**Supplemental material for "Haplotype-phased synthetic long reads from short-read sequencing"**

**Table S1.** Synthetic long read assembly statistics.

| Sample | Trimmed, filtered 2x150 bp read pairs | Synthetic reads > 1 kb | N50 length (kb) | Short-read bases per synthetic read base[a] |
|---|---|---|---|---|
| *E. coli* MG1655 | 8,124,591 | 2,878 | 6.0 | 221.4 |
| *G. sempervirens* | 112,289,622 | 149,447[b] | 3.9 | 75.8 |
| *G. sempervirens* #2 | 10,019,885 | 28,574 | 2.8 | 43.6 |
| *G. gallus* (chicken) | 103,601,271 | 125,203 | 2.0 | 113.3 |
| *S. tuberosum* (potato)[c] | 2,789,741 | 1,528 | 3.3 | 188.9 |
| Recombinant *E. coli* from evolution experiment (sum of 24 strains) | 201,717,764 | 87,395 | 4.0 | 224.7 |
| HCT116 mRNA | 85,118,973 | 11,707 | 1.5 | N/A |
| HepG2 mRNA | 43,827,058 | 6,640 | 1.6 | N/A |
| HIV *env* mixture | 51,127,680 | 7,723[d] | 2.3 | N/A |

[a]Calculated as (total short-read nucleotides) / (total nucleotides in synthetic reads > 1 kb). Not calculated for mRNA samples, where many synthetic reads were shorter than 1 kb due to RNA degradation and the natural length distribution of mammalian mRNA, or for the HIV sample, which was deliberately over-sequenced. Low-quality or adapter-sequence nucleotides trimmed from short reads were included, i.e., the numerator was (2*150*number of read pairs).

[b]Of these,111,054 synthetic reads longer than 1.5 kb with an N50 of 4.3 kb were used to scaffold the draft genome.

[c]100 bp were trimmed from the ends of these synthetic reads prior to alignment, yielding 1,411 reads > 1 kb with an N50 length of 3.1 kb.

[d]Additional steps to remove duplicate synthetic reads reduced this number to 1,173.

**Table S2.** Accuracy of synthetic long reads aligned against the MG1655 genome.

| Reference genome | Aligned bases |
|---|---|
| Aligned bases | 10,162,249 |
| Mismatches | 3,897 |
| Mismatch rate | 0.00038 |
| Insertions | 23 |
| Insertion rate | 2.263e-6 |
| Deletions | 545 |
| Deletion rate | 5.363e-5 |
| Clipped bases | 1,205,861 |
| Percent clipped (hard + soft) | 10.61 |
| A → C | 2.3% |
| A → G | 25.1% |
| A→ T | 6.3% |
| C → A | 2.7% |
| C → G | 1.5% |
| C → T | 10.9% |
| G → A | 10.5% |
| G → C | 1.3% |
| G → T | 2.7% |
| T → A | 5.3% |
| T → C | 29.2% |
| T → G | 1.9% |

**Table S3.** *S. tuberosum* synthetic read alignment statistics.

| Mapping Quality (MapQ) cutoff | Total alignments | % Total synthetic reads |
|---|---|---|
| Unaligned | 38 | 2.7% |
| Multimapping (MapQ = 0) | 11 | 0.8% |
| MapQ >= 30 | 1,329 | 94.2% |
| MapQ >= 60 | 1,291 | 91.5% |

1,411 synthetic reads were aligned to the potato draft genome (v 4.03) with BWA-MEM.

**Table S4.** Summary of synthetic reads used in the synthetic read scaffolded *G. sempervirens* assembly.

| Feature | Metric |
|---|---|
| Number of Reads | 111,054 |
| Total Read Size | 397.8 Mb |
| Estimated Genome Coverage | 1.3x |
| Maximum Read Length | 15,260 bp |
| Minimum Read Length | 1,500 bp |
| Average Read Length | 3,581 bp |
| Median Read Length | 3,257 bp |

**Table S5.** Evaluation of the *G. sempervirens* assemblies using the CEGMA pipeline.

| | Shotgun contigs | Synthetic read scaffolds |
|---|---|---|
| Complete (no, %) | 201 (81.05%) | 203 (81.85%) |
| Partial (no, %) | 239 (96.37%) | 241 (97.18%) |

Genome assembly quality was assessed using the CEGMA pipeline pipeline (Parra et al. 2007), which identifies 248 highly conserved eukaryotic genes in an assembly. Target genes identified with > 70% coverage are indicated as complete, while those identified with less than 70% coverage are indicated as partial. The number of genes identified from each genome assembly is shown, with the percentage of the total 248 in parentheses.

**Table S6.** Alignment of RNA-seq reads to the *G. sempervirens* assemblies.

| | | Shotgun contigs | | Synthetic read scaffolds | |
|---|---|---|---|---|---|
| Tissue | Total no. of cleaned reads | No. of reads aligned in multiple mapping (%) | No. of reads aligned in single mapping (%) | No. of reads aligned in multiple mapping (%) | No. of reads aligned in single mapping (%) |
| Immature leaf | 48,752,558 | 40,298,622 (82.66) | 38,865,127 (79.72) | 40,442,197 (82.95) | 38,893,695 (79.78) |
| Stem | 77,935,967 | 54,580,534 (70.03) | 52,047,211 (66.78) | 55,007,503 (70.58) | 52,284,194 (67.09) |
| Stamens | 61,775,435 | 48,801,351 (79.00) | 46,386,521 (75.09) | 48,964,224 (79.26) | 46,376,610 (75.07) |
| Pistils | 43,707,006 | 35,017,604 (80.12) | 33,605,249 (76.89) | 35,118,309 (80.35) | 33,644,121 (76.98) |
| Petal | 57,708,901 | 44,982,737 (77.95) | 42,838,719 (74.23) | 45,133,271 (78.21) | 42,915,754 (74.37) |

Cleaned RNA-seq reads from five tissues were aligned to shotgun contigs and synthetic read scaffolds. For each assembly, the number of reads that aligned to genome in multiple and single mapping mode is shown with the percentage of reads in parentheses.

**Table S7.** Multiplexed synthetic long read assembly statistics.

| *E. coli* strain number | Trimmed, filtered 2x150 bp read pairs | Contigs >1kb | N50 length (kb) | *E. coli* genome coverage |
|---|---|---|---|---|
| Overall combined | 201,717,764 | 87,395 | 4.006 | 2.44 |
| REL11734 | 8,872,508 | 2,868 | 3.921 | 1.94 |
| REL11735 | 11,930,637 | 2,366 | 4.218 | 1.74 |
| REL11736 | 10,809,493 | 7,262 | 3.866 | 4.77 |
| REL11737 | 7,954,228 | 4,673 | 3.914 | 3.09 |
| REL11738 | 10,911,001 | 7,782 | 3.746 | 4.86 |
| REL11739 | 6,665,185 | 1,454 | 4.319 | 1.10 |
| REL11740 | 13,681,197 | 5,766 | 3.749 | 3.78 |
| REL11741 | 8,853,848 | 4,806 | 3.976 | 3.21 |
| REL11742 | 9,302,047 | 800 | 4.487 | 0.57 |
| REL11743 | 7,254,760 | 508 | 4.204 | 0.33 |
| REL11744 | 8,048,677 | 5,253 | 3.776 | 3.32 |
| REL11745 | 8,929,129 | 6,325 | 3.796 | 3.96 |
| REL11746 | 9,800,591 | 3,654 | 4.137 | 2.59 |
| REL11747 | 10,318,263 | 2,634 | 5.084 | 2.10 |
| REL11748 | 14,214,396 | 5,031 | 4.243 | 3.64 |
| REL11749 | 8,038,491 | 2,985 | 4.269 | 2.11 |
| REL11750 | 4,071,121 | 830 | 5.231 | 0.68 |
| REL11751 | 9,047,392 | 4,720 | 4.123 | 3.20 |
| REL11752 | 2,093,789 | 2,860 | 2.457 | 1.33 |
| REL11753 | 6,834,299 | 3,251 | 4.124 | 2.26 |
| REL11754 | 7,798,683 | 3,257 | 4.142 | 2.27 |
| REL11755 | 1,060,475 | 976 | 3.643 | 0.60 |
| REL11756 | 8,673,609 | 3,136 | 4.175 | 2.21 |
| REL11757 | 6,553,945 | 4,198 | 4.137 | 2.83 |

All twenty-four strains are clones isolated from the twelve recombination treatment populations in Souza et al. 1997, with two clones from each population.

**Table S8.** Comparison of synthetic long read approaches for genome assembly and phasing.

| | Metrics | | | | | |
|---|---|---|---|---|---|---|
| Method | Ref. | Demonstrated N50 (kb) | Short-read bases per synthetic read base | Format | Single-tube Multiplexing | Reagent cost per sample prep[b] |
| This work | This work | 6.0 | 43-225 | 1-2 tubes | Demonstrated | $65 |
| TruSeq Synthetic Long Reads[a] | 8, 13, 14 | 8.2 | 73-91 | 384-well plate | Incompatible | $848[c] |

[a]Alternately called Moleculo or LRseq.
[b]Labor is not included in the calculations. The per-sample cost for the method described in this work is calculated assuming multiplexing of 4 samples. Further multiplexing will further reduce the cost per sample.
[c]Cost of Illumina TruSeq kit along with necessary reagents not supplied by the kit.

**Table S9.** Human mRNA splice variant analysis.

| Cell line | Total splicing junctions | Known splicing junctions | Partial novel splicing junctions (alternative 5' or 3') | Novel splicing junctions |
|---|---|---|---|---|
| HCT116 | 12,739 | 12,357 | 192 | 190 |
| HepG2 | 9,122 | 8840 | 129 | 153 |

**Table S10.** Best-supported HCT116 mRNA synthetic reads spanning novel splice junctions.

| Chromosome | Intron start position | Intron stop position | No. of supporting synthetic long reads | Type[a] |
|---|---|---|---|---|
| 10 | 47377806 | 47387310 | 4 | Partial novel |
| 16 | 88425214 | 88425694 | 4 | Complete novel |
| 10 | 88902957 | 88903646 | 4 | Partial novel |
| 7 | 176762677 | 176764141 | 8 | Complete novel |
| 6 | 64445080 | 64447943 | 3 | Complete novel |
| 2 | 176761404 | 176762672 | 5 | Partial novel |
| 7 | 88424101 | 88425210 | 8 | Partial novel |
| 19 | 64440441 | 64444965 | 4 | Partial novel |

[a]Partial novel: alternative 5' or 3'. Complete novel: alternative 5' and 3'.

**Table S11.** Best-supported HepG2 mRNA synthetic reads spanning novel splice junctions.

| Chromosome | Intron start position | Intron stop position | No. of supporting synthetic long reads | Type[a] |
|---|---|---|---|---|
| 2 | 73583653 | 73585593 | 4 | Complete novel |
| 2 | 85839465 | 85840338 | 4 | Partial novel |
| 4 | 73581764 | 73583644 | 6 | Complete novel |
| 5 | 99015215 | 99015503 | 4 | Complete novel |
| 3 | 145244890 | 145249544 | 4 | Complete novel |
| 5 | 223489494 | 223493388 | 4 | Partial novel |
| 2 | 99015506 | 99017013 | 4 | Complete novel |
| 3 | 1105509 | 1105663 | 4 | Complete novel |

[a]Partial novel: alternative 5' or 3'. Complete novel: alternative 5' and 3'.

**Table S12.** Comparison of application categories enabled by different synthetic long read methods.

| | | Application category | | |
|---|---|---|---|---|
| Method | Ref. | Genome assembly and phasing | Full-length mRNA splice variants | Phasing similar individuals (e.g. viruses) |
| This work | This work | Shown | Shown | Shown |
| BAsE-Seq | Hong et al. 2014 | Incompatible | Incompatible | Shown |
| TruSeq Synthetic Long Reads | Voskoboynik et al. 2013 | Shown | Possible | Incompatible |

**Table S13.** Oligonucleotides used in library preparation.

| Oligo # | Function | Sequence |
|---|---|---|
| Oligo 1 | Barcode adapter | 5'-/5Phos/NNN GTTCAGAGTTCTACAGTCCGACGATC NNNNNNNNNNNNNNN CC AGGAATAGTTATGTGCATTAATGAATGG CCGC-3' or 5'-/5Phos/NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN AC AGGAATAGTTATGTGCATTAATGAATGG CCGC-3' (a mixture of this and the above were used in the *E. coli* MG1655 experiment) or 5'-/5Phos/NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN AC AATTCCTATCGTTCACGTCGTGT CGCCATTTAGTGTCCAGTCTGA-3 (used in the *env* experiment) or 5'-/5Phos/NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN CC AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| Oligo 2 | Barcode adapter, PCR primer (Rungpragayphan et al. 2002) | 5'-CCATTCAT/ideoxyU/AATGCACA/ideoxyU/ AACTATTCC/3deoxyU/G*G-3' or 5'-CCATTCAT/ideoxyU/AATGCACA/ideoxyU/ AACTATTCC/ideoxyU/G-3' or 5’-ACACGACG/ideoxyU/GAACGA /ideoxyU/AGGAAT/ideoxyU/G*T-3’ (used in *env* experiment) |
| Oligo 3 | lcPCR adapter | 5'-CCGAGAATTCCA*T-3' |
| Oligo 4 | lcPCR adapter | 5'-/5Phos/TGGAATTCTCGG GTGCCAAGG-3' |
| Oligo 5 | lcPCR primer | 5’-CAAGCAGAAGACGGCATACGAGAT (Index) GTGACTGGAGTT CCTTGGCACCCGAGAATTCCA-3' |
| Oligo 6 | lcPCR primer | 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTC CCTACACGACGCTCTTCCGATC*T-3' |
| Oligo 7 | Barcode pairing lcPCR adapter | 5’-ACACTCTTTCCCTACACGAC GCTCTTCC-3’ |
| Oligo 8 | Barcode pairing | 5’-/5Phos/A*TC GGAAGAGC ACACGTCT |

| | | |
|---|---|---|
| | lcPCR adapter | |
| Oligo 9 | Barcode pairing lcPCR primer | 5'-CAAGCAGAAGACGGCATACGAGAT (Index) GTGACTGGAGTTC AGACGTGTGCTCTTCCGATC*T-3' |
| Oligo 10 | Single-tube barcode pairing lcPCR primer | 5'-AATGATACGGCGACCACCGAGATCTACACGTTCAG AGTTCTACAGTCCGA-3' |
| Oligo 11 | Complexity quantification | 5'-CAAGCAGAAGACGGCATACGAGAT (Index) GTGACTGGAGTTC AGACGTGTGCTCTTCCGATC CCATTCATTAATGCACATAACTATTCC-3' |
| Oligo 12 | mRNA RT barcoding oligo-dT primer | 5′-CCATTCATTAATGCACATAACTATTCCT GGNNNNNNNNNNNNNNN GATCGTCGGACTGTAGAACTCTGAAC $T_{30}$VN-3′ |
| Oligo 13 | mRNA RT barcoding TSO primer | 5′- GCGGCCATTCATTAATGCACATAACTATTCCT GTNNNNNNNNNNNNNNN AGATCGGAAGAGCGTCGTGTAGG TrGrG+G-3′ |
| Probe 1 | Quenched fluorescent qPCR probe (IDT) | 5'-/56-FAM/CCT ACA CGA /ZEN/CGC TCT TCC GAT CT/3IABkFQ/-3' |

**Key:**
/5Phos/ = 5' phosphate group
/ideoxyU/ = internal deoxyuracil base
/3deoxyU/ = 3' deoxyuracil base
* = phosphorothioate linkage
rG = riboG
+G = locked nucleic acid G
N = mixture of A, T, G, and C
V = mixture of A, G, and C
$T_{30}$ = 30 consecutive Ts
lcPCR = limited-cycle PCR
Index = 6-base Illumina TruSeq Small RNA multiplexing index sequence

/56-FAM/ = probe fluorophore
/ZEN/ = probe quencher
/3IABkFQ/ = probe quencher

**Table S14.** Barcode adapter oligonucleotides (Oligo 1 in Table S13) for multiplexed library preparation.

| Name | Sequence |
|---|---|
| MULTIPLEX_RPI01 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN ATCACG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI02 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN CGATGT C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI03 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN TTAGGC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI04 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN TGACCA C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI05 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN ACAGTG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI06 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GCCAAT C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI07 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN CAGATC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI08 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN ACTTGA C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI09 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GATCAG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI10 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN TAGCTT C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3' |
| MULTIPLEX_RPI11 | 5'-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GGCTAC C AGGAATAGTTATGTGCATTAATGAATGG |

| | |
|---|---|
| | CGCC-3’ |
| MULTIPLEX_RPI12 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN CTTGTA C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI13 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN AGTCAA C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI14 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN AGTTCC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI15 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN ATGTCA C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI16 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN CCGTCC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI17 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GTAGAG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI18 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GTCCGC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI19 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GTGAAA C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI20 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GTGGCC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI21 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GTTTCG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI22 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN CGTACG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI23 | 5’-NNN CCTACACGACGCTCTTCCGATCT |

| | |
|---|---|
| | NNNNNNNNNNNNNNN GAGTGG C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |
| MULTIPLEX_RPI24 | 5’-NNN CCTACACGACGCTCTTCCGATCT NNNNNNNNNNNNNNN GGTAGC C AGGAATAGTTATGTGCATTAATGAATGG CGCC-3’ |

Multiplexing index regions are underlined.

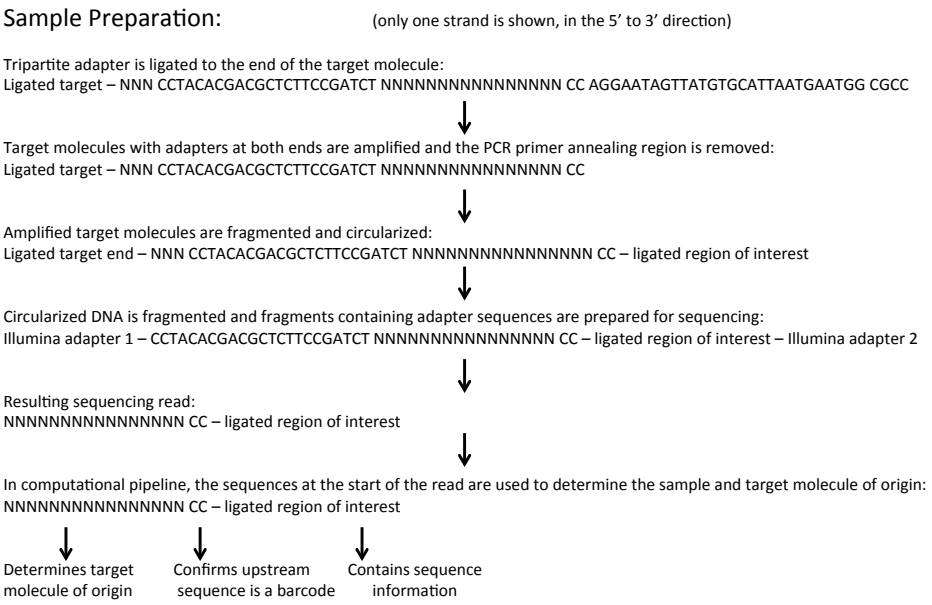



**Figure S1.** Detail showing the function of the regions of the barcode adapter during sample preparation. In the multiplexed version of the protocol, the 'CC' adjacent to the barcode is replaced by a sample-specific multiplexing index.

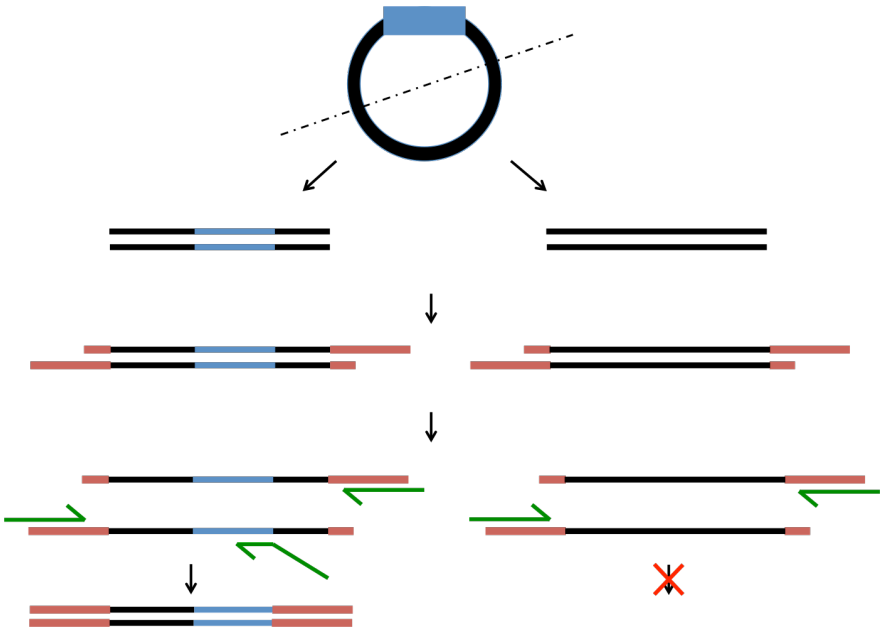


**Figure S2.** Schematic of the steps to convert sheared circular DNA into a sequencing-ready library. Circularized DNA (black) containing barcode and annealing sequences (blue) is fragmented (dotted line) into molecules about 500 bp in length. Some of the resulting molecules contain a barcode and others do not. Asymmetric adapters are ligated to each end of the molecules. Limited-cycle PCR is performed with a first primer complementary to the asymmetric adapter and a second primer complementary to the internal annealing sequence from the tripartite adapter. The primers add the full sequencing adapter sequences to the PCR product. Only molecules containing internal annealing sequences and barcodes are exponentially amplified in the PCR.

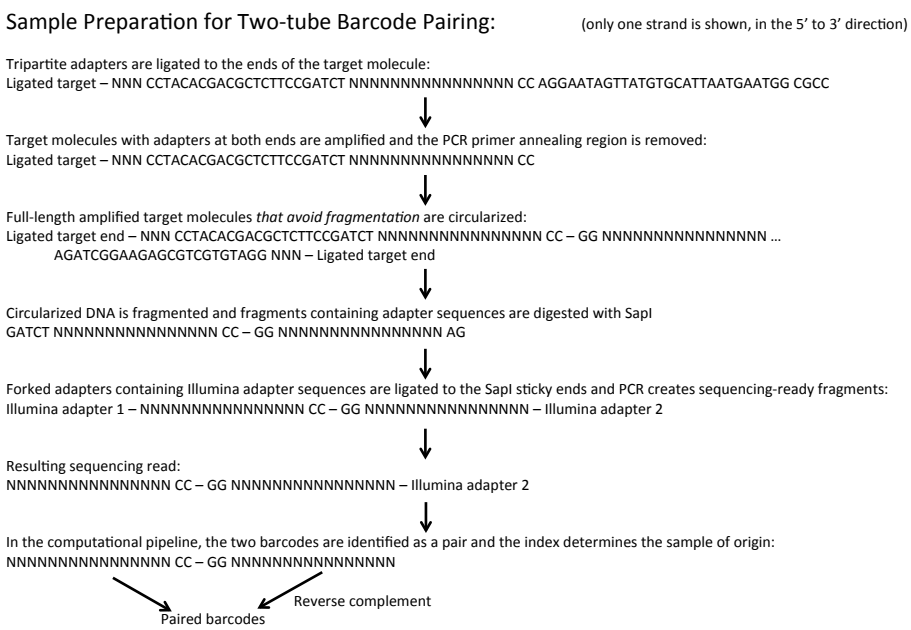



**Figure S3.** Detail showing the function of the regions of the tripartite adapter during sample preparation for two-tube barcode pairing.

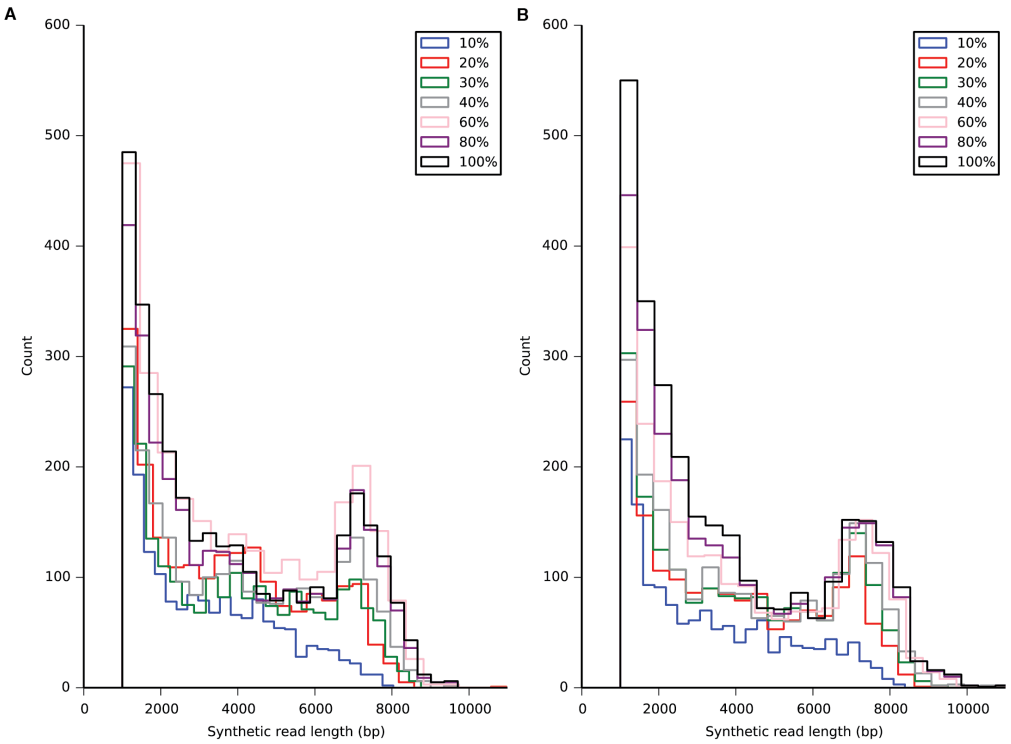



**Figure S4.** Overlaid length histograms of synthetic long reads assembled from increasing fractions of the *E. coli* MG1655 sequencing data show assembly improvement from barcode pairing. *(A)* Synthetic reads assembled without barcode pairing. *(B)* Synthetic reads assembled with barcode pairing. Barcode pairing improves assembly of long synthetic reads, particularly at low coverage (i.e., low fractions of the dataset used).

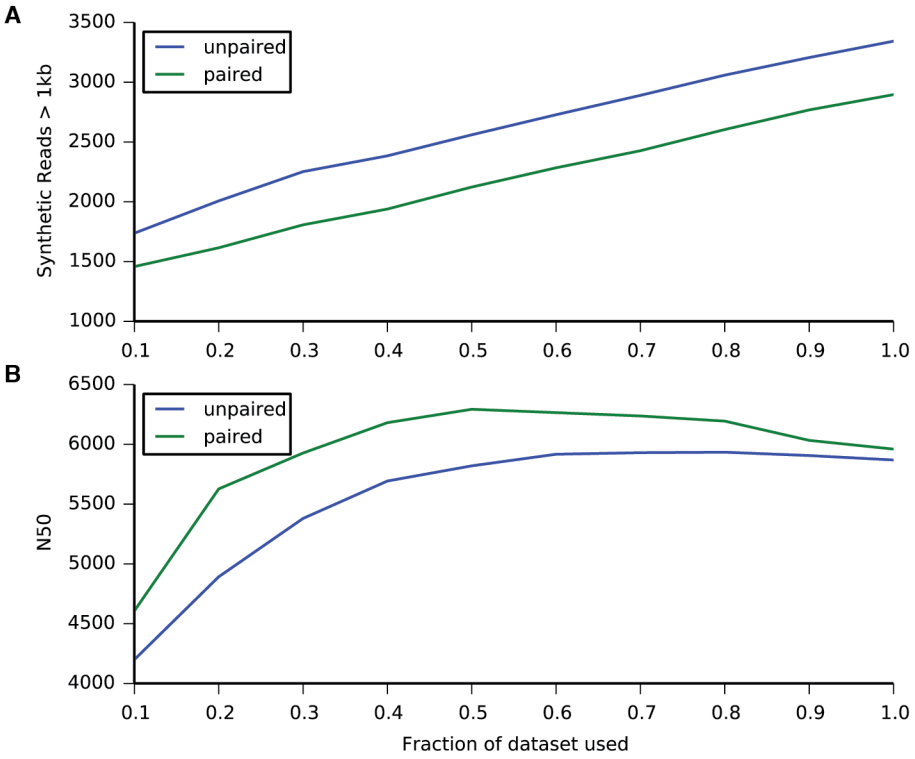



**Figure S5.** Barcode pairing improves assembly N50 length. Shown are assembly statistics of synthetic long reads assembled from increasing fractions of the *E. coli* MG1655 sequencing data. Blue = without barcode pairing, green = with barcode pairing. *(A)* The number of synthetic reads longer than 1 kb. Barcode pairing removes duplicate synthetic reads that result from two unpaired barcodes assembling the same or overlapping target fragments. *(B)* The N50 length of the assembled synthetic reads longer than 1 kb. Barcode pairing increases the N50 length of the assemblies.

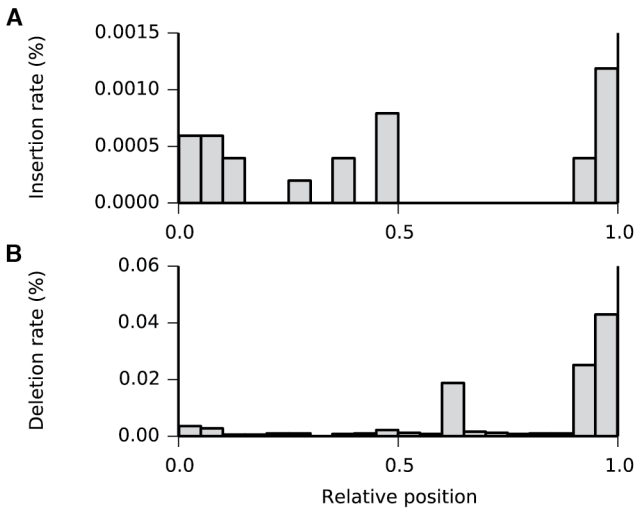



**Figure S6.** *(A)* Insertion and *(B)* deletion rates (inserted or deleted nucleotides per aligned position) of synthetic long reads from the *E. coli* MG1655 dataset, plotted as a function of relative position. Both distributions indicate indels are most likely in the low-confidence regions near the ends of the assembled synthetic long reads.

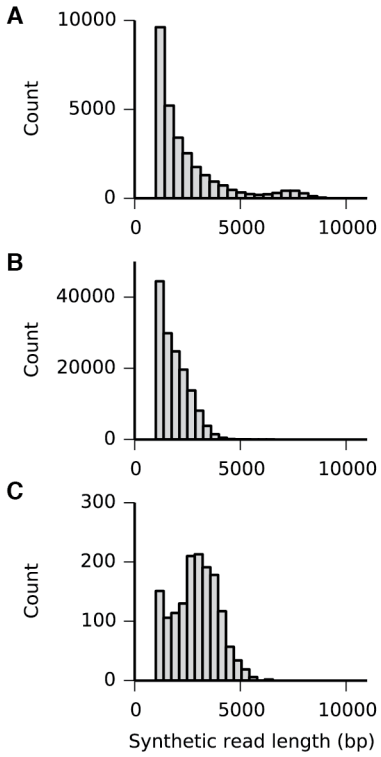



**Figure S7.** *(A)* Length histogram of synthetic long reads assembled from short reads from a second, independent sample of *G. sempervirens* genomic DNA (minimum length 1 kb). The N50 length of the assembly is 2.8 kb. *(B)* Length histogram of synthetic long reads assembled from *G. gallus* genomic reads (minimum length 1 kb). The N50 length of the assembly is 2.2 kb. *(C)* Length histogram of the synthetic long reads assembled from *S. tuberosum* genomic reads (minimum length 1 kb). The N50 length of the assembly is 3.3 kb.

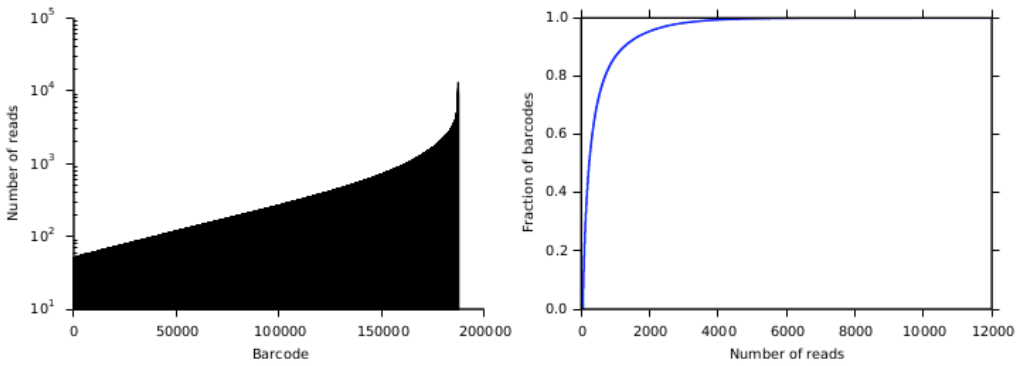
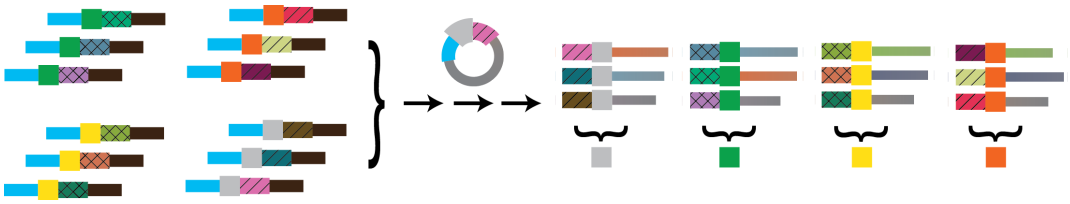



**Figure S8.** *(A)* The number of read pairs associated with each barcode in the *G. sempervirens* dataset, with a minimum of 50 read pairs. Ideally, the same number of reads would be associated with each barcode. The observed two-log range is likely due to PCR bias, and may be reduced by optimizing PCR conditions or primer sequences, or by introducing a linear amplification phase prior to exponential PCR. *(B)* Cumulative probability graph of the read distribution.



ATCCTATTTTATCATCATGTCACCATTATCCATTTTAAAGAATTACTGCATAGAACAAA

TCTCATGCCAAAAAAAGGTTCAAACTCAAATGAAATTAGTAATGCAATTGTACAATCCT

ATTATCAGTAAGAAGACGAAGACCAAG

**Key: barcode, multiplexing index**

**Figure S9.** Incorporation of a multiplexing index into the barcode-containing adapter allows independently barcoded samples to be mixed and processed in a single tube. Adapter sets containing distinct 6-bp multiplexing indexes (green, orange, yellow, and grey) are ligated to sample DNA in separate, parallel reactions and PCR amplified. The purified, quantified PCR products are mixed, and the intramolecular nature of the key circularization step enables multiplexed library preparation. After sequencing, short reads are demultiplexed according to the 6-bp index sequence that follows the barcode region. A representative forward read is shown. Because the multiplexing index is contained in the forward read, standard Illumina sample multiplexing using a 6- to 8-bp multiplexing read can additionally be used.

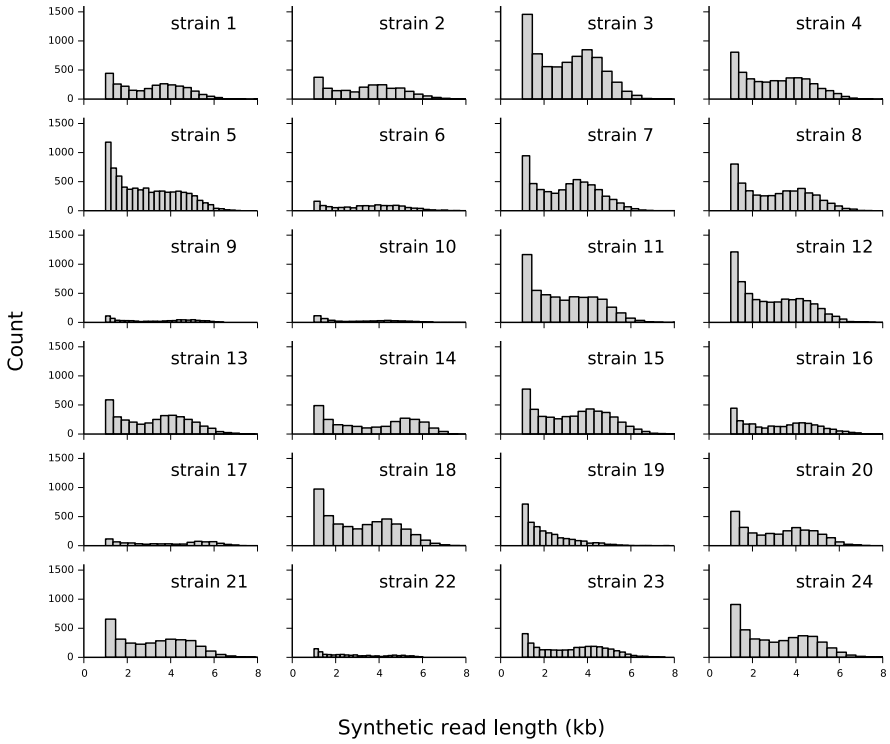



**Figure S10.** Length histograms of twenty-four independent *E. coli* genomic samples prepared for sequencing in a single tube using a multiplexed protocol.

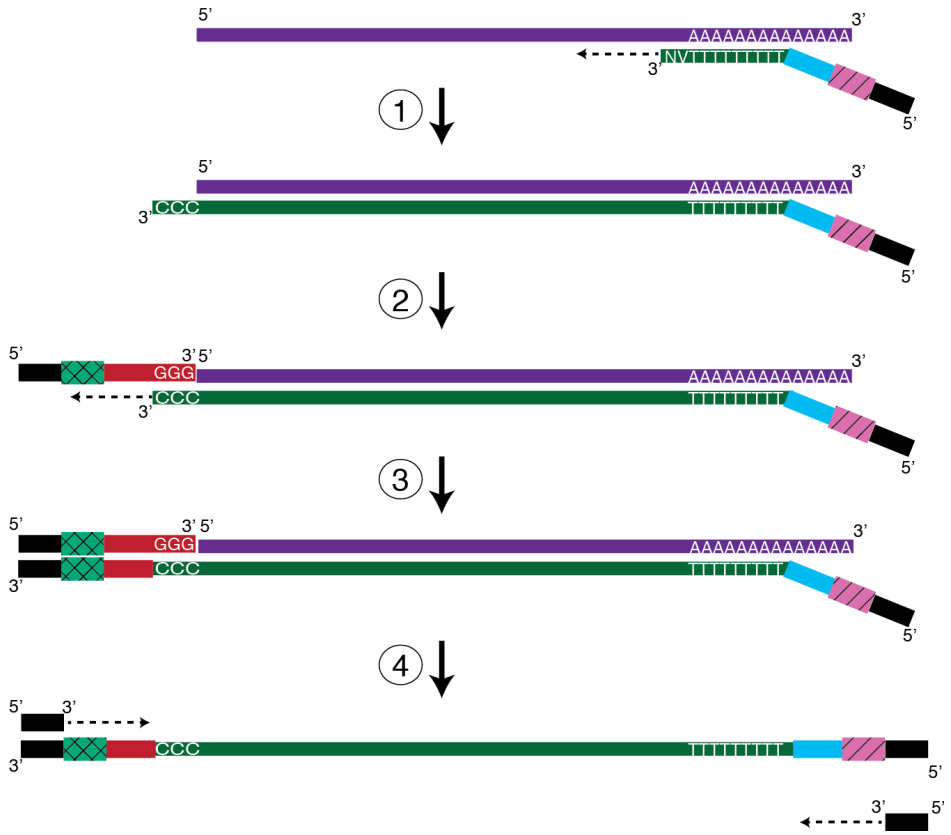



**Figure S11.** Schematic diagram of the approach for adding barcodes to full-length cDNA during the reverse-transcription step (Picelli et al. 2013). (1) RNA (purple) is reverse transcribed from a primer consisting of a poly-T annealing region (green) and an overhang containing an Illumina adapter sequence (blue), a barcode (pink stripes), and a PCR primer annealing region (black). The reverse transcriptase adds several non-templated dC bases to the 3' end of the newly synthesized strand. (2) dG bases at the 3' end of a template-switching oligonucleotide (TSO) anneal to the overhanging non-templated dC bases. The TSO consists of a PCR annealing region (black), a second barcode region (green hashed), a second Illumina adapter sequence (red), and the 3' dG bases. (3) The reverse transcriptase template-switches to copy the TSO and further extend

the 3’ end of the first DNA strand. (4) The second strand is synthesized and full-length cDNA is exponentially amplified by PCR with a single primer (black).

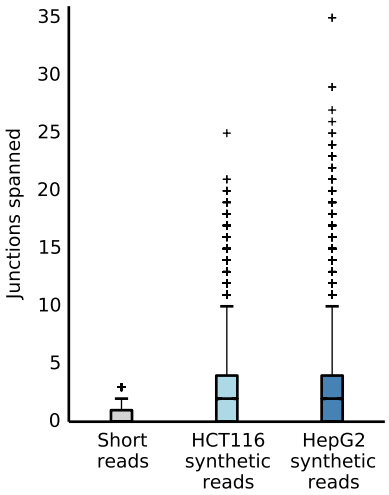



**Figure S12.** Version of Figure 2C with a standard axis. Box plots showing the number of splice junctions spanned by short reads and synthetic long reads.

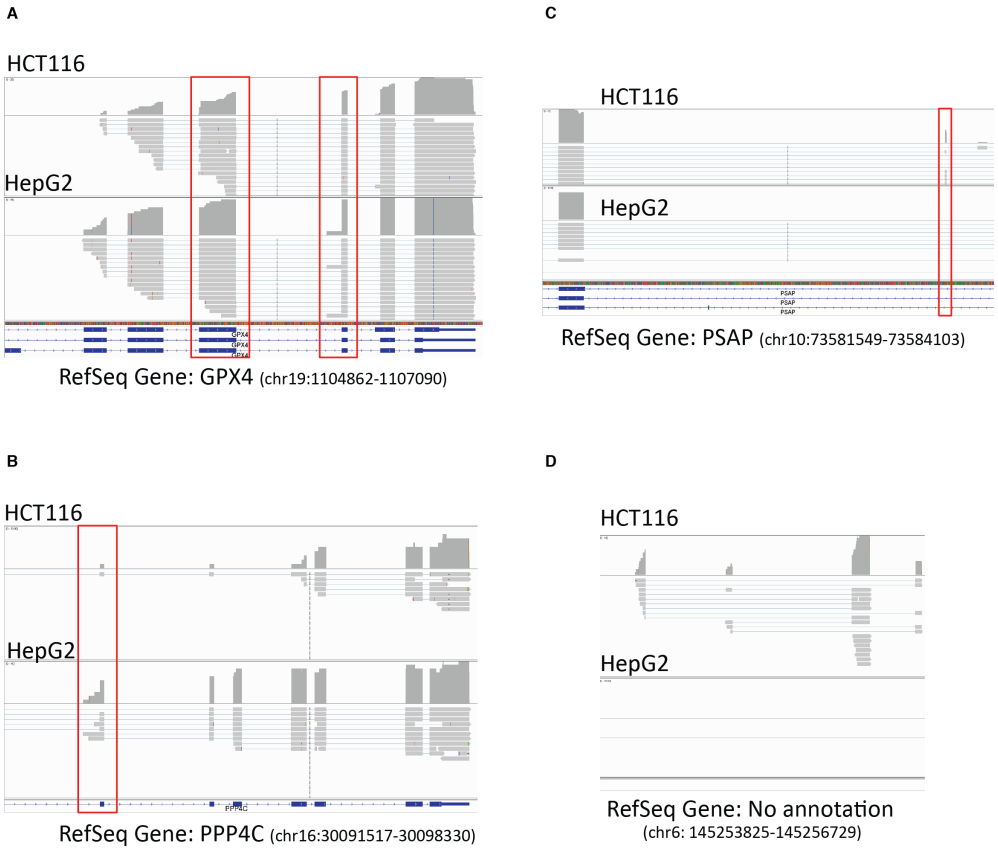



**Figure S13.** Visualization of alignments shows examples of junctions where splicing differs between HCT116 and HepG2 cell lines *(A, B, C)*, and a novel transcript in the HCT116 cell line *(D)*. The aligned long reads are shown in IGV (Integrative Genomics Viewer). *(A)* Synthetic reads indicate novel alternative 5' splice sites of two exons on gene GPX4, which are differentially spliced in the HCT116 and HepG2 cell lines. *(B)* Novel variable 5' splice sites on gene PPP4C expressed in the HepG2 cell line. *(C)* A novel exon on gene PSAP expressed in the HCT116 cell line. *(D)* Assembled long reads identify a novel transcript on chromosome 6 expressed in the HCT116 cell line.

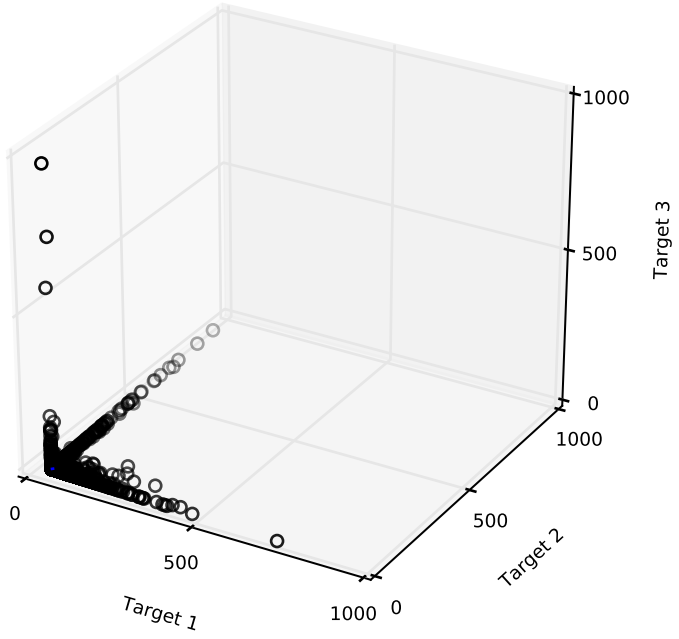



**Figure S14.** 3D scatter plot showing barcode fidelity in sequencing results from a mixture of six plasmids. The reads associated with each barcode were searched for short sequences unique to each variant. Each point represents a different barcode (8,108 total) and its position indicates the number of times sequences unique to each of three of the mixed target molecules were found within that set of barcode-grouped reads. Counting the barcodes associated with each target provides a measurement of mixture composition. Note that although Target 3 is rare in the mixture, the barcodes that tag it have as many counts as barcodes tagging more abundant targets.

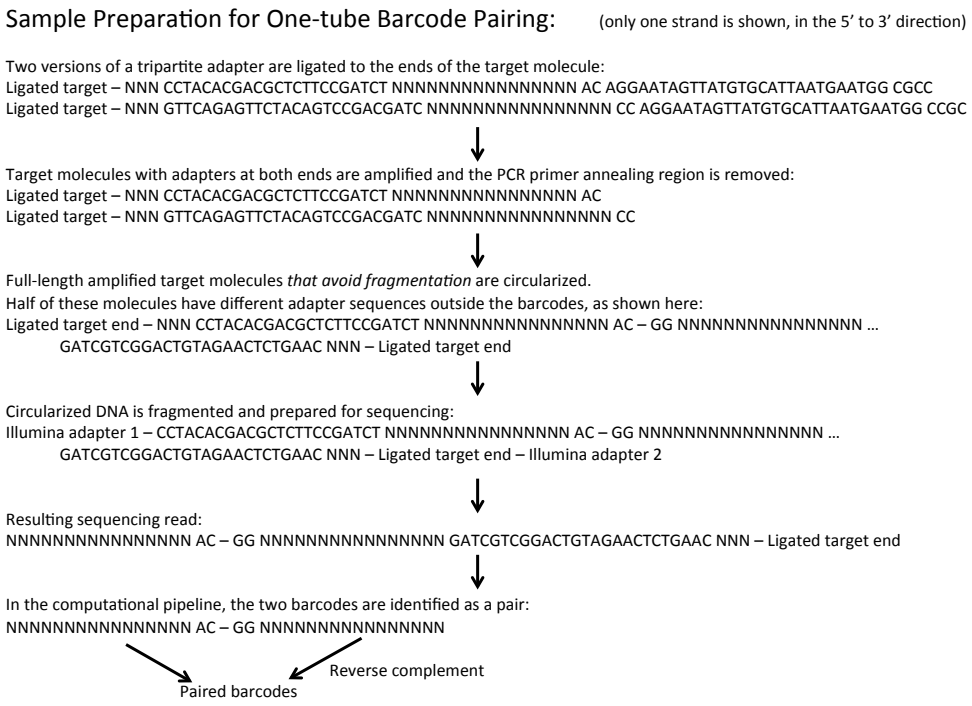



**Figure S15.** Detail showing the function of the regions of the tripartite adapter during sample preparation for one-tube barcode pairing.

**Supplementary Note.** Two separate protocols were developed for barcode pairing, the two-tube protocol and the one-tube protocol. The one-tube protocol was used for the MG1655 experiment, and the two-tube protocol was used for the multiplexed *E. coli*, mRNA, and *env* experiments. Barcode pairing was not used in the *G. gallus*, *S. tuberosum*, or *G. sempervirens* experiments.

The one-tube protocol (Supplementary Fig. 15) has the advantage of sample preparation occurring entirely in a single tube. However, only a fraction of the total barcodes can be paired (1/2 when two adapters are used). A mixture of two or more barcode-containing adapters is ligated to the dT-tailed target fragments. The adapters differ in their sequencing primer region. We used sequences derived from the Illumina Universal and Index primer sequences, respectively. As a result, approximately half of the target fragments will have different sequencing regions in the adapters that ligate to the two ends. Following PCR, some fraction of the full-length copies will avoid fragmentation, and circularization will bring the two barcodes together. Downstream limited-cycle PCR (lcPCR) will fail to amplify molecules that have the same adapter at each end because the identical sequencing regions outside the barcode regions will form a tight hairpin upon becoming single stranded. However, in molecules with different adapters at the ends, no hairpin will form, and addition of a primer complementary to the second sequencing region enables amplification of the paired barcodes. In the computational pipeline, paired-barcode reads are identified, trimmed of adapter sequences, and parsed to extract the barcode pairs.

The two-tube protocol (Supplementary Fig. 3**)** adds the complexity of splitting the library preparation into two tubes for the last third of the protocol, one tube to generate barcoded target reads and a second solely to generate paired barcode reads. The advantage is improved control of the fraction of the eventual short reads of each type. In this protocol, only one adapter sequence is used, so all target molecules ligate the same adapter at both ends. As a result, all molecules derived from circularized full-length amplicons will form a tight hairpin during lcPCR, and no paired-barcode reads will be present in the main sequencing sample. Following attachment to streptavidin-coated beads and prior to ligation of asymmetric adapters, a fraction (~15%) of the beads are moved to a second tube. SapI digestion cuts a site in the sequencing region (taken from the Illumina Multiplexing Sample Prep Oligo Only Kit), leaving sticky ends. Y-shaped adapters are ligated to the sticky ends to provide PCR annealing regions, and subsequent lcPCR adds the requisite sequencing adapter regions and a multiplexing index that allows barcode-pairing reads to be identified during analysis.