RNA CaptureSeq reveals the full complexity of the human transcriptome

Tim R. Mercer1, Daniel J. Gerhardt2, Marcel E. Dinger1, Joanna Crawford1, Cole Trapnell3, Jeffrey A. Jeddeloh2,4, John S. Mattick1,4 and John L. Rinn3,4

1Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4067, Australia.
2Roche NimbleGen Inc, Research and Development, Madison, WI 53719, U.S.A.
3Department of Stem Cell & Regenerative Biology, Harvard University, Cambridge, MA 02138, U.S.A.

4To whom correspondence should be addressed:
Jeffrey A. Jeddeloh
Tel: +1 608 218 7600
Fax: +1 608 218 7601
Email: jeffrey.jeddeloh@roche.com

John S. Mattick
Tel: +61 7 3346 2079;
Fax: +61 7 3346 2101;
Email: j.mattick@uq.edu.au

John L. Rinn
Tel: +1 617 496 3561
Fax: +1 617 496 9679;
Email: john_rinn@harvard.edu

Transcriptomic analyses have revealed unexpected complexity to the human transcriptome, whose breadth and depth exceeds current RNA sequencing capability1-4. Using tiling arrays to target and sequence select portions of the transcriptome, an approach we term RNA CaptureSeq, we identify and characterize novel transcripts whose rare or transient expression is below the detection limits of conventional sequencing approaches. We apply the unprecedented depth of coverage afforded by this technique to reach the deepest limits of the human transcriptome, exposing widespread, regulated and remarkably complex non-coding transcription in ‘gene deserts’, as well as novel exons and splicing patterns in even intensively-studied protein-coding loci such as p53 and HOX. The data show that intermittent sequenced reads observed in conventional RNA sequencing datasets, previously dismissed as noise, are in fact indicative of unassembled rare transcripts. Collectively these results reveal the range, depth and complexity of a human transcriptome that is far from fully characterized.
RNA sequencing technologies can provide an unbiased profile of the human transcriptome, with techniques of *ab initio* transcript assembly identifying novel transcripts and expanding our catalog of genes and their expressed isoforms. These technologies provide an opportunity to assemble a complete annotation of the human transcriptome, thereby providing a full account of the functional output of the genome and the identification of the differences in gene expression that drive and specify variation between cells. These include not only protein-coding transcripts but also an expanding catalog of long non-coding RNAs (lncRNAs) that are intergenic, overlapping or antisense to annotated genes. However, despite recent technological advances, we have still not yet reached the limits of the transcriptome nor realized its full scale and complexity, fuelling ongoing debate as to the extent of the transcriptome and the biological relevance of transcripts that are expressed at low levels.

To profile such rare transcriptional events and thereby assess the full depth of the transcriptome, we employed an augmented RNA capture sequencing strategy that allowed us to sequence select portions of the transcriptome to saturating depth. Briefly, RNA CaptureSeq involves the construction of tiling-arrays across genomic regions of interest, against which cDNAs are hybridized, eluted and sequenced. While this ability to isolate and target RNA has been employed in genetic analysis for some time, here we combine this ability with deep-sequencing technology in an approach analogous to exome sequencing approaches and previous in-solution capture methods that have been used to identify variation within human protein-coding genes.

To inform the design of arrays and as a comparative reference, conventional RNAseq was initially performed on a primary human foot fibroblast cell line using the Illumina GAII platform (Supplementary Table 1). *Ab initio* transcript assembly of the resulting ~20.4 million alignable paired-end reads yielded 48,091 multi-exon transcripts, of which 88.3% correspond to annotated gene models (Supplementary Data 1). From these annotations, we selected ~50 loci that included both annotated protein-coding genes and functionally characterized long noncoding RNAs (lncRNAs), such as HOTAIR, TUG1 and MEG3, for inclusion on the array (Supplementary Fig. 1a; Supplementary Tables 2,3). In addition, we also included intergenic regions that exhibited little or no transcriptional activity (see below). In total, 2,265 contiguous regions that together comprise ~0.77Mb were represented on the array. To validate the array design we firstly conducted capture sequencing of matched foot fibroblast genomic DNA (Supplementary Results, Supplementary Fig. 1b-d), confirming the specificity, sensitivity, uniformity and reproducibility of the capture arrays, comparable to previous DNA capture and sequencing studies.

RNA CaptureSeq was then performed from matched foot fibroblast cDNA (see Supplementary Methods for detailed protocol). To permit direct comparison, the same sequencing and alignment methods were performed as for pre-capture RNAseq libraries, yielding ~25.8 million alignable paired-end reads using the Illumina GAII platform. In total, 80.7% of captured reads aligned within probed regions, resulting in a mean ~4,607-fold coverage. By comparison, only 0.21% of pre-capture reads aligned to probed regions.
providing a ~8.2-fold alignment coverage (Supplementary Fig. 2a). A comparison between RNAseq and CaptureSeq sequenced libraries showed that the capture protocol did not significantly diminish library diversity nor introduce PCR amplification bias (Supplementary Results, Supplementary Fig. 2b). Given RNA CaptureSeq achieved a ~380-fold enrichment for alignment coverage across targeted regions of the transcriptome, we extrapolate that ~10 billion aligned sequenced reads from a single sample by conventional RNAseq would be required to achieve an equivalent coverage depth across this targeted transcriptional region (Supplementary Fig. 2c).

We next investigated the advantage conferred by the increased sequencing depth of RNA CaptureSeq in de novo transcript assembly, initially focusing on regions containing well annotated protein-coding genes. We reconstructed all genes assembled within pre-capture RNAseq data with a similar uniformity of transcript coverage (100% of transcript chains reconstructed; Figure 1a; Supplementary Fig. 2d; Supplementary Data 2). We identified an additional 204 novel isoforms of 55 protein-coding loci, alone representing a 2.8-fold increase over the current catalog of isoforms for these loci and demonstrating that for even well characterized loci considerable complexity remains to be resolved. Indeed, many of the newly identified exons were entirely undetected within our initial RNAseq libraries (24.7% undetected with a further 10.4% only detected by a single read). For example, previously 3 splicing variations generating up to nine alternative isoforms, each with alternate functional consequences, have been described for the p53 gene. By RNA CaptureSeq we identified an additional 4 alternative isoforms of p53 that were subsequently validated by RT-PCR (Figure 2a; Supplementary Fig. 3a), three of which modified the domain structure of the protein, such as the exclusion of the tetramerization domain required for intra-p53 interactions or modification of the p53 transactivation domain. As a class, the newly identified isoforms exhibit weaker expression (mean 2.4-fold decrease) and conservation (mean 1.8-fold decrease) relative to previously annotated isoforms, but a similarly stringent enrichment for canonical splice junctions (Supplementary Fig. S4a-g). A subset of these rare isoforms also have limited coding potential, representing noncoding variants of the dominant mRNA transcript. Lastly, we also resolved an additional 163 neighboring and antisense IncRNAs around protein-coding genes.

The sequencing depth of RNA CaptureSeq permitted us to assemble ab initio transcripts exhibiting a complex array of splicing patterns. To confirm the intricate structure of assembled isoforms, we performed matched RNA CaptureSeq using the 454 GSFLX Titanium, whose longer read length provides greater power to resolve complex gene structures, yielding ~314,707 reads that aligned to the genome (Supplementary Table 1). Despite this much shallower sequencing depth, ab initio assembly of these longer reads validated the existence of most (64.8% transcript chains reconstructed) newly described isoforms and neighboring IncRNAs. This approach also revealed that, like mRNAs, alternative splicing of IncRNAs can modulate the inclusion or exclusion specific functional domains. For example, the IncRNA HOTAIR exhibited an alternative splice site that eliminates the PRC2 binding domains, as well as small exon length variations supported by canonical intronic polypyrimidine tracts and splice junctions (Figure 3c). Lastly, we also undertook an assembly that incorporated both long 454 and short Illumina reads
(Supplementary Data 4). The synergistic combination of long and short reads has been previously shown to provide additional accuracy in delineating complex and rare spliced isoforms and estimating their relative abundance.\(^{26}\)

We next assessed whether RNA CaptureSeq retains the differential gene expression profile of the original uncaptured sample, thereby permitting the quantitative analysis of captured transcripts. We firstly confirmed the reproducibility between two CaptureSeq technical replicates by qRT-PCR \((r^2=0.99)\) and within sequenced libraries \((r^2=0.94\) and \(r^2=0.97)\) and the uniform enrichment of transcripts following capture (Supplementary Results, Supplementary Fig. 5). Next, to determine the ability of RNA CaptureSeq to compare gene expression between alternative cell types, we applied RNA CaptureSeq to human fetal lung (FL) fibroblast cells whose distinct gene expression program reflects their alternative location within the body (Table S1).\(^ {11}\) Sequencing and assembly using the 454 platform were performed as before, in total assembling 430 multi-exon transcripts (Supplementary Data 5) captured in association with genic probed regions. In comparison to foot fibroblasts, we find 37% of captured genes undergo significant differential expression (Supplementary Fig. 6a).\(^ {27}\) This is aptly illustrated by the opposing polar transcriptional enrichment across the \(HOX\) loci that reflect the positional differences in the origin of foot and fetal lung fibroblasts along the body axis (Figure 3a). Despite high cross-hybridization potential, RNA CaptureSeq faithfully maintains the transcriptional boundary that pivots between the \(HOXA7\) and \(HOXA9\) genes, consistent with previous reports using alternative methods.\(^ {11, 13}\) Next, we confirmed by qRT-PCR that the relative enrichment of \(HOX\) genes along this linear axis was closely maintained following capture (Figure 3b). We observed a close correlation between differential expression profiles for \(HOX\) gene pre- and post-capture that was additionally concordant with estimates of gene abundance obtained from CaptureSeq (Figure 3B, Supplementary Results and Supplementary Fig. 6b). Validations by qRT-PCR were extended to confirm the differential expression of a six lowly expressed intergenic transcripts between foot and lung fibroblasts (see below). Together, this indicates that following both phases of the CaptureSeq approach, capture and sequencing, the gene expression profiles of the original sample are maintained with fidelity, permitting the application of this technique for quantitative analysis.

The existence of intermittent sequenced reads that align within intergenic regions has fuelled recent controversy as to whether they represent the low-frequency sampling of authentic transcripts, or simply reflect biological noise from spurious nascent transcription or technical noise from sequencing and alignment.\(^ {4-6}\) Having established the fidelity of the RNA CaptureSeq approach (Supplementary Results), we applied its sensitivity to characterize these rare transcriptional events within intergenic regions and thereby help resolve this ongoing debate. To achieve this we included numerous intergenic regions for interrogation within our array which, despite overlapping active chromatin domains \((H3K4me3/H3K36me3)\)\(^ {27}\), showed little or no evidence of transcription according to publicly available transcriptomic resources or our own initial pre-capture RNAseq analysis (Figure 1b, Supplementary Fig. 6c).
We found aligned captured reads covered almost all intergenic probed bases (98.1%), comparable in extent to genic regions (94.1% of bases, Supplementary Fig. 7a). However, for our analysis we considered only those regions with evidence of post-transcriptional splicing, retaining in total 45% (443) of intergenic probed regions. The rationale was two-fold: firstly it removed the potential for genomic DNA contamination (we found less than 1.3% of sequenced reads showed evidence of artifactual spliced alignments within our corresponding control capture using genomic DNA); and secondly it omitted the potential for ‘spurious’ transcriptional noise since we reasoned formulaic and reproducible splicing of transcripts necessitates attentive post-transcriptional regulation. Indeed, these regions with evidence of splicing exhibit a 37.5-fold enrichment in aligned read frequency relative to excluded regions with no evidence of splicing (Supplementary Fig. 7b). Together, we captured 798 splice junctions within intergenic probed regions28, of which 95.7% were not identified in pre-captured libraries or pre-existing gene annotations. Despite their novelty, these junctions exhibit similar enrichment for canonical splice motifs as annotated genes (Supplementary Fig. 4). To resolve the complex isoforms that utilize such intricate splicing parameters we performed ab initio transcript assembly, constructing 257 multi-exonic captured transcripts (Figure 1b, Supplementary Data 2). The full-length of the almost all (76.7%) transcripts was independently verified by the longer read 454 sequencing (Supplementary Data 3). Captured intergenic transcripts comprise an average of 3.6 exons, with an average size of 428 bp and mature full length of 1.54 kb. Lastly, RT-PCR and sequencing of amplified products independently validated the existence and structure of almost all tested (13 of 15) assembled intergenic IncRNAs (Supplementary Fig. 3b). Though lower than proteins, captured intergenic transcripts exhibit similar conservation to other annotated functional IncRNAs, and higher than intronic or surrounding intergenic sequence (Supplementary Fig. 7c). A range of metrics, including the presence, size and structure of open-reading frames, homology of predicted open reading frames to known proteins, and synonymous to non-synonymous nucleotide substitution rate confidently ascribed the majority (92.3%) of these transcripts as noncoding RNAs (Figure 3d, Supplementary Fig. 7d).

To contextualize the rarity of captured intergenic transcripts in relation to the whole human transcriptome, expression profiles were firstly normalized between conventional RNAseq and RNA CaptureSeq libraries according to shared genes (Supplementary Fig. 5i). Captured IncRNAs exhibited a mean expression of only 0.011 rpkm29, 463-fold less than the median gene expression within fibroblasts (Figure 3c). We performed quantitative RT-PCR using our pre-capture RNA sample to provide an informed estimate of IncRNA transcript copy number. Assuming an average human fibroblast cell contains ~300 fg of mRNA per cell30, we estimate that the newly discovered IncRNAs were present at an average of ~0.0006 transcripts per cell, indicating expression in only a small sub-population of the cells sampled. By comparison, we calculate HOXA genes to be present at an average ~0.13 transcripts per cell, consistent with previous calculations that estimate around 40% of mammalian genes, and in particular transcription factors, to be expressed at less than one copy per cell30.

Given these intergenic transcripts represent some of the rarest transcriptional events characterized to date, we next considered whether our application of RNA CaptureSeq had
achieved full coverage and therefore reached the limits of the fibroblasts transcriptome. Within our initial (pre-capture) RNAseq we found only minimal and intermittent coverage of the lowly expressed transcripts indicative of low-frequency sampling and non-saturating coverage (the bottom 5th percentile of transcripts are each represented by average 5 aligned reads). Indeed, only 31.3% of captured intergenic transcripts were even detected in pre-capture RNAseq libraries, with a further 9.7% detected by a single alignment. By comparison even lowly expressed transcripts are represented by large numbers of aligned reads in RNA CaptureSeq dataset indicating an asymptotic transcript discovery rate associated with the approach of coverage saturation (the bottom 5th percentile of transcripts are each represented by an average 160.8 aligned reads; Figure 3f, Supplementary Fig. 8a-b).

During pre-capture RNAseq, we found a significant component (14.8%) of alignable reads that could not be assigned into assembled or previously annotated gene models (Figure 3e). These unassigned reads have been previously thought to represent biological or technical artifacts since they superficially display characteristics of random sampling from a low-level background6. We firstly considered unassigned reads aligning to intronic regions that have been previously dismissed as collateral from splicing by-products6. We found a significant overlap between unassigned intronic reads from the pre-capture RNAseq and newly identified isoforms (7.02-fold enrichment; two-tailed p-value < 0.0001 chi-squared test expecting random distribution of read alignments throughout introns; Figure 3g), thereby rescuing 61.5% of unassigned reads aligning within captured transcript introns. We further validated the existence of 4 (of 4) of these novel exons that rescue intronic reads (Supplementary Fig. 8c), confirming they are mature spliced transcripts rather than background unprocessed intronic intermediates. In addition, 53.4% of unassigned reads aligning to probed intergenic regions were also incorporated within assembled intergenic lncRNAs identified by RNA CaptureSeq. We reason that a similarly significant proportion of unassigned reads from our initial RNAseq dataset that fall outside captured regions also correspond to the low frequency sampling of intergenic transcription in a small subset of the cell population. When projected across the genome as a whole and across the range of positionally and functionally distinct cells in the body, as well as during the dynamic processes of development, physiology, and brain function, this in turn suggests a sizeable expansion to the borders of the human transcriptome and anticipates a scale of transcriptional complexity that surpasses even previous reports1.

In this context of an expanded transcriptome, the RNA CaptureSeq approach assumes considerable value by virtue of its ability to focus and magnify the interrogation of regions of interest. For example, it can comprehensively profile haplotype blocks identified by genome-wide association studies to be associated with complex diseases or phenotypes, many of which occur outside of coding genes31, in order to identify all gene products emanating from these regions as the next step in determining causality. In addition, the combination of CaptureSeq with multiplex sample preparation can permit high-throughput transcriptional profiling of large numbers of samples at a fraction of conventional sequencing costs (Supplementary Fig. 8d), thereby providing molecular signatures across a wide range of samples in a single sequencing run. Given these advantages, and the
challenge of understanding the full range of gene products expressed from the human and other genomes, we foresee RNA CaptureSeq as an important new tool with a wide range of research and clinical applications.

Our data clearly demonstrate that the full extent of the human transcriptome dynamically expressed in different cells, tissues and developmental stages is still far from being characterized. Indeed the low expression of many bona fide transcripts imputes that there are significant transcriptomic differences between cells, even those in clonal cell culture, suggesting that each cell has an individual if not unique transcriptomic signature. This in turn challenges the notion that there may be a single, stable transcriptome by which a cell can be characterized, although broad cell types, such as fibroblasts or myoblasts, may show similar patterns. These conclusions converge with recent findings from single-cell transcriptomics and a transcriptional model characterized by rapid-bursting dynamics, and advocate a model of the human transcriptome that embraces highly specific ontogeny and positional identity, dynamism, plasticity and diversity.

SHORT METHODS

Cell Culture
Primary human female fetal lung fibroblasts and human male foot fibroblasts in DMEM supplemented with 10% FBS at 37 °C with 5% CO2 as previously described.

Double Stranded cDNA library preparation
RNA was oligo-dT reverse transcribed with SuperScript III Reverse Transcriptase (Invitrogen), RNaseH digested and second-strand synthesis using DNA polymerase according to the manufacturer’s instructions.

RNA sequencing (by Illumina sequencing)
Illumina paired-end libraries, were constructed from double stranded cDNA preparation using Illumina’s PE Kit according to manufacturers instructions with the following modifications: prescribed Agarose gel excision was performed at 350-300 base pairs to produce libraries with an approximate insert size of 240bp. Illumina RNAseq libraries were used in cluster formation on an Illumina cBot and paired end sequencing was done using the Genome Analyzer IIX. Both cluster formation and 76bp paired-end sequencing were performed using the manufacturers provided protocols.

Array-Capture for (by Illumina sequencing)
Illumina paired-end libraries were constructed from double stranded cDNA preparation as above. Samples were purified and used in pre-hybridization library amplification via the Illumina sequencing adapters. Following amplification, DNA were quantified and evaluated electrophoretically to show a mean library fragment size of 340bp. Amplified libraries were combined with COT-1 DNA, primers, Nimblegen hybridization buffer, and Component A.
The capture array, manufactured by Nimblegen, contained 385,000 features and was overlaid with library mixture and hybridized for 72 hours at 42 degrees C according to manufacturers instructions. Slide washing and sample library elution were performed as previously published \[^{34}\]. Post-hybridization was performed with a modified version of the Illumina PE 1.0 and 2.0 primers. Post hybridization amplification employed the same reaction and cycling parameters as for pre-capture amplification, except for a lower annealing temperature. Following completion of amplification, samples were purified by Qiaquick column (QIAGEN) and electrophoretically evaluated with an Agilent Bioanalyzer 2100. Deep-sequencing was performed as above.

**Read alignment and transcript assembly (by Illumina sequencing)**

Illumina 72bp paired-end sequenced reads from pre-capture RNAseq and post-capture RNA CaptureSeq were aligned and assembled using identical parameters. Illumina*.fastq files were firstly aligned to the human genome (hg19) using TopHat \[^{28}\] with the following non-default parameters; -r 242 --min-isofrac=0.001 -G RefSeq.gtf (downloaded from UCSC hg19 October 2010). Cufflinks\[^{12}\] was employed to assembled transcripts from resultant .sam files according to the following parameters; --min-isofrac=0.001, --min-infract=0.01, -r hg19.fa, --min-frags-per-transfrag = 5.

**Array-Capture (by 454 sequencing)**

454 GS-FLX Titanium Sequencing libraries from c45 and FL fibroblast cDNA were constructed using the 454 LifeSciences GSFLX Titanium Kit according to manufacturers instructions. All single-stranded DNA product from library preparation was utilized as template in a Pre-Hybridization Linker Mediated PCR reaction to ensure that the plurality of the molecules contained adapters on both sides of the putative cDNA inserts. The amplified material was recovered with a Qiaquick column and DNA evaluated electrophoretically by an Agilent Bioanalyzer 2100. Library fragment sizes were found to be between 500-700 bp. Samples were purified and combined with COT-1 DNA, enhancing oligos, NimbleGen hybridization buffer and Component A prior to capture. Hybridization and post-hybridization LMPCR were performed with identical conditions. The resulting post capture enriched sequencing libraries were sequenced on 454 Genome Sequencer FLX System using Titanium chemistry.

**Read alignment and transcript assembly (by 454 sequencing)**

Roche 454 reads were firstly aligned to the human genome (hg19) using Blat (http://www.soe.ucsc.edu/~kent) with the following non-default parameters; minIdentity=90, minScore=100. Resultant *.psl files were converted into *.sam files using bedTools \[^{35}\] and samtools \[^{36}\]. Gaps smaller than 30nt were removed from alignments. The direction of reads spanning putative introns was inferred according to the direction of the canonical splice motifs (GT-AG). Reads spanning introns with non-canonical introns from which direction could not be inferred were discarded and reads not spanning introns were retained as unstranded. Cufflinks\[^{12}\] was employed to assembled transcripts from resultant *.sam files according to the following non-default parameters; --min-isofrac=0.01, --min-infract=0.01, -r hg19.FA , -r --small-anchor-fraction = 0.05, --min-frags-
per-transfrag = 5. These options were chosen given the longer read length and lower read depth of 454 sequencing and our aim to identify minor isoform.

Data Accession
All sequencing data has been submitted to GEO (GSE29041).

ACKNOWLEDGMENTS
The authors would like to thank Manuel Garber for his design of the arrays and constructive contribution to the manuscript; Magdalena Koziol and Kelly Thomas for preparing cDNA and sequencing libraries; Thomas Albert, Todd Arnold, Jason Affourtit, Brian Dessany, Thomas Jarvie, Dawn Green and Tracy Millard provided sequencing support. The authors would like to thank the following funding sources: Human Frontiers Science Program (to T.R.M.); Queensland Government Department of Employment, Economic Development and Innovation Smart Futures Fellowship (to M.E.D.); Australian Research Council/University of Queensland co-sponsored Federation Fellowship (FF0561986; to J.S.M.); Australian National Health and Medical Research Council Australia Fellowship (631668; to J.S.M) and Career Development Award (CDA631542; to M.E.D.); Damon Runyon-Rachleff, Searle, Smith Family Foundation and Richard Merkin Foundation Scholar (to J.L.R); and National Institutes of Health (1DP2OD00667-01; to J.L.R and C.T.).

AUTHOR CONTRIBUTIONS

COMPETING FINANCIAL INTERESTS
REFERENCES


FIGURE LEGENDS

Figure 1. Circle plot illustrating the prevalence and complexity of captured transcripts at genic (left panel) and intergenic (right panel) loci. Successive tracks from outer edge indicate the following features: (1) Genomic position (colored bars indicate different chromosomes and black ticks demarcate 5kb); (2) Previous gene annotations (black bars on green background); (3) Frequency distribution of sequenced read alignments from pre-capture library (green histogram on grey background); (4) Assembled transcript structures from pre-capture library (green bars indicate exons and links indicate splice junctions); (5) Probed regions represented on capture-array (black bars on blue background); (6) Frequency distribution of sequenced read alignments from CaptureSeq library (blue histogram on grey background); and (7) Assembled transcript structures from CaptureSeq library (green bars and links corresponds to exons and splice junctions identified in both pre- and CaptureSeq libraries, blue bars and links correspond to exons and splice junctions exclusively identified in CaptureSeq libraries). Insert shows detail of selected regions.

Figure 2. Resolution of novel p53 isoforms. (a) Genome-browser view of the p53 gene showing previously annotated alternative isoforms (red; α, β, γ). The coverage and relative expression as determined by conventional RNAseq is indicated by upper red histogram. (b) Genome-browser view showing novel alternative splicing (blue; i–iv) identified using RNA CaptureSeq. The relative coverage and expression as determined by RNA CaptureSeq are also indicated by upper histogram (blue with red cap on values exceeding scale). (c) Relative expression of alternative novel p53 isoforms. The annotated (blue) and novel (red) isoforms of p53, along with subsequent modifications to characterized protein domains are indicated in left panel. The relative expression of annotated and novel isoforms is indicated in right panel (error bars indicate upper and lower bound of 95% confidence interval).

Figure 3. Identification of novel exons variants and rare intergenic noncoding RNAs by RNA CaptureSeq. (a) Differential expression across HOXA loci (black bars show gene annotations) between lung (green) and foot (blue) fibroblasts, reflecting the different anatomical origin of each cell line. Coverage and relative abundance by RNA CaptureSeq (histograms) is indicated for each cell line. (b) Relative enrichment of HOXA genes and IncRNAs (1–7) between Foot (F) and Lung (L) fibroblasts as determined by CaptureSeq (dark grey) or qRT-PCR using pre-capture (light grey) or post-capture (mid grey) RNA samples. (c) Genome browser view of HOTAIR showing 6 novel isoforms (i), including fine-scale alternate splicing events (ii; zoom detail) that generate 16 novel isoforms. Relative abundance and coverage in RNAseq (upper blue histogram) and CaptureSeq (upper red histogram) libraries from foot fibroblast cell line indicated. Relative abundance of exon-variants indicated in lower histogram. (d) Cumulative frequency distribution showing codon substitution frequency of full-length transcripts assembled from captured libraries (blue) and coding genes (green) and known noncoding RNAs (red) for reference. (e) Cumulative frequency distribution indicates the normalized expression of full-length novel intergenic ncRNAs (red) relative to subset of genes captured on array (blue; captured) or genes identified by conventional RNAseq (green; all). (f) Cumulative frequency distribution showing the raw sequenced read frequency aligning to captured intergenic transcripts from
both RNAseq (dashed red) and CaptureSeq (blue) and all assembled transcripts from RNAseq (solid red). The large difference in raw alignment frequency between RNAseq and CaptureSeq indicates saturated coverage achieved by CaptureSeq. Pie-chart (right) indicating the proportion of RNAseq reads assigned to assembled transcripts, previous gene annotations, or unassignable reads occurring in intronic or intergenic regions. Histogram (left panel) indicates the proportion of unassigned intronic or intergenic reads 'rescued' by incorporation into rare transcript exons.
FIGURE 1

a Annotated gene loci

b Intergenic loci

Tracks (from outer edge):
- Annotated gene
- RNAseq
- Assembled exon
- Splice junction
- Probed region
- RNAseq
- Novel assembled exon
- Novel splice junction
FIGURE 2

a  
RNAseq

b  
Capture Seq

393aa
354aa
386aa
393aa
369aa

p53

Transactivation
Oligomerisation
DNA-binding
Nuclear local signal
Oligomerisation

Intron Exon

Known:  
Novel:

Expression RPKM (x10^6)

Foot fibroblast  
Lung fibroblast
**FIGURE 3**

**a**

Foot
(F/raw)

Fetal
Lung
(F/ raw)

HOX A1
text

chr7: 27140000 27240000

3,537

Foot
(F/ raw)

Fetal
Lung
(F/ raw)

3,625

HOX A13

**b**

Fold change (log2) between lung (L) and foot (F) fibroblasts

HOXA6
HOXA7
HOXA8
HOX A13

**c**

107

RNAseq
(rare)

236,799

CaptureSeq
(raw)

HOTAIR

**d**

Expression (RPKM)

Cumulative fraction

CSF score

**e**

Fraction

CaptureSeq

qRT-PCR:

Pre-capture

Post-capture

**f**

Fraction

Aligned reads (raw count)

**g**

Distribution of foot fibroblast aligned reads

- Previous gene annotations (8%)
- Introns (5%)
- Intergenic (10%)
- Assembled transcripts (77%)

Fraction 'rescued' within captured regions