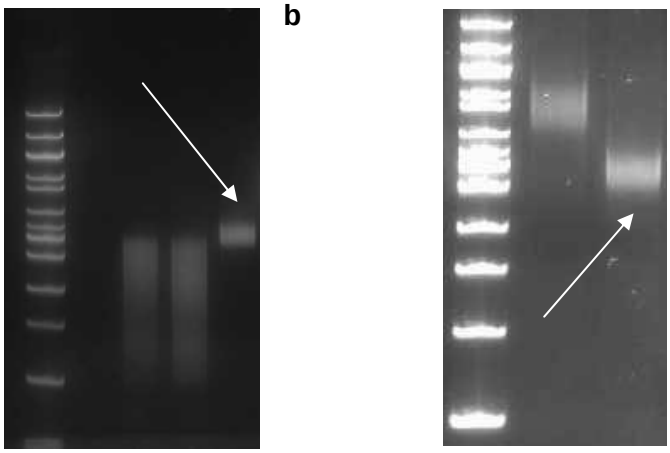
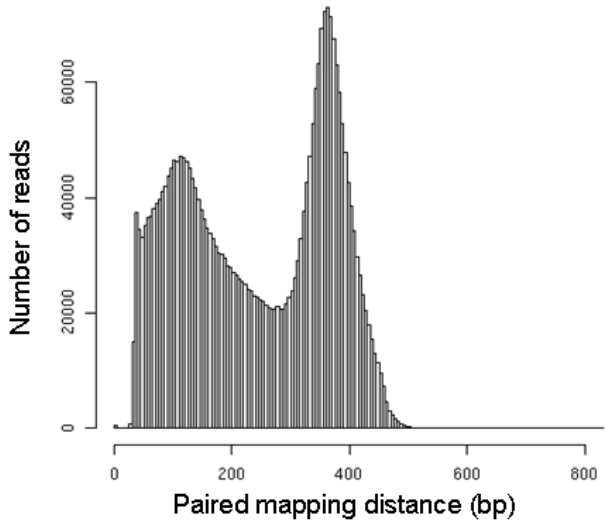
Supplementary Figure 1. Length of library fragments by PAGE

a **b**



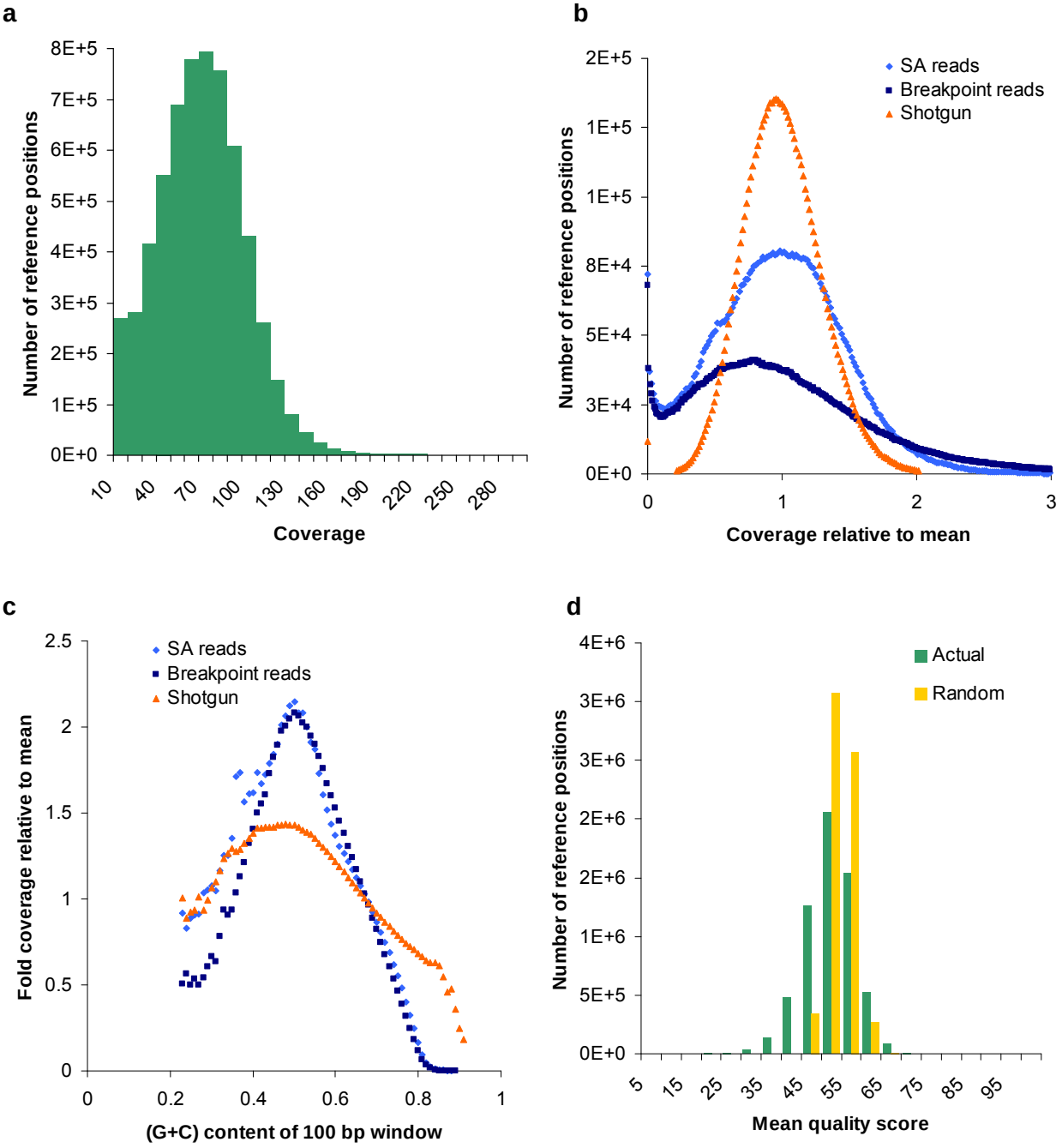
(a) PAGE of NEB 100 bp ladder and nebulized and size-selected ~550 bp *P. aeruginosa* fragments. (b) PAGE of NEB 100 bp ladder and Biorupted and size-selected Methylamine metagenomic fragments.

Supplementary Figure 2. Length distribution of subassembly fragments by paired-end sequencing



Histogram of mapping distance separating tag and breakpoint reads from the 450-600 bp size-selection performed at the end of the subassembly library construction protocol of a representative subset of the *Pseudomonas* data. Paired 20x76 bp reads were mapped to the PAO1 reference genome using *maq.* Shorter mapping distances are thought to arise from over-amplification during PCR, which causes shorter fragments to migrate with longer fragments during PAGE. Retained shorter fragments are then preferentially amplified and sequenced during the Illumina sequencing protocol. Careful PCR amplification is essential to prevent small fragments from completely dominating the sequencing reaction. The non-uniform nature of this distribution may contribute to the bimodal distribution of subassembled read length that we observed for this sample.

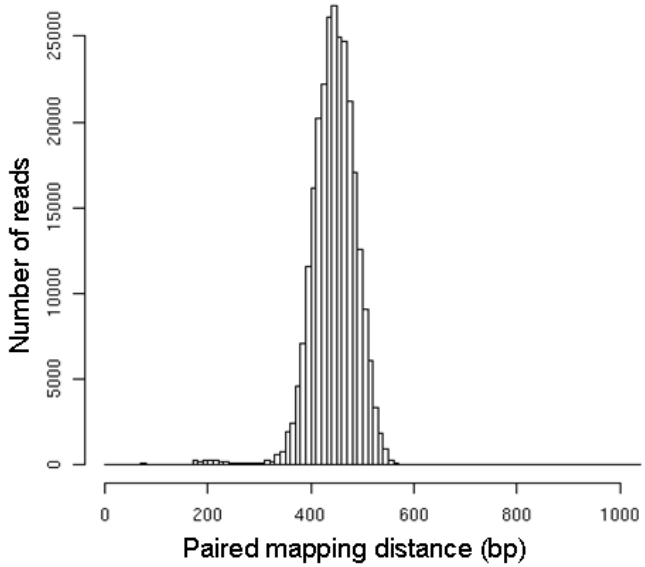
Supplementary Figure 3. Coverage of the PAO1 reference by SA reads



(a) Histogram of coverage of the PAO1 reference by SA reads as determined by BLAST alignment (bin size 10 bp). (b) Histogram of coverage of the PAO1 reference by SA reads, a standard Illumina paired-end 36 bp shotgun library, and the 76 bp breakpoint reads that contributed to SA reads. (c) Mean (G+C) content in the 100 bp window around reference positions with a given coverage on the x-axis by SA reads, a standard Illumina paired-end 36 bp shotgun library, and the 76 bp breakpoint reads that contributed to SA reads. A strong relationship between coverage and (G+C) content is observed. That is, reference bases in very high (G+C) content regions tend to have reduced coverage relative to the mean, and regions with intermediate (G+C) content are correspondingly overrepresented. This is likely due to

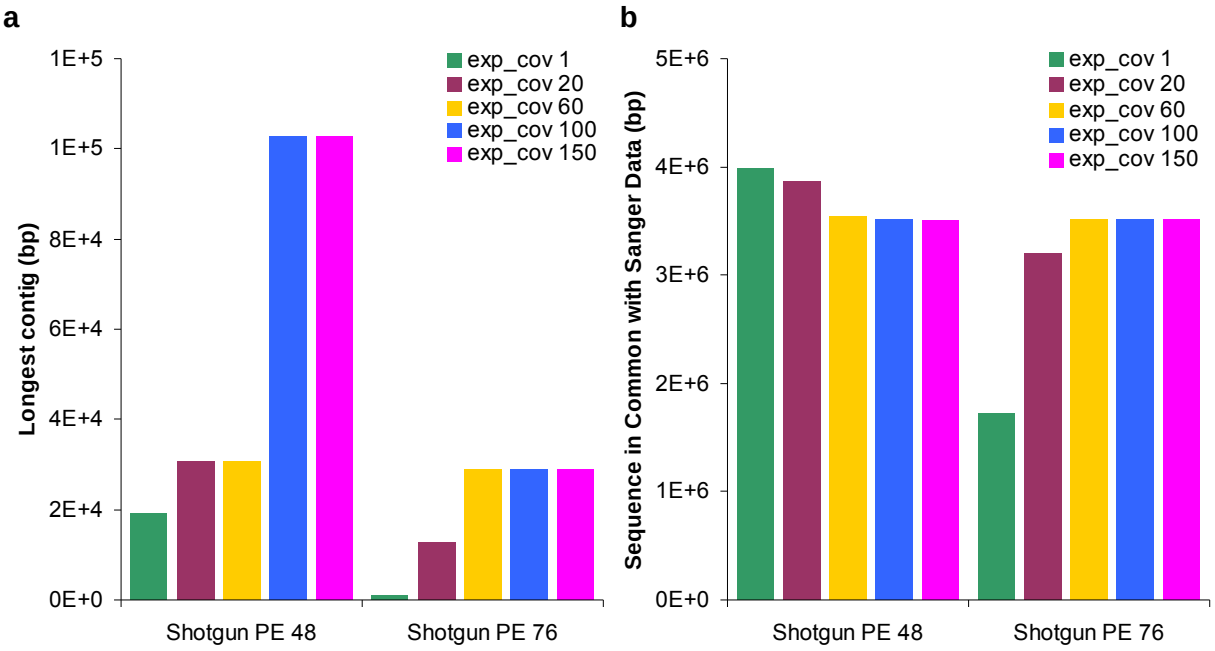
(G+C) content biases present during the PCR steps of library construction, as a similar relationship is observed for the contributing 76 bp reads, and could likely be mitigated by PCR conditions designed to reduce (G+C) bias. (d) Distribution of mean quality score (and therefore predicted error rate) across the reference. The number of reference positions with a given mean quality score is plotted in green (“Actual”), while a simulated distribution was made by randomizing the full set of quality score assignments in SA reads and then recalculating mean quality scores for reference positions, and is plotted in yellow (“Random”). The standard deviation of the actual distribution was six compared to three for the random distribution, indicating a small systematic bias in quality score (and therefore error) distribution across the PAO1 genome.

Supplementary Figure 4. Length distribution of metagenomic fragments



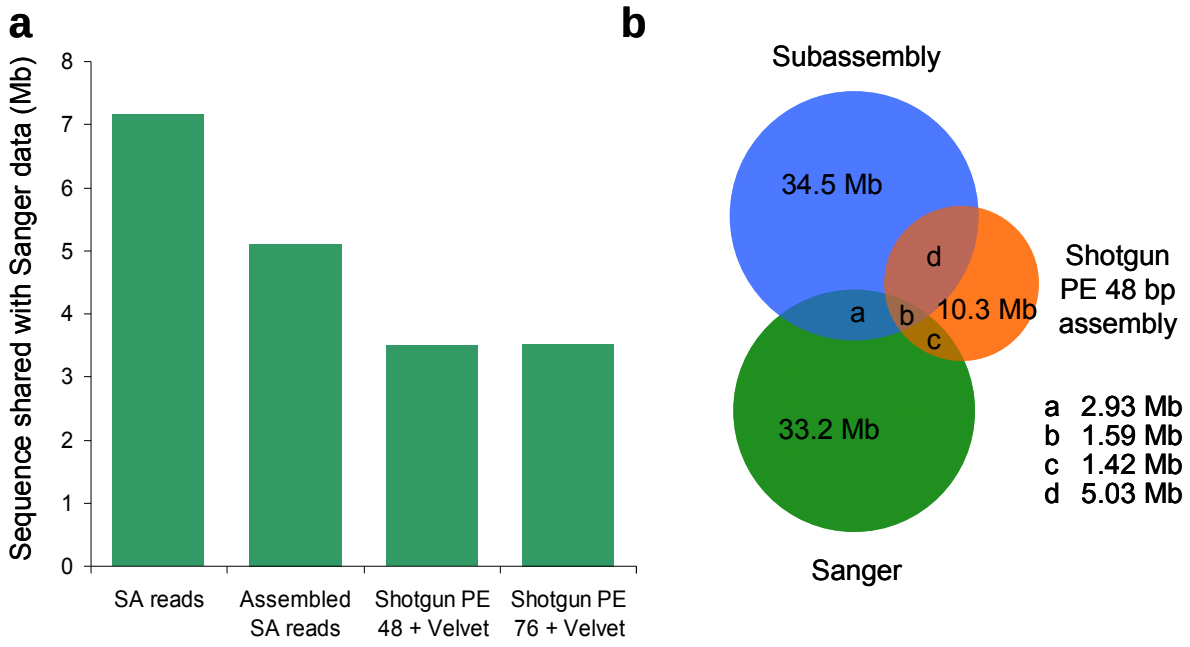
Histogram of the mapping distance separating paired tag reads from 36 bp paired-end sequencing data of the metagenomic library fragments (used to pair and merge TDRGs). Paired-end reads were mapped to the recently obtained Sanger data from the same sample using *maq.* Some selection for shorter molecules during the Illumina sequencing protocol may have taken place, shifting the peak of the distribution somewhat shorter than would be expected based on PAGE of the original fragments (**Supplementary Fig.1**).

Supplementary Figure 5. Optimization of Velvet parameters for shotgun metagenomic assembly



We optimized Velvet parameters for shotgun metagenomic assembly with respect to contig length and sequence shared with available Sanger data. (a) Maximum contig length as a function of changing Velvet parameters for assembly of shotgun paired-end 48 bp and paired-end 76 bp reads. Contig length was found to be very sensitive to the exp_cov parameter. (b) Sequence in common with the available Sanger data from the same sample as a function of changing Velvet parameters as in (a). Shared sequence was found to be somewhat sensitive to the exp_cov parameter in an unpredictable fashion, with shared sequence decreasing with increased exp_cov for the 48 bp reads and increasing with increased exp_cov for the 76 bp reads. To optimize length and coverage, we chose to perform subsequent analyses with the exp_cov = 100.

Supplementary Figure 6. Coverage overlap of metagenomic sample between sequencing methods



(a) Sequence shared with Sanger data for SA reads, assembled SA reads, and Velvet-assembled shotgun paired-end reads. Shared sequence was estimated by considering BLAST alignments with at least 98% identity across at least 100 bp. SA reads covered more than twice as much of the Sanger data as either shotgun assembly. (b) Venn diagram illustrating reciprocal coverage across data sets as determined by stringent BLAST analysis. Contigs produced by Celera assembly of SA reads, Velvet assembly of a 48 bp paired-end shotgun library with exp_cov=100, and the recently obtained Sanger sequencing data were compared to one another using BLAST. Coverage was defined as the best pair-wise match between bases as determined by the bit score of the alignment as long as the alignment had at least 98% identity and was at least 100 bp long. The bases in common shown here are not in exact agreement with those presented in (a) because, for the purposes of constructing this diagram, each base was only allowed to align to one corresponding base in another dataset. Circles are drawn to scale; regions of overlap not to scale.

Supplementary Table 1. *Phrap* optimization

Min match	Min score	Force level	Index word size	# of TDRGs	Mean longest SA read	Median longest SA read	Fraction of non-BLASTing SA's	Fraction of SA's BLASTing <90% of length	Fraction of mismatches among BLASTing bases
12	12	1	10	2619	361.6	403	0.004964	0.02993	0.001513
10	12	1	10	2619	364.4	406	0.004964	0.0284	0.001543
10	12	1	8	2619	364.4	406	0.004964	0.0284	0.001543
10	10	1	8	2619	369.5	409	0.004964	0.04106	0.001551
8	10	1	8	2619	371.9	411	0.004964	0.04643	0.001579

A representative subset of 10,000 Pseudomonas TDRGs was randomly selected and subjected to phrap assembly using different parameters and the resulting lengths and qualities of the longest subassemblies from each TDRG were assessed. We determined that parameters of minmatch 10, minscore 12, force level 1, and index word size 8, achieved the optimal balance between assembly accuracy, measured as the fraction of subassembled reads BLASTing across at least 90% of their length in a single BLAST hit (and the fraction removed because of oppositely oriented reads, not shown), and subassembled read length.

Supplementary Table 2. Summary statistics for subassembled reads

Sample	Original fragment size	# of read-pairs	# of filtered TDRGs	Median length
P. aeruginosa	~550 bp	56.8M	1,031,537	338 bp
Metagenomic	~450 bp	21.8M	262,298	256 bp
Metagenomic (merged)	~450 bp	21.8M+1.8M	180,008 (90,004 pairs)	408 bp

For the two samples used and the two analyses performed of the methylamine-enriched metagenomic sample, listed is the approximate size of long fragments from which subassembly libraries were generated, the number of Illumina read-pairs that were used to generate subassembled (SA) reads (merged analysis also shows the number of reads used to pair tags), the number of TDRGs after filtering for successful assembly and properly oriented contributing reads, and the median length of the longest SA read from each filtered TDRG.

Supplementary Table 3. Summary statistics from assembly of metagenomic SA reads versus assembly of a standard shotgun library

Input	Assembly strategy	# of contigs	Median contig length	Sequence in contigs ≥ 200 bp	Longest contig
SA reads	Celera	86,418	390 bp	35.7 Mb	6,000 bp
Shotgun PE 48 bp	Velvet (exp_cov = 100)	17,618	332 bp	9.9 Mb	102,806 bp
Shotgun PE 76 bp	Velvet (exp_cov = 100)	33,374	315 bp	16.0 Mb	28,861 bp

Comparison of assembly of short reads from a standard Illumina shotgun library prepared from the metagenomic sample to Celera assembly of the full complement of SA reads from the same sample. Listed is the assembly input, the assembly strategy used, and, for contigs at least 200 bp long, the number of contigs produced, the median contig length, the total amount of sequence contained in such contigs, and the longest contig. 76 bp paired-end (PE) reads were collected from a standard shotgun library and were trimmed to 48 bp reads to match the amount of sequence collected per read-pair for subassembly (20+76). Velvet assembly was performed using both 48 bp and 76 bp paired-end reads, but the same total amount of raw sequence as collected for subassembly (2.2 gigabases) was used in each shotgun assembly. Notably, while the shotgun assemblies achieve greater contiguity at the longest lengths, potentially due to deep sampling of abundant genomes or to misassemblies, subassembly produces at least twice as much sequence at the lengths necessary for phylogenetic analysis and gene prediction.

Supplementary Table 4. Oligo sequences

	Name	Sequence
Tag-adjacent adaptor oligos	Ad1	TCGCAATACAGAGTTTACCGCATT
	Ad1_rc	/5Phos/ATGCGGTAAACTCTGTATTGCGA
	Ad2	CTCTTCCGCATCTCACAACCTACT
	Ad2_rc	/5phos/GTAGGTTGTGAGATGCGGAAGAG
Breakpoint-adjacent adaptor oligos	Illum_rev	CTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T
	Illum_rev_rc	/5Phos/GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG
Bottleneck PCR primers	Ad1_amp	/5phos/TCGCAATACAGAGTTTACCGCATT
	Ad2_amp	/5phos/CTCTTCCGCATCTCACAACCTACT
TDRG merging PCR primer	Illum_amp_r_Ad2	CAAGCAGAAGACGGCATACGAGATATCGAGAGCCTCTTCCGC ATCTCACAACCTACT
Sequencing PCR primers	Illum_amp_f_Ad1	AATGATACGGCGACCACCGAGATCTACACCAATGGAGCTCGC AATACAGAGTTTACCGCATT
	Illum_amp_f_Ad2	AATGATACGGCGACCACCGAGATCTACAC ATCGAGAGCCTCTTCCGCATCTCACAACCTACT
	Illum_amp_r	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTG CTGAACCGCTCTTCCGATCT
Oligos used in sequencing	Ad1_seq	CAATGGAGCTCGCAATACAGAGTTTACCGCATT
	Ad2_seq	ATCGAGAGCCTCTTCCGCATCTCACAACCTACT
	Illum_seq_r	CGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT

Oligos were obtained from Integrated DNA Technologies. An asterisk indicates a phosphorothioate bond. /5Phos/ indicates a five-prime phosphate modification.

Supplementary Note 1. The importance of a complexity bottleneck

A complexity bottleneck is needed so that multiple overlapping, randomly positioned breakpoint reads can be observed for each member of the long fragment library with a reasonable amount of sequencing. In other words, it is necessary to sample each nested sub-library in sufficient depth to reconstruct the sequence of the parent long molecule. For example, we obtained approximately 60 million read-pairs across six lanes of Illumina sequencing that enabled us to reconstruct part or all of the sequence of approximately one million long molecules. If we had used a library containing 100 million long molecules, we would only have observed, on average, less than one read-pair per long molecule, preventing any subassembly from taking place within most if not all sub-libraries. The only exception to this principle is in the case of a very small effective genome size and if the ends of molecules (and not degenerate synthetic adaptors) are used as tags. For example, in the case of a genome of only 500 kilobases, the maximum number of unique tag reads (assuming no repetitive sequence at the scale of the tag read) is one million, which we have shown to be a tractable library complexity. In such a situation, it might not be formally necessary to restrict library complexity.

Supplementary Note 2. Filtering of predicted misassemblies

Manual inspection of predicted misassemblies revealed four contigs that were incorrectly called misassemblies because of differences between the strain that we sequenced and the reference PAO1 strain. Three of these (2548, 2129 and 2115) exhibited extremely high sequence identity with a phage-like insertion in PAO1 that was recently added to GenBank (ID GQ141978.1) and that we have observed in independent shotgun sequencing data from our strain. Notably, the same phage-like insertion seems to have caused the lone misassembly in our scaffolds (Scaffold_LR7_3). The fourth contig (2622) spans a ~1 kb deletion in our strain that we have also observed in independent shotgun sequencing data (data not shown).

Supplementary Note 3. Comparison of *de novo* assembly to hybrid 454-Illumina approach

We compared the performance of our method to a recently published, high quality *de novo* assembly from a similar but significantly lower (G+C) content organism (66.6% versus 58.5%), which was generated by combining both long-read and long-range mate-paired 454 data with short distance paired-end Illumina data[1]. We find that our method compares very favorably to that approach with respect to N50 (445 kb versus 92 kb), longest scaffold (915 kb versus 389 kb), substitution error rate (~1/14,000 versus ~1/7,000), and number of rearrangements (one versus twenty). It should be noted that the authors of that study also performed sequencing and assembly of a related organism without a reference genome and achieved apparently better performance (N50 of 532 kb, longest contig of 794 kb), which they attempted to validate with limited Sanger sequencing. However, it is difficult to make a direct comparison with respect to accuracy in the absence of a reference genome. Our method also used significantly more raw data than that study, but only required a single sequencing platform, which may increase its general utility.

Supplementary Note 4. Estimated cost of subassembly protocol

Although it is difficult to draw firm conclusions in the face of rapidly changing costs associated with many second-generation sequencing platforms, it is clear that subassembly is significantly more expensive than standard shotgun Illumina sequencing if only the total amount of sequence produced is considered. However, as subassembly produces much longer reads at much higher per-base accuracy than the raw reads from the Illumina platform, such a comparison is not valid. Even the comparison to Roche/454 sequencing, which produced reads in the hundreds of base-pairs, is difficult because of the decreased accuracy of that method relative to the method we present here. Still, we estimate that our method is roughly cost-comparable to Roche/454 sequencing. For example, if a lane of sequencing is assumed to cost ~$2,000, from six lanes of sequencing we generated 405 Mb of long SA reads for the *P. aeruginosa* sample, which corresponds to a cost of ~$30/Mb, or about half that of recently published estimates of the cost of Roche/454 Sequencing (Shendure and Ji, 2008). However, the reduced error rate is a critical differentiator, making the cost comparison tenuous. A major advantage of subassembly is that a very low error rate and long effective read length is maintained independent of sample complexity. In the case of short read sequencing (Illumina, AB SOLiD, Helicos), read length and error limitations can be overcome through the use of very high coverage. The ability to achieve high coverage depends implicity on sample complexity and can be complicated by relatedness of sequences therein. With Roche/454 sequencing, read lengths are longer, but once again, errors can only be overcome with high coverage, which again may be impossible in the case of either very high sample complexity or the presence of highly related sequences. We therefore conclude that subassembly produces equivalently long sequences at below or equal to the cost of Roche/454 sequencing with length and error performance that remains independent of sample complexity and sequence relatedness, a feature of no other currently available second-generation sequencing method.

Supplementary Protocol 1. General subassembly protocol for library construction, sequencing, and data analysis

The following protocol details the series of steps and timing needed to construct, sequence, and analyze a subassembly library. A brief discussion of cost can be found in Supplementary Note 3. For more detailed information about reagents used and reaction conditions, see above in the section entitled Supplementary Protocol 2. Note: Sample is stable and can be stored at -20'C indefinitely after the column purification component of any step.

Overview:

1. Fragmentation of source DNA (45 m)
2. End repair (1 h)
3. Size selection to define long fragment size (4 h)
4. A-tailing (30 m)
5. Ligation to tag-adjacent adaptors (30 m)
6. Size selection of ligated fragments to remove free adaptor (4 h)
7. PCR amplification to impose bottleneck (3 h)
8. PCR amplification to prepare merging library (3 h; Optional)
9. Blunt ligation of PCR products (30 m)
10. Fragmentation of high molecular weight concatemers (30 m)
11. End repair (1 h)
12. A-tailing (30 m)
13. Ligation to breakpoint-adjacent adaptor (30 m)
14. PCR amplification (3 h)
15. Size selection to define distribution of breakpoint reads across long fragments (4 h)
16. Illumina sequencing (~7 days)
17. Grouping and parallelized assembly (2-7 days)

1. Fragmentation of source DNA (45 m)

Use a Bioruptor sonicator to fragment 1-5 ug of high-molecular weight DNA. Perform two fifteen minute cycles of 30 second intervals at high power (Supplementary Protocol 3). Recover sheared DNA using QIAquick PCR purification kit (Supplementary Protocol 3).

2. End repair (1 h)

Use the Epicentre Biosciences End-It DNA End Repair Kit to polish fragment ends according to manufacturer's specifications (Supplementary Protocol 3). Purify end-polished DNA using QIAquick PCR purification kit.

3. Size selection to define long fragment size (4 h)

Perform PAGE-based size-selection (Supplementary Protocol 3) and QIAQuick purification of sheared, end-repaired DNA to define the size of long fragments that will later be subassembled. Fragments as short as ~300 bp or as long as 1-2 kb can be used; a relatively narrow size distribution (+/- 5%) should be selected at this stage to facilitate downstream size-selections.

4. A-tailing (30 m)

Perform A-tailing reaction (Supplementary Protocol 3) to facilitate downstream ligation of tag-adjacent adaptors. Purify A-tailing reaction with QIAQuick kit.

5. Ligation to tag-adjacent adaptors (30 m)

Ligate annealed tag-adjacent adaptors (Ad1+Ad1_rc and Ad2+Ad_rc, Supplementary Table 4) to sheared, end-repaired, size-selected, A-tailed long fragment library to facilitate downstream PCR bottleneck. For details about this step, see step 6 of Supplementary Protocol 2.

6. Size selection of ligated fragments to remove free adaptor (4 h)
Perform PAGE-based size-selection (Supplementary Protocol 3) and QIAQuick purification of adaptor-ligated fragments to remove free adaptor and adaptor dimers that will disrupt PCR amplification. The size-range should be similar to the range selected in Step 3, with ~50 bp added to account for the addition of adaptors.

7. PCR amplification to impose bottleneck (3 h)
Real-time PCR amplification of a serial dilution of adaptor-ligated fragments should be performed. For a discussion of the significance of the complexity bottleneck, see Supplementary Note 1. The molarity of the adaptor-ligated, size-selected fragments can be estimated with a Qubit fluorometer or a Nanodrop (Supplementary Protocol 3), but, because ligation efficiency is low and unpredictable, a range of input molecule concentrations should be used. For example, if the desired library complexity is ~1 million long fragments, PCR amplifications should be performed containing 1, 5, 10, 50 and 100 million molecules as estimated by fluorometry or UV absorbance. Effective library complexity should then be estimated using the amplification profile from the real-time PCR instrument and gel electrophoresis. For details about PCR amplification, see step 8a of Supplementary Protocol 2.

NOTE: It may be helpful to perform an additional PCR reaction using the product of this PCR reaction that has the desired complexity in order to produce sufficient material for downstream processing **(+3 h)**. The product should be divided across 8 PCR reactions. For details about this PCR reaction, see step 8c of Supplementary Protocol 2.

8. PCR amplification to prepare merging library (3 h; Optional)
Real-time PCR amplification of the PCR product from step 7 should be performed if TDRG merging capability is desired. For details about this PCR reaction, see step 8d of Supplementary Protocol 2.

9. Blunt ligation of PCR products (30 m)
Concatemerize PCR products from step 7 to generate a high-molecular weight sample for subsequent shearing. For details about this step, see step 9 of Supplementary Protocol 2.

10. Fragmentation of high molecular weight concatemers (30 m)
Shear high-molecular weight concatemers from step 9 using a Bioruptor sonicator (Supplementary Protocol 3).

11. End repair (1 h)
Use the Epicentre Biosciences End-It DNA End Repair Kit to polish fragment ends according to manufacturer's specifications (Supplementary Protocol 3). Purify end-polished DNA using QIAquick PCR purification kit.

12. A-tailing (30 m)
Perform A-tailing reaction (Supplementary Protocol 3) to facilitate downstream ligation of the breakpoint-adjacent adaptor. Purify A-tailing reaction with QIAQuick kit.

13. Ligation to breakpoint-adjacent adaptor (30 m)

Ligate annealed breakpoint-adjacent adaptors (Illum_rev,Illum_rev_rc, see Supplementary Table 4) to sheared, end-repaired, A-tailed fragments to facilitate downstream PCR. For details about this step, see step 13 of Supplementary Protocol 2.

14. PCR amplification (3 h)
Perform PCR amplification of adaptor-ligated fragments from step 13 using two sets of primers as described in step 14 of Supplementary Protocol 2.

15. Size selection to define distribution of breakpoint reads across long fragments (4 h)
Perform PAGE-based size-selection (Supplementary Protocol 3) and QIAQuick purification of PCR product to define the distribution of breakpoint reads across the original long fragments (see Supplementary Figure 2). If shorter fragments were used (<=500 bp), it is likely that only one size-range will be needed to achieve a sufficient distribution of breakpoint reads (~250 bp – fragment length). If especially long fragments were selected in step 3 (>500 bp), it may be necessary to perform multiple size-selections and sequence these in separate lanes of the Illumina flow-cell. Because of biases during cluster generation, short fragments are preferentially sequenced within a given lane, so that it is necessary to segregate fragments based on size to achieve a more uniform distribution (**Supplementary Fig. 2**).

16. Illumina sequencing (~7 days)
Perform paired-end Illumina sequencing of the libraries from step 15 (20x76 bp reads) and step 6 (36x36 bp reads) according to manufacturer's specifications. For more details, see step 16 of Supplementary Protocol 2.

17. Grouping and parallelized assembly (2-7 days)
Perform error-correction of tag reads and computational grouping of reads based on the identity of tag reads as described in the Online Methods. After read-grouping, perform parallelized assembly. The time to perform tag read error correction scales $O(N^2)$ with the number of tag reads, as an all-by-all comparison is made to collapse tag reads differing by only one base. Assembly scales roughly linearly with the number of TDRGs, and was observed to require approximately 12 hours per 100,000 TDRGs when parallelized across 13 CPUs. After assembly, perform consensus quality score estimation and tag- and adaptor-masking. All software available upon request.

Supplementary Protocol 2. Detailed description of methods used for library construction

Note: Shotgun libraries from PAO1 and from metagenomic samples were constructed according to standard Illumina protocols (details available upon request).

Overview of subassembly library production strategy

1. Isolation of source DNA
2. Fragmentation of source DNA
3. End repair
4. Size selection
5. A-tailing
6. Ligation to tag-adjacent adaptors
7. Size selection of ligated fragments
8. PCR amplification
9. Blunt ligation of PCR products
10. Fragmentation of high molecular weight concatemers
11. End repair
12. A-tailing
13. Ligation to breakpoint-adjacent adaptor
14. PCR amplification
15. Size selection
16. Illumina sequencing

1. Isolation of source DNA

Genomic DNA from *Pseudomonas aeruginosa* PAO1 was generously provided by Colin Manoil.

Metagenomic source DNA, generously provided by Ludmila Chistoserdova, was isolated from a microbial population that was obtained from sediment 63 m below the surface of Lake Washington and subsequently enriched using Stable Isotope Probing for organisms that utilized methylamine as a food source.

2. Fragmentation of source DNA

Pseudomonas: ~2 ug of genomic DNA was randomly fragmented using nebulization, as described in Supplementary Protocol 3.

Metagenomic: ~2 ug of metagenomic source DNA was randomly fragmented using a Bioruptor, as described in Supplementary Protocol 3.

3. End repair

Fragmented template was end-repaired with the Epicentre Biosciences End-It DNA End Repair Kit as described in Supplementary Protocol 3. The end-repaired mixture was purified and eluted in 30 µL Buffer EB by QIAGEN QIAquick column.

4. Size selection

500-600 bp fragments (Pseudomonas) and 400-500 bp fragments (Metagenomic) of sheared DNA were selected by 6% TBE gel electrophoresis and recovered by ethanol precipitation as described in Supplementary Protocol 3.

5. A-tailing

Terminal 3' adenosines were added to size-selected DNA as described in Supplementary Protocol 3 to allow ligation to the T-tailed adaptors. A-tailed DNA was purified by QiaQuick column and eluted in 50 uL of Buffer EB.

6. Ligation to tag-adjacent adaptors

50 uM tag-adjacent adaptors were prepared by mixing equal volumes of Ad1 with Ad1_rc and Ad2 with Ad2_rc (initially diluted to 100 uM), heating to 95°C, then turning off the thermal cycler block and cooling passively to room temperature.

Genomic fragments were quantified using a Qubit fluorometer (Invitrogen, Q32857) and the Quant-IT dsDNA HS kit (Invitrogen, Q32854). Fragments were ligated to adaptors using the Quick Ligation Kit (NEB, M2200) at a molar ratio of 1:10 as follows.

	Pseudomonas	Metagenomic
Genomic fragments	13 uL (~1 ng/uL → 36 fm)	13 uL (0.25 ng/uL → 9.1 fm)
Annealed adaptor	1.44 uL (500 nM → 720 fm)	1.8 uL (100 nM → 180 fm)
dH2O	0.56 uL	0.2 uL
Quick Ligation buffer (2x→1x)	15 uL	15 uL
Quick Ligase	1.5 uL	1.5 uL

-All components were mixed by brief vortexing and centrifugation.
-The reaction was carried out at room temperature for 15 minutes.
-The reaction was stored on ice.

7. Size selection

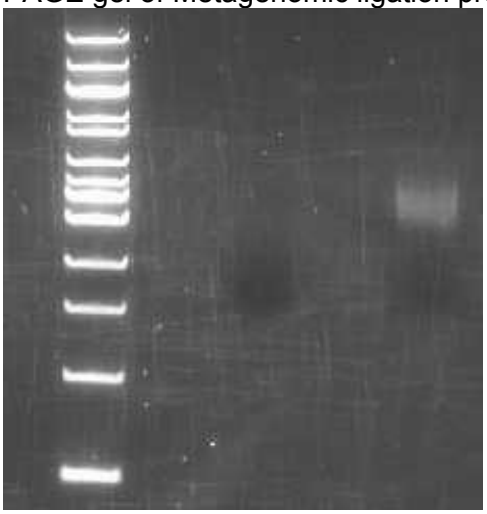
To remove excess unligated adapter, 400-800 bp fragments of ligated DNA were selected by 6% TBE gel electrophoresis and recovered by ethanol precipitation as described in Supplementary Protocol 3.

PAGE gel of Pseudomonas ligation product:

Left: NEB 100 bp ladder

Right: Pseudomonas ligation products

PAGE gel of Metagenomic ligation product:



Left: NEB 100 bp ladder
Right: Metagenomic ligation products

8.a. PCR amplification

To impose a complexity bottleneck and generate multiple copies of genomic fragments, quantitative real-time PCR amplification was performed using Phusion Hot-Start polymerase (Finnzymes, F-540S) and SYBR Green (Invitrogen, S-7563) in a Bio-Rad MiniOpticon thermal cycler. Five-prime phosphorylated primers and the Pfu polymerase were used to facilitate concatemerization in the next step.

Complexity was limited by serially diluting the DNA recovered from size selection. For the Pseudomonas sample, undiluted, 10-fold, and 100-fold diluted samples were subjected to PCR. Amplification of the 100-fold dilution was split across ten reactions, each containing a 1000-fold dilution, to improve yield. Because of the lower concentration of the Metagenomic sample during ligation, PCR was performed with both 1 uL (+ 9 uL H2O, “1X”) and 10 uL (“10X”) of the adaptor-ligated, size-selected fragments. A given dilution was chosen for further processing based on an assessment of the gel. In general, the least complex sample that did not demonstrate banding on the gel was chosen. Alternatively, a sequencing library can be produced as in 8.d. and sequenced on one lane of a standard paired-end 36 bp to estimate complexity.

Care was taken to ensure that reactions were removed from the thermal cycler prior to the completion of log-phase amplification, since “over-amplification” results in aberrantly slow gel migration of small fragments that will contaminate downstream size-selections.

	Pseudomonas	Metagenomic
Template	1	10
Phusion HF Buffer (5x → 1x)	10	10
dNTPs (25 mM → 200 uM)	0.4	0.4
SYBR Green I (1x → 0.1x)	5	5
Ad1_amp (10 uM → 500 nM)	2.5	2.5
Ad2_amp (10 uM → 500 nM)	2.5	2.5
dH2O	28.1	19.1

Phusion Hot-Start polymerase	0.5	0.5

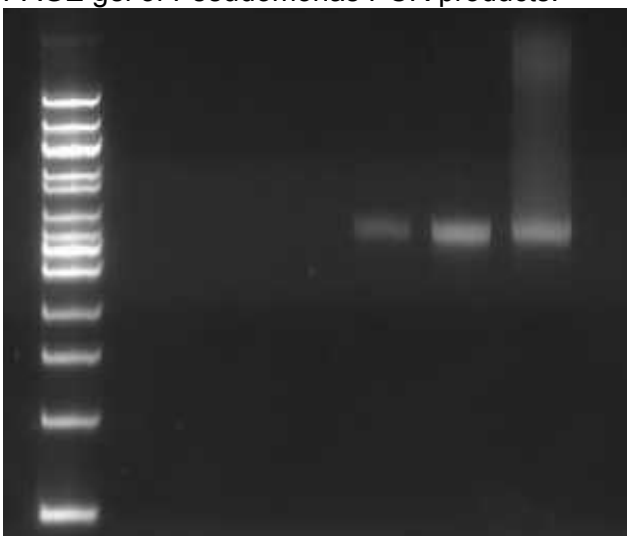
-All components were mixed by brief vortexing and centrifugation.
-Thermal cycling in a Bio-Rad MiniOpticon was performed as follows:

1. 98°C for 30 sec
2. 98°C for 10 sec
3. 60°C for 30 sec
4. 72°C for 50 sec
5. plate read
6. 72°C for 10 sec
7. go to 2, 24 times
8. 72°C for 5 mins
9. hold 16°C

-Reactions were removed from the cycler as soon as log phase amplification appeared to be ending.
-Reactions were stored at 4°C.

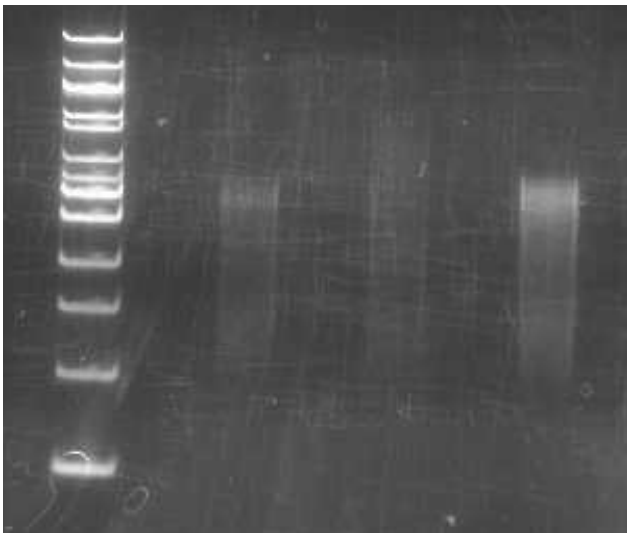
PCR reactions were purified by QiaQuick column and eluted in 30 uL of Buffer EB. For the 100-fold dilution sample, reactions were pooled prior to purification.

PAGE gel of Pseudomonas PCR products:



1. NEB 100 bp ladder
5. 100-fold dilution
6. 10-fold dilution
7. undiluted

PAGE gel of Metagenomics PCR products:

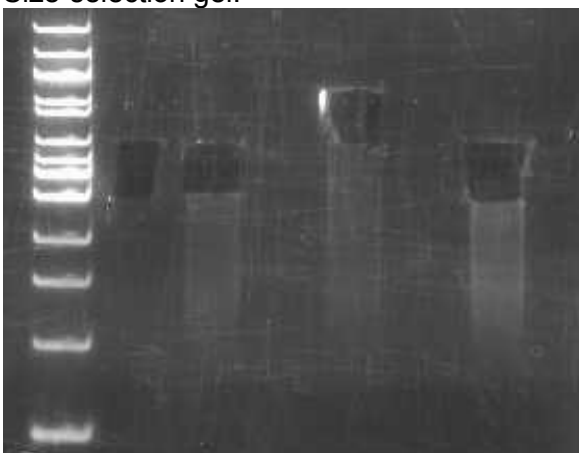


1. NEB 100 bp ladder
3. 1X PCR
7. 10X PCR

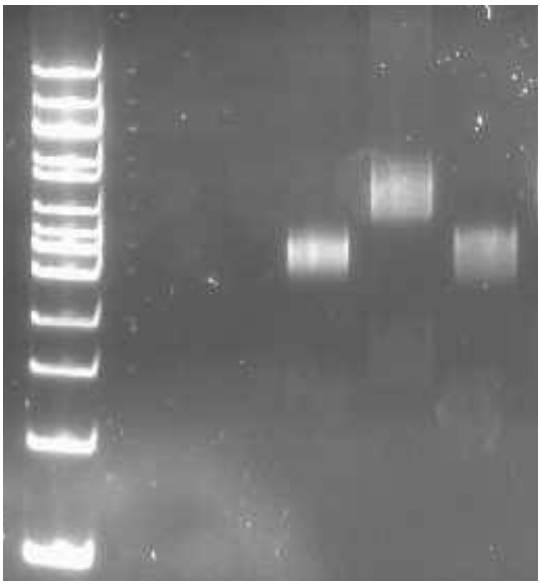
8.b. Size selection of Metagenomic PCR products

Because of length heterogeneity in the PCR products of the Metagenomic library, and to maintain a long population of fragments, the purified PCR products were again size-selected from 500-600 bp as described in Supplementary Protocol 3, then amplified as in step 8.a.

Size-selection gel:



PCR gel:

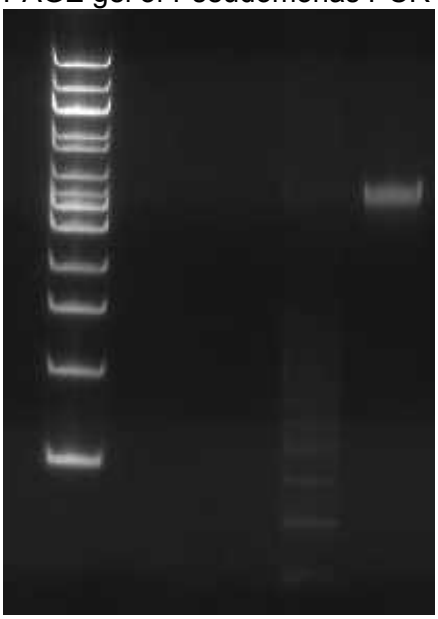


1. NEB 100 bp ladder
4. 1X PCR
6. 10X PCR

8.c. Reamplification to produce adequate template for downstream steps

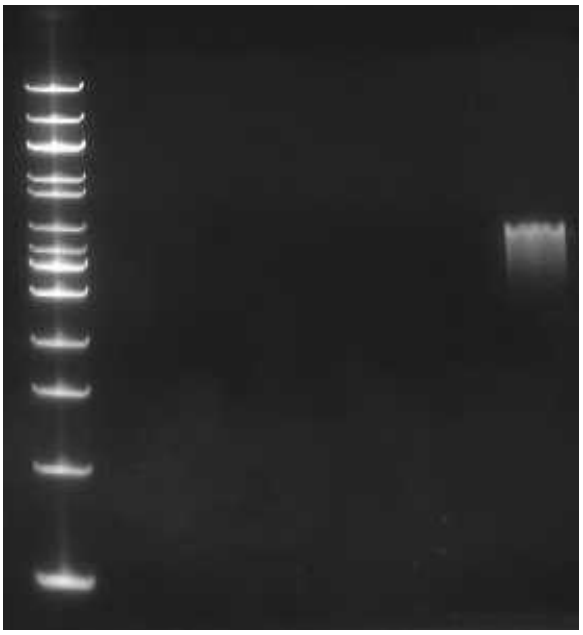
To produce sufficient material to avoid a complexity bottleneck in subsequent steps, 1 uL (Pseudomonas) or 10 uL (Metagenomic) of the above PCR product (after step 8.b. for the Metagenomic sample) was split across eight PCR reactions and amplified again as above, then pooled and purified as above.

PAGE gel of Pseudomonas PCR product:



1. NEB 100 bp ladder
5. Bulk amplification of 100-fold dilution

PAGE gel of Metagenomics PCR products:



1. NEB 100 bp ladder
6. Bulk amplification of 10-fold concentrated sample (10 uL into PCR in step 8.a.)

8.d. PCR of bottlenecked fragment library for paired-end sequencing

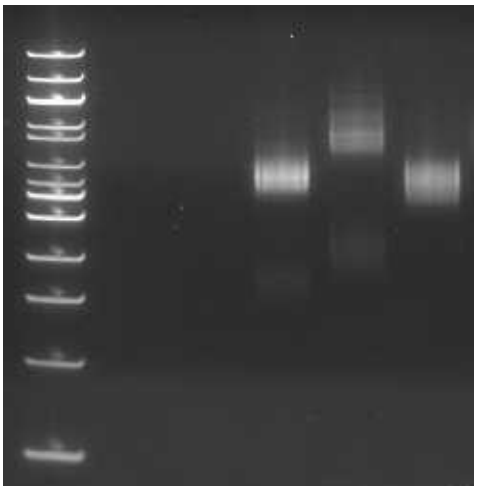
To enable pairing of TDRGs from opposite ends of the same original fragment, Metagenomic PCR products from step 8.c. were amplified with oligos that encoded compatibility with the Illumina flowcell, using iProof HF Master Mix (Bio-Rad #172-5311) in a Bio-Rad MiniOpticon thermal cycler as below.

	Metagenomic
Template	1
SYBR Green I (1x → 0.1x)	5
Illum_amp_f_Ad1 (10 uM → 500 nM)	2.5
Illum_amp_r_Ad2 (10 uM → 500 nM)	2.5
dH2O	14
iProof HF master mix (2x → 1x)	25

-All components were mixed by brief vortexing and centrifugation.
-Thermal cycling in a Bio-Rad MiniOpticon was performed as follows:

1. 98'C, 30 sec
2. 98'C, 10 sec
3. 58'C, 15 sec
4. 72'C, 15 sec
5. plate read
6. 72'C, 5 sec
7. go to 2, 29 times
8. 72'C, 10 mins
9. 16'C for ever

PAGE gel of Metagenomic TDRG merging libraries:



1. NEB 100 bp ladder
4. 1X PCR
6. 10X PCR

Sequencing of the TDRG merging library was performed on an Illumina GA-II with 36 bp paired-end reads according to manufacturer's specifications, except that the following oligos were used: Ad1_seq for the first read and Ad2_seq for the second read.

9. Blunt ligation of PCR products

To generate high molecular weight concatemers of PCR products, blunt ligation was performed using the Quick Ligation Kit (NEB, M2200) as below. Reaction components were mixed by brief vortexing and centrifugation, the reaction was carried out at room temperature for 15 minutes, and then stored at 4°C.

	Pseudomonas	Metagenomic
Sample	30 uL (of 30 uL QiaQuick elution)	25 uL (of 50 uL QiaQuick elution)
Quick Ligation Buffer (2x → 1x)	30 uL	25 uL
Quick Ligase Enzyme	6 uL	5 uL

10. Fragmentation of high molecular weight concatemers

PCR product ligations were randomly fragmented using the Bioruptor, as described in Supplementary Protocol 3.

11. End repair

Fragmented template was end-repaired with the Epicentre Biosciences End-It DNA End Repair Kit as described in Supplementary Protocol 3. The end-repaired mixture was purified and eluted in 30 µL Buffer EB by QIAGEN QIAquick column.

12. A-tailing

Terminal 3' adenosines were added to end repaired DNA as described in Supplementary Protocol 3 to allow ligation to the T-tailed adaptors. A-tailed DNA was purified by QiaQuick column and eluted in 50 uL of Buffer EB.

13. Ligation to breakpoint-adjacent adaptor

50 uM breakpoint-adjacent adaptor was prepared by mixing equal volumes of Illum_rev and Illum_rev_rc (initially diluted to 100 uM), heating to 95°C, then turning off the thermal cycler block and cooling passively to room temperature.

Fragments were quantified using a Qubit fluorometer (Invitrogen, Q32857) and the Quant-IT dsDNA HS kit (Invitrogen, Q32854). Fragments derived from the Pseudomonas PCR were quantified at 20 femtomoles/microliter; A-tailed Metagenomic fragments were quantified at 9 femtomoles/microliter. Fragments were ligated to the Illumina reverse adaptors using the Quick Ligation Kit (NEB, M2200) at a molar ratio of 1:20 as follows.

	Pseudomonas	Metagenomic
Template	10 uL	11 uL
Adaptor	1.14 uL (@ 5 uM)	4 uL (@ 500 nM)
Quick Ligation Buffer (2x → 1x)	15 uL	15 uL
dH2O	1.16 uL	0
Quick Ligase	1.5 uL	1.5 uL

-All components were mixed by brief vortexing and centrifugation.
-The reaction was carried out at room temperature for 15 minutes.
-The reaction was stored on ice.

Ligated DNA was purified by QiaQuick column and eluted in 30 uL of Buffer EB.

14. PCR amplification

To prepare molecules for Illumina paired-end sequencing, adaptor-ligated DNA was subjected to real-time quantitative PCR amplification using Phusion Hot-Start polymerase (Finnzymes, F-540S) and SYBR Green (Invitrogen, S-7563) in a Bio-Rad MiniOpticon thermal cycler. Each sample was amplified in two separate reactions using different pairs of primers to enable amplification of fragments containing sequence from each end of the original fragment.

After amplification, size-selection and PCR was performed to enrich for fragments that contained a random break-point at least 150-300 bp distal to the tag read, as shorter fragments will outcompete for cluster formation on the flowcell and dominate sequencing. For this reason, real-time monitoring of amplification is essential to prevent overamplification, which results in aberrant migration of the PCR products on the gel and interferes with downstream size-selection. Care should be taken to ensure that PCR is stopped while the reaction is still in log phase.

The first primer in the mixture below was always Illum_amp_r, while the second primer was Illum_amp_f_Ad1 in one reaction and Illum_amp_f_Ad2 in the other. Four reactions were performed for each primer combination, using in total 10 uL of the 30 uL eluate from the adaptor ligation.

	Pseudomonas	Metagenomic
Template	1.25	1.25
Phusion HF Buffer (5x → 1x)	10	10
dNTPs (25 mM → 200 uM)	0.4	0.4
SYBR Green I	5 (1x → 0.1x)	2.5 (10x → 0.5x)
Illum_amp_r (10 uM → 500 nM)	2.5	2.5

Illum_amp_f_Ad* (10 uM → 500 nM)	2.5	2.5
dH2O	27.85	30.35
Phusion Hot-Start polymerase	0.5	0.5

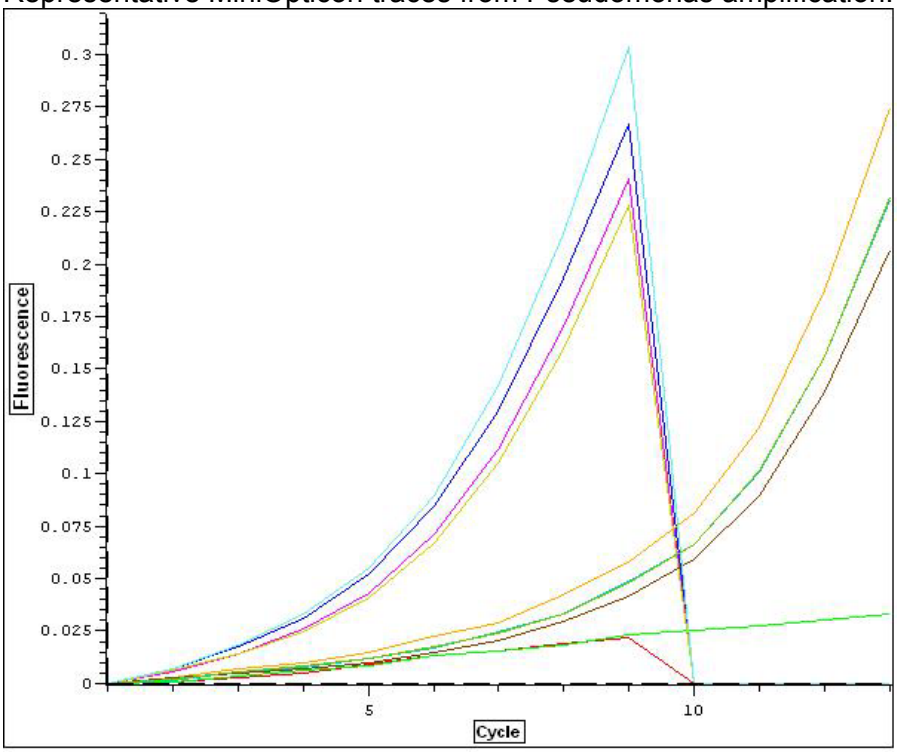
-All components were mixed by brief vortexing and centrifugation.
-Thermal cycling in a Bio-Rad MiniOpticon was performed as follows:

1. 98'C, 30 sec
2. 98'C, 10 sec
3. 58'C, 15 sec
4. 72'C, 50 sec
5. plate read
6. 72'C, 15 sec
7. go to 2, 39 times

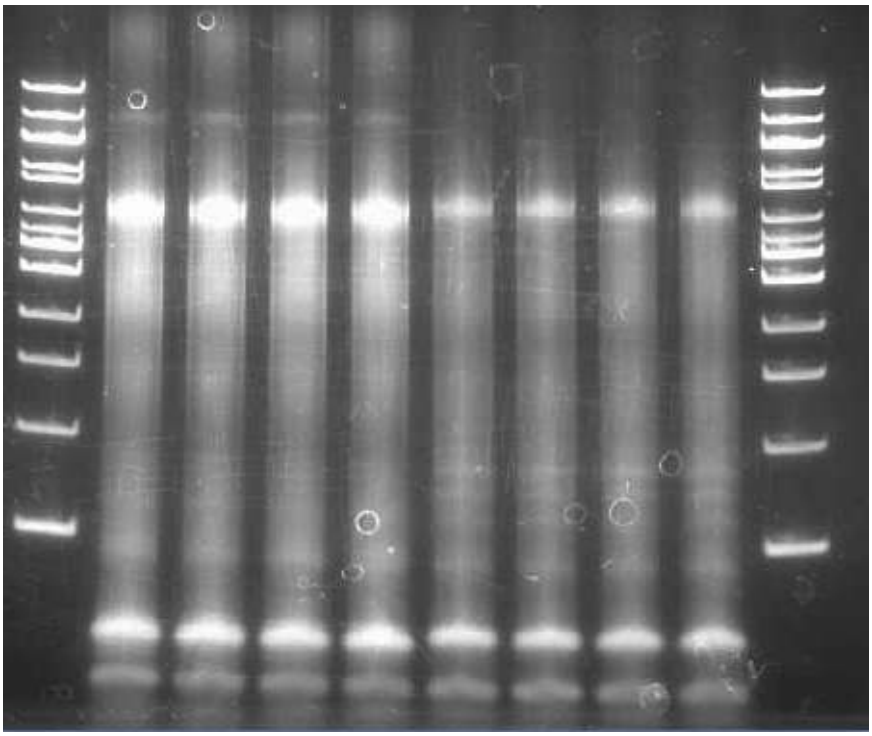
-Reactions were removed from the cycler as soon as log phase amplification appeared to be proceeding robustly.
-Reactions were stored at 4°C.

PCR reactions were purified by QiaQuick column and eluted in 30 uL of Buffer EB.

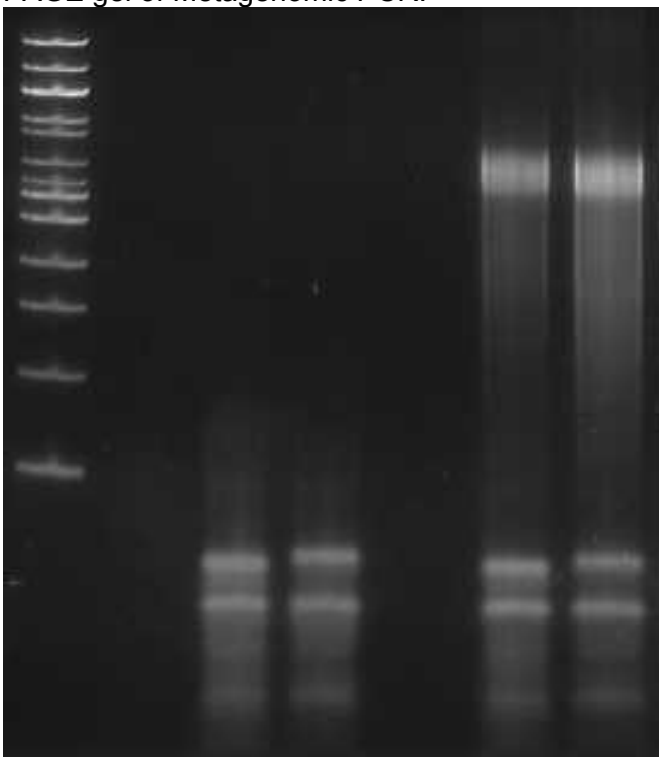
Representative MiniOpticon traces from Pseudomonas amplification:



PAGE gel of Pseudomonas PCR:



PAGE gel of Metagenomic PCR:

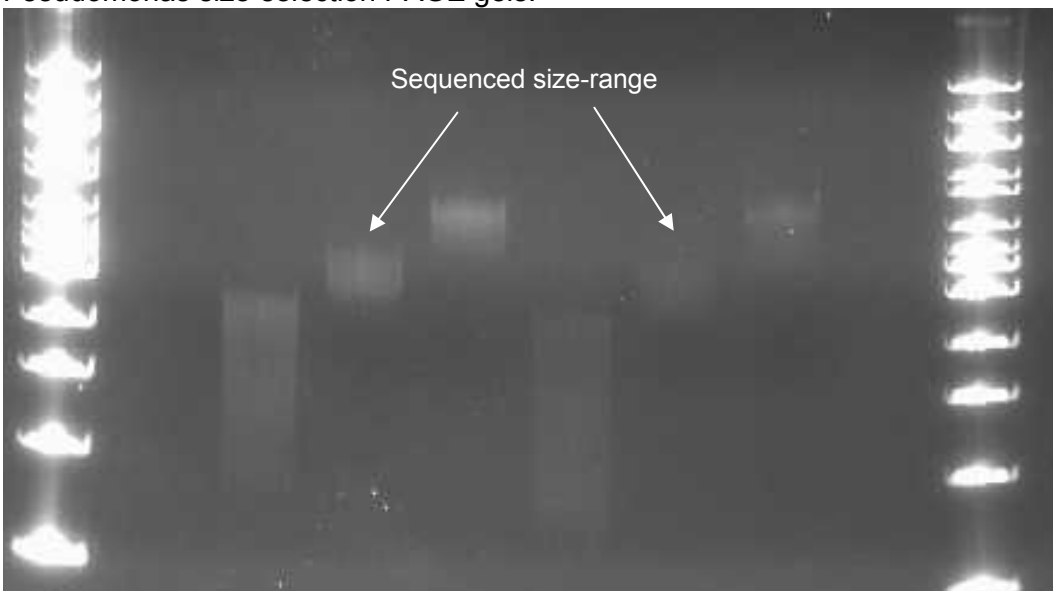


15. Size selection

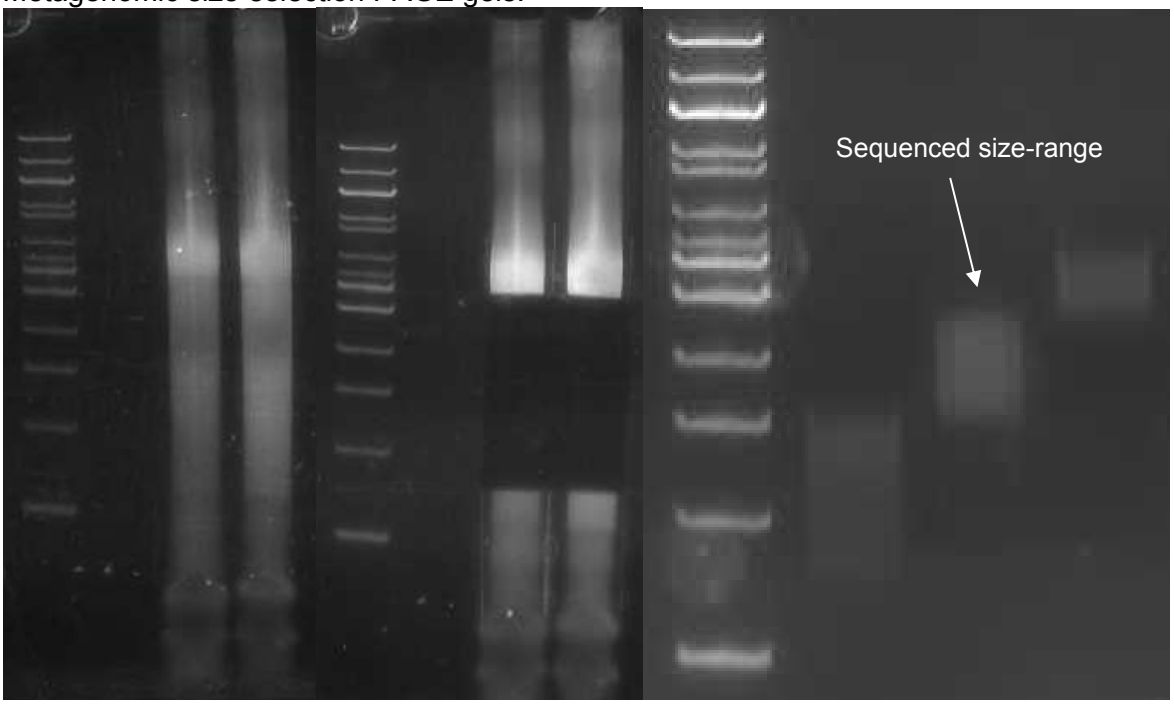
Amplified template was size-selected to ranges of 450-600 bp (Pseudomonas) and 300-450 bp (Metagenomic) as described in Supplementary Protocol 3; removal of short fragments improves

cluster formation uniformity on the flowcell and improves the distribution of reads across the original fragments **(Supplementary Figure 2)**.

Pseudomonas size-selection PAGE gels:



Metagenomic size-selection PAGE gels:



Following size-selection, a final PCR was performed as below to obtain adequate material for Illumina paired-end sequencing.

	Pseudomonas	Metagenomic
Template	5	10

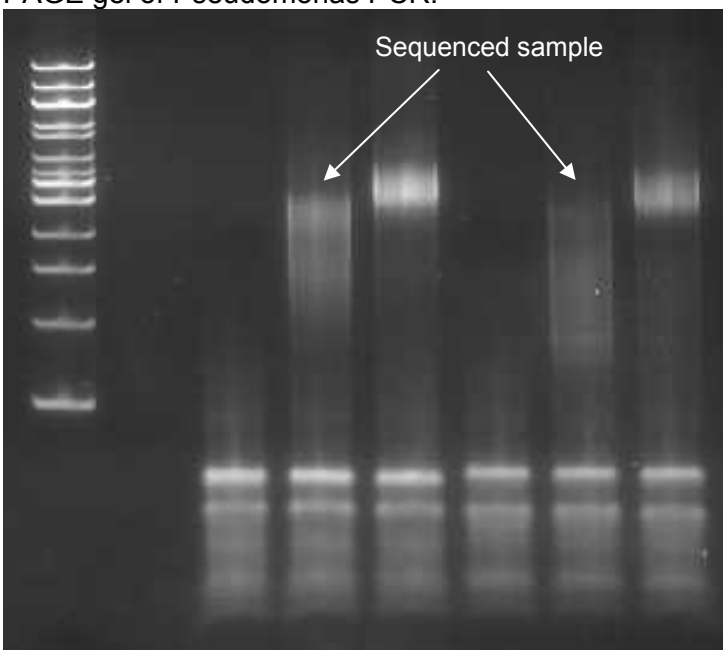
Phusion HF Buffer (5x → 1x)	10	10
dNTPs (25 mM → 200 uM)	0.4	0.4
SYBR Green I (10x → 0.5x)	2.5	2.5
Illum_amp_r (10 uM → 500 nM)	2.5	2.5
Illum_amp_f_Ad* (10 uM → 500 nM)	2.5	2.5
dH2O	26.6	21.4
Phusion Hot-Start polymerase	0.5	0.5

-All components were mixed by brief vortexing and centrifugation.
-Thermal cycling in a Bio-Rad MiniOpticon was performed as follows:

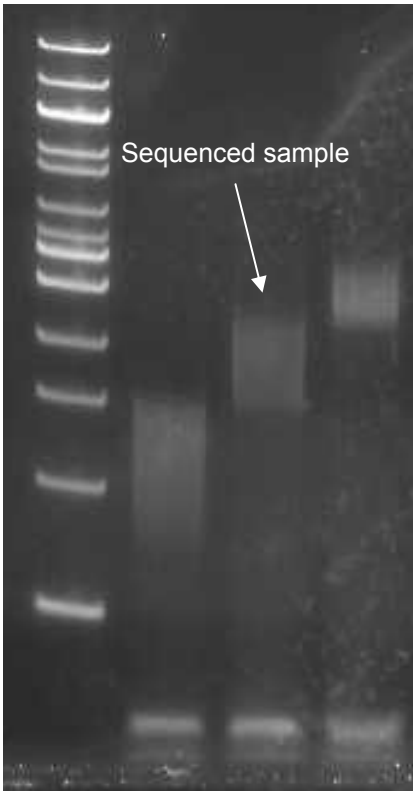
1. 98'C, 30 sec
2. 98'C, 10 sec
3. 58'C, 15 sec
4. 72'C, 50 sec
5. plate read
6. 72'C, 15 sec
7. go to 2, 39 times

PCR reactions were purified by QiaQuick column and eluted in 30 uL of Buffer EB.

PAGE gel of Pseudomonas PCR:



PAGE gel of Metagenomic PCR:



16. Illumina sequencing
After PCR and QiaQuick cleanup, amplicons from the desired size range (450-600 bp for Pseudomonas, 300-450 bp for Metagenomic) were subjected to paired-end Illumina sequencing according to manufacturer's specifications for a 20 bp first read and a 76 bp second read, except that we used the following sequencing oligos: Ad1_seq and Ad2_seq on the first read, and Illum_seq_r on the second read (Supplementary Table 4).

Supplementary Protocol 3. Common techniques

Nebulization

High molecular weight DNA was diluted to 50 µL in TE Buffer, pH 7.5-8 before being added to the 40% glycerol nebulizing solution.

- x µL template
- 50 – x µL TE Buffer, pH 7.5
- 325 µL EB
- 375 µL 80% glycerol

The nebulizing mixture was pipetted to the bottom of the Invitrogen Nebulizer (45-0072). The lid was tightly closed and wrapped with ParaFilm to limit sample loss. Nebulizing was performed on ice for 15-90 seconds with 6 psi pressurized air. A slow centrifuge spun down the sample mixture, which was collected by pipette. Repetitive centrifugation/collection was necessary to ensure adequate recovery. DNA was purified using QIAquick columns and eluted in 30 µL Buffer EB.

Bioruptor

High molecular weight DNA was placed in a 1.6 µL Eppendorf and diluted to 300 µL in TE. The sample was sheared in the Bioruptor (Diagenode) for 8x15 minute cycles, with 30 second sonication intervals at high power. DNA was purified using QIAquick columns and eluted in 30 µL Buffer EB.

QIAquick Purification

QIAGEN QIAquick PCR Purification Kits (28106 and 28304) were used for purification of library components. Purification by QIAGEN QIAquick columns occurred as follows.

- x µL sample + 5x µL Buffer PBI.
- Load 600 µL of sample/PBI mixture to column.
- Spin down 30 seconds, discard eluate.
- Repeat until entire sample/PBI mixture is loaded.
- Load 750 µL Buffer PI to column.
- Spin down 30 seconds, discard eluate.
- Spin down 120 seconds, discard eluate.
- Replace round-bottom tube with clean 1.6 mL Eppendorf tube.
- Load desired elution volume of Buffer EB to column.
- Wait 5 minutes.
- Spin down 120 seconds.
- Collect purified product in eluate.

End repair

The Epicentre Biosciences End-It DNA End Repair Kit (ERK-70823) was used to produce blunt-end fragments, following the manufacturer's directions as shown below. The kit has a maximum input DNA capacity of 5 µg; for reactions containing >5 µg, the reaction was increased proportionally.

- x µL DNA
- 34 - x µL water
- 5 µL 10x End_It buffer (10x → 1x)
- 5 µL 10x dNTP (10x → 1x)
- 5µL 10x ATP (10x → 1x)

- 1 µL End-It enzyme

All of the reactants except the enzyme were combined, then briefly vortexed and centrifuged. The enzyme was added, and the reaction was carried out in a PCR tube at room temperature for 45 minutes.

A-tailing
This reaction added dATP to each 3'-end of the library fragments, and requires template with blunt ends.

- x µL DNA
- 89 - x µL water
- 10 µL 10x PCR buffer w/ Mg2+ (10x → 1x) (Invitrogen)
- 0.5 µL 100 mM dATP (100 mM → 0.5 mM) (Invitrogen)
- 0.5 µL Taq (5U/µL → 0.025 U/µL) (Invitrogen)

All of the reactants except the enzyme were combined, then briefly vortexed and centrifuged. The enzyme was added, and reactions were carried out at 70°C for 20 minutes.

Size selection by gel electrophoresis
Samples were loaded with 1x BlueJuice loading dye (Invitrogen), and separated at 180V. DNA was visualized with Sybr Gold. Each size range was collected with fresh razor blades and plasticware. The size range of interest was excised by razor blade and transferred to a siliconized 0.6 mL Eppendorf tube with a hole punched in the bottom. The 0.6 mL tube was placed inside a siliconized 1.6 mL Eppendorf tube and centrifuged in a tabletop microcentrifuge at 13,200 rpm for 5 minutes. Following centrifugation, the 0.6 mL tube was discarded, and the gel slurry dissolved in 200 µL TE Buffer, pH 7.5. DNA was passively eluted from the gel slurry by incubation in a 65ºC water bath for 2 hours, with periodic vortexing. Eluted DNA was separated from gel fragments by centrifugation through 0.2 µm NanoSep columns (PALL Life Sciences, ODM02C34). NanoSep columns were pre-wetted with 5 µL water before being loaded with the gel slurry, and the slurry was centrifuged at 13,200 rpm for 5 minutes. Post-centrifugation, eluate was transferred to a 1.6 mL Eppendorf for ethanol precipitation. NOTE: QIAQuick purification can also be performed in place of ethanol precipitation to reduce the time required to prepare the library. We did not observe any yield disadvantages with QIAQuick recovery relative to ethanol precipitation.

Ethanol precipitation
DNA was purified and concentrated using ethanol precipitation. To an x volume of input solution, ethanol, ammonium acetate, and glycogen were added:

- x µL DNA solution
- 2.5x µL 100% cold ethanol
- 0.1x µL ammonium acetate (7.5 mM)
- 0.01x µL glycogen

Precipitation was carried out at -80ºC overnight. Following precipitation, the solution was centrifuged at 13,200 rpm for 30 minutes. Supernatant was discarded, and the pellet washed with 750 µL cold 75% ethanol. The solution was centrifuged at 13,200 rpm for 5 minutes; the supernatant was discarded, and the 75% ethanol wash repeated. Following the second wash, the pellet was dried by Speed-Vac at 30ºC for 20 minutes. DNA was resuspended in 30 µL Buffer EB.

References

1. J. A. Reinhardt, D. A. Baltrus, M. T. Nishimura et al., *Genome Res* **19** (2), 294 (2009).