DuplexDiscoverer : a computational method for the analysis of experimental duplex RNA-RNA interaction data

Egor Semenchenko[1,2], Volodymyr Tsybulskyi[1,2], 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, Robert-Rössle-Str. 10, 13125 Berlin, Germany
[2]Department of Biology, Chemistry and Pharmacy, Institute of Chemistry and Biochemistry, Thielallee 63, Freie Universität Berlin, 14195 Berlin, Germany
[3]Department of Mathematics and Computer Science, Institute of Computer Science, Takustraße 9, Freie Universität Berlin, 14195 Berlin, Germany

*To whom correspondence should be addressed. Email: irmtraud.meyer@cantab.net

1 Supplementary Tables

1.1 RNAnue installation and runtime issues

Problem	Encountered	Program type	Step	Operation system	Libraries	link to Github issue/error log
segmentation fault	runtime	Docker container	RNAnue detect	Red Hat Enterprise Linux (RHEL9)	Singularity: apptainer version 1.3.6-1.el9;	https://github.com/Ibvt/RNAnue/issues/25
				Ubuntu 20.04	Docker: 28.0.1, build 068a01e;	
error when there is no controls			RNAnue align	Ubuntu 20.04		https://github.com/Ibvt/RNAnue/issues/24
possible removed sub-call			RNAnue complete	RHEL9	Seqan 3.3.0; Boost 1.83.0; ViennaRNA 2.5.1; Segemehl 0.3.4; htslib 1.19;	https://doi.org/10.5281/zenodo.10789913
missing htslib unlisted in dependencies	installation	compiled from source	cmake			
compilation error while building			make			

Supplementary Table S1: Description of the installation and runtime problems one can encounter for RNAnue version 0.2.3. We provide a link to respective pages on RNAnue repository for the known issues. Error logs of several runtime and installation problems are uploaded to Zenodo. We note that our HPC utilises Singularity to run Docker containers which may have affected the execution on RHEL9. The run-time errors (Segmentation fault) persisted for versions 0.2.1 and 0.2.3 compiled from the source.

1.2 List of the samples analysed with DuplexDiscoverer

SRA accession	Celltype	Protocol	Sample	Pre-processing	Layout
SRR3404929	HeLa	SPLASH	HeLa1	none, already processed in SRA	SINGLE
SRR3404931	HeLa	SPLASH	HeLa2		SINGLE
SRR3404943	hES1	SPLASH	ES_1		SINGLE
SRR3404926	hES2	SPLASH	ES_2		SINGLE
SRR3404927	RA5	SPLASH	RA1		SINGLE
SRR3404928	RA5	SPLASH	RA2		SINGLE
SRR3404939	LMB	SPLASH	LBpoly1		SINGLE
SRR3404940	LMB	SPLASH	LBpoly2		SINGLE
SRR3404941	LMB	SPLASH	LBpoly3		SINGLE
SRR3404942	LMB	SPLASH	LBpoly4		SINGLE
SRR3404924	LMB	SPLASH	LBtotal1		SINGLE
SRR3404925	LMB	SPLASH	LBtotal2		SINGLE
SRR3404936	LMB	SPLASH	LBtotal3		SINGLE
SRR3404937	LMB	SPLASH	LBtotal4		SINGLE
SRR3361013	HEK293T	LIGR-seq	LIGR_rep1	Phread >=Q15 filter with fastp	SINGLE
SRR3361017	HEK293T	LIGR-seq	LIGR_rep2		SINGLE
SRR8632820	HeLa	RIC-seq	RIC_rrna1		PAIRED
SRR8632821	HeLa	RIC-seq	RIC_rrna2		PAIRED
SRR2814761	HeLa	PARIS	PARIS_hela_low	Duplicates and barcodes removed with PARIS scripts `https://github.com/qczhang/icSHAPE/`	SINGLE
SRR2814762	HeLa	PARIS	PARIS_hela_high		SINGLE
SRR2814763	HEK293T	PARIS	PARIS_HEK_1		SINGLE
SRR2814764	HEK293T	PARIS	PARIS_HEK_2		SINGLE
SRR2814765	HEK293T	PARIS	PARIS_HEK_3		SINGLE
SRR6811718	HeLa	RNA-seq	Read data for artificial chimeric reads	Phread >=Q15 filter with fastp	SINGLE
SRR6811722	HeLa	RNA-seq			SINGLE
SRR6811723	HeLa	RNA-seq			SINGLE
SRR6811728	HeLa	RNA-seq			SINGLE
SRR6811719	HeLa	RNA-seq			SINGLE

Supplementary Table S2: List of the samples for RNA duplex probing experiments analyzed with DuplexDiscovereR. Except for the SPLASH and PARIS data, we applied quality control via fastp [1] with default parameters. SPLASH data is sequenced in pair-end mode and deposited to the NCBI Sequence Read Archive (SRA) already preprocessed with overlapping read mates merged into single reads. PARIS data contains barcodes and adapters, which needed to be removed by icSHAPE scripts, as in the original PARIS pipeline.

1.3 List of simulated RNA duplex-probing libraries

Sample	Num. chimeric	Num. background	Num. non chimeric	Num. total	Arm A-B lengths	Num. *cis* DG	Num. *trans* DG	Num. *cis* chimeric reads	Num. *trans* chimeric reads	*cis*:*trans* ratio
sim_1	1053076	12498896	0	13551972	20-20	10819	99961	950003	103073	9.2
sim_2	919524	12498897	0	13418421	30-30	41822	55000	522424	397100	1.3
sim_3	1049450	12498898	0	13548348	40-40	55394	55000	522894	526556	1
sim_4	921385	12498899	0	13420284	50-50	41939	55000	522533	398852	1.3
sim_5	569112	12498896	100000	13168008	20-20	30000	30000	284556	284556	1
sim_6	570882	12498896	100000	13169778	30-30	30000	30000	285412	285470	1
sim_7	569990	12498896	100000	13168886	40-40	30000	30000	285256	284734	1
sim_8	569014	12498896	100000	13167910	50-50	30000	30000	284020	284994	1
sim_9	570077	12498896	100000	13168973	20-30	30000	30000	285037	285040	1
sim_10	570272	12498896	100000	13169168	20-40	30000	30000	284901	285371	1
sim_11	571155	12498896	100000	13170051	20-50	30000	30000	285721	285434	1
sim_12	570287	12498896	100000	13169183	30-40	30000	30000	285497	284790	1
sim_13	570315	12498896	100000	13169211	30-50	30000	30000	285130	285185	1
sim_14	569194	12498896	100000	13168090	40-50	30000	30000	284674	284520	1

Supplementary Table S3: Characteristics of the samples with simulated chimeric reads Four sets of artificially created chimeric reads with variable lengths were simulated using publicly available RNA-seq libraries of HeLa cells `SRR6811718, SRR6811722, SRR6811723, SRR6811728`. `SRR6811722` library .fastq file was concatenated with each set of chimeric reads, resulting in samples containing chimeric reads to be tested and "normal" reads as the background. Samples sim1-4 were used to test the mapping scheme, see Figure S4. Samples 4-14 with imbalanced arm lengths were used to benchmark the accuracy of predictions. Non-chimeric splice-junction (SJ) spanning reads from the background library were added separately to observe SJ filtering in DuplexDiscoverer and CRSSANT explicitly

1.4 Chimeric read types in the background of simulated dataset

Read type	**Count**	**% of chimeric input**	**% of the background library size**
two arm splice junc.	6392	0.9	0.05
multi-map	38906	5.2	0.31
multi-split	12721	1.7	0.10
multi-split&map	5033	0.7	0.04
bad junction	5415	0.7	0.04
too short junction <5nt	479	0.1	0.00
self-overlap antisense	10064	1.3	0.08
self-overlap	4659	0.6	0.04
two arm no DG	48111	6.4	0.38
two arm clustered to DG	15455	2.1	0.12

Supplementary Table S4: Counts of read types obtained by DuplexDiscoverer in the background RNA-seq library `SRR6811722`

1.5 Artificially created non-chimeric reads in the simulated dataset

DuplexDiscoverer			CRSSANT		
Classification type	Count	% of simulated SJ reads	Intermediate file	Count	% of simulated SJ reads
2arm – not DG	77	0.08	gap1.sam	96046	96
2arm – SJ	267	0.27			
multi map	316	0.32			
multi split	209	0.21	rri.sam	364	0.36
multi split&map	31	0.03			
not chimeric	99053	99.05			
Formed DG	42	**0.04**	Formed DG	7398	**7.4**

Supplementary Table S5: A total of 100000 reads which span 15473 splice junctions (SJ) were added to simulated samples *sim 4-14*. For DuplexDiscoverer and CRSSANT it is possible to observe how these were processed and whether they were included in the final DGs. For DuplexDiscoverer most of such reads were removed at the mapping stage, while CRSSANT filters them internally. Overall, both methods correctly filter out most of the SJ-spanning reads. DGs formed by artificial SJ reads comprise between 3-4% of DGs reported by CRSSANT in simulated dataset

1.6 STAR mapping configurations evaluated for mapping of the chimeric reads

DuplexDiscoverer	ARRIBA
	—chimOutType Junctions
	—chimOutJunctionFormat 1
—chimOutType Junctions	—alignIntronMin 1
—chimOutJunctionFormat 1	—alignIntronMax 10
—alignIntronMin 1	—chimMultimapNmax 10
—alignIntronMax 10	—chimSegmentMin 10
—outSJfilterReads All	—chimScoreMin 1
—chimSegmentMin 15	—chimJunctionOverhangMin 10
—chimMultimapNmax 10	—chimScoreSeparation 1
—chimScoreDropMax 30	—chimMultimapNmax 50
—chimScoreJunctionNonGTAG 0	—chimScoreDropMax 30
	—chimSegmentReadGapMax 3
	—chimScoreJunctionNonGTAG 0
CRSSANT	**RNAContacts**
—chimMultimapNmax 10	
—chimOutType Junctions	
—chimOutJunctionFormat 1	—chimOutType Junctions
—alignIntronMin 1	—chimOutJunctionFormat 1
—alignIntronMax 10	—alignIntronMin 1
—scoreGap 0	—alignIntronMax 10
—outFilterMultimapNmax 10	—chimSegmentMin 15
—scoreGapNoncan 0	—chimJunctionOverhangMin 15
—scoreGapGCAG 0	—chimScoreMin 1
—scoreGapATAC 0	—chimScoreDropMax 25
—chimFilter None	—chimSegmentReadGapMax 3
—scoreGenomicLengthLog2scale −1	—chimScoreJunctionNonGTAG −1
—chimSegmentMin 5	—scoreGapNoncan −1
—chimJunctionOverhangMin 5	—scoreGapATAC −1
—chimScoreJunctionNonGTAG 0	—scoreGapGCAG −1
—chimScoreDropMax 80	—outFilterMatchNminOverLread 0.5
—chimNonchimScoreDropMin 20	—outFilterScoreMinOverLread 0.5
—outFilterScoreMinOverLread 0	
—outFilterMatchNminOverLread 0	

Supplementary Table S6: STAR [2] mapping configurations evaluated for mapping of the chimeric reads. Parameters used in ARRIBA [3], CRSSANT [4] and RNAContacts [5] pipelines were adapted to output chimeric reads into `Chimeric.out.junction` file. By setting `---alignIntronMax 10`, predicting de-novo splice junctions were disabled and new junctions were treated as chimeric. The default parameters for DuplexDiscoverer were chosen by determining the best combination of the PPV and sensitivity estimated on the simulated data.

2 Supplementary Notes

2.1 Supplementary Note S1. Simulating artificial duplex groups

To create the samples with artificial duplex groups, we used sequences from "normal" RNA-seq reads libraries. The general outline of the procedure consists of two steps. First, we arrange individual RNA-seq reads into pairs, forming a pool of the paired sequences, which serve as the sequence sources or "backbones" for the artificial duplex groups (DGs). Second, we form synthetic chimeric reads by extraction and concatenation of sub-sequences from these backbones. We expand on the details of these steps below.

2.1.1 Forming a pool of source read pairs

Four samples `SRR6811718, SRR6811722, SRR6811723, SRR6811728` from a publicly available dataset of 100 bp single-ended RNA-seq of HeLa cells were merged into a single file and aligned with STAR [2] using default parameters. We selected reads that map to the genes with expression counts (values from GeneCounts.tab of STAR output) above 30 that map without gaps or ambiguities. There are two categories of DGs: *cis* - where both arms map to the same gene (approximating true *cis* RNA-RNA interaction, where the RNA duplex is formed within the single transcript) and *trans* - where the two arms map to different genes. To construct read pairs for *cis* DGs, we first split all genes (that were filtered by the expression count cutoff 30) - into 50 subgroups of 400 genes. This partitioning was arbitrarily selected for technical purposes to reduce the sampling space and to improve processing speed through parallelization.

For each gene subgroup, we considered only reads mapped to the corresponding subset of genes. One million reads were randomly sampled (with replacement) two times - for arms A and B, respectively. Reads were arranged in parallel and formed pairs and their paired read mapping coordinates were converted into `GInteractions` object of the R `InteractionsSet` package [6]. Read pairs were filtered to retain only those mapping to the same gene and consisting of two distinct reads. To further refine the dataset and remove redundant pairs, we applied the clustering function of DuplexDiscovereR. Artificial read pairs were treated as chimeric reads to identify and filter out overlapping pairs. Within each subgroup, only unique (non-overlapping) read pairs and one representative pair per cluster were retained.

After the procedure above was repeated for each subgroup, we stacked derived read pairs into a single dataset and performed a second round of clustering to remove potential rare or complicated cases where loci mapped to overlapping genes. The final *cis* DG dataset consists of non-redundant read pairs mapping to the same genes. For the read pairs for *trans* DGs, we followed a similar approach, which in this case required no gene sub-groups. We sampled 1.5 million reads twice, arranged them into pairs and removed redundant pairs through clustering with DuplexDiscoverer. For the *trans* DGs, we selected read pairs where arms A and B mapped to different genes. After applying these procedures, we obtained a pool of 179,840 *cis* read pairs and 1.4 million *trans* read pairs. Among these, 96.2% reads were used only once and formed a single pair, while the other reads were used at most twice - in *cis* and in *trans* pairs.

2.1.2 Simulating reads for artificial DGs

The number of artificial DGs in the simulated sample is determined by the selected number of source read pairs from the pre-arranged pool. For samples sim5–sim14, we selected 30,000 *cis* and 30,000 *trans* read pairs by taking two random samples from the respective subsets of *cis* and *trans* source read pairs. Each simulated sample featured a single combination of chimeric arm lengths.

To generate chimeric reads belonging to one DG, we used the following approach: For each read in a source pair, we extracted a sub-sequence starting at the read's midpoint with a length corresponding to the target arm length. The extracted sub-sequences were then concatenated to form synthetic two-segment sequences. This process was repeated 5 to 15 times, with the number of repeats randomly assigned for each DG, defining the number of reads in that DG. During each iteration, random shifts were introduced at the start of a new sub-sequence within the range [-chimeric arm length $\times$ 0.3; chimeric arm length $\times$ 0.3]. The arm lengths of the extracted sub-sequences were drawn from a discrete normal distribution, with the mean set to the desired arm length and a standard deviation of 1 nt. The final duplex group consisted of non-identical chimeric reads sharing at least 2/3 of their sequence in both segments.

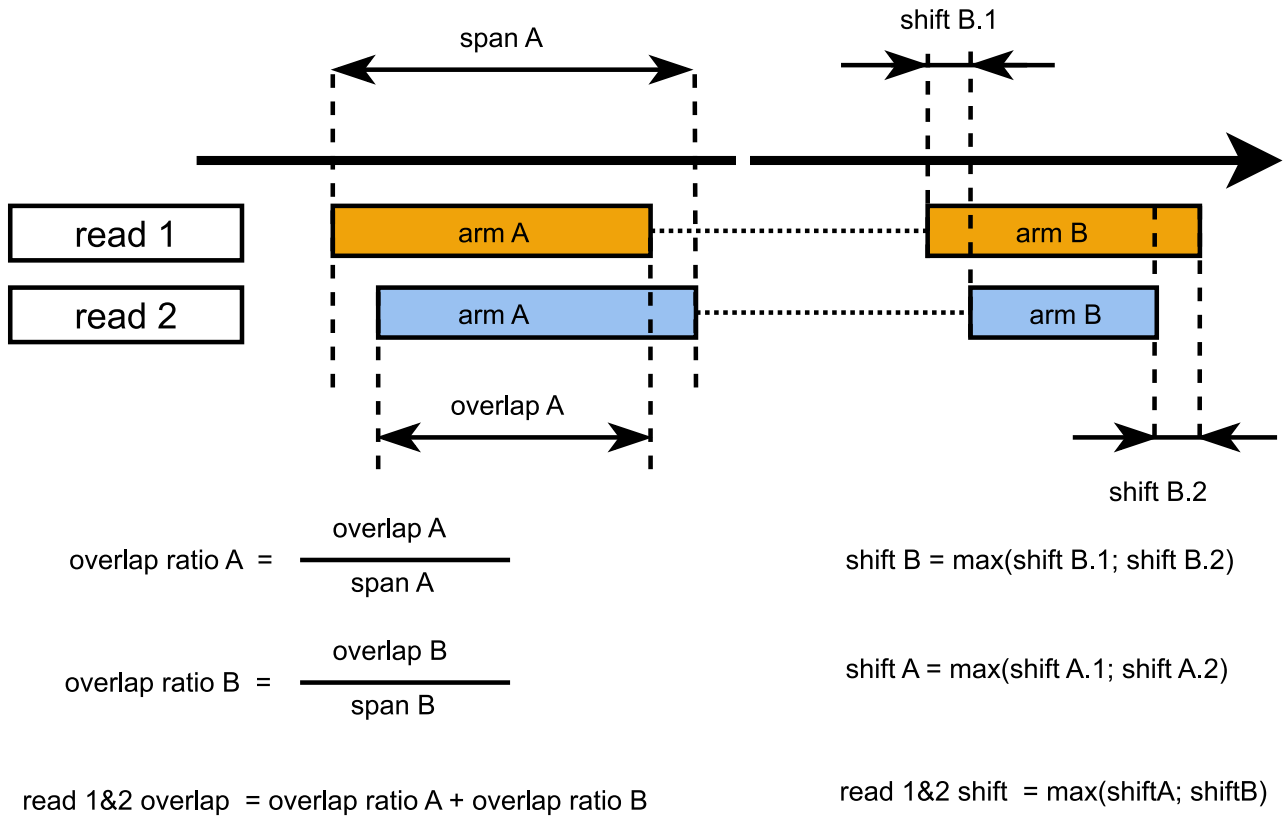
For samples sim1–sim4, the same procedure was used to generate chimeric reads from source read pairs. However, only *trans* source read pairs were pre-arranged, and *trans* DGs were generated first. *Cis* source read pairs were sampled dynamically ("on the fly") with no restriction to any gene sub-group. After sampling, redundant read pairs were removed through clustering, as described above, and *cis* chimeric reads were then generated. Sampling was repeated iteratively multiple times until the total number of *cis* and *trans* chimeric reads approached 1.2 million. These samples were primarily used to investigate optimal STAR parameters for chimeric read mapping rather than for benchmarking.

After the chimeric reads were simulated, they were concatenated with the `SRR6811722` RNA-seq library serving as background reads. We used a single background sample to prevent background- induced variability in benchmarks and

to reduce computational cost. Additionally, we extracted 100.000 reads which span 15473 splice junctions (SJs) and added them to simulated samples sim5-sim14 to obtain the statistics on the filtering of SJ for DuplexDiscoverer and CRSSANT, see Supplementary Table S5.

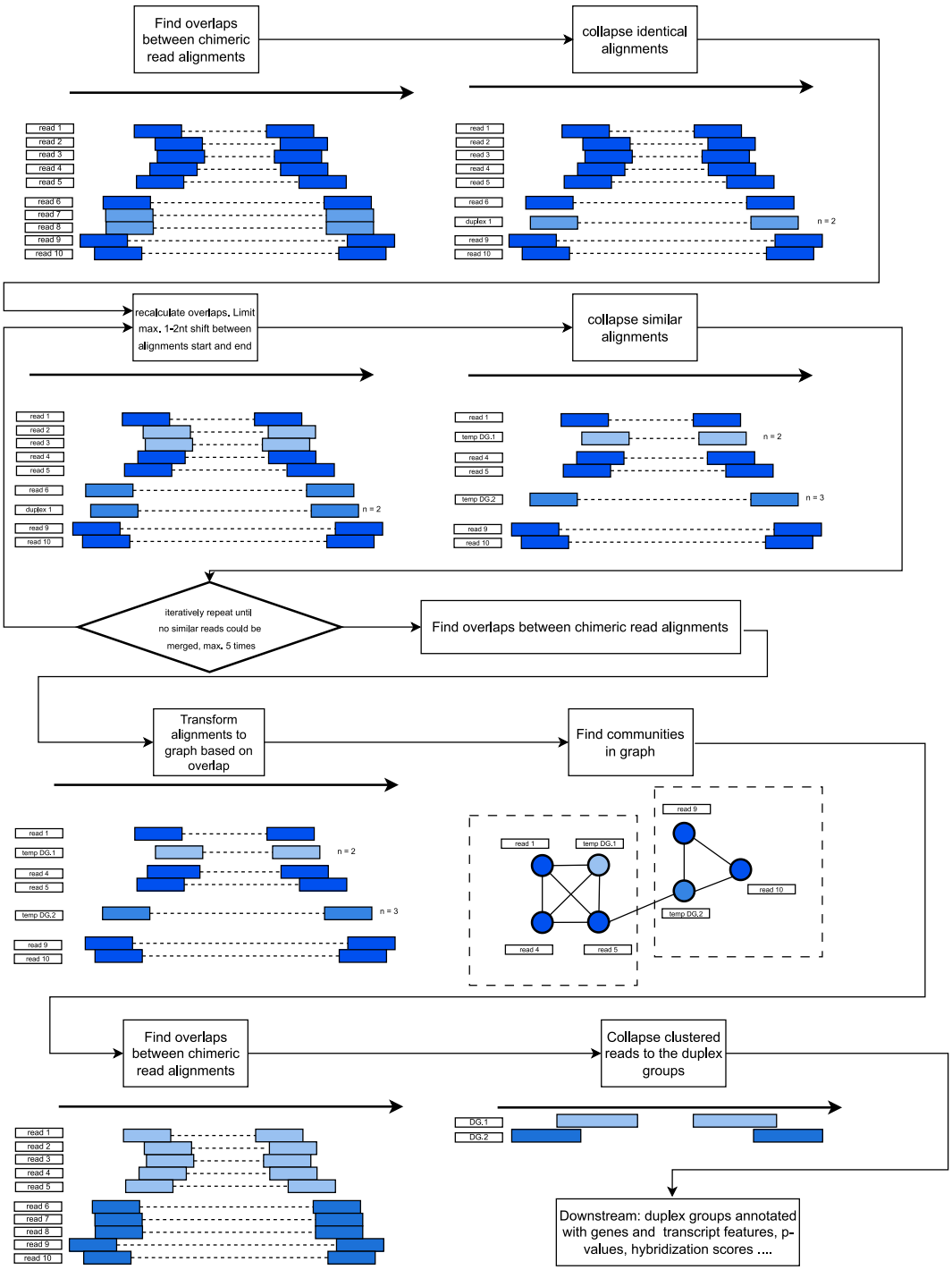
3 Supplementary Figures

3.1 Parameters for clustering the duplex groups



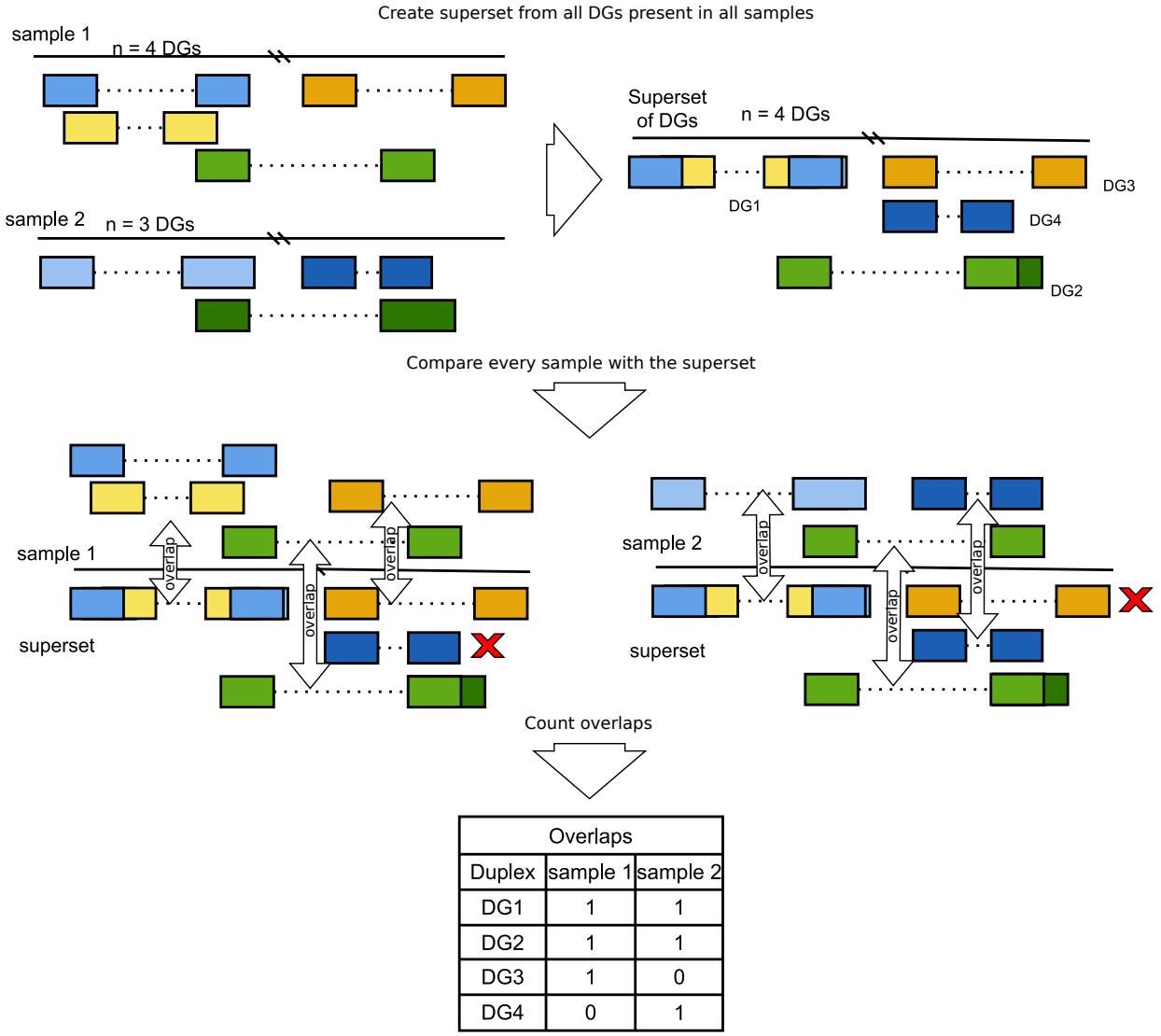
Supplementary Figure S1: Definitions of the overlap and shift between chimeric reads. Split-read-based overlap and shifts are parameters which can be adjusted for read comparison and clustering procedures of DuplexDiscoverer.

3.2 Iterative read clustering procedure



Supplementary Figure S2: The iterative DG merging procedure. The first step collapses identical chimeric reads and is equivalent to the deduplication. Collapsing similar alignments uses user-defined shift thresholds to find temporary duplex groups and repeats until all small-shifted reads are clustered, or a maximum of five times, whichever condition is reached first. Finally, the full graph based on all reads is built, where each read becomes a node and edges define overlap, weighted by the overlap ratio. After communities in the graph are found, graph representation is reverted to the reads. Reads are collapsed to the DGs with alignment boundaries re-defined as min- and max-coordinates of the reads within the group. Note that the "collapse similar alignments" step internally uses the same method of read merging based on the graph. For samples of small size with fewer than one million reads, iterative merging can be disabled, which slightly increases the speed of the clustering.

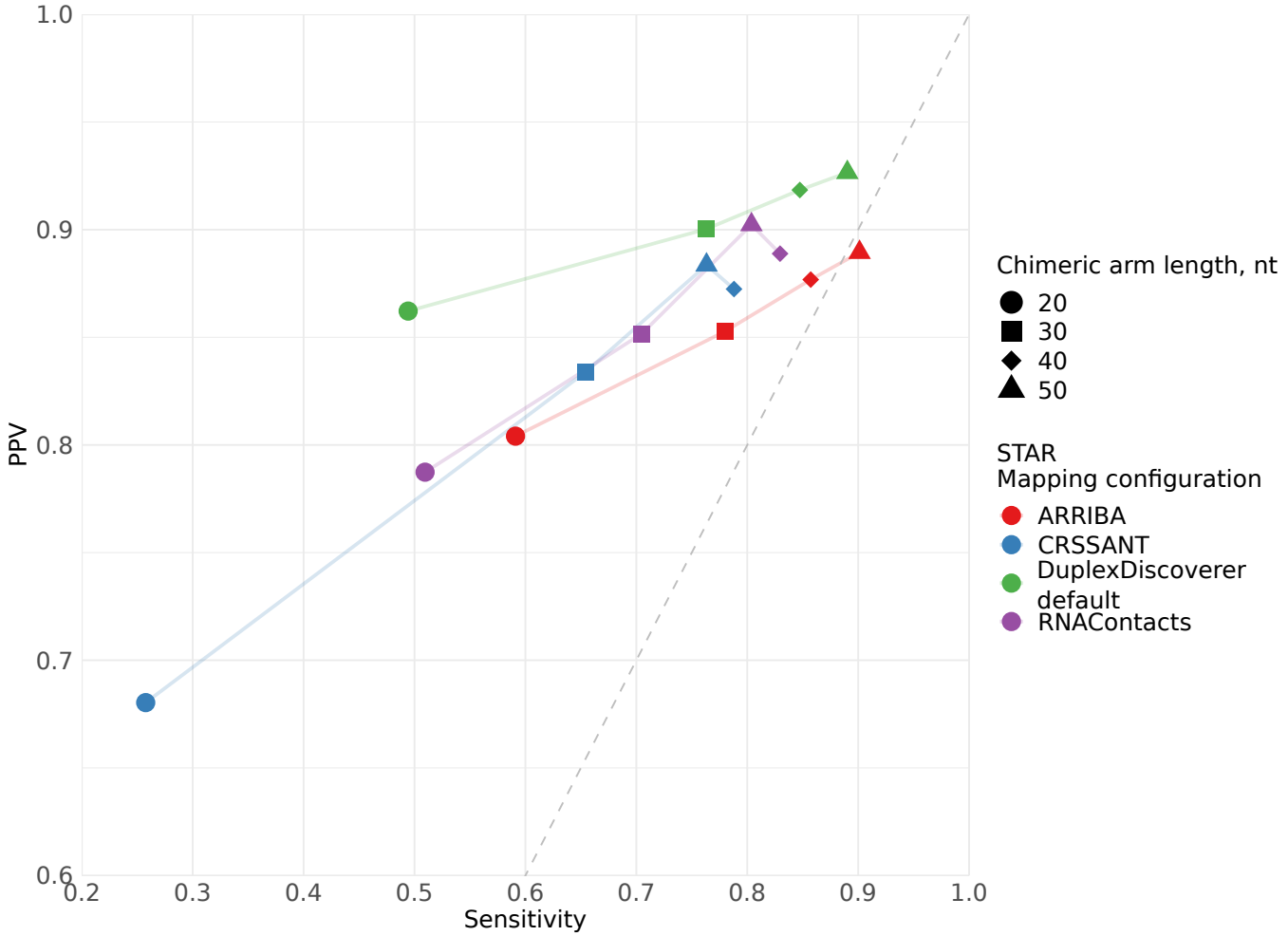
3.3 Comparisons between multiple sets of duplex groups



Overlaps		
Duplex	sample 1	sample 2
DG1	1	1
DG2	1	1
DG3	1	0
DG4	0	1

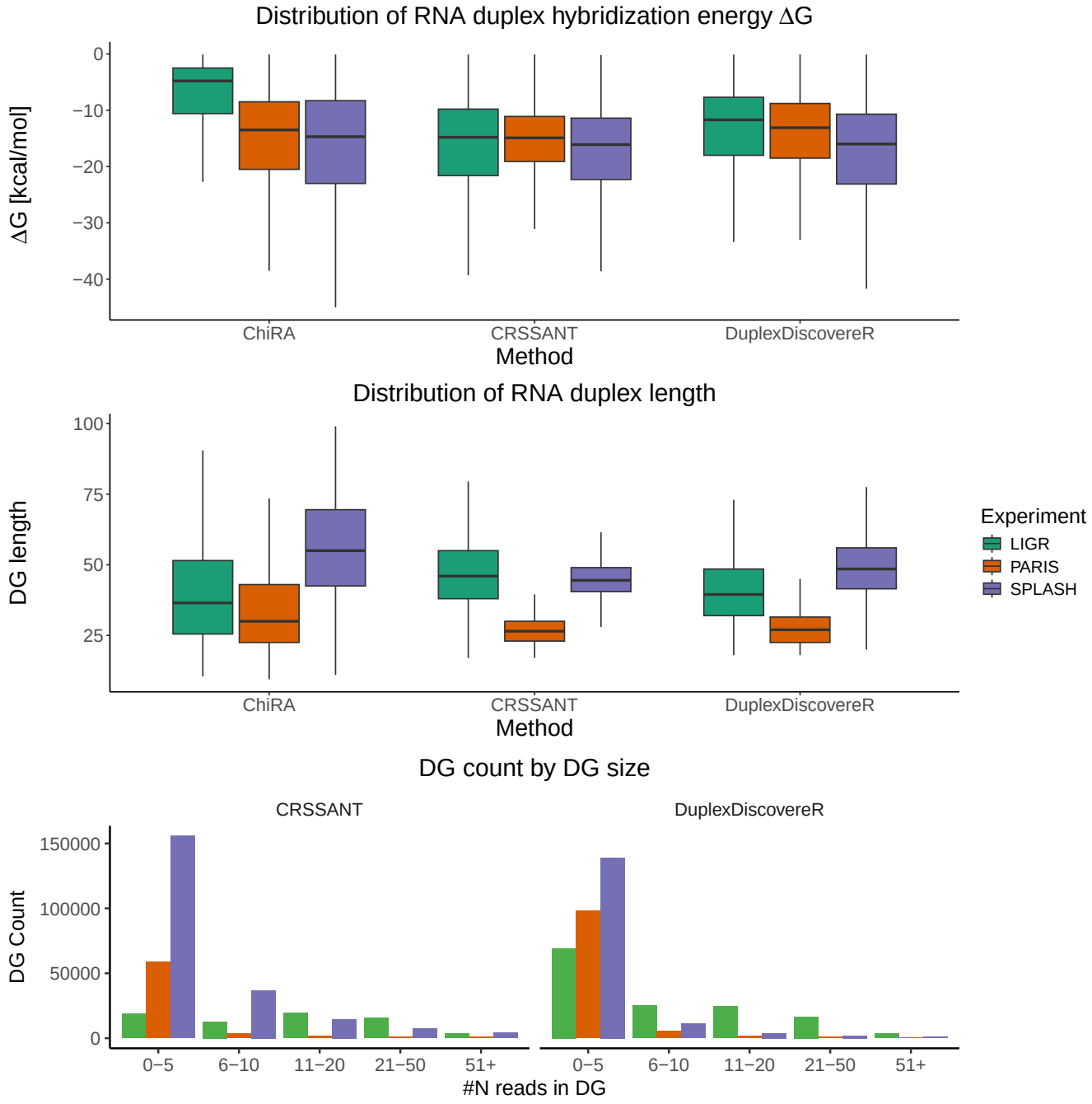
Supplementary Figure S3: Strategy to compare duplex groups derived from different experiments. First, the non-redundant superset of duplex groups (DGs) is created. If DGs overlap, they form a new DG with extended boundaries in a superset, defining the total number of DGs. Each DG in the superset could be found in at least one sample. Every sample is then compared to the superset. Finally, overlaps between the superset and the samples are recorded once per DG in the superset. The resulting table can be used for further per-sample comparisons.

3.4 Mapping the chimeric reads with different STAR configurations



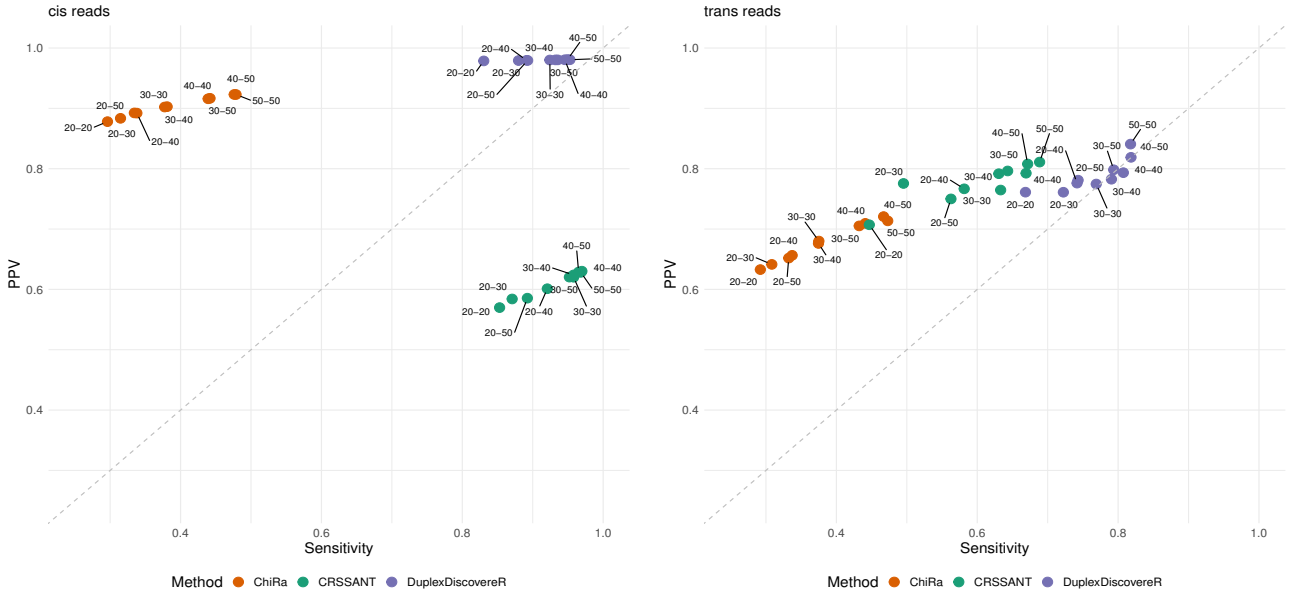
Supplementary Figure S4: Mapping accuracy of different STAR configurations. Several alignment configurations were tested to map the simulated chimeric read samples, see Supplementary Tables S6 and S3. Coloured connecting lines are drawn to highlight the groups of values. All mapping configurations performed best for the samples with chimeric arms longer than 30 nt. Both strategies with changed gap score penalties – RNAContacts and CRSSANT – demonstrated lower sensitivity for samples with 50 nt chimeric arms.

3.5 Distribution of DG features

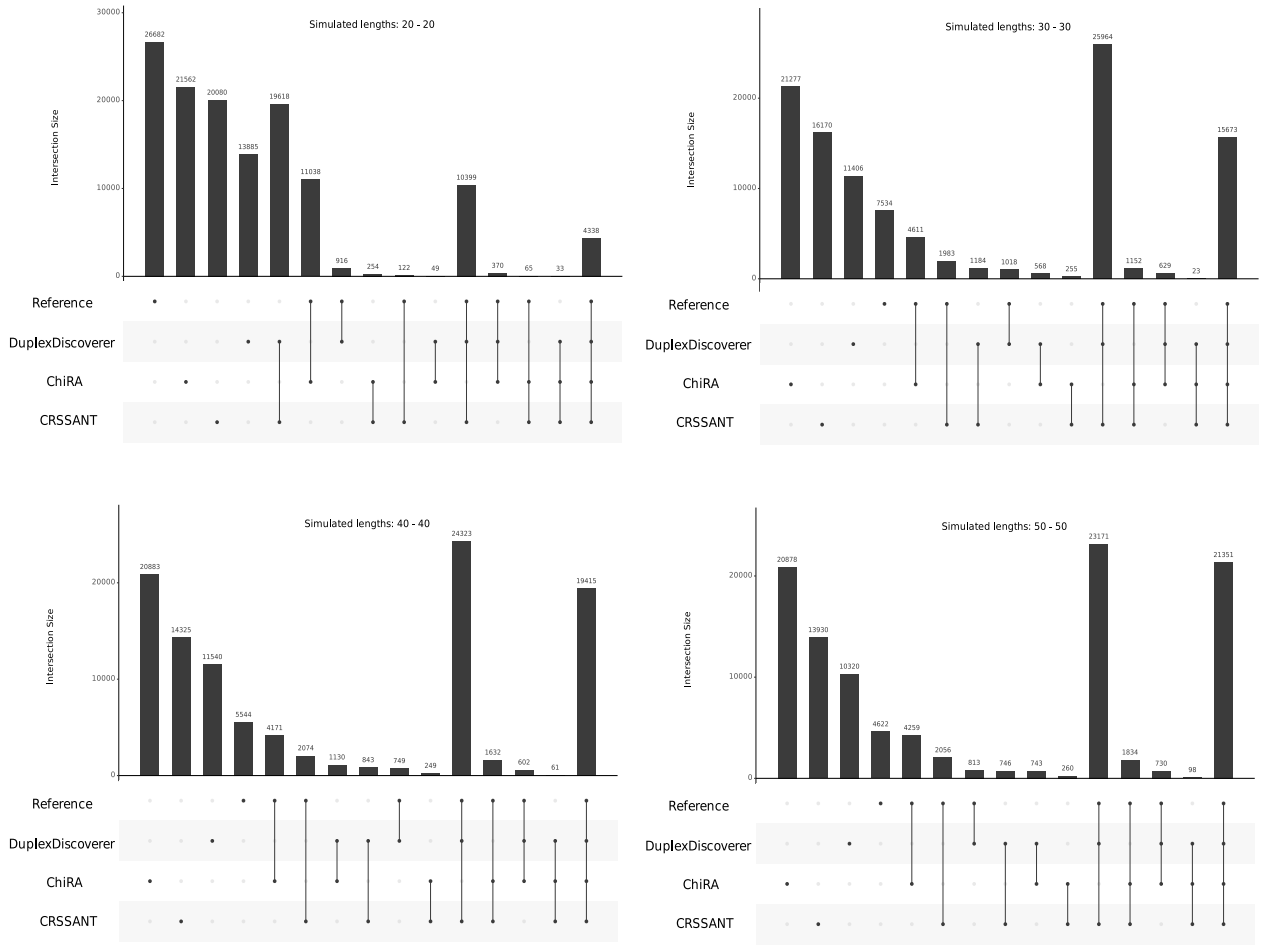


Supplementary Figure S5: Distribution of hybridisation energies and DG lengths for different pipelines. n=2 replicate SPLASH ES cells data was aggregated into a single set and only DGs supported by both replicates left.

3.6 Benchmarking of the DG detection

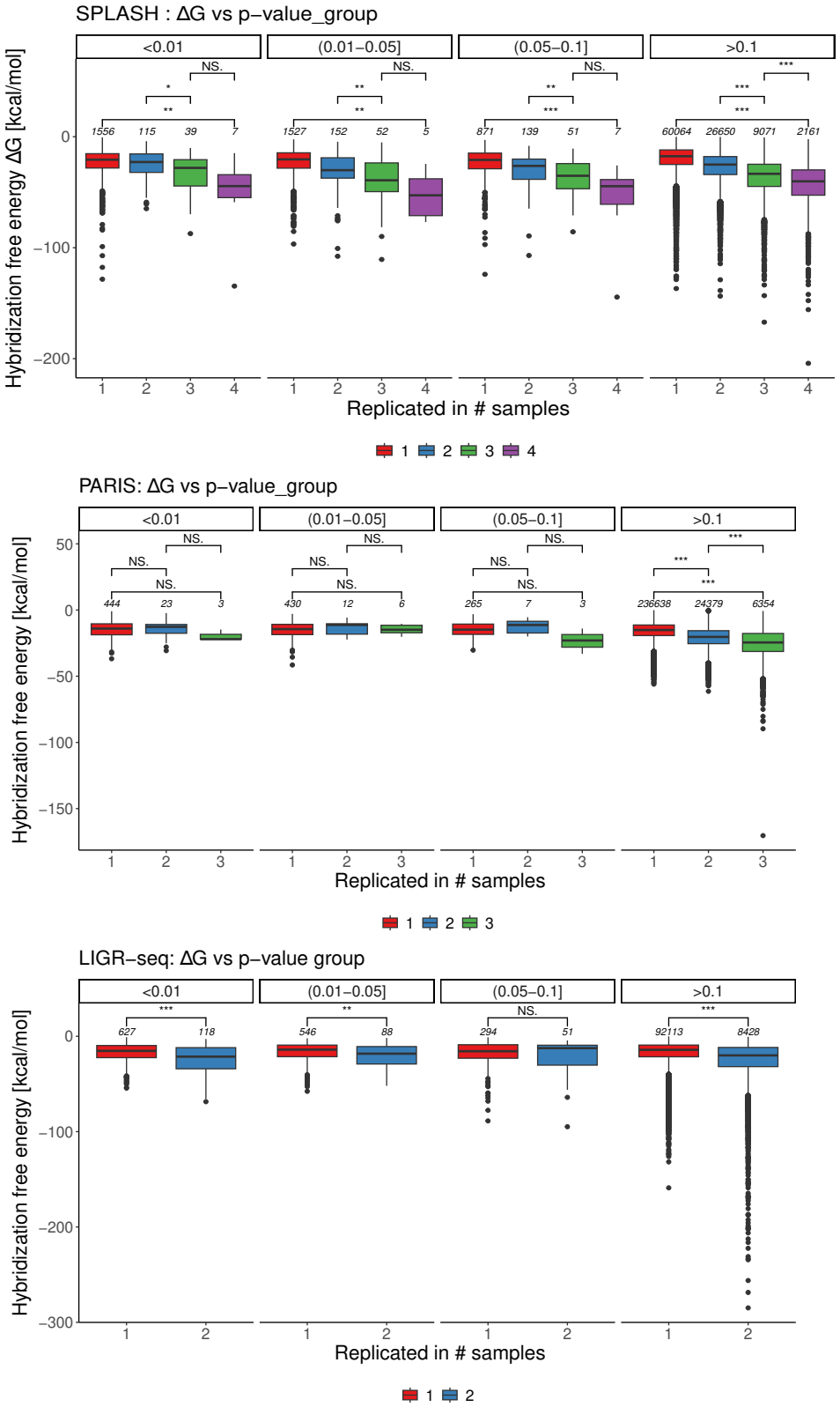


Supplementary Figure S6: Benchmarking the predictive performance of DuplexDiscoverer, ChiRA and CRSSANT on the *cis*-DGs subset of the simulated data. For all methods, the correct detection of the *trans* chimeric reads is a more challenging task.



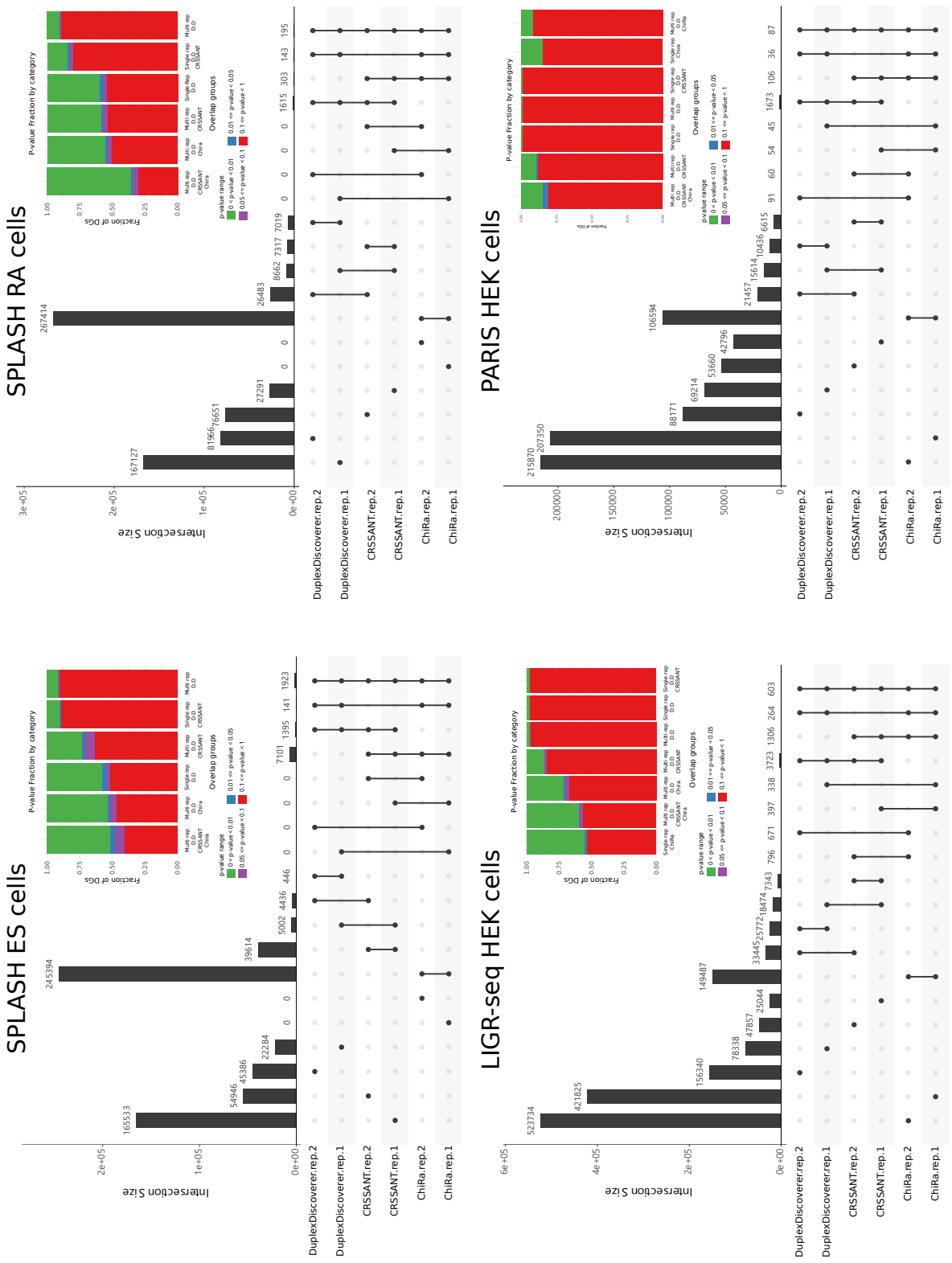
Supplementary Figure S7: Comparisons of simulated DG detected by DuplexDiscoverer, ChiRA and CRSSANT pipelines for different DG arm lengths.

3.7 Distributions of hybridisation energies and p-values



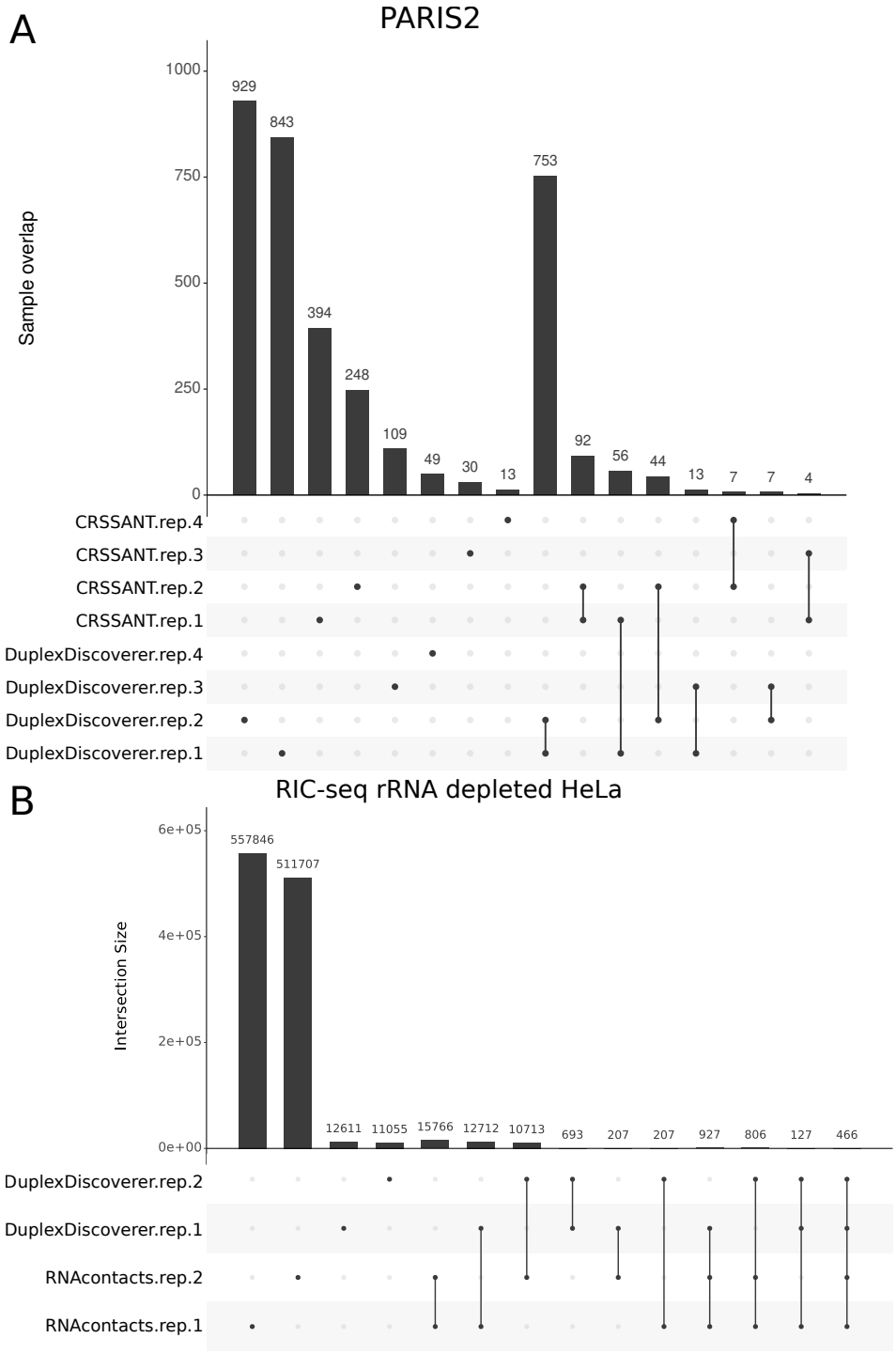
Supplementary Figure S8: Distribution of hybridisation energies separated into groups per replication level and size of the DG. n=4 replicates SPLASH polyA- selected Lymphoblastoid cells data. n=3 replicates of PARIS HEK cells, n = 2 replicates of LIGR-seq. P-values for Wilcoxon rank-sum test are coded as "***" $\leq$ 0.001, "**" $\leq$ 0.01, "*"$\leq$0.05, "NS" - not significant.

3.8 Comparisons of RNA interaction probing results produced by different methods

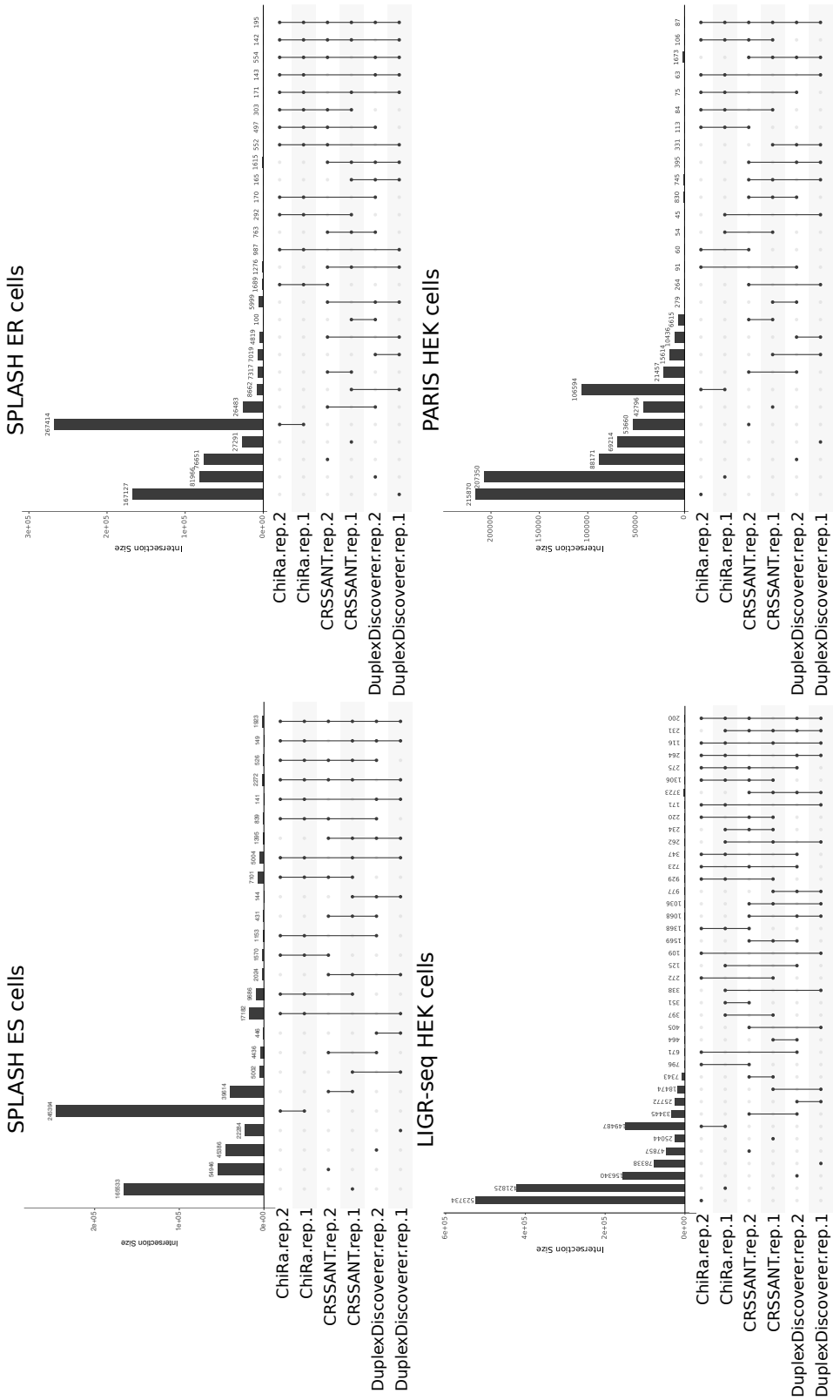


Supplementary Figure S9: Per-replicate intersection of RNA-RNA interactions devised by three different computational methods. Fraction of p-values below 0.01, 0.01 and 0.05 are shown for the intersections with DuplexDiscoverer.

3.9 Comparisons of RNA interaction probing results produced by different methods — all replicate categories



Supplementary Figure S10: Per-replicate intersection of RNA-RNA interactions in PARIS2 detected with DuplexDiscoverer and CRSSANT and RIC-seq with DuplexDiscoverer and RNAContacts.



Supplementary Figure S11: Per-replicate intersection of RNA-RNA interactions in SPLASH, PARIS, LIGR-seq. All replicate categories with n > 50 DGs in the intersection are shown.

References

[1] Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics,* **34**, i884–i890.

[2] Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* **29**, 15–21.

[3] Uhrig,S., Ellermann,J., Walther,T., Burkhardt,P., Fröhlich,M., Hutter,B., Toprak,U.H., Neumann,O., Stenzinger,A., Scholl,C. *et al.* (2021) Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome research,* **31**, 448–460.

[4] Zhang,M., Hwang,I.T., Li,K., Bai,J., Chen,J.F., Weissman,T., Zou,J.Y. and Lu,Z. (2022) Classification and clustering of RNA crosslink-ligation data reveal complex structures and homodimers. *Genome Research,* **32**, 968–985.

[5] Margasyuk,S.D., Vlasenok,M.A., Li,G., Cao,C. and Pervouchine,D.D. (2023) RNAcontacts: A Pipeline for Predicting Contacts from RNA Proximity Ligation Assays. *Acta Naturae,* **15**, 51–57.

[6] Lun,A.T., Perry,M. and Ing-Simmons,E. (2016) Infrastructure for genomic interactions: Bioconductor classes for Hi-C, ChIA-PET and related experiments. R package available at `https://doi.org/doi:10.18129/B9.bioc.InteractionSet`