



**HAL**  
open science

# Extraction and selection of high-molecular-weight DNA for long-read sequencing from *Chlamydomonas reinhardtii*

Frédéric Chaux-Jukic, Nicolas Agier, Stephan Eberhard, Zhou Xu

► **To cite this version:**

Frédéric Chaux-Jukic, Nicolas Agier, Stephan Eberhard, Zhou Xu. Extraction and selection of high-molecular-weight DNA for long-read sequencing from *Chlamydomonas reinhardtii*. 2022. hal-03850623

**HAL Id: hal-03850623**

**<https://hal.science/hal-03850623v1>**

Preprint submitted on 14 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **Title:** Extraction and selection of high-molecular-weight DNA for long-read sequencing from  
2 *Chlamydomonas reinhardtii*

3

4 **Short title:** High-molecular-weight DNA extraction and selection from *Chlamydomonas*  
5 *reinhartii*

6

7 **Authors:** Frédéric Chaux-Jukic<sup>1,\*</sup>, Nicolas Agier<sup>1</sup>, Stephan Eberhard<sup>2</sup>, Zhou Xu<sup>1,\*</sup>

8

9 **Affiliations:**

10 <sup>1</sup>Sorbonne Université, CNRS, UMR7238, Institut de Biologie Paris-Seine, Laboratory of  
11 Computational and Quantitative Biology, 75005 Paris, France

12 <sup>2</sup>Sorbonne Université, CNRS, UMR7141, Institut de Biologie Physico-Chimique, Laboratory  
13 of Chloroplast Biology and Light-Sensing in Microalgae, 75005 Paris, France

14

15 \*Corresponding authors

16 E-mails: [zhou.xu@sorbonne-universite.fr](mailto:zhou.xu@sorbonne-universite.fr); [frederic.chaux-jukic@sorbonne-universite.fr](mailto:frederic.chaux-jukic@sorbonne-universite.fr)

17

18 **Keywords:** Algae DNA extraction, high-molecular weight DNA, long-read sequencing, size  
19 selection

## 20 Abstract

21 Recent advances in long-read sequencing technologies have enabled the complete assembly  
22 of eukaryotic genomes from telomere to telomere by allowing repeated regions to be fully  
23 sequenced and assembled, thus filling the gaps left by previous short-read sequencing  
24 methods. Furthermore, long-read sequencing can also help characterizing structural variants,  
25 with applications in the fields of genome evolution or cancer genomics. For many organisms,  
26 the main bottleneck to sequence long reads remains the lack of robust methods to obtain high-  
27 molecular-weight (HMW) DNA. For this purpose, we developed an optimized protocol to  
28 extract DNA suitable for long-read sequencing from the unicellular green alga  
29 *Chlamydomonas reinhardtii*, based on CTAB/phenol extraction followed by a size selection  
30 step for long DNA molecules. We provide validation results for the extraction protocol, as  
31 well as statistics obtained with Oxford Nanopore Technologies sequencing.

32

## 33 Introduction

34 In recent years, long-read sequencing technologies, such as the ones developed by Pacific  
35 Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore), have emerged as a  
36 solution to the pitfalls of short-read technologies in the detection of structural variants and in  
37 assembling repeated sequences and other complex regions (1). Additionally, because native  
38 DNA is used, long-read technologies can directly detect a variety of modified bases, including  
39 the most commonly studied methylated cytosines (2, 3). For their applications in genome  
40 assembly and structural variant detection, these technologies typically sequence DNA  
41 molecules ranging in size from kilobases to hundreds of kilobases as a continuous read. Reads  
42 traversing repeated sequences are necessary to correctly assemble neighboring regions, with  
43 longer reads enabling more contiguous genome assemblies. Today, the major bottleneck to  
44 sequence long reads comes from the ability to extract high-quality DNA devoid of polyphenol  
45 and polysaccharide contaminants with sizes compatible with this purpose. This is especially  
46 true for most plant tissues and algae cells, because polyphenols and polysaccharides are often  
47 co-extracted with DNA and can inhibit downstream applications such as sequencing (4, 5).

48  
49 *Chlamydomonas reinhardtii* is a unicellular green alga that is widely used as a model  
50 organism to study photosynthesis and cellular motility (6), and is an organism of choice for  
51 biotechnological application, with many synthetic biology tools being currently developed (7,  
52 8). In *C. reinhardtii*, as for other plants and algae, contending with phenolic and  
53 polysaccharide contaminants while preserving HMW DNA is a major challenge and requires  
54 an optimized protocol. PacBio and Nanopore sequencing have been performed on this  
55 organism, contributing to important advances in our understanding of its genome structure  
56 and content, base modifications and evolution (9-13). However, previous protocols did not  
57 include a size selection step, which can substantially enrich for longer molecules.

58 Here, we present a detailed protocol dedicated to efficiently extract and select HMW DNA  
59 from *C. reinhardtii* cells. The protocol minimizes DNA-shearing manipulations and  
60 comprises an additional step to enrich for HMW DNA. We validated the method by pulse-  
61 field gel electrophoresis (PFGE) and measurement of read length from Nanopore sequencing.  
62

## 63 Materials and Methods

64 The protocol described here is also published on [protocols.io](https://www.protocols.io),  
65 [dx.doi.org/10.17504/protocols.io.8epv59j9jg1b/v2](https://doi.org/10.17504/protocols.io.8epv59j9jg1b/v2) and is included as supporting information  
66 file 1 with this article.

67

### 68 *Nanopore sequencing*

69 Sequencing libraries were prepared as per manufacturer's recommendations, using NEBNext  
70 companion module (E7180S, NEB) and Ligation Sequencing Kit SQK LSK-109  
71 (Nanoporetech), except for the ligation time, which we increased to 30 min. For each run, 500  
72 ng were loaded on MinION flow cells (R9.4.1, Nanoporetech) and sequenced for 6h to 16h,  
73 depending on flow-cell kinetics. Libraries were loaded at least twice, with 1h wash using the  
74 manufacturer's washing buffer (EXP-WSH004) between runs. Basecalling was performed  
75 using Guppy (version 4.3.4) with parameters set to "high accuracy".

76

## 77 Results

78 We extracted genomic DNA following the presented protocol ([S1 Supplementary Protocol](#))  
79 and applied size selection using the Short Read Eliminator (SRE) kit (Circulomics), an easy-  
80 to-use method that does not require dedicated devices which is based on a length-dependent  
81 precipitation of nucleic acids driven by polyvinylpyrrolidone crowding. Large amounts of  
82 small DNA fragments can be detrimental for long-read Nanopore sequencing (14), not only  
83 because the subsequent reads are short, but also because these molecules can outcompete the  
84 longer ones, both for adapter ligation and pore usage, thus yielding suboptimal results.

85

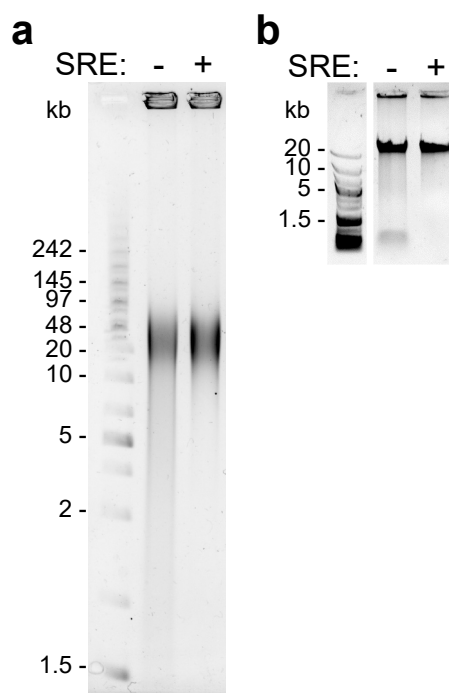
86 The size distribution of the extracted DNA was assessed by PFGE and Nanopore sequencing,  
87 with and without size-selection for HMW DNA. Samples were migrated in a pulse field,

88 stained by ethidium bromide and imaged with UV light (Fig. 1a). The DNA molecules  
89 extracted without size selection migrated as a large smear spread between approximately 1.5  
90 and 150 kb. After size selection with the SRE kit, the upper part of the distribution remained  
91 unchanged while the low-molecular-weight fragments (< 10 kb) were visibly reduced. We  
92 made a similar observation after electrophoresis and staining of the samples in a 0.3% agarose  
93 gel (Fig. 1b).

94

95 Size-selection of DNA fragments before preparation of libraries for Nanopore sequencing  
96 reproducibly led to a substantially decreased number of shorter molecules and an enrichment  
97 of longer ones (Fig. 2a-c), without negatively affecting read quality (Fig. 2d) and with no  
98 effect on genome-wide sequencing depth (S3 Supplementary Figure). Size-selection doubled  
99 the mean read length, increased the N50 from 12 kb to 17 kb (20 kb in two other  
100 experiments), with reads in the top decile being longer than 21 kb (26 kb and 27 kb in two  
101 other experiments) (S2 Table). The longest molecules we sequenced were over 100 kb, which  
102 are instrumental for genome assemblies.

103

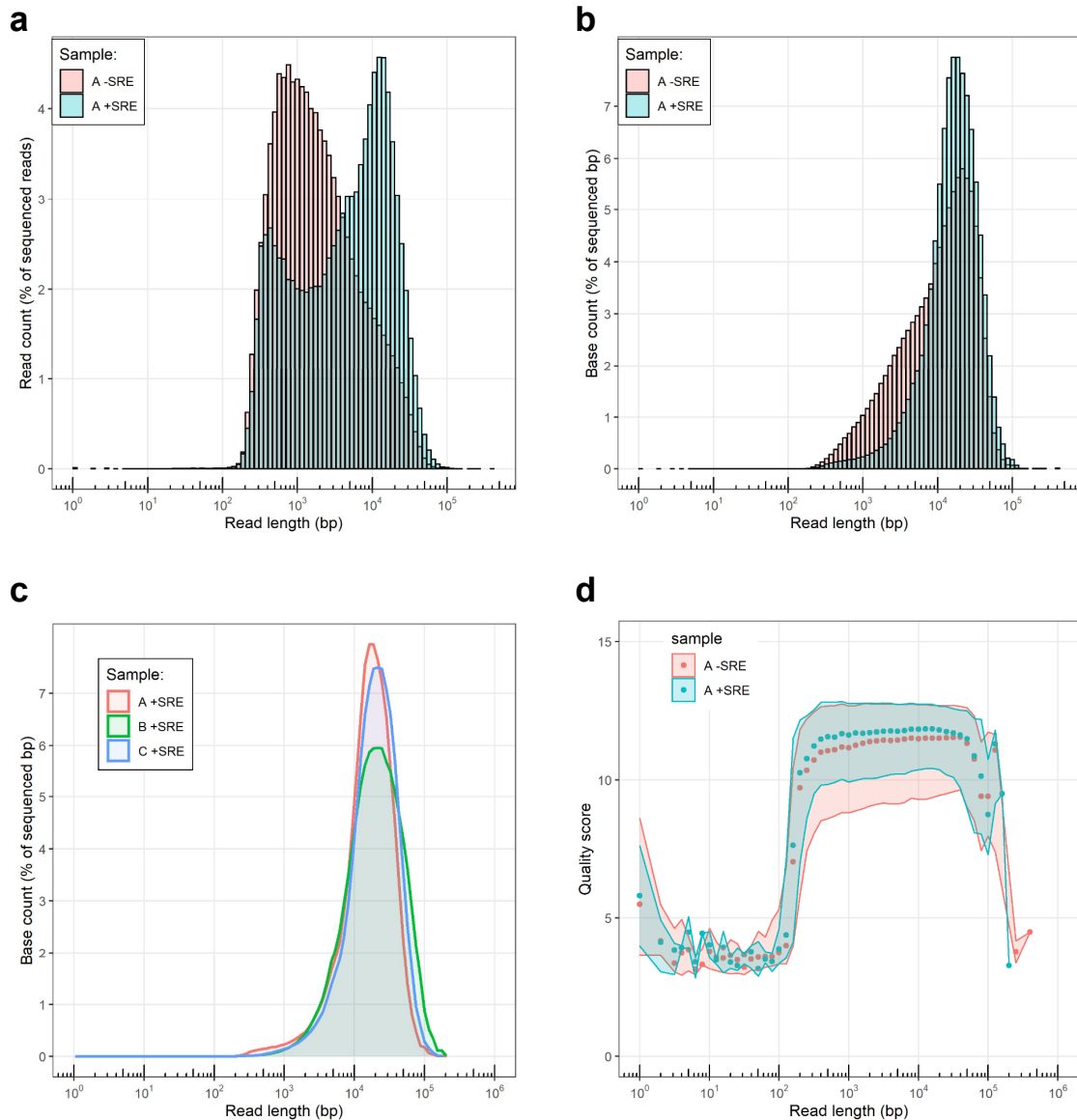


104

105 **Figure 1.** Visualization of extracted genomic DNA size distributions. (a) PFGE using 0.5  $\mu\text{g}$   
 106 of DNA prepared with (+) or without (-) SRE size-selection, embedded in 30  $\mu\text{l}$  of 0.5% low-  
 107 melting agarose plugs, migrated in a 1% SeaKem GTG agarose (Lonza) gel. The ladder is a  
 108 mix of PFG mid-range (N0342S, NEB) and GeneRuler 1 kb Plus (SM1331, ThermoFischer).  
 109 Electrophoresis conditions: 0.5X TBE (Tris Borate EDTA) buffer, 6  $\text{V}\cdot\text{cm}^{-1}$ , 120° angle, for  
 110 11h, switching time ramp from 1 to 60 seconds. Gel stained in ethidium bromide and imaged  
 111 with UV. (b) Standard gel electrophoresis (0.3% agarose) of the indicated samples. GeneRuler  
 112 1 kb Plus (SM1331, ThermoFischer) is used as the ladder.

113





114

115 **Figure 2.** Distributions of read length in Nanopore-sequenced datasets. (a-b) Count percentage  
 116 of (a) reads and of (b) bases as a function of read length obtained from genomic DNA of *C.*  
 117 *reinhardtii* (experiment “A”, see Table S2) with or without size selection (+SRE and -SRE).  
 118 (c) Count of bases after size-selection (+SRE) as a function of read length obtained from three  
 119 different samples (see S2 Table and S4 Supplementary Figure). (d) Quality score for individual  
 120 reads, grouped into bins of 0.1 log unit for samples “A-SRE” and “A+SRE”. The shaded areas  
 121 represent the values between the 1<sup>st</sup> and 3<sup>rd</sup> quartiles.

122

## 123 Acknowledgements

124 We thank Samuel O'Donnell for his help in the initial development of this protocol.

125

## 126 References

- 127 1. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its  
128 applications. *Nat Rev Genet.* 2020;21(10):597-614.
- 129 2. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al.  
130 Mapping DNA methylation with high-throughput nanopore sequencing. *Nat Methods.*  
131 2017;14(4):411-3.
- 132 3. Feng Z, Fang G, Korlach J, Clark T, Luong K, Zhang X, et al. Detecting DNA  
133 modifications from SMRT sequencing data by modeling sequence context dependence of  
134 polymerase kinetic. *PLoS Comput Biol.* 2013;9(3):e1002935.
- 135 4. Porebski S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol  
136 for plants containing high polysaccharide and polyphenol components. *Plant molecular*  
137 *biology reporter.* 1997;15(1):8-15.
- 138 5. Healey A, Furtado A, Cooper T, Henry RJ. Protocol: a simple method for extracting  
139 next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant*  
140 *Methods.* 2014;10:21.
- 141 6. Harris EH. *The Chlamydomonas Sourcebook*: Elsevier/Academic Press; 2009.
- 142 7. Scaife MA, Nguyen G, Rico J, Lambert D, Helliwell KE, Smith AG. Establishing  
143 *Chlamydomonas reinhardtii* as an industrial biotechnology host. *Plant J.* 2015;82(3):532-46.
- 144 8. Crozet P, Navarro FJ, Willmund F, Mehrshahi P, Bakowski K, Lauersen KJ, et al.  
145 Birth of a Photosynthetic Chassis: A MoClo Toolkit Enabling Synthetic Biology in the  
146 Microalga *Chlamydomonas reinhardtii*. *ACS Synth Biol.* 2018;7(9):2074-86.
- 147 9. O'Donnell S, Chaux F, Fischer G. Highly Contiguous Nanopore Genome Assembly of  
148 *Chlamydomonas reinhardtii* CC-1690. *Microbiol Resour Announc.* 2020;9(37).
- 149 10. Liu Q, Fang L, Yu G, Wang D, Xiao CL, Wang K. Detection of DNA base  
150 modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat*  
151 *Commun.* 2019;10(1):2449.
- 152 11. Chaux-Jukic F, O'Donnell S, Craig RJ, Eberhard S, Vallon O, Xu Z. Architecture and  
153 evolution of subtelomeres in the unicellular green alga *Chlamydomonas reinhardtii*. *Nucleic*  
154 *Acids Res.* 2021;49(13):7571-87.
- 155 12. Craig RJ, Yushenova IA, Rodriguez F, Arkhipova IR. An Ancient Clade of Penelope-  
156 Like Retroelements with Permuted Domains Is Present in the Green Lineage and Protists, and  
157 Dominates Many Invertebrate Genomes. *Mol Biol Evol.* 2021;38(11):5005-20.
- 158 13. Craig RJ, Hasan AR, Ness RW, Keightley PD. Comparative genomics of  
159 *Chlamydomonas*. *Plant Cell.* 2021.
- 160 14. Delahaye C, Nicolas J. Sequencing DNA with nanopores: Troubles and biases. *PLoS*  
161 *One.* 2021;16(10):e0257521.

162

163

164

165 Supporting information

166

167 **S1 Protocol.** Step-by-step protocol, also available on protocols.io:

168 [dx.doi.org/10.17504/protocols.io.8epv59j9jg1b/v2](https://doi.org/10.17504/protocols.io.8epv59j9jg1b/v2)

169

170 **S2 Table. Summary statistics for 6 DNA preparations and sequencing experiments.**

171 Major limiting outputs are shown in red.

Sample	CLIP library <sup>a</sup>	DNA extraction	Library preparation	Read count	Base count (Gb)	Read length quantiles (kb)			Median read length weighted by base count ("N50") (kb)	Top 3 longest reads <sup>b</sup> (kb)		
						Median	Q90	Q99				
A +SRE	LMJ.RY0402.077111	This protocol	LSK109 +barcodes	282 076	2,0	5,0	21	44	17	142	121	120
A -SRE		This protocol (except SRE)	LSK109	850 066	3,6	1,5	12	34	12	127	117	112
B +SRE	CC-4533 (cMJ030)	This protocol	LSK109	280 614	3,1	6,7	26	64	20	213	201	191
B -SRE		Monarch HMW DNA kit <sup>c</sup>	LSK109	177 794	1,1	2,5	15	53	15	199	177	175
C +SRE	LMJ.RY0402.209904	This protocol	LSK110	344 223	3,9	8,0	27	54	20	175	166	161
C -SRE		DNeasy Plant/Tip100 <sup>d</sup>	LSK109	81 078	0,6	3,0	22	54	21	121	119	108

172

173 <sup>a</sup> <https://www.chlamylibrary.org> and reference (15).

174 <sup>b</sup> with quality > 7.

175 <sup>c</sup> as per manufacturer's protocol (Monarch<sup>®</sup> HMW DNA Extraction Kit for Tissue Cat. no. T3060L, New England  
176 Biolabs).

177 <sup>d</sup> cell lysis using DNeasy Maxi Plant (Cat. no. 68163, Qiagen) as in (16) and purification using Genomic-tip 100/G  
178 (Cat. no. 10243, Qiagen), then AMPure beads (Cat. no. A63880, Beckman Coulter).

179

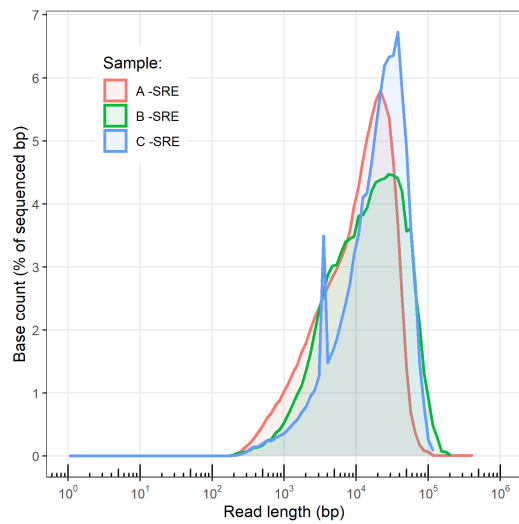


180

181 **S3 Supplementary Figure.** Sequencing depth normalized to the median across the whole

182 genome of the sequencing reads for all chromosomes, using DNA obtained with (+) or

183 without (-) SRE size selection.



184

185 **S4 Supplementary Figure.** Count percentage of bases as a function of read length with

186 alternative sample preparations without size selection (-SRE). See [Table S2](#) for details.

187 Sample C was sequenced in the presence of control DNA (“DNA CS” from Oxford Nanopore

188 sequencing), which peaked at 3 kb.

189