

1 ***A comprehensive toolkit to enable MinION sequencing in any laboratory***

2

3 Authors:

4 Miriam Schalamun^{a,b}, David Kainer^a, Eleanor Beavan^a, Ramawatar Nagar^a, David Eccles^c,
5 John P. Rathjen^a, Robert Lanfear^{a,\$}, Benjamin Schwessinger^{a,\$}

6

7 Research School of Biology, The Australian National University, Acton 2601, ACT,
8 Australia^a; Current address: University of Natural Resources and Life Sciences, 1190
9 Vienna, Austria^b, Malaghan Institute of Medical Research, Wellington, New Zealand^c

10 ^{\$} Corresponding authors:

11 Robert Lanfear, email: rob.lanfear@anu.edu.au, Benjamin Schwessinger, email:
12 benjamin.schwessinger@anu.edu.au

13 Keywords: Long-read sequencing, Genomics, Nanopore, high molecular weight DNA

14

15 **Abstract**

16 Long-read sequencing technologies are transforming our ability to assemble highly complex
17 genomes. Realising their full potential relies crucially on extracting high quality, high
18 molecular weight (HMW) DNA from the organisms of interest. This is especially the case for
19 the portable MinION sequencer which potentiates all laboratories to undertake their own
20 genome sequencing projects, due to its low entry cost and minimal spatial footprint. One
21 challenge of the MinION is that each group has to independently establish effective protocols
22 for using the instrument, which can be time consuming and costly. Here we present a
23 workflow and protocols that enabled us to establish MinION sequencing in our own
24 laboratories, based on optimising DNA extractions from a challenging plant tissue as a case
25 study. Following the workflow illustrated we were able to reliably and repeatedly obtain > 8.5
26 Gb of long read sequencing data with a mean read length of 13 kb and an N50 of 26 kb.
27 Our protocols are open-source and can be performed in any laboratory without special
28 equipment. We also illustrate some more elaborate workflows which can increase mean and
29 average read lengths if this is desired. We envision that our workflow for establishing
30 MinION sequencing, including the illustration of potential pitfalls, will be useful to others who
31 plan to establish long-read sequencing in their own laboratories.

32

33 **Introduction**

34 Single-molecule nanopore sequencing records changes in electrical current as individual
35 tagged DNA molecules pass through an engineered pore across a chemical gradient (Jain,
36 Olsen, Paten, & Akeson, 2016). Groups of consecutive bases cause a characteristic shift in
37 current, and this can be deconvoluted to infer the individual base sequence of the DNA
38 molecule, a process referred to as basecalling. This technology can sequence DNA
39 fragments of varied lengths, from a few hundred bases to over a megabase (Mb), which
40 compares favorably to sequencing by synthesis (e.g. Illumina), which is limited to hundreds
41 of bases (Leggett & Clark, 2017). Long reads have a number of important applications,

42 including: improving the accuracy and efficiency of genome assembly, especially for
43 genomes that contain long low-complexity regions; detailed investigation of segmental
44 duplications and structural variation (Jain et al., 2018); major histocompatibility complex
45 (MHC) typing (Liu et al., 2017); and detecting methylation patterns (Simpson et al., 2017).
46 The number of genome assemblies using nanopore data either exclusively or in combination
47 with other sequencing data is steadily increasing, for example the 3.5 gigabase (Gb) human
48 genome, the 860 Mb European eel genome, the 1 Gb genome of the wild tomato species
49 *Solanum pennellii*, and the 135 Mb genome of *Arabidopsis thaliana* (Jain et al., 2018;
50 Jansen et al., 2017; Michael et al., 2018; Schmidt et al., 2017). In short, nanopore
51 sequencing solves the technical challenges of reading long DNA fragments, while still having
52 room for improvement in terms of accuracy. One of the primary remaining challenges is to
53 extract and purify very long DNA fragments from the organisms or tissues of interest.

54 The Oxford Nanopore Technologies (ONT) MinION makes long-read sequencing accessible
55 to most laboratories outside of a dedicated genome facility. It has very low capital cost; has
56 the potential to generate more than 1 Gb of sequence data per 100 USD; has a footprint
57 about the size of an office stapler; and runs on a standard desktop or laptop computer. The
58 MinION uses small consumable flowcells for sequencing, which contain fluid channels that
59 flow samples onto a sequencing matrix, and provide a small amount of fluid waste storage.

60 This democratization of sequencing brings the challenge that every laboratory has to
61 establish the sequencing platform and concomitantly, new DNA extraction and library
62 preparation protocols. This can be challenging and time consuming. Here we illustrate the
63 workflow we applied to establish MinION sequencing in our laboratories using the tree
64 species *Eucalyptus pauciflora* as a case study. It is challenging to extract high purity and
65 high molecular weight DNA from *E. pauciflora* because the mature leaf tissue is physically
66 tough, and because it contains very high levels of secondary metabolites which are known to
67 reduce the efficacy of DNA extraction protocols (Healey, Furtado, Cooper, & Henry, 2014).
68 We illustrate reliable and repeatable ways of measuring DNA purity to optimise output from
69 the MinION sequencer. We discuss important considerations for DNA library preparation,
70 and methods to control and optimise the final distribution of read lengths. We show that
71 during DNA extraction, small alterations in sample homogenisation protocols can drastically
72 alter DNA fragment lengths; introduce a novel low-tech size selection protocol based on
73 Solid Phase Reversible Immobilization (SPRI) beads; and compare size selection methods
74 using electrophoresis versus DNA shearing. Finally, we introduce an open-source MinION
75 user group that shares DNA extraction, size-selection, and library preparation protocols for
76 many additional organisms, making our workflow applicable well beyond the case study
77 presented here.

78

79

80

81

82

83

84

85 Results

86 Optimizing sequencing output

87 DNA Sample Purity

88 The first goal of our project was to optimize extraction protocols to yield highly intact and
89 high purity DNA suitable for long-read sequencing. High purity of DNA is defined by
90 Nanodrop spectrophotometer (Thermo Fisher) absorbance of DNA with a 260/280 nm ratio
91 between 1.8 to 2.0, and a 260/230 nm ratio between 2 and 2.2 (Desjardins & Conklin, 2010;
92 Mackey & Chomczynski, 1997). In addition to this we found it critical that the ratio of DNA
93 concentrations measured on the Qubit and Nanodrop instruments respectively should be
94 1:1.5. This ratio indicates that most DNA molecules are double-stranded and that no other
95 molecules (e.g. RNA) are present that absorb at 260 nm (e.g. Qubit: 100 ng/μl; Nanodrop:
96 150 ng/μl gives an acceptable ratio of 1:1.5; (O'Neill, McPartlin, Arthure, Riedel, & McMillan,
97 2011).

98 In our workflow, we first aimed to recover high molecular weight DNA with a Nanodrop/Qubit
99 concentration ratio that was close to one. We then optimized DNA purity based on 260/280
100 nm ratios, which are indicative of protein contamination, and 260/230 nm ratios, which are
101 indicative of contamination by salts, phenol, and carbohydrates (O'Neill et al., 2011). To
102 achieve this, we first tested a well-established hexadecyltrimethylammonium bromide
103 (CTAB) extraction protocol to extract DNA from *E. pauciflora* leaves collected in June 2017
104 from adult trees in the Kosciuszko National park near Thredbo, New South Wales, Australia
105 (Healey et al., 2014; Schwessinger & Rathjen, 2017). While the CTAB protocol returned
106 good yields of double stranded DNA (~5 μg DNA per g tissue), the Qubit/Nanodrop ratio of
107 0.05 indicated significant contamination with RNA or single-stranded DNA. Nanodrop
108 absorption spectra from 220 to 350 nm (Figure 1A) revealed the presence of contaminants
109 as the curve deviated drastically from pure DNA absorption curves (Figure 1D). In such
110 cases it is often recommended to clean the DNA using SPRI paramagnetic beads in
111 combination with a polyethylene glycol (PEG) and sodium chloride (NaCl) mixture, such as
112 AMPure XP beads (Beckman Coulter). These beads bind to the DNA but most contaminants
113 do not and can be washed away (Krinitzina, Sizova, Zaika, Speranskaya, & Sukhorukov,
114 2015; Mayjonade et al., 2016). We were able to improve sample quality slightly by adding
115 the standard measure of 0.45 vol (V/V) AMPure XP beads (Figure 1B), but repeating this
116 step did not increase the purity of the DNA further. Next we tested an extraction method
117 employing the detergent sodium dodecyl sulfate (SDS) which contains a PEG-NaCl
118 precipitation step to capture the DNA onto SPRI beads. This approach has been reported to
119 work well with many species including sunflower, human, and *Escherichia coli* (Mayjonade et
120 al., 2016). Using this approach we recovered high levels of double-stranded DNA
121 (Qubit/Nanodrop = 1:1.5), but the Nanodrop absorption curves still indicated the presence of
122 contaminants in the final DNA extract (Figure 1C). Again, we were unable to improve the
123 DNA purity by repeated SPRI clean-up steps as the 260/280 nm ratios did not improve. As
124 an alternative method we cleaned the crude DNA obtained from the SDS-based method
125 using a chloroform:isoamylalcohol extraction followed by isopropanol precipitation of the
126 DNA, as described for some fungal DNA samples (Dong, 2017). This consistently resulted in
127 high purity DNA with Qubit/Nanodrop ratios of 1:1-1.5, 260/280 nm ratios of ~1.8, 260/230
128 nm ratios of ~2.0, and excellent Nanodrop absorbance curves (Figure 1D).

129 ONT 1D library preparations involve the ligation of sequencing adapters at both 3' ends of
130 end-repaired double-stranded DNA. Sequencing adapters carry a motor protein that guides
131 the DNA to the pore and regulates the translocation speed of the DNA across the pore. In
132 addition, they carry a characteristic DNA sequence which is used by basecallers to
133 recognize the translocation start of a new DNA molecule (Jain et al., 2016; Leggett & Clark,

134 2017). We tested the effect of sample impurities on MinION output using the 1D ligation
135 protocol. Our three samples differed primarily in their 260/230 nm ratios. One sub-optimal
136 sample (sample 5, Table 1) had a low ratio of 1.0, and the other two samples (samples 10
137 and 27, Table 1) had close-to-optimal ratios of 2.1 and 2.3 respectively. The sample with the
138 low 260/230 nm ratio yielded an order of magnitude less sequence data from a single
139 flowcell compared to the other two samples (0.7 Gb vs. ~7 Gb respectively, Table 1,
140 Supplemental Table 1). It seems likely that the contaminants causing the reduced 260/230
141 nm ratio inhibited the library preparation or the sequencing itself. Based on this observation
142 we recommend adhering to the DNA quality measures nominated above whenever possible,
143 or else to assume reduced sequencing outputs. We also advise establishing suitable DNA
144 extraction methods well in advance of ordering sequencing materials; our experience
145 suggests that optimizing DNA extraction protocols can take several months.

146 *Sequencing library preparation*

147 The manufacturer-recommended library preparations involving DNA repair and end-prep are
148 optimized for 0.2 pmol of input DNA with an average fragment size of 8 kb, which in turn
149 requires 1 µg of double-stranded DNA. This implies that the DNA input as expressed in
150 mass needs to be adapted according to the concentration of free DNA ends available for
151 adapter ligation, which is a function of fragment length (Mayjonade, 2018; Schwessinger,
152 2018). The molarity of the DNA sample can be calculated using the Promega BioMath
153 calculator (<http://www.promega.com/a/apps/biomath/>) which requires the average fragment
154 length to calculate the respective DNA mass for 0.2 pmol. For example, 0.2 pmol of DNA of
155 mean length 24 kb requires a DNA input of 3 µg. In our case we estimated a mean DNA
156 fragment length of ~30 kb based on 1% agarose gel electrophoresis (Figure 2) and pulsed
157 field gel electrophoresis (PFGE) which provides higher resolution in the high molecular
158 weight range (Figure 3). When estimating mean DNA fragment length based on fluorescent
159 intensity (e.g. after staining with SYBR red or ethidium bromide) it is important to consider
160 that smaller DNA molecules incorporate less dye so appear fainter during imaging. For
161 example, even faint DNA smears below 10 kb can indicate the significant presence of short
162 DNA fragments that are best avoided if long-read lengths are a primary goal of the
163 sequencing effort (see below.) Failure to account for this can easily lead to overestimation of
164 mean DNA fragment length, and miscalculation of the true concentration of DNA fragments.

165 As a starting point we defined the optimal DNA input based on our initial mean fragment
166 length estimate of 30 kb. This was followed by empirical adjustments from plotting
167 sequencing outputs vs. the DNA input into adapter ligation (Figure 4). This approach
168 revealed an optimum of ~2 µg dsDNA (Figure 4), which required an input of 2.9 µg DNA for
169 the DNA preparation stage considering typical losses of 30% after clean-up using in house
170 SPRI beads (see below). Neither decreasing or increasing the DNA input improved the
171 sequencing output, due to too few adapter-DNA molecules, or too many free DNA molecules
172 damaging DNA pores. Assuming that 2.9 µg input DNA was the equivalent of 0.2 pmol, we
173 estimate a mean DNA fragment length of 23 kb for our sample preparation. This suggests
174 we initially overestimated the mean DNA fragment length, and highlights the difficulty of
175 estimating these values based on gel imaging. At the same time, we stress that it is best to
176 establish optimal DNA inputs empirically for each DNA extraction and/or shearing protocol.
177 In addition, one can use the sequence read-length distribution from the initial flowcells to
178 improve the estimate of the mean fragment length of the DNA extracted from the tissue. This
179 approach can help to quickly optimise the amount of input DNA added to the ligation step.

180

181 **Altering DNA fragment length and DNA read length**

182 Several factors influence DNA stability during extraction, including chemical properties of the
183 buffer and the physical forces applied during tissue homogenisation, phase separation, and
184 pipetting (Klingstrom, Bongcam-Rudloff, & Pettersson, 2018). The buffer composition is the
185 least flexible factor, especially for difficult tissues such as field samples of eucalyptus leaves
186 that require complex buffers for DNA extraction (see above). In contrast, the conditions
187 during tissue homogenisation can be adjusted more easily by changing treatment type and
188 length. Optimizing these parameters is very important when optimizing DNA fragment length.

189 To demonstrate this effect, we compared DNA fragment length with sequencing read lengths
190 between two sets of samples that were subjected to different tissue homogenisation
191 procedures. Our standard tissue homogenisation method for eucalyptus leaves consisted of
192 crushing frozen samples for 35 seconds with two 5 mm metal beads in a Qiagen tissuelyser
193 at 24 Hz. To maintain the frozen state, each Eppendorf tube as well as the grinding rack
194 were frozen in liquid nitrogen before the homogenisation step. In an attempt to improve
195 throughput, we tested the effect of homogenizing samples in larger batches, which likely led
196 to a situation where not all samples were completely frozen throughout the procedure. This
197 small change in handling clearly impacted the DNA fragment length distribution as estimated
198 by 0.8% agarose gel electrophoresis. DNA samples extracted using our standard method
199 migrated largely as a single high molecular weight DNA band at the upper limit of resolution
200 (~23 kb), and well above the 10 kb size standard. For this sample we observed only a light
201 smear visible to 2.5 kb. In contrast, the tissue sample treated in larger batches showed an
202 enhanced low molecular weight smear visible to 1 kb (Figure 2) in addition to the large HMW
203 band. This suggests that the average DNA fragment length was reduced in this sample. To
204 more accurately assess the effect of the change in tissue handling, we ran the second DNA
205 extraction on a single flowcell, and compared the results to those of two flowcells loaded with
206 DNA prepared using the standard (constantly frozen) tissue handling method. The relatively
207 subtle increase in visible DNA smearing on the agarose gel (Figure 2) belied a drastic shift in
208 read length distributions; the mean read length dropped from ~13 kb to 4.9 kb, and the
209 median from ~7 kb to 2.5 kb (Table 2). This illustrates that even a slight change in DNA
210 smearing can have a huge impact on sequencing output. It is therefore important to carefully
211 assess DNA fragment length, if possible by comparison to other samples, by agarose gel or
212 PFGE to avoid short sequence reads.

213 Because our focus for this project was on generating reads >5 kb to assemble a repeat-rich
214 genome *de novo*, we reasoned that it would be beneficial to remove smaller DNA fragments
215 (<1-2 kb) from all samples. AMPure XP beads can be used to size-select DNA fragments in
216 the range 100-500 bp (He, Zhu, & Gu, 2013; Schmitz & Riesner, 2006). However, it is not
217 possible to remove DNA fragments larger than ~500 bp with AMPure XP beads (Figure 5) as
218 adding less than 0.4 vol (V/V) of bead solution causes the NaCl concentration to fall below
219 0.4 M, leading to complete sample loss (He et al., 2013). We reasoned that by adjusting the
220 PEG and NaCl concentrations, which precipitate DNA in a cooperative manner, we might be
221 able to select a higher average DNA fragment length, and thereby remove unwanted smaller
222 DNA fragments (Lis & Schleif, 1975; Ramos, de Vries, & Ruggiero Neto, 2005). Using 0.8
223 vol (V/V) of our adjusted SPRI beads mixture (which translates to final PEG concentrations
224 of 4.8% (v/v) and 0.7 M NaCl) we were able to remove DNA fragments of up to 1.5 kb
225 (Figure 5) (Schalamun & Schwessinger, 2017). We used this adapted SPRI beads mixture
226 subsequently for DNA sample clean up and during library preparation.

227 We next assessed the effect of DNA shearing and gel-based size-selection procedures on
228 sequencing throughput and read length distribution. In the case of DNA shearing our
229 hypothesis was that a more unimodal size distribution of shorter DNA fragments with a peak
230 of about 20 kb (Figure 3) would increase sequencing throughput. We used g-TUBEs with a
231 benchtop centrifuge to shear DNA by forcing it through a μm mesh. DNA shearing did not
232 increase yield, but did affect the read length distribution (Table 3). Compared with non-
233 sheared samples, the sequence read length distribution from sheared reads shifted to
234 smaller values and peaked at about 11 kb (Figure 6), with an $N50_{Q7}$ of 18 kb, compared to
235 an $N50_{Q7}$ of ~ 26 kb from the unsheared samples (Table 3). Whereas Q7 presents the default
236 quality threshold from the basecaller. Interestingly, the median read length from the sheared
237 DNA samples increased to 7.5 kb from 6.5 kb when compared to unsheared DNA. At the
238 same time low quality short reads were reduced in the sheared samples. Possibly, removing
239 long DNA fragments (> 50 kb) leads to fewer low quality reads caused by long DNA
240 molecules being stuck in the pore, at least when using the R9.5 pore chemistry. This
241 highlights that filtering reads based on their q-scores, as well as removing short sequencing
242 reads, may help to avoid error propagation during downstream analyses of the data.

243 We also tested the effect of removing DNA fragments below 20 kb by size selection using
244 the BluePippin system in the high-pass mode which enables the collection of DNA molecules
245 above a certain size. When we applied the 20 kb high-pass filter we were able to remove
246 DNA fragments less than 20 kb while maintaining the high molecular weight size distribution
247 (Figure 3). After sequencing, the read length $N50_{Q7}$ increased to 35 kb from 26 kb, while the
248 mean and median read-lengths increased to 26 and 23 kb from 12 and 6.5 kb respectively
249 (Table 4 and Figure 3). The main drawbacks of BluePippin high-pass size selection were the
250 high sample loss (65-75%), the increase in cost, and prolonged sample handling.

251 Overall, we did not employ DNA shearing using g-TUBEs or size selection via BluePippin in
252 our final sequencing protocol even though they improved sequencing read-length
253 distributions. In our case, the high quality sequencing results achieved with our standard
254 protocol using the improved SPRI beads mixture did not warrant the additional time and
255 financial investment required when incorporating g-TUBEs DNA shearing or BluePippin size
256 selection into our workflow.

257 **Real time and between run evaluation**

258 The software MinKNOW makes it possible to perform a real time monitoring during the
259 MinION sequencing run. Interpreting the pore signal statistics and the length graph during
260 the first two hours of sequencing gives the user a clear idea if the run should be continued or
261 stopped. We used this feature of MinKNOW to optimize our runs. First, we evaluated pore
262 occupancy, defined as the ratio of 'in strand' (light green) to the sum of 'in strand' plus 'single
263 pores', after one hour. A high pore occupancy ($>70\%$) indicates successful library
264 preparation and is predictive of a high final sequencing output. If the pore occupancy was
265 below 70% we stopped the run, washed the flowcell and loaded a new library to ensure high
266 throughput per flowcell (Figure 8). We reasoned that these low throughput runs were usually
267 due to insufficient DNA molecules being ligated to sequencing adapters during the library
268 preparation. We found that we had to load at least 1 μg library DNA onto a flowcell to
269 achieve acceptable yields (Figure 4). To ensure sufficient adapter ligated DNA, we started
270 library preparation with at least 4 μg of high quality starting DNA to account for potential
271 losses during the SPRI bead clean-up steps. A second pore statistic to consider is the
272 number of unavailable pores, e.g. 'zero' (black), 'unavailable' (light blue), or 'active feedback'

273 (pink) (Mayjonade, 2018). If these numbers increase too quickly in the first few hours of the
274 run it is likely that the library is contaminated and the pores are being irreversibly blocked or
275 damaged, or that the membrane has ruptured. If recognized early enough the flowcell can be
276 washed and a new library loaded, but the pores cannot always be recovered. Furthermore,
277 the length distribution from the length graph can be easily assessed and, if unsatisfactory,
278 the library exchanged for a separately prepared sample (Figure 8). We also recommend to
279 track the sequencing run with a continuous screenshot application (e.g. newlapse for linux,
280 <https://github.com/mtib/newlapse>), in addition to visual inspection during the first few hours
281 of the sequencing run. This enables continuous monitoring and assessment of unusual
282 sequencing behaviours out of hours.

283 One key to ongoing optimisation of flowcells in our laboratories was the tracking of all
284 parameters for each sequencing run using our monitoring spreadsheet (Supplemental Table
285 2) and a continued comparison of the output of each additional flowcell. After running each
286 flowcell, we used ONT's Albacore 2.0 basecaller to convert the raw signal data from the
287 MinION into DNA sequence data in fastq format. Albacore 2.0 produces a
288 sequencing_summary.txt file which contains a summary of every sequencing read, and can
289 be used for rapid assessment of each flowcell using the minionQC script
290 (https://github.com/roblanf/minion_qc). After basecalling each flowcell, we ran this script and
291 examined in detail the length and mean quality distributions of the reads, and the physical
292 performance map of the flowcell. This allowed us to continually evaluate and improve our
293 protocols for each flowcell. Before we were halfway through our project, we were able to
294 reliably and repeatedly obtain more than 6 Gb of data from each flowcell, with mean read
295 lengths consistently above 12 kb.

296 **Discussion**

297 Here we present a complete workflow to establish MinION long read sequencing in any
298 laboratory. We highlight the importance of clean high molecular weight DNA for successful
299 sequencing runs and provide detailed wet lab DNA extraction and purification protocols that
300 include size selection. All these protocols and many others applicable to different starting
301 material, some provided by other community members, are freely available on the open-
302 access protocol sharing repository protocols.io in form of a MinION user group
303 (<https://www.protocols.io/groups/minion-user-group-with-fungi-and-plants-on-their-mind>)
304 (Schwessinger, 2016) . We encourage others to contribute to this open science platform to
305 accelerate research and for the community to save costs when establishing long read DNA
306 sequencing in their own laboratory. High quality 'living' protocols with careful run and run-to-
307 run evaluations as described here will facilitate knowledge generation instead of constant
308 'reinvention of the wheel'.

309 **Acknowledgment**

310 We would like to acknowledge fruitful discussion, leading to and improving this manuscript,
311 with the following; Louise Judd, Ken McGrath, Baptiste Mayjonade, David Hayward, Josh
312 Quick, and Megan McDonald. We would also like to acknowledge all contributors of the
313 MinION usergroup on protocols.io for sharing their protocols openly.

314

315 **Methods**

316 ***Tissue collection***

317 *Eucalyptus pauciflora* leaf tissue was collected from Thredbo, New South Wales (NSW),
318 Australia. After harvesting the young twigs were transported in plastic bags and stored in
319 darkness at 4°C in water until DNA extraction.

320

321 ***High molecular weight DNA extraction and clean up***

322 We extracted high molecular weight DNA based on Mayjonade's DNA extraction protocol
323 optimized for our eucalyptus samples (Mayjonade et al., 2016). Each extraction was carried
324 out with 800 - 1000 mg leaf tissue which was cut into small pieces and split between 8
325 separate 2 mL Eppendorf tubes, each containing 2 metal beads of 5 mm in diameter, before
326 freezing in liquid nitrogen. We lysed the tissue mechanically by grinding using the Qiagen
327 tissue lyzer II for 35 seconds at 25 Hz. Pulverised tissue was suspended in 700 µL SDS lysis
328 buffer (1% w/v PVP40, 1% w/v PVP10, 500 mM NaCl, 100 mM Tris-HCl pH 8.0, 50 mM
329 EDTA, 1.25% w/v SDS, 1% w/v sodium metabisulfite, 5 mM DTT, Milli-Q water and heated
330 to 64°C for 30 minutes to inactivated DNases. One µL RNase A (10 mg/mL) (Thermo Fisher)
331 per 1 mL lysis buffer was added to the mixture after the heat treatment, followed by
332 incubation at 37°C for 50 minutes at 400 rpm on a thermomixer. Twenty minutes into the
333 incubation we added 10 µl Proteinase K (800 Units/mL) (NEB). To precipitate proteins, the
334 tubes were cooled on ice for 2 min before adding 0.3 vol (210 µL) 5 M potassium acetate pH
335 7.5 and mixed by inverting the tube 20 times. The precipitates containing leaf material and
336 proteins were removed by centrifugation at 8000 g for 12 min at 4°C. We transferred the
337 supernatants to new tubes and purified the DNA from solution as described below in
338 "Removal of small DNA fragments < 1.5 kb with optimized SPRI beads".

339 We further purified the samples using a chloroform:isoamylalcohol extraction. The eight
340 aqueous DNA solutions were pooled to a total of 400 µL to which one volume of
341 chloroform:isoamylalcohol was added, mixing by inversion for 5 minutes. The phases were
342 separated by centrifugation at 5000 g for 2 minutes at room temperature (RT). We
343 transferred the upper, DNA containing phase to a fresh tube performing another round of the
344 chloroform:isoamylalcohol purification. After the second extraction the DNA was precipitated
345 by adding 0.1 volume 3 M sodium acetate pH 5.3 and 1 volume 100% cold ethanol, followed
346 by centrifugation at 5000 g for 2 min at RT. The short centrifugation at low speed may
347 reduce DNA yields but potentially precipitates longer fragments in favor of shorter fragments.
348 The transparent pellet was washed with 70% ethanol and resuspended in 50 µL 10 mM Tris-
349 HCl pH 8.0 for 2 h at room temperature. The solubilised DNA was stored at 4°C until library
350 preparation, for a maximum of 10 days.

351 ***DNA size selection***

352 g-TUBE shearing

353 We processed 5 µg of pure HMW DNA through a g-TUBE (Covaris) in an Eppendorf 5418
354 centrifuge at 3800 rpm for a total of 2 minutes.

355 BluePippin size selection

356 We used the BluePippin system (Sage science) with 0.75% dye-free Agarose cassettes and
357 S1 marker, selecting for fragments > 20 kb using 6 µg sample for each lane following the
358 manufacturer's instructions.

359

360 **Removal of small DNA fragments < 1.5 kb with optimized SPRI beads**

361 In order to purify and remove small fragments from our DNA samples we optimized a SPRI
362 beads solution which we used for clean ups and library preparations. The improved beads
363 solution consists of 11% PEG 8000, 1.6 M NaCl, 10 mM Tris-HCl pH 9.0, 1 mM EDTA, 0.4%
364 Sera-Mag SpeedBeads (GE Healthcare PN 65152105050250) (Schalamun & Schwessinger,
365 2017). For the clean up procedure, 0.8 vol of this beads solution was mixed with the DNA
366 sample and incubated on a hula mixer for 10 min. After a brief (pulse) centrifugation step in a
367 microcentrifuge, we placed the tube in a magnetic stand so that the beads bound to the rear
368 of the tube, allowing for removal of the supernatant. We then washed the beads twice with 1
369 mL 70% ethanol, keeping the tube on the magnetic stand throughout the wash procedure to
370 avoid loss of DNA bound to the beads (the tube can be rotated 360° within the stand,
371 allowing comprehensive washing while ensuring bead retention). After the second wash we
372 centrifuged the tube briefly again to remove the last traces of ethanol. The beads were dried
373 for no longer than 30 s before elution of the DNA in 50 µL Tris-HCl pH 8.0 preheated to
374 50°C, for 10 min.

375 ***DNA Quality control***

376 DNA concentrations were determined using the Qubit dsDNA BR (Broad Range) assay kit
377 (ThermoFisher). The purity of the sample was measured with the NanoDrop, assessing
378 curve shape, the 260/280 nm and 260/230 nm values, and congruence of concentrations
379 with the Qubit values. The DNA was examine after 0.8 % agarose gel electrophoresis
380 containing 0.001% (v/v) SYBR Safe dye (ThermoFisher) in 1X TBE buffer (10.8 g/L Tris
381 base, 5.5 g/L boric acid, 0.75 g/L EDTA, pH 8.3) for 45 minutes at 100 V. For higher
382 resolution, pulsed field gel electrophoresis (PFGE) was used with a 1.5% agarose gel in
383 0.5X TBE running buffer, run for 17.7 hours at 6 V/cm and 1.4 s initial and 13.5 s final switch
384 time. The gel was stained after the electrophoresis with 5 µL SYBR Safe dye in
385 approximately 200 mL Milli-Q water.

386 ***Library preparation and sequencing***

387 We followed the 1D ligation protocol SQK-LSK108 selecting for long reads but instead of the
388 recommended AMPure XP beads we used our optimized SPRI beads solution (Schalamun,
389 2017). We started the library preparation with 4 µg HMW DNA and used 1 µg of the resultant
390 library DNA for sequencing on a R9.5 flowcell. The sequencing software MinKNOW version
391 1.7.3 was installed on a computer with minimum of 4 cores running a Linux operating system
392 (Ubuntu 14.4).

393

394 **Tables**

395 **Table 1. DNA purity impacts sequencing yields**

396 Comparison of yield per flowcell for different quality samples. Impact of sample quality
397 measured by 260/280 and 260/230 nm ratios (Nanodrop data) on the final sequence output

398 measured in Gb per flowcell (Figure 1). Sample #10 and #27 are two representative
399 sequencing runs. #5 is a run with low input DNA purity.

Sample	Qubit [ng/μl]	Nanodrop [ng/μl]	260/280	260/230	Yield [Gb]	Yield _{Q7} [Gb]
10	178	203	1.8	2.1	6.0	5.9
27	142	188	1.8	2.3	7.8	7.4
5	57	80	1.7	1.0	0.7	0.7

400

401

402 **Table 2. DNA integrity impacts sequencing read length**

403 **Read length comparison for samples sheared during the extraction process.**
404 Comparison of N50_{Q7}, mean read length_{Q7} and median read length_{Q7} between untreated
405 samples (#10 and #27) and the DNA sample sheared during DNA extraction as shown in
406 Figure 2 #3 (#9).

Sample	Size selection	N50 _{Q7} [kb]	Mean _{Q7} [kb]	Median _{Q7} [kb]	Yield [Gb]	Yield _{Q7} [Gb]
10	NO	25.8	12.4	6.2	6.0	5.9
27	NO	26	13.2	7.5	7.8	7.4
9	sheared during extraction	9.2	4.9	2.5	3.5	3.5

407

408

409

410 **Table 3. Targeted mechanical DNA shearing does not increase sequencing**
411 **throughput**

412 Read length comparisons for unsheared and sheared samples. Comparison of N50_{Q7}, mean
413 read length_{Q7} and median read length_{Q7} of untreated samples (#10 and #27) and sheared (g-
414 covaris tube) samples (#4 and #23) (Figure 3).

Sample	Size selection	N50 _{Q7} [kb]	Mean _{Q7} [kb]	Median _{Q7} [kb]	Yield [Gb]	Yield _{Q7} [Gb]
10	NO	25.8	12.4	6.2	6.0	5.9

27	NO	26	13.2	7.5	7.8	7.4
4	g-covaris	18.4	11.8	9.5	4.8	4.7
23	g-covaris	17.9	11.2	8.5	7.2	7.0

415

416

417 **Table 4. High-pass size selection increases read length statistics**

418 Read-length comparisons for BluePippin size-selected samples. Comparison of N50_{Q7}, mean
 419 read-length_{Q7} and median read-length_{Q7} of untreated samples (10) and (27) and Blue-Pippin
 420 size-selected samples (2) (Figure 3).

Sample	Size selection	N50 _{Q7} [kb]	Mean _{Q7} [kb]	Median _{Q7} [kb]	Yield [Gb]	Yield _{Q7} [Gb]
10	NO	25.8	12.4	6.2	6.0	5.9
27	NO	26	13.2	7.5	7.8	7.4
2	BluePippin	35.1	26.5	23.9	3.5	3.5

421

422 **Figures**

423 **Figure 1: Illustration of different purity DNA preparations.** Nanodrop readings of different
 424 DNA preparations. (A) DNA extraction with CTAB lysis buffer followed by
 425 phenol:chloroform:isoamylalcohol extraction (Schwessinger & Rathjen, 2017). (B) Sample A
 426 after SPRI beads clean-up. (C) DNA extraction using SDS lysis buffer and SPRI beads
 427 purification (Mayjonade et al., 2016). (D) Sample C followed by an additional
 428 chloroform:isoamylalcohol purification step .

429 **Figure 2. Illustration of the impact on DNA extraction procedures on DNA fragment
 430 length**

431 0.8 % agarose gel of 100 ng DNA prepared with two different DNA extraction procedures as
 432 explained in the main text. #1 HyperLadder 1 kb (Bioline). #2 DNA extracted following the
 433 default HMW DNA extraction protocol with mean read length of 13 kb as shown in Table 2.
 434 #3 DNA accidentally sheared during the extraction procedure with mean read length of 5 kb
 435 as shown in Table 2.

436 **Figure 3. Purposeful mechanical shearing and high-pass filtering alters DNA fragment
 437 length distribution**

438 Pulsed field gel electrophoresis of differently treated DNA samples. Lane #1 and #5
439 MidRange II PFG marker (BioLabs). Lane #2 DNA extracted following the default HMW DNA
440 extraction protocol (mean read length of 13 kb as shown in Table 4). Lane #3 same DNA
441 extraction as in #2 followed by size selection with the Blue Pippin using 20 kb high pass
442 filtering (a mean read length of 26 kb as shown in Table 4). Lane #4 same same DNA
443 extraction as in #2 followed by mechanical shearing with the a g-TUBE (a mean read length
444 of 11.8 kb as shown in Table 3).

445 **Figure 4. Optimized DNA input into the sequencing adapter ligation reaction.**

446 DNA input [μg] into the adapter-ligation reaction of the 1D library preparation (x-axis) versus
447 final sequence yields [Gb]. We achieved highest sequencing throughput by adding
448 approximately 2 μg of FFPE and end-repaired DNA into the adapter-ligation step. This
449 optimum was identified empirically and is likely related to the best ratio of free DNA-ends to
450 available DNA sequencing adapters in the ligation reaction. The red points mark outliers with
451 low yields due to a broken flowcell membrane and low yield due bad quality DNA input
452 (Table 1). These points were excluded from the calculation of the smoothed line.

453

454 **Figure 5. Improved DNA size selection using an adapted PEG-NaCl-SPRI beads**
455 **protocol**

456 Each lane represents 80 ng DNA before size selection. Lanes C contain the HyperLadder 1
457 kb (BioLine) as untreated control. Lane A is DNA ladder size selected with 0.45 vol (V/V)
458 Agencourt AMPure XP beads. Lane 1.0, 0.9, and 0.8 are DNA ladder size selected with the
459 adapted PEG-NaCl-SPRI beads solution.

460 **Figure 6. The impact of DNA extraction protocol on the distribution of read lengths**
461 **from ONT sequencing.** Each line represents the read length distribution for a single
462 flowcell. The x-axis shows the read lengths on a log scale, and the y-axis shows the density
463 of reads at a particular length. The top panel shows data for all reads, and the bottom panel
464 shows the same data, but with reads that have a mean quality (Q) score less than 7
465 removed.

466 **Figure 7. The impact of DNA extraction protocol on the yield of ONT sequencing.** Each
467 line represents a single flowcell. The y axis shows the yield of in bases, and the x axis shows
468 the minimum readlength at which the yield was calculated. For example, the yield of reads
469 longer than 20KB from each flowcell can be compared by comparing the height of the lines
470 at the 20KB point on the x axis. The plot shows that while using the Blue Pippin improved
471 the distribution of read lengths (insofar as the red line is relatively flat below 20KB, showing
472 that only a small proportion of the sequenced bases were in reads shorter than 20KB), we
473 were able to obtain higher yields of reads >20KB from the libraries that were prepared
474 without the Blue Pippin (blue and pink lines, labelled 'no shearing'). These flowcells also
475 contained a considerable yield of reads between 1KB and 20KB, which may be useful for
476 many applications.

477

478 **Figure 8. Real time analysis of sequencing runs via the MinKNOW graphical user**
479 **interface**

480 Both panels (A and B) show the MinKNOW interface two hours into a run. Panel A illustrates
481 an unsatisfactory sequencing run where read length is short, pore occupancy poor (~40%)

482 and many pores are not available for sequencing any more (see main text for details). This
483 run was aborted after two hours to not to waste this flow cell and to reload an improved
484 library. Panel B illustrates a satisfactory sequencing run with excellent read length
485 distribution, good pore occupancy (~80%), and most pores still readily available for
486 sequencing.

487 **Figure 9. MinION Nanopore sequencing workflow to optimize sequencing output**

488

489 **Supplemental material**

490 **Supplemental Table 1. Raw data of all quality control variables tracked during DNA**
491 **extraction, library preparation, and sequencing run summary statistics described in**
492 **this study.**

493 **Supplemental Table 2. Template spreadsheet to use for tracking quality control**
494 **variables for DNA extraction, library preparation, and sequencing run summary**
495 **statistics.**

496 **Supplemental Figure 1: Sequencing yield depends on quality control statistics**

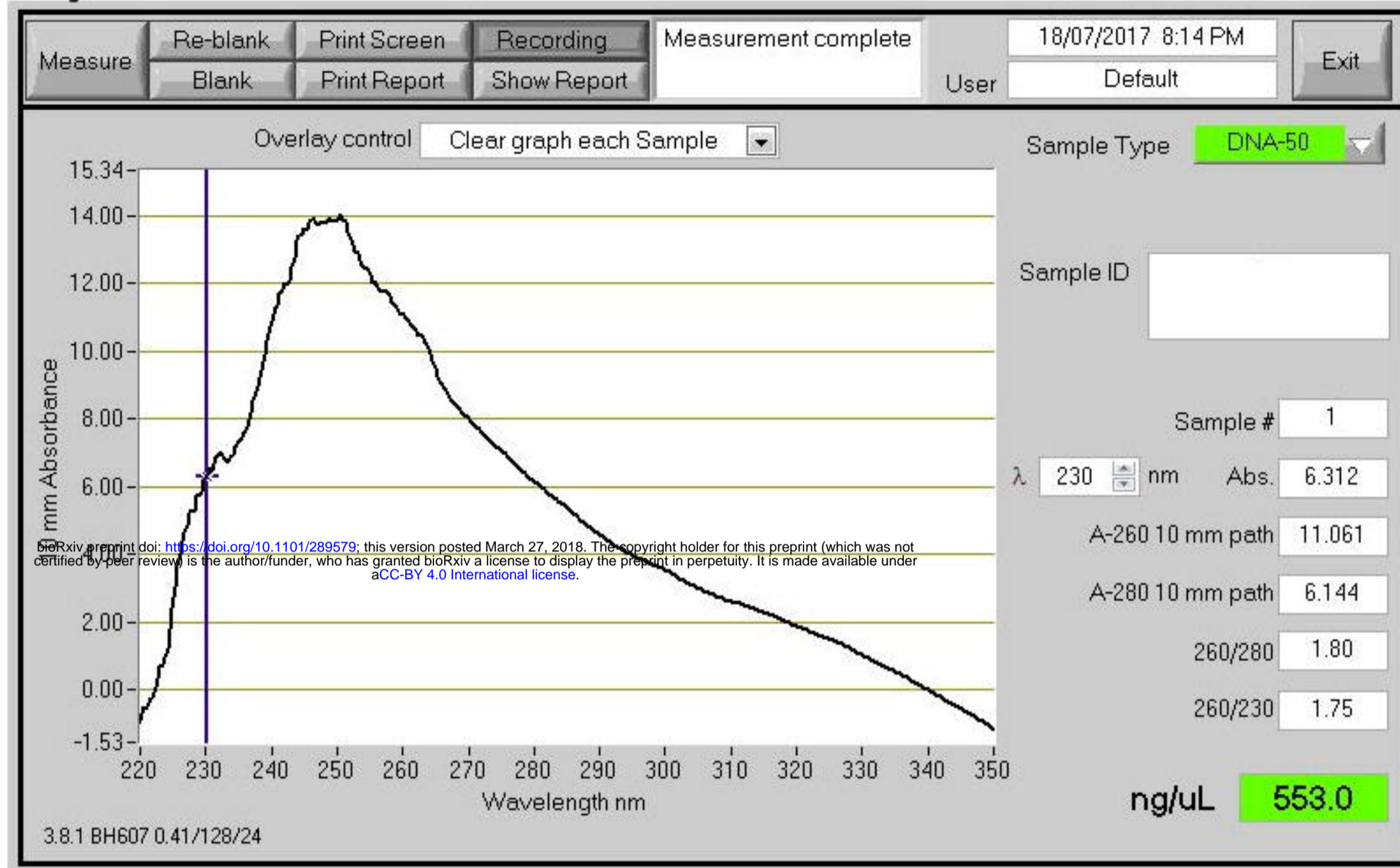
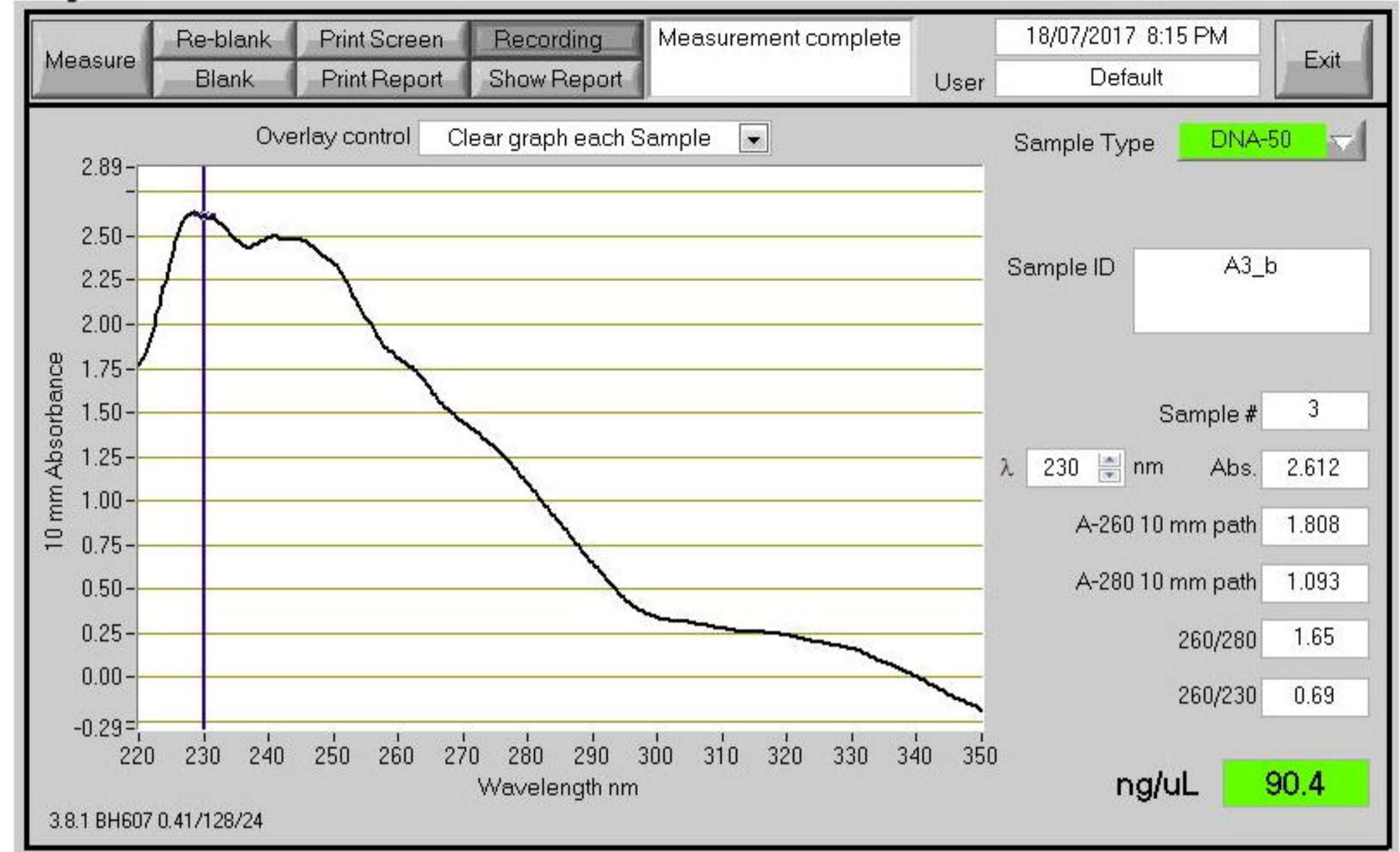
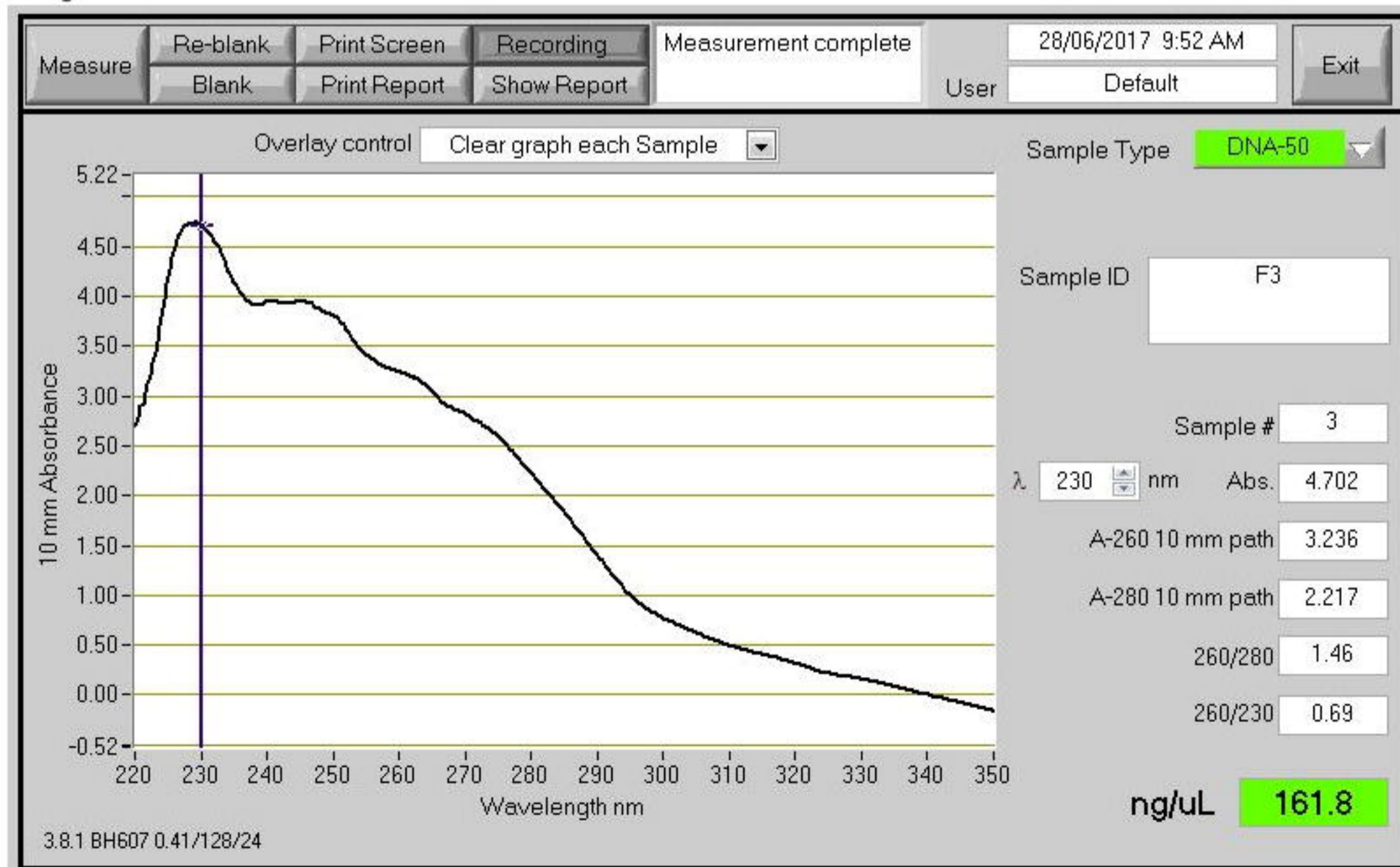
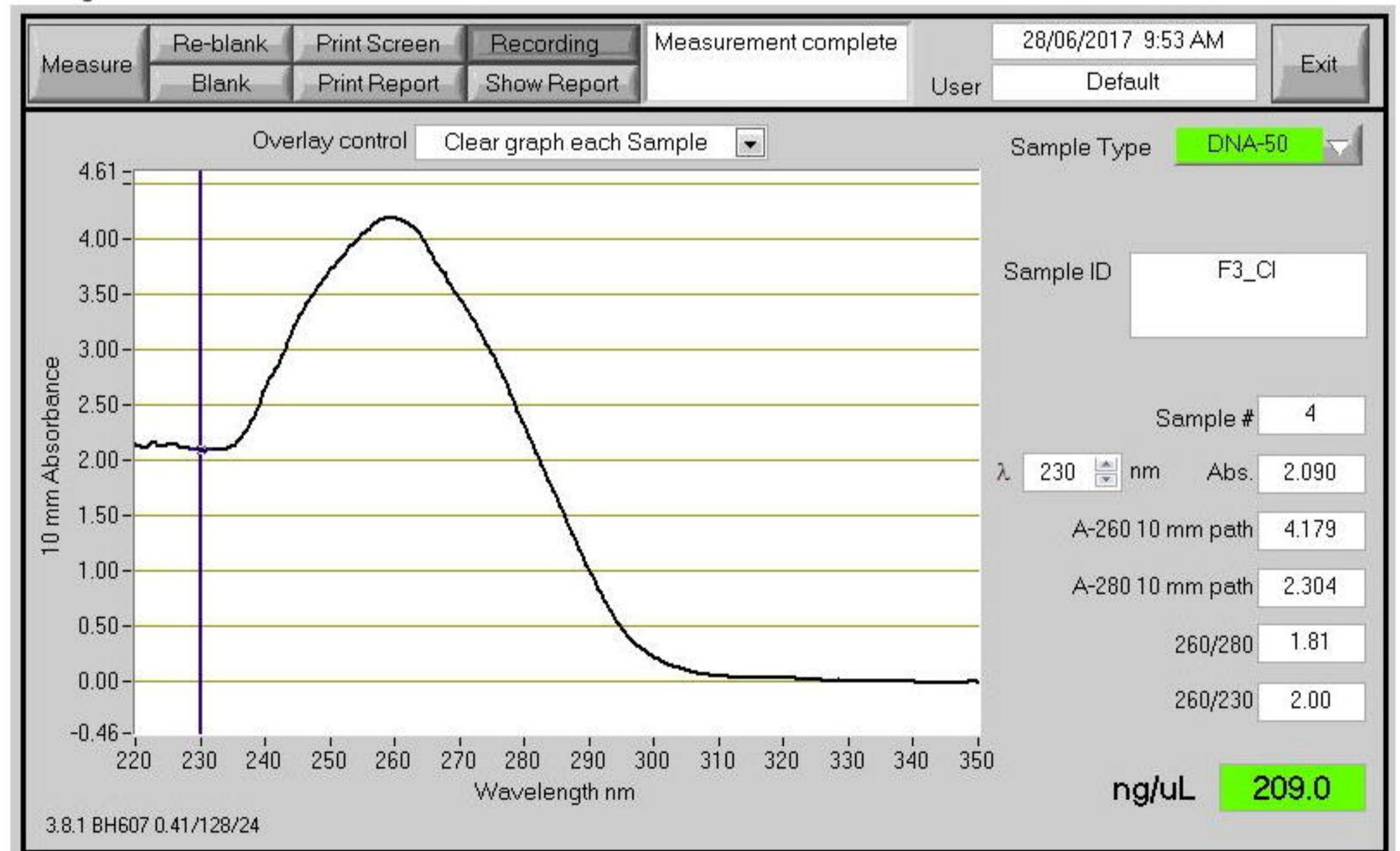
497 Comparison of sequencing yield (as GB of called data) versus all measured QC statistics.
498 Positive correlations that are not directly related to yield are seen with sample number, the
499 number of channels seen during flow cell QC, and the measured initial input library
500 concentration. Lines indicate fitted quantile regression lines at 25%, 50% and 75% (i.e. lower
501 quartile, median, and upper quartile respectively). The script for plotting the quality control
502 statistics vs sequencing yield is provided on github ([https://github.com/gringer/minion-user-](https://github.com/gringer/minion-user-group)
503 [group](https://github.com/gringer/minion-user-group)).

504

505 **Bibliography**

- 506 Desjardins, P., & Conklin, D. (2010). NanoDrop Microvolume Quantitation of Nucleic Acids.
507 *Journal of Visualized Experiments*, (1). doi:10.3791/2565
- 508 Dong, C. (2017). Purification of HMW DNA from Fungi for long read sequencing.
509 *Protocols.io*. doi:dx.doi.org/10.17504/protocols.io.hbvb2n6
- 510 He, Z., Zhu, Y., & Gu, H. (2013). A new method for the determination of critical polyethylene
511 glycol concentration for selective precipitation of DNA fragments. *Applied*
512 *Microbiology and Biotechnology*, 97(20), 9175–9183. doi:10.1007/s00253-013-5195-
513 0
- 514 Healey, A., Furtado, A., Cooper, T., & Henry, R. J. (2014). Protocol: a simple method for
515 extracting next-generation sequencing quality genomic DNA from recalcitrant plant
516 species. *Plant Methods*, 10(1), 21.
- 517 Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., ... Loose, M. (2018).
518 Nanopore sequencing and assembly of a human genome with ultra-long reads.
519 *Nature Biotechnology*. doi:10.1038/nbt.4060
- 520 Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION:
521 delivery of nanopore sequencing to the genomics community. *Genome Biology*,
522 17(1). doi:10.1186/s13059-016-1103-0
- 523 Jansen, H. J., Liem, M., Jong-Raadsen, S. A., Dufour, S., Weltzien, F.-A., Swinkels, W., ...
524 Henkel, C. V. (2017). Rapid de novo assembly of the European eel genome from

- 525 nanopore sequencing reads. *Scientific Reports*, 7(1). doi:10.1038/s41598-017-
526 07650-6
- 527 Klingstrom, T., Bongcam-Rudloff, E., & Pettersson, O. V. (2018). A comprehensive model of
528 DNA fragmentation for the preservation of High Molecular Weight DNA. *BioRxiv*,
529 254276.
- 530 Krinitsina, A. A., Sizova, T. V., Zaika, M. A., Speranskaya, A. S., & Sukhorukov, A. P. (2015).
531 A rapid and cost-effective method for DNA extraction from archival herbarium
532 specimens. *Biochemistry (Moscow)*, 80(11), 1478–1484.
533 doi:10.1134/S0006297915110097
- 534 Leggett, R. M., & Clark, M. D. (2017). A world of opportunities with nanopore sequencing.
535 *Journal of Experimental Botany*, 68(20), 5419–5429. doi:10.1093/jxb/erx289
- 536 Lis, J. T., & Schleif, R. (1975). Size fractionation of double-stranded DNA by precipitation
537 with polyethylene glycol. *Nucleic Acids Research*, 2(3), 383–389.
- 538 Liu, C., Xiao, F., Hoisington-Lopez, J., Lang, K., Quenzel, P., Duffy, B., & Mitra, R. D. (2017).
539 Accurate typing of class I human leukocyte antigen by Oxford nanopore sequencing.
540 *BioRxiv*, 178590.
- 541 Mackey, K., & Chomczynski, P. (1997). Effect of pH and ionic strength on the
542 spectrophotometric assessment of nucleic acid purity. *Biotechniques*, 22(3), 474–
543 481.
- 544 Mayjonade, B. (2018, March). *BEST PRACTICE TO MAXIMIZE THROUGHPUT WITH*
545 *NANOPORE TECHNOLOGY & DE NOVO...* Technology. Retrieved from
546 [https://www.slideshare.net/BaptisteMayjonade/best-practice-to-maximize-throughput-](https://www.slideshare.net/BaptisteMayjonade/best-practice-to-maximize-throughput-with-nanopore-technology-de-novo-sequencing-of-genetic-lines-of-arabidopsis-thaliana)
547 [with-nanopore-technology-de-novo-sequencing-of-genetic-lines-of-arabidopsis-](https://www.slideshare.net/BaptisteMayjonade/best-practice-to-maximize-throughput-with-nanopore-technology-de-novo-sequencing-of-genetic-lines-of-arabidopsis-thaliana)
548 [thaliana](https://www.slideshare.net/BaptisteMayjonade/best-practice-to-maximize-throughput-with-nanopore-technology-de-novo-sequencing-of-genetic-lines-of-arabidopsis-thaliana)
- 549 Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., ... Muños, S.
550 (2016). Extraction of high-molecular-weight genomic DNA for long-read sequencing
551 of single molecules. *BioTechniques*, 61(4). doi:10.2144/000114460
- 552 Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., ... Ecker, J. R.
553 (2018). High contiguity Arabidopsis thaliana genome assembly with a single
554 nanopore flow cell. *Nature Communications*, 9(1). doi:10.1038/s41467-018-03016-2
- 555 O'Neill, M., McPartlin, J., Arthure, K., Riedel, S., & McMillan, N. (2011). Comparison of the
556 TLDA with the Nanodrop and the reference Qubit system. *Journal of Physics:*
557 *Conference Series*, 307, 012047. doi:10.1088/1742-6596/307/1/012047
- 558 Ramos, J. É. B., de Vries, R., & Ruggiero Neto, J. (2005). DNA Ψ -Condensation and
559 Reentrant Decondensation: Effect of the PEG Degree of Polymerization. *The Journal*
560 *of Physical Chemistry B*, 109(49), 23661–23665. doi:10.1021/jp0527103
- 561 Schalamun, M., & Schwessinger, B. (2017). DNA size selection (>1kb) and clean up using
562 an optimized SPRI beads mixture. doi:dx.doi.org/10.17504/protocols.io.idmca46
- 563 Schmidt, M. H.-W., Vogel, A., Denton, A. K., Istace, B., Wormit, A., van de Geest, H., ...
564 Usadel, B. (2017). De Novo Assembly of a New *Solanum pennellii* Accession Using
565 Nanopore Sequencing. *The Plant Cell*, 29(10), 2336–2348. doi:10.1105/tpc.17.00521
- 566 Schmitz, A., & Riesner, D. (2006). Purification of nucleic acids by selective precipitation with
567 polyethylene glycol 6000. *Analytical Biochemistry*, 354(2), 311–313.
568 doi:10.1016/j.ab.2006.03.014
- 569 Schwessinger, B. (2018). <p>DNA quality and nanopore sequencing</p>. *F1000Research*,
570 7. doi:10.7490/f1000research.1115331.1
- 571 Schwessinger, B., & Rathjen, J. P. (2017). Extraction of High Molecular Weight DNA from
572 Fungal Rust Spores for Long Read Sequencing. In *Wheat Rust Diseases* (pp. 49–
573 57). Humana Press, New York, NY. doi:10.1007/978-1-4939-7249-4_5
- 574 Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017).
575 Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*,
576 14(4), 407–410. doi:10.1038/nmeth.4184
- 577

A)**B)****C)****D)**

[kb]

1

2

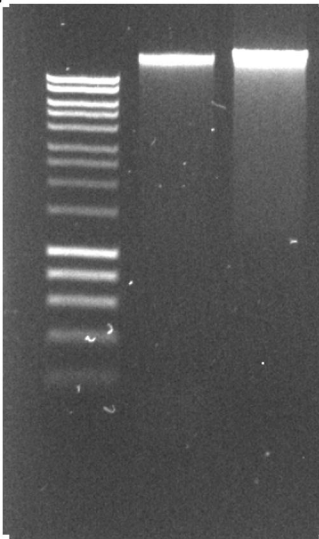
3

10

1.5

1.0

0.4



[kb] 1 2 3 4 5

270

194

145

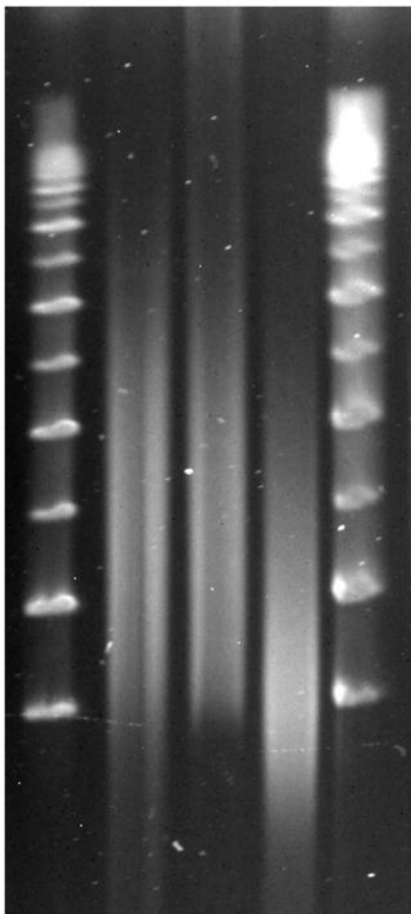
121

97

73

48

24



GB data [Gb]

6

4

2

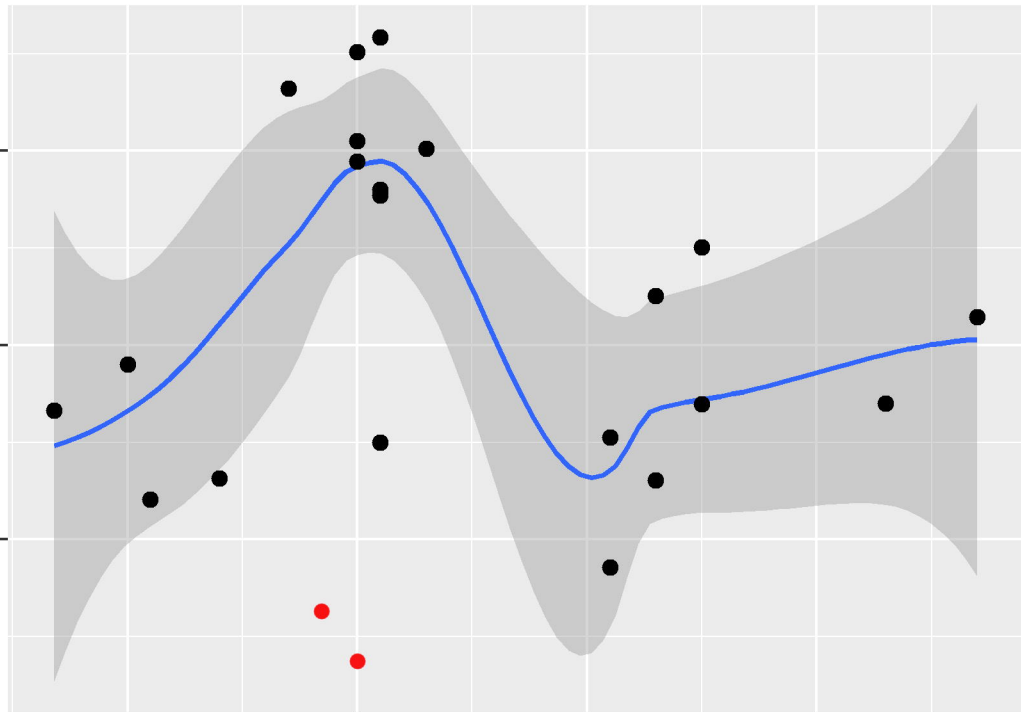
1

2

3

4

DNA input into ligation reaction [μg]



[kb]

C

A

1.0

0.9

0.8

C

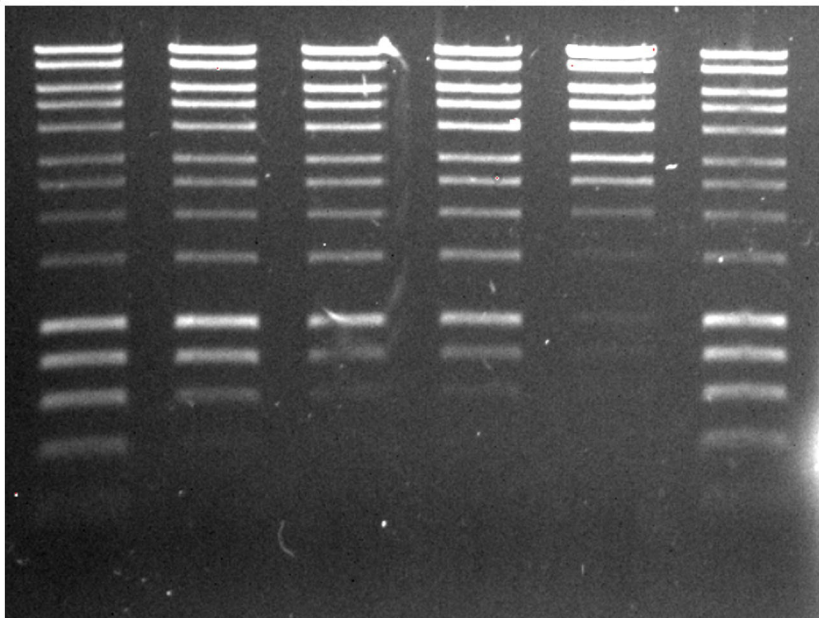
10

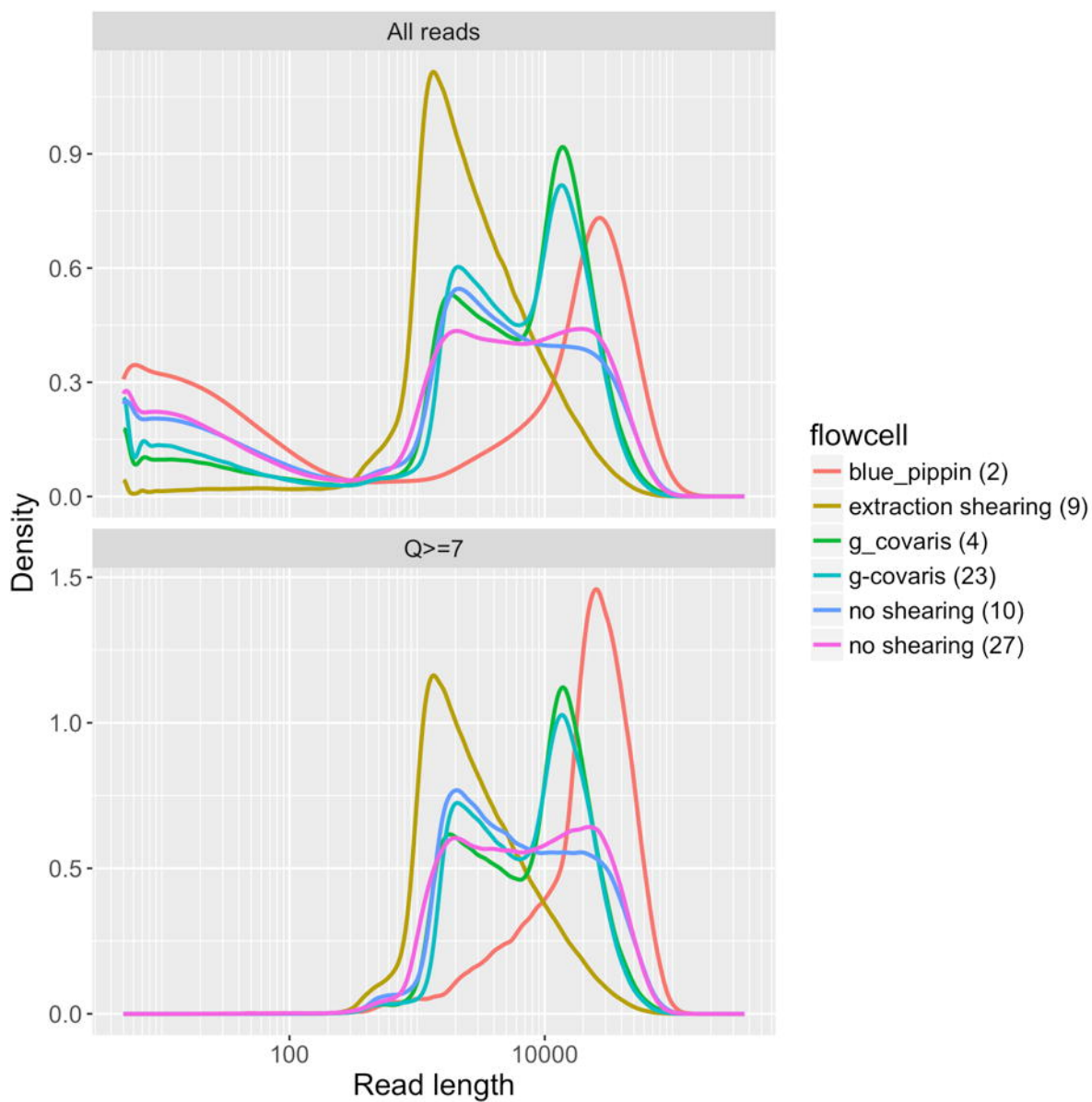
2.0

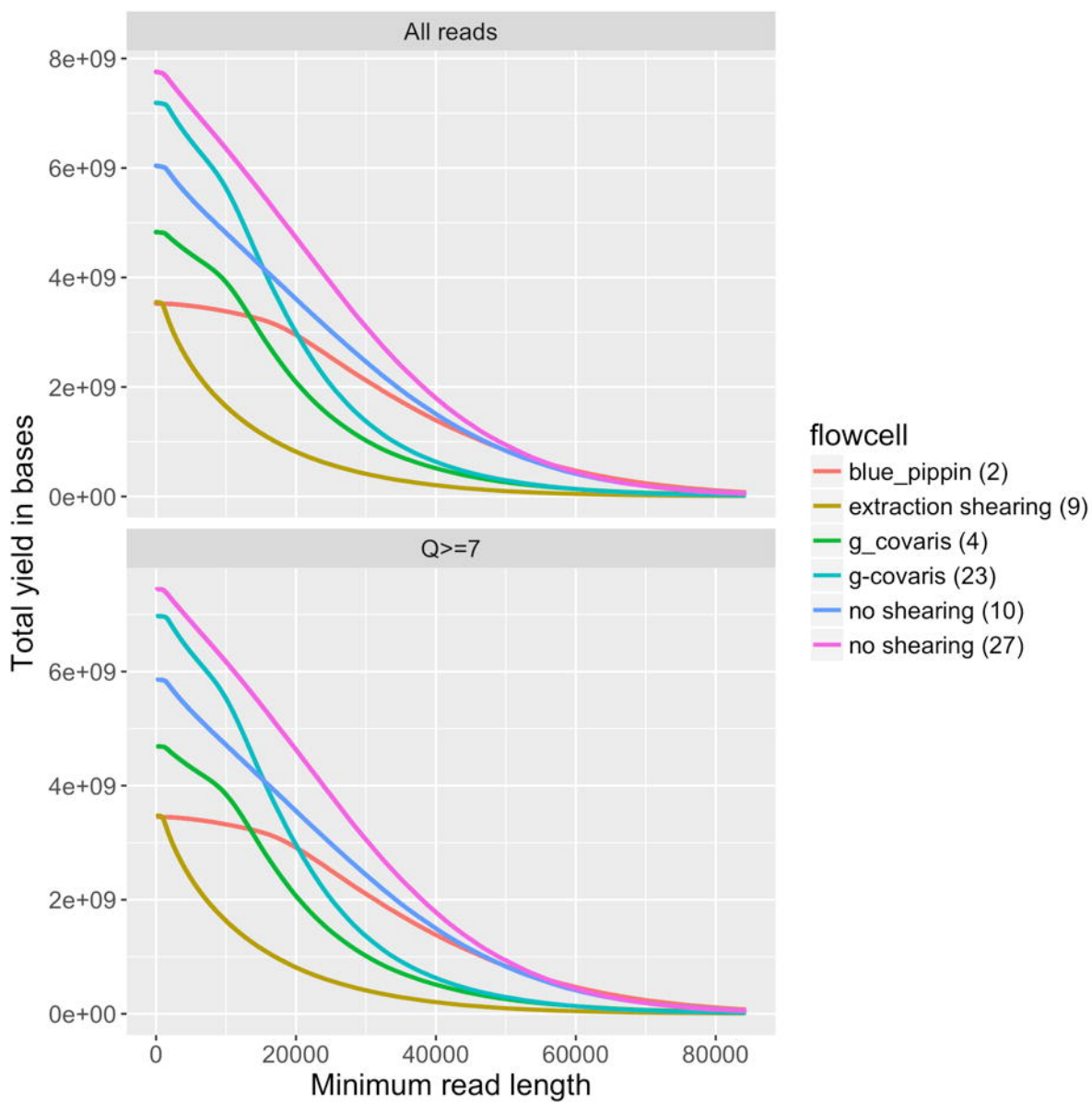
1.5

1.0

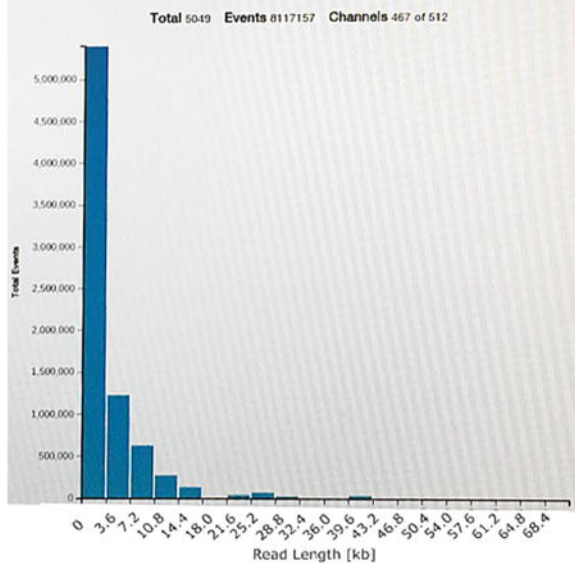
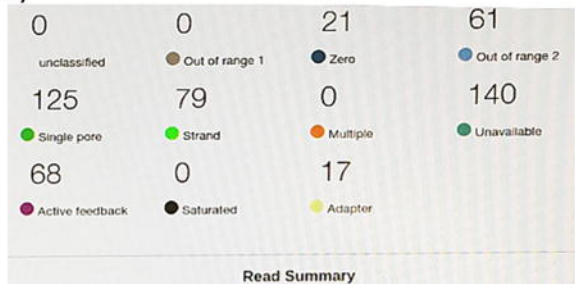
0.4



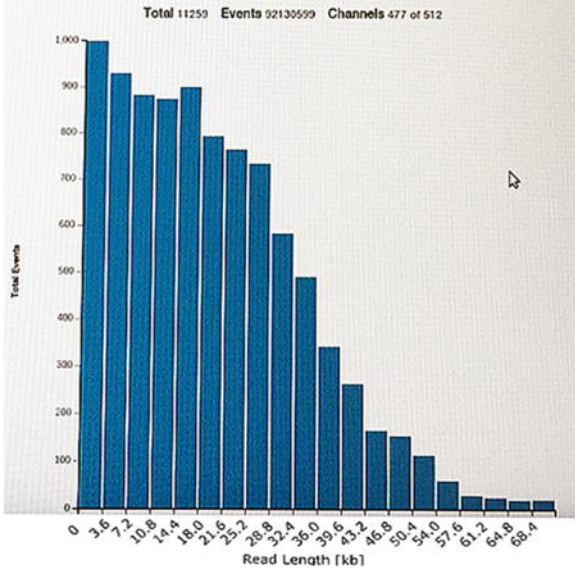
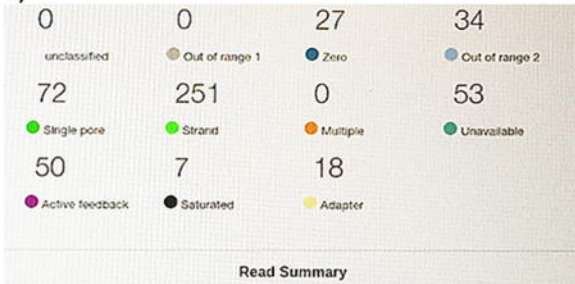




A)



B)



Work flow from tissue to sequence for 1D nanopore sequencing

