Monocyte Lineage Expansion Drives Transcriptomic Individuality in Genetically Identical Armadillo Quadruplets

Short title: Transcriptomic Individuality in Armadillo Quadruplets.

Authors

Risa Karakida Kawaguchi,[1,2,3]† Sara Ballouz,[1,4]† Maria T. Pena[5], Leon French[6], Frank M. Knight[7], Linda B. Adams[8], Jesse Gillis[6*]

Affiliations

1 Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA
2 Center for iPS Cell Research and Application, Kyoto University, Kyoto, Japan
3 Graduate School of Pharmaceutical Science, The University of Tokyo, Tokyo, Japan
4 Present address: School of Computer Science and Engineering, Faculty of Engineering, University of New South Wales Sydney, Sydney, NSW, Australia
5 United States Department of Health and Human Services, Health Resources and Services Administration, Health Systems Bureau, National Hansen's Disease Program, Baton Rouge, LA, USA.
6 Department of Physiology and Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON, Canada
7 University of the Ozarks, Clarksville, AR, 72830, USA
8 Former United States Department of Health and Human Services, Health Resources and Services Administration, Health Systems Bureau, National Hansen's Disease Program, Baton Rouge, LA, USA.

† These authors contributed equally to this work.
* Correspondence. jesse.gillis@utoronto.ca

Abstract

Genetic diversity shapes phenotypes, yet even genetically identical individuals differ. In nine-banded armadillo (*Dasypus novemcinctus*) quadruplets, we previously showed that allele-specific expression (ASE) imbalances provide a stable molecular fingerprint of individuality. Here, we test whether such transcriptomic individuality reflects functional biological differences. We profiled bulk blood RNA from five cohorts of genetically identical quadruplets across three time points, and found persistent gene-expression signatures that predict individual identity. Focusing on a highly variable cohort, we then performed single-cell RNA-seq and ATAC-seq. In this litter, the most stably distinct individual showed an expanded monocyte-lineage compartment and gene-expression programs enriched for inflammatory and differentiation pathways. These cell-type and regulatory differences were stable over time and robust to experimental leprosy infection. Together, our results link transcriptomic individuality to lasting differences in immune-cell composition, illustrating how early stochastic events can produce persistent, biologically meaningful divergence among genetically identical individuals.

Teaser

Genetically identical armadillo quadruplets are identifiable by transcriptomic signatures shaped by cellular composition variation.

Introduction

Genetically identical individuals raised in similar environments can nevertheless diverge markedly in their phenotypes (*1–4*). Understanding the origins of this non-genetic individuality is challenging because it reflects influences that fall outside both genotype and measurable environmental exposures. One plausible contributor is stochasticity in developmental processes, meaning random molecular or cellular events whose consequences can propagate across scales (*5*). A useful analogy comes from Waddington's depiction of cell-fate decisions as trajectories shaped by an underlying "epigenetic landscape" (*6*). Although originally formulated for cellular differentiation, this idea illustrates how small early perturbations could, in principle, bias organism-level outcomes. Experimental systems support this view, showing that noise in gene expression or protein abundance can generate persistent divergence among genetically identical cells or organisms (*7*, *8*). Random monoallelic expression offers another potential source of such divergence (*9*). In previous work on *Dasypus novemcinctus* (the nine-banded armadillo), we found that stable allele-specific expression (ASE) imbalance patterns distinguish genetically identical siblings, indicating that early stochastic events can leave persistent epigenetic signatures (*10*).

It is not known whether the transcriptomic individuality we previously described corresponds to stable differences in biological function. In general, much molecular variation is expected to be neutral, and many allele-specific expression (ASE) imbalances likely fall into this category, meaning they can persist as individual markers without altering total gene output or phenotype. Total expression of a gene (from both alleles), in contrast, has more direct links to cellular physiology, although it is more dynamic and sensitive to environmental influences. Several routes could, in principle, connect non-genetic individuality to functional differences. One possibility is that early stochastic events bias the relative abundance of specific cell types, producing persistent but subtle differences in cellular composition. Another possibility is that individuals differ primarily in the activation state or regulatory programs of the same cell types, leading to stable differences in transcriptional tone without changes in lineage proportions. Clonal haematopoiesis in aging humans illustrates how small lineage or state biases can have long-term consequences for immune function (*11*). Similar non-heritable processes could contribute to stable individuality among genetically identical siblings.

In this study, we ask whether transcriptomic individuality extends beyond neutral molecular signatures to include stable functional differences. We analyzed bulk blood transcriptomes from five sets of genetically identical armadillo quadruplets raised under controlled conditions, and found that each sibling maintained a consistent gene expression profile across three sampling points. These profiles contain thousands of genes whose ranked expression distinguishes individuals despite identical genotype and shared environment. To test whether these differences reflect underlying physiology, we generated single-cell RNA-seq and single-cell ATAC-seq data from one litter that showed particularly high phenotypic and transcriptomic variability. In this cohort, the most distinctive sibling displayed a clear shift in monocyte-lineage abundance together

with an associated innate immune expression program, whereas its siblings showed more modest lineage- or state-specific differences. These findings link transcriptomic individuality to coherent differences in cellular composition and regulatory activity. They also illustrate how early stochastic influences can give rise to lasting, biologically meaningful individuality among genetically identical animals.

Results

Bulk Transcriptome Analysis of Genetically Identical Quadruplets Reveals Structured Phenotypic Variability

Under controlled conditions, genetically identical individuals might generally be expected to show limited phenotypic divergence. In the five armadillo quadruplet litters we examined (20 animals in total, each litter derived from a single embryo), we nevertheless observed consistent differences in basic physiological traits (Fig. 1A). We quantified within-litter variability using the coefficient of variation for measures such as body weight and white blood cell counts obtained from clinical chemistry and complete blood count (CBC) assays. These traits differed in a structured way across litters. Some groups showed little dispersion, whereas others were noticeably more variable. The 16-90 litter, for example, had one of the highest coefficients of variation in body weight and the greatest variability in white blood cell counts (Fig. 1B-E). These observations indicate that factors other than genotype and shared environment contribute to stable phenotypic differences among siblings.

We next asked whether transcriptomic profiles showed evidence of the same structured individuality. Bulk RNA sequencing of peripheral blood mononuclear cells was performed for all five quadruplet sets at three time points over roughly one year (*10*) (Fig. 1F). Gene expression correlations confirmed that each sibling maintained a highly stable transcriptome across time and that these within-individual profiles were more similar to one another than to profiles from unrelated animals (Fig. 1G). Intra-individual correlations were uniformly high (median Spearman $\rho = 0.979$) and close in magnitude to correlations among siblings, although the small difference between these two groups was statistically significant ($\rho = 0.982$ on average for sibling pairs, $p = 0.008$; Fig. 1H). To place these values in context, we compared log fold changes of homologous genes in armadillos to data from human monozygotic and dizygotic twins. Armadillo siblings showed substantially lower transcriptomic dispersion than human monozygotic twins (Fig. S1), a result consistent with reduced environmental variation in the controlled armadillo colony relative to human cohorts.

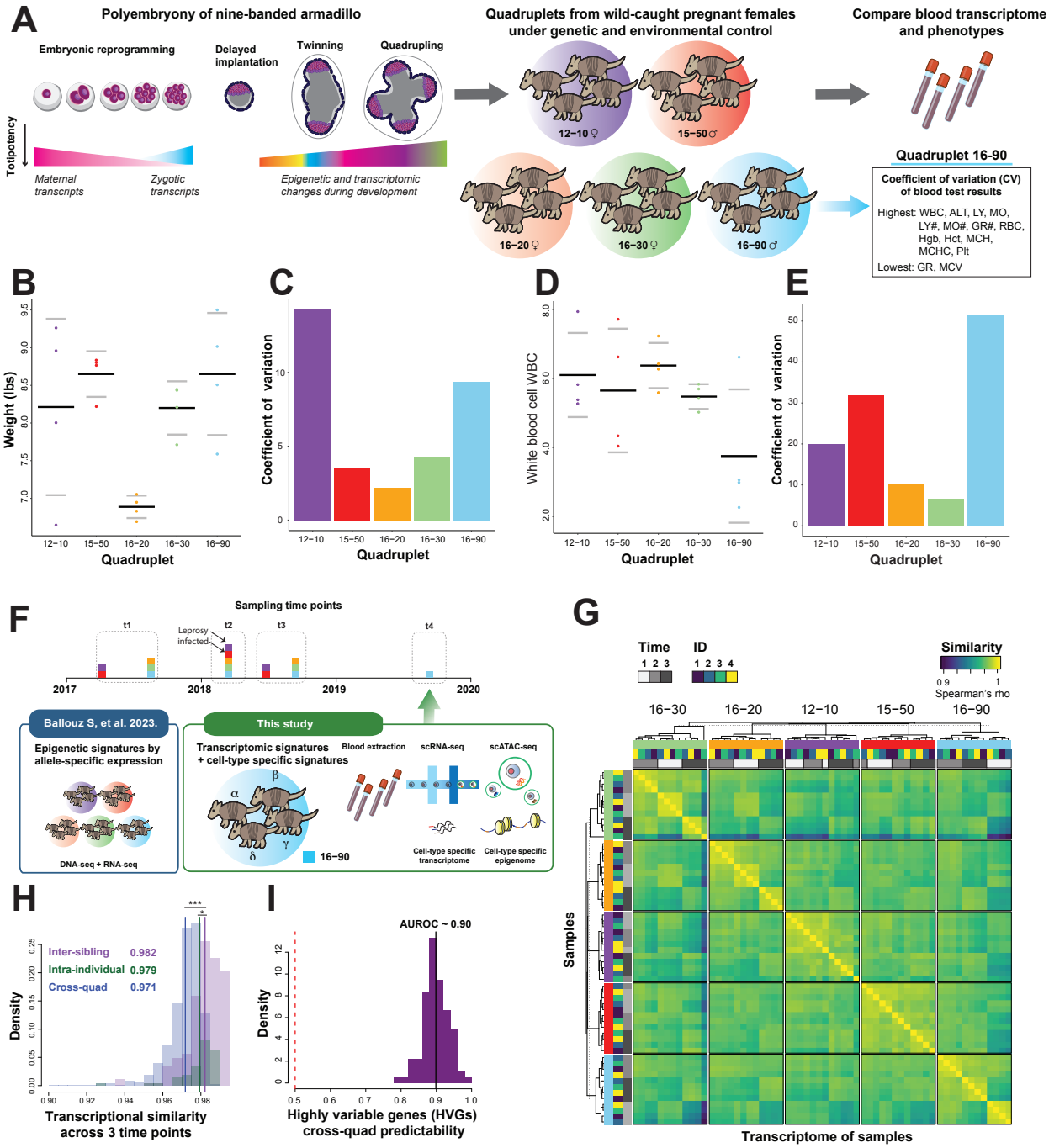


Fig. 1. Dissecting non-genetic variation in armadillo sibling cohort data. (**A**) Polyembryony in the nine-banded armadillo enables efficient sampling of genetically identical siblings under environmental conditions. Quadruplets from five different lineages were raised in the same colony at the laboratory, and blood samples were collected for transcriptomic analysis and phenotypic assessment. Quadruplet group 16-90 shows the highest phenotypic variation across 13 blood-based traits. (**B-E**) Phenotypic variations of each quadruplet group: 1) the weight distribution and its coefficient of variation (**B**) and (**C**), and 2) white blood cell (WBC) count ($\times 10^3$ cells/mm^3) for individuals from five quadruplets (**D**) and (**E**). (**F**) Timeline depicting previously collected blood samples and newly obtained samples used for scRNA-seq and scATAC-seq analyses of quadruplet 16-90. (**G**)

Correlation matrix of bulk gene expression profiles for five quadruplet sets at three time points. (**H**) Distributions of correlations from (**G**) categorized as intra-individual across time (intra-ind), within quadruplet sets (inter-ind), and across quadruplet sets (cross-quad). (**I**) AUROC distribution from leave-one-out analysis to evaluate the consistency of highly variable genes (HVGs) across quadruplet groups.

Highly Variable Genes Within Sibling Cohorts Are Shared Across Cohorts

We next asked whether certain genes contributed more strongly than others to the observed individuality. If transcriptomic differences among siblings were purely diffuse noise, we would not expect the same genes to vary consistently within different litters. For each litter at each time point, we ranked genes by their coefficient of variation across siblings and selected the top 100 highly variable genes. We then evaluated how well the HVG sets from one litter predicted the HVGs identified in another by calculating AUROC scores across all cross-litter comparisons (Fig. 1I). These scores were high, with a mean AUROC of 0.90, indicating that the same genes tend to show elevated within-litter variability across independent groups. This pattern is unlikely under a random model and suggests that individuality is partly encoded in a subset of genes that reliably exhibit greater variance among genetically identical siblings.

Rank-Ordered Gene Expression Patterns Identify Stable Identity Signatures

We next examined whether individuality was reflected not only in variability at single time points but also in the stability of expression differences across time. To do this, we assessed whether the relative ordering of siblings by gene expression, within a litter, was preserved across the three sampling points. For example, if a sibling had the highest expression of a gene at the first time point, we asked whether it tended to remain highest at the later time points (Fig. 2A). This rank-based approach parallels our earlier analysis of allele-specific imbalance and reduces sensitivity to shifts in absolute expression levels.

Clustering based on raw expression values grouped samples primarily by time point rather than by individual, indicating that global expression levels changed systematically as the armadillos developed and aged over the sampling period (Fig. 2A, left). In contrast, clustering based on rank-ordered expression grouped samples more strongly by individual identity (Fig. 2A, right). Adjusted Rand Index scores confirmed this pattern: rank-based clustering aligned more closely with true individual labels, whereas raw-expression clustering reflected developmental time (Fig. 2B). These results suggest that relative expression ordering captures stable, individual-specific transcriptomic features that are not apparent from absolute expression alone.

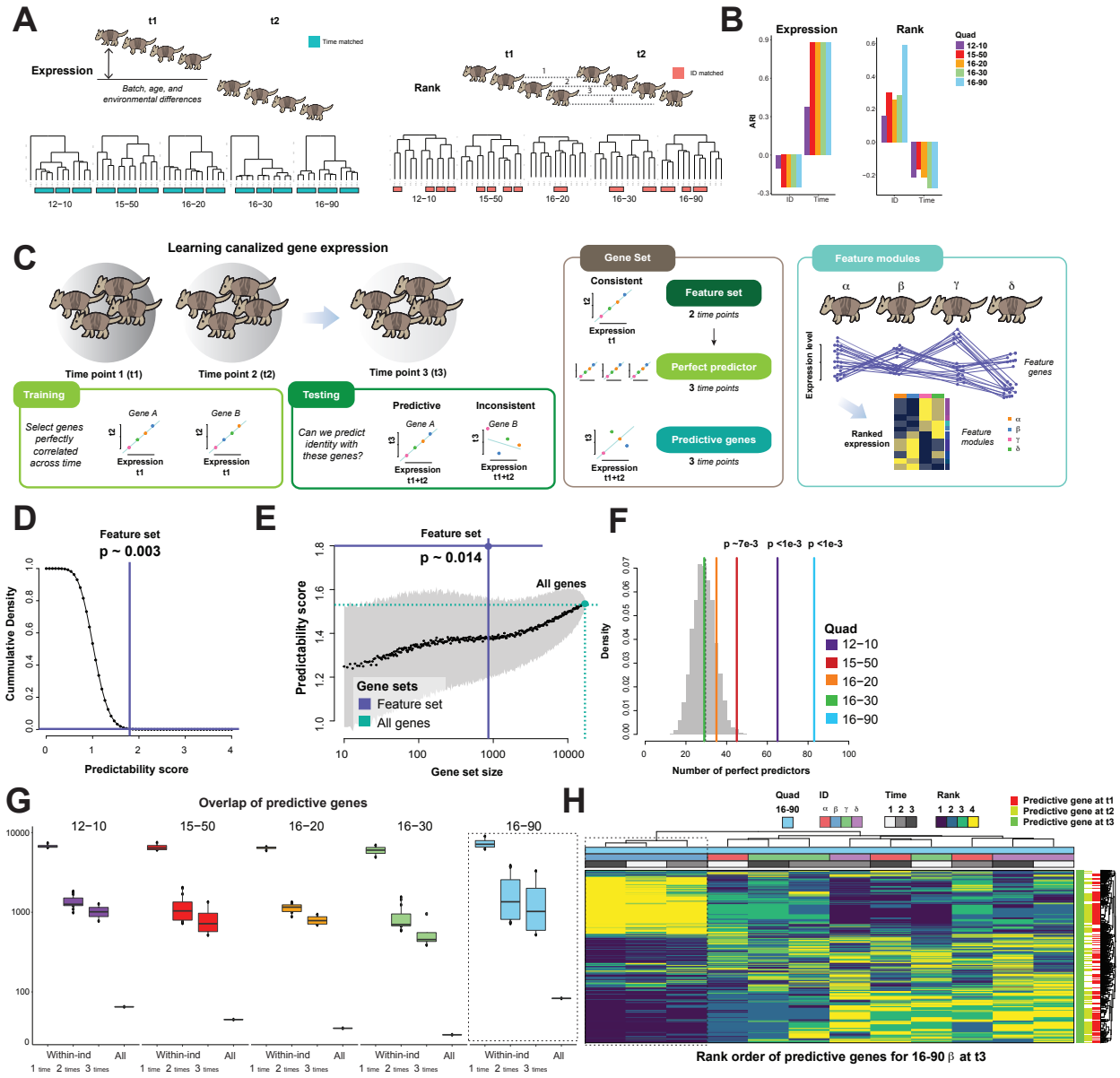


Fig. 2. Gene expression-based identity prediction by sibling-specific signatures. (**A**) Clustering of transcriptome profiles among siblings before (left) and after (right) ranking within a given time point (**B**) Adjusted Rand Index (ARI) measuring how well the samples from the same individual (ID) or time point are assigned to the same cluster as shown in (**A**). (**C**) Workflow illustrating the procedure for predicting individual identity and defining identity-predictable genes. A feature set is optimized for the prediction of each individual to make the gene expression ranking consistent between two time points. A perfect predictor can further predict the third point based on the average rank of two time points. For functional analysis, each gene is classified into the feature modules based on the expression patterns. (**D,E**) Analytical (**D**) and empirical (**E**) null of identity predictability scores with the score of feature gene sets 1.8. (**F**) Number of perfect predictors for each quadruplet with p-values computed by random background distribution (gray). (**G**) Number of common predictive genes within the same individual and within the siblings from five lineages. (**H**) Example of

ranked gene expression profiles for predictive genes identified at time point t_3 from the quadruplet group exhibiting the highest predictability (16-90).

To quantify how well gene expression patterns could identify individual siblings, we built a supervised classifier based on ranked expression (Fig. 2C). For each litter, we used two time points as training data and asked whether the ranked expression of selected genes could correctly match each sibling in the remaining time point. Genes were included as features if their ranked expression was consistent between the two training time points. For each gene, we compared the ordering of the four siblings in the training data with their ordering in the held-out time point. A gene contributed one point for each sibling whose position in the ranked order matched between training and test data. Summing across genes yielded a score for each sibling, and the final identity score for a litter was the number of siblings, out of four, that were correctly assigned to themselves on average across all training–test splits.

Across litters, the classifier achieved a mean identity score of 1.8, which is significantly higher than the score of approximately 1.0 expected by chance (Fig. 2D; empirical p-value ~0.014). Randomly selected gene sets produced much lower scores, and their predictability increased only gradually with gene-set size (Fig. 2E). The strength of the identity signal differed across litters. Two groups, 12-10 and 16-90, had especially strong signatures, each containing more than 60 genes that maintained a perfectly consistent rank order across all three time points (Fig. 2F). These "perfect predictors" did not generally overlap with the highly variable genes identified earlier and were not distinguished by higher average expression (Fig. S2). Stable rank ordering of gene expression therefore identifies individual-specific signatures that differ across litters and are not explained by variability or expression level alone.

Environmental Robustness of Individuality: The Role of Leprosy Exposure

Two of the five quadruplet groups (12-10 and 15-50) were experimentally infected with *Mycobacterium leprae* after the second time point as part of an unrelated study at the National Hansen's Disease Program, where the colony is maintained for leprosy research. Although this exposure was not a focus of our experimental design, it provides a natural test of whether individuality signals persist in the presence of physiological disturbance. Even modest immune challenges can reshape blood transcriptomes, so the infection offers a useful opportunity to examine robustness, even if the perturbation is not severe.

To assess the impact of infection, we applied a k-mer based classifier to detect pathogen-derived sequences in the RNA-seq data (*12*). No *M. leprae* sequences were detected, consistent with early or localized infection, and detection of other organisms did not differ by infection status. Differential expression analysis identified some infection-responsive genes, but these did not overlap with established leprosy-associated signatures (Fig. S3). Although the infection was mild, individuality signatures remained clearly detectable in the exposed litters. This suggests that the identity-specific patterns are stable under at least modest immune perturbation, while leaving open the possibility that stronger challenges could produce additional effects.

Total Expression-Based Individuality is Independent of Allelic Imbalance Signatures

We then asked how the total-expression identity signatures relate to the allele-specific expression (ASE) signatures described in our earlier study (*10*). ASE-based individuality arises from persistent monoallelic biases and can distinguish genetically identical siblings without altering total transcript abundance. Consistent with this expectation, the allelically imbalanced markers identified previously did not show elevated variability in their overall expression across individuals (Fig. S4). These genes therefore maintain stable total expression despite their allelic differences, indicating that the ASE-based and total-expression-based individuality signals represent largely independent axes of variation.

Broadening Identity Signatures Through Predictive Genes

Perfect predictors are highly stringent, since they require an identical rank order across all three time points, and therefore identify only a limited set of genes. To capture weaker but still informative signals, we defined a broader class of "predictive genes." These genes are specific to an individual and a time point, and they correctly rank that individual when the ordering inferred from the other two time points is used as a reference. Unlike perfect predictors, they do not require the remaining siblings to maintain a consistent order.

Using this definition, each litter contained several thousand predictive genes. The numbers ranged from 5,983 in litter 16-30 to 7,380 in litter 16-90 (Fig. 2G). On average, about 1,000 predictive genes were shared across different time points within a given individual, whereas perfect predictors were far fewer and required full consistency across all time points. For illustration, the predictive genes for sibling β of the 16-90 litter at time point t3 showed largely preserved rank ordering at the earlier time points t1 and t2, particularly for genes at the top or bottom of the ranking (Fig. 2H).

Overall, these results show that identity signatures are not confined to the small set of perfect predictors. Instead, thousands of genes contribute to individual-specific ranked expression patterns, with a substantial fraction showing consistency across time within the same sibling.

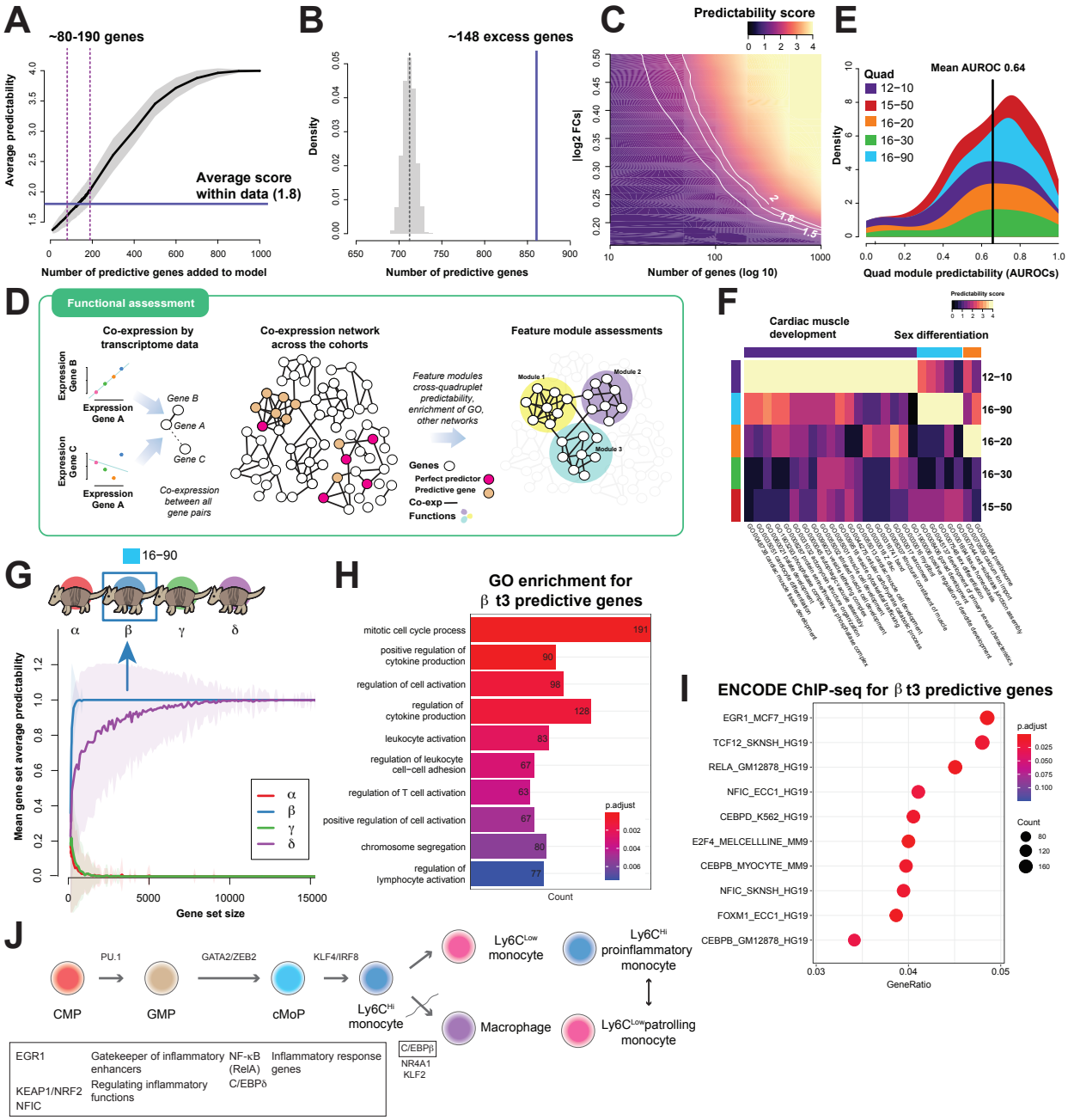


Fig. 3. Functional characterization of identity-predictive gene modules. (**A**) A model and (**B**) our empirical estimates suggest ~150 excess genes would contribute to the identity predictable signal. (**C**) This reflects ~0.3 |log2FC| according to our model. (**D**) Feature genes mark identity and form feature modules (co-expressed). (**E**) Module predictability signal has weak cross-quad predictability. (**F**) These identity genes contribute to known pathways and processes that are also quadruplet specific. **G**) The identity predictability comparison among 16-90 highlights the distinctive predictability of individual β. Enrichment analysis of 16-90 β predictive genes at t_3 for (**H**) GO annotations and (**I**) ENCODE ChIP-seq binding targets for human. (**J**) TFs enriched for 16-90 predictive genes are associated with the differentiation process of monocytes.

Individual Identity Arises from Modest Shifts in a Small Co-expressed Gene Set

To dissect the structure of identity signatures, we asked whether individuality emerges from a handful highly informative "hub" genes or from the cumulative effect of many modestly varying genes. To test this, we performed two complementary simulations based on our empirical model. First, we progressively spiked the expression data with increasing numbers of perfectly predictive genes and calculated the resulting predictability score (Fig. 3A). The empirical predictability score observed in the real data (mean = 1.8) was achieved when incorporating only 80-190 predictive genes. Relative to a rank-order null model, armadillo quadruplets contained an excess of approximately 150 predictive genes (Fig. 3B). In the second simulation, we systematically varied two parameters - the number of affected genes (range: 10 to 1,000) and their absolute effect size ($|\log_2 FC| = 0.1$ to 0.5) - and assessed the resulting predictability (Fig. 3C). The empirical score of 1.8 was reached at 150 genes, each with a modest effect size of about $|\log_2 FC| \approx 0.3$. Thus, subtle yet coordinated expression shifts across a relatively small subset of about 150 genes appear sufficient to encode individual identity.

To investigate whether identity signatures were associated with distinctive gene-gene relationships, we constructed gene co-expression networks for each cohort and time point separately, resulting in a total of 15 networks, which were subsequently aggregated into a single network (Fig. 3D). To assess whether identity-predictive genes exhibited stronger co-expression patterns, we applied EGAD (Extending 'Guilt-by-Association' by Degree), a method designed to evaluate the performance of guilt-by-association predictions within gene networks (*13*). Cross-quadruplet testing revealed that the sets of perfect predictor genes were only modestly predicted by these co-expression relationships (average AUROC = 0.64, Fig. 3E), suggesting limited consistency of interactions among these genes across different quadruplets. Additionally, analysis of expression levels indicated that perfect predictor genes were not distinct in their average expression compared to other genes (Fig. S2).

Functionally, pathway enrichment analysis identified several pathways related to cardiac muscle development and sex differentiation as differentially regulated in specific quadruplet sets (Fig. 3F), whereas their co-expression relationships are conserved between human and armadillo (Fig. S5), highlighting unique biological processes potentially underlying identity-associated variation.

Immunological Phenotypes Linked to Predictable Individual Identity in 16-90

Identity-predictive gene sets overlapped little between the five cohorts, implying that individuality signatures are largely litter-specific. To pinpoint the origins of one such signature, we focused on quadruplet 16-90, which was phenotypically unusual: it exhibited the highest CV across haematological measures and had more perfect predictor genes than expected. Hereafter, we denote the four siblings of quadruplet 16-90 as α, β, γ and δ.

Rather than averaging predictability across the litter, we assessed each sibling separately. By evaluating the individual transcriptomic predictability rather than quadruplet predictability, individuals β and δ showed the greatest predictability (Fig. 3G), with β emerging as easily identifiable, requiring only a small number of genes. Concordantly, β displayed elevated platelet and white-blood-cell counts, consistent with an activated immune profile among the quadruplets (Fig. S6). Predictive

genes for β at t_3 were significantly enriched for GO terms related to cytokine production and to leukocyte and lymphocyte activation (Fig. 3H; Fig. S7), linking β's transcriptomic distinctiveness to an immune-biased physiological state.

To uncover the upstream drivers of this immune-skewed signature, we next examined whether the β-specific identity genes share common regulatory inputs. Promoter analysis revealed significant enrichment of 18 transcription-factor (TF) target sets ($q < 0.05$; Fig. 3I, Fig. S8), with the strongest signals arising from factors that guide blood-cell differentiation - particularly within the monocyte lineage. Across individuals, the most consistently enriched targets belonged to TFs that regulate myeloid development. Foremost among them were targets of EGR1, a master regulator of inflammatory macrophage enhancers and monocyte maturation, implying a central role for EGR1 in establishing divergent expression programs (Fig. 3K and J) (*14–16*). Additional enriched TFs included RELA (NF-κB), CEBPδ, NFIC, and FOXM1, each with well-documented functions in immune and inflammatory pathways (*17–19*). Together, these findings suggest that individuality signatures are orchestrated by specific regulatory networks, particularly those governing innate immune responses.

Combined, our bulk-transcriptomic analyses show that genetically identical armadillos acquire stable, individual-specific expression profiles that persist over time and withstand moderate environmental variation. These individuality signatures are encoded by largely non-overlapping gene sets and can be explained by differences in regulatory networks rather than technical noise. The most striking example is sibling β of the 16-90 litter, whose transcriptomic distinctiveness coincides with elevated white-blood-cell counts observed in the hematological tests, hinting at an underlying shift in immune-cell composition. To investigate this hypothesis at higher resolution, we next employed single-cell sequencing analyses.

Single-Cell Profiling Links Individuality to Immune Cell Composition Differences

To probe the cellular basis of individuality, we collected peripheral blood mononuclear cells from the 16-90 quadruplets, selected for its pronounced bulk-level variability, at a later time point (t_4; Fig. 1F). After stringent filtering, we obtained high-quality profiles for 16,093 single cells by scRNA-seq and 19,003 nuclei by scATAC-seq. The scRNA-seq cells were almost evenly distributed across the four siblings (Fig. 4A; Fig. S9, S10), whereas the scATAC-seq dataset was skewed, with sibling β contributing a disproportionate share of nuclei both before and after filtering (Fig. S11, S12).

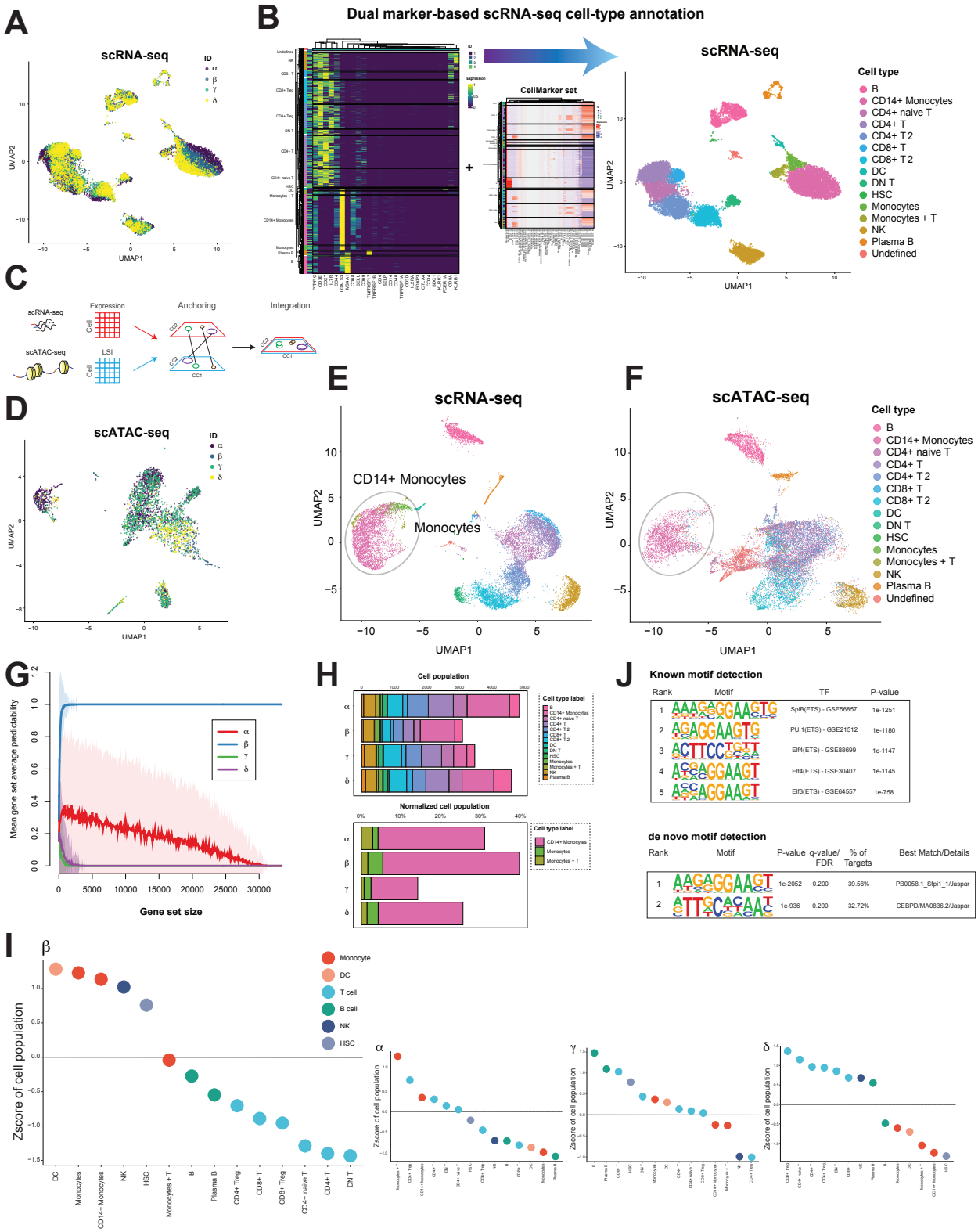


Fig. 4. Single-cell transcriptome and epigenome analysis reveals consistent leukocyte population changes. (A) UMAP embedding of scRNA-seq data from the 16-90 lineage. Each point represents a cell, colored by individual. (B) Each cluster from scRNA-seq data is annotated according to the expression of known marker genes and module activity of CellMarker gene sets. (C) Integration of scRNA-seq and scATAC-seq was performed using Seurat. (D) UMAP embedding of scATAC-seq data. (E,F) UMAP visualization of scRNA-seq (E) and scATAC-seq gene activity profiles (F) in the

integrated space with the predicted cell type annotation transferred from scRNA-seq. (G) The identity predictability comparison among 16-90 based on scRNA-seq pseudo-bulk profile demonstrates the predictability of β. (H) Horizontal stacked bars showing the distribution of cell-type counts for each individual (top) and normalized monocyte-related cell-type proportions (bottom). (I) Deviation of each cell-type population against global population for β and the other three individuals. (J) Motif analysis of CD14+ monocyte cluster using HOMER.

To identify cell types, we leveraged known marker genes and a curated database of cell type signatures. No established armadillo-specific markers for blood cell types exist, and marker gene expression can vary across species. We therefore applied a dual annotation strategy: we manually selected a panel of 27 canonical marker genes for mammalian blood cell types, and we also consulted the CellMarker database of human/mouse cell type markers (*20*). By examining the expression of these markers, we successfully annotated 14 of the 15 transcriptional clusters as known blood cell types (Fig. 4B; one small cluster remained ambiguous). The use of cross-species marker data increased our confidence in these annotations despite species differences.

We next integrated the scRNA-seq and scATAC-seq datasets to corroborate the cell type identities and investigate chromatin differences. Using Seurat-based integration, we co-embedded the transcriptomic and chromatin profiles in a common space (Fig. 4C-F). In this integrated visualization, cells grouped by cell type rather than by modality, indicating that the ATAC profiles of, say, monocytes aligned well with the RNA profiles of monocytes, etc. This allowed us to transfer labels and confirm that the scATAC-seq clusters corresponded to the same cell types identified in scRNA-seq.

To directly link the single-cell findings back to the bulk transcriptome signatures, we created “pseudo-bulk” profiles for each individual by aggregating their single-cell transcriptomes. We then asked whether the identity predictability from the bulk analysis was consistent in the pseudo-bulk profiles. Indeed, applying our supervised learning approach to the pseudo-bulk profiles successfully reproduced the strong predictability of individual β (Fig. 4G), whereas the other individuals were less distinctly classified. The integrated analysis also revealed that the overall cell type composition captured by scATAC-seq was similar, with one caveat: we noticed that the raw cell yields in scATAC-seq showed a slight bias (before filtering, one sibling had fewer high-quality nuclei). After filtering, however, the distribution of cells by individual was roughly even in scRNA-seq and only modestly uneven in scATAC-seq (Fig. S10, S11), so any composition differences are likely biological rather than technical.

With cell type labels in hand for each cell, we calculated the Z-score of each major cell type proportion for each sibling. The results were striking: the four siblings showed noticeable differences in the makeup of their blood immune cells (Fig. 4H). In particular, sibling β (the transcriptomically distinctive individual) had a substantially higher fraction of monocytes compared to the others (Fig. 4I). Monocyte-lineage cells (including classical $CD14^+$ monocytes and a cluster of cells we labeled "Monocytes+T") were overrepresented in β. This pattern is consistent with the blood hematology results including more white blood cells at t2 and t3 (Fig. S13). Correspondingly, β had slightly lower percentages of lymphocytes (T and B cells) than some of its siblings. Notably, sibling δ (the second-most predictable individual in bulk expression) showed a mild increase in T cells relative to the others, and sibling γ showed a mild increase in B cells.

These shifts were smaller than the monocyte expansion in β but followed a pattern: each sibling appeared to have a unique immune cell type profile, with β's deviation in the myeloid compartment being the most pronounced. Thus, the single-cell data confirm that functional differences in blood cell composition exist among the siblings and contribute to their transcriptomic individuality.

Single-Cell Chromatin Accessibility Profiles Support Cell Typing Through TF Motif Enrichment Analysis

We further analyzed the scATAC-seq chromatin landscapes to identify TF motifs enriched within each cell cluster. Due to sampling bias and data sparsity, several populations detected in scRNA-seq - DC, HSC, and the mixed monocyte + T cluster - were absent from the scATAC-seq data. While DN T, Monocytes, and Plasma B clusters were captured in numbers too small for motif analysis. Focusing on clusters annotated as CD14+ monocytes, we found significant enrichment of PU.1 (encoded by SPI1) and Spi-B motifs in differentially accessible regions (DAR); these factors bind to highly similar motifs (Fig. 4J). ELF-family motifs rank within the top five enrichments for CD14+ monocytes—a contrast to human datasets where these motifs are usually more accessible in B cells than in monocytes - an effect we do not observe (Fig. S14). Collectively, the chromatin signatures reinforce our monocyte annotation across the matched scRNA-seq and scATAC-seq profiles.

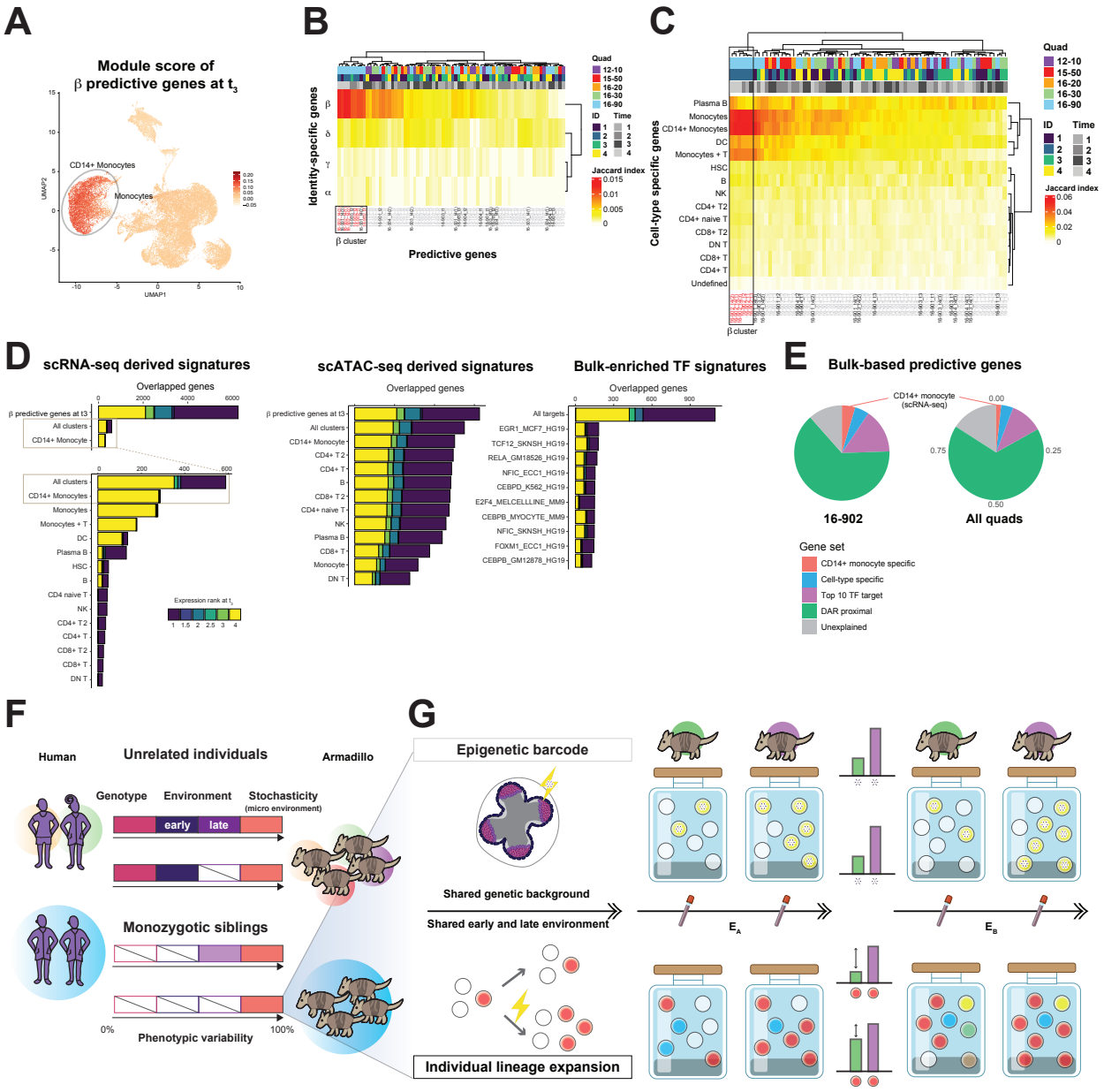


Fig. 5. Cross-modality categorization of predictable genes indicates overlap of monocyte-derived signatures and β predictive genes. (**A**) UMAP of scRNA-seq data at t_3 showing module scores for β predictive genes. (**B,C**) Overlap between the predictive gene sets and sibling (**B**) and cluster-specific (**C**) gene sets shows the enriched signal in β and monocyte-related clusters. The sample from 16-90 β, any other 16-90 sibling, and unrelated individual is shown in red, black, and gray, respectively. (**D**) The ranked expression profiles of overlapping genes with β predictive genes at t_3 for the signatures derived from scRNA-seq (left), scATAC-seq (center) cluster-specific genes and top 10 bulk-enriched TF target genes (right). (**E**) Proportion of β predictive genes overlapping with: CD14+ monocyte-specific genes, other cell-type specific genes in scRNA-seq data, top 10 TF target genes, differentially accessible region (DAR) proximal genes in scATAC-seq data, and none of the above (labelled as 'Unexplained'). (**F**) Phenotypic variability arises from three major sources: genetic variation, environmental variation (both early and late), and stochasticity. (**G**) Two potential

449 mechanisms to promote stochastic developmental canalization: an epigenetic fingerprint
450 established during early development, and lineage expansion occurring in later stages.

451 **Potential Role of Cell-Type Compositional Changes in Driving Transcriptomic**
452 **Identity Signatures**

453 To trace the bulk-level identity signatures to their cellular origins, we compared bulk
454 RNA-seq signatures with the single-cell data from the 16-90 litter. Aggregate expression
455 of β's predictive genes at t_3 and other time points is markedly enriched in monocyte
456 clusters (Fig. 5A; Fig. S15). Intersecting the 16-90 sibling-specific identity genes with
457 the full set of predictive genes identified across all litters and time points (bulk and
458 pseudo-bulk) again yields the strongest overlap for β (Fig. 5B). A parallel comparison
459 with curated cell-type markers shows that β's predictors intersect most with monocyte-
460 specific genes (Fig. 5C).

461 To further characterize the identity signature of sibling β, we examined the distribution
462 of expression ranks for its predictive genes at t_3. The monocyte-specific genes tended to
463 occupy the extremes—either highest (rank 4) or lowest (rank 1)—relative to the other
464 siblings (Fig. 5D). The number of those genes surpassed the number of theoretically
465 estimated identity signatures for observed bulk data, specifically 80-190 or 150 genes by
466 two different simulations, suggesting the monocyte enrichment possesses a sufficient
467 power to make sibling β predictable in bulk. Notably, the subset of monocyte-specific
468 genes within this predictive set was predominantly highly expressed in β, reinforcing the
469 cell-type–biased pattern.

470 In contrast, genes located near differentially accessible regions (DAR-proximal genes)
471 showed more modest differences across cell types, with a rank distribution similar to that
472 of the full β predictive set (Fig. 5D). Although the sizes of these gene sets vary, this
473 trend suggests that peak-level signals from scATAC-seq may not always reflect distinct
474 cell-type–specific regulation, possibly due to redundancy in regulatory elements.
475 Additionally, the targets of the top 10 TFs identified from the bulk data were enriched
476 for genes with extreme expression ranks with a bias toward lower-ranked genes.

477 To integrate these findings, we organized identity-predictive genes for β and all
478 individuals into four mutually exclusive categories, applied in this order: $CD14^+$
479 monocyte-specific genes, genes specific to other cell types, top 10 TF targets, and DAR-
480 proximal genes (Fig. 5E). Among β's predictive genes, 4.6% were classified as $CD14^+$
481 monocyte-specific, compared to just 1.7% across all cohorts, further emphasizing the
482 monocyte-skewed nature of β's transcriptomic identity. Together, these findings indicate
483 that the transcriptomic signature distinguishing sibling β is driven primarily by a
484 monocyte-associated gene program.

485

486 **Discussion**

487 Our results show that transcriptomic individuality in genetically identical animals
488 corresponds to stable differences in immune-cell composition. In nine-banded armadillo
489 quadruplets raised under controlled conditions, each sibling developed a distinct gene
490 expression profile that persisted over time. These profiles were not limited to neutral
491 molecular markers. Instead, they reflected coordinated differences in cell-type
492 abundance that were evident at both bulk and single-cell levels. The most striking case

was sibling β of the 16-90 litter, which displayed a consistent expansion of monocyte-lineage cells and an associated innate immune program. This establishes that non-genetic individuality can produce durable physiological divergence in immune composition.

Our earlier work identified allele-specific expression (ASE) imbalance as a stable molecular fingerprint of individuality in the same animals (*10*). That signal arose from persistent monoallelic biases and did not require changes in total gene output. In contrast, the individuality described here is driven by differences in total expression levels. We found little overlap between ASE-based and total-expression-based signatures, which suggests that they represent independent axes of non-genetic variation. ASE variation likely captures neutral epigenetic differences, while the expression signatures identified here reflect functional shifts that influence cell physiology and abundance.

The nature of these shifts offers insight into possible mechanisms. The single-cell data point to subtle differences in lineage production or survival rather than dramatic reprogramming. Modest changes affecting about 150 genes were sufficient to recapitulate the observed identity signal in simulation, and these genes exhibited coordinated but small expression shifts. These patterns are compatible with the idea that early stochastic events in hematopoietic development bias the relative representation of particular lineages. Such processes need not involve somatic mutations. Reanalysis of ASE data found no evidence for reproducible allelic imbalance at canonical clonal hematopoiesis genes, and the animals were much younger than the age at which clonal expansions typically arise in species with comparable lifespans. Although somatic mosaicism cannot be fully excluded, the available evidence suggests that the individuality observed here derives mainly from developmental variation rather than mutational clonality.

The broader implications of these findings relate to non-genetic individuality more generally. Stable biases in immune-cell composition have been documented among monozygotic human twins, and these differences are often attributed to non-heritable influences early in life (*1*, *21*). Our results provide a model in which such variation can arise without genetic or major environmental differences (Fig. 5F and G). Early stochastic cues during development can produce long-lasting shifts in lineage balance, which then manifest as stable differences in transcriptomes and immune profiles. This framework provides a potential explanation for why genetically identical individuals sometimes diverge in disease susceptibility, including autoimmune and inflammatory conditions. More speculatively, developmental sources of heterogeneity among genetically identical individuals could serve as a form of bet hedging, increasing the resilience of a population to diverse pathogens (*22*). Although our study was not designed to test adaptive hypotheses, the presence of individualized immune profiles in a clonal sibling group demonstrates that developmental stochasticity can generate phenotypic diversity with potential functional consequences.

In summary, genetically identical armadillos raised in nearly identical environments acquire distinct and stable transcriptomic signatures that reflect coherent differences in immune-cell composition. These differences are consistent over time and are detectable at both bulk and single-cell resolution. They arise independently of ASE variation and do not require somatic mutations in known clonal expansion pathways. Together, these findings illustrate how early stochastic influences can become embedded in the transcriptomic and cellular landscape, giving rise to lasting and biologically meaningful individuality.

Materials and Methods

543 **Armadillo Collection and Sample Acquisition**
544 Five sets of nine-banded armadillo (*Dasypus novemcinctus*) quadruplets (20 animals
545 total) were used in this study. These animals were obtained and reared under controlled
546 conditions as described previously (*10*). Briefly, pregnant females were captured from
547 the wild and immediately transported to the holding facility at the University of the
548 Ozarks, Clarksville, Arkansas. The pups were kept with the mothers until about 6-10
549 weeks postnatal age, where most litters shared the pen with another litter (2 litters per
550 pen). After separation from the mothers, at four to five months of age, the animals were
551 delivered to the National Hansen's Disease Program (NHDP) facility in Baton Rouge,
552 LA, where they were placed in pairs in modified rabbit cages.
553 Blood samples for transcriptome analysis were collected at three time points per
554 quadruplet staggered over the course of a year (Fig. 1F), starting from March 2017 until
555 August 2018. The pilot study was performed for two sets of quadruplets (12-10 and 15-
556 50), and then later quadruplets 16-20, 16-30 and 16-90 were included into the subjects.
557 Only for 16-90, the blood samples were additionally extracted for single-cell analyses in
558 Aug 2019. The original two of the five sets of armadillo quadruplets were infected
559 intravenously in the saphenous vein with $1x10^9$ *Mycobacterium leprae* after the first
560 time point. The animals were humanely sacrificed when they developed heavy *M. leprae*
561 dissemination with severe hypochromic microcytic anemia. Otherwise, the animals will
562 eventually succumb to secondary complications of persistent bacteremia because of the
563 bacteria located in the bone marrow.
564
565 **Human twin samples**
566 Transcriptomic data from blood samples in the TwinsUK cohort were used for the
567 human twin analysis (*23*). Briefly, BAM files were obtained from the EUROBATS
568 project (EGAD00001001088) following approval for access to controlled data. The
569 dataset includes 391 blood samples collected from monozygotic (MZ) and dizygotic
570 (DZ) female twin pairs. After quality control, 66 MZ and 96 DZ twin pairs were used for
571 our analysis.
572
573 **Armadillo DNA- and RNA-sequencing data**
574 We used RNA-seq data generated in our previous study (*10*), which is publicly available
575 through the Gene Expression Omnibus (GEO) under accession number GSE141951,
576 with further details on read processing and genome construction provided in that study.
577 For quantification, RNA-seq reads were aligned to personalized genomes constructed
578 from matched DNA-seq data using STAR. Transcript abundance was normalized as
579 counts per million (CPM) for comparison across samples.
580
581 **Armadillo single-cell RNA-seq and single-cell ATAC-seq**
582 For each sample, single-cell RNA-seq (scRNA-seq) libraries were prepared with the 10x
583 Genomics Chromium Single Cell 3′ Gene Expression (Next GEM) kit, targeting ~20,000
584 cells per sample. Single-cell ATAC-seq (scATAC-seq) libraries were generated with the
585 10x Genomics Chromium Single Cell ATAC Library & Gel Bead Kit (Next GEM).
586 Multiplexed paired-end sequencing (PE76) was carried out on an Illumina NextSeq 500
587 on multiple flow cells.
588 Demultiplexing and barcode processing were performed using Cell Ranger (v3.0.2, 10x
589 Genomics), with each sample processed individually. Reads were mapped to the
590 armadillo genome (DasNov3, Ensembl release v95), and genome indices were generated
591 using STAR (v2.7) (*24*). For scRNA-seq analysis, we used STARsolo with the FASTQ
592 files as input to generate gene-count matrices (*25*). Following careful inspection of the

593 cell-typing results, we decided to replace the Ensembl gene definition of CD4
594 (JH562564:228,438–255,355) with the RefSeq annotation (JH562564:226,782–
595 265,885). This decision was based on the weak signal observed in our annotated scRNA-
596 seq data with the Ensembl definition, likely due to a >1,600 nucleotide difference
597 between the transcription start sites (TSS) of the two annotations. Raw sequencing data
598 and processed object files of scRNA-seq and scATAC-seq analysis are publicly
599 available at the GEO under accession number GSE273197.
600
601 **Single-cell RNA-seq data analysis**
602 Gene count matrices from PBMC samples of the 16-90 siblings were processed using
603 the Scanpy toolkit (*26*). After preprocessing, 16,093 cells were retained for downstream
604 analysis and assigned to 15 clusters using Leiden clustering (*27*). Detailed preprocessing
605 steps are provided in the accompanying Jupyter notebook available in our GitHub
606 repository. Summary statistics and basic properties of the scRNA-seq data are presented
607 in Fig. S11 and S12.
608
609 **Single-cell ATAC-seq analysis**
610 Raw BAM files were imported and converted into Snap objects using SnapATAC
611 (v1.0.0) (*28*). From the mapped reads, SnapATAC generated bin-by-cell, peak-by-cell,
612 and gene-by-cell matrices. For the bin-by-cell matrix, a bin size was set to 5 kb. Peak-
613 by-cell matrices were constructed using peak calls from MACS2. To generate gene-by-
614 cell matrices, the 1 kb bin-by-cell matrix was first binarized, and counts were summed
615 for each Ensembl gene (and for CD4, the RefSeq annotation was used) by aggregating
616 bins located within the gene body and promoter region (defined as 2,000 bp upstream of
617 the transcription start site).
618 To remove low-quality cells, we applied a filtering threshold of ≥1,500 reads per cell
619 and a UQ score ≥2.35, representing the number of unique molecular identifiers (UMIs)
620 in the mapped reads. After filtering, 19,003 cells were retained for downstream
621 integration with the scRNA-seq data. Normalization, clustering, and UMAP
622 visualization were performed with the SnapATAC software package. Summary statistics
623 and quality metrics of the scATAC-seq data are shown in Fig. S11 and S12.
624
625 **Cell-type annotation**
626 To assign cell-type identities to the 15 scRNA-seq clusters, we first computed a pseudo-
627 bulk expression profile for each cluster and visualized expression patterns of manually
628 curated marker genes (Fig. 4B). After aligning gene synonyms and excluding genes with
629 low or missing expression, 27 marker genes were used to identify blood cell types:
630 PTPRC, CD3D, CD3E, IL7R, IL2RA, CD4, FOXP3, CTLA4, CD8A, CD27, MS4A1,
631 TNFRSF17, SDC1, FCER1A, CD40, KLRB1, KLRK1, CD34, SELL, CD14, SELP,
632 LGALS3, CD68, CD44, CD69, TNFRSF1A, and TNFRSF1B. As a complementary
633 approach, we retrieved marker gene sets for human and mouse blood cells from the
634 CellMarker database (*20*). To assess the enrichment of these gene sets among
635 upregulated genes in each cluster, we computed module scores using the
636 AddModuleScore function in Seurat v4 (*29*) (Fig. 4B, middle panel).
637
638 **Single-cell integrative analysis and label transfer**
639 To jointly embed the scRNA-seq and scATAC-seq datasets, we performed canonical
640 correlation analysis (CCA)-based integration using Seurat (*29*). For the scATAC-seq
641 data, gene-level chromatin activity was represented by the gene-by-cell matrix generated
642 by SnapATAC, which was incorporated as a feature in the corresponding Seurat object.

Cell-type labels were then transferred from the annotated scRNA-seq data to the scATAC-seq cells using Seurat's FindTransferAnchors and TransferData functions.

Measuring transcriptional stochasticity

To assess transcriptional stochasticity, we conducted a transcriptome-wide clustering analysis of the samples using Spearman's rank correlation to quantify pairwise similarity. Within each quadruplet set, highly variable genes (HVGs) were identified by ranking genes based on their coefficient of variation and mean expression levels. To evaluate the recurrence of transcriptional variability across individuals, we assessed the overlap of HVGs between different quadruplet sets. For each set, we selected the top 100 HVGs at a given time point and computed their AUROC scores in cross-quadruplet comparisons. Additionally, we performed ANOVA across all samples within each quadruplet to identify genes exhibiting the highest intra-quadruplet variability (Fig. S6).

Supervised learning to quantify the quadruplet identity predictability from expression

To evaluate the significance of transcriptomic individuality, we applied a supervised learning framework designed to evaluate the relative consistency of gene expression across time points, as described previously (*10*). Briefly, two time points were selected to form a training set, and the rest is left for testing. For each gene, a gene scoring matrix (4 by 4) is built by comparing the ordering of the test and training CPM data averaged for two samples. Each individual gives a score of 1 to the test data individual whose rank it matched, and 0 otherwise. Then all the matrices for the selected genes feature are aggregated to produce one scoring matrix, where we associate each individual in the test data with that in training according to the scores following a winner-takes-all strategy. We repeat this three times for the different combinations of the training and test data. The final score of "predictability score" is between 0 and 4, with 4 as perfect predictability i.e., each armadillo correctly identifies its future (or past) self. The feature set consists of the genes whose order is consistent across the training samples without the information of test data while a perfect predictor gene is perfectly correlated across all time points. We obtained analytic and empirical p-values for this score by convolution of the expected distributions or repeating the learning task on randomly selected genes, respectively.

Empirical models to estimate the number of genes and lineage contributions

To investigate the underlying mechanisms driving individual identity signatures, we estimated the number of genes contributing to the observed signal using a series of empirical models (Fig. 3A–C). To generate empirical distributions of expression-based identity scores, we simulated two models. In the first model, we assumed the identity signal is localized, and introduced a perfectly predictable signal to a subset of genes in the original expression matrix. In contrast, the second model assumed the identity signal is distributed across a larger number of genes, each contributing a weaker effect. Specifically, we simulated expression-based identity predictability by adding a fixed proportion of variance (denoted M) to a defined number of genes (N). For each simulation, the average identity prediction performance was computed. The introduced signal strength (M) was converted to an absolute log2 fold change (|log2FC|) for interpretability.

Gene co-expression network construction and functional module evaluation

For each quadruplet at each time point, we constructed a gene co-expression network (*30*) (Fig. 3D). Briefly, Spearman correlation coefficients were calculated for all gene pairs based on expression across the four individuals within each quadruplet. These correlation values were used as edge weights in a gene-gene co-expression network, which was subsequently rank-standardized. The resulting individual networks (15 in total) were summed to produce an aggregate co-expression network. For all networks, we first applied unsupervised clustering analyses and tested for functional enrichment of the network modules identified by using our supervised machine learning tool EGAD (*13*) on the aggregate network and the individual networks. In EGAD, we report performance using AUROCs, as is typical within machine learning (Fig. 3E). An AUROC is the area under the receiver operator characteristic curve, and is the probability that a positive result is ranked higher than negative result by the algorithm (e.g., does this list of DE genes rank genes involved with "function X" higher than other genes). We used the functional annotations we mapped from GO for this assessment. To test for cross-quadruplet generalizability, we employed a leave-one-out validation approach: individual networks were aggregated while excluding one quadruplet, and modules predictive of identity in the excluded set were evaluated for performance in the aggregate network using EGAD (Fig. 3E).

Functional analysis of identity signatures in the 16-90 Cohort

Predictability scores were calculated for each individual in the 16-90 cohort and averaged across three different combinations of training and test time points (Fig. 3G). To functionally characterize genes contributing to individual identity signatures in this cohort, we performed enrichment analysis using the R interface to enrichR. Gene Ontology (GO) and ENCODE TF ChIP-seq 2015 were selected as reference databases for this analysis (*31*, *32*).

Detecting the significance of cell-type compositional variation

After cell-type annotation of scRNA-seq, we assessed whether specific cell types were significantly enriched in specific individuals. We converted the percentage of each cell-type composition in the individual i c_i into the z-score. Using the cell type composition data of all four siblings, the z-score was computed based on the equation $z_i = (c_i - \mu)/\sigma$, where μ and σ is the mean and standard deviation of the distribution of c across four quadruplets, respectively.

Enrichment of identity signatures across gene sets

The functional classification of the identity signatures was carried out by comparing the gene sets obtained in this study. To evaluate the overlap of two gene sets g_i and g_j., Jaccard index is obtained by $Jaccard(g_i, g_j) = \frac{|g_i \cap g_j|}{|g_i \cup g_j|}$. First, we compared the identity predictive genes obtained from bulk time course data with those from scRNA-seq pseudo-bulk data. For cell-type specific genes, top 1,000 genes were selected for each cell type as the cluster-specific genes based on Wilcoxon rank-sum test p-values.

Disclaimer

The content and views expressed in this article are solely the opinions of the authors and do not necessarily reflect the official policies of the National Institutes of Health, the U.S. Department of Health and Human Services or the Health Resources and Services Administration, nor does mention of the department or agency names imply endorsement by the U.S. Government.

741
742 **References**
743

744 1. J. E. Castillo-Fernandez, T. D. Spector, J. T. Bell, Epigenetics of discordant monozygotic
745 twins: implications for disease. *Genome Med* **6**, 60 (2014).

746 2. C. Bycroft, C. Freeman, D. Petkova, G. Band, L. T. Elliott, K. Sharp, A. Motyer, D.
747 Vukcevic, O. Delaneau, J. O'Connell, A. Cortes, S. Welsh, A. Young, M. Effingham, G.
748 McVean, S. Leslie, N. Allen, P. Donnelly, J. Marchini, The UK Biobank resource with
749 deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

750 3. T. J. C. Polderman, B. Benyamin, C. A. De Leeuw, P. F. Sullivan, A. Van Bochoven, P.
751 M. Visscher, D. Posthuma, Meta-analysis of the heritability of human traits based on fifty
752 years of twin studies. *Nat Genet* **47**, 702–709 (2015).

753 4. P. Bateson, D. Barker, T. Clutton-Brock, D. Deb, B. D'Udine, R. A. Foley, P. Gluckman,
754 K. Godfrey, T. Kirkwood, M. M. Lahr, J. McNamara, N. B. Metcalfe, P. Monaghan, H. G.
755 Spencer, S. E. Sultan, Developmental plasticity and human health. *Nature* **430**, 419–421
756 (2004).

757 5. A. Raj, A. van Oudenaarden, Nature, nurture, or chance: stochastic gene expression and its
758 consequences. *Cell* **135**, 216–226 (2008).

759 6. C. H. Waddington, Canalization of development and genetic assimilation of acquired
760 characters. *Nature* **183**, 1654–1655 (1959).

761 7. S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, P. K. Sorger, Non-genetic origins of
762 cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–432 (2009).

763 8. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a
764 single cell. *Science* **297**, 1183–1186 (2002).

765 9. A. Gimelbrant, J. N. Hutchinson, B. R. Thompson, A. Chess, Widespread monoallelic
766 expression on human autosomes. *Science* **318**, 1136–1140 (2007).

767 10. S. Ballouz, R. K. Kawaguchi, M. T. Pena, S. Fischer, M. Crow, L. French, F. M. Knight,
768 L. B. Adams, J. Gillis, The transcriptional legacy of developmental stochasticity. *Nat*
769 *Commun* **14**, 7226 (2023).

770 11. S. Jaiswal, B. L. Ebert, Clonal hematopoiesis in human aging and disease. *Science* **366**,
771 eaan4673 (2019).

772 12. K. S. Katz, O. Shutov, R. Lapoint, M. Kimelman, J. R. Brister, C. O'Sullivan, STAT: a
773 fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation
774 sequence submissions. *Genome Biol* **22**, 270 (2021).

775 13. S. Ballouz, M. Weber, P. Pavlidis, J. Gillis, EGAD: ultra-fast functional analysis of gene
776 networks. *Bioinformatics* **33**, 612–614 (2017).

777 14. G. Genovese, A. K. Kähler, R. E. Handsaker, J. Lindberg, S. A. Rose, S. F. Bakhoum, K.
778 Chambert, E. Mick, B. M. Neale, M. Fromer, S. M. Purcell, O. Svantesson, M. Landén, M.
779 Höglund, S. Lehmann, S. B. Gabriel, J. L. Moran, E. S. Lander, P. F. Sullivan, P. Sklar, H.

780 Grönberg, C. M. Hultman, S. A. McCarroll, Clonal hematopoiesis and blood-cancer risk
781 inferred from blood DNA sequence. *N Engl J Med* **371**, 2477–2487 (2014).

782 15. M. Trizzino, A. Zucco, S. Deliard, F. Wang, E. Barbieri, F. Veglia, D. Gabrilovich, A.
783 Gardini, EGR1 is a gatekeeper of inflammatory enhancers in human macrophages. *Sci.*
784 *Adv.* **7** (2021).

785 16. L. Ziegler-Heitbrock, P. Ancuta, S. Crowe, M. Dalod, V. Grau, D. N. Hart, P. J. M.
786 Leenen, Y.-J. Liu, G. MacPherson, G. J. Randolph, J. Scherberich, J. Schmitz, K.
787 Shortman, S. Sozzani, H. Strobl, M. Zembala, J. M. Austyn, M. B. Lutz, Nomenclature of
788 monocytes and dendritic cells in blood. *Blood* **116**, e74–e80 (2010).

789 17. Z. Zhang, M. Li, T. Sun, Z. Zhang, C. Liu, FOXM1: Functional Roles of FOXM1 in Non-
790 Malignant Diseases. *Biomolecules* **13**, 857 (2023).

791 18. M. Braunstein, M. K. Anderson, HEB in the spotlight: Transcriptional regulation of T-cell
792 specification, commitment, and developmental plasticity. *Clin Dev Immunol* **2012**, 678705
793 (2012).

794 19. P. O. Humbert, C. Rogers, S. Ganiatsas, R. L. Landsberg, J. M. Trimarchi, S. Dandapani,
795 C. Brugnara, S. Erdman, M. Schrenzel, R. T. Bronson, J. A. Lees, E2F4 is essential for
796 normal erythrocyte maturation and neonatal viability. *Mol Cell* **6**, 281–291 (2000).

797 20. C. Hu, T. Li, Y. Xu, X. Zhang, F. Li, J. Bai, J. Chen, W. Jiang, K. Yang, Q. Ou, X. Li, P.
798 Wang, Y. Zhang, CellMarker 2.0: an updated database of manually curated cell markers in
799 human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res* **51**, D870–D876
800 (2023).

801 21. H. Jonsson, E. Magnusdottir, H. P. Eggertsson, O. A. Stefansson, G. A. Arnadottir, O.
802 Eiriksson, F. Zink, E. A. Helgason, I. Jonsdottir, A. Gylfason, A. Jonasdottir, A.
803 Jonasdottir, D. Beyter, T. Steingrimsdottir, G. L. Norddahl, O. Th. Magnusson, G. Masson,
804 B. V. Halldorsson, U. Thorsteinsdottir, A. Helgason, P. Sulem, D. F. Gudbjartsson, K.
805 Stefansson, Differences between germline genomes of monozygotic twins. *Nat Genet* **53**,
806 27–34 (2021).

807 22. A. T. Tate, J. Van Cleve, Bet-hedging in innate and adaptive immune systems. *Evol Med*
808 *Public Health* **10**, 256–265 (2022).

809 23. S. Verdi, G. Abbasian, R. C. E. Bowyer, G. Lachance, D. Yarand, P. Christofidou, M.
810 Mangino, C. Menni, J. T. Bell, M. Falchi, K. S. Small, F. M. K. Williams, C. J. Hammond,
811 D. J. Hart, T. D. Spector, C. J. Steves, TwinsUK: The UK Adult Twin Registry Update.
812 *Twin Res Hum Genet* **22**, 523–529 (2019).

813 24. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M.
814 Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**,
815 15–21 (2013).

816 25. B. Kaminow, D. Yunusov, A. Dobin, STARsolo: accurate, fast and versatile
817 mapping/quantification of single-cell and single-nucleus RNA-seq data. [Preprint] (2021).
818 https://doi.org/10.1101/2021.05.05.442755.

26. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**, 15 (2018).

27. V. A. Traag, L. Waltman, N. J. van Eck, From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).

28. R. Fang, S. Preissl, Y. Li, X. Hou, J. Lucero, X. Wang, A. Motamedi, A. K. Shiau, X. Zhou, F. Xie, E. A. Mukamel, K. Zhang, Y. Zhang, M. M. Behrens, J. R. Ecker, B. Ren, Comprehensive analysis of single cell ATAC-seq data with SnapATAC. *Nat Commun* **12**, 1337 (2021).

29. Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, P. Hoffman, M. Stoeckius, E. Papalexi, E. P. Mimitou, J. Jain, A. Srivastava, T. Stuart, L. M. Fleming, B. Yeung, A. J. Rogers, J. M. McElrath, C. A. Blish, R. Gottardo, P. Smibert, R. Satija, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

30. S. Ballouz, W. Verleyen, J. Gillis, Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* **31**, 2123–2130 (2015).

31. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).

32. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

Acknowledgments

We thank Paul Pavlidis, Tony Zador, Peter Koo, Megan Crow and Dan Levy for helpful discussion and comments. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Armadillo blood for the study was obtained from the National Hansen's Disease Program. TwinsUK is funded by the Wellcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. We used ChatGPT for AI-assisted editing.

Funding:

National Institutes of Health R01LM012736 (JG)
National Institutes of Health R01MH113005 (JG)
National Institute of Allergy and Infectious Diseases Interagency Agreement IAA 15006-004 (MTP, LBA)
HRSA/NIH/NIAID/ IAA AAI25003-001-00000 (MTP)
JSPS KAKENHI Grant Number 23K14165 (RKK)
The Uehara Memorial Foundation Postdoctoral fellowship (RKK)
CiRA foundation research grant 2022 startup (RKK)

iPS academia Japan research grant 2022 (RKK)

Author contributions:

Conceptualization: RKK, SB, JG
Methodology: RKK, SB, JG
Armadillo collection: FMK
Blood collection: MTP
Investigation: RKK, SB, MTP, LF, JG
Visualization: RKK, SB
Supervision: RKK, SB, LF, JG
Writing—original draft: RKK, SB, LF, JG
Writing—review & editing: RKK, SB, MTP, LF, FMK, LBA, JG

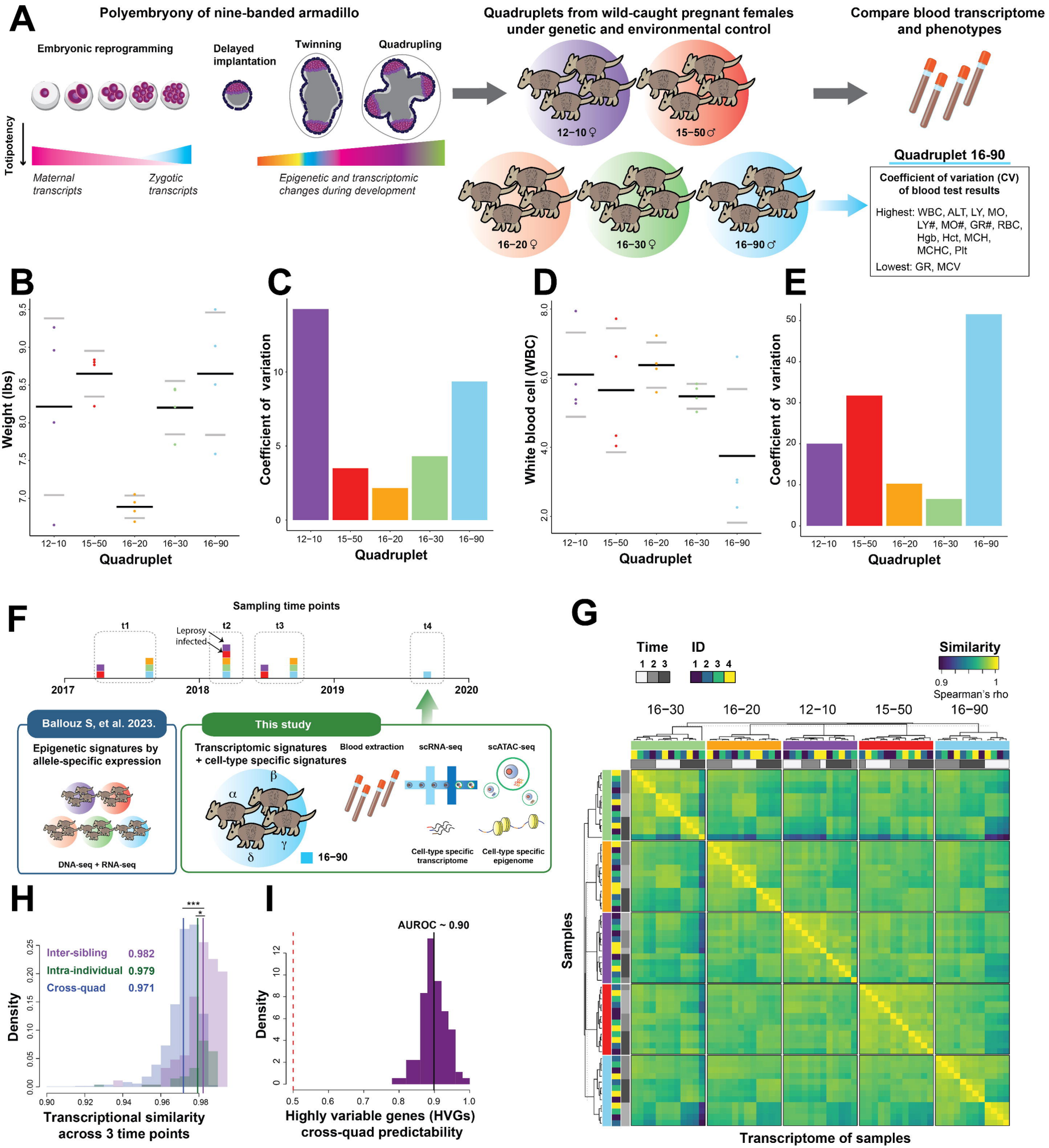
All authors discussed, critically revised, and approved the final version of the manuscript.

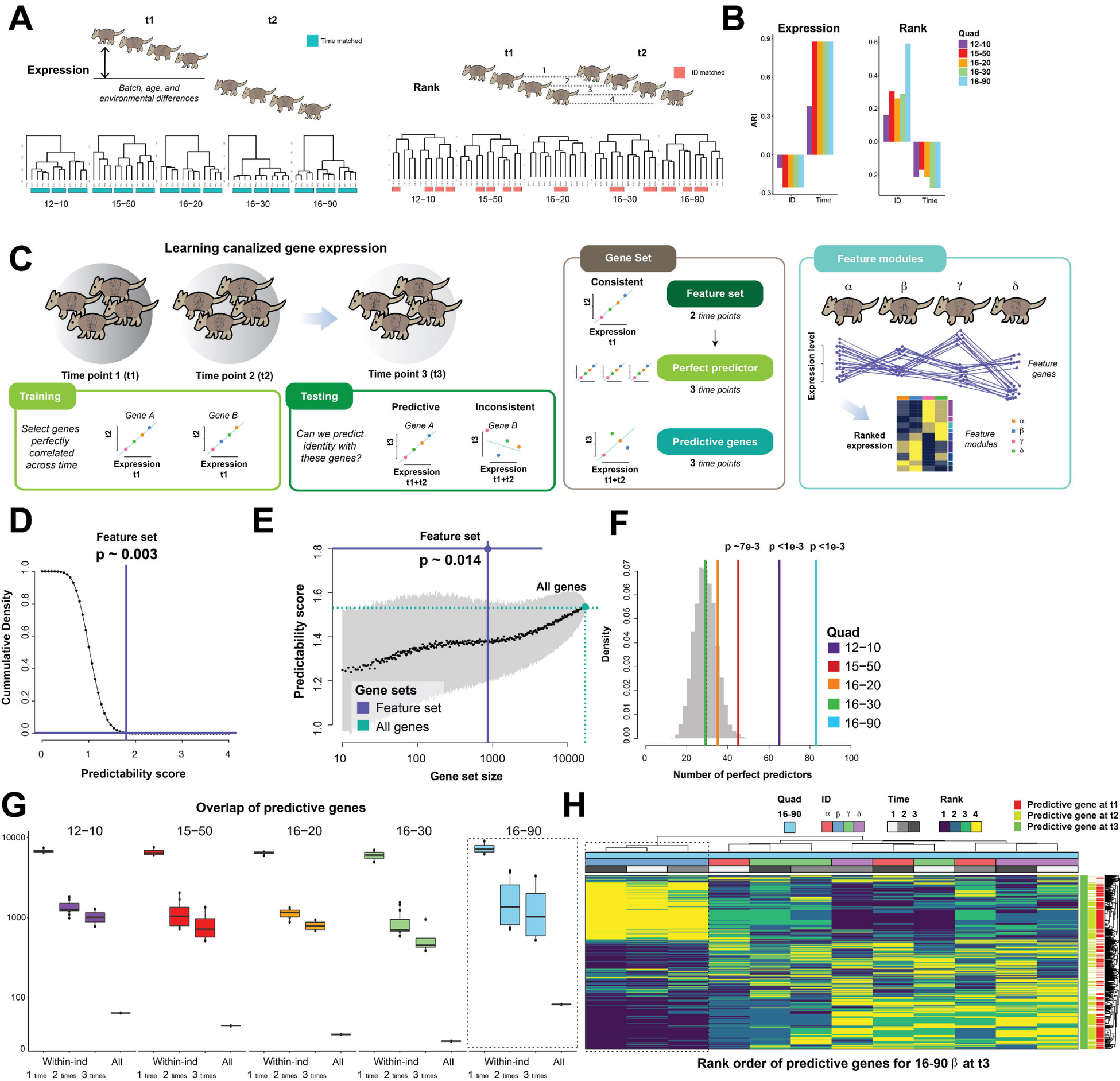
Competing interests: L.F. serves as a scientific advisor for Lighthouse Pharma, and has received consulting fees from PeopleBio Co. and GC Therapeutics Inc. The remaining authors declare no competing interests.

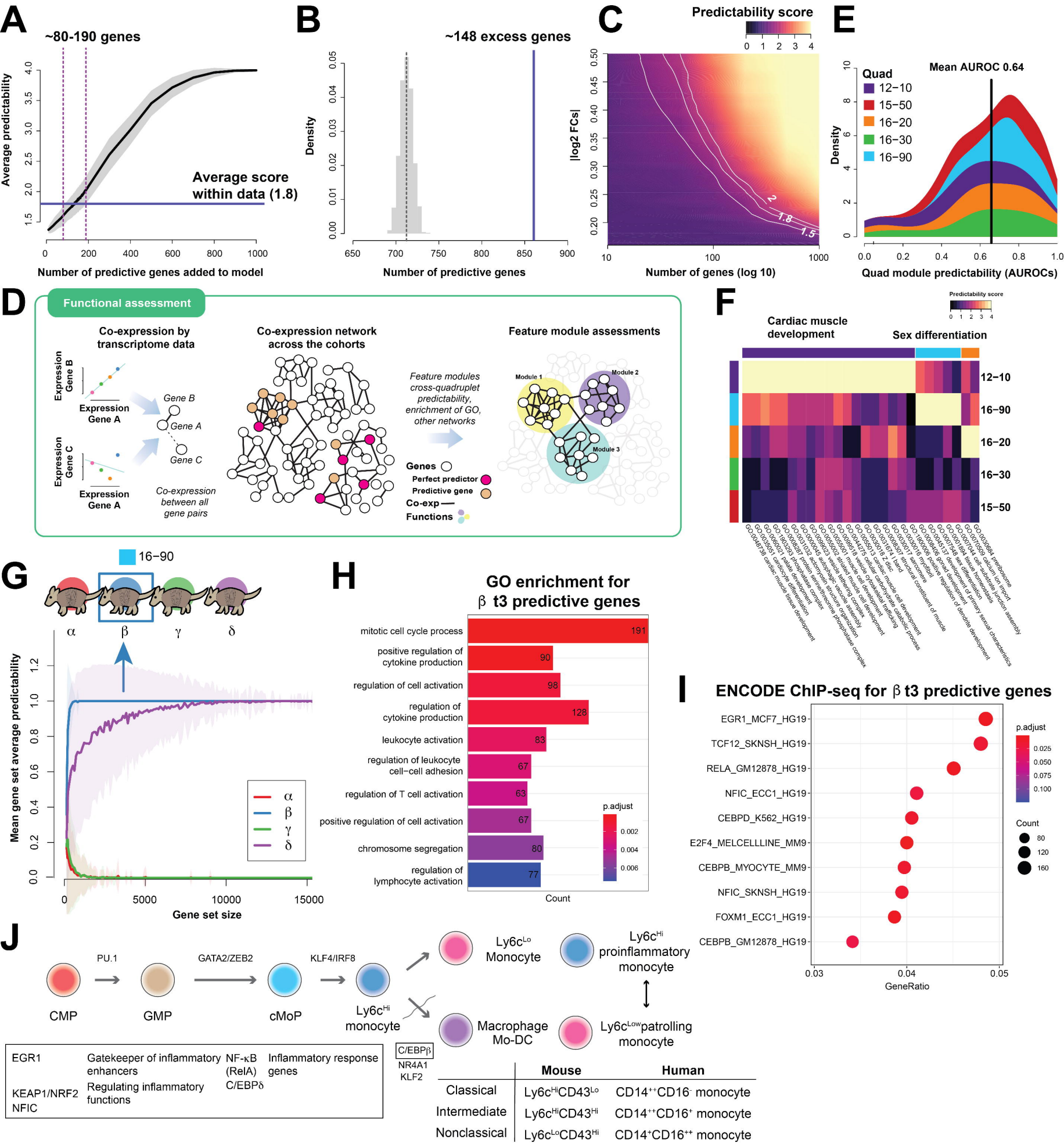
Data and materials availability: RNA-seq data repurposed from our previous study (*10*) is published at GEO under accession number GSE141951. Raw sequencing data and processed object files of scRNA-seq and scATAC-seq analysis are available under accession number GSE273197. Armadillo's trait data such as body size and blood test results were published together with the previous study (*10*). Other intermediate files can be downloaded or recreated by running the scripts found at https://github.com/carushi/ayotochtli.

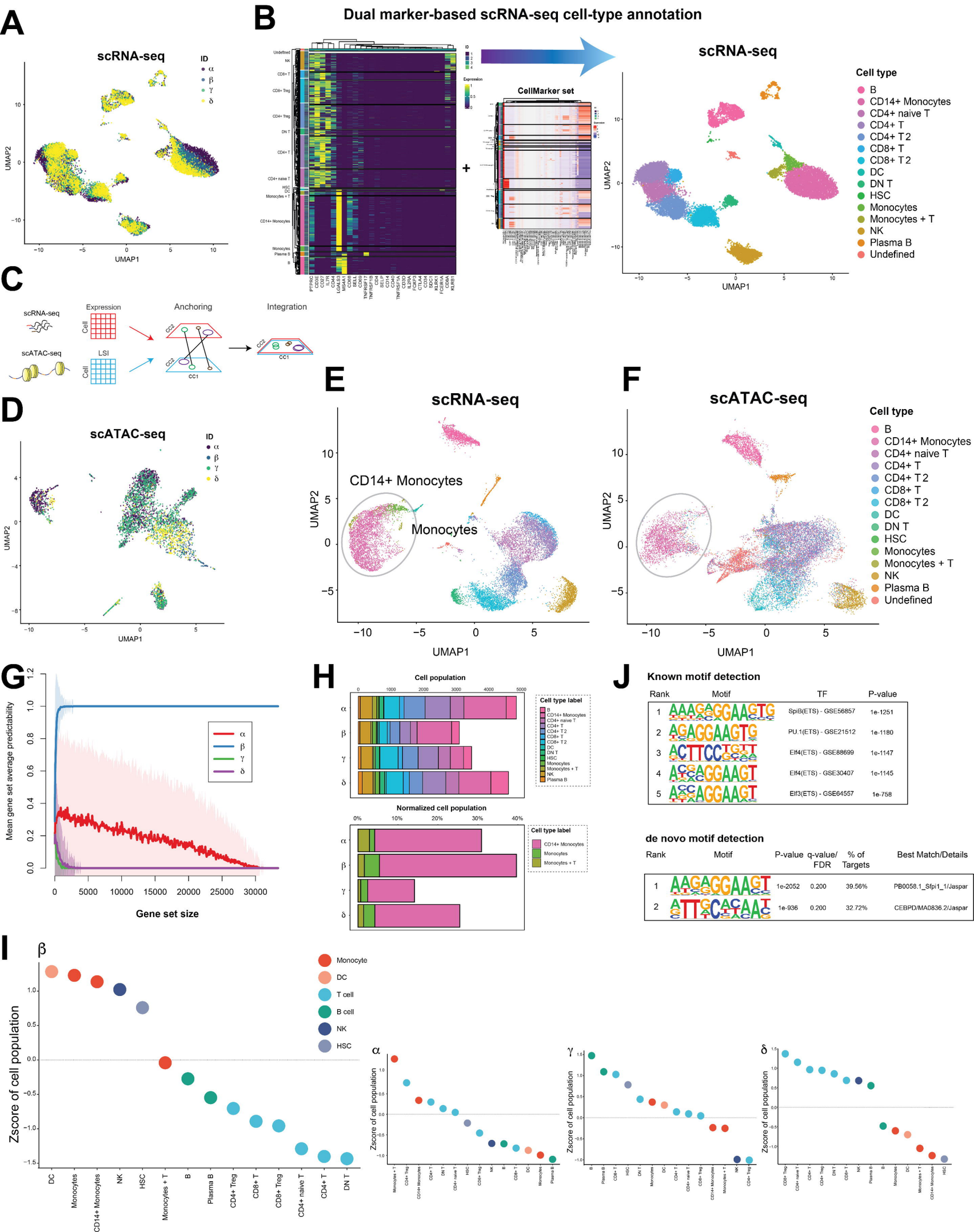
Supplementary Materials

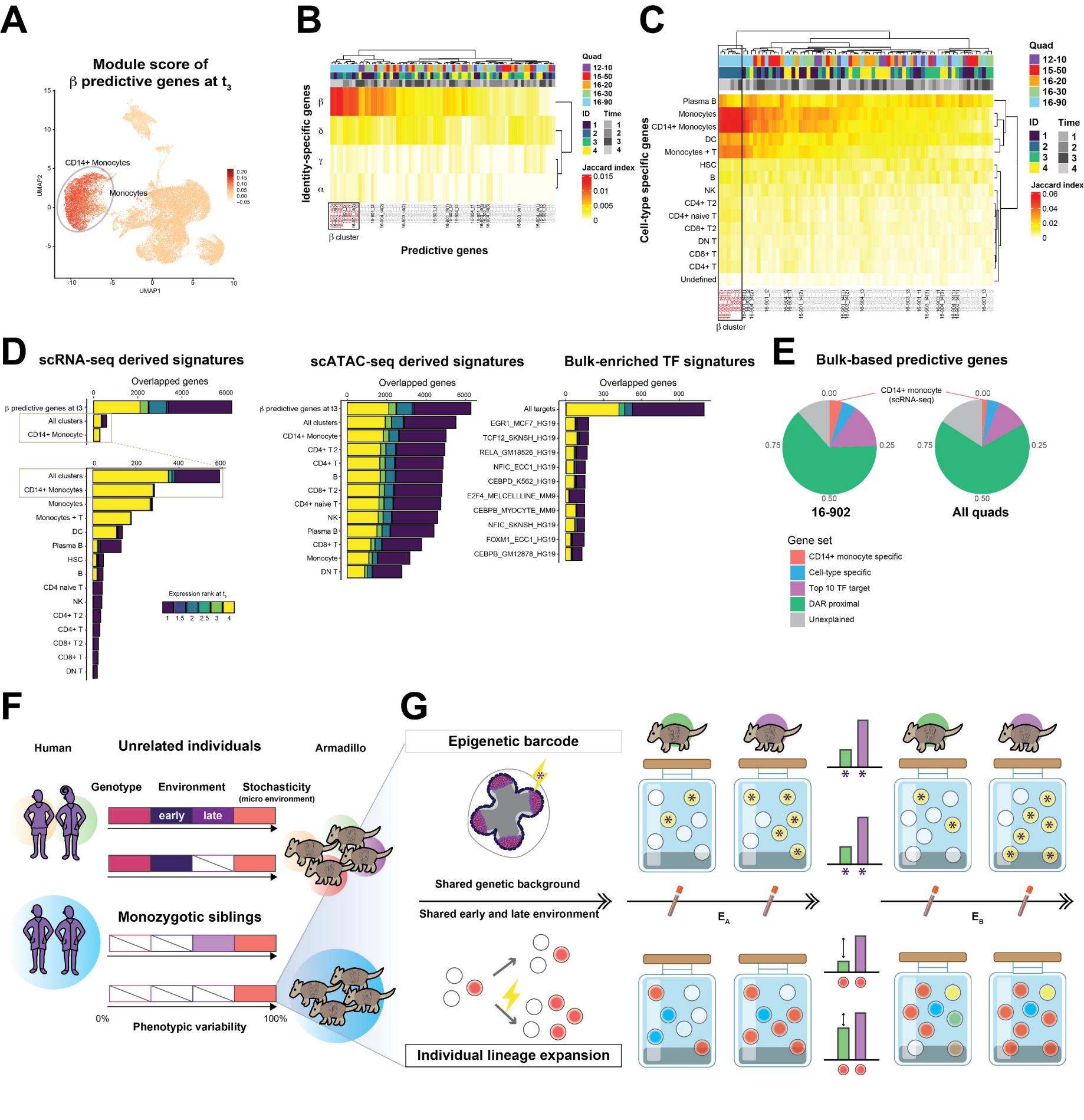
Supplementary_data
Supplementary_table 1
Supplementary_table 2
Supplementary_table 3
Supplementary_table 4


A
Polyembryony of nine-banded armadillo
Embryonic reprogramming
Delayed implantation
Twinning
Quadrupling
Totipotency
Maternal transcripts
Zygotic transcripts
Epigenetic and transcriptomic changes during development
Quadruplets from wild-caught pregnant females under genetic and environmental control
12–10 ♀
15–50 ♂
16–20 ♀
16–30 ♀
16–90 ♂
Compare blood transcriptome and phenotypes
Quadruplet 16-90
Coefficient of variation (CV) of blood test results
Highest: WBC, ALT, LY, MO, LY#, MO#, GR#, RBC, Hgb, Hct, MCH, MCHC, Plt
Lowest: GR, MCV
B
Weight (lbs)
Quadruplet
C
Coefficient of variation
Quadruplet
D
White blood cell (WBC)
Quadruplet
E
Coefficient of variation
Quadruplet
F
Sampling time points
t1
t2
t3
t4
Leprosy infected
2017
2018
2019
2020
Ballouz S, et al. 2023.
Epigenetic signatures by allele-specific expression
DNA-seq + RNA-seq
This study
Transcriptomic signatures + cell-type specific signatures
Blood extraction
scRNA-seq
scATAC-seq
16–90
Cell-type specific transcriptome
Cell-type specific epigenome
G
Time
ID
Similarity
Spearman's rho
16–30
16–20
12–10
15–50
16–90
Samples
Transcriptome of samples
H
Inter-sibling 0.982
Intra-individual 0.979
Cross-quad 0.971
Density
Transcriptional similarity across 3 time points
I
AUROC ~ 0.90
Density
Highly variable genes (HVGs) cross-quad predictability


A
t1
t2
Time matched
Expression
Batch, age, and environmental differences
Rank
ID matched
12−10
15−50
16−20
16−30
16−90
B
Expression
Rank
Quad
12-10
15-50
16-20
16-30
16-90
ARI
ID
Time
C
Learning canalized gene expression
Time point 1 (t1)
Time point 2 (t2)
Time point 3 (t3)
Training
Select genes perfectly correlated across time
Gene A
Gene B
Expression t1
Testing
Can we predict identity with these genes?
Predictive
Inconsistent
Expression t1+t2
Gene Set
Consistent
Feature set
2 time points
Perfect predictor
3 time points
Predictive genes
3 time points
Feature modules
α
β
γ
δ
Expression level
Feature genes
Ranked expression
Feature modules
D
Feature set
p ~ 0.003
Cummulative Density
Predictability score
E
Feature set
p ~ 0.014
All genes
Predictability score
Gene sets
Feature set
All genes
Gene set size
F
p ~7e-3
p <1e-3
p <1e-3
Density
Quad
12−10
15−50
16−20
16−30
16−90
Number of perfect predictors
G
Overlap of predictive genes
12−10
15−50
16−20
16−30
16−90
Within-ind
All
1 time
2 times
3 times
H
Quad
ID
Time
Rank
Predictive gene at t1
Predictive gene at t2
Predictive gene at t3
Rank order of predictive genes for 16-90 β at t3


A
~80-190 genes
Average predictability
Average score within data (1.8)
Number of predictive genes added to model
B
~148 excess genes
Density
Number of predictive genes
C
Predictability score
|log2 FCs|
Number of genes (log 10)
E
Mean AUROC 0.64
Quad
12−10
15−50
16−20
16−30
16−90
Density
Quad module predictability (AUROCs)
D
Functional assessment
Co-expression by transcriptome data
Expression Gene B
Expression Gene A
Expression Gene C
Expression Gene A
Gene B
Gene A
Gene C
Co-expression between all gene pairs
Co-expression network across the cohorts
Feature modules cross-quadruplet predictability, enrichment of GO, other networks
Feature module assessments
Module 1
Module 2
Module 3
Genes
Perfect predictor
Predictive gene
Co-exp
Functions
F
Predictability score
Cardiac muscle development
Sex differentiation
12−10
16−90
16−20
16−30
15−50
GO:0048738 cardiac muscle tissue development
GO:0003205 cardiocyte differentiation
GO:0060021 palate development
GO:0090263 phosphatase complex
GO:0008097 protein serine/threonine phosphatase complex
GO:0033032 autophagy of mitochondrion
GO:0000045 autophagic vacuole organization
GO:0098023 vesicle tethering complex
GO:0055007 striated muscle cell development
GO:0055013 cardiac muscle cell development
GO:0044275 cellular carbohydrate catabolic process
GO:0031674 I band
GO:0030018 Z disc
GO:0030017 sarcomere
GO:0030016 myofibril
GO:0005198 structural constituent of muscle
GO:0008406 gonad development
GO:0048537 gonad development
GO:0007548 sex differentiation
GO:0007894 tissue homeostasis
GO:0007044 cell-substrate junction assembly
GO:0030509 calcium ion import
GO:0030684 preribosome
G
16−90
α
β
γ
δ
Mean gene set average predictability
Gene set size
H
GO enrichment for β t3 predictive genes
mitotic cell cycle process
positive regulation of cytokine production
regulation of cell activation
regulation of cytokine production
leukocyte activation
regulation of leukocyte cell−cell adhesion
regulation of T cell activation
positive regulation of cell activation
chromosome segregation
regulation of lymphocyte activation
191
90
98
128
83
67
63
67
80
77
p.adjust
Count
I
ENCODE ChIP-seq for β t3 predictive genes
EGR1_MCF7_HG19
TCF12_SKNSH_HG19
RELA_GM12878_HG19
NFIC_ECC1_HG19
CEBPD_K562_HG19
E2F4_MELCELLLINE_MM9
CEBPB_MYOCYTE_MM9
NFIC_SKNSH_HG19
FOXM1_ECC1_HG19
CEBPB_GM12878_HG19
p.adjust
Count
GeneRatio
J
PU.1
GATA2/ZEB2
KLF4/IRF8
CMP
GMP
cMoP
Ly6c^Hi monocyte
Ly6c^Lo Monocyte
Ly6c^Hi proinflammatory monocyte
Macrophage Mo-DC
Ly6c^Low patrolling monocyte
C/EBPβ
NR4A1
KLF2
EGR1
Gatekeeper of inflammatory enhancers
NF-κB (RelA)
Inflammatory response genes
KEAP1/NRF2
Regulating inflammatory functions
C/EBPδ
NFIC
Mouse
Human
Classical
Ly6c^HiCD43^Lo
CD14^++CD16^- monocyte
Intermediate
Ly6c^HiCD43^Hi
CD14^++CD16^+ monocyte
Nonclassical
Ly6c^LoCD43^Hi
CD14^+CD16^++ monocyte


A
scRNA-seq
ID
α
β
γ
δ
UMAP2
UMAP1
B
Dual marker-based scRNA-seq cell-type annotation
CellMarker set
scRNA-seq
Cell type
B
CD14+ Monocytes
CD4+ naive T
CD4+ T
CD4+ T 2
CD8+ T
CD8+ T 2
DC
DN T
HSC
Monocytes
Monocytes + T
NK
Plasma B
Undefined
C
scRNA-seq
scATAC-seq
Expression
Anchoring
Integration
Cell
LSI
CC1
CC2
D
scATAC-seq
E
scRNA-seq
CD14+ Monocytes
Monocytes
F
scATAC-seq
G
Mean gene set average predictability
Gene set size
H
Cell population
Cell type label
Normalized cell population
J
Known motif detection
Rank
Motif
TF
P-value
SpiB(ETS) - GSE56857
PU.1(ETS) - GSE21512
Elf4(ETS) - GSE88699
Elf4(ETS) - GSE30407
Elf3(ETS) - GSE64557
de novo motif detection
q-value/ FDR
% of Targets
Best Match/Details
PB0058.1_Sfpi1_1/Jaspar
CEBPD/MA0836.2/Jaspar
I
Zscore of cell population
Monocyte
DC
T cell
B cell
NK
HSC


A
Module score of
β predictive genes at t3
CD14+ Monocytes
Monocytes
UMAP2
UMAP1
B
Identity-specific genes
β
δ
γ
α
β cluster
Predictive genes
Quad
12-10
15-50
16-20
16-30
16-90
ID
Time
Jaccard index
C
Cell-type specific genes
Plasma B
Monocytes
CD14+ Monocytes
DC
Monocytes + T
HSC
B
NK
CD4+ T2
CD4+ naive T
CD8+ T2
DN T
CD8+ T
CD4+ T
Undefined
β cluster
D
scRNA-seq derived signatures
Overlapped genes
β predictive genes at t3
All clusters
CD14+ Monocyte
Monocytes
Monocytes + T
DC
Plasma B
HSC
B
CD4 naive T
NK
CD4+ T2
CD4+ T
CD8+ T2
CD8+ T
DN T
Expression rank at t3
scATAC-seq derived signatures
Overlapped genes
β predictive genes at t3
All clusters
CD14+ Monocyte
CD4+ T2
CD4+ T
B
CD8+ T2
CD4+ naive T
NK
Plasma B
CD8+ T
Monocyte
DN T
Bulk-enriched TF signatures
Overlapped genes
All targets
EGR1_MCF7_HG19
TCF12_SKNSH_HG19
RELA_GM18526_HG19
NFIC_ECC1_HG19
CEBPD_K562_HG19
E2F4_MELCELLLINE_MM9
CEBPB_MYOCYTE_MM9
NFIC_SKNSH_HG19
FOXM1_ECC1_HG19
CEBPB_GM12878_HG19
E
Bulk-based predictive genes
CD14+ monocyte
(scRNA-seq)
16-902
All quads
Gene set
CD14+ monocyte specific
Cell-type specific
Top 10 TF target
DAR proximal
Unexplained
F
Human
Unrelated individuals
Armadillo
Genotype
Environment
Stochasticity
(micro environment)
early
late
Monozygotic siblings
0%
100%
Phenotypic variability
G
Epigenetic barcode
Shared genetic background
Shared early and late environment
Individual lineage expansion
E_A
E_B