Epigenetics & Chromatin

RESEARCH Open Access

RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships

Julie Nadel[1], Rodoniki Athanasiadou[1,5], Christophe Lemetre[1,6], N. Ari Wijetunga[1], Pilib Ó Broin[1], Hanae Sato[1], Zhengdong Zhang[1], Jeffrey Jeddeloh[2], Cristina Montagna[1], Aaron Golden[1], Cathal Seoighe[3] and John M. Greally[1,4*]

Abstract

Background: RNA:DNA hybrids represent a non-canonical nucleic acid structure that has been associated with a range of human diseases and potential transcriptional regulatory functions. Mapping of RNA:DNA hybrids in human cells reveals them to have a number of characteristics that give insights into their functions.

Results: We find RNA:DNA hybrids to occupy millions of base pairs in the human genome. A directional sequencing approach shows the RNA component of the RNA:DNA hybrid to be purine-rich, indicating a thermodynamic contribution to their in vivo stability. The RNA:DNA hybrids are enriched at loci with decreased DNA methylation and increased DNase hypersensitivity, and within larger domains with characteristics of heterochromatin formation, indicating potential transcriptional regulatory properties. Mass spectrometry studies of chromatin at RNA:DNA hybrids shows the presence of the ILF2 and ILF3 transcription factors, supporting a model of certain transcription factors binding preferentially to the RNA:DNA conformation.

Conclusions: Overall, there is little to indicate a dependence for RNA:DNA hybrids forming co-transcriptionally, with results from the ribosomal DNA repeat unit instead supporting the intriguing model of RNA generating these structures *in trans*. The results of the study indicate heterogeneous functions of these genomic elements and new insights into their formation and stability in vivo.

Keywords: RNA:DNA hybrid, R-loop, Chromatin, DNA methylation, Transcription factor, Transcription, Non-coding RNA, Mass spectrometry

Background

The complex regulatory process leading to gene expression involves, as a major upstream influence, the effects of transcription factors (TFs) binding to specific DNA motifs. The targeting of TFs to specific locations is an informational puzzle, as the number of potential binding sites represented by their generally short sequence binding motifs vastly exceeds the minority used in vivo. This observation suggests that there is additional information present in genomic organization that determines the selection of this subset of sequence motifs. Studies aiming to identify these extra layers of genomic information have revealed influences of chromatin organization [1–4] and DNA methylation [4–6], each of which can facilitate or reduce TF binding to cognate motifs, but the role of the conformation of the DNA molecule in vivo is less well studied. While it is known that nucleic acids can form numerous non-canonical conformations [7], the

*Correspondence: john.greally@einstein.yu.edu
[4] Department of Genetics, Center for Epigenomics and Division of Computational Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA
Full list of author information is available at the end of the article

influence of these conformations in living cells remains under-studied. There is, however, evidence from in vitro assays that DNA conformation influences binding of proteins [8]. As examples, the SP1 transcription factor binds preferentially to the intra-strand G-quadruplex structure in vitro [9], while we have found the methyl-binding domain of the Mecp2 protein to bind preferentially to single-stranded DNA (ssDNA), also in vitro [10]. These observations indicate that the exploration of these and other non-canonical structures occurring in vivo may be fruitful in adding a layer of information to the transcriptional regulatory processes. The potential for ssDNA to occur in living cells, prompted by the results of our Mecp2 studies [10], raised the question about how such structures could be created and maintained stably in vivo. One candidate process to mediate the stable formation of ssDNA is the generation of an RNA:DNA hybrid on one DNA strand leaving the other strand in a single-stranded conformation, a nucleic acid structure referred to as an R-loop [11].

Formation of an R-loop has multiple potential consequences in terms of local organization of transcriptional regulatory elements. The helical conformation of the RNA:DNA hybrid differs from the B-form typical of double-stranded DNA (dsDNA), instead creating a conformation intermediate with the A-form associated with dsRNA [12]. A locus forming an RNA:DNA hybrid therefore creates a double-stranded A/B intermediate conformation, with a second target for single-stranded nucleic acid binding proteins on the complementary, displaced DNA strand. Another property of the R-loop is the displacement by the RNA of G-rich ssDNA [13, 14], allowing the formation of intramolecular G-quadruplex structures [15]. The potential that RNA:DNA hybrids may be resistant to the activity of DNA methyltransferases has previously been proposed [16], as has their failure to organize DNA into a nucleosomal conformation [17], further adding to their local influence on nucleic acid organization.

Formation and maintenance of an RNA:DNA hybrid is subject to many influences [13, 14, 18–21]. Transcription of a locus has been positively associated with RNA:DNA hybrid formation [22, 23], presumably by the RNA acting *in cis* with the DNA from which it was transcribed, but there is evidence in yeast that Rad51 can facilitate RNA molecules *in trans* also forming RNA:DNA hybrids [24]. Mutations of enzymes such as RNase H [25], which specifically hydrolyzes the RNA in RNA:DNA hybrids, RNA helicases [26] and topoisomerases [27] have been found to be associated with the increased formation of RNA:DNA hybrids, supporting a model in which these enzymes normally function to remove these structures from the genome. The presence of RNA:DNA hybrids at ribosomal DNA repeats appears to be a conserved feature from yeast [28] to human cells [16], for which any associated physiological role remains unclear. Functionally, RNA:DNA hybrids and their associated ssDNA regions have been found to have numerous properties in vitro and in vivo in a range of organisms. These include involvement in immunoglobulin class switching [29, 30], regulation of gene expression [31], constitutive formation in yeast telomeres [32] and at the origin of replication in mitochondrial DNA [21]. Additionally, these structures have been linked with epigenetic modifications, such as chromatin organization through enrichment at condensed chromatin marked by histone H3 serine 10 (H3S10) phosphorylation in yeast, *C. elegans* and human HeLa cells [33], centromeric heterochromatin [34], and formation at promoter CpG islands lacking DNA methylation [16]. The functions attributed to RNA:DNA hybrids are thus diverse and appear to have a major degree of dependence upon their genomic context.

RNA:DNA hybrids are being increasingly associated with human diseases, with a major concern that their presence predisposes a locus to chromosomal breakage. For example, it has been shown that R-loops are processed by the nucleotide excision repair endonucleases XPF and XPG into double strand breaks [35], and both BRCA1 [36] and BRCA2 [37] have been implicated as major processing enzymes involved in the resolution of RNA:DNA hybrids. The formation of RNA:DNA hybrids has also been associated with a number of neurological diseases. Mutations in the RNA:DNA helicase senataxin (*SETX*) mutations are implicated in the dominant juvenile form of amyotrophic lateral sclerosis type 4 (ALS4) and a recessive form of ataxia oculomotor apraxia type 2 (AOA2) [38]. RNase H2 (*RNASEH2*) mutations are among those associated with Aicardi-Goutières syndrome, in which the accumulation of unusual nucleic acids triggers inflammatory and autoimmune responses [39]. In addition, it is known that triplet repeats are prone to forming unusual nucleic acid structures, including R-loops and RNA:DNA hybrids, a phenomenon conserved in organisms from prokaryotes [40] to mammalian cells [22]. Trinucleotide repeat expansion diseases are therefore being evaluated for a potential contribution of nucleic acid structures to disease pathogenesis, with accumulating evidence that R-loops are involved in Fragile X syndrome [22, 41, 42] and Friedreich's ataxia [42], with similar events also occurring in hexanucleotide repeat expansions [43]. We refer the reader to a number of excellent recent reviews of this topic for more complete insights into these unusual nucleic acid structures and their disease associations [11, 44, 45].

To establish a foundation for understanding their function, we mapped RNA:DNA hybrids genome-wide in vivo in two human cell lines with parallel transcriptional and proteomic studies. These studies provide new insights into how specific loci are preferentially selected as sites of formation of these structures, and allow the inference of some of their likely functional properties. These non-canonical nucleic acid structures occur in ribosomal DNA and at tens of thousands of loci in the remainder of the genome, with sequence characteristics indicating a polypurine-richness of the RNA in the hybrid that is likely to increase the thermodynamic stability of these structures. RNA:DNA hybrids appear to have heterogeneous and context-dependent properties, with subgroups showing relationships with local transcription and chromatin structural features, and a general trend towards decreased DNA methylation. On a more regional scale of hundreds of kilobases, RNA:DNA hybrids are enriched in regions of the genome with a greater abundance of L1 LINEs and CpG islands, and the chromatin modifications indicative of heterochromatin organization. These findings also support the possibility that the RNA generating these RNA:DNA hybrids is frequently generated *in trans*, a set of results that combines to provide new insights into these non-canonical nucleic acid structures in human cells.

Results

RNA:DNA immunoprecipitation (RDIP)

We optimized an assay previously published as DNA:RNA immunoprecipitation (DRIP) [16] to map RNA:DNA hybrids, changing several components of the protocol. These updates include the pre-treatment of the cellular nucleic acid with RNase I, the use of sonication with the goal of minimizing bias in fragmenting the nucleic acid, and the addition of directional information about the strand derived from the RNA component of the hybrid. Given the extensive changes made, we distinguish the updated assay with the new acronym RDIP (RNA:DNA immunoprecipitation). The assay is based on the use of the S9.6 antibody, which is believed to recognize the intermediate A/B helical RNA:DNA duplex conformation, with little to no sequence specificity [46]. We performed extensive in vitro testing of the antibody to reconfirm these properties, including electrophoretic mobility shift assays and South-Western blots of oligonucleotides (including RNase H pre-treatment) that confirmed the necessary RNA:DNA hybrid specificity of the antibody (Additional file 1: Figure S1a–e).

The in vivo studies were focused on the primary, non-transformed, diploid IMR-90 lung fibroblast cell line because of the substantial genome-wide data available from the Roadmap Epigenomics Program [47]. For comparison, we isolated a clone of HEK 293T cells that we found to have the least copy number variability of several tested as determined by array comparative genomic hybridization (Additional file 1: Figure S1f). The immunoprecipitation using sonicated whole cell nucleic acid, pre-treated with RNase I, was optimized, and tested using a Southern dot blot using a $(TTAGGG)_n$ probe to confirm enrichment of the telomeric TERRA-associated R-loop [32] (Additional file 1: Figure S1g). This pre-treatment with RNase I was recently shown to be necessary to reduce noise due to the S9.6 antibody detecting RNA in unusual conformations [48]. To allow the immunoprecipitated RNA:DNA hybrid to be ligated into sequencing adapters, an approach derived from RNA-seq library preparation was used. This provided the opportunity to introduce dUTP during second strand synthesis to reveal directional information about the strand on which the RNA molecule was located [49]. To confirm the RDIP-seq assay worked, we used peak calling analytical methodologies borrowed from ChIP-seq to identify the locations of RNA:DNA hybrids, followed by the use of single locus quantitative PCR to confirm enrichment in the immunoprecipitated material at these loci (Additional file 1: Figure S1h). Peaks were also verified at further loci using the orthogonal approach of bisulphite sequencing of non-denatured DNA to demonstrate the presence of the ssDNA that occurs at R-loops [50] (Additional file 1: Figure S1i).

Subcellular localization studies

The subcellular localization of RNA:DNA hybrids has been studied in multiple organisms using a number of techniques [16, 24, 37, 51] and was investigated in the current study using two separate approaches. The first used limited amplification of the HEK 293T RDIP-seq library with a PCR primer to which the Texas Red fluorophore had been conjugated. This was hybridized to control human metaphases for visualization. As early results suggested that the pericentromeric region of chromosome 9 was generating signal, a locus-specific probe targeting the subtelomeric region of the p arm of this specific chromosome was included in the fluorescence in situ hybridization (FISH) study. Figure 1a depicts the results of these studies. A strong signal at the centromere of chromosome 9 is observed, as well as from the p arms of the acrocentric chromosomes, indicating enrichment at the Nucleolar Organising Regions (NORs), where ribosomal DNA (rDNA) repetitive sequences are located.

The second subcellular localization approach employed was to use the S9.6 antibody for immunofluorescence of the HEK 293T cells. Consistent with previously published studies [16, 52], a subnuclear enrichment within nucleoli (confirmed with an anti-fibrillarin antibody, Fig. 1b) was observed. Of note was the additional cytoplasmic signal that has also been noted in prior studies [52]. This signal

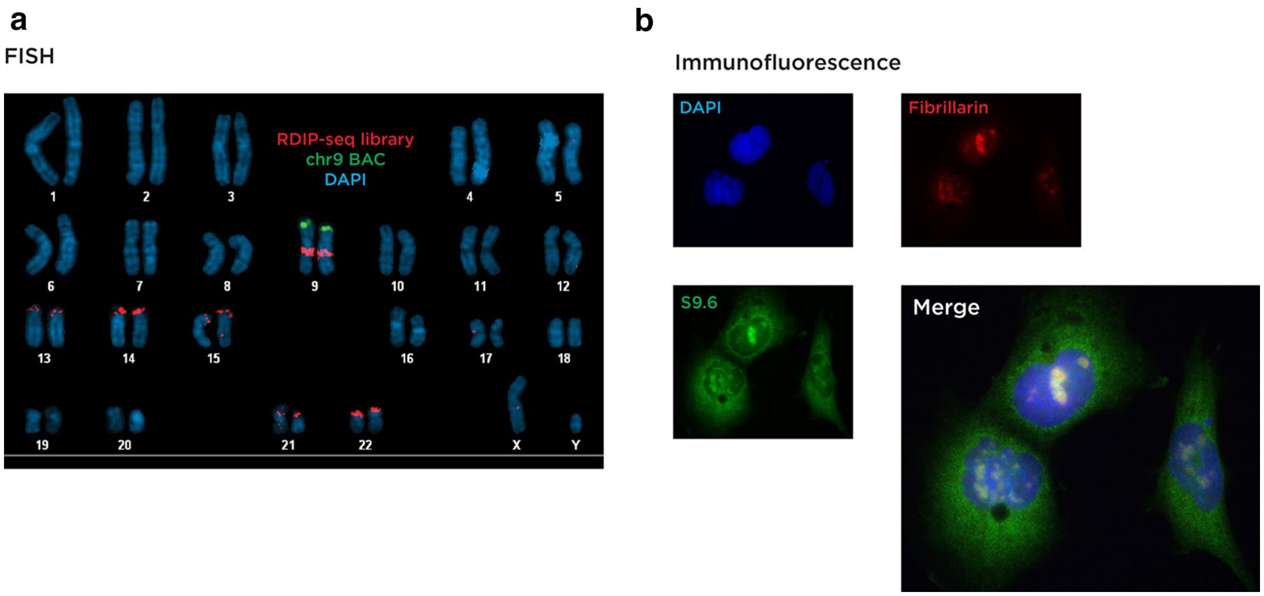

Fig. 1 Subcellular localization studies. In panel **a** we show the results of hybridization of the fluorescently-labeled RDIP-seq library to a control male metaphase preparation. The RDIP-seq library is shown in *red*, a bacterial artificial chromosome (BAC) probe mapping to chromosome 9 in *green*, and DNA counterstained by DAPI in *blue*. We observe a specific strong signal from the RDIP-seq library mapping to the p arms of acrocentric chromosomes (HSA13-15 and HSA21-22), indicating enrichment at the nucleolar organizing regions (NORs) encoding ribosomal RNAs, and at the pericentromeric region of chromosome 9. In panel **b** we show the results of immunofluorescence using the S9.6 antibody (*green*) with an antibody to fibrillarin (*red*), demonstrating co-localization with the intranuclear S9.6 antibody signal (merge) and therefore enrichment in nucleoli. Further signal from the nuclear periphery and the cytoplasm using S9.6 is also observed, which may represent detection by this antibody of RNA conformations rather than RNA:DNA hybrids specifically [48]

may in part reflect signals from mitochondrial DNA [53] or the S9.6 antibody detecting ssRNA in unusual conformations [48].

Ribosomal DNA studies

Prompted by the co-localization with the NORs seen in the subcellular localization studies, further investigation into RNA:DNA hybrid formation within ribosomal DNA was undertaken. The IMR-90 RDIP library was sequenced and mapped to a human reference genome including the consensus ribosomal DNA repeat unit [54] (accession number gi|555853|gb|U13369.1|HSU13369), following the same approach as Zentner and colleagues [55]. The results showed that ~2 % of reads mapped to the ribosomal DNA repeat unit and the remainder to the sequenced majority of the human genome. The mapping of reads to the rDNA repeat unit is shown in Fig. 2. The immunoprecipitated RNA:DNA hybrids map heterogeneously within this repeat unit, with accumulation of reads at the known exons of the rDNA gene, and others in the intergenic spacer (IGS) region.

To determine the relationship between the RNA:DNA hybrids and the transcribed sequences, RNA-seq on total RNA from the IMR-90 cells was performed without polyA selection or depletion of ribosomal RNA. This allowed deep sequencing of the expressed rRNA and co-localization with the RNA:DNA hybrid reads (Fig. 2). The RDIP-seq reads in the 5′ end of the repeat unit are precisely co-localized with the RNA-seq reads, but there is RNA:DNA hybrid formation with comparable read enrichment in the IGS region. Using K562 cell ChIP-seq data provided by Zentner and colleagues [55], the RNA:DNA hybrids are found to be located upstream from the rDNA promoter and flanking the candidate *cis*-regulatory sequence in the IGS region (Fig. 2). The intergenic candidate *cis*-regulatory sequence was also shown to occur in embryonic stem cells, umbilical vein cells and normal human epidermal keratinocytes [55], and thus appears to be constitutive. It is therefore reasonable to predict that the element is also present in the IMR-90 cells. Some of the rDNA RDIP-seq signal is attributable to RNA:DNA hybrid formation involving the canonical rRNA transcript, but further RNA:DNA hybrids are formed in the IGS ribosomal DNA region sparing the regions containing candidate *cis*-regulatory elements.

Genome-wide studies

Having defined the source of the rDNA signal, the focus turned to the majority of reads that mapped to the remainder of the sequenced genome. There are tens of

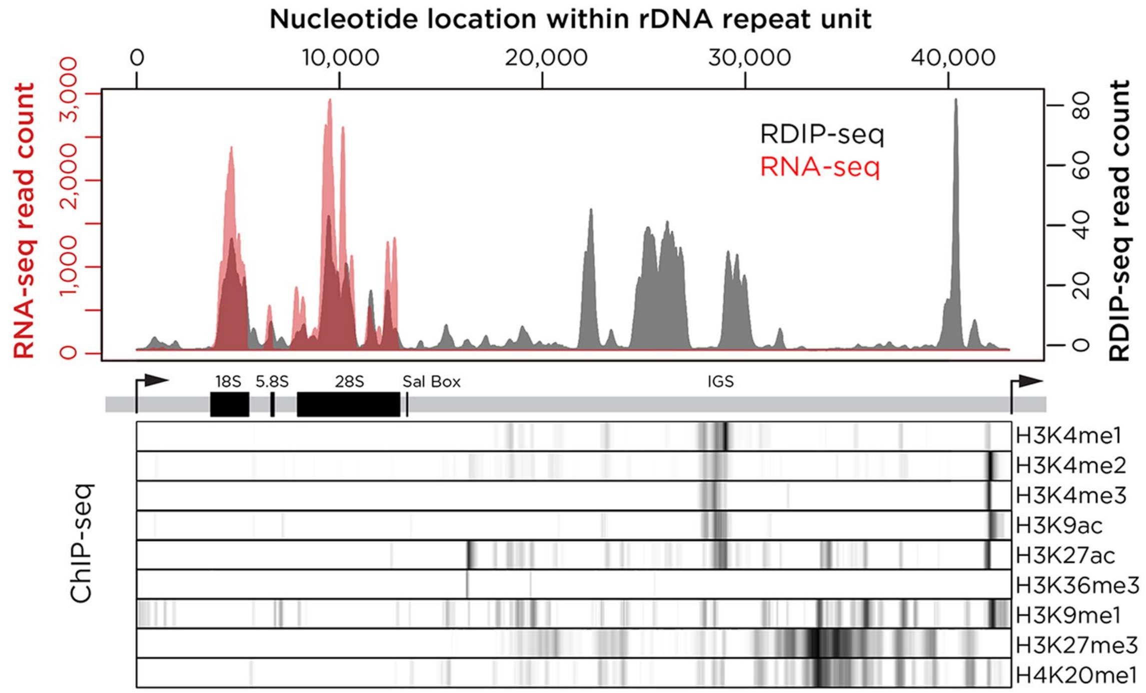

Fig. 2 Mapping of RNA:DNA hybrids within the ribosomal DNA repeat unit. The *upper panel* shows the results of RDIP-seq (*gray*) and RNA-seq (*red*), with genomic annotations and results of ChIP-seq analysis in K562 cells [55] plotted below. RDIP-seq and RNA-seq data are both represented using a smoothed plot showing the number of reads aligned to each basepair of the repeating unit, while the ChIP-seq data signal intensity represents the mean value of non-overlapping 50 bp windows. RDIP-seq values were normalized by subtracting the frequencies of aligned reads of the input sample in each window. We find that RNA:DNA hybrids co-localize with the rRNA transcripts, but that there are also RDIP-seq peaks of comparable magnitude in the intergenic spacer (IGS) where no transcriptional activity is apparent from RNA-seq. The RNA:DNA hybrids in the IGS are upstream of the promoter region and flank the upstream candidate *cis*-regulatory sequence where there is H3K4 methylation and acetylation of H3K9 and H3K27

thousands of RNA:DNA hybrid-forming loci (mapped as peaks using a ChIP-seq analytical approach) throughout the human genome (Additional file 1: Figure S2), the same magnitude observed previously in DRIP-seq experiments [16]. There is a significant enrichment for loci shared by IMR-90 and HEK 293T cells, indicating that many RNA:DNA hybrid-forming loci may be constitutive across cell types. Focusing on the loci in the human diploid IMR-90 fibroblast cell line, RNA:DNA hybrids are demonstrated to be distributed genome-wide, with most of the peaks located in intergenic regions (Fig. 3a). The enrichment of peaks in each of these major genomic contexts was calculated and the significance of enrichment was tested based on overlap (nucleotide occupancy) using permutation analyses. Figure 3b shows that promoters (and the highly correlated CpG island feature) are strongly enriched for RNA:DNA hybrids, and that they are distributed elsewhere in the genome at close to expected frequencies, apart from a modest but significant depletion at RefSeq gene bodies and intergenic regions (excluding promoter and lncRNA sequences).

As RNA:DNA hybrids in yeast have been shown to be enriched at transposons [28], their representation within sequences annotated as repetitive within the human genome was explored. In Fig. 3b, the sequences annotated as low complexity and simple repeats by RepeatMasker are shown to be the most strongly overrepresented, but satellite repeats are also found to be enriched in RNA:DNA hybrids. When the low complexity repeats were explored in greater detail, the strand on which the RNA component of the RNA:DNA hybrid was located was found to be composed of GA-rich, G-rich, and A-rich families of low complexity repeats. Additionally, within the satellite repeats that co-localized with the RNA of RNA:DNA hybrids, 76.5 % of the repeats were $(GAATG)_n$ sequences.

It is known that purine-rich RNA binds in vitro with greater affinity to its pyrimidine-rich DNA complement than the equivalent purine-rich DNA sequence [12, 20], which may indicate a role for biochemical stability maintaining RNA:DNA hybrids in vivo. As the analyses of repetitive sequences suggested enrichment of purine-rich RNA in these RNA:DNA hybrids, this finding

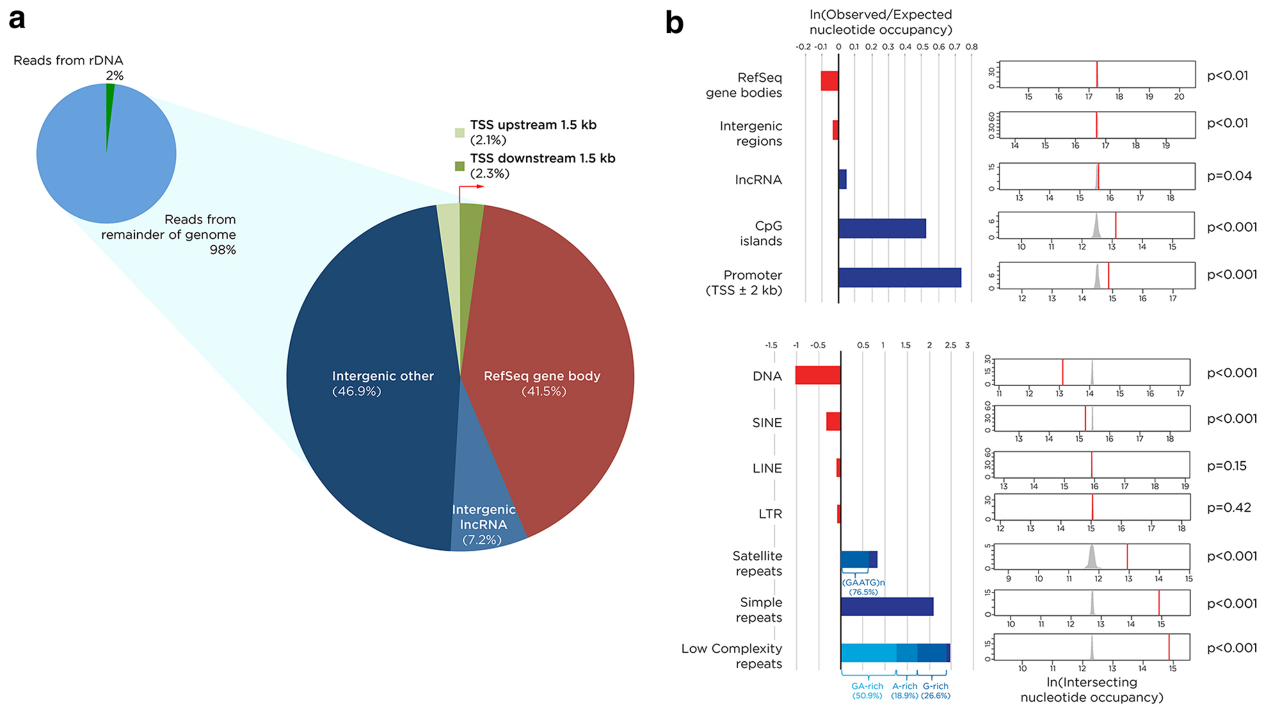

Fig. 3 Genomic distribution of RNA:DNA hybrids. In panel **a** we show that the proportion of reads mapping to rDNA is 2 %, and break down the remaining 98 % by genomic context, showing the majority of RNA:DNA hybrids (called as *peaks* using ChIP-seq analytical approaches) to be located in intergenic regions. To understand these RNA:DNA hybrid distributions, we calculated observed/expected ratios based on nucleotide occupancy of genomic features, and performed permutation analyses testing for the likelihood of randomized intersection (**b**), the results of which are shown in Additional file 2: Table S1. We found depletion of RNA:DNA hybrids at RefSeq gene bodies, intergenic regions, and SINE and DNA transposable elements but significant enrichment at promoters and CpG islands, and a number of purine-rich repetitive sequences

was explored more fully, testing for and finding from the genome-wide data a strong intramolecular skewing towards GA:CT enrichment (Fig. 4a). To test globally whether this purine (GA) enrichment was present on the RNA-containing strand, the directional sequence information was used to examine nucleotide skewing on each strand at RNA:DNA hybrids, confirming the RNA-derived sequence to be strongly purine-enriched (Fig. 4b). The 10 % of peaks with the least tendency towards having the RNA enriched on one strand were removed from further analyses as being likely to over-represent experimental noise.

Relationship of RNA:DNA hybrids to local transcription

As some RNA:DNA hybrids have been found to have transcriptional termination properties [56, 57], it was tested whether the RDIP directional sequencing allowed the observation of the an orientation bias within genes. This tendency has been observed for transposable elements, which are believed to have different effects on gene function depending on their insertion orientation in gene bodies [58, 59]. The nucleotide skewing within each peak was visualized, revealing the purine-enriched component to be displaced 5′ from the mid-point of the peak (Additional file 1: Figure S3), which is consistent with the RDIP protocol using the RNA component of the RNA:DNA hybrid to prime second strand synthesis, proceeding unidirectionally 3′ and relatively under-representing the region 5′ to the RNA. This observation is independently supportive of the RNA component of the RNA:DNA hybrid being purine-enriched. There was a modest orientation bias against purine-rich sequences in the same orientation as the gene (Additional file 1: Figure S3b), indicating that most but not all genes tolerate an RNA:DNA hybrid with the RNA on the transcribed strand.

To explore the relationship between RNA:DNA hybrid formation and transcription further, the proportions of genes with peaks were tested for transcription states from the RNA-seq data, finding that most transcribed RefSeq genes do not contain RNA:DNA hybrids but that the transcribed genes have a higher frequency of RNA:DNA hybrids than non-transcribed genes (7.75 % compared with 6.09 %; Fig. 5a). The locations of these RNA:DNA hybrids within genes was defined using a metaplot, identifying the first ~ 1.5 kb downstream from

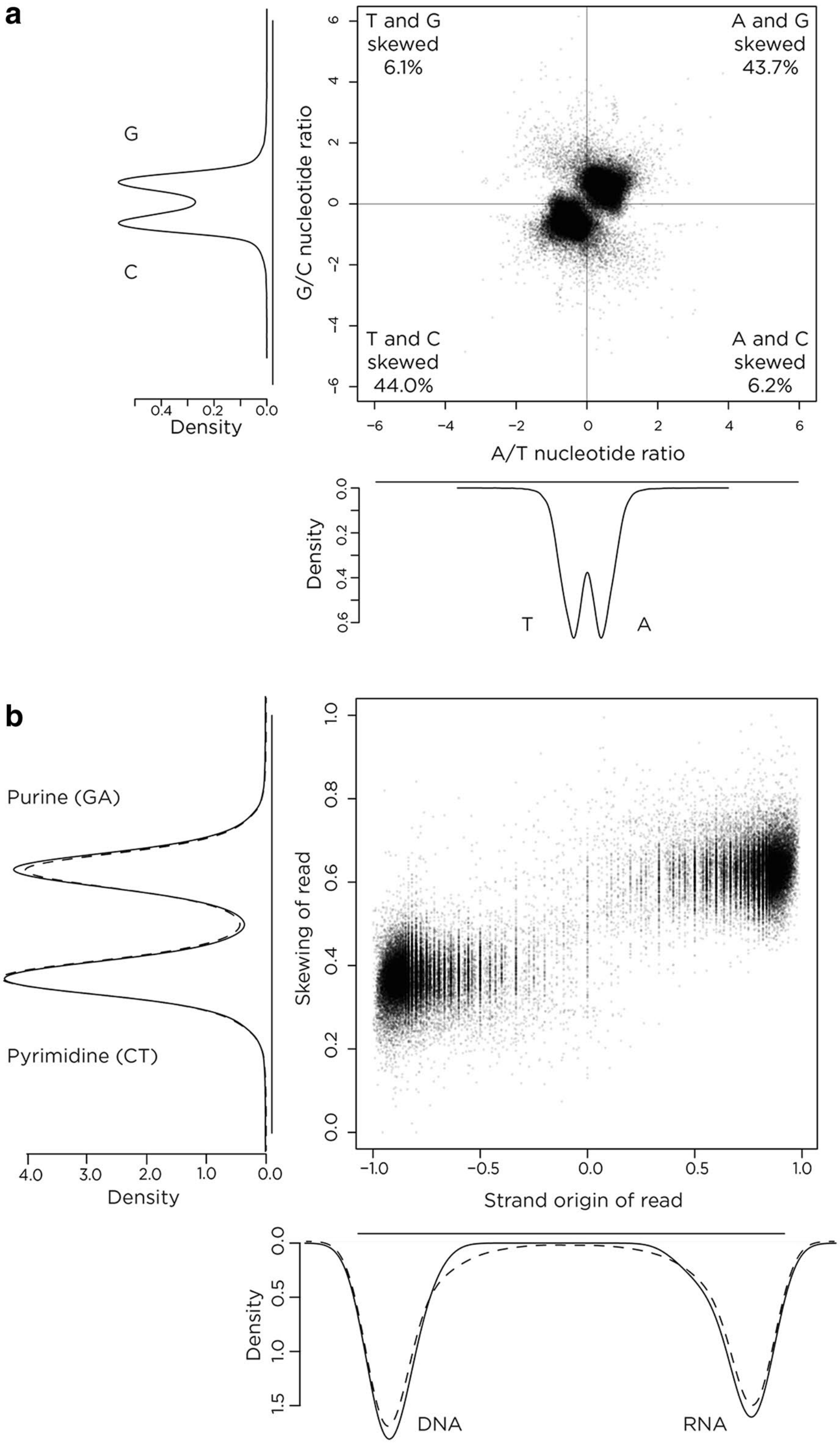

Fig. 4 Nucleotide skewing analyses. In panel **a** we plot the skewing within a strand of A compared to T (*x* axis) or G compared to C (*y* axis) in the RNA:DNA hybrid peaks genome-wide. We find that the peaks are strongly over-represented for purine (G+A) and pyrimidine (C+T) skewing. As our sequencing approach allowed us to identify the RNA and DNA-derived strands separately in the RNA:DNA hybrid, in **b** we proceeded to test whether there was a relationship between skewing (based on the number of G+A divided by the total number of nucleotides) and each type of nucleic acid-derived sequence, finding a clear enrichment for purine skewing on the RNA-derived strand

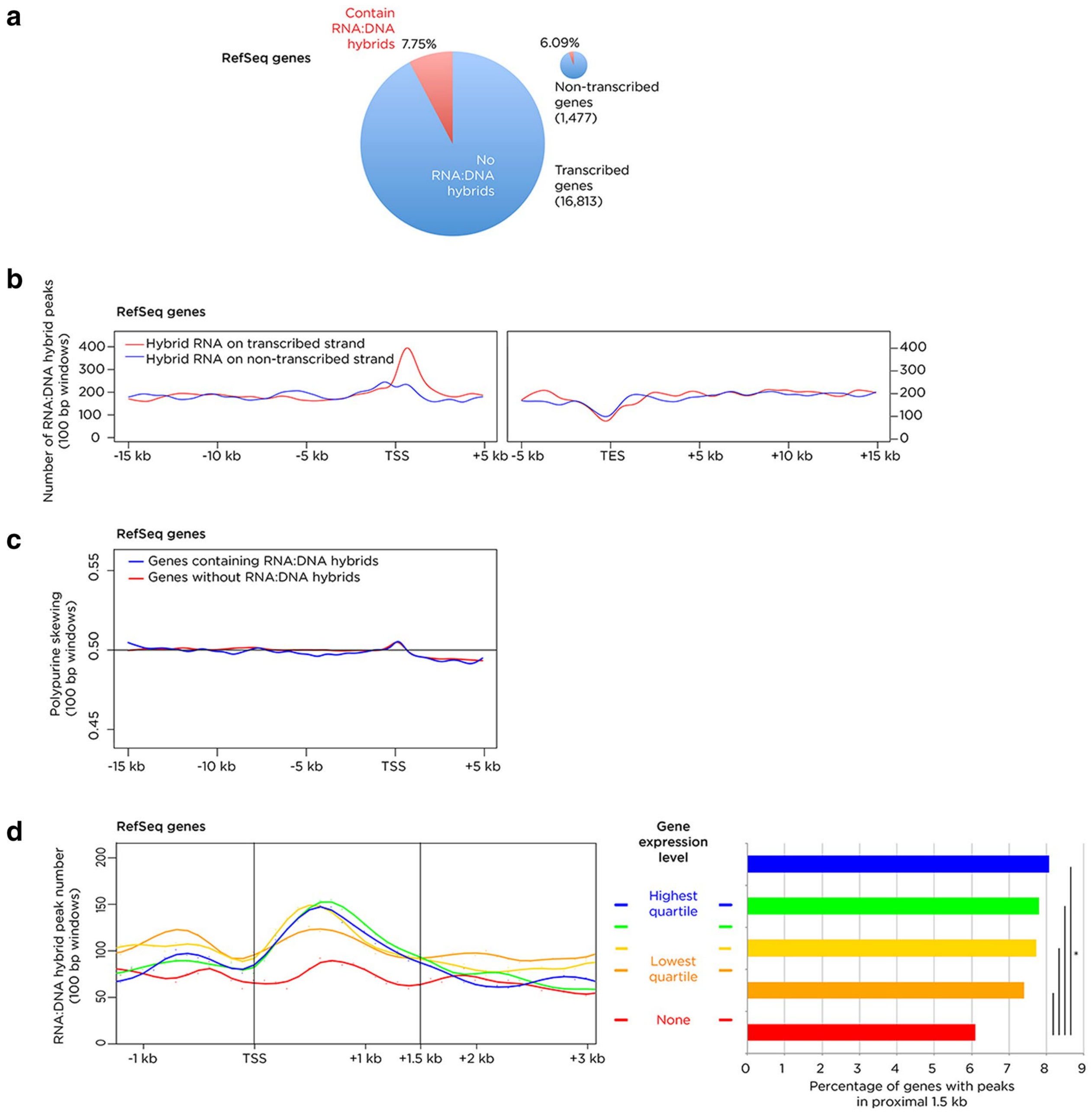

Fig. 5 Transcriptional relationships of RNA:DNA hybrids. In **a** the proportion of RNA:DNA hybrid peaks in transcribed genes is shown to be higher than in non-transcribed genes, but that the majority of genes do not contain RNA:DNA hybrids. In **b** a metaplot of RNA:DNA hybrid peaks is shown, illustrating the number of peaks intersecting with 100 bp windows, with the RNA of the hybrid on the transcribed strand of the gene (*red*) or the opposite strand (*blue*). This revealed an enrichment of the RNA-derived sequence on the transcribed strand in the first ~1.5 kb downstream from the transcription start site (TSS). A depletion of RNA:DNA hybrids is found at the transcription end site (TES). In **c** we show that the region immediately downstream from the TSS is purine-skewed, represented by skewing values of 100 bp windows averaged for all genes, but that this is to the same degree in genes that form RNA:DNA hybrids (*blue*) as those genes that do not form these structures (*red*). In **d** a metaplot of RefSeq genes (*left*) shows that the transcription level of genes (as measured by RNA-seq) is positively associated with the number of RN:DNA hybrids intersecting with 100 bp windows immediately downstream of the TSS. This reflects only modest increases in the small proportions of genes forming peaks (*right*), though found to be a significant relationship using a proportions test

the transcription start site (TSS) as the region most consistently enriched (Fig. 5b). This region is also found to be modestly enriched in purine skewing for genes with and without RNA:DNA hybrids (Fig. 5c). Surprisingly, given the transcriptional termination properties attributed to RNA:DNA hybrids [56, 57], the transcriptional end site is notable for a slight depletion of these structures (Fig. 5b). The information from lncRNAs also suggests a modest enrichment for RNA:DNA hybrids in the immediate vicinity of the TSS (Additional file 1: Figure S4). The local generation of RNA:DNA hybrids has previously been described to be associated with transcription of the region [14, 22], so the genes were stratified by expression level, finding that the proximal 1.5 kb region downstream from the TSS showed an increase in peaks associated with increasing quantiles of gene expression states (Fig. 5d). The conclusion is that transcriptional levels have effects on the likelihood of forming RNA:DNA hybrids, and that local purine enrichment may increase the tendency of these structures to be formed in the ~1.5 kb immediately downstream of the TSS in a small subset of genes.

As RNA-seq measures the steady state of RNA in the cell, and does not necessarily reflect active transcription, we added an analysis of global run-on sequencing (GRO-seq) data generated previously for IMR-90 cells [60]. We found 24,875 RNA:DNA hybrids in these cells to map to RefSeq genes, 24,174 to active genes of which 23,717 had evidence for transcription from GRO-seq data. Interestingly, of the 788 RNA:DNA hybrids mapping to genes with no evidence for transcriptional activity from our RNA-seq data, 733 mapped to loci where GRO-seq indicated local transcription. Exploration of the 57,308 RNA:DNA hybrids mapping to intergenic regions revealed 11,825 to map to loci where GRO-seq indicated transcription. These data suggested GRO-seq to be a more sensitive indicator of transcription than RNA-seq or genomic context, prompting us to explore how many of the RNA:DNA hybrids mapped to loci where GRO-seq indicated transcription, finding evidence for transcription at 27,352 (47.7 %) of the RNA:DNA hybrids.

Relationship of RNA:DNA hybrids to regulators of transcription

To begin to infer any transcriptional regulatory function of the RNA:DNA hybrids from their genomic locations, studies were performed correlating RNA:DNA hybrid locations with enrichment or depletion for other chromatin and transcriptional regulators directly overlapping the RNA:DNA hybrids. Using IMR-90 bisulphite sequencing data from the Roadmap Epigenomics Project (accession number NA000020923.1), a modest decrease in DNA methylation within RNA:DNA hybrids was found compared with genome-wide levels, a finding which is consistent with the hypomethylation of DNA previously observed for RNA:DNA hybrids at CpG islands [16] (Additional file 1: Figure S5a). In vitro studies have shown RNA:DNA hybrids to be refractory to the formation of nucleosomal structures [17], a finding supported by the observation that 7.46 % of all RNA:DNA hybrids overlap DNase hypersensitive sites, representing a significant association genome-wide (Additional file 1: Figure S5b, c). We also tested whether there may be unrecognized transcription at enhancer RNAs [61, 62] at sites of formation of RNA:DNA hybrids. We found that of the 57,308 RNA:DNA hybrids, 4277 (7.46 %) map to DNase hypersensitive sites, of which 3560 have the H3K27ac modification, most of which (3434) also have the H3K4me1 modification. Of these H3K27ac/H3K4me1 DNase hypersensitive sites with RNA:DNA hybrids, 2154 have evidence for transcription from GRO-seq. As enhancer RNAs are inherently unstable [63], it is possible that transcription is actually occurring at a higher proportion of these loci that the GRO-seq data would indicate, potentially linking transcription with RNA:DNA hybrid formation at as many as the 7.46 % located at DNase hypersensitive sites. A motif analysis of RNA:DNA hybrid-forming loci genome-wide revealed an enriched polypurine $(GGAA)_n$ sequence, which has been associated with binding by the FLI1 transcription factor [64] (Additional file 1: Figure S6a).

A notable macro-scale organization of RNA:DNA hybrids was apparent in the human genome, with regions of dense and sparse RNA:DNA hybrid formation (example shown in Additional file 1: Figure S6b). Using publicly-available ChIP-seq data from the IMR-90 cell line, it was possible to ask whether RNA:DNA hybrids in the human genome occur in regions of distinctive regulatory characteristics. We have previously noted that there is extensive inter-correlation of genomic features [65], making it difficult to discriminate specific associations when there are multiple correlating genomic variables. In order to explore the transcriptional and regulatory context of RNA:DNA hybrid peaks, regression models were fitted to the data, regularized using the least absolute shrinkage and selection operator (LASSO; [66]) with the peak density as the response variable. Least angle regression (LARS; [67]) was used, progressively adding covariates to the model and testing the significance of each added predictor using the covariance test statistic proposed by Lockhart et al. [68]. The results of this procedure are shown in Fig. 6. The first covariate to enter the model as significantly enriched in co-localization with RNA:DNA hybrids in 500 kb windows is the repressive histone mark, H3 lysine 27 trimethylation (H3K27me3), followed by CpG islands, L1 LINE retroelements and a further

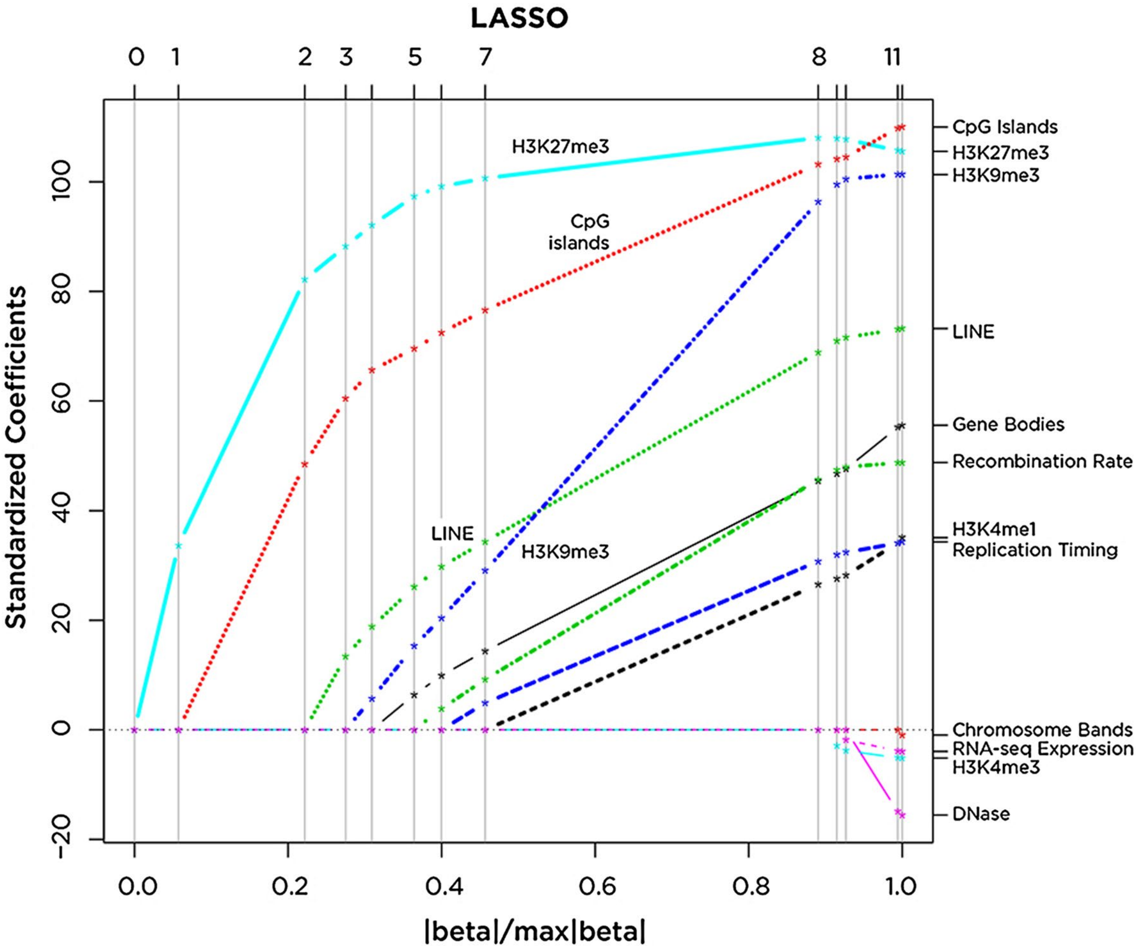

Fig. 6 Macro-scale genomic associations of RNA:DNA hybrids. We used a least absolute shrinkage and selection operator (LASSO) adaptive regression approach to explore the association of genomic sequence features with RNA:DNA hybrid density in 500 kb windows. The figure shows the order in which covariates enter the model as the constraint on the sum of the regression coefficients (*x* axis) is progressively relaxed from 0 to its maximum value (corresponding to the ordinary least squares regression vector)

repressive histone mark, H3K9me3. The first eight covariates to enter the model all gave significant values of the covariance test statistic.

Local chromatin organizational studies using mass spectrometry

Finally, characterization of chromatin located at RNA:DNA hybrids was performed to identify the proteins enriched at these loci. Chromatin from HEK 293T cells was sonicated and a fraction immunoprecipitated with the S9.6 antibody, eluting the protein complexes using RNA:DNA hybrid oligonucleotides, and identifying local proteins through mass spectrometry (Fig. 7a). These results and Western blotting validation of candidate proteins of interest are shown in Fig. 7b and Additional file 2: Table S2. A number of different specific proteins plausibly associated with RNA:DNA hybrids were identified. RNA helicase A (encoded by *DHX9*) is a protein known to be involved in resolving RNA:DNA hybrids [69] and is a necessary partner for FLI1 in tumourigenesis [70], while DNA binding protein B (YBX1) is known to bind to ssDNA [71] which should be part of R-loops formed at these loci. ILF2 and ILF3 are also found in the chromatin at RNA:DNA hybrids. These are transcription factors known to recognize a purine-rich motif [72], with our results raising the possibility that their binding may depend on the target nucleic acid existing in an RNA:DNA conformation.

The presence in local chromatin of RNA helicases and topoisomerases is consistent with prior reports that these enzymes are involved in the removal of RNA:DNA hybrids [26, 42]. The question arose whether the IMR-90 and HEK 293T cells express the genes encoding the broader group of proteins implicated in removal of RNA:DNA hybrids in vivo. Using the RNA-seq data, nine

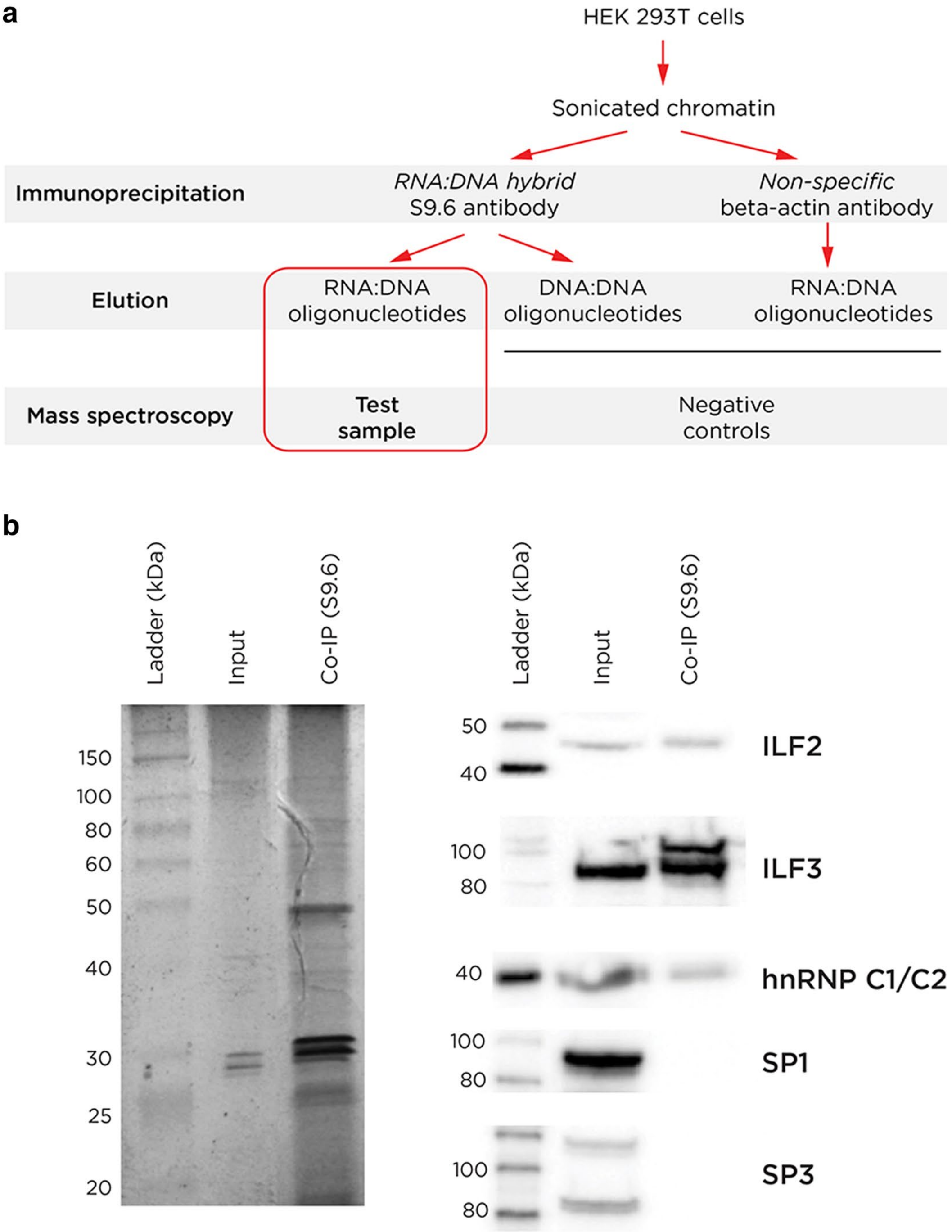

Fig. 7 Chromatin organizational studies at RNA:DNA hybrids using mass spectrometry. In panel **a** we show the experimental approach used for these proteomic studies. In **b** the altered pattern of enriched proteins compared with the input sample is seen using gel electrophoresis, and the results of Western blots confirming the enrichment of specific candidate proteins identified by mass spectrometry (ILF2, ILF3, hnRNP C1/C2), with SP1 and SP3 as controls known to bind to G-skewed DNA motifs

of these genes were categorized into quartiles of expression, finding that all of the genes were expressed at high levels (Additional file 1: Figure S7). The presence of the RNA:DNA hybrids in these cells is therefore in spite of robust levels of expression of genes encoding proteins that should actively remove them.

Discussion

Mapping RNA:DNA hybrids in human cells has allowed new insights into the properties of these non-canonical nucleic acid structures. We confirm through subcellular localization studies prior observations that the ribosomal DNA harbors these structures [16] (Fig. 1). Additionally, we expand on findings in yeast [28] by mapping RNA:DNA hybrid locations within the human rDNA repeating unit, revealing these structures to be formed not only at the expressed rDNA gene but also in the intergenic spacer sequence (Fig. 2). The signal from this repetitive sequence is necessarily composed of all rDNA repeat units in the genome, so we cannot distinguish events occurring within individual alleles, but we can make several inferences. Firstly, that the enrichment of RNA:DNA hybrids within the rDNA repeat unit is not uniform but is enriched at two types of loci, the exons of the rRNA genes and the intergenic spacer sequence where they spare candidate *cis*-regulatory loci (Fig. 2). The mapping of RNA:DNA hybrids to the rDNA gene exons is an interesting finding as it implies that the RNA associating with the rDNA is already spliced and not the primary transcript through the region. This is less supportive of a co-transcriptional model for RNA:DNA hybrid formation [29, 30] and more indicative of rRNA acting *in trans* to generate these structures, as has been found for RNA:DNA hybrids in yeast [24].

The mapping of reads to the rDNA repeat was consistent with the imaging data indicating the presence of RNA:DNA hybrids in nucleoli (Fig. 1), allowing us to proceed with confidence to assess the distribution of the majority of the reads elsewhere in the genome. The first observation was that the RNA:DNA hybrids were not enriched in gene bodies relative to intergenic sequences (Fig. 3a, b), again failing to support their presence being solely a function of recognized transcription. Furthermore, the rRNA model would suggest that spliced mRNAs might associate *in trans* with their genes of origin, but this is not reflected by over-representation of RNA:DNA hybrids in RefSeq genes (Fig. 3b). Instead we observe that a small proportion of genes have peaks within their bodies (Fig. 5a), with a significantly higher proportion of expressed genes than silent genes containing RNA:DNA hybrids (Fig. 5d). These tend to form in the ~1.5 kb immediately downstream of the transcription start site, where they are influenced by the level of transcription (Fig. 5d) but can be found even in genes that are not measurably expressed by RNA-seq (Fig. 5a, d), and are overall depleted in RefSeq gene bodies (Fig. 3b). Analysis of GRO-seq data from IMR-90 cells [60] reveals evidence for transcriptional activity at genes categorized as silent by RNA-seq, and at loci of open chromatin in the genome, which should include sites of transcription of enhancer RNAs [61, 62]. Overall, however, even these sensitive indicators of active transcription only account for a subset of RNA:DNA hybrids in the genome, indicating that transcription through a locus is therefore only moderately influential in generating these structures.

Adding to the tendency of the proximal 1.5 kb to form RNA:DNA hybrids is the enrichment at this location genome-wide for purine-skewed DNA in the transcriptional orientation of the gene (Fig. 5c). We first noticed that purine enrichment may be a property of RNA:DNA hybrids in vivo when we found a strong enrichment for repetitive sequences composed of polypurines in our RepeatMasker analysis (Fig. 3b). We confirm the purine skewing to be a general property of these sequences (Fig. 4; Additional file 1: Figure S3), which extends prior observations that suggested isolated G density [14] or GC [19] skewing, to be characteristic of these loci. As purine-rich RNA binds to complementary pyrimidine-rich DNA with greater affinity than the same purine-rich DNA sequence in vitro [12, 20], this is likely to be a factor in the ability of the RNA to maintain displacement of the ssDNA in the R-loop structure.

While transcriptional termination has been described to be a property of RNA:DNA hybrids [56] (reviewed in [11]), we observe that RNA:DNA hybrids are not enriched at the annotated ends of RefSeq genes and are, in fact, relatively depleted (Fig. 5b). However, we see a small orientation bias in RefSeq genes, with a shift away from RNA:DNA hybrids with the RNA in the same orientation as transcription (Additional file 1: Figure S3). We interpret this to indicate that a subset of RNA:DNA hybrids may cause transcriptional disruption effects, but that it is not a universal property throughout the genome.

We can infer some likely functional properties of RNA:DNA hybrids by genomic co-localization and proteomic approaches. The genomic co-localization studies were both immediately at the RNA:DNA hybrid location and more broadly in their flanking regions, the latter prompted by what appeared to be higher-scale organization of the distribution of these loci (Additional file 1: Figure S6b) and by prior studies in yeast [34]. The immediate local features included DNase hypersensitivity (Additional file 1: Figure S5b, c), which is consistent with prior in vitro published findings that nucleosomes do not readily form on these structures [17]. The tendency of RNA:DNA hybrids to be resistant to acquisition of DNA methylation [16] finds some support from our data, but the modest degree of relative hypomethylation indicates that the effects occur at only a small subset of loci. In the regional analysis of the co-localization of RNA:DNA hybrids and genomic sequence features within 500 kb windows of the genome, the enrichment found for CpG islands was not surprising given our observations that promoter-proximal sequences are enriched in RNA:DNA hybrids (Fig. 3b). However, the enrichment in the same broader regions for the repressive H3K27me3 and H3K9me3 marks was unexpected for structures with the possibility of being co-transcriptionally generated. We interpret this to indicate one of the following three possibilities: that these regions are more transcribed than we can appreciate using the data available to us, allowing co-transcriptional formation of RNA:DNA hybrids, or that RNA forming RNA:DNA hybrids *in trans* is better able to target these regions, or that these structures are more stable in the context of repressive heterochromatin, with a causal model prompted by observations in fission yeast [34] that would involve the RNA:DNA hybrids having a mechanistic role to induce the regional repressive organization.

The proteins revealed by the proteomic studies were consistent with the local presence of RNA:DNA hybrids and R-loops (Fig. 7; Additional file 2: Table S2), including RNA helicase (DHX9) and single-stranded DNA binding properties. We were especially intrigued by the presence of the ILF2 and ILF3 components of the Nuclear Factor of Activated T-cells (NF-AT) transcription factor, which is required for T cell expression of interleukin 2 and represents a target of the immunosuppressive Cyclosporin A and FK506 drugs [73]. ILF2 (NF45) and ILF3 (NF90) are characterized by their binding to polypurine-rich interleukin gene enhancers [72], and are described to have the property of being able to bind to dsRNA in vitro [74]. This property, when combined with our finding of enrichment in chromatin at RNA:DNA hybrids, suggests that the selective binding of NF-AT at specific genomic locations may be dependent upon those sites being in an RNA:DNA hybrid conformation, which is structurally more similar to A-form dsRNA than B-form dsDNA [12]. The sequence motif $(GGAA)_n$ that we found to be enriched at RNA:DNA hybrids (Additional file 1: Figure S6a) closely resembles that of the binding site for the FLI1 transcription factor [75]. FLI1 is a master regulator of hematopoiesis [76] in the ETS family, and has been causally implicated in pediatric Ewing's sarcoma [77]. The oncogenic effect of FLI1 (as a fusion protein with EWS) is enhanced by RNA helicase A [70] which it appears to inhibit [78], an interaction that can in turn be inhibited by small molecules with therapeutic potential [79]. Expression of EWS-FLI1 induces chromatin opening at sequences with the $(GGAA)_n$ motif [80]. The combination of the findings of binding to a polypurine-rich motif and interaction with RNA helicase A combine to suggest that FLI1 may also bind to an RNA:DNA nucleic acid conformation.

The model for RNA:DNA hybrid physiology that results from our studies indicates that they form as a result of an equilibrium between formation, stability and removal, with increased transcription having only a modest influence for the subset we believe to be formed co-transcriptionally. Once formed, those at purine-skewed loci are likely to be more stable thermodynamically, resisting the presence of enzymes like RNA helicase A in the local chromatin and the robust expression of genes encoding proteins that remove RNA:DNA hybrids (Additional file 1: Figure S7), reflecting how these structures persist despite active processes dedicated to their removal. The RNA:DNA hybrids form DNase hypersensitive structures which may facilitate or reflect binding of transcription factors with preferences for either the A/B form RNA:DNA duplex or the ssDNA in the R loop, and exist in large scale domains of repressed chromatin, with which their causal relationship is uncertain. We propose that the weight of evidence supports many of the RNA:DNA hybrids being formed *in trans*, by RNA transcripts originating from regions of the genome other than the location of the RNA:DNA hybrid itself. The ability of RNA to invade a double stranded DNA molecule *in trans* is being strikingly highlighted at present by CRISPR/Cas technology, which creates an RNA:DNA hybrid as part of an R-loop [81]. We find little evidence for the majority of the RNA:DNA hybrids in vivo to be located at recognizably transcribed sequences. More persuasively supporting a *trans* hypothesis is the finding that the RNA:DNA hybrids in the rDNA repeat unit map to processed rather than primary rRNA transcripts. The simplicity of the polypurine-skewed sequences at RNA:DNA hybrids potentially allows a limited number of transcripts to target a large number of loci. The nuclear-retained polypurine-rich RNAs found in mammalian cells represent a type of non-coding RNA of unclear function [82] that could mediate such *trans* effects in vivo. Overall, it appears that there are numerous influences upon physiological RNA:DNA hybrid formation, the dissection of which will be essential if we are to understand the roles ascribed to them in disease states [83].

Conclusions

A systematic analysis of RNA:DNA hybrids in human cells reveals their presence throughout the genome, including in the ribosomal DNA repeat unit, cumulatively representing millions of base pairs of DNA. The results help to resolve a number of conflicting theories about the formation of RNA:DNA hybrids, with only small influences of local transcription found, and evidence indicative of their formation *in trans*. Their sequence characteristics are clearly shown to be defined by purine enrichment for the RNA component of the hybrid, supporting a thermodynamic characteristic of RNA:DNA hybrids that should favor their stability. Functionally, we find evidence linking the presence of these structures to local DNA methylation and local and regional chromatin organizational states, with proteomic studies revealing the presence of transcription factors that may be binding preferentially to the RNA:DNA conformation. The contribution of non-canonical nucleic acid structures in transcriptional regulation is underexplored but warrants further investigation, adding a new layer of information in understanding transcriptional regulation in mammalian cells.

Methods

S9.6 antibody production

The S9.6 antibody-producing hybridoma line was purchased from ATCC (HB08730), and the hybridoma line was grown in Integra Flasks by our institution's monoclonal antibody core facility in serum-free medium. The S9.6 antibody was then purified by the macromolecular therapeutics core facility using a Protein-G column and size exclusion. The antibody was validated using an electrophoretic mobility shift assay (EMSA) and southwestern blotting to test for specificity to RNA:DNA hybrid oligonucleotides. A full description of these experiments is provided in the Additional file 1: Supplemental experimental procedures.

Immunofluorescence

HEK 293T cells were fixed in 4 % paraformaldehyde for 10 min at room temperature, and then permeabilized for 10 min with 0.5 % Triton-X-100. The cells were immunostained with anti-S9.6 antibody and anti-Fibrillarin antibody (Cell Signaling) for 1 h, washed three times with phosphate buffered saline (PBS), and incubated with Alexa Fluor 488-labeled anti-mouse IgG antibody and Alexa Fluor 568 labeled anti-rabbit IgG antibody (Invitrogen) for 30 min at room temperature. Finally, cells were mounted in mounting solution ProLong Gold with DAPI (Invitrogen).

FISH

Fluorescence in situ hybridization (FISH) was performed using our previously published approach [84]. For the experiment described, 2 μg of DNA from the Illumina RDIP-seq library were labeled by nick translation using spectrum orange-dUTP (Invitrogen, Carlsbad, CA). A locus-specific BAC clone (9p TelVysion probe #05J03-009) mapping to chromosome 9 was labeled in green using Spectrum Green (Vysis, Abbott Molecular, Des Plaines, IL). Both probes were hybridized to 46, *XY* control metaphases. The slides were denatured with 50 % formaldehyde/2× SSC at 80 °C for 1.5 min and then dehydrated with serial ethanol washing steps (ice cold 70, 90, and 100 % for 3 min each). The probes were denaturated in the hybridization solution (50 % dextran sulfate/2× SSC) at 85 °C for 5 min, applied to the slides, and incubated overnight at 37 °C in a humidified chamber. The slides were then washed 3 times for 5 min with 50 % formamide/2× SSC, 1× SSC and 4× SSC/0.1 % Tween. Slides were dehydrated with serial ethanol washing steps (see above) and mounted with ProLong Gold antifade reagent with DAPI (Invitrogen, Carlsbad, CA, USA) for imaging. Image acquisition is described in Additional file 1: Supplemental experimental procedures.

RNA:DNA hybrid immunoprecipitation (RDIP)

The cell culture conditions for IMR-90 and HEK 293T cells are described in the Additional file 1: Supplemental experimental procedures. Whole cell nucleic acid was isolated from HEK 293T cells and IMR-90 cells through a modified salting out extraction protocol [85]. Nucleic acid was sonicated to an average size of 400–600 bp using the Covaris sonicator. The fragmented nucleic acid was then treated with RNase I (Ambion AM2294) to remove any ssRNA from the sample, phenol/chloroform purified and re-suspended in EB buffer. Part of the nucleic acid sample was set aside as an untreated input sample for comparative sequencing. Three micrograms of nucleic acid sample was then incubated overnight with the S9.6 antibody, following which the RNA:DNA hybrids were enriched by immunomagnetic precipitation using Dynabeads (M-280 Sheep anti-mouse IgG). The sample was then extracted through phenol/chloroform purification, precipitated in the presence of glycogen and re-suspended in EB buffer. A complete detailed protocol is available in the Additional file 1: Supplemental experimental procedures. Enrichment of predicted peaks in the RDIP product was validated using quantitative PCR (Quanta PerfeCTa SYBR Green Fastmix). The primer sequences used are provided in Additional file 2: Table S3.

Directional RDIP-seq

Using RDIP and input material, directional RDIP-seq libraries were prepared using elements of a directional RNA-seq protocol modified from a previously published approach [49]. Starting the library preparation at the second strand synthesis step, the RNA of the RNA:DNA hybrid was nicked using RNase H treatment to serve as a primer for the DNA polymerase. The second strand was formed while incorporating dUTP to allow for directional sequencing and the identification of the RNA strand of the RNA:DNA hybrid. Next, the ends of fragments were repaired, adenosine tails added, and Illumina TruSeq strand-specific adaptors ligated (adaptor sequences in Additional file 2: Table S4). UNG treatment was utilized to degrade the dUTP-containing RNA strand of the

RNA:DNA hybrid, and barcoded PCR primers were used to amplify the library while maintaining directionality. The complete RDIP-seq protocol is available in the Additional file 1: Supplemental experimental procedures.

Prior to sequencing, the libraries were analyzed for quality of preparation using an Agilent Bioanalyzer high-sensitivity chip. Libraries were multiplexed and combined for sequencing using Illumina HiSeq 2500 150 bp paired-end sequencing in our institutional Epigenomics Shared Facility. FASTQ files were generated through the Illumina CASAVA pipeline (v1.8). Sequencing reads were then run through the Wasp System (WASP v3.1.5 rev. 6632) hosted pipeline for primary data processing, as follows. The reads were aligned to the hg19 reference genome using Bowtie (v0.12.7), using non-default parameters of –tryhard (increasing the number of attempts bowtie uses to find an alignment and number of backtracks), -I 50 (the minimum insert size in basepairs for valid paired-end alignments) and -X 650 (the maximum insert size for valid paired-end alignments). Alignments were generated in SAM format, which were then transformed into BAM files using Samtools (version 0.1.8). The aligned sequences in BAM format had PCR duplicates removed, and peaks were called based on input and IP files using MACS v1.4.2 [86]. RDIP-seq peaks for IMR-90 cells and two datasets for HEK 293T cells were then analyzed using the program CHANCE for quality of immunoprecipitation [87]. Based on the results of CHANCE, we discarded one of the HEK 293T datasets and continued on with one set of peaks for each cell line. All peaks containing "N" nucleotides were discarded. Custom code and parameters for this analysis can be found on our GitHub resource in the file "Peak Calling". Motif analysis of RNA:DNA hybrid peaks is described in the Additional file 1: Supplemental experimental procedures.

R-loop validation through non-denaturing bisulphite conversion

RDIP-seq peaks were validated through non-denaturing bisulphite conversion. Whole cell nucleic acid was isolated from HEK 293T cells through a modified salting out extraction protocol as outlined in the Additional file 1: Supplemental experimental procedures. Nucleic acid was digested with EcoRV-HF. Non-denaturing bisulphite treatment was performed according to a previously published protocol [50]. Regions of interest were amplified through PCR after denaturing or non-denaturing bisulphite treatment using primers to converted or unconverted DNA. The PCR product was purified, cloned using a TOPO-TA cloning kit (Life Technologies) and sequenced. The primer sequences used in non-denaturing bisulphite validation for this study are provided in Additional file 2: Table S5.

Directional RDIP-seq strandedness analysis

Due to using directional sequencing through the incorporation of dUTP, we were able to determine the RNA-derived sequence of the RNA:DNA hybrids. To do this, we used the BAM flag information describing our aligned sequences (http://broadinstitute.github.io/picard/). The second read in the pair, representing the sequence derived from the RNA strand following degradation using UNG of the dUTP-incorporated complementary strand, has the bit flag identifiers of 163 or 147, indicating that it maps to the top or bottom strand of the reference genome, respectively. By measuring the number of RNA reads aligned to the top or bottom reference strand for each peak, we could assign each RDIP-seq peak a "strandedness" value, with +1 being all RNA-derived reads aligned to the top strand and −1 all RNA-derived reads aligned to the bottom strand. We removed the small minority (10 %) of peaks with intermediate values of strandedness to decrease what we presumed to be experimental noise in our data set. Custom code for this analysis can be found on our GitHub resource in the file "Determining RNA Strand and Minus10 files".

RNA-seq of HEK 293T cells and IMR-90 cells

RNA was isolated from HEK 293T and IMR-90 cells using TRIzol extraction. Four biological replicates from each cell line were DNase treated, and Ribo-Zero rRNA removal (Ribo-Zero, Epicentre) was utilized for three of the four RNA samples, leaving a non-Ribo-Zero depleted sample for rRNA expression analysis. RNA-seq libraries were prepared using a directional RNA-seq protocol modified from a prior published approach [49] and detailed in the Additional file 1: Supplemental experimental procedures, Directional whole transcriptome sequencing protocol. Prior to sequencing, the libraries were assessed for quality using an Agilent Bioanalyzer high-sensitivity chip. The samples were multiplexed and sequenced using 100 bp single-end read sequencing on the Illumina HiSeq 2500 in our institutional Epigenomics Shared Facility. The TruSeq adaptor sequences used in this assay are provided in Additional file 2: Table S6.

After sequencing, FASTQ file generation was completed using the Illumina CASAVA pipeline (v1.8). Post-sequencing analysis was performed using the WASP pipeline (v3.1.5 rev. 6632), involving read alignment using gsnap (2012-07-20), with htseq (v0.5.3p3) used to determine read quantitation. Biological replicates were normalized using DESeq (Bioconductor) and RefSeq gene identifiers were assigned using biomaRt. Only gene expression assigned a RefSeq identifier was used for further analysis. Custom code for this analysis can be found on our GitHub resource under the file "RNAseq analysis".

Ribosomal DNA analysis

In order to align our RDIP-seq reads to the rDNA repeating unit, we used the alignment approach of Zentner and colleagues [55]. We added the rDNA repeating unit FASTA file (gi|555853|gb|U13369.1|HSU13369) to the start of the hg19 chromosome 13, replacing the telomeric "N" nucleotides. Duplicate reads were removed from the IMR-90 RDIP-seq and input FASTQ files using a custom perl script provided by Zentner and colleagues [55], and the remaining reads were aligned to the hg19+rDNA genome file using Bowtie. Wiggle tracks were then created using FSeq, and counts representative of the reads aligned to the rDNA portion of chromosome 13 were isolated. The RDIP-seq wiggle track values were normalized by subtracting the input values from the RDIP values. The same pipeline was used to align the IMR-90 RNA-seq samples that did not have prior Ribo-Zero depletion to the rDNA sequence. Processed histone mark datasets from K562 cells for rDNA were provided by Zentner and colleagues [55], and averaged across 50 bp windows across the rDNA repeating unit. Custom code for this analysis can be found on our GitHub resource under the file "Fig. 2—rDNA figure with Zentner histone marks" and the custom perl script under "Zentner removeDupsFromFASTQ Perl Script".

Regression models of RNA:DNA hybrid peak density

We used LASSO regularized linear regression to explore the relationship between the density of RNA:DNA hybrid peaks in 500 kb windows and genomic features associated with transcription and regulation. LASSO regression fits a linear model subject to a constraint on the sum of the regression coefficients [66]. The LARS algorithm, implemented in the *LARS* R package, was applied to determine the Lasso path. This algorithm provides the optimal values of the regression coefficients as the constraint on the sum of the coefficients is progressively relaxed [67]. Tight constraint on the sum of the coefficients enforces sparseness on the model with the number of covariates in the model increasing as this constraint is relaxed. The covariance test statistic [68], implemented in the *covTest* R package, was used to test the significance of each additional covariate when it enters the model.

Co-immunoprecipitation of RNA:DNA hybrid binding proteins (CoIP)

Native chromatin was isolated using a sucrose gradient from HEK 293T cells. Chromatin was incubated overnight with S9.6 antibody or a non-specific control antibody (β-actin, Sigma A5441), following which immunoprecipitation was performed on each sample using immunomagnetic precipitation (Dynabeads M-280 Sheep anti-mouse IgG). RNA:DNA hybrid-binding protein complexes were then eluted using RNA:DNA hybrid oligonucleotides, with DNA:DNA oligonucleotides as a control. The oligonucleotide sequences used in this assay are provided in Additional file 2: Table S7. The resulting enriched proteins were run on a 12 % polyacrylamide gel, stained with GelCode Blue (Life Technologies 24594) and tested using Mass Spectrometry (MS). Proteins which were considered to bind specifically to RNA:DNA hybrids were defined as those only present in the S9.6 immunoprecipitated sample and eluted with the RNA:DNA oligonucleotides, removing any proteins also present in the control samples (those isolated with the β-actin antibody, and with the S9.6 antibody eluted with the DNA:DNA oligonucleotides). This analysis was performed using Scaffold3 proteome software [88]. Peptide counts were assigned to each protein identified through mass spectrometry by measuring the quantity of the identified peptides by their spectra, and filtered by those peptides that also occurred in negative control experimental samples. Candidate proteins identified by mass spectrometry were then validated using Western blotting using the antibodies described in Additional file 2: Table S8.

Custom code

Analysis of RDIP-seq, RNA-seq, and code for all figures are included and annotated at: https://github.com/GreallyLab/Nadel-et-al.-2015.

Data access

The data generated are all available through the Gene Expression Omnibus, accession number GSE68953 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68953).

Additional files

Additional file 1. Supplemental Results and Methods.

Additional file 2. Supplemental Tables.

Abbreviations

BAC: Bacterial artificial chromosome; ChIP: Chromatin immunoprecipitation; CpG: Cytosine-guanine dinucleotide; DAPI: 4′,6-Diamidino-2-phenylindole; DNA: Deoxyribonucleic acid; DRIP: DNA:RNA immunoprecipitation; dsDNA: Double-stranded DNA; dUTP: Deoxyuridine triphosphate; EMSA: Electrophoretic mobility shift assay; FISH: Fluorescence in situ hybridization; H3K27me3: Histone H3 lysine 27 trimethylation; H3K9me3: Histone H3 lysine 9 trimethylation; H3S10: Histone H3 serine 10; IGS: Intergenic spacer of ribosomal DNA repeat unit; LARS: Least angle regression; LASSO: Least absolute shrinkage and selection operator; LINE: Long interspersed nuclear element; lncRNA: Long non-coding RNA; mRNA: Messenger RNA; NOR: Nucleolar organising region; PCR: Polymerase chain reaction; polyA: Polyadenylation; RDIP: RNA:DNA hybrid immunoprecipitation; rDNA: Ribosomal DNA; RNA: Ribonucleic acid; ssDNA: Single-stranded DNA; TERRA: Telomeric repeat containing RNA; TF: Transcription factor; UNG: Uracil *N*-glycosylase.

Authors' contributions

JMG and JJ designed the original project. JMG, RA and JN designed the original experiments and the analytical approaches. JN executed the experiments and analyzed results. JMG and JN wrote the manuscript. CL, NAW and ZZ contributed to analysis. PO and AG performed the motif analysis. HS performed the immunofluorescence studies, and the group of CM performed the FISH studies. CS performed the LASSO analysis. All authors read and approved the final manuscript.

Author details

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA. [2] Roche-NimbleGen, Madison, WI 53711, USA. [3] School of Mathematics, Statistics and Applied Mathematics, National University of Ireland Galway, Galway, Ireland. [4] Department of Genetics, Center for Epigenomics and Division of Computational Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA. [5] Present Address: Department of Biology, Center for Genomics and Systems Biology, New York University, 12 Waverly Place, New York, NY 10003, USA. [6] Present Address: Integrated Genomics Operation, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA.

Acknowledgements

The project was funded by NIH R21 GM101880 Grant to JMG. Einstein core facilities involved were the Genome Imaging Core, the High-Performance Computing Core, the Epigenomics Shared Facility, the Proteomics Core Facility, the Monoclonal Antibody Core Facility and the Genomics Core Facility, with support from the Albert Einstein Cancer Center (P30CA013330) and the Center for Epigenomics. JN was supported by the Training Program in Cellular and Molecular Biology and Genetics (NIH T32 GM007491) and NAW by the Medical Student Training Program (NIH T32 GM007288). We thank Drs. Gabe Zentner and Peter Scacheri at Case Western Reserve University for sharing code and data to allow us to perform the rDNA studies as comparably as possible with their prior work.

Competing interests

The authors declare that they have no competing interests.

Received: 22 July 2015 Accepted: 29 October 2015
Published online: 16 November 2015

References

1. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;22:1798–812.
2. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. 2012;489:83–90.
3. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res. 2012;22:1711–22.
4. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, Rozowsky J, Birney E, Bickel P, Snyder M, Gerstein M. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012;13:R48.
5. Hu S, Wan J, Su Y, Song Q, Zeng Y, Nguyen HN, Shin J, Cox E, Rho HS, Woodard C, et al. DNA methylation presents distinct binding sites for human transcription factors. eLife. 2013;2:e00726.
6. Medvedeva YA, Khamis AM, Kulakovskiy IV, Ba-Alawi W, Bhuyan MS, Kawaji H, Lassmann T, Harbers M, Forrest AR, Bajic VB. Effects of cytosine methylation on transcription factor binding sites. BMC Genom. 2014;15:119.
7. Schlick T. Topics in nucleic acids structure: noncanonical helices and rna structure. In: Molecular Modeling and simulation: an interdisciplinary guide. Interdisciplinary Applied Mathematics, vol. 21. New York: Springer; 2010. pp. 205–36.
8. Zhou T, Shen N, Yang L, Abe N, Horton J, Mann RS, Bussemaker HJ, Gordan R, Rohs R. Quantitative modeling of transcription factor binding specificities using DNA shape. Proc Natl Acad Sci USA. 2015;112:4654–9.
9. Raiber EA, Kranaster R, Lam E, Nikan M, Balasubramanian S. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. Nucleic Acids Res. 2012;40:1499–508.
10. Khrapunov S, Warren C, Cheng H, Berko ER, Greally JM, Brenowitz M. Unusual characteristics of the DNA binding domain of epigenetic regulatory protein MeCP2 determine its binding specificity. Biochemistry. 2014;53:3379–91.
11. Aguilera A, Garcia-Muse T. R loops: from transcription byproducts to threats to genome stability. Mol Cell. 2012;46:115–24.
12. Roberts RW, Crothers DM. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. Science. 1992;258:1463–6.
13. Roy D, Yu K, Lieber MR. Mechanism of R-loop formation at immunoglobulin class switch sequences. Mol Cell Biol. 2008;28:50–60.
14. Roy D, Lieber MR. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. Mol Cell Biol. 2009;29:3124–33.
15. Murat P, Balasubramanian S. Existence and consequences of G-quadruplex structures in DNA. Curr Opin Genet Dev. 2014;25:22–9.
16. Ginno PA, Lott PL, Christensen HC, Korf I, Chedin F. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell. 2012;45:814–25.
17. Dunn K, Griffith JD. The presence of RNA in a double helix inhibits its interaction with histone protein. Nucleic Acids Res. 1980;8:555–66.
18. Roy D, Zhang Z, Lu Z, Hsieh CL, Lieber MR. Competition between the RNA transcript and the nontemplate DNA strand during R-loop formation in vitro: a nick can serve as a strong R-loop initiation site. Mol Cell Biol. 2010;30:146–59.
19. Ginno PA, Lim YW, Lott PL, Korf I, Chedin F. GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res. 2013;23:1590–600.
20. Ratmeyer L, Vinayak R, Zhong YY, Zon G, Wilson WD. Sequence specific thermodynamic and structural properties for DNA.RNA duplexes. Biochemistry. 1994;33:5298–304.
21. Wanrooij PH, Uhler JP, Shi Y, Westerlund F, Falkenberg M, Gustafsson CM. A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary stability of the mitochondrial R-loop. Nucleic Acids Res. 2012;40:10334–44.
22. Loomis EW, Sanz LA, Chedin F, Hagerman PJ. Transcription-associated R-loop formation across the human FMR1 CGG-repeat region. PLoS Genet. 2014;10:e1004294.
23. Tracy RB, Lieber MR. Transcription-dependent R-loop formation at mammalian class switch sequences. EMBO J. 2000;19:1055–67.
24. Wahba L, Gore SK, Koshland D. The homologous recombination machinery modulates the formation of RNA-DNA hybrids and associated chromosome instability. eLife. 2013;2:e00505.
25. Wahba L, Amon JD, Koshland D, Vuica-Ross M. RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. Mol Cell. 2011;44:978–88.
26. Mischo HE, Gomez-Gonzalez B, Grzechnik P, Rondon AG, Wei W, Steinmetz L, Aguilera A, Proudfoot NJ. Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol Cell. 2011;41:21–32.
27. Tuduri S, Crabbe L, Conti C, Tourriere H, Holtgreve-Grez H, Jauch A, Pantesco V, De Vos J, Thomas A, Theillet C, et al. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat Cell Biol. 2009;11:1315–24.
28. Chan YA, Aristizabal MJ, Lu PY, Luo Z, Hamza A, Kobor MS, Stirling PC, Hieter P. Genome-wide profiling of yeast DNA:RNA hybrid prone sites with DRIP-chip. PLoS Genet. 2014;10:e1004288.
29. Reaban ME, Griffin JA. Induction of RNA-stabilized DNA conformers by transcription of an immunoglobulin switch region. Nature. 1990;348:342–4.
30. Daniels GA, Lieber MR. RNA:DNA complex formation upon transcription of immunoglobulin switch regions: implications for the mechanism and regulation of class switch recombination. Nucleic Acids Res. 1995;23:5006–11.
31. Sun Q, Csorba T, Skourti-Stathaki K, Proudfoot NJ, Dean C. R-loop stabilization represses antisense transcription at the *Arabidopsis* FLC locus. Science. 2013;340:619–21.
32. Pfeiffer V, Crittin J, Grolimund L, Lingner J. The THO complex component Thp2 counteracts telomeric R-loops and telomere shortening. EMBO J. 2013;32:2861–71.

33. Castellano-Pozo M, Santos-Pereira JM, Rondon AG, Barroso S, Andujar E, Perez-Alegre M, Garcia-Muse T, Aguilera A. R loops are linked to histone H3 S10 phosphorylation and chromatin condensation. Mol Cell. 2013;52:583–90.
34. Nakama M, Kawakami K, Kajitani T, Urano T, Murakami Y. DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation. Genes Cells. 2012;17:218–33.
35. Sollier J, Stork CT, Garcia-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA. Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. Mol Cell. 2014;56:777–85.
36. Hatchi E, Skourti-Stathaki K, Ventz S, Pinello L, Yen A, Kamieniarz-Gdula K, Dimitrov S, Pathania S, McKinney KM, Eaton ML, et al. BRCA1 recruitment to transcriptional pause sites is required for R-loop-driven DNA damage repair. Mol Cell. 2015;57:636–47.
37. Bhatia V, Barroso SI, Garcia-Rubio ML, Tumini E, Herrera-Moyano E, Aguilera A. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. Nature. 2014;511:362–5.
38. Chen YZ, Bennett CL, Huynh HM, Blair IP, Puls I, Irobi J, Dierick I, Abel A, Kennerson ML, Rabin BA, et al. DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). Am J Hum Genet. 2004;74:1128–35.
39. Gunther C, Kind B, Reijns MA, Berndt N, Martinez-Bueno M, Wolf C, Tungler V, Chara O, Lee YA, Hubner N, et al. Defective removal of ribonucleotides from DNA promotes systemic autoimmunity. J Clin Invest. 2015;125:413–24.
40. Lin Y, Dent SY, Wilson JH, Wells RD, Napierala M. R loops stimulate genetic instability of CTG.CAG repeats. Proc Natl Acad Sci USA. 2010;107:692–7.
41. Colak D, Zaninovic N, Cohen MS, Rosenwaks Z, Yang WY, Gerhardt J, Disney MD, Jaffrey SR. Promoter-bound trinucleotide repeat mRNA drives epigenetic silencing in fragile X syndrome. Science. 2014;343:1002–5.
42. Groh M, Lufino MM, Wade-Martins R, Gromak N. R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. PLoS Genet. 2014;10:e1004318.
43. Haeusler AR, Donnelly CJ, Periz G, Simko EA, Shaw PG, Kim MS, Maragakis NJ, Troncoso JC, Pandey A, Sattler R, et al. C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature. 2014;507:195–200.
44. Groh M, Gromak N. Out of balance: R-loops in human disease. PLoS Genet. 2014;10:e1004630.
45. Skourti-Stathaki K, Proudfoot NJ. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. Genes Dev. 2014;28:1384–96.
46. Boguslawski SJ, Smith DE, Michalak MA, Mickelson KE, Yehle CO, Patterson WL, Carrico RJ. Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids. J Immunol Methods. 1986;89:123–30.
47. Chadwick LH. The NIH roadmap epigenomics program data resource. Epigenomics. 2012;4:317–24.
48. Zhang ZZ, Pannunzio NR, Hsieh CL, Yu K, Lieber MR. Complexities due to single-stranded RNA during antibody detection of genomic rna:dna hybrids. BMC Res Notes. 2015;8:127.
49. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009;37:e123.
50. Yu K, Chedin F, Hsieh CL, Wilson TE, Lieber MR. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat Immunol. 2003;4:442–51.
51. Yeo AJ, Becherel OJ, Luff JE, Cullen JK, Wongsurawat T, Jenjaroenpoon P, Kuznetsov VA, McKinnon PJ, Lavin MF. R-loops in proliferating cells but not in the brain: implications for AOA2 and other autosomal recessive ataxias. PLoS One. 2014;9:e90219.
52. Koo CX, Kobiyama K, Shen YJ, LeBert N, Ahmad S, Khatoo M, Aoshi T, Gasser S, Ishii KJ. RNA Polymerase III regulates cytosolic RNA:DNA hybrids and intracellular MicroRNA expression. J Biol Chem. 2015;290:7463–73.
53. Brown TA, Tkachuk AN, Clayton DA. Native R-loops persist throughout the mouse mitochondrial DNA genome. J Biol Chem. 2008;283:36743–51.
54. Gonzalez IL, Sylvester JE. Complete sequence of the 43-kb human ribosomal DNA repeat: analysis of the intergenic spacer. Genomics. 1995;27:320–8.
55. Zentner GE, Saiakhova A, Manaenkov P, Adams MD, Scacheri PC. Integrative genomic analysis of human ribosomal DNA. Nucleic Acids Res. 2011;39:4949–60.
56. Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol Cell. 2011;42:794–805.
57. Belotserkovskii BP, Liu R, Tornaletti S, Krasilnikova MM, Mirkin SM, Hanawalt PC. Mechanisms and implications of transcription blockage by guanine-rich DNA sequences. Proc Natl Acad Sci USA. 2010;107:12816–21.
58. Nellaker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 2012;13:R45.
59. Medstrand P, van de Lagemaat LN, Mager DL. Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. 2002;12:1483–95.
60. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science. 2008;322:1845–8.
61. Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, et al. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010;465:182–7.
62. De Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 2010;8:e1000384.
63. Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat Genet. 2014;46:1311–20.
64. Mao X, Miesfeldt S, Yang H, Leiden JM, Thompson CB. The FLI-1 and chimeric EWS-FLI-1 oncoproteins display similar DNA binding specificities. J Biol Chem. 1994;269:18216–22.
65. Fazzari MJ, Greally JM. Epigenomics: beyond CpG islands. Nat Rev Genet. 2004;5:446–55.
66. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc. 1996;58:267–88.
67. Efron B, Hastie T. Johnstone I. Tibshirani R: Least angle regression; 2004. pp. 407–99.
68. Lockhart R, Taylor J. Tibshirani RJ. Tibshirani R: A significance test for the lasso; 2014. p. 413–68.
69. Chakraborty P, Grosse F. Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes. DNA Repair (Amst). 2011;10:654–65.
70. Toretsky JA, Erkizan V, Levenson A, Abaan OD, Parvin JD, Cripe TP, Rice AM, Lee SB, Uren A. Oncoprotein EWS-FLI1 activity is enhanced by RNA helicase A. Cancer Res. 2006;66:5574–81.
71. Stein U, Jurchott K, Walther W, Bergmann S, Schlag PM, Royer HD. Hyperthermia-induced nuclear translocation of transcription factor YB-1 leads to enhanced expression of multidrug resistance-related ABC transporters. J Biol Chem. 2001;276:28562–9.
72. Aoki Y, Zhao G, Qiu D, Shi L, Kao PN. CsA-sensitive purine-box transcriptional regulator in bronchial epithelial cells contains NF45, NF90, and Ku. Am J Physiol. 1998;275:L1164–72.
73. Kao PN, Chen L, Brock G, Ng J, Kenny J, Smith AJ, Corthesy B. Cloning and expression of cyclosporin A- and FK506-sensitive nuclear factor of activated T-cells: NF45 and NF90. J Biol Chem. 1994;269:20691–9.
74. Langland JO, Kao PN, Jacobs BL. Nuclear factor-90 of activated T-cells: a double-stranded RNA-binding protein and substrate for the double-stranded RNA-dependent protein kinase, PKR. Biochemistry. 1999;38:6361–8.
75. Boeva V, Surdez D, Guillon N, Tirode F, Fejes AP, Delattre O, Barillot E. De novo motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. Nucleic Acids Res. 2010;38:e126.
76. Pimkin M, Kossenkov AV, Mishra T, Morrissey CS, Wu W, Keller CA, Blobel GA, Lee D, Beer MA, Hardison RC, Weiss MJ. Divergent functions of hematopoietic transcription factors in lineage priming and differentiation during erythro-megakaryopoiesis. Genome Res. 2014;24:1932–44.
77. Li Y, Luo H, Liu T, Zacksenhaus E, Ben-David Y. The ets transcription factor Fli-1 in development, cancer and disease. Oncogene. 2015;34:2022–31.

78. Erkizan HV, Schneider JA, Sajwan K, Graham GT, Griffin B, Chasovskikh S, Youbi SE, Kallarakal A, Chruszcz M, Padmanabhan R, et al. RNA helicase A activity is inhibited by oncogenic transcription factor EWS-FLI1. Nucleic Acids Res. 2015;43:1069–80.
79. Erkizan HV, Kong Y, Merchant M, Schlottmann S, Barber-Rotenberg JS, Yuan L, Abaan OD, Chou TH, Dakshanamurthy S, Brown ML, et al. A small molecule blocking oncogenic protein EWS-FLI1 interaction with RNA helicase A inhibits growth of Ewing's sarcoma. Nat Med. 2009;15:750–6.
80. Riggi N, Knoechel B, Gillespie SM, Rheinbay E, Boulay G, Suva ML, Rossetti NE, Boonseng WE, Oksuz O, Cook EB, et al. EWS-FLI1 utilizes divergent chromatin remodeling mechanisms to directly activate or repress enhancer elements in Ewing sarcoma. Cancer Cell. 2014;26:668–81.
81. Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, Siksnys V, Seidel R. Direct observation of R-loop formation by single RNA-guided Cas9 and cascade effector complexes. Proc Natl Acad Sci USA. 2014;111:9798–803.
82. Zheng R, Shen Z, Tripathi V, Xuan Z, Freier SM, Bennett CF, Prasanth SG, Prasanth KV. Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. J Cell Sci. 2010;123:3734–44.
83. Bacolla A, Cooper DN, Vasquez KM. Non-B DNA structure and mutations causing human genetic disease. In: eLS. New York: John Wiley & Sons Ltd; 2001.
84. Montagna C, Andrechek ER, Padilla-Nash H, Muller WJ, Ried T. Centrosome abnormalities, recurring deletions of chromosome 4, and genomic amplification of HER2/neu define mouse mammary gland adenocarcinomas induced by mutant HER2/neu. Oncogene. 2002;21:890–8.
85. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. 1988;16:1215.
86. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.
87. Diaz A, Nellore A, Song JS. CHANCE: comprehensive software for quality control and validation of ChIP-seq data. Genome Biol. 2012;13:R98.
88. Searle BC. Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. Proteomics. 2010;10:1265–9.