ORF Capture-Seq as a versatile method for targeted identification of full-length isoforms

Gloria M. Sheynkman1,2,3, Katharine S. Tuttle1,2,3,4,5,6, Florent Laval7,8, Elizabeth Tseng9, Jason G. Underwood9, Liang Yu10, Da Dong10, Melissa L. Smith5,6, Robert Sebra5,6, Luc Willems7,8, Tong Hao1,2,3, Michael A. Calderwood1,2,3, David E. Hill1,2,3 & Marc Vidal1,2

Most human protein-coding genes are expressed as multiple isoforms, which greatly expands the functional repertoire of the encoded proteome. While at least one reliable open reading frame (ORF) model has been assigned for every coding gene, the majority of alternative isoforms remains uncharacterized due to (i) vast differences of overall levels between different isoforms expressed from common genes, and (ii) the difficulty of obtaining full-length transcript sequences. Here, we present ORF Capture-Seq (OCS), a flexible method that addresses both challenges for targeted full-length isoform sequencing applications using collections of cloned ORFs as probes. As a proof-of-concept, we show that an OCS pipeline focused on genes coding for transcription factors increases isoform detection by an order of magnitude when compared to unenriched samples. In short, OCS enables rapid discovery of isoforms from custom-selected genes and will accelerate mapping of the human transcriptome.
Mechanisms that enable production of multiple isoforms from a single gene—such as alternative transcriptional start sites, splicing, and polyadenylation—contribute to expanding the functional capacity of the encoded proteome. The full extent of this capacity is unknown, as we are currently unable to generate an accurate and comprehensive catalog of the human transcriptome. Although advances in high-throughput sequencing have enabled mapping of local elements (e.g., individual splice junctions), how these elements combine to form full-length isoforms is largely unknown. Short-read RNA-seq data from currently popular platforms (<250 bp, Illumina) fail to resolve such sequences. Consequently, the majority of annotated isoform models remain as predictions derived from partial transcripts, particularly for context-specific, disease-specific, or low-abundance isoforms.

Long-read sequencing platforms—PacBio, Oxford Nanopore, and those based on adaptations to next-generation sequencing—that produce synthetic long reads such as 10X and Moleculo sequencing—can return unambiguous, full-length isoform sequences that fully resolve transcriptome complexity. However, in comparison to short-read sequencing platforms, these methods suffer from lower sampling sensitivity and can miss low (<10 copies/cell) to moderate (10–50 copies/cell) abundance transcripts. A number of transcripts at these abundance levels are responsible for producing disease-associated or important regulatory proteins (e.g., transcription factors and kinases). The sensitivity problem is exacerbated by the wide dynamic range of the human transcriptome across at least six orders of magnitude, causing an inordinate amount of sequencing effort to be used for detecting the most abundant isoforms. Therefore, most transcripts of lower abundance are not sequenced satisfactorily due to this biased sampling.

An established solution to increase detectability of isoforms is targeted sequencing, which involves enriching for transcripts of desired genes. DNA or RNA hybridization-based enrichment followed by high-throughput sequencing is a particularly efficient, robust, and cost-effective solution. These approaches were initially developed for targeted sequencing of protein-coding regions of genomic DNA (i.e., whole exome sequencing) and RNA fragments from short-read RNA-seq experiments (e.g., CaptureSeq). Such approaches have been adapted for targeted sequencing of long genomic fragments (>2 kb) or full-length cDNA molecules. In two notable studies described recently, complex pools of biotinylated oligos are used to enrich for full-length cDNA corresponding to thousands of protein-coding and non-coding RNA targets. The enriched material is subjected to long-read sequencing, leading to considerable gains in full-length isoform detection and insights about the nature of transcriptomic complexity.

The success of these targeted full-length sequencing methods hints at the potential of using this approach in a more general framework. Previous studies employ a single panel of probes with many targets. This schema worked well for exome sequencing, in which a single probe set designed against all protein-coding exons yields high coverage of all DNA targets, each of which are present at identical concentration (2 copies/cell). However, such high coverage is challenging to attain from transcriptome sequencing, because transcriptomes are highly heterogeneous, with a wide dynamic range and variation in composition (i.e., set of genes expressed), depending on cellular or disease contexts. Therefore, a single probe set enables increased sequence sampling of isoforms from genes of interest, but the expression patterns within the genes of interest will still be skewed, reducing coverage. It is thus essential to devise a flexible strategy to generate with ease multiple, distinct probe sets that match the particular transcriptome context, specifically, to enable facile detection of isoforms from any set of genes from any set of samples.

Here we present ORF Capture-Seq (OCS), a generalizable method for direct synthesis of biotinylated capture oligos from existing or newly designed ORF clones followed by targeted enrichment and sequencing of full-length cDNAs to discover new isoforms. The unique combination of low cost, time, ease, and versatility (any pool of ORFs or clones, up to thousands at a time) of the method offers the experimental flexibility needed to rapidly characterize any desired subset of the transcriptome. Using reagents and instruments available in most molecular biology laboratories, a user can synthesize probes from one or a set of amplicons or clones in less than 24 h. We envision this method will be of broad utility in many applications, from single-gene studies to system-scale applications seeking to characterize whole transcriptomes. Here, we compare OCS probes against a commercial standard, benchmark the method using spike-in standards, and apply it towards characterization of novel isoforms of ~800 human transcription factors (TFs).

Results

OCS method for flexible targeted sequencing. The OCS pipeline begins with flexible and straightforward synthesis of biotinylated capture probes (Fig. 1a). PCR is performed on any number of pooled templates (e.g., plasmids, amplicons) using universal primers in the presence of biotin-dUTP. The resulting pool of biotinylated PCR products, with biotin-dUMP incorporated throughout, are randomly sheared into overlapping DNA fragments of ~150 bp. This generates a set of overlapping fragments from each PCR amplicon. After purification and removal of PCR primers and unincorporated nucleotides, the resulting OCS probe set is used for hybridization-based capture of target nucleotide sequences.

We demonstrate the application of OCS for enrichment and sequencing of full-length transcripts from protein-coding genes, though the method can also be applied to non-coding RNAs as well (Fig. 1a). Probes are derived from one or more ORF(s) or PCR amplicon(s). Though each ORF represents just one isoform of a gene, the corresponding probes are expected to capture all isoforms, due to the high sequence overlap between isoforms of the same gene; probes need to only target a portion of a full-length cDNA for enrichment. We capitalized on the availability of our human ORFeome collection, a resource of freely available Gateway Entry clone ORFs for ~17.5 K of the ~20 K protein-coding genes in human, creating from this resource customized pools of ORF clones to use as templates for biotin-labeling PCR. All clones share a common vector backbone, enabling production of any amplicon from universal primers. Tens to hundreds to thousands of ORF clones may be pooled and processed together so that complex and customized probe sets can be generated with relative ease.

We emphasize that though the demonstration of OCS in this paper involves a large-scale application where templates are derived from a comprehensive ORFeome collection, our OCS method is highly applicable to smaller-scale studies using a small number of clones or amplicons.

OCS probes perform comparably to a commercial standard. We first established that OCS probes are comparable to commercially synthesized biotinylated probes in terms of enrichment efficiency.

To benchmark, we selected three low abundance human TF genes (ARNTL, STAT1, and ZNF268) expressed in brain. The enrichment of these TFs was compared against two high abundance housekeeping genes (MBP and HSP90A1) to serve
as a measure of off-target binding (Fig. 1b). For each of the three TFs, OCS and commercial probes were synthesized (Supplementary Fig. 1a). Probes were not generated for the two housekeeping genes. OCS probes were sequenced on an Illumina MiSeq, confirming an even distribution of probe coverage (estimated ~150X tiling density) and high purity (Supplementary Fig. 1b). Commercially available probes were synthesized as 5’ biotinylated DNA vectors derived from ORF inserts in which each ORF is ranked by ~100 bp concentration were not targeted and served as controls. ERCC vector sequences can cause non-specific binding. To investigate this, we compared background binding profiles derived from OCS versus IDT capture experiments. The profiles are displayed as enrichment of each transcript as a function of initial abundance, because higher abundance transcripts have been observed to non-specifically bind to the beads (i.e., streptavidin beads which bind to and enrich probe-target complexes) to a greater extent than low abundance transcripts (Supplementary Fig. 1h, see No Probe Control). We found no systematic bias using either probe type.

Analytical benchmarking of OCS using spike-in standards. Next, we benchmarked the analytical performance of OCS by employing External RNA Controls Consortium (ERCC) standards, which are 92 synthetic ORFs of concentrations spanning 10-orders of magnitude. To assess specificity and reproducibility, we measured enrichment of a subset of ERCC ORFs in human reference RNA (Fig. 2a). OCS probes were synthesized for the 64 ERCC ORFs of lowest concentration (ERCC64, Supplementary Fig. 2a, Supplementary Table 2). The remaining 28 ORFs of highest concentration were not targeted and served as controls. ERCC
RNA standards were spiked into Universal Human Reference RNA (UHRR) at high (1:10, i.e., 10X dilution of ERCC standard), medium (1:80), and low (1:5120) concentrations, in technical duplicates (Supplementary Fig. 2b). Input cDNA and capturant cDNA were sequenced on an Illumina MiSeq and abundance values estimated using the Kallisto software package\(^{31}\), in which a subsampling of 100 K reads were subjected to analysis to allow for comparison of ORF detection at comparable sequencing depth.

We show that OCS probes successfully enriched all 64 targeted ERCCs at uniformly elevated levels (Fig. 2b). We computed overall on-target rates by mapping reads to a reference transcriptome (GENCODE) and computing the total abundance, in transcripts per million (TPM), arising from ORFs targeted for enrichment. Fold enrichment was calculated by dividing on-target percentages detected in the capturant by those found in the unenriched input. For the high- and medium-spike capture, on-target rates were above 80% with fold enrichments of 195- and 763-fold (Fig. 2c). For the low-spike capture, on-target rates were lower (~6%) but enrichment was high (730-fold). We observed that the relative abundances of targeted ERCCs remained linear post-capture, indicating potential for quantification as long as probes are in excess concentration compared to the target cDNA (Supplementary Fig. 2c, see Clark et al. for more details\(^{32}\)). Furthermore, though the correlation passed statistical significance, the extent of enrichment was not markedly affected by properties such as starting concentration, GC content, ORF length, and probe representation (Supplementary Fig. 2d). Importantly, we did not detect a significant drop in enrichment efficiency for longer ORFs (up to ~2 kb) nor ORFs with lower probe representation, as may be expected. Technical reproducibility was also high, with Pearson’s $R^2$ above 0.97 for all technical replicates (Supplementary Fig. 2e).

**OCS enriches a family of isoforms using only one isoform.** The use of OCS for isoform discovery relies on the assumption that probes derived from a single isoform can efficiently enrich for all isoforms of a given gene. Indeed, there is typically sufficient overlap between any given isoform sequence and all other isoforms of that gene. However, to address the concern that there could be lower capture efficiency due to low sequence overlap, we measured the relationship between overlap and capture efficiency.

For this purpose, we used Spike-In RNA Variant (SIRV) standards (Lexogen), consisting of 69 synthetic isoforms with highly complex splicing patterns from seven artificial genes (Fig. 2d)\(^{33}\). We synthesized OCS probes from one representative SIRV isoform per gene (SIRV7, Supplementary Fig. 2f, g, Supplementary Table 3) and used this probe set to enrich for all SIRV isoforms that were spiked into UHRR (Supplementary Fig. 2h). We found no appreciable difference in capture efficiency when sequence overlap ranged between 45 and 2500 nt (Fig. 2e). Only in one extreme case, where overlap was only 35 nt, did
capture efficiency sharply decline to the level of negative controls. These results are consistent with our observation that captures employing a single probe have good enrichment. For example, use of a single probe from the ARNTL IDT probes enabled large enrichments of ARNTL isoforms (0% in input cDNA, 8 and 10% in the capturant).

Based on the SIRV results, we calculated how well the OCS approach could cover the human transcriptome. Using GENCODE-annotated transcripts, we calculated the degree of overlap between a single representative isoform and all other annotated isoforms of the same gene (Supplementary Fig. 2i, see "Methods" section, Isoform Overlap Estimation). The overlap between pairs of isoforms (one representative versus all isoforms of a gene) was calculated by taking the intersection of genomic ranges. If the representative isoform is set to the principle isoform in the APPRIS database, as defined by GENCODE34 (version 29), we estimate that 99.7% of all isoforms are potentially captured by OCS probes designed against the principal isoform (overlap of 50 bp or more, based on results from the previous SIRV). If the representative isoform is set to a randomly chosen isoform, we estimate that 99.3% of all isoforms are covered. The slight decrease in coverage can be explained by the fact that that the APPRIS principle isoform tends to be longer in length than randomly chosen isoforms, and longer isoforms would more likely overlap other isoforms of that gene.

**Applying OCS to characterize human TF isoforms.** Alternative transcriptional start sites, splicing, and polyadenylation can modulate the activity of TFs by altering sequences corresponding to DNA binding, co-factor binding, and other properties such as availability of phosphorylation sites35,36. Despite being heavily studied, many TF isoforms remain uncharacterized due to low abundance (<10 copies/cell), complex splicing patterns, or expression in cell-, tissue-, or disease-specific contexts35,37. Here, we applied OCS to characterize alternative isoforms of human TFs.

First, we sought to explore the relationship between the number of genes targeted and sensitivity. We synthesized OCS probe sets for 2, 12, 88, and 682 randomly chosen TF genes and applied it to human cDNA derived from cerebral cortex (Supplementary Fig. 3a, Supplementary Tables 4–6, Methods). To unambiguously distinguish different TF isoforms, we subjected the enriched cDNA to long-read sequencing in addition to short read RNA-seq.

We were able to maintain high capture efficiency even at increased probe set complexity (Fig. 3a, b). However, the limiting sequencing depth reduced the number of genes, as well as isoforms per gene, that were detected (Fig. 3c, d). This was further shown by saturation–discovery curve analysis in which we plotted the number of unique genes and isoforms detectable at different sequencing depths. At 20,000 full-length reads, we reached saturation for the TF2 and TF12 probe sets, but were still discovering new isoforms for the TF88 and TF682 (Supplementary Fig. 3b, Methods). Our results show that additional sequencing is required to reach saturation using these more complex probe sets. We note that, as of this writing, emerging platforms (e.g., PacBio Sequel II) provides much higher throughput than the older platforms (e.g., PacBio RS II, Sequel) used in this study. With increased throughput, there is a potential to saturate the samples at even the isoform-level and also increase the level of sample-multiplexing, further reducing the cost for full-length transcript studies.

**Discovery of isoforms in tissue using transcript sequencing.** We applied OCS towards discovering novel TF isoforms in a diverse set of human tissues. We created an OCS probe set consisting of 763 randomly chosen TFs to enrich from a pool of barcoded cDNA libraries stemming from 7 distinct human tissues -- adult brain, fetal brain, heart, liver, pancreas, placenta, and testes (Supplementary Fig. 4a, Supplementary Tables 7 and 8, Methods). We sequenced the unenriched and enriched cDNA on the PacBio Sequel platform, resulting in 118,872 and 476,589 full-length reads, respectively. Raw data was processed using the Iso-Seq v3 bioinformatic pipeline38 followed by SQANTI for isoform annotations39. The on-target rate for the capturant sample was ~60%, with most of the top ranked genes being the target TFs (Fig. 4a).

To compare enrichment efficiency, ~100 K full-length reads were sampled from both the input and capturant datasets and sequencing statistics were calculated. We found that the number of genes, isoforms, and full-length reads increased 2-, 7-, and 43-fold after enrichment, respectively, emphasizing the need for enrichment to fully sample the isoform space (Fig. 4b). In theory, if sequencing depth is increased to appropriate levels, any low abundance isoform should be detectable from unenriched cDNA. However, in this case, approximately 40 times more coverage using bulk long-read sequencing would have been needed to attain the coverage achieved using OCS, showing the benefit of target enrichment strategies.

We analyzed the extent to which we recovered full-length isoforms that are annotated in GENCODE. Overall, the recovery of GENCODE genes was 74%, increasing to 86% or higher when considering only genes which were well-represented in the probe set (1 TPM or higher), or genes well-expressed (10 TPM or higher) in the 7 tissues (Fig. 4c, Supplementary Table 9). The recovery of GENCODE isoforms was 37%, also increasing when accounting for probe representation or tissue expression. Note, however, that given the limited tissues analyzed in this study, 100% GENCODE isoform recovery is not expected.

In GENCODE annotations, every isoform is tagged with a transcript support level (TSL) 1–5, denoting the extent of experimental information underlying each isoform model. Isoforms with a TSL of 1 contain high-quality experimental, full-length mRNA sequence support. Isoforms with a TSL of 5 are computational predictions with no experimental support. The full-length sequencing data provided confirmation of 398 GENCODE isoforms of TSL 2–5, which includes 85 computationally predicted isoforms (TSL 5) (Supplementary Fig. 4b, Supplementary Table 10). Thus, OCS is an invaluable tool for confirming isoform models in gene annotations.

We found that OCS-based enrichment not only enables significant increases in detection of annotated isoforms, but discovery of novel isoforms as well. To increase the confidence of novel isoforms found in the TF763 dataset, we employed an orthogonal sequencing approach to validate any novel junctions. We subjected the original RNA to dication-catalyzed RNA fragmentation and random hexamer priming to generate cDNA fragments which were subjected to Illumina sequencing. This random hexamer approach for short-read sequencing, which is not subject to artifacts that come from full-length cDNA amplification, provides orthogonal support for individual junctions found in the long-read data39. We required each novel junction found in the PacBio isoforms to be supported by at least three short reads, resulting in a population of novel isoforms with quality features (e.g., non-canonical junction rate), intrinsic sequencing properties (e.g., number of predicted RT artifacts), and functional genomics evidence (e.g., overlap of 5′ end with CAGE peaks) that are indistinguishable from the high-quality, known isoforms that match GENCODE annotations (Supplementary Fig. 4c-k). All subsequent analyses involved this orthogonally-validated set of novel isoforms (Supplementary Table 11).
Fig. 3 Multiplexing parameters for enrichment of human transcription factors. a Rank abundance bar plots for unenriched (input) and enriched (capturant) cDNA. Data is shown for (i) input brain cDNA, and (ii) the series of capturants prepared using probe sets with increasing number of TF genes. The 5′ splice site corresponds to the splice donor and the 3′ splice site corresponds to the splice acceptor. SS, splice site.

b Absolute number of targeted genes (dark purple) and isoforms (light purple) detected from each capture reaction. Match categories are defined by the isoform annotation tool SQANTI. c Fraction of all GENCODE genes and transcripts detected in the capturant. A gene is considered detected if at least one full-length read is detected for that locus. Genes for which there was evidence of expression in the tissues interrogated (10 TPM or higher).

d Fraction of novel splice sites, junctions, and full-length isoforms in the TF enrichment experiment. Unique splice sites and junctions are only counted once. The 5′ splice site corresponds to the splice donor and the 3′ splice site corresponds to the splice acceptor. SS, splice site. e Proportions of known and novel isoforms. Known isoforms are further divided by completeness. Novel isoforms are further divided by whether all splice sites are found in GENCODE (novel in catalog, NIC) or if the isoform contains a novel splice site (novel not in catalog, NNC). Match categories are defined by the isoform annotation tool SQANTI.

f Example of isoforms from the gene MITF identified from the TF763 capture experiment. Source data are available in the Source Data file.

Fig. 4 Full-length transcription factor isoforms across diverse human tissues. a Rank abundance bar plot for cDNA enriched for human TFs. Only the top 50 most abundantly expressed genes are shown. Each bar corresponds to a single gene, colored by whether that gene is targeted (purple) or not targeted (gray). Abundances as shown on the y-axis are computed by dividing the number of full-length reads mapped to the gene by the total number of full-length reads. On-target rates for input and capturant are shown. b Gains in coverage of target genes upon enrichment. Increases in number of genes, isoforms, and full-length reads are shown. Data used to generate these numbers used an equal number of full-length raw reads that were subsampled from the unenriched (input) or enriched (capturant) cDNA. c Fraction of all GENCODE genes and transcripts detected in the capturant. A gene is considered detected if at least one full-length read is detected for that locus. isoforms are considered detected if the full set of junctions are identical between the capturant and the unenriched control. d Fraction of genes detected from each capture reaction. e Genes ranked by abundance (unit = rank of the gene). On-target rates, as calculated for the entire sample, are displayed on the upper right-hand position of the plots. f Absolute number of targeted genes (dark purple) and isoforms (light purple) detected from each capture reaction.
One potential concern with the use of high-density OCS probes is the formation of chimeric products during subsequent rounds of PCR, which may lead to the formation of artifactual isoforms. Currently, it is not possible to distinguish de facto novel isoforms from chimeric products formed by stitch-like PCR of partial fragments. To experimentally assess the percentage of novel isoforms that are non-chimeric, we employed a dual barcoding strategy used previously to assess this type of artifact. The strategy includes a CDNA preparation step in which each cDNA molecule is synthesized such that known matching barcodes are appended at the 3' and 5' ends prior to whole-cDNA PCR amplification. Using this library, cases in which the barcodes are swapped occur only through formation of chimeric products and can be directly detected in the long read sequencing data. We subjected a dual barcoded library generated from human brain cDNA to enrichment with either IDT probes or OCS probes directed against the same three transcription factor genes, as described in Fig. 1. The libraries were sequenced on a PacBio Sequel II system. When considering all isoforms, the rate of non-chimeric isoforms was 99% for IDT and 98% for OCS. When considering only novel isoforms, 98% of the IDT isoforms and 94% of the OCS isoforms were non-chimeric. Therefore, compared to IDT, OCS captures have a slightly elevated rate of chimeric isoforms. Overall, however, both technologies have very low rates of chimera formation and users may find that the cost and flexibility benefits of the OCS method are advantageous.

We classified the isoform set against GENCODE v29. Approximately 4% and 10% of all distinct splice sites and junctions were novel, respectively, but a much higher fraction of isoforms, 37%, were novel (Fig. 4d). This can be explained by the fact that a single local event (e.g., one novel splice site) leads to an entirely distinct transcript, in terms of the full-length sequence. Overall, the total number of isoforms detected dramatically increased upon enrichment and at the same time the relative fraction of novel isoforms increased in proportion (Fig. 4e, Supplementary Fig. 4f). A substantial proportion of the reads arose from novel isoforms, though the number of full-length reads counted was slightly lower for novel isoforms as compared to known isoforms (Supplementary Fig. 4m). Overall, 1,528 distinct novel isoform sequences from 306 TF genes were detected. Among the TF genes targeted, at least one novel isoform was found for 39% (306/782). A majority of these isoforms (68%) arise from unique combinations of annotated junctions or splice sites, representing a change in the connectivity of known transcript elements. Most of the remaining isoforms (26%) contained novel donor or acceptor splice sites, effectively creating novel exons. A small fraction (5%) of isoforms contained intron retention events. We show an example for the gene MITF, which display isoforms which perfectly match a GENCODE transcript, has novel combinations of junctions, and has novel splice sites (Fig. 4f).

Discussion

Eukaryotic transcriptomes remain unresolved at full-length resolution, and the extent of transcript diversity is unknown. Recently, targeted full-length sequencing methods, in which pools of biotinylated oligos are used to enrich for full-length cDNAs for sequencing, have characterized focused subsets of the transcriptome with great depth and accuracy. Here, we establish that ORF Capture-Seq is a versatile method to synthesize probes that can be used for comprehensive enrichment of cDNAs for targeted full-length sequencing studies. Compared to traditional or microarray-based synthesis of biotinylated oligos, OCS stands as a complementary strategy for synthesis of simple to complex probe sets for desired genes. OCS enables direct characterization of isoform sequences without prior information about transcript boundaries or exons, unlike with RACE- or PCR-based sequencing. This facilitates discoveries of the full array of known and novel isoforms for other classes of proteins (e.g., kinases, G protein-coupled receptors) or targets involved in a biological pathway of interest or implicated in human genetic diseases. Applying OCS to detect TF isoforms in human tissues, we found a preponderance of novel isoforms, detecting over a thousand novel TF isoforms.

Several applications involving OCS can be envisioned. First, isoform expression information could guide high-throughput cloning efforts. Second, OCS can help define full-length isoform sequences from genes exhibiting differential splicing at the local level, using programs like Leafcutter. Third, it also provides opportunities for increased accuracy in isoform quantification workflows. Here, knowledge of isoforms expressed in a sample, as informed by long-read data, can serve as the scaffolds (i.e., gene models) upon which short reads rely to estimate isoform abundances. Finally, OCS probes could conceptually be applied at the single cell level, as pools of barcoded single-cell cDNA could be subjected to enrichment similarly to bulk cDNA preparations. Indeed, a method using commercial probes to enrich for T-cell-receptor and B-cell-receptor transcripts from single cells was recently described.

The ease of making probes using the OCS method opens doors for novel strategies in capture experiments. For example, a series of captures can be designed in an iterative manner, in which the initial capture returns the first batch of detected genes, and subsequent captures use probe sets that include only genes that failed to return isoforms in the first round. Alternatively, multiple gene panels may be created, stratified by endogenous abundance of genes (e.g., separation of probe sets for low and high abundance transcripts) or priority of disease genes (e.g., low/high confidence of association). Furthermore, the representation of genes within a probe set is customizable to the greatest extent. For example, the concentration of individual ORFs may be titrated based on desired factors, such as priority or endogenous abundance in the sample, so that the relative coverage of isoforms from different genes is normalized, allowing for greater sequencing coverage.

Some limitations of OCS remain. For example, since the probe synthesis step relies on PCR, one constraint of the method lies in the length of ORF clones that can be used as templates. Indeed, while probe synthesis and full-length sequencing for transcripts up to 4 kb is attainable, employing the method for long transcripts above 5–6 kb is a challenge. For longer transcripts, other mechanisms to generate biotinylated probes may be required (e.g., 5’ biotinylated directed or random primers, ligation or nick translation). Alternatively, we generated probes from individual segments of an ultra-long ORF clone (unpublished). Another limitation is some targeted transcript may be missed if there is insufficient overlap between the ORF used for probe synthesis and targeted transcripts (isoforms) of the same gene, although we estimate this will occur only very rarely based on annotated transcripts.

Another challenge is not just related to OCS but all long-read cDNA sequencing methods. Given the high number of novel isoforms routinely detected in long-read sequencing studies, an open question in the field is how to best assess the quality of these isoforms and to understand all sources of artifacts that could arise from the biochemical and analytical preparation of samples for sequencing. For example, the OCS protocol, as well as any targeted long-read method, requires PCR to amplify the enriched cDNA, and a valid concern is the extent of PCR artifact formation. A recent study employed a dual-barcoding approach to find
that ~1% of cDNAs sequenced were PCR chimeras (i.e., hybrid molecules from recombinational events during PCR)\(^\text{40}\). In our own study, we found a very low rate of false positive isoform identifications resulting from chimeric reads.

At a more general level, few studies have systematically investigated the sources and frequencies of artifacts. One of the most comprehensive assessments to date is work from Tardaguila and colleagues\(^\text{39}\), in which they evaluated intrinsic and sequence-related properties that contribute to isoform artifacts. They found that non-canonical or RT-template switching junctions underlie poor quality novel isoforms, but that experimental validation using an orthogonal approach effectively detects and removes these events. We employed a similar orthogonal validation in this study. Overall, until full understanding of artifacts is achieved, we recommend that any novel isoform must be viewed as a candidate isoform until it is validated by another assay or combined with strongly supporting functional data. Such evaluation was done for several TF isoform sequences in this study, which led to their acceptance as official transcript models in GENCODE (personal communication with Adam Frankish).

OCS stands as a complementary approach to previously described targeted methods, specifically for synthesizing probes, and delivers unique benefits in certain contexts. Previously described methods for targeted long-read sequencing employed probes that tile across target elements (e.g., genes, predicted lnRNAs) which were designed in silicone and synthesized by commercial vendors to generate biotinylated oligo pools. OCS is a viable alternative to generate probes in a flexible manner. For the many laboratories in which researchers have at hand a clone or the ability to generate a PCR amplicon representing a transcript from a gene of interest, probes can be made within 24 hours at nominal cost using widely available molecular biology reagents (Taq polymerase, NTPs, biotin-dUTP, etc.). For larger-scale applications, such as what was demonstrated in this work, some researchers may be interested in investing in a clone collection to capitalize on the ability to rapidly generate probe sets for a near-infinite combination of genes, such as those from a disease pathway. Regardless of the scale, a key benefit is that OCS probes have ultra-high tiling density and can be generated at near inexhaustive supply (e.g., 100+ enrichment reactions per PCR).

We recognize that commercial probes (e.g., IDT probes) could be more useful in certain circumstances, such as when clones are unavailable (e.g., novel lnRNAs) or when targeting predicted genes. Some researchers may prioritize the convenience of direct-to-order commercial probes, without additional labor required to make OCS probes.

In conclusion, OCS is a highly generalizable strategy to synthesize probes for use in full-length capture experiments. Though we demonstrated OCS as applied towards characterization of isoforms originating from protein-coding genes, it can be adapted for use in capture and characterization of different types of genetic and post-transcriptional variants, such as genetic variations, segmental duplications, or lnRNAs. For example, multiple isoforms of lnRNAs are routinely being characterized\(^\text{28}\). It is also possible that ORFs from one species could be used to enrich for isoforms from another species, given high sequence conservation of protein-coding regions (e.g., human ORFs to enrich for mouse isoforms). Overall, we envision this approach will be of broad utility for application within both basic research and the clinical and diagnostic fields\(^\text{31}\).

**Methods**

**ORF capture-Seq probe synthesis.** ORF amplicons corresponding to ARNTL, STAT1, ZNF668 (three TFs) were generated. Our human ORFeome contains one representative ORF, in the form of a Gateway entry clone in pDONR223\(^\text{\textregistered}\), for ~17.5 k of the ~20 k human protein-coding genes\(^\text{29}\). These ORF clones are available as bacterial (DH5α) culture glycerol stocks. Bacterial stocks corresponding to the three TFs were cherry-picked from the human ORFeome. Using ~1 μl of culture to template, the ORF inserts were PCR amplified with Platinum Taq DNA Polymerase High Fidelity (Invitrogen) using pDONR223 universal primers with the following sequences: FOR: CCGAATTCGACACGTTAAACCGC
REV: GTAAACGAGCTTGGAGGAC
PCR was performed for 35 cycles, each cycle consisting of denaturation at 94°C for 30 s, annealing at 57°C for 30 s, and extension at 72°C for 5 min. The final extension was for 15 min. PCR products were analyzed via agarose gel electrophoresis to confirm that amplicons were of the expected size. Products were purified using Agencourt AMPure XP beads (Beckman Coulter) and quantified using the Qubit dsDNA assay (Thermo Fisher Scientific).

A biotin-labeling PCR was done to generate amplicons which contain randomly incorporated biotin-dUTP. Both control and biotin-dUTP PCRs were performed for each TF. Using as template 1 ng of the ORF amplicon, biotin-spiked PCR was done using Taq polymerase (NEB). The dNTP mixture was modified so that a third of the dNTPs were substituted with Biotin-16-Aminomethyl-2'-dUTP (Thelink), referred to heretofore as biotin-dUTP. The program was run for 30 cycles, with each cycle consisting of denaturation at 95°C for 15 s, annealing at 57°C for 30 s, and extension at 68°C for 5 min. The final extension was for 10 min.

The biotin-labeled amplicons were randomly fragmented via sonication. Control and biotinylated amplicons were fragmented on an agarose gel to confirm successful PCR. Products were transferred to Covaris AFA Fibercramp Cap microTUBEs and fragmented on a Covaris E220 sonicator to size distribution of ~150 bp. The sonication method parameters are as follows: peak power of 175 W, duty cycle of 10%, 200 cycles per burst, and duration of 480 s. Fragmented DNA was verified using SPRISelect beads (Beckman Coulter) using a 1:50 ratio of sample to beads to remove high mass fragments above ~300 bp. Concentration of fragments were measured with the Qubit dsDNA assay (Thermo Fisher Scientific).

Equally fragmented samples (i.e., probe set) were mixed to generate a pool, or probe set. An equal weight mixture of the three TFs probes was prepared. The final concentration of the probe set was adjusted to 0.5 ng/μl.

For the following gene sets, OCS probes were synthesized using the same protocol as for the three TFs, with exceptions described below.

Using a protocol similar to that used for synthesis of the three TFs probeset, an OCS probe set was created which correspond to ORFs from the External RNA Controls Consortium, or ERCC. Differences in the protocol are described in the following section. The ERCC has compiled a collection of 92 synthetic ORF sequences, in the form of plasmids, from which RNA standards have been prepared by various vendors. We obtained the ERCC DNA Sequence Library for External RNA Controls (SRM 2374, NIST), a collection of all ERCC ORFs in the pT7T318 vector.

ERCC ORF inserts were PCR amplified with hot-start KOD polymerase (Invitrogen) using M13 primers, sequences below:

**M13_canon_FOR: GTAAAACGACGGCCAGT**

**M13_canon_REV: CAGGAAACAGCTATGAC**

PCR was performed for 18 cycles, with each cycle consisting of denaturation at 94°C for 30 s, annealing at 55°C for 30 s, and extension at 72°C for 5 min. The final extension was for 15 min.

An amplicon pool was prepared for 64 ERCC ORF sets (Supplementary Table 2). To make the ERCC64 amplicon pool, PCR reactions were pooled with adjustment based on length of the ORF, where higher volumes were used for longer ORFs. Using the pool of amplicons as template, a single biotin-labeling PCR was done. The final probe set is designated as ERCC64 in this manuscript.

Using a protocol similar to that used for synthesis of the three TFs probeset, an OCS probe set was created which correspond to ORFs from the Spike-in RNA Variant Control Mixes, or SIRV Mixes (Lexogen). The probe set is designated as SIRV7 in the manuscript. Differences in the protocol are described in this section. The SIRV Mixes are 69 synthetic transcripts from seven genes which mimic the highly complex splicing patterns found in the human transcriptome\(^\text{32}\). We obtained PCR products of SIRVs constructs corresponding to SIRV101, SIRV201, SIRV301, SIRV403, SIRV510, SIRV601, and SIRV701. ORF-specific primers were used for the biotin-labeling PCR and were designed to anchor the ATG-start and just upstream of the stop codon. All SIRV primer sequences used may be found in Supplementary Table 3.

Using a protocol similar to that used for synthesis of the three TFs probeset, OCS probe sets were created from different size pools of ORFs derived from transcription factor (TF) genes. For example, TF2 is a probe set that corresponds to two TFs. The probe sets are designated as TF2, TF12, TF88, TF682, TF763 in the manuscript. These probe sets were synthesized for the purpose of enriching transcripts from different sets of human transcription factors (TFs). To make pools, ORF clones in the form of bacterial stocks were cherry-picked from the human ORFeome collection using Genesis Automated Liquid Handler (Tecan). PCR success was checked by running a subset of the reactions on an E-Gel 96 agarose gel (Thermo Fisher Scientific). Pools corresponding to 2, 12, 88, and 682 ORF-specific primers were used as templates for PCR for TF763, several pools were made, each containing ORFs of a similar length. Each pool underwent a separate biotin-labeling PCR. TFs belonging to each probe set may be found in Supplementary Table 4.
Sequence validation of OCS probes. Each OCS probe set was subjected to Illumina sequencing to verify the probe identities and abundances across the source templates. The Illumina sequencing data for one of the Clontech primers (Roche) was used, in which barcoded Illumina adapters were ligated directly to the probes. Samples were prepared on a Beckman Coulter Biomek FX. For each sample, approximately 50,000 paired-end reads of length 150 bp were generated on an Illumina MiSeq.

To estimate the representation of each OCF within a given probe set, Kallisto[31] (version 1.4.2, default parameters) was used to estimate gene-level abundances. Paired-end reads were aligned. Alignment indices were prepared from a FASTA file containing all human ORFeome sequences. For analysis of probe sets involving ERCC or SIRV standards, the relevant sequences were included in the FASTA file. To generate read coverage plots across the ORF, a set of landmarks were defined: a “local” option with default parameters. The alignment file, in SAM format, was parsed using SAMTools[54] (version 1.2) and custom Python scripts were used to extract read coverage across the ORF on a per-nucleotide basis.

Quantitative PCR (qPCR) method development. Standard solutions of ORF inserts from ARNTL, STAT1, and ZNF268 (Supplementary Table 1) were used. The probes are high purity oligonucleotides (120-mers) with a biotin conjugated at the 5′ end. For each target ORF, a ~1X diluting density was maintained by designing probes that randomly tile the forward and reverse complement sequences. Following reconstitution to 0.75 pmol/µl with TE buffer, probes from the target ORF were combined in equal ratios.

For the single probe enrichments, ARNTL probes “ARNTL_forward_7” (returned 8% on-target) and “ARNTL_recomp_17” (returned 10% on-target) were used. See Supplemental Table 1 for sequences.

Enrichment of TFs from cDNA from 7 human tissues. A capture protocol was developed to allow for quantification of target (three TFs) and background (from high abundance housekeeping genes) ORFs (see Fig. 1). TaqMan PCR assays (Integrated DNA Technologies) were designed against 450-500 bp regions within each of the five genes. The long target region length enables quantification of full-length target cDNAs with minimized background interference from the OCS or IDT oligonucleotide probes. The Primetime Gene Expression Master Mix and accompanying protocol was used as per manufacturer’s protocols, except for the extension time, which was increased to 120 s. Because of vector backbone.

The hybridization experiments, in which the probes are allowed to bind to the target cDNA, were performed using the following conditions. The cDNA was heated to 95 °C for 10 minutes, followed by a rapid ramp down to 65 °C. Either 4 ng of OCS probes or 3 pmol of IDT probes were added and the mixture was incubated at 65 °C for 4 hours. 50 µl of M-270 streptavidin beads (Invitrogen) were added and a series of washes were performed according to the IDT xGen Lockdown protocol version 2, except that initial washes used wash buffer preheated to 72 °C instead of 65 °C to reduce non-specific binding.

For the hybridization-based capture, an on-bead PCR was performed to generate adequate amounts of captured material for sequencing. After the washes, the final bead solution was resuspended in 50 µl of TE buffer. To amplify the full-length cDNAs that were captured on the beads, on-bead PCR was performed with 5 µl of resuspended beads in a 30 µl reaction using KAPA HiFi HotStart 2X mix (KAPA) and the universal Clontech primer. The program was run for 30 cycles, with each cycle consisting of denaturation at 98 °C for 20 s, annealing at 65 °C for 13 s, and extension at 72 °C for 5 min. The final extension was for 10 min.

In some cases, rather than an on-bead PCR, a heat elution was performed for the purpose of quantifying abundances of bound cDNA via qPCR. An aliquot of beads was diluted 10-fold with buffer EB (10 mM Tris-HCl, pH 8.0) (Qiagen) and heated at 95 °C for 5 min. Beads were placed on a magnet and supernatant recovered for subsequent qPCR analysis.

Illumina library preparation and analysis. The transcript abundances in human brain RNA were quantified using RNA-Seq (data used in Fig. 1). This study, in which the RNA-Seq data was obtained from total RNA from the vertebral cortex region (Biochain) using the protocol described in Preparation of cDNA. cDNA was converted into an Illumina library using the NEBNext protocol (New England Bionsciences) and ~20 million PE75 reads were sequenced on an Illumina NextSeq, in duplicate.

Sequencing data was collected at the Center for Cancer Computational Biology (CCCB) at the Dana-Farber Cancer Institute.

To estimate expression values for each gene, RSEM was used with the STAR aligner. The STAR genome index was built based on hg38 and using annotation obtained from GENCODE (version 27). Transcripts per million (TPM) values were calculated using RSEM (version 1.2.29).

The following section describes the sequencing and quantification of enriched cDNA from capture experiments. Sequencing data was collected for the enrichment of three TFs (Fig. 1d). ERCC ORFs (Fig. 2b), and the other TFs studied (Fig. 4a). Illumina sequencing data was collected following the procedure described in Sequence validation of OCS probes. To quantify isoform- and gene-level expression, Kallisto (version 0.44.0) was used with default parameters. To estimate values for all human genes (as in Figs. 1d, 4a), Kallisto indices based on GENCODE (version 27) transcript sequences were used. To estimate expression values for each ERCC ORF, Kallisto indices based on ERCC and GENCODE (version 27) sequences were used.

For the TF multiplexing experiment described in Fig. 3, Illumina sequencing data was collected on cDNA subjected to a workflow similar to Nextera sequencing (Plexwell sequencing, SeqWell). For gene quantification, Kallisto (version 0.44.0) was used with default parameters.

For each reaction, ~1 µg of either input cDNA or captured cDNA was converted into a SMRTbell library using the SMRTbell
PacBio data analysis with Iso-Seq 3. Bioinformatics analysis was done by using the Iso-Seq 3 application in the PacBio SMRT Analysis v6.0 to obtain high-quality, full-length transcript sequences, followed by downstream analysis, as described below.

First, full-length reads were identified from the raw data. Full-length reads were determined as CAGE consensus sequence (CCS) reads that contained both the 5' and 3' primer and the polyA tail preceding the 3' primer. The 5' primer consists of the Clontech SMARTer cDNA primer with an ATGGG overhang. The 3' primer consists of a 16 bp-long PacBio barcode that is sample-specific followed by the Clontech SMARTer cDNA primer.

Second, isoform-level clustering was done to obtain high-quality transcript sequences. To increase detection of rare isoforms, the de-multiplexed full-length reads were pooled to perform isoform-level clustering analysis38. After clustering, consensus sequences were called using the Arrow algorithm (within SMRT Link version 7 software) and only polished sequences with predicted consensus accuracy ≥99% were considered high-quality and retained for the next step.

Third, clustered sequences were mapped to the reference genome and filtered for on-target isoforms. The high quality transcript sequences were mapped to hg38 using minimap225 (version 2.11-r797) using parameters --ax splice -t 30 -uf --secondary=no -c5. We then filtered transcripts mapped to targeted probe region with ≥99% coverage and ≥95% identity.

Fourth, the final isoforms were de-multiplexed by sample barcodes. We recovered the relative abundance of each of the final isoforms in each sample by extracting the fraction of full-length reads supporting each isoform from each sample.

Last, isoforms were annotated and assessed for quality with SQANTI2. SQANTI is a computational tool for annotation and quality assessment of full-length reads sequenced on long-read platforms39. We adapted SQANTI so that it is a computational tool for annotation and quality assessment of full-length reads were drawn at random and, for each subsampled pool of reads, the number of unique genes or transcript isoforms detected was determined. For each sampling depth, 100 iterations of sampling were done, and the average of the number of unique genes or isoforms observed was computed. Only isoforms exactly matching GENCODE (v29) were considered. The saturation-discovery curve analysis uses ‘subsample.py’ and ‘sub-sample_with_category.py’ scripts in https://github.com/Magdoll/cDNA_Cupcake.

Gencode recoveries. For each Gencode isoform (i.e., ENST) in version 29, we determined if there was an exact match in the PacBio transcripts. An exact match is defined as cases in which the detected full-length isoform contains an exact sequence of junctions (i.e., introns) as found in the GENCODE transcript. This was accomplished using a modified version of the SQANTI program, SQANTI2.

Functional features of isoforms annotated within SQANTI2. In this section is described the various functional features assessed within SQANTI2 to evaluate the quality of novel isoforms, which includes CAGE peak overlap, junction conservation, and polyA motifs.

For the CAGE peak overlap, the overlap between the 5' end of isoforms and CAGE peak data was assessed. CAGE peak annotations27 were downloaded from: http://fantom.gsc.riken.jp/5/datasets/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_coord.bed.gz.

Genomic coordinates were converted from hg19 to hg38 using the liftOver program from the UCSC Genome Browser35. The genomic position of the 5' end of the isoforms was compared to all CAGE peaks and the following peaks were described: (i) the distance between the 5' end and the center position of the closest CAGE peaks, and (ii) whether the 5' resided within a range of a CAGE peak.
8. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single molecule long-read survey of the human transcriptome. Nat. Biotechnol. 31, 1009–1014 (2013).

9. Vrga, E. R. et al. Improved read accuracy with the R2C2 method enables the sequencing of highly multiplexed full-length single-cell cDNA. Proc. Natl. Acad. Sci. USA 115, 9726–9731 (2018).

10. Tilgner, H. et al. Comprehensive transcriptome analysis using synthetic long-Read sequencing reveals molecular co-association of distantly spliced events. Nat. Biotechnol. 33, 736–742 (2015).

11. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656 (2019).

12. Spataro, N., Rodriguez, J. A., Navarro, A. & Bosch, E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. Hum. Mol. Genet. 26, 489–500 (2017).

13. Djebali, S. et al. Landscape of transcription in human cells. Nature 489, 101–108 (2012).

14. Mamanova, L. et al. Target-enrichment strategies for next-generation sequencing. Nat. Methods 7, 111–118 (2010).

15. Gnikue, A. et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 27, 182–189 (2009).

16. Haiden, J., Zaghool, A. & Feuk, L. Exome RNA sequencing reveals rare and novel alternative transcripts. Nucleic Acids Res. 41, e6 (2013).

17. 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, 8 (2009).

18. Mercer, T. R. et al. Targeted long-read sequencing for gene discovery and quantification using RNA Capture-Seq. Nat. Protoc. 9, 989–1009 (2014).

19. Ueno, T. et al. High-throughput resequencing of target-captured DNA in cancer cells. Cancer Sci. 103, 131–135 (2012).

20. Mercer, T. R. et al. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat. Biotechnol. 30, 99–104 (2012).

21. Scaglioni, P. P. et al. Solution hybrid selection capture for the recovery of functional full-length eukaryotic cDNAs from complex environmental samples. DNA Res. 21, 685–694 (2014).

22. Giali, M. et al. Comparative analysis of targeted long read sequencing approaches for characterization of a plant’s immune receptor repertoire. BMC Genomics 18, 564 (2017).

23. Karamitos, T. & Mørkøv, G. A novel method for the multiplexed target enrichment of MinION next generation sequencing libraries using PCR-generated baits. Nucleic Acids Res. 43, e152 (2015).

24. Wang, M. et al. PacBio-LITS: a large-insert targeted sequencing method for improved nanopore read accuracy with the R2C2 method. Nucleic Acids Res. 42, e82 (2014).

25. Bashir, S. M. et al. Dynamic genome organization in the context of human disease. Nature 521, 44–51 (2015).

26. Zheng, Y. et al. Improved nanopore read accuracy with the R2C2 method. Bioinformatics 33, i347–i350 (2017).

27. Wang, M. et al. PacBio-LITS: a large-insert targeted sequencing method for improving nanopore read accuracy with the R2C2 method. Nucleic Acids Res. 42, e82 (2014).

28. Bashir, S. M. et al. Dynamic genome organization in the context of human disease. Nature 521, 44–51 (2015).

29. Zheng, Y. et al. Improved nanopore read accuracy with the R2C2 method. Bioinformatics 33, i347–i350 (2017).

30. Wang, M. et al. PacBio-LITS: a large-insert targeted sequencing method for improving nanopore read accuracy with the R2C2 method. Nucleic Acids Res. 42, e82 (2014).

31. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic enrichment of MinION next generation sequencing libraries using PCR-based baits. Nucleic Acids Res. 43, e152 (2015).

32. Wang, M. et al. PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. BMC Genomics 16, 12 (2015).

33. Witek, M. et al. Accelerated cloning of a potato late blight-resistance gene using RemSeq and SMRT sequencing. Nat. Biotechnol. 34, 656–660 (2016).

34. Giali, M. et al. Targeted capture and sequencing of gene-sized DNA molecules. Biotechniques 61, 315–322 (2016).

35. Lagarde, J. et al. High-throughput annotation of full-length long noncoding RNA with capture long-read sequencing. Nat. Genet. 49, 1731–1740 (2017).

36. Deveson, I. W. et al. Universal alternative splicing of noncoding exons. Cell 6, 245–255 (2018).

37. ORFeome Collaboration. The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nat. Biotechnol. 245, 245–255 (2018).

38. Gordon, S. P. et al. Widespread polycistronic transcripts in fungi revealed by an unbiased high-throughput annotation pipeline. Nucleic Acids Res. 47, D383–D388 (2019).

39. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. 50, 151–158 (2018).

40. Tong, M. et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. 17, 74 (2016).

41. Singh, M. et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. 10, 3120 (2019).

42. Noonan, J. P. et al. Sequencing and analysis of Neanderthal genomic DNA. Science 314, 1113–1118 (2006).

43. Salehi-Ashtiani, K. et al. Isoform discovery by targeted cloning, deep-well pooling and parallel sequencing. Nat. Methods 5, 597–600 (2008).

44. Li, Y. I. et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. 50, 151–158 (2018).

45. Noonan, J. P. et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. 17, 74 (2016).

46. Singh, M. et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. 10, 3120 (2019).

47. Noonan, J. P. et al. Sequencing and analysis of Neanderthal genomic DNA. Science 314, 1113–1118 (2006).

48. Salehi-Ashtiani, K. et al. Isoform discovery by targeted cloning, deep-well pooling and parallel sequencing. Nat. Methods 5, 597–600 (2008).
**Additional information**

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16174-z.

Correspondence and requests for materials should be addressed to G.M.S. or D.E.H.

Peer review information Nature Communications thanks Rickard Sandberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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