Herbarium Genomics, Skimming and Plastome Sequencing

Freek T. Bakker, Di Lei and Rens Holmer

Abstract

Herbarium genomics is a promising field, as next generation sequencing approaches are well suited to deal with the usually fragmented nature of archival DNA. We show that routine assembly of plastome sequences from herbarium specimens is feasible, from total DNA extracts and apparently only slightly depending on specimen age. We used genome skimming and an automated assembly pipeline, iterative organelle genome assembly (IOGA), that assembles paired-end reads into a series of candidate assemblies, the best one of which is selected based on assembly likelihood estimation. We used 93 specimens from 12 angiosperm families, 73 of which were from herbaria with specimen ages up to 146 years old. For 84 specimens, a sufficient amount of paired-end reads were generated (at least 50,000), yielding successful plastome assemblies for 74. Differences in plastome assemblies between herbarium and fresh specimens were modest, but the same assembly lengths were obtained. Specimens from wet-tropical conditions appear to have a higher number of contigs per assembly and lower median contig length, indicating they need more editing compared with specimens collected from dry areas. Using fungal rDNA sequences as reference in IOGA we retrieved limited anounts of reads from our samples, both silica-gel dried and herbarium, and find that fungal rDNA is not easily assembled. We conclude that routine plastome sequencing from herbarium specimens using genome skimming is feasible and cost-effective and can be performed with highly limited sample destruction.

Key Words: DNA sequence data, herbarium specimens, IOGA, museomics, organellar genomes

Freek T. Bakker, Di Lei & Rens Holmer, Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands. Corresponding author, E-mail: freek.bakker@wur.nl

SCI.DAN.B. 6 · TROPICAL PLANT COLLECTIONS
THE ROYAL DANISH ACADEMY OF SCIENCES AND LETTERS · 2017

Obtaining DNA sequence data from museum specimens has been an intriguing endeavour ever since the first attempts proved successful in the 1990s (Pääbo 1989; Savolainen *et al.* 1995; Shapiro *et al.* 2002). The notion that museum collections actually represent 'dead DNA repositories', hence enabling the testing of historical biological hypotheses, has inspired many workers to exploit collections further (e.g. Neubig *et al.* 2014) whilst minimising destructive sampling. This has led to an increase in activities in optimising sampling DNA from museum specimens, as for instance in the EU FP7-funded SYNTHESYS II programme (see http://www.synthesys.info/joint-research-activities/), where efforts have focussed on such issues as optimising DNA extraction from muco-polysaccharide-rich tissues, or minimising sampling damage (by taking small samples) from rare archaeological bone fragments. In addition, modeling of DNA decay in such material enabled quantifying the risks associated with destructive analysis of specimens prior to DNA extraction (Smith *et al.* 2003; see thermal-age.eu). All in all, the term 'museomics', i.e. the large-scale analysis of the DNA content of museum collections, has become established in several research programmes (e.g. Der Sarkissian *et al.* 2015; Gushansky *et al.* 2013; Chomicki & Renner 2015; Fabre *et al.* 2014), in addition to 'palaeogenomics' (Hofreiter *et al.* 2015).

Over the past decades museomics has been primarily focussed on improving polymerases and PCR reagents used in working with archival DNA (e.g. Hajibabaei *et al.* 2005). Efforts included developing (recombinant) polymerases with lower error rates than is typical (1 in 10,000–50,000 base pairs, which is too high for most applications), by revisiting hot springs and hydrothermal vents from where the original *Thermus acquaticus* was isolated (Chien *et al.* 1976) in persuit of thermostable polymerases with 3'-5' exonuclease proofreading capacity. Or using Restorase (Sigma-Aldrich, St Louis, MO, USA), capable of handling fragment lengths from 200–20,000 bp, in case of damaged DNA. Whereas these efforts have had mixed succes, and PCR inhibition in ancient DNA samples remains a significant problem (Kemp *et al.* 2014), this has now all been taken over by the emergence of next generation sequencing (NGS) technology. 'Suddenly' the fragmented nature of archival DNA is not a problem anymore as the massive parallel sequencing approach followed in most 'second generation' sequencing platforms uses a fragmented template anyway (Metzker 2010); in contrast, 'third generation' sequencing involves single molecule sequencing instead (Hörandl & Appelhans 2015), making it less suitable for archival DNA.

Since the application of NGS, spectacular results have been obtained in museomics, with, e.g. discovering new hominids from sequencing of small bone fragments (Reich *et al.* 2010), sequencing genomes from extinct lineages such as the Tasmanian tiger (Miller *et al.* 2009), or placing Caribbean endemic lineages of rodent (Fabre *et al.* 2014). All in all, next-generation sequencing has opened up tremendous possibilities for sequencing museum specimens due to increased output and power, but also because of ever-decreasing costs (Millar *et al.* 2008; Metzker 2010; Glenn 2011; Rowe *et al.* 2011; Buerki & Baker 2015).

The Botanical Perspective

From a botanical perspective, things are a little different given that, apart from the presence of a third genomic compartment, the plastid genome or 'plastome', the angiosperm nuclear genome is usually of much larger size than that from animals or fungi (Gregory *et al.* 2007; and see also below) and contains many repeats, which hampers genome sequence assembly. Nevertheless, herbaria do take a special place in museomics as the possession of cell walls in plant (and fungal) material provides much better protection for DNA than is the case in animal tissues (Mateiu & Rannala 2008; Roldán-Arjona & Ariza 2009), for instance for damage due to oxidative stress. On the other hand, herbarium specimens are often dried with heat, which can have adverse effects on the immediate survival of DNA. It is fairly well understood that applying heat to DNA when it is in a desiccating specimen is not favourable and can cause a range of irrepairable damage, both single- and double-stranded (Staats *et al.* 2011; Bakker 2015). Double-stranded damage causes the number of amplifiable template molecules to be reduced, as herbarium DNA is typically highly degraded into low molecular weight fragments (Doyle & Dickson 1987; Pyle & Adams 1989; Harris 1993). Single-stranded damage, however, leads to the generation of erroneous sequence information or mis-coding lesions. Thus, damaged nucleotides in herbarium DNA may result in damage-specific nucleotide mis-incorporations (miscoding lesions) by DNA polymerases during amplification (Hofreiter *et al.* 2001; Gilbert *et al,* 2003; Stiller *et al.* 2006). This includes the occurrence of a-puric sites, de-aminated cytosine residues, and oxidized guanine residues, as found in studies in vivo and on ancient DNA (Lindahl

1993; Pääbo *et al.* 2004). This type of damage is in principle polymerase-bypassible, leading to incorrect bases in the inferred sequence. Studies involving experimental preparation of herbarium specimens and the use of next generation sequencing (Staats *et al.* 2011, 2013; summarised in Bakker 2015) indicated no evidence for increased post-mortem single-stranded damage in herbarium specimens up to 100 years old. These specimens were compared with fresh DNA of the same individuals (trees growing in the Botanical Garden Leiden, The Netherlands), allowing the assertion that herbarium DNA sequence data are accurate. Whereas quantitative PCR assays indicated 90% of the DNA to be inaccessible to polymerases, probably due to double-stranded breaks directly after heat treatment, the remaining molecules are sequenced without apparent mis-coding lesions (single-stranded damage) irrespective of specimen age (Staats *et al.* 2011). Based on these data, 'DNA repair protocols' such as those suggested by Yoshida *et al.* (2015) for herbarium DNA are therefore probably not nessecary.

In a follow-up study, Staats *et al.* (2013) demonstrated that by using Illumina HiSeq technology, herbarium DNA is perfectly amenable to plastome sequencing (in spite of the 90% DNA 'lock-up'), and in case of a 43-year-old *Arabidopsis thaliana* (L.) Heynh. specimen, a full nuclear genome was sequenced as well (at 12 × average coverage). Indeed, herbarium genomics has already yielded valuable data and contributed importantly in testing historical biological hypotheses: for instance, genomes were sequenced from type specimens and rare or extinct species stored in herbaria by Zedane *et al.* (2015). Herbarium DNA was used for finding previously unknown sister groups for important crops (Sebastian *et al.* 2010; Chomicki & Renner 2015), or in SNP analysis in genotyping by sequencing of species in *Solidago* (Asteraceae) (Beck & Semple 2015). To study historical pathogens, Yoshida *et al.* (2014, 2015) determined the genotype of the *Phytophtera infestans* (Mont.) de Bary strain that caused the great Irish potato famine in the 19th century. Likewise, herbarium DNA was crucial in discovering ancient alleles in *Alopecurus myosuroides* Huds. that are relevant to herbicide resistance but pre-dating human influence (Délye *et al.* 2013). Reconstructing the shift to C4 photosynthesis in grasses could be conducted using DNA from a 100 year old Malagasy herbarium specimen for which both its phylogenetic placement and its 'genetic make-up' with regards C4 photosynthesis could be assessed (Besnard *et al.* 2014). For taxonomy and DNA barcoding herbaria collectively represent a potential treasure trove ready to be exploited (e.g. Xu *et al.* 2015). Bebber *et al.* (2010) estimated that around 70,000 new species are already in herbarium collections, 'waiting to be described'.

Therefore, it is probably fair to say that we are currently at the dawn of a herbarium genomics era (Buerki & Baker 2015), and chances are high that a large body of plant archival genomic data will be generated in the years to come. This can only underline the vital importance of securing our herbarium collections for further molecular exploitation. In addition, there is an unprecedented need for more or less automated bioinformatics pipelines for genome sequence assembly as well as for annotation and gene sequence compilation and alignment. Obviously, such tools will greatly expedite the process of massive herbarium plastome sequencing (and of other genomic compartments).

In this chapter, we discuss recent findings on generating plastome sequences from a range of fresh and herbarium angiosperm specimens, and outline challenges and issues relating to assembly accuracy, possible contamination and the use of (tropical) plant specimens.

Herbarium DNA Extraction: Garbage in, garbage out?

Challenges to extracting genomic data from herbarium specimens abound, starting with DNA extraction. Various studies (Erkens *et al.* 2008; Särkinen *et al.* 2012; Drábková *et al.* 2002; Telle & Thines 2008) focus explicitly on the efficiency of extraction and on the quality of herbarium DNA, mostly measured by PCR amplification. The expectation was that heat treatment (see above) but also the 'Schweinfurth method' (Schrenk 1888), which includes spraying specimens with ethanol in order to stop fungal growth, prior to heat treatment, will have been used in preparation of herbarium specimens. The general consensus is that when extracting DNA from herbarium leaf material, most commercially available solutions are fine as long as some combination of CTAB protocols (Doyle & Dickson 1987; Doyle & Doyle 1987) and anion exchange purification is applied. Yields are usually low, which can obviously be a problem when dealing with small, historic specimens, especially types.

In addition, and perhaps not unexpectedly, short PCR fragments were always found to amplify better using herbarium DNA (Särkinen *et al.* 2012) which is due to the fact that extracted herbarium DNA is almost always highly fragmented (Staats *et al.* 2011). As mentioned above, this double-stranded type of damage is most likely the result of herbarium specimen preparation, which is known to induce high levels of metabolic and cellular stress responses and ultimately cell death (Savolainen *et al.* 1995). The high temperatures (60-70 °C) at which herbarium specimens are typically dried cause cells to rupture quickly, releasing nucleases and other cellular enzymes (Gill & Tuteja 2010), as well as reactive oxygen species. Such physiological conditions resemble necrosis, and this cellular stress typically causes DNA to degrade randomly into smaller fragments, running as a smear on agarose gels (Reape *et al.* 2008; McCabe *et al.* 1997).

After the PCR Era

Precisely this aspect, fragmentation of herbarium DNA, transfigured from 'nuisance' to 'blessing in disguise' in the NGS world, as targeted (Sanger) sequencing of amplified fragments has been replaced by massive parallel sequencing (Metzker 2010), which requires fragmentation of the template genomic DNA. Therefore, the problems associated with traditional herbarium DNA extraction in the PCR era, i.e. low yields and DNA fragmentation, came into a new light with fragments now being incorporated directly into NGS libraries, and the generally low yields sometimes being overcome by whole genome amplification (WGA). Whereas WGA can help obtaining enough DNA strands for proper library building, it can in principle, however, cause artefacts in the representation of the target genomes and hence in genome sequence assembly. The alternative is to use more starting herbarium material, but generally speaking, for plastome sequencing one square centimeter of herbarium leaf tissue suffices for successful extraction, library preparation and (Illumina) sequencing, which will be feasible for most specimens.

On the other hand, herbarium DNA fragmentation can sometimes have happened to such an extent that the efficiency of paired-end sequencing using Illumina HiSeq is affected. In such cases, the effective insert size in the sequencing libraries becomes so small that the actual sequencing reads 'meet in the middle' of the insert and start to overlap, therefore reducing the power of the paired-end information used in the assembly. Furthermore, it is clear that in such cases the use of third generation technologies such as provided by Pacific Biosciences (www.pacb.com) using whole molecule sequencing is prevented.

Genome Skimming

The angiosperm genome size ranges from a minute 65 Mb (parasitic *Genlisea*, Lentibulariaceae) up to a staggering 150,000 Mb (octaploid *Paris japonica*, Melianthaceae) and is on average considered to be 6000 Mb long (Litt 2013). Well over half the angiosperm genomes estimated to date were found to be smaller than 5000 Mb and about one-third to be under 1000 Mb (Murray *et al.* 2010). Therefore angiosperm genome sequence assembly represents a huge challenge (e.g. The Tomato Genome Consortium 2012) and is by far not as routine an undertaking as it is in animal and fungal genomics. Some parts of the angiosperm genome, however, are present in high copy number, notably the rDNA cistron repeats, the organellar genomes, i.e. the plastome and the chondrome (mitochondrial genome), and the different classes of highly repeated elements among which we distinguish microsatellite regions and long terminal repeats or transposable elements. Because of their repetitive nature, such regions will collectively be relatively well represented, even in a limited or 'skimmed' second generation sequencing sample that, by itself, would be too small to cover the entire nuclear genome. 'Genome skimming' has therefore been coined for the approach where superficial sequencing is performed and only genomic repeats or organellar genomes are represented with sufficient sequencing depth (Straub *et al.* 2012; Dodsworth *et al.* 2015). Usually this results in relatively low costs compared with full genome sequencing (although the cost for sequencing library preparation remains the same), and therefore it is an approach well suited for comparative studies involving many specimens. Another advantage of a skimming approach is that it prevents introducing rare variants and errors from various sources (Lonardi *et al.* 2015), whilst at the same time maintaining sufficient coverage for each repetitive genomic compartment. In a sense, it makes genome skimming comparable again with Sanger sequencing, in which 'rare variants' are marginalised in light of a main, average signal peak in Sanger trace files.

IOGA

In a paper in a special issue on 'Collection-based research in the genome era' in the *Biological Journal of the Linnean Society* we described an automated bioinformatics assembly pipeline for angiosperm organellar genomes, including iterative organelle genome assembly (IOGA) based on genome skimming data (Bakker *et al.* 2016). Our approach is similar to the 'baiting and iterative mapping' MitoBIM pipeline described by Hahn *et al.* (2013) for mitochondrial genomes, the difference being that IOGA does not require closely related reference organelle genome sequences, and in addition that best assemblies are selected from multiple candidate assemblies. The IOGA Python script can be obtained from Github (https://github.com/holmrenser/IOGA), and is usually run after first taking a random subsample of reads R from the overall read pool in order to avoid excessive plastome coverage (and hence excessive processing time); the subsample typically includes 1M forward and 1M reverse reads. R is then subjected to IOGA which includes the following steps : (1) low quality, adapter and other Illumina-specific sequences are trimmed from individual reads; (2) plastid genome-derived reads ('R_{Pl}') are filtered out of R by aligning the latter to a panel of reference angiosperm (and land plant) plastid genome sequences, using Bowtie (Langmead *et al.* 2009). R_{Pl} is then subjected to the following steps: (3) using SOAPdenovo2 (https://github.com/aquaskyline/SOAPdenovo2) assemblies are made from the filtered, trimmed and corrected plastid reads contained in R_{Pl}, using k-mer values ranging from 37–97; and (4) 'best assemblies' are selected using the N50 criterion and then used as a 'new reference' in order to find target-specific reads from R that were not selected in the first iteration. (N50 is defined as the median length-weighted contig length or the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs.) Step (4) is then repeated until no further R_{Pl} reads are found, followed by (5), assembly of the final set of reads with SPAdes3.0 (Bankevich *et al.* 2012). This assembler applies a bi-directional De Bruijn graph, solving 'complex knots', under a range of different k-mer settings. Finally, (6) in order to select among candidate assemblies from SPAdes (step (3)) we apply a 'read'-driven test named 'assembly likelihood estimation' (Clark *et al.* 2013), which calculates the likelihood of the fit of the original reads to each candidate assembly, using a model that includes parameters such as 'read quality', 'mate pair orientation', 'read alignment' and 'sequence coverage'. The ALE test therefore assures assembly quality at the read level (Clark *et al.* 2013) and the one with the best -LnL score is selected as final assembly, (5), which is then subjected to further genome annotation (for instance using DOGMA; Wyman *et al.* 2004). After scaffolding, i.e. correcting the relative orientation and order of contigs using 'map to reference' in Geneious (www.geneious.com), final assemblies are then compared with available 'nearest' reference plastome sequences in order to check accuracy of our assemblies. This is done in pair-wise alignments using MUMmer plots (Kurtz *et al.* 2004), as implemented in MAFFT using default settings (Katoh & Standley 2013); basically, one would expect co-linearity of assembly and reference plastome in case of conspecifics. For further technical information on IOGA, scripts, updates and programmes used, see the Github mentioned above and Bakker *et al.* (2016).

A Herbarium Genomics Test-case

Using the IOGA pipeline described above, we compared 93 specimens from 12 angiosperm families, 73 of which were herbarium specimens up to 146 years old, to explore the feasibility of herbarium genomics (Bakker *et al.* 2016). After DNA extraction and quantification, carried out under standard conditions (i.e. not in an ancient DNA lab), sequence library preparation, index PCR and equimolar pooling of indexed libraries were conducted and all libraries were then sequenced on four lanes on an Illumina HiSeq 2000 platform using paired-end chemistry. For 84 out of our 93 specimens, sufficient numbers of paired-end reads were generated (at least 50,000), with all but two of the failed specimens being from historical herbarium material. A significant negative correlation was found between total reads per sample and specimen age, indicating that despite PCR enhancement of poor samples in the library preparation, older specimens still give fewer reads. The 84 successful samples were then subjected to IOGA, which yielded (after filtering out all contigs < 1000 bp) successful plastome assemblies for 74 specimens (80% of the specimens), at a rate of approximately one hour per specimen using IOGA on a 64GB RAM Linux workstation with 16 cores. The fact that 19 of our 93 specimens did not yield plastome assemblies we feel may have been due to the fact that not enough copies of these plastomes were present in the first place or the required equimolar mixing of specimens in the Illumina flow cell may have been unsuccessful, causing libraries for those specimens not to be sequenced successfully.

Assembly lengths varied from 6–220 kb with an overall average total assembly length of 136,167 bp, which is consistent with reported average angiosperm plastome length, 120–170 kb (e.g. Downie & Palmer 1992), including two inverted repeat (IR) regions of, on average, 25 kb each. In one case, *Pelargonium elegans* Willd., a 117-year-old herbarium specimen, using only 24 ng of herbarium DNA, yielded a 167,770 bp assembly; from another, *Aethionema membranaceum* DC., a 146-year-old herbarium specimen, a complete plastome sequence was obtained. After checking pair-wise alignments (MUMmer plots) of best assemblies in selected samples, we found good co-linearity with the published reference plastome sequence in cases for which reference and target were the same species, indicating accurate plastome sequence assembly. Reduced co-linearity was found in case of congenerics, which reflects phylogenetic distance between target and reference rather than mis-assembly.

When comparing fresh and herbarium specimens in terms of plastome assembly, it was found that differences were modest, with herbarium specimens yielding lower fractions of plastome-derived reads (4%) compared with those from fresh and silica-gel dried specimens (13%; Fig. 1). This would suggest that plastids may be lost preferentially, after herbarium specimen fixation with high temperatures. This seems to contradict the studies by Staats *et al.* (2011), who did not find evidence for preferential degradation of organellar DNA in herbarium tissue based on quatitative PCR assays. In any case, herbarium specimens appear to yield enough reads for effective plastome assembly; we found that total assembly length did not differ significantly between fresh and herbarium specimens, but that fresh samples on average yielded longer individual assemblies. This indicates that the specimen preparation process, which often included heat treatment, causes plastome assemblies to be more fragmented compared with fresh samples, possibly in additional fragments <1000 bp. Nevertheless, total assembly length from herbarium DNA is the same, and herbarium assemblies just need slightly more more editing and 'scaffolding'.

Unexpectedly, specimen age per se does not seem to correlate with plastome assembly succes. Of the 74 succesful specimens in Bakker *et al.* (2016), there were eight specimens older than 80 years, half of which gave plastome assemblies (>125kb) that may be complete (or excluding one IR region). For all other spec-

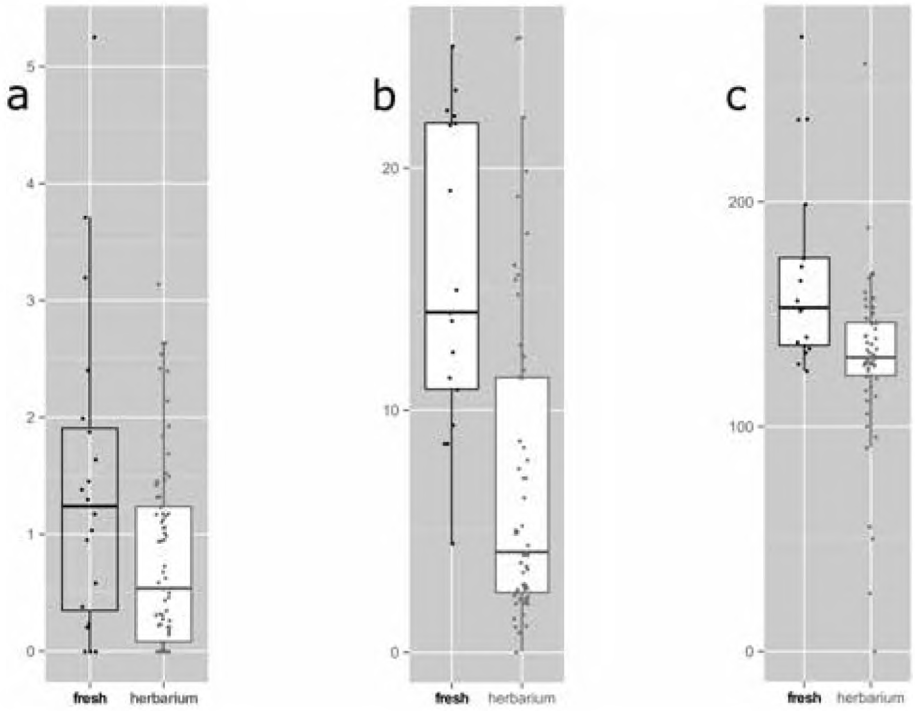

Fig. 1. Median, first and third quartile and 95% confidence interval of median of total number of reads (× 10^7) for fresh or silica-dried samples vs. those from herbarium samples (a); the same for plastid-derived reads R_{Pl} (b); total assembly length (in kb) for 74 successful assemblies derived from fresh or silica-dried vs. herbarium specimens (c). Re-drawn from Bakker *et al.* (2016).

imens (i.e. younger than 80 years), this proportion was just over half (55%). Although there were more young than old specimens, which prevents making direct comparisons, it still appears that assembly success does not depend on specimen age. This is of course promising for the near-future further exploitation of herbarium collections world-wide, as many older (type) specimens are available.

A special note needs to be made about herbarium specimens from wet-tropical conditions, of which there were 13 included in our study. Given the potentially different conditions under which these specimens have been collected and preserved, it is worthwhile determining if this correlates with herbarium genomics success, i.e. plastome assembly efficiency. Whereas 'dry collected' specimens sometimes may not even have been subjected to heat treatment (other than the sun) and ususally do not get 'Schweinfürted' (Schrenk 1888) i.e. sprayed with ethanol in order to stop any fungi growing, for wet-tropical specimens this may be the opposite. It appears that preserving such specimens by immersion in ethanol prevents any DNA from being recovered later on (Mark Chase *pers. com.*). Bressan *et al.* (2014) however, found no difference in neither quality nor quantity of nuclear DNA recovered from tropical plant leaf tissue stored in liquid nitrogen versus 96% ethanol, but also show how storage in ethanol causes cytoplasmic contents (including plastids) to be cleared from the leaf tissue cells. Therefore, in our opinion ethanol preservation is best to be avoided for herbarium genomics when targeting plastomes or chondromes. The Schweinfürth treatment in wet-tropical conditions nowadays usually entails keeping specimens inside a plastic bag under a saturated ethanol atmosphere, which can last for days before a drier is reached. Alternatively, specimens are somtimes dried directly on a kerosine or gas-stove (Jan Wieringa *pers. com.*).

When we compare our wet-tropical samples with the rest, we see generally a higher number of contigs per assembly and lower N50 values (Fig. 2). When plotted against specimen age it appears as if the wet-tropical specimens seem to 'age' more quickly in terms of increased plastome assembly fragmentation

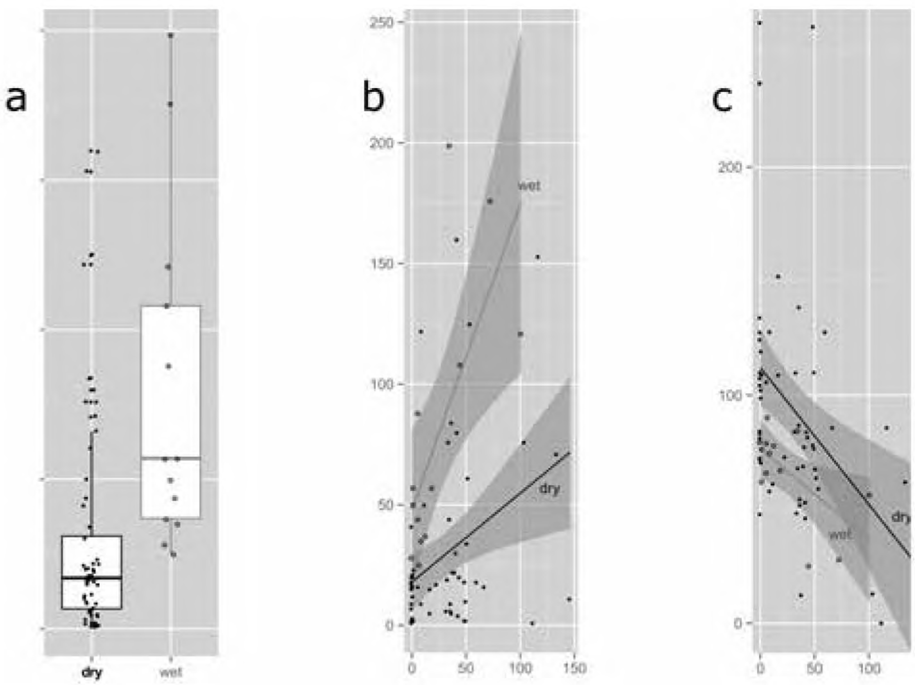

Fig. 2. A comparison between specimens collected from wet-tropical and dry conditions with regard to success in plastome assembly, in terms of number of contigs (a), number of contigs plotted against specimen age in years (b), and (c) N50 (in kb) plotted against specimen age in years. Darker shaded areas indicate confidence intervals of the linear model fitted. Re-drawn from Bakker *et al.* (2016).

when compared with dry habitat specimens. As the exact preservation histories cannot be reconstructed for most herbarium specimens, we cannot draw firm conclusions here but suggest that wet-tropical herbarium specimens may need some extra effort in terms of plastome assembly and possibly require more additional Sanger sequencing-based confirmation of assembly boundaries.

Worry about Contamination?

Another concern with using harbarium DNA could be the presence of contaminant DNA in samples, for instance from either endophytic or 'post-mortem' fungi. In case of post-mortem contamination of the specimen, we would expect the contaminant DNA to be much less fragmented than that of the specimen, as only the latter would have been heat-treated. Fungal contamination in plant (herbarium) samples has been reported to be fairly widespread (Álvarez & Wendel 2003; Miranda *et al.* 2010), and the extent to which plant rDNA ITS sequences in public databases such as GenBank are actually fungal can be questioned. Because the genome skimming/IOGA approach is in theory suitable for other high copy number compartments such as chondromes and rDNA, it is relevant to know to what extend non-target rDNA could be picked-up using this approach. Therefore, to assess the proportion of fungal-derived reads in a selection of our samples we re-ran IOGA using a panel of fungal SSU rDNA and ITS1-5.8SrDNA-ITS2 sequences, comprising both asco- and basidiomycetes. In case fungal ITS sequences were assembled, they were identified using the UNITE database (Köljalg *et al.* 2013) that currently holds 354,465 annotated fungal rDNA ITS sequences (http://unite.ut.ee/). BLAST was used to match our target ITS sequences against a subset library of 20,000 fungal ITS sequences from UNITE.

The results (Lei 2015) were unexpected in that only modest numbers of reads (ranging up to appr. 73,000) were found in the selected herbarium specimen read samples by using these fungal references, and when assembled into rDNA sequences, the majority of contigs turned out to be plant rDNA not fungal rDNA. In only a minority of cases were 'non-plant' contigs found, usually of <2 kb, which could only in some cases be identified as fungal. In addition, when repeating the analyses, but this time using plant rDNA sequences (both SSU and rDNA ITS), in some cases a minority of fungal contigs were assembled that could be identified using the UNITE data base as *Cladosporium*, *Aureobasidium*, *Fibulobasidium* species and in one case a 'human skin community' type fungus. Whereas the first three matches would make sense given the ecology of these fungi (leaf parasites or endophytic fungi), the latter would be consistent with a scenario of human fungal contamination. The cross-assembly results can probably be explained by the high conservation at the nucleotide sequence level of parts of the fungal and plant rDNA cistron. However, in practical terms, given the difficulty we encountered in obtaining fungal reads and assembling fungal rDNA from these herbarium samples, and in the same time given the ease with which plant rDNA reads and assemblies could be obtained, we consider fungal cross-contamination artefacts in herbarium DNA to be of minimal importance.

Conclusions

We conclude that effective plastome sequence assembly using genome skimming is feasible using small amounts of herbarium specimen tissue, roughly one square centimetre of leaf, and show that the results are only in some aspects different from those obtained from fresh or silica-gel-dried material. We are confident that most of our specimens have been sampled non-destructively and therefore are optimistic that this approach can be used more widely for future genomic exploitation of herbarium collections.

The IOGA automated pipeline established previously in Bakker *et al.* (2016) appears to be working effectively, with draft plastome assemblies being completed in one or a few hours only. Obviously, subsequent gene annotation and quality check of contigs, which may include Sanger verification of contig boundaries, is still a formidable task but is (time-wise) probably less so than the curation of a large scale comparative sequence project using traditional Sanger sequencing. Using a panel of land-plant-wide plastome sequences as reference proves to be efficient, and no closely related reference plastome is needed. For instance, no Brassicaceae reference plastome was included (*Medicago* was probably the closest reference included phylogenetically), but all Brassicaceae samples in our study were assembled correctly. The fact that our IOGA plastome assemblies could be aligned without any problem to their reference plastome sequences indicates that assembly was accurate. Nevertheless, additional analysis by re-mapping reads to finally selected assemblies and checking whether anomalies exist is still important, but this is general 'good genomic practice'. For specimens collected and preserved in wet-tropical conditions we conclude that more effort into contig assembly, scaffolding and editing of plastome sequences is probably required but is expected to yield fully comparable final results compared with dry-collected specimens. Finally, we found possible contamination of herbarium specimens with fungal DNA not to be an (important) issue. Therefore, herbarium genomics is promising and further makes continued support and curation of herbarium collections around the world important.

Acknowledgements

We thank the organising committee of the *Tropical Plant Collections Symposium* for inviting this contribution. We thank Mark Chase and an anonymous reviewer for valuable comments on the manuscript.

References

Alvarez, I. & Wendel, J.F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417-434.

Bakker, F.T., Lei, D., Yu, J., Mohammadin, S., Wei, Z., Van de Kerke, S., Gravendeel, B., Nieuwenhuis, M., Staats, M., Alquezar-Planas, D.E. & Holmer, R. (2016). Herbarium genomics: Plastome sequence assembly from a range of herbarium specimens using an terative organelle genome assembly (IOGA) pipeline. *Biological Journal of the Linnean Society* 117: 33-43. http://dx.doi.org/10.1111/bij.12642

Bakker, F.T. (2015). DNA sequences from plant herbarium tissue. *In*: E. Hörandl. & M.S. Appelhans (eds.), *Next-generation Sequencing in Plant Systematics. Regnum Vegetabile* 158: 271–288. http://dx.doi.org/10.14630/000009.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A. & Pevzner, P.A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 19(5): 455–477. doi:10.1089/cmb.2012.0021.

Bebber, D.P., Carine, M.A., Wood, J.R.I., Wortley, A.H., Harris, D.J., Prance, G.T., Davidse, G., Paige, J., Pennington, T.D., Robson, N.K.B. & Scotland, R.W. (2010). Herbaria are a major frontier for species discovery. *Proceedings of the National Academy of Sciences* 107: 22169–22171.

Beck, J.B. & Semple, J.C. (2015). Next-generation sampling: Pairing genomics with herbarium specimens provides species-level signal in *Solidago* (Asteraceae). *Bio One. Applications in Plant Sciences* 3(6): 1500014. doi: http://dx.doi.org/10.3732/apps.1500014

Besnard, G., Christin, P.-A., Malé, P.-J.G., L'huillier, E., Lauzeral, C., Coissac, E. & Vorontsova, M.S. (2014). From museums to genomics: Old herbarium specimens shed light on a C3 to C4 transition. *Journal of Experimental Botany* 65(22): 6711–6721 doi:10.1093/jxb/eru395

Bressan, E.A., Rossi, M.L., Gerald, L.T. & Figueira, A. (2014). Extraction of high-quality DNA from ethanol-preserved tropical plant tissues. *BMC Research Notes* 24(7): 268. doi: 10.1186/1756-0500-7-268.

Buerki, S. & Baker, W.J. (2015). Collections-based research in the genomic era. *Biological Journal of the Linnean Society* 117: 5–10. doi: 10.1111/bij.12721

Chien, A., Edgar, D.B., & Trela, J.M. (1976). Deoxyribonucleic acid polymerase from the extreme thermophile *Thermus aquaticus. Journal of Bacteriology* 127(3): 1550–1557.

Chomicki, G. & Renner, S.S. (2015). Watermelon origin solved with molecular phylogenetics including Linnaean material: Another example of museomics. *New Phytologist* 205: 526–532.

Clark, S.C., Egan, R., Frazier, P.I. & Wang, Z. (2013). ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* 29: 435–443.

Délye, C., Deulvot, C. & Chauvel, B. (2013). DNA analysis of herbarium specimens of the grass weed *Alopecurus myosuroides* reveals herbicide resistance pre-dated herbicides. *PLoS ONE* 8(10): e75117.

Der Sarkissian, C., Allentoft, M.E., Ávila-Arcos, M.C., Barnett, R., Campos, P.F., Cappellini, E., Ermini, L., Fernández, R., Fonseca, R. de, Ginolhac, A., Hansen, A.J., Jónsson, H., Korneliussen, T., Margaryan, A., Martin, M.D., Moreno-Mayar, J.V., Raghavan, M., Rasmussen, M., Sandoval Velasco, M., Schroeder, H., Schubert, M., Seguin-Orlando, A., Wales, N., Gilbert, M.T.P., Willerslev, E. & Orlando, L. (2015). Ancient genomics. *Philosophical Transactions of the Royal Society B* 370: 20130387. http://dx.doi.org/10.1098/rstb.2013.0387

Dodsworth, S., Chase, M.W., Kelly, L.J., Leitch, I.J., Macas, J., Novák, P., Piednoel, Weiss-Schneeweiss, H. & Leitch, A.R. (2015). Genomic repeat abundances contain phylogenetic signal. *Systematic Biology* 64(1): 112–126. doi: 10.1093/sysbio/syu080

Downie, S.R. & Palmer, J.D. (1992). Use of chloroplast DNA rearrangements in reconstruction plant phylogeny. *In*: P.S. Soltis, D.E. Soltis & J.J. Doyle (eds.), *Molecular Systematics of Plants*. Chapman and Hall, New York. P.p. 14–35.

Doyle, J.J. & Dickson, E.E. (1987). Preservation of plant samples for DNA restriction endonuclease analysis. *Taxon* 36: 715 722.

Doyle, J.J. & Doyle, J.L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19: 11 15.

Drábková, L., Kirschner, J. & Vlcek, C. (2002). Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of Juncaceae. *Plant Molecular Biology Reporter* 20: 161 175.

Erkens, R.H.J., Cross, H., Maas, J.W., Hoenselaar, K. & Chatrou, L.W. (2008). Age and greenness of herbarium specimens as predictors for successful extraction and amplification of DNA. *Blumea* 53: 407–428.

Fabre, P.H., Vilstrup, J.T., Raghavan, M., Sarkissian, C.D., Willerslev, E., Douzery, E.J.P. & Orlando, L. (2014). Rodents of the Caribbean: Origin and diversification of hutias unravelled by next-generation museomics. *Biology Letters* 10: Article number 20140266.

Gilbert, M.T.P., Willerslev, E., Hansen, A.J., Barnes, I., Rudbeck, L., Lynnerup, N. & Cooper, A. (2003). Distribution patterns of postmortem damage in human mitochondrial DNA. *American Journal of Human Genetics* 72: 32–47.

Gill, S.S. & Tuteja, N. (2010). Reactive oxygen species and antioxidant machinery in abiotic stress tolerance in crop plants. *Plant Physiology and Biochemistry* 48: 909–930.

Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. *Molecular Ecology Resources* 11: 759–769.

Gregory, T.R., Nicoll, J.A., Tamm, H., Kullman, B., Kullman, K., Leitch, I.J., Murray, B.G., Kapraun, D.F., Johann Greilhuber, J. & Bennett, M.D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research* 35(Database issue): D332–D338. doi:10.1093/nar/gkl828.

Guschanski, K., Krause, J., Sawyer, S., Valente, L.M., Bailey, S., Finstermeier, K., Sabin, R., Gilissen, E., Sonet, G., Nagy, Z.T., Lenglet, G., Mayer, F. & Savolainen, V. (2013). Next-generation museomics disentangles one of the largest primate radiations. *Systematic Biology* 62(4): 539–554. doi: 10.1093/sysbio/syt018

Hahn, C., Bachmann, L. & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads – a baiting and iterative mapping approach. *Nucleic Acids Research* 41(13): e129.

Hajibabaei, M., deWaard, J.R., Ivanova, N.V., Ratnasingham, S., Dooh, R.T., Kirk, S.L., Mackie, P.M. & Hebert, P.D.N. (2005). Critical factors for assembling a high volume of DNA barcodes. *Philosophical Transactions of the Royal Society B* 360: 1959–1967. doi:10.1098/rstb.2005.1727

Harris, S.A. (1993). DNA analysis of tropical plant species: An assessment of different drying methods. *Plant Systematics and Evolution* 188: 57–64.

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

Hofreiter, M., Paijmans, J.L.A., Goodchild, H., Speller, C.F., Barlow, A., Fortes, G.G., Thomas, J.A., Ludwig, A. & Collins, M.J. (2015). The future of ancient DNA: Technical advances and conceptual shifts. *BioEssays* 37: 284–293. doi: 10.1002/bies.201400160

Hörandl, E. & Appelhans, M.S. (2015). Introduction to chapters and methodological overview. *In:* E. Hörandl & M. Appelhans (eds.), *Next Generation Sequencing in Plant Systematics*. *Regnum Vegetabile* 158: 1–8. http://dx.doi.org/10.14630/000010.

Katoh, K. & Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30: 772–780.

Kemp, B.M., Monroe, C., Judd K.G., Reams, E. & Grier, C. (2014). Evaluation of methods that subdue the effects of polymerase chain reaction inhibitors in the study of ancient and degraded DNA. *Journal of Archaeological Science* 42: 373e380. https://doi.org/10.1016/j.jas.2013.11.023.

Köljalg, U., Nilsson, R.H., Abarenkov, K., Tedersoo, L., Taylor, A.F.S., Bahram, M., Bates, S.T., Bruns, T.D., Bengtsson-Palme, J., Callaghan, T.M., Douglas, B., Drenkhan, T., Eberhardt, U., Dueñas, M., Grebenc, T., Griffith, G.W., Hartmann, M., Kirk, P.M., Kohout, P., Larsson, E., Lindahl, B.D., Lücking, R., Martín, M.P., Matheny, P.B., Nguyen, N.H., Niskanen, T., Oja, J., Peay, K.G., Peintner, U., Peterson, M., Põldmaa, K., Saag, L., Saar, I., Schüßler, A., Scott, J.A., Senés, C., Smith, M.E., Suija, A., Taylor, D.L., Telleria, M.T., Weiß, M. & Larsson, K.-H. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* 22 (21): 5271–5277. doi: 10.1111/mec.12481.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. *Genome Biology* 5: Article R12.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology 2009* 10: R25. doi:10.1186/gb-2009-10-3-r25

Lei, D. (2015). Assembling ribosomal DNA and mitochondrial genomic sequences from Illumina read libraries for species-level phylogenetic reconstruction. *Research Minor Report*. [Student report.] Hogeschool Arnhem Nijmegen, Arnhem and Nijmegen.

Lonardi, S., Mirebrahim, H., Wanamaker, S., Alpert, M., Ciardo, G., Duma, D., Close, T.J. (2015). When less is more: 'slicing' sequencing data improves read decoding accuracy and de novo assembly quality. *Bioinformatics* 2015: 1–9.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature* 362: 709–715.

Litt, A. (2013). Comparative evolutionary genomics of land plants. *Annual Plant Reviews* 45: 227–276. doi: 10.1002/9781118305881.ch8

Mateiu, L. M. & Rannala, B.H. (2008). Bayesian inference of errors in ancient DNA caused by postmortem degradation. *Molecular Biology and Evolution* 25(7): 1503–1511. doi:10.1093/molbev/msn095

McCabe, P.F., Levine, A., Meijer, P.J., Tapon, N.A. & Pennell, R.I. (1997). A programmed cell death pathway activated in carrot cells cultured at low cell density. *The Plant Journal* 12: 267–280.

Metzker, M.L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* 11: 31–46.

Millar, C.D., Huynen, L., Subramanian, S., Mohandesan, E. & Lambert, D.M. (2008). New developments in ancient genomics. *Trends in Ecology and Evolution* 7: 386–393.

Miller, W., Drautz, D.L., Janecka, J.E., Lesk, A.M., Ratan, A., Tomsho, L.P., Packard, M., Zhang, Y., McClellan, L.R., Qi, J., Zhao, F., Gilbert, M.T.P., Dalén, L., Arsuga, J.L., Ericson, P.G.P., Huson, D.H., Helgen, K.M., Murphy, W.J., Götherström, A. & Schuster, S.C. (2009). The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). *Genome Research* 19: 213–220.

Miranda de, V.F.O., Martins, V.G., Furlan, A. & Bacci jr., M. (2010). Plant or fungal sequences? An alternative optimized PCR protocol to avoid ITS (nrDNA) misamplification. *Brazilian Archives of Biology and Technology* 53: 141–152.

Murray, B.G., Leitch, I.J. & Bennett, M.D. (2010). Gymnosperm DNA C-values database. Release 4.0, Oct. 2010. Available from: http://data.kew.org/cvalues/

Neubig, K.M., Whitten, W.M., Abbott, J.R., Elliott, S., Soltis, D.E. & Soltis, P.S. (2014). Variables affecting DNA preservation in archival DNA specimens. *In*: W.L. Applequist & L.M. Campbell (eds.), *DNA Banking for the 21st Century. Proceedings of the US Workshop on DNA Banking*. Missouri Botanical Garden, St. Louis. Pp. 81–136.

Pääbo, S. (1989). Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences* 86: 1939. doi:10.1073/pnas.86.6.1939 pmid:2928314

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

Pyle, M.M. & Adams, R.P. (1989). In situ preservation of DNA in plant specimens. *Taxon* 38: 576–581.

Reape, T.J., Molony, E.M. & McCabe, P.F. (2008). Programmed cell death in plants: Distinguishing between different modes. *Journal of Experimental Botany* 59: 435–444.

Reich, D., Green, R.E., Kircher, M., Krause, J., Patterson, N., Durand, E.Y., Viola, B., Briggs, A.W., Stenzel, U., Johnson, P.L.F., Maricic, T., Good, J.M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E.E., Stoneking, M., Richards, M., Talamo, S., Shunkov, M.V., Derevianko, A.P. & Hublin, J.-J. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* 468: 1053–1060. doi:10.1038/nature09710

Roldán-Arjona, T. & Ariza, R.R. (2009). Repair and tolerance of oxidative DNA damage in plants. *Mutation Research* 681: 169–179.

Rowe, K.C., Singhal, S., MacManes, M.D., Ayroles, J.F., Morelli, T.L., Rubidge, E.M., Bi, K. & Moritz, C.C. (2011). Museum genomics: Low-cost and high-accuracy genetic data from historical specimens. *Molecular Ecology Resources* 11: 1082–1092.

Särkinen, T., Staats, M., Richardson, J.E., Cowan, R.S. & Bakker, F.T. (2012). How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS ONE* 10: 1371/journal.pone.0043808. https://doi.org/10.1371/journal.pone.0043808

Savolainen, V., Cuénoud, P., Spichiger, R., Martinez, M.D.P., Crèvecoeur, M. & Manen, J.-F. (1995). The use of hebarium specimens in DNA phylogenetics: Evaluation and improvement. *Plant Systematics and Evololution* 197: 87–98.

Sebastian, P., Schaefer, H., Telford, I.R.H. & Renner, S.S. (2010). Cucumber (*Cucumis sativus*) and melon (*C. melo*) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proceedings of the National Academy of Sciences* 107: 14269–14273.

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

Schrenk, J. (1888). Schweinfurth's method of preserving plants for herbaria. *Bulletin of the Torrey Botanical Club* 15: 292–293.

Smith, C.I., Chamberlain, A.T., Riley, M.S., Stringer, C. & Collins, M. (2003). The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution* 45: 203–217.

Staats, M., Cuence, A., Richardson, J.E., Vrielink-van Ginkel, R., Petersen, G., Seberg, O. & Bakker, F.T. (2011). DNA damage in plant herbarium tissue. *PLoS ONE* 6: *e28448*.

Staats, M., Erkens, R.H.J., van de Vossenberg, B., Wieringa, J.J., Kraaijeveld, K., Stielow, B., Geml, J., Richardson, J.E. & Bakker, F.T. (2013). Genomic treasure troves: Complete genome sequencing of herbarium and insect museum specimens. *PLoS ONE* 8(7): e69189. doi:10.1371/journal.pone.0069189

Stiller, M., Green, R.E., Ronan, M., Simons, J.F., Du, L., He, W., Egholm, M., Rothberg, J.M., Keates, S.G., Ovodov, N.D., Antipina, E.E., Baryshnikov, G.F., Kuzmin, Y.V., Vasilevski, A.A.,Wuenschell, G.E., Termini, J., Hofreiter, M., Jaenicke-Després, V. & Pääbo, S. (2006). Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. *Proceedings of the National Academy of Sciences* 103: 13578–13584.

Straub, S.C.K., Parks, M., Weitemeir, K., Fishbein, M., Cronn, R. & Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. *American Journal of Botany* 99: 349–364.

Telle, S. & Thines, M. (2008). Amplification of *cox2* (~620 bp) from 2 mg of up to 129 years old herbarium specimens, comparing 19 extraction methods and 15 polymerases. *PLoS ONE* 3: e3584.

The Tomato Genome Consortium. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485: 635. doi:10.1038/nature11119

Wyman, S.K., Jansen, R.K. & Boore, J.L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20(17): 3252–3255.

Xu, C., Dong, W., Shi, S., Cheng, T., Li, C., Liu, Y., Wu, P., Wu, H., Gao, P. & Zhou, S. (2015). Accelerating plant DNA barcode reference library construction using herbarium specimens: Improved experimental techniques. *Molecular Ecology Resources* 15: 1366–1374. doi: 10.1111/1755-0998.12413

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

Yoshida, K., Sasaki, E. & Kamoun, S. (2015). Computational analyses of ancient pathogen DNA from herbarium samples: Challenges and prospects. *Frontiers in Plant Science* 6: Article number 771.

Zedane, L., Hong-Wa, C., Murienne, J., Jeziorsky, C., Baldwin, B.G. & Besnard, G. (2015). Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. *Biological Journal of the Linnean Society* 117: 44–57.