nature biotechnology

Targeted RNA sequencing reveals the deep complexity of the human transcriptome

Tim R Mercer[1], Daniel J Gerhardt[2], Marcel E Dinger[1], Joanna Crawford[1], Cole Trapnell[3], Jeffrey A Jeddeloh[2], John S Mattick[1] & John L Rinn[3]

Transcriptomic analyses have revealed an unexpected complexity to the human transcriptome, whose breadth and depth exceeds current RNA sequencing capability[1–4]. Using tiling arrays to target and sequence select portions of the transcriptome, we identify and characterize unannotated transcripts whose rare or transient expression is below the detection limits of conventional sequencing approaches. We use the unprecedented depth of coverage afforded by this technique to reach the deepest limits of the human transcriptome, exposing widespread, regulated and remarkably complex noncoding transcription in intergenic regions, as well as unannotated exons and splicing patterns in even intensively studied protein-coding loci such as *p53* and *HOX*. The data also show that intermittent sequenced reads observed in conventional RNA sequencing data sets, 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 (RNA-Seq) technologies can provide an unbiased profile of the human transcriptome, with techniques of *ab initio* transcript assembly capable of 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[1], 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 noncoding RNAs (lncRNAs) that are intergenic, overlapping or antisense to annotated genes[2,3]. However, despite recent technological advances, we have still not yet reached the limits of the transcriptome nor realized its full scale and complexity, fueling ongoing debate as to the extent to which the genome is transcribed and the biological relevance of transcripts that are expressed at low levels[4–6].

To profile such rare transcriptional events and thereby assess the full depth of the transcriptome, we employed a targeted RNA capture and sequencing strategy, which for brevity we term RNA CaptureSeq, that is similar to previous in-solution capture methods[7] and analogous to exome sequencing approaches[8]. Briefly, RNA CaptureSeq involves the construction of tiling arrays across genomic regions of interest, against which cDNAs are hybridized, eluted and sequenced. Although this ability to isolate and target RNA has been used in genetic analysis for some time[9,10], here we combine this ability with deep-sequencing technology to provide saturating coverage and permit the robust assembly of rare and unannotated transcripts.

To inform the design of arrays and as a comparative reference, conventional RNA-Seq was initially performed on a primary human foot fibroblast cell line[11] using the Illumina GAII platform (**Supplementary Table 1**). *Ab initio* transcript assembly[12] of the resulting ~20.4 million alignable paired-end reads yielded 48,091 multiexon 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 lncRNAs (such as *HOTAIR*[13], *TUG1* and *MEG3*) for inclusion on the array (**Supplementary Fig. 1a** and **Supplementary Tables 2** and **3**). In addition, we also included intergenic regions that exhibited little or no transcriptional activity. In total, 2,265 contiguous regions that together comprise ~0.77 Mb were represented on the array. To validate the array design we first conducted capture sequencing of matched foot fibroblast genomic DNA (**Supplementary Results** and **Supplementary Fig. 1b–d**), confirming the specificity, sensitivity, uniformity and reproducibility of the capture arrays, comparable to previous DNA capture and sequencing studies[14,15].

Targeted RNA capture and sequencing was then carried out on matched foot fibroblast cDNA. To permit direct comparison, we applied the same sequencing and alignment methods as for precapture RNA-Seq libraries, yielding ~25.8 million alignable paired-end reads generated on an Illumina GAII instrument. 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 precapture reads aligned to probed regions (**Supplementary Fig. 2a**). A comparison between RNA-Seq- and CaptureSeq-sequenced libraries showed that the capture protocol did not substantially diminish library diversity or introduce PCR amplification bias (**Supplementary Results** and **Supplementary Fig. 2b**). Given that RNA CaptureSeq achieved a ~380-fold enrichment for alignment coverage across targeted

[1]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. [2]Roche NimbleGen Inc., Research and Development, Madison, Wisconsin, USA. [3]Department of Stem Cell & Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. Correspondence should be addressed to J.A.J. (jeffrey.jeddeloh@roche.com), J.S.M. (j.mattick@uq.edu.au) or J.L.R. (john_rinn@harvard.edu).

Received 5 August; accepted 4 October; published online 13 November 2011; doi:10.1038/nbt.2024

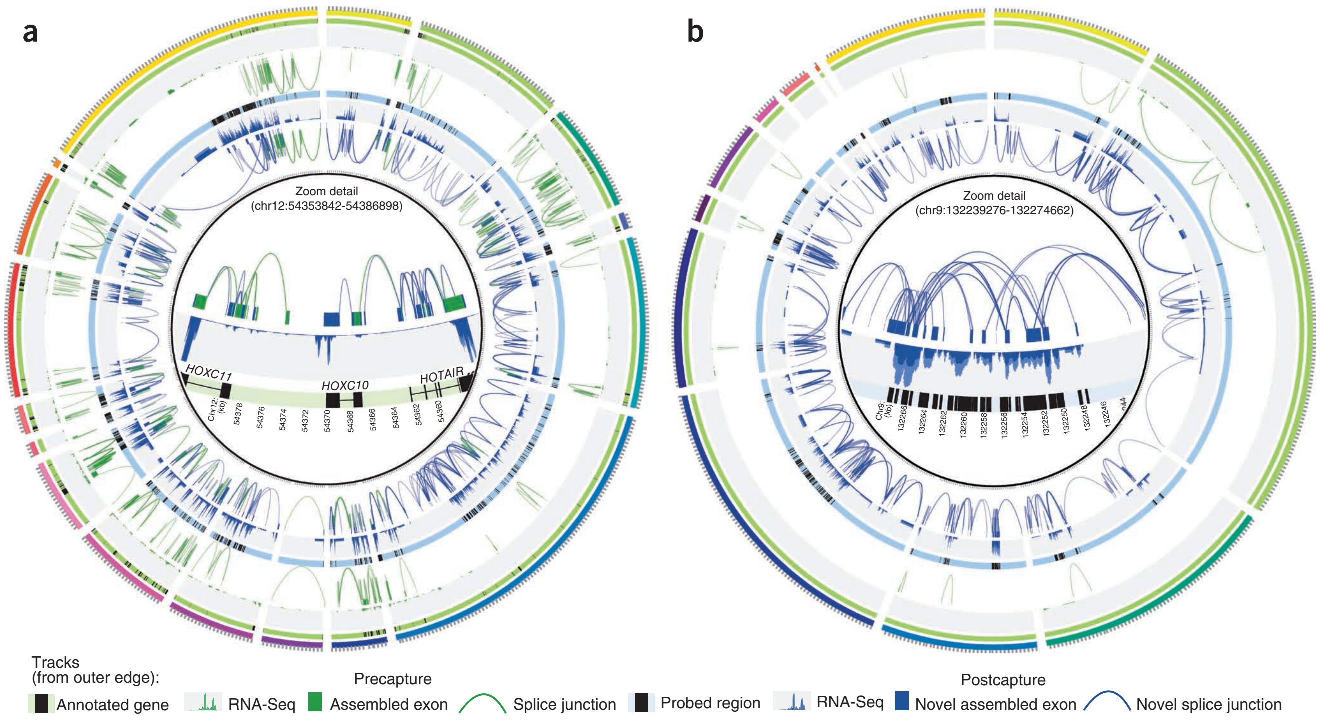


Figure 1 Circle plots illustrating the prevalence and complexity of captured transcripts at genic (**a**) and intergenic (**b**) loci. Successive tracks from outer edge indicate the following features: (i) genomic position (colored bars indicate different chromosomes and black ticks demarcate 5 kb); (ii) previous gene annotations (black bars on green background); (iii) frequency distribution of sequenced read alignments from precapture library (green histogram on gray background); (iv) assembled transcript structures from precapture library (green bars indicate exons and links indicate splice junctions); (v) probed regions represented on capture array (black bars on blue background); (vi) frequency distribution of sequenced read alignments from CaptureSeq library (blue histogram on gray background); and (vii) assembled transcript structures from CaptureSeq library (green bars and links correspond 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). Inset shows detail of selected regions. Plot generated using Circos software (http://www.circos.ca/).

regions of the transcriptome, we extrapolate that ~10 billion aligned sequenced reads from a single sample by conventional RNA-Seq 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 *ab initio* transcript assembly (**Fig. 1**), initially focusing on regions containing well-annotated protein-coding genes. We reconstructed all genes assembled within precapture RNA-Seq data with a similar uniformity of transcript coverage (100% of transcript chains reconstructed; **Fig. 1a**, **Supplementary Fig. 2d** and **Supplementary Data 2**). We identified an additional 204 unannotated 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[16]. Indeed, many of the newly identified exons were entirely undetected within our initial RNA-Seq libraries (24.7% undetected with a further 10.4% only detected by a single read)[17]. For example, previously three splicing variations generating up to nine alternative isoforms, each with alternate functional consequences, have been described for the *p53* gene[18] (**Fig. 2a**). By RNA CaptureSeq we identified additional alternative isoforms of *p53*, whose unannotated exon junctions were subsequently validated by RT-PCR and sequencing (**Supplementary Table 4** and **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[19]. As a class, the newly identified isoforms exhibit weaker expression (mean 2.4-fold decrease) (**Fig. 2b,c**) and conservation (mean 1.8-fold decrease) relative to previously annotated isoforms, but a similarly stringent enrichment for canonical splice junctions (**Supplementary Fig. 4a–g**). A subset of these rare isoforms also has limited coding potential, representing noncoding variants of the dominant mRNA transcript[20]. Lastly, we also resolved an additional 163 neighboring and antisense lncRNAs around protein-coding genes[21].

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 a 454 GSFLX Titanium instrument, 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 lncRNAs (**Supplementary Data 3**). This approach also revealed that, like mRNAs, alternative splicing of lncRNAs can modulate the inclusion or exclusion of specific functional domains. For example, the lncRNA *HOTAIR* exhibited an alternative splice site that eliminates the polycomb repressor complex binding domains[22,23], as well as small exon length variations supported by canonical intronic polypyrimidine tracts and splice junctions (**Fig. 3a**)[22,23]. 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[24].

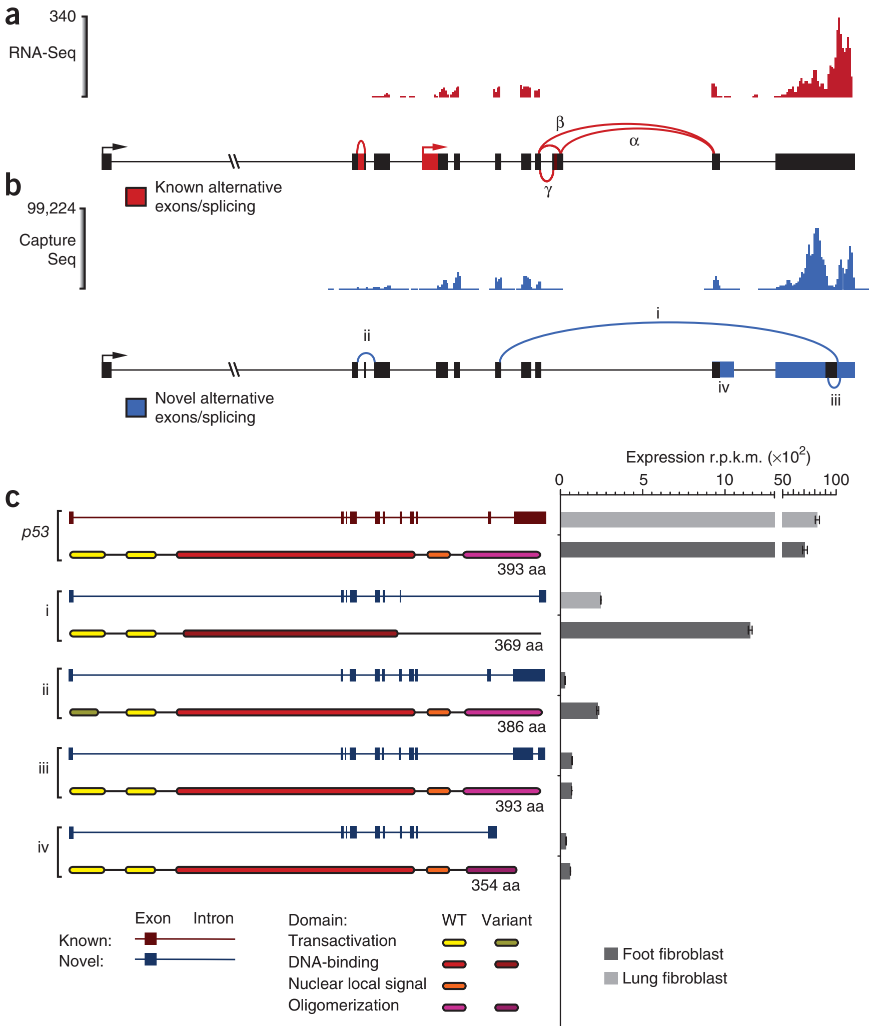


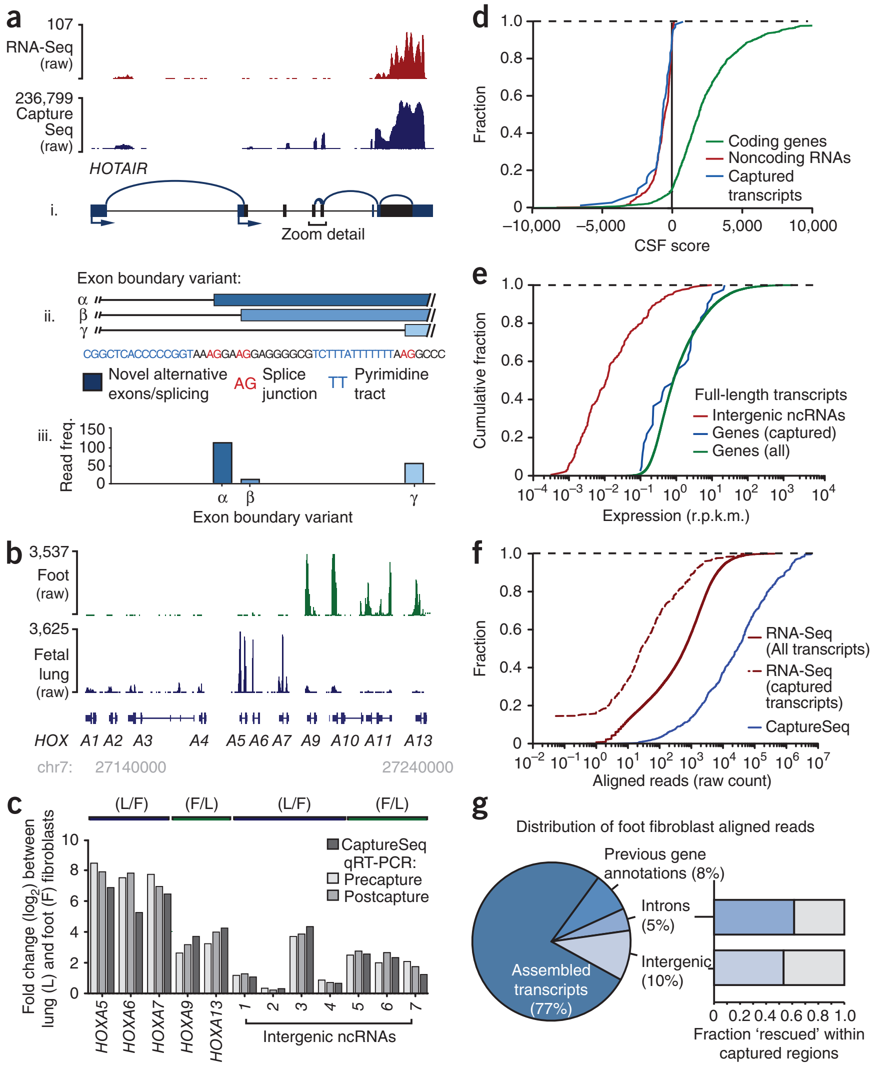
Figure 2 Resolution of unannotated p53 isoforms. (**a**) Genome-browser view of the p53 gene. The coverage and relative expression as determined by conventional RNA-Seq is indicated by upper red histogram. (**b**) Genome-browser view showing unannotated alternative splicing (blue; i–iv) identified using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

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 first confirmed the reproducibility between two CaptureSeq technical replicates by quantitative reverse transcriptase (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 after capture (**Supplementary Results** and **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 fibroblast cells, which show a distinct gene expression program consistent with their alternative location within the body (**Supplementary Table 1**)[11]. We performed sequencing and assembly using the 454 platform as before, assembling in total 430 multiexon transcripts (**Supplementary Data 5**) captured in association with genic probed regions. In comparison to foot fibroblasts, we find 37% of captured genes undergo significant (P < .05) differential expression (**Supplementary Fig. 6a**)[25]. 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 (**Fig. 3b**). 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 after capture (**Fig. 3c**). We observed a close correlation between differential expression profiles for *HOX* before and after capture that was additionally concordant with estimates of gene abundance obtained from CaptureSeq (**Fig. 3c**, **Supplementary Results** and **Supplementary Fig. 6b**). In addition, we confirmed by qRT-PCR the differential expression between foot and lung fibroblasts of six intergenic transcripts expressed at low levels. Taken together, these results indicate that after both phases of the CaptureSeq approach, capture and sequencing, are completed, 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 fueled recent controversy as to whether they represent the low-frequency sampling of authentic transcripts, biological noise from spurious nascent transcription or technical noise from sequencing and alignment[4–6]. Having established the fidelity of the RNA CaptureSeq approach, we applied its sensitivity to characterize these rare transcriptional events within intergenic regions and thereby help resolve this ongoing debate. To do this, we included numerous intergenic regions for interrogation within our array which, despite overlapping active chromatin domains (marked by H3K4me3 and H3K36me3)[25], showed little or no evidence of transcription according to publicly available transcriptomic resources or our own initial precapture RNA-Seq analysis (**Fig. 1b** and **Supplementary Fig. 6c**).

We found aligned captured reads that covered almost all intergenic probed bases (98.1%), similar 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 twofold. First, this filter removed the potential for genomic DNA contamination (we observed that <1.3% of sequenced reads showed evidence of artifactual spliced alignments within our corresponding control capture using genomic DNA); and, second, it omitted the potential for 'spurious' transcriptional noise because we reasoned that 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

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



of splicing (**Supplementary Fig. 7b**). In total, we captured 798 splice junctions[26] within intergenic probed regions, of which 95.7% were not identified in precaptured libraries or preexisting gene annotations. Despite being unreported previously, 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 multiexonic captured transcripts (**Fig. 1b** and **Supplementary Data 2**). The full length of 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 of almost all tested (13 of 15) assembled intergenic transcripts (**Supplementary Fig. 3b**). Although the captured intergenic transcripts exhibit lower evolutionary conservation than protein-coding sequences (as evidenced by coverage of phastCons elements; see Online Methods), they exhibit a similar level of conservation to that of annotated functional lncRNAs, and are more conserved than intronic or surrounding intergenic sequences (**Supplementary Fig. 7c**). A range of metrics, including the presence, size and structure of ORFs, homology of predicted ORFs to known proteins, and synonymous-to-nonsynonymous nucleotide substitution rate confidently ascribed the majority (92.3%) of these transcripts as noncoding RNAs (**Fig. 3d** and **Supplementary Fig. 7d**).

To contextualize the rarity of captured intergenic transcripts in relation to the whole human transcriptome, we first normalized expression profiles between conventional RNA-Seq and RNA CaptureSeq libraries according to shared genes (**Supplementary Fig. 5i**). Captured lncRNAs exhibited a mean expression of only 0.011 reads per kilobase per million reads, 463-fold less than the median gene expression within fibroblasts (**Fig. 3e**). We performed quantitative RT-PCR using our precapture RNA sample to provide an informed estimate of lncRNA transcript copy number. Assuming an average human fibroblast cell contains ~300 fg of mRNA per cell[27], we estimate that the lncRNAs we discovered were present at an average of ~0.0006 transcripts per cell, indicating expression in only a small subpopulation of the cells sampled. By comparison, we calculate *HOXA* to be present at an average ~0.13 transcripts per cell, consistent with previous estimates[27].

Given that 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 fibroblast transcriptome. Within our initial (precapture) RNA-Seq we found only minimal and intermittent coverage of the transcripts expressed at low levels, which is indicative of low-frequency sampling and nonsaturating coverage. Indeed, only 31.3% of captured intergenic transcripts were even detected in precapture RNA-Seq libraries, with a further 9.7% detected by a single alignment. By comparison, even transcripts expressed at low levels are represented by large numbers of aligned reads in the RNA CaptureSeq data set, 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; **Fig. 3f** and **Supplementary Fig. 8a**,**b**).

During precapture RNA-Seq, we found a substantial component (14.8%) of alignable reads that could not be assigned into assembled or previously annotated gene models (**Fig. 3g**). These unassigned reads have been previously thought to represent biological or technical artifacts because they appear to have characteristics of random sampling from a low-level background[6]. We first considered unassigned reads aligning to intronic regions that have been previously dismissed as collateral from splicing by-products[6]. We found a significant overlap between unassigned intronic reads from the precapture RNA-Seq and newly identified isoforms (7.02-fold enrichment; two-tailed $P < 0.0001$ χ^2 test expecting random distribution of read alignments throughout introns; **Fig. 3g**), thereby rescuing 61.5% of unassigned reads aligning within captured transcript introns. We further validated the existence of 4 (of 4) of these unannotated 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 RNA-Seq data set 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, this suggests a sizeable expansion to the borders of the human transcriptome and anticipates a scale of transcriptional complexity that surpasses even previous reports[1].

In this context of an expanded transcriptome, the RNA CaptureSeq approach provides considerable value because it allows one to focus on and comprehensively interrogate 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 genes[28] so as to identify all gene products produced 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 approach with a wide range of research and clinical applications.

Our data strongly suggest 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 implies that there are substantial 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, may show similar patterns. These conclusions converge with recent findings from single-cell transcriptomics[29] and a transcriptional model characterized by rapid-bursting dynamics[30], and advocate a model of the human transcriptome that embraces highly specific ontogeny and positional identity, dynamism, plasticity and diversity.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. All sequencing data have been submitted to GEO (GSE29041).

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

The authors would like to thank M. Garber for his design of the arrays and constructive contribution to the manuscript; M. Koziol and K. Thomas for preparing cDNA and sequencing libraries; T. Albert, T. Arnold, J. Affourtit, B. Dessany, T. Jarvie, D. Green and T. 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 US National Institutes of Health (1DP2OD00667-01; to J.L.R. and C.T.).

AUTHOR CONTRIBUTIONS

T.R.M., J.A.J., J.S.M. and J.L.R. designed the experiments. D.J.G. performed array capture, quality assessments and supported the sequencing teams. J.C. performed RT-PCR. M.E.D., T.R.M. and C.T. performed alignment, transcript assembly and analysis. T.R.M., M.E.D., J.A.J., J.S.M. and J.L.R. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/nbt/index.html.

Published online at http://www.nature.com/nbt/index.html.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
2. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
3. Katayama, S. *et al.* Antisense transcription in the mammalian transcriptome. *Science* **309**, 1564–1566 (2005).
4. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Response to "the reality of pervasive transcription". *PLoS Biol.* **9**, e1001102 (2011).
5. Clark, M.B. *et al.* The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625 (2011).
6. van Bakel, H., Nislow, C., Blencowe, B.J. & Hughes, T.R. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).
7. Levin, J.Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* **10**, R115 (2009).
8. Teer, J.K. *et al.* Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. *Genome Res.* **20**, 1420–1431 (2010).
9. Yehle, C.O. *et al.* A solution hybridization assay for ribosomal RNA from bacteria using biotinylated DNA probes and enzyme-labeled antibody to DNA:RNA. *Mol. Cell. Probes* **1**, 177–193 (1987).
10. Crider-Miller, S.J. *et al.* Novel transcribed sequences within the BWS/WT2 region in 11p15.5: tissue-specific expression correlates with cancer type. *Genomics* **46**, 355–363 (1997).
11. Rinn, J.L., Bondre, C., Gladstone, H.B., Brown, P.O. & Chang, H.Y. Anatomic demarcation by positional variation in fibroblast gene expression programs. *PLoS Genet.* **2**, e119 (2006).

12. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
13. Rinn, J.L. *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
14. Li, Y. *et al.* Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.* **42**, 969–972 (2010).
15. Ng, S.B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
16. Kapranov, P. *et al.* Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **15**, 987–997 (2005).
17. Pan, Q., Shai, O., Lee, L.J., Frey, B.J. & Blencowe, B.J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
18. Khoury, M.P. & Bourdon, J.C. The isoforms of the p53 protein. *Cold Spring Harb. Perspect. Biol.* **2**, a000927 (2010).
19. Olivares-Illana, V. & Fahraeus, R. p53 isoforms gain functions. *Oncogene* **29**, 5113–5119 (2010).
20. Kloc, M., Foreman, V. & Reddy, S.A. Binary function of mRNA. *Biochimie* **93**, 1955–1961 (2011).
21. Mercer, T.R., Dinger, M.E. & Mattick, J.S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).
22. Tsai, M.C. *et al.* Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693 (2010).
23. Hiller, M. & Platzer, M. Widespread and subtle: alternative splicing at short-distance tandem sites. *Trends Genet.* **24**, 246–255 (2008).
24. Schatz, M.C., Delcher, A.L. & Salzberg, S.L. Assembly of large genomes using second-generation sequencing. *Genome Res.* **20**, 1165–1173 (2010).
25. Khalil, A.M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. USA* **106**, 11667–11672 (2009).
26. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
27. Carter, M.G. *et al.* Transcript copy number estimation using a mouse whole-genome oligonucleotide microarray. *Genome Biol.* **6**, R61 (2005).
28. Hindorff, L.A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **106**, 9362–9367 (2009).
29. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
30. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622–634 (2011).

ONLINE 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[11]. Total RNA was purified from each cell culture using TRIzol according to the manufacturer's instructions (Invitrogen).

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

Custom microarray design. Targeted regions were selected from annotated protein coding genes and uncharacterized human intergenic regions exhibiting H3K4/H3K36 domains[25] according to their transcriptional status as determined by RNA-Seq (**Supplementary Table 1**). We employed the Titanium Optimized Sequence Capture 385K Array, designed by Nimblegen, for RNA capture. Array design and probe selection for tiling designs was conducted by Nimblegen using window-based rank selection, retaining probes that received the highest score as determined from a combination of frequency, T_m and uniqueness information. A detailed description of the strategies by which probes are appraised can be found in *Technical Note: Roche Nimblegen Probe Design Fundamentals* (http://www.nimblegen.com/products/lit/probe_design_2008_06_04.pdf). This includes the filtering of probes with variable T_m and repetitive sequences. In addition to the WindowMasker program[31] employed by Nimblegen during array design, we also omitted any sequences overlapping RepeatMasker annotated elements from our design.

Capture library preparation and prehybridization amplification (for 454 sequencing). The 454 GS-FLX Titanium Sequencing library was constructed using the 454 LifeSciences (454 hereafter) GSFLX Titanium Kit as described in the user's guide. All of the single-stranded DNA product from this library preparation (e.g., sst Library) was used as a template in a PreHybridization Linker Mediated PCR (LMPCR) reaction to ensure that the plurality of the molecules contained adaptors on both sides of the putative cDNA inserts. The LMPCR conditions consisted of five reactions each containing 5 µl 100058 Platinum High Fidelity Polymerase Buffer from Invitrogen, 2.5 µl $MgSO_4$, 1 µl 25nM dNTP's from Epicentre, 1 µl of 25 µM Primer A 5′-CCATCTCATCCCTGCGTGTC, 1 µl of 25 µM Primer B 5′-CCTATCCCCTGTGTGCCTTG and 0.4 µl Platinum High Fidelity Polymerase. DNA in equal amounts was apportioned for each of the five reactions and water added to 50 µl. The master mix was pipetted into 0.2 ml strip tubes and then placed into a thermal cycler. The reactions were then subjected to 94 °C for 4 min followed by 8 cycles of the following pattern: 94 °C for 30 s, 1 min at 58 °C and 1.5 min at 68 °C. The last step was an extension at 72 °C for 5 min. The reactions were then kept at 4 °C until further processing. The amplified material was recovered with a Qiagen Qiaquick column according to the manufacturer's instructions except the DNA were eluted in 50 µl water instead of the elution buffer. The DNA was quantified using the NanoDrop-1000 and the library was evaluated electrophoretically with an Agilent Bioanalyzer 2100 using a DNA 7500 chip. The library fragment sizes were found to be between 500–700 bp.

Optimized cDNA sequence capture array processing (for 454 sequencing). Prior to array hybridization the following components were added to a 1.5 ml tube: 3 µg of library material, 0.65 µl of 1,000 µM Enhancing Oligo A 5′-CCATCTCATCCCTGCGTGTCTCCGACTCAG/3ddc/ and 0.65 µl of 1000 µM Enhancing Oligo B 5′-CCTATCCCCTGTGTGCCTTGGCAGTCTCAG/3ddc/, and 100 µg of CoT-1 DNA, Invitrogen. Samples were dried down by puncturing a hole in the 1.5 ml tube cap with a 20 gauge needle and processing in an Eppendorf Vacufuge set to 60 °C for 40 min. To each dried sample 4.8 µl of water was added and it was then placed in a heating block at 70 °C for 10 min to resuspend. Samples were subjected to vigorous vortex mixing for 30 s and centrifuged to recollect any dispersed sample. To each sample tube 8 µl NimbleGen SC Hybridization Buffer and 3.2 µl NimbleGen Hybridization component A was added, and the sample was vortexed for 30 s, centrifuged and placed in a heating block at 95 °C s C for 10 min. The samples were again mixed for 10 s, spun down and placed in a Roche NimbleGen Hybridization System at 42 °C until ready for hybridization. The capture array contained 385,000 features and was overlaid with an X1 mixer according to manufacturer's instructions, and 16 µl of the hybridization mixture (library, C0t-1, enhancing oligos, SC Hybridization Buffer and SC Component A) was pipetted onto the array field. The loading and vent holes were covered with port seals, and each array sample was hybridized for 72 h at 42 °C on Hybridization Station setting "B." Slide washing and sample library elution were done as previously described[32].

Posthybridization LMPCR (for 454 sequencing). Posthybridization amplification (e.g., LMPCR via 454 adapters) consisted of ten reactions for each sample using the same enzyme and primer concentrations as the precapture amplification. Posthybridization amplification consisted of 16 cycles of PCR with identical cycling conditions as used in the prehybridization LMPCR. Following the completion of the amplification reaction, the samples were purified using a 2 Qiagen Qiaquick column according to the manufacturer's recommended protocol and eluate from each column was combined into one tube. DNA was quantified spectrophotometrically using the NanoDrop-1000, and electrophoretically evaluated with an Agilent Bioanalyzer 2100 using a DNA 7500 chip. The resulting postcapture enriched sequencing libraries were sequenced on 454's Genome Sequencer FLX System using Titanium chemistry.

Read alignment and transcript assembly (by 454 sequencing). Roche 454 reads were first aligned to the human genome (hg19) using Blat (http://users.soe.ucsc.edu/~kent/) with the following nondefault parameters; minIdentity = 90, minScore = 100. Highest scoring alignments were selected and resultant *.psl files were converted into *.sam files using bedTools[33] and SAMtools[34]. Gaps smaller than 30 nt were removed from alignments. The direction of reads spanning putative alignments was inferred according to the direction of the canonical splice motifs (GT-AG). Reads spanning introns with noncanonical introns, from which direction could not be inferred, were discarded. Reads not spanning intron were retained as unstranded. Cufflinks[12] was employed to assembled transcripts from resultant *.sam files according to the following nondefault parameters: --min-isoform-fraction = 0.01,--min-intron-fraction = 0.01, -r hg19.FA, --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 variants. Cuffdiff[12] was employed to determine differences in transcript abundance between foot and lung fibroblast libraries using foot transcript annotations as reference. Cuffcompare[12] was employed to compare structural differences between foot and lung fibroblast libraries using foot transcript annotations as reference.

cDNA capture library preparation and prehybridization amplification (for Illumina sequencing). Illumina paired-end libraries were constructed from the same double-stranded cDNA prep using Illumina's PE Kit with the following modifications. The prescribed agarose gel excision was done at 350–300 base pairs to produce libraries with an approximate insert size of 340 bp. DNA was purified from the agarose using a Qiagen, Qiaquick column and eluted in 30 µl of water. The entire recovery product was used as template in the prehybridization library amplification by the Illumina sequencing adapters (LMPCR). Prehybridization LMPCR consisted of one reaction containing 50 µl Phusion High Fidelity PCR Master Mix (New England BioLabs), 2 µM of primers Illumina PE 1.0: 5′-AATGATACGGCGACCACCGAGATCTACACTCTT TCCCTACACGACGCTCTT CCG ATC* T and 2.0: 5′-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCAT TCCTGCTGAACCGCTCTTCCGATC* T (asterisk denotes phosphorothioate bond), 30 µl DNA, and water up to 100 µl. PCR cycling conditions were as follows: 98 °C for 30 s, followed by 8 cycles of 98 °C for 10 s, 65 °C for 30 s, and 72 °C for 30 s. The last step was an extension at 72 °C for 5 min. The reaction was then kept at 4 °C until further processing. The amplified material was recovered with a Qiagen Qiaquick column according to the manufacturer's instructions, except the DNA was eluted in 50 µl water. The DNA was quantified using the NanoDrop-1000 and the library was evaluated electrophoretically with an Agilent Bioanalyzer 2100 using a DNA1000 chip. The mean library fragment size was found to be 328 bp.

Capture array processing (for Illumina sequencing). Before array hybridization the following components were added to a 1.5 ml tube: 3 µg of library material, 6.5 μl of 100 μM Illumina primer PE 1.0 and PE 2.0 at, and 100 μl of CoT-1 DNA (Invitrogen). Samples were dried down by puncturing a hole in the 1.5 ml tube cap with a 20 gauge needle and processing in an Eppendorf Vacufuge set to 60 °C for 20 min. To each dried sample 4.8 μl of water was added and, it was then placed in a heating block at 70 °C for 10 min to resuspend sample. Samples were subjected to vigorous vortex mixing for 30 s and centrifuged to recollect any dispersed sample. To each sample tube 8 μl NimbleGen SC Hybridization Buffer and 3.2 μl NimbleGen Hybridization component A was added, and the sample was vortexed for 30 s, centrifuged and placed in a heating block at 95 °C for 10 min. The samples were again mixed for 10 s, spun down and placed in a Roche NimbleGen Hybridization System at 42 °C until ready for hybridization. The capture array contained 385,000 features and was overlayed with an X1 mixer according to manufacturer's instructions, and 16 μl of the hybridization mixture (library, C0t-1, enhancing oligos, SC Hybridization Buffer, and SC Component A was pipetted onto the array field. The loading and vent holes were covered with port seals, and each array sample was hybridized for 72 h at 42 °C on Hybridization Station setting "B." Slide washing and sample library elution were done as previously described[34].

Posthybridization LMPCR (for Illumina sequencing). Posthybridization amplification (e.g., LMPCR via Illumina adaptors) consisted of two reactions for each sample using the same enzyme and primer concentrations as the precapture amplification, but a modified version of the Illumina PE 1.0 and 2.0 primers were employed: forward primer 5′-AATGATACGGCGACCACCGAGA and reverse primer 5′-CAAGCAGAAGACGGCATACGAG. Posthybridization amplification consisted of 16 cycles of PCR with identical cycling conditions as used in the prehybridization LMPCR; with one exception the annealing temperature was lowered from 65 °C to 60 °C. Following the completion of the amplification reaction, the samples were purified using a Qiagen Qiaquick column, using the manufacturer's recommended protocol, and the DNA was quantified spectrophotometrically using the NanoDrop-1000, and electrophoretically evaluated with an Agilent Bioanalyzer 2100 using a DNA1000 chip. The resulting postcapture-enriched sequencing libraries were diluted to 10 nM and used in cluster formation on an Illumina cBot, and paired-end sequencing was done using the Genome Analyzer II_X. Both cluster formation and 76 bp paired-end sequencing were done using the manufacturer's provided protocols.

Read alignment and transcript assembly (by Illumina sequencing). Illumina 76 bp paired-end sequenced reads from precapture RNA-Seq and postcapture RNA CaptureSeq were aligned and assembled using identical parameters. Illumina*.fastq files were first aligned to the human genome (hg19) using TopHat[26] with the following nondefault parameters: -r 242 --min-isoform-fraction 0.01 -G RefSeq.gtf (downloaded from UCSC hg19 October 2010). Cufflinks[12] was employed to assemble transcripts from resultant *.sam files according to the following parameters: --min-isoform-fraction = 0.01,--min-intron-fraction = 0.01, -r hg19.fa,--min-frags-per-transfrag = 5. Cuffdiff was employed to determine differences in transcript abundance between precapture and postcapture libraries using precapture annotations as reference. Cuffcompare was employed to compare structural differences between precapture and postcapture libraries and identify isoforms using precapture annotations as reference.

Transcript characterization. PhastCons annotations (Vertebrate Conserved Elements, 28-Way Multiz Alignment;[35]) were retrieved from the UCSC (October 2010; http://hgdownload.cse.ucsc.edu/downloads.html) and intersected with transcript annotations using overlapSelect (http://users.soe.ucsc.edu/~kent/) to determine fractional coverage.

Coding potential of transcripts was assessed by two approaches. First, transcript sequences were submitted to the Coding Potential Calculator[36] (CPS) that scores each transcript according to the potential to encode a protein according to a range of different metrics including the presence, size and integrity of an ORF, matches to known protein domains and the conservation of these matches with a single frame, transcript coverage and structure with reference to predicted ORF. Second, we performed a codon substitution frequency analysis[37] (CSF) to determine synonymous to nonsynonymous substitutions within transcripts and thereby provide evidence of selective evolutionary pressure acting on transcript sequences to preserve putative ORF. Input *.maf files were retrieved from UCSC (October 2010; http://hgdownload.cse.ucsc.edu/downloads.html). Both CPC and CSF were installed and implemented locally using the UniRef90 database (November 2010; (ref. 38)) for BLASTX searches.

Gene expression by RT-PCR. For nonquantitative expression analysis, 1 ng postcapture cDNA was PCR amplified for 35 cycles and products visualized after electrophoresis in a 2.5% agarose gel (primer sequences are documented in **Supplementary Table 4**). Quantitative PCR reactions were done with a final 0.1 ng/μl concentration of cDNA using SYBRTM green PCR master mix (Applied Biosystems). Amplification and cycling conditions were as recommended by the manufacturer. The standard curve method, using a 1 pg to 10 ng serial dilution, was used for absolute quantization of transcript expression.

31. Morgulis, A., Gertz, E.M., Schaffer, A.A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
32. Fu, Y. *et al.* Repeat subtraction-mediated sequence capture from a complex genome. *Plant J.* **62**, 898–909 (2010).
33. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
34. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
35. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
36. Kong, L. *et al.* CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **35**, W345 (2007).
37. Lin, M.F. *et al.* Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Res.* **17**, 1823–1836 (2007).
38. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C.H. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**, 1282–1288 (2007).

doi:10.1038/nbt.2024