**READ ME**

**PeakScanner\_peakFilter\_v3\_5\_1**

**Version 3\_5\_1**

**Authors: Christoph Metzendorf and Eugenia Mazzaferro**

**Objective**

To determine the fragment sizes, peak heights and relative wild type-peak area of potentially CRISPR/Cas9-edited samples of interest (unknown genotype), based on peaks observed in wild type control samples obtained using PCR products run through capillary electrophoresis on an Applied Biosystems® 3730XL DNA analyzer (Applied Biosystems, Waltham, USA) and data processed by Peak Scanner version 2.0 (ThermoFisher, Waltham, USA).

**Prerequisites and software requirements**

**Hardware requirements**

PeakScanner\_peakFilter\_v3\_5\_1 only requires a “normal” desktop or laptop computer with enough RAM to support operations.

**Software requirements**

**OS requirements**

This script is supported for macOS and Windows. The script has been tested on the following systems:

* macOS Ventura 13.5
* macOS Catalina 10.15

**Software**

R version 4.1.0

RStudio 2022.07.2+576

Peak Scanner version 2.0 (also tested on Peak Scanner version 1.0)

R packages and dependencies:

* stringr
* tidyverse
* viridis
* plotly

A loading step for all required packages is included on the script.

**Installation guide:**

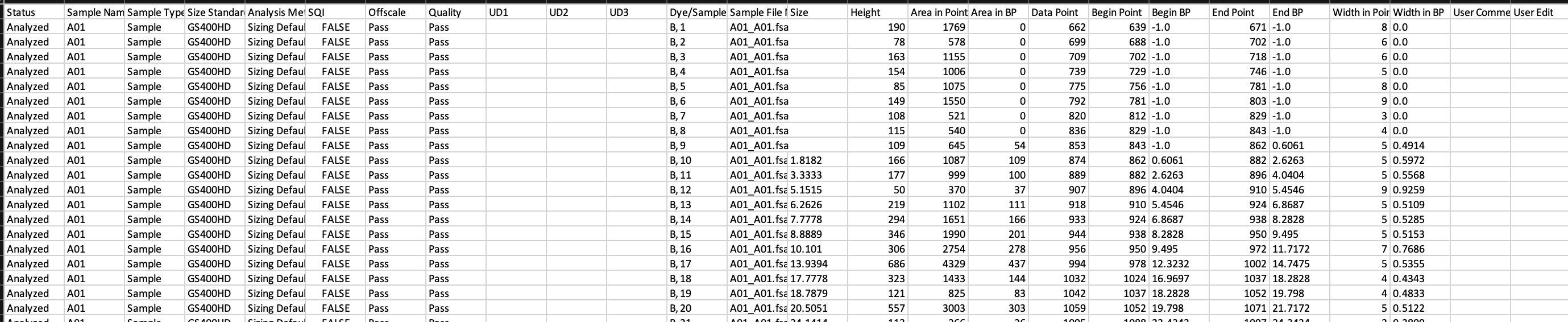
No installation required. Open the .Rmd file using R Studio.

**DEMO**

In this demo, files from the "**Targeting *kita* exerts limited effects on cardiometabolic traits**" experiment were selected. Transgenic larvae with fluorescence labelled cell types of interest were micro-injected with sgRNAs targeting de *kita* gene at either site 1 (*kita-*T1) or site 2 (*kita-*T2). Un-injected sibling controls were collected as Wild Type samples. A multiplexed PCR approach was used to amplify the two different targeted DNA regions tagged M13 with two different fluorophores (M13\_atto for *kita-*T1 and M13\_fam for *kita-*T2) The steps to prepare the FLA are described in the main manuscript.

**List of files in repository:**

1. Scripts: You can find two versions of the script. One designed for *kita*-T1 and a second one for *kita*-T2, since the amplicons are labelled by different fluorophores and can be analyzed separately.
2. Raw data: Provided are raw files obtained from DNAanalyzer. There is a .fsa file for each well (1.Raw\_files/).
3. Input folder: Contains files needed to run the script (input/).
   1. Processed DNA data or Fragment Length Analysis: Raw data were analyzed and exported into a .txt file using Peak Scanner version 2, which is the main input file for the script (Fig. 1). (input/20220404\_MDH\_kitaT1vsT2\_mpompeg\_PLATE01.txt)
   2. Plate layout: Layout of the 96-well plate that indicates 1) the sample ID for each unknown well and; 2) the location of wild type samples and blanks (Fig. 2). (input/20220404\_MDH\_kitaT1vsT2\_mpompeg\_PLATE01\_PLATE\_LAYOUT.txt)
4. Output folder (output/): Contains results from the two scripts in different formats:
   1. OVERVIEW: provides a visual overview of each step performed by the script (Fig. 3). (output/ 20220404\_MDH\_kita-T1vsT2\_mpompeg\_PLATE01\_FLA\_OVERVIEW\_T1\_atto.html)
   2. DATA-FRAME: An intermediate file that provides a dataframe with the samples used for the analysis. (output/20220404\_MDH\_kita-T1vsT2\_mpompeg\_PLATE01\_peakScanner\_DF\_F\_DATA-FRAME\_T1\_atto.txt)
   3. RESULTS: A .txt file providing – for each sample ID and for each peak – the indelsize, total area, label, number of peaks, presence of WT peaks and relative WT area (i.e., the fraction of total peak area that belongs to WT peaks) (Fig. 4). (output/20220404\_MDH\_kita-T1vsT2\_mpompeg\_PLATE01\_RESULTS\_T1\_atto.txt)

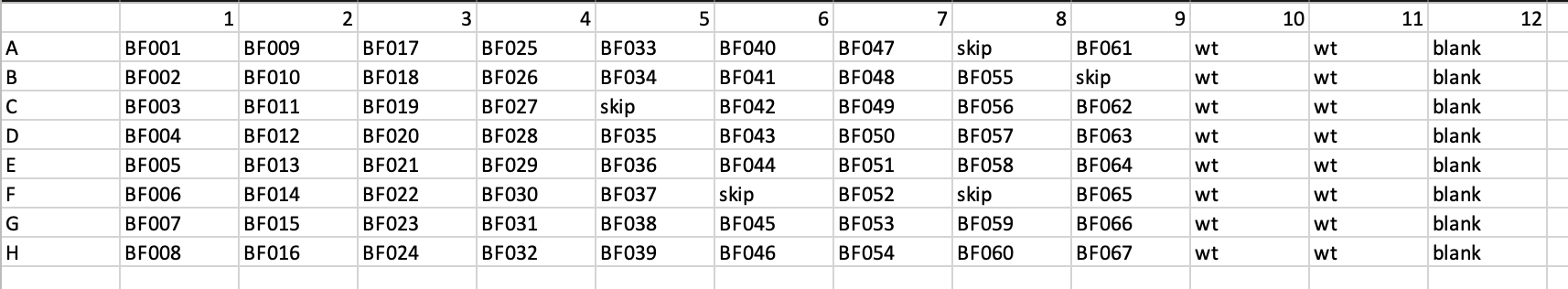


**Fig 1.** Example of the PeakScanner exported file (input file for the script). It contains information on all wells and all peaks detected on each of the 4 fluorophores available.

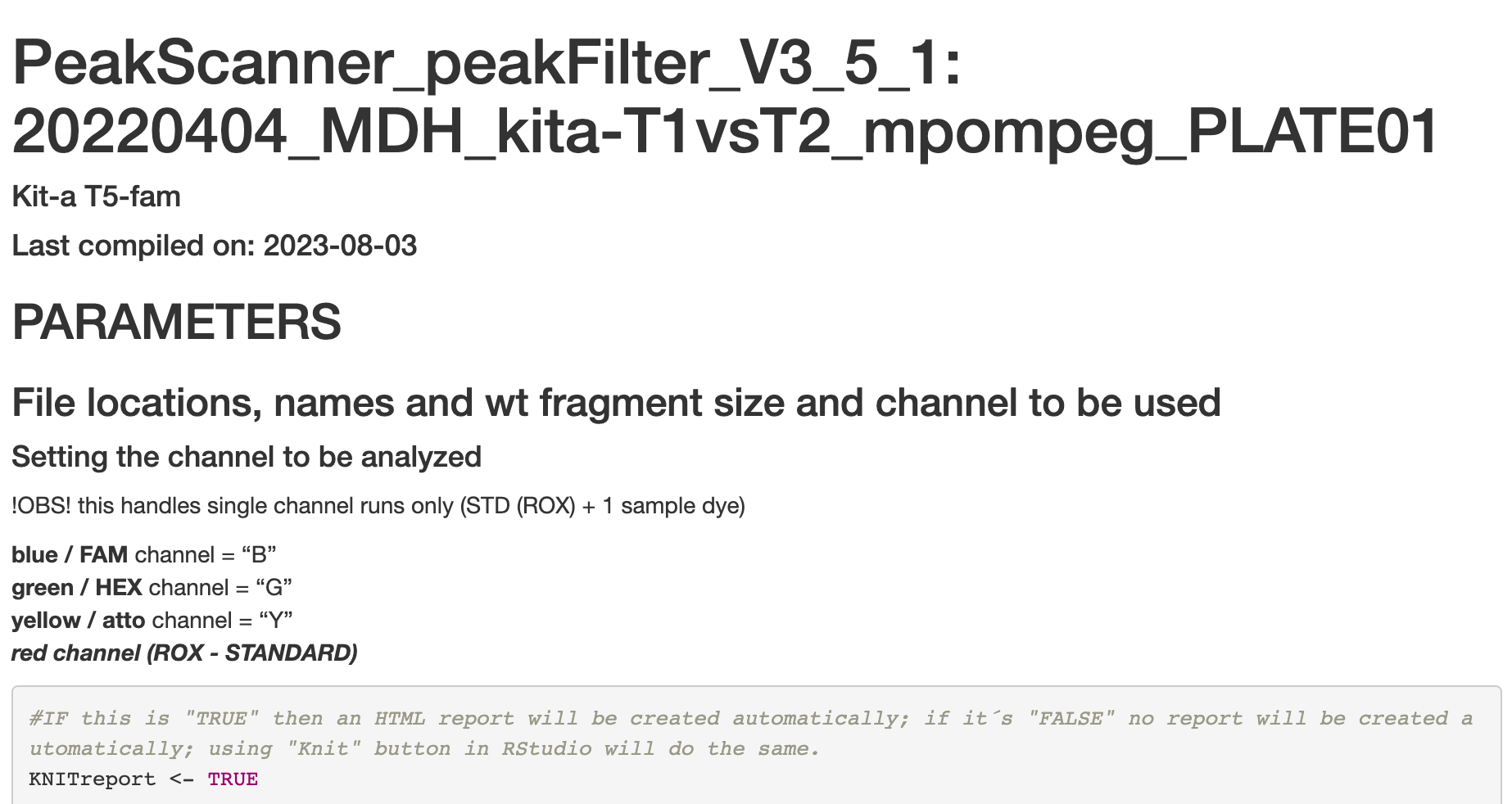
**Notes on preparing the input files for your own data:**

The plate layout file should contain the format of a 96-well plate and each well can accept one of the following categories:

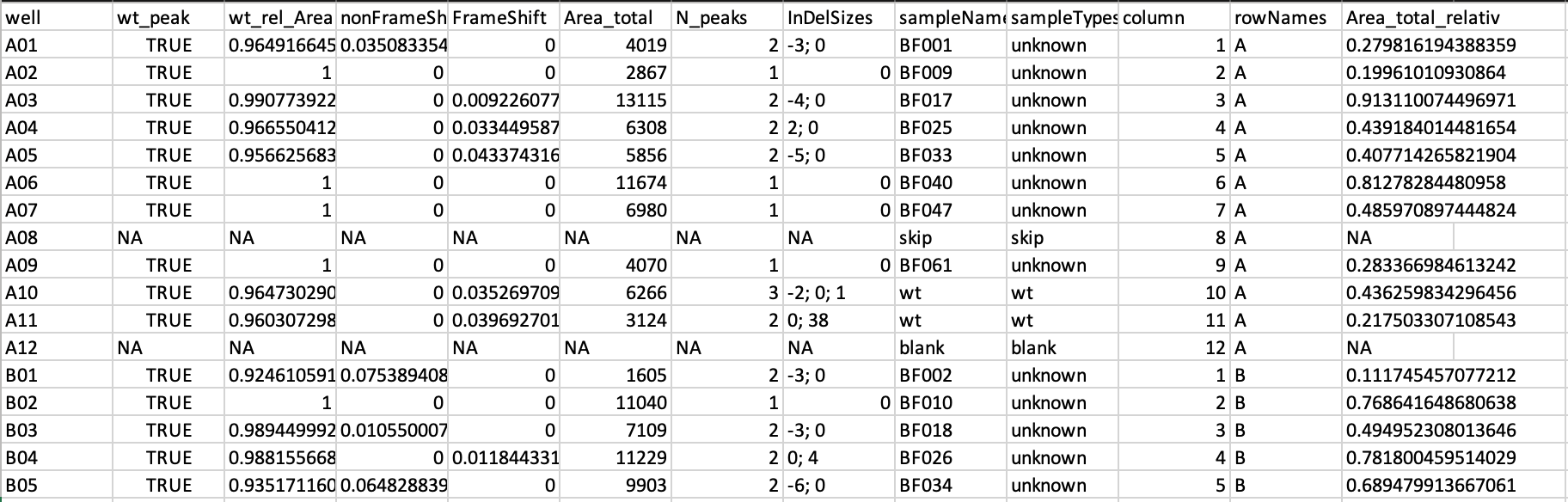
* A unique code identifying the sample (e.g., **BF001**)
* **“wt”** to indicate the position of negative controls (un-injected larvae)
* **“blank”** to indicate the position in the plate of wells loaded with HiDi and ladder without DNA
* **skip** if the well is empty



**Fig 2.** Example of the PLATE\_LAYOUT file



**Fig 3** Example of the OVERVIEW file

****

**Fig 4** Example of the RESULTS file

**Running the script in demo dataset and own data**

Expected run time for demo on a “normal” desktop computer: 3 minutes

**Steps:**

* Before starting, make sure you have an input folder containing the two files per plate to be analyzed (PeakScanner exported data and plate layout), and an output folder in the same directory as your script.

The script runs a total of 9 steps:

1. **Select the channel to be analyzed ("B" blue/FAM, "G" green/HEX, "Y" yellow/ATTO).** The script can only analyse one channel per time. In these demonstrations, two scripts are provided: PeakScanner\_peakFilter\_v3\_5\_1\_kita-T1\_atto\_211.Rmd to analyze the ATTO ("yellow" channel) PeakScanner\_peakFilter\_v3\_5\_1\_kita-T2\_fam\_302.Rmd to analyze the data for the FAM ("blue" channel).

Specify the file name on the PeakScanner output file and provide the theoretical wt peak size (*WTsize*), which in this case is expected to be 211 for *kita-*T1, and 302 for *kita-*T2.

1. **Parameters setting.**  All scripts are performed by running the standard settings:

* *WTsize margin* (default = 0.4). A value between 0 and 0.4 should be included. It defines how wide the size range should be around the 'wt size'.
* *IgnorePeakSize* (default = -1000). List of peaks to be ignored by the script (Only include peaks to ignore if they are present in sibling wt samples and are not the main WT peak)
* *smallSizeCutOff* (default = 100). It ensures that too small fragments are not included in the analysis (e.g., primer dimers). Everything below 100 is excluded.
* *smallPeakHeightCutOff* (default = 800). It sets the minimum value of "acceptable" peak height. Everything above 800 is included.
* *manualPeakHeightCutOFF* (default = TRUE). when set to “True”, it uses the value provided by the user; when set to “False”, the Peak Height CutOff is determined by experimental data.
* *smallPeakHeightCutOff\_quantile* (default = 0.99).
* *Size\_span* (default = 50). Set the window of peak sizes to be considered, by default ±50 the size of the wt peak.

1. **Reading of data files.** The PeakScanner exported files and the plate layout files are read by the script.
2. **Determination of noise peak**. Noise is defined as the presence of peaks in the wells labelled as "blank" in the plate layout file. If any peak is detected across the "blank" samples, this is used as a new threshold to exclude artifacts in the other samples. If no peak is detected, the manually selected *smallPeakHeightCutOff*  is used.
3. **Filtering and processing data.** From the PeakScanner exported file, only the samples with the predefined channel are maintained. Excluded are: i) all samples with missing values in the size column; ii) peaks smaller than the *smallPeakHeightCutOff; iii)* peaks outside the *Size\_span*; peaks included in the *IgnorePeakSize* list; and iv) peaks corresponding to samples labelled as "skip".
4. **Determining relative PeakArea and TotalPeakArea per sample.** Using the filtered dataset, the following parameters are calculated for each peak in each well: Area.in.BP.RelToMax, Area.in.BP.RelToSum and Area.in.BP.Sum or total peak area.
5. **Estimating the size range of Wt-peaks.** Peak sizes of all samples are rounded to full base pairs. In samples labelled as "wt", the highest peak with a fragment size of ±1 bp of the theoretical size is selected to be the wt peak. If two peaks are within this size range, the peak closest to the theoretical wild-type fragment size is selected. The median fragment size across all the samples labelled as "wt" is assigned to be the wt size and *WTsize* is updated if it differs. Afterwards, it applies this new\_wt size to detect WT peaks in all samples and identify non-wt peaks. The fragment size distribution amongst wt samples and unknows are summarized in plots.
6. **Calculating the size of indels.** All non-wild-type peaks are assigned to be in-frame or frameshift, depending on whether the difference in bp between the peak and the theoretical wild-type peak is a multiple of 3 or not.
7. **Summarizing data and plotting.** The relative wt area is calculated for each well. A series of visualizations illustrate the performance of the analysis, and a visual representation of genotype by well is provided. (**Fig. 5**).

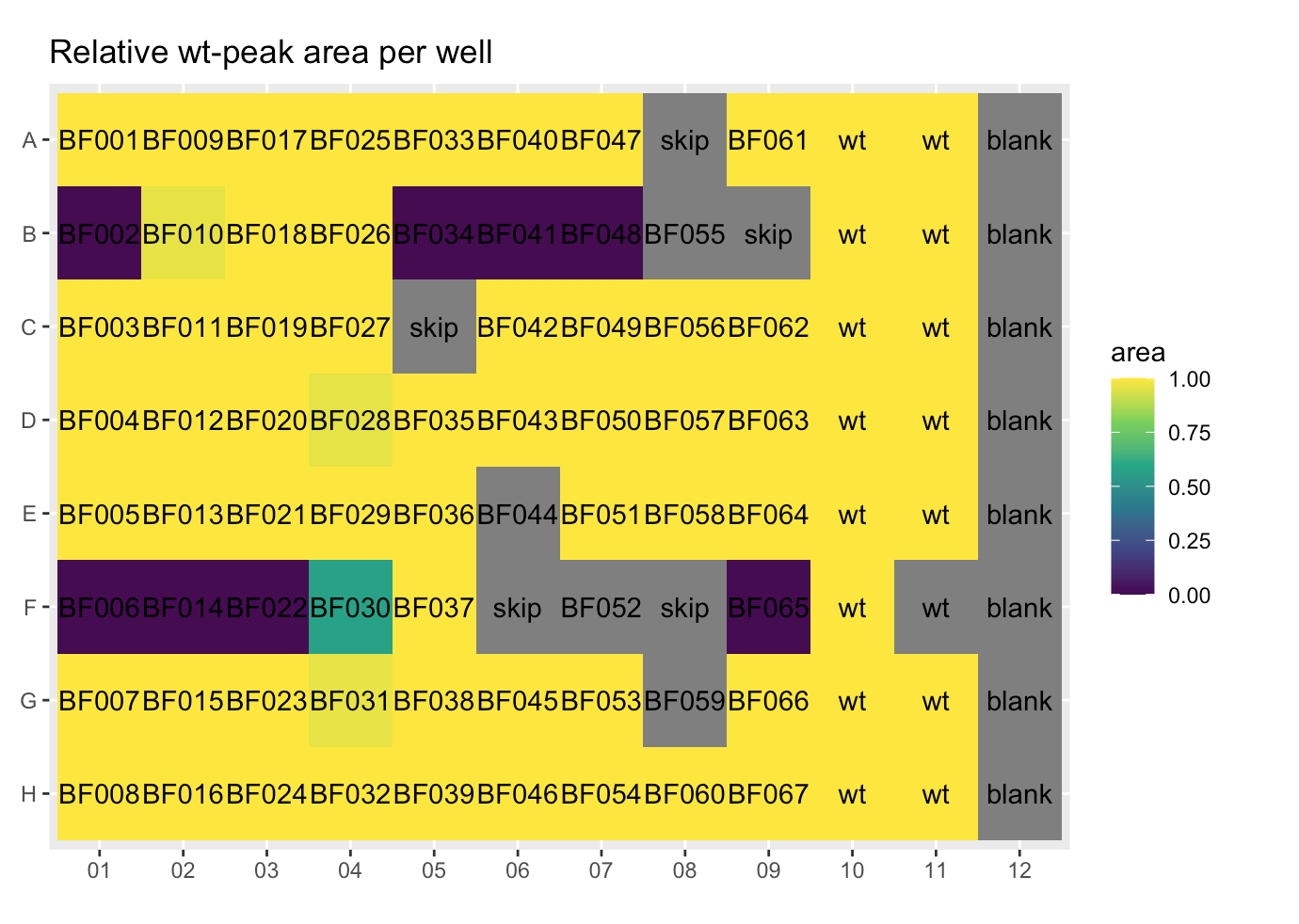
**Instructions how to run software on your own data:**

-Prepare the input files and directories as explained above

-Open the script and only modify steps 1 and 2 from above according to your own data requirements.

-Run all

Note: Multiple plates of the same experiment can be analyzed simultaneously and will be compiled in the same output file.

****

**Fig 5.** Example of visual representation of the presence/absence of wt peaks in each well in a 96-well plate. Yellow indicates the relative wt-peak area per well is closer to 1. Purple shows the relative wt-peak area is closer to 0 (no wt-peak). Grey denotes samples that were excluded from the analysis due to low quality of PCR product or as indicated in the layout.