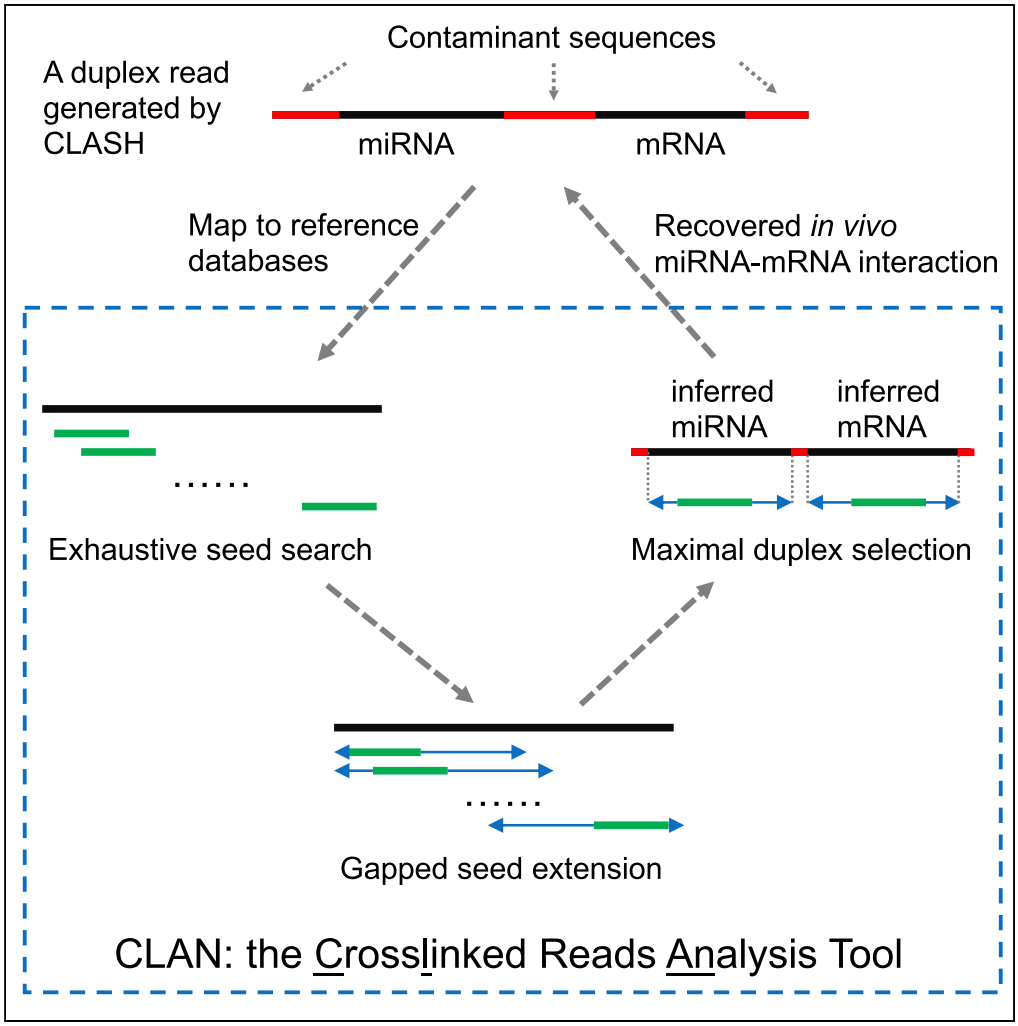
iScience

CellPress

Special Issue: RECOMB-Seq 2019

Article

Accurate and Efficient Mapping of the Cross-Linked microRNA-mRNA Duplex Reads



Cuncong Zhong, Shaojie Zhang

cczhong@ku.edu

HIGHLIGHTS

Cross-linked miRNA-mRNA read may contain artificial sequences and is difficult to map

We developed CLAN for miRNA-mRNA duplex read mapping

CLAN aims at maximizing the total length of the mapped segments of the read

CLAN was benchmarked with other mapping tools and showed improved performances

Zhong & Zhang, iScience *18*, 11–19
August 30, 2019
https://doi.org/10.1016/j.isci.2019.05.038

iScience

CellPress

Special Issue: RECOMB-Seq 2019

Article

Accurate and Efficient Mapping of the Cross-Linked microRNA-mRNA Duplex Reads

Cuncong Zhong[1,3,*] and Shaojie Zhang[2]

SUMMARY

MicroRNA (miRNA) trans-regulates the stability of many mRNAs and controls their expression levels. Reconstruction of the miRNA-mRNA interactome is key to the understanding of the miRNA regulatory network and related biological processes. However, existing miRNA target prediction methods are limited to canonical miRNA-mRNA interactions and have high false prediction rates. Other experimental methods are low throughput and cannot be used to probe genome-wide interactions. To address this challenge, the Cross-linking Ligation and Sequencing of Hybrids (CLASH) technology was developed for high-throughput probing of transcriptome-wide microRNA-mRNA interactions *in vivo*. The mapping of duplex reads, chimeras of two ultra-short RNA strands, poses computational challenges to current mapping and alignment methods. To address this issue, we developed CLAN (CrossLinked reads ANalysis toolkit). CLAN generated a comparable mapping of singular reads to other tools, and significantly outperformed in mapping simulated and real CLASH duplex reads, offering a potential application to other next-generation sequencing-based duplex-read-generating technologies.

INTRODUCTION

MicroRNA (miRNA) is a class of important regulator non-coding RNA that interacts with its target mRNAs through sequence complementarity (often observed at the 3′ UTR of the mRNA), subsequently regulating the corresponding mRNAs' translation level by degrading the targeted mRNAs (Bartel, 2009). Mature miRNA has a length between 21 and 25 nucleotides (nts), usually with the 28th nucleotide perfectly complementing its target mRNA, serving as the seed region of the binding. The miRNA-mRNA binding and subsequent mRNA degradation are facilitated by the RNA-induced silencing complex (RISC), a microRNA ribonucleoprotein complex. The Argonaute protein (AGO) within RISC contains two RNA-binding domains (PAZ and PIWI) that bind the miRNA and mRNA, respectively, and plays key roles in facilitating the miRNA-mRNA interaction. The biological function of the miRNA is often understood through the function of its targeted mRNAs. For example, cancer-causing miRNA point mutations or aberrant expression can lead to the positive regulation of the targeted cancer-causing genes or the negative regulation of the targeted cancer-repressing genes (Zhang et al., 2006; Volinia et al., 2006; Calin and Croce, 2006). As a result, elucidating the miRNA target is key to the understanding of miRNA function and its regulating biological processes.

MicroRNA targets can be identified in three ways primarily. First, they can be predicted computationally through models (Wang, 2016) that summarize, e.g., the sequence complementarity (Agarwal et al., 2015) and site accessibility (Kertesz et al., 2007) information. Although computational approaches have successfully recovered many genuine miRNA targets, their false-positive rates remain high. Second, the next-generation sequencing (NGS)-based CLIP (Crosslinking Immunoprecipitation)-seq on the AGO protein is capable of probing the potential miRNA-binding sites across the entire genome, but fails in specifying the targets of a specific miRNA family (Chi et al., 2009). Finally, experimental approaches can be used to validate a specific miRNA-mRNA interaction, yet in a low-throughput manner (Jin et al., 2013).

Recently, the Cross-linking Ligation and Sequencing of Hybrids (CLASH) technology was developed for high-throughput genome-wide probing of *in vivo* miRNA-mRNA interactions (see Figure 1). CLASH (Kudla et al., 2011; Helwak and Tollervey, 2014) first pulls down the interacting miRNA-mRNA strands through AGO immunoprecipitation, followed by the covalent cross-linking of the interacting miRNA-mRNA strands, and eventually sequences the cross-linked miRNA-mRNA duplex. As the covalent cross-linking

[1]Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA

[2]Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA

[3]Lead Contact

*Correspondence: cczhong@ku.edu

https://doi.org/10.1016/j.isci.2019.05.038

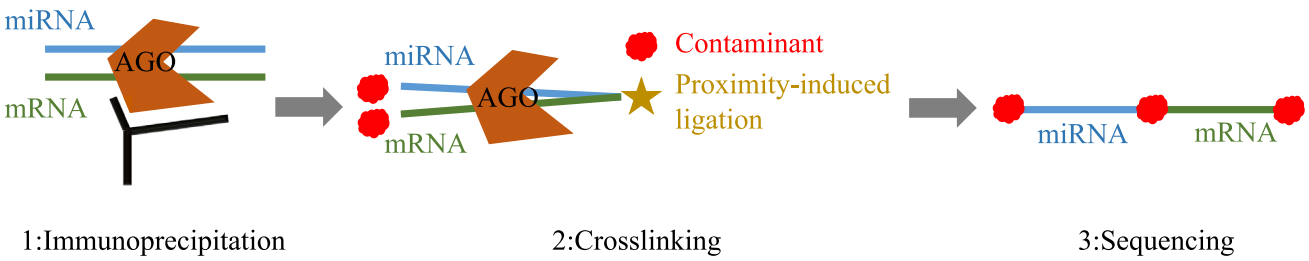


Figure 1. Schematic Illustration of the Generation Process of the CLASH Reads

Step 1 (immunoprecipitation): the interacting miRNA and mRNA are enriched through immunoprecipitation of the Argonaute (AGO) protein. Step 2 (cross-linking): the interacting miRNA and mRNA are covalently cross-linked through proximity-induced ligation. Potential contaminants (floating short nucleotide fragments) could be incorporated at the terminus of the miRNA or mRNA. Step 3 (sequencing): the bound AGO protein is washed and the cross-linked miRNA-mRNA duplex is subjected to standard library preparation and sequencing to generate the CLASH reads.

is proximity induced, the miRNA/mRNA strands could ligate to spatially adjacent free-floating nucleotide fragments (red clouds in Figure 1) before being cross-linked to each other. These nucleotide fragments could also be ligated to the terminus of the miRNA/mRNA strands. Intuitively, the mapping of the resulting duplex reads will reveal direct evidence for the corresponding miRNA-mRNA interaction. In practice, the inclusion of the random nucleotide fragment contaminants complicates the CLASH read mapping.

In the original CLASH analysis (Helwak et al., 2013; Helwak and Tollervey, 2014), BLAST (Altschul et al., 1997) was used to map the CLASH reads, leading to a 2%–3% mapping rate. The same group further tested BOWTIE2 (Langmead and Salzberg, 2012) as the aligner in a subsequent analysis pipeline called Hyb (Travis et al., 2014). Although BOWTIE2 greatly improved the computational efficiency of the CLASH read mapping, it was comparable with BLAST in terms of mapping sensitivity. The main reasons for the low-sensitivity mapping of the CLASH reads are (1) random nucleotide fragment contaminants lack apparent pattern and therefore are difficult to detect and can mislead the read mapping and (2) the resulting CLASH reads can be very short (owing to the intrinsic length of the AGO-binding site). For example, the average read length of a real CLASH dataset (SRR959751) is 20 nt (after adapter trimming), corresponding to 10 nts per miRNA/mRNA strand.

In addition to BLAST and BOWTIE2, other read-mapping tools such as BWA-MEM (Li and Durbin, 2009), STAR (Dobin et al., 2013), and HISAT2 (Kim et al., 2015), are expected to perform similarly because they share a similar read-mapping objective. Subsequent analyses (such as Hyb, Travis et al., 2014) take the mapping results as the input and attempt to prioritize confident miRNA-mRNA interactions. However, they cannot create new read mappings and hence cannot be used to tackle the low-sensitivity issue. A dedicated CLASH reads aligner is in demand.

In consideration of the unique CLASH read layout and the limitations of the existing alignment and mapping tools, we provide a novel formulation for the CLASH read-mapping problem. We term a single miRNA/mRNA an ***arm*** of the read, and we seek to find a single arm (when the miRNA-mRNA cross-linking fails) or two non-overlapping arms (when the miRNA and mRNA cross-linking is successful) from a read such that (1) each arm can be mapped to the reference database under a given sequence similarity threshold and (2) the total length of the mapped single or two arms is maximized. This formulation assumes that, under successful experimental conditions, bona fide miRNA/mRNA sequences should dominate contaminants in the CLASH reads. This problem is subsequently solved through efficient identification of the set of all candidate arms through Burrows-Wheeler Transformation (BWT)-assisted searches against the reference database, followed by a dynamic programming chaining algorithm to identify the arm(s) with maximized total length.

We implemented the algorithm into a program called CLAN (the CrossLinked reads ANalysis toolkit). We summarized the details of the CLAN algorithms in Figure S1 and the Transparent Methods section of the Supplemental Information. We benchmarked CLAN with popular alignment and mapping tools BLAST (Altschul et al., 1997), BWA-MEM (Li and Durbin, 2009), STAR (Dobin et al., 2013), HISAT2 (Kim et al., 2015), and BOWTIE2 (Langmead and Salzberg, 2012). All mapping tools are capable of handling spliced alignments, which approximates the duplex mapping problem we are attempting to address here.

Dataset	Arm Length	Total Length	Contaminant	Error	Source
duplex.[l_a].noInsert	l_a	$2l_a$	No	Q30	miRBase, TargetScan
duplex.[l_a].Insert	l_a	$l \geq 2l_a$	Yes	Q30	miRBase, TargetScan
singular.[l]	l	L	No	Q30	hg38 3′UTR

Table 1. The Simulated Benchmark Datasets

The variable l_a represents the length of each arm in the corresponding dataset, with values of 10, 12, 15, 18, and 20. The variable l represents the full length of each simulated read in the corresponding dataset.

On both simulated and real datasets, we have demonstrated that CLAN, BLAST, and BWA can satisfactorily map duplex CLASH reads, with CLAN outperforming the other tools by >25% F-score (a comprehensive measure of both sensitivity and specificity).

RESULTS

We constructed three test datasets to benchmark the performance of CLAN. The first dataset was generated by cross-linking *in silico* one arm randomly sampled from miRBase (Kozomara and Griffiths-Jones, 2011) and another from the TargetScan- (Agarwal et al., 2015) predicted targets. The lengths of each RNA arm, i.e. l_a, took series values of 10, 12, 15, 18, and 20. If the selected mature miRNA or miRNA target site has full length less than l_a, its complete sequence was taken. This dataset was thereafter referred to with a pattern "duplex.[l_a].noInsert." The second dataset further inserted random sequences to the cross-linked reads to simulate experimental contaminants; the contaminants were randomly placed in between or at the terminus of the two RNA arms. The total lengths of the resulted simulated reads, including the inserted contaminants, took series values of 25, 30, 35, 40, and 45, corresponding to the RNA arm lengths of 10, 12, 15, 18, and 20, respectively. This dataset was thereafter referred to with a pattern "duplex.[l_a].Insert." Finally, to test CLAN's performance on mapping singular reads, the third dataset was generated by sequencing *in silico* the human genome build 38 (hg38) 3′ UTR with lengths l, which took values of 20, 30, 40, and 50. This dataset was thereafter referred to with a pattern "singular.[l]." An error rate of 0.1% (Q30) was introduced for all datasets described above. Each dataset contained 1 million reads. See Table 1 for the summary of these benchmark datasets.

We tested CLAN, BLAST (Altschul et al., 1997), BWA-MEM (Li and Durbin, 2009), STAR (Dobin et al., 2013), HISAT2 (Kim et al., 2015), and BOWTIE2 (Langmead and Salzberg, 2012) on the simulated datasets, and a real CLASH dataset (SRR959751). All experiments were run on an in-house server equipped with an Intel(R) Xeon(R) CPU E7-4850 v4 @ 2.10 GHz and 1 TB physical memory. Details regarding the parameters chosen for each program can be found from the Supplemental Information. For each program, the same set of parameters was used for mapping all simulated datasets (singular and duplex) and the real CLASH dataset.

Mapping Performance of the Simulated Singular Reads

We selected CLAN, BLAST, BWA-MEM, STAR, HISAT2, and BOWTIE2 for the benchmark. To evaluate singular-read mapping, we define four mapping categories. The "***perfect***" category indicates that the read is uniquely mapped to the correct location; the "***multi***" category indicates that the read is mapped to multiple locations, and at least one of the mapped locations is correct; the "***wrong***" category indicates that the read is mapped, but none corresponds to the correct location; the "***miss***" category contains reads that are not mapped. The mapping results for the simulated singular datasets are summarized in Figure 2.

Among the programs that have been tested, BWA-MEM produced not only the highest number of perfect mappings (blue bars in Figure 2) but also the highest number of wrong mappings. The other programs, including CLAN, produced relatively more multi-mappings, but much fewer wrong mappings. Overall, CLAN performed reasonably well by generating the second-largest number of perfect mappings (following BWA) while maintaining a low level of wrong mapping rate.

We further analyzed the consistency among the reads perfectly mapped by CLAN, BLAST, and BWA-MEM (see the Venn diagrams in Figure 3). The overall consistency is high among all three programs. CLAN and

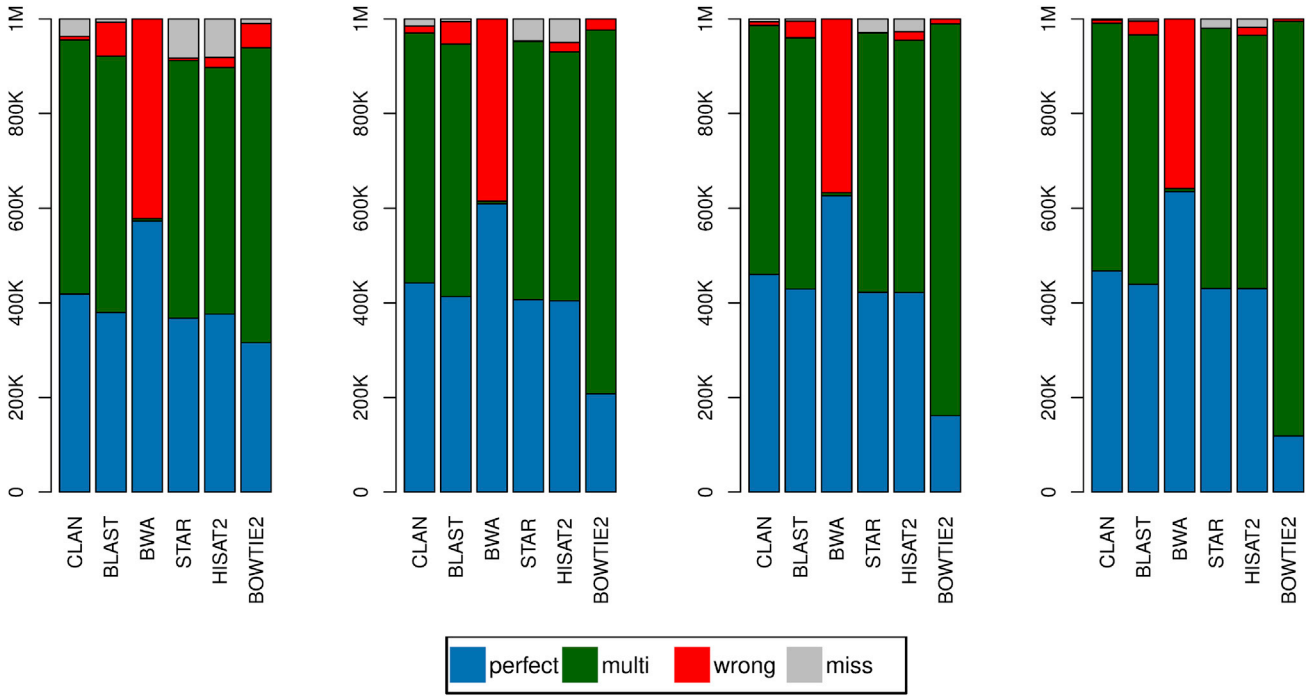


Figure 2. Performance of Different Programs when Mapping Simulated Singular Reads
From left to right: performance for simulated datasets singular.20, singular.30, singular.40, and singular.50.

BWA-MEM showed higher consistency when compared with BLAST, potentially because both of them employed the BWT data structure for alignment seeding.

Mapping Performance of the Simulated Duplex Reads

We further benchmarked the performances of CLAN, BLAST (Altschul et al., 1997), BWA-MEM (Li and Durbin, 2009), STAR (Dobin et al., 2013), HISAT2 (Kim et al., 2015), and BOWTIE2 (Langmead and Salzberg, 2012) on mapping the simulated duplex reads. The best and second-best hits (as measured by bit score) of BLAST alignment were taken to allow the consideration of both arms. To measure the performance of duplex read mapping, we define the following mapping categories: (1) ***perfect***: both arms are mapped correctly (>80% overlap with the ground-truth interval) and uniquely; (2) ***partial multi***: *one* arm is mapped correctly and uniquely, the other one is mapped to multiple locations, and the correct location is included in the multi-mapping; (3) ***both multi***: both arms are mapped to multiple locations, and both correct locations are included in the multi-mappings; (4) ***partial wrong***: one arm is mapped correctly and uniquely, and the other one is mapped (disregarding whether the arm is unique- or multi-mapped) but not to the correct location; (5) ***both wrong***: both arms are mapped (disregarding whether the arm is unique- or multi-mapped) but not to the correct locations; (6) ***partial miss***: one arm is mapped correctly and uniquely, and the other one is not mapped. (7) ***both miss***: both arms are not mapped. The performances of the programs are summarized in Figure 4.

We noted that CLAN, BLAST, and BWA-MEM produced satisfactory mappings in all datasets. CLAN and BLAST performed robustly, whereas the performance of BWA-MEM seemed to be hampered by the included sequence contaminants (comparing the top and bottom panels in Figure 4). In the datasets with added sequence contaminants, CLAN topped the performance among all tested programs.

We performed a similar analysis of the perfectly mapped duplex reads generated by CLAN, BLAST, and BWA-MEM. The Venn diagrams of their overlaps are shown in Figure 5. Similar to the singular read datasets, all programs generated consistent mappings. CLAN and BWA-MEM also showed the highest consistency for all datasets, except in duplex.18.Insert and duplex.20.Insert, where BWA-MEM seemed to be

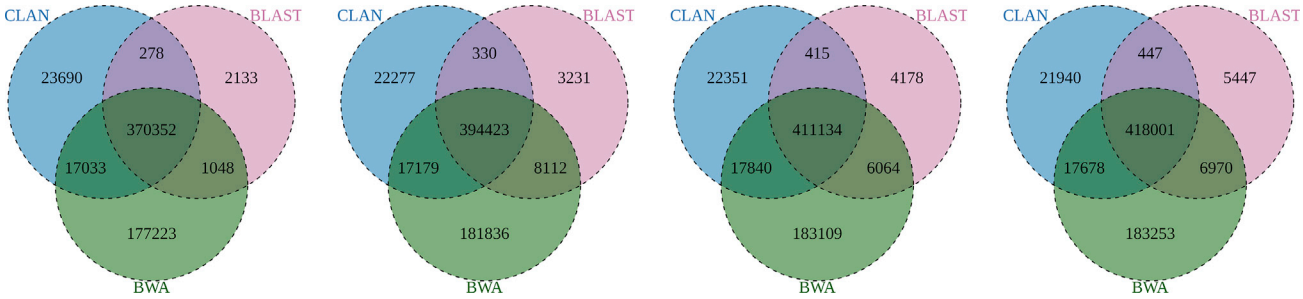


Figure 3. Venn Diagrams of the Perfectly Mapped Reads Generated by CLAN, BLAST, and BWA

From left to right: performances for simulated datasets singular.20, singular.30, singular.40, and singular.50.

hampered by the sequence contaminants. CLAN also mapped more duplex reads than the other aligners when sequence contaminants were present (Figure 5, bottom panels).

As the uniquely mapped reads contain the most reliable and interpretable information, we further evaluated the recall and precision of the uniquely mapped arms produced by CLAN, BLAST, and BWA-MEM. We define *TP* (True-positive) as the number of arms that are mapped correctly, *FP* (False-positive) as the number of arms that are mapped incorrectly, and *FN* (False-negative) as the number of arms that are not mapped or multi-mapped. We further define *Recall, Precision*, and *F-score* as:

$$Recall = \frac{TP}{TP + FN},\ Precision = \frac{TP}{TP + FP},\ F - score = \frac{2 * Recall * Precision}{Recall + Precision}.$$

The performance of CLAN and BLAST on the uniquely mapped reads for the simulated duplex read datasets are summarized in Table 2. CLAN demonstrated the best overall F-score in all datasets. CLAN also showed significantly higher recall and precision in mapping short CLASH reads (duplex.10.noInsert and duplex.10.Insert). For longer reads, BLAST showed the highest precision, followed by CLAN, which was marginally lower. BWA-MEM performed the best in terms of recall when mapping long CLASH reads without sequence contaminants. Taken together, the uniquely mapped reads produced by CLAN demonstrated high recall and precision, outperforming either BLAST or BWA-MEM.

Mapping a Real CLASH Dataset

We further compared the mapping produced by CLAN, BLAST, BWA-MEM, STAR, HISAT2, and BOWTIE2 on a real CLASH dataset (SRR959751), which was generated from a human kidney cancer cell line (Helwak et al., 2013). The first 2 million reads of the dataset (so that BLAST can finish within a reasonable amount of time) were mapped to a comprehensive database consisting of miRBase- (Kozomara and Griffiths-Jones, 2011) and TargetScan- (Reczko et al., 2011) predicted targets. As there was no ground-truth knowledge for the real dataset, we only considered reads that were mapped (either as singular or duplex reads) for >60% of their total lengths. For reads that were mapped as duplexes, we required that the two arms overlap for <4 nt (as recommended in Hyb). Furthermore, we required that one arm of duplex be mapped as miRNA, and the other as mRNA 3′ UTR. The counts of the successfully mapped reads (those passed the above filters) generated by different programs are summarized in Figure 6A.

The mapping results of the real CLASH dataset were largely consistent with the observation made from the simulated datasets. CLASH mapped the highest number of duplex reads, twice the second-performing mapper BWA-MEM (65,026 by CLAN and 31,598 by BWA-MEM, note that the y axis of Figure 6A is log scaled). STAR, HISAT2, and BOWTIE2 could barely map any duplex reads (although not zero), but as expected performed well in mapping singular reads. We further analyzed the mapping consistency of CLAN, BLAST, and BWA-MEM, and again, observed that the consistency was high among the three programs (see Figure 6B). Many reads were uniquely mapped by BLAST, with only their internal sequences mapped, a mapping configuration rarely seen in real CLASH reads.

Speed Comparison

Last, we benchmarked the speed of CLAN, BLAST, BWA-MEM, STAR, HISAT2, and BOWTIE2 on the simulated and real datasets. The wall-clock running time for the programs is summarized in Table 3. CLAN and

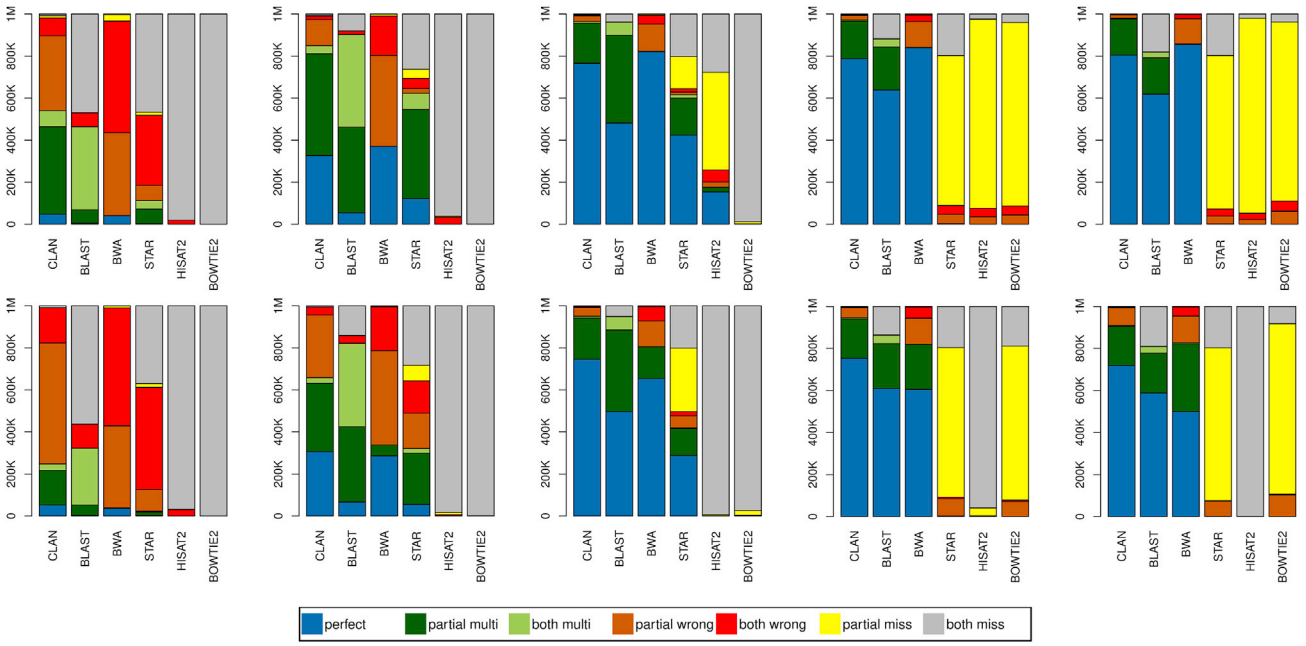


Figure 4. Performance Summary of CLAN, BLAST, BWA-MEM, STAR, and HISAT2 on the Simulated Duplex Datasets

The y axis represents the number of reads, and the X axis represents different programs. Top panels: duplex.[l_a].noInsert. Bottom panels: duplex.[l_a].Insert. From left to right: l_a with values of 10, 12, 15, 18, and 20.

BWA-MEM are much faster than BLAST, potentially because they utilize the BWT data structure to speed up the search. STAR, HISAT2, and BOWTIE2 are comparably faster, but cannot properly map the CLASH reads.

DISCUSSIONS

CLASH is an innovative NGS-empowered technology for high-throughput and unbiased *in vivo* probing of the miRNA-mRNA interactome. CLASH read mapping is a challenging computational problem that limits the broader applications of CLASH. Here we present a novel mapping algorithm, CLAN, for CLASH reads.

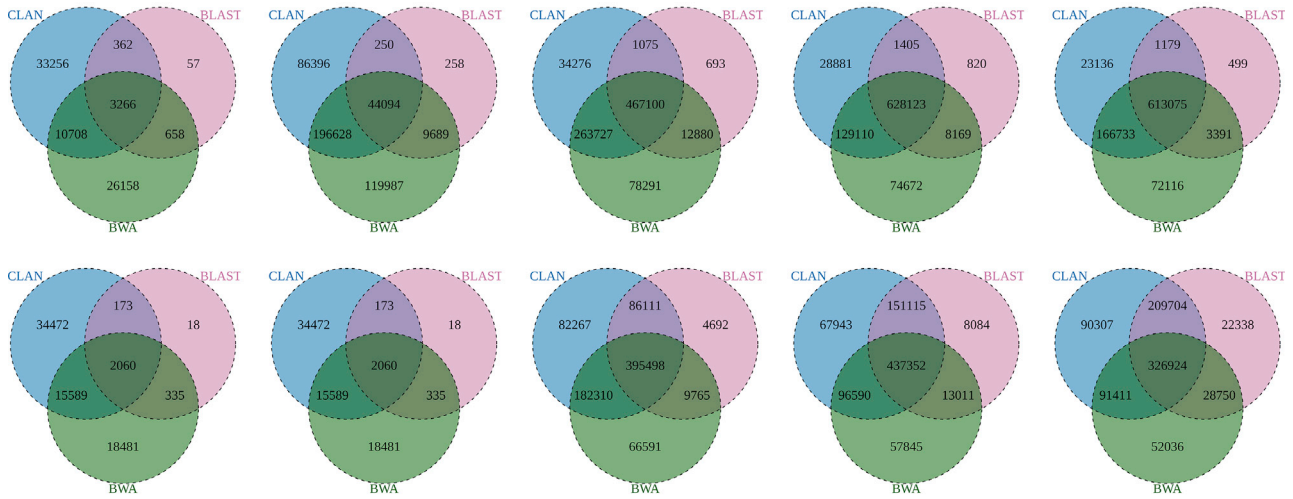


Figure 5. Venn Diagrams for the Perfectly Mapped Duplex Reads by CLAN, BLAST, and BWA

Top panels: duplex.[l_a].noInsert. Bottom panels: duplex.[l_a].Insert. From left to right: l_a with values of 10, 12, 15, 18, and 20.

Datasets	CLAN			BLAST			BWA		
	Re.	Pre.	F.	Re.	Pre.	F.	Re.	Pre.	F.
duplex.10.noInsert	**0.44**	**0.89**	**0.59**	0.04	0.79	0.07	0.28	0.31	0.30
duplex.15.noInsert	0.88	0.99	**0.93**	0.69	**1.00**	0.82	**0.91**	0.91	0.91
duplex.20.noInsert	0.90	0.99	**0.94**	0.71	**1.00**	0.83	**0.93**	0.93	0.93
duplex.10.Insert	**0.42**	**0.69**	**0.52**	0.03	0.62	0.05	0.26	0.27	0.27
duplex.15.Insert	**0.87**	0.98	**0.92**	0.69	**1.00**	0.82	0.81	0.90	0.86
duplex.20.Insert	**0.85**	0.96	**0.91**	0.68	**1.00**	0.81	0.73	0.92	0.82

Table 2. The Recall (Re.), Precision (Pre.), and F-score (F.) of the Uniquely Mapped Reads Generated by CLAN, BLAST, and BWA on Various Simulated Duplex Datasets

The highest performances in each category are in bold.

Benchmark results on both simulated and real CLASH datasets showed that CLAN outperformed the other major aligners and can map CLASH reads efficiently and accurately.

The important next step would be to reconstruct the miRNA-mRNA interactome from the CLAN mappings. We note that the reference dataset we used in this study comprises miRBase miRNA and TargetScan mRNA targets. This reference, however, does not allow the identification of novel miRNA-mRNA interactions. We plan to further expand the reference with the miRNA targetome revealed by AGO-targeted CLIP-seq experiments (Clark et al., 2014). Meanwhile, we also plan to incorporate additional miRNA-mRNA interaction features, such as the accessibility of the target site, minimum free energy of the binding, and length of the seed region, to compile high-confidence data on miRNA-mRNA interactions.

Recently, more NGS-based technologies that rely on the cross-linking of RNA strands have been developed, including CLASH (Kudla et al., 2011; Helwak and Tollervey, 2014), iPAR-CLIP (Jungkamp et al., 2011), MARIO (Nguyen et al., 2016), hiCLIP (Sugimoto et al., 2015), RPL (Ramani et al., 2015), PARIS (Lu et al., 2016), and LIGR-seq (Sharma et al., 2016). These technologies also generate duplex reads, whose mapping can potentially be improved by CLAN. We will further test CLAN on these datasets for a systematic and rigorous benchmark, to recommend technology-dependent best practices for CLAN.

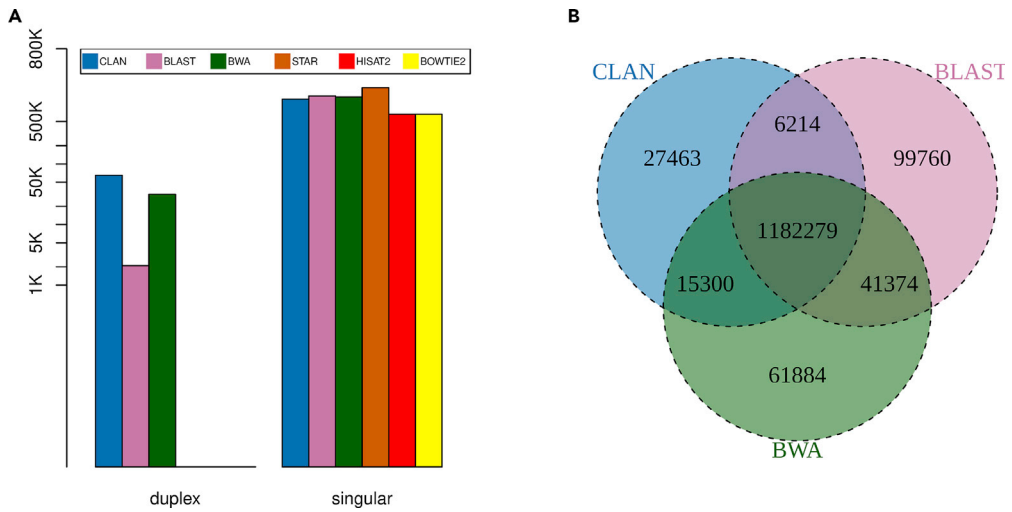


Figure 6. The Mapping Results of a Real CLASH Dataset SRR959751 Generated by the Listed Programs

(A) The number of mapped duplex and singular reads produced by the listed programs. Note that the y axis is log scaled.

(B) The Venn diagram of the mapped reads produced by CLAN, BLAST, and BWA-MEM.

Dataset	CLAN	BLAST	BWA	STAR	HISAT2	BOWTIE2
duplex.10.noInsert	27 s	1 h 19 min 41 s	20 s	14 s	8s	2s
duplex.20.noInsert	53 s	2 h 24 min 30 s	24 s	28 s	7s	3s
duplex.10.Insert	35 s	1 h 33 min 01 s	27 s	35 s	8s	2s
duplex.20.Insert	1 min	2 h 44 min 44 s	38 s	49 s	6s	3s
singular.20	1 min 49 s	12 h 52 min 28 s	39 s	3 min 48 s	12s	12s
singular.50	2 min 48 s	31 h 39 min 24 s	1 min 09 s	1 min 49 s	6s	28s
SRR959751 (1M reads)	14 s	14 min 26 s	20 s	28 s	16s	3s

Table 3. Comparison of the Wall-Clock Running Time of the Programs on the Simulated Datasets

Limitations of Study

We would like to note that CLAN is designed for mapping duplex reads that resulted from cross-linking experiments. To ensure high sensitivity, CLAN exhaustively looks for identical subsequences shared between the reads and the reference sequences to seed the alignment. This process is comparably slower than the seeding process implemented in many of the other aligners on the market. Therefore the users would expect slower running of CLAN when compared with the other tools (see Table 3 for running time benchmark). As a real duplex read dataset may contain majorly singular reads (>90%, see Figure 6), we recommend first using regular mapping tools to map and detect the singular reads, followed by mapping the remaining duplex reads with CLAN. In this way, the majority of the singular reads can first be filtered out using faster aligners, leaving much fewer reads to be aligned using CLAN.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

DATA AND SOFTWARE AVAILABILITY

CLAN is implemented in GNU C++, and is freely available from https://sourceforge.net/projects/clan-mapping.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2019.05.038.

ACKNOWLEDGMENTS

C.Z. was supported by the University of Kansas New Faculty General Research Fund allocation #2302114 and the National Science Foundation EPSCoR First Awards in Microbiome Research. S.Z. was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM102515. Publication costs have been funded by the National Science Foundation EPSCoR First Awards in Microbiome Research. The authors would also like to thank the reviewers for their insightful comments.

AUTHOR CONTRIBUTIONS

C.Z. and S.Z. conceived the study. C.Z. developed the algorithm, implemented the software, performed the benchmark experiments, and analyzed the results. C.Z. and S.Z. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 23, 2019
Revised: April 14, 2019
Accepted: April 22, 2019
Published: August 30, 2019

REFERENCES

Agarwal, V., Bell, G.W., Nam, J.W., and Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. Elife *4*, e05005.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Calin, G.A., and Croce, C.M. (2006). MicroRNA-cancer connection: the beginning of a new tale. Cancer Res. *66*, 7390–7394.

Chi, S.W., Zang, J.B., Mele, A., and Darnell, R.B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. Nature *460*, 479–486.

Clark, P.M., Loher, P., Quann, K., Brody, J., Londin, E.R., and Rigoutsos, I. (2014). Argonaute CLIP-Seq reveals miRNA targetome diversity across tissue types. Sci. Rep. *4*, 5947.

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.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. Cell *153*, 654–665.

Helwak, A., and Tollervey, D. (2014). Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). Nat. Protoc. *9*, 711–728.

Jin, Y., Chen, Z., Liu, X., and Zhou, X. (2013). Evaluating the microRNA targeting sites by luciferase reporter gene assay. Methods Mol. Biol. *936*, 117–127.

Jungkamp, A.C., Stoeckius, M., Mecenas, D., Grun, D., Mastrobuoni, G., Kempa, S., and Rajewsky, N. (2011). In vivo and transcriptome-wide identification of RNA binding protein target sites. Mol. Cell *44*, 828–840.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. Nat. Genet. *39*, 1278–1284.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. *39*, D152–D157.

Kudla, G., Granneman, S., Hahn, D., Beggs, J.D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. Proc. Natl. Acad. Sci. U S A *108*, 10010–10015.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Lu, Z., Zhang, Q.C., Lee, B., Flynn, R.A., Smith, M.A., Robinson, J.T., Davidovich, C., Gooding, A.R., Goodrich, K.J., Mattick, J.S., et al. (2016). RNA duplex map in living cells reveals higher-order transcriptome structure. Cell *165*, 1267–1279.

Nguyen, T.C., Cao, X., Yu, P., Xiao, S., Lu, J., Biase, F.H., Sridhar, B., Huang, N., Zhang, K., and Zhong, S. (2016). Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. Nat. Commun. *7*, 12023.

Ramani, V., Qiu, R., and Shendure, J. (2015). High-throughput determination of RNA structure by proximity ligation. Nat. Biotechnol. *33*, 980–984.

Reczko, M., Maragkakis, M., Alexiou, P., Papadopoulos, G.L., and Hatzigeorgiou, A.G. (2011). Accurate microRNA target prediction using detailed binding site accessibility and machine learning on proteomics data. Front. Genet. *2*, 103.

Sharma, E., Sterne-Weiler, T., O'hanlon, D., and Blencowe, B.J. (2016). Global mapping of human RNA-RNA interactions. Mol. Cell *62*, 618–626.

Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'ambrogio, A., Luscombe, N.M., and Ule, J. (2015). hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature *519*, 491–494.

Travis, A.J., Moody, J., Helwak, A., Tollervey, D., and Kudla, G. (2014). Hyb: a bioinformatics pipeline for the analysis of CLASH (crosslinking, ligation and sequencing of hybrids) data. Methods *65*, 263–273.

Volinia, S., Calin, G.A., Liu, C.G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., et al. (2006). A microRNA expression signature of human solid tumors defines cancer gene targets. Proc. Natl. Acad. Sci. U S A *103*, 2257–2261.

Wang, X. (2016). Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. Bioinformatics *32*, 1316–1322.

Zhang, L., Huang, J., Yang, N., Greshock, J., Megraw, M.S., Giannakakis, A., Liang, S., Naylor, T.L., Barchetti, A., Ward, M.R., et al. (2006). microRNAs exhibit high frequency genomic alterations in human cancer. Proc. Natl. Acad. Sci. U S A *103*, 9136–9141.

ISCI, Volume *18*

Supplemental Information

Accurate and Efficient Mapping of the Cross-Linked microRNA-mRNA Duplex Reads

Cuncong Zhong and Shaojie Zhang

Supplementary Figure

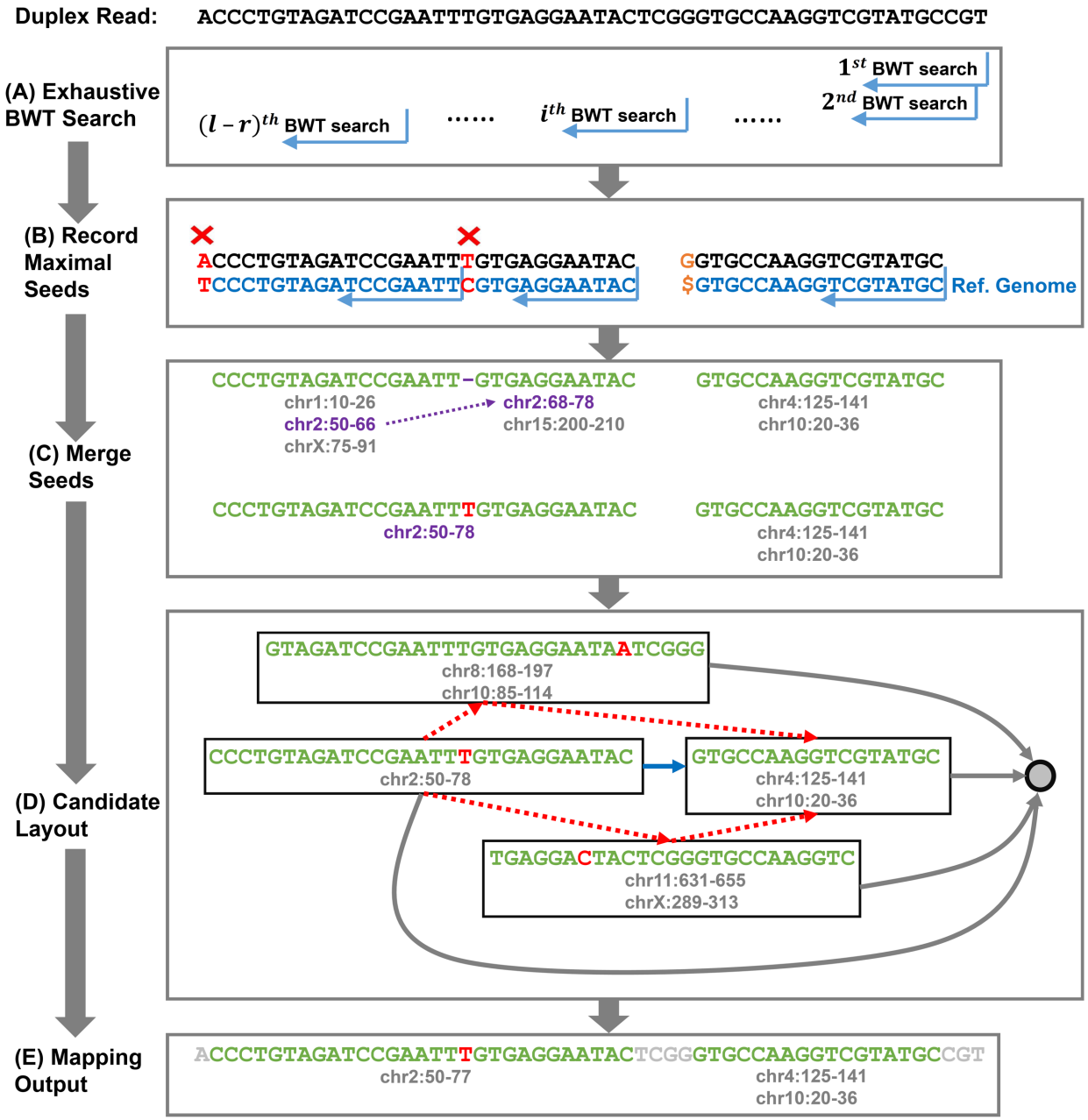


Figure S1: An overview of the CLAN algorithm with an artificial example. Related to Figure 2, Figure 3, Figure 4, Figure 5, and Figure 6.

Transparent Methods

The CLAN algorithm

We formulate the duplex read mapping problem as finding two non-overlapping arms whose total mapping length is maximized (see Figure S1 for the high-level summary of the CLAN algorithm).

CLAN first identifies a set of seeds that satisfy: (1) at least r-nt long (default r =10); (2) mapped to less than m reference locations (default m =100); (3) mapped to the reference database perfectly (no mismatch/gap). To identify the seeds, CLAN constructs the Burrows-Wheeler Transformation (BWT) and the corresponding Full-text index in Minute space (FM-index) from the reference database. Then, for a read s with length l, CLAN exhaustively searches all of its prefixes backward against the indexed database (Figure S1A), until a mismatch is encountered or the prefix/reference is exhausted (Figure S1B). The perfectly mapped substring is called a ***seed***.

The second step is to merge seeds that are potentially broken due to errors/SNP/indels, using a gapped-BLAST-like two-hit strategy (e.g. the purple intervals in Figure S1C). We assume that each RNA arm can be broken by no more than k such non-consecutive mismatches (by default k=1). To describe the merging step, let an arbitrary seed $s(i,j)$ be mapped to a set of genomic locations, with the xth denoted as $T(w^x, z^x)$. Here, T is the reference database, and w^x and z^x are the start and end of the mapped genomic interval. For two non-overlapping seeds $s(i_1, j_1)$ and $s(i_2, j_2)$ (without loss of generality, assume $i_2 > j_1$), CLAN attempts to merge the seeds by looking for two adjacent mapped locations, i.e. $T(w_1^x, z_1^x)$ and $T(w_2^y, z_2^y)$, such that:

$$(1)\ 1 \le i_2 - j_1 \le h,\ (2)\ 1 \le w_2^y - z_1^x \le h, \text{ and } (3)\ \left|\left(w_2^y - z_1^x\right) - (i_2 - j_1)\right| \le g.$$

The first two conditions ensure that the two seeds are adjacent in both the duplex read and the reference (at most h-nt apart, default h=5); the third condition ensures that the gap (if any) for the corresponding alignment is small (default value of g is set to 5). CLAN will exhaustively test all combinations of mapped genomic locations, and merge both seeds into a ***candidate*** (i.e., $s(i_1, j_2)$, with a new mapping location $T(w_1^x, z_2^y)$, see Figure S1C, the purple genomic interval in the second row) if all conditions are satisfied. CLAN progressively iterates this merging process for k times to allow k mismatches. To maximize sensitivity, the original set of seeds, disregarding whether they were subsequently merged, were kept in the final candidate set.

The third step is to find f non-overlapping arms with maximized total mapping length (Figure S1D). Note that the parameter f is reserved to allow future consideration of the crosslinking of more than two RNA

molecules, for example, the crosslinking of the small nuclear RNAs U1, U2, U4, U5, and U6 that corporate in the spliceosome complex. (We note that this is not currently supported by CLASH.) For CLASH duplex read mapping, f is set to 2. Conceptually, the candidates and their relationships can be represented by a directed acyclic graph (DAG). In the graph, each node corresponds to a candidate (Figure S1D). Partially order the candidates by the increasing order of their starting locations, and break ties in decreasing order of their ending locations. Also, define two nodes as ***compatible*** if their corresponding candidates do not overlap. For two arbitrary nodes u and v, a $\{u, v\}$ edge (Figure S1D) is added if the following three conditions are satisfied: (1) u is partially ordered before v; (2) u and v do not overlap; and (3) no node exists between u and v and is simultaneous compatible with both of them. (For example, the red dotted edges in Figure S1 do not exist because the corresponding nodes overlap with each other.) For each edge (u, v), CLAN sets its length $l_{\{u,v\}}$ as follows: $l_{\{u,v\}} = l_u - c$. The parameter c is the penalty (default c=2) for including an additional arm in the solution set, which prioritizes the mapping as a singular read unless necessary. Finally, a dummy node d succeeding every other node is included (Figure S1D), receiving an incoming edge from each node u with an edge weight of l_u (Figure S1D, gray solid edges; we do not include the penalty c because d does not correspond a real candidate). In this case, the problem of finding two non-overlapping candidates whose total length is maximized can be transformed as finding the longest path in the DAG that involves no more than f edges.

CLAN solves this problem using a dynamic programming (DP) approach. Denote the resulting DAG as $G = (V, E)$, where V corresponds to the node set and E corresponds to the edge set. Also let $L[v, f]$ record the length of the longest path that ends with v and involves at most f edges. CLAN computes $L[v, f]$ with the following recursive functions:

$$L[v, f] = \max \begin{cases} \max\limits_{u:\{u,v\}\in E} \{L[u, f-1] + l_{\{u,v\}}\} \ (\text{if } f > 0) \\ \max\limits_{u:\{u,v\}\in E} \{L[u, f]\} \\ 0 \end{cases}$$

The first condition considers cases where the path is extended from u to v with the candidate arm u taken. The second condition considers the case that u is not taken into the solution. The third condition corresponds to boundary cases where v is the starting node of the path. The final solution can be found in $L[d, f]$, where d is the dummy node. The output of the mapping contains the selected arms and their corresponding locations in the duplex read and the reference (Figure S1E).

Finally, we analyze the time complexity of CLAN. Recall that l is the length of a duplex read. Clearly, each BWT search of an l-long sequence against the reference requires $O(l)$ time. Because CLAN searches every prefix of the duplex read and there are at most l prefixes, the total time required for the exhaustive BWT

search step is thus $O(l^2)$. For candidate merging, CLAN tests the merging of every pair of candidate seeds in the worst-case scenario, which leads to an $O(m^2 l^2)$ complexity, where m is the maximum number of reference locations associated with each candidate (a constant default to 100). CLAN hashes the mappings according to their genomic locations into a fixed number of bins, and only attempts to merge candidates within the same bin. This makes the merging step practically efficient. Finally, for the DP-based chaining, each node v has at most l nodes that precede it; as a result, computing each value in L requires $O(l)$ time. Since there are $O(fl)$ entries in L, and the total time required for the chaining step is thus $O(fl^2)$ (where f is a constant). Taken together, CLAN requires $O(l^2)$ to map a single duplex read. Note that the duplex read length l is determined by the length of the miRNA-mRNA binding site, therefore can be viewed as a constant. CLAN thus requires a constant time to map a single duplex read, and the overall running time is linear with respect to the throughput of the experiment (or the number of reads in the dataset).

To summarize, we highlight the major algorithmic contributions of CLAN: (1) CLAN employs a DP-based chaining step to detect duplex mappings with desirable configuration, which prevents the mapping of the internal sequence of a read, while leaving long segments of prefix and suffix unmapped; (2) CLAN exhaustively searches for all seeds, which maximizes the sensitivity that is currently lacking in CLASH read mapping; and (3) CLAN accounts for mismatches through merging seeds, rather than directly identifying imperfect seeds. This reduces the computational cost of CLAN's seeding step and makes CLAN efficient overall.

Parameters used for running different programs

(Parameters that are not listed here were used as the defaults.)

CLAN:

```
-p 2 -v 2 -m 100 -t 16 -c 0.6
```

BLAST:

```
-num_threads 16 -task blastn-short
```

BWA-MEM:

```
-k 10 -r 1 -L 0 -T 10 -t 16
```

STAR:

Indexing:

```
--runThreadN 16 --runMode genomeGenerate --genomeSAindexNbases 10
```

Align:

```
--outSAMattributes All -runThreadN 16 --outFilterMismatchNmax 999 --alignIntronMax 1000000 --outFilterMatchNmin 8 --seedSearchStartLmax 10 --chimSegmentMin 10 --chimSegmentReadGapMax 10 --chimScoreJunctionNonGTAG 0 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0
```

HISAT2:

```
-p 16 --score-min L,0,-1 --pen-noncansplice 0 --no-temp-splicesite --min-intronlen 20 --max-intronlen 1000000 --pen-canintronlen G,0,0 --pen-noncanintronlen G,0,0
```

BOWTIE2:

```
-D 20 -R 3 -N 0 -L 16 -k 20 --local -i S,1,0.50 --score-min L,18,0 --ma 1 --np 0 --mp 2,2 --rdg 5,1 --rfg 5,1 -p 16
```