**Readme Barcoding data Eisele et al. EPO**

1. **Calling barcodes**

**1.0 Getting the raw sequencing files**

The raw sequencing files are available on xenodon see link the corresponding folder.

**1.1 Xcalibur run to extract barcodes**

The raw barcode sequencing files will be filtered for a perfect match to the input index- and common-sequences of the barcodes and filtered against the barcode reference list using the program XCALIBR available on github (<https://github.com/NKI-GCF/xcalibr)>. To use Xcalibr, the following files will have to be put in one same folder :

* Plate\_index\_list.txt
* Barcode\_reference\_list.txt
* newXcalibr.bash
* fastq files

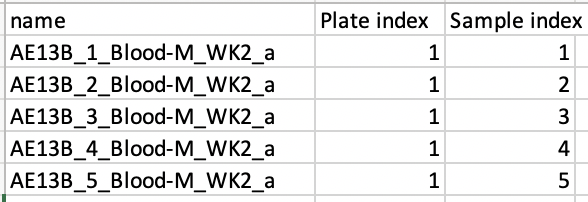
Barcode\_reference\_list.txt is the reference list of barcodes created by duplicate sequencing of the barcode plasmids and further filtering as described in Material and Methods. The Plate\_index\_list.txt file contains a list of plate indices (added during PCR 2 of barcode amplification) used in the specific experiment. In the script file adapt the min and max values, the name of the file to be handled, the number of samples per plate index, as well as f1 and all paths. The bash script can be run in terminal. It will call xcalibr analyze and xcalibr extract (see also Xcalibur documentation). Output will be for every fastq, a file with ending -res-all.txt -> these are the extracted barcodes. Furthermore, two files with ending -plate\_index\_stats.txt and -barcode\_stats.txt will be generated, these give for the most abundant plate index and barcodes respectively, the percentage of reads attributed.

**1.2 Process Xcalibur output**

After the Xcalibur run, the “ProcessXcaliburoutput.R” file can be used for the last processing steps- merging all -res-all.txt files. Adjust the input arguments, and the plate indices in the file and run it in R. Output will be the file “barcode\_data\_processed.csv” which is a table with per barcode a column of reads per sample index and plat index.

**1.3 Add names to processed Xcalibur output**

In the next step we add a more meaningful name to each sample\_index – plate\_index combi. This is done in R using the file “Addingnames.R”. Input are the “barcode\_data\_processed.csv” and a txt file with as below a column with the new name, plate index and sample index. The names we give usually have the setup Experiment\_(EPO does)\_mouse number\_organ\_lineage\_timepoint\_replicate.



The resulting named files can be found in the folder unfiltered\_barcoding\_data and are named “ExperimentX\_1\_witnames”. These are the files on which filtering is performed, first outside of R using excel, then using R. This is explained in the next sections. We advise to make a word file to take notes during filtering for every experiment.

1. **Filtering outside R**

**2.1 Check number of reads and proportion that could be matched**:

* Open “Experiment\_1\_witnames” file in excel.
* Under the row and column of reads not matching both barcode and index library, put a row and column of sum of available barcodes as well as a row and column to check the proportion of reads that could be matched. This should be quite high. Only if under 40-50% take it out. Be sure to consider the total read count. If very low overall read count, the percentage is not important.
* Change the name by adding “with nohits” ending and number 2\_ at start.

**2.2. Make it the raw reads file:**

* Before analyzing further take out anything you wanted to filter out, delete the no hit rows.
* Safe new table “3\_ withnames\_without\_nohits”

1. **Filtering in R**

The “3\_withnames\_without\_nohits” files will be further processed in R. First, all reads belonging to a same sample (different plate indices) are summed to have one readcount value per sample. After this, further filtering steps are performed. These are:

* Normalization of readcounts to 100000 per sample
* Checking correlation between replicates and filter out samples with low correlation
* Removal of barcodes not present in one replicate of a sample
* Getting 95th quantile value of noise
* Filtering based on 95th quantile noise value
* Re-run previous filtering steps on quantile-filtered data

These steps are distributed over three different R scripts. In the first R scripts, reads belonging to a same sample are summed, and filtering is performed till and including the removal of barcodes not present in one replicate. In a second script, the 95th quantile of noise reads is determined. In the third script, filtering using the 95th quantile value is performed and all other filtering steps repeated to gain the final filtered data as found in the “filtered\_barcoding\_data” folder. The individual filtering steps are further explained in the Rscripts.