Supporting Information

Tilgner et al. 10.1073/pnas.1400447111

SI Methods

Cell Culture and RNA Sample Preparation. Human lymphoblastoid cells (LCLs) were grown to a density of 6–8 × 10^5/mL in RPMI1640 (Life Technologies; 11875-093; with 2mM L-glutamine) with 15% (vol/vol) heat-inactivated fetal bovine serum (Life Technologies; 10082-147) and 1% PenStrep (Life Technologies; 15140-122). Cells were lysed and RNA was extracted and subjected to double PolyA selection by using the mRNA direct kit (Life Technologies) according to the manufacturer's instructions, then purified by using the Qiagen RNeasy kit (Qiagen) and treated with RNase-free DNase (Qiagen). RNA integrity was checked on a Bioanalyzer (Agilent) and only RNA with an RNA integrity number of >9.5 was used for subsequent library constructions.

101-bp Paired-End Sequencing. Approximately 100 ng of PolyA RNA were fragmented with 10× fragmentation buffer (Life Technologies; AM8740) for 2 min at 70 °C. cDNA was synthesized by using Life Technologies' SuperScript double-stranded cDNA synthesis kit (11917-010), and reverse transcription (RT) was primed with random hexamers. The cDNA was end-repaired and A-tailed, and Illumina Paired-end adapters were added. After size selection on an agarose gel and PCR amplification, samples were sequenced on the Illumina Hi-Seq 2000, generating paired-end (PE) reads with a length of 2 × 101-bp.

cDNA Synthesis. PolyA-selected RNA from the same batch as for the 101-bp PE-read library was used to synthesize unfragmented cDNA using Life Technologies' SuperScript double-stranded cDNA synthesis kit (11917–010) and RT was primed with anchored oligo(dT)20 primers (Life Technologies; 12577-011). The cDNA was purified by using Qiagen MinElute PCR purification kit (28004) and eluted with 50 μL of RNase-free water. The cDNA was quantified by using the Qubit HS dsDNA kit (Life Technologies; Q32851), and quality was assessed by using the Agilent 2100 bioanalyzer.

Library Preparation, Sequencing, and Data Collection. SMRT bell libraries were generated by using Pacific Biosciences' 2.0 template prep kit (part 001-540-726) and Pacific Biosciences' template preparation and sequencing protocol for 0.25- to 3-kb libraries. SMRT bell templates were bound to polymerases by using the DNA/polymerase binding kit 2.0 (part 001-672-551) and v2 primers. Polymerase–template complexes were bound to magbeads by using Pacific Biosciences' Magbead binding kit (part 100-134-800), and sequencing was carried out on the Pacific Biosciences' real-time sequencers RTII by using C2 sequencing reagents. Movie lengths were 120 min for each SMRT cell. Subread filtering was performed by using Pacific Biosciences' SMRT analysis software (v1.3.3).

Long-Read Mapping. Mapping of and primary analysis of PacBio CCS reads was carried out as described (1, 2) by using GMAP (3).

Criteria for Considering CSMMs Full-Length. We used four different and increasingly relaxed criteria to define, if a CSMM was likely to represent all splice sites of the RNA molecule it originated from: (*i*): The first splice site of the read-mapping is the first splice site of an annotated transcript and the last splice site of the read is the last splice site of an annotated transcript. (*ii*) The first exon of the read-mapping overlaps the first exon of an annotated transcript and the last exon of the read-mapping overlaps the last exon of an annotated transcript. Note, that these two criteria logically include the corresponding criterion *i*. (*iii*) Criterion *ii* applies for the first exon of the read or the first exon of the read overlaps no exon at all, wereas for the last exon of the read, criterion 2 applies or the last exon of the read overlaps no exon at all. (*iv*) If criterion *iii* does not apply for the most 3′ exon (but criterion *iii* holds for the most 5′ exon), we consider the read mapping full-length if and only if the read shows a polyA tail of 15 bp or longer.

Construction and Quantification of a Long-Read–Enhanced Annotation. CSMMs that showed attributes of full-length molecules were added to the annotation if and only if (*i*) they were novel (1) with respect to the Gencode-version 15 annotation and (*ii*) they could be attributed to a gene by splice site identity (1).

When multiple CSMMs showed identical splice sites, we chose a random CSMM to represent them. The first exon start and the last exon end were set to be the most upstream and downstream nucleotides of all these CSMMs.

101-bp PE Read Mapping. Mapping of these reads was performed by using the STAR mapper (4) by using the same parameters as in the ENCODE project (5) with the following exceptions: (*i*) we required a minimal intron length of 25 (–alignIntronMin 25) for consistency with our previous work (1, 2), (*ii*) the guide annotation was the Gencode 15 annotation (6), and (*iii*) mappings to annotated junctions were only considered when ≥5 matching bp were found on either side (–alignSJDBoverhangMin 5).

Gene and Transcript Quantification Through Illumina Reads. Quantification of genes and transcripts was performed by using Cufflinks version 2.1.1 based on the above STAR mappings: (*i*) defining a minimal intron length of 25 (–min-intron-length 25) and (*ii*) giving the Gencode 15 annotation as the source of genes and transcripts to quantify (-G gencode_file_name).

Selection of 166 Genes for Principal Component Analysis. We considered all spliced genes that (*i*) did not share any splice sites with other genes and (*ii*) for which two or more annotated heterozygous SNPs existed that were overlapped by at least 10 full-length CSMMs and at least 80% of all full-length CSMMs attributed to this gene.

Annotated Heterozygous SNVs. SNV calls for the GM12878 cell line were downloaded from the University of California, Santa Cruz browser on July 18, 2013. Only variations affecting a single nucleotide and which substituted it by a single nucleotide were retained.

Gene Set for Comparison Between Illumina Sequencing and Long-Read Sequencing. We selected all Gencode (version 15) annotated genes with at least one spliced annotated isoform, which (*i*) did not share any splice sites with other genes and (*ii*) were classified in the Gencode-15 annotation, either as "protein_coding" or as "lincRNA."

PCR and Sanger Sequencing Analysis. cDNA molecules that include exon 2 were amplified by using a forward primer spanning the exon 1-exon 2 junction (AGATGCTACTGGCAGCTGGATGTC) and a reverse primer located on the last exon (CAAAACACTGGGACAACTTGTAA). cDNA molecules that skip exon 2 were amplified by using a forward primer spanning the exon 1-exon 3 junction (AGATGCTACTGGCTGCCAGTTTTG) and the same reverse primer as above. Both PCRs were performed as follows: (*i*) denaturation at 98 °C for 30 s, (*ii*) stage 2 (40 times) for 98 °C for 10 s, 56 °C for 20 s, and 72 °C for 40 s, and (*iii*) stage 3 72 °C for 120 s.

1. Tilgner H, et al. (2013) Accurate identification and analysis of human mRNA isoforms using deep long read sequencing. *G3 (Bethesda)* 3(3):387–397.
2. Sharon D, Tilgner H, Grubert F, Snyder M (2013) A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* 31(11):1009–1014.
3. Wu TD, Watanabe CK (2005) GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859–1875.
4. Dobin A, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1): 15–21.
5. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414): 101–108.
6. Harrow J, et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.

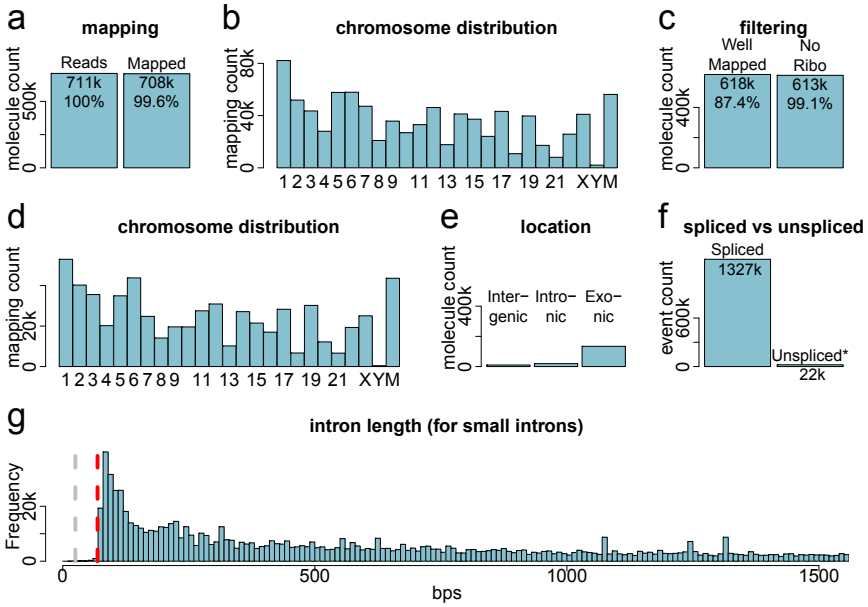

Fig. S1. Mapping statistics. (*A*) Number of molecules submitted for mapping (*Left*) and number (and percentage) of molecules that could be mapped to the hg19 genome using GMAP (*Right*). (*B*) Chromosome distribution of molecule mappings including multiple mappings for a single molecule if GMAP produced such. (*C*) Number and percentage of molecules for which we could determine a single high-confidence mapping ("well-mapped,"; *Left*) and those that did not overlap ribosomal RNA genes (*Right*, percentage with respect to previous bar). (*D*) Chromosome distribution of high confidence read mappings. (*E*) Number of molecules falling entirely into intergenic, intronic, and exonic regions. Note that the definition of intergenic used here is based on the Gencode-v15 annotation, which defines lncRNA genes, ribosomal RNA genes, and many other kinds of short RNA genes as "genes." (*F*) Number of introns in high-confidence mappings (*Left*) and number of events of incomplete splicing (intron retention and/or partially spliced nuclear RNAs). (*G*) Intron length distribution for introns in consensus split-molecule mappings, showing only introns of up to 1.5 kb. The gray dashed line (at 25 bp) indicates that we asked GMAP to consider splits below 25 bp as insertions/deletions. The red dashed line (at 70 bp) indicates a cutoff under which very few annotated human introns could be found (1), suggesting a minimal intron size of this length, under which human introns might be difficult to process for the splicing machinery. Reassuringly GMAP reported almost no introns between 25 and 70 bp, indicating that intron calls are generally high-quality calls.

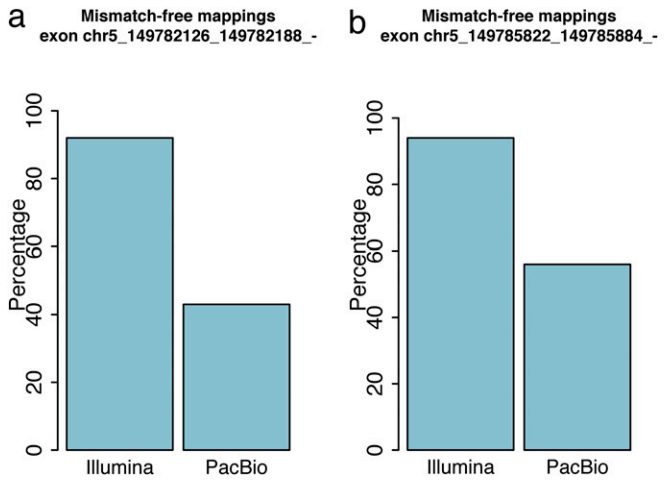

Fig. S2. Comparison of sequence quality in a highly controlled fashion. Illumina and PacBio reads differ greatly in their length and, therefore, also in the number of times they span introns. We selected the two most abundantly found exons in the PacBio data that were between 60 and 70 bp. The rationale for the two cutoffs is to select exons for which Illumina reads can have multiple introns just as the PacBio reads but which are not much shorter than exons typically found. The two exons in question (for which we have 3,556 and 3,383 PacBio reads in total) are given by the following locations: chr5:149782126–149782188 and chr5:149785822–149785884. We remapped all Ilumina reads in this region by using the same aligner (GMAP) as for the PacBio reads. Using this aligner for short reads takes longer but is for the sake of this one-time experiment (ca. 280,000 Illumina reads) advantageous, because there is no variability in mapping stemming from the use of different mappers. We then selected all reads that contained the entire exon for the two above exons and determined the number of reads that are error-free. *A* and *B* shows the percentage of reads for both platforms that mapped perfectly to exon chr5:149785822–149785884 (*A*) and for exon chr5:149785822–149785884 (*B*). This analysis shows that PacBio has more errors per mapped region, whereas for Illumina, one could, in principle, remove imperfectly mapping reads.

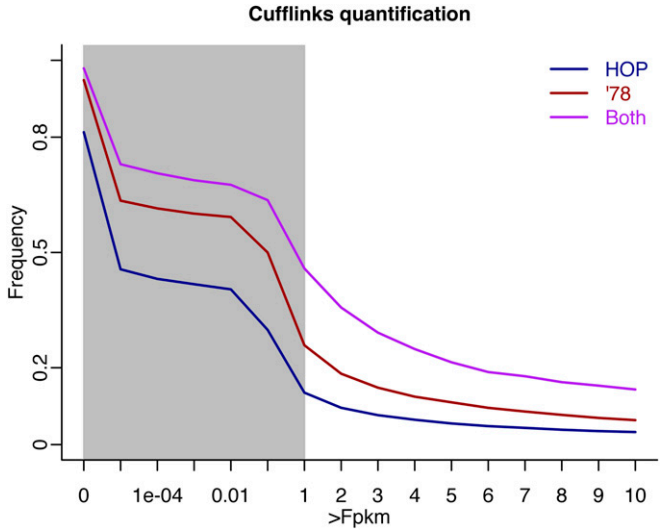

Fig. S3. Quantification of an enhanced annotation by using novel alignments irrespectively of whether all junctions are supported by the existing annotation or short-read sequencing. The quantification of the enhanced annotation with Illumina 101-bp PE sequencing was performed by using the same approach as in Fig. 3*C*.