TITLE:

RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases

AUTHORS:

László Halász[1,2#], Zsolt Karányi[1,3#], Beáta Boros-Oláh[1,2], Tímea Kuik-Rózsa[1,2], Éva Sipos[1,4], Éva Nagy[1], Ágnes Mosolygó-L[1,2], Anett Mázló[5], Éva Rajnavölgyi[5], Gábor Halmos[4], Lóránt Székvölgyi[1,2 *]

AFFILIATIONS:

[1] MTA-DE Momentum, Genome Architecture and Recombination Research Group, Research Centre for Molecular Medicine, University of Debrecen, Hungary

[2] Department of Biochemistry and Molecular Biology, University of Debrecen, Hungary

[3] Department of Internal Medicine, University of Debrecen, Hungary

[4] Department of Biopharmacy, University of Debrecen, Hungary

[5] Department of Immunology, University of Debrecen, Hungary

[*] Corresponding author:

Lóránt Székvölgyi, Ph.D.

Faculty of Medicine, University of Debrecen, Hungary

Address: 4032 Debrecen Hungary, Nagyerdei krt. 98.

Tel.: +36 52 411 717 / 50225

Fax: +36 52 255 990

URL: http://geneart.med.unideb.hu/

E-mail: lorantsz@med.unideb.hu

[#] Equal contribution: joint first authors

ABSTRACT

The impact of R-loops on the physiology and pathology of chromosomes has been demonstrated extensively by chromatin biology research. The progress in this field has been driven by technological advancement of R-loop mapping methods that largely relied on a single approach, DNA-RNA immunoprecipitation (DRIP). Most of the DRIP protocols use the experimental design that was developed by a few laboratories, without paying attention to the potential caveats that might affect the outcome of RNA-DNA hybrid mapping. To assess the accuracy and utility of this technology, we pursued an analytical approach to estimate inherent biases and errors in the DRIP protocol. By performing DRIP-sequencing, qPCR and receiver operator characteristic (ROC) analysis, we tested the effect of formaldehyde fixation, cell lysis temperature, mode of genome fragmentation, and removal of free RNA on the efficacy of RNA-DNA hybrid detection, and implemented workflows that were able to distinguish complex and weak DRIP signals in a noisy background with high confidence. We also show that some of the workflows perform poorly and generate random answers. Furthermore, we found that the most commonly used genome fragmentation method (restriction enzyme digestion) led to the overrepresentation of lengthy DRIP fragments over coding ORFs, and this bias was enhanced at the first exons. Biased genome sampling severely compromised mapping resolution and prevented the assignment of precise biological function to a significant fraction of R-loops. The revised workflow presented herein is established and optimized using objective ROC analyses and provides reproducible and highly specific RNA-DNA hybrid detection.

KEYWORDS

RNA-DNA hybrid, R-loop, DRIP-seq, chromatin immunoprecipitation, ROC analysis

INTRODUCTION

R-loops are three stranded nucleic acid structures that are composed of an RNA-DNA hybrid and a displaced single-stranded DNA. Under physiological conditions R-loops are prevalent along the chromosomes, constituting 5-8% of the genome and impacting many cellular processes (Sanz et al. 2016; Wahba et al. 2016). For instance, R-loops *i.* drive embryonic stem cell differentiation via modulating the chromosomal binding of chromatin-regulatory complexes (Chen et al. 2015), *ii.* ensure the optimal binding of transcriptional activators to the promoter of the human vimentin (*VIM*) gene (Boque-Sastre et al. 2015), *iii.* massively form on estrogen-responsive genes in human breast and other tissues upon estrogen-hormone stimulation (Stork et al. 2016a), *iv.* induce heterochromatin formation in *Schizosaccharomyces pombe* (Nakama et al. 2012), *v.* inhibit the expression of an antisense non-coding RNA in *Arabidopsis thaliana*, associated to the flowering process (Sun et al. 2013). In a pathological context, perturbation or mutation of any of the following factors causes the chromosomal accumulation of RNA–DNA hybrids and consequent genomic instability: *i.* mRNA splicing factors and RNA export factors (e.g. *THO2*, *HPR1*, *MFT1*, *THP2*, *THOC1-7*, *SRSF1*) (Huertas and Aguilera 2003; Domínguez-Sánchez et al. 2011; Gan et al. 2011; Li and Manley 2005), *ii.* RNA-DNA hybrid helicases (e.g. *SETX* / *SEN1*, *AQR*, *PIF1*) (Alzu et al. 2012; Mischo et al. 2011; Sollier et al. 2014; Boulé and Zakian 2007), *iii.* RNA-DNA ribonucleases (*RNASEH1 / RNH1*, *RNASEH2A-C* / *RNH201*) (El Hage et al. 2010; Chon et al. 2013; Stuckey et al. 2015), *iv.* homologous recombination proteins (e.g. *BRCA1*, *BRCA2*, *RTEL1, SRS2*), (Hatchi et al. 2015; Bhatia et al. 2014), *v.* Fanconi anaemia proteins (*FANCA*, *FANCB*, *FANCC*) (García-Rubio et al. 2015; Schwab et al. 2015), *vi.* topoisomerases (*TOP1*, *TOP3B*) (Wilson-Sali and Hsieh 2002; El Hage et al. 2010; Yang et al. 2014; Marinello et al. 2016)

The above examples clearly illustrate the massive progress in the field that has been driven by technological advancements of R-loop detection methods. These techniques involve, for instance, electrophoretic mobility shift assays (Yu et al. 2006), atomic force microscopy (Brown et al. 2008), transmission electron microscopy (Pohjoismäki et al. 2010), fluorescent microscopy (Székvölgyi et al. 2007), fluorescence in situ hybridization (Nadel et al. 2015), native bisulfite modification (Yu et al. 2003), immunoprecipitation (Skourti-Stathaki et al. 2011; Ginno et al. 2012b) and computational prediction (Jenjaroenpun et al. 2015). The increasing number of R-loop mapping data relied on a single approach, DNA-RNA immunoprecipitation (DRIP) and its variations (RDIP, DRIPc, S1-DRIP, DRIP-RNA, DIP, ChIP). The DRIP method applies the S9.6 anti-RNA-DNA hybrid antibody (Hu et al. 2006) to capture RNA-DNA hybrids in their native chromosomal context, followed by mapping the enriched DNA fragments on a selected number of loci or across the whole genome, using quantitative PCR, microarray hybridization or deep sequencing.

Having surveyed the published RNA-DNA hybrid mapping studies (Chan et al. 2014; Ohle et al. 2016; Alzu et al. 2012; Cloutier et al. 2016; El Hage et al. 2010; Salvi et al. 2014; Sanz et al. 2016; Bhatia et al. 2014; Boque-Sastre et al. 2015; García-Rubio et al. 2015; Groh et al. 2014; Hatchi et al. 2015; Herrera-Moyano et al. 2014; Jenjaroenpun et al. 2015; Loomis et al. 2014; Pefanis et al. 2015; Rigby et al. 2014; Romanello et al. 2016; Skourti-Stathaki et al. 2011; Sun et al. 2013; Yang et al. 2014; Zhang et al. 2014a, 2014b, 2015; Castellano-Pozo et al. 2013; Wahba and Koshland 2013; Yang et al. 2016; Mischo et al. 2011; Chen et al. 2015; Ginno et al. 2012b; Lim et al. 2015; Stork et al. 2016b; Marinello et al. 2016; Nadel et al. 2015; Wahba et al. 2016; El Hage et al. 2014; Zeller et al. 2016) (**Supplemental Table S1**) we found that most DRIP protocols used the experimental design that was developed by a few laboratories (**Supplemental Material, Figure S1**). The original protocols are still being used without paying attention to their potential caveats: several critical points have remained exceedingly heterogeneous among the DRIP studies (**Supplemental Table S1**) that might account for at least some of the contradictory results (El Hage et al. 2014; Chan et al. 2014; Wahba et al. 2016; Nadel et al. 2015; Ginno et al. 2012b). One can reveal technical heterogeneities *i.* in terms of the studied model organisms and cell types, *ii.* whether the cells were fixed by formaldehyde (HCHO) or not, *iii.* whether the immunoprecipitation was chromatin-based or DNA-based (ChIP *vs*. DIP), *iv.* in the cell lysis temperature (65 °C, 55 °C, 37 °C), *v.* in the mode of DNA fragmentation (restriction enzyme digestion *vs.* sonication), *vi.* in total nucleic acid extraction (solid-phase purification *vs.* organic extraction, or salting out extraction), *vii.* in the application of ribonuclease A digestion to eliminate free RNA from the nucleic acid prep. Obviously, each of these variables can introduce substantial bias that might obscure the overall outcome of the experiment, but their consequence, alone or in combination, has remained unexplored.

In the current study, we aimed to assess possible confounding effects related to key experimental variables of the DRIP procedure. Combining DRIP-qPCR, DRIP-sequencing and receiver operator characteristic (ROC) calculation, we devised an unbiased and systematic analytical pipeline in human T lymphoblastoid cells regarding the most important DRIP variables so that we suggest a reproducible and specific RNA-DNA hybrid detection, underlain by the objective criteria of ROC analysis.

RESULTS

Introducing DRIP classifiers to assess true and false R-loop associations

Based on the available workflows of published DRIP protocols and considering the main technical variables that might contribute to the observed heterogeneities, we designed forty DRIP experimental schemes (binary classifiers) so that we assess how they rank different test loci according to their known RNA-DNA hybrid status (**Fig. 1**). The classifiers ("DRIP experiments" or "dependent variables") were designed to systematically explore the main factors that might create experimental bias associated with the DRIP procedure.

Experiments 1-16 considers the effect of *i.* formaldehyde (HCHO) fixation, *ii.* the method of nucleic acid isolation, *iii.* removal of free RNA, *iv.* the mode of nucleic acid fragmentation (**Fig. 1A**), and *v.* cell lysis temperature (65 °C as default vs. 37 °C; not shown in **Fig. 1.A**, but referred throughout the text as "37 °C").

Step 1: Formaldehyde fixation

The basic assumption behind HCHO-crosslinking is to maximize the DRIP yield while preserving biologically meaningful RNA-DNA hybrid interactions. But formaldehyde has some well-known adverse effects: *i.* the DNA accompanies a conformational change upon crosslinking, involving local denaturation or “breathing” of the double helix (McGhee and von Hippel 1977). This might create ectopic R-loop sites or abolish physiological R-loop contacts. *ii.* HCHO-treatment can reduce antigen accessibility or mask epitopes recognized by the antibody used for the immunoprecipitation. This might prevent a fraction of R-loops from being detected. *iii*. HCHO-fixation elicits spurious localization of irrelevant proteins at highly expressed genes (Baranello et al. 2016), and induces massive poly(ADP)ribose polymer formation in live cells (Beneke et al. 2012). These examples warrant deeper investigation of the usage of HCHO-fixation in RNA-DNA hybrid mapping, therefore we classified our DRIP samples as HCHO-treated and non-treated categories (**Fig. 1A-B**).

Step 2: Nucleic acid purification

Two common methods were compared: organic (phenol/chloroform) extraction versus solid-phase (silica membrane) purification of total nucleic acids (**Fig. 1A-B**).

Step 3: Ribonucleolytic treatment (RNase A, RNase H and sodium hydroxide)

Most DRIP protocols do not treat the isolated nucleic acid with ribonucleases to remove free RNA, however the S9.6 antibody can recognize RNA duplexes with a ~5-fold reduced affinity compared to RNA-DNA hybrids (Phillips et al. 2013). At this point, four kinds of ribonucleoleolytic digestion were incorporated into our DRIP pipelines: *i.* RNase H1 digestion that removes RNA-DNA hybrids (negative control #1), *ii.* alkaline hydrolysis by sodium hydroxide that degrades free RNA and RNA-DNA hybrids (negative control #2), *iii.* RNase A digestion at high (300 mM) NaCl concentration that removes free RNA, *iv.* RNase A digestion at low (25 mM) NaCl concentration that removes free RNA and RNA-DNA hybrids.

RNase H1 treatment is an accepted negative control of the DRIP procedure since it degrades the RNA strand in the hybrids preventing their recognition by the S9.6 antibody. Half of the nucleic acid prep was digested by RNase H1 before the DNA fragmentation step that made us estimate the bulk level of RNA-DNA hybrids (dot blot setting; **Supplemental Material, Figure S2A**). The other half was digested just before the S9.6 immunoprecipitation step that let us obtain crucial information about the specificity of the IP signal (see DRIP-qPCR). As expected, RNA-DNA hybrids were sensitive to RNase H1 digestion *in vitro*. Similarly to RNase H1, alkaline hydrolysis by 50 mM NaOH also efficiently eliminated the RNA-DNA hybrid signal (**Supplemental Material, Figure S2A**). Less is known about the salt-dependent RNase H-like activity of RNase A that is supposed to digest RNA-DNA hybrids as an efficient hybridase at low ionic strength (https://www.thermofisher.com/order/catalog/product/EN0531). As shown in **Supplemental Figure S2B** (**Supplemental Material**), the hybrids were indeed resistant to RNase A digestion at high ionic strength, but they became highly sensitive to RNase A as a function of decreasing monovalent concentration. The RNase H-like activity of RNase A at low salt condition was confirmed by an independent method (**Supplemental Material, Figure S2C-D**) applying fluorescent microscopic detection. Based on these experiences, RNase A digestion at high salt concentration (300 mM NaCl) was integrated into our DRIP protocol to test if removal of competing free RNA improves the specificity of the RNA-DNA hybrid signal. Also, RNase H1 digestion of the fragmented nucleic acid was kept as an obligatory negative control of the immunoprecipitation.

<u>Step 4: Nucleic acid fragmentation</u>

The choice of restriction enzymes defines the cleavage pattern of DNA that is critical to achieve optimal fragment length distribution and mapping resolution. Based on the original DRIP protocol (Ginno et al. 2012b), we combined five enzymes (Hind**III**, EcoR**I**, BsrG**I**, Xba**I** and Ssp**I**) for *in silico* digestion, resulting in a median restriction fragment length of 314 bp (**Supplemental Material, Figure S3A**). In contrast to the theoretical fragment size distribution, we observed a broad DNA size range in a real digestion reaction (between 100-10.000 bp; **Supplemental Material, Figure S4A**). As a control, we repeated the restriction enzyme cleavage in varying reaction conditions, without detecting any improvement in the digestion efficacy (**Supplemental Material, Figure S3B**). When a budding yeast genomic DNA was digested in a parallel experiment, we managed to obtain the expected (*in silico*) fragment size distribution (**Supplemental Material, Figure S3C**). These observations necessitate for the proper control of DNA fragment length distribution in DRIP samples that derive from restriction enzyme fragmented nucleic acid.

As opposed to restriction enzyme digestion, sonication creates random DNA fragments with a typical size of 150-500 bp that dictate the spatial resolution of the DRIP assay (**Supplemental Material, Figure S4B**). However, excessive sonication can introduce strand breaks in the DNA or simply shake off a subset of R-loops from the chromosomes, potentially compromising their detectability by qPCR. Because of the above, the mode of DNA fragmentation (restriction enzymes and sonication) was introduced as an important parameter in our DRIP pipeline (**Fig. 1A**).

Fragmenting chromatin rather than purified genomic DNA (experiments 17-24)

In comparison to the original DRIP protocol, classical chromatin immunoprecipitation (ChIP) involves the capture of RNA-DNA hybrids by immunoprecipitation from cross-linked and sonicated chromatin (rather than naked DNA) followed by phenol/chloroform purification (**Fig. 1C**). Since sonication, performed on purified genomic DNA, led to loss of ~80% of the DRIP signal in yeast (Wahba et al. 2016), we tested if acoustic shearing performed on a chromatin prep rather than on naked nucleic acid (**Supplemental Material, Figure S4C**) could improve the signal to noise ratio of the DRIP measurement (**Fig. 1B**).

Varying the cell lysis temperature

Published DRIP protocols apply various cell lysis temperatures, ranging from 37 °C to 65 °C and lasting from a couple of hours to overnight. To test the effect of temperature on the specificity of RNA-DNA hybrid detection, we lysed the samples at 65 °C for 7 hrs, or at 37 °C overnight. Experiments 1-16 were processed in parallel at both temperatures, while exp. 17-24 were omitted from the temperature analysis since crosslink reversal typically occurs at 65 °C.

Taken together, the above experimental variables resulted in forty (16x2+8) autonomous DRIP classifiers (schemes) for which RNA-DNA hybrid enrichment scores were determined at several test loci. This allowed us to assess whether the S9.6 signal represented true or false R-loop associations within the applied condition.

Making a reference R-loop set for benchmarking the DRIP classifiers

To derive the parameters of the DRIP classifiers, known positive and negative examples (genomic sites) could be chosen from the scientific literature based on their known R-loop profiles; however, the heterogeneity of the available DRIP-qPCR and DRIP-seq datasets (see Introduction) prompted us to establish our independent R-loop training set. We performed DNA-RNA hybrid mapping (DRIP-seq) in two closely related human cell types (Jurkat T cell leukemia cell line and naive CD4$^+$ T lymphocytes) and identified 88.830 and 99.337 R-loop enriched regions, respectively (**Fig. 2A**). A high-confidence R-loop peak set was generated from the identified binding sites and their chromosomal distribution was characterized. The peaks were significantly enriched at gene promoters and repetitive elements (**Fig. 2B**), consistent with previously published DRIP-seq results (Ginno et al. 2012a; Nadel et al. 2015). R-loop sites were underrepresented at protein coding exons, similarly to earlier DRIP experiments performed with sonicated nucleic acid, however restriction enzyme fragmented DRIP samples were positively biased towards exons. Sonicated and restriction enzyme digested samples were strikingly different in their R-loop length distributions (narrow: 179-2.369 bp *vs.* wide: 178-22.479 bp; **Fig. 2C**), and the identified R-loop binding sites significantly overlapped within each group, but sharply stood apart between the two groups (**Fig. 2D**). We attribute these differences to the extensive variation of R-loop lengths and heterogeneities of the studied cell types. Biological implications of having too wide peak sizes will be discussed later. With the observed variances in mind, our consensus R-loop set was regarded as an amenable reference to benchmark the DRIP classifiers.

Measuring RNA-DNA hybrid enrichment over the DRIP classifiers

Positive and negative test regions were selected from the identified R-loop set (**Supplemental Material, Figure S5**) and were systematically probed for RNA-DNA hybrid enrichment across the DRIP classifiers (**Supplemental Material, Figure S6**). Five test regions were frequently used as positive and negative controls in various published DRIP studies (*SNRPN, ZNF554, MYADM, FMR1, APOE*; (Yang et al. 2014; Marinello et al. 2016; Groh et al. 2014; García-Rubio et al. 2015; Loomis et al. 2014; Bhatia et al. 2014; Herrera-Moyano et al. 2014; Ginno et al. 2012b; Boque-Sastre et al. 2015), while the remaining sites were picked at random from the consensus R-loop set (*PRR5L, LOC440704, NOP58, VIM, ING3*). The reference DRIP-seq signal (benchmarking the classifiers) is shown over selected test regions along with DRIP-seq patterns taken from published studies (**Supplemental Material, Figure S5**). DRIP-qPCR yields were measured in control and RNase H-treated samples for forty (16x2+8) DRIP classifiers, at ten test regions, in five independent experiments. The resulting 4000 (40x2x10x5) DRIP enrichment scores were then readily used as an input parameter of receiver operator characteristics (ROC) calculation.

Determining the sensitivity and specificity of RNA-DNA hybrid detection: ROC analysis

We quantitated the relative trade-offs between true positive hits and experimental errors (false R-loop associations) by performing ROC analysis (Robin et al. 2011) on the DRIP-qPCR screen characterizing the classifiers (**Supplemental Material, Figures S6-S10**). The sensitivity, specificity and the area under the curve (AUC) values were extracted from the ROC plots (**Supplemental Table S2**) and used as an objective measure of the robustness of the forty experiments. High (>0.7) AUC values were obtained for ten DRIP classifiers (exp. 5, 6, 13, 15, 17, 18, 19, 21, and 24), implying that those experiments could predict the presence or absence of an RNA-DNA hybrid with high efficacy (**Fig. 3A**). AUC values close to 0.5 were obtained in four experiments (exp. 2, 10, 11, and 16), implying that the classifiers gave random answers without any predictive power as to the presence of an R-loop. Based on these considerations, the top four DRIP classifiers were: exp. 5, 13, 17, and 19 (**Fig. 3B-C**) with a sensitivity of 68.5-75 % and specificity of 68-79 %. Similar (or even higher) ROC parameters were obtained in a repeated experiment using a B lymphoblastoid cell line

(**Supplemental Material, Figure S8**), demonstrating the reliability of the tested DRIP protocols in other cell types.

Pairwise comparison of the main experimental variables (**Fig. 3D**) revealed no significant difference between *i.* formaldehyde-fixed *vs.* unfixed samples, *ii.* phenol-chloroform extracted *vs*. silica membrane purified nucleic acid samples, and *iii.* DNA-fragmented (exp. 1-16) *vs*. chromatin-fragmented DRIP samples (exp. 17-24). Cell lysis temperature (65°C *vs.* 37°C) did not change the specificity and sensitivity of the DRIP assay (**Supplemental Material, Figures S9-S10**). Statistically significant difference was obtained for RNase A-treated *vs.* untreated samples (p=0.03), suggesting that addition of RNase A does not improve the efficacy of RNA-DNA hybrid detection (Step 3, **Fig. 3D**). We explain the adverse effect of RNase A by its reported DNA binding activity (Benore-Parsons and Ayoub 1997; Dona and Houseley 2014) that selectively eliminates a vast amount (micrograms) of melted DNA regions upon nucleic acid purification (Dona and Houseley 2014). We confirmed the strong DNA binding of RNase A as migration defects on DNA gels, when a plasmid DNA was incubated with the enzyme (**Supplemental Material, Figures S11**). The observed electrophoretic mobility shift was prevalent on supercoiled, nicked-circular and linearized DNA templates.

Finally, by comparing sonicated and restriction enzyme fragmented DRIP samples (Step 4, **Fig. 3D**) we found a statistically significant difference (p=0.0002) in the ROC parameters, suggesting that sonication in more efficient in discriminating true positive signals from false positives, at least within the tested conditions.

Good DRIP practice: impact on the annotation and basic biological function of R-loops

Suboptimal DRIP conditions might prevent the assignment of precise biological function to a significant fraction of R-loops. Although the average DNA fragment size resulting from restriction enzyme digestion fits the requirements of the DRIP assay, we found that the frequency of cutting sites was significantly higher within intergenic regions, producing lengthy restriction fragments over protein coding ORFs (**Fig. 4**). Biased genome sampling, related to the non-random distribution of restriction enzyme recognition sequences, was even more pronounced over exons (**Fig. 4C**), especially over the first exons (**Fig. 4D**). In 82% of first exons there were only 0-1 suitable restriction sites compared to intergenic regions (59%). We estimated the digestion efficiency of restriction enzyme cutting sites to ~50 % over intergenic regions (based on the proportion of zero reads over restriction enzyme cutting sequences, representing cleaved sites), which was significantly reduced over gene coding regions (**Fig. 4E-F**). Consequently, genic regions void of suitable restriction sites appear as long DRIP fragments that potentially compromise mapping resolution. The *MYC*, *BCL6*, and *VIM* genes are shown as representative examples for large, restriction fragment-sized DRIP peaks (**Fig. 5**). Precise genomic position of R-loops could be resolved by sonication.

DISCUSSION

The increasing recognition of RNA-DNA hybrid structures in the physiology and pathology of chromosomes has prompted us to develop an analytical approach to estimate the inherent biases and errors of existing DRIP protocols and to assess the power of the technology. The determined ROC parameters (AUC, sensitivity, specificity, threshold) served as an objective measure for the efficacy of predicting the presence or absence of RNA-DNA hybrids. In the tested experimental conditions, we managed to find and verify DRIP workflows that were able to distinguish complex or weak DRIP-qPCR signals from a noisy background with high confidence across a number of genomic regions (exp. 5, 13, 17, and 19). On the contrary, some DRIP workflows performed unreliably and generated random answers (exp. 2, 10, 11, and 16). Under our experimental conditions, we highlight these groups as “preferred” and “not preferred”. By testing the main parameters of the DRIP experimental scheme - involving formaldehyde fixation, cell lysis temperature, nucleic acid isolation, free RNA removal, and DNA fragmentation – we found that fragmenting the nucleic acid by sonication and omitting RNase A digestion could improve the precision and specificity of RNA-DNA hybrid detection (**Fig. 3D**). At this point we emphasize the lack of correlation between the DRIP scores (IP/input ratios) and AUC values, as these quantities are not related to each other. The former highlights the yield of immunoprecipitation, while the latter is a quantitative measure of true and false R-loop associations. For instance, the worse and best DRIP schemes (exp 2 and exp 5) had a qPCR yield of 10%-95% and 1%-18% over the studied regions, respectively (**Supplemental Material, Figures S6 and S8**). Consequently, high DRIP enrichment does not necessarily accompany with increased accuracy, and *vice versa*.

We also showed that genome fragmentation by restriction enzymes led to the overrepresentation of long DRIP fragments over ORFs, which was especially enhanced over the first exons of protein coding genes (**Figs. 4-5**). Biased genome sampling severely compromised mapping resolution and, as a consequence, the assignment of clear biological function to a fraction of R-loops. For instance, correct estimation of evolutionary conservation between R-loop binding sites, relying on sequence homologies of exons that are associated with R-loops (Sanz et al. 2016), becomes uncertain.

Based on the above experiences, we suggest the following refinements of DRIP workflows to obtain accurate estimates of RNA-DNA hybrid occupancies: 1. Omission of HCHO-fixation and RNase A treatment, isolation of nucleic acid by silica membrane (kit) purification, nucleic acid fragmentation by sonication, followed by immunoprecipitation with the S9.6 antibody (see Methods). 2. If formaldehyde-fixation is applied, we recommend preparing soluble chromatin and fragmenting the prep by sonication (similarly to the ChIP protocol), followed by organic extraction and immunoprecipitation with the S9.6 antibody. 3. If restriction enzyme fragmentation needs to be applied (e.g. in some cases sonication might be too harsh to capture transient or very week RNA-DNA hybrid interactions), we advise the careful control of DNA fragment size distribution before immunoprecipitation.

An important premise is that our recommendations apply to the experimental conditions investigated by this study. Generalization should be avoided since altering critical parameters in the experiment (e.g. incorporating S1 nuclease (S1-DRIP) (Wahba et al. 2016) or lambda exonuclease digestion (DRIP-exo) (Ohle et al. 2016), or changing the model organism) might significantly affect the outcome of RNA-DNA hybrid detection.

In conclusion, the DRIP method remains a gold-standard for identifying *bona fide* R-loop binding sites across individual chromosomes, but a continued effort is needed to find alternatives and test complementary protocols. We hope that this aim has been achieved, at least in part, by this study that will help recognize real R-loop binding events and enable a better interpretation of DRIP-seq mapping data.

METHODS

Detection of RNA-DNA hybrids by DNA-RNA immunoprecipitation (DRIP)

I. DRIP classifiers 1-16

Crosslinking (Step 1)

Crosslinking of Jurkat cells (experiments 1-8) was done with 1% paraformaldehyde (UP) for 10 minutes, then quenched with 2.5 M glycine (pH 6, final concentration: 500 mM) for 5 minutes at room temperature. Crosslinking was omitted from experiments 9-16.

Cell lysis

Cells were lysed in 1 ml lysis buffer composed of 500 µl 2x lysis buffer (1% SDS, 20 mM Tris-HCl pH 7.5, 40 mM EDTA pH 8, 100 mM NaCl, ddH2O) plus 500 µl TE buffer (100 mM Tris-HCl pH 8, 10 mM EDTA pH 8) per 5 million cells. Cell lysis was performed at two different temperatures: either at 65 °C for 7 hours, or at 37 °C overnight, as indicated in the text.

Phenol chloroform extraction of total nucleic acid (Step 2)

In experiments 1-4, and 9-12, total nucleic acid was prepared by phenol-chloroform extraction. Before the phenol-chloroform extraction step, the nucleic acid preps were treated with 10 µl of Proteinase K (20 mg/ml; Thermo Fisher Scientific) at 65 °C for 7 hours, or at 37 °C overnight, to remove the proteins. The extracted DNA was precipitated with 1/10 volume 3 M Na-acetate (pH 5.2) plus 1 volume of isopropanol. The DNA pellet was dissolved in 200 µl of 10 mM Tris-HCl pH 8.

Silica membrane-based (kit) extraction of total nucleic acid (Step 2)

In experiments 5-8 and 13-16, total nucleic acid was isolated by the NucleoSpin Tissue Kit (Macherey-Nagel) according to the manufacturer's protocol, except the cell lysis step that was performed either at 65 °C for 7 hours (according to the kit protocol), or at 37 °C overnight, where indicated in the text. Nucleic acids were eluted in 500 µl of elution buffer (5 mM Tris-HCl pH 8.5).

Removal of free RNA by RNase A treatment (Step 3)

In experiments 3-4, 7-8, 11-12, and 15-16, the DNA purification step was directly followed by the RNase A digestion of free ribonucleic acids. The purified DNA preps (from Step 2) were supplemented with 18 µl of 5 M NaCl and 2 µl of RNase A (10 mg/ml; UD-GenoMed Ltd.) in a buffer containing 10 mM Tris-HCl (pH 8) and 300 mM NaCl (V=300 µl) at 37 °C for 1 hour. RNase A-treated samples were re-purified either by phenol-chloroform extraction (experiments 4, 12) or by the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) (experiments 8, 16). Phenol-chloroform extracted DNA was dissolved in 100 µl of 5 mM Tris-HCl pH 8.5. The DNA purified with the kit was eluted in 5 mM Tris-HCl pH 8.5.

Nucleic acid fragmentation by sonication (Step 4)

In experiments 1, 3, 5, 7, 9, 11, 13, 15, the purified nucleic acid preps were sonicated in a buffer of 10 mM Tris-HCl pH 8.5 supplemented with 300 mM NaCl (V=300 µl) for 2 x 5 min (30 sec ON, 30 sec OFF, LOW; Bioruptor, Diagenode) to yield an average DNA fragment size of ~300 bp.

Nucleic acid fragmentation by restriction enzyme digestion (Step 4)

In exp. 2, 4, 6, 8, 10, 12, 14, 16, purified DNA samples (~25 µg each) were fragmented using a restriction enzyme cocktail of 1 µl HindIII (20 U/µl), 1 µl EcoRI (20 U/µl), 2 µl BsrGI (10 U/µl), 1 µl XbaI (20 U/µl), 4 µl SspI (5 U/µl)) in NEB Buffer 2 (NEB) (V=300 µl) at 37 °C, for 4 hours.

The fragmented DNA samples were re-purified either by phenol-chloroform extraction (experiments 1-4; 9-12) or by the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) (experiments 5-8; 13-16). The DNA was dissolved in 100 µl of 5 mM Tris-HCl pH 8.5.

Two percent (V/V%) of the DNA preps were kept as input DNA for the DRIP-qPCR measurement. Half of samples were treated by 8 µl of RNase H (5000 U/ml; NEB) in a total volume of 80 µl, at 37 °C, overnight.

II. DRIP classifiers 17-24

Crosslinking (Step 1)

Crosslinking of Jurkat cells (experiments 17-20) was done with 1% paraformaldehyde (UP) for 10 minutes, then quenched with 2.5 M glycine (pH 6, final concentration: 500 mM) for 5 minutes at room temperature. Crosslinking was omitted from experiments 21-24.

Chromatin preparation (Step 2)

Cell lysis

Cells were lysed 750 µl of ChIP lysis buffer (50 mM HEPES-KOH at pH 7.5, 140 mM NaCl, 1 mM EDTA at pH 8, 1% Triton X-100, 0.1% Na-Deoxycholate, 1% SDS) per 10 million cells and homogenized using Fast Prep-24 5G (MP Biomedicals, speed: 6 m/s; time: 40 sec; 2 cycles; pause time: 120 sec; A lysing matrix).

Chromatin fragmentation by sonication (Step 3)

300 µl of chromatin preps were sonicated for 2 x 5 min (30 sec ON, 30 sec OFF, LOW, Bioruptor) to yield an average DNA fragment size of ~300 bp.

Removal of free RNA by RNase A treatment (Step 4)

In experiments 19, 20, 23, 24, the sonication step was directly followed by the RNase A digestion of free ribonucleic acids. The fragmented chromatin was supplemented with 270 µl of 5 M NaCl (300 mM) and 10 µl of RNase A (10 mg/ml; UD-GenoMed Ltd.) in 4500 µl of TE buffer (10 mM Tris-HCl pH 8, 10 mM EDTA pH 8) at 37 °C for 1 hour.

Before Step 5, the chromatin preps were treated with 30 µl of Proteinase K (20 mg/ml; Thermo Fisher Scientific) at 65 °C overnight to remove the proteins and reverse the cross-links.

Phenol chloroform extraction of total nucleic acid (Step 5)

In experiments 17, 19, 21, 23, total nucleic acid was prepared by phenol-chloroform extraction. The extracted DNA was precipitated with 1/10 volume 3 M Na-acetate (pH 5.2) plus 1 volume of isopropanol. The DNA pellet was dissolved in 100 µl of 5 mM Tris-HCl pH 8.5.

Silica membrane-based (kit) extraction of total nucleic acid (Step 5)

In experiments 18, 20, 22, 24, total nucleic acids were isolated by the NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) according to the manufacturer's protocol. Nucleic acids were eluted in 100 µl of elution buffer (5 mM Tris-HCl pH 8.5).

Two percent (V/V%) of the DNA preps were kept as input DNA for the DRIP-qPCR measurement. Half of samples were treated by 8 µl of RNase H (5000 U/ml; NEB) in a total volume of 80 µl, at 37 °C, overnight.

III. RNA-DNA hybrid Immunoprecipitation with the S9.6 antibody

Dynabeads Protein A magnetic beads (ThermoFisher Scientific) were pre-blocked with PBS/EDTA containing 0.5% BSA. To immobilize the S9.6 antibody, 50 µl pre-blocked Dynabeads Protein A was incubated with 10 µg of S9.6 antibody in IP buffer (50 mM Hepes/KOH at pH 7,5; 0,14 M NaCl; 5 mM EDTA; 1% Triton X-100; 0,1 % Na-Deoxycholate, ddH2O) at 4°C for 4 hours with rotation. Six micrograms of digested genomic DNA was added to the mixture and gently rotated at 4°C, overnight. Beads were recovered and washed successively with 1ml lysis buffer (low salt, 50 mM Hepes/KOH pH 7.5, 0.14 M NaCl, 5 mM EDTA pH 8, 1% Triton X-100, 0.1 % Na-Deoxycholate), 1ml lysis buffer (high salt, 50 mM Hepes/KOH pH 7.5, 0.5 M NaCl, 5 mM EDTA pH 8, 1% Triton X-100, 0.1 % Na-Deoxycholate), 1ml wash buffer (10 mM Tris-HCl pH 8, 0.25M LiCl, 0.5% NP-40, 0.5% Na-Deoxycholate, 1 mM EDTA pH 8) and 1ml TE (100 mM Tris-HCl pH 8, 10 mM EDTA pH 8) at 4°C, two times. Elution was performed in 100 µl of elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1 % SDS) for 15 min at 65 ˚C. After purification by NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel), nucleic acids were eluted in 55 µl of elution buffer (5 mM Tris-HCl pH 8.5). The recovered DNA was then analyzed by quantitative real-time PCR (qPCR). qPCR was performed with LightCycler 480 SYBR Green I Master (Roche) and analyzed on QuantStudio 12K Flex Real-Time PCR System (Thermo Fisher Scientific). Primer sequences are listed on Supplemental Table S3. qPCR results were analyzed using the comparative C_T method. The RNA-DNA hybrid enrichment was calculated based on the IP/Input ratio.

DRIP-sequencing

DRIP-sequencing was performed in human Jurkat cells and naive T CD4+ lymphocytes. Full description of the DRIP-seq experiment and bioinformatics analysis can be found in the Supplemental Material.

Receiver Operating Characteristic (ROC) Analysis

ROC curves were obtained for each DRIP variables (DRIP experiments) by ranking the studied genomic loci having known RNA-DNA hybrid states (based on the training set) according to their DRIP-qPCR profile, starting from the lowest to the highest estimated DRIP scores and then calculating sensitivity and specificity. The ROC curves plotted the sensitivity or true positive rate (TPR) against the false-positive rate (FPR) or 1-specificity, estimated as follows: TPR = P(positive DRIP-qPCR result | R-loop present), FPR = P(positive DRIP-qPCR result | R-loop absent), where P means conditional probability. The AUC values were then calculated from the observed DRIP-qPCR (IP/input) yields using the pROC algorithm.

DATA ACCESS

DRIP Sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra) under accession number SRP095885.

ACKNOWLEDGMENTS

L.Sz. received funding from the Hungarian Academy of Sciences (Lendület programme, LP2015-9/2015), from CRP-ICGEB, Italy (CRP/HUN13-01), from the European Union (FP7/Marie Cure Actions/CIG_#292259), from IMéRA / Inserm and the Aix-Marseille University (France), from NKFIH_ERC_HU_#117670, and from H2020/NKFIH_GINOP-2.3.2-15-2016-00024. We are grateful for Ibolya Fürtös for the excellent technical assistance and for the Genomic Medicine and Bioinformatics Core Facility (University of Debrecen) for the NGS service. We thank Dr. György Fenyőfalvi for critical discussions on the topic and for his idea about the RNase H-like activity of RNase A. We thank Dr. Gábor Szabó for providing us with the S9.6 antibody.

AUTHOR CONTRIBUTIONS

B.B.O., T.K.R., É.S., É.N. and Á.M.L. conceived and performed the experiments, A.M., É.R., G.H. and Zs.K. provided reagents, expertise and feedback, L.H., Zs.K. and L.Sz. analyzed the data, L.Sz. wrote the manuscript and secured funding.

DISCLOSURE DECLARATION

The authors declare no conflict of interest.

REFERENCES

Alzu A, Bermejo R, Begnis M, Lucca C, Piccini D, Carotenuto W, Saponaro M, Brambati A, Cocito A, Foiani M, et al. 2012. Senataxin associates with replication forks to protect fork integrity across RNA-polymerase-II-transcribed genes. *Cell* **151**: 835–46.

Baranello L, Kouzine F, Sanford S, Levens D. 2016. ChIP bias as a function of cross-linking time. *Chromosom Res* **24**: 175–181.

Beneke S, Meyer K, Holtz A, Hüttner K, Bürkle A. 2012. Chromatin composition is changed by poly(ADP-ribosyl)ation during chromatin immunoprecipitation. *PLoS One* **7**: 1–10.

Benore-Parsons M, Ayoub MA. 1997. Presence of RNase a causes aberrant DNA band shifts. *Biotechniques* **23**: 128–131.

Bhatia V, Barroso SI, García-Rubio ML, Tumini E, Herrera-Moyano E, Aguilera A. 2014. BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511(7509)**: 362-365.

Boque-Sastre R, Soler M, Oliveira-Mateos C, Portela A, Moutinho C, Sayols S, Villanueva A, Esteller M, Guil S. 2015. Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc Natl Acad Sci* **112**: 201421197.

Boulé J-B, Zakian V a. 2007. The yeast Pif1p DNA helicase preferentially unwinds RNA DNA substrates. *Nucleic Acids Res* **35**: 5809–18.

Brown T a, Tkachuk AN, Clayton D a. 2008. Native R-loops persist throughout the mouse mitochondrial DNA genome. *J Biol Chem* **283**: 36743–51.

Castellano-Pozo M, Santos-Pereira JM, Rondón AG, Barroso S, Andújar E, Pérez-Alegre M, García-Muse T, Aguilera A. 2013. R Loops Are Linked to Histone H3 S10 Phosphorylation and Chromatin Condensation. *Mol Cell* **52**: 1–8.

Chan Y a, Aristizabal MJ, Lu PYT, Luo Z, Hamza A, Kobor MS, Stirling PC, Hieter P. 2014. Genome-Wide Profiling of Yeast DNA:RNA Hybrid Prone Sites with DRIP-Chip. *PLoS Genet* **10**: e1004288.

Chen PB, Chen H V, Acharya D, Rando OJ, Fazzio TG. 2015. R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nat Struct Mol Biol* **22**: 999–1007.

Chon H, Sparks JL, Rychlik M, Nowotny M, Burgers PM, Crouch RJ, Cerritelli SM. 2013. RNase H2 roles in genome integrity revealed by unlinking its activities. *Nucleic Acids Res* 1–14.

Cloutier SC, Wang S, Ma WK, Al Husini N, Dhoondia Z, Ansari A, Pascuzzi PE, Tran EJ. 2016. Regulated Formation of lncRNA-DNA Hybrids Enables Faster Transcriptional Induction and Environmental Adaptation. *Mol Cell* **61**: 393–404.

Domínguez-Sánchez MS, Barroso S, Gómez-González B, Luna R, Aguilera A. 2011. Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. *PLoS Genet* **7**: e1002386.

Dona F, Houseley J. 2014. Unexpected DNA loss mediated by the DNA binding activity of ribonuclease A. *PLoS One* **9**: 1–11.

El Hage A, French SL, Beyer AL, Tollervey D. 2010. Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. *Genes Dev* **24**: 1546–58.

El Hage A, Webb S, Kerr A, Tollervey D. 2014. Genome-Wide Distribution of RNA-DNA Hybrids Identifies RNase H Targets in tRNA Genes, Retrotransposons and Mitochondria. *PLoS Genet* **10**.

Gan W, Guan Z, Liu J, Gui T, Shen K, Manley JL, Li X. 2011. R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev* **25**: 2041–56.

García-Rubio ML, Pérez-Calero C, Barroso SI, Tumini E, Herrera-Moyano E, Rosado I V., Aguilera A. 2015. The Fanconi Anemia Pathway Protects Genome Integrity from R-loops. *PLoS Genet* **11**: 1–17.

Ginno P a, Lott PL, Christensen HC, Korf I, Chédin F. 2012a. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol Cell* **45**: 814–25.

Ginno PA, Lott PL, Christensen HC, Korf I, Chédin F. 2012b. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol Cell* **45**.

Groh M, Lufino MMP, Wade-Martins R, Gromak N. 2014. R-loops Associated with Triplet Repeat Expansions Promote Gene Silencing in Friedreich Ataxia and Fragile X Syndrome. *PLoS Genet* **10**: e1004318.

Hatchi E, Skourti-Stathaki K, Ventz S, Pinello L, Yen A, Kamieniarz-Gdula K, Dimitrov S, Pathania S, McKinney KM, Eaton ML, et al. 2015. BRCA1 Recruitment to Transcriptional Pause Sites Is Required for R-Loop-Driven DNA Damage Repair. *Mol Cell* **57**: 636–647.

Herrera-Moyano E, Mergui X, García-Rubio ML, Barroso S, Aguilera A. 2014. The yeast and human FACT chromatin-reorganizing complexes solve R-loop-mediated transcription-replication conflicts. *Genes Dev* **1**: 735–748.

Hu Z, Zhang A, Storz G, Gottesman S, Leppla SH. 2006. An antibody-based microarray assay for small RNA detection. *Nucleic Acids Res* **34**: 1–7.

Huertas P, Aguilera A. 2003. Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol*

Cell **12**: 711–721.

Jenjaroenpun P, Wongsurawat T, Yenamandra SP, Kuznetsov VA. 2015. QmRLFS-finder: A model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res* **43**: W527–W534.

Li X, Manley JL. 2005. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122**: 365–378.

Lim YW, Sanz LA, Xu X, Hartono SR, Chédin F. 2015. Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutières syndrome. *Elife* **4**: e08007.

Loomis EW, Sanz L a, Chédin F, Hagerman PJ. 2014. Transcription-Associated R-Loop Formation across the Human FMR1 CGG-Repeat Region. *PLoS Genet* **10**: e1004294.

Marinello J, Bertoncini S, Aloisi I, Cristini A, Tagliazucchi GM, Forcato M, Sordet O, Capranico G. 2016. Dynamic effects of topoisomerase i inhibition on R-loops and short transcripts at active promoters. *PLoS One* **11**: 1–18.

McGhee JD, von Hippel PH. 1977. Formaldehyde as a Probe of DNA Structure. 4. Mechanism of the Initial Reaction of Formaldehyde with DNA. *Biochemistry* **16**: 3276–3293.

Mischo HE, Gómez-González B, Grzechnik P, Rondón AG, Wei W, Steinmetz L, Aguilera A, Proudfoot NJ. 2011. Yeast Sen1 helicase protects the genome from transcription-associated instability. *Mol Cell* **41**: 21–32.

Nadel J, Athanasiadou R, Lemetre C, Wijetunga NA, Broin PÓ, Sato H, Zhang Z, Jeddeloh J, Montagna C, Golden A, et al. 2015. RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics Chromatin* 1–19.

Nakama M, Kawakami K, Kajitani T, Urano T, Murakami Y. 2012. DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation. *Genes Cells* **17**: 218–33.

Ohle C, Tesorero R, Schermann G, Dobrev N, Sinning I, Fischer T. 2016. Transient RNA-DNA Hybrids are Required for Efficient Double-Strand Break Repair. *Cell* **167**: 1001–1013.

Pefanis E, Wang J, Rothschild G, Lim J, Kazadi D, Sun J, Federation A, Chao J, Elliott O, Liu Z-P, et al. 2015. RNA Exosome-Regulated Long Non-Coding RNA Transcription Controls Super-Enhancer Activity. *Cell* **161**: 774–789.

Phillips DD, Garboczi DN, Singh K, Hu Z, Leppla SH, Leysath CE. 2013. The sub-nanomolar binding of DNA-RNA hybrids by the single-chain Fv fragment of antibody S9.6. *J Mol Recognit* **26**: 376–381.

Pohjoismäki JLO, Holmes JB, Wood SR, Yang M-Y, Yasukawa T, Reyes A, Bailey LJ, Cluett TJ, Goffart S, Willcox S, et al. 2010. Mammalian mitochondrial DNA replication intermediates are essentially duplex but contain extensive tracts of RNA/DNA hybrid. *J Mol Biol* **397**: 1144–55.

Rigby RE, Webb LM, Mackenzie KJ, Li Y, Leitch A, Reijns M a M, Lundie RJ, Revuelta A, Davidson DJ, Diebold S, et al. 2014. RNA:DNA hybrids are a novel molecular pattern sensed by TLR9. *EMBO J* **33**: 542–58.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**: 77.

Romanello M, Schiavone D, Frey A, Sale JE. 2016. Histone H3.3 promotes IgV gene diversification by enhancing formation of AID-accessible single-stranded DNA. *EMBO J* 1–13.

Salvi JS, Chan JNY, Szafranski K, Liu TT, Wu JD, Olsen JB, Khanam N, Poon BPK, Emili A, Mekhail K. 2014. Roles for Pbp1 and Caloric Restriction in Genome and Lifespan Maintenance via Suppression of RNA-DNA Hybrids. *Dev Cell* **30**: 177–191.

Sanz LA, Hartono SR, Lim YW, Ginno PA, Sanz LA, Hartono SR, Lim YW, Steyaert S, Rajpurkar A, Ginno PA, et al. 2016. Prevalent , Dynamic , and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell* **63**: 167–178.

Schwab RA, Nieminuszczy J, Shah F, Langton J, Lopez Martinez D, Liang CC, Cohn MA, Gibbons RJ, Deans AJ, Niedzwiedz W. 2015. The Fanconi Anemia Pathway Maintains

Genome Stability by Coordinating Replication and Transcription. *Mol Cell* **60**: 351–361.

Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* **42**: 794–805.

Sollier J, Stork CT, García-Rubio ML, Paulsen RD, Aguilera A, Cimprich KA. 2014. Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability. *Mol Cell* **56**: 777–785.

Stork CT, Bocek M, Crossley MP, Sollier J, Sanz LA, Chédin F, Swigut T, Cimprich KA. 2016a. Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife* **5**: e17548.

Stork CT, Bocek M, Crossley MP, Sollier J, Sanz LA, Chédin F, Swigut T, Cimprich KA. 2016b. Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife* **5**: e17548.

Stuckey R, García-Rodríguez N, Aguilera A, Wellinger RE. 2015. Role for RNA:DNA hybrids in origin-independent replication priming in a eukaryotic system. *Proc Natl Acad Sci* **2015**: 201501769.

Sun Q, Csorba T, Skourti-Stathaki K, Proudfoot NJ, Dean C. 2013. R-loop stabilization represses antisense transcription at the Arabidopsis FLC locus. *Science* **340**: 619–21.

Székvölgyi L, Rákosy Z, Bálint BL, Kókai E, Imre L, Vereb G, Bacsó Z, Goda K, Varga S, Balázs M, et al. 2007. Ribonucleoprotein-masked nicks at 50-kbp intervals in the eukaryotic genomic DNA. *Proc Natl Acad Sci U S A* **104**: 14964–9.

Wahba L, Costantino L, Tan FJ, Zimmer A, Koshland D. 2016. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev* **30**: 1327–38.

Wahba L, Koshland D. 2013. The Rs of biology: R-loops and the regulation of regulators. *Mol Cell* **50**: 611–2.

Wilson-Sali T, Hsieh T-S. 2002. Preferential cleavage of plasmid-based R-loops and D-loops by Drosophila topoisomerase IIIbeta. *Proc Natl Acad Sci U S A* **99**: 7974–9.

Yang Y, La H, Tang K, Miki D, Yang L, Wang B, Duan C-G, Nie W, Wang X, Wang S, et al. 2016. SAC3B, a central component of the mRNA export complex TREX-2, is required for prevention of epigenetic gene silencing in *Arabidopsis*. *Nucleic Acids Res* gkw850.

Yang Y, McBride KM, Hensley S, Lu Y, Chedin F, Bedford MT. 2014. Arginine Methylation Facilitates the Recruitment of TOP3B to Chromatin to Prevent R Loop Accumulation. *Mol Cell* **53**: 484–497.

Yu K, Chedin F, Hsieh C-L, Wilson TE, Lieber MR. 2003. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat Immunol* **4**: 442–51.

Yu K, Roy D, Huang F-T, Lieber MR. 2006. Detection and structural analysis of R-loops. *Methods Enzymol* **409**: 316–29.

Zeller P, Padeken J, van Schendel R, Kalck V, Tijsterman M, Gasser SM. 2016. Histone H3K9 methylation is dispensable for Caenorhabditis elegans development but suppresses RNA:DNA hybrid-associated repeat instability. *Nat Genet* **48**.

Zhang ZZ, Pannunzio NR, Han L, Hsieh C-L, Yu K, Lieber MR. 2014a. The Strength of an Ig Switch Region Is Determined by Its Ability to Drive R Loop Formation and Its Number of WGCW Sites. *Cell Rep* **8**: 1–13.

Zhang ZZ, Pannunzio NR, Hsieh C-L, Yu K, Lieber MR. 2015. Complexities due to single-stranded RNA during antibody detection of genomic rna:dna hybrids. *BMC Res Notes* **8**.

Zhang ZZ, Pannunzio NR, Hsieh C-L, Yu K, Lieber MR. 2014b. The role of G-density in switch region repeats for immunoglobulin class switch recombination. *Nucleic Acids Res* 1–8.

FIGURE LEGENDS

Figure 1. Experimental Design: Constructing DRIP Schemes.

(*A*) Experiments 1-16 explore the effect of formaldehyde-fixation (Step 1), nucleic acid isolation (Step 2), removal of free RNA (Step 3), and nucleic acid fragmentation (Step 4) on the outcome of RNA-DNA hybrid detection. Each experiment was performed at two parallel cell lysis temperatures (65°C and 37°C), respectively. The temperature variable is not depicted in the cartoon, but it is referred in the main text. (*B*) Experiments 17-24 test the impact of acoustic sharing performed on a chromatin prep rather than on naked nucleic acid, similarly to the ChIP protocol. Each experiment was performed at 65°C cell lysis temperature. (*C*) Workflow of a ChIP experiment (shown only for comparison with the DRIP pipeline). Abbreviations: HCHO: Formaldehyde fixation; Phe/Chl: Phenol/Chloroform extraction; Kit: silica membrane-based nucleic acid purification; RNase A: Ribonuclease A digestion performed at high (300 mM) NaCl concentration. Son: Sonication; RE: restriction enzyme cocktail digestion (HindIII, EcoRI, BsrGI, XbaI and SspI). As a negative control, RNase H digestion was applied in all DRIP experiments (not indicated in the cartoon).

Figure 2. Summary of Available Human DRIP-seq Experiments.

(*A*) Bar chart showing the number of identified R-loop peaks in human Jurkat cells and naive T cells (this study). (*B*) Annotation of R-loop binding sites over functional genomic elements. DRIP-seq peaks were determined in Jurkat cells and naive T cells, and in other published cell types (NTERA2, K562, Fibroblast, MCF7, IMR90, HEK293T). The upper four rows represent DRIP experiments fragmenting the nucleic acid by sonication, while the lower five rows highlight restriction enzyme-digested DRIP samples. The difference between the two groups is especially noticeable over exons (associated to 14%-27% and 1%-3.5% of R-loops, respectively) and repeat elements (SINEs, LINEs, LTRs, simple and low complexity repeats) that involve 22%-38% and 54%-67% of the R-loop peaks, respectively. At other annotation categories (gene body, introns and promoters) the difference was not significant between the two groups. (*C*) Density plots showing the distribution of R-loop peaks sizes, classified by fragmentation method (restriction enzyme *vs.* sonication). Median peak length and 2.5%-97.5% quantiles are indicated. Peak length distributions differ significantly between the two fragmentation methods. (*D*) Heatmap showing the overlap of R-loop binding sites between independent DRIP-seq experiments. Values and cell colors represent pairwise and unique overlap ratios between each peak set. The difference between the two nucleic acid fragmentation methods is clearly apparent, as peak sets from the same fragmentation process better resemble to each other (highlighted in black).

Figure 3. Good DRIP Practice.

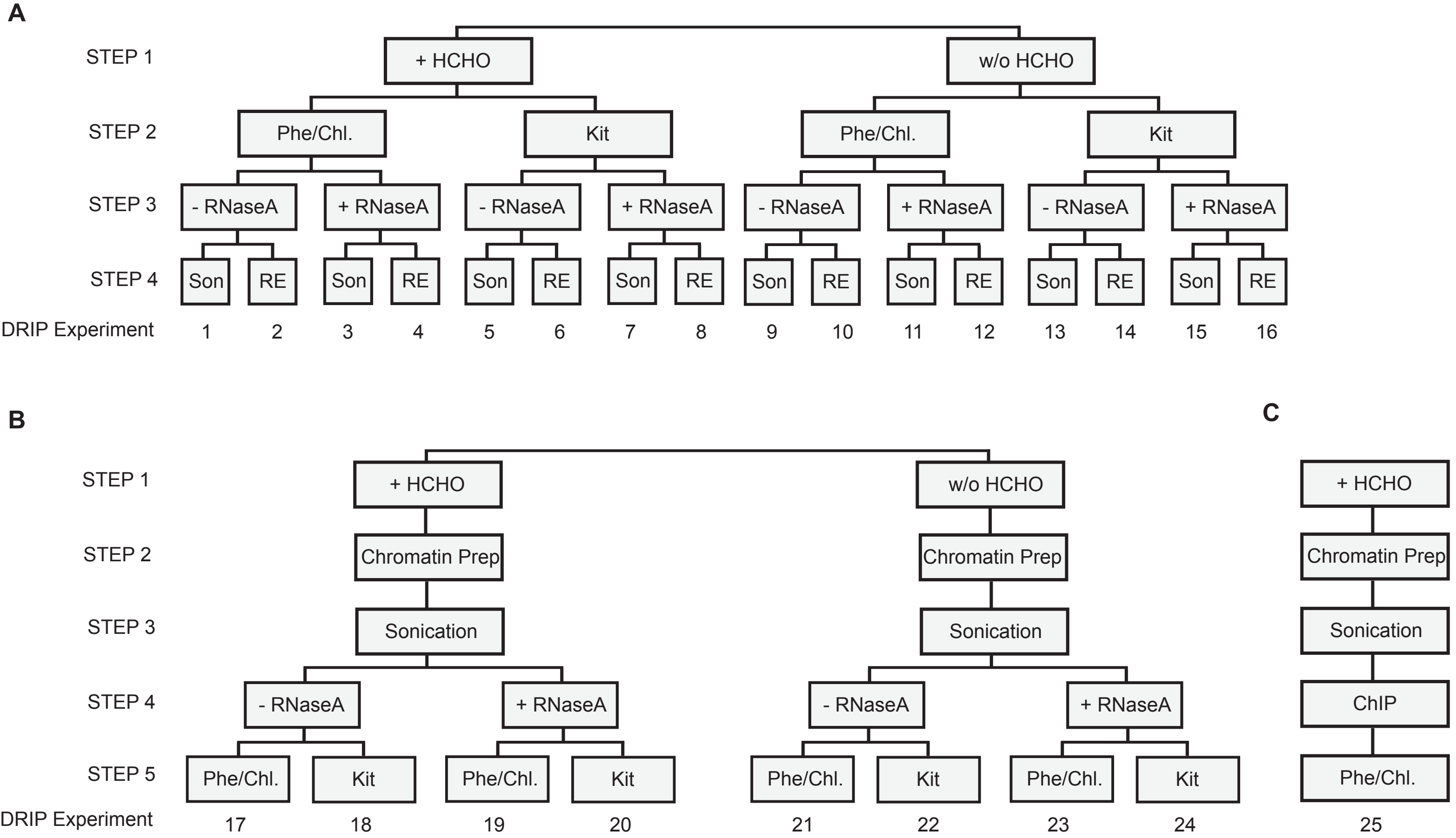
(*A*) Bar charts showing the distribution of AUC (area under the curve) values of ROC plots for twenty-four DRIP classifiers. Error bars represent the confidence interval of AUCs. High (>0.7) AUC values were obtained for ten DRIP classifiers (exp. 5, 6, 13, 15, 17, 18, 19, 21, and 24). Low (~0.5) AUC values were obtained in four DRIP experiments (exp. 2, 10, 11, and 16). We highlight these groups as “preferred” and “not preferred”, respectively. (*B*-*C*) The top four DRIP experiments ranked by AUCs (exp 5, 13, 17, 19). (*B*) DRIP-qPCR enrichment scores are displayed over the test regions. Horizontal dotted lines represent the cutoff value (calculated from the ROC curves) separating the true R-loop signal from background. (*C*) ROC curves of the top four experiments. (*D*) Paired-ROC plots, comparing the main variables (steps) of the DRIP experiments. The level of statistical significance was 0.05.

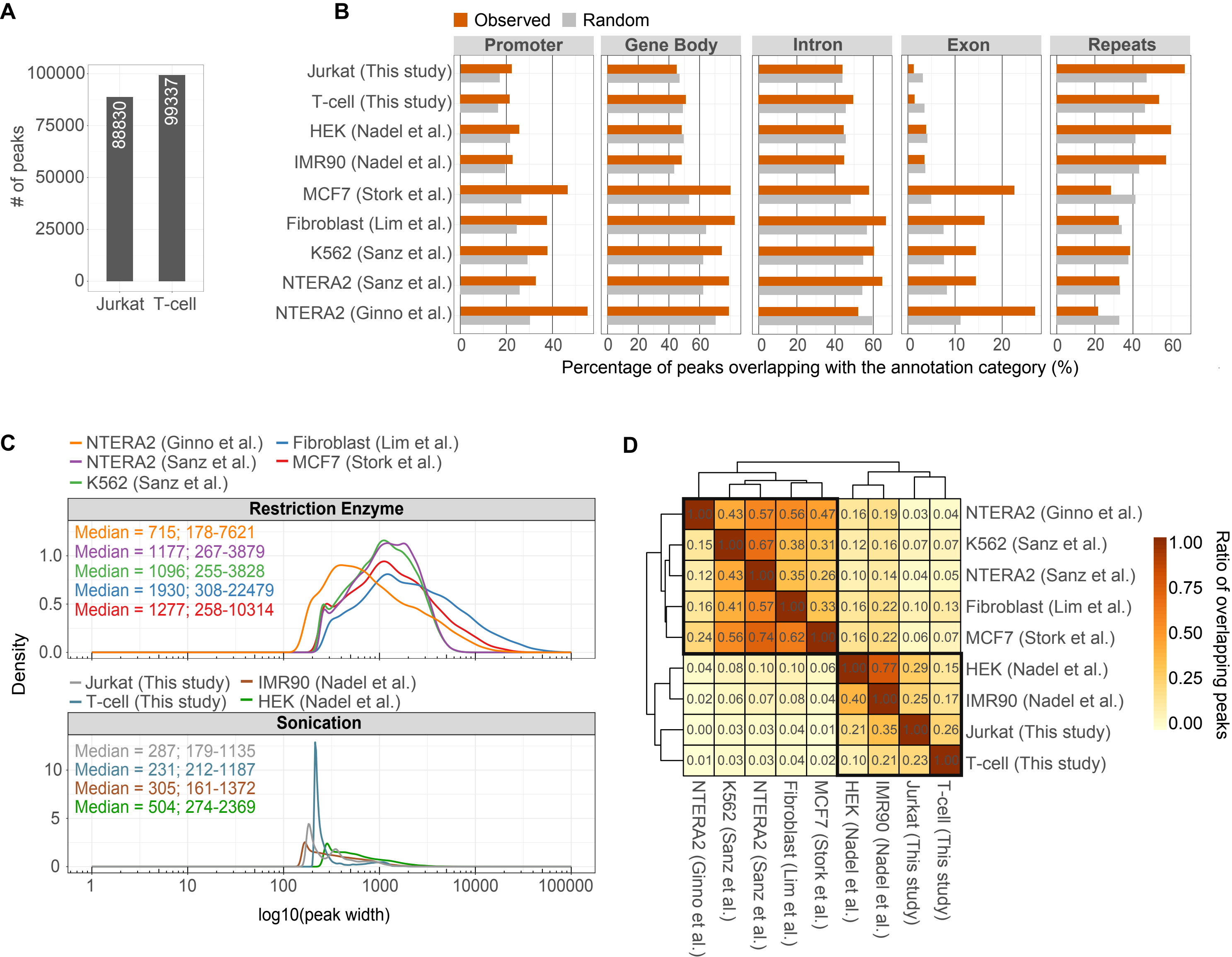
Figure 4. Analysis of restriction sites over genic and intergenic regions.

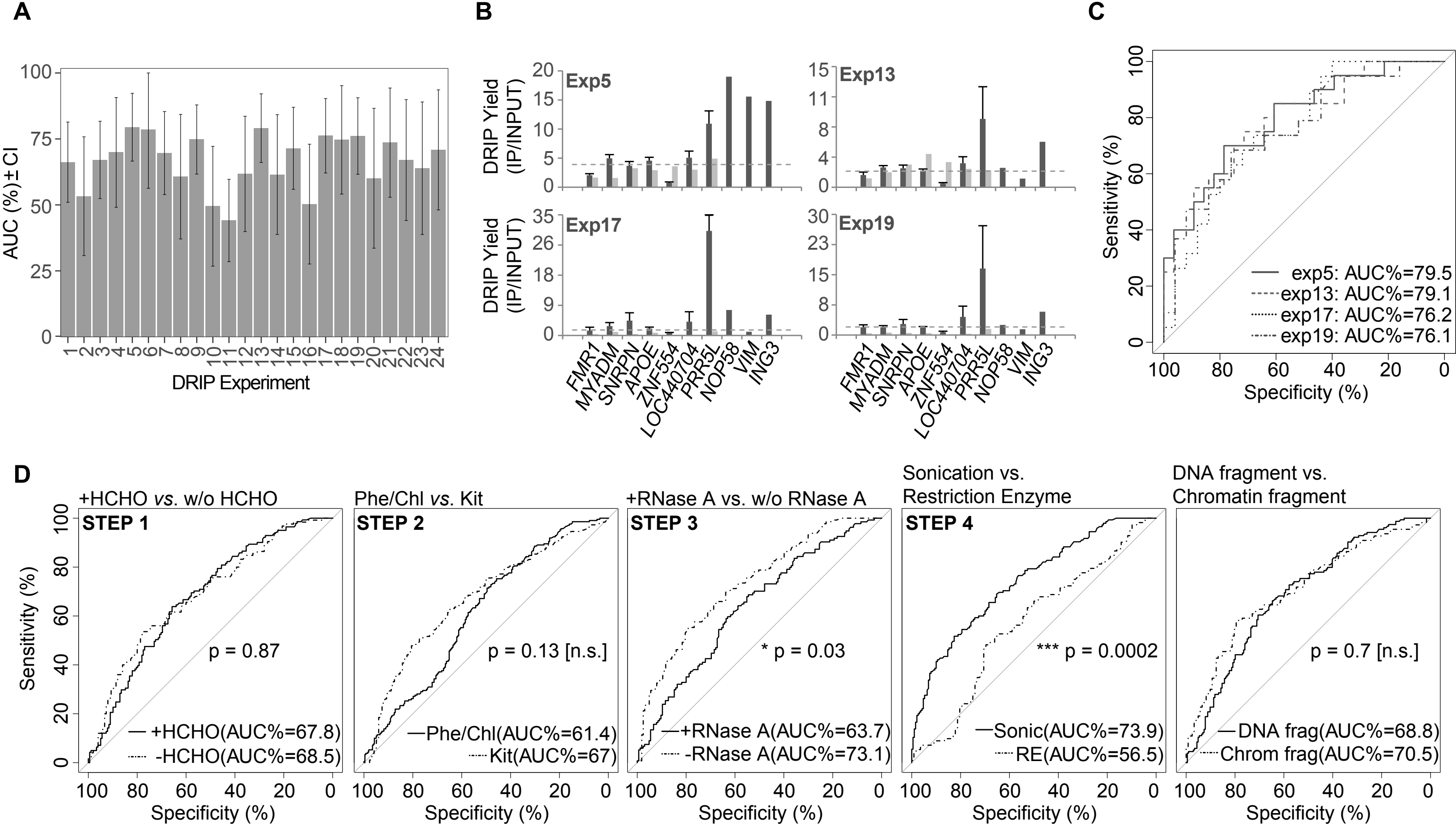
(*A*) Restriction fragment lengths over genic regions (gene bodies, exons, first exons) are significantly larger compared to intergenic regions. The plot shows the difference of genic (observed) and intergenic (expected) fragment sizes in base pairs. The following enzymes were applied in combination: HindIII, EcoRI, BsrGI, XbaI and SspI. (*B*-*C*-*D*) The number of restriction sites over genic regions is significantly lower compared to intergenic regions. Colors indicate the proportion of cutting sites in each category. Red and blue slices, marking the rarest restriction site frequencies, are prevalent over genic elements in each pie chart. (*E*) Cutting efficiency of restriction enzymes applied in the indicated DRIP-seq experiments. Zero read: the restriction site was cut. Greater equal than one read: the restriction site was uncut in a fraction of cells. There were uncut reads (sites) over half of the theoretical restriction sites. The proportion of uncut reads was even higher within gene coding regions compared to intergenic regions. See the model of cutting efficiency in panel (*F*).

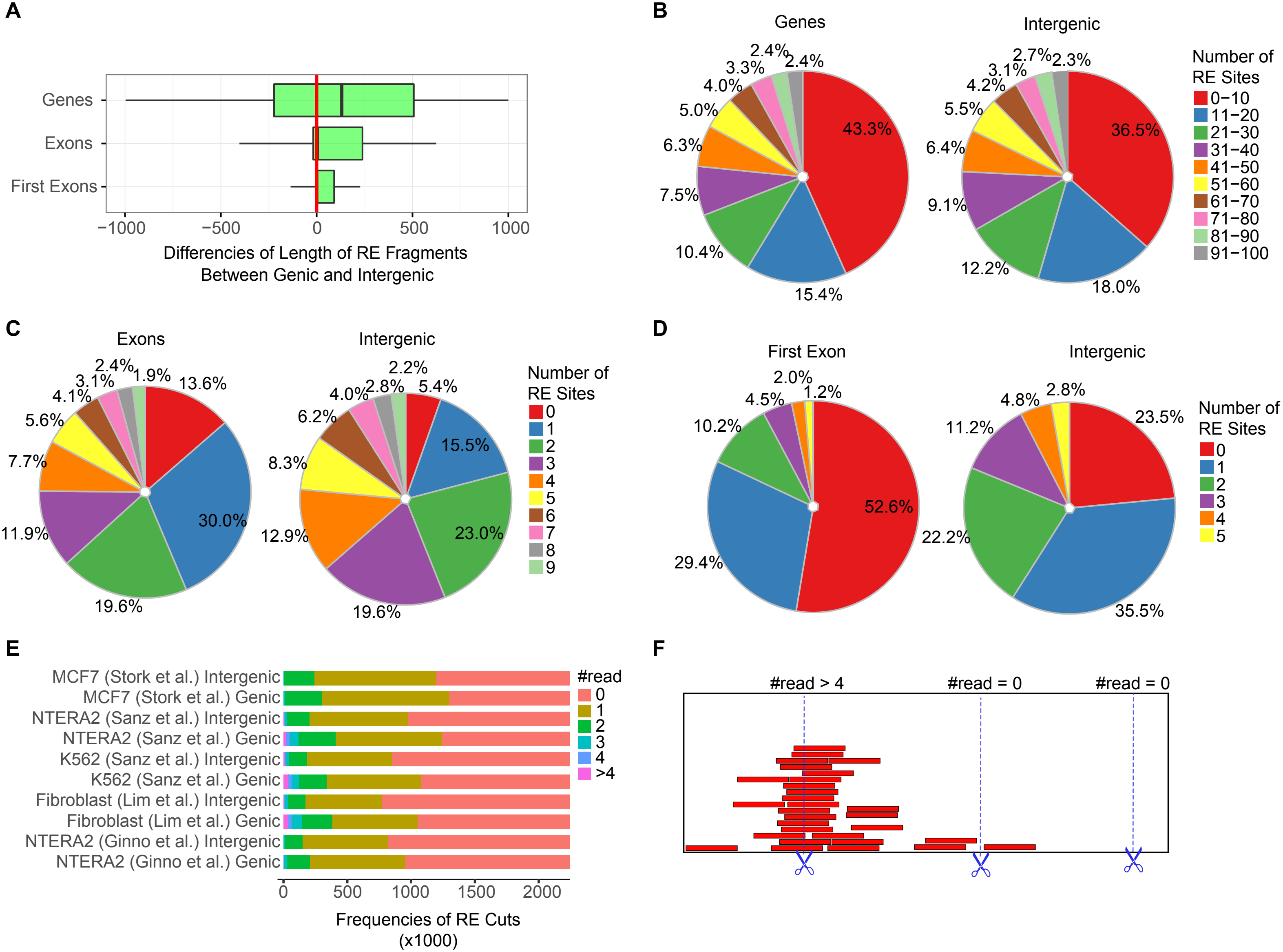
Figure 5. Large restriction fragments over gene bodies causes uncertainty in the precise localization of R-loops, potentially impeding their functional annotation.

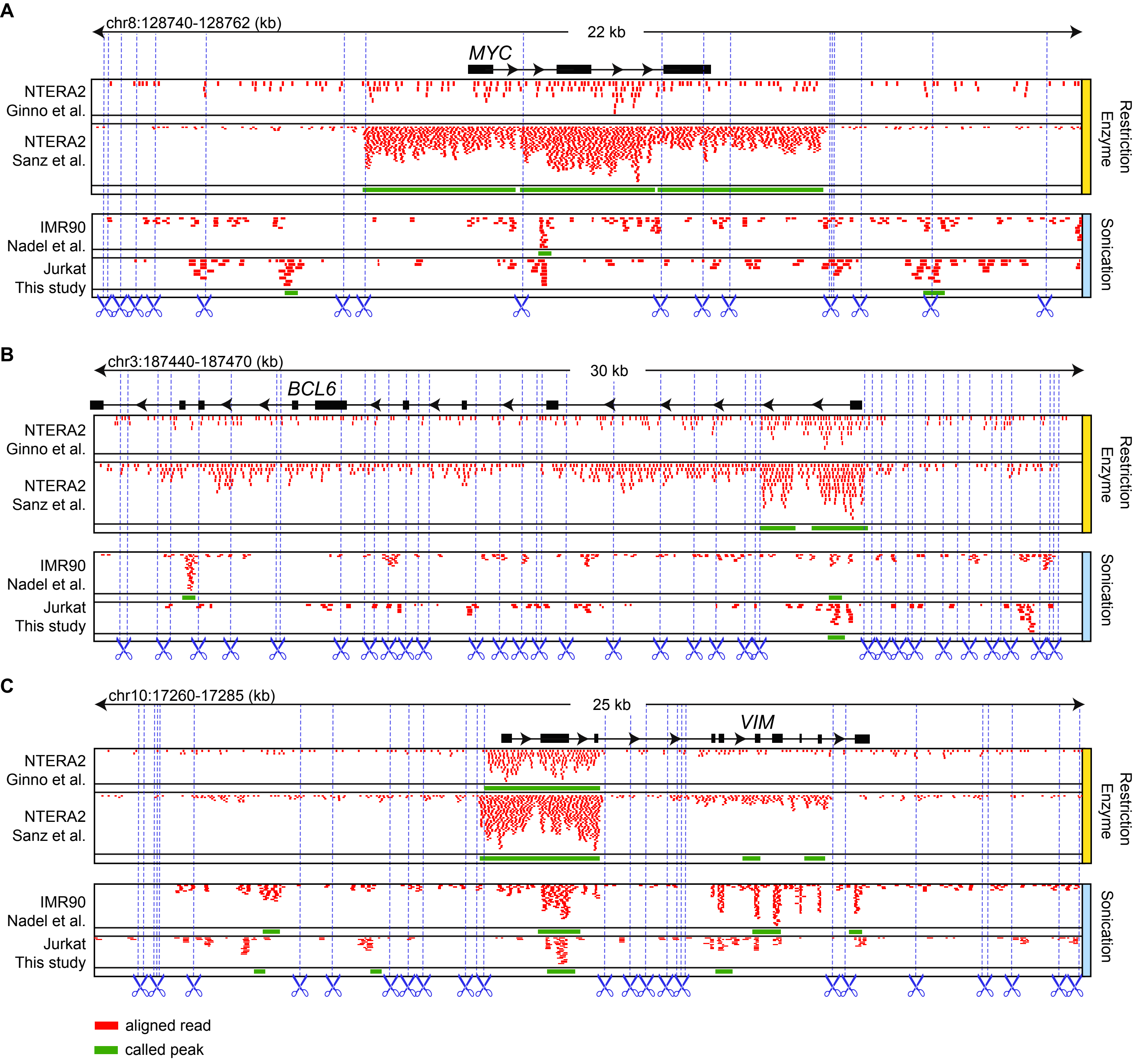
(*A*-*C*) Genome browser tracks showing three representative examples (*MYC*, *BCL6* and *VIM*). Upper two tracks: restriction fragment-sized R-loops are prevalent over the 5' prime end of genes, vastly exceeding the gene borders in the case of *MYC*. Lower two tracks: the precise genomic position of R-loops was resolved in the sonicated group of samples. Green boxes represent R-loop enriched regions predicted by the peak callers. Blue dashed lines represent cutting sites for restriction enzymes (HindIII, EcoRI, BsrGI, XbaI and SspI).


A
STEP 1
+ HCHO
w/o HCHO
STEP 2
Phe/Chl.
Kit
Phe/Chl.
Kit
STEP 3
- RNaseA
+ RNaseA
- RNaseA
+ RNaseA
- RNaseA
+ RNaseA
- RNaseA
+ RNaseA
STEP 4
Son
RE
Son
RE
Son
RE
Son
RE
Son
RE
Son
RE
Son
RE
Son
RE
DRIP Experiment
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
B
STEP 1
+ HCHO
w/o HCHO
STEP 2
Chromatin Prep
Chromatin Prep
STEP 3
Sonication
Sonication
STEP 4
- RNaseA
+ RNaseA
- RNaseA
+ RNaseA
STEP 5
Phe/Chl.
Kit
Phe/Chl.
Kit
Phe/Chl.
Kit
Phe/Chl.
Kit
DRIP Experiment
17
18
19
20
21
22
23
24
C
+ HCHO
Chromatin Prep
Sonication
ChIP
Phe/Chl.
25


A
# of peaks
100000
75000
50000
25000
0
88830
99337
Jurkat
T-cell
B
Observed
Random
Promoter
Gene Body
Intron
Exon
Repeats
Jurkat (This study)
T-cell (This study)
HEK (Nadel et al.)
IMR90 (Nadel et al.)
MCF7 (Stork et al.)
Fibroblast (Lim et al.)
K562 (Sanz et al.)
NTERA2 (Sanz et al.)
NTERA2 (Ginno et al.)
0 20 40
0 20 40 60
0 20 40 60
0 10 20
0 20 40 60
Percentage of peaks overlapping with the annotation category (%)
C
NTERA2 (Ginno et al.)
Fibroblast (Lim et al.)
NTERA2 (Sanz et al.)
MCF7 (Stork et al.)
K562 (Sanz et al.)
Restriction Enzyme
Median = 715; 178-7621
Median = 1177; 267-3879
Median = 1096; 255-3828
Median = 1930; 308-22479
Median = 1277; 258-10314
1.0
0.5
0.0
Density
Jurkat (This study)
IMR90 (Nadel et al.)
T-cell (This study)
HEK (Nadel et al.)
Sonication
Median = 287; 179-1135
Median = 231; 212-1187
Median = 305; 161-1372
Median = 504; 274-2369
10
5
0
1
10
100
1000
10000
100000
log10(peak width)
D
NTERA2 (Ginno et al.) 1.00 0.43 0.57 0.56 0.47 0.16 0.19 0.03 0.04
K562 (Sanz et al.) 0.15 1.00 0.67 0.38 0.31 0.12 0.16 0.07 0.07
NTERA2 (Sanz et al.) 0.12 0.43 1.00 0.35 0.26 0.10 0.14 0.04 0.05
Fibroblast (Lim et al.) 0.16 0.41 0.57 1.00 0.33 0.16 0.22 0.10 0.13
MCF7 (Stork et al.) 0.24 0.56 0.74 0.62 1.00 0.16 0.22 0.06 0.07
HEK (Nadel et al.) 0.04 0.08 0.10 0.10 0.06 1.00 0.77 0.29 0.15
IMR90 (Nadel et al.) 0.02 0.06 0.07 0.08 0.04 0.40 1.00 0.25 0.17
Jurkat (This study) 0.00 0.03 0.03 0.04 0.01 0.21 0.35 1.00 0.26
T-cell (This study) 0.01 0.03 0.03 0.04 0.02 0.10 0.21 0.23 1.00
NTERA2 (Ginno et al.)
K562 (Sanz et al.)
NTERA2 (Sanz et al.)
Fibroblast (Lim et al.)
MCF7 (Stork et al.)
HEK (Nadel et al.)
IMR90 (Nadel et al.)
Jurkat (This study)
T-cell (This study)
1.00
0.75
0.50
0.25
0.00
Ratio of overlapping peaks


A
AUC (%) ± CI
100
75
50
25
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
DRIP Experiment
B
DRIP Yield (IP/INPUT)
Exp5
20 15 10 5 0
Exp13
15 11 8 4 0
Exp17
35 26 18 9 0
Exp19
30 23 15 8 0
FMR1
MYADM
SNRPN
APOE
ZNF554
LOC440704
PRR5L
NOP58
VIM
ING3
C
Sensitivity (%)
Specificity (%)
100 80 60 40 20 0
exp5: AUC%=79.5
exp13: AUC%=79.1
exp17: AUC%=76.2
exp19: AUC%=76.1
D
+HCHO vs. w/o HCHO
STEP 1
p = 0.87
+HCHO(AUC%=67.8)
-HCHO(AUC%=68.5)
Phe/Chl vs. Kit
STEP 2
p = 0.13 [n.s.]
Phe/Chl(AUC%=61.4)
Kit(AUC%=67)
+RNase A vs. w/o RNase A
STEP 3
* p = 0.03
+RNase A(AUC%=63.7)
-RNase A(AUC%=73.1)
Sonication vs.
Restriction Enzyme
STEP 4
*** p = 0.0002
Sonic(AUC%=73.9)
RE(AUC%=56.5)
DNA fragment vs.
Chromatin fragment
p = 0.7 [n.s.]
DNA frag(AUC%=68.8)
Chrom frag(AUC%=70.5)
Sensitivity (%)
Specificity (%)
100 80 60 40 20 0


A
Genes
Exons
First Exons
−1000
−500
0
500
1000
Differencies of Length of RE Fragments
Between Genic and Intergenic
B
Genes
43.3%
15.4%
10.4%
7.5%
6.3%
5.0%
4.0%
3.3%
2.4%
2.4%
Intergenic
36.5%
18.0%
12.2%
9.1%
6.4%
5.5%
4.2%
3.1%
2.7%
2.3%
Number of
RE Sites
0−10
11−20
21−30
31−40
41−50
51−60
61−70
71−80
81−90
91−100
C
Exons
13.6%
30.0%
19.6%
11.9%
7.7%
5.6%
4.1%
3.1%
2.4%
1.9%
Intergenic
5.4%
15.5%
23.0%
19.6%
12.9%
8.3%
6.2%
4.0%
2.8%
2.2%
Number of
RE Sites
0
1
2
3
4
5
6
7
8
9
D
First Exon
52.6%
29.4%
10.2%
4.5%
2.0%
1.2%
Intergenic
23.5%
35.5%
22.2%
11.2%
4.8%
2.8%
Number of
RE Sites
0
1
2
3
4
5
E
MCF7 (Stork et al.) Intergenic
MCF7 (Stork et al.) Genic
NTERA2 (Sanz et al.) Intergenic
NTERA2 (Sanz et al.) Genic
K562 (Sanz et al.) Intergenic
K562 (Sanz et al.) Genic
Fibroblast (Lim et al.) Intergenic
Fibroblast (Lim et al.) Genic
NTERA2 (Ginno et al.) Intergenic
NTERA2 (Ginno et al.) Genic
0
500
1000
1500
2000
Frequencies of RE Cuts
(x1000)
#read
0
1
2
3
4
>4
F
#read > 4
#read = 0
#read = 0


A
chr8:128740-128762 (kb)
22 kb
MYC
NTERA2
Ginno et al.
NTERA2
Sanz et al.
IMR90
Nadel et al.
Jurkat
This study
Restriction
Enzyme
Sonication
B
chr3:187440-187470 (kb)
30 kb
BCL6
NTERA2
Ginno et al.
NTERA2
Sanz et al.
IMR90
Nadel et al.
Jurkat
This study
Restriction
Enzyme
Sonication
C
chr10:17260-17285 (kb)
25 kb
VIM
NTERA2
Ginno et al.
NTERA2
Sanz et al.
IMR90
Nadel et al.
Jurkat
This study
Restriction
Enzyme
Sonication
🟥 aligned read
🟩 called peak