

Antibody microarray data analysis pipeline



Introduction

This document will describe the usage of and concepts behind the applications developed by Bionamic for Immunovia to reproducibly and rapidly develop signatures in antibody microarray datasets. The pipeline is designed to start with a list of expression sets in the standard format used at Immunovia, where the user may subset, handle missing data, normalize, and QC the data. The data is then prepared for a backward elimination (BE) run, after which the results can be visualized, studied and used to develop a signature. Once the signature is locked, an SVM model is trained and validated on independent data. The results from the validated signature can then further be used in simple simulations to estimate and visualize positive and negative predictive values at user defined sensitivity and specificity thresholds. The analytical pipeline will automatically export interactive HTML reports at two key points in the workflow, where all selections and decisions are tracked together with key results, figures and statistics.

The workflow is split into two Shiny applications, where the first (here referred to as '*normqc*') hands the data and user settings over to an R script running the computationally expensive BE process. The BE script utilizes parallel processing to reduce the processing time, and when completed, hands the data over to the second Shiny app (here referred to as '*sigdev*'). There, the BE result can be studied closely, and used to design an antibody signature, both in terms of length and content.

This guide will take you through each workflow step, and explain each feature. It will also go into some detail regarding some of the most central concepts.

App 1 – Normalization, QC, and BE preparation

1. Open <http://bionamic-normqc.awsimmunovia.com/app1/> in your favorite web browser.
2. A page with six tabs will open. You will work your way through all six, starting with the **Select data** tab. Select the expression set you want to work with and click the **LOAD DATA** button.

The screenshot shows the 'Select data' tab of the Immunovia application. At the top, there is a navigation bar with the Immunovia logo and six tabs: 'Select data', 'Data split', 'Imputation', 'Normalization', 'Analytes', and 'Report/Export'. The 'Select data' tab is active. Below the navigation bar, there is a 'Select data set' dropdown menu with 'HTest-expressions-06' selected. To the right of the dropdown, the text 'HTest-expressions-06.json' is displayed. Below the dropdown, there is a 'LOAD DATA' button. To the right of the button, the text 'Clinical variables' is displayed. At the bottom, there is a note: 'Clinical variables available in data set.'

After loading a data set, stats on the samples' clinical and batch variables will be displayed in three tables.

3. **Switch to the Data split tab.**
 - a. Choose one or more classification variables. These will allow you to compose case and control groups. It is important to assign the patient group you are trying to detect in the **Cases** group, as these will be the target in later calculations of positive predictive value and similar statistics. In the example below, **status_detailed** was selected and **stage II**, **stage III**, and **stage IV** combined into the **Cases** group to be compared to **Normal** in the **Controls** group.

Create training and test sets

Select classification variable

- ☐ status
- ☒ status_detailed

Cases

- ☐ Normal (783)
- ☐ Pancreatitis (22)
- ☐ Stage_IA (1)
- ☐ Stage_IB (5)
- ☒ Stage_IIA (32)
- ☒ Stage_IIB (89)
- ☒ Stage_III (55)
- ☒ Stage_IV (199)

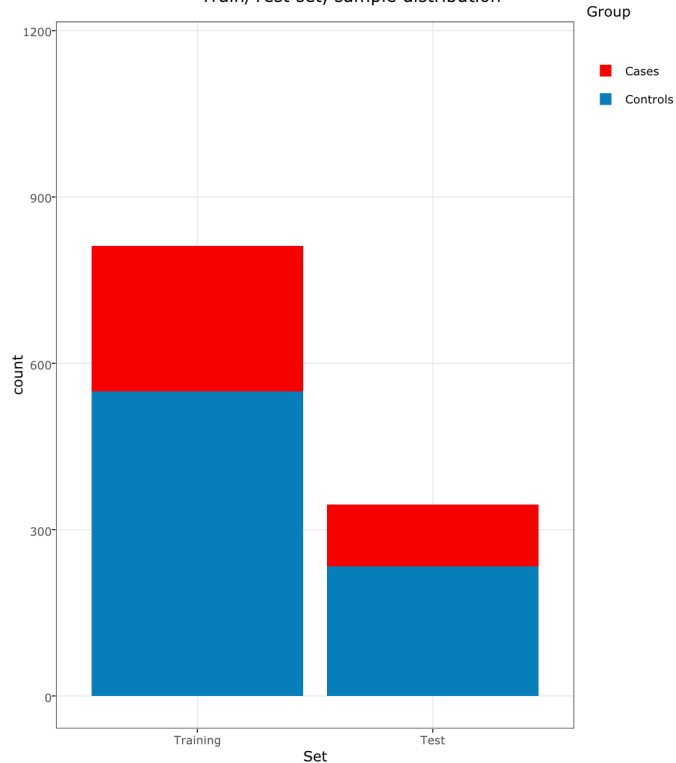
Controls

- ☒ Normal (783)
- ☐ Pancreatitis (22)
- ☐ Stage_IA (1)
- ☐ Stage_IB (5)
- ☐ Stage_IIA (32)
- ☐ Stage_IIB (89)
- ☐ Stage_III (55)
- ☐ Stage_IV (199)

Data split

- ☒ Split on variable
- ☐ Split by variable

Train/Test set, sample distribution

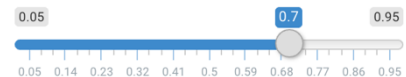


4. **Select proportion of data to withhold** as independent test set. This can be performed in two different ways. **Split on variable** will divide **Cases** and **Controls** with equal frequencies into two sets defined by the slider **Fraction of data for training**. The default setting is 0.7 (*i.e.* 70 % in the training set). The plot shows the training and test set sizes and their Cases/Controls composition.

Data split

- ☒ Split on variable
- ☐ Split by variable

Fraction of data for training



The second option is to use the **Split by variable** method. This allows the user to define training and test set based on a batch variable. Select a variable and choose which batches should be included in the training. In doing so, samples from the selected batches will be used for training, and the remaining for the independent test set.

Data split

- ☐ Split on variable
- ☒ Split by variable

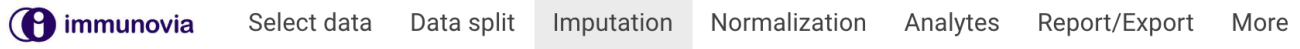
Select variable to split by

- ☐ slide
- ☐ array_block
- ☒ source_plate
- ☐ well
- ☐ sample_name
- ☐ scan_position
- ☐ scan_serial
- ☐ scan_date

Select subgroups for training set

- ☒ P07
- ☒ P10
- ☒ P05
- ☐ P12
- ☐ P15

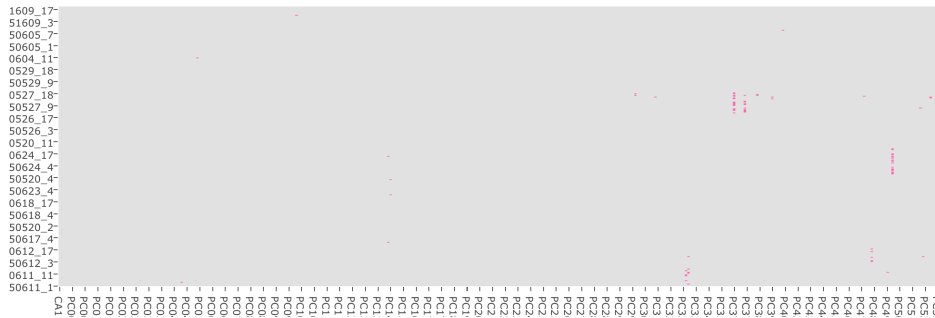
5. Switch to the **Imputation** tab to handle missing values.



Missing values can here be handled by either removing the entire variable or trying to impute the missing values. The first is recommended if a “significant fraction” of the values for a specific variable is missing (*e.g.* above 5 or 10 %), otherwise the benefit of keeping the variable may be higher than the risk associated with introducing artificial values in the dataset.

Three plots and a selectable list of antibodies will be created upon opening the **Imputation** tab.

a. **Heatmap where missing values are shown as pink bars.**



To impute missing values, select **Imputation method**.

Imputation method

- ☐ None
☐ Bag (slow)
☒ Median (fast)

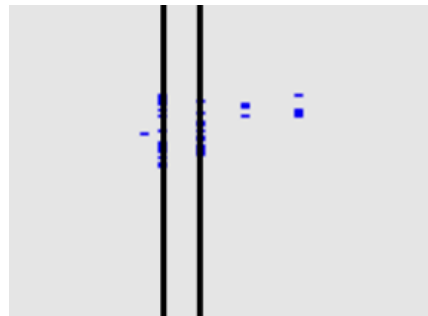
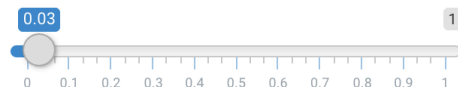
The **Median** option is fast and will replace each missing value with the median of the non-missing values for the same variable. The **Bag** option will perform a more sophisticated calculation by first combining samples into groups that are similar, and then only use the data from those samples when estimating the missing value. This is a *very* slow method and will take a long time to do in a large data set.

To remove variables with too many missing values, click the **Yes** button as shown below, and set an **Impute/remove cut-off** using the slider (defaulting to 0.05). A value of 0.03 will remove all variables with more than 3 % of values missing. Variables with less than 3 % missing values will be kept. Imputed values will change from pink to blue, and removed variables will be marked with a black line in the heatmap.

Remove analytes with number of missing values
above threshold

- ☐ No
☒ Yes

Impute/remove cut-off



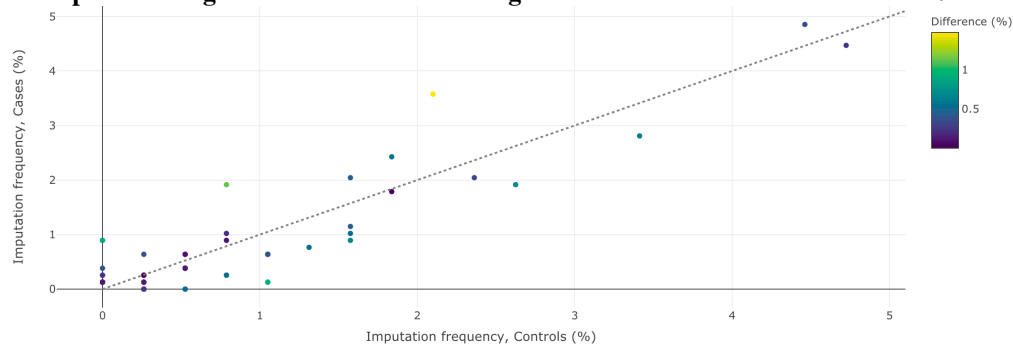
b. **Antibodies to include list**

This list can be used to manually remove antibodies from the dataset. **Importantly**, changes to the dataset using this list should be done *after* any removal of antibodies using the **Impute/remove cut-off** has been made. Adjusting the cut-off or clicking the **Yes** button as described above will negate any manual adjustments made to the list. Thus, *always* perform step 5 a before 5 b.

Antibodies to include

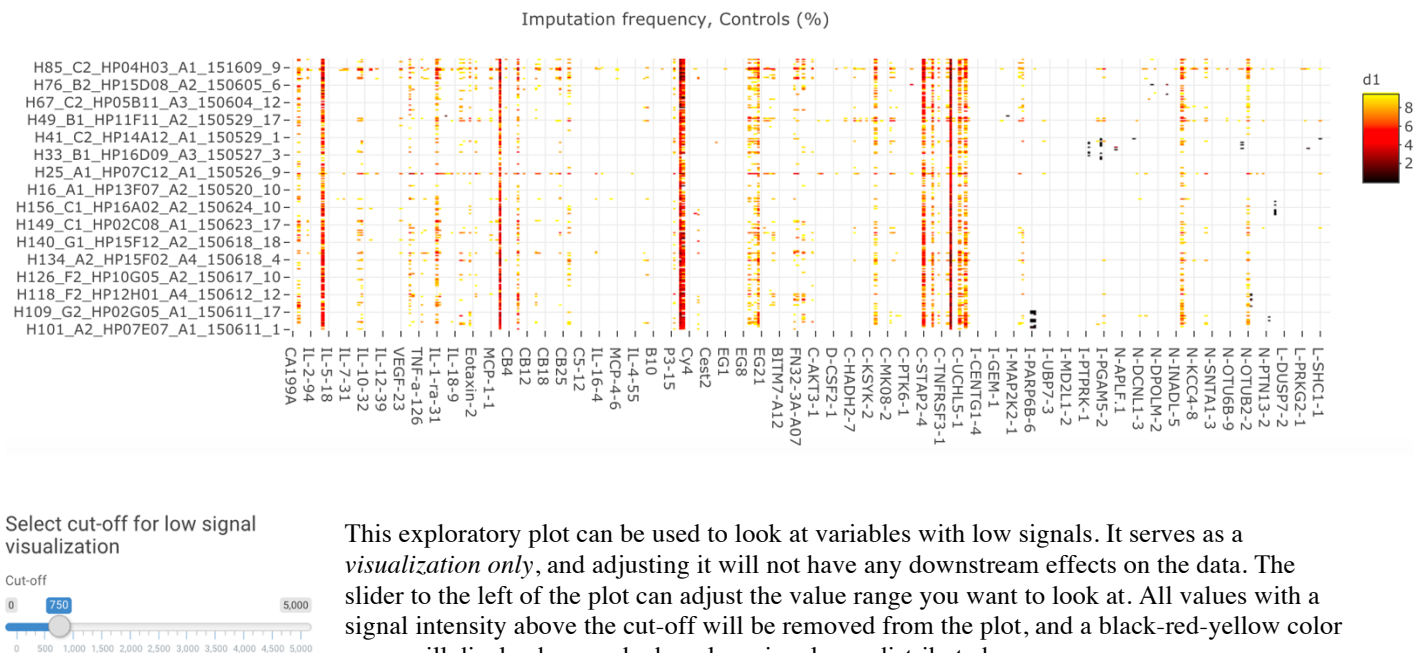
- ☒ CA199A
☒ CA199B
☐ PC001
☒ PC002
☒ PC003
☒ PC004

c. Scatter plot showing the distribution of missing values for each variable across cases/controls.



This plot can be used to identify antibodies with uneven imputation frequency across the cases and controls groups. Points are plotted by their imputation frequency in both groups and colored by the relative difference of the number of imputed values. Points should not deviate significantly from the dotted line. Points in the plot will disappear as their corresponding antibodies are removed from the dataset.

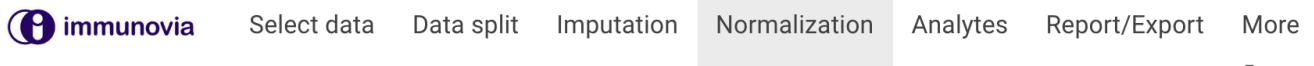
d. A heatmap visualizing low (but not missing) values.



This exploratory plot can be used to look at variables with low signals. It serves as a *visualization only*, and adjusting it will not have any downstream effects on the data. The slider to the left of the plot can adjust the value range you want to look at. All values with a signal intensity above the cut-off will be removed from the plot, and a black-red-yellow ramp will display how and where low signals are distributed.

It is recommended to study this heatmap in the context of the heatmap described in **5a**. It is very important that missing values are *randomly* missing, and not due to low signal. If values that appear as missing in **5a** are situated in a cluster of very low signal antibodies, they should not be imputed. Instead it is recommended to re-quantify the signals on the array or remove them from the dataset completely.

6. Switch to the Normalization tab.



The goal of the work in this tab is to perform normalization on and QC of the data. There are a few functions and concepts that are important to understand for this process to work well.

- Training and test set samples.** The assignment of a sample to either training or test set is made randomly if the data splitting method in **Step 5** was set to **Split on variable**. Further, this random distribution is performed every time any changes are made to the data set. If, for example, a sample is identified to be an outlier and subsequently removed from the data set, all sample are automatically redistributed into training and test set. The same is true if a normalization setting is changed or added – all samples will randomly be redistributed across training and test sets. The main reason for this is to remove the possibility of biasing the dataset by optimizing normalization settings to achieve certain features in the training or test set.

- b. *Normalization factor calculation.* The normalization factor is a factor that describes how a particular sample relates to the average sample. They are hence calculated on a per sample basis, both in the training and test set. One key thing that separates the samples in the test set from the samples in the training set is that parameters used to calculate the normalization factors are defined using the training set samples only. This is important in making sure that the independent test set truly remains independent, *i.e.* that no information from the test set influences the model they will be used to validate. Considering point **6a.** above, this means that every time a sample is removed or a setting changed, the normalization process starts over from the beginning.
7. The left-side panel in the **Normalization** tab lets you set all settings necessary for the normalization.
- a. Select **Scale**, defaulting to **Logarithmic**.

Scale and normalize data

Scale

- ☒ Logarithmic
- ☐ Arithmetic

- b. In the drop down menu below **Select normalization** you may select **CV** and/or **ComBat**. If both are selected, they will be applied in the order in which they were selected. To deselect a method, put your cursor in the white box and delete it. If **CV** is selected, a slider titled **Fraction of analytes to use for normalization** will appear. The default of 0.15 will select the 15 % of antibodies with the lowest standard deviation (if using the logarithmic scale) or CV (if using arithmetic scale).

Select normalization

CV ComBat

Fraction of analytes to use for normalization



- c. If you are using **ComBat** normalization, under **Choose batch variable**, select the batch variable that you want to normalize. At the bottom you may see a list under **The following batch parameters included too few observations** that will display batch variables that didn't have enough samples to function in the stage normalization. This is for information only and not actionable.

Choose batch variable

- ☐ array_block
- ☒ source_plate
- ☐ scan_position
- ☐ scan_date

The following batch parameters included too few observations:

slide, well, sample_name, scan_serial

In addition, at the very bottom, a **REMOVE** button together with a dropdown list of all samples can be found. Here, samples can manually be removed from the dataset at any time. Select them in the drop down list and click the **REMOVE** button. Once clicked, all removed samples will be listed in the **Samples currently removed** table to the right of the **Remove sample list**. The table also shows the total number of removed samples at the bottom.

REMOVE

Remove samples

H101_B2_HP05B03_A1_150611_1
 H101_B1_HP10A11_A1_150611_1
 H101_C2_HP12E10_A1_150611_1 |

H101_A2_HP07E07_A1_150611_1
 H101_F2_HP15B01_A1_150611_1
 H101_G2_HP04D11_A1_150611_1
 H102_A2_HP07F06_A2_150611_2
 H102_B1_HP12E11_A2_150611_2
 H102_C1_HP05B06_A2_150611_2
 H102_C2_HP15B02_A2_150611_2
 H102_F1_HP07F04_A2_150611_2

Samples currently removed:

Show **All** entries Search:

Removed
H101_A2_HP07E07_A1_150611_1
H101_B1_HP10A11_A1_150611_1
H101_B2_HP05B03_A1_150611_1
H101_C1_HP03H09_A1_150611_1
H101_C2_HP12E10_A1_150611_1

Showing 1 to 5 of 5 entries Previous 1 Next

8. The right-side panel has six tabs.

Mean and SD Batch effect Heatmap Sample PCAs Outliers Analyte stats

Each of the six tabs have their own specific plots and functions. The message: **Error: missing value where TRUE/FALSE needed** will appear if the missing values in the previous step were not handled. Go back and remove or impute missing values to correct this.

The six tabs has the following plots and functionalities:

a . Mean and SD

- i. Top plot
 - x-axis: **samples** sorted by mean (scaled and normalized) signal
 - y-axis: minimum, mean, and maximum signal
 - color: blue = training set; red = test set.
- ii. Bottom plot
 - x-axis: **antibodies** sorted by mean (raw) signal
 - y-axis: mean raw signal with standard deviation
 - color: yellow if included in calculating the CV normalization factor

b . Batch effect

- x-axis: samples in the first PCA component
- y-axis: samples in the second PCA component
- color: by *selected batch variable*.
- symbol: 1 = cases; 2 = controls

c . Heatmap

- x-axis: antibodies in alphabetical order
- y-axis: samples in alphabetical order
- color: signal intensity Z-score; green > yellow

d . Sample PCAs

Tab '2D PCA'

- left: 2D PCA of non-normalized data
- right: 2D PCA of normalized data
- color: by selected batch variable

Tab '3D PCA'

- left: 3D PCA of non-normalized data
- right: 3D PCA of normalized data
- color: by selected batch variable

e . Outliers

Three outlier detection methods can be selected

Method

☒ KS ☐ sum ☐ upperquartile

The **KS** option will calculate the Kolmogorov-Smirnov test statistic for the intensities of all arrays, while **sum** and **upperquartile** will calculate the sum or the 75 percent quantile, respectively. The distribution of these values are then used to identify sample outliers. Read more about the `arrayQualityMetrics` package implementing these here:

<https://bioconductor.riken.jp/packages/3.2/bioc/manuals/arrayQualityMetrics/man/arrayQualityMetrics.pdf>

The values for all included samples are plotted (purple for training samples and green for test samples), together with an outlier threshold value (red line). Values above the threshold are listed under **Classified as outliers**, and can easily be added to the **Remove sample list** by clicking the **Add outliers to list** button. Once added to the list, the samples are removed from the dataset by clicking the **Remove** button.

Classified as outliers:

H102_B1_HP12E11_A2_150611_2, H105_G1_HP05B10_A1_
H143_C2_HP01H05_A3_150623_3, H143_F2_HP03C08_A3_
H16_C2_HP06C01_A2_150520_10, H32_A1_HP02H03_A2_1

ADD OUTLIERS TO LIST

REMOVE

Remove samples



f . Analyte stats

The antibody correlations and their significances (p-values) can be viewed both as correlation matrices and as histograms. Tabs are used to switch between plot types and radio buttons to change between correlation factors and p-values.

Correlation matrix

Correlation distribution

Display

☒ Correlation factors

☐ P-values

9. Switch to the Analytes tab



Select data

Data split

Imputation

Normalization

Analytes

Report/Export

More

This tab shows a table with an overview of the antibodies. For each analyte, the minimum, mean, maximum intensity is shown together with the standard deviation. In addition, the column **N_imputed** gives the number of imputed values for each analyte, **Removed** tells you if the analyte was removed due to having too many imputed values (depending on your earlier settings), and **CV_ab** will display if the analyte was included in the panel used for CV normalization (if used). The table can be sorted by clicking the arrows next to the column names.

Analyte	Min	Mean	Max	SD	N_imputed	Removed	CV_ab
CA199A	1300	4908	40231	2423	0		Yes
CA199B	1345	4613	39368	2478	0		Yes
PC001	163	1200	4897	593	0		Yes
PC002	556	3002	14799	1367	0		Yes
PC003	703	5658	55432	3705	0		

10. **Switch to the Report/Export tab.** When first opened, you are required to select a user and enter a run ID. A run description can also optionally be entered. This is recommended to do. You also have the option to download the preprocessed and normalized data in the form of a CSV file by clicking the **DOWNLOAD NORMALIZED DATA** button.

 DOWNLOAD NORMALIZED DATA

Select user

Who to notify after completed BE run

Select user

Run ID

Assign a run ID

Run description


General description of run and content

After doing so, a number of options for the subsequent BE run will appear:

- a. **Number of CV-splits (parallelized):**
“Outer loops” in the backward elimination. The dataset is split into N parts, where N-1 parts are used for training.
- b. **Number of repeats (parallelized):**
The number of time each cross-validation will be repeated in each elimination round.
- c. **Number of inner folds (sequential):**
The number of splits within each outer fold.
- d. **Correlation cut-off:**
For antibody pairs with correlation factor higher than the selected thresholds, the one with highest correlation factor to any other antibodies will be eliminated. The number of antibodies that will be removed is displayed just below the slider.
- e. **Error metric:**
The metric to optimize against during BE. K-L (Kullback-Leibler) or ROC (ROC AUC).
- f. **Number of cores:**
The number of (virtual) cores to be used during the BE run. To save time, the outer loops and the cross-validation repeats are performed using parallel processing. Thus, an optimal number of cores to use is calculated by multiplying the two, which is done automatically. This number should generally not be changed manually.

A plot is generated to display the data splits. The bar heights represent the proportion of samples that are kept as validation data for each fold. It should be remembered that this is still within the training data, and that the test set withheld in the **Data split** tab is not included in the BE at all.

 GENERATE REPORT

With all settings made, press  GENERATE REPORT. This will do two things – first an HTML report is rendered and downloaded to your local workstation. In addition, the normalized data together with the settings for BE will be saved on AWS. This will in turn trigger a server with a suitable number of processors to be spun up and execute the BE script. An e-mail will be sent to the selected user within a few minutes to confirm this. Depending on BE settings and the size of the data set, this process will run up to several hours. After completion, the results will be stored on AWS and another e-mail set to the selected user as notification.

11. HTML report

The file name of the report defaults as:

<Date>_<Number>_<RunID>_<Dataset name>_normalization.html. The date and run ID will remain also during the downstream signature development steps.