scientific reports

Check for updates

OPEN

Transcriptome-wide high-throughput mapping of protein–RNA occupancy profiles using POP-seq

Mansi Srivastava[1], Rajneesh Srivastava[1] & Sarath Chandra Janga[1,2,3]✉

Interaction between proteins and RNA is critical for post-transcriptional regulatory processes. Existing high throughput methods based on crosslinking of the protein–RNA complexes and poly-A pull down are reported to contribute to biases and are not readily amenable for identifying interaction sites on non poly-A RNAs. We present Protein Occupancy Profile-Sequencing (POP-seq), a phase separation based method in three versions, one of which does not require crosslinking, thus providing unbiased protein occupancy profiles on whole cell transcriptome without the requirement of poly-A pulldown. Our study demonstrates that ~ 68% of the total POP-seq peaks exhibited an overlap with publicly available protein–RNA interaction profiles of 97 RNA binding proteins (RBPs) in K562 cells. We show that POP-seq variants consistently capture protein–RNA interaction sites across a broad range of genes including on transcripts encoding for transcription factors (TFs), RNA-Binding Proteins (RBPs) and long non-coding RNAs (lncRNAs). POP-seq identified peaks exhibited a significant enrichment (*p* value < 2.2e−16) for GWAS SNPs, phenotypic, clinically relevant germline as well as somatic variants reported in cancer genomes, suggesting the prevalence of uncharacterized genomic variation in protein occupied sites on RNA. We demonstrate that the abundance of POP-seq peaks increases with an increase in expression of lncRNAs, suggesting that highly expressed lncRNA are likely to act as sponges for RBPs, contributing to the rewiring of protein–RNA interaction network in cancer cells. Overall, our data supports POP-seq as a robust and cost-effective method that could be applied to primary tissues for mapping global protein occupancies.

Interaction of proteins with RNA is crucial for post-transcriptional gene regulation such as capping, splicing, polyadenylation and localization which is indispensable for cellular homeostasis and survival[1,2]. Despite the increasingly appreciated role of protein–RNA interactions, the global occupancy profiles of proteins in a cellular environment is not fully elucidated. For instance dysregulated expression of RNA binding proteins (RBPs) has been associated with a broad spectrum of human pathologies including cancers, neurological and hereditary diseases[3–6]. Therefore, it is critical to investigate the diverse protein occupancy sites and their functional impact on physiology and diseases.

Experimental approaches such as crosslinking followed by immunoprecipitation (CLIP) have been widely used to identify the binding pockets of specific RBPs across the transcriptome[7,8]. CLIP based methods exploit the stability of crosslinked protein–RNA complexes by ultraviolet (UV) irradiation followed by immunoprecipitation and sequencing of the co-purified RNA[9–11]. However, these methods are reported to contribute to biases in the interaction profiles due to the inherent nature of UV crosslinking[12–14]. Other methods that employ antibody pulldown of protein–RNA complexes such as RIP-seq and DO-RIP-seq[15–17] are difficult to scale up for detecting hundreds of RNA interacting proteins at the same time. Methods like POPPI-seq[18] and others have enabled the capture of protein–RNA interaction sites by incorporation of photoreactive nucleosides in UV irradiated cells followed by poly-A pull down and sequencing of the bound RNA. Although, the use of photoreactive nucleosides

[1]Department of BioHealth Informatics, School of Informatics and Computing, Indiana University Purdue University, Informatics and Communications Technology Complex, IT475H, 535 West Michigan Street, Indianapolis, IN 46202, USA. [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 5021 Health Information and Translational Sciences (HITS), 410 West 10th Street, Indianapolis, IN 46202, USA. [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Medical Research and Library Building, 975 West Walnut Street, Indianapolis, IN 46202, USA. ✉email: scjanga@iupui.edu

is known to induce cellular stress that may result in non-physiological protein–RNA interactions and thus limits their application to only in-vitro cultures[8]. In addition, poly-A pulldown requirement in these methods excludes their application to non-polyadenylated RNAs such as lncRNAs, miRNAs and histone mRNAs[18,19]. Several methods also employ formaldehyde crosslinking to capture protein–RNA complexes, however formaldehyde is also known to introduce biases by capturing non-specific interactions[8,20–22]. Therefore, there is a need to develop unbiased and cost-effective method that can map the global occupancy profiles of protein bound sites in a transcriptome wide manner with its application to diverse species of RNA.

Trizol based phase separation strategy has emerged as a robust technology that has expanded the identification of protein binding sites independent of the poly-A capture[23–25]. Trizol extraction is deployed as a prevalent method to purify total RNA from the cell lysates. This involves solubilization of the biological material by phenol and guanidium isothiocyanate followed by chloroform induced phase separation. After the phase separation, proteins migrate to the organic phase, RNA migrates to the aqueous phase and the DNA/RNA–protein adducts are trapped in the interphase.

In this study, we propose a method called POP-seq (Protein Occupancy Profile-sequencing) that incorporates a multi-step phase separation strategy using trizol, followed by high-throughput sequencing of small RNAs, without the requirement of poly-A pulldown. Current study reports three versions of POP-seq; NPOP-seq (no-crosslinking), FPOP-seq (formaldehyde crosslinking) and UPOP-seq (UV crosslinking), among which NPOP-seq can efficiently capture the interactions without crosslinking mediated biases under physiological conditions. Computational analysis of POP-seq data revealed a significant enrichment of clinically relevant somatic variants in the protein–RNA interaction sites. Further, this study also demonstrates that highly expressed lncRNAs act as sponges to titrate the abundance of RBPs thereby altering the protein–RNA regulatory networks. Overall, POP-seq is a robust and cost-effective method that can be utilized by researchers to capture the protein occupied sites on all RNA types.

Results

POP-seq captures protein bound RNA fragments with three transcriptome wide approaches. We aimed to generate an unbiased transcriptome wide protein–RNA occupancy profiles using a trizol based phase separation method, POP-seq (Protein Occupancy Profile-sequencing) in K562 cells. Recent studies have demonstrated that phase separation using trizol yields abundance of RNA–protein interactions at the interphase[23,24]. However, identification of precise protein occupied pockets across transcriptome remains obscure.

POP-seq is employed in three different versions: NPOP-seq (no crosslinking), FPOP-seq (Formaldehyde crosslinking) and UPOP-seq (UV crosslinking) in K562 cells (Fig. 1A). POP-seq employs trizol lysis of cells that generates three phases: aqueous phase, interphase, and organic phase. After removal of aqueous and organic phases, interphase is subjected to RNase A/T1 digestion, to remove the unprotected RNA from the RNA–Protein complexes trapped in the interphase. This is followed by degradation of the bound protein using proteinase K leaving behind the small RNA pockets. Further, DNase treatment ensures the sample quality by eliminating any DNA traces that might arise from the interphase. This is followed by removal of highly abundant ribosomal RNA (Fig. 1B). Implementation of RNase digestion creates a 5′-hydroxyl, and 3′-phosphate ends in purified RNA making it inappropriate for adapter ligation during library preparation. Therefore, we modified RNA ends using T4 Polynucleotide kinase (PNK) and Calf intestine alkaline phosphatase (CIAP) to add 5′-phosphate and 3′-hydroxyl to the ends (Fig. 1B). RNA integrity was assessed by Bioanalyzer QC and small RNA libraries were prepared for high throughput short read sequencing by Illumina, next-seq platform (Fig. 1B). Together, this method allows identification of global protein occupancy across transcriptome.

POP-seq reproducibly capture the abundance of protein–RNA interactions on the exonic regions in a human leukemia cell line. To characterize the transcriptome wide binding of the proteins, POP-seq libraries (in replicates) were sequenced to generate ~ 20 million reads each. We implemented our NGS pipeline to facilitate the analysis of the POP-seq data which includes quality control and read alignment followed by peak calling, resulting in the identification of 319,657, 288,129, and 320,310 unique peaks in the respective protocols (Fig. 2A). Overall, ~ 85% of total peaks had length below 50 bp (Fig. 2B). Since reproducibility is an important aspect to estimate the robustness of high throughput methods, we compared the aligned reads per 10 kb genome in replicates for each protocol using deepTools[26] 'plot correlation' command. Our data showed a correlation of ~ 64% (spearman correlation R^2 values; 0.65, 0.64, 0.64 for NPOP, FPOP and UPOP respectively) between the replicates as shown in Figure S1. Comparison of POP-seq peaks with the combined eCLIP[27] profiles of 97 RBPs from the ENCODE project[28], with an end to end 50% peak overlap, revealed support for 68.2%, 67.3% and 66.4% in NPOP, FPOP and UPOP-seq respectively. Additionally, we observed that ~ 64% of the genes exhibited by POP-seq peaks are common across the three protocols (Fig. 2C). We observed that even the NPOP-seq can capture a significant fraction of genes targeted by proteins, while missing ~ 19% of the total identified genes by the other two versions of POP-seq (Fig. 2C).

Next, we examined the distribution of POP-seq peaks across transcript regions annotated by HOMER[29] and observed that majority of the peaks were mapped to the exons (~ 48%) while a relatively lower proportion were mapped to intronic regions (~ 19%) and 3′UTR's (~ 16%) (Fig. 2D). Our observation supporting the higher proportions of the peaks in exonic regions was in agreement with the previous reports in MCF7 and HEK293 cell lines[18,19]. Next, we examined the fraction of gene types exhibiting the POP-seq peaks which revealed that majority of the genes (~ 67.3%) mapped to protein coding followed by ~ 28.4% lncRNA and ~ 18.6% snoRNA with respect to the total genes annotated in the human reference genome (hg38) (Fig. 2E). These proportions were generally consistent across all the POP-seq protocols.

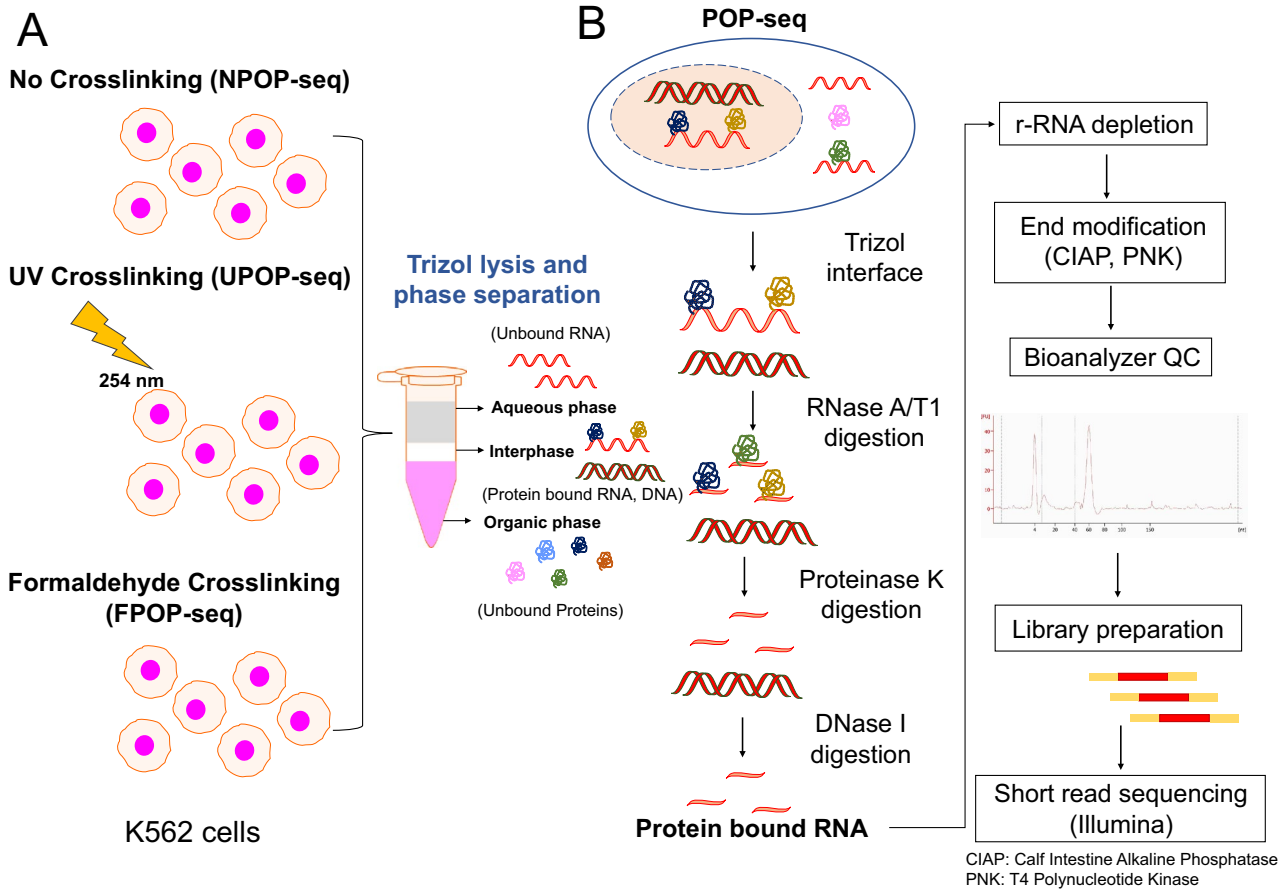

Figure 1. Experimental workflow of POP-seq in K562 cells. (**A**) Three versions of POP-seq, No-crosslinking (NPOP-seq), UV crosslinking, 254 nm (UPOP-seq) and formaldehyde crosslinking (FPOP-seq) generate three phases (aqueous, interphase and organic phase) upon trizol lysis. (**B**) Cell lysates from the POP-seq are digested with RNase A/T1 mix, Proteinase K and DNase followed by r-RNA depletion, RNA quality check and library preparation for short read Illumina sequencing.

POP-seq illustrates a significant enrichment for protein–RNA interactions. To rigorously evaluate whether POP-seq identifies any non-RBP binding events, we performed three comprehensive analyses as summarized below (i) To estimate non-RBP interactions, we compared the POP-seq peaks with ChIP-seq data of 67 proteins and CLIP-seq data of 79 proteins available for K562 cells (from ENCODE project). We observed a significantly higher overlap (p value < 2.2e−16) of POP-seq peaks with CLIP-seq peaks compared to ChIP-seq peaks as shown in Figure S2A, (ii) To estimate the protein–DNA interactions (false positives) that could be captured by POP-seq, we systematically compared the POP-seq peaks (and 5 random peak profiles separately) with the binding profile of 18 proteins for which both ChIP-seq and CLIP-seq data was available in K562 cells (from ENCODE project). Our results showed a significant enrichment (Odds ratio > 20 averaged across 5 random controls, p value < 2.2e−26, Fishers Exact test) of POP-seq signals overlapping with the CLIP-seq profile than the ChIP-seq profile of these 18 proteins (Figure S2B), indicating that POP-seq peaks are enriched for protein–RNA interactions. Among the 18 tested proteins, NONO (Non-POU domain-containing octamer-binding protein), which is known to bind both DNA and RNA[30,31], expectedly demonstrated relatively similar significance of binding to both DNA and RNA compared to random locations. Overall, our analysis shows that irrespective of POP-seq protocols, signals are underrepresented in ChIP-seq data while overrepresented in the CLIP-seq data indicating a clear enrichment for RNA binding events compared to publicly available protein–DNA maps. (iii) To estimate the ribosomal protein interactions captured by POP-seq, we compared the POP-seq peaks with publicly available ribo-seq data in K562 cells. Our analysis showed that ~ 20% of the total POP-seq peaks (with peak length ≤ 50 bp) exhibited 50% end-to-end overlap with the ribo-seq peaks indicating that some ribosomal protein–RNA interactions are captured by POP-seq (Table S3).

Comparison of POP-seq data with Formaldehyde and UV crosslinked RBPs reveals high quality of POP-seq peaks. POP-seq is a technique which provides occupancy levels for proteins on a global transcriptome-wide scale. The functionality of thousands of binding sites that are generated resulting from this method could correspond to scores of RBPs and RNP complexes. Since there are no large-scale assays currently available to validate the RNA–protein interaction sites globally, and the low throughput assays will not cover large number of interaction sites, therefore, we employed orthogonal methods using publicly available data for

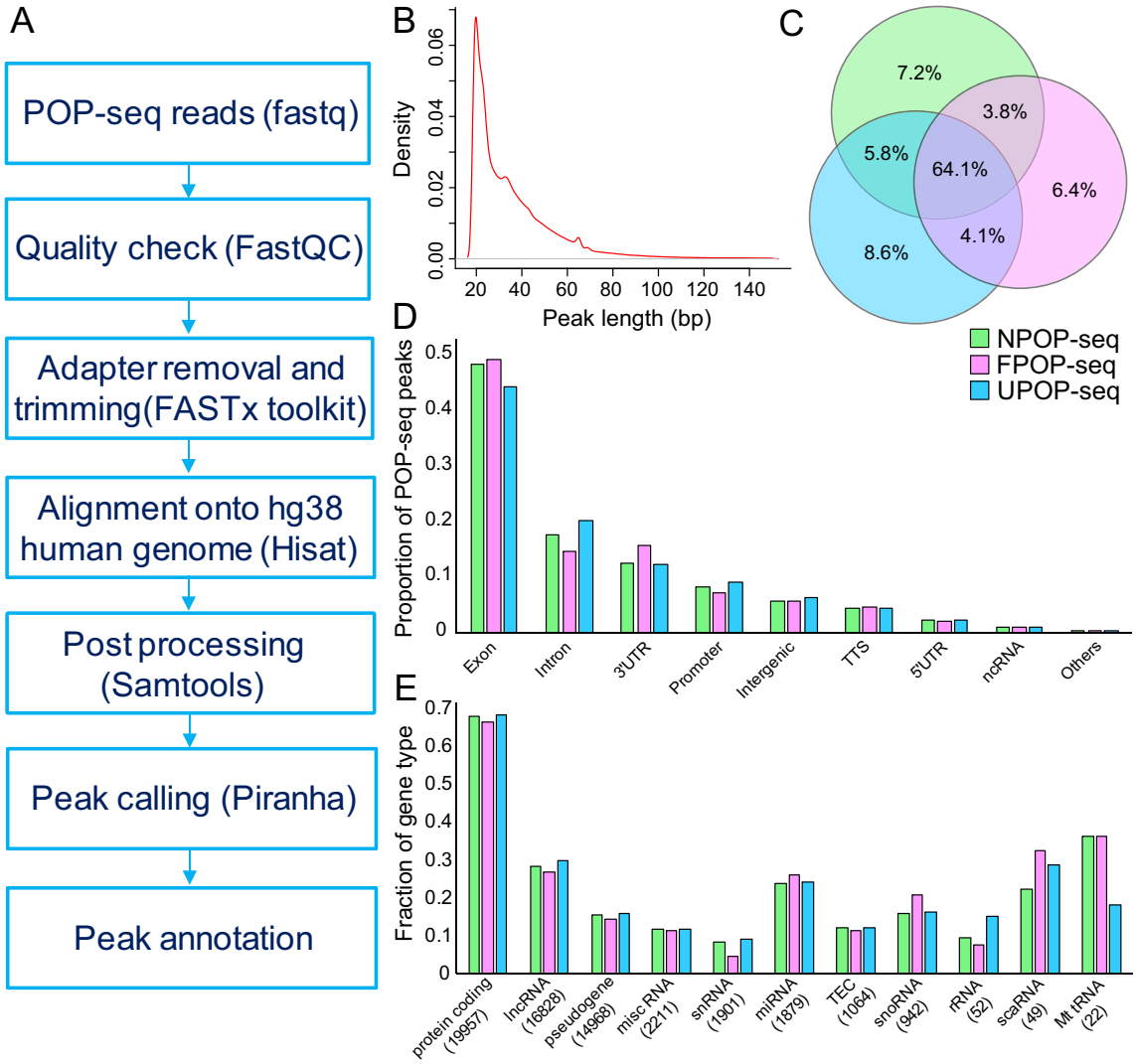

Figure 2. Statistical analysis of POP-seq dataset. (A) workflow for POP-seq data processing and downstream analysis. (**B**) A density plot showing the distribution of POP-seq peaks length (bp). (**C**) Venn diagram showing the overlap of genes exhibiting POP-seq peaks across the protocols. (**D**) Proportion of POP-seq peaks in genomic elements. (**E**) Fraction of gene types captured by POP-seq.

24 formaldehyde crosslinked RBPs[32] with respect to five random non-peak files (See "Methods") and eCLIP profile of 97 RBPs[28] in the K562 cells from ENCODE (see "Methods"). Our analysis indicates a significant enrichment of POP-seq peaks in both CLIP-seq (Fig. 3A for top 24 RBPs, supplementary Figure S3) and fRIP-seq profile (Fig. 3B) of individual RBP's compared to 5 random non-peak profiles. Confusion matrix for this analysis is documented in supplementary Table S1 and S2. Overall, our analysis shows that POP-seq can recover high quality peaks corresponding to specific RBPs identified from individual crosslinking protocols.

CRISPR knock out RNA-sequencing data of RBPs supports the functionality of POP-seq peaks. Development of targeted genome editing using CRISPR has revolutionized the genomic research[33], particularly to understand the molecular mechanism involved in gene regulation and expression[34,35]. A recent study by our research group demonstrated that the functional relevance of protein–RNA interactions can be estimated by the expression of the exons upon perturbation of RBP binding site in their neighborhood using CRISPR-Cas9 system[36]. Therefore, we aimed to interrogate the functional impact of POP-seq captured protein–RNA interactions by its systematic comparison with the eCLIP profile of RBP's, for which knockout data is publicly available in ENCODE project[28,37]. For this analysis, we used two RBPs; (a) DGCR8 (DiGeorge Syndrome Critical Region 8), which is involved in microRNA processing and is implicated in the pathogenesis of cancer[38,39] and (b) IGF2BP1 (insulin like growth factor 2 mRNA binding protein 1) which is a critical post-transcriptional regulator of various mRNA involved in cancer progression[40]. First, we identified the POP-seq peaks from the individual protocol that showed > 50% base-to-base overlap with eCLIP profile of respective RBP (obtained from ENCODE[28]). Next, we extracted the expression levels of exons 'proximal' (< 1000 bp) to overlapped peaks from CRISPR knock out data set (Material and Methods). We observed that the cumulative expression level of 'proxi-

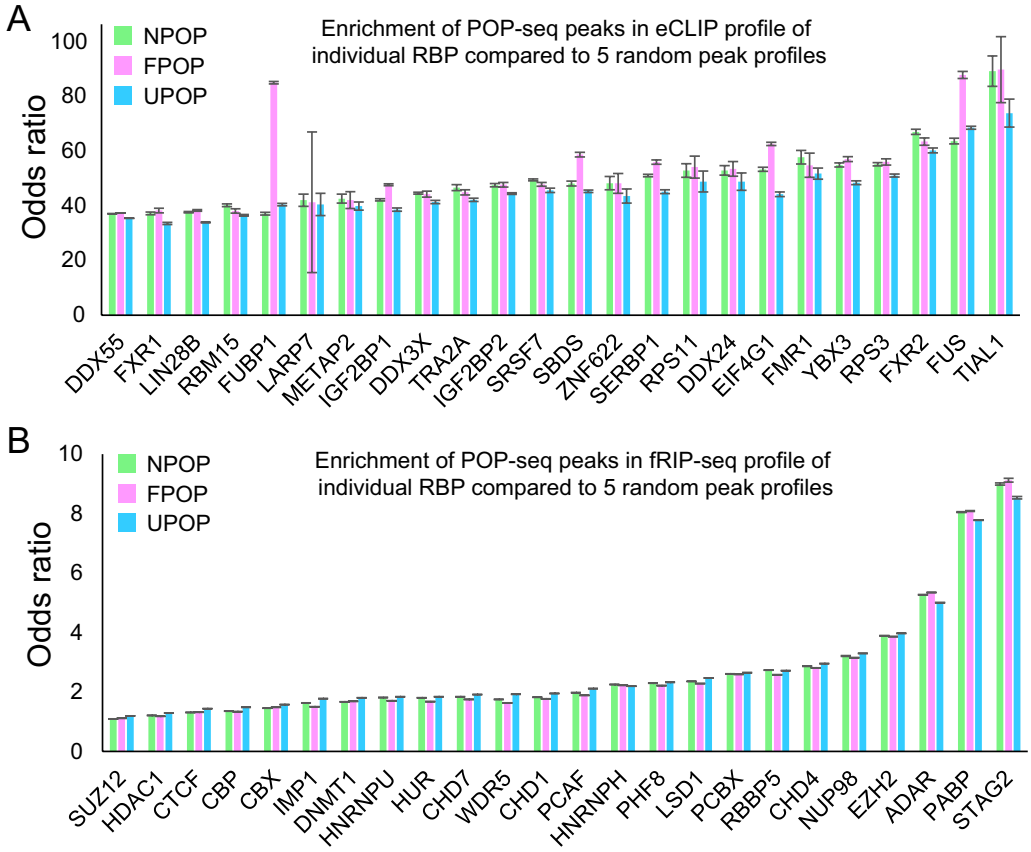

Figure 3. Comparison of POP-seq peaks with RBP centric peaks derived from orthogonal assay. Bar plot showing the enrichment of POP-seq peaks in (**A**) eCLIP and (**B**) fRIP-seq profiles of individual RBP in K562 cells, compared to 5 random peak profiles and statistically tested using Fisher's exact test.

mal' exons was significantly dysregulated with respect to the non-targeting control. We observed that there was a significant reduction in the expression of 'proximal' exons in DGCR8 KO and a significant increase in IGF2BP1 KO with respect to their non-targeting CRISPR control (Fig. 4). However, there could be alternative hypotheses such as the contribution of other binding sites from same or different RBPs that could account for the compensatory effects in expression levels in our CRISPR analysis, which could explain why not all the proximal exon levels are altered. More importantly, RBP binding does not always alter the expression of the target exon/transcript but instead may contribute to editing, structure and localization of bound RNA. However, despite these alternate possibilities, it is promising to observe that the loss of binding sites has a significant impact on the target exons. Overall, this analysis suggests that POP-seq can capture the functionally relevant protein bound sites and indicates that the dysregulation of exons proximal to the functional binding site occur in an RBP dependent manner.

POP-seq supports the protein–RNA binding sites across regulatory gene families. To obtain a detailed perspective of the peaks captured by POP-seq, we examined the occurrence of the protein–RNA binding sites across different regulatory genes, classified as RNA binding protein, ENO1 (Enolase1)[41], lncRNA MALAT1 (metastasis associated lung adenocarcinoma transcript 1)[42] and a transcription factor, Jun[43]. ENO1 is a crucial glycolytic enzyme involved in cell growth and is also reported as an oncogene that promotes metastasis by facilitating cell proliferation in multiple cancers including colorectal, lung, and prostate cancer[44–48]. We observed the abundance of protein bound pockets in the genomic loci of ENO1 across the POP-seq protocols (Fig. 5A). We also investigated the protein–RNA interactions in MALAT 1, a highly conserved lncRNA that governs a variety of functions including regulation of gene expression, alternative splicing, neural development and vascular growth[49–51]. Several studies report the abundant expression of MALAT1 in multiple cancers such as lung cancer, bladder cancer, breast cancer, colorectal cancer and others[49,52–55]. Therefore, identification of regulatory sites targeted by proteins in MALAT1 is crucial for understanding its pathogenesis in cancers. Our data suggest that all the three protocols can capture the abundance of regulatory sites targeted by proteins in the genomic boundary of MALAT1 (Fig. 5B). We observed relatively enhanced signals for protein occupancy sites in MALAT1 locus in UPOP-seq compared to the NPOP-seq and FPOP-seq (Fig. 5B). To our observation, NPOP-seq also captured the physiologically relevant protein–RNA interactions in MALAT1 suggesting its application for unbiased protein occupancy profiling (Fig. 5B).

We further explored the regulatory binding sites in AP1 transcription factor subunit 'Jun' which is a proto-oncogene actively involved in cell proliferation, apoptosis, inflammation and carcinogenesis[56–60]. We found

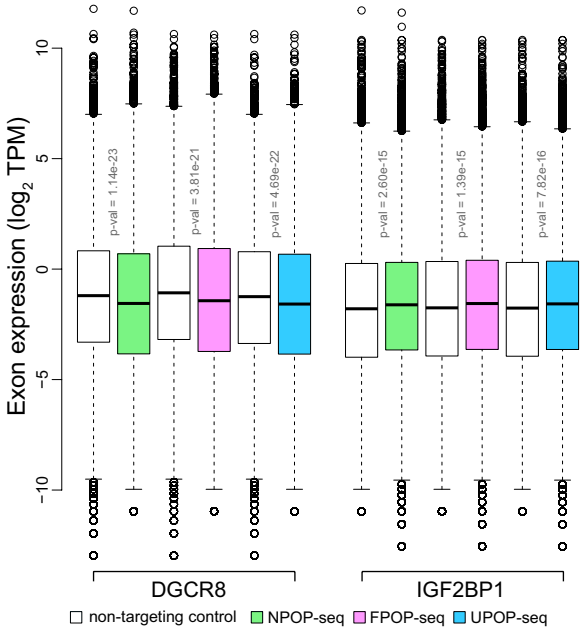

Figure 4. Comparison of POP-seq data with CRISPR knock out RNA-sequencing data of RBPs. Box plot showing the cumulative exon expression levels proximal (< 1000 bp) to POP-seq peaks overlapped (50% end to end peak overlap) with the eCLIP profile of DGCR8 and IGF2BP1 in K562 cells (ENCODE project). The exon expression levels of respective RBP CRISPR Cas9 knock out was compared with respective non-targeting Crispr control and statistically tested using Wilcoxon test.

the occurrence of POP-seq peaks in Jun's genomic boundary across all the three protocols (Fig. 5C) consistent with the observations in ENO1 and MALAT1. Overall, the results demonstrate that POP-seq can capture the protein–RNA interaction sites across the regulatory gene families.

POP seq identifies germline and somatic variants that potentially contribute to altered post transcriptional regulation. Single nucleotide polymorphisms (SNPs) are reported as the most common form of somatic variations and are widely associated with metabolism, cell cycle regulation and DNA mismatch repair[61,62]. In past years, SNPs has emerged as a potential diagnostic biomarker for several cancer types[63–66]. Therefore, it is imperative to investigate the somatic variations arising due to SNPs and their effect on transcriptome wide protein RNA interaction sites.

In order to detect the somatic variations captured by POP-seq, we calculated the proportion of known somatic variations (see "Materials and methods") occurring in the equivalent genomic loci of the peaks. We tested the enrichment of genomic variations from GWAS catalog[67] and Ensembl Variation database[68] (PhenVar, ClinVar and somatic variations) in POP-seq peaks than expected by chance (i.e. 5 random non-peak profiles) using Fisher's exact test. For this analysis, random 'non-peak' files were generated as described previously. Our results indicate a significant enrichment for each SNP cohort with relatively lesser enrichment for GWAS SNP (averaged odds ratio = 1.45, 1.42, 1.35 for NPOP, FPOP and UPOP-seq respectively, p value < 2.2e−16). Similar test for other genomic variations including PhenVar, SomaticVar and ClinVar indicated relatively higher enrichment (Odds ratio ~ 22 averaged across 5 random controls for each cohort, p value < 2.2e−16, Fishers Exact test) in POP-seq peaks compared to non-peaks (Fig. 6A). This observation provides support for the enrichment of both germline and somatic SNPs including those reported with clinical significance to be prominent on protein RNA interaction sites, implying the need for deeper understanding of their functional consequences. Indeed, we identified the occurrence of two clinically relevant genetic risk loci from GWAS; rs45461499 in CDC20 (Cell-division cycle protein 20) and rs7578199 in HDLBP (High Density Lipoprotein Binding Protein) genes that are reported in acute and chronic lymphoblastic leukemia respectively[69,70] (Fig. 6B,C) to harbor protein–RNA interaction sites. Thus, our implementation of POP-seq in K562 cells demonstrate a novel and robust approach to elucidate the occurrence of somatic variants in leukemic patients.

Highly expressed lncRNAs exhibit abundance of POP-seq peaks in K562 cells. Long non-coding RNAs (lncRNAs) have been widely documented with diverse roles in the transcriptional and post-transcriptional regulation of gene expression[71,72]. Aberrant expression of lncRNA is associated with the pathogenesis of various diseases including cancer[73] and have been profoundly recognized as pivotal targets in cancer therapeutics[74]. However, the mechanism underlying lncRNA regulation is not well elucidated[75]. Therefore, we speculated that associating the expression of lncRNA with the occurrence of POP-seq peaks would provide an insight into the transcriptome wide regulation of lncRNAs. We observed that the highly expressed lncRNAs exhibit abundance of regulatory binding sites in K562 cells (Fig. 7A, "Materials and methods"), suggesting that lncRNAs dynami-

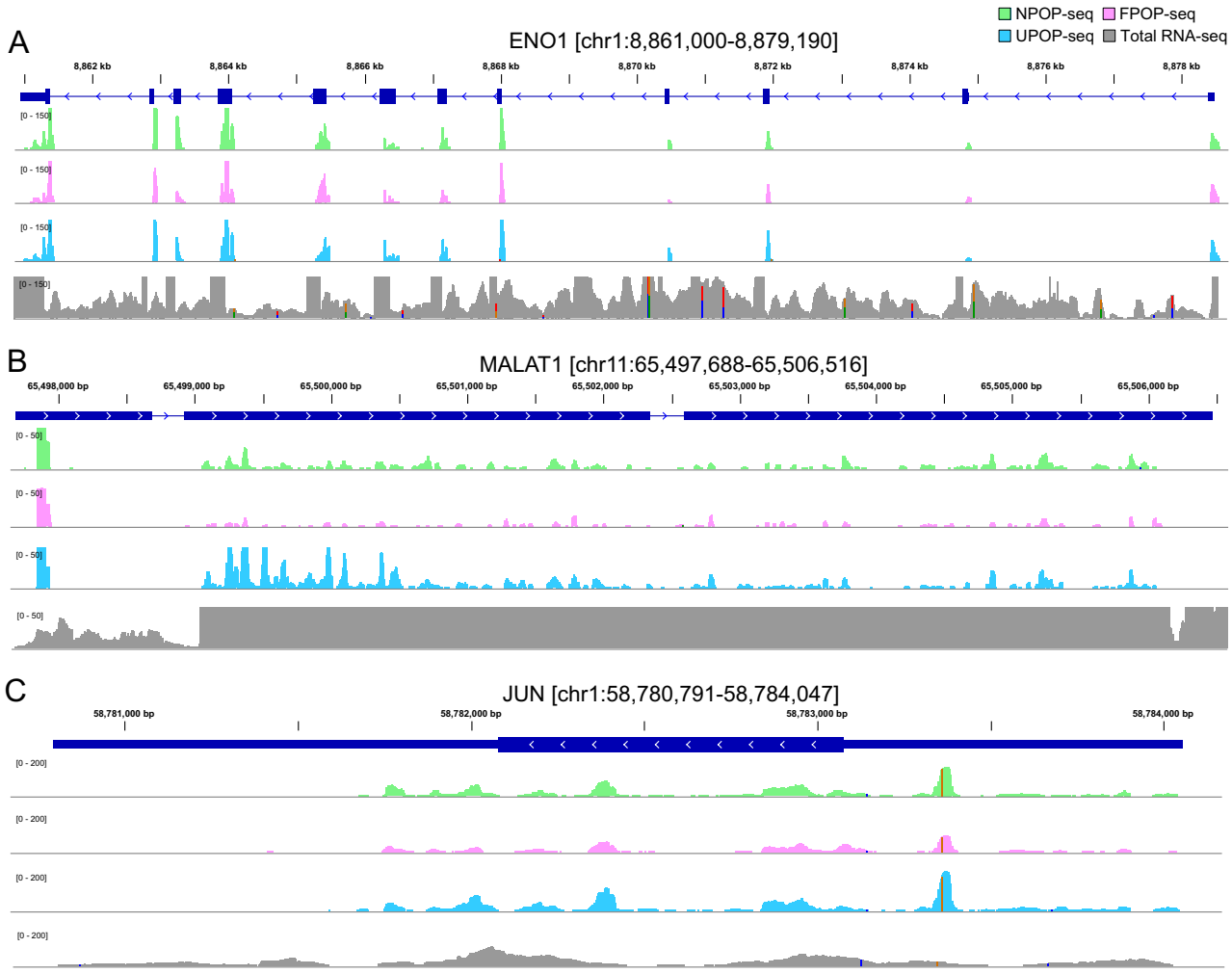

Figure 5. Genomic tracks showing the POP-seq peaks spanning the genomic loci of regulatory genes; (**A**) RBP – Enolase1 (ENO1) (**B**) LncRNA – MALAT1 and (**C**) transcription factor – Jun. Total RNA-seq data (from K562 cells) is included in the track as a background control (Y-axis adjusted to POP-seq scale).

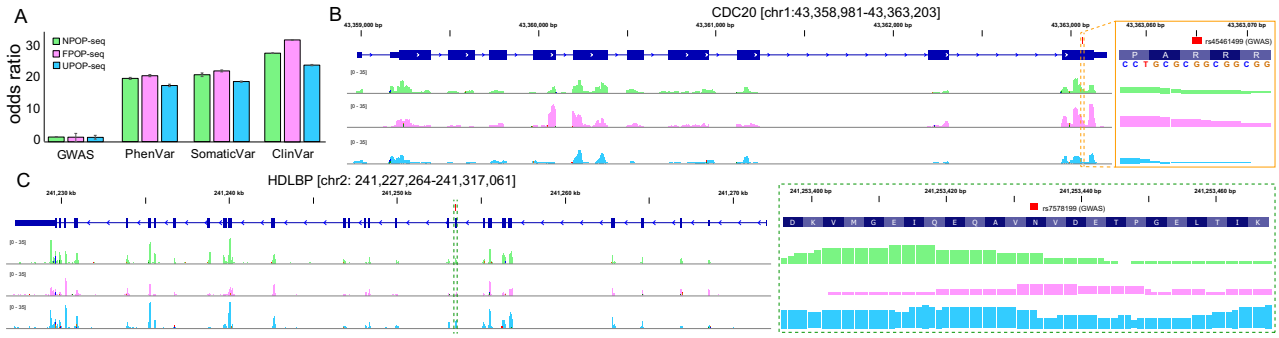

Figure 6. Somatic variation captured in POP-seq peaks. (**A**) Bar plot showing the enrichment (odds ratio) of genomic variations (GWAS, PhenVar, ClinVar and somatic variations) in POP-seq peaks, statistically tested using Fisher's exact test. Genomic tracks showing the POP-seq peaks spanning the genetic risk loci (GWAS SNP) associated to (**B**) acute myeloid leukemia in CDC20 and (**C**) chronic myeloid leukemia in HDLBP.

cally interact with RBPs in pathological conditions. In general, highly expressed RNAs are expected to be more available for binding by proteins and therefore exhibit higher RNA binding events. Therefore, we tested the association between the expression levels (high and low) with the number of POP-seq peaks per unit length for both lncRNA and non-lncRNA genes across the technical replicates. We found that the replicates exhibited a reproducibility in the trend (Figure S4A), irrespective of the POP-seq protocols. To test whether such a trend can

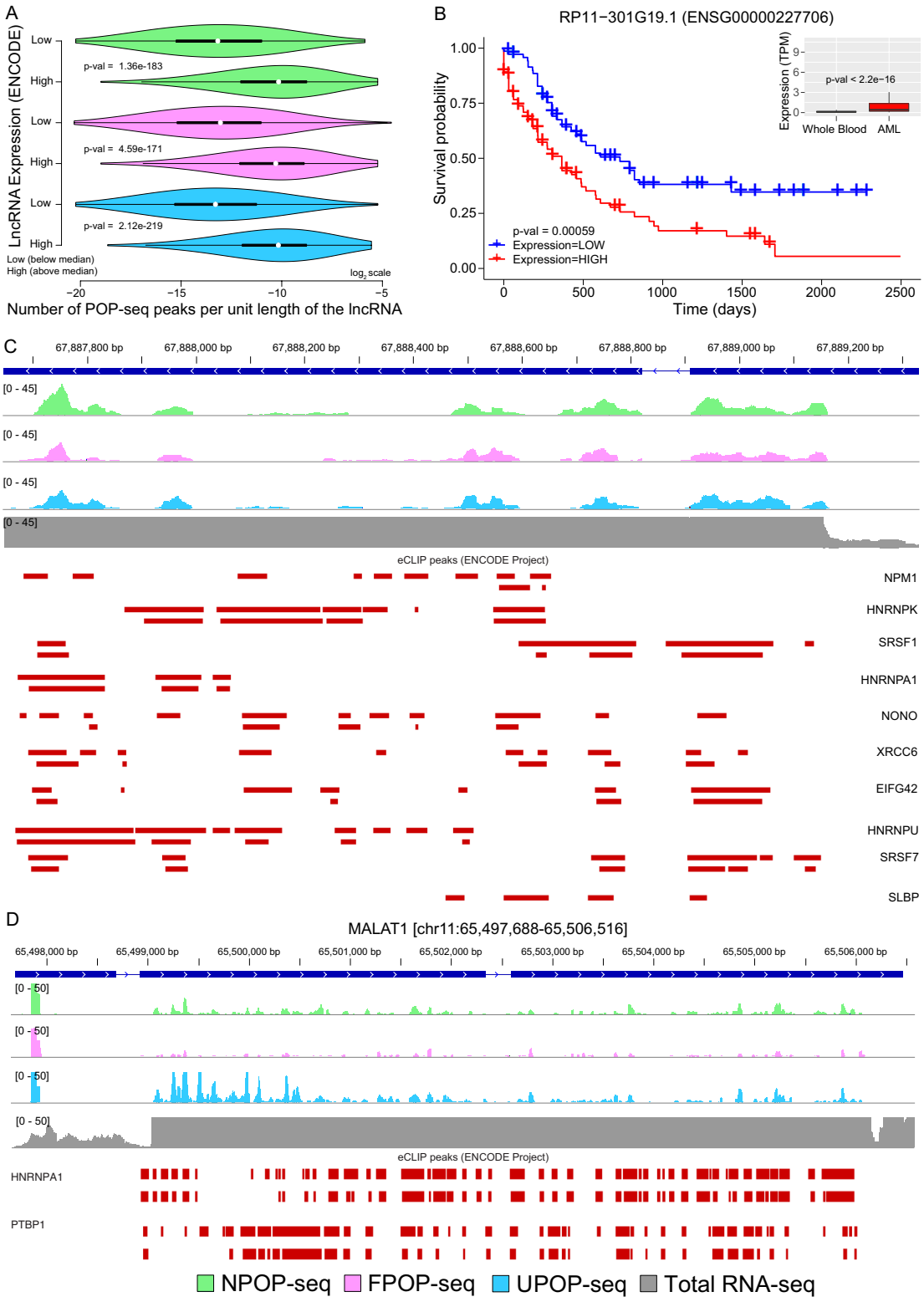

Figure 7. Comparative analysis of POP-seq peaks with lncRNA expression. (**A**) Violin plot showing the number of POP-seq peaks (normalized per unit length of the lncRNA) binned in low and high expression group of lncRNAs in K562 cells. The difference in normalized peak counts between the two groups was statistically tested using Wilcoxon test. (**B**) Kaplan–Meier plot showing the association of lncRNA RP11 – 301G19.1 expression with the survival of AML patients (in days). An inset showing the expression of RP11 – 301G19.1 in whole blood (GTEx) cohort and AML patients (from TCGA) where the difference between the two groups was statistically tested using Wilcoxon test. (**C**) Genomic tracks showing the distribution of POP-seq variant peaks in RP11 – 301G19.1 along with the eCLIP profile of highly expressed RBPs in K562 cells. (**D**) Genomic tracks showing the distribution of POP-seq peaks in MALAT1 along with the eCLIP profile of HNRNPA1 and PTBP1 in K562 cells. Total RNA-seq data (from K562 cells) is included in the track as a background control (Y-axis adjusted to POP-seq coverage scale).

also be observed for non-lncRNA, we carried out the same analysis across the replicates. We observed that there is tendency for even non-lncRNAs to exhibit higher expression with more binding sites, however the trend is not as robust with lower significance (Figure S4B) compared to lncRNA as shown in Figure S4A.

Several studies have reported that the majority of the lncRNAs exhibit a 'sponge effect' to titrate the abundance of regulatory proteins such as RBPs in a cell type specific manner[76–78]. These studies proposed that lncRNA sponges can extensively rewire the post-transcriptional gene regulatory networks by altering the protein–RNA interaction landscape in cell-type and phenotype specific manner. Based on this finding we tested the hypothesis that highly expressed lncRNAs could sponge the regulatory RBPs that would further provide an insight into lncRNA mediated regulation of RBPs in disease context. We uncovered RP11 – 301G19.1, a highly expressed lncRNA in leukemia[79,80], to illustrate the 'sponge effect' in K562 cells. We found an abundance of this lncRNA in AML patients from TCGA[81] compared to its expression level in normal whole blood from GTEx cohort[82] (Fig. 7B, inset). In order to predict whether the expression of this lncRNA contributes to the survival of AML patients, we employed Kaplan–Meier survival analysis[83] for RP11 – 301G19.1. We found that this lncRNA exhibits a significant (False Discover Rate (FDR) < 0.00059) prognostic impact in AML patients (Fig. 7B). Next, we interrogated the regulatory sites captured by POP-seq in the genomic loci of RP11 – 301G19.1 and observed a consistent occurrence of peaks across all POP-seq protocols. Our results demonstrate that majority of the POP-seq peaks overlapped with the eCLIP profile of multiple RBPs (Figure S5) which further supports the "sponge effect" for RP11 – 301G19.1. A subset of highly expressed RBPs in K562 cells illustrates a general agreement of their eCLIP profile with POP-seq peaks in RP11 – 301G19.1 gene (Fig. 7C). To further elaborate the sponge effect, we investigated a well-studied lncRNA MALAT1 which has been shown to interact with numerous RBPs[84–88]. As illustrated in Fig. 7D, we observed that POP-seq captured peaks were sparsely distributed along the length of MALAT1 in contrast to the fairly uniform distribution of total RNA-seq reads showing higher expression of MALAT1 in K562 cells. We also included a track of ENCODE eCLIP peaks for RNA binding proteins HNRNPA1 and PTBP1 with known binding to MALAT1[84] shown in Fig. 7D. The results suggest that HNRNPA1 and PTBP1 are most likely being sponged by highly expressed lncRNA MALAT1 in K562 cells. Additionally, we also observed that other RBPs with transcriptome wide interaction maps available from ENCODE project exhibited several binding sites overlapping with POP-seq peaks along the length of the MALAT1 and NEAT1 lncRNAs (Supplementary Figure S6A and B). Overall, the results suggest that POP-seq can capture the occupancy sites of RBPs on lncRNAs and thus advance our understanding of lncRNA regulation in diseases.

Discussion

Protein–RNA interaction is a vital phenomenon regulating crucial transcriptional and post-transcriptional processes starting from intercalation of the DNA-RNA juncture to RNA metabolism, translation and decay[89,90]. RNA binding proteins have been widely recognized as the key regulatory proteins for these processes[91]. Although the past few decades have seen a surge in the number of methods for capturing protein–RNA interaction sites occupied by RBPs on a transcriptome-wide scale[18,19,21]. Majority of these protocols employ UV or formaldehyde cross linking and poly-A pull followed by sequencing of RNA. However, this excludes their application to non-polyadenylated RNAs and so far, a systematic analysis to depict the effect of crosslinking on the captured interaction sites has not been investigated.

In this study, we present POP-seq to assess the protein bound RNA fragments using trizol based phase separation. POP-seq can uniquely map the protein bound RNA pockets in a transcriptome wide manner. We implemented three versions of POP-seq to generate an unbiased profile of protein–RNA interactions in K562 cells. We demonstrated that POP-seq captured peaks generally agreed with the eCLIP profile of several RBPs. The abundance of protein–RNA interactions captured by POP-seq was mostly observed in exonic followed by intronic regions, with consistent overlap across the POP-seq protocols in K562 cells. Further analysis of publicly available CRISPR KO dataset of RBPs from ENCODE project[92], illustrate that POP-seq can capture the functionally relevant protein bound pockets implying that the dysregulation of exons proximal to the functional binding site occur in RBP dependent manner. POP-seq also enables the identification of clinically relevant somatic variations associated to leukemia. Further, POP-seq provides a comprehensive evidence of the potential protein binding sites in most of the regulatory gene families such as TFs, RBPs and lncRNAs.

Since ribosomal protein complexes constitute the basic translation machinery and hence are expected to be highly abundant in cells. To evaluate the prevalence of ribosomal protein occupied sites in our data, we compared POP-seq peaks with publicly available ribo-seq data[93] in K562 cells. We observed that ~ 20% of the total POP-seq peaks (with peak length ≤ 50 bp) exhibited 50% end-to-end overlap with the ribo-seq peaks and this fraction was significantly reduced with increase in % peak overlap as shown in supplementary Table S3. These observations suggest that a fraction of POP-seq peaks correspond to the regions occupied by ribosomal complexes indicating a potential for further optimization of the protocol to enhance the stringency for selectively capturing sites occupied by regulatory RBPs under physiological conditions.

Interestingly, POP-seq provides evidence for the 'sponge effect' depicted by lncRNA on multiple RBPs thus advancing our understanding of the post-transcriptional regulatory mechanism controlled by lncRNAs interacting with RBPs in cancer. In summary, POP-seq is a cost-effective and robust approach to elucidate the binding sites of proteins in a transcriptome-wide manner. Thus, it should stand as a generic framework for mapping the global protein–RNA interactions, widening the scope and application of this technique to primary tissues for rapid profiling of protein occupancies.

Materials and methods

Cell culture. K562 cells were obtained from the American Type Culture Collection (ATCC). Cells were cultured in Dulbecco's minimal essential medium (DMEM, Gibco) supplemented with 10% heat-inactivated fetal bovine serum (FBS, Atlanta Biologicals) along with 1% antibiotics (penicillin 5000 Units/ml, Streptomycin 5000 µg/ml). All cells were maintained at 37 °C and 5% CO_2 in a humidified incubator and fresh media was replenished every alternate day until confluent.

Crosslinking. Cells were cultured in T-175 flask until a maximum of 90% confluency was reached. A total of 20 million cells per replicate of each sample were used for UV, formaldehyde, and no-crosslinking approaches. Cells were washed twice with 1X PBS, the supernatant was removed by pipetting and cells were resuspended in 1X PBS. For UV crosslinking, cells in PBS suspension were transferred to 100 mm dishes and UV irradiated at 254 nm with 400 mJ/cm^2 dosage (UV Stratalinker 1800). Immediately after crosslinking, cells were collected in 15 ml tubes and pelleted at 1500 rpm for 5 min. Supernatant was discarded and cells were lysed in 1 ml trizol reagent (Life Technologies) by pipetting up and down to obtain a homogenous cell lysate.

For Formaldehyde crosslinking, cells were first washed in 1X PBS twice until all the media is removed. Next, cells were resuspended in 30 ml PBS and crosslinked with 0.5% formaldehyde for 10 min by gentle shaking at room temperature. To stop crosslinking, 1 M Glycine was added to the cell suspension for 5 min by gentle shaking at room temperature. Cells were pelleted down at 1500 rpm for 5 min, lysed in 1 ml trizol and homogenized by pipetting up and down. For non-crosslinked samples, cells were pelleted down, washed in 1X PBS and immediately lysed in trizol for phase separation.

Guanidinium thiocyanate–phenol–chloroform (TRIZOL) extraction. Trizol lysed cells were incubated at room temperature for 5 min to dissociate the weak RNA–protein interactions. Phase separation was achieved by adding 200 µl chloroform and thoroughly mixed by vortexing, followed by incubation at room temperature for 5 min. Samples were then centrifuged at 12,000*g* for 10 min at 4 °C to obtain three phases: aqueous phase (top), interphase (middle) and organic phase (bottom). The aqueous layer was discarded by pipetting and the organic layer was discarded by passing the tip through the interphase leaving behind up to 100 µl of the organic layer. The interphase was resolubilized in trizol followed by phase separation with chloroform three times. After the third phase separation, the interphase was precipitated by adding 1 ml methanol, spun down to remove supernatant containing methanol.

POP-seq strategy. Following the trizol lysis and phase separation, POP-seq was implemented on the three versions in replicates. Interphase pellet was subjected to RNase A/T1 (Thermo Scientific) degradation in RNase buffer (10 mM Tris–HCl, pH 7.5, 300 mM NaCl and 5 mM EDTA, pH 7.5). 2 µg of RNase A/T1 mix was added to the interphase pellet, mixed by pipetting, and incubated at 37 °C for 1 h. Interphase-RNase mixture was resolubilized in trizol reagent to recover the RNA–protein complexes. The aqueous and organic layers were discarded as described previously. Interphase pellet was precipitated in 1 ml methanol. Next, the interphase was mixed with Proteinase K (Ambion) in appropriate buffer (0.1 M NaCl, 10 mM Tris–HCl, pH 8, 1 mM EDTA, 0.5% SDS and 200 µg/ml proteinase K) and incubated at 50 °C for 2 h. After proteinase K digestion, free RNA was recovered from the aqueous layer by trizol extraction as described previously.

Purified RNA concentration was estimated using Nanodrop and up to 1 µg of RNA was incubated with DNase I, 1U (Thermo Scientific) at 37 °C for 30 min to remove any traces of DNA contamination. 1 µl of 50 mM EDTA was added to the reaction mixture and incubated at 65 °C for 10 min to terminate the reaction. RNA was purified using trizol extraction from the aqueous layer. At this point, r-RNA depletion was performed with 1 µg input RNA using Ribo-cop kit (Lexogen) as per manufacturer instructions. Further, the ends of r-RNA depleted RNA were modified by treating with Calf intestine alkaline phosphatase (CIAP, Invitrogen) and T4 polynucleotide kinase (T4 PNK, Thermo Scientific) as per manufacturer protocol. The end modification enabled the library preparation of these RNA fragments.

RNA integrity, library preparation and sequencing. RNA purity and concentration were assessed at each step using Nanodrop, based on the absorbance ratio 260/280 > 2. RNA integrity was evaluated using Agilent 2100 bioanalyzer system. At least 50 ng of r-RNA depleted RNA was used to generate sequencing libraries using the True-seq small RNA library prep kit (Illumina). All libraries were barcoded and sequenced in parallel on a Next-seq platform for 400 million reads to obtain 75 bp single end reads.

Data processing and statistical analysis of POP-seq peaks. We implemented NGS data processing pipeline to facilitate the analysis of the POP-seq data as shown in Fig. 2A. Firstly, we investigated the quality of sequenced reads using FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and deployed FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) for removal of adapters and low quality read fragments wherever applicable. Next, we aligned the high quality reads onto human reference genome (GRCh38.p12) using HISAT[94] followed by post processing using Samtools[95]. To ensure the reproducibility between the replicates, we compared the aligned reads per 10 kb genome using deepTools 'plot correlation' module[26]. We employed Piranha[96] for peak calling with default parameters and obtained the resulting POP-seq peaks in bed format. Source code of the POP-seq data processing pipeline is accessible at GitHub (https://github.com/Janga-Lab/POP-seq). We merged the replicate bed files of respective POP-seq protocols and used several tools such as bedtools[97], HOMER[29] (annotatePeaks.pl), and R (https://www.r-project.org/) for annotation, statistical testing and other downstream analysis. We also downloaded the publicly available formaldehyde RNA ImmunoPrecipi-

tation (fRIP-seq) data[32] for 24 RBPs from GSE67963 and raw ribo-seq data[93] from GSE125218, both generated in K562 cells and processed them using the same pipeline and identified the peaks. Next, we computed the fraction of POP-seq peaks overlapping with the identified peaks from both the datasets independently using bedtools[97].

Similarly, we downloaded the eCLIP[98] profiles of 97 RBPs from ENCODE project[28] and concatenated the unique coordinates into a bed file. Resulting coordinates of the binding sites of RBPs were compared with the POP-seq peaks using bedtools[97]. All POP-seq data have been deposited under Gene Expression Omnibus (GEO) accession number GSE142460. We also downloaded the raw FATSQ reads of total RNA (accession no. ENCLB822JYE from ENCODE project) and processed using the standard NGS pipeline as described in below section.

Comparison of POP-seq peaks with publicly available CLIP-seq and ChIP-seq data. We downloaded the ChIP-seq and CLIP-seq profile of available proteins for K562 cells from ENCODE project[28]. We computed the overlap of POP-seq peaks with CLIP-seq peaks and ChIP-seq peaks using bedtools[97]. This dataset also includes 18 proteins for which both ChIP and CLIP-seq data is available for K562 cells. We generated 5 random peak profiles for each POP-seq protocols using bedtools 'shuffle' function. Each random peak profile contains equal but "non-peak" locations (with consistent peak width distribution), within the gene boundary. Then, we compared the POP-seq derived peaks (and the random peaks separately) with the ChIP and CLIP-seq profile of 18 proteins. We employed Fisher's exact test to estimate the level of significance for the enrichment of POP-seq in protein bound DNA(ChIP-seq)/RNA(CLIP-seq) locations compared to non-peaks.

Integrated data analysis of POP-seq peaks with the CRISPR knock out studies. CRISPR/Cas9 Knock Down (KD) followed by expression profiling on several RBPs have been conducted as part of the ENCODE project[28] to facilitate the understanding of downstream biological processes associated to loss of function of the respective RBP. We downloaded the raw RNA-sequencing dataset of CRISPR experiments of DGCR8 (n = 6), IGF2BP1 (n = 2) where gRNAs were used to deplete the functional form of RBPs and their non-specific CRISPR control (n = 8) in K562 cells[92]. We processed the raw sequencing reads using standard NGS data analysis pipeline (as described previously). Briefly, we filtered the low-quality reads (Phred Score < 30) using FASTQC tool and aligned them onto human reference genome (hg38) using HISAT[94]. After post processing (using samtools), we used StringTie[99] to quantify the expression levels in Transcripts Per Million (TPM) reads for all the genes annotated in human genome (hg38). Thereafter, we used an ad-hoc script to calculate the exon levels from the resulting gtf files of StringTie and converted the resulting files into an exon expression matrix.

To further investigate the functional relevance of protein–RNA interactions, we identified the POP-seq peaks from the individual protocols that exhibited at least 50% overlap with an eCLIP profile of the respective RBP (as described previously). Next, we extracted the expression levels of exons, proximal (< 1000 bp) to overlapped peaks, from exon expression matrix of the knock down experiments. We compared the distribution of the expression levels of these proximal exons in non-targeting control to that in respective KD and statistically examined the condition specific expression differences using Wilcoxon test[100].

Identification of somatic variants in protein–RNA interacting sites. Several studies suggest that single nucleotide variations (SNVs) play an important role in gene regulation via riboSNitches'[101] i.e. by altering RNA secondary structure or TAM (Transcript associated mutation)[102] that further contribute to transcriptome complexity in higher eukaryotes. Therefore, it is imperative to investigate the genomic variations occurring in protein–RNA interaction sites identified by POP-seq protocols. Hence, we downloaded the somatic variants reported in the GWAS catalog[67] and Ensembl Variation database[68] including phenotype and clinically associated somatic variations (ftp://ftp.ensembl.org/pub/release-97/variation/vcf/homo_sapiens/). In order to detect the somatic variations captured by POP-seq, we tested the enrichment of genomic variations from GWAS catalog[67] and Ensembl Variation database[68] (PhenVar, ClinVar and somatic variations) in POP-seq peaks than expected by chance (i.e. 5 random non-peak profiles) using Fisher's exact test. For this analysis, random 'non-peak' files were generated as described previously. To gain disease specific understanding of the role of SNPs in impacting protein–RNA interactions, we also investigated the POP-seq peaks overlapping with SNPs associated with leukemia (from GWAS) and generated genomic tracks in Integrative Genomics Viewer (IGV)[103] for CDC20 and HDLBP genes.

Comparative analysis of POP-seq peaks across lncRNAs and their association with lncRNA expression. We mapped the protein–RNA interaction sites captured by POP-seq protocols onto known lncRNAs using bedtools. For each lncRNA, the number of POP-seq peaks normalized by respective gene length was calculated. To obtain the expression levels, we processed the raw RNA sequencing dataset (paired end reads, n = 5, in replicates) of K562 cells from ENCODE using a standard NGS data analysis pipeline described earlier and generated a gene expression matrix. TPM values of known lncRNAs[104] were extracted from the resulting matrix and averaged for each lncRNA across the replicates. Further, we binned all the expressed lncRNAs into two groups based on their median TPM value. We compared the number of POP-seq peaks (normalized per unit length of the lncRNA) mapped to the two groups of lncRNAs categorized based on low and high median expression in K562 cells. The difference in normalized peak counts between the two groups was statistically tested using Wilcoxon test[100].

Additionally, we downloaded the raw RNA sequencing dataset of 'whole blood' cohort from 141 individuals from the GTEx[105] and 174 AML patient samples from The Cancer Genome Atlas (TCGA)[81]. We processed the dataset using the NGS data processing pipeline, to generate expression levels for all human genes annotated in the human genome (hg38). We extracted the expression level of lncRNA—RP11 – 301G19.1 (ENSG00000227706)

from the two groups; AML and GTEx 'whole blood'. The difference in expression levels was statistically examined between the two groups using Wilcoxon test. We also calculated the patient's survival over time using the expression levels of this lncRNA in AML patients using the Kaplan–Meier method implemented in 'The survival', an R package[83]. We generated a genomic track for this lncRNA using IGV and illustrated the regulatory sites identified by POP-seq. We re-investigated the genomic coordinates of this gene in SliceIt[36] and added a panel to enable all the possible regulatory sites captured by eCLIP of RBPs in ENCODE project[28].

Data availability

All POP-seq data have been deposited under Gene Expression Omnibus (GEO) accession number GSE142460. Source code of the POP-seq data processing pipeline and genome track browser shots were made available in GitHub (https://github.com/Janga-Lab/POP-seq).

Received: 15 June 2020; Accepted: 28 December 2020
Published online: 13 January 2021

References

1. Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341. https://doi.org/10.1038/nrm.2017.130 (2018).
2. Halbeisen, R. E., Galgano, A., Scherrer, T. & Gerber, A. P. Post-transcriptional gene regulation: from genome-wide studies to principles. *Cell Mol. Life Sci.* **65**, 798–813. https://doi.org/10.1007/s00018-007-7447-6 (2008).
3. Lukong, K. E., Chang, K. W., Khandjian, E. W. & Richard, S. RNA-binding proteins in human genetic disease. *Trends Genet.* **24**, 416–425. https://doi.org/10.1016/j.tig.2008.05.004 (2008).
4. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet.* **29**, 318–327. https://doi.org/10.1016/j.tig.2013.01.004 (2013).
5. Neelamraju, Y., Hashemikhabir, S. & Janga, S. C. The human RBPome: from genes and proteins to human disease. *J. Proteom.* **127**, 61–70. https://doi.org/10.1016/j.jprot.2015.04.031 (2015).
6. Kechavarzi, B. & Janga, S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol.* **15**, R14. https://doi.org/10.1186/gb-2014-15-1-r14 (2014).
7. Ule, J., Jensen, K., Mele, A. & Darnell, R. B. CLIP: a method for identifying protein–RNA interaction sites in living cells. *Methods (San Diego, Calif.)* **37**, 376–386. https://doi.org/10.1016/j.ymeth.2005.07.018 (2005).
8. Lee, F. C. Y. & Ule, J. Advances in CLIP technologies for studies of protein–RNA interactions. *Mol. Cell* **69**, 354–369. https://doi.org/10.1016/j.molcel.2018.01.005 (2018).
9. Darnell, R. B. HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA* **1**, 266–286. https://doi.org/10.1002/wrna.31 (2010).
10. Spitzer, J. *et al.* PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods Enzymol.* **539**, 113–161. https://doi.org/10.1016/B978-0-12-420120-0.00008-6 (2014).
11. Huppertz, I. *et al.* iCLIP: protein–RNA interactions at nucleotide resolution. *Methods (San Diego, Calif.)* **65**, 274–287. https://doi.org/10.1016/j.ymeth.2013.10.011 (2014).
12. Wheeler, E. C., Van Nostrand, E. L. & Yeo, G. W. Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. *Wiley Interdiscip. Rev. RNA*. https://doi.org/10.1002/wrna.1436 (2018).
13. Meisenheimer, K. M. & Koch, T. H. Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. Biol.* **32**, 101–140. https://doi.org/10.3109/10409239709108550 (1997).
14. Wurtmann, E. J. & Wolin, S. L. RNA under attack: cellular handling of RNA damage. *Crit. Rev. Biochem. Mol. Biol.* **44**, 34–49. https://doi.org/10.1080/10409230802594043 (2009).
15. Nicholson, C. O., Friedersdorf, M. B., Bisogno, L. S. & Keene, J. D. DO-RIP-seq to quantify RNA binding sites transcriptome-wide. *Methods (San Diego, Calif.)* **118–119**, 16–23. https://doi.org/10.1016/j.ymeth.2016.11.004 (2017).
16. Zambelli, F. & Pavesi, G. RIP-Seq data analysis to determine RNA–protein associations. *Methods Mol. Biol.* **1269**, 293–303. https://doi.org/10.1007/978-1-4939-2291-8_18 (2015).
17. Jayaseelan, S., Doyle, F. & Tenenbaum, S. A. Profiling post-transcriptionally networked mRNA subsets using RIP-Chip and RIP-Seq. *Methods (San Diego, Calif.)* **67**, 13–19. https://doi.org/10.1016/j.ymeth.2013.11.001 (2014).
18. Schueler, M. *et al.* Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biol.* **15**, R15. https://doi.org/10.1186/gb-2014-15-1-r15 (2014).
19. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46**, 674–690. https://doi.org/10.1016/j.molcel.2012.05.021 (2012).
20. Singh, G., Ricci, E. P. & Moore, M. J. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. *Methods (San Diego, Calif.)* **65**, 320–332. https://doi.org/10.1016/j.ymeth.2013.09.013 (2014).
21. Foley, S. W. & Gregory, B. D. Protein interaction profile sequencing (PIP-seq). *Curr. Protoc. Mol. Biol.* **116**, 27–25. https://doi.org/10.1002/cpmb.21 (2016).
22. Popova, V. V., Kurshakova, M. M. & Kopytova, D. V. Methods to study the RNA–protein interactions. *Mol. Biol. (Mosk)* **49**, 472–481. https://doi.org/10.7868/S0026898415020111 (2015).
23. Queiroz, R. M. L. *et al.* Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat. Biotechnol.* **37**, 169–178. https://doi.org/10.1038/s41587-018-0001-2 (2019).
24. Trendel, J. *et al.* The human RNA-binding proteome and its dynamics during translational arrest. *Cell* **176**, 391–403. https://doi.org/10.1016/j.cell.2018.11.004 (2019).
25. Urdaneta, E. C. *et al.* Purification of cross-linked RNA–protein complexes by phenol–toluol extraction. *Nat. Commun.* **10**, 990. https://doi.org/10.1038/s41467-019-08942-3 (2019).
26. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-165. https://doi.org/10.1093/nar/gkw257 (2016).
27. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514. https://doi.org/10.1038/nmeth.3810 (2016).
28. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726-732. https://doi.org/10.1093/nar/gkv1160 (2016).
29. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 (2010).
30. Yang, Y. S., Yang, M. C., Tucker, P. W. & Capra, J. D. NonO enhances the association of many DNA-binding proteins to their targets. *Nucleic Acids Res.* **25**, 2284–2292. https://doi.org/10.1093/nar/25.12.2284 (1997).

31. Hudson, W. H. & Ortlund, E. A. The structure, function and evolution of proteins that bind DNA and RNA. *Nat. Rev. Mol. Cell Biol.* **15**, 749–760. https://doi.org/10.1038/nrm3884 (2014).
32. Hendrickson, D. G., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biol.* **17**, 28. https://doi.org/10.1186/s13059-016-0878-3 (2016).
33. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911. https://doi.org/10.1038/s41467-018-04252-2 (2018).
34. Pham, H., Kearns, N. A. & Maehr, R. Transcriptional regulation with CRISPR/Cas9 effectors in mammalian cells. *Methods Mol. Biol.* **1358**, 43–57. https://doi.org/10.1007/978-1-4939-3067-8_3 (2016).
35. Pickar-Oliver, A. *et al.* Targeted transcriptional modulation with type I CRISPR-Cas systems in human cells. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-019-0235-7 (2019).
36. Vemuri, S. *et al.* SliceIt: a genome-wide resource and visualization tool to design CRISPR/Cas9 screens for editing protein–RNA interaction sites in the human genome. *Methods* https://doi.org/10.1016/j.ymeth.2019.09.004 (2019).
37. Consortium, T. E. P. *RNA-seq Profiling of CRISPR/Cas9 Based Knock Outs of RNA-Binding Proteins in Human Cell Line K562*. https://www.encodeproject.org/search/?type=Experiment&assay_title=CRISPR+RNA-seq&replicates.library.biosample.life_stage=adult. (2017).
38. Wen, J., Lv, Z., Ding, H., Fang, X. & Sun, M. Association of miRNA biosynthesis genes DROSHA and DGCR8 polymorphisms with cancer susceptibility: a systematic review and meta-analysis. *Biosci. Rep.* https://doi.org/10.1042/BSR20180072 (2018).
39. Guo, Y. *et al.* Silencing the double-stranded RNA binding protein DGCR8 inhibits ovarian cancer cell proliferation, migration, and invasion. *Pharm. Res.* **32**, 769–778. https://doi.org/10.1007/s11095-013-1219-9 (2015).
40. Huang, X. *et al.* Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) in cancer. *J. Hematol. Oncol.* **11**, 88. https://doi.org/10.1186/s13045-018-0628-y (2018).
41. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406. https://doi.org/10.1016/j.cell.2012.04.031 (2012).
42. Zhao, M. *et al.* MALAT1: a long non-coding RNA highly associated with human cancers. *Oncol. Lett.* **16**, 19–26. https://doi.org/10.3892/ol.2018.8613 (2018).
43. Rahmsdorf, H. J. Jun: transcription factor and oncoprotein. *J. Mol. Med. (Berl.)* **74**, 725–747. https://doi.org/10.1007/s001090050077 (1996).
44. Strein, C., Alleaume, A. M., Rothbauer, U., Hentze, M. W. & Castello, A. A versatile assay for RNA-binding proteins in living cells. *RNA* **20**, 721–731. https://doi.org/10.1261/rna.043562.113 (2014).
45. Yu, L. *et al.* Estrogen promotes prostate cancer cell migration via paracrine release of ENO1 from stromal cells. *Mol. Endocrinol.* **26**, 1521–1530. https://doi.org/10.1210/me.2012-1006 (2012).
46. Ji, H. *et al.* Progress in the biological function of alpha-enolase. *Anim. Nutr.* **2**, 12–17. https://doi.org/10.1016/j.aninu.2016.02.005 (2016).
47. Castello, A., Hentze, M. W. & Preiss, T. Metabolic enzymes enjoying new partnerships as RNA-binding proteins. *Trends Endocrinol. Metab.* **26**, 746–757. https://doi.org/10.1016/j.tem.2015.09.012 (2015).
48. Zhan, P. *et al.* FBXW7 negatively regulates ENO1 expression and function in colorectal cancer. *Lab. Invest.* **95**, 995–1004. https://doi.org/10.1038/labinvest.2015.71 (2015).
49. Zhang, X., Hamblin, M. H. & Yin, K. J. The long noncoding RNA Malat 1: its physiological and pathophysiological functions. *RNA Biol.* **14**, 1705–1714. https://doi.org/10.1080/15476286.2017.1358347 (2017).
50. Li, Z. X. *et al.* MALAT1: a potential biomarker in cancer. *Cancer Manag. Res.* **10**, 6757–6768. https://doi.org/10.2147/CMAR.S169406 (2018).
51. Sun, Y. & Ma, L. New insights into long non-coding RNA MALAT1 in cancer and metastasis. *Cancers (Basel)* https://doi.org/10.3390/cancers11020216 (2019).
52. Gutschner, T., Hammerle, M. & Diederichs, S. MALAT1—a paradigm for long noncoding RNA function in cancer. *J. Mol. Med. (Berl.)* **91**, 791–801. https://doi.org/10.1007/s00109-013-1028-y (2013).
53. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031–8041. https://doi.org/10.1038/sj.onc.1206928 (2003).
54. Zheng, H. T. *et al.* High expression of lncRNA MALAT1 suggests a biomarker of poor prognosis in colorectal cancer. *Int. J. Clin. Exp. Pathol.* **7**, 3174–3181 (2014).
55. Zhao, Z., Chen, C., Liu, Y. & Wu, C. 17beta-Estradiol treatment inhibits breast cell proliferation, migration and invasion by decreasing MALAT-1 RNA level. *Biochem. Biophys. Res. Commun.* **445**, 388–393. https://doi.org/10.1016/j.bbrc.2014.02.006 (2014).
56. Eckhoff, K. *et al.* The prognostic significance of Jun transcription factors in ovarian cancer. *J. Cancer Res. Clin. Oncol.* **139**, 1673–1680. https://doi.org/10.1007/s00432-013-1489-y (2013).
57. Vogt, P. K. & Bos, T. J. jun: oncogene and transcription factor. *Adv. Cancer Res.* **55**, 1–35. https://doi.org/10.1016/s0065-230x(08)60466-2 (1990).
58. Hartl, M., Bader, A. G. & Bister, K. Molecular targets of the oncogenic transcription factor jun. *Curr. Cancer Drug Targets* **3**, 41–55. https://doi.org/10.2174/1568009033333781 (2003).
59. Kappelmann-Fenzl, M. *et al.* C-Jun drives melanoma progression in PTEN wild type melanoma cells. *Cell Death Dis.* **10**, 584. https://doi.org/10.1038/s41419-019-1821-9 (2019).
60. Srivastava, M. *et al.* Inhibition of the TIRAP-c-Jun interaction as a therapeutic strategy for AP1-mediated inflammatory responses. *Int. Immunopharmacol.* **71**, 188–197. https://doi.org/10.1016/j.intimp.2019.03.031 (2019).
61. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74. https://doi.org/10.1038/nature15393 (2015).
62. Deng, N., Zhou, H., Fan, H. & Yuan, Y. Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget* **8**, 110635–110649. https://doi.org/10.18632/oncotarget.22372 (2017).
63. Poduri, A., Evrony, G. D., Cai, X. & Walsh, C. A. Somatic mutation, genomic variation, and neurological disease. *Science* **341**, 1237758. https://doi.org/10.1126/science.1237758 (2013).
64. Iourov, I. Y., Vorsanova, S. G. & Yurov, Y. B. Somatic genome variations in health and disease. *Curr. Genom.* **11**, 387–396. https://doi.org/10.2174/138920210793176065 (2010).
65. Erickson, R. P. Somatic gene mutation and human disease other than cancer. *Mutat. Res.* **543**, 125–136. https://doi.org/10.1016/s1383-5742(03)00010-3 (2003).
66. Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–718. https://doi.org/10.1038/nrg3539 (2013).
67. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901. https://doi.org/10.1093/nar/gkw1133 (2017).
68. Hunt, S. E. *et al.* Ensembl variation resources. *Database* https://doi.org/10.1093/database/bay119 (2018).
69. Liu, C. *et al.* Clinical and genetic risk factors for acute pancreatitis in patients with acute lymphoblastic leukemia. *J. Clin. Oncol.* **34**, 2133–2140. https://doi.org/10.1200/JCO.2015.64.5812 (2016).
70. Berndt, S. I. *et al.* Meta-analysis of genome-wide association studies discovers multiple loci for chronic lymphocytic leukemia. *Nat. Commun.* **7**, 10933. https://doi.org/10.1038/ncomms10933 (2016).

71. Fernandes, J. C. R., Acuna, S. M., Aoki, J. I., Floeter-Winter, L. M. & Muxel, S. M. Long non-coding RNAs in the regulation of gene expression: physiology and disease. *Noncoding RNA* https://doi.org/10.3390/ncrna5010017 (2019).
72. Melissari, M. T. & Grote, P. Roles for long non-coding RNAs in physiology and disease. *Pflugers Arch.* **468**, 945–958. https://doi.org/10.1007/s00424-016-1804-y (2016).
73. Batista, P. J. & Chang, H. Y. Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**, 1298–1307. https://doi.org/10.1016/j.cell.2013.02.012 (2013).
74. Jiang, M. C., Ni, J. J., Cui, W. Y., Wang, B. Y. & Zhuo, W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am. J. Cancer Res.* **9**, 1354–1366 (2019).
75. Yang, Y. X. *et al.* Long non-coding RNA p10247, high expressed in breast cancer (lncRNA-BCHE), is correlated with metastasis. *Clin. Exp. Metastasis* **35**, 109–121. https://doi.org/10.1007/s10585-018-9901-2 (2018).
76. Porto, F. W., Daulatabad, S. V. & Janga, S. C. Long non-coding RNA expression levels modulate cell-type-specific splicing patterns by altering their interaction landscape with RNA–binding proteins. *Genes (Basel)* https://doi.org/10.3390/genes10080593 (2019).
77. HafezQorani, S., Houdjedj, A., Arici, M., Said, A. & Kazan, H. RBPSponge: genome-wide identification of lncRNAs that sponge RBPs. *Bioinformatics* **35**, 4760–4763. https://doi.org/10.1093/bioinformatics/btz448 (2019).
78. Kim, J. *et al.* LncRNA OIP5-AS1/cyrano sponges RNA-binding protein HuR. *Nucleic Acids Res.* **44**, 2378–2392. https://doi.org/10.1093/nar/gkw017 (2016).
79. Thiel, D. *et al.* Identifying lncRNA-mediated regulatory modules via ChIA-PET network analysis. *BMC Bioinform.* **20**, 292. https://doi.org/10.1186/s12859-019-2900-8 (2019).
80. Casero, D. *et al.* Long non-coding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. *Nat. Immunol.* **16**, 1282–1291. https://doi.org/10.1038/ni.3299 (2015).
81. Tomczak, K., Czerwinska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* **19**, A68-77. https://doi.org/10.5114/wo.2014.47136 (2015).
82. Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660. https://doi.org/10.1126/science.1262110 (2015).
83. Terry M. Therneau, P. M. G. *Modeling Survival Data: Extending the Cox Model.* (2000).
84. Scherer, M., Levin, M., Butter, F. & Scheibe, M. Quantitative proteomics to identify nuclear RNA-binding proteins of Malat1. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21031166 (2020).
85. Jonas, K., Calin, G. A. & Pichler, M. RNA-binding proteins as important regulators of long non-coding RNAs in cancer. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21082969 (2020).
86. Lubelsky, Y. & Ulitsky, I. Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* **555**, 107–111. https://doi.org/10.1038/nature25757 (2018).
87. Nguyen, T. M. *et al.* The SINEB1 element in the long non-coding RNA Malat1 is necessary for TDP-43 proteostasis. *Nucleic Acids Res.* **48**, 2621–2642. https://doi.org/10.1093/nar/gkz1176 (2020).
88. Tripathi, V. *et al.* The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Mol. Cell* **39**, 925–938. https://doi.org/10.1016/j.molcel.2010.08.011 (2010).
89. Glisovic, T., Bachorik, J. L., Yong, J. & Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.* **582**, 1977–1986. https://doi.org/10.1016/j.febslet.2008.03.004 (2008).
90. Zanzoni, A., Spinelli, L., Ribeiro, D. M., Tartaglia, G. G. & Brun, C. Post-transcriptional regulatory patterns revealed by protein–RNA interactions. *Sci. Rep.* **9**, 4302. https://doi.org/10.1038/s41598-019-40939-2 (2019).
91. Castello, A. *et al.* Comprehensive identification of RNA-binding domains in human cells. *Mol. Cell* **63**, 696–710. https://doi.org/10.1016/j.molcel.2016.06.029 (2016).
92. RNA-seq profiling of CRISPR/Cas9 based knock outs of RNA-binding proteins in human cell line K562. https://www.encodeproject.org/search/?type=Experiment&assay_title=CRISPR+RNA-seq&replicates.library.biosample.life_stage=adult. (2017) (ENCODE Project).
93. Martinez, T. F. *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468. https://doi.org/10.1038/s41589-019-0425-0 (2020).
94. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360. https://doi.org/10.1038/nmeth.3317 (2015).
95. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 (2009).
96. Uren, P. J. *et al.* Site identification in high-throughput RNA–protein interaction data. *Bioinformatics* **28**, 3013–3020. https://doi.org/10.1093/bioinformatics/bts569 (2012).
97. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. https://doi.org/10.1093/bioinformatics/btq033 (2010).
98. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* https://doi.org/10.1038/nmeth.3810 (2016).
99. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295. https://doi.org/10.1038/nbt.3122 (2015).
100. Bauer, D. F. Constructing confidence sets using rank statistics. *J. Am. Stat. Assoc.* **67**, 687–690. https://doi.org/10.1080/01621459.1972.10481279 (1972).
101. He, F. *et al.* Integrative analysis of somatic mutations in non-coding regions altering RNA secondary structures in cancer genomes. *Sci. Rep.* **9**, 8205. https://doi.org/10.1038/s41598-019-44489-5 (2019).
102. Pan, S. *et al.* Transcription-associated mutation promotes RNA complexity in highly expressed genes-a major new source of selectable variation. *Mol. Biol. Evol.* **35**, 1104–1119. https://doi.org/10.1093/molbev/msy017 (2018).
103. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26. https://doi.org/10.1038/nbt.1754 (2011).
104. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789. https://doi.org/10.1101/gr.132159.111 (2012).
105. Consortium, G. T. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585. https://doi.org/10.1038/ng.2653 (2013).

Acknowledgements

We thank the lab members for their valuable suggestions over the course of this project. We are thankful to Dr. Mark Kaplan for providing access to space and equipment to conduct this study. We also thank Swapna Vidhur Daulatabad for providing the Kaplan–Meier plot of lncRNA RP11–301G19.1.

Author contributions

M.S., R.S. and S.C.J. conceived and designed the study. M.S. developed the POP-seq method with three versions in K562 cells and generated the NGS library for sequencing. R.S. implemented the bioinformatic tools and

integrated the datasets for downstream data analysis. M.S., R.S. and S.C.J. interpreted the data and wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM123314 (SCJ).

Competing interest
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-80846-5.

Correspondence and requests for materials should be addressed to S.C.J.

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

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021