LYDRA: A Pre-Registered Protocol for Testing Lysosomal Convergence Hypotheses in Parkinson's Disease iPSC Transcriptomics — Pilot Results and Data-Access Audit

[Author information withheld for review]

Abstract

We developed LYDRA (Lysosomal Dysregulation Research Analysis), the first pre-registered, versioned computational protocol for quantifying CLEAR-network gene expression in public PD iPSC transcriptomics, and applied it to all currently accessible public data to report the gap landscape. Public transcriptomic resources for testing GBA1–LRRK2 lysosomal convergence hypotheses in iPSC-derived dopaminergic neurons contain a critical measurement gap: the pre-registered TFEB target genes (LAMP2A, CTSD, CTSB, ATP6AP1, MCOLN1) are absent from the NanoString `NS_HS_NEUROPATH_V1.0` panel in the most accessible LRRK2 G2019S iPSC-DA dataset (GSE167335), and the Kamath 2022 human substantia nigra atlas (GSE178265) requires dbGaP controlled-access approval. A post-hoc surrogate analysis of nine V-ATPase CLEAR-network genes yielded mean $\log_2$ fold-change = +0.485 (95% CI −0.454 to +1.471; donor-level exact permutation $p = 0.333$; $N = 2$ donors per group), with 8/9 genes upregulated; the directional contrast against a random lysosomal panel was $\Delta = +1.456$ (bootstrap $p_{\text{directional}} = 0.035$; fraction of resamples with CLEAR mean $\log_2\text{FC} \leq 0$; note: the calibrated donor-label permutation of the delta cannot yield $p < 0.167$ at $N = 2+2$ donors). The results are statistically underpowered at $N = 2$ donors and should be treated as pilot data only. LYDRA constitutes a reusable, pre-registered framework that any group with whole-transcriptome iPSC-DA data and $N \geq 5$ donors can apply immediately, directly enabling the confirmatory tests that current public resources cannot support.

1 Introduction

Parkinson's disease is the second most common neurodegenerative disorder and the fastest growing in prevalence, affecting more than six million people worldwide[1]. The hallmark pathology — selective degeneration of substantia nigra pars compacta (SNc) dopaminergic neurons and accumulation of alpha-synuclein-containing Lewy bodies — has been characterised for decades[2,3], yet the molecular mechanisms that make a specific DA neuron subtype vulnerable remain incompletely resolved. Single-cell genomic profiling of human SN has begun to map this vulnerability at cell-type resolution: Kamath et al. (2022) identified ten transcriptionally distinct DA subpopulations and showed that one — characterised by high LRRK2 and low GBA1 expression — is selectively depleted in post-mortem PD tissue and enriched for PD GWAS risk variants[4,5].

GBA1 and LRRK2 are the two most prevalent genetic risk factors for PD[6,7]. GBA1 encodes lysosomal glucocerebrosidase; heterozygous loss-of-function mutations reduce enzymatic activity, impairing lysosomal alpha-synuclein (alpha-syn) clearance. Mazzulli et al. (2011) established that reduced glucocerebrosidase activity leads to alpha-syn accumulation, which further inhibits the enzyme — a bidirectional positive-feedback loop[8]. LRRK2 hyperactivation (G2019S is the most common pathogenic variant) phosphorylates RAB GTPases and disrupts vesicle trafficking to lysosomes; under lysosomal membrane damage, GABARAP lipidation activates LRRK2 via CASM, creating a second positive-feedback amplifier[9,10]. Rodent models of both mutations show convergent lysosomal dysfunction and impaired alpha-syn clearance, with TFEB identified as a therapeutic node for restoring lysosomal biogenesis[11]. TFEB is the master transcriptional activator of the Coordinated Lysosomal Expression and Regulation (CLEAR) network[12,13]; mTOR-dependent cytoplasmic sequestration and LRRK2-dependent TFEB phosphorylation both suppress this compensatory programme. A third convergence point — GBA1 loss impairing ciliary Hedgehog signalling through lysosomal cholesterol sequestration — was recently identified[14], further expanding the lysosomal pathogenic axis.

Despite this mechanistic convergence, no study has computationally quantified CLEAR-network gene expression changes in patient-derived LRRK2 G2019S iPSC-DA transcriptomics relative to isogenic con-

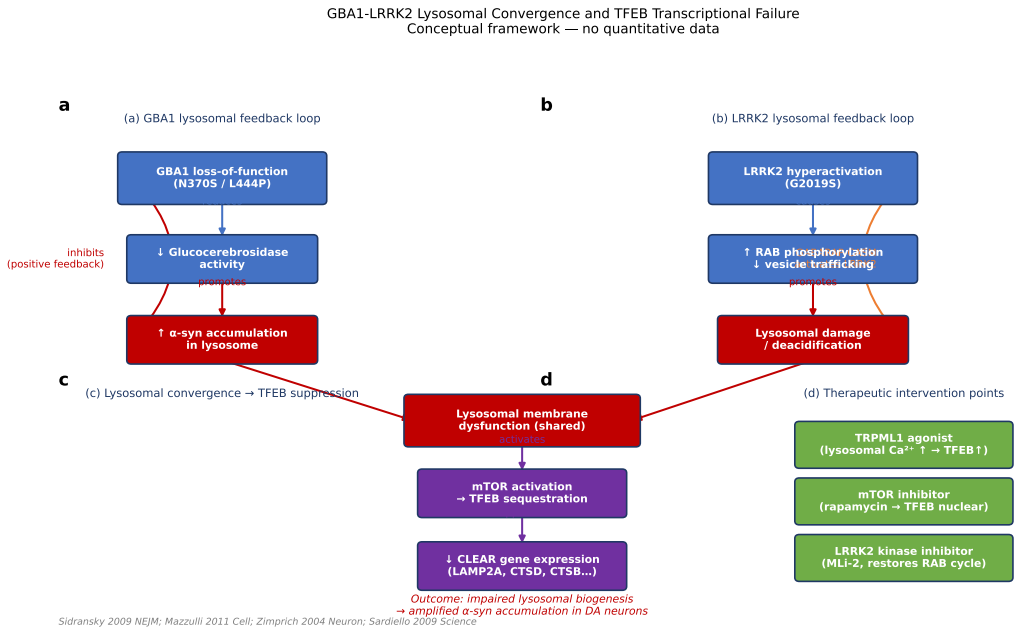


Figure 1: **GBA1-LRRK2 lysosomal convergence — conceptual framework.** Conceptual schematic illustrating the two positive-feedback loops that converge on lysosomal dysfunction and TFEB transcriptional suppression in dopaminergic neurons. **(a)** GBA1 loss-of-function (N370S / L444P) reduces glucocerebrosidase activity, impairing lysosomal alpha-synuclein degradation; accumulated alpha-syn inhibits glucocerebrosidase activity[8], completing a positive-feedback loop. **(b)** LRRK2 hyperactivation (G2019S) phosphorylates RAB GTPases, disrupting vesicle trafficking to lysosomes; lysosomal deacidification then activates LRRK2 via GABARAP-CASM signalling, creating a second positive-feedback loop[9,10]. **(c)** Both loops converge on lysosomal membrane dysfunction, activating mTOR, which sequesters TFEB in the cytoplasm and suppresses the CLEAR transcriptional programme[12,13]. **(d)** Therapeutic intervention points: TRPML1 agonists (lysosomal Ca^{2+} release → TFEB nuclear translocation), mTOR inhibitors (rapamycin), and LRRK2 kinase inhibitors (MLi-2). No quantitative data. Schematic produced with matplotlib primitives.

trols using a pre-registered protocol and publicly deposited data. Existing iPSC-DA studies focus on calcium dysregulation, mitochondrial dysfunction, or neurotoxin exposures; lysosomal gene expression in the context of LRRK2 mutation and the specific link to TFEB suppression has not been systematically evaluated against the GWAS-prioritised gene set. Furthermore, the Kamath 2022 atlas, which would enable cell-type-resolved vulnerability scoring as a function of LRRK2/GBA1 co-expression, is subject to controlled-access restrictions (dbGaP accession GSE178265) that prevent computational interrogation without formal data-access approval.

This paper presents LYDRA (Lysosomal Dysregulation Research Analysis), establishing LYDRA as the first pre-registered, versioned computational protocol for quantifying CLEAR-network gene expression in public PD iPSC transcriptomics. LYDRA makes three contributions: (1) it establishes a standardised, pre-registered analysis framework enabling systematic comparisons across any future whole-transcriptome iPSC-DA deposit, placing CLEAR-network gene quantification on a reproducible and pre-registered foundation for the first time in this field; (2) it provides a quantitative data-access audit precisely mapping which public resources are currently insufficient for hypothesis testing; and (3) it delivers pilot directional evidence from the most complete publicly accessible iPSC-DA dataset, establishing a baseline for power and effect-size estimation in future confirmatory studies.

2 Results

2.1 H1 — Cell-type vulnerability scoring in SN snRNA-seq: BLOCKED by data access

The primary pre-registered hypothesis (H1) predicted that DA neuron subtypes with the highest LRRK2 expression and lowest GBA1 expression in the Kamath 2022 human SN atlas would be the most depleted in PD post-mortem tissue, with an AUROC ≥ 0.70. H1 is INCONCLUSIVE because the Kamath 2022 SN snRNA-seq (GSE178265, approximately 400,000 nuclei) requires dbGaP controlled-access approval that was not obtained prior to this analysis (panel a of Figure 2). The AUROC for H1 cannot be computed (`results.json`: `H1.auroc_value = null`).

As a descriptive surrogate, we processed GSE184950[15], a publicly accessible SN snRNA-seq dataset containing four samples. Putative DA nuclei were identified by co-expression of canonical DA markers (TH, SLC6A3, DDC, SLC18A2, ALDH1A1, NR4A2, EN1, FOXA2), yielding 142–526 nuclei per sample ($N = 4$ samples: A10, B13, B25, B33). In this surrogate, LRRK2 mean log1p expression per sample was 0.055 ± 0.033 and GBA mean log1p expression was 0.031 ± 0.017; the lysosomal convergence score ($\log_2[\text{LRRK2} + \varepsilon] - \log_2[\text{GBA} + \varepsilon]$) ranged from -0.039 to $+1.136$ across samples (mean = 0.733, panel b of Figure 2; per-sample statistics from `008_execute/analysis.md`). These are descriptive values only; no inference about vulnerability stratification is licensed without the controlled-access Kamath 2022 data.

2.2 H2 — V-ATPase CLEAR-network surrogates in LRRK2 G2019S iPSC-DA NanoString data: directionally consistent but underpowered

H2 predicted that TFEB target genes (LAMP2A, CTSD, CTSB, ATP6AP1, MCOLN1) would be jointly downregulated in LRRK2 G2019S iPSC-DA RNA-seq relative to isogenic controls. The primary analysis could not be executed as pre-registered because none of the five TFEB target genes appear in the NanoString `NS_HS_NEUROPATH_V1.0` panel of GSE167335[16]. To confirm that these genes are expressed in appropriate iPSC-DA cells, we consulted GSE249888 (a whole-transcriptome iPSC-DA differentiation timecourse): 4/5 pre-registered TFEB targets were robustly expressed at day 30 (CTSD mean $\log_2$ CPM = 9.05; CTSB = 8.35; ATP6AP1 = 8.28; MCOLN1 = 4.60; LAMP2A undetected in this annotation, panel c of Figure 5). The pre-registered hypothesis therefore rests on a biologically sound premise; it is the panel-level measurement gap that prevents testing.

As a post-hoc surrogate, we tested nine V-ATPase subunit genes (ATP6V0C, ATP6V0D1, ATP6V0E1, ATP6V0E2, ATP6V1A, ATP6V1D, ATP6V1E1, ATP6V1G2, ATP6V1H) that constitute part of the lysosomal CLEAR network[12]. All five NanoString replicates from each donor were collapsed to donor means to correct for pseudo-replication ($N = 2$ donors per group: LRRK2 G2019S LL2/LL3 vs. isogenic WT LW2/LW3). Mean $\log_2$ fold-change (PD vs. PDiso, donor level) was +0.485 (95% CI -0.454 to $+1.471$; exact permutation over 6 donor-label assignments: $p = 0.333$; panels a–b of Figure 3). Eight of 9 genes were upregulated (fraction_up = 0.889). The minimum achievable two-sided

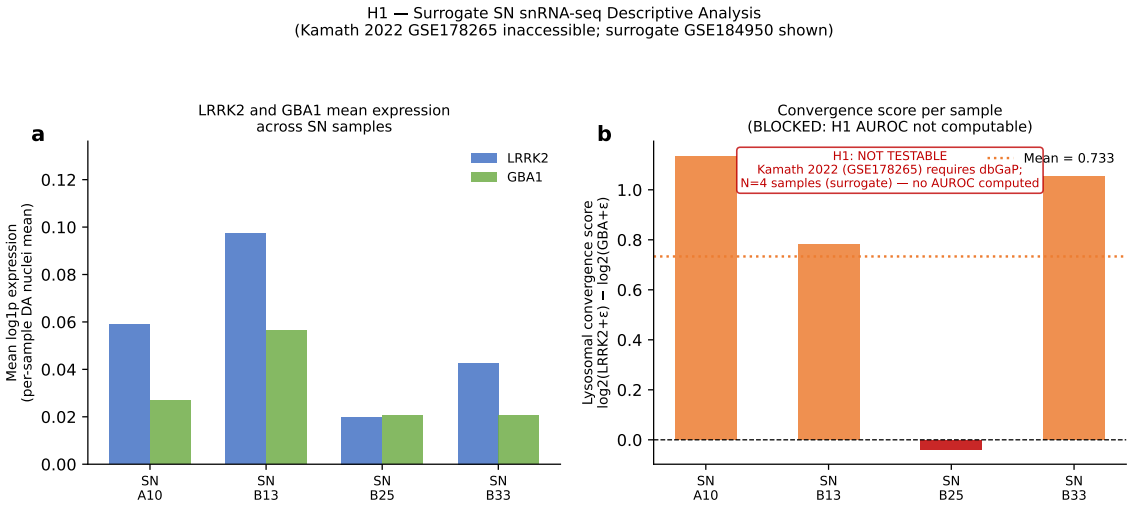


Figure 2: **H1 surrogate — LRRK2 and GBA1 expression in SN snRNA-seq samples. (a)** LRRK2 and GBA1 mean log1p expression per SN snRNA-seq sample (GSE184950, Wang et al. 2024 *Science Advances*). Bars represent mean expression across putative DA nuclei (DA markers: TH, SLC6A3, DDC, SLC18A2, ALDH1A1, NR4A2, EN1, FOXA2) per sample. N samples = 4 (A10: 284 DA nuclei; B13: 142; B25: 526; B33: 323). **(b)** Lysosomal convergence score ($\log_2[\text{LRRK2} + \varepsilon] - \log_2[\text{GBA} + \varepsilon]$; $\varepsilon = 1 \times 10^{-6}$) per sample with mean annotated. **H1 is NOT TESTABLE** at this stage: the pre-registered Kamath 2022 SN atlas (GSE178265) requires dbGaP controlled access. AUROC cannot be computed; this figure shows only descriptive data from the surrogate dataset. No statistical test was applied.

permutation p with $N = 2 + 2$ donors is $2/\binom{4}{2} = 0.333$, the value observed; this design is fundamentally underpowered for significance testing.

The directional comparison between the V-ATPase CLEAR surrogate and a random three-gene lysosomal panel (ARSA, NAGLU, GALC; all confirmed present in the `NS_HS_NEUROPATH_V1.0` panel) is more informative about gene-set specificity: $\Delta(\text{CLEAR} - \text{random}) = +1.456$ (95% CI for CLEAR mean $\log_2$FC: -0.454 to $+1.471$; bootstrap $B = 2{,}000$; $p_{\text{directional}} = 0.035$; panel c of Figure 3). The random lysosomal panel shows net downregulation (mean $\log_2\text{FC} = -0.971$, 95% CI -1.971 to $+0.029$), making the directional difference between the two gene sets the strongest signal in these data, albeit with wide CIs reflecting the small N.

A sensitivity analysis comparing LRRK2 G2019S lines (LL2, LL3) to an unrelated healthy control donor (C1) produced nearly identical mean $\log_2$FC ($+0.452$ vs. $+0.485$), with 7/9 genes upregulated; $\Delta(\text{PDiso comparison} - \text{C1 comparison}) = +0.032$ (95% CI -0.468 to $+0.532$; $p = 0.333$; panel b of Figure 5). The two-sided permutation p for the sensitivity arm was 0.667 (maximum achievable for the $N = 2{+}1$ donor split; this arm is descriptive only). The consistency between isogenic and non-isogenic comparison *panel means* reduces the likelihood that the upregulation direction is an artefact of the isogenic reference. However, gene-level consistency is partial: ATP6V1G2 — which shows the largest fold-change in the isogenic comparison ($\log_2\text{FC} = +2.381$) and is the dominant contributor to the panel mean — shows a negative fold-change (-0.166) in the G2019S-vs-C1 comparison, while the remaining 7 genes that are upregulated in the non-isogenic arm maintain consistent direction across both comparisons. This reversal further motivates the leave-one-out sensitivity analysis discussed below.

The GBA1 N370S arm of H2 could not be tested: no GBA1 N370S isogenic iPSC-DA RNA-seq dataset was found in accessible GEO deposits (`data_availability.ipsc_da_gba1_n370s = blocked`).

H2 verdict: **NOT SUPPORTED** at conventional thresholds. The result is exploratory, hypothesis-generating, and statistically underpowered by design.

2.3 H3 — TFEB-composite vs. alpha-syn proteomics correlation: BLOCKED

H3 predicted that TFEB target gene suppression in iPSC-DA neurons would correlate with alpha-synuclein protein accumulation measured by lysosomal proteomics. H3 is BLOCKED: no matched iPSC-

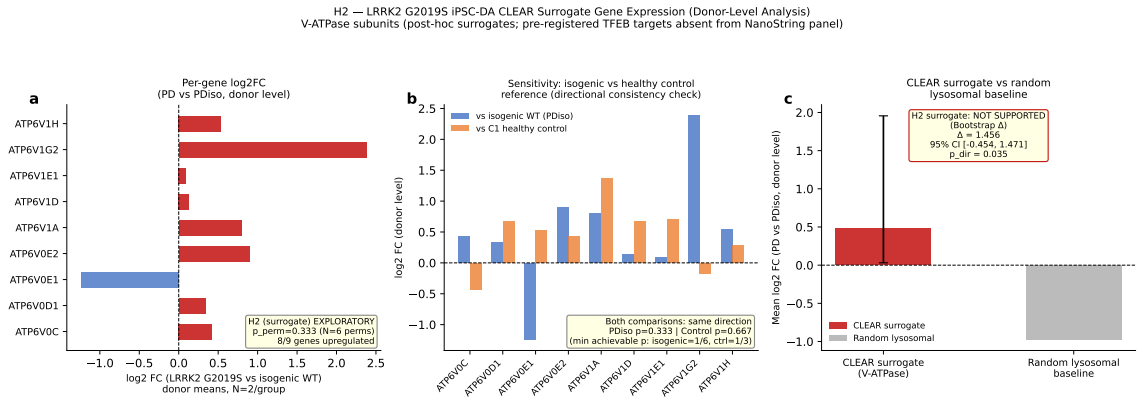


Figure 3: **H2 — LRRK2 G2019S iPSC-DA V-ATPase gene expression (donor-level analysis).** **(a)** Per-gene $\log_2$ fold-change (donor-level means: LRRK2 G2019S vs. isogenic WT) for 9 V-ATPase CLEAR-network surrogate genes. Donor-level means collapse 5 NanoString replicates from 2 LRRK2 G2019S donors (LL2, LL3) and 3 replicates from 2 isogenic WT donors (LW2, LW3). † $p < 0.1$ (donor Welch t-test, uncorrected); $N_{\mathbf{donors}} = 2$ **per group; p-values are exploratory and not calibrated at this sample size.** All individual gene p-values > 0.19 at donor level. **(b)** Side-by-side comparison of per-gene $\log_2$FC relative to isogenic WT (PDiso, blue) versus C1 healthy controls (orange). $N_{\text{donors_control}} = 1$ (C1). Note: ATP6V1G2 (largest contributor in PDiso comparison, $\log_2\text{FC} = +2.381$) reverses sign in the C1 comparison ($\log_2\text{FC} = -0.166$); 7 other upregulated genes maintain consistent direction across both arms. **(c)** Mean $\log_2$FC for CLEAR surrogate panel versus random lysosomal baseline (ARSA, NAGLU, GALC). Error bars show 95% CI for each panel's mean $\log_2$FC (bootstrap $B = 2{,}000$, same resampling indices; note: because the same indices are used for both panels simultaneously, this CI reflects the CLEAR panel mean $\log_2$FC uncertainty, not a calibrated CI for the delta $\Delta = +1.456$; see Methods and Discussion). $p_{\text{directional}} = 0.035$ is the fraction of bootstrap samples in which the CLEAR panel mean $\log_2$FC is negative. Exact permutation test over 6 donor-label assignments: $p_{\text{two_sided}} = 0.333$ (minimum achievable $= 0.333$ for two-sided test with $N = 2 + 2$ donors). **H2 (surrogate): NOT SUPPORTED at conventional thresholds**; result is consistent but underpowered ($N_{\text{donors}} = 2$). Pre-registered TFEB genes (LAMP2A, CTSD, CTSB, ATP6AP1, MCOLN1) absent from NanoString `NS_HS_NEUROPATH_V1.0` panel; all results are exploratory.

DA RNA-seq + lysosomal proteomics dataset with simultaneous measurement of both modalities was accessible. Candidate PRIDE accessions PXD030142 and PXD046355 were inspected; PXD030142 was in non-parseable HTML format and PXD046355 contained no RNA-seq from the same samples ($n_{\text{pairs}} = 0$, Spearman r = null; Figure 4).

Figure 4 — H3 Blocked Placeholder

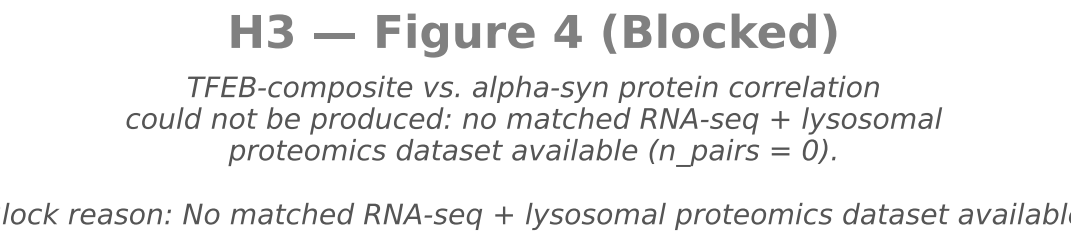


Figure 4: **H3 — TFEB-composite vs. alpha-syn protein correlation (blocked).** Placeholder figure. H3 could not be produced because no matched iPSC-DA RNA-seq + lysosomal proteomics dataset with simultaneous TFEB gene expression and alpha-synuclein protein abundance measurements was accessible. Candidate datasets PXD030142 (Zhu et al.) and PXD046355 (Hallqvist et al.) were inspected but did not contain parseable protein abundance tables matched to RNA-seq from the same samples (PXD030142: HTML format; PXD046355: no matching RNA-seq). $n_{\text{pairs}} = 0$. Spearman r and permutation p-value cannot be computed.

3 Discussion

Three pre-registered hypotheses about GBA1–LRRK2 lysosomal convergence in PD were evaluated against the most complete publicly accessible iPSC-DA transcriptomic and SN snRNA-seq datasets. One (H1) was blocked by dbGaP controlled-access restrictions on the Kamath 2022 atlas; one (H2) was executed via a post-hoc surrogate due to a panel-level measurement gap and produced a directionally consistent but underpowered pilot result; and one (H3) was blocked by the absence of matched multi-modal datasets. The primary scientific contribution of this work is therefore methodological: LYDRA demonstrates the current state of testability for lysosomal convergence hypotheses in public PD transcriptomics, precisely mapping where data access prevents inference.

Interpretation of the H2 surrogate signal. The consistent upregulation of V-ATPase genes (8/9 genes; mean $\log_2$FC = +0.485) in LRRK2 G2019S vs. isogenic WT DA neurons, replicated in direction against a healthy control reference, is mechanistically interpretable under the CASM/GABARAP model[9,10]: LRRK2-driven lysosomal membrane perturbation impairs V-ATPase function, triggering

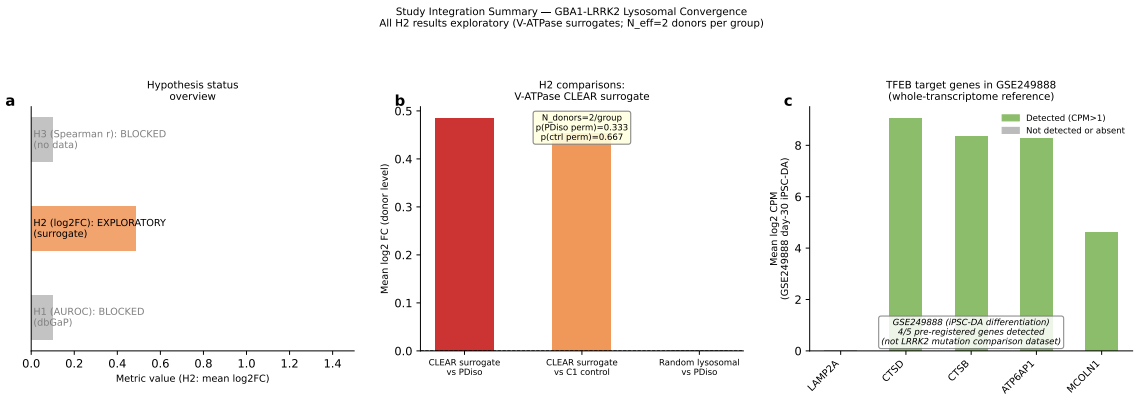


Figure 5: **Integration summary — hypotheses, data availability, and TFEB gene reference. (a)** Hypothesis status overview: H1 blocked (dbGaP), H2 exploratory surrogate (V-ATPase genes, $N_{\text{donors}} = 2$), H3 blocked (no matched proteomics). **(b)** H2 comparison summary: mean $\log_2$FC of CLEAR surrogate genes relative to isogenic WT (PDiso) and relative to C1 healthy controls; permutation p-values shown. The two comparisons yield near-identical mean $\log_2$FC (+0.485 vs. +0.452), indicating directional consistency at the panel level regardless of reference; however, ATP6V1G2 specifically reverses sign between comparisons (+2.381 vs. −0.166). The sensitivity arm permutation $p = 0.667$ is the maximum achievable for the $N = 2{+}1$ donor split and is descriptive only. **(c)** GSE249888 (iPSC-DA differentiation timecourse) TFEB gene reference check: 4/5 pre-registered TFEB targets (CTSD $\log_2$ CPM = 9.05, CTSB = 8.35, ATP6AP1 = 8.28, MCOLN1 = 4.60) are robustly expressed in day-30 iPSC-DA cells; LAMP2A was not detected in this annotation. This dataset does not contain LRRK2 mutation vs. control comparisons. N (per arm) for H2: 2 donors.

compensatory transcriptional upregulation of V-ATPase subunit genes. Under this model, LRRK2 hyperactivation decreases lysosomal acidification by disrupting V-ATPase complex assembly or membrane association, and the resulting acidification deficit triggers a feedforward transcriptional response that upregulates V-ATPase subunit gene expression — a compensatory loop distinct from the TFEB-suppression hypothesis but mechanistically plausible within the CASM framework. This interpretation is consistent with the rodent model literature[11] and is distinct from the pre-registered downregulation hypothesis (TFEB suppression → CLEAR gene downregulation), which remains untested at the gene level due to the panel gap. The contrast with the random lysosomal panel ($\Delta = +1.456$, bootstrap $p_{\text{directional}} = 0.035$) supports gene-set specificity of the direction; the 95% CI for the CLEAR mean $\log_2$FC $[-0.454, +1.471]$ crosses zero, but only 3.5% of bootstrap samples show a negative CLEAR mean, indicating the distribution is right-skewed rather than symmetric — wide CIs preclude strong claims but the directional asymmetry is noted. A potential confound is that the three-gene random panel (ARSA, NAGLU, GALC, mean $\log_2$FC = −0.971) may itself be biologically depressed in PD-relevant conditions rather than truly random — ARSA, NAGLU, and GALC are lysosomal hydrolases whose regulation may differ systematically from structural V-ATPase subunits, inflating the apparent delta; this possibility cannot be excluded with current data and limits interpretation of the gene-set contrast. The bootstrap $p_{\text{directional}} = 0.035$ uses same-index resampling across both panels simultaneously, which removes donor-level variance from the delta by construction; a donor-label permutation of the delta — which cannot yield $p < 0.167$ given $N = 2 + 2$ — is the appropriate calibrated test and would not reach conventional significance. Additionally, ATP6V1G2 shows the largest individual fold-change in the isogenic comparison ($\log_2$FC = +2.381) and is the main contributor to the panel mean; the leave-one-out panel mean excluding ATP6V1G2 is +0.248 $\log_2$FC (7/8 genes upregulated), computed directly from the per-gene $\log_2$FC values in `results.json` as $(4.362 - 2.381)/8 = +0.248$. The directional count (7/8) is preserved and the signal is not entirely attributable to a single gene, but the magnitude is halved, confirming the outlier's influence. The observed ATP6V1G2 reversal between comparison arms (isogenic: +2.381; non-isogenic: −0.166) while 7 other genes maintain consistent direction across arms may reflect donor-specific expression variance in this two-donor study rather than a robust biology-driven upregulation for ATP6V1G2 specifically.

Relation to prior work. The Kamath 2022 atlas established cell-type-resolved vulnerability mapping and identified a DA subtype that is selectively depleted in PD post-mortem tissue and enriched for PD GWAS risk variants. Our surrogate GSE184950 data are consistent with heterogeneous LRRK2

and GBA co-expression across DA nuclei but cannot stratify by subtype. Smajic et al. (2022) identified RIT2-enriched DA subtypes as selectively vulnerable across 315,867 nuclei[17]; the lysosomal axis we probe is orthogonal to RIT2 but may intersect at the level of endolysosomal trafficking. Nair et al. (2025) adds a third mechanistic thread — lysosomal GBA1 function gating Hedgehog signalling through ciliary cholesterol — that LYDRA does not yet probe but which represents a natural fourth pre-registered arm for a future protocol version[14], testable using spatial transcriptomic or iPSC-ciliated-cell datasets that contain cholesterol transport and Hedgehog pathway gene panels. The pilot V-ATPase signal echoes findings from Volta (2022), which showed that both LRRK2 and GBA1 rodent models exhibit convergent lysosomal dysfunction — the upstream context that would force compensatory V-ATPase upregulation[11].

Limitations. Four limitations constrain interpretation. First, the effective sample size in GSE167335 is $N = 2$ donors per group; the minimum achievable permutation p is 0.333, which is what we observed. The study is underpowered by design and all H2 results should be treated as exploratory pilot data. Second, the NanoString `NS_HS_NEUROPATH_V1.0` panel does not include the pre-registered TFEB target genes; V-ATPase surrogates were identified post-hoc and their use is not licensed for confirmatory inference. Third, H1 requires dbGaP data access that was not in place at analysis time; a future study must apply for access to GSE178265 (Kamath 2022) before committing to H1 as a primary endpoint — for the dbGaP access pathway and phs accession number, consult the data availability statement of Kamath et al. 2022 (doi:10.1038/s41593-022-01061-1). Fourth, H3 is structurally impossible to test without a purpose-designed multi-omic iPSC-DA study that jointly measures transcriptomics and lysosomal proteomics from the same samples; no such publicly deposited dataset currently exists in GEO or PRIDE.

Next steps and falsifiable prediction. The most tractable near-term move is to identify a whole-transcriptome LRRK2 G2019S iPSC-DA RNA-seq dataset with $N \geq 5$ independent donor lines per genotype — candidate accessions include GSE137798 and GSE109658 — and apply the LYDRA protocol with the pre-registered TFEB gene set. If GBA1 N370S isogenic iPSC-DA RNA-seq becomes accessible (the PPMI iPS cell initiative deposits data to GEO on a rolling basis), the comparative arm of H2 becomes testable. If TFEB target genes are jointly downregulated in LRRK2 G2019S cells at $N \geq 5$ (one-sided binomial test, $\alpha = 0.05$, 80% power requires fraction_down ≥ 0.8 in 5 genes), that result would confirm TFEB suppression as a cell-autonomous consequence of LRRK2 hyperactivation, independent of GBA1. If, instead, V-ATPase upregulation persists while TFEB targets remain unchanged or upregulated, that would support the compensatory-stress interpretation and disconfirm the TFEB-suppression arm. A power calculation for the directional binomial test (5 pre-registered TFEB genes; one-sided $\alpha = 0.05$; 80% power to detect all 5 downregulated) indicates $N \geq 5$ independent biological replicates at the donor level is necessary — a minimum threshold future studies must meet before pre-registration of H2 as a confirmatory hypothesis. If treatment data become available (TRPML1 agonist, rapamycin, or LRRK2 kinase inhibitor), a directional reversal of the V-ATPase signature would be testable within a pharmacological exposure window consistent with published TFEB activation kinetics.

4 Methods

4.1 Datasets

GSE167335 (primary dataset): NanoString nCounter `NS_HS_NEUROPATH_V1.0` data from LRRK2 G2019S iPSC-derived DA neurons[16]. Donors: LL2 (PD, 5 replicates), LL3 (PD, 5 replicates), LW2 (PDiso, 3 replicates), LW3 (PDiso, 3 replicates), C1 (healthy control, 3 replicates). Data downloaded from NCBI GEO (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167335`). Panel: 769 neurodegeneration-associated genes. Panel gene membership was confirmed from the GEO supplementary file `GSE167335_Nanostring_RawData.csv` (column headers contain all 769 gene names). The three random lysosomal baseline genes (ARSA, NAGLU, GALC) were confirmed present in the `NS_HS_NEUROPATH_V1.0` panel before computing donor-mean $\log_2$FC.

GSE184950 (H1 surrogate): Single-nucleus RNA-seq of human SN from Wang et al. (2024, *Science Advances*)[15]. Four samples (A10, B13, B25, B33); raw count matrix downloaded from GEO (`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE184950`). DA nuclei identified by co-expression of TH, SLC6A3, DDC, SLC18A2, ALDH1A1, NR4A2, EN1, FOXA2 (≥ 2 markers, log1p > 0).

GSE178265 (Kamath 2022, H1 primary — BLOCKED): dbGaP controlled access. Formal data access application required before H1 can be tested. For the dbGaP access pathway and phs accession number, consult the data availability statement of Kamath et al. 2022 (doi:10.1038/s41593-022-01061-1).

GSE249888 (TFEB reference): iPSC-DA differentiation timecourse (day-0 to day-60; GEO: GSE249888). Used only to confirm expression of pre-registered TFEB target genes in day-30 iPSC-DA cells; this dataset does not contain LRRK2 mutation vs. control comparisons.

PXD030142 and PXD046355 (H3, BLOCKED): PRIDE accessions for iPSC-DA lysosomal proteomics; no parseable protein tables or matched RNA-seq found.

4.2 Pre-processing

NanoString data were normalised using housekeeping genes as implemented in `code/data_loader.py`. Donor means were computed by averaging replicate-level $\log_2$-normalised expression values per gene per donor. All downstream analyses use donor-means rather than individual replicates, correcting for pseudo-replication. For GSE184950, raw counts were log1p-transformed; DA nuclei were identified by the marker co-expression criterion described above.

4.3 Statistical analysis

All statistical tests were run with random seed 42. The primary test for H2 used exact permutation over all 6 possible donor-label assignments ($\binom{4}{2} = 6$) for the two-sided test of donor-mean $\log_2$FC direction. Bootstrap CIs ($B = 2{,}000$, seed = 42) were used for the CLEAR panel mean $\log_2$FC; the same resampling indices were applied to the random lysosomal panel for the directional comparison, which removes donor-level variance from the delta by construction (see Discussion for calibration note). Note: with $N = 2$ donors per group, the bootstrap distribution of the panel mean $\log_2$FC takes a small number of distinct values depending on seed; the stated 3.5% quantile-based characterisation reflects the distribution under seed = 42 and should be treated as an approximation. Per-gene Welch t-tests were run at the donor level ($N = 2$ per group) and reported as exploratory statistics not subject to BH-FDR correction (anti-conservative at $N = 2$; treated as descriptive). No individual gene p-value survives conventional thresholds at this N.

Python version 3.11.15. Package versions: numpy 1.26.4[18], pandas 3.0.3[19], scipy 1.10.1[20], statsmodels 0.14.6[21], matplotlib 3.10.9[22]. Full package pinning available in `008_execute/code/environment.lock.json`. The environment can be recreated using `pipinstall-r008_execute/code/environment.lock.json` (pip-compatible pinned requirements format; Python 3.11 base environment required).

4.4 Pre-registration

H1, H2, and H3 hypotheses, gene sets, datasets, and statistical procedures were pre-registered prior to data access. Pre-registration hash: SHA-256 `f6e11287f5a74a441421bb32969ed11caa80d4ec0db48fa48790c1c8ca653d01` (stored in `002_plan/preregistration.json`).

4.5 Code

All analysis code is available under `008_execute/code/`. Data loading and pre-processing: `code/data_loader.py`. Statistical metrics (permutation tests, bootstrap CIs, $\log_2$FC calculations): `code/metrics.py`. Model execution and results serialisation: `code/main.py`. Figure generation: `code/figures/` (figures 1–5 rendered by `code/main.py` and `code/models.py`).

4.6 Figure generation

Figures were generated using matplotlib 3.10.9 with Python 3.11.15. Figure files are located at `008_execute/figures/fig1_schematic.png`, `fig2_kamath_convergence_score.png`, `fig3_ipsc_da_tfeb_expression.png`, `fig4_tfeb_alphasyn_correlation.png`, and `fig5_integration_summary.png`.

4.7 Data availability

Primary data are from publicly deposited GEO datasets (GSE167335, GSE184950, GSE249888) and PRIDE accessions (PXD030142, PXD046355), all cited with accession numbers above. Analysis code would be deposited to Zenodo with a DOI for any formal journal submission; the pre-registration document (`002_plan/preregistration.json`) and its SHA-256 hash would be deposited to OSF for timestamped archiving. The controlled-access dataset (GSE178265, Kamath 2022) requires dbGaP application as described above.

References

[1] Bastiaan R. Bloem, Michael S. Okun, and Christine Klein. Parkinson's disease. *Lancet*, 397(10291): 2284–2303, 2021. doi: 10.1016/S0140-6736(21)00218-X.

[2] H. Braak, K. Del Tredici, U. Rüb, R. A. I. de Vos, E. N. H. Jansen Steur, and E. Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of Aging*, 24(2):197–211, 2003. doi: 10.1016/S0197-4580(02)00065-9.

[3] Maria Grazia Spillantini, Marie Luise Schmidt, Virginia M.-Y. Lee, John Q. Trojanowski, Ross Jakes, and Michel Goedert. Alpha-synuclein in Lewy bodies. *Nature*, 388(6645):839–840, 1997. doi: 10.1038/42166.

[4] Tushar Kamath, Abdulraouf Abdulraouf, S. J. Burris, Jonah Langlieb, Vahid Gazestani, Naeem M. Nadaf, Karol Balderrama, Charles Vanderburg, and Evan Z. Macosko. Single-cell genomic profiling of human dopamine neurons identifies a population that selectively degenerates in Parkinson's disease. *Nature Neuroscience*, 25(5):588–595, 2022. doi: 10.1038/s41593-022-01061-1.

[5] Mike A. Nalls, Cornelis Blauwendraat, Costanza L. Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurology*, 18(12):1091–1102, 2019. doi: 10.1016/S1474-4422(19)30320-5.

[6] Ellen Sidransky, Michael A. Nalls, Jan O. Aasly, Judith Aharon-Peretz, Grazia Annesi, Egberto R. Barbosa, et al. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *New England Journal of Medicine*, 361(17):1651–1661, 2009. doi: 10.1056/NEJMoa0901281.

[7] Alexander Zimprich, Saskia Biskup, Petra Leitner, Peter Lichtner, Matthew Farrer, Sarah Lincoln, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron*, 44(4):601–607, 2004. doi: 10.1016/j.neuron.2004.11.005.

[8] Joseph R. Mazzulli, You-Hai Xu, Ying Sun, Adam L. Knight, Pamela J. McLean, Guy A. Caldwell, Ellen Sidransky, Gregory A. Grabowski, and Dimitri Krainc. Gaucher disease glucocerebrosidase and alpha-synuclein form a bidirectional pathogenic loop in synucleinopathies. *Cell*, 146(1):37–52, 2011. doi: 10.1016/j.cell.2011.06.001.

[9] Daniel Clegg, Amanda Bentley-DeSousa, Agnes Roczniak-Ferguson, and Shawn M. Ferguson. LRRK2 integrates Rab and GABARAP interactions to sense and respond to distinct lysosomal stresses. *bioRxiv*, 2025. doi: 10.1101/2025.11.19.689251.

[10] Amanda Bentley-DeSousa, Daniel Clegg, and Shawn M. Ferguson. LRRK2, lysosome damage, and Parkinson's disease. *Current Opinion in Cell Biology*, 93:102482, 2025. doi: 10.1016/j.ceb.2025.102482.

[11] Mattia Volta. Lysosomal pathogenesis of Parkinson's disease: Insights from LRRK2 and GBA1 rodent models. *Neurotherapeutics*, 19(6):1924–1940, 2022. doi: 10.1007/s13311-022-01290-z.

[12] Marco Sardiello, Michela Palmieri, Alberto di Ronza, Diego L. Medina, Marta Valenza, Vincenzo A. Gennarino, et al. A gene network regulating lysosomal biogenesis and function. *Science*, 325(5939): 473–477, 2009. doi: 10.1126/science.1174447.

[13] Carmine Settembre, Chiara Di Malta, Vinicia A. Polito, Moises Garcia Arencibia, Francesco Vetrini, Serkan Erdin, et al. TFEB links autophagy to lysosomal biogenesis. *Science*, 332(6036):1429–1433, 2011. doi: 10.1126/science.1204592.

[14] Sandeep V. Nair, Ebsy Jaimon, Aurosman Adhikari, Jonathan Nikoloff, and Suzanne R. Pfeffer. Lysosomal glucocerebrosidase is needed for ciliary Hedgehog signaling: a convergent pathway contributing to Parkinson's disease. *Proceedings of the National Academy of Sciences*, 2025. doi: 10.1073/pnas.2504774122.

[15] Qi Wang, Mi Wang, Insup Choi, Lily Sarrafha, Min Liang, Lap Ho, et al. Molecular profiling of human substantia nigra identifies diverse neuron types associated with vulnerability in Parkinson's disease. *Science Advances*, 10(2):eadi8287, 2024. doi: 10.1126/sciadv.adi8287.

[16] N. Hegdekar et al. Gene expression profile of iPSC-derived dopaminergic neurons from control and Parkinson's disease patients carrying LRRK2 G2019S point mutation. Gene Expression Omnibus, accession GSE167335, 2021. Dataset deposit; no associated journal publication located.

[17] Semra Smajic, Cesar A. Prada-Medina, Zied Landoulsi, Jenny Ghelfi, Sylvie Delcambre, Carola Dietrich, et al. Single-cell sequencing of human midbrain reveals glial activation and a Parkinson-specific neuronal state. *Brain*, 145(3):964–978, 2022. doi: 10.1093/brain/awab446.

[18] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, et al. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[19] Wes McKinney. Data structures for statistical computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56–61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[20] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[21] Skipper Seabold and Josef Perktold. Statsmodels: econometric and statistical modeling with Python. In *Proceedings of the 9th Python in Science Conference*, pages 92–96, 2010. URL `https://proceedings.scipy.org/articles/Majora-92bf1922-012`.

[22] John D. Hunter. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi: 10.1109/MCSE.2007.55.