CellSexID: Sex-Based Computational Tracking of Cellular Origins in Chimeric Models

Huilin Tai [1,2,3,†], Qian Li[1,4,5,†], Jingtao Wang[1,4,†], Jiahui Tan[1,4], Ryann Lang[1,4], Basil J. Petrof[1,4,*], Jun Ding[1,4,6,7,8,*]

[1] Meakins-Christie Laboratories, Research Institute of McGill University Health Centre, 1001 Decarie Blvd, Montreal, H4A 3J1, Quebec, Canada

[2] Department of Mathematics and Statistics, McGill University, 845 Sherbrooke St W, Montreal, H3A 0G4, Quebec, Canada

[3] Department of Computer Science, Columbia University, New York, NY, USA

[4] Department of Medicine, Division of Experimental Medicine, McGill University, 1001 Decarie Blvd, Montreal, H4A 3J1, Quebec, Canada

[5] Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences & Peking Union Medical College, Suzhou, 215123, Jiangsu, China.

[6] Quantitative Life Sciences, McGill University, 845 Rue Sherbrooke Ouest, Montreal, H3A 0G4, Quebec, Canada

[7] School of Computer Science, McGill University, 3480 Rue University, Montreal, H3A 2A7, Quebec, Canada

[8] Mila-Quebec AI Institute, 6666 Rue Saint-Urbain, Montreal, H2S 3H1, Quebec, Canada

[†]These authors contributed equally to this work.

[*]Corresponding authors: Basil J. Petrof (petrof.basil_j@mcgill.ca) and Jun Ding (jun.ding@mcgill.ca).

Abstract

Cell tracking in chimeric models is essential yet challenging, particularly in developmental biology, regenerative medicine, and transplantation studies. Existing methods, such as fluorescent labeling and genetic barcoding, are technically demanding, costly, and often impractical for dynamic, heterogeneous tissues. To address these limitations, we propose a computational framework that leverages sex as a surrogate marker for cell tracking. Our approach uses a machine learning model trained on single-cell transcriptomic data to predict cell sex with high accuracy, enabling clear distinction between donor (male) and recipient (female) cells in sex-mismatched chimeric models. The model identifies specific genes critical for sex prediction and has been validated using public datasets and experimental flow sorting, confirming the biological relevance of the identified cell populations. Applied to skeletal muscle macrophages, our method revealed distinct transcriptional profiles associated with cellular origins. This pipeline offers a robust, cost-effective solution for cell tracking in chimeric models, advancing research in regenerative medicine and immunology by providing precise insights into cellular origins and therapeutic outcomes.

Keywords: Chimera models, Single-cell, Sex Prediction, Cell origin tracking, Machine learning, Sex biomarkers

INTRODUCTION

Chimeric models, which contain cell populations derived from genetically distinct origins, are critical tools for understanding how a cell's developmental origin—or cellular ontogeny—influences its function within adult tissues. By allowing researchers to trace cell lineages throughout an organism's lifespan, these models enable valuable insights into processes like development, aging, and disease progression. For example, studies of macrophages—immune cells that arise from both prenatal progenitors and postnatal hematopoietic stem cells (HSCs)—have revealed distinct functional roles based on their developmental origins bone marrow [1]. Prenatal-origin macrophages are seeded into tissues early in development and can self-renew, while HSC-derived macrophages are continuously replenished by blood-derived monocytes originating from the bone marrow [2]. Despite macrophages' prevalence and functional diversity across all tissues, the specific contributions of these distinct populations of different origins to tissue maintenance, immune responses, and disease remain poorly understood.

Accurate tracking of cell origin in chimeric models is essential for dissecting the functional roles of distinct cell populations, especially at single-cell resolution. Mouse chimeras generated through bone marrow transplantation into recipient hosts of different origin are a frequently employed and powerful tool in biological experimentation. However, existing methods for distinguishing the donor from recipient cells face significant restraints. Traditional approaches rely on tracking genetically engineered artificial markers, such as distinct CD45 alleles or fluorescent reporters. Unfortunately, these markers are often not readily available in the mouse models under study, leading to the need for labor-intensive breeding programs and preparation steps [3, 4]. Another strategy which does not require genetic engineering employs sex-mismatched chimeras, using biological sex as a surrogate marker to determine cell origin by detecting either sex chromosomes [5, 6] or their associated gene products [7]. Techniques like fluorescence in situ hybridization (FISH) can detect Y chromosomes in individual cells, providing single-cell resolution, but are only semi-quantitative and incompatible with high-throughput single-cell transcriptomic analysis [5, 6]. On the other hand, PCR-based methods to detect sex-related genes allow for bulk-level quantification of chimerism but lack the single-cell resolution necessary for individual cell origin identification [7]. The inherent limitations of these current techniques hinder our ability to fully characterize and quantify ontogenically distinct cell populations within tissues at the single-cell level, and consequently restrict advances in fields such as immunology, oncology, developmental biology, and regenerative medicine.

To overcome these limitations, we introduce **CellSexID**, an innovative machine learning framework designed to leverage sex as a surrogate marker to identify cellular origin in sex-mismatched chimeric models at single-cell resolution. One of the core innovations of CellSexID is its ensemble feature selection scheme, which rigorously identifies a minimal yet informative set of 12 marker genes that optimize classification performance. By training on this carefully selected gene panel, CellSexID achieved robust classification accuracy (AUPRC > 0.95), effectively distinguishing hematopoietic stem cell (HSC)-derived donor (male) from non-HSC-derived host (female) macrophages. Our findings show that among the 9 macrophage clusters identified in healthy adult skeletal muscle by single-cell RNA sequencing (scRNAseq) [8], only one was almost exclusively associated with a non-HSC origin, and exhibited a unique anti-inflammatory gene expression profile. Our findings show that among the 9 macrophage clusters identified in healthy adult skeletal muscle by single-cell RNA sequencing

(scRNAseq), only one was almost exclusively associated with a non-HSC origin, and exhibited a unique anti-inflammatory gene expression profile.

Our findings underscore CellSexID's ability to uncover ontogeny-specific gene signatures and functional characteristics. This scalable and efficient approach enables high-resolution cell origin tracking across diverse biological contexts. Beyond macrophages, CellSexID holds significant potential for advancing studies of ontogeny-related functions in various cell types and tissues, including applications in chronic inflammatory diseases, cancer, and stem cell therapies, where accurate cell tracking could yield critical insights into disease mechanisms [9]. By enabling precise, single-cell analysis of cellular origin, CellSexID provides a versatile tool for developmental biology, immunology, and regenerative medicine, offering new avenues for understanding cellular function, lineage-specific responses, and disease progression in complex tissues.

RESULTS

Method overview

CellSexID, as illustrated in Figure 1, identifies sex-specific gene features using a committee of four machine learning classifiers: XGBoost [10-12], Support Vector Machine [13-15], Random Forest [16-20], and Logistic Regression [21-23]. The committee determines important features based on classifier consensus (Fig. 1a). Using publicly available mouse adrenal gland scRNAseq datasets containing cells from both sexes [24], the committee identified 12 critical gene features for sex prediction. Cross-validation demonstrated near-perfect prediction performance across all classifiers, validating the robustness of the selected features and confirming the feasibility of using single-cell gene expression for sex prediction with machine learning.

To evaluate the real-world applicability of CellSexID, we created sex-mismatched chimeric mice through bone marrow transplantation experiments and generated a scRNAseq dataset in macrophages isolated from the diaphragm muscle (Fig. 1b). This de novo validation dataset was used to assess the predictive performance of the models trained on the above public datasets in the determination of cell sex origin. Flow cytometry was performed on this validation dataset to establish ground truth labels, ensuring rigorous evaluation of model predictions. Models trained on the public datasets achieved high accuracy in distinguishing donor (male) from recipient (female) cells when applied to the validation dataset. Standard performance metrics confirmed the generalizability of CellSexID across datasets (Fig. 1c).

Finally, to demonstrate its broad utility in biological research, we applied CellSexID to annotate diaphragm macrophages in chimeric mice from a separate experimental dataset (Fig. 1d). Differential gene expression (DEG) and pathway enrichment analyses were performed on donor and recipient macrophages, revealing distinct transcriptional profiles and pathways. These findings illustrated the tool's capacity to accurately track cell origins and facilitate the investigation of transplanted cells in complex biological systems. This controlled chimeric model also highlights how tracking donor and recipient cells provides insights into cell-specific behaviors and disease-associated pathways. Such capabilities can support the development of targeted therapies and enhance our understanding of molecular mechanisms underlying pathogenesis.

CellSexID identifies a minimal and effective 12-gene marker set for sex prediction through ensemble feature selection

Use of a limited number of key gene features for sex-based cell tracking decreases both the complexity

and cost of future experimental applications. Therefore, to enhance the real-world applicability of our sex-based cell origin model, we focused on identifying a minimal gene set capable of ensuring accurate sex determination. As shown in Supplementary Fig. 1, our feature selection strategy achieves optimal prediction performance compared to models without feature selection or those with varying numbers of top features. Additionally, optimizing feature selection can improve the computational efficiency and accuracy of our model. To identify the most predictive gene features, we employed an ensemble feature selection approach [25-28], combining four machine learning algorithms: XGBoost, Support Vector Machine, Random Forest, and Logistic Regression. Supplementary Fig. 2 demonstrates that this ensemble approach outperforms individual models, resulting in superior prediction accuracy.

Our ensemble approach was then trained and tested on a publicly available scRNAseq datasets from the Gene Expression Omnibus (GEO), featuring adrenal gland cells from female (GSM6153751; 3,963 cells) and male (GSM6153750; 2,685 cells) mice [24]. Figures 2a-2d display the top 20 genes with the highest feature importance scores across each model. Certain genes consistently ranked as top predictors across different models. By analyzing the committee votes of each gene, we identified a robust set of 12 genes selected by at least 3 of the 4 models (Fig. 2e; see Supplementary Fig. 3 for justification of this threshold). These 12 genes (Fig. 2f) were used in all subsequent analyses due to their consistently high performance level across multiple algorithms.

It is notable that among the 12 selected genes, only 5 are found on sex chromosomes whereas the rest are autosomal. The sex chromosome-linked genes have a combination of X chromosome (*Xist*) and Y chromosome (*Ddx3y*, *Eif2s3y*, *Uty, Kdm5d*) locations. *Xist* plays a central role in X chromosome inactivation in females [29]. *Ddx3y* and *Eif2s3y* dictate male-specific functions such as spermatogenesis [30] [31]. *Uty* [32] and *Kdm5d* [33], located on Y chromosome, are also recognized as reliable sex markers, with their expression limited to males and their roles linked to histone modifications and chromatin organization [34]. Among the 7 autosomal genes identified, several are also known to exhibit sexual dimorphism in other cell types [35] [36]. To further validate our findings, we compared the performance of our 12 selected genes against features limited to only the Y chromosome or the two sex chromosomes together. As shown in Supplementary Fig. 4, the genes selected by our model, including those located on autosomal chromosomes, led to a superior predictive performance. This underscores the value of our end-to-end feature selection approach, powered by the machine learning model committee, in identifying robust sex markers beyond sex chromosomes. A more comprehensive list of identified features is available in Supplementary Table 1.

In summary, our ensemble feature selection approach not only identified a biologically relevant and minimal gene set but also demonstrates the power of combining multiple machine learning algorithms for robust and reproducible feature selection. This consensus-based selection method was able to restrict the number of predictors to an easily manageable level without sacrificing performance.

Cross-validation shows accurate sex prediction using the minimal 12-gene set on public single-cell data

To further test whether our feature selection approach is a practical and effective tool for various experimental applications, we next cross-validated its performance on the public adrenal gland single-cell datasets from Dolfi et al. (GSM6153751 for female mice and GSM6153750 for male mice) [24]. In these datasets, hematopoietic cells (CD45+) from adrenal glands of 7-week-old male and female C57Bl/6 mice were isolated by fluorescence-activated cell sorting (FACS)[37] and scRNAseq was

performed using the 10x Genomics platform. The sex origin of the cells was definitively known based on the biological sex of the source animals under controlled laboratory conditions. Using our identified 12 most predictive gene features, we evaluated the performance of our machine learning models based on these selected features using XGBoost, Support Vector Machine, Random Forest, and Logistic Regression.

Using these curated datasets, where cell sex origin is reliably known based on the source animal, the models were evaluated through 5-fold cross-validation, with the selected 12 gene features as inputs. As shown in Figure 3, all models demonstrated excellent predictive performance across multiple metrics with those 12 selected features. Specifically, they achieved high accuracy (Fig. 3a), F1 scores (Fig. 3b), area under the receiver operating characteristic curve (AUC, Fig. 3c), and area under the precision-recall curve (AUPRC, Fig. 3d), confirming the effectiveness of the 12 selected features for sex prediction. Notably, XGBoost emerged as the top performer across all metrics, showcasing its particularly high classification accuracy.

Validation of CellSexID for accurate cell origin tracking using cell sorting

After demonstrating the strong sex prediction performance of CellSexID on public datasets, we next evaluated its applicability for tracking cell origin in a real-world setting by performing "de novo validation" in sex-mismatched chimeric mice. The chimeric mice were created by transplanting bone marrow from male donor mice (CD45.1 allele) into female (CD45.2 allele) host recipient mice (Fig 4a). The resident macrophages present in the diaphragm muscle of female recipient mice were protected from the pre-transplant irradiation protocol through the use of lead shielding as we have recently described [38]. At 8 weeks post-transplantation, macrophages were sorted and pooled from the diaphragms of two chimeric mice by FACS based on their CD45 allele status (CD45.1 or CD45.2) and then labelled with Cell Multiplexing Oligo (CMO) barcodes. Using our scRNAseq analysis pipeline, these distinct CMO barcodes established the ground truth for sex-based macrophage origin, providing a robust benchmark for experimentally evaluating CellSexID's annotation performance in this "validation dataset".

The results of this validation, summarized in Figure 4, demonstrate the high accuracy of CellSexID in classifying cell origin. Figure 4b shows the confusion matrix for the best-performing model, XGBoost, highlighting its precision and recall by detailing true positive, true negative, false positive, and false negative rates. Figure 4c presents the ROC curves for the four machine learning algorithms, illustrating each model's ability to distinguish between male and female cells, with the area under the curve (AUC) as a primary metric of performance. Figure 4d shows the AUPRC score curves for the same models, offering a balanced perspective on precision and recall by evaluating the trade-off between false positives and false negatives. To further assess model stability, Figures 4e and 4f present bootstrapped AUROC performance across multiple samples for each model. These bar plots confirm the robustness of the AUROC scores, with XGBoost demonstrating particularly consistent performance. Finally, Figure 4g compares the predicted male-to-female ratio with the actual true ratio based on the validation dataset, underscoring the accuracy of CellSexID in estimating population composition in a real-world setting. This comprehensive evaluation of CellSexID with its minimal 12-gene set in both publicly available and de novo validation data, supports its reliability and robustness for accurate cell origin tracking in multiple scRNA-seq dataset scenarios.

CellSexID provides novel biological Insights into macrophage ontogeny

The steady-state macrophage pool in adult skeletal has two principal origins: 1) a prenatal population established during muscle development from embryonic progenitor cells; and 2) a postnatal population derived from definitive monocytes recruited to the muscles [1]. The former macrophage population undergoes self-renewal in the tissue and is bone marrow-independent, whereas the latter population turns over rapidly and requires continuous replenishment by monocytes from the bone marrow. Accordingly, we next sought to determine whether CellSexID could ascertain the relative contributions of these two populations to macrophage subsets and their associated transcriptomic profiles in the main respiratory muscle, the diaphragm – where the role of macrophage ontogeny remains underexplored. To address this question, we generated additional chimeric mice in which CD45.1 male bone marrow was transplanted into CD45.2 female recipients, generating another "experimental dataset" (Fig.5a). Using our pre-trained XGBoost model to predict cell origins from scRNAseq analysis of sorted diaphragm macrophages at 8 weeks post-transplantation, the predicted percentages of donor (bone marrow-dependent) and recipient (bone marrow-independent) macrophages were 66% and 34%, respectively. This closely matched the percentages obtained from flow cytometry staining of CD45.1+ and CD45.2+ cells (70% and 30%, respectively) (Fig. 5b). This confirms that CellSexID can be used to determine cellular origin in chimeric mice without the need for conventional tracking methods such as the CD45.1 allele or fluorescent marker genes, which are not readily available in most genetically modified mouse model strains.

The scRNAseq UMAP plot revealed 9 distinct macrophage clusters in the diaphragm (Fig. 5c, Supplementary Fig. 5). The sex origin of each macrophage cluster in the muscle was determined using the XGBoost model, and there was a very high degree of overlap between male origin and expression of *Ccr2* (Fig. 5d), which is a marker for bone marrow-derived macrophages[39, 40]. All of the clusters were predominantly derived from donor (male) bone marrow monocytes with the exception of cluster C4, in which 93% of the macrophages were of recipient (female) origin (Fig. 5e). The donor bone marrow-dependent (male) macrophages displayed a dispersed and heterogeneous clustering pattern, consistent with a diverse population having varying gene expression profiles and functional states. In contrast, the bone marrow-independent (female) macrophages formed a tight, homogeneous cluster, indicating greater uniformity in their gene expression and most likely their function.

To gain further insights into the differential functions of bone marrow-dependent versus bone marrow-independent macrophages, we performed differential gene expression (DEG) analysis. In macrophage cluster C4, which is primarily composed of bone marrow-independent macrophages, 522 genes were significantly upregulated, and 266 genes were downregulated compared to all other clusters (Fig. 5f). The top 10 upregulated genes included *Cfh, Colec12, Timd4, Mir99ahg, Selenop, C4b, Cpne8, Tanc2*, *Stab1*, and *Mctp1*, while the top 10 downregulated genes were *Cd74, H2-Eb1, H2-Ab1, H2-Aa, Ccr2, H2-DMa, Cd52, H2-DMb1, Mpeg1*, and *Tmsb10*. Pathway enrichment analysis revealed that the upregulated genes in cluster C4 were significantly associated with pathways such as "Endocytosis," "Adherens Junction," "Ras Signaling Pathway," and "Rap1 Signaling Pathway" (Fig. 5g). These pathways suggest enhanced cellular uptake processes, cell adhesion, and growth factor signaling, likely contributing to increased cell communication and proliferation. In contrast, the downregulated pathways included "Hematopoietic Cell Lineage," "TNF Signaling Pathway," and "NF-kappa B Signaling Pathway," indicating reduced pro-inflammatory immune responses in this bone marrow-independent macrophage population.

To determine whether similar differences could be observed within a macrophage cluster containing

cells from both origins, we applied our method, CellSexID, to macrophage cluster C3, which contains a mixture of donor-origin (63%) and recipient-origin (37%) macrophages. By enabling the identification of cellular origins within the same cluster, CellSexID allowed us to perform differential expression analysis between donor-derived (bone marrow-dependent) and recipient-derived (bone marrow-independent) macrophages. Without CellSexID, such comparisons would not have been possible, as the mixed origins of cells within the cluster obscure their distinctions. Using CellSexID, we identified 81 upregulated and 46 downregulated genes in recipient-origin macrophages compared to donor-origin macrophages (Fig. 5h). Notably, pathway enrichment analysis revealed that the downregulated pathways in recipient-origin macrophages mirrored those observed in cluster C4, including pathways associated with antigen presentation, T-cell activation, and other pro-inflammatory immune responses (Fig. 5i). These findings reinforce the distinction between bone marrow-dependent and bone marrow-independent macrophages and highlight the utility of CellSexID in enabling origin-specific analyses within mixed populations.

DISCUSSION

This study addressed the critical challenge of tracking cell origins in chimeric models by developing CellSexID, a machine learning-based tool that uses sex-specific gene expression in scRNA-seq data as a natural surrogate marker. By applying CellSexID to sex-mismatched chimeric models we have created a simple, low cost, and high-precision method that accurately distinguishes donor and recipient cells at single-cell resolution. Through careful feature selection, we identified 12 essential genes that include but are not limited to those located on sex chromosomes, which allow for robust sex prediction and thereby enable reliable cell origin tracking. CellSexID simplifies the cell tracking process by eliminating the need for biased labelling techniques dependent on exogenous marker genes, which commonly have variable levels of gene expression under different experimental conditions and can also lack cellular specificity (so-called “leaky expression”) [41, 42]. In addition, expensive and labor-intensive cross-breeding strategies are often required to incorporate exogenous cell origin markers into existing animal models [3, 4]. By taking advantage of the endogenous gene expression profile associated with biological sex, CellSexID is well positioned as an efficient, scalable solution in experimental as well as clinical settings.

The primary innovation of this study lies in CellSexID’s consensus-based feature selection approach, which leverages four machine learning models — XGBoost, Support Vector Machine, Random Forest, and Logistic Regression — to select a minimal yet highly informative set of genes for sex-based cell origin tracking. Unlike traditional cell tracking methods, CellSexID achieves high classification accuracy with a limited gene panel, enabling efficient tracking in complex tissue environments. Interestingly, our feature selection approach revealed that the most relevant markers for sex determination are not limited to sex chromosomes, as commonly assumed. While genes like *Xist, Ddx3y*, *Eif2s3y*, *Uty,* and *Kdm5d* found on the sex chromosomes were expectedly selected, a majority of the top features were located on autosomes. This challenges the intuitive belief that genes on sex chromosomes (particularly the Y chromosome) are inherently more predictive but it is consistent with recent literature. For example, *Uba52* encodes a ubiquitin-ribosomal fusion protein that has been reported to play an important role in sex-related behavioral differences mediated by the amygdala [35]. The *Lars2* gene product, leucine-tRNA-synthase 2, was recently found to be highly enriched in endothelial cells of male mice [36]. Several ribosomal proteins of the large (Rpl) and small (Rps) subunits, the two major components of the ribosome which are critical for ribosomal assembly and

function, were also identified. These findings, validated through comparisons to models restricted to sex chromosome features, highlight the power of CellSexID's end-to-end machine learning approach to uncover robust sex markers beyond traditional assumptions. By including autosomal markers, we achieved higher predictive performance metrics, such as AUROC and AUPRC, demonstrating the strength of this integrative feature selection framework.

To demonstrate CellSexID's real-world applicability, we validated its performance in chimeric mice with sex-mismatched bone marrow transplantation. In tracking the sex origin of macrophages in skeletal muscle using our tool, we were able to show high concordance with ground truth flow cytometry results, confirming its accuracy in distinguishing bone marrow-derived from bone marrow-independent cells. Furthermore, here we show the presence of important transcriptional differences in skeletal muscle macrophages based on cell origin (ie., between bone marrow HSC-derived and embryonic-derived macrophages), thus revealing how ontogeny helps to shape immune responses. Notably, we identified a unique subset of macrophages in the diaphragm which are overwhelming bone marrow-independent and display an anti-inflammatory and pro-regenerative gene expression profile. This discovery highlights CellSexID's capacity to reveal ontogeny-driven functional differences within tissues where traditional methods have struggled to provide sufficient resolution. The characteristics of embryonic-derived tissue-resident macrophages are particularly relevant to various pathological states in which these cells are often lost from the diseased tissue [2, 43], and CellSexID provides a powerful new tool for tracking their behavior at different stages of pathology.

A limitation of CellSexID is its reliance on sex differences between donor and recipient animals, which could potentially influence cell states or functions in certain contexts. While this dependency does not affect the main goals of this study, it may introduce bias in specific cell types where sex differences play an important role in determining phenotype. To address this limitation, future studies could implement bi-directional chimeric experiments with both male and female donors and recipients. This approach not only strengthens the robustness of our method but could also promote a greater understanding of the more nuanced aspects of sex differences in biological systems.

CellSexID has the potential to transform research by providing a precise, scalable, and cost-effective tool for tracking cell origins at single-cell resolution. By leveraging endogenous sex-specific gene expression, it eliminates the need for exogenous labeling or genetic engineering, overcoming limitations such as high costs and technical complexity in traditional methods. This innovation is broadly applicable across diverse research settings and addresses a critical gap in experimental biology. By enabling high-resolution tracking of cellular dynamics, CellSexID empowers researchers to explore ontogeny-driven differences in cell function across contexts such as immune regulation, tissue repair, and developmental biology. Its ability to identify functionally distinct cell populations, such as embryonic-derived macrophages with pro-regenerative properties, underscores its utility in uncovering fundamental biological insights. Furthermore, CellSexID democratizes cell-tracking methods, making advanced single-cell analyses more accessible and fostering innovation across disciplines. By providing a robust framework for understanding cellular dynamics, it holds great promise for advancing biomedical research and enabling the development of targeted therapeutic strategies.

METHODS

Experimental animals: Experiments utilized mice carrying CD45.2 and CD45.1 alleles, obtained from The Jackson Laboratories in Bar Harbor, ME, USA. The animals were housed under controlled

conditions in the vivarium of the McGill University Health Centre, following the standards of the Canadian Council on Animal Care, the McGill University Animal Ethics Committee, and ARRIVE guidelines. Before starting the study, the mice underwent a 7-day acclimation period in our facility. Skilled personnel provided high-quality care, including regular checks, to ensure the animals' well-being throughout the research.

Bone marrow transplantation: Chimeric mice were created following our previously outlined protocol [38]. In brief, 6-week-old female CD45.2 mice received whole-body X-ray irradiation using the X-RAD SmART irradiator (Precision X-ray, USA). Lead shielding was applied to the diaphragm area to protect resident macrophages. The irradiation was fractionated into two doses of 6 Gy, delivered 4 hours apart, with settings of 225kV, 13mA, and 1.0265 Gy/min. Six hours after the final irradiation, an intraperitoneal injection of 30 mg/kg busulfan (Sigma, USA) was administered to eliminate any remaining bone marrow protected by diaphragm shielding. Twenty-four hours post-irradiation, CD45.2 mice received an intravenous injection of $2x10^7$ bone marrow cells from age-matched CD45.1 male donors. To reduce infection risk, the recipient mice were provided with enrofloxacin in their drinking water for 7 days following irradiation. An 8-week recovery period was then allowed to ensure complete bone marrow reconstitution. Two female mice were used as recipients for the validation dataset and eight female mice were used as recipients for the experimental dataset.

Flow cytometry: To quantify and identify macrophages in diaphragm, flow cytometry was performed as described in prior studies [38, 44] . Mice were anesthetized using isoflurane before euthanasia, followed by heart perfusion with 20 ml of phosphate-buffered saline (PBS) (Wisent, Cat. #811-425), with an additional 20 ml administered after cutting the dorsal aorta. The diaphragm tissue was dissected, minced, and digested in PBS containing 0.2% collagenase B (Roche, Cat. #11088815001) for 1.5 hours to create a cell suspension. Diaphragm cells were stained with a viability dye to distinguish live from dead cells. For surface marker staining, cells were incubated at 4°C in the dark with primary antibodies, including anti-CD45.1 (clone A20, Cat. #553776), anti-CD45.2 (clone 104, Cat. #562129), anti-SiglecF (clone E50-2440, Cat. #562757), and anti-CD11c (clone HL3, Cat. #564986) (all from BD Biosciences, NJ, USA); and anti-CD11b (clone M1/70, Cat. #101226) and anti-F4/80 (clone BM8, Cat. #123114) (both from Biolegend, CA, USA). To prevent non-specific binding, cells were treated with Fc blocking solution (anti-CD16/CD32, clone Ab93, BD Biosciences, Cat. #567021). To identify diaphragm macrophages, we defined them as cells positive for CD45 (either CD45.1 or CD45.2), CD11b and F4/80, and negative for SiglecF and CD11c. For the model validation experiment, we specifically added Cell Multiplexing Oligo (CMO, 10X Genomics, cat. # 1000261) to sorted populations of CD45.1 and CD45.2 macrophages to differentiate between donor and recipient mouse macrophages. In experiments addressing our research question using the model, CMOs were added to the unstained muscle cell suspension before sorting to distinguish cells originating from different mice. All macrophages were sorted in a PBS solution containing 20% fetal bovine serum (FBS) to maintain cell viability and integrity throughout the process.

Validation dataset: For the validation experiment, we utilized a ratio of 80% Cell Multiplexing Oligo (CMO)-labelled CD45.1 macrophages to 20% CMO-labelled CD45.2 macrophages. This mixture was loaded onto a 10x Genomics Chip G (Cat. #1000127) and processed using the 10x Chromium Controller, following the manufacturer's instructions.

Experimental dataset: For the experiments aimed at addressing the research question, a mix of CD45.1

and CD45.2 positive cells was added to the chip without distinguishing between the two types. The library preparation was then conducted in accordance with 10x Genomics protocol CG000388.

All libraries were sequenced by Illumina NovaSeq6000 S4 v1.5.

Single-cell data acquisition: For our generated single-cell data (validation and experimental datasets), at 6 weeks post-transplantation, we used the same bone marrow transplantation protocol as described for generating chimeric mice. Wild-type (WT) female recipient mice were transplanted with bone marrow cells from WT male donors. Following the transplantation protocol, the mice were allowed to recover for 8 weeks to ensure complete bone marrow reconstitution. Subsequently, macrophages from the diaphragms of these mice were isolated via cell sorting, specifically targeting populations expressing CD45+ and F4/80+ markers to ensure high purity. Single-cell RNA sequencing was then performed on these isolated macrophages. Sequencing reads were processed using 10x Genomics Cell Ranger 7.1.0. Each library was processed using the Cell Ranger [45] multi pipeline. Specifically, we used the reference genome refdata-gex-mm10-2020-A with a minimum assignment confidence of 0.7. The libraries included gene expression and multiplexing capture data. The samples analyzed were WT_to_WT samples. Using Cell Ranger multi, each library was aligned to an indexed mm10 genome, generating an aggregated matrix of WT cells. This aggregation normalized the number of confidently mapped reads per cell in each library, preparing the data for downstream analysis.

Public single-cell data: The public single-cell datasets with sex labels for training the machine learning models are collected from the Gene Expression Omnibus (GEO) under the accession number GSM6153751 (female mice cells) and GSM6153750 (male mice cells). They provide a comprehensive single-cell RNA profile of AG CD45+ cells isolated from the adrenal glands of 7-week-old female (3,963 cells) and male (2,685 cells) C57Bl/6 mice. The two datasets are integrated and used for feature selection and training the classifier.

Single-cell data preprocessing: The SCANPY [46] (version 1.9.5) was applied to preprocess single-cell raw count matrices for both all three datasets. For the experimental dataset, we first performed quality control filtering genes not expressed in at least three cells. We filter cells expressing more than 6200 genes or fewer than 200 genes. Cells with more than 20% mitochondrial reads are regarded as dead cells and also filtered. Immunocytes were specifically filtered by selecting cells expressing Ptprc. Subsequently, the single-cell data matrix was normalized to a total count of 10,000 per cell and log1p-transformed. Then, for cell origin classification, only the 12 sex markers' expressed are used for model input. For clustering analysis, highly variable genes were detected using SCANPY's tool 'pp.highly_variable_genes' with default parameters, resulting in 1895 genes. The effects of total counts and mitochondrial gene percentage were regressed out. The data was then scaled, capping the values at 10.

For the validation dataset, we performed initial filtering, removing cells expressing fewer than 100 genes and genes expressed in fewer than 3 cells. The data was then normalized to a total count of 10,000 per cell and log1p-transformed. Then the 12 sex markers expression are used for sex prediction.

The public dataset was preprocessed by the dataset provider, where low-quality cells and genes were filtered out: genes expressed in fewer than 2 cells and cells with fewer than 200 non-zero gene counts were removed. Additionally, cells with a mitochondrial gene fraction above 1.2% were excluded to eliminate stressed or low-quality cells. Following these initial filtering steps, the dataset was normalized using the SCTransform method from the Seurat package, which performed library size normalization

and adjusted for sequencing depth. Building on this preprocessed dataset, we applied additional filtering to remove genes expressed in fewer than 3 cells and conducted a log1p transformation.

Classifier committee: We employed a classifier committee to predict sex based on gene expression data. The committee consists of four distinct machine learning classifiers, each contributing unique perspectives to the task. By working together, the committee helps to identify the most relevant features for maximizing prediction accuracy. While the best-performing classifier is ultimately chosen to make the final prediction on the test dataset, the combined efforts of the committee enhance the overall robustness and effectiveness of the feature selection process.

The choices of the models are based on their capabilities in feature selection and dealing with our classification tasks. **XGBoost Classifier** is a highly scalable model grounded in gradient boosting frameworks and is capable of capturing complex nonlinear patterns [47, 48]. It evaluates feature importance through a combination of metrics that describe how feature enhances accuracy and decision-making, which includes Gain – the improvement in accuracy a feature provides to the decision points it influences, Cover – the proportion of data points a feature impacts, and Frequency – how often a feature is utilized across the ensemble of decision trees. **Random Forest Classifier [49]** leverages an ensemble of decision trees and quantifies feature importance through the mean decrease in impurity, often expressed in terms of Gini impurity or entropy. This approach calculates the extent to which each gene reduces uncertainty across the trees, with a higher reduction indicating a stronger influence on model predictions. This method is particularly effective in discerning the genes that consistently refine the model's outputs by creating clearer, more defined paths in the decision process. For the **Support Vector Machines (SVM)** [50-52] with a linear kernel used in our method, feature importance is derived directly from the model's coefficients. These coefficients are integral to the hyperplane that separates the classes in the dataset. Each coefficient quantifies the influence of a corresponding gene on the classification boundary, with its magnitude indicating the strength of the impact and its sign denoting the direction (positive or negative). This linear approach provides a clear and quantifiable measure of each gene's contribution to the model's decision-making process, facilitating the identification of key genetic features crucial for predictive accuracy. Linear SVMs are preferred for their ability to perform well in high-dimensional spaces with clear margins of separation, and their simplicity and effectiveness in binary classification scenarios. Lastly, **Logistic Regression** [53, 54]determines the importance of genes through its coefficients within a linear context, where the magnitude of each coefficient indicates the extent to which a gene affects the model's output. Larger absolute values suggest a stronger influence, making this approach straightforward for gauging the impact of individual genes on the prediction of the target variable. Logistic Regression is favored for its interpretability and the provision of probability scores for outcomes, which can be crucial for understanding risk factors in medical or biological studies. These methods, grounded in robust statistical and mathematical principles, allow for a detailed exploration of gene significance in predictive modeling. In our implementation, Logistic Regression, Random Forest and Support Vector Machine are implemented in Python packages Scikit-learn [55] and XGBoost is implemented using package 'XGBoost' [10].

Sex-relevant feature selection: We used the public mouse adrenal gland dataset to select the most significant gene features for predicting the sex of individual cells. First, each model in the committee is trained on the public dataset using all gene features and employed the built-in feature selection functions provided by their respective libraries to compute feature importance scores. The top K (default 20, see justification of this hyperparameter in Supplementary Fig. 1) important features are selected. Then

features selected by at least C committee members are used as final selected features. We set C to be 3 out of 4 (75%) to claim a majority and Supplementary Fig. 3 justifies our choice. For any gene feature g, it will be selected in the final feature set $\boldsymbol{G_{final}}$ if and only if it is selected by at least three models.

$$g \in \boldsymbol{G}_{\mathbf{final}} \Leftrightarrow \sum_{model \in \{RF,SVM,XGB,LR\}} \mathbb{1}_{\text{model}}(g) \geq C \quad (1)$$

Where each $\mathbb{1}_{\text{model}}(g)$ represents an indicator function indicating if the gene feature is selected by a specific model (SVM: Support Vector Machine; LR: Logistic Regression; RF: Random Forest; XGB: XGBoost).

$$\begin{cases} \mathbb{1}_{\text{model}}(g) = 1, if\ g \in \arg\ max_{\boldsymbol{G}} \left(\sum_{i=1}^{K} |\boldsymbol{w}_{\text{model},\boldsymbol{G}_i}| \right) \\ \mathbb{1}_{\text{model}}(g) = 0, otherwise \end{cases} \quad (2)$$

Different $\boldsymbol{w}$ corresponds to different models' feature importance score returned by the Python library after training using all gene features on the public dataset. $|\boldsymbol{w}_{SVM}|$ represents the absolute value of the trained SVM's weights. $|\boldsymbol{w}_{LR}|$ denotes the absolute value of the trained Logistic Regression classifier's coefficients. $\boldsymbol{w}_{RF}$ is the importance of the trained Random Forest derived from the mean decrease in impurity across all trees in the forest. $\boldsymbol{w}_{XGB}$ is the importance of the score of trained XGBoost derived from the Gain, Cover, and Frequency as mentioned in the previous paragraph. The subscription $\boldsymbol{G}_i$ specifies the gene that the weight corresponds to.

The model committee also computes a scaled feature importance score for each gene feature, designed to reflect consensus among the committee members. This score is determined as follows: first, each committee member's score is normalized to a range between 0 and 1. The normalized scores are then summed across all committee members. To further refine the score, a bonus of 0.25 is added for each committee member that selects the feature, while a penalty of 0.25 is subtracted for each committee member that does not select the feature. As a result, the final bonus/penalty component of the score will range between -1 (if none of the four committee members select the feature) and +1 (if all four committee members select the feature), providing a balanced measure of agreement across the committee. All the gene features will be ranked based on their feature selection scores.

Once the feature selection process is completed, we use XGBoost as the final sex predictor. XGBoost is an ensemble learning method that builds an additive model using decision trees [56, 57], where each tree T_i represents a series of hierarchical splits on the selected feature set $\boldsymbol{G_{final}}$ to partition the data space for improved predictive performance. The prediction score z is computed as a weighted sum of the tree outputs: $z = \sum_{i=1}^{N} \theta_i T_i(\boldsymbol{G_{final}})$ Where the θ_i is the i-th tree's coefficient and N is the total number of decision trees. Then the output prediction score z is converted to prediction probability using the logistic function.

Sex classifier evaluation: Using the selected 12 final gene features, we tested the same models (Logistic Regression, Random Forest, Support Vector Machine, XGBoost) both on the public datasets and the validation dataset. For testing on the public dataset, we used 5-fold cross validation to compute Accuracy and F1 score. Standard deviations are shown as error bars. For calculating Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPRC) we used functions in scikit-learn. For testing the model on the validation dataset, the full public dataset was used for training the model. For generating the bar plots in Fig. 4 d and e, AUPRC

and AUROC are calculated using bootstrapping with 100 iterations. In each iteration, the public dataset is randomly resampled with replacement to get the same number of samples as the training set for the current iteration, but the test set is unchanged, ensuring a consistent basis for performance comparison. These metrics, along with their 95% confidence intervals derived from the bootstrap samples, provided comprehensive insights into the models' predictive capabilities. Based on these metrics, XGBoost has the highest performance and is selected as our final model. The final XGBoost classifier was subsequently used for annotating the cells in the single-cell dataset from our experiments.

UMAP and clustering analysis for the single-cell data: Principal Component Analysis (PCA)[58-60] was performed with 50 components. For the experimental dataset, Harmony [61] was applied to remove batch effects across 8 samples and adjust the PCA basis. The neighborhood graph of cells was computed using the PCA representation using the SCANPY's tool 'pp.neighbors' by considering 10 nearest neighbors and using the first 40 principal components, followed by UMAP [62] embedding using the 'tl.umap' tool with default parameters and Leiden clustering using SCANPY's 'tl.leiden' tool with a resolution of 0.6. After further filtering out clusters with fewer than 200 cells, 9 macrophage clusters were identified.

Enrichment analysis: Differentially expressed genes (DEGs) were identified using SCANPY's tool 'tl.rank_genes_groups' with the Wilcoxon rank-sum test method [63]. Sex-related genes were removed from the analysis. Up-regulated genes were selected with Benjamini-Hochberg adjusted p-values less than 0.05 and log2 fold changes greater than 0.6, while down-regulated genes were selected with Benjamini-Hochberg adjusted p-values less than 0.05 and log2 fold changes smaller than -0.6. Top regulated genes were visualized through volcano plots, generated using the custom scatter plot function implemented in Python tool Matplotlib. If the number of top regulated genes exceeded 200, the top 200 genes were used; if fewer than 200 top regulated genes were identified, all top genes were used. The top up- and down-regulated genes were then subjected to KEGG 2019 Mouse [64-67] enrichment analyses using the GSEApy [68] tool 'enrich' with KEGG for enrichment analysis.

DATA AVAILABILITY

Three datasets are used in this study. The public mouse adrenal gland dataset is from Dolfi et al.'s [24] research and can be downloaded from the Gene Expression Omnibus (GEO) repository under accession number GSM6153751 for female mice and GSM6153750 for male mice. All the single-cell sequencing data from our experiments will be made publicly available upon acceptance.

CODE AVAILABILITY

Code for the models and results reproduction is publicly available on GitHub: https://github.com/mcgilldinglab/CellSexID

ACKNOWLEDGEMENT

This work was funded by grants awarded to JD and BJP. We gratefully acknowledge the support from the Canadian Institutes of Health Research (CIHR) under Grant Nos. PJT-180505 and 180644; the Funds de recherche du Québec - Santé (FRQS) under Grant Nos. 295298 and 295299; the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. RGPIN2022-04399; and the Meakins-Christie Chair in Respiratory Research.

FIGURES

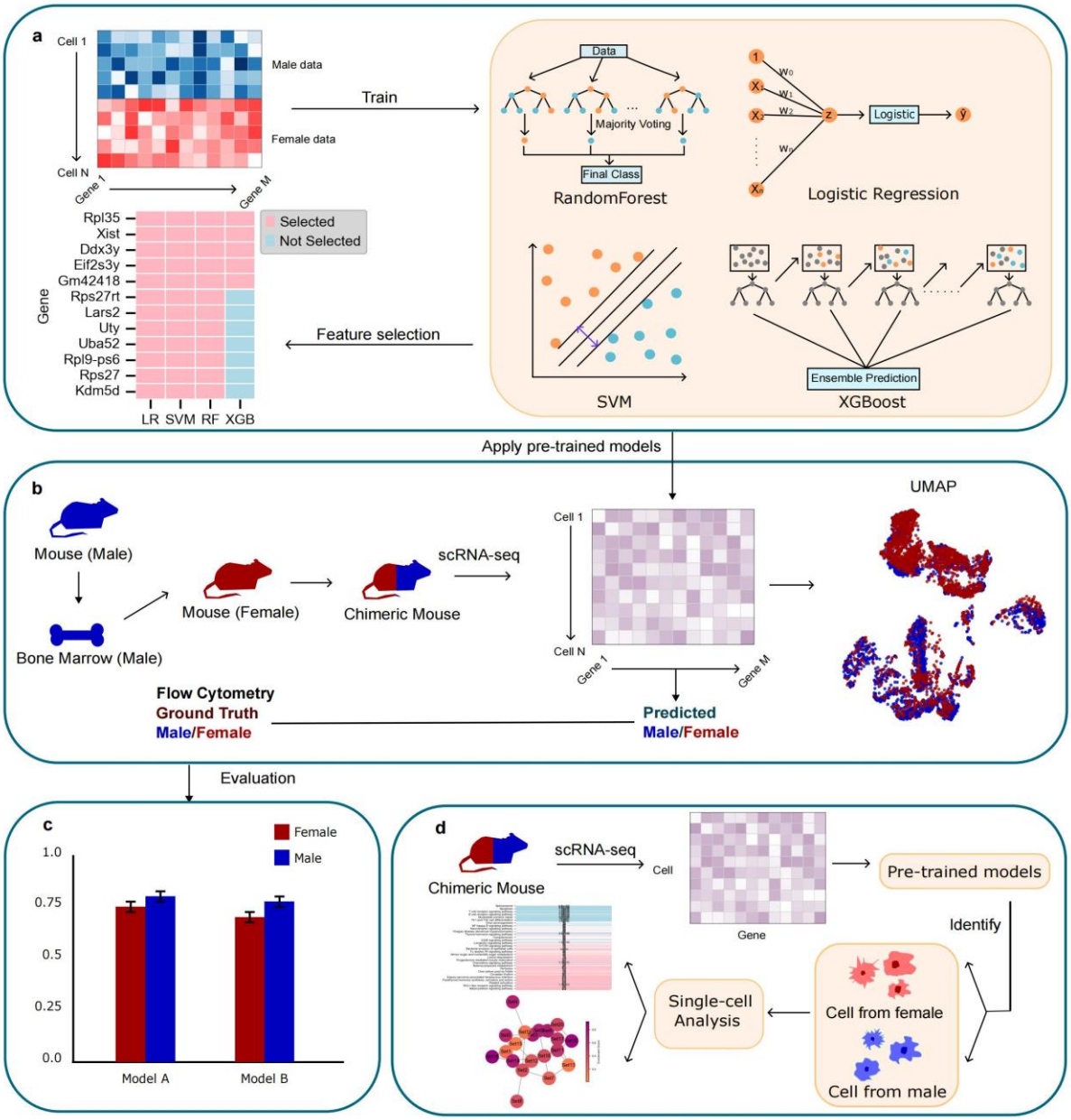
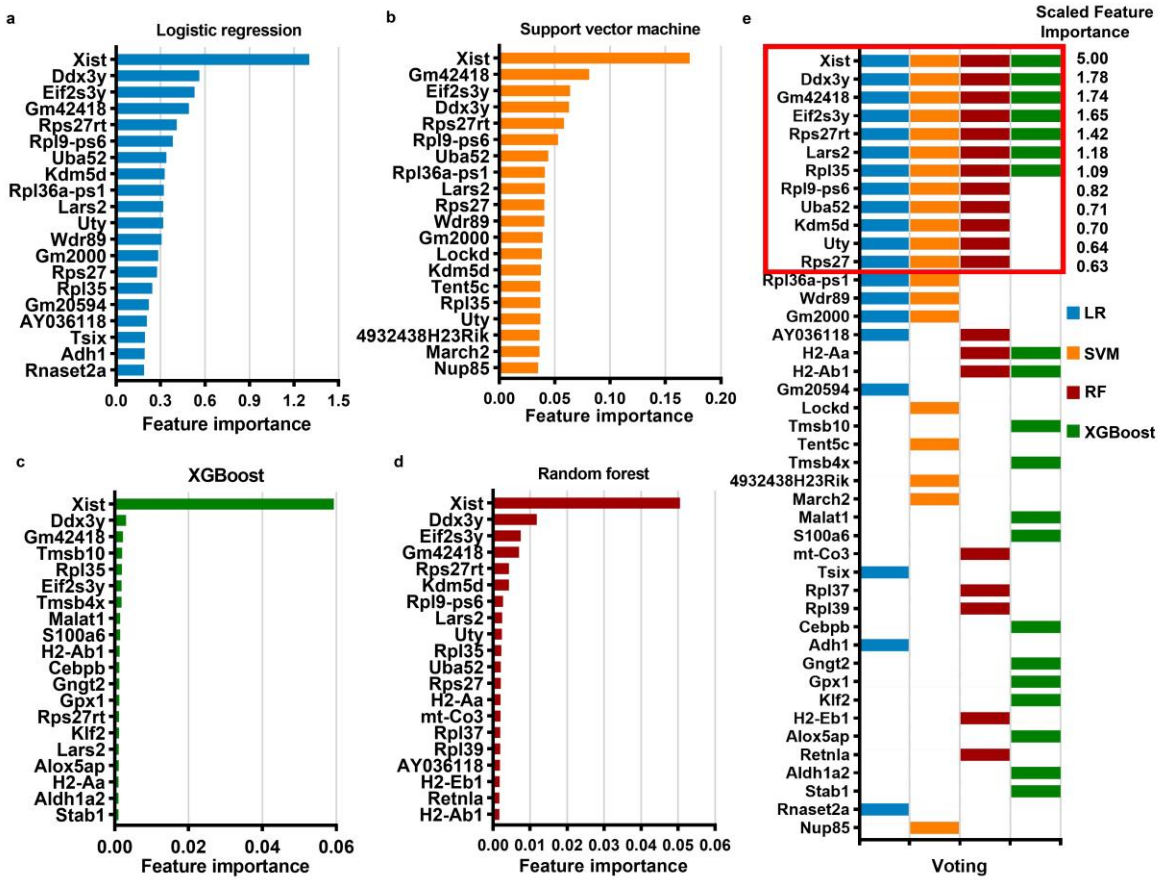


Figure 1: Overview of the CellSexID method. a, Identification of sex-specific gene features for predicting cell origin in single-cell RNA sequencing data. A committee of four classifiers (Random Forest, Logistic Regression, SVM, XGBoost) selects genes deemed critical for distinguishing cell sex, based on majority voting across models. **b,** Experimental validation of CellSexID using chimeric mouse models. The single-cell RNA-seq data from chimeric mice are used to test classifiers pre-trained on a public dataset, with predictions based on the selected gene features. Diaphragm macrophages from female mice transplanted with male bone marrow are isolated, flow-sorted, barcoded, and sequenced, serving as ground truth for validating CellSexID predictions. **c,** Evaluation of the model's predictive performance using various metrics calculated against flow cytometry-derived ground truth, providing an assessment of reliability. **d,** Application of CellSexID for annotating cell origin in chimeric mice enables a range of single-cell analyses, supporting studies of cellular dynamics and differences between recipient and donor cells in diverse research contexts.



f

Rank	Gene	Chromosome	Gene name	NCBI Gene ID
1	Xist	X	X-inactive specific transcript	213742
2	Ddx3y	Y	DEAD box helicase 3, Y-linked	26900
3	Gm42418	17	18s RNA, related sequence 5	110183
4	Eif2s3y	Y	Eukaryotic translation initiation factor 2, subunit 3, structural gene Y-linked	26908
5	Rps27rt	9	Ribosomal protein S27, retrogene	100043813
6	Lars2	9	Leucyl-tRNA synthetase, mitochondrial	102436
7	Rpl35	2	Ribosomal protein L35	66489
8	Rpl9-ps6	19	Ribosomal protein L9, pseudogene 6	100042823
9	Uba52	8	Ubiquitin A-52 residue ribosomal protein fusion product 1	22186
10	Kdm5d	Y	Lysine demethylase 5D	20592
11	Uty	Y	Ubiquitously transcribed tetratricopeptide repeat containing, Y-linked	22290
12	Rps27	3	ribosomal protein S27	57294

Figure 2: The machine learning model committee identified 12 gene markers most relevant to sex prediction using the public mouse adrenal gland dataset (GSM6153751 and GSM6153750). a-d, Top 20 gene features relevant to sex prediction identified by each committee member (Logistic Regression-LR, SVM, XGBoost, Random Forest-RF) using an 8:2 train-test split on the dataset. e, Bar plot summarizing gene selection voting across models, indicating strong consensus among committee members on robust features. The scaled feature importance is calculated by first applying min-max normalization to feature scores within each model to ensure comparability, then summing these normalized scores across all committee members to derive the overall importance for each feature. **f,** Final selection of 12 gene features with details on rank, chromosomal position, full gene name, and NCBI Gene IDs.

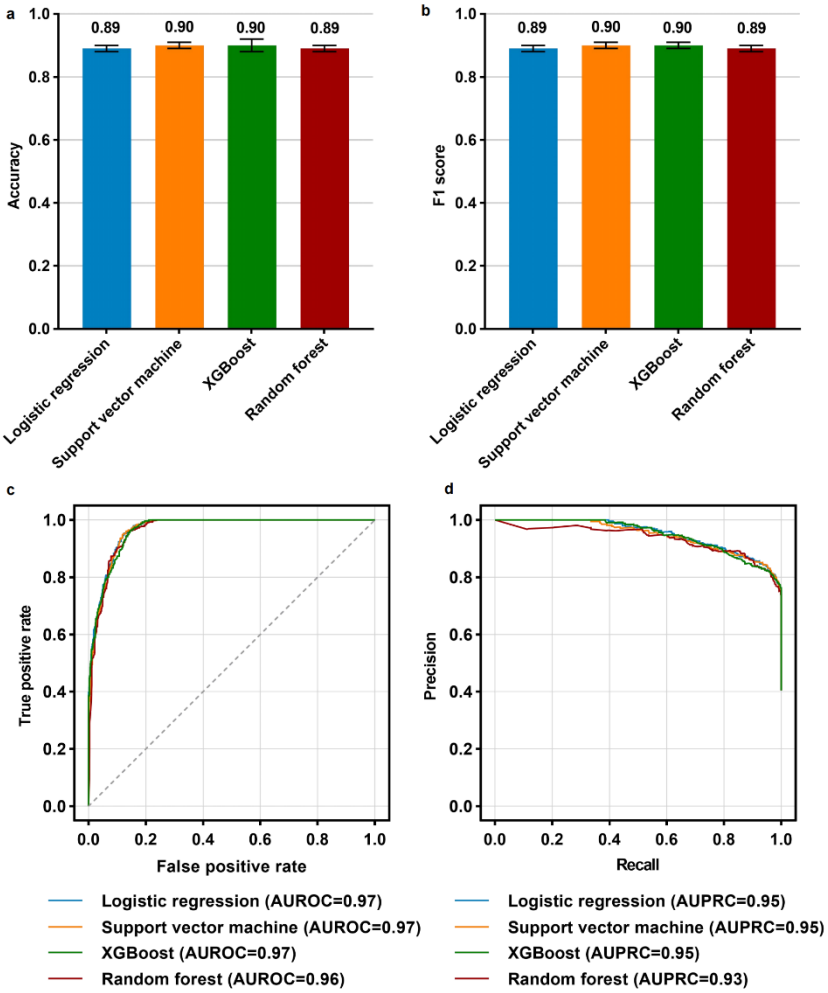


Figure 3: Validation of the cell origin classifiers and the 12 selected markers using a public mouse adrenal gland dataset with ground truth labels (GSM6153751 and GSM6153750). a, Bar plot showing Accuracy for the four models (Logistic Regression, SVM, Random Forest, and XGBoost), evaluated using 5-fold Stratified Cross-Validation on the adrenal gland tissue dataset, indicating strong discriminatory power across models. Error bars represent standard deviations. **b,** Average F1 macro scores for each classifier from 5-fold Stratified Cross-Validation, demonstrating consistent, high prediction accuracy across models. Error bars represent standard deviations. **c,** ROC curves of the four models under an 8:2 train-test split, showing near-perfect performance. **d,** AUPRC indicates strong predictive power for all models, with XGBoost slightly outperforming the others.

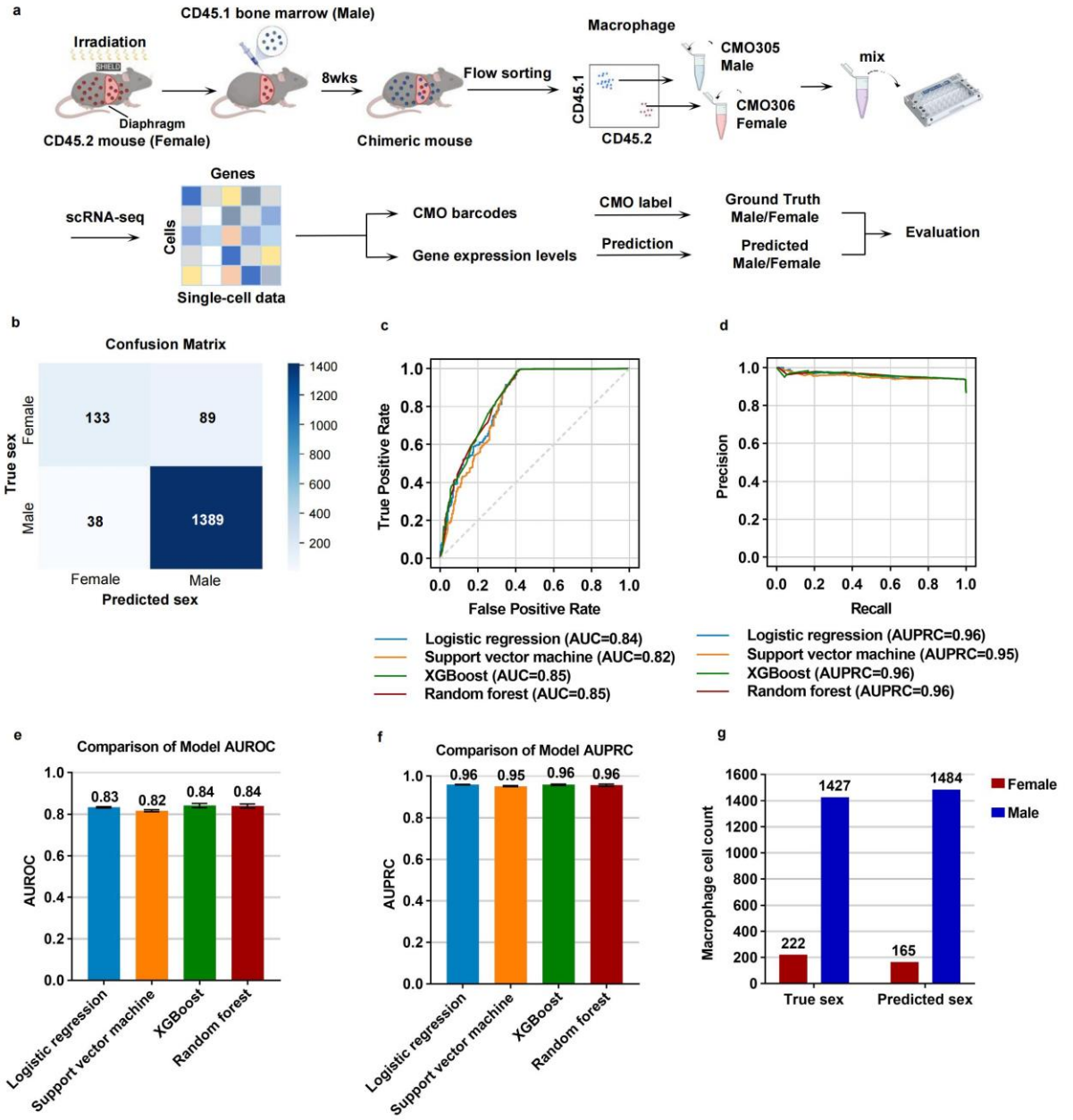


Figure 4: Reliable performance of cell origin classifiers on additional single-cell validation data from a chimeric mouse with cell sorting as ground truth. a, Generation of a validation dataset by creating chimeric mouse, collecting single-cell data, and using cell sorting as ground truth. A female CD45.2 mouse, irradiated with diaphragm shielding, was transplanted with male CD45.1 bone marrow. After 8 weeks, diaphragm macrophages were isolated, stained to identify live macrophages (CD11c-, SiglecF-, F4/80+, CD11b+), flow-sorted by CD45.1 and CD45.2 expression, barcoded (CD45.2: CMO306, CD45.1: CMO305), pooled, and sequenced using 10x Genomics. **b,** Confusion matrix from XGBoost's predictions based on the 12 selected genes, demonstrating high predictive accuracy for cell origin. **c,** Reliable prediction performance of all four classifiers (pre-trained on the public dataset) based on AUROC. **d,** Precision-Recall curves for the same models, illustrating robust prediction performance. **e-f,** Comparative analysis of AUROC and AUPRC values for each model on the validation dataset using bootstrapping, with error bars representing 2.5–97.5% confidence intervals obtained through bootstrap resampling (100 iterations, with replacement) of the training data, while maintaining the original test set for consistent evaluation. For each bootstrap iteration, models were retrained on the resampled data, and performance metrics were calculated on the fixed test set. The superior performance of XGBoost justifies our selection of it as our final prediction model. **g,** Bar plot comparing XGBoost's predicted sex (cell origin) ratio with ground truth, underscoring the model's accuracy in predicting cell origin using the selected gene features."

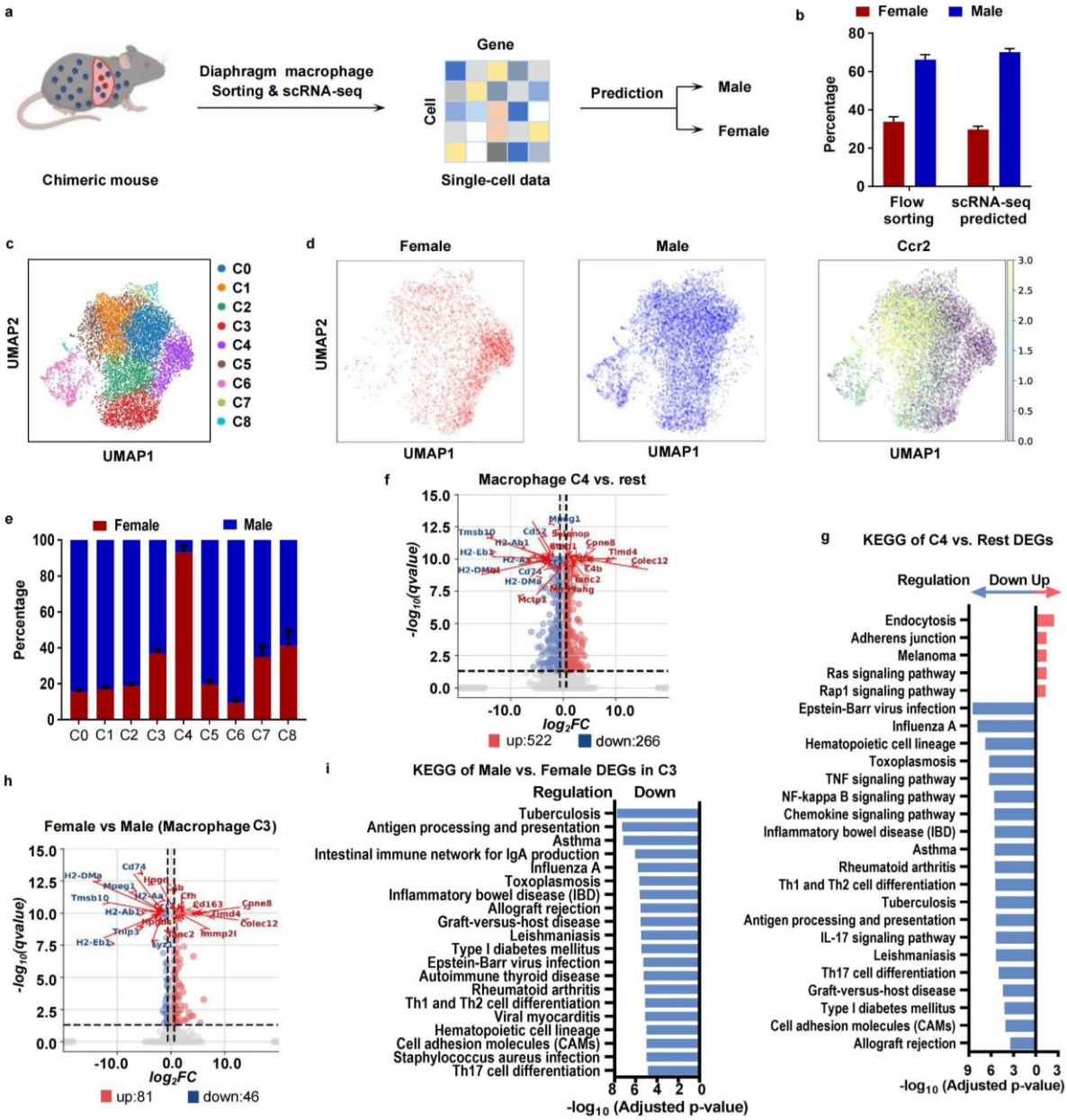


Figure 5: CellSexID Enables Origin-Specific Analysis of Macrophage Subpopulations, Revealing Unique Gene Expression Profiles and Functional Pathways. a, Schematic overview of the experimental design. **b**, Proportional distribution of macrophages from recipient and donor origins in each mouse, determined by flow cytometry sorting and scRNA-seq predictions using the XGBoost model. Data represent eight biological replicates (n=8, mean ± SEM). **c**, Unsupervised clustering analysis identifies eight distinct macrophage groups, segregating cells from recipient (female) and donor (male) origins based on unique gene expression profiles. **d**, UMAP visualization showing the spatial distribution of recipient- and donor-derived macrophages. Cell origins were annotated using the XGBoost model and validated against the monocyte-derived macrophage marker *Ccr2*, supporting the accuracy of the predictions. **e**, Proportion of recipient- and donor-derived macrophages (annotated using the pre-trained XGBoost model) within each identified cluster, highlighting origin-specific predominance in certain clusters (n=8, mean ± SEM). **f**, Volcano plot displaying the top 10 upregulated and downregulated genes in macrophage cluster C4 compared to all other clusters, highlighting significant gene expression differences. **g**, Pathway enrichment analysis of cluster C4 (top 20 pathways), indicating enhanced cell communication and proliferation alongside reduced immune and inflammatory responses compared to other clusters. **h**, Volcano plot highlighting genes with significant upregulation or downregulation in recipient- versus donor-derived macrophages within macrophage cluster C3. **i**, Pathway analysis of macrophage cluster C3 (top 20 pathways), revealing significantly downregulated immune-related pathways in recipient-derived macrophages compared to donor-derived macrophages.

REFERENCES

1. Mass, E., et al., *Tissue-specific macrophages: how they develop and choreograph tissue biology.* Nature Reviews Immunology, 2023. **23**(9): p. 563-579.
2. Park, M.D., et al., *Macrophages in health and disease.* Cell, 2022. **185**(23): p. 4259-4279.
3. Swirski, F.K., et al., *Ly-6C hi monocytes dominate hypercholesterolemia-associated monocytosis and give rise to macrophages in atheromata.* The Journal of clinical investigation, 2007. **117**(1): p. 195-205.
4. Misharin, A.V., et al., *Monocyte-derived alveolar macrophages drive lung fibrosis and persist in the lung over the life span.* Journal of Experimental Medicine, 2017. **214**(8): p. 2387-2404.
5. Gussoni, E., et al., *Dystrophin expression in the mdx mouse restored by stem cell transplantation.* Nature, 1999. **401**(6751): p. 390-394.
6. Trotman, W., et al., *Dual Y chromosome painting and in situ cell-specific immunofluorescence staining in lung tissue: an improved method of identifying donor marrow cells in lung following bone marrow transplantation.* Histochemistry and cell biology, 2004. **121**: p. 73-79.
7. Byrne, P., et al., *Chimerism analysis in sex-mismatched murine transplantation using quantitative real-time PCR.* Biotechniques, 2002. **32**(2): p. 278-286.
8. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nature methods, 2009. **6**(5): p. 377-382.
9. Kircher, M.F., S.S. Gambhir, and J. Grimm, *Noninvasive cell-tracking methods.* Nature reviews Clinical oncology, 2011. **8**(11): p. 677-688.
10. Chen, T. and C. Guestrin. *Xgboost: A scalable tree boosting system*. in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
11. Vasighizaker, A., Y. Trivedi, and L. Rueda, *Cell Type Annotation Model Selection: General-Purpose vs. Pattern-Aware Feature Gene Selection in Single-Cell RNA-Seq Data.* Genes, 2023. **14**(3): p. 596.
12. Lieberman, Y., L. Rokach, and T. Shay, *CaSTLe–classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments.* PloS one, 2018. **13**(10): p. e0205499.
13. Platt, J., *Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.* Advances in large margin classifiers, 1999. **10**(3): p. 61-74.
14. Le, H., et al., *Machine learning for cell type classification from single nucleus RNA sequencing data.* Plos one, 2022. **17**(9): p. e0275070.
15. Nguyen, V. and J. Griss, *scAnnotatR: framework to accurately classify cell types in single-cell RNA-sequencing data.* BMC bioinformatics, 2022. **23**(1): p. 44.
16. Breiman, L., *Random forests.* Machine learning, 2001. **45**: p. 5-32.
17. Caruana, R. and A. Niculescu-Mizil. *An empirical comparison of supervised learning algorithms*. in *Proceedings of the 23rd international conference on Machine learning*. 2006.
18. Díaz-Uriarte, R. and S. Alvarez de Andrés, *Gene selection and classification of microarray data using random forest.* BMC bioinformatics, 2006. **7**: p. 1-13.
19. Yang, P., H. Huang, and C. Liu, *Feature selection revisited in the single-cell era.* Genome Biology, 2021. **22**: p. 1-17.
20. Pouyan, M.B. and D. Kostka, *Random forest based similarity learning for single cell RNA sequencing data.* Bioinformatics, 2018. **34**(13): p. i79-i88.
21. Yu, H.-F., F.-L. Huang, and C.-J. Lin, *Dual coordinate descent methods for logistic regression and maximum entropy models.* Machine Learning, 2011. **85**: p. 41-75.
22. Cox, D.R., *The regression analysis of binary sequences.* Journal of the Royal Statistical Society Series

B: Statistical Methodology, 1958. **20**(2): p. 215-232.

23. Wei, Z. and S. Zhang, *CALLR: a semi-supervised cell-type annotation method for single-cell RNA sequencing data.* Bioinformatics, 2021. **37**(Supplement_1): p. i51-i58.
24. Dolfi, B., et al., *Unravelling the sex-specific diversity and functions of adrenal gland macrophages.* Cell reports, 2022. **39**(11).
25. Hoque, N., M. Singh, and D.K. Bhattacharyya, *EFS-MI: an ensemble feature selection method for classification: An ensemble feature selection method.* Complex & Intelligent Systems, 2018. **4**: p. 105-118.
26. Saeys, Y., T. Abeel, and Y. Van de Peer. *Robust feature selection using ensemble feature selection techniques*. in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19*. 2008. Springer.
27. Yang, X., et al., *Ensemble feature selection with block-regularized m× 2 cross-validation.* IEEE Transactions on Neural Networks and Learning Systems, 2021. **34**(9): p. 6628-6641.
28. Pes, B., *Ensemble feature selection for high-dimensional data: a stability analysis across multiple domains.* Neural Computing and Applications, 2020. **32**(10): p. 5951-5973.
29. Brockdorff, N., J.S. Bowness, and G. Wei, *Progress toward understanding chromosome silencing by Xist RNA.* Genes & development, 2020. **34**(11-12): p. 733-744.
30. Liu, W., et al., *The role of Eif2s3y in mouse spermatogenesis and ESC.* Current Stem Cell Research & Therapy, 2022. **17**(8): p. 750-755.
31. Dicke, A.-K., et al., *DDX3Y is likely the key spermatogenic factor in the AZFa region that contributes to human non-obstructive azoospermia.* Communications Biology, 2023. **6**(1): p. 350.
32. Villesen, P. and T. Fredsted, *Fast and non-invasive PCR sexing of primates: apes, old world monkeys, new world monkeys and Strepsirrhines.* BMC ecology, 2006. **6**: p. 1-6.
33. Nucera, S., et al., *A novel flow-cytometric based method to assess post-HSCT donor chimerism exploiting RNA hybridization.* Bone Marrow Transplantation, 2024. **59**(2): p. 171-177.
34. Pottmeier, P., et al., *Sex-biased gene expression during neural differentiation of human embryonic stem cells.* Frontiers in Cell and Developmental Biology, 2024. **12**: p. 1341373.
35. Devulapalli, R., et al., *Males and females differ in the regulation and engagement of, but not requirement for, protein degradation in the amygdala during fear memory formation.* Neurobiology of learning and memory, 2021. **180**: p. 107404.
36. Paik, D.T., et al., *Single-cell RNA sequencing unveils unique transcriptomic signatures of organ-specific endothelial cells.* Circulation, 2020. **142**(19): p. 1848-1862.
37. Herzenberg, L.A., et al., *The history and future of the fluorescence activated cell sorter and flow cytometry: a view from Stanford.* Clinical chemistry, 2002. **48**(10): p. 1819-1827.
38. Li, Q., et al., *Dynamic equilibrium of skeletal muscle macrophage ontogeny in the diaphragm during homeostasis, injury, and recovery.* Scientific Reports, 2024. **14**(1): p. 9132.
39. Serbina, N.V., et al., *TNF/iNOS-producing dendritic cells mediate innate immune defense against bacterial infection.* Immunity, 2003. **19**(1): p. 59-70.
40. Willenborg, S., et al., *CCR2 recruits an inflammatory macrophage subpopulation critical for angiogenesis in tissue repair.* Blood, The Journal of the American Society of Hematology, 2012. **120**(3): p. 613-625.
41. Fischer, K.B., H.K. Collins, and E.M. Callaway, *Sources of off-target expression from recombinase-dependent AAV vectors and mitigation with cross-over insensitive ATG-out vectors.* Proceedings of the National Academy of Sciences, 2019. **116**(52): p. 27001-27010.

42. Dick, S.A., et al., *Self-renewing resident cardiac macrophages limit adverse remodeling following myocardial infarction.* Nature immunology, 2019. **20**(1): p. 29-39.
43. Heidt, T., et al., *Differential contribution of monocytes to heart macrophages in steady-state and after myocardial infarction.* Circulation research, 2014. **115**(2): p. 284-295.
44. Bhattarai, S., et al., *TLR4 is a regulator of trained immunity in a murine model of Duchenne muscular dystrophy.* Nature communications, 2022. **13**(1): p. 879.
45. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nature communications, 2017. **8**(1): p. 14049.
46. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome biology, 2018. **19**: p. 1-5.
47. Deng, X., et al., *Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification.* Medical & Biological Engineering & Computing, 2022. **60**(3): p. 663-681.
48. Cao, J., et al. *Specific Emitter Identification Based on ACO-XGBoost Feature Selection*. in *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. 2022. Springer.
49. Iranzad, R. and X. Liu, *A review of random forest-based feature selection methods for data science education and applications.* International Journal of Data Science and Analytics, 2024: p. 1-15.
50. Liu, Q., et al., *Feature selection for support vector machines with RBF kernel.* Artificial Intelligence Review, 2011. **36**(2): p. 99-115.
51. Lessmann, S., R. Stahlbock, and S.F. Crone. *Genetic algorithms for support vector machine model selection*. in *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. 2006. IEEE.
52. Hermes, L. and J.M. Buhmann. *Feature selection for support vector machines*. in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*. 2000. IEEE.
53. Civitelli, E., et al., *An effective procedure for feature subset selection in logistic regression based on information criteria.* Computational Optimization and Applications, 2021. **80**(1): p. 1-32.
54. Saarela, M. and S. Jauhiainen, *Comparison of feature importance measures as explanations for classification models.* SN Applied Sciences, 2021. **3**(2): p. 272.
55. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
56. Breiman, L., *Classification and regression trees*. 2017: Routledge.
57. Quinlan, J.R., *C4. 5: programs for machine learning*. 2014: Elsevier.
58. Tipping, M.E. and C.M. Bishop, *Probabilistic principal component analysis.* Journal of the Royal Statistical Society Series B: Statistical Methodology, 1999. **61**(3): p. 611-622.
59. Halko, N., P.-G. Martinsson, and J.A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions.* SIAM review, 2011. **53**(2): p. 217-288.
60. Martinsson, P.-G., V. Rokhlin, and M. Tygert, *A randomized algorithm for the decomposition of matrices.* Applied and Computational Harmonic Analysis, 2011. **30**(1): p. 47-68.
61. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nature methods, 2019. **16**(12): p. 1289-1296.
62. McInnes, L., J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction.* arXiv preprint arXiv:1802.03426, 2018.
63. Wilcoxon, F., *Individual comparisons by ranking methods*, in *Breakthroughs in statistics: Methodology and distribution*. 1992, Springer. p. 196-202.
64. Thomas, P.D., et al., *PANTHER: Making genome-scale phylogenetics accessible to all.* Protein Science,

2022. **31**(1): p. 8-22.
65. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
66. Carbon, S., et al., *AmiGO: online access to ontology and annotation data.* Bioinformatics, 2009. **25**(2): p. 288-289.
67. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences, 2005. **102**(43): p. 15545-15550.
68. Fang, Z., X. Liu, and G. Peltz, *GSEApy: a comprehensive package for performing gene set enrichment analysis in Python.* Bioinformatics, 2023. **39**(1): p. btac757.