Cost-effective high-throughput single-haplotype iterative mapping and sequencing for complex genomic structures

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

[1]Whitehead Institute, Cambridge, Massachusetts, USA. [2]Howard Hughes Medical Institute, Whitehead Institute, Cambridge, Massachusetts, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to D.W.B. (bellott@wi.mit.edu).

Published online 22 March 2018; doi:10.1038/nprot.2018.019

The reference sequences of structurally complex regions can be obtained only through highly accurate clone-based approaches. We and others have successfully used single-haplotype iterative mapping and sequencing (SHIMS) 1.0 to assemble structurally complex regions across the sex chromosomes of several vertebrate species and to allow for targeted improvements to the reference sequences of human autosomes. However, SHIMS 1.0 is expensive and time consuming, requiring resources that only a genome center can provide. Here we introduce SHIMS 2.0, an improved SHIMS protocol that allows even a small laboratory to generate high-quality reference sequence from complex genomic regions. Using a streamlined and parallelized library-preparation protocol, and taking advantage of inexpensive high-throughput short-read-sequencing technologies, a small laboratory with both molecular biology and bioinformatics experience can sequence and assemble 192 large-insert bacterial artificial chromosome (BAC) or fosmid clones in 1 week. In SHIMS 2.0, in contrast to other pooling strategies, each clone is sequenced with a unique barcode, thus enabling clones containing nearly identical sequences to be multiplexed in a single sequencing run and assembled separately. Relative to SHIMS 1.0, SHIMS 2.0 decreases the required cost and time by two orders of magnitude while preserving high sequencing accuracy.

INTRODUCTION

Ampliconic sequences, euchromatic repeats that display >99% identity over more than 10 kb, are the most structurally complex regions in the genome and are notoriously difficult to assemble[1]. These complex repetitive structures mediate deletions, duplications, and inversions associated with human disease[2,3], but the absence of an accurate reference sequence for these regions has impeded comprehensive studies of genomic structural variation and of the mechanisms that govern the rearrangements associated with ampliconic sequences. Furthermore, experiments based on aligning short reads to existing reference sequences—such as genome and exome resequencing, RNA sequencing, and chromatin immunoprecipitation–sequencing—are necessarily limited by the quality and completeness of the reference sequence. Reanalysis of short-read data sets in light of improved reference sequences can immediately allow for rich annotation of structurally complex regions for studying their roles in human variation in health and disease.

Only extremely long and accurate reads can discriminate between amplicon copies and generate a correct reference sequence from structurally complex regions. The human reference sequence was assembled from a patchwork of BAC clones derived from the genomes of 16 diploid individuals[4]. Each BAC clone was shotgun sequenced in Sanger reads and assembled into a synthetic long read of ~150 kb with error rates as low as 1 in 1,000,000 nt (ref. 5). Guided by software tools such as Consed[6] or GAP[7], highly skilled technicians would inspect computer-generated draft assemblies for errors and anomalies; design experiments, including subcloning, PCR reactions, restriction digests, and transposon bombing; and then carry out additional sequencing to correct each assembly and produce a contiguous, high-quality 'finished' sequence. This process was both slow and expensive, and subsequent generations of sequencing technology have prioritized driving down sequencing costs at the expense of read length and accuracy. Whole-genome shotgun (WGS) strategies based on Sanger reads forfeit the ability to assemble duplications with >97% identity or comprising more than 15 kb (ref. 8). Assemblies of shorter Illumina reads and sequencing by oligonucleotide ligation and detection (SOLiD) reads struggle to traverse smaller and more numerous genome-typical interspersed repeats[9]. Single-molecule sequencing technologies, such as PacBio and Nanopore sequencing, offer longer read lengths that can span most genome-typical repeats, but they lack the accuracy to assemble ampliconic sequences[10].

Although the extremely long and accurate reads produced by BAC sequencing are necessary to assemble ampliconic sequences, they are not sufficient. In the multihaplotype assembly of the human genome, amplicons have been misassembled or mistakenly skipped for being redundant[1,11–15]. Ampliconic sequences can differ from each other by <1 bp in 10,000, an order of magnitude less than the average difference between alleles[16]. Only these rare differences, which we call sequence family variants (SFVs), distinguish truly overlapping BACs from those that belong to paralogous ampliconic sequences. When a mix of haplotypes is used, the noise of frequent differences between alleles overwhelms the signal of rare SFVs. We developed our SHIMS approach, which is the only sequencing technology capable of assembling ampliconic regions, to cope with the ampliconic sequences of the human Y chromosome[17]. We deliberately restricted ourselves to BAC clones from one man's Y chromosome. Mapping and sequencing became coupled and iterative processes, so that sequencing an initial tiling

path of clones with substantial overlaps (preferably 30 kb, but always at least 10 kb) revealed SFVs that allowed us to refine our map and select BACs for subsequent rounds of sequencing. This painstaking approach produced a complete and accurate representation of the repeat architecture of a single Y chromosome[18], and as a result, we were able to predict and characterize rearrangements mediated by that architecture throughout the human population[19–26].These recurrent Y-chromosome rearrangements are the most common known genetic cause of spermatogenic failure in men and have been shown to play important roles in sex reversal, Turner syndrome, and testicular germ-cell tumors. Our SHIMS approach has been instrumental in producing the reference sequences of structurally complex sex chromosomes from several vertebrate species[1,27–32]. Here we describe how we advanced this technique to combine the advantages of a hierarchical clone-based strategy with new high-throughput sequencing technologies (**Fig. 1**). SHIMS 2.0 decreases both the time and the cost by two orders of magnitude while maintaining read length and accuracy[32].

Development

SHIMS 1.0 has been successful in generating reference sequence across several vertebrate sex chromosomes, and it remains the only technique capable of accessing structurally complex ampliconic regions. However, it relies on the Sanger-sequencing pipelines and the expertise of dedicated genome finishers at genome centers to assemble each BAC clone. This process is expensive and time consuming; each BAC clone costs ~$9,000 to sequence and assemble, and each iteration of mapping and sequencing takes ~6 months. We therefore sought to adopt new technologies to decrease costs, increase speed and efficiency, and bring SHIMS within the reach of a single laboratory (**Fig. 1**). SHIMS 2.0 takes advantage of the low cost and high consensus accuracy of Illumina reads to sequence indexed pools of hundreds of BAC clones. We streamlined and parallelized library production to bring sequencing costs down to $50 per BAC clone and to shorten mapping and sequencing iterations to a single week. In this strategy, in contrast to earlier BAC-pooling strategies[33–36], each clone is tagged with a unique barcode to assemble the reads from each clone separately. This process allows for pooling of clones from the same amplicon without endangering the integrity of the assembly of the entire pool. It is rare to encounter closely related interspersed repeats within a single BAC clone, and therefore long interspersed repeats typically do not confound the assembly of individual BAC clones from Illumina reads, as they do in WGS approaches. When internally repetitive clones are encountered, long reads from single-molecule sequencing technologies are used to scaffold the short-read assemblies, thus eliminating much of the need for manual finishing. Together, these optimizations allow a small independent laboratory to do work that once required a fully staffed genome center.

Overview of SHIMS 2.0

The most critical resource for a SHIMS project is a large-insert clone library derived from a single haplotype. Several academic and commercial suppliers[37] offer large-insert clone libraries from hundreds of species, but not all libraries are suitable for SHIMS projects. For sex chromosomes, in which ampliconic sequences are abundant, any library constructed from an individual of the heterogametic sex (males for an X and Y system; females for a Z and W system) contains a single haplotype for each sex chromosome (at half the coverage of the autosomes). For many model organisms, a library constructed from an inbred strain provides a single haplotype of the autosomes. For diploid organisms in which inbreeding is not possible, special sources of single-haplotype DNA are required. A single-haplotype BAC library has been constructed for the human genome, by using DNA from a hydatidiform mole, an abnormal conceptus that arises when an enucleated egg is fertilized by a single sperm bearing an X chromosome[12–15]. For some plant species, including many important cereal crops, haploid cell lines or haploid plants can be used as a source of single-haplotype DNA[38,39]. Ideally, the BAC library should have >10× coverage of the chromosome of interest; coverage <5× inevitably results in a fragmented assembly, owing to gaps in library coverage. When there is prior knowledge about the size of ampliconic repeat units, choosing a library with an average insert size smaller than the repeat unit of the amplicon is useful, because the presence of multiple amplicon copies within a single insert causes the clone assembly to collapse. For smaller ampliconic repeat units, fosmids, related cloning vectors that can accommodate inserts of ~40 kb, can substitute for BACs.

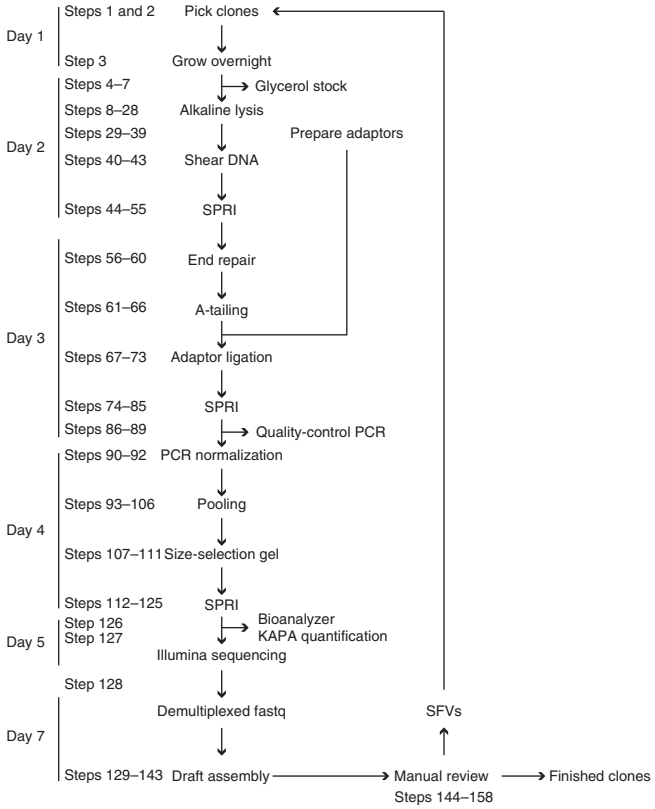


Figure 1 | Overview of the SHIMS 2.0 protocol. A timeline of a single iteration of the SHIMS 2.0 protocol, showing the major protocol steps, with key quality controls on the right. During a single week-long iteration, 192 clones are processed in parallel, and the resulting draft clone sequences are used to identify SFVs that distinguish paralogous ampliconic sequences. A single technician can proceed from a list of clones to completed Illumina libraries in 5 d. After a 2-d-long MiSeq run, a bioinformatics specialist assembles demultiplexed fastq sequences into draft clone assemblies and identifies SFVs to select clones for the next iteration.

Constructing or acquiring a copy of a BAC library can require a substantial investment. Library construction is relatively inexpensive (~$3,000), but picking and arraying the clones into 384-well plates is costly, and for a given insert size, this cost scales linearly with genome size and library depth. For a BAC library constructed from a typically sized mammalian genome of ~3 Gb at 10× depth, ~$10,000–$15,000 is typically required for a copy of an existing library, and ~$60,000–$70,000 is necessary to construct and array a new library. A set of high-density filters for screening may cost ~$3,000–$4,000. When sequencing a single chromosome, it is often more cost effective to order individual clones from an existing library than to order a complete copy. In general, approximately ten BAC clones are required to tile across per megabase. Costs for individual orders vary from $20 to $100 per clone; some suppliers offer discounted prices for larger orders.

After choosing or constructing a single-haplotype BAC library, the next step is to select an initial tiling path for sequencing and iterative refinement. A variety of mapping methods can be used to identify clones of interest from a BAC library, including fingerprint maps, end sequences, screening of high-density filters by hybridization, or screening of high-dimensional pools of BACs for sequence-tagged-site content by PCR. Typically, the cost to confirm each positive clone by another round of end-sequencing or PCR exceeds the cost to obtain a draft sequence with the SHIMS 2.0 protocol, thus making sequencing the most efficient way to confirm the identity of clones.

SFVs that distinguish between amplicon copies can be identified by the use of draft sequences from the initial tiling path of clones. We scrutinize the differences in the apparent overlaps between clones by using a graphical editor, such as Consed[6] or Gap5 (ref. 40). We typically limit ourselves to single-nucleotide differences supported by high-quality ($Q \geq 40$) bases in the majority of reads. Variants in short tandem repeats are not reliable; these are not always accurately assembled, and differences between clones often represent mutations that occur during propagation of the BACs in *Escherichia coli* rather than true differences between paralogous amplicons. After using newly identified SFVs as markers to refine the sequence map and resolve all paralogous amplicons, we order and orient the resulting sequence contigs through a complementary method, such as radiation hybrid (RH) mapping or metaphase fluorescence *in situ* hybridization (FISH). Whenever possible, we estimate the size of the remaining gaps, through either RH mapping or extended chromatin FISH.

Early attempts to adapt next-generation sequencing technologies for BAC assembly pooled many clones in a single sequencing run. Although this approach was faster and more cost effective than traditional Sanger sequencing, it had a major shortcoming in that assemblies either collapsed at genome-typical repeats shared among clones or, worse, contained chimeric contigs from two or more clones in the pool. Therefore, we added unique 'barcodes' or 'indexes' to each clone during library preparation[41,42] and pooled material from different clones only after these indexes were added. This procedure allows each read to be automatically assigned to a single clone, thus making the assembly of each clone less prone to artifacts. We typically use a set of Illumina TruSeq–compatible adaptors with 384 unique 8-nt indexes (**Supplementary Table 1**) and pool 192 clones in a single Illumina MiSeq run; however, in principle, this procedure could be extended to use larger numbers of indexes, or even dual indexes, to sequence larger pools of clones on higher-throughput sequencing machines while holding the level of coverage constant. For example, an Illumina NovaSeq instrument could sequence 38,400 clones in a single run, enough to provide more than 2× BAC coverage of a mammalian-sized genome.

A major challenge in pooling hundreds of samples is ensuring an adequate representation of each sample. When each sample in the pool has a widely varying concentration, some samples will have wasteful coverage that is greater than necessary for an adequate assembly, whereas others will have too few reads to generate any assembly, thus necessitating another round of library preparation to join the next pool. We use a short course of library amplification (20 cycles of PCR) with limiting primers. This procedure is sufficient to normalize each library within two- to threefold of the median concentration. Although PCR amplification can introduce errors and biases into Illumina libraries, we have found that the consensus-sequence accuracy does not suffer, and this procedure saves a large amount of tedious and exacting labor in measuring each library and diluting it to the same concentration.

Simple sequence repeats (SSRs) are the chief obstacle to contiguous BAC assemblies with short Illumina reads. Reads that cover SSRs are subject to stutter noise from replication slippage in library preparation (producing reads with inaccurate SSR-array length) and cluster generation (decreasing quality scores in the SSR)[43]. As a result, most assemblers fail to assemble SSRs, thus leaving gaps flanked by SSR sequence. We use a combination of long (300 bp) reads and large (1,000–1,200 bp) fragment sizes to scaffold over most (>99.8%) SSRs. One drawback to having fragment sizes greater than ~600 bp is that they must be size selected by gel purification to eliminate any smaller fragments. This gel purification would be onerous and expensive if each clone were purified separately, but our indexing and normalization procedures allow for a single size selection on a pool of libraries from hundreds of clones.

Advantages of SHIMS 2.0

SHIMS produces *de novo* sequence assemblies with greater accuracy and contiguity than provided by any other technique, and it is the only technique that has successfully produced accurate reference sequences of ampliconic regions. These advantages are rooted in the clone-based nature of SHIMS. Each clone assembly represents the highly accurate sequence of a single long molecule—with error rates as low as 1 in 1,000,000 nt. Any observed SFV can be verified by resequencing a clone of the same molecule, thereby increasing the confidence and resolution of the SFV map.

In contrast to WGS sequencing with short-read or even Sanger technologies, clone-based approaches produce a much more contiguous and accurate assembly[9]. Although genome-typical interspersed repeats, such as short interspersed nuclear elements, long interspersed nuclear elements, or endogenous retroviruses, are the primary obstacle to WGS assembly, they only rarely confound the assembly of individual clones. Furthermore, a hierarchical clone-based approach ensures that all sequence contigs are unambiguously mapped within a single clone in the assembly, and that clone is in turn mapped by long perfect overlaps with neighboring clones.

Continuous long-read technologies offer improvements in contiguity relative to short-read shotgun sequencing, but their accuracy (~85%)[44,45] is far too low to resolve ampliconic sequences, which are >99% identical. Error correction with short reads can improve the accuracy in single-copy regions[44,46,47], but this process tends to obscure SFVs by correcting long reads to the consensus of paralogous amplicons, thus resulting in collapsed assemblies. For example, the recent assembly of the gorilla Y chromosome with a mixture of Illumina and PacBio reads identified ampliconic sequences and estimated their copy number but could not resolve their structural organization[48].

Synthetic long-read technologies produce more accurate reads than do continuous long-read technologies, but they are not sufficiently accurate or long to assemble ampliconic sequences. Synthetic long reads have an error rate of ~1 in 10,000 nt (refs. 49,50), two orders of magnitude higher than that of clone-based approaches. Furthermore, synthetic long reads produced without cloning afford no opportunity to resequence the same molecule to resolve discrepancies between reads. Synthetic long reads are also one to two orders of magnitude shorter than BAC clones[49,50], thus limiting their power to resolve long ampliconic sequences that can differ by <1 in 10,000 nt (refs. 18,30,31).

Optical mapping techniques provide long-range scaffolding information that can help to increase the contiguity of genome assemblies by generating restriction maps of DNA fragments 0.1–1 Mb in size that can be compared against *in silico* restriction maps of WGS contigs[51]. In general, these restriction maps lack sufficient resolution to uncover the single-nucleotide differences that constitute reliable SFVs, and they do not sample molecules long enough to resolve many ampliconic sequences. Even when combined with PacBio and Illumina reads, optical mapping did not resolve the ampliconic sequences on the human Y chromosome[52].

Limitations of SHIMS 2.0

The SHIMS approach provides access to longer and higher-identity ampliconic sequences than any other sequencing technique, but the clone-based nature of this approach imposes several limitations. First, the maximum size of BAC inserts limits SHIMS to resolving duplications with <99.999% identity. Second, SHIMS is limited to sequences that can be cloned into *E. coli*. Third, SHIMS is confounded by repeated sequences shorter than a single clone.

The average BAC-clone size limits the power to resolve paralogous amplicons to those that differ by >1 nt in 100,000, so that each clone can be mapped by one or more SFVs. Clones with longer inserts, such as yeast artificial chromosomes (YACs), could potentially capture SFVs that distinguish paralogous amplicons at lower rates of divergence, but, in practice, YACs are subject to high rates of chimerism[53], deletion, and rearrangement[54], thus making them far too unreliable for sequencing ampliconic regions. This limitation will remain until long-read technologies are able to surpass BAC sequencing in both read length and accuracy, or a reliable cloning technology emerges that exceeds the insert size of BACs.

SHIMS is also limited to sequences that can be cloned. Sequences that are toxic to *E. coli* are underrepresented in BAC and fosmid libraries. An exhaustive search through the library will sometimes reveal deleted clones, when the cloning process has selected for clones with rearrangements that eliminate the toxic sequences. Gaps of this nature can be resolved through directed efforts that avoid cloning in *E. coli*. For example, orthologous ampliconic sequences on the human and chimpanzee Y chromosomes contain an unclonable sequence that has led to deleted clones in both human and chimpanzee BAC libraries. This ~30-kb unclonable region was eventually sequenced from a long-range PCR product[18,30].

Arrays of repeated sequences shorter than the clone insert size present special problems in clone-based sequencing. The centromere (171-bp repeat in 6-kb secondary unit), long-arm heterochromatin (degenerate pentamer repeat in 3.5-kb secondary unit), and TSPY gene array (20.4-kb unit) have not been resolved on the human Y chromosome[18]. Arrays with short repeat units may cause library gaps, because restriction sites appear either too frequently or not at all, such that no fragments covering the array are successfully cloned in the library. In that case, libraries produced by random shearing, such as fosmid libraries, may produce better coverage than those generated by restriction digests. The presence of many highly identical repeats within a single clone causes the sequence assembly to collapse multiple repeats into a single short contig. Whenever possible, it is best to choose clone libraries with an insert size that matches the expected repeat-unit size. However, sequencing with fosmids, as compared with BACs, is more expensive because many more clones are required to cover the same amount of sequence. In some cases, continuous long-read technologies applied to individual BAC clones can resolve internal repeats, albeit at higher costs and with lower per-base accuracy.

Applications of SHIMS 2.0

SHIMS has been repeatedly applied to resolve ampliconic sequences across vertebrate genomes, particularly the sex chromosomes, which contain the most abundant and elaborate ampliconic sequences. SHIMS has been used to resolve the ampliconic sequences of the human, chimpanzee, and mouse Y chromosomes; the human X chromosome; and the chicken Z and W chromosomes[1,17,18,27,28,30,31]. SHIMS has also been applied to the human immunoglobulin gene cluster[14] and other structurally complex regions on human autosomes[12,13,15], by using a single haplotype library derived from a hydatidiform mole.

The SHIMS approach is applicable to any genomic sequence for which amplicons and other structurally complex regions complicate the WGS assembly. A library of large-insert clones from a haploid or inbred diploid source is required to resolve ampliconic sequences, but a library derived from an outbred diploid source could be used to generate a phased diploid genome assembly covering nonampliconic regions.

Existing instruments (Illumina HiSeq series) already generate sufficient numbers of reads to sequence a tiling path of BAC clones across the human genome in a single run costing ~$14,000 (ref. 55); the costs of sample preparation and library generation therefore dominate cost considerations. With our current SHIMS 2.0 approach, assembling the entire human genome would cost ~$2,000,000, three orders of magnitude less than the cost of generating the original BAC-based reference sequence. The cost could be decreased further with future extensions of the indexing and pooling strategy described here to reduce the reagent costs and labor required for sample preparation and library generation.

Thus, SHIMS 2.0 could potentially be cost effective to apply across whole genomes, even in the absence of the extensive resources, such as BAC fingerprint maps and end sequences, available for common model organisms.

Experimental design

The primary time and cost savings of SHIMS 2.0 over traditional BAC sequencing come from its ability to process many clones in parallel and sequence them in a single pool. Although each step can be performed by hand with a multichannel pipette, all operations, particularly size selection with solid-phase reversible immobilization (SPRI) beads (Steps 44–55 and 74–85), are more accurate with a liquid-handling robot. We use the Zephyr NGS Workstation manufactured by PerkinElmer. A liquid-handling robot need not be expensive—adequate used instruments can be purchased for <$5,000, and many core facilities offer access to a liquid-handling robot.

We have made several optimizations to adapt standard DNA-extraction and library-preparation techniques[56] for our purposes. To support the growth of *E. coli* carrying single-copy BACs, clones are grown in 2× Luria broth (LB) (Step 1). SPRI beads are added directly to the isopropanol-precipitation step to recover more DNA than is yielded by a standard alkaline lysis preparation (Steps 17–27). Crude preparations of low-copy plasmids, such as BACs or fosmids, contain 10–25% *E. coli* genomic DNA contamination; sequencing costs are sufficiently low that special measures to decrease this contamination fraction any further are not cost effective.

In our experience, the smallest fragments present in the library determine the average sequenced fragment size. We have found that a Covaris Focused-ultrasonicator is invaluable for the generation of DNA fragments with a reproducibly tight size distribution as input for the library-generation protocol (Steps 40–43). We have optimized our shearing conditions for a Covaris LE220 Focused-ultrasonicator; other makes and models may require slightly different conditions to achieve 1-kb fragments. The Covaris 96 microTUBE plates (Step 42) are costly but necessary for consistent shearing across wells. After library generation is complete, a gel-based size selection assures the tightest size distribution around 1 kb (Steps 107–111). Because each sample is individually barcoded, all samples can be pooled in a single well before size selection, thus drastically decreasing the labor involved (Steps 93–106).

We use a custom set of 384 octamer indexes for barcodes (**Supplementary Table 1**); Illumina offers sets of 96 and 384 barcodes through a dual-indexing scheme. More elaborate dual-index schemes[57] could potentially allow for larger pools on higher-throughput Illumina machines. We selected the MiSeq because of its combination of long reads, short run times, and low cost, and we believe that it offers the right combination of features for a single lab to perform SHIMS 2.0 on targeted genomic regions. Scaling up to higher-throughput instruments, such as the Illumina HiSeq or NovaSeq, is certainly possible for genome-scale sequencing projects, but sequencing-reagent costs would decrease by only a modest amount, because the bulk of the cost is in the library preparation.

A SHIMS 2.0 project requires substantial bioinformatics expertise to proceed from raw reads to finished annotated sequence. State-of-the-art software advances rapidly, such that specific software recommendations are likely to become outdated very quickly. Demultiplexed fastq-format files should first be trimmed to remove Illumina adaptor sequences and low-quality bases, then be screened for contamination from the host genome and vector sequences. Filtered sequences are then used for assembly, scaffolding, and gap closure. We use cutadapt[58] to trim adaptors and low-quality sequences (Step 130), bowtie2 (ref. 59) to screen out vector and *E. coli* genomic-DNA contamination (Steps 131 and 133), SPAdes[60] for assembly (Step 136), BESST[61] for scaffolding (Step 138), and Gap2Seq (ref. 62) to fill gaps (Step 139). Some clone assemblies require manual finishing; we rely on Consed[6] for visualizing discrepant bases, separating collapsed duplications, and merging overlapping contigs (Steps 144–158).

Several controls ensure the accuracy and quality of a SHIMS 2.0 assembly. A cell line derived from the same individual or strain as the BAC or fosmid library allows for FISH experiments and long-range PCR. An independent FISH, RH map, or optical map can be used to confirm the order and orientation of contigs in the clone map, as well as to estimate the size of any remaining gaps. Finally, the error rate of the assembly can be calculated from the number of discrepancies observed in the long (>10 kb) redundant overlaps between adjacent clones; for BACs, it is possible to achieve error rates as low as 1 in 1,000,000 nt.

There are important quality-control checkpoints in both the library-preparation and the bioinformatic-analysis stages. After adaptor ligation, but before pooling, it is useful to reserve a sample from each clone's library for 40 cycles of PCR with universal Illumina primers and subsequent gel electrophoresis, to ensure that each library contains PCR-amplifiable material in the expected size range, before proceeding with sequencing (Steps 86–69) (**Fig. 1**). After pooling and gel purification, we recommend testing the fragment-size distribution with a Bioanalyzer before sequencing (Step 126) (**Fig. 1**). The front of the fragment-size distribution will be the peak of the sequenced fragment-size distribution. During assembly, reads from *E. coli* genomic DNA serve as an internal control to assess library-insert size and sequencing error rates (Steps 132–135). As each clone is assembled, putative overlapping clones are aligned to identify the differences with Consed[6] and to verify that the reads in each clone support the difference (Steps 145–147). Most of these high-quality differences will be SFVs that distinguish paralogous amplicons, but some will be genuine errors due to mutations in the BAC or fosmids.

Because the structures of ampliconic regions are often unclear until a nearly complete tiling path is assembled, seeding the first iteration with as many clones as possible is preferable, to take full advantage of the high throughput of SHIMS 2.0 to identify SFVs early and decrease the total number of iterations. In our experience, even beginning with a near-complete tiling path in mostly single-copy sequence, at least three iterations are still required to create a contiguous chromosome sequence. In subsequent iterations, it is best to select several of the clones that may potentially prolong a contig or close a gap, favoring those that seem to add the most new sequence. Library gaps will become apparent in later iterations, as repeated attempts to extend a contig fail to identify new clones or identify only artifacts such as deleted, chimeric, or mislabeled clones. Because draft clone assemblies are accurate enough to identify SFVs, redundant clones can be abandoned at the draft stage. Ideally, each clone in the final tiling path should be 'finished' into a contiguous sequence, with all sequences ordered

and oriented, all gaps closed, and any ambiguities (e.g., SSRs) marked as unresolved (Steps 151–159).

A major benefit of this protocol is that it allows a small team to carry out a SHIMS project that, only a few years ago, would have required the cooperation of a large genome center. A single technician can process 192 clones from frozen library plates to Illumina libraries in a single week, and a bioinformatics specialist can set up a pipeline to automatically process the resulting reads into draft assemblies, identify SFVs, and manually finish complex clone assemblies. It is particularly helpful to have a team member or collaborator with experience in metaphase FISH to help resolve the order and orientation of contigs within the final assembly.

MATERIALS

REAGENTS

General reagents

- 10× ligase buffer (New England Biolabs, cat. no. B0202S)
- 100 µM barcode adaptor (Integrated DNA Technologies)
- Isopropanol (VWR, cat. no. BDH1133-1LP) ! CAUTION Isopropanol is flammable and an irritant. Keep away from flame, and wear gloves and eye protection when handling it.
- Taq DNA polymerase with standard Taq buffer (New England Biolabs, cat. no. M0273L)
- T4 polynucleotide kinase (New England Biolabs, cat. no. M0201L)
- End Repair Module (New England Biolabs, cat. no. E6050L)
- Klenow fragment (3′→5′ exo⁻) (New England Biolabs, cat. no. M0212L)
- T4 DNA ligase (New England Biolabs, cat. no. M0202L)
- 100 mM dATP (New England Biolabs, cat. no. N0440S)
- 100-bp DNA ladder (New England Biolabs, cat. no. N3231S)
- Rediload dye (Invitrogen, cat. no. 750026)
- SeaKem ME agarose (Lonza, cat. no. 50014)
- 5× Phusion buffer (New England Biolabs, cat. no. M0530L)
- Phusion enzyme (2 U/µl; New England Biolabs, cat. no. M0530L)
- Thermo Scientific dNTP set (Thermo Fisher Scientific, cat. no. R0186)
- Library Quantification Kit–Illumina/ABI Prism (Kapa Biosystems, cat. no. K4835)
- MiSeq Reagent Kit v3 (Illumina, cat. no. MS-102-2023)
- Seal-Rite 1.5-ml microcentrifuge tubes (USA Scientific, cat. no. 1615-5500)
- Glycerol (EMD Millipore, cat. no. 356350-1000ML)
- RNase A (17,500 U; Qiagen, cat. no. 19101)
- E-Gel SizeSelect agarose gels (2%; Invitrogen, cat. no. G661002)
- GE Healthcare Sera-Mag SpeedBeads carboxyl magnetic beads, hydrophobic (Thermo Fisher Scientific, cat. no. 09981123)
- Sodium hydroxide (NaOH; AmericanBio, cat. no. AB01916-050000) ! CAUTION NaOH is corrosive. Wear gloves and eye protection when handling it.
- PEG 8000 (Sigma-Aldrich, cat. no. P2139-2KG)
- Chloramphenicol (Sigma-Aldrich, cat. no. C0378-5G) ! CAUTION Chloramphenicol powder is hazardous. Handle this reagent in a ventilated fume hood with gloves and eye protection.
- RNase A (17,500 U; Qiagen, cat. no. 19101)
- Ethylenediaminetetraacetate, disodium, dihydrate ($Na_2EDTA{\cdot}2H_2O$; AmericanBio, cat. no. AB00500-01000) ! CAUTION Irritant. Wear gloves and eye protection when handling this reagent.
- Bacto-tryptone (BD Biosciences, cat. no. 211705)
- Yeast extract (BD Biosciences, cat. no. 211929)
- Sodium chloride (NaCl; AmericanBio, cat. no. B01915-10000)
- Sodium dodecyl sulfate (SDS; Sigma-Aldrich, cat. no. L3771-100G) ! CAUTION SDS powder is hazardous. Handle this reagent in a ventilated fume hood with gloves and eye protection.
- Potassium acetate (KOAc; Sigma-Aldrich, cat. no. P1190-100G)
- Tris base (AmericanBio, cat. no. AB020000-05000)
- Acetic acid, glacial (EMD Millipore, cat. no. AX0073-9) ! CAUTION Glacial acetic acid is corrosive. Wear gloves and eye protection when handling it.
- Ethyl alcohol, 200 proof (ethanol; Pharmco-Aaper, cat. no. 11000200) ! CAUTION Ethyl alcohol is flammable; keep away from flame when handling it.
- Hydrochloric acid (HCl; VWR, cat. no. BDH3026-500MLP) ! CAUTION HCl is corrosive. Wear gloves and eye protection when handling it.
- Tris-EDTA buffer (TE buffer; VWR, cat. no. E112-500ML)

Custom oligonucleotides

▲ CRITICAL All oligonucleotides (oligos) were ordered from the commercial vendor Integrated DNA Technologies (https://www.idtdna.com/site/) with standard desalting ('Reagent Setup' section)

- Solexa primer 1.0 (25 nmol; Integrated DNA Technologies)
- Solexa primer 2.0 (25 nmol; Integrated DNA Technologies)

Adaptor oligos

▲ CRITICAL Adaptor oligos are annealed to form a Y-shaped adaptor ('Prepare Barcoded Adaptor' section)

- Universal adaptor oligo (250 nmol; Integrated DNA Technologies)
- Barcode adaptor oligos (25 nmol; Integrated DNA Technologies)

EQUIPMENT

- AirPore tape sheets (Qiagen, cat. no. 19571)
- Eppendorf Twin.tec plates (USA Scientific, cat. no. 4095-2624Q)
- Nunc 96 DeepWell plates (Thermo Fisher Scientific, cat. no. 278743)
- TempPlate Semi-Skirt PCR plates (USA Scientific, cat. no. 1402-9700)
- TempPlate XP PCR sealing film (USA Scientific, cat. no. 2972-2100)
- Adhesive PCR plate seals (Thermo Fisher Scientific, cat. no. AB0558)
- 96 MicroTUBE plates (Covaris, cat. no. 520078) ▲ CRITICAL We have found that these Covaris plates are essential to shear DNA evenly across all samples.
- Costar assay plates, 96-well (Corning, cat. no. 3797)
- 12-strip 0.2-ml PCR tubes (Neptune, cat. no. 3426.12.X)
- MicroAmp fast optical 96-well reaction plates (Thermo Fisher Scientific, cat. no. 4346906)
- MicroAmp optical adhesive film (Thermo Fisher Scientific, cat. no. 431197)
- Aluminum adhesive foil (Bio-Rad, cat. no. MSF1001)
- Costar 50-ml reagent reservoirs (Corning, cat. no. 4870)
- 96-well-format plate magnet (Alpaqua, cat. no. 003011)
- 2100 Bioanalyzer (Agilent Technologies, cat. no. G2938A)
- 7500 Fast Real-Time PCR System (Thermo Fisher Scientific, cat. no. 4351107)
- NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific, cat. no. ND 1000) or Qubit 4 fluorometer (Thermo Fisher Scientific, cat. no. Q33226)
- SimpliAmp Thermal Cycler (Applied Biosystems, cat. no. A24811)
- Centrifuge 5810 R (Eppendorf, cat. no. 00267023)
- New Brunswick Innova 2300 shaker (Eppendorf, cat. no. M1191-0022)
- E-Gel Precast Agarose Electrophoresis System (Thermo Fisher Scientific, cat. no. G6465)
- DynaMag-2 magnet (Thermo Fisher Scientific, cat. no. 12321D)
- Vortex-Genie 2 (Scientific Industries, cat. no. SI-0236)
- LE220 Focused-ultrasonicator (Covaris, cat. no. 500219)
- Zephyr NGS Workstation (PerkinElmer, cat. no. 133750)
- ErgoOne pipette, 10–100 µl, 12 channel (USA Scientific, cat. no. 7108-1100)
- ErgoOne pipette, 30–300 µl, 12 channel (USA Scientific, cat. no. 7112-3300)
- VWR Signature multichannel electronic pipettor (VWR, cat. no. 89000-674)
- VWR aerosol filter pipet tips, sterile, 1,000 µl (VWR, cat. no. 89003-060)
- VWR aerosol filter pipet tips, sterile, 200 µl (VWR, cat. no. 89003-056)
- Kimble graduated cylinders, 250 ml (Fisher Scientific, cat. no. 08-548-205)
- Kimble graduated cylinders, 1 l (Fisher Scientific, cat. no. 08-548-207)
- Parafilm M film (Bemis, PM996)
- Pyrex media bottles, graduated, 1 l (VWR, cat. no. 1395-1L)
- VWR laboratory bottle, polypropylene, wide mouth (VWR, cat. no. 414004-127)
- Falcon centrifuge tubes, polypropylene, sterile (VWR, cat. no. 21008-951)
- AMSCO 2021 gravity sterilizer (Amsco, cat. no. R78ES-2)

- VWR standard magnetic stirrer, 120 V (VWR, cat. no. 97042-670)
- ALPS 25 manual heat sealer, 100 V (Thermo Scientific, cat. no. AB-0384-110)
- ThermoMixer temperature control device (Eppendorf, 5382000023)
- Thermolyne 16500 Dri-Bath (Thermolyne, cat. no. TH-16500) (optional)

Software

- cutadapt (https://github.com/marcelm/cutadapt/)
- flash (https://ccb.jhu.edu/software/FLASH/)
- bowtie2 (https://github.com/BenLangmead/bowtie2/)
- SPAdes (http://cab.spbu.ru/software/spades/)
- samtools (https://github.com/samtools/samtools/)
- BESST (https://github.com/ksahlin/BESST/)
- Gap2Seq (https://www.cs.helsinki.fi/u/lmsalmel/Gap2Seq/)
- Consed (http://www.phrap.org/consed/consed.html)
- BLAST+ (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)
- Covaris SonoLab v7.3 control software

REAGENT SETUP

70% (vol/vol) ethanol Mix 30 ml of 100% (vol/vol) ethanol with 70 ml of distilled, deionized water (ddH_2O). ▲ **CRITICAL** 70% (vol/vol) ethanol should be prepared on the day of the experiment.

1 N NaOH Dissolve 40 g of NaOH in 1 l of ddH_2O. 1 N NaOH can be prepared in advance and stored at room temperature (22 °C) in an airtight, dark plastic bottle for up to 3 months.

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

10 mM Tris-Cl, pH 8.5 Mix 0.5 ml of 1 M Tris-Cl with 49.5 ml of ddH_2O. This solution can be prepared in advance and stored at room temperature for up to 1 year.

80% (vol/vol) glycerol solution Add 400 ml of glycerol to a graduated cylinder; adjust the volume to 500 ml with ddH_2O. Seal the cylinder with Parafilm M film, and mix by inversion. Transfer to a bottle, and autoclave for 20 min in liquid cycle. This solution can be prepared in advance and stored at room temperature for up to 1 year.

Solution 1 Dissolve 6.06 g of Tris base and 3.72 g of $Na_2EDTA{\cdot}2H_2O$ in 800 ml of ddH_2O. Adjust the pH to 8.0 with concentrated HCl, and then adjust the volume to 1 l with ddH2O. Add 100 mg of RNase A to the solution. Solution 1 can be prepared in advance and stored at 4 °C for up to 1 year. ▲ **CRITICAL** Add fresh RNase A after 6 months of storage.

Solution 2 Dissolve 8 g of NaOH in 950 ml of ddH_2O. Add 30 ml of 20% (wt/vol) SDS solution. ▲ **CRITICAL** Solution 2 should be prepared on the day of the experiment.

Solution 3 Dissolve 294.5 g of potassium acetate in 500 ml of ddH_2O. Adjust the pH to 5.0 with glacial acetic acid. Adjust the volume to 1 l with ddH_2O. Solution 3 can be prepared in advance and stored at 4 °C for up to 1 year.

5 M NaCl Dissolve 292 g of NaCl in 800 ml of ddH_2O. Adjust the volume to 1 l with ddH_2O. 5 M NaCl can be prepared in advance and stored at room temperature for up to 1 year.

SPRI beads Add 135 g of PEG 8000 powder to a 1-L bottle. Add 150 ml of 5 M NaCl, 7.5 ml of Tris-HCl, 1.5 ml of 0.5 M EDTA, and 450 ml of ddH_2O to make PEG buffer. Resuspend the stock solution of Sera-Mag beads by vortexing. Transfer 15 ml of Sera-Mag beads to a conical tube. Pellet the beads in a magnetic rack (e.g. DynaMag-2). Remove the storage buffer, and wash the beads twice with 20 ml of TE buffer. Resuspend the beads in 25 ml of ddH_2O and add to the PEG buffer. Wash the conical tube with another 25 ml of ddH_2O and add the wash to the PEG buffer. Wrap the bottle in aluminum foil to protect the solution from light. ▲ **CRITICAL** Store at 4 °C, protected from light, for up to 1 year. Bring the solution to room temperature and mix well before use.

0.5 M EDTA Add 186.1 g of $Na_2EDTA{\cdot}2H_2O$ to 800 ml of ddH_2O. Stir vigorously on a magnetic stirrer. Adjust the pH to 8.0 with NaOH. 0.5 M EDTA solution can be stored at room temperature for up to 1 year.

2× LB Add 20 g of bacto-tryptone, 10 g of yeast extract, and 20 g of NaCl to ddH_2O, and adjust the volume to 1 l. Mix well with a magnetic stirrer. After mixing, distribute 500-ml aliquots into 1-L bottles. Cap loosely, prewarm to 50 °C, and autoclave for 20 min on liquid cycle. Store at room temperature for up to 1 year.

Chloramphenicol Dissolve 0.34 g of chloramphenicol into 10 ml of 100% (vol/vol) ethanol. Chloramphenicol stock can be stored at −20 °C for up to 1 year.

20% (wt/vol) SDS Dissolve 200 g of SDS into 800 ml of ddH_2O by stirring. After mixing, adjust the volume to 1 l with ddH_2O. 20% (wt/vol) SDS can be prepared in advance and stored at room temperature for up to 1 year.

Solexa P1 oligo (10 µM) The sequence of this oligo is 5′-AATGATAC GGCGACCACCGAGATCTACACTCTTTCCCTACACGA-3′. Add 2.5 ml of TE buffer to 25 nmol of dry, lyophilized oligo to make 10 µM Solexa P1 oligo stock solution. Solexa P1 oligo can be stored at 4 °C for up to 3 months. For long-term storage, store Solexa P1 oligo at −20 °C for up to 1 year.

Solexa P2 oligo (10 µM) The sequence of this oligo is 5′-CAAGCAGA AGACGGCATACGAGAT-3′. Add 2.5 ml of TE buffer to 25 nmol of dry, lyophilized oligo to make 10 µM Solexa P2 oligo stock solution. Solexa P2 oligo can be stored at 4 °C for up to 3 months. For long-term storage, store Solexa P2 oligo at −20 °C for up to 1 year.

Universal adaptor oligo (100 µM) The sequence of this adaptor is 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACA CGACGCTCTTCCGATCT-3′. Add 2.5 ml of TE buffer to 250 nmol of dry, lyophilized oligo to make 100 µM universal adaptor oligo stock solution. Universal adaptor oligo can be stored at 4 °C for up to 3 months. For long-term storage, store universal adaptor oligo at −20 °C for up to 1 year.

Barcode adaptor oligo (100 µM) The sequence of this adaptor is 5′- ATCG GAAGAGCACACGTCTGAACTCCAGTCAC[NNNNNNNN]ATCTCGTATG CCGTCTTCTGC-3′. Add 250 µl of TE buffer to each 25 nmol of dry, lyophilized oligo to make 100 µM barcode adaptor oligo stock solution. Barcoded adaptor oligo can be stored at −20 °C for up to 1 year.

PROCEDURE

Pick clones and grow cultures ● TIMING 18 h

1 | Fill each well of a Nunc 96 DeepWell plate with 1.9 ml of 2× LB containing 34 µg/ml chloramphenicol.
▲ **CRITICAL STEP** Rich medium (2× LB) is appropriate for single-copy plasmids such as BACs or fosmids, which use chloramphenicol resistance as a selectable marker. For high-copy-number plasmids, standard 1× LB will prevent overgrowth. Use the appropriate antibiotic to select other plasmids.

2 | Use a clean pipette tip to scrape the surface of a frozen glycerol stock, and drop the tip directly into the DeepWell plate to inoculate a well.

3 | Seal the plates with AirPore tape sheets, and incubate them at 37 °C for 16–17 h, shaking at 220 r.p.m.
▲ **CRITICAL STEP** Overgrowth of cultures (cell density >3 × 10^9 or 4 × 10^9 cells per ml) will decrease the yield of BAC DNA.

Glycerol stock plate ● TIMING 30 min

4 | Dispense 150 µl of 80% (vol/vol) glycerol solution into each well of a Costar assay plate.

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

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

7 | Store the glycerol stock plate at –80 °C. Glycerol stock plates can be stored indefinitely.

Alkaline lysis ● TIMING 2–3 h

8 | Cover the 96 DeepWell plate from Step 3 with adhesive PCR plate seal. Pellet the BAC cultures by centrifugation for 20 min at 2,500*g* at 4 °C.

9 | Peel back the seal quickly, and invert each DeepWell Plate over a waste container to dispose of the spent medium. Tap the DeepWell Plate firmly on a paper towel to remove any remaining droplets.

10 | Use a 12-channel pipettor with large fill volume to add 0.2 ml of Solution 1 to each well of the 96 DeepWell Plate.

11 | Apply an adhesive PCR plate seal to each DeepWell Plate, and resuspend the bacterial pellets by vortexing (Vortex-Genie 2). Incubate at room temperature for 30 min.

12 | Add 0.2 ml of Solution 2 to each well, apply a fresh seal, mix gently but thoroughly by inverting ten times, and incubate at room temperature for 5 min.
▲ **CRITICAL STEP** Do not vortex the lysates at this stage, to avoid shearing of the bacterial genomic DNA. Do not incubate for more than 5 min.

13 | Add 0.2 ml of Solution 3 to each well, apply a fresh seal, and mix immediately by inverting 20 times.
▲ **CRITICAL STEP** Thorough mixing ensures uniform precipitation.

14 | Centrifuge for 20 min at 6,000*g* at 4 °C.

15 | Transfer 480 µl of supernatant from Step 14 to a new 96 DeepWell plate. Use a multichannel pipettor with a sufficiently large fill volume (e.g., VWR Signature). Most of the precipitated material will stick to the walls of the culture DeepWell plate. Discard the old plate.
▲ **CRITICAL STEP** Avoid transferring cell debris to the new plate.

16 | Add 350 µl of isopropanol to each well.

17 | Add 50 µl of SPRI beads to each well. Mix ten times by pipetting up and down with a multichannel pipettor.

18 | Place the plate on a 96-well-format plate magnet until the wells are clear, for ~5–10 min.
▲ **CRITICAL STEP** Keep the plate on the magnet through Steps 19–23.

19 | Quickly invert the plate over a waste container to remove and discard the supernatant.

20 | Keeping the plate on the magnet, use fresh pipette tips to add 1,000 µl of 70% (vol/vol) ethanol. Incubate the plate on the magnet for 30 s.
▲ **CRITICAL STEP** Gently touch the side of the well, and dispense ethanol into wells without disturbing the beads.

21 | Quickly invert the plate over a waste container to remove and discard the ethanol without disturbing the beads.

22 | Repeat Steps 20 and 21.

23 | Tap the DeepWell plate, still on the magnet, firmly on a paper towel to remove any remaining droplets.

24 | Remove the plate from the magnet. Let the sample plate air dry for 10 min, or, for faster drying, place on a 37 °C heating block for 3 min.
▲ **CRITICAL STEP** Make sure that all ethanol is evaporated, but do not overdry the beads, to avoid decreasing the efficiency of DNA elution.

25 | Add 140 µl of 10 mM Tris-HCl to each well to elute the DNA from the beads.

26 | Mix 15 times by pipetting up and down. Place the plate back onto the magnet until the wells are clear, for ~5 min.

27 | Aspirate 135 µl of the resuspended DNA into a new Eppendorf Twin.tec plate. Avoid aspirating any beads.

28 | Label and seal the plate with a heat sealer, and store at –20 °C.
■ **PAUSE POINT** Store at –20 °C for up to 1 week.

Prepare barcoded adaptor ● TIMING 4 h
29 | For 384 barcoded adaptors, prepare a master mix for 400 reactions according to the table below, and add 35 µl of master mix to each well of a PCR plate.
▲ **CRITICAL STEP** Set up all reactions on ice, and keep on ice at all times between steps.

Reagents	Volume per reaction (µl)	Final concentration (after Step 30)
NEB 10× ligase buffer supplemented with 10 mM ATP	4.25	1×
T4 polynucleotide kinase	1	10 U/42.5 µl
ddH_2O	29.75	

30 | Add 7.5 µl of each 100 µM barcode adaptor to the respective well on the PCR plate.

31 | Seal the PCR plate with an adhesive PCR plate seal.

32 | Place the PCR plate in a thermocycler (e.g., SimpliAmp Thermal Cycler), and run the following program.

Cycle number	Temperature (°C)	Time
1	37	60 min
2	4	Hold

33 | Transfer the plate from the thermocycler to the centrifuge on ice. Centrifuge the plate at 1,500*g* for 3 min at 4 °C.

34 | Add 7.5 µl of 100 µM universal adaptor to each well. Mix by pipetting up and down ten times.

35 | Incubate the adaptor mixture in a thermocycler with a heated lid at 105 °C to anneal the barcode adaptors to the universal adaptors by using the following program.

Cycle number	Temperature (°C)	Time
1	98	5 min
2–95	–1/min	1 min
96	4	Hold

36 | Remove the plate from the thermocycler, and place on ice for transfer to a centrifuge.

37 | Centrifuge the plate at 1,700*g* for 3 min at 4°C.

38 | Add 50 µl of 10 mM Tris-HCl to each well for a total volume of 100 µl and a final concentration of 7.5 µM. Mix by pipetting up and down ten times.

39 | Dispense 20-µl aliquots of each 100-µl adaptor mixture into five PCR plates. Clearly label each plate, and add the date. Store at –20°C for up to 6 months.
▲ **CRITICAL STEP** Aliquots of adaptors are divided among five plates to minimize the number of freeze–thaw cycles for each plate.

DNA shearing ● TIMING 1 h
40 | Fill the Covaris ultrasonicator water bath to water level 10, as marked on the water-bath container. Degas the water for ~45 min before shearing according to the manufacturer's instructions.

41 | Load a 12-channel pipettor (ErgoOne, 30–300 µl) with clean tips, and use them to prepierce the foil on each well of the Covaris 96 microTUBE plate for easier liquid transfer.

42 | Transfer 130 µl of the DNA from Step 28 to the Covaris 96 microTUBE plate, and seal the plate.
▲ **CRITICAL STEP** Take care when dispensing samples into wells. Gently touch the tips to the bottom of the well, and dispense the samples slowly. Double-check the tips when removing them from the wells, because acoustic fibers sometimes stick to the sides of the tips.

43 | Shear the DNA to 1,200 bp in glass Covaris tubes by using the following settings:

Parameter	Setting
Duty cycle	5%
Peak incident power	450
Cycle per burst	200
Time	25 s

▲ **CRITICAL STEP** These settings are optimized for Covaris SonoLab v7.3 control software; different versions may require adjustments to the cycle time.
? TROUBLESHOOTING

SPRI cleanup ● TIMING 1 h
44 | Transfer 100 µl of the 130-µl sheared sample to a new Eppendorf Twin.tec 96-well plate, and discard the Covaris microTUBE plate. SPRI cleanup can be performed manually or with a liquid-handling robot (e.g., Zephyr NGS Workstation).

45 | Add 60.7 µl of SPRI beads to each well, and thoroughly mix by pipetting the mixture up and down. Incubate for 5 min at room temperature.
▲ **CRITICAL STEP** Aspirate the SPRI beads slowly. The SPRI solution is very viscous. A slow pipetting speed helps to ensure that accurate volumes are dispensed and decreases the formation of air bubbles, which markedly decrease the yield. Ensure that DNA–bead mixtures are thoroughly mixed. Check the pipette tips at every step to ensure that the volume is accurate. Pipette tips are particularly prone to retaining extra drops of solution inside or at the point.

46 | Place the plate on a 96-well-format plate magnet until the wells are clear, for ~5 min.

47 | Remove and discard the supernatant, which now contains most fragments smaller than 1,000 bp.

48 | Keeping the plate on the magnet, use fresh pipette tips to add 200 µl of 70% (vol/vol) ethanol.
▲ **CRITICAL STEP** Always use freshly prepared 70% (vol/vol) ethanol for SPRI cleanup. Gently touch the side of the well, and dispense ethanol into wells without disturbing the beads.

49 | Incubate the plate on the magnet for 30 s.

50 | Remove and discard the ethanol without disturbing the beads.

51 | Repeat Steps 48–50.

52 | Use a pipette to remove any drops of ethanol remaining in each well.

53 | Remove the plate from the magnet. Let the sample plate air dry for 10 min, or, for faster drying, place on a 37 °C heating block.

54 | Add 20 µl of 10 mM Tris-HCl to each well to elute the DNA. Mix 15 times by pipetting, and place the plate back on the magnet.

55 | Carefully aspirate 17 µl of the resuspended DNA fragments into a new PCR plate, making sure not to aspirate any beads. Discard the old plate.

? TROUBLESHOOTING

■ **PAUSE POINT** Store at –20 °C for up to a week before library construction.

Library construction ● **TIMING 3–4 h**

56 | Thaw the plate from Step 55, and centrifuge for 3 min at 300*g* at 4 °C.

57 | Place a 12-strip of PCR tubes on ice to cool. Prepare the End Repair Module master mix, following the table below, for 110 reactions (96 reactions plus 14 extra to account for pipetting error), and dispense 27 µl of master mix into each tube.

Reagents	Volume per reaction (µl)	Final concentration
End-repair buffer	2	1×
End-repair enzyme mix	1	1×

58 | With a 12-channel pipettor (ErgoOne, 10–100 µl), add 3 µl of master mix to each well of the thawed plate from Step 56, and mix each sample ten times by pipetting up and down.

59 | Seal the PCR plate, vortex, and then centrifuge at 1,700*g* for 3 min at 4 °C.

60 | Incubate the plate in a thermocycler with a heated lid (100 °C), using the following program:

Cycle number	Temperature (°C)	Time
1	20	25 min
2	75	15 min
3	4	Hold

61 | Centrifuge the plate at 1,700*g* for 3 min at 4 °C.

62 | Prepare A-Base Addition master mix for 110 reactions according to the table below. Flick to mix and centrifuge briefly.

Reagents	Volume per reaction (µl)	Final concentration
ddH_2O	1	
100 mM dATP	0.5	2.27 mM
Klenow fragment (3′→5′ exo–)	0.5	0.11 U/µl

63 | Dispense 18 µl of the master mix into each tube of a 12-strip PCR tube on ice.

64 | With a 12-channel pipette, add 2 µl of master mix to each well of the plate, and mix each sample ten times by pipetting up and down.

65 | Seal the PCR plate. Quickly vortex the plate, and centrifuge at 1,700*g* for 3 min at 4 °C.

66 | Incubate the plate in a thermocycler with a heated lid (100 °C), using the following program:

Cycle number	Temperature (°C)	Time
1	37	25 min
2	75	15 min
3	4	Hold

67 | Centrifuge the plate from Step 66 at 1,700*g* for 3 min at 4 °C.

68 | Prepare Adaptor Ligation master mix for 110 reactions by following the table below. Flick to mix and centrifuge briefly.

Reagents	Volume per reaction (µl)	Final concentration
10× ligase buffer supplemented with 10 mM ATP	3	1×
ddH_2O	1	
T4 DNA ligase	1	13 U/µl

69 | Dispense 45 µl of the master mix into each tube of a 12-strip PCR tube on ice.

70 | With a 12-channel pipette, add 5 µl of master mix to each well of the plate from Step 67, and mix each sample ten times by pipetting up and down.

71 | Add 3 µl of annealed adaptor from Step 39 to the respective well. Seal the PCR plate.

72 | Quickly vortex the plate, and centrifuge at 1,700*g* for 3 min at 4 °C.

73 | Incubate the plate in a thermocycler with a heated lid (100 °C), using the following program:

Cycle number	Temperature (°C)	Time
1	16	20 min
2	75	15 min
3	4	Hold

SPRI size selection ● TIMING 1 h

74 | Add 70 µl of water to the plate from Step 73 to adjust the total volume to 100 µl. SPRI size selection can be performed manually or with a liquid-handling robot (e.g., Zephyr NGS Workstation).

75 | Add 60.7 µl of SPRI beads to each well, and incubate for 5 min at room temperature.

76 | Place the plate on the magnet until the wells are clear (~5 min).

77 | Remove and discard the supernatant.

78 | Keeping the plate on the magnet, use fresh tips to add 200 µl of 70% (vol/vol) ethanol.
▲ **CRITICAL STEP** Gently touch the side of the well, and dispense ethanol into wells without disturbing the beads.

79 | Incubate the plate on the magnet for 30 s.

80 | Remove and discard the ethanol without disturbing the beads.

81 | Repeat Steps 78–80.

82 | Use a pipette to remove any drops of ethanol remaining in each well. Air dry for 10 min, or, for faster drying, place on a 37 °C heating block.

83 | Remove the plate from the magnet, and add 23 µl of 10 mM Tris-HCl to each well to elute the DNA. Mix 15 times by pipetting up and down, and incubate the plate at room temperature for 2 min.

84 | Place the plate back onto the magnet, and wait until the supernatant is clear, for ~3 min.

85 | Carefully aspirate 20 µl of the resuspended DNA into a new PCR plate. Discard the old plate.
■ **PAUSE POINT** Store completed libraries at –20 °C for up to 1 month.

Quality control ● TIMING 4 h

86 | Prepare master mix for 110 reactions according to the table below, and dispense 17 µl of master mix into each well of a TempPlate Semi-Skirt PCR plate.

Reagents	Volume per reaction (µl)	Final concentration
10× PCR buffer	2	1×
Solexa P1 (100 µM)	0.05	0.25 µM
Solexa P2 (100 µM)	0.05	0.25 µM
dNTP (1 mM)	2	0.1 mM
Taq polymerase	0.2	2 U/20 µl of PCR
Rediload dye	2	1×
ddH_2O	10.7	

87 | Add 3 µl of library from Step 85 to each respective well.

88 | Seal the plate with TempPlate XP PCR sealing film, and incubate the plate in a thermocycler with a heated lid (100 °C), using the following program:

Cycle number	Denature	Anneal	Extend
1	94 °C, 3 min		
2–36	94 °C, 10 s	60 °C, 45 s	72 °C, 90 s
37			72 °C, 5 min

89 | Run 15 µl from each PCR reaction on a 2% (wt/vol) Tris–borate–EDTA gel (SeaKem ME agarose) to determine library quality (**Fig. 2**). If there is no amplification, or the amplified product is not the right size, then the library preparation has failed.

? TROUBLESHOOTING

■ **PAUSE POINT** Store the plate at –20 °C for up to 1 month.

Library enrichment and normalization ● TIMING 2 h

90 | Prepare and dispense 40 µl of master mix into each well of the TempPlate Semi-Skirt PCR plate according to the table below.

▲ **CRITICAL STEP** Solexa P1 and Solexa P2 must be the limiting reagents in the PCR reactions, to ensure that each library reaches the same concentration, so be careful to use the correct volumes.

Reagents	Volume per reaction (µl)	Final concentration
5× NEB Phusion buffer	10	1×
Solexa P1 (100 µM)	0.1	0.2 µM
Solexa P2 (100 µM)	0.1	0.2 µM
dNTP (10 mM)	1	0.2 µM
ddH_2O	28.3	
Phusion enzyme	0.5	1 U/50 µl of PCR

91 | Transfer 10 µl of DNA from each of the purified libraries from Step 88 to the PCR plate.

92 | Incubate the plate in a thermocycler with a heated lid (100 °C), using the following program:

Cycle number	Denature	Anneal	Extend
1	98 °C, 30 s		
2–21	98 °C, 10 s	65 °C, 30 s	72 °C, 45 s
22			72 °C, 5 min

Library pooling ● TIMING 1 h

93 | Take 20 µl from each well of the plate from Step 92, and dispense it into one Costar reagent reservoir. Divide the pooled library into 800-µl aliquots placed in 1.5-ml microcentrifuge tubes.

94 | Add 498 µl of SPRI beads to each tube, and pipette up and down 15 times to mix.

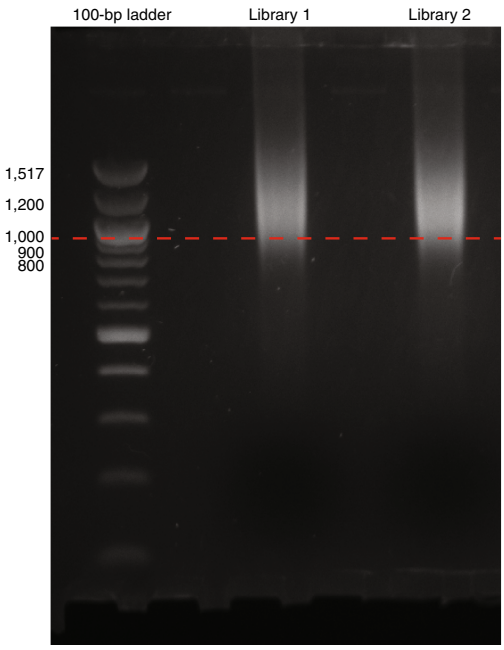


Figure 2 | Example QC agarose gel with >1-kb fragments. Left lane, 100-bp ladder; middle and right lanes, example libraries after PCR amplification in Step 88; dashed line, 1,000 bp.

95 | Place the tubes on an Eppendorf ThermoMixer for 5 min, 1,100 r.p.m., at room temperature.

96 | Place the tubes on a magnetic rack until they are clear.

97 | Aspirate and discard the supernatant from each tube.

98 | With tubes still on the rack, add 1.5 ml of 70% (vol/vol) ethanol to each tube, and wash for 30 s.

99 | Aspirate and discard the solution from each tube.

100 | Repeat Steps 98 and 99.

101 | Dry each tube on a 37 °C block for 5–10 min to evaporate the remaining 70% (vol/vol) ethanol.
▲ **CRITICAL STEP** Make sure that all ethanol is evaporated, but do not overdry the beads, to avoid decreasing the efficiency of DNA elution.

102 | Add 20 µl of 10 mM Tris-HCl to each tube to elute the DNA.

103 | Mix 15 times, and return the tubes to the magnetic rack, until the solution becomes clear, for ~15–20 min.

104 | Aspirate and combine all the samples from the tubes into one new microcentrifuge tube. Discard the tubes with beads.
▲ **CRITICAL STEP** Avoid aspirating any beads into the new tube.

105 | Label the tube with the date and sample ID.

106 | Use a NanoDrop or Qubit instrument to determine the concentration of the library. The concentration of the pooled library should be >120 ng/µl.
? TROUBLESHOOTING
■ **PAUSE POINT** Store the library at –20 °C for up to 1 month.

E-gel size selection ● **TIMING 1 h**
▲ **CRITICAL** We use E-gel for size selection, but other methods, such as gel extraction or Pippin Prep, can also be used.

107 | Follow the manufacturer's instructions to set up the E-Gel SizeSelect agarose gels and E-Gel Precast Agarose Electrophoresis System.

108 | Load 120–750 ng of pooled library from Step 105 into each well of the E-gel.

109 | To collect fragments of ~1 kb, run program 2 'E-Gel 4%' for ~32 min.
▲ **CRITICAL STEP** The exact collection time may vary between runs; use the E-gel ladder as a guide to select the correct collection time, and collect multiple fractions to ensure that at least one fraction contains the desired size.

110 | Manually collect the sample from the E-gel collection well into one tube per fraction. After each fraction, refill the collection wells with 25 µl of water, run for 1 min, and collect again. Repeat three times for a total of four fractions per sample.

111 | Rinse the collection wells one by one with an additional 25 µl of water. Add this rinse to the respective collection tubes.

SPRI cleanup ● **TIMING 1 h**
112 | Add water to each tube containing the size-selected pooled library to adjust the total volume to 150 µl.

113 | Add 225 µl of SPRI beads to each tube, and incubate for 5 min at room temperature.

114 | Place the tubes on a magnetic rack until the wells are clear, for ~5 min.

115 | Remove and discard the supernatant.

116 | Keeping the tubes on the magnetic rack, use fresh pipette tips to add 1,500 µl of 70% (vol/vol) ethanol.
▲ **CRITICAL STEP** Gently touch the side of the well, and dispense ethanol into wells without disturbing the beads.

117 | Incubate the tubes on the magnetic rack for 30 s.

118 | Remove and discard ethanol without disturbing the beads.

119 | Repeat Steps 116–118.

120 | Use a pipette to remove any drops of ethanol remaining in each tube.

121 | Let the sample tubes air dry for 10 min, or, for faster drying, place on a 37 °C heating block.

122 | Remove the tubes from the magnetic rack, and add 18 µl of 10 mM Tris-HCl to each tube to elute the DNA.

123 | Mix 15 times by pipetting up and down, and incubate the tubes at room temperature for 2 min.

124 | Return the tubes to the magnetic rack, and wait until the supernatants are clear, for ~3 min.

125 | Carefully aspirate 15 µl of the resuspended DNA into new tubes.
■ **PAUSE POINT** Store the library at –20 °C for up to 1 month.

BioAnalyzer sizing ● **TIMING 2 h**

126 | Determine the fragment-size distribution of the pooled library with an Agilent 2100 BioAnalyzer, according to the manufacturer's instructions. The average fragment will be ~120 bp longer than the sequenced fragment, owing to the adaptors (**Fig. 3**).
? TROUBLESHOOTING

Library quantification ● **TIMING 2 h**

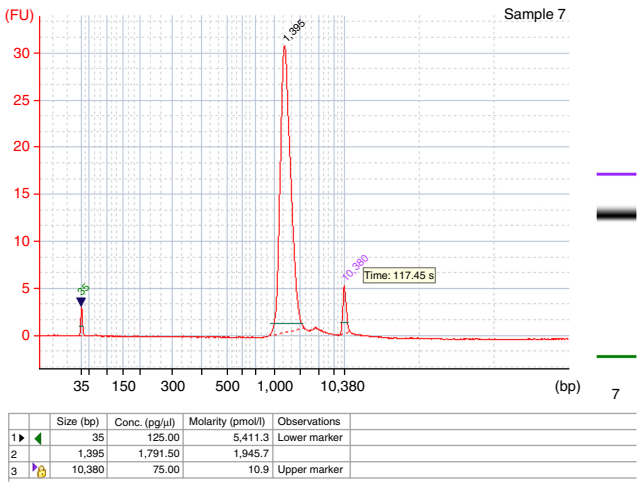
127 | The Kapa Library Quantification Kit is used to quantify the pooled library. Follow the manufacturer's instructions for the reaction setup. Change the cycling conditions as shown below to accommodate a longer insert. Calculate the library concentration with the corrected average size fragment from Step 126.

Cycle number	Denature	Annealing/ extension/ data acquisition
1	95 °C, 5 min	
2–36	95 °C, 30 s	60 °C, 90 s

? TROUBLESHOOTING

Sample loading ● **TIMING 30 min**

128 | Follow the standard MiSeq protocol, using MiSeq Reagent Kit v3, and load the library at a 15 ± 2 pM concentration. In the Illumina Experiment Manager Sample Sheet Wizard, set both cycles Read 1 and Read 2 to 340.
? TROUBLESHOOTING



	Size (bp)	Conc. (pg/µl)	Molarity (pmol/l)	Observations
1	35	125.00	5,411.3	Lower marker
2	1,395	1,791.50	1,945.7	
3	10,380	75.00	10.9	Upper marker

Figure 3 | Example Agilent Bioanalyzer electropherogram plot. The library fragments are between 900 and 2,000 bp, with an average size of 1,395 bp. Lower and upper markers are 35 and 10,380 bp, respectively. The table shows Bioanalyzer-calculated size in base pairs, concentration in pg/µl, and molarity in pmol/l, for a lower marker (1), a representative library (2), and an upper marker (3).

Draft assembly ● TIMING 1 h

▲ **CRITICAL** We have automated Steps 130–143 with a custom Perl script (available at https://github.com/dwbellott/shims2_assembly_pipeline/), but the workflow is described below to allow for direct use of the individual software tools or substitution of alternative tools.

129 | Download the compressed fastq-format reads from the MiSeq directly or via Illumina BaseSpace.
? TROUBLESHOOTING

130 | Run cutadapt on paired input to remove Illumina adaptors and trim low quality bases using the following command:

```
cutadapt --mask-adapter --quiet --match-read-wildcards -q 10 --minimum-length 22
-b AGATCGGAAGAGC -B AGATCGGAAGAGC -o library_1.cutadapt.fq.gz -p library_2.cutadapt.
fq.gz library_1.fastq.gz library_2.fastq.gz
```

131 | Align reads to *E. coli* genome with bowtie2 using the following command:

```
bowtie2 --very-sensitive-local --n-ceil L,0,1 -I 0 -X 2501 -x
ecoli_genome_bowtie_index -1 library_1.cutadapt.fq.gz -2 library_2.cutadapt.fq.gz
-S ecoli.sam
```

132 | Parse the SAM format output to extract unaligned read pairs using the following commands:

```
samtools view -f 77 ecoli.sam | samtools sort -n - | samtools view - | awk '{print
"@"$1"\n"$10"\n+\n"$11}' | gzip -9c >library_1.ecoli.fq.gz

samtools view -f 141 ecoli.sam | samtools sort -n - | samtools view - | awk '{print
"@"$1"\n"$10"\n+\n"$11}' | gzip -9c >library_2.ecoli.fq.gz
```

Use the aligned reads to estimate the average fragment size and s.d. with the following command:

```
samtools view -f 66 ecoli.sam | cut -f 9 | awk '{sum += sqrt($ 1^2); sumsq +=
$ 1^2} END {printf " %f %f \n " , sum/NR, sqrt((sumsq - sum^2/NR)/NR)}'
```

▲ **CRITICAL STEP** Fewer than 25% of reads should align to the *E. coli* genome; >25% contamination indicates a problem with cell lysis in Step 12.
? TROUBLESHOOTING

133 | Align the filtered reads to the BAC-cloning vector with bowtie2 with the following command:

```
bowtie2 --very-sensitive-local --n-ceil L,0,1 -I 0 -X 2501 --x cloning_vector_bow-
tie_index --1 library_1.ecoli.fq.gz --2 library_2.ecoli.fq.gz --S vector.sam
```

134 | Parse the SAM-format output to extract unaligned read pairs with the following commands:

```
samtools view -f 77 vector.sam | samtools sort -n - | samtools view - | awk '{print
"@"$1"\n"$10"\n+\n"$11}' | gzip -9c >library_1.vector.fq.gz

samtools view -f 141 vector.sam | samtools sort -n - | samtools view - | awk '{print
"@"$1"\n"$10"\n+\n"$11}' | gzip -9c >library_2.vector.fq.gz
```

Use the aligned reads to estimate the average fold coverage of the cloning vector with the following command:

```
samtools view -f 66 ecoli.sam | cut -f 10 | tr -d '\n' | wc --m
```

Divide the result by the length of the vector sequence to obtain the fold coverage.

135 | Optional: if the average fragment length minus the s.d. measured in Step 134 is less than twice the average read length, overlap the forward and reverse reads with flash by running the following command:

```
flash bowtie2 library_1.vector.fq.gz library_2.vector.fq.gz -f average_fragment_size
-s standard_deviation_fragment_size -r averae_read_length -o library_name -d output_
directory
```

136 | If you ran flash in Step 135, assemble the reads with SPAdes with the following command:

```
spades.py -1 library.notCombined_1.fastq -2 library.notCombined_2.fastq
-s library.extendedFrags.fastq --only-assembler --careful -o output_directory --cov-
cutoff fold_coverage
```

Otherwise, assemble with the following command:

```
spades.py -1 library_1.vector.fq.gz -2 library_2.vector.fq.gz
--only-assembler --careful -o output_directory --cov-cutoff fold_coverage
```

137 | Align the quality-trimmed reads to `scaffolds.fasta` produced by SPAdes, and generate a sorted BAM-format alignment output with the following command:

```
bowtie2build -q scaffolds.fasta scaffolds_bowtie_index
```

```
bowtie2 -x scaffolds_bowtie_index -1 library_1.cutadapt.fq.gz
-2 library_2.cutadapt.fq.gz | samtools view -b -S - | samtools sort -
>scaffolds.bowtie.sorted.bam
```

138 | Use SPAdes scaffolds and sorted BAM as input to run BESST with the following command:

```
runBESST -c scaffolds.fasta -f scaffolds.bowtie.sorted.bam -o output_directory
--orientation fr
```

139 | Use quality-trimmed reads and BESST scaffolds as input to run Gap2Seq with the following command:

```
Gap2Seq -scaffolds BESST_output/pass1/Scaffolds_pass1.fa -filled output_file -reads
library_1.cutadapt.fq.gz,library_2.cutadapt.fq.gz
```

140 | Order and orient the scaffolds by using BLAST to align the final scaffolds to clone end sequences and known peptides, if available. Clone end sequences will align to the first and last scaffolds in the assembly in opposite orientation. Peptide sequences with partial alignments to several scaffolds can be used to order and orient scaffolds between the ends. Each partial alignment represents a coding exon, which should run in order, on the same strand, from the N to the C terminus of the peptide.

141 | Align the reads to the final scaffolds with bowtie2 with high gap-opening and extension penalties, to generate a sorted BAM-format alignment for Consed with the following command:

```
bowtie2-build -q final.fasta final_bowtie_index
```

```
bowtie2 -I 0 -X 2501 --rdg 502,502 --rfg 502,502 -x final_bowtie_index -1 library.
notCombined_1.fastq -2 library.notCombined_2.fastq -s library.extendedFrags.fastq |
samtools view -b -S - | samtools sort - >final.bowtie.sorted.bam
```

142 | Use `makeRegionsFile.perl` from the Consed package to prepare a regions file for Consed with the following command:

```
makeRegionsFile.perl final.fasta
```

143 | Use the bam2ace command in Consed to generate the files and directory structures used by Consed with the following command:

```
consed -bam2ace -bamFile final.bowtie.sorted.bam -regionsFile finalRegions.txt -dir
consed_ouput
```

? TROUBLESHOOTING

Identifying SFVs ● TIMING 0–2 h

144 | Align contigs from your clone of interest to contigs from putative neighbors in the tiling path with BLAST.
▲ CRITICAL STEP Complete draft assemblies for all neighboring clones before examining overlaps for SFVs.

145 | Identify the positions of discrepancies between the putative neighbors from the alignment.
▲ CRITICAL STEP Focus on single-base mismatches rather than variations in microsatellites. Microsatellite repeats are unstable, and differences between clones are much more likely to represent assembly artifacts, sequencing errors, or mutations during BAC culture.

146 | Open the assembly from Step 143 of each putative neighbor in Consed. For each clone, navigate to the consed_output/edit_dir directory with the following command:

```
cd consed_output/edit_dir
```

Then type the following command to open the assembly:

```
consed
```

True SFVs will be supported by high-quality bases in most reads. Correct any errors in the consensus sequence by manually editing the consensus sequence to the correct base.

147 | Correct the tiling path to account for new SFV information by inserting a gap between neighbors that can be distinguished by SFVs. Keep track of the tiling path in a spreadsheet or tab-delimited text file. Some researchers may wish to use either the TPF[63] or AGP[64] formats, but even a simple three-column list (clone name, clone length, and length of overlap with the clone on the next line) should suffice for most projects.

148 | Use the newly identified SFVs to search for new neighbors for each clone among other sequenced clones with BLAST. True neighbors will have a long exact match that includes the SFV. If there are no neighbors that match each SFV, screen the BAC library (e.g., with overgo hybridization[65]) for additional overlapping clones, and sequence them to find clones that share each SFV and extend each contig.

149 | Correct the tiling path to account for newly identified overlaps, and remove the gaps between clones that share SFVs in a long (>10 kb) overlap. Joins between large contigs in the tiling path may require edits to many lines of the file, especially if the new overlap requires the orientation of one contig to be reversed. Be sure to make a backup of the file before each change, or preferably use a version-control system.

Finishing ● TIMING 0–8 h

150 | Identify any collapsed duplications according to aberrantly high (greater than twice the clone average) read depth in Consed's 'Assembly View' window.
▲ CRITICAL STEP Resolve all discrepancies between neighboring clones, and identify SFVs (Steps 144–149) before finishing.

151 | Resolve any collapsed duplications near the ends of the clone insert by reassembling the clone from reads that include vector sequences (before Step 133), with the following command:

```
spades.py -1 library_1.ecoli.fq.gz -2 library_2.ecoli.fq.gz
--only-assembler --careful -o output_directory --cov-cutoff fold_coverage
```

SPAdes will be able to correctly identify repeat variants near the cloning vector by using a mate in the vector sequence.

152 | Resolve the remaining collapsed duplications by pulling apart deep contigs in Consed's 'Aligned Reads' window. Use consistent differences between two or more sets of reads to assign each to a separate contig.

153 | Order and orient contigs on the basis of mate pairs that connect adjacent contigs, which are visible as red and blue lines in Consed's 'Assembly view', and by looking for contigs that end in similar SSRs.

154 | Close short gaps by padding the ends of contigs with 100–500 Ns, and then repeating Steps 141–143. Call the consensus sequence at contig ends again; join newly extended contigs on the basis of overlaps identified in Consed's 'Assembly view'. To view overlaps between contigs, select 'Sequence Matches' from the 'What to Show' menu.

155 | Resolve SSRs at contig ends by using the raw reads (before the quality trimming in Step 130) and an overlap–layout–consensus assembler (such as phrap in the Consed package).
▲ **CRITICAL STEP** Stutter noise from replication slippage in SSRs causes divergent reads and low-quality base calls, thus disrupting the short, perfect matches used by *k*-mer-based assemblers such as SPAdes. Most gaps will be at SSRs. In some cases, unambiguous resolution of these repeats may not be possible, and they should be annotated as unresolved in Step 158.

156 | Use Consed's autofinish feature to design PCR primers to amplify the sequence of any remaining gaps, and determine the sequence of the PCR products. PCR products of approximately the same size as the average library fragment can be processed in parallel with clones, starting at Step 62.

157 | After all contigs are ordered and oriented, and all gaps are closed, export the finished clone sequence from Consed to fasta format with the 'Export Consensus Sequence' command from the 'File' menu.

158 | Remove any vector-sequence contamination at the ends of the clone. These may be too short to be identified by BLAST, so inspect and edit the consensus sequence manually, either in Consed or a text editor. Annotate any remaining ambiguities in the clone sequence (e.g., unresolved SSRs) by compiling a feature table[66], which will be useful when finished clone sequences are submitted to GenBank.

● **TIMING**
Steps 1–3, pick clones and grow cultures: 18 h
Steps 4–7, glycerol-stock plates: 30 min
Steps 8–28, alkaline lysis: 2–3 h
Steps 29–39, prepare barcoded adaptor: 4 h
Steps 40–43, DNA shearing: 1 h
Steps 44–55, SPRI cleanup: 1 h
Steps 56–73, library construction: 3–4 h
Steps 74–85, SPRI size selection: 1 h
Steps 86–89, quality control: 4 h
Steps 90–92, library enrichment/normalization: 2 h
Steps 93–106, library pooling: 1 h
Steps 107–111, E-gel size selection: 1 h
Steps 112–125, SPRI cleanup: 1 h
Step 126, BioAnalyzer sizing: 2 h
Step 127, library quantification: 2 h
Step 128, sample loading: 30 min
Steps 129–143, draft assembly: 1 h
Steps 144–149, identifying SFVs: 0–2 h per clone
Steps 150–158, finishing: 0–8 h per clone

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

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
43	Wrong-sized fragments after shearing	Too much/little DNA in Covaris tubes	Limit DNA in microTube to 50–800 ng per tube
55	No DNA after SPRI cleanup	70% (vol/vol) ethanol not made freshly	Make fresh 70% (vol/vol) ethanol immediately before cleanup
		SPRI solution added incorrectly	Mix SPRI solution well, and pipette slowly to aspirate the correct volume
89	Quality control failed/no product was amplified	Library construction failed	Perform agarose gel electrophoresis to check barcode adaptors for degradation Adaptor freeze–thaw cycles should be limited. It is better to store them in a freezer without automatic defrosting
106	Library concentration too low	Library enrichment failed	Repeat Steps 90–106
126	BioAnalyzer trace indicates the presence of smaller-than-desired fragments	SPRI cleanup performed incorrectly	Perform Steps 112–125 again to remove smaller fragments
127	qPCR indicates that library concentration is too low	Enrichment failed	Run the enriched product on a gel along with the unenriched product If quality control (Steps 86–89) passed but the enrichment step failed, perform the enrichment again with new reagents
128	Sequenced indexes vary in abundance by more than twofold	Normalization failed because universal primers were not limiting	Increase dNTP concentration so that primers are the limiting reagent
		Library construction failed because of degraded barcode adaptors	Check barcode adaptors for degradation; adaptor quality affects ligation efficiency and library yield
129	Individual samples have no reads or very few reads	DNA extraction failed	Regrow the clone for an additional round of sequencing Regrow and add to the next run, or replace the clone with another
132	>25% *E. coli* genomic-DNA contamination	Lysis (Step 12) failed	Use fresh Solution 2, and mix gently to avoid shearing *E. coli* genomic DNA
143	Insert sequences of two clones present at low coverage	BAC or fosmid culture contaminated with another clone	Streak out mixed culture, pick individual clones, and resequence
	Clone sequence does not match predicted sequence from end sequences or neighboring clones	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
	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

ANTICIPATED RESULTS

We typically pool 192 clones for a single MiSeq run, generating ~20 million 2 × 340 reads. Approximately 80% of reads pass Illumina quality filters, for a total of ~15 Gb of sequence data, with ~55% of bases at Q30 or higher. Each sample typically receives ~0.5% of the total reads and has ~20% *E. coli* genomic DNA contamination. We usually observe average fragment sizes of ~1,100–1,200 bp, with an s.d. of ~100–120 bp. If sample abundances vary by more than twofold, normalization (Steps 90–92) has failed (see troubleshooting information for Step 128). If *E. coli* contamination is greater than 25%, then the alkaline lysis (Step 12) was performed improperly.

In our experience in sequencing vertebrate sex chromosomes, fosmids almost always assemble into a single contig. Approximately 15–20% of BACs assemble into a single contiguous finished sequence without any human intervention. Another 35–40% are in only two to three contigs that are easily ordered and oriented. The remaining 40–50% of BACs are still highly contiguous (more than 80% of BACs have *N*50 (shortest sequence length at 50% of the genome) >50 kb, and more than 50% have *N*50 >100 kb), but they may contain collapsed repeats that require scaffolding or manual review to arrive at a finished assembly (Steps 151–159). Most gaps are at SSRs and should be relatively simple to resolve with an alternate assembler.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS This work was supported by the National Institutes of Health and the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS D.W.B., H.S., J.F.H., and D.C.P. designed the study. D.W.B. and T.-J.C. developed the experimental methods. D.W.B. wrote the scripts for computational analysis. D.W.B., T.-J.C., and D.C.P. wrote the manuscript.

COMPETING INTERESTS
The authors declare no competing interests.

1. Mueller, J.L. *et al.* Independent specialization of the human and mouse X chromosomes for the male germ line. *Nat. Genet.* **45**, 1083–1087 (2013).
2. Lupski, J.R. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14**, 417–422 (1998).
3. Stankiewicz, P. & Lupski, J.R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
4. Ross, M.T. *et al.* The DNA sequence of the human X chromosome. *Nature* **434**, 325–337 (2005).
5. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
6. Gordon, D. & Green, P. Consed: a graphical editor for next-generation sequencing. *Bioinformatics* **29**, 2936–2937 (2013).
7. Bonfield, J.K., Smith, K. & Staden, R. A new DNA sequence assembly program. *Nucleic Acids Res.* **23**, 4992–4999 (1995).
8. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
9. Alkan, C., Sajjadian, S. & Eichler, E.E. Limitations of next-generation genome sequence assembly. *Nat. Methods* **8**, 61–65 (2011).
10. Gordon, D. *et al.* Long-read sequence assembly of the gorilla genome. *Science* **352**, aae0344 (2016).
11. Eichler, E.E. Segmental duplications: what's missing, misassigned, and misassembled—and should we care? *Genome Res.* **11**, 653–656 (2001).
12. Dennis, M.Y. *et al.* Evolution of human-specific neural SRGAP2 genes by incomplete segmental duplication. *Cell* **149**, 912–922 (2012).
13. Steinberg, K.M. *et al.* Single haplotype assembly of the human genome from a hydatidiform mole. *Genome Res.* **24**, 2066–2076 (2014).
14. Watson, C.T. *et al.* Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am. J. Hum. Genet.* **92**, 530–546 (2013).
15. Mohajeri, K. *et al.* Interchromosomal core duplicons drive both evolutionary instability and disease susceptibility of the chromosome 8p23.1 region. *Genome Res.* **26**, 1453–1467 (2016).
16. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
17. Kuroda-Kawaguchi, T. *et al.* The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat. Genet.* **29**, 279–286 (2001).
18. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–837 (2003).
19. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat. Genet.* **38**, 463–467 (2006).
20. Lange, J. *et al.* Intrachromosomal homologous recombination between inverted amplicons on opposing Y-chromosome arms. *Genomics* **102**, 257–264 (2013).
21. Lange, J., Skaletsky, H., Bell, G.W. & Page, D.C. MSY Breakpoint Mapper, a database of sequence-tagged sites useful in defining naturally occurring deletions in the human Y chromosome. *Nucleic Acids Res.* **36**, D809 D (2008).
22. Lange, J. *et al.* Isodicentric Y chromosomes and sex disorders as byproducts of homologous recombination that maintains palindromes. *Cell* **138**, 855–869 (2009).
23. Repping, S. *et al.* Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nat. Genet.* **35**, 247–251 (2003).
24. Repping, S. *et al.* Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am. J. Hum. Genet.* **71**, 906–922 (2002).
25. Repping, S. *et al.* A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8-Mb deletion in the azoospermia factor c region. *Genomics* **83**, 1046–1052 (2004).
26. Rozen, S.G. *et al.* AZFc deletions and spermatogenic failure: a population-based survey of 20,000 Y chromosomes. *Am. J. Hum. Genet.* **91**, 890–896 (2012).
27. Bellott, D.W. *et al.* Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. *Nature* **508**, 494–499 (2014).
28. Bellott, D.W. *et al.* Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* **466**, 612–616 (2010).
29. Hughes, J.F. *et al.* Strict evolutionary conservation followed rapid gene loss on human and rhesus Y chromosomes. *Nature* **483**, 82–86 (2012).
30. Hughes, J.F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463**, 536–539 (2010).
31. Soh, Y.Q. *et al.* Sequencing the mouse Y chromosome reveals convergent gene acquisition and amplification on both sex chromosomes. *Cell* **159**, 800–813 (2014).
32. Bellott, D.W. *et al.* Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat. Genet.* **49**, 387–394 (2017).

33. Li, G. *et al.* Comparative analysis of mammalian Y chromosomes illuminates ancestral structure and lineage-specific evolution. *Genome Res.* **23**, 1486–1495 (2013).
34. Sato, K., Motoi, Y., Yamaji, N. & Yoshida, H. 454 Sequencing of pooled BAC clones on chromosome 3H of barley. *BMC Genom.* **12**, 246 (2011).
35. Quinn, N.L. *et al.* Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genom.* **9**, 404 (2008).
36. Rounsley, S., Lin, X. & Ketchum, K.A. Large-scale sequencing of plant genomes. *Curr. Opin. Plant Biol.* **1**, 136–141 (1998).
37. National Center for Biotechnology Information. *Commercial and Academic Suppliers of Clones, Libraries and Other Reagents Described in Clone DB* https://www.ncbi.nlm.nih.gov/clone/content/distributors/ (2017).
38. Guha, S. & Maheshwari, S.C. Cell division and differentiation of embryos in pollen grains of *Daturain vitro*. *Nature* **212**, 97–98 (1966).
39. Jain, S.M., Sopory, S.K. & Veilleux, R.E. *In vitro haploid production in higher plants* (Kluwer Academic Publishers, 1996).
40. Bonfield, J.K. & Whitwham, A. Gap5: editing the billion fragment sequence assembly. *Bioinformatics* **26**, 1699–1703 (2010).
41. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
42. Wilkening, S. *et al.* Genotyping 1000 yeast strains by next-generation sequencing. *BMC Genom.* **14**, 90 (2013).
43. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 1154–1162 (2012).
44. Goodwin, S. *et al.* Oxford nanopore sequencing, hybrid error correction, and *de novo* assembly of a eukaryotic genome. *Genome Res.* **25**, 1750–1756 (2015).
45. Berlin, K. *et al.* Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
46. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
47. Madoui, M.A. *et al.* Genome assembly using nanopore-guided long and error-free DNA reads. *BMC Genom.* **16**, 327 (2015).
48. Tomaszkiewicz, M. *et al.* A time- and cost-effective strategy to sequence mammalian Y chromosomes: an application to the *de novo* assembly of gorilla Y. *Genome Res.* **26**, 530–540 (2016).
49. McCoy, R.C. *et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly-repetitive transposable elements. *PLoS One* **9**, e106689 (2014).
50. Li, R. *et al.* Illumina synthetic long read sequencing allows recovery of missing sequences even in the "finished" *C. elegans* genome. *Sci. Rep.* **5**, 10814 (2015).
51. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*). *Nat. Biotechnol.* **31**, 135–141 (2013).
52. Seo, J.S. *et al.* *De novo* assembly and phasing of a Korean human genome. *Nature* **538**, 243–247 (2016).
53. Nagaraja, R. *et al.* Characterization of four human YAC libraries for clone size, chimerism and X chromosome sequence representation. *Nucleic Acids Res.* **22**, 3406–3411 (1994).
54. Venter, J.C., Smith, H.O. & Hood, L. A new strategy for genome sequencing. *Nature* **381**, 364–366 (1996).
55. Glenn, T.C. Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour.* **11**, 759–769 (2011).
56. Agencourt Bioscience Corporation. *Agencourt CosMCPrep High and Low Copy Plasmid Purification* https://www.beckmancoulter.com/wsrportal/bibliography?docname=Protocol 000381v012.pdf (2006).
57. Lange, V. *et al.* Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genom.* **15**, 63 (2014).
58. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
59. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
60. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
61. Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J. & Arvestad, L. BESST: efficient scaffolding of large fragmented assemblies. *BMC Bioinform.* **15**, 281 (2014).
62. Salmela, L., Sahlin, K., Makinen, V. & Tomescu, A.I. Gap filling as exact path length problem. *J. Comput. Biol.* **23**, 347–361 (2016).
63. Church, D.M. *Tiling Path File (TPF) Specification v1.4* https://www.ncbi.nlm.nih.gov/projects/genome/assembly/TPF_Specification_v1.4_20110215.pdf (2011).
64. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/assembly/agp/AGP_Specification/ (2014).
65. McPherson, J.D. *et al.* A physical map of the human genome. *Nature* **409**, 934–941 (2001).
66. National Center for Biotechnology Information. *What is tbl2asn?* https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/ (2017).