerged in relation to the definition of 'related parents' as two individuals from the same population share at least one ancestor (Curik et al., 2014). Another cause of disagreement is the vast array of phenomenons related to the emergence of inbreeding: changes in effective population size, genetic drift, deviation from the Hardy-Weinberg equilibrium, decrease of genetic diversity. An alternative is to focus on the genetic consequences of inbreeding to define it: it changes genotype frequencies (increasing homozygosity, decreasing heterozygosity) but do not affect alleles frequencies (Curik et al., 2014). This situation emerges when maternal and paternal haplotypes are similar, as a consequence of autozygosity. Autozygosity is the breeding of related individuals, in which the two alleles inherited by the progeny are identical by descent (IBD). Two alleles or haplotypes are IBD if they have been inherited from the same ancestral haplotype without recombination. Several methods have been set up to detect autozygosity, some rely for example on comparing the frequencies of haplotypes identical-by-state (IBS) to the frequencies for IBD (Cockerham et al., 1968). Others calculate the expected genotypic frequency in a population and compare it to the alleles frequency at a locus to see if there are frequencies disequilibrium (Curik et al., 2014).

However, the method of measuring Runs of homozygosity (ROH) can disentangle the autozygosity of other methods genotypic effects. ROH are long chromosomal homozygous segments resulting from the mating of related individuals (Ceballos et al., 2018). They were first detected in the late 90s from the efforts to build the first human genetic map, where several long stretches of non-informative homozygous markers were detected in one of the reference family part of the genotyping experiment (Broman et al., 1999). Homozygous segment in the genome can be caused by linkage disequilibrium, as homozygous linked sequences would be inherited altogether. But this phenomenon would only be local and cause short homozygous segments, at the contrary the length of ROHs observed in an inbred population would be longer (Broman et al., 1999, Curik et al., 2014). The high density of markers (8,000 short tandem-repeat polymorphisms) in this early study and a genetic linkage map have enabled to distinguish ROH caused by linkage disequilibrium and inbreeding caused by autozygosity (Broman et al., 1999).

It is also noticeable that ancestral ROH are shorter as homozygous haplotypes are broken up by recombination during meiosis (Broman et al., 1999, Curik et al., 2014).

ROH are created when an individual's parent share haplotypes (Ceballos et al., 2018). Longer haplotypes are expected when the parents are more closely related. Patterns of ROH can therefore be very informative on population history. Large populations will have small numbers of short ROH, small populations will have more and longer. Many ROH of varying sizes are expected after a population bottleneck, populations with extensive inbreeding will have an over representation of long ROH (Ceballos et al., 2018). They are several methods available to detect ROH. There are different methods you can use to make the measurements. Observational genotype-counting approach with PLINK (Purcell et al., 2007) or Hidden Markov models (with GERMLINE, Beagle), this last one taking in account background LD. For short-reads datasets it seems that HM models have better resolution (Meyermans et al., 2020). Nonetheless, PLINK is still the most popular tool used (Meyermans et al., 2020). While traditionally restricted to the exploration of cattle inbreeding, PLINK is increasingly used to study inbreeding in plant model (Kumar et al., 2021).

The Runs of Homozygosity (ROH) were estimated with PLINK v1.90b6.21



Nature Reviews | Genetics

Figure 3.6: Demographic origins of ROH. Demographic history of six diverse hypothetical populations. The dots are individuals connected by dark blue lines which are pedigree. Population size is represented by the blue area, and loops in the lineage represent crossing between individuals of the same population. SROH is the sum of total length of ROH, and NROH is the total number of ROH. The different signatures of demographic scenarios on ROH are shown in the NROH versus SROH plot (Figure from Ceballos et al., 2018).

(Purcell et al., 2007). It uses a sliding-window approach to detect homozygosity, where a preset number of SNPs are checked to detect extent of homozygosity. The ROH segments are defined using a threshold comparing the window's average to the average of overlapping windows for each SNP. To prevent an ROH to be called by incomplete data or by chance, the minimum number of SNPs to consider a ROH (l) is calculated with the following equation (Lencz et al., 2007, Kumar et al., 2021).

$$l = \frac{\log_e \frac{\alpha}{n_s n_i}}{\log_e(1 - het)} \tag{3.1}$$

Where α is the percentage of false-positive ROH (set at 0.05 here). n_s is the average number of SNPs per individual, n_i is the number of individuals, and het is the mean ratio of the number of heterozygous SNPs on the number of all SNPs. To avoid low SNP density that could introduce a bias in the ROHs measured, we set a maximum gap of 1Mb between consecutive SNPs (Meyermans et al., 2020).

3.2.7 Estimating homozygosity level in a low-coverage dataset

Estimating homozygosity level in a low-coverage dataset is problematic due to the amount of missing or incorrect data. It is especially challenging to detect homozygous sites in aDNA because of the generally lower depth and high error rate specific to these datasets. For many sites, the sequencing depth will be too shallow to accurately detect polymorphisms. Low coverage amplifies the problems which may result from incorrect mapping of hybrid capture data (Bryc et al., 2013). Different models and tools have been set up specifically to analyse shallow aDNA datasets. Most of the methods include algorithms that stochastically learn alleles frequencies from the same population for inference (Kim et al., 2011). This described as leveraging the joint information across the panel of samples to estimate the alleles distribution in the genome including sequencing errors. This method allows one to disentangle the sequencing errors from rate of homozygosity using expected-maximization algorithms, it is helpful to know the expected relatedness between samples (Bryc et al., 2013). The most up-to-date tool on for calling variant in a low-coverage dataset is the Genome Analysis Toolkit (GATK) HaplotypeCaller algorithm (Brouard et al., 2019).

3.3 Results

3.3.1 Baits with multiple matches

A BLAST has been made with the baits set on different genomes to visualize the potentiality of our baits to capture multiple sequences as microsatellites or paralogous duplicated sequences. The results show a minority of baits matching up to 100 different sequences across genomes, with different behaviours depending of the species (Fig.3.7). The 36 multi-scoring most problematic baits were removed from the analysis. More information have been collected on the capture of microsatellites sequences with the second batch of Socotran specimens (Chapter 3).

3.3.2 Target capture and genome skimming set of data

The target capture method did not work evenly across all samples after sequencing. A total of 169,091,031 reads were produced for the dataset, with an average of 9,393,946 reads per sample. The average proportion of PCR duplicates was 0.978 duplicates in each sample, reducing the average number of unique hits per sample to 81,651 with an average coverage of 4.59 per sample.

The capture rate of each bait has been uneven in cause of the problematic baits showing microsatellites sequences, but the other baits show irregular patterns of capture as well. We aligned the sequencing reads to the bait set, and the average coverage for each bait is uneven across samples of the mapping population (Fig. 3.8, 3.9a). A closer look to the rate of target recovered (rate of nucleotides with a minimal coverage of 1 on target capture bait sequence) shows similarly an incomplete capture of the sequences targetted for a majority of the baits (Fig. 3.9b).

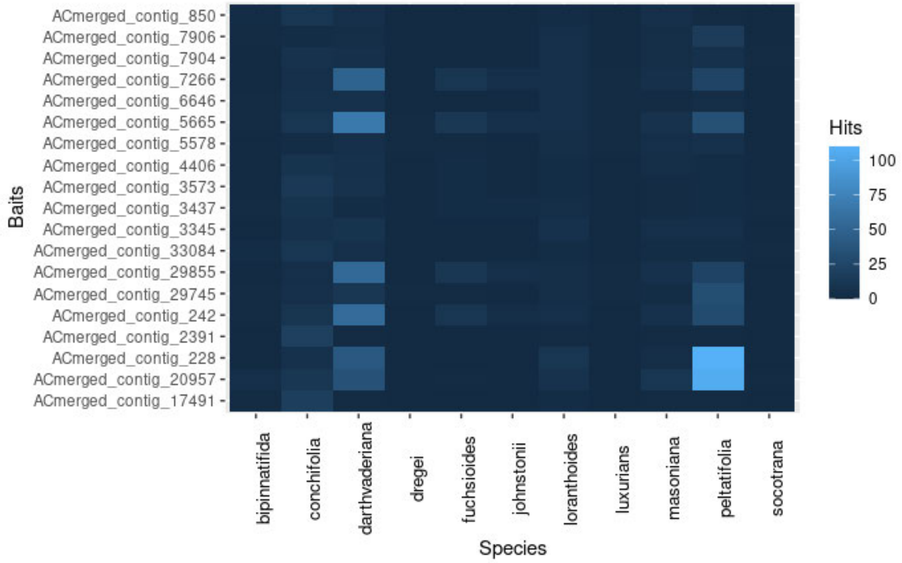


Figure 3.7: Baits with multiple matches across genomes.

3.3.3 Paralogous baits detected

Several methods were tried and compared to detect baits capturing paralogous sequences. The genotype frequency method is based on the use of the mapping population Parents, F1, and hybrids backcrosses (BC). The expected genotype of the Parents would be homozygous for a majority of sites, heterozygous for the F1 hybrids, and 25% homozygous 50% heterozygous for the BC hybrids. As our target capture set has not provided the even coverage expected, we have used the 82 specimens of the genome skimming set of samples from the mapping population to calculate the genotype frequencies. All the SNPs captured by baits which did not match the genotypes frequencies expected were marked as potential paralogs (Fig.3.10b).

The expected heterozygosity method is assuming an average rate of heterozygosity for each variant detected. This ratio has been calculated with HDplot (McKinney et al., 2017), an R package plotting H the heterozygosity for each SNP, and D the deviation from even read-ratios in heterozygous individuals (Fig.3.10a). The outliers on this plot have been marked as potential paralogs.

The HybPiper (Johnson et al., 2016) module paralog_investigator.py has been run to detect potential paralogs and compare them to the other methods results. The final results show only an overlap of 73 baits between the three methods, while HDplot detect 310 potential paralogs, HybPiper 136, and the deviation from expected genotype frequencies 635 (Fig.3.11). The different number of putative paralogous loci found with the three methods as well as the poor overlap of the results suggest that a more robust data set is necessary, with

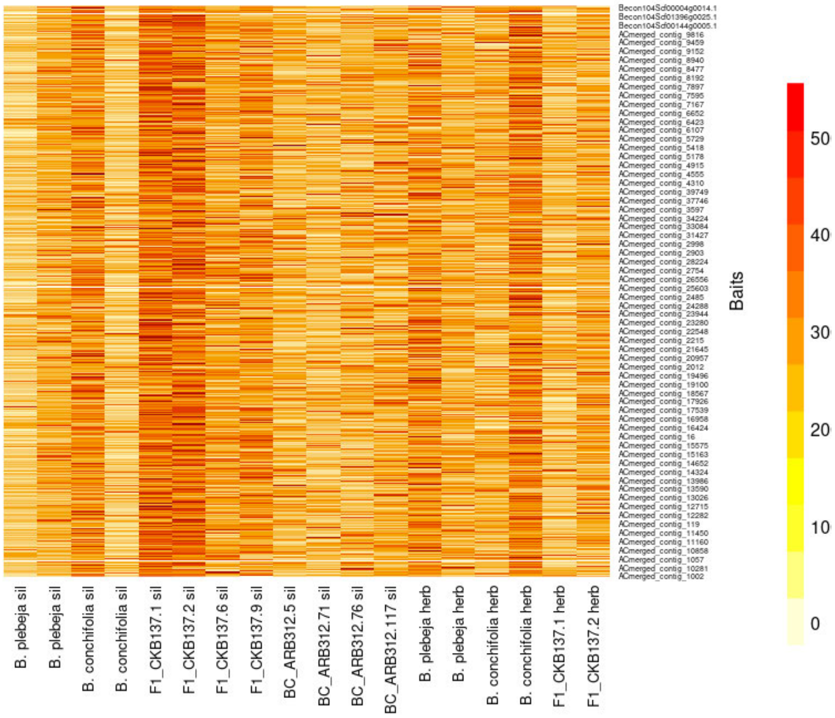
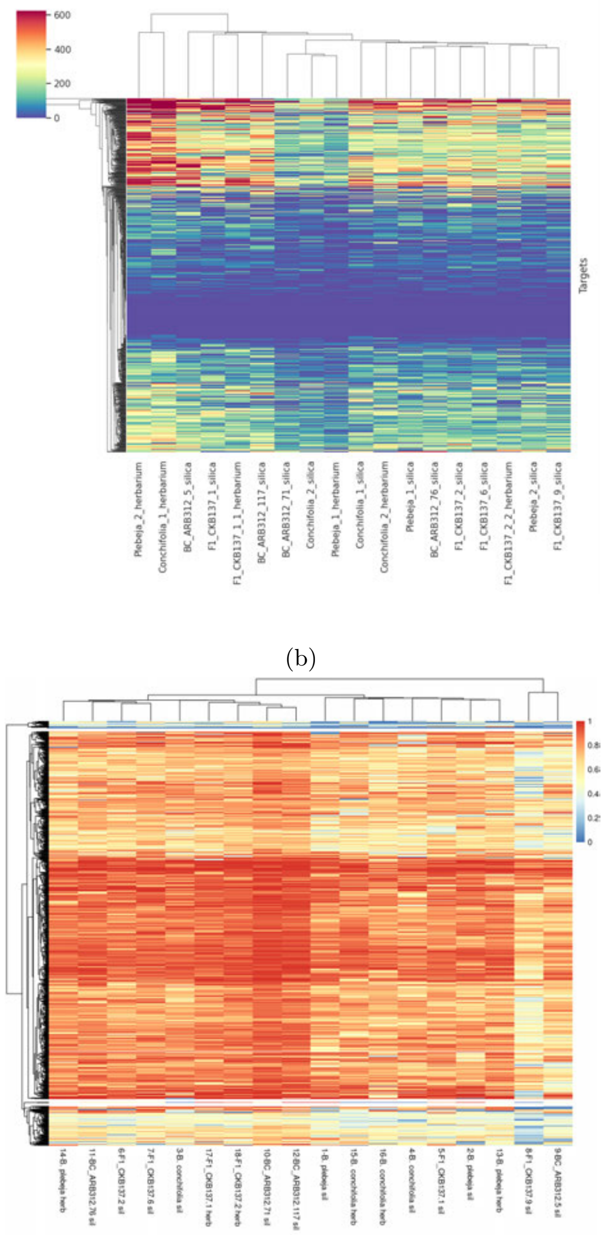


Figure 3.8: Average coverage for each bait per mapping population sample. The heat map shows the average coverage per sample and on each baits (labels shown on the right, not all the baits are presented).

better coverage and more individuals to enable finer tuning of our analysis tools.

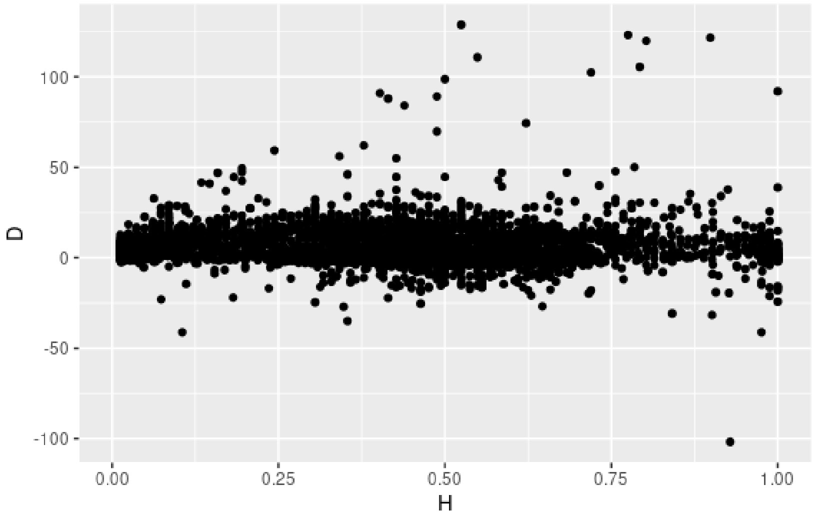


(a)

(b)

Figure 3.9: Target capture reads recovered. a) Average depth of sequencing coverage, b) rate of target recovered by bait and sample.

(a)



(b)

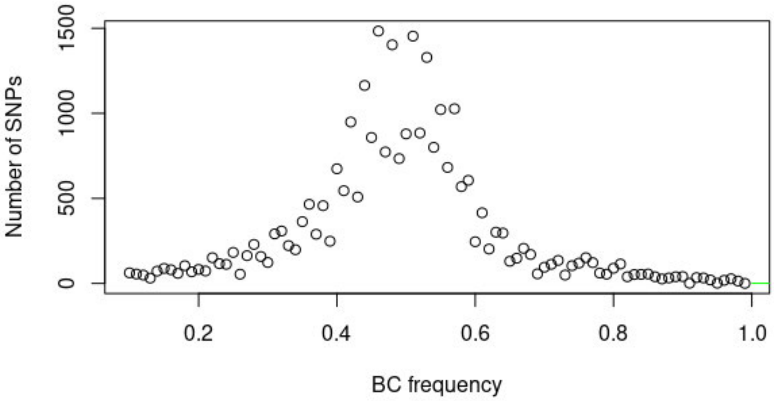


Figure 3.10: Paralogous sequences captured by the bait set. (a) SNPs plotted with HDplot, (b) Genotypes frequencies in the backcross specimens.

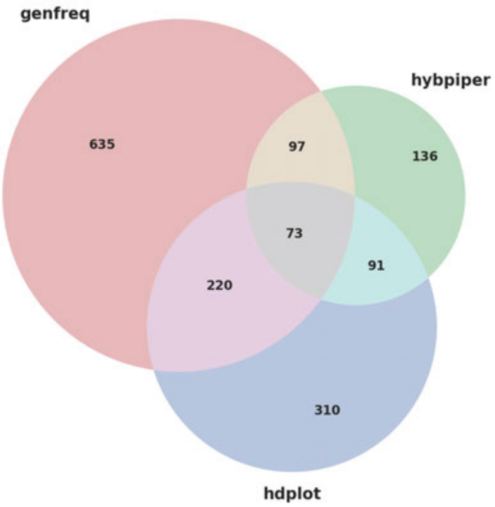


Figure 3.11: Venn diagram of the numbers of paralogous baits detected by three different methods.

3.3.4 Syntenic sequences

The detection of runs of homozygosity required genotyping of extensive lengths of chromosome. With over 1,000 target sequences and generally good synteny across *Begonia* we expected to find several regions of target loci syntenous across *Begonia* (Michel et al., 2022), Li et al., 2022). Four long *Begonia* genomes were assembled recently by Li et al., 2022, with pseudo-chromosomes totalizing between 331.75 Mb (*B. peltatifolia*) and 799.83 Mb (*B. masoniana*). We selected them as reference genome for our study on ROH in order as longer chromosomes would allow to observe more and longer ROH. An assay was set up to observe the size of the sequences covered by the baits, and if they would be syntenic block between different species. We have matched our bait set to the genomes and for each match counted the number of contiguous baits and the length covered by the baits (Fig.3.12). We have observed 54 syntenic blocks shared between the four species and captured by our bait set. The average size of a block is 16 contiguous baits covering 16.9Kb, and a maximum of 58 contiguous baits covering 104.4Kb (Fig.3.12).

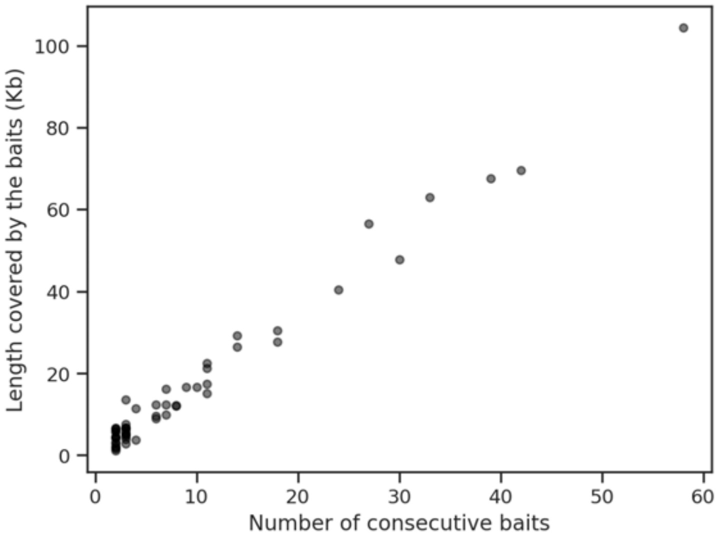


Figure 3.12: Syntenic contiguous baits shared by *B. darthvaderiana*, *B. peltatifolia*, *B. loranthoides*, and *B. masoniana.* Each dot represent a sequence of baits following each other in the same order in the four genomes.

3.3.5 Detection of ancient DNA

We used mapDamage2 (Jónsson et al., 2013) on our herbarium-dried samples to see if we could detect the signature changes associated with ancient DNA. We found similar rates of nucleotides substitution than is the Socotran dataset (Fig.4.7) and did not see any indication of aDNA damage patterns (Fig. 3.13). This is likely due to our library preparation methods which included an end-repair step that might have obscured the signal, or to the recentness of our samples that would display fragmentations patterns, but not nucleotides substitution..

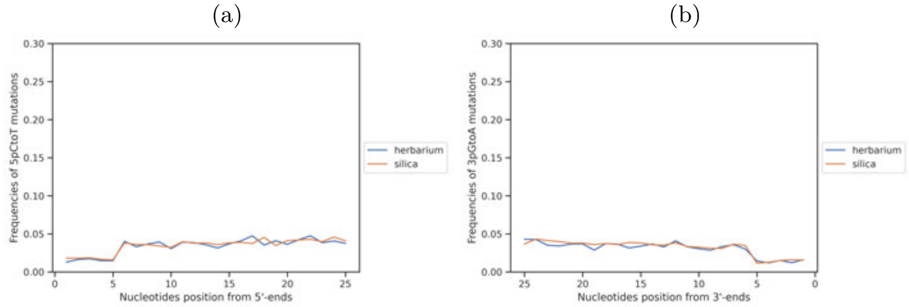


Figure 3.13: Nucleotides substitutions in 5' and 3' of DNA fragments (a) 5' substitutions, (b) 3' substitution. Frequencies of nucleotides substitutions for silica-dried specimens (blue) and herbarium-dried specimens (red).

3.4 Discussion

3.4.1 Partial recovery of targeted sequences

We did not achieve a total recovery of all the targeted sequences in the mapping population specimens with the Hyb-Seq protocol and the *Begonia* baits. The causes of the poor targets recovery are still unclear, but the sequence of the baits should not be the cause of this issue, as our set of baits has been proved able to capture enriched libraries on other similar samples batches (Michel et al., 2022, Wilson et al., 2020). We suspect a degradation of the RNA target capture baits, as they were outdated at the time of the capture. Nonetheless, the sequences recovered were sufficient to set up our methods, and test our analysis pipeline. We have used an additional batch of genome skimming data on the same mapping population specimens for the paralogs detection part of the pipeline (Fan, 2023, Section 3.2.2).

Furthermore, we set up this analysis pipeline to process either fresh specimens or historical herbarium specimen, with degraded aDNA. We can call variants with a high degree of confidence despite low coverage and fragmentary data. The uneven depth of sequencing in this batch might mimic highly degraded specimens and help to understand what are the limits of such an analysis. We might be able to approximate when degraded DNA data are too damaged to yield any genomic and demographic information.

3.4.2 Paralogous sequences detection

Before the analysis of the dataset, a preliminary study on the *Begonia* baits set and multiple genomes has revealed baits with multiples matches across chromosomes/scaffolds and species. These baits have been filtered out from the analysis pipeline as they might capture paralogous sequences or repeated DNA motifs as microsatellites, or Short Tandem Repeats (STRs).

We decided to go further and to detect other problematic baits that might capture paralogous duplicated sequences but are more difficult to detect. Three different methods were used to detect these baits capturing paralogs, based on genotype frequencies ('genfreq' method with home-made scripts), expected heterozygosity rate (HDplot), and multiple matches of assembled contigs on one of the bait (HybPiper). A poor overlap of detection has been observed between the three methods used, which can be explained by several bias in our study.

- The genome skimming dataset used for the genotype frequency method might not be statistically representative, including only one parent and F1 plant, for 80 BC individuals.
- The threshold of genotype frequency to consider a bait capturing orthologous sequences could be selected with more robust statistical tools.
- The identification of problematic baits with HDplot could be improved by the use of clustering analysis to properly select clusters of orthologous sequences.

Nonetheless, 73 baits identified as capturing paralogous sequences for the three methods are likely to represent paralogous regions in our dataset, and will be filtered out of the pipeline in subsequent results.

3.4.3 Syntenic sequences captured by the baits

We used the 4 long genomes produced by Li et al., 2022 to measure the number and the length of ROH that might be detected with our bait set on all the genomes. The Figure 3.12 shows the syntenic regions across species on which our pipeline will be able to measure ROH with Hyb-Seq dataset using the *Begonia* bait set. This mean that analysis of specimens from different species related to the four genomes investigated (*B. darthvaderiana*, *B. peltatifolia*, *B. loranthoides*, and *B. masoniana*) would capture at least the potential ROH in the syntenic sequences described and would be less likely to involve a genetic distance bias in the results.

3.4.4 Next analysis with the pipeline

With these caveats in view we analysed the *B. socotrana* and *B. samhaensis* data to determine whether these species are in genetic danger (chapter 4), and we used a more successful capture to examine demographics of *Begonia* form Papua New Guiana (chapter5).

Chapter 4

Socotran Begonia

4.1 Introduction

4.1.1 The Socotran archipelago

Socotra is an archipelago located in the western Indian Ocean, near the Horn of Africa, 350km to the South of Yemen mainland. It includes three inhabited island: Socotra, Abd al Kuri, and Samha. The fourth island, Darsa, is uninhabited. The main island is Socotra, which cover 95% of the total land of the archipelago, and has an array of geo diversity that includes the Haggeher (or Haggier) mountains, surrounded by limestone plateaux and coastal plains (Brown et al., 2012). The climate is arid, with periodic rainfalls during the monsoon season, and a permanent fog in the highlands (Král et al., 2006, Banfield et al., 2011). This main island has been separated from the Arabian mainland millions of years ago, and a high number of the species found on the island are endemic doe to this isolation. The archipelago counts 835 vascular plant species, of which 37% are endemic, representing 15 endemic genera (Scholte et al., 2011, Riccardi et al., 2020).

4.1.2 Anthropogenic impact on biodiversity

In 1970 the population on the main island of Socotra was estimated at 500 inhabitants. Since, 100,000 inhabitants has been recorded during a 2018-2020 survey. Growth of the population is due to rapid urbanization of the main island, notably of its capital Hadibu (Scholte et al., 2011, Damme, Kay van, 2022). The origin of this demographic change can be attributed to a cultural shift in traditions and practices, including increased fishery, expanding agriculture, and logging. These societal changes are due to political and economic shifts in local communities, but are also for a large part a consequence of the development of foreign tourism. The number of tourist has doubled every 18 months since 2003, and to support the industry the artisanal fishing has doubled over the last decade (Scholte et al., 2011). Where traditional communities were relying on sustainable use of resources for subsistence, an increasing pressure of logging, fishery, and cattle grazing impact the flora and fauna of the Socotra island (Scholte et al., 2011, Riccardi et al., 2020). A migration of local communities in valleys of the mountainous area has increased local goat grazing in a permanent manner, where the valley floor was grazed only after rainfall formerly (Pietsch et al., 2010). Grazing by goat is a serious threat for vegetation, especially for forests as it impacts tree regeneration and soil content (Scholte et al., 2011, Attorre et al., 2007). Restoration programs have been designed especially to address the issue (Rezende et al., 2022).

4.1.3 Case study for conservation

Socotra is under both environmental and anthropic impact. Climate change increases periods of drought and decreases precipitations during rainy seasons, and might increase the occurrences of floods and cyclones (Riccardi et al., 2020, Damme, Kay van, 2022). Drought has already impacted negatively the tree coverage of the Hageher mountains, which were considered as a 'wet refugia' (Scholte et al., 2011). Furthermore, a study on tree coverage on the island found that climate and pedological factors are affecting more tree coverage than roads and settlements (Riccardi et al., 2020).

As several conservation programs have been raised to address climatic and anthropic pressures on the environment, it has been noted that the island is an interesting case study of the impact of anthropogenic activities on woodlands and other plants, notably due to the high level of endemism and potential vulnerability of plant communities (Attorre et al., 2014).

As presence of endemic and rare vegetation on the islands has been well documented since the 19th century, and several historical specimens are available with consequent records (geographical and ecological data, herbarium specimens, living plants collection), an exploration of the influence of climate and anthropic threat on vegetation can be explored with tools such as population genetics, biogeography, and reconstruction of demographic history.

4.1.4 The *Begonia* species from Socotra

Socotran *Begonia* are under serious threat due to their endemism, leading to geographic and genetic isolation combined with other external factors such as direct anthropogenic activity, and climate change. *Begonia socotrana* (Fig.4.1a) and *B. samhaensis* (Fig.4.1b) are restricted to small populations with restricted distribution. Pollen flow in these species is limited and there is no evidence of interbreeding between the populations (Hughes et al., 2002c). Species are suspected to be very homozygous, and possibly at risk of an extreme case of deleterious genetic drift called mutational meltdown. This inbreeding-related syndrome is caused by fixation of deleterious alleles in the population, leading to reduction in population size, faster fixation of deleterious alleles and further reduction in population size (Lynch et al., 1995, Gralka et al., 2016). This study aims to assess the degree of homozygosity within populations over time, and assess their vulnerability for conservation purposes.

While the use of living plants for genome analysis provides sequences of optimum quality, herbarium specimens provide precious information about plant populations in the recent past. The rich herbarium resources of RBGE can be exploited to provide information about the genomes of extinct species, enabling the analysis of rare plants or species not easily available in a living collection (Fig.4.1c, 4.1d). Herbarium specimens collected at different time points and places can be markers of the evolution of plant populations through time and space (Chapter 1).

This project aims to use the pipeline described in Chapter 3 to investigate the genetic health of *Begonia* populations on the Socotra islands, and determine how it has changed over time.

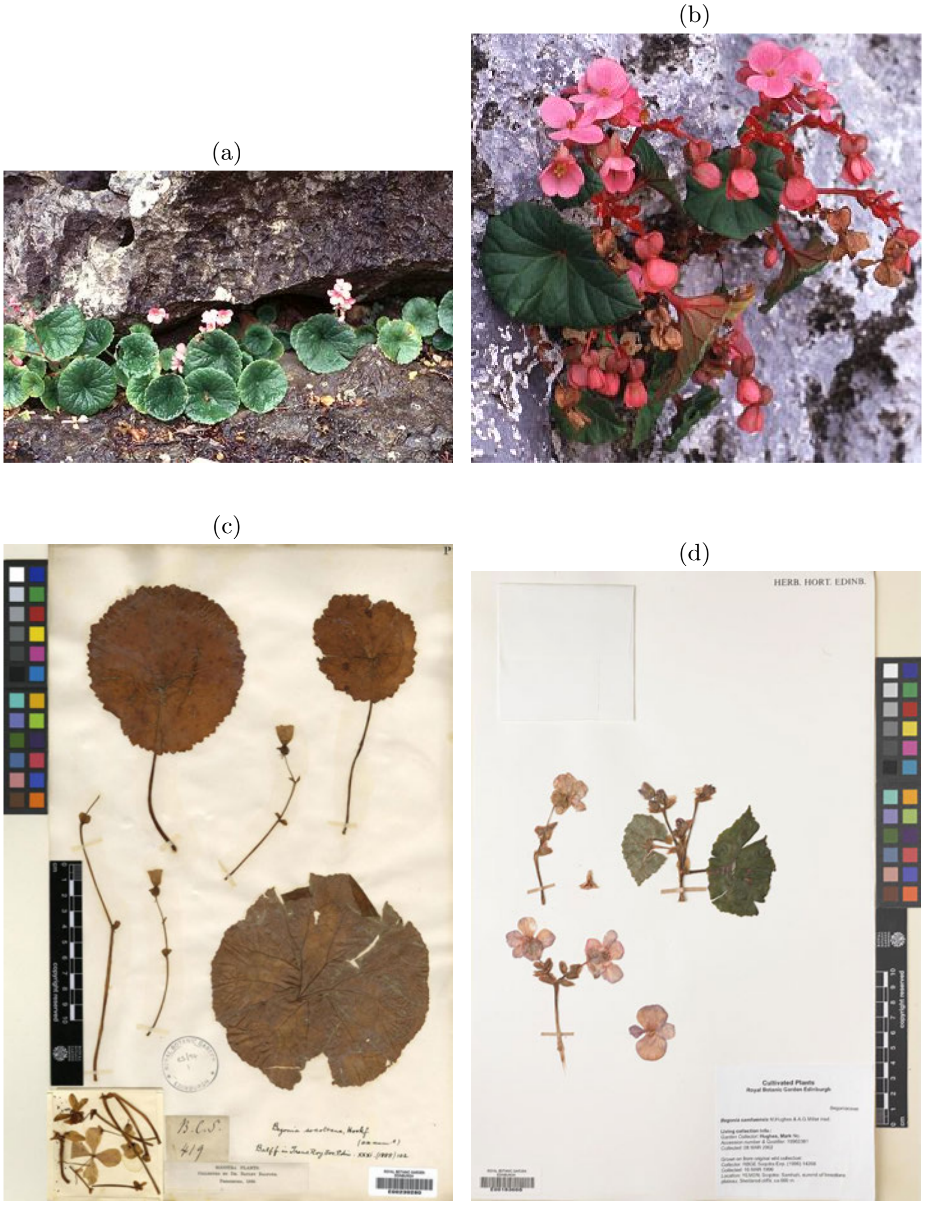


Figure 4.1: Socotran *Begonia* included in the study. (a) *B. socotrana* growing in rock crevice (Mark Hughes), (b) *B. samhaensis* growing on a cliff side (Mark Hughes), (c) *B. socotrana* herbarium specimen (1888), RBGE (Herbarium catalog), (d) *B. samhaensis* herbarium specimen (2002), RBGE (Herbarium catalog).

4.2 Methods

4.2.1 Sampling plan

Two batch of samples were set up for this study. The first batch included specimens from species *B. socotrana* and *B. samhaensis*, the two *Begonia* species present in the Socotra archipelago. The *B. socotrana* species is endemic from the Reyged and Rewgid plateau, and from the Haggeher mountains. We used the living collection at RBGE and herbarium material from partner institutions to form a representative subset of the populations from these area, including present-day and historical material. The *B. samhaensis* species is endemic from the Samha island north-Caing cliff. In reason of its restricted area, the several *B. samhaensis* specimens collected were not included as part of a specific population but included as a subset of the main population representative of the genetic diversity on the island.

The second batch of specimens has been selected later and processed to a target capture sequencing set. The target capture method did not work properly, and they were processed for genome skimming sequencing. Samples selected were horticultural hybrids added to assess the genetic background of these lines, extra Socotrana specimens, rare and widespread related species, and distantly related species as outgroup for phylogenetic studies.

4.2.2 First batch of target capture dataset

In total, 43 specimens have been included in this dataset for the Hyb-Seq analysis. We sampled individuals from the two Socotran species: 9 *B. socotrana* individuals, and 7 *B. samhaensis* individuals from the RBGE living collection chosen based on previous microsatellite genotyping (Hughes et al., 2002c) and 8 *B. socotrana* and 1 *B. samhaensis* from herbarium collections dating from 1880 to 1999. The living specimens of *B. socotrana* have been selected in priority to match the populations sampled by Dr. Mark Hughes and Dr. Antony Miller (Hughes et al., 2002c, Hughes et al., 2002b) and check the genetic distances formerly calculated with polymorphic microsatellite markers (Hughes et al., 2002c, Hughes et al., 2003). The living specimens of *B. samhaensis* have been selected among the available specimens in the living collection. The historical specimens of both species have been donated by partner institutes including RBGE, the Royal Botanic Garden Kew, and the Museum of Evolution of Uppsala University (Table 4.1). Access to historical herbarium specimens was limited for destructive sampling to the content of the capsule, when available.

Silica-dried and herbarium-dried replicates of all the non-historical samples have been included to study the effect of herbarium specimen preparation on the DNA and detect and identify ancient DNA (aDNA) damage patterns.

Table 4.1: First batch of Socotran specimens for target capture sequencing.

Species	Sampling	Condition	ID	Collector	Date
B. socotrana	Skand	Living	19990434		
B. socotrana	Dicksam	Living	19990433		
B. socotrana	Reiged	Living	19990422		
B. socotrana	Reiged	Living	19990424		
B. socotrana	Reiged	Living	19990425		
B. socotrana	West Haggier	Living	20000299		
B. socotrana	West Haggier	Living	20000303		
B. socotrana	West Haggier	Living	20000304		
B. socotrana	West Haggier	Living	20000308		
B. samhaensis	Samha	Living	19990395		
B. samhaensis	Samha	Living	19990396		
B. samhaensis	Samha	Living	19990398		
B. samhaensis	Samha	Living	19990400		
B. samhaensis	Samha	Living	19990409		
B. samhaensis	Samha	Living	19990410		
B. samhaensis	Samha	Living	19990405		
B. socotrana	no indications	Herbarium		Balfour	1880
B. socotrana	no indications	Herbarium	E00299131	T.M. Bent	1897
B. socotrana	no indications	Herbarium		T.M. Bent	1897
B. socotrana	no indications	Herbarium	8667	A.G. Miller	1989
B. socotrana	Aduno pass	Herbarium	E00299132	A.G. Miller	1989
B. socotrana	Reiged plateau	Herbarium		A.G. Miller	1989
B. socotrana	Reiged plateau	Herbarium	M.8335	A.G. Miller	1989
B. socotrana	Reiged plateau	Herbarium		M.Thulin A.N.Gifri	1994
B. samhaensis	Samha	Herbarium	E00239279	M. Hughes, AG. Miller	1999

4.2.3 Second batch of genome skimming sequencing

A second set of samples was added to our study in 2021 (Table 4.2) in order to investigate the genome of horticultural hybrids from a *B. socotrana* cross, add other population of *B. Socotrana* specimens, add rare and widespread related species to compare their level of homozygosity, and add an outgroup for phylogenetic studies. We attempted three captures with these which all failed, so the remaining libraries were sequenced with genome skimming sequencing. This second batch included samples for horticultural lines to detect Socotran-specific alleles in breeding material, and assess the genetic robustness of those lines. This material was kindly provided by the Royal Botanic Garden Kew, the Royal Botanic Garden of Edinburgh, Dr. Mark Hughes, and an anonymous source. Several *Begonia* from other groups were also added to test hypothesis about their degree of inbreeding and to contrast rare and widespread related species, notably *B. maxwelliana.* Finally, a few species were included to fill in sampling gaps for the full *Begonia* phylogeny reconstruction, including African specimens from Tanzania, Kenya, Gabon, and one Southern Asian specimen from Sri Lanka.

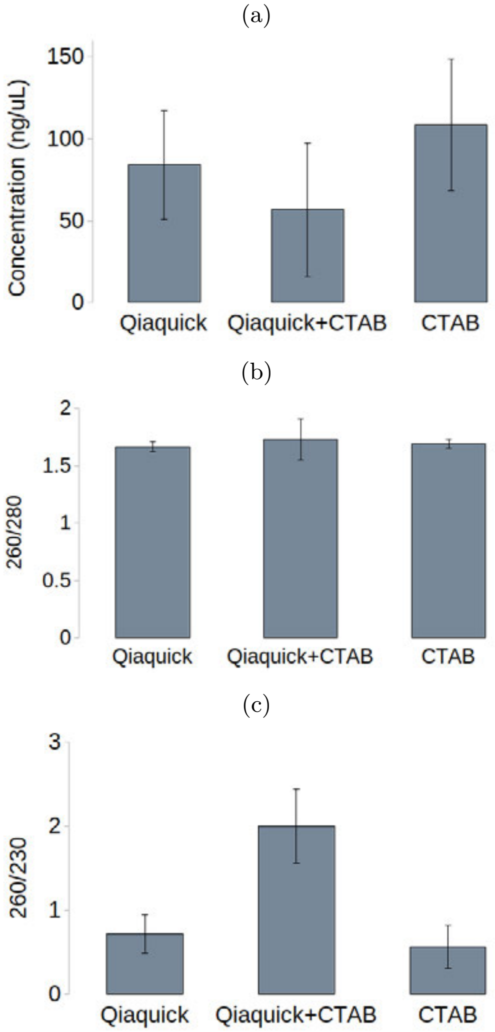
Table 4.2: Second batch of specimens for genome skimming sequencing.

Species	Sampling	Condition	ID	Date	Collector
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
Horticulture hybrid old	Horticulture	Silica	None	12/2020	Anonymous
B. sutherlandii	Tanzania	Silica	SGNO198.196	Unknown	M.Hughes
B. meyeris-johannis	Kenya	Silica	SGNO196.8790	Unknown	J.J.de Wilde
B. socotrana	Yemen	Silica	SGNO201.48	22/02/1999	M.Hughes
B. oxyloba	Gabon	Silica	SGNO196.744	Unknown	J.J.de Wilde
B. tenera	Sri Lanka	Herbarium	E00656886	20/02/2013	L.Kumarage
B. maxwelliana	Malaysia	Herbarium	E00300547	16/09/1949	J.Sinclair
B. maxwelliana	Malaysia	Herbarium	E00879845	13/02/2003	R.Kiew

4.2.4 DNA extraction

Several protocols have been established to isolate DNA from plant tissues with maximum quantity and quality. The difficulties of plant DNA extraction are linked to the presence of polysaccharides, polyphenols, and others secondary metabolites (Aboul-Maaty et al., 2019). Historically, CTAB methods have been widely used and specific protocols have been dedicated to specific group of plants, as CTAB extraction for *Musa* and *Ipomoea* (Gawel et al., 1991), rain forest plants (Scott et al., 1996), or *Zingiberales* (Devi et al., 2013). More recent extraction methods involve the use of silicate extraction columns kits that have gained popularity for their simplicity of use, being time efficient, and increasing potential yields (Abdel-Latif et al., 2017). On this study we have used the later method for increased yields as the batch of samples included historical specimens from herbarium specimens.

Herbarium specimens are precious museum items available in very limited quantity if no unique, and the amount of plant material involved is scarce, often limited to a few grams of leaves or stem tissue. We decided to try different extraction methods to know what would be the best one to isolate DNA from precious *Begonia* historical specimens. Several extraction optimization assays were tested on the herbarium-dried non-historical specimens of the mapping population, including a protocol with CTAB extraction (Särkinen et al., 2012), silicate column extraction (DNeasy Plant Mini Kit from Qiagen, catalogue number 69104), and a compound of the two methods (Gutaker et al., 2016) (Fig.4.2, Fig.4.3). DNA preps were checked for quantity and quality on gels, with a Tapestation Genomic DNA ScreenTape, and with a Quant-iT Qubit dsDNA HS Assay Kit. We tested three factors related to the quantity and quality of the DNA extracted: concentration, spectrophotometry ratio 260/280 indicative of the purity of DNA of protein and phenol contaminants, and spectrophotometry ratio 260/230 indicative of the nucleic acid purity. Three methods achieved significantly similar results for concentration and spectrophotometry ratio 260/280. The mixed approach Quiaquick+CTAB achieved significantly better score for spectrophotometry ratio 260/230 than for the Quiaquick columns or CTAB extraction alone. The silicate column extraction method was eventually used for all specimens in reason of its effectiveness on the samples processing workflow with no loss of quantity of DNA extracted compared to the other methods (Fig.4.2, Fig.4.3). A very low amount of genetic material, 100pg minimum, can be sufficient for library preparation, and achieve good coverage (Durvasula et al., 2017; Nicholls et al., 2015; Kopperud et al., 1995). Despite a high level of degradation in the historical herbarium samples (Fig.4.4), the concentration of the DNA isolated was ranging from 3.78 to 7.68 ng/uL and comparable to the silica-dried samples concentration.


(a)

(b)

(c)

Figure 4.2: Parameters of DNA extraction per method. (a) Concentration, (b) 260/280 ratio, (c) 260/230 ratio.

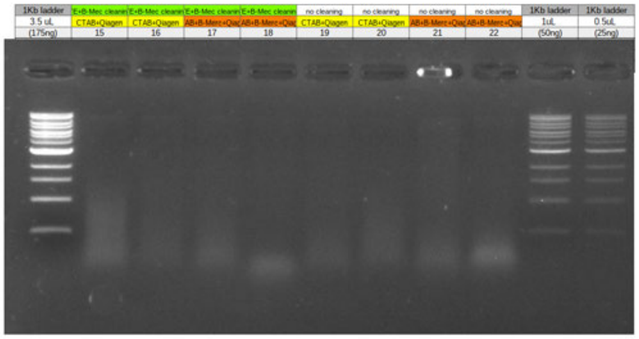


Figure 4.3: Agarose gel analysis of herbarium DNA extracted with a range of protocols.

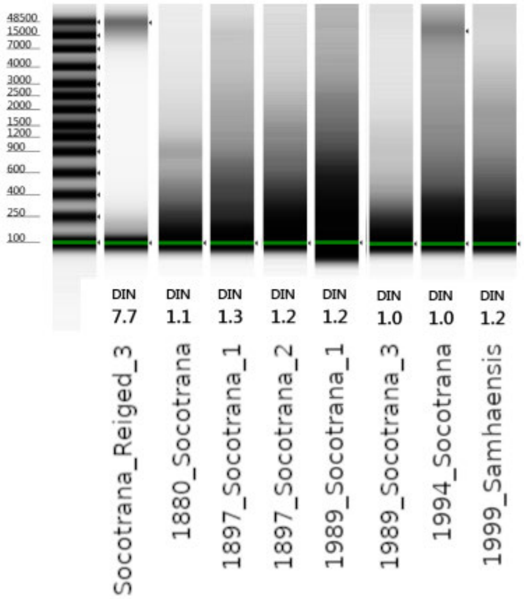


Figure 4.4: Electrophoresis of Socotran specimens DNA. Socotran_Reiged_3 has been silica-dried from fresh sample, the other samples are from historical herbarium specimens.

4.2.5 Library preparation

All samples were normalized to 2 ng/uL before fragmentation step. Fresh and recent historical specimens were fragmented to 350bp using a Covaris M220

Focused-ultrasonicator. Library preparation followed the protocol of the NebNext Ultra II DNA Library Prep for Illumina Kit. Seramag Sample Purification beads were used for size selection of samples above 50 ng, and clean up for less concentrated samples. An Agilent Tapestation with High Sensitivity kit was used for libraries quality check. Subsequently, libraries have been normalised to 10 nM, then pooled according to fragment size and quality. Three pools of 10, 14, and 19 libraries were made.

4.2.6 Target capture on Socotran *Begonia*

The bait set used for the target capture has been designed based on the transcriptome of *B. luzhaiensis* with additional baits from the *B. conchifolia* genome and is described in (Michel et al., 2022). The hybridisation and sequencing is described in Chapter 3, it followed the the MyBaits Hybridization Capture for Targeted NGS Manual version 4.01. Pools were sequenced by Edinburgh Genomics on a single lane of NovaSeq6000 SP with 250bp paired end reads. The number of reads expected was 375 million paired-end.

4.2.7 Sequencing

To estimate the rate of success of the target capture, the trimmed reads of the Hyb-Seq batch were aligned with the sequences of the bait set using Bowtie2 (Li et al., 2009b) with standard settings for local sensitive alignment. The alignment statistics were obtained with Samtools flagstat from the BAM files (Li et al., 2010) and are shown in Table 4.3.

4.2.8 Second capture and genome skimming

Samples for the second capture (Table 4.2) were prepared as for the first, and hybrid capture followed the same protocol, but used a different bait kit, which had already been used successfully by a previous student. Although the libraries looked good before capture, the post-capture Tapestation analysis showed capture had failed completely. Repeated attempts did not remedy this. It was decided to use genome skims on the remaining libraries to obtain some data for these samples, but coverage was not sufficient for useful analysis (Table 4.3). The samples have been sequenced for genome skimming on a single lane of a NovaSeq SP 150PE, yielding 375M reads pairs.

4.2.9 Historical patterns of damage

Two types of DNA degradation are common in herbaria material. DNA fragmentation is the most common, observed at various degrees in silica-dried and herbarium material (Forrest et al., 2019). Due to the process of depurination of the DNA sugar phosphate backbone and β-elimination, it occurs in slowly-dried specimens, specimens exposed to heat, or exposed to moisture during storage (Dabney et al., 2013). Substitution of C to T and G to A, caused by

Table 4.3: Statistics of first and second batch sequencing.

Specimen	Batch	Retained reads	Unique hits	Unique hits fraction	Coverage
Socotrana__Skand	1	22,754,035	945,799	4.16%	24.02
Socotrana__Dicksam	1	25,747,056	1 ,224,551	4.76%	29.28
Socotrana__Reiged__1	1	23,379,363	1,029,067	4.40%	26.06
Socotrana__Reiged__2	1	30,722,542	1,124,063	3.66%	29.65
Socotrana__Reiged__3	1	11,161,873	642,353	5.75%	14.87
Socotrana__West__Haggier__1	1	20,437,560	720,010	3.52%	19.42
Socotrana__West__Haggier__2	1	24,091,198	953,371	3.96%	23.65
Socotrana__West__Haggier__3	1	17,796,774	922,075	5.18%	21.66
Socotrana__West__Haggier__4	1	25,324,837	972,673	3.84%	26.75
Samhaensis__1	1	28,272	1,831	6.48%	0.05
Samhaensis__2	1	16,178,717	924,910	5.72%	21.75
Samhaensis__3	1	16,168,785	1,079,928	6.68%	23.90
Samhaensis__4	1	15,950,461	788,115	4.94%	19.08
Samhaensis__5	1	15,068,842	697,936	4.63%	18.10
Samhaensis__6	1	17,965,483	679,365	3.78%	18.27
Samhaensis__7	1	23,916,437	815,646	3.41%	23.19
1880__Socotrana	1	3,193,204	8,094	0.25%	0.20
1897__Socotrana__1	1	1,986,583	5,050	0.25%	0.11
1897__Socotrana__2	1	4,088,499	102,089	2.50%	2.28
1989__Socotrana__1	1	21,857,396	932,428	4.27%	21.19
1989__Socotrana__2	1	3,747,683	102,035	2.72%	2.74
1989__Socotrana__3	1	16,837,994	540,985	3.21%	12.77
1989__Socotrana__4	1	14,709,730	621,635	4.23%	14.28
1994__Socotrana	1	9,242,712	160,106	1.73%	3.96
1999__Samhaensis	1	13,357,415	398,315	2.98%	9.78
Hort__socotrana	2	32	2	6.25%	0.00
Hort__hybrid__old	2	452,458	7,852	1.74%	0.30
Hort__hybrid__old	2	25,403	465	1.83%	0.02
Hort__hybrid__old	2	6	1	16.67%	0.00
Hort__hybrid__young	2	912,073	17,656	1.94%	0.64
Hort__hybrid__young	2	1,386,219	25,985	1.87%	1.00
Hort__hybrid__young	2	366,938	4,890	1.33%	0.19
Hort__hybrid__young	2	1,000,215	15,956	1.60%	0.60
Hort__hybrid__young	2	4,612	75	1.63%	0.00
B.suterlandii	2	15,458	129	0.83%	0.01
B.meyeris-Johannis	2	3,599,448	22,939	0.64%	0.90
B.socotrana	2	2,925	94	3.21%	0.00
B.oxyloba	2	630	23	3.65%	0.00
B.tenera	2	18,617	60	0.32%	0.00
B.tenera	2	2,334,988	2,435	0.10%	0.08
B.maxwelliana	2	7,471	46	0.62%	0.00

nucleotide deamination and located at extremities of DNA fragments occurs during amplification of aDNA, and is a more reliable marker of aDNA.

In the Socotran set of samples, DNA fragmentation of historical specimens was observed during molecular work (Fig.4.3), and nucleotide substitution rate was measured with MapDamage (Jónsson et al., 2013, Chapter 3).

4.2.10 Selection of Socotran-specific alleles

To accurately compare the different populations of Socotran *Begonia* in the dataset and subsequent work on horticultural hybrids, the loci with reliable SNPs called by our pipeline were selected for population genetics and phylogenetic reconstruction. Only nuclear loci were considered in this study instead of plastids or mitochondrial loci as the nuclear target capture allow to increase the phylogenetic resolution (Nicholls et al., 2015, Michel et al., 2022), to proceed to subsequent population genetics analysis using ROH on nuclear pseudo-chromosomes (Li et al., 2022), and to discriminate different *B. socotrana* haplotypes. Socotran-specific alleles were identified as fixed shared alleles in specific populations (Fig.4.5). Only silica-dried samples from 2018 were considered for this as only these samples had sufficient depth of coverage to have full confidence in the SNP calling for a reference set.

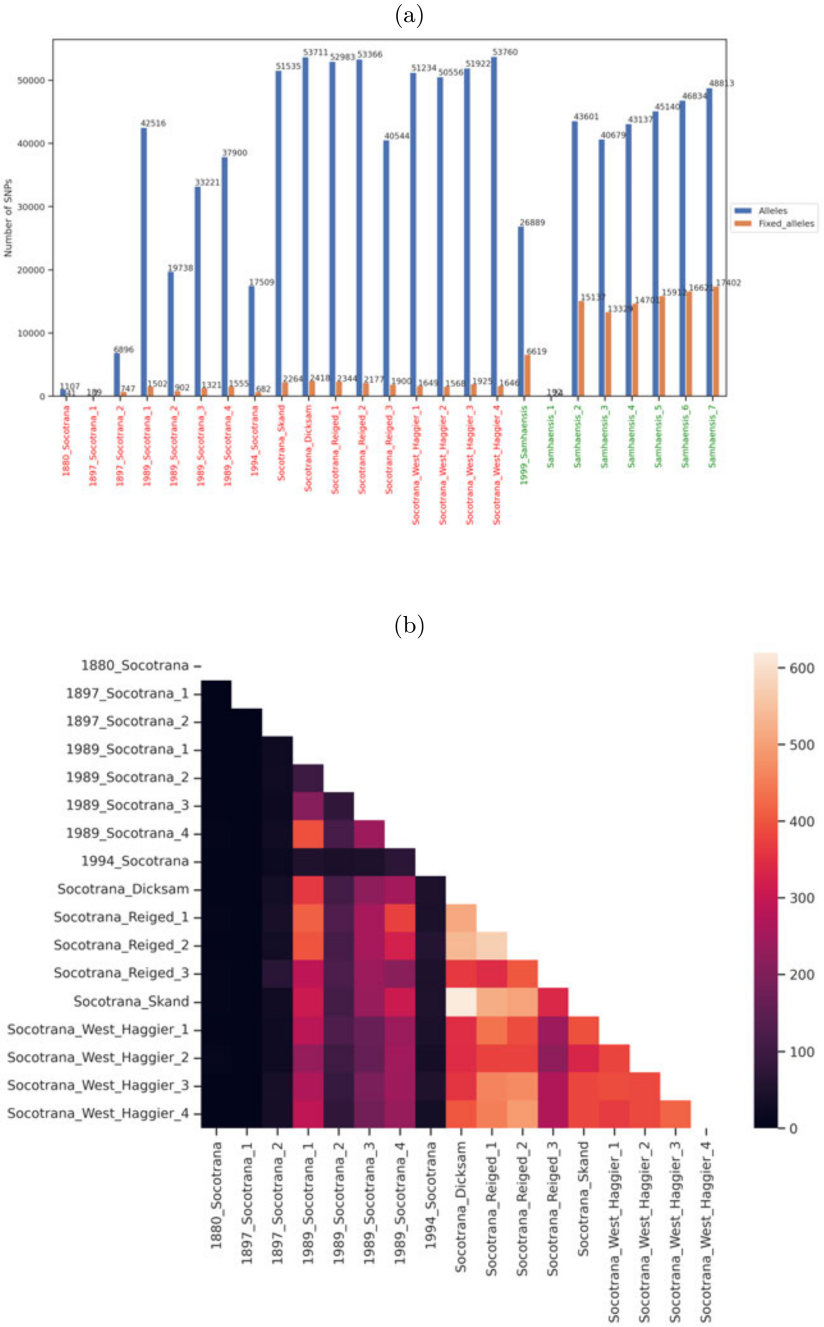


Figure 4.5: Socotran-specific fixed alleles shared among specimens. (a) Total number of SNPs and number of fixed Socotran-specific SNPs in the specimens, (b) Number of fixed Socotran-specific SNPs shared between specimens.

4.2.11 Phylogenetic analysis

After processing the dataset through our pipeline, the VCF variants files were processed by the tool vcf2phylip (Ortiz, 2019) to produce a PHYLogeny Inference Package (PHYLIP) file. Following this the data has been taken as an input by IQTREE (Nguyen et al., 2015, Minh et al., 2020). The substitution model selected running ModelFinder (Kalyaanamoorthy et al., 2017) is PMB+F+R2. The phylogenetic trees were drawn using Figtree (Rambaut, 2022).

4.3 Results

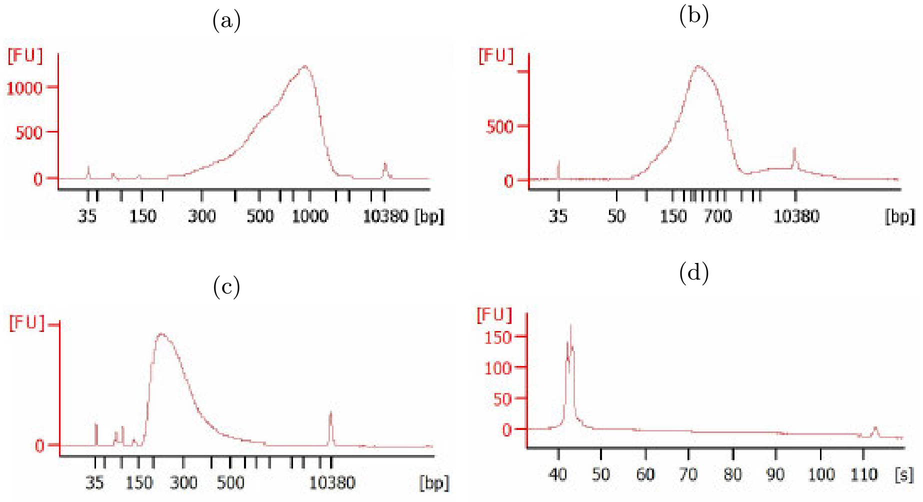


Figure 4.6: Electrophoregram of pre and post capture genomic libraries. (a) Pre-capture *B.socotrana* Skand library batch 1, (b) Post-capture pool 2 of genomic libraries batch 1, (c) Pre-capture *B.socotrana* library batch 2, (d) Post-capture pool 2 of genomic libraries batch 2.

4.3.1 Historical patterns of damage

As expected from herbarium slow-dried specimens, DNA fragmentation what observed in the Socotran set of samples during molecular work (Fig.4.3), To estimate the amount of aDNA-specific patterns of damages in the dataset, nucleotide substitution rate was measured with MapDamage (Jónsson et al., 2013, Chapter 3). Results show a very low level of base substitutions in 5' (Fig.4.7a), and 3' (Fig.4.7b) of the DNA fragments. The expected curve of increasing substitution rate at both extremities is not observable. It has been hypothesized that the USER enzyme, part of the Illumina library kits used during library preparation, prevents the amplification of deaminated bases, and therefore lowers the aDNA substitution damage patterns.

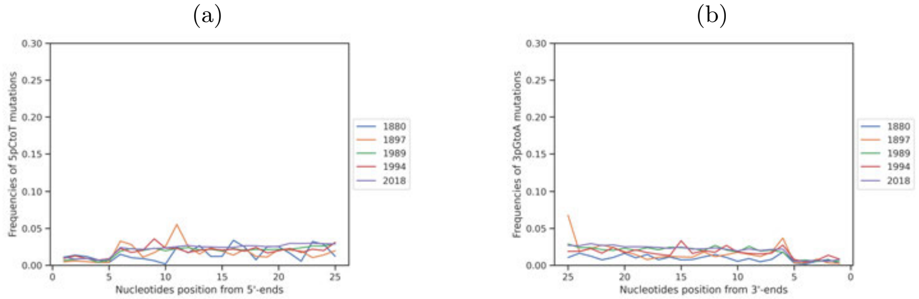


Figure 4.7: Ancient DNA damages patterns in the Socotran dataset. (a) Substitution C to T in 5' region of the reads, (b) Substitution G to A in 3' region of the reads.

Target capture reads recovery

We proceeded to the capture of the first batch of samples 4.1, which provided fragmentary, patchy data. Then, a subsequent second batch of samples was set up for target capture and sequencing 4.2. This time, the capture did not work at all despite multiple trials (Fig.4.6), the remaining genomic libraries were processed through genome skimming, and were not used in the rest of this study.

Alignment of the first dataset to the bait set revealed a poor and variable capture efficiency. The number of reads on-target was 14% on average (Fig.4.8a). No clear reasons were found for the failure of the target capture protocols. The MyBaits protocol recommendations were followed, the genomic libraries had looked good pre- and post-capture (Fig.4.6), and the bait set worked efficiently for other captures (Michel et al., 2022). The possible explanations are a slightly outdated bait set for the first capture, and degradation of the baits during transportation on dry ice from one facility to the other for the second capture.

PCR duplicates

The detection and removal of the PCR duplicates in collapsed reads was done with Picard MarkDuplicates (*Picard Toolkit* 2019). We estimated that for each sample, 85.3% of the reads on average are PCR duplicates (Fig.4.8). This very high level could be due to low capture efficiency generating a low-complexity pool for the post-capture PCRs (Fig.4.8b).

On-target unique hits

Post-filtering, the average number of on-target hits on total number of reads was 3.1% (Table 4.3). On the 564,804,482 reads collapsed by the pipeline, 17,816,448 were unique hits in-target. Despite the relatively low amount of reads recovered after filtering, the average read depth for all samples was 5.6X (Fig.4.9).

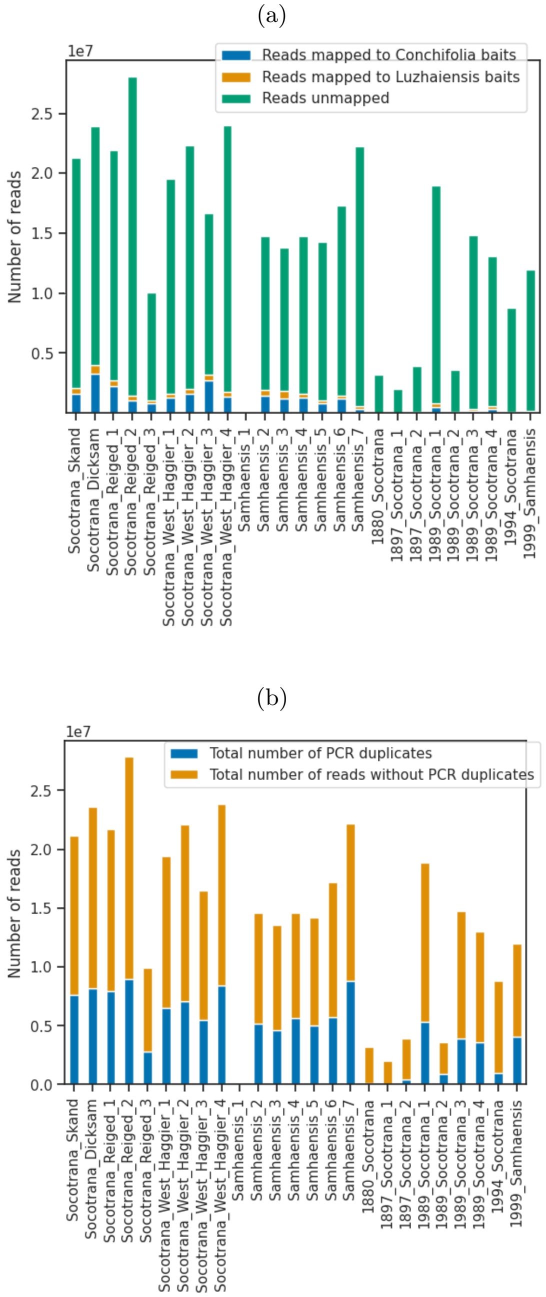


Figure 4.8: Sequencing read content. (a) Number of reads aligned to the sequence of the target capture baits and number of reads not aligned at all, (b) Number of PCR duplicates and non-duplicates reads in the dataset.

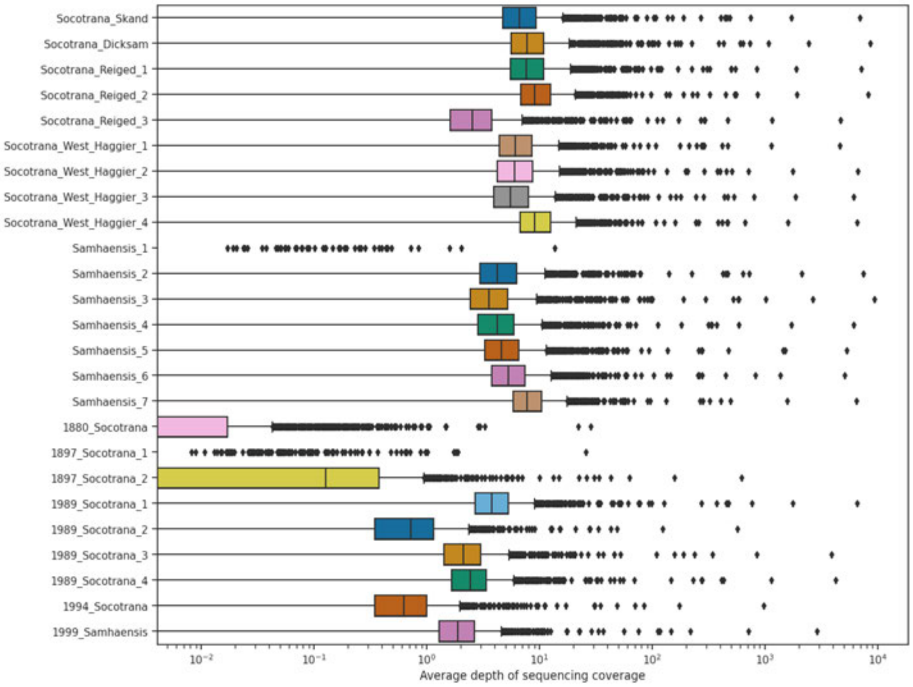


Figure 4.9: Average depth of sequencing coverage in the Socotran data set.

Patchiness of capture

A closer inspection of the reads mapped to the original target sequences shows that several baits have a dramatically higher rate of capture than the rest of the set. The depth of coverage measured on each bait is uneven and vary drastically between specimens (Fig.). The highest value of average depth of sequencing is 243,289 on a single bait. No value per bait is uniformly distributed across all samples, suggesting that the capture efficiency varies randomly between samples.

'Bad baits'

Considering the baits with high depth of sequencing, just three baits captured 31% of the reads from the Socotran living samples, 48% of the Socotran historical samples reads. These multiple matches may be indicative of problematic bait behaviour such as capture of repetitive sequences and seems to be particularly an issue in the genome of *B. socotrana*. Examining the sequences revealed that repeated dinucleotides sequences are present on these three baits of the *Begonia* set and are likely responsible for the uneven rate of capture (Table 4.4). It seems that many of the Socotran reads are actually microsatellites captured by at least 8 baits of the target capture baits set.

Figure 4.12 shows the coverage per target for the eight targets with highest coverage. These eight targets did not have correspondingly high capture rates

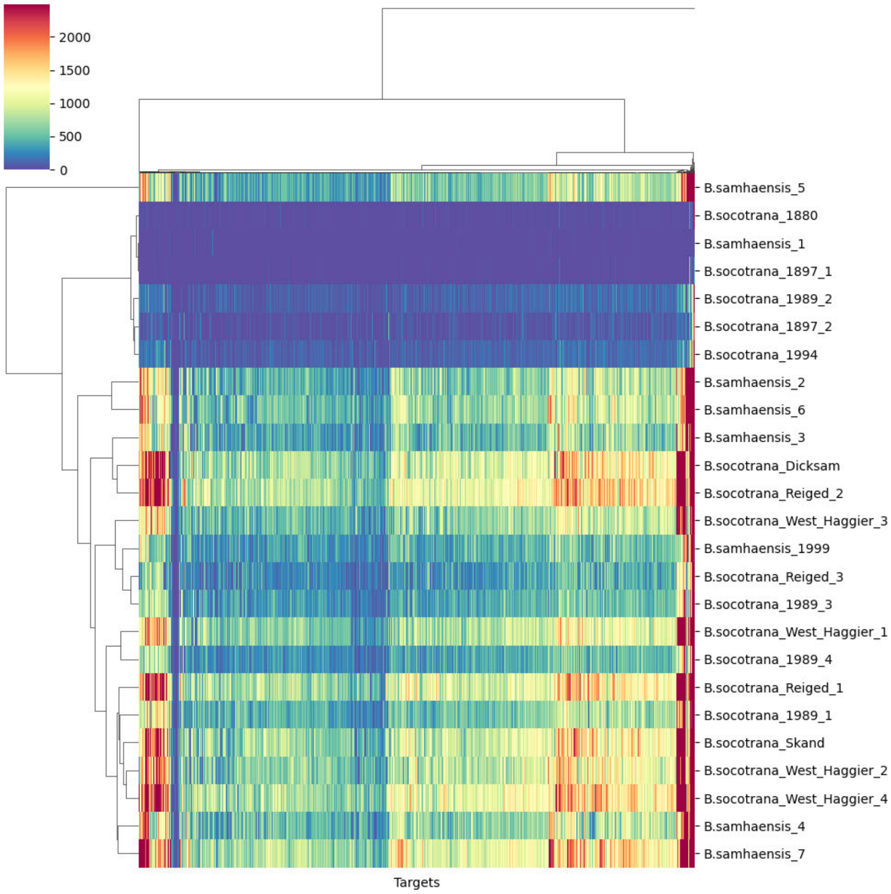


Figure 4.10: Average depth of sequencing coverage of Socotran specimens per bait.

in any other capture with the same bait set (Michel et al., 2022). Examination of these target loci revealed simple sequence repeats (SSRs) in six of them (Table.4.4).

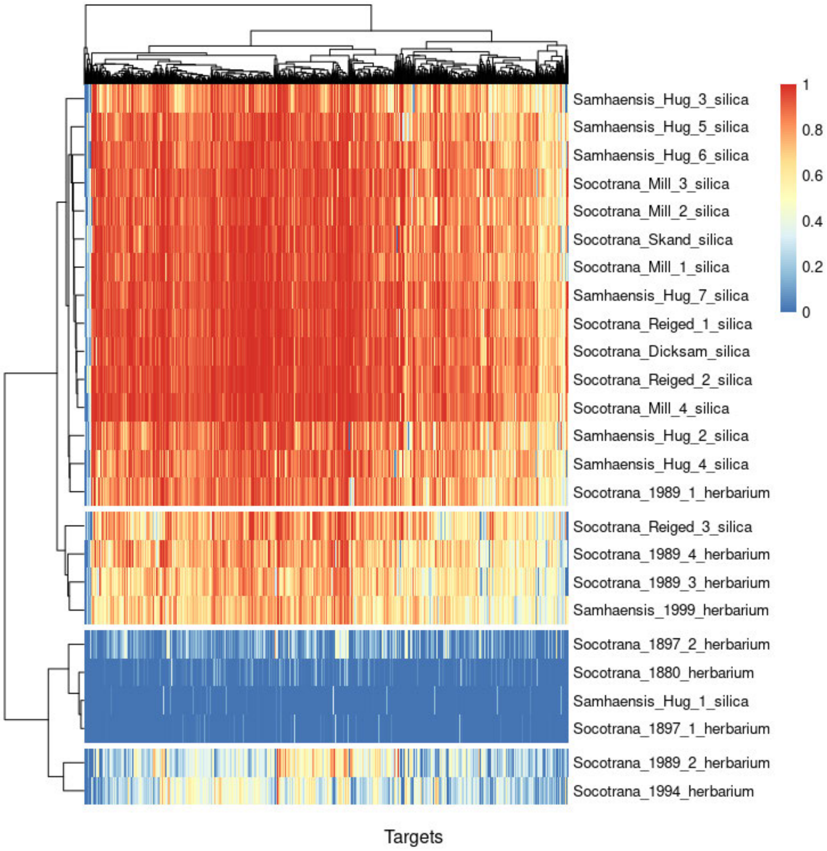


Figure 4.11: Rate of the bait reference sequences covered by the sequencing reads for each baits.

Table 4.4: Single sequence repeats in the *Begonia* bait set.

Bait ID	SSRs
Becon104Scf00540g0006.1	CT_{32}
Becon104Scf00540g0002.1	CT_{26}
ACmerged_contig_9951	AG_{19}
ACmerged_contig_1166	AG_{17}
ACmerged_contig_2307	CT_{11}
Becon104Scf01167g0029.1	TG_{10} AG_{12}
ACmerged_contig_5451	T_{37}
ACmerged_contig_20957	Myb domain

The SSRs showed high capture rates in samples from *B. socotrana Hook.f.* and *B. samhaensis M.Hughes & A.G.Mill.*. The poly-T motive and the myb domain showed relatively high capture rates in all samples.

To understand the bias introduced introduced by this discrepancy in the capture recovery rate, we looked to see if targets from a closely related species were captured better than targets from a more distant species. Figure 4.13 shows the comparison between log percentage capture by target for species in the this dataset. There was a greater range of capture efficiency in the targets based on sequences from *B. luzhaiensis* as there were many more targets (1,192 compared to 47 *B. conchifolia* targets) but the baits designed from *B. conchifolia* targets did not capture better than the ones derived from *B. luzhaiensis*, even in *B. conchifolia* samples.

Eventually, the problematic 'bad baits' were removed from our analysis as they can affect the reliability of the variant calling step.

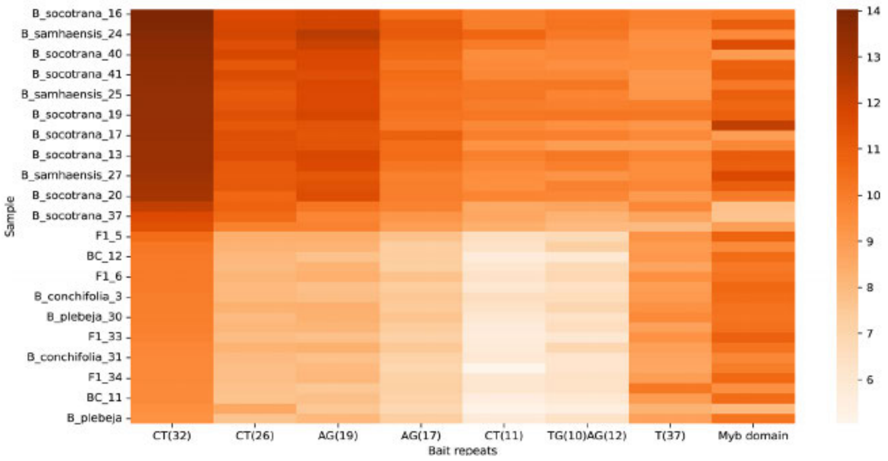


Figure 4.12: Heat map of log read capture for the eight targets with exceptionally high mean capture rates with the *Begonia* baits set. (Michel et al., 2022).

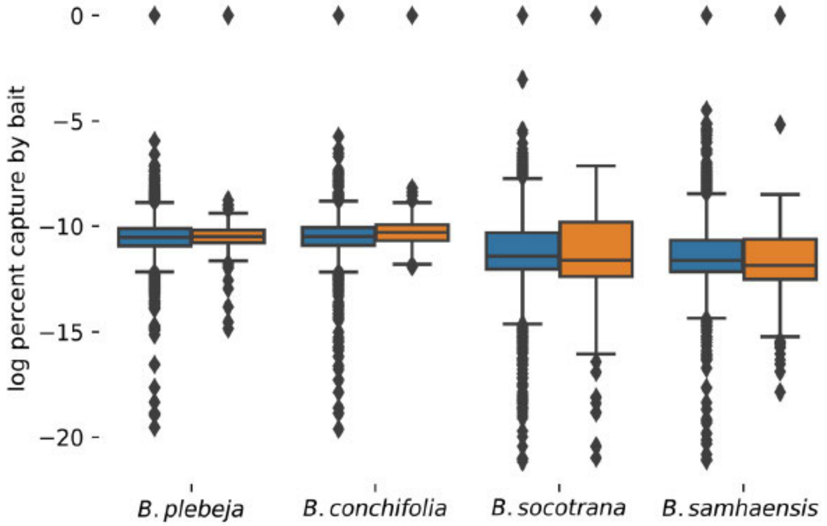


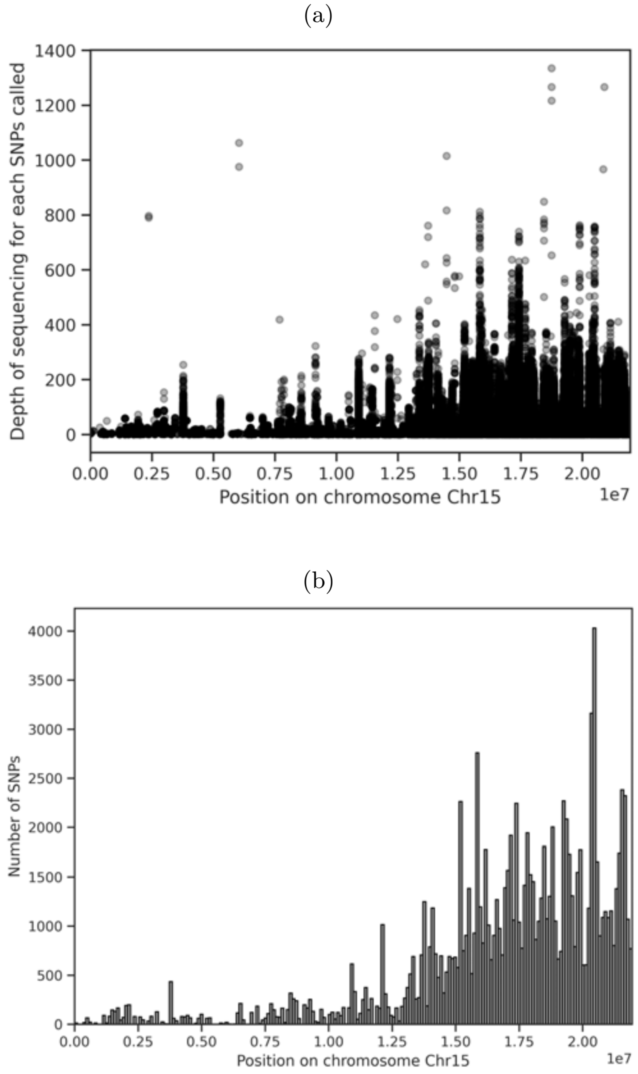
Figure 4.13: Capture by phylogenetic distance between target and sample. Log percentage read capture per target per species. Blue: baits designed on *Begonia* luzhaiensis. Orange: baits designed on *Begonia* conchifolia.
(Michel et al., 2022).

4.3.2 The effect of uneven coverage on variant calling for ROH

Socotran captured reads were aligned to the genome of *Begonia peltatifolia* (Li et al., 2022). This genome was been chosen as the most closely related genome with a chromosome-level assembly, to enable discovery of longer ROHs. The variants were called in the dataset following the pipeline standard (Chapter 3). The location of called SNPs on chromosomes was uneven, several chromosomic regions displayed far higher SNP frequencies than others (Fig.4.14a) Comparing these regions with the read depth of sequencing, we observe that the SNPs number called is simply proportional to the read depth of sequencing, so will vary along the chromosome with bait distribution, linked to genetic content (Fig.4.14b).

4.3.3 The effect of poor target capture on ROH detection

ROH were detected using the base parameters of PLINK. The size of the ROH detected ranged from 1kB to 654 kB, however the distribution of ROH size is generally low, as the mean size is 2.5 kB (Fig.4.15a). The few very large ROH observed were due to the low thresholds used in PLINK (minimum 1kB to consider an SNP). The largest ROH identified by PLINK had only 66 SNPs, while the mean value is 161 SNPs per ROH (Fig.4.15b).


(a)

(b)

Figure 4.14: Parameters of SNPs called with the pipeline on *B. peltatifolia* chromosome 15. (a) Depth of sequencing coverage of the SNPs called and their distribution on *B. peltatifolia* Chromosome 15. The depth of sequencing associated with each SNPs was extracted from the VCF file and represented at their location on the chromosome, (b) SNPs distribution on *B. peltatifolia* Chromosome 15. The number of SNPs called on each position on the chromosome has been represented on the barplot.

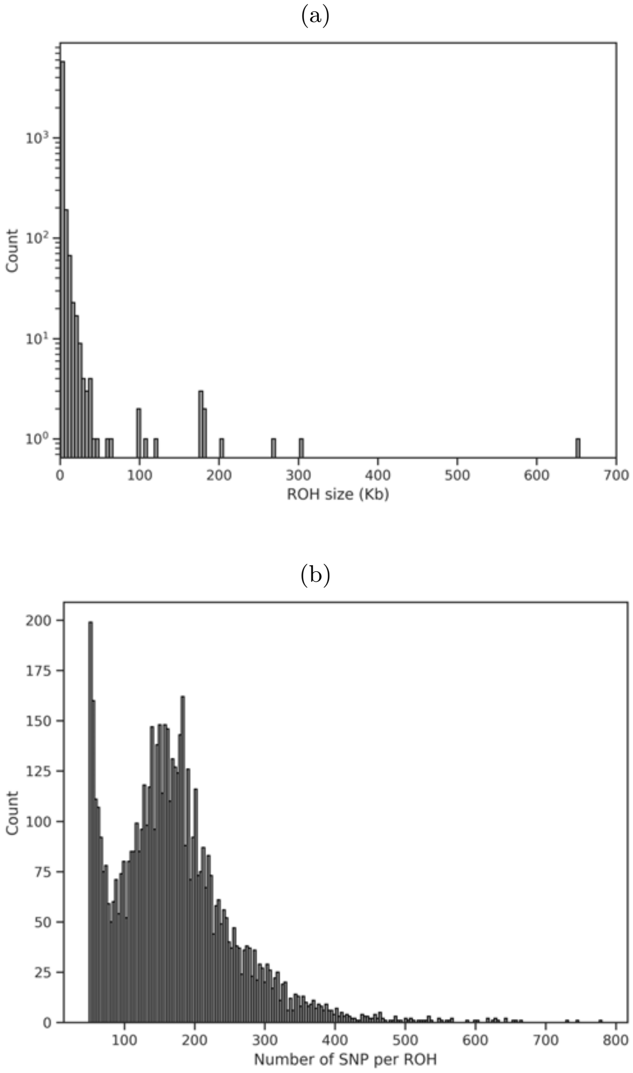


Figure 4.15: Distribution of ROH parameters for the Socotran dataset. (a) Distribution of the ROH by size expressed in Kilobases (Kb) in the Socotran dataset, (b) Distribution of the ROH by number of SNPs in the Socotran dataset.

4.3.4 Filtering Runs of Homozygosity (ROH) in the dataset

The equation given by Kumar et al., 2021 filters the dataset to get rid of ROH appearing by chance in the VCF files and calculates the average rate of homozygous SNPs in the dataset with an interval of confidence of 5%. The threshold calculated for the whole set of samples is l = 856 SNPs to consider a ROH. The number of SNPs per ROH in the dataset is on average 161, and the highest value is 780 (Fig.4.15a). These relatively low values are certainly due to the fragmented gene recovery caused by the failed target capture protocol. Therefore, as these values are below the threshold of minimal SNPs defined, no robust assessment can be made on the homozygosity of the samples based on the ROHs in the set. The average heterozygosity observed in the dataset is 2%, therefore there is 98% probability that any SNP observed is homozygous. Given there are on average 308,892 SNPs observed per individual and 43 individuals, the probability to observe randomly generated ROHs across all subject is: 0.98^{856} x 308892 x 43=0.05. A minimum length of 856 SNPs would be required to randomly generate less than 5% of randomly generated ROHs. Unfortunately, our dataset has short ROHs (average 2.5 kB) defined with low number of SNPs (average 161), due to the uneven capture and unequal distribution of SNP called (4.15). For comparison, the Papua New Guinea dataset had 3,331 ROH after filtration, with an average size of 4.6 kB. Consequently, we cannot use these ROH measurements without considering that more of 5% of the ROH observed would be randomly generated, and do not represent a biological reality. Keeping this limitation in mind, subsequent analysis have been made to explore the limitation of ROH analysis with biased data and assess the limitations of the method.

4.3.5 Phylogenetic reconstruction

A phylogenetic reconstruction of the Socotran specimens was made with the set of SNP data generated by the pipeline in order to compare our result to the results of Mark Hughes (Hughes et al., 2002c) and assess whether the SNPs contain useful biological data. Three reconstructions have been made: one with the full set of data (60,978 SNPs involved) (Fig.4.16a), one limited to analysis of the loci present in 1989 specimens (19,738 SNPs) (Fig.4.16b), and one with loci present in the 1880 specimen, the older and most degraded specimen in the set, providing the lowest number of loci considered (Fig.4.16c).

In all datasets *B. socotrana* and *B. samhaensis* resolved separately except for samhaensis_1, which grouped with the 1,897 *B. socotrana* samples. All three samples had very low coverage and this is likely an artifact based on missing data patterns. The historical *B. samhaensis* was correctly placed with fresh samples from that species, despite a capture efficiency of less than 10%. The historical *B. socotrana* samples all grouped together, this could also be due to missing data, as outside this set we were able to resolve some geographic variation with samples from Skand and Dicksam as sisters in the full dataset as in Hughes et al., 2002c, but where there are multiple samples per population (West Haggier, Reiged) we do not see the clustering of samples from the same populations noted in the microsatellite data. The support values on the smaller datasets are lower than those seen on the full dataset, where most nodes have 100% support. This suggests that this very poor data set is adequate for species assignment but not sufficient for analysis of population-level relationships.

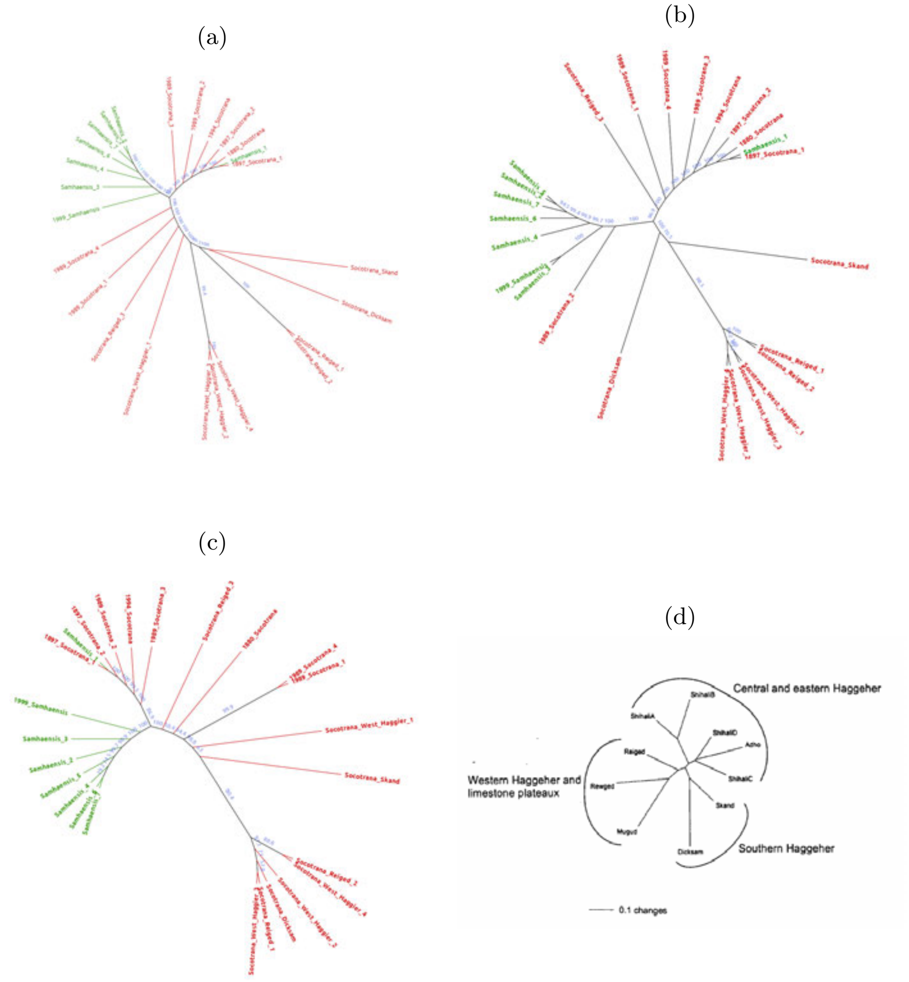


Figure 4.16: Cladograms of the Socotran specimens generated with all set of SNPs, with SNPs shared between fresh samples and historical sample from 1989 and 1880, and drawn with microsatellites data. The red lines and labels are B. socotrana specimens, the green lines and labels are B. samhaensis specimens. (a) Cladogram generated with all loci from 1880 to 2018 (60,978 SNPs), (b) cladogram generated with number of loci limited by 1989 specimen (19,738 SNPs), (c) cladogram generated with loci limited to 1880 specimen (1107 SNPs), (d) neighbour joining tree of pairwise 1-D_{ps} distance of *B. socotrana* microsatellites data (Hughes et al., 2002c).

4.3.6 Demographic history estimation

After ROH analysis through the pipeline and filtration of identified ROH based on numbers of SNPs, the full-set of ROH were plotted and analysed to explore the limits of the method.

The lower-scoring *B. socotrana* at the bottom-left of the plot are from historical specimens, and we can observe that almost all of them display very low SROH and NROH (Fig. 4.17a). This can be explained by the low rate of sequences captured from the most ancient historical specimens (1880, 1897, 1989, 1994, and 1999), fewer SNPs called, and shorter and less numerous SNPs detected than for the silica-dried specimens from 2018. One historical specimen Socotrana_4_1989_herbarium does display higher numbers and longer lengths of ROH, but the lower number of SNPs supporting the detection of this ROH by Plink would suggest that the measure is not significant and would have been filtered out with the conventional settings of the pipeline.

Further investigations on the Socotran populations analysed by the pipeline reveal that the different individuals are scattered across the NROH SROH plot (Fig.4.17b). Several West Haggier specimens seem to cluster, the distribution of Reiged specimens is scattered throughout the plot.

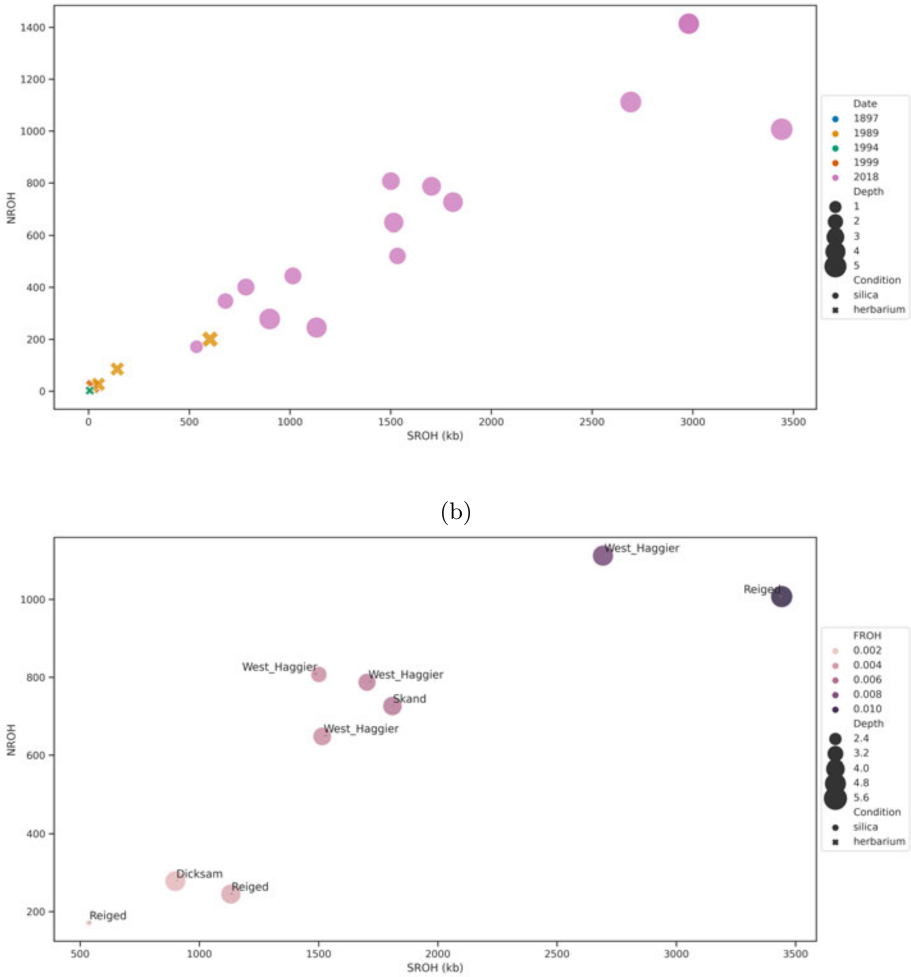


Figure 4.17: Distribution of ROH in the Socotran specimens, total length of ROH (SROH) versus total number of ROH (NROH). (a) Specimens coloured by date of collection of the Socotran *Begonia* specimens, (b) Population of *B. socotrana* in the dataset.

4.3.7 Phylogenetic analysis comparison with ROH estimation

To determine if the clusters of individuals with similar genetic structures are phylogenetically close, the Socotran specimens have been divided into three clusters: High F_{ROH} (purple colour), Medium F_{ROH} (brown colour), and low F_{ROH} (orange colour) (Fig. 4.19a).

These clusters have been used to colour-code the labels of a phylogenetic tree reconstructed with the SNPs of the 1,107 loci shared between present-day and historical specimens (Fig.4.19b). This set was chosen to avoid any bias between low-SNPs historical specimens and high number of SNPs silica-dried specimens. The most homozygous individuals in both *B. socotrana* and *B. samhaensis* are paired, suggesting they are products of the same bottle neck events, despite the two *B. socotrana* individuals being from two different collections: Reiged and West Haggier. However Reiged is an areas within the West Haggier mountains so the sites may have been very close (Fig.4.18).

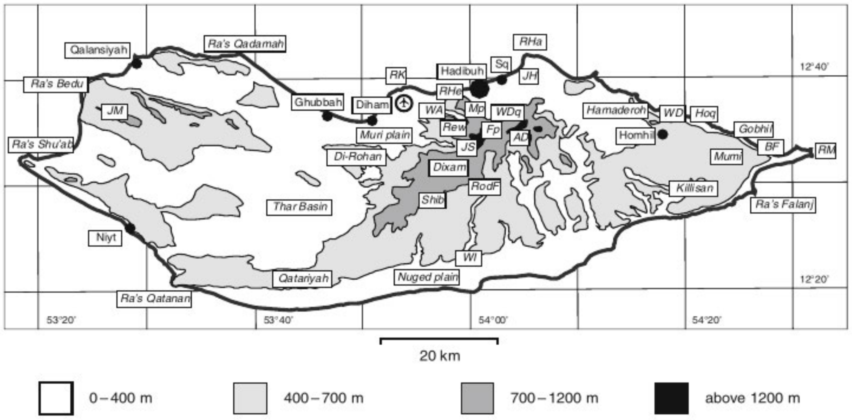


Figure 4.18: Sites of *B. socotrana* sampling on Socotra island (Brown et al., 2012).

4.4 Discussion

4.4.1 Variable capture efficiency and aDNA analysis

In this project we hoped to discover the genetic health of two endemic *Begonia* species and examine change through time. Unfortunately the captures did not achieve a full recovery of the targeted sequences, so the analysis we can do is limited. The reason for the failure of the capture is likely to be degradation of the baits. The first capture was done with a new kit, but although stored promptly on arrival we had to transport it across the city to perform the captures in the

University labs and it is possible that the baits degraded during that transport. Post-capture Bioanalyser traces looked good, but it is clear that this was due to PCR amplification of a subset of sequences, rather than the full range of targets. Differential degradation of baits could explain some of the patchy capture results. The patchy capture also reflects the presence of dinucleotide repeats in a small number of baits. These repeats make up microsatellites common in the genome of *B. socotrana*, but not in the genomes of other *Begonia* (Campos-Domınguez, 2022). Although these baits were not problematic in other captures (Michel et al., 2022), this emphasises the importance of bait design and avoidance of any repeat-like sequences or multi-gene families.

4.4.2 Ancient DNA patterns

We used duplicated samples of silica-dried and herbarium specimens as well as historical herbarium specimens to examine the effect of herbarium preparation on our ability to derive populations genetic parameters. The silica-dried specimens of the dataset shown very low patterns of fragmentation, where the historical specimens shown moderate patterns of fragmentation. Looking at the patterns of aDNA damage in the herbarium-dried and historical herbarium specimens, the same average level of C to T and A to G substitution was found in these specimens and silica-dried specimens from 2018. This lack of historical damage patterns was also found in the Papua New Guinean set of samples that included historical specimens in the same age range than the Socotran dataset (Chapter 5). The increase of substitution rate on the extremities of DNA fragment, usually found in Palaeobotanical remains, has been demonstrated in herbarium plant specimens from the same time frame than our Socotran dataset (Bieker et al., 2020). However, the library preparation enzyme used in our study Uracil-Specific Excision Reagent enzyme (USER) is known to prevent amplification of aDNA pattern of damages (Briggs et al., 2010, Rohland et al., 2015). The USER enzyme is a mix of Uracil DNA Glycosylase (UDG) and the DNA glycosylase-lyase Endonuclease VIII (Bitinaite et al., 2007). UDG treatment cleave deaminated cytosines (uracils) and cut the fragment, lowering the number of nucleotides substitutions observed, and making it more difficult to detect aDNA damage patterns (Rohland et al., 2015). We are planning to test our pipeline on datasets non-UDG treated to observe the signature of herbarium aDNA damages with our pipeline settings.

4.4.3 Variant call and phylogenetic reconstruction

The patchy read depth of sequencing of the Socotran dataset has affected our ability to access all the molecular markers targeted by the *Begonia* bait set. Nonetheless, the loci covered and filtered by our pipeline proved informative, correctly grouping species and allowing limited analysis of population genetic parameters.

Species delimitation

Three different phylogenetic trees were reconstructed, one with the full-set of SNPs (2018 tree), one reduced number of SNPs to match the historical specimens from 1989 (1989 tree), and one to match the historical and older specimen of 1880 (1880 tree). The trees all show the same clustering of *B. socotrana* and *B. samhaensis*. The only exception in all cases is the grouping of sample Samhaensis_1 with historical *B. socotrana*. This is possibly due to the high level of missing data in these samples, but further investigations on the exact loci covered by *B. samhaensis* specimens might explain this behaviour.

Population delimitation for *B. socotrana*

The 2018 phylogenetic tree shows clustering of specimens by population, the only exception being Socotrana_Reiged_3 which appears genetically distant from the other Reiged collections. This discrepancy was seen in all trees, it might be due to the lower number of reads recovered for this sample than for the other Reiged specimens. Once again, a closer look at the loci captured from this specimen will be informative.

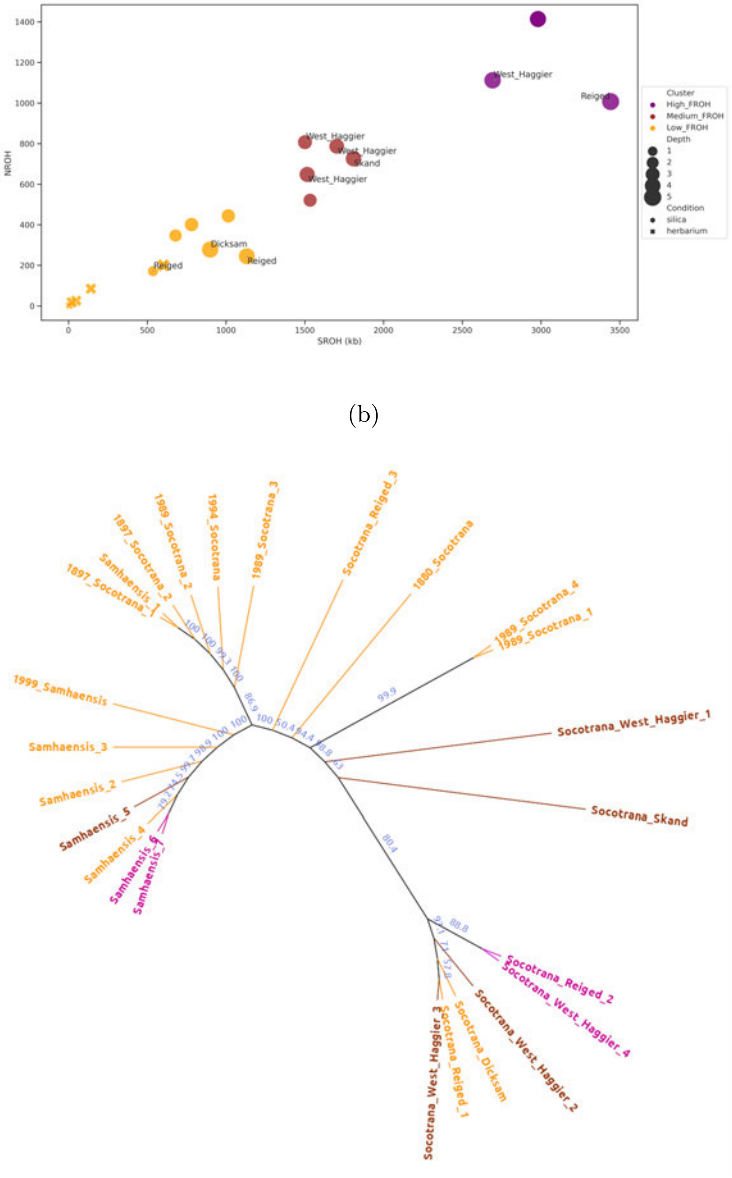
The genetic distance between populations varies depending on the tree considered. Skand and Dicksam populations, part of the Southern Haggier group, are closely related in the 2018 tree, a similar result to that in the microsatellite study (Hughes et al., 2002c). While their genetic distance increases when the number of included SNPs drops in the 1989 and 1880 tree, these two groups stay closely related.

Both the Reiged and West Haggier specimens were collected from the West Haggier and limestone plateaux, and actual distances between the collections are unknown. Even though the relationships of these specimens changes in the three trees, most of these specimens are grouped together in the 1989 tree.

We had hoped to identify the origin of the early collections, but the lack of data make a confident identification difficult. All the historical collections group together and the closest living collection sample is Reiged_3 in all three trees. However this may reflect only the low number of reads in this sample.

4.4.4 Further population genetics analysis

The homozygosity estimators we were using to asses the degree of endemism of the different accessions based on runs of homozygosity were not robustly estimated due to the incomplete dataset. The limit of estimating demographic history with ROH is the need for a relatively even read depth of sequencing, which is not the case in this dataset. However, other homozygosity estimators use overall genomic homozygosity, and can be assessed with this type of data. F_{SNP} for example is an inbreeding coefficient based on measures of inbreeding in the most recent generation, using observed and expected number of SNPs without relying on preserved homozygous segments. Other methods could be used to estimate relatedness between populations, as F-statistics, F_{ST}, or admixture analysis. However the poor overlap between samples in confidently called SNPs could also affect these measures.



(a)

(b)

Figure 4.19: Socotran specimens grouped by F_ROH score. (a) Scatter plot showing the number of ROH against total length of ROH for the specimens from the Socotran dataset. The specimens have been categorized in three clusters of high, medium, and low F_{ROH} represented respectively in purple, brown, and yellow, (b) Cladogram of the Socotran specimens, with lines and name label coloured with the same color code than the scatter plot.

Chapter 5

Begonia from Papua New Guinea

5.1 Introduction

This chapter uses the pipeline established in chapter 3 on data collected by a previous PhD student, Dr. Hannah Wilson. The dataset covers *Begonia* species from Papua New Guinea (PNG) and includes silica-dried and historical herbarium specimens. We expected to discover high levels of endemism in populations from the PNG highlands, and evidence of bottlenecks. Several of the samples are technical duplicates from poor quality extractions so we were able to verify the reliability of our results.

5.1.1 The Papua New Guinea biodiversity

Papua New Guinea is equatorial, located between the intertropical and South Pacific rainfall convergence zones. The climate is monsoonal, characterized by high temperature and humidity throughout the year, although some contrasts can be drawn between the north-west where monsoon occurs from December to April, and the South-East where monsoon occurs from May to October (Pereira et al., 2019). The island has a high level of biodiversity, being registered in eight of the nine globally recognised biodiversity conservation priority templates. This is the largest island in the region, the highest island, and displays a large array of different climates and geological landscapes, which form a vast panel of different ecological niches. The most species-rich ecosystem of these niches is the lowland rainforest, which is the wettest, least seasonal tropical biome, and with the highest levels of sunlight (Koenen et al., 2015, Kreft et al., 2008). Several models have been proposed to explain the high levels of biodiversity in the rainforests. They can be seen as stable ecosystems where biodiversity can gradually accumulate ('museum' model), or seen as dynamic ecosystems with high diversification and extinction rates ('recent cradle' model). A study on *Meliaceae* Juss. (Koenen et al., 2015) has shown that most of the Meliaceae species diversity in Papua New Guinea rainforest is recent, and is attributed to a higher rate of speciation for rainforest clades than for other clades. This observation, in conjunction with the previous two models 'museum' and 'recent cradle' has helped to suggest a new model where rainforests are dynamic ecosystems with high rates of speciation and extinction, and from which species radiates. The new model where PNG rainforests are source of species for the rest of the island might be more complex, as subsequent studies made on Neotropics Rainforests mention that 'cradle-like' and 'museum-like' speciation rate and extinction rates can be related to the clade studied, and change over time with climate conditions (Couvreur et al., 2011, Eiserhardt et al., 2017).

5.1.2 Previous work on PNG *Begonia*

The diversity of PNG sections of *Begonia* has been investigated by Moonlight et al., 2018, Hughes et al., 2015, and Wilson, 2021. This group is known for its high level of species diversity and endemism. Wilson's collections from fieldwork in 2018 included specimens from section *Petermannia*, section *Symbegonia*, and section *Oligandrae*, as well as samples attributed to section *Diploclinium* (Fig.5.1).

The data set we use here includes field collected samples, but in order to produce as full an account of PNG *Begonia* as possible Wilson also included in the capture set herbarium specimens from RBGE, Lyon Botanic Gardens and Harvard herbarium as well as from the RBGE living collection (Supplementary Table 2 from Wilson, 2021). Captures used the same bait set as described Chapter 3 and described in Michel et al., 2022.

The previous work undertaken by Hannah Wilson was to study the evolution and maintenance of tropical diversity using PNG *Begonia* as model to study species diversification and accumulation of biodiversity (Wilson, 2021). The study has included a phylogenetic investigation of New Guinea *Begonia* to explain their radiation over time, geographical locations, and different niches.

5.1.3 PNG *Begonia* clades

The findings of Dr. Wilson have resolved several taxonomical and phylogentic issues, here are her main findings.

Two clear clades were recovered with this set of samples: an early diverging New Guinea clade (EDNG) emerging 7.85 Million Years Ago (MYA) and originating from Borneo, and a 'large New Guinea clade' (LNG), with a crown age of 4.9 MYA, originating from the Philippines (Wilson, 2021). Species from both clades grow sympatrically (Wilson, personal observation).

Most of the species in EDNG are from section *Oligandrae*, section *Diploclinum*, one species of *Symbegonia*, and one species of section *Petermannia*. These are generally succulent lithophytic herbs with few stamens, and distinctive limestone niches and fragmented distributions (Hughes et al., 2015, Wilson et al., 2020).

A low genetic diversity would be expected from these species geographically restricted.

The species from sect.*Oligandrae* are locally common, but have more restricted distribution than in other sections. Two species: *B. pentandra* and *B. chambersiae* from Sect. *Oligandrae* show evidence of limestone endemism. The EDNG lineage might have been isolated from the ancestral species by the time Papua New Guinea central range was a series of ophiolite island emerging from the sea. It could explain its fragmented distribution and predilection for limestone. Section *Diploclinum* includes *B. kaniensis* which may be wind pollinated. We would expect to see high levels of genetic diversity in this species. The other PNG species of section *Diploclinum* are much more limited in distribution but are not included in this dataset. The high levels of morphological variation seen in some species in this group could reflect genetic drift in small populations and genetic vulnerability.

The more recently evolved LNG includes sections *Symbegonia* and *Petermannia*. It has a high diversification rate (0.33) and both widespread (many *Petermannia*) and endemic species (in *Symbegonia*).

5.1.4 Using target capture datasets with ROH pipeline

These investigations have used the same *Begonia* target capture baits than we have used in the case of the Socotran *Begonia* investigation (Chapter 4). We decided to use our Hyb-Capture analysis pipeline to generate demographic metrics from each sample, and compare them to the results previously found by the Dr. Wilson. We expected to see a diversity of genetic patterns, and hoped to discover how variable populations are across PNG (Fig.5.1).

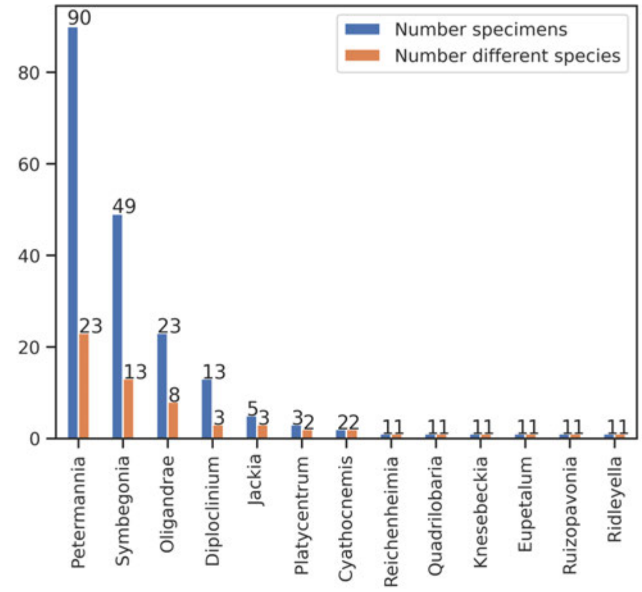


Figure 5.1: Sections and number of specimens in the Papua New Guinea dataset.

5.1.5 Subsequent analysis using F-statistics and heterozygosity rate

Further enquiries has been made to check the outputs of our pipeline after having shown the first results in the present document. We planned to measure the homozygosity level of the PNG specimens with other estimators to check the limitations associated with ROH measurements in the case of a target capture dataset. We demonstrated in Section 3.3.4 that maximal length of the genome captured with consecutive baits is 104.4Kb, and it is uncertain if the number and length of ROH detected by our tool would be representative of the genome targeted. To assess the representativeness of our results, we decided to compare our ROH estimators to other estimators related to inbreeding and population structure.

F-statistics are a commonly used population genetic tool, notably to estimate population differentiation with fixation index F_{ST}, and individual inbreeding coefficient F_{IS} (Weir et al., 1984). They have been used in conservation genomics to track inbreeding depression in natural plant populations (Chaves et al., 2011, Aravanopoulos et al., 2015, Edwards et al., 2021) and their use is documented for seed bank collection maintenance (Schoen et al., 2001). The calculation of these estimators does not match the constrains of our dataset as a minimum number n of individuals is required by population to establish allele frequencies and calculate these indices. Acknowledging this limitation, we state the hypothesis that F-statistics can be used on the PNG dataset, at species scale, and the results of these analysis will inform this hypothesis.

Another estimator used to check our results is the overall heterozygosity rate of each specimen. This estimator does not require several specimens from the same population for calculation, and has been used for conservation genomic purposes. Furthermore, overall heterozygosity has been used along with ROH estimation as a two-dimensional representation of genetic diversity for discriminating IUCN conservation categories (Genereux et al., 2020). We decided to use the same method than described in this publication to visualize the relationship between overall heterozygosity level and ROH index F_{ROH}. If our pipeline is working properly, we expect to see a negative relationship between these two indices, which would demonstrate that the number and length of the ROH detected matches the overall heterozygosity rate captured with the baits. F_{ROH} is dependant of the contiguity of the *Begonia* baits while overall heterozygosity is not, relying on variants captured by the baits independently of their position in the genome. Finding a relationship between these two indices would support that the ROH sizes and lengths captured by the baits are significant and that our method can estimate the relative inbreeding and demographic history of the specimens of a dataset.

5.2 Methods

5.2.1 Data generation (Hannah Wilson)

DNA was extracted with Qiagen kits and 191 library preps made using NebNext Ultra II. Capture was as described in Arbor Biosciences MyBaits kit version 4, with 16 hours hybridisation at 62°C and 9-22 post-capture PCR cycles. Sequencing was 150bp paired end Illumina HiSeq, and an average coverage of 105 read per bait per sample was achieved (Wilson, 2021, Michel et al., 2022). HybPiper was used to extract data for a gene tree analysis using ASTRAL (Wilson, 2021).

5.2.2 Hyb-Seq dataset analysis

The samples were processed via our pipeline and aligned to the reference genome of *B. peltatifolia* 'Begonia_peltatifolia_scaffold.fasta' (Li et al., 2022). The alignment was done with the Paleomix BAM pipeline (Schubert et al., 2014), using BWA (Li et al., 2009a), and the BAM files were recalibrated using MapDamage 2 (Jónsson et al., 2013). The G to A and C to T misincorporation rates have been measured at the end of sequencing reads with MapDamage2 (Jónsson et al., 2013). We used reference genome of *B. peltatifolia* as it was the most closely related to the PNG dataset species and providing the longest scaffolds or pseudo-chromosomes in order to proceed to subsequent Runs of Homozygosity measurements (Li et al., 2022).

Variants were called using GATK-4.2.0.0 with pipeline base parameters. The data was processed in three batches comprising sections *Oligandrae*, *Symbegonia*, and *Petermannia.* This subsampling was necessary as GATK tools gatk HaplotypeCaller and gatk GenomicsDBImport were not able to process the full set of samples (Brouard et al., 2019).

5.2.3 ROH filtering

For the filtering of ROH in the PNG dataset, the *l* threshold has been calculated with a value of 693 SNPs for the full dataset (Section 3.2.6).

The Papua New Guinea dataset included 191 accessions, of those, 131 samples were found to have ROH longer than 1kB, and only 73 of those had ROH above 1kb after minimal-SNPs filtering. These represent only 3.3% of all ROH in the initial dataset.

Table 5.1: Number of ROH in sections before and after SNP filtering

Sections	Samples	Species	Samples filtered	Species filtered
Petermannia	90	23	43	10
Symbegonia	49	13	17	8
Oligandrae	23	8	5	5
Diploclinium	13	3	7	1
Jackia	5	3	0	0
Platycentrum	3	2	0	0
Cyathocnemis	2	2	0	0
Eupetalum	1	1	0	0
Knesebeckia	1	1	0	0
Ruizopavonia	1	1	0	0
Ridleyella	1	1	0	0
Reichenheimia	1	1	0	0
Quadrilobaria	1	1	0	0

While a minimum of ROH length of hundreds of kb is usually required to consider an ROH, the use of baits reduce the length of ROH observed. As seen in Figure 5.2 the ROH size distribution for most of the Papua New Guinea dataset is below 10kB. There are 1820 filtered ROH in the full dataset, and only 78 of them are above 10kB. Therefore the ROH minimum length threshold has not been set up as 500Kb as presented in literature (Kumar et al., 2021, Ceballos et al., 2018), but to a 1 kb minimum ROH length.

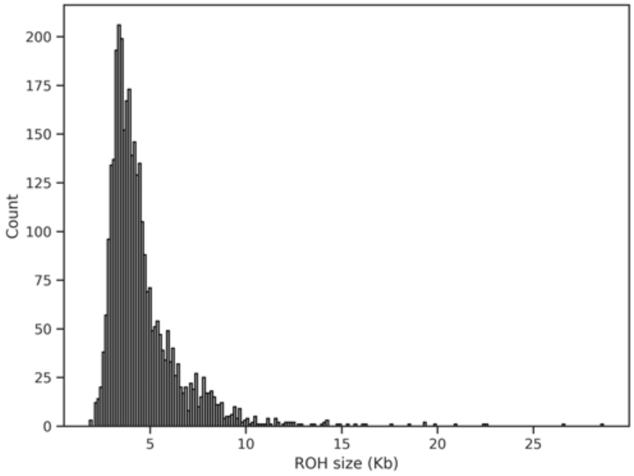


Figure 5.2: Distribution of ROH by size.

5.2.4 Inbreeding estimators

Several estimators have been derived from ROH statistics to assess the demographic history of each accessions of the dataset. Among them, the sum of ROH (SROH) is the total genomic length covered by ROH and Number of ROH (NROH) is the total number of ROH per individual (Kumar et al., 2021). The relationship between these two estimators allows the estimation of the recent demographic history of a population, and disentangling demographic events such as bottlenecks and high consanguinity, as both increase the overall genomic homozygosity. Another estimate related to the inbreeding coefficient (F) has been considered: F_{ROH}, which has been described as the fraction of each genome in ROH longer than 0.5Mb (Kumar et al., 2021, Ceballos et al., 2018).

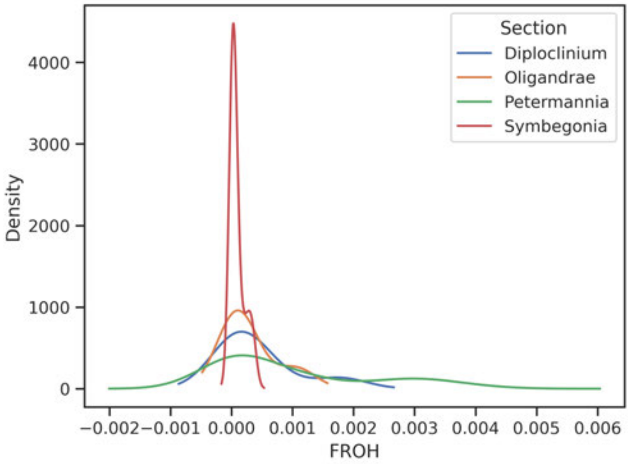


Figure 5.3: Inbreeding coefficient density in Papua New Guinean sections.

5.2.5 Phylogenetic tree

To check if the data filtered through the pipeline were phylogenetically comparable to the previous results of Hannah Wilson on the same dataset (Wilson, 2021), a cladogram has been produced with the output of the pipeline. To reconstruct specimens phylogeny, aligning the reads to a larger genome as *B. peltatifolia* was not necessary as calling SNPs did not require a contiguous scaffold to be used. Therefore, the raw reads of the Papua New Guinea Hyb-Seq dataset were aligned using BWA (Li et al., 2010) to the *B. bipinnatifida* sequences orthologous to the bait targets. Variants were processed through the pipeline to make a joint variant call, and through similar filtering steps described for the ROH measurements. To make the species tree, the vcf2phylip script (Ortiz, 2019) has been used to generate a .phy file, and then IQ-TREE Nicholls et al., 2015 to build the phylogenic tree. To enable comparison with the phylogeny shown in Hannah Wilson work (Wilson, 2021), similar substitution model has been selected to build the tree: GTR+G+I, with General Time Reversible model (GTR), gamma option to rescale the branch length (G), and an option for including proportion of invariable sites (I).

5.2.6 Biogeography

The specimens collected by Hannah Wilson and Mark Hughes are scattered all across the central mountain range of the New Guinea Island. To understand if the patterns of homozygosity observed arise not only from phylogenic history but as well from geographic background, correlation analysis has been done between altitude and F_{ROH} for the specimens displaying detectable ROH. A population-scale observation has been included as well to understand if geographic features as rivers of cliffs play a role in the homozygosity detected. The GPS coordinates of the different specimens were extracted from field records for the silica-dried specimens and herbarium sheets records for historical specimens or calculated using a Digital Elevation Model (DEM) map with the specimen coordinates. Subsequently, the relation between altitude and inbreeding coefficient has been explored plotting these parameters (Fig.5.14a). To check if a relationship could be established between altitude and inbreeding coefficient F_{ROH} the whole set of specimens has been subjected to a Pearson, Kendall, and Spearman correlation measures. The Person correlation index (r) assumes that the data is normally distributed, which is apparent on the density map (Fig.5.14b). A p-value can be calculated for this test, but the scarcity of data points has provided low scores as a dataset of 500 or more individuals is required for a robust analysis (Bujang et al., 2016). The Spearman correlation coefficient (ρ) and Kendall correlation coefficient (τ) have been calculated for rank correlation analysis, which does not require the data distribution to be linear.

5.2.7 Subsequent analysis using F-statistics and heterozygosity rate

To confirm findings made with ROH estimators, F-statistics estimators including fixation index F_{ST} and inbreeding coefficient F_{IS} were used to check our results. The population structure of the PNG specimens has been calculated using the protocol described in Bieker et al., 2022 using Angsd 0.910 (Korneliussen et al., 2014) and PCAngsd 1.10 (Meisner et al., 2018). Considering that in the majority of cases only one individual was available by population, different groupings of the specimens have been considered for calculations of the F-statistics estimators.

For F_{ST} calculation, specimens have been grouped by species. The minimal number of specimens per population to estimate genetic differentiation has been estimated as small as n=4-6 for a large number of genetic markers (k>1,000) (Willing et al., 2012). We selected a value of n=5 specimens minimum by species to include them in our study. Angsd has been used to calculate the Sample Allele Frequency (SAF), Site Frequency Spectrum (SFS), and 2D-SFS for each individual, and the F_{ST} index between each species.

For F_{IS} calculation, the specimen have been grouped using PCAngsd to produce a covariance matrix and estimate the number of population clusters K=50. Considering these clusters as populations, Angsd has been used to calculate genotype likelihood and F_{IS} for each specimen.

The rate of heterozygous sites on the total length of the sequences captured has been calculated for all the specimens using home-made scripts based on the Variants Call Format (VCF) file output of our pipeline. The comparison between heterozygosity rate and F_{ROH} has been run on the present set of PNG *Begonia* and on a set of Asian *Begonia* belonging to the section *Coelocentrum* not described in detail in this study. Although the dataset of *Begonia* section *Coelocentrum* have not been detailed, the results of their analysis through our pipeline have been added in Figure C.5b as a proof of concept of our pipeline.

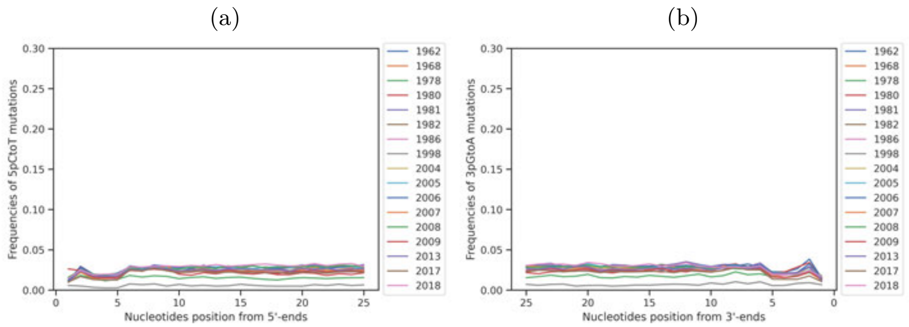
5.3 Results

5.3.1 aDNA damages

The damages observed in figure 5.4 are lower than those expected for aDNA from herbaria.

The overall average number of aDNA substitutions observed is 0.020, which is comparable to the rate of other studies on herbarium specimens (Staats et al., 2011, Bieker et al., 2020), however the exponential increase of substitutions observed at the extremities of the fragments is not present here.

A likely explanation for this is that the NebNext Ultra II kit uses Uracil hairpin adaptors followed by USER treatment, containing Uracil DNA Glycosylase (UDG), to break the hairpin into two adaptors (NEB, 2022). This will also remove Uracils from the rest of the molecule, and cut the fragment, so no substitution could be observed in subsequent analysis steps (Rohland et al., 2015).


(a)

(b)

Figure 5.4: Nucleotides substitutions in 5' and 3' of DNA fragments. (a) 5' substitutions, (b) 3' substitution.

5.3.2 Target capture efficiency

The average depth of sequencing coverage has been calculated for each baits and each samples in the set to check the homogeneity of target capture. The maximum depth of sequencing coverage per bait for the Papua New Guinea set is 366,827, with an average of 226 per bait. These are scores comparable with the Socotran set that shown a maximum depth of sequencing coverage of 243,289, with an average of 466 per bait. Some 'bad baits' are still visible on the heat map, represented by white squares (Fig.5.5). However, they are more 'bad baits' observable in the Socotrana set. This can be explained by the behaviour of several baits to capture microsatellites in high number in the *B. socotrana* genome (Chapter 4). Furthermore, the depth of coverage is more evenly distributed per sample in the Papua New Guinea set than in the Socotran set. This could reflect the poor target capture efficiency observed on the Socotran set (Chapter 4) compared to the high sequences recovery rate that has been effected on the Papua New Guinea samples (Wilson, 2021).

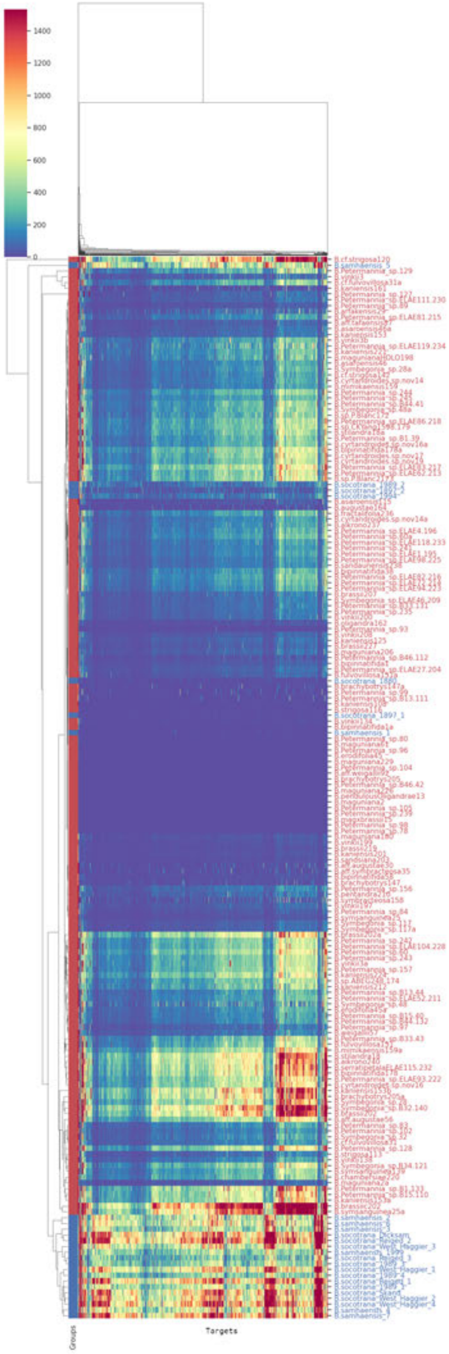


Figure 5.5: Cluster map of the average depth of sequencing coverage per baits between PNG and Socotran target capture sets. The PNG samples are labelled in red and the Socotran samples in blue.

5.3.3 Phylogenetic reconstruction

To assess the congruence of our pipeline output and previous phylogenetic reconstruction of the PNG batch of specimens (Wilson, 2021), a maximum-likelihood phylogenic cladogram was produced using the SNPs called by our pipeline (Fig.5.6). The cladogram shows the expected clustering of the accessions by section, and the same split in *Petermannia* seen in the ASTRAL tree generated by Wilson, 2021 (Fig.5.7). The specimen B.Petermannia_sp.ELAE119.234 is noticeable for having a long branch compared to all other specimens. It is a specimen collected at Lae Botanic garden, and described as New Guinea specimen. Hannah Wilson noted that in her analysis this high-quality sample grouped with South American clade and is present in 710 of the 713 gene trees generated. It is therefore a misidentified specimen of unknown origin, but seems to be closer to South American clades than New Guinean ones (Wilson, 2021).

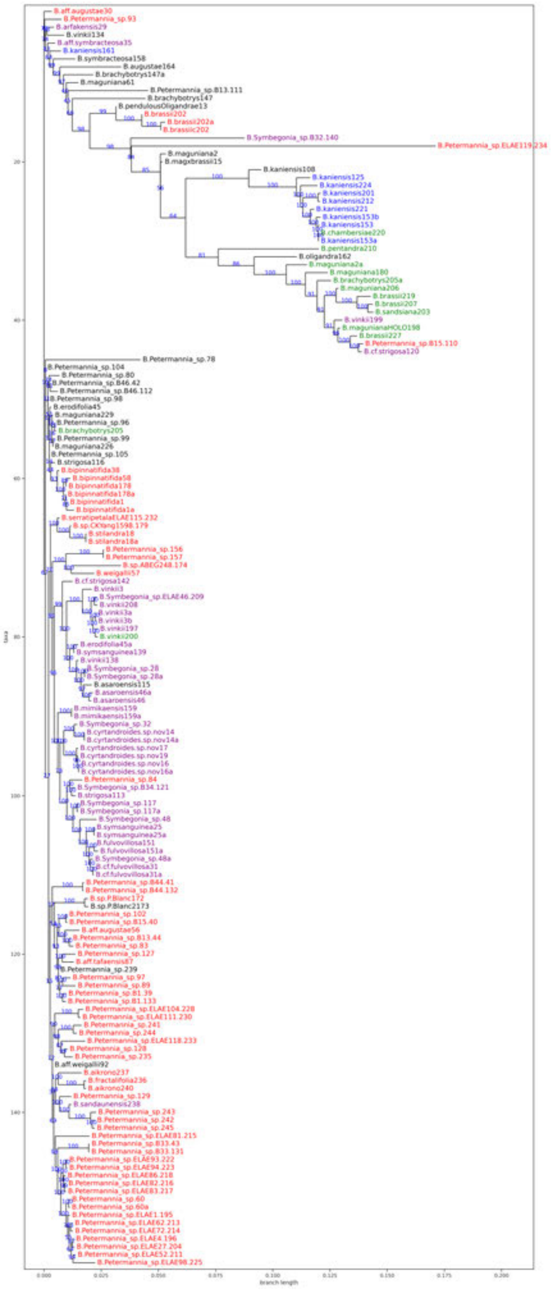


Figure 5.6: Phylogenetic reconstruction of the Papua New Guinea full dataset. Tip labels coloured by sectional placement,: *Petermannia* in red, *Symbegonia* in purple, *Oligandrae* in green, *Diploclinium* in blue, and other outgroup sections in black.

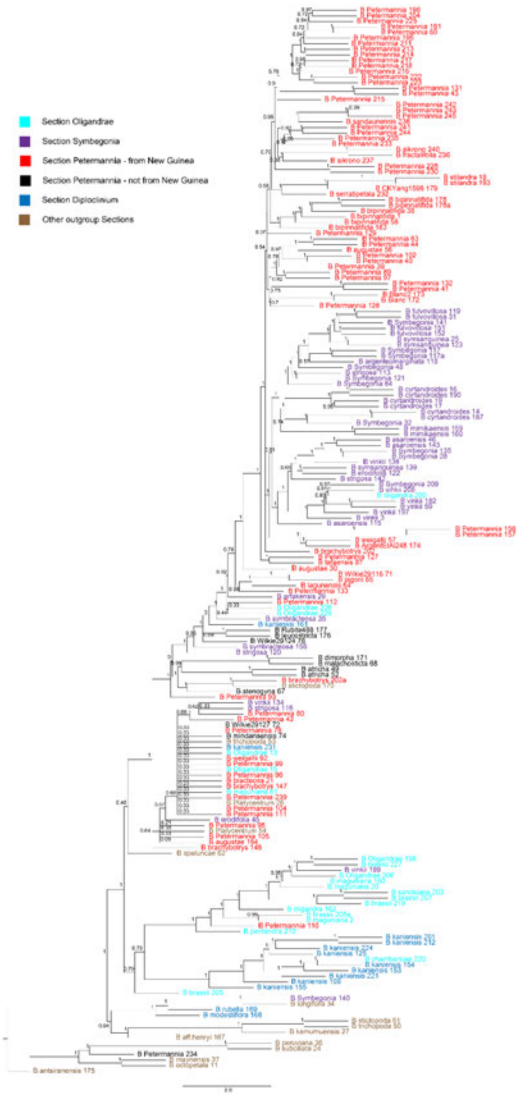

🟦 Section Oligandrae
🟪 Section Symbegonia
🟥 Section Petermannia - from New Guinea
⬛ Section Petermannia - not from New Guinea
🟦 Section Diploclinium
🟫 Other outgroup Sections

Figure 5.7: Complete ASTRAL phylogeny of the Papua New Guinea dataset from Appendix 2 of Wilson, 2021.

5.3.4 Demographic history

The estimators F_{ROH}, NROH, and SROH have been plotted for all accessions after ROH filtering on minimum SNPs per ROH and ROH size above 1kB (Fig.5.9a). To check the method, average sequencing depth and the type of material (historical herbarium material, or silica-dried fresh material) have been plotted along with the ROH estimators (Fig.5.9).

The F_{ROH} estimator follows the trend of the NROH/SROH table estimators, though it is less informative. Specimens with higher F_{ROH} are predicted to be from populations which have undergone bottlenecks or have high levels of consanguinity, and those with lower scores are predicted to be from large populations with possible introgression promoting high heterozygosity (Fig.5.9a, Wilson, 2021).

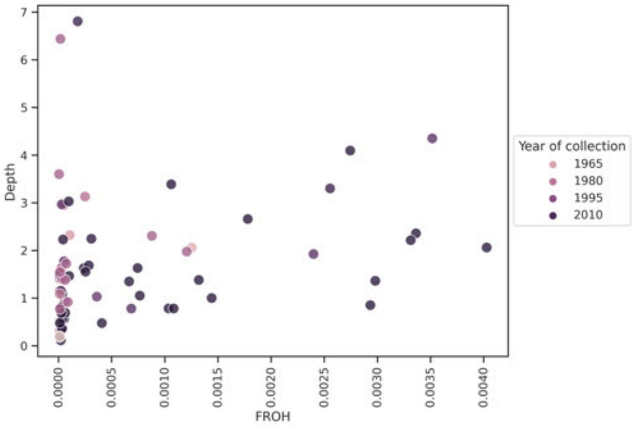
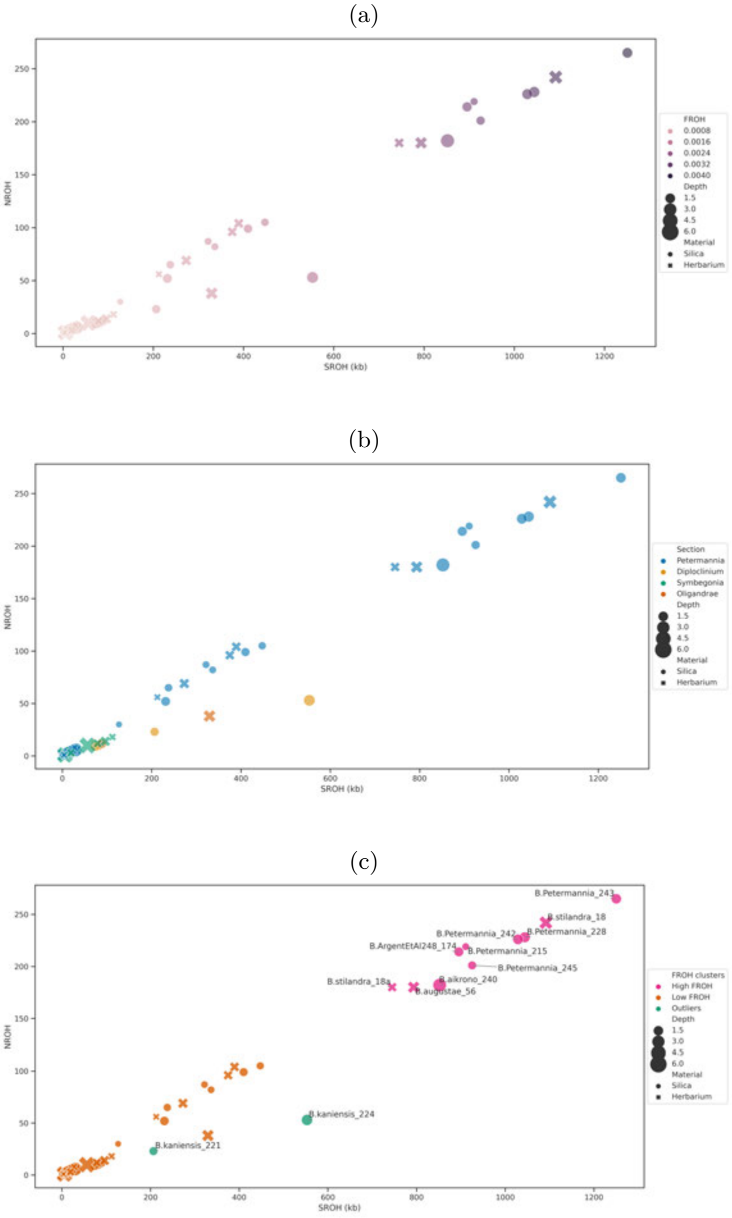


Figure 5.8: Relationship between depth of coverage, F_{ROH}, and date of sampling of the specimen.


(a)

(b)

(c)

Figure 5.9: Demographic history estimation for the whole set of samples. (a) with F_{ROH}, (b) with all Sections. (a) Plot SROH, NROH, and F_{ROH} for all accessions, (b) Plot SROH and NROH for all accessions with sections, (c) Plot SROH and NROH clusters.

5.3.5 Estimating ROH in historical specimens

The plot A.1 (Appendix A) shows more specifically the relationship between depth of coverage, F_{ROH}, and date of sampling of the specimen. Both old historical specimens and recent silica-dried specimens are found in the low-F_{ROH} cluster. Several historical specimens show a high-F_{ROH} with very low coverage. The two replicates of B.stilandra_18 and B.augustae_56 are present in the high-F_{ROH} cluster while showing low depth of coverage 5.9. Historical specimens present in the low-F_{ROH} cluster show various levels of depth of coverage, similar or far above the high-F_{ROH} coverage. Age of the specimen seems not to be a limitation to detect ROH in a genome, even though further studies will need be to done to establish clear limitations of depth of coverage in ROH detection.

5.3.6 Technical replicates

The specimens collected in 2018 during the expedition of Hannah Wilson and Mark Hughes included duplicated accessions in cases where the initial extraction produced a poor library (Wilson, 2021). These can be used for checking the repeatability of the pipeline. Considering a batch of 23 replicates: 21 specimens are duplicates, and 2 are triplicates (Table 5.2). After measurement of the ROH and filtration of the low-SNPs ROH, 4 sets of replicates display ROH in the same cluster of F_{ROH} and in 11 no reliable ROH are detected. However, 8 replicated specimens show ROH where their duplicates do not show any. In the high-F_{ROH} cluster, the *B. stilandra_18* duplicates show ROH with a variability of 60 ROH. For the low-F_{ROH} cluster, the average difference between duplicates or triplicates is 6.9 ROH. The variation in the number of ROH detected in the low-F_{ROH} cluster can affect our ability to quantify accurately ROH in this cluster. Noticeably, 38 ROH have been detected in B. brassi_202, and none in its duplicate, as most of them have been removed with the minimum-SNPs threshold of the pipeline.

The filtering of ROH involved in our protocol has excluded several specimens from the analysis. Among them the specimen B.asaroensis_46a, replicate of B.asaroensis_46 which has been rejected from the analysis while this last specimen is still in the dataset with 3 ROH detected. More surprising is the case of B.brassii_2002a displaying a higher number of 38 ROH remarkable for their length.

The scatter plot 5.10 shows the replicates along with the other specimens of same species. The *B. Petermannia* group of unidentified species is shown to give an indication of levels seen for of low-F_{ROH} and high-F_{ROH} groups. *B. vinkii* and *B. fulvovillosa* replicates are both in the low-F_{ROH} group, and display similar values that are within the interquartile range of the species distribution. A two-tailed significance test on F_{ROH} values has found all replicates to be not statistically different, while other samples of the same species were found significantly not similar to B.vinkii_208.

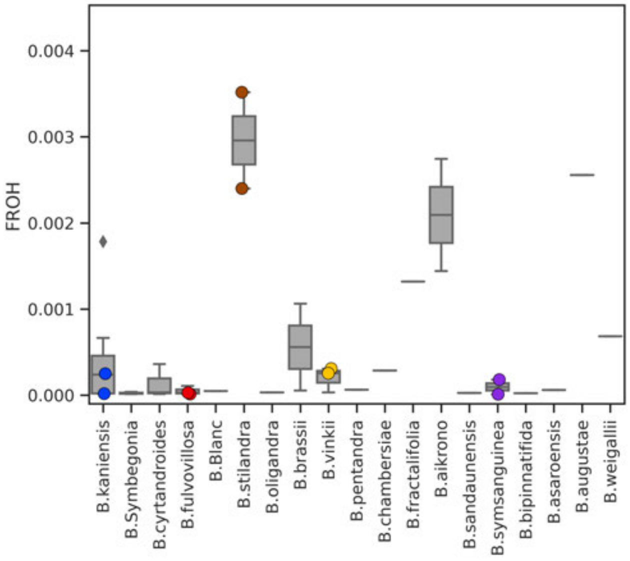


Figure 5.10: Technical replicates in the Papua New Guinea Dataset.

5.3.7 Population parameters in PNG Begonia

We can distinguish two clusters for the Papua New Guinea dataset: one on the upper part of the plot with high-F_{ROH} , and a more numerous group with low-F_{ROH} in the lower part of the plot (Fig.5.9c). No clear correlation can be seen with sequencing depth or with type of material. The specimens B.stilandra_18 and B.bracssii_202 group with the to the high-F_{ROH} group, with fewer depth in their dataset than the whole dataset average depth.

The specimen B.kaniensis_224 is part of the outliers cluster (green colour) and is remarkable for having longer ROH (high SROH), but less than other specimens (low NROH). This particular genomic situation can be interpreted as signature of inbreeding, while B.Petermannia_228 has the signature of a bottleneck event, and other specimens B.Petermannia_228 and B.Petermannia_230 have low number and size of ROH typical from larger or admixed populations.

5.3.8 Interspecies homozygosity rates

One of the main questions we wished to answer about the different lineages within the Papua New Guinea dataset is whether the homozygosity patterns observed here are due to genetic drift driven by microevolution at the population scale, whether they are related to phylogenetically stable aspects of the clade's biology. For this purpose, we want to see if individual specimen's values of F_{ROH} differed from the those of the rest of the species. F_{ROH} values for within species specimens were plotted and tested with a statistical two-tailed test . The

Table 5.2: Number of ROH detected in replicates.

Specimen	ROH_Rep1	ROH_Rep1 coverage	ROH_Rep2	ROH_Rep2 coverage	ROH_Rep3	ROH_Rep3 coverage
B.stilandra_18	242	4.3	180	1.9	-	-
B.fulvovillosa_151	6	2.3	0	0.3	-	-
B.Petermannia_44	4	1.4	0	0	-	-
B.asaroensis_46	3	1.3	0	0.5	-	-
B.cyrtandroides_16	2	2.9	0	1.6	-	-
B.kaniensis_153	1	0.4	11	3.1	0	4.7
B.symsanguinea_25	1	0.1	10	6.8	-	-
B.fulvovillosa_31	1	0.7	2	1.6	-	-
B.brassi_202	0	5.1	38	8.4	-	-
B.cyrtandroides_14	0	0.9	18	1	-	-
B.vinkii_3	0	2.7	14	2.2	12	1.55
B.Symbegonia_48	0	1.3	1	1.5	-	-
B.bipinnatifida_1	0	0.2	0	0.03	-	-
B.Symbegonia_28	0	4.5	0	0.9	-	-
B.Petermannia_60	0	2.5	0	1.1	-	-
B.Symbegonia_117	0	0.3	0	0.3	-	-
B.brachybotrys_147	0	0.3	0	0.3	-	-
B.mimikaensis_159	0	0.9	0	2.9	-	-
B.bipinnatifida_178	0	3.4	0	4.3	-	-
B.brassii_205	0	4.3	0	4.3	-	-
B.maguniana_2	0	0	0	0.4	-	-
B.erodifolia_45	0	0	0	1.3	-	-

p-value test the hypothesis that a specimen's F_{ROH} value is significantly different than the rest of the species (H1), or that the F_{ROH} value of a specimen is not distinct from the average of the population (H0). Within the *B. Petermannia* group specimens are unidentified to species, and therefore the scattered values observed are not a test of inter-specific variation. Almost all specimens were found not significantly different than the rest of the species. The only specimen significantly different from the rest of the species is B.vinkii_208. The value of this result is moderated by the fact that the two other samples from same species examined are replicates from the same individual with very similar scores. Across all our tests the number of specimens are usually two or three, and might be too few to draw strong conclusions. However, members of a single species are always part of the same clusters previously observed (low-F_{ROH}, high-F_{ROH}), with the exception of *B. aikrono* which displaying intermediate values.

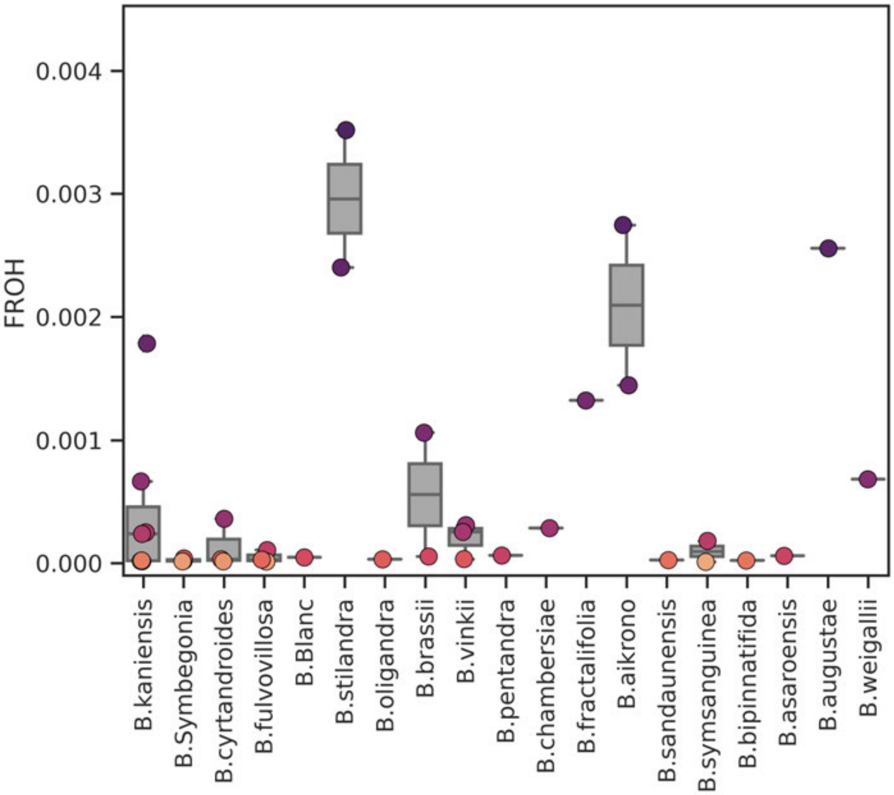


Figure 5.11: F_{ROH} homozygosity estimator for all samples per species.

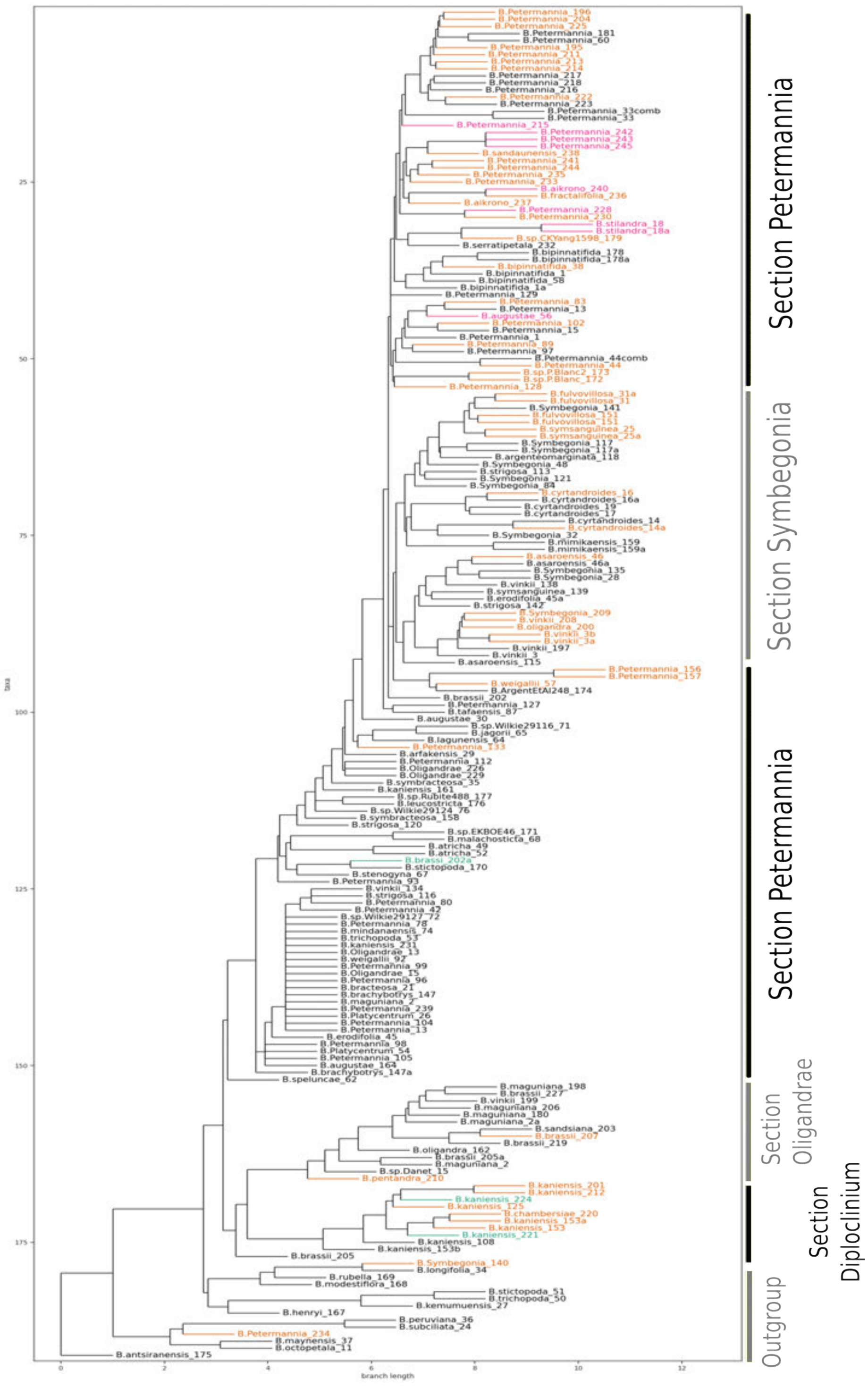


Figure 5.12: ASTRAL PNG phylogeny with F_{ROH} estimator

5.3.9 Biogeography

The relationship between homozygosity level and phylogenetic genetic distance having been considered, we decided to check if the geographical locations of the specimens would be related to their patterns of ROH. The specimens have been mapped to their location of the central mountain range of the New Guinea Island (5.13), and the relationship between altitude of sampling and F_{ROH} estimator has been calculated.

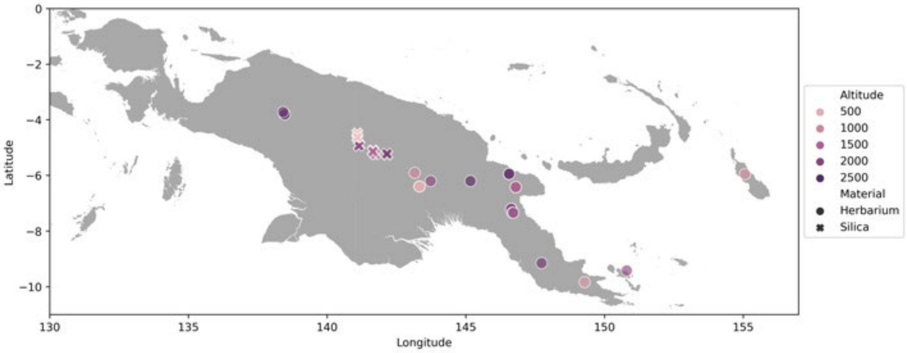
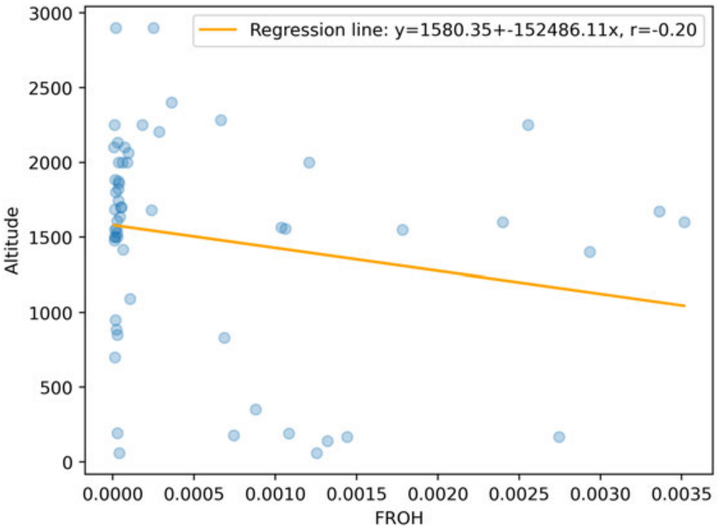


Figure 5.13: Specimens distribution on New Guinea Highlands, including the herbaria specimens and silica-dried specimens collected during the Wilson and Hughes expedition of 2018 (Wilson, 2021). The distribution follows the central range of the New Guinea Highlands.

Plotting altitude in relation to F_{ROH} show a negative relationship between these parameters (Fig. 5.14a). To check the strength of this correlation, we calculated linear correlation and rank correlation coefficients. A preliminary analysis of the density estimation of the specimens altitude and F_{ROH} score show normal distribution of these parameters (5.14b), meeting the conditions required to calculate a Pearson correlation coefficient. Rank correlation coefficients of Spearman and Kendall have been considered as well. The results show a slightly negative correlation (r=-0.197) between altitude and the F_{ROH} coefficient (Table 5.3). The p-value indicate that this statistic is not significant (p-value > 0.1). The Spearman and Kendall correlation coefficients show a negative correlation as well, even though the p-values indicate a weak significance as well. As our results failed to find a clear relation between altitude and patterns of homozygosity detected on a global scale, we decided to visualize the F_{ROH} values at population scale to see if the geographic context can explain the results of our pipeline.

(a)



(b)

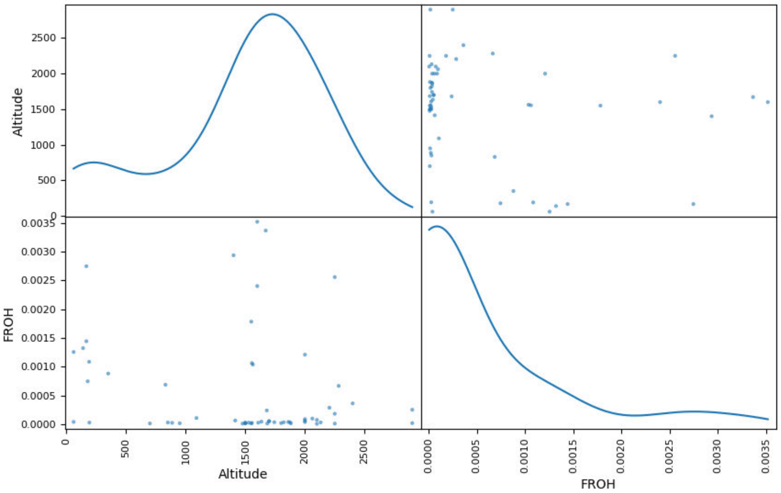


Figure 5.14: Correlation between the altitude on which each specimen has been collected and the inbreeding coefficient F_{ROH} for specimens with detectable ROH. (a) Scatter plot matrix with trend line, (b) Kernel density estimation of the specimens altitude and F_{ROH} scores.

Table 5.3: Correlation coefficients between altitude and F_{ROH}.

Coefficient	Value	P-value
Pearson (r)	-0.197	0.137
Spearman (ρ)	-0.123	0.357
Kendall (τ)	-0.065	0.468

We have looked for the patterns of homozygosity of the specimens in five areas located at three locations of Papua New Guinea.

- In the North-Wester area, the populations of Busilmin and Kwima.
- In the centre-Western area, the populations of Telefomin and Tekin.
- In the Eastern area, the populations of Teptep.

In the maps provided, the specimens have been coloured accordingly to their belonging to a demographic history cluster (Fig. 5.9).

The specimens scattered around Busilmin have all been sampled in montane forest, at altitude ranging from 1,550 to 1,882 meters high (Fig. 5.15a). They are all silicate-dried specimens collected in 2018, and do not differ in the degradation of the DNA material (Wilson, 2021). B.Petermannia_228 collected 'in open scrubby forest, by steam side' is in the high-F_{ROH} cluster (pink colour), displaying a high level of homozygosity compared to the rest of the *Petermannia* section specimens, and twice as high as the *B. kaniensis* specimen located further North in the same area (Fig.5.9) .

The low-F_{ROH} specimens located in the same area are B.Petermannia_230 sampled in montane forest, noted as 'male on thick moss by stream, growing with *Begonia* section *Oligandrae*', B.kaniensis_224 collected on mossy rock by stream, and B.Petermannia_225 sampled in montane forest.

No apparent ecological reason here can explain the high homozygosity level of B.Petermannia_228 compared to the other specimens. The particularity of B.kaniensis_224 as an outlier in our set is as well not apparent related to its sampling location.

All specimens from Kwima were sampled on the North bank of the August river, collected in 2018 and silica-dried, at altitude ranged between 139m and 167m (Fig.5.15b) (Wilson, 2021). All specimens have relatively high F_{ROH} compared to the rest of the dataset, however B.aikrono_240 is the only one which is part of the higher F_{ROH} cluster. This specimen is interesting to compare to the B.aikrono_237 specimen, from the same species, location and altitude of sampling. It is unclear if they are from the same population, but are expected to come from populations closely related. Furthermore, all these specimens are phylogenetically closely related (Fig.5.12). These specimens might come from closely related populations with an intermediate average level of homozygosity, where higher ROH are found in several more isolated groups or individuals.

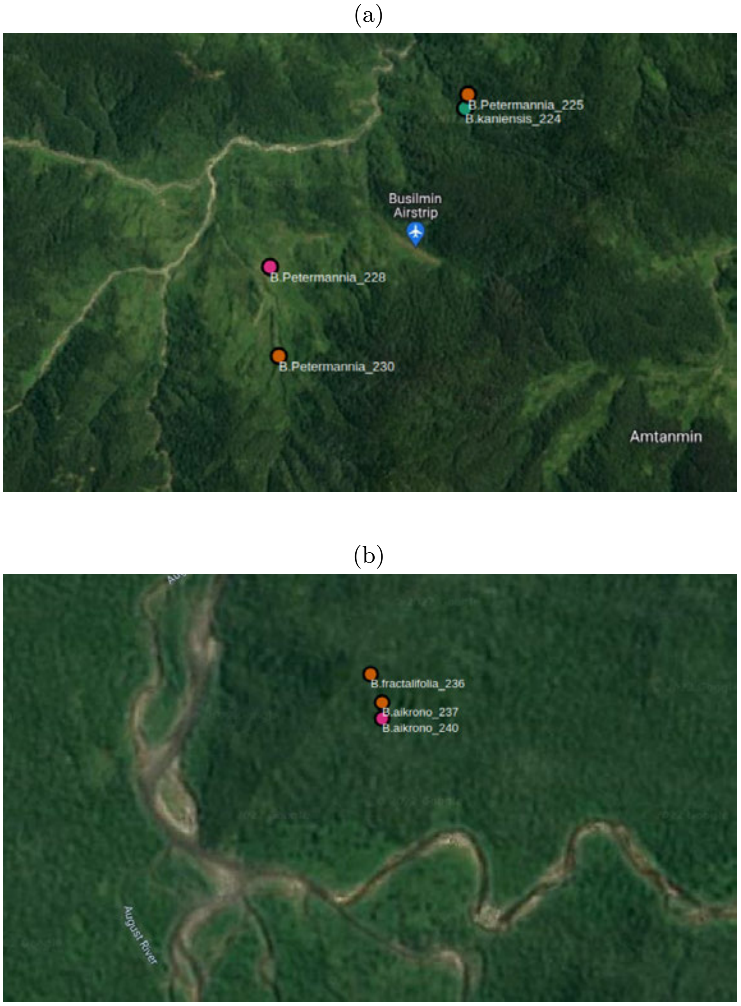


Figure 5.15: Distribution of the specimens in two North-Western areas. (a) Busilmin, (b) Kwima.

The specimens from the Telefomin populations come from the Sandaun province, on a mountain chain located close to the Telefomin station town (Fig.5.16a). They have been collected in 2018 and silica-dried (Wilson, 2021). The collection places were located at high altitude ranging from 1,401m to 1,555m, along river banks or rainforests floor, and limestone rocks. B.Petermannia_215 was collected along a river bank in gulley and is the only specimen part of the high-F_{ROH} cluster (Fig. 5.9c), and is part of the group of *B. Petermannia* specimens with lower ROH but taxonomically close (Fig. 5.12). This level of high homozygosity could therefore be indicative of recombination in a population displaying already an intermediate level of homozygosity by isolation or stochastic events. B.brassi_202a is part of the the outliers cluster (Fig. 5.9c), with longer ROH than expected indicative of possible inbreeding. Taking in account that B.brassi_202a replicate, a sample with similar coverage as well (Table 5.2) has been rejected by the pipeline, the significance of this sample is questionable.

The specimens collected in the heights nearby Tekin were collected between 2,062m and 2,205m in montane forest (Fig.5.16b , Wilson, 2021). In this section, B.kaniensis_221 and B.chambersiae_220 are closely related taxonomically and have been both collected in montane forest though there is no direct explanation for B.kaniensis_221 being part of the cluster of outlier, the specimens having same ecological location, and are closely related.

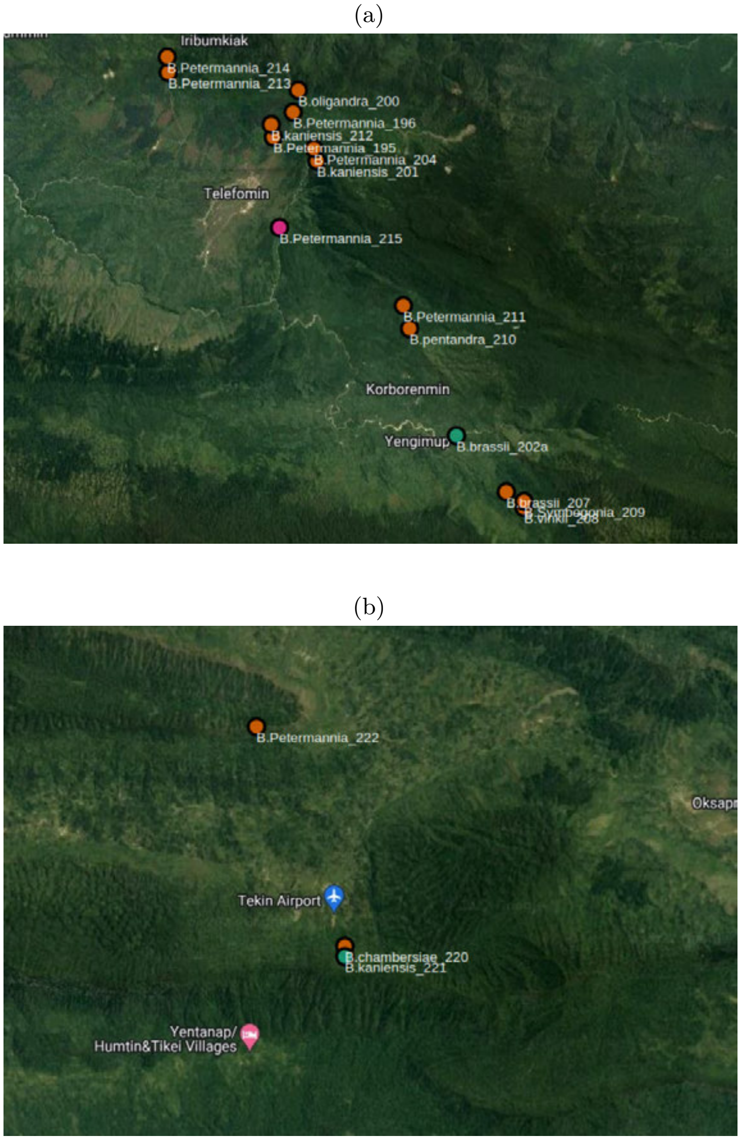


Figure 5.16: Distribution of the specimens in two of the centre-Western areas. (a) Telefomin, (b) Tekin.

The specimens collected at the north of Teptep are historical herbarium specimens coming from a location at estimated altitude of 2,577m (Fig.5.17, Wilson, 2021). The two samples of *B. Symsanguinea* are replicates, and B.augustae_56 is not closely related, being part of the section *Petermannia*, and *B. symsanguinea* part of the section *Symbegonia*. In this case, the impact of taxon-related homozygosity and topological or ecological background is difficult to disentangle to explain B.augustae_56 high level of homozygosity. We can still notice that the closest relative of this specimen are B.stilandra_18 or B.Petermannia_228, which are in the high-F_{ROH} cluster as well, and that the evolutionary history of B.augustae_56 might be the key factor explaining high-homozygosity with geographical isolation contributing to this situation.

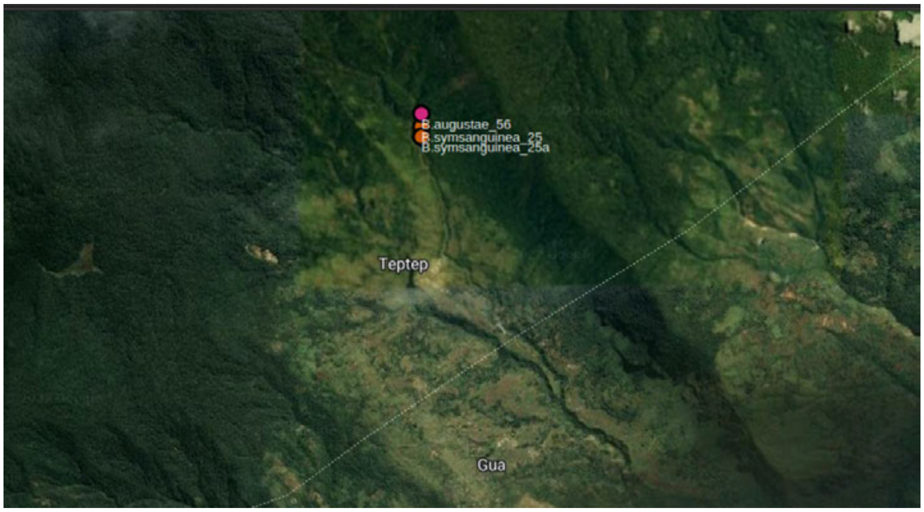


Figure 5.17: Distribution of the specimens in a Eastern area of Teptep.

5.3.10 Subsequent analysis using F-statistics and heterozygosity rate

As our pipeline detect a limited number of loci, and cover a fragments of the targeted genomes, other estimators have been used to check our results after the initial findings of this project. F-statistics have been calculated for the PNG specimens to compare populations structures to the patterns of homozygosity found with ROH, as they are not limited by the acquisition of long segments of the target genome. However, the calculation of these indices rely on alleles diversity found within populations, requiring several individual of the same population being analysed, which is not the case on this dataset.

Keeping this limitation in mind, F_{ST} values have been calculated, grouping the specimens by species, and considering each species as a population, and the whole batch of specimens as a metapopulation.

The Figure C.1 describes the F_{ST} values between species of the PNG specimens. Comparing these results to the phylogenetic tree established in Hannah Wilson studies (Fig. 5.7, Wilson, 2021) and to the ROH measurements made previously (Fig. 5.12), no relationship could be established between the F_{ST} values found and phylogenetic relationship between the species or patterns of homozygosity.

The comparison between individual F_{IS} and F_{ROH} values previously found do not show a relationship between the inbreeding coefficient and the ROH coefficient (Fig.C.2). Similarly, a phylogenetic reconstruction of the PNG dataset annotated with F_{ROH} and F_{IS} (Fig.C.3) does not show similar patterns of high-scoring indices. There is no clear overlap either between specimens showing low F_{ROH} and low F_{IS}.

Comparison of heterozygosity ratio and F_{ROH} provides a weak correlation between these two indices for the PNG specimens (Fig.C.5a, Pearson r=0.068, p-value=0.556, Spearman ρ=0.011, p-value=0.923). However, processing a similar dataset of Asian *Begonia* to the pipeline to produce the same comparison show a strong correlation between the two indices (Fig.C.5b, Pearson r=-0.910, p-value=3.280, Spearman ρ=-0.780, p-value=2.574).

5.4 Discussion

5.4.1 Genetic health in Papua New Guinea Begonia

We have successfully estimated the homozygosity level of samples from historical herbarium and silica-dried specimens. The condition of specimens (historical or silica-dried) or depth of sequencing seems not to be related to the level of homozygosity found in the results (Fig. 5.9). Several of the oldest sample as B.Petermannia_157 from 1963 and B.Petermannia_140 from 1987 are in the high-F_{ROH} cluster, and the *Begonia* bait set and our pipeline has captured and detected long ROH compared to the rest of the dataset.

While the EDNG clade of New Guinean begonia was expected to show high level of genetic diversity, as well as section *Petermannia*, other section were still have to be investigated.

Section *Oligandrae* whose species are locally common, have restricted distributions compared to other sections.

This is especially pertinent to *B. pentandra* and *B. chambersiae* that have evidence of limestone endemism. Our results show a surprising range of genetic structure in PNG begonia, identifying low levels of homozygosity, and suggesting that most of the species in excellent genetic health.

5.4.2 Clusters of F_{ROH}

While high-F_{ROH} provides unambiguous results, the outliers cluster gives questionable results in terms of relation with the low-F_{ROH} cluster. The outliers group of samples are at the limit of detection of our pipeline, B.kaniensis_224 displaying 53 ROHs, B.kaniensis_221 23 ROHs, and B.brassi_202a 38 ROHs, where the high-F_{ROH} cluster starts at 180 ROHs. Furthermore, the B.brassi_202 duplicate has been rejected by the pipeline while having a coverage value similar to B.brassi_202a.

The *B. kaniensis* outliers belong to the group of lower-F_{ROH} specimens from section *Diploclinium.* This species is a monoecious vine with distinct sections of the stem bearing huge clusters of male and female flowers (Fig.5.18). In *Begonia* separation of sexes in different inflorescences is usually associated with different flowering times for each sex, providing an isolation barrier for cross-pollination. However, several cases of overlapping flowering times for male and female flowers have been recorded (Wilson, personal observation). The two outliers of this group may be suffering from localized inbreeding due to same-population cross pollination, being found at high altitudes that might act as an isolating barrier for pollination. Another possibility is that the localized inbreeding might be the result of selfing by geitonogamy, resulting from overlapping flowering times.

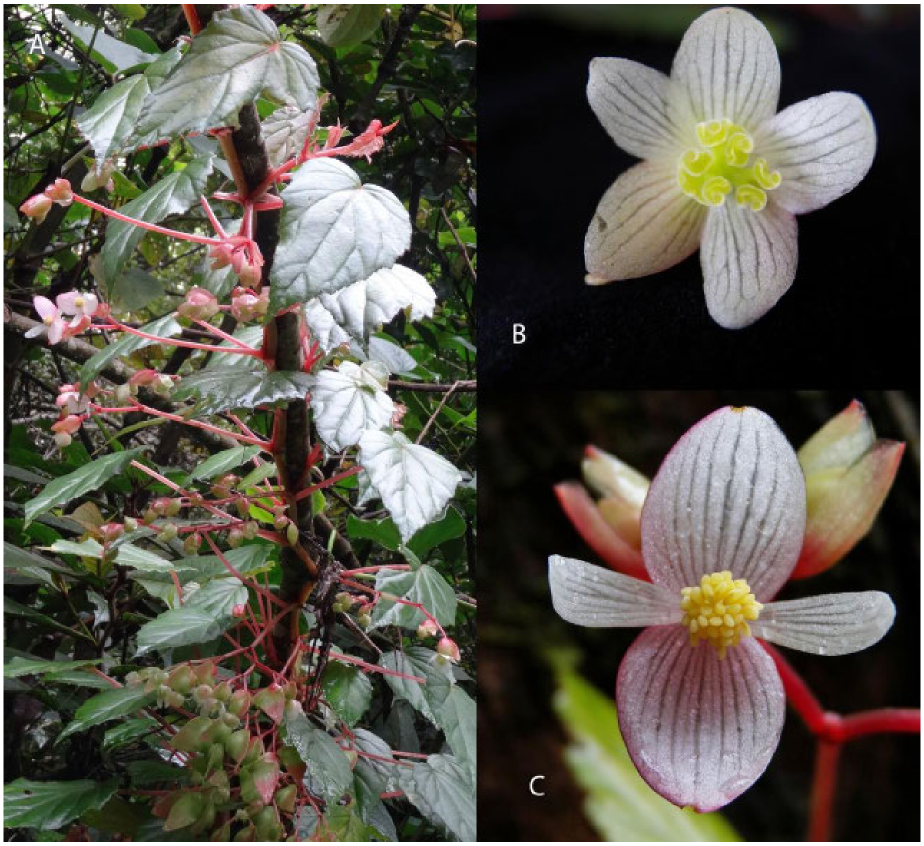


Figure 5.18: *Begonia kaniensis*. (a) lianescent habit, (b) female flower, (c) male flower (Wilson, 2021) .

5.4.3 Historical material

The analysis of Papua New Guinea data has revealed a low influence of age of the specimens on detection of ROH on this set of specimens. The oldest specimens are B.Petermannia_89 (1953), B.kaniensis_125 (1962), and B.Petermannia_156 (1963). All show low-F_{ROH}, while B.Symbegonia_140 (1987), B.Petermannia_133 (1987), and B.kaniensis_153a (1981) show among the highest scores of F_{ROH} and depth of coverage of the entire set. It is still unclear if the most ancient specimens did not show high-F_{ROH} due to insufficient depth of coverage or due to their genetic background. Further studies are necessary to assess the influence of age on ROH quantification. Nonetheless, these specimens can be analysed to produce genetic structure detectable by our pipeline. More recent specimens, collected in the 1980's show various levels of homozygosity, clearly more related to their genome architecture than to their age.

5.4.4 Limits of the pipeline

We have observed that a majority of the specimens with low-depth of sequencing coverage show low-F_{ROH} or have been removed from the analysis by our filters. It is due to the low number of SNPs confidently called by our pipeline for these specimens. The *Begonia* bait set used in this study has shown its capacity to capture long contiguous segment of the genome, usable for ROH analysis. The critical requirement to successfully estimate size and number of ROH in a genome seems to be success of target capture and sufficient read depth of coverage during the sequencing process. While this parameter is crucial for robustness of analysis, we demonstrate here that a depth of coverage as low as 2.5 is enough to detect unambiguously ROH in fresh and historical specimens.

5.4.5 Further studies

We have observed that most of the *Begonia* species part of the central mountain range of Papua New Guinea show an unexpected level of genetic diversity, and only a few show high level of homozygosity. The demographic history estimation we made does not show a situation of inbreeding, but rather a constriction in the population size and in some cases a bottleneck event in their recent demographic history. This does not match the expectations for high-altitude populations with restricted distribution, and isolated from each other. Further studies will investigate the relationships between these populations and track possible signature of introgression between them. Moreover, this study will be contrasted with the analysis of other *Begonia* groups to compare patterns of homozygosity among them. A differential study might explain the surprising genetic health of *Begonia* from Papua New Guinea, and help understanding the mechanisms of inbreeding or preventing inbreeding in this very diverse genus.

5.4.6 Subsequent analysis using F-statistics and heterozygosity rate

The F-statistics estimators, species F_{ST} and individuals F_{IS} have failed to show any relationship with the previously calculated F_{ROH} index, or with the phylogeny that has been established by previous studies (Wilson, 2021). First, the F_{ST} by species values were not congruent with the genetic distance between species or the average level of F_{ROH} for each species. The clusters of High-F_{ROH}, low-F_{ROH}, and outliers were not overlapping with the scores of F_{IS} found. Furthermore, the F_{IS} results do not match the phylogenetic reconstruction established previously and shown labelled in Figure C.3. We therefore have to invalidate the hypothesis that F-statistics could be used at species-scale on all the species of the PNG dataset.

The comparison between heterozygosity rate and F_{ROH} has provided contrasting results. While the PNG set has shown a null correlation between the two indices, running the pipeline on another clade of Asian *Begonia* section *Coelocentrum* has revealed a strong correlation between heterozygosity rate and F_{ROH}. While this dataset is quoted here as note added in proof, and should be described further for these results to be properly considered, this correlation can suggest our approach may work, but there is a bias in the PNG results explaining why no correlation cannot be seen between the indices.

In the light of this observation, and as the heterozygosity ratio varies across all the group of PNG specimens, it is unlikely that the few high values of F_{ROH} would be a biological reality, but more likely a technical issue with our pipeline. The main difference in processing the PNG dataset compared to the Asian dataset is that the variants have been called in three batches for technical reasons. It is as well noticeable that many specimens are showing null F_{ROH} as most of the ROH detected are filtered out of the pipeline by the minimum SNPs threshold of the pipeline. Therefore Further experiments could be set up to explore this possibility and test if a separate variant call might impact the final results of the ROH pipeline. The next step of our project should include as well the analysis of a *Begonia* long reads dataset with our pipeline, as a full-length pseudo-chromosomes assembly can display ROH closer to the biological reality than our target capture dataset. Detecting the ROH on this dataset would allow to give an accurate estimation of the number and length of ROH in *Begonia* genome and estimate precisely if our estimation with target capture data can reflect the overall genomic content in ROH.

Chapter 6

conclusion

6.1 Conservation genomics

The results of this project have a direct application in the field of conservation genetics (CG). CG is the study of the genetic factors that affect extinction risk, the genetic management regimes to minimize these risks, and the use of molecular genetics to determine aspects of species important for their conservation (Fox et al., 2006).

One of the major issue in conservation genetics is to detect loss of genetic diversity and ability to evolve in response to environmental change. Genetic diversity and allelic diversity at population level are key concept in CG, as reduced genetic or allelic diversity has been associated with reduced fitness and linked to higher extinction rate (Fox et al., 2006, Ollivier et al., 2013). Several methods can be used to detect lower values of genetic variation in populations, and potentially inbreeding depression : identification of the Evolutionary Significant Units (ESUs) (Moritz, 1999), calculation of effective population size (N_e) (Willi et al., 2022), direct measures of heterozygosity (Genereux et al., 2020), assess genes of fitness among populations (Wayne et al., 2004), or exploring the admixture between populations (Supple et al., 2018).

Population size in particular is an important concept in CG, as the articles that have founded the basis of the field have linked the concept of effective population size and genetic marker diversity (Wright, 1931, Fraser, 1972).

Drift-effective population size (N_e) is linked to the loss of heterozygosity and fixation or loss of alleles (Willi et al., 2022). Contrary to the census population size (N), N_e represents the size of an idealized population to display the same genetic diversity or inbreeding rate observed in the actual population. It it idealized because in this case we consider random mating, simultaneous birth of each generation, constant population size, equal number of children per parent (Charlesworth, 2009). This parameter is indicative of the effect of genetic drift on a population, and its drop can indicate the loss of heterozygosity and fixation or loss of alleles (Willi et al., 2022). Calculation of N_e can be made with demographic or genetic data, and with different algorithms (Charlesworth, 2009).

Unfortunately, the different models to calculate N_e all rely on alleles frequencies. Other tools that might help to discriminate genetic erosion and inbreeding in vulnerable populations are F-statistics (Kramer et al., 2009). Unfortunately again, alleles frequencies are still required, which means that several individuals have to be available for any assessment about the genetic health of a population.

Several other issues are problematic for vulnerable species detection: assessments directed to loci with neutral variation or fitness-related genes (Teixeira et al., 2021), or the lack of detailed genealogical information and their influence on behaviour and demography (Wayne et al., 2004).

Our method to detect inbreeding depression has been selected to take in account such limitations as low number of specimens from the populations studied and a lack of phenotypic informations on the individuals (Genereux et al., 2020). We rely on ROH and overall heterozygosity rate, which does not require the segregation of neutral or fitness-related loci. As specimens and robust information about rare tropical herbaceous plants are difficult to collect, such a method is a precious tool for CG assessments.

6.2 Mutational meltdown as a vector of extinction in the *Begonia* genus

Our work started with the hypothesis that inbreeding and mutational meltdown reduce *Begonia* population size and eventually drive them to extinction. As many tropical herbaceous plants, the genus *Begonia* is susceptible to form small and isolated population, with low gene flow between populations. High level of homozygosity would affect their fertility and fitness, shrinking the number of individuals in a population. This would favour the fixation of deleterious alleles in the population and reduce genetic diversity, decreasing their capacity to adapt to new threats. To assess this possibility, we planned to investigate the genome of present and historical *Begonia* specimens. If our hypothesis is valid, the genomes of rare or extinct species would be highly homozygous, this would be recent, and the homozygous alleles would be distributed evenly across chromosomes.

6.3 Setting up pipeline for complicated genome architecture.

We firstly wanted to check several parameters in the *Begonia* mapping population:

- Can we robustly call variants?
- Are the baits capturing contiguous segment of the genome?
- Are the baits capturing syntenic sequences across *Begonia* groups?

- Can we detect paralogous SNPs accurately?

The target capture protocol failed on the mapping population set, and provided a ragged NGS dataset difficult to use to set up our analysis pipeline. However, a genome skimming dataset from F1 hybrids of the same population had been produced by Cynthia Fan and was used to test parameters. The SNPs set called by different tools are variable, therefore we selected the most robust method of variant calling possible. We joint-called SNPs with the GATK toolkit, GATK haplotype-caller being reputed the most efficient tool for joint calling. SNPs were filtered out of the analysis in order to minimize the fake positives SNPs. Subsequently, a subset of the baits were identified as capturing paralogous sequences using three different methods of detection based on heterozygosity level, genotype frequencies in the mapping population, and occurrence of baits on different scaffolds of a genome. The overlap between the baits identified to capture paralogs between the three methods is comparatively small. Only 73 baits are in this overlap, where 635 were tagged paralog by the genotype frequency method, 135 by a module of HybPiper, and 310 by HDplot, a tool detecting higher SNPs heterozygosity level than expected. Even though problematic baits have been excluded from our analysis, the comparatively narrow overlap between these method is unexpected. Further assays with variation of the sensitivity and specificity of these paralogous baits detection tools might provide a better overlap and identify the whole set of problematic baits.

6.4 Socotran *Begonia* dataset: when target capture goes wrong

Begonia from the Socotran archipelago have been studied since the 19th century along with all the vegetation of the island well known for its exceptional degree of endemism. Herbarium specimens were collected on the island from the 19th century until present day, and a living collection has been established in the RBGE from the Dr. Mark Hughes sampling in the island in 1999. The two species considered were selected early for study. First as a microsatellites study of genetic distances between *B.socotrana* populations was available (Hughes et al., 2002b) and a re-assessment of the phylogeny of the group involving more molecular markers was relevant to understand the demographic history of the species on the island. Secondary, populations of *B. socotrana* a relatively isolated and *B. samhaensis* is exceptional for its high degree of endemism. Testing their genetic health, and the evolution of it over time would has provided clues about how to re-assess their conservation status. Lastly, herbarium and silica-dried specimens availability could have enabled to build a time series of samples to study evolution of their demographic history over time. Unfortunately, the target capture protocol failed, and the uneven coverage did not allow to recover all targeted regions of the genome. The dataset was nonetheless processed through the pipeline to explore the limits of our pipeline. The final estimation of homozygosity level in the specimens has not been possible as a result of the fragmented data, and the few number of consecutive SNPs supporting each ROH detected. Regardless, our data were analysed and plotted to show what patterns can be still be found in this incomplete data set.

The analysis of demographic history using ROH as marker of endemism has failed to give statistically significant results. The limit of detection of this method has been reached with short length of the genome having an even coverage, and ROH detection not supported with sufficient markers. Future implementation of other inbreeding estimators in our pipeline might however overcome this limitation and enable to re-analyse this dataset.

The phylogenetic analysis of our set of samples confirms the findings of Hughes et al., 2002c with respect to relations between populations, However, several discrepancies were observed, notably almost all historical *B. socotrana* specimens grouping with a single specimen of *B. samhaensis*. Two historical specimens from 1989 cluster with *B. socotrana* from the West Haggier mountains, close to a specimen from Reiged. But the overall uncertainty about the phylogenetic results would indicate this relationship as questionable until new studies could corroborate this assumption.

6.5 Exploring the genetic diversity of Papua New Guinea Begonia

Collaboration with Dr. Hannah Wilson and Dr. Mark Hughes have provided access to an Hyb-Seq dataset of specimens from Papua New Guinea originally collected to explore diversity of the different clades on the island. In contrast to this study, we have explored patterns of endemism in this group to assess the genetic diversity of the different taxa and place them in their geographic context. We expected most species to be generally low in genetic diversity due to isolation. Several species such as *B. kaniensis* which is wind pollinated and widespread were expected to show high genetic diversity. However, species like *B. pentandra* and *B. chambersiae* are limestone endemic species, and expected to show very low genetic diversity. Species from section Oligandrae have been observed to be locally common, but with restricted distribution, and a lower rate of homozygosity was expected from them as well. Our results show a surprising range of genetic structure in Papua New Guinea Begonia. Some species were found with low patterns of homozygosity and therefore in excellent genetic health. Several specimens showed higher patterns of homozygosity, possibly coming from populations that are recent colonies, and suffering from geitonogamy inbreeding (family=Jong et al., 1993).

We have observed the formation an outlier group composed of *B. brassi* and *B. kaniensis* specimens with low-F_{ROH} longer ROH than seen in the other member of these species. These specimens could be the results of inbreeding by geitonogamy. However, the low rate of overall homozygosity in these specimens place them at the limit of detection of our pipeline, as demonstrated by the rejection of the replicate of one of these *B. brassi* sample. The significance of these samples can therefore be questioned, and the relatively long ROH observed can be either a biological reality, or noise in the data. In the later case, these outliers may be the remain of low-F_{ROH} ROH that have not been filtered out by the pipeline SNPs number threshold.

Studying the geographic distribution of these samples does not reveal an obvious relation between geographical context and homozygosity level, nor does the topography of sampling sites, or altitudinal distribution of the specimens. We do not observe similar patterns of homozygosity within populations or populations closely related. Therefore, it seems that homozygosity patterns observed are not due to genetic drift driven by microevolution, but rather a stochastic trait inherited phylogenetically.

A bias in our analysis is that the specimens have been collected along the central mountain range of Papua New Guinea. Most of the specimens have been collected at high altitude ranging between 1,000m to 2,000m, and further sampling in the lowland could reveal patterns of endemism or diversity in contrast to the present set. The central mountain range expanding from the North-West to South-East of the island, further sampling in other sites may reveal patterns of migration or colonisation not observed with our set of specimens. Our pipeline has been able to track patterns of endemism from silica-dried and herbarium historical specimens without apparent bias due to the age of the specimens. Further analysis including time series of samples might indicate the evolution of genetic diversity over time and indicate movement of populations in the island.

6.6 Publication of the pipeline and future work

Our pipeline will be adapted to the workflow management system Snakemake, and made available to the larger scientific community via publication on GitHub. Large datasets of NGS *Begonia* are being prepared to run on the pipeline, and more are being produced now as candidates for it. The genetic health of the incoming specimens will be assessed by the pipeline, and comparison between different groups of *Begonia* could be made. Preliminary comparative studies Chinese *Begonia* from section *Coelocentrum* showed a specimens displaying higher pattern of homozygosity than the two other clades studied to this day. This would corroborate our assumptions on the patterns of endemism of these lineage as a many of the species in this set are cave-dwelling and have a very restricted area of dispersal.

The pipeline will be used to make a quick assessment of the genetic vulnerability of these species, observe the conservation of this trait across their lineages, and the influence it could have on the radiation of the different taxa studied. Furthermore , the analysis of *Begonia* genetic diversity and patterns of endemism could help to solve ecological mysteries, as the co-existence of widespread and endemic species morphologically similar in the same ecological context (Chan et al., 2018, Chan et al., 2019).

Other functionalities of the pipeline have still to be developed for further enquiries. The absence of observable aDNA patterns of damage, even moderate, was surprising for the historical specimens analysed in the two sets of samples. Other similar studies on plant herbarium specimens involving specimens of comparable age show obvious aDNA patterns of damages (Bieker et al., 2020, Canales et al., 2022). However, this low rate of aDNA signature is most probably due to the method of library preparation involving an UDG enzyme and is not representative of the actual nucleotides substitution rate in our specimens. This method of library preparation was avoided in the studies quoted above, and we expect to see aDNA signature in the future analysis on UDG-free libraries dataset with our pipeline.

Other population genetics tools will be implemented as well to corroborate the ROH estimation. F-statistics will be used to confirm the presence of genetic drift and F_{ST} compared to F_{ROH} to observe if a linear correlation can be traced. A particular interest will be taken to integrate D-statistics to the pipeline, as gene flow between *Begonia* species might provide the explanation for genetic isolation.

One purpose of this study has not been reached yet, to study the evolution of genetic diversity in *Begonia* over time. As climate changes as faster rate and biodiversity drops down there is a need to re-assess quickly the conservation status of potential threatened species, but moreover to monitor this decline to understand the future changes in biodiversity. Using a time series of pre and post-industrial historical specimens could show the evolution of genetic fitness over time and be used to estimate rates of extinction due to the Anthropocene. Studying endemism could as well unravel plant populations dynamics as changes of distribution, patterns of colonisation, or persistence in a refugia.

Bibliography

Abbott, Richard J. et al. (Aug. 25, 2000). “Molecular Analysis of Plant Migration and Refugia in the Arctic”. In: *Science (New York, N.Y.)* 289.5483, pp. 1343–1346. DOI: 10.1126/science.289.5483.1343. URL: https://www.science.org/doi/10.1126/science.289.5483.1343 (visited on 11/29/2022).

Abdel-Latif, Amani and Gamal Osman (Jan. 3, 2017). “Comparison of Three Genomic DNA Extraction Methods to Obtain High DNA Quality from Maize”. In: *Plant Methods* 13.1, p. 1. ISSN: 1746-4811. DOI: 10.1186/s13007-016-0152-4. URL: https://doi.org/10.1186/s13007-016-0152-4 (visited on 04/25/2023).

Aboul-Maaty, Nadia Aboul-Ftooh and Hanaa Abdel-Sadek Oraby (Feb. 12, 2019). “Extraction of High-Quality Genomic DNA from Different Plant Orders Applying a Modified CTAB-based Method”. In: *Bulletin of the National Research Centre* 43.1, p. 25. ISSN: 2522-8307. DOI: 10.1186/s42269-019-0066-1. URL: https://doi.org/10.1186/s42269-019-0066-1 (visited on 04/25/2023).

Agren, Jon and Douglas W. Schemske (Sept. 1991). “Pollination by Deceit in a Neotropical Monoecious Herb, Begonia Involucrata”. In: *Biotropica* 23.3, p. 235. ISSN: 00063606. DOI: 10.2307/2388200. JSTOR: 2388200. URL: https://www.jstor.org/stable/2388200?origin=crossref (visited on 11/25/2022).

Albani Rocchetti, Giulia et al. (Apr. 2021). “Reversing Extinction Trends: New Uses of (Old) Herbarium Specimens to Accelerate Conservation Action on Threatened Species”. In: *New Phytologist* 230.2, pp. 433–450. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.17133. URL: https://onlinelibrary.wiley.com/doi/10.1111/nph.17133 (visited on 08/18/2022).

Alsos, Inger Greve et al. (Sept. 30, 2016). “The Role of Sea Ice for Vascular Plant Dispersal in the Arctic”. In: *Biology Letters* 12.9, p. 20160264. DOI: 10.1098/rsbl.2016.0264. URL: https://royalsocietypublishing.org/doi/10.1098/rsbl.2016.0264 (visited on 11/29/2022).

Alsos, Inger Greve et al. (May 1, 2021). “Ancient Sedimentary DNA Shows Rapid Post-Glacial Colonisation of Iceland Followed by Relatively Stable Vegetation until the Norse Settlement (Landnám) AD 870”. In: *Quaternary Science Reviews* 259, p. 106903. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2021.106903. URL: https://www.sciencedirect.com/science/article/pii/S0277379121001104 (visited on 11/29/2022).

Andermann, Tobias et al. (July 13, 2018). “SECAPR—a Bioinformatics Pipeline for the Rapid and User-Friendly Processing of Targeted Enriched Illumina

Sequences, from Raw Reads to Alignments". In: *PeerJ* 6, e5175. ISSN: 2167-8359. DOI: 10.7717/peerj.5175. URL: https://peerj.com/articles/5175 (visited on 07/22/2021).

Anderson-Carpenter, Lynn L et al. (Dec. 2011). "Ancient DNA from Lake Sediments: Bridging the Gap between Paleoecology and Genetics". In: *BMC Evolutionary Biology* 11.1, p. 30. ISSN: 1471-2148. DOI: 10.1186/1471-2148-11-30. URL: https://bmcevolbiol.biomedcentral.com/articles/10.1186/1471-2148-11-30 (visited on 04/10/2023).

APG (2003). "An Update of the Angiosperm Phylogeny Group Classification for the Orders and Families of Flowering Plants: APG II". In: *Botanical Journal of the Linnean Society* 141.4, pp. 399–436. ISSN: 1095-8339. DOI: 10.1046/j.1095-8339.2003.t01-1-00158.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1095-8339.2003.t01-1-00158.x (visited on 07/07/2023).

Aravanopoulos, Filippos A., Ioannis Ganopoulos, and Athanasios Tsaftaris (Jan. 1, 2015). "Chapter Four - Population and Conservation Genomics in Forest and Fruit Trees". In: *Advances in Botanical Research*. Ed. by Christophe Plomion and Anne-Françoise Adam-Blondon. Vol. 74. Land Plants - Trees. Academic Press, pp. 125–155. DOI: 10.1016/bs.abr.2015.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0065229615000300 (visited on 08/04/2023).

Ash, Jeremy D., Thomas J. Givnish, and Donald M. Waller (Mar. 2017). "Tracking Lags in Historical Plant Species' Shifts in Relation to Regional Climate Change". In: *Global Change Biology* 23.3, pp. 1305–1315. ISSN: 1354-1013, 1365-2486. DOI: 10.1111/gcb.13429. URL: https://onlinelibrary.wiley.com/doi/10.1111/gcb.13429 (visited on 09/15/2022).

Attorre, F. et al. (May 4, 2014). "Analysing the Relationship between Land Units and Plant Communities: The Case of Socotra Island (Yemen)". In: *Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology* 148.3, pp. 529–539. ISSN: 1126-3504. DOI: 10.1080/11263504.2014.900127. URL: https://doi.org/10.1080/11263504.2014.900127 (visited on 08/06/2023).

Attorre, Fabio et al. (Sept. 1, 2007). "Will Dragonblood Survive the next Period of Climate Change? Current and Future Potential Distribution of Dracaena Cinnabari (Socotra, Yemen)". In: *Biological Conservation* 138.3, pp. 430–439. ISSN: 0006-3207. DOI: 10.1016/j.biocon.2007.05.009. URL: https://www.sciencedirect.com/science/article/pii/S0006320707002273 (visited on 08/07/2023).

Bakker, Freek T. (2022). "Herbarium DNA Degradation: Different Ways of Falling to Pieces". In: Bauhin_2022, 400 Years Botanical Collections, Implications for Present-Day Research. Basel, Swisserland: Wageningen University & Research.

Bakker, Freek T. et al. (Jan. 2016). "Herbarium Genomics: Plastome Sequence Assembly from a Range of Herbarium Specimens Using an Iterative Organelle Genome Assembly Pipeline". In: *Biological Journal of the Linnean Society* 117.1, pp. 33–43. ISSN: 00244066. DOI: 10.1111/bij.12642. URL: https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/bij.12642 (visited on 11/05/2020).

Banfield, Lisa M., Kay Van Damme, and Anthony G. Miller (2011). “Evolution and Biogeography of the Flora of the Socotra Archipelago (Yemen)”. In: *The Biology of Island Floras*. Ed. by David Bramwell and Juli Caujapé-Castells. Cambridge: Cambridge University Press, pp. 197–225. isbn: 978-0-521-11808-8. doi: 10.1017/CBO9780511844270.009. url: https://www.cambridge.org/core/books/biology-of-island-floras/evolution-and-biogeography-of-the-flora-of-the-socotra-archipelago-yemen/136A078884889E401942503CE69FBF7E (visited on 08/07/2023).

Bank, Claudia et al. (Dec. 1, 2014). “Thinking Too Positive? Revisiting Current Methods of Population Genetic Selection Inference”. In: *Trends in Genetics* 30.12, pp. 540–546. issn: 0168-9525. doi: 10.1016/j.tig.2014.09.010. pmid: 25438719. url: https://www.cell.com/trends/genetics/abstract/S0168-9525(14)00158-9 (visited on 11/29/2022).

Barrett, Craig F. et al. (2016). “An Introduction to Plant Phylogenomics with a Focus on Palms”. In: *Botanical Journal of the Linnean Society* 182.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/boj.12399, pp. 234–255. issn: 1095-8339. doi: 10.1111/boj.12399. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/boj.12399 (visited on 11/28/2022).

“Begonia Bekopakensis” (1983). In: *Flora of Madagascar* 24, p. 144.

Beissinger, Timothy M. et al. (June 13, 2016). “Recent Demography Drives Changes in Linked Selection across the Maize Genome”. In: *Nature Plants* 2.7, pp. 1–7. issn: 2055-0278. doi: 10.1038/nplants.2016.84. url: https://www.nature.com/articles/nplants201684 (visited on 11/28/2022).

Bellorini, Cristina (Sept. 16, 2016). *The World of Plants in Renaissance Tuscany: Medicine and Botany*. London: Routledge. 280 pp. isbn: 978-1-315-55139-5. doi: 10.4324/9781315551395.

Bieker, Vanessa C. and Michael D. Martin (Oct. 2, 2018). “Implications and Future Prospects for Evolutionary Analyses of DNA in Historical Herbarium Collections”. In: *Botany Letters* 165.3-4, pp. 409–418. issn: 2381-8107, 2381-8115. doi: 10.1080/23818107.2018.1458651. url: https://www.tandfonline.com/doi/full/10.1080/23818107.2018.1458651 (visited on 02/09/2021).

Bieker, Vanessa C. et al. (Sept. 2020). “Metagenomic Analysis of Historical Herbarium Specimens Reveals a Postmortem Microbial Community”. In: *Molecular Ecology Resources* 20.5, pp. 1206–1219. issn: 1755-098X, 1755-0998. doi: 10.1111/1755-0998.13174. url: https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13174 (visited on 12/16/2020).

Bieker, Vanessa C. et al. (Aug. 26, 2022). “Uncovering the Genomic Basis of an Extraordinary Plant Invasion”. In: *Science Advances* 8.34, eabo5115. issn: 2375-2548. doi: 10.1126/sciadv.abo5115. url: https://www.science.org/doi/10.1126/sciadv.abo5115 (visited on 08/27/2022).

Bilinski, Paul et al. (May 10, 2018). “Parallel Altitudinal Clines Reveal Trends in Adaptive Evolution of Genome Size in Zea Mays”. In: *PLOS Genetics* 14.5. Ed. by Gregory P. Copenhaver, e1007162. issn: 1553-7404. doi: 10.1371/journal.pgen.1007162. url: https://dx.plos.org/10.1371/journal.pgen.1007162 (visited on 11/06/2020).

Bitinaite, Jurate et al. (Mar. 2007). "USER™ Friendly DNA Engineering and Cloning Method by Uracil Excision". In: *Nucleic Acids Research* 35.6, pp. 1992–2002. ISSN: 0305-1048. DOI: 10.1093/nar/gkm041. pmid: 17341463. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1874603/ (visited on 12/01/2022).

Bjune, Anne E. et al. (July 30, 2021). "Rapid Climate Changes during the Lateglacial and the Early Holocene as Seen from Plant Community Dynamics in the Polar Urals, Russia". In: *Journal of Quaternary Science*, jqs.3352. ISSN: 0267-8179, 1099-1417. DOI: 10.1002/jqs.3352. URL: https://onlinelibrary.wiley.com/doi/10.1002/jqs.3352 (visited on 05/27/2022).

Boast, Alexander P. et al. (Feb. 13, 2018). "Coprolites Reveal Ecological Interactions Lost with the Extinction of New Zealand Birds". In: *Proceedings of the National Academy of Sciences* 115.7, pp. 1546–1551. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1712337115. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1712337115 (visited on 11/05/2020).

Bosser, Aymonin & (1983). "Begonia Monicae". In: *Begonia monicae Aymonin & Bosser, Fl. Madagasc. 144: 19 (1983).* Flora of Madagascar.19, p. 144. URL: https://www.gbif.org/occurrence/438448314.

BPG et al. (Aug. 18, 2022). "RESOLVING PHYLOGENETIC AND TAXONOMIC CONFLICT IN BEGONIA". In: *Edinburgh Journal of Botany* 79, pp. 1–28. ISSN: 1474-0036, 0960-4286. DOI: 10.24823/ejb.2022.1928. URL: https://journals.rbge.org.uk/ejb/article/view/1928 (visited on 11/25/2022).

Brennan, Adrian Christopher et al. (Dec. 2012). "Genomic Resources for Evolutionary Studies in the Large, Diverse, Tropical Genus, Begonia". In: *Tropical Plant Biology* 5.4, pp. 261–276. ISSN: 1935-9756, 1935-9764. DOI: 10.1007/s12042-012-9109-6. URL: http://link.springer.com/10.1007/s12042-012-9109-6 (visited on 08/16/2021).

Brewer, Grace E. et al. (Sept. 18, 2019). "Factors Affecting Targeted Sequencing of 353 Nuclear Genes From Herbarium Specimens Spanning the Diversity of Angiosperms". In: *Frontiers in Plant Science* 10, p. 1102. ISSN: 1664-462X. DOI: 10.3389/fpls.2019.01102. URL: https://www.frontiersin.org/article/10.3389/fpls.2019.01102/full (visited on 05/16/2022).

Bridson D, L Forman (2000). *The Herbarium Handbook*. Royal Botanic Gardens, United Kingdom. ISBN: ISBN 10: 1900347431 ISBN 13: 9781900347433.

Briggs, A. W. et al. (Sept. 11, 2007). "Patterns of Damage in Genomic DNA Sequences from a Neandertal". In: *Proceedings of the National Academy of Sciences* 104.37, pp. 14616–14621. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0704665104. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0704665104 (visited on 11/06/2020).

Briggs, Adrian W. et al. (Apr. 2010). "Removal of Deaminated Cytosines and Detection of in Vivo Methylation in Ancient DNA". In: *Nucleic Acids Research* 38.6, e87–e87. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkp1163. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp1163 (visited on 12/01/2022).

Briggs, D. et al. (July 1983). "MERCURY VAPOUR: A HEALTH HAZARD IN HERBARIA". In: *New Phytologist* 94.3, pp. 453–457. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/j.1469-8137.1983.tb03458.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.1983.tb03458.x (visited on 09/26/2022).

Broman, Karl W. and James L. Weber (Dec. 1999). "Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude Du Polymorphisme Humain". In: *The American Journal of Human Genetics* 65.6, pp. 1493–1500. ISSN: 00029297. DOI: 10.1086/302661. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929707626779 (visited on 04/11/2023).

Brouard, Jean-Simon et al. (June 21, 2019). "The GATK Joint Genotyping Workflow Is Appropriate for Calling Variants in RNA-seq Experiments". In: *Journal of Animal Science and Biotechnology* 10.1, p. 44. ISSN: 2049-1891. DOI: 10.1186/s40104-019-0359-0. URL: https://doi.org/10.1186/s40104-019-0359-0 (visited on 10/21/2022).

Brown, A. G. et al. (June 3, 2021). "Ancient DNA, Lipid Biomarkers and Palaeoecological Evidence Reveals Construction and Life on Early Medieval Lake Settlements". In: *Scientific Reports* 11.1, p. 11807. ISSN: 2045-2322. DOI: 10.1038/s41598-021-91057-x. URL: https://www.nature.com/articles/s41598-021-91057-x (visited on 10/10/2022).

Brown, Gary and Bruno A. Mies (2012). *Vegetation Ecology of Socotra*. Vol. 7. Plant and Vegetation. Dordrecht: Springer Netherlands. ISBN: 978-94-007-4140-9 978-94-007-4141-6. DOI: 10.1007/978-94-007-4141-6. URL: http://link.springer.com/10.1007/978-94-007-4141-6 (visited on 12/01/2022).

Brown, Terence A. (Jan. 29, 1999). "How Ancient DNA May Help in Understanding the Origin and Spread of Agriculture". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 354.1379. Ed. by M. K. Jones et al., pp. 89–98. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.1999.0362. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.1999.0362 (visited on 01/22/2021).

Bryc, Katarzyna, Nick Patterson, and David Reich (Oct. 2013). "A Novel Approach to Estimating Heterozygosity from Low-Coverage Genome Sequence". In: *Genetics* 195.2, pp. 553–561. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.113.154500. URL: http://www.genetics.org/lookup/doi/10.1534/genetics.113.154500 (visited on 01/29/2021).

Bujang, Mohamad Adam and Nurakmal Baharum (Mar. 10, 2016). "Sample Size Guideline for Correlation Analysis". In: *World Journal of Social Science Research* 3.1, p. 37. ISSN: 2332-5534, 2375-9747. DOI: 10.22158/wjssr.v3n1p37. URL: http://www.scholink.org/ojs/index.php/wjssr/article/view/398 (visited on 07/12/2023).

Cabassi, Jacopo et al. (Jan. 2020). "100 Years of High GEM Concentration in the Central Italian Herbarium and Tropical Herbarium Studies Centre (Florence, Italy)". In: *Journal of Environmental Sciences* 87, pp. 377–388. ISSN: 10010742. DOI: 10.1016/j.jes.2019.07.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S1001074219312082 (visited on 09/26/2022).

Caetano-Anolles, Derek (Oct. 27, 2022). *Hard-Filtering Germline Short Variants*. GATK. URL: https://gatk.broadinstitute.org/hc/en-us/articles/360035890471-Hard-filtering-germline-short-variants (visited on 11/29/2022).

Campos-Dominguez (2020). “Does a Dynamic Genome Drive Speciation in a Mega-Diverse Genus?” University of Edinburgh.

Campos-Dominguez, Lucia et al. (Aug. 18, 2022). “EVOLUTIONARY PATTERNS OF GENOME SIZE AND CHROMOSOME NUMBER VARIATION IN BEGONIACEAE”. In: *Edinburgh Journal of Botany* 79, pp. 1–28. issn: 1474-0036, 0960-4286. doi: 10.24823/ejb.2022.1876. url: https://journals.rbge.org.uk/ejb/article/view/1876 (visited on 11/30/2022).

Campos-Domınguez, Lucıa (2022). “A Study on the Evolution of Genome Dynamics across the Mega-Diverse Genus Begonia L. (Begoniaceae)”. In: p. 259.

Canales, Nataly Allasi et al. (Apr. 28, 2022). *Museomic Approaches to Genotype Historic* Cinchona *Barks*. preprint. Genomics. doi: 10.1101/2022.04.26.489609. url: http://biorxiv.org/lookup/doi/10.1101/2022.04.26.489609 (visited on 11/27/2022).

Capo, Eric et al. (Feb. 13, 2021). “Lake Sedimentary DNA Research on Past Terrestrial and Aquatic Biodiversity: Overview and Recommendations”. In: *Quaternary* 4.1, p. 6. issn: 2571-550X. doi: 10.3390/quat4010006. url: https://www.mdpi.com/2571-550X/4/1/6 (visited on 11/04/2021).

Cappellini, Enrico et al. (June 20, 2018). “Ancient Biomolecules and Evolutionary Inference”. In: *Annual Review of Biochemistry* 87, pp. 1029–1060. issn: 1545-4509. doi: 10.1146/annurev-biochem-062917-012002. pmid: 29709200.

Carranza-Rojas, Jose et al. (Dec. 2017). “Going Deeper in the Automated Identification of Herbarium Specimens”. In: *BMC Evolutionary Biology* 17.1, p. 181. issn: 1471-2148. doi: 10.1186/s12862-017-1014-z. url: http://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-017-1014-z (visited on 09/15/2022).

Ceballos, Francisco C. et al. (Apr. 2018). “Runs of Homozygosity: Windows into Population History and Trait Architecture”. In: *Nature Reviews Genetics* 19.4, pp. 220–234. issn: 1471-0056, 1471-0064. doi: 10.1038/nrg.2017.109. url: http://www.nature.com/articles/nrg.2017.109 (visited on 05/05/2022).

Chan, Ym et al. (Nov. 15, 2019). “Understanding Rarity in a Narrow Endemic Begonia through Biological Comparison with a Common Species”. In: *JOURNAL OF TROPICAL FOREST SCIENCE* 31.4, pp. 422–432. issn: 01281283, 25219847. doi: 10.26525/jtfs2019.31.4.422. url: https://info.frim.gov.my/infocenter_applications/jtfsonline/jtfs/v31n4/422-432.pdf (visited on 01/12/2021).

Chan, Yoke Mui et al. (Jan. 2, 2018). “Limited Dispersal and Geographic Barriers Cause Population Differentiation and Structuring in *Begonia Maxwelliana* at Both Large and Small Scales”. In: *Plant Ecology & Diversity* 11.1, pp. 69–83. issn: 1755-0874, 1755-1668. doi: 10.1080/17550874.2018.1471625. url: https://www.tandfonline.com/doi/full/10.1080/17550874.2018.1471625 (visited on 01/12/2021).

Charlesworth, Brian (Mar. 2009). “Effective Population Size and Patterns of Molecular Evolution and Variation”. In: *Nature Reviews Genetics* 10.3 (3), pp. 195–205. issn: 1471-0064. doi: 10.1038/nrg2526. url: https://www.nature.com/articles/nrg2526 (visited on 08/07/2023).

Charlesworth, D and B Charlesworth (1987). “Inbreeding Depression and Its Evolutionary Consequences”. In.

Chaves, Lázaro José et al. (Apr. 1, 2011). “Estimating Inbreeding Depression in Natural Plant Populations Using Quantitative and Molecular Data”. In: *Conservation Genetics* 12.2, pp. 569–576. ISSN: 1572-9737. DOI: 10.1007/s10592-010-0164-y. URL: https://doi.org/10.1007/s10592-010-0164-y (visited on 08/04/2023).

Chen, Zhongsheng, Michael Boehnke, and Christian Fuchsberger (Jan. 2020). “Combining Sequence Data from Multiple Studies: Impact of Analysis Strategies on Rare Variant Calling and Association Results”. In: *Genetic epidemiology* 44.1, pp. 41–51. ISSN: 0741-0395. DOI: 10.1002/gepi.22261. pmid: 31520493. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7231418/ (visited on 11/29/2022).

Clark, Susie H. (Nov. 1986). “PRESERVATION OF HERBARIUM SPECIMENS: AN ARCHIVE CONSERVATOR’S APPROACH”. In: *TAXON* 35.4, pp. 675–682. ISSN: 0040-0262, 1996-8175. DOI: 10.2307/1221610. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/1221610 (visited on 09/30/2022).

Clarke, C. L. et al. (Dec. 23, 2019). “Persistence of Arctic-Alpine Flora during 24,000 Years of Environmental Change in the Polar Urals”. In: *Scientific Reports* 9.1, pp. 1–11. ISSN: 2045-2322. DOI: 10.1038/s41598-019-55989-9. URL: https://www.nature.com/articles/s41598-019-55989-9 (visited on 10/10/2022).

Clarke, Charlotte L. et al. (Nov. 1, 2020). “A 24,000-Year Ancient DNA and Pollen Record from the Polar Urals Reveals Temporal Dynamics of Arctic and Boreal Plant Communities”. In: *Quaternary Science Reviews* 247, p. 106564. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2020.106564. URL: https://www.sciencedirect.com/science/article/pii/S0277379120305266 (visited on 11/29/2022).

Clement, Wendy L. et al. (June 2004). “Phylogenetic Position and Biogeography of *Hillebrandia Sandwicensis* (Begoniaceae): A Rare Hawaiian Relict”. In: *American Journal of Botany* 91.6, pp. 905–917. ISSN: 0002-9122, 1537-2197. DOI: 10.3732/ajb.91.6.905. URL: https://onlinelibrary.wiley.com/doi/10.3732/ajb.91.6.905 (visited on 07/07/2023).

Cockerham, C. C. and B. S. Weir (Nov. 1968). “Sib Mating with Two Linked Loci”. In: *Genetics* 60.3, pp. 629–640. ISSN: 0016-6731. DOI: 10.1093/genetics/60.3.629. pmid: 5728745.

Cooper, A. (Aug. 18, 2000). “Ancient DNA: Do It Right or Not at All”. In: *Science (New York, N.Y.)* 289.5482, 1139b–1139. ISSN: 00368075, 10959203. DOI: 10.1126/science.289.5482.1139b. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.289.5482.1139b (visited on 04/09/2021).

Corlett, Richard T. (Feb. 2016). “Plant Diversity in a Changing World: Status, Trends, and Conservation Needs”. In: *Plant Diversity* 38.1, pp. 10–16. ISSN: 24682659. DOI: 10.1016/j.pld.2016.01.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2468265916300300 (visited on 11/25/2022).

Couvreur, Thomas L. P. et al. (2019). “Phylogenomics of the Major Tropical Plant Family Annonaceae Using Targeted Enrichment of Nuclear Genes”. In:

Frontiers in Plant Science 9. ISSN: 1664-462X. URL: https://www.frontiersin.org/articles/10.3389/fpls.2018.01941 (visited on 04/04/2023).

Couvreur, Thomas LP, Félix Forest, and William J. Baker (June 16, 2011). “Origin and Global Diversification Patterns of Tropical Rain Forests: Inferences from a Complete Genus-Level Phylogeny of Palms”. In: *BMC Biology* 9.1, p. 44. ISSN: 1741-7007. DOI: 10.1186/1741-7007-9-44. URL: https://doi.org/10.1186/1741-7007-9-44 (visited on 12/01/2022).

Cowman, Peter F. et al. (Dec. 2020). “An Enhanced Target-Enrichment Bait Set for Hexacorallia Provides Phylogenomic Resolution of the Staghorn Corals (Acroporidae) and Close Relatives”. In: *Molecular Phylogenetics and Evolution* 153, p. 106944. ISSN: 1095-9513. DOI: 10.1016/j.ympev.2020.106944. pmid: 32860973.

Cronn, Richard et al. (Feb. 2012). “Targeted Enrichment Strategies for Next-generation Plant Biology”. In: *American Journal of Botany* 99.2, pp. 291–311. ISSN: 0002-9122, 1537-2197. DOI: 10.3732/ajb.1100356. URL: https://onlinelibrary.wiley.com/doi/10.3732/ajb.1100356 (visited on 10/18/2022).

Crump, Sarah E. et al. (Mar. 30, 2021). “Ancient Plant DNA Reveals High Arctic Greening during the Last Interglacial”. In: *Proceedings of the National Academy of Sciences* 118.13, e2019069118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2019069118. URL: https://pnas.org/doi/full/10.1073/pnas.2019069118 (visited on 05/27/2022).

Cruzan, Mitchell B. and Alan R. Templeton (Dec. 2000). “Paleoecology and Coalescence: Phylogeographic Analysis of Hypotheses from the Fossil Record”. In: *Trends in Ecology & Evolution* 15.12, pp. 491–496. ISSN: 01695347. DOI: 10.1016/S0169-5347(00)01998-4. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534700019984 (visited on 02/15/2021).

Culley, Theresa M. (Nov. 2013). “Why Vouchers Matter in Botanical Research”. In: *Applications in Plant Sciences* 1.11, p. 1300076. ISSN: 2168-0450. DOI: 10.3732/apps.1300076. URL: http://doi.wiley.com/10.3732/apps.1300076 (visited on 05/08/2022).

Curik, Ino, Maja Ferenčaković, and Johann Sölkner (Aug. 1, 2014). “Inbreeding and Runs of Homozygosity: A Possible Solution to an Old Problem”. In: *Livestock Science*. Genomics Applied to Livestock Production 166, pp. 26–34. ISSN: 1871-1413. DOI: 10.1016/j.livsci.2014.05.034. URL: https://www.sciencedirect.com/science/article/pii/S1871141314003060 (visited on 04/11/2023).

Dabney, Jesse, Matthias Meyer, and Svante Pääbo (July 1, 2013). “Ancient DNA Damage”. In: *Cold Spring Harbor Perspectives in Biology* 5.7, a012567. ISSN: 1943-0264. DOI: 10.1101/cshperspect.a012567. pmid: 23729639.

daFonseca, RuteR et al. (Jan. 11, 2015). *The Origin and Evolution of Maize in the American Southwest.* preprint. Evolutionary Biology. DOI: 10.1101/013540. URL: http://biorxiv.org/lookup/doi/10.1101/013540 (visited on 11/06/2020).

Damme, Kay van (2022). *Nature and People in the Socotra Archipelago - UNESCO Digital Library.* UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381003 (visited on 08/06/2023).

Danecek, Petr et al. (Feb. 2021). “Twelve Years of SAMtools and BCFtools”. In: *GigaScience* 10.2. giab008. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab008.

eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. URL: https://doi.org/10.1093/gigascience/giab008.

De Summa, Simona et al. (Mar. 23, 2017). "GATK Hard Filtering: Tunable Parameters to Improve Variant Calling for next Generation Sequencing Targeted Gene Panel Data". In: *BMC Bioinformatics* 18.5, p. 119. ISSN: 1471-2105. DOI: 10.1186/s12859-017-1537-8. URL: https://doi.org/10.1186/s12859-017-1537-8 (visited on 11/29/2022).

Dehasque, Marianne et al. (Apr. 2020). "Inference of Natural Selection from Ancient DNA". In: *Evolution Letters* 4.2, pp. 94–108. ISSN: 2056-3744, 2056-3744. DOI: 10.1002/evl3.165. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/evl3.165 (visited on 11/05/2020).

Devi, Khumallambam Devala et al. (Dec. 13, 2013). "An Efficient Protocol for Total DNA Extraction from the Members of Order Zingiberales- Suitable for Diverse PCR Based Downstream Applications". In: *SpringerPlus* 2, p. 669. ISSN: 2193-1801. DOI: 10.1186/2193-1801-2-669. pmid: 24363983. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3867630/ (visited on 04/25/2023).

Dewitte, A. et al. (Dec. 2, 2011). "The Origin of Diversity in Begonia: Genome Dynamism, Population Processes and Phylogenetic Patterns". In: *The Dynamical Processes of Biodiversity - Case Studies of Evolution and Spatial Distribution*. Ed. by Oscar Grillo. InTech. ISBN: 978-953-307-772-7. DOI: 10.5772/23789. URL: http://www.intechopen.com/books/the-dynamical-processes-of-biodiversity-case-studies-of-evolution-and-spatial-distribution/the-origin-of-diversity-in-begonia-genome-dynamism-population-processes-and-phylogenetic-patterns (visited on 01/16/2023).

Durvasula, Arun et al. (May 16, 2017). "African Genomes Illuminate the Early History and Transition to Selfing in Arabidopsis Thaliana". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.20, pp. 5213–5218. ISSN: 1091-6490. DOI: 10.1073/pnas.1616736114. pmid: 28473417.

Edwards, Christine E. et al. (Mar. 11, 2021). "Conservation Genetics of the Threatened Plant Species Physaria Filiformis (Missouri Bladderpod) Reveals Strong Genetic Structure and a Possible Cryptic Species". In: *PLOS ONE* 16.3, e0247586. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0247586. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0247586 (visited on 08/04/2023).

Eggli, Urs and Beat Ernst Leuenberger (1996). "A Quick and Easy Method of Drying Plant Specimens, Including Succulents, for the Herbarium". In: *TAXON* 45.2, pp. 259–261. ISSN: 1996-8175. DOI: 10.2307/1224665. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/1224665 (visited on 09/15/2022).

Eidesen, Pernille Bronken, Inger Greve Alsos, and Christian Brochmann (Aug. 2015). "Comparative Analyses of Plastid and AFLP Data Suggest Different Colonization History and Asymmetric Hybridization between Betula Pubescens and B. Nana". In: *Molecular Ecology* 24.15, pp. 3993–4009. ISSN: 1365-294X. DOI: 10.1111/mec.13289. pmid: 26113148.

Eiserhardt, Wolf L., Thomas L. P. Couvreur, and William J. Baker (June 2017). “Plant Phylogeny as a Window on the Evolution of Hyperdiversity in the Tropical Rainforest Biome”. In: *New Phytologist* 214.4, pp. 1408–1422. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.14516. URL: https://onlinelibrary.wiley.com/doi/10.1111/nph.14516 (visited on 12/01/2022).

Elbaum, Rivka et al. (2009). “New Methods to Isolate Organic Materials from Silicified Phytoliths Reveal Fragmented Glycoproteins but No DNA”. In: *Quaternary International*, p. 9.

Eleanor Green and Camilla Speller (July 13, 2017). “Novel Substrates as Sources of Ancient DNA: Prospects and Hurdles”. In: *Genes* 8.7, p. 180. ISSN: 2073-4425. DOI: 10.3390/genes8070180. URL: http://www.mdpi.com/2073-4425/8/7/180 (visited on 01/27/2021).

Elgar, Mark A. and Danielle Clode (2001). “Inbreeding and Extinction in Island Populations: A Cautionary Note”. In: *Conservation Biology* 15.1, pp. 284–286. ISSN: 0888-8892. JSTOR: 2641670. URL: https://www.jstor.org/stable/2641670 (visited on 10/17/2022).

Epp, L. S. et al. (June 1, 2015). “Lake Sediment Multi-Taxon DNA from North Greenland Records Early Post-Glacial Appearance of Vascular Plants and Accurately Tracks Environmental Changes”. In: *Quaternary Science Reviews* 117, pp. 152–163. ISSN: 0277-3791. DOI: 10.1016/j.quascirev.2015.03.027. URL: https://www.sciencedirect.com/science/article/pii/S0277379115001341 (visited on 11/29/2022).

Eserman, Lauren A. et al. (May 17, 2021). “Target Sequence Capture in Orchids: Developing a Kit to Sequence Hundreds of Single-copy Loci”. In: *Applications in Plant Sciences* 9.7, e11416. ISSN: 2168-0450. DOI: 10.1002/aps3.11416. pmid: 34336404. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8312744/ (visited on 11/28/2022).

Faircloth, Brant C (2015). “PHYLUCE Is a Software Package for the Analysis of Conserved Genomic Loci”. In: p. 3.

Faircloth, Brant C. et al. (Oct. 1, 2012). “Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales”. In: *Systematic Biology* 61.5, pp. 717–726. ISSN: 1076-836X, 1063-5157. DOI: 10.1093/sysbio/sys004. URL: https://academic.oup.com/sysbio/article/61/5/717/1735316 (visited on 10/19/2022).

family=Jong given=TomJ., given-i=TomJ prefix=de useprefix=true, Nickolas M. Waser, and Peter G.L. Klinkhamer (Sept. 1993). “Geitonogamy: The Neglected Side of Selfing”. In: *Trends in Ecology & Evolution* 8.9, pp. 321–325. ISSN: 01695347. DOI: 10.1016/0169-5347(93)90239-L. URL: https://linkinghub.elsevier.com/retrieve/pii/016953479390239L (visited on 11/26/2022).

Fan, Cynthia (2023). “The Genetics of Leaf Shape Variation in Begonia L. (Begoniaceae)”. In.

Fijarczyk, Anna and Wiesław Babik (July 2015). “Detecting Balancing Selection in Genomes: Limits and Prospects”. In: *Molecular Ecology* 24.14, pp. 3529–3545. ISSN: 09621083. DOI: 10.1111/mec.13226. URL: http://doi.wiley.com/10.1111/mec.13226 (visited on 01/17/2021).

Fordyce, Sarah L. et al. (Jan. 11, 2013). “Deep Sequencing of RNA from Ancient Maize Kernels”. In: *PLoS ONE* 8.1. Ed. by Dorian Q. Fuller, e50961. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0050961. URL: https://dx.plos.org/10.1371/journal.pone.0050961 (visited on 11/05/2020).

Forrest, Laura L. et al. (Nov. 26, 2019). “The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques”. In: *Frontiers in Ecology and Evolution* 7, p. 439. ISSN: 2296-701X. DOI: 10.3389/fevo.2019.00439. URL: https://www.frontiersin.org/article/10.3389/fevo.2019.00439/full (visited on 02/11/2022).

Forrest, Laura Lowe (2000). “A Phylogeny of Begoniaceae Bercht. & J. Presl”. PhD thesis. Ann Arbor : ProQuest Dissertations & Theses, 391 pp. URL: https://eleanor.lib.gla.ac.uk/record=b1966036 (visited on 08/13/2023).

Fox, Charles W. and Jason B. Wolf (Apr. 27, 2006). *Evolutionary Genetics: Concepts and Case Studies*. Oxford University Press. 618 pp. ISBN: 978-0-19-977504-0. Google Books: 4UzDyXz7a9IC.

Frankham, Richard (1998). “Inbreeding and Extinction: Island Populations”. In: *Conservation Biology* 12.3, p. 12.

— (2001). “Inbreeding and Extinction in Island Populations: Reply to Elgar and Clode”. In: *Conservation Biology* 15.1, pp. 287–289. ISSN: 0888-8892. JSTOR: 2641671. URL: https://www.jstor.org/stable/2641671 (visited on 10/17/2022).

Fraser, Alex S. (1972). “An Introduction to Population Genetic Theory. By J. F. Crow and M. Kimura. Harper and Row, New York. 656 Pp. 1970”. In: *Teratology* 5.3, pp. 386–387. ISSN: 1096-9926. DOI: 10.1002/tera.1420050318. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/tera.1420050318 (visited on 07/28/2023).

Frodin, David G. (Aug. 2004). “History and Concepts of Big Plant Genera”. In: *TAXON* 53.3, pp. 753–776. ISSN: 0040-0262, 1996-8175. DOI: 10.2307/4135449. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/4135449 (visited on 01/29/2021).

Funk, Vicki (2003). *100 Uses for an Herbarium (Well at Least 72)*. American Society of Plant Taxonomists Newsletter.

Fér, Tomáš and Roswitha E Schmickl (Jan. 2018). “HybPhyloMaker: Target Enrichment Data Analysis From Raw Reads to Species Trees”. In: *Evolutionary Bioinformatics* 14, p. 117693431774261. ISSN: 1176-9343, 1176-9343. DOI: 10.1177/1176934317742613. URL: http://journals.sagepub.com/doi/10.1177/1176934317742613 (visited on 10/18/2022).

Gawel, N. J. and R. L. Jarret (Aug. 1, 1991). “A Modified CTAB DNA Extraction Procedure forMusa andIpomoea”. In: *Plant Molecular Biology Reporter* 9.3, pp. 262–266. ISSN: 1572-9818. DOI: 10.1007/BF02672076. URL: https://doi.org/10.1007/BF02672076 (visited on 04/25/2023).

GBIF (2023). URL: https://www.gbif.org/ (visited on 04/04/2023).

Genereux, Diane P. et al. (Nov. 2020). “A Comparative Genomics Multitool for Scientific Discovery and Conservation”. In: *Nature* 587.7833 (7833), pp. 240–245. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2876-6. URL: https://www.nature.com/articles/s41586-020-2876-6 (visited on 07/19/2023).

Giguet-Covex, Charline et al. (Feb. 3, 2014). “Long Livestock Farming History and Human Landscape Shaping Revealed by Lake Sediment DNA”. In: *Nature Communications* 5.1, p. 3211. ISSN: 2041-1723. DOI: 10.1038/ncomms4211. URL: https://www.nature.com/articles/ncomms4211 (visited on 10/10/2022).

Gilbert, M. Thomas P. et al. (May 9, 2008). “DNA from Pre-Clovis Human Coprolites in Oregon, North America”. In: *Science* 320.5877, pp. 786–789. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1154116. URL: https://www.science.org/doi/10.1126/science.1154116 (visited on 10/10/2022).

Gillespie, John H. (Aug. 6, 2004). *Population Genetics: A Concise Guide*. JHU Press. 240 pp. ISBN: 978-1-4214-0170-6. Google Books: KAcAfiyHpcoC.

Ginolhac, Aurelien et al. (Aug. 1, 2011). “mapDamage: Testing for Damage Patterns in Ancient DNA Sequences”. In: *Bioinformatics (Oxford, England)* 27.15, pp. 2153–2155. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btr347. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr347 (visited on 10/20/2022).

Global Plants on JSTOR (2023). URL: https://plants.jstor.org/ (visited on 04/04/2023).

Glover, Beverley J. and Richard J. Abbott (Jan. 1995). “Low Genetic Diversity in the Scottish Endemic *Primula Scotica* Hook.” In: *New Phytologist* 129.1, pp. 147–153. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/j.1469-8137.1995.tb03018.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1469-8137.1995.tb03018.x (visited on 11/25/2022).

Gnirke, Andreas et al. (Feb. 2009). “Solution Hybrid Selection with Ultra-Long Oligonucleotides for Massively Parallel Targeted Sequencing”. In: *Nature Biotechnology* 27.2, pp. 182–189. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.1523. URL: http://www.nature.com/articles/nbt.1523 (visited on 11/06/2020).

Gralka, Matti, Diana Fusco, and Oskar Hallatschek (July 2016). “Watching Populations Melt Down”. In: *Biophysical Journal* 111.2, pp. 271–272. ISSN: 00063495. DOI: 10.1016/j.bpj.2016.06.020. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006349516304660 (visited on 12/04/2022).

Grass, Robert N. et al. (Feb. 16, 2015). “Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes”. In: *Angewandte Chemie International Edition* 54.8, pp. 2552–2555. ISSN: 14337851. DOI: 10.1002/anie.201411378. URL: http://doi.wiley.com/10.1002/anie.201411378 (visited on 11/05/2020).

Greeff, JacoM and Bettine vanVuuren (Apr. 1, 2003). “Introduction to Conservation Genetics”. In: *African Zoology* 38.1, pp. 192–192. ISSN: 1562-7020. DOI: 10.1080/15627020.2003.11657212. URL: https://doi.org/10.1080/15627020.2003.11657212 (visited on 10/15/2022).

Gutaker, Rafal M and Hernán A Burbano (Apr. 2017). “Reinforcing Plant Evolutionary Genomics Using Ancient DNA”. In: *Current Opinion in Plant Biology* 36, pp. 38–45. ISSN: 13695266. DOI: 10.1016/j.pbi.2017.01.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S136952661630156X (visited on 05/02/2022).

Gutaker, Rafal M. et al. (Sept. 20, 2016). *Extraction of Ultrashort DNA Molecules from Herbarium Specimens*. preprint. Genomics. DOI: 10.1101/076299. URL: http://biorxiv.org/lookup/doi/10.1101/076299 (visited on 11/30/2022).

Gutaker, Rafal M. et al. (July 2019). "The Origins and Adaptation of European Potatoes Reconstructed from Historical Genomes". In: *Nature Ecology & Evolution* 3.7, pp. 1093–1101. ISSN: 2397-334X. DOI: 10.1038/s41559-019-0921-3. URL: http://www.nature.com/articles/s41559-019-0921-3 (visited on 08/13/2021).

Hall, A. V. (Nov. 1988). "PEST CONTROL IN HERBARIA". In: *TAXON* 37.4, pp. 885–907. ISSN: 0040-0262, 1996-8175. DOI: 10.2307/1222094. URL: https://onlinelibrary.wiley.com/doi/abs/10.2307/1222094 (visited on 09/27/2022).

Hallam, N. D., B. E. Roberts, and D. J. Osborne (Dec. 1972). "Embryogenesis and Germination in Rye (Secale Cereale L.) : II. Biochemical and Fine Structural Changes during Germination". In: *Planta* 105.4, pp. 293–309. ISSN: 0032-0935. DOI: 10.1007/BF00386767. pmid: 24477844.

Harris, Bob (n.d.). *LASTZ - Pairwise DNA Sequence Aligner*. URL: https://github.com/lastz/lastz (visited on 11/29/2022).

Hart, Michelle L. et al. (Oct. 26, 2016). "Retrieval of Hundreds of Nuclear Loci from Herbarium Specimens". In: *Taxon* 65.5, pp. 1081–1092. ISSN: 00400262. DOI: 10.12705/655.9. URL: http://doi.wiley.com/10.12705/655.9 (visited on 08/03/2021).

Havermans, John, René Dekker, and Ron Sportel (Dec. 2015). "The Effect of Mercuric Chloride Treatment as Biocide for Herbaria on the Indoor Air Quality". In: *Heritage Science* 3.1, p. 39. ISSN: 2050-7445. DOI: 10.1186/s40494-015-0068-8. URL: http://www.heritagesciencejournal.com/content/3/1/39 (visited on 09/26/2022).

Henn, Brenna M. et al. (June 2015). "Estimating the Mutation Load in Human Genomes". In: *Nature Reviews Genetics* 16.6, pp. 333–343. ISSN: 1471-0056, 1471-0064. DOI: 10.1038/nrg3931. URL: http://www.nature.com/articles/nrg3931 (visited on 01/10/2021).

Higuchi, Russell et al. (Nov. 1984). "DNA Sequences from the Quagga, an Extinct Member of the Horse Family". In: *Nature* 312.5991, pp. 282–284. ISSN: 1476-4687. DOI: 10.1038/312282a0. URL: https://www.nature.com/articles/312282a0 (visited on 10/09/2022).

Hofman, Courtney A. et al. (Sept. 2015). "Conservation Archaeogenomics: Ancient DNA and Biodiversity in the Anthropocene". In: *Trends in Ecology & Evolution* 30.9, pp. 540–549. ISSN: 01695347. DOI: 10.1016/j.tree.2015.06.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534715001597 (visited on 12/16/2020).

Hofreiter, M. et al. (May 2001). "Ancient DNA". In: *Nature Reviews. Genetics* 2.5, pp. 353–359. ISSN: 1471-0056. DOI: 10.1038/35072071. pmid: 11331901.

Huang, S. et al. (Oct. 2020). "Genetic and Morphologic Determination of Diatom Community Composition in Surface Sediments from Glacial and Thermokarst Lakes in the Siberian Arctic". In: *Journal of Paleolimnology* 64.3, pp. 225–242.

ISSN: 0921-2728, 1573-0417. DOI: 10.1007/s10933-020-00133-1. URL: https://link.springer.com/10.1007/s10933-020-00133-1 (visited on 05/27/2022).

Hughes, M. and P. M. Hollingsworth (June 2008). "Population Genetic Divergence Corresponds with Species-Level Biodiversity Patterns in the Large Genus *Begonia*: POPULATION DIFFERENTIATION IN *BEGONIA*". In: *Molecular Ecology* 17.11, pp. 2643–2651. ISSN: 09621083. DOI: 10.1111/j.1365-294X.2008.03788.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-294X.2008.03788.x (visited on 10/14/2022).

Hughes, M., P. M. Hollingsworth, and A. G. Miller (Oct. 1, 2003). "Population Genetic Structure in the Endemic Begonia of the Socotra Archipelago". In: *Biological Conservation* 113.2, pp. 277–284. ISSN: 0006-3207. DOI: 10.1016/S0006-3207(02)00375-0. URL: https://www.sciencedirect.com/science/article/pii/S0006320702003750 (visited on 03/21/2023).

Hughes, M., P. M. Hollingsworth, and J. Squirrell (2002a). "Isolation of Polymorphic Microsatellite Markers for Begonia Sutherlandii Hook. f." In: *Molecular Ecology Notes* 2.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1471-8286.2002.00201.x
_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1471-8286.2002.00201.x, pp. 185–186. ISSN: 1471-8286. DOI: 10.1046/j.1471-8286.2002.00201.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1471-8286.2002.00201.x (visited on 11/25/2022).

Hughes, M. and A. G. Miller (June 25, 2002b). "A NEW ENDEMIC SPECIES OF BEGONIA (BEGONIACEAE) FROM THE SOCOTRA ARCHIPELAGO". In: *Edinburgh Journal of Botany* 59.2, pp. 273–281. ISSN: 1474-0036. DOI: 10.1017/S0960428602000082. URL: https://journals.rbge.org.uk/ejb/article/view/1082 (visited on 11/25/2022).

Hughes, M., J. Russell, and P. M. Hollingsworth (June 2002c). "Polymorphic Microsatellite Markers for the Socotran Endemic Herb Begonia Socotrana". In: *Molecular Ecology Notes* 2.2, pp. 159–160. ISSN: 1471-8278, 1471-8286. DOI: 10.1046/j.1471-8286.2002.00185.x. URL: http://doi.wiley.com/10.1046/j.1471-8286.2002.00185.x (visited on 11/21/2022).

Hughes, Mark and Wayne Takeuchi (Feb. 4, 2015). "A New Section (Begonia Sect. Oligandrae Sect. Nov.) and a New Species (Begonia Pentandra Sp. Nov.) in Begoniaceae from New Guinea". In: *Phytotaxa* 197.1, p. 37. ISSN: 1179-3163, 1179-3155. DOI: 10.11646/phytotaxa.197.1.4. URL: https://biotaxa.org/Phytotaxa/article/view/phytotaxa.197.1.4 (visited on 11/10/2022).

Humbert (1972). "Begonia Antaisaka". In: *Bulletin du Museum National d'Histoire Naturelle* 76, p. 3.

Humphreys, Aelys M. et al. (July 2019). "Global Dataset Shows Geography and Life Form Predict Modern Plant Extinction and Rediscovery". In: *Nature Ecology & Evolution* 3.7 (7), pp. 1043–1047. ISSN: 2397-334X. DOI: 10.1038/s41559-019-0906-2. URL: https://www.nature.com/articles/s41559-019-0906-2 (visited on 11/28/2022).

Ibrahim, Anan et al. (July 2021). "Anthropogenic Impact on the Historical Phytoplankton Community of Lake Constance Reconstructed by Multimarker

Analysis of Sediment-Core Environmental DNA". In: *Molecular Ecology* 30.13, pp. 3040–3056. ISSN: 0962-1083, 1365-294X. DOI: 10.1111/mec.15696.

iDigBio Home | iDigBio (2023). URL: https://www.idigbio.org/ (visited on 04/04/2023).

Jackson, Chris, Todd McLay, and Alexander N. Schmidt-Lebuhn (Nov. 10, 2021). *Hybpiper-Rbgv and Yang-and-Smith-Rbgv: Containerization and Additional Options for Assembly and Paralog Detection in Target Enrichment Data.* preprint. Bioinformatics. DOI: 10.1101/2021.11.08.467817. URL: http://biorxiv.org/lookup/doi/10.1101/2021.11.08.467817 (visited on 11/29/2022).

Jaenicke-Despres, V. (Nov. 14, 2003). "Early Allelic Selection in Maize as Revealed by Ancient DNA". In: *Science (New York, N.Y.)* 302.5648, pp. 1206–1208. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1089056. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.1089056 (visited on 01/22/2021).

Johnson, Matthew G. et al. (July 2016). "HybPiper: Extracting Coding Sequence and Introns for Phylogenetics from High-throughput Sequencing Reads Using Target Enrichment". In: *Applications in Plant Sciences* 4.7, p. 1600016. ISSN: 2168-0450, 2168-0450. DOI: 10.3732/apps.1600016. URL: https://onlinelibrary.wiley.com/doi/10.3732/apps.1600016 (visited on 08/15/2021).

Johnson, Matthew G et al. (July 1, 2019). "A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using K-Medoids Clustering". In: *Systematic Biology* 68.4, pp. 594–606. ISSN: 1063-5157. DOI: 10.1093/sysbio/syy086. URL: https://doi.org/10.1093/sysbio/syy086 (visited on 05/16/2022).

Jónsson, Hákon et al. (July 2013). "mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters". In: *Bioinformatics (Oxford, England)* 29.13, pp. 1682–1684. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btt193. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt193 (visited on 11/05/2020).

Kalyaanamoorthy, Subha et al. (June 2017). "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates". In: *Nature Methods* 14.6, pp. 587–589. ISSN: 1548-7105. DOI: 10.1038/nmeth.4285. pmid: 28481363.

Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte (Feb. 21, 2017). "Dynamics of Genome Size Evolution in Birds and Mammals". In: *Proceedings of the National Academy of Sciences* 114.8, E1460–E1469. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1616702114. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1616702114 (visited on 11/05/2020).

Kates, Heather R. et al. (June 2021). "The Effects of Herbarium Specimen Characteristics on Short-Read NGS Sequencing Success in Nearly 8000 Specimens: Old, Degraded Samples Have Lower DNA Yields but Consistent Sequencing Success". In: *Frontiers in Plant Science* 12, p. 669064. ISSN: 1664-462X. DOI: 10.3389/fpls.2021.669064.

Kim, Su Yeon et al. (June 11, 2011). "Estimation of Allele Frequency and Association Mapping Using Next-Generation Sequencing Data". In: *BMC Bioinformatics* 12.1, p. 231. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-231. URL: https://doi.org/10.1186/1471-2105-12-231 (visited on 10/21/2022).

Kirkpatrick, John B., Emily A. Walsh, and Steven D'Hondt (Aug. 2016). "Fossil DNA Persistence and Decay in Marine Sediment over Hundred-Thousand-Year to Million-Year Time Scales". In: *Geology* 44.8, pp. 615–618. ISSN: 0091-7613. DOI: 10.1130/G37933.1.

Kistler, L. et al. (Feb. 25, 2014). "Transoceanic Drift and the Domestication of African Bottle Gourds in the Americas". In: *Proceedings of the National Academy of Sciences* 111.8, pp. 2937–2941. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1318678111. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1318678111 (visited on 11/05/2020).

Kistler, Logan and Beth Shapiro (Dec. 2011). "Ancient DNA Confirms a Local Origin of Domesticated Chenopod in Eastern North America". In: *Journal of Archaeological Science* 38.12, pp. 3549–3554. ISSN: 03054403. DOI: 10.1016/j.jas.2011.08.023. URL: https://linkinghub.elsevier.com/retrieve/pii/S0305440311003025 (visited on 01/23/2021).

Kistler, Logan et al. (Dec. 8, 2015). "Gourds and Squashes (*Cucurbita* Spp.) Adapted to Megafaunal Extinction and Ecological Anachronism through Domestication". In: *Proceedings of the National Academy of Sciences* 112.49, pp. 15107–15112. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1516109112. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1516109112 (visited on 11/05/2020).

Kistler, Logan et al. (Dec. 14, 2018). "Multiproxy Evidence Highlights a Complex Evolutionary Legacy of Maize in South America". In: *Science (New York, N.Y.)* 362.6420, pp. 1309–1313. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aav0207. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aav0207 (visited on 11/05/2020).

Kistler, Logan et al. (Apr. 29, 2020). "Ancient Plant Genomics in Archaeology, Herbaria, and the Environment". In: *Annual Review of Plant Biology* 71.1, pp. 605–629. ISSN: 1543-5008, 1545-2123. DOI: 10.1146/annurev-arplant-081519-035837. URL: https://www.annualreviews.org/doi/10.1146/annurev-arplant-081519-035837 (visited on 11/05/2020).

Koenen, Erik J. M. et al. (2015). "Recently Evolved Diversity and Convergent Radiations of Rainforest Mahoganies (Meliaceae) Shed New Light on the Origins of Rainforest Hyperdiversity". In: *New Phytologist* 207.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.13490, pp. 327–339. ISSN: 1469-8137. DOI: 10.1111/nph.13490. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/nph.13490 (visited on 11/09/2022).

Kopperud, Cornelia and John W. Einset (June 1995). "DNA Isolation from Begonia Leaves". In: *Plant Molecular Biology Reporter* 13.2, pp. 129–130. ISSN: 0735-9640, 1572-9818. DOI: 10.1007/BF02668783. URL: http://link.springer.com/10.1007/BF02668783 (visited on 10/13/2022).

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen (Dec. 2014). "ANGSD: Analysis of Next Generation Sequencing Data". In: *BMC Bioinformatics* 15.1, p. 356. ISSN: 1471-2105. DOI: 10.1186/s12859-014-0356-4. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-014-0356-4 (visited on 11/06/2020).

Kozlov, Alexey M., Andre J. Aberer, and Alexandros Stamatakis (Aug. 1, 2015). "ExaML Version 3: A Tool for Phylogenomic Analyses on Supercomputers". In: *Bioinformatics* 31.15, pp. 2577–2579. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv184. URL: https://doi.org/10.1093/bioinformatics/btv184 (visited on 03/28/2023).

Kramer, Andrea T. and Kayri Havens (Nov. 2009). "Plant Conservation Genetics in a Changing World". In: *Trends in Plant Science* 14.11, pp. 599–607. ISSN: 13601385. DOI: 10.1016/j.tplants.2009.08.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S1360138509002040 (visited on 02/09/2023).

Kreft, Holger et al. (Feb. 2008). "Global Diversity of Island Floras from a Macroecological Perspective". In: *Ecology Letters* 11.2, pp. 116–127. ISSN: 1461-0248. DOI: 10.1111/j.1461-0248.2007.01129.x. pmid: 18036182.

Král, Kamil and Jindřich Pavliš (Aug. 10, 2006). "The First Detailed Land-cover Map of Socotra Island by Landsat/ETM+ Data". In: *International Journal of Remote Sensing* 27.15, pp. 3239–3250. ISSN: 0143-1161. DOI: 10.1080/01431160600646557. URL: https://doi.org/10.1080/01431160600646557 (visited on 08/06/2023).

Kumar, Satish et al. (Jan. 5, 2021). "Homozygosity Mapping Reveals Population History and Trait Architecture in Self-Incompatible Pear (Pyrus Spp.)" In: *Frontiers in Plant Science* 11, p. 590846. ISSN: 1664-462X. DOI: 10.3389/fpls.2020.590846. URL: https://www.frontiersin.org/articles/10.3389/fpls.2020.590846/full (visited on 05/05/2022).

Lammers, Youri, Peter D. Heintzman, and Inger Greve Alsos (Feb. 16, 2021). "Environmental Palaeogenomic Reconstruction of an Ice Age Algal Population". In: *Communications Biology* 4.1, pp. 1–11. ISSN: 2399-3642. DOI: 10.1038/s42003-021-01710-4. URL: https://www.nature.com/articles/s42003-021-01710-4 (visited on 11/29/2022).

Lang, Patricia L. M. et al. (Jan. 2019). "Using Herbaria to Study Global Environmental Change". In: *New Phytologist* 221.1, pp. 110–122. ISSN: 0028646X. DOI: 10.1111/nph.15401. URL: https://onlinelibrary.wiley.com/doi/10.1111/nph.15401 (visited on 09/09/2022).

Larridon, Isabel et al. (Jan. 9, 2020). "Tackling Rapid Radiations With Targeted Sequencing". In: *Frontiers in Plant Science* 10, p. 1655. ISSN: 1664-462X. DOI: 10.3389/fpls.2019.01655. URL: https://www.frontiersin.org/article/10.3389/fpls.2019.01655/full (visited on 05/16/2022).

Larson, G. et al. (Apr. 2014). "Current Perspectives and the Future of Domestication Studies". In: *Proceedings of the National Academy of Sciences* 111.17, pp. 6139–6146. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1323964111.

Lavrentovich, Maxim O. et al. (June 21, 2016). "Spatially Constrained Growth Enhances Conversional Meltdown". In: *Biophysical Journal* 110.12, pp. 2800–2808. ISSN: 0006-3495. DOI: 10.1016/j.bpj.2016.05.024. URL: https://www.sciencedirect.com/science/article/pii/S0006349516303332 (visited on 08/08/2023).

Lemmon, Emily Moriarty and Alan R. Lemmon (Nov. 23, 2013). "High-Throughput Genomic Data in Systematics and Phylogenetics". In: *Annual Review of Ecology, Evolution, and Systematics* 44.1, pp. 99–121. ISSN:

1543-592X, 1545-2069. DOI: 10.1146/annurev-ecolsys-110512-135822. URL: https://www.annualreviews.org/doi/10.1146/annurev-ecolsys-110512-135822 (visited on 11/28/2022).

Lencz, Todd et al. (Dec. 11, 2007). “Runs of Homozygosity Reveal Highly Penetrant Recessive Loci in Schizophrenia”. In: *Proceedings of the National Academy of Sciences* 104.50, pp. 19942–19947. DOI: 10.1073/pnas.0710021104. URL: https://www.pnas.org/doi/full/10.1073/pnas.0710021104 (visited on 10/27/2022).

Li, Heng and Richard Durbin (July 15, 2009a). “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform”. In: *Bioinformatics (Oxford, England)* 25.14, pp. 1754–1760. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp324. pmid: 19451168.

— (Mar. 1, 2010). “Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform”. In: *Bioinformatics (Oxford, England)* 26.5, pp. 589–595. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp698. pmid: 20080505.

Li, Heng et al. (Aug. 15, 2009b). “The Sequence Alignment/Map Format and SAMtools”. In: *Bioinformatics (Oxford, England)* 25.16, pp. 2078–2079. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp352. pmid: 19505943.

Li, Lingfei et al. (Apr. 2022). “Genomes Shed Light on the Evolution of *Begonia* , a Mega-Diverse Genus”. In: *New Phytologist* 234.1, pp. 295–310. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.17949.

Liu, Sisi et al. (May 20, 2021). “Sedimentary Ancient DNA Reveals a Threat of Warming-Induced Alpine Habitat Loss to Tibetan Plateau Plant Diversity”. In: *Nature Communications* 12.1, p. 2995. ISSN: 2041-1723. DOI: 10.1038/s41467-021-22986-4. URL: https://www.nature.com/articles/s41467-021-22986-4 (visited on 10/10/2022).

Livio, M. and A. Kopelman (Jan. 1990). “Life and the Sun’s Lifetime”. In: *Nature* 343.6253, pp. 25–25. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/343025a0. URL: http://www.nature.com/articles/343025a0 (visited on 08/03/2022).

Lynch, M. et al. (Sept. 1, 1993). “The Mutational Meltdown in Asexual Populations”. In: *Journal of Heredity* 84.5, pp. 339–344. ISSN: 0022-1503. DOI: 10.1093/oxfordjournals.jhered.a111354. URL: https://doi.org/10.1093/oxfordjournals.jhered.a111354 (visited on 08/08/2023).

Lynch, Michael, John Conery, and Reinhard Burger (Dec. 1995). “Mutational Meltdowns in Sexual Populations”. In: *Evolution* 49.6, p. 1067. ISSN: 00143820. DOI: 10.2307/2410432. JSTOR: 2410432. URL: https://www.jstor.org/stable/2410432?origin=crossref (visited on 12/04/2022).

Malaspinas, Anna-Sapfo (Jan. 2016). “Methods to Characterize Selective Sweeps Using Time Serial Samples: An Ancient DNA Perspective”. In: *Molecular Ecology* 25.1, pp. 24–41. ISSN: 09621083. DOI: 10.1111/mec.13492. URL: https://onlinelibrary.wiley.com/doi/10.1111/mec.13492 (visited on 09/18/2022).

Malaspinas, Anna-Sapfo et al. (Oct. 15, 2014). “Bammds: A Tool for Assessing the Ancestry of Low-Depth Whole-Genome Data Using Multidimensional Scaling (MDS)”. In: *Bioinformatics (Oxford, England)* 30.20, pp. 2962–2964. ISSN: 1460-2059, 1367-4803. DOI: 10.1093/bioinformatics/btu410. URL: https:

//academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu410 (visited on 02/08/2021).

Mascher, Martin et al. (Sept. 2016). “Genomic Analysis of 6,000-Year-Old Cultivated Grain Illuminates the Domestication History of Barley”. In: *Nature Genetics* 48.9, pp. 1089–1093. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3611. URL: http://www.nature.com/articles/ng.3611 (visited on 01/23/2021).

Matolweni, Luzuko Orlyn, Kevin Balkwill, and Tracy McLellan (Mar. 2000). “Genetic Diversity and Gene Flow in the Morphologically Variable, Rare Endemics *Begonia Dregei* and *Begonia Homonyma* (Begoniaceae)”. In: *American Journal of Botany* 87.3, pp. 431–439. ISSN: 00029122. DOI: 10.2307/2656639. URL: http://doi.wiley.com/10.2307/2656639 (visited on 04/10/2023).

McKain, Michael R. et al. (2018). “Practical Considerations for Plant Phylogenomics”. In: *Applications in Plant Sciences* 6.3, e1038. ISSN: 2168-0450. DOI: 10.1002/aps3.1038. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/aps3.1038 (visited on 07/16/2023).

McKenna, Aaron et al. (Sept. 2010). “The Genome Analysis Toolkit: A MapReduce Framework for Analyzing next-Generation DNA Sequencing Data”. In: *Genome Research* 20.9, pp. 1297–1303. ISSN: 1549-5469. DOI: 10.1101/gr.107524.110. pmid: 20644199.

McKinney, Garrett J. et al. (July 2017). “Paralogs Are Revealed by Proportion of Heterozygotes and Deviations in Read Ratios in Genotyping-by-Sequencing Data from Natural Populations”. In: *Molecular Ecology Resources* 17.4, pp. 656–669. ISSN: 1755098X. DOI: 10.1111/1755-0998.12613. URL: http://doi.wiley.com/10.1111/1755-0998.12613 (visited on 03/09/2021).

Meineke, Emily K. et al. (Jan. 7, 2019). “Biological Collections for Understanding Biodiversity in the Anthropocene”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1763, p. 20170386. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2017.0386. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0386 (visited on 05/07/2022).

Meisner, Jonas and Anders Albrechtsen (Oct. 1, 2018). “Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data”. In: *Genetics* 210.2, pp. 719–731. ISSN: 1943-2631. DOI: 10.1534/genetics.118.301336. URL: https://doi.org/10.1534/genetics.118.301336 (visited on 08/02/2023).

Meyermans, R. et al. (Dec. 2020). “How to Study Runs of Homozygosity Using PLINK? A Guide for Analyzing Medium Density SNP Data in Livestock and Pet Species”. In: *BMC Genomics* 21.1, p. 94. ISSN: 1471-2164. DOI: 10.1186/s12864-020-6463-x. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6463-x (visited on 04/10/2023).

Michel, Thibauld et al. (Aug. 18, 2022). “A Hybrid Capture Bait Set for Begonia”. In: ISSN: 1474-0036. DOI: 10.24823/ejb.2022.409. URL: https://research-scotland.ac.uk/handle/20.500.12594/20011 (visited on 10/22/2022).

Minh, Bui Quang et al. (May 1, 2020). “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era”. In: *Molecular Biology and Evolution* 37.5, pp. 1530–1534. ISSN: 1537-1719. DOI: 10.1093/molbev/msaa015. pmid: 32011700.

Mitchell, Kieren J. and Nicolas J. Rawlence (Mar. 2021). “Examining Natural History through the Lens of Palaeogenomics”. In: *Trends in Ecology & Evolution* 36.3, pp. 258–267. ISSN: 01695347. DOI: 10.1016/j.tree.2020.10.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0169534720302822 (visited on 05/27/2022).

Modi, Alessandra et al. (Dec. 2021). “Successful Extraction of Insect DNA from Recent Copal Inclusions: Limits and Perspectives”. In: *Scientific Reports* 11.1, p. 6851. ISSN: 2045-2322. DOI: 10.1038/s41598-021-86058-9. URL: http://www.nature.com/articles/s41598-021-86058-9 (visited on 05/27/2022).

Molecular Ecology, 3rd Edition | Wiley (2020). Wiley.com. URL: https://www.wiley.com/en-gb/Molecular+Ecology%2C+3rd+Edition-p-9781119426158 (visited on 03/13/2023).

Monchamp, Marie-Eve, Piet Spaak, and Francesco Pomati (May 14, 2019). “High Dispersal Levels and Lake Warming Are Emergent Drivers of Cyanobacterial Community Assembly in Peri-Alpine Lakes”. In: *Scientific Reports* 9.1, p. 7366. ISSN: 2045-2322. DOI: 10.1038/s41598-019-43814-2. URL: https://www.nature.com/articles/s41598-019-43814-2 (visited on 10/09/2022).

Moonlight, Peter W. et al. (Apr. 2018). “Dividing and Conquering the Fastest–Growing Genus: Towards a Natural Sectional Classification of the Mega–Diverse Genus *Begonia* (Begoniaceae)”. In: *TAXON* 67.2, pp. 267–323. ISSN: 0040-0262, 1996-8175. DOI: 10.12705/672.3. URL: https://onlinelibrary.wiley.com/doi/10.12705/672.3 (visited on 11/25/2022).

Moritz, Craig (1999). “Conservation Units and Translocations: Strategies for Conserving Evolutionary Processes”. In: *Hereditas* 130.3, pp. 217–228. ISSN: 1601-5223. DOI: 10.1111/j.1601-5223.1999.00217.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1601-5223.1999.00217.x (visited on 08/07/2023).

Mouttham, Nathalie et al. (July 2015). “Surveying the Repair of Ancient DNA from Bones via High-Throughput Sequencing”. In: *BioTechniques* 59.1, pp. 19–25. ISSN: 0736-6205. DOI: 10.2144/000114307. URL: https://www.future-science.com/doi/10.2144/000114307 (visited on 10/20/2022).

Murray, Gemma G. R. et al. (Nov. 2017). “Natural Selection Shaped the Rise and Fall of Passenger Pigeon Genomic Diversity”. In: *Science (New York, N.Y.)* 358.6365, pp. 951–954. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aao0960.

Neale, S, W Goodall-Copestake, and C A Kidner (2006). “The Evolution of Diversity in Begonia”. In: p. 7.

NEB (2022). *NEBNext® Multiplex Oligos for Illumina® (96 Unique Dual Index Primer Pairs Set 2) NEB #E6442S/L*. New England Biolabs. URL: https://international.neb.com/-/media/nebus/files/manuals/manuale6442.pdf?rev=9905fdcd8c4c476189db84137a950910&hash=8BF40719EB37F5AA930CC59636D379F0.

Neukamm, Judith, Alexander Peltzer, and Kay Nieselt (Oct. 2021). “DamageProfiler: Fast Damage Pattern Calculation for Ancient DNA”. In: *Bioinformatics (Oxford, England)* 37.20. Ed. by Janet Kelso, pp. 3652–3653. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btab190.

Nguyen, Lam-Tung et al. (Jan. 2015). “IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies”. In: *Molecular Biology and Evolution* 32.1, pp. 268–274. ISSN: 1537-1719, 0737-4038. DOI: 10.1093/molbev/msu300. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msu300 (visited on 04/10/2023).

Nguyen, Nam, Siavash Mirarab, and Tandy Warnow (Jan. 26, 2012). “MRL and SuperFine+MRL: New Supertree Methods”. In: *Algorithms for Molecular Biology* 7.1, p. 3. ISSN: 1748-7188. DOI: 10.1186/1748-7188-7-3. URL: https://doi.org/10.1186/1748-7188-7-3 (visited on 03/28/2023).

Nho, Kwangsik et al. (Oct. 2014). “Comparison of Multi-Sample Variant Calling Methods for Whole Genome Sequencing”. In: *IEEE International Conference on Systems Biology : [proceedings]. IEEE International Conference on Systems Biology* 2014, pp. 59–62. ISSN: 2325-0704. DOI: 10.1109/ISB.2014.6990432. pmid: 26167514. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4496949/ (visited on 11/29/2022).

Nic Lughadha, Eimear et al. (Jan. 7, 2019). “The Use and Misuse of Herbarium Specimens in Evaluating Plant Extinction Risks”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1763, p. 20170402. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2017.0402. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0402 (visited on 09/09/2022).

Nicholls, James A. et al. (Sept. 17, 2015). “Using Targeted Enrichment of Nuclear Genes to Increase Phylogenetic Resolution in the Neotropical Rain Forest Genus Inga (Leguminosae: Mimosoideae)”. In: *Frontiers in Plant Science* 6. ISSN: 1664-462X. DOI: 10.3389/fpls.2015.00710. URL: http://journal.frontiersin.org/Article/10.3389/fpls.2015.00710/abstract (visited on 12/16/2020).

Nishii, Kanae et al. (Apr. 2022). “The First Genome for the Cape Primrose Streptocarpus Rexii (Gesneriaceae), a Model Plant for Studying Meristem-Driven Shoot Diversity”. In: *Plant Direct* 6.4, e388. ISSN: 2475-4455. DOI: 10.1002/pld3.388. pmid: 35388373.

Nistelberger, H. M. et al. (Dec. 2016). “The Efficacy of High-Throughput Sequencing and Target Enrichment on Charred Archaeobotanical Remains”. In: *Scientific Reports* 6.1, p. 37347. ISSN: 2045-2322. DOI: 10.1038/srep37347. URL: http://www.nature.com/articles/srep37347 (visited on 11/05/2020).

Nordling, Linda (Nov. 2022). “Seeding an Anti-Racist Culture at Scotland’s Botanical Gardens”. In: *Nature* 611.7937, pp. 835–838. ISSN: 1476-4687. DOI: 10.1038/d41586-022-03797-z. pmid: 36414779.

O’Connor, J. P. (Oct. 1, 1991). “D. PINNIGER. Insect Pests in Museums. Institute of Archaeology Publications, London: 1989. Pp Ii, 45; Illustrated. Price: None Stated. ISBN 0-905853-25-3.” In: *Archives of Natural History* 18.3, pp. 423–424. ISSN: 0260-9541. DOI: 10.3366/anh.1991.18.3.423b. URL: https://www.euppublishing.com/doi/abs/10.3366/anh.1991.18.3.423b (visited on 02/23/2023).

O’Connor BD, Van der Auwera GA (2020). *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition)*. O’Reilly Media.

O’Leary, Shannon J. et al. (2018). “These Aren’t the Loci You’e Looking for: Principles of Effective SNP Filtering for Molecular Ecologists”. In: *Molecular*

Ecology 27.16, pp. 3193–3206. ISSN: 1365-294X. DOI: 10.1111/mec.14792. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14792 (visited on 11/29/2022).

Ollivier, Louis and Jean-Louis Foulley (Dec. 1, 2013). "A Note on the Partitioning of Allelic Diversity". In: *Conservation Genetics* 14.6, pp. 1285–1290. ISSN: 1572-9737. DOI: 10.1007/s10592-013-0508-5. URL: https://doi.org/10.1007/s10592-013-0508-5 (visited on 08/08/2023).

Orlando, Ludovic and Alan Cooper (Nov. 23, 2014). "Using Ancient DNA to Understand Evolutionary and Ecological Processes". In: *Annual Review of Ecology, Evolution, and Systematics* 45.1, pp. 573–598. ISSN: 1543-592X, 1545-2069. DOI: 10.1146/annurev-ecolsys-120213-091712. URL: http://www.annualreviews.org/doi/10.1146/annurev-ecolsys-120213-091712 (visited on 02/07/2021).

Ortiz, Edgardo M. (Jan. 2019). *Vcf2phylip v2.0: Convert a VCF Matrix into Several Matrix Formats for Phylogenetic Analysis.* Zenodo. DOI: 10.5281/zenodo.2540861.

Otto-Bliesner, Bette L. et al. (Oct. 28, 2013). "How Warm Was the Last Interglacial? New Model–Data Comparisons". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371.2001, p. 20130097. DOI: 10.1098/rsta.2013.0097. URL: https://royalsocietypublishing.org/doi/10.1098/rsta.2013.0097 (visited on 10/10/2022).

Paer, Céline Van de et al. (Dec. 2016). "Mitogenomics of Hesperelaea, an Extinct Genus of Oleaceae". In: *Gene* 594.2, pp. 197–202. ISSN: 03781119. DOI: 10.1016/j.gene.2016.09.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378111916307223 (visited on 11/05/2020).

Palmer, S. A. et al. (Aug. 1, 2012). "Archaeogenomic Evidence of Punctuated Genome Evolution in Gossypium". In: *Molecular Biology and Evolution* 29.8, pp. 2031–2038. ISSN: 0737-4038, 1537-1719. DOI: 10.1093/molbev/mss070. URL: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss070 (visited on 01/22/2021).

Parducci, L. et al. (Aug. 2005). "Ancient DNA from Pollen: A Genetic Record of Population History in Scots Pine". In: *Molecular Ecology* 14.9, pp. 2873–2882. ISSN: 0962-1083, 1365-294X. DOI: 10.1111/j.1365-294X.2005.02644.x. URL: http://doi.wiley.com/10.1111/j.1365-294X.2005.02644.x (visited on 11/05/2020).

Parducci, Laura et al. (May 2017). "Ancient Plant DNA in Lake Sediments". In: *New Phytologist* 214.3, pp. 924–942. ISSN: 0028646X. DOI: 10.1111/nph.14470. URL: http://doi.wiley.com/10.1111/nph.14470 (visited on 11/05/2020).

Parducci, Laura et al. (2019). "Shotgun Environmental DNA, Pollen, and Macrofossil Analysis of Lateglacial Lake Sediments From Southern Sweden". In: *Frontiers in Ecology and Evolution* 7. ISSN: 2296-701X. URL: https://www.frontiersin.org/articles/10.3389/fevo.2019.00189 (visited on 10/09/2022).

Pedersen, Mikkel Winther et al. (Sept. 2013). "A Comparative Study of Ancient Environmental DNA to Pollen and Macrofossils from Lake Sediments Reveals Taxonomic Overlap and Additional Plant Taxa". In: *Quaternary Science Reviews* 75, pp. 161–168. ISSN: 02773791. DOI: 10.1016/j.quascirev.2013.06.0

06. URL: https://linkinghub.elsevier.com/retrieve/pii/S0277379113002187 (visited on 11/05/2020).

Pereira, F. B. et al. (2019). "A Study of Climate Variability in Papua New Guinea". In: *Journal of Geoscience and Environment Protection* 07.05, pp. 45–52. ISSN: 2327-4336, 2327-4344. DOI: 10.4236/gep.2019.75005. URL: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/gep.2019.75005 (visited on 10/25/2022).

Peris, David et al. (Sept. 28, 2020). "DNA from Resin-Embedded Organisms: Past, Present and Future". In: *PLOS ONE* 15.9. Ed. by David Caramelli, e0239521. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0239521. URL: https://dx.plos.org/10.1371/journal.pone.0239521 (visited on 11/05/2020).

Peyrégne, Stéphane and Benjamin M. Peter (Dec. 2020). "AuthentiCT: A Model of Ancient DNA Damage to Estimate the Proportion of Present-Day DNA Contamination". In: *Genome Biology* 21.1, p. 246. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02123-y. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02123-y (visited on 10/20/2022).

Phillipson, J.David (Jan. 1982). "Chemical Investigations of Herbarium Material for Alkaloids". In: *Phytochemistry* 21.10, pp. 2441–2456. ISSN: 00319422. DOI: 10.1016/0031-9422(82)85239-4. URL: https://linkinghub.elsevier.com/retrieve/pii/0031942282852394 (visited on 08/19/2022).

Picard Toolkit (2019). URL: https://broadinstitute.github.io/picard/ (visited on 11/30/2022).

Pietsch, Dana and Miranda Morris (2010). "Modern and Ancient Knowledge of Conserving Soils in Socotra Island, Yemen". In: *Land degradation and desertification: assessment, mitigation and remediation*, pp. 375–386.

Plana, Vanessa (2003). "Phylogenetic Relationships of the Afro-Malagasy Members of the Large Genus Begonia Inferred from trnL Intron Sequences". In: *Systematic Botany* 28.4, pp. 693–704. ISSN: 0363-6445. JSTOR: 25063916. URL: https://www.jstor.org/stable/25063916 (visited on 08/13/2023).

Poinar, H. N. (July 17, 1998). "Molecular Coproscopy: Dung and Diet of the Extinct Ground Sloth Nothrotheriops Shastensis". In: *Science (New York, N.Y.)* 281.5375, pp. 402–406. DOI: 10.1126/science.281.5375.402. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.281.5375.402 (visited on 11/05/2020).

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin (July 1, 2009). "FastTree: Computing Large Minimum Evolution Trees with Profiles Instead of a Distance Matrix". In: *Molecular Biology and Evolution* 26.7, pp. 1641–1650. ISSN: 0737-4038. DOI: 10.1093/molbev/msp077. URL: https://doi.org/10.1093/molbev/msp077 (visited on 03/28/2023).

Purcell, Shaun et al. (Sept. 2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *American Journal of Human Genetics* 81.3, p. 559. DOI: 10.1086/519795. pmid: 17701901. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/ (visited on 10/27/2022).

Racimo, Fernando, Jeremy J. Berg, and Joseph K. Pickrell (Apr. 2018). "Detecting Polygenic Adaptation in Admixture Graphs". In: *Genetics* 208.4, pp. 1565–1584. ISSN: 0016-6731, 1943-2631. DOI: 10.1534/genetics.117.300489. URL: http://www.genetics.org/lookup/doi/10.1534/genetics.117.300489 (visited on 01/09/2021).

Rambaut, Andrew (2022). *Rambaut/Figtree*. URL: https://github.com/rambaut/figtree (visited on 11/30/2022).

Ramos-Madrigal, Jazmín et al. (June 2019). "Palaeogenomic Insights into the Origins of French Grapevine Diversity". In: *Nature Plants* 5.6, pp. 595–603. ISSN: 2055-0278. DOI: 10.1038/s41477-019-0437-5. URL: http://www.nature.com/articles/s41477-019-0437-5 (visited on 11/05/2020).

Raveloaritiana, Estelle et al. (2021). "Land-Use Intensification Increases Richness of Native and Exotic Herbaceous Plants, but Not Endemics, in Malagasy Vanilla Landscapes". In: *Diversity and Distributions* 27.5, pp. 784–798. ISSN: 1472-4642. DOI: 10.1111/ddi.13226. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ddi.13226 (visited on 10/13/2022).

RBGE, Royal Botanic Garden (2020). *350th Anniversary | What's On*. RBGE. URL: https://www.rbge.org.uk/about-us/our-history/ (visited on 11/28/2022).

Reflora | Kew (2023). URL: https://www.kew.org/science/our-science/projects/reflora (visited on 04/04/2023).

Rezende, Marcelo et al. (Aug. 2022). "Identifying Suitable Restoration and Conservation Areas for Dracaena Cinnabari Balf.f. in Socotra, Yemen". In: *Forests* 13.8 (8), p. 1276. ISSN: 1999-4907. DOI: 10.3390/f13081276. URL: https://www.mdpi.com/1999-4907/13/8/1276 (visited on 08/06/2023).

Riccardi, Tullia et al. (Sept. 1, 2020). "Environmental Factors and Human Activity as Drivers of Tree Cover and Density on the Island of Socotra, Yemen". In: *Rendiconti Lincei. Scienze Fisiche e Naturali* 31.3, pp. 703–718. ISSN: 1720-0776. DOI: 10.1007/s12210-020-00923-9. URL: https://doi.org/10.1007/s12210-020-00923-9 (visited on 10/13/2022).

Rijal, Dilli P. et al. (July 30, 2021). "Sedimentary Ancient DNA Shows Terrestrial Plant Richness Continuously Increased over the Holocene in Northern Fennoscandia". In: *Science Advances* 7.31, eabf9557. DOI: 10.1126/sciadv.abf9557. URL: https://www.science.org/doi/10.1126/sciadv.abf9557 (visited on 10/10/2022).

Rohland, Nadin et al. (Jan. 19, 2015). "Partial Uracil–DNA–Glycosylase Treatment for Screening of Ancient DNA". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1660, p. 20130624. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2013.0624. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2013.0624 (visited on 11/05/2020).

Rollo, F. (Dec. 1985). "Characterisation by Molecular Hybridization of RNA Fragments Isolated from Ancient (1400 B.C.) Seeds". In: *Theoretical and Applied Genetics* 71.2, pp. 330–333. ISSN: 0040-5752, 1432-2242. DOI: 10.1007/BF00252076. URL: http://link.springer.com/10.1007/BF00252076 (visited on 11/05/2020).

Rollo, F. et al. (Oct. 1, 2002). "Otzi's Last Meals: DNA Analysis of the Intestinal Content of the Neolithic Glacier Mummy from the Alps". In: *Proceedings of the National Academy of Sciences* 99.20, pp. 12594–12599. ISSN: 0027-8424, 1091-6490. doi: 10.1073/pnas.192184599. url: http://www.pnas.org/cgi/doi/10.1073/pnas.192184599 (visited on 11/05/2020).

Rollo, Franco, Franco Maria Venanzi, and Augusto Amici (Dec. 1991). “Nucleic Acids in Mummified Plant Seeds: Biochemistry and Molecular Genetics of Pre-Columbian Maize”. In: *Genetical Research* 58.3, pp. 193–201. issn: 0016-6723, 1469-5073. doi: 10.1017/S0016672300029943. url: https://www.cambridge.org/core/product/identifier/S0016672300029943/type/journal_article (visited on 01/22/2021).

Rønsted, Nina, Olwen M. Grace, and Mark A. Carine (Aug. 21, 2020). “Editorial: Integrative and Translational Uses of Herbarium Collections Across Time, Space, and Species”. In: *Frontiers in Plant Science* 11, p. 1319. issn: 1664-462X. doi: 10.3389/fpls.2020.01319. url: https://www.frontiersin.org/article/10.3389/fpls.2020.01319/full (visited on 05/07/2022).

Salick, Jan, Katie Konchar, and Mark Nesbitt (2014). *Curating Biocultural Collections: A Handbook.* isbn: 978-1-84246-509-7.

Schaefer, Hanno and Susanne S. Renner (Feb. 2011). “Phylogenetic Relationships in the Order Cucurbitales and a New Classification of the Gourd Family (Cucurbitaceae)”. In: *TAXON* 60.1, pp. 122–138. issn: 00400262. doi: 10.1002/tax.601011. url: https://onlinelibrary.wiley.com/doi/10.1002/tax.601011 (visited on 07/07/2023).

Schlumbaum, Angela and Viviane Jaenicke-Després (Mar. 2008). “Ancient Plant DNA in Archaeobotany”. In: *Vegetation History and Archaeobotany* 17.2, pp. 233–244. issn: 0939-6314, 1617-6278. doi: 10.1007/s00334-007-0125-7. url: http://link.springer.com/10.1007/s00334-007-0125-7 (visited on 01/18/2021).

Schoen, Daniel J. and Anthony H. D. Brown (Nov. 1, 2001). “The Conservation of Wild Plant Species in Seed Banks: Attention to Both Taxonomic Coverage and Population Biology Will Improve the Role of Seed Banks as Conservation Tools”. In: *BioScience* 51.11, pp. 960–966. issn: 0006-3568. doi: 10.1641/0006-3568(2001)051[0960:TCOWPS]2.0.CO;2. url: https://doi.org/10.1641/0006-3568(2001)051[0960:TCOWPS]2.0.CO;2 (visited on 08/04/2023).

Scholte, Paul, Abdulraqueb Al-Okaishi, and Ahmed Saed Suleyman (July 2011). “When Conservation Precedes Development: A Case Study of the Opening up of the Socotra Archipelago, Yemen”. In: *Oryx* 45.3, pp. 401–410. issn: 1365-3008, 0030-6053. doi: 10.1017/S0030605310001535. url: https://www.cambridge.org/core/journals/oryx/article/when-conservation-precedes-development-a-case-study-of-the-opening-up-of-the-socotra-archipelago-yemen/F85C7CE07A472C2D7F9154FF03ED8E7A# (visited on 08/06/2023).

Schubert, Mikkel et al. (May 2014). “Characterization of Ancient and Modern Genomes by SNP Detection and Phylogenomic and Metagenomic Analysis Using PALEOMIX”. In: *Nature Protocols* 9.5, pp. 1056–1082. issn: 1754-2189, 1750-2799. doi: 10.1038/nprot.2014.063. url: http://www.nature.com/articles/nprot.2014.063 (visited on 12/16/2020).

Schulte, Luise et al. (Apr. 2021). “Hybridization Capture of Larch (Larix Mill.) Chloroplast Genomes from Sedimentary Ancient DNA Reveals Past Changes of Siberian Forest". In: *Molecular Ecology Resources* 21.3, pp. 801–815. ISSN: 1755-0998. DOI: 10.1111/1755-0998.13311. pmid: 33319428.

Schwarzbach, A. E. and R. E. Ricklefs (Apr. 2000). "Systematic Affinities of Rhizophoraceae and Anisophylleaceae, and Intergeneric Relationships within Rhizophoraceae, Based on Chloroplast DNA, Nuclear Ribosomal DNA, and Morphology". In: *American Journal of Botany* 87.4, pp. 547–564. ISSN: 0002-9122. pmid: 10766727.

Schwörer, Christoph et al. (Apr. 5, 2022). "The Untapped Potential of Macrofossils in Ancient Plant DNA Research". In: *New Phytologist*, nph.18108. ISSN: 0028-646X, 1469-8137. DOI: 10.1111/nph.18108. URL: https://onlinelibrary.wiley.com/doi/10.1111/nph.18108 (visited on 05/27/2022).

Scott, Kirsten D. and Julia Playford (June 1996). "DNA Extraction Technique for PCR in Rain Forest Plant Species". In: *BioTechniques* 20.6, pp. 974–978. ISSN: 0736-6205. DOI: 10.2144/96206bm07. URL: https://www.future-science.com/doi/abs/10.2144/96206bm07 (visited on 04/25/2023).

Scott, Michael F. et al. (Nov. 2019). "A 3,000-Year-Old Egyptian Emmer Wheat Genome Reveals Dispersal and Domestication History". In: *Nature Plants* 5.11, pp. 1120–1128. ISSN: 2055-0278. DOI: 10.1038/s41477-019-0534-5. URL: http://www.nature.com/articles/s41477-019-0534-5 (visited on 02/03/2021).

Shepherd, Lara D. (Aug. 31, 2017). "A Non-Destructive DNA Sampling Technique for Herbarium Specimens". In: *PLOS ONE* 12.8, e0183555. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0183555. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0183555 (visited on 10/10/2022).

Silva, Christian et al. (Feb. 2017). "Museomics Resolve the Systematics of an Endangered Grass Lineage Endemic to North-Western Madagascar". In: *Annals of Botany* 119.3, pp. 339–351. ISSN: 0305-7364, 1095-8290. DOI: 10.1093/aob/mcw208. URL: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcw208 (visited on 05/07/2022).

Silvertown J, D Charlesworth (2001). *Introduction to Plant Population Biology.* Blackwell Science.

Slon, Viviane et al. (May 12, 2017). "Neandertal and Denisovan DNA from Pleistocene Sediments". In: *Science (New York, N.Y.)* 356.6338, pp. 605–608. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam9695. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aam9695 (visited on 11/05/2020).

Smith, Oliver et al. (July 4, 2014). "Genomic Methylation Patterns in Archaeological Barley Show De-Methylation as a Time-Dependent Diagenetic Process". In: *Scientific Reports* 4.1, p. 5559. ISSN: 2045-2322. DOI: 10.1038/srep05559. URL: https://www.nature.com/articles/srep05559 (visited on 08/08/2023).

— (May 2015). "Genomic Methylation Patterns in Archaeological Barley Show De-Methylation as a Time-Dependent Diagenetic Process". In: *Scientific Reports* 4.1, p. 5559. ISSN: 2045-2322. DOI: 10.1038/srep05559. URL: http://www.nature.com/articles/srep05559 (visited on 11/05/2020).

Smith, Oliver et al. (2017). "Small RNA Activity in Archeological Barley Shows Novel Germination Inhibition in Response to Environment". In: p. 8.

Spielman, Derek, Barry W. Brook, and Richard Frankham (Oct. 19, 2004). “Most Species Are Not Driven to Extinction before Genetic Factors Impact Them”. In: *Proceedings of the National Academy of Sciences* 101.42, pp. 15261–15264. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0403809101. URL: https://pnas.org/doi/full/10.1073/pnas.0403809101 (visited on 10/15/2022).

Staats, Martijn et al. (Dec. 5, 2011). “DNA Damage in Plant Herbarium Tissue”. In: *PLoS ONE* 6.12. Ed. by Carles Lalueza-Fox, e28448. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0028448. URL: https://dx.plos.org/10.1371/journal.pone.0028448 (visited on 10/26/2022).

Stull, Gregory W. et al. (May 2020). “Nuclear Phylogenomic Analyses of Asterids Conflict with Plastome Trees and Support Novel Relationships among Major Lineages”. In: *American Journal of Botany* 107.5, pp. 790–805. ISSN: 1537-2197. DOI: 10.1002/ajb2.1468. pmid: 32406108.

Supple, Megan A. and Beth Shapiro (Sept. 11, 2018). “Conservation of Biodiversity in the Genomics Era”. In: *Genome Biology* 19.1, p. 131. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1520-3. URL: https://doi.org/10.1186/s13059-018-1520-3 (visited on 07/28/2023).

Swarts, Kelly et al. (Aug. 4, 2017). “Genomic Estimation of Complex Traits Reveals Ancient Maize Adaptation to Temperate North America”. In: *Science (New York, N.Y.)* 357.6350, pp. 512–515. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aam9425. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.aam9425 (visited on 11/05/2020).

Sweeney, Patrick W. et al. (2018). “Large–Scale Digitization of Herbarium Specimens: Development and Usage of an Automated, High–Throughput Conveyor System”. In: *TAXON* 67.1. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.12705/671.10, pp. 165–178. ISSN: 1996-8175. DOI: 10.12705/671.10. URL: https://onlinelibrary.wiley.com/doi/abs/10.12705/671.10 (visited on 11/28/2022).

Särkinen, Tiina et al. (Aug. 28, 2012). “How to Open the Treasure Chest? Optimising DNA Extraction from Herbarium Specimens”. In: *PLOS ONE* 7.8, e43808. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0043808. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0043808 (visited on 10/10/2022).

Søe, Martin Jensen et al. (2018). “Ancient DNA from Latrines in Northern Europe and the Middle East (500 BC-1700 AD) Reveals Past Parasites and Diet”. In: *PloS One* 13.4, e0195481. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0195481. pmid: 29694397.

Sønstebø, J. H. et al. (2010). “Using Next-Generation Sequencing for Molecular Reconstruction of Past Arctic Vegetation and Climate”. In: *Molecular Ecology Resources* 10.6. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1755-0998.2010.02855.x, pp. 1009–1018. ISSN: 1755-0998. DOI: 10.1111/j.1755-0998.2010.02855.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1755-0998.2010.02855.x (visited on 11/29/2022).

Teixeira, João C. and Christian D. Huber (Mar. 9, 2021). “The Inflated Significance of Neutral Genetic Diversity in Conservation Genetics”. In: *Proceedings of the National Academy of Sciences* 118.10, e2015096118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2015096118. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.2015096118 (visited on 02/20/2021).

Thiers, Barbara M. (Feb. 2022). *The_Worlds_Herbaria_Jan_2022.Pdf*. Index Herbariorum website. URL: http://sweetgum.nybg.org/science/ih/annual-report/.

Trucchi, Emiliano et al. (Feb. 2021). “Ancient Genomes Reveal Early Andean Farmers Selected Common Beans While Preserving Diversity”. In: *Nature Plants* 7.2, pp. 123–128. ISSN: 2055-0278. DOI: 10.1038/s41477-021-00848-7. URL: https://www.nature.com/articles/s41477-021-00848-7 (visited on 10/09/2022).

Tseng, Yu-Hsin et al. (2017). “Development and Characterization of EST-SSR Markers for Begonia Luzhaiensis (Begoniaceae)”. In: *Applications in Plant Sciences* 5.5, p. 1700024. ISSN: 2168-0450. DOI: 10.3732/apps.1700024. URL: https://onlinelibrary.wiley.com/doi/abs/10.3732/apps.1700024 (visited on 07/17/2023).

Twyford, A. D., C. A. Kidner, and R. A. Ennos (Apr. 2014a). “Genetic Differentiation and Species Cohesion in Two Widespread Central American Begonia Species”. In: *Heredity* 112.4 (4), pp. 382–390. ISSN: 1365-2540. DOI: 10.1038/hdy.2013.116. URL: https://www.nature.com/articles/hdy2013116 (visited on 08/16/2023).

Twyford, Alex D. et al. (Jan. 2013). “Population History and Seed Dispersal in Widespread Central American *Begonia* Species (Begoniaceae) Inferred from Plastome-Derived Microsatellite Markers: POPULATION BIOLOGY OF *BEGONIA*”. In: *Botanical Journal of the Linnean Society* 171.1, pp. 260–276. ISSN: 00244074. DOI: 10.1111/j.1095-8339.2012.01265.x. URL: https://academic.oup.com/botlinnean/article-lookup/doi/10.1111/j.1095-8339.2012.01265.x (visited on 03/15/2023).

Twyford, Alex D. et al. (Feb. 2014b). “The Evolution of Sex Ratio Differences and Inflorescence Architectures in *Begonia* (Begoniaceae)”. In: *American Journal of Botany* 101.2, pp. 308–317. ISSN: 00029122. DOI: 10.3732/ajb.1300090. URL: http://doi.wiley.com/10.3732/ajb.1300090 (visited on 11/25/2022).

Vachaspati, Pranjal and Tandy Warnow (2015). “ASTRID: Accurate Species TRees from Internode Distances”. In: *BMC genomics* 16 Suppl 10 (Suppl 10), S3. ISSN: 1471-2164. DOI: 10.1186/1471-2164-16-S10-S3. pmid: 26449326.

VanAndel, Tinde et al. (Jan. 17, 2022). “Sixteenth-Century Tomatoes in Europe: Who Saw Them, What They Looked like, and Where They Came From”. In: *PeerJ* 10, e12790. ISSN: 2167-8359. DOI: 10.7717/peerj.12790. URL: https://peerj.com/articles/12790 (visited on 09/16/2022).

Vatanparast, Mohammad et al. (Mar. 2018). “Targeting Legume Loci: A Comparison of Three Methods for Target Enrichment Bait Design in Leguminosae Phylogenomics”. In: *Applications in Plant Sciences* 6.3, e1036. ISSN: 2168-0450. DOI: 10.1002/aps3.1036. pmid: 29732266.

Villaverde, Tamara et al. (Oct. 2018). “Bridging the Micro- and Macroevolutionary Levels in Phylogenomics: Hyb-Seq Solves Relationships from Populations to Species and Above”. In: *New Phytologist* 220.2, pp. 636–650. ISSN:

0028646X. DOI: 10.1111/nph.15312. URL: https://onlinelibrary.wiley.com/doi/10.1111/nph.15312 (visited on 05/16/2022).

Wagner, Stefanie et al. (Mar. 2018). "High-Throughput DNA Sequencing of Ancient Wood". In: *Molecular Ecology* 27.5, pp. 1138–1154. ISSN: 09621083. DOI: 10.1111/mec.14514. URL: http://doi.wiley.com/10.1111/mec.14514 (visited on 11/05/2020).

Wales, Nathan and Logan Kistler (2019). "Extraction of Ancient DNA from Plant Remains". In: *Methods in Molecular Biology (Clifton, N.J.)* 1963, pp. 45–55. ISSN: 1940-6029. DOI: 10.1007/978-1-4939-9176-1_6. pmid: 30875043.

Wales, Nathan et al. (Aug. 2016). "The Limits and Potential of Paleogenomic Techniques for Reconstructing Grapevine Domestication". In: *Journal of Archaeological Science* 72, pp. 57–70. ISSN: 03054403. DOI: 10.1016/j.jas.2016.05.014. URL: https://linkinghub.elsevier.com/retrieve/pii/S0305440316300772 (visited on 11/06/2020).

Walker, Barnaby E., Allan Tucker, and Nicky Nicolson (Jan. 13, 2022). "Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning". In: *Frontiers in Plant Science* 12, p. 806407. ISSN: 1664-462X. DOI: 10.3389/fpls.2021.806407. URL: https://www.frontiersin.org/articles/10.3389/fpls.2021.806407/full (visited on 02/21/2023).

Wayne, Robert K. and Phillip A. Morin (Mar. 2004). "Conservation Genetics in the New Molecular Age". In: *Frontiers in Ecology and the Environment* 2.2, pp. 89–97. ISSN: 1540-9295. DOI: 10.1890/1540-9295(2004)002[0089:CGITNM]2.0.CO;2. URL: http://doi.wiley.com/10.1890/1540-9295(2004)002[0089:CGITNM]2.0.CO;2 (visited on 08/07/2023).

Weir, B. S. and C. Clark Cockerham (Nov. 1984). "ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE". In: *Evolution; International Journal of Organic Evolution* 38.6, pp. 1358–1370. ISSN: 1558-5646. DOI: 10.1111/j.1558-5646.1984.tb05657.x. pmid: 28563791.

Weitemier, Kevin et al. (Sept. 2014). "Hyb-Seq: Combining Target Enrichment and Genome Skimming for Plant Phylogenomics". In: *Applications in Plant Sciences* 2.9, p. 1400042. ISSN: 2168-0450. DOI: 10.3732/apps.1400042. URL: http://doi.wiley.com/10.3732/apps.1400042 (visited on 05/16/2022).

Weiß, Clemens L. et al. (June 2016). "Temporal Patterns of Damage and Decay Kinetics of DNA Retrieved from Plant Herbarium Specimens". In: *Royal Society Open Science* 3.6, p. 160239. ISSN: 2054-5703. DOI: 10.1098/rsos.160239. URL: https://royalsocietypublishing.org/doi/10.1098/rsos.160239 (visited on 05/02/2022).

Willerslev, E. (May 2, 2003). "Diverse Plant and Animal Genetic Records from Holocene and Pleistocene Sediments". In: *Science (New York, N.Y.)* 300.5620, pp. 791–795. ISSN: 00368075, 10959203. DOI: 10.1126/science.1084114. URL: https://www.sciencemag.org/lookup/doi/10.1126/science.1084114 (visited on 11/05/2020).

Willerslev, Eske et al. (Feb. 2014). "Fifty Thousand Years of Arctic Vegetation and Megafaunal Diet". In: *Nature* 506.7486, pp. 47–51. ISSN: 1476-4687. DOI: 10.1038/nature12921. URL: https://www.nature.com/articles/nature12921 (visited on 10/10/2022).

Willi, Yvonne et al. (Jan. 4, 2022). “Conservation Genetics as a Management Tool: The Five Best-Supported Paradigms to Assist the Management of Threatened Species”. In: *Proceedings of the National Academy of Sciences* 119.1, e2105076119. DOI: 10.1073/pnas.2105076119. URL: https://www.pnas.org/doi/10.1073/pnas.2105076119 (visited on 07/27/2023).

Williams, M. V., T. Winters, and K. S. Waddell (Feb. 1, 1987). “In Vivo Effects of Mercury (II) on Deoxyuridine Triphosphate Nucleotidohydrolase, DNA Polymerase (Alpha, Beta), and Uracil-DNA Glycosylase Activities in Cultured Human Cells: Relationship to DNA Damage, DNA Repair, and Cytotoxicity.” In: *Molecular Pharmacology* 31.2, pp. 200–207. ISSN: 0026-895X, 1521-0111. pmid: 3027530. URL: https://molpharm.aspetjournals.org/content/31/2/200 (visited on 11/28/2022).

Willing, Eva-Maria, Christine Dreyer, and prefix=van useprefix=false family=Oosterhout given=Cock (Aug. 14, 2012). “Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers”. In: *PLOS ONE* 7.8, e42649. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0042649. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0042649 (visited on 08/03/2023).

Wilson, Hannah (2021). “Megadiversity and the New Guinea Orogeny”. In: p. 158.

Wilson, H.P. et al. (Dec. 15, 2020). “Three New Species of Begonia Sect. Petermannia (Begoniaceae) from Sandaun Province, Papua New Guinea”. In: *Gardens’ Bulletin Singapore* 72.2, pp. 275–284. ISSN: 03747859, 23825812. DOI: 10.26492/gbs72(2).2020-10. URL: https://www.nparks.gov.sg/sbg/research/publications/gardens-bulletin-singapore/-/media/sbg/gardens-bulletin/gbs_72_02_y2020/72_02_10_y2020_v7202_gbs_pg275.pdf (visited on 11/10/2022).

Wood, Jamie R. et al. (Apr. 2016). “Microscopic and Ancient DNA Profiling of Polynesian Dog (Kurī) Coprolites from Northern New Zealand”. In: *Journal of Archaeological Science: Reports* 6, pp. 496–505. ISSN: 2352409X. DOI: 10.1016/j.jasrep.2016.03.020. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352409X16300906 (visited on 11/05/2020).

Woudstra, Yannick et al. (Dec. 21, 2021). “A Customised Target Capture Sequencing Tool for Molecular Identification of Aloe Vera and Relatives”. In: *Scientific Reports* 11.1, p. 24347. ISSN: 2045-2322. DOI: 10.1038/s41598-021-03300-0. URL: https://www.nature.com/articles/s41598-021-03300-0 (visited on 11/28/2022).

Wright, Sewall (Mar. 1, 1931). “EVOLUTION IN MENDELIAN POPULATIONS”. In: *Genetics* 16.2, pp. 97–159. ISSN: 1943-2631. DOI: 10.1093/genetics/16.2.97. URL: https://doi.org/10.1093/genetics/16.2.97 (visited on 07/28/2023).

Yang, Luming et al. (2012). “Chromosome Rearrangements during Domestication of Cucumber as Revealed by High-Density Genetic Mapping and Draft Genome Assembly”. In: *The Plant Journal* 71.6, pp. 895–906. ISSN: 1365-313X. DOI: 10.1111/j.1365-313X.2012.05017.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2012.05017.x (visited on 07/17/2023).

Zedane, Loubab et al. (2015). "Museomics Illuminate the History of an Extinct, Paleoendemic Plant Lineage (Hesperelaea, Oleaceae) Known from an 1875 Collection from Guadalupe Island, Mexico". In: *Biological Journal of the Linnean Society*, p. 14.

Zedane, Loubab et al. (Jan. 2016). "Museomics Illuminate the History of an Extinct, Paleoendemic Plant Lineage (*Hesperelaea* , Oleaceae) Known from an 1875 Collection from Guadalupe Island, Mexico: Biogeographic History of *Hesperelaea*". In: *Biological Journal of the Linnean Society* 117.1, pp. 44–57. ISSN: 00244066. DOI: 10.1111/bij.12509. URL: https://academic.oup.com/biolinnean/article-lookup/doi/10.1111/bij.12509 (visited on 05/07/2022).

Zhang, Chao et al. (May 8, 2018). "ASTRAL-III: Polynomial Time Species Tree Reconstruction from Partially Resolved Gene Trees". In: *BMC Bioinformatics* 19.6, p. 153. ISSN: 1471-2105. DOI: 10.1186/s12859-018-2129-y. URL: https://doi.org/10.1186/s12859-018-2129-y (visited on 03/28/2023).

Zhang, Ning et al. (2012). "Highly Conserved Low-Copy Nuclear Genes as Effective Markers for Phylogenetic Analyses in Angiosperms". In: *New Phytologist* 195.4, pp. 923–937. ISSN: 1469-8137. DOI: 10.1111/j.1469-8137.2012.04212.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.2012.04212.x (visited on 07/17/2023).

Zheng, Zhuo et al. (Oct. 5, 2021). "Anthropogenic Impacts on Late Holocene Land-Cover Change and Floristic Biodiversity Loss in Tropical Southeastern Asia". In: *Proceedings of the National Academy of Sciences* 118.40, e2022210118. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2022210118. URL: https://pnas.org/doi/full/10.1073/pnas.2022210118 (visited on 11/25/2022).

Zhou, Boyan et al. (Dec. 2017). "AntCaller: An Accurate Variant Caller Incorporating Ancient DNA Damage". In: *Molecular Genetics and Genomics* 292.6, pp. 1419–1430. ISSN: 1617-4615, 1617-4623. DOI: 10.1007/s00438-017-1358-5. URL: http://link.springer.com/10.1007/s00438-017-1358-5 (visited on 11/08/2021).

Zimmermann, Heike H. et al. (Oct. 13, 2017). "The History of Tree and Shrub Taxa on Bol'shoy Lyakhovsky Island (New Siberian Archipelago) since the Last Interglacial Uncovered by Sedimentary Ancient DNA and Pollen Data". In: *Genes* 8.10, p. 273. ISSN: 2073-4425. DOI: 10.3390/genes8100273. pmid: 29027988.

Ågren, Jon and Douglas W. Schemske (Feb. 1993). "OUTCROSSING RATE AND INBREEDING DEPRESSION IN TWO ANNUAL MONOECIOUS HERBS, *BEGONIA HIRSUTA* AND *B. SEMIOVATA*". In: *Evolution; international journal of organic evolution* 47.1, pp. 125–135. ISSN: 00143820. DOI: 10.1111/j.1558-5646.1993.tb01204.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1993.tb01204.x (visited on 11/25/2022).

Appendices

Appendix A

Target capture reads depth of coverage for PNG and Socotran specimens

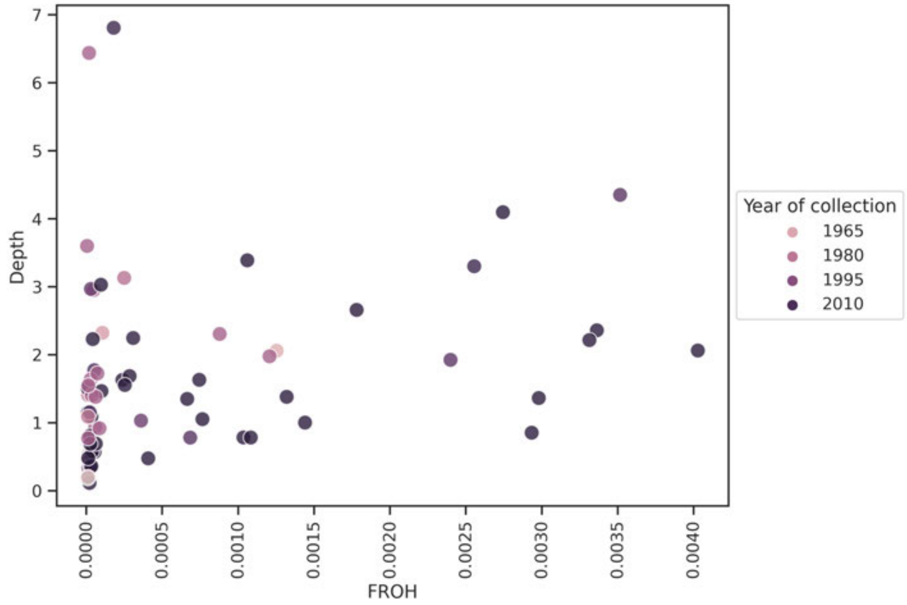


Figure A.1: Relationship between depth of coverage, F_{ROH}, and date of sampling of the specimen.

Appendix B

A Hybrid Capture Bait Set For *Begonia*

T Michel (1,2), Y-H Tseng (3), Hannah Wilson (4,2) , K-F Chung (3), C Kidner (1,2)*

1 Institute of Molecular Plant Sciences, University of Edinburgh, The King's Buildings, Edinburgh EH9 3BF, UK

2 Royal Botanic Garden Edinburgh, 20A Inverleith Row, Edinburgh EH3 5LR, UK

3 Research Museum and Herbarium (HAST), Biodiversity Research Center, Academia Sinica, 128 Academia Road, Section 2, Taipei 115201, Taiwan

4 Institute of Biodiversity Animal Health & Comparative Medicine, Glasgow University, Graham Kerr Building, Glasgow G12 8QQ, UK

*** Correspondence:**

Catherine Kidner

Keywords: Begoniaceae, Target Capture Sequencing, Enrichment Panel, Probe set

Running title: Bait Set for *Begonia*

Length: 5852 words, two tables, 8 figures

Abstract

Hybrid capture with baits has proven to be a rich source of genetic data for many genera. The depth of information provided allows resolution of rapid radiations and of deep phylogenetic patterns. Retrieved data can also be used for population genetic studies and analysis of functional genetic diversity. To gain a better understanding of the evolutionary patterns across this large, diverse and fascinating genus through phylogenetics, population genetics and sequence analysis, we have designed and tested a set of 1239 baits covering low copy number and functionally annotated genes involved in shade adaptation and development and genetically linked to key traits. We demonstrate successful recovery of sequence data from species across *Begonia* and from fresh, silica

dried and older herbarium material.

B.1 Introduction

Hybrid capture is now a common method for retrieving sequence data for phylogenetics and population genetics in plants (Cronn *et al.*, 2012; Dodsworth *et al.*, 2019; Hale *et al.*, 2020; Larridon *et al.*, 2020; Slimp *et al.*, 2021). The method gives high coverage for hundreds of chosen loci across the genome. The large numbers of loci recovered gives a fuller picture of evolutionary history and allows exploration of reticulate patterns produced by hybridisation events. The high coverage gives confidence in variant calling and analysis of variation at population levels. The ability to choose loci allows testing of evolutionary hypotheses about the role of specific types of genes (McKain *et al.*, 2018).

The phylogenetic reach of bait sets varies. Cross angiosperm baits have been produced and are becoming widely used for both deep and shallow phylogenomic studies (Johnson *et al.*, 2019; Larridon *et al.*, 2020; Slimp *et al.*, 2021). Family-wide sets have been useful in untangling relationships, particularly in large groups and recent radiations (for example; Compositae - (Mandel *et al.*, 2015), Euphorbiaceae (Villaverde *et al.* 2018), mimosoid legumes - (Koenen *et al.*, 2020), and Annonaceae (Couvreur *et al.*, 2018)). Sets focused on specific genera have been useful in resolving relationships in difficult groups (Folk *et al.*, 2015; Pezzini, 2019; Soto Gomez *et al.*, 2019), and even species-specific sets have been designed for population genetics and breeding (for example barley (*Hordeum vulgar*e L.) (Hill *et al.*, 2019) and wheat (*Triticum aestivum* L.) (Gardiner *et al.*, 2019)).

Begonia L. is one of the largest plant genera (Frodin, 2004; Moonlight *et al.*, 2018) and presents a number of phylogenetic problems with increased resolution needed at both deep (sectional divisions) and shallow (recent radiations) levels. In addition there is a need to better understand the influence of past hybridisation events on present *Begonia* species diversity given the evidence of multiple ancient and recent hybridisations (Goodall-Copestake *et al.*, 2010; Thomas *et al.*, 2012; Moonlight *et al.*, 2015; Tseng *et al.*, 2017; Liu *et al.*, 2019a). Hybrid capture provides a wealth of genetic data which can help resolve these issues and provide data for further studies of diversity and evolution across the genus.

In this paper we aim to provide background and advice on a genus-specific bait set for *Begonia* to encourage the use of this technology in addressing some of the question about this exceptional genus. We describe how we designed a specific bait set for *Begonia* which incorporates developmental genes, differentially expressed genes and genes linked to traits. Four projects have already used the bait set (Forrest *et al.*, 2019, Wilson, 2021, and two currently unpublished projects). We use data generated by these projects to characterise the performance of the bait set. We compare samples, captures and baits to determine the range of sample quality and species the bait set works on, the range of hybridisation conditions and the performance of individual baits in different captures. To compare sequence analysis pipelines we analyse data from one capture using 5 pipelines to compare assembly metrics and phylogenies generated as concatenated data and gene the analysis. The comparison of captures and analyses in this methods-focused paper will provide guidance for future projects using this bait set to answer the many questions about *Begonia* biology.

B.2 Methods

B.2.1 Bait design

Our starting point was a transcriptome produced from mature leaves and male flower buds of the Asian species *B. luzhaiensis* T.C.Ku (Tseng *et al.*, 2017) (Fig.B.1). BLASTN of the transcriptome to itself identified 15,349 sequences with 98% identity (or less) over 100bp to another sequence. These were taken to be likely single copy genes.

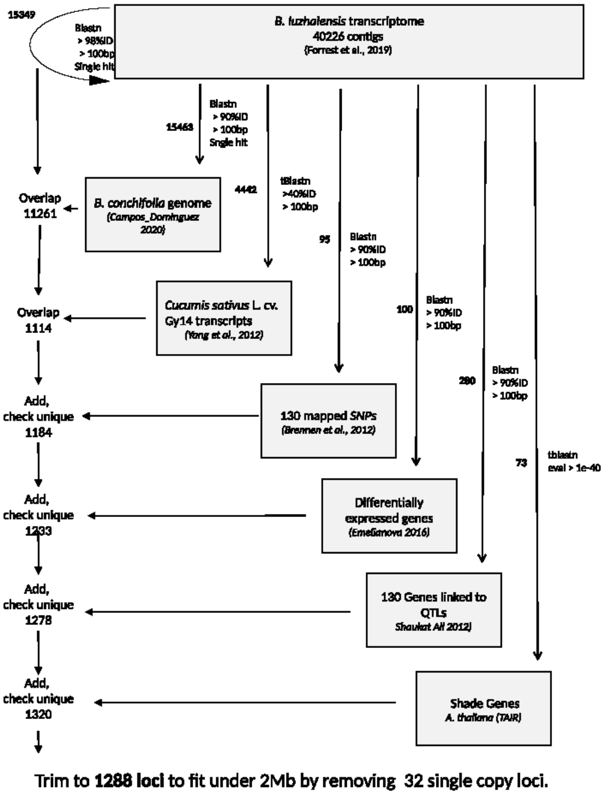


Figure B.1: Design of a bait set for Begonia. Sequences from the Begonia luzhaiensis transcriptome were self-blasted to obtain possible single-copy genes; the overlap with the set of genes annotated in both the B. conchifolia genome and the *Cucumis sativus* genome was obtained. To this set were added sequences of the markers from a genetic map of Begonia, differentially expressed transcriptional factors, genes linked to quantitative trait loci (QTL), and a set of candidate genes for shade growth. Numbers in bold are the numbers of sequences retained at each step. ID, identity; SNP, single- nucleotide polymorphism.

We used BLASTN to compare the *B. luzhaiensis* transcriptome sequences with the set of annotated genes from the genome of the Central American *Begonia conchifolia* A.Dietr. (Campos-Dominguez, 2020). We identified 15,463 sequences with 'good' matches of greater than 90% identity over greater than 100bp. We took the overlap of this set with the set of 'single copy' sequences identified in step one for further analysis - 11,261 sequences. We filtered the 11,261 sequences to just those with matches at > 90% and >100bp to an annotated cucumber gene using the Yang genome assembly (Yang *et al.*, 2012). This resulted in 1,114 sequences.

We added 3 sets of sequences which may be useful for functional studies in *Begonia*. The first were the matches to 130 single nucleotide polymorphisms (SNP) markers used to generate the first *Begonia* genetic map (Brennan *et al.*, 2012); the second set were matches to a hundred transcription factors differentially expressed between *B. conchifolia* and *B. plebeja* Liebm. (Emelianova & Kidner 2021); the third set were 280 sequences with good matches to genes falling within one LOD (logarithm of odds) drop of significant QTLs (quantitative trait locus) for a range of traits in B. *conchifolia* × *B. plebeja*, (Twyford *et al.*, 2014).

The final set of chosen sequences were a list of shade-associated genes. We surveyed the literature on genes associated with light responses, in particular shade tolerance and picked *PhyA PhyB PhyC PhyD PhyE CRY1 CRY2 PIF3 PIF4 PIF5 GLK PLASTID MOVEMENT HAT4* (see supplementary table 1 for details of these genes in *Arabidopsis* Schur). We used the *Arabidopsis* proteins to search for orthologs in the *B. luzhaiensis* transcriptome using tblastn and a cut off of 1e-40. These sequences were compared to the set already picked and the new sequences included to give a final set of 1320 target loci. This set of sequences went over the 2MB of a standard kit size, so it was trimmed down to 1,288 loci by removing some sequences that had been identified only as single copy with a cucumber annotated match. Daicel Arbor Biosciences (Ann Arbor, MI, USA) designed and synthesised 100-mer nucleotide baits, with 2.1× tiling across our target sequences.

The first capture with this bait set was performed on samples from Begonia section *Coelocentrum* Irmsch. Genomic DNAs were extracted from fresh or dried materials using the DNeasy Plant Mini Kit (Qiagen, Germany). Approximately 1 µg of DNA was sonicated in a Bioruptor Pico machine (Cosmo-bio Inc., Tokyo, Japan) with a program to generate fragment sizes of 400–500 bp. Using 50 µl of sonicated DNA, a dual-indexed library for each sample was prepared using NEBNext Ultra II DNA Library Prep kit (New England BioLabs, MA, USA), following the manufacturer's protocol and selecting DNA in the range of 400–500 bp. Each of the eight libraries was pooled into a 2 µg pool to perform hybridization capture of target DNA using biotinylated RNA baits from the first custom-designed MYbaits kit for *Begonia* (Arbor Biosciences, Ann Arbor, MI, USA). The custom bait set included 100mer baits with 2.1× tiling density across 1,288 loci with 1,990,537 bp. The hybridization procedure was performed at 65°C for 19 hours following the MYbaits v2.3.2 protocol. For post-capture PCR amplification, pools were amplified using for 11-12 cycles. Target-enriched libraries were quantified using QubitTM 3.0 Fluorometer (Thermo Scientific, MA, USA) and quality checked with an Agilent Bioanalyzer (Agilent, CA, USA). All samples were sequenced on the Illumina Hiseq platform (250 bp paired-end) at the High Throughput Genomics Core at Biodiversity Research Centre, Academia Sinica, Taiwan. This dataset is referred to as COEL.

Analysis of the targeted capture sequencing output from this capture identified 13 loci with high paralogy and 83 loci with >90% missing data. These loci were removed from the set and replaced with target sequences from the following two sets. For the first set we identified interesting genes involved in regulating anthocyanin synthesis, flowering, leaf form, and epigenetic regulation, among other functions. We then used the *Arabidopsis* proteins to search for orthologs in the *B. conchifolia* genome using TBLASTN and a cut-off of 1E-40. The second set were sequences matching the angiosperm353 panel (Johnson *et al.*, 2019). The original set of targets had an overlap of 23 loci with the angiosperms353 enrichment panel. We added five further overlapping loci to bring this number up to 28. The final target sequences are presented in Supplemental Data 1, with annotation in Supplemental Table 1.

B.2.2 Capture experiments

The bait set has been used in four captures so far, two published, Forrest *et al.*, 2019;(HAIR) Wilson 2021, (PNG), and two unpublished (COEL and POP). We used the sequence data from these experiments to characterise the performance of the bait set (Table 1).

HAIR This data derives from a paper testing the degree to which various preservation techniques, including using a hairdryer, affected our ability to derive useful data from samples of three exemplar species (Forrest *et al.*, 2019). Sample extraction library prep and capture methods are detailed in Forrest *et al.*, 2019 and summarised in Table 1.

PNG This data is taken from research that sampled across New Guinea begonias for a PhD study of *Begonia*'s colonisation of the island (Wilson, 2021). It had wide taxonomic sampling including many closely related species, multiple individuals for some species and technical duplicates. Extractions used the standard Qiagen DNeasy Plant Mini Kit-columns following the manufacturer's protocol. Extracted DNA was quantified, using Qubit dsDNA HS chemistry with duplicate reads taken for all samples. Quality was assessed on a DeNovix DS-11 and an Agilent TapeStation. All DNA extractions were normalized to 1.9 ngul. Samples containing HMW DNA were sheared to 300bp using a Covaris M220 Focused-ultrasonicator. Library preparation was carried out as half reactions using one NebNext Ultra II DNA Library Prep for Illumina Kit following the manufacturer's protocol. Samples that contained over 50 ng of DNA and had broad fragment size peaks on their Screentape were size selected using Seramag Sample Purification beads. The samples were then quality checked using Agilent Tapestation High Sensitivity screen tapes, and then re-quantified on Invitrogen Qubit 2.0 fluorometer using Qubit dsDNA HS chemistry, such that each library could be normalized to 10 nM. Libraries were grouped by quality into 32 pools (each containing 4 to 10 libraries). Hybridization of the library pools followed Forrest *et al.* (2019) with some modifications, using a 16 h hybridization at 62°C. The number of post- capture PCR cycles was dependent upon pool input DNA, varying from 9 to 22 cycles. Sequencing was carried out by NovogeneAIT Genomics, Singapore, on a HiSeq X, with 150bp paired end reads

COEL This data comes from a study of the radiation of *B.* section *Coleocentrum*Irmsch.which included a wide range of samples including closely related species but only a single sample per species. Sample preparation, library production and hybridisation followed the same workflow as above. Details of these samples are in preparation for publication by Yu-Hsing Tseng and Kou-Fung Chung.

POP This data derives from a study of variant calling for population genetics using a mapping population (Brennan *et al.*, 2012) and closely related samples from *B. soccotrana* Hook.f. and *B. samhaiensis* M. Hughes and A G Mill. DNA extraction followed the standard Qiagen DNeasy Plant Mini Kit protocol. DNA was quantified using a Qbit 4 Fluorometer with the dsDNA HS chemistry kit, and a quality check performed on an Agilent TapeStation. All samples were normalized to 2 ng/uL before fragmentation step. Fresh and recent historical specimens were fragmented to 350bp using a Covaris M220 Focused-ultrasonicator. Library preparation followed the protocol of the NebNext Ultra II DNA Library Prep for Illumina Kit. Seramag Sample Purification beads were used for size selection of samples above 50 ng, and clean up for less concentrated samples. An Agilent Tapestation with High Sensitivity kit was used for libraries quality check. Subsequently, libraries have been normalised to 10 nM, then pooled according to fragment size and quality. Three pools of 10, 14, and 19 libraries were made. The hybridization step followed the MyBaits Hybridization Capture for Targeted NGS Manual version 4.01. According to the guidelines of the manual relating to degraded or contaminated DNA libraries, the hybridization time was extended to 24 hours with a temperature of 62°C. 16 post-amplification cycles were performed on all the samples. Pools were sequenced by Edinburgh Genomics on a single lane of NovaSeq600 SP with 250 paired end reads. Details of these samples are in preparation for publication by T. Michel and C Kidner.

An overview of the hybridisations used is in Table 1. Comparisons of the capture by sample and by bait are in Figures 2 and 3.

Table 1. Capture experiments carried out using the Begonia bait set.

	HAIR	**PNG**	**COEL**	**POP**
Reference	Forrest et al., 2019	H Wilson PhD thesis 2021	Y-H Tseng in prep	T Michel in prep
Number of sections	2	13	3	2
Material	Fresh, silica, dried, herbarium	Fresh, silica, dried, herbarium	Fresh, silica	Fresh, herbarium
DNA prep	Qiagen	Qiagen	Qiagen	Qiagen
Library	NEBnext/TruSeq	NebNext Ultra II	NEBNext Ultra II	NEBnext
MyBaits kit version	4	4	3	4
Pooled sample per capture	11 - 17	4-10	8	10 - 19
Hybridisation time/temp	14 hours/60°C	16 hours/62°C	19 hours/65°C	24 hours/60°C
Post capture PCR cycles	13 - 20	9 - 22	11 - 12	16
Sequencing	MiSeq 250bp	HiSeq PE 150 bp	HiSeq PE 250 bp	NanoSeq6000 150bp
Max. Capture efficiency (Bowtie2 to baits)	86%	72%	85%	52%
Capture efficiency (average coverage)	634.9	105.0	206.9	437.3
Sample number (species, individuals)	45 (3,3)	191 (152,170)	67 (67,67)	43 (4, 35)

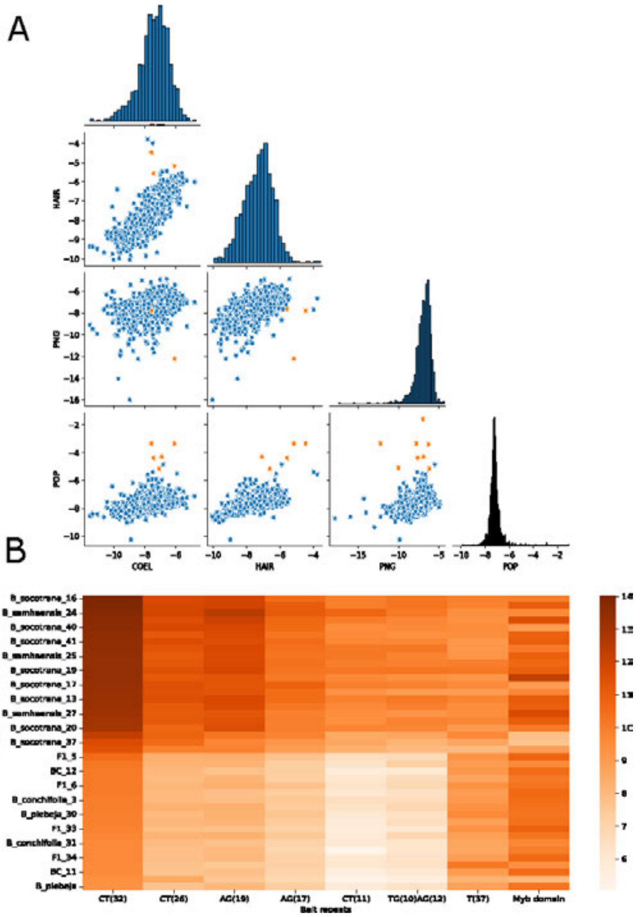


Figure B.2: Variation in capture by target across experiments. A, Log percentage read capture per target per experiment. Orange data points are the eight targets with exceptionally high mean capture rates in experiment POP. B, Heat map of log read capture for experiment POP for the eight targets with exceptionally high mean capture rates.

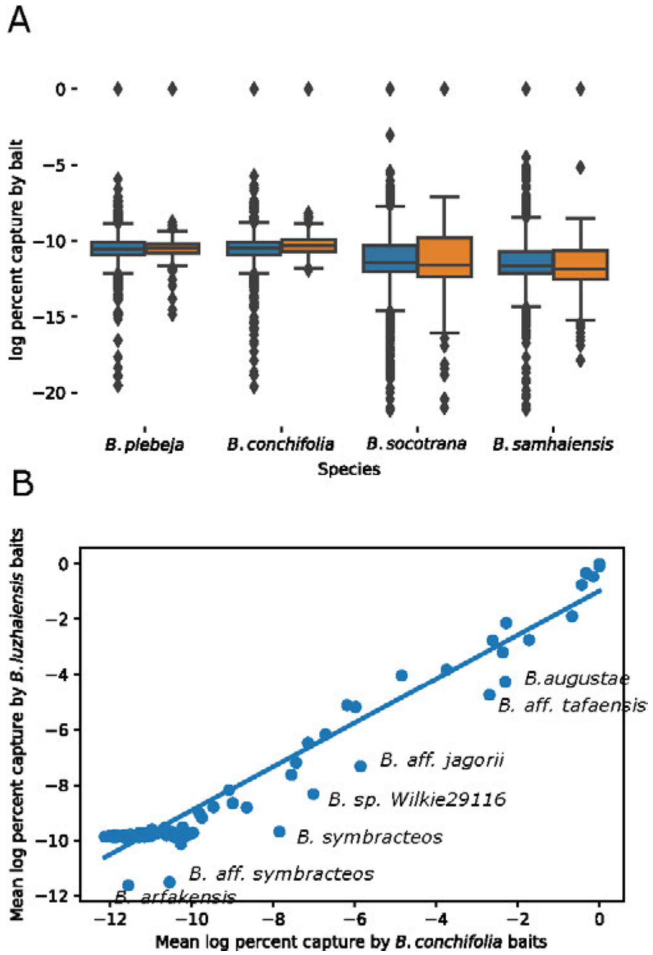


Figure B.3: Capture by phylogenetic distance between target and sample. A, Log percentage read capture per target per species (data set POP). Blue: baits designed on *Begonia luzhaiensis*. Orange: baits designed on *Begonia conchifolia*. B, Mean log percentage read capture per target per species for data set PNG. Samples that show less capture by baits from Begonia conchifolia are labelled below the line.

B.2.3 Pipeline comparisons

We compared five pipelines for generation of consensus sequence from captured reads on a subset of data from Wilson (2021) (BASIC (Nicholls *et al.*, 2015, PALEOMIX (PAL) (Schubert *et al.*, 2014), HybPiper (HP) (Johnson *et al.*, 2016), SECAPR (Andermann *et al.*, 2018) and HybPhyloMaker (HPM) (Fér & Schmickl, 2018)). Pipelines were run on the Crop Diversity server (funded by BBSRC BB/S019669/1), except for HybPhyloMaker which was run on the server of the Biodiversity Research Center, Academia Sinica.

The data we chose to compare the pipelines is from the PNG dataset of Wilson (2021), a group of 47 samples of *Begonia* section *Symbegonia* (Warb.) L.L.Forrest & Hollingsw., a small section endemic to New Guinea. This data set was chosen to allow comparison of technical replicates with samples from the same populations and from closely related species. We have data for 15 species, with more than one sample for eight of these and eleven technical replicates. The choice of technical replicates are not optimal, but rather cases where low yield required the generation of a second set of data, or where a silica sample was used to confirm results from a herbarium sample.

The BASIC pipeline has been used for *Inga* Mill, *Begoni*a and *Ceiba* Mill. (Nicholls *et al.*, 2015; Hart *et al.*, 2016; Forrest *et al.*, 2019; Pezzini, 2019). It uses a very conservative approach to align reads to the bait sequences using Bowtie2, using samtools and bcftools (Li *et al.*, 2009, Li & Durban 2010). We followed the method described in Nicholls *et al.*, (2015). Reads were cleaned with Trimmomatic v0.30 (Bolger *et al.*, 2014). Bowtie2 v2.0.2 (Langmead & Salzberg, 2012), was used to align the reads back to bait sequences with an alignment score parameter of 140 to minimise mapping of paralogs. A VCF file was generated using bcftools and filtered for quality score of >36 and to remove indels. A consensus sequence was generated with a custom Perl script (https://github.com/ckidner/Targeted_enrichment.git) and ambiguity codes converted to Ns.

HybPiper (Johnson *et al.*, 2016) maps reads to reference target sequences using BWA (Li & Durbin, 2009). BLASTx (Altschul *et al.*, 1990) then extracts the reads which map to each locus using samtools (Li *et al.*, 2009) and performs a de-novo assembly for each locus using SPAdes (Bankevich *et al.*, 2012). It then removes target flanking regions using exonerate (Slater & Birney, 2005) and picks the best contig to represent each locus, Alternatively, the intronerate.py script can be used to keep said flanking regions and create "supercontigs", flagging putative paralogs on the process, which can later be investigated. Our run of Hyb Piper (Johnson *et al.*, 2016) used the default settings and the script 'reads_first.py' as described in https://github.com/mossmatters/HybPiper/wiki with BWA (Li & Durbin, 2009) as aligner and SPADES for assembly (Bankevich *et al.*, 2012). The "supercontigs" were extracted for analysis with the corresponding Python script.

HybPhyloMaker (Fér & Schmickl, 2018) is a complete sequence to phylogeny pipeline which has been used for phylogenomic studies at different taxonomic levels, such as *Ranunculus* L. (Tomasello *et al.*, 2020), Asteraceae (Jones *et al.*, 2019), and Zingiberales (Carlsen *et al.*, 2018). Our run of the HybPhyloMaker pipeline (Fér & Schmickl, 2018) used the following settings: reads were cleaned with Trimmomatic v0.30 (Bolger *et al.*, 2014); mapped to the 'pseudoreference' based on the bait sequences using BWA (Li & Durbin, 2009); consensus sequence per locus generated with minimum relative abundance of the alternative base ("plurality" in the setting file of HybPhyloMaker) of 0.3; maximum number of heterozygous sites per exon of four; a minimum read coverage for ambiguity calling ("mincov") of 10; 51% majority consensus for base calling using Kindel v.0.1.4 (Constantinides & L. Robertson, 2017).

SECAPR has been used to analyse datasets including Palms (Helmstetter *et al.*, 2020), *Alchemilla* L. (Morales-Briones *et al.*, 2018) and Ochnaceae DC.(Schneider *et al.*, 2021). In this study it was run with default settings. Trimmed and cleaned reads were de-novo assembled and contigs matching target regions were extracted using the original bait sequence as a reference. Sequence alignments were built using MAFFT (Katoh & Toh, 2008) for all loci present in at least 3 samples. Cleaned reads from the start of the analysis were then aligned to a file containing all alignments and a consensus produced for each locus, for each sample.

To investigate an approach designed for damaged DNA (potentially useful for herbarium samples), we used PALEOMIX (Schubert *et al.*, 2014) following the methods described online (https://paleomix.readthedocs.io/en/latest/) and generating consensus sequences per locus per sample using samtools (Li *et al.*, 2009). PALEOMIX has been widely used on animals (e.g., Frantz *et al.*, 2016; Schubert *et al.*, 2017) but less frequently on plants (Vallebueno-Estrada *et al.*, 2016). The key step is the use of MapDamage2.0 (Jónsson *et al.*, 2013) to identify the signatures of typical ancient DNA and rescale the quality scores based on this analysis. We ran PALEOMIX using default parameters for cleaning and trimming reads and with BWA (Li & Durbin, 2009) for aligning reads to the reference bait sequences. PCR duplicates were marked and removed. MapDamage2 (Jónsson *et al.*, 2013) was used to recalibrate the BAM files in order to reduce the errors related to aDNA damage patterns. BCFtools was used to call variants, to normalize indels, filter adjacent indels within 5bp, and call consensus sequences for each locus for each sample (Li & Durban 2009.

A consensus per locus per sample was derived from each pipeline. The multifastas were re-arranged by locus rather than by sample using a custom python script (https://github.com/ckidner/Targeted_enrichment.git), aligned using MAFFT v7.475 (Katoh & Toh, 2008) and trimmed using trimAl v1.4.rev15 with strict settings (Capella-Gutiérrez *et al.*, 2009).

The alignments of each target were concatenated using AMAS (Borowiec, 2016). Phylogenies were produced for the concatenated matrix using IQ-TREE multicore version 2.1.2 (parameters: -B 1000 -m MFP+MERGE -alrt 1000) (Nguyen *et al.*, 2014). We portioned the analysis by target locus except for SECAPR and BASIC, where there was too much missing data to allow this. Using our standard pipeline, individual trees for each target were generated using FastTree version 2.1.10 (Price *et al.*, 2010) to produce a species tree with ASTRAL.5.7.7 (Zhang *et al.*, 2018). Metrics were collected using AMAS (Borowiec, 2016) and IQ_TREE (Nguyen *et al.*, 2014) (Table 2). We used phyparts to analyse gene tree bipartitions (Smith *et al.*, 2015) and ETE3 to analyse Robinson Foulds (RF) distances (Huerta-Cepas *et al.*, 2016). Phylogenies were visualised using FigTree v.1.4.4 (Rambaut, 2012: https://github.com/rambaut/figtree/releases/tag/v1.4.4) and ETE3 following directions in https://github.com/mossmatters/MJPythonNotebooks/blob/master/PhyParts_PieCharts.ipynb to visualise bipartition analysis results. Tree comparisons used hierarchical clustering analysis carried out in scipy.cluster.hierarchy on the RF values and SciPy Version: 1.7.1 and tree space (Jombart *et al.*, 2017) using Kendall-Colijn distances (Kendall & Colijn, 2016).

B.3 Results

B.3.1 Design of the bait set

We have produced a bait set which works across *Begonia* and includes sequences for genes likely to be of interest in the genus. We started with likely single copy genes from a transcriptome produced from the Asian species *B. luzhaiensis* (Tseng *et al.*, 2017)

We were concerned that the baits ought not to be anonymous as for many downstream analyses it is important to understand the function of the genes used. At the time we were designing the baits the closest related species with a well annotated genome available was cucumber (*Cucumis sativus* L.) so we limited target sequences to those annotated in the cucumber genome assembly (Yang *et al.*, 2012) (Fig.B.1).

We wished to link the bait set to work already done and maximise our ability to use it for functional studies. We added sequences which matched markers used to generate the first *Begonia* genetic map (Brennan *et al.*, 2012) along with genes linked to QTLs from an analysis of species-level variation (Twyford *et al.*, 2014) and a DESeq analysis (Emelianova & Kidner 2021) and sequences from transcription factors differentially expressed between *B. conchifolia* and *B. plebeja* Liebm. (Emelianova, 2018).

Most *Begonia* are shade adapted and we wished to allow sequence analysis of key genes in the pathways of light perception and response. We added the *B. luzhaiensis* T.C.Ku orthologs of genes associated with light responses, in particular shade tolerance. Based on performance in the *Coelocentrum* capture (Table 1), we refined the bait set, removing baits which captured many paralogs, or which captured poorly and replacing them with sequences of developmental genes and matches to the angiosperm353 baits. We identified thirteen target sequences with high paralogy and eighty three which failed to capture. These sequences were removed from the set and replaced with sequences of several more developmental genes sequences matching the angiosperm353 bait set (Johnson *et al.*, 2019). This set was used for the three further captures which generated the data analysed here.

B.3.2 Comparison of hybridisation protocols

To determine how well the bait set works we compare four captures using the initial set and the revised set of baits (Table 1.). All the captures worked although there was considerable variation depending on the quality of the sample as also reported in Forrest *et al.*, 2019) (Table 1). In particular this affected very poor herbarium samples in the POP set (four with <10ng DNA input) and the PNG set (eight with <10ng input DNA).

B.3.3 Variation between baits

Fig.B.2 shows the log percentage of reads captured for each target compared across experiments. Eight of the targets (labelled ‘odd’) showed very high capture rates in the POP dataset. These eight targets did not have correspondingly high capture rates in the other experiments. Examination of these target loci revealed simple sequence repeats (SSRs) in six of them (Becon104Scf00540g0006.1;CT(32), Becon104Scf00540g0002.1;CT(26), ACmerged_contig_9951;AG(19), ACmerged_contig_1166;AG(17), ACmerged_contig_2307;CT(11), Becon104Scf01167g0029.1;TG(10)AG(12)) a poly T motif in a seventh one (ACmerged_contig_5451;T(37)), and the eighth contained a myb domain (ACmerged_contig_20957). The SSRs showed high capture rates in samples from *B. socotrana* Hook.f. and *B. samhaensis* M.Hughes & A.G.Mill. The poly-T motive and the myb domain showed high capture rates in three degraded samples from a specific species (Fig.B.2).

We wish these baits to work across *Begonia* without bias from how related the samples are to the species used for bait design. We looked to see if targets from a closely related species were captured better than targets from a more distant species. Fig.B.3shows the comparison between log percentage capture by target for species in the POP dataset. There was a greater range of capture efficiency in the targets based on sequences from *B. luzhaiensis* as there were many more targets (1192 compared to 47 *B. conchifolia* targets) but the baits designed from *B. conchifolia* targets did not capture better than the ones derived from *B. luzhaiensis*, even in *B. conchifolia* samples (Fig.B.3). Some species in the PNG dataset had fewer reads captured by the baits designed from *B. conchifolia* targets, but these species are not more phylogenetically distant to *B. conchifolia* than those which did not show this difference (Fig.B.3).

B.3.4 Comparison of assembly pipelines

We used sequence data from *Begonia* sect. *Symbegonia* generated as part of the PNG data set to compare the performance of five approaches to assembling consensus sequences from the captured reads. We chose to compare HybPiper (Johnson *et al.*, 2019), HybPhyloMaker (Fér & Schmickl, 2018), SECAPR (Andermann *et al.*, 2018), PALEOMIX (Schubert *et al.*, 2014) and the basic pipeline we had previously used on *Inga* Mill., (referred to here as BASIC) (Nicholls *et al.*, 2015). We compare ease of installation and use, and the amount and consistency of the results produced.

The BASIC pipeline was simple to set up and very fast. It generated a highly conservative consensus with relatively high proportions of missing data due to ambiguous (heterozygous) sites and indels being removed.

HybPiper (Johnson *et al.*, 2019) had the advantage of excellent instructions for installing and running (https://github.com/mossmatters/HybPiper/), but required very specific formatting of input files and extensive memory space due to requiring unzipped read files and generation of many interim files. Several samples needed to be re-run as initial memory settings on the cluster (32G) were insufficient. Two paralogs (or allelic variants) were identified for most targets.

HybPhyloMaker (Fér & Schmickl, 2018) is also supported by an informative github (https://github.com/tomas-fer/HybPhyloMaker) and is relatively simple to set up and run. The pipeline performs all steps of Hyb-seq data analysis from raw reads to species tree reconstruction, calculates and summarizes the alignment and gene tree, and implements several species tree reconstruction methods. The preparation steps for renaming the raw reads and folder structure are more time consuming in HybPhyloMaker than in other pipelines, but as it covers the whole process from raw-reads to phylogeny the time spent is worthwhile.

SECAPR is semi-automated and designed to be as easy as possible for new users, as such can be installed using conda (Andermann *et al.*, 2018). Individual consensus sequences generated using SECAPR were shorter than using the other pipelines possibly due to the two step process of de-novo assembly followed by mapping to the assembled contigs, so the final length of consensus sequences is dependant on the success of the de-novo assembly as well as the mapping step.

To investigate an approach designed for damaged DNA, which may be useful for herbarium samples we also used the bam pipelines of PALEOMIX (Schubert *et al.*, 2014) to process bams from BWA alignment (Li & Durbin, 2009), followed by a basic consensus calling using samtools and bcftools (Li *et al.*, 2009, Li & Durban 2010). This approach was fast and made no excessive memory demands.

We recovered extensive sequence using all approaches (Fig.B.4 Table 2). The Paleomix pipeline generated the most data and the BASIC pipeline the least (Table 2). The BASIC pipeline removes all ambiguous sites and all indels, whereas the Paleomix pipeline assembles as long a sequence as possible using the reads which map to the reference. The BASIC pipeline also produced a concatenated matrix with the fewest number of PI sites and a high proportion of constant sites (Table 2). Paleomix produces the targets with the highest distribution of PI sites (Figure 4C). Both HybPiper and HybPhyloMaker only output sequences where a certain amount of data is present, therefore some targets are missing sequences for samples with poor data (Fig.B.4). Many of the SECAPR sequences are short (Fig.B.4) and as expected recovered targets have a low number of PI sites.

Table 2. Alignment metrics per pipeline (post trimming).

Metric	BASIC	HybPiper	HPM	SECAPR	Paleomix
Length of alignment (bp)	1806453	1552006	1814262	1002748	1940728
Patterns	261684	381325	194954	19133	78046
All gap/ ambiguous sites	188706	0	5518	79510	965
Phylogenetic informative sites	28110	78788	194954	42555	143386
Singletons	47654	242602	70742	53583	122762
Constant sites	1730689	1230616	1692575	906610	1674580
Samples with >50% gaps	13	16	9	0	7

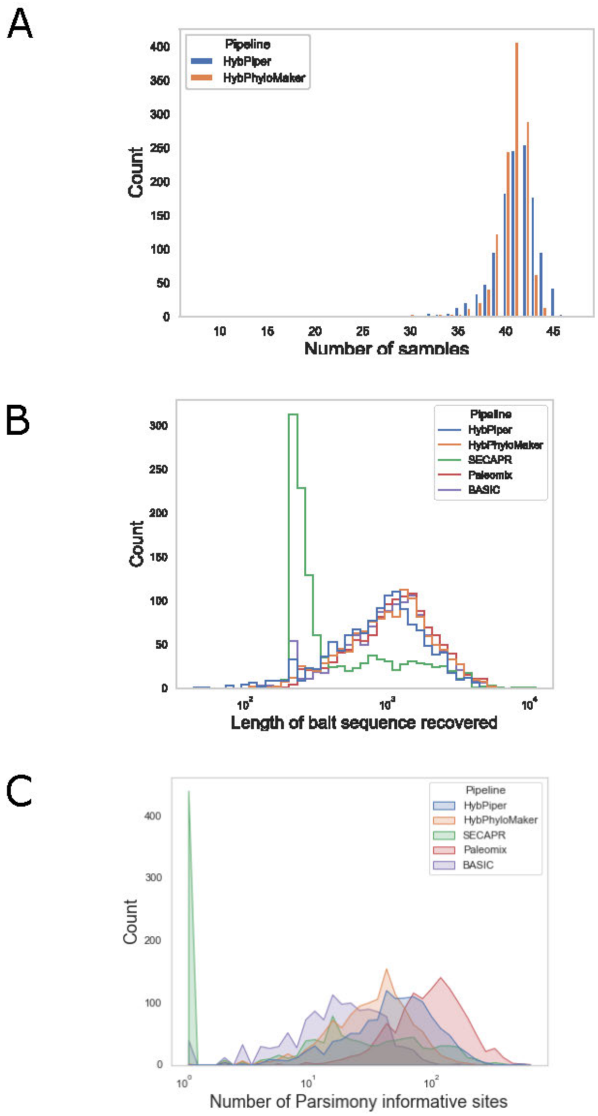


Figure B.4: Alignment metrics for 47 *Begonia* sect. *Symbegonia* samples by pipeline. A, Number of consensus sequences recovered per sample for HybPiper and HybPhyloMaker pipelines (other approaches produced data for every locus). B, Distribution of mean length of sequence recovered per target across samples using each approach. C, Distribution of number of parsimony-informative sites per target using each approach.

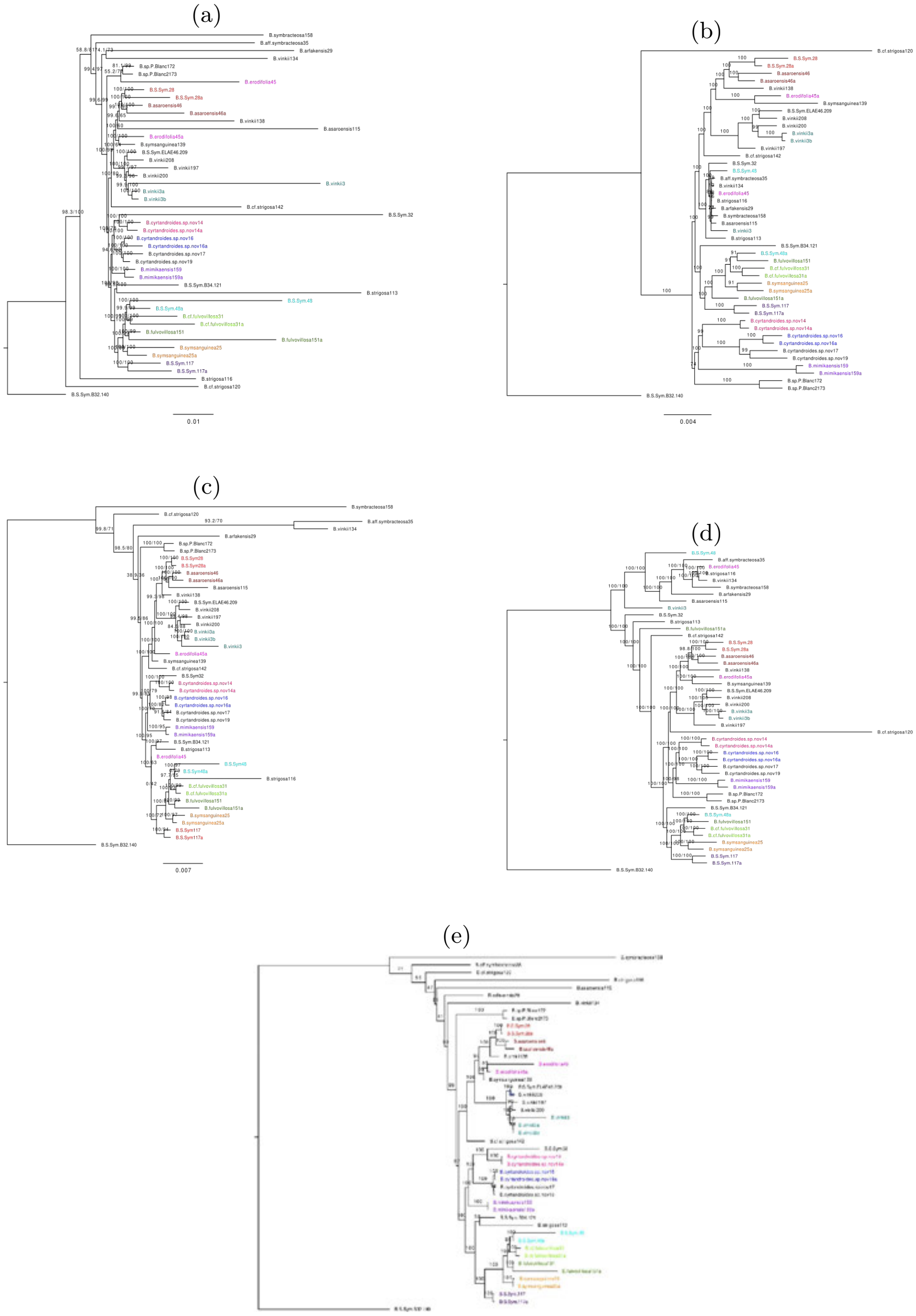


Figure B.5: Phylogenies for *Begonia* sect. *Symbegonia* samples by pipeline. IQ-TREE concatenated maximum likelihood trees for each pipeline: (a) HybPiper; (b) SECAPR; (c) HybPhyloMaker; (d) PALEOMIX; (e) BASIC.

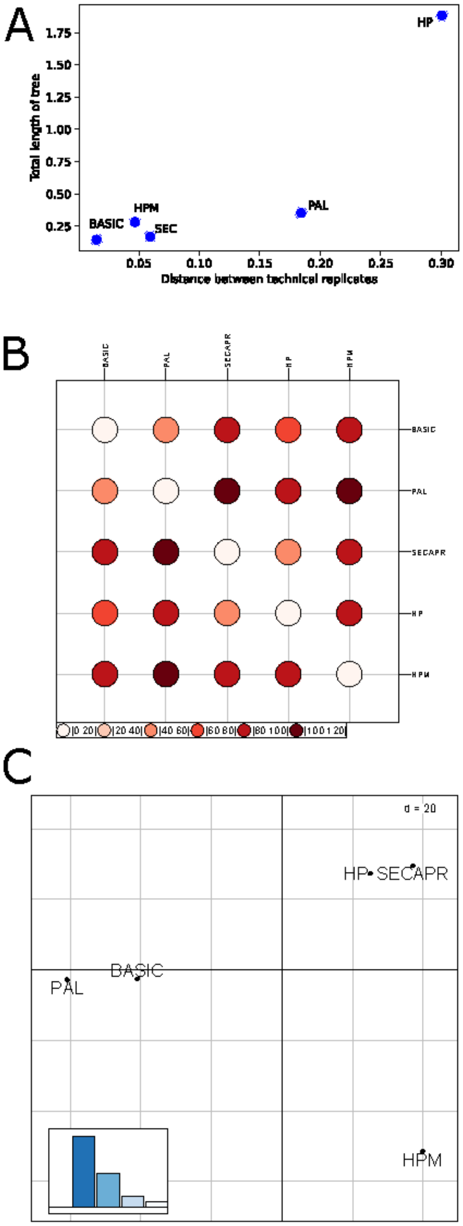


Figure B.6: Comparisons between phylogenies produced from each pipeline. A, Distance between technical replicates by total tree length for each pipeline; B, pairwise comparison of trees by Kendall– Colijn metrics; C, principal coordinates analysis of tree distances. HP, HybPiper; HPM, HybPhyloMaker; PAL, PALEOMIX; SEC, SECAPR..

The ML trees produced from the concatenated alignments vary between pipelines (Fig.B.5 Fig.B.6. HybPhyloMaker and Paleomix are least similar (RF distance 0.69) and HP and SECAPR are the most similar, despite the differences in tree length (RF distance 0.27) (Fig.B.6B and C).

The BASIC pipeline has the least distance between technical replicates and the HybPiper pipeline the most (Fig.B.6). This is likely due to the very conservative approach of the BASIC pipeline and our use of all consensus contigs from the Hyb-Piper output (to facilitate comparison with other pipelines for the consensus calling) rather than a full paralog-sensitive analysis. The placement of technical replicates as anything other than close sisters is concerning, but in this case it may be due to the poor quality and quantity of the samples for the technical replicates.

HybPiper produced the longest tree and the Basic pipeline the shortest. HybPiper produced long branches for many samples, contributing to a very long tree overall (Fig.B.5). These might be reduced by careful selection of which loci to include, removing all those identified as having paralogs. It is notable that different samples are placed on long branches by different pipelines, suggesting that the issue is not simply a high rate of gene duplications in a subset of species. This could represent random noise in the mapping leading to calling different paralogs, or sensitivity to different types of errors in HybPiper and Paleomix.

B.3.5 Variation amongst gene trees

The ASTRAL quartet score analysis and the Phyparts bipartition analysis show the very high ratio of noise to signal in our data (Fig.B.7. Some nodes in our species trees are supported by very few gene trees; in all analyses we see a minimum of 2/1239 gene trees supporting a given node, while only three nodes in each analysis are supported by a majority of genes. The well-supported nodes vary between analyses, and in many cases even the technical replicates are not supported by a majority of gene trees, suggesting that variation between gene trees of target loci reflects not only biological processes such as hybridisation and incomplete lineage sorting, but also noise and error in sequence assembly .

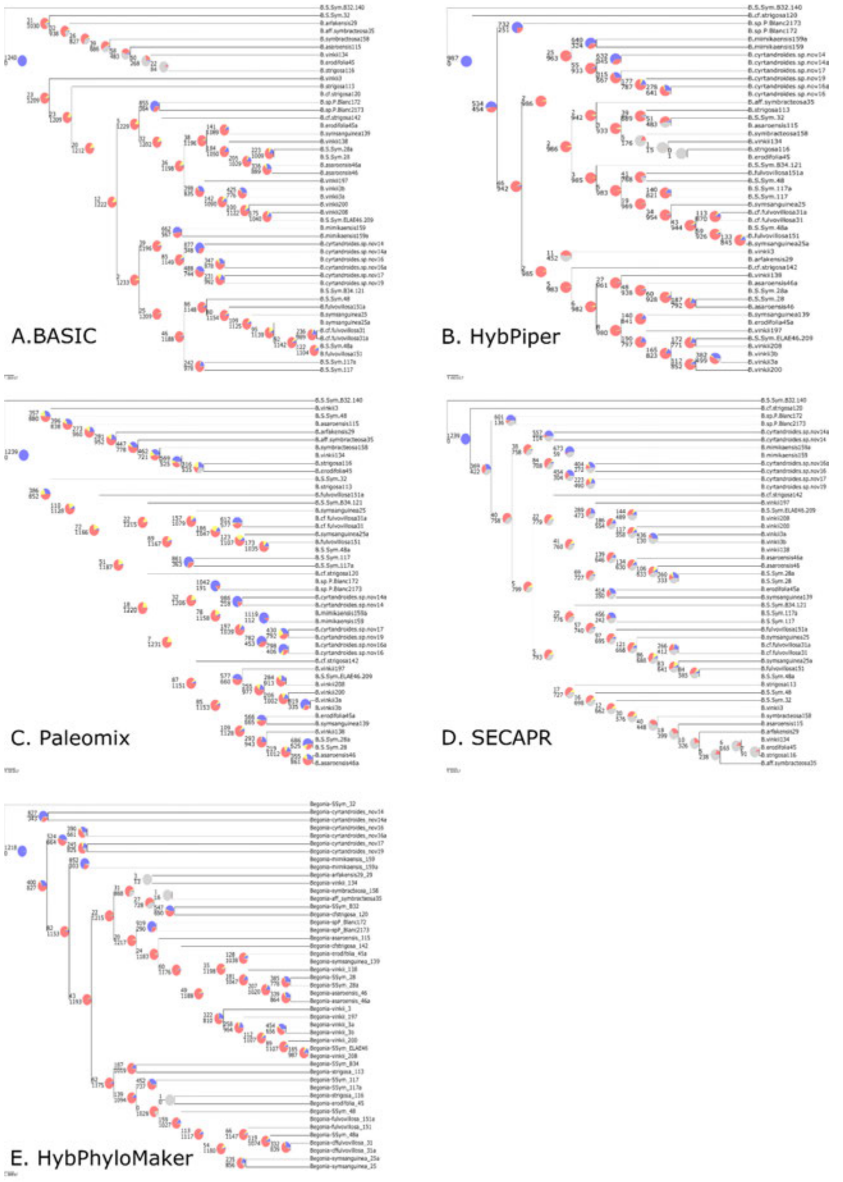


Figure B.7: ASTRAL trees with PhyParts analysis for each pipeline: A, BASIC; B, HybPiper; C, PALEOMIX; D, SECAPR; and HybPhyloMaker. Numbers at each node show the numbers of supporting gene trees over the number of conflicting gene trees. At each node is a supporting gene trees/conflicting gene trees pie chart in which blue indicates the proportion supporting the topology; yellow, the proportion supporting the next most common bipartition; red, all other conflicting gene trees; and grey, the proportion with no support for a conflicting bipartition.

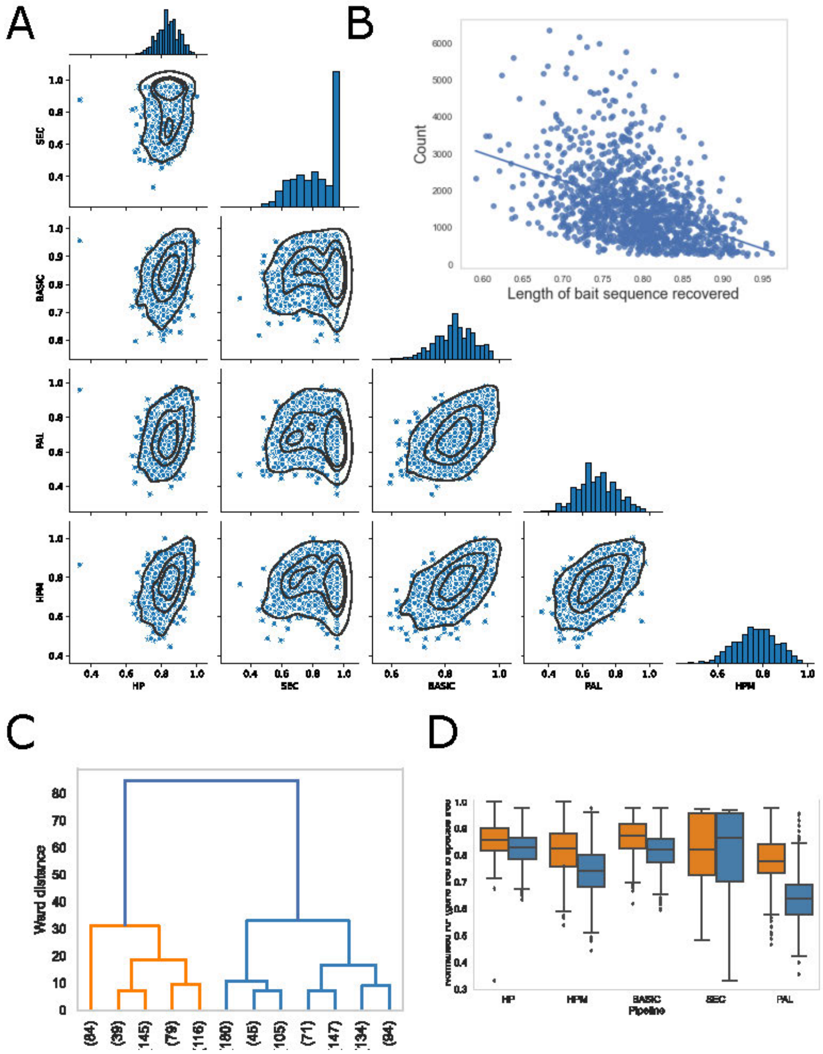


Figure B.8: Patterns across gene trees. A, Robinson–Foulds (RF) tree distances between each gene tree and the species tree for each pipeline. As the points overlap, contour lines show the point density; B, mean RF distance across pipelines and the length of the bait sequence; C, hierarchical clustering of RF distance between individual gene trees and the species tree in the PALEOMIX pipeline; D, RF distance gene tree to species tree by pipeline for baits in the two groups identified in the PALEOMIX hierarchical clustering. Group 1, orange; group 2, blue. HPM, HybPhyloMaker; PAL, PALEOMIX; SEC, SECAPR.

We investigated this further by examining the RF distance between each gene tree and the ASTRAL tree in each pipeline (Fig.B.8). SECAPR had a large number of gene trees very different to the species tree and HybPiper had a single gene tree (for target Becon104Scf03147g0009.1, a Dicer-like3 ortholog) that was a very close match to the ASTRAL tree (RF 0.33). Overall there is a positive correlation between the gene trees from each pipeline, supporting the conclusion that an overlapping set of genes are contributing to the ASTRAL species tree in each case. Loci which have a higher-than-average similarity to the species tree across all pipelines include Phytochrome A (ACcontig_8273), Phytochrome C (ACcontig_4025), and Plastid Movement Impaired (ACcontig_6745, AT5G26160.2).

Some of the similarities between pipelines in which gene trees are best related to the species tree derives from the influence of locus length. Longer loci produce genes trees which are closer to the species tree (Fig.B.8). Consensus sequences produced by SECAPR were 30% shorter than the reference bait sequences, resulting in many gene trees which individually had little phylogenetic information. The variation seen between normalised RF scores for each bait points to the influence of noise, error and variation between consensus calling in each pipeline on the phylogeny produced.

We used the data from the pipeline with highest PI per target to look for structure within the gene tree space. We calculated the reciprocal RF between all the gene trees from the Paleomix pipeline and ran a hierarchical clustering analysis (in SciPy) to determine if there were distinct groups of trees. At least two and possibly four, or more groups were detected (Fig.B.8). Group one (orange in Fig.B.8) comprises 465 trees generally distant to the species tree, group two (blue in Figure 8C) comprises 778 trees more similar to the species tree. To see if a similar structure existed in the gene trees produced by the other pipelines we used the grouping from the Paleomix analysis to codify the targets across all pipelines and plotted the normalised RF distances between gene and species trees as box plots (Fig.B.8). The group one set (the set of targets in orange in Fig.B.8) were generally more distant to the species tree for all pipelines except for SECAPR. This suggests that the groups detected in the clustering analysis reflect a true pattern and are not an artefact of a particular pipeline.

To examine the two groups of gene trees in more detail we ran an ASTRAL analysis on each group, followed by Phyparts to illustrate the gene tree support at each node (Supplemental Fig.B.1A,B). The two trees differ with an RF of 0.49 (44/90). Nine of the eleven sets of technical replicates map as sisters on the cluster 2 tree, seven with support from the majority of gene trees, but only six are sisters on the cluster 1 tree, and of these three have support from the majority of gene trees. This suggests that cluster 1 contains trees with more noise and error than cluster 2.

B.4 Discussion

B.4.1 Why use a target capture approach in Begonia?

The results presented here show that the *Begonia* bait set resolves species-level phylogenies well and the large number of variants identified supports its use on a population level. However, sequencing costs are decreasing to the point that it is feasible to use genome skims for phylogenetics, which begs the question: why deal with the extra lab time and expense of hybrid capture? For some questions genome skims are a useful approach, but the complexity and large size of plant genomes means that skims are an inefficient way of gathering functional genetic data. The prevalence of gene family expansions and partial and whole genome duplications in plants is also easier to deal with for a limited set of loci than with skim data.

With over 2000 *Begonia* species few living collections hold anywhere near a representative collection needed for genomic studies. In addition, field collection of samples is hampered by the remote and often difficult terrain in which many species are found. This has meant that *Begonia* phylogenetic studies often have to rely on data from herbarium collections. For such old, degraded DNA, hybrid capture represents the only way to extract reliable sequence data across the nuclear genome (Hart *et al.*, 2016), and the *Begonia* baits set has been shown to retrieve useful data from even very poor herbarium samples (Forrest *et al.*, 2019).

Current phylogenetic studies on *Begonia* indicate a large number of rapid radiations and likely hybridisation events (Goodall-Copestake *et al.*, 2010; Thomas *et al.*, 2012; Moonlight *et al.*, 2015; Tseng *et al.*, 2017; Liu *et al.*, 2019a). The large number of unlinked SNPs recovered from hybrid capture (Fig.B.4B and C) is ideal for analysis of rapid radiations and reticulate lineages (Shee *et al.*, 2020; Thomas *et al.*, 2021) and offers our best hope of understanding the complexities of *Begonia* evolution.

Studies have shown that broad baits sets such as the angiosperms353 can resolve recent radiations and have the potential for population genetic analysis even in the case of polyploids (Kates *et al.*, 2018; Larridon *et al.*, 2019; van Andel *et al.*, 2019; Melichárková *et al.*, 2020; Šlenker *et al.*, 2021; Slimp *et al.*, 2021). However, although general bait sets give excellent overlap between studies, providing useful data matrices, they can capture less efficiently than specific baits (Kadlec *et al.*, 2017; Liu *et al.*, 2019b; Larridon *et al.*, 2020). Given the numbers of *Begonia* species and the poor preservation of DNA likely in many herbarium samples it is important to focus on getting the highest proportion of usable data possible in each sequencing run. This is best provided by using a specific bait set. Such a set also has the advantage of allowing focus on particular aspects of *Begonia* biology and inclusion of genomic regions known to vary between species (as recommended in Lee *et al.*, 2021). The resolution seen in recent Coelocentraum radiation shows the utility of the Begonia bait set in this respect, but also emphasises the need to careful consideration of analysis pipelines (Fig.B.5.

Integration with previous *Begonia* phylogeny studies can be achieved through use of the off-target reads to assemble plastid, mitochondrial sequences, and nuclear repeats. This may require 'spiking' of the sequencing reactions with uncaptured libraries as a high capture efficiency, as seen in some cases here (Fig.B.2), reduces the off-target reads to levels too low for plastid genome assembly (Weitemier *et al.*, 2014)).

Only 23 of the loci in the *Begonia* bait set are also present in the angiosperms353 enrichment panel. Better integration with other studies could be achieved through combining our bait set with the angiosperms353 set in the hybridisation step. This approach has worked well in Brassicaceae (Hendriks *et al.*, 2021).

B.4.2 Library prep, hybridisation and sequencing

The four capture experiments using our bait set have shown effective capture under a range of conditions. Forrest *et al.*, (2019) showed no clear differences between NEB Next Seq and TruSeq library prep kits. Pooling is recommended to reduce costs. Here we report pooling up to 19, but up to 48 should be possible (Hale *et al.*, 2020). We suggest that the bait set is fairly robust to small changes in the hybridisation protocol and that further work to optimise may be required for the most difficult samples. Unfortunately the most difficult samples are usually those with the least material, limiting the possibilities of multiple hybridisation attempts. Arbor Biosciences suggest 65°C and 16-24 hours as the standard hybridisation conditions with lower temperatures (55°C for very fragmented samples and up to 40 hours incubation for samples with very low ratios of target to off-target DNA (Arbor Biosciences https://arborbiosci.com/wp-content/uploads/2019/08/myBaits-Manual-v4.pdf). Based on results presented here we recommend 19 hours at 60°C with 10–12 rounds of post-capture PCR as a good starting point.

Both 250bp and 150bp reads have been generated in the studies examined here. As the reads are mapped to a known reference there is little advantage to the longer length of read unless off-target sequences (either introns or organellar/ITS data) are also required. Given the depth of sequencing observed for the on-target reads we suggest that the cheapest approach be used regardless of the read length generated. The *Begonia* bait set is 1.9MB, longer than some others, such as the angiosperm353 set, so more sequencing is required to give comparable sequence depth to that obtained with shorter bait sets. With 60% capture efficiency 180 samples could be sequenced in one lane of MiSeq to 20× average coverage (good for species-level phylogenetics, although given the variation seen across loci and samples half this number (90 samples per lane) might be recommended.

B.4.3 Analysis pipelines

Of the five pipelines we trailed for calling consensus sequences from captured reads, the BASIC pipeline gave least data but also the least distance between technical replicates with all the technical replicates resolving as sisters in the concatenated tree (Table 3, Fig.B.5 Figure 6A). However, the bipartition analysis shows that even these nodes at technical replicates were not supported by all the gene trees (Figure 7). This confirms that all the pipelines tested produce errors which contribute to long branches and poor node support, despite the high level of support for each node in the ML analysis. It is possible that some of the discordance in our test set derives from hybridisation and incomplete lineage sorting in *Symbegonia*. Paleomix in particular shows quite strong support for an alternative branching pattern in several of the deeper nodes in the tree, which could be related to hybridisation early in the colonisation of New Guinea, as suggested by morphology (Wilson, 2021). We would recommend using at least two methods for deriving consensus bait sequences to allow comparison of results. We suggest one using reference-based consensus calling (such as BASIC or Paleomix) and one using a de-novo assembly step (such as HybPiper, HybPhyloMaker or SECAPR). Analysis of the gene tree support in each approach could then be used to exclude 'noisey' loci.

B.4.4 Phylogenetic approaches using Hybrid Capture data

One standard phylogenetic approach with hybrid capture data is to concatenate baits and generate a species tree using maximum likelihood (ML) analysis. A second approach infers individual gene trees from each target (also under ML) and produces a species tree from the gene trees under the multispecies coalescent (MSC) using a program like ASTRAL. Our bipartition analysis on the MSC phylogeny revealed very high variation between gene trees (Fig.B.7. Picking which loci to use for analysis and which to reject will clearly make large differences to support for particular nodes and the shape of the tree, even with the concatenated approach as a few loci can have disproportionate effects on topology (Shen *et al.*, 2017) (Supplemental Fig.B.1,B).

There is clearly variation between captures in which baits capture well and which poorly (Fig.B.2). Eight baits with some repetative sequences gave very high capture in one experiments (POP, Fig.B.2), but behaved normally in the other three experiments (Fig.B.2). There is also extensive variation between baits in the number of parsimony informative sites obtained (Fig.B.4). Patterns in bait capture need to be considered for each experiment and baits with exceptionally high, low or variable coverage excluded from further analysis. Analysis of patterns of gene trees is key to reducing noise. Phyparts can be used to reveal the variation between gene trees and species trees and this can be further explored using other software such as treespace (https://thibautjombart.github.io/treespace/) (Fig.B.7 supp Fig.B.1. This study shows that some targets have a consistently good match to the species tree across pipelines (Fig.B.8). We have used clustering analysis to show that the targets can be divided into two groups, one of which is close to the species tree and the other (which appears noisier-technical replicates are more distant) which is more distant to the species tree. It is possible that one set of targets is reflecting the ‘true tree’ and the other is capturing more paralogs so generating distorting noise. However, it is also possible that a single ‘true tree’ does not exist.

Striving to derive a single species tree from the mass of data in hundreds of loci is possibly not the best use of this data. The power of a Hybrid Capture approach lies in the ability to resolve complex evolutionary histories and we suggest the use of approaches which include incongruence amongst gene trees in the analysis. In particular, given the prevalence of hybridisation and gene duplication in the evolution of *Begonia* (Brennan *et al.*, 2012; Hughes *et al.*, 2018; Liu *et al.*, 2019a; Tseng *et al.*, 2019) it is only by acknowledging the complex and reticulate nature of the phylogenies that we will begin to understand the evolutionary patterns in this genus. Such approaches are becoming more widespread (Morales-Briones *et al.*, 2018; Harris, 2019; Gagnon *et al.*, 2021; Lee *et al.*, 2021). The published pipelines accompanying these papers make this type of analysis more accessible though the computing resources required can still be considerable for large numbers of trees and larger bait sets.

B.4.5 Further work

Our choice of loci includes a set of key developmental, physiological and stress-related genes. The depth of coverage obtained in the captures from good quality material (including many herbarium samples) is sufficient to allow sequence analysis comparing evolutionary patterns in functional genes. We hope that the wealth of data produced from capture experiments will not be limited to phylogenetic studies but will provide a greater knowledge of functional evolutionary patterns across the group.

B.5 Conclusions

We recommend this bait set for *Begonia* phylogenetics and population-level studies. Arbor Biosciences can produce a copy of this bait set within 2 weeks. It is nearly as quick to obtain as an off-the-shelf general kit such as the Angiosperms 353 set, but has the advantages of better capture, more loci and loci chosen to study issues of *Begonia* biology.

General advice on a cost-effective approach to hybrid capture protocol has been published by Hale *et al.* (2020) , making hybrid capture possible for budget constrained labs and studies using hundreds of samples. We hope that by using a coordinated approach the usefulness of the data will be increased and novel comparisons and overviews will be possible.

B.6 Acknowledgements:

The authors thank Michelle Hart and Laura Forest (RBGE) for lab help and general support, the staff of the Experimental Greenhouse and NGS High Throughput Genomic Core of BRCAS., Mark Hughes (RBGE) for *Begonia* expertise, M.L. MacIntyre Begonia Trust for funding for HW, PlantID (EU Horizon 2020 DTP 765000) for funding for TM and general support and advice, CropDiversity servers and administrators for computing support (BBSRC grant

BBSRC BB/S019669/1). Yu-Hsing Tseng acknowledges the grant from the Minister of Science and Technology, Taiwan (Grant no. 107-2917-I-564-026) for a postdoctoral grant for a visit to RBGE. Yu-Hsing Tseng and Kuo-Fang Chung were supported by research grant from Academia Sinica, Taiwan.

B.7 References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410. DOI: https://doi.org/10.1016/S0022-2836(05)80360-2

Andermann T, Cano Á, Zizka A, Bacon C, Antonelli A (2018) SECAPR-a bioinformatics pipeline for the rapid and user-friendly processing of targeted enriched Illumina sequences, from raw reads to alignments. PeerJ(6):e5175

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., andPevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology, 19(5), 455-477.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (30):2114–2120

Borowiec ML (2016) AMAS: a fast tool for alignment manipulation and computing of summary statistics. PeerJ (4):e1660

Brennan AC, Bridgett S, Shaukat Ali M, Harrison N, Matthews A, Pellicer J, Twyford AD, Kidner CA (2012) Genomic Resources for Evolutionary Studies in the Large, Diverse, Tropical Genus, *Begonia*. Tropical Plant Biology (5):1–16

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (25):1972–1973

Campos-Dominguez L (2020) Does a dynamic genome drive speciation in a mega-diverse genus? PhD Thesis University of Edinburgh

Carlsen MM, Fér T, Schmickl R, Leong-Škorničková J, Newman M, Kress WJ (2018) Resolving the rapid plant radiation of early diverging lineages in the tropical Zingiberales: Pushing the limits of genomic data. Molecular and Phylogenetic Evolution (128):55–68

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics (2008):619832 doi: 10.1155/2008/619832.

Constantinides B, L. Robertson D (2017) Kindel: indel-aware consensus for nucleotide sequence alignments. Journal of Open Source Software (2):282

Couvreur TLP, Helmstetter AJ, Koenen EJM, Bethune K, Brandão RD, Little SA, Sauquet H, Erkens RHJ (2018) Phylogenomics of the Major Tropical Plant Family Annonaceae Using Targeted Enrichment of Nuclear Genes. Frontiers in Plant Science (9):1941

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J (2012) Targeted enrichment strategies for next-generation plant biology. American Journal of Botany (99):291–311

Dodsworth S, Pokorny L, Johnson MG, Kim JT, Maurin O, Wickett NJ, Forest F, Baker WJ (2019) Hyb-Seq for Flowering Plant Systematics. Trends in Plant Science (24):887–891

Emelianova K, Martínez Martínez A, Campos-Dominguez L, Kidner C. (2021) Multi-tissue transcriptome analysis of two *Begonia* species reveals dynamic patterns of evolution in the chalcone synthase gene family. Sci Rep. 11(1):17773. doi: 10.1038/s41598-021-96854-y.

Fér T, Schmickl RE (2018) HybPhyloMaker: Target Enrichment Data Analysis From Raw Reads to Species Trees. Evolutionary Bioinformatics Online (14):1176934317742613

Folk RA, Mandel JR, Freudenstein JV (2015) A protocol for targeted enrichment of intron-containing sequence markers for recent radiations: A phylogenomic example from *Heuchera* (Saxifragaceae). Applications in Plant Sciences doi: 10.3732/apps.1500039

Forrest LL, Hart ML, Hughes M, Wilson HP, Chung K-F, Tseng Y-H, Kidner CA (2019) The Limits of Hyb-Seq for Herbarium Specimens: Impact of Preservation Techniques. Frontiers in Ecology and Evolution. doi: 10.3389/fevo.2019.00439

Frantz LAF, Mullin VE, Pionnier-Capitan M, Lebrasseur O, Ollivier M, Perri A, Linderholm A, Mattiangeli V, Teasdale MD, Dimopoulos EA, et al (2016) Genomic and archaeological evidence suggest a dual origin of domestic dogs. Science (352):1228–1231

Frodin DG (2004). History and concepts of big plant genera. *Taxon* 53(3):753–776.

Gagnon E, Hilgenhof R, Orejuela A, McDonnell A, Sablok G, Aubriot X, Giacomin L, Gouvêa Y, Bragionis T, Stehmann JR, Bohs L, Dodsworth S, Martine C, Poczai P, Knapp S, Särkinen T. Phylogenomic discordance suggests polytomies along the backbone of the large genus Solanum. Am J Bot. 2022 Feb 16. doi: 10.1002/ajb2.1827. Epub ahead of print. PMID: 35170754.

Gardiner L-J, Brabbs T, Akhunov A, Jordan K, Budak H, Richmond T, Singh S, Catchpole L, Akhunov E, Hall A (2019) Integrating genomic resources to present full gene and putative promoter capture probe sets for bread wheat. Gigascience. doi: 10.1093/gigascience/giz018

Goodall-Copestake WP, Pérez-Espona S, Harris DJ, Hollingsworth PM (2010) The early evolution of the mega-diverse genus *Begonia* (Begoniaceae) inferred from organelle DNA phylogenies. Biologial Journal of the Linnean Society, London (101):243–250

Hale H, Gardner EM, Viruel J, Pokorny L, Johnson MG (2020) Strategies for reducing per-sample costs in target capture sequencing for phylogenomics and population genomics in plants. Applications in Plant Science (8):e11337

Harris K (2019) From a database of genomes to a forest of evolutionary trees. Nature Genetics (51):1306–1307

Hart ML, Forrest LL, Nicholls JA, Kidner CA (2016) Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon (65):1081–1092

Helmstetter AJ, Kamga SM, Bethune K, Lautenschläger T, Zizka A, Bacon CD, Wieringa JJ, Stauffer F, Antonelli A, Sonké B, et al (2020) Unraveling the Phylogenomic Relationships of the Most Diverse African Palm Genus *Raphia* (Calamoideae, Arecaceae). Plants. doi: 10.3390/plants9040549

Hendriks KP, Mandáková T, Hay NM, Ly E, Hooft van Huysduynen A, Tamrakar R, Thomas SK, Toro-Núñez O, Pires JC, Nikolov LA, Koch MA, Windham MD, Lysak MA, Forest F, Mummenhoff K, Baker WJ, Lens F, Bailey CD.2021) The best of both worlds: Combining lineage-specific and universal bait sets in target-enrichment hybridization reactions. Applications in Plant Science. doi: 10.1002/aps3.11438

Hill CB, Wong D, Tibbits J, Forrest K, Hayden M, Zhang X-Q, Westcott S, Angessa TT, Li C (2019) Targeted enrichment by solution-based hybrid capture to identify genetic sequence variants in barley. Scientific Data (6):12

Huerta-Cepas J, Serra F, Bork P (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Molecular Biology and Evolution (33):1635–1638

Hughes M, Peng C-I, Lin C-W, Rubite RR, Blanc P, Chung K-F (2018) Chloroplast and nuclear DNA exchanges among *Begonia* sect. Baryandra species (Begoniaceae) from Palawan Island, Philippines, and descriptions of five new species. PLoS One (13):e0194877

Johnson MG, Malley C, Goffinet B, Shaw AJ, Wickett NJ (2016) A phylotranscriptomic analysis of gene family expansion and evolution in the largest order of pleurocarpous mosses (Hypnales, Bryophyta). Molecular Phylogenetics and Evolution (98):29–40

Johnson MG, Pokorny L, Dodsworth S, Botigué LR, Cowan RS, Devault A, Eiserhardt WL, Epitawalage N, Forest F, Kim JT, et al (2019) A Universal Probe Set for Targeted Sequencing of 353 Nuclear Genes from Any Flowering Plant Designed Using k-Medoids Clustering. Systematic Biology (68):594–606

Jombart T, Kendall M, Almagro-Garcia J, Colijn C. (2017) treespace: Statistical exploration of landscapes of phylogenetic trees. Mol Ecol Resour (6):1385-1392

Jones KE, Fér T, Schmickl RE, Dikow RB, Funk VA, Herrando-Moraira S, Johnston PR, Kilian N, Siniscalchi CM, Susanna A, Slovák M, Thapa R, Watson LE, Mandel JR. (2019) An empirical assessment of a single family-wide hybrid capture locus set at multiple evolutionary timescales in Asteraceae. Applications in Plant Science (7):e11295

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L (2013) mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics (29):1682–1684

Kadlec M, Bellstedt DU, Le Maitre NC, Pirie MD (2017) Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? PeerJ (5):e3569

Kates HR, Johnson MG, Gardner EM, Zerega NJC, Wickett NJ (2018) Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus.* American Journal of Botany (105):404–416

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics (9):286–298

Kendall M, Colijn C, (2016) Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution, Molecular Biology and Evolution, (33):2735–2743,

Koenen EJM, Ojeda DI, Steeves R, Migliore J, Bakker FT, Wieringa JJ, Kidner C, Hardy OJ, Pennington RT, Bruneau A, et al (2020) Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. New Phytologist (225):1355-1369

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods (9):357–359

Larridon I, Villaverde T, Zuntini AR, Pokorny L, Brewer GE, Epitawalage N, Fairlie I, Hahn M, Kim J, Maguilla E, et al., (2020) Tackling Rapid Radiations With Targeted Sequencing. Frontiers in Plant Science (10):1655

Lee AK, Gilman IS, Srivastav M, Lerner AD, Donoghue MJ, Clement WL (2021) Reconstructing Dipsacales phylogeny using Angiosperms353: issues and insights. Am J Bot 108: 1122–1142

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (25):1754–1760

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):2078–2079

Liu S-H, Tseng Y-H, Zure D, Rubite RR, Balangcod TD, Peng C-I, Chung K-F (2019a) *Begonia balangcodiae* sp. nov. from northern Luzon, the Philippines and its natural hybrid with *B. crispipila*, *B. × kapangan* nothosp. nov. Phytotaxa (407):5–21

Liu Y, Johnson MG, Cox CJ, Medina R, Devos N, Vanderpoorten A, Hedenäs L, Bell NE, Shevock JR, Aguero B, et al (2019b) Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. Nature Communications (10):1–11

Mandel JR, Dikow RB, Funk VA (2015) Using phylogenomics to resolve mega-families: An example from Compositae. Journal of Systematics and Evolution (53):391–402

McKain MR, Johnson MG, Uribe-Convers S, Eaton D, Yang Y (2018) Practical considerations for plant phylogenomics. Applications in Plant Sciences (6):e1038

Melichárková A, Šlenker M, Zozomová-Lihová J, Skokanová K, Šingliarová B, Kačmárová T, Caboňová M, Kempa M, Šrámková G, Mandáková T, Lysák MA, Svitok M, Mártonfiová L, Marhold K. (2020) So Closely Related and Yet So Different: Strong Contrasts Between the Evolutionary Histories of Species of the *Cardamine pratensis* Polyploid Complex in Central Europe. Frontiers in Plant Sciences (11):588856

Moonlight PW, Ardi WH, Padilla LA, Chung K-F, Hughes M (2018) Dividing and conquering the fastest-growing genus: Towards a natural sectional classification of the mega-diverse genus *Begonia* (Begoniaceae). Taxon. doi: 10.12705/672.3

Moonlight PW, Richardson JE, Tebbitt MC (2015) Continental-scale diversification patterns in a megadiverse genus: the biogeography of Neotropical *Begonia.* Journal of Biogeography (42):1137-1149

Morales-Briones DF, Liston A, Tank DC (2018) Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). New Phytologist (218)):1668–1684

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2014) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution (32):268–274

Nicholls JA, Pennington RT, Koenen EJM, Hughes CE, Hearn J, Bunnefeld L, Dexter KG, Stone GN, Kidner CA (2015) Using targeted enrichment of nuclear genes to increase phylogenetic resolution in the neotropical rain forest genus *Inga* (Leguminosae: Mimosoideae). Fronteirs in Plant Science (6):710

Pezzini FF (2019) Phylogeny, taxonomy and biogeography of *Ceiba* Mill. (Malvaceae: Bombacoideae). PhD Thesis, The University of Edinburgh

Price MN Dehal PS and Arkin AP (2010) FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 5(3):e9490. doi:10.1371/journal.pone.0009490.

Schneider JV, Jungcurt T, Cardoso D, Amorim AM, Töpel M, Andermann T, Poncy O, Berberich T, Zizka G (2021) Phylogenomics of the tropical plant family Ochnaceae using targeted enrichment of nuclear genes and 250+ taxa. Taxon (70):48–71

Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, 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):1056–1082

Schubert M, Mashkour M, Gaunitz C, Fages A, Seguin-Orlando A, Sheikhi S, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, Chuang R, et al (2017) Zonkey: A simple, accurate and sensitive pipeline to genetically identify equine F1-hybrids in archaeological assemblages. Journal of Archaeological Science (78):147–157

Shee ZQ, Frodin DG, Cámara-Leret R, Pokorny L. (2020). Reconstructing the Complex Evolutionary History of the Papuasian *Schefflera* Radiation Through Herbariomics. *Front Plant Sci.* 11:258. doi:10.3389/fpls.2020.00258

Shen X-X, Hittinger CT, Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nature Ecology and Evolution (1):126

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC bioinformatics, 6(1), 1-11. DOI: https://doi.org/10.1186/1471-2105-6-31

Šlenker M, Kantor A, Marhold K, Schmickl R, Mandáková T, Lysak MA, Perný M, Caboňová M, Slovák M, Zozomová-Lihová J (2021) Allele Sorting as a Novel Approach to Resolving the Origin of Allotetraploids Using Hyb-Seq Data: A Case Study of the Balkan Mountain Endemic Cardamine barbaraeoides. Fronteris in Plant Science (12):659275

Slimp M, Williams LD, Hale H, Johnson MG (2021) On the potential of Angiosperms353 for population genomic studies. Applications in Plant Science. doi: 10.1002/aps3.11419

Smith SA, Moore MJ, Brown JW, Yang Y (2015) Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. BMC Evolutionary Biology (15):150

Soto Gomez M, Pokorny L, Kantar MB, Forest F, Leitch IJ, Gravendeel B, Wilkin P, Graham SW, Viruel J (2019) A customized nuclear target enrichment approach for developing a phylogenomic baseline for *Dioscorea* yams (Dioscoreaceae). Applications in Plant Science (7):e11254

Thomas, A. E., Igea, J., Meudt, H. M., Albach, D. C., Lee, W. G., and Tanentzap, A. J.. 2021. Using target sequence capture to improve the phylogenetic resolution of a rapid radiation in New Zealand *Veronica. American Journal of Botany* 108(7): 1289– 1306.

Thomas DC, Hughes M, Phutthai T, Ardi WH, Rajbhandary S, Rubite R, Twyford AD, Richardson JE (2012) West to east dispersal and subsequent rapid diversification of the mega-diverse genus *Begonia* (Begoniaceae) in the Malesian archipelago. Journal of Biogeography (39):98–113

Tomasello S, Karbstein K, Hodač L, Paetzold C, Hörandl E (2020) Phylogenomics unravels Quaternary vicariance and allopatric speciation patterns in temperate-montane plant species: A case study on the *Ranunculus auricomus* species complex. Mol Ecol 29: 2031–2049

Tseng Y-H, Huang H-Y, Xu W-B, Yang H-A, Liu Y, Peng C-I, Chung K-F (2017) Development and characterization of EST-SSR markers for *Begonia luzhaiensis* (Begoniaceae). Applications in Plant Science. doi: 10.3732/apps.1700024

Tseng Y-H, Huang H-Y, Xu W-B, Yang H-A, Peng C-I, Liu Y, Chung K-F (2019) Phylogeography of Begonia luzhaiensis suggests both natural and anthropogenic causes for the marked population genetic structure. Botanical Studies (60):20

Twyford AD, Ennos RA, White CD, Ali MS, Kidner CA (2014) The evolution of sex ratio differences and inflorescence architectures in *Begonia* (Begoniaceae). American Journal of Botany (101):308–317

Vallebueno-Estrada M, Rodríguez-Arévalo I, Rougon-Cardoso A, Martínez González J, García Cook A, Montiel R, Vielle-Calzada J-P (2016) The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. Proceedings of the Natural Academy of Science U S A (113):14151–14156

Van Andel T, Veltman MA, Bertin A, Maat H, Polime T, Hille Ris Lambers D, Tjoe Awie J, De Boer H, Manzanilla V. Hidden Rice Diversity in the Guianas. Front Plant Sci. 2019 Sep 20;10:1161. doi: 10.3389/fpls.2019.01161. PMID: 31616452; PMCID: PMC6764085.

Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M.G., Gardner, E.M., Wickett, N.J., Molero, J., Riina, R. and Sanmartín, I. (2018), Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. New Phytol, 220: 636-650. https://doi.org/10.1111/nph.15312

Weitemier K, Straub SCK, Cronn RC, Fishbein M, Schmickl R, McDonnell A, Liston A (2014) Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Applications in Plant Science. doi: 10.3732/apps.1400042

Wilson H, 2021 Megadiversity and the New Guinea Orogeny. PhD thesis, University of Glasgow

Yang L, Koo D-H, Li Y, Zhang X, Luan F, Havey MJ, Jiang J, Weng Y (2012) Chromosome rearrangements during domestication of cucumber as revealed by high-density genetic mapping and draft genome assembly. The Plant Journal (71):895–906

Zhang C, Rabiee M, Sayyari E, Mirarab S (2018) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics (19):153

Appendix C

F-statistics applied to the PNG dataset

C.1 F_{ST} analysis of the PNG species

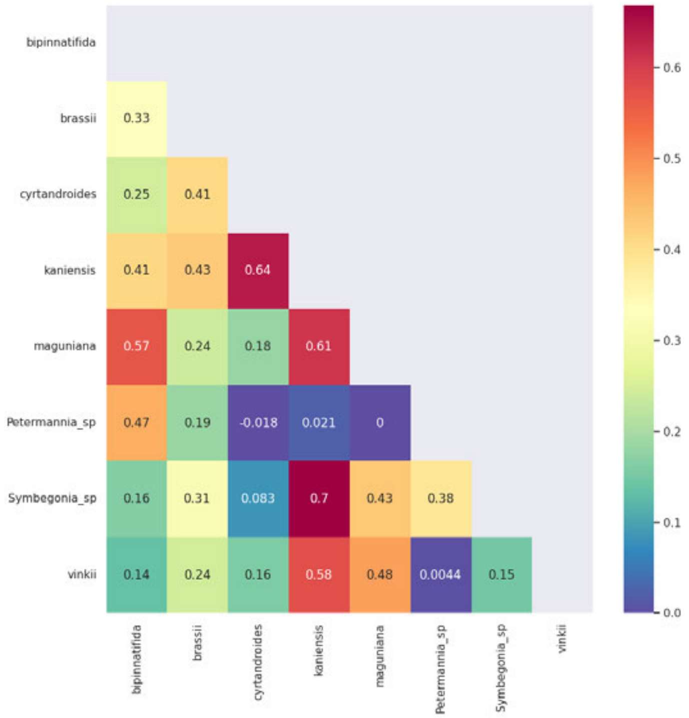


Figure C.1: F_{ST} analysis of PNG specimens by Species.

C.2 F_{IS} analysis of the PNG specimens

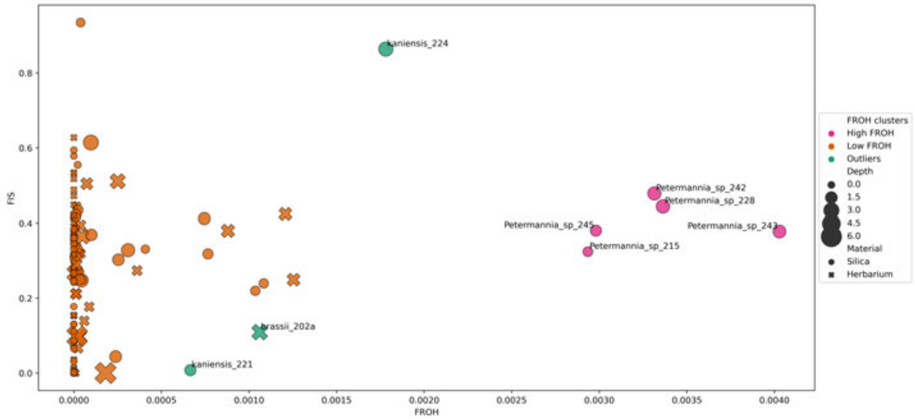


Figure C.2: Comparison between F_{IS} and F_{ROH} for PNG specimens. The color code of the dots represent the F_{ROH} clusters discussed in section 5.3.4, and represented in Fig. 5.9c

C.3 PNG Begonia phylogenetic trees annotated with F_{IS} and F_{ROH} indices

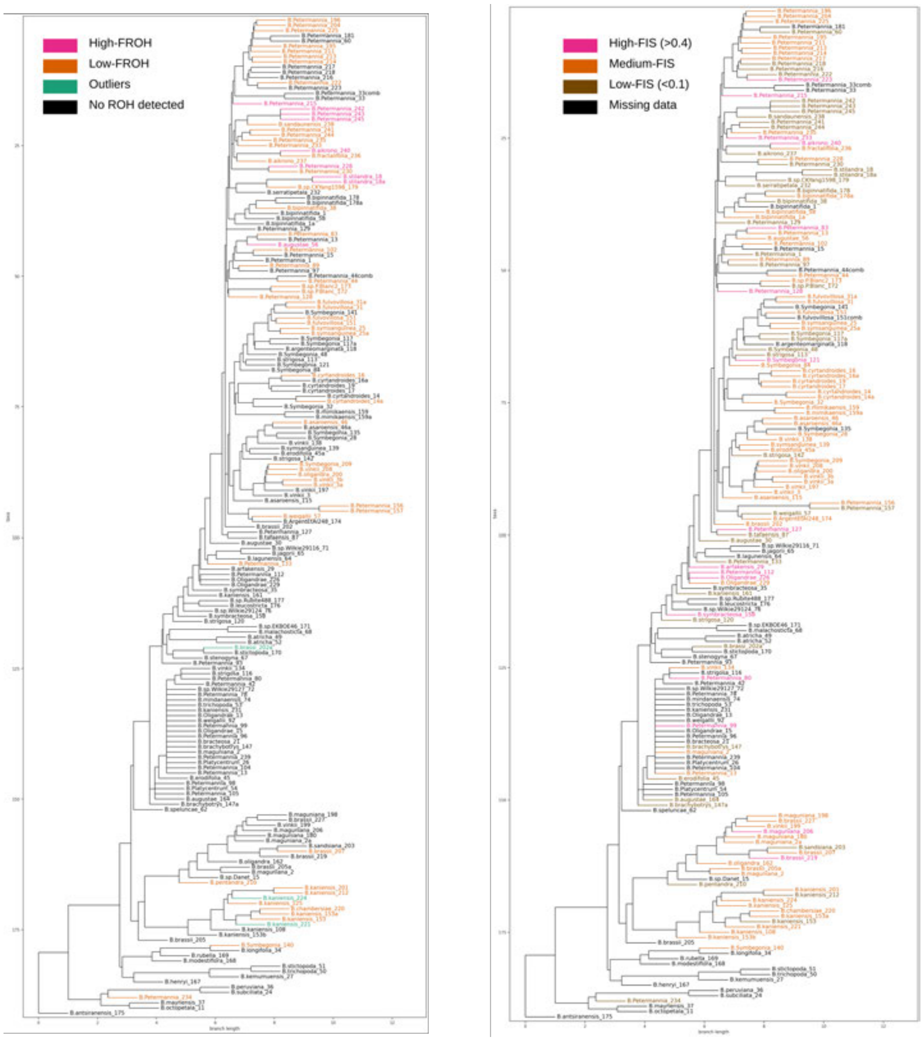
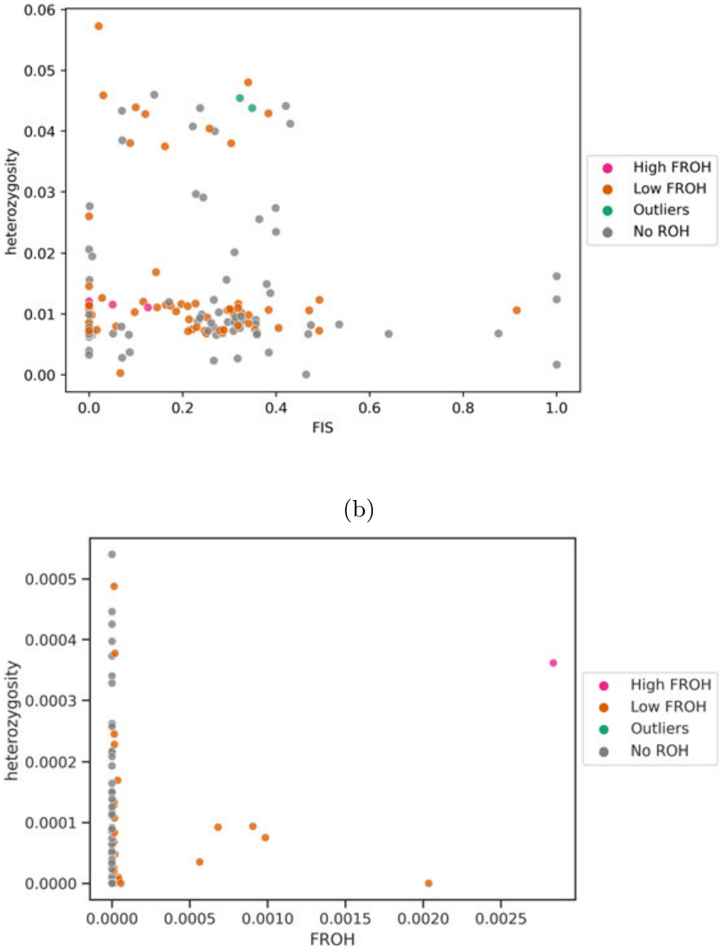


Figure C.3: Phylogenetic reconstruction of the PNG dataset. The colors represent different level of estimators. F_{ROH} (left), and F_{IS} (right) estimators are represented.

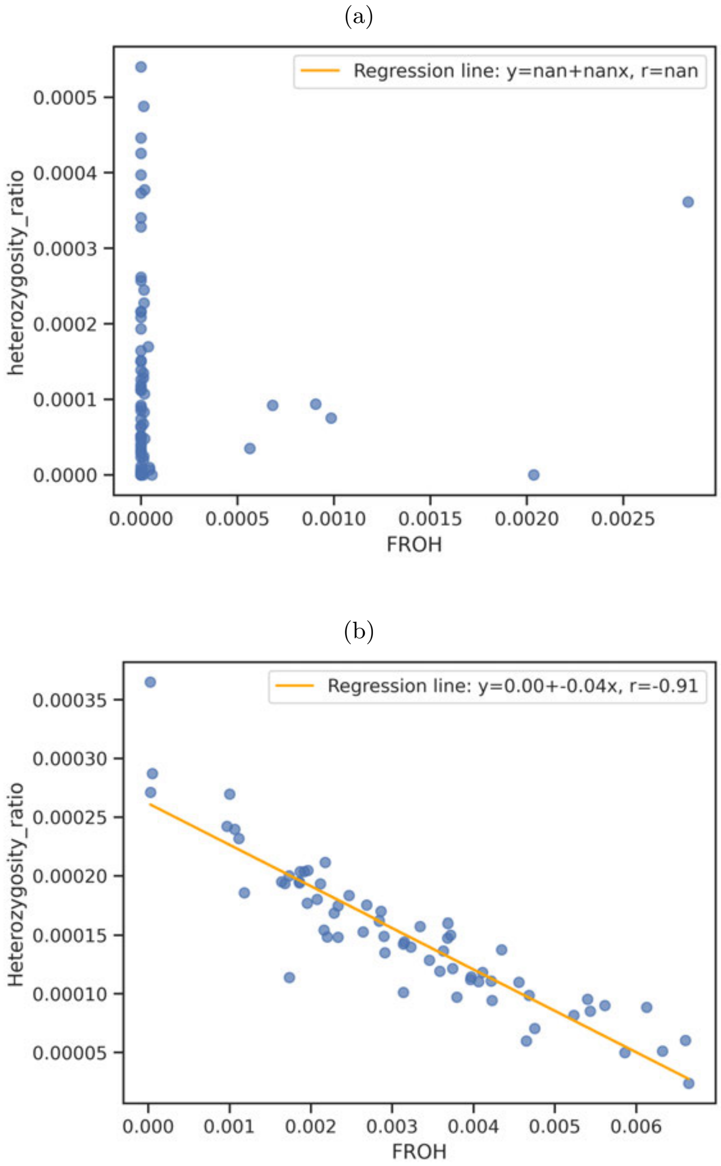
C.4 Heterozygosity, F_{IS}, and F_{ROH} indices comparison


(a)

(b)

Figure C.4: Comparison between F_{IS} calculated for each PNG specimens, F_{ROH}, and heterozygosity rate. (a) Comparison between heterozygosity rate and F_{IS}, the color code represent the F_{ROH} clusters discussed in section 5.3.4, and represented in Fig. 5.9c.

C.5 Heterozygosity and F_{IS} indices for two different clades of *Begonia*


(a)

(b)

Figure C.5: Comparison between individual F_{ROH} and heterozygosity rate for two different clade of *Begonia*. (a) PNG *Begonia*, (b) Asian *Begonia* from section *Coelocentrum* (note added in proof).