phaeno

WHITE PAPER

PSEQ WHITE PAPER SERIES · PAPER 1 OF 4

PSeq: Whole-Molecule Phased RNA Sequencing on Standard Short-Read NGS

Platform architecture, workflow, and data model

July 2026 | Version 1.0

Executive summary

PSeq is a research-use only RNA-sequencing platform that brings whole-molecule resolution and end-to-end phasing to standard short-read next-generation sequencing (NGS). During reverse transcription, a source-molecule tagging reagent labels each RNA sequence, and a library strategy preserves the tag marker across random fragments. Paired-end sequencing runs on unmodified instruments, while an automated pipeline groups reads by Source Molecule Identifier (SMID), identifies the source gene, assembles a consensus sequence, and retains supporting records in a run-specific SQL database.

This paper, and the three papers that follow, presents PSeq's core innovation, architecture, and operating model.

Central proposition: PSeq is designed to combine the throughput and base-level accuracy of short-read NGS with molecule-resolved, end-to-end transcript reconstruction and evidence traceability.

PSeq at a glance

- Runs on unmodified short-read NGS instruments.
- Positions a high-diversity SMID plus strand-sense discriminator (SD) for each molecule on random internal fragments derived from each individual transcript.
- Assorts fragment-reads by source molecule tag followed by identifying source gene and assembling the consensus sequence.
- Retains raw reads, intermediate analyses, consensus sequences and run metadata in a sample-run database.

The short-read linkage problem

Conventional RNA-Seq achieves high throughput by sequencing short fragments prepared from a heterogeneous RNA sample. It can identify expressed genes, quantify read abundance, and discover exon junctions. However, when two sequence features are separated by more than the fragment read length, the data no longer directly discern whether those features occurred on the same source molecule.

In RNA-Seq, transcript identity and isoform abundance are inferred from populations of fragments with statistical models and sample ground-truth prior information. While powerful, with this approach inferred transcripts conflate evidence from multiple molecules, sometimes including multiple isoforms. RNA-Seq does not sequence single molecules.

Long-read instruments address linkage more directly but differ NGS in throughput, data assembly and error characteristics.

PSeq combines the accuracy and throughput of NGS with the length capabilities of long-read platforms. This is accomplished by adding a source molecule identifier that survives library preparation and sequencing, and guises computational reconstruction with higher accuracy and scale.

PSeq Clear-Signal Architecture™

PSeq Clear-Signal Architecture™ integrates molecular information in DNA libraries with computational software required for true single-molecule sequencing, at very large scale. This architecture further permits in depth traceability of each step in data assembly for each molecule of a sample.

CSA is based on three linked principles:

- **Identity is established before fragmentation.**
- **Reads are grouped by source molecule before reconstruction.**
- **The reconstructed molecule remains connected to its supporting evidence.**

Together, these principles convert a mixed short-read data set into distinct, independently inspectable molecular records.

PSeq design innovation principles

- Identify before fragmenting. A source-molecule tag is introduced during reverse transcription, before the cDNA is converted into a sequencing library.
- Preserve molecular provenance. Random fragments derived from the same tagged cDNA carry the same marker block and can be grouped together after sequencing.
- Separate identity from sequence inference. The SMID establishes the molecular grouping; gene identification and data assembly then operate within that group.
- Retain the evidence chain. Final molecular records are designed to remain traceable back through alignments and intermediate files to the originating FASTQ records and base-quality information.
- Scale on familiar infrastructure. Library sequencing uses paired-end NGS, while computation is designed to scale with highly parallel processing on cloud resources.

End-to-end operating model

To initiate sample preparation, heterogeneous RNA is reverse transcribed, attaching a multifunctional tagging reagent. Each tag contains a high-diversity SMID plus structural elements that report strand sense. The tagged reverse transcript is amplified, fragmented through reactions confined to intramolecular rearrangements, coupled to conventional sequencing adaptors, and converted into a paired-end PSeq DNA library.

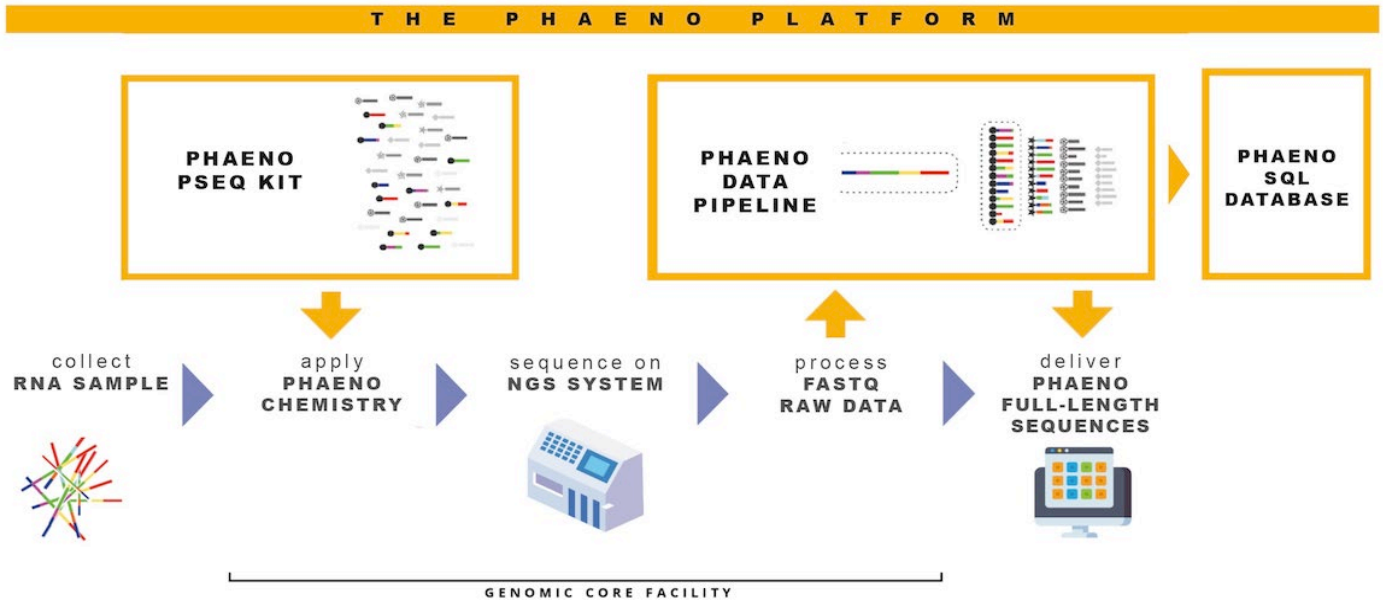

Figure 1. PSeq fits a familiar genome-core workflow: proprietary library preparation and automated data assembly are added around conventional paired-end NGS.

After sequencing, raw FASTQ files enter the PSeq data pipeline. Marker tags are localized, reads are clustered by SMID, tag sequences are trimmed, strand sense is recorded, redundant reads are removed, and the remaining sequences are used for source-gene identification and transcript assembly. Records of all Intermediate and final steps in data assembly populate a run-specific SQL database.

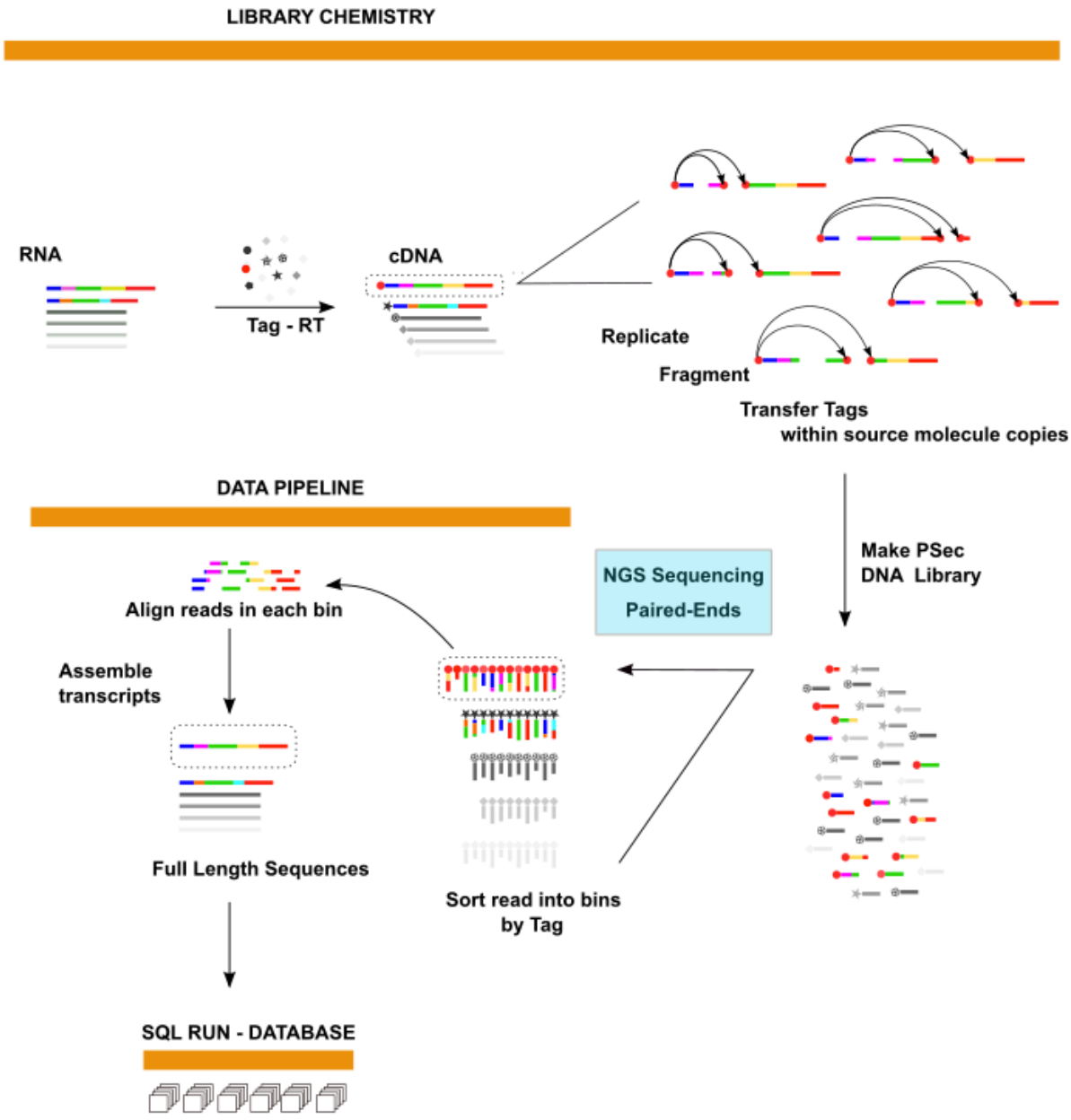

Figure 2. Conceptual PSeq workflow from RNA tagging and library construction through paired-end sequencing, SMID-based read grouping, sequence assembly, and SQL run-database output.

Platform components

Molecular tagging and library chemistry

PSeq DNA library protocols attach a tagging reagent to the reverse-transcript of each RNA. Subsequently, a marker from the reagent is redistributed to random fragments of the cDNA. The marker includes one SMID barcode plus strand discriminator. The resulting DNA library therefore comprises possesses random cDNA fragment sequences linked to source molecule identity and the source molecule's strand sense.

Automated data pipeline

After paired-end NGS sequencing, the data pipeline sorts the mixed population of read pairs into molecule-specific bins. For each SMID bin, the source gene is identified, and the reference gene sequence is retrieved. The pipeline generates gene and transcript alignments and a final, de novo consensus sequence of the individual transcript, across all transcripts.

Run-specific SQL data product

The SQL database produced by the pipeline compiles client and run meta data, total raw FASTQ data records for the entire run and files for each sequenced molecule, identified by SMID barcode.

SMID-specific files retain records from raw FASTQ data through library protocols, sequencing and data assembly. Formatted files can be used to visualize and validate each molecular assembly. The SQL database further provides links to external gene resources.

What changes when the molecule becomes the unit of analysis

Conventional NGS expression profiling aggregates evidence across a population of inferred transcripts. PSeq adds a second level of analysis in which each source molecule has its own evidence bundle.

The new level of information supports in-depth inquiries:

- Which reads contributed to this reconstructed molecule?
- Which gene and strand were assigned to the molecule, and on what evidence?
- Which (phased) splice junctions, mutations or other sequence variations occur in the same molecule.
- Where do individual aligned reads disagree with the consensus or reference sequence?
- Can an unusual base call, junction, or assembly decision be traced back to its original FASTQ record and quality score?

Deployment

PSeq operates within conventional workflows of sequencing laboratories and genome cores equipped for conventional NGS service.

PSeq reagents, kits and protocols are comparable to those for RNA-Seq, WGS and WES.

The PSeq v1 version of the automated data pipeline is installed on Amazon EC2 instances with local reference databases, and is highly scalable. A 6 to12-hour analysis cycle is predicted, independent of scale, but which could vary with compute configuration.

The run-specific SQL database produced by the pipeline compiles information for client reports and provides complete, traceable information for in depth data validation.

PSeq Deliverables

Client meta data (FASTA)

Run meta data (FASTA)

Demultiplexed raw files (FASTQ)

Tagged FASTQ records

Transcript inventory (.gtf)

- Novel isoforms
- Registered isoforms

Source Gene Inventory (.gtf)

- Novel annotation
- Registered, expanded annotation
- Registered, unmodified

Isoform sequences (.bam)

- Consensus
- trimmed reads

RSeQC (.html)

Deliverables summarized in a PDF report.

Intended context and current boundaries

PSeq v1 will be limited to research-use only (RUO) implementation. Detailed kit compositions, reagent specifications, laboratory recipes, and operational quality controls will be archived in design control documentation. Performance documentation will be presented separately.

The PSeq workflow is currently configured to conform to protocols for RNA-Seq, WGS, WES but not scRNA-Seq. PSeq will be adapted to single-cell workflows in subsequent versions.

Interpretation boundary: PSeq should be evaluated as an integrated chemistry-and-computation system. Platform diagrams describe the intended workflow; they do not by themselves establish accuracy, reproducibility, sensitivity, or fitness for a clinical purpose.

Conclusion

High-throughput short-fragment NGS is an immensely advance form of shotgun sequencing. NGS has not previously been capable of true single-molecule sequencing. .

Phaeno Clear Sequencing Architecture integrates molecular information in specialized DNA libraries with computational software required for true single-molecule sequencing., at very large scale. The partitioning of data assembly for individual molecules enables execution of SQL inquiries for in depth data traceablity and QC validation. Population level single-molecule sampling permits validation of library parameters, such as occurrence of chimeric constructs, missed events and proportional sampling.

If the molecular labeling, read grouping, and reconstruction steps perform as intended, the result is a scalable route to phased, molecule-resolved RNA records executed on familiar sequencing infrastructure.

Companion white papers in this series discuss library chemistry, automated data assembly, and and initial validation of platform performance.

Notes

1. The platform can be implemented on a diversity of NGS sequencers, e.g., but not limited to Illumina NGS, Complete Genomics DNB sequencers, Element Biosciences AVITI instruments.
2. The platform overview is derived from the internal PSeq v1 technical report dated July 21, 2026. Operational specifications remain subject to finalizing PSeq v1 SOPs and expanded QC metrics.
3. Research-use context only. This paper does not infer clinical validity or regulatory clearance.
4. RSeQC refers to a comprehensive open-source RNA-Seq Quality Control package to be modified for PSeq data structure. (see https://rseqc.sourceforge.net/index.html).

PSeq white paper series

PAPER 1	PAPER 2	PAPER 3	PAPER 4
PSeq: Whole-Molecule Phased RNA Sequencing on Standard Short-Read NGS	PSeq Molecular Tagging and Library Architecture	PSeq Data Pipeline	Initial Technical Validation of the PSeq Whole-Molecule RNA Sequencing Platform.