**SUPPLEMENTARY MATERIAL FOR:**

**Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing**

Ryan D. Morin, Matthew Bainbridge, Anthony Fejes, Martin Hirst, Martin Krzywinski, Trevor J. Pugh, Helen McDonald, Richard Varhol, Steven J.M. Jones, and Marco A. Marra  
Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada

*BioTechniques* 45:81-94 (July 2008)

---

**Supplementary Figure S1.** Comparing the distribution of reads across the lengths of transcripts for two libraries (preparation differed only in sonication time).
Supplementary Figure S2. Demonstrating the reproducibility in exon quantitation between two libraries.

Supplementary Figure S3. Allelic expression ratio for known and novel single-base changes.
Supplementary Figure S4. Enrichment of noncoding RNA (ncRNA) genes (red) over protein-coding genes (black) in polyA- fraction of RNA (y-axis).
SUPPLEMENTARY METHODS, RESULTS, AND DISCUSSION

Computing the Mappability of Genomic Bases

Considering a given genomic position, \( i \), let \( s_i \) be an \( n \)-mer subsequence that starts at this position. For all of our calculations \( n = 31 \). Let \( P_i \) be the set of genomic positions to which the \( n \)-mer \( s_i \) maps. If the \( n \)-mer is unique, its position set contains a single entry, \( P_i = \{ i \} \). If the \( n \)-mer appears more than once in the genome it cannot be unambiguously mapped and \( |P_i| > 1 \).

Let \( u_i \) be 1 if \( P_i = \{ i \} \) (i.e., \( n \)-mer at \( i \) is unique) and \( u_i = 0 \) otherwise.

Let \( Q_i \) be the set of all genomic positions that neighbor on position \( i \) and start an \( n \)-mer that overlaps with position \( i \). Mappability, \( m_i \), is defined for each genomic position, \( i \), as the fraction

\[
m_i = \frac{\sum_{j \in Q_i} u_j}{n},
\]

which gives the number of unambiguously mappable \( n \)-mers that overlap position \( i \). If each \( n \)-mer that overlaps with position \( i \) is unique in the genome (i.e., appears only once), then the mappability is 1.

Analysis of polyA- RNA for Novel Exons

The sequences generated from the polyA- fraction of RNA should largely represent noncoding RNA (ncRNA) species, most of which we expected to be known. We searched the peaks generated from this library for evidence of novel ncRNA genes in the genome by eliminating those overlapping known protein coding or RNA genes or regions identified by RepeatMasker (www.repeatmasker.org) derived from rRNA or tRNA genes. After removing small (<32 nt) peaks, we noted that 7862 remained and had an average height of 47 (149 peaks had heights >30). This was considerably lower than the peaks corresponding to known ncRNAs (1165, mean height). Of these peaks, only 1077 (13.6%) were highly conserved based on their phastCons score with the remainder showing minimal conservation. The physical distribution of the 149 peaks across the genome is shown in Supplementary Figure S5.

The plot in Supplementary Figure S5 shows a region of apparent enrichment for novel polyA- peaks on chromosomes 8 and 17 (spikes in red line). These two regions are of particular interest because they suggest enhanced transcriptional activity in these regions yielding unpolyadenylated transcripts.

The region on chromosome 17 (Supplementary Figure S6) appears to represent an unannotated gene related to (or derived from) the human SCD gene. The peak structure, which suggests five or six distinct exons, does not correspond to the Ensembl (red), RefSeq (black), or GenScan (yellow) models.

The second region (Supplementary Figure S7), residing on chromosome 8, has a different structure and explanation. This locus represents a ~20 kb region nearly consistently covered by peaks. This region lies within a gene desert with the two nearest genes being a pseudogene of \( POU5F1 \) and the \( MYC \) proto-oncogene. This region of chromosome 8 (128,290,875-128,312,000) lies within the known fragile site FRA8C. This is a recurrent integration site for the human papillo-

REFERENCES

Supplementary Figure S5. Distribution of polyA- peaks across all chromosomes. This karyogram (produced by the KaryoView program at Ensembl) shows the relative density of polyA- peaks with a vertical red line. The two regions with the highest concentration of peaks are on 8q and 17p.
Supplementary Figure S6. Novel gene model supported by polyA- peaks. Up to six exons in this region of chromosome 17 (20,628,400–20,631,500) are revealed by polyA- peaks, and these show partial overlap with gene models from both Genscan and Ensembl. Both the primary and in silico translated sequence of this region show high similarity to the human gene SCD.
Supplementary Figure S7. Widespread transcription near HPV18 integration site. This region of chromosome 8 (128,290,875–128,312,000) showed a large enrichment for peaks in both the polyA- and polyA+ fractions. This is a known fragile site and a preferred integration site for the HPV18 genome. The clone mapped to this region (BC106081) was obtained from an unknown cervical cancer cell line. The clone is flagged as “chimeric” and includes a portion of the HPV genome, suggesting it results from similar transcription of this region in that cell line.