## Cell type prediction

Ahmed Mahfouz

Department of Human Genetics, Leiden University Medical Center Pattern Recognition and Bioinformatics, TU Delft



mahfouzlab.org







# Comparing cells across studies/species/conditions...



1165 genes

Bakken et al. bioRxiv 2020

Several marker genes are species-specific



Hodge, Bakken et al. Nature 2019 Keefe & Nowakowski, Nature 2019

## Cell identity



Morris Development 2019

### How can we identify cell populations?





### How can we identify cell populations?





### Cytosplore Transcriptomics



Abdelaal et al., bioRxiv 2020

https://transcriptomics.cytosplore.org/

### Unsupervised cell identification is problematic



#### Not reproducible



#### Subjective







#### Clustering

- Unsupervised learning
- Discovering structure/relations
- Clusters are defined by a decision boundary



#### Classification

- Supervised learning
- Prior information available about different groups
- Classifiers find descriptions of decision boundaries

### Classification







## Classifier training

- Dataset: for *j* <sup>th</sup> cell:
  - gene expressions **x**<sub>i</sub>
  - class label:  $y_j \in \{1=T, -1=B\}$
- Classifier:  $\hat{y}_j = W(x_j)$

• Errors: 
$$E = \operatorname{sum}(E_j)$$
  $E_j = \begin{cases} 1 & \text{if } \hat{y}_j \neq y_j \\ 0 & \text{if } \hat{y}_j = y_j \end{cases}$ 

(



• Place decision boundary (i.e. change W) s.t. E is minimal

### Instance Based Learning (Lazy Classification)

- Example: Nearest neighbor (k-NN)
- Keep the whole training dataset
- A query example (vector) comes
- Find closest example(s)
- Predict
- No actual training



### Nearest Neighbor (k-NN)

- To make Nearest Neighbor work we need 4 things:
- 1) Distance metric:
- 2) How many neighbors to look at?
- 3) Weighting function (optional)
- 4) How to fit with the local points?



## Nearest Neighbor (k-NN)

- Distance metric:
  - Euclidean
- How many neighbors to look at?
  - k
- Weighting function (optional):
  - Unused
- How to fit with the local points?
  - Predict the average output among k nearest neighbors



### Effect of k



## Weighted Nearest Neighbor

(kernel regression)

- Distance metric:
  - Euclidean
- How many neighbors to look at?
  - All of them (!)
- Weighting function:

• 
$$w_i = \exp(-\frac{d(x_i,q)^2}{K_w})$$

- Nearby points to query q are weighted more strongly.  $K_w$ :kernel width.
- How to fit with the local points?
  - Predict weighted average:  $\frac{\sum_{i} w_{i} y_{i}}{\sum_{i} w_{i}}$



$$d(x_i, q) = 0$$

### Comparison: K=1, K=2, kernel



### Support Vector Machine (SVM)



### Support Vector Machine (SVM)

Which boundary is better?



### Support Vector Machine (SVM)

Which boundary is better?

The one that maximizes the margins from both labels.

i.e. The one whose distance to the nearest element of each label is the largest.











### 16 existing classifiers (April 2019)



## 16 existing classifiers (April 2019)



### 16 existing + 6 off-the-shelf classifiers



### Experiment 1: intra-dataset evaluation

• Stratified 5-fold cross validation

- Performance evaluation
  - Median F1-score:  $F1 = 2 \frac{precision.recall}{precision+recall}$
  - % unlabelled cells



### Most classifiers work well

#### Median F1-score



### Most classifiers work well

#### Median F1-score



#### % Unlabeled

		F	CellBench				
SVM <sub>rejection</sub> -	2.3	1.5	1.6	1.9	0	0	0
scPred-	6.7	10.8	8.5	10	11.1	0.4	1.1
SVM-	0	0	0	0	0	0	0
singleCellNet-	0.1	0	0	0	0	0	0
ACTINN-	0	0	0	0	0	0	0
CaSTLe-	0	0	0	0	0	0	0
scmapcell-	5.8	4.2	3.8	6.4	8.6	0	0
LDA-	0	0	0	0	0	0	0
scmapcluster-	14.6	7.9	1.1	3.6	4	0	0.2
RF-	0	0	0	0	0	0	0
SingleR -	0	0	0	0	0	0	0
LAmbDA-	0	0	0	0	0	0	0
NMC-	0	0	0	0	0	0	0
CHETAH-	0.5	0.5	0.9	1.1	0.6	0.1	0
scVI-	0	0	0	0	0	0	0
scID-	23.6	8.3	17.3	32.1	0.2	24.2	9.8
Cell_BLAST-	20.3	3.2	19.6	23.1	4.1	0.1	68.1
kNN-	0	0	0	0	0	0	0
SCINA-							
DigitalCellSorter-							
Garnett <sub>CV</sub> -							
Garnett <sub>pretrained</sub> -							
Moana-							
Garnett <sub>DE</sub> -							
SCINA <sub>DE</sub> -							
DigitalCellSorter <sub>DE</sub> -							
	Baron Mouse-	Baron Human-	Muraro-	Segerstolpe-	Xin-	10X-	CEL-Seq2-
			Unial	beled ('	%)		
	0		25	50	75	1	00

### Performance drops with deeper annotation

Median F1-score





% Unlabeled





34

### Performance drops with deeper annotation

Median F1-score





% Unlabeled



Unlabeled (%) 25 50 75 100

### Trade-off between high performance and rejecting cells Median F1-score % Unlabeled



100





Median F1-score

### Prior knowledge is not beneficial

1

% Unlabeled



•

Lower

classes!

#### Median F1-score PBMC SVM<sub>rejection</sub> 0.99 0.92 scPred 0.96 SVM 0.95 0.7



### Off-the-shelf SVM outperforms dedicated single cell classifiers Median F1-score % Unlabeled

		Pancreas		CellBench TM		Allen Mouse Brain		PBMC						
	SVM <sub>rejection</sub> -	0.99	0.99	0.98	1	0.98	1	1	0.99	1	1	0.98	0.99	0.92
	scPred-	1	0.98	0.98	1	0.95	1	1	0.97	1	1	0.69	0.96	
	SVM-	0.98	0.98	0.97	1	0.99	1	1	0.98	1	0.99	0.89	0.95	0.7
	singleCellNet-	0.97	0.96	0.97	0.99	1	1	1	0.94	1	0.99	0.87	0.88	0.74
	ACTINN-	0.97	0.98	0.97	1	0.95	1	1	0.97	1	0.99	0.86	0.88	0.74
	CaSTLe-	0.93	0.94	0.96	0.98	0.96	1	0.99	0.94	1	0.99	0.79	0.84	0.79
	scmapcell-	0.98	0.98	0.97	1	0.73	1	1	0.98	1	1	0.91	0.73	0.64
	LDA-	0.94	0.97	0.96	0.99	0.89	1	1	0.95	1	0.99	0.88	0.63	0.66
	scmapcluster-	0.99	0.95	0.97	1	1	1	1	0.87	1	0.98	0.88	0.73	0.44
	RF-	0.94	0.94	0.96	0.98	0.85	1	1	0.91	1	0.99	0.73	0.81	0.66
	SingleR -	0.96	0.97	0.95	0.97	0.99	1	1	0.88	1	0.97	0.86	0.66	0.32
	LAmbDA-	0.92	0.8	0.95	0.96	0.97	1	1	0.62	1	0.99	0.84		0.4
	NMC-	0.92	0.91	0.84	0.93	0.99	0.92	0.9	0.69	0.99	0.97	0.81	0.71	0.55
	CHETAH-	0.91	0.94	0.96	0.97	0.96	1	1	0.83	1	0.96	0.81	0.65	0.11
	scVI-	0.98	0.56	0.97	0.99	1	1	1	0	1	0.97	0	0.97	0.64
	scID-	0.75	0.59	0.95	0.85	0.8	1	1	0.42	1	0.95	0.63	0.61	0.42
	Cell_BLAST-	0.11	0.89	0.79	0.08	0.63	1	0.99	0.97	1	0.99	0.76	0.91	0.74
	kNN-	0.91	0.95	0.95	0.85	0.03	1	0.98	0.92	1	0.64	0.13	0.45	0.54
	SCINA-												1*	1*
[	DigitalCellSorter-												0.99*	0.78*
	Garnett <sub>CV</sub> -												0.94*	0.6*
	Garnett <sub>pretrained</sub> -												0.98*	0.54*
	Moana-												0.93*	0.5*
	Garnett <sub>DE</sub> -												0.65	0.37
	SCINA <sub>DE</sub> -												0.38	0.47
Dig	italCellSorter <sub>DE</sub> -												0	0
		Baron Mouse-	Baron Human-	Muraro-	Segerstolpe-	- vix	-V01	CEL-Seq2-	score	AMB3-	AMB16-	AMB92-	Zheng sorted -	Zheng 68K-
						0	0.25	0.5	0.7	5	1			

#### Pancreas CellBench TΜ Allen Mouse Brain PBMC SVM<sub>rejection</sub>-23.5 61.8 0.4 scPred 10.8 61.9 SVMsingleCellNet ACTINN-CaSTLescmapcell-8.6 58.2 70.2 LDA-20.2 scmapcluster-RF SingleR-LAmbDA-0 NMC-10.9 CHETAHscVI 0.2 scID-Cell BLAST 68.1 kNN-0 SCINA DigitalCellSorter Garnett<sub>CV</sub> 70 Garnettpretrained 55.2 Moana Garnett 50.9 SCINA<sub>DE</sub> DigitalCellSorter Zheng 68K-Seq2-AMB3-AMB16-AMB92sorted-Muraro rstolpe Xin 10X Σ Baron Mouse **Baron Human** Zheng Ы (n Unlabeled (%)

25

75

50

100

### Performance depends on dataset complexity



### Experiment 2: inter-dataset evaluation

- Train on one dataset, evaluate on another
- More realistic scenario
- More challenging, data is not aligned



### Experiment 2: inter-dataset evaluation



Jiarui Ding et al. Nature Biotechnology 2020



Training set



SM2 -SM2 -CL -DR -ID -SW -



SM2 -10XV3 -CL -DR iD -SW -SW -10XV2 - Training set





Lest set













### Experiment 3: rejection evaluation





### **Experiment 3: rejection evaluation**



## Performance Summary



53

### Conclusions so far

- Simple, off-the-shelf classifiers outperform dedicated single cell methods (see also Köhler et al. bioRxiv 2019)
- Prior-knowledge does not improve performance (highly dependent on selected markers)
- Rejection is difficult
- SnakeMake pipeline: <u>https://github.com/tabdelaal/scRNAseq\_Benchmark/</u>

Abdelaal\*, Michielsen\* et al. Genome Biology 2019

## Still, challenges remain

- Incomplete/missing reference atlas
- Inconsistent labels across datasets
- Sharing data is an issue (scArches, Lotfollahi et a;. bioRxiv 2020)



### Hierarchical Progressive Learning



### Hierarchical Progressive Learning



### Hierarchical Progressive Learning



### Tree construction

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS
CD19+ B	812	676	2,000
Monocytes (MC)		1,194	
CD14+	2,081		2,000
CD16+	274		
CD4+ T	13,523	1,458	
Reg.			2,000
Naive			2,000
Memory			2,000
CD8+ T	4,195	2,128	
Naive			2,000
Megakaryocyte (MK)	142	433	
NK cell		429	2,000
CD56+ bright	355		
CD56+ dim	2,415		
Dendritic			
Plasmacytoid (pDC)	101		
Myeloid (mDC)	455		
CD34+			2,000



### Tree construction

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS
CD19+ B	812	676	2,000
Monocytes (MC)		1,194	
CD14+	2,081		2,000
CD16+	274		
CD4+ T	13,523	1,458	
Reg.			2,000
Naive			2,000
Memory			2,000
CD8+ T	4,195	2,128	
Naive			2,000
Megakaryocyte (MK)	142	433	
NK cell		429	2,000
CD56+ bright	355		
CD56+ dim	2,415		
Dendritic			
Plasmacytoid (pDC)	101		
Myeloid (mDC)	455		
CD34+			2,000



### Tree construction

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS
CD19+ B	812	676	2,000
Monocytes (MC)		1,194	
CD14+	2,081		2,000
CD16+	274		
CD4+ T	13,523	1,458	
Reg.			2,000
Naive			2,000
Memory			2,000
CD8+ T	4,195	2,128	
Naive			2,000
Megakaryocyte (MK)	142	433	
NK cell		429	2,000
CD56+ bright	355		
CD56+ dim	2,415		
Dendritic			
Plasmacytoid (pDC)	101		
Myeloid (mDC)	455		
CD34+			2,000





### Classification performance

Cell population	Batch1 eQTL	Batch2 Bench 10Xv2	Batch3 FACS	Testset Bench 10Xv3
CD19+ B	812	676	2,000	346
Monocytes (MC)		1,194		
CD14+	2,081		2,000	354
CD16+	274			98
CD4+ T	13,523	1,458		960
Reg.			2,000	
Naive			2,000	
Memory			2,000	
CD8+ T	4,195	2,128		962
Naive			2,000	
Megakaryocyte (MK)	142	433		270
NK cell		429	2,000	194
CD56+ bright	355			
CD56+ dim	2,415			
Dendritic				38
Plasmacytoid (pDC)	101			
Myeloid (mDC)	455			
CD34+			2,000	



### Summary

- Cell identification is moving from unsupervised (clustering/visualization) to supervised (classification) learning
- Comprehensive benchmark of classifiers for single-cell RNA-seq data helps both users and developers
- Continuous learning from a growing reference atlas by combining multiple annotated datasets into a hierarchical classifier (scHPL)









