A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome

Marc Sultan,[1]* Marcel H. Schulz,[2,3]* Hugues Richard,[2]* Alon Magen,[1] Andreas Klingenhoff,[4] Matthias Scherf,[4] Martin Seifert,[4] Tatjana Borodina,[1] Aleksey Soldatov,[1] Dmitri Parkhomchuk,[1] Dominic Schmidt,[1] Sean O'Keeffe,[2] Stefan Haas,[2] Martin Vingron,[2] Hans Lehrach,[1] Marie-Laure Yaspo[1]†

The functional complexity of the human transcriptome is not yet fully elucidated. We report a high-throughput sequence of the human transcriptome from a human embryonic kidney and a B cell line. We used shotgun sequencing of transcripts to generate randomly distributed reads. Of these, 50% mapped to unique genomic locations, of which 80% corresponded to known exons. We found that 66% of the polyadenylated transcriptome mapped to known genes and 34% to nonannotated genomic regions. On the basis of known transcripts, RNA-Seq can detect 25% more genes than can microarrays. A global survey of messenger RNA splicing events identified 94,241 splice junctions (4096 of which were previously unidentified) and showed that exon skipping is the most prevalent form of alternative splicing.

Global analysis of gene expression has mostly relied on RNA hybridization on high-density arrays (*1*–*3*), allowing the profiling of many tissues (*4*, *5*) but detecting only specific sequences. Whole-genome tiling arrays theoretically allow the capture of much of the complexity of the transcriptome (*6*, *7*), but they ignore splice-junction information and are associated with high costs and difficulties in data analysis. Arrays that specifically detect alternative splicing (AS) events (*8*, *9*) have been hampered by issues of completeness and specificity.

Digital transcript-counting approaches overcome many of the inherent limitations of array-based systems and bypass problems inherent to analog measurements, including complex normalization procedures and limitations in detecting low-abundance transcripts. However, the expressed sequence tag (EST) approach, providing partial sequences of individual cDNA clones, is sensitive to cloning biases and has high costs. Serial analysis of gene expression (*10*) and massively parallel signature sequencing (*11*) are also costly and cannot be used for splicing events.

The potential of RNA-Seq (short-read high-throughput sequencing) was first demonstrated by the polony multiplex analysis of gene expression, allowing the detection of 0.3 RNA copies per cell (*12*). Illumina-based RNA-Seq technology has recently been applied to yeast and *Arabidopsis thaliana* (*13*–*15*), providing transcriptome surveys at single-nucleotide resolution.

We present here a snapshot of the human transcriptome at base-pair resolution via RNA-Seq (*16*). Briefly, poly(A) RNA was extracted from human embryonic kidney (HEK) 293T and Ramos B cells and used to generate double-stranded cDNA using random hexamers as primers. The double-stranded DNA was sheared by sonication for preparing the sequencing libraries according to the Illumina protocol (*16*). Illumina deep sequencing was used to generate 27–base pair (bp) reads from replicate samples for each cell line. Reads were mapped to the human genome (hg18, National Center of Biotechnology Information build 36.1) using the Eland software, allowing up to two mismatches (*16*). Of the total reads, 50% matched to unique genomic locations, 16 to 18% showed multiple matches, and 25% had no match to the genome (Table 1 and table S1). 6000 reads from HEK were adenovirus or SV40 sequences, reflecting the origin of this cell line. We mapped the unique reads to known genes based on both ENSEMBL (*17*) and RefSeq/ElDorado (Tables 1 and 2 and tables S1 and S2) (*16*): 80% of the unique reads mapped to known exons.

Digital expression levels were normalized (NE values) by taking into account the theoretical number of unique 27-mers (sequences that are 27 bases long) contained in each exon and the total number of reads generated in each experiment (table S2) (*16*).

To assess whether NE values were a reliable indicator of gene activity, we correlated these values with hypophosphorylated RNA polymerase II (PolIIa) occupancy, used as a landmark of transcription initiation (*18*). For HEK, we identified PolIIa islands by chromatin immunoprecipitation and sequencing (ChIP-Seq) (*16*). Figure 1 shows that the density of PolIIa reads correlates positively with gene expression levels. However, in contrast to a study reporting that 37% of the silent promoters contained PolII islands (*19*), we observed virtually no PolIIa near the promoters of silent genes. This apparent contradiction is most likely due to the higher sensitivity of RNA-Seq, detecting gene expression that would be scored silent with arrays (see below). The current model of the pre-recruitment of PolIIa at the promoter of silent genes (*20*) may be lacking sufficiently sensitive expression data. In Fig. 1, the peaks for low and moderately expressed genes exhibit a more pronounced shoulder than those for highly expressed genes. This might reflect the presence of a large preinitiation complex where PolIIa is parked upstream of the transcription start site (TSS) of the less active genes until activated, or the existence of alternative TSS. In clustering the reads specifying PolIIa-bound regions, we identified 9710 PolIIa-bound regions, of which 80% associated to known promoters (table S3) (*16*). Of the remaining 1936 PolIIa-bound regions, more than half were supported by Cap-analysis of gene expression (CAGE) tags (*21*), and 567 were either located within genes or less than 1 kb upstream of the next annotated transcript, representing putative alternative promoters.

In evaluating the dynamic range and sensitivity of RNA-Seq, we predicted the number of genes present within a cell type by applying a Poisson mixture statistical analysis on the number of reads mapped to genes (*16*, *22*). We showed that the performances achieved for each sample corresponded to a gene identification score of 83 to 92% for HEK and 70 to

[1]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany. [2]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany. [3]International Max Planck Research School for Computational Biology and Scientific Computing. [4]Genomatix Software Gmbh, Bayerstrasse 85a, 80335 Munich, Germany.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: yaspo@molgen.mpg.de

84% for B cells (fig. S1) (*16*). RNA-Seq was significantly more sensitive than microarrays on the same RNA source, detecting 25% more genes (Fig. 2, A and B) (*16*). Genes detected exclusively by RNA-Seq were in the lowest range of NE values, corresponding to rarely expressed genes (Fig. 2B). Between 100 and 200 transcripts were only detected on arrays (Fig. 2A), exhibiting intensity values close to the background and hence increasing the chance that those were artefacts, a known issue in array analysis (*23*, *24*).

We analyzed expression of all ENSEMBL genes expressed simultaneously on both platforms and in both cell lines (7043 genes) (*16*). Correlation between the two platforms was high [Pearson correlation coefficient (PCC) = 0.88], in spite of a compression effect resulting in smaller ratios in microarrays (Fig. 2C). This feature was reported previously and is partly due to the limited dynamic range of array experiments (*23*–*26*). Microarrays detected 3421 genes whose levels of expression were different between the two cell lines, whereas RNA-Seq detected 4376 such genes. The overlap between the two approaches was 2685 genes (table S4). For the latter, levels-of-expression differences between the two cell lines were highly concordant between the two approaches (PCC, 0.94; Kendall rank correlation coefficient, 0.75).

We carried out a functional analysis of differentially regulated and cell type–specific genes (table S5) (*16*). Among the 55 genes most over-

Table 1. Summary of genes, splice junctions, and previously unrecognized TUs identified by RNA-Seq; mapping of the read for the merged lanes.

Mapping summary	HEK 293	B cells
Total reads	8,638,919	7,682,230
Low-quality reads	234,160	194,999
Reads with multiple matches	1,546,361	1,324,770
Reads with unique matches	4,640,112	3,895,643
Reads mapping to annotated RNAs (ENSEMBL + Eldorado)	3,712,476	2,902,387
ENSEMBL genes with at least five reads	12,567	10,668
ENSEMBL genes with at least one read	14,963	13,739
Reads in intronic clusters	38,598	44,781
ENSEMBL genes with intronic read clusters	1445	1409
Introns with read clusters	1862	1847
Reads with no match to the genome	2,218,286	2,266,818
Reads aligned to splice junctions	307,904	229,453
Identified junctions	78,880	62,596
(expected)	(81,302)	(66,981)
Genes (at least five reads) with junctions	10,292	8655
Genes (at least one read) with junctions	10,558	8910
Genes (at least one read) with previously unknown junctions	2078	1732
Previously unknown junctions	2397	1965
Previously unknown junctions identified by less than one read	203	182

Table 2. Summary of genes, splice junctions, and previously unrecognized TUs identified by RNA-Seq; features associated to the 352 previously unknown intergenic TUs identified in HEK and B cells.

Features	Number of TUs
CAGE tags	253
CAGE tags + PolIIa-bound regions	22
PolIIa-bound regions (without CAGE)	2
Contain repeated elements	50
Match to human pseudogenes	9
Identity to human full-length cDNA	28
Similarity to nonhuman sequences	16
Similarity to known proteins	121
Exon-intron structure	24
Uniquely expressed in HEK293	134
Uniquely expressed in B cells	153
Expressed in both cell types	66

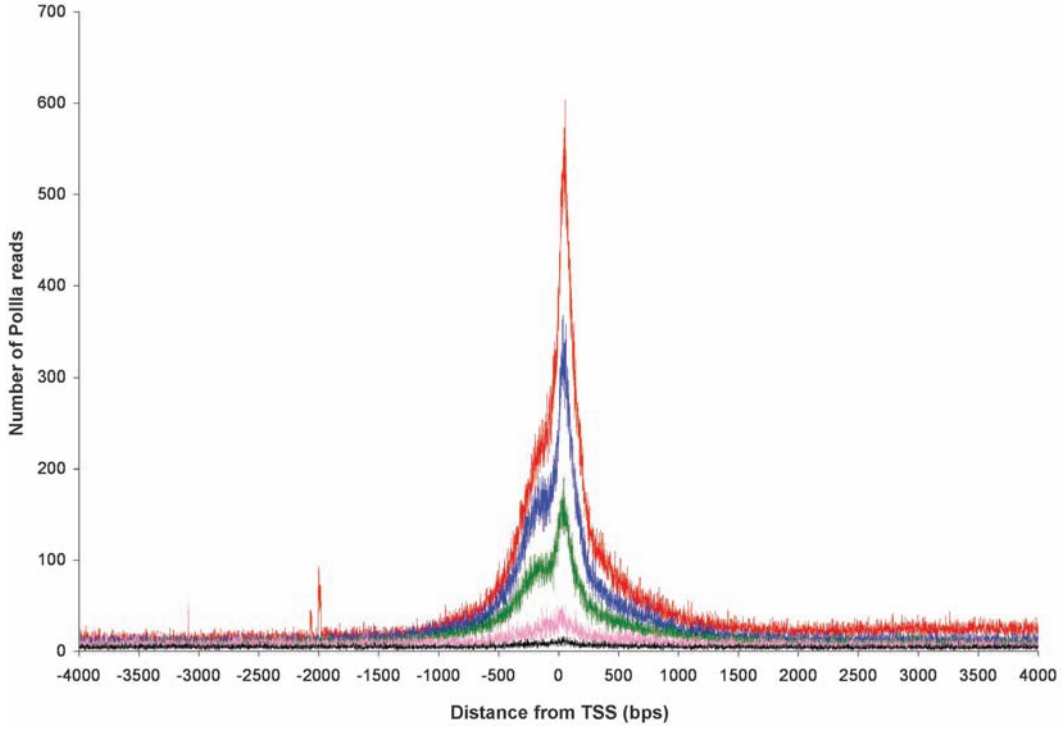

Fig. 1. Correlation of RNA PolIIa read density with TSS. The plot shows the number of RNA PolIIa reads relative to the TSS for all 12,567 ENSEMBL-expressed genes distributed in five groups: (i) high (4189 genes with 0.0889 < NE < 47.8; red); (ii) moderate (4189 genes with 0.0263 < NE < 0.0889; green); (iii) low (4189 genes with 0.0003 < NE < 0.0263; blue); (iv) uncertain (2396 genes with one to four reads; pink); and (v) silent (7333 genes with no read; black) expression. bps, base pairs.

expressed in the lymphoma cells, we found an enrichment of factors involved in Ras protein signal transduction pathway and immune system processes. The 271 most active genes specific to B cells were significantly enriched for MHC class II receptors and factors belonging to the CD38, LCK (lymphocyte-specific protein tyrosine kinase), ZAP70 (zeta chain–associated protein kinase 70 kD), CD19, and BLK (B lymphoid tyrosine kinase) signaling pathways. Of the 2669 genes specific to HEK, the top 1000 were enriched for factors involved in DNA binding and for cytoskeletal proteins binding the extracellular matrix.

To more precisely define 5′ and 3′ gene boundaries and identify all transcribed regions, we analyzed reads in intronic and intergenic regions. We assessed noise levels by means of a Poisson model of the noncoding part of the genome; the probability to observe more than four random reads per 100-bp window was $<10^{-12}$ (*16*). We scanned the DNA regions 5 kb upstream and downstream of all transcripts, only considering read clusters that displayed a density similar to that of the neighboring exon. Approximately 500 genes were extended at the 5′ end by at least 50 bp (table S6), 300 of which were supported by CAGE tag(s) (*21*), and ~300 genes were extended at their 3′ end in each cell line. Only 15% of these were common to both cell types. Furthermore, we searched for read clusters in the 39% of the genome corresponding to intronic regions, using a stringent algorithm that required a minimum of five reads in 100-bp sliding windows (*16*). We identified 2751 and 2862 clusters (average length of 191 bp, totaling 38,598 and 44,781 reads) located within 1862 and 1847 introns of 1445 and 1409 genes in HEK and B cells, respectively (table S7). A large fraction (87%) of these clusters mapped to human ESTs (table S7). ESTs were used to infer previously unidentified exons of known genes when clusters and genes mapped to the same EST (e.g., 1500 and 1358 previously unknown exons were connected within 916 and 834 genes in HEK and B cells, respectively). Of these exons, 70% were unique to one cell type and likely to be differentially spliced. Remaining clusters could either represent rare exons not represented in ESTs or hallmarks of transcriptional activity in both the DNA strands.

Similarly, we scanned for transcriptional activity in the 58% of the human genome corresponding to intergenic regions (*16*). We identified 531 clusters totaling 13,805 reads distributed in 280 intergenic regions (table S8). Using inferences from EST mapping, 237 out of 531 (237/531) clusters collapsed within 58 transcribed units (TUs), whereas 294/531 clusters remained individual units, identifying altogether 352 TUs (size range: 92 to 182 bp) (table S8). An exon-intron structure was found in 24 TUs, based on a minimal distance of 1 kb between two clusters (*16*). Additional attributes supporting the importance of the previously unrecognized TUs—including CAGE tags, PolIIa-Bound regions, and similarities to known expressed sequences identified by BlastN and BlastX (*27*) analysis—are summarized in Table 2 (details in table S8).

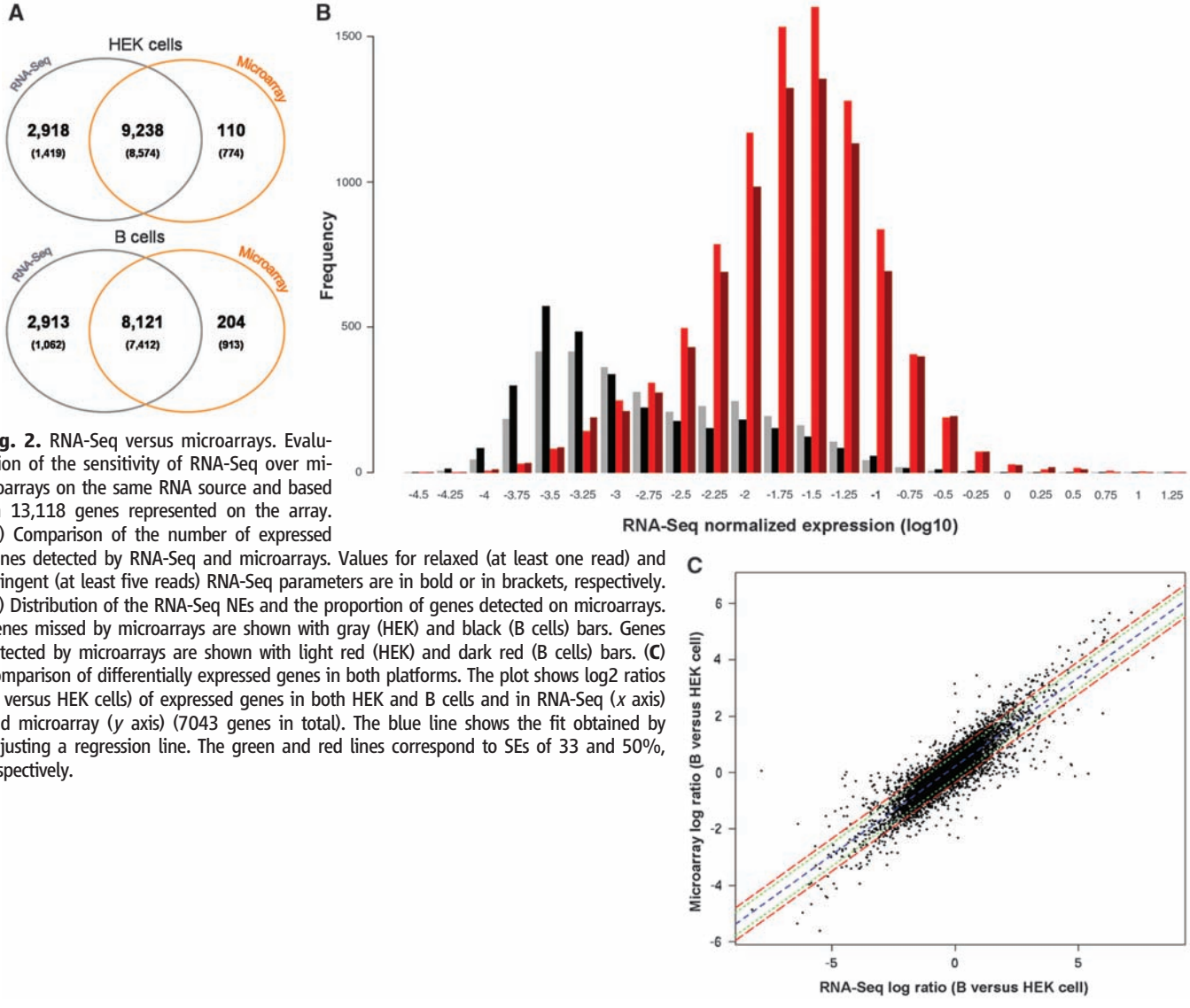

Fig. 2. RNA-Seq versus microarrays. Evaluation of the sensitivity of RNA-Seq over microarrays on the same RNA source and based on 13,118 genes represented on the array. (**A**) Comparison of the number of expressed genes detected by RNA-Seq and microarrays. Values for relaxed (at least one read) and stringent (at least five reads) RNA-Seq parameters are in bold or in brackets, respectively. (**B**) Distribution of the RNA-Seq NEs and the proportion of genes detected on microarrays. Genes missed by microarrays are shown with gray (HEK) and black (B cells) bars. Genes detected by microarrays are shown with light red (HEK) and dark red (B cells) bars. (**C**) Comparison of differentially expressed genes in both platforms. The plot shows log2 ratios (B versus HEK cells) of expressed genes in both HEK and B cells and in RNA-Seq (*x* axis) and microarray (*y* axis) (7043 genes in total). The blue line shows the fit obtained by adjusting a regression line. The green and red lines correspond to SEs of 33 and 50%, respectively.

More than 80% of the TUs were unique to one cell line. Similarities to known proteins were found in 121 TUs, although some contained a stop codon. In addition, Blast analysis grouped 13 TUs into six larger units (table S8). For instance, TU 33 and TU 34 defined a highly transcribed region spanning 190 kb on chromosome 2 (fig. S2). Overall, 7% of the orphan reads were clustered in potentially active regions. The bulk of orphan reads seems to reflect a moderate-to-low transcriptional activity more diffusely distributed in the genome. Relaxing the parameters to 3 reads per 400-bp window embedded 328,683 reads, roughly equally distributed between intronic and intergenic regions, covering a total of 27.5 Mb of DNA (0.9% of the human genome). Taking this figure, 66% of the polyadenylated transcriptome of the two cell lines mapped to known genes and 34% to nonannotated genomic regions.

Approximately 14% of the unmatched reads (Table 1) could be mapped to a set of synthetically computed splice junctions enumerating all theoretical constitutive and AS junctions within annotated transcripts Table 1 and table S9, A and B) (*16*). We observed, on average, 7.2 junctions per gene and a mean density of 3.8 reads

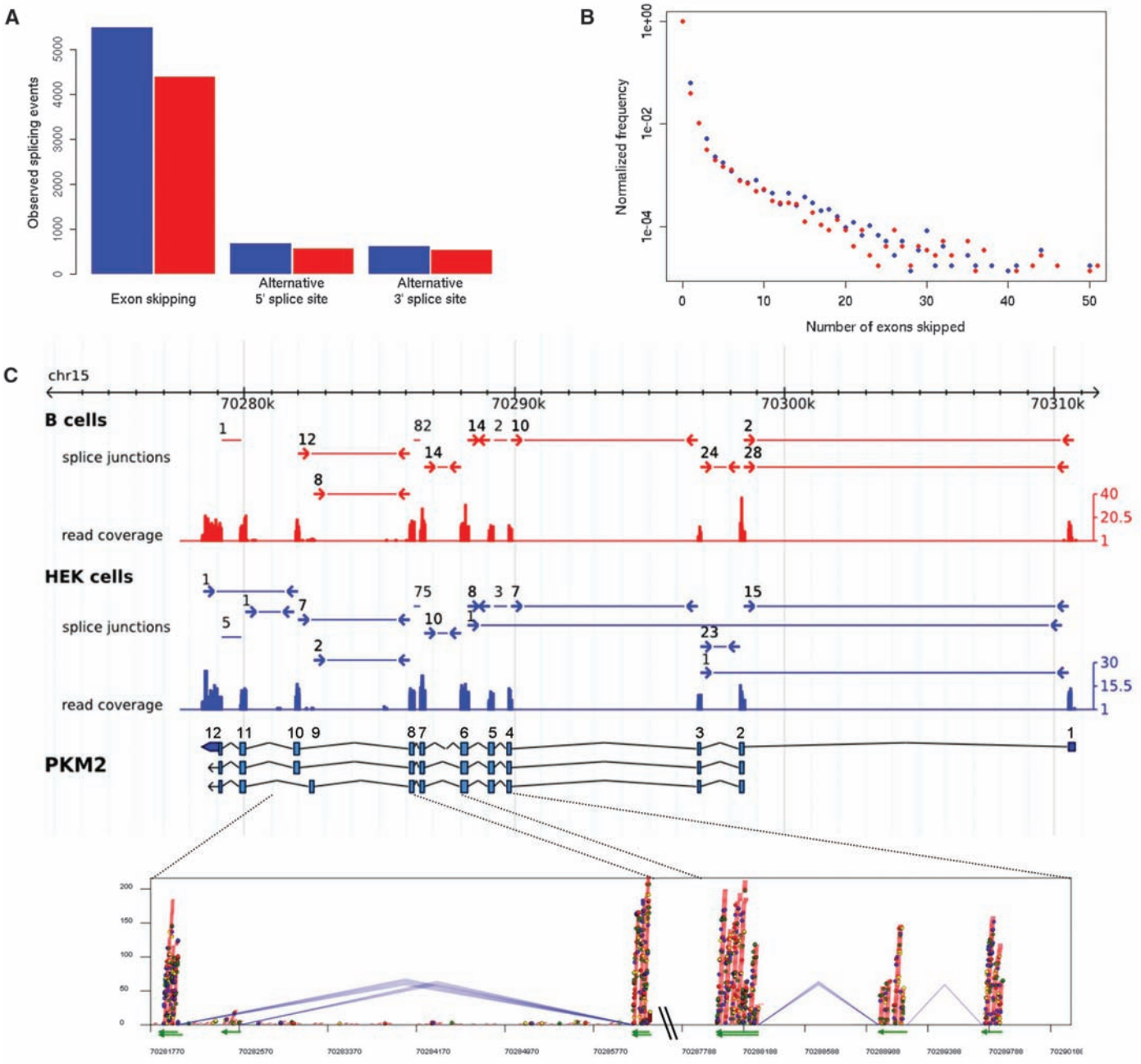

Fig. 3. AS events observed by junction reads. (**A**) Distribution of the three major types of AS: (i) cassette exons, (ii) alternative 5′ splice sites, and (iii) alternative 3′ splice sites. (**B**) Frequency of skipped exons normalized to the number of junctions that do not skip any exon in HEK (blue) and in B cells (red). (**C**) Example of AS in the PKM2 gene. Three isoforms annotated in ENSEMBL (ENST00000335181, ENST00000389092, ENST00000389091) are shown next to the gene name, and exons are numbered. The read coverage is shown for each exon (blue for HEK and red for B cells). Splice-junction reads are shown as arrows; the numbers above the arrows represent the number of reads at junctions. The bottom box shows base-pair resolution coverage in HEK cells of the gene's regions containing exons 8 to 10 (green arrows at left) and 4 to 6 (green arrows at right). The blue lines denote splice junctions. (Left) Two different sequenced junctions connecting either exon 9 or exon 10 and identifying alternative transcripts with mutually exclusive exons in HEK and in B cells. Colored dots represent sequence differences.

per junction. Although 29,689 junctions in HEK and 24,848 in B cells had only one read, those were considered highly notable, as we expect at most 23 reads hitting a junction by chance in the entire data set (*16*). Splice junctions were associated with 81% of the expressed genes. We also observed splice junctions for ~260 genes in each cell line that were not classified as expressed (Tables 1 and 2). Of those, 70% had between 1 and 4 reads and 30% were silent, suggesting a very low activity. The fact that 2275 expressed genes in HEK and 2013 in B cells had no splice-junction reads correlated with the fact that those genes contained fewer exons and a lower activity than the average, reducing the probability to hit a splice junction.

We observed 95% of the splicing events expected in this data set, given the current sequencing depth (Table 1) (*16*). We identified 4096 previously unknown splice junctions in 3106 genes, mostly called by single reads and unique to one cell type (Table 1). Many of these junctions were associated with actively transcribed genes exhibiting more exons than average, pointing to rare splicing events. Approximately 6% of all splice-junction reads identified AS events (6416 junctions in 3916 genes HEK and 5195 junctions in 3262 genes in B cells) (table S9). In a parallel study surveying the mouse transcriptome, AS forms were observed for 3462 genes in three tissues (*28*), but no attempts were made to search for previously unrecognized junctions. Within a cell type, junction reads identify AS in 30% of the expressed genes, where exon skipping was largely overrepresented (Fig. 3A). Skipping events affected mostly one or two exons, with a sharp decline between one and five exons (Fig. 3B). An illustrative example of AS is given for PKM2, also showing that the read density reflects the exon usage (Fig. 3C). Very complex patterns of AS could be detected. For instance, with the use of EIF4G1 coding for the eukaryotic translation initiation factor 4 gamma 1, we showed 12 AS junctions in B cells, of which five have not yet been identified (fig. S3). Although AS is known to regulate the expression of EIF4G1 (*29*, *30*), such a complex pattern had never been described before.

References and Notes

1. H. Lehrach *et al.*, in *Genome Analysis: Genetic and Physical Mapping*, vol. 1, K. Davies, Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1990), pp. 39–81.
2. G. G. Lennon, H. Lehrach, *Trends Genet.* **7**, 314 (1991).
3. E. M. Southern, *Curr. Opin. Genet. Dev.* **2**, 412 (1992).
4. G. M. Hampton, H. F. Frierson, *Trends Mol. Med.* **9**, 5 (2003).
5. T. R. Hughes *et al.*, *Cell* **102**, 109 (2000).
6. P. Bertone *et al.*, *Science* **306**, 2242 (2004), published online 11 November 2004; 10.1126/science.1103388.
7. J. Cheng *et al.*, *Science* **308**, 1149 (2005), published online 24 March 2005; 10.1126/science.1108625.
8. Q. Pan *et al.*, *Mol. Cell* **16**, 929 (2004).
9. J. A. Calarco, A. L. Saltzman, J. Y. Ip, B. J. Blencowe, *Adv. Exp. Med. Biol.* **623**, 64 (2007).
10. V. E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler, *Science* **270**, 484 (1995).
11. C. V. Jongeneel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 4702 (2003).
12. J. B. Kim *et al.*, *Science* **316**, 1481 (2007).
13. U. Nagalakshmi *et al.*, *Science* **320**, 1344 (2008), published online 30 April 2008; 10.1126/science.1158441.
14. R. Lister *et al.*, *Cell* **133**, 523 (2008).
15. B. T. Wilhelm *et al.*, *Nature* **453**, 1239 (2008).
16. Materials and methods are available as supporting material on *Science* Online.
17. T. J. Hubbard *et al.*, *Nucleic Acids Res.* **35**, D610 (2007).
18. A. S. Brodsky *et al.*, *Genome Biol.* **6**, R64 (2005).
19. A. Barski *et al.*, *Cell* **129**, 823 (2007).
20. T. Margaritis, F. C. Holstege, *Cell* **133**, 581 (2008).
21. T. Shiraki *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15776 (2003).
22. J.-P. Z. Wang, B. G. Lindsay, *J. Am. Stat. Assoc.* **100**, 942 (2005).
23. F. Liu *et al.*, *BMC Genomics* **8**, 153 (2007).
24. M. Barnes, J. Freudenberg, S. Thompson, B. Aronow, P. Pavlidis, *Nucleic Acids Res.* **33**, 5914 (2005).
25. C. V. Jongeneel *et al.*, *Genome Res.* **15**, 1007 (2005).
26. L. Shi *et al.*, *Nat. Biotechnol.* **24**, 1151 (2006).
27. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *J. Mol. Biol.* **215**, 403 (1990).
28. A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, *Nat. Methods* **5**, 621 (2008).
29. M. P. Byrd, M. Zamora, R. E. Lloyd, *J. Biol. Chem.* **280**, 18610 (2005).
30. M. J. Coldwell, S. J. Morley, *Mol. Cell. Biol.* **26**, 8448 (2006).

31. The Gene Expression Omnibus accession number for the microarrays and sequence data is GSE11892. Data are displayed in a public version of browser interfaces developed by the Max Planck Institute for Molecular Genetics (http://promotion.molgen.mpg.de/cgi-bin/gbrowse/Hs.Solexa) and Genomatix (www.genomatix.de/MPI). This work was supported in part by the Max Planck Society, the European Union [AnEUploidy (LSHG-CT-2006-037627) and BioSapiens (LHSG-CT-2003-503265)], the National Genome Research Network, and the Federal Ministry for Education and Research of Germany [BioChancePLUS-3 (0313724A) to A.K. and M.S.].

Supporting Online Material

www.sciencemag.org/cgi/content/full/1160342/DC1
Materials and Methods
Figs. S1 to S3
Tables S1 to S9
References

12 May 2008; accepted 27 June 2008
Published online 3 July 2008;
10.1126/science.1160342
Include this information when citing this paper.