# Data Processing Information for M295

## Processing Pipeline Overview

1. Raw paired-end sequencing data is parsed and assembled into single blocks consisting of barcode, UMI, and SSI.  
   During this stage all resulting reads with ’N’s are filtered, along with any reads containing sections longer than 7 bp containing the same base (since homopolymers introduce errors).
2. We take the barcodes of all remaining reads, and perform an all-against-all alignment and find all sets of barcodes with Hamming distance of 3 or less from each other (based on 30 bp of barcode length). These sets are then used to ‘collapse’ the barcode in all the full sequences (VBC+UMI+SSI) to a single unique sequence, essentially ‘fixing’ replication and sequencing errors in the barcodes.
3. We then take the full sequences and sort them all according to SSIs (which correspond to unique brain areas), and then determine the molecule number of a barcode in a certain brain area by counting UMIs. As this is done we aggregate the original read counts for each VBC+UMI+SSI.
4. Within each SSI we separate out spike-in and real sequences based on their properties.
5. We then create a full aggregate table M295.vbctable.tsv. This is the step at which we apply a minimum read count threshold of 2.
6. We then apply all filters and thresholds to create M295.vbfiltered.tsv. This table represents only data that will be included in the per-brain matrices.
7. The filtered table can now be converted to connection matrices, with the number of each real and spike-in barcode in each brain area. In this matrix, each row is one barcode, each column is a brain area, and the element of the matrix corresponds to the molecule number of the barcode in the brain area.
8. Lastly, we normalized the number of molecules of each barcode in each brain area to the total number of spike-in molecules in the corresponding brain area. This is to compensate RT/PCR variations during sequencing library preparation. This produces the normalized barcode matrix above.

## Table and Matrix information

M295.vbcfiltered.tsv: All viral barcode (VBC) sequences, broken down by type (real, spike-in) and target region.  
Barcode sequences may appear more than once, if found in multiple target areas. ``  
umi\_count is the number of unique UMIs seen, with read\_count representing the total number of reads behind those UMIs (the allocation of reads to UMIs is not included).

For each brain:  
<brain>.rbcm.tsv raw (real) barcode matrix  
Raw real barcode UMI counts from filtered VBC table.  
Only includes barcodes where:  
At least one target area had > molecules (UMIs)  
Rows represent barcodes and columns represent areas.

<brain>.sbcm.tsv spike-in barcode matrix.  
Spike in barcode UMI counts.  
Rows represent barcodes and columns represent areas.

<brain>.nbcm.tsv normalized barcode matrix  
Filtered barcode values normalized by spike-in counts per SSI area.

The columns for all the matrices represent the ‘ourtube’ ID contained in the ‘sampleinfo.tsv’ file. That file includes the mappings between user-supplied tube labels, our tube labels, and the rtprimer label for each sample.

## Quality Control

For this MAPseq data set, we have checked:

1. Sequencing depth for projection sites. There are a minimum of 2 reads for each VBC+UMI+SSI combination in target site samples. This sequencing depth is considered enough for target sites.

Spike-in counts are mostly uniform across all projections sites.  
We have seen at most 2-3 folds difference in spike-in counts between projection sites, which is considered normal.

1. We have checked barcode amount in controls.

The UMI counts from negative controls are listed in the TSV named “vbc\_controls.tsv”. This background looks normal.

1. We have calculated false positive rates. For target sites, the false positive rate is 0 when the UMI threshold is set to 2.

## Further Analysis

What we haven’t done but MAPseq users should do is to:

1. Check distribution/mean barcode counts in projection areas: does it fit what’s known?
2. Population-level correlation with ABA or other bulk tracing data
3. Check area-to-area correlations. They should fit what’s known (e.g. cortical neurons mostly don’t project to contralateral cortex and brain stem at the same time). Check for potential contaminations/fibers of passage/bad dissections.
4. Check consistency across multiple brains: Are neurons mixed in t-sne? Are the min distance within a brain similar to min distance across brains?
5. Clustering: do you get known cell types?

## Data and sample storage policy

The MAPseq/BARseq Core Facility is pleased to offer secure storage of sequencing data and processed samples for a period of six months (0.5 years) following the completion of your project. During this time, we kindly encourage you to back up any sequencing data you wish to retain. If you would like to have your samples returned, please feel free to request shipment within this six month window. After this period, all sequencing data will be permanently deleted unless you have made other arrangements with us. We respectfully remind you that it is your responsibility to request data retrieval or sample return within the allotted timeframe. We regret that we are unable to retain data or samples beyond the six month period without prior notice.

## Acknowledgement Policy

If the services from the MAPseq Core Facility are used to generate data in any publications, please acknowledge the MAPseq Core Facility in CSHL for processing samples for MAPseq sequencing.