Supplementary information

InteRRact: a web server for the interactive exploration and comparison of transcriptome-wide RNA-RNA interactions.

Egor Semenchenko[1,2], Jingwen Luo[1,3], Volodymyr Tsybulskyi[1,2], Charlie Rettig[1,3], Irmtraud M. Meyer[1,2,3]

[1] Laboratory of Bioinformatics of RNA Structure and Transcriptome Regulation, Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany

[2] Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Freie Universität Berlin, Berlin, Germany

[3] Department of Mathematics and Computer Science, Institute of Computer Science, Freie Universität Berlin, Berlin, Germany

List of pre-processed human duplex datasets

dataset name	protocol	cell type	selection	layout	replicates	sra accessions
SPLASH H1	SPLASH	H1	polyA	SINGLE	2	SRR3404943;SRR3404926
SPLASH RA		RA	polyA		2	SRR3404927;SRR3404928
SPLASH Lymphoblastoid total		GM12892	polyA		4	SRR3404939; SRR3404940; SRR3404941; SRR3404942
SPLASH Lymphoblastoid polyA		GM12892	total		4	SRR3404924; SRR3404925; SRR3404936; SRR3404937
SPLASH HeLa		HeLa	total		2	SRR3404929, SRR3404931
LIGR-seq HEK293T	LIGR-seq	HEK293T	total		2	SRR3361013, SRR3361017
PARIS HEK293T	PARIS	HEK293T	total		3	SRR2814763,SRR2814764, SRR2814765
PARIS HeLa rRNA(-)		HeLa	rRNA-low		1	SRR2814761
PARIS HeLa rRNA(+)		HeLa	rRNA-high		1	SRR2814762
RIC-seq HeLa rRNA(+)	RIC-seq	HeLa	rRNA-high	PAIRED	2	SRR8632826,SRR8632827
RIC-seq HeLa rRNA(-)		HeLa	total and rRNA depleted		2	SRR8632820,SRR8632821
RIC-seq HepG2		HepG2			8	SRR17136192, SRR17136193, SRR2412914, SRR2412915, SRR17136194, SRR17136195, SRR24172916, SRR24172917
RIC-seq GM12878		GM12878			6	SRR17136184, SRR17136185; SRR24172912, SRR17136186; SRR17136187, SRR24172913
RIC-seq IMR90		IMR90			4	SRR17136200,SRR17136201, SRR17136202,SRR17136203
RIC-seq K562		K562			6	SRR17136204, SRR17136205, SRR24172920, SRR17136206, SRR17136207, SRR24172921
RIC-seq H1		H1			4	SRR17136188, SRR17136189, SRR17136190, SRR17136191
RIC-seq hNPC		hNPC			6	SRR17136196, SRR17136197, SRR24172918, SRR17136198, SRR17136199, SRR24172919

Table S1. RNA-RNA interaction probing datasets pre-processed for InterRRact

Pre-processing the RNA-RNA interaction data

Both paired-end and single-end reads were pre-processed with fastp v1.01. (1), using the default parameters, which exclude the low-quality reads that are allowed to have a maximum 40% of the read length with Phred score <Q15. Additionally, PARIS reads were pre-processed with icSHAPE pipeline scripts (status: Oct 2024), available at https://github.com/qczhang/icSHAPE/, according to the procedures employed by the authors of this protocol.

Reads were mapped with STAR (2) using the set of parameters, which were previously determined based on the simulated chimeric read data (3)

```
--chimOutTypeJunctions
--chimOutJunctionFormat 1
--alignIntronMin 1
--alignIntronMax 10
--outSJfilterReadsAll
--chimSegmentMin 15
--chimMultimapNmax 10
--chimScoreDropMax 30
--chimScoreJunctionNonGTAG 0
```

Chimeric.out.Junction files were processed with DuplexDiscovereR (3) using the single pipeline call with the configuration outlined in Table S2.

Pipeline call	Variables

`res <- runDuplexDiscoverer(`
`  data = df,`
`  junctions_gr = junctions_gr,`
`  anno_gr = annotation_gr,`
`  anno_gr_keys = c('gene_id','gene_name','gene_type', 'repName', 'repClass', 'repFamily','encode_bad'),`
`  df_counts = dfcounts,`
`  sample_name = sample_name,`
`  fafile = NULL,`
`  min_junction_len = 10,`
`  collapse_n_inter = 1,`
`  lib_type = layout,`
`  table_type = "STAR",`
`  min_arm_ratio = 0.4,`
`  min_overlap = 10,`
`  max_gap = MAXGAP,`
`  trim_alignments = TRIM,`
`  trim_length = TRIMLEN,`
`  max_sj_shift = 15,`
`  compute_p_values = PVALUE_BOOL1,`
`  compute_odds_test_p_values = PVALUE_BOOL2`
`  )`
	df: data.frame Chimeric.out.junction output of STAR
	junctions_gr: GRanges object with the splice junction coordinates
	annotation_gr: GRanges object with combined ENCODE v48 gene annotations, UCSC Repeatmasker and ENCODE blacklist loci coordinates.
	df_counts: data.frame with per-gene counts used for the p-value estimation. Extracted from the corresponding FeatureCounts output. Contains gene ID and read count columns.

	For **single-end** reads:
	`LAYOUT = "SE"`
	`TRIM=FALSE`
	`TRIMLEN = 0`
	`MAXGAP = 40`
	`PVALUE_BOOL1 = TRUE`
	`PVALUE_BOOL2 = TRUE`

	For **paired-end** reads:
	`LAYOUT = "PE"`
	`TRIM=TRUE`
	`TRIMLEN = 100`
	`MAXGAP = 50`
	`PVALUE_BOOL1 = FALSE`
	`PVALUE_BOOL2 = TRUE`

Table S2. Configuration of DuplexDiscovereR used to infer duplex groups

We opted to use direct read counting instead of a pseudo-alignment-based method. The sequencing library preparation for duplex-probing methods varies by protocol and is not similar to the procedures typically used for bulk RNA-seq. Hence, the assumptions about fragment length – which are fundamental for pseudo-alignment-based transcript-level quantification – are likely invalid. Multi-mapped reads were included in the counting mode, where each ambiguous alignment contribution is divided by the total number of alignments for the read. Counts for both single-end and paired-end libraries were obtained with FeatureCounts, called through the Rsubread R package (4), with the following parameters.

```
featureCounts(<bam file list>,
GTF.featureType = 'exon',
GTF.attrType  = 'gene_id',
isGTFAnnotationFile = TRUE,
annot.ext = <path_to gencode annotation>,
fracOverlap = 0.3,
useMetaFeatures = TRUE,
allowMultiOverlap = TRUE,
countMultiMappingReads = TRUE,
fraction  = TRUE,
verbose = TRUE,
minOverlap = 10,
strandSpecific = c(1))
```

The coverage per sample was computed with bedtools v 2.31.1 (5)

bedtools genomecov -ibam <bam file> -bg

To make coverage more comparable across samples, raw coverage counts of duplex groups were additionally normalised for each sample using sample-specific size factors estimated by DESeq2 (6) from count matrices generated with featureCounts.

Duplex groups detected in each replicate within the same condition (dataset in Table S1) were clustered into duplex groups per-condition with

```
DuplexDiscovereR::clusterDuplexGroups(
<input gi object per-sample>,
maxgap=60,
minoverlap=10,
min_arm_ratio = 0.4)
```

To predict the loci of RNA duplex and hybridisation energy, we used a wrapper for the RNAHybrids of ViennaRNA (7)

```
DuplexDiscoverer::getRNAHybrids(
<input gi object per-condition>,
fafile = <path to hg38 fasta file>)
```

To summarise the information on the confidence measures for the per-condition duplex groups, the mean values of the per-sample duplex group were used for coverage score and gene expression, and the maximum - i.e. most conservative value - was used for the p-values.

Counts of duplex groups detected in pre-processed human datasets

Dataset	Total	Cis	Trans	Filtered out
SPLASH HeLa	7793	983	1032	5778
SPLASH H1	242113	57911	111044	73158
SPLASH RA	285562	70806	122884	91872
SPLASH Lymphoblastoid PolyA	204143	50751	72152	81240
SPLASH Lymphoblastoid total	29902	1466	1411	27025
LIGR-seq HEK293T	328743	57439	61742	209562
PARIS HeLa rRNA(-)	373096	164950	72057	136089
PARIS HeLa rRNA(+)	414482	172175	77755	164552
PARIS HEK293T	582237	164180	92787	325270
RIC-seq HeLa rRNA(-)	153133	4113	3617	145403
RIC-seq HeLa rRNA(+)	23595	316	371	22908
RIC-seq HepG2	567678	126759	164524	276395
RIC-seq IMR90	226495	38565	87929	100001
RIC-seq K562	645052	166227	174739	304086
RIC-seq hNPC	861019	164517	391202	305300
RIC-seq GM12878	1365943	351256	454037	560650
RIC-seq H1	450507	77303	205249	167955

Table S3. RNA-RNA interaction probing datasets pre-processed for InterRRact. Total count represents the number of recovered RRI, which were subsequently filtered to retain only RRI mapping to annotated genes, excluding those mapping to non-transcribed pseudogenes, to the repeat blacklist categories “Simple repeat” and “Low complexity”, or overlapping the ENCODE “High Mapping Signal” annotation.

Over-representation analysis of trans RRI communities

To identify over-represented gene ontology (GO) terms in sub-networks of trans RRIs, we rely on the Louvain community detection algorithm implemented in the R igraph package (8). For network communities containing more than four genes, over-representation analysis is performed at the user's request using the R clusterProfiler package.

```
results = clusterProfiler::enrichGO(gene = gene_query,
                            OrgDb = org.Hs.eg.db,
                            keyType = 'SYMBOL',
                            ont           = <ontology type>,
                            universe = <background>,
                            pAdjustMethod = "BH",
                            pvalueCutoff  = 1,
                            qvalueCutoff  = 1,
                            minGSSize = 1,
                            maxGSSize = 1000)
```

Many trans RRIs involve lncRNAs or other genes that lack any GO annotation. Before starting the over-representation analysis, these genes are excluded from the analysed community. As the universe of background genes, we use the per-dataset list of expressed genes, retaining genes with more than ten mapped reads. To simplify the results, we use the GOSemSim package (9), which prunes redundant GO terms and keeps the representative ontologies.

```
SEM_LIST <- list(
  BP = GOSemSim::godata(annoDb = 'org.Hs.eg.db', ont = 'BP'),
  MF = GOSemSim::godata(annoDb = 'org.Hs.eg.db', ont = 'MF'),
  CC = GOSemSim::godata(annoDb = 'org.Hs.eg.db', ont = 'CC')
)

clusterProfiler::simplify(
              x = results,
              cutoff = 0.7,
              by = "Count",
              select_fun = max,
              measure = "Wang",
              semData = <matching set from SEM_LIST>
            )
```

The resulting GO terms are cached on the InteRRact server backend and are automatically used without re-calculation based on the current network community structure.

References

1. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884–90. doi:10.1093/bioinformatics/bty560

2. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21. doi:10.1093/bioinformatics/bts635

3. Semenchenko E, Tsybulskyi V, Meyer IM. DuplexDiscoverer: a computational method for the analysis of experimental duplex RNA–RNA interaction data. Nucleic Acids Res. 2025 Apr 24;53(7):gkaf266. doi:10.1093/nar/gkaf266

4. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019 May 7;47(8):e47. doi:10.1093/nar/gkz114

5. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2. doi:10.1093/bioinformatics/btq033

6. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec;15(12):550. doi:10.1186/s13059-014-0550-8

7. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011 Nov 24;6(1):26. doi:10.1186/1748-7188-6-26

8. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, et al. igraph: Network Analysis and Visualization [Internet]. 2026 [cited 2026 Apr 9]. Available from: https://cran.r-project.org/web/packages/igraph/index.html

9. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010 Apr 1;26(7):976–8. doi:10.1093/bioinformatics/btq064