Long-adapter single-strand oligonucleotide probes for the massively multiplexed cloning of kilobase genome regions

Lorenzo Tosi†, Viswanadham Sridhara†, Yunlong Yang†, Dongli Guan†, Polina Shpilker†, Nicola Segata‡, H. Benjamin Larman* and Biju Parekkadan†,‡,§*  

As the catalogue of sequenced genomes and metagenomes continues to grow, massively parallel approaches for the comprehensive and functional analysis of gene products and regulatory elements are becoming increasingly valuable. Current strategies to synthesize or clone complex libraries of DNA sequences are limited by the length of the DNA targets, throughput and cost. Here, we show that long-adapter single-strand oligonucleotide (LASSO) probes can capture and clone thousands of kilobase DNA fragments in a single reaction. As proof of principle, we simultaneously cloned over 3,000 bacterial open reading frames (ORFs) from Escherichia coli genomic DNA (spanning 400–5,000-bp targets). Targets were enriched up to a median of around 60-fold compared with non-targeted genomic regions. At a cutoff of three times the median non-target reads per kilobase of genetic element per million reads, around 75% of the targeted ORFs were successfully captured. We also show that LASSO probes can clone human ORFs from complementary DNA, and an ORF library from a human-microbiome sample. LASSO probes could be used for the preparation of long-read sequencing libraries and for massively multiplexed cloning.

DNA sequencing has created an unprecedented wealth of biological information. With dramatic advances in parallelization, ‘reading’ DNA sequences has now become orders of magnitude more efficient and cost-effective than ‘writing’ them (for instance, synthesizing or cloning sequences of interest). Consequently, a bottleneck has formed between our knowledge of DNA sequences and our understanding of their functional significance. Massively parallel technologies enabling the synthesis and cloning of long DNA sequences are therefore required to bridge the widening gap from sequence to significance.

Highly multiplexed cloning of long target sequences has not been previously demonstrated. Traditional multiplexed PCR is generally not feasible for this purpose, due to the unpredictable interactions among large numbers of primer sets. The parallel amplification of hundreds of DNA targets (~200 base pairs (bp)) in a single reaction using short selector oligonucleotides (~70bp) that act as a template for the circularization of specific target sequences has been shown. Despite its usefulness for highly multiplex PCR-based enrichment, this technique cannot be used for the recovery of desired DNA sequences because it relies on the presence of specific restriction enzyme sites in the target sequences. Molecular inversion probes (MIPs) have proven to be a useful tool for short DNA target capture and enrichment, since they exhibit high specificity and can be massively multiplexed. MIPs are short single-stranded DNA (ssDNA) molecules (~150bp), which become circularized by gap filling after annealing to target sequences that flank a desired DNA fragment. However, traditional MIPs are inefficient at capturing larger target sequences (greater than ~200bp) due to the persistence length (‘stiffness’) of double-stranded (dsDNA) (Supplementary Fig. 1). This constraint has prevented their use for the capture of larger fragments and for the cloning of open reading frames (ORFs) that encode full-length proteins or large protein domains. In an attempt to address this target size limitation, increasing the length of the MIP linker backbone has been shown to permit the capture of larger targets (up to ~400bp). However, the method used to construct these probes requires a separate PCR reaction for each individual probe, thus limiting its scalability. Furthermore, these long MIPs were constructed as dsDNA probes, resulting in the capture of both sense and anti-sense DNA strands, which is an undesirable feature for certain applications, such as the cloning of ORFs from genomic DNA.

Here, we report the construction and use of long-adapter single-strand oligonucleotide (LASSO) probe libraries (Fig. 1), which enable the massively multiplexed capture of kilobase-sized fragments for downstream sequencing or expression. Our approach was developed to permit the assembly of LASSO probes from a complex pool of shorter, synthetic oligonucleotides, which can be readily obtained using programmable DNA microarray synthesis technology. Using LASSO probe libraries, we show the simultaneous capture of thousands of target ORFs, including those over four kilobases (kb) in length (tenfold longer than previously reported). LASSO-based ORFeome cloning captured 75% of targeted bacterial ORFs (~3,000 ORFs; over three megabases in total), enriching targeted regions over non-targeted regions by an average of ~60-fold. The functional utility of our cloned ORFs is illustrated by the LASSO-based cloning of a kanamycin gene, which conferred antibiotic resistance to host cells. LASSO cloning can also be used on human complementary DNA (cDNA) libraries. Finally, the capture and cloning of Escherichia coli ORFs using human stool-derived...
DNA as the input demonstrate the application of LASSO cloning to studies of the human microbiome at the molecular level.

**LASSO construction and single-target ORF cloning**

LASSO probe construction begins with the fusion of a precursor probe (pre-LASSO probe), designed to hybridize with sequences that flank the targeted region, with a common long-adapter sequence ([Fig. 1a,b](#)). The fusion of a long adaptor and a pre-LASSO probe ([Fig. 1c](#)) occurs with better specificity when the hybridized complex is extended before amplification and is robust to varying concentrations of adapter and pre-LASSO probe lengths (Supplementary Fig. 2a,b). The resulting pre-LASSO fusion product is then circularized ([Fig. 1d](#) and Supplementary Fig. 3) and subjected to inverse PCR, so that the LASSO annealing arms are made to flank the long-adapter sequence in the final configuration ([Fig. 1e](#) and Supplementary Fig. 4). The external primer sites are next removed and the final ssDNA LASSO probe is produced by exonuclease digestion (Methods). The final LASSO probe pool is then purified and ready to use in massively parallel target capture reactions.

The LASSO probes were first evaluated individually for their ability to clone long DNA targets. The capture reaction involves a multistep process of annealing, extension, ligation, digestion and amplification of the probe–target complex ([Fig. 2a](#)). Starting with a 100-bp target, we used single target reactions to determine the optimal conditions for gap filling and ligation (Supplementary Fig. 5). Four LASSO probes (fused with a 442-bp long adapter) were designed to capture four different target DNA sequences of approximately 0.6 kb, 1 kb, 2 kb and 4 kb in size, located within the ssDNA genome of the M13 bacteriophage. All four probes were able to capture their targets with high specificity ([Fig. 2b](#)). Notably, the LASSO probe successfully captured a 4 kb fragment, which is tenfold longer than any previously reported MIP-captured target.

---

**Figure 1** | **Synthesis of the DNA LASSO probe components.**

- **a**, Schematic of a final ssDNA LASSO probe. Two sequences complementary to regions that flank a target are linked to a universal adapter by a series of processing reactions. **b**, Schematic of the starting components for LASSO probe synthesis, consisting of a pre-LASSO probe and a long adapter. **c**, Schematic of the PCR reaction used to fuse the long adapter and the pre-LASSO probe. Gel electrophoresis results illustrate successful fusion. **d**, Schematic of the intramolecular circularization reaction of the fusion PCR product. Gel electrophoresis results illustrate successful ligation-dependent circularization. **e**, Inverted PCR is used to create linear probe precursors. Gel electrophoresis results confirm the product of inverse PCR. A 125-bp pre-LASSO probe was used with either a 242-bp long adapter or a 442-bp long adapter in the example shown.
To mimic target capture within a complex background of unrelated DNA, we assessed the influence of target DNA strandedness and background matrix complexity. The same concentration of LASSO probe was applied to M13 ssDNA, the corresponding M13 dsDNA target sequence produced by PCR, and M13 dsDNA, in the presence or absence of sheared *E. coli* whole genomic DNA. In the absence of background genomic DNA, the dsDNA target capture was less efficient compared with the ssDNA target capture. Efficiency was recovered, however, when the dsDNA template was first melted within a complex background of sheared genomic DNA (Fig. 2c). This finding is consistent with dsDNA target renaturation, which is expected to compete with LASSO probe annealing. Next, capture was performed using a dilution series of LASSO probe to test the sensitivity of the reaction, and the feasibility of performing massively multiplexed reactions that include thousands of LASSO probes (individually at low concentration) in the same reaction. A 1-kb dsDNA target sequence (500 fM) was spiked into an equimolar background of *E. coli* genomic DNA to simulate the capture of a single copy target and upstream of each ORF's start and stop codons, respectively. Using parameters defined by our optimization experiments, we developed a LASSO probe design algorithm, which we used to generate thousands of pre-LASSO probe sequences. Of the 3,999 designed ORFs from *E. coli* genomic DNA using two different adapter lengths. These results were compared with capture using traditional MIP probes (~120 bp in length). A schematic of the workflow is presented in Fig. 4a. ORFeome cloning is a particularly stringent test of multiplexed long-sequence capture, since the design of probes is highly constrained by the target sequences downstream and upstream of each ORF's start and stop codons, respectively. Using parameters defined by our optimization experiments, we developed a LASSO probe design algorithm, which we used to generate thousands of pre-LASSO probe sequences. Of the 3,999 annotated *E. coli* K12 (ATCC 27325) ORFs, the algorithm produced 3,664 pre-LASSO probe sequences that satisfied our algorithm's probe design requirements (~92% of targets). Adjusting the thresholds for target length, melting temperature or the length of the ligation/extension arms determines the number of acceptable probes. Of the 3,664 acceptable probes, we removed those corresponding to ORF targets smaller than 400 bp as a precaution to avoid potentially skewing our capture library during its subsequent PCR amplification. Approximately 20% of the *E. coli* K12 ORFeome (835 ORFs) was thus left untargeted and used as an internal, negative control for our experiments. A programmable DNA microarray was used to synthesize the pool of 3,108 × 160-bp pre-LASSO probes assembly into a pET-21(+) vector. Dual selection for ampicillin resistance (present in the pET-21(+) backbone) and kanamycin resistance demonstrated that 94% of the captured *KanR2* genes were functionally expressed (Fig. 3b,c and Supplementary Fig. 7).

**Multiplex LASSO cloning of the *E. coli ORFeome*.** We next evaluated the ability of LASSO probes to capture a library of kilobase-sized ORFs from *E. coli* genomic DNA using two different adapter lengths. These results were compared with capture using traditional MIP probes (~120 bp in length). A schematic of the workflow is presented in Fig. 4a. ORFeome cloning is a particularly stringent test of multiplexed long-sequence capture, since the design of probes is highly constrained by the target sequences downstream and upstream of each ORF's start and stop codons, respectively. Using parameters defined by our optimization experiments, we developed a LASSO probe design algorithm, which we used to generate thousands of pre-LASSO probe sequences. Of the 3,999 annotated *E. coli* K12 (ATCC 27325) ORFs, the algorithm produced 3,664 pre-LASSO probe sequences that satisfied our algorithm's probe design requirements (~92% of targets). Adjusting the thresholds for target length, melting temperature or the length of the ligation/extension arms determines the number of acceptable probes. Of the 3,664 acceptable probes, we removed those corresponding to ORF targets smaller than 400 bp as a precaution to avoid potentially skewing our capture library during its subsequent PCR amplification. Approximately 20% of the *E. coli* K12 ORFeome (835 ORFs) was thus left untargeted and used as an internal, negative control for our experiments. A programmable DNA microarray was used to synthesize the pool of 3,108 × 160-bp pre-LASSO probes assembly into a pET-21(+) vector. Dual selection for ampicillin resistance (present in the pET-21(+) backbone) and kanamycin resistance demonstrated that 94% of the captured *KanR2* genes were functionally expressed (Fig. 3b,c and Supplementary Fig. 7).
These precursor probes were then converted into mature LASSO probe libraries of two different lengths (~350 bp or ~550 bp) by fusion with two different length adapters (242 bp or 442 bp in length, respectively; sequences available in Supplementary Table 1) during probe assembly. A series of library capture optimization experiments were performed using a single adapter and a partial ORFeome LASSO probe library (Supplementary Fig. 8a–c).

A corresponding library of 3,108 pre-MIPs (~160 bp in length) was obtained from a second programmable DNA microarray and amplified with common primers. The MIPs of this library targeted the same ORFs and were prepared to be identical to the mature LASSO probe libraries described above, aside from the shorter conserved linker (sequences available in the Supplementary Dataset 'Pre-MIP Library'). The amplicon corresponded to the expected size (data not shown), and was converted to mature MIPs as described in the Methods.

The E. coli ORFeome capture was performed using the two LASSO probe sets and corresponding MIP probes. The post-capture PCR amplicons were sheared and sequenced on an Illumina NextSeq instrument to obtain 50 nucleotide single end reads. When captures were performed using LASSO probe libraries, more than half of the sequencing reads aligned with the E. coli genome (65% for LASSO-242 bp and 50% for LASSO-442 bp). In comparison, only 20% of the reads from the MIP capture libraries could

![Diagram](https://via.placeholder.com/150)

**Figure 3 | Functional assessment of a LASSO-captured ORF target.** a, PCR of circles obtained from the capture of kanamycin resistance determinant (KanR2) from total genomic DNA or plasmid DNA. The negative control for capture was total genomic DNA extracted from an E. coli clone without a vector. b, Kanamycin-resistant E. coli colonies obtained by cloning the post-capture PCR of KanR2 into a pET21 expression vector and transformation into BL21 kanamycin susceptible E. coli cells by electroporation. Scale bars, 6 mm. c, Percentage of functional KanR2 ORFs present in the kanamycin-resistant E. coli colonies.

**Figure 4 | Comparison of ORFeome capture using LASSO or MIP probe libraries.** a, Schematic of the workflow of ORFeome capture using LASSO or MIP probe libraries. A genome database was used to guide the design of the probe library, which was synthesized as pre-LASSO or pre-MIP DNA oligonucleotides on a programmable array. The pre-probe pools were converted into the mature probe pools in pooled format. The libraries of the probes were hybridized on target DNA. Closed DNA circles containing captured ORFs were selected by exonuclease digestion and then PCR amplified using universal primers. b, Median RPKM enrichment ratios of targeted ORFs versus non-targeted genetic elements for LASSO-242 bp, LASSO-442 bp and MIP captures. When comparing the enrichment ratios of LASSO probes with those of MIP probes, 100 bases on either end of the ORFs were omitted for computational purposes, as described further in the Methods. c, Quantification of unique ORFs cloned and sequenced from MIP and LASSO-242 bp capture transformations. d, Positions of captured reads mapped across the length-normalized target ORFs for LASSO-242 bp and MIP captures. All ORFs with a size greater than 1 kb were included in the graphs.
be aligned to the *E. coli* genome; most of the remaining 80% arose from self-circularization reactions (ligation of an extension arm to a ligation arm without an intervening sequence). For reads mapping to the *E. coli* genome, we calculated target enrichment factors, which we defined as the reads per kilobase of genetic element per million reads (RPKM), which were mapped to the targeted ORFs versus non-targeted ORFs or other genetic elements. Furthermore, RPKM targeted/non-targeted ratios were analysed for different length genetic elements by binning (Fig. 4b). In this experiment, LASSO targeted ORFs were enriched in all bins (up to ~60x for 1–2 Kb ORFs). Standard MIPs exhibited little to no target enrichment (<2x), regardless of ORF length.

We performed recombination-based cloning on the post-capture PCR product to move the captured *E. coli* ORFeome library into the pDONR221 entry vector. We obtained far fewer colonies when using LASSO capture material compared with MIP capture material (~20,000 versus hundreds); the same was true for complete coverage of unique target ORFs observed by sequencing (640 versus 14; Fig. 4c).

For the LASSO-242 bp and MIP capture libraries, we plotted the frequency of mapped sequence reads according to their normalized positions within the corresponding target for all ORFs greater than 1 kb in length in the LASSO-242 bp capture library (Fig. 4d). LASSO captured sequences were uniformly distributed across the full length of the target ORFs. In contrast, MIP captured sequences were strongly enriched at the end of the ORFs, suggesting that MIP capture products were largely derived from incomplete or misprimed polymerase extensions that were then able to ligate to the ligation arm and form closed circular products. Importantly, the size distribution of the LASSO capture amplicon corresponded well with that of the targeted ORFs (Supplementary Fig. 9). In summary, our data illustrate a linear regression of the logged sequencing depth as a function of target length. Sanger sequencing analysis of a randomly selected *E. coli* clone obtained from the capture library (NP_414738.1). The top inset shows the long-adapter sequence, the ligation arm of the LASSO probe and the start codon of the ORF. The bottom inset shows the end of the long-adapter sequence, the extension arm of the LASSO probe and the stop codon of the ORF.
were 465.3 and 50.3, respectively, whereas the mean and median RPKM of the non-targeted genomic regions were 2.9 and 0, respectively; fold-enrichment of targets was calculated to be between 17.5- and 160-fold (by the median or mean of the target RPKM, respectively, over the mean non-target RPKM). At a cutoff of three times the median non-target RPKM, around 75% of the targeted ORFs were successfully captured. There was an excellent positive predicted value (area under the curve: 0.959) for LASSO targets as a function of normalized read depth by receiver operating characteristic (ROC) analysis (Fig. 5c). In terms of absolute sequence coverage, the majority of the targeted ORF sequences were fully covered by the mapped reads, whereas close to none of the non-targeted sequences were fully covered (Fig. 5d). We observed a negative correlation between the normalized abundance of each target ORF and its length; ORF representation was observed to decline by 60% with each doubling of length (Fig. 5e). This bias may reflect target length-dependent capture efficiency, post-capture PCR bias or a combination of the two effects. Importantly, however, 89.4% of the cloned library was present within tenfold normalized abundance of the median, indicating a relatively uniform representation of the captured ORFs. Using a greater than 2-kb target probe subpool from the original pre-LASSO library, we observed a much more homogeneous distribution of ORF abundance, as expected (Fig. 5f). Target size-based subpooling may be a useful strategy for creating more uniform capture libraries. The integrity of several ORFs was additionally confirmed by Sanger sequencing of capture library clones. An abridged sequence of the start and stop regions of a representative cloned ORF is shown in Fig. 5g. As shown, the sequence contains the long adapter between the primer used for post-capture PCR and the ligation arm, the ATG start codon followed by the complete captured ORF, and the sequence of the long adapter between the STOP codon and the primer used for PCR. These data provide additional evidence that the cloned sequences are indeed derived from the desired LASSO capture product.

LASSO cloning of human and commensal ORFs. Finally, we evaluated the utility of LASSO cloning in the context of two important biomedical applications: (1) the cloning of a human ORFeome from cDNA and (2) the capture of bacterial ORF libraries from human gut microbiome samples. Using capture conditions optimal for massively multiplexed cloning, we evaluated the ability of LASSO probes to capture two individual full-length ORFs from a mammalian cell cDNA library. The tumour protein p53 (TP53) and ribosomal protein lateral stalk subunit P0 (RPLP0) genes were successfully captured in this manner (Fig. 6a), as confirmed by Sanger sequencing, thus demonstrating the potential utility of our method for the multiplexed construction of human protein expression libraries. We also applied the E. coli LASSO-242 bp probe library (designed using a K12 reference strain) to DNA extracted from a human stool sample. Given the extreme complexity of this DNA sample, which likely included hundreds of bacterial species and host genomic DNA, one might expect increased off-target capture. To control for this, we performed a parallel capture from the same microbiome sample, but using the MIP probe library instead. The capture amplicons obtained from the LASSO-242 bp and MIP libraries showed band patterns consistent with a successful and unsuccessful capture, respectively (Supplementary Fig. 10). The amplicons were cloned in pDONR.221 and electroporated into E. coli cells as described before. MIP-based cloning produced hundreds of colonies, while LASSO-based cloning produced several thousand (Fig. 6b). Quantification of the colonies and next-generation sequencing analysis of the pDONR plasmids from E. coli colonies recovered from agar plates revealed that 1,129 ORFs from E. coli K12 were captured with various degrees of coverage (Fig. 6c). The top 500 ORFs sequenced from pDONR plasmids showed approximately twofold enrichment (Fig. 6d; median RPKM 17.41; median RPKM non-target: 5.23), suggesting that while the LASSO cloning worked, this experiment was probably impacted by the extreme complexity of the sample. Using the MIP library in parallel, we only detected a few ORFs from E. coli, all at very low coverage, indicating that the LASSO library was effectively able to capture and clone ORFs from a microbiome sample.

Outlook
We have demonstrated that LASSO probes can be used to clone thousands of kilobase-sized DNA fragments from a prokaryotic genome in a single reaction (over three megabases in this study), and that this technique can be adapted to cDNA-derived expression libraries, as well as ORFeome cloning from microbiome samples. Conventional MIPs, by comparison, were unable to demonstrably capture ORF targets of greater than 400 nucleotides in length. LASSO cloned ORFs include their native start and stop codons, while maintaining their intended reading frames; resulting libraries can thus be expressed using standard vectors for functional biomedical screening applications. By design, new libraries of protein domains (for example, extracellular, catalytic, DNA binding, and so on) may be produced in this way. LASSO probe-based cloning may also be used to construct libraries of promoters, enhancers, long non-coding RNAs, untranslated regions of messenger RNAs, and so on, for use in high-throughput studies of gene expression12. We expect that the ability to produce inexpensive
large-frame DNA libraries will find many additional applications, including the targeted construction of long-read sequencing libraries and the assembly of chromosome-scale synthetic DNA fragments.23

Methods

Pre-LASSO probes and long adapters. Pre-LASSO probes were obtained as dsDNA oligonucleotides (about 1800 bp) or as pools of ssDNA oligonucleotides (about 1500 bp) from programmable DNA microarray (Custom Array). The pre-LASSO probes were approximately 160-bp long. For the single LASSO probes and for the 3,108 E. coli K12 ORF LASSO library subpool, the design was: 5′-GGATTATCCGGGGAGAATTC, ligation arm (variable), AAACACTTCTGCGGATGGTTCCTGGCTCTTCGATC, extension arm (variable), AGAGAAGTTCTCAGGGAGAATTC, ligation arm (variable), TGGTCTCTGCGGATGGTTCCTGGCTCTTCGATC, extension arm (variable), TGGTCTCTGCGGATGGTTCCTGGCTCTTCGATC. The ORFs of the E. coli K12 genome that were longer than 400 nucleotides were targeted with ligation and extension arms beginning at the beginning and end of the ORFs, respectively, and extended until the desired melting temperature was reached. Specifically, the algorithm first selected the ORF’s leading and trailing 32-nucleotide sequences for the two arms at melting temperatures for the ligation and extension arms of between 65°C and 85°C and 55°C and 80°C, respectively. If at least one of these conditions were not satisfied, the algorithm increased the length of the nucleotide and re-tested the condition until they were satisfied or until the end of the ORF was reached. Since an EcoRI digestion step was used to assemble the LASSO probes, the algorithm discarded the design of pre-LASSO probes where an EcoRI restriction site was present in the ligation or extension arm. A full list of the ORFs with valid ligation and extension arms and their corresponding pre-LASSO probes (subpool > 400 bp) and the pre-LASSO subpool length was found in Supplementary Dataset.

Inverted LASSO fusion PCR. The mature ssDNA form of the LASSO probes was purified using AMPure beads (1.4X, washed with 70% ethanol twice and eluted with 40 μl of nuclelease-free water. The concentration of purified inverted PCR product was measured by Nanodrop.

Production of mature LASSO probes. Approximately 1 μg of purified inverted PCR product was digested by adding 4 μl of 10X CutSmart buffer (NEB) and 1 μl of BspQI restriction enzyme (NEB). Digestion was performed at 50°C for 1 h followed by 10 min at 80°C. After digestion, 1 μl of the BspQI digested DNA and incubated for 30 min at 37°C followed by 10 min at 80°C. At this point, 2 μl (1 unit μl−1) of USER enzyme (NEB) was added in solution and incubated for 30 min at 37°C. The mature ssDNA form of the LASSO probes was purified using AMPure beads (1.4X and washed twice with 70% ethanol) and eluted in 40 μl of water. The concentration of purified mature LASSO probes was determined by Nanodrop.

MIP probe library assembly. The MIP library was designed and synthesized with flanking adapters for PCR amplification and primer removal as above. The pre-MIP probes were approximately 160-bp long and had the following design: 5′-ATGCGCGGAAAGAATTC, ligation arm (variable), TGGATTATCCGGGGAGAATTC, ligation arm (variable), AGAGAAGTTCTCATCGGGAGAATTC, ligation arm (variable), TGGTCTCTGCGGATGGTTCCTGGCTCTTCGATC. The ligation and extension arms of the MIP probes were identical to the ligation and extension arms of the LASSO probes. The sequences of the pre-MIP probes are available in Supplementary Dataset.

The pre-MIP library was PCR amplified with the TiolNew and SapINew primers. The PCR reactions were performed in 25 μl of 1× Omni Klentaq LA buffer containing 0.2 μl of Omni Klentaq LA (DNA Polymerase Bio). The PCR thermal profile was initiated for 5 min at 95°C, followed by 30 cycles of 15 s at 95°C, 20 s at 55°C and 40 s at 72°C, and then 5 min at 72°C. The PCR product was loaded in a 1% agarose gel and DNA bands corresponding to the expected size of the long adapters were cut and purified from the gel using the Wizard SV Gel and PCR Clean-Up System (Promega). The sequences of the long adapter 242 bp and long adapter 442 bp are listed in Supplementary Table 1 and all the DNA sequences of the primers are listed in Supplementary Table 2.

LASSO probe assembly. Fusion PCR. The fusion PCR reactions were performed in 25 μl reaction volumes containing 2.5 μl of 10X Klentaq Mutant Buffer, 0.6 μl of 10 mM dNTPs, 0.2 μl of Omni Klentaq LA, around 20 ng of the pre-LASSO probe (a single dsDNA pre-LASSO probe or a pool of ssDNA pre-LASSO probes), and around 20 ng of the long adapter. The solution was denatured for 4 min at 95°C and subjected to ten thermal cycles as follows: 15 s at 95°C, 20 s at 50°C and 40 s at 72°C. After the ten cycles, the PCR was stopped and 1 μl of 10X fusion forward primer BLaF and 1 μl of 10X fusion reverse primer RFPR200 EcoRI for long adapter 242 bp or RFPR400 EcoRI for long adapter 442 bp. The PCR reactions were performed in 25 μl of 1× Omni Klentaq LA buffer (NEB) containing approximately 5 μg of EcoRI digested fusion PCR product and 400 units of T4 DNA ligase; the DNA ligase was added last. Ligation was performed in a 15 μl conical tube in a cold water bath (16°C) for 16h and then concentrated to a volume of approximately 20 μl in a SpeedVac Concentrator (Thermo Scientific). The concentrated DNA was adjusted to a final volume of 100 μl by adding water, purifying with AMPure beads (1.4X and washed twice with 70% ethanol), and finally eluted into 50 μl of water. Uncircularized linear DNA was digested by adding 2 μl of solution containing 1 μl of Lambda Exonuclease (5 μl−1) and 1 μl of Exonuclease I (20 μl−1) (both from NEB) directly into the 50 μl volume containing the self-circularized DNA. Digestion was performed at 37°C for 30 min followed by 20 min at 80°C.

Inverted PCR. Inverted PCR was performed in a 25 μl total volume containing 10 μl of the circularized LASSO precursors as described above, 2.5 μl of 10X Klentaq Mutant Buffer, 0.2 μl of Omni Klentaq LA, 0.6 μl of dNTPs, and 1 μl of 0.125 μM of forward primer SapINew. BspQI and TiolNew anneal with opposite orientations in the conserved central section of the pre-LASSO probe (AACACTTCTGCGGATGGTTCCTGGCTCTTCGATC). The first three bases of TiolNew were phosphorothioate bonds to prevent digestion during subsequent Exonuclease treatment, and a 3′-terminal uracil base for subsequent primer removal using Uracil-DNA Glycosylase (USER enzyme); SapINew includes the SapI (Type II side-cleaving restriction enzyme) site for primer removal via digest with the isoschizomer BspQI. The PCR thermal profile was initiated for 4 min at 95°C, followed by 30 cycles of 10 s at 95°C, 20 s at 55°C and 40 s at 72°C, and then 4 min at 72°C.

The inverted PCR product was subsequently purified using AMPure beads (1.4X), washed with 70% ethanol twice and eluted with 40 μl of nuclelease-free water. The concentration of purified inverted PCR product was measured by Nanodrop.

DNA templates used in the capture experiments. For the LASSO probe capture optimization experiments, we used a 7,249 bp circular dsDNA isolated from the M13mp18 phage (NEB) or, alternatively, a double-stranded, covalently closed, circular form of DNA derived from bacteriophage M13 (NEB). For the capture experiments of the E. coli ORFeome by the MIP or LASSO probes, total genomic DNA from the E. coli strain K12, strain W3110 (Migula) Castellani and Chalmers (ATCC 27325) was extracted from 500 μl of LB broth (Salko Aldrich) overnight culture using a Charge Switch genomic DNA Mini Bacteria Kit (Life Technology). Shared total genomic DNA of E. coli K12 was obtained by sonicating 1 μg of total DNA in a volume of 200 μl in a 1.5 ml Eppendorf tube on ice using a Branson sonifier 450 (WWR Scientific) at an output control of 2 and a duty cycle of 50% for 40 s. For the capture of the 815 bp long kanamycin resistance gene KanR2, we used
total DNA of the E. coli clone n 29664 (Addgene), which contained the pET StrEpII TEV LIC cloning vector harbouring the Kmr2 gene.

**E. coli ORFome capture.** The LASSO libraries or the MIP library were hybridized on E. coli genomic DNA. The hybridization was performed in 15 μl of 1X AmpliTagc DNA Ligase Buffer (Epicentre) containing 250 ng of unsheared or 250 ng of sheared E. coli K12 total genomic DNA and 5 ng LASSO-242 bp or 9 ng of LASSO-442 bp or 1.7 ng of MIP library. In the hybridization volume, the concentration of Z-dNA homopolymers was approximately 10 pM. The concentration of individual LASSO probes or MIP was approximately 14 pM (44 nM for the complete LASSO or MIP libraries). The solution (15 μl) containing the MIP or LASSO probe pool and the E. coli DNA was denatured for 5 min at 95 °C in a PCR thermocycler (Eppendorf Mastercycler), then incubated at 65 °C for 60 min. After hybridization, 50 μl of the freshly prepared hybridization solution was added into the hybridization solution while maintaining the reaction at 65 °C in the thermocycler. Gap filling and ligation was performed for 30 min at 65 °C. After capture, the DNA samples were denatured for 3 min at 95 °C, and the temperature was reduced to 57 °C. Next, 4 μl linear DNA digestion solution was added immediately. Digestion was performed for 1 h at 37 °C, followed by 20 min at 80 °C. After digestion, the capture reaction was purified using AMPure beads (1.8X and washed with 70% ethanol) and eluted in 25 μl of DNase-free water. The gap filling mix was prepared fresh for each capture experiment and the composition for 50 μl of gap filling mix was as follows: 2 μl of 1 mM dNTPs, 1 μl of AmpliTagc DNA Ligase (5 U μl⁻¹), 2 μl of Omni KlenTaq LA that was previously diluted 1/10 in 1× Turbo DNA Ligase Buffer (10X AmpliTagc DNA Ligase Buffer and 40 μl of DNAase-free water. The linear DNA digestion solution (volume: 48 μl) was composed of 24 μl of nuclease-free water, 6 μl of Exonuclease I (20 units μl⁻¹), 6 μl of Exonuclease III (100 units μl⁻¹), 6 μl of Lambda Exonuclease (5 μl units⁻¹) and 6 μl of Exonuclease VII (10 μl units⁻¹) (all from NEB).

Capture of DNA targets from phage M13 using single LASSO probes. The capture of the 620-bp, 1-kb, 2-kb and 4-kb target sequence located in the DNA of the phage M13 was performed with the same gap filling mix composition and the same thermal profile for hybridization and capture used for the LASSO probe libraries, as described above. We used approximately 0.3 fmol of the single LASSO probes and 4 fmol of M13mp18 dsDNA or ssDNA. The E. coli K12 total genomic DNA background was 18 ng (in 25 μl capture volume).

For the LASSO probe sensitivity test, the E. coli K12 total genomic DNA background was around 500 ng (25 ng in a 15 µl capture volume). The concentration of M13mp18 dsDNA was around 500 ng (0.03 mg in 15 µl). The serial dilution concentrations of the LASSO 1 kb probe were zero pM, 50 pM, 5 pM and 500 fM. Capture of the Kmr2 gene was performed using 20 ng of total genomic DNA of the bacteria, which was digested with EcoRI and HindIII and ligated with the long adapter 442 bp. The solution (15 μl) was composed of 48 μl of nuclease-free water, 6 μl of Exonuclease I (20 units μl⁻¹), 6 μl of Exonuclease III (100 units μl⁻¹), 6 μl of Lambda Exonuclease (5 μl units⁻¹) and 6 μl of Exonuclease VII (10 μl units⁻¹) (all from NEB).

Expression cloning of the Kmr2 gene. Post-capture PCR amplitons were cloned by Gibson Assembly in the template pET21(+)-J (Novagen) that was previously linearized by PCR using the template primers pET21R/Gbson and pET21Gfbson. A Gibson assembly reaction was performed as described by the vendor (NEB). Transformation of the plasmid electrocompetent E. coli cells was performed using a 0.1 cm cuvette (Bio Rad) and electroporated using a Micro Pulsar (BioRad).

Next-generation sequencing of the capture libraries. Illumina library construction. Post-capture PCR products (25 μl) were purified using magnetic beads in an Agencourt AMPure XP system and eluted in 40 μl of water. The captured clones into the pDONR211 vector were minipreped and eluted into 50 μl water. The DNA concentration was measured by Nanodrop. Purified post-capture PCR (50 ng) of the capture libraries were visualized by agarose gel electrophoresis using 3% agarose gel in 1X TBE buffer (100 mM Tris-HCl, 10 mM Sodium Bicarbonate and 1 mM EDTA). The presence of DNA inserts was determined using the colony as a DNA template for PCR with the primers provided with the kit. The PCR products (5 μl) were visualized by agarose gel electrophoresis and purified using AMPure beads. Sanger sequencing of colony PCR amplitons was performed by capillary electrophoresis on the 96-well capillary matrix of an ABI3730XL DNA analyser.
version 1.3.1 (ref. 12), we filtered the reads to include only those satisfying MAPQ scores of at least 30 and then sorted the resulting bam file. Since the probes were made for 3,108 genes that satisfy the requirements of the current protocol, we considered these genes as targets. The rest of the genes, along with the intergenic regions, were considered as non-targets.

When comparing the enrichment ratios of the LASSO probes with those of the MIP probes (Fig. 4b), we trimmed 100 bases on either end of the ORFs during data preprocessing, as the MIP capture products were largely derived from incomplete or misprimed polymerase extensions that were then able to ligate to the ligation arm and form closed circular products. For ROC analysis, we considered targets as true positives and non-targets as false positives. For each of these regions; for instance, targets and non-targets, we created two bed files separately using the custom scripts. We then used bedtools version 2.19.0 (ref. 12) to estimate the depth of the regions. We then estimated the RPMK for each of these regions. Using RPMK allows us to easily compare the metrics across different samples that might have genes sequenced at different depths and a different number of total reads. We also calculated the fraction coverage; that is, the fraction of the bases covered in the regions of interest. We repeated all the above steps for the LASSO and MIP analyses of E. coli, pDONR and the human stool sample. We then looked at various analyses; for instance, ROC analysis, the plot fraction coverage for targets and non-targets, scatter plots of depth and RPKM versus length. We have prepared an R script for all the above analyses.

Statistical analysis. All the data are presented as mean or median ± s.e.m., as stated in the figure legends. Statistical significance was assessed using a Student’s t-test for pair-wise comparison, and one-way ANOVA for comparisons between multiple (≥3) conditions. P < 0.05 was considered significant.11,17.

Data availability. The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information. Next-generation sequencing data of the captures performed with the LASSO and MIP libraries are available at https://www.ncbi.nlm.nih.gov/sra/?term=SRP079889.

Code availability. All the parameters used in Trimmomatic, Bowtie 2 and SAMtools, along with custom-based scripts used with bedtools and the R script, are provided in github: https://github.com/viswam78/LASSOprobes.

Received 18 July 2016; accepted 10 May 2017; published 26 June 2017

References

1. Van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. Trends Genet. 30, 418–426 (2014).
2. Nilsson, M., Dahl, F., Larsson, C., Gullberg, M. & Stenberg, J. Analyzing genes using closing and replicating circles. Trends Biotechnol. 24, 83–88 (2006).
3. Dahl, F. et al. Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc. Natl Acad. Sci. USA 104, 9387–9392 (2007).
4. Turner, E. H., Lee, C., Ng, S. B., Nickerson, D. A. & Shendure, J. Massively parallel exon capture and library-free resequencing across 16 genomes. Nat. Methods 6, 315–316 (2009).
5. Nilsson, M. et al. Padlock probes: circularizing oligonucleotides for localized DNA detection. Science 265, 2085–2088 (1994).
6. Landegren, U. et al. Molecular tools for a molecular medicine: analyzing genes, transcripts and proteins using padlock and proximity probes. J. Mol. Recognit. 17, 194–197 (2004).
7. Hagerman, P. J. Flexibility of DNA. Annu. Rev. Biophys. Biophys. Chem. 17, 265–286 (1988).
8. Krishnakumar, S. et al. A comprehensive assay for targeted multiplex amplification of human DNA sequences. Proc. Natl Acad. Sci. USA 105, 9296–9301 (2008).
9. Shen, P. et al. Multiplex target capture with double-stranded DNA probes. Genome Med. 5, 50 (2013).
10. Shen, P. et al. High-quality DNA sequence capture of 524 disease candidate genes. Proc. Natl Acad. Sci. USA 108, 6549–6554 (2011).
11. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. Nat. Methods 11, 499–507 (2014).
12. Tewhey, R. et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell 165, 1519–1529 (2016).
13. Boeke, J. D. et al. The genome project—write. Science 353, 126–127 (2016).
14. Bolger, A. M. et al. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
15. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
16. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).
17. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

Acknowledgements

This work was supported in part by the Shriners Hospitals for Children (B.P. and L.T.), a Prostate Cancer Foundation Young Investigator award (H.B.L.), and National Institutes of Health Grants R01EB012521 (B.P.), K01DK087770 (B.P.) and 1U24AI118633 (H.B.L.).

Author contributions

L.T., H.B.L. and B.P. conceived and designed the study. L.T., V.S., Y.Y., D.G., P.S. and N.S. performed the experiments, and analysed and interpreted the data. L.T., H.B.L. and B.P. wrote the manuscript.

Additional information

Supplementary information is available for this paper.

Correspondence and requests for materials should be addressed to H.B.L. and B.P.

How to cite this article: Tosi, L. et al. Long-adapter single-strand oligonucleotide probes for the massively multiplexed cloning of kilobase genome regions. Nat. Biomed. Eng. 1, 0092 (2017).

Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Competing interests

A patent application on the technology has been filed (PCT/US2016/035919). The authors declare no other competing financial interests.