# How To BatchLR

## Introduction

Batchelor implements three methods for batch correction of single-cell RNA sequencing data. Batch effects refer to differences between data sets that are often an unavoidable product of differences in samples we wish to compare, including: time-points, laboratories, and even sequencing pipelines. The fastMNN method implemented in this plugin is based on detecting mutally nearest neighbors as well as simple sparsity-preseving translation of the population means. This approach does not rely on equal population compositions across batches, instead it requires that only a subset of the population be shared between batches. This allows the method to be scaled to large numbers of cells. The algorithm returns a matrix of corrected principal components that can be used for downstream analyses such as visualization and clustering, without altering the underlying data matrices. MNN batch effect correction has been implemented as an R/BioConductor library and published: Haghverdi L, Lun ATL, Morgan MD, Marioni JC (2018). "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors." Nat. Biotechnol., 36(5), 421–427. https://www.nature.com/articles/nbt.4091.

The latest version of BatchLR now includes Harmony Batch correction. You can read more about Harmony in the publication: Korsunsky et al "Fast, sensitive, and accurate integration of single cell data with Harmony" https://www.biorxiv.org/content/10.1101/461954v2

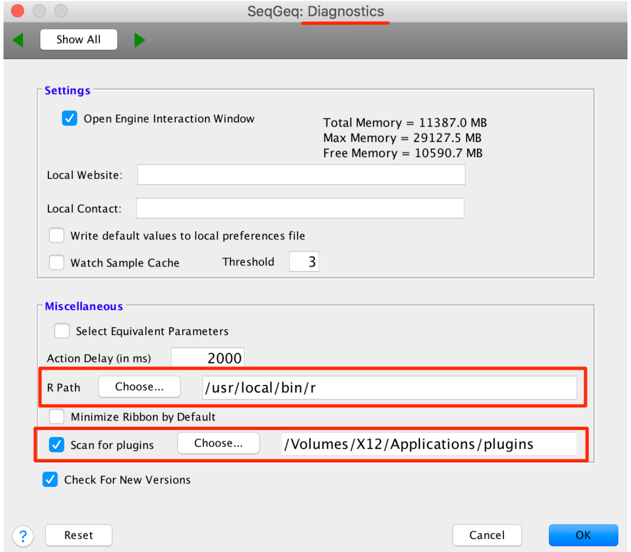
We have developed a BatchLR plugin that integrates this functionality directly into SeqGeq. A video tutorial is available here: https://tinyurl.com/BatchLR-v0-3-Demo.

Please review FlowJo documentation for installing plugins http://docs.flowjo.com/d2/plugins/installing-plugins/.

### Download and installation

1. Place the plugin .jar file in your Plugins folder, and direct SeqGeq to that folder using the Diagnostics section of the Preferences.

2. Make sure you have R installed and the R path is specified in the R Path field of the Diagnostics section of the Preferences.



Settings

☑ Open Engine Interaction Window

Total Memory = 11387.0 MB
Max Memory = 29127.5 MB
Free Memory = 10590.7 MB

Local Website:

Local Contact:

☐ Write default values to local preferences file

☐ Watch Sample Cache Threshold 3

Miscellaneous

☐ Select Equivalent Parameters

Action Delay (in ms) 2000

R Path Choose... /usr/local/bin/r

☐ Minimize Ribbon by Default

☑ Scan for plugins Choose... /Volumes/X12/Applications/plugins

☑ Check For New Versions

3. The batchelor package needs to be installed and will run in the R environment. This plugin was tested in R version 4.0.5 & 4.1.0 and batchelor version 1.8.0. To install the required R packages, use the following commands in R:

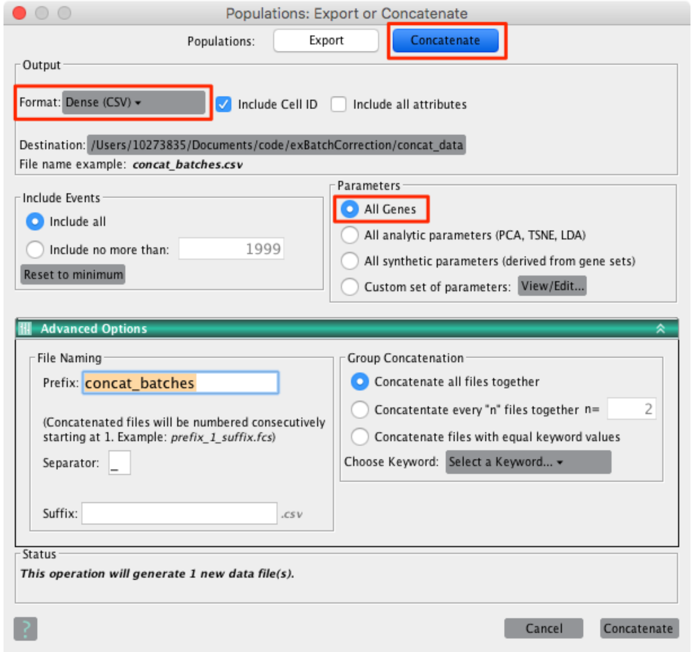
```
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install()
```

```
install.packages(c("Seurat", "BiocManager", "dplyr", "readr", "tibble", "cowsay",
                   "data.table", "cowplot", "devtools"))
BiocManager::install(c("batchelor", "SingleCellExperiment", "scran", "scater"))
library(devtools)
install_github("immunogenomics/harmony")
```

**Note1:** For older versions of R, please refer to the appropriate Bioconductor release: https://bioconductor.org/about/release-announcements/

## Usage

In order to run the BatchLR plugin you will first need to concatenate the different sample files together that you want to perform batch correction on. Use cmd-click(Mac) or Ctrl-click(Windows) to select multiple samples to concatentate. Right-click and choose Export/Concatenate Populations. In the resulting window, make sure to select ‘All Genes’, or a Custom geneset of interest:



Populations: Export or Concatenate

Populations: Export [Concatenate]

Output

Format: Dense (CSV)  ☑ Include Cell ID  □ Include all attributes

Destination: /Users/10273835/Documents/code/exBatchCorrection/concat_data

File name example: *concat_batches.csv*

Include Events

● Include all

○ Include no more than: 1999

Reset to minimum

Parameters

● All Genes

○ All analytic parameters (PCA, TSNE, LDA)

○ All synthetic parameters (derived from gene sets)

○ Custom set of parameters: View/Edit...

Advanced Options

File Naming

Prefix: concat_batches

(Concatenated files will be numbered consecutively starting at 1. Example: *prefix_1_suffix.fcs*)

Separator: _

Suffix: .csv

Group Concatenation

● Concatenate all files together

○ Concatentate every "n" files together n= 2

○ Concatenate files with equal keyword values

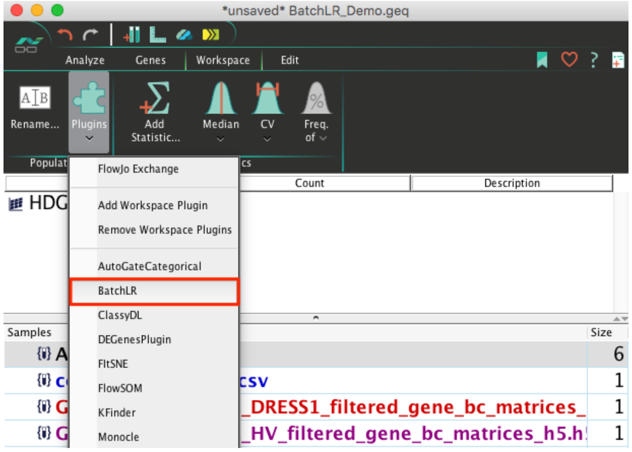
Choose Keyword: Select a Keyword...

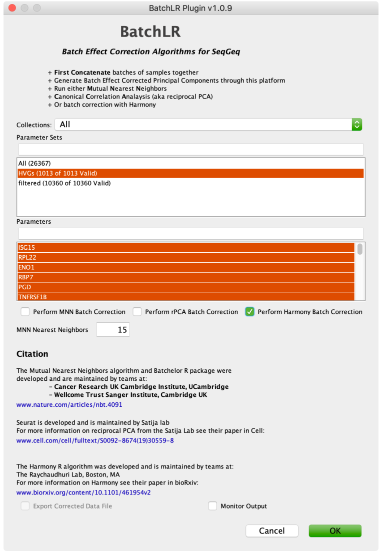
Status

***This operation will generate 1 new data file(s).***

**Note:** The order in which you add samples to the workspace will determine the order in which they are concatenated and assigned SampleIDs. The SampleID keywords will be used to create the batches to correct. The order in which the batches are corrected will affect the final results. The first batch is used as a reference batch against which the subsequent batches will be corrected. Corrected values of the second batch are added to the reference batch, against which the third batch is corrected, and so on.

To run the algorithm, select the concatenated data matrix, go to the Workspace tab of the SeqGeq workspace and select the **BatchLR** plugin from within the plugins drop-down menu. Note that plugins will be unavailable (greyed out) if no file is selected:





BatchLR Plugin v1.0.9

**BatchLR**

***Batch Effect Correction Algorithms for SeqGeq***

+ **First Concatenate** batches of samples together
+ Generate Batch Effect Corrected Principal Components through this platform
+ Run either **M**utual **N**earest **N**eighbors
+ **C**anonical **C**orrelation **A**nalaysis (aka reciprocal PCA)
+ Or batch correction with Harmony

Collections: All

Parameter Sets

All (26367)
HVGs (1013 of 1013 Valid)
filtered (10360 of 10360 Valid)

Parameters

ISG15
RPL22
ENO1
RBP7
PGD
TNFRSF1B

☐ Perform MNN Batch Correction ☐ Perform rPCA Batch Correction ☑ Perform Harmony Batch Correction

MNN Nearest Neighbors 15

**Citation**

The Mutual Nearest Neighbors algorithm and Batchelor R package were
developed and are maintained by teams at:
**– Cancer Research UK Cambridge Institute, UCambridge**
**– Wellcome Trust Sanger Institute, Cambridge UK**
www.nature.com/articles/nbt.4091

Seurat is developed and is maintained by Satija lab
For more information on reciprocal PCA from the Satija Lab see their paper in Cell:
www.cell.com/cell/fulltext/S0092-8674(19)30559-8

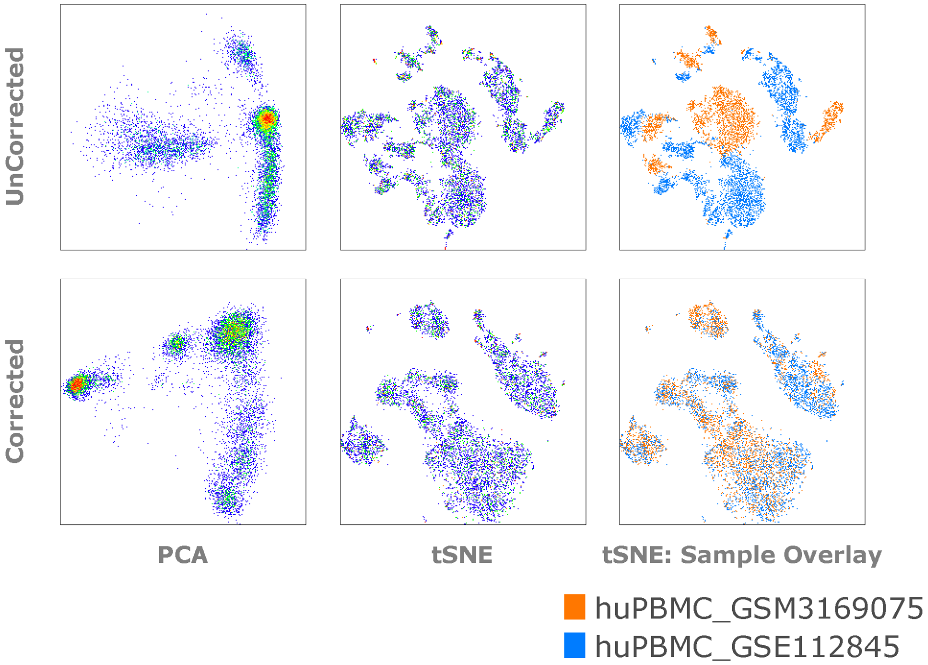
The Harmony R algorithm was developed and is maintained by teams at:
The Raychaudhuri Lab, Boston, MA
For more information on Harmony see their paper in bioRxiv:
www.biorxiv.org/content/10.1101/461954v2

☐ Export Corrected Data File ☐ Monitor Output

**Note:** We generally recommend a parameterset of the data's most highly dispersed features to correct for batch effects, for best possible results downstream, based on preliminary testing.

## Benchmarking

In **Fig1** batch effect correction performed between similar samples, human PBMC datasets, which were collected for entirely separate studies.



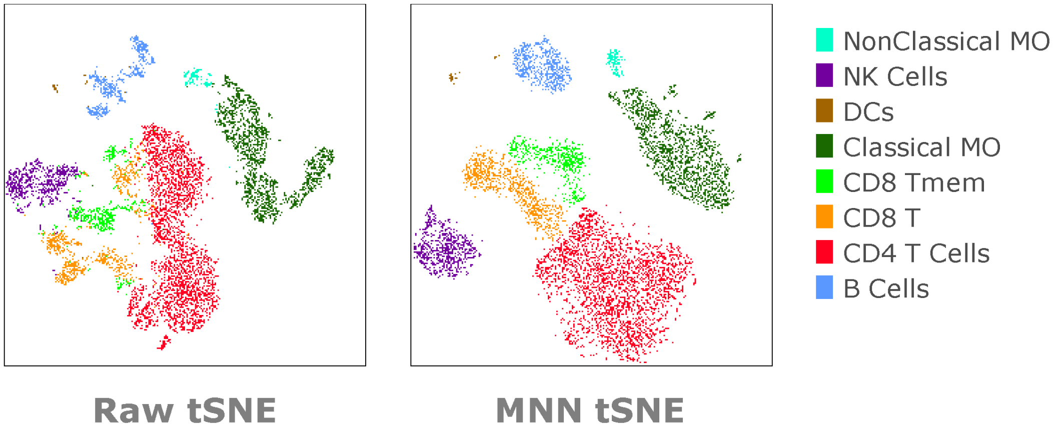


**Figure 1 -** Initial the combined data matrices show stark batch effects in raw PCA space, despite similar sample types, but these are nicely mitigated after MNN batch effect correction is applied.

**Fig2** shows the subtle effect of batch effect correction when samples are placed in different order, by visualizing human Dendritic cells sequenced after enrichment by sorting in Flow Cytometry, combined with a dataset from total human PBMCs (PBMCs followed by DCs above, and DCs followed by PBMCs below):

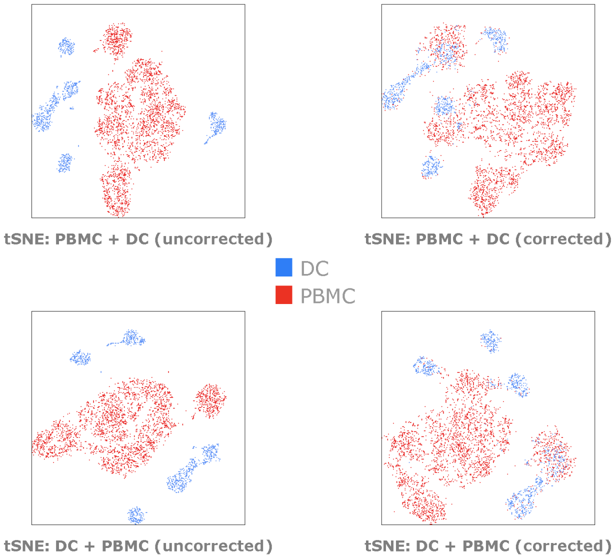


**Figure 2 -** The DC subsets are slightly better preserved and more easily distinguished when the PBMC batch is included first.

**Fig3** illustrates the effect of batch correcting many samples (x5) from a variety of randomly selected data matrices including cutanious T cells, monocytes, and different total PBMC files:

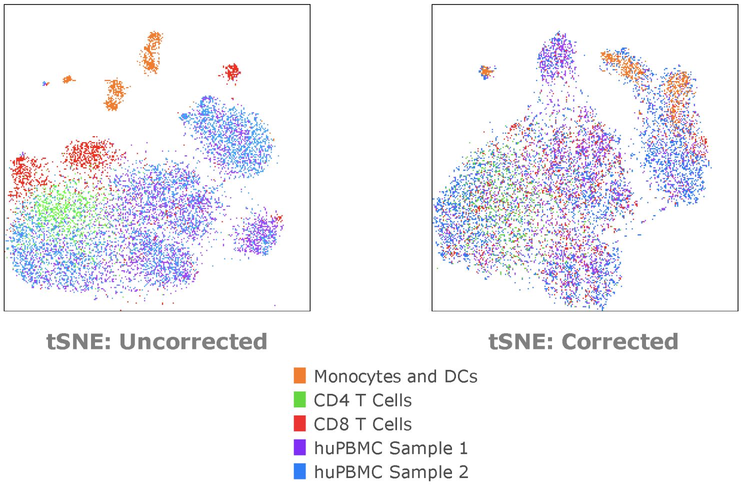
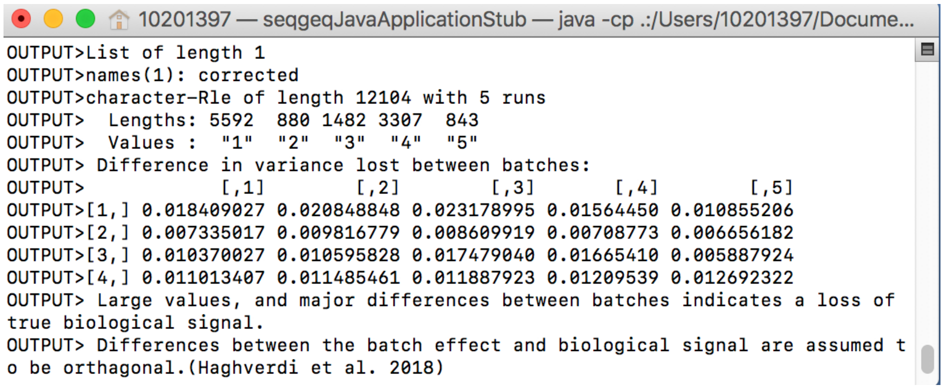


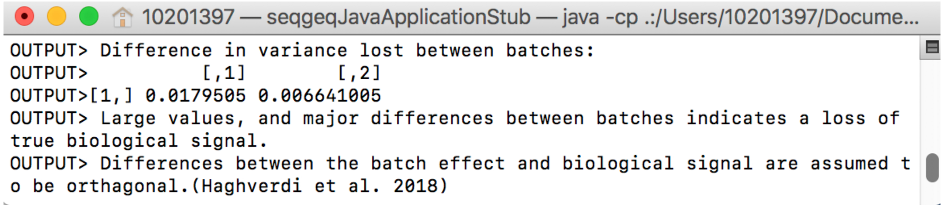
**Figure 3 -** Too many decidedly different sample batches, from different: Subjects, studies, and even sequencing technologies - can lead to poor quality corrections. In this case the distinction between different subsets appears to have become less clear-cut.

**Statistical scores** calculated per batch give some indication of the goodness of batch effect correction, **Table 1** (poor) and **Table 2** (good). By checking the **Variance Lost** as a result of the algorithm, accross batches comparatively, we can get some idea of the amount of biological signal that has been sacrified. These scores can be viewed by accessing the terminal while BatchLR is running in SeqGeq:


10201397 — seqgeqJavaApplicationStub — java -cp .:/Users/10201397/Docume...

```
OUTPUT>List of length 1
OUTPUT>names(1): corrected
OUTPUT>character-Rle of length 12104 with 5 runs
OUTPUT>  Lengths: 5592  880 1482 3307  843
OUTPUT>  Values :  "1"  "2"  "3"  "4"  "5"
OUTPUT> Difference in variance lost between batches:
OUTPUT>            [,1]        [,2]        [,3]       [,4]        [,5]
OUTPUT>[1,] 0.018409027 0.020848848 0.023178995 0.01564450 0.010855206
OUTPUT>[2,] 0.007335017 0.009816779 0.008609919 0.00708773 0.006656182
OUTPUT>[3,] 0.010370027 0.010595828 0.017479040 0.01665410 0.005887924
OUTPUT>[4,] 0.011013407 0.011485461 0.011887923 0.01209539 0.012692322
OUTPUT> Large values, and major differences between batches indicates a loss of
true biological signal.
OUTPUT> Differences between the batch effect and biological signal are assumed t
o be orthagonal.(Haghverdi et al. 2018)
```

**Table 1 -** The table above is taken directly from batch effect correction shown in **Figure 3**, and is given in units: frequency of total variance.


10201397 — seqgeqJavaApplicationStub — java -cp .:/Users/10201397/Docume...

```
OUTPUT> Difference in variance lost between batches:
OUTPUT>          [,1]        [,2]
OUTPUT>[1,] 0.0179505 0.006641005
OUTPUT> Large values, and major differences between batches indicates a loss of
true biological signal.
OUTPUT> Differences between the batch effect and biological signal are assumed t
o be orthagonal.(Haghverdi et al. 2018)
```

**Table 2 -** The table above was calculated for batch effect correction shown in **Figure 1**.

## Leave us your feedback

Please write to seqgeqq@bd.com with any questions or concerns.