phaeno

WHITE PAPER

PSEQ WHITE PAPER SERIES · PAPER 2 OF 4

PSeq Molecular Tagging and Library Architecture

Design principles for preserving source-molecule identity and strand sense

July 2026 | Version 1.0

Executive summary

Whole-molecule reconstruction from short reads requires more than sequencing depth: the fragments must retain a reliable connection to the molecule from which they originated. PSeq addresses that requirement by introducing a multifunctional tag during reverse transcription and redistributing its marker among random fragments generated from the tagged cDNA. The marker carries a high-diversity **S**ource **M**olecule **Id**entifier (SMID), fixed sequence landmarks, and strand-sense information, all of which can be recognized after paired-end sequencing.

This paper describes the molecular design and expected read structure of the PSeq library. It is an architectural overview of the PSeq chemistry.

Design objective: Fragment a source cDNA into a conventional short-read library without losing the identity or strand sense of the RNA molecule from which each fragment arose.

The identity-preservation problem

Random fragmentation is central to short-read library preparation, but it destroys physical continuity. Once fragments from many transcripts are mixed, local read sequences alone cannot always establish which distant exons, sequence variations, or untranslated functional domains belonged to the same original RNA. PSeq Clear-Signal Architecture™ is designed to preserve that continuity by establishing source-molecule identity before fragmentation. During reverse transcription, each source RNA is assigned a Source Molecule Identifier (SMID). Subsequent reactions amplify the tagged cDNA and distribute a recognizable marker to random fragments through reactions intended to remain intramolecular. After sequencing, fragments carrying the same SMID can be grouped computationally and used to reconstruct a molecule-specific RNA record.

Anatomy of the PSeq Tag

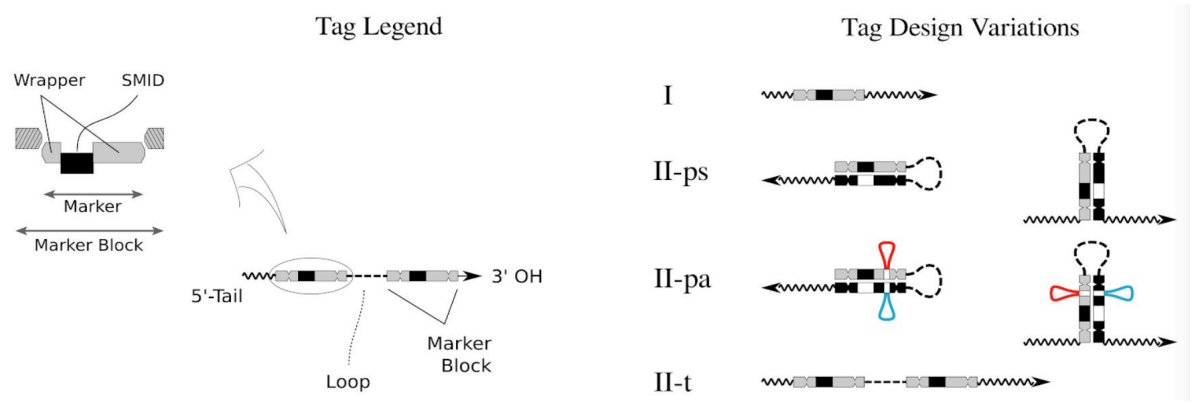

Figure 1. PSeq tag legend and representative design variations. The marker block contains the sequence elements used to recover molecular identity after sequencing.

Source Molecule Identifier

The SMID is a high-diversity barcode composed of random-base triplets separated by invariant check bases. The source report describes reagent designs with barcode diversity ranging from approximately 68.7 billion to 4.3 trillion possible identifiers, depending on the number of random triplets used. The practical purpose of that diversity is to make the identifier space much larger than the number of molecules expected in a sequencing run.

Wrapper and check bases

Fixed 5' and 3' wrapper sequences provide landmarks for locating the marker in a raw read. Invariant check bases distributed through the barcode supply additional positional structure. Together, these elements help the pipeline distinguish a valid marker from unrelated cDNA sequence and establish the boundaries of the tag before trimming.

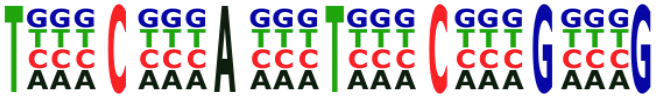

Figure 2. Example SMID motif with six random-base triplets punctuated by invariant check bases.

Strand-sense discriminators

Strand-sense discriminators (SDs) encode the orientation of the source molecule. In the Type II-pa design illustrated below, asymmetric structural features provide the positive- and negative-strand signal. Retaining strand sense is important for gene assignment, transcript interpretation, and the detection of antisense species.

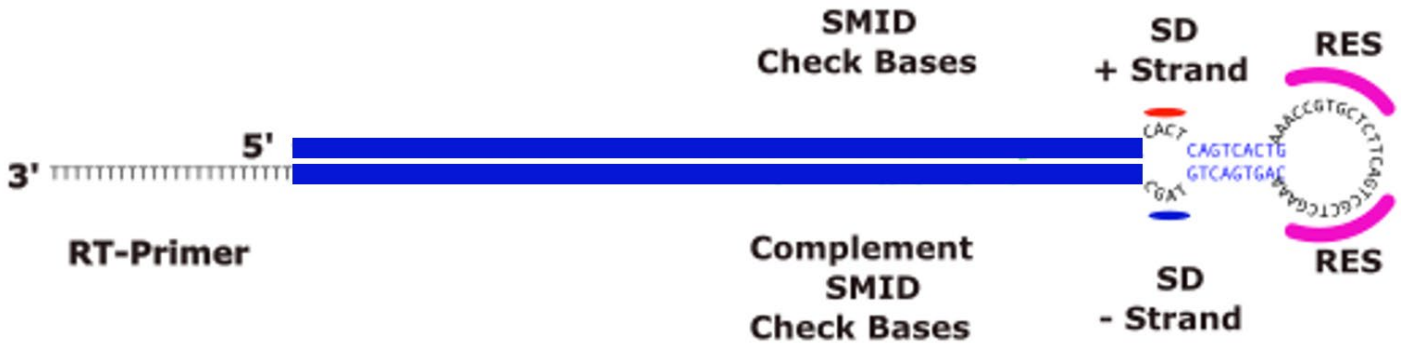

Figure 3. Representative Type II-pa tagging reagent with RT primer, SMID/check-base region, strand discriminators, and restriction-enzyme sites.

From RNA to a PSeq DNA library

1. Reverse transcription and tagging. Heterogeneous RNA is reverse transcribed in a reaction that incorporates a multifunctional PSeq tagging reagent into each cDNA.
2. Identity encoding. The tag assigns a SMID and strand-sense information (SD) before the source molecule is fragmented.
3. Non-PCR isothermal amplification. Tagged reverse transcripts are amplified to provide material for library construction while preserving the tag-to-molecule relationship.
4. Intramolecular fragmentation and marker transfer. Amplified cDNA is fragmented in free solution under reaction conditions intended to redistribute the marker block among fragments from the same source cDNA without exchanging markers between molecules.
5. Adaptor coupling and library amplification. Tagged fragments are coupled to conventional sequencing adaptors and amplified to produce the PSeq DNA library.
6. Paired-end sequencing. The completed library is sequenced on an unmodified paired-end NGS instrument.

Expected sequencing-read contract

The library design creates a specific relationship among the marker, Read 1, Read 2, and the source transcript. That relationship is the data contract consumed by the PSeq pipeline.

Two structural exceptions require explicit handling. First, a genuine gene-fusion transcript can contain sequence from more than one gene. Second, a short insert can cause Read 1 and Read 2 to overlap or capture part of the marker block more than once. These cases should not automatically be classified as library failure.

Element	Expected structure	Computational use
Read 1 start	Marker block located at or near the 5′ start	Locate the wrapper, SMID, check bases, and strand discriminator
Read 1 remainder	Random cDNA fragment sequence	Contributes random internal transcript sequence after the marker is trimmed
Read 2	cDNA fragment sequence	Provides paired sequence evidence without an expected leading marker
Read pair	Read 1 and Read 2 ordinarily derive from the same source gene and strand	Supports gene assignment and molecule-specific assembly
Shared SMID	Multiple fragments carry the same source-molecule code	Defines the molecule-specific bin used by downstream algorithms

Design variants and customization

The PSeq platform describes multiple tag architectures for different applications. Type II designs contain two copies of one barcode in tandem marker blocks, while palindromic, asymmetric structures provide strand-sense reporting. The common requirement is not a single physical shape, but a machine-recognizable marker that preserves source-molecule identity across library transformation.

Kits, reagents, and protocols may support end-user customization of tagging reagents and sample-preparation workflows. However, customization must preserve the sequence landmarks consumed by the data pipeline and should be qualified against the same structural and molecular structure criteria used for the standard design.

Interfaces with the data pipeline

Library chemistry and computation are intentionally coupled. The wet-lab process creates the identity-bearing read structure; the data pipeline must then locate the marker, validate its landmarks, recover the SMID and strand signal, trim the tag, and group the resulting cDNA reads. A chemistry change that moves a wrapper, changes an invariant base, or alters strand encoding is therefore also a software-contract change.

The most critical interface is preserving intramolecular association. If marker tags were exchanged among unrelated cDNAs, one SMID bin could contain fragments from multiple source molecules, and the apparent reconstruction would be chimeric. Demonstrating single-gene and single-molecule confinement therefore requires downstream sequencing and alignment evidence; it cannot be established from the reagent diagram alone.

Controlled information and operational SOPs

This architectural paper intentionally excludes proprietary reagent compositions, reaction concentrations, incubation conditions, acceptance thresholds, lot-release criteria, and troubleshooting instructions. Those details are in controlled documents that will be detailed with the specific kit and data pipeline release.

Change-control principle: The tag sequence specification, library protocol, marker parser, and SMID/strand data model work as one integrated interface. A change in any one component will trigger a compatibility review to ensure assay performance.

Conclusion

PSeq library architecture is designed to make original molecule identity survive short-read library preparation. A structured, high-diversity marker is installed before fragmentation; random fragments inherit the marker; paired-end reads carry sequence evidence; and the pipeline uses the recovered SMID to rebuild molecule-specific records. The design is conceptually simple, but its success depends on tag recognizability, barcode diversity, intramolecular marker transfer, and a stable chemistry-to-software contract. Paper 4 examines the initial evidence for those properties.

Notes

1. Code-space values, tag names, and molecular structures are reproduced from the internal PSeq v1 overview dated July 21, 2026 and should be confirmed against the controlled reagent specification before external publication.
2. The SMID design provides a large theoretical code space; empirical collision rate, synthesis bias, sequencing-error tolerance, and effective usable diversity require separate quantitative validation.
3. Research-use context only. This paper is an architectural description and does not disclose operational SOP parameters.

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.