

Supplementary methods and results

Sections S1–S10 cover the original study and earlier experiments; S11–S17 cover the external study and follow-up analyses; S18 compares related designs; S19 assesses measured bulk against flow cytometry.

S1. Original study: data, annotation and eligibility

In both sources and all follow-ups, deposited labels were used to restrict target cells to the held-out donor and reference cells to other donors. Reuse across folds and targets creates dependence; label errors and cross-source donor aliases remain possible.

The original comparison used GSM2560248 from the control arm of GSE96583 [S1]. Cells from eight donors with systemic lupus erythematosus were cultured for six hours without interferon stimulation. “Control” refers to culture condition, not healthy donors. We used the deposited UMI counts and annotations without realigning reads or re-inferring genotypes or cell types.

We used GSM2560248_2.1.mtx.gz, GSM2560248_barcodes.tsv.gz, GSE96583_batch2.genes.tsv.gz and GSE96583_batch2.total.tsne.df.tsv.gz. Source URLs and hashes are recorded in reproduction_integrated/environment/geo_sources.json. The matrix contains 35,635 genes and 14,619 control barcodes. Barcode identity and order matched the annotations, whose barcode row-name column was read as the index. Ensembl identifiers were unique; duplicate gene symbols were not merged. Deposited processing used hg19 and Cell Ranger 1.2.

Filtering the deposited annotations to multiplets=singlet and stim=ctrl retained 12,315 cells before eligibility selection. The publication reported 12,138 control singlets; this discrepancy remains unresolved. Deposited donor IDs matched the authors’ stored first-choice IDs. Two singlets had an ambiguous BEST classification in the original output; both belonged to donor 1039, excluded below. These checks cannot establish error-free labels. The later formula correction [S2] did not change the tool implementation or resolve the singlet-count discrepancy.

Before PBMC predictions, eligibility required at least 50 cells in every retained donor–type group, at least four donors and at least three types. We maximized the number of donor–type entries, breaking ties by more donors and then lexicographic IDs. The unique maximum comprised six donors, six types and 11,066 cells. Donors 107 and 1039, dendritic cells and megakaryocytes were excluded from the simulation but retained in the source files. Simulated proportions therefore describe the six included types.

Table S1. Eligible cell inventory. Numbers are cells labelled as singlets in the current deposited annotations. Mono., monocytes; NK, natural killer.

Donor	B	CD14+ Mono.	CD4 T	CD8 T	FCGR3A+ Mono.	NK
101	107	202	293	74	80	70
1015	453	779	848	179	252	177
1016	126	369	422	536	115	128
1244	116	415	1,068	58	56	105
1256	216	390	1,001	131	69	261
1488	218	295	1,182	58	101	116

S2. Original study: target construction and exact randomization

Donor order was 101, 1015, 1016, 1244, 1256, 1488; cell-type order was B cells, CD14+ Monocytes, CD4 T cells, CD8 T cells, FCGR3A+ Monocytes, NK cells. Both orders are fixed in the input manifest. For each held-out donor, references used all ten three-donor combinations of the remaining five donors, in lexicographic order.

The 60 target compositions were generated once with NumPy `default_rng(2026091901)`. A vector with six equal probabilities was followed by 59 draws from `Dirichlet(1,1,1,1,1,1)`. Multinomial(300, probabilities) generated cell counts for each vector; the first count vector was then set to exactly 50 cells per type. The same 60 count-composition vectors were used for all six held-out donors. Actual cell draws used separate `default_rng(2026091901 + 1000 + fold)` generators, with zero-based fold indices. Within each type, cells were sampled with replacement from the held-out donor. Thus target compositions were paired across folds, whereas target cell identities differed. Zero counts for a type were allowed.

Each target summed raw counts from its 300 saved cell selections, including repeated cells. Ground truth was the realized count of each cell type divided by 300. All 60 targets per donor entered each `music_prop` call together, giving 360 target columns with library totals of 301,568–750,343. Every allocation and both outputs reused these targets.

Earlier v3.1 targets averaged cells after dividing each cell by its complete-gene library total and multiplying by 10,000. The present targets sum original counts from the same selected cells, so cells with different library sizes contribute differently. Shared cell IDs and composition truth do not make absolute MAEs comparable across these experiments. Both constructions produce synthetic mixtures.

S3. Original study: paired references and experimental size

Each reference contained 60 cells per type, hence 360 cells in total, from the same three donors. Allocation was either 20:20:20 or 50:5:5, with each reference donor taking the dominant position. Reference cells were selected without replacement within each donor–type pool.

The three reference seed blocks were 2026091902, 2026091903 and 2026091904. The original v3.1 generator was `default_rng(SeedSequence([block_seed, fold, triple_id, repeat]))`. For the MuSiC comparison we used `repeat=0` from each block, fixed in the protocol before empirical MuSiC predictions. Within a draw, every donor–type pool was permuted once. Each allocation used the required prefix of the same permutation, producing paired selections whose donor/type prefixes were nested; complete reference sets were not nested. All three blocks and all dominant-donor choices were retained, without choosing draws by their errors.

The endpoint design used 6 held-out donors $\times$ 10 triples $\times$ 3 draws $\times$ 4 allocations per draw (one balanced and three dominant-donor choices), giving 720 reference configurations. Each was applied to 60 targets with two outputs, yielding 86,400 prediction rows. Unequal allocations shared the balanced baseline within each donor–triple–draw; reusing it in paired contrasts does not add independent information.

S4. Original study: official MuSiC computation

We used MuSiC 1.0.0 at official commit `f21fe67f5670d5e9fca0ad7550abaae3423eb59c` [S3], through the flat `music_prop` workflow. Full original-gene reference counts entered a `SingleCellExperiment`. Parameters were `markers=NULL`, `clusters="cell_type"`, `samples="donor"`, the fixed six `select.ct` labels, `cell_size=NULL`, `ct.cov=FALSE`, `iter.max=1000`, `nu=0.0001`, `eps=0.01`, `centered=FALSE` and `normalize=FALSE`. We did not run the recursive workflow or tune the defaults.

Let $c(g,i)$ be the original count for gene g in reference cell i , and $L(i)$ its complete-gene count total. For cell type k and donor d , MuSiC constructs its basis from:

- $\theta(g, k, d) = \frac{\sum_i c(g, i)}{\sum_i L(i)}$, summing cells in that donor-type group;
- $S(k, d) = \frac{\sum_i L(i)}{n(k, d)}$;
- the equal-donor mean $\bar{\theta}(g, k)$, mean $\bar{S}(k)$, and sample variance $\Sigma(g, k)$ of $\theta(g, k, d)$ across the three reference donors;
- $D(g, k) = \bar{\theta}(g, k) \times \bar{S}(k)$.

Both MuSiC outputs use the factorized donor-equal basis D and the same estimated cell sizes. Increasing a donor's cell quota does not increase its coefficient in the donor mean. `Est.prop.allgene` normalizes ordinary NNLS coefficients; `Est.prop.weighted` uses iterative weighted NNLS with gene weights

$\frac{1}{v + r(g)^2 + \sum_k (q(k)S(k))^2 \Sigma(g, k)}$. Here q denotes unnormalized fitted coefficients and r denotes gene-wise residuals. Initial weights use ordinary-fit residuals, and later weights use the preceding weighted fit. We used the unchanged implementation. Iteration stops when the summed absolute change in normalized proportions falls below 0.01 or reaches the iteration limit; convergence refers to this stopping rule.

MuSiC removes genes with zero reference expression, intersects the remaining genes with those expressed in the target batch, calculates target relative abundance over these common genes, and removes zero-expression genes separately for each target fit. With `normalize=FALSE`, `music.iter` also multiplies target relative abundance by 100. Target-specific support contained 7,921–9,417 genes. Both outputs share these genes within a fit, but allocation can change which genes are retained. Allocation also changes D , S , Σ and residual weights, so the comparison measures their combined effect.

S5. Original study: outcomes and aggregation

For a six-type prediction $\hat{p}$ against sampled proportions p , we used:

- $MAE = \frac{100}{6} \sum_k |\hat{p}(k) - p(k)|$;
- $RMSE = 100 \times \sqrt{\frac{1}{6} \sum_k (\hat{p}(k) - p(k))^2}$.

Both errors are reported in percentage points. RMSE was calculated for each prediction before averaging; it was not derived from squared errors pooled across predictions.

Errors were averaged over 60 targets and three draws for each held-out donor, triple, allocation and output. Each unequal allocation was paired with the same triple's balanced baseline; contrasts were then averaged equally over ten triples, three dominant-donor choices and six held-out donors. A positive weighted-minus-ordinary difference means that the signed allocation contrast is larger for weighted MuSiC, not that its absolute error is larger.

Conditional SD was the sample SD (ddof=1) of predicted proportions across three draws for each fixed donor, triple, allocation, target and type. We averaged these SDs over types and targets, then arrangements and donors. This quantity differs from MuSiC's `Var.prop` and does not measure uncertainty in paired MAE contrasts. Its precision is limited by the three draws. Shared references, compositions and cells also prevent treating prediction rows or arrangements as independent biological replicates; no population tests or confidence intervals were calculated.

Table S2. MAE summaries for all six held-out donors, in percentage points. B denotes balanced allocation and I denotes 10:1:1 allocation. Weighted Δ is I–B; the last column subtracts the ordinary NNLS contrast from the weighted contrast. Calculations use unrounded values, and the aggregate gives each donor equal weight.

Held-out donor	NNLS B	NNLS I	Weighted B	Weighted I	Weighted Δ	Difference in differences
101	9.324	9.431	5.138	5.867	0.729	0.622
1015	8.693	8.984	5.604	6.468	0.864	0.573
1016	10.061	9.871	6.220	7.065	0.845	1.036
1244	8.841	8.907	5.790	6.400	0.610	0.544
1256	8.832	9.221	5.666	6.454	0.788	0.400
1488	8.908	8.950	5.384	5.914	0.530	0.488
Equal-donor aggregate	9.110	9.227	5.634	6.361	0.728	0.610

Ordinary NNLS had a negative allocation contrast for donor 1016 (−0.191 points), although weighted MAE was lower under both allocations. Weighted-minus-ordinary differences in allocation contrasts were negative in 38/180 arrangements and positive in 142/180. Seventeen of 18 donor–seed-block differences were positive; the remaining difference was −0.00588 points. Weighted allocation penalties were positive in all 18 blocks. Weighted conditional SD decreased for two donors despite increasing overall.

Ordinary RMSE was 11.566 under balanced and 11.625 under unequal allocation; weighted RMSE was 7.106 and 7.941. Full donor, arrangement, block and type-specific summaries are retained in results/music_review/ (Table S9).

S6. Original study: computational verification

Synthetic-mixture controls, source matching and traced/untraced comparisons passed. All 43,200 weighted fits reached the default stopping criterion, and errors were recalculated for all 86,400 predictions. A separate code path reconstructed counts and selections and checked 12 bases, 720 ordinary predictions and 36 weighted predictions. Maximum proportion differences in these checks were 3.07×10^{-14} and 6.99×10^{-15} .

The recorded environment used Python 3.12.14, R 4.5.2 and MuSiC 1.0.0 on Windows. Package details are retained with the analysis resources (Table S9). Cross-platform reproduction has not been demonstrated.

S7. Exploratory history and contrary evidence

The original MuSiC comparison followed experiments with normalized-cell targets and marker panels (Table S3). Its endpoint protocol was fixed before MuSiC predictions, but was not publicly preregistered. Earlier pancreas analyses used four GSE84133 donors [S4] and five eligible types. Their marker ranking used training-donor cells beyond the 60-cell signature budget, so that budget did not represent total data acquisition. These results were not pooled with the raw-count analysis.

Table S3. Earlier experiments and their implications. In the first four rows, equal and pooled refer to earlier reference averaging rules, not the two official MuSiC outputs.

Stage	Design	Result	Interpretation
Pancreas pilot	Four donors; 30,240 predictions; normalized NNLS	Strong-allocation MAE was 5.786 pooled versus 5.438 equal; 4/12 dominant-donor arrangements worsened. SD was a post hoc diagnostic.	Equal-donor averaging did not improve accuracy in every arrangement.
Pancreas v2	Four panels, two solvers, 24 draws; 645,120 predictions	Top100 NNLS equal-minus-pooled MAE was -0.285 points, but removing five abundant genes reversed it to $+0.182$. The reversal occurred in all three seed blocks. Across eight settings, conditional SD differences were $+0.993$ to $+1.907$.	The direction of the accuracy difference depended on gene selection.
External PBMC v3 failure and v3.1	Six donors; two panels/two solvers; full v3.1 grid of 2,419,200 predictions	Initial SLSQP oracle discrepancy 6.722×10^{-5} exceeded 5×10^{-5} . Uniformly tightening ftol from 10^{-12} to 10^{-14} yielded a complete passing run; scientific design and acceptance limits stayed fixed. Top100 NNLS MAE improved 0.433 points while conditional SD increased 1.694.	The initial failed run was retained. The pancreas gene-removal reversal was not reproduced across the four PBMC settings.
Finite-population calibration	Reconstructed saved v2 reference signatures; no new proportion predictions	Strong-allocation equal/pooled signature-variance trace ratio was 3.466 theoretically and 3.411 empirically.	The signature-variance result was consistent with finite-population sampling; its contribution to variation in estimated proportions was not quantified.
Official MuSiC follow-up	Same saved PBMC cell IDs; raw counts; 720 references; 86,400 predictions	Weighted MAE $5.634 \rightarrow 6.361$; ordinary MAE $9.110 \rightarrow 9.227$; difference in differences $+0.610$ points. Weighted absolute error and conditional SD remained lower.	Tests donor allocation within the existing MuSiC workflow; the comparison does not identify a mechanism or a generally optimal allocation.

In the calibration, independent sampling within each donor-type inventory gives $\text{Var}(\text{mean}) = (1/n - 1/N)S^2$. For reference weights $w(d)$, the signature-coordinate variance is $\sum_d w(d)^2(1/n(d) - 1/N(d))S(d)^2$. Summing coordinates gives a covariance trace without requiring independence between genes. This concerns normalized reference signatures, not the variance of nonlinear estimated proportions. This calibration led us to narrow the question from equal-donor instability to allocation effects in the MuSiC workflow.

S8. Reproduction resources

Code, processed inputs, saved predictions and full-precision results are archived at <https://doi.org/10.5281/zenodo.22902701>. Extract both ZIP files into one directory; see the README for commands and Table S9 for resource locations.

A compact code and result collection is available at <https://github.com/ReInhard-1/bioinformatics-paper> (release v1.1.0). Diagnostic resources retain selected-target outputs, summary data and control raw weights; a validated portable diagnostic refit and all production raw weight vectors are not provided.

S9. Intermediate fixed-budget allocations

After inspecting the endpoints, we specified the intermediate-allocation protocol and added 30:15:15 and 40:10:10 before generating new predictions. The six donors, six types, $B=60$, ten triples, three saved draws, dominant-donor rotations and 360 targets were retained. Each donor–type’s ordered 50-cell prefix was reconstructed from its original dominant reference; shorter prefixes were checked against saved selections. All 540 ratio4 selections matched the earlier cell lists. Whole references were not nested across ratios because increasing one donor’s quota reduced the others’ quotas.

The two intermediate allocations added 1,080 references and 129,600 predictions, alongside the reused 86,400 endpoint predictions. All 64,800 new weighted fits met the default stopping criterion; effective support ranged from 7,945 to 9,416 genes. Original-case replay and traced/untraced controls matched exactly.

An independent reconstruction checked reference selections and target truths. Twelve full-target replays reproduced 1,440 predictions, with maximum proportion and MAE/RMSE differences of 5.55×10^{-16} and 5.00×10^{-16} , respectively.

Table S4. All held-out donor MAEs across four allocations (percentage points). Each cell averages all reference triples, draws, targets and dominant choices where applicable.

Output	Donor	20:20:20	30:15:15	40:10:10	50:5:5
Weighted	101	5.138	5.191	5.358	5.867
Weighted	1015	5.604	5.641	5.893	6.468
Weighted	1016	6.220	6.337	6.543	7.065
Weighted	1244	5.790	5.844	5.949	6.400
Weighted	1256	5.666	5.765	5.928	6.454
Weighted	1488	5.384	5.404	5.499	5.914
Ordinary	101	9.324	9.380	9.371	9.431
Ordinary	1015	8.693	8.669	8.694	8.984
Ordinary	1016	10.061	10.122	9.865	9.871
Ordinary	1244	8.841	8.828	8.791	8.907
Ordinary	1256	8.832	8.833	9.044	9.221
Ordinary	1488	8.908	8.988	9.118	8.950

Across the four allocations, weighted and ordinary MAE profiles were nondecreasing for six and one of the six donor means, respectively. The corresponding counts were 16 and two of 18 donor–block means, and 95 and 31 of 180 arrangements. Weighted MAE decreased at successive allocation steps in 56, 34 and two arrangements; ordinary MAE decreased in 79, 91 and 84. No ties occurred at a tolerance of 10^{-10} percentage points. Complete paired and adjacent contrasts are in results/music_gradient_review/.

The smallest eligible donor–type inventory was 56 FCGR3A+ monocytes from donor 1244. The 10:1:1 allocation requires at least 5B/6 cells in every group when all dominant choices are tested without replacement. Requiring integral counts for all four allocations makes B divisible by 12; the largest common supported budget in this cohort was therefore 60.

S10. Balancing donors by discarding reference cells

A separately specified exploratory analysis retained the first five cells from each donor and type in the original 50:5:5 references. This produced 5:5:5 references, discarded 75% of cells and reduced B from 60 to 15. The three dominant parent allocations shared the resulting subset, giving 180 distinct references and 21,600 predictions. Thinning 50:5:5 cannot produce 20:20:20 without adding cells from the minor donors.

Weighted MAE increased from 6.361 to 6.908 points (+0.547), with increases for all six donor averages and 179 of 180 arrangements. One arrangement improved. Ordinary MAE decreased from 9.227 to 9.181 (−0.046), with three donors improving and three worsening. All 10,800 new weighted fits converged under unchanged defaults. A separate untraced replay checked 720 predictions, with maximum proportion difference 5.55×10^{-16} .

Thinning changes total reference size, profiles, estimated cell sizes, variance and gene support together. Earlier work balanced cell types by downsampling [S5]; this analysis balances donors within each type while reducing reference size. It therefore tests a different choice from redistributing cells at fixed B.

S11. External source, selection and fixed design

The external analysis used the public CELLxGENE release associated with Perez et al. [S6]: collection 436154da-bcf1-4130-9c8b-120ff9a888f2, collection version 0340d8f6-a8ca-4ad8-8664-07338a8bf55f, dataset 218acb0f-9f2f-4f76-b90b-15a4b7c7f629, H5AD object c55dc602-d168-4d15-acc1-5de4f2f5d551. The release links GSE174188 and healthy-source GSE137029; it is not a count-matrix attachment downloaded directly from GSE174188.

The metadata contain 1,263,676 cells, 261 donor labels (162 SLE and 99 normal) and 30,172 unique Ensembl identifiers. Screening all eight original-source donor labels found matches for 1015, 1039 and 1256 among SLE donors; these were excluded. Label matching cannot establish participant identity or rule out aliases. Before examining predictions, we selected all 14 normal/na donors with at least 250 cells in every retained type. Each has one sample and suspension identifier; libraries pool donors and are not biological replicates. Some donors span processing cohorts. Selection did not use expression similarity or outcomes.

Author labels B, cM, T4, T8, ncM and NK were mapped, in that order, to B cells, CD14+ monocytes, CD4 T cells, CD8 T cells, FCGR3A+ monocytes and NK cells. They are corresponding broad annotations, not proven identical truth across sources. The 2024 source erratum corrected swapped cDC1/cDC2 labels and revised data-processing documentation [S7]. These dendritic subtypes are excluded here; the retained broad labels were used as deposited. We reviewed the erratum and the original Methods summary. The revised supplementary PDF was unavailable for detailed preprocessing comparison.

We used raw.X, retained all 30,172 Ensembl identifiers and full count totals, and did not merge symbols or intersect genes across studies. The initial selected-cell union contained 93,892 cells and 75,749,702 nonzero counts. Selected counts were finite, nonnegative integers; sorting sparse indices preserved values and removed no entries. Counts ranged from 1 to 767 and cell totals from 498 to 20,320. Exported matrices were checked entrywise, all 840 targets reconstructed, and 15 selected rows independently reread from the H5AD. These checks cover the selected subset; no whole-H5AD SHA256 was obtained.

Table S5. Design differences between the two sources. Cell counts refer to the saved analysis inputs, not the original publications' cohort sizes. Both initial analyses use six retained types, three reference draws and 60 target compositions per donor. Differences in error between sources cannot isolate disease or processing effects.

Source	Eligible donors / selected cells	Genes supplied	Reference triples per held-out donor	Budgets per type	Within-target cell sampling	Reference draws
Original cultured SLE	6 / 11,066	35,635	All 10 of the other 5 donors	60	With replacement within type	Fold/triple-specific paired prefixes
External normal/na	14 / 93,892 selected	30,172	Four predefined triples of 12 of the other 13 donors	60, 120, 300	Without replacement within each target	Donor/type/block prefixes reused across folds/triples

The protocol analysis/MUSIC_EXTERNAL_PROTOCOL.md was frozen before external predictions, after the original results and metadata audit. A deterministic donor ring sorted labels by ascending SHA256 of UTF-8 `music-external-v1|ring|{donor}`. Its order was HC-502, IGTB256, IGTB514, IGTB1506, IGTB508, IGTB1650, IGTB195, IGTB1913, HC-572, HC-014, IGTB1921, IGTB826, IGTB1966, IGTB498. For held-out position i , the next 12 cyclic donors formed four consecutive triples; the thirteenth was omitted. Each donor was omitted from one other fold and served as a reference donor in 12 folds. This tests 14 unique triples reused across folds, not all possible triples.

For each source donor/type/block (blocks 0–2), sorted H5AD row indices were permuted and the first 250 retained. NumPy Generator(PCG64(seed)) used a seed obtained as the unsigned big-endian integer of the first 16 SHA256 bytes of UTF-8 `music-external-v1|reference|{donor}|{author_type}|{block}`. The exact prefixes were reused wherever the group appeared. At B=60 the four allocations were 20:20:20, 30:15:15, 40:10:10 and 50:5:5; at B=120 and 300 they were multiplied by two and five. All three dominant donors were rotated. Thus references contained 360, 720 or 1,800 cells, with nested selections across budgets for a fixed allocation. Whole references were not nested across allocation levels.

Only the 60 original composition vectors were reused, not original cells or expression. Their largest required cell numbers were 179, 159, 201, 185, 201 and 200 by type, below the external minimum inventory of 250. For donor/type/target, a fresh permutation used the analogous seed key `music-external-v1|target|{donor}|{author_type}|{target_index}`, selecting the required prefix without replacement within a target. Cells could recur in different targets or in reference/target roles across distinct folds. Each fit excluded its target donor from the reference. All budgets and allocations shared the same saved 840 raw-count target sums and truths.

The external grid comprised 5,040 logical held-out-donor/reference configurations and 1,260 unique reference matrices. Each matrix was evaluated against four eligible target donors, producing 604,800 prediction rows across both outputs. Unequal allocations shared their matched balanced baseline. The cohorts were analyzed separately.

Each external fold covered four of 286 possible triples among the 13 non-target donors, whereas the original cohort covered all ten possible triples per fold. Donor summaries share reference sources. We did not independently resample those summaries for intervals or tests because doing so would omit this dependence.

S12. External outcomes and reversals

MAE, RMSE and conditional SD follow S5. Means give equal weight to targets, three draws, four triples, dominant-donor choices where applicable and finally 14 donors. Each balanced baseline matches the target, triple, block and budget of its unequal allocation. $I(B)=MAE(10:1:1)-MAE(balanced)$ is signed; the primary

external contrast, weighted $I(300)-I(60)$, was fixed before predictions. Secondary summaries cover other allocations, budget and output differences, cell-type errors and reversals.

Table S6. Weighted allocation contrasts for every external donor, in percentage points and frozen ring order. Positive $I(B)$ indicates higher MAE under 10:1:1 allocation; negative $I(B)$ indicates lower MAE. A negative budget difference means that the signed contrast decreased, not necessarily its absolute magnitude or the prediction error.

Donor	$I(60)$	$I(120)$	$I(300)$	$I(300)-I(60)$
HC-502	+0.962	+0.484	-0.047	-1.009
IGTB256	+0.845	+0.458	-0.013	-0.858
IGTB514	+0.861	+0.657	+0.370	-0.490
IGTB1506	+0.655	+0.094	-0.388	-1.043
IGTB508	+0.829	+0.406	+0.049	-0.780
IGTB1650	+0.972	+0.517	+0.243	-0.729
IGTB195	+0.780	+0.280	-0.203	-0.983
IGTB1913	+0.955	+0.722	+0.443	-0.511
HC-572	+0.059	-0.009	-0.130	-0.189
HC-014	-0.434	-0.346	-0.216	+0.218
IGTB1921	+0.855	+0.302	-0.191	-1.045
IGTB826	+0.446	-0.048	-0.706	-1.152
IGTB1966	+0.224	-0.087	-0.073	-0.297
IGTB498	+0.690	+0.425	-0.062	-0.752

The primary contrast averaged -0.687 points and was negative for 13/14 donors. HC-014 had a positive budget change while $I(B)$ remained negative. At $B=60$, 120 and 300, weighted $I(B)$ was positive for 13/14, 10/14 and 4/14 donors; ordinary $I(B)$ was positive for 10/14, 7/14 and 7/14. Weighted-minus-ordinary contrasts averaged $+0.454$, $+0.209$ and -0.036 points, with positive differences for 12/14, 11/14 and 5/14 donors. The more negative weighted mean at $B=300$ does not imply a smaller absolute effect.

Weighted MuSiC had lower MAE in all 168 donor $\times$ budget $\times$ allocation comparisons. These differ from the 168 donor $\times$ reference-triple $\times$ dominant-donor arrangements at each fixed budget, of which 84/168 still had positive weighted contrasts at $B=300$. Across $B=60$, 120 and 300, weighted MAE profiles were nondecreasing across all four allocation ratios for 10/14, 7/14 and 3/14 donor means; 81/168, 51/168 and 30/168 arrangements; and 30/42, 22/42 and 7/42 donor-block summaries. Ordinary NNLS counts were 4/14, 3/14 and 2/14; 19/168, 25/168 and 14/168; and 7/42, 10/42 and 8/42, respectively.

Weighted budget contrasts for 120-60 and 300-120 averaged -0.346 and -0.342 points and were negative for 13/14 and 12/14 donors. The ordinary 300-60 contrast averaged -0.198 points and was negative for 10/14 donors. Weighted RMSE contrasts between 10:1:1 and balanced allocation were $+0.713$, $+0.344$ and -0.119 points. At $B=300$, the corresponding cell-type MAE contrasts were B $+0.054$, CD14+ $+0.029$, CD4 T -0.424 , CD8 T -0.160 , FCGR3A+ -0.016 and NK $+0.122$. Complete CSVs retain all cell-type and intermediate-allocation contrasts.

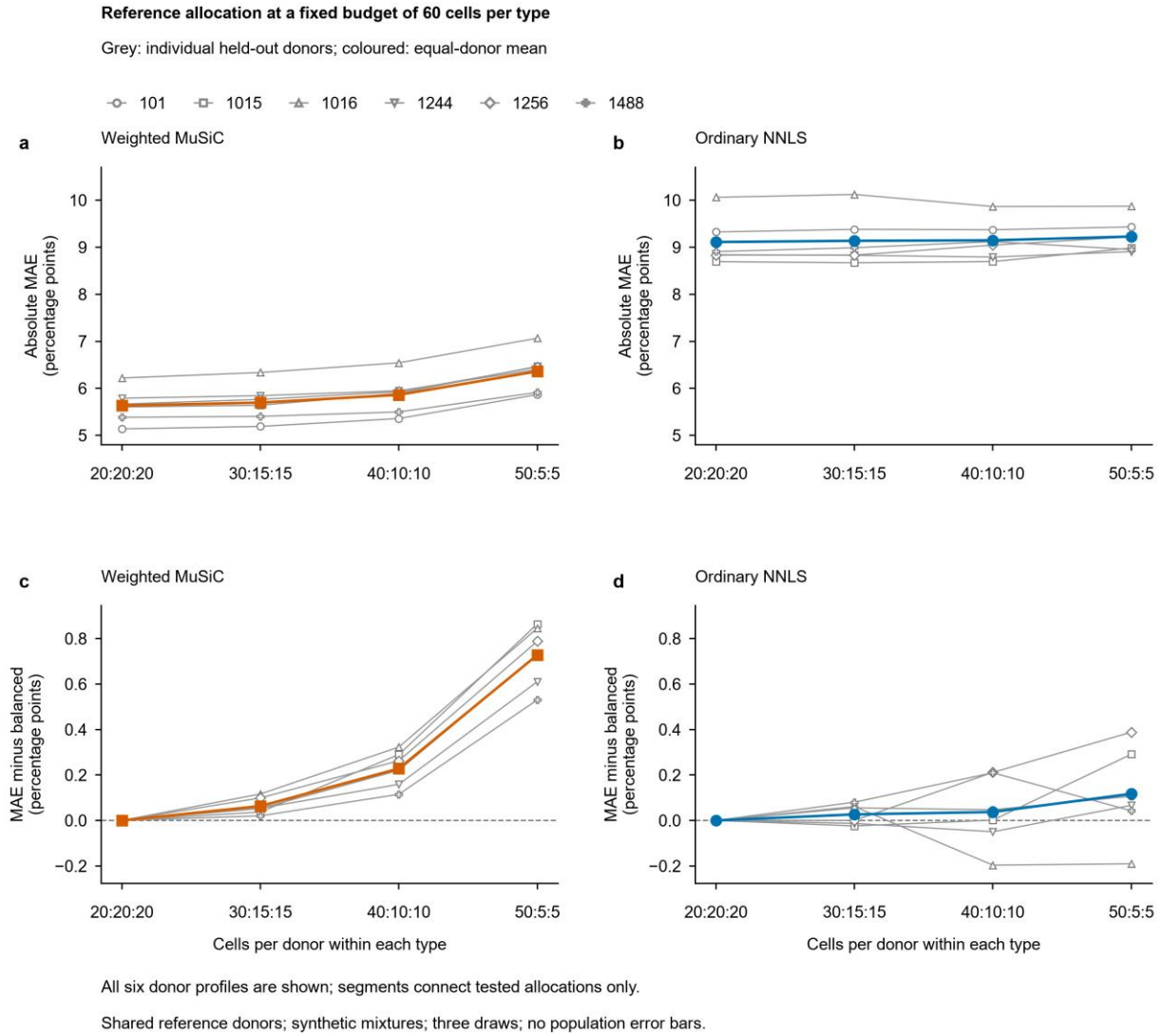
Table S7. Secondary outcomes in the original cohort across all allocations. Values are equal-donor means in percentage points. Conditional SD describes variation across three reference draws, not population uncertainty.

Output	Allocation	RMSE	Conditional SD
Weighted MuSiC	1:1:1	7.106	1.968
Weighted MuSiC	2:1:1	7.178	1.993
Weighted MuSiC	4:1:1	7.365	1.981
Weighted MuSiC	10:1:1	7.941	2.056
Ordinary NNLS	1:1:1	11.566	5.139
Ordinary NNLS	2:1:1	11.581	5.308
Ordinary NNLS	4:1:1	11.567	5.503
Ordinary NNLS	10:1:1	11.625	5.884

Table S8. External secondary outcomes across all budgets and allocations, using the units and definitions in Table S7. Mean conditional SD decreased with budget at each allocation, although absolute accuracy did not consistently improve.

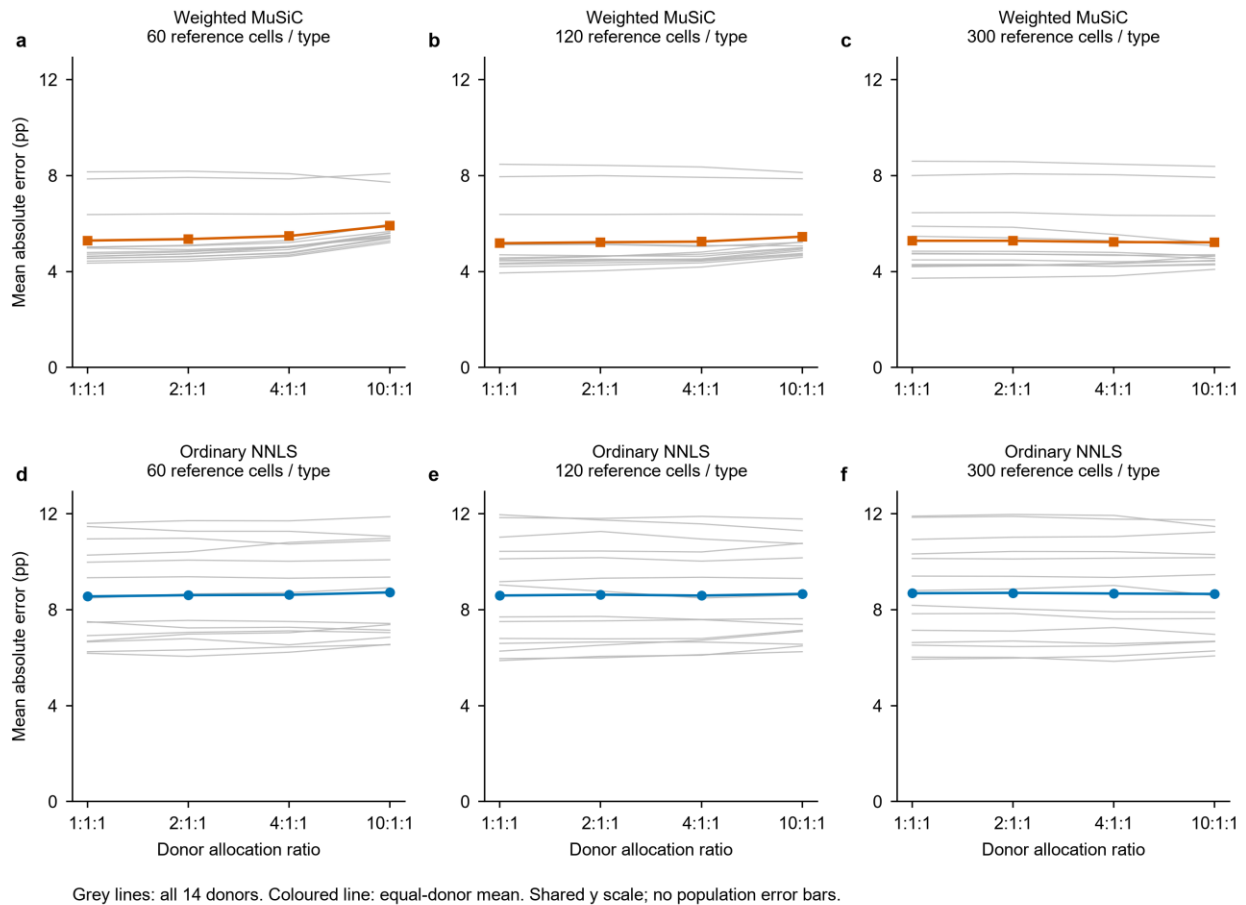
Output	B	Allocation	RMSE	Conditional SD
Weighted MuSiC	60	1:1:1	6.803	1.627
Weighted MuSiC	60	2:1:1	6.865	1.648
Weighted MuSiC	60	4:1:1	7.014	1.758
Weighted MuSiC	60	10:1:1	7.516	1.877
Weighted MuSiC	120	1:1:1	6.643	1.460
Weighted MuSiC	120	2:1:1	6.706	1.496
Weighted MuSiC	120	4:1:1	6.749	1.598
Weighted MuSiC	120	10:1:1	6.988	1.746
Weighted MuSiC	300	1:1:1	6.833	1.190
Weighted MuSiC	300	2:1:1	6.829	1.186
Weighted MuSiC	300	4:1:1	6.736	1.271
Weighted MuSiC	300	10:1:1	6.714	1.504
Ordinary NNLS	60	1:1:1	10.908	3.496
Ordinary NNLS	60	2:1:1	10.961	3.750
Ordinary NNLS	60	4:1:1	10.951	4.269
Ordinary NNLS	60	10:1:1	11.044	4.997
Ordinary NNLS	120	1:1:1	10.968	2.566
Ordinary NNLS	120	2:1:1	11.015	2.828
Ordinary NNLS	120	4:1:1	10.960	3.241
Ordinary NNLS	120	10:1:1	11.001	4.156
Ordinary NNLS	300	1:1:1	11.118	1.902
Ordinary NNLS	300	2:1:1	11.132	1.853
Ordinary NNLS	300	4:1:1	11.096	2.022
Ordinary NNLS	300	10:1:1	11.045	2.902

Figure S1. Original six-donor allocation profiles.



a,b, Mean absolute error (MAE) for weighted MuSiC and ordinary nonnegative least squares (NNLS), at 60 reference cells per type across three donors. c,d, Paired MAE differences from balanced allocation. All values are percentage points. Grey lines show six donor means; coloured lines show their equal-weight mean. Values average ten triples, three draws, 60 targets and three dominant-donor choices for unequal allocations. Scales are shared within each outcome. Lines connect tested allocations; S9 reports reversals within individual arrangements.

Figure S2. External absolute-error profiles.



a–c, Weighted MuSiC at budgets of 60, 120 and 300 reference cells per type across three donors; d–f, ordinary NNLS. All panels share a mean absolute error (MAE) scale in percentage points. Grey lines show 14 donor means; coloured squares or circles show their equal-weight mean. Each value averages 60 targets, three draws, four triples and three dominant-donor choices for unequal allocations. Main Figures 1 and 2 show the corresponding allocation contrasts.

S13. External verification and reproduction scope

Pooled-versus-separate target-batch controls matched exactly. All 604,800 predictions passed identifier, truth, proportion and error checks, and all 302,400 weighted fits reached the default stopping criterion. A separate untraced code path reconstructed 14 prespecified cases and reproduced 1,680 predictions and errors exactly. Complete error recalculation differed by at most 9.85323×10^{-16} .

Table S9. Archived analysis resources. Paths are relative to the extracted Zenodo release; directory names retain the original analysis layout.

Entry	Contents
A1	Earlier normalized-cell exploratory findings: results/manuscript_tables/Table_S3.csv; full description in Section S7.
A2	Original endpoint summaries: results/music_review/; software and source information: reproduction_integrated/environment/.
A3	Four-level gradient: results/music_gradient_review/ and results/music_gradient_figures/
A4	Thinning comparison: results/music_downsampling_review/.
A5	External design: results/music_external_selection/, results/music_external_input/, results/music_external/ and results/music_external_review/
A6	Alternative grouping: results/music_reference_composition/ and results/music_reference_composition_review/
A7	Profile and weight diagnostics: results/music_mechanism/ and results/music_mechanism_review/.
A8	Parameterized reconstruction and fitting: reproduction_integrated/code/; processed inputs and saved predictions: data.zip.

The five-stage integrated package was tested with relocated processed inputs in the existing environment. Its 4,818 control rows matched saved values, and 1,530 summaries recalculated from 1,003,680 predictions agreed within 7.11×10^{-15} percentage points. Fresh installation, a second machine and reconstruction from the complete upstream H5AD were not tested. This package did not include the later diagnostics, whose checks are described in S15.

S14. One alternative reference-donor grouping

This sensitivity analysis was specified after the initial external results and before follow-up predictions. It retained the cohort and existing cell draws and was not publicly preregistered.

The same 14 external donors were ordered by ascending SHA256 of UTF-8 `music-external-reference-composition-v1|ring|{donor}`. The single resulting order was IGTB1966, IGTB256, HC-014, HC-502, IGTB195, IGTB1913, HC-572, IGTB498, IGTB1650, IGTB1506, IGTB508, IGTB514, IGTB826, IGTB1921. For each held-out donor, the next 12 labels formed four consecutive triples and the thirteenth was omitted, using the cyclic rule from S11. Each donor was omitted once and served as reference in 12 folds. Ten of 14 unique triples differed from the original ring; four overlapped, and 11/14 omitted-donor choices changed. No alternative salt was searched and overlap did not trigger reselection. The change includes both triple membership and the omitted donor.

The alternative grouping reused all 840 targets, truth proportions, three donor–type reference draws, 30,172 genes and MuSiC settings. It tested $B=60$ and 300 with balanced and 10:1:1 allocations and all dominant choices: 336 distinct matrices and 161,280 predictions. Intermediate budgets and allocations were not included.

Under the alternative grouping, weighted $I(300)-I(60)$ averaged -0.635 points (range -1.223 to $+0.242$) and was negative for 13/14 donors. HC-014 remained positive; HC-572 was close to zero (-0.017). Relative to the original grouping, the budget contrast changed by $+0.053$ points on average (range -0.180 to $+0.293$), with ten positive and four negative changes (Table S11). These results describe sensitivity to grouping; no equivalence test was performed.

Weighted I(60) was +0.602 points, positive for 13/14 donors. At B=300, the mean was -0.033 despite positive contrasts for 8/14 donors and 81/168 arrangements, compared with 4/14 donors and 84/168 arrangements in the original grouping. Individual penalties thus remained under both near-zero means. Ordinary NNLS I(60) and I(300) were +0.184 and +0.049, with nine positive donors at each budget; its high-budget mean changed sign from the original -0.030. Weighted-minus-ordinary contrasts were +0.418 and -0.082. Weighted donor-average MAE was lower in all 56 donor/budget/allocation conditions.

Weighted RMSE allocation contrasts were +0.688 at B=60 and -0.114 at B=300. Weighted endpoint contrasts within blocks 0, 1 and 2 were -0.505, -0.698 and -0.701 points. Complete type, block and arrangement summaries are retained in the resources listed under A6.

Table S10. Alternative-grouping absolute MAE and signed allocation contrasts. Equal-donor means in percentage points; calculations use unrounded values. B denotes cells/type across three donors.

Output	B	Balanced MAE	10:1:1 MAE	I(B)
Weighted MuSiC	60	5.325	5.927	+0.602
Weighted MuSiC	300	5.225	5.192	-0.033
Ordinary NNLS	60	8.712	8.896	+0.184
Ordinary NNLS	300	8.651	8.701	+0.049

Table S11. Paired weighted budget contrasts across grouping designs. Each value is I(300)-I(60), in percentage points. The last column subtracts the original grouping from the alternative grouping. Shared donors, targets and prefixes prevent interpreting the rows or designs as independent populations.

Held-out donor	Original grouping	Alternative grouping	Alternative minus original
HC-014	+0.218	+0.242	+0.024
HC-502	-1.009	-0.716	+0.293
HC-572	-0.189	-0.017	+0.171
IGTB1506	-1.043	-1.223	-0.180
IGTB1650	-0.729	-0.701	+0.028
IGTB1913	-0.511	-0.356	+0.156
IGTB1921	-1.045	-1.174	-0.129
IGTB195	-0.983	-1.062	-0.079
IGTB1966	-0.297	-0.285	+0.012
IGTB256	-0.858	-0.618	+0.240
IGTB498	-0.752	-0.706	+0.046
IGTB508	-0.780	-0.763	+0.016
IGTB514	-0.490	-0.562	-0.072
IGTB826	-1.152	-0.943	+0.209

Pooled and separate target-batch controls agreed exactly. All 80,640 weighted fits met the default stopping criterion; effective support ranged from 9,793 to 12,929 genes. Predictions passed identifier, quota, truth, proportion and error checks. The four triples shared with the original grouping produced 14,400 identical-condition predictions that matched exactly.

A separate code path rebuilt one prespecified case per donor and called untraced MuSiC; all 1,680 replay predictions and errors matched exactly.

S15. Post-hoc reference-profile and fitted-weight diagnostics

Diagnostic design

The diagnostic protocol was fixed after the external results but before diagnostic fits. It retained all 14 donors, six types, three draws, three budgets and balanced/10:1:1 allocations with every dominant choice. We selected target 0 plus the seven smallest SHA256 values of music-mechanism-v1|mixture|{id} among IDs 1–59: 0, 6, 12, 31, 32, 38, 46 and 48. Selection did not use outcomes. This gave 112 targets and 16,128 weighted fits with corresponding ordinary outputs. The main endpoint continued to use all 60 targets per donor.

Profile precision and statistical connection

The exported 93,892-cell input contains the union of selected reference/target cells, not all 114,524 eligible donor/type cells; only 23/84 strata are complete. No exported-union profile was treated as full-inventory truth. Using the exact three reference draws, profiles θ were reconstructed on all 30,172 genes at quotas $n=5,10,20,25,40,50,100,250$. We recorded $V(n)=\Sigma_g Var_{draw}(\theta_g)$, with sample-variance denominator two; the mean of three pairwise profile TV distances; and the coefficient of variation of mean library size across draws. Repeated profiles across triples were checked for exact equality before deduplicating to 672 donor/type/quota summaries. Source inventory N came from the frozen eligibility table.

For interpretation, fix a gene/type and $m=3$ donors and write an estimated profile as $\theta_d=\theta_d+\varepsilon_d$. Under independent, zero-mean errors with variance v_d , the sample variance across donor estimates satisfies $E[\Sigma | \theta] = s^2_\theta + (\frac{1}{m})\Sigma_d v_d$. This follows by applying the centering matrix $P = I - \frac{11^T}{m}$: the expected noise contribution is $\frac{tr(P diag(v))}{(m-1)}$. Thus sampling variation can enter the quantity used as cross-subject variance; s^2_θ itself includes all fixed donor-profile differences and is not identified as purely biological variation.

As an illustrative approximation only, if $v_d = \frac{c}{n_d}$ with a common c , the added term is $\frac{c(\Sigma_d \frac{1}{n_d})}{3}$. At B=60 it equals 0.050c for 20:20:20 and 0.140c for 50:5:5; at B=300 the values are 0.010c and 0.028c. This calculation illustrates the approximation rather than fitting it to the data. MuSiC profiles are ratios of count sums; ratio bias, heterogeneous donor variances and finite-population sampling can invalidate the approximation, particularly at small n . It does not estimate an MAE effect or identify a sufficient number of cells.

Official variance and weight measurements

For each reference and type, we summed MuSiC's between-donor variance Σ over genes and subtracted the balanced value matched by triple, draw and budget. All-zero reference genes contributed zero. Each comparison was assigned to its four eligible held-out donors and averaged equally over types, draws, dominant choices, triples and donors. This quantity measures between-donor variation; $V(n)$ measures within-donor variation across draws.

For each matched target, we restricted fitted weights to shared genes with positive, finite weights and normalized each vector on that intersection. Total-variation distance (TV) was half the summed absolute difference between the vectors. We also recorded Spearman correlation, gene-support Jaccard similarity and the retained weight fraction, without filling missing weights with zeros. Summaries gave equal weight to

eight targets, draws, dominant choices, triples and held-out donors. As in the official code, weight denominators used the preceding iteration's residuals and coefficients.

Results and contrary observations

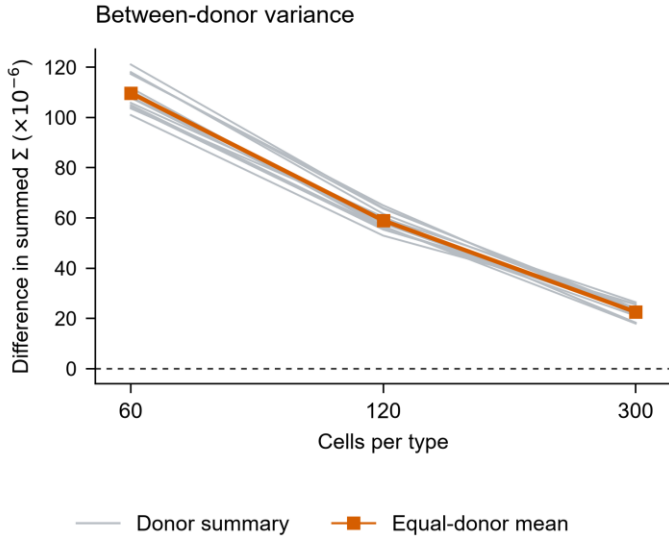
Table S12. Conditional diagnostic summaries. Equal-donor means; V and gene-summed Σ are multiplied by 10^6 and have squared relative-expression units. Minor-donor quotas are 5, 10 and 25. Other measures are dimensionless, except $I(B)$, which is in percentage points. Profile rows average six types per reference donor; Σ and weight rows use the held-out-donor aggregation described above.

Diagnostic	B=60	B=120	B=300
Minor-profile $V \times 10^6$	235.914	129.069	48.179
Gene-summed Σ allocation contrast $\times 10^6$	109.727	58.981	22.558
Shared-support weight TV	0.404	0.346	0.288
Shared-support weight Spearman	0.564	0.623	0.713
Weight-support Jaccard	0.946	0.968	0.984
Balanced weight fraction retained	0.947	0.968	0.986
Unequal weight fraction retained	0.945	0.973	0.989
Eight-target weighted $I(B)$, points	+0.750	+0.247	-0.141

Variance across draws decreased at both quota increases for all 84 donor-type groups contributing the smaller cell samples. The ratio of overall means, $V(25)/V(5)$, was 0.20422; the mean of individual ratios was 0.22340. Balanced and dominant trajectories also had lower endpoint variance in all 84/84 groups, although 5/84 balanced trajectories increased at an intermediate step. Library-size CV increased between endpoints in 10/84 minor, 15/84 balanced and 9/84 dominant trajectories. Profile variance and library-size variability therefore did not improve uniformly together.

All 14 held-out donor means had positive Σ allocation contrasts that decreased at both budget steps. Individual reference-type contrasts, however, were negative in 105/756, 145/756 and 186/756 cases. Weight TV decreased at both steps for all 14 donor means, while gene-support overlap increased. The eight-target weighted budget contrast was -0.890891 points and was negative for 13/14 donors; the corresponding contrast using all 60 targets was -0.687 points. These diagnostic associations do not establish the cause of the error changes.

Figure S3. Allocation contrasts in between-donor variance.



Gene-summed between-donor variance Σ under 10:1:1 minus balanced allocation, in units of 10^{-6} squared relative expression. Grey lines show 14 held-out donor summaries, each averaging types, draws, dominant choices and reference triples; orange squares show their equal-weight mean. The dashed line marks zero. Configurations overlap, and budgets are categorical. Main Figure 3 shows profile dispersion and fitted-weight distances; Table S12 gives all diagnostic summaries.

Verification and limitations

Batch-size and traced/untraced controls passed, and direct basis reconstruction matched MuSiC. All 16,128 weighted fits reached the default stopping criterion; predictions matched the original run within 6.11×10^{-16} . A separate reconstruction checked all 672 profile summaries (maximum differences: variance 1.08×10^{-19} , TV 1.05×10^{-14} , library-size CV 5.55×10^{-17}). Independent checks also covered 36 control bases, raw-weight support and batching, 24 production weight summaries, 12 paired summaries and donor aggregation. Weight checks covered these subsets rather than every production weight vector.

The dispersion estimates use only three draws. At $n=250$, sampling fractions ranged from 4.27% to 97.28%, so finite-population saturation may contribute to the decline. Profiles were not assessed against complete-inventory truth. Because the basis, cell sizes, Σ , residuals and gene support change together, their individual contributions to the MAE pattern remain unresolved.

S16. Magnitude and distribution of arrangement-level contrasts

For each held-out donor, reference triple and dominant donor, we averaged paired MAE differences over 60 targets and three draws. Each budget and output therefore has 168 arrangement-level contrasts: 12 for each of 14 donors. These are distinct from the 168 donor–budget–allocation absolute-error comparisons in S12.

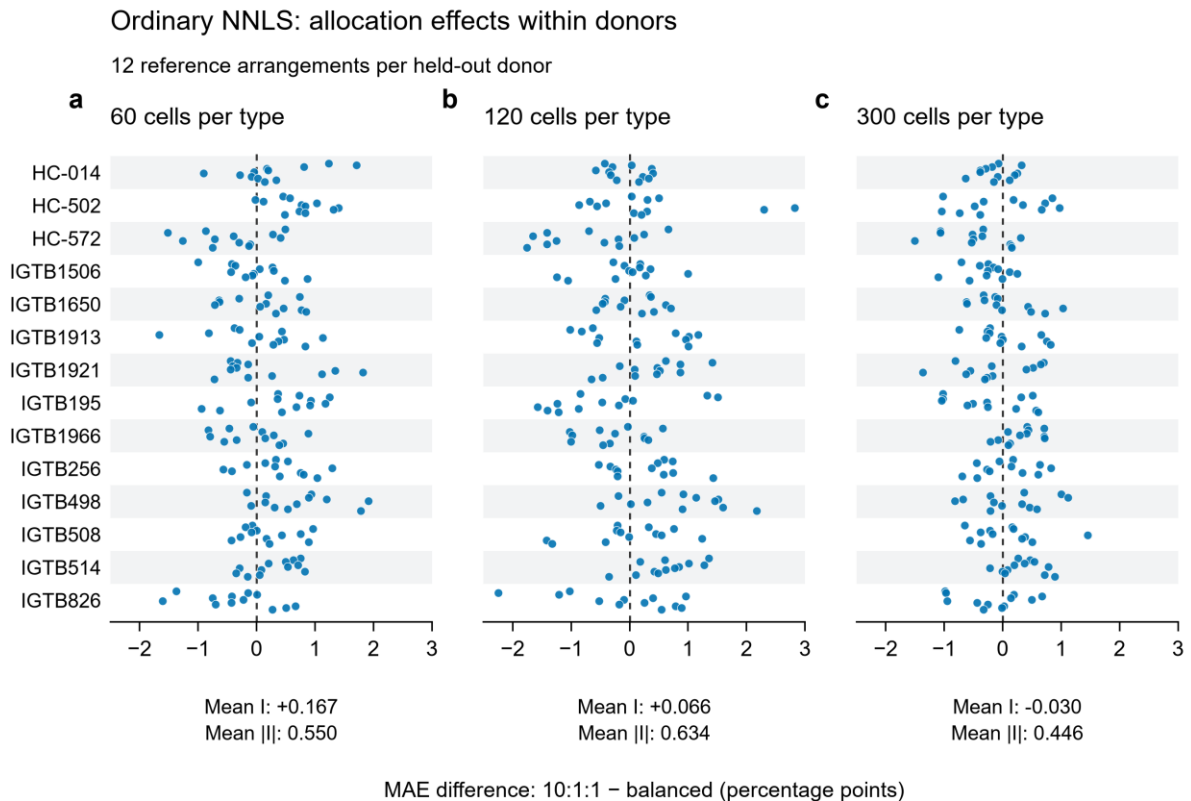
We report signed and absolute means, median absolute contrasts and linearly interpolated empirical quantiles. Absolute values were taken after averaging within arrangements, so they describe changes in accuracy rather than distances between prediction vectors or absolute changes for individual targets. Signs were classified with a tolerance of 10^{-10} percentage points; Table S13 has no ties. Quantiles describe the tested configurations, not population uncertainty.

Table S13. Arrangement-level signed contrasts and magnitudes. All effect values are percentage points. The 10th–90th interval contains empirical signed quantiles, not confidence limits. Both groupings use their original three draws; rows retain all 168 arrangements. Signed means may be close to zero while absolute means remain nonzero.

Grouping	Output	B	Mean signed	Mean absolute	Median absolute	10th–90th percentiles	Positive/168
Original	Weighted MuSiC	60	+0.621	0.708	0.751	−0.043 to +1.168	147/168
Original	Weighted MuSiC	120	+0.276	0.439	0.445	−0.345 to +0.739	125/168
Original	Weighted MuSiC	300	−0.066	0.402	0.321	−0.722 to +0.524	84/168
Original	Ordinary NNLS	60	+0.167	0.550	0.436	−0.708 to +0.944	100/168
Original	Ordinary NNLS	120	+0.066	0.634	0.492	−1.023 to +1.007	91/168
Original	Ordinary NNLS	300	−0.030	0.446	0.371	−0.715 to +0.704	77/168
Alternative	Weighted MuSiC	60	+0.602	0.688	0.739	−0.085 to +1.143	143/168
Alternative	Weighted MuSiC	300	−0.033	0.419	0.378	−0.769 to +0.605	81/168
Alternative	Ordinary NNLS	60	+0.184	0.635	0.544	−0.814 to +1.197	103/168
Alternative	Ordinary NNLS	300	+0.049	0.453	0.363	−0.584 to +0.852	86/168

In the original grouping, weighted mean absolute contrasts decreased from 0.708 to 0.439 to 0.402 points, while signed 10th–90th percentile spans were 1.211, 1.085 and 1.246. The alternative grouping gave mean absolute contrasts of 0.688 and 0.419 at B=60 and 300, with spans of 1.227 and 1.373. In both groupings, the lower average magnitude did not consistently narrow the central signed range, and individual allocation penalties remained.

Figure S4. Ordinary NNLS allocation contrasts in the original external grouping.



Panels show budgets B of 60, 120 and 300 reference cells per type across three donors. Each point is $I(B) = \text{MAE}(10:1:1) - \text{MAE}(\text{balanced})$, the mean absolute error difference in percentage points, averaged over 60 targets and three draws for a fixed held-out donor, triple and dominant donor. Each donor contributes 12 arrangements; vertical offsets separate points without scientific meaning. Panels share a horizontal scale, and dashed lines mark zero. Mean I and mean $|I|$ summarize signed contrasts and absolute magnitudes. Shared sources and targets make these points dependent. Main Figure 2 shows weighted MuSiC.

S17. Additional reference sampling and conditional precision

Design and randomization unit

After inspecting the initial results and arrangement distributions, we specified 12 new global reference-draw blocks (IDs 3–14) before selecting cells or generating predictions. The follow-up retained all 14 donors, the original grouping, 840 targets, gene list and MuSiC settings. It tested $B=60$ and 300 with balanced and 10:1:1 allocations and every dominant-donor choice. This exploratory follow-up was not publicly preregistered.

Each donor/type/block permutation uses the complete eligible inventory of 114,524 cells, not the old selected-cell union. SHA256 of `music-external-v1|reference|{donor}|{author_type}|{block}`, interpreted from its first 16 bytes as an unsigned big-endian integer, initializes NumPy PCG64. Prefixes of up to 250 cells are reused across triples and folds, with nested budgets within allocation. Randomization is independent between complete global blocks; reuse of cells across blocks is permitted. Within-block donor, target and reference dependence is preserved.

Within each block, MAEs were averaged over targets and paired with the matched balanced baseline. Dominant-donor choices, triples and held-out donors received equal weight. For block estimates of I(60), I(300) and C=I(300)–I(60), MCSE was the sample SD divided by the square root of the number of blocks [S8]. It describes reference-randomization precision conditional on the fixed inventory, donors, grouping and targets, not population uncertainty.

All 12 blocks were required, with an MCSE target of 0.05 points for each of the three weighted estimates. Summaries at 3, 6, 9 and 12 blocks served as diagnostics, not stopping rules. Original, new and combined blocks are reported separately; combining them was an additional follow-up. Ordinary outputs and all donor, type and arrangement summaries were retained.

Results and verification

All 12 blocks completed, yielding 1,344 reference matrices and 645,120 prediction rows across both outputs. All 322,560 weighted fits met the default stopping criterion. Pooled-versus-separate target controls agreed exactly. Predictions passed identifier, truth, proportion and diagnostic checks; an independent summary implementation agreed within 2.22×10^{-16} points across 2,316 comparisons.

The expanded export contains 106,359 cells: the original 93,892 plus 12,467 newly selected cells. Shared counts matched exactly, and targets, truth and genes were unchanged. All 1,260 old and new donor–type–block prefixes were reproduced from the complete eligible inventory of 114,524 cells. Randomization requires this full inventory rather than the exported subset.

Table S14. Conditional reference-sampling precision. Means and MCSEs are in percentage points. Each global block retains all 14 donors, 840 targets, reference triples and dominant choices. MCSE is block SD divided by the square root of block count. Only the three New 12 weighted MCSEs had the prespecified 0.05-point target; original and combined MCSEs are additional descriptive summaries.

Draw set	Output	I(60)	MCSE	I(300)	MCSE	I(300)-I(60)	MCSE
Original 3	Weighted MuSiC	+0.621	0.0146	–0.066	0.0291	–0.687	0.0203
Original 3	Ordinary NNLS	+0.167	0.0930	–0.030	0.0954	–0.198	0.1860
New 12	Weighted MuSiC	+0.591	0.0319	–0.073	0.0296	–0.665	0.0486
New 12	Ordinary NNLS	+0.061	0.0812	–0.089	0.0429	–0.149	0.1025
Combined 15	Weighted MuSiC	+0.597	0.0256	–0.072	0.0240	–0.669	0.0388
Combined 15	Ordinary NNLS	+0.082	0.0672	–0.077	0.0382	–0.159	0.0873

All three MCSEs from the new blocks met the precision target. The budget-contrast MCSE was 0.0486 points, larger than the original three-block estimate; dispersion estimates based on few blocks can be unstable.

The new-block weighted budget contrast was negative in every block (12/12; range –0.974 to –0.394 points). After averaging blocks, it was negative for 13/14 donors in both the new and combined sets. HC-014 remained the exception, with budget contrasts of +0.087 and +0.113 points, respectively. Weighted donor-

average absolute MAE remained below ordinary NNLS in all 56 endpoint donor/budget/allocation conditions in each set.

At B=300, positive I(B) remained for 6/14 donors and 84/168 arrangements in the new blocks, and for 85/168 arrangements in the combined set. New-block mean absolute contrasts were 0.650 and 0.395 points at B=60 and 300, with signed 10th–90th percentile spans of 1.008 and 1.203. Combined values were 0.659 and 0.382 for absolute means, and 0.990 and 1.171 for spans. As in S16, a smaller mean absolute contrast accompanied a wider central signed range.

Complete block, donor and arrangement summaries are retained in results/music_mc_extension_review/, with full-inventory mapping in results/music_mc_reproduction_inventory/. These additional draws apply only to the endpoint allocations and budgets; S15 diagnostics remain based on three draws.

S18. Experimental controls in selected related work

Table S15. Selected reference-design experiments. Unspecified means that a control was not specified for the inspected experiment. B denotes cells per type.

Study and experiment	Reference donor identities	Cells per type	Within-type donor quotas	Multiple cell budgets	Evidence
Zhao: condition matching [S9]	Fixed within a reference; differ across conditions/runs	Fixed B unspecified	Condition composition varied; quota intervention unspecified	20 versus 40 donors, not controlled B	Simulation study; Fig. 1
Huuki-Myers: type balancing [S5]	Same source cohort; donor retention per draw unspecified	1599 (Bisque) or 1601 (hspe) after downsampling	Types equalized; donor quotas not controlled	Full versus one reduced B per method	Equal proportion reference simulation; Figs. S29–S30
Huuki-Myers: donor subsampling [S5]	Number and selected identities change	Fixed B unspecified	Donor inclusion varied	Ten donor counts, 3–52	Donor subsampling simulation; Fig. S33
Dietrich: context comparison [S10]	Reference datasets/sources differ	500 in training references	Quota intervention unspecified	One B in this experiment	Technology/tissue-context methods; Fig. 6B–C
Dietrich: efficiency [S10]	Same Hao source; donor retention unspecified	Not the stated control	Quota intervention unspecified	1000–100000 whole-reference cells	Computational-efficiency methods
Malik: preview and fixed code [S11]	Biological donor identities not established in the inspected code	Code caps each type at 200	Not established from available evidence	Allocation-by-B grid not established	Publisher preview; fixed MuSiC runner
Huang: scValue deconvolution [S13]	Two fixed reference donors	Selected by subsampling method	Within-type quotas not controlled	Subsampling evaluated; no quota-by-B grid	T&ILC synthetic mixtures; deconvolution experiment
Present external experiment	Same three donors per paired comparison	Fixed B within each comparison	Four ratios; all dominant choices	B=60/120/300; follow-ups at 60/300	S11, S14 and S17

Huuki-Myers’ cell-type balancing and donor subsampling are separate experiments [S5]. Dietrich’s reference-context comparison and efficiency analysis also address different outcomes [S10]. scValue evaluates cell selection for deconvolution using a two-donor reference [S13]. The present grouping and

sampling follow-ups retain the controlled quota comparison but cover only the endpoint budgets and allocations.

For Malik et al., we inspected the publisher preview and author code at a fixed commit [S11,S12], not the full journal PDF. The inspected MuSiC runner assigns per-cell dummy subject IDs and caps cells at 200 per type; its relationship to every published result remains unresolved. Full methods and selected supplementary material were available for the other experiments in Table S15.

S19. Measured PBMC RNA-seq and flow-cytometry proportions

We used the Monaco S13 cohort: 13 healthy participants with PBMC RNA-seq, of whom 12 had released flow-cytometry proportions [S14]. DZQV was absent from the flow table; the source reports one failed flow quality check. Sample titles linked the remaining 12 participants to GSE107011/SRP125125 runs. The source article’s Table S6, worksheet ‘GSE107019 (Our S13 cohort)’, supplied measured proportions; neither its expression-based estimates nor its calibrated signatures were used.

recount3 gene coverage counts used GRCh38 and GENCODE v26 [S15]. We divided each run by STAR’s average mapped length and rounded, following `compute_read_counts`; GEO TPM values were not used in this initial assessment. Removing Ensembl version suffixes while retaining PAR suffixes matched 30,165 of 30,172 reference genes. MuSiC selected shared genes internally; complete single-cell counts and library sizes were retained.

We combined the six predicted types into four lineages: B cells, CD4 plus CD8 T cells, NK cells, and CD14+ plus FCGR3A+ monocytes. The flow counterparts were ‘B Naive+Memory’, ‘T cells’, ‘NK’ and ‘Monocytes’. B proportions excluded separately annotated plasmablasts. Combining T cells accommodated MAIT cells within the Perez CD8 compartment [S6], although gamma-delta annotation remained unresolved. Combining monocytes avoided assigning the flow intermediate class to one reference subtype. These four groups represented 82.975–93.536% of reported PBMCs and were normalized to sum to one for scoring. Other populations still contributed to bulk expression.

Before fitting, we specified all 1,344 saved references from blocks 3–14: 14 unique triples, two budgets and four allocation arrangements per triple and budget. Each reference predicted every bulk participant once, without duplicating synthetic held-out assignments. Parameters followed S4. Four-lineage MAE and paired contrasts gave equal weight to participants, blocks, triples and dominant-donor choices. MCSE was calculated from the 12 whole-block contrasts. We did not tune against flow measurements, draw new reference cells or exclude samples by outcome.

All 16,128 weighted fits reached the default stopping criterion with 11,119–15,245 effective genes and finite variance estimates. Yet weighted B estimates were at most 10^{-10} in every fit, and T estimates in 16,127. Ordinary NNLS estimated zero B and T proportions throughout. After observing this failure, we reconstructed NNLS using exact gene IDs for participant 453W and the first balanced reference. Agreement with the saved ordinary estimate was within 1.12×10^{-15} . This targeted check found no row-order error but did not explain failed transfer.

Table S16. Measured-bulk errors and allocation contrasts in percentage points. MAE is calculated across four represented lineages. Mean |I| is taken across 42 triple/dominant arrangements after averaging participants and blocks. MCSE describes conditional reference-draw precision.

Output	B	Balanced MAE	10:1:1 MAE	I(B)	MCSE I(B)	Mean I 42 arrangements
Weighted	60	31.676	31.675	-0.0013	0.0026	0.0058
Weighted	300	31.672	31.672	+0.0007	0.0006	0.0007
Ordinary NNLS	60	37.557	37.477	-0.0793	0.0486	0.0983
Ordinary NNLS	300	37.574	37.556	-0.0181	0.0129	0.0227

The weighted budget contrast was +0.0019 points (MCSE 0.0027), compared with +0.0612 (MCSE 0.0521) for ordinary NNLS. The near-zero weighted contrast accompanied severe error and does not validate the simulation budget trend in measured samples. Assay differences, cell-size estimation and omitted populations remain possible explanations that this comparison cannot distinguish.

Table S17. Participant-level measured-bulk contrasts in percentage points. Coverage is the sum of the four reported flow fractions before normalization. W and N denote weighted MuSiC and ordinary NNLS. $\Delta = I(300) - I(60)$. Each contrast averages all 12 blocks, 14 triples and three dominant choices. Values below displayed precision are rounded to zero.

Participant	Coverage (%)	W I(60)	W I(300)	W Δ	N I(60)	N I(300)	N Δ
36TS	87.5	-0.0013	+0.0000	+0.0013	+0.0000	+0.0000	+0.0000
453W	90.9	-0.0022	+0.0004	+0.0025	-0.0158	+0.0000	+0.0158
4DUY	83.0	+0.0025	+0.0000	-0.0025	-0.0005	+0.0000	+0.0005
684C	89.6	-0.0010	+0.0069	+0.0079	-0.0517	-0.0007	+0.0511
925L	87.5	-0.0039	+0.0000	+0.0039	-0.0309	+0.0000	+0.0309
9JD4	93.5	+0.0000	+0.0000	+0.0000	-0.1428	-0.0599	+0.0829
CR3L	87.9	+0.0000	+0.0000	+0.0000	-0.3217	-0.0370	+0.2847
CYFZ	87.9	-0.0001	+0.0000	+0.0001	-0.1861	-0.0531	+0.1330
CZJE	92.9	+0.0000	+0.0000	+0.0000	-0.1615	-0.0645	+0.0969
FLWA	87.9	+0.0000	+0.0000	+0.0000	-0.0184	-0.0004	+0.0179
FY2H	92.4	-0.0092	+0.0008	+0.0100	+0.0000	+0.0000	+0.0000
G4YW	93.4	+0.0000	+0.0000	+0.0000	-0.0220	-0.0011	+0.0209

Processed inputs and sample mappings are in data/processed/real_bulk/; predictions and summaries are in results/real_bulk/. Preparation, fitting, plotting and diagnostic scripts are in analysis/, alongside REAL_BULK_PROTOCOL.md.

Measured PBMC RNA-seq and independent flow cytometry

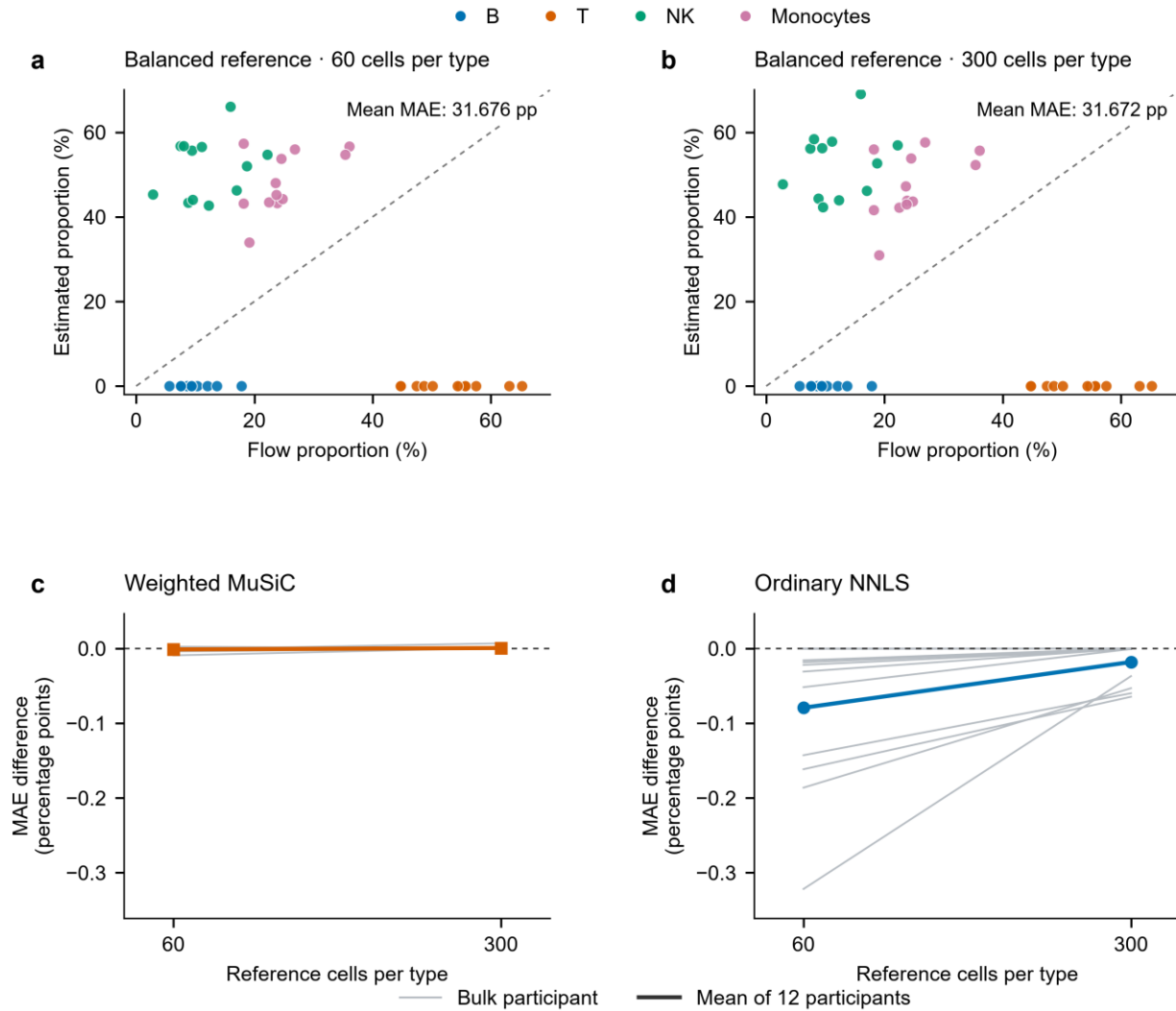


Figure S5. Measured PBMC RNA-seq compared with flow cytometry. B denotes reference cells per type across three donors. a,b, Weighted estimates from balanced references at B=60 and 300 versus flow proportions normalized within B-cell, T-cell, NK-cell and monocyte lineages. Each point is one participant and lineage, averaging predictions over 12 blocks and 14 triples; dashed lines indicate equality. Mean absolute error (MAE) annotations average errors across fits before predictions are averaged. c,d, $I(B) = \text{MAE}(10:1:1) - \text{MAE}(\text{balanced})$ for weighted MuSiC and internal ordinary NNLS. Grey lines show participants; coloured lines show their equal-weight mean. Each contrast averages 12 blocks, 14 triples and all three dominant-donor choices. Near-zero contrasts accompany severe B- and T-cell underestimation. pp, percentage points.

Post-hoc input-scale diagnostic

After failed transfer, we specified two input-scale comparisons using the 14 balanced references from block 3 at B=300 and all 12 participants. One divided recount3 counts by GENCODE v26 exonic length in kb from the annotation GTF score field [S15]; the other used author-supplied PBMC TPM [S14]. Reference cells, raw single-cell UMI counts, type mapping and fitting parameters were unchanged, and saved direct-count predictions supplied the baseline. No flow calibration or marker tuning was used. TPM also changes quantification and gene coverage, so this comparison does not isolate length scaling.

Table S18. Post-hoc comparison of bulk input scales. Four-lineage MAE in percentage points, averaged over the same 14 references within each of 12 participants and then over participants. This subset uses one draw block and balanced allocation at B=300.

Bulk input	Weighted MuSiC	Ordinary NNLS
Read-equivalent counts	31.672	37.595
Counts / exonic length	31.763	37.644
Author-supplied TPM	31.672	37.421

All 168 weighted fits converged for each new input, but both outputs retained B and T estimates at most 10^{-10} throughout. Neither input reduced participant-average weighted MAE. Length scaling alone did not resolve the failure.

Post-hoc residual and gene-exclusion diagnostic

For one balanced reference (d00_d01_d02_b3_n300_balanced) and all 12 participants, we inspected gene-level residuals from saved fits. Genes with RPL or RPS prefixes accounted for 25.7–50.5% of expression in the six reference profiles, versus 0.8–1.7% in read-equivalent bulk counts on the same genes. We excluded all 95 genes whose GENCODE v26 symbols began with RPL or RPS, including related genes and pseudogenes. Reference cells, full single-cell library sizes, direct bulk input and fitting parameters were retained. No further genes were selected for exclusion.

Weighted MAE fell from 31.672 to 26.741 points, improving in all 12 participants. B estimates became positive, but T estimates remained zero. Ordinary NNLS was unchanged at 37.644 points, with B and T estimates zero throughout.

Single-sample quantification diagnostic

We re-quantified the already selected sample 453W (SRR6298350), retaining the same balanced reference. Paired 50-base reads were processed with fastp 0.23.4 [S16], using paired-end adapter detection, a minimum length of 25 and otherwise default filters. Transcripts were reconstructed with gffread 0.12.7 from Ensembl GRCh37 release 87, retaining protein_coding, lincRNA and antisense genes as in the 10x hg19 reference 3.0.0 build recipe. Salmon 1.10.3 [S17] used full-genome decoys, k=23, duplicate transcripts retained, sparse indexing, automatic library detection, validated mappings, sequence and GC-bias correction, and six threads. Transcript NumReads were summed by stable gene ID and rounded to integers, with ties to even.

One fit reused the saved basis; the other selected shared annotated genes before rebuilding the basis from the same cells. Both retained the original parameters and official music_basis/music.iter implementation, with saved predictions as the baseline. Before quantification, expansion required at least 70% for both read retention and mapping, converged weighted fits with finite variances, B and T estimates above 0.0001, and lower weighted MAE than baseline. This diagnostic did not reproduce Zhao et al.’s different reference and processing workflow [S9].

Read retention was 99.44% and mapping 75.16%; both weighted fits converged after four iterations. Shared-gene selection retained 99.9757% of reference UMIs. B and T estimates remained zero and weighted MAE remained 34.358 points. NK estimates fell from 56.52% to 46.87%, with a corresponding monocyte increase, so unchanged MAE did not mean identical predictions. Ordinary NNLS remained at 38.087 points (Figure S6). The recovery and accuracy criteria were not met, and processing was not expanded.

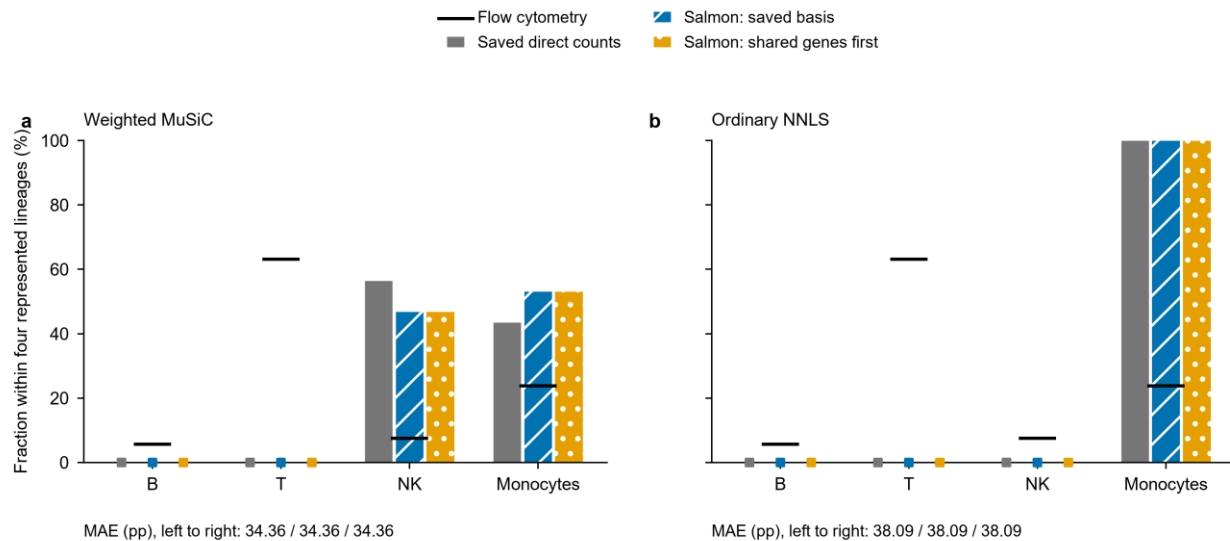


Figure S6. Quantification diagnostic for sample 453W with one balanced three-donor reference containing 300 cells per type. Bars show predictions from saved direct counts, Salmon counts with the saved basis, and Salmon counts after shared-gene selection. Black marks show flow proportions normalized to the four represented lineages (90.911% of reported PBMCs); small squares mark zero estimates. Mean absolute error (MAE) is in percentage points. Ordinary NNLS is MuSiC's internal unweighted comparator.

Designs and parameters are in analysis/REAL_BULK_SCALE_PROTOCOL.md; inputs and results are in the real_bulk_scale and real_bulk_salmon directories under data/processed/ and results/. Preparation, fitting and summary scripts are in analysis/.

Supplementary references

- S1. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol.* 2018;36(1):89-94. doi:10.1038/nbt.4042. Deposited data: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96583>.
- S2. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Author Correction: Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol.* 2020;38(11):1356. doi:10.1038/s41587-020-0715-9.
- S3. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun.* 2019;10(1):380. doi:10.1038/s41467-018-08023-x. Fixed code: <https://github.com/xuranw/MuSiC/tree/f21fe67f5670d5e9fca0ad7550abae3423eb59c>.
- S4. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst.* 2016;3(4):346-360.e4. doi:10.1016/j.cels.2016.08.011. Deposited data: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84133>.
- S5. Huuki-Myers LA, Montgomery KD, Kwon SH, Cinquemani S, Eagles NJ, Gonzalez-Padilla D, et al. Benchmark of cellular deconvolution methods using a multi-assay dataset from postmortem human prefrontal cortex. *Genome Biol.* 2025;26(1):88. doi:10.1186/s13059-025-03552-3. Methods "Equal proportion reference simulation", Supplementary Figures S29–S30.
- S6. Perez RK, Gordon MG, Subramaniam M, Kim MC, Hartoularos GC, Targ S, et al. Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus. *Science.* 2022;376(6589):eabf1970. doi:10.1126/science.abf1970.
- S7. Erratum for the Research Article "Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus" by R. K. Perez et al. *Science.* 2024;385(6705):eadr4064. doi:10.1126/science.adr4064.

- S8. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074–2102. doi:10.1002/sim.8086.
- S9. Zhao Y, Vestal BE, Moore CM. Evaluating reference-mixture matching in cell-type deconvolution with single-cell RNA-seq references. *Brief Bioinform*. 2026;27(2):bbag184. doi:10.1093/bib/bbag184.
- S10. Dietrich A, Merotto L, Pelz K, Eder B, Zackl C, Reinisch K, et al. omnideconv: a unifying framework for using and benchmarking single-cell-informed deconvolution of bulk RNA-seq data. *Genome Biol*. 2026;27:6. doi:10.1186/s13059-026-03955-w.
- S11. Malik AA, Alif MN, Bratton A, Sun J, Li Q, Zhang W. Benchmarking reference-based cellular deconvolution algorithms to predict cell proportions. *Comput Biol Chem*. 2026:109370. Published online 15 September 2026; journal pre-proof. doi:10.1016/j.compbiolchem.2026.109370.
- S12. CompBioLab UCF. Benchmarking-Deconvolution-Algorithms. GitHub; file MuSiC/run_MuSiC.R, commit 35f641156efdd2ec3dd68d5b4453a3616c3d2147. Source code. Accessed 20 September 2026.
- S13. Huang L, Gong W, Chen D. scValue: value-based subsampling of large-scale single-cell transcriptomic data for machine and deep learning tasks. *Brief Bioinform*. 2025;26(3):bbaf279. doi:10.1093/bib/bbaf279.
- S14. Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep*. 2019;26(6):1627–1640.e7. doi:10.1016/j.celrep.2019.01.041. Flow proportions: supplementary Table S6. RNA-seq: GSE107011/SRP125125.
- S15. Wilks C, Zheng SC, Chen FY, Charles R, Solomon B, Ling JP, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol*. 2021;22:323. doi:10.1186/s13059-021-02533-6. Count conversion: https://research.libd.org/recount3/reference/compute_read_counts.html. Accessed 22 September 2026.
- S16. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–i890. doi:10.1093/bioinformatics/bty560.
- S17. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417–419. doi:10.1038/nmeth.4197.