SHIMS 3.0: Highly efficient single-haplotype iterative mapping and sequencing using ultra-long nanopore reads

Daniel W. Bellott[1&4], Ting-Jan Cho[1&4], Emily K. Jackson[2], Helen Skaletsky[1&3], Jennifer F. Hughes[1], David C. Page[1-3]

1. Whitehead Institute, Cambridge, Massachusetts, USA
2. Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
3. Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts, USA
4. These authors contributed equally to this work.

Correspondence should be addressed to D. W. B. (bellott@wi.mit.edu)

Keywords: genome assembly, repeat sequences, nanopore sequencing, high-throughput nucleotide sequencing, sequence finishing, bacterial artificial chromosome

Abstract

The reference sequence of structurally complex regions can only be obtained through a highly accurate clone-based approach that we call Single-Haplotype Iterative Mapping and Sequencing (SHIMS). In recent years, improvements to SHIMS have reduced the cost and time required by two orders of magnitude, but internally repetitive clones still require extensive manual effort to transform draft assemblies into reference-quality finished sequences. Here we describe SHIMS 3.0, an extension of our SHIMS sequencing strategy, using ultra-long nanopore reads to augment the Illumina data from SHIMS 2.0 assemblies and resolve internally repetitive structures. This greatly minimizes the need for manual finishing of Illumina-based draft assemblies, allowing a small team with no prior finishing experience to sequence challenging targets with high accuracy. This protocol proceeds from clone-picking to finished assemblies in 2 weeks for about 80 dollars per clone. We recently used this protocol to produce reference sequence of structurally complex palindromes on chimpanzee and rhesus macaque X chromosomes; as a further demonstration of the capabilities of SHIMS 3.0, we finish the *TSPY* array on the human Y chromosome, which could not be resolved by previous sequencing methods. Our protocol provides access to structurally complex regions that would otherwise be inaccessible from whole-genome shotgun data or require an impractical amount of manual effort to generate an accurate assembly.

40 # Introduction

41 ## Background and applications

42 Reference genome sequence quality is of central importance to modern biological
43 research. Experiments based on aligning cheap and abundant short reads to existing
44 reference sequences have become commonplace, permitting studies of variation by
45 genome and exome resequencing, transcription by RNA sequencing, and epigenetic
46 modifications by chromatin immunoprecipitation–sequencing. However, these
47 experiments are limited by the quality and completeness of the underlying
48 reference sequence, so that new insights may emerge from reanalyzing short-read
49 datasets in the light of an improved reference sequence. The foremost obstacles to
50 accurate reference genome assembly are repeated sequences within the genome.
51 The most structurally complex repeats are ampliconic sequences – euchromatic
52 repeats with greater than 99% identity over more than 10 kb[1]. The complex
53 repetitive structures in amplicons mediate deletions, duplications, and inversions
54 associated with human disease[2]. Amplicons pose special challenges for genome
55 assembly, requiring extremely long and accurate reads to discriminate between
56 amplicon copies and produce a correct reference sequence.

57 We developed our Single Haplotype Iterative Mapping and Sequencing (SHIMS)
58 approach to cope with the ampliconic sequences of the human Y chromosome[3].
59 Because paralogous ampliconic repeats are more similar than alleles, we sequenced
60 large-insert clones from a single haplotype, allowing us to confidently identify the

61 rare sequence family variants (SFVs) that distinguish paralogous repeats in highly
62 accurate (less than 1 error per megabase) synthetic long reads[3]. Mapping and
63 sequencing were coupled; newly sequenced clones provide novel SFVs that refine
64 the clone map and serve as markers to select new clones. SHIMS has been
65 instrumental to producing reference sequences of structurally complex sex
66 chromosomes from several species[4-10], as well as the human immunoglobulin gene
67 cluster[11], and other structurally complex regions on human autosomes[12]. SHIMS
68 remains the only sequencing approach that can reliably disentangle ampliconic
69 repeats. Whole genome shotgun (WGS) strategies are constrained by a tradeoff
70 between read length and accuracy among existing sequencing technologies. Sanger
71 or Illumina reads are accurate, but are not long enough to traverse interspersed
72 repeats, much less ampliconic sequence[13]. Single-molecule sequencing technologies
73 like PacBio or nanopore sequencing offer reads long enough to span interspersed
74 repeats and smaller ampliconic sequences, but lack the accuracy to disentangle
75 nearly identical ampliconic repeats[14]. As originally implemented, SHIMS 1.0
76 required the resources of a fully-staffed genome center to generate Sanger reads,
77 assemble draft sequences, and manually 'finish' each clone. We developed SHIMS 2.0
78 to combine the advantages of a hierarchical clone-based strategy with high-
79 throughput sequencing technologies, allowing a small team to generate sequence,
80 while reducing time and cost by two orders of magnitude, while maintaining high
81 accuracy[15]. However, SHIMS 2.0 still required intensive manual review to resolve
82 internally repetitive clones, and in some cases – particularly short, nearly perfect,
83 tandem repeats – complete resolution remained impossible.

84 Here we describe SHIMS 3.0, an extension of our SHIMS sequencing strategy that we
85 recently employed to produce reference sequence of structurally complex regions of
86 chimpanzee and rhesus macaque X chromosomes[10]. SHIMS 3.0 uses a combination
87 of nanopore and Illumina sequencing technologies to resolve repetitive structures
88 within individual large-insert clones. We describe a protocol for generating full-
89 length nanopore reads for pools of clones, and combining the structural information
90 from these full-length reads with highly accurate short-read data to automatically
91 produce assemblies of internally repetitive clones (**Fig. 1**). This protocol proceeds
92 from clone-picking to finished assemblies in 2 weeks for about 80 dollars per clone,
93 an improvement of 2 orders of magnitude compared with 24 months and 4000
94 dollars under SHIMS 1.0. As further demonstration of the performance of SHIMS 3.0,
95 we apply this protocol to resolve the *TSPY* array on the human Y chromosome. The
96 *TSPY* array is one of the largest and most homogeneous protein-coding tandem
97 arrays in the human genome[16], and it could not be completely resolved in the SHIMS
98 1.0 reference sequence of the human Y chromosome[4].

99 Methodology

100 Large-insert clone libraries derived from a single haplotype are essential to the
101 SHIMS strategy, and are discussed in detail in our description of SHIMS 2.0[15]. In
102 brief, any library derived from an individual of the heterogametic sex will provide a
103 single haplotype source for sequencing sex chromosomes, albeit at half the coverage
104 of the autosomes. Libraries created from inbred strains can provide a single-
105 haplotype source for autosomes. When inbreeding is not possible, special measures

106 may be necessary to obtain a single-haplotype source of DNA[12]. The ideal library
107 will have greater than 10x coverage of the chromosome of interest to minimize the
108 number of gaps in library coverage. In SHIMS 1.0 and 2.0 it was important to match
109 the average library insert size to the expected amplicon unit size, such that it was
110 rare for two units to be present in the same clone[15]. SHIMS 3.0 uses ultra-long
111 nanopore reads to span entire BAC clones[17]; therefore, we now recommend striving
112 for the largest possible insert size, to minimize library construction, screening, and
113 sequencing costs, as fewer clones will be necessary to achieve the required level of
114 coverage.

115 All SHIMS strategies begin by selecting an initial tiling path of large-insert clones for
116 sequencing and iterative refinement. Depending on the resources available for each
117 library, it may be possible to identify clones of interest electronically, using
118 fingerprint maps or end sequences; screening high-density filters by hybridization
119 with labeled oligos, or high-dimensional pools for sequence-tagged-site content by
120 PCR. It is most cost-effective to confirm the identity of each clone by generating
121 draft sequence with the SHIMS 2.0 protocol, rather than designing specific assays for
122 each clone[15]. In brief, this highly parallel method involves shearing BAC DNA to
123 generate large (~1 kb) fragments for individually indexed Illumina TruSeq–
124 compatible libraries to sequence and assemble pools of 192 clones in a single
125 week.[15] In our experience, the structure of ampliconic regions is often unclear until
126 a nearly complete tiling path is assembled, as the sequence map gradually unfolds as
127 new variants are identified by sequencing. It is therefore preferable to seed the first

128 iteration with as many clones as possible to identify sequence family variants early,
129 and minimize the total number of iterations.

130 Draft clone assemblies generated from Sanger or Illumina reads are accurate
131 enough to identify sequence family variants, and identify a minimum tiling path of
132 clones. In previous iterations of SHIMS, each clone in this path would be
133 painstakingly 'finished' to produce as correct and contiguous a sequence as possible.
134 Highly skilled technicians would inspect draft assemblies for errors and anomalies,
135 order and orient all draft sequence contigs, close all gaps, and resolve or annotate all
136 sequence ambiguities (e.g. SSRs). SHIMS 3.0 departs from this approach, instead
137 relying on the ability of nanopore-based sequencing technology to generate full-
138 length reads to scaffold short-read assemblies and eliminate the need for laborious
139 and time-consuming experiments, such as subcloning, PCR reactions, restriction
140 digests, and transposon bombing, that were used to correct draft assemblies in the
141 past.

142 We have typically employed SHIMS 3.0 on regions where we have prior knowledge
143 of ampliconic structures that would render sequence assembly from short reads
144 either very difficult (palindromes on primate X chromosomes) or impossible (the
145 *TSPY* array on the human X chromosome). When sequencing a novel ampliconic
146 target, it is often the case that the underlying genomic structures are known to be
147 too complex for whole genome shotgun sequencing, but the exact repetitive
148 structure is not known. In this case, we recommend sequencing a redundant tiling
149 path of clones by SHIMS 2.0 first, and to deploy SHIMS 3.0 on the subset of clones

150 that fail to assemble automatically from Illumina reads alone. This strategy provides
151 the least expensive means for assembling the full range of ampliconic structures
152 that we have encountered in vertebrate genomes, from tandem arrays of repeats
153 tens of kilobases long to multi-megabase palindromes.

154 We adapted existing methods for generating ultra-long reads[17] for use with pools of
155 large-insert clones on the Oxford Nanopore Technologies (ONT) MinION platform.
156 Successful generation of full-length reads requires intact DNA of high concentration
157 and purity. We optimized our protocol to avoid unnecessary manipulations that
158 could damage DNA or introduce contamination. We culture 24 clones separately,
159 and then pool all cultures for DNA isolation, library preparation, and sequencing. In
160 contrast to conventional plasmids, BACs and fosmids are present in only a single
161 copy per host cell, and common reagents for increasing the efficiency of DNA
162 precipitation, such as glycogen or SPRI beads, are incompatible with nanopore
163 sequencing. We compensate for this by starting with large volume (~15 ml) BAC
164 and fosmid cultures to ensure that we harvest a sufficient amount of intact DNA. To
165 preserve the integrity of high-molecular-weight (HMW) DNA, we handle it as little
166 as possible, pipetting very slowly, using only wide-bore tips. We allow precipitated
167 DNA to resuspend in water slowly over several days, rather than mixing by
168 vortexing, or pipetting up-and-down. We have had the best results generating
169 libraries from 7.5 to 15 μg of HMW DNA in 15 μl using the transposase-based RAD-
170 004 library preparation kit from ONT. At these concentrations, solutions of HMW
171 DNA will be extremely viscous, and it is difficult to measure the concentration
172 precisely; some trial-and-error may be required to get the correct ratio of

173 transposase to DNA, but we find that 0.5 µl of FRA for 15 µg is generally a good
174 starting point to ensure that most BAC clones are cut only once. It is important to
175 wait 45 minutes between loading the nanopore flow cell and starting the run, to
176 allow time for full-length molecules to diffuse to the pores, otherwise the run will be
177 dominated by shorter molecules. In contrast to our approach for generating
178 Illumina reads, indexing or barcoding individual clones is not necessary, as full-
179 length nanopore reads can be uniquely assigned to clones, even within the same
180 amplicon.

181 Full-length nanopore reads transform clone finishing into a purely computational
182 exercise. The tool chain for handling ultra-long nanopore reads is not yet fully
183 mature, but it is developing rapidly. We rely on Minimap2 for alignments involving
184 full-length reads[18]. This includes assigning nanopore reads to clones based on
185 SHIMS 2.0 draft sequences, identifying full-length reads that start and end in vector
186 sequence, and aligning a mix of long and short reads to generate a consensus. We
187 use Racon for polishing the consensus sequence[19], and SAMtools and custom scripts
188 to manipulate read alignments[20]. We use Gap5 and Consed for visualizing discrepant
189 bases and manually editing the consensus[21,22]. While full-length reads guarantee the
190 correct overall sequence structure, a variety of alignment artifacts may occur in
191 clones with highly identical internal repeats. In this case, we find it is best to
192 electronically split the clone sequence into individual repeat units, and correct each
193 unit separately, before merging them together to create a finished clone sequence.

194 ## Performance

195 The *TSPY* array on the Y chromosome is the largest and most homogeneous protein-
196 coding tandem array in the human genome[16], consisting of a 20.4-kb unit present in
197 a highly variable number of copies, ranging from 11 to 72 per individual[23]. *TSPY*
198 encodes a testis-specific protein, implicated in gonadoblastoma[24], that regulates cell
199 proliferation[25]; *TSPY* copy number is positively correlated with sperm count and
200 sperm concentration[26]. As a demonstration of the expected performance of SHIMS
201 3.0, we fully resolved this array for the first time, using clones from the RP11 BAC
202 library that we previously employed for our SHIMS 1.0 assembly of the male-specific
203 region of the human Y chromosome[4] (**Fig. 2**). This array spans 600 kb and contains
204 29 repeat units (**Fig. 2a & b**). We sequenced a redundant path of 19 clones and
205 identified 94 sequence family variants that allowed us to select a non-redundant
206 tiling path of 9 clones for finishing (**Fig. 2c**). On average, each unit differs from the
207 others by 1 in 100 bases. The array encodes 14 distinct *TSPY* transcripts, encoding
208 10 different proteins, and it includes one pseudogene (**Fig. 2d**). We observed that at
209 least 4 transcript variants were expressed in published testis RNA-seq datasets from
210 other males[27], indicating that multiple copies are expressed.

211 Using only Sanger or Illumina reads, the presence of multiple amplicon copies
212 within a single insert causes the clone assembly to collapse. The median BAC clone
213 in the *TSPY* array contains 9 repeat units, making it impossible to accurately
214 assemble even a single clone from this region using SHIMS 1.0 without months of
215 manual finishing efforts at a genome center. By incorporating full-length nanopore

216 reads, SHIMS 3.0 permits a small team – a technician and bioinformatics specialist –
217 to finish these challenging sequencing targets. Our SHIMS 3.0 protocol can improve
218 24 SHIMS 2.0 draft sequences to finished quality in 2 weeks at a cost of $80 per
219 clone. Despite the enormous reduction in staffing, cost, and time, the quality of
220 finished sequence remains extremely high. We observe less than 1 error per
221 megabase in overlaps between clones, on par with previous versions of SHIMS.

222 Comparison with other methods

223 SHIMS produces de novo sequence assemblies with higher accuracy than any other
224 technique, making it possible to produce accurate reference sequence of the most
225 extreme repetitive regions, from ampliconic sequences like the nearly-perfect multi-
226 megabase duplications on the mouse Y chromosome[9], to the thousands of
227 centromeric satellite repeats that form the centromere of the human Y
228 chromosome[28]. This extremely high accuracy is due to the clone-based nature of
229 SHIMS. Each clone represents a single long molecule that can be sequenced
230 repeatedly, with comple mentary technologies, to generate an assembly that is
231 accurate at the level of overall structure as well as the identity of individual
232 nucleotides. This property makes it possible to identify and repeatedly verify the
233 rare sequence family variants that distinguish ampliconic repeats, and build a high-
234 confidence map from individual clones.

235 The impressive advances in single-molecule sequencing technologies that enabled
236 SHIMS 3.0 have also increased the capabilities of whole genome shotgun
237 approaches[29]. It is now routine to generate nanopore sequencing runs where half of

238 the bases are in reads longer than 100 kb, so that interspersed repeats and smaller
239 ampliconic structures can be spanned by a single long read. Whole genome shotgun
240 with nanopore reads enabled the complete assembly of the human X chromosome
241 from a single haplotype source, the CHM13hTERT cell line[30]. This effort required
242 deep coverage from nanopore reads as well as from a broad array of complementary
243 sequencing and mapping technologies, combined with manual review of structurally
244 complex regions[30]. Error rates were still orders of magnitude higher than clone
245 based strategies – 1 error in 10 kb in single-copy sequence, and 7 errors per
246 kilobase in sequences present in more than one copy[30]. This elevated error rate in
247 multi-copy sequence is due to the inherent difficulties of uniquely mapping short
248 reads to long paralogous repeats; instead of reconstructing the true sequence, error
249 correction with short reads blurs all paralogs together into an erroneous consensus.
250 A second, orthogonal quality control measure indicates that shotgun sequencing
251 with nanopore reads still lags behind clone-based approaches; 18% of CHM13 BAC
252 sequences from segmental duplications and other difficult-to-assemble regions
253 were missing from the whole genome shotgun assembly of CHM13hTERT[30]. While
254 sequencing technologies continue to improve in read length and accuracy, clone-
255 based approaches will continue to be relevant for generating highly accurate
256 reference sequence, particularly in otherwise inaccessible ampliconic regions.

257 Relative to previous versions of SHIMS, SHIMS 3.0 greatly reduces the resources
258 required to successfully generate finished sequence. Under SHIMS 1.0, the cost to
259 produce draft sequence averaged about $5000 per clone, while finishing averaged
260 around $4000. Weeks of bench experiments and many days of expert review were

261 required to transform each low coverage (5-8x) Sanger draft sequence with
262 frequent gaps into a complete and contiguous assembly. In SHIMS 2.0, we reduced
263 the need for costly finishing activities by opting for much higher coverage (50-80x)
264 in much cheaper Illumina reads. We encounter fewer coverage gaps at this higher
265 depth, and also fewer library gaps because of differences in the library preparation
266 protocol. We rely on sonication to provide random shearing, and amplify library
267 fragments by PCR, as opposed to cloning fragments in *E. coli*. Although this deep and
268 relatively even coverage ensured that wet-bench experiments were rarely required
269 for finishing, structurally complex clones still required several days of expert review
270 using an assembly editor like Consed. In the most complex cases, involving many
271 paralogous repeats within a single clone, such as the *TSPY* array or centromeric
272 satellite repeats, it was still impossible to completely resolve the correct structure.

273 By incorporating full-length nanopore reads from each clone, SHIMS 3.0 now makes
274 it possible to assemble even the most internally repetitive clones. Full-length
275 nanopore reads provide complete certainty about the overall clone structure; there
276 is no doubt about the order and orientation of sequences, and no question about the
277 copy number of complex repeats. This limits finishing activity to the simple matter
278 of resolving the few remaining discrepancies between nanopore and Illumina reads.
279 In most clones, this requires less than an hour of effort for even inexperienced
280 finishers, and results in highly accurate sequences, with less than 1 error per
281 megabase. Highly repetitive clones require more attention, but they can be resolved
282 by a simple divide-and-conquer strategy, where each paralogous repeat is finished
283 separately, with special attention to SFV sites, and then merged to create the full

284 finished sequence. Correctly mapping short reads to repeated sequences becomes
285 more difficult as the number of paralogs increases, increasing the chances that each
286 paralogous repeat unit is blurred toward the consensus. In contrast to WGS
287 strategies, in SHIMS 3.0, this blurring is confined to the boundaries of a single clone,
288 and comparisons with neighboring clones can be used to resolve the position of
289 paralogous SFVs. SHIMS 3.0 dramatically decreases the time, cost, and effort
290 required to obtain finished sequence; using an optimized protocol for preparing
291 HMW DNA in parallel from pools of BAC clones, a small team can finish 24 Illumina
292 draft assemblies in 2 weeks for $80 per clone.

293 Limitations of SHIMS 3.0

294 SHIMS 3.0 exceeds the capabilities of previous iterations of the SHIMS technique,
295 providing access to the longest, most highly identical ampliconic sequences, as well
296 as arrays of repeated sequences shorter than a single clone. However, SHIMS 3.0
297 shares two of the same limitations as previous versions of SHIMS and other clone-
298 based approaches. First, the maximum size of BAC inserts limits SHIMS to resolving
299 duplications with <99.999% identity. This limitation will remain until long-read
300 technologies are able to surpass BAC sequencing in both read length and accuracy,
301 or a reliable cloning technology emerges that exceeds the insert size of BACs.
302 Second, SHIMS is limited to sequences that can be cloned into *E. coli*. Sequences that
303 are toxic to *E. coli* are underrepresented in BAC and fosmid libraries. These library
304 gaps can be resolved by directed efforts that avoid cloning in *E. coli*, like sequencing

305 long-range PCR products[4], or using the emerging selective sequencing ("ReadUntil")
306 capability of nanopore-based sequencers to enrich for reads flanking the gap.

307 Practitioners of SHIMS 3.0 also face new challenges due to their reliance on
308 nanopore reads spanning 100-300 kb BAC inserts. Bioinformatics tools for aligning,
309 visualizing, and editing reads of this length are not fully mature. SAM and BAM files
310 both encode alignment details in the CIGAR format, however, the BAM format is
311 limited to 65535 operations, which is frequently too few to encode the many
312 transitions between matches, mismatches, insertions and deletions encountered in
313 alignments of ultra-long nanopore reads[31]. Moreover, Consed does not reliably
314 display alignments of reads longer than 1 kb. Our workaround has been to split SAM
315 formatted alignments of nanopore reads into uniquely named sub-alignments every
316 1000 match operations, and convert the resulting SAM files to BAM format, which is
317 accepted by Consed. This permits full visualization of nanopore reads alongside
318 Illumina reads during finishing.

319 Expertise

320 As with our previous protocol for SHIMS 2.0, we have designed the SHIMS 3.0
321 protocol to be carried out by a small team. A single technician can process 24 BAC
322 clones from frozen stocks to nanopore sequencing libraries in 5 days with common
323 molecular biology lab equipment. A bioinformatics specialist can set up a pipeline to
324 identify full-length reads for each clone, generate a consensus sequence,
325 automatically correct most errors using alignments with short reads, and manually
326 review the resulting assembly for errors and identify SFVs. It is important to keep

327 abreast of new developments in software for processing nanopore data, as all
328 aspects from base-calling to alignment and error correction are continuously being
329 improved.

330

331 # MATERIALS

332

333 ## REAGENTS

334 - Ethyl Alcohol 200 Proof (Ethanol; Pharmco-Aaper, cat. no. 11000200)

335 **!CAUTION** Ethyl Alcohol is flammable; keep away from flame when handling

336 it.

337 - Tris Base (AmericanBio, cat. no. AB020000-05000)

338 - Hydrochloric Acid (HCl; VWR, cat. no. BDH3026-500MLP)

339 **!CAUTION** Hydrochloric Acid is corrosive. Wear gloves and eye protection

340 when handling it.

341 - Polyethylene glycol 8000(PEG-8000; Sigma-Aldrich, cat. no. P2139-2KG)

342 - Sodium Chloride (NaCl; AmericanBio, cat. no. AB01915-10000)

343 - Bacto-tryptone (BD Biosciences, cat. no 211705)

344 - Yeast Extract (BD Biosciences, cat. no 211929)

345 - Chloramphenicol (Sigma-Aldrich, cat. no. C0378-5G)

346 **!CAUTION** Chloramphenicol powder is hazardous. Handle this reagent in

347 ventilated fume hood with gloves and eye protection.

348 - Glycerol (EMD Millipore Corp., cat. no. 356350-1000ML)

349 - ZymoPURE II Plasmid Maxiprep Kit (Zymo Research, cat. no. D4203)

350 - Rapid Sequencing Kit (Oxford Nanopore Technologies, cat. no. SQK-RAD004)

351

352 **EQUIPMENT**

353 - EZ-Vac Vacuum Manifold (Zymo Research, cat. no. S7000)
354 - MinION (Oxford Nanopore Technologies)
355 - NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, cat. no. ND
356 1000)
357 - Centrifuge 5810 R (Eppendorf, cat. no. 00267023)
358 - Microcentrifuge 5425 (Eppendorf, cat. no. 5405000107)
359 - Vortex-Genie 2 (Scientific Industries, cat. no. SI-0236)
360 - Portable Pipet-Aid XP2 Pipette (Drummond, cat. no. 4-000-501)
361 - Eppendorf ThermoMixer (Eppendorf, cat. no. 5380000028)
362 - Beckman Coulter Avanti J-E centrifuge (Beckman Coulter, cat, no. A20698)
363 - Fisherbrand Low-Retention Microcentrifuge Tubes (Fisher Scientific, cat. no.
364 02-681-320)
365 - ART Barrier Specialty Pipette Tips, 1000, wide bore (Fisher Scientific, cat. no.
366 2069GPK)
367 - 50 ml Falcon Tube (Fisher Scientific, cat. no. 14-959-49A)
368 - Nalgene™ PPCO Centrifuge Bottles with Sealing Closure (Fisher Scientific, cat.
369 no 3141-0250)
370 - Costar Assay Plate 96-well (Corning, cat. no. 3797)
371 - Nunc 96 DeepWell Plate (Thermo Fisher Scientific, cat. no. 278743)
372 - AirPore Tape Sheets (Qiagen, cat. no. 19571)
373 - Parafilm M film (Bemis, PM996)
374 - Aluminum Adhesive Foil (Bio-rad, cat. no. MSF1001)

SOFTWARE

- minimap2 (https://github.com/lh3/minimap2)
- racon (https://github.com/lbcb-sci/racon)
- samtools (http://www.htslib.org)
- tg_index (http://staden.sourceforge.net)
- gap5 (http://staden.sourceforge.net)
- (optional) consed (http://www.phrap.org/consed/consed.html)

REAGENT SETUP

70% (vol/vol) Ethanol

Mix 30 ml of 100% (vol/vol) ethanol with 70 ml of of distilled, deionized water (ddH_2O).

▲CRITICAL 70% (vol/vol) ethanol should be prepared on the day of the experiment.

1M Tris-Cl, pH 8.5

Dissolve 121 g of Tris base in 800 ml of ddH_2O. Adjust pH to 8.5 with concentrated HCl, then adjust volume with ddH_2O to 1 l. 1M Tris-Cl can be prepared in advance and stored at room temperature (22 °C) for up to a year.

395 **10 mM Tris-Cl, pH 8.5**

396 Mix 0.5 ml of 1M Tris-Cl with 49.5 ml of ddH_2O. This solution can be prepared in

397 advance and stored at room temperature for up to a year.

398

399 **PEG buffer (18% PEG(wt/vol)/1M NaCl Solution)**

400 Add 135 g of PEG-8000 powder into 1 l bottle. Add 150 ml of 5M NaCl, 7.5 ml of Tris-

401 HCl, 1.5 ml of 0.5M EDTA and 450 ml of ddH_2O to make PEG buffer. Store at room

402 temperature for up to 1 year.

403

404 **2X LB**

405 Add 20 g of bacto-tryptone, 10 g of yeast extract, and 20 g of NaCl to ddH2O, and

406 adjust the volume to 1 l. Mix well with a magnetic stirrer. After mixing, distribute

407 500-ml aliquots into 1 l bottles. Cap loosely, prewarm to 50 °C, and autoclave for 20

408 min on liquid cycle. Store at room temperature for up to 1 year.

409

410 **Chloramphenicol**

411 Dissolve 0.34 g of chloramphenicol into 10 ml of 100% (vol/vol) ethanol.

412 Chloramphenicol stock can be stored at −20 °C for up to 1 year.

413

414 **80% (vol/vol) Glycerol solution**

415 Add 400 ml of glycerol to a graduated cylinder; adjust the volume to 500 ml with

416 ddH2O. Seal the cylinder with Parafilm M film, and mix by inversion. Transfer to a

417 bottle, and autoclave for 20 min in liquid cycle. This solution can be prepared in

418 advance and stored at room temperature for up to 1 year

419

420 # PROCEDURE

421

422 ## Pick Clones and Grow Cultures ⏱TIMING 18 h

423 1| Fill each well of a Nunc 96 DeepWell plate with 1.9 ml of 2X LB containing 34
424 µg/ml chloramphenicol.
425 ▲**CRITICAL STEP** Rich media (2X LB) is appropriate for single-copy plasmids
426 like BACs or fosmids, which use chloramphenicol resistance as a selectable
427 marker.
428 2| Use a clean pipette tip to scrape the surface of a frozen glycerol stock and
429 drop the tip directly into the DeepWell plate to inoculate a well. Inoculate
430 each sample 8 times for a total of 15.2 ml/sample. 24 samples in total for
431 each library prep.
432 ▲**CRITICAL STEP** Start with a glycerol stock of a clone verified to be correct by
433 PCR for a known sequence-tagged site or previous sequencing experiments
434 (e.g. cultures created during Step 7 of the SHIMS 2.0 protocol).
435 3| Seal plates with AirPore Tape Sheets and place at 37 °C for 16-17 h, shaking
436 at 220 RPM.
437 ▲**CRITICAL STEP** Overgrowth of cultures (cell density > 3 x 10^9 cells per ml)
438 will decrease yield of BAC DNA.

439 **Glycerol Stock Plate ?TIMING 30 min**

440 4| Dispense 150 µl of 80% (vol/vol) glycerol solution into two rows of a Costar
441 Assay Plate.

442 5| Transfer 150 µl of each sample culture from Step 3 to a corresponding well of
443 the assay plate and mix by pipetting up and down 20 times.

444 6| Seal the glycerol stock plate with aluminum adhesive foil.

445 7| Store the glycerol stock plate at -80 °C.

446 **Pooling Clones ?TIMING 1-2h**

447 8| Pour overnight cultures from Step 3 into a large beaker to combine pool.

448 9| Divide pooled culture into two 250 ml Nalgene bottles and spin down culture
449 at 6000*g* for 30 min at 4 °C.

450 10| Remove media by pouring into a waste-collecting container. Be careful not to
451 disturb the pellets.

452 **PAUSE POINT** Store at -20 °C for up to a week.

453 **Alkaline Lysis ?TIMING 1-2 h**

454 11| Add 7 ml of ZymoPURE P1 (Red) to each bacterial cell pellet and resuspend
455 completely by pipetting. Combine into one bottle when cells are completely
456 resuspended.

457 12| Add 14 ml of ZymoPURE P2 (Green) and immediately mix by gently inverting
458 the tube 6 times.
459 ▲**CRITICAL STEP** Do not vortex! Let sit at room temperature for 3 min. Cells
460 are completely lysed when the solution appears clear, purple, and viscous.
461 13| Add 14 ml of ZymoPURE P3 (Yellow) and mix gently but thoroughly by
462 inversion.
463 ▲**CRITICAL STEP** Do not vortex! The sample will turn yellow when the
464 neutralization is complete, and a yellowish precipitate will form.
465 14| Ensure the plug is attached to the Luer Lock at the bottom of the ZymoPURE
466 Syringe Filter. Place the syringe filter upright in a tube rack and load the
467 lysate into the ZymoPURE Syringe Filter and wait 8 min for the precipitate to
468 float to the top.
469 15| Remove the Luer Lock plug from the bottom of the syringe and place it into a
470 clean 50 ml conical tube. Place the plunger in the syringe and push the
471 solution through the ZymoPURE Syringe Filter in one continuous motion
472 until approximately 33-35 ml of cleared lysate is recovered. Save the cleared
473 lysate!
474 16| Add 14ml ZymoPURE Binding Buffer to the cleared lysate from step 5 and
475 mix thoroughly by inverting the capped tube 10 times.
476 17| Ensure the connections of the Zymo-Spin V-P Column Assembly are finger-
477 tight and place onto a vacuum manifold.

478 18| With the vacuum off, add the entire mixture from step 6 into the Zymo-Spin
479 V-P Column Assembly, and then turn on the vacuum until all the liquid has
480 passed completely through the column.
481 19| Remove and discard the 50 ml reservoir from the top of the Zymo-Spin V-P
482 Column Assembly.
483 20| With the vacuum off, add 5 ml of ZymoPURE Wash 1 to the 15 ml Conical
484 Reservoir in the Zymo-Spin V-P Column Assembly. Turn on the vacuum until
485 all the liquid has passed completely through the column.
486 21| With the vacuum off, add 5 ml of ZymoPURE Wash 2 to the 15ml Conical
487 Reservoir. Turn on the vacuum until all the liquid has passed completely
488 through the column. Repeat this wash step.
489 22| Remove and discard the 15 ml Conical Reservoir and place the Zymo-Spin V-
490 P Column in a Collection Tube. Centrifuge at ≥10,000*g* for 1 minute, in a
491 microcentrifuge, to remove any residual wash buffer.
492 23| Pre-warm 450 µl of 10 mM Tris-Cl to 50 °C. Transfer the column into a clean
493 1.5ml microcentrifuge tube and add the 450 µl of 10 mM Tris-Cl directly to
494 the column matrix. Wait 10 min, and then centrifuge at ≥ 10,000*g* for 1
495 minute in a microcentrifuge.
496 24| Add 450 µl of PEG buffer to the tube containing sample. Mix by flicking and
497 rotating the 1.5 ml microcentrifuge tube.
498 25| Centrifuge at ≥10,000*g* for 30 min at 4 °C, in a microcentrifuge.
499 26| Remove supernatant from the tube without disturbing the pellet.
500 27| Add 1 ml of 70% EtOH and spin for 10 min at 4 °C.

501 28| Repeat step 27 and 28.

502 29| Remove supernatant and any left over 70% EtOH from microcentrifuge tube.

503 30| Air dry for 10 min or until no visible liquid is left in the tube.

504 ▲**CRITICAL STEP** Do not overdry the pellet.

505 31| Dissolve DNA pellet in 18 µl 10 mM Tris-Cl.

506 32| Store DNA at 4 °C for several days until pellet completely dissolves into

507 solution.

508 33| Check DNA concentration and quality with Qubit or NanoDrop. Aim for

509 concentration ≥ 1 µg/µl, an A260/280 ratio ~ 1.8, and an A260/230 ratio

510 between 2.0 and 2.2.

511 **?TROUBLESHOOTING**

512

513 **MinION Library prep ?TIMING 30 Minutes**

514 34| Adjust sample concentration from step 33 to 1 µg/µl with 10 mM Tris-Cl.

515 **?TROUBLESHOOTING**

516 35| Using a wide bore pipette tip, slowly aspirate 15 µl into a low retention

517 microcentrifuge tube.

518 ▲**CRITICAL STEP** Pipetting can shear fragile high-molecular-weight DNA.

519 Pipette slowly using wide bore pipette tips.

520 36| In a separate microcentrifuge tube, add 0.5 µl FRA to 4.5 µl 10 mM Tris-Cl.
521 Flick the tube to mix well.

522 ▲**CRITICAL STEP** The FRA solution is included in the Rapid Sequencing Kit.

523 37| Add the diluted FRA solution from step 36 into sample tube from step 35.
524 38| Gently flick the tubes a few times to mix.

525 ▲**CRITICAL STEP** Vortexing can shear fragile high-molecular-weight DNA.
526 Flick solutions to mix.

527 39| Incubate sample on 30 °C heat block for 35 seconds, then move the tube to 80
528 °C heat block. Incubate for 2 min at 80 °C.
529 40| Remove the tube from heat block and incubate on ice for 1 minute, then
530 move the tube off the ice. Equilibrate to room temperature, about 1 minute.
531 41| While the sample is equilibrating to room temperature, add 4.5 µl 10 mM
532 Tris-Cl to 0.5 µl of RAP. Flick to mix well.

533 ▲**CRITICAL STEP** The RAP solution is included in the Rapid Sequencing Kit.

534 42| Add RAP dilution from step 41 into sample tube. Slowly flick the tube a few
535 times to mix. Keep the sample at room temperature before loading.

536 ▲**CRITICAL STEP** Vortexing can shear fragile high-molecular-weight DNA.
537 Flick solutions to mix.

538 **MinION Library loading ⍰TIMING 30 Minutes**

539 43| Add 30 µl of FLT to one tube of FLB, to make flush buffer, according to the
540 Rapid Sequencing Kit instructions. Vortex the solution to mix , then
541 centrifuge briefly.
542 ▲**CRITICAL STEP** The FLT and FLB solutions are included in the Rapid
543 Sequencing Kit.
544 44| Perform QC on a new MinION flow cell to check available pores and ensure
545 that a sufficient number of pores are present. If there are fewer than 800
546 available pores, place the flow cell in storage and use a new MinION flow cell.
547 45| Use a P1000 pipette to remove about 20-30 µl of storage buffer from priming
548 pore. Load 800 µl flush buffer via the pore slowly. Wait 5 min.
549 46| Lift SpotON cover and load 200 µl flush buffer slowly. Try to dispense at a
550 speed where each bead of liquid is siphoned into the SpotON port as soon as
551 it is visible.
552 47| Add 34 µl SQB and 15 µl ddH_2O to the sample tube from step 42.
553 ▲**CRITICAL STEP** The SQB solution is included in the Rapid Sequencing Kit.
554 48| Flick the tube gently to mix, then centrifuge briefly to collect library to the
555 bottom of the tube.

556 ▲**CRITICAL STEP** Vortexing can shear fragile high-molecular-weight DNA.
557 Flick solutions to mix.

558 49| Slowly aspirate 75 µl of library with a wide bore tip. Very slowly, load the
559 library into SpotON pore drop by drop.

560 ▲**CRITICAL STEP** Pipetting can shear fragile high-molecular-weight DNA.
561 Pipette slowly using wide bore pipette tips.

562 50| Close both priming pores and put the SpotON cover back onto the pore.
563 51| After loading the library, leave the flow cell on bench for 45 min before
564 starting the run.

565 ▲**CRITICAL STEP** Wait at least 45 minutes between loading the flow cell and
566 starting the run. This allows time for full-length molecules to diffuse to the
567 pores. Starting the run earlier will favor the sequencing of shorter molecules.

568 **?TROUBLESHOOTING**

569

570 **Demultiplex Reads ?TIMING 30 Minutes**

571 52| Prepare file of draft clone sequences in fasta format: `draft_clones.fa`
572 ▲**CRITICAL STEP** When concatenating draft sequence assemblies, ensure that
573 each sequence has a unique name.

574 53| Prepare file of vector sequence in fasta format: `vector.fa`

575 54| Download fastq formatted reads from the device running MinION control
576 software: `nanopore.fq`

55| Align nanopore reads to file of draft sequences to assign nanopore reads to clones by best match:

?TROUBLESHOOTING

```
minimap2 -x map-ont draft_clones.fa nanopore.fq | sort -r -n -k 10 | awk '!seen[$1]++' > best_clone_match.paf
```

```
grep clone_name best_clone_match.paf | cut -f 1 >clone_name.txt
```

```
grep -A 3 -f clone_name.txt nanopore.fq | grep -v "^--$" > clone_name.nanopore.fq
```

Identify Full-length Reads ?TIMING 30 Minutes

▲**CRITICAL** We have automated Steps **56-65** with a custom Perl script (available at https://github.com/dwbellott/shims3_assembly_pipeline/), but the workflow is described below to allow for direct use of the individual software tools or substitution of alternative tools.

56| For each clone, align nanopore reads to file of vector sequence:

```
minimap2 -x map-ont vector.fa clone_name.nanopore.fq –o clone_name.vector.paf
```

57| Search for reads that begin and end with high-quality matches to vector sequence on the same strand – these are full-length reads.

```
cut -f 1,5,6 clone_name.vector.paf | sort | uniq -c |
sed 's/^//' | grep "^2" | cut -f 2 -d ' '
>clone_name.2x.txt
```

```
awk '$2 - $3 < $7 && $12 == 60' clone_name.vector.paf |
cut -f 1,5,6 | grep -f clone_name.2x.txt
>clone_name.2x.right.txt
```

```
awk '$4 < $7 && $12 == 60' clone_name.vector.paf | cut
-f 1,5,6 | grep -f clone_name.2x.right.txt | cut -f 1 |
sort | uniq >clone_name.fl.txt
```

58| For each clone, generate a fastq file of full-length reads, as well as a fasta file of the longest full-length read to use as a scaffold for final assembly.

?TROUBLESHOOTING

```
grep -A 3 -f fl.txt clone_name.nanopore.fq | grep -v
"^--$" > clone_name.fl.fq
```

```
grep -A 1 `head -n 1 clone_name.fl.txt` nanopore.fq |
sed 's/^\@.*/\>clone_name/' >clone_name.longest.fl.fa
```

Generate Consensus Sequence ?TIMING 30 Minutes

59| Polish the longest read twice, using the other full-length nanopore reads:

```
minimap2 -x map-ont clone_name.longest.fl.fa
clone_name.fl.fq >clone_name.longest.fl.paf

racon clone_name.fl.fq clone_name.longest.fl.paf
clone_name.longest.fl.fa
>clone_name.longest.fl.racon.fa

minimap2 -x map-ont clone_name.longest.fl.racon.1.fa
clone_name.fl.fq >clone_name.longest.fl.racon.paf

racon clone_name.fl.fq
clone_name.longest.fl.racon.1.paf
clone_name.longest.fl.racon.1.fa
>clone_name.fl.consensus.fa
```

60| Gather up Illumina, nanopore, and (if available) PacBio reads for each clone.

```
cat clone_name.illumina.forward.fq
clone_name.illumina.reverse.fq
clone_name.illumina.single.fq clone_name.nanopore.fq
clone_name.pacbio.fq >> clone_name.allreads.fq
```

61| Polish the nanopore consensus sequence, using both long and short reads.

```
minimap2 -x asm20 clone_name.nanopore.consensus.fa
clone_name.illumina.single.fq >>
clone_name.polish.1.paf

minimap2 -x sr clone_name.nanopore.consensus.fa
clone_name.illumina.forward.fq
```

```
clone_name.illumina.reverse.fq >>
clone_name.polish.1.paf

minimap2 -x map-ont clone_name.nanopore.consensus.fa
clone_name.nanopore.fq >> clone_name.polish.1.paf

minimap2 -x map-pb clone_name.nanopore.consensus.fa
clone_name.pacbio.fq >> clone_name.polish.1.paf

racon clone_name.allreads.fq clone_name.polish.1.paf
clone_name.nanopore.consensus.fa
>clone_name.polish.1.fa
```

62| Repeat Step 61 four more times, for a total of 5 rounds of polishing

63| Align reads one last time to generate SAM format alignments suitable for assembly editors.

▲**CRITICAL STEP** The BAM file format cannot accommodate CIGAR strings with greater than 65535 operations. Alignments involving nanopore reads spanning the full length of a BAC clone will exceed this limit. We strongly recommend storing alignments in SAM or CRAM format to avoid the loss of detailed alignment information.

```
minimap2 -x asm20 -a -L --sam-hit-only -R
'@RG\tID:S\tSM:S\tPL:ILLUMINA' clone_name.polish.5.fa
clone_name.illumina.single.fq | samtools sort -O SAM -
>clone_name.single.sorted.sam

minimap2 -x sr -a -L --sam-hit-only -R
'@RG\tID:FR\tSM:FR\tPL:ILLUMINA' clone_name.polish.5.fa
```

```
clone_name.illumina.forward.fq
clone_name.illumina.reverse.fq | samtools sort -O SAM -
>clone_name.paired.sorted.sam
```

```
minimap2 -x map-pb -a -L --sam-hit-only -R
'@RG\tID:P\tSM:P\tPL:PACBIO' clone_name.polish.5.fa
clone_name.pacbio.fq | samtools sort -O SAM -
>clone_name.pacbio.sorted.sam
```

```
minimap2 -x map-ont -a -L --sam-hit-only -R
'@RG\tID:N\tSM:N\tPL:PACBIO' clone_name.polish.5.fa
clone_name.nanopore.fq | samtools sort -O SAM -
>clone_name.nanopore.sorted.sam
```

64| Combine alignments

```
samtools merge -f clone_name.allreads.sorted.sam
clone_name.single.sorted.sam
clone_name.paired.sorted.sam
clone_name.pacbio.sorted.sam
clone_name.nanopore.sorted.sam
```

65| Generate database for Gap5

▲**CRITICAL STEP** We now recommend Gap5 over Consed, because Gap5 natively supports loading data directly from SAM files and displaying full-length nanopore reads. It is possible to split SAM alignments of full-length nanopore reads into smaller fragments that can be encoded in a BAM file and displayed by Consed without loss of information. For those who wish to use

685 Consed, we implement this work-around in a custom Perl script (available at
686 https://github.com/dwbellott/shims3_assembly_pipeline/).

```
687 tg_index -o clone_name.g5d -p -9 -s
688 clone_name.allreads.sorted.sam
```

689

690 **Finishing TIMING 0-8 h**

691 66| Open the assembly in Gap5:

```
692 gap5 clone_name.g5d
```

693 67| Select 'Edit Contig' from the 'Edit' menu.

694 68| Resolve discrepancies between Illumina reads and full-length nanopore
695 reads (**Fig. 3**).

696 ▲**CRITICAL STEP** In Gap5, it is not possible to directly edit the consensus
697 sequence. Instead, indicate which readings are authoritative by marking
698 bases as high quality with the ']' key, and the consensus will automatically
699 update.

700 ▲**CRITICAL STEP** We usually resolve the consensus in favor of the Illumina
701 reads. The vast majority of discrepancies between these technologies occur
702 at homopolymer repeats, where nanopore reads are especially prone to
703 insertion and deletion errors (**Fig. 3a**). More rarely, we encounter systematic
704 errors in nanopore base calling that generate a consensus that is not
705 supported by any Illumina read.

706 ▲**CRITICAL STEP** We resolve disagreements among Illumina reads in favor of
707 the consensus of full-length nanopore reads. In clones that contain duplicated
708 sequences, short Illumina reads can be mapped to the wrong repeat unit, but
709 full-length nanopore reads are not subject to this artifact, and will usually
710 have the correct base at each SFV.

711 69| Resolve SSRs by realigning reads around the SSR region. Select reads by
712 clicking on their names on the left hand side of the edit window, and choose
713 'Realign Selection' from the 'Command' menu.

714 ▲**CRITICAL STEP** Stutter noise from replication slippage in SSRs causes
715 divergent reads and low-quality base calls. In some cases, unambiguous
716 resolution of these repeats may not be possible, and they should be
717 annotated as unresolved in Step 71.

718 70| Remove any vector-sequence contamination at the ends of the clone. In the
719 Gap5 edit window, use the 'Search' button to search the consensus sequence
720 for the sequences at the cloning site of your vector. Trim away the vector
721 sequence outside of the restriction sites used to generate your clone library
722 (usually EcoRI, BamHI, or MboI).

723 71| Annotate any remaining ambiguities in the clone sequence (e.g., unresolved
724 simple sequence repeats, where neither Illumina or nanopore reads are
725 completely accurate) by compiling a feature table[32], which will be useful
726 when finished clone sequences are submitted to GenBank.

727 **?TROUBLESHOOTING**

?TIMING

Steps 1-3, pick clones and grow cultures: 18 h

Steps 4-7, glycerol stock plate: 30 min

Steps 8-10, pooling clones: 1-2 h

Steps 11-33, alkaline lysis: 1-2 h

Steps 34-42, MinION library prep: 30 min

Steps 43-51, MinION library loading: 30 min

Steps 52-55, demultiplex reads: 30 min

Steps 56-58, Identify full-length reads: 30 min

Steps 59-65, generate consensus sequence: 30 min

Steps 66-71, finishing: 0-8 h

?TROUBLESHOOTING

Troubleshooting advice can be found in **Table 1**.

744 **Table 1 | Troubleshooting Table.**

Step	Problem	Possible Reason	Solution
33	Low DNA concentration	Culture undergrowth or overgrowth	Check culture OD600 is between 0.2-0.35
		Incomplete Lysis	Make sure to thoroughly mix the solution until the color is uniform
		Incomplete neutralization	Solution from step 13 should not appear viscous and precipitate should float to the surface
		Incomplete DNA elution	Pre-warm elution buffer to 50°C
34	Concentration varies when checking with NanoDrop or Qubit	DNA is not completely mixed	After adjusting concentration from step 33, leave DNA solution on a heated shaker at the gentlest setting at 50°C until DNA is completely mixed
51	Pores decrease rapidly	Impure DNA sample	Re-check DNA concentration. Extract DNA again if NanoDrop and Qubit results are discordant, 260/280 < 1.7, 260/280 > 2.0, 260/230 < 2.0, or 260/230 > 2.2
		Bubbles introduced during loading	Pipette very slowly and take care not to introduce bubbles during flow cell priming and library loading

55	No reads for one or more clones	Clone culture failed	Regrow and add to the next run, or replace the clone with another
			Regrow the clone for an additional round of sequencing
		Bookkeeping error; some common bookkeeping errors result from transposing digits, rotating a plate by 180°, or contamination from a clone in an adjacent well	Resolve bookkeeping error, and rerun a new clone or replace with another clone
58	Low fraction of long reads	FRA treatment time too long	Promptly heat-inactivate FRA at 35 seconds
			Adjust the FRA incubation time below 35 seconds
		Shearing during library prep	Use wide-bore tips for all mixing and loading steps
71	Clone sequence is shorter than expected or missing known sequence	Deletion during culture	Regrow the clone from the original culture or another library copy, and replace with the alternate clone
		Sequence toxic to *E. coli*	Close the gap by long-range PCR or region-specific extraction

745

Anticipated Results

We typically pool 24 clones for a single MinION run, generating about 300,000 reads with a read n50 of 20 kb, and a total of about 1.5 Gb of sequence data. Each clone typically receives 1-5% of the total reads. Occasionally some clones will have no reads; this usually indicates that the culture of the clone (Steps 1-3) has failed (see troubleshooting information for Step 55).

Expect to obtain 3-10 full-length reads per clone. Because of the high rate of insertions and deletions in individual nanopore reads, full-length reads may differ in length by 10 kb or more. Occasionally, a clone will have no reads that start and end in vector sequence, but the clone length will be apparent from a peak in the tail of the distribution of read lengths. It may still be possible to reconstruct a full-length consensus sequence by rotating one of these putative full-length reads to place the vector sequence at the beginning. However, we do not recommend this procedure for internally repetitive clones, particularly tandem arrays. Instead, sequence the clone again, and use these ambiguous reads to help polish the consensus.

Acknowledgements

This work was supported by the Howard Hughes Medical Institute, and generous gifts from Brit and Alexander d'Arbeloff and Arthur W. and Carol Tobin Brill

764 ## Author Contributions

765 D.W.B., H.S., J.F.H., and D.C.P. designed the study. D.W.B., T.-J.C., and E.K.J. developed
766 the experimental methods. D.W.B. wrote the scripts for computational analysis. T.J.C.
767 carried out the sequencing of the *TSPY* region. H.S. assembled the *TSPY* region.
768 D.W.B., T.-J.C., and D.C.P. wrote the manuscript.

769 ## Competing Interests

770 The authors declare no competing interests.

771 ## Data Availability

772 Data that support the findings of this study have been deposited in Genbank with
773 the accession codes AC279301, AC279302, AC279304, AC279305, AC279306,
774 AC279307, AC279309, AC279311, AC279312, and AC279315.

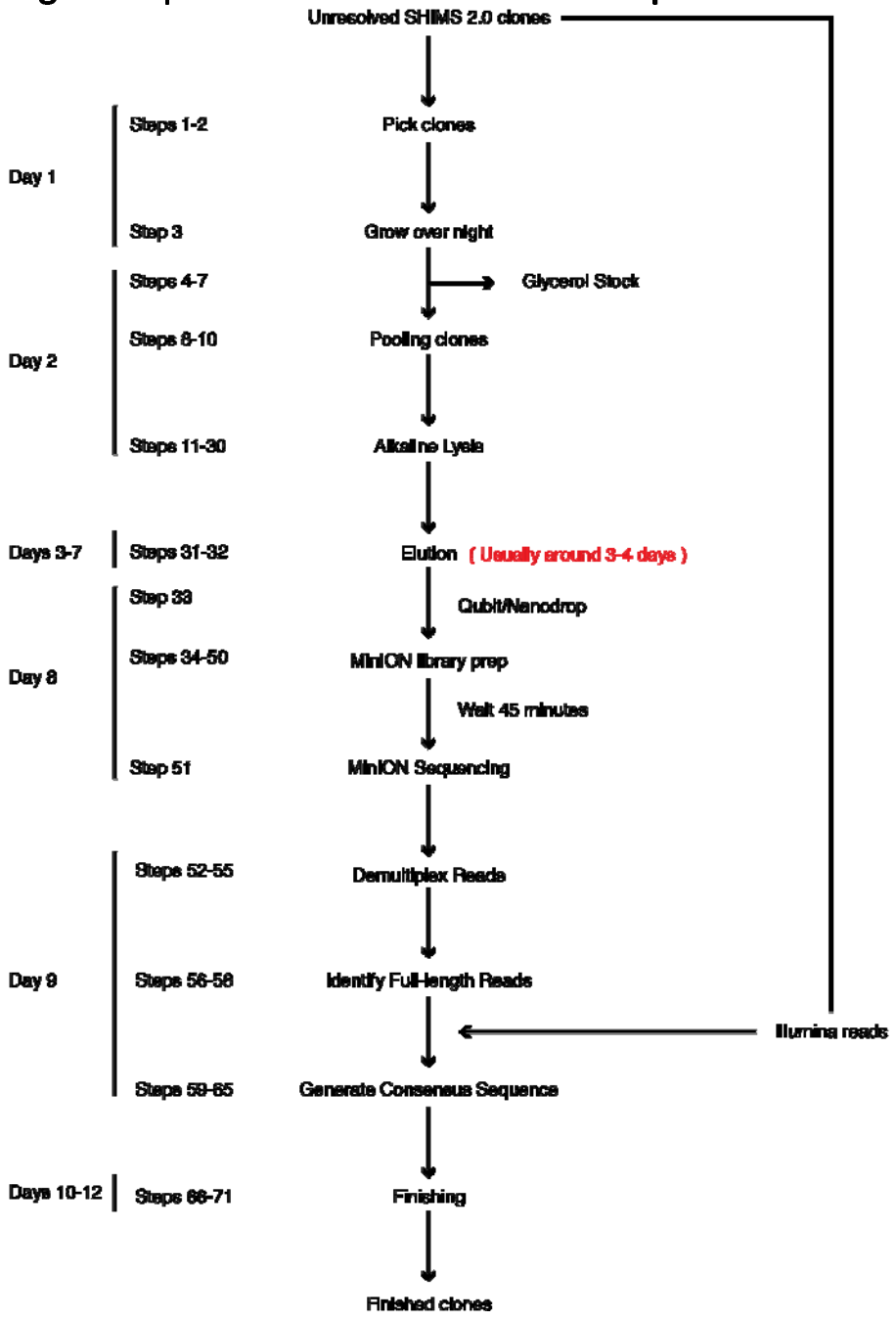
775 ## Code Availability

776 We have automated Steps **56-65** with a custom Perl script, available at:
777 https://github.com/dwbellott/shims3_assembly_pipeline/

778

Figures

781 **Figure 1 | Overview of the SHIMS3.0 protocol.**



782

783 **Figure 1** | Overview of the SHIMS3.0 protocol. A timeline of a single iteration of the
784 SHIMS 3.0 protocol, showing the major protocol steps, with key quality controls on
785 the right. During a two-week iteration, 24 clones are processed in parallel to rapidly
786 generate finished sequence from structurally complex clones. A single technician
787 can proceed from a list of clones to full-length nanopore libraries in 8 d. After a brief
788 MinION run overnight, a bioinformatics specialist can demultiplex fastq sequences,
789 identify full-length reads, then polish and edit the consensus of these reads to
790 generate finished clone sequence.

Figure 2 | Structure of the Human *TSPY* array.

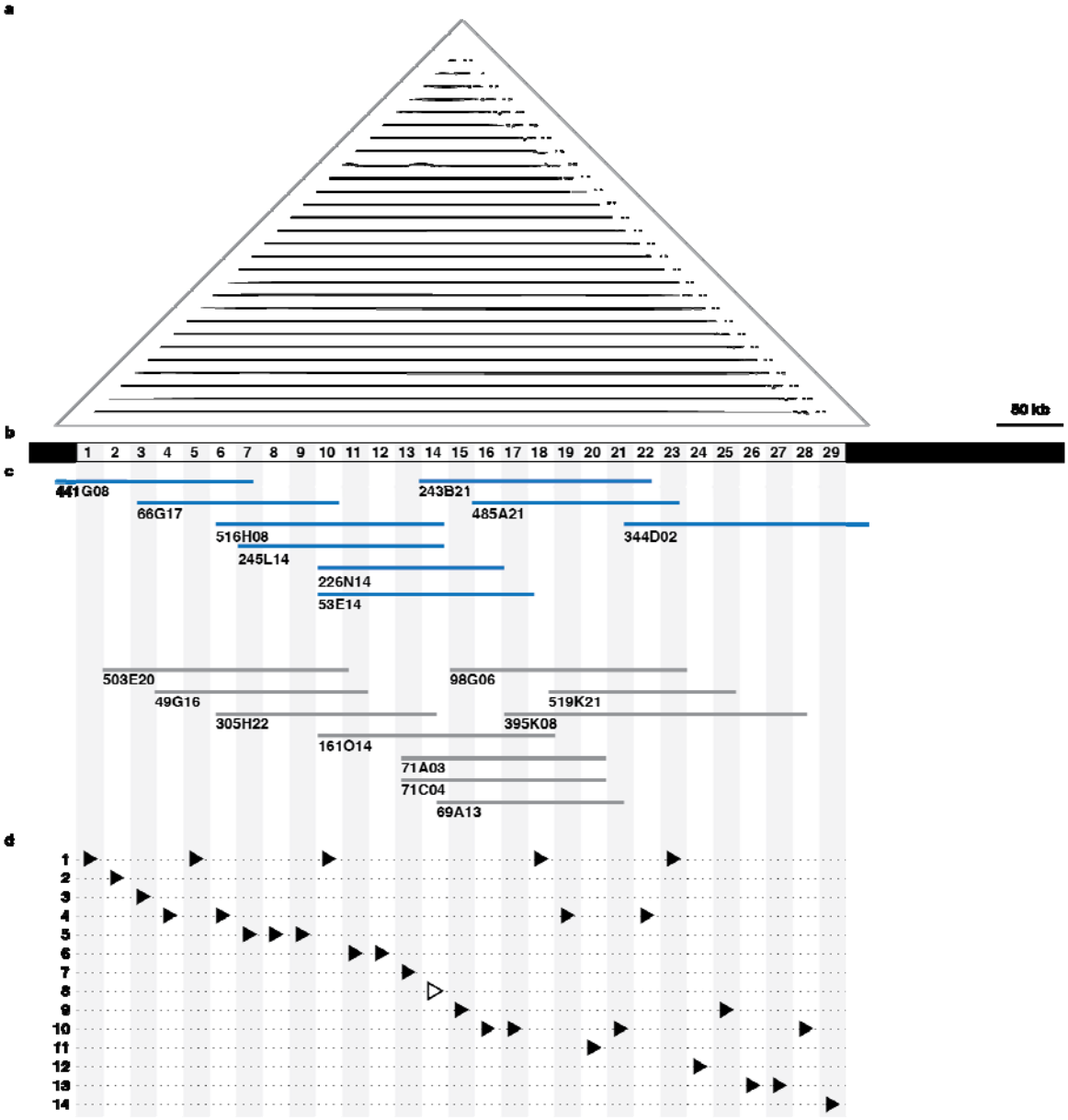


Figure 2 | Structure of the Human *TSPY* array. a) Triangular dot plot of *TSPY* array region (AC279304) assembled using SHIMS3.0; each dot represents a 100 nucleotide perfect match between sequences within the array; scale bar: 50 kb. b) Schematic representation of the 29 repeat units of the *TSPY* array. c) Clones from the RP11 BAC library sequenced with SHIMS 2.0 and SHIMS 3.0 (blue) to obtain finished sequence, or SHIMS 2.0 alone (grey) to identify sequence family variants used to map the array. d) There are 14 distinct *TSPY* transcripts (triangles), including 1 pseudogene (open triangle).

Figure 3 | Editing clone assemblies in Gap5.

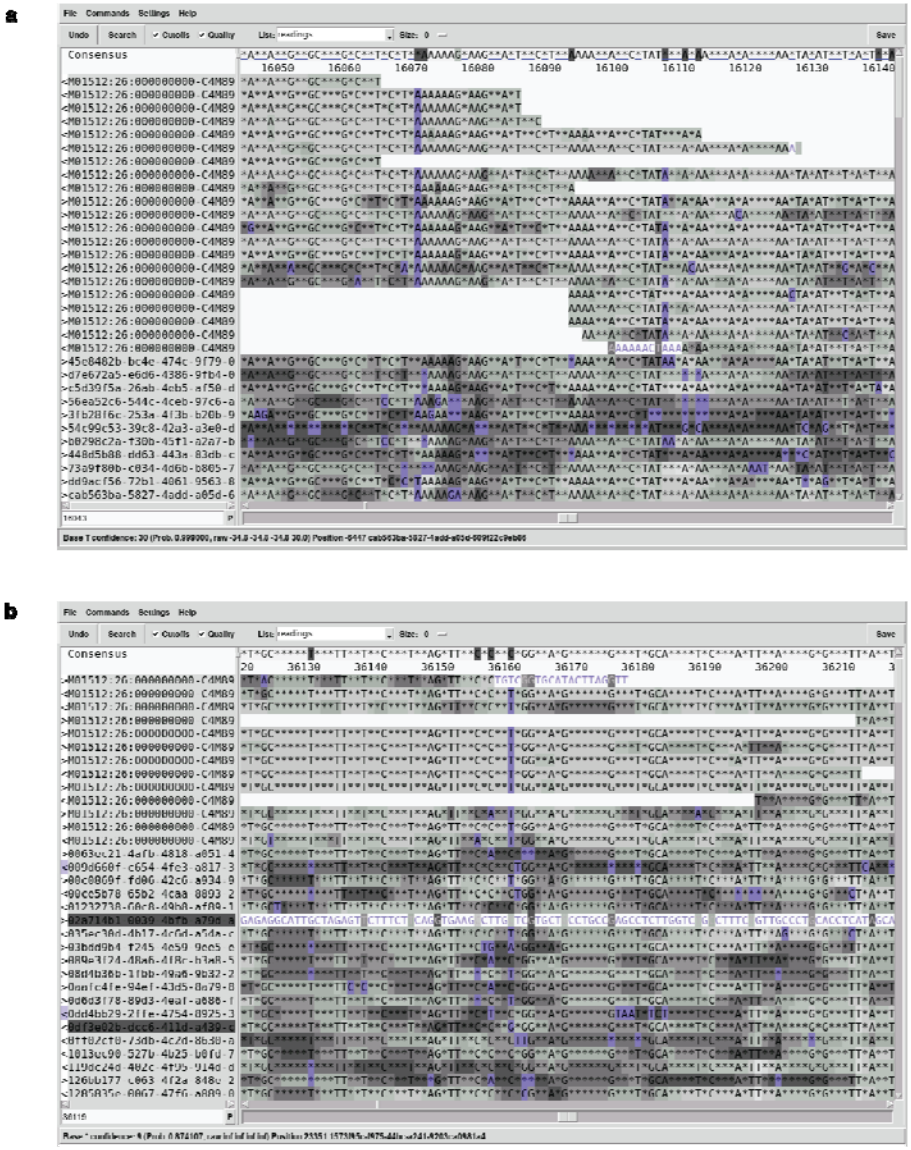


Figure 3 | Editing clone assemblies in Gap5. Screenshots from Gap5 with reads sorted by technology (Illumina on top; nanopore on bottom), showing two instances where errors in the consensus can be resolved by correcting to the consensus of the Illumina reads: a) frequent insertion and deletion errors at homopolymer runs, and b) more rare substitution errors.

810 ## References

811 1. Mueller, J. L. *et al.* Independent specialization of the human and mouse X
812 chromosomes for the male germ line. *Nature Genetics* **45**, 1083–1087 (2013).

813 2. Lupski, J. R. Genomic disorders: Structural features of the genome can lead to DNA
814 rearrangements and human disease traits. *Trends in Genetics* **14**, 417–422 (1998).

815 3. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features
816 massive palindromes and uniform recurrent deletions in infertile men. *Nature*
817 *Genetics* **29**, 279–286 (2001).

818 4. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a
819 mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).

820 5. Bellott, D. W. *et al.* Convergent evolution of chicken Z and human X chromosomes
821 by expansion and gene acquisition. *Nature* **466**, 612–616 (2010).

822 6. Hughes, J. F. *et al.* Chimpanzee and human Y chromosomes are remarkably
823 divergent in structure and gene content. *Nature* **463**, 536–539 (2010).

824 7. Hughes, J. F. *et al.* Strict evolutionary conservation followed rapid gene loss on
825 human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012).

826 8. Bellott, D. W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-
827 sensitive regulators. *Nature* **508**, 494–499 (2014).

828 9. Soh, Y. Q. S. *et al.* Sequencing the mouse Y chromosome reveals convergent gene
829 acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).

830 10. Jackson, E. K. *et al.* Large palindromes on the primate X Chromosome are
831 preserved by natural selection. *Genome Research* 31, 1337–1352 (2021).

832 11. Watson, C. T. *et al.* Complete Haplotype Sequence of the Human Immunoglobulin
833 Heavy-Chain Variable, Diversity, and Joining Genes and Characterization of Allelic
834 and Copy-Number Variation. *The American Journal of Human Genetics* **92**, 530–546
835 (2013).

836 12. Dennis, M. Y. *et al.* Evolution of Human-Specific Neural SRGAP2 Genes by
837 Incomplete Segmental Duplication. *Cell* **149**, 912–922 (2012).

838 13. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications
839 within the human genome. *Nature* **431**, 927–930 (2004).

840 14. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science*
841 **352**, aae0344–aae0344 (2016).

842 15. Bellott, D. W., Cho, T.-J., Hughes, J. F., Skaletsky, H. & Page, D. C. Cost-effective
843 high-throughput single-haplotype iterative mapping and sequencing for complex
844 genomic structures. *Nature Protocols* **13**, 787–809 (2018).

845 16. Warburton, P. E. *et al.* Analysis of the largest tandemly repeated DNA families in
846 the human genome. *BMC Genomics* **9**, 533 (2008).

847 17. Quick, J. Ultra-long read sequencing protocol for RAD004 v3
848 (protocols.io.mrxc57n). (2018) doi:10.17504/protocols.io.mrxc57n.

849 18. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**,
850 3094–3100 (2018).

851 19. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome
852 assembly from long uncorrected reads. *Genome Research* **27**, 737–746 (2017).

853 20. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics*
854 **25**, 2078–2079 (2009).

855 21. Bonfield, J. K. & Whitwham, A. Gap5--Editing the billion fragment sequence
856 assembly. *Bioinformatics* **26**, 1699–1703 (2010).

857 22. Gordon, D. & Green, P. Consed: A graphical editor for next-generation
858 sequencing. *Bioinformatics* **29**, 2936–2937 (2013).

859 23. Tyler-Smith, C., Taylor, L. & Müller, U. Structure of a hypervariable tandemly
860 repeated DNA sequence on the short arm of the human Y chromosome. *Journal of*
861 *Molecular Biology* **203**, 837–848 (1988).

862 24. Tsuchiya, K., Reijo, R., Page, D. C. & Disteche, C. M. Gonadoblastoma: Molecular
863 definition of the susceptibility region on the Y chromosome. *American Journal of*
864 *Human Genetics* **57**, 1400–1407 (1995).

865 25. Oram, S. W., Liu, X. X., Lee, T.-L., Chan, W.-Y. & Lau, Y.-F. C. TSPY potentiates cell
866 proliferation and tumorigenesis by promoting cell cycle progression in HeLa and
867 NIH3T3 cells. *BMC Cancer* **6**, 154 (2006).

868 26. Giachini, C. *et al.* TSPY1 Copy number variation influences spermatogenesis and
869 shows differences among Y lineages. *The Journal of Clinical Endocrinology &*
870 *Metabolism* **94**, 4016–4022 (2009).

871 27. Lin, S. *et al.* Comparison of the transcriptional landscapes between human and
872 mouse tissues. *Proceedings of the National Academy of Sciences* **111**, 17224–17229
873 (2014).

874 28. Jain, M. *et al.* Linear assembly of a human centromere on the Y chromosome.
875 *Nature Biotechnology* **36**, 321–323 (2018).

876 29. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-
877 long reads. *Nature Biotechnology* **36**, 338–345 (2018).

878 30. Miga, K. H. *et al.* Telomere-to-telomere assembly of a complete human X
879 chromosome. *Nature* (2020) doi:10/gg4vw8.

880 31. Li, H. A versatile pairwise aligner for genomic and spliced nucleotide sequences.
881 *minimap2* https://github.com/lh3/minimap2 (2020).

882 32. National Center for Biotechnology Information. What is tbl2asn?
883 https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/ (2017).