Cover Page

Official Title: Combined Low-Coverage Whole-Genome and Methylome Testing of Nipple Discharge to Differentiate Benign and Malignant Breast Lesions: A Prospective Diagnostic Study

Department: Breast Surgery

Institution/Sponsor: Hubei Cancer Hospital

Study Location: Breast Surgery Department, Hubei Cancer Hospital, Wuhan, China

NCT Number: [NCTXXXXXXXX]

Document Type: Study Protocol

Version: v1.0

Document Date: 05 September 2025

1. Background and Rationale

Abnormal nipple discharge—especially bloody discharge—is a common clinical presentation of breast disease. Etiologies range from benign lesions (e.g., intraductal papilloma) to malignancies (e.g., ductal carcinoma in situ, DCIS; or invasive ductal carcinoma, IDC). Although most pathologic nipple discharge cases are benign, surgical duct excision for histopathology remains the diagnostic gold standard to rule out cancer. This strategy leads to a considerable proportion of patients undergoing unnecessary surgery. Reported malignancy rates among pathologic nipple discharge cases are approximately 11%–16% (range 3%–29% across studies), meaning a majority of operated patients are ultimately diagnosed with benign disease. Common benign causes include intraductal papilloma (35%–48%) and duct ectasia (17%–36%). Thus, many patients experience surgical trauma, anesthesia, and postoperative recovery despite benign pathology, creating avoidable physical, psychological, and resource burdens.

Non- or minimally invasive diagnostic tools for nipple discharge remain limited. Conventional cytology of nipple discharge has low diagnostic value, with reported sensitivity for breast cancer of 11%–34.6% and a false-negative rate up to 65% [1], which precludes reliable rule-out and explains why many guidelines do not recommend routine cytology screening [1]. Imaging (ultrasound, mammography, ductography, MRI) can help localize ductal lesions but has limited overall sensitivity/specificity and cannot definitively distinguish benign from malignant disease without surgical biopsy. There is a pressing need for molecular diagnostics that can accurately classify nipple discharge as benign or malignant preoperatively. Reliable identification of benign cases would avoid unnecessary operations, while timely detection of malignancy would expedite appropriate surgery, reducing harm and cost and improving decision-making.

Liquid biopsy approaches offer new opportunities. Nipple discharge is a ductal fluid containing cells and cell-free DNA (cfDNA) proximal to intraductal lesions and may be a promising biospecimen for early breast cancer detection [2]. Studies have explored markers in nipple aspirate fluid (NAF) for screening and diagnosis [2], including whole-genome sequencing of cfDNA and targeted DNA methylation assays to aid tumor

diagnosis [1,3]. DNA methylation is a key epigenetic hallmark of cancer; promoter hypermethylation of breast cancer–related genes (e.g., Cyclin D2) is frequent in malignancy but uncommon in benign disease [1,3]. In fine-needle aspirates, methylation of selected tumor suppressors has achieved high diagnostic specificity (up to 100%) for breast cancer [1,3]. Low-coverage whole-genome sequencing (lcWGS) of cfDNA can capture tumor-associated genomic instability, such as somatic copy-number alterations (CNAs) and aberrant fragmentomic patterns. In plasma studies, lcWGS has enabled noninvasive cancer detection, especially in chromosomally unstable tumors, via abnormal CNA profiles [4]. Fragmentomics (e.g., cfDNA fragment length distributions, end motifs, nucleosome footprints) can further increase sensitivity [4]. Combining CNAs with fragment features improves detection versus either alone (by ~20% in some contexts) [4]. Therefore, integrating genomic (lcWGS) and epigenomic (methylome) features may substantially enhance detection of tumor signals in nipple discharge.

This study proposes a minimally invasive diagnostic model that integrates lcWGS- and methylation-derived features from nipple discharge to distinguish benign from malignant intraductal breast lesions preoperatively. If validated, the model could inform surgical decision-making, reduce unnecessary operations in benign disease, and ensure timely treatment of malignancy, providing meaningful clinical and societal value.

2. Objectives

Overall objective:

Develop and validate a molecular diagnostic model based on nipple discharge to discriminate benign versus malignant breast lesions and guide surgical decisions.

Primary objective:

Construct a binary classifier integrating lcWGS and DNA methylome features to distinguish benign from malignant disease based on nipple discharge, and evaluate

diagnostic performance (AUC, sensitivity, specificity) via model training and internal validation.

Secondary objectives:

(1) Feature discovery: Identify malignancy-associated molecular features in nipple discharge, including genomic (CNA profiles, cfDNA fragment length distributions, other fragmentomics) and methylation markers; determine features enriched or elevated in malignancy.

(2) Model optimization and performance: Explore different feature sets and algorithms (e.g., logistic regression, machine learning) to improve robustness; determine optimal thresholds via ROC analysis; compute sensitivity, specificity, PPV, NPV; compare single-marker versus combined models.

(3) Independent validation and clinical utility: Test the trained model in an independent validation cohort to assess generalizability; estimate potential clinical benefit (e.g., proportion of benign surgeries avoided at a high-sensitivity operating point).

3. Study Design and Overview

Design: Single-center, prospective diagnostic accuracy study.

Setting: Hubei Cancer Hospital, Breast Surgery Department.

Population: Patients with suspected pathologic nipple discharge scheduled for diagnostic surgery; postoperative pathology serves as the reference standard.

Sample size: Approximately 30 consecutive eligible participants (enriched by routine clinical flow), including both benign (e.g., intraductal papilloma, duct ectasia) and malignant cases (e.g., DCIS, IDC).

Enrollment period: September 2025 to May 2026, or until target accrual is reached.

Study flow:

(1) Screening and consent.

(2) Preoperative nipple discharge collection (and NAF if needed) and one peripheral blood draw (~5 mL, EDTA).

(3) Processing, storage, and molecular assays (lcWGS and methylation profiling).

(4) Collection of clinical data and postoperative pathology.

(5) Feature extraction, model training, and validation.

4. Study Population

Inclusion criteria

(1) Age ≥18 years; primarily female (male cases may be considered but are not the main recruitment population).

(2) Unilateral, spontaneous, single-duct abnormal nipple discharge, predominantly bloody or serosanguineous (pathologic discharge), recurrent or persistent with clinical concern.

(3) Planned diagnostic surgery (e.g., microdochectomy/duct excision biopsy or tumor excision) after specialist evaluation; cases post-ductoscopy still proceeding to surgery are eligible.

(4)Willing and able to provide written informed consent, nipple discharge sample, and blood sample; agree to clinical data and postoperative pathology access.

Exclusion criteria

(1) Physiologic or nonspecific discharge (e.g., bilateral, multi-duct discharge on expression; small amounts of milky/clear/green discharge suggestive of physiologic or benign etiologies such as hyperprolactinemia or duct ectasia).

(2) Discharge due to overt infection/inflammation (e.g., abscess drainage) likely to confound DNA analysis.

(3) Prior diagnosis and treatment of breast cancer (surgery/radiation/chemotherapy) that could confound current ductal secretion and genomic characteristics.

(4) Recent invasive ductal procedures (e.g., ductoscopy with instrumentation, duct cannulation/irrigation) that may introduce artifacts before sampling.

(5) Pregnancy or lactation, due to physiologic breast changes and safety considerations.

(6) Severe hematologic disorders or coagulopathy precluding safe sampling.

(7) Refusal or withdrawal of consent; subjects later found ineligible will be discontinued.

5. Study Procedures and Sample Handling

Nipple discharge collection

(1) Timing: Preoperative, after consent.

(2) Visible discharge: Gently identify the secreting duct orifice with a sterile slide, then collect the fluid directly using a sterile swab or microtube.

(3) If no visible discharge at the visit but history supports pathologic discharge: Use nipple aspirate fluid (NAF) collection [2] with a sterile nipple cup and 10 mL syringe to apply gentle vacuum for several minutes; light breast massage may be used to facilitate flow. Literature-based adjuncts such as brief warm compress or oxytocin nasal spray may be considered per institutional policy to increase yield [2].

(4) Target volume: ≥100 μL per participant when feasible; any amount obtained will be processed using cfDNA workflows if volume is limited.

(5) Documentation: Record date/time, side, macroscopic appearance

(bloody/serosanguineous/other), and estimated volume.

(6) Storage: Process promptly. Short-term 4°C if immediate processing; otherwise snap-cool and store at −80°C to minimize DNA degradation.

Blood collection

Peripheral blood (~5 mL, EDTA) for germline leukocyte DNA and/or plasma cfDNA as reference.

Sample processing

(1) If sufficient cellular material is present: Low-speed centrifugation to separate cell pellet and supernatant; extract genomic DNA from pellet; concentrate cfDNA from supernatant (e.g., filtration or cfDNA kits).

(2) For low-volume or acellular specimens: Process entire sample with a cfDNA extraction kit.

(3) Label all materials with coded study IDs only. Log preanalytical variables.

6. Outcome Measures

Primary endpoint

Diagnostic accuracy of the combined lcWGS + methylome model to discriminate benign vs malignant disease in nipple discharge, quantified by area under the ROC curve (AUC) with 95% confidence intervals in training and validation analyses. Sensitivity and specificity will be emphasized; operating points will prioritize high sensitivity.

Secondary endpoints

(1) Predictive values: Positive predictive value (PPV) and negative predictive value (NPV).

(2) Calibration: Calibration plots and/or Hosmer–Lemeshow test to assess agreement

between predicted probabilities and observed outcomes.

(3) Threshold/clinical impact analyses: Operating point selection (e.g., maximal Youden index) and projected clinical effects (e.g., proportion of benign surgeries potentially avoided, expected miss rate) under predefined sensitivity constraints (e.g., ≥95% sensitivity).

(4) Subgroup performance: Stratified analyses by pathology subtype (DCIS vs IDC), tumor size, or imaging findings.

(5) Safety metrics: False-negative rate (missed malignancy) and characterization of any false-negative cases, with root-cause analysis.

7. Laboratory Methods

Sequencing and quality control

(1) Library prep and sequencing per standard operating procedures. Generate raw FASTQ files.

(2) Quality control: FastQC or equivalent for read quality, adapters, and error profiles.

Read alignment

(1) lcWGS data: Align to GRCh38/hg38 using bwa-mem2; produce BAM files; mark/remove PCR duplicates; apply quality filtering.

(2) Methylation data: For whole-genome bisulfite sequencing (WGBS), align with Bismark or equivalent; extract CpG-level methylation metrics.

Genomic variation (lcWGS) and fragmentomics

(1) CNA calling from low-depth data using algorithms suited for lcWGS (e.g., ichorCNA, Control-FREEC) with read-depth normalization and segmentation [4].

(2) Derive tumor-associated CNA features: genome instability indices (e.g., fraction of genome altered), recurrent breast cancer–related CNAs (e.g., 8q24 gains),

presence/absence indicators.

(3) Fragmentomics: Compute cfDNA fragment length distributions, short-fragment fractions, end motif bias, and additional indices (e.g., fragmentation uniformity, TSS-associated patterns) where feasible [4].

Methylation analysis

(1) Compute site/region-level methylation (beta values).

(2) Differential methylation analysis between benign and malignant groups to identify differentially methylated cytosines (DMCs) and regions (DMRs) using R/Bioconductor packages (e.g., limma, DSS); multiple-testing adjustment; example thresholds: $|\Delta\beta|$ above preset cutoff with adjusted $p<0.01$.

(3) Functional annotation/enrichment for genes overlapping significant DMRs to support biological plausibility.

(4) If sequencing depth is limited, prioritize literature-supported breast cancer methylation loci (e.g., RARβ2, RASSF1A, Cyclin D2) for targeted comparison [1,3].

(5) Example tools: CpGassoc, DMRcate (R).

8. Data Management and Statistical Analysis

Data management

De-identified datasets will be stored on secure hospital servers with access control. Linkage files will be kept separately by the PI. Paper records stored in locked cabinets.

Statistical analysis

(1) Group comparisons: For continuous features, t-test or Mann–Whitney U as appropriate; for categorical features, chi-square or Fisher's exact tests. Two-sided tests; $\alpha=0.05$.

(2) Feature screening: Univariate logistic regression to estimate odds ratios and p-values; variables with $p<0.10$ considered candidates.

(3) Feature selection and modeling: Stepwise selection (AIC) or LASSO to finalize multivariable models. Core algorithms: logistic regression; exploratory: support vector machine, random forest. The reference standard is postoperative pathology (benign vs malignant).

(4) Performance metrics: ROC curves, AUC with 95% CI; sensitivity, specificity, PPV, NPV; confusion matrices at prespecified thresholds. DeLong test to compare AUCs between combined models and single markers.

(5) Internal validation: k-fold cross-validation (e.g., 5-fold) or bootstrap resampling to assess robustness and estimate variance.

(6) Independent validation: Apply the locked model to an independent validation set (held-out or subsequently accrued cases); compute AUC, sensitivity, specificity, and 95% CIs; recalibrate thresholds if prespecified.

Sample size considerations

This is an exploratory single-center study ($n\approx30$) aimed at feasibility, feature discovery, and preliminary effect-size estimation rather than definitive hypothesis testing. Findings will inform power calculations for subsequent larger and/or multicenter studies.

9. Ethical Considerations and Subject Protection

(1) Ethics approval: The protocol and informed consent process have been reviewed and approved by the Ethics Committee of Hubei Cancer Hospital prior to initiation. Amendments will be submitted for ethics review.

(2) Informed consent: Participants will receive clear verbal and written information, and consent will be obtained before any study procedures.

(3) Risk/benefit: Nipple discharge collection is noninvasive with minor potential discomfort. Venipuncture may cause brief pain, bruising, rarely infection or vasovagal episodes. There is no direct clinical benefit; potential societal benefit lies in improving diagnostic strategies that may reduce unnecessary surgery in the future.

(4) Privacy and confidentiality: Coded identifiers will be used for all samples and data. The linkage key is stored separately with restricted access. Publications will use aggregate, de-identified results only.

(5) Costs and compensation: Participants will not be charged for study-specific procedures; no monetary compensation is provided.

(6) Voluntary participation and withdrawal: Participation is voluntary; withdrawal is permitted at any time without impact on clinical care. Upon request, stored samples/data will be destroyed or no longer used.

(7) Incidental findings: Because this is exploratory and not intended for clinical decision-making, individual molecular results will not be routinely returned. Any unexpected finding with clear, significant, and actionable clinical relevance will be considered by the study team and Ethics Committee before potential disclosure to the participant and treating physician.

10. Quality Assurance and Monitoring

The study will follow institutional SOPs for sample handling, data security, and laboratory procedures. Routine oversight will be performed by the PI and study coordinator. Given minimal risk and non-interventional design, a formal DSMB is not planned.

11. Dissemination

Results will be analyzed and disseminated via conference presentations and peer-

reviewed publications. Authorship and reporting will follow international guidelines and institutional policies.

12. References

[1] Pitarch M, Alcantara R, Comerma L, et al. An update on multimodal imaging strategies for nipple discharge: from detection to decision. Insights Imaging. 2025;16(1):70.

[2] Patuleia SIS, Suijkerbuijk KPM, van der Wall E, et al. Nipple Aspirate Fluid at a Glance. Cancers (Basel). 2021;14(1):[pagination not provided in source].

[3] Liu GY, Lu JS, Shen KW, et al. Fiberoptic ductoscopy combined with cytology testing in patients with spontaneous nipple discharge. Breast Cancer Res Treat. 2008;108(2):271–277.

[4] Janke F, Gasser M, Angeles AK, et al. Low-coverage whole genome sequencing of cell-free DNA to predict and track immunotherapy response in advanced non-small cell lung cancer. J Exp Clin Cancer Res. 2025;44(1):87.

Administrative Information

Version: v1.0

Version Date: 05 September 2025

Department: Breast Surgery, Hubei Cancer Hospital

Sponsor/Institution: Hubei Cancer Hospital