CellBender `remove-background` report

This output report from `cellbender remove-background` contains a summary of the run, including counts remaining, counts removed, further analyses, and any warnings or suggestions if the run seems to be abnormal.

This HTML report is created from a jupyter notebook at

`cellbender/cellbender/remove-background/report.ipynb`

within the CellBender codebase. Feel free to run the notebook yourself and make any changes you see fit, or use it as a starting point for further analyses.

The commentary in this report is generated using automated heuristics and best guesses based on hundreds of real datasets. If any of the automated commentary in this report seems incorrect for your dataset, please submit a question or an issue at our github repository https://github.com/broadinstitute/CellBender

Cellarium Lab .. Methods Group .. Data Sciences Platform .. Broad Institute

---

Input and output files

(Modify this section if you run this notebook yourself.)

```
Input file: /raven/u/oknight/data/INT2a_DOGMA/outs/raw_feature_bc_matrix.h5
Output file: /raven/u/oknight/data/INT2a_DOGMA/outs/cellbender_output.h5
```

Report

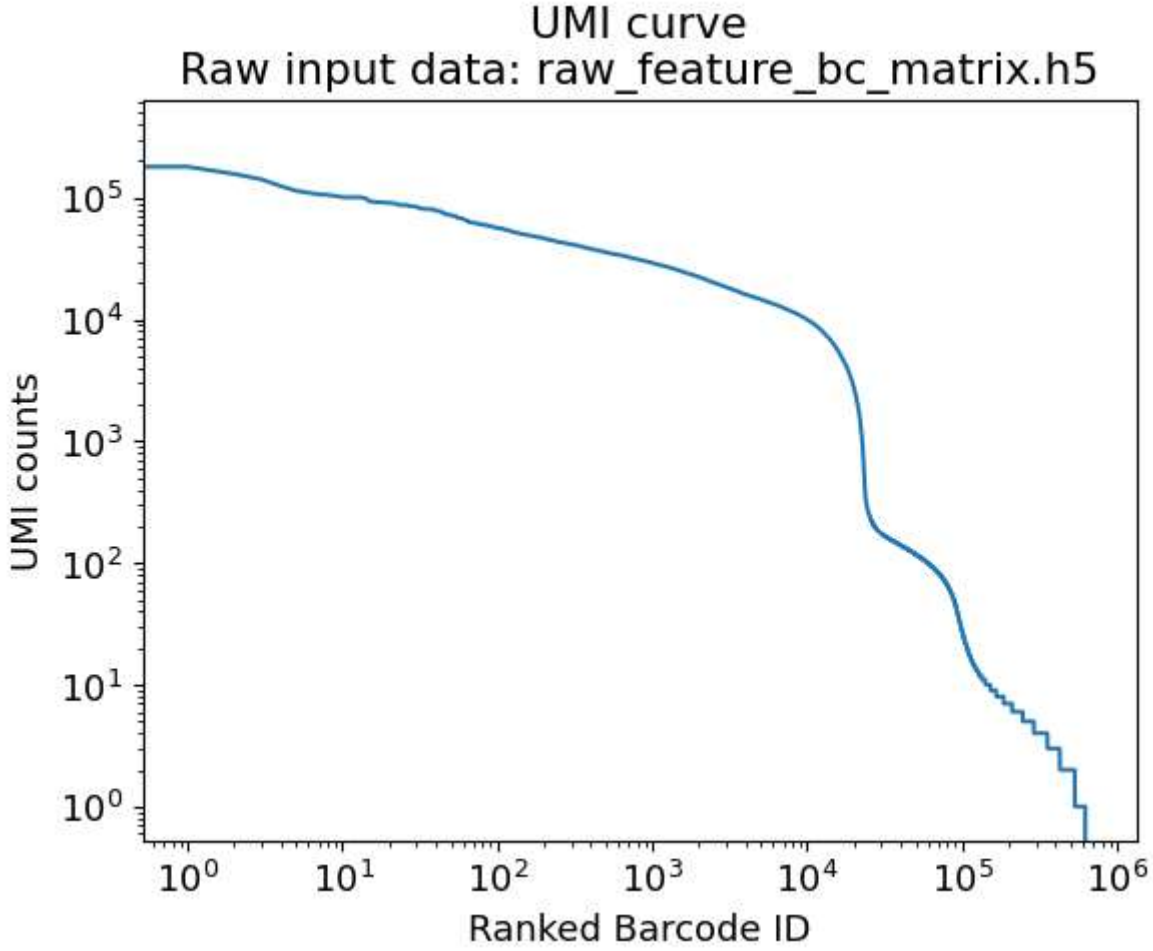
CellBender version 0.3.0

2023-04-29 02:12:26

cellbender_output.h5

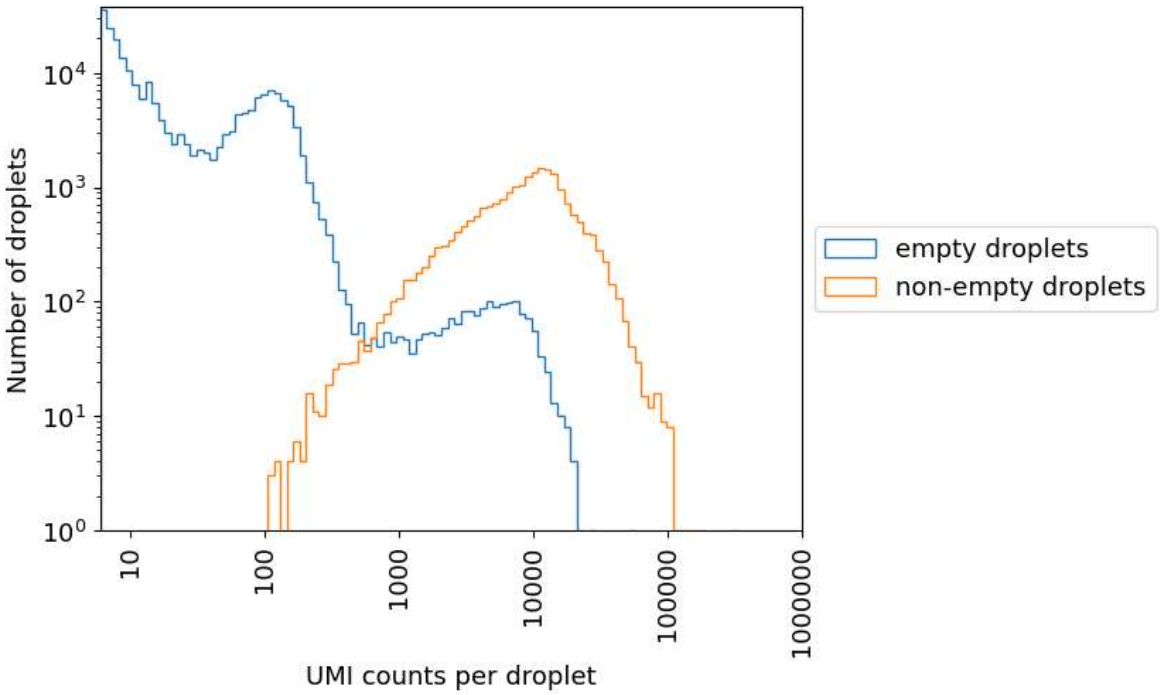
Loaded dataset

```
AnnData object with n_obs × n_vars = 28603 × 162756
    obs: 'background_fraction', 'cell_probability', 'cell_size', 'droplet_efficie
ncy', 'n_raw', 'n_cellbender'
    var: 'ambient_expression', 'feature_type', 'genome', 'gene_id', 'cellbender_a
nalyzed', 'n_raw', 'n_cellbender'
    uns: 'cell_size_lognormal_std', 'empty_droplet_size_lognormal_loc', 'empty_dr
oplet_size_lognormal_scale', 'swapping_fraction_dist_params', 'estimator', 'featu
res_analyzed_inds', 'fraction_data_used_for_testing', 'learning_curve_learning_ra
te_epoch', 'learning_curve_learning_rate_value', 'learning_curve_test_elbo', 'lea
rning_curve_test_epoch', 'learning_curve_train_elbo', 'learning_curve_train_epoc
h', 'target_false_positive_rate'
    obsm: 'cellbender_embedding'
    layers: 'raw', 'cellbender'
```



Examine how many counts were removed in total

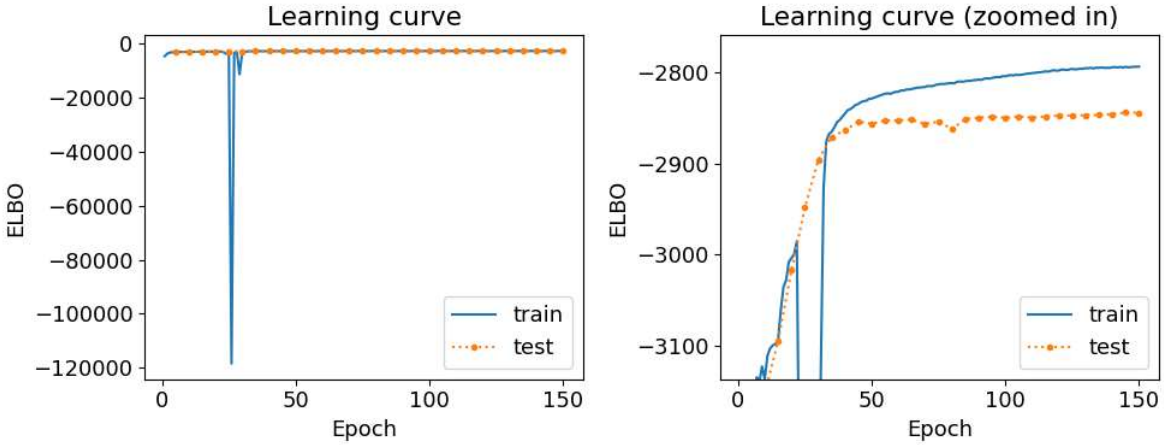
```
removed 1998447 counts from non-empty droplets
removed 0.84% of the counts in non-empty droplets
```




```
Rough estimate of expectations based on nothing but the plot above:
roughly 1597503 noise counts should be in non-empty droplets
that is approximately 0.67% of the counts in non-empty droplets
with a false positive rate [FPR] of 1.0%, we would expect to remove about 1.67% o
f the counts in non-empty droplets
```

It looks like the algorithm did a decent job meeting that expectation.

Assessing convergence of the algorithm



The learning curve tells us about the progress of the algorithm in inferring all the latent variables in our model. We want to see the ELBO increasing as training epochs increase. Generally it is desirable for the ELBO to converge at some high plateau, and be fairly stable.

What to watch out for:

1. large downward spikes in the ELBO (of value more than a few hundred) 2. the test ELBO can be smaller than the train ELBO, but generally we want to see both curves increasing and reaching a stable plateau. We do not want the test ELBO to dip way back down at the end. 3. lack of convergence, where it looks like the ELBO would change quite a bit if training went on for more epochs.

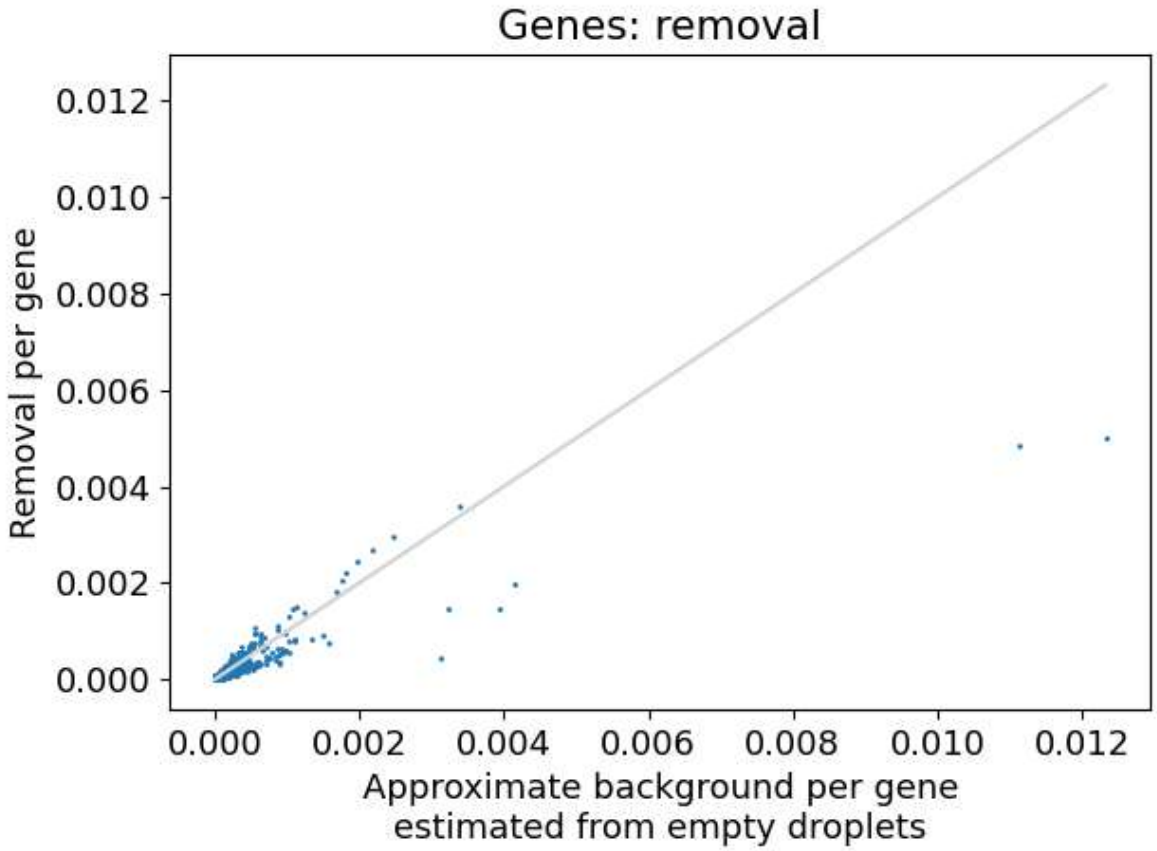
Automated assessment --------

- We typically expect to see the training ELBO increase almost monotonically. This curve seems to have a concerted period of motion in the wrong direction near epoch 25. If this is early in training, this is probably okay.

Summary:

This learning curve looks normal.

Examine count removal per gene



R^2 value for the fit of y=x for removal is 0.2534

WARNING: This deviates from expectations, and may indicate that the run did not go well

Table of top genes removed

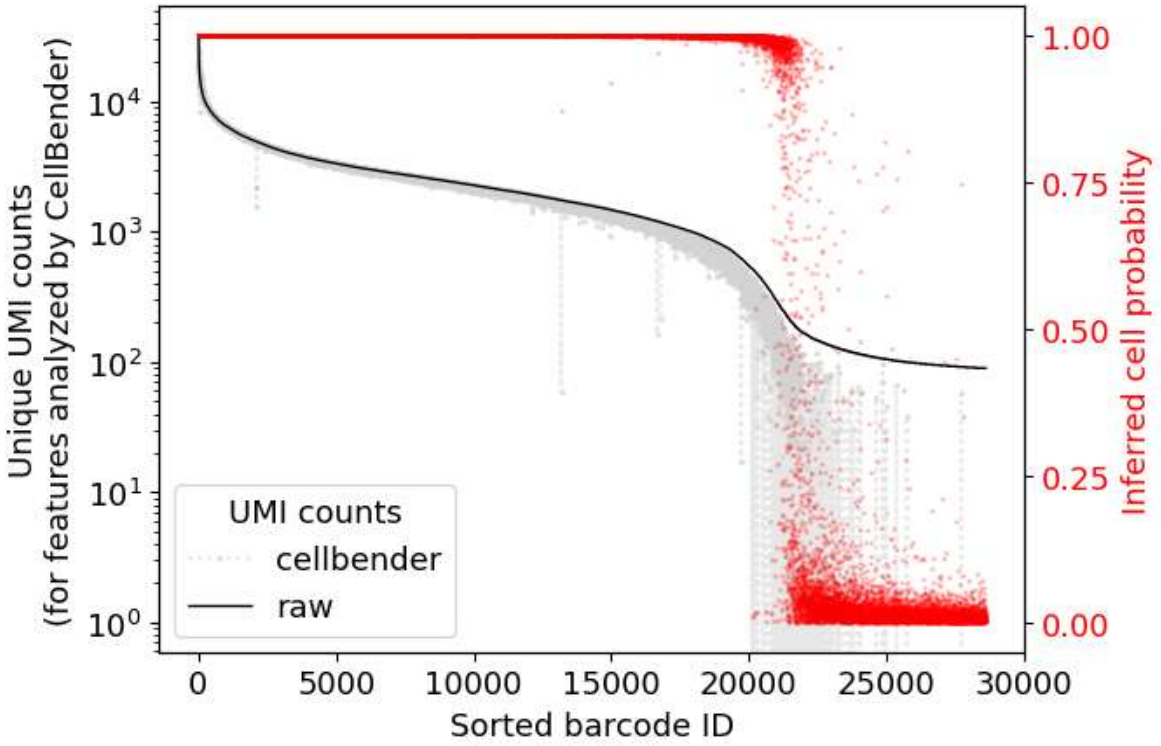
Ranked by fraction removed, and excluding genes with fewer than 3674 total raw counts (90th percentile)

gene_name	ambient_expression	feature_type	genome	gene_id	cellbender
MT-ND3	0.005912	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000198840	
RPS29	0.003810	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000213741	
RPL31	0.001679	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000071082	
MT-ND4	0.009294	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000198886	
RPS27	0.003726	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000177954	
MT-ND1	0.004331	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000198888	
MT-CO2	0.010229	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000198712	
RPS21	0.001946	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000171858	
MT-CYB	0.005304	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000198727	
ATP5F1E	0.000635	Gene Expression	GRCh38-arc-hardmasked-optimised-v2	ENSG00000124172	

Cell probabilities

The inferred posterior probability that each droplet is non-empty.

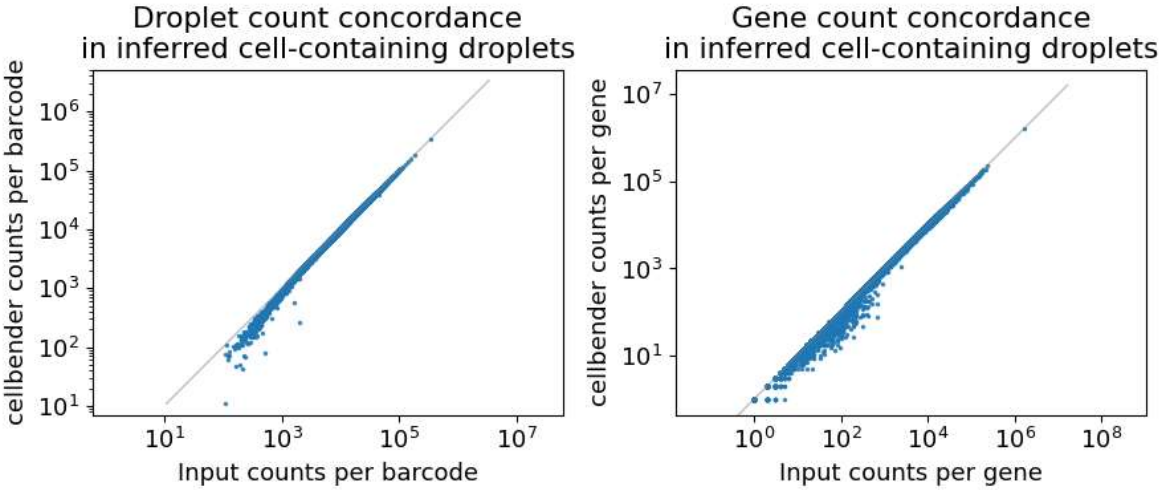
We sometimes write "non-empty" instead of "cell" because dead cells and other cellular debris can still lead to a "non-empty" droplet, which will have a high posterior cell probability. But these kinds of low-quality droplets should be removed during cell QC to retain only high-quality cells for downstream analyses.



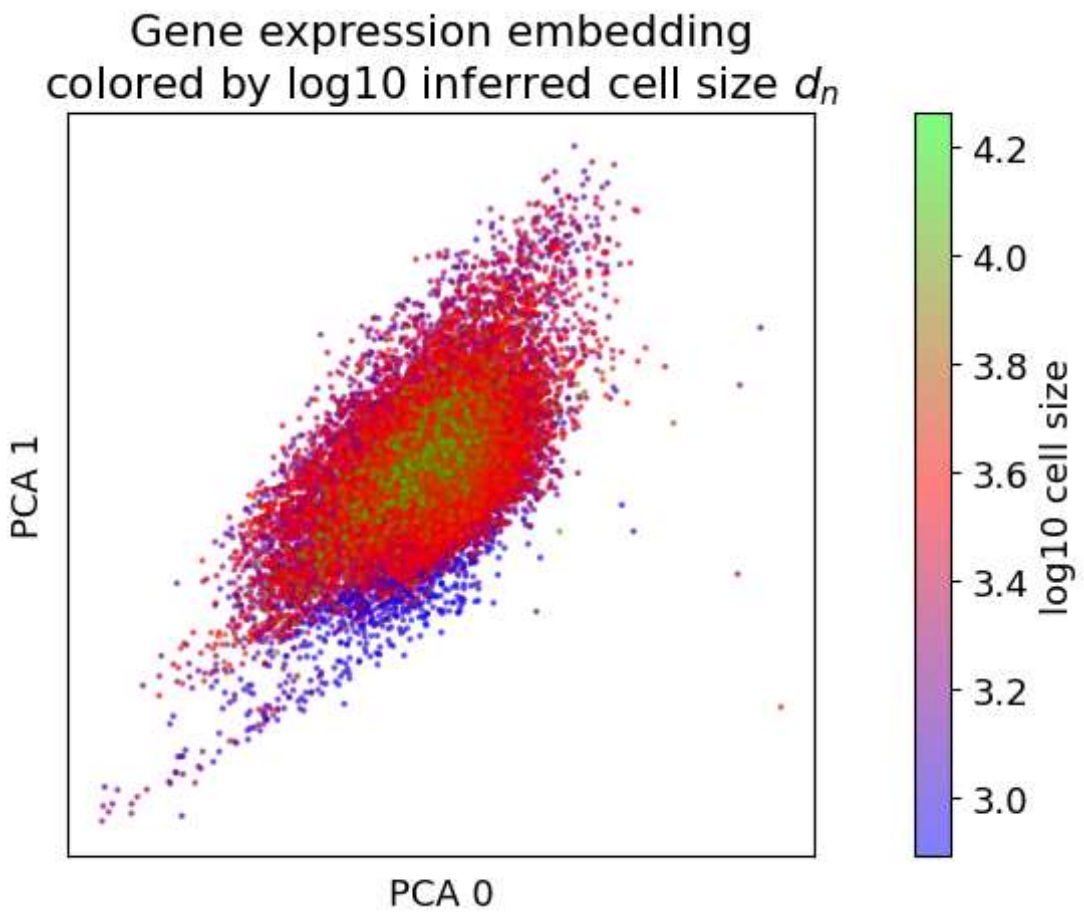
Concordance of data before and after `remove-background`

The intent is to change the input data as little as possible while achieving noise removal. These plots show general summary statistics about similarity of the input and output data. We expect to see the data lying close to a straight line (gray). There may be outlier genes/features, which are often those highest-expressed in the ambient RNA.

The plots here show data for inferred cell-containing droplets, and exclude the empty droplets.



PCA of encoded gene expression



We are not looking for anything specific in the PCA plot of the gene expression embedding, but often we see clusters that correspond to different cell types. If you see only a single large blob, then the dataset might contain only one cell type, or perhaps there are few counts per droplet.

Summary of warnings:

Back-tracking in training ELBO.

Per-gene removal does not closely match a naive estimate of ambient RNA from empty