Accurate sample deconvolution of pooled snRNA-seq using sex-dependent gene expression patterns

Guy M. Twa[1], Robert A. Phillips III[1,2], Nathaniel J. Robinson[1], Jeremy J. Day[1]*

Summary: Single nucleus RNA sequencing (snRNA-seq) technology offers unprecedented resolution for studying cell type-specific gene expression patterns. However, snRNA-seq poses high costs and technical limitations, often requiring the pooling of independent biological samples and loss of individual sample-level data. Deconvolution of sample identity using inherent features would enable the incorporation of pooled barcoding and sequencing protocols, thereby increasing data throughput and analytical sample size without requiring increases in experimental sample size and sequencing costs. In this study, we demonstrate a proof of concept that sex-dependent gene expression patterns can be leveraged for the deconvolution of pooled snRNA-seq data. Using previously published snRNA-seq data from the rat ventral tegmental area, we trained a range of machine learning models to classify cell sex using genes differentially expressed in cells from male and female rats. Models that used sex-dependent gene expression predicted cell sex with high accuracy (90-92%) and outperformed simple classification models using only sex chromosome gene expression (69-89%). The generalizability of these models to other brain regions was assessed using an additional published data set from the rat nucleus accumbens. Within this data set, model performance remained highly accurate in cell sex classification (89-90% accuracy) with no additional re-training. This work provides a model for future snRNA-seq studies to perform sample deconvolution using a two-sex pooled sample sequencing design and benchmarks the performance of various machine learning approaches to deconvolve sample identification from inherent sample features.

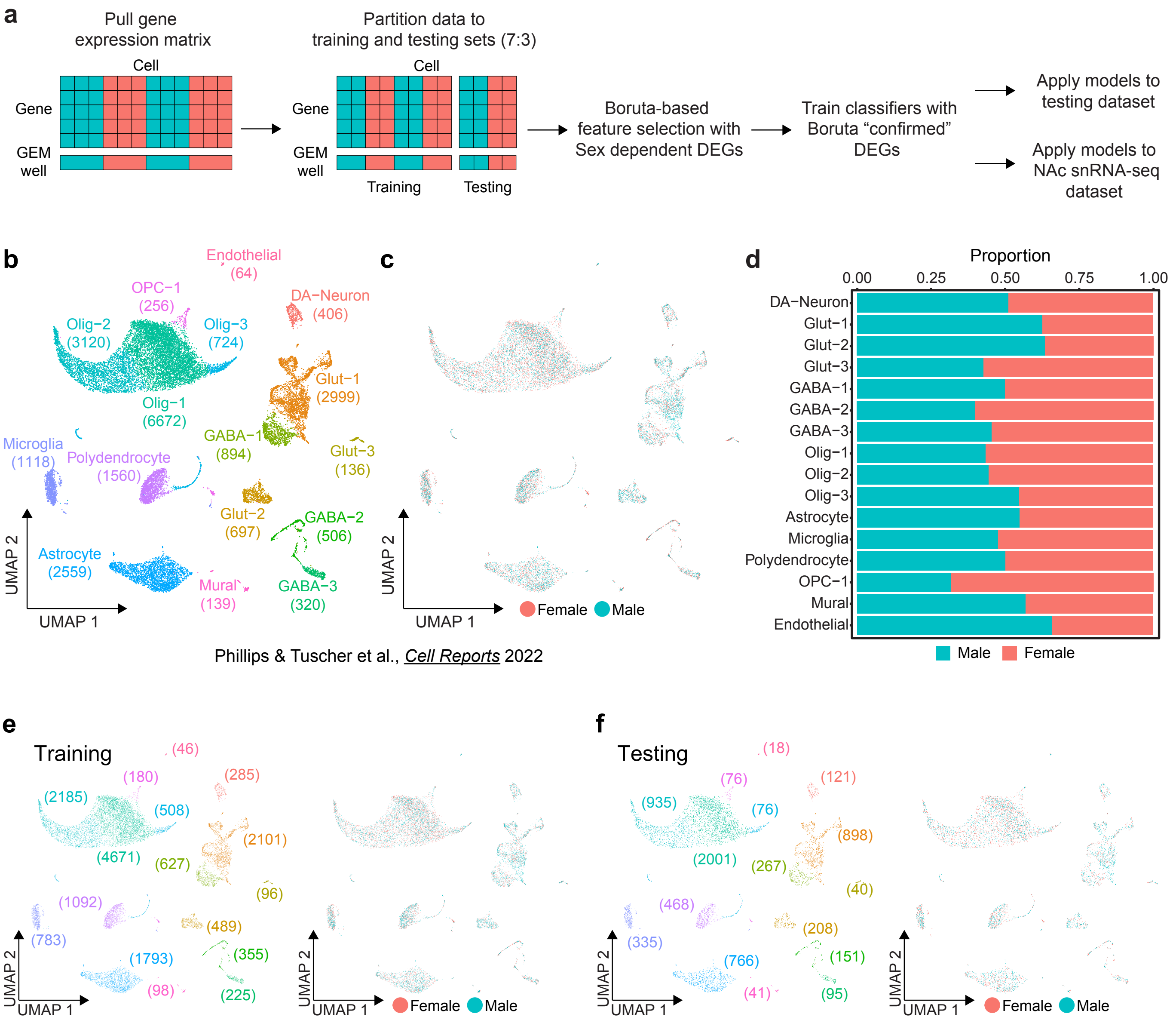
INTRODUCTION

Single nucleus RNA sequencing (snRNA-seq) technologies enable the investigation of cell type-specific gene expression patterns, which are highly relevant to brain health and disease mechanisms [1–3]. Sample size is a key factor in snRNA-seq experimental design, as the number of experimental samples directly influences experimental costs and statistical power [4–6]. Additionally, droplet-based snRNA-seq technologies often require pooling tissue samples from multiple animals to achieve appropriate nuclei concentrations for maximum capture. Without the ability to deconvolute these samples, pooling decreases the effective sample size and causes individual-level data loss, thereby increasing the experimental costs required for sufficient statistical power with current gold-standard pseudobulking analysis approaches [7]. Additionally, the loss of individual-level data prevents the correlation of genetic contributions to molecular, physiological, or behavioral observations. Therefore, techniques to efficiently deconvolve pooled sample data are necessary to improve the power and cost-effectiveness of snRNA-seq technologies.

Currently, there are two popular methods for pooled sample deconvolution: nucleus hashing and genotype-based multiplexing [5,8,9]. Nuclei hashing requires tagging individual sample nuclei with oligonucleotide barcoded antibodies, before pooling with other samples in sequence library preparation [9]. Following sequencing, cells are assigned to their respective samples based on the presence of detected hashing tags. This approach is limited by ambient hashing tag signal in suspension and attachment of hashing antibodies to lysis debris, both of which can make sample assignment non-trivial and noisy [9,10]. Genotype-based multiplexing deconvolutes cells by assigning them to samples based on shared genomic variants observed in sequenced nuclei libraries and sample genotypes [8]. This approach requires the collection of an additional data modality (sample genotype) and is limited by insufficient coverage of variant loci in snRNA-seq data. Multiplexing strategies utilizing both techniques may allow each to compensate for the other's weaknesses [11,12]. However, this complicates sample processing by compounding sample preparation and data analysis requirements.

Here, we outline a model of pooled sample deconvolution centered on the classification of cell sex by sex-dependent transcriptome features inherent within the data, which can enable sample assignment

[1] Department of Neurobiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[2] Present affiliation: Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205, USA.
*Correspondence to Jeremy Day (jjday@uab.edu | day-lab.org | @daylab.bsky.social | @DayLabUAB)



Phillips & Tuscher et al., *Cell Reports* 2022

Figure 1. Sex classification model training workflow and ground truth data set. a, Schematic of classifier workflow including data preparation, feature selection, model training, and performance evaluation. b, UMAP of previously published rat ventral tegmental area (VTA) single nucleus RNA-seq data set including annotation of transcriptionally defined cell types (22,170 cells, 16 cell types) c, Distribution of sexes within UMAP. d, Stacked bar chart of the proportion of each sex within each cell type. e, UMAP of the training partition (70%) of the dataset with distributions of cell types (left) and sexes (right). f, UMAP of the testing partition (30%) of the VTA dataset with distributions of cell types (left) and sexes (right).

without the need for additional data modalities or sample preprocessing. We demonstrate that cell sex can be reliably recovered from snRNA-seq data by machine learning (ML) models, providing a proof of concept for sex-based sample pooling and deconvolution. This approach requires pooling nuclei of two animals of different sexes within a single microfluidic well and *post hoc* model-assisted cell sex classification for sample deconvolution. To determine the feasibility of ML models in cell sex classification, we trained and evaluated the performance of several widely used models with a range of complexities. When applied to previously published snRNA-seq data from the ventral tegmental area (VTA) of the rat brain, models trained using sex-dependent differentially expressed genes (DEGs) accurately classified up to 96% of cells. Models maintained high performance (90% accuracy) when applied to independent snRNA-seq data from the rat nucleus accumbens (NAc), demonstrating their generalizability to distinct cell types in a different brain area. This analysis supports the pooling and deconvolution of samples in snRNA-seq assays based on sex, effectively halving experimental costs and doubling analytical power.

RESULTS

Determination of sex-dependent transcriptome features

To learn rules for predicting cell sex using transcriptome features, ML models require a relevant ground-truth training set of cells with known sex. To this end, we used previously published snRNA-seq data from the rat VTA containing 22,170 cells

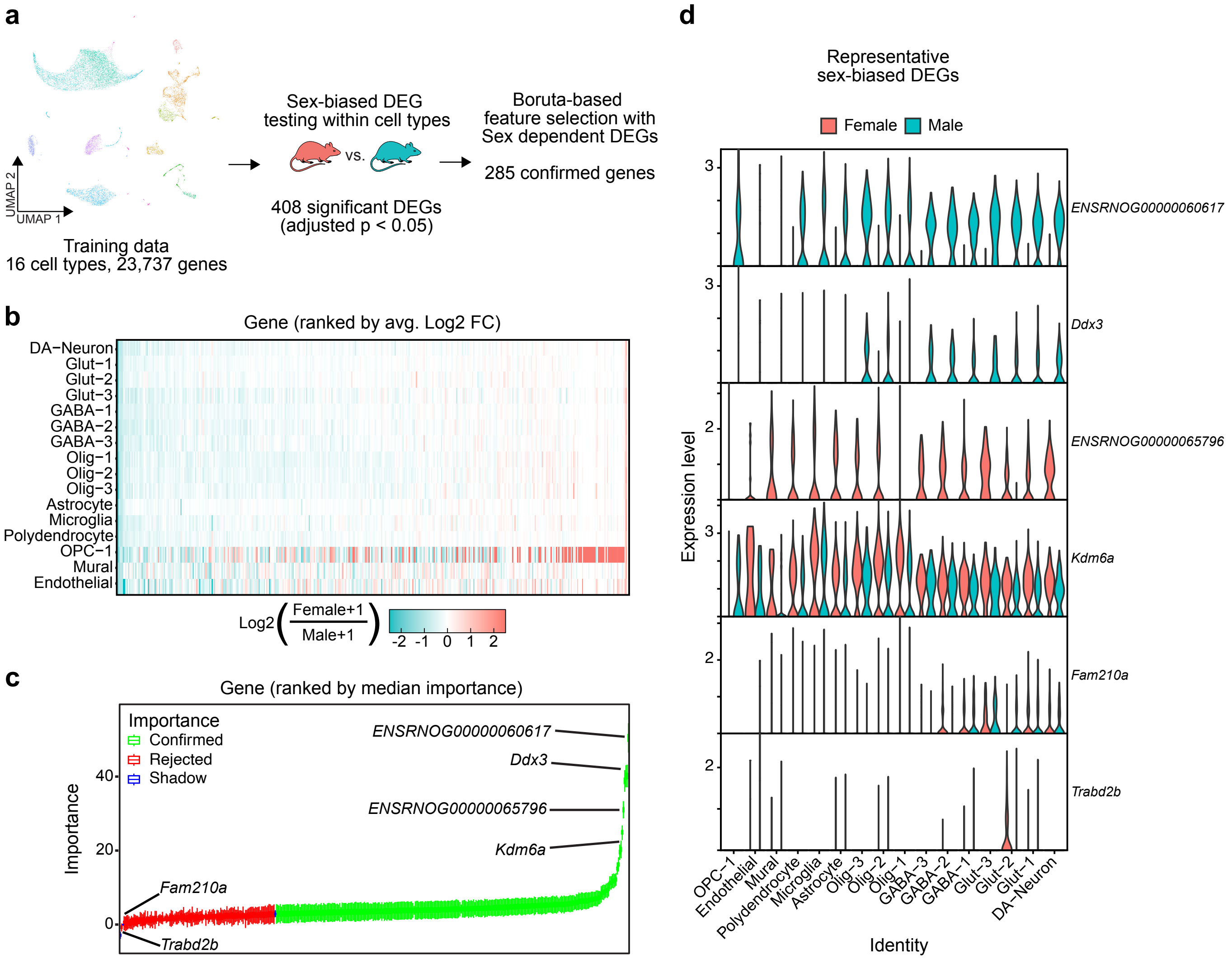


Figure 2. 285 Sex-dependent DEGs confirmed as important features for cell sex classification model **a**, Schematic of feature selection using the VTA training data partition. **b**, Heatmap of log2 fold changes of genes significantly differentially expressed (adjusted p < 0.05) in at least one cell type (408 genes), color scale is capped from -2.5 to 2.5. Genes are ordered along the x-axis by ranked mean log2 fold change across cell types. **c**, Boxplots of gene importance scores as determined through 2000 iterations of the Boruta algorithm. Genes are ordered along the x-axis by median importance score. Boxplots are colored by the final importance decision by Boruta. ("Shadow" features are generated by Boruta from randomly shuffled gene data, used to evaluate the importance of real feature data). **d**, Violin plots of gene expression levels, within cell type split by sex, of the two highest confirmed importance male (ENSRNOG00000060617 (Uty), Ddx3) and female (ENSRNOG00000065796 (Xist), Kdm6a) biased, and the two lowest importance rejected (Fam210a, Trabd2b) sex-dependent DEGs determined by Boruta.

and 16 transcriptionally defined cell types (**Fig. 1b**) [13]. This data set was obtained from pooled male or female samples that were assigned to separate 10X GEM wells and contained a roughly equal number of male and female cells of each annotated cell type (**Fig. 1c,d**). Cells were split into training and testing partitions of 70% and 30% respectively (15,534:6,636) while maintaining cell type and cell sex ratios (**Fig. 1e,f**). The testing partition was set aside until model evaluation, and we leveraged the training partition to determine relevant sex-dependent transcriptome features and train classification models.

Selecting variables informative for class prediction is important for efficient model training and high performance. To select the transcriptome features most relevant for cell sex classification, we implemented two successive feature selection steps: sex-dependent differentially expressed gene (DEG) testing and Boruta feature selection (**Fig. 2a**) [14]. Of 25,732 detected genes, DEG testing identified 407 genes with significant sex-dependent expression in at least one cell type (FDR < 0.05) (**Fig. 2b**). Sex-dependent directionality of expression of these genes was broadly replicated across cell types. To refine the set of genes used for prediction models, we applied the Boruta algorithm, which identifies important features by comparing their impact on classification accuracy to shuffled "shadow" features [14]. Boruta rejected 123 genes and confirmed 285 as more important than shadow features (**Fig. 2c**). Genes with known sex-dependent gene expression, such as X chromosome genes *Xist* (*ENSRNOG00000065796*) and *Kdm6a* and Y chromosome genes *Uty* (ENSRNOG00000060617) and *Ddx3*, replicated sex-specific expression patterns in our data and were some of the most important features selected by Boruta (**Fig. 2c,d**). Conversely, genes expressed sparsely and without prior evidence of sex-biased transcription were rejected (**Fig. 2c,d**). Together, these processes narrowed the set of predictor variables to 285 transcripts (~1% of all

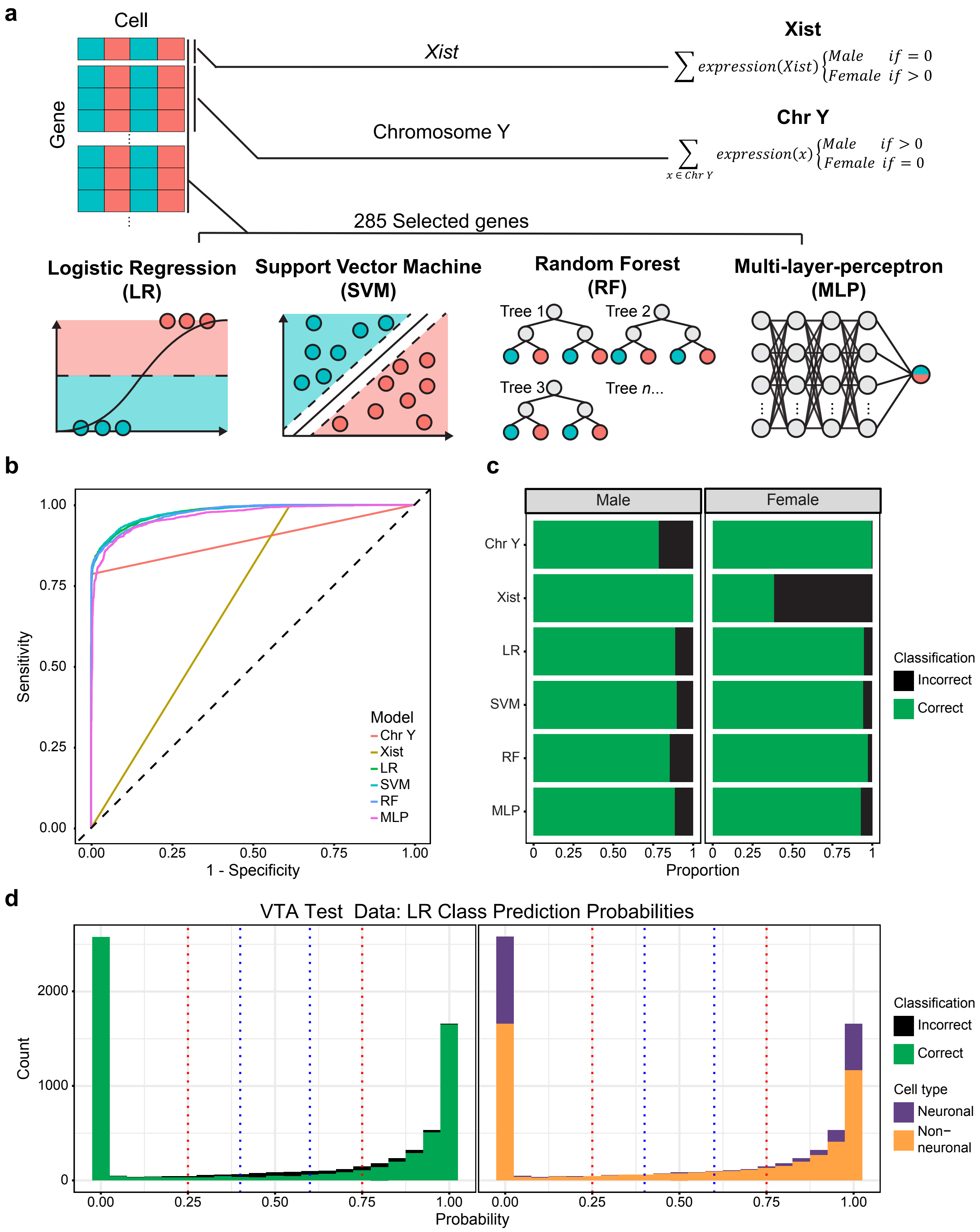


Figure 3. Machine learning models using selected genes outperform simple classifiers in sex classification. **a**, Schematic of models evaluated, the predictor genes used, and classification framework. **b**, Receiver operating characteristic curve of model sensitivity and specificity for classification of the VTA testing data partition. **c**, Stacked bar chart of the proportion of correct and incorrect classifications of the VTA testing data partition sexes. **d**, Histogram of logistic regression predicted cell sex probabilities (closer to 1 = high probability of being female, closer to 0 = high probability of being male) for the VTA test partition, bin size = 0.05. Blue and red dotted lines represent thresholds of 0.4-0.6 and 0.25-0.75 respectively. Bins are colored by the proportion of incorrect/correct classifications (left) and cell types (right).

detected genes) with sex-dependent transcription that demonstrate a greater impact on cell sex classification accuracy than null shadow features.

Machine learning models perform accurate cell sex classification

A wide range of approaches have been developed for classification tasks. To assess the potential for various models in cell sex classification,

Table 1. Ventral tegmental area model performance. Cell sex classification of the ventral tegmental area (VTA) testing partition was performed using two non-ML and four ML models. Model performance was assessed by accuracy and AUC-ROC. Model: Name of classification model; Predictors: set of predictor variables in model; Training time: total model training time in hours (hr), minutes (min), and seconds (s); Overall: performance measured across all cells; Neuronal: performance measured across only neuronal cell types; Non-neuronal: performance measured across only non-neuronal cell types; AUC-ROC: area under the curve of the receiver operating characteristic curve; Accuracy: proportion of correct classifications out of all classifications.

			Overall		Neuronal		Non-Neuronal	
Model	**Predictors**	**Training time (hr:min:s)**	**AUC-ROC**	**Accuracy**	**AUC-ROC**	**Accuracy**	**AUC-ROC**	**Accuracy**
Xist	*Xist* (ENSRNOG00000065796)	N/A	0.692	0.689	0.768	0.799	0.670	0.648
Chr Y	Chromosome Y genes	N/A	0.892	0.892	0.965	0.954	0.861	0.869
Logistic Regression (LR)	285 Selected features	00:00:29	0.978	0.917	0.994	0.971	0.968	0.898
Support Vector Machine (SVM)	285 Selected features	26:12:51	0.978	0.920	0.995	0.970	0.969	0.902
Random Forest (RF)	285 Selected features	14:47:29	0.976	0.913	0.996	0.970	0.963	0.893
Multi-Layer-Perceptron (MLP)	285 Selected features	07:31:27	0.966	0.907	0.988	0.948	0.955	0.892

we surveyed four common ML models of varying complexity: logistic regression (LR), random forest (RF), support vector machine (SVM), and neural network multilayer-perceptron (MLP) models [15–22]. Additionally, we prepared two non-ML cell sex classification models based on the binary expression of sex chromosome genes: Xist, based on the expression of *Xist*, and Chr Y, based on the expression of chromosome Y genes. Together, these models represent various levels of complexity in approaches to cell sex classification strategies (**Fig. 3a**).

When applied to the testing partition of the VTA data set, all ML models outperformed non-ML classification models based only on sex chromosome gene predictions. We evaluated models based on their overall accuracy, as well as sensitivity (true positive rate) and specificity (true negative rate) using the receiver operating characteristic curve (ROC). The area under the curve of the ROC (AUC-ROC) provides a quantitative measure of how well a classifier assigns observations, where a perfectly balanced classifier has an AUC-ROC of 1 and a random classifier has an AUC-ROC of 0.5. ML models achieved high overall accuracy (91-92%) and had balanced sensitivity and specificity (0.97-0.98 AUC-ROC) in cell sex classification (**Fig. 3b,c, Table 1**). LR and SVM models outperformed all other methods (overall accuracies ~92%, AUC-ROCs ~0.98), despite the SVM model taking over 3,000x longer to train (**Fig. 3b,c, Table 1**). RF and MLP models only slightly underperformed LR and SVM models, with overall accuracies of ~91% and AUC-ROCs of 0.98 and 0.97, respectively. Non-ML classifiers suffered from poor accuracy, sensitivity, and specificity relative to the ML models. The Xist model was perfectly sensitive (misclassifying no male cells), but lacked specificity (misclassifying nearly 60% of female cells; **Fig. 3b,c**). Conversely, the Chr Y model was nearly perfectly specific, misclassifying only 17 female cells (potentially due to misaligned transcripts), and poorly sensitive, misclassifying ~22% of male cells (**Fig. 3b,c**). The high proportion of misclassifications by the non-ML models highlights a potential disadvantage of relying on a few predictors, as single nuclei data is prone to sparsity and gene drop-out. Accuracy and the number of UMIs (unique molecular identifiers) were significantly positively associated with one another ($p < 0.05$) for all models (**Fig. S1, Table S1**). With the exception of the MLP model, all ML models required fewer UMIs to reach a 95% predicted probability of correct classification as compared to non-ML classifiers. Together, these results demonstrate the ability of ML models, utilizing a robust set of sex-dependent genes, to accurately classify snRNA-seq cell sex.

Table 2. Predicted probability thresholding improves accuracy. Ventral tegmental area cells with tenuous (~0.5) logistic regression (LR) predicted class probabilities were either retained or omitted using either narrow (0.4-0.6) or wide (0.25-0.75) thresholds. None: No omission by thresholding; 0.4-0.6: Omission of cells with class probabilities between 0.4 and 0.6; 0.25-0.75: Omission of cells with class probabilities between 0.25 and 0.75; N cells included: number of cells included after thresholding; N cells excluded: number of cells excluded after thresholding; N female cells included: number of female cells included after thresholding; N female cells included: number of female cells excluded after thresholding; Accuracy of included: Proportion of correct classifications out of all classifications made on included cells, outside of the threshold; Accuracy of excluded: Proportion of correct classifications out of all classifications made on excluded cells, inside of the threshold.

		Threshold	
	None	**0.4 - 0.6**	**0.25 - 0.75**
N cells included	6636	6310	5790
N cells excluded	0	326	846
N female cells included	3357	3182	2887
N female cells excluded	0	175	470
Accuracy of included	0.917	0.939	0.964
Accuracy of excluded	N/A	0.506	0.596

Machine learning models predict cell sex probabilities from 0 (likely male) to 1 (likely female), enabling us to refine accuracy by omitting tenuous predictions near the classification threshold of 0.5. We assessed the number of cells omitted and improvements to accuracy for two thresholds centered around 0.5: narrow (0.4-0.6) and wide (0.25-0.75) (**Fig. 3d**). We found that the accuracy of cell sex predictions within the narrow threshold was 51% as compared to 94% accuracy outside, and within the wide threshold was 60% as compared to 96% outside (**Table 2**). The omission of cells within the narrow and wide thresholds would exclude 4.9% to 12.75% of cells from sex prediction, respectively (**Table 2**). Notably, cells within these low-accuracy thresholds tended to be female and non-neuronal cells (**Table 2**). The replication of this trend across all models highlights a broader challenge in cell sex classification – accurate classification of non-neuronal cells (**Table 1**).

Models performance is limited by information content in non-neuronal cell transcriptome

To explore the performance disparity between neuronal and non-neuronal cell populations, we assessed the performance of models trained specifically for either cell class. Class-specific models were trained with the same workflow as pan-cellular models, using the VTA training dataset's cell type-specific data partitions (**Fig. 4a**). Independent feature selection confirmed the importance of 55/70 neuronal and 232/364 non-neuronal DEGs, with 24 features

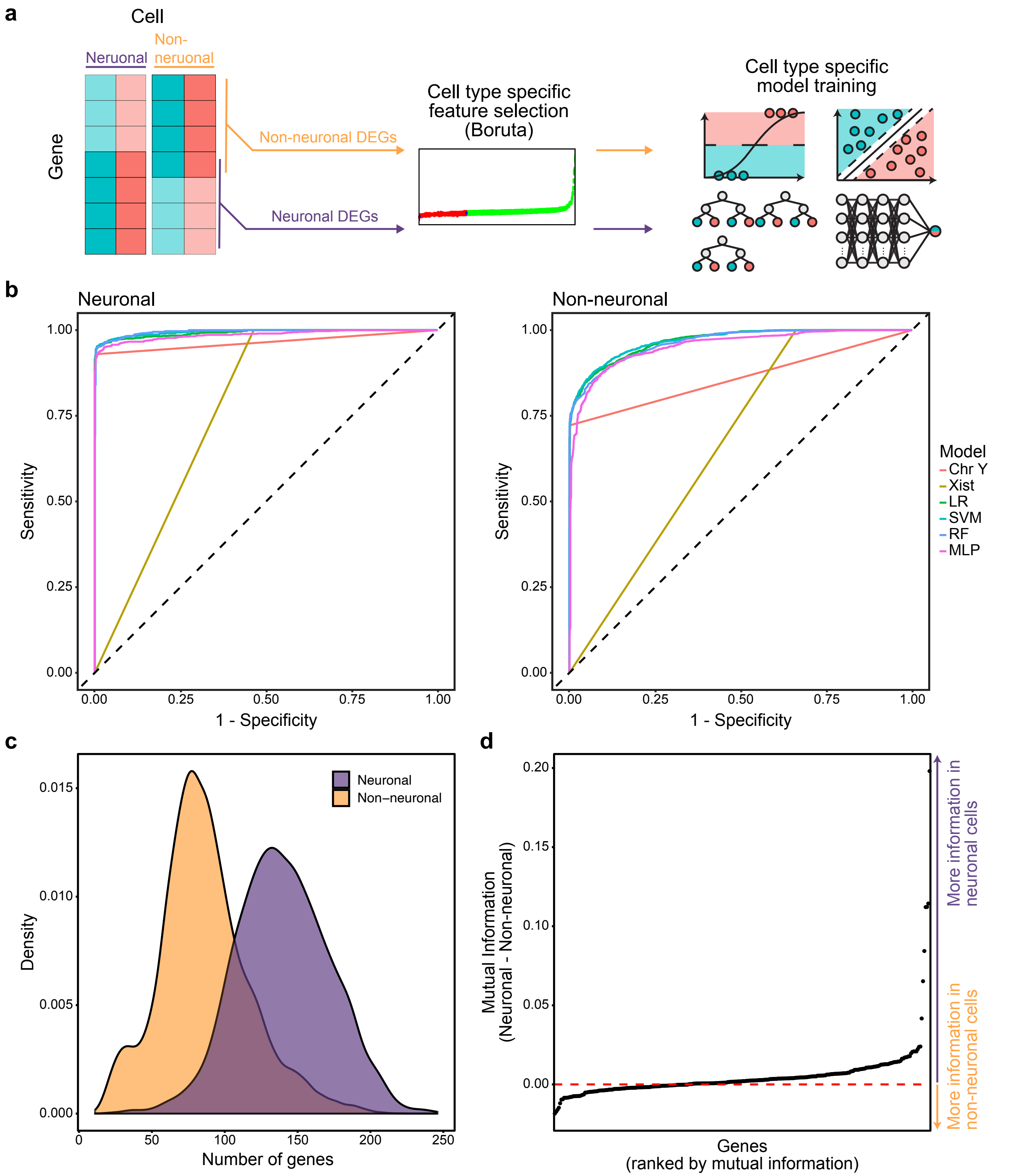


Figure 4. Cell type-specific models mirror the performance trends of models trained on all cells. **a**, Schema for cell type-specific model training strategy. **b**, Reciever operating characteristic curves of cell type-specific model sensitivity and specificity for classification of neuronal (left) and non-neuronal cells (right) of the VTA testing data partition. **c**, Distribution of the number of model predictor genes expressed in neuronal and non-neuronal cell populations of the VTA training data. **d**, Difference of mutual information score, calculated as mutual information of gene expression and sex, between neuronal and non-neuronal cells of the VTA training data partition.

overlapping. Despite differences in selected genes and training partitions, cell type-specific models performed similarly to pan-cellular models on neuronal and non-neuronal cell sex classification (**Fig. 4b, Table 3**). Differences in AUC-ROC and accuracy of neuronal and non-neuronal classification by either class-specific or pan-cellular models were within 0.002 AUC-ROC and 0.9% overall accuracy (**Table 3**). This result suggests that the disparity in cell class-specific model performance was not driven by an imbalance in the cell class composition of the training data, but rather due to inherent limitations in the ability of gene expression in non-neuronal cells to be used for cell sex classification.

Consistent with this hypothesis, we observed

Table 3. Cell type-specific classification performance. Cell sex classification of ventral tegmental area (VTA) testing partition was performed using either neuronal or non-neuronal specific models. Cell type-specific predictor variable selection and model training was conducted using respective subsets of the VTA training partition. Model performance was assessed by accuracy and AUC-ROC. Model: Name of classification model; Predictors: the set of predictor variables in the model; Neuronal: performance of neuronal-specific models measured across neuronal cells only; Non-neuronal: performance of non-neuronal models measured across only non-neuronal cell types; AUC-ROC: area under the curve of the receiver operating characteristic curve; Accuracy: proportion of correct classifications out of all classifications.

		Neuronal		Non-Neuronal	
Model	**Predictors**	**AUC-ROC**	**Accuracy**	**AUC-ROC**	**Accuracy**
Logistic Regression (LR)	55 neuronal or 232 non-neuronal selected features	0.992	0.967	0.968	0.902
Support Vector Machine (SVM)	55 neuronal or 232 non-neuronal selected features	0.994	0.970	0.969	0.904
Random Forest (RF)	55 neuronal or 232 non-neuronal selected features	0.995	0.969	0.964	0.895
Multi-Layer-Perceptron (MLP)	55 neuronal or 232 non-neuronal selected features	0.986	0.957	0.953	0.873

distinct trends in neuronal and non-neuronal sex-dependent gene expression. In addition to having fewer detected genes on average, non-neuronal cells tended to express fewer model predictor genes than non-neuronal cells (**Fig. 4c**). Furthermore, we calculated the mutual information of model gene's expression and cell sex to quantify the information gained about cell sex by observing its expression in either neuronal or non-neuronal cell types [23,24]. Model predictor gene expression was more informative of cell sex in neuronal cells for 65% of genes (**Fig. 4d**). The six genes with the largest difference in mutual information (*ENSRNOG00000060617*, *Eif2s3y*, *Kdm5d*, *ENSRNOG00000065796*, *Ddx3*) were all located on sex chromosomes, or paralogues of sex chromosome genes (*AC239701.1* with *Med14*), and conveyed more information about cell sex in neurons than non-neurons. Together, these observations underscore the inherent limitations in non-neuronal cell sex classification, as genes important for cell sex classification tend to be more sparsely detected in non-neuronal cells and their expression tends to be less informative of cell sex.

Model performance generalizes to independent data

Performance on new or unseen data, called generalizability, is a critical measure of a model's utility. To validate that models learned robust cell sex classification rules that could be applied to another dataset, we measured the AUC-ROC and accuracy of model predictions in an orthogonal snRNA-seq experiment from the rat nucleus accumbens (NAc) [25]. Unlike the VTA, the NAc is largely composed of GABAergic medium spiny neurons (in addition to cholinergic and GABAergic interneurons), yet also contains similar glial cell types [26]. This dataset contains 39,254 cells of 16 transcriptionally defined cell types from 32 rats (16M/16F) (**Fig 5a,b**). While performance estimates were lower compared to the VTA dataset, all ML models achieved high overall accuracy (0.87-0.90) and balanced sensitivity and specificity (AUC-ROC 0.94-0.97) for NAc cell sex classification (**Fig. 5b-c, Table 4**). Of all ML models, the Random Forest model achieved the best overall performance with an AUC-ROC of 0.97 and an accuracy of 0.90 (**Fig. 5b-c, Table 4**). In comparison, the non-ML models remained poorly balanced for sensitivity and specificity, as shown by AUC-ROC scores of 0.78 for Xist and 0.88 for Chr Y classifiers (**Fig. 5b, Table 4**). However, the Chr Y model matched the ML models in terms of overall accuracy (0.89), even outperforming the MLP model (**Fig. 5b, Table 4**). Together, these results demonstrate that our feature selection and training strategies did not overfit models to the VTA data set and that the resulting models are generalizable to cell sex classification of a distinct dataset from a different brain region.

DISCUSSION

Sample size and sequencing costs are practical limitations to snRNA-seq experimental design, especially when working with samples of limited starting material [4]. Pooling nuclei prior to sequencing is a popular method to overcome these limitations, but recovering individual-level data after sequencing poses additional challenges. Several methods for snRNA-seq demultiplexing have previously been proposed [5,8–10]. Our work expands the available toolkit for demultiplexing by providing a powerful, low-cost method that leverages information inherent in the transcriptome to deconvolve pooled nuclei through sex prediction with machine learning models.

The use of sex-dependent transcripts is a key advantage of our demultiplexing strategy. In contrast, barcode hashing and genotype-based strategies require additional sample preparation steps before sequencing and involve non-trivial processing to assess sample identity post-sequencing [11,27]. Sex-based sample pooling eliminates the need for additional sample preparation or the collection of other modalities. Moreover, publicly available datasets can be used to identify sex-dependent transcriptional features and fit classification models (**Fig. 1**, **Fig. 2**). Highly accurate and lightweight machine learning models, such as logistic regression, can be quickly trained and applied to new data (**Fig. 3**, **Table 1**). Furthermore, we have demonstrated that these models generalize well to new datasets from a distinct brain region, supporting their reusability across experiments (**Fig. 5**). Previous study designs with sex-balanced samples but pooled same-sex samples or different-sex samples that were not deconvolved might benefit from sex-dependent feature-based deconvolution with ML models [25,28]. Our sex-based classification approach thus complements and expands on previous strategies by leveraging orthogonal data inherent in sex-balanced snRNA-seq experiments.

Although this work directly evaluates the feasibility of sample deconvolution using sex-dependent transcriptome data, we also hypothesize that machine learning models could learn sex-

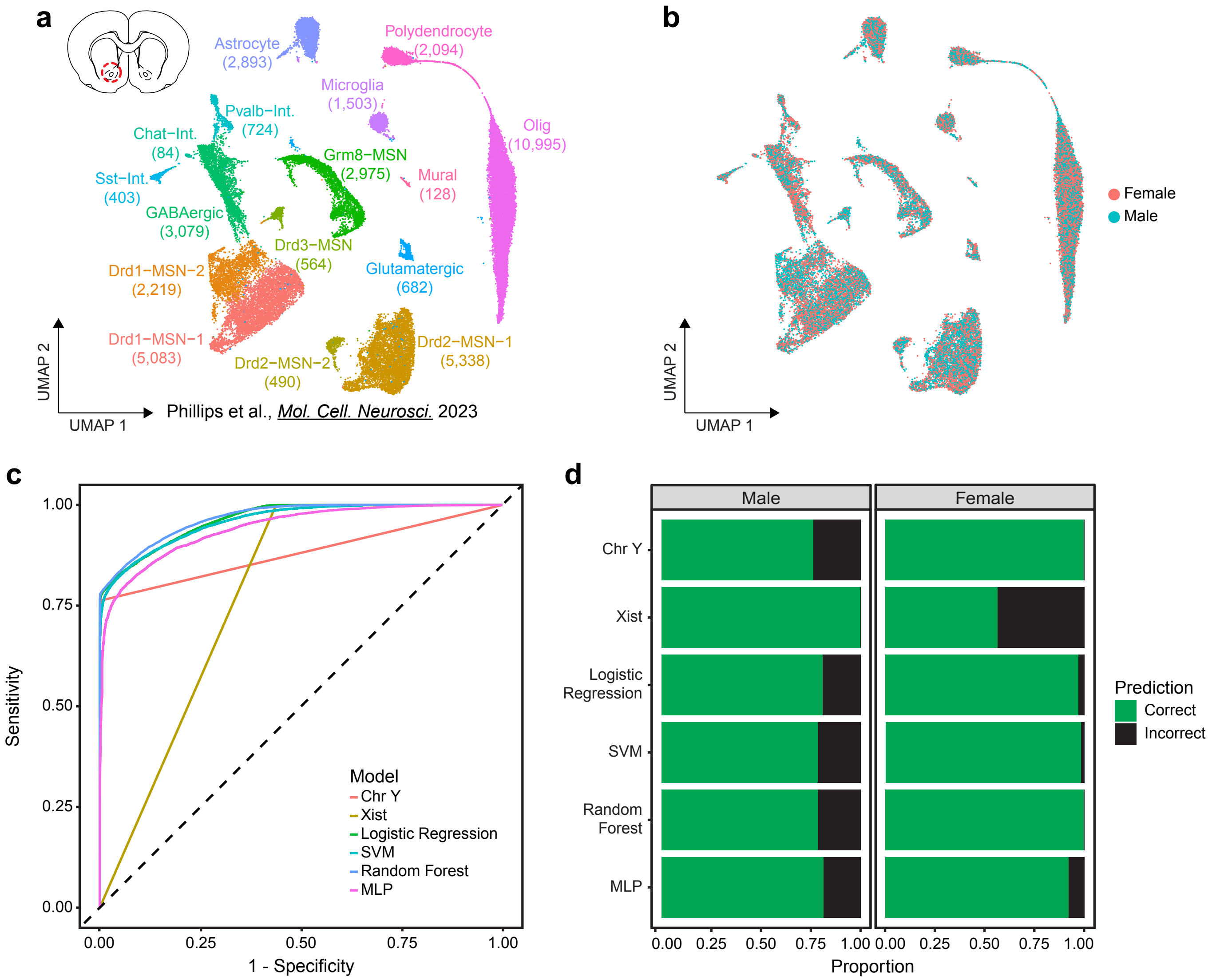


Figure 5. Sex classification models are generalizable to independent data sets. **a**, UMAP of previously published independent snRNA-seq data set of nucleus accumebns (NAc, top left diagram) samples from 16 (8 male, 8 female) rats including annotation of transcriptionally defined cell types (39,254 cells, 16 cell types). **b**, Distribution of sexes within UMAP. **c**, Reciever operating characteristic curve of VTA trained models' sensitivity and specificity for classification of NAc cells. **d**, Stacked bar chart of proportion of correct and incorrect classifications of NAc sexes.

dependent DNA accessibility features. This would enable their application in snATAC-seq experiments, expanding the potential for sex-based deconvolution in other modalities where sex differences in chromatin accessibility are observed. Additionally, this approach could be particularly valuable in multiomic datasets, such as those combining snRNA-seq and snATAC-seq, where both modalities could be used synergistically to enhance sex prediction accuracy. However, further analysis and validation would be necessary to assess the robustness of such models with snATAC-seq data. Future work in this direction would represent a meaningful extension of sex-based deconvolution to additional sequencing modalities and multiomic datasets.

Our evaluation of cell sex classification models identified poorer performance in non-neuronal cells as a potential limitation. While model performance remained above 87% accuracy, it was consistently lower than the performance observed with neuronal cells (**Fig. 4**, **Table 1**, **Table 2**). This discrepancy may be attributed to the observation that non-neuronal cells express fewer sex-predictive features, and that expressed sex-predictive features tended to convey less information than in neuronal cell populations (**Fig. 4**). Additionally, our analysis revealed a significant relationship between UMI count and accuracy across all models (**Fig. S1**), suggesting that datasets with lower UMI counts – a common feature of non-neuronal populations in other brain snRNA-seq studies — may also exhibit reduced performance in sex-based classification [25,29,30].

One potential strategy to address this limitation is the integration of sex-based pooling and deconvolution with another demultiplexing strategy, such as barcode hashing, which does not exhibit the same cell-type constraints [9]. Using multiple orthogonal methods for sample deconvolution offers a practical way to resolve ambiguous sample assignments and identify potential doublets [5]. Integration of multiple deconvolution approaches would permit future studies to take advantage of the complementary strengths of each method and mitigate their limitations [11,12].

Table 4. Nucleus accumbens classification performance. Cell sex classification of the nucleus accumbens dataset was performed using two non-ML and four ML models trained with the ventral tegmental area (VTA) training partition. Model performance was assessed by accuracy and AUC-ROC. Model: Name of classification model; Predictors: the set of predictor variables in the model; Overall: performance measured across all cells.

		Overall	
Model	**Predictors**	**AUC-ROC**	**Accuracy**
Xist	ENSRNOG00000065796 (Xist)	0.781	0.767
Chr Y	All detected Y Chromosome genes	0.881	0.886
Logistic Regression (LR)	285 Selected features	0.964	0.894
Support Vector Machine (SVM)	285 Selected features	0.960	0.889
Random Forest (RF)	285 Selected features	0.968	0.896
Multi-Layer-Perceptron (MLP)	285 Selected features	0.942	0.870

In conclusion, we demonstrate that sex-dependent transcripts can be leveraged to train accurate cell sex prediction models, supporting the feasibility of sex-based pooling and sample demultiplexing. This work represents a meaningful expansion of previously proposed strategies for sample deconvolution. Our approach does not require additional sample preprocessing or modalities, and demultiplexing can be achieved using lightweight machine-learning models. Moreover, because it does not rely on additional data modalities, our method is highly compatible with other sample deconvolution strategies allowing them to complement one another's strengths. Accurate and easy-to-implement sex-based sample deconvolution enables future works to carry out less expensive and more powerful snRNA-seq analyses.

METHODS

Computational Resource and Environment

To maintain a consistent model training and evaluation environment, all analyses were run on UAB's Cheaha research computing cluster (8 CPUs with 16 GB of RAM per CPU). All analyses used R version 4.2.0 for the x86_64-pc-linux-gnu platform, with a consistent random seed (set_seed(1234)).

snRNA-seq data for model training and evaluation

To train and evaluate models data was acquired as Seurat (version 4.3.01) objects from previous publications of snRNA-seq from the ventral tegmental area (VTA) and nucleus accumbens (NAc) [31]. Data were quality-controlled, integrated, clustered, and annotated by cell type as described in their respective publications and no further processing was applied for analyses. Objects were created using the *Rattus norvegicus* reference genome version mRatBN7.2. The genetic sex of nuclei was determined by the originating GEM (gel bead in emulsion) well used for nuclei capture, as sample nuclei of same-sex samples were pooled before capture. Analyses were conducted using RNA counts log normalized and scaled by a factor of 10,000. Training and testing partitions of the VTA data set were created with a ratio of 7:3 by randomly assigning cells to either partition based on cell type and sex, maintaining distributions between partitions.

Differential expression

Differentially expressed genes (DEGs) between male and female cells were determined for each cell type by the Wilcoxon Rank Sum test implemented in the FindMarkers() function from Seurat with default parameters. Significant DEGs were identified as those with Bonferroni-adjusted p-values < 0.05 in at least one cell type.

Boruta feature selection

The Boruta feature selection algorithm further refined the set of significant DEGs by identifying genes with significant importance for cell sex classification within the VTA training partition. Importance was determined by iteratively comparing the importance, measured as Z-scores of mean decrease accuracy measure, of real genes to a set of null "shadow features" constructed from shuffled gene counts using the Boruta() function from the Boruta package (version 8.0.0). At each iteration, genes with significantly lower scores are rejected and removed while genes with significantly higher scores are confirmed and retained. After 2000 iterations, final decisions of genes with tentative importance were made using the TentativeRoughFix() Boruta function. The final set of genes with confirmed importance was used as predictive variables for cell sex classification models.

Mutual Information

To compare the information shared between a gene's expression and cell sex between neuronal and non-neuronal cell types, the mutual information of the two variables was calculated for the two broad cell types separately. Mutual information of transcript count data and cell sex, for all selected model genes, was calculated using the mmi.pw() function from the mpmi package in R (version 0.43.2.1) [24]. For calculation of gene's expression and cell sex mutual information content in neuronal or non-neuronal populations, the VTA training partition was separated into neuronal (Glut-Neuron-1, Glut-Neuron-2, Glut-Neuron-3, GABA-Neuron-1, GABA-Neuron-2, GABA-Neuron-3, DA-Neuron) and non-neuronal (Olig-1, Olig-2, Olig-3, Astrocyte, Polydendrocyte, Microglia, OPC-Olig-1, Mural, Endothelial) populations.

Model Training

To train cell sex prediction models, models were fitted using the VTA training partition log normalized gene counts as predictor variables and cell sex as the outcome variable. The Xist model made binary female or male cell sex predictions based on the presence or absence of *Xist* (*ENSRNOG00000065796*) counts. Reciprocally, the Chromosome Y classifier made binary male or female cell sex predictions based on the presence or absence of counts from any gene on the Y chromosome. The logistic regression model was fit to the training data with the glm() function from the stats package (version 4.2.0). To assess training for a range of hyperparameters the random forest, support vector machine, and multilayer perceptron models were trained with the train() function of the caret package (version 6.0-94) [32]. Training parameters were set to output cell sex class probabilities and perform 3x repeated 10-fold cross-validation using trainControl() with parameters method = "repeatedcv", number = 10, repeats = 3, classProbs = T, allowParallel = T. Ranges for model hyperparameters for each model were specified using tuning grids defined for each model, described in detail below. Model accuracy was used to select the optimal configuration for each model. Random forest: The random forest model was trained with method = "rf" and a constant number of 1000 trees ntrees = 1000 as train() function parameters. To tune the number of variables randomly sampled at each split, the custom tuning grid assessed a range of "mtry" values from 2-334 by steps of 12. The final optimal model selected for accuracy used ntree = 1000, and mtry = 34. Support vector machine: The support vector machine (SVM) model was trained with method = "svmRadial" as a parameter of the train() function, to specify an SVM model with a radial basis function. A custom tuning grid of hyperparameters specifying "sigma" values from 0.0001 to 1, where steps increase by a factor of 10 with each subsequent term, and C values from 0.1 to 10 (0.01,0.1,0.2,0.5,1,1.5,2,5,10), was utilized for model training. The final optimal model selected for accuracy used sigma = 1e-3 and C = 5. Multilayer perceptron: A multi-layer perceptron with weight decay was trained with method = "mlpWeightDecayML" as train() function parameters. Model hyperparameters for training were tuned using a custom tuning grid specifying ranges for the number of nodes for each layer 1 (1,5,10,15,20), layer 2 (0,2,5,8,10), layer 3 (0,1,2,4,5), and weight decay `decay` values (0,0.05,0.1,0.15,0.2). Only hyperparameter configurations with decreasing nodes in successive layers were evaluated. The final optimal model selected for accuracy was trained using layer1 = 20, layer2 = 2, layer3 = 0, and decay = 0. Neuronal/Non-neuronal models: Before fitting cell type-specific models, the VTA training data partition was split into subsets for either neuronal (Glut-Neuron-1, Glut-Neuron-2, Glut-Neuron-3, GABA-Neuron-1, GABA-Neuron-2, GABA-Neuron-3, DA-Neuron) or non-neuronal (Olig-1, Olig-2, Olig-3, Astrocyte, Polydendrocyte, Microglia, OPC-Olig-1, Mural, Endothelial) cell type groups. Feature selection with Boruta was repeated with significant DEGs within each cell type subset. LR, RF, SVM, and MLP models were fit as described above for neuronal and non-neuronal cell type subsets.

Model evaluation

Model performance was evaluated using the VTA testing partition and the NAc dataset. Neuronal and non-neuronal models were evaluated using respective subsets of the VTA testing data partition. Overall model classification accuracy was calculated as the number of correct classifications divided by the number of total classifications made. For all models, the receiver operating characteristic (ROC) curve and the area under the ROC (AUC-ROC) were used to evaluate the trade-off of sensitivity and specificity rates. ROC curve and AUC values were calculated using the roc() function of the pROC package in R. True-positive and false-positive rates were calculated using males as the "positive" class.

DATA AVAILABILITY

All relevant data that support the findings of this study are available by request from the corresponding author (J.J.D.). Sequencing data that support the findings of this study are available in Gene Expression Omnibus. Accession numbers of specific datasets are outlined below.
Ventral tegmental area snRNA-seq VTA: GSE168156
Nucleus accumbens snRNA-seq: GSE137763, GSE222418

Custom code can be found at https://github.com/Jeremy-Day-Lab/Twa_etal_2024

ACKNOWLEDGEMENTS

We thank all current and former Day Lab members for their assistance and support. This work was supported by NIH grants DP1DA039650, R01MH114990, R01DA053743, and R01DA054714 and the McKnight Foundation Neurobiology of Brain Disorders Award (JJD).

AUTHOR CONTRIBUTIONS

Conceptualization: RAP, JJD
Methodology: RAP, GT, NJR
Data Curation: GT, RAP
Formal Analysis: GT
Visualization: GT
Supervision: JJD
Funding acquisition: JJD
Writing – original draft: GT
Writing – review & editing: GT, RAP, NJR, JJD

CONFLICTS OF INTEREST

The authors declare no competing interests, financial or otherwise.

REFERENCES

1. Siletti, K., Hodge, R., Mossi Albiach, A., Lee, K.W., Ding, S.-L., Hu, L., Lönnerberg, P., Bakken, T., Casper, T., Clark, M., *et al.* (2023). Transcriptomic diversity of cell types across the adult human brain. Science *382*, eadd7046.

2. Li, Y.E., Preissl, S., Miller, M., Johnson, N.D., Wang, Z., Jiao, H., Zhu, C., Wang, Z., Xie, Y., Poirion, O., *et al.* (2023). A comparative atlas of single-cell chromatin accessibility in the human brain. Science *382*, eadf7044.

3. Tian, W., Zhou, J., Bartlett, A., Zeng, Q., Liu, H., Castanon, R.G., Kenworthy, M., Altshul,

J., Valadon, C., Aldridge, A., *et al.* (2023). Single-cell DNA methylation and 3D genome architecture in the human brain. Science *382*, eadf5357.

4. Schmid, K.T., Höllbacher, B., Cruceanu, C., Böttcher, A., Lickert, H., Binder, E.B., Theis, F.J., and Heinig, M. (2021). scPower accelerates and optimizes the design of multi-sample single cell transcriptomic studies. Nat. Commun. *12*, 6625.

5. Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B.Z., Mauck, W.M., Smibert, P., and Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. Genome Biol. *19*, 224.

6. Cost Per Cell | Satija Lab Available at: https://satijalab.org/costpercell/ [Accessed June 28, 2024].

7. Squair, J.W., Gautier, M., Kathe, C., Anderson, M.A., James, N.D., Hutson, T.H., Hudelle, R., Qaiser, T., Matson, K.J.E., Barraud, Q., *et al.* (2021). Confronting false discoveries in single-cell differential expression. Nat. Commun. *12*, 5692.

8. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., *et al.* (2018). Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat. Biotechnol. *36*, 89–94.

9. Gaublomme, J.T., Li, B., McCabe, C., Knecht, A., Yang, Y., Drokhlyansky, E., Van Wittenberghe, N., Waldman, J., Dionne, D., Nguyen, L., *et al.* (2019). Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. Nat. Commun. *10*, 2907.

10. Howitt, G., Feng, Y., Tobar, L., Vassiliadis, D., Hickey, P., Dawson, M.A., Ranganathan, S., Shanthikumar, S., Neeland, M., Maksimovic, J., *et al.* (2023). Benchmarking single-cell hashtag oligo demultiplexing methods. NAR Genom. Bioinform. *5*, lqad086.

11. Curion, F., Wu, X., Heumos, L., André, M.M.G., Halle, L., Ozols, M., Grant-Peters, M., Rich-Griffin, C., Yeung, H.-Y., Dendrou, C.A., *et al.* (2024). hadge: a comprehensive pipeline for donor deconvolution in single-cell studies. Genome Biol. *25*, 109.

12. Li, L., Sun, J., Fu, Y., Changrob, S., McGrath, J.J.C., and Wilson, P.C. (2024). A hybrid demultiplexing strategy that improves performance and robustness of cell hashing. Brief. Bioinformatics *25*.

13. Phillips, R.A., Tuscher, J.J., Black, S.L., Andraka, E., Fitzgerald, N.D., Ianov, L., and Day, J.J. (2022). An atlas of transcriptionally defined cell populations in the rat ventral tegmental area. Cell Rep. *39*, 110616.

14. Kursa, M.B., and Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. J. Stat. Softw. *36*.

15. Cox, D.R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological) *20*, 215–232.

16. Breiman, L. (2001). Random Forests. Springer Science and Business Media LLC *45*, 5–32.

17. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Mach. Learn. *20*, 273–297.

18. Rosenblatt, F. (1961). Principles of neurodynamics. perceptrons and the theory of brain mechanisms (US Dept of the Navy).

19. Amari, S. (1967). A theory of adaptive pattern classifiers. IEEE Trans. Electron. Comput. *EC-16*, 299–307.

20. Werbos, P.J. (1982). Applications of advances in nonlinear sensitivity analysis. In System modeling and optimization, R. F. Drenick and F. Kozin, eds. (Berlin/Heidelberg: Springer-Verlag), pp. 762–770.

21. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). Learning representations by back-propagating errors. Nature *323*, 533–536.

22. Linnainmaa, S. (1976). Taylor expansion of the accumulated rounding error. BIT *16*, 146–160.

23. Shannon, C.E. (1948). A mathematical theory of communication. Bell System Technical Journal *27*, 379–423.

24. Pardy, C., Galbraith, S., and Wilson, S.R. (2018). Integrative exploration of large high-dimensional datasets. Ann. Appl. Stat. *12*, 178–199.

25. Phillips, R.A., Tuscher, J.J., Fitzgerald, N.D., Wan, E., Zipperly, M.E., Duke, C.G., Ianov, L., and Day, J.J. (2023). Distinct subpopulations of D1 medium spiny neurons exhibit unique transcriptional responsiveness to cocaine. Mol. Cell. Neurosci. *125*, 103849.

26. Scofield, M.D., Heinsbroek, J.A., Gipson, C.D., Kupchik, Y.M., Spencer, S., Smith, A.C.W., Roberts-Wolfe, D., and Kalivas, P.W. (2016). The Nucleus Accumbens: Mechanisms of Addiction across Drug Classes Reflect the Importance of Glutamate Homeostasis. Pharmacol. Rev. *68*, 816–871.

27. Sayed, M., Wang, Y.J., and Lim, H.-W. (2024). Systematic benchmark of single-cell hashtag demultiplexing approaches reveals robust performance of a clustering-based method. Brief. Funct. Genomics.

28. Simon, R.C., Loveless, M.C., Yee, J.X., Goh, B., Cho, S.G., Nasir, Z., Hashikawa, K., Stuber, G.D., Zweifel, L.S., and Soden, M.E. (2024). Opto-seq reveals input-specific immediate-early gene induction in ventral tegmental area cell types. Neuron *112*, 2721-2731.e5.

29. Phan, B.N., Ray, M.H., Xue, X., Fu, C., Fenster, R.J., Kohut, S.J., Bergman, J., Haber, S.N., McCullough, K.M., Fish, M.K., *et al.* (2024). Single nuclei transcriptomics in human and non-human primate striatum in opioid use disorder. Nat. Commun. *15*, 878.

30. Ottenheimer, D.J., Simon, R.C., Burke, C.T., Bowen, A.J., Ferguson, S.M., and Stuber, G.D. (2024). Single-cell sequencing of rodent ventral

pallidum reveals diverse neuronal subtypes with non-canonical interregional continuity. BioRxiv.

31. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., *et al.* (2021). Integrated analysis of multimodal single-cell data. Cell *184*, 3573-3587.

32. Kuhn, M. (2008). Building Predictive Models in *R* Using the caret Package. J. Stat. Softw. *28*.

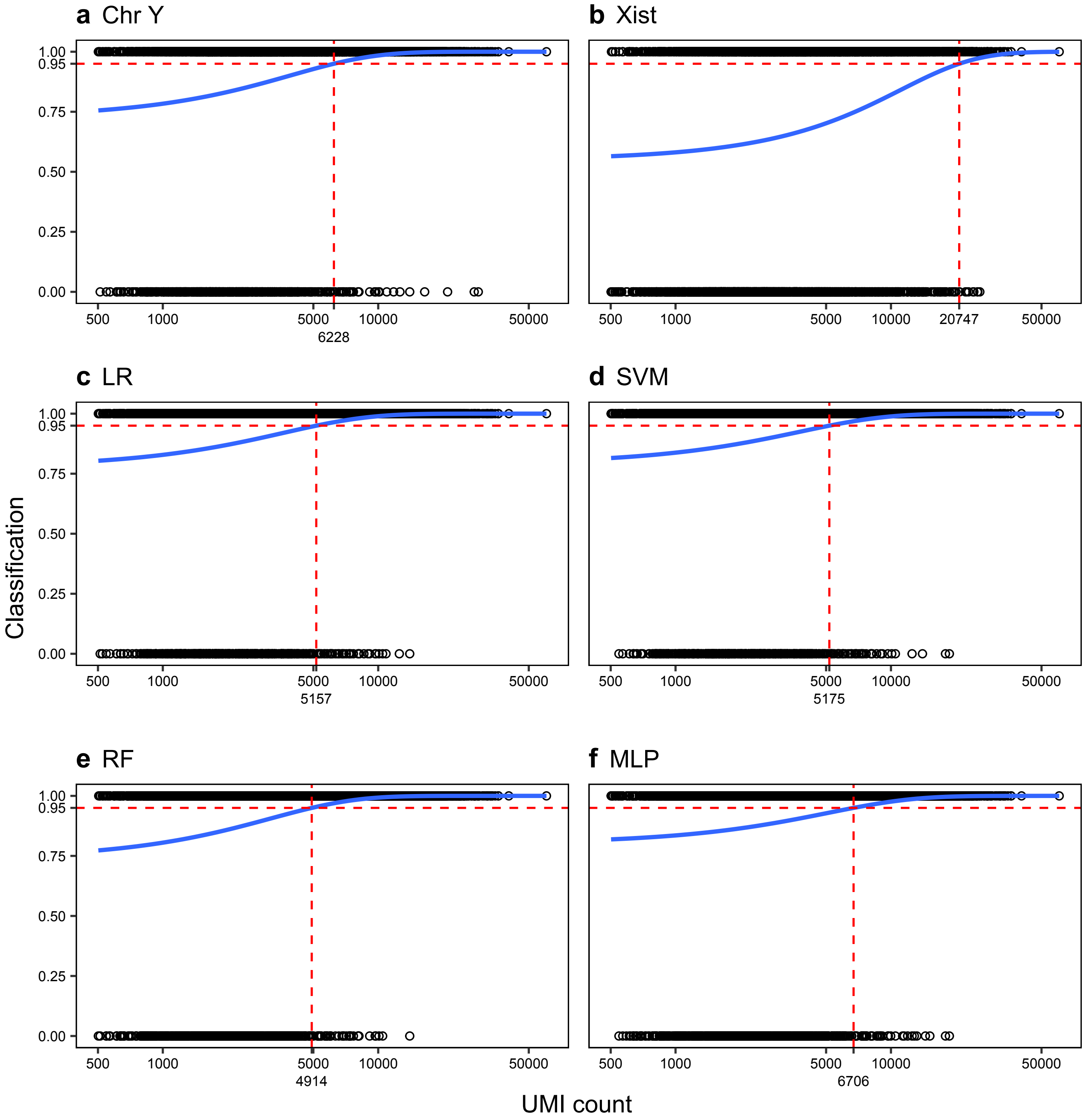


Figure S1. Increased RNA count per cell increases probability of correct classification for all models. Logistic regression of ventral tegmental area test partition cell UMI (unique molecular index) count and model classification (correct: 1, incorrect: 0) for chromosome Y (**a**), Xist (**b**), logistic regression (**c**), support vector machine (**d**), random forest (**e**), and multilayer-perceptron (**f**) models. Dashed red lines indicate the RNA count corresponding to a 95% probability of correct classification by a model.

Table S1. Increased UMI count improves the likelihood of accurate cell sex classification. The relationship between transcript UMI (unique molecular index) count and accuracy of model classification (incorrect: 0, correct: 1) in the ventral tegmental area test partition data was modeled using logistic regression for all cell sex classification models. Model: model classifications used to fit logistic regression; Intercept: fit logistic regression model intercept; UMI: fit logistic regression model coefficient for UMI count; Estimate: model estimate for either intercept or UMI terms; StdErr: standard error of either Intercept or UMI terms; Z-score: Estimate / Std Error; pval: p-value associated with the value Z-score column; 95% Probability Intercept: UMI value of where probability of correct classification reaches 95% as predicted by fit model.

	Intercept				UMI count				
Model	**Estimate**	**StdErr**	**Z-score**	**pval**	**Estimate**	**StdErr**	**Z-score**	**pval**	**95% Probability Intercept**
Chr Y	0.969	0.074	13.121	2.49E-39	3.17E-04	2.21E-05	14.364	8.68E-47	6228
Xist	0.194	0.044	4.462	8.14E-06	1.33E-04	8.48E-06	15.635	4.20E-55	20747
LR	1.245	0.083	14.919	2.49E-50	3.29E-04	2.58E-05	12.770	2.43E-37	5157
SVM	1.328	0.084	15.852	1.36E-56	3.12E-04	2.54E-05	12.307	8.26E-35	5175
RF	1.029	0.085	12.161	5.02E-34	3.90E-04	2.79E-05	13.964	2.59E-44	4914
MLP	1.393	0.074	18.747	2.04E-78	2.31E-04	1.98E-05	11.662	2.00E-31	6706