Targeted sequencing for gene discovery and quantification using RNA CaptureSeq

Tim R Mercer[1,5], Michael B Clark[1,2,5], Joanna Crawford[2,5], Marion E Brunck[3], Daniel J Gerhardt[4], Ryan J Taft[2], Lars K Nielsen[3], Marcel E Dinger[1] & John S Mattick[1]

[1]Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [2]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. [3]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Queensland, Australia. [4]Mater Research Institute—The University of Queensland, Translational Research Institute, Woolloongabba, Queensland, Australia. [5]These authors contributed equally to this work. Correspondence should be addressed to J.S.M. (j.mattick@garvan.org.au) or M.E.D. (m.dinger@garvan.org.au).

Published online 4 April 2014; doi:10.1038/nprot.2014.058

RNA sequencing (RNAseq) samples the majority of expressed genes infrequently, owing to the large size, complex splicing and wide dynamic range of eukaryotic transcriptomes. This results in sparse sequencing coverage that can hinder robust isoform assembly and quantification. RNA capture sequencing (CaptureSeq) addresses this challenge by using oligonucleotide probes to capture selected genes or regions of interest for targeted sequencing. Targeted RNAseq provides enhanced coverage for sensitive gene discovery, robust transcript assembly and accurate gene quantification. Here we describe a detailed protocol for all stages of RNA CaptureSeq, from initial probe design considerations and capture of targeted genes to final assembly and quantification of captured transcripts. Initial probe design and final analysis can take less than 1 d, whereas the central experimental capture stage requires ~7 d.

INTRODUCTION

RNAseq is a global technique for simultaneously measuring gene abundance, detecting unannotated genes and reconstructing the complex gene isoforms (both known and novel) that result from splicing[1–5]. However, owing to the wide dynamic range of the transcriptome, in which a minority of highly expressed genes constitute the majority of RNA molecules within a cell[6], RNAseq achieves only sparse coverage of weakly expressed transcripts, impairing accurate transcript assembly and quantification.

Targeted RNAseq can overcome the challenge posed by the wide dynamic range of the cellular RNA population by focusing sequencing on targeted genes of interest, thereby providing a huge enrichment of sequencing read coverage[7,8] (**Fig. 1**). This enables more sensitive gene discovery, quantification and assembly of even weakly expressed transcripts. Furthermore, in combination with multiplex library preparation, the increased efficiency of targeted RNAseq can also reduce reagent costs.

Here we describe a protocol for RNA CaptureSeq that facilitates the targeted and focused sequencing of RNAs of interest[7]. This protocol uses labeled in-solution DNA oligonucleotides to capture RNA targets of interest that are then purified and subjected to sequencing.

Applying CaptureSeq to investigations of RNA biology

The increased sequencing coverage afforded by CaptureSeq can be applied to a wide range of transcriptional analyses. CaptureSeq is a highly sensitive tool for novel gene discovery, with the contiguous probing of genomic regions often revealing the existence of novel and weakly expressed transcripts[7]. CaptureSeq can also be used to selectively target exons to achieve improved coverage of known genes, identify novel splicing events and exons[7] and even distinguish allele-specific gene expression[9].

We recently demonstrated the use of CaptureSeq for quantitative gene profiling (M.B.C. *et al.*, unpublished data), finding that CaptureSeq accurately retained the quantitative abundance of the original RNA samples for all but the most abundant genes. Each gene of interest is targeted by a diversity of probes that are present with excessive abundance relative to the targeted gene; hence, probe availability does not impose a limit on gene sampling (with the potential exception of the most abundant genes). Therefore, although nontargeted genes are omitted, those genes that are targeted are sampled representatively. Indeed, for weakly expressed genes, the sequence coverage and high sampling rate provided a more accurate measurement of abundance than matched RNAseq, which achieved only a low coverage and variable sampling rate. This quantitative capacity permits the use of CaptureSeq to profile specific gene pathways, marker genes, classes of genes (such as long noncoding RNAs) and genes associated with disease[8].

CaptureSeq can also be applied as an analytical tool to investigate aspects of RNA biology, such as dissecting the pathways involved in RNA synthesis, processing and degradation. CaptureSeq can sufficiently enrich transient or intermediate RNA species for analysis, or it can resolve specific transcriptional features with high coverage and resolution. For example, padlock capture has been used to achieve sufficient coverage for the high-confidence detection and analysis of RNA-editing events[10].

Alternative platforms for CaptureSeq

This protocol has been developed by modifying and combining RNAseq platforms with the target enrichment platforms that are commonly used for targeted DNA sequencing (often referred to as exome sequencing)[11]. These platforms have been well developed for the analysis of genetic variation, and they involve a range of alternative enrichment strategies. Direct comparisons between the technical features of these alternative strategies (including coverage, enrichment, uniformity, specificity and so on) have been previously reported[11,12], and here we discuss the application of the three major strategies for targeting RNA transcripts[13,14].

Mutliplex PCR amplification uses a pool of primer pairs to simultaneously amplify regions of interest for sequencing[15], and it has been successfully applied for the large-scale validation of splice isoforms[16]. Despite relying on previous gene annotations, this approach can also identify novel exons that are amplified

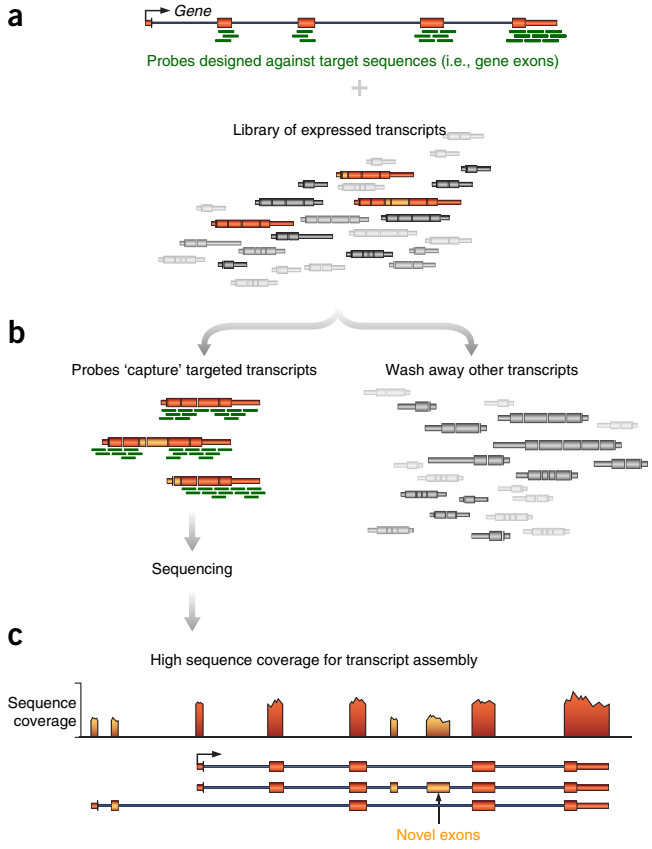

Figure 1 | Schematic overview of targeted RNAseq. (**a**) Oligonucleotide probes that are designed to target the exons of a gene are added to a library of expressed transcripts. (**b**) Probes hybridize and capture targeted RNAs of interest, whereas other nontargeted RNAs are washed away. The purified RNA of interest is then subjected to deep sequencing. (**c**) The resulting highly enriched sequencing coverage enables robust transcript assembly, abundance quantification and sensitive detection of novel exons and isoforms.

between primer pairs. Although primer design must be performed carefully, as differences between individual primers may result in heterogeneous coverage, this is a relatively fast protocol, and it can be potentially used for quantitative gene profiling. Commercially available kits, such as the Life Technologies Ion Ampliseq gene panels, have recently become available for multiplex PCR-targeted RNAseq.

Capture by circularization targets two paired regions with a chimeric probe that forms a circle of the targeted region for amplification[17,18]. This strategy enjoys high specificity, owing to the requirement for two complementary sequences in the correct orientation, but it relies on previous gene annotations, and it has similar disadvantages to multiplex PCR.

In-solution capture uses labeled RNA or DNA oligonucleotides that can be hybridized to cRNA or cDNA of interest. Only part of a gene needs to be targeted for successful capture, and this approach can resolve alternative splicing, transcription initiation and termination events. Multiple overlapping probes that target a single sequence help average the differential performance of individual probes, thereby resulting in a relatively uniform coverage and accurate measurement of quantitative abundance. However, because of hybridization and washing steps, the protocol is longer and more involved than the other approaches.

The protocol described here is based on the use of TruSeq stranded mRNA sample preparation (Illumina) in combination with SeqCap EZ library (Roche/NimbleGen). This platform uses in-solution DNA oligonucleotides for targeted enrichment, is fully customizable and can target a large area with high coverage (2.1 million in-solution oligonucleotide probes that cover up to 200 Mb of sequence). Alternative targeted in-solution oligonucleotide capture platforms are available, and they include Agilent SureSelect or Illumina TruSeq custom/exome enrichment, both of which have recently released commercially available dedicated kits for targeted RNAseq.

Protocol overview

This protocol provides detail on the application of CaptureSeq for gene discovery and quantification, as previously described[7]. CaptureSeq consists of three major phases (**Fig. 2**). The first phase involves the design of DNA oligonucleotide probes (**Box 1**). This design phase determines which portion of the transcriptome will be enriched, and careful design and forethought at this phase are critical for the subsequent success of the CaptureSeq method. The second phase constitutes the preparation of cDNA libraries, hybridization of libraries to probes, washing and removal of nontargeted cDNA and elution of targeted cDNA for sequencing. This phase encompasses the majority of experimental work

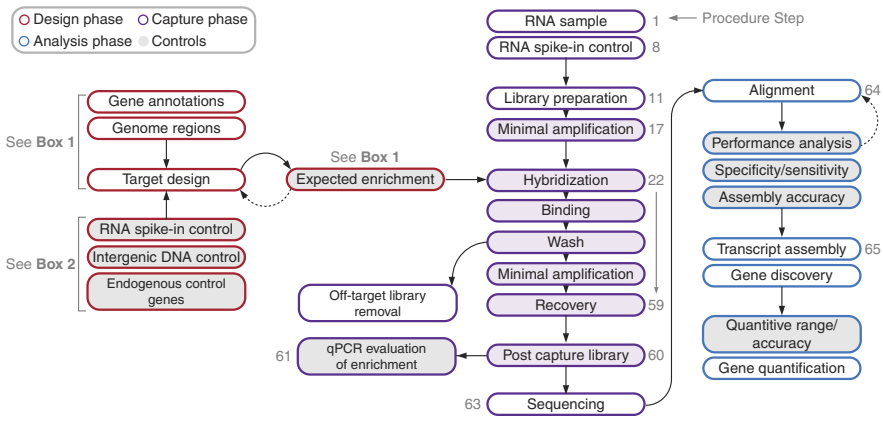

Figure 2 | Schematic overview of targeted RNAseq in three stages: design (red), capture (purple) and analysis (blue). Controls to assess and inform experimental performance are indicated (shaded gray). During the initial design phase (red), complementary oligonucleotide probes are targeted to RNA transcripts or genomic regions of interest. **Box 1** provides additional detail for probe design. A suite of control probes are also included within the design and, before manufacture, we recommend iterative design validation and estimation of expected fold enrichment by comparison with RNAseq libraries. The capture phase (purple) encompasses the experimental steps of the protocol, including sample preparation and central capture steps (shaded purple). PCR amplification cycles are minimized during capture, and qPCR is used to evaluate enrichment before sequencing. The final analysis phase (blue) involves the alignment, transcript assembly and quantification of sequenced reads. The parallel evaluation of control reads (**Box 2** provides further detail on how assessment is performed) permits an assessment of targeted RNAseq performance and informs the computational parameters used for experimental sequence read analysis.

Box 1 | Probe design

This initial design phase is only necessary for users requiring custom probe manufacture; users using fixed designs can proceed directly from Step 1 of the PROCEDURE. Design consideration may vary among platforms, and this protocol is designed for use with SeqCap EZ probe library (Roche/NimbleGen).

Retrieving a custom gene list

Users designing custom probes are required to generate a tab-delimited file (0-based) containing chromosome start and stop coordinates of regions to be targeted by probes. As an example of how to generate a custom target gene list, we describe the retrieval of exon coordinates for GENCODE long intergenic noncoding RNAs (lincRNAs) by using the UCSC Genome Browser (http://genome.ucsc.edu).

1. Go to the 'Table Browser' tool (click 'Tables' in top blue header menu) and select the latest GENCODE annotation (select the following from drop-down fields; clade: mammal / genome: human / assembly Feb 2009 (GRCh37/hg19)/group: Genes and Gene Prediction Tracks / track: GENCODE Genes V17/table: Basic (wgEncodeGencodeBasicV17). Select 'create' in the filter field, scroll down and select to activate the 'Linked Tables' (hg19 wgEncodeGencodeAttrsV17 Basic set of attributes associated with all GENCODE transcripts) and select 'allow filtering using fields in checked tables' and type into the field 'geneType does match: lincRNA' and select 'submit'.
(Optional) Gene identifiers (including names or accession IDs) can be pasted in the identifier (names/accession) field to return a selected gene list.
2. Select 'output format: BED' and select 'get output'. On the following screen, 'Output wgEncodeGencodeBasicV17 as BED', select 'create one BED record per: Exons' and select 'get Bed'. The first three columns of the retrieved tab-delimited file provide the chromosome, start and stop coordinates of exons required for probe design. The GENCODE (v17) lincRNAs output file should comprise 16,525 exons (that represent 6,020 full-length transcripts).

Visualizing targeted sequences

3. Upload coordinate files for review and visualization in the UCSC Genome Browser by using the 'Add custom Tracks' tool.
We recommend thorough review of coordinates to confirm that they correspond to user requirements. Although we do not screen for repetitive regions (screening of nonunique sequences is performed during Roche/NimbleGen probe sequence manufacture), we recommend omitting any regions that overlap structural ncRNA genes (such as tRNAs, snRNAs and rRNAs) that can be visualized with the RepeatMasker track in the UCSC Genome Browser.

Estimating expected fold enrichment

4. First, build an index from the probed region sequences. The sequences of GENCODE (v17) lincRNAs can be retrieved as a .fasta file by using the Table Browser settings listed for 'Retrieving a custom gene list' and selecting output format: 'sequence'.
(Optional) The fasta sequence file for an uploaded custom track can be selected from the table browser by selecting group: 'Custom Tracks' and selecting output format: 'sequence'.

```
$ bowtie2-build probe_sequence.fa probe_index
```

and then align reads from a matched RNAseq library to this probe sequence index:

```
$ bowtie2 --no-head -x probe_index -U RNAseq_library.fastq \
  -S alignments.sam
```

The output file ('alignments.sam') contains the alignment of reads to the probe sequence. For further details on the SAM file format, see Li *et al.*[40].

5. To count all alignments to probe sequences, type the following command:

```
$ awk '$2 != 4' alignments.sam | wc -l
```

Dividing the number of reads that align to the genome divided by the number of reads that align to probe sequences provides an estimate of fold enrichment.

6. Provide the tab-delimited coordinate file of probe regions to a commercial vendor to manufacture the probe sequences. Supplied probe libraries should be prepared as described in Reagent Setup.

that is conducted within the laboratory, and it is described in detail within this protocol. The final stage is the computational analysis of the resultant sequenced libraries for which there are a range of RNAseq analysis packages available according to the user's requirements[3,19]. Here we have focused on data analysis features specific to CaptureSeq.

Experimental design

Probe design. Users may use fixed probe designs or may wish to customize probe design according to their specific requirements. Fixed designs target commonly used content, do not require the user to undertake probe design and often have the advantages of previous validation and optimization. However, if fixed designs do not overlap the genes of interest, users may be required to customize designs.

Probes can be designed to tile contiguous genomic regions, with no reference to prior gene annotations, permitting the discovery of entirely novel genes transcribed from these regions. Current oligonucleotide probe platforms can target up to ~6% of the human genome when contiguously tiled. Alternatively, exons

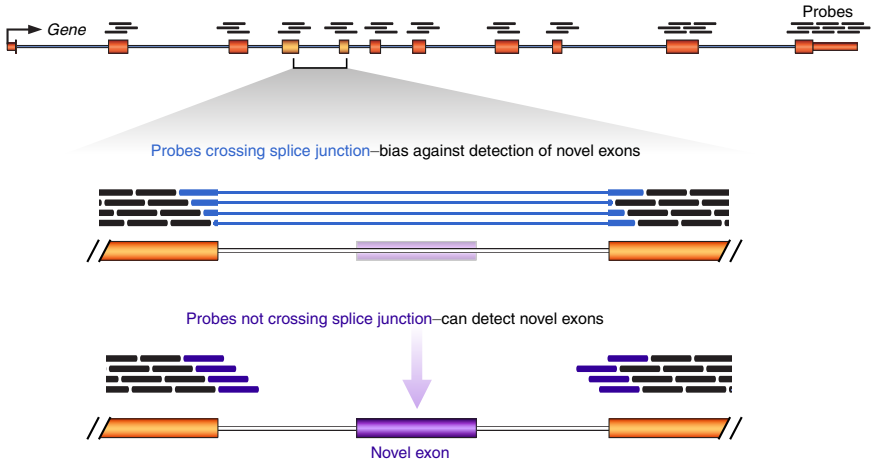

Figure 3 | Designing probes at splice junctions. Designing probes that traverse exon-exon junctions (blue) excludes novel cassette exons. Designing probes that target exons but do not traverse exon-exon junctions permits the identification of novel exons (purple).

can be selectively targeted, with intervening introns ignored. Targeting only part of a transcript is often sufficient to capture the full-length transcript, including additional spliced exons within the same transcript. Indeed, this is a major advantage of CaptureSeq that allows the identification of novel exons and splice isoforms for the expansion of known annotations. Specific isoforms can be also targeted by spanning probes across an exon-exon junction (**Fig. 3**); however, this biases against the identification of intervening novel exons.

Strategies for optimizing probe sequences that were originally developed for microarray technologies can be similarly applied to CaptureSeq[20]. Efficient probe design aims to minimize differences between individual probe melting temperature and nucleotide content and to avoid stable secondary structures. High probe coverage, with multiple probes overlapping a single region, also provides redundancy, averaging the differential performance between probes. Probe features vary substantially between different enrichment platforms, but they are typically longer than 100 nt, which affords high specificity[11].

Estimating fold enrichment. The scale of enrichment and sequencing coverage realized by CaptureSeq is inversely proportional to the collective expression of targeted transcripts. Therefore, to maintain a high enrichment, a large number of weakly expressed genes or a smaller number of moderately expressed genes can be targeted. Targeting all known exons to 1,000 randomly selected 'average' human genes corresponds to an expected ~55-fold target enrichment (**Fig. 4**), and we recommend aiming to achieve at least ~20-fold enrichment, which corresponds to targeting ~2,700 genes[21].

Before manufacture, we recommend estimating the expected maximal fold enrichment of probe designs by using RNAseq libraries from matched or closely similar tissue sources (**Box 1**). These can often be retrieved from a public sequence data archive (such as Gene Expression Omnibus (GEO)) or consortium (such as Encyclopedia of DNA Elements (ENCODE))[22]. The fractional overlap of read alignments to targeted regions (including control probes) allows an estimate of fold enrichment, and fractional overlap with individual targeted genes can inform iterative design amendments that omit highly expressed genes or identify off-target hybridization from homologous transcribed sequences. In practice, we generally return a lower enrichment than the estimated fold enrichment because of the additional capture of novel exons and isoforms and contamination by off-target transcripts.

The estimated fold enrichment can also inform the sequencing depth required to achieve the user's aim. Gene discovery applications generally require high sequence coverage for robust *de novo* assembly of spliced transcripts, with an estimated minimum eightfold coverage required for the assembly of spliced genes[6]. Sequence coverage is also a function of read length, and sequencing with paired-end reads of maximum length is highly advantageous for efficient transcript assembly[3]. For gene profiling applications, the sequencing depth may be lower, as coverage only needs to be sufficient for robust abundance measurements. Accordingly, users may prefer to profile larger numbers of multiplexed samples, with a commensurately lower coverage.

Targeting repeats. Repetitive sequences require special consideration during probe design. Repetitive sequences span a range of uniqueness, from repeats with a unique sequence (albeit highly similar to others) to identical sequences found in high numbers throughout the genome. Targeting a transcript that harbors a repetitive or similar sequence can result in off-target hybridization and, if probes become saturated, incomplete coverage of the original targeted RNA. Although some designs may specifically target and exploit a repeat element to target a class of RNA[23], we generally recommend masking out highly transcribed repetitive genes (such as rRNAs, tRNAs and small nuclear RNAs (snRNAs)) whose off-target capture will substantially reduce enrichment.

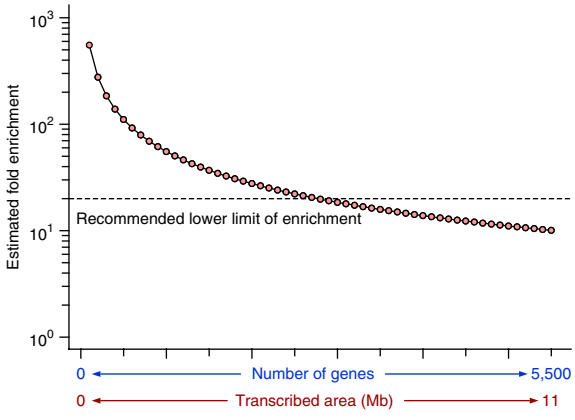

Figure 4 | Estimated fold enrichment achieved relative to the number of genes targeted. The estimated fold enrichment was calculated according to the number of genes targeted by the probe design. Estimated fold enrichment was calculated based on mean expression of a random selection of genes in three replicates of the K562 human cell type[45].

Control probe design. We recommend the inclusion of control probes to appraise capture performance. For example, control probes targeting nontranscribed genomic regions can indicate DNA contamination and measure false-positive alignment and transcript assembly. The inclusion of probes to housekeeping genes provides positive controls that also help evaluate correct transcript assembly

Box 2 | Control probe design and analysis

We recommend the inclusion of the following control probes that permit the assessment of CaptureSeq performance:

- Probes targeting a nontranscribed intergenic region (<100 kb) to identify gDNA contamination.
- Control probes targeting a subset (>5) of endogenous genes that cover a range of expression levels (avoiding highly expressed genes). These genes can be used to determine library enrichment by qPCR before sequencing and evaluate read alignment and transcript assembly parameters.
- Sequences from *E. coli* or other common laboratory species to detect library contamination.
- If investigating the expression of single exon transcripts or sites of transcriptional initiation or termination, users can include nearby intergenic regions (of at least a random subset), to help distinguish signal from noise.
- Control probes targeting a subset of ERCC RNA spike-in standards[24] to assess performance, sensitivity, dynamic linear range, enrichment and off-target capture. The RNA spike-ins standards consist of 92 *in vitro*–transcribed polyadenylated transcripts, combined at varying abundance spanning an ~10^6-fold range in concentration, that are included during initial RNA sample preparation according to the manufacturer's instructions. To minimize the reduction in global enrichment caused by sequencing of highly expressed control probes instead of captured gene targets, we suggest not targeting the ERCC spike-in probes present at the highest three or four concentrations (unless the user aims to analyze and profile highly expressed genes). Capturing a subset of the controls also allows the measurement of off-target capture. Details of probes are available in the ERCC RNA spike-in control mixes user guide.

Linear dynamic range

CaptureSeq can perform quantitative measurement of targeted genes. However, targeting abundant genes can saturate probes, imposing an upper limit where the CaptureSeq dynamic range flattens. To determine this upper limit, we plot the FPKM (fragments per kilobase per million mapped reads) measurement for each RNA standard against its known molar concentration, and use linear regression to determine the best-fit line and then perform segmented regression to determine the upper limit to the CaptureSeq dynamic range. Detection of ERCC probes also indicates the limit of sensitivity with which weakly expressed transcripts can be detected. CaptureSeq is sensitive enough to routinely detect the lowest molar amount of ERCC probes present in the sample.

To illustrate this, we have plotted the measured abundance of ERCC RNA spike-ins (in FPKM) against their known concentrations within the example data (**Fig. 6a**), observing a linear relationship across the full dynamic range. This reveals that probes targeting even the most abundant ERCC RNA spike-ins have not become saturated. Similarly, we do not detect the least abundant ERCC spike-ins, indicating the lower limit of sensitivity. This indicates that the enriched sequence coverage of the example library has not achieved saturation, probably owing to both the large number of transcripts targeted for captured within this design and also the relatively shallow sequenced library depth. By comparison, we have included a plot of ERCC concentration and measured abundance derived from a CaptureSeq experiment with sufficient sequencing depth and enrichment to approach saturated sequencing coverage of ERCC RNA spike-ins (**Fig. 6b**). Robust sequencing coverage is achieved for all concentrations of ERCC spike-ins, indicating high sensitivity and low variability.

An inflection point determined by segmental linear regression identifies an upper limit to the CaptureSeq quantitative range owing to probe saturation.

Endogenous gene quantification

Comparison of the measured FPKM of endogenous transcripts with the FPKM of ERCC spike-ins at known concentrations provides an estimate of endogenous transcript concentration within the initial RNA sample.

Off-target capture

The capture of novel exons that are spliced to captured exons can prevent a clear measure of off-target capture. Given that highly abundant ERCC spike-ins may not be targeted because they may reduce global enrichments, these abundant ERCC spike-ins provide an ideal measure of off-target capture, and they indicate the stringency of the CaptureSeq enrichment.

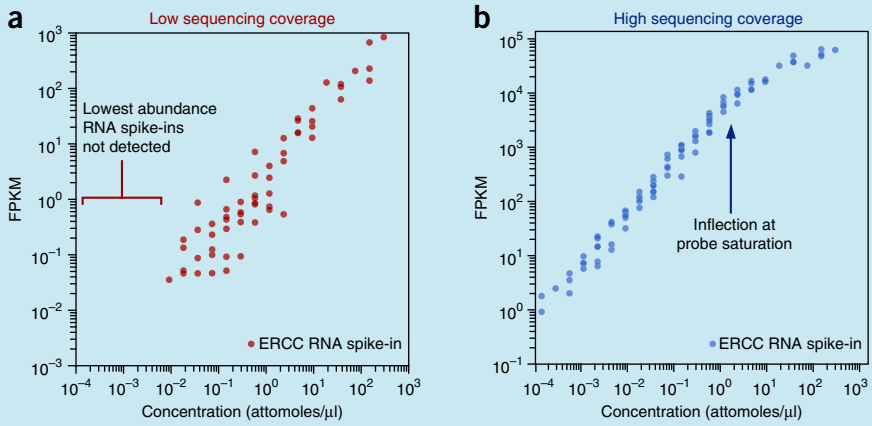

Figure 6 | Analysis of ERCC RNA spike-ins to assess the performance of CaptureSeq experiments. (**a**,**b**) Scatter plot indicating the measured abundance of ERCC RNA spike-ins relative to known concentration within a sample, for CaptureSeq example data with low sequencing coverage (**a**) and high sequence coverage (**b**). (**a**) The absence of detection of low-abundance ERCC spike-ins and high variability of measured abundance indicates low sequencing coverage due to shallow library sequencing depth and/or targeting a large portion of transcriptome. (**b**) Low variability and representative sampling and the presence of an inflection point followed by increasing underestimation of measured abundance owing to probe saturation indicate high sequencing coverage achieved by CaptureSeq.

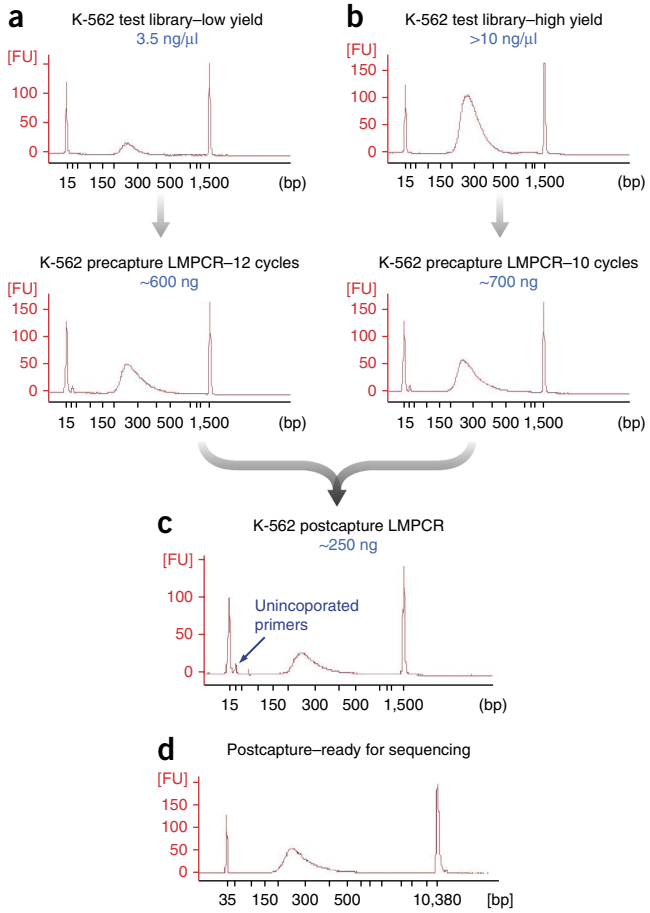

Figure 5 | Example Agilent Bioanalyzer results for pre- and postcapture libraries. (**a**,**b**) Examples of K-562 test libraries with high and low yields owing to different RNA inputs into the Illumina TruSeq library preparation and the generation of sufficient library yield with different numbers of cycles during pre-capture LMPCR. Libraries show typical fragment size ranges. (**c**) Agilent high-sensitivity chip showing K-562 postcapture LMPCR library. The library shows typical size range, with blue arrow indicating small nucleic acid species, probably unincorporated primers, which, if present, should be removed by a second round of purification. (**d**) Example of a purified postcapture LMPCR library ready for sequencing.

parameters. Probes targeting RNA Spike-In controls added during sample preparation can also indicate the quantitative accuracy, dynamic range and sensitivity of each capture and aid comparative analysis between different CaptureSeq experiments. This protocol includes the use of External RNA Control Consortium (ERCC) ExFold RNA spike-in mixes, which are added to the RNA sample before capture (see **Box 2** for further details)[24].

RNA input. Although many RNAseq library preparation protocols require as little as 100 ng of total RNA or 10 ng of rRNA-depleted RNA, the resultant cDNA library yield may be insufficient for subsequent capture steps (encompassing hybridization, wash and elution steps). We find that 5 µg of total RNA will give a final yield of at least 250 ng (and more commonly ~500 ng) of amplified cDNA library that is ready for capture. However, we recommend preparing at least 1.15 µg of cDNA library per capture (1 µg for capture steps, 150 ng for later quantitative PCR (qPCR)). Although additional library amplification can increase the library yield, this increases the effect of PCR amplification biases and there is an added risk of transcript 'drop-out' when targeting weakly expressed transcripts from a small RNA sample input. Therefore, one effective strategy to generate sufficient library input is to combine multiple multiplex-prepared samples in a single capture hybridization (see 'Multiplexing' below).

Minimizing PCR amplification cycles. PCR amplification is required for both cDNA library preparation before capture and for the amplification of the postcapture library before sequencing. These two stages of PCR amplification can risk unwanted PCR amplification artifacts or biases. To minimize the effect of PCR amplification artifacts, we reduce the number of amplification cycles as much as possible during the procedure. For example, before precapture ligation-mediated PCR (LMPCR; using Phusion polymerase) we create test libraries to estimate the minimum number of cycles required. In our experience, low test library concentrations of 2.5–4 ng/µl require 12 precapture LMPCR cycles to generate >450–500 ng of cDNA library yield (**Fig. 5**). Concentrations between 5 and 9 ng/µl require 11 cycles to generate >500 ng of cDNA library, and a test library concentration of >10 ng/µl requires 10 cycles of pre-capture LMPCR to provide more than 500 ng of library yield. Contrary to expectations, higher test library concentrations often do not result in proportionally higher yield (for example, 20 ng/µl doesn't generally provide >500 ng after nine cycles). Note that differing conditions and polymerases will affect the number of cycles required.

Evaluating the postcapture library. We recommend using qPCR to evaluate the fold enrichment and capture performance before commencing sequencing. Endogenous control genes, capture targets and/or ERCC RNA spike-ins can be compared between pre- and postcapture libraries to estimate the fold enrichment achieved. Similarly, measuring the fold depletion of transcripts not targeted by the capture probes provides a measure of capture stringency.

Multiplexing. CaptureSeq is compatible with multiplexed library preparation, permitting multiple libraries to be processed in a single capture (hybridization/wash/elution) reaction. This permits CaptureSeq to be used to efficiently profile a gene set of interest in large numbers of samples at reduced sequencing costs, which is a major advantage for gene expression profiling applications. A further benefit is that multiple uniquely bar-coded libraries can be combined to provide sufficient library yield input for capture. However, it is important that blocking or hybridization-enhancing (HE) oligonucleotides are coordinated with multiplex bar-code sequences (**Supplementary Data**). Despite these advantages of multiplexing, users should be aware that precapture pooling could potentially result in PCR recombination artifacts. These artifacts can be mitigated by the use of double index primers, if required[25].

Performing a control DNA capture experiment. Although it is not strictly required, a control capture using matched genomic DNA (gDNA) can be used to identify anomalous probe hybridization artifacts and normalize RNA expression by the efficiency with which each oligonucleotide probe captures DNA. This control is a standard gDNA capture, and it should not be performed by using

the protocol described below. Instead, we recommend following the standard gDNA capture protocol, described in the Roche/NimbleGen SeqCap EZ library SR User's Guide (see Equipment), with the minor amendments of fragmenting gDNA to a similar size range to RNAseq libraries and making corresponding adjustments to AMPure DNA cleanup steps. Similarly to cDNA capture, extra precapture LMPCR cycles may also be required to generate sufficient DNA for capture.

CaptureSeq data analysis. The computational analysis of CaptureSeq data is similar to that of other RNAseq applications. The wide range of tools for sequence read quality control, aligning reads to the genome or transcriptome and across splice junctions, assembling aligned reads into full-length transcripts and quantifying genes are applicable and well documented[12,26–28]. We recommend that users familiarize themselves with RNAseq analysis protocols that provide detailed information for downloading, installing and running the required software[3,19]. Within this protocol, we have only focused on computational considerations specific to the analysis of CaptureSeq data.

We have assumed that users operate software through the UNIX shell command line. Users unfamiliar with the UNIX shell, or without access to sufficient computational resources, can perform data analysis within GenePattern[29] or the Galaxy Project[30], which provides web-interface access to cloud-computing bioinformatic resources, including TopHat2 (ref. 31) and Cufflinks[26], in an intuitive graphical format.

MATERIALS

REAGENTS

▲ CRITICAL Although we generally recommend that most reagents be purchased from the listed companies, generic reagents (such as Cot-1, nuclease-free water and so on) can be purchased from alternative vendors. Most equipment can also be purchased from manufacturers other than those listed, and some procedures (such as nucleic acid quantification or qPCR) can be equivalently performed with alternative protocols. For software packages listed below, we recommend using the most recent release version, as well as reading corresponding documentation.

- Sodium acetate, 3 M, pH 5.0
- Agarose (for DNA electrophoresis)
- Agencourt AMPure XP, 60-ml kit (Beckman Coulter Genomics, cat. no. A63881)
- Agilent DNA 1000 kit (Agilent Technologies, cat. no. 5067-1504)
- Agilent high-sensitivity DNA kit (Agilent Technologies, cat. no. 5067-4626)
- Agilent RNA 6000 nano kit (Agilent Technologies, cat. no. 5067-1511)
- Agilent RNA 6000 pico kit (Agilent Technologies, cat. no. 5067-1513)
- Cot-1 human DNA, fluorometric grade, 1 mg/ml; 1 ml (Life Technologies, cat. no. 15279-011)
- DNA electrophoresis buffer (1× sodium boric acid buffer (35 mM boric acid, pH to 8.5 with sodium hydroxide))
- DNA electrophoresis loading dye and ladder
- dNTP set (Life Technologies, cat. no. 10297-018)
- Dynabeads M-270 (Life Technologies, cat. no. 65305)
- EPH buffer (2× FPF buffer; Illumina—supplied as special request item)
- ERCC ExFold RNA spike-in mixes (Life Technologies, cat. no. 4456739)
- Ethanol (absolute)
- NimbleGen SeqCap EZ choice/choice XL libraries (solution-based captures)
- NimbleGen SeqCap EZ hybridization and wash kit (Roche Diagnostics, cat. no. 05634261001 24rxns)
- Oligonucleotides (for sequences and concentrations, see **Supplementary Data**)
- Phusion high-fidelity (HF) DNA polymerase with 5× HF buffer, 100 reactions (NEB Finnzymes, cat. no. M0530S)
- QIAquick PCR purification kit, 50 (Qiagen, cat. no. 28104)
- Ribo-Zero rRNA magnetic kit (Epicentre, cat. no. MRZH116)
- RNA sample(s) for analysis (sample(s) should be of sufficient quality (RNA integrity number (RIN) >6–7))
- RNase-/DNase-free water (Life Technologies, cat. no. 10977-023)
- RNeasy MinElute cleanup kit (Qiagen, cat. no. 74204)
- Species-specific gDNA (or similar)
- SuperScript II reverse transcriptase (Life Technologies, cat. no. 18064-014)
- SYBR Green PCR master mix, 5ml (Applied Biosystems, cat. no. 4309155) (alternative quantitative PCR methods can be similarly used for accurate quantification[32])
- SYBR Safe DNA gel stain (Applied Biosystems, cat. no. S33102)
- Taq DNA polymerase (Fisher Biotec, cat. no. TAQ-1) (Taq polymerase from alternative vendors can be similarly used for DNA contamination testing)
- TruSeq stranded mRNA sample prep kit (Illumina, cat. no. RS-122-2101 kit A)
- TURBO DNase (Life Technologies, cat. no. AM2238)

EQUIPMENT

- Nuclease-free PCR tubes, 0.2 ml (Eppendorf, cat. no. 951010006)
- DNA LoBind tubes, 1.5 ml (Eppendorf, cat. no. 022431021)
- RNase/DNase-free Falcon tubes, 15 ml (Corning, cat. no. 352095)
- RNase/DNase-free Falcon tubes, 50 ml (Corning, cat. no. 352070)
- PCR plates, 96 wells, 0.3 ml (Bio-Rad, cat. no. Hss-9601)
- Agarose gel electrophoresis equipment
- BD PrecisionGlide needle (18-G × 1 1/2 in) (Becton Dickinson, cat. no. 302032)
- Benchtop centrifuges or microcentrifuges for 1.5-ml and 0.2-ml tubes
- Benchtop centrifuge (that can centrifuge 96-well plates)
- Bioanalyzer (Agilent Technologies)
- DynaMag-2 magnet, to hold 1.5-ml tubes (Invitrogen, cat. no. 123-21D)
- Freezers (−20 °C, −80 °C)
- Magnetic stand 96 (Ambion, cat. no. AM10027)
- Microseal 'B' adhesive seals (Bio-Rad, cat. no. MSB-1001)
- NanoDrop (Thermo Scientific) (Alternative nucleotide acid quantification methods, such as the QuBit fluorometer, can be equivalently used)
- PCR thermocycler(s) suitable for 0.2-ml tubes, 0.3-ml 96-well plates
- Pipettors, l–10 µl, 20 µl, 200 µl, 1,000 µl
- qPCR machine
- Refrigerator, 4 °C
- Vacuum concentrator
- Vortex mixer
- Water bath and/or heating blocks

Software

- Bowtie2 (ref. 33) (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml) is used to align reads by TopHat2 and also used to estimate the fold enrichment
- TopHat2 (ref. 31) (http://tophat.cbcb.umd.edu/) is used within this protocol to align sequenced reads to the genome and across known and novel splice junctions. Alternative software, such as STAR[34], SpliceMap[35] and GSNAP[36], can be similarly used
- Cufflinks[37] (http://cufflinks.cbcb.umd.edu/) is used within this protocol to assemble full-length transcripts from short-read alignments. Alternative software, such as iReckon[38] and SLIDE[39], can be similarly used
- SAMtools[40] (http://samtools.sourceforge.net/) permits manipulation of SAM- or BAM-formatted files
- BEDTools[41] (https://github.com/arq5x/bedtools2/) permits the manipulation BAM-, GTF- and BED-formatted files
- Index for human reference genome and gene annotations (available for download at http://tophat.cbcb.umd.edu/igenomes.shtml)
- (Optional) A gene annotation file (.gtf format) to aid Cufflinks-mediated transcript assembly: we recommend using the most recent and comprehensive GENCODE annotation available at http://www.gencodegenes.org/releases/
- (Optional) Genome browser for visualization of read alignments and assembled transcripts: e.g., the Integrated Genome Viewer (http://www.broadinstitute.org/software/igv/) or the University of

California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu). The UCSC Genome Browser also comprises a useful repository for gene annotations and permits simple bioinformatics queries. We recommend that users become familiar with using the UCSC Genome Browser[42] by completing the tutorial and training modules (http://genome.ucsc.edu/training.html)

Additional documentation referred to within the protocol

- Agilent DNA 1000 kit quick start guide, part no. G2938-90015 (revision B)
- Agilent high-sensitivity DNA kit guide, part no. G2938-90322 (revision C)
- Agilent RNA 6000 nano kit quick start guide, part no. G2938-90037 (revision C)
- Agilent RNA 6000 pico kit quick start guide, part no. G2938-90049 (revision C)
- ERCC RNA spike-in control mixes user guide, part no. 4455352 (revision D)
- Illumina TruSeq stranded mRNA sample preparation guide, part no. 15031047 (revision D; September 2012)
- QIAquick spin handbook (May 2012)
- Ribo-Zero magnetic kit protocol. lit. no. 335 (April 2013)
- Roche/NimbleGen SeqCap EZ library SR user's guide, v 4.0 (January 2013)
- TURBO DNase, part no. 1907M (revision G)

REAGENT SETUP

Minimizing PCR contamination Because of the sensitivity of CaptureSeq, it is recommended that reagents required for PCR amplification be divided into separate stocks for pre- and postcapture LMPCR amplification steps. Ideally, all pre-PCR amplification steps should be performed in a separate location and with separate equipment for post-PCR steps.

Probe library preparation Upon receipt of SeqCap EZ probe library, thaw the tube(s) on ice. Once they have thawed, vortex the tube(s) for 3 s and centrifuge each tube at 10,000*g* for 30 s at room temperature (20–25 °C) to accumulate the solution at the bottom of the tube. Transfer 4.5-µl aliquots of SeqCap EZ probe library into 0.2-ml PCR tubes to create single-use aliquots and store them at −20 °C until use. The sequence capture libraries are potentially sensitive to multiple freeze-thaw cycles, and they can be stored at −20 °C for at least 6 months.

HE oligo preparation If you are starting from lyophilized oligos, make up 1,000 µM stocks of HE oligos in nuclease-free water. Divide the stocks into smaller volumes to minimize future freezing and thawing. For all HE oligos corresponding to indexes, also prepare several aliquots at 100 µM. Freeze those aliquots that are not required for immediate capture hybridization at −20 °C.

EQUIPMENT SETUP

Maintaining hybridization and wash temperatures Several steps during capture, including hybridization and washing, are highly sensitive to time delays and temperature changes. To prevent delays and to prevent samples from cooling when being transferred between equipment, we recommend that these steps be performed with localized equipment. If this is not possible, the samples should be stored in a heated block to maintain temperature when being transferred between equipment. Thermocyclers should be programmed with the required reaction programs before use.

Example data Data from M.B.C. *et al.* (unpublished data) are used here to demonstrate probe design and data analysis. The library comprises a single replicate of ~20 million reads from human K562 cells with a long noncoding RNA (lncRNA)–specific probe design. This library is chosen for its small memory requirements and fast processing time; however, the effective sequencing depth is less than the depth we recommend for novel gene discovery.

Raw sequencing reads and probe design coordinates, as well as anticipated results (coverage and assembled transcripts), are available through the Gene Expression Omnibus (accession no. GSE52503). Probe design coordinates targeting long noncoding RNAs are based on publicly available annotations from GENCODE (v12), lncRNAdb[43], Cabili *et al.*[44] and proprietary annotations. This includes GENCODE long intergenic noncoding RNA (lincRNA) annotations whose retrieval is described within the protocol (**Box 1**), but it also targets additional annotated lncRNAs.

An RNAseq library of the human K562 cell line is used to estimate fold enrichment before probe manufacture; it can be downloaded from UCSC/ENCODE[45] (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq/wgEncodeCaltechRnaSeqK562R1x75dFastqRep1.fastq.gz).

PROCEDURE

Preparation of DNA-free, rRNA-depleted sample RNA ● TIMING ~1 d

▲ CRITICAL For all steps involving RNA, perform experiments under RNase-free conditions with RNase-free reagents and labware. RNA samples should be kept on ice when thawed.

1| CaptureSeq is an extremely sensitive technique, and DNA contamination can confound interpretation of RNAseq results. Validate that the sample RNA is free from gDNA contamination by conducting PCR amplification for a short region of gDNA (~100 nt) by using 200 ng of RNA as a template. Include a positive gDNA and a negative H_2O control. Primer sequences to amplify a 95-nt gDNA product within the human nuclear paraspeckle assembly transcript 1 (*NEAT1*) loci are provided within the **Supplementary Data**. A variety of Taq polymerases and buffer conditions will be suitable; an example PCR assembly is as follows:

Component	Amount (µl)	Final concentration
RNA, 200 ng/µl	1	10 ng/µl
dNTP, 10 mM (2.5 mM each)	0.8	0.4 mM
Fisher Biotec 10× reaction buffer	2	1×
Primer (forward, 10 µM)	0.5	250 nM
Primer (reverse, 10 µM)	0.5	250 nM
$MgCl_2$, 25 mM	1.2	1.5 mM
Fisher Biotec Taq DNA polymerase, 5 U/µl	0.25	1.25 U
Distilled water	13.75	
Total	20	

2| Perform a 35-cycle PCR by using the following cycling conditions:

Cycle number	Denature	Anneal	Extend
1	94 °C, 3 min		
2–36	94 °C, 30 s	55 °C, 30 s	72 °C, 25 s
37			72 °C, 5 min
Final			4 °C, hold

3| Confirm that the RNA samples are free of DNA by agarose gel electrophoresis of the PCR product(s). Cast a 2.5% (wt/vol) agarose gel in DNA electrophoresis buffer (1× sodium boric acid buffer) with either 1× SYBR Safe dye or ethidium bromide. Run the gel at 170 V for 20 min to verify that amplification occurred for the positive control sample alone, and not for the RNA samples. If no DNA contamination is present in the RNA sample, proceed directly to Step 4. However, if DNA contamination is present, treat the RNA with TURBO DNase (as per the manufacturer's instructions), purify it by using either standard phenol/chloroform extraction or an RNeasy column (as per the manufacturer's instructions) and then repeat Steps 1–3 to recheck for DNA contamination.

4| Check the RNA sample integrity with an Agilent Bioanalyzer RNA Nanochip. Follow the Agilent RNA 6000 nano kit guide and select a Total RNA assay in the Bioanalyzer software. High RIN scores (>8) are ideal. Partially degraded RNA samples (RIN scores between 6 and 8) can be used but may benefit from a shorter fragmentation time (reduction from 8 to 6 min) (Step 10) before Illumina RNA TruSeq library construction. We have not validated this protocol on RNA with RINs <6; other library construction methods may perform better for degraded RNA[46].
■ **PAUSE POINT** DNA-free RNA samples can be stored at –80 °C indefinitely.

5| rRNA-deplete the sample RNA by using Ribo-Zero purification as per the Ribo-Zero magnetic core kit protocol. Purify the rRNA-depleted sample RNA by using RNeasy MinElute columns according to the protocol in Appendix A of the Ribo-Zero magnetic kit. The ERCC RNA spike-in control mix can be added at this step before Ribo-Zero purification (instead of at Step 8 below); see the ERCC RNA spike-in control mixes user guide to calculate the correct amount to be added.
▲ **CRITICAL STEP** Given the abundance of rRNA compared with capture targets (often 1% or less of the 'mRNA' fraction), rRNA depletion is important to ensure that off-target rRNA does not erode enrichment levels. An additional concern is that highly abundant rRNA may preclude amplification of the less-abundant species during the initial PCR amplification step that precedes capture, probably resulting in a less-informative capture outcome.
▲ **CRITICAL STEP** If required, the samples can be divided into smaller volumes for more efficient parallel rRNA depletion.
■ **PAUSE POINT** rRNA-depleted RNA samples can be stored at –80 °C for at least 1 week.

6| Determine the concentration of rRNA-depleted RNA samples. Analyze 1 µl from each sample on a NanoDrop according to the manufacturer's instructions to provide a concentration estimate for correct Bioanalyzer loading. Dilute 1 µl of each rRNA-depleted RNA sample for analysis and quantification with an Agilent Bioanalyzer according to the Agilent RNA 6000 pico kit guide and by selecting mRNA assay in the Bioanalyzer software. Confirm successful rRNA depletion of RNA. The expected yield is ~2–10% of the original input RNA.
▲ **CRITICAL STEP** If multiple ribo-depletions were performed from the same total RNA sample (at Step 5), pool rRNA-depleted samples before preparing sequencing libraries.
? TROUBLESHOOTING

Preparation of sequencing libraries with the Illumina TruSeq stranded mRNA kit ● TIMING ~2 d

7| A maximum of 400 ng of rRNA-depleted sample can be used as input for each sequencing library preparation. If the rRNA-depleted RNA yield is greater than 400 ng, add 400 ng of RNA into a new tube and adjust the volume to 8 µl by adding nuclease-free water. If the rRNA-depleted RNA yield is less than 400 ng, reduce the RNA volume to 8 µl with a vacuum concentrator (do not apply heat).

8| Add 1 µl of ERCC RNA spike-in control at an appropriate dilution, as indicated in the user guide for the ERCC RNA spike-in control mixes.
▲ **CRITICAL STEP** ERCC RNA spike-in mixes provide important information for sequencing analysis (**Box 2** and **Fig. 6**), and they are added in this step (if not previously added before rRNA depletion).

9| Add 9 µl of each rRNA-depleted RNA sample (now including ERCC RNA spike-in) to a well in a 96-well plate. Add 9 µl of EPH buffer and mix it thoroughly by pipetting. Return the EPH buffer to –20 °C and seal the 96-well plate with Microseal 'B' adhesive seal.

10| Fragment the RNAs in a thermocycler (heat the plate to 94 °C for 8 min and then cool it to 4 °C). Centrifuge the plate at 280*g* for 10 s, if required. Proceed immediately to the next step.

11| Prepare RNA-sequencing libraries according to the Illumina TruSeq stranded mRNA sample preparation guide. Start at chapter 'Synthesize first strand cDNA'. Proceed through the Illumina TruSeq stranded mRNA sample preparation guide protocol (including 'Synthesize first strand cDNA', 'Synthesize second strand cDNA', 'Adenylate 3' ends' and 'Ligate adaptors' chapters) to a final step of 'Ligate adaptors' where unamplified libraries are ready to be transferred from the Clean-up ALP (CAP) plate.
▲ **CRITICAL STEP** If libraries are being prepared for multiplex capture or sequencing, it is important to ensure that each library has a different index. Correct multiplex index combinations (provided in the Illumina TruSeq stranded mRNA sample preparation guide) are important, and they must be coordinated with corresponding blocking oligonucleotides used in Step 29.

12| Transfer 20 µl of the supernatant from each well of the CAP plate to a new 96-well plate and store the unamplified cDNA library for future precapture LMPCR (Step 19 below).
▲ **CRITICAL STEP** A volume of 20 µl of unamplified cDNA library is saved for precapture LMPCR (Step 19), whereas 1 µl of the library remaining in the CAP plate is used to create an amplified test library for assessing library quality.
■ **PAUSE POINT** The unamplified 20-µl sample can be stored for at least 2 weeks at –20 °C.

13| The CAP plate should now have ~2.5 µl of supernatant and beads remaining. Carefully remove 1 µl of supernatant (without beads), and transfer it to a new 96-well plate labeled with an Illumina PCR bar code.

14| Dispense 19 µl of Illumina resuspension buffer into each well containing 1 µl of supernatant and mix it well by pipetting. This plate is used to amplify a test library.
■ **PAUSE POINT** The 96-well plate of unamplified test libraries can be stored for at least 1 week at –20 °C.

15| Create the test library (to assess the quality of the sequencing library before performing precapture LMPCR) by following the protocol in the Illumina TruSeq stranded mRNA sample preparation guide/'Enrich DNA fragments' chapter.
■ **PAUSE POINT** Purified amplified test library PCR plate can be stored for at least 1 week at –20 °C.

16| Measure the purified test library yield by loading 1 µl of the test library on an Agilent DNA 1000 Chip according to the Agilent DNA 1000 kit guide. The library size should range from 180 to 500 nt, with a peak at ~260–280 nt (**Fig. 5a**,**b**). Performing a 'Smear analysis' on this size range by using the 2100 Expert software provides a ng/µl value for the test library. Although concentrations above 10 ng/µl are desirable, libraries that provide much lower yields can still be successfully used; see 'Minimizing PCR amplification cycles' in Experimental design for further guidance.
? TROUBLESHOOTING

Precapture LMPCR ● TIMING ~3 h

17| Use the yield from the amplified test library to calculate the number of cycles required for precapture LMPCR of the 20-µl unamplified library from Step 12. See 'Minimizing PCR amplification cycles' in Experimental design for guidance.
▲ **CRITICAL STEP** Performing as few PCR amplification cycles as possible to generate precapture LMPCR libraries will reduce the effect of PCR amplification artifacts. The number of cycles required will depend on the polymerase used and the amount of library required for capture.

18| Thaw the reagents and prepare LMPCR reactions on ice. A water control is required for each precapture LMPCR, and thus even a single library will need a minimum 2× LMPCR master mix. The volumes necessary for a 1× master mix are given below; however, creating 5–10% additional PCR master mix is recommended.

Component	Amount (µl)	Final concentration
Phusion HF buffer, 5×	20	1×
dNTPs, 10 mM	2	200 µM
TS-PCR oligo1, 100 µM	2	2 µM
TS-PCR oligo2, 100 µM	2	2 µM
Nuclease-free water	53	
Phusion DNA polymerase, 2 U/µl	1	2 U
Total	80	

19| Add the 20 µl of unamplified cDNA library (from Step 12), or nuclease-free water (for the negative control), into a PCR tube or well. Add 80 µl of the LMPCR master mix into each tube or well. Mix the sample and PCR master mix by gentle pipetting (about five times). Perform precapture LMPCR with the following cycling conditions:

Cycle number	Denature	Anneal	Extend
1	98 °C, 30 s		
2–user defined	98 °C, 30 s	60 °C, 30 s	72 °C, 30 s
Final			72 °C, 5 min
			4 °C, hold

■ **PAUSE POINT** Amplified precapture LMPCR cDNA library can be stored for 3 d at 4 °C, or at least for 1 week at –20 °C.

20| Transfer PCR reactions to separate 1.5-ml tubes, and clean up each amplified precapture LMPCR cDNA library and negative control by using either QIAquick PCR purification (option A) or AMPure XP beads (option B)

(A) QIAquick PCR purification kit

(i) Follow the QIAquick PCR purification kit user guide, ensuring that the library is eluted in 50 µl of nuclease-free water (pH 7.0–8.5).

(B) AMPure XP beads

(i) Place AMPure XP beads at room temperature. Allow them to warm for 30 min before use.
(ii) While the beads are warming in Step 20B(i), prepare fresh 80% (vol/vol) ethanol sufficient for Step 20B(vi–viii) below.
(iii) Resuspend room temperature beads by vortexing to ensure a homogeneous mixture.
(iv) Add a 1.8× volume of beads to each PCR (i.e., add 180 µl to a 100-µl PCR), mix by brief vortexing and then incubate the beads for 15 min at room temperature.
(v) Place each tube on a magnetic plate or stand. Once the solution has cleared, without disturbing the beads, remove the supernatant and discard it.
(vi) Keep the tubes on the magnetic stand and add 200 µl of 80% (vol/vol) ethanol (from Step 20B(ii)) to each tube.
(vii) Incubate the tubes for 30 s, and then, without disturbing the beads, remove all of the ethanol.
(viii) Repeat Step 20B(vi, vii) for a total of two ethanol washes.
(ix) Leave the tubes in the magnetic plate/stand with lids open for up to 15 min to remove all ethanol traces and to dry the beads. Once the beads or the tube is dry, continue with the next step. Do not overdry (the bead cluster appears cracked) the beads, as this inhibits elution of DNA from the beads.
(x) Take the tubes off the magnetic stand and resuspend the beads in 52 µl of nuclease-free water by pipette-mixing ten times. If the beads were overdried, perform extra pipette mixes.
(xi) Incubate the tubes for 2 min at room temperature, and then place them on a magnetic stand. Once the solution has cleared, collect 50 µl of supernatant (contains cleaned-up LMPCR library) and transfer it to a new 1.5-ml tube.

■ **PAUSE POINT** Purified precapture LMPCR library can be stored for at least 3 months at –20 °C.

21| Validate successful precapture LMPCR library construction by analyzing 1 µl of purified DNA with an Agilent DNA 1000 chip according to the Agilent DNA 1000 kit guide. The library size should range from 180 to 500 nt, with a peak at ~260–280 nt (**Fig. 5a**,**b**). Performing a smear analysis on this size range by using the 2100 Expert software provides a ng/µl value for the precapture LMPCR library. Negative-control PCR reactions should contain no amplicon peaks between 180 and 500 nt; however, a small peak corresponding to unincorporated primers is not a concern.

? TROUBLESHOOTING

Hybridization of precapture LMPCR Libraries to SeqCap EZ probe library ● TIMING ~2 h plus 3 d of incubation

▲ **CRITICAL** Evaporation during the 3-d incubation can result in the failed enrichment of target genes. Before performing an experimental hybridization, confirm that tubes or plates to be used experience little or no sample evaporation during a 72-h incubation at 47 °C (a maximum 3 µl of evaporation from 15 µl is acceptable). Although we find that the plates, seals and 0.2-ml tubes listed in the Equipment section perform well, and other suppliers are similarly appropriate, we strongly recommend performing a prior test for evaporation.

22| Set a thermocycler to 47 °C (with the lid at 57 °C) and a 1.5-ml tube heat block to 95 °C. If required or available, set a second thermocycler or 0.2-ml tube heat block to 47 °C. Thaw Cot-1 DNA, as well as the 2× hybridization buffer and hybridization component A (from the NimbleGen SeqCap EZ hybridization and wash kit) at room temperature and place the reagents on ice once thawed.

23| Thaw the required number of 4.5-µl aliquots of SeqCap EZ capture oligonucleotide probes on ice.

24| Thaw the required precapture LMPCR libraries on ice.

25| Thaw 1,000 µM Universal HE oligo and required 100 µM index HE oligo(s) on ice.
▲ **CRITICAL STEP** If multiple libraries are being captured in the same multiplex hybridization, it is crucial to ensure that each library has a different index and that matched HE index oligonucleotides are used in Step 29.

26| Prepare the libraries; a total of 1.15 µg per capture is recommended (1 µg for capture hybridization and 150 ng for qPCR). When you are multiplexing libraries in capture hybridization, mix the library aliquots together to obtain 1.15 µg. Adding equal nanogram amounts of each library is recommended.

27| Add 1 µg of prepared library DNA to a 1.5-ml tube. Add 5 µg of Cot-1 DNA.

28| Add 1 µl of 1,000 µM Universal HE oligo to the tube.

29| Add 10 µl (in total) of 100 µM index HE oligos to the tube. When the capture contains multiplexed libraries, ensure that index HE oligos match indexes on DNA libraries and are added in the same proportions. For example, if the multiplexed pool of libraries contains 200 ng of five libraries, add 2 µl of the correct five index HE oligos at 100 µM.

30| Close the 1.5-ml tube and use an 18-gauge (or similar) needle to make a hole in the lid.

31| Dry the tube containing libraries, Cot-1 and indexes at 60 °C in a vacuum concentrator.

32| Once dry, add the following from the NimbleGen SeqCap EZ hybridization and wash kit to the 1.5-ml tube: 7.5 µl of 2× hybridization buffer and 3 µl of hybridization component A. Cut the pierced cap off the 1.5-ml tube and replace it with the cap from another tube. Vortex the 1.5-ml tube for 10 s and centrifuge it at full speed for 10 s.

33| Denature the capture samples (precapture LMPCR library/Cot-1/HE oligo pool/hybridization cocktail) for 10 min at 95 °C in a heating block. If you are performing multiple capture hybridizations, consider starting the denaturation of each sample at intervals of 30 s^{-1} min. Continue with Steps 34 and 35 while the samples are denaturing.
▲ **CRITICAL STEP** Steps 34 and 35 should be performed rapidly to prevent sample evaporation or to prevent the reaction from cooling. We recommend that users familiarize themselves with the protocol and localize equipment to rapidly proceed through protocol steps.

34| When 5 min of the 95 °C denaturation step remain, transfer 0.2-ml PCR tube(s) containing thawed 4.5-µl aliquot(s) of SeqCap EZ capture probes from ice to room temperature.

35| Mix the capture sample(s) with SeqCap EZ capture probe aliquot(s) to begin hybridization. This step can be performed by using either (option A) a 96-well plate or (option B) 0.2-ml tubes.
▲ **CRITICAL STEP** When performing multiple capture hybridizations, consider starting hybridizations sequentially.

(A) Plate, 96 wells

(i) Place a 96-well plate with a microseal cover in a thermocycler to equilibrate it to 47 °C during the 10-min 95 °C denaturation step (Step 33).
(ii) With ~45 s of the 95 °C denaturation remaining, remove the microseal cover and leave the thermocycler lid open so that the plate is ready for sample loading (Step 35A(v)).
(iii) At the completion of the 95 °C denaturation step, immediately transfer the 1.5-ml tube containing the capture sample library to a benchtop microcentrifuge and spin for 10 s at full speed (16,000*g*) at room temperature to accumulate the solution at the bottom of the tube.
(iv) If a second thermocycler or 0.2-ml tube heat block is available, place SeqCap EZ capture probe aliquot to warm to 47 °C during centrifugation (Step 35A(iii)). Otherwise, leave the probe aliquot at room temperature.
(v) Immediately transfer the entire 10.5-µl capture sample to the 4.5-µl SeqCap EZ capture probe aliquot. Quickly but gently mix by pipetting (about five times) and transfer the solution to the 47 °C 96-well plate.
(vi) Ensure that the sample is well-mixed by pipette mixing a further 5–10 times with the plate in the thermocycler.
(vii) Seal the plate with a microseal cover, or, if several hybridizations are being performed, seal only the column of wells containing the sample. Ensure that the seal is tight to prevent evaporation. If several samples are being captured,

repeat Step 35A(iii–vii) until all samples have been processed. Close the thermocycler lid and begin hybridization (Step 36).

(B) Tube(s), 0.2 ml

(i) At the completion of the 95 °C denaturation, immediately transfer the 1.5-ml tube containing the capture sample to a benchtop microfuge and spin for 10 s at full speed (16,000*g*) at room temperature to accumulate the solution at the bottom of the tube.

(ii) During centrifugation (Step 35B(i)), place SeqCap EZ Capture probe aliquot in the thermocycler to warm to 47 °C.

(iii) Immediately transfer the entire 10.5-µl of capture sample(s) to 4.5-µl SeqCap EZ capture probe aliquot. Mix by pipetting (about ten times). Close the thermocycler lid. If several samples are being captured, repeat Step 35B(i–iii) until all samples have been processed, and then begin hybridization (Step 36).

36| Incubate the samples at 47 °C for 64–72 h in the thermocycler with the lid set to 57 °C. We recommend starting a timer upon hybridization to record the actual incubation time.

Binding captured DNA to Dynabeads and washing to remove nontarget DNA ● TIMING ~2.5 h, including incubations

▲ **CRITICAL** Many reagents require time to equilibrate to the required temperature. We recommend preheating the water baths or heat blocks to 47 °C, and equilibrating 47 °C buffers for 2 h.

▲ **CRITICAL** Do not let the capture sample temperature drop below 47 °C during binding and washing steps. Keep the sample tubes in a heated block at 47 °C if they need to be transferred between equipment. This protocol is designed for the concurrent washing of one or two capture reactions. If you are performing more reactions, consider mixing each sample with Dynabeads consecutively and staggering the start of the wash steps.

37| Assemble the following solutions (per reaction) by using the buffer concentrates from the SeqCap EZ hybridization and wash kit:

Buffer	Amount (µl)	Water (µl)	Total (µl)	Temperature (°C)
Stringent wash buffer	44	396	440	47
Wash buffer I, 10×	11	99	110	47
Wash buffer I, 10×	20	180	200	Room temperature
Wash buffer II, 10×	20	180	200	Room temperature
Wash buffer III, 10×	20	180	200	Room temperature
Bead wash buffer, 2.5×	210	315	525	Room temperature

Split the stringent wash buffer into two tubes of 220 µl. Equilibrate the 47 °C buffers for 2 h in a water bath or heat block.

38| At 40 min into the 2-h buffer equilibration, place the streptavidin Dynabeads at room temperature. Allow the beads to warm for 30 min before use.

39| Prepare and wash the Dynabeads. Resuspend the beads by vortexing for 15 s to ensure a homogeneous mixture. For each capture to be performed, transfer 100 µl of resuspended beads to a 1.5-ml tube (the beads for up to six captures can be prepared and washed in a single tube).

40| Place the tube on a magnetic plate or stand. Once the beads have bound and the solution has cleared, without disturbing the beads, remove the supernatant and discard it. While keeping the tubes on the magnetic stand, add 200 µl of 1× bead wash buffer for each capture being performed.

41| Remove the tube from the magnetic stand and resuspend it thoroughly by medium-speed vortexing for 10 s or by pipette mixing.

▲ **CRITICAL STEP** Dynabeads can adhere to the walls of some tubes under certain buffer conditions. This can be minimized by pipette mixing rather than by vortexing.

42| Perform Steps 40 and 41 a second time with 200 µl of 1× bead wash buffer per capture and then a third time with 100 µl of 1× bead wash buffer per capture, leaving the Dynabeads resuspended in 100 µl of 1× bead wash buffer per capture being performed.

43| Transfer 100-µl aliquots of the resuspended beads into 0.2-ml tubes.
▲ **CRITICAL STEP** It is crucial that the following steps (Steps 44 and 45, 48–52) be carried out quickly so that Dynabeads do not dry out and the sample temperature remains as close to 47 °C as possible. Allowing the sample to cool decreases the capture efficiency and hence enrichment. When multiple hybridizations are conducted, we recommend performing Step 45 (binding captured DNA to Dynabeads) one hybridization at a time, as small variations in incubation time (Step 46) are acceptable.

44| Place the 0.2-ml tube(s) containing 100 µl of Dynabeads on a magnetic plate or stand to clear the beads.

45| Remove the supernatant and resuspend Dynabeads with hybridization samples containing captured cDNA. If the hybridization was performed with a 96-well plate, follow option A, and if 0.2-ml tubes were used, follow option B.

(A) Plate (96 wells) hybridization

(i) Remove as much supernatant from the Dynabeads as possible (a residual amount of supernatant remaining is acceptable) and close the tube cap to prevent desiccation.
(ii) Remove the adhesive seal from the 96-well plate.
(iii) If a second thermocycler or 0.2-ml tube heat block is available, remove the 0.2-ml tube with Dynabeads from the magnetic stand and place it at 47 °C in the thermocycler or 0.2-ml tube heat block. Otherwise, remove it from the magnetic stand and leave it at room temperature.
(iv) Immediately add the hybridization sample to the Dynabeads. Note whether evaporation was substantial. Place the 0.2-ml tube at 47 °C in a thermocycler (if it is not there already) with the lid open and set the temperature to 57 °C. Mix by pipetting gently but thoroughly (about ten times). Optionally, if the hybridization sample and beads are not mixing well, vortex the 0.2-ml tube for 1 s, pulse-spin it for 1 s and return it to a 47 °C thermocycler and gently resuspend by pipette mixing. This needs to be completed within a few seconds to prevent the sample from cooling below 47 °C.
(v) Close the lid and begin incubation.

(B) Tube(s) (0.2 ml) hybridization

(i) Remove as much supernatant from the Dynabeads as possible (a residual amount of supernatant remaining is acceptable), and close the tube cap to prevent desiccation.
(ii) Transfer the 0.2-ml tube with Dynabeads to the 47 °C thermocycler (with the open lid set to 57 °C).
(iii) Immediately add the hybridization sample to Dynabeads and mix by pipetting gently but thoroughly (about ten times) in the thermocycler with the lid open. Note whether evaporation was substantial. Optionally, if the hybridization sample and beads are not mixing well, vortex the 0.2-ml tube for 1 s, pulse-spin it for 1 s and return it to the 47 °C thermocycler; gently resuspend the mixture by pipette mixing. This needs to be completed within a few seconds to prevent the sample from cooling below 47 °C.
(iv) Close the lid and begin incubation.
? TROUBLESHOOTING

46| Incubate the beads with the hybridization sample for 45 min at 47 °C in a thermocycler (lid, 57 °C). Resuspend 5–10 times every 15 min by pipette mixing with the tube lid open while the tube is still in the thermocycler.

47| During the 45-min incubation (Step 46), label and warm a 1.5-ml tube for each capture sample and place them in the 47 °C water bath or heat block.

48| After 45 min of incubation, add 100 µl of wash buffer I (preheated to 47 °C) to the 0.2-ml tube containing the hybridization sample and Dynabeads while the tube is still in the thermocycler. Mix the entire volume by gentle pipetting (about ten times) in the thermocycler block set to 47 °C.

49| Transfer the contents of the 0.2-ml tube(s) to the 1.5-ml tube(s) preheated in a 47 °C water bath or heat block (from Step 47).

50| Transfer the 1.5-ml tube to the magnetic stand to bind Dynabeads with a magnet. Remove the buffer once it is clear.

51| Return the 1.5-ml tube to a 47 °C water bath or heat block. Immediately add 200 µl of 1× stringent wash buffer (preheated to 47 °C) to the beads. Mix the entire volume by gentle pipetting (about ten times) while at 47 °C, and incubate it for 5 min at 47 °C.

52| Place the tube on a magnetic stand and remove the buffer once it is clear. If some Dynabeads do not initially aggregate to the tube side nearest the magnet, but accumulate at the bottom of the tube, use a pipette tip to gently blow the beads from the bottom of the tube to the side nearest to the magnet.

53| Repeat the washing step with 1× stringent wash buffer (Steps 51 and 52) for a total of two washes.

54| Perform the following washes at room temperature; each wash is followed by binding of Dynabeads with the magnet and removal of buffer on the magnetic stand.

Buffer	Amount (μl)	Mix
Wash buffer I	200	2 min vortexing
Wash buffer II	200	1 min vortexing
Wash buffer III	200	30 s pipette mix

▲ **CRITICAL STEP** Ensure that the Dynabeads do not stick to the sides of the tube and dry out, especially in buffers II and III. If necessary, decrease the speed of vortexing or use the pipette tip to push the Dynabeads back into the buffer.

55| After the final wash, remove the buffer and resuspend the Dynabeads (containing captured cDNA) in 50 μl of PCR-grade water.
■ **PAUSE POINT** Dynabeads containing captured cDNA can be stored at –20 °C for at least 2 weeks.

Postcapture LMPCR ● TIMING ~3 h
▲ **CRITICAL** To prevent cross-contamination, postcapture LMPCR should be performed with separate aliquots of reagents from those used for precapture LMPCR.
▲ **CRITICAL** To minimize amplification artifacts, it is important to perform as few cycles of PCR as possible when generating the postcapture LMPCR library. We recommend optimizing the number of PCR cycles required; however, if this is not possible, we recommend performing 17 cycles.
56| Thaw the PCR components and prepare LMPCR reactions on ice. For each sample, two reactions are performed that are subsequently combined. A water control is required for each postcapture LMPCR, and thus a single capture reaction will require a 3× master mix. Preparing additional PCR master mix (5–10%) is recommended. Assemble the following (for a 1× mix):

Component	Amount (μl)	Final concentration
Phusion HF buffer, 5×	20	1×
dNTPs, 10 mM	2	200 μM
TS-PCR oligo1, 100 μM	2	2 μM
TS-PCR oligo2, 100 μM	2	2 μM
Nuclease-free water	53	
Phusion DNA polymerase, 2U/μl	1	2 U
Total	80	

57| Resuspend the Dynabeads from Step 55 (containing captured cDNA) by pipette mixing. Add 20 μl of resuspended beads into two PCR tubes or wells. Add 20 μl of the nuclease-free water to a third tube or well (for a negative control).

58| Pipette 80 μl of postcapture LMPCR master mix into each tube or well. Mix by gentle pipetting (about five times). Proceed with postcapture LMPCR by using the following cycling conditions:

Cycle number	Denature	Anneal	Extend
1	98 °C, 30 s		
2–user defined	98 °C, 30 s	60 °C, 30 s	72 °C, 30 s
Final			72 °C, 5 min
			4 °C, hold

■ **PAUSE POINT** Amplified LMPCR DNA can be stored for 3 d at 4 °C, or at least for 1 week at –20 °C.

59| Clean up each postcapture LMPCR library and negative control by using either the QIAquick PCR purification kit (option A) or AMPure XP beads (option B).

(A) QIAquick PCR purification kit

(i) Pool the two reactions from each sample into 1.5-ml tubes.

(ii) Follow the QIAquick PCR purification kit user guide to purify DNA, noting that the maximum volume that can be added to a spin column at one time is 750 µl. To elute DNA, add 50 µl of EB (included in the kit) to each spin column and incubate it for 1 min before centrifugation.

(B) AMPure XP beads

(i) Place AMPure XP beads at room temperature. Allow them to warm for 30 min before use.

(ii) While the beads from Step 59B(i) are warming, prepare fresh 80% (vol/vol) ethanol sufficient for Step 59B(vii–ix) below.

(iii) While the beads from Step 59B(i) are warming, pool the two reactions from each sample into 1.5-ml tubes.

(iv) Resuspend the room temperature beads by vortexing to ensure a homogeneous mixture.

(v) Add a 1.8× volume of beads to each PCR reaction (i.e., add 360 µl to 200 µl of pooled PCR reactions), mix them by brief vortexing and then incubate them for 15 min at room temperature.

(vi) Place the tubes on a magnetic plate or stand. Once the solution has cleared, without disturbing the beads, remove the supernatant and discard it.

(vii) While keeping the tubes on the magnetic stand, add 200 µl of 80% (vol/vol) ethanol (from Step 59B(ii)) to each tube.

(viii) Incubate the tubes for 30 s, and then, without disturbing the beads, remove all of the ethanol.

(ix) Repeat Step 59B(vii, viii) for a total of two ethanol washes.

(x) Leave the tubes in the magnetic stand with lids open for up to 15 min to remove all ethanol traces and to dry the beads. Once the beads are dry, continue with the next step. Do not overdry the beads (the bead cluster appears cracked), as this inhibits elution of DNA from the beads.

(xi) Remove the tubes from the magnetic stand and resuspend the beads in 52 µl of nuclease-free water by pipette-mixing ten times. If the beads were overdried, perform extra pipette mixes.

(xii) Incubate the beads for 2 min at room temperature, and then place the tubes on a magnetic stand. Once the solution has cleared, collect 50 µl of the supernatant (contains amplified capture DNA) and transfer it to a new 1.5-ml tube.

■ **PAUSE POINT** The purified library can be stored for at least 3 months at –20 °C.

60| Validate successful postcapture LMPCR with 1 µl of purified captured DNA on Agilent Bioanalyzer according to the Agilent high-sensitivity DNA kit guide. The library size should range from 180 to 500 nt, with a peak at 260–280 nt (**Fig. 5c**,**d**). Negative-control PCRs should show no signal between 180 and 500 nt. The expected yield is greater than 250 ng, and typical yields range from 250 ng to 1 µg.

▲ **CRITICAL STEP** Samples containing a peak of unincorporated primers (extra peaks partially overlapping or near to the lower marker) will require an additional round of purification (repeat Step 59) before sequencing.

? TROUBLESHOOTING

qPCR for capture enrichment ● TIMING ~3 h, including qPCR run time

▲ **CRITICAL** Ensure that each primer used for qPCR has been efficiency-tested under the reaction conditions used in the experiment (including Roche/NimbleGen control primers). We recommend only using primers with primer efficiency (PE) above 1.9 (see Step 62 for further details).

61| Determine the capture enrichment by comparing transcript abundance between pre- and postcapture LMPCR samples by qPCR. We recommend testing 6–8 amplicons in triplicate and testing both enrichment of captured transcripts and the depletion of nontarget transcripts. The Roche/NimbleGen control amplicons 237, 268 and 272 are suitable for cDNA capture enrichment analysis, as are ERCC controls and design-specific capture transcripts. 3–5 ng of cDNA per well is generally sufficient for qPCR analysis of a wide range of targets. Users should consult the product literature for the qPCR machine available to them because reaction components, concentrations and volumes can vary substantially between qPCR models. For an introduction to qPCR, consult Derveaux *et al.*[32] or the qPCR handbook (Life Technologies).

62| For each enrichment amplicon tested, determine the relative fold enrichment. For CaptureSeq, the 'delta Ct' method is sufficient where enrichment = $PE^{\Delta Ct}$. Briefly, average the cycle threshold (Ct) for the triplicate samples and calculate the ΔCt value (average postcapture Ct value – average precapture Ct value). PE values can be measured by using the Ct slope method, where $PE = 10^{(-1/slope)}$, with a perfect efficiency giving a value of 2. Depletion ratios are calculated similarly. Depletion = $-1/(PE^{\Delta Ct})$. Additional details for calculating qPCR fold enrichments and testing the PE are available[47]. Although it is dependent on user requirements and probe design, we aim for a minimum tenfold enrichment value, and we routinely achieve greater than 50-fold enrichment.

? TROUBLESHOOTING

Library sequencing (Illumina HiSeq 2000) ● **TIMING ~11 d**

63| Having established successful enrichment and that libraries pass Agilent Bioanalyzer quality control (Step 60), subject libraries to sequencing by using the Illumina platform. Please refer to the 'Estimation of fold enrichment' section (INTRODUCTION) for guidance on sequencing type and depth.

CaptureSeq data analysis ● **TIMING 24 h**

64| Sequence reads are provided as a FASTQ-format file. To align sequenced reads from each sample to reference the human genome, type the following in the command line:

```
$ tophat2 --library-type fr-firststrand \
  -G gencode.v17.annotation.gtf -o tophat_output \
  hg19_index reads_1.fastq reads_2.fastq
```

▲ **CRITICAL STEP** Optional quality control of sequenced reads can be performed by using a range of tools, including FastQC and FASTX, before analysis to confirm successful sequencing. Refer to associated documentation for detail.

65| Assemble the aligned reads into full-length transcript models:

```
$ cufflinks -g gencode.v18.annotation.gtf -o cufflinks_output \
  tophat_output/accepted_hits.bam
```

The inclusion of a reference annotation ('-g/--GTF-guide') aids in transcript assembly.
▲ **CRITICAL STEP** Cufflinks has an option '-F/--min-isoform-fraction' that suppresses isoforms expressed below this abundance relative to dominant expressed isoform. To identify novel isoforms, we recommend lowering the value of this option (0.1 by default) so that weakly expressed isoforms that may be of interest are not omitted.
▲ **CRITICAL STEP** Optional quality control of RNAseq reads and alignment can be performed by using RNA-SeQC[28]. Refer to the associated documentation for details.

66| Remove off-target transcripts that do not overlap the probe design:

```
$ intersectBed -u -a transcripts.gtf -b probed_regions.bed \
  >captured_transcripts.gtf
```

This step reports all transcripts that have exonic sequence overlapping of any of the regions targeted for capture.
▲ **CRITICAL STEP** The '–u' option returns the entire transcript entry that overlaps the probe, rather than only the region that directly overlaps the probe design. This permits the user to identify all exons that are spliced into the mature mRNA (of which only part may be targeted) and thereby identify novel exons and isoforms.

67| If you are visualizing transcript assemblies in the UCSC Genome browser (http://genome.ucsc.edu/FAQ/FAQcustom.html), transcripts corresponding to ERCC RNA spike-ins must first be removed.

```
$ grep -v ERCC capture_transcripts.gtf
  >captured_transcripts_noERCC.gtf
```

Optionally, the number of reads overlapping the captured genes can be determined for estimating the fold enrichment achieved (as described in **Box 1**). By comparing the number of overlapping reads after capture to matched RNAseq, the fold enrichment and target specificity can be ascertained.

? TROUBLESHOOTING

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

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
6	Poor rRNA removal	Progression through time-sensitive steps is too slow	Repeat rRNA depletion. However, yield may be very low and more sample may be needed
		Temperature of the water bath or heat block may be incorrect	Validate temperatures with an external, calibrated thermometer and repeat rRNA depletion
16	Low library yield	Low RNA input into library preparation	Increase the number of precapture LMPCR cycles, plan to multiplex libraries or repeat library preparation with increased RNA input
		Degraded RNA	Confirm RNA quality using an Agilent Bioanalyzer. If degraded RNA has been detected, the fragmentation time can be decreased to improve library yield and insert size
		Errors in library preparation method; reagents degraded	Consult the TruSeq sample preparation best practices and troubleshooting guide at http://support.illumina.com/ for more details
	cDNA libraries are the incorrect size	RNA was fragmented for an incorrect length of time	Capture may still work, although different-sized libraries may capture with different kinetics and efficiencies. You may also need to adjust any size-based purification steps
21	Low precapture LMPCR yield	Too few cycles of LMPCR were performed	Make more libraries and combine LMPCRs or multiplex more samples for capture. In the future, confirm that the test library (steps 15 and 16) yields are correctly predicting precapture LMPCR (steps 17–21) yields by initially performing precapture LMPCR on a few select samples
45	Low volume of sample after hybridization	Evaporation during hybridization	Identify equipment (plates, tubes, thermocycler, etc.) that causes less than 2–3 µl of evaporation during hybridization. Then repeat the experimental capture hybridization with improved equipment
60	Low postcapture LMPCR yield	LMPCR was performed poorly	Repeat LMPCR with 5 µl of resuspended beads
		Capture failed	Repeat the capture
62	Low qPCR enrichments	Capture design targets a greater-than-optimal percentage of the transcriptome	Re-design the capture and repeat experiment using existing libraries (if available)
		qPCR primers are targeting a highly expressed gene	Design new qPCR primers against moderately or weakly expressed captured transcripts
		Evaporation prevented DNA hybridization to capture probes	Repeat the capture hybridization once a method to prevent excess evaporation has been confirmed
		Temperature dropped during binding and/or wash steps, allowing nonspecific DNA binding to the Dynabeads	Practice performing steps quickly with dummy reagents. Validate the temperature of the water bath or heat block with an external, calibrated thermometer. If you need to move tubes between equipment, put them in a 47 °C block
		HE oligos are degraded	Use a new aliquot in future hybridizations
	Noncaptured qPCR control shows little depletion after capture	Temperature dropped during binding and/or wash steps, allowing non-specific DNA binding to the Dynabeads	See potential solutions above. However, if captured amplicons show expected enrichment, the sample may still be suitable for sequencing
		Target amplicon has sequence similarity to captured regions	Test other amplicons

(continued)

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
62	Roche/NimbleGen control amplicons show much higher enrichment than design-specific amplicons	This appears to be a consistent result for different designs and samples and is not a concern	Ensure that the design-specific amplicons also demonstrate acceptable enrichment before proceeding to sequencing
	Design-specific amplicons show widely variable enrichment	Wide variability between design-specific amplicons could have many causes. If all show acceptable enrichment, proceed to sequencing	After sequencing, investigate ERCC controls to ensure that they show a good correlation between known and measured abundance. Performing a control DNA capture experiment may help to decrease the variability of enrichment measured by sequencing

● TIMING

Steps 1–6, preparation of DNA-free, rRNA-depleted sample RNA: ~1 d (if no DNA contamination detected)
Steps 7–16, preparation of sequencing libraries with the Illumina TruSeq stranded mRNA kit: ~2 d
Steps 17–21, precapture LMPCR: ~3 h
Steps 22–36, hybridization of the sample to SeqCap EZ library: ~2 h, plus 3 d of incubation
Steps 37–55, binding captured DNA to Dynabeads and washing to remove nontarget DNA: ~2.5 h, including incubations
Steps 56–60, postcapture LMPCR: ~3 h
Steps 61 and 62, performing qPCR for capture enrichment: ~3 h, including qPCR run time
Step 63, sequence libraries using Illumina HiSeq 2000: ~11 d
Steps 64–67, CaptureSeq data analysis: ~24 h (Timing is highly dependent on the computational resources available, and the depth of sequencing performed. However, we estimate that data analysis will take ~36 h for read alignment and transcript assembly on a machine with eight processing cores and at least 8 GB of RAM.)

ANTICIPATED RESULTS

The example results discussed below are derived from the capture of lncRNAs from human K562 cells. User results may vary if different versions of software (we used Bowtie 2.1.0, SAMtools 1.18, TopHat 2.0.9 and Cufflinks 2.1.1), reference genomes or gene annotations are used.

Estimated fold enrichment

From **Box 1**, we estimate a 1.9% overlap between read alignments from the example human K562 cell total RNA-sequencing data set downloaded from ENCODE[45] and the probe design containing 6,020 lincRNAs annotated from GENCODE annotations. This corresponds to an estimated maximum 52-fold (1/0.019) enrichment, should we target these transcripts in the K562 cell line. In practice, we generally observe less than this estimated enrichment owing to capture of both novel exons and off-target transcripts.

Read alignment

In Step 64, we observe ~14.9 million reads (74.3%) that align at least once to the human genome. This is 10–15% lower than the number routinely returned by conventional RNAseq and is due, in part, to the larger number of reads that span novel splice junctions that are not present within the reference gene annotation or identified by TopHat2, and that therefore do not align to the genome. More expansive reference gene annotations or *de novo* splice junction identification (by deeper library sequencing or more relaxed assembly parameters) can reduce this difference.

Transcripts assembled

From the read alignments, we are then able to assemble the transcript structures of all lncRNAs that have been captured with the design (not just the GENCODE lincRNA subset specified in **Box 1**). Performing assembly (Steps 65–67) should return ~23,825 transcripts that overlap the probe design, of which 6,359 represent novel isoforms. 20.8 million (64%) alignments (a single read can have more than one alignment by using the TopHat2 parameters specified) correspond to these assembled transcripts models (this includes targeted gene models and novel isoforms). However, the proportion of alignments to 'captured' transcripts from probed regions can vary, and it is dependent on the stringency of capture reaction

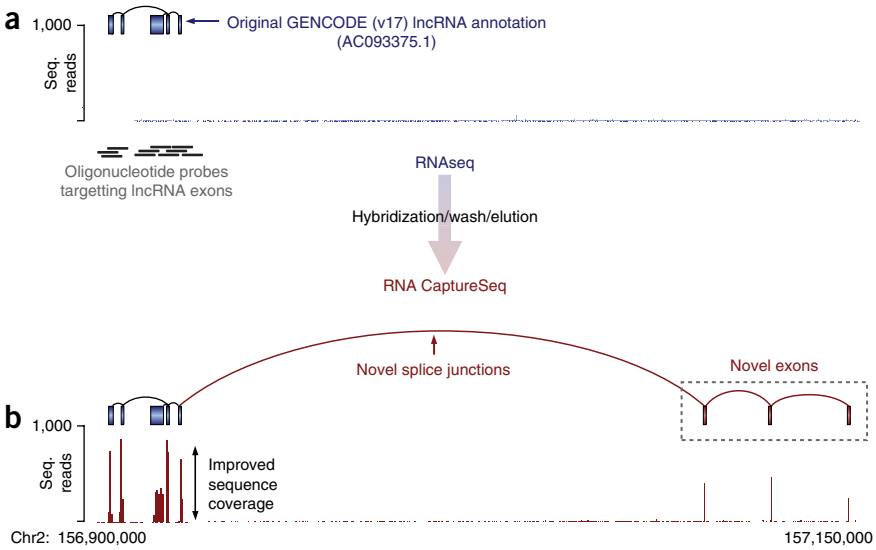

Figure 7 | Example of targeted RNAseq expanding the annotation of an lncRNA. (**a**) RNAseq achieves low coverage (blue histogram) of an lncRNA. (**b**) The previously annotated exons of the lncRNA (from GENCODE[48] and Cabili *et al.*[44]) are targeted by probes. Performing CaptureSeq provides improved sequence coverage (lower red histogram) that supports the identification of novel exons and splice junctions.

achieved and the fraction of transcription targeted. This compares favorably to conventional RNAseq, in which 8.2% of alignments correspond to captured lncRNAs transcripts. A remaining ~11.4 million alignments from the CaptureSeq experiment fall outside these transcript assemblies and represent off-target transcripts, multiple-aligning reads or reads derived from novel captured lncRNAs that have not been successfully assembled into full-length transcript models.

An example of a captured lncRNA is provided in **Figure 7**. After CaptureSeq, there is a large enrichment of reads aligning to the lncRNA with additional novel isoforms and exons assembled relative to the initial reference gene annotation. Novel terminal exons that are only supported by a single splice junction should be treated carefully, and we recommend that a random subset of novel isoforms and selected isoforms of interest be independently validated by reverse-transcription (RT)-PCR.

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

ACKNOWLEDGMENTS We thank the following funding sources: the Australian National Health and Medical Research Council (Australia Fellowship 631668; to J.S.M., T.R.M. and M.B.C.) and the Queensland State Government (National and International Research Alliance Program; to L.K.N.). We also thank the Institute for Molecular Bioscience core sequencing facility; we thank P. Danoy, J. Jeddeloh (Roche/NimbleGen) and T. Bruxner (Queensland Centre for Medical Genomics) for technical advice and assistance with capture sequencing; and we thank R. Bannen (Roche/NimbleGen) for helping with the design of capture arrays.

AUTHOR CONTRIBUTIONS T.R.M. and M.E.D. jointly conceived the CaptureSeq strategy. J.C. and M.B.C. designed, optimized and performed all stages of the protocol. T.R.M. and M.B.C. performed the analysis. M.E.B. and D.J.G. contributed to protocol development and optimization. T.R.M., J.C., M.B.C., M.E.B., L.K.N., R.J.T., M.E.D. and J.S.M. prepared the manuscript. L.K.N., R.J.T. and J.S.M. provided funding support.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

1. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods* **5**, 621–628 (2008).
2. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* **5**, 613–619 (2008).
3. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
4. Martin, J.A. & Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **12**, 671–682 (2011).
5. Ozsolak, F. & Milos, P.M. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* **12**, 87–98 (2011).
6. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
7. Mercer, T.R. *et al.* Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **30**, 99–104 (2012).
8. Levin, J.Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* **10**, R115 (2009).
9. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).
10. Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
11. Clark, M.J. *et al.* Performance comparison of exome DNA sequencing technologies. *Nat. Biotechnol.* **29**, 908–914 (2011).
12. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
13. Turner, E.H., Ng, S.B., Nickerson, D.A. & Shendure, J. Methods for genomic partitioning. *Annu. Rev. Genomics Hum. Genet.* **10**, 263–284 (2009).
14. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).
15. Craig, D.W. *et al.* Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5**, 887–893 (2008).
16. Howald, C. *et al.* Combining RT-PCR–seq and RNA-seq to catalog all genic elements encoded in the human genome. *Genome Res.* **22**, 1698–1710 (2012).
17. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
18. Dahl, F., Gullberg, M., Stenberg, J., Landegren, U. & Nilsson, M. Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. *Nucleic Acids Res.* **33**, e71 (2005).
19. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8**, 1765–1786 (2013).

20. Kreil, D.P., Russell, R.R. & Russell, S. Microarray oligonucleotide probes. *Methods Enzymol.* **410**, 73–98 (2006).
21. Harrow, J. *et al.* GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* **7** (suppl. 1), S4: 1–9 (2006).
22. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
23. Baillie, J.K. *et al.* Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* **479**, 534–537 (2011).
24. ERC Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* **6**, 150 (2005).
25. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
26. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
27. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
28. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
29. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
30. Goecks, J., Nekrutenko, A. & Taylor, J. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* **11**, R86 (2010).
31. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
32. Derveaux, S., Vandesompele, J. & Hellemans, J. How to do successful gene expression analysis using real-time PCR. *Methods* **50**, 227–230 (2010).
33. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
34. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
35. Au, K.F., Jiang, H., Lin, L., Xing, Y. & Wong, W.H. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.* **38**, 4570–4578 (2010).
36. Wu, T.D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26**, 873–881 (2010).
37. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31**, 46–53 (2013).
38. Mezlini, A.M. *et al.* iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.* **23**, 519–529 (2013).
39. Li, J.J., Jiang, C.R., Brown, J.B., Huang, H. & Bickel, P.J. Sparse linear modeling of next-generation mRNA sequencing (RNA-seq) data for isoform discovery and abundance estimation. *Proc. Natl. Acad. Sci. USA* **108**, 19867–19872 (2011).
40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
41. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
42. Kuhn, R.M. *et al.* The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.* **37**, D755–761 (2009).
43. Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E. & Mattick, J.S. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res.* **39**, D146–D151 (2011).
44. Cabili, M.N. *et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **25**, 1915–1927 (2011).
45. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
46. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
47. Citri, A., Pang, Z.P., Sudhof, T.C., Wernig, M. & Malenka, R.C. Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nat. Protoc.* **7**, 118–127 (2012).
48. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).