phaeno

WHITE PAPER

PSEQ WHITE PAPER SERIES · PAPER 4 OF 4

# Initial Technical Validation of the PSeq Whole-Molecule RNA Sequencing Platform

DNA library structure • Computational assembly • Data traceability

July 2026 | Version 1.0

## Executive summary


The PSeq platform begins with preparing DNA libraries with a specialized informational architecture. Conventional NGS paired-end sequencing then provides the automated data pipeline with all the information required to assemble each individual RNA.

Together, these elements form PSeq Clear-Signal Architecture™, an integrated chemistry-to-data framework for whole molecule sequencing.

This paper presents initial technical validation for PSeq v1 across the entire chain of operations.

In a case study, data assembled for a single transcript are readily visualized with an open-source gene viewer (IGV) designed for examining whole-genome sequences and RNA-Seq data.

Stepping through this information for a single transcript from the gene PRPF31 visualizes the pipeline performance at a level of detail not otherwise available. This exercise further illustrates the traceability of all steps in the platform chain of operation.

This white paper constitutes initial technical validation, not to be construed as a statistically complete platform validation, clinical validation or regulatory performance analysis.

# Library validation

1. PSeq DNA library protocols reproducibly generate DNA libraries with the intended size distribution.
2. PSeq markers are efficiently incorporated at the optimal 5' position of Read 1.
3. Paired-end FASTQ reads conform to the intended marker-plus-cDNA structure.

Confirming that library chemistry confines source molecule identifying SMIDs are confined to (1) one gene and (2) one transcript from that source gene requires automated pipeline data assembly.

Confirming additional population-level library structure requires running the data pipeline after library sequencing: viz, are reads linked to each SMID confined to one source gene and one transcript molecule from that gene?

# Evidence Summary

| Validation Question | Evidence reported | Initial observation | Primary limitation |
|---|---|---|---|
| Library quality | Bioanalyzer trace from two independently prepared pooled-library replicates | Library distribution in the intended sequencing range | Quantitative replicate and lot summary to be reported |
| Marker location | Visual display of Read 1/Read 2 marker randomly sampled in a 9-million-read-pair block; 300 sampled records presented | Markers concentrated in Read 1 positioned near 5′ start | Records susceptible to inadvertent computational selection or filtering. |
| Marker chemistry | Bulk-library Sanger sequencing from the Read 1 adaptor primer | Expected wrapper/check-base pattern is visible ruling out computational distortion. | Random barcode confirmed, true barcode diversity is not quantified. |
| Read structure | Representative paired FASTQ record assigned to NUDT18 | Marker at Read 1 start followed by cDNA; Read 2 contains cDNA originating from same gene. | Single illustrative record |
| Molecule confinement | PRPF31 SMID-bin BAM review in IGV | Reads align over one gene and one transcript, covering all eon splice junctions: full phased sequencing. True shotgun sequencing of single molecule from heterogeneous samples. | Single representative transcript case study |
| Traceability | IGV mismatch, quality, read-detail, and SQL-record inspection | Base-level discrepancies can be trace to FASTQ read and quality score | Evident technology leading consensus accuracy to be confirmed in independent experiments. |

# Library size-distribution quality control

A representative Bioanalyzer scan of two pooled, independently prepared library-replicates generated from human universal reference RNA standards shows size distributions compatible with paired-end NGS sequencing employed by the PSeq platform.

This observation confirms well resolved library size distribution but does not illuminate information embodied in the read fragments. A formal release-method will include size-range acceptance criteria and failure-thresholds, together with a standard profile of read quality vs position, adapter burden and duplication rates

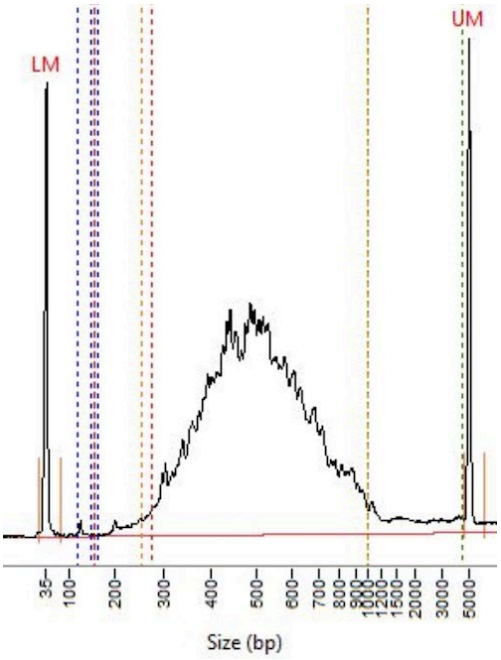


*Figure 1. Representative Bioanalyzer size-distribution trace for pooled PSeq library replicates. LM and UM denote lower and upper size markers.*

# Marker placement in PSeq library fragments

A file of 9 million read pairs was sampled at intervals of 30,000 records, displayed as 300 read pairs. Tagged reads are highlighted and the 5' wrapper of marker tags are in bold. Reproducibly, tagging efficiency ranges from 85% to 98%. Essentially all markers occur in Read 1. Routinely, more than 96% of markers lie within the first five bases of Read 1. These values are lower limits given that only exact marker matches are highlighted.

This result is consistent with the intended library structure. The tagging efficiency and positional statistics are computed from the total 9 million read population, but encompassing similar samples from separate experiments. While the image highlights exact matches, tag efficiency is computed allowing up to 2 random errors per 57 invariant bases in the attached marker

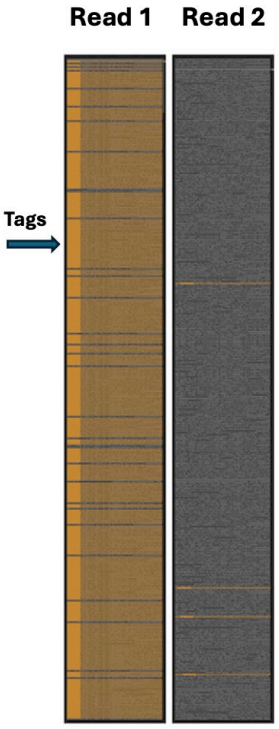


*Figure 2. Representative positional survey of marker matches in paired-end reads. Signal is concentrated at the beginning of Read 1 rather than Read 2.*

# Sanger confirmation of tagging uniformity

NGS records are susceptible to inadvertent filtering or selection. To assess true uniformity of the DNA libraries, the solution of library fragments can be subjected to Sanger end-sequencing from the A primer adaptor site for Read 1.

If markers occur uniformly at the expected 5' position, the bulk chromatogram will show fixed wrapper sequences of the marker flanking the distinctive SMID pattern of random base triplets and invariant check-base signals at their designed positions. Barcode triplets will comprise A, T, C, and G peaks of equal abundance. This is fully confirmed.

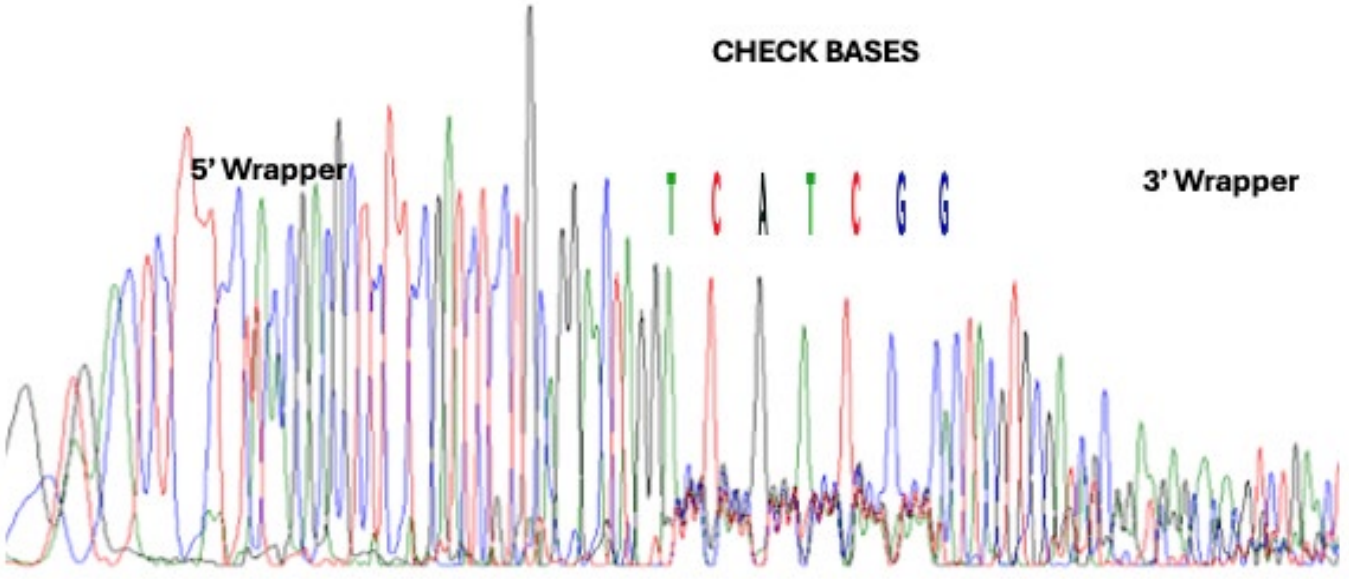


*Figure 3. Representative bulk-library Sanger chromatogram used as an independent qualitative check of marker position and barcode/check-base structure.*

The observed pattern indicates consistent preferential marker incorporation at the start of Read 1 and the SMID barcodes with random triplets punctuated by invariant bases.

The method cannot quantify barcode diversity; it does confirm that all four bases occur with roughly equal likelihood at the intended random base positions and that check bases exhibit the expected periodicity. This confirms the absence of systematic computational distortion of the library profile.

# FASTQ read composition

The remaining structural specifications for PSeq library fragments are:

1. Following the marker block of Read 1, the remainder of Read 1 will embody a random fragment of a transcript cDNA.
2. Read 2 will consist entirely of cDNA random fragment sequence
3. Read 1 and 2 will report the same source gene and strand sense, except:
    a. in the instance of rare transcripts from gene fusion events.
    b. when short-library fragments result in Reads 1 and Reads 2 reporting all or part of the same marker: that is, when the paired ends cover the same SMID from both ends.

A representative raw data FASTQ record illustrates:

1. the marker (tag) motif is located at the beginning of Read 1
2. The continuation of Read 1 is a cDNA sequence
3. All Read 2 is a cDNA sequence.
4. Read 1 and Read 2 report the same transcript gene source and strand sense.
5. To the extent tested to date, single read pairs containing cDNAs from two genes (chimeras) have not been encountered.


*Figure 4. Representative paired FASTQ record illustrating the Read 1 tag, adjoining cDNA sequence, and cDNA sequence in Read 2.*

This record illustrates the prescribed canonical structure, e.g.

1. 5' tag motif linked via a short poly A connector to a cDNA in Read 1
2. Read 2 entirely comprising a cDNA fragment
3. Read 1 and read 2 cDNA components from the same source gene with the same strand sense.

This read pair incorporates cDNA sequences from NUDT18, a gene located at chromosome 8p21.3 (GRCh38 coordinates: 8:22,106,878–22,110,482), which encodes a nudix hydrolase enzyme that hydrolyzes oxidized nucleoside diphosphates, particularly those containing 8-oxo-7,8-dihydroguanine (8-oxo-Gua).

Higher order information in the PSeq libraries require running the data pipeline.

1. Does each SMID and DS associate with one an only source gene and one and only one RNA transcript from that gene and with the same strand sense?
2. Does this apply to every SMID, DR-identified read cluster in the library?

Failure to conform to these criteria could indicate that library chemistry is not confined to intramolecular rearrangements in every case – with the possibility of producing false-positives (chimeric constructs).

A final QC metric will ascertain whether the libraries reliably sample transcript populations with accurate proportionality, independent of transcript size distributions. This metric will await analyses of cumulative sample databases.

# Pipeline Validation: a PRPF31 molecule-level case study

## Case-study rationale

The pipeline-validation example examines BAM files generated for one SMID-tagged transcript originating from PRPF31 on chromosome 19. Figures are generated with the open-source IGV gene viewer designed to examine genome sequences and RNA-Seq records: (https//IVG.org/).

PRPF31 encodes a pre-mRNA splicing factor with more than a dozen annotated isoforms. As a splicing regulatory factor that, itself, is subject to regulated splicing, the canonical isoform profiled in this case study is an apt example with which to illustrate pipeline performance.

## Confinement to one source gene

The IGV view of this segment of chromosome 19 harboring PRPF31 identifies multiple neighboring genes.

Trimmed read alignments from this SMID, DR-identified transcript result in base coverage exclusively over PRPF31. In contrast, RNA-Seq records never align to only one gene.

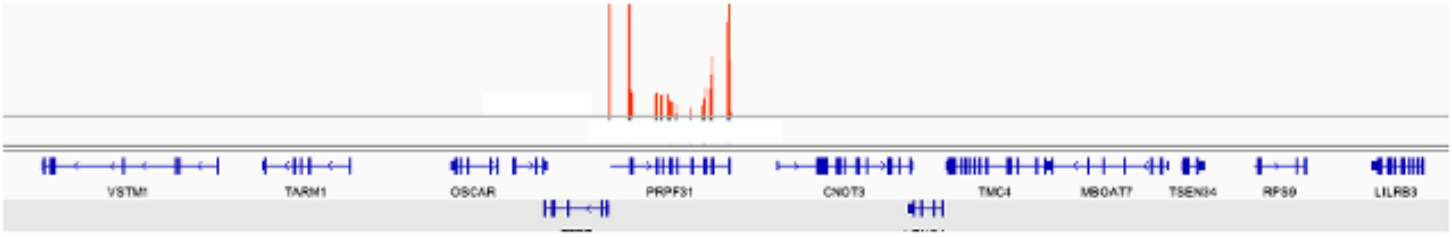


*Figure 5. The chromosome 19 neighborhood and the PSeq alignment signal for the selected SMID bin (red histogram). The displayed reads are concentrated over PRPF31 with no reads over adjacent genes.*

## Exon alignment

Bases from trimmed reads align exclusively over exons independently annotated in Ref Seq. In this isoform, all 14 Ref Seq annotated exons are present.

Thicker exon segments denote open reading frame (ORF) coding sequences, thinner exons denote 3' and 5' UTRs. The height of the red histograms may be misleading: the base with the maximum coverage of was sequenced independently 5,129 times.

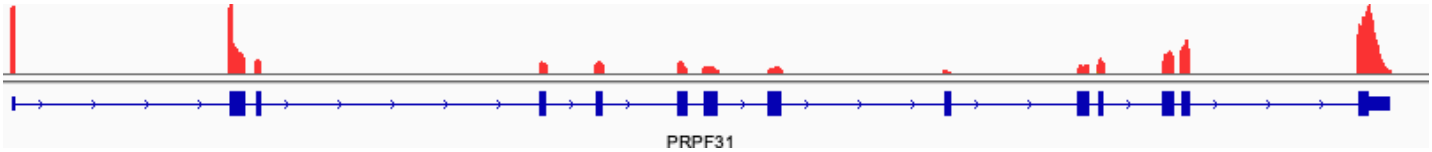


*Figure6. Frequency histograms of individual bases from trimmed reads in this SMID cluster accumulate exclusively over the exons of the PRPF31 gene. In this variant, all 14 exons are included.*

## Trimmed read alignment

Alignments of trimmed reads confirms the absence of extraneous sequences in the SMID-clustered FASTQ records for this transcript, e.g., the absence of chimeric library fragments in this data set.

The hairlines connecting aligned segments denote the coverage of splice junctions in the same trimmed read. Only a representative fraction of the aligned reads can be seen at the scale shown.

Although representative of individual gene alignments, this does not establish the complete absence of chimeras in entire single-molecule data sets

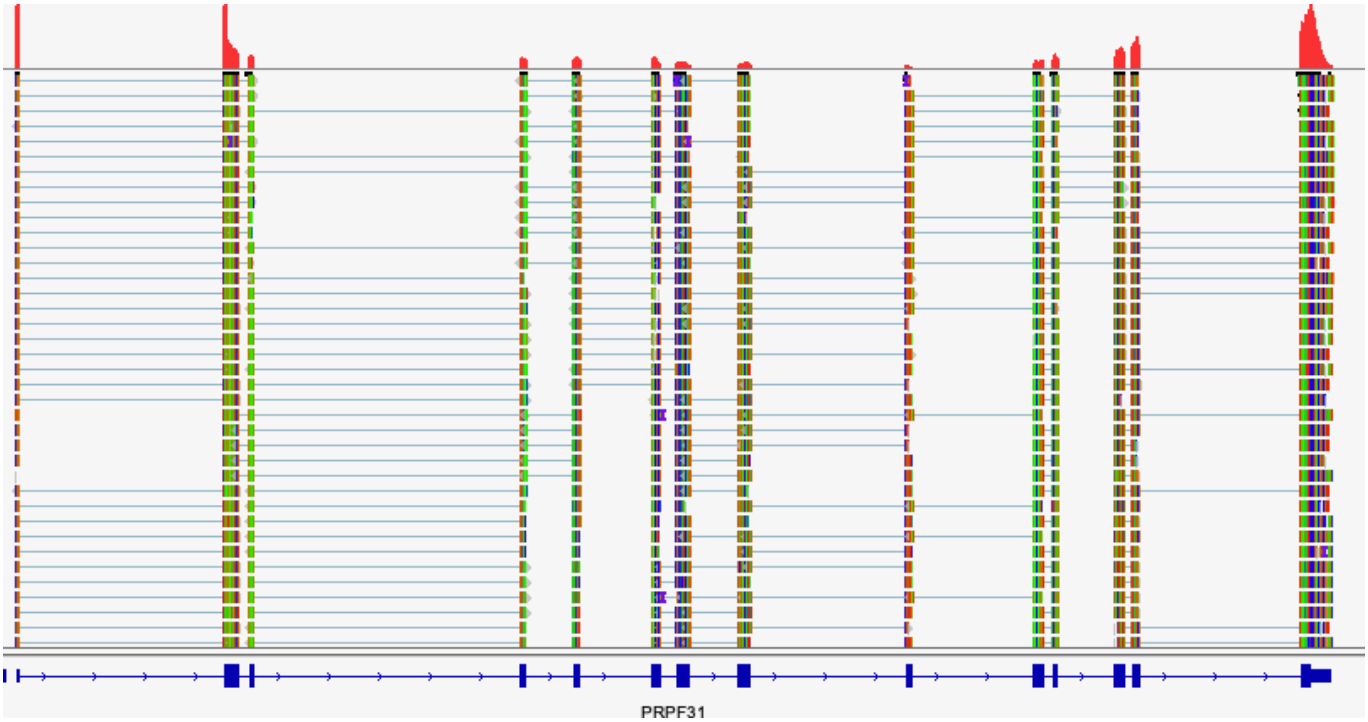


*Figure7. Frequency histograms of individual bases from trimmed reads in this SMID cluster accumulate exclusively over the exons of the PRPF31 gene. In this variant, all 14 exons are included.*

# End-to-end exon and splice-junction coverage

A compressed view of all read alignments documents the depth of coverage of exon junctions. Because all reads are identified by a single SMID barcode, one of >68 B variations, this is strong support for end-to-end phased sequencing of a single molecule

The pattern in this depiction denotes the information provided by PSeq library construction not provided by the RNA-Seq.

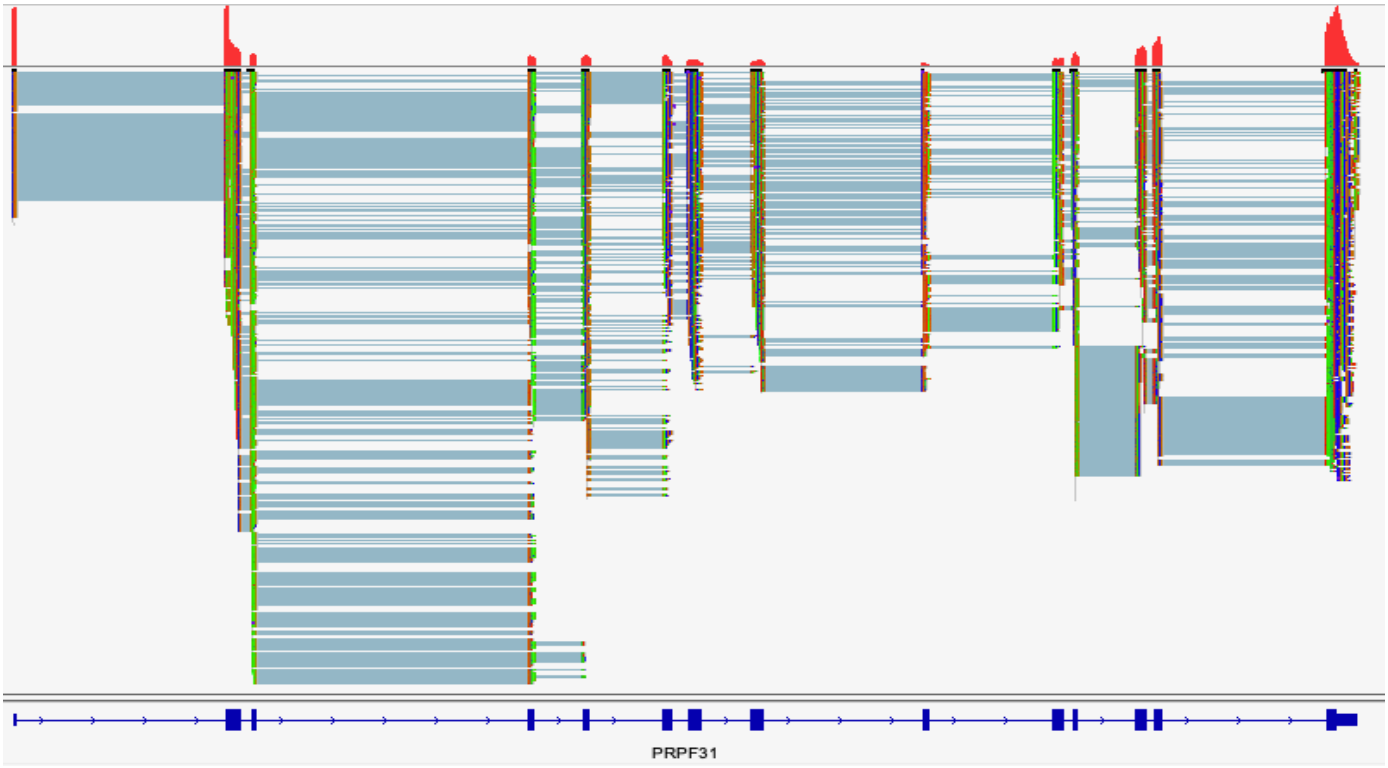


*Figure 6. Compressed IGV view of the PRPF31 SMID bin showing repeated coverage across exons and junction-spanning alignments for the selected molecule.*

An alternative view of end-to-end splice variant coverage is a Sashimi plot of trimmed read alignments.

Splice junction depth of coverage is indicated in the number written over the junctional loops. Sashimi plots for RNA-Seq data generally show multiple possible linkages, and numbers are harvested to help with statistical estimation of the occurrence of multiple isoforms. PSeq identified variants are specific and do not conflate more than one molecule, whether multiple copies of the same variant, or statistically weighted guesses

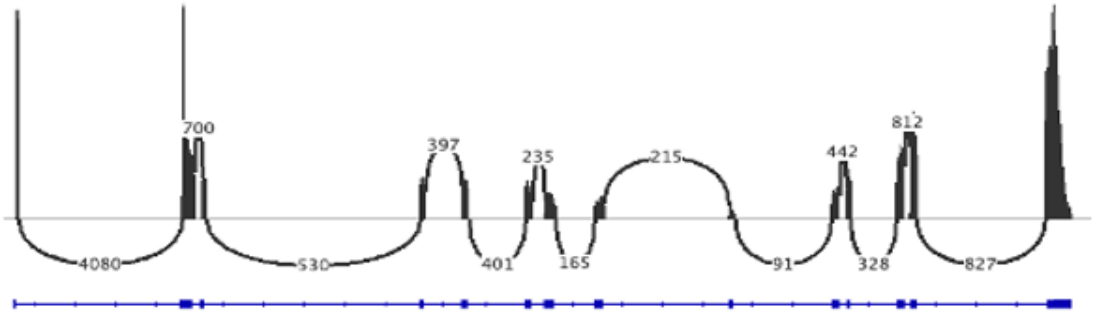


*Figure 7. Sashimi plot of the selected PRPF31 read alignment. Numbers overlying splice linkages indicated the number of splice junction spanning trimmed reads.*

Splice junction depth of coverage is indicated in the number written over the junctional loops. Sashimi plots for RNA-Seq data generally show multiple possible linkages, and numbers are harvested to help with statistical estimation of the occurrence of multiple isoforms. PSeq identified variants are specific and do not conflate more than one molecule, whether multiple copies of the same variant, or statistically weighted guesses.

A further QC metric will be assessing the average depth of coverage of the transcript population as a function of transcript length. These metrics can help adjust steps in the library protocol to adjust for target length, e.g. sequencing very long RNA virus genomes.

## True single-molecule NGS shotgun sequencing

The PRPF31 case study illustrates that the consensus sequence results from true shotgun sequence of individual transcripts. With progressive expansion of read alignments over a portion of exon 14, the independent sequencing of each base is fully evident.

NGS base calling exhibits the expected **intrinsic accuracy** of the Illumina platform of > Q30, e.g. one error per 1,000 called bases. Specifically, because this applies to a single molecule for which literally the same base is sequenced repetitively, the **consensus accuracy** is predicted to scale as the intrinsic accuracy scaled exponentially according to depth. To be confirmed independently by experiment, this suggests that reverse transcripts are sequenced effectively without platform error.

When accuracy, defined as the average length of sequences before the encountering the first error, greatly exceeds the size of the target RNA, true single molecule sequencing, can be achieved, e.g. whole-genome virus RNA surveillance.

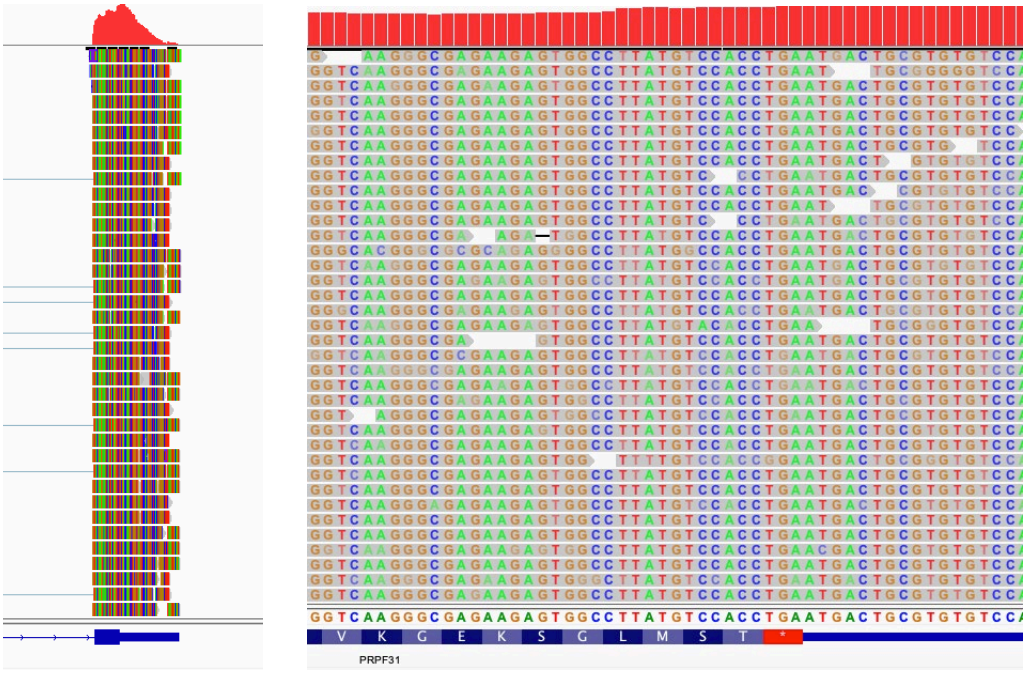


*Figure 8. Progressive inspection of PRPF31 exon 14, from a compact pileup (left) to base-resolved alignment and reference context (right).*

## Base-level error visualization and traceability

The PRPF31 case study also illustrates how errors remain visible rather than disappearing behind the consensus. Errors are visualized as mis-matched bases. Here, color intensity encodes the quality score with which the base was called in the original FASTQ record. In the assembled data, each called base is linked to the individual raw data FASTQ read pair, with the quantitative measure of read quality.

Traceability permits a reviewer to distinguish a recurrent biological difference from isolated sequencing errors and to examine how coverage depth supports a consensus.

A single case study does not quantify the distribution of molecular consensus accuracies in an entire sample. Calibration of confidence scores remain required QC endpoints.

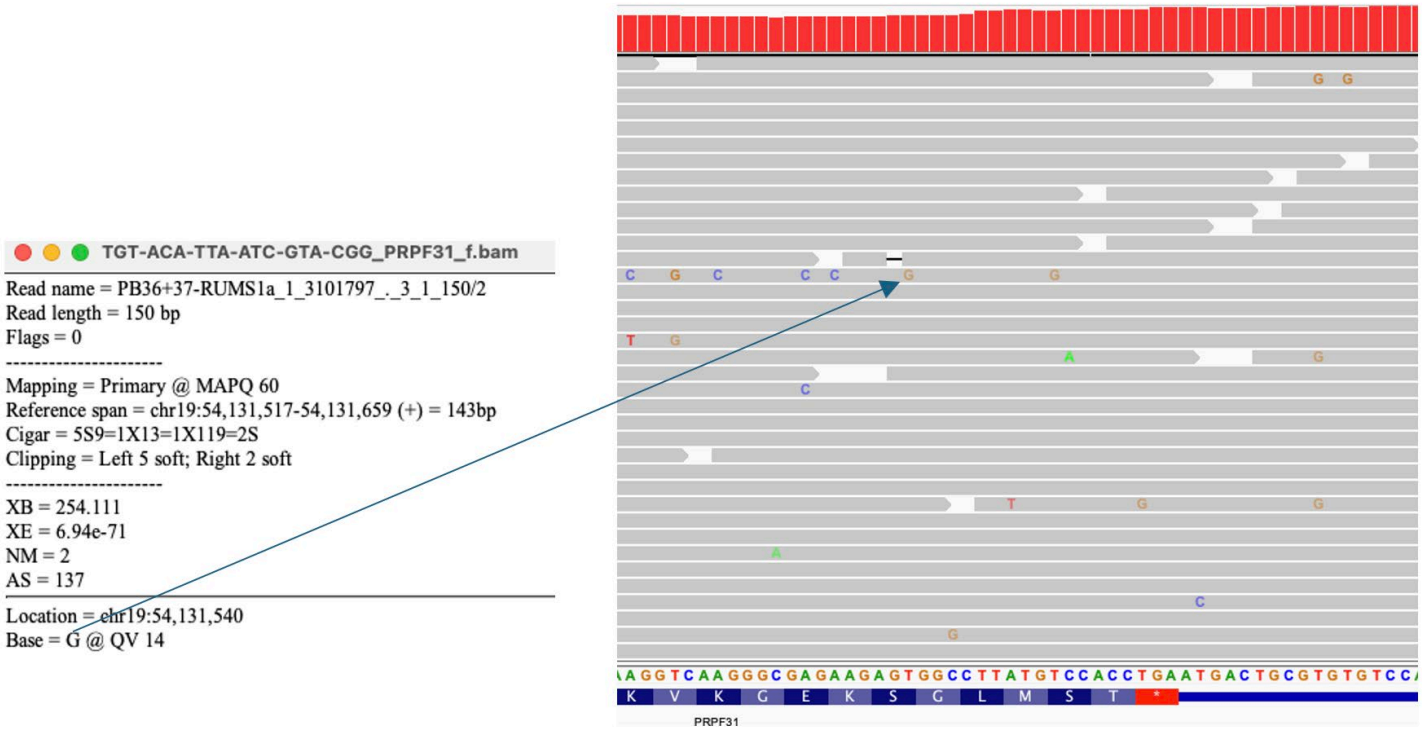


*Figure 9. Read-level metadata (left) and the corresponding alignment context (right) illustrate how discrepant bases can be traced to individual records and quality evidence.*

# Interpretation of the current evidence

Initial technical validation studies support these key conclusions:

1. PSeq chemistry produces DNA libraries of the desired structure.
    a. Average read length
    b. Positionally tagged with marker including sequencing landmarks, SMIDs, SDs, for determining the identity and strand source of individual RNA source molecules.
    c. Random internal fragments of source molecule cDNAs.
    d. No or minimal evidence for chimerization.

2. Pipeline processing of paired-end FASTQ raw data:
    a. Assembles consensus end-to-end phased sequences of individual transcriptome RNAs.
    b. Compiles records of information for each molecule, from raw FASTQ data files, through source-gene identification and all steps in consensus sequence assembly.
    c. Conforms to true shotgun end-to-end shotgun sequencing of individual molecules present in heterogeneous samples.
    d. Assembles data files that provide links for coherent, in-depth sample validation.
    e. Performs all data processing without reliance on statistical algorithms or prior ground-truth information for target samples.

These initial findings merit full technical validation of platform performance that will be required for RUO commercialization. A future report should provide a pre-specified sample panel, complete run list, replicate statistics, blinded truth set, formal comparator, quantified sampling statistics, quantified chimera rate, isoform-level precision and recall, consensus-accuracy distribution, reproducibility across operators, instruments, reagent lots, and pipeline versions.

# Planned next-phase validation program

1. Define the intended-use claims and map each claim to a measurable acceptance criterion.
2. Use characterized reference materials and synthetic spike-ins with known isoforms, concentrations, and sequence differences.
3. Run independent library replicates across operators, days, reagent lots, instruments, and pipeline versions.
4. Quantify marker detection, positional accuracy, effective barcode diversity, SMID collisions, malformed tags, cross-gene chimeras, and within-gene molecule mixing.
5. Measure gene assignment, transcript reconstruction, splice-junction phasing, isoform precision/recall, base accuracy, and consensus-confidence calibration against truth sets.
6. Benchmark runtime, failure recovery, database completeness, and cost across run sizes and compute configurations.
7. Pre-specify analysis methods and report all molecules or a defined sampling frame rather than selecting only representative visual examples.

# Conclusion

PSeq v1 initial validation confirms a plausible end-to-end technical chain.

1. Libraries are prepared with the intended informational architecture.
2. The data pipeline exploits the molecular information of the libraries to assemble individual end-to-end transcripts.
3. The pipeline bundles the evidence chain for each molecule, providing an unprecedented depth of information.

Next steps are to expand validation tests for large data sets, replicate reproducibility, scale independence, and benchmarking performance with alternative instruments, platform operators, etc.

# Notes

1. Human universal reference RNA source noted in the internal report: Thermo Fisher Scientific catalog QS0639, https://www.thermofisher.com/order/catalog/product/QS0639.
2. IGV alignment viewer: https://igv.org/.
3. Sashimi-plot method reference supplied in the source report: https://arxiv.org/pdf/1306.3466.
4. The source report states that, to the extent tested, single read pairs containing cDNA from two genes had not been encountered. This qualitative observation should not be treated as a quantified chimera-rate result.
5. The internal report indicates that sample BAM files can accompany the document for independent viewing; they were not embedded in the reviewed DOCX and are outside the scope of this draft.

## 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.** |