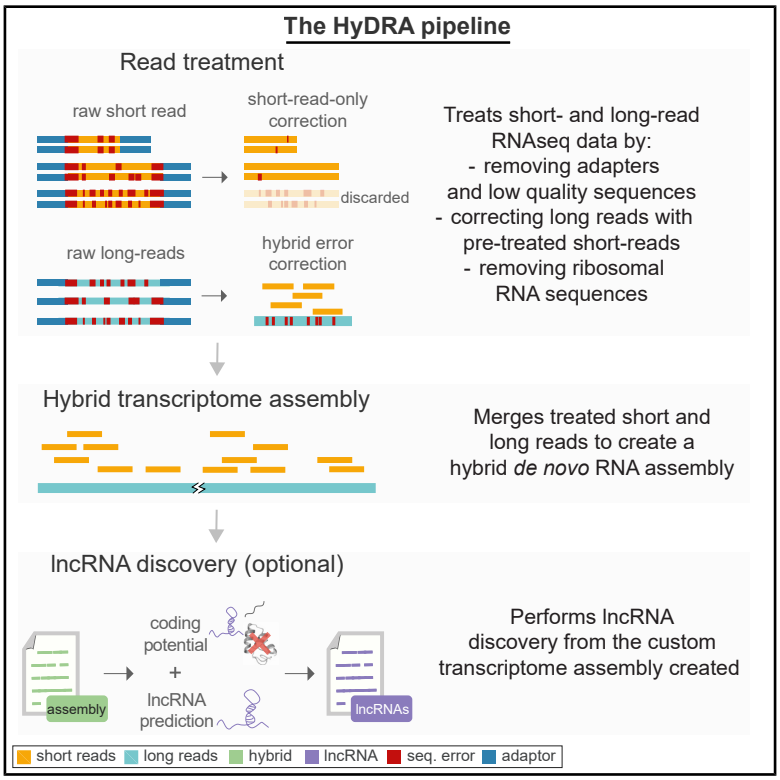
iScience

Article

HyDRA: A pipeline for integrating long- and short-read RNAseq data for custom transcriptome assembly

Graphical abstract



Highlights

- HyDRA leverages the accuracy of short reads and the resolution of long reads
- HyDRA supports the discovery of high-confidence transcripts in specific conditions
- Benchmarking showed HyDRA to outperform competing methods by up to 80%
- We identified >50 k high-confidence lncRNAs from the HyDRA ovarian metatranscriptome

Authors

Isabela Almeida, Xue Lu, Stacey L. Edwards, Juliet D. French, Mainá Bitar

Correspondence

isabela.almeida@qimrb.edu.au (I.A.), maina.bitar@qimrb.edu.au (M.B.)

In brief

Molecular biology; Bioinformatics; Systems biology

Almeida et al., 2026, iScience *29*, 116131
June 19, 2026
https://doi.org/10.1016/j.isci.2026.116131

CellPress

iScience

CellPress
OPEN ACCESS

Article

HyDRA: A pipeline for integrating long- and short-read RNAseq data for custom transcriptome assembly

Isabela Almeida,[1,2,*] Xue Lu,[1] Stacey L. Edwards,[1,2,3] Juliet D. French,[1,2,3,4] and Mainá Bitar[1,2,3,4,5,*]
[1]Cancer Program, QIMR Berghofer, Brisbane, QLD 4029, Australia
[2]Faculty of Health, Medicine and Behavioural Sciences, The University of Queensland, Brisbane, QLD 4006, Australia
[3]Faculty of Health, Queensland University of Technology, Brisbane, QLD 4006, Australia
[4]These authors contributed equally
[5]Lead contact
*Correspondence: isabela.almeida@qimrb.edu.au (I.A.), maina.bitar@qimrb.edu.au (M.B.)
https://doi.org/10.1016/j.isci.2026.116131

SUMMARY

Short-read RNA sequencing (RNAseq) remains a cornerstone for transcriptome profiling, but is limited in reconstructing full-length transcripts and capturing transcript diversity. While long-read RNAseq spans entire transcripts and resolves complex structures, this technology is hindered by its high error rates. In parallel, noncoding RNA transcripts remain underrepresented in current references. Here, we present hybrid *de novo* RNA assembly (HyDRA), a pipeline that integrates the accuracy of short reads with the structural resolution of long reads to produce more complete *de novo* transcriptome assemblies. Benchmarking showed HyDRA to outperform existing methods by up to 40%. Using the HyDRA human ovarian metatranscriptome, we identified >50,000 high-confidence long noncoding RNAs, most of which have not been previously detected using traditional methods. Although long-read RNAseq is advancing, the vast availability of short reads ensures HyDRA's ongoing role in capturing high-confidence, cell-type-specific transcripts and advancing our understanding of transcriptomic complexity and the noncoding genome.

INTRODUCTION

Short-read RNA sequencing (RNAseq) has revolutionized the transcriptomic era due to its high-throughput, affordability, and low error rates.[1] However, a limitation of short-read RNAseq lies in its dependency on fragmenting the native transcript molecules. Reassembling and quantifying these sequenced reads, typically of ~50 to ~500 nt in length, still poses significant computational challenges.[2] The majority of RNAseq studies measure the expression of genes and transcripts by mapping the sequenced reads to a reference annotated transcriptome and removing those that fail to map. Notably, reference transcriptomes such as those annotated by ENSEMBL and GENCODE are far from complete, leaving a large proportion of transcripts unquantified by standard RNAseq analysis methods.[3] To overcome these limitations, a *de novo* custom transcriptome assembly can be generated and used in the substitution of a reference, to reconstruct the sequenced fragments of transcripts expressed in the sample of interest in the substitution of a reference. Mapping sequenced reads onto a custom transcriptome allows the quantitation of expression levels from both annotated and unannotated transcripts.[4] However, to date, only a small fraction of publications make use of *de novo* custom assemblies, accounting for ~1% of the total PubMed publications that use RNAseq. In addition, *de novo* assemblies obtained using only short reads cannot accurately resolve all RNA transcripts.[5]

Long-read sequencing platforms such as those from Pacific Biosciences (PacBio) and Oxford Nanopore Technology (ONT) have the potential to produce reads that span the full-length of transcripts.[6] These platforms can perform end-to-end sequencing of single complementary DNA (cDNA) or RNA molecules, generating long reads that ameliorate the issues caused by transcript fragmentation.[7] ONT platforms include a range of devices in which single molecules thread through a nanopore containing a nanoscale sensor able to detect each nucleotide within a single run.[8] The produced long reads are one order of magnitude longer than typical short reads, providing better resolution of splice junctions, increasing correct isoform identification and the discovery of unannotated transcripts.[8] However, compared to short reads, long reads have much higher basecalling error rates[7,9,10] and remain a costly and comparatively less used technology. Importantly, although both long-read sequencing and basecaller technologies are continually evolving, most facilities are not equipped to support long-term storage of the large raw data files due to associated costs. Therefore, most users cannot recall previously sequenced reads when an improved basecalling algorithm is released to increase the quality of long-read data.

Emerging studies have shown that hybrid transcriptome assembly approaches, which integrate short- and long-read RNAseq data, are more accurate than approaches that use data from either

method alone.[11–16] Considering that the average human transcript length is one kilobase (kb) long[17] and that long-read sequences are on average 1–3 kb in length,[18,19] long reads should capture the majority of human transcripts within a single read and ideally bypass the need for fragment assembly. However, considering the low quality of long reads,[7,9,10] it is still recommended to correct for intrinsic sequencing errors.[9] Although there are different strategies to achieve this, a pre-assembly hybrid error correction using both short and long reads was recently shown to be the best-performing method.[10] Additionally, information from both long and short reads may be integrated at the assembly stage, to help reconstruct different isoforms.[20] To the best of our knowledge, none of the available tools fully benefit from the two types of read integration, but instead adopt either a hybrid-correction-only or a hybrid-assembly-only approach.

Notably, for long noncoding RNAs (lncRNAs), which constitute the largest class of unannotated RNA transcripts, their lower abundance in bulk tissues and high content of repetitive elements mean the assembly step is even more challenging.[21–23] The few lncRNA-focused hybrid assembly studies that have been performed indicate that RNAseq data integration can enhance the accuracy and reliability of lncRNA discovery.[24,25] However, no automated method for hybrid *de novo* assembly to date allows for accurate lncRNA discovery.

To address the need for an efficient method to perform comprehensive discovery of unannotated transcripts, we developed hybrid *de novo* RNA assembly (HyDRA), a true-hybrid pipeline that integrates short- and long-read RNAseq data for *de novo* transcriptome assembly, with additional steps for lncRNA discovery. Our pipeline combines read treatment, assembly, filtering, and parallel quality control (QC) steps to ensure the reconstruction of high-quality transcripts. Comprehensive benchmarking showed HyDRA to outperform best-in-class short-read-only, long-read-only, and hybrid *de novo* transcriptome approaches by up to 40%, having identified more than 50,000 high-confidence lncRNAs. In contrast with long-read sequencing, a vast amount of short-read RNAseq data is readily available for many species, tissues, and conditions. Pipelines such as HyDRA can make the best use of available data in its totality, allowing users to achieve high-quality transcriptome assemblies while long-read sequencing technologies continue to advance. We anticipate that HyDRA will facilitate the generation of sample- and condition-specific custom transcriptomes, providing a valuable resource for expression analyses across different cell types and tissues.

RESULTS AND DISCUSSION

Overview of the HyDRA pipeline

We developed HyDRA (Figure 1A), a true-hybrid pipeline that integrates bulk short- and long-read RNAseq data for generating custom transcriptomes. This is achieved through (1) read treatment steps to correct sequencing errors by treating low-frequency *k*-mers and removing contaminants (e.g., adaptors and reads from ribosomal RNAs (rRNAs)), (2) steps to *de novo* assemble the filtered and corrected reads and further process the resulting assembly, and (3) optional steps to discover a high-confidence set of lncRNAs supported by multiple machine-learning model predictions (Figures 1B–1D). We consider HyDRA to be a true-hybrid *de novo* transcriptome assembly pipeline given that it combines short- and long-read data at two stages, read correction and assembly, and because it does not rely on reference files to reconstruct transcripts. This section contains a detailed explanation of HyDRA. Additional details from development and implementation, including the tools and algorithms underlying each step, recommended parameters, and thresholds are also available in the supplementary files (Table S1) and on the HyDRA GitHub repository[26] (https://github.com/isabela42/HyDRA).

Input files

There are several input files required to run HyDRA. In summary, it requires ONT long reads as input files together with short-read data. Users are advised to use deep RNAseq data for improved lncRNA discovery. Although users may use PacBio long-read data, this will require them to refrain from using Porechop to remove adapter sequences and to modify the Cutadapt command-line. Additionally, users are required to provide a series of step-specific input files (see the complete user guide), e.g., adapter sequences and experimentally confirmed lncRNA fasta files. Although some of these are reference files, the steps in which they are used do not alter the inputs for generating the *de novo* transcriptome assembly, rather providing informative quality metrics to the user, such as flagging for contamination.

Pipeline availability

HyDRA is available in a range of different script options, including (1) a Dockerfile to build a Docker container image with all tools/versions used in this study (2) step-by-step command lines files to guide users that wish to run the pipeline on a different system and/or with the Docker container; and (3) the BASH scripts that write and submit portable batch system (PBS) jobs. Providing a Dockerfile dramatically reduces implementation time for users by eliminating manual dependency installation and environment configuration across diverse computer environments. Additionally, for users without access to an infrastructure that supports Docker, the Docker container image itself has also been provided and can be converted into a Singularity/Apptainer container image, as described on the HyDRA GitHub repository.

User guide

A complete user guide is available within the HyDRA GitHub repository[26] (https://github.com/isabela42/HyDRA), including all pipeline required tools/versions, input files, and detailed pipeline files usage guidelines. Importantly, the computational resource usage varies with input data size and complexity for each tool. Therefore, to some extent, users can control the computational resources usage according to the computational power available. Additionally, to offer users a baseline, we have indicated in all scripts the resources used for pipeline development. In line with this, although we have used default parameters for most tools with minor adjustments (Table S1), HyDRA provides a well-defined set of options to be used with each tool, optimizing its behavior for its main steps. If desired, users may easily edit these BASH scripts to make use of a complex range of options for the different tools integrated in HyDRA.[26]

Read treatment

Sequencing errors are known to introduce artificial nodes in de Bruijn graphs during *de novo* isoform resolution[27] and interfere

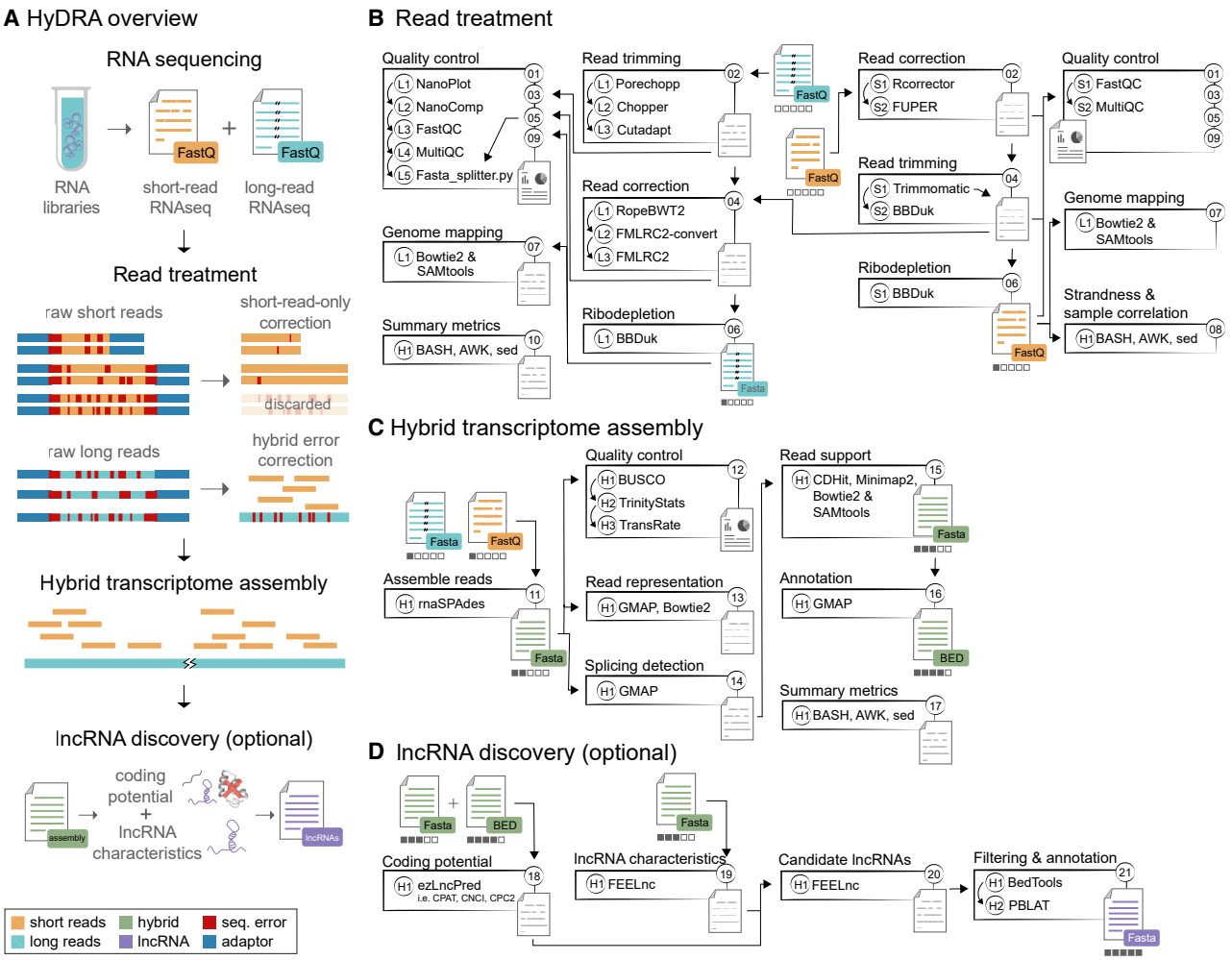


Figure 1. The HyDRA pipeline

(A) Overview of the pipeline, from RNA library preparation to sequencing and availability of raw FASTQ files for both short- and long-read samples.

(B) Both short and long reads first undergo extensive quality control and processing, including hybrid error-correction of long reads and short-read-only correction of short reads. These steps are important to assess low-frequency *k*-mers for error correction and to remove contaminants (e.g., adaptors and reads from ribosomal RNAs).

(C) Treated reads undergo a hybrid *de novo* transcriptome assembly and further filtering and quality assessment. HyDRA generates summary metrics for steps in (B and C).

(D) Optional steps can be performed for the discovery of high-confidence lncRNAs.

with all downstream steps.[28,29] In addition to correcting sequenced reads, common read processing practice includes removing adaptor sequences identified during the raw quality assessment of the raw data.[30] As traditional tools optimized for short-read data fail to correctly treat longer sequences,[7,9,31] HyDRA includes scripts and subroutines carried out by best-performing tools specifically designed to process these data separately (Figure 1B). As a result, HyDRA's read treatment phase includes 38 scripts that perform the first ten steps. The short-read processing steps of HyDRA follow our previously published *de novo* assembly pipeline, which is currently the best-in-class approach for short-read-only *de novo* transcriptome assembly.[4] Processed in parallel, long-read treatment steps are dependent on the pre-processed short-reads for hybrid error correction using FMLRC2 v.0.1.7.[10,32] This tool was selected for being the best ranked at benchmarking experiments,[32] including an in-house assessment where multiple tools failed to perform at the same level (see development-and-implementation at the HyDRA GitHub repository - https://github.com/isabela42/HyDRA). QC routines are interspersed throughout the read treatment steps to guarantee high read quality, including an in-house Python script (fasta_splitter.py) to assess long-read length, allowing the user to implement personalized cut-offs to exclude ultra-long reads (e.g., > 35 kb). These QC steps were designed to enhance the quality of input read data and are performed after each key processing step.

Custom transcriptome assembly

After careful consideration and benchmarking of alternatives (see development-and-implementation at the HyDRA GitHub

repository - https://github.com/isabela42/HyDRA), we selected rnaSPAdes v.3.14.1[12] as the assembler for HyDRA. Of note, rnaSPAdes was specifically designed to integrate short and long RNAseq reads and is the only available assembler that uses a genome-independent process (Figure 1C and Table S1). This assembler was developed from the foundational algorithms SPAdes and hybridSPAdes, enabling the integration of both paired-end short-read RNAseq data and single-end long-reads, from either PacBio or ONT. This approach facilitates the construction of a high-quality transcriptome assembly that represents full-length transcripts and their alternative isoforms.[12] Next in the HyDRA pipeline, a step is included to remove highly redundant transcripts and differentiate between multiexonic and monoexonic sequences in the assembly. This allows users to set appropriate read support thresholds for each subset of transcripts, with monoexonic transcripts requiring higher read support to differentiate from sequencing noise or genomic DNA contamination. HyDRA uses reads per kilobase per million (RPKM) values from independent short- and long-read alignments to estimate read support (i), with user-defined thresholds for filtering.

$$\text{(i) RPKM} = \frac{\text{Read count} \times 10^9}{\text{Total reads} \times \text{Feature length}}$$

Assembled transcripts are then aligned to the reference transcriptome to identify unannotated transcripts in the custom transcriptome. Similar to our QC routine for input reads, we use biologically supported quality evaluation tools (BUSCO v.20161119,[33] Trinity Stats v.2.8.4[34] and TransRate v.1.0.3[28]), to assess completeness and other metrics that characterize the generated custom transcriptome. With that, this section of the pipeline includes nine scripts performing seven steps.

LncRNA discovery

A custom transcriptome assembly can help in the discovery of a variety of transcript types, with lncRNAs representing a substantial portion of the unannotated transcriptome. LncRNAs are highly specific to different tissues, cell types, conditions, and developmental stages.[4,35] Despite their significance, lncRNAs are often underrepresented in transcriptional studies due to their lack of annotation in reference transcriptomes. This is partially due to short-read RNAseq's inherent biases and inability to capture full-length transcripts. Using a combination of long and short reads, HyDRA is well-equipped to support the annotation of lncRNAs. We have therefore included four optional steps after the core assembly, allowing HyDRA to perform lncRNA discovery. Using a combination of three machine learning models from ezLncPred v.1.0,[36] i.e., CPAT, CNCI, CPC2, we first predict the coding potential of the transcripts identified in the assembly. These transcripts are also assessed in parallel by FEELnc v.0.2 to detect lncRNAs.[37] FEELnc is a suite of machine learning algorithms that requires the user to supply both a reference annotation of protein-coding and experimentally confirmed lncRNA transcripts (e.g., GENCODE annotation) to train the model in classifying transcripts assembled by HyDRA. A transcript is then considered to be a candidate lncRNA based on FEELnc's prediction in combination with the predicted absence of detected coding potential from at least two different ezLncPred tools. These tools were used in combination to decrease false positives among the discovered lncRNAs.[4] To further remove false-positive lncRNAs (e.g., transcripts that match annotated protein-coding transcripts), HyDRA aligns the sequences of candidate lncRNAs to the reference transcriptome. Candidate lncRNAs matching annotated protein-coding transcripts with at least 75% identity and a minimum bidirectional overlap of 85% on either strand are identified as false-positives and removed. Coordinates of candidate lncRNAs are also intersected with annotated protein-coding genes using BedTools,[38] resulting in a final set of high-confidence lncRNAs. Finally, HyDRA maps the candidate lncRNAs to a comprehensive database of confirmed lncRNAs that can be the internal default (containing 112,439 lncRNAs from multiple sources, as described in Bitar et al.[4]) or a user-defined database. This allows the user to pinpoint which lncRNAs have been detected in the custom assembly, and which were already known, either from the reference transcriptome or from additional databases (see development-and-implementation at the HyDRA GitHub repository - https://github.com/isabela42/HyDRA).

HyDRA improves the quality of both short- and long-sequenced reads

HyDRA was developed and benchmarked using data obtained from short- and long-read RNAseq on normal and tumorigenic data, including normal primary and immortalized fallopian tube secretory epithelial cells (FTSEC) and ovarian surface epithelial cells (OSEC), breast cancer and uveal melanoma datasets (Table S2). All three datasets presented similar read and custom assembly quality metrics (Tables S3 and S4). Therefore, we will only describe in detail the quality metrics obtained from one of these, the normal primary and immortalized FTSEC and OSEC dataset, which we generated in-house. We used the Illumina NovaSeq 6000 to generate short reads from the normal primary and immortalized FTSEC and OSEC samples, which have the lowest error rates for high-throughput sequencing.[39] For long-read sequencing, we used ONT with current error rates predicted at 5–10%.[9,40] Common practice with long-read RNAseq includes the removal of reads below a mean minimum Phred quality of 7 (Q7), which is a much lower threshold than for short-read data. In our dataset, all long reads were over Q7, with 93% of them surpassing Q10 (Table S3).

The short-read treatment steps begin with the correction of raw reads and subsequent trimming. The majority of the short reads survived the correction step (average of 9.16% uncorrectable reads; Figure 2 and Table S3; development-and-implementation at the HyDRA GitHub repository - https://github.com/isabela42/HyDRA), and about 60% of the sequence ends were above the minimum quality set for trimming Q30 (Table S3), indicating a high quality of corrected and trimmed short reads. Median short-read quality measured in the Phred scale increased from 35.65 to 36.12 after treatment steps were performed (Figure S1 and S2, Table S3), with a concomitant decrease in the calculated error rate from one base in ~3,000 to one in ~4,000. In HyDRA, due to the prerequisites of the selected tools, the long-read treatment steps follow the opposite order, with trimming (adaptor removal followed by quality trimming) performed before correction. Median long-read quality measured

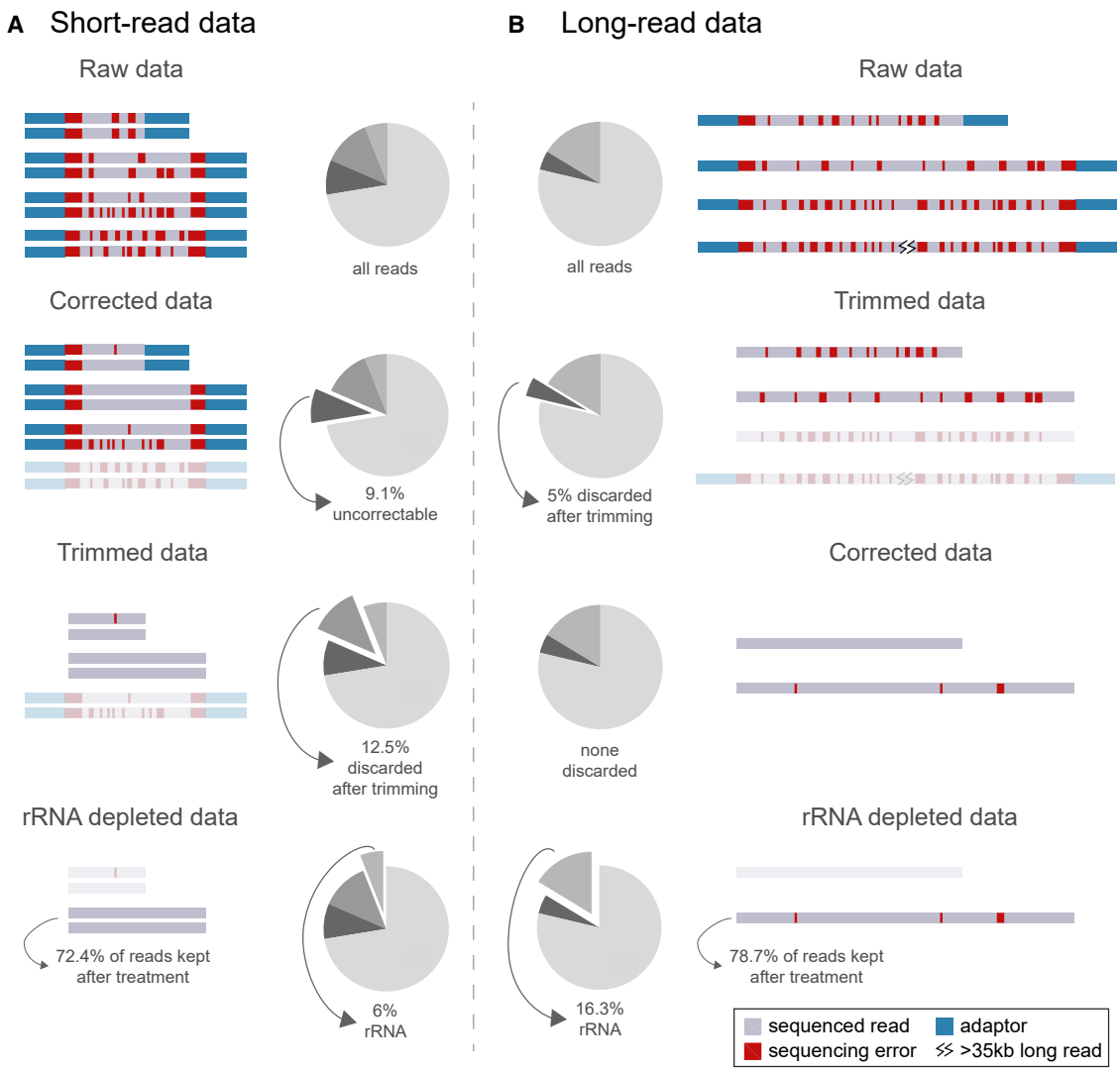


Figure 2. Short- and long-read data treatment in HyDRA

(A) Paired-end short-read data, followed by pie charts highlighting the proportion of reads discarded in each step relative to the total number of raw reads.
(B) Single-end long-read data, preceded by pie charts highlighting the proportion of reads discarded in each step relative to the total number of raw reads.

in the Phred scale showed an increase from 12.90 to 14.50 after trimming. Approximately 5% of the reads were discarded during adaptor removal and quality trimming (Figure 2). From the remaining reads, 99% were above Q10 and 81% were above Q12 (Figure S3 and Table S3). Next, we integrated the pre-processed short- and long-read sequencing data to perform a hybrid error correction of the long reads. We observed a balanced base composition in the Burrows-Wheeler transform created from all pre-processed short reads. However, we consistently noticed that RopeBWT2,[20] one of the tools used in the hybrid error correction steps (more details on development-and-implementation at the HyDRA GitHub repository - https://github.com/isabela42/HyDRA), outputs a base count report in which thymine base counts and N (undefined) base counts are swapped. This has been addressed in HyDRA, which now outputs the correct base counts to the user (Table S3). Despite base quality information being lost after long-read correction, no sequences were discarded at this point, implying that all reads that survived trimming were corrected and kept for further processing (Table S3).

Most RNAseq library preparation protocols include a ribodepletion or poly(A) selection step, but rRNAs still represent a large portion of the sequenced data.[30] These are considered cognate contaminants, meaning they are reads originating from undesired RNA types and must be removed prior to the *de novo* assembly. Using a database of known rRNA sequences (Table S1), we have included a step in HyDRA where pre-processed reads are computationally filtered to remove ribosomal contamination. During quality assessment with FastQC,[41] two long-read sequences identified as overrepresented were confirmed through BLAT searches to be human rRNAs.[42] On average, short-read data contained 6.00% of rRNA-derived

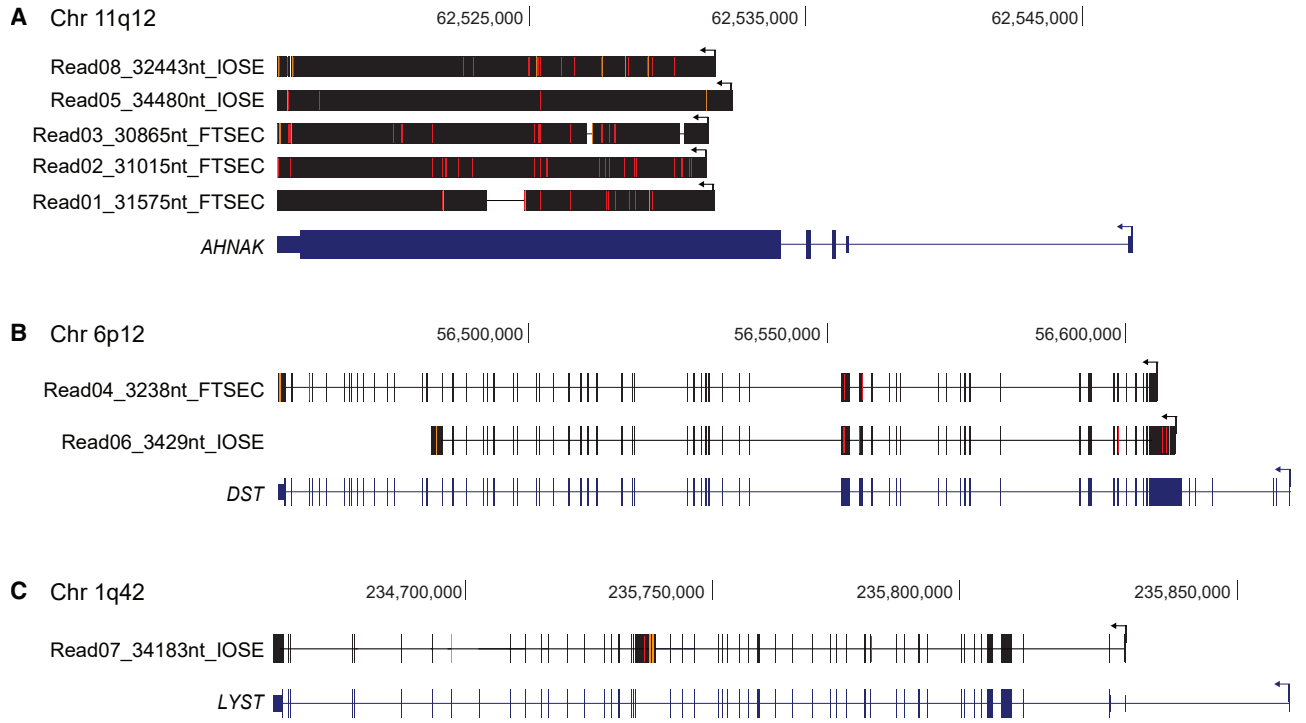


Figure 3. Treated long-read sequences reaching up to 35 kb aligned to the human genome using BLAT

These eight pre-processed reads (four from FTSEC and four from OSEC samples) were aligned against the human genome (GRCh38) in a UCSC BLAT search to confirm they were valid sequences. The sequencing reads aligned to (A) *AHNAK*, (B) *DST,* or (C) *LYST*.

reads and long-read data contained 16.30% (Figure 2 and Table S3). These numbers align with expected rRNA sequencing levels, even after ribodepletion during library preparation.[43]

Long-read sequences of up to 35 kb were kept and used for assembly

The longest known human transcript is *TTN* (titin), with 109,224 nt,[44] thus we anticipated that certain long-read sequences in our dataset would be significantly longer than the reported average of 1–3 kb,[18,19] potentially including ultra-long reads over 100 kb in length. Indeed, although the average length of the raw long reads was 1024 nt, the longest was 103,744 nt (Figures S4A–S4C and Table S3). FastQC analysis indicated that, while reads with up to ~30,000 nt had the expected base composition (i.e., balanced proportions of A, T, C, and G), longer sequences presented a distinctively biased pattern of nucleotide composition, rich in thymines and guanines (Figure S4B). In light of this, we implemented an additional QC step to analyze sequence lengths throughout the pipeline and allow users to remove sequences with unexpected nucleotide composition (fasta_splitter.py; Figure S4D). After all read treatment steps were performed, the eight remaining longest sequenced reads (Table S3; 30–35 kb) were aligned to the GRCh38 reference genome using BLAT[42] (Figure 3). All were confirmed as valid human sequences, aligning to *AHNAK* (desmoyokin), *DST* (dystonin), or *LYST* (lysosomal trafficking regulator). All treated long-read sequences were used for *de novo* assembly, including those reaching 35 kb. Importantly, this maximum length is dependent on the input data, the tissue(s), developmental stage, and species being analyzed. Additionally, through fasta_splitter.py, HyDRA gives users the option to strictly keep sequences that have up to *n* nt in length.

HyDRA based hybrid transcriptome assembly performs better than alternative approaches

To assess HyDRA's overall performance, we generated several assemblies using the same set of treated reads but different assemblers: hybrid approaches (rnaSPAdes[12] and StringTie2[45]); short-read-only approaches (Trinity v.2.8.4[34] and StringTie2[45]); and a long-read-only approach (StringTie2[45]). The resulting assemblies were subjected to the same processing steps as their HyDRA counterpart. To evaluate the quality of these assemblies, we used a subset of the metrics reported in a recent benchmark study[46] and respective normalized score (0–1). These metrics included transcript length, N50, reference coverage, open reading frame (ORF) percentage, undefined base count, and conserved orthologs representation. Benchmarking showed the HyDRA-rnaSPAdes assembly to be the best-performing method, even considering the nuances of the different tested datasets. Based on the normalized score, the HyDRA-rnaSPAdes assembly performed over 80% better than the hybrid approach obtained with StringTie2,[45] with a performance two times better compared to StringTie2's short-read-only approach and almost 60% better compared to its long-read-only assembly approach. Compared to other methods, the HyDRA-rnaSPAdes assembly performed on median ~28% better than the best-performing

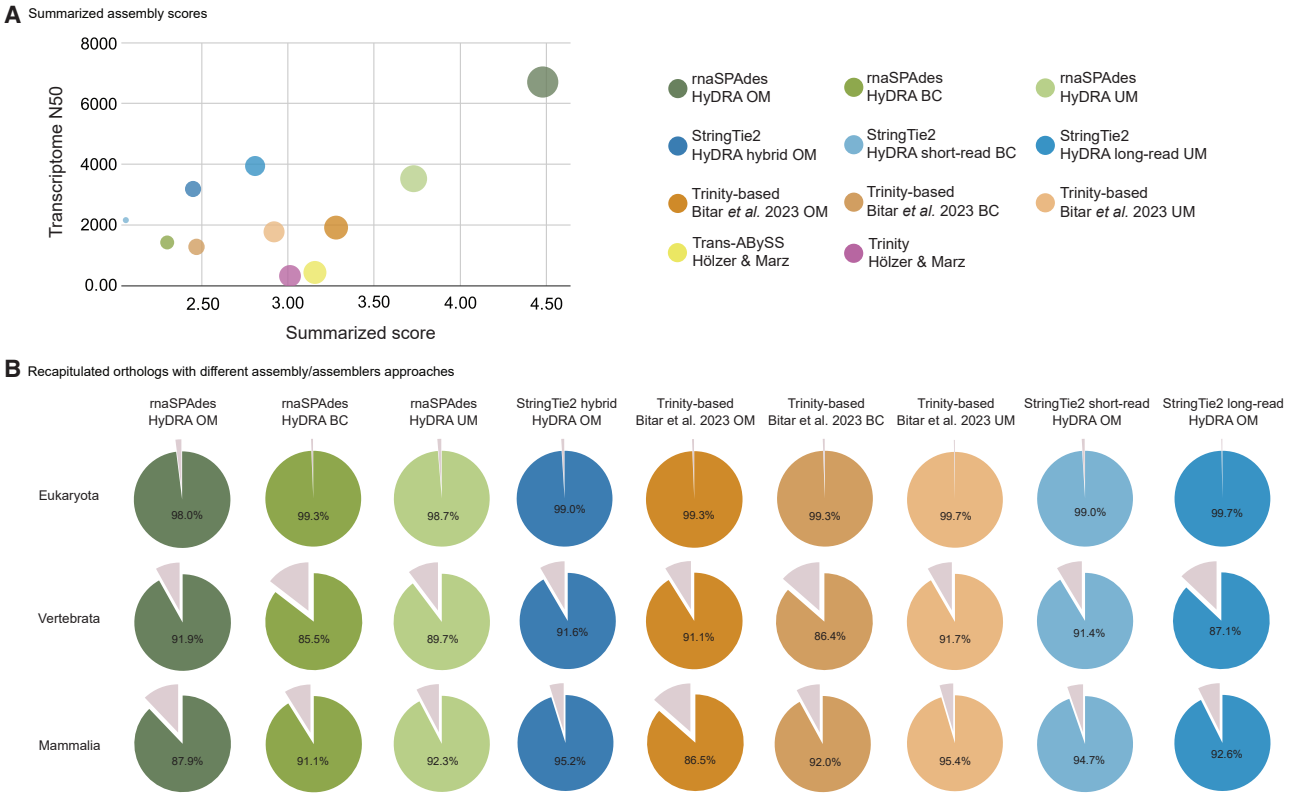


Figure 4. Overall assessment of the HyDRA-generated assembly

(A) Normalized assembly scores. Bubble sizes vary according to the N50 value. The graph shows scores for the assemblies produced by two hybrid assemblers, (i) rnaSPAdes and (ii) StringTie2, with HyDRA-treated reads as input; (ii) two short-read-only assemblers, Trinity and StringTie2, with Bitar et al.[4] pipeline-treated reads as input; (iii) the best (*Trans*-ABySS) and iv) second-best *de novo* assembly tools alone (Trinity). Both (i) and (ii) are based on data described here (OM – ovarian metatranscriptome from human ovarian and fallopian tube samples; BC – breast cancer; UM – uveal melanoma); (iii) and (iv) were based on data by Hölzer and Marz's.[46]

(B) BUSCO analysis shows assembly completeness based on the percentage of recapitulated orthologs. Colored regions show recapitulated orthologs, opposed to the gray segments in the pie charts, which represent the known orthologs in the database that were not recapitulated in the analyzed assembly.

short-read-only approach[4] reaching up to ~35% in one of the benchmarked datasets and outperformed the top-ranked *de novo* assembly tool alone by over 40% (Figure 4A and Table S4).[46] Our hybrid approach with rnaSPAdes generated 857,736 transcript sequences, reaching up to 67,466 nt, with an average transcript length of 2,409 nt and GC content of ~44% (Table S4), which agrees with the reported human GC content of coding (~52%) and noncoding (~44%) isoforms.[47]

In terms of assembly contiguity, the N50 is an important metric defined as the length of the sequence at which 50% of the total assembly size is contained in sequences of at least that length. For the benchmarked datasets, HyDRA produced assemblies with N50 varying from 1283 to 6708 nt, which reflects how a hybrid approach can represent full-length human transcripts (Table S4). For comparison, Hölzer and Marz's best performing assembler produced a transcriptome with an N50 of 441 nt (15.21 times smaller than HyDRA's assembly)[46] and Bitar et al. an N50 of 1383 (4.85 times smaller than HyDRA's assembly).[4] The highest N50 observed by Hölzer and Marz was 2381 nt (2.82 times smaller than HyDRA's assembly), but this study showed that the assembler performed poorly compared to the other tools and metrics.[46] Here, while investigating the contribution of adding long reads to transcriptome assembly, we used multiple short-read-only (Bitar et al.[4] pipeline-based and StringTie2[45]), long-read-only (StringTie2[45]) and hybrid (rnaSPAdes[12] and StringTie2[45]) assembly methods applied to HyDRA-treated data. The calculated N50 of the short-read-only assembly was up to three times smaller than HyDRA's, and the assembly had double the number of transcripts. This suggests HyDRA can generate less fragmented assemblies that are likely to better recapitulate full-length transcripts while maintaining high overall quality. Compared to the long-read-only assembly, HyDRA showed an N50 almost twice as long and ~25% more transcripts, indicating an advantage in integrating short and long reads for custom transcriptome assembly. Similar to the N50, the N90 metric corresponds to the transcript length at which 90% of the total assembly size is contained in sequences of at least that length. Using our hybrid approach, we achieved an N90 up to ~1000 nt, meaning that 90% of the transcripts in the HyDRA assembly are sequences matching the average human transcript length.[17] This demonstrates the overall contiguity of the produced custom transcriptome (Table S4).

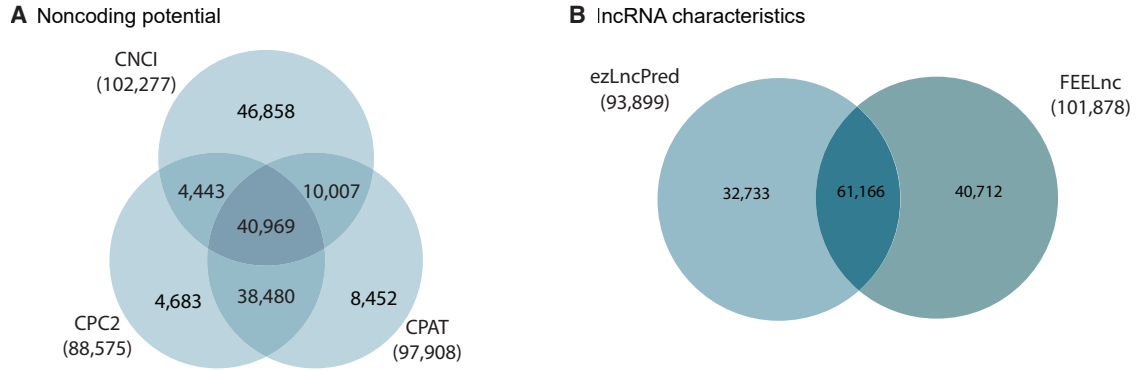


Figure 5. LncRNA discovery

(A) Intersection of lncRNA candidates predicted by three different ezLncPred machine learning models (CNCI, CPAT, and CPC2).[36]

(B) Intersection between FEELnc lncRNA predictions[37] and the list of candidates predicted by at least two of the ezLncPred machine learning models.

The HyDRA-generated assembly accurately recapitulated several aspects of the human transcriptome. For example, BUSCO analysis revealed that >98% of the eukaryotic (297/303), 91.9% of the vertebrata (2376/2586), and 87.8% of the mammalian (3606/4104) conserved orthologs were captured in our hybrid assembly of the ovarian metatranscriptome, indicating overall completeness (Figure 4B). These values were consistent with those obtained across all the benchmarked hybrid, short-read-only, and long-read-only assemblies/assemblers (Figure 4B and Table S4). According to TransRate, the custom hybrid assembly covered 24% of the reference human transcriptome (GENCODE), which is comparable to the 23–26% achieved with the best-performing assembler tools found by Hölzer and Marz[46] (Table S4). For perspective, HyDRA's transcripts cover approximately 12% of the reference genome, while the exons and UTRs in the reference annotation (GENCODE v36) cover approximately 5% (genome coverages were calculated with BedTools genomecov[38]).

Splicing assessment showed 30% of transcripts to be multiexonic

Most assemblies to date disregard monoexonic transcripts, but recent evidence has shown this class to contain conserved lncRNAs of functional relevance.[48–51] Similar to Bitar et al.,[4] we have kept the monoexonic transcripts in our assembly, as long as they had high read support. Transcripts aligned to the reference human genome (GRCh38) were classified as monoexonic or multiexonic according to the presence of "N" tags in the alignment file. A minimum length of 50 nt was defined to differentiate introns from insertions and deletions (indels), which aligns with current knowledge about human introns. Before filtering out low read support transcripts, HyDRA's hybrid assembly showed a ratio of multiexonic:monoexonic transcripts of 3:7 (~232,000 transcripts were classified as multiexonic and ~619,000 as monoexonic; Figure 4C and Table S4). For comparison, the short-read-only approach showed a ratio of 3:17 (~143,000 were multiexonic and ~740,000 were monoexonic). This difference is likely guided by the power of long reads to resolve transcript architecture and improve overall isoform assembly.

Removing transcripts with low read support helps remove technical artifacts and transcriptional leakage products, as well as problematic transcripts arising from misassembly. As HyDRA integrates short and long reads, read support for each transcript was calculated based on a combination of both subsets, which is computationally and biologically challenging. In HyDRA, redundant transcripts are collapsed prior to read support calculations. This redundancy reduction step removed ~8,500 transcripts from the original assembly (Table S4). From the remaining ~224,500 multiexonic and ~618,000 monoexonic transcripts, ~189,000 (84.34%) and ~13,000 (2.12%) respectively, passed the more permissive RPKM cut-off for read support (0.3 and 3 RPKM). As expected, the number of supported transcripts was much lower when using the stricter RPKM cut-off (1 and 5 RPKM), with a 90% decrease in multiexonic (~20,000) and 50% decrease in monoexonic transcripts (~7,300) passing the filtering step. Since we previously validated transcripts with low read support by qPCR, confirming that the less stringent cut-off still identifies bona fide transcripts,[4] we opted to use these transcripts for further analysis. The ratio of multiexonic to monoexonic transcripts in the assembly is 1:14, maintaining the expected lncRNA ratio observed in the telomere-to-telomere (T2T) human genome (T2T-CHM13).[52] In total, the final filtered custom assembly consists of 202,459 transcripts, providing a comprehensive representation of the normal ovarian metatranscriptome.

Identification of unannotated lncRNAs in HyDRA's custom transcriptome

To assess the coding potential of the 202,459 transcripts, we ran three machine learning models from the ezLNCpred package, CPAT, CNCI, and CPC2.[36] On average, at least two of the models agree on 61% of the noncoding predictions (93,899), suggesting that these methods are more effective at confirming the absence of ORFs rather than detecting their presence. Additionally, 26.6% (40,969) had no coding potential detected by any of the three models (Figure 5A and Table S5). A total of 47,281 transcripts were predicted by at least two models as noncoding and not by any model as protein-coding. We decided to include all 93,899 transcripts predicted as

noncoding by at least two of the tools in our further analysis (Figure 5B and Table S5).

FEELnc was trained using the GENCODE GRCh38 transcriptome annotation of protein-coding and lncRNA transcripts.[53] This enabled us to identify which of the 202,459 transcripts assembled by HyDRA had lncRNA characteristics. To ensure robust predictive capability, we exclusively used experimentally validated GENCODE lncRNAs. FEELnc predicted 101,878 transcripts to be lncRNAs (Table S5). This represents about half the total number of transcripts in our custom transcriptome, indicating that lncRNAs constitute a significant proportion of expressed transcripts in normal ovarian and fallopian tube tissues. Importantly, more than 60% of these (61,166) lncRNAs were supported by at least two of the three ezLncPred machine-learning models (Figure 5B) and selected as candidate lncRNAs.

The majority of these candidate lncRNAs were multiexonic, which may reflect the biased training dataset. From the 61,166 candidate lncRNAs, 629 showed a bidirectional overlap of at least 85% with a protein-coding transcript and a minimum identity of 75%. We believe these to be either false positives that our methods failed to detect, or noncoding isoforms of protein-coding genes. Although monoexonic and sense genic transcripts (i.e., those overlapping protein-coding genes in the same strand) are functionally relevant, it is difficult to differentiate these from technical artifacts or transcriptional leakage. Furthermore, sense genic lncRNAs cannot easily be uncoupled from the corresponding protein-coding gene, and we have included a restrictive alignment step in HyDRA to facilitate their removal. From the remaining 60,537 lncRNAs, 228 were already annotated in GENCODE (GRCh38: 19 multiexonic; 1 monoexonic) or the in-house database of over 112,000 lncRNA transcripts (124 multiexonic; 84 monoexonic) as exact matches. We also intersected the coordinates of the remaining 60,309 lncRNAs with those of protein-coding genes using BedTools,[38] revealing 7,615 exon-overlapping transcripts, resulting in the identification of 53,551 high-confidence lncRNA transcripts. Importantly, HyDRA was designed to split monoexonic and multiexonic sequences after assembly, which allows users to easily focus on either or both sets of transcripts.

To demonstrate the functional relevance of these 53,551 high-confidence lncRNA transcripts from the normal ovaries and fallopian tubes, we assessed their expression profiles in a different subset of sequenced RNA samples (unrelated but biologically similar to those used for assembly). We used publicly available RNAseq data from normal and cancerous ovarian and fallopian tube tissues found in the RNA Atlas project.[54] Most of HyDRA's lncRNAs (46,317 or 86.49% with at least 0.1 TPM and 27,257 or 50.90% with at least 0.5 TPM) were detected in at least one of these samples, demonstrating its efficiency in assembling both annotated and unannotated lncRNA transcripts, through the integration of long- and short-read RNAseq data. We expect the remaining lncRNAs to have not been detected due to their specificity regarding expression in limited cell types, conditions, or at low levels, which are likely not entirely represented in the RNA Atlas project.

Limitations of the study

This study has several strengths, including the application of high-throughput long-read technology, benchmarking on different RNAseq datasets, and comprehensive QC throughout pipeline design and implementation. However, there are limitations worth discussing. First, while benchmarking was performed for choosing a hybrid assembler and a long-read error correction tool, we relied on published benchmarking studies to choose other tools integrated in HyDRA. It is worth noting that some of the existing tools had issues or specificities which prevented us to consider them our benchmarking (see development-and-implementation at the HyDRA GitHub repository - https://github.com/isabela42/HyDRA). Reflecting bias from the available data, HyDRA was implemented and benchmarked using poly(A)-captured RNAseq reads. Future applications will likely test HyDRA's efficiency in assembling comprehensive (e.g., ribo-depleted) RNAseq. Although this is not a limitation specific to HyDRA, we expect it to have an impact on the number of identified lncRNA transcripts, as current estimates suggest about 40% of all lncRNAs are non-polyadenylated. Additionally, the majority of datasets used in this work were derived from cancer samples. Although we do not expect any changes in performance in non-cancerous datasets, as we have seen with the normal ovarian and fallopian tube data, there is a limitation in our study for not including other disease types. HyDRA's performance in different datasets was compared at the assembled transcriptome stage, before lncRNA identification. To the best of our knowledge, this was the most adequate comparison, as the number and the isoforms of lncRNAs are expected to vary widely across different tissues, cell types, diseases, and stages. This would not support a direct comparison between the lncRNA sets, and the absence of a gold standard means lncRNA discovery was not a useful measure of HyDRA's performance. We therefore only performed lncRNA discovery as a proof of principle for the ovarian metatranscriptome.

Despite these limitations, HyDRA is a comprehensive pipeline that integrates short- and long-read sequencing data for a true-hybrid *de novo* transcriptome assembly and lncRNA discovery. We used deep, short- and long-read RNAseq from ovarian and fallopian tube epithelial cells, breast and uveal melanoma samples to develop, validate, and assess the efficacy of the pipeline in generating a high-quality custom transcriptome. We have shown that HyDRA's assembly performed up to 40% better than the benchmarked alternative approaches. Based on this custom assembly, we identified 61,166 candidate lncRNAs, among which 60,309 have not been previously annotated, and 53,551 showed no overlap with protein-coding transcripts. In summary, HyDRA is a high-performing hybrid-assembly tool capable of facilitating accurate transcriptome reconstruction and advancing lncRNA annotation.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Dr. Mainá Bitar (maina.bitar@qimrberghofer.edu.au).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All original codes have been deposited at GitHub at https://github.com/isabela42/HyDRA and https://github.com/isabela42/GRADE2 and are publicly available in source and executable form without charge for nonprofit, academic, and personal uses under MIT licenses.[26,55]
- All data generated and used in this study are available in the key resources table and details in Table S2.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This work was supported by grants from the National Health and Medical Research Council of Australia (NHMRC; 2019101). S.L. Edwards was supported by an NHMRC Investigator Grant (2033097). J.D. French was supported by an NHMRC Investigator Grant (2016826). M. Bitar was supported by The Donald and Joan Wilson Foundation. I. Almeida was supported by a QIMR Berghofer International PhD scholarship and a University of Queensland Research Training scholarship. We would like to acknowledge the QIMR Berghofer HPC team for their support in maintaining the institutional computational infrastructure. Finally, we would like to acknowledge Bryan Khelven for his support in the creation of Docker containers.

AUTHOR CONTRIBUTIONS

Conceptualization: I.A., J.D.F., S.L.E., and M.B.; data curation: I.A.; formal analysis: I.A.; funding acquisition: J.D.F., S.L.E., and M.B.; investigation: I.A. (pipeline development, implementation, benchmarking, analyses, computational code revision and GitHub repository creation and maintenance), M.B. (benchmarking of long read correction methods) and X.L. (preparation of RNA libraries for short- and long-read RNA sequencing); methodology: I.A. and M.B.; project administration: M.B. and J.F.; resources: I.A.; software: I.A.; supervision: J.D.F., S.L.E. and M.B.; validation: M.B.; visualization: I.A.; writing: . I.A., J.D.F., S.L.E., and M.B.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - FTSEC and OSEC cell lines culture/growth conditions
- METHOD DETAILS
  - RNAseq sample preparation and sequencing
  - Additional datasets for benchmarking
  - Databases and reference genome versions
  - Parameters used for the ovarian and fallopian tube custom assembly
  - Expression profiles of lncRNA transcripts
  - Bioinformatics tools used for HyDRA's development and benchmarking
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2026.116131.

Received: November 24, 2025
Revised: January 23, 2026
Accepted: May 12, 2026

REFERENCES

1. Hu, T., Chitnis, N., Monos, D., and Dinh, A. (2021). Next-generation sequencing technologies: An overview. Hum. Immunol. *82*, 801–811. https://doi.org/10.1016/j.humimm.2021.02.012.
2. Garber, M., Grabherr, M.G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. Nat. Methods *8*, 469–477. https://doi.org/10.1038/nmeth.1613.
3. Wu, P.-Y., Phan, J.H., and Wang, M.D. (2013). Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinf. *14*, S8. https://doi.org/10.1186/1471-2105-14-s11-s8.
4. Bitar, M., Rivera, I.S., Almeida, I., Shi, W., Ferguson, K., Beesley, J., Lakhani, S.R., Edwards, S.L., and French, J.D. (2023). Redefining normal breast cell populations using long noncoding RNAs. Nucleic Acids Res. *51*, 6389–6410. https://doi.org/10.1093/nar/gkad339.
5. Foord, C., Hsu, J., Jarroux, J., Hu, W., Belchikov, N., Pollard, S., He, Y., Joglekar, A., and Tilgner, H.U. (2023). The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing. Nat. Methods *20*, 20–24. https://doi.org/10.1038/s41592-022-01715-9.
6. Carbonell Sala, S., Uszczyńska-Ratajczak, B., Lagarde, J., Johnson, R., and Guigó, R. (2021). Annotation of Full-Length Long Noncoding RNAs with Capture Long-Read Sequencing (CLS). In Functional Analysis of Long Non-Coding RNAs, H. Cao, ed. (Springer US), pp. 133–159. https://doi.org/10.1007/978-1-0716-1158-6_9.
7. Dong, X., Du, M.R.M., Gouil, Q., Tian, L., Jabbari, J.S., Bowden, R., Baldoni, P.L., Chen, Y., Smyth, G.K., Amarasinghe, S.L., et al. (2023). Benchmarking long-read RNA-sequencing analysis tools using in silico mixtures. Nat. Methods *20*, 1810–1821. https://doi.org/10.1038/s41592-023-02026-3.
8. Workman, R.E., Tang, A.D., Tang, P.S., Jain, M., Tyson, J.R., Razaghi, R., Zuzarte, P.C., Gilpatrick, T., Payne, A., Quick, J., et al. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. Nat. Methods *16*, 1297–1305. https://doi.org/10.1038/s41592-019-0617-2.
9. Amarasinghe, S.L., Su, S., Dong, X., Zappia, L., Ritchie, M.E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. Genome Biol. *21*, 30. https://doi.org/10.1186/s13059-020-1935-5.
10. Zhang, H., Jain, C., and Aluru, S. (2020). A comprehensive evaluation of long read error correction methods. BMC Genom. *21*, 889. https://doi.org/10.1186/s12864-020-07227-0.
11. Shumate, A., Wong, B., Pertea, G., and Pertea, M. (2022). Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput. Biol. *18*, e1009730. https://doi.org/10.1371/journal.pcbi.1009730.
12. Prjibelski, A.D., Puglia, G.D., Antipov, D., Bushmanova, E., Giordano, D., Mikheenko, A., Vitale, D., and Lapidus, A. (2020). Extending rnaSPAdes functionality for hybrid transcriptome assembly. BMC Bioinf. *21*, 302. https://doi.org/10.1186/s12859-020-03614-2.
13. Fu, S., Ma, Y., Yao, H., Xu, Z., Chen, S., Song, J., and Au, K.F. (2018). IDP-denovo: de novo transcriptome assembly and isoform annotation by hybrid sequencing. Bioinformatics *34*, 2168–2176. https://doi.org/10.1093/bioinformatics/bty098.
14. Puglia, G.D., Prjibelski, A.D., Vitale, D., Bushmanova, E., Schmid, K.J., and Raccuia, S.A. (2020). Hybrid transcriptome sequencing approach improved assembly and gene annotation in Cynara cardunculus (L.). BMC Genom. *21*, 317. https://doi.org/10.1186/s12864-020-6670-5.
15. Kainth, A.S., Haddad, G.A., Hall, J.M., and Ruthenburg, A.J. (2023). Merging short and stranded long reads improves transcript assembly. PLoS Comput. Biol. *19*, e1011576. https://doi.org/10.1371/journal.pcbi.1011576.
16. Vilperte, V., Lucaciu, C.R., Halbwirth, H., Boehm, R., Rattei, T., and Debener, T. (2019). Hybrid de novo transcriptome assembly of poinsettia (Euphorbia pulcherrima Willd. Ex Klotsch) bracts. BMC Genom. *20*, 900. https://doi.org/10.1186/s12864-019-6247-3.

17. Sharon, D., Tilgner, H., Grubert, F., and Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. Nat. Biotechnol. *31*, 1009–1014. https://doi.org/10.1038/nbt.2705.

18. Xia, Y., Jin, Z., Zhang, C., Ouyang, L., Dong, Y., Li, J., Guo, L., Jing, B., Shi, Y., Miao, S., and Xi, R. (2023). TAGET: a toolkit for analyzing full-length transcripts from long-read sequencing. Nat. Commun. *14*, 5935. https://doi.org/10.1038/s41467-023-41649-0.

19. Schwenk, V., Leal Silva, R.M., Scharf, F., Knaust, K., Wendlandt, M., Häusser, T., Pickl, J.M.A., Steinke-Lange, V., Laner, A., Morak, M., et al. (2023). Transcript capture and ultradeep long-read RNA sequencing (CAPLRseq) to diagnose HNPCC/Lynch syndrome. J. Med. Genet. *60*, 747–759. https://doi.org/10.1136/jmg-2022-108931.

20. Li, H. (2014). Fast construction of FM-index for long sequence reads. Bioinformatics *30*, 3274–3275. https://doi.org/10.1093/bioinformatics/btu541.

21. Wan, Y., Liu, X., Zheng, D., Wang, Y., Chen, H., Zhao, X., Liang, G., Yu, D., and Gan, L. (2019). Systematic identification of intergenic long-noncoding RNAs in mouse retinas using full-length isoform sequencing. BMC Genom. *20*, 559. https://doi.org/10.1186/s12864-019-5903-y.

22. Kelley, D., and Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. *13*, R107. https://doi.org/10.1186/gb-2012-13-11-r107.

23. Johnson, R., and Guigó, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA *20*, 959–976. https://doi.org/10.1261/rna.044560.114.

24. Zhang, X., and Meyerson, M. (2020). Illuminating the noncoding genome in cancer. Nat. Cancer *1*, 864–872. https://doi.org/10.1038/s43018-020-00114-3.

25. Soneson, C., Yao, Y., Bratus-Neuenschwander, A., Patrignani, A., Robinson, M.D., and Hussain, S. (2019). A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat. Commun. *10*, 3359. https://doi.org/10.1038/s41467-019-11272-z.

26. Almeida, I. (2026). A Hybrid de novo RNA assembly pipeline: Release 1 (Zenodo). https://doi.org/10.5281/ZENODO.19043714.

27. Laehnemann, D., Borkhardt, A., and McHardy, A.C. (2016). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. Brief. Bioinform. *17*, 154–179. https://doi.org/10.1093/bib/bbv029.

28. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J.M., and Kelly, S. (2016). TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res. *26*, 1134–1144. https://doi.org/10.1101/gr.196469.115.

29. Liao, X., Li, M., Zou, Y., Wu, F., Yi-Pan, and Wang, J. (2019). Current challenges and solutions of de novo assembly. Quant. Biol. *7*, 90–109. https://doi.org/10.1007/s40484-019-0166-9.

30. Raghavan, V., Kraft, L., Mesny, F., and Rigerte, L. (2022). A simple guide to de novo transcriptome assembly and annotation. Brief. Bioinform. *23*, bbab563. https://doi.org/10.1093/bib/bbab563.

31. Dong, X., Tian, L., Gouil, Q., Kariyawasam, H., Su, S., De Paoli-Iseppi, R., Prawer, Y.D.J., Clark, M.B., Breslin, K., Iminitoff, M., et al. (2021). The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. NAR Genom. Bioinform. *3*, lqab028. https://doi.org/10.1093/nargab/lqab028.

32. Mak, Q.X.C., Wick, R.R., Holt, J.M., and Wang, J.R. (2023). Polishing De Novo Nanopore Assemblies of Bacteria and Eukaryotes With FMLRC2. Mol. Biol. Evol. *40*, msad048. https://doi.org/10.1093/molbev/msad048.

33. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics *31*, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351.

34. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. *29*, 644–652. https://doi.org/10.1038/nbt.1883.

35. Mattick, J.S., Amaral, P.P., Carninci, P., Carpenter, S., Chang, H.Y., Chen, L.-L., Chen, R., Dean, C., Dinger, M.E., Fitzgerald, K.A., et al. (2023). Long non-coding RNAs: definitions, functions, challenges and recommendations. Nat. Rev. Mol. Cell Biol. *24*, 430–447. https://doi.org/10.1038/s41580-022-00566-8.

36. Xu, X., Liu, S., Yang, Z., Zhao, X., Deng, Y., Zhang, G., Pang, J., Zhao, C., and Zhang, W. (2021). A systematic review of computational methods for predicting long noncoding RNAs. Brief. Funct. Genomics *20*, 162–173. https://doi.org/10.1093/bfgp/elab016.

37. Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., et al. (2017). FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. Nucleic Acids Res. *45*, e57. https://doi.org/10.1093/nar/gkw1306.

38. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

39. Stoler, N., and Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. NAR Genom. Bioinform *3*, lqab019. https://doi.org/10.1093/nargab/lqab019.

40. Chandak, S., Neu, J., Tatwawadi, K., Mardia, J., Lau, B., Kubit, M., Hulett, R., Griffin, P., Wootters, M., Weissman, T., et al. (2020). Overcoming High Nanopore Basecaller Error Rates for DNA Storage via Basecaller-Decoder Integration and Convolutional Codes. In ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), P. Closas and M. Bugallo, eds. (IEEE), pp. 8822–8826. https://doi.org/10.1109/icassp40776.2020.9053441.

41. Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., and Wingett, S. (2018). FastQC Tool (Babraham Institute). http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

42. Kent, W.J. (2002). BLAT - The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664. https://doi.org/10.1101/gr.229202.

43. Tariq, M.A., Kim, H.J., Jejelowo, O., and Pourmand, N. (2011). Whole-transcriptome RNAseq analysis from minute amount of total RNA. Nucleic Acids Res. *39*, e120. https://doi.org/10.1093/nar/gkr547.

44. Lopes, I., Altab, G., Raina, P., and de Magalhães, J.P. (2021). Gene Size Matters: An Analysis of Gene Length in the Human Genome. Front. Genet. *12*, 559998. https://doi.org/10.3389/fgene.2021.559998.

45. Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L., and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. *20*, 278. https://doi.org/10.1186/s13059-019-1910-1.

46. Hölzer, M., and Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience *8*, giz039. https://doi.org/10.1093/gigascience/giz039.

47. Niazi, F., and Valadkhan, S. (2012). Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3′ UTRs. RNA *18*, 825–843. https://doi.org/10.1261/rna.029520.111.

48. Wang, L., Bitar, M., Lu, X., Jacquelin, S., Nair, S., Sivakumaran, H., Hillman, K.M., Kaufmann, S., Ziegman, R., Casciello, F., et al. (2024). CRISPR-Cas13d screens identify KILR, a breast cancer risk-associated lncRNA that regulates DNA replication and repair. Mol. Cancer *23*, 101. https://doi.org/10.1186/s12943-024-02021-y.

49. Ilott, N.E., Heward, J.A., Roux, B., Tsitsiou, E., Fenwick, P.S., Lenzi, L., Goodhead, I., Hertz-Fowler, C., Heger, A., Hall, N., et al. (2014). Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes. Nat. Commun. *5*, 3979. https://doi.org/10.1038/ncomms4979.

50. Roux, B.T., Heward, J.A., Donnelly, L.E., Jones, S.W., and Lindsay, M.A. (2017). Catalog of Differentially Expressed Long Non-Coding RNA following Activation of Human and Mouse Innate Immune Response. Front. Immunol. *8*, 1038. https://doi.org/10.3389/fimmu.2017.01038.

51. Khachane, A.N., and Harrison, P.M. (2010). Mining Mammalian Transcript Data for Functional Long Non-Coding RNAs. PLoS One *5*, e10316. https://doi.org/10.1371/journal.pone.0010316.

52. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science (1979) *376*, 44–53. https://doi.org/10.1126/science.abj6987.

53. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J.M., Sisu, C., Wright, J., Armstrong, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. *47*, D766–D773. https://doi.org/10.1093/nar/gky955.

54. Lorenzi, L., Chiu, H.-S., Avila Cobos, F., Gross, S., Volders, P.-J., Cannoodt, R., Nuytens, J., Vanderheyden, K., Anckaert, J., Lefever, S., et al. (2021). The RNA Atlas expands the catalog of human non-coding RNAs. Nat. Biotechnol. *39*, 1453–1465. https://doi.org/10.1038/s41587-021-00936-1.

55. Almeida, I. (2026). General RNAseq Analysis for Differential Expression version 2 (Zenodo). https://doi.org/10.5281/ZENODO.19043655.

56. Zhang, Z., Li, C., Li, Q., Su, X., Li, J., Zhu, L., Lin, X.J., and Shen, J. (2023). Structure prediction of novel isoforms from uveal melanoma by AlphaFold. Sci. Data *10*, 513. https://doi.org/10.1038/s41597-023-02429-z.

57. De Coster, W., and Rademakers, R. (2023). NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics *39*, btad311. https://doi.org/10.1093/bioinformatics/btad311.

58. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

59. Li, H. (2018). GitHub Repository: Seqtk, A Toolkit for Processing Sequences in FASTA/Q Formats (GitHub). https://github.com/lh3/seqtk.

60. Wick, R. (2018). Porechop: Adapter Trimmer for Oxford Nanopore Reads (GitHub). https://github.com/rrwick/porechop.

61. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. *17*, 10. https://doi.org/10.14806/ej.17.1.200.

62. Song, L., and Florea, L. (2015). Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. GigaScience *4*, 48. https://doi.org/10.1186/s13742-015-0089-y.

63. Bushnell, B. (2014). BBMap: A Fast, Accurate, Splice-Aware Aligner (9th Annual Genomics of Energy & Environment Meeting).

64. Freedman, A.H., and Guan, L. (2016). Transcriptome Assembly Tools (GitHub). https://github.com/harvardinformatics/TranscriptomeAssemblyTools.

65. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

66. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.

67. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

68. Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185. https://doi.org/10.1093/bioinformatics/bts356.

69. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

70. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinf. *10*, 421. https://doi.org/10.1186/1471-2105-10-421.

71. Edwards, J.A., and Edwards, R.A. (2019). Fastq-pair: efficient synchronization of paired-end fastq files. Preprint at bioRxiv. https://doi.org/10.1101/552885.

72. Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310.

73. Broad Institute. (2019). Picard Toolkit: A Set of Command Line Tools (in Java) for Manipulating High-Throughput Sequencing (HTS) Data and Formats such as SAM/BAM/CRAM and VCF (GitHub). https://github.com/broadinstitute/picard.

74. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

75. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

76. Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920. https://doi.org/10.1093/bioinformatics/bts277.

77. NullModel, H. (2014). kentUtils, UCSC Command Line Bioinformatic Utilities (GitHub). https://github.com/ENCODE-DCC/kentUtils.

78. Bonfield, J.K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. GigaScience *10*, giab007. https://doi.org/10.1093/gigascience/giab007.

79. Moore, B. (2012). Genome Annotation Library, A Perl Toolkit for Working with SO Compliant Genome Annotations (GitHub). https://github.com/The-Sequence-Ontology/GAL.

80. Wang, M., and Kong, L. (2019). pblat: a multithread blat algorithm speeding up aligning sequences to genomes. BMC Bioinf. *20*, 28. https://doi.org/10.1186/s12859-019-2597-8.

81. Morisse, P., Lecroq, T., and Lefebvre, A. (2018). Hybrid correction of highly noisy long reads using a variable-order de Bruijn graph. Bioinformatics *34*, 4213–4222. https://doi.org/10.1093/bioinformatics/bty521.

82. Salmela, L., and Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. Bioinformatics *30*, 3506–3514. https://doi.org/10.1093/bioinformatics/btu538.

83. 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. https://doi.org/10.1093/bioinformatics/bts635.

84. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.

85. Heberle, H., Meirelles, G.V., da Silva, F.R., Telles, G.P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinf. *16*, 169. https://doi.org/10.1186/s12859-015-0611-3.

86. Mattick, J.S., and Rinn, J.L. (2015). Discovery and annotation of long noncoding RNAs. Nat. Struct. Mol. Biol. *22*, 5–7. https://doi.org/10.1038/nsmb.2942.

87. Seifuddin, F., Singh, K., Suresh, A., Judy, J.T., Chen, Y.-C., Chaitankar, V., Tunc, I., Ruan, X., Li, P., Chen, Y., et al. (2020). lncRNAKB, a knowledgebase of tissue-specific functional annotation and trait association of long noncoding RNA. Sci. Data *7*, 326. https://doi.org/10.1038/s41597-020-00659-z.

88. Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. *25*, 1915–1927. https://doi.org/10.1101/gad.17446611.

89. Yuan, H., Yan, M., Zhang, G., Liu, W., Deng, C., Liao, G., Xu, L., Luo, T., Yan, H., Long, Z., et al. (2019). CancerSEA: a cancer single-cell state

atlas. Nucleic Acids Res. *47*, D900–D908. https://doi.org/10.1093/nar/gky939.

90. Lanzós, A., Carlevaro-Fita, J., Mularoni, L., Reverter, F., Palumbo, E., Guigó, R., and Johnson, R. (2017). Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. Sci. Rep. *7*, 41544. https://doi.org/10.1038/srep41544.
91. Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res. *47*, D1034–D1037. https://doi.org/10.1093/nar/gky905.
92. The RNAcentral Consortium, Petrov, A.I., Kay, S.J.E., Kalvari, I., Howe, K.L., Gray, K.A., Bruford, E.A., Kersey, P.J., Cochrane, G., Finn, R.D., et al. (2017). RNAcentral: a comprehensive database of non-coding RNA sequences. Nucleic Acids Res. *45*, D128–D134. https://doi.org/10.1093/nar/gkw1008.
93. Moradi Marjaneh, M., Beesley, J., O'Mara, T.A., Mukhopadhyay, P., Koufariotis, L.T., Kazakoff, S., Hussein, N., Fachal, L., Bartonicek, N., Hillman, K.M., et al. (2020). Non-coding RNAs underlie genetic predisposition to breast cancer. Genome Biol. *21*, 7. https://doi.org/10.1186/s13059-019-1876-z.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
DMEM/F-12 medium	Thermo Fisher Scientific	Cat. # 11320033
Fetal bovine serum (FBS)	Thermo Fisher Scientific	Cat. # 26140079
L-Glutamine	Thermo Fisher Scientific	Cat. # 25030081
ReproLife™ Complete Kit	Lifeline Cell Technology	Cat. # LL0068
Fallopian tube epithelial cell Medium Kit	AcceGen	Cat. # ABM-TM5521
Ovarian epithelial cell Medium Kit	AcceGen	Cat. # ABM-TM3743
Media 199	Thermo Fisher Scientific	Cat. # 11825015
MCDB 105 Medium	Sigma-Aldrich	Cat. #M6395
Ovarian epithelial cell Medium	Sciencell	Cat. # 7311
Critical commercial assays		
RNeasy Plus Mini Kit	Qiagen	Cat. # 74134
High Sensitivity RNA ScreenTape	Agilent	Cat. # 5067-5579
TruSeq stranded mRNA library prep kit	Illumina	Cat. # 20020594
Ribo-Zero™ Plus rRNA Depletion Kit	Illumina	Cat. # 20040526
Nanopore Barcoding Kit 24 V14	Oxford Nanopore Technologies	Cat. # SQK-NBD114.24
PromethION™ P48 Flow Cell	Oxford Nanopore Technologies	Cat. #FLO-PRO114M
Deposited data		
Short-read RNAseq of FTSEC/ OSEC cell lines	This study	GEO: GSE327033
Long-read RNAseq of FTSEC/ OSEC cell lines	This study	GEO: GSE327033
Short- and long-read RNAseq of uveal melanoma	Zhang et al.[56]	GEO: GSE206464
Short- and long-read RNAseq of breast cancer	Bitar et al.[4]	N/A
Short-read RNAseq from the RNA Atlas	Lorenzi et al.[54]	GEO: GSE138734
Experimental Models: Cell Lines		
Human: IOSE397	Canadian Tissue Bank	RRID: CVCL_5541
Human: IOSE7576	Canadian Tissue Bank	RRID: CVCL_E227
Human: Ovarian epithelial cells (Lot#24593)	ScienCell	Cat. # 7310
Human: Ovarian epithelial cells (Lot#7859)	ScienCell	Cat. # 7310
Human: Ovarian epithelial cells	AcceGen	Cat. # ABC-TC3743
Human: FT282	ATCC	Cat. # CRL-3449; RRID: CVCL_A4AX
Human: FT194	ATCC	Cat. # CRL-3445; RRID: CVCL_UH58
Human: Fallopian tube epithelial cells (Lot#2839)	BioScientific	Cat. # FC-0081
Human: Fallopian tube epithelial cells (Lot#4583)	BioScientific	Cat. # FC-0081
Human: Fallopian tube epithelial cells	AcceGen	Cat. # ABC-TC5521

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
NanoPlot (v 1.41.6)	De Coster et al.[57]	https://github.com/wdecoster/nanopack
NanoComp (v 1.41.6)	De Coster et al.[57]	https://github.com/wdecoster/nanopack
FastQC (v 0.12.1)	Andrews et al.[41]	http://www.bioinformatics.babraham.ac.uk/projects/fastqc
MultiQC (v 1.14)	Ewels et al.[58]	https://github.com/MultiQC/MultiQC
Fasta_splitter.py (v v1.0.6)	This study[26]	https://github.com/isabela42/HyDRA
seqtk (v 1.3)	Li[59]	https://github.com/lh3/seqtk
Porechop (v 0.2.4)	Wick[60]	https://github.com/rrwick/Porechop/
Chopper (v 0.5.0)	De Coster et al.[57]	https://github.com/wdecoster/nanopack
Cutadapt (v 3.9.13)	Martin[61]	https://github.com/marcelm/cutadapt
Rcorrector (v 1)	Song and Florea[62]	https://github.com/mourisl/Rcorrector
Reformat (v 39.01)	Bushnell[63]	https://sourceforge.net/projects/bbmap/
FilterUncorrectabledPEfastq.py (v 2016)	Freedman[64]	https://github.com/harvardinformatics/TranscriptomeAssemblyTools
RopeBWT2 (v r187)	Li[20]	https://github.com/lh3/ropebwt2.git
FMLRC2 (v 0.1.7)	Mak[32]	https://github.com/jwanglab/fmlrc2
Trimmomatic (v 0.36)	Bolger et al.[65]	https://github.com/usadellab/trimmomatic
BBDuk (v 39.01)	Bushnell[63]	https://sourceforge.net/projects/bbmap/
Bowtie2 (v 2.2.9)	Langmead and Salzberg[66]	https://github.com/BenLangmead/bowtie2
SAMtools (v 1.9 – for HyDRA)	Danecek et al.[67]	https://github.com/samtools/samtools
RSeQC (v 2.6.4)	Wang et al.[68]	http://code.google.com/p/rseqc/
DeepTools2 (v 3.5.0)	Ramirez et al.[69]	deeptools.ie-freiburg.mpg.de
RnaSPAdes (v 3.14.1)	Prjibelski et al.[12]	http://cab.spbu.ru/software/spades/
BUSCO (v 20161119)	Simão et al.[33]	http://busco.ezlab.org
BLAST (v 2.2.31+)	Camacho et al.[70]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+
Trinity (v 2.8.4)	Grabherr et al.[34]	http://TrinityRNASeq.sourceforge.net
TransRate (v 1.0.3)	Smith-Unna et al.[28]	http://github.com/Blahah/transrate
Fastq-pair (v 20231003)	Edwards and Edwards[71]	https://doi.org/10.1101/552885
GMAP (v 45127)	Wu and Watanabe[72]	http://www.gene.com/share/gmap
Picard (v 2.19.0)	Broad Institute[73]	https://github.com/broadinstitute/picard
Minimap2 (v 2.26)	Li[74]	https://github.com/lh3/minimap2
CD-HIT (v 4.6.8)	Fu et al.[75]	http://cd-hit.org
BEDOPS (v 2.4.41)	Neph et al.[76]	http://code.google.com/p/bedops/
BedTools (v 2.29.0)	Quinlan and Hall[38]	http://code.google.com/p/bedtools
ezLncPred (v 1)	Xu et al.[36]	https://pypi.org/project/ezLncPred/
UCSC Tools (v 20160223)	NullModel[77]	https://github.com/ENCODE-DCC/kentUtils
FEELnc (v 0.2)	Wucher et al.[37]	https://github.com/tderrien/FEELnc
HTSlib (v 1.19.1)	Bonfield et al.[78]	htslib.org
gtf2gff (v 0.1)	Moore[79]	https://github.com/The-Sequence-Ontology/GAL
genestats (v 1)	This study[26]	https://github.com/isabela42/HyDRA
PBLAT (v 2.5.1)	Wang and Kong[80]	http://icebert.github.io/pblat
StringTie2 (v 3.0.0)	Kovaka et al.[45]	https://github.com/skovaka/stringtie2
HGCoLoR (v 20210120)	Morisse et al.[81]	https://github.com/morispi/HG-CoLoR

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
LoRDEC (v 0.9)	Salmela and Rivals[82]	https://github.com/cmonjeau/docker-lordec
STAR (v 2.7.1a)	Dobin et al.[83]	https://github.com/alexdobin/STAR
SAMtools (v 1.3 – for RNAseq)	Danecek et al.[67]	https://github.com/samtools/samtools
NovoSort (v 3.05.01)	Novocraft Technologies Sdn Bhd, Malaysia	http://www.novocraft.com/
RSEM (v 1.2.30)	Li and Dewey[84]	https://github.com/deweylab/RSEM
InteractiVenn	Herberle et al.[85]	https://heberleh.github.io/interactivenn/web/
Adobe Illustrator (v 28.5)	Adobe Inc. 2024	https://www.adobe.com/products/illustrator.html
HyDRA (v 1.0.0)	This study[26]	https://github.com/isabela42/HyDRA
GRADE (v 2.0.0)	This study[55]	https://github.com/isabela42/GRADE2

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

FTSEC and OSEC cell lines culture/growth conditions

Primary and immortalized fallopian tube secretory epithelial cells (FTSEC) and ovarian surface epithelial cells (OSEC) were used in this study. Thaw and culture of primary and immortalized FTSECs and OSECs was conducted in specific growth media (Table S2). Cell lines were maintained under standard conditions (37°C, 5% CO_2), tested for *Mycoplasma*, profiled for short tandem repeats, and harvested at 70–80% confluence for RNA extraction (Table S2).

METHOD DETAILS

RNAseq sample preparation and sequencing

RNA was extracted from primary and immortalized FTSEC and OSEC cell lines using the QIAGEN RNeasy Plus Mini kit (Table S2). One microgram of total RNA was rRNA depleted with the Ribo-Zero Plus kit according to the manufacturers' instructions (Illumina). Short-read RNAseq libraries were prepared using the Truseq Stranded mRNA Library Prep Kit (Illumina), and sequenced at high depth (PE150, > 75 million reads per sample; Table S2) on the Illumina Novaseq 6000 (Australian Genome Research Facility, Melbourne, Australia). High sequence depth is considered best practice for lncRNA discovery, as they are often expressed at low levels and have poor isoform representation.[86] For long-read sequencing, RNA was extracted from one FTSEC and one OSEC cell line (Table S2). ONT cDNA libraries were generated with polyadenylation enrichment and SQK-NBD114.24 native barcoding kit at the Garvan Institute's Nanopore Sequencing Facility (Australia). Samples were barcoded using the supplied PCR barcodes and sequenced at high depth (> 69 million reads per sample; Table S2) on the PromethION P48 flowcells (FLO-PRO114M - R10.4.1). The slow5 files were base-called using Guppy v.6.4.6+ae70e8f and MinKNOW v.22.12.5 by the Garvan Institute's Nanopore Sequencing Facility (Australia).

Additional datasets for benchmarking

In addition to the in-house RNAseq dataset described above, we have used two additional datasets for benchmarking HyDRA. These include uveal melanoma data from GSE206464 (BioProject PRJNA851027)[56] and breast cancer data described in Bitar et al.[4]

Databases and reference genome versions

We used the human genome GCRh38 release 79[53] for lncRNA identification and annotation, and the T2T-CHM13 genome[52] for long-read effects on the produced transcriptome assembly. The GENCODE annotation for GCRh38 was used as the reference transcriptome. A previously published database of 169 human rRNA sequences,[4] with the addition of two sequences identified from our FastQC analysis (Table S3), was used to filter pre-processed reads for ribosomal contamination. To identify which of the discovered lncRNAs were known from the reference and which were present only on the custom assembly, we used all annotated lncRNAs in GENCODE GRCh38 together with a comprehensive database of >112,000 known lncRNA genes from 8 public repositories (BIGtranscriptome, MiTranscriptome and LNCipedia from lncRNAKB[87]; Cabili et al.[88]; CancerSEA[89]; Lanzos et al.[90]; LncRNADisease[91] and RNAcentral[92]), as described in Bitar et al.[4] Experimentally confirmed protein-coding and lncRNAs annotated in the same GENCODE version were used to train the machine-learning algorithms FEELncfilter, FEELnc_codpot and FEELnc_classifier.[37]

Parameters used for the ovarian and fallopian tube custom assembly

A comprehensive list of the parameters used in each step is available in Table S1). Importantly, we used default parameters in most of the tools and steps throughout HyDRA, having only changed these to be stricter during the quality control threshold for long-read data. In line with this, we defined both a restrictive and permissive set of cut-offs for read support. A strict read support of 3 RPKM was enforced for monoexonic transcripts, but we relaxed the cut-off to 0.3 RPKM for multiexonic transcripts. This is in agreement with Bitar et al.[4] and maintained the expected ratio of multiexonic to monoexonic lncRNAs of 14:1, consistent with annotations based on the T2T genome. A stricter threshold of 5 RPKM and 1 RPKM, respectively, was also tested. Importantly, lncRNAs expressed at ~0.5 RPKM had previously been experimentally confirmed by our group with an 80% success rate.[93]

Expression profiles of lncRNA transcripts

RNAseq analysis was run based on the GRADE (General RNAseq Analysis for Differential Expression) pipeline[4] and TPM counts derived from STAR alignments,[83] sorted with SAMtools[67] and NovoSort (Novocraft Technologies Sdn Bhd, Malaysia, http://www.novocraft.com/), and quantified with RSEM.[84] Modifications to the updated version (GRADE2) scripts now allow the user to quantify reads based on any user-provided transcriptome sequence and are available online.[55] The expression profiles of lncRNAs were assessed in (i) ovarian and fallopian tube whole tissue; (ii) high-grade serous ovarian carcinoma (HGSOC) tumor samples, including homologous recombination (HR)-deficient and HR-proficient; and (iii) HGSOC cell lines. Public RNAseq of ovarian and fallopian tube samples, sequenced at high depth, were obtained from the RNA Atlas project.[54] Transcripts with at least 0.1 TPM counts in a sample were considered as expressed.

Bioinformatics tools used for HyDRA's development and benchmarking

Our pipeline integrates 39 open-source tools[12,20,28,32–34,36–38,41,57–80] and runs through a series of BASH scripts. A comprehensive list of the all tools used in each step is available HyDRA GitHub repository[26] (https://github.com/isabela42/HyDRA, Table S1). HyDRA was developed using an x86_64 GNU/Linux operational system in an HPC environment where computational tasks are allocated in PBS. We have also implemented a Docker container image with all tools/versions used in this study, which can be converted into a Singularity/Apptainer container image, as described on the HyDRA GitHub repository. Resources used for pipeline development are indicated in each script and can be controlled by the user according to their available computational power. The length evaluation script, fasta_splitter.py, was developed in Python 2.7+. To assess HyDRA's assembly, we have benchmarked several assembly methods using the normal primary and immortalized FTSEC and OSEC, breast cancer and uveal melanoma datasets. We have tested short-read-only Bitar et al.[4] pipeline-based and StringTie2[45]), long-read-only (StringTie2[45]) and hybrid (rnaSPAdes[12] and StringTie2[45]) assembly methods based on HyDRA treated reads of the described datasets. Multiple long-read error correction tools were also benchmarked (FMLRC2,[20] HGCoLoR[81] and LoRDEC[82]). BLAT (BLAST-like alignment tool implemented at the UCSC genome browser) searches[42] were performed to confirm rRNA sequences, long-read sequences and investigate identified lncRNAs. Plots were produced either directly by the underlying tools (referenced in text) or with Python 2.7+. [26]Venn Diagrams were generated with InteractVenn.[85] Figures were edited in Adobe Illustrator v.28.5.

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analysis performed are underlying the tools used for the development and benchmarking of HyDRA. The used datasets are described in Table S2. The profiling of the HyDRA lncRNAs discovered from the ovarian metatranscriptome was performed using the GRADE2 pipeline (RSEM-based). Their expression profiles were assessed in public RNAseq of ovarian and fallopian tube samples from the RNA Atlas project[54] (Table S5). Transcripts with at least 0.1 TPM counts in a sample were considered as expressed.