**Guideline to run RIMA**

Contents

[Introduction 2](#_Toc516228841)

[NGS data analysis 2](#_Toc516228842)

[CLC Genomics Workbench NGS analysis workflow 2](#_Toc516228843)

[Understanding RIMA 5](#_Toc516228844)

[Installation and requirements 5](#_Toc516228845)

[Access the codes 5](#_Toc516228846)

[Organization of worksheets 6](#_Toc516228847)

[Graphical interface (“Single” worksheet) 7](#_Toc516228848)

[A: user inputs and settings for graphical representation of variants 8](#_Toc516228849)

[B: Table of variants 12](#_Toc516228850)

[C: Control buttons 12](#_Toc516228851)

[Inputs 14](#_Toc516228852)

[Single run analysis 15](#_Toc516228853)

[Step-by-step example for batch analysis 15](#_Toc516228854)

[Future directions 21](#_Toc516228855)

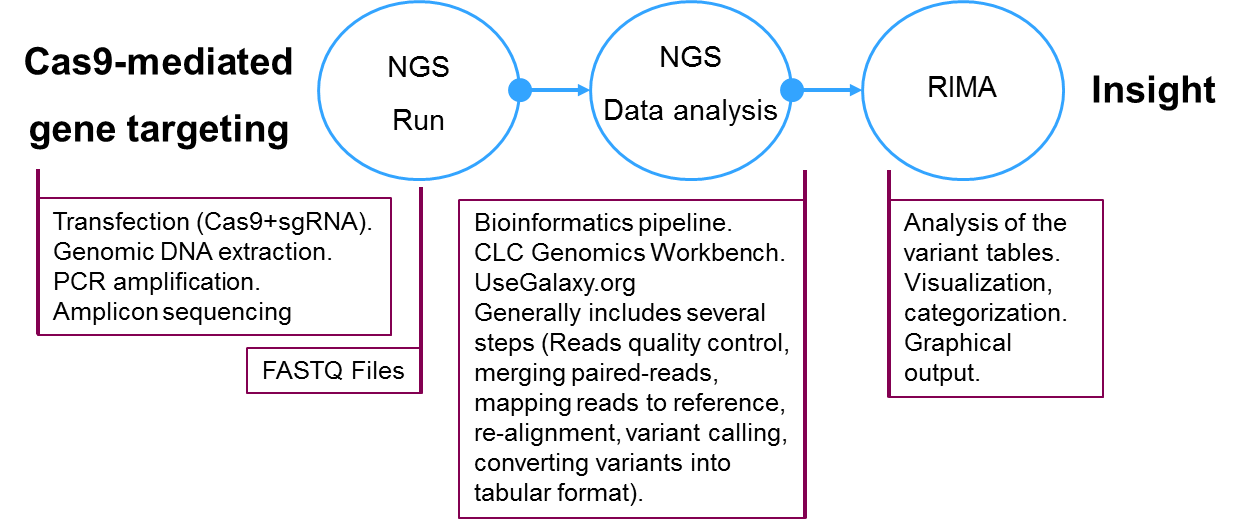
[How to cite RIMA 21](#_Toc516228856)

# Introduction

Double strand breaks (DSBs) in mammalian cells are the most deleterious DNA lesions that if remained unrepaired can cause cell death or genome instability. A variety of programmable nucleases can be used in the context of genome-editing to induce DSBs into the human genome in a targeted manner. These nucleases include Zinc Finger Nuclease (ZFNs), Transcription Activator Like Effector Nucleases (TALENs) and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-Cas9 system. The latter becomes the most versatile system applicable for genome-editing in many organisms including bacteria, plants, animals and human cells.

Repair of a Cas9-induced cut in human cells results in introduction of non-random mutations at the targeted site. Variation among the mutational signatures at different target sites can be due to many reasons. However, studying the mutational signatures and their variations could results a better understanding of the repair pathways, kinetics of the cleavage and possible interactions between the DNA lesions (as substrate) and the DNA repair pathways.

Rational Indel Meta-Analysis (RIMA) is a tool that has been developed in the course of our study and would allow biologists to analyse their data without prior Bioinformatics knowledge. A general workflow for studying the mutation signatures is provided below (**Figure 1**):



**Figure 1.** Schematic representation for a general workflow to study the mutation patterns after a Cas9 cut in human cells.

# NGS data analysis

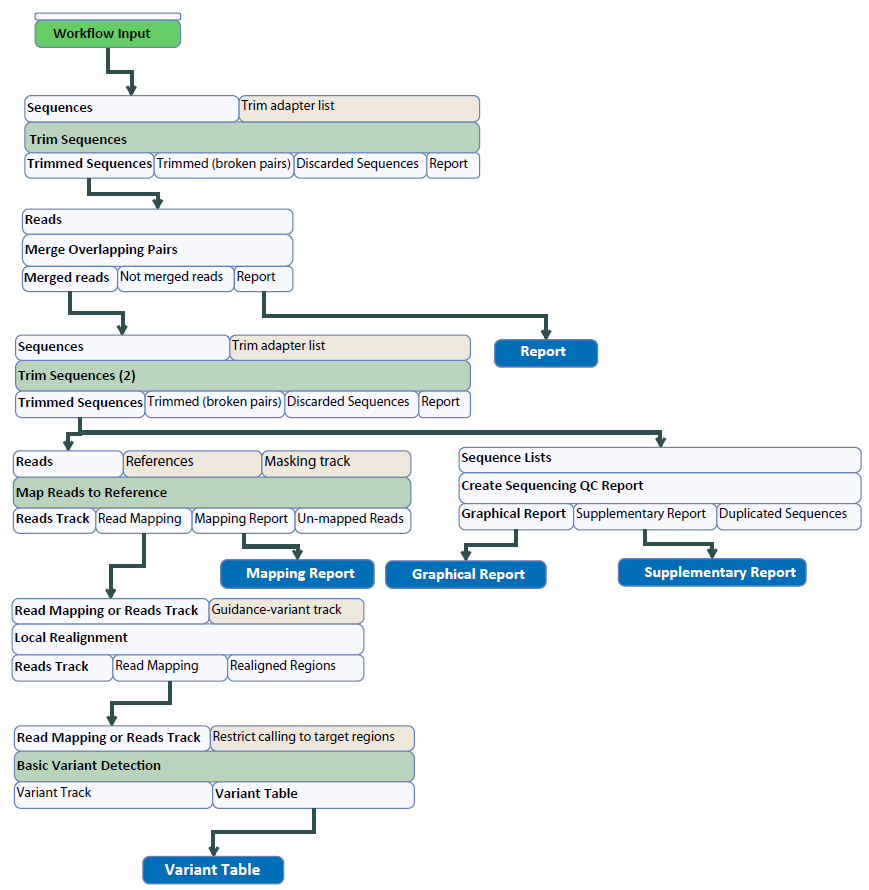
The amplicon next generation sequencing (NGS) can generate thousands to million reads per sample. The raw data is then output as FASTQ files, which contain both the sequence and the sequencing quality for each base. Analysis of the FASTQ files requires a multistep workflow. Different workflows can be used to detect the mutated reads and quantify them. Here we provide guidelines to setup workflows in CLC Genomics Workbench that do not require prior bioinformatics knowledge.

## CLC Genomics Workbench NGS analysis workflow

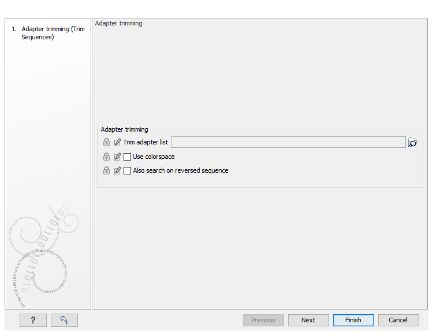
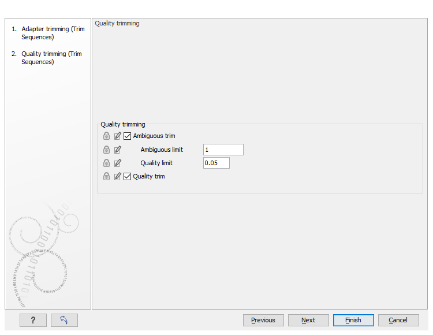
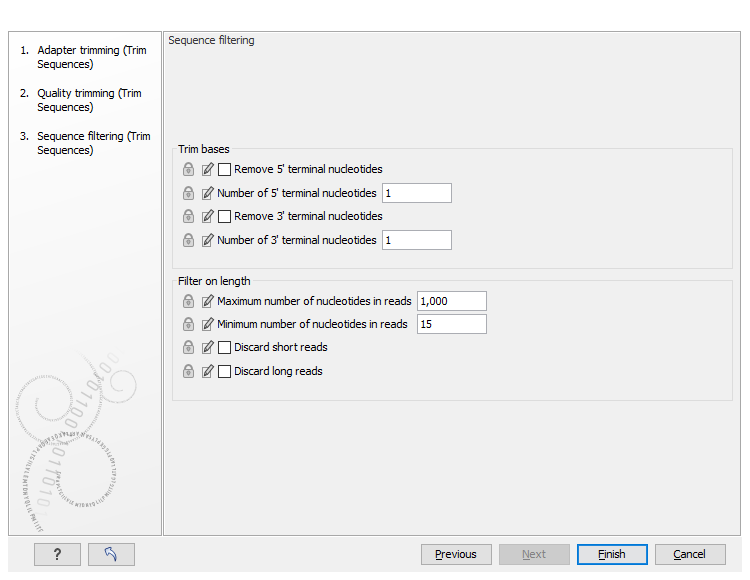
Latest version of CLC Genomics Workbench can be purchased from QIAGEN at:

<https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/>

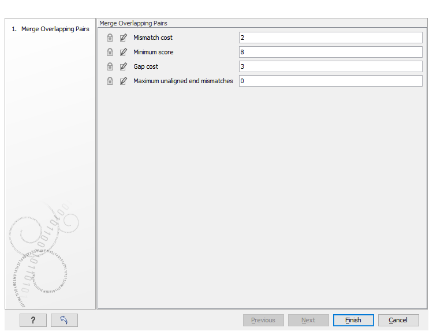
For information about creating and running workflows, please see the user manual at: <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Workflows.html>



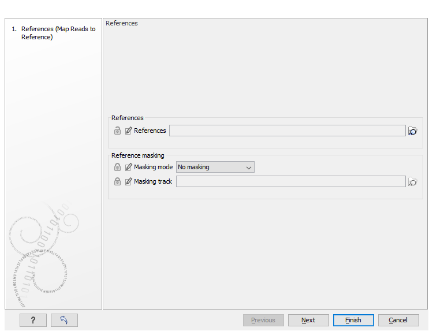
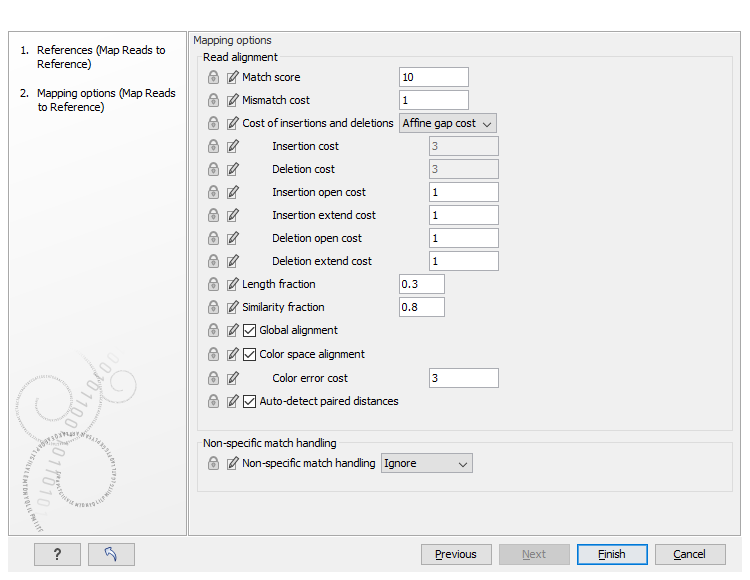
**Figure 2.** Example of a CLC Genomics Workbench NGS analysis workflow for paired-end reads. The workflow starts with reads as input. Screenshots for each of the steps and their settings is shown in the following figures.

🡪🡪

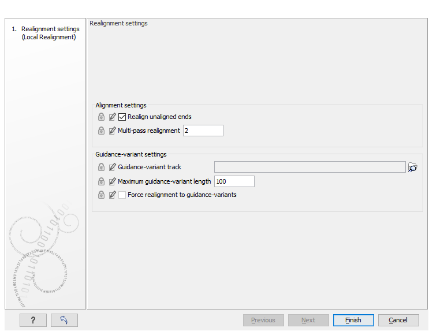
**Figure 3.** Configuration of the “Trim Sequences” step in CLC Genomics Workbench workflow.



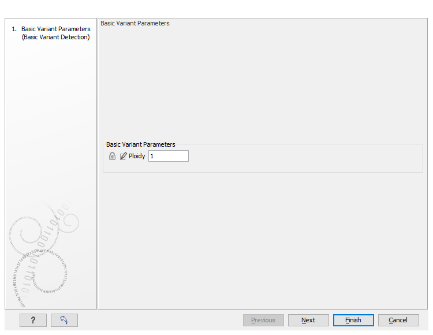
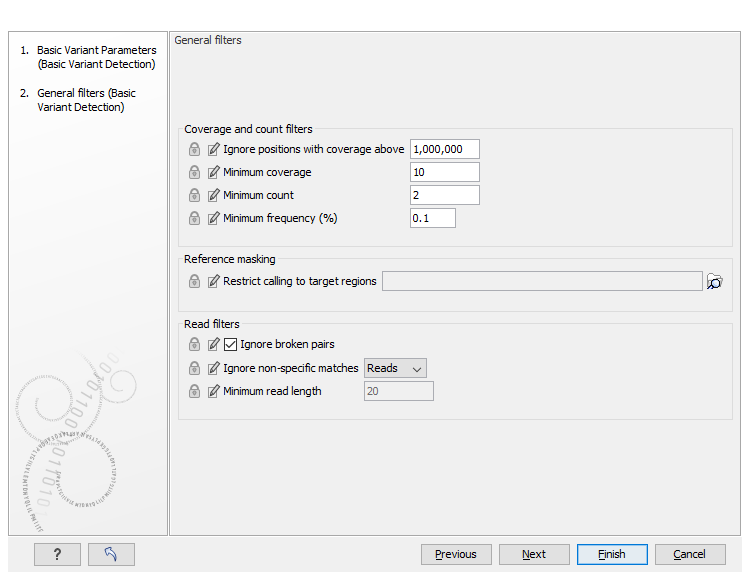
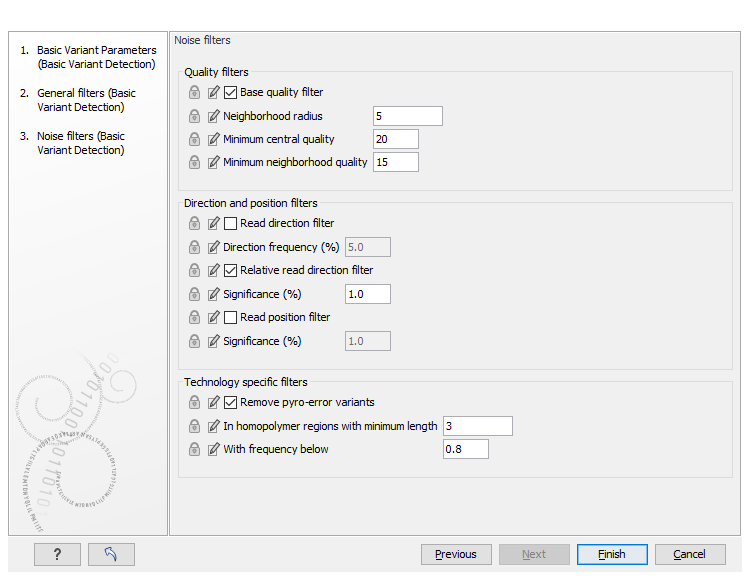
**Figure 4**. Configuration of the “Merge Overlapping Pairs” step in CLC Genomics Workbench workflow.

🡪

**Figure 5**. Configuration of the “Map reads to reference” step in CLC Genomics Workbench workflow. A fasta sequence must be provided at the first window as the reference sequence. The same sequence must be used later for RIMA analysis.



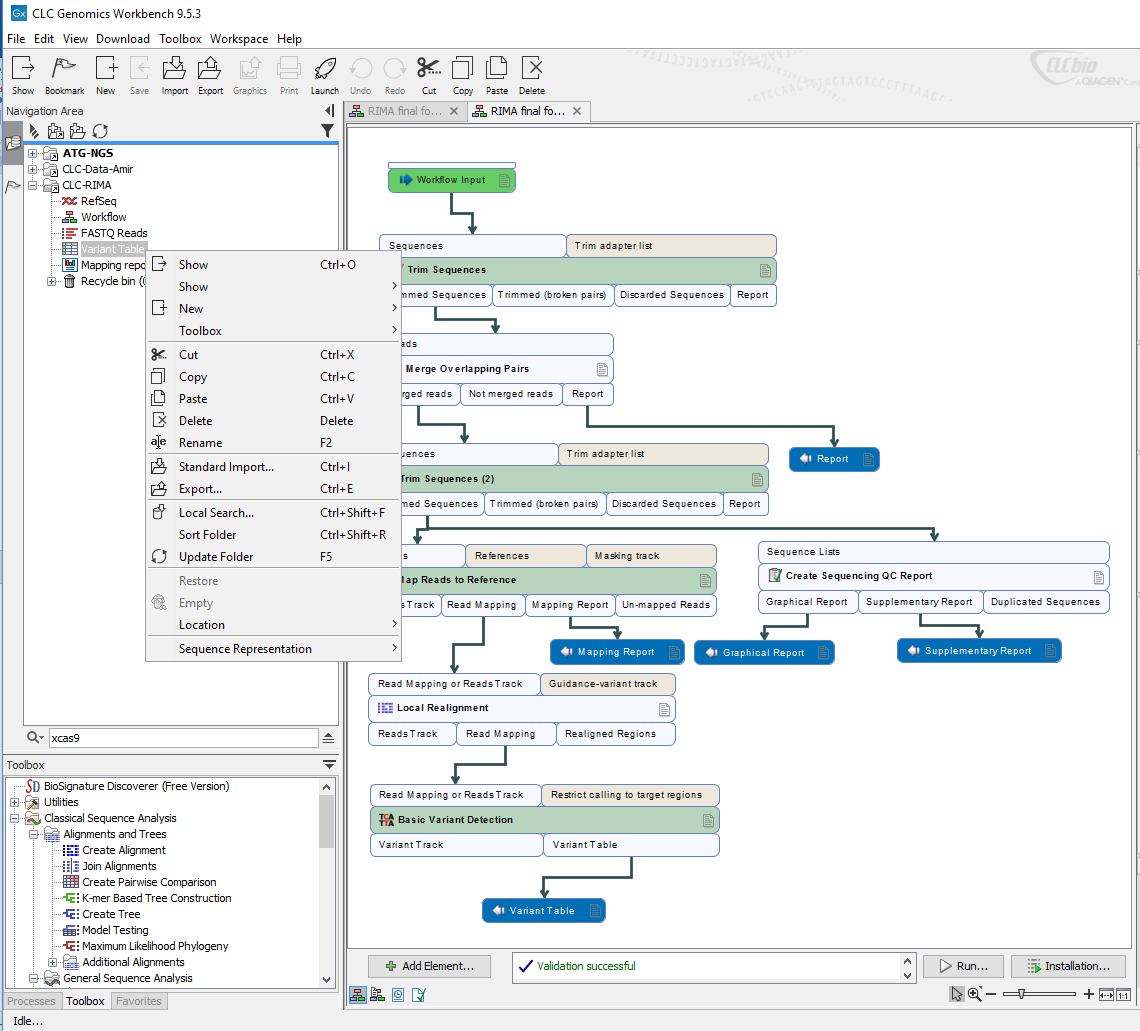
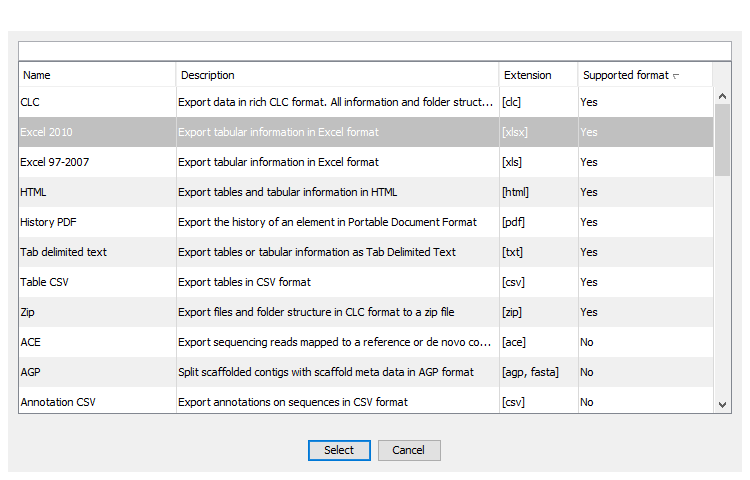
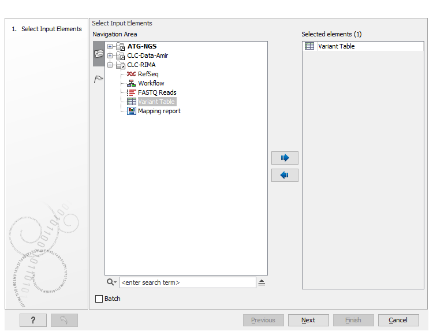
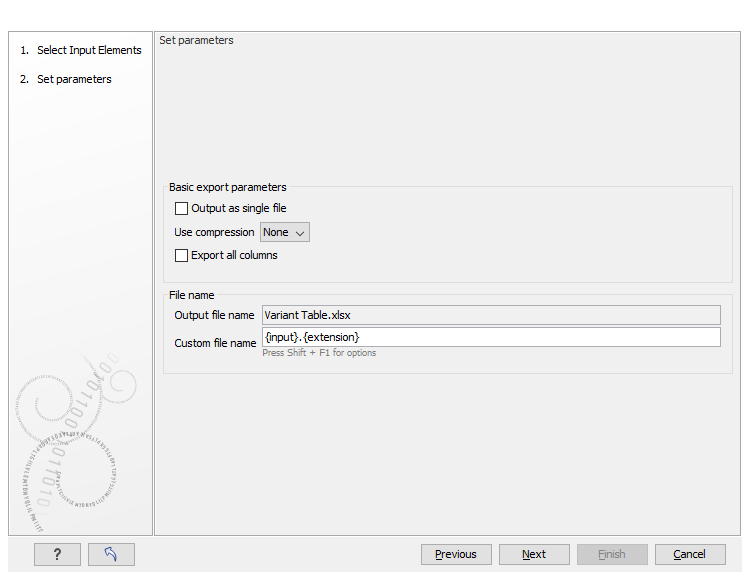
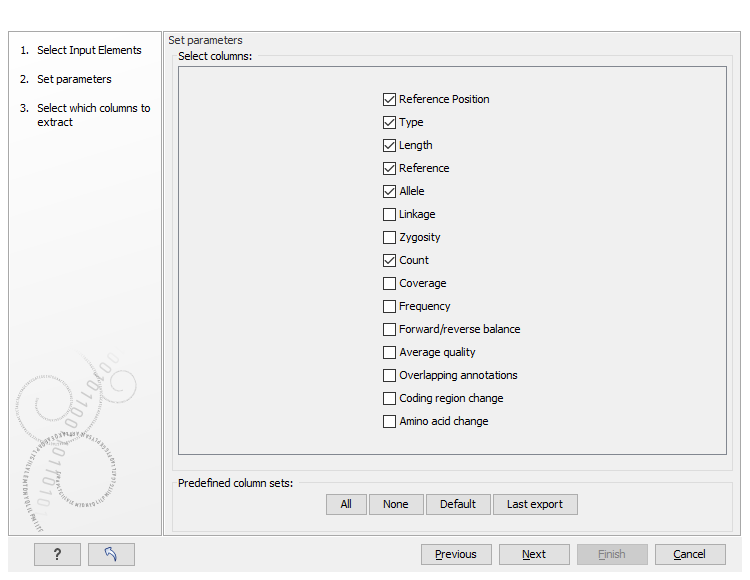
**Figure 6**. Configuration of the “Local Realignment” step in CLC Genomics Workbench workflow.

🡪🡪

**Figure 7**. Configuration of the “Basic Variant Detection” step in CLC Genomics Workbench workflow.

The CLC NGS analysis workflow shown in the example above will generate several reports for each step, which can be very helpful to control the outcome of each step. For example, how many reads were in the dataset, how many reads were quality trimmed, how many reads were successfully merged, how many reads were mapped to the reference (this can be used later to calculate the percentage of modified reads), and finally what are the detected variations in a tabular format.

Screenshots for exporting the variant tables as Excel 2010 are shown in **Figure 8**:

 🡪 🡪 🡪 🡪

**Figure 8**. Exporting variant tables as Excel 2010 from CLC Genomics Workbench. For RIMA to be to analyse the variant table, it is important that the output be in the format that is shown in this figure. Right click on the variant table in the Navigation area and select Export. On the following windows, deselect the “Export all columns” and press next. In the Select columns window select only the options that are shown on the last panel, and save the file on your computer.

The NGS analysis workflow can run for both individual FASTQ file analysis and batch mode, which in this case separate reports and graphical reports will be generated per each batch unit. Variant tables then can be exported as Excel 2010 (shown in **Figure 8**).

# Understanding RIMA

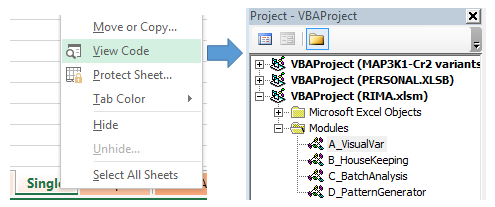
RIMA is a software for the analysis of the mutation patterns at CRISPR-Cas9 targeted sites. Mutations detected in deep targeted next generation sequencing (NGS) data can be analysed using RIMA for different purposes. The algorithm allows the quantification of classical-Microhomology mediated end joining (c-MMEJ) repair events after a Cas9 cut. It also allows the categorization of insertions, duplications and deletions according to their length and positions. RIMA offers a “batch mode” which enables analysis of hundreds of NGS runs in one go. The purpose of this document is to provide the essential information on how RIMA works as well as a step-by-step guideline to run RIMA.

## Installation and requirements

RIMA is a stand-alone Excel Workbook and does not require any specific installation. However, it is essential to make sure that the file has the “.xlsm” as the file extension and will be opened as Macro-enabled. It has been validated using Microsoft Excel 2013 and 2016 installed on Windows operating system.

## Access the codes

RIMA’s codes are written in Excel Visual Basic for Application (VBA) language. All the codes are available in the developer mode of Microsoft Excel Program. To enable the developer mode and accessing the codes, you must configure it to show the **Developer** tab as it does not appear by default (see: <https://msdn.microsoft.com/en-us/library/bb608625.aspx>). Alternatively, pressing Alt+F11 opens the “Visual Basic for Application (VBA)” window and the list of the Modules and their codes will be accessible. It is also possible to right click on one of the tabs inside the workbook and select “View code” (**Figure 9**) to reach to the VBA window.



**Figure 9.** Accessing the codes in the VBA window. RIMA’s codes are categorized into four Modules A, B, C and D.

***Note***: to run RIMA there is no need to be familiar with the codes and programming. All the analysis can be performed using the command bottoms provided in the Excel sheets. Refer to the below sections for detailed info.

## Organization of worksheets

A screenshot of the opened RIMA.xlsm file is shown in **Error! Reference source not found.**. Five worksheets in the main workbook include Experiment, Single, Samples, BatchAnalysis, GroupAnalysis, HeatMaps and info. Details and a step-by-step example is available in the following sections of this document.



**Figure 10**. Screenshot of the RIMA Excel workbook containing 8 worksheets.

**Single**: The main interface of RIMA is built in this worksheet. All the analysis for an individual table of variants is done in the Single worksheet. During a batch analysis, a copy of this worksheet after analysis will be saved locally on the computer. Once the whole set of variant tables are analysed, then it will be possible to re-open all the Single worksheets and read the results and collect them back into the main workbook.

**Experiment**: Use this worksheet to list all the table of variants that needs to be analysed in one go. It is possible to browse files and provide RefSeq, sgRNAs, and the number of mapped reads associated with an individual table of variants.

**Samples**: List and address of the analysed Single worksheets will be populated in this worksheet. Then it is possible to import the analysed data back to the main workbook.

**BatchAnalysis**: Once data from individual single files imported into the “Samples” worksheet, then it will be possible to collect and organize the results in a single table. This table then can be used for statistical analysis.

**GroupAnalysis**: Yet to be developed. This worksheet can be used for dividing al of the samples into sub groups and analyse them separately.

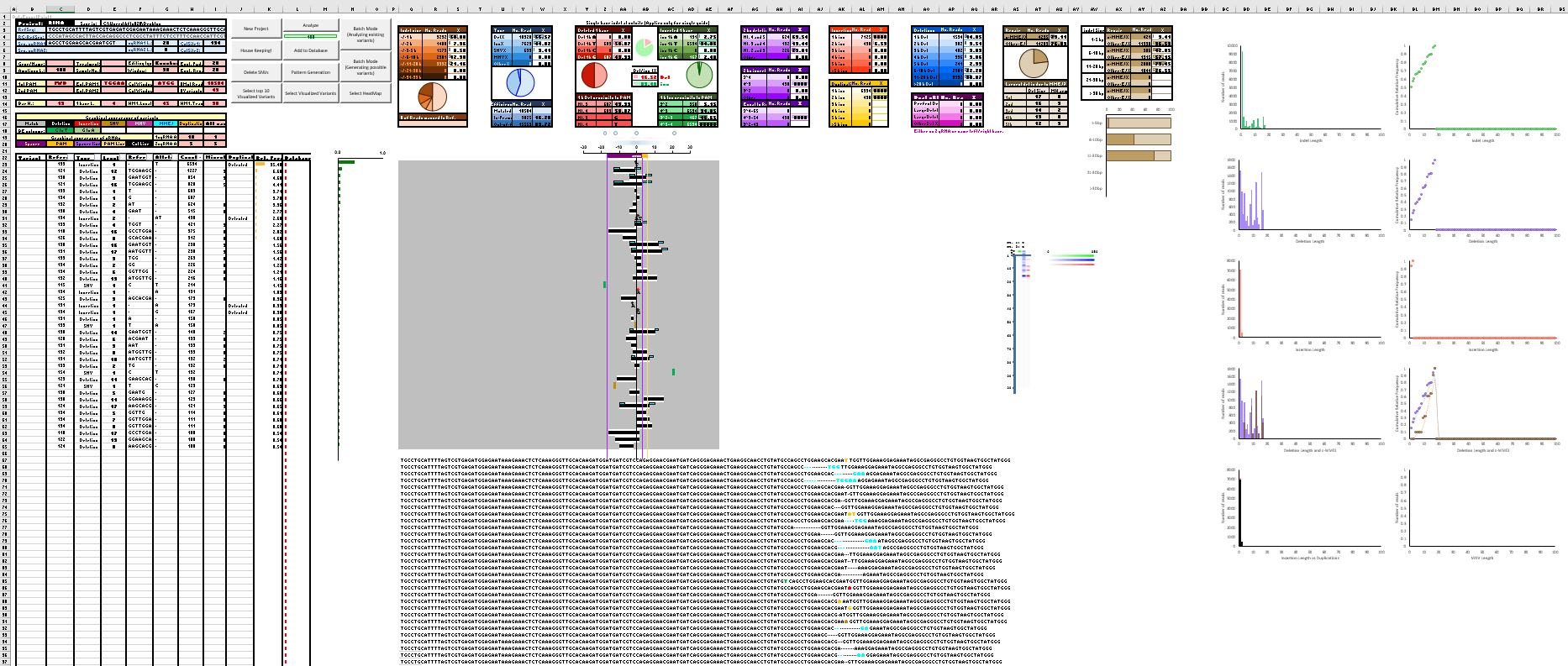
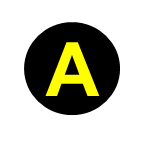
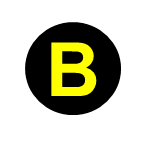
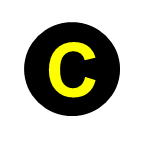
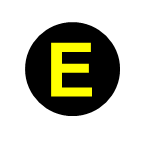
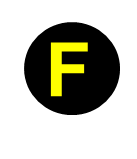
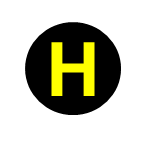
**Heatmaps**: For side-by-side of comparison of samples, a heatmap for selected samples can be generated using this worksheet.

**UseGalaxy**: A temporary worksheet to be used for analysing tabular VCF files generated by UseGalaxy workflows. This section is under development.

**Info**: Supporting information related to the development of the worksheet and citation is provided.

### Graphical interface (“Single” worksheet)

The main graphical interface of RIMA is built in the “Single” worksheet and consists of a few cells that are dedicated for inputs, some for setting the graphical appearance of the output, and some buttons associated with Macros that run the codes and perform the analysis. A large picture of an analysed table of variants is provided in **Figure 11**.



**Figure 11**. The organization of inputs, outputs and control buttons in the “Single” worksheet. Details for each section is provided below.

**A**: Inputs (RefSeq and sgRNAs) and setting for the graphical output.

**B**: Table of variants to be analysed.

**C**: Control buttons to run the analysis, housekeeping, select graphics, etc.

**D**: The graphical representation of the mutations.

**E**: An alignment of the mutations.

**F**: Categorization of the mutations based on type, size, length, etc.

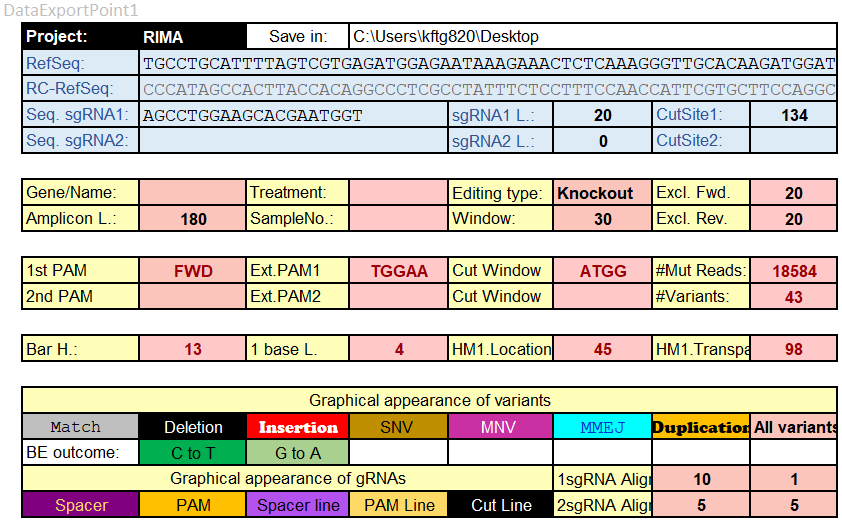
**G**: Heatmap representation of the length and frequency of the mutations.

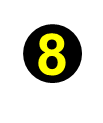
**H**: Distribution of the mutations based on their size, type, etc.

A brief description is provided below:

#### **A**: user inputs and settings for graphical representation of variants

#### 





**Figure 12**. Screenshot of the RIMA’s interface for RefSeq and sgRNA inputs, and graphical settings. Among all of the indicated fields, only 2 and 4 are required to be provided. A description for each field is provided below:

**1-** Provide a local address for saving the analysis reports.

**2-** Provide the RefSeq (reference sequence, 5´to 3´). This should be the same sequence that has been used for mapping the reads during NGS data analysis.

**3-** The Reverse Complement for the RefSeq. No need to fill, this will be converted from RefSeq during analysis.

**4 and 5-** Provide the sequence of the sgRNA (5´to 3´) without PAM. The orientation of the sgRNA will be determined by RIMA. This sequence of sgRNA1 is required for the analysis.

**6** **and 7**- The length of sgRNA1 and 2 will be calculated by RIMA.

**8 and 9**- The expected cut sites for each of the sgRNAs will be calculated by RIMA.

**10, 11 and 15**- Optional fields to indicate the name, treatment and sample number.

**12**- Purpose of the experiment. Knockout is the only option for the current version of RIMA.

**13 and 17**- Primer regions to exclude from the analysis. Mutations (if any) at the beginning or the end of the sequence will be removed from the analysis. This helps to avoid false-positive variant calls.

**14**- Length of the amplicon will be calculated by RIMA.

**15**- see 10.

**16**- The window around the cutsite (in base pair) for a mutation to be included in the analysis. Mutations out of this window will be removed from the analysis. Mutations spanning this region (e.g. large deletions) will be included in the analysis.

**17**- see 13.

**18 and 19**- The PAM sequences for sgRNAs will be calculated by RIMA.

**20 and 21**- The extended PAM sequences for sgRNAs will be calculated by RIMA.

**22 and 23**- Nucleotides at positions 2,3,4, and 5 before each of the PAM sequences. These sequences will appear on the graphical output.

**24**- Total number of mutated alleles, will be calculated based on the provided table of variants.

**25**- Total number of variants provided in the table of variants.

**26, 27, 28, and 29**- Graphical settings. See **Figure 13** for details.

**30**- Change the background colour and/or the font colour (applicable for Match, Deletion, Insertion, MMEJ, Duplication, MNVs, and SNVs) to set the colour of the graphical output.

**31**- Setting for limiting the graphical output to “only deletions”. See **Figure 14** for details.

**32**- Setting for the window of alignment shown as text. See



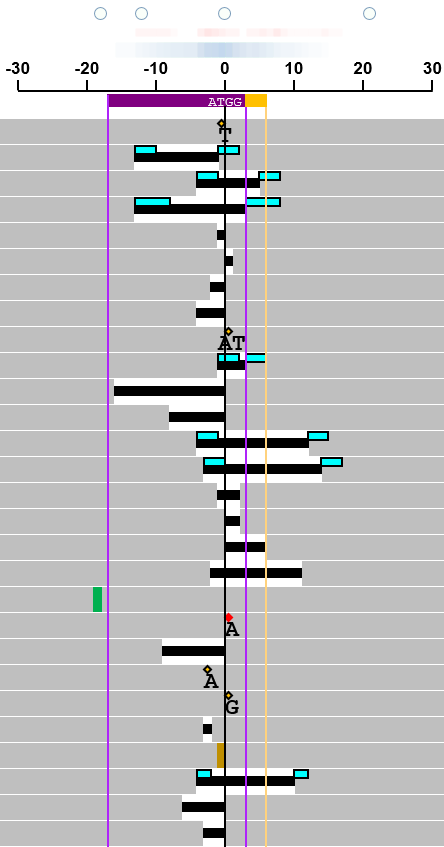
Transparency for deletion heatmap.

Heatmap for microhomologies.

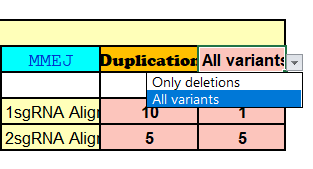
Distance.

Bar Height.

SNVs

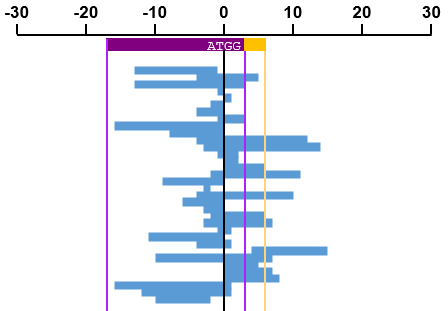


**Figure 13**. Graphical output for setting 26, 28, and 29 shown in **Figure 12**.

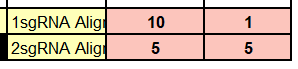


Select “Only deletions” and analyse the worksheet again to produce output below:





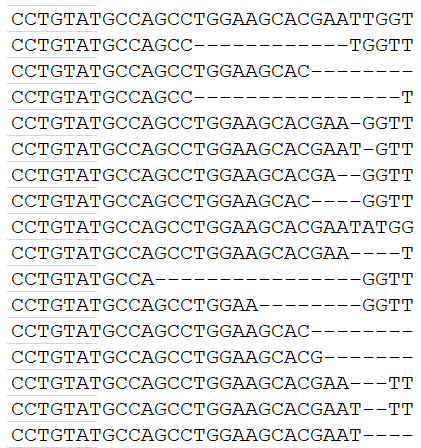
**Figure 14**. Setting for limiting the graphical output to deletions.





10

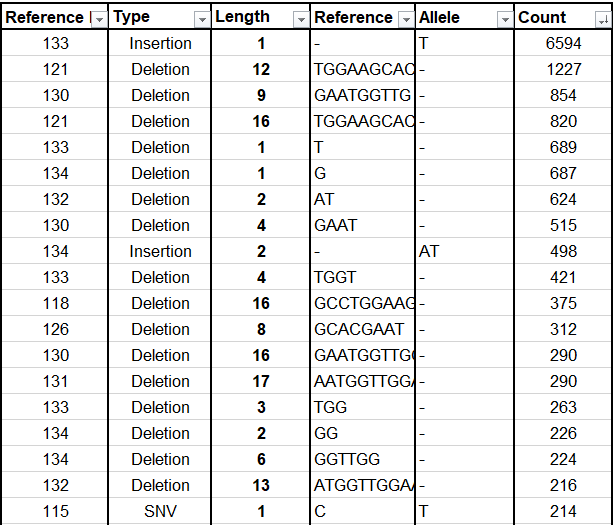
1



**Figure 15**. Setting for the alignment window shown for single (1sgRNA) or dual sgRNAs (2sgRNAs). Left number (10 in this case) is the total number of nucleotides shown before sgRNA start site and the right number (1 in this case) is the number of nucleotide after target site. For 2sgRNAs (where a pair of sgRNAs were used simultaneously) the left and right number (5 in this case) is total number before and after the first and the second sgRNA.

#### **B**: Table of variants

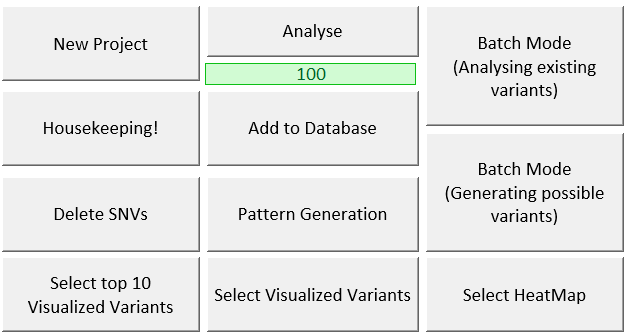
The table of variants needs to be provided in the same format as shown **Figure 16**.



**Figure 16**. The format for the table of variants required in RIMA.

#### **C**: Control buttons

Control buttons to run the Macros are shown in **Figure 17**.



**Figure 17**. Control buttons within the “Single” worksheet.

**New Project**: opens a dialogue box to provide a name for the project under analysis. The project name then will be used as the name of the database folder, where a copy of “Single” worksheets will be saved.

**Analyse**: if all the required inputs (RefSeq, sgRNA1, and Table of variants) are provided, then this button runs all the Macros written to analyse the data. It is not possible to use Excel program while the analysis is running. The green bar under the “Analyse” shows the progress of the analysis (100 is when the analysis is done).

**Housekeeping**: it is essential to clean all the previous data from the previous run before starting a new analysis. Housekeeping cleans all the fields that need a refresh.

**Add to Database**: this will save a copy of the current “Single” worksheet on the computer. The address of the saved file will be added to the “Samples” worksheet. This allows to analyse multiple table of variants manually (without using the batch mode).

**Delete SNVs**: this removes all the SNV variants from the table of variants.

**Pattern Generation**: this allows to generate a virtually possible pattern for a given RefSeq and sgRNA. The generated patterns are not ranked, but can possibly provide a prediction of the genome editing outcome.

**Select top 10 Visualized Variants**: before using this option, it is essential to first sort the variants based on their “count” and ensure that the analysis has been performed. This control buttons select the top 10 variants, then it will be easy to copy and paste then into other programs e.g. PowerPoint, Illustrator, etc.

**Select Visualized Variants**: this allows to select the visualized variants, to copy and paste into another program (e.g. PowerPoint, illustrator, etc).

**Select Heatmap**: this allows to select the heatmap, to copy and paste into another program (e.g. PowerPoint, illustrator, etc).

**Batch Mode (Analysing existing variants)**: this allows to automate the analysis for 100s of table of variants that their address and corresponding RefSeq and sgRNAs listed in the “Experiment” worksheet. For each single analysis, a copy of the “Single” worksheet will be saved in the project folder and addresses will be populated into the “Samples” worksheet. A step-by-step guide on how to run RIMA in batch mode is provided in this document. It will not be possible to use the Excel program, while the batch mode analysis is running. However, it is possible to use other programs meanwhile. It is important also to not minimize the Excel during this analysis.

**Batch Mode (Generate possible variants)**: this allows to generate virtual mutation patterns for RefSeq and sgRNA sequences listed in the “Experiment” worksheet.

## Inputs

For each NGS run the following information must be provided:

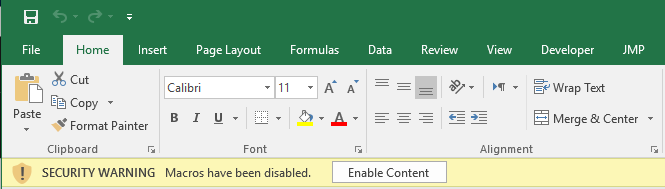
1. Reference sequence: The exact same sequence that has been used for mapping the reads and variant calling step in the NGS analysis workflow. The orientation of the sequence should be 5’ to 3’.
2. Guide RNA: The 5’ to 3’ sequence of the guide-RNA. This must be the protospacer sequence upstream of the PAM (excluding the PAM itself). It’s orientation on the reference sequence will be recognized by the software.
3. The variant table: This must be provided as same as the format that has been indicated in **Figure 18**.



**Figure 18**. RIMA’s essential inputs include the RefSeq (Reference Sequence), Target site sequence (excluding PAM) and the “Table of Variants”. The targeted sequence must be 5´to 3´ and its location on the positive or negative strand of the RefSeq will be determined by software.

If CLC Genomics Workbench software used for variant calling, then it is possible to export the variant tables as Excel file. Other inputs are optional and are explained in the following sections. Default values are shown in **Error! Reference source not found.**.

# Single run analysis

An example for RefSeq, sgRNA sequence and the Table of variant is provided in the default settings of the RIMA workbook. To perform the analysis, open the “RIMA.xlsm” file from Desktop or where the file has been saved. After opening the file, enable the contents. 

**Figure 19**. The message bar displays an option to “Enable Content”.

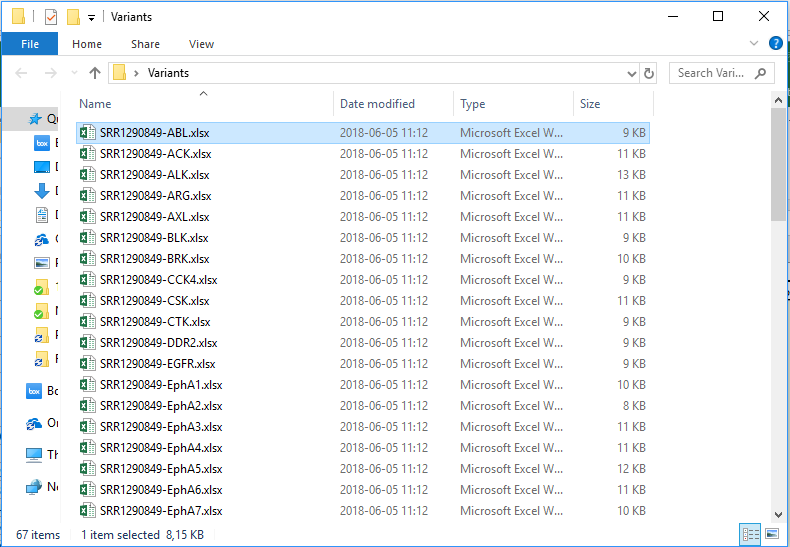
After opening the file and enabling the contents, click on the “Analyse” button. This triggers the execution of codes that generate the graphical representation of the mutations, categorization of the mutation into several classes etc.

To perform a similar analysis for another sample, then it simply enough to press “Housekeeping” and provide variant table and corresponding sgRNA and RefSeq.

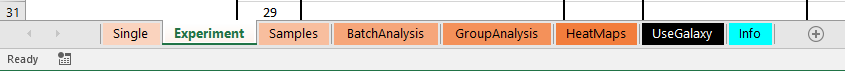
To save a copy of the “Single” worksheet for each sample, click on the “Add to database” button.

# Step-by-step example for batch analysis

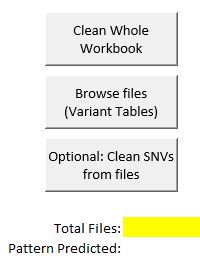
1. Download and unzip the supplementary “Example\_Variants\_for\_batch\_analysis.zip” file. This is a data set previously reported by Bae et al. (2014, doi:10.1038/nmeth.3015) for 67 sgRNAs tested in HeLa cells. The Fastq file can be downloaded from SRA databank (SRR1290849). The RefSeq and sgRNA sequences are provided in the **Supplementary Table 1**.



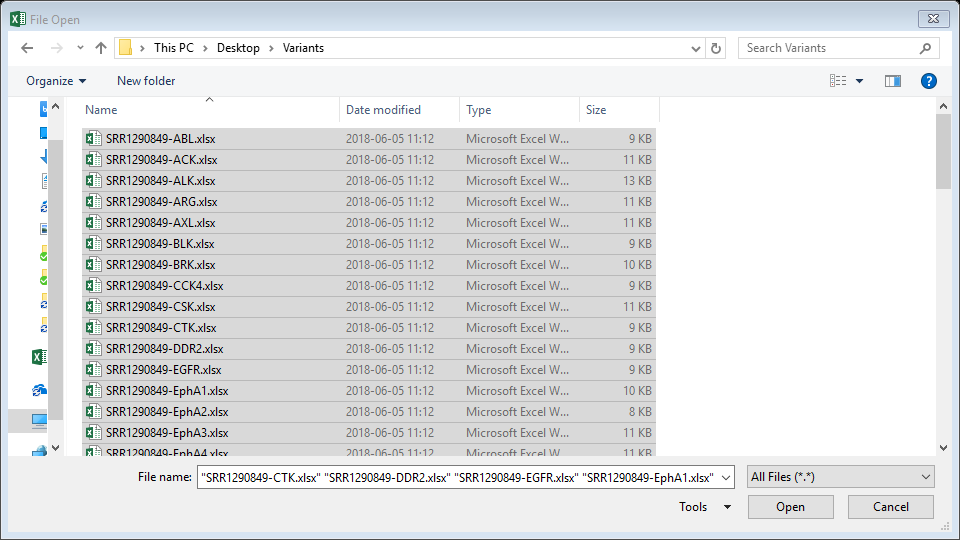
1. Open the “RIMA.xlsm” file and enable the content (see **Figure 19**).
2. Go to the “Experiment” worksheet.



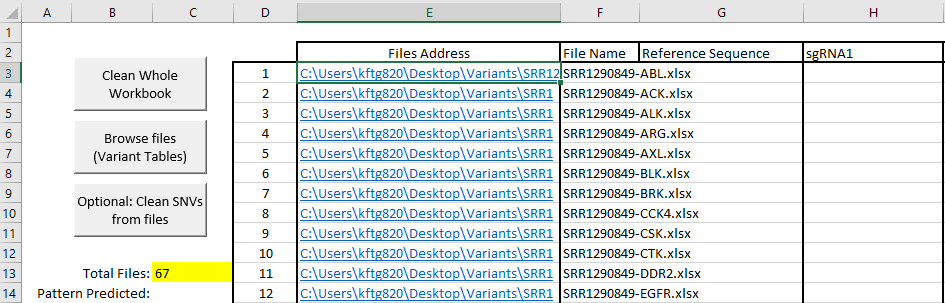
1. Click on “Browse files (Variant Tables).



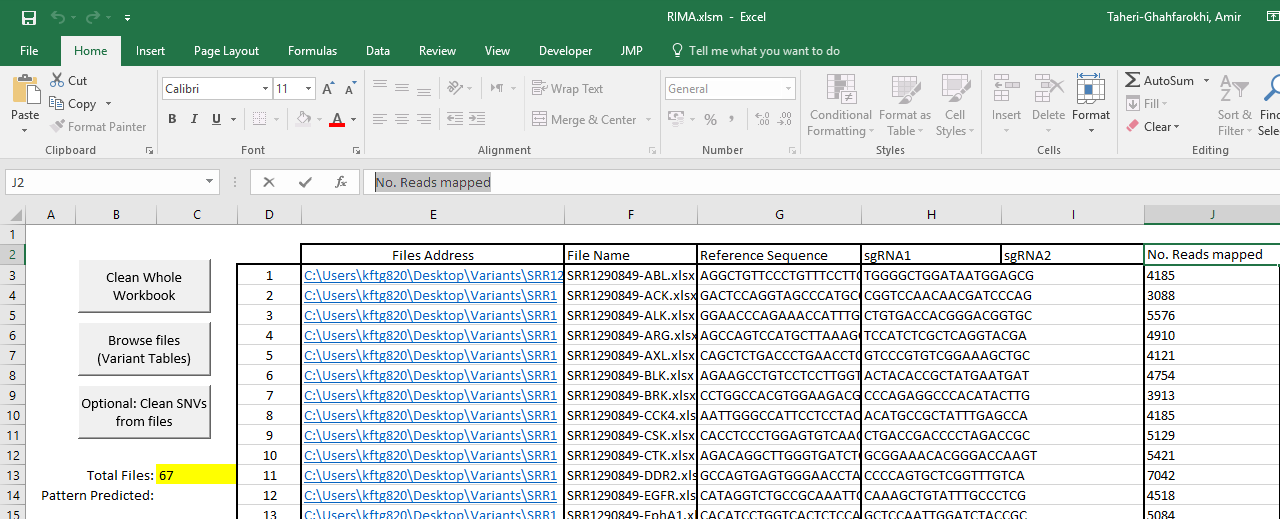
1. Navigate to the folder that the unzip variant files are stored, select all the 67 files and press “open”:



1. Now the address for all the files are populated in the “Experimet” worksheet. Note that Total Files: 67.



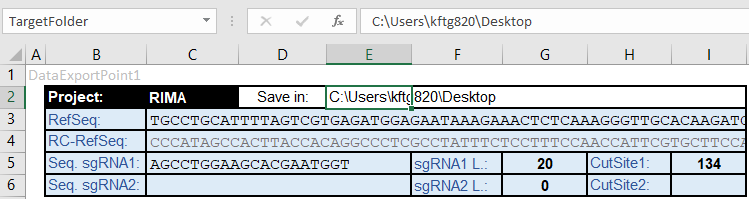
1. Press “Optional: Clean SNVs from files”. This is recommended to perform as most of the SNVs are due to amplification errors. This step takes ~70 sec for all the files to be processed.
2. Copy the No. Reads mapped, RefSeq, and sgRNA sequences from **Supplementary Table 1** and paste into the corresponding columns in “Experiment” worksheet.



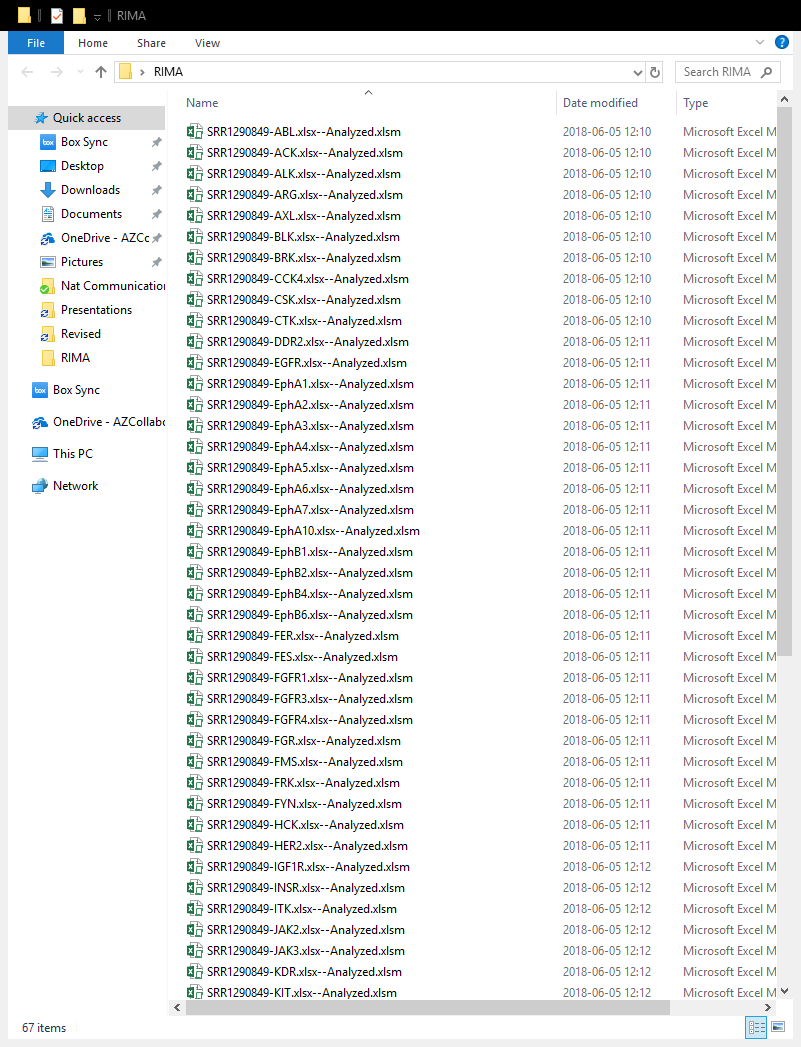
1. Go to the “Single” Worksheet.

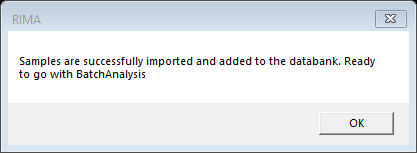
Screen Clipping

1. Provide a valid address in cell “E2” for saving the analysed files.

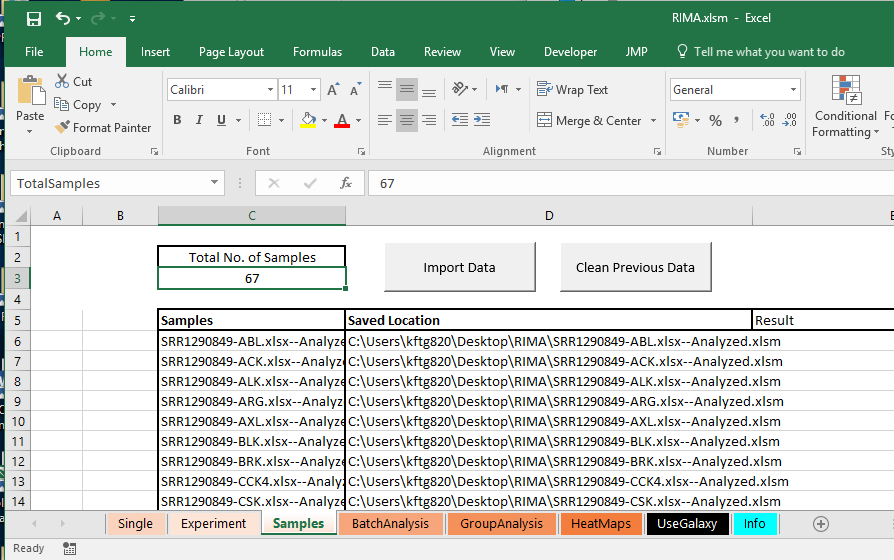


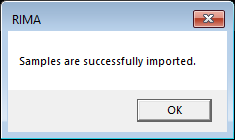
1. You are ready now to press the “Batch Mode (Analysing existing variants)”. This will automate the import from variant files into the “Single” worksheet, as well as reading the RefSeq, sgRNA and mapped reads from the “Experiment” worksheet, perform the analysis, and save a copy of the “Single” worksheet for each sample. It is possible to use the computer in the meanwhile that this analysis is performing (e.g. for web surfing, etc), but it is not possible to use Excel program. Also, it is important to avoid minimizing the main RIMA workbook, since this can interrupt the analysis. The whole analysis for all 67 table of variants takes ~15 min. RIMA has been tested to analyse up to 500 files in one run. The following message will be displayed when the Batch analysis is over. For each file, a copy is saved in the database.



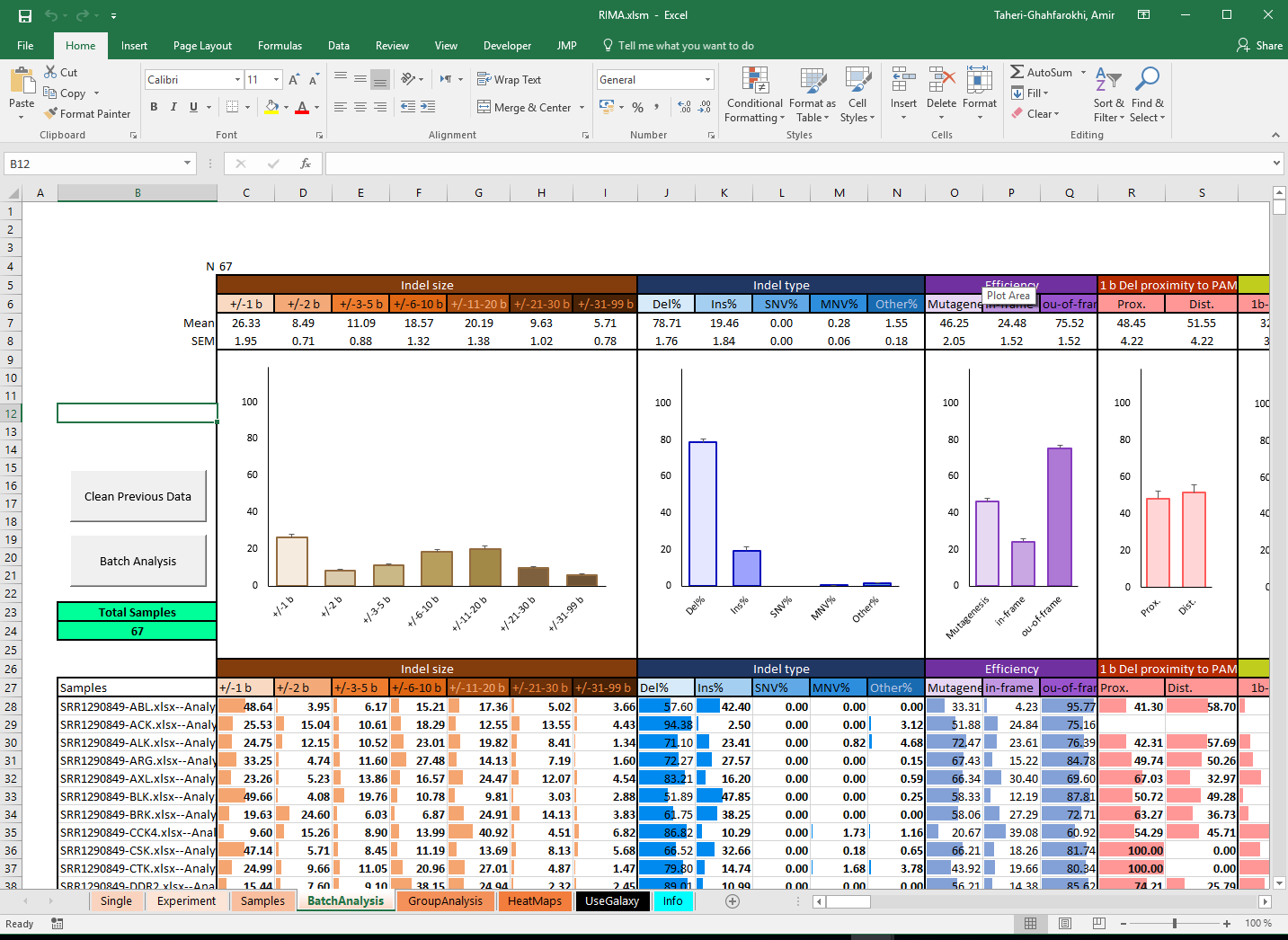


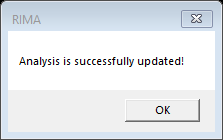
1. Go to “Samples” Worksheet. The addresses for the saved files are populated in this sheet, and the Total No. of Samples indicated. Click on “Import Data” to import the data from each single file back into the main workbook. This process takes ~3 min for all the 67 files.



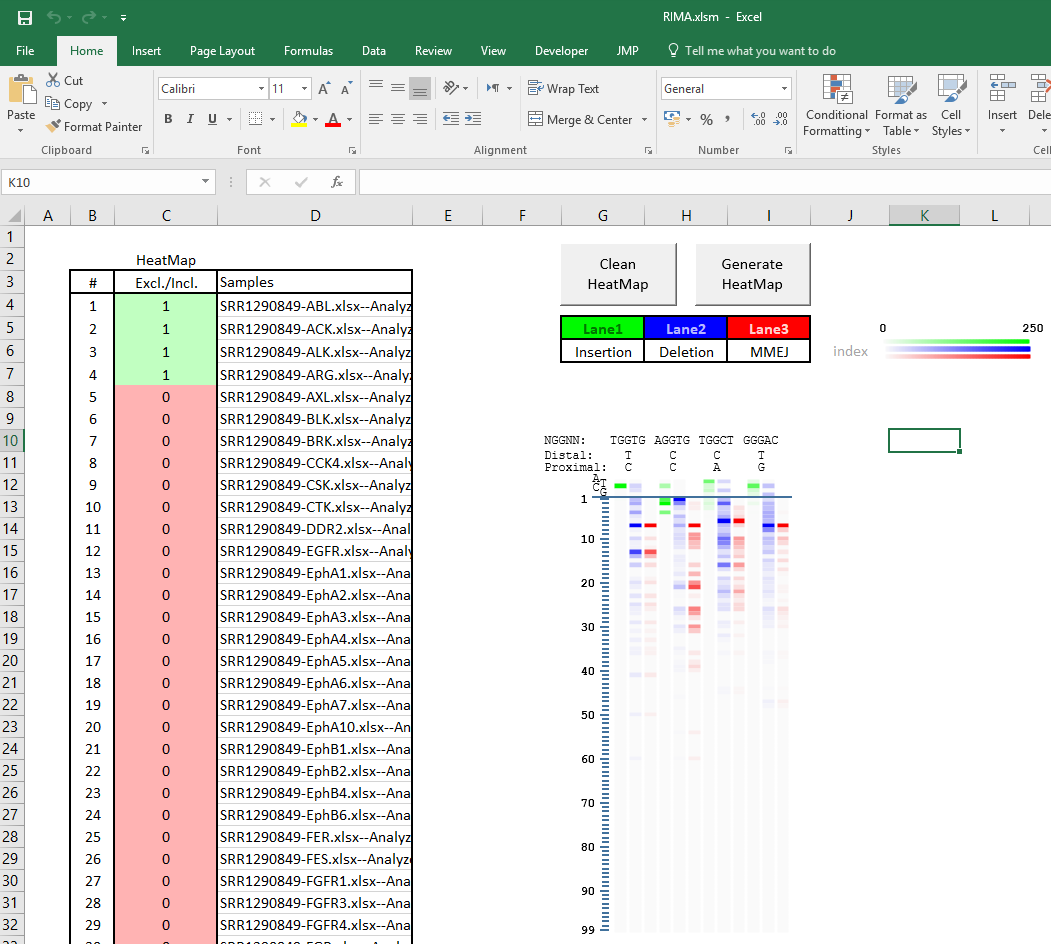


1. Go to the “BatchAnalysis” Worksheet and click on the “Batch Analysis”. This organizes the imported data from “Samples” worksheet into the “BatchAnalysis” worksheet, making it easy to copy the data into statistical analysis software.





1. Go to “HeatMaps” worksheet and select the samples that you want to generate heatmaps for them by simply replacing “0” with “1” and clicking on the “Generate Heatmap”.



1. Save the main RIMA.xlsm workbook. Single files can be individually opened and if you keep the main Workbook open, it will be possible to use the control buttons from the single worksheets. For example, variant tables can be sorted based on the count and generate the graphical representation based on the rank of the mutations.

# Future directions

There are several possibilities to develop RIMA further:

* Profiling DNA repair pathways based on the most recent DNA repair models (e.g. DS-MMEJ and TMEJ insertions).
* Visualization of Base-Editing outcome (or other flavour of nucleases).
* Using other NGS analysis platforms (e.g. UseGalaxy.org) to generate RIMA-compatible inputs.
* Developing web-based interface for RIMA.

# How to cite RIMA

Taheri-Ghahfarokhi A., et al. (2018) Decoding non-random mutational signatures at Cas9 targeted sites. Nucleic Acids Research. doi: 10.1093/nar/gky653

**GitHub address:**

*https://github.com/Ghahfarokhi/RIMA/*