Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Independent references reveal how neutrophil annotation and filtering reshape patient-level estimates

Minhao Liu

minhliu@ttu.edu

Texas Tech University https://orcid.org/0009-0009-3268-8499

Article

Keywords:

Posted Date: October 22nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11109514/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Abstract

Reliable neutrophil quantification can connect single-cell measurements to patient-level immune changes. Here we benchmark fixed annotation and filtering procedures against flow cytometry, surface proteins and physical sorting across four public cohorts comprising 629,185 cells. In whole blood from 39 participants, SingleR with Blueprint/ENCODE recovers 99.10% of protein-adjudicated positives, versus 82.42% for a marker rule and less than 0.1% for two CellTypist models. Matched flow measurements reveal technology-dependent composition errors: additional filtering increases Parse error from 4.63 to 25.04 percentage points. Across nine acute-to-convalescent pairs, an unchanged alternative filter retains 78.52% of protein-positive but 3.43% of protein-negative cells, reducing estimated neutrophil-fraction decreases from nine patients to three. Physical and tissue controls identify where recovery gains transfer. These benchmarks connect independent identity validation to patient-level interpretation and show why positive recovery, negative-control errors and denominator retention should be evaluated together.

Introduction

Neutrophil abundance is a clinically interpretable dimension of immune variation that single-cell studies can connect to molecular state. During infection and inflammatory disease, circulating and tissue neutrophils occupy diverse developmental and activation states, making their representation important for both immune atlases and patient comparisons. Single-cell profiling has resolved this heterogeneity across maturation, infection and tissue recruitment, while joint RNA and surface-protein measurements provide an opportunity to test whether transcriptional annotations recover independently measured cell compartments. [1,2,3,4,5,6]

The numerical representation of a cell population is determined by several operations. Preparation and capture define which cells enter the measurement; quality control defines which measured cells enter the analysis; annotation defines which of the retained cells enter a named population. Each operation affects the estimated fraction, and its effect depends on the other cells that remain in the denominator. This distinction is particularly relevant to neutrophils, whose transcript abundance and experimental recovery differ from those of other leukocytes. Comparisons of neutrophil isolation and single-cell capture systems demonstrate that the experimental workflow can alter the recovered population. [7,8] Previous work has shown that modifying single-cell analysis can substantially increase neutrophil detection, and recent cross-platform experiments have directly established large differences in capture and transcriptome recovery. [9,10] Computational annotation adds a further choice because a model's available reference states, feature representation and rejection procedure influence the labels it can assign. [11,12,13]

An increase in the number of neutrophil calls is therefore most informative when linked to an independent measurement. Purified libraries provide a useful enrichment challenge, but cannot by themselves measure the specificity of a classifier in mixed blood. Protein measurements can support

cell-level positive and negative compartments; physical sorting can challenge transfer across lineages and developmental stages; matched flow cytometry can evaluate sample composition. These references address different questions and need not give identical rankings. Their combined value is that they can distinguish annotation recovery from selective cell loss and connect both to the biological quantity being estimated. [6,14,15,16]

Here we use these complementary references to evaluate fixed neutrophil annotation procedures and two previously specified filtering rules. The study combines four public reference cohorts with a separate purified-library depth challenge. We follow the measurement chain from independently adjudicated cells, through sample composition, to changes in nine paired patients. The central contribution is an empirical test of a usable analysis principle: recovery of the population of interest should be evaluated alongside negative-control errors and retention of the cells that define its denominator. This approach identifies a reference procedure with high recovery in whole blood and physical controls, and shows how a different filter can alter paired estimates even when that procedure remains unchanged.

Results

Complementary references connect cell identity to sample composition

The four reference cohorts comprised 629,185 source-retained cells and 99 computational sample or sorting partitions (Table 1). Hatje and colleagues contributed 322,323 cells from 21 preparations with matched flow measurements; the whole-blood study by Kwok and colleagues contributed 272,993 RNA–protein-matched cells from 48 samples and 39 participants; the Atlas of Blood Cells (ABC) contributed 7,643 expression-profiled cells with public sorting lists; and an atherosclerotic-plaque study contributed 26,226 cells from six samples with measured surface proteins. The cohorts covered distinct experimental and biological settings rather than interchangeable replicates. [5,10,15,16]

Table 1

Cohorts and independent measurement domains. The four external-validation cohorts contain 629,185 source-retained or source-provided cells across 99 computational partitions. The seven enriched GSE188288 libraries provide a separate depth/context benchmark; their 43,643 QC-retained cells are not pooled with those input-cell totals. Hatje donors 1–3 recur across 10x 3', Flex and Parse; HIVE uses separate donors A–C. ABC partitions and sorting-list donor aliases cannot be counted as distinct people across lineages. Its two mature-positive aliases and 12 mature-negative aliases define the controls in Fig. 5; unresolved mappings and other developmental populations remain unassigned to binary truth. Kwok includes 26 acute sepsis samples, nine linked recovery samples, seven surgical controls and six healthy controls. Three acute–recovery pairs share batch and six change batch. Only the six plaque samples with measured antibody features are used here; twelve RNA-only samples do not contribute to the protein reference. Reference negatives are measured or physically sorted controls, not all cells lacking a positive label. UMI, unique molecular identifier; ADT, antibody-derived tag; RBC, red blood cell; QC, quality control.

Dataset	Public source	Analyzed cells	Biological units	Independent reference	Main role
Enriched neutrophils	GSE188288	43,643 After QC	7 donors; 7 libraries	Physical enrichment; no per-cell truth label	Calling fractions; technical depth perturbation
Hatje	Zenodo 15373124[38]	322,323 Input cells	6 donors; 21 preparations	Matched flow; 12 RBC-depleted preparations in Fig. 3	Composition error; selective retention
Kwok	Zenodo 7924238 v1.1[39]	272,993 Input cells	39 people; 48 samples; 9 acute–recovery pairs	Measured ADT; 148,138 positives; 19,026 negatives	Identity performance; patient changes
ABC	GSE149938	7,643 Input cells	24 partitions; 2 mature-positive donor aliases	Physical sorting; 103 mature / 382 earlier positives; 2,008 negatives	Maturation and other-lineage controls
Plaque	GSE253904	26,226 Input cells	6 patients; 6 RNA–ADT samples	Measured ADT plus isotype background; 68 positives; 6,363 negatives	Tissue transfer controls

Protein references were constructed from measured antibody counts without RNA values, predicted RNA labels or patient outcomes. In whole blood, the resulting compartments contained 148,138 positive cells,

19,026 negative-control cells and 105,829 unadjudicated cells. In plaque, the corresponding counts were 68, 6,363 and 19,795. Reference coverage varied between samples (Fig. 1a); retaining the unadjudicated category made the scope of each measurement explicit. The whole-blood examples illustrate the CD15–CD16 distribution of the positive compartment assigned by the complete multi-marker rule across acute and convalescent samples (Fig. 1c–f). Plaque additionally provided CD66b and isotype measurements, whereas ABC supplied physical sorting labels independently of the RNA comparisons.

We evaluated a fixed marker rule (M4), two frozen CellTypist immune models and SingleR with Blueprint/ENCODE, saving both pruned and unpruned SingleR calls. No classifier was trained on the evaluation cohorts. The primary, frozen RNA quality-control rule required at least 300 unique molecular identifiers (UMIs), at least 200 detected genes and at most 20% mitochondrial counts. An alternative retained the previously specified requirement of 100–2,500 genes and less than 10% mitochondrial counts. Both rules were applied unchanged to all cohorts, and their effects on cell retention differed substantially (Fig. 1b). Thus, the benchmark tested complete method–reference combinations and explicit filtering choices, with coverage, rejection and exclusion reported separately.

Independent protein measurements establish a recovery gain beyond calling frequency

A separate enrichment challenge first illustrated the scale of annotation dependence. Seven purified neutrophil libraries from GSE188288 contained 43,643 cells passing the primary QC rule. Across donors, the median fraction called neutrophils was 87.27% for pruned SingleR, 25.73% for M4 and approximately 0.08% and 0.07% for the two CellTypist models (Fig. 2a). These are calling fractions in enriched libraries. Technical thinning of the same counts reduced median SingleR calling from 87.3% at full depth to 31.1% at half depth and 1.52% at quarter depth; M4 declined from 25.7% to 5.2% and 0.22% (Fig. 2b). Keeping the original QC cell set in the denominator made losses attributable to thinning visible. [9]

Table 2

External-reference performance and rejection counts under frozen QC. Values are pooled cell counts and percentages under frozen QC (at least 300 UMIs, at least 200 genes and mitochondrial fraction at most 20%); independent participant-level distributions are shown in Figs. 2 and 5. Positive denominators include all source-retained reference positives, including QC exclusions. Negative denominators contain only QC-retained reference negatives. Before QC, there were 19,026 Kwok and 6,363 plaque negatives; 17,327 and 6,267 remain. The corresponding positive counts retained are 147,620/148,138 and 66/68. All 103 mature positives, 382 earlier-stage cells and 2,008 negatives in ABC pass frozen QC. Earlier-stage recovery is separate from mature identity; a dash means not applicable. Rejections denote retained cells receiving no accepted native label after SingleR pruning, and are neither QC exclusions nor correct negative assignments. SingleR is the pruned primary method; its unpruned results are shown separately. CT Low/High and M4 use the frozen configurations described in Fig. 2. ABC negative errors include the unchanged NK-sorted controls; for SingleR, 64 of 65 negative calls arise in 230 NK-sorted cells. All primary methods have zero calls among the retained plaque negative controls. These pooled rates are descriptive agreement measures within the stated reference compartments, not tests based on independent cell replicates.

Cohort	Method	Primary positive recovery n/N (%)	Earlier-stage recovery n/N (%)	Negative-control calls n/N (%)	Rejected cells positive / earlier / negative
Kwok	CT Low	113 / 148,138 (0.08%)	—	0 / 17,327 (0.00%)	0 / — / 0
	CT High	49 / 148,138 (0.03%)	—	0 / 17,327 (0.00%)	0 / — / 0
	M4	122,100 / 148,138 (82.42%)	—	96 / 17,327 (0.55%)	0 / — / 0
	SingleR	146,808 / 148,138 (99.10%)	—	6 / 17,327 (0.03%)	777 / — / 259
	SingleR unpruned	147,582 / 148,138 (99.62%)	—	6 / 17,327 (0.03%)	0 / — / 0
ABC	CT Low	14 / 103 (13.59%)	98 / 382 (25.65%)	16 / 2,008 (0.80%)	0 / 0 / 0
	CT High	1 / 103 (0.97%)	19 / 382 (4.97%)	19 / 2,008 (0.95%)	0 / 0 / 0
	M4	50 / 103 (48.54%)	217 / 382 (56.81%)	20 / 2,008 (1.00%)	0 / 0 / 0
	SingleR	99 / 103 (96.12%)	363 / 382 (95.03%)	65 / 2,008 (3.24%)	2 / 8 / 25

Cohort	Method	Primary positive recovery n/N (%)	Earlier-stage recovery n/N (%)	Negative-control calls n/N (%)	Rejected cells positive / earlier / negative
	SingleR unpruned	101 / 103 (98.06%)	371 / 382 (97.12%)	67 / 2,008 (3.34%)	0 / 0 / 0
Plaque	CT Low	0 / 68 (0.00%)	—	0 / 6,267 (0.00%)	0 / — / 0
	CT High	0 / 68 (0.00%)	—	0 / 6,267 (0.00%)	0 / — / 0
	M4	16 / 68 (23.53%)	—	0 / 6,267 (0.00%)	0 / — / 0
	SingleR	42 / 68 (61.76%)	—	0 / 6,267 (0.00%)	0 / — / 84
	SingleR unpruned	42 / 68 (61.76%)	—	0 / 6,267 (0.00%)	0 / — / 0

The whole-blood protein reference then tested whether higher calling corresponded to recovery of independently measured positives. With primary QC, pruned SingleR recovered 146,808 of all 148,138 protein-positive cells (99.10%), including QC losses in the denominator. M4 recovered 122,100 cells (82.42%), CellTypist Low recovered 113 (0.076%) and CellTypist High recovered 49 (0.033%; Table 2). Averaging repeated samples within participants gave mean recovery of 99.08% for SingleR and 82.12% for M4. SingleR recovery exceeded each comparator in all 39 participants, with Holm-adjusted $P = 2.55 \times 10^{-11}$ for each two-sided paired sign test (Fig. 2c).

Recovery was accompanied by a low error rate in the measured negative compartment. SingleR assigned neutrophil labels to six of 17,327 retained negative controls (0.035%), compared with 96 for M4 (0.554%) and zero for both CellTypist models. Thirty-four participants had retained negative controls and contributed to participant-level comparisons. SingleR had lower participant-level negative-control call rates than M4 (adjusted $P = 7.63 \times 10^{-6}$), whereas comparisons with the two CellTypist models did not reach significance (adjusted P = 0.125 each; Fig. 2d). The recovery and false-positive panels together identify the operating behavior of these fixed combinations; a zero negative-control error rate accompanied near-zero positive recovery for the two CellTypist references.

The pruned SingleR procedure rejected 777 retained positive and 259 retained negative cells, with these outcomes preserved separately from accepted predictions (Table 2). The unpruned comparator

recovered 147,582 positives and made the same six false-positive calls. Pruned SingleR remained the prespecified primary reference procedure for composition analyses; both versions were retained to assess query-dependent pruning.

Matched flow measurements expose technology-dependent filtering losses

The matched-flow cohort tested composition in a different way. Three donors contributed 10x 3', Flex and Parse preparations, each with peripheral blood mononuclear cell and red-cell-depleted processing, while three additional donors contributed HIVE red-cell-depleted preparations. The principal comparison used the 12 red-cell-depleted preparations, with a non-platelet RNA denominator to approximate the flow leukocyte compartment (Fig. 3). The RNA data retained the supplied shared-feature representation, so this analysis measures the effect of additional filtering on the public count export.

Before additional QC, pruned SingleR's mean absolute deviation from matched flow was 5.36 percentage points for Flex, 4.63 for Parse, 7.87 for HIVE and 28.07 for 10x 3' (Fig. 3a). After primary QC, the corresponding errors were 4.58, 25.04, 8.33 and 37.59 percentage points (Fig. 3b). Parse therefore provided a direct example in which close initial composition agreement was substantially reduced under the additional QC domain. The same method that recovered nearly all whole-blood protein positives could not ensure stable sample composition across technologies and filtering domains.

Source-annotation retention separated a descriptive indicator of selection from the external flow endpoint. Retention of cells annotated by the source as neutrophils differed from retention of other cells, with particularly large losses in some preparation–technology combinations (Fig. 3c). The flow comparison supplied the independent composition measurement; the source labels described which annotated compartment was removed. Excluding a donor with a documented discrepancy between public flow summaries preserved the direction of the Parse result, with mean error changing from 6.73 to 28.37 percentage points in the remaining two donors. Together, these measurements show why annotation recovery and cell-selection effects need to be examined on the same sample denominator.

Selective loss of the denominator changes paired patient estimates

The whole-blood cohort included nine explicitly linked acute-to-convalescent patient pairs. Under primary QC, pruned SingleR estimated a decrease in neutrophil fraction in all nine patients, with a mean change of − 17.20 percentage points and a descriptive patient-bootstrap 95% confidence interval of − 20.30 to − 14.35. M4 also estimated decreases in all nine pairs, with mean change − 14.49 percentage points. The two CellTypist models gave near-zero changes, consistent with their low recovery of the protein-positive compartment (Fig. 4a; Table 3).

Table 3

Patient-pair changes across all methods and fixed input domains. Changes are recovery minus acute in RNA-estimated neutrophil fractions, in percentage points (pp). Decrease counts indicate any negative estimate, not significance. Existing descriptive 95% percentile intervals use 10,000 patient-pair bootstrap draws (seed 140918); none are computed for the three same-batch pairs.

Source-retained cells receive no additional QC. Frozen QC requires at least 300 UMIs, at least 200 genes and mitochondrial fraction at most 20%; alternative QC requires 100–2,500 genes and mitochondrial fraction below 10%. Singlet analyses retain frozen-QC scDblFinder singlets, reusing SingleR pruning without another pass. Both QC rules were fixed before external method comparison. All nine protein-adjudicated subset changes are negative, but whole-population protein bounds span zero in every pair. These method-dependent estimates therefore do not establish fully calibrated clinical changes. CT, CellTypist; CI, confidence interval. SingleR denotes pruned results unless marked unpruned.

Input domain	Method	All nine pairs Mean change (pp) [95% bootstrap CI]	Pairs with decrease	Same-batch pairs Mean change (pp)	Pairs with decrease
Source-retained	CT Low	−0.058 [− 0.142, + 0.016]	5 / 9	−0.039	2 / 3
	CT High	−0.035 [− 0.090, + 0.007]	4 / 9	+ 0.008	1 / 3
	M4	−15.888 [− 21.126, − 11.101]	9 / 9	−19.427	3 / 3
	SingleR	−18.897 [− 22.214, − 15.813]	9 / 9	−19.526	3 / 3
	SingleR unpruned	−20.072 [− 23.427, − 16.948]	9 / 9	−20.403	3 / 3
Frozen QC	CT Low	−0.053 [− 0.129, + 0.017]	5 / 9	−0.042	2 / 3
	CT High	−0.033 [− 0.082, + 0.004]	4 / 9	+ 0.004	1 / 3
	M4	−14.488 [− 19.692, − 9.663]	9 / 9	−17.689	3 / 3
	SingleR	−17.200 [− 20.297, − 14.351]	9 / 9	−17.398	3 / 3
	SingleR	−18.324	9 / 9	−18.202	3 / 3

Input domain	Method	All nine pairs Mean change (pp) [95% bootstrap CI]	Pairs with decrease	Same-batch pairs Mean change (pp)	Pairs with decrease
	unpruned	[− 21.465, − 15.417]			
Alternative QC	CT Low	−0.001 [− 0.025, + 0.031]	3 / 9	+ 0.011	2 / 3
	CT High	−0.013 [− 0.033, + 0.000]	2 / 9	−0.009	1 / 3
	M4	−0.449 [− 5.765, + 4.408]	3 / 9	−3.893	2 / 3
	SingleR	+ 0.123 [− 0.453, + 0.658]	3 / 9	−0.257	1 / 3
	SingleR unpruned	−1.159 [− 1.775, − 0.540]	7 / 9	−1.334	3 / 3
Frozen QC singlets	CT Low	−0.059 [− 0.137, + 0.009]	5 / 9	−0.053	2 / 3
	CT High	−0.032 [− 0.083, + 0.005]	4 / 9	+ 0.010	1 / 3
	M4	−14.798 [− 19.795, − 10.226]	9 / 9	−18.101	3 / 3
	SingleR	−17.175 [− 20.398, − 14.377]	9 / 9	−17.704	3 / 3
	SingleR unpruned	−18.352 [− 21.474, − 15.405]	9 / 9	−18.522	3 / 3

Independent protein measurements provided a complementary direction check. Neutrophil fractions within the protein-adjudicated compartment decreased in all nine pairs; the fraction of all cells represented by confidently protein-positive cells decreased in eight pairs. Three pairs were processed within the same batch, and all three showed decreases under these protein summaries. Because the protein reference did not adjudicate every cell, these checks concern the measured compartments; the interpretation of complete-sample changes is considered below.

Applying the unchanged alternative QC rule to the same RNA data altered the estimated trajectory. This rule retained 116,319 of 148,138 protein-positive cells (78.52%), but only 652 of 19,026 negative-control cells (3.43%; Fig. 4b). Primary QC retained 99.65% and 91.07%, respectively. With the alternative, pruned SingleR estimated decreases in only three of nine patients, and the mean paired change was + 0.12 percentage points (95% confidence interval, − 0.45 to + 0.66). Every patient's recovery-minus-acute estimate shifted upward relative to the primary-QC estimate; the adjusted paired sign-test P value was 0.0117 (Fig. 4c). This comparison tests the effect of the analytical choice on paired estimates.

An explanatory decomposition localized the change. Holding the original source-population SingleR predictions fixed, selecting cells with the alternative QC mask shifted the mean paired estimate by + 18.77 percentage points relative to the source-cell analysis. Recalculating the query-dependent pruning contributed a further + 0.25 percentage points (Fig. 4d). Of 18,374 protein-negative cells excluded by the alternative rule, 18,264 met the mitochondrial exclusion criterion and 110 failed exclusively because of the upper gene-count limit. The magnitude of the selection component explains how a fixed annotator can yield different paired estimates when the retained population changes.

A computational singlet restriction preserved the primary result: all nine paired estimates decreased, with mean change − 17.18 percentage points (95% confidence interval, − 20.40 to − 14.38; Table 3). Within retained protein-positive singlets, SingleR recovered 141,359 of 142,167 cells (99.43%); using all source protein positives as the denominator gave 95.42% end-to-end recovery. Three of 17,124 retained negative controls received neutrophil calls. This sensitivity reused the primary-QC pruning and therefore tested the additional cell restriction without another pruning change.

Physical and tissue controls define where the recovery gain transfers

Public sorting labels extended the comparison beyond the whole-blood protein panel. Among 103 physically sorted mature neutrophils from two bone-marrow donor aliases, pruned SingleR recovered 99 cells (96.12%), M4 recovered 50 (48.54%), CellTypist Low recovered 14 (13.59%) and CellTypist High recovered one (0.97%; Fig. 5a; Table 2). Earlier neutrophil developmental stages were evaluated separately: SingleR recovered 363 of 382 cells (95.03%), compared with 217 (56.81%) for M4. These positive controls supported transfer of the recovery gain across a distinct source of cell labels.

The corresponding negative controls localized a lineage-dependent discordance. SingleR called 65 of 2,008 mature B-, T-, NK- or monocyte-sorted cells as neutrophils (3.24%). Sixty-four of these calls occurred among 230 NK-sorted cells, while one occurred among 1,778 cells in the other three lineages (Fig. 5b). Original sorting labels were retained throughout. Many discordant cells expressed both neutrophil and NK RNA panels, and computational doublet restriction did not generally resolve the discrepancy. The pattern identifies a specific transfer condition to examine when using this reference in physically selected populations.

In plaque, the protein and isotype measurements defined 68 positives across six patients. SingleR recovered 42 (61.76%), M4 recovered 16 (23.53%) and neither CellTypist model recovered a positive (Fig. 5c; Table 2). All four procedures made zero neutrophil calls among 6,267 QC-retained negative controls (Fig. 5d). Thus, the protein-positive recovery advantage extended to tissue, while its absolute magnitude differed from whole blood. Small positive compartments, ranging from four to 29 cells per patient before RNA QC, are displayed individually to make this sampling context visible.

A post-comparison label-scope check additionally credited native myelocyte and promyelocyte predictions as broader granulocytic-lineage calls. This left ABC recovery unchanged and increased whole-blood CellTypist Low and High recovery only to 0.292% and 0.074%, respectively; in plaque each model detected three additional lineage-positive cells. The primary acceptance labels and reported comparisons remained unchanged. These controls establish the behavior of the tested model–reference combinations across contexts and distinguish improved positive recovery from a claim that one algorithm will be uniformly preferable.

Discussion

Independent references connect the cell-level performance of an annotation workflow to the patient-level quantity that motivates its use. Across the present cohorts, protein measurements and physical sorting supported substantial neutrophil recovery by SingleR with Blueprint/ENCODE. Matched flow and paired patient analyses then showed that recovering a measured positive compartment does not by itself stabilize composition estimates. The decisive additional quantity was retention of the other cells in the denominator. Measuring recovery, negative-control errors and denominator retention together made this connection observable and provided a practical basis for evaluating an analysis before interpreting patient differences.

The paired whole-blood analysis gives a concrete example. Under primary QC, the reference procedure recovered 99.10% of the operational protein-positive compartment and estimated decreases in all nine paired patients. The alternative rule retained most positive cells but removed nearly the entire measured negative compartment. The resulting fractions approached a different compositional regime, and only three patients retained a decrease. Holding source-population SingleR predictions fixed attributed most of the shift to cell selection, with a much smaller contribution from query-dependent pruning. These results extend the established importance of QC beyond the number of retained cells: a rule can materially change the comparison of patients by changing which populations contribute to its denominator. [17,18,19,20]

This interpretation extends studies that established the feasibility of recovering neutrophil transcriptomes through experimental and analytical choices. Wigerblad and colleagues showed that changes to data analysis increased neutrophil detection; our analysis asks when a detection gain agrees with independent labels and how the retained denominator affects a patient comparison. Hatje and colleagues directly compared preparation and sequencing technologies; our reanalysis uses their

matched flow measurements to ask how additional computational selection propagates into composition error. The pronounced Parse change shows that an analysis can move away from an external composition reference after a conventional QC step. The persistent 10x 3' error also situates the contribution of annotation within the measurement chain: computational recovery acts on cells that are present in the supplied matrix. The experiment, exported feature space and annotation procedure therefore need to be considered jointly when transferring a workflow. [8,9,10,21]

The comparison also clarifies what a model benchmark should resolve. CellTypist and SingleR were evaluated as fixed models or reference combinations with their native decision procedures. Their marked separation in these data should motivate scrutiny of represented reference states, feature overlap and rejection, rather than an algorithm-wide ranking. The purified-library thinning result shows a further dependence on count depth, while ABC and plaque show that both positive recovery and negative-control behavior can change with context. Previous studies have demonstrated the value of reference mapping and systematic annotation benchmarks; the present work adds independent measured compartments and follows the consequences into sample and patient estimates. [11,12,13,22,23]

A practical workflow follows from these observations. Before comparing disease states, analysts can define the intended denominator and identify available independent reference measurements. Cell-level checks should then report how many measured positives survive selection and are recovered, how often retained negatives receive target labels, and how many observations are rejected or remain unadjudicated. A sample-level composition measurement, when available, tests the combined effect of these decisions. Finally, repeating the patient comparison under a small set of explicitly retained analytical choices reveals whether the biological direction depends on a selective loss. No single recovery statistic substitutes for these linked checks. This framework provides a starting point for other rare or technically sensitive populations whose abundance is inferred from a selected single-cell sample. Independent histological validation of eosinophil recovery provides a related example of testing low-RNA cell representation against a distinct measurement. [24] Population-composition methods can then address changes in a defined, evaluated cell set. [25,26,27]

Several features define the scope in which these conclusions can be used. The protein gates provide operational high-confidence compartments, with different panel coverage in blood and plaque. Work on ADT background further supports accounting for assay-specific noise when interpreting antibody counts. [28] The whole-blood panel lacks CD66b and isotype controls, and neither panel independently determines the identity of every measured cell. Accordingly, protein agreement supports the direction within adjudicated compartments; allowing all unadjudicated cells to take either identity leaves all nine whole-sample paired bounds spanning zero. These are reference-coverage bounds, not sampling confidence intervals. Additional flow measurements on the same aliquots would strengthen the link from computational fractions to complete-sample changes.

Patient and donor structure also guides interpretation. Six of the nine clinical pairs changed processing batch, while the same-batch analysis contains three pairs. HIVE uses a distinct set of donors, and the

matched-flow comparisons contain three donors per technology. The physical-control aliases cannot be assumed to identify the same people across lineages, and the plaque-positive compartment is small. We therefore retain individual observations, within-participant averaging and patient-level resampling, and use the sign tests as explanatory assessments of analytical differences. The comparisons show that the estimated trajectory depends on analysis; they do not attribute the underlying clinical trajectory exclusively to neutrophil biology. [29,30]

The processed public exports support reproducible evaluation of the retained cell sets, but do not reconstruct all original droplets or resolve viability and ambient-RNA contributions. The common Hatje feature space affects available QC measurements, and the NK-sorted discordances warrant measurements that can distinguish mixed identity, sorting impurity and transcriptional overlap. These are well-defined extensions of the current benchmark. They leave intact its directly tested finding: independent reference measurements can identify both a useful annotation recovery gain and the selective cell loss that changes its patient-level interpretation. For neutrophil composition studies, retaining this link from measured cells to the analysis denominator is a necessary part of making patient comparisons reproducible.

Methods

Study design and analysis status

This retrospective computational study used public human single-cell datasets and measurements released by the source studies. The benchmark protocol and reference-construction rules were locked before generating the new method-performance comparisons. Source inventories, published fractions and metadata availability had already been examined, so the lock was an analysis specification rather than prospective registration. Earlier control-derived marker rules, model files and reference data were preserved. Functional measurements, patient outcomes and RNA-method performance were not used to select the protein gates or revise annotation rules. All evaluated domains, rejection states and control discordances were retained. Explanatory pruning, label-scope and statistical-annotation analyses were labeled separately from the fixed primary comparisons.

Cohorts and sample linkage

Hatje data were obtained from Zenodo record 15373124. The integer-count export contained 753,435 rows and a shared 14,966-feature space. Exact barcode linkage identified 322,323 author-retained cells for the benchmark. The scaled 577-feature representation was not used as raw counts. Donors 1–3 contributed 18 preparations across 10x 3', Flex and Parse with two preparation types; HIVE donors A–C contributed three red-cell-depleted preparations. Their mapping was recovered from the public notebook. Matched-flow medians were extracted from cached notebook tables and checked against public error summaries. A discrepancy between the notebook and supplement for donor 2 was retained in the audit and addressed by a donor-exclusion sensitivity. The main RNA denominator excluded source-labeled

platelets to approximate the flow leukocyte compartment; results using all source-retained cells were also saved.

Kwok data used the complete v1.1 Zenodo record 7924238. RNA and antibody-derived tag (ADT) measurements were exactly matched by barcode for 272,993 cells. The 48 samples comprised 26 acute sepsis samples, nine linked convalescent samples, seven surgical controls and six healthy controls, representing 39 participants. Paired analyses used the nine explicit source links, with same-batch pairs S066, S099 and S104 identified in advance of paired summaries. The ABC expression matrix from GSE149938 contained 7,643 cells, of which 5,714 matched public sorting lists exactly. Mature neutrophils and earlier neutrophil stages were separated; mature B-, T-, NK- and monocyte-sorted cells formed the negative control. Other stages and unresolved mappings retained their original status. GSE253904 contributed the six plaque RNA–ADT samples; twelve RNA-only samples were not used as protein-reference observations. The separate GSE188288 challenge used seven purified neutrophil libraries and retained their own donor and QC accounting.

RNA processing, QC and annotation

Sparse integer count matrices were used throughout. Exact duplicate gene symbols were summed, and protein features were excluded from RNA normalization and QC. The primary domain required at least 300 UMIs, at least 200 detected genes and mitochondrial count fraction at most 20%. The alternative required 100–2,500 detected genes and mitochondrial fraction below 10%, without changing its pre-existing thresholds. MT-prefixed symbols defined mitochondrial features within each supplied gene universe. Kwok RNA totals, detected genes and mitochondrial numerators matched source metadata; differences between the stored mitochondrial percentage and the exported denominator changed no cell's eligibility under either rule.

M4 required detection of at least four members of a fixed ten-gene neutrophil panel, at least one member of a five-gene anchor panel, fewer than two monocyte-panel genes and fewer than two genes in each competing-lineage panel. Detection meant a nonzero count. The neutrophil panel comprised FCGR3B, CSF3R, CXCR2, CEACAM8, FPR1, LTF, MPO, ELANE, MMP8 and MMP9; the anchor subset was FCGR3B, CSF3R, CXCR2, CEACAM8 and LTF. The monocyte panel was CD14, FCGR3A and MS4A7. Competing panels were CD3D, CD3E, TRAC, IL7R and CD8A for T cells; NKG7, GNLY, PRF1 and KLRD1 for NK cells; MS4A1, CD79A and CD79B for B cells; EPCAM, KRT8, KRT18 and KRT19 for epithelial cells; and HBA1, HBA2, HBB and ALAS2 for erythroid cells. This rule was selected previously using training controls and was not refitted on any benchmark cohort. CellTypist used the preserved Immune_All_Low and Immune_All_High v2 files, normalization to 10,000 counts per cell, log1p transformation and native shared-feature centering/scaling and an upper clip of 10, without majority voting. The accepted primary labels were Neutrophils, Granulocytes and Neutrophil-myeloid progenitor. Native implementation parity was checked on 32 cells from each of four cohorts for both models. A separate explanatory comparison added Myelocytes and Promyelocytes to a broader lineage acceptance set. [12,31]

SingleR used the preserved Blueprint/ENCODE reference with 259 profiles, including 23 neutrophil profiles, classic reference-derived differential markers, native fine tuning and default pruning. Initial scores were generated on the source-provided cell set. Native pruning was recalculated within each sample and RNA domain because rejection depends on the query distribution. Both raw and pruned labels were retained. Direct native calls on four complete QC-restricted samples reproduced saved labels, scores and pruning. The computational singlet analysis reused primary-QC predictions without another pruning pass. [11,32,33]

Protein-only adjudication

Protein reference construction used measured ADTs without RNA values, RNA predictions, author joint RNA–protein labels or clinical outcomes. In Kwok, one- and two-component Gaussian mixture models were fitted to each marker's log1p counts within each sample. A separable two-component distribution required a Bayesian information criterion improvement of at least 10, log-mean separation of at least one, standardized separation of at least two, component weight at least 0.001 and at least ten expected cells per component. High signal required upper-component posterior probability at least 0.95 and at least ten counts; low signal required lower-component posterior probability at least 0.95. When a marker was nonseparable, counts of zero or one could establish low signal, but no high call was made. Both mixture fits had to converge. Fitting used seed 140918, five initializations, at most 300 iterations and covariance regularization of 10^{-4}.

Kwok positives required high CD15 and at least one of CD16, CD10 or CD11b, with low CD14, HLA-DR, CD19, CD56 and CD8. Negatives required low CD15 together with either high CD4, CD8, CD19 or CD56, or high CD14 and HLA-DR jointly. CD64 was retained as a diagnostic and did not exclude positives. Other cells remained unadjudicated. In plaque, the maximum count across nine isotype controls defined a per-cell background envelope b. High signal required at least ten counts and at least 3(b + 1); low signal required counts no greater than b + 1. Positives required high CD66b and either CD15 or CD16, together with low CD193, CD3, CD19, CD56, CD14 and HLA-DR. Negatives required low CD66b and CD15 with high CD3, CD19, CD56 or CD193, or high CD14 together with high HLA-DR. Mixed and unresolved signals were retained. These procedures define measured reference compartments and their coverage.

Depth and cell-artifact diagnostics

The purified-library challenge used binomial thinning of each nonzero count at fractions 1, 0.5 and 0.25, recalculating QC while keeping the original full-depth QC-qualified cell set as the denominator. Each reduced-depth condition used one thinning per library with seed 120918 plus the integer encoded by the first eight hexadecimal digits of SHA256(GSM identifier + textual depth fraction). The thinned libraries are technical perturbations of the same donors. scDblFinder was applied to primary-QC cells by sample or explicit partition, using deterministic barcode-hash chunks containing at most 15,000 cells. Barcodes were ordered by a base-131 polynomial hash modulo 2,147,483,647, with barcode tie-breaking; each partition seed was 140918 plus the sum of sample-identifier Unicode code points and the chunk index. Execution used clusters = FALSE and serial BiocParallel, with other scDblFinder settings left at their

defaults. Partitions with fewer than 100 cells were recorded as unevaluated. The diagnostic produced 519,810 singlet calls, 46,484 doublet calls and 138 unevaluated cells, with zero runtime failures; 62,753 cells had failed primary QC. For plate-sorted and demultiplexed data, these are computational mixed-profile diagnostics. Doublet removal did not redefine the primary population. The public processed exports did not supply an independent empty-droplet background for an empty-droplet or ambient-RNA reconstruction. [34,35,36]

Estimands and decomposition

End-to-end positive recovery was the number of accepted neutrophil predictions among reference positives divided by all source reference positives, including those excluded by RNA QC. Conditional recovery used only retained positives. The false-positive rate was the number of accepted neutrophil predictions among retained reference negatives divided by the retained negative count. A rejected label was not converted into a correct negative assignment. Sample-level flow errors were signed or absolute differences, in percentage points, between RNA-estimated neutrophil fractions and the fixed matched-flow reference. The source RNA annotations were descriptive comparators, not truth labels.

Clinical paired changes were convalescent minus acute. For protein measurements, the fraction positive among adjudicated cells was recorded separately from the all-cell lower bound positive/N. Whole-sample fractions were bounded by positive/N and (positive + unadjudicated)/N; subtracting the appropriate upper and lower endpoints gave paired difference bounds. These intervals describe unresolved reference coverage. To separate selection from pruning, source-population SingleR labels were held fixed while applying each existing QC mask. The paired estimate under the mask was compared with the source-cell estimate; the further change after domain-specific re-pruning was recorded separately. No thresholds were optimized from this decomposition.

Statistics and Reproducibility

Biological units were participants or donors where explicitly identified. Repeated samples were first averaged within participant for protein-performance comparisons, and complete patient pairs were retained for longitudinal estimates. Sorting aliases were reported as aliases and were not assumed to be independent donors across lineages. Descriptive percentile-bootstrap 95% confidence intervals used 10,000 draws with seed 140918, resampling participants or complete patient pairs. Intervals were not computed for groups with fewer than five biological units. The intervals condition on the operational references and do not include all possible antibody or sorting-label errors.

Seven explanatory paired sign tests were specified during preparation of the figures after descriptive results were available: three comparisons of positive recovery against pruned SingleR in 39 participants, three comparisons of false-positive rate in 34 participants with an eligible negative denominator, and one comparison of the nine paired clinical estimates under alternative versus primary QC. Tests were two-sided and exact, omitted differences with absolute value at most 10^{-12} as ties and used Holm correction across the seven-test family. The tests assess direction of within-unit differences; plotted means

summarize magnitude. Adjusted P < 0.05, 0.01, 0.001 and 0.0001 correspond to one, two, three and four stars, respectively; non-significant comparisons are shown. No individual cells were treated as independent biological replicates for these tests. [37]

All original method outputs, exclusions, rejected labels and source-to-panel data were preserved, and the final presentation was checked against frozen numeric exports. The workflow was a complete reanalysis of the selected public exports rather than a powered prospective experiment. No randomized allocation or intervention was performed. Computational procedures were fixed before the primary comparisons, and measured protein references were constructed independently of RNA predictions.

Software and implementation verification

Archived execution used R 4.6.1, SingleR 2.14.2, scDblFinder 1.26.7, SingleCellExperiment 1.34.0, BiocParallel 1.46.0 and Matrix 1.7-5. Python 3.13.5 annotation dependencies included CellTypist 1.7.1, NumPy 2.5.3, pandas 3.0.5, SciPy 1.18.1, scikit-learn 1.9.1, anndata 0.13.3.post0 and Scanpy 1.12.4. General processing ran in a separately recorded environment with NumPy 2.4.3, pandas 2.3.3, SciPy 1.16.3 and scikit-learn 1.7.2. Model-file hashes, reference hashes and both environment records are included in the reproducibility archive; current package versions were not substituted for the archived execution versions. Native-API parity checks, exact barcode linkage, input accounting and panel-level numerical checks were completed before manuscript assembly.

Declarations

Data availability

All datasets analyzed in this study are publicly available. The Hatje technology-comparison data are available from Zenodo (https://doi.org/10.5281/zenodo.15373124), and the Kwok whole-blood RNA and antibody-derived tag data from Zenodo version 1.1 (https://doi.org/10.5281/zenodo.7924238). The Atlas of Blood Cells physical-sorting data, atherosclerotic-plaque RNA–protein data and enriched-neutrophil libraries are available from the Gene Expression Omnibus under accession numbers GSE149938, GSE253904 and GSE188288, respectively. [9,15,16,38,39] The controls used in the earlier development of the frozen M4 rule are available under GSE293018 and GSE127465.

Code availability

Code for data preparation, reference construction, annotation, diagnostics and figure generation is publicly available. Repository and release details are withheld for double-anonymous peer review. The repository includes frozen analysis settings, environment records, source hashes and instructions for reproducing the figures and statistical comparisons.

Acknowledgements

The author thanks the investigators and participants of the source studies for making their data available. No funding was received for this work.

Author contributions

The author conceived and designed the study, performed the computational analyses, interpreted the results, prepared the figures and tables, and wrote and approved the manuscript.

Competing interests

The author declares no competing interests.

References

1. Palomino-Segura, M., Sicilia, J., Ballesteros, I., Hidalgo, A.: Strategies of neutrophil diversification. Nat. Immunol. **24**, 575–584 (2023). https://doi.org/10.1038/s41590-023-01452-x
2. Xie, X., et al.: Single-cell transcriptome profiling reveals neutrophil heterogeneity in homeostasis and infection. Nat. Immunol. **21**, 1119–1133 (2020). https://doi.org/10.1038/s41590-020-0736-z
3. Evrard, M., et al.: Developmental Analysis of Bone Marrow Neutrophils Reveals Populations Specialized in Expansion, Trafficking, and Effector Functions. Immunity. **48**, 364–379 (2018). https://doi.org/10.1016/j.immuni.2018.02.002 .e8
4. Grieshaber-Bouyer, R., et al.: The neutrotime transcriptional signature defines a single continuum of neutrophils across biological compartments. Nat. Commun. **12**, 2856 (2021). https://doi.org/10.1038/s41467-021-22973-9
5. Kwok, A.J., et al.: Neutrophils and emergency granulopoiesis drive immune suppression and an extreme response endotype during sepsis. Nat. Immunol. **24**, 767–779 (2023). https://doi.org/10.1038/s41590-023-01490-5
6. Stoeckius, M., et al.: Simultaneous epitope and transcriptome measurement in single cells. Nat. Methods. **14**, 865–868 (2017). https://doi.org/10.1038/nmeth.4380
7. Connelly, A.N., et al.: Optimization of methods for the accurate characterization of whole blood neutrophils. Sci. Rep. **12**, 3667 (2022). https://doi.org/10.1038/s41598-022-07455-2
8. Salcher, S., et al.: Comparative analysis of 10X Chromium vs. BD Rhapsody whole transcriptome single-cell sequencing technologies in complex human tissues. Heliyon 10, e28358 (2024). https://doi.org/10.1016/j.heliyon.2024.e28358
9. Wigerblad, G., et al.: Single-Cell Analysis Reveals the Range of Transcriptional States of Circulating Human Neutrophils. J. Immunol. **209**, 772–782 (2022). https://doi.org/10.4049/jimmunol.2200154
10. Hatje, K., et al.: Comparison of single-cell RNA-seq methods to enable transcriptome profiling of neutrophils in clinical samples. Cell. Rep. Methods. **5**, 101173 (2025). https://doi.org/10.1016/j.crmeth.2025.101173

11. Aran, D., et al.: Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat. Immunol. **20**, 163–172 (2019). https://doi.org/10.1038/s41590-018-0276-y
12. Domínguez Conde, C., et al.: Cross-tissue immune cell analysis reveals tissue-specific features in humans. Science. **376**, eabl5197 (2022). https://doi.org/10.1126/science.abl5197
13. Abdelaal, T., et al.: A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biol. **20**, 194 (2019). https://doi.org/10.1186/s13059-019-1795-z
14. Peterson, V.M., et al.: Multiplexed quantification of proteins and transcripts in single cells. Nat. Biotechnol. **35**, 936–939 (2017). https://doi.org/10.1038/nbt.3973
15. Xie, X., et al.: Single-cell transcriptomic landscape of human blood cells. Natl. Sci. Rev. **8**, nwaa180 (2021). https://doi.org/10.1093/nsr/nwaa180
16. Bashore, A.C., et al.: High-Dimensional Single-Cell Multimodal Landscape of Human Carotid Atherosclerosis. Arterioscler. Thromb. Vasc Biol. **44**, 930–945 (2024). https://doi.org/10.1161/ATVBAHA.123.320524
17. Luecken, M.D., Theis, F.J.: Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. **15**, e8746 (2019). https://doi.org/10.15252/msb.20188746
18. McCarthy, D.J., Campbell, K.R., Lun, A.T.L., Wills, Q.F.: Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq. data R Bioinf. **33**, 1179–1186 (2017). https://doi.org/10.1093/bioinformatics/btw777
19. Hippen, A.A., et al.: miQC: An adaptive probabilistic framework for quality control of single-cell RNA-sequencing data. PLoS Comput. Biol. **17**, e1009290 (2021). https://doi.org/10.1371/journal.pcbi.1009290
20. Quinn, T.P., Erb, I., Richardson, M.F., Crowley, T.M.: Understanding sequencing data as compositions: an outlook and review. Bioinformatics. **34**, 2870–2878 (2018). https://doi.org/10.1093/bioinformatics/bty175
21. Denisenko, E., et al.: Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. **21**, 130 (2020). https://doi.org/10.1186/s13059-020-02048-6
22. Pasquini, G., Rojo Arias, J.E., Schäfer, P., Busskamp, V.: Automated methods for cell type annotation on scRNA-seq data. Comput. Struct. Biotechnol. J. **19**, 961–969 (2021). https://doi.org/10.1016/j.csbj.2021.01.015
23. Hao, Y., et al.: Integrated analysis of multimodal single-cell data. Cell. **184**, 3573–3587e29 (2021). https://doi.org/10.1016/j.cell.2021.04.048
24. Doni Jayavelu, N., et al.: Single-cell transcriptomic profiling of eosinophils and airway immune cells in childhood asthma. J. Allergy Clin. Immunol. **156**, 923–936 (2025). https://doi.org/10.1016/j.jaci.2025.06.034
25. Heumos, L., et al.: Best practices for single-cell analysis across modalities. Nat. Rev. Genet. **24**, 550–572 (2023). https://doi.org/10.1038/s41576-023-00586-w

26. Büttner, M., Ostner, J., Müller, C.L., Theis, F.J., Schubert, B.: scCODA is a Bayesian model for compositional single-cell data analysis. Nat. Commun. **12**, 6876 (2021). https://doi.org/10.1038/s41467-021-27150-6
27. Dann, E., Henderson, N.C., Teichmann, S.A., Morgan, M.D., Marioni, J.C.: Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat. Biotechnol. **40**, 245–253 (2022). https://doi.org/10.1038/s41587-021-01033-z
28. Mulè, M.P., Martins, A.J., Tsang, J.S.: Normalizing and denoising protein expression data from droplet-based single cell profiling. Nat. Commun. 13, (2099). (2022) https://doi.org/10.1038/s41467-022-29356-8
29. Zimmerman, K.D., Espeland, M.A., Langefeld, C.D.: A practical solution to pseudoreplication bias in single-cell studies. Nat. Commun. **12**, 738 (2021). https://doi.org/10.1038/s41467-021-21038-1
30. Büttner, M., Miao, Z., Wolf, F.A., Teichmann, S.A., Theis, F.J.: A test metric for assessing single-cell RNA-seq batch correction. Nat. Methods. **16**, 43–49 (2019). https://doi.org/10.1038/s41592-018-0254-1
31. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. **19**, 15 (2018). https://doi.org/10.1186/s13059-017-1382-0
32. Martens, J.H.A., Stunnenberg, H.G.: BLUEPRINT: mapping human blood cell epigenomes. Haematologica. **98**, 1487–1489 (2013). https://doi.org/10.3324/haematol.2013.094243
33. ENCODE Project Consortium: An integrated encyclopedia of DNA elements in the human genome. Nature. **489**, 57–74 (2012). https://doi.org/10.1038/nature11247
34. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W., Robinson, M.D.: Doublet identification in single-cell sequencing data using scDblFinder. F1000Res. **10**, 979 (2022). https://doi.org/10.12688/f1000research.73600.2
35. Lun, A.T.L., et al.: EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. **20**, 63 (2019). https://doi.org/10.1186/s13059-019-1662-y
36. Young, M.D., Behjati, S.: SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. GigaScience. **9**, giaa151 (2020). https://doi.org/10.1093/gigascience/giaa151
37. Holm, S.A., Simple Sequentially Rejective Multiple Test Procedure: Scand. J. Stat. **6**, 65–70 (1979). https://www.jstor.org/stable/4615733
38. Hatje, K., Bell, E.: Comparison of Fixed Single Cell RNA-seq Methods to Enable Transcriptome Profiling of Neutrophils in Clinical Samples. Zenodo (2025). https://doi.org/10.5281/zenodo.15373124
39. Kwok, A.J., et al.: Neutrophil and emergency granulopoietic drivers of sepsis immune suppression and an extreme response to infection. Zenodo (2023). https://doi.org/10.5281/zenodo.7924238

Figures

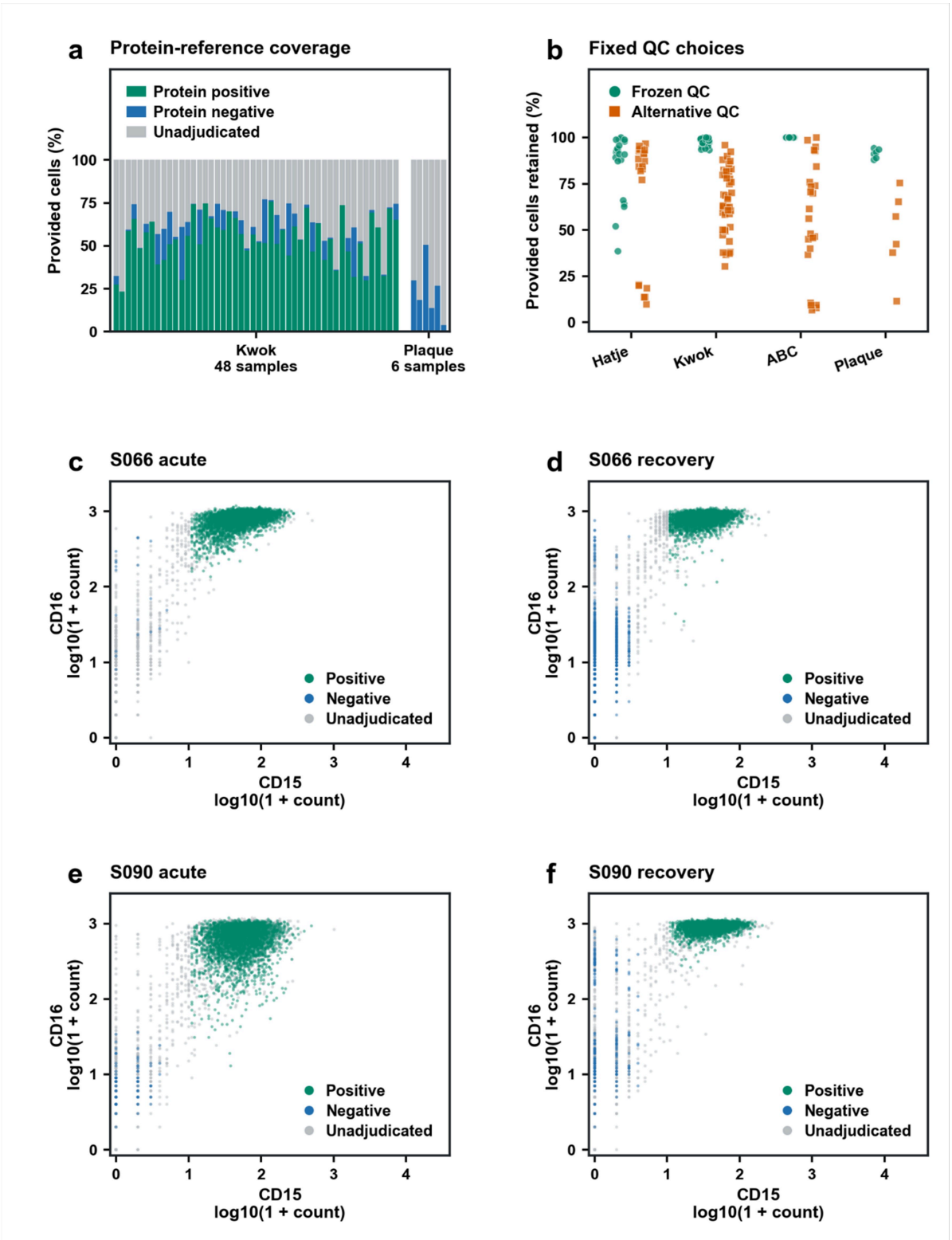


Figure 1

Independent measurement domains and fixed RNA quality control. a, Protein-reference coverage in all 48 Kwok whole-blood samples and six plaque samples with measured antibody-derived tags (ADT). Each bar is one sample; all source-retained cells remain in the denominator. Protein-positive, protein-negative and unadjudicated compartments were assigned using protein measurements alone. b, Retention under the frozen RNA quality-control rule (at least 300 UMIs, at least 200 detected genes and mitochondrial

fraction at most 20%) and the pre-existing alternative rule (100–2,500 detected genes and mitochondrial fraction below 10%). Points represent all 21 Hatje preparations, 48 Kwok samples, 24 ABC computational partitions and six plaque samples. Partitions and preparations are not additional independent donors. c–f, Measured CD15 and CD16 counts in two illustrative acute–recovery pairs, transformed as log10(1 + count). S066 was processed in the same batch at both time points; S090 changed batch. Every cell in each displayed sample is plotted, without subsampling. Colors show the complete protein-only reference assignment, which also uses other measured markers; these two-dimensional views do not define new gates. Protein values were measured rather than imputed from RNA. Unadjudicated cells remain explicit and contribute to whole-sample composition bounds. No RNA result or patient outcome was used to select these protein rules or the two RNA QC settings.

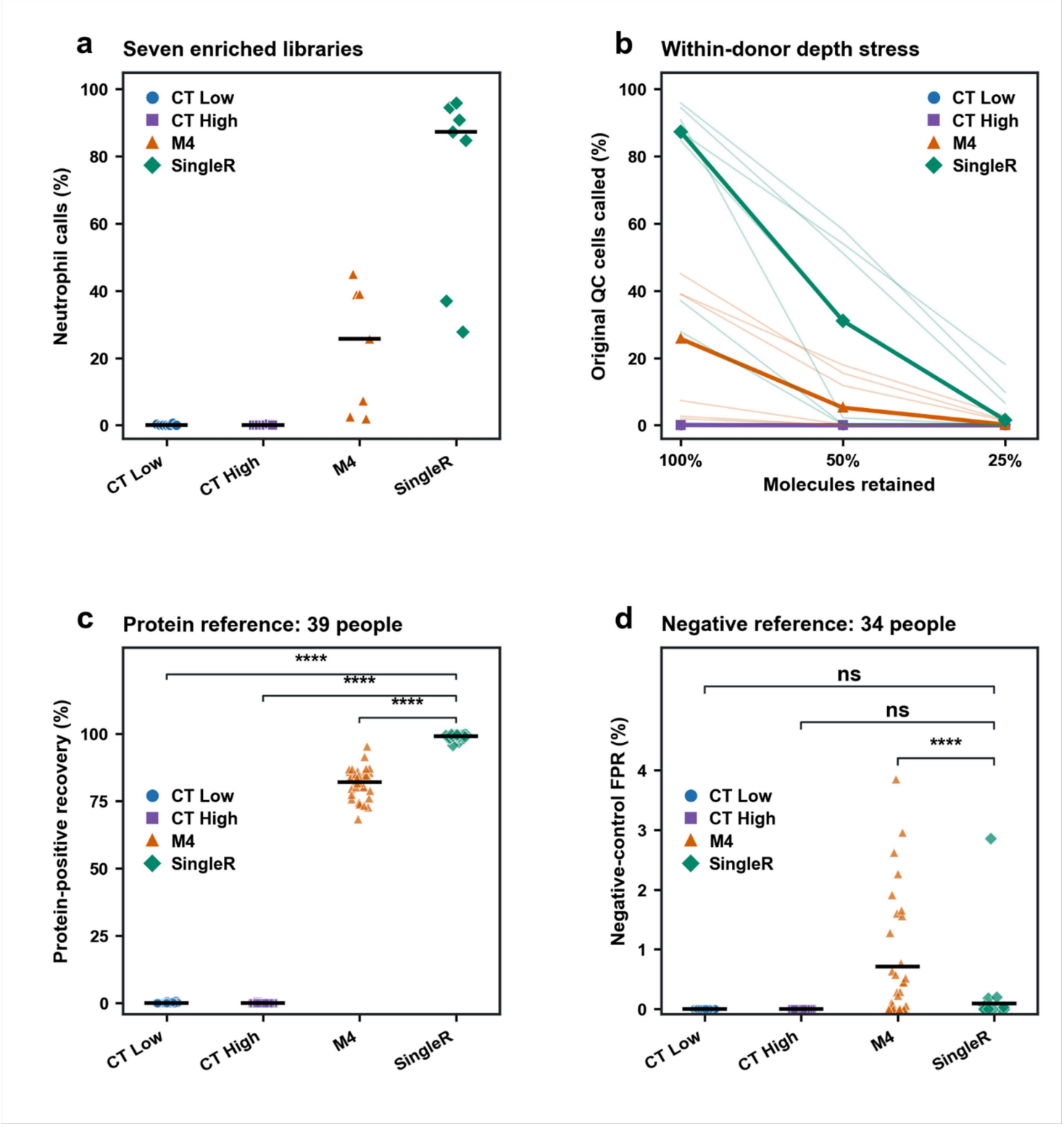


Figure 2

Reference-dependent calling and independently measured neutrophil recovery. a, Neutrophil calling fractions in seven physically enriched neutrophil libraries from seven GSE188288 donors. Points are donors and black segments are medians. The denominator comprises the 43,643 cells retained by frozen QC across these libraries; enrichment does not provide an individual-cell truth label, so these are calling fractions. b, Existing within-library molecule-thinning results at 100%, 50% and 25% depth. Thin

lines connect each donor and thick lines connect donor medians. The denominator remains the original full-depth QC cell set, so the endpoint includes both QC loss and annotation loss. Each reduced-depth condition is one technical perturbation per library, not a new biological replicate. c, Recovery of protein-positive Kwok cells after frozen QC and annotation, with all source-retained protein positives in the denominator. d, Neutrophil calls among QC-retained protein-negative controls, expressed as the false-positive rate (FPR). Each point in c and d is a participant mean across available repeat samples; black segments are means of these participant means. There are 39 participants for recovery and 34 with a usable negative-control denominator. Horizontal jitter only separates marks. CT Low and CT High denote the frozen CellTypist Immune_All_Low and Immune_All_High configurations; M4 denotes the frozen marker-and-anchor rule; SingleR denotes Blueprint/ENCODE-based annotation with native pruning. The comparisons concern these specific method–reference combinations. Zero observed negative-control calls do not imply complete specificity outside the measured reference compartments. Exact pooled counts and rejection counts are given in Table 2. Brackets in c and d use two-sided exact paired sign tests, Holm-adjusted across seven display-stage comparisons (six here and one in Fig. 4c). Nonzero differences number 39 for each c contrast and 5, 5 and 20 for d comparisons against CT Low, CT High and M4, respectively. Adjusted P values are 2.55×10^{-11} for each c contrast, and 0.125, 0.125 and 7.63×10^{-6} in d. ****$P < 0.0001$; ns, $P \geq 0.05$. These explanatory tests concern paired directions; displayed means remain descriptive.

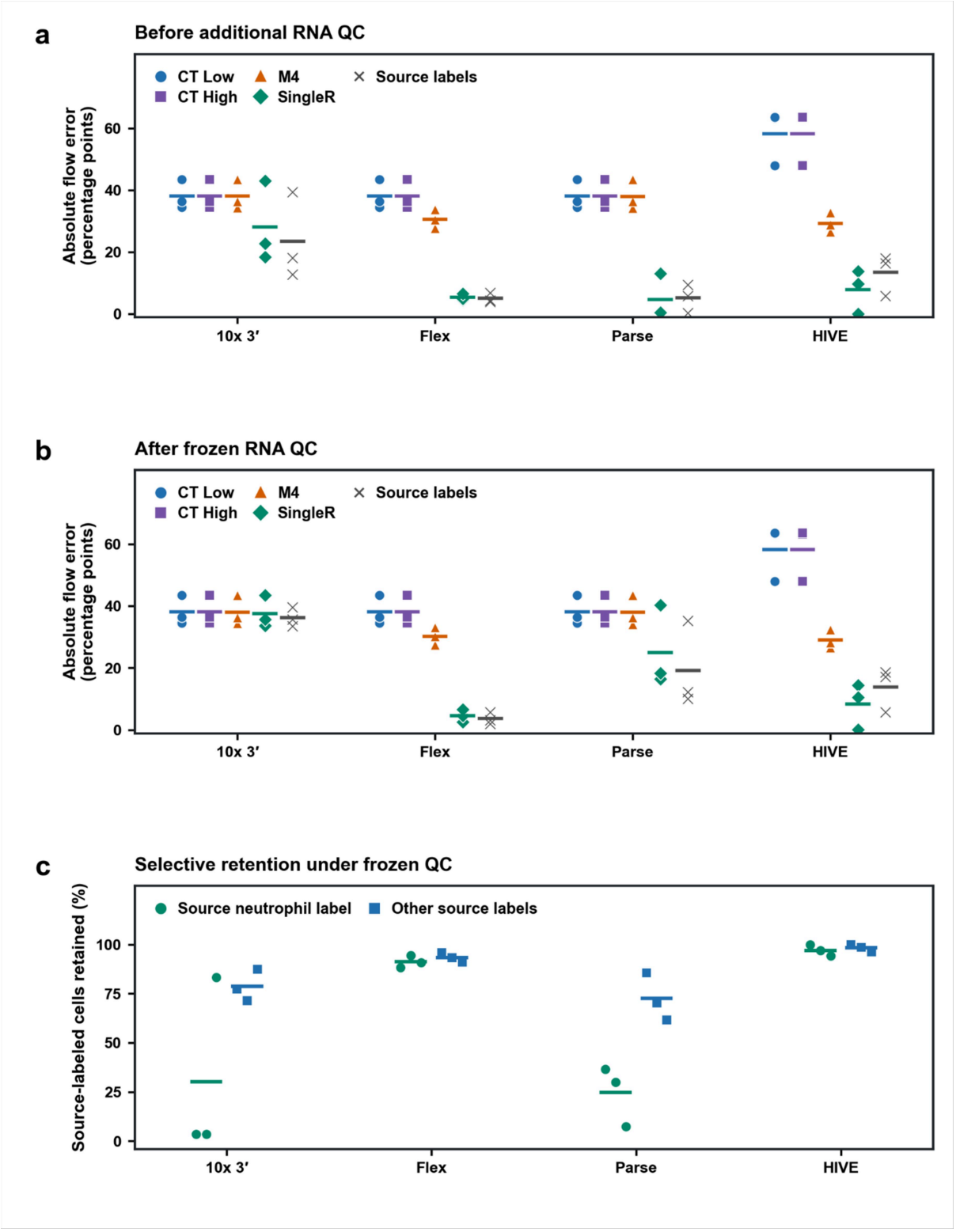


Figure 3

Annotation and RNA filtering contribute distinct composition errors against matched flow. a and b, Absolute neutrophil-composition errors in red-cell-depleted Hatje preparations before additional RNA QC (a) and after the frozen QC rule (b). The denominator excludes source-labeled platelets to approximate the flow CD45 compartment. Points are individual donors and short horizontal lines are arithmetic means. Donors 1–3 recur across 10x 3', Flex and Parse, whereas HIVE uses separate donors A–C. The

comparison therefore contains six biological donors and three observations per technology, rather than twelve independent platform–donor observations. Flow references are the provenance-linked author medians stored in the public analysis notebook. The source-annotation comparator is descriptive. c, Retention after frozen QC of source-annotated neutrophils and other source-annotated cells in the same red-cell-depleted preparations. Points are donors and segments are arithmetic means; these source labels describe selective filtering and are not an independent cell-identity reference. All RNA QC calculations use the published 14,966-feature intersection. The source-retained input is already processed and does not reconstruct all originally captured droplets. A published donor-2 discrepancy is retained in the audit; excluding that donor changes the SingleR Parse mean absolute error from 6.73 to 28.37 percentage points across the same before/after comparison, preserving its direction. Frozen QC requires at least 300 UMIs, at least 200 genes and mitochondrial fraction at most 20%. Method colors and symbols follow Figure 2; SingleR is pruned.

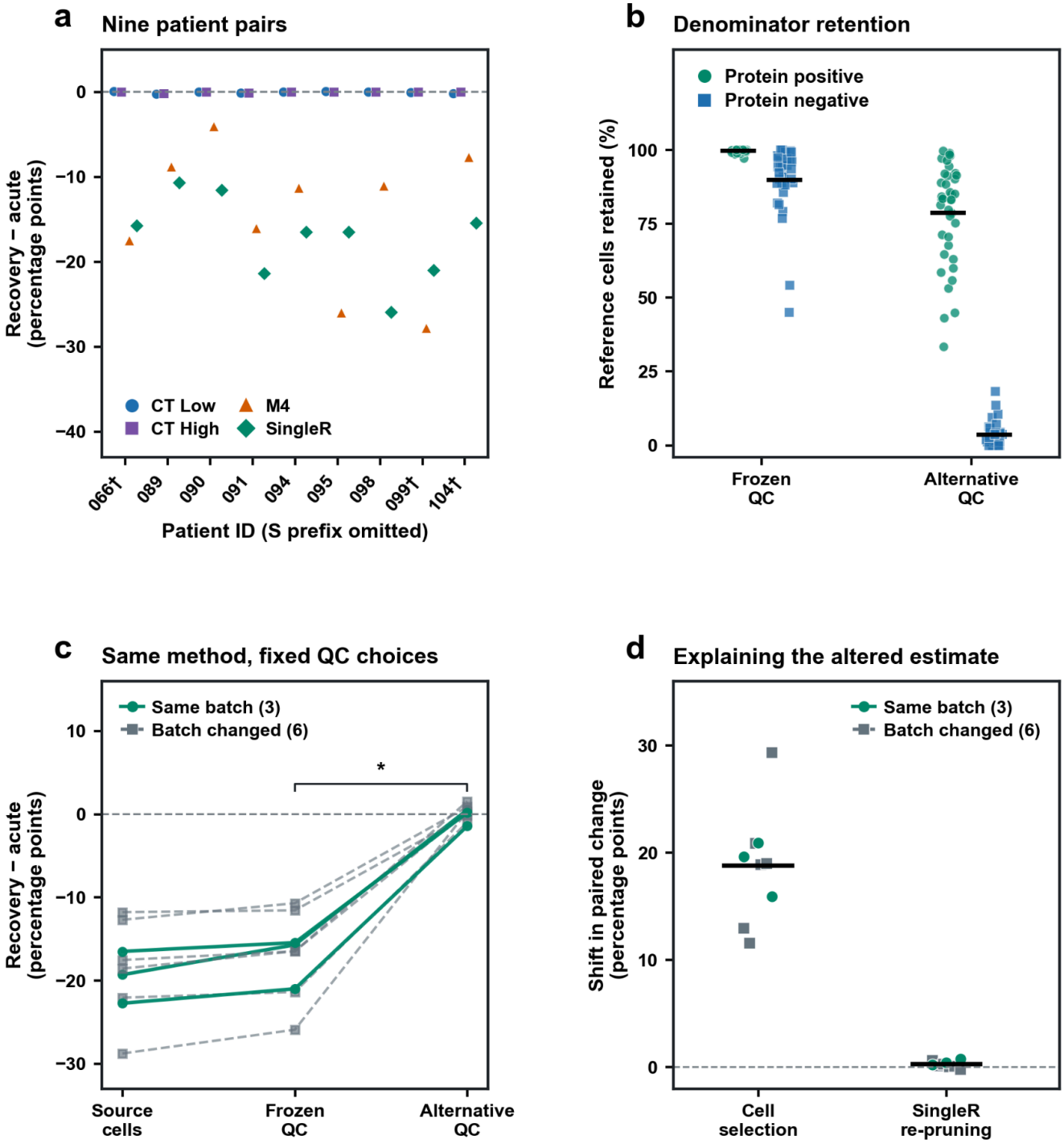


Figure 4

Selective denominator loss alters within-patient estimates despite an unchanged annotator. a, Recovery-minus-acute changes in the RNA-estimated neutrophil fraction for all nine explicitly linked Kwok patients under frozen QC. The letter S is omitted from patient IDs; daggers (†) identify the three same-batch pairs. The remaining six pairs changed processing batch. Each symbol is one patient–method result; the dashed line denotes no change. b, Retention of protein-positive and protein-negative

reference cells under frozen QC and the fixed alternative rule. Points are participant means across repeat samples (39 with positives and 34 with negatives); black segments are means of participant means. The pooled cell retentions under alternative QC are 116,319/148,138 positives (78.52%) and 652/19,026 negatives (3.43%). c, SingleR paired changes across the source-retained, frozen-QC and alternative-QC domains. Lines connect the same patient across computational choices, not additional clinical time points. Same-batch pairs are green solid lines; cross-batch pairs are gray dashed lines. SingleR yields decreases in nine of nine pairs under frozen QC and three of nine under alternative QC. d, Existing algebraic decomposition of the change from the source-retained to alternative-QC estimate into cell selection and query-dependent SingleR re-pruning. Points are patients; black segments are arithmetic means. Selection and re-pruning contribute mean shifts of +18.77 and +0.25 percentage points, respectively. This is a post-comparison explanatory calculation, without retuning labels or thresholds. Frozen QC requires at least 300 UMIs, at least 200 genes and mitochondrial fraction at most 20%; alternative QC requires 100–2,500 genes and mitochondrial fraction below 10%. The protein-adjudicated subset decreases in all nine pairs, but whole-sample protein bounds span zero for every pair. The figure therefore evaluates sensitivity of patient estimates, not calibrated whole-blood changes or a causal treatment effect. All methods, input domains, same-batch results and singlet sensitivity are retained in Table 3. Bracket in c: two-sided exact paired sign test across nine patients, Holm-adjusted within the seven display-stage comparisons; $P = 0.0117$ (*$P < 0.05$). This explanatory test concerns the shift caused by changing QC, not recovery versus acute at either single QC setting.

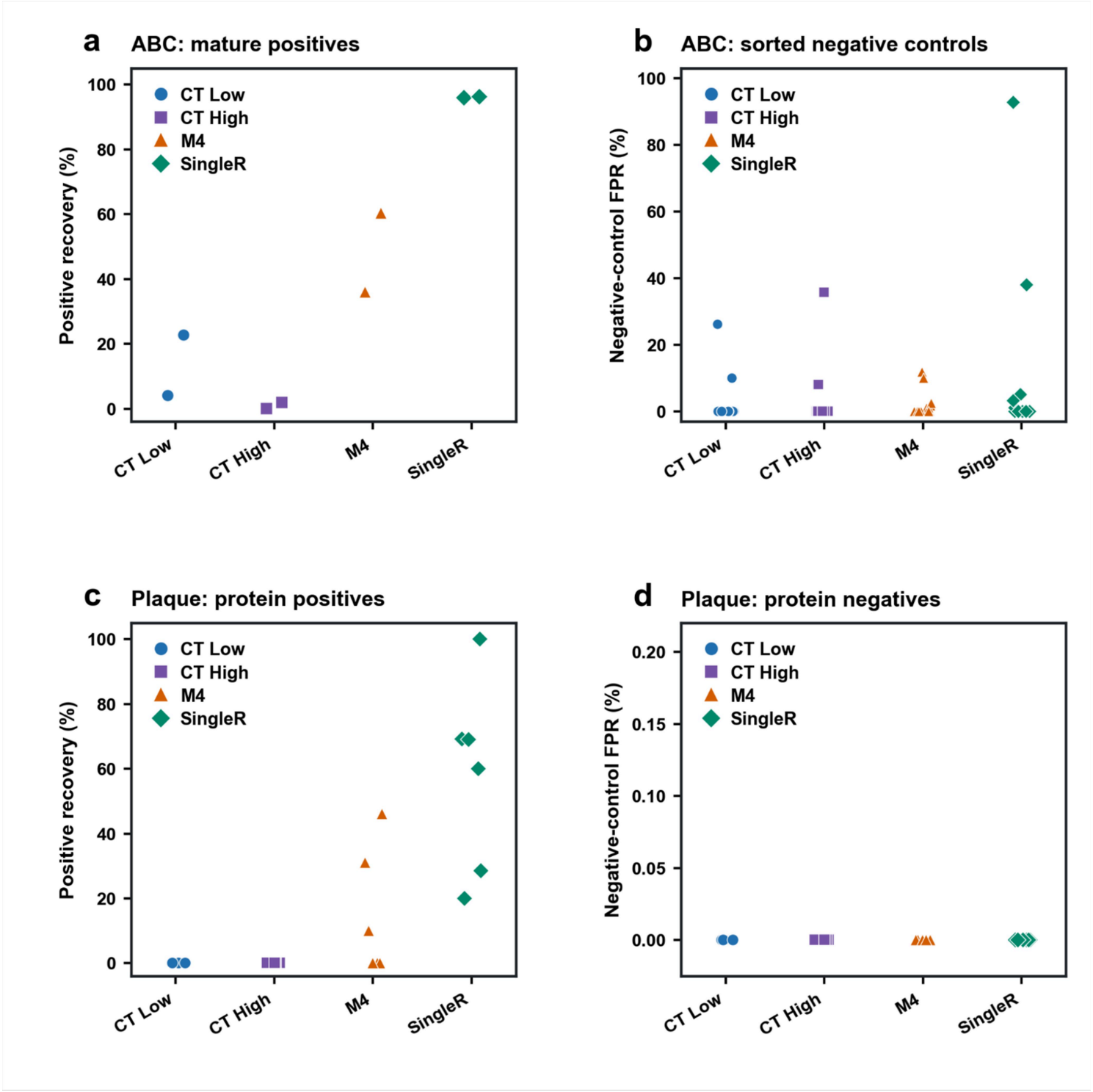


Figure 5

Physical sorting and tissue protein controls test transfer across measurement domains. a, Recovery of 103 physically sorted mature ABC neutrophils from two bone-marrow donor aliases, with one point per alias. b, Neutrophil calls among 2,008 predeclared mature B-, T-, NK- and monocyte-sorted controls, shown for all 12 negative-control donor aliases. These aliases are sorting-list identifiers and are not asserted to represent 12 distinct people. The high-discordance NK-sorted groups remain visible and were not relabeled using RNA. SingleR calls 65/2,008 negatives, including 64/230 NK-sorted cells and

1/1,778 B-, T- or monocyte-sorted cells. c, Recovery of 68 protein-positive cells across six plaque samples. Each point is one sample; positive denominators range from 4 to 29 source-retained cells. d, False-positive rates among the 6,267 QC-retained protein-negative plaque cells, with one point per sample. All four primary methods yield zero calls in that negative-control compartment. Positive recovery uses all source-retained positives in the denominator, including QC exclusions; negative-control rates use QC-retained negatives. All plotted comparisons use the frozen QC rule, and horizontal jitter only separates marks. Method colors and symbols follow Figure 2; SingleR is pruned. Table 2 also reports unpruned SingleR, rejection counts, and the separately evaluated 382 physically sorted earlier neutrophil-stage cells. Earlier-stage cells are not added to the mature-neutrophil denominator. The operational physical and protein labels are preserved across methods.