nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-019-0166-3

RNA proximity sequencing reveals the spatial organization of the transcriptome in the nucleus

Jörg Morf[1*], Steven W. Wingett[1,2], Irene Farabella[3,4], Jonathan Cairns[1,12], Mayra Furlan-Magaril[5], Luis F. Jiménez-García[6], Xin Liu[7], Frank F. Craig[7], Simon Walker[8], Anne Segonds-Pichon[2], Simon Andrews[2], Marc A. Marti-Renom[3,4,9,10] and Peter Fraser[1,11]

The global, three-dimensional organization of RNA molecules in the nucleus is difficult to determine using existing methods. Here we introduce Proximity RNA-seq, which identifies colocalization preferences for pairs or groups of nascent and fully transcribed RNAs in the nucleus. Proximity RNA-seq is based on massive-throughput RNA barcoding of subnuclear particles in water-in-oil emulsion droplets, followed by cDNA sequencing. Our results show RNAs of varying tissue-specificity of expression, speed of RNA polymerase elongation and extent of alternative splicing positioned at varying distances from nucleoli. The simultaneous detection of multiple RNAs in proximity to each other distinguishes RNA-dense from sparse compartments. Application of Proximity RNA-seq will facilitate study of the spatial organization of transcripts in the nucleus, including non-coding RNAs, and its functional relevance.

The spatial organization of nucleic acids in cell nuclei is critical for gene expression and ultimately cell physiology[1]. Different scales of the spatial organization of DNA and chromatin have been analysed; from promoter–enhancer interactions to megabase-sized topological association domains (TADs) and large-scale, higher-order folding of chromosomes into distinct compartments composed of transcriptionally active or inactive regions[2–4]. However, there is little understanding of where specific transcripts are synthesized, processed and/or sequestered in relation to nuclear landmarks. Large parts of transcriptomes have been spatially resolved in tissues[5] and single cells[6–8] with imaging techniques, but these methods do not infer colocalization or spatial associations between different RNA molecules. Pairwise probing of RNA–RNA interactions has thus far been limited to direct base-paired contacts[9] or to short-range distances between RNA ends allowing enzymatic ligations[10–13]. We therefore devised a widely applicable, massive-throughput method that determines spatial associations between pairs or groups of transcripts irrespective of the nature of their interaction to provide a functional readout of transcriptional genome activity in the nucleus.

Proximity RNA-seq uniquely barcodes RNAs in millions of subnuclear particles in parallel by a simple, rapid vortexing step combining fragmented nuclear particles with barcoded beads in a water-in-oil emulsion (Fig. 1a and Supplementary Fig. 1). Barcoded complementary DNAs are then sequenced to enable the reconstruction of proximities between chromatin-associated, nascent RNAs and non-coding RNAs in nuclei. We use Proximity RNA-seq to identify, characterize and map specific transcript families in relation to a major recognizable nuclear landmark, providing a cytological spatial map of RNAs in cells.

Results

Massive-throughput barcoding in emulsion. We first developed an on-bead PCR protocol in emulsion to uniquely barcode up to 1 billion beads individually for an experiment. We used conditions that favor the amplification of a single synthesized DNA template containing a 26-base-long random barcode on each bead in the emulsion[14,15] (Fig. 1b). The encapsulation of a single barcode with one bead can be approximated by diluting templates and beads sufficiently before emulsification. To estimate the fraction of singly barcoded beads, we first used two barcode templates of defined sequences (Supplementary Table 1) for PCR amplification on beads. Fluorescent probes were then hybridized to the amplified barcodes on beads before flow cytometry analysis and comparison with expected Poisson distributions (Fig. 1c,d). To optimize the yield of beads with individual random barcodes, we chose encapsulation conditions and template amounts according to a Poisson model that aimed at 50% barcoded beads, of which around 70% were covered with copies of a single barcode and 30% had copies of multiple barcodes. After PCR, barcode copies on beads were extended with 15 random bases to generate reverse transcription (RT) primers (Fig. 1b,e). Subnuclear particles from SH-SY5Y human neuroblastoma cells were obtained by minimal sonication to disrupt nuclei isolated from chemically crosslinked cells. After each sonication cycle, disruption of nuclei was examined by microscopy. We used ethylene glycol-bis(succinimidylsuccinate) (EGS), with a 16 Å spacer between reactive groups, in combination with formaldehyde for crosslinking, with the aim to increase the fraction of particles containing multiple RNA molecules. After sonication of nuclei, all the homogenate was encapsulated into droplets without prior centrifugation. In-droplet reverse transcription of

[1]Laboratory of Nuclear Dynamics, Babraham Institute, Cambridge, UK. [2]Bioinformatics, Babraham Institute, Cambridge, UK. [3]CNAG-CRG, Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain. [4]CRG, Centre for Genomic Regulation, Barcelona, Spain. [5]Departamento de Genética Molecular, Instituto de Fisiología Celular, Universidad Nacional Autónoma de México, Mexico City, Mexico. [6]Department of Cell Biology, Faculty of Sciences, National Autonomous University of Mexico, Mexico City, Mexico. [7]Sphere Fluidics Limited, Babraham Research Campus, Cambridge, UK. [8]Imaging Facility, Babraham Institute, Cambridge, UK. [9]Universitat Pompeu Fabra, Barcelona, Spain. [10]Catalan Institution for Research and Advanced Studies, Barcelona, Spain. [11]Department of Biological Science, Florida State University, Tallahassee, FL, USA. [12]Present address: Quantitative Biology department, Discovery Sciences, IMED Biotech Unit, AstraZeneca, Cambridge, UK. *e-mail: jorg.morf@gmail.com

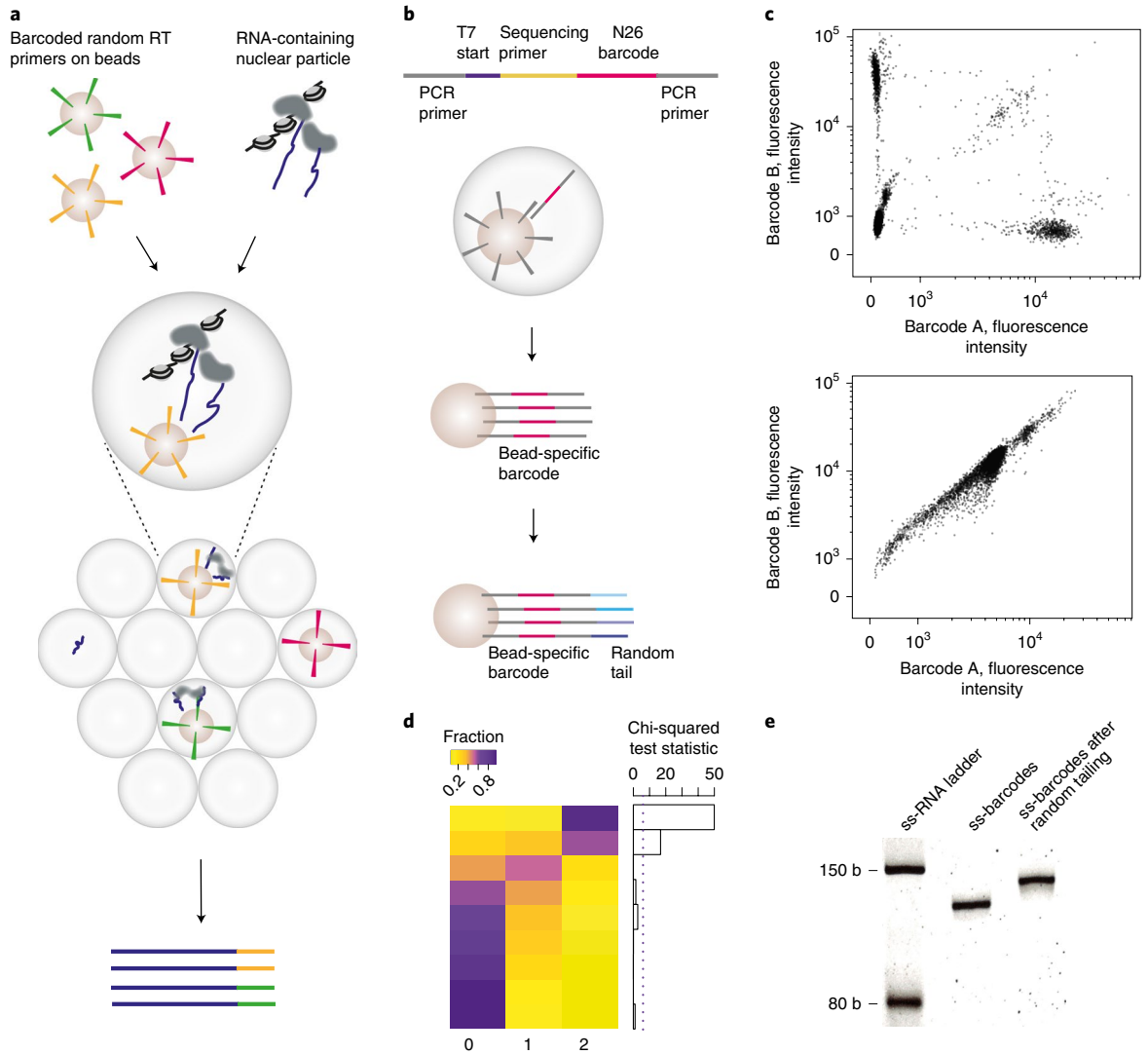

Fig. 1 | Proximity RNA-seq. a, Massive-throughput barcoding by RT of RNA-containing particles in water-in-oil droplets. **b**, Barcoding of magnetic beads with immobilized primers and diluted random DNA templates by emulsion PCR. Barcodes on beads were end-tailed with random nucleotides to subsequently serve as RT primers. **c**, To control barcoding and emulsion integrity, two barcodes of known sequence were amplified on beads in emulsion or solution prior to hybridization with complementary fluorescent probes and fluorescence-activated cell sorting (FACS) analysis to count empty, single and mixed barcoded beads. Top panel, PCR in emulsion; bottom panel, PCR in solution. Axes specify fluorescence signals of hybridized probes. **d**, Different two-barcode experiments, rows, were ordered according to increasing fractions of non-barcoded beads (yellow, low fraction; purple, high fraction). Fractions of beads containing no, one or both barcodes (columns) were compared with expected Poisson distributions (chi-squared test statistic, dashed line indicates $P = 0.05$). $n = 10{,}000$ beads per experiment were analysed. **e**, Acrylamide gel of single-stranded barcodes (ss-barcodes) before (lane 2) and after (lane 3) the addition of 15 random nucleotides. Single-stranded RNA ladder (ss-RNA ladder) was run in lane 1 ($n = 2$).

RNAs in crosslinked particles with barcoded beads followed by cDNA library amplification (Supplementary Fig. 1) resulted in a fragment length suitable for Illumina sequencing.

Proximity RNA-seq characterization and validation. Sequencing and mapping of RNAs in nuclear-enriched particles resulted in a fourfold increase in intron-to-exon read ratio compared with ribosomal RNA-depleted RNA-seq. In addition, the characteristic overrepresentation of reads at the 5′ ends in introns compared with the 3′ ends[16] demonstrated a clear enrichment for nascent transcripts (Fig. 2a,b). We trimmed three bases on either end of the barcode to account for offsets by a few bases of otherwise identical barcodes. The resulting barcode length of 20 random bases corresponds to a theoretical barcode complexity of 10^{12}, which exceeded an estimated 10^9 nascent RNA molecules in the input material. Multiple reads with the same barcode and mapping to the same transcript were dubbed a proxy read and counted as one transcript observation (Supplementary Fig. 2). Up to 25% of proxy reads were cobarcoded with proxy reads mapping to other transcripts identifying spatial RNA associations (Supplementary Fig. 3). Proximity RNA-seq benefits from sequencing into saturation to increase the number of RNAs identified on beads and cobarcoded proxy reads (Supplementary Fig. 4).

To validate spatial associations, we first analysed cobarcoding events of the abundant small RNA U3, which resides in the nucleolus together with hundreds of other small non-coding, nucleolar RNAs (snoRNAs)[17]. We found overwhelmingly significant cobarcoding of snoRNAs with U3 compared with the non-snoRNA transcriptome (Mann–Whitney U-test: $P = 4 \times 10^{-56}$; Cliff's delta effect size: 0.9; Fig. 2c). In contrast, control experiments using multi-barcoded

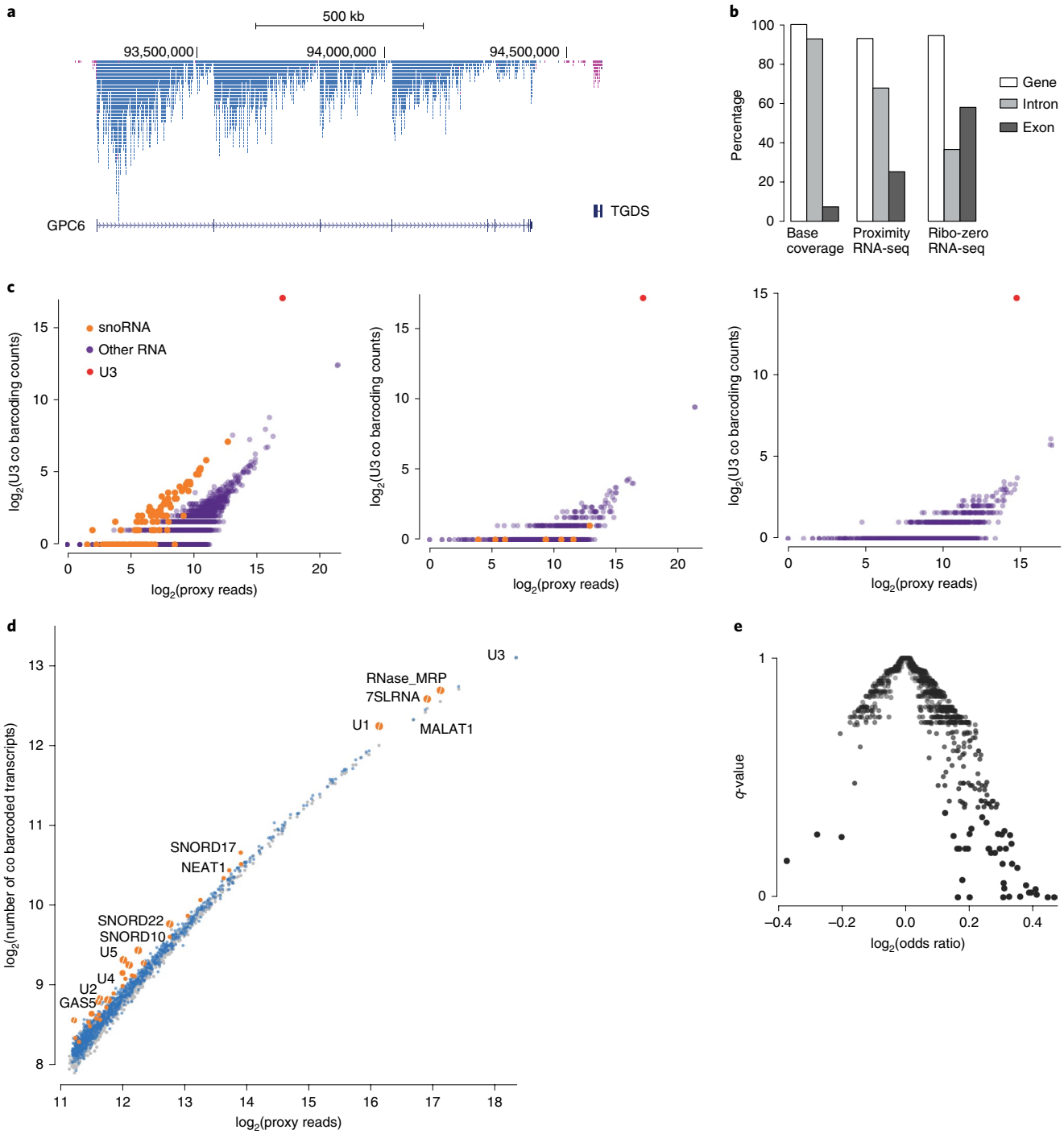

Fig. 2 | RNA cobarcoding and proximal transcriptomes. a, Proximity RNA-seq reads mapping to GPC6 (blue) and adjacent transcripts on the opposite strand (pink) illustrate both high intronic read densities and the saw-tooth read pattern along introns. **b**, Fractions in percentage of reads in transcript features. Proximity RNA-seq was compared with total RNA-seq after ribosomal depletion[39] and the base coverage of exons, introns and genes in the genome (set as 100), respectively. **c**, Counts of cobarcoding events involving U3 RNA plotted against the number of proxy reads of the other RNA. Left panel, crosslinked sample with uniquely barcoded beads; middle panel, control using crosslinked sample and randomly barcoded beads; right panel, control with purified RNA after crosslink reversal and uniquely barcoded beads. Of note, most snoRNAs are not plotted for control data, as no cobarcoding with U3 was detected. **d**, For the top 1,000 RNAs (combined data from $n = 3$ independent experiments), the number of cobarcoded transcripts was plotted against the number of proxy reads per RNA. Gray, randomized; blue, observed; orange, RNAs with high complexity of cobarcoded transcripts ($P < 0.01$); orange with white slash, $q < 0.05$ as derived in **e**. **e**, Transcripts with more complex composition of cobarcoded transcripts than expected at random were identified by Fisher's exact test, two-sided, and FDR adjustment (x axis, odds ratio with positive values indicating high complexity of cobarcoded transcripts).

beads produced by PCR amplification in a droplet-free solution showed essentially no cobarcoding between U3 and snoRNAs. Similarly, using crosslink-reversed and purified RNA and uniquely barcoded beads also showed no cobarcoding between U3 and snoRNAs. We next compared the crosslink-reversed control, the standard crosslinked condition used throughout the manuscript and a

prolonged crosslinked sample to assess the effects of sample preparation on particle encapsulation, RNA read counts and nuclear proximities. Global correlation in transcript abundance as measured by Proximity RNA-seq between standard crosslinked and crosslink-reversed samples (Spearman's rank correlation coefficient: 0.8) was found to be similarly high to that between replicates of crosslinked samples (Supplementary Fig. 3). Equal transcript abundances in both datasets suggested unbiased encapsulation of crosslinked RNA particles into droplets. As exemplified by U3 (Fig. 2c), the crosslink-reversed sample lacked information on RNA proximities, and global correlation in pairwise transcript cobarcoding between crosslinked and crosslink-reversed samples was poor (Spearman's rank correlation coefficient: 0.29). We then compared the sample with prolonged fixation with the standard sample preparation. The same number of sonication cycles was used for both samples to disrupt nuclei. We found that the fraction of ribosomal RNA reads increased from 43% in standard conditions to 87% with prolonged crosslinking. Of note, nuclei contribute less than the cytoplasm to a total of around 90% rRNA in whole cells[18,19]. Intronic reads decreased from 74% to 47% and exonic reads increased from 18% to 47% with prolonged crosslinking compared with standard conditions (Supplementary Fig. 5). We conclude that prolonged crosslinking copurified larger amounts of cytoplasmic RNA compared with standard crosslinking, which hampers the analysis of nuclear RNA organization. Nuclear particle preparation is therefore likely a trade-off between a high crosslinking efficiency, to increase cobarcoding of transcripts at the cost of cytoplasmic RNA contamination, and aiming for a cleaner preparation of nuclei with less crosslinking and less frequent cobarcoding of transcripts. Using our standard sample preparation in SH-SY5Y cells, high intron and modest rRNA contents suggested that the conditions were suitable to probe nuclear RNA proximities. Furthermore, correlations in pairwise transcript cobarcoding between different standard crosslinked replicates ranged between 0.71 and 0.83 (Spearman's rank correlation coefficient, Supplementary Fig. 3). These results demonstrate that our method reproducibly detects RNAs present in individual subnuclear particles. However, we cannot rule out the possibility that other conditions may be more suitable.

To calculate the significance of spatial associations between transcripts while taking their vastly different abundances into account, we randomized pairings of proxy reads and their barcodes 100,000 times to obtain simulated cobarcoding counts for pairs of transcripts (see Methods, Supplementary Fig. 3). Observed cobarcoded pairwise RNA proximities were compared with simulated counts to determine statistical significance (Supplementary Tables 2,3). These analyses verified known contacts between 18S and 28S rRNAs, MALAT1 and U1 (ref. [20]) and between spliceosomal RNAs, respectively ($q < 0.002$). To further test the analysis pipeline, we performed Proximity RNA-seq using a mixture of particles from different species. To a standard input of human particles equivalent to 50 ng RNA we added the same amount of fruit fly particles to obtain a conservative estimate of false positive RNA proximities. We first selected barcode groups with exactly two transcripts and found that 20% of all two-transcript barcode groups consisted of a fly and a human transcript (Supplementary Fig. 6). We then compared the whole dataset with randomizations and estimated the false positive rates of pairwise, inter-species RNA proximities. *P*-value cutoffs for pairwise RNA proximities at 0.01, 0.05 and 0.1 resulted in 0%, 2% and 6.4% false positive rates, respectively. This shows that Proximity RNA-seq and the analysis pipeline produce few false positives even in conditions with considerable species mixing in droplets.

Main RNA compartments and transcript positioning in nuclei. Next, we analysed the number of unique transcripts cobarcoded once or multiple times and irrespective of the significance of pairwise associations, with a given RNA molecule of the 1,000 highest expressed transcripts. We found that RNAs present at higher levels had higher cobarcoded transcript counts (Fig. 2d and Supplementary Table 4). Two factors likely underlie this trend. First, in cells, highly expressed RNAs have more opportunities to encounter other transcripts, and second, in droplets, abundant RNAs are more often cobarcoded with transcripts from distinct particles, according to the Poisson distribution of particles into droplets, thereby further increasing the number of unique cobarcoded transcripts. Notably, we identified RNAs cobarcoded with more unique transcripts than would be expected by their abundance. Such RNAs were often non-coding, rather than protein-coding, and included spliceosomal (U1, U2, U4 and U5), paraspeckle (NEAT1) and nucleolar (snoRNAs) transcripts (Fig. 2d,e). The more complex compositions of cobarcoded transcripts likely reflected transcriptome-wide functions in intron excision and exon splicing in the case of spliceosomal RNAs and supported the notion of some non-coding RNAs as spatial organizers or markers of larger RNA groups in nuclei.

In DNA contact assays, close distance on the linear genome largely increases contact strength between genomic regions. Significant but lower contact strength between regions further apart is indicative of higher-order chromatin folding. We therefore assumed that pairs of nascent RNAs, although not physically linked like DNA of the same chromosome, would show increased spatial association when the distance between their genes was small. Using randomization-derived *P* values for RNA pairs, we found higher levels of associations between transcripts encoded by genes located nearby in the linear genome than between transcripts from genes further apart or on different chromosomes (Supplementary Fig. 7). This shows that one factor that specifies spatial associations of a nascent transcript with other RNAs is linear genome proximity of the respective genes.

We then asked whether larger groups of nuclear-retained, non-coding and nascent coding transcripts preferentially associate with each other and thereby partition the nuclear transcriptome. Principal component analysis (PCA) on *P* values of pairwise associations between any RNA and the top 100 connected transcripts identified two main compartments for RNA in nuclei. Principal component 2 (PC2) separated snoRNAs as well as a set of protein-coding transcripts from the bulk of mostly protein-coding transcripts (Fig. 3a, Supplementary Table 5). We dubbed this transcript group, which is highly enriched for nucleolar transcripts, compartment I, based on the prevalent RNA polymerase I activity for rRNA synthesis within the nucleolus. Accordingly, compartment II was named after RNA polymerase II, which is predominant in the nucleoplasm (Fig. 3b). We then performed dual RNA fluorescence in situ hybridization (FISH) using U3 as a nucleolar marker and sets of 24 probes hybridizing to intronic sequences of 13 different RNAs (Fig. 3c,d). The results confirmed that nascent transcripts from compartment I are indeed preferentially synthesized in close proximity to the nucleolus, whereas nascent transcripts from compartment II are transcribed in the nucleoplasm at greater distances from the nucleolus. This validates the accuracy of Proximity RNA-seq to predict spatial distances between different transcripts and a major nuclear structure, the nucleolus, providing a 'magnetic north' reference point for genome function that is lacking in chromosome conformation capture (Hi-C) data. On average, median distances between the edge of the nearest nucleolus and transcripts of compartment I were $0.82 \pm 0.24\,\mu m$ (mean ± s.d.). Compartment II transcripts, on the other hand, were found at on average twofold increased distances from the nucleoli ($1.68 \pm 0.37\,\mu m$). Our findings suggest that the perinucleolar region, which has previously been associated with transcriptionally silent, compact chromatin[3,21], expresses a specific subgroup of RNAPII-transcribed genes.

We next analysed features of RNAs grouped into eight quantiles according to their relative nuclear position derived from PC2 values (Fig. 3b,e–g). Compartment I (quantiles 1–3) contained 370 (70%) and compartment II (quantiles 4–8) 784 protein-coding

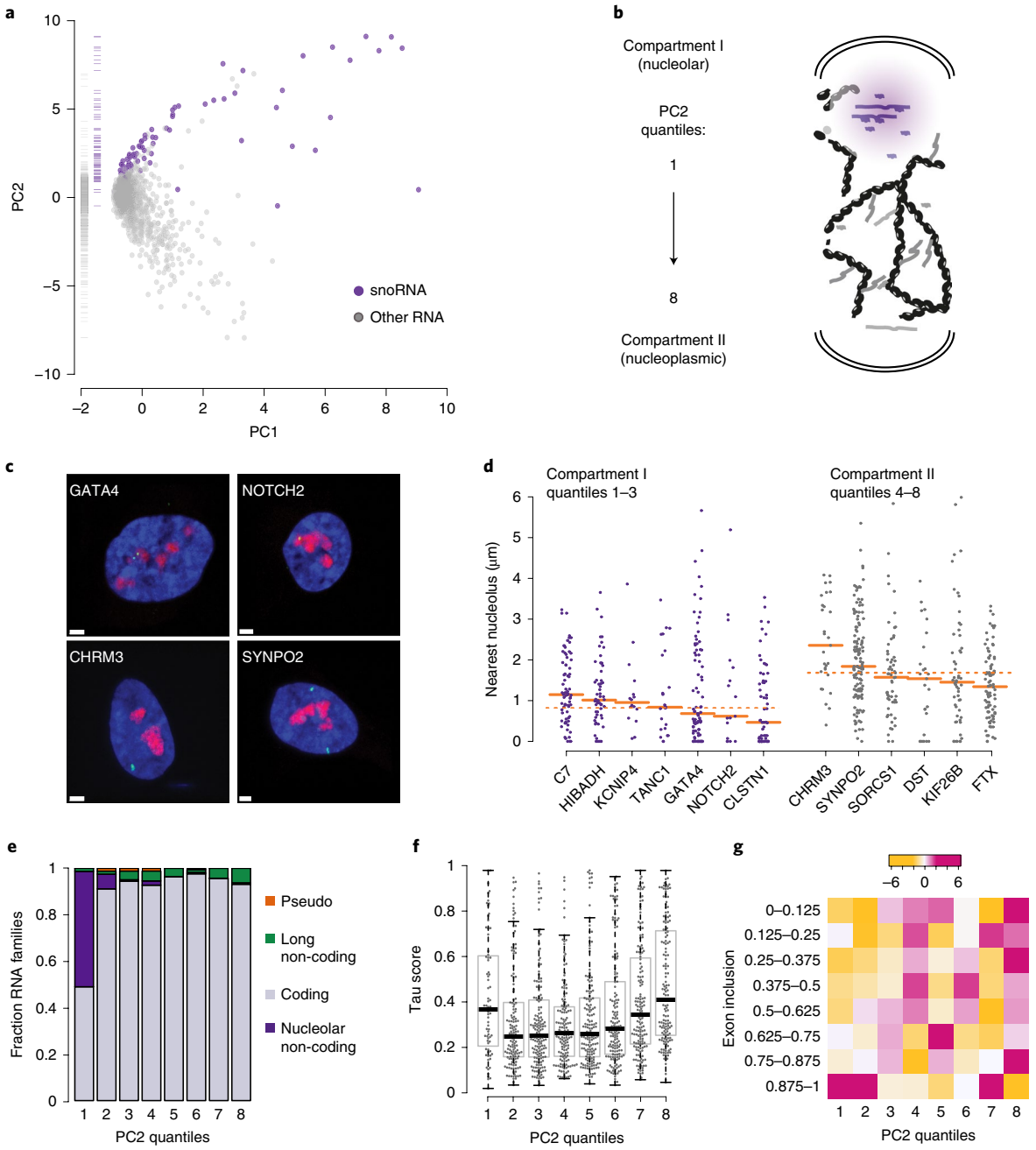

Fig. 3 | Bipartite nuclear transcriptome. a, PCA of transcripts (n=1,388) based on their pairwise proximity with the top 100 connected transcripts. **b**, Schematic of compartment I with transcripts in purple and compartment II with transcripts in gray. PC2 quantiles 1–8 are indicated. **c**, Examples of dual RNA-FISH. Red, U3; green, intronic probes against candidate RNA; blue, DAPI; scale bars, 2 µm. **d**, Spatial distances of intronic transcript spots to nearest nucleolus as measured by dual RNA-FISH. Compartment I transcripts are shown in purple, compartment II in gray (I versus II nested analysis of variance (ANOVA): P=0.002). C7 n=72 intronic spots, HIBADH 58, KCNIP4 16, TANC 24, GATA44 86, NOTCH2 17, DST 63, CHRM3 29, SNPO2 129, SORCS1 59, CLSTN1 27, KIF26B 55, FTX 88. Orange bar, median; dashed orange lines, mean of medians. **e**, Fractions of RNA families along PC2 axis (PC2 quantiles 1, 3, 6, 8: n=174, quantiles 2, 4, 5, 7: n=173). **f**, Tissue specificity (Tau scores, 0 for broadly expressed and 1 for tissue specific[22,23]) for each transcript in different proximity quantiles. The borders, bar and whiskers of the box plot represent the first (Q1) and third (Q3) quartiles, the median and the most extreme data points within 1.5× interquartile range from Q1 to Q3, respectively. **g**, Heatmap of Pearson residuals (yellow–magenta) from regression against PC2 quantiles (columns, PC2 quantiles 1–8: n=1,214, 2,017, 2,926, 3,070, 3,324, 3,884, 3,417, 3,225 exons) and bins of exon inclusion scores (rows).

transcripts (90%). We observed an accumulation of gene ontology terms, many specific to the neuronal cell type, in compartment II but few enriched terms in compartment I (Supplementary Fig. 8). Similarly, the number of transcripts with high tissue specificity[22,23] increased from compartment I to compartment II, with the exception of nucleolar quantile 1 (Fig. 3f). Consistently, genes encoding compartment II transcripts were closer in the linear genome to multi-enhancer domains crucial for cell identity (super-enhancers[24,25]) than genes whose transcripts were assigned to compartment I (Supplementary Fig. 8). Finally, alternative splicing of exons in protein-coding transcripts occurred less frequently in compartment I, as indicated by high exon inclusion[26]. In contrast, transcripts in

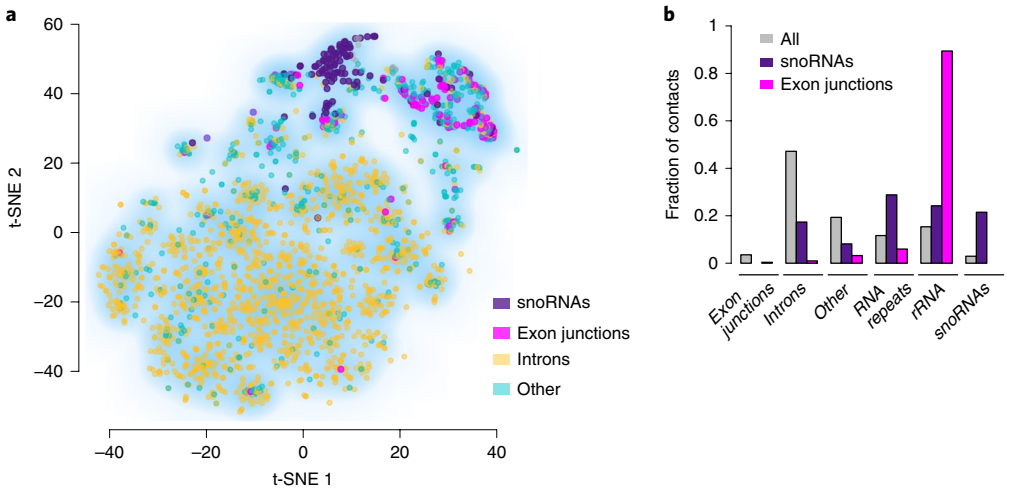

Fig. 4 | RNA proximities of nascent and processed transcripts. a, t-SNE visualization of intronic ($n=1,666$), exon junction ($n=105$) and other, ambiguous ($n=608$) RNA features based on their pairwise proximity with the top 100 connected features. Additionally, snoRNAs are colored in purple ($n=111$). **b**, Comparison of spatial RNA associations ($P \leq 0.1$) between all features, snoRNAs and exon junctions with different classes of features (number of associations between features: all - exon junctions $n=1,022$, snoRNAs - exon junctions $n=0$, between pairs of exon junctions $n=4$, all - introns: $n=13,596$, snoRNAs - introns $n=147$, exon junctions - introns $n=10$, all - other $n=5,571$, snoRNAs - other $n=69$, exon junctions - other $n=33$, all - RNA repeats: $n=3,360$, snoRNAs - RNA repeats $n=244$, exon junctions - RNA repeats $n=61$, all - rRNA $n=4,426$, snoRNAs - rRNA $n=205$, exon junctions - rRNA $n=914$, all - snoRNAs $n=849$, between pairs of snoRNAs $n=182$, exon junctions - snoRNAs $n=0$).

compartment II exhibited variable and often lower exon inclusion (Fig. 3g). This suggests that the nuclear transcriptome is partitioned into broadly expressed, so-called housekeeping, transcripts in the vicinity of nucleoli in compartment I and tissue-specific RNAs and transcripts that are more frequently subject to alternative splicing in compartment II.

Precursor and processed RNA localization. To distinguish different spatial sites for precursor and processed RNA, we split reads for any transcript into three groups: overlapping introns for precursors, exon junctions for mature RNAs and a group of other, ambiguous stretches of transcripts. After comparison with random simulations (Supplementary Table 6), we visualized the high-dimensional spatial RNA association data by giving each RNA a position in a two-dimensional map using t-distributed Stochastic Neighbor Embedding (t-SNE[27]) (Fig. 4a). The clustering of snoRNAs apart from the bulk of precursor transcripts recapitulated the two compartments revealed by PCA and using gene-based transcript annotations. In addition, we found exon junctions clustered together, separated from nascent RNA and nucleolar transcript clusters. Strikingly, most pairwise RNA associations involving exon junctions were formed with ribosomal RNA (89%). Unlike snoRNAs, which contact each other and rRNA in nucleoli, exon junctions showed no association with snoRNAs (Fig. 4b). These results indicate that rRNA was captured at two distinct stages and locations: during synthesis and processing with snoRNAs in nucleoli, and as functional ribosomes, likely situated at the outer nuclear envelope and copurified from the cytoplasm by crosslinking to nuclei, in complex with spliced protein-coding RNAs as represented by exon junction reads.

Compartment- and transcript-specific local RNA density. Proximity RNA-seq enables the simultaneous detection of two or more proximal transcripts. We reasoned that frequent or rare simultaneous detection of multiple proximal transcripts is characteristic of a specific RNA and the nuclear region from which crosslinked particles originate and reflects high or low local RNA density, respectively. To derive local RNA density and/or connectivity for individual transcripts, we introduced a measure describing how often a transcript was detected singly or cobarcoded with one or multiple other RNA molecules. This so-called 'valency' of a given transcript was inferred from relative enrichments or depletions (z-score) in the number of reads mapping to the transcript in barcode groups encompassing one, two and three transcripts, respectively, compared with the nuclear transcriptome average (Fig. 5a, Supplementary Fig. 9, Supplementary Table 7). After assigning valency to individual transcripts, we analysed the distribution of valency in compartments I and II. We found that transcripts in compartment I overall exhibited high valency (that is, z-scores of valency $1<0$ and z-scores of valency 2 and $3>0$), indicating the increased RNA density in nucleoli[28]. Of note, spatial regions adjacent to nucleoli contained in PC2 quantiles 2 and 3 also exhibited high valency despite much reduced numbers of snoRNAs in such regions. In contrast, low-valency transcripts prevailed in compartment II (Fig. 5b). Further, we analysed valencies of whole proximal transcriptomes (association $P \leq 0.1$) specific to the abundant non-coding RNAs 7SLRNA, 7SK and nuclear speckle-resident MALAT1, respectively (Fig. 5c). We first examined transcript assignments to PC2 quantiles and found 7SLRNA enriched in proximity to compartment I transcripts (quantiles 1–3). 7SK and MALAT1 colocalization with proximal transcripts peaked near the border of or within compartment II (quantiles 4–8), respectively (Fig. 5d). We then combined transcript-specific valencies for each proximal transcriptome. All three proximal transcriptomes showed decreased valency 1 and increased valency 2 and 3 compared with the entire transcriptome, indicative of locally increased RNA density (Fig. 5e) (tests for valency 1, compared with entire transcriptome, Kolmogorov–Smirnov, two-sided: 7SL $P=5\times10^{-5}$, 7SK $P=0.001$, MALAT1 $P=3\times10^{-5}$, Cliff's delta effect size: 7SL 0.25, 7SK 0.22, MALAT1 0.21). We next asked how the valency of proximal transcriptomes is affected through associations with high-valency nucleoli and snoRNAs. Removal of snoRNAs from the analysis reduced valencies 2 and 3 and increased valency 1 of 7SLRNA (Fig. 5e) (for valency 1 with and without snoRNAs, Kolmogorov–Smirnov, two-sided: $P=0.003$, Cliff's delta effect size: −0.23). Therefore, the proximal transcriptome of 7SLRNA increased RNA density through association with nucleolar transcripts. The valency of the 7SK transcriptome showed little change upon removal of snoRNAs (for valency 1 with and without

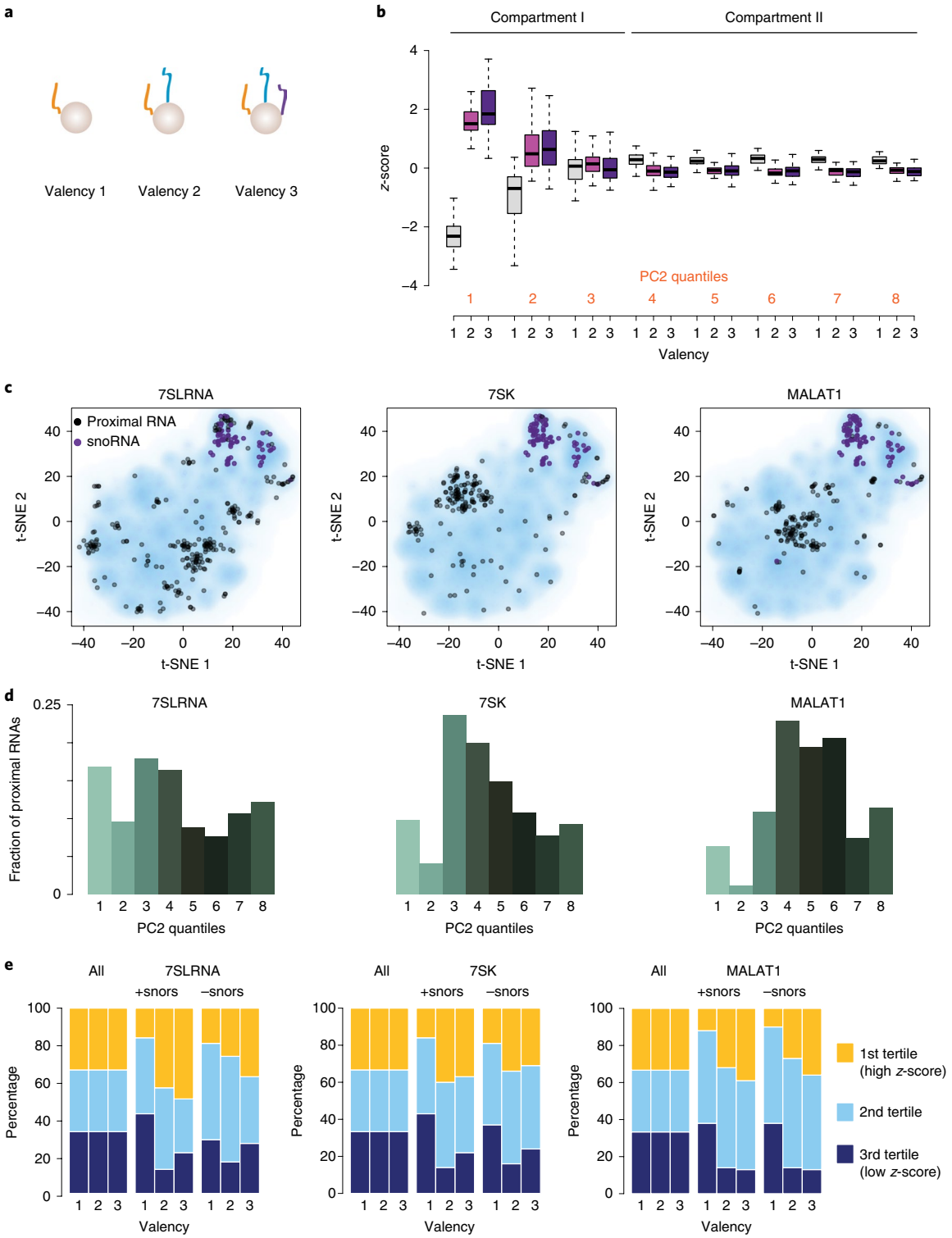

Fig. 5 | RNA valency. a, RNA valency was based on read counts of a given transcript in barcode groups with one, two or three transcripts. **b**, *z*-scores (*y* axis) of valencies (*x* axis: black numbers) along the PC2 axis (*x*: orange numbers, for PC2 quantiles 1 and 8 $n=58$, for quantiles 2–7 $n=57$ transcripts). The borders, bar and whiskers of the box plot represent the first (Q1) and third (Q3) quartiles, the median and the most extreme data points within 1.5× the interquartile range from Q1 to Q3, respectively. **c**, t-SNE visualization of transcripts (black) proximal ($P \leq 0.1$) to 7SLRNA ($n=415$), 7SK ($n=333$) and MALAT1 ($n=255$). SnoRNAs are indicated in purple. **d**, Fractions of transcripts proximal to 7SLRNA, 7SK and MALAT1, respectively, in different PC2 quantiles along the compartment I to II axis. **e**, Valency of proximal transcriptomes of 7SLRNA, 7SK, MALAT1 and of the whole transcriptome (all) before ($n=150$, 120, 103 and 2,486 transcripts, respectively) and after ($n=115$, 104, 95 and 2,411, respectively) removal of all snoRNAs from analysis (+snors, −snors). Tertiles were derived from whole-transcriptome distributions for valency 1, 2 and 3. Occurrences of transcripts proximal to 7SLRNA, 7SK and MALAT1 were counted in the tertiles and shown as percentages. For example, high valency is represented by a smaller first (yellow) and a larger third tertile (dark blue) for valency 1 and vice versa for valency 2 and 3 compared with whole-transcriptome tertiles.

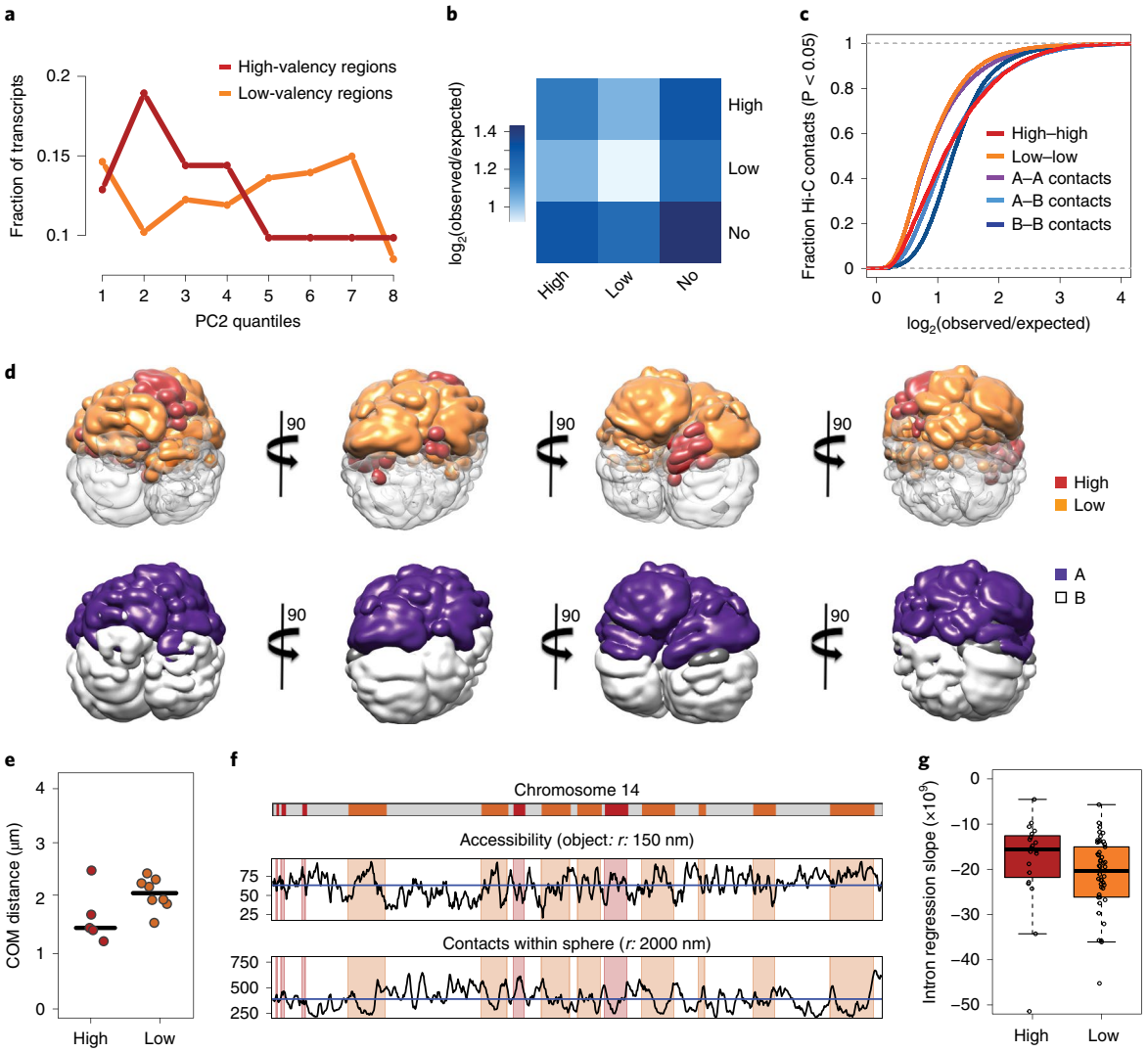

Fig. 6 | RNA valency and chromatin territories. a, Factions of transcripts with assigned valency in different PC 2 quantiles for high- and low-valency genomic regions. **b**, Mean Hi-C contact enrichments between high-, low- and no-valency regions. **c**, Cumulative distributions of Hi-C contact enrichments ($P \leq 0.05$) for high ($n = 5179$ contacts) and low ($n = 17{,}630$) valency segments in comparison with A–A ($n = 42{,}351$), A–B ($n = 22{,}217$) and B–B contacts ($n = 43{,}290$). Kolmogorov–Smirnov, two-sided, pairs of high-valency contacts versus pairs of low-valency contacts: $P = 0$, Cliff's delta effect size: 0.25. **d**, Density map of chromosome 14 ensemble model. **e**, Distances, median indicated as bar, between the center of mass (COM) of the B compartment and COM for high- ($n = 5$) and low-valency ($n = 8$) regions in chromosome 14, respectively (Mann–Whitney *U*-test, two-sided, $P = 0.1$). **f**, Local accessibility for a virtual object with radius of 150 nm and local contact density number within spherical volumes with radii 2,000 nm. High-valency regions in red, low-valency orange. **g**, Genome-wide intronic read density decays indicated faster transcription elongation in high- (red, $n = 20$) than low-valency (orange, $n = 50$) genomic regions (Kolmogorov–Smirnov, two-sided: $P = 0.03$, Cliff's delta effect size: 0.33). The borders, bar and whiskers of the box plot represent the first (Q1) and third (Q3) quartiles, the median and the most extreme data points within 1.5× the interquartile range from Q1 to Q3, respectively.

snoRNAs, Kolmogorov–Smirnov, two-sided: $P = 0.3$, Cliff's delta effect size: −0.13). Similarly to 7SK and in line with its architectural role in membrane-less nuclear bodies, autonomous from nucleoli, the proximal transcriptome of MALAT1 retained its valency state irrespective of snoRNAs (for valency 1 with and without snoRNAs, Kolmogorov–Smirnov, two-sided: $P = 0.9$, Cliff's delta effect size: −0.07). In conclusion, we found compartment I to be clearly distinguished from compartment II based on RNA density. Yet, the proximal transcriptome of MALAT1 was identified as a high-valency, RNA-dense body apart from compartment I, suggesting a heterogeneous RNA valency distribution throughout the relatively RNA-sparse compartment II.

RNA valency identifies dense, fast-transcribing chromatin. We next applied valency to differentiate transcriptionally active genomic subcompartments. Given that up to 80% of Proximity RNA-seq reads mapped to introns, we reasoned that Proximity RNA-seq-derived valency could be deployed to further classify transcript-encoding genome regions. We therefore performed Hi-C in SH-SY5Y cells and assigned average RNA valency scores to genomic regions. Hi-C studies have described open, transcriptionally active domains that preferentially contact other active domains in higher-order A compartments, while compact, poorly expressed domains preferentially contact each other in B compartments[2,29–31]. We found genomic A regions with high average RNA valency to be enriched in compartment I transcripts. Low-valency A regions more frequently encoded compartment II transcripts (Fig. 6a). As expected, genomic regions without assigned valency encompassed most B regions with low transcript expression (Supplementary Fig. 10) and showed very strong enrichment of chromatin contacts between

them by Hi-C (Fig. 6b). When comparing high- and low-valency genomic regions by Hi-C, we identified stronger DNA contact enrichments for pairs of high-valency compartment I regions, which resembled A to B contacts, whereas contact enrichments between low-valency regions were weaker, similar to A to A contacts (Fig. 6b,c, Kolmogorov–Smirnov, two-sided, pairs of high-valency contacts versus pairs of low-valency contacts: $P=0$, Cliff's delta effect size: 0.25). To gain further insights into distinct spatial distributions and local chromatin properties for different valency regions, we created three-dimensional (3D) models of a chromosome using Hi-C contact frequencies as spatial restraints[32]. In line with the contact enrichment profiles, high-valency domains in chromosome 14 were spatially closer to the B compartment than low-valency regions (Fig. 6d,e). Electron micrographs of nuclei likewise provided evidence for RNA-dense regions adjacent to compact chromatin (Supplementary Fig. 11)[33]. 3D model-derived estimates of local accessibility of a virtual object and of chromatin contact counts per volume suggested, furthermore, that high-valency regions were less accessible and more compact than low-valency regions (Fig. 6f, Supplementary Fig. 12). Finally, we observed increased transcription elongation rates in high-valency compared with low-valency regions (Fig. 6g, Supplementary Fig. 13) as estimated by regression of 5′ to 3′ read density within long introns measured by Proximity RNA-seq (Fig. 2a)[16,34]. This suggests that high- and low-valency territories are also distinguishable by local, apparent catalytic activities. In a reciprocal analysis in K562 haematopoietic cells we confirmed that fast-transcribing genomic regions, defined by BruDRB-seq[35], showed stronger Hi-C contact enrichments than slowly transcribed regions[36] (Supplementary Fig. 14).

Discussion

Much insight into RNA–RNA associations has so far been gained from methods using psoralen derivatives to specifically crosslink base-paired interactions. This approach has proved invaluable to map RNA secondary structure and to identify pairs of RNAs with complementary and hybridized sequence patches[9–13,37]. However, the larger 3D context of where RNA molecules are located and form contacts in cells has remained unaddressed. Proximity RNA-seq measures co-localization and positioning of RNAs in cellular 3D space at the resolution of individual transcripts, while psoralen methods identify base-paired regions within a transcript or a pair of transcripts. Dimensionality reduction of pairwise spatial RNA associations enabled us to accurately position transcripts to compartment I, encompassing the nucleolus and adjacent regions, and compartment II, the nucleoplasm and nuclear periphery, in strong agreement with 3D RNA-FISH. Previously, the detection and identification of the RNA-dense compartment I based on DNA proximity ligations has been hampered due to the low and repetitive DNA content in nucleoli. Furthermore, distances between genes distributed over the surface area of nucleoli are probably too large for ligation[3]. Measuring RNA proximities in subnuclear particles circumvents this limitation of DNA ligation assays to describe RNA-dense regions in nuclei. Ligation-free DNA proximity measurements mirror our finding of transcript compartments at the genome level[3].

Positioning of a gene to dense heterochromatin around nucleoli has been associated with gene repression[3,21]. However, the direct spatial mapping of transcriptional output here identified active RNA synthesis from specific genes at the periphery of nucleoli. Little is known about what defines the positioning of active genes and their nascent transcripts to either the nucleoplasm or nucleoli. Gene editing experiments, in which a compartment I gene, whose transcripts are close to nucleoli, replaces a gene encoding compartment II transcripts positioned further away from nucleoli, promise insights into whether DNA regions and/or the expressed transcript play roles in gene and RNA positioning.

Our in-emulsion barcoding and sequencing method enables not only the identification of pairwise spatial associations but also the simultaneous detection of more than two RNAs in proximity to each other. This allows transcripts to be characterized by valency, which can be interpreted as a local RNA density or connectivity. Transcripts can therefore be located and assigned to RNA-dense or sparse neighborhoods. For example, we mapped the proximal transcriptomes of 7SLRNA overlapping or at the border to high-valency compartment I, which explained the increased valency of the network. In contrast, the proximal transcriptome of MALAT1 in speckle bodies assigned to the often RNA-sparse compartment II showed increased valency independently of associations with nucleoli in compartment I.

The combined analysis of Hi-C chromatin contacts and RNA valency suggested that chromatin regions encoding compartment I transcripts with increased RNA valency exhibit stronger DNA contacts, reflecting proximity to the perinucleolar B compartment. These nucleic-acid-dense territories displayed faster transcription elongation compared with low-valency compartment II regions. We can only speculate that, besides genomic sequence and chromatin determinants, sequestering of certain protein factors might favor rapid transcription elongation in high-valency nuclear regions. Interestingly, transcription elongation factors have been shown to associate with chromatin in an RNA-dependent manner[38], which raises the possibility that transcripts themselves in crowded RNA-dense regions might accumulate elongation factors critical to speed up transcription and thereby sustain local high RNA valency.

The utility of Proximity RNA-seq will lie in its versatility to sequence different subcellular RNA-containing structures. Our understanding of the composition and functioning of RNA compartments beyond the nucleus, such as RNA-rich phase separations or aggregates in disease states, will benefit from the spatial RNA measurements introduced here. We envision that large-scale proximity measurements of RNA, a molecule with ubiquitous but non-random distribution throughout cells, will achieve a map of whole-cell spatial organization by means of sequencing.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41587-019-0166-3.

Received: 12 July 2018; Accepted: 22 May 2019;
Published online: 1 July 2019

References

1. Zhao, R., Bodnar, M. S. & Spector, D. L. Nuclear neighborhoods and gene expression. *Curr. Opin. Genet. Dev.* **19**, 172–179 (2009).
2. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
3. Quinodoz, S. A. et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744–757 (2018).
4. Beagrie, R. A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).
5. Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78–82 (2016).
6. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
7. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
8. Shah, S. et al. Dynamics and spatial genomics of the nascent transcriptome by Intron seqFISH. *Cell* **174**, 363–376.e16 (2018).
9. Weidmann, C. A., Mustoe, A. M. & Weeks, K. M. Direct duplex detection: an emerging tool in the RNA structure analysis toolbox. *Trends Biochem. Sci.* **41**, 734–736 (2016).
10. Nguyen, T. C. et al. Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. *Nat. Commun.* **7**, 12023 (2016).

11. Kudla, G., Granneman, S., Hahn, D., Beggs, J. D. & Tollervey, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl Acad. Sci. USA* **108**, 10010–10015 (2011).
12. Sugimoto, Y. et al. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature* **519**, 491–494 (2015).
13. Ramani, V., Qiu, R. & Shendure, J. High-throughput determination of RNA structure by proximity ligation. *Nat. Biotechnol.* **33**, 980–984 (2015).
14. Dressman, D., Yan, H., Traverso, G., Kinzler, K. W. & Vogelstein, B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl Acad. Sci. USA* **100**, 8817–8822 (2003).
15. Shendure, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
16. Ameur, A. et al. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
17. Scheer, U. & Hock, R. Structure and function of the nucleolus. *Curr. Opin. Cell Biol.* **11**, 385–390 (1999).
18. Neve, J. et al. Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation. *Genome Res.* **26**, 24–35 (2016).
19. Gondran, P., Amiot, F., Weil, D. & Dautry, F. Accumulation of mature mRNA in the nuclear fraction of mammalian cells. *FEBS Lett.* **458**, 324–328 (1999).
20. Engreitz, J. M. et al. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
21. Padeken, J. & Heun, P. Nucleolus and nuclear periphery: Velcro for heterochromatin. *Curr. Opin. Cell Biol.* **28**, 54–60 (2014).
22. Fagerberg, L. et al. Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13**, 397–406 (2014).
23. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* **18**, 205–214 (2017).
24. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
25. van Groningen, T. et al. Neuroblastoma is composed of two super-enhancer-associated differentiation states. *Nat. Genet.* **49**, 1261–1266 (2017).
26. Busch, A. & Hertel, K. J. HEXEvent: a database of Human EXon splicing Events. *Nucleic Acids Res.* **41**, D118–124 (2013).
27. van der Maaten, L. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2015).
28. Edstrom, J. E., Grampp, W. & Schor, N. The intracellular distribution and heterogeneity of ribonucleic acid in starfish oocytes. *J. Biophys. Biochem. Cytol.* **11**, 549–557 (1961).
29. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
30. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012).
31. Sexton, T. et al. Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell* **148**, 458–472 (2012).
32. Serra, F. et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLOS Comput. Biol.* **13**, e1005665 (2017).
33. Bernhard, W. A new staining procedure for electron microscopical cytology. *J. Ultrastruct. Res.* **27**, 250–265 (1969).
34. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3**, e02407 (2014).
35. Veloso, A. et al. Rate of elongation by RNA polymeraseII is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014).
36. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
37. Ziv, O. et al. COMRADES determines in vivo RNA structures and interactions. *Nat. Methods* **15**, 785–788 (2018).
38. Battaglia, S. et al. RNA-dependent chromatin association of transcription elongation factors and Pol II CTD kinases. *eLife* **6**, e25637 (2017).
39. Rybak-Wolf, A. et al. Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell* **58**, 870–885 (2015).

Acknowledgements

J.M. was supported by a Swiss National Science Foundation early postdoc mobility fellowship, a Human Frontier Science Program long-term fellowship and the Babraham Science Policy Committee. The work of I.F. and M.A.M.-R. was partially supported by the European Research Council under the 7th Framework Program FP7/2007-2013 (ERC grant no. 609989) and the European Union's Horizon 2020 research and innovation program (grant no. 676556) to M.A.M.-R. M.A.M.-R. also acknowledges the support of the Spanish Ministry of Economy and Competitiveness (grant nos. BFU2013-47736-P and BFU2017-85926-P), Centro de Excelencia Severo Ochoa 2013-2017 (grant no. SEV-2012-0208) and the Agency for Management of University and Research Grants (AGAUR). M.F.-M. was supported by UNAM Technology Innovation and Research Support Program PAPIIT IA201817 and PAPIIT IN207319. We acknowledge Sphere Fluidics for their contribution of microfluidic knowhow and time and their free donation of surfactants. We thank Babraham sequencing, fluorescence-activated cell sorting and imaging facilities for technical support, and Peter Rugg-Gunn, Paulo Amaral and Lucas Edelman for helpful discussions.

Author contributions

J.M. designed the study, performed experiments, analysed data and wrote the manuscript with contributions from all authors. S.W.W. provided conceptual advice and analysed data. I.F., J.C., A.S.-P. and M.A.M.-R. analysed data. M.F.-M., L.F.J.-G. and S.W. performed experiments. X.L. and F.F.C. provided technical help. S.A. provided conceptual advice. P.F. designed and supervised the study and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-019-0166-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Preparation of nuclear homogenates for Proximity RNA-seq. SH-SY5Y cells were cultured in high-glucose DMEM medium (Thermo Fisher Scientific) containing 10% fetal bovine serum (Hyclone) and Streptomycin, Penicillin (Thermo Fisher Scientific). At 70–80% confluency in 150 mm dishes, cells were crosslinked. Culture medium was replaced by 19.8 ml of pre-warmed 1× PBS and 0.2 ml of freshly prepared 10 mM EGS (Thermo Fisher Scientific) in dimethyl sulfoxide (DMSO) was added drop-wise for 1 mM final concentration. Cells were incubated for 10 min at 37 °C. Then, 1.4 ml of 16% formaldehyde (Agar Scientific) was added to dishes for a final concentration of 1%, and the incubation was extended for another 10 min at room temperature. The addition of 3.5 ml of 1 M glycine quenched crosslinking, and cells were scraped immediately and pelleted in a 50 ml Falcon tube at 210*g* for 5 min at 4 °C. Cell pellets were washed with 1× PBS supplemented with 125 mM glycine. A second wash with 1× PBS was carried out subsequently. Cell pellets were flash frozen in liquid nitrogen and stored at −80 °C until use.

Frozen cell pellets from single 150 mm dishes were thawed on ice in 1 ml of 20 mM Tris buffer pH 7.2, spun and resuspended in 1.5 ml of hypotonic Igepal C-630 lysis buffer (20 mM Tris, pH 7.2, 5 mM NaCl, 0.2% Igepal C-630, 2 mM EDTA, 1 mM EGTA) supplemented with 1× Complete, EDTA-free protease inhibitor cocktail (Roche Applied Science) and 0.5 units μl^{-1} SUPERase IN RNase inhibitor (Thermo Fisher Scientific). Cells were kept on ice for 30 min with occasional mixing. Nuclei were spun at 2,000 r.p.m. for 5 min in a benchtop centrifuge at room temperature, and the buffer exchanged with SDS washing buffer (20 mM Tris, pH 7.2, 5 mM NaCl, 0.3% SDS, 2 mM EDTA, 1 mM EGTA) supplemented with inhibitors as listed above. Samples were incubated for 10 min at room temperature in a thermoblock with constant mixing at 750 r.p.m. Triton X-100 was added for a final concentration of 1.7% and incubation prolonged for 10 min. Nuclei were then washed once in 10 mM Tris pH 7.2, 5 mM NaCl, 0.5 mM EDTA, 1% Triton X-100, supplemented with inhibitors as specified above, and once in the same buffer with only 0.05% Triton X-100. Nuclei were resuspended in 0.2 ml of wash buffer with 0.05% Triton X-100 and sonicated in 15 ml Falcon tubes using a bioruptor UCD-200 sonicator (Diagenode) with power set to medium for three to five cycles of 10 s on followed by 10 s off at 8 °C. After three cycles of sonication, the disruption of nuclei was inspected using Trypan blue staining and light microscopy. Sonicated nuclear homogenates were flash frozen in liquid nitrogen and stored at −80 °C until use.

Content and fragment length of RNA and DNA, respectively, were estimated after crosslink reversal and purification of nucleic acids using a spectrophotometer (NanoDrop) and agarose gel electrophoresis as follows. For DNA extraction, 20 µl of nuclear homogenate was supplemented with the following reagents to final concentrations of 50 mM Tris pH 8.0, 50 mM NaCl, 2 mM EDTA, 0.2% SDS and 5 µl of proteinase K (Roche Applied Science, 10 mg ml^{-1} stock concentration) for a total volume of 40 µl. DNA extraction mixes were first incubated for 4 h at room temperature followed by 2 h at 70 °C. Then, 2 µl of RNase A (Roche Applied Science, 10 mg ml^{-1} stock concentration) was added and samples were incubated for 1 h at 37 °C. DNA was purified by phenol/chloroform extraction and salt/isopropanol precipitation and resuspended in 10 µl of nuclease-free water (Ambion). To extract RNA, final concentrations of 100 mM sodium citrate pH 6.2, 2 mM EDTA, 0.2% SDS and 5 µl of proteinase K were added for a total volume of 40 µl. RNA extraction mixes were incubated for 4 h at room temperature followed by 1 h at 70 °C. 1 ml of Trizol LS (Thermo Fisher Scientific) was added and RNA isolated according to the manufacturer's manual and resuspended in 10 µl of water.

Barcoding of beads by emulsion PCR. For eight PCR reactions, 140 µl of MyOne streptavidin C1 magnetic beads (Thermo Fisher Scientific) was washed twice in polypropylene tubes (Treff) on a magnetic rack with high salt buffer (20 mM Tris pH 8.0, 1 M NaCl, 1 mM EDTA). After resuspension in 280 µl of high salt buffer, 48 µl of 100 µM dual-biotinylated primer R (Integrated DNA technologies) was added and beads were mixed briefly by vortexing. Binding was allowed for 20 min at room temperature with occasional vortexing. Beads were then washed twice in high salt buffer and once in TTLE buffer (10 mM Tris pH 8.0, 0.5 mM EDTA, 0.05% Triton X-100) with 0.04 µg ml^{-1} of molecular biology grade BSA (NEB). Beads were resuspended in 280 µl of TTLE buffer with 0.04 µg μl^{-1} BSA and stored for up to 2 d at 4 °C.

The aqueous phase for eight PCR reactions using AccuPrime HiFi Taq DNA polymerase and reagents (Thermo Fisher Scientific) was prepared on ice in DNA LoBind tubes (Eppendorf). To 943 µl of water were added 128 µl of 10× PCR buffer I, 32 µl of $MgSO_4$ (50 mM stock), 25 µl of dNTP mix (stock of 25 mM each dNTP, Thermo Fisher Scientific), 32 µl of Primer F (100 µM stock, Sigma) and 40 µl of 1 µM non-biotinylated primer R (Sigma). Then, 35 µl of beads carrying dual-biotinylated primer R were pipetted extensively and heated at 95 °C for 40 s before addition to the PCR mix on ice. After mixing by pipetting, 20 µl of AccuPrime HiFi Taq polymerase was added, and the mix was pipetted again. Next, 25 µl of 1 nM random barcode template was heated for 1 min at 95 °C before immediate addition to the PCR mix on ice. The mix was extensively pipetted on ice. The oil phase for one PCR reaction consisted of 480 µl of Pico-Surf 1–5% in Novec 7500 (Sphere Fluidics). To 480 µl of Pico-Surf in polypropylene tubes (Treff), 160 µl of the PCR mix was added. The water–oil phase-separated mixture was then emulsified by vortexing using a Vortex Genie 2 vortexer (Scientific Industries) on a horizontal tube holder at 4 °C for 20 min at maximum speed, and 50 µl aliquots of the emulsion were pipetted into each well of a 96-well hard-shell plate (Biorad). The PCR was carried out with a lid temperature of 100 °C. After 1 min at 94 °C, 35 cycles of 15 s at 94 °C, 30 s at 58 °C and 45 s at 68 °C were performed, followed by 5 min at 68 °C.

The 50 µl emulsion PCR reactions were pooled and spun at 2,000 r.p.m. for 1–2 min in a benchtop centrifuge or until non-emulsified oil phase appeared at the bottom. The lower oil phase was removed, three volumes of the initial aqueous phase of 5% Ficoll 400 was added, and the mixture was vortexed. Then, 350 µl of PFOH (1H,1H,2H,2H-Perfluorooctan-1-ol) was added, and after vortexing, the mixture was incubated at 37 °C for 5 min. To separate beads from the mixture, the tube was placed on a magnet and incubated at 37 °C for 5–10 min. The PFOH phase at the bottom and the aqueous upper phase, if transparent, were removed. Beads were resuspended in the remaining aqueous phase and TLE buffer (10 mM Tris pH 8.0, 0.5 mM EDTA) was added up to a total volume of around 300 µl. The PFOH extraction was repeated. Beads were then washed three to four times in TLE buffer and transferred into new tubes. Further washes as specified followed: three times in 1% SDS buffer (20 mM Tris pH 8.0, 5 mM NaCl, 1% SDS), once in TLE buffer supplemented with 1% Triton X-100, once in TTLE buffer with 0.04 µg μl^{-1} BSA. Beads from eight PCR reactions were pooled and resuspended in 50 µl of TTLE buffer with 0.04 µg μl^{-1} BSA.

Barcoding quality controls. Two-barcode quality control experiments (Fig. 1c,d) were carried out as described above except that random barcode pools were replaced with two barcode templates with defined sequences. To count barcoded beads of two-barcode experiments, fluorescently labeled probes base-pairing to barcode sequences were hybridized. Briefly, 2 µl of beads from a 50 µl batch of eight PCR reactions were mixed with 3 µl of 10× Accuprime PCR buffer I, 2 × 2 µl of the two barcode-specific probes or 1 × 2 µl of T7 probe (100 µM) and water up to 30 µl. Hybridization was carried out using the following temperature program: 2 min incubations at 94 °C and 80 °C, then from 75 °C to 61 °C a temperature decrease by 1° every 2 min, followed by 10 min at 60 °C and 2 min at 55 °C, 50 °C and 45 °C, respectively. Beads were washed once with 1% SDS buffer, once in high salt buffer, then resuspended in 1× Accuprime PCR buffer I, which was pre-warmed at 67 °C and incubated for 3 min at 67 °C. These washes were repeated two more times. After resuspension in TTLE buffer with 0.04 µg μl^{-1} BSA, beads were pipetted extensively before analysis by flow cytometry. We used a FACSAria III machine (BD Biosciences) with a 70 µm diameter nozzle and removed neutral density filter. For two-barcode experiments, fluorophores were detected with settings for PE 582/15 nm (Cy3) and APC 670/14 nm (Cy5).

To control the fraction of barcoded beads when using random template pools, we only used a Cy5-labeled DNA probe (T7 probe) hybridizing to a non-random region of the barcodes. We retained bead batches with 40–60% barcoding.

Random tailing of barcodes on beads. First, unused primers immobilized on beads were digested by an exonuclease I (NEB) treatment. One batch of beads from eight PCR reactions was resuspended in 65 µl of 10× Exo I buffer, 520 µl of water and 65 µl of exonuclease I and incubated at 37 °C for 90 min. Beads were then washed once in1% SDS buffer, once in TLE buffer with 1% Triton X-100 and once in TTLE buffer with 0.04 µg μl^{-1} BSA. To remove untemplated adenosine overhangs added by *Taq* polymerase, beads were incubated for 30 min at 20 °C in 10 µl of NEB Next End Repair Reaction Buffer, 5 µl of NEB Next End Repair Enzyme Mix (NEB) and 85 µl of water. Wash steps were repeated as outlined after exonuclease I treatment. Beads were then subjected to a T7 exonuclease treatment in 20 µl of NEB 4 buffer, 10 µl of T7 exonuclease (NEB) and 170 µl of water for 15 min at 25 °C to generate single-stranded barcodes immobilized via their dual-biotinylated 5′ ends, which are inert to exonuclease activity. The same wash steps as described above were repeated. To add random bases to barcode ends on beads, a primer with a 5′ random overhang was hybridized to the 3′ ends of barcodes and the overhang used as template to introduce 3′ random bases to barcodes. To do so, beads were resuspended in a Pfx polymerase mix (Thermo Fisher Scientific) consisting of 10 µl of 10× Platinum Pfx buffer, 84 µl of water, 3 µl of 10 mM nucleotides, 2 µl of 50 mM $MgSO_4$, 1 µl of Random tail primer (100 µM) and 1 µl of Pfx polymerase enzyme. The mix was incubated for 2 min at 94 °C, 5 min at 55 °C and 10 min at 68 °C. The wash steps as described above were repeated. The T7 exonuclease treatment to generate single-stranded barcodes was repeated as described above. Beads were resuspended in 50 µl of TTLE buffer with 0.04 µg μl^{-1} BSA and stored at −20 °C or 4 °C until use.

In-emulsion reverse transcription of RNA-containing particles and crosslink reversal. Two batches of beads, equivalent to 16 PCR reactions, were used to reverse-transcribe RNA-containing particles equivalent to 50 ng of purified RNA. The beads were resuspended in 30 µl of TTLE with 0.04 µg μl^{-1} BSA, and freshly prepared Actinomycin D (Sigma) in DMSO was added for a final concentration of 6 ng μl^{-1} to inhibit reverse transcriptase activity on DNA templates. Possible aggregates of beads were disrupted by extensive pipetting, followed by sonication using a bioruptor UCD-200 sonicator (Diagenode) with power set to medium for two cycles of 5 s on followed by 5 s off. The 160 µl of RT mixes (Thermo Fisher

Scientific) prepared in LoBind tubes (Eppendorf) contained 32 µl of 5× First-Strand Buffer, 8 µl of 10 mM dNTP Mix (each), 8 µl of 0.1 M dithiothreitol (DTT), 2 µl of RNase inhibitor, 0.32 µl of BSA for a final concentration of 0.04 µg µl^{-1} (NEB), 1.6 µl of $MgCl_2$ for a final concentration of 0.5 mM,16 µl of Superscript III, 2 µl of Actinomycin D, final concentration of 5 ng µl^{-1}, nuclear homogenate corresponding to 50 ng of RNA, and water. The beads were heated at 90 °C for 1 min before addition to the RT mix on ice. The emulsion was prepared with 550 µl of Pico-Surf on a vortexer as described above at 4 °C for 40 min at maximum speed, and 50 µl per well was added to a 96-well plate. A cycling temperature program was used to increase the probability of RNA priming. In a thermocycler with lid temperature at 55 °C, samples were incubated at 55 °C for 2 min, then placed on ice for 2 min, followed by 30 min at 22 °C. Then, the following cycle was repeated 60 times: 22 °C for 1 min, ramping to 50 °C with a rate of 0.5 °C s^{-1}, 50 °C for 1 min. Emulsions were broken as described earlier. Beads were resuspended in 200 µl of 20 mM Tris pH 8.0, 50 mM NaCl and incubated overnight at 65 °C. cDNA on beads was treated with 1 µl of RNase H (Thermo Fisher Scientific) and 0.5 µl of RNase A (Roche Applied Science, 10 mg ml^{-1} of stock concentration) in 5 µl of 10× RNaseH buffer l (Thermo Fisher Scientific) and 43.5 µl of water for 30 min at 37 °C. Beads were then washed once in 1% SDS buffer, once in TLE buffer with 1% Triton X-100 and once in TTLE buffer with 0.04 µg µl^{-1} BSA and resuspended in 1× TTLE buffer with 0.04 µg µl^{-1} BSA.

Proximity RNA-seq library preparation. The second Illumina sequence was introduced by PCR with a primer flanked by a poly-C stretch after G-tailing of the barcoded and immobilized cDNA. For G-tailing, beads were resuspended in 5 µl of 10× TdT buffer (NEB), 5 µl of cobalt chloride, 2.5 µl of 100 µM dGTP/ddGTP mix (95 µM dGTP (Thermo Fisher Scientific), 5 µM ddGTP (Sigma)), 5 µl of TdT enzyme (NEB, 2 units µl^{-1} final) and 32.5 µl of water and incubated for 20 min at 37 °C. Beads were washed once in 1% SDS buffer, once in TLE buffer with 1% Triton X-100, once in 10 mM Tris pH 8.0, 0.5 M NaCl, 0.5 mM EDTA, 0.05% Triton X-100 and twice in TTLE buffer with 0.04 µg µl^{-1} BSA.

Beads were resuspended in 5 µl of 10× Accuprime buffer I, 42.5 µl of water, 0.5 µl of RP1_long primer (50 µM stock), 0.5 µl of polyC12_cDNA adapter (50 µM stock) and 0.5 µl of Accuprime HiFi Taq pol (Thermo Fisher Scientific). After 1 min incubation at 94 °C, five cycles of 15 s at 94 °C, 45 s at 52 °C and 2 min 30 s at 68 °C were carried out. The beads were captured, the supernatant containing the barcoded cDNA library was transferred to a new tube and 150 µl of 20 mM Tris pH 8, 50 mM NaCl, 1 mM EDTA was added. Size selection of PCR products using AMPure XP beads (0.75× the sample volume, Beckman Coulter) was repeated twice.

To reduce non-specific amplification and PCR duplicates during library amplification with low input, an emulsion PCR was performed. The eluate from the preamplification was mixed with 20 µl of 10× Accuprime buffer I, 2 µl of RP1 long (50 µM), 2 µl of Index primer (50 µM), 1 µl of dNTPs (stock of 25 mM each dNTP, Thermo Fisher Scientific), 1 µl of $MgSO_4$ (50 mM stock), 4 µl of Accuprime HiFi Taq pol and water for a final volume of 200 µl. The mix was emulsified for 20 min with 600 µl of Pico-Surf as described above. The emulsion was transferred into wells of a 96-well plate and the following PCR program run: 94 °C for 1 min, 20–24 cycles of 15 s at 94 °C, 30 s at 52 °C and 2 min 30 s at 68 °C. After pooling the PCR reactions, the aqueous phase was recovered and DNA was size-selected in one round with 0.65× AMPure beads and a second round with 0.8× AMPure beads. The final libraries were eluted in 35 µl of TLE buffer. The concentration and fragment size distribution of libraries were determined by Bioanalyzer profiles (Agilent Technologies) and Kapa Illumina SYBR green qPCR (Kapa Biosystems) according to manufacturer's instructions.

Species-mixing control experiment. For the species-mixing experiment, 1182-4H *Drosophila melanogaster* cells were cultured in M3 medium (10% FCS) according to recommendations of the Drosophila Genomics Resource Center (DGRC). Nuclear particles were prepared as described above for human cells. Fly particles equivalent to 50 ng of purified RNA were added to the same amount of human particles before encapsulation into droplets together with barcoded beads for reverse transcription. All steps were carried out as described for libraries with human particles only.

Hi-C library generation. Hi-C libraries were generated for biological duplicates as described previously[40,41]. Briefly, after fixation in 2% formaldehyde for 10 min, nuclei from 2–10 million SH-SY5Y cells were digested with HindIII at 37 °C overnight, and DNA ends were labeled with biotin-14-dATP (Life Technologies) in a Klenow fill-in reaction and then re-ligated overnight. After treatment with proteinase K (Roche) and crosslink reversal at 65 °C for 16 h, DNA was purified and sheared to an average size of 400 base pairs (bp), following manufacturer's instructions (Covaris). The sheared DNA was end-repaired, A-tailed and size-selected using AMPure XP beads (Beckman Coulter) to isolate DNA ranging from 250 to 550 bp in size. Accessible biotin at fragment ends was removed, and internally biotinylated fragments were immobilized on MyOne Streptavidin C1 DynaBeads (Invitrogen). Paired-end adapters (Illumina) were ligated to fragment ends and Hi-C libraries were amplified by PCR.

Dual RNA-FISH. The smiFISH (single molecule inexpensive FISH) strategy adapted here has been developed by Tsanov et al[42]. The transcript-specific probes (usually 24) are unlabeled but contain a fixed sequence complementary to a fluorescently labeled detector probe (FLAP oligo, Supplementary Table 1). Transcript probe design was performed with Stellaris probe designer (www.biosearchtech.com) with the following settings: probe length 20 bases, human masking level 5, minimal spacing 2 bases. The first and/or second intron of transcripts was used for probe design. An equimolar mix of probes (100 µM) was prepared, and 2 µl of this mix and 200 pmol of the FLAP oligo were added to 50 mM Tris pH 8, 100 mM NaCl in a final volume of 20 µl. The mix was incubated at 85 °C for 1 min. To hybridize probes with FLAP oligo, the temperature was then gradually decreased at 1 °C intervals from 66 to 59 °C and the mix incubated for 1 min at each temperature. After a final incubation at 55 °C for 1 min, labeled probes were frozen. Cells were seeded on a coverglass to reach around 80% confluency the following day. Then, cells on coverglasses were washed in 1× PBS and crosslinked with 3.7% formaldehyde in 1× PBS for 10 min at room temperature followed by two 1× PBS washes. Cells were permeabilized in 1× PBS with 0.3% SDS for 10 min at room temperature followed by 1× PBS with 1% Triton X-100 for 5 min at room temperature. Subsequently, samples were washed in wash buffer (2× saline sodium citrate (Thermo Fisher Scientific), 10% deionized formamide (Ambion), 10 mM Ribonucleoside Vanadyl Complex (NEB)) for 2–5 min. Coverglasses with fixed cells were then incubated upside down on 200 µl of hybridization buffer (2× saline sodium citrate (Thermo Fisher Scientific), 10% dextran sulfate, 10% deionized formamide (Ambion), 10 mM Ribonucleoside Vanadyl Complex (NEB)) with U3 probes (100 ng) and candidate transcript probes (3 µl of mix) in a light-tight humidified chamber for 4 h at 37 °C. After in situ hybridization, samples were washed twice for 30 min in wash buffer at 37 °C. During the second wash step, 10 ng ml^{-1} of DAPI was added. Then, coverglasses with cells were washed twice in 1× PBS and mounted on microscope slides using Prolong Diamond Antifade Mountant (Thermo Fisher Scientific). Stacked images were taken with a confocal Zeiss 780 microscope (Zeiss) and 3D distances measured using the Imaris software (Bitplane).

Electron microscopy of nuclei after EDTA regressive staining. EDTA regressive staining was conducted as originally reported[33] with modifications as follows. Briefly, cell pellets were fixed in 2.5% glutaraldehyde and 4% paraformaldehyde. Samples were then dehydrated with increasing concentrations of ethanol (50%, 70%, 80%, 90%, 96%, for 10 min each and 100% ethanol three times for 10 min). Then, samples were treated three times with propylene oxide for 5 min each. Samples were embedded in epoxy resin and polymerized at 60 °C for 16 h. Thin sections of about 50 nm width were mounted on copper grids covered with formvar and treated with 3% uranyl acetate for 3 min, rinsed and then floated on 0.2 M EDTA for 18 min at room temperature. After rinsing, cells were contrasted with 0.2% lead citrate for 2 min. The grids were washed with bi-distilled water and air-dried. Imaging was done with a transmission electron microscope (Jeol JEM-1010, Peabody, MA) operating at 80 kV and a CCD (charge-coupled device) camera (CCD-300-RC, MT 1).

Filtering and mapping of barcode and cDNA reads generated by Proximity RNA-seq. Proximity RNA-seq libraries were sequenced on two to three sequencing lanes. We combined the FASTQ files, 150 bp single-end, derived from the same library into one FASTQ file.

Properly constructed sequence reads started at the 5′ end with a 26 bp random barcode, immediately followed by a fixed 20 bp sequence, which specified one of the primer sequences used to amplify barcodes on beads. The pipeline confirmed that this fixed sequence was present, allowing for a 2 bp mismatch. Reads lacking the fixed primer sequence were discarded.

Low-complexity polynucleotides, the same nucleotide 13 or more times in a 20 bp sequence, in barcodes were removed. We chose this threshold because the probability of a randomly generated 20 bp sequence containing the same nucleotide at least 13 times was less than 1 in 1,000 (it was assumed that each nucleotide occurred with equal frequency and that any combination of nucleotides was equiprobable). In addition, the software filtered out putative adapter artefacts by screening barcodes against adapter sequences used in the generation of the library.

To account for errors in barcodes introduced throughout the protocol, which could lead to an inflation of the barcode complexity of a library and reduce the number of identified spatial RNA associations, the pipeline allocated barcodes to groups of barcodes of very closely related sequences, most likely derived from the same barcode.

To create these groups, barcode sequences were extracted from all the sequenced reads. This dataset was then de-duplicated and concatenated into a single string, in which barcodes were separated from one to another by 24 unspecified bases (*N*). The produced file was in FASTA format and was subsequently converted to a Bowtie 2 FM-index (based on the Burrows–Wheeler transform). Each barcode was then mapped back to this index using Bowtie 2 in FASTA mode, reporting all alignments (-a)[43]. Barcodes either mapped uniquely to this index or mapped to multiple locations. In the latter scenario, the two locations to which the read mapped were considered members of the same barcode group. After mapping, all barcode groups sharing a common barcode were collapsed into a single barcode group comprising all the observed, slightly varying barcodes.

We noticed barcode sequences that were identical to one another but offset by a nucleotide. If artefactual, this would cause new barcodes to be generated and

potentially split otherwise cobarcoded groups of transcripts. For this reason, we trimmed barcodes by 3 bp at either end after extracting from the 150 bp sequenced read and before mapping to the virtual barcode genome, which consisted of the full 26 bp units.

Finally, to remove large, likely false positive transcript groups, we chose a conservative cutoff in which the pipeline modeled the variation in increasing barcode group size, that is, the number of cobarcoded transcripts, as a Poisson distribution. Barcode groups of a size so large that they would have been expected to occur with a probability of less than 0.001 were removed from the dataset.

For mapping cDNA sequences, the first 67 bp from the start of a read were removed (26 bp barcode, 20 bp barcode PCR primer, 15 bp random reverse transcription primer and six extra base pairs, which improved mapping efficiency); the next 50 bp were retained and mapped to the human genome using HISAT2 with default parameters[44] and known splice-sites (GRCh38.83).

Ribosomal RNA sequences in the reference human genome 38 were masked. The rRNA sequence was subsequently re-introduced into the genomic sequence as a separate chromosome (NCBI Reference Sequence: NR_046235.1).

Following mapping, exact read duplicates (reads sharing identical barcodes and mapping to the same position) were removed, but one representative copy was retained.

Custom transcriptome annotation. We used gene annotations from ENSEMBL Human Genome 38 (Ensembl 78: December 2014) as transcript units. We divided the reference annotation so that reads could be aligned unambiguously to a single feature. Since genes may overlap, we defined a hierarchy to establish how regions were classified (Supplementary Fig. 2). First, the pipeline used the ENSEMBL Human Genome 38 to retrieve gene coordinates and repeat masker to define RNA repeat regions in Human Genome 38 (http://www.repeatmasker.org). Opposing strands were treated separately with respect to overlapping genes, but regions of overlap between genes on the same strand were excluded. An exception to that rule was when genes were contained entirely within a surrounding gene. In such instances, the area of overlap was assigned to the inner gene. This was particularly important, since many genes, such as snoRNAs, are often located entirely within larger genes. Reads mapping to either strand of a RNA repeat region in the genome were assigned to single representations of 5S, HY1, HY3, HY4, HY5, U1, U2, U3, U4, U5, U6, U7, U8, U13, U14, U17, BC200, transfer RNA (all t-RNAs were collapsed into one feature), 7SLRNA and 7SK.

For gene features, only uniquely mapping reads were kept. The pipeline allowed multi-mapping as long as the best possible alignment corresponded to a predefined repeat region. Reads not mapping to any feature were discarded.

The pipeline grouped features into RNA particles or transcript groups based on the barcode read sequences. Multiple reads with the same barcode and mapping to the same transcript were counted as one observation, which we named a proxy read, due to the ambiguity of the multiple reads originating from a single RNA molecule or from several copies of the same transcript.

Monte Carlo simulation to identify preferential pairwise RNA contacts. We created a Monte Carlo simulation to identify preferential spatial associations or contacts between pairs of transcripts. The simulation took as input: (1) the number of transcript groups or RNA particles observed, which equals the number of unique barcodes in a dataset, (2) the barcode group size of each of those particles and (3) the frequency with which each transcript was observed in the dataset (number of proxy reads). The simulation first constructed frequency distributions for barcode group size and transcript abundance. Then, virtual RNA particles were created in silico by selecting a random value from the barcode group size distribution and then randomly selecting different transcripts from the transcript abundance distribution to fill up the barcode group. Again, the simulation allowed any combination of transcripts within a barcode group except combinations with multiple observations of the same transcript. On the occasion when a transcript was selected that was already present in the particle, the selection of the new transcript was rejected and instead added to a priority pool. Transcripts inside the priority pool were randomly picked for the generation of the following virtual RNA particle.

Following random particle generation, the cobarcoded pairwise RNA contacts were counted. In total, 100,000 simulations were performed, and the number of simulations in which a pairwise RNA contact occurred at least as many times as in the observed data was recorded. This gave a P value for the probability that an observed RNA–RNA contact in the real data could have occurred with that frequency by chance. One of the simulated randomized datasets was also passed to the Monte Carlo simulation to undergo 100,000 simulations. We intended that this would give us a measure of the variation between random datasets (but still retain comparable barcode group size and transcript abundance distributions to those of the observed datasets). There were now two groups of Monte Carlo simulations: the first the result of performing 100,000 simulations on the observed dataset (group 1) and the second the result of performing 100,000 simulations on one of the randomized datasets (group 2) of group 1.

Each P value assigned to an observed contact after comparison with group 1 was then added to a local distribution of 500 P values from the group 2 simulations. The local distributions were chosen to most closely match the RNA–RNA contact being examined in terms of the observed and simulated counts. Each distribution was modeled as a normal distribution to assign a probability of a given contact falling within that distribution. Consequently, for infrequent contacts the selected distributions generally exhibited high variability, whereas for frequent contacts the variability was lower, so the new P values calculated took this change in variability into account. These new, local-background-corrected P values obtained from the normal distribution, referred to simply as P values throughout the manuscript, were used when filtering spatial RNA associations and for principal component and t-SNE analyses. Furthermore, background-corrected P values were adjusted using Benjamini–Hochberg multiple testing correction.

Nascent versus processed transcriptome analysis. We categorized the Binary Alignment Map (BAM) format reads mapped by HISAT2 into one of three categories: 'Exon Junction', 'Nascent' or 'Other'. To achieve this, we defined regions in human genome assembly 38.78 as unambiguously intronic: that is, notated as an intron but never also notated as an exon in the case of overlapping isoforms or genes. After determining the genomic positions of these sequences, we identified BAM reads in which both ends overlapped an unambiguous intron sequence, and analysis of the Compact Idiosyncratic Gapped Alignment Report (CIGAR) string showed that the read was not split by HISAT2 during mapping. These were considered nascent reads. In contrast, we defined exon junction reads as those that that were split by HISAT2 during mapping and neither of the split read sequences overlapped with a predefined unambiguous intron region. All reads classified neither as exon junction nor as nascent were labeled as other.

Poisson distribution in bead barcoding. Bead barcoding, and in general the integrity of emulsions generated as described earlier, was controlled by applying a Poisson model to two-barcode PCR assay data (see section "Barcoding quality controls"). Specifically, flow cytometry analysis of barcoded beads was used to count the number of beads without barcodes, beads with both barcodes and beads with signal from only one of the two barcodes. However, beads for which we detected only one type of barcode could originate from PCR reactions with initially a single or multiple copies of one of the barcode templates. Similarly, beads with both barcode types could originate from more than two templates. We estimated the number of initial barcodes per droplet as follows. Let X and Y be the hidden number of initial barcode templates of each type on a bead. We assumed that $X \approx \text{Poisson}(\lambda_X)$ and $Y \approx \text{Poisson}(\lambda_Y)$, with X and Y independent. For each bead, we observe $I_X = I(X > 0)$ and $I_Y = I(Y > 0)$—that is, binary indicators of whether each Poisson variable is greater than zero. It can be shown that the maximum likelihood estimate of λ_X is $-\log(1 - \text{mean}(I_X))$, with a similar result for λ_Y. From this information, we estimated the number of beads with unique barcodes, which is $P(X+Y=1) = (\lambda_X + \lambda_Y)\exp(-\lambda_X - \lambda_Y)$ (since $X+Y \approx \text{Poisson}(\lambda_X + \lambda_Y)$). Similarly, the number of beads with multiple barcodes is $P(X+Y>1) = 1 - P(X<2) = 1 - (1+\lambda_X+\lambda_Y)\exp(-(\lambda_X+\lambda_Y))$. Hence, we estimated the ratio of single-to-multiple barcodes as $P(X+Y=1)/P(X+Y>1)$. This assumption can be verified with the chi-squared test of independence—we used two degrees of freedom, since the total number of beads is not fixed and two parameters are estimated from the data.

Mapping statistics. We compared the percentage of reads mapping to genes, exons and introns (Ensembl 78: December 2014) from Proximity RNA-seq and ribosomal RNA-depleted total RNA library from SH-SY5Y cells[39] (SRX1007599) with the percentage of base coverage of the same features. Genes were set as 100.

U3 RNA proximities and experimental controls. The number of cobarcoding events was plotted against the number of proxy reads for each transcript. A Mann–Whitney U-test was performed on the log ratios ($\log_2$(cobarcoding counts) $-$ $\log_2$(proxy reads)) for snoRNAs and non-snoRNAs.

Number of cobarcoded transcripts of top 1,000 RNAs. We restricted the analysis to the top 1,000 transcripts (based on the number of proxy reads in the observed dataset). For each of these transcripts, the number of unique, cobarcoded transcripts was counted irrespective of how many times a given pair of transcripts was observed together. The number of unique, cobarcoded transcripts was plotted against the number of proxy reads for the corresponding transcript of the top 1,000 list. Using the top 1,000 transcripts from the observed dataset, unique cobarcoded transcripts were retrieved from random simulations as well. Fisher's exact tests were performed for each transcript on 2 by 2 tables, which included the number of contacted, that is, cobarcoded, transcripts and the number of transcripts that were not cobarcoded of the top 1,000 transcriptome for the observed and the randomized data. The top 1,000 proximal transcriptome was defined as the union between observed and simulated datasets of all transcripts contacted at least once by any top 1,000 RNA. P values were false discovery rate (FDR)-adjusted to control for multiple testing.

RNA proximities and genomic distance. Monte Carlo simulation-derived $-\log_{10}$ P values, capped at 10, were plotted for RNA association pairs with more than three observations and after filtering out RNA repeats with multiple gene loci against genomic distance between the RNA-encoding genes.

Principal component analysis on pairwise RNA contacts. We performed PCA on a matrix of $-\log_{10} P$ values within each cell from pairwise spatial RNA associations. The top 100 connected transcripts were used as input variables (columns) and all transcripts as observations (rows). Including more transcripts as input variables resulted in failure of PCA due to violation of the normality assumption. Furthermore, we only used pairwise spatial RNA associations with more than three observations and a P value cutoff ≤ 0.1 as derived by the Monte Carlo simulations. We removed transcripts with only one reported pairing with another transcript and associations that involved mitochondrial ribosomal RNA. We restricted $-\log_{10} P$ values to a plateau at 5 and then subtracted the matrix mean from cells before performing PCA (R: princomp, using the correlation matrix). Transcripts were assigned to eight quantiles based on PC2.

Analysis of PC2 quantiles. Transcripts contained within the eight PC2 quantiles were assigned to RNA classifications retrieved from HGNC (HUGO gene nomenclature committee, http://www.genenames.org/).

We analysed enrichments of biological processes from gene ontology for transcripts in PC 2 quantiles using ToppCluster[45] (https://toppcluster.cchmc.org/). A P value cutoff ≤ 0.05 and Bonferroni correction were selected. Furthermore, limits of transcript number n (minimum and maximum number of transcripts allowed for an annotation) were set to $5 \leq n \leq 1{,}500$. Identified terms of biological processes were clustered for visualization in Supplementary Fig. 8 based on semantic grouping using the R package GOSemSim[46].

Tissue specificity of transcripts in PC2 quantiles based on Tau scores, where 0 means broadly expressed and 1 is specific, was retrieved from Kryuchkova-Mostacci and Robinson-Rechavi[23]. The expression data are based on RNA-seq measurements of 27 human tissues[22].

Exon inclusion scores, a measure for how often an exon is included in the mature transcript molecule, were used to estimate alternative splicing of transcripts in PC2 quantiles. For exons of known mRNA isoforms (excluding non-coding RNAs) as defined by the UCSC Genome Browser (GRCh37/hg19) scores were obtained from the HEXEvent database[26] (http://hexevent.mmg.uci.edu/). Furthermore, the exon inclusion scale from 0 to 1 was divided into eight equal interval groups. Thus, each exon had an assigned exon inclusion group and also a PC2 quantile. To identify overrepresented or underrepresented combinations of exon inclusion group and PC2 quantile, we used a Poisson generalized linear model (GLM). We counted the number of transcripts with each combination. This count was regressed against a linear combination of exon-inclusion-group-specific and PC2-quantile-specific parameters via a logarithmic link function. The Pearson residuals from this regression model were plotted in a heatmap.

t-SNE using introns, exon junctions and other features of transcripts or gene-based annotations. The same matrix as for PCA with $-\log_{10} P$ values within each cell from pairwise spatial RNA associations was used for t-SNE. The top 100 connected transcripts were used as input variables (columns) and all transcripts as observations (rows). Again, we only used pairwise spatial RNA associations with more than three observations and a P value cutoff ≤ 0.1 as derived by the Monte Carlo simulations. We removed transcripts with only one reported pairing with another transcript and associations that involved mitochondrial rRNA. We restricted $-\log_{10} P$ values to a plateau at 5 and then subtracted the matrix mean from cells before performing t-SNE (Rtsne package: seed 96, perplexity = 50, theta = 0.2). Similarly, for gene-based annotations the PCA matrix was used for t-SNE visualization (seed 96, perplexity = 46, theta = 0.1).

Assignment of valency to transcripts. Proxy reads for each transcript were counted in barcode groups of different size (named valencies). We then selected transcripts with the sum of valency 1, 2 and 3 (1: an observation of a single transcript, 2: two different transcripts detected with one barcode, 3: three different transcripts detected with one barcode) greater than 10 proxy reads. For each transcript, counts in valency 1, 2 and 3 were divided by the sum of all three valencies of that given transcript. Subsequently, the transcriptome-wide distributions of the three valencies were separately transformed into z-scores. Transcripts were then assigned to high- and low-valency classes, respectively. High-valency transcripts were defined based on z-scores of valency 1 <0 and a mean of valency 2 and 3 z-scores >0. Low-valency transcripts had z-scores of valency 1 >0 and a mean of valency 2 and 3 z-scores <0. Furthermore, only transcripts reproducibly assigned to high- and low-valency classes in Proximity RNA-seq triplicates were retained and used for further analysis. To analyse the valency of a group of transcripts proximal to a specific RNA, the mean valency z-scores were used.

Hi-C data processing. Sequencing data generated by Hi-C of two biological replicates were processed using HiCUP, which includes mapping to the reference genome, di-tag filtering and removal of artefacts[47].

Each chromosome contact matrix was normalized using TADbit[32]. Briefly, low-quality bins (those presenting low contacts numbers) were removed and ICE normalization[48]—also known as 'vanilla' normalization[36]—was performed using the default settings.

Identification of subnuclear genomic compartments. To segment the genome into A and B regions, the normalized Hi-C matrices at 100 kb resolution were transformed into correlation matrices using the Pearson product-moment correlation. The first component of a PCA (PC1) on each of these matrices was used as a quantitative measure of compartmentalization, and gene expression values (rpkm) from Proximity RNA-seq data were used to assign negative and positive PC1 categories to the correct compartments. If necessary, the sign of the PC1 (which is randomly assigned) was inverted so that positive PC1 values corresponded to A compartment regions and vice versa for the B compartment.

Integrative 3D chromosome modeling with TADbit. To assess the potential of the Hi-C data for modeling, we computed the MMP score (0.6) of the matrix and the predicted accuracy of the models, named distance Spearman correlation coefficient (0.5), using the MMP score as implemented in TADbit[32,49]. The MMP score is based on the matrix size, the contribution of significant eigenvectors in the matrix, and the skewness and kurtosis of the z-scores distribution of the matrix.

The normalized interaction matrix of chromosome 14 was used for modeling at a resolution of 100 kb. The short chromosome arm and the centromere were omitted due to the poor quality of the contact information. TADbit generates 3D models using a restraint-based modeling approach, where the experimental frequencies of interaction are transformed into a set of spatial restraints. The size of each particle in the models was defined by the relationship 0.01 nm bp^{-1} assuming the canonical 30 nm fibre[50]. Using a grid search approach, TADbit identified empirically the three optimal parameters to be used for modeling: (1) maximal distance between two non-interacting particles (maxdist set as 1,600 nm); (2) a lower-bound cutoff to define particles that do not interact frequently (lowfreq set as -1.6); and (3) an upper-bound cutoff to define particles that do interact frequently (upfreq set as 0.0). Once the three optimal parameters are defined, TADbit sets the type of restraints between each pair of particles considering an inverse relationship between the frequencies of interactions of the contact map and the corresponding spatial distances. Two consecutive particles were spatially restrained by a harmonic oscillator with an equilibrium distance that corresponds to the sum of their radii. Non-consecutive particles with contact frequencies above the upper-bound cutoff were restrained by a harmonic oscillator at an equilibrium distance, while those below the lower-bound cutoff were maintained further than an equilibrium distance by a lower-bound harmonic oscillator. To identify 3D models that best satisfy all the imposed restraints, the optimization procedure was then performed using a Monte Carlo simulated annealing sampling protocol as implemented in TADbit. A total of 5,000 models were generated and only the 1,000 that best satisfied the input restraints were used for further analysis. The contact map obtained from the final 1,000 models ensemble resulted in a Pearson correlation of 0.72 with the input Hi-C interaction matrix, which is indicative of good model accuracy[49].

Clustering was performed on the models ensemble to assess its structural similarity using a pairwise rigid-body superposition that minimizes the root mean squared deviation (r.m.s.d.) between the superimposed conformations, as implemented in TADbit.

Structural analysis of the 3D chromosome model. Using TADbit, a set of descriptive measures were calculated to analyse the structural properties of each particle in the ensemble model: (1) accessibility, measuring how accessible a particle is from an external object of radius of 150 nm; (2) density, measuring the local compactness of the chromatin fibre; (3) contact density, counting the number of particles within a given spatial distance (2× particle size) from a specified particle; and (4) walking angle, measuring the angle formed by a particle and its two immediate neighbor particles. Using these descriptive measurements, we calculate local proprieties of the ensemble model. We define local accessibility of a particle as the mean accessibility value of a given particle over all N models in the ensemble. Similarly, we define local contact density of a particle as the mean contact density of a given particle over all N models in the ensemble; this is used as a proxy for local packing density.

A density map is generated for each cluster in the ensemble using the *molmap* command in Chimera[51]. The density map is represented by intensities at points i (ρ_i) on a cubic grid with a grid space of resolution in kb divided a factor of 3. Each structure is defined by its beads coordinate in Cartesian space and a mass corresponding to 1 bead unit.

The density ρ_i is defined with a Gaussian function as:

$$\rho_i = \sum_N \frac{Z_N}{(\sigma\sqrt{2\pi})^3} e^{-\frac{(x-x_n)^2+(y-y_n)^2+(z-z_n)^2}{2\sigma^2}}$$

x, y and z are the Cartesian coordinates of particle N, Z_N is the mass and σ is set equal to the resolution in kb (100). Density maps for specific compartments or regions are generated accordingly.

The distances between the centers of mass (COMs) of different density map regions were calculated to analyse the relative positioning of high- and low-valency regions with respect to the B compartment. Analysis was performed with Chimera[51].

Hi-C contact enrichments between genomic valency segments. We computed contact enrichments between all PCA-derived Hi-C regions (A and B regions,

excluding chromosome X) using a statistical model based on a binomial test to estimate the significance of contacts as implemented in SeqMonk (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). Contacts with P values ≤ 0.05 and at least five observations were reported. High- and low-valency transcripts (see method paragraph on valency) were scored as 1 and −1, respectively, and for every A, B genomic region the mean score of valency transcripts overlapping the region was calculated. We selected regions with at least three transcripts with assigned valency for further analysis. Low-valency regions were defined based on scores smaller than or equal to the median of all region valency scores; high-valency regions had bigger scores than the median. Regions that did not overlap with any valency transcript were classified as no-valency regions. Empirical cumulative distribution functions and means of $\log_2$ observed/expected contacts (contact enrichment) were plotted for groups of contact pairs as indicated (Fig. 6c).

Regression of intronic read density decay. We implemented a linear regression analysis to estimate the read decrease along introns into the SeqMonk suite (https://www.bioinformatics.babraham.ac.uk/projects/seqmonk/). Briefly, Proximity RNA-seq, pool of p2, p5, p8 and p7 libraries, was used as a 1D RNA-seq dataset after read de-duplication based on genomic position and ignoring RNA particle-specific barcodes. We only considered introns of at least 30 kb in length. Reads were binned into 500 bp windows for linear regression. Regressions were only reported if the slope was negative and the P value did not exceed 0.05. All slope values were reported multiplied by a factor of 1,000,000. For genomic intervals overlapped by multiple intron isoforms, the mean slope of all isoforms was used. Genomic regions with two or more valency transcripts and three or more introns with estimated elongation rates were included in the analysis.

Statistics. Proximity RNA-seq libraries were generated from multiple independent SH-SY5Y cell cultures. The frequency with which pairwise RNA cobarcoding occurred at least as many times in simulations as in observed data generated P values for RNA–RNA proximities. P values were then background-corrected based on local distributions of observed and simulated RNA–RNA proximity counts modeled as normal distributions.

In follow-up analyses exact sample sizes and statistical tests, all two-sided, are noted.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Proximity RNA-seq and Hi-C raw sequencing data are available on Gene Expression Omnibus accession: GSE129732.

Code availability

Code for Hi-C and Proximity RNA-seq analysis is available on github: https://github.com/3DGenomes/TADbit and https://github.com/StevenWingett/CloseCall, respectively.

References

40. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
41. Rubin, A. J. et al. Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation. *Nat. Genet.* **49**, 1522–1528 (2017).
42. Tsanov, N. et al. smiFISH and FISH-quant – a flexible single RNA detection approach with super-resolution capability. *Nucleic Acids Res.* **44**, e165 (2016).
43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
44. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
45. Kaimal, V., Bardes, E. E., Tabar, S. C., Jegga, A. G. & Aronow, B. J. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res.* **38**, W96–102 (2010).
46. Yu, G. et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).
47. Wingett, S. et al. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
48. Imakaev, M. et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
49. Trussart, M. et al. Assessing the limits of restraint-based 3D modeling of genomes and genomic domains. *Nucleic Acids Res.* **43**, 3465–3477 (2015).
50. Gerchman, S. E. & Ramakrishnan, V. Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proc. Natl Acad. Sci. USA* **84**, 7802–7806 (1987).
51. Pettersen, E. F. et al. UCSF Chimera – a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

nature research

Corresponding author(s): Jörg Morf

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Read sequences were acquired using Illumina instrumentation and software. For dual RNA-FISH, probe design was performed with Stellaris probe designer (www.biosearchtech.com), 3D FISH distances were measured using the Imaris software (Bitplane).
Data analysis	All software and code used in this study has been described in published literature (Bowtie v2.2.6, HICUP v0.5.8, Bowtie v2.3.2, Hisat2 v2.1.0, SeqMonk v1.35.0, R v3.1.3, ToppCluster) or are custom made and available on github, https://github.com/3DGenomes/TADbit and https://github.com/StevenWingett/CloseCall.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Proximity RNA-seq and Hi-C sequencing data are available on Gene Expression Omnibus (GEO) accession: GSE129732.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Proximity RNA-seq with identical experimental parameters was performed using 3 biological replicates. Results were pooled for subsequent analysis. Hi-C experiments were performed in duplicates and reads pooled for further analysis.
Data exclusions	No data was excluded.
Replication	All attempts to replicate data were successful.
Randomization	No randomization was required for this study since no comparisons were made between samples/experimental groups.
Blinding	No blinding was required for this study since no comparisons were made between samples/experimental groups.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	SH-SY5Y human neuroblastoma cell line (Sigma)
Authentication	SH-SY5Y cells were identified based on RNA expression (Proximity RNA-seq expression) and gene ontology.
Mycoplasma contamination	Cell line was tested for mycoplasma contamination by vendor only.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell line was used.