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.

When do mouse models transfer? Cross-species prediction of Huntington's disease from brain transcriptomes

Abdul Ahad

ahadraj@ieee.org

North South University https://orcid.org/0009-0005-1253-5314

Nurjahan Begum

Universiti Sains Islam Malaysia

Nusrat Jahan

Chittagong Medical College

Research Article

Keywords: Huntington's disease, cross-species transfer, transcriptomics, permutation testing, cell-type deconvolution, software

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11860740/v1

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

Additional Declarations: The authors declare no competing interests.

When do mouse models transfer? Cross-species prediction of Huntington's disease from brain transcriptomes

Abdul Ahad[1*], Nurjahan Begum[2], Nusrat Jahan[3]

[1] Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh

[2] Universiti Sains Islam Malaysia, Nilai, Negeri Sembilan, Malaysia

[3] Chittagong Medical College, Chattogram, Bangladesh

*Corresponding author: Abdul Ahad | ahad.abdul01@northsouth.edu

Abstract: Mouse models are the main preclinical testbed for Huntington's disease (HD), yet whether their brain transcriptomes predict human disease is usually judged by overlapping gene lists. A stricter test trains a classifier on mouse data and applies it, unchanged, to human tissue. We show that this test is easily misread, and we provide methods and software to audit and improve it. TransferAudit, an open-source Python package, combines source-label permutation, random-gene and composition-adjusted nulls with an inflation index, τ, that is computed before training and flags targets in which conventional p-values fail (up to 43% false positives once $\tau \geq 0.4$). In HD, a pre-specified knock-in striatum model reached an area under the curve (AUC) of 0.87 in human caudate but not in cortex or cerebellum; its conventional p-value (7×10^{-8}) became 0.018–0.035 under permutation, and deconvolution against sorted human nuclei explained the bulk result by projection-neuron loss and gliosis. Single mouse cohorts transferred inconsistently (AUC 0.37–0.97). A new consensus method pooling five cohorts from four model families cut this variability fivefold, gave the strongest evidence of transfer (AUC 0.94, $p = 0.001$) and was the first model to transfer significantly into sorted human projection neurons (donor-level AUC 1.00, $p = 0.025$), but not interneurons. Its signature shared 493 genes with HD changes inside sorted projection neurons, eleven times the chance expectation (exact $p = 0.012$), yet this per-cell signal was invisible in independent bulk caudate once composition was removed. Cross-species prediction needs auditing, multi-cohort training and cell-resolved targets.

Keywords: Huntington's disease; cross-species transfer; transcriptomics; permutation testing; cell-type deconvolution; software.

1 Introduction

Huntington's disease is caused by an expanded CAG repeat in *HTT*, the gene encoding huntingtin [1]. The disease first strikes the projection neurons of the striatum, also called medium spiny neurons (MSNs), and the caudate nucleus loses them earliest [2, 3]. Single-nucleus studies have recently added a mechanism to this picture. In each neuron the inherited repeat keeps lengthening over decades, and once it passes roughly 150 CAGs the cell loses its transcriptional identity and soon dies [4]. Cortical neurons seem to follow a similar path [5, 6].

Nearly every candidate therapy for HD is tested in mice first, so it matters how much of the human disease these animals reproduce. The literature gives a mixed answer. Striatal expression changes in several mouse models resemble those in human caudate [7, 8], and a knock-in allelic series showed that repeat length and age drive networks dominated by falling MSN identity genes [9]. Cell-resolved comparisons, however, find severe MSN loss in patients that the mice do not show [10, 11]. Bulk-tissue resemblance could therefore reflect shared molecular biology, or simply two different processes that happen to lower the same genes.

Most studies compare species through shared gene lists or correlated fold changes. A more demanding test asks whether a classifier trained only on mouse data can recognise disease in human tissue it has never

seen. That framing turns a qualitative comparison into a measurable one, but it also brings statistical traps that overlap studies never face. Leakage inflates accuracy [14, 15], batch effects can pass for biology [16], and in cancer, random gene signatures predict outcome as well as published ones because they all correlate with proliferation [17–20]. When a model is trained in one species and scored in another, even the right null hypothesis is unclear, and to our knowledge no available software checks for these problems.

This article contributes two methods, a tool, a resource and a case study. The first method is an analytic null distribution for the AUC of a transferred classifier. Its single parameter, the inflation index τ, can be computed before training and tells the analyst whether conventional p-values will be trustworthy. The tool, TransferAudit, packages this index together with source-label permutation, random-gene and composition-adjusted nulls, and returns a plain-language verdict. The second method, multi-cohort consensus transfer, trains on several mouse cohorts at once and keeps only genes whose disease changes agree across them, because we found single cohorts to be unreliable sources. The resource is a conserved per-cell signature: genes changed in the same direction in the mouse consensus and inside sorted human projection neurons, with an exact permutation test. The case study applies the audit to HD across three human brain regions from the same donors, fluorescence-sorted human nuclei and five mouse cohorts, including mice in which genetic suppressors of somatic expansion were introduced.

Some of our observations confirm earlier work, and we say so where they arise: mouse striatal signatures do resemble human caudate [7, 8], and patients lose MSNs that mice keep [10]. What is new is how much of the resemblance survives a proper audit. Conventional statistics overstate it by orders of magnitude; in bulk caudate it is fully explained by cell composition; a smaller per-cell signal exists only in projection neurons; and how well a mouse cohort transfers is not predicted by how faithfully it models HD genetics. Pooling cohorts solves the last problem: the consensus model transfers more strongly and far more consistently, and it reaches human projection neurons. The genes it shares with them, however, cannot be seen in bulk tissue once composition is accounted for.

2 Related work

Several methods translate expression changes between species. Found In Translation learns gene-wise mappings from paired public mouse and human datasets [21]. TransComp-R projects human data onto principal components of mouse data [22] and has since been used for tuberculosis and Alzheimer's disease [23], and autoencoders can translate signatures without an ortholog map [24]. In Alzheimer's disease, sparse models built on mouse data have been projected onto human cohorts [25], and single-cell foundation models now offer cross-species embeddings, although simple representations often perform as well [26]. These approaches are judged by accuracy or correlation in the target species. None, to our knowledge, is tested against nulls that retrain on shuffled source labels, and none checks whether the target's cell composition drives the result.

Within HD, machine learning has mostly been applied inside human cohorts or from human to mouse [27], and robust striatal signatures from the allelic series were validated only in other mouse models [28]. Our statistical concerns have precedents elsewhere. Permutation tests ask whether a classifier exploits genuine class structure [29], and restricted permutations can expose confounding [30]. In clinical prediction, case mix changes discrimination at external validation [31], small validation sets give imprecise AUCs [32], and TRIPOD+AI sets reporting standards [33]. Deconvolution methods separate cell composition from per-cell regulation in bulk tissue [34, 35]. TransferAudit brings these ideas together for the cross-species setting and adds an analytic null that the earlier work lacks.

3 Materials and methods

3.1 Data

All datasets are public in the Gene Expression Omnibus [36] and are listed in Table 1. Training data came from the knock-in allelic series of Langfelder et al. [9]: heterozygous mice with 20, 80, 92, 111, 140 or 175 CAG repeats, profiled at 2, 6 and 10 months with eight animals per group, plus wild-type littermates. Cortex (144 knock-ins and 24 wild-type mice) and striatum (the same knock-in design and 64 wild-type mice) were analysed separately.

Human caudate, frontal cortex and cerebellum came from one postmortem cohort profiled on Affymetrix U133A arrays [37]. The caudate set had 38 HD and 32 control samples. Its HD cases were mostly early-stage (Vonsattel grades 0, 1, 2, 3 and 4 numbered 3, 13, 16, 5 and 1; 2) and age-matched to controls (59.0 and 58.4 years). Frontal cortex and cerebellum from the same donors contributed 37 and 39 HD samples against 28 and 27 controls. A separate RNA-seq cohort of prefrontal cortex (Brodmann area 9; 20 HD and 49 controls) represented late-stage disease, with a mean grade of 3.3 and HD deaths about ten years earlier than controls (58.3 and 68.4 years) [38].

Sorted nuclei came from Mätlik et al. [39], who used fluorescence-activated nuclear sorting on postmortem striatum. Their 134 RNA-seq samples come from 14 donors. HD nuclei exist only for neurons: D1 MSNs from four HD donors, D2 MSNs from five and interneurons from four. The eight control donors also contributed astrocytes, microglia and oligodendrocytes, which we used as a deconvolution reference. Four further mouse cohorts, all profiled at 6 months, formed a genetic benchmark: Q140 knock-ins crossed with Msh3 or Pms1 knockouts [12], zQ175 knock-ins with or without Msh3 [13], and translating-ribosome profiles of D1 and D2 MSNs from R6/2 mice [11].

3.2 Pre-processing and cross-species mapping

Sequencing counts were converted to $\log_2$ counts per million plus one, using library sizes over all genes. Microarray signals were $\log_2$-transformed. We dropped probes annotated to more than one gene and, where several probes measured the same gene, kept the most highly expressed; in caudate, 19,933 probes collapsed to 12,502 genes. Genes were matched between species through 16,293 one-to-one orthologs from the HGNC Comparison of Orthology Predictions [40, 41], accessed through babelgene [42], with at least three of twelve source databases in agreement. A gene counted as expressed in a sequencing dataset if it reached one count per million in a quarter of samples, and in an array dataset if its mean exceeded the 25th percentile. Only genes expressed in both datasets of an analysis were kept, leaving 7,721–12,308 genes. Every dataset was then standardised gene by gene with its own mean and standard deviation, without looking at labels. Centring adds the same constant to every score of a linear model, so it cannot change that model's AUC, whatever the class balance.

3.3 Training designs and the confirmatory analysis

Mice with at least 111 repeats and aged at least 6 months were labelled as diseased. In the primary design, Q20 knock-ins processed alongside them served as controls; one sensitivity design added wild-type littermates, and another kept only processing batches that contained both classes at the same age. Before touching the striatal data we fixed one confirmatory analysis — striatum to caudate, all eligible mice, Q20 controls — and its success criterion. Everything else in this article is exploratory. The 32 exploratory transfer tests (five further region pairs and the prefrontal cohort, composition-invariant transfer, six sorted-nucleus targets and their cell-type contrast, and eighteen benchmark tests) were corrected together by the Benjamini–Hochberg procedure [43].

3.4 Classifiers

The default classifier selects 200 genes by an analysis-of-variance F-test inside every training fold and fits an L2-penalised logistic regression (inverse penalty 0.1, balanced class weights); a random forest of 500 trees served as an alternative [44, 45]. Gene number and penalty were fixed in advance. Tuning them on human labels would leak target information, and tuning on mouse labels alone would optimise within-mouse accuracy, which we show to be uninformative. Within-mouse performance was estimated by five-fold stratified cross-validation repeated five times. The final model was refitted on all eligible mice and applied without modification to human samples, with AUC confidence intervals from 2,000 stratified bootstrap resamples [46].

3.5 TransferAudit

TransferAudit takes standardised source and target matrices with matched genes, the two sets of labels and, optionally, source strata and target cell proportions. It trains the classifier on the source, scores the target and then runs four checks.

Inflation index and analytic null. Write the target score as $s = \mathbf{w}^\top \mathbf{x}$. Let $\boldsymbol{\delta}$ be the difference between disease and control means in the target and $\boldsymbol{\Sigma}_t$ the pooled within-class covariance. If the scores are roughly normal with equal variance in both classes, AUC $= \Phi\left(d'/\sqrt{2}\right)$ with $d' = \mathbf{w}^\top \boldsymbol{\delta}/(\mathbf{w}^\top \boldsymbol{\Sigma}_t \mathbf{w})^{1/2}$. Suppose the source labels carry no transferable information. The fitted weights then behave like a random draw shaped only by the source covariance, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_s)$ on the selected genes. The numerator is normal with variance $\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_s \boldsymbol{\delta}$, and with many genes the denominator settles near $\mathrm{tr}(\boldsymbol{\Sigma}_s \boldsymbol{\Sigma}_t)^{1/2}$. Hence

$$d' \sim \mathcal{N}(0, \tau^2), \qquad \tau^2 = \frac{\boldsymbol{\delta}^\top \boldsymbol{\Sigma}_s \boldsymbol{\delta}}{\mathrm{tr}(\boldsymbol{\Sigma}_s \boldsymbol{\Sigma}_t)}, \quad (1)$$

and an observed AUC A has the one-sided analytic p-value $1 - \Phi\{\sqrt{2}\,\Phi^{-1}(A)/\tau\}$. The index τ grows when the disease shift in the target runs along many genes that also co-vary in the source. It needs target labels and source data but no trained model. TransferAudit averages it over 2,000 random gene subsets of the selected size, using only $n \times n$ matrix products. With true labels and random genes, the weights approximate the source mean difference, so d' becomes a sum of many weak gene-level agreements; this is why random gene sets can transfer.

Permutation and random-gene nulls. Conventional tests over target samples treat the trained model as fixed and ignore the main source of uncertainty, namely which direction the source data push the model. TransferAudit therefore retrains the whole pipeline 1,000 times on source labels shuffled within strata (here, mouse age) and, separately, 1,000 times on 200 random genes with the true labels. Empirical p-values are $(1 + \#\{A_b \geq A_{\mathrm{obs}}\})/(B + 1)$, and because they carry Monte Carlo error we report their range across independent runs.

Composition audit. Target cell proportions can be supplied or estimated by the package's deconvolution routine. For each of five classes (MSNs, interneurons, astrocytes, microglia and oligodendrocytes), genes averaging at least ten counts per million in control sorted nuclei [39] were ranked by the $\log_2$ ratio of their expression in that class to their highest expression elsewhere. The top fifty per class formed a 250-gene signature. Bulk signals were returned to a linear scale, each marker was divided by its mean reference level, samples were median-scaled to absorb platform differences, and proportions were fitted by non-negative least squares and rescaled to sum to one [34]. The classifier's target scores are then regressed on these proportions, and the residual AUC is compared with the residual AUCs of the same permuted models.

Verdict. The report states whether τ exceeds 0.4, the threshold above which our simulations showed inflated conventional tests, and whether transfer beats the permutation null, the random-gene null and the composition-adjusted null at $\alpha = 0.05$. TransferAudit is written in Python on scikit-learn and SciPy, has a command-line interface and a test suite on synthetic data with known answers, and audits the confirmatory HD analysis, including 1,000 permutations, 1,000 random gene sets and deconvolution, in about 25 seconds on a laptop processor.

Baselines and sensitivity analyses. Two published strategies served as baselines: a TransComp-R-style projection [22], which learnt ten principal components from either species without labels and trained logistic regression on the mouse projections, and CORAL domain adaptation [47], which matches the second-order statistics of the selected mouse genes to those of the human data. Sensitivity analyses restricted the ortholog map to genes supported by eleven or twelve of twelve databases, or replaced standardisation by a within-sample rank-based normal-scores transform. As cruder composition checks we also used marker-gene scores [35], removed composition-correlated genes, and residualised every gene on cell scores in both species before training. MSN identity genes (for example PPP1R1B, DRD1, PENK and PDE10A) were compared with pan-neuronal genes by Cohen's d.

3.6 Simulations

Two simulations tested the index. The first estimated type-I error: each target had 1,000 genes in 70 samples, with a hidden factor acting on 30% of genes and displaced by $\Delta \in \{0,1,2,3\}$ standard deviations in disease. The 64 source samples had labels unrelated to their expression, and, to mimic conserved co-expression, each gene inherited its loading from the target with probability 0 or 0.8. The second compared analytic and permutation p-values (99 permutations each) across 800 simulated datasets from eight settings, some containing transferable signal and some not.

3.7 Sorted nuclei and genetic benchmark

In sorted nuclei, each cell class was normalised on its own, samples were scored with the primary model and scores were averaged per donor, so the AUC compares HD with control donors rather than samples. Controls had also been sampled from posterior regions, so region-matched analyses used only anterior caudate, anterior putamen and caudate. Cell-type specificity was tested by comparing the difference between MSN and interneuron AUCs with the differences produced by the same permuted models.

For the genetic benchmark, the primary model first scored every mouse in the four additional cohorts. A new classifier was then trained on each contrast and applied to caudate and sorted MSNs. The contrasts were disease versus wild type in each model; the expansion-dependent contrast, Q140 versus Q140 lacking Msh3 or Pms1; and five controls: zQ175 with or without Msh3, which reportedly does not restore striatal transcription [13], and Msh3 or Pms1 knockouts versus wild type, which carry no disease allele. Each cohort was normalised on its own, and the translating-ribosome data were standardised separately within D1 and D2 neurons.

3.8 Multi-cohort consensus transfer

Single mouse cohorts gave contradictory answers (see Results), so we developed a training procedure that pools them. Five striatal cohorts were used: the allelic series (48 diseased and 16 Q20 mice), zQ175 (8 and 8), the two Q140 cohorts (8 and 8; 8 and 7) and R6/2 translating-ribosome profiles (11 and 11). Each was standardised on its own and restricted to the 6,846 genes shared by all cohorts and human targets. For each cohort and gene we computed a Welch t statistic for disease against control, and combined the cohorts by Stouffer's method, $z_g = \sum_c t_{gc} / \sqrt{C}$, giving every cohort an equal vote. Genes whose direction agreed in

at least $C - 1$ cohorts were ranked by $|z_g|$, and the top 200 entered a logistic regression trained on the pooled cohorts, with sample weights that gave every cohort, and both classes within it, the same total weight. Significance came from a cohort-stratified permutation null: labels were shuffled within each cohort (and within age or cell-type strata), and the whole procedure, including the meta-analysis and gene selection, was refitted 1,000 times. Stability was assessed by leaving out each cohort in turn and by comparison with single-cohort models built on the same genes.

3.9 Conserved per-cell signature

To find disease changes shared by the two species within a cell type, we averaged the standardised sorted-MSN profiles per donor (5 HD and 8 control donors) and computed a Welch t statistic per gene. A gene was called conserved if its mouse consensus statistic satisfied $|z_g| \geq 3$, its human sorted-MSN statistic $|t_g| \geq 2.5$, and the two agreed in sign. The null distribution of the number of conserved genes was obtained exactly, by repeating the procedure for all $\binom{13}{5} = 1{,}287$ assignments of HD labels to donors with the mouse statistics fixed. The signature was then scored in the independent bulk caudate cohort as the mean standardised expression of its genes, signed by their direction, and adjusted for deconvolved composition; this adjusted AUC was compared with the same 1,287 donor-permuted signatures, with 1,000 random gene sets of equal size, and with equally sized signatures built from the human or mouse data alone. The eight tests of the consensus and signature analyses form a second exploratory family, corrected by the Benjamini–Hochberg procedure. Both procedures are implemented in TransferAudit (ConsensusTransfer and conserved_signature).

3.10 Implementation and reproducibility

Analyses used Python with scikit-learn, SciPy and pandas [45] and fixed random seeds. Seventeen analysis scripts reproduce every number, table and figure, and TransferAudit reproduces the confirmatory audit from the processed data. Running them requires the public GEO datasets in Table 1. Claude (Anthropic) was used to assist with writing analysis code and drafting text; the authors reviewed and verified all code, results and text, and take full responsibility for the content.

4 Results

4.1 An audit of the confirmatory transfer

The pre-specified classifier met its criterion: trained on knock-in mouse striatum, it recognised HD in human caudate with an AUC of 0.87 (95% CI 0.77–0.95). Table 2 shows the full audit. The conventional Mann–Whitney test gave $p = 7 \times 10^{-8}$, but the inflation index was 0.69, well above the 0.4 threshold, and permutation of the mouse labels gave $p = 0.018$–0.035 across independent runs. Random sets of 200 genes, trained on the true labels, reached an average AUC of 0.70 ($p_{\mathrm{rand}} = 0.12$ for the selected genes). After adjustment for deconvolved cell composition, the AUC fell to 0.47 ($p_{\mathrm{perm}} = 0.77$). TransferAudit's verdict, generated automatically, summarises the rest of this section: the transfer is real by permutation, but it is carried by many genes rather than specific ones, and composition explains it.

The headline AUC itself was robust. Four training designs gave 0.73–0.93; restricting the ortholog map to the best-supported genes gave 0.84 and 0.88; and a rank-based transform gave 0.95. The random forest reached 0.85 but did not beat its own permutation null ($p_{\mathrm{perm}} = 0.085$), and within-mouse cross-validation was saturated at 1.00, which tells us little. The genes the classifier chose included classic MSN identity genes, all lower in HD: DRD1, PENK, PDE10A, PDE1B, CNR1, ARPP21 and PCP4. Within HD

donors, scores did not follow age at death or inherited repeat length, but they rose with neuropathological grade in six of eight design and model combinations (Spearman correlations 0.43–0.55).

4.2 Transfer is confined to caudate

Caudate, frontal cortex and cerebellum came from the same donors on the same arrays, so region could be separated from cohort and platform (Figure 2, Table 3). Classifiers trained on either mouse region recognised HD in caudate (0.87–0.88), whereas AUCs were 0.50–0.59 in frontal cortex and 0.34–0.66 in cerebellum. The late-stage prefrontal cohort gave 0.35–0.69 across designs; its best result (a random forest with Q20 controls, AUC 0.69, Mann–Whitney $p = 0.008$) did not survive correction for eight combinations, and the logistic-regression version had $p_{\mathrm{perm}} = 0.29$.

We read cortex and cerebellum as low-signal tissues rather than negative controls. Cortical neurons undergo expansion-driven changes much like those in striatum [5, 6, 48], but in early-stage cortex few have crossed the critical repeat length, and a model trained and tested within human frontal cortex reached only 0.75. Cerebellum (0.85) and late-stage prefrontal cortex (0.93) did carry human signals that no mouse model captured.

The training data held a trap of their own. In the cortex series, 3,654 genes separated wild-type from Q20 mice at 6 months, although neither group carries a pathogenic allele, and every 6-month comparison against wild-type controls pointed the wrong way relative to human disease, whatever the repeat length (Figure 9). Training only within shared processing batches lowered within-mouse cross-validated AUC from 0.99 to 0.82.

4.3 Why conventional statistics fail

The width of the permutation null in caudate (95% range 0.15–0.88; Figure 3) has a simple cause. The deconvolved fraction of projection neurons alone separates HD from control caudate with an AUC of 0.96, so almost any classifier direction that happens to align with this axis scores well. Across 200 random gene sets, alignment with human neuron content explained nearly all variation in AUC (Spearman correlation 0.95). Neither baseline escaped the problem. The projection model reached 0.83 with mouse components, but its own null extended to 0.94 ($p_{\mathrm{perm}} = 0.30$); with human components it reached 0.61 ($p_{\mathrm{perm}} = 0.32$), and CORAL reached 0.76 ($p_{\mathrm{perm}} = 0.081$).

The same effect appeared when we tried to match mouse conditions to human disease stages. Separate classifiers for each of 15 combinations of repeat length and age (eight against eight mice) were scored against four human grade groups. By the Mann–Whitney test, 34 of these 60 comparisons were significant; after permutation testing with false-discovery control, none were (smallest permutation $p = 0.015$; Figure 10). Several conditions with hardly any mouse signal gave extreme values: 2-month Q80–Q111 mice, which differed from controls in only 5–57 genes, produced human AUCs of 0.07–0.33, far from 0.5 but in the wrong direction. With eight mice per group, no single mouse condition can be matched to a human stage.

4.4 The inflation index predicts which targets mislead

Caudate targets had inflation indices of 0.69–0.78, against 0.24 for frontal cortex, 0.29 for cerebellum and 0.37 for prefrontal cortex, and these values tracked the widths of the permutation nulls (Figure 4a). The rank order of the seven analyses was almost identical under analytic and permutation testing (rank correlation 0.96), and no analytic value differed from its permutation counterpart by more than a factor of 3.2, whereas conventional p-values were up to six orders of magnitude too small (Figure 4b). The analytic null was a little narrow, mainly because of the normal approximation: exact AUCs of random directions

drawn from the source covariance reproduced the caudate permutation null closely (97.5th percentiles 0.883 and 0.879). For six of seven analyses, the random-gene approximation predicted the average random-set AUC within 0.04.

The simulations show how to use the index (Figures 4c and 5). Without transferable signal, conventional tests stayed near the nominal 5% only while τ was below 0.4 (false-positive rate 0.03); above it they rejected 27–29% of null datasets in the calibration study, and up to 43% in the first simulation when co-expression modules were shared between species. Permutation held 1–4%, and so did the analytic test (0–4%). With real signal, however, the analytic test was underpowered (14–77% against 100% for permutation). We therefore use τ as an early warning, computed before any model is trained, and leave inference to permutation. Every caudate analysis lay above the threshold, and every cortical and cerebellar analysis below it.

4.5 Cell composition explains bulk transfer

Deconvolution recovered the known pathology of HD caudate (Table 4). The estimated MSN fraction was roughly halved in HD (0.22 against 0.46 in controls), astrocytes rose from 0.26 to 0.42, microglia and oligodendrocytes rose modestly, and interneurons held steady (0.12 against 0.11), as expected from their relative sparing. The classifier's score tracked the MSN fraction (Spearman correlation -0.63), and once the score was adjusted for the estimated proportions, transfer vanished (AUC 0.47, 95% CI 0.33–0.61; $p_{\mathrm{perm}} = 0.77$; Figure 6a). Cruder marker-based corrections left AUCs of 0.59–0.77, and composition-invariant transfer 0.62. Adjustment may also remove real changes in surviving neurons that progress with disease stage, so these values are conservative. Still, the bulk signal cannot be credited to shared per-cell biology.

4.6 A smaller per-cell signal remains in projection neurons

Sorting removes cell loss from the picture, and in sorted nuclei the primary classifier still separated HD from control donors using MSN nuclei (AUC 0.95), beating random gene sets ($p_{\mathrm{rand}} = 0.006$). It failed in interneurons from the same donors (0.28; Table 5, Figure 6b). With only four or five HD donors, the permutation nulls stayed wide even after sorting ($\tau = 1.06$–1.26), so no single MSN analysis reached $p_{\mathrm{perm}} < 0.05$ (0.051–0.13). The comparison between cell types was more telling: the MSN-minus-interneuron difference (+0.67) exceeded the same permuted models ($p_{\mathrm{perm}} = 0.027$). Random gene sets with true labels also favoured MSNs, by 0.51 on average, so the specificity comes from the mouse disease labels rather than from particular genes. Scores did not follow postmortem interval among controls.

Inside sorted MSNs, identity genes were lower in HD donors (Cohen's $d = -1.43$, $p = 0.03$), so in human MSNs this decline is not only a matter of cell loss. Pan-neuronal genes fell too ($d = -0.93$), however, and identity relative to pan-neuronal expression did not differ ($d = -0.13$). Mouse striatum behaved differently: identity genes fell ($d = -2.36$) while pan-neuronal genes rose ($d = +1.41$; Figure 6c). Mice lose MSN identity selectively; human MSNs undergo a broader decline.

4.7 How faithful a mouse cohort is does not predict how well it transfers

The primary classifier generalised to all four independent cohorts (Table 6, Figure 7a), scoring Q140, zQ175 and R6/2 mice at 0.91–1.00 and every wild-type and knockout group at 0.01 or less. Its scores followed the genetics of somatic expansion. Q140 mice heterozygous or homozygous for an Msh3 deletion scored 0.74 and 0.23, and Q140 mice lacking Pms1 scored 0.54, whereas zQ175 mice without Msh3 stayed at 0.98, in line with the report that deleting Msh3 does not normalise striatal transcription in

this model [13]. Within mice, then, the classifier tracks expansion-dependent pathology rather than strain or batch.

Transfer to humans was a different matter (Figure 7b). zQ175-trained classifiers reached 0.94 in caudate ($p_{\mathrm{perm}} = 0.012$) and 0.97 in sorted MSNs ($p_{\mathrm{perm}} = 0.018$), and a classifier trained on R6/2 MSN profiles reached 0.97 in caudate ($p_{\mathrm{perm}} = 0.001$). Two Q140 cohorts of identical design disagreed, however (0.89 and 0.36), and across the four disease-versus-wild-type models the caudate AUCs varied with a standard deviation of 0.29, seven times their sampling error. The expansion-dependent contrasts were inconsistent: Q140 versus Q140 lacking Pms1 transferred to caudate (0.95, $p_{\mathrm{perm}} = 0.003$) but not to sorted MSNs (0.40), and the Msh3 version transferred to neither. All five control contrasts failed permutation, yet one with only 59 differentially expressed mouse genes reached a human AUC of 0.84, which a Mann–Whitney test over 70 human samples would call significant at $p = 7 \times 10^{-7}$. Of the 32 exploratory transfer tests, seven had $p_{\mathrm{perm}} < 0.05$ and two survived false-discovery correction: R6/2 to caudate ($q = 0.032$) and Q140 versus Q140 lacking Pms1 to caudate ($q = 0.048$). Transfer from zQ175, the MSN-versus-interneuron contrast and composition-invariant transfer reached $q = 0.13$–0.15 and remain suggestive.

4.8 Consensus training stabilises transfer and reaches projection neurons

Pooling the five cohorts changed the picture (Table 7, Figure 8a). Single-cohort models built on the same genes transferred to caudate with AUCs from 0.37 to 0.97 (standard deviation 0.25). The consensus model reached 0.94, and the five models that each left one cohort out stayed between 0.83 and 0.96 (standard deviation 0.05), a fivefold reduction in variability. Under the cohort-stratified permutation null the consensus transfer to caudate was the strongest in the study ($p_{\mathrm{perm}} = 0.001$; null 97.5th percentile 0.85).

In sorted nuclei, the consensus model separated HD from control donors perfectly using MSN nuclei (AUC 1.00, $p_{\mathrm{perm}} = 0.025$), the first single analysis of projection neurons to pass permutation, whereas it did not transfer to interneurons from the same donors (0.38, $p_{\mathrm{perm}} = 0.65$; Figure 8b). The MSN-minus-interneuron difference (+0.63) was borderline ($p_{\mathrm{perm}} = 0.052$). Composition still explained the bulk result: after adjustment for deconvolved proportions, the consensus model's caudate AUC fell to 0.48 ($p_{\mathrm{perm}} = 0.66$). Within this second family of eight tests, the consensus transfer to caudate ($q = 0.008$) survived correction and the MSN transfer was suggestive ($q = 0.067$).

4.9 A conserved per-cell signature that bulk tissue cannot reveal

Four hundred and ninety-three genes changed in the same direction in the mouse consensus and inside sorted human MSNs (Table 7). Across all 1,287 possible assignments of HD labels to the 13 donors, the median count was 43 and the 95th percentile 287, so the observed overlap was eleven times the chance expectation (exact $p = 0.012$, $q = 0.047$; Figure 8c). The agreement was concentrated in the strongly changed genes: across all genes with $|z_g| \geq 3$, the rank correlation between species was only 0.05 ($p = 0.13$). The signature contained 107 genes that fell and 386 that rose in disease. Those that fell included MSN and synaptic genes such as PDE10A, RGS9, ARPP21, PCP4, CNR1, DCLK1, GRIN2A, GRIK2 and CACNA2D3, and the most strongly shared also included PDE8B, PLCB1, DGKB and ATP2B1; those that rose included signalling and stress-response genes such as BCL6, SHC1, PEA15 and TYK2. The full list is provided as supplementary material.

In the independent bulk caudate cohort, the signature separated HD from controls with an AUC of 0.83, but once composition was removed its AUC was 0.55, no better than donor-permuted signatures ($p = 0.21$) or random genes ($p = 0.60$). Signatures of the same size built from the human sorted data alone or

from the mouse consensus alone fared no better (0.50 and 0.51). Per-cell changes that the two species share are therefore real and detectable in sorted cells, but they cannot be recovered from bulk tissue of this kind once cell loss is accounted for.

5 Discussion

Our confirmatory result agrees with earlier reports that mouse striatal signatures resemble human HD caudate [7–9], and the sorted-nucleus data agree that patients lose MSNs that mice keep [10]. The audit changes how that resemblance should be read. Transfer is confined to the most affected tissue in this early-stage cohort; its significance depends almost entirely on how it is tested; and in bulk caudate it reflects the depletion of projection neurons and the accompanying glial expansion, not a shared per-cell program.

The inflation index explains why the testing matters so much. When one biological axis dominates the target, as MSN loss dominates HD caudate, a classifier that lines up with that axis by chance will score well, and only a null built by retraining on shuffled source labels captures this. The parallel with random prognostic signatures in cancer, where proliferation plays the role that neuron content plays here, is close [17, 19]. The difference is practical: τ can be computed before any model is trained, so an analyst knows in advance whether a conventional p-value is worth reporting. In our simulations the warning sign was $\tau \geq 0.4$.

The per-cell signal that survives sorting is real but modest, and it arises in partly different ways in the two species. Knock-in mice lower MSN identity genes selectively, whereas human MSNs show a broader transcriptional decline. That fits evidence that human projection neurons keep near-normal expression until their repeat expands past about 150 CAGs, after which identity collapses shortly before death [4]; a sorted population averages over neurons at every stage of this process. The genetic benchmark adds a caution. A classifier that faithfully tracks expansion-dependent pathology across mouse cohorts, including its suppression by mismatch-repair knockouts, still transfers to humans for some cohorts and not others, and contrasts that barely change mouse expression can yield high human AUCs. Our data cannot show that the expansion-dependent component removed by Msh3 or Pms1 loss is what carries over to humans.

The consensus results point to a constructive remedy. Single cohorts of eight mice per group are noisy sources, and their noise, not just their biology, decides whether they transfer. Combining cohorts by meta-analysis, and keeping only genes whose direction agrees across them, removed most of that noise: transfer became stronger, five times more stable, and for the first time significant inside human projection neurons. The conserved per-cell signature shows what this transfer consists of. It is enriched for the synaptic and signalling genes that define projection neurons, it is shared far beyond chance between species within one cell type, and it is invisible in bulk tissue once cell loss is accounted for. Testing the fidelity of mouse models at the level of single cell types is therefore not a refinement but a requirement.

These findings bear on how mouse models are chosen and validated. A signature that replicates across mouse cohorts and responds to genetic modifiers is not, by that alone, a model of the human disease process, and agreement with human bulk tissue is especially hard to interpret when that tissue has lost the cells in which the mouse changes occur. Where human cell-resolved data exist, transfer should be tested inside the relevant cell type and against a spared one. Where they do not, deconvolution offers a conservative check. Because single cohorts of eight animals per group gave contradictory answers, conclusions about a model should rest on several independent cohorts, ideally combined by consensus training, with the spread between them reported. The same caution applies to translation methods in general [21, 22, 24]: neither the projection baseline nor domain adaptation escaped dominant-axis inflation, and more flexible models may fit such axes even more readily. Neuronal loss, gliosis and inflammation also dominate tissue in Alzheimer's and Parkinson's diseases, so we expect the problem,

and TransferAudit's usefulness, to extend well beyond HD. Table 8 lists the controls that changed our conclusions, as a checklist that complements TRIPOD+AI [33] for cross-species work.

The study has limitations. The confirmatory caudate cohort was profiled on arrays in 2006, and the sorted-nucleus and benchmark analyses rest on four or five HD donors and seven to eleven mice per group, so their AUCs are imprecise [32] and few of their tests withstand correction for multiple comparisons. The deconvolution reference covers five cell classes from nuclear RNA on a different platform and lacks endothelial and precursor cells. The analytic null is an approximation: conservative when signal is present, and mildly anti-conservative in some low-index real analyses. Ortholog definitions were varied only within one resource, and the mouse cohorts represent a few model families. The consensus and signature analyses inherit the small number of HD donors with sorted nuclei, and their eight tests were corrected as a separate family; the signature thresholds were fixed before the analysis but remain a choice. A recent atlas of early-grade HD striatum with paired single-cell repeat lengths [49], along with larger cohorts [4, 50], now allows external validation at donor level. Applying the audit separately to dorsal and ventral striatum, and to neurons with short and long expansions, would test whether mouse signatures track the transition that human projection neurons undergo.

6 Conclusions

Classifiers trained on HD mouse brain recognise HD in human caudate but not in cortex or cerebellum. In bulk tissue that success comes from the loss of projection neurons and the glial response; a weaker per-cell signal persists in sorted projection neurons but not interneurons, and the shared readout arises from selective identity loss in mice but a broader decline in humans. Conventional statistics overstate such transfer by orders of magnitude, and an inflation index computed before training tells the analyst when they will. Training on several mouse cohorts at once makes transfer stronger and far more stable, and reveals a conserved per-cell signature in human projection neurons that bulk tissue cannot show. TransferAudit makes these checks routine, and we recommend that cross-species studies report source-label permutation, random-gene comparisons, the inflation index, a composition analysis, multi-cohort training and, wherever possible, cell-resolved targets.

Statements and Declarations

Funding. No funds, grants, or other support were received.

Competing interests. The authors declare that they have no competing interests.

Ethics approval. This secondary analysis used publicly available, de-identified data only. No new participants or animals were recruited. Ethical approvals for original data collection are described in the original reports [37–39].

Consent to participate. Not applicable to this secondary analysis; consent for the original human studies was handled by the data-collecting investigators.

Consent for publication. Not applicable; no identifiable individual-level information is presented.

Data availability. Public transcriptomic datasets are available from Gene Expression Omnibus; accession numbers appear in Table 1.

Code availability. The TransferAudit Python package (with tests and a worked example) and the fifteen analysis scripts are provided as a supplementary archive and will be released in a public repository under an open-source licence on publication.

Author contributions. Abdul Ahad: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing, Supervision, Project administration. Nurjahan Begum: Validation, Writing – review & editing. Nusrat Jahan: Validation, Writing – review & editing. All authors read and approved the final manuscript.

Acknowledgements. Not applicable.

References

[1] The Huntington's Disease Collaborative Research Group (1993) 'A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes', *Cell*, Vol. 72, No. 6, pp.971–983.

[2] Vonsattel, J.-P., Myers, R.H., Stevens, T.J., Ferrante, R.J., Bird, E.D. and Richardson, E.P. (1985) 'Neuropathological classification of Huntington's disease', *Journal of Neuropathology and Experimental Neurology*, Vol. 44, No. 6, pp.559–577.

[3] Ross, C.A. and Tabrizi, S.J. (2011) 'Huntington's disease: from molecular pathogenesis to clinical treatment', *The Lancet Neurology*, Vol. 10, No. 1, pp.83–98.

[4] Handsaker, R.E. et al. (2025) 'Long somatic DNA-repeat expansion drives neurodegeneration in Huntington's disease', *Cell*, Vol. 188, No. 3, pp.623–639.

[5] Kashin, S., Lee, W.-S., McDonald, T.M. et al. (2026) 'Somatic DNA-repeat expansions initiate distinct yet analogous changes in Huntington's disease cortical and striatal neurons', *Cell* (published online 28 October 2026), https://doi.org/10.1016/j.cell.2026.10.001.

[6] Vazquez-Sanchez, S. et al. (2026) 'Multimodal spatial transcriptomics redefines cortical pathology in Huntington's disease', *Cell* (published online 28 October 2026), https://doi.org/10.1016/j.cell.2026.09.064.

[7] Kuhn, A., Goldstein, D.R., Hodges, A. et al. (2007) 'Mutant huntingtin's effects on striatal gene expression in mice recapitulate changes observed in human Huntington's disease brain and do not differ with mutant huntingtin length or wild-type huntingtin dosage', *Human Molecular Genetics*, Vol. 16, No. 15, pp.1845–1861.

[8] Neueder, A. and Bates, G.P. (2014) 'A common gene expression signature in Huntington's disease patient brain regions', *BMC Medical Genomics*, Vol. 7, p.60.

[9] Langfelder, P., Cantle, J.P., Chatzopoulou, D. et al. (2016) 'Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice', *Nature Neuroscience*, Vol. 19, No. 4, pp.623–633.

[10] Matsushima, A., Pineda, S.S., Crittenden, J.R. et al. (2023) 'Transcriptional vulnerabilities of striatal neurons in human and rodent models of Huntington's disease', *Nature Communications*, Vol. 14, p.282.

[11] Lee, H., Fenster, R.J., Pineda, S.S. et al. (2020) 'Cell type-specific transcriptomics reveals that mutant huntingtin leads to mitochondrial RNA release and neuronal innate immune activation', *Neuron*, Vol. 107, No. 5, pp.891–908.

[12] Wang, N. et al. (2025) 'Distinct mismatch-repair complex genes set neuronal CAG-repeat expansion rate to drive selective pathogenesis in HD mice', *Cell*, Vol. 188, pp.1524–1544.

[13] Aldous, S.G. et al. (2024) 'A CAG repeat threshold for therapeutics targeting somatic instability in Huntington's disease', *Brain*, Vol. 147, No. 5, https://doi.org/10.1093/brain/awae063.

[14] Varma, S. and Simon, R. (2006) 'Bias in error estimation when using cross-validation for model selection', *BMC Bioinformatics*, Vol. 7, p.91.

[15] Kapoor, S. and Narayanan, A. (2023) 'Leakage and the reproducibility crisis in machine-learning-based science', *Patterns*, Vol. 4, No. 9, p.100804.

[16] Leek, J.T., Scharpf, R.B., Bravo, H.C. et al. (2010) 'Tackling the widespread and critical impact of batch effects in high-throughput data', *Nature Reviews Genetics*, Vol. 11, No. 10, pp.733–739.

[17] Venet, D., Dumont, J.E. and Detours, V. (2011) 'Most random gene expression signatures are significantly associated with breast cancer outcome', *PLoS Computational Biology*, Vol. 7, No. 10, p.e1002240.

[18] Shimoni, Y. (2018) 'Association between expression of random gene sets and survival is evident in multiple cancer types and may be explained by sub-classification', *PLoS Computational Biology*, Vol. 14, No. 2, p.e1006026.

[19] Goh, W.W.B. and Wong, L. (2018) 'Why breast cancer signatures are no better than random signatures explained', *Drug Discovery Today*, Vol. 23, No. 11, pp.1818–1823.

[20] Tschodu, D., Lippoldt, J., Gottheil, P. et al. (2023) 'Re-evaluation of publicly available gene-expression databases using machine-learning yields a maximum prognostic power in breast cancer', *Scientific Reports*, Vol. 13, p.16402.

[21] Normand, R., Du, W., Briller, M., Gaujoux, R., Starosvetsky, E., Ziv-Kenet, A., Shalev-Malul, G., Tibshirani, R.J. and Shen-Orr, S.S. (2018) 'Found In Translation: a machine learning model for mouse-to-human inference', *Nature Methods*, Vol. 15, No. 12, pp.1067–1073.

[22] Brubaker, D.K., Proctor, E.A., Haigis, K.M. and Lauffenburger, D.A. (2019) 'Computational translation of genomic responses from experimental model systems to humans', *PLoS Computational Biology*, Vol. 15, No. 1, p.e1006286.

[23] Pullen, K.M. et al. (2024) 'Cross-species transcriptomics translation reveals a role for the unfolded protein response in Mycobacterium tuberculosis infection', *npj Systems Biology and Applications*, https://doi.org/10.1038/s41540-024-00487-6.

[24] Meimetis, N. et al. (2024) 'AutoTransOP: translating omics signatures without orthologue requirements using deep learning', *npj Systems Biology and Applications*, https://doi.org/10.1038/s41540-024-00341-9.

[25] Foster-Powell, A., Meno-Tetang, G., Rostami-Hodjegan, A., Ogungbenro, K. and Mager, D.E. (2025) 'Translatability of animal models for Alzheimer's disease using a machine learning based workflow', *Clinical and Translational Science*, Vol. 18, No. 11, e70387, https://doi.org/10.1111/cts.70387.

[26] Souza, H. and Mehta, P. (2026) 'Parameter-free representations outperform single-cell foundation models on downstream benchmarks', arXiv:2602.16696.

[27] Dash, N., Bashar, M.A., Lee, J. and Dash, R. (2025) 'Identification of novel biomarkers in Huntington's disease based on differential gene expression meta-analysis and machine learning approach', *Applied Sciences*, Vol. 15, No. 15, p.8286.

[28] Obenauer, J.C., Chen, J., Andreeva, V., Aaronson, J.S., Lee, R., Caricasole, A. and Rosinski, J. (2022) 'Expression analysis of Huntington disease mouse models reveals robust striatum disease signatures', bioRxiv, doi:10.1101/2022.02.04.479180.

[29] Ojala, M. and Garriga, G.C. (2010) 'Permutation tests for studying classifier performance', *Journal of Machine Learning Research*, Vol. 11, pp.1833–1863.

[30] Chaibub Neto, E. (2018) 'Using permutations to detect, quantify and correct for confounding in machine learning predictions', arXiv:1805.07465.

[31] van Klaveren, D. et al. (2016) 'Assessing discriminative performance at external validation of clinical prediction models', *PLoS One*, Vol. 11, No. 2, p.e0148820.

[32] Riley, R.D., Snell, K.I.E., Archer, L. et al. (2024) 'Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study', *BMJ*, Vol. 384, p.e074821.

[33] Collins, G.S. et al. (2024) 'TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods', *BMJ*, Vol. 385, p.e078378.

[34] Newman, A.M., Liu, C.L., Green, M.R. et al. (2015) 'Robust enumeration of cell subsets from tissue expression profiles', *Nature Methods*, Vol. 12, No. 5, pp.453–457.

[35] McKenzie, A.T., Wang, M., Hauberg, M.E. et al. (2018) 'Brain cell type specific gene expression and co-expression network architectures', *Scientific Reports*, Vol. 8, p.8868.

[36] Edgar, R., Domrachev, M. and Lash, A.E. (2002) 'Gene Expression Omnibus: NCBI gene expression and hybridization array data repository', *Nucleic Acids Research*, Vol. 30, No. 1, pp.207–210.

[37] Hodges, A., Strand, A.D., Aragaki, A.K. et al. (2006) 'Regional and cellular gene expression changes in human Huntington's disease brain', *Human Molecular Genetics*, Vol. 15, No. 6, pp.965–977.

[38] Labadorf, A., Hoss, A.G., Lagomarsino, V. et al. (2015) 'RNA sequence analysis of human Huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression', *PLoS One*, Vol. 10, No. 12, p.e0143563.

[39] Mätlik, K. et al. (2024) 'Cell-type-specific CAG repeat expansions and toxicity of mutant Huntingtin in human striatum and cerebellum', *Nature Genetics*, Vol. 56, pp.383–394.

[40] Eyre, T.A., Wright, M.W., Lush, M.J. and Bruford, E.A. (2007) 'HCOP: a searchable database of human orthology predictions', *Briefings in Bioinformatics*, Vol. 8, No. 1, pp.2–5.

[41] Yates, B., Gray, K.A., Jones, T.E.M. and Bruford, E.A. (2021) 'Updates to HCOP: the HGNC comparison of orthology predictions tool', *Briefings in Bioinformatics*, Vol. 22, No. 6, p.bbab155.

[42] Dolgalev, I. (2022) 'babelgene: Gene Orthologs for Model Organisms in a Tidy Data Format', R package version 22.9, CRAN.

[43] Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society: Series B*, Vol. 57, No. 1, pp.289–300.

[44] Breiman, L. (2001) 'Random forests', *Machine Learning*, Vol. 45, No. 1, pp.5–32.

[45] Pedregosa, F., Varoquaux, G., Gramfort, A. et al. (2011) 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research*, Vol. 12, pp.2825–2830.

[46] Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*, Chapman & Hall, New York, NY, USA.

[47] Sun, B., Feng, J. and Saenko, K. (2016) 'Return of frustratingly easy domain adaptation', *Proc. AAAI Conf. Artif. Intell.*, pp.2058–2065.

[48] Pressl, C., Mätlik, K., Kus, L. et al. (2024) 'Selective vulnerability of layer 5a corticostriatal neurons in Huntington's disease', *Neuron*, Vol. 112, No. 6, pp.924–941.

[49] Linville, R.M., James, B.T., Galani, K. et al. (2026) 'Cross-species single-cell atlas of the striatum defines cell-type and subregion disease vulnerabilities', *Cell*, Vol. 189, No. 20, pp.6490–6521.e19, https://doi.org/10.1016/j.cell.2026.08.006.

[50] Paryani, F. et al. (2024) 'Multi-omic analysis of Huntington's disease reveals a compensatory astrocyte state', *Nature Communications*, Vol. 15, p.6742.

Tables

Table 1: Datasets used in this study.

Accession	Species	Tissue or design	Platform	Samples
GSE65770	Mouse	Cortex, allelic series	RNA-seq	168
GSE65774	Mouse	Striatum, allelic series	RNA-seq	208
GSE3790	Human	Caudate; frontal cortex; cerebellum	Microarray (U133A)	70; 65; 66
GSE64810	Human	Prefrontal cortex (BA9)	RNA-seq	69
GSE227729	Human	Sorted striatal nuclei	RNA-seq	134 (14 donors)
GSE237571	Mouse	Striatum, Q140 × Msh3	RNA-seq	48
GSE288739	Mouse	Striatum, Q140 × Pms1	RNA-seq	48
GSE245510	Mouse	Striatum, zQ175 × Msh3	RNA-seq	32
GSE152058	Mouse	D1/D2 MSN translating RNA, R6/2	RNA-seq	22

Table 2: TransferAudit report and robustness checks for the confirmatory analysis (mouse striatum to human caudate). Permutation nulls use 1,000 runs unless stated.

Model or variant	Human AUC	p_{perm}
Primary: ANOVA-200 + logistic regression	0.87 (0.77–0.95)	0.018–0.035
Naive Mann–Whitney test (model treated as fixed)	–	7×10^{-8}
Inflation index $\tau = 0.69$; analytic p	–	0.016
Within-mouse cross-validation (same model)	1.00	–
Four training designs	0.73–0.93	–
Orthologs supported by $\geq$11 of 12 databases	0.84	0.061
Orthologs supported by 12 of 12 databases	0.88	0.015
Rank-based inverse-normal transform	0.95	0.001
Random forest (200 permutations)	0.85	0.085
Random 200-gene sets (mean of null)	0.70	–
Projection baseline, mouse components	0.83	0.30
Projection baseline, human components	0.61	0.32
CORAL domain adaptation	0.76	0.081
Adjusted for deconvolved proportions	0.47 (0.33–0.61)	0.77

Table 3: Transfer to three human regions from the same donors and to an independent prefrontal cohort (logistic regression, all eligible mice, Q20 controls). Upper bound: cross-validated AUC of a model trained within the human data; τ: inflation index; q: Benjamini–Hochberg over 32 exploratory tests.

Mouse	**Human**	**AUC**	p_{MW}	p_{perm}	τ	**Upper bound**	q
Striatum	Caudate	0.87	7×10^{-8}	0.018–0.035	0.69	0.96	confirmatory
Cortex	Caudate	0.88	4×10^{-8}	0.031	0.78	0.96	0.15
Striatum	Frontal cortex	0.50	0.51	0.52	0.24	0.75	0.67
Cortex	Frontal cortex	0.59	0.11	0.21	0.24	0.75	0.40
Striatum	Cerebellum	0.34	0.98	0.91	0.29	0.85	0.92
Cortex	Cerebellum	0.66	0.015	0.095	0.29	0.85	0.26
Cortex	Prefrontal (BA9)	0.63	0.051	0.29	0.37	0.93	0.43

Table 4: Deconvolved cell-type proportions in human caudate (mean per group).

Cell class	**Control**	**HD**
Medium spiny neurons	0.46	0.22
Interneurons	0.11	0.12
Astrocytes	0.26	0.42
Microglia	0.05	0.07
Oligodendrocytes	0.12	0.17

Table 5: Transfer inside sorted human striatal nuclei (donor level, logistic regression).

Target	**HD/control donors**	**AUC**	**Random forest**	p_{perm}	p_{rand}	τ
MSNs, all regions	5/8	0.95	1.00	0.051	0.006	1.06
MSNs, region-matched	5/8	0.80	1.00	0.13	0.021	1.20
D1 MSNs, region-matched	4/8	0.81	0.97	0.10	0.040	1.26
D2 MSNs, region-matched	5/7	0.83	1.00	0.11	0.009	1.10
Interneurons, all regions	4/8	0.28	0.88	0.76	0.092	0.76
Interneurons, region-matched	4/3	0.25	0.83	0.74	0.068	1.01
MSN minus interneuron	–	+0.67	–	0.027	0.14	–

Table 6: Mouse cohorts and genetic controls: human AUC with permutation p (1,000 runs). DE: differentially expressed mouse genes ($|t| > 3$). Cohort 1, Pms1 study; cohort 2, Msh3 study.

Mouse contrast	**DE**	**Caudate**	**Sorted MSNs**
zQ175 vs. wild type	2,786	0.94 (0.012)	0.97 (0.018)
R6/2 vs. wild type	2,947	0.97 (0.001)	0.80 (0.11)
Q140 vs. wild type, cohort 1	3,285	0.89 (0.061)	0.90 (0.086)
Q140 vs. wild type, cohort 2	2,465	0.36 (0.63)	0.75 (0.23)
Q140 vs. Q140 lacking Pms1	1,174	0.95 (0.003)	0.40 (0.62)
Q140 vs. Q140 lacking Msh3	1,780	0.47 (0.50)	0.70 (0.30)
zQ175 vs. zQ175 lacking Msh3 (control)	59	0.84 (0.20)	0.70 (0.27)
Wild type vs. Msh3 knockout (control)	65	0.75 (0.25)	0.62 (0.41)
Wild type vs. Pms1 knockout (control)	878	0.24 (0.74)	0.15 (0.92)

Table 7: Multi-cohort consensus transfer and the conserved per-cell signature. Permutation p: cohort-stratified (consensus, 1,000 runs) or exact over 1,287 donor labelings (signature); q: Benjamini–Hochberg within these eight tests.

Model or test	**Value**	p_{perm}	q
Transfer to human caudate (AUC)			
Single cohorts on the same genes (5 models)	0.37–0.97 (SD 0.25)	–	–
Consensus, each cohort left out (5 models)	0.83–0.96 (SD 0.05)	–	–
Consensus, all five cohorts	0.94	0.001	0.008
Consensus, adjusted for deconvolved composition	0.48	0.66	0.66
Transfer to sorted human nuclei (donor-level AUC)			
Consensus, MSNs	1.00	0.025	0.067
Consensus, interneurons	0.38	0.65	0.66
MSN minus interneuron	+0.63	0.052	0.10
Conserved per-cell signature			
Conserved genes (null median 43)	493	0.012	0.047
Rank concordance across species	0.05	0.13	0.20
Bulk caudate AUC, raw / composition-adjusted	0.83 / 0.55	0.21	0.28

Table 8: Reporting checklist for cross-species transfer studies, with the effect of each control in this study.

	Item
1	Report performance in the target species, not source cross-validation (1.00 here; batch-inflated 0.99 vs. 0.82).
2	Normalise each dataset separately and select features inside training folds.
3	Pre-specify one confirmatory test and correct the exploratory remainder (7 nominal → 2 significant).
4	Test against source-label permutation ($7 \times 10^{-8} \rightarrow 0.02$–0.04) and random gene sets (null mean 0.70).
5	Compute the inflation index before training; naive tests fail when $\tau \geq 0.4$.
6	Report a target-only upper bound and interpret low-signal tissues accordingly (0.75–0.96).
7	Audit source batches with control-versus-control contrasts (3,654 genes).
8	Deconvolve target composition with a cell-type reference and test sorted cell types (0.87 → 0.47; MSNs 0.95 vs. interneurons 0.28).
9	Replicate across independent source cohorts and include genetic negative controls (0.89 vs. 0.36; control 0.84).
10	Report every design, model, baseline and sensitivity analysis.

Figures

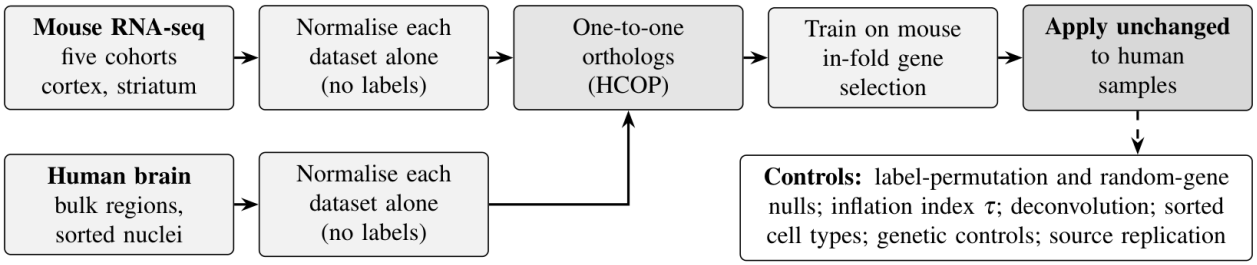

Figure 1: The TransferAudit workflow. Each dataset is normalised separately without labels and mapped through one-to-one orthologs; a classifier trained on source data alone is applied unchanged to the target, and the result is audited with the inflation index τ, source-label permutation and random-gene nulls, and a composition-adjusted null, followed by sorted-cell and source-replication checks.

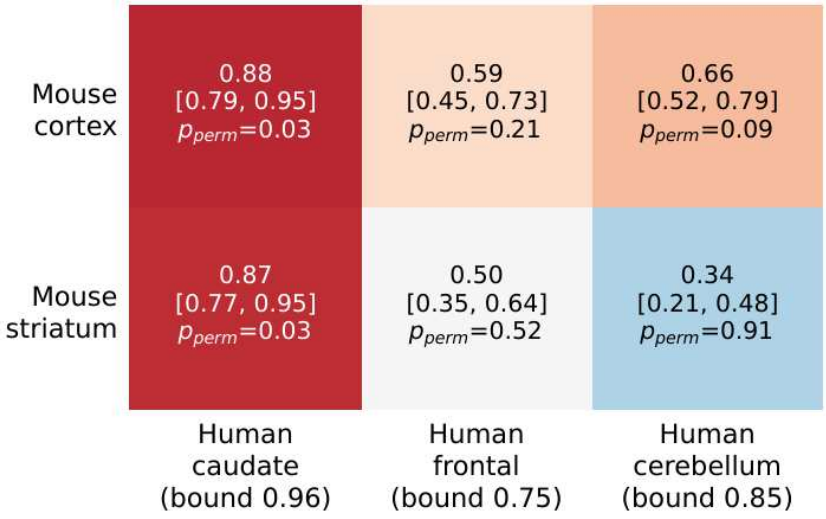

Figure 2: Transfer from two mouse regions to three human regions of the same donors (logistic regression). Each cell gives the human AUC, its 95% confidence interval and the label-permutation p-value; 'bound' is the AUC of a model trained within the human data.

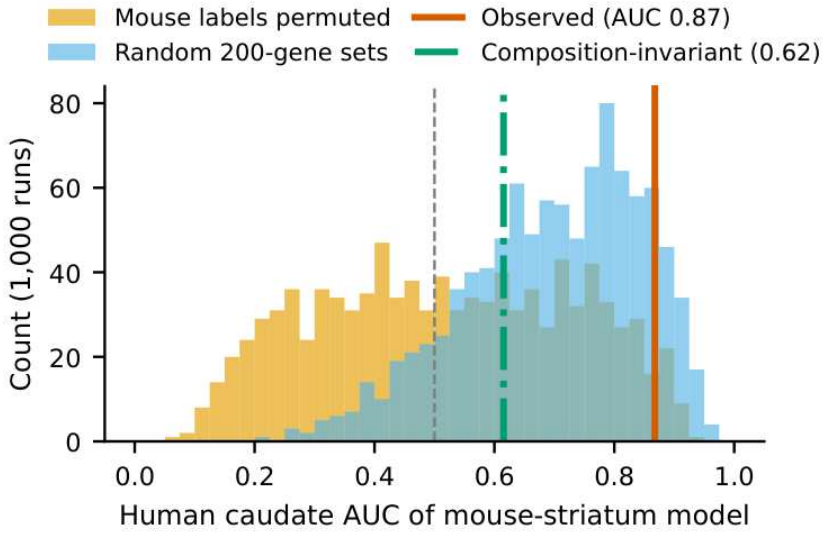

Figure 3: Null distributions for the confirmatory analysis: classifiers trained on permuted mouse labels and on random 200-gene sets, with the observed and composition-invariant AUCs.

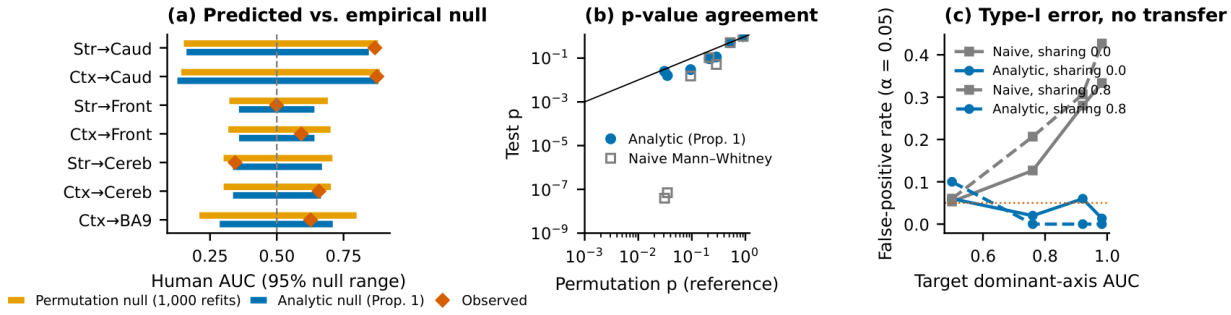

Figure 4: Analytic null for transferred AUC. (a) 95% null ranges from label permutation and from the analytic approximation. (b) Analytic and naive p-values against permutation p-values. (c) False-positive rates in simulations without transferable signal, with co-expression modules shared between species with probability 0 (solid) or 0.8 (dashed).

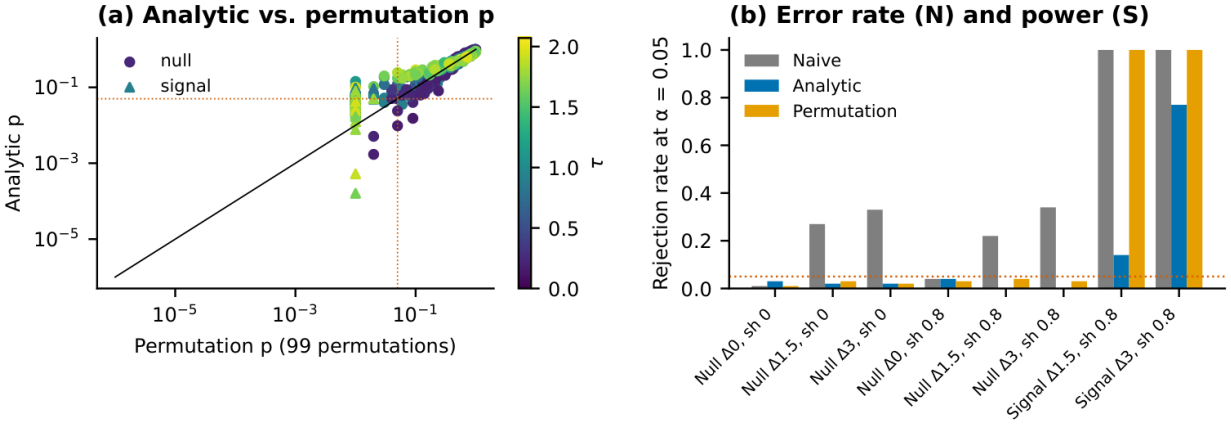

Figure 5: Calibration of the analytic test against full permutation in 800 simulated replicates. (a) Analytic against permutation p-values, shaded by the inflation index. (b) Rejection rates by method: false-positive rates without signal and power with signal.

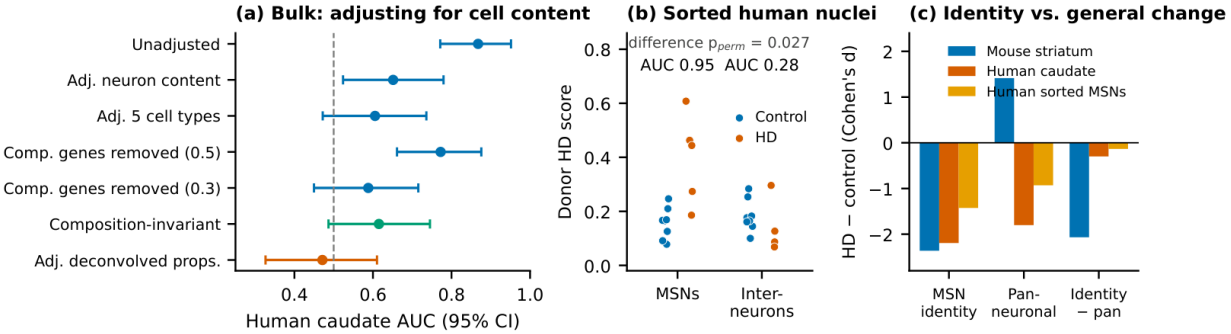

Figure 6: What the transferred signal consists of. (a) Caudate AUC after successive adjustments for cell composition, including reference-based deconvolution. (b) Donor-level scores in sorted human MSNs and interneurons. (c) Effect sizes of MSN identity and pan-neuronal scores in mouse striatum, human caudate and sorted human MSNs.

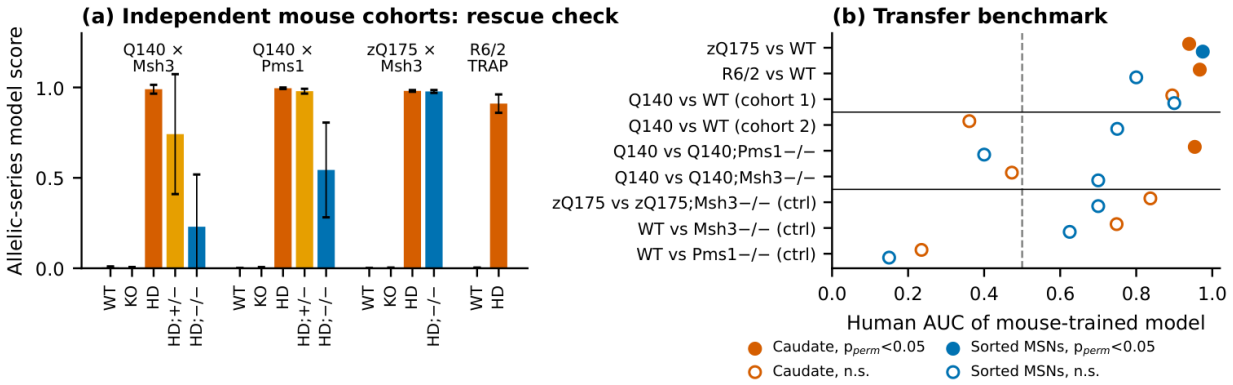

Figure 7: Independent mouse cohorts. (a) Scores of the primary classifier (mean ± standard deviation); HD;+/− and HD;−/− denote disease mice lacking one or both copies of Msh3 or Pms1. (b) Human AUC of classifiers trained on each mouse contrast; filled markers denote permutation $p < 0.05$.

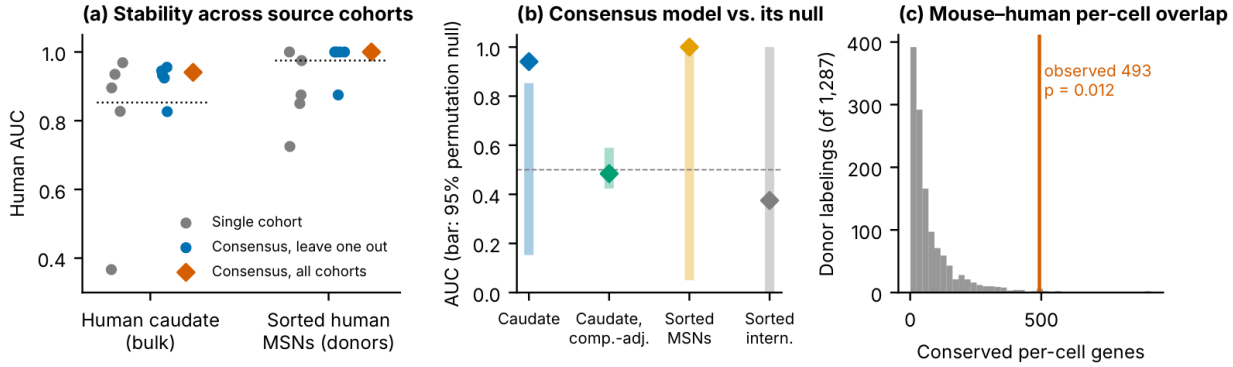

Figure 8: Multi-cohort consensus transfer. (a) Human AUC of single-cohort models, consensus models leaving one cohort out, and the full consensus model; dotted lines mark the 97.5th percentile of the consensus permutation null. (b) Consensus-model AUCs against their cohort-stratified permutation nulls (bars: central 95%). (c) Number of genes changed in the same direction in the mouse consensus and in sorted human MSNs, against the exact null over all 1,287 donor labelings.

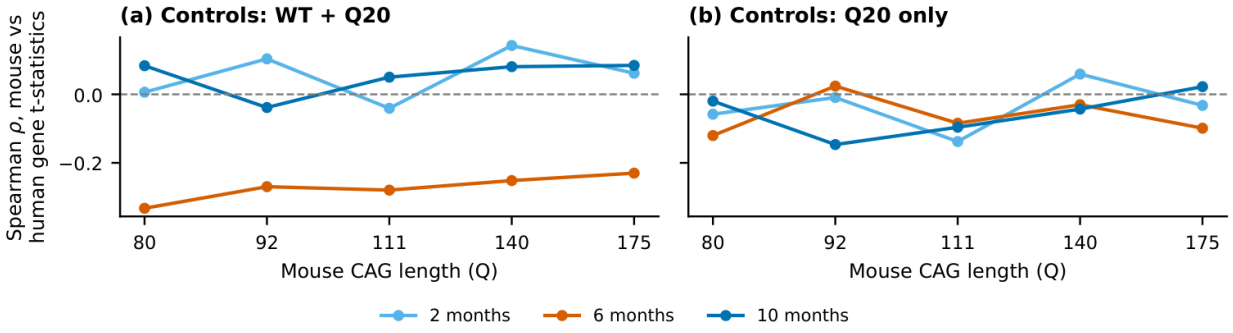

Figure 9: Batch audit of the cortex allelic series: correlation between mouse and human per-gene disease effects with (a) wild-type plus Q20 controls and (b) Q20-only controls.

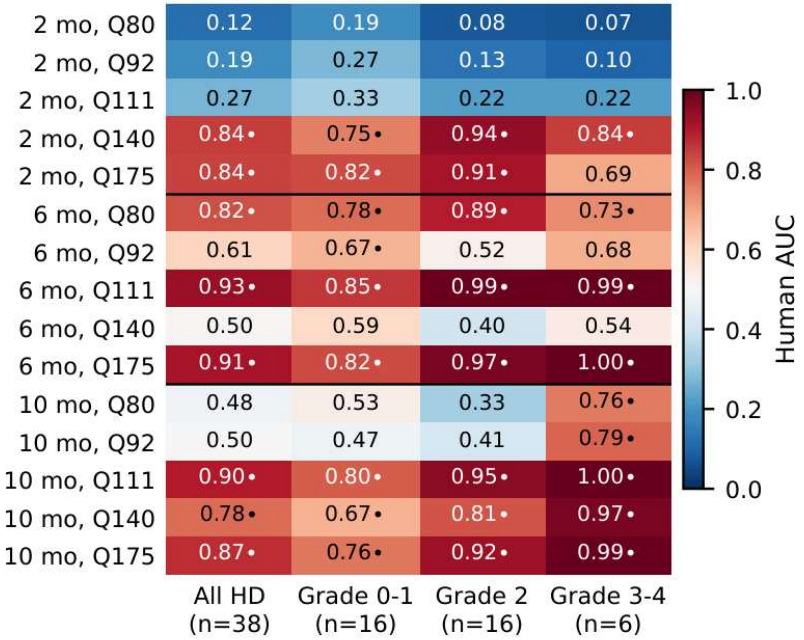

Figure 10: Cross-species staging map. Each row is a classifier trained on one mouse condition against Q20 mice of the same age; columns are human grade groups against all controls. Dots mark nominal Mann–Whitney $p < 0.05$ (34 of 60); no cell survived permutation testing with false-discovery control.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 03SupplementaryInformation.pdf