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.

cSVI-subtype: a GWAS-anchored, vascular-centered framework for molecular subtype barcode discovery in cerebral small vessel disease

Yuqian Li

yuqian-li@hotmail.com

Independent Researcher https://orcid.org/0009-0003-9256-9420

Method Article

Keywords:

Posted Date: June 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9937210/v1

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

Additional Declarations: The authors declare potential competing interests as follows: The author is the developer of the cSVI-subtype framework and the csviSubtype R package described in this preprint, and has filed related software copyright records. The package is publicly available for research use. The author declares no current financial competing interests.

cSVI-subtype: a GWAS-anchored, vascular-centered framework for molecular subtype barcode discovery in cerebral small vessel disease

Yuqian Li[1*]

[1] Independent Researcher, Beijing, China

*Corresponding author:

Yuqian Li, Independent Researcher, Beijing, China. Email: yuqian-li@hotmail.com

Abstract

Motivation:

Cerebral small vessel disease (cSVD) is still classified mainly by clinical phenotypes and neuroimaging features, which provide limited molecular resolution for robust and transferable subtype definition. We developed cSVI-subtype to derive compact and reproducible molecular subtype barcodes for cSVD by integrating human genetics with cross-species single-cell evidence.

Results:

cSVI-subtype is a GWAS-anchored, vascular-centered framework for molecular subtype barcode discovery. It prioritizes human candidate genes through colocalization and supportive Mendelian randomization, and then projects these signals into cross-species single-cell disease models. The framework integrates a layered vascular evidence chain comprising vascular abundance gating (CSS), source-to-vascular interaction strength (CIS), and ligand-receptor-to-gene allocation with pathway/GO projection (LRS), enabling subtype-specific prioritization while mitigating receptor-level sparsity. Applied to cSVD, cSVI-subtype identified compact and reproducible barcodes for A1/N3KO-like and B1/BCAS-like states and remained stable across GWAS perturbation, single-chain ablation, and internal control analyses. These results support cSVI-subtype as a practical framework for molecular subtype definition in cSVD and suggest its potential extensibility to other vascular-related disorders.

Availability and implementation:

cSVI-subtype is implemented in R and distributed as the package csviSubtype (version 0.1.0). Source code, documentation, shipped example input files, and a quickstart vignette are freely available at https://github.com/YuqianLii/csviSubtype. An archived release will be deposited in Zenodo for the journal version.

Contact:

yuqian-li@hotmail.com

Supplementary information:

Supplementary data are available at https://github.com/YuqianLii/csviSubtype.

Introduction

Cerebral small vessel disease (cSVD) is one of the most common cerebrovascular disorders in older adults and an important pathological basis of ischemic stroke, intracerebral hemorrhage, vascular cognitive impairment, and dementia (Chen, et al., 2021; Duering, et al., 2023; Hilkens, et al., 2024; Li, et al., 2025; Li, et al., 2025; Markus and Joutel, 2025). It primarily affects small arteries, venules, capillaries, and their surrounding structures in the brain(Liu, et al., 2024; Pantoni, 2010). As populations age, the burden of cSVD is expected to increase further, making mechanistic dissection and precise subtype stratification increasingly important.

Recent efforts to investigate the molecular basis of cSVD have moved beyond descriptive clinical and neuroimaging characterization toward integrative analyses of human genetics and disease-relevant cellular states(Deng, et al., 2025; Sargurupremraj, et al., 2020). Advances in large-scale population genetics and multi-omic integration have enabled prioritization of cSVD-related candidate genes, cell types, and pathways from multiple layers of evidence, including genome-wide association studies (GWAS), quantitative trait locus (QTL) resources, colocalization analysis, Mendelian randomization, and single-cell transcriptomics(Ma, et al., 2023; Sun, et al., 2025; Townsend, et al., 2024; Traylor, et al., 2021). However, most existing studies remain focused on gene discovery or risk prioritization. They are more effective at identifying cSVD-associated genes than at defining how these signals can be organized into reproducible and mechanistically interpretable molecular subtype structures.

Consequently, current cSVD classification frameworks still rely predominantly on clinical presentation and neuroimaging features, with limited molecular resolution for robust and transferable subtype definition.

This limitation is particularly important in cSVD, where disease progression is likely driven by coordinated dysfunction across multiple cellular compartments rather than abnormalities in a single cell type. Previous studies have implicated inflammatory regulation, vascular homeostasis disruption, extracellular matrix remodeling, immune signaling, and neurovascular unit dysfunction in cSVD pathogenesis(Hong, et al., 2024; Le Grand, et al., 2024; Li, et al., 2025; Li, et al., 2025; Li, et al., 2022; Markus and Joutel, 2025). A central unresolved challenge is therefore how to couple genetically anchored disease signals with layered mechanistic evidence from pathological cell states to derive interpretable and reproducible molecular subtype barcodes.

To address this challenge, we developed cSVI-subtype, a GWAS-anchored and vascular-centered framework for molecular subtype definition in cSVD that integrates human genetic evidence with cross-species single-cell disease models. The framework first prioritizes disease-relevant candidate genes through colocalization analysis with supportive Mendelian randomization, and then projects these signals into disease-model single-cell datasets for mechanistic reinforcement and subtype assignment through a layered vascular evidence chain. Rather than returning only extended candidate-gene lists or enriched pathways, cSVI-subtype is designed to

distill compact, stable, and cross-dataset reproducible molecular subtype barcodes. Here, we applied this framework to representative cSVD disease states to identify subtype-specific molecular footprints for N3KO-like and BCAS-like conditions, and to evaluate its feasibility and robustness for molecular subtype definition.

Materials and methods

Overview of the cSVI-subtype framework

We developed cSVI-subtype, a GWAS-anchored and vascular-centered framework for molecular subtype barcode discovery in cerebral small vessel disease (cSVD). The framework integrates stable human genetic evidence with cross-species single-cell disease-state signals to derive compact and reproducible subtype barcodes. Briefly, human GWAS summary statistics for white matter hyperintensity and lacunar stroke were processed through genetic prioritization, colocalization, and supportive Mendelian randomization to define genetically anchored candidate features. These signals were then projected into mouse single-cell disease models representing N3KO-like and BCAS-like vascular states, where a layered vascular evidence chain was used to derive subtype-specific scRNA features. Finally, genetic and scRNA evidence were fused to generate subtype-specific molecular footprints and evaluated through robustness and baseline analyses.

Human genetic anchoring and cross-species candidate mapping

Four human GWAS summary-statistics datasets from HuGeAMP were used, including discovery and validation datasets for white matter hyperintensity and lacunar stroke. Discovery datasets served as the primary human genetic anchor, whereas validation datasets were reserved for out-of-sample perturbation and replication analyses. GWAS loci were interrogated against cis-eQTL resources from eQTLGen whole blood and GTEx v7 vascular and brain tissues to prioritize genetically supported candidate genes. Colocalization was performed within lead $SNP \pm 1000kb$ regions using the Approximate Bayes Factor framework implemented in coloc.abf, and PP4 was used as the main metric of shared causal signal support. Mendelian randomization was incorporated as a supportive prioritization layer rather than a stand-alone causal gate. Candidate genes from blood and vascular-relevant tissues were then summarized to define the GWAS-backed vascular feature layer. Finally, one-to-one human-mouse ortholog mapping was applied to project human candidate genes into downstream mouse scRNA-seq disease models, whereas genes without reliable one-to-one orthologs were retained only in human-side summary tables.

Single-cell disease models and vascular-centered evidence extraction

Two mouse scRNA-seq datasets from GEO were used to represent cross-species disease states relevant to cSVD: GSE204803 for the Notch3-deficient (N3KO-like) axis and GSE263191 for the BCAS-like chronic hypoperfusion axis. Samples were

grouped as A1/A2 for the Notch3 axis and B1/B2 for the BCAS axis, with A1 versus B1 used as the primary subtype-discovery contrast and within-study contrasts reserved for internal replication analyses. Single-cell objects were integrated using Harmony to obtain a stable low-dimensional manifold for annotation and downstream vascular-gate definition. To maximize robustness under the current data structure, major cell-type annotation was restricted to four classes: vascular, immune, glia, and neuron, with the primary analysis gate defined as the vascular compartment.

Within this framework, the scRNA layer was modeled as a vascular-centered source-to-target transmission system, in which non-vascular cell classes acted as candidate upstream sources and vascular cells served as the target layer. Three linked evidence modules were used: cell subtype score (CSS), cell interaction score (CIS), and ligand-receptor score (LRS). Together, these modules quantified vascular abundance, source-to-vascular interaction strength, and gene-level signaling allocation, and were integrated with vascular differential-expression information to derive phenotype-specific scRNA feature rankings.

Cell subtype score (CSS)

MCSS was used as a vascular-capacity gate rather than as a direct cross-group comparison metric. For each phenotype p, vascular abundance $R_{v,p}$ was calculated as the fraction of vascular cells in that group, and a gating factor C_p was defined using a frozen threshold $\tau_{vascular} = 0.10$. Higher vascular abundance therefore

increased the effective transmission capacity of downstream CIS and LRS components. In parallel, differential expression within the vascular gate was used to quantify gene-level vascular state bias, and the magnitude term M_g was retained for downstream feature scoring. In the current implementation, the main vascular differential table was derived from the B1-versus-A1 comparison, with direction labels used to assign subtype orientation.

$$C_p = g(R_{v,p}) = min(1, \frac{R_{v,p}}{\tau_{vascular}})$$

$$M_g = \left| logFC_g \right|$$

Cell interaction score (CIS)

CIS quantified the influence of each source cell class on the vascular target layer. Cell-cell communication was estimated separately within A1 and B1 using CellChat. For each source-to-vascular edge, two summary quantities were extracted: the number of inferred interactions and the summed interaction weight. These quantities were normalized within each phenotype to control for network-size effects and combined into a base interaction score. Disease-conditioned source priors and the CSS-derived gate were then incorporated to obtain the final CIS value for each source-to-vascular relationship. In the main v1.0 analysis, the top three source cell classes per phenotype were retained as the dominant upstream contributors to vascular-state modulation.

$$\widetilde{N}^{cell}_{s \to v,p} = \frac{N^{cell}_{s \to v,p}}{\sum_{s \in S} N^{cell}_{s \to v,p} + \varepsilon}$$

$$\widetilde{W}^{cell}_{s\to v,p} = \frac{W^{cell}_{s\to v,p}}{\sum_{s\in S} W^{cell}_{s\to v,p} + \varepsilon}$$

$$I^{base}_{s\to v,p} = \alpha \widetilde{N}^{cell}_{s\to v,p} + (1-\alpha)\widetilde{W}^{cell}_{s\to v,p}$$

$$S^{int}_{s\to v,p} = \pi_s \cdot C_p^{\gamma} \cdot I^{base}_{s\to v,p}$$

$$S^{topK}_{v,p} = TopK_{s\in S}(S^{int}_{s\to v,p}),\ \ K = 3$$

Ligand-receptor score (LRS) and gene-level allocation

LRS was used to allocate source-to-vascular interaction support to specific vascular genes. Pair-level ligand-receptor communication was extracted from CellChat for the retained $topK$ source classes targeting vascular cells. To avoid underestimation caused by receptor-complex strings, receptor complexes were decomposed into single-gene receptor components, and edge weights were equally divided among the component receptors so that total weight was conserved. Because direct receptor-level matching led to excessive sparsity, the operational v1 framework adopted a pathway-mediated allocation strategy, in which receptor-level support was first projected onto GO/pathway terms and then redistributed to direction-matched vascular DEG genes within each term. Final gene-level scRNA feature strength was computed by integrating upstream CIS support, pathway-mediated gene allocation, and the vascular DEG magnitude term M_g.

$$\hat{S}^{LR}_{(l,r),s\to v,p} = \frac{S^{LR}_{(l,r),s\to v,p}}{\sum_{(l',r')\in \tilde{L}_{s\to v,p}} S^{LR}_{(l',r'),s\to v,p} + \varepsilon}$$

$$\delta^{LR}_{(l,r),s\to v,p} = C_p \cdot \hat{S}^{LR}_{(l,r),s\to v,p}$$

$$F_{g,v,p}^{gene} = M_g \cdot \sum_{s \in S_{v,p}^{topK}} S_{s \to v,p}^{int} \cdot A_{g,s,p}$$

scRNA evidence stratification and subtype-specific feature extraction

To standardize scRNA evidence strength across genes, a combined percentile score was computed using both vascular DEG magnitude and gene-level feature rank. For each phenotype, percentile calculations were restricted to the direction-matched vascular DEG universe, and zero-valued gene-feature scores were assigned a percentile of zero to avoid inflation from tied values. The final integrated scRNA percentile was defined as the maximum of the DEG percentile and the gene-feature percentile, and genes were stratified into high, mid, and low evidence tiers. Directional subtype-specific feature sets were then defined according to the frozen contrast orientation, yielding phenotype-specific candidate features for A1-like and B1-like states.

GWAS-scRNA fusion and subtype barcode definition

GWAS and scRNA evidence were fused into a common molecular-footprint space. In the original evidence-map design, genes were classified into red, orange, yellow, and grey zones according to their GWAS and scRNA evidence strengths, with direction labels retained but not used to alter strength classification. Because strict cross-modal overlap was sparse under the current v1.0 setting, the main subtype barcode output was defined using a binary-core rule. Under this operational definition, genes

supported by both the strict GWAS vascular backbone and the subtype-specific scRNA feature set were labeled red, and the resulting red-zone genes constituted the final binary subtype barcode for each phenotype.

$$Z_{g,p}^{binary} = 1(zone\_binary(g,p) = "red")$$

$$Barcode_p = \left\{g : Z_{g,p}^{binary} = 1\right\}$$

Framework evaluation and robustness analyses

Framework robustness was evaluated from three complementary directions. First, to assess sensitivity to the human genetic anchor, the full GWAS-colocalization pipeline was rerun on independent validation GWAS datasets, and discovery-only, validation-only, and pooled GWAS-backed candidate sets were separately fused with the fixed scRNA layer to generate alternative barcode versions. Second, to address the possibility that the main A1-versus-B1 contrast reflected cross-study effects rather than subtype biology, frozen A1 and B1 barcodes were tested within each original study and against cross-negative controls to examine within-study directionality and non-specific vascular-stress signals. Third, single-chain ablation analyses were performed by disabling CSS, CIS, or LRS individually while keeping the GWAS feature layer and fusion rules unchanged.

Baseline comparison and reproducibility

Several baseline strategies were used to benchmark the value of the integrated framework, including GWAS-only, scRNA-only, coloc-only, MR-only, and naïve GWAS-scRNA set-operation baselines. Comparator performance was assessed primarily by barcode stability across discovery and validation GWAS runs, together with candidate-set size, subtype specificity, and GWAS anchoring rate.

All analyses were implemented in R 4.4.0 using standard packages for GWAS, colocalization, Mendelian randomization, and single-cell analysis, and the core framework has been released as the R package csviSubtype (version 0.1.0). Main pipeline outputs were versioned and organized into standardized intermediate and summary tables to ensure reproducibility and traceability.

Results

cSVI-subtype integrates human genetic anchoring with vascular-centered scRNA evidence

We first established the overall architecture of cSVI-subtype, which links a human genetic anchor layer to a vascular-centered single-cell disease layer for subtype barcode discovery (Fig. 1). Starting from WMH and lacunar stroke GWAS signals, the framework defined a genetically constrained candidate space and derived subtype-resolved vascular evidence from cross-species scRNA disease models. Integration of these two layers yielded the final A1/N3KO-like and B1/BCAS-like barcode outputs, which served as the basis for the downstream analyses.

Human genetic anchor construction produced a compact and stable GWAS-backed candidate universe

To establish the human genetic anchor layer, WMH and lacunar stroke GWAS datasets were analyzed under discovery, validation, and pooled settings. Across these three runs, the GWAS-backed candidate universe remained compact, comprising 27 genes in the discovery run, 26 genes in the validation run, and 32 genes in the pooled run, with only modest expansion in the pooled setting and broadly similar tier composition overall (Fig. 2A). This pattern indicates that the human genetic layer converged to a small and stable candidate universe rather than a diffuse list of weak associations.

Colocalization support within this compact universe was concentrated in a limited subset of strongly supported genes rather than being broadly distributed across all candidates (Fig. 2B). Several genes showed consistently high PP4 support across runs, including *Zcchc14*, *Nbeal1*, *Sh3pxd2a*, and *Nmt1*, whereas other loci displayed more moderate or run-specific support. Notably, several of the strongest pooled anchors were directly retained in the final subtype barcodes, linking the human genetic layer to the downstream subtype-definition output (Fig. 2C). Run-level comparison further showed that most top anchor genes were preserved across discovery, validation, and pooled summaries (Fig. 2D), indicating that the GWAS-backed candidate universe was robust to changes in the input human dataset. Together, these results show that the

human anchor layer of cSVI-subtype is compact and cross-run stable, providing a genetically constrained starting point for subsequent vascular-centered scRNA integration.

Vascular-centered scRNA evidence identified differential vascular abundance, sender-to-vascular relay structure, and compact subtype-resolved feature genes

We next examined vascular-centered scRNA evidence that projected the human genetic anchor into disease-state cellular contexts. At the cell-composition level, CSS revealed marked branch differences in vascular abundance: vascular cells accounted for approximately 27–34% of all cells in the A1/N3KO-like branch, but only ~3–4% in the B1/BCAS-like branch (Fig. 3A). Within the vascular gate, both branches were predominantly endothelial, although A1 showed a stronger endothelial bias, whereas B1 contained a relatively larger mural component (Fig. 3B).

We then examined sender-to-vascular communication using CIS and LRS. Relay burden was concentrated in glial input rather than broadly distributed across all upstream compartments. In A1, glia-to-vascular communication comprised four interactions with the highest summed interaction weight, whereas immune and neuronal contributions were absent. In B1, glia-to-vascular communication expanded to nine interactions, and a small immune-to-vascular component newly appeared, but the total interaction burden was lower and more diffuse than in A1. Sender-level summaries further showed that relay efficiency remained high in the glial branch,

consistent with a structured rather than nonspecific communication architecture (Fig. 3C).

Finally, these layered scRNA signals were distilled into compact subtype-resolved vascular feature genes. The resulting feature layer separated A1-enriched and B1-enriched programs and prioritized a small set of top-ranked genes by footprint score, including *TTR*, *AES*, *PTGDS*, *MGP*, *ATP5O.1*, and *USMG5* on the A1 side, and *NRG3*, *CDR1OS*, *DPP10*, *TLE5*, *GAS5*, and *KCNIP4* on the B1 side (Fig. 3D). These compact vascular feature programs were retained as the downstream scRNA evidence layer for cross-modal fusion rather than as the final subtype barcode itself.

Cross-modal fusion yields compact subtype barcodes

Using the operational v1.0 binary-core rule, we retained genes jointly supported by the strict GWAS vascular backbone and the subtype-specific scRNA feature layer as the final red-zone barcode for each phenotype. In the discovery analysis, the GWAS pipeline produced a 27-gene vascular backbone for both branches, which was fused with substantially larger scRNA candidate spaces comprising 2,660 A1 genes and 3,562 B1 genes. Despite these broad upstream candidate layers, cross-modal fusion converged to a compact 4-gene A1/N3KO-like barcode and a 9-gene B1/BCAS-like barcode, indicating strong compression from candidate space to final subtype output (Fig. 4A).

The final A1/N3KO-like barcode consisted of *Sh3pxd2a*, *Gnb2*, *Sfr1*, and *Mrpl38*, whereas the B1/BCAS-like barcode consisted of *Zcchc14*, *Nmt1*, *Nbeal1*, *Celf1*, *Gigyf1*, *Banp*, *Slk*, *Tab2*, and *Sptbn1* (Fig. 4B). Notably, the two barcode sets showed no overlap in the discovery analysis, indicating complete subtype separation at the final binary output layer (Fig. 4B). At the gene level, barcode members retained variable but interpretable support profiles across GWAS and scRNA evidence, rather than being driven by a uniform signal pattern (Fig. 4C). By contrast, the stricter multi-zone core remained empty for both subtypes under the discovery setting, whereas the binary-core rule retained 4 and 9 genes, respectively (Fig. 4D). Together, these results support the binary-core formulation as the primary reported output of the current v1.0 framework, with the stricter formulation retained as a sensitivity layer rather than the main subtype definition.

External validation and pooled GWAS perturbation support barcode stability

We next asked whether the final subtype barcodes were robust to changes in the human genetic anchor by rerunning the GWAS-colocalization pipeline on independent validation datasets and under a pooled setting that combined discovery and validation support. At the backbone level, the validation run produced only a modest reduction in the GWAS-supported vascular candidate set relative to discovery, whereas the pooled run modestly expanded the backbone. Despite these upstream changes, the final binary barcode output remained highly stable across all three analyses (Fig. 5A).

For the A1/N3KO-like subtype, the discovery-derived and validation-derived barcodes shared 3 of 4 genes (Jaccard = 0.75), whereas the pooled barcode fully matched the discovery-derived barcode and retained the same 4-gene output (Fig. 5B). For the B1/BCAS-like subtype, the discovery-derived and validation-derived barcodes showed high overlap (Jaccard = 0.80), and both remained highly concordant with the pooled barcode, which expanded modestly to a 10-gene output (Fig. 5B). Gene-level comparison showed that these differences were restricted to a small number of edge genes rather than broad restructuring of the barcode layer. In A1, the discrepancy was limited to transient loss of *Sfr1* in the validation run, whereas in B1 the discovery-versus-validation difference reflected boundary-level exchange between *Tab2* and *Plekhg1*, with the pooled analysis mainly absorbing both into a slightly expanded barcode (Fig. 5C,D). Thus, barcode perturbation under independent GWAS input was confined to a few boundary genes, while the core subtype structure remained preserved across discovery, validation, and pooled analyses.

Benchmarking against simpler baselines supports a favorable compactness – stability regime

We next asked whether the stability of the integrated subtype barcodes could be reproduced by simpler alternatives or instead reflected a specific advantage of the layered fusion framework. Across GWAS-only, scRNA-only, coloc-only, MR-only, and naive set-operation baselines, the final cSVI barcodes occupied a favorable compactness-stability regime rather than achieving reproducibility through

candidate-set expansion (Fig. S2A). The integrated outputs remained small (A1=4 genes; B1=9 genes) while preserving strong discovery-to-validation concordance, whereas naive union baselines achieved similarly high or higher Jaccard values only by expanding to much larger candidate sets of roughly 50 – 60 genes. By contrast, MR-only outputs were empty under the current v1.0 setting, and coloc-only baselines showed only intermediate performance, indicating that neither sparse causal filtering nor large-set aggregation alone could recover the observed balance of compactness and stability.

A particularly informative negative result came from the direct GWAS-scRNA overlap analysis. Under the current data setting, the strict intersection baseline was empty, and the scRNA top50 candidate set showed essentially no overlap with either the GWAS-backed universe or the final barcode layer (Fig. S2A,B). Instead, the final barcodes retained partial but selective overlap with the GWAS universe, consistent with a constrained genetic prior rather than a simple transcriptomic ranking rule (Fig. S2B). At the gene-evidence level, PP4 values for barcode genes followed an approximately diagonal discovery-versus-validation pattern overall, with discordant points restricted to boundary genes such as Sfr1 in A1 and Tab2 in B1 (Fig. S2C). Together, these comparisons indicate that the final cSVI barcodes are not merely large stable gene sets or direct GWAS-scRNA intersections, but compact outputs produced by layered fusion of human genetic anchoring and vascular-centered scRNA refinement.

Internal scRNA replication provided asymmetric within-study support for subtype barcode behavior

To further test whether the frozen subtype barcodes could reproduce the expected within-study behavior beyond the primary A1-versus-B1 discovery contrast, we performed internal scRNA replication using sample-level vascular module scores within each original study. The A1/N3KO-like barcode was evaluated along the *Notch3* axis by comparing A1 (N3KO) with A2 (matched controls), whereas the B1/BCAS-like barcode was evaluated along the BCAS axis by comparing B1 (BCAS) with B2 (sham controls). In parallel, cross-negative-control analyses were performed by applying the A1 barcode to the BCAS axis and the B1 barcode to the Notch3 axis (Fig. 6A–D).

Under this validation scheme, the B1/BCAS-like barcode showed clearer internal support than the A1/N3KO-like barcode. On the BCAS axis, the B1 barcode score was higher in B1 than in B2, supporting directional consistency of the BCAS-like program despite limited sample size (Fig. 6B). By contrast, on the Notch3 axis, the A1 barcode score showed only minimal separation between A1 and A2, indicating weaker internal support for the N3KO-like program under the current data setting (Fig. 6A). Thus, within-study replication was present but asymmetric, with clearer support for the B1/BCAS-like barcode than for the A1/N3KO-like barcode.

The cross-negative controls further argued against a fully nonspecific vascular response. In the BCAS dataset, application of the A1 barcode did not reproduce the

matched-axis separation observed for the B1 barcode and showed only weak discrimination between B1 and B2 (Fig. 6C). Conversely, when the B1 barcode was applied to the *Notch3* axis, only minimal separation was observed between A1 and A2 (Fig. 6D). Together, these findings support partial subtype specificity within the vascular compartment while indicating unequal internal replication strength across the two subtype programs.

Additional stratification by vascular subcompartment did not materially strengthen the A1/N3KO-like signal, whereas the B1/BCAS-like program remained directionally consistent in both endothelial and mural cells and showed clearer separation in endothelial cells (Fig. S1).

Single-chain ablation perturbed upstream scRNA ranking but left the final binary barcode unchanged

To assess whether the final subtype barcodes depended critically on any single scRNA evidence chain, we performed single-chain ablation analyses by removing CSS, CIS, or LRS while keeping the remainder of the framework unchanged. We then compared each ablation setting against the full model at two levels: the upstream scRNA feature-ranking layer and the final binary barcode layer. This design allowed us to distinguish sensitivity of intermediate ranking from robustness of the final subtype definition.

At the ranking layer, single-chain ablation substantially perturbed the Top50 scRNA feature set, with the strongest effect observed after removal of LRS (Fig. 7A). For the A1 branch, the Top50 Jaccard relative to the full model was 0.695 under both noCSS and noCIS, but dropped to 0.020 under noLRS. For the B1 branch, the corresponding Jaccard values were 0.515, 0.333, and 0.064, respectively (Fig. 7A). These results indicate that the upstream scRNA ranking layer remained responsive to removal of individual evidence chains, particularly the ligand – receptor-to-gene allocation component.

By contrast, the final binary barcode output was completely invariant under all three ablation settings. For both A1 and B1, the binary-footprint Jaccard relative to the full model remained 1.000 under noCSS, noCIS, and noLRS (Fig. 7B), and the final footprint sizes were unchanged at 4 genes for A1 and 9 genes for B1 across all ablation settings (Fig. 7C). Thus, although single-chain ablation altered the ordering of upstream scRNA candidates, it did not alter the final subtype barcodes. Together, these findings indicate that cSVI-subtype is sensitive at the intermediate ranking layer yet robust at the final barcode layer, supporting the stability of the binary subtype definition against removal of any single scRNA evidence chain.

Discussion

Current cSVD stratification is still dominated by clinical presentation, neuroimaging burden, and pathological subtype, whereas operational molecular subtype definition remains underdeveloped(Dupre, et al., 2024; Jacob, et al., 2023; Staals, et al., 2014; Wardlaw, et al., 2024). Here, we developed cSVI-subtype as a GWAS-anchored and vascular-centered framework to address this gap by converting broad multi-layer evidence into compact subtype barcodes rather than another extended list of candidate genes. Conceptually, the framework bridges two levels of disease representation that are usually analyzed separately: human large-cohort genetic support and disease-state cellular organization in single-cell models. In this sense, cSVI-subtype is not intended to replace clinical or imaging classification, but to provide a complementary molecular endpoint that can be stress-tested across perturbation settings and used as a starting point for downstream mechanistic refinement.

A key motivation for the framework is that neither statistical genetics nor scRNA-based disease modeling alone naturally yields a compact and robust subtype barcode. Post-GWAS prioritization, colocalization, and MR are powerful for nominating loci, genes, and putative causal relationships, but their outputs remain fundamentally association- or prioritization-oriented and depend on assumptions such as shared causal-variant structure in colocalization and valid instrument structure in MR(Sanderson, et al., 2022; Sun, et al., 2025). Conversely, scRNA data provide

high-resolution cellular context, but they do not directly resolve the gap between transcriptional responsiveness and disease relevance, particularly in cross-species disease models in which dropout, batch effects, dissociation bias, sampling imbalance, and model-specific biology can distort apparent signal strength(Gollihue, et al., 2025). More broadly, animal single-cell models and human large-cohort genetics capture different layers of disease biology: one reflects mechanistic cellular states under a defined perturbation, whereas the other reflects population-level association architecture across heterogeneous human backgrounds(Deng, et al., 2025). cSVI-subtype was developed precisely because these evidence streams are complementary but non-equivalent, and because subtype barcode discovery requires a filtered endpoint rather than separate lists of genetically supported genes and transcriptionally responsive genes.

Recent large-cohort and integrative studies have substantially advanced the field by identifying cSVD-associated loci, MRI-marker genetics, shared vascular risk architecture, and disease-relevant cell types or molecular programs(Duperron, et al., 2023; Sargurupremraj, et al., 2020). More broadly, current analytical frontiers increasingly combine GWAS with QTL resources, transcriptomic atlases, and single-cell expression specificity to move from variant discovery toward gene and cell-type prioritization(Sargurupremraj, et al., 2024). However, even these frontier approaches still tend to output associated loci, prioritized genes, enriched cell types, or candidate pathways rather than compact subtype barcodes that can be compared

directly across perturbation settings. In other words, current large-cohort analysis is becoming increasingly effective at identifying which signals matter, but remains less suited to determining how multiple evidence layers should be compressed into an operational subtype-defining endpoint(Yang, et al., 2024). This distinction is especially important for cSVD, where biological heterogeneity is substantial but molecular subtype standards remain immature.

This context also clarifies how cSVI-subtype differs from existing tool classes. Tools such as FUMA are highly effective for annotation and gene prioritization from GWAS loci, but their natural endpoint remains biologically interpreted candidate genes rather than disease subtype barcodes(Watanabe, et al., 2017). Methods such as MAGMA, scDRS, and related GWAS-single-cell integration frameworks are well suited for prioritizing trait-relevant cell types, cell states, or polygenic enrichment patterns, yet they are not primarily designed to produce compact gene-level subtype outputs that remain stable under independent genetic perturbation(de Leeuw, et al., 2015; Zhang, et al., 2022). Likewise, communication-oriented tools such as CellChat and NicheNet are powerful for reconstructing ligand-receptor or ligand-target architecture, but they optimize communication inference rather than a genetics-anchored subtype-definition endpoint(Browaeys, et al., 2020; Dimitrov, et al., 2022; Jin, et al., 2021). From this perspective, the main advantage of cSVI-subtype is not that it universally outperforms these tools on their own tasks, but that it reorganizes outputs from gene prioritization, cell-state relevance, and communication

evidence toward a different objective: compact subtype-resolved barcode discovery under dual human-genetic and vascular-mechanistic constraints. The trade-off is that the framework becomes more conservative, more task-specific, and more sensitive to the design choices used to compress layered evidence into a final barcode.

Within this design, an important conceptual feature of cSVI-subtype is that robustness operates differently across analytical layers. The upstream scRNA ranking layer remained sensitive to perturbation, as shown by substantial Top50 changes after single-chain ablation, particularly after removal of the ligand-receptor allocation component (Fig. 7A). By contrast, the final binary barcode layer remained invariant across all tested ablations and highly stable under external GWAS perturbation (Figs. 5B-D, 7B,C). This distinction suggests that the framework is not rigid throughout the full pipeline; rather, it allows flexibility at the intermediate evidence-ranking stage while preserving stability at the final subtype-definition stage. We therefore interpret the current v1.0 implementation as deliberately compressive at the endpoint: many upstream details may shift, but the final barcode remains small and stable unless multiple evidence layers move concordantly.

The internal scRNA evaluation also requires a nuanced interpretation. The B1/BCAS-like barcode showed clearer within-study directional support than the A1/N3KO-like barcode, whereas the A1 branch showed only limited same-direction separation on the Notch3 axis (Fig. 6A,B). We do not interpret this asymmetry as a failure of the framework, but neither should it be overstated as equally strong internal

replication for both branches. The most defensible interpretation is that the current data provide stronger internal support for the BCAS-like program and weaker but still directionally compatible support for the N3KO-like program. At the same time, the cross-negative controls were informative: the A1 barcode did not reproduce the matched-axis separation observed for the B1 program on the BCAS axis, arguing against a fully nonspecific vascular-stress interpretation (Fig. 6C,D). Thus, the present evidence supports subtype specificity, but with unequal internal replication strength across branches. In practical terms, the B1 program is currently the better-supported within-study phenotype, whereas the A1 program should be regarded as stable at the final barcode level but still somewhat provisional in its internal biological expression pattern. Additional stratification by vascular subcompartment did not materially strengthen the A1/N3KO-like signal, whereas the B1/BCAS-like program remained directionally consistent in both endothelial and mural cells and showed clearer separation in endothelial cells (Fig. S1).

Several limitations of the current v1.0 framework should therefore be made explicit. First, the present implementation is based on two mouse disease-model datasets and does not yet span the broader phenotypic spectrum of cSVD. Second, the scRNA layer was intentionally frozen at the major-cell-class and vascular-centered level to favor robustness, but this necessarily sacrifices finer vascular substate resolution, bidirectional multicellular feedback, and temporal disease transitions. Third, the framework uses colocalization and MR as supportive filtering layers rather

than definitive causal adjudication, and cross-species projection through ortholog mapping inevitably excludes some biologically relevant genes. Fourth, the final binary barcode was more strongly supported for the B1/BCAS-like branch than for the A1/N3KO-like branch in internal scRNA evaluation, indicating unequal branch-level confidence despite final barcode stability (Fig. 6A,B). Finally, the strict multi-zone formulation remained empty in the present setting, meaning that the reported binary-core barcode should be regarded as an operational v1.0 output rather than a definitive molecular atlas (Fig. 4D). Future work should therefore extend the framework toward finer vascular-state stratification, additional disease models, longitudinal or trajectory-aware subtype transitions, and validation in human single-nucleus and spatial datasets.

Taken together, these results support cSVI-subtype as a proof-of-principle framework for operational molecular subtype definition in cSVD. Rather than relying on direct GWAS-scRNA overlap or candidate-set expansion, the framework uses layered filtering to generate compact subtype-resolved outputs that remain stable under multiple perturbation settings (Figs. 5, 7; Fig. S2). Although the current implementation remains deliberately conservative and incomplete, it provides a tractable starting point for subtype-oriented molecular mapping in cSVD and suggests a general strategy for linking human disease genetics to vascular-centered cellular-state modeling in complex cerebrovascular disorders.

Conclusion

In conclusion, we developed cSVI-subtype as a GWAS-anchored and vascular-centered framework for molecular subtype barcode discovery in cerebral small vessel disease. By integrating colocalization-supported human genetic signals with layered vascular scRNA evidence, the framework compresses broad candidate spaces into compact, subtype-resolved binary barcodes. In the current v1.0 setting, cSVI-subtype identified distinct A1/N3KO-like and B1/BCAS-like barcodes that remained stable under independent GWAS perturbation and single-chain ablation, while showing asymmetric but directionally supportive behavior in internal scRNA evaluation. Together, these results establish cSVI-subtype as a proof-of-principle framework for moving cSVD stratification beyond candidate-gene prioritization toward operational molecular subtype definition. Further validation in additional models and human single-nucleus or spatial datasets will strengthen this framework, but the current implementation already provides a compact and testable starting point for subtype-oriented molecular mapping in cSVD.

Acknowledgements

None.

Author contributions

Y.L. conceived and designed the study, developed the methodology, performed the analyses, generated the figures, and wrote the manuscript. **Y.L.** approved the final version of the manuscript.

Funding

None

Data Availability

All data analyzed in this study are publicly available. Human GWAS summary statistics were obtained from HuGeAMP, and the mouse single-cell RNA-seq datasets were obtained from GEO under accession numbers GSE204803 and GSE263191. The csviSubtype R package, shipped example files, and reproducibility scripts are publicly available at https://github.com/YuqianLii/csviSubtype. Additional processed intermediate outputs used in the current manuscript are available from the corresponding author upon reasonable request.

Declarations

Consent to participate

Not applicable.

Consent for publication

Not applicable.

Conflict of Interest

The authors declare that they have no competing interests.

References

Browaeys, R., Saelens, W. and Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat Methods* 2020;17(2):159-162.

Chen, J.*, et al.* Dysregulation of Principal Circulating miRNAs in Non-human Primates Following Ischemic Stroke. *Front Neurosci* 2021;15:738576.
de Leeuw, C.A.*, et al.* MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput Biol* 2015;11(4):e1004219.
Deng, H.*, et al.* Integrating scRNA-seq and GWAS data reveals potentially critical endothelial cells in large artery atherosclerotic stroke. *Front Neurosci* 2025;19:1646993.
Dimitrov, D.*, et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nat Commun* 2022;13(1):3224.
Duering, M.*, et al.* Neuroimaging standards for research into small vessel disease-advances since 2013. *Lancet Neurol* 2023;22(7):602-618.
Duperron, M.G.*, et al.* Genomics of perivascular space burden unravels early mechanisms of cerebral small vessel disease. *Nat Med* 2023;29(4):950-962.
Dupre, N., Drieu, A. and Joutel, A. Pathophysiology of cerebral small vessel disease: a journey through recent discoveries. *J Clin Invest* 2024;134(10).
Gollihue, J.L.*, et al.* Inhibition of astrocyte signaling leads to sex-specific changes in microglia phenotypes in a diet-based model of cerebral small vessel disease. *J Neuroinflammation* 2025;22(1):202.
Hilkens, N.A.*, et al.* Stroke. *Lancet* 2024;403(10446):2820-2836.
Hong, H., Tozer, D.J. and Markus, H.S. Relationship of Perivascular Space Markers With Incident Dementia in Cerebral Small Vessel Disease. *Stroke* 2024;55(4):1032-1040.
Jacob, M.A.*, et al.* Cerebral Small Vessel Disease Progression and the Risk of Dementia: A 14-Year Follow-Up Study. *Am J Psychiatry* 2023;180(7):508-518.
Jin, S.*, et al.* Inference and analysis of cell-cell communication using CellChat. *Nat Commun* 2021;12(1):1088.
Le Grand, Q.*, et al.* Diffusion imaging genomics provides novel insight into early mechanisms of cerebral small vessel disease. *Mol Psychiatry* 2024;29(11):3567-3579.
Li, X.*, et al.* Molecular Mediators of Neutrophil Primary Granule Release Following Acute Ischemic Stroke and their Associated Epigenetic Modulation by HDAC2. *Mol Neurobiol* 2025;62(5):6544-6561.
Li, Y.*, et al.* The High-Affinity IL-2 Receptor Affects White Matter Damage after Cerebral Ischemia by Regulating CD8 + T Lymphocyte Differentiation. *J Neuroimmune Pharmacol* 2025;20(1):8.
Li, Y.*, et al.* The IL-2A receptor pathway and its role in lymphocyte differentiation and function. *Cytokine Growth Factor Rev* 2022;67:66-79.
Liu, P.*, et al.* Cerebrovascular reactivity MRI as a biomarker for cerebral small vessel disease-related cognitive decline: Multi-site validation in the MarkVCID Consortium. *Alzheimers Dement* 2024;20(8):5281-5289.
Ma, Y.*, et al.* Polygenic regression uncovers trait-relevant cellular contexts through pathway activation transformation of single-cell RNA sequencing data. *Cell Genom* 2023;3(9):100383.

Markus, H.S. and Joutel, A. The pathogenesis of cerebral small vessel disease and vascular cognitive impairment. *Physiol Rev* 2025;105(3):1075-1171.
Pantoni, L. Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *Lancet Neurol* 2010;9(7):689-701.
Sanderson, E.*, et al.* Mendelian randomization. *Nat Rev Methods Primers* 2022;2.
Sargurupremraj, M.*, et al.* Genetic Complexities of Cerebral Small Vessel Disease, Blood Pressure, and Dementia. *JAMA Netw Open* 2024;7(5):e2412824.
Sargurupremraj, M.*, et al.* Cerebral small vessel disease genomics and its implications across the lifespan. *Nat Commun* 2020;11(1):6285.
Staals, J.*, et al.* Stroke subtype, vascular risk factors, and total MRI brain small-vessel disease burden. *Neurology* 2014;83(14):1228-1234.
Sun, Z.*, et al.* Proteins Involved in Endothelial Function and Inflammation Are Implicated in Cerebral Small Vessel Disease. *Stroke* 2025;56(3):692-704.
Townsend, H.A.*, et al.* Evaluating methods for integrating single-cell data and genetics to understand inflammatory disease complexity. *Front Immunol* 2024;15:1454263.
Traylor, M.*, et al.* Genetic basis of lacunar stroke: a pooled analysis of individual patient data and genome-wide association studies. *Lancet Neurol* 2021;20(5):351-361.
Wardlaw, J.M.*, et al.* European stroke organisation (ESO) guideline on cerebral small vessel disease, part 2, lacunar ischaemic stroke. *Eur Stroke J* 2024;9(1):5-68.
Watanabe, K.*, et al.* Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* 2017;8(1):1826.
Yang, X.Z.*, et al.* Genome-Wide Mendelian Randomization Study Reveals Druggable Genes for Cerebral Small Vessel Disease. *Stroke* 2024;55(9):2264-2273.
Zhang, M.J.*, et al.* Polygenic enrichment distinguishes disease associations of individual cells in single-cell RNA-seq data. *Nat Genet* 2022;54(10):1572-1580.

Figure Legends

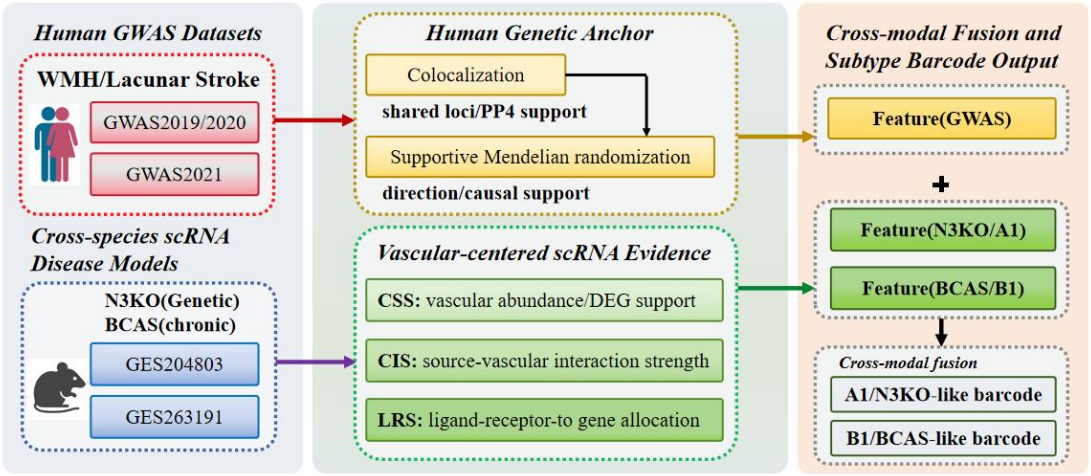

Fig. 1. Overview of the cSVI-subtype framework.

Schematic overview of cSVI-subtype, a GWAS-anchored and vascular-centered framework for molecular subtype barcode discovery in cerebral small vessel disease.

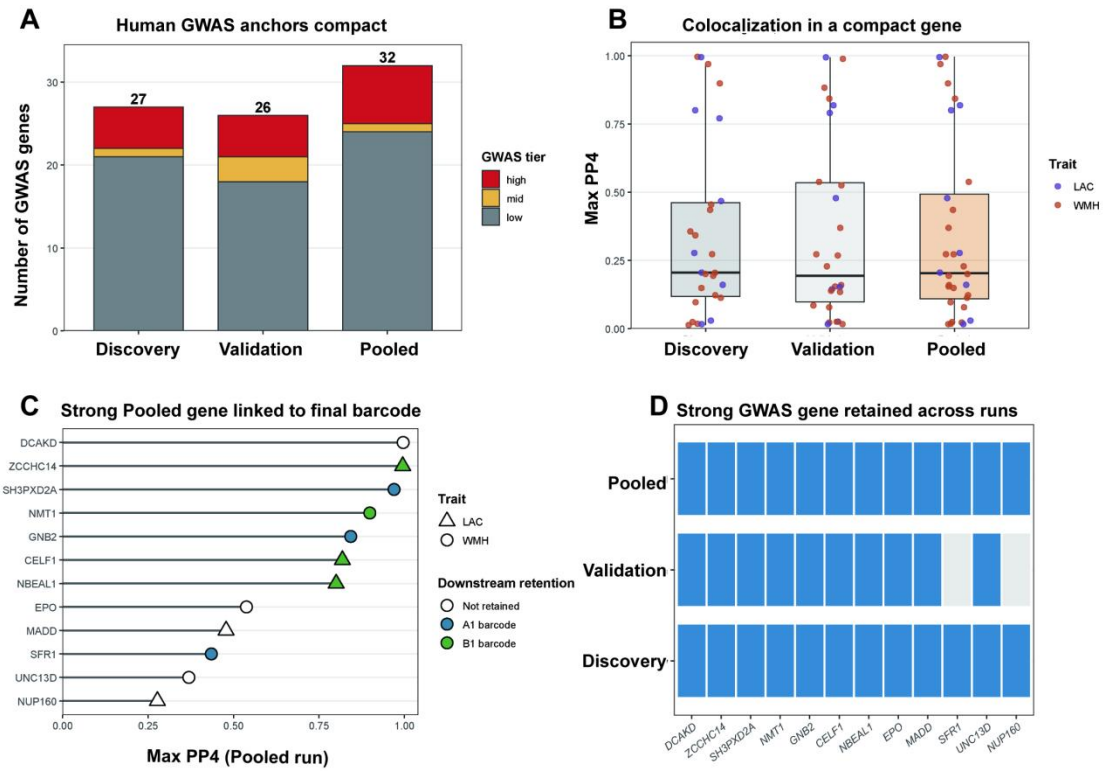

Fig. 2. Human genetic anchor construction produced a compact and stable GWAS-backed candidate universe.

(A) Total size of the GWAS-backed vascular candidate universe across discovery, validation, and pooled analyses, showing that the anchor layer remained compact across runs.

(B) Distribution of gene-level colocalization support (PP4) across runs, illustrating that strong shared-causal support was concentrated in a limited subset of genes rather than broadly distributed across the full anchor space.

(C) Top pooled-anchor genes ranked by maximum PP4, with point fill indicating whether each gene was retained in the final discovery-derived subtype barcodes.

(D) Run-level retention of top anchor genes across discovery, validation, and pooled summaries, showing that most of the strongest GWAS-supported genes were

preserved across independent analyses. Together, these panels show that the human genetic layer converged to a small, cross-run stable candidate universe that directly seeded downstream subtype-barcode construction.

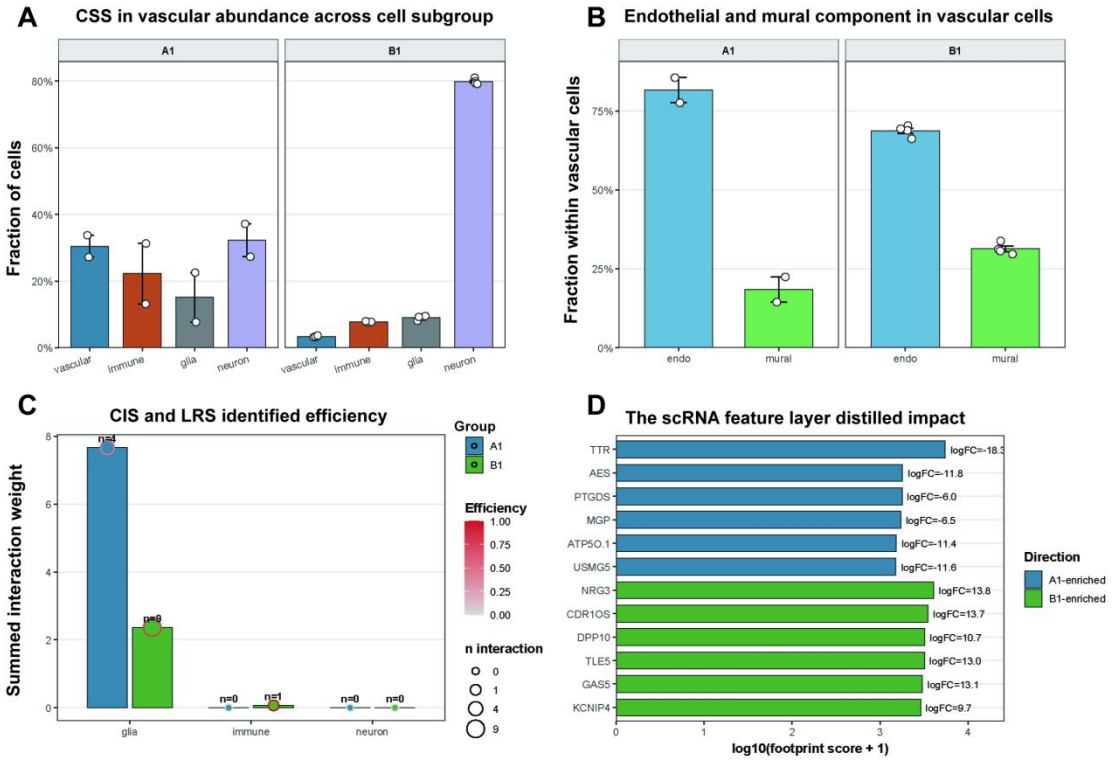

Fig. 3. Vascular-centered scRNA evidence was built through CSS, CIS, LRS, and feature ranking.

(A) CSS highlighted marked differences in vascular abundance across the two subtype branches, with a substantially higher vascular fraction in the A1/N3KO-like branch than in the B1/BCAS-like branch.

(B) Decomposition of the vascular gate into endothelial and mural compartments showed that both branches were predominantly endothelial, while the B1/BCAS-like branch contained a relatively larger mural component.

(C) CIS and LRS summarized sender-to-vascular communication burden and relay efficiency. Glial input dominated the source-to-vascular interaction structure in both branches; B1 additionally showed immune-to-vascular input and a broader but weaker communication architecture.

(D) The scRNA feature layer distilled compact subtype-resolved vascular feature genes by combining vascular differential-expression magnitude with relay-derived support, thereby generating the downstream scRNA evidence layer used for fusion rather than the final barcode itself.

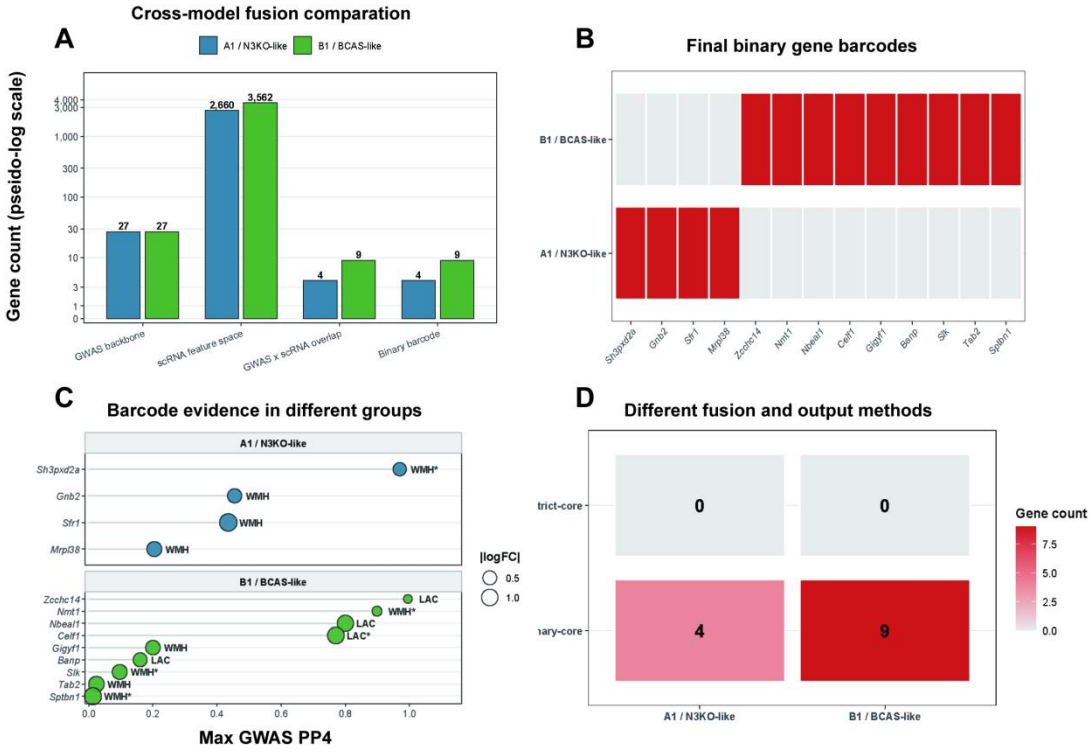

Fig. 4. Cross-modal fusion distilled compact discovery-only subtype barcodes.

(A) Compression of candidate space from the GWAS vascular backbone and broad subtype-specific scRNA candidate spaces into compact final barcode outputs.

(B) Final binary barcode membership for the A1/N3KO-like and B1/BCAS-like states under the operational v1.0 binary-core rule.

(C) Barcode evidence profile for the retained subtype genes, linking final barcode membership to upstream GWAS support.

(D) Comparison between the strict multi-zone formulation and the binary-core operational output. In the discovery setting, fusion converged to a 4-gene A1/N3KO-like barcode and a 9-gene B1/BCAS-like barcode with no overlap between branches, whereas the stricter multi-zone core remained empty. These results support the binary-core formulation as the primary reported subtype output in v1.0.

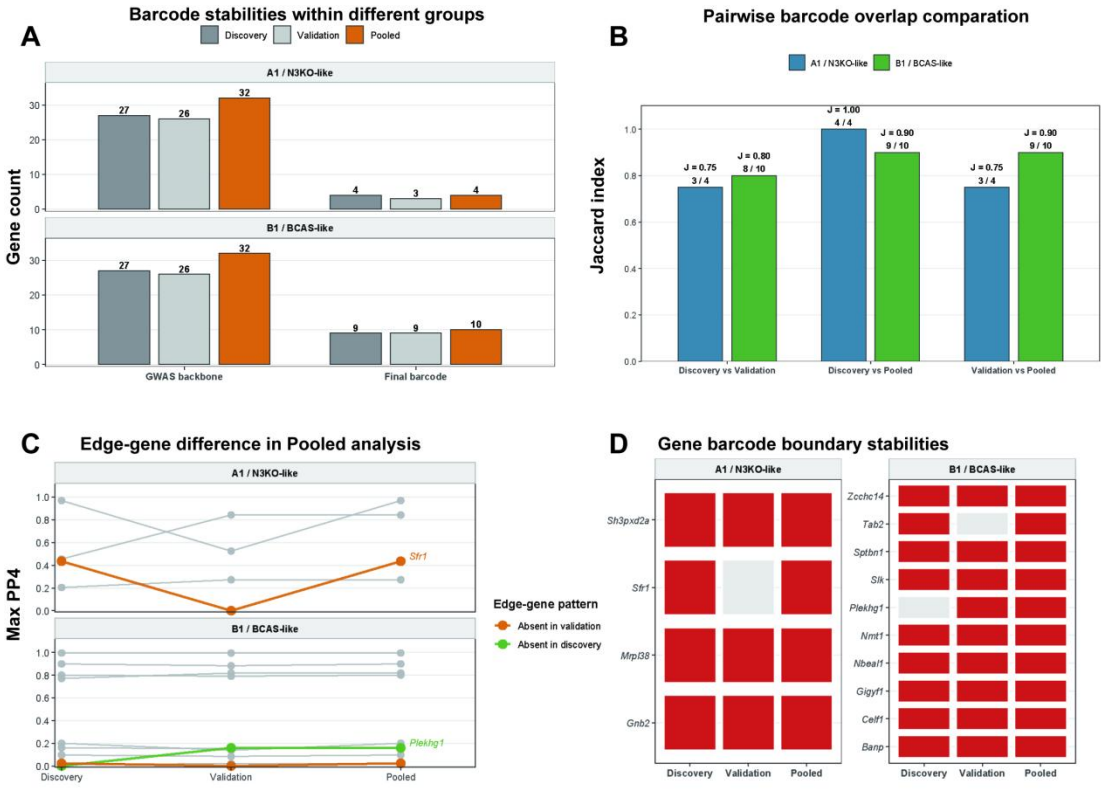

Fig. 5. Independent validation and pooled GWAS perturbation support barcode stability.

(A) Sizes of the GWAS backbone and final barcode outputs across discovery, validation, and pooled analyses. The GWAS anchor layer showed only modest run-to-run change, and final barcode size remained compact.

(B) Pairwise barcode overlap across runs, showing high concordance between discovery-derived, validation-derived, and pooled outputs for both A1/N3KO-like and B1/BCAS-like branches.

(C) Gene-level pooled analysis illustrating that pooled GWAS support mainly absorbed boundary-level differences rather than reshaping the core barcode structure.

(D) Heatmap of barcode membership across runs, showing that barcode differences were restricted to a small number of edge genes. Together, these panels indicate that the final binary barcode layer remained stable under independent changes in the human genetic anchor.

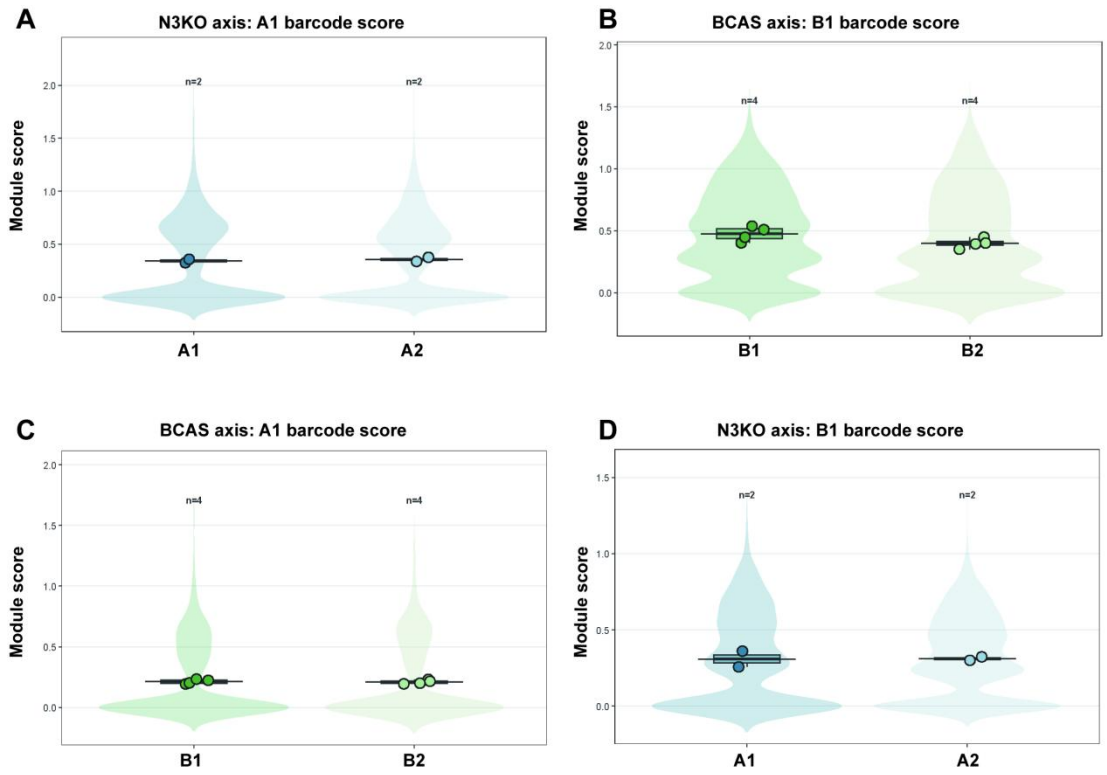

Fig. 6. Internal scRNA replication evaluated subtype barcode behavior across matched disease axes.

Sample-level vascular module scores are shown as primary points and summaries, with cell-level score distributions displayed in the background.

(A) A1/N3KO-like barcode behavior on the matched Notch3 axis (A1 versus A2).

(B) B1/BCAS-like barcode behavior on the matched BCAS axis (B1 versus B2).

(C) Cross-negative-control analysis applying the A1 barcode to the BCAS axis.

(D) Cross-negative-control analysis applying the B1 barcode to the Notch3 axis. The B1/BCAS-like barcode showed clearer within-study support than the A1/N3KO-like barcode, whereas cross-negative controls did not show concordant elevation, supporting subtype specificity rather than a generic vascular-stress signal.

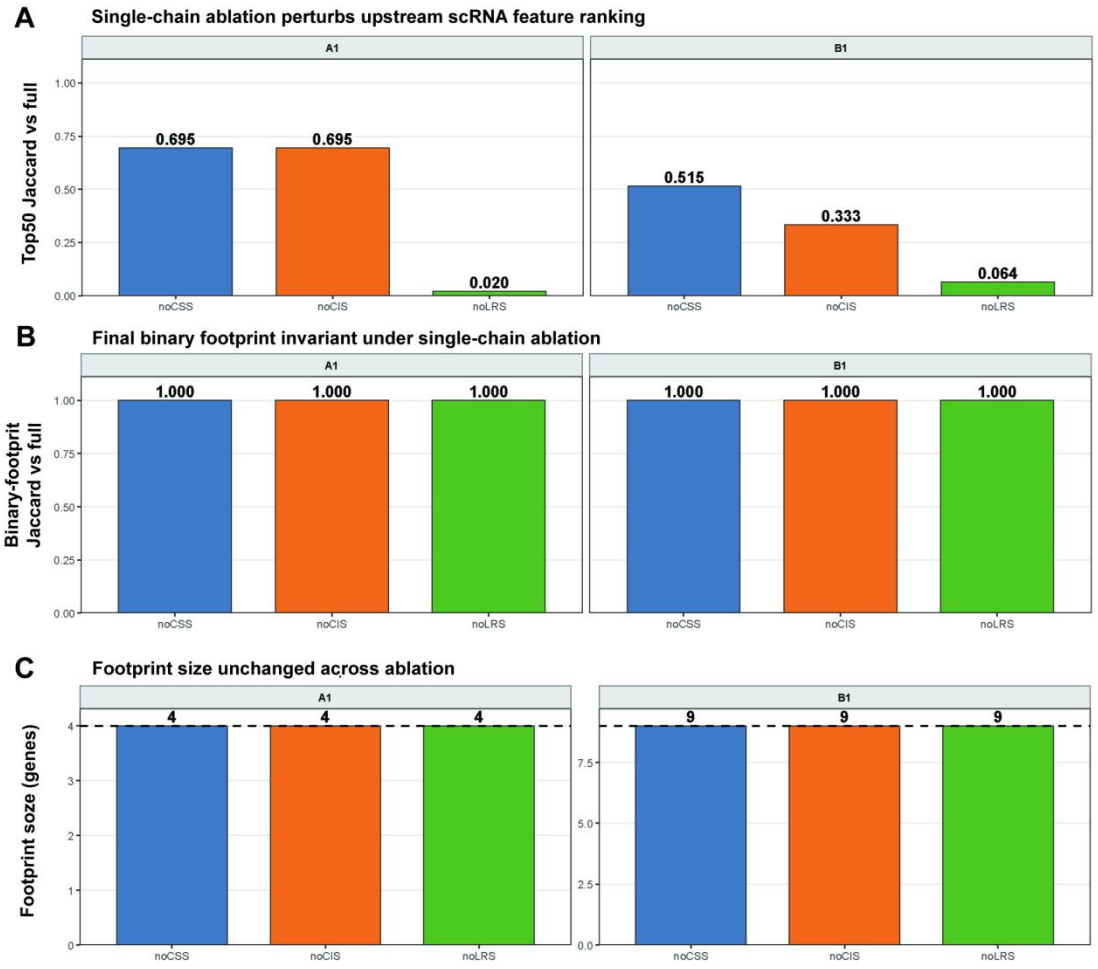

Fig. 7. Single-chain ablation perturbs upstream scRNA ranking but leaves the final barcode unchanged.

(A) Jaccard overlap of Top50 scRNA feature rankings after removal of CSS, CIS, or LRS, showing that single-chain ablation substantially perturbed the intermediate ranking layer, with the strongest disruption observed after removal of LRS.

(B) Jaccard overlap of final binary barcodes under the same ablation settings, showing complete invariance of the final subtype barcode layer for both A1 and B1 branches.

(C) Final footprint size across ablation settings, demonstrating that the number of retained barcode genes remained unchanged relative to the full model. Together, these results show that cSVI-subtype is sensitive at the upstream ranking layer but robust at the final binary subtype-definition layer.

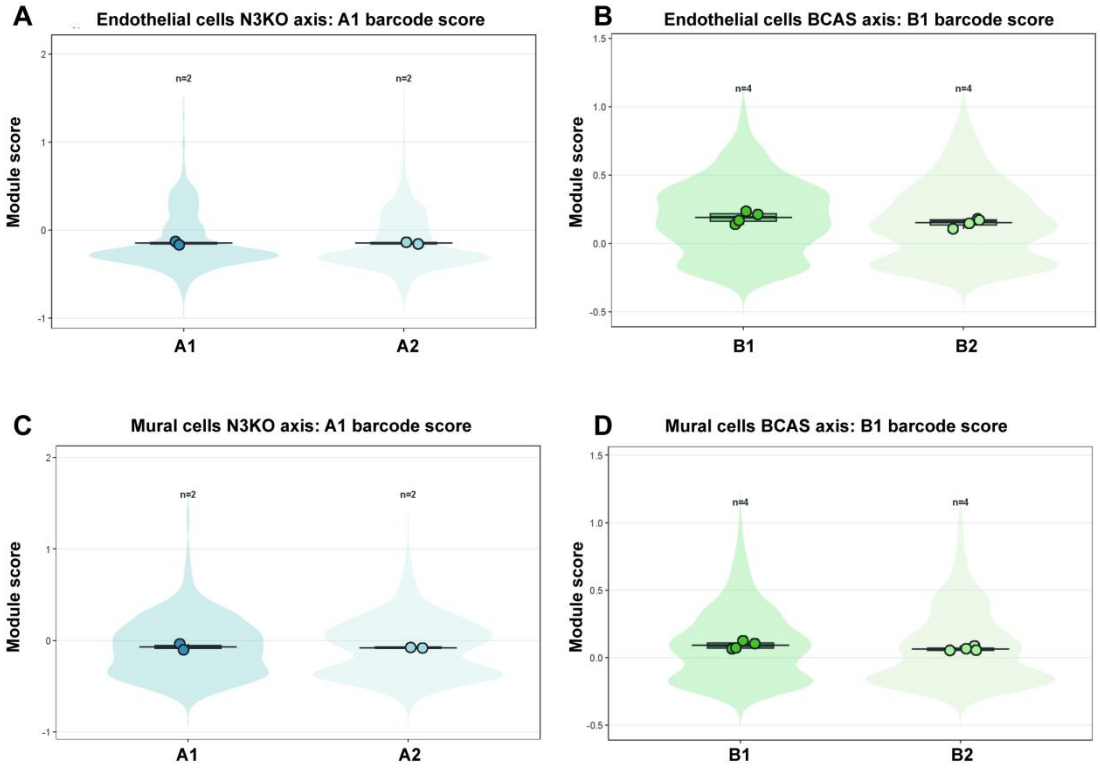

Fig. S1. Vascular-subcompartment stratification of internal barcode behavior.

(A) A1/N3KO-like barcode score within endothelial cells on the Notch3 axis.

(B) B1/BCAS-like barcode score within endothelial cells on the BCAS axis.

(C) A1/N3KO-like barcode score within mural cells on the Notch3 axis.

(D) B1/BCAS-like barcode score within mural cells on the BCAS axis. Subcompartment stratification did not materially strengthen the A1/N3KO-like internal signal, whereas the B1/BCAS-like program remained directionally consistent in both endothelial and mural compartments and was visually clearer in endothelial cells.

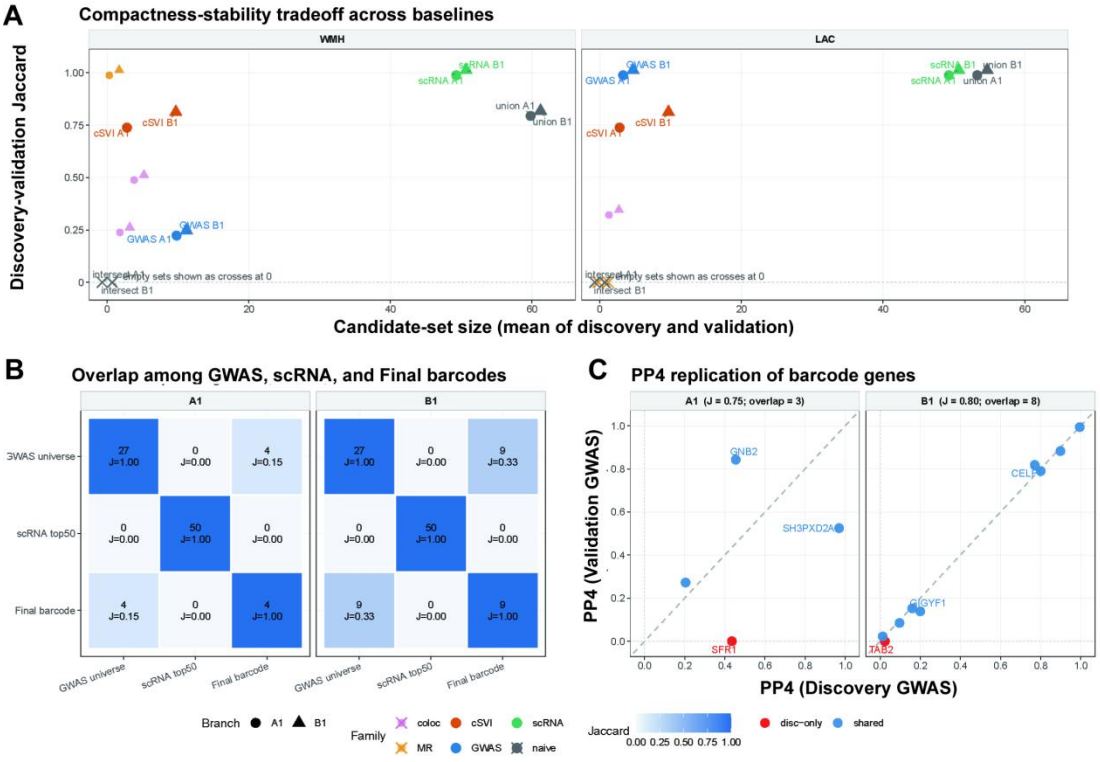

Fig. S2. Baseline and comparator benchmarking support compact and specific subtype barcodes.

(A) Compactness-stability tradeoff across baseline strategies, including GWAS-only, scRNA-only, coloc-only, MR-only, and naive set-operation baselines. The final cSVI

barcode occupied a favorable regime by remaining small while preserving strong discovery-to-validation stability; empty baselines are plotted as crosses at Jaccard = 0. (B) Pairwise overlap heatmaps among the GWAS-backed universe, scRNA Top50 candidate set, and final barcode for A1 and B1. The near-zero overlap between GWAS and scRNA-only layers shows that the final subtype barcodes could not be recovered by simple direct intersection.

(C) Discovery-versus-validation PP4 replication for genes in the barcode union, showing that most genes followed an approximately diagonal pattern, whereas discordant points corresponded mainly to edge-gene substitutions. Together, these analyses show that cSVI-subtype does not outperform simpler alternatives by producing a larger output, but by converging on a compact, genetically anchored, and subtype-specific barcode.