

Phase 1: In Silico Validation

Prepared by Shashank Padala · N1 Therapeutics

1. Executive summary

Retriever v1 beta is N1 Therapeutics’ bioinformatics pipeline for identifying and prioritizing candidate cancer-vaccine targets from tumour and matched normal sequencing data. Its output is a ranked set of candidate targets for laboratory evaluation and subsequent mRNA vaccine design.

The hackathon prototype, Project Rosie, demonstrated how mutation analysis, target ranking and a proposed vaccine-sequence design could fit together in one application. Phase 1 focused on the evidence needed to assess that workflow: traceable study data, checks on intermediate outputs, comparisons with experimental measurements, and repeatable analyses.

Phase 1 delivered Retriever v1 beta and its documented computational evaluation. We can inspect how candidates were constructed, which tumour measurements support them, and how the evaluated ranking policies performed. Some checks supported the implementation; the ranking comparisons did not establish a general improvement over a simpler baseline.

Retriever v1 beta is the proposed starting pipeline for Phase 2: in vitro and ex vivo validation with linked canine specimens. The core team’s next decision is the study scope, including patient-specific immune typing, sequencing and experimental measurements. The sections below explain the product changes, data choices, results and requirements for that study.

2. From the hackathon prototype to Phase 1

Project Rosie already connected mutation inputs, binding prediction, fixed-weight ranking, explanatory reports and proposed mRNA sequence design. Phase 1 retained useful scoring logic and the original prototype history, while adding a research workspace and evaluations around specific failure points. Existing demonstration runs were not treated as evidence for every new endpoint.

Retriever v1 beta names the pipeline release; Phase 1 names the in silico evaluation stage. Phase 2 will evaluate candidate biology using a recorded pipeline version, with any subsequent method changes tracked as new versions.

Table 1. Prototype capabilities, evaluation gaps and Phase 1 additions

Area	Hackathon baseline	Gap to address	Phase 1 addition
------	--------------------	----------------	------------------

Area	Hackathon baseline	Gap to address	Phase 1 addition
Sequencing inputs	The pipeline started from a VCF: a list of mutations already called from sequencing.	The selected tumour/normal sequencing inputs needed a traceable path to mutation calls.	Added and executed a raw sequencing-to-variant workflow on one matched canine case. FASTQ files contain raw reads; VCF files record called variants.
Mutation-to-peptide construction	Annotation and peptide-generation tools connected mutations to candidate targets.	We needed to check sample/reference assumptions, altered sequences and unsupported cases.	Explicit sample and reference handling, independent software checks of emitted sequences, and recorded exclusions.
Ranking and tumour evidence	Predicted binding, DNA variant frequency and a configured score contributed to prioritization.	A plausible score did not establish improved immune-target ranking or support for mutant RNA.	Linked gene RNA and selected-case mutant-read evidence; compared fixed policies with simpler and established approaches.
Canine model support	The prototype generated DLA binding predictions.	Patient typing, model allele coverage, displayed peptides and immune recognition needed separate evaluation.	Explicit supported-allele handling, measured-ligand component checks and no patient nomination when DLA typing is absent.
Research interface	Case submission, progress, candidate explanation and sensitivity views demonstrated the workflow.	Core-team review needed source datasets, exclusions, run provenance and comparisons.	Local dataset/case/run browsing, candidate evidence, comparative runs and exports.
Evaluation and release	Prototype demonstrations showed the components could run together.	Target-selection performance and repeated outputs needed endpoint-specific references and records.	Curated public-data comparisons, recorded settings, repeatability checks and a public evidence supplement.
Vaccine design	The prototype produced proposed mRNA construct and synthesis-specification outputs.	Sequence design alone did not qualify candidate biology or a manufactured product.	Phase 1 concentrated on candidate selection and evaluation. Product/formulation qualification remains later work.

This compares the inspected prototype with the Retriever v1 beta capabilities evaluated during Phase 1. Existing prototype capabilities and new additions are distinguished; prototype demonstrations did not establish biological validity.

- [Background: from hackathon prototype to clinical validation](#)

3. Evaluation approach and dataset selection

We selected data by the measurement needed for each component. Mutation calls and RNA can evaluate sequence construction and molecular support. Experimentally displayed peptides can evaluate a presentation component. Human assay-labelled peptides can

evaluate ranking under human immune settings. None of these resources supplies all endpoints for the same dogs.

DLA is the dog's peptide-presenting immune-molecule system; human studies use HLA. A peptide is a short protein fragment. Predicted binding, measured display on cells and an immune response are distinct measurements, so each requires its own reference.

Table 2. Resources considered, selected purpose and selection rationale

Resource	What it provides	Why we selected it	Limit of its role
Canine mammary tumour study [1]	Tumour/normal sequencing, published mutation calls, tumour RNA and sample metadata.	Permitted canine input-to-sequence checks and same-case molecular evidence, subject to verified data overlap.	Published calls are an imperfect reference. Patient DLA typing and linked immune-response labels were unavailable.
Canine DLA presentation study [2]	Experimentally observed peptides displayed by cells, with reported DLA context.	Provided measured canine examples for evaluating the presentation predictor.	Not an answer key for mutated candidates from the mammary dogs; training exposure and artificial backgrounds constrain interpretation.
TESLA human study [3]	Peptide/HLA records with experimental recognition labels, predicted binding and expression; some measured binding values.	Allowed separate checks of ranking against recognition labels and association with binding measurements.	Human evidence; selected assay panels; public peptide tables lack the DNA variant frequency required by the original Rosie scorer.
IMPROVE human resource [4]	Recognition labels and features including expression and DNA variant frequency.	Supplied the inputs needed to evaluate the original Rosie scoring logic alongside simpler policies.	Selected after earlier results were inspected; an exploratory comparison, with study/training overlap retained.
ICDC discovery route	A route to identify canine cancer studies and available data.	Used for source discovery and feasibility assessment; original study files supplied the evaluated resources.	No separate ICDC cohort was evaluated or added to the sample counts.

- [\[1\] Canine mammary study: dataset publication](#)
- [\[2\] Canine presentation: study and observed peptide resource](#)
- [\[3\] TESLA: human neoantigen selection study](#)
- [\[4\] IMPROVE: human recognition and feature resource](#)
- [Scientific report: study design, data acquisition and cohort selection](#)

4. Data acquisition, cleaning and the evaluated subsets

Selection proceeded from available metadata to actual eligible records. We preserved original files and checksums, matched explicit specimen/source labels, normalized supported variant and immune-allele names, checked peptide validity and duplicate keys,

retained missing values, and recorded exclusions. A similar-looking sample identifier was not sufficient to join two files.

The mammary inventory contained 197 cases. Applying the malignancy, matched-DNA, tumour-RNA and exact published-call-label requirements left ten eligible source-labelled cases. We also selected CMT-002 for raw DNA/RNA execution using source BioSample linkage. Its mapping to a legacy published mutation-call alias remains unresolved.

Table 3. Sequential mammary cohort selection

Selection stage	Retained cases	Excluded at this stage
Inventoried cases	197	0
Source malignant diagnosis	140	57
Matched DNA metadata	132	8
Tumour RNA column available	107	25
Exact paired published-VCF source labels	10	97

The 97 cases excluded at the final step lacked the required exact paired source-label join; this is not a claim that their biological material is unusable.

Table 4. Acquired resources and analysis-specific subsets

Source	Acquired and curated inputs	Evaluated subset and purpose
Canine mammary study [1]	197 inventoried cases; clinical/sequencing metadata, published VCF and gene RNA matrix; selected DNA/RNA FASTQs	10 exact-source-label cases for sequence/gene-RNA checks; one separate raw-read case, CMT-002.
Canine DLA study [2]	16,359 peptide observations across canine and engineered human-host contexts; source DLA identities retained	Three fully covered canine donors for ligand ranking; fourth donor excluded for incomplete model allele support. Human-host contexts analyzed separately.
Human TESLA [3]	Table S4: 608 rows, six patients, 37 positive labels; Table S7: 310 rows, three patients, four positive labels	S4 development; S7 separately evaluated fixed policy. Common-feature S4 comparison: 404 rows. Measured binding: 496 eligible S4 rows.
Human IMPROVE [4]	17,520 rows from 70 patients; 467 source-positive labels; original pre-imputation features	17,412 common-feature rows for exploratory binding-only, fixed expression policy and configured original Rosie comparison.

These populations measure different endpoints. Counts must not be pooled into a single validation sample size.

- [Scientific report: acquisition and curation methods](#)
- [Recorded dataset inventory and cohort-flow counts](#)

5. Scope and workstream delivery

Phase 1 evaluated Retriever v1 beta’s data curation, supported computational workflow, research interface, component comparisons and repeatability. New biological assays, qualified vaccine formulation and treatment are later programme stages.

The workstreams below describe delivery of Retriever v1 beta and its evaluation resource. Implemented capabilities, executed analyses and independent scientific review are recorded separately.

Table 5. Workstreams, delivered artifacts and practical value

Workstream	Delivered	Value and remaining constraint
W1: Scope and design	Programme baseline, endpoint separation, component protocols and claim boundaries	Documents which Retriever v1 beta capabilities were tested; exploratory comparisons retain their original status.
W2: Retriever v1 beta workspace	Local dataset/case/run inspection, candidate evidence, ranking comparisons and exports	Lets researchers inspect inputs and decisions. Independent user review is pending.
W3: Data acquisition and curation	Original processed files, selected raw reads, eligible-case accounting and frozen dataset snapshot	Provides traceable evaluation inputs; unresolved aliases remain excluded from accuracy claims.
W4: Scientific workflow	FASTQ-to-VCF, supported peptide construction, gene RNA and mutant-fragment evidence	Connects inputs to checked sequences for supported variant/reference contexts; patient DLA typing remains absent in the evaluated mammary cases.
W5: Evaluation	Canine component and human comparator results, including negative outcomes	Measures components separately and identifies where ranking improvements were not demonstrated.
W6: Repeatability and release	Input/model/reference fingerprints, downstream repeats and separate-environment replay	Detects output drift; full raw execution and external reproduction were not repeated.
W7: Publication and handoff	Evidence supplement, review register and proposed biological-study brief	Supports scientific scrutiny and specific collaboration discussions; no external endorsement recorded.

6. Evaluation outcomes and their implications

With the resources introduced above, we evaluated each component separately. Sequence agreement means that emitted peptide fragments matched an independent software reconstruction within the supported scope. It does not measure how many true mutations were missed. RNA support means that the declared tumour-read threshold was met; it does not establish peptide display.

For human ranking, binding-only is the simpler comparator. The fixed expression policy combines predicted binding with tumour expression. Configured original Rosie preserves its scoring logic with documented inputs. pVACbind is an established peptide-input workflow used as an additional exploratory comparator. Each comparison has its own eligible population.

Table 6. Recorded results and programme decisions

Endpoint	Recorded outcome	Programme implication
Sequence construction	14,109 windows across ten cases; zero recorded reference mismatches; emitted windows agree with independent software	Supports scoped construction, with unsupported contexts explicitly excluded.
Raw-read execution	One matched DNA case: 70 PASS variants and 605 checked windows	Demonstrates execution of the selected workflow; mutation sensitivity and specificity remain unmeasured.
Matched mutant RNA	7/70 variants meet the declared support threshold; only one creates supported candidate windows (38)	RNA can reduce the supported candidate set; low coverage must remain distinguishable from a negative result.
Canine presentation	Three donor recovery rates approximately 69.0%, 72.6% and 60.0% against artificial shuffles	Useful component evidence; training exposure limits independence.
Human ranking	57 top-ten positives: binding-only 1, expression policy 3, pVACbind 3; IMPROVE: 42, 39 and 34 for binding-only, expression policy and configured Rosie	No general ranking superiority; broader evaluation is needed before adding complexity.
Measured human binding	496 rows across 11 HLA alleles; within-allele rank associations	Does not establish calibrated canine affinity or binding stability.

- [Scientific report: results, denominators, figures and limitations](#)

7. Milestone completion and outstanding review

The internal programme grouped delivery into milestones M0-M6, from defining the work to publishing evidence. The table below records their acceptance evidence for Retriever v1 beta. Completion refers to implementation, execution and reporting of the stated evaluations; a negative comparison is still a completed evaluation.

Independent reproduction by another person is outstanding. Repeated outputs for Retriever v1 beta were produced by the implementer, and the separate-environment replay reused recorded model predictions.

Table 7. Milestone delivery assessment

Milestone	Acceptance evidence	Observed outcome
-----------	---------------------	------------------

Milestone	Acceptance evidence	Observed outcome
M0: Scope established	Source audit, programme baseline and endpoint-specific evaluation design	Complete for Retriever v1 beta; component protocol timing remains explicit.
M1: Data feasibility	Source inventory, modality overlap, exclusions and acquisition records	Complete: four study resources; ten eligible source-labelled canine cases and one selected raw DNA/RNA case.
M2: Complete supported case	Traceable execution and independently checked sequences; unresolved allele scope labelled	Complete: 70 DNA PASS records and 605 checked peptide windows; no DLA-based patient nomination.
M3: Locks and versioned inputs	Component input/code/protocol locks and frozen processed snapshot	Inputs and settings recorded for the stated evaluations; later exploratory comparisons were not retroactively locked as confirmation.
M4: Evaluations reported	Selected tracks executed with comparators and exclusion accounting	Complete: 14,109 ten-case windows checked; presentation and human benchmarks reported, including mixed results.
M5: Retriever v1 beta repeatability	Exact downstream outputs or declared prediction tolerances; review pathway recorded	Complete for repeatability: implementer repeats and separate-environment cached-prediction replay. External reproduction remains pending.
M6: Evidence published	Public methods/results, figures, provenance, review register and biological-study brief	Published October 18; reports revised October 19. No reviewer response or journal acceptance claimed.

8. Taking Retriever v1 beta into Phase 2

The proposed Phase 2 study would use Retriever v1 beta as its starting candidate-selection pipeline. Matched tumour/normal sequencing, tumour RNA and qualified DLA typing would support a traceable candidate list for peptide-DLA binding/stability assays and functional immune assays where specimen quality and laboratory capability allow.

OVC Veterinary Biobank is a prospective specimen-access route, University of Guelph AAC a prospective sequencing route, and an immunology laboratory would provide assay expertise. Participation, specimen access, study design, data rights and assay feasibility require agreement.

The pipeline version and candidate-selection rules should be recorded before testing the new cohort. New reference contexts or unsupported alleles may require additional qualification. Changes informed by the biological results should produce a separately versioned pipeline update, preserving the original predictions and assay outcomes.

- Phase 2 exit evidence: interpretable molecular and assay outcomes, including negatives, sufficient to justify the next development decision.
- Phase 3 entry: a separate product-specific formulation, quality, safety and authorized clinical-development plan; positive computational results alone do not meet this requirement.

Table 8. Requirements before a biological study can begin

Requirement	Required deliverable	Current status
Study ownership and review	Named scientific lead, protocol reviewers and agreed responsibilities	Assignments and acceptance pending; no collaborator is assigned by this report.
Specimens and rights	Inventory of matched tumour/constitutional material, consent/reuse terms and identity controls	Access and agreements pending.
Molecular measurements	Qualified DLA genotype; tumour RNA and mutant support; relevant DLA expression assessment	Patient-specific DLA typing/expression unavailable in Phase 1.
Binding and stability	Lab-qualified peptide-DLA assay scope, supported alleles, controls and success criteria	Feasibility, reagents and laboratory pending.
Functional assays	Qualified viable cells where needed; peptide-response controls; tumour recognition/killing assessed separately	Material and assay feasibility pending.
Execution readiness	Agreed cohort, endpoints, sample handling, budget and data management	Not yet established.

9. Core-team review and supporting materials

The team review should first assess Retriever v1 beta and its computational evidence, then agree the next study's requirements. Shashank Padala owns programme priorities and release decisions. Engineering and analysis were agent-assisted. Scientific review responsibilities, biological-study leadership and partner participation remain to be agreed.

- Review the prototype-to-Phase-1 comparison and confirm that the supported capabilities and remaining gaps are described accurately.
- Review the dataset choices, sample joins and exclusions in the scientific report before interpreting benchmark outcomes.
- Agree which canine molecular and assay measurements the next cohort must connect, with patient DLA typing as an explicit requirement.
- Identify a scientific lead and suitable laboratory collaborators; establish specimen feasibility, controls, data rights, budget and study endpoints before execution.
- [Scientific report: complete methods, results and limitations](#)
- [Evidence supplement and provenance](#)
- [Proposed biological-study brief](#)
- [Scientific review and corrections register](#)

References and supporting materials

- [\[1\] Canine mammary tumour sequencing resource \(2019\).](#)
- [\[2\] Canine DLA immunopeptidomics study \(2026\).](#)

- [\[3\] TESLA consortium study \(2020\)](#)
- [\[4\] IMPROVE study and feature resource \(2024\)](#)
- [\[5\] MHCflurry 2.0.6 software and model distribution](#)
- [\[6\] Versioned evidence supplement: methods, results and provenance](#)
- [\[7\] Supplement manifest and file hashes](#)
- [\[8\] Download the original evidence archive](#)
- [\[9\] Scientific review and corrections register](#)
- [Companion computational research report](#)
- [Proposed biological-study brief](#)