

1 **Extraction of ultrashort DNA molecules from herbarium**

2 **specimens**

3 Rafal M. Gutaker¹, Ella Reiter², Anja Furtwängler², Verena J. Schuenemann^{2,3}, Hernán

4 A. Burbano^{1,*}

5 ¹Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck

6 Institute for Developmental Biology, Tuebingen 72076, Germany

7 ²Institute of Archaeological Sciences, University of Tuebingen, Tuebingen 72076, Germany

8 ³Senckenberg Center for Human Evolution and Paleoenvironment, University of Tübingen, Tübingen

9 72076, Germany.

10 *Address correspondence to Hernán A. Burbano, Research Group for Ancient Genomics and

11 Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology. E-

12 mail: hernan.burbano@tuebingen.mpg.de

13 Keywords: ancient DNA; herbarium; DNA extraction; paleogenomics; next-

14 generation sequencing

15

16 **Abstract**

17 DNA extracted from herbarium specimens is highly fragmented and decays at a faster
 18 rate than DNA from ancient bones. Therefore, it is crucial to utilize extraction
 19 protocols that retrieve short DNA molecules. Improvements in extraction and library
 20 preparation protocols for animal remains have allowed efficient retrieval of molecules
 21 shorter than 50 bp. We adapted those improvements to extraction protocols for
 22 herbarium specimens and evaluated their performance by shotgun sequencing, which
 23 allows an accurate estimation of the distribution of fragment lengths. Extraction with
 24 PTB buffer decreased median fragment length by 35% when compared to CTAB.
 25 Modifying the binding conditions of DNA to silica allowed for an additional decrease
 26 of 10%. We did not observe a further decrease in length when we used single-
 27 stranded instead of double-stranded library preparation methods. Our protocol enables
 28 the retrieval of ultrashort molecules from herbarium specimens and will help to
 29 unlock the genetic information stored in herbaria.

30 **Method summary**

31 We optimized the extraction procedure for isolating ultrashort DNA fragments from
 32 herbarium specimens through combination of PTB lysis buffer and modifications
 33 previously used for ancient bones. We show the advantage of this protocol over others
 34 by estimating the DNA fragment length through shotgun sequencing.

35

36 Since ancient DNA (aDNA) is highly fragmented, it is particularly important to
 37 employ extraction protocols that retrieve ultrashort molecules (< 50 bp). It has been
 38 shown that a recently developed extraction protocol for animal remains efficiently

39 recovers those molecules (1), which has allowed sequencing highly fragmented
40 hominin (2) and cave bear remains (1) that are hundreds of thousands of years old.
41 DNA retrieved from herbarium specimens is also highly fragmented because it decays
42 six times faster than in bones (3). Consequently, DNA from century-old herbarium
43 specimens is as short as that of thousands of years old animal remains. To take full
44 advantage of the genetic information stored in those samples it is important to
45 optimize the extraction of ultrashort molecules from desiccated plant tissue.

46 We assessed the impact of extraction and library preparation methods on the
47 distribution of DNA fragment lengths in 20 *Arabidopsis thaliana* herbarium
48 specimens, which were collected between 1839 and 1898 (Table S1). We used a
49 hierarchical experimental design that includes three different phases due to limited
50 availability of tissue per sample (Figure 1). In phase one and two we used 10 *A.*
51 *thaliana* samples (~20 mg of leaf tissue each), which were subjected to two different
52 extraction protocols (~10 mg of tissue per treatment), followed by double-stranded
53 library preparation. To compare the performance of double- and single-stranded
54 library preparation methods, in phase three we applied single-stranded library
55 preparation method to DNA extracts produced by the most efficient DNA extraction
56 protocol in phase two. In each phase we evaluated the performance of the methods by
57 sequencing the genomic libraries with the Illumina MiSeq platform (Table S2).

58 Extraction buffers used for ancient bones and teeth are commonly composed
59 predominantly or exclusively of EDTA and proteinase K (4), reagents that are not
60 optimal for DNA extraction from plant tissue. Hence, in the first phase we tested two
61 commonly used DNA extraction buffers for historical plant specimens, which contain
62 either cetyl-trimethyl ammonium bromide (CTAB), or a mixture of N-
63 phenacylthiazolium bromide (PTB) and dithiothreitol (DTT) (5) (Figure 1). CTAB is

a strong detergent that under high salt concentrations binds to polysaccharides and aids their removal from the solution (6). Although CTAB is highly used in DNA extractions from modern plants, it has been shown that it does not have a detectable effect when applied to non-carbonized archaeobotanical remains (7). PTB is a substance that cleaves glucose-derived protein cross-links (8) and can help to release DNA trapped within sugar-derived condensation products (9); it has been effectively used to retrieve DNA from archaeobotanical remains (10). DTT digests disulfide bonds releasing thiolated DNA from cross-link complexes (11). In order to allow better comparison of the CTAB and PTB protocols, we replaced the ethanol precipitation step of the CTAB method with silica column binding (12) provided with the DNeasy® Plant Mini Kit. Subsequently, libraries were prepared using a double stranded DNA library protocol (13).

Based on qPCR measurements on unamplified libraries, the PTB protocol recover a higher number of unique library molecules than CTAB protocol (paired t-test $p = 0.007$) (Figure 2F and Table S3). We found that PTB decreases the median fragment length by 35% (from 88 to 57 bp) (paired t-test $p = 2.8e-06$) when compared to CTAB (Figure 2A and 2B). This decrease in length was also manifested as a higher proportion of damaged sites (λ) (paired t-test $p = 1.3e-06$) (Figure 2C), which represents the fraction of bonds broken in the DNA backbone (14, 15). In addition, DNA molecules extracted with PTB buffer showed more cytosine (C) to thymine (T) substitutions at the 5' end (paired t-test $p = 1.2e-06$; Figure 2G). C-to-T substitutions are typical damage patterns of aDNA and result from spontaneous deamination of C to uracil (U), which is read as T by the polymerase (16, 17). It is possible that shorter and more damaged fragments of DNA were released after cross-links were resolved by PTB and DTT, since there is a strong negative correlation between median

89 fragment length and C-to-T substitutions at first base ($R^2 = 0.44$; $p = 1.5e-07$; $N = 50$)
 90 (Figure S4). Alternatively, the observed variation in fragment length distribution
 91 could be explained by unknown chemical incompatibilities of lysis and binding
 92 buffer, i.e. certain reagents could in principle reduce DNA-binding properties of the
 93 buffer. Finally, in the CTAB protocol we apply a chloroform-isoamyl alcohol wash,
 94 which could also reduce recovery of short molecules.

95 In the second phase, to further increase the recovery of short fragments, we
 96 used PTB/DTT, which was the most successful extraction buffer in phase 1, and
 97 evaluated two systems for binding DNA to silica. We tested DNeasy® mini spin
 98 columns (Qiagen) in combination with the binding buffer used in the Plant Mini kit
 99 and MinElute® silica spin columns in conjunction with a binding buffer optimized for
 100 the recovery of short molecules from animal remains (1) (Figure 1). We found that the
 101 latter method decreased the median fragment length by 10% (from 60 to 54 bp)
 102 (paired t-test $p = 1.9e-04$), which shows that it is suitable to recover very short
 103 sequences also from herbarium specimens (Figure 2A and B). The frequency of C-to-
 104 T substitutions at the first base differed significantly between the two DNA binding
 105 methods (paired t-test $p = 3.3e-03$) (Figure 2G), with a decrease in median fragment
 106 length again being accompanied by an increase in C-to-T substitutions.

107 To investigate whether library preparation has an effect on fragment length
 108 distribution, in the third phase we produced single-stranded DNA (ssDNA) libraries
 109 using the extracts from the modified PTB/DTT extraction (18,19) and compared them
 110 to the dsDNA libraries constructed from the modified PTB extraction material of
 111 phase 2 (Figure 1). We did not observe a significant decrease of the median of the
 112 fragment length distribution in ssDNA libraries (paired t-test $p = 0.44$) (Figure 2A and
 113 B). Instead, the shape of the distribution changed towards larger numbers of longer

114 and shorter molecules at the cost of intermediate-size molecules, which is reflected in
115 decreased lambda (Figure 2C) and congruent with previous findings (19). Similarly to
116 Gansauge and Meyer (2013), we also detected a reduction of GC content in ssDNA
117 libraries when compared to dsDNA (Figure 2D). This phenomenon can be attributed
118 to a known bias in dsDNA libraries towards molecules with higher GC content
119 (20,21). We detected uniform GC content across the distribution of fragment lengths,
120 which suggests that the ssDNA library preparation protocol excels in reducing those
121 biases (Figure S8). In contrast to previous reports (19), the ssDNA library method did
122 not produce an increase in the proportion of endogenous DNA (Figure 2E, Figure S1,
123 S6). However, it has been suggested that increase in the proportion of endogenous
124 DNA occurs only when the initial content of endogenous DNA is lower than 10% (22,
125 23). Our *A. thaliana* samples have endogenous DNA between 16% and 94%, which
126 could explain why we did not detect a gain in endogenous DNA.

127 In summary, we demonstrate that the choice of extraction buffer has a great impact on
128 the length distribution of molecules recovered from herbarium specimens. Ultrashort
129 molecules are most efficiently retrieved using a combination of PTB/DTT mixture for
130 DNA extraction and the buffers and conditions suggested by Dabney *et al.* (2013) for
131 DNA binding. The two library preparation methods tested here appear to be equally
132 efficient in retaining short DNA fragments, however, while single stranded method
133 reduces GC bias in library it also decreases the fraction of endogenous DNA. We
134 present the DNA extraction protocol that increases the recovery of short fragments
135 and thus the accessibility of precious herbarium specimens for genetic analyses.

136

137

138 Author contributions

139 R.M.G, V.J.S, and H.A.B designed the experiments. R.M.G., E.R. and A.F.
140 performed the experiments. R.M.G analyzed the data. R.M.G and H.A.B wrote the
141 manuscript with contributions from all authors.

142 Acknowledgements

143 We thank curators from the Missouri Botanical Garden, University of Illinois,
144 National Museum of Natural History, West Virginia University, Harvard University
145 and New York Botanical Gardens for kindly providing samples for this study; Marco
146 Thines, Charles B. Fenster and Matthew T. Rutter for sampling the herbarium
147 specimens; Daniel Koenig for advice in the experimental design; Patricia Lang,
148 members of the Research Group for Ancient Genomics and Evolutions, and especially
149 Matthias Meyer for comments on the manuscript. This work was funded by the
150 Presidential Innovation Fund of the Max Planck Society.

151 Competing interests

152 The authors declare no competing interests

153

References

- 1.Dabney, J., M. Knapp, I. Glocke, M.T. Gansauge, A. Weihmann, B. Nickel, C. Valdiosera, N. Garcia, et al. 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America* 110:15758-15763.
- 2.Meyer, M., J.L. Arsuaga, C. de Filippo, S. Nagel, A. Aximu-Petri, B. Nickel, I. Martinez, A. Gracia, et al. 2016. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* 531:504-507.
- 3.Weiss, C.L., V.J. Schuenemann, J. Devos, G. Shirsekar, E. Reiter, B.A. Gould, J.R. Stinchcombe, J. Krause, and H.A. Burbano. 2016. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. *R Soc Open Sci* 3:160239.
- 4.Rohland, N. and M. Hofreiter. 2007. Comparison and optimization of ancient DNA extraction. *Biotechniques* 42:343-352.
- 5.Kistler, L. 2012. Ancient DNA extraction from plants. *Methods in molecular biology* 840:71-79.
- 6.Rogers, S.O. and A.J. Bendich. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant molecular biology* 5:69-76.
- 7.Wales, N., K. Andersen, E. Cappellini, M.C. Avila-Arcos, and M.T. Gilbert. 2014. Optimization of DNA recovery and amplification from non-carbonized archaeobotanical remains. *PloS one* 9:e86827.

- 177 **8.Vasan, S., X. Zhang, X. Zhang, A. Kapurniotu, J. Bernhagen, S. Teichberg, J.**
178 **Basgen, D. Wagle, et al.** 1996. An agent cleaving glucose-derived protein
179 crosslinks in vitro and in vivo. *Nature* 382:275-278.
- 180 **9.Poinar, H.N., M. Hofreiter, W.G. Spaulding, P.S. Martin, B.A. Stankiewicz, H.**
181 **Bland, R.P. Evershed, G. Possnert, and S. Paabo.** 1998. Molecular
182 coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops*
183 *shastensis*. *Science* 281:402-406.
- 184 **10.Jaenicke-Despres, V., E.S. Buckler, B.D. Smith, M.T. Gilbert, A. Cooper, J.**
185 **Doebley, and S. Paabo.** 2003. Early allelic selection in maize as revealed by
186 ancient DNA. *Science* 302:1206-1208.
- 187 **11.Gill, P., A.J. Jeffreys, and D.J. Werrett.** 1985. Forensic application of DNA
188 'fingerprints'. *Nature* 318:577-579.
- 189 **12.Palmer, S.A., Moore, J.D., A.J. Clapham, P. Rose and R.G. Allaby.** 2009.
190 Archaeogenomic evidence of ancient Nubian Barley evolution from six to
191 two-row indicates local adaptation. *PloS one* 4:e6301.
- 192 **13.Meyer, M. and M. Kircher.** 2010. Illumina sequencing library preparation for
193 highly multiplexed target capture and sequencing. Cold Spring Harbor
194 protocols 2010:pdb prot5448.
- 195 **14.Deagle, B.E., J.P. Eveson, and S.N. Jarman.** 2006. Quantification of damage in
196 DNA recovered from highly degraded samples--a case study on DNA in
197 faeces. *Frontiers in zoology* 3:11.
- 198 **15.Allentoft, M.E., M. Collins, D. Harker, J. Haile, C.L. Oskam, M.L. Hale, P.F.**
199 **Campos, J.A. Samaniego, et al.** 2012. The half-life of DNA in bone:
200 measuring decay kinetics in 158 dated fossils. *Proceedings. Biological*
201 *sciences / The Royal Society* 279:4724-4733.

- 202 **16.Briggs, A.W., U. Stenzel, P.L. Johnson, R.E. Green, J. Kelso, K. Prufer, M.**
203 **Meyer, J. Krause, et al.** 2007. Patterns of damage in genomic DNA
204 sequences from a Neandertal. *Proceedings of the National Academy of*
205 *Sciences of the United States of America* 104:14616-14621.
- 206 **17.Brotherton, P., P. Endicott, J.J. Sanchez, M. Beaumont, R. Barnett, J. Austin,**
207 **and A. Cooper.** 2007. Novel high-resolution characterization of ancient DNA
208 reveals C > U-type base modification events as the sole cause of post mortem
209 miscoding lesions. *Nucleic acids research* 35:5717-5728.
- 210 **18.Meyer, M., M. Kircher, M.T. Gansauge, H. Li, F. Racimo, S. Mallick, J.G.**
211 **Schraiber, F. Jay, et al.** 2012. A high-coverage genome sequence from an
212 archaic Denisovan individual. *Science* 338:222-226.
- 213 **19.Gansauge, M.T. and M. Meyer.** 2013. Single-stranded DNA library preparation
214 for the sequencing of ancient or damaged DNA. *Nature protocols* 8:737-748.
- 215 **20.Green, R.E., J. Krause, A.W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N.**
216 **Patterson, H. Li, et al.** 2010. A draft sequence of the Neandertal genome.
217 *Science* 328:710-722.
- 218 **21.Briggs, A.W., J.M. Good, R.E. Green, J. Krause, T. Maricic, U. Stenzel, C.**
219 **Lalueza-Fox, P. Rudan, et al.** 2009. Targeted retrieval and analysis of five
220 Neandertal mtDNA genomes. *Science* 325:318-321.
- 221 **22.Bennett, E.A., D. Massilani, G. Lizzo, J. Daligault, E.M. Geigl, and T. Grange.**
222 2014. Library construction for ancient genomics: single strand or double
223 strand? *Biotechniques* 56:289-290, 292-286, 298, passim.
- 224 **23.Wales, N., C. Caroe, M. Sandoval-Velasco, C. Gamba, R. Barnett, J.A.**
225 **Samaniego, J.R. Madrigal, L. Orlando, and M.T. Gilbert.** 2015. New

226 insights on single-stranded versus double-stranded DNA library preparation

227 for ancient DNA. Biotechniques 59:368-371.

228

229 **Figure 1. Experimental design for testing the effect of DNA extraction and**
 230 **library preparation protocols on properties of sequenced libraries from**
 231 **herbarium specimens.** Experiments were conducted in three phases. In phase one we
 232 subject 10 herbarium specimens of *Arabidopsis thaliana* to extraction with two
 233 different lysis buffers and compare sequencing results. In phase two we tested two
 234 DNA-binding methods on second set of 10 *A. thaliana* specimens. In phase three we
 235 compared the libraries constructed with double- and single-stranded methods.

236 **Figure 2. The effect of DNA extraction and library preparation protocols on**
 237 **different properties of DNA sequencing libraries.** The figure depicts the results
 238 from experiments in phases 1-3 (Figure 1). (A) Distribution of fragment lengths of
 239 merged reads mapped to the *Arabidopsis thaliana* reference genome. The y-axis
 240 shows the kernel density estimates. (B-G) Distributions represented as box and
 241 whisker plots; medians are depicted by thick black lines, boxes represent data
 242 between quartile Q1 and Q3, whiskers extend to 1.5 times the interquartile range
 243 between Q1 and Q3, and points symbolize outliers. Comparisons within experiments
 244 that result in significant differences in a paired t-test are connected with black lines
 245 ('***' indicates an alpha level of 0.005) (B) Fragment length medians. (C)
 246 Proportion of broken DNA fragments (lambda). (D) Proportion of GC content. (E)
 247 Proportion of endogenous DNA (proportion of reads mapped to *A. thaliana* reference
 248 genome). (F) Number of unique molecules per base of *A. thaliana* reference genome
 249 (molecule coverage of DNA extract) calculated from qPCR measurements on
 250 unamplified libraries. (G) Percentage of cytosine to thymine substitutions at first base
 251 at the 5' end.

Fig. 1

Library
preparation

Binding

DNA extraction

Lysis

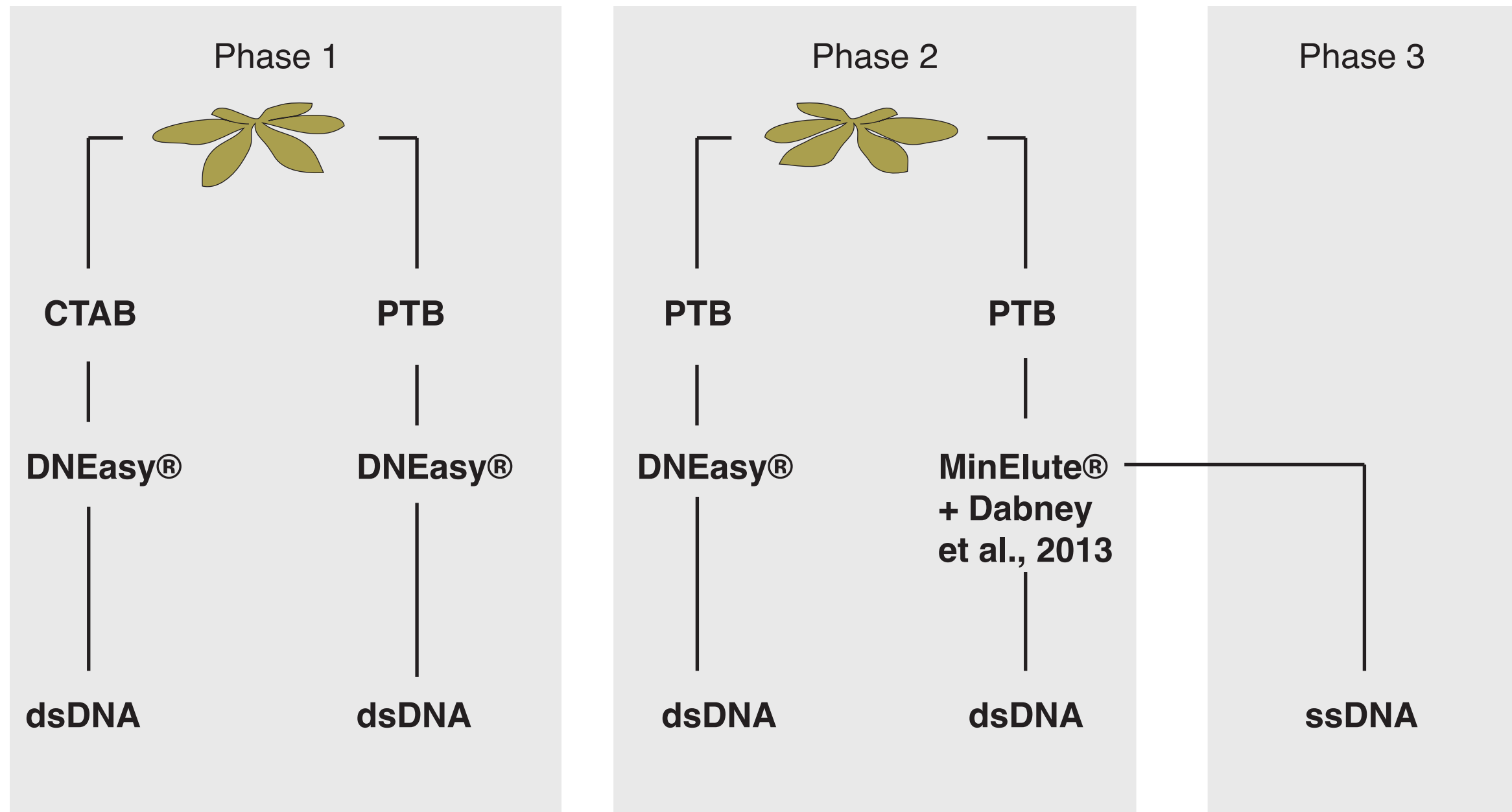


Fig. 2 A

