Supplemental Material

Genetic regulation of nascent RNA maturation revealed by direct RNA nanopore sequencing

Karine Choquet, Louis-Philippe Chaumont, Simon Bache, Autum R. Baxter-Koenigs, L. Stirling Churchman

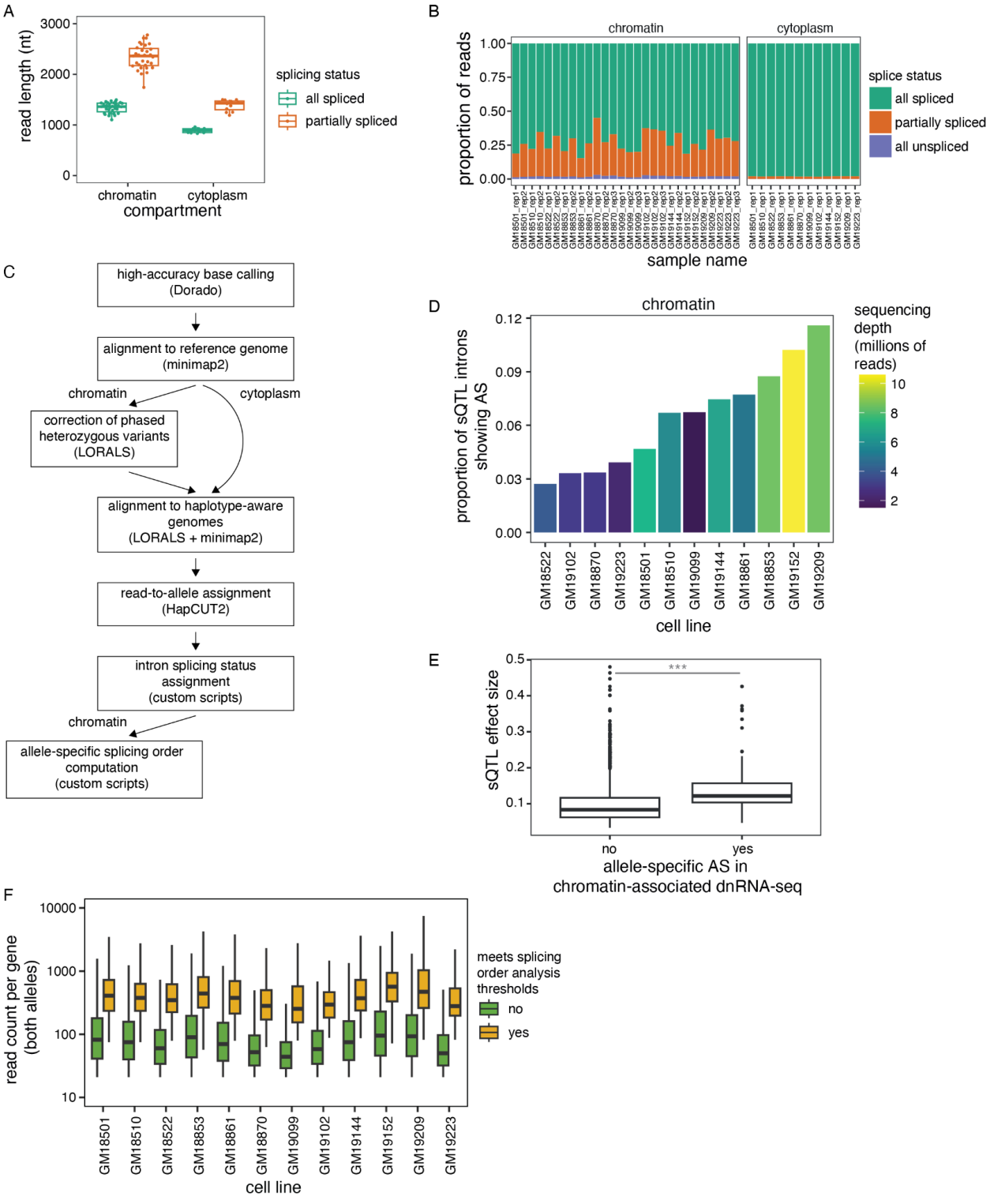

Figure S1. Quality control metrics of subcellular dnRNA-seq (related to Figure 1). A) Distribution of average read lengths per sample as a function of the read splicing status and the

subcellular compartment. Each dot represents one sample. B) Proportion of reads spanning at least two introns that are all spliced, partially spliced, or all unspliced, in each subcellular compartment. C) Workflow for the processing and allele-specific analysis of dnRNA-seq data. Correction of phased heterozygous variants is performed once per cell line using reads from chromatin-associated RNA, since they contain more introns and thus more information to determine phasing compared to cytoplasm. D) Proportion of GTEx sQTL introns showing alternative splicing (AS) in chromatin-associated dnRNA-seq. Bars are colored based on the total sequencing read depth for each cell line. E) GTEx sQTL effect size for introns that showed allele-specific AS or not in chromatin-associated dnRNA-seq. The two groups were compared using a two-sided Wilcoxon rank-sum test. ***: p-value < 2.2e-16. F) Distribution of the number of reads assigned to either allele for genes containing intron groups that meet the thresholds for splicing order analysis versus those that do not.

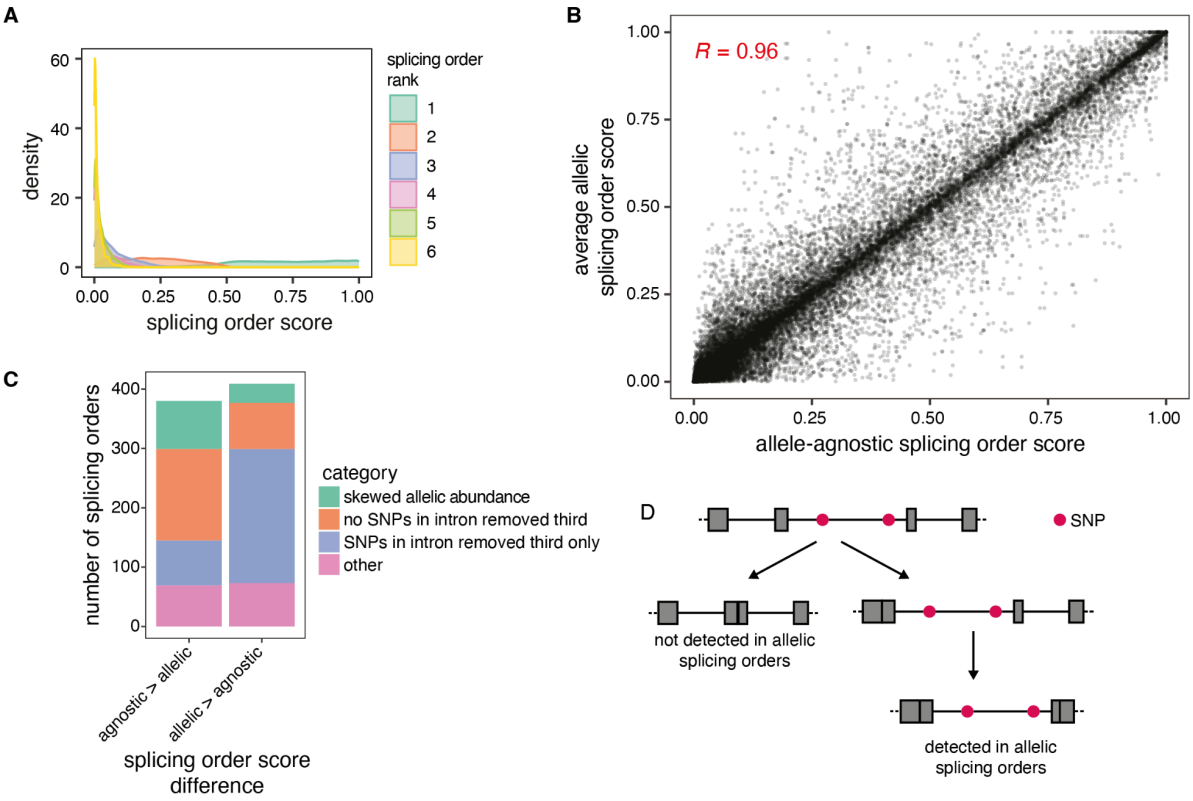

Figure S2. Characterization of allelic splicing orders (related to Figure 1). A) Distribution of splicing order scores across all LCLs and alleles, colored by splicing order rank. B) Correlation between allele-agnostic and average allelic splicing order scores across LCLs. Each dot represents one splicing order in one LCL. C) Classification of splicing orders based on whether there is a skewed abundance of pre-mRNA per allele (ratio of allele 1 reads / all reads < 0.4 or > 0.6) or whether there are SNPs in the intron that are removed third. Each observation refers to one splicing order in one cell line. Only splicing orders with a difference > 0.1 between allelic and allele-agnostic splicing order scores, and a score > 0.25 in at least one of the two approaches, are included. D) Schematic representing why some partially spliced reads are not detected (left) in allelic splicing orders, leading to a higher splicing order score for allele-agnostic splicing orders, or enriched (right) in allelic splicing orders, leading to a higher splicing order score for allelic splicing orders.

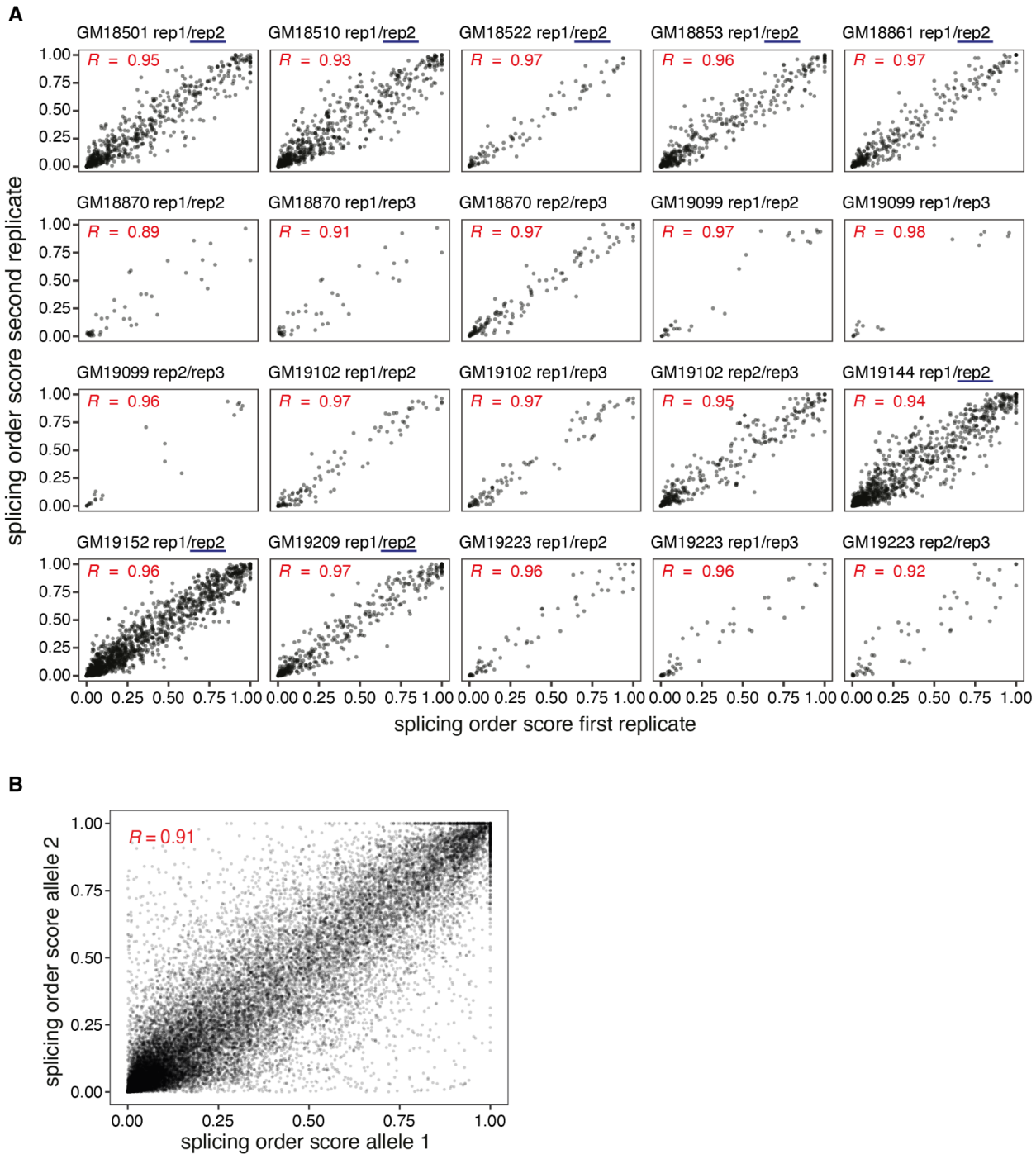

Figure S3. Correlation in splicing order scores between replicates and between alleles (related to Figure 1). A) Correlation in splicing order scores between biological or technical replicates of each LCL. Each dot represents one splicing order on one allele. Replicates sequenced with SQK-RNA004 are underlined in dark blue. For simplicity, "rep1", "rep2", and "rep3" refer to the first, second and third listed replicates in Table S1. B) Correlation in splicing order scores between alleles for all LCLs. Each dot represents one splicing order in one LCL.

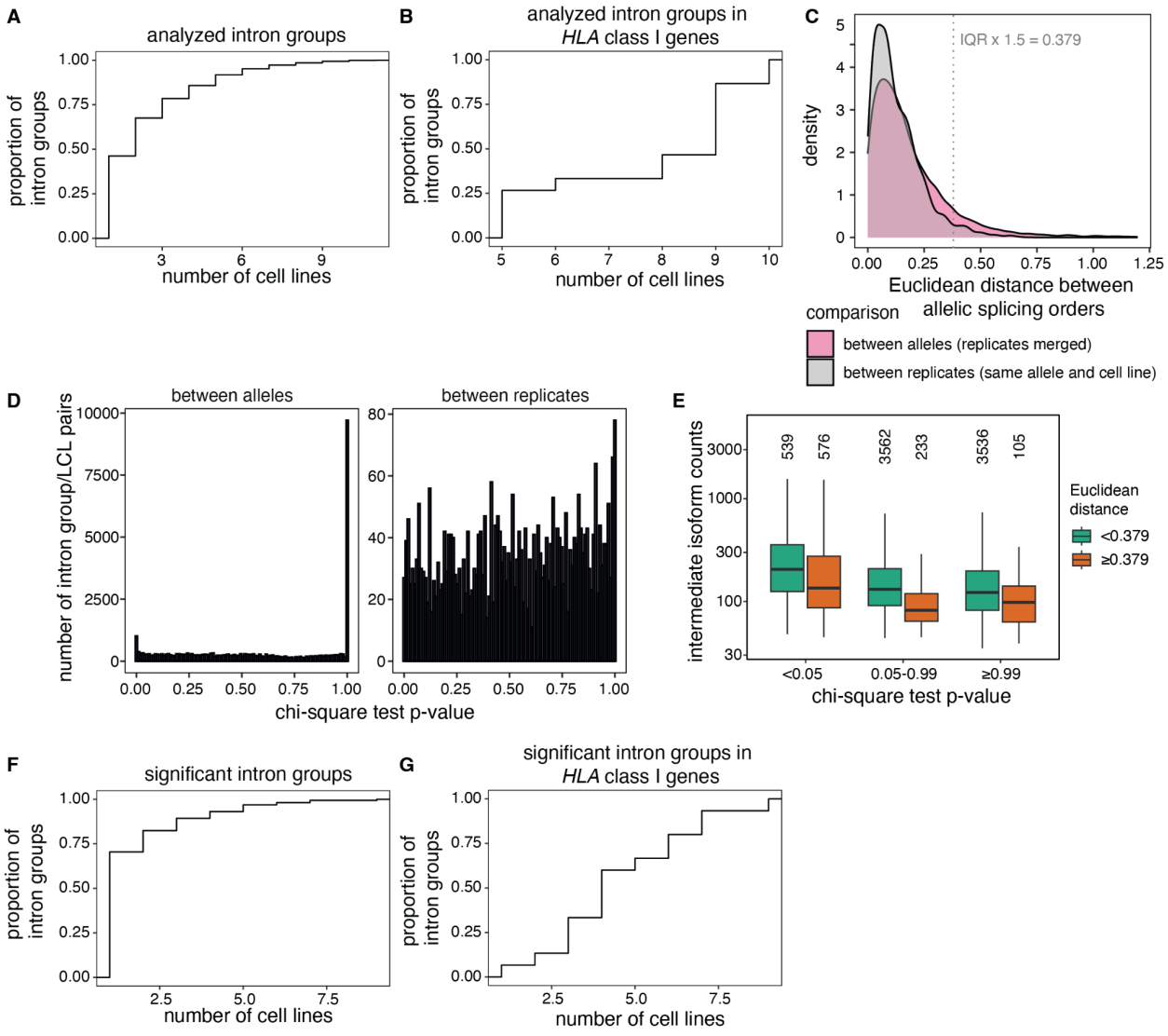

Figure S4. Identification of allele-specific splicing orders (related to Figure 1). A) Cumulative distribution frequency (CDF) of the number of cell lines in which each analyzed intron group could be detected in an allele-specific manner. B) Same as A), but only for intron groups in *HLA* class I genes. C) Distribution of the Euclidean distance between splicing order scores between alleles in merged replicates (pink) or between replicates of the same allele (grey) across all analyzed intron groups and LCLs. The threshold for a significant difference between alleles is shown as a grey dotted line (IQR: interquartile range of the distribution of splicing order scores between replicates). D) Distributions of p-values from chi-square tests comparing intermediate isoform counts between alleles in merged replicates (left) or between replicates of the same allele (right) across all analyzed intron groups and LCLs. E) Distribution of intermediate isoform counts per intron group as a function of chi-square test p-value and Euclidean distance per intron group. F) CDF of the number of cell lines in which each analyzed intron group showed significant allele-specific splicing order. G) Same as F), but only for intron groups in *HLA* class I genes.

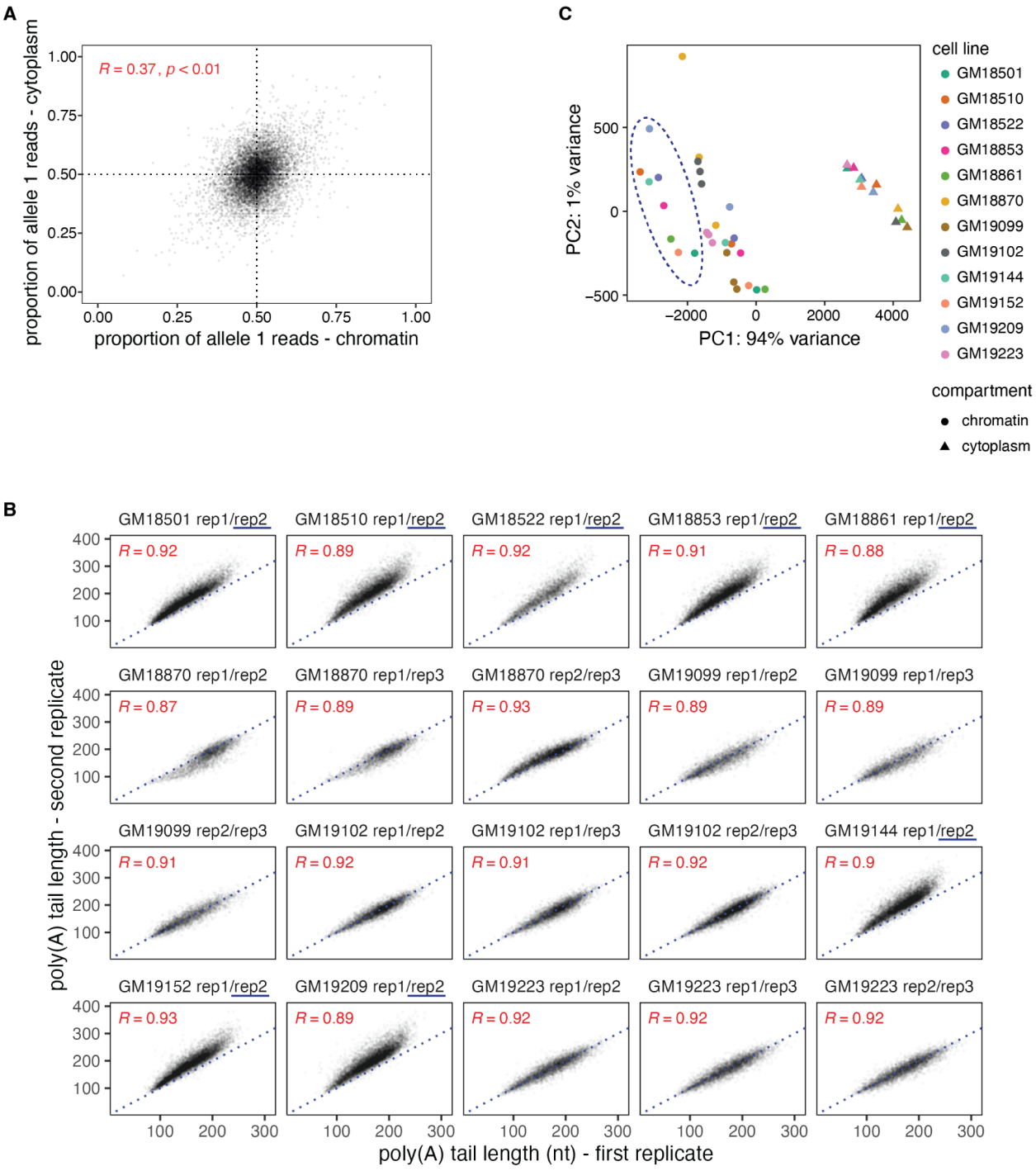

Figure S5. Allele-specific mRNA abundance and poly(A) tail length quality control (related to Figures 3 and 4). A) Correlation in allele-specific mRNA abundance between chromatin and cytoplasm. The proportion of allele 1 reads divided by the total number of reads for alleles 1 and 2 is shown for each subcellular compartment. Each dot represents one gene in one LCL. All genes that met the coverage threshold in both compartments are shown. B) Correlation in chromatin-associated poly(A) tail lengths between biological or technical replicates. Each dot represents one gene. Replicates sequenced with SQK-RNA004 are underlined in dark blue. For simplicity,

“rep1”, “rep2”, and “rep3” refer to the first, second and third listed replicates in Table S1. C) Principal component analysis of poly(A) tail lengths across replicates and subcellular compartments. Replicates sequenced with SQK-RNA004 are circled with a dotted dark blue line.

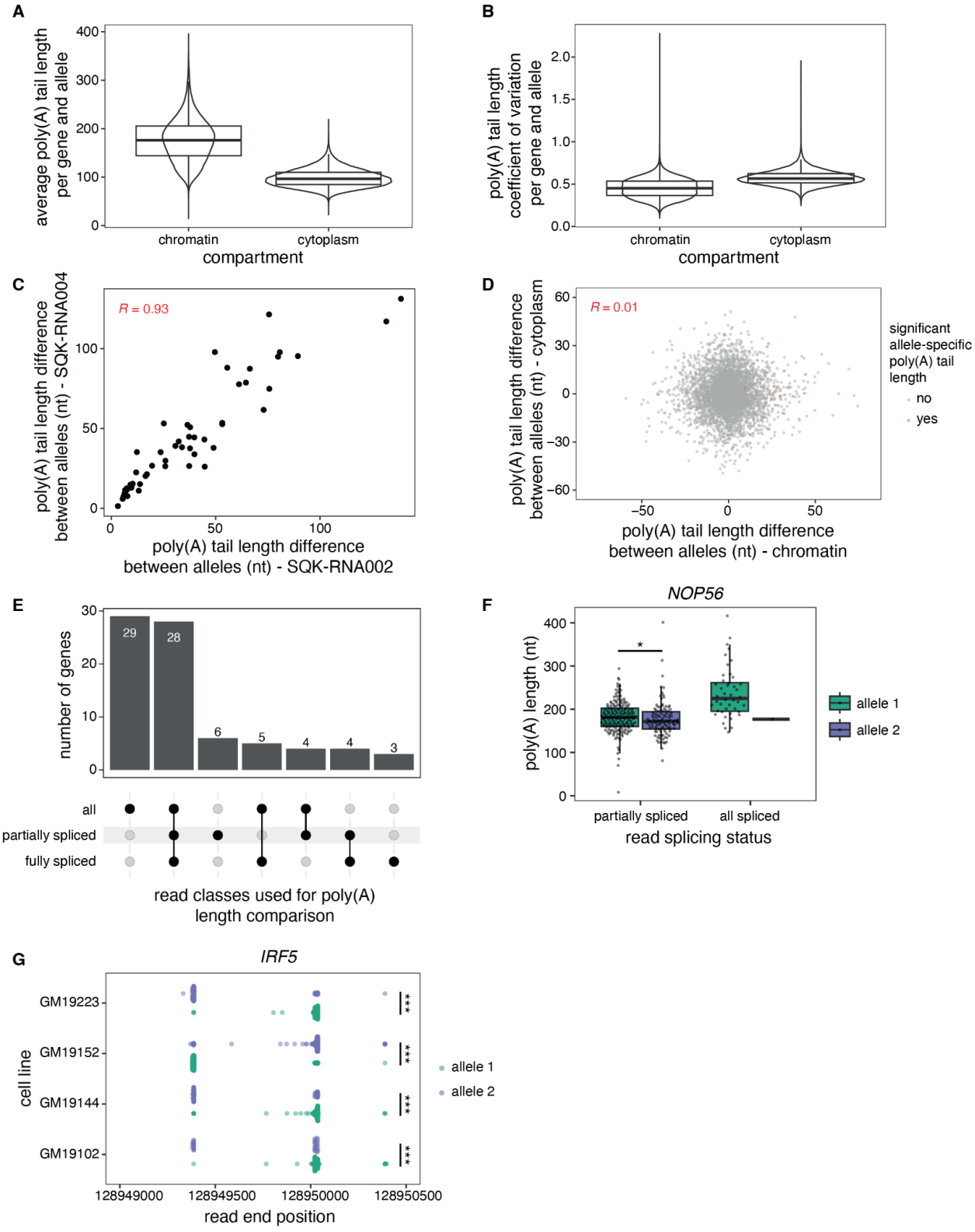

Figure S6. Allele-specific poly(A) tail length and 3′-end position (related to Figure 4). A) Distribution of the average poly(A) tail length per gene and allele across all LCLs for each

compartment. B) Distribution of the poly(A) tail length coefficient of variation per gene and allele across all LCLs for each compartment. C) Correlation in the absolute poly(A) tail length difference between alleles for genes that showed a significant difference in chromatin-associated poly(A), in LCLs for which one technical replicate was sequenced with SQK-RNA002 and the other with SQK-RNA004. Each dot represents one LCL/gene pair. D) Correlation in poly(A) tail length difference between alleles (allele 1 - allele 2) between chromatin-associated and cytoplasmic RNA. Genes with significant differences in poly(A) tail length on chromatin are shown in orange. E) UpSet plot showing the number of genes with a significant difference in poly(A) tail length between alleles for different classes of reads (all reads, partially spliced reads and fully spliced reads). F) Poly(A) tail length distribution for partially spliced and all spliced reads from *NOP56* in GM19102. A two-sided Wilcoxon rank-sum test was used to compare tail length distributions between alleles for each read splicing status. *: adjusted p-value < 0.05. G) 3′-end position of reads mapping to *IRF5* in four LCLs. Each dot represents one read. The x-axis shows allele-specific differential accumulation of reads at two distinct positions. A two-sided Wilcoxon rank-sum test was used to compare 3′-end position distributions between alleles for each cell line. ***: p-value < 0.001.

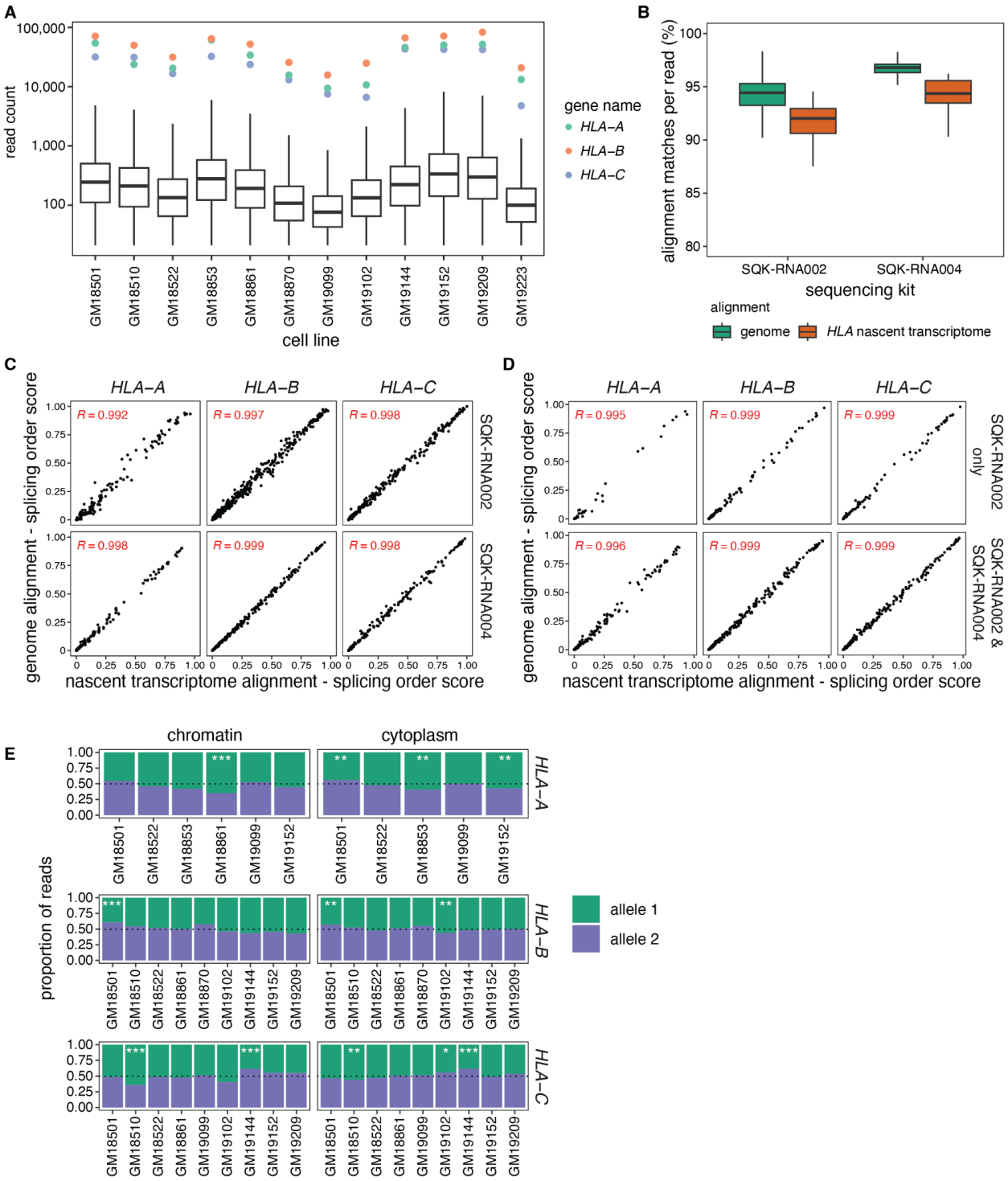

Figure S7. Allele-specific analysis of *HLA* class I transcripts (related to Figure 5). A) Boxplots representing the read count distribution of genes with at least 20 reads in each LCL. Colored dots indicate the read counts for *HLA* class I genes. B) Percent of nucleotides per read matching the reference sequence for alignment to the genome or to the *HLA* nascent transcriptome, as a function of the Oxford Nanopore Technologies sequencing chemistry used. C) Correlation in splicing order

scores obtained from aligning reads to the genome or to the nascent transcriptome (see Methods) for individual replicates, separated according to the sequencing chemistry used. Each dot represents one splicing order on one allele. D) Same as C), but for splicing order scores obtained from merged replicates, separated based on whether all replicates were sequenced with SQK-RNA002 or with a combination of the two chemistries. E) Proportion of reads mapping to each allele of *HLA-A, HLA-B* and *HLA-C* in chromatin and cytoplasm. Only LCLs for which allele-specific RNA abundance met the required coverage thresholds are shown. The number of reads mapping to each allele on chromatin-associated or cytoplasmic RNA was compared using Qllelic (Mendelevich et al. 2021) or a two-sided binomial test, respectively. ***: p-value < 0.001 and proportion of allele 1 reads < 0.4 or > 0.6; **: p-value < 0.01 and proportion of allele 1 reads < 0.45 or > 0.55; *: p-value < 0.05 and proportion of allele 1 reads < 0.45 or > 0.55.

Supplemental Methods

Cell fractionation and RNA extraction

Cellular fractionation was performed as described in steps 6 to 19 of Drexler et al. 2021. For each LCL, 120-200 million cells were collected and centrifuged for 5 minutes at 500g, then split into aliquots of 10 million cells per cellular fractionation reaction (12-20 reactions total). At the end of the cellular fractionation, 3-4 chromatin pellets or 200 uL of cytoplasmic fraction were resuspended in 1 mL of QIAzol lysis reagent and stored at -80ºC. For RNA extraction, samples were thawed for 2 min at 65ºC, followed by addition of 200 uL of chloroform. Samples were vortexed for 15 seconds and centrifuged at 12,000g for 15 minutes at 4°C. The aqueous phase was transferred to a new tube, mixed with an equal volume of isopropanol, incubated at room temperature for 10 minutes and centrifuged at 12,000g for 10 minutes at 4°C. Pellets were washed twice with 1 mL of 75% ethanol, centrifuged at 7,500g for 5 minutes at 4ºC and resuspended in nuclease-free water. Cell fractionations were verified by western blot as described in (Mayer and Churchman 2016) with antibodies against RNA polymerase II CTD phospho Ser2 (Active Motif, 61984) and GAPDH (ThermoFisher MA5-15738).

Haplotype-aware alignment

Phased genotypes from LCLs were downloaded in VCF format from the International Genome Sample Resource (https://www.internationalgenome.org/data, GRCh38, release 20170504, downloaded in May 2019). We used the script process_vcf.sh from LORALS (Glinos et al. 2022) (https://github.com/LappalainenLab/lorals) to obtain per-individual VCF files that included only heterozygous SNPs. We then used the LORALS script make_new_vcf.sh and the initial alignments from chromatin-associated RNA samples (all replicates merged together for each cell line) to correct phased haplotypes in the VCF files and to generate two haplotype-specific reference genome FASTA files per individual. The LORALS script hap_aligner.sh was then used to align reads (from chromatin-associated or cytoplasmic RNA) to these two genomes using minimap2 (Li 2018) and to select the highest scoring alignments as previously described (Glinos et al. 2022) to create a final BAM file from the two haplotype-aware alignments. The command extractHAIRS from HapCUT2 v1.3.3 (Edge et al. 2017) with options --nf 1 and --ont 1 and the script https://github.com/nanopore-wgs-consortium/NA12878/blob/master/nanopore-human-transcriptome/scripts/ase.py from Workman et al. 2019 were used to assign reads to their alleles of origin, requiring that at least two heterozygous SNPs be present. Reads were classified as "not determined" if less than two informative variants were present or less than 75% of identified variants agreed with one of the two alleles.

Determination of the excision status of introns

The excision status of introns was determined using our previously published method (Drexler et al. 2021). Reads that overlap introns from the hg38 RefSeq annotation were extracted using BEDTools intersect (Quinlan and Hall 2010) with strand specificity. For each read/intron pair, we extracted the portion of the CIGAR string corresponding to the 50 nucleotides surrounding the 5′ and 3′ splice sites (SS) of the intron. Introns were classified as "excised" if the CIGAR string "N" (splicing event) started and ended within 50 nt of the 5′ and 3′SS and the size of this splicing event was within 10% of the annotated intron size. Introns were classified as "not excised" if there was no evidence of the CIGAR string "N", and there was more than 50% coverage (CIGAR string "M") in the 50 nt surrounding each splice site and more than 75% coverage within the region of

intron that the read mapped to. Intron/read pairs where the read started within the intron (no coverage over the 5′SS) and met these criteria for the 3′SS were classified as "not excised". Introns were classified as "skipped splice sites" if a splicing event (CIGAR string "N") overlapped with greater than 50% of the 50 nt surrounding the 5′SS and/or the 3′SS and the portion of the intron within the splicing event was within 10% of the annotated intron size. These included alternative 5′ or 3′ SSs that are in an annotated exon and introns flanking skipped exons. Intron/read pairs that did not meet any of these criteria were classified as "undetermined". Only introns with the splicing statuses "excised" or "not excised" were considered for splicing order analyses. To identify the proportion of partially spliced reads (Fig. S1B), all reads spanning two introns or more were considered. Reads that contained only "not excised" introns were considered "all unspliced", reads that contained both "not excised" and "excised" introns were considered "partially spliced", and reads that contained "excised introns" and no "not excised" introns were considered "all spliced".

Computation of splicing order

Splicing order plots for groups of three introns were produced in R (R Core Team 2021) with ggplot2 (https://ggplot2.tidyverse.org) with the following command:

```
ggplot(aes(x=splicing_level, y=new_intron_spliced)) +
geom_line(aes(group=order_name, size=score, alpha=score))
```

Splicing order for intron pairs was computed as described in Drexler et al. 2020 using the excision statuses determined as outlined above. Reads spanning two consecutive introns with different excision statuses (one excised, one not excised) were extracted. The frequency of reads with the upstream intron removed or with the downstream intron removed was calculated for each allele. Intron pairs were used for downstream analyses if the number of reads per allele was higher than 10 and greater than twice the number of reads for which the allele could not be determined. To identify statistically significant differences in allelic splicing order for intron pairs, we compared the distribution of intermediate isoform (partially spliced) read counts between alleles using a two-sided Fisher's exact test, followed by multiple testing correction with the Benjamini-Hochberg method. Intron pairs were considered to display allele-specific splicing when FDR < 0.05 and the absolute difference in the frequency of the upstream intron excised first between alleles was ≥ 0.1.

Comparison of allelic and allele-agnostic splicing orders

Allele-agnostic splicing orders were computed in the same way as allelic splicing orders, but using all reads (assigned to allele 1, allele 2, or undetermined). Comparison of allelic and allele-agnostic splicing orders showed high correlation (Fig. S2B-D). The small number of outliers was mainly due to the absence or over-representation of some intermediate isoforms in the allelic splicing orders because of the presence of SNPs in only some introns of the group (Fig. S2C-D). Nevertheless, since the same SNPs are used to differentiate between alleles, this limitation did not prevent us from investigating splicing order differences between alleles.

Identification of splicing order differences between alleles

We computed the Euclidean distance between the two vectors using the math.dist function in Python. To identify the proper threshold for splicing order differences between alleles, we computed the Euclidean distance for the same allele between biological or technical replicates of the same cell line for all intron groups that met coverage thresholds (N=970 intron group/LCL combinations). We calculated the interquartile range (IQR) of the combined distribution from all cell lines and defined the threshold for allele-specific splicing order as IQR * 1.5 = 0.379. Intron

groups were considered to display allele-specific splicing when chi-square contingency test FDR < 0.05 for one or both splicing levels and Euclidean distance between allelic splicing order vectors > 0.379. We found that the majority of intron groups with chi-square test p-value > 0.05 also had low Euclidean distances, while a smaller number had a Euclidean distance higher than our threshold for identifying allele-specific orders (>0.379). The latter group tended to have lower coverage (Fig. S4E), suggesting that lower counts lead to larger variation between alleles when calculating the Euclidean distance, which may be false positive differences that are filtered out by also considering statistical significance through the chi-square test. By contrast, approximately half of intron groups with p-value < 0.05 had Euclidean distance >0.379. The other half had low Euclidean distances but higher read coverage (Fig. S4E), consistent with high sample sizes yielding smaller p-values despite small effect sizes (Lin et al. 2013). This analysis demonstrates the pertinence of using both the chi-square test p-value and the Euclidean distance as a measure of effect size to identify robust splicing order changes.

Classification of splicing order differences

For each intron group with significant allele-specific splicing order, the top ranked splicing order was extracted for each allele and separated into individual introns. If two consecutive positions (first and second or second and third) were different between alleles while the remaining position (first or third) was the same (e.g. 1->2->3 vs. 2->1->3), the intron group was categorized as "reversal of first or last two positions". If the first and last positions were different between alleles while the middle position was the same (e.g. 1->2->3 vs. 3->2->1), the intron group was categorized as "reversal of first and last positions". If all three positions were different (e.g. 1->2->3 vs. 2->3->1), the intron group was categorized as "reversal of 3 positions". If all three positions were the same, the intron group was categorized as "different score for same order". The number of intron groups per category was counted and represented as in Fig. 1D.

Characterizing genetic variants associated with splicing order

For SNPs located in splice sites, the impact on splice site strength was assessed with MaxEnt (Yeo and Burge 2004). sQTL SNPs from (Li et al. 2016) were downloaded from http://eqtl.uchicago.edu/jointLCL/ on 2019/06/19 and SNP coordinates were converted from hg19 to hg38 using CrossMap (Zhao et al. 2014). sQTL SNPs identified in lymphoblastoid cell lines from the GTEx Consortium v8 were obtained from GTEx_Analysis_v8_sQTL.tar, which was downloaded from https://www.gtexportal.org/home/downloads/adult-gtex/qtl. SNPs associated with splicing order and sQTL SNPs with the same coordinates were considered to overlap.

Identification of allele-specific alternative splicing (AS) events

Using the chromatin-associated dnRNA-seq data from each LCL, two BAM files were created corresponding to reads mapping to each allele from the haplotype-aware alignment. Transcript isoforms were discovered and quantified from each BAM file using ESPRESSO (v1.4.0) with default settings and human genome reference and annotation [Ensembl GRCh38 (release-86)]. Transcripts with less than 10 reads on both alleles were excluded. Known and novel transcripts overlapping with genes with allele-specific splicing orders were extracted. For each pair of transcripts in each gene, the abundance between alleles was compared using a two-sided Fisher's exact test. Multiple testing correction was performed using the Benjamini-Hochberg method. Transcripts were considered to have statistically significant allele-specific AS if the corrected p-value was < 0.05 and the odds ratio was < 0.5 or > 2. To identify the type of AS event between

transcripts, we used the script classify_isoform_differences.py from rMATS-long v1.0.0 (https://github.com/Xinglab/rMATS-long) where the parameters –main-transcript-id and –second-transcript-id were the pairs of transcript isoforms considered to display allele-specific AS. AS categories "intron retention", "alternative first exon", "alternative last exon" and "complex" were excluded from further analyses. AS events in *HLA* genes were also excluded due to frequent genome alignment artifacts (see below) because of the small size of exon 6, leading to erroneous novel transcripts in ESPRESSO. For each intron group displaying allelic splicing orders, we defined a window corresponding to the start of the first intron -50 nt to the end of the last intron +50 nt and identified overlapping AS events using a custom Python script.
For comparing allele-specific splicing to known sQTLs, we used sQTLs identified in lymphoblastoid cell lines from the GTEx Consortium v8, obtained from GTEx_Analysis_v8_sQTL.tar downloaded from https://www.gtexportal.org/home/downloads/adult-gtex/qtl. For each cell line, we identified sQTL SNPs that were heterozygous using the phased VCF files from LORALS. We extracted the corresponding sQTL introns and identified overlapping reads using BEDTools intersect. For introns with more than 20 reads per allele, we compared alternative splicing between alleles using an intron-centric approach, as previously described (Choquet et al. 2023). Intron excision status was determined as described above. We compared the number of reads with "excised" vs. "skipped" or "undetermined" statuses on each allele using a two-sided Fisher's exact test. Multiple testing correction was performed using the Benjamini-Hochberg method. The ratio of excised reads divided by the total number of reads (excluding "not excised" reads) was computed. Introns with an adjusted p-value < 0.05 and a difference between allelic ratios ≥ 0.1 were considered to display allele-specific alternative splicing. The same approach was used to analyze alternative splicing of *HLA-C*.

Allele-specific transcript abundance
Reads were assigned to the genes they mapped to using BEDTools intersect -s -F 0.5 -wo with the BAM file from haplotype-aware alignment and the hg38 RefSeq gene coordinates in BED format as input. The output file was merged with read to allele assignments from above and the number of reads per gene and allele was counted. Genes were included in this analysis if 1) they had more than 20 reads assigned to either allele, and 2) the number of reads assigned to each allele was at least two times higher than the number of undetermined reads. The allelic ratio was defined as the number of reads assigned to allele 1 divided by the number of reads assigned to alleles 1 or 2. We used Qllelic (Mendelevich et al. 2021) with 2 or 3 replicates per sample to identify genes with allelic imbalance in chromatin-associated RNA. We used a two-sided binomial test with $p=0.5$ to identify genes with allelic imbalance in cytoplasmic RNA due to the absence of replicates. For both approaches, multiple testing correction was performed with the Benjamini-Hochberg method. Genes with an adjusted p-value < 0.05 and an allelic ratio < 0.4 or > 0.6 were considered to have unbalanced or skewed RNA abundance. To assess the correlation in allelic ratio between subcellular compartments (Fig. 3A and S5A), only genes for which the ratio could be computed in both chromatin and cytoplasm were considered.

Comparison of 3′-end positions
The BAM file was converted to a BED file using BEDTools bam_to_bed (Quinlan and Hall 2010). For each read, the genomic position of the 3′-end was recorded in a strand-specific manner (end of the read on positive strand, start of the read on negative strand). The same coverage thresholds

and treatment of replicates were used as for poly(A) tail length analysis. Distributions of 3′-end positions were compared using a two-sided Wilcoxon rank-sum test. Multiple testing correction was performed with the Benjamini-Hochberg method. Genes with adjusted p-value < 0.05 and at least 10 nt between the mean 3′-end position of each allele were considered to have statistically significant differences in 3′-end position (i.e. APA).

Visualization

The coverage track in Fig. 1A was produced with pyGenomeTracks (Lopez-Delisle et al. 2021). The package ComplexUpset (https://github.com/krassowski/complex-upset) was used to represent the overlap of allele-specific features in each gene (Fig. 4E, S6E).

Alignment to the HLA nascent transcriptome

In *HLA* class I genes, the short size of exon 6 led to artifacts suggesting skipping of this exon following alignment to the reference genome. Though we excluded any reads with exon skipping from splicing order analysis, to ensure that this did not impact splicing order measurements,, we re-aligned reads to personalized *HLA* class I nascent transcriptomes and computed splicing order. For each *HLA* class I gene (*HLA-A*, *HLA-B*, *HLA-C*), we extracted the hg38 intron and exon annotations. Genomic coordinates were converted to coordinates from the start of each gene. We identified all possible combinations of retained introns, from 0 to 7 (completely unspliced), and re-constructed the nascent isoform sequences corresponding to each combination for each LCL using the haplotype-aware genome reference sequences generated above. From the haplotype-aware genome alignments, we extracted reads mapping to *HLA* class I loci and aligned them to the haplotype-aware *HLA* nascent transcriptomes using hap_aligner.sh as described above but with the minimap2 parameter -ax map-ont instead of -ax splice -uf -k14. For splicing order computation, the same strategy was used as above. For measuring the proportion of alignment matches per read in Fig. S7B, the CIGAR string was extracted with pysam (https://github.com/pysam-developers/pysam) (Li et al. 2009). The edit distance (NM tag) was subtracted from the total of matches (M) and insertions (I) and this number was divided by the total number of matches and insertions in the read (M + I - NM / M + I * 100). Our analysis showed a very strong correlation (R > 0.99) between splicing orders calculated following genome or transcriptome alignment, irrespective of the proportion of alignment mismatches, which improved with updated dnRNA-seq chemistry (SQKRNA-002 vs. SQK-RNA004) (Fig. S7B-D).

HLA typing

To identify the type of each *HLA-A, HLA-B* and *HLA-C* allele in our dataset, we used *HLA* typing data from (Abi-Rached et al. 2018) available through the 1000Genomes project FTP: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/. After retrieving the *HLA* types for the 12 LCLs in our study, we extracted the sequences from the corresponding *HLA* alleles from the IPD-IMGT/HLA database (Robinson et al. 2015). For each LCL, we assembled a FASTA file composed of the sequences of all possible variants of their assigned *HLA* alleles (hereafter "personalized *HLA* database"). We used BEDTools (Quinlan and Hall 2010) getfasta to obtain the *HLA* sequences from the two haplotype-specific reference genomes for each LCL generated above from the dnRNA-seq data. These haplotype-specific *HLA* sequences were aligned to the corresponding personalized *HLA* database using minimap2 and the option --secondary=no. For each *HLA* class I gene, we used the primary alignment to assign each allele to one of the two known *HLA* types per individual.

References

Abi-Rached L, Gouret P, Yeh J-H, Di Cristofaro J, Pontarotti P, Picard C, Paganini J. 2018. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS One* **13**: e0206512.

Choquet K, Baxter-Koenigs AR, Dülk S-L, Smalec BM, Rouskin S, Churchman LS. 2023. Pre-mRNA splicing order is predetermined and maintains splicing fidelity across multi-intronic transcripts. *Nat Struct Mol Biol* **30**: 1064–1076.

Drexler HL, Choquet K, Churchman LS. 2020. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol Cell* **77**: 985–998.e8.

Drexler HL, Choquet K, Merens HE, Tang PS, Simpson JT, Churchman LS. 2021. Revealing nascent RNA processing dynamics with nano-COP. *Nat Protoc* **16**: 1343–1375.

Edge P, Bafna V, Bansal V. 2017. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res* **27**: 801–812.

Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K, et al. 2022. Transcriptome variation in human tissues revealed by long-read sequencing. *Nature* **608**: 353–359.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lin M, Lucas H, Shmueli G. 2013. Research commentary - too big to fail: Large samples and the p-value problem. *Inf Syst Res* **24**: 906–917.

Li YI, van de Geijn B, Raj A, Knowles DA, Petti AA, Golan D, Gilad Y, Pritchard JK. 2016. RNA splicing is a primary link between genetic variation and disease. *Science* **352**: 600–604.

Lopez-Delisle L, Rabbani L, Wolff J, Bhardwaj V, Backofen R, Grüning B, Ramírez F, Manke T. 2021. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics* **37**: 422–423. http://dx.doi.org/10.1093/bioinformatics/btaa692.

Mayer A, Churchman LS. 2016. Genome-wide profiling of RNA polymerase transcription at nucleotide resolution in human cells with native elongating transcript sequencing. *Nat Protoc* **11**: 813–833.

Mendelevich A, Vinogradova S, Gupta S, Mironov AA, Sunyaev SR, Gimelbrant AA. 2021.

Replicate sequencing libraries are important for quantification of allelic imbalance. *Nat Commun* **12**: 3370.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. http://dx.doi.org/10.1093/bioinformatics/btq033.

R Core Team. 2021. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/.

Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. 2015. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res* **43**: D423–31.

Workman RE, Tang AD, Tang PS, Jain M, Tyson JR, Razaghi R, Zuzarte PC, Gilpatrick T, Payne A, Quick J, et al. 2019. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat Methods* **16**: 1297–1305.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**: 377–394.

Zhao H, Sun Z, Wang J, Huang H, Kocher J-P, Wang L. 2014. CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**: 1006–1007.