Advantages of RNA CaptureSeq for profiling genes with alternative splicing or low expression.

(a) Schematic figure indicating limitations of qRT-PCR for quantifying alternative splicing events. (b,c) Dynamic range of K562 cell transcriptome populations demonstrated by transcript (b) or exon (c) expression. Notably, the top 1% of transcripts comprises 38.4% of the total expressed mRNA population. (d) Calculated maximal fold enrichment achieved by CaptureSeq relative to number of genes (combining all known isoforms) targeted (estimated gene expression based on average gene expression in human K562 cell line). Note that higher enrichments can be maintained by removing highly expressed isoforms and gene loci from CaptureSeq targets.
Supplementary Figure 2

Comparative analysis of ERCC spike-in quantification using RNA sequencing and CaptureSeq.

(a) Fold enrichment achieved by CaptureSeq for each ERCC standard. High and variable enrichment at low ERCC concentrations results from low and sporadic alignment of RNAseq reads to ERCC standards. Decreasing enrichment at high ERCC concentrations is due to CaptureSeq saturation. Each technical replicate capture hybridization contained three biological replicate samples. (b) Spearman correlation of measured abundance by CaptureSeq of ERCC probes for three biological replicate samples in technical replicate. (c,d) Average Spearman correlation of measured abundance of ERCC probes for three biological replicates of CaptureSeq (c) and RNA-seq (d). (e) Segmented regression analysis indicates inflection point in the measured abundance of ERCC probes by CaptureSeq at an ERCC concentration of 2.34 attomol/μl (dotted line). n = 3 biological replicates; error bars are s.d. (f) RNA sequencing exhibits a linear profile across the range of ERCC concentrations it detects. n = 3 biological replicates; error bars are s.d.
Supplementary Figure 3

Enrichment in read coverage of ERCC RNA spike-in by CaptureSeq.

(a,b) Averaged read coverage for each ERCC probe from RNA-seq and CaptureSeq. \( n = 3 \) biological replicates; error bars are s.d. Horizontal dotted line shows eightfold coverage. RepA and RepB are technical replicate capture hybridizations containing three biological replicate samples. (a) Number of ERCC transcripts required for eightfold coverage. (b) Concentration of ERCC transcripts required for eightfold coverage. Vertical dotted line marks concentrations above which CaptureSeq is saturated. Lowest three concentrations of probes (<0.00114 attomol/μl) have zero coverage in more than 50% of RNA-seq libraries. (c) Fold difference in variability between RNA-seq measurement of ERCC abundance and CaptureSeq technical replicates (\( n = 3 \) biological replicates). Horizontal dotted line is at 1 and –1 (no difference in variability). Values above 1 show RNA-seq is more variable; values below –1 show CaptureSeq is more variable. Vertical dotted line is the ERCC concentration that allows consistent eightfold coverage by RNA-seq. RNA-seq is more variable at low expression levels. (d) Mean difference between CaptureSeq and RNA-seq accuracy in measuring ERCC abundance. RNA-seq provided less accurate expression measurements at low levels but was more accurate at high levels. \( n = 3 \) biological replicates; error bars are s.d.
Comparative analysis of technical bias between RNA-seq and CaptureSeq.

(a, b) Relationships among ERCC length (a), GC% (b) and CaptureSeq performance on moderately expressed probes compared to RNA-seq (enrichment residuals shown). Spearman correlation shown; line is nonlinear regression fit. RepA and RepB are technical replicate capture hybridizations containing three biological replicate samples. (c) Combined sequence read coverage across ERCC all standards merged (left) or two representative ERCC controls (middle and right) by RNAseq (blue) and CaptureSeq (red). Difference between read coverage indicated by gray shaded area. (d) Relative nucleotide enrichment for ERCC sequences that exhibit differential coverage between RNA-seq and CaptureSeq. No
specific nucleotide bias is observed in regions exhibiting differential coverage. (e) Sequenced read coverage profile of *SMPD2* by RNA-seq (blue) and CaptureSeq (red). Only minor variation is observed between the two profiles.
Analysis of differential gene expression between samples with CaptureSeq.

(a) Pearson correlations of measured abundance of ERCC probes for one representative sample versus all others containing the same ERCC mix. Top, ERCC mix 1; bottom, ERCC mix 2. Two multiplexed capture hybridizations were performed containing a mix of ERCC mix 1 and 2 samples. Slightly higher correlations equate to samples present in the same hybridization. (b) Clustering of ERCC read counts following variance stabilizing transformation. ERCC mixes 1 (n = 5) and 2 (n = 4) clearly separate followed by separation by capture hybridization. Samples present in same hybridization shown in red and black, respectively. (c,d) Relationship between ERCC concentration and detected ERCC abundance. Segmental linear regression to determine the ERCC concentration at which saturation occurs (dotted line). Error bars are s.d. Linear slopes from segmental linear regression and the Pearson correlation for non-saturating concentrations are provided. (c) ERCC mix 1 samples; n = 5 biological replicates. Saturation at 1.30 attomol/μl. (d) ERCC mix 2 samples; n = 4 biological replicates. Saturation at 0.976 attomol/μl. (e) Averaged read coverage for each ERCC probe from ERCC mix 1 (n = 5) and ERCC mix 2 (n = 4) pools. Error bars are s.d. Y-axis dotted line shows eightfold coverage. (f) edgeR MA plot of log fold change for each ERCC control between the two mixes against transcript expression in log CPM (counts per million) expression.
million). Differentially expressed (DE) controls colored red; non-DE colored black. Zero fold change between two samples shown by blue line. edgeR performed using TMM normalization.
Supplementary Figure 6

Comparison of CaptureSeq and RNA-seq for differential gene expression analysis.

Comparison of CaptureSeq and RNA-seq for differential gene expression analysis. (a) Quantification of fold changes in ERCC standard abundances between two distinct samples (ERCC 1, \(n = 5\) biological replicates; and ERCC 2, \(n = 4\) biological replicates) for CaptureSeq and RNA-seq (with a matched number of reads). CaptureSeq records values for all ERCC standards (92); expression values were not obtained for 11 standards with RNA-seq. Slopes from nonlinear regression with a straight-line fit. (b) Variability in fold-change measurements for each ERCC fold-change category between CaptureSeq and matched RNA-seq. For each category RNA-seq showed greater variation. (c,d) edgeR MA plot of log fold change for each ERCC control against transcript expression in log CPM (counts per million). Differentially expressed (DE) controls colored red; non-DE colored black. Zero fold change between two samples shown by blue line.
Matched RNA-seq (c), RNA-seq all reads, no downsampling (d). (e) Relationship between ERCC expression level (log CPM) and ability of CaptureSeq and RNA-seq to detect DE, given various levels of expression differences between two groups. Left, CaptureSeq; middle, matched RNA-seq; right, RNA-seq all reads, no downsampling. FDR, false discovery rate. 1% FDR shown by dashed line. FDR values limited to minimum value of $10^{-37}$. 
Supplementary Figure 7
Targeted sequencing of IncRNAs with CaptureSeq.

Frequency distribution of expression for different gene classes according to biotype (a), gene ontology biological function (b) or annotation in disease database (c) in K562 cells. (d) Frequency distribution of probes relative to fraction of length with overlapping alignments from captured genomic DNA. We found greater than onefold coverage across the entirety of 96.5% of probes, thereby validating the ability to capture gDNA. (e) Plot showing measured relative to known abundance of ERCC standards by CaptureSeq (orange) and RNA-seq (dark blue). We have plotted measured abundance before (orange, light blue) and after (red, dark blue) removing duplicate reads. Although removing duplicate reads may reduce the impact of PCR amplification artifacts, it also causes the abundance of ERCC spike-ins to be underestimated, decreasing the quantitative range of CaptureSeq, and is therefore not recommended. (f) Genome browser view showing read alignment profile and assembled transcripts from RNA-seq (upper) and CaptureSeq (lower) across the Titin-antisense IncRNA locus. CaptureSeq read alignment shows higher specificity for exons, with fewer reads derived from nascent transcription present, resulting in more accurate transcript assembly. By contrast, RNA-seq shows a large amount of nascent transcription, resulting in the misassembly of the transcript locus with ‘retained’ introns.
Supplementary Figure 8

Quantitative accuracy of each tissue within expression atlas.

Measured expression (FPKM) of ERCC standards in each human tissue library analyzed. Pearson’s correlation indicates the quantitative accuracy of libraries following capture. Despite enhanced coverage, some ERCC probes (red) remained undetected, indicating that sequencing had not proceeded to saturation.
Supplementary Figure 9

Analysis of novel captured exons and isoforms.

(a) Proportion of introns with canonical splice junctions in previous coding and lncRNA exons is similar to new introns...
identified using CaptureSeq. (b) Sequence motif at 3’ intron end shows similar enrichment for poly-pyrimidine tract and splice elements in previously annotated introns and new introns identified by CaptureSeq. (c) Example of multiple previous IncRNA annotations that are merged into single higher-order contiguous IncRNA loci following more complete and accurate assembly with CaptureSeq. (d,e) Frequency distribution of open-reading-frame length and hexamer score indicates distinction between coding and noncoding transcripts analyzed from CaptureSeq assembled transcripts. (f) Box-whisker plot showing that CaptureSeq assembled gene models contained more exons and were more complete than previous annotations (based on GENCODE v19, Cabili et al. (2011), and IncRNAdb) used to design the capture array.
Supplementary Figure 10

Comparative analysis of captured and annotated coding and noncoding exons.
(a) Cumulative frequency distribution indicating the conservation (according to 100-way MutliZ Alignment) of previously annotated (based on GENCODE v19, Cabili et al., and IncRNAdb) coding and IncRNA exons relative to novel exons identified using CaptureSeq. (b) Conservation at 3’ exon boundary showing 3-nt periodicity characteristics of previously known coding gene exons (red) relative to new coding gene exons (orange) identified by CaptureSeq and (c) similar conservation of splice elements in previous IncRNA annotations relative to new IncRNA exons identified by CaptureSeq. (d) Comparison of SNP, repeat and predicted RNA secondary structure density between previous gene annotations (based on GENCODE v19, Cabili et al. (2011), and IncRNAdb) and new annotations assembled from CaptureSeq experiments.
Supplementary Figure 11

Examples of transcripts assembled following CaptureSeq.

(a) Targeting previous lncRNA annotations (blue) integrates them into a single complex locus. (b) CaptureSeq ensnares additional novel exons into the initial annotation, thereby expanding the previous annotation to annotated TSS’s. (c) CaptureSeq revises previous lncRNA annotations to identify a 1021 amino acid ORF. An assembly gap in GRCh37 (hg19) means the protein N-terminal may not be present. A new contig in GRCh38 places MGC50722 6kb upstream suggesting the possibility these two loci form one gene. (d) LncRNA can be erroneously annotated when only transcript fragments are available as demonstrated in example showing a lncRNA locus contains distal coding exons for downstream NPAS4 gene. Arrows indicate direction of transcription. Fantom 5 TSS on forward strand (red) and reverse strand (blue).
Supplementary Figure 12

Tissue-specific expression of captured IncRNAs.

(a) Hierarchal clustering of IncRNA loci according to expression. (b) Example of brain-specific IncRNA Evf2 correctly assembled and quantified using CaptureSeq.
Supplementary Figure 13

Examples of captured novel coding exons.

Examples of novel coding exons within GENCODE genes assembled following CaptureSeq. (a) Identification of a novel transcription start site for GLIS1 gene well supported by chromatin marks for transcriptional initiation. New first exon adds 175 amino acids to 5' of protein. (b) Novel internal coding exons in TNXB. Novel exon(s) are conserved and maintain the TNXB reading frame. (c) Targeting novel exons solely identified by evolutionary conservation enables the identification of novel exons that help assemble multiple HMNC2 annotation fragments into a contiguous gene locus. (d) Putative novel coding locus in bi-directional orientation with ZNF593 contains a 186 amino acid ORF.
SUPPLEMENTARY RESULTS

Design validation with genomic DNA controls
We performed an initial genomic DNA capture to confirm the efficacy of the lncRNA capture design. We found that 74.5% of gDNA sequenced reads overlapped probes, with 96.5% of probes having more than 1-fold coverage across their entire length (Figure S7d).

The lncRNA design included probes targeting intergenic regions to measure potential DNA contamination. By comparing the relative enrichment of reads aligning to experimental probes (that target exons) to the control probes can provide an indication as to the existence of DNA contamination. We found a ~1,874 fold relative enrichment of experimental probes targeting exons relative to control probes when capture is performed with cDNA libraries. By comparison, we observe a 1.57-fold enrichment for reads aligning to control probes relative to experimental probes when capture is performed with genomic DNA. Taken together these results demonstrate a very low level of potential DNA contamination that could affect our lncRNA capture results.

Read Coverage across transcripts is similar between RNAseq and CaptureSeq
The single exon nature of ERCC standards makes them well suited for the analysis of read coverage heterogeneity. Both RNAseq and CaptureSeq exhibit similar and highly reproducible sequence coverage across the full length of standards, with only minor (7.8%) divergence observed (Figure S4c). In those regions of ERCC transcripts that exhibit different coverage, we observe no significant association with nucleotide content (paired t-test, p = 0.86; Figure S4d). We also examined read coverage across annotated control genes that were included within probe designs. We observed similar and highly reproducible sequence coverage across the full length of control genes in both RNAseq and CaptureSeq (Figure S4e). This collectively suggests that sequence-specific biases in RNAseq library preparation, sequencing and alignment introduce greater technical variation than CaptureSeq specific protocol steps of hybridization and washing.

Potential biases for CaptureSeq enrichment
For moderately abundant ERCC standards (0.0366-2.34 attomoles/μl) we compared the quantification of ERCC standards relative to their length and GC%. For CaptureSeq, we observe an inverse relationship between standard length and enrichment p <0.001 (Figure S4a). Given that short sequences are generally under-represented in RNAseq libraries, this suggests CaptureSeq may perform better in the detection and quantification of short transcripts. We observe no significant relationship between enrichment and average transcript GC% (Figure S4b), though this may not hold true at GC percentages not covered by the ERCC standards.

Duplication bias
We considered whether PCR amplification artifacts may be introduced by the additional PCR step in the CaptureSeq. We previously found little divergence in
sequence read coverage across genes between RNAseq and CaptureSeq, and no evidence of ‘block’ coverage characteristic of PCR duplication (Figure S4e).

Given the higher coverage of exons by reads in CaptureSeq, duplicated reads likely comprise both PCR duplication and bona-fide identical sequenced reads. We assessed the quantitative accuracy for measuring ERCC standards following the removal of duplicate reads, finding a slightly lower correlation following duplicate removal ($R^2$-0.9966 to 0.9989) that is exacerbated for abundant ERCC standards. Therefore, we recommend not removing duplicate reads, however, digital PCR methods\(^2\) will be required to distinguish relative contributions of PCR duplication biases and bona-fide identical sequenced reads.

**Linear quantitation of RNAseq and CaptureSeq**

In addition to the correlations observed between the measured and known abundance of ERCC standards, we also examined the linearity of quantitation (how increases in ERCC abundance equate to a proportional increase in measured expression) for RNAseq and CaptureSeq by non-linear regression with a straight line fit. As previously reported, RNAseq shows a linear quantitation over six orders of magnitude\(^1\). CaptureSeq displayed regression slopes of 0.050 (95% CI 0.0464 to 0.05334) for technical replicate one and 0.033 (95% CI 0.03063 to 0.03581) for technical replicate two (each technical replicate capture hybridization contained three biological replicate RNA samples, see **Online Methods**). These lower slopes were largely due to CaptureSeq saturation at high expression levels; below the saturation cutoff in the CaptureSeq dynamic range the linear slopes were 0.7832 (95% CI 0.7391 to 0.8273) and 0.6358 (95% CI 0.6007 to 0.6710) respectively. Given that CaptureSeq provided a linear response over ~4 orders of magnitude (and possibly more as the limit of the dynamic range was below the lowest ERCC concentration), the quantitative efficiency with which CaptureSeq detected a 10 fold change in transcript abundance below the saturation cutoff was $\geq$ 90%.

At low expression levels the linear quantitation of RNAseq was above 1 (1.24 95% CI 1.006 to 1.474), demonstrating the change in measured expression was larger than the actual change in ERCC abundance. This is due to the paucity of detected reads for many low level ERCC standards, including a number that were not detected in individual replicates. This results in artificially low RPKM values for many of the rarest ERCC standards, maintaining the appearance of dynamic range at these concentrations. However, the lower correlation values and high variability between replicates (Figure1, S2, S3) make it clear these values are not accurate quantitative measurements of transcript abundance.

**Measuring the fold differences in gene expression between samples with Capture Sequencing**

We utilized a second set of Capture samples to examine the potential for CaptureSeq to identify gene expression differences between groups of samples. RNA from 9 individuals was spiked with either the previously utilized ERCC mix number 1 (5 samples) or ERCC mix number 2 (4 samples). The ERCC controls in mix 1 and 2 differ by known amounts allowing the determination of both true and false positive and negative calls for differential transcript abundance. The samples used were randomized to one of two multiplex capture hybridizations.
The correlations between the expression levels of each ERCC control were very high (average mix 1 Pearson $R^2 = 0.968$, average mix 2 Pearson $R^2 = 0.970$), with representative comparisons shown in Figure S5a. Similar results were observed with sample clustering (Figure S5b). These results confirm samples from different multiplex captures have very high replicability with only minor effects from the capture process (hybridization, washing, post capture LMPCR). Nevertheless, we recommend randomising samples to different captures as this will ensure the true (although small) technical variability of the procedure is represented in the sample results.

Similar to Figure 1 and S2, we investigated the CaptureSeq dynamic range and saturation point for the 9 samples. Despite a lower overall sequencing depth (average of ~8.4 fold less ERCC reads per sample), all ERCC controls were consistently detected Figure S5c-e. Segmental linear regression supported a linear slope up to a saturation point of 1.3 and 0.976 attomoles/μl for mix 1 and mix 2 respectively, similar to the 2.34 attomoles/μl saturation point identified for the first set of samples. Hence CaptureSeq provides a linear response over ~4 orders of magnitude and possibly more because the limit of the dynamic range was below the lowest ERCC concentration. Linear slopes (out of 1) prior to saturation of 0.45 (mix 1) and 0.35 (mix 2) suggest the quantitative efficiency with which CaptureSeq detected a 10 fold change in transcript abundance was around or over 80%.

We next examined the performance of CaptureSeq in measuring the differences in transcript abundance between ERCC mixes. Although CaptureSeq performed well in detecting RNA abundance differences, the detected fold changes and linear slope suggest CaptureSeq can underestimate the abundance differences between samples (Figure 1e, S6a). Notably, no relationship was found between ERCC concentration and the accuracy of fold change measurement, i.e.: the differences in ERCC abundance between the mixes were measured just as accurately above the saturation threshold as below it. An explanation for this observation is that in a capture hybridization containing multiple libraries, each ERCC transcript is competing with the same transcript from the other libraries to bind a capture probe. When the total amount of probe is limiting, the same transcript in each library would compete and bind to the probes in proportion to its abundance and therefore accurate fold change information is obtained even when expression level information is not.

To determine if the detected fold differences were sufficient to identify differential expression (DE) we tested the CaptureSeq results with edgeR. Under standard conditions with a 1% FDR, 100% of the ERCC controls with 4 fold or 2 fold changes were correctly identified as DE, as were 52% of controls with 1.5 fold changes (Figure S5f). Two (of 23) ERCC controls with no change between the mixes were called DE by edgeR, giving a false positive rate of 9% and a recall value of 0.84. Sensitivity was highest for ERCC controls at moderate concentrations, but CaptureSeq also identified DE at the lowest and highest concentrations.

We compared the CaptureSeq results for transcript abundance and differential gene expression profiling to those of standard RNAseq performed on the same libraries prior to capture. Precapture control libraries were sequenced to a greater depth and had, on average, 6.3 times more mapped reads than the CaptureSeq libraries. This allowed us to compare CaptureSeq and RNAseq results with a matched number of
mapped reads (using downsampling), but also to investigate if RNAseq could match CaptureSeq in identifying DE at low expression levels though deeper sequencing.

Unlike CaptureSeq, matched RNAseq showed no general underestimation of fold change differences (linear slope 0.98, 95% CI 0.81 to 1.1, compared to 0.71 95% CI 0.66 to 0.75, Figure S6a). However, matched RNAseq provided no results for 11 lowly expressed ERCC controls and the variability of the measured fold changes was larger as a result of less accurate fold change measurements for many other standards (Figure S6b). Testing for DE, with matched RNAseq and a 1% FDR, revealed 65% of the ERCC controls with 4 fold, 57% of 2 fold and 52% of 1.5 fold changes were correctly identified as DE. One (of 23) ERCC controls with no change between the mixes was called DE, giving a false positive rate of 4% and a recall value of 0.58 (Figure S6c). This is a similar false positive, but much lower recall value, than obtained with CaptureSeq due to the decreased ability of RNAseq to identify low level transcripts as DE. Repeating this analysis with all RNAseq reads (6.3 times more mapped reads per library than CaptureSeq) gave the same false positive rate as CaptureSeq, but the recall rate was still lower (albeit improved) at 0.73 (Figure S6d).

This strengths and weaknesses of each technique to identify DE can be visualized by plotting the expression of each differentially expressed standard in comparison to the likelihood, from edgeR, it is DE (Figure S6e). RNAseq shows a clear decrease in its ability to detect DE as expression levels decrease, although (and consistent with RNAseq showing a more linear response to expression level changes), RNAseq appears to perform better at providing high confidence calls when expression level changes are small (see 1.5 fold change results). Collectively, these results demonstrate that CaptureSeq can be used to identify DE over a wide range of gene expression levels and that CaptureSeq provides a more accurate method for quantifying DE of lowly expressed genes, such as IncRNAs. While performing RNAseq at increased read depths does not provide an efficient means of emulating the advantages of CaptureSeq.

Proportion of cellular transcriptome better quantified by CaptureSeq

To provide a context to the advantages of CaptureSeq, we next compared the concentrations of ERCC standards to the total RNA population from K562 cells. Based on our above analysis, we estimate that 42.1% of transcripts would be better quantified using CaptureSeq (Figure 1d). RNAseq and CaptureSeq would perform comparably well for 53.2% transcripts, and RNAseq would perform better for the top 4.6% expressed transcripts that are enriched for housekeeping, structural or metabolic genes. By contrast, the lowly expressed genes in K562 cells for which CaptureSeq provides superior quantitative accuracy are enriched for transcription factors, and genes causally associated with cancer (CancerCensus) and human disease (OMIM, HGMD and BIOBASE) (Figure S7a-c). Although these estimates may vary in other cell-types or tissues according to transcriptome size and complexity, they nevertheless demonstrates the scope of genes for which CaptureSeq may be advantageous.
Identification of novel coding exons

A large-scale study of mammalian conserved elements previously predicted ~3800 novel protein-coding exons in human\textsuperscript{4}. Of these, 2060 exons had no prior evidence of expression. To determine if these putative coding exons create novel isoforms and/or completely novel coding genes, we examined their expression simultaneously with lncRNAs across 20 human tissues and 4 cell lines by CaptureSeq.

We found 543 (26\%) of the putative novel coding exons were expressed in at least one tissue. Most exons (457 or 84.2\%) were incorporated into transcripts that overlapped 333 GENCODE gene annotations (208 coding genes, 72 pseudogenes and 29 lincRNAs). This reveals that only a minority of the expressed novel exons could potentially form completely novel coding genes, with most instead forming novel exons of annotated coding genes or suggesting coding potential in transcripts otherwise thought to be noncoding. The vast majority (>80\%) of putative coding exons associated with pseudogenes match known pseudogene exons, suggesting these putative novel coding exons predictions could be detecting residual pieces of open read frame.

Focusing our analysis on the 315 novel exons to GENCODE v19 coding genes and lncRNAs (lincRNAs and antisense transcripts). We find 72 new 5’, 51 new 3’ and 192 new internal exons, with examples of a new 5’ exons correcting transcriptional and translation start sites and novel alternative internal exons (Table S2 Figure S13A-B). A standout example are the 17 novel exons found to connect together several HMCN2 gene fragments on chr 9 to create a 2836 amino acid ORF. This may not be full length, using the existing exons and splice junctions from our CaptureSeq transcripts, ILL_076783\textsuperscript{5} and ENST00000428715.1 predicts a putative 3703 amino acid ORF (Figure S13C).

Delineating whether transcripts are coding or lncRNAs can be challenging, especially when transcript structures are incomplete. We re-examined 27 GENCODE lincRNA loci with 1 or more putative coding exons, identifying 16/27 as potential coding transcripts, including 4 which contained Pfam\textsuperscript{6} domains (TableS3).

The remaining 86 novel coding exons with no overlap to known genes formed 51 novel transcribed loci. The assembly of putatively coding transcripts without any overlap with GENCODE genes or pseudogenes allows the identification completely novel coding genes, but also has a high probability of identifying false positives. Therefore we selected all novel transcript loci with 2 or more putative coding regions for closer examination. This filtered the novel loci from 51 to 11, 10 of which potentially encode a peptide/protein (TableS3). Due to the limitation in assembling whole transcripts when only capturing a small number of exons 6/10 ORFs were incomplete at the 3’ end and it is plausible four of these form novel extensions to known coding genes. This leaves an estimate of 6 novel coding loci (TableS3, Figure S13D), 3 of which contained Pfam\textsuperscript{6} domains. Probable novel coding loci included ORFs with similarity to mouse fatty acid desaturase 2-like protein and keratins. While we identify possible ORF altering extensions to proteins such as absent in melanoma 1-like (AIM1L) and the poorly annotated protein C4orf50.

These results clearly demonstrate the ability of CaptureSeq to characterise poorly annotated genes, integrating novel exons into existing coding genes as well as identifying novel coding genes missed by previous annotation efforts.
SUPPLEMENTARY REFERENCES