SIQ User guide

Version: 1.0

Date: 30 June 2022

Author: Robin van Schendel

Sequence Interrogation and Qualification (SIQ) is specially designed to create mutation profiles for experiments where DNA damage is inflicted on a specific location. This includes (but is not limited to) CRISPR induced double-strand breaks; I-SceI, Talens, ZNFs and CRISPR nickase experiments can also be analysed by SIQ. All that is required are NGS files (or Sanger) containing sequence reads that span the intended target site. Below we provide a schematic overview of how to define a cut site (or two) by using the left and right flank, which are recommended inputs for SIQ. This instructs SIQ where to search for modifications in the reads compared to the provided reference DNA.

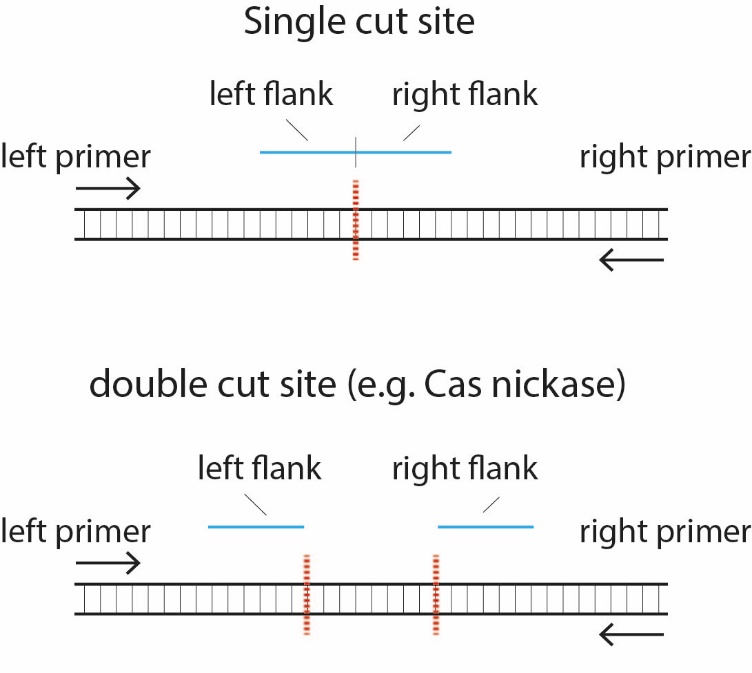


Figure 1 Illustration of how to specify flanks and primers

SIQ is written in **Java** so that it can be run on various operating systems, such as Windows, Linux and MacOS. The only software you require in addition to the latest version of SIQ is **Java** which can be downloaded and installed for all operating systems. See [Download](#_###_Installation) for details.

We have designed and implemented this tool for anyone to use and you do not require a background in (bio-)informatics to be able to analyse NGS data. We think this User Guide will help you through all steps to analyse your data. If you run into any problems, please make an issue of this [here](https://github.com/RobinVanSchendel/SIQ/issues), so we can help you solve your problems.

Contents

[Download SIQ 3](#_Toc107830727)

[Starting SIQ 3](#_Toc107830728)

[Windows 3](#_Toc107830729)

[Linux 5](#_Toc107830730)

[MacOS 6](#_Toc107830731)

[Running SIQ 7](#_Toc107830732)

[Quick Start 8](#_Toc107830733)

[SIQ on Illumina data 10](#_Toc107830734)

[SIQ on PacBio data 12](#_Toc107830735)

[SIQ on Sanger sequences 14](#_Toc107830736)

[SIQ optional parameters (Illumina & PacBio) 18](#_Toc107830737)

[SIQ output 19](#_Toc107830738)

[rawData 19](#_Toc107830739)

[Information 21](#_Toc107830740)

[RunInfo 21](#_Toc107830741)

[Top100BadReads 22](#_Toc107830742)

[SIQPlotteR 23](#_Toc107830743)

[SIQPlotteR input 23](#_Toc107830744)

[Dashboard (left panel) 23](#_Toc107830745)

[Tornado plot 25](#_Toc107830746)

[Allele plot 25](#_Toc107830747)

[Efficiency plot 26](#_Toc107830748)

[Type plot 26](#_Toc107830749)

[Homology plot 26](#_Toc107830750)

[Size plot 26](#_Toc107830751)

[SNVs plot 26](#_Toc107830752)

[Target plot 27](#_Toc107830753)

[Sample Info 27](#_Toc107830754)

[Templated insertions plot 27](#_Toc107830755)

[Outcome plot 27](#_Toc107830756)

[About 28](#_Toc107830757)

[Troubleshooting 28](#_Toc107830758)

[SIQ does not start on MacOS 28](#_Toc107830759)

[SIQ does not work on files that need merging of paired-end reads 28](#_Toc107830760)

# Download SIQ

* Ensure Java (version 1.8 or higher is required) is installed which can be installed from [here](https://www.java.com/en/download/)
* Download latest Java .jar file from the GitHub repository [here](https://github.com/RobinVanSchendel/SIQ/releases/latest). For example the file SIQ\_1.1.jar

# Starting SIQ

## Windows

Navigate to the location where you downloaded the .jar file. Let’s assume your downloaded SIQ\_1.1.jar then you can now double click it. Some users report that Windows gives a warning message about possible malware when they download the .jar file. This can be ignored and should only occur the first time you start SIQ.

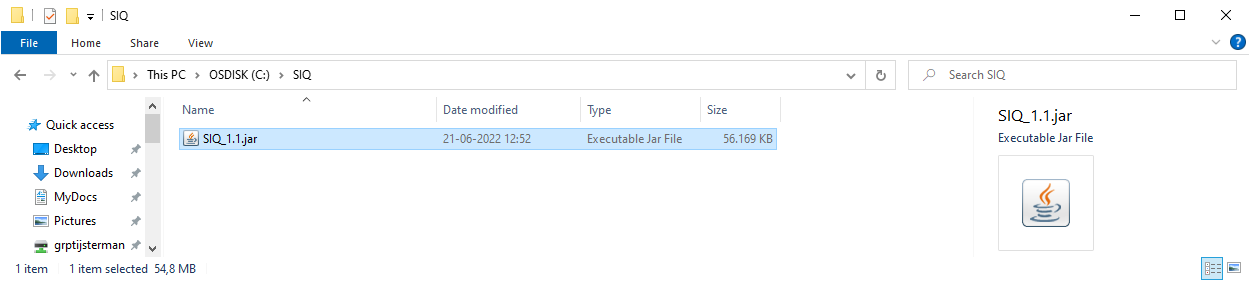


Figure 2 Windows screenshot of where jar file is located

If Java is not installed you will see:

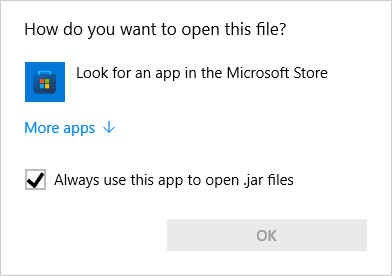


Figure 3 Java is not installed

Please download and install Java, see [Download SIQ](#_Download_SIQ)

You can (optionally) check if Java is correctly installed by starting a Command Prompt and type:

java -version

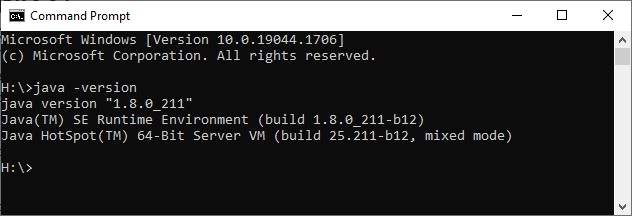


Figure 4 Java 1.8 is correctly installed

Once Java is correctly installed you can double-click on the SIQ .jar file and SIQ will launch:

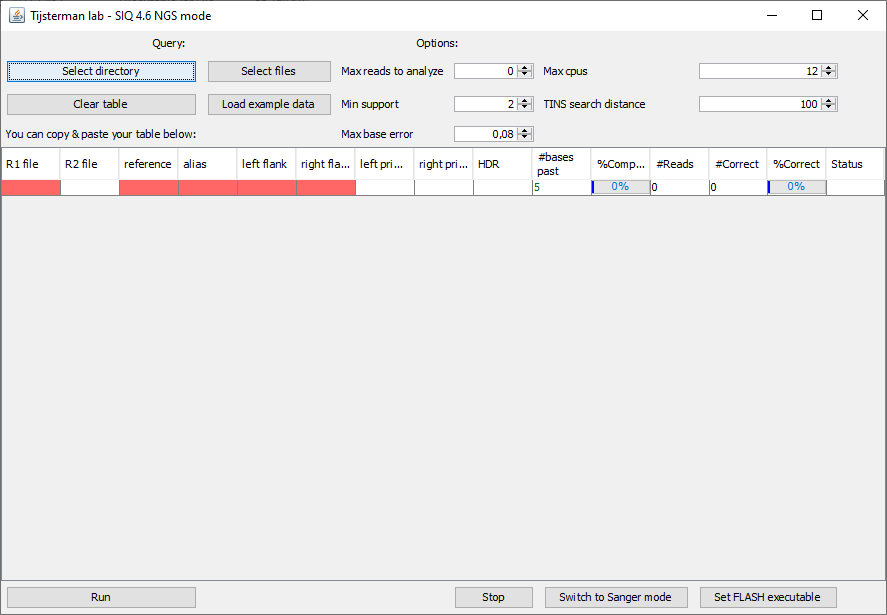
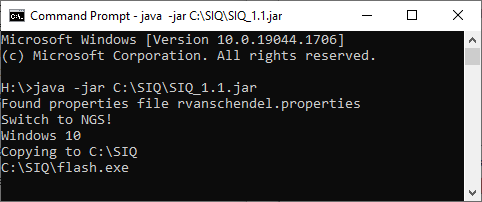


Figure 5 SIQ is ready to run

If Java is correctly installed and the window does not launch try to launch SIQ from the Command Prompt (in this case SIQ was put in C:\SIQ, but it can also be for example C:\Downloads\SIQ\_1.1.jar):

java -jar C:\SIQ\SIQ\_1.1.jar



If SIQ fails to start there should be an error here describing the problem. If you cannot solve the problem, please file an issue [here](https://github.com/RobinVanSchendel/SIQ/issues).

## Linux

For Linux users, please ensure Java is installed. If you prefer to install OpenJDK you can also use that. Please check [here](https://openjdk.org/install/) for a detailed description on how to install OpenJDK. If java is correctly installed, you should type this in a terminal:

java -version

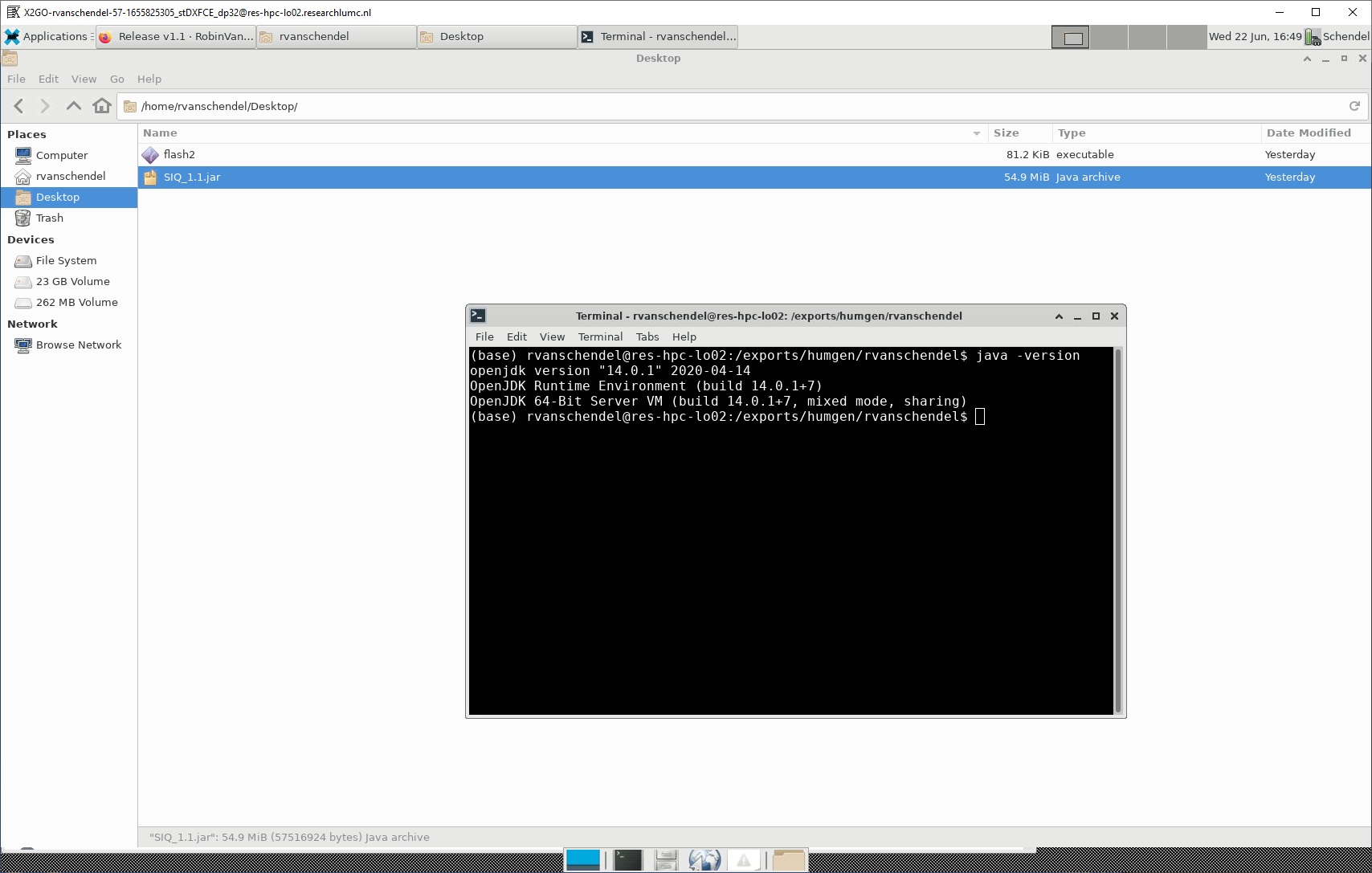
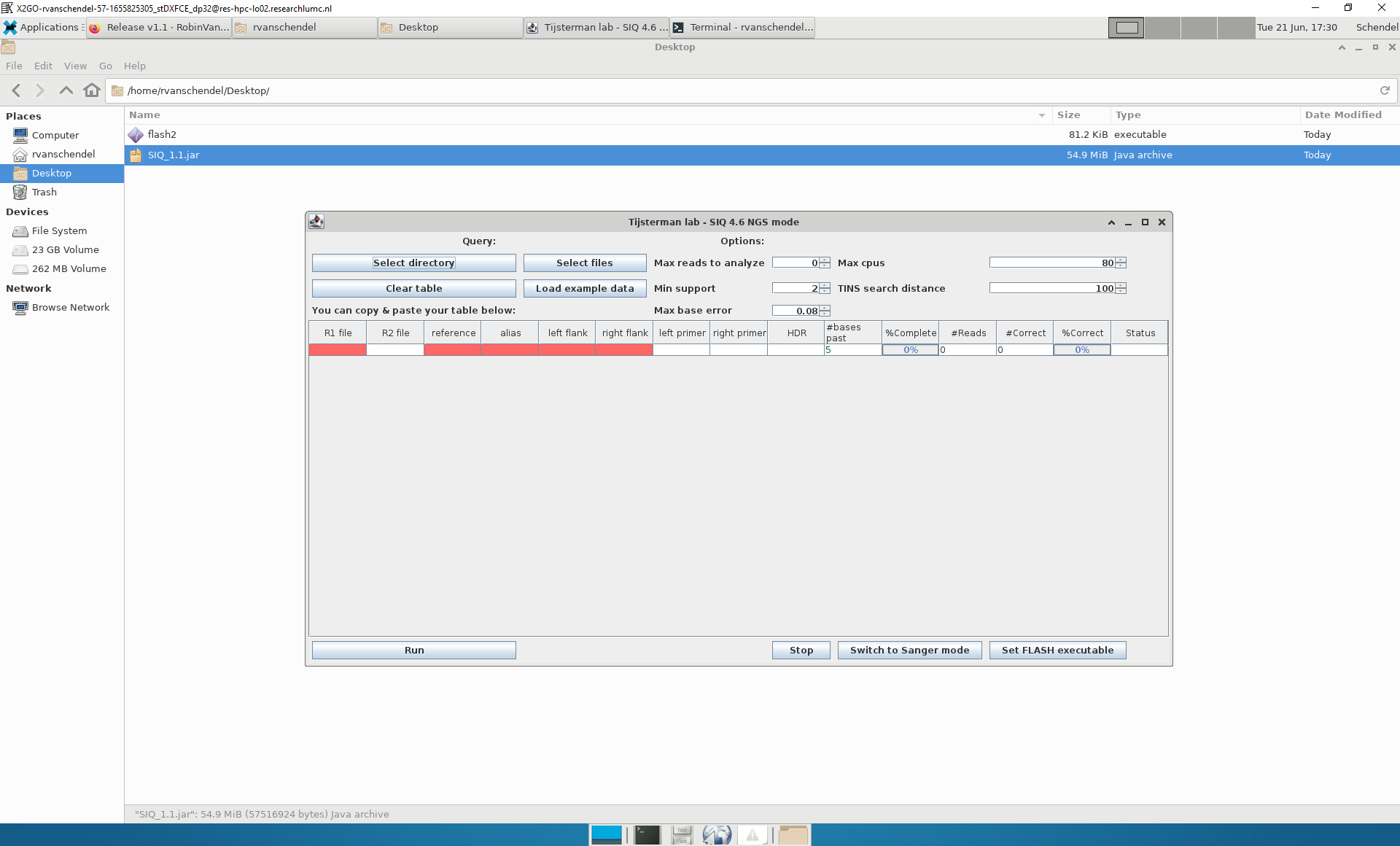


Figure 6 Example of installed OpenJDK version 14

To run SIQ double click the SIQ icon or type in a terminal (assuming SIQ is in the location /home/rvanschendel/):

java -jar /home/rvanschendel/SIQ\_1.1.jar

When SIQ has started you should see the following screen:



## MacOS

For MacOS make sure Java has been installed. You can download and install Java [here](https://www.java.com/en/download/).

You can optionally check that java is working correctly by opening a Terminal and enter

java -version

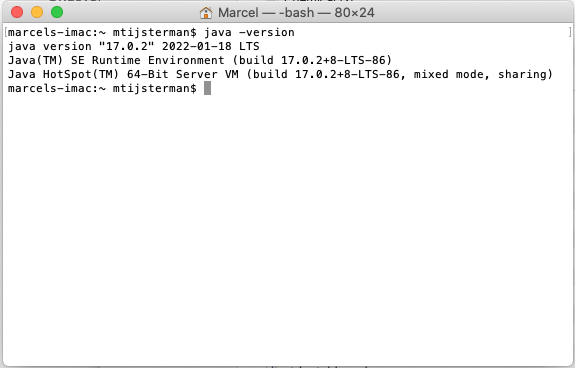


Figure 7 Terminal screenshot showing current version of Java

Once Java is installed you can double click the .jar file.

MacOS does not trust any software that is not downloaded from the App store. It could therefore be that you get the following notification:



Figure 8 MacOS warning that SIQ cannot be started

To solve this please press ‘Control’ + click on the jar file (e.g. SIQ\_1.1.jar) and then the following should appear:



Figure 9 MacOS warning which allows you to open SIQ

You can now ignore this message and start SIQ by clicking ‘Open’. The next time you want to open SIQ the warning message will not appear and SIQ launches directly. Once SIQ has launched it will appear like this:

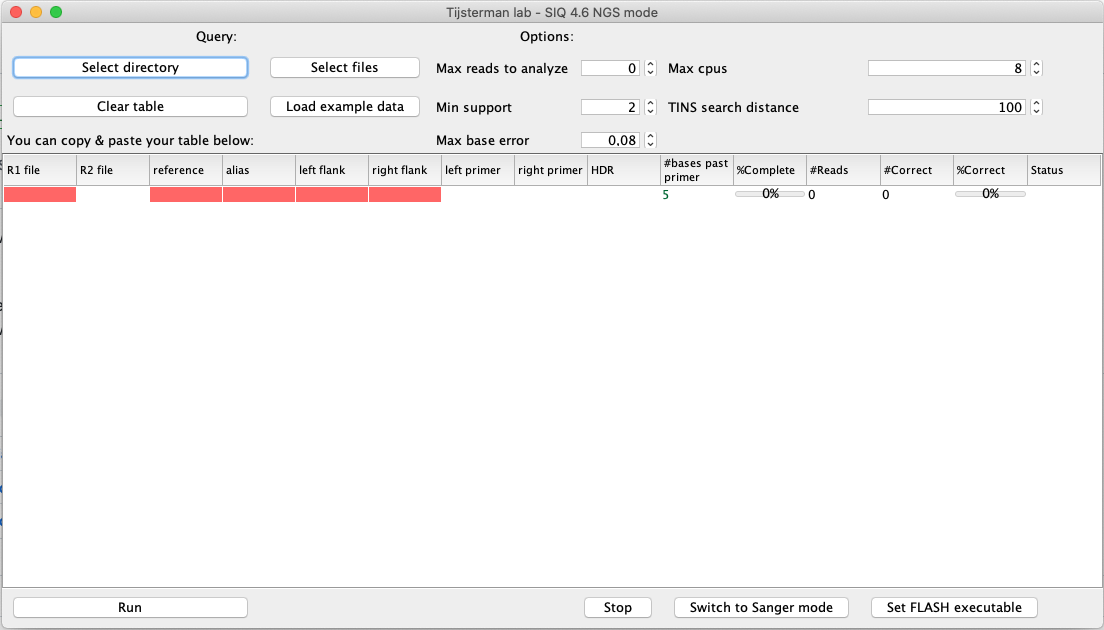
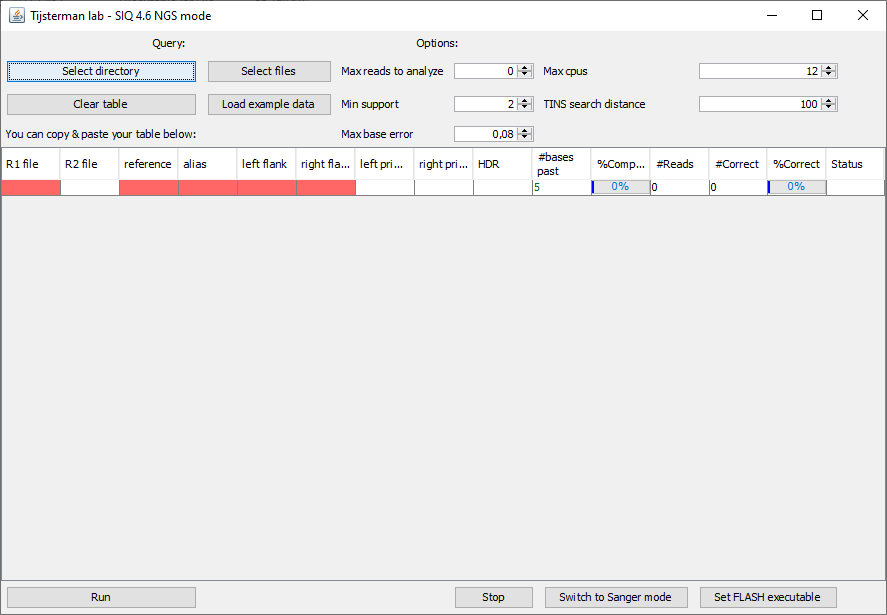


Figure 10 SIQ is running on MacOS

You are now ready to run SIQ which will be discussed in the following chapters.

# Running SIQ

In this section we assume you have SIQ up and running and see a screen similar to this:



SIQ can be used to analyse the following type of data:

\* Sanger sequences (.ab1 files). Note: only Sanger sequences containing **a single mutation** can be analyzed. SIQ does not decompose Sanger sequences containing multiple mutations. If you require such an analysis, please use another tool such as Tide, ICE or DECODR. For information on analysing Sanger sequences, please take a look at this [section](#_SIQ_on_Sanger).

\* Illumina single-end and paired-end sequence data (.fastq or .fastq.gz files)

\* PacBio data (.fastq or .fastq.gz files)

## Quick Start

If you want to quickly start analysing some NGS files click on ‘Load example data’ to get the following screen:

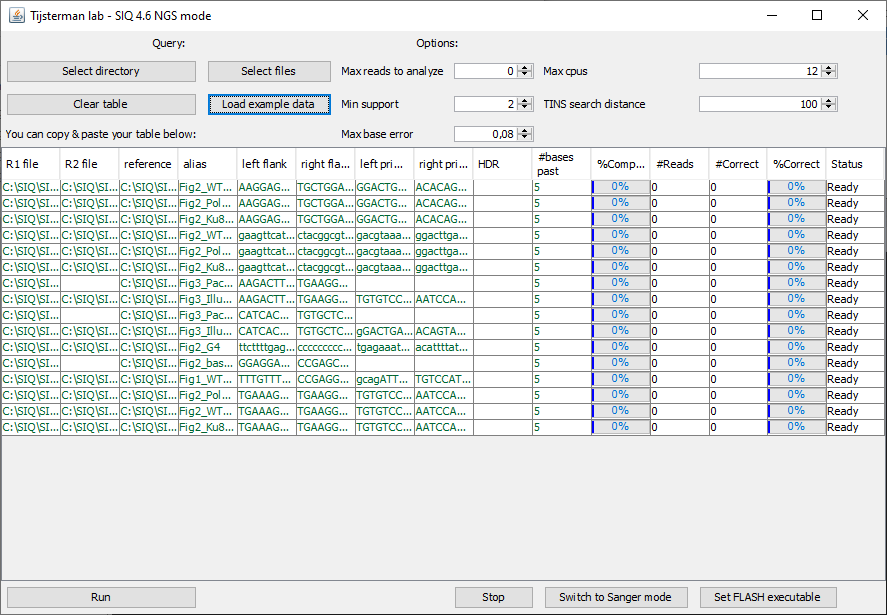


Figure 11 SIQ after clicking 'load example data'

SIQ has unpacked some files containing data that is also shown in the SIQ paper. You can now press ‘Run’ to start the analysis. You can fill in an output directory and the name of the resulting Excel file:

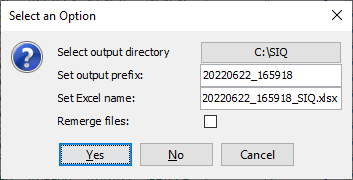


Figure 12 Output location options

You can change the directory or name of the Excel file. Press ‘Yes’ to start the run:

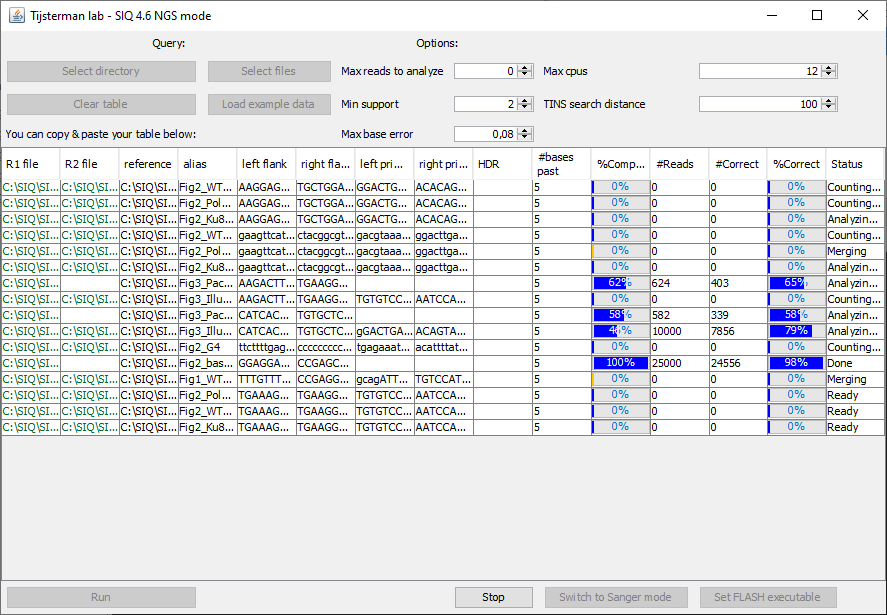


Figure 13 SIQ is running on example data

Once SIQ is complete it will open a folder with the resulting Excel file:

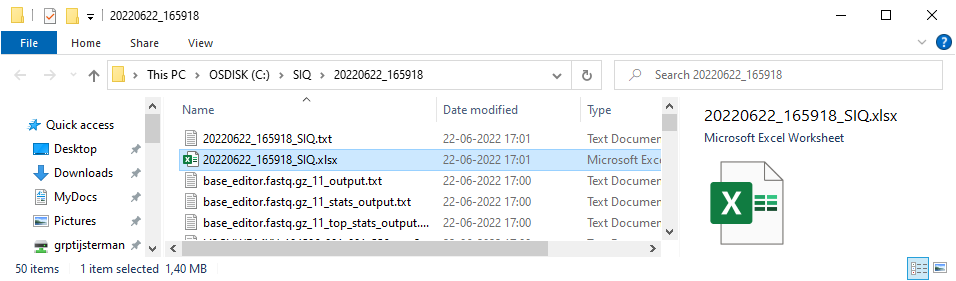


Figure 14 SIQ output folder with Excel file

You can analyse the Excel file manually or, which is the recommended option, analyse it with [SIQPlotteR](https://siq.researchlumc.nl/SIQPlotter/). For a detailed instruction on how to use SIQplotteR, please go [here](#_###_SIQPlotteR).

## SIQ on Illumina data

To run SIQ with Illumina data you need the following:

* *R1* - sequencing file (required). This needs to be in either fastq or fastq.gz format. The full path is required, so if the file myR1.fastq.gz is location in C:\raw\myR1.fastq.gz then you need to fill in: C:\raw\myR1.fastq.gz
* *R2* - paired end sequencing file. If provided SIQ will merge R1 and R2 using FLASH (optional). R2 needs to be also in fastq or fastq.gz format, similar to R1. The full path is required, so if the file myR2.fastq.gz is location in C:\raw\myR2.fastq.gz then you need to fill in: C:\raw\myR2.fastq.gz
* *reference* - reference file containing the DNA reference sequence in FASTA format (required). Needs to include the primer sequences as well. Keep the reference sequence small (preferably <1kb) as this partly determines the runtime of SIQ.
  + FASTA format (DNA sequence should at least include the left and right primer). Note that the name of the DNA (here: refDNA) will be used as a Target name in SIQPlotteR.

>refDNA

ACCTAATGTTAGAGCAGTAGATAGA

* *alias* - sample name (required). If you have files that use the same sample name these will be merged by SIQPlotteR later. So ensure you use unique names for different samples, unless you want the files to be merged later. It is fine to use the same sample name at different target-sites.
* *left flank* - the stretch of DNA that directly flanks the expected target site at the left side. Needs to be at least 15 nucleotides long. See above for a graphical example (recommended).
* *right flank* - the stretch of DNA that directly flanks the expected target site at the right side. Needs to be at least 15 nucleotides long. See above for a graphical example. In the case of for example Cas9 nickases this can designate the second sgRNA target site (recommended).
* *left primer* – the forward primer used to amplify the DNA. The entire primer needs to be present in the reference. Specify the primer from 5’ to 3’ (recommended).
* *right primer* – the reverse primer used to amplify the DNA. The entire primer needs to be present in the reference. Specify the primer from 5’ to 3’ (recommended).
* *HDR* – an optional file containing the sequence of a HDR event, including the left and right primer. The HDR sequence needs to be in FASTA format, similar to the reference.
  + FASTA format (normally this DNA is longer and at least includes the left and right primer):

>hdrDNA

ACCTAATGTTA**C**AGCAGTAGATAGA

* *bases past primer* - the number of bases your sequence reads at least have to pass the primer to be included as a real event. This filter is there to make sure your primers annealed at the expected target site in the DNA (default: 5, 0 disables this filter). In most situations the default value works fine.

To run SIQ you need to provide input to the table. There are a few ways to input the information in SIQ. You can either **drag & drop** the files into SIQ or you can **copy & paste** a table (for example from Excel). You can also click ‘Select files’ or ‘Select directory’ to fill in all R1 and R2 files. An example input looks like this:

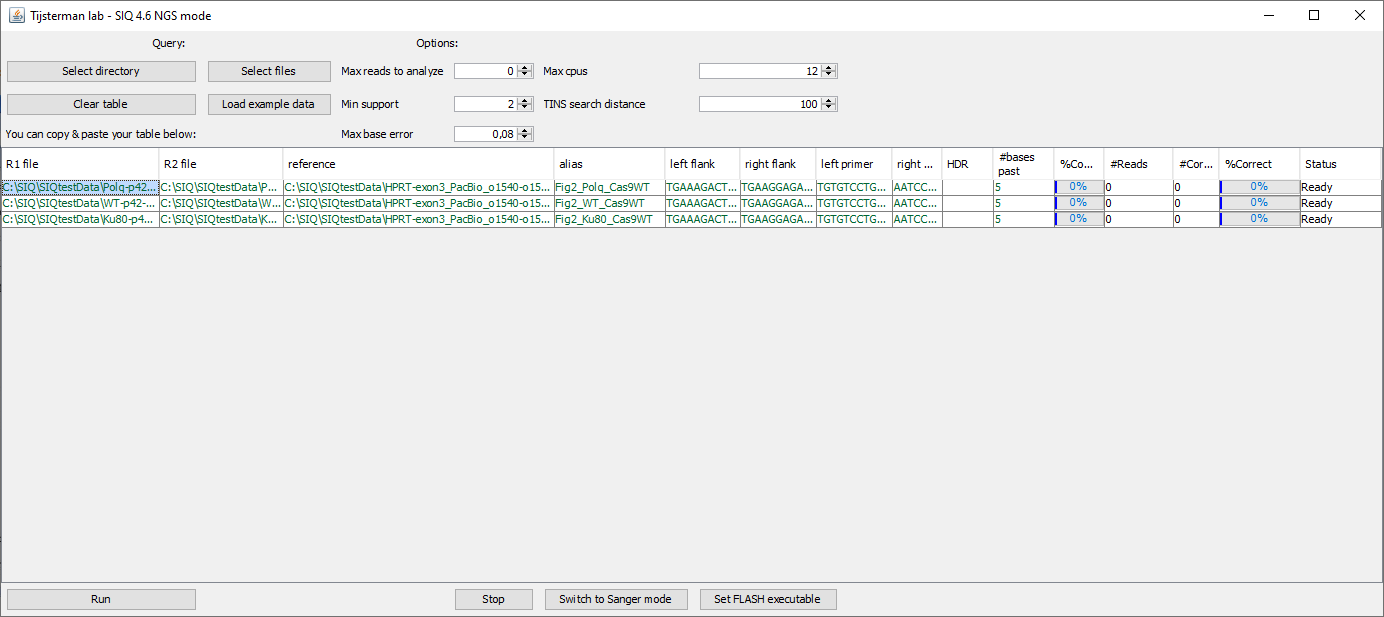


Figure 15 SIQ is ready to start analyzing 3 samples

The Excel output file also contains the input table so you can easily rerun your analysis if that is required.

The resulting Excel file can be analysed by [SIQPlotteR](#_SIQPlotteR).

## SIQ on PacBio data

To run SIQ with PacBio data you need the following:

* R1 - sequencing file (required). This needs to be in either fastq or fastq.gz format.
* reference - reference file containing the DNA reference sequence in FASTA format. Needs to contain the primer sequences as well if supplied. Keep the reference sequence small (preferably <10kb) as this partly determines the runtime of SIQ (required).
  + FASTA format:

>refDNA

ACCTAATGTTAGAGCAGTAGATAGA

* alias - sample name (required). If you have files that use the same sample name these will be merged by SIQPlotteR. So ensure you use unique names for different samples.
* left flank - the stretch of DNA that directly flanks the expected target site at the left side. Needs to be at least 15 nucleotides long. See above for a graphical example (required). For PacBio you need to provide this information otherwise SIQ will call the first difference between the references and the PacBio sequence which is generally not at the target site.
* right flank - the stretch of DNA that directly flanks the expected target site at the right side. Needs to be at least 15 nucleotides long. See above for a graphical example (required). For PacBio you need to provide this information otherwise SIQ will call the first difference between the references and the PacBio sequence which is generally not at the target site.
* left primer – currently not used for PacBio sequence data
* right primer – currently not used for PacBio sequence data
* HDR – an optional file containing the sequence of a HDR event. The HDR sequence needs to be in FASTA format, similar to the reference.
* bases past primer – this number is ignored for PacBio sequence data as no primers are set

This analysis is almost identical to a Illumina run, only the R2 files is not present and the primers are not included. An example input and subsequent run for a PacBio set looks like:

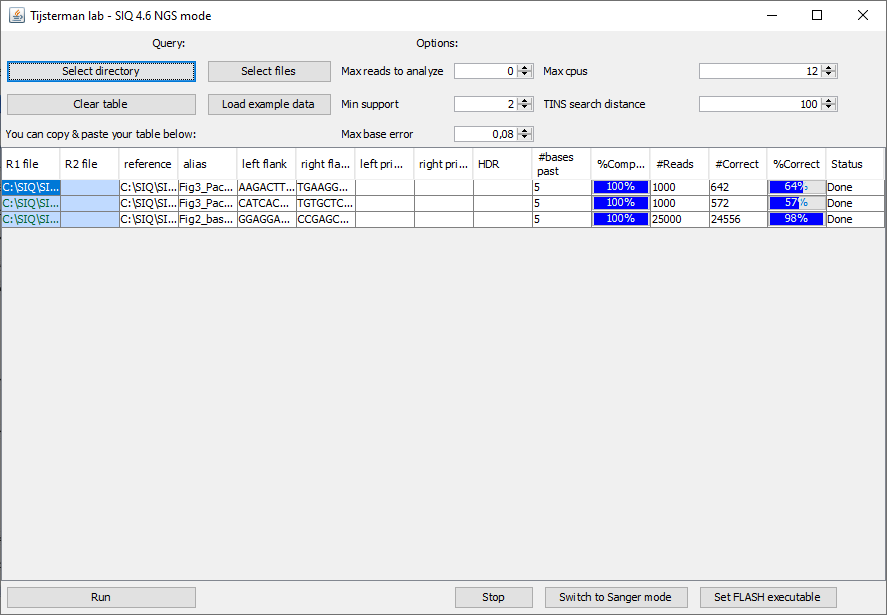


Figure 16 Example of PacBio data analyzed by SIQ

Note: Illumina and PacBio can be mixed, there is no need to run SIQ separately on each format. For example you could do this:

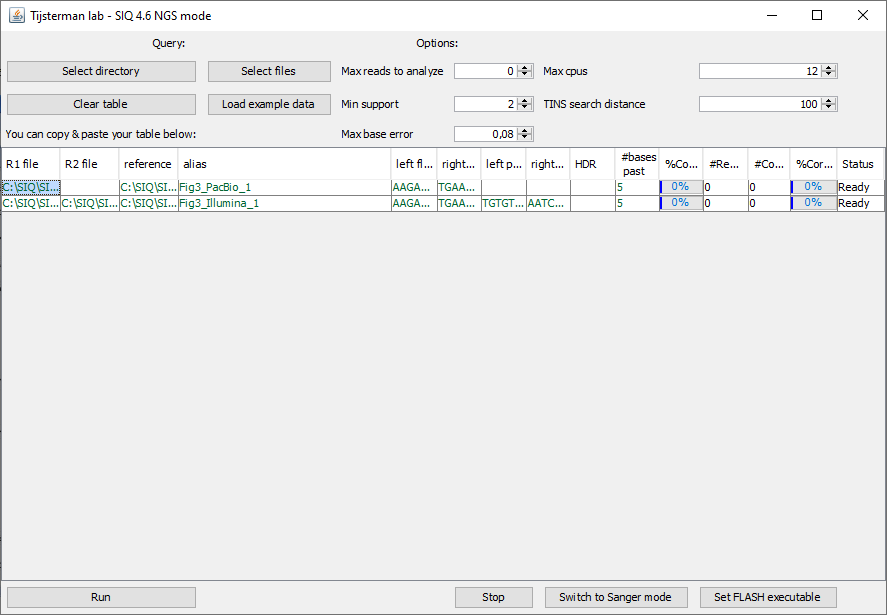


Figure 17 Illumina and PacBio files ready to be analysed

## SIQ on Sanger sequences

Important: SIQ can only analyse Sanger sequences containing a single mutation. If you performed for example a CRISPR\Cas9 experiment and you mixed the DNA prior to Sanger-sequencing you need another tool to create mutation profiles. For example Tide, ICE or DECODR would be the way to go as they can decompose Sanger sequences. If you do have Sanger sequence files containing a single variation (e.g. by colony picking & PCR & Sanger) you can switch SIQ to Sanger mode by clicking ‘Switch to Sanger mode’.

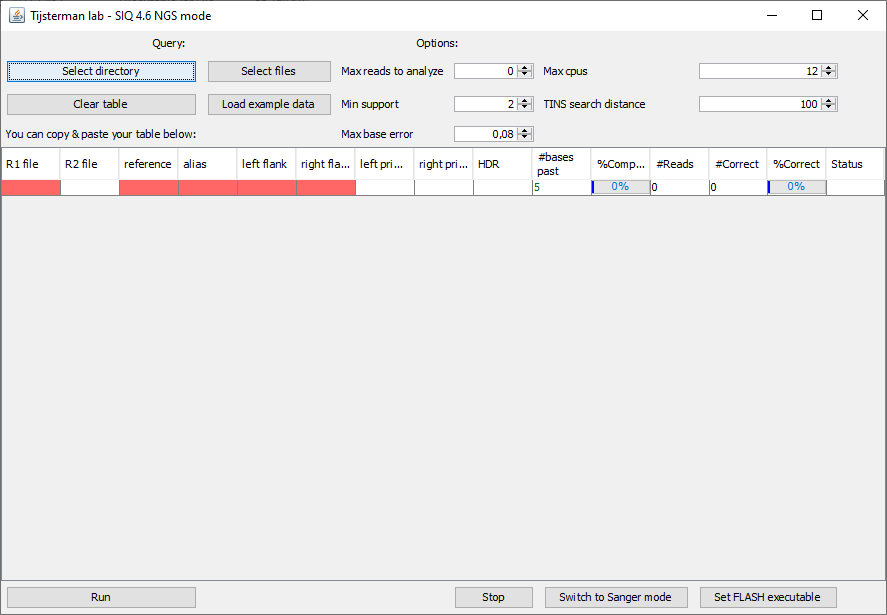


Figure 18 Go to Sanger functionality by pressing 'Switch to Sanger Mode'

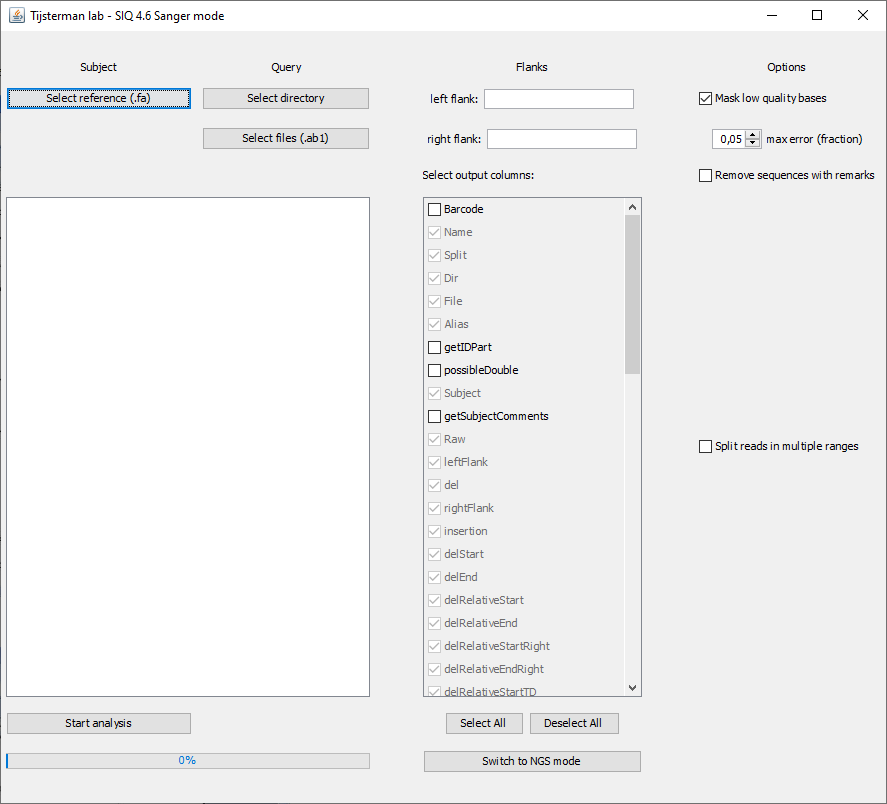


Figure 19 SIQ in Sanger mode

In order to analyse Sanger sequences please perform the following steps/fill in the information:

* Select the reference DNA file using ‘Select reference (.fa)’.
* Select the Sanger files by ‘Select directory’ (also subdirectories are searched)
  + Or by ‘Selecting files (.ab1)’
* Fill in the 15 nucleotides prior to the predicted break site (optional, but recommended)
* Fill in the 15 nucleotides past the predicted break site (optional, but recommended)

If everything is filled in the screen should look similar to this. Note that in the middle column you can check various output columns. Most of them are mandatory and cannot be unchecked. For an explanation on these columns, please take a look [here](#_rawData).

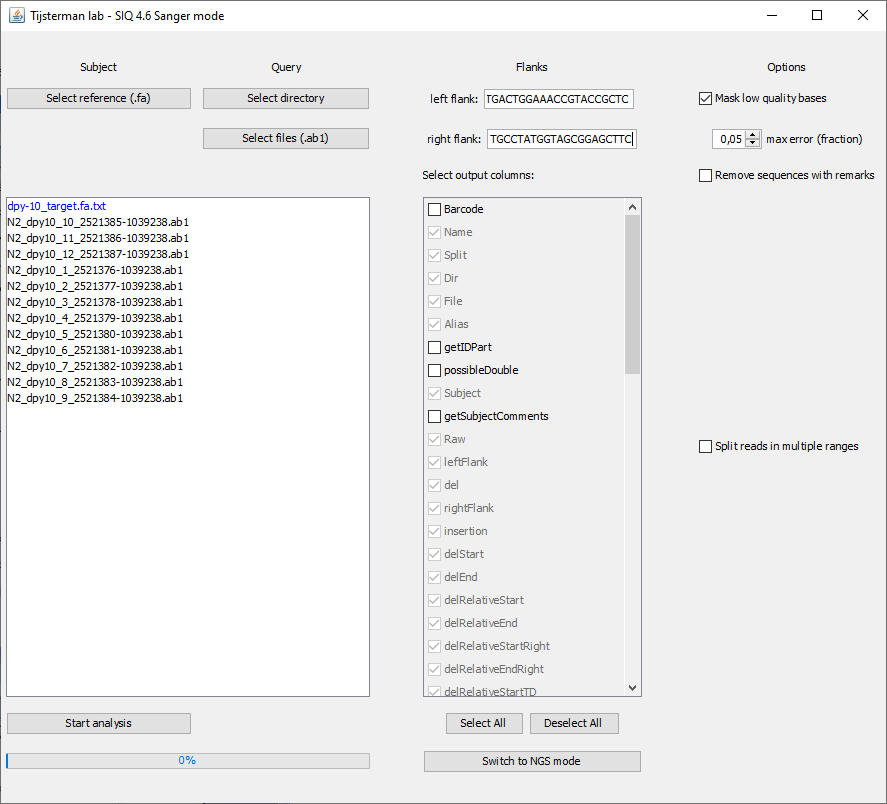


Figure 20 SIQ Sanger analysis is ready to run

When you are ready with choosing all input you can click ‘Start analysis’. When the analysis is complete a new window will appear:

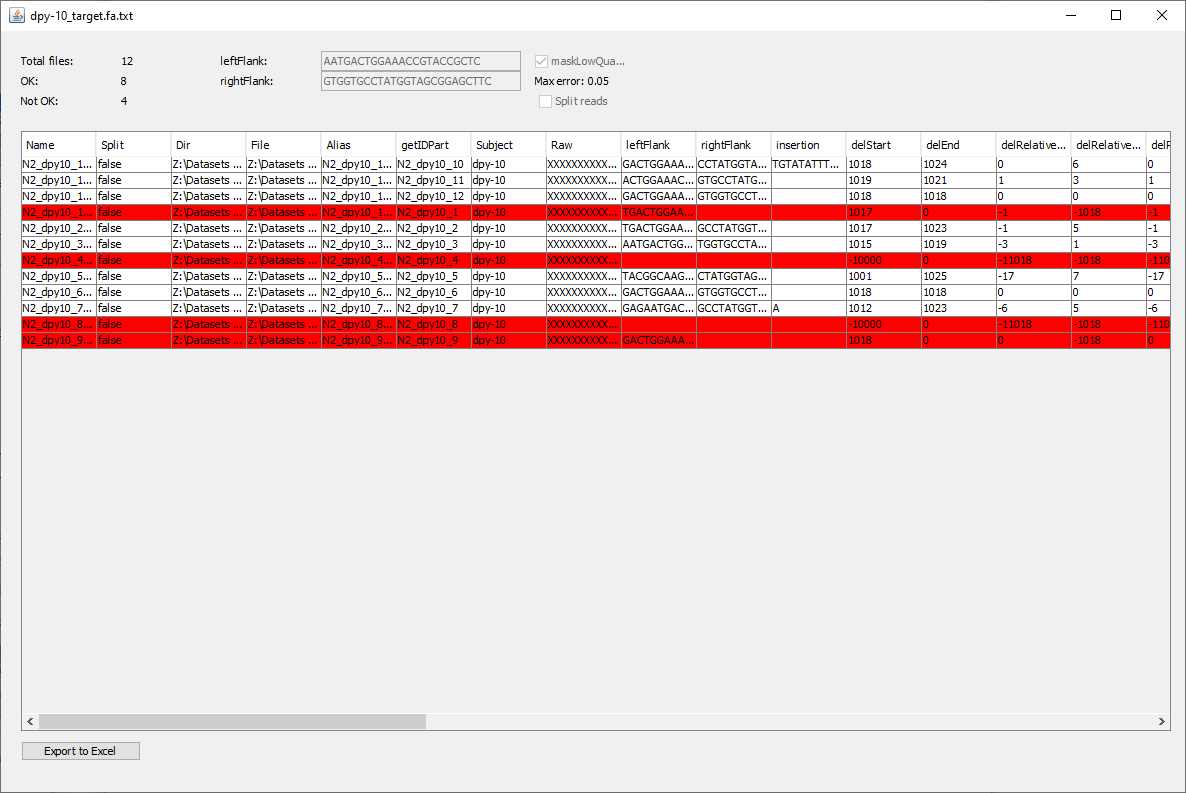


Figure 21 SIQ Sanger analysis complete

All rows in red are Sanger sequence files where SIQ was unable to find a mutation. This could have various reasons such as: bad sequence quality, heterozygous mutation or too short sequence. You can now export all data to Excel and then analyse all data in SIQPlotteR.

Note: in order for SIQPlotteR to group all sequences together it is recommended to change the Alias column to contain the same value for each related experiment/genotype. In our case we would change it to ‘N2’ (a *C. elegans* strain):

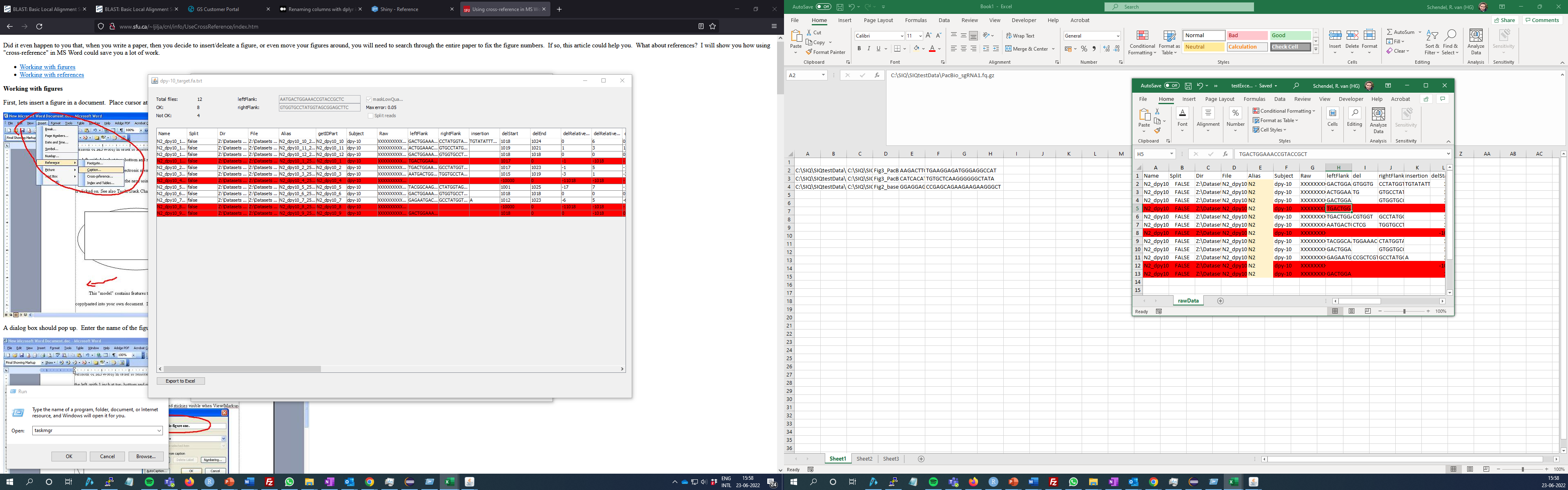


Figure 22 Alias column changed to 'N2'

Next we can view the data in SIQPlotteR:

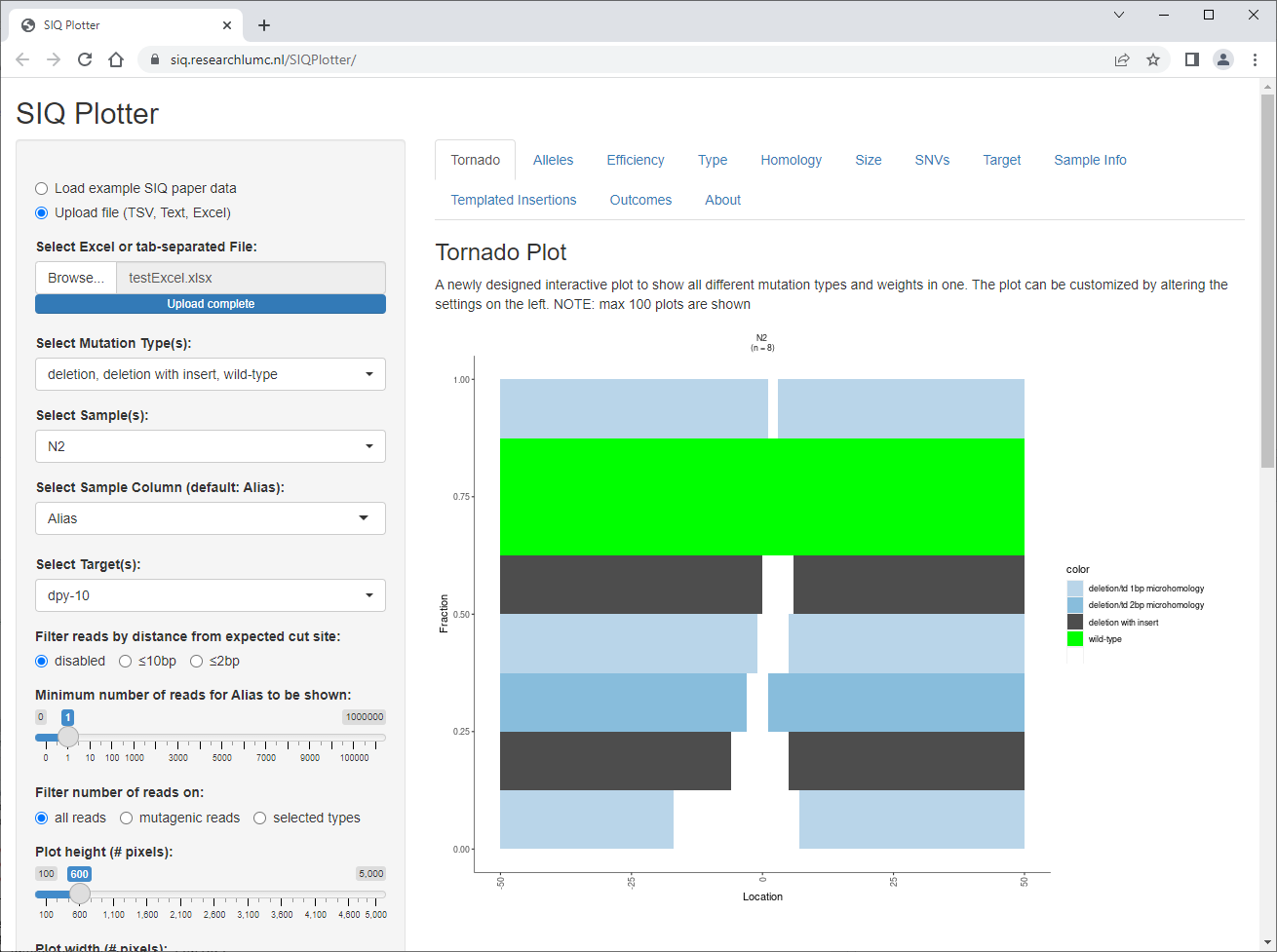


Figure 23 a tornado plot visualisation of Sanger sequence data shown in SIQPlotteR

Note: all rows that have an entry in the Remarks column are excluded by SIQPlotteR.

## SIQ optional parameters (Illumina & PacBio)

Optional settings:

* *Max reads to analyse* - the maximum number of reads you want SIQ to analyse per sample (0 mean all reads will be analysed). This option allows you to only analyse a subset of reads.
* *Min support* - the minimum number an event has to be seen to be part of SIQ's output (default: 2). To ensure only real events are called we require an event to be seen by at least two sequence. A higher number can also be used e.g. 10, but may result in infrequent mutagenic events to be missed. This setting will very much depend on the sequencing depth in your experiments. In our hands we usually keep it at 2 as infrequent events don’t add much to the total.
* *Max base error* - the maximum per base error that is allowed in a read to be analyzed (default: 0.08). Bases that have a too high base error rate are removed from the read. In combination with the ‘bases past primer’ setting this generally results in low-quality reads to be dropped from the final profile.
* Max cpus - the maximum number of CPUs SIQ can use at any given time. Note that SIQ uses a maximum of 1 CPU per NGS file (default: All)
* TINS search distance - the distance relative to the event junction to be included in the search space to search for the origin of insertions (default: 100). A TINS is a templated insertion and is generally a signature of repair by POLQ (Theta-mediated end-joining).

# SIQ output

Once SIQ successfully completes it will create an Excel file with the combined analysis. If you do not particularly care about the SIQ output, but want to directly visualize mutation outcomes, please upload the Excel file in [SIQPlotteR](#_SIQPlotteR_1).

Below we will discuss the various sheets that SIQ produces:

## rawData

This is the core of the analysis SIQ produced and contains all the events detected by SIQ and their frequencies. Each row represents one event from a single analysis file. Many columns are self-explanatory, but some are less so. We will discuss the most important ones here:

* *countEvents*: the number of reads that support this event
* *fraction*: the fraction of total reads for an Alias (Sample name) that supports this event. This totals to 1 for each Alias
* *Barcode*: experimental field, please ignore
* *Name*: read name of first read that supports this event. Can be used to look a read up in the original file
* *Split*: only applies to Sanger analysis. If true, than all high quality parts of a Sanger read are split into multiple parts that are all compared to the reference and result in an event. So each high quality part of a read results in an event. Experimental code.
* *Dir*: the directory where the read files are located
* *File*: the file that was used to call the events. If R1 and R2 was provided (as is the case for Illumina data) then this is the merged (merging is done by FLASH) file.
* *Alias*: the name given to a sample by the user. The file can contain multiple Aliases per file and Aliases with the same name and Subject are combined by SIQPlotteR. The same Alias with a different Subject is not merged
* *possibleDouble*: will be removed in future versions, was part of Sanger sequence analysis
* *Subject*: the reference DNA name, which is the first word behind ‘>’ in the provided fasta file (see [Running SIQ](#_Running_SIQ)).
* *getSubjectComments:* you can add a comment to the reference DNA file by adding a *space* to the provided reference fasta file:
  + *>Subject Subject\_Comment*
* *Raw:* the entire sequence read. If multiple reads contain the same event SIQ will show the largest read here
* *leftFlank:* 15 bp of the left part of the event junction
* *del:* the deleted piece of the reference, if found
* *rightFlank:* 15 bp on the right part of the event junction
* *delStart, delEnd:* the start and end location of the event. The start of the DNA reference is at 0.
* *delRelativeStart, delRelativeEnd:* the start and end location of the event, but now relative to the flanks set for the analysis. The location is relative to the end of the chosen leftFlank.
* *delRelativeStartRight, delRelativeEndRight:* the start and end location of the event, but now to relative the flanks set for the analysis. The location is relative to the start of the chosen rightFlank. Only if two target sites are set (for example with two nickases) these values differ from delRelativeStart and delRelativeEnd.
* *delRelativeStartTD,* delRelativeEndTD: for plotting purposed the TDs the values are set here relative to the TD detected
* *getHomologyColor:* initial coloring, but is overwritten by SIQPlotteR later
* *homology:* homology found between the junctions. Homology is only searched in Deletions and TDs
* *homologyLength:* length of the homology
* *homologyMismatch10%:* the homology if we allow 10% of mismatches in the homology. Only useful for large stretches of homology
* *homologyLengthMismatch10:* the homology length for *homologyMismatch10%*
* *homologyMismatch10%ref:* the homology if we allow 10% of mismatches in the homology. Only useful for large stretches of homology. The homology is based on the reference, which might be different from using the read.
* *homologyLengthMismatch10%ref:* the length for *homologyMismatch10%ref*
* *delSize:* the deletion size of the event
* *insSize:* the insert size of the event
* *Mod3:* is this event dividable by 3? If so, this number is 0. If not it will be 1 or 2. 1 or 2 indicates a frameshift. Note that SIQ is not aware of introns so if your target site overlaps or is in an intron this information is less useful
* *SNVMutation:* if the event type is SNV (delSize = 1, insSize = 1) the mutation is presented here
* *Type:* SIQ discriminates between the following types: WT (wild-type), DELETION, DELINS (deletions with an insertion), INSERTION, TANDEMDUPLICATION, TANDEMDUPLICATION\_COMPOUND (a TD with some additional inserted sequence), SNV (1bp substitution), TINS (templated insertion: a deletion with an insertion where the insert is copied from the flank), HDR (homology directed repair event), HDR\_1MM (HDR event with 1 indel or mismatch which is frequently the result of sequencing errors). Note that HDR and HDR\_1MM are only called if a HDR reference was supplied.
* *SecondaryType*: some Types have a subdivision. This is experimental and not used by SIQPlotteR at the moment.
* *isFlankInsert:* can the insert be reliably found around the event junction? The region around the left and right junction is searched in both forward and reverse complement orientation. Only if the insert is large enough (typically >5) and a large enough match is found this value becomes TRUE.
* *Ranges:* only used for Sanger analysis and represents the subpart of the Sanger reads that is used for event calling
* *Masked:* only high quality read parts are used and masking of nucleotides is performed that have low base quality. This value should be TRUE.
* *Remarks:* any problems found with this event are written here. Note that for NGS analyses only events without a remark are part of the output. For Sanger sequences there can be an indication of a problem. Any event with a remark should be excluded from further analysis as many fields are not correct.
* *Classname:* a short notation of this event. Not used by SIQPlotteR at the moment.
* *InZone:* Experimental column, will likely disappear from subsequent versions
* *leftFlankMatch:* which part of the raw read matches the reference DNA to the left side of the junction.
* *rightFlankMatch:* which part of the raw read matches the reference DNA to the right side of the junction.
* *matchStart:* the start location of the read in the reference DNA
* *matchEnd:* the end location of the read in the reference DNA
* *jumpedLeft, jumpedRight:* PacBio and Sanger reads are generally so long that they will always contain a difference with the reference sequence. Therefore SIQ jumps over these variations as long as they are not in the vicinity of the specified target site.
* *entireQueryused:* column currently not in use

The remaining columns are only there to give information on the presence of an insert and how much of that can be found back in the reference DNA

* *isGetLargestMatch*: size of largest match of the detected insert in the reference sequence
* *isGetLargestMatchString:* the largest matching DNA
* *isGetSubS:* This is a representation of the matching DNA. All matches nucleotides are in uppercase. Any remaining nucleotides are kept in lowercase.
* *isGetSubS2:* This is a representation of the matching DNA. <R0> means the largest match was found on the right. <0L> the largest match on the left. <R1> the second largest match from the right. Matching is performed as long as the remaining insert part is long enough (>5 nucleotides). Any remaining nucleotides are kept in lowercase.
* *isGetType:* If the entire insert is found the type is SOLVED, if some nucleotides remain it is either PARTIALLY SOLVED or ALMOST SOLVED. This field is not further used.
* *isGetLengthS:* the matching lengths of the insert. Ordered from largest to smaller
* *isPosS:* the relative positions where the insert match was found
* *isFirstHit:* the location and orientation of the largest hit: <R0>: right and forward, <rcR0> right reverse complement, <0L>: left and forward, <0Lrc>: left and reverse complement
* *getFirstPos:* the location of the largest match relative to the junction
* *isStartPos:* the start location of the largest match relative to the junction
* *isEndPos:* the end location of the largest match relative to the junction
* *isStartPosRel*:same as isStartPos, but now relative to the target site
* *isEndPosRel*: same as isEndPos, but now relative to the target site

## Information

This sheet contains information about the total number of reads analysed, the number of merged reads and the reads that failed various filters. This sheet is used to create the SampleInfo tab in [SIQPlotteR](#_SIQPlotteR_1).

## RunInfo

This sheet contains the input table that was used to run SIQ. You can copy & paste this table to SIQ:

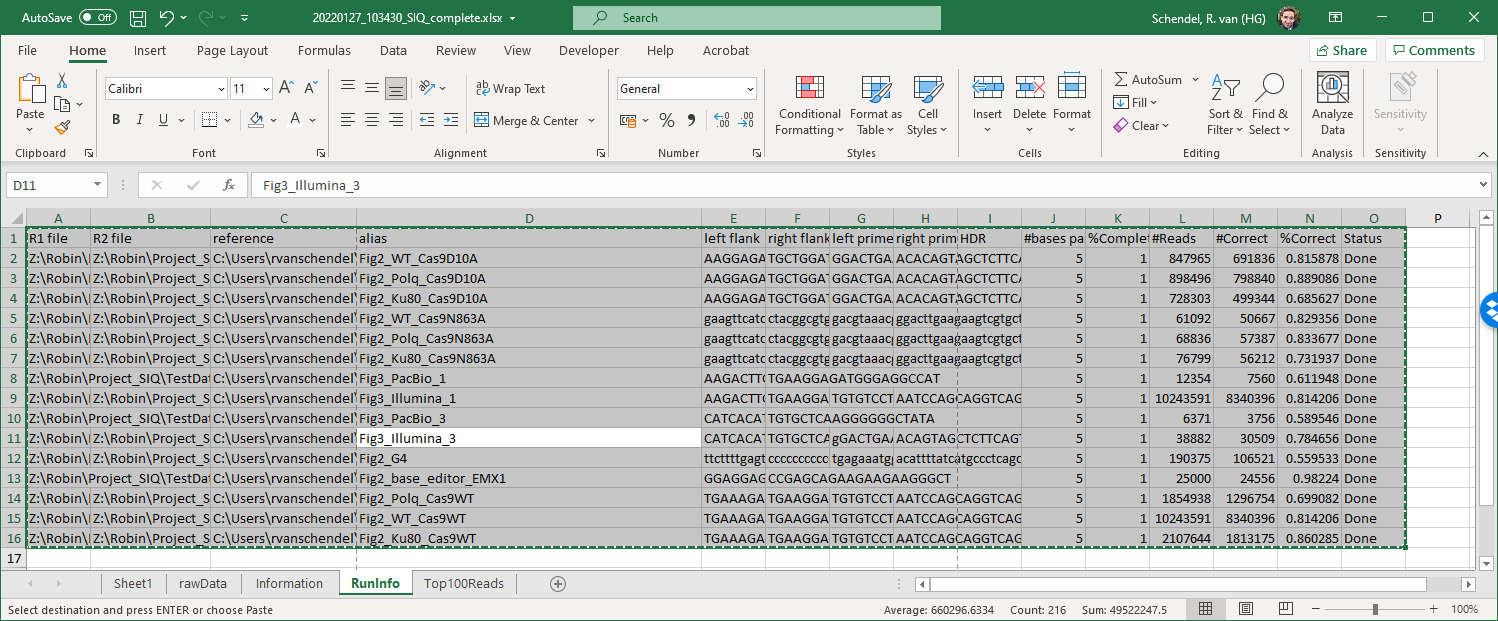


Figure 24 RunInfo sheet from a SIQ output file

And paste it in SIQ:

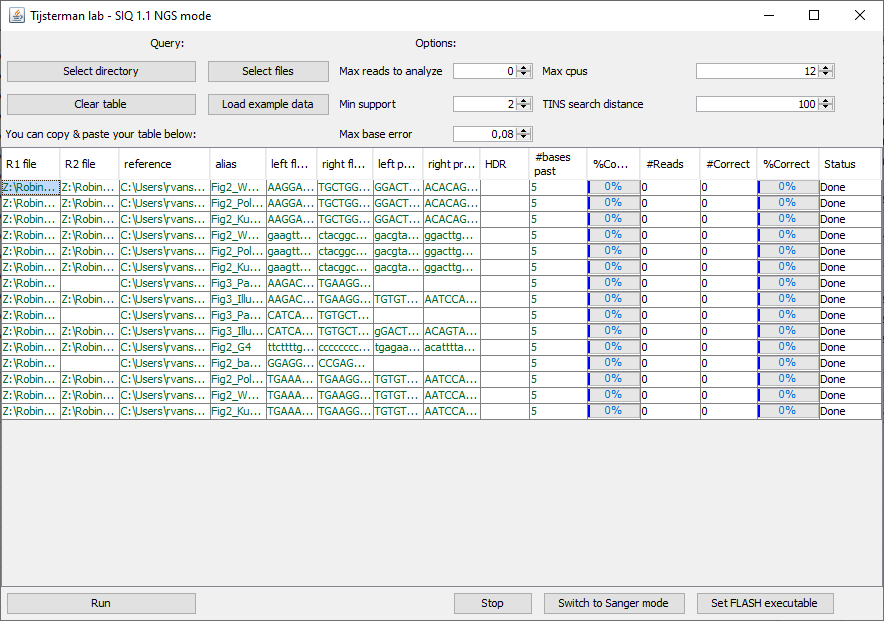


Figure 25 entire RunInfo sheet was copy&paste to SIQ, which can now reanalyse the data

Note: other parameters such as: ‘Min support’ and ‘Max base error’ are not saved in the Excel file.

## Top100BadReads

This sheet contains the top 100 most frequent reads that are not part of the final outcome as they failed one or more filters. This sheet can be used to inspect the reads for clues to why they failed the filters. If you get for example very low fractions of correct reads this is the right place to search for clues for the reason. Some common problems found in practice are:

* *Provided primer sequences are not identical to the used primers* - If the used primer starts earlier in the reference DNA then the majority of reads will not be seen as correct
* *Provided Reference sequence does not match* *amplicon* – it is good practice to Sanger sequence your wild-type amplicon to ensure the reference DNA indeed matches the amplicons sequence

# SIQPlotteR

<https://siq.researchlumc.nl/SIQPlotter/>

SIQPlotteR is the interactive visualisation tool of SIQ. SIQPlotteR allows you to visualize mutation profiles according to your own requirements. SIQPlotteR can visualize multiple samples and multiple target sites simultaneously. Below we will shortly explain all the visualisations which are at the moment present in SIQPlotteR. We might extend the type of visualisations based on our own insight or based on user request. If you experience any difficulties or problems with SIQPlotteR, please let us know:

* Github: <https://github.com/RobinVanSchendel/SIQ/issues>
* E-mail: [r.van\_schendel@lumc.nl](mailto:r.van_schendel@lumc.nl)

In addition it could be that old SIQ analyzed files no longer work correctly in SIQPlotteR. Before contacting us it is always a good idea to first rerun SIQ on your input (see Excel Worksheet ‘RunInfo’ and copy that to SIQ).

## SIQPlotteR input

SIQPlotteR requires input from SIQ analysed data in the following format:

* Excel (recommended)
* Text file

The Excel file is recommended as it contains additional information on the analysis that is not present in the text file. This concerns mostly the total number of reads analysed and the reads that were found to be correct. If your Excel file is too large (>100Mb) you might consider using the text file which is also part of the output produced by SIQ.

## Dashboard (left panel)

The dashboard allows you to customize all the plots in various ways. Here is a brief explanation of the shared options between the plots.

* Input file
  + Load example SIQ paper data
    - This is a subset of the data described in the SIQ paper. This is for users to practice with a test-set of data
  + Upload file
    - Drag & drop your SIQ file or click ‘Browse’
      * Excel file or text file
* Select Mutation Type(s)
  + You can subselect the mutation types that you want to include in the plot. E.g. you can remove WT or SNV events as desired
* Select Sample(s):
  + Select the samples that you want to be displayed. The selectable Samples depend on the selected target(s). If there are >100 samples they will not be automatically selected to prevent SIQPlotteR from creating a large number of plots.
* Select Sample Column (default: Alias):
  + If you want to use another column to be used as the Sample column you can do that here. For example if you want to change names of your samples without rerunning SIQ you can add a column and put the altered names there.
* Select Target(s):
  + Select the target sites that you want to include in the plots. Most plots work fine with multiple target sites.
* Filter reads by distance from expected cut site:
  + In most cases you only want to include events that overlap with the expected cut site (as defined by leftFlank and rightFlank in SIQ). This filter allows you to do so. For most plots ≤2bp works fine. For example SNVs, which are often the result of a PCR/sequence mistake, are usually noise that you want to exclude in the plots.
* Minimum number of reads on:
  + The number of reads that each sample should have to be included. Depending on the setting ‘Filter number of reads on’ this could be based on total reads, mutagenic reads (non-WT read) or selected types (all types currently displayed)
* Filter number of reads on:
  + Select the criterium on which Sample(s) are included in the plots. Occasionally you will have samples with very low reads or mutagenic reads that you would like to exclude from the analysis. Note that there is no predetermined way to define which samples should be excluded. We usually look at the ‘Sample Info’ and determine whether some samples have very few reads compared to other samples. Those samples are then excluded.
* Plot height (#pixels):
  + Height of each plot
* Plot width (#pixels):
  + Width of each plot
* Export to PDF
  + Export the current plot to a PDF file
* Separate targets:
  + Group the Sample by their respective target. Note: this does not work for all plots
* Sort Type:
  + Determine the order of the types by drag & drop
* Sort Samples:
  + Determine the sample order. Note that if samples from multiple target sites are selected the ordering might not work as expected and samples are still grouped by target site.
* Select colors:
  + Choose the desired color per mutation type.
  + You can also alter the homology colors here

## Tornado plot

This visualisation tries to combine several aspects of mutagenic events into one customizable plot:

* Mutation type
* Contribution to total
* Deletion size
* Homology usage (for deletions and tandem duplications)
* Tandem duplication length

There are various options to change the tornado plots.

* *X-axis range:* you can zoom in and out on the x-axis
* *Set y-*value: this becomes of importance when you deselect types. Should the plot then go to 1 or to the actual value (option: ‘not set’)? Sometimes it is also of importance to see the relative contribution between samples. You can then use ‘max of plots’ to ensure the scale is the same for each plot.
* *Select Sort:* sorting of the data set can be altered. Examples are on Start position, end position and size
* *Select tornado type:* Regular shows the deleted sequence in white. Inverted shows the deletion event in the color of that type. Inverted and ‘Closest to 0’ are experimental and might be changed later
* *Y-*axis value: show the absolute number of reads or the fraction of reads

If you hover over an event you will see the details of each event.

## Allele plot

The goal of this plot and table is to provide you with the top X (default: 10) outcomes of a target site. At the moment the table underneath the plot functions best by selecting only a single target. You can copy the data table to the clipboard and save it in for example an Excel sheet.

The allele plot provides the following additional options:

* *Set the number of alleles to be shown:* this defaults to the top 10 most frequent events.
* *Set top alleles based on: ‘*Total’ means the top X frequency is determined based on the total frequency of all selected samples. ‘Sample’ means you select the sample of which you want to see the top X and the frequency of those events in other samples is visualized.

## Efficiency plot

This plot will show you the targeting efficiency for each sample

## Type plot

For each sample only the event types are shown in this plot and a table is shown underneath the plot. This allows the user to inspect the different types of events called and compare samples without using the details of each event. So for examples all deletions are grouped into ‘deletion’ regardless of size and homology.

A few extra options are available for this plot:

* *Set fraction:* ‘relative’ sets the total to 1 also when you deselect certain types. ‘Absolute’ keeps the actual fraction and so if you deselect types such as WT, samples total may be lower than 1.
* *Set counts:* ‘relative’ displays the fraction of reads per type in the table below, ‘absolute’ sets the actual number of reads per type.
* *Show data labels:* display the fraction of reads per bar. Will likely disappear in a new version as the information is already in the table underneath.

## Homology plot

This plot displays the homology found at the deletion and tandem duplication junctions.

## Size plot

This plot shows the size changes for all samples as either a heat map or as a violin plot. The user can adjust various settings:

* *Set manual y-axis*: if checked you can use the y-axis maximum slider to set the limits of the y-axis
* *Y-axis maximum:* only functions if ‘Set manual y-axis’ is checked
* *Select Column to plot:* by default both insert and deletion size are shown. You can select also only deletion size or only insertion size
* *Select type of plot:* either show a heat map or a violin representation of the data
* *Set fraction: ‘*relative’ means that all selected types combined add up to 1, whereas ‘absolute’ means if you deselect types the total will be less than 1.

## SNVs plot

Here you can visualize the SNVs that are detected in each sample. Note that you can alter the x-axis length by changing the ‘Filter reads by distance from expected cut site’ setting. SNVs are either shown per sample or combined into a single plot. You can use these additional settings:

* *Plot type:* Combined plot puts all SNVs in a single plot. For a limited number of samples this works fine, but combining many samples looks a bit messy. The alternative is to plot per Alias (per sample). SNVs are then color coded into separate bar graphs.
* *Set y-value:* not set – each graph uses its own limits. ‘Max of plots’ sets the maximum of all plots to the maximum of the most frequent SNVs.
* *SNVs of size:* This option allows you to also include di- and trinucleotide SNVs (up to 5). We noticed that for example in base editor experiments you may want to include these type of events.

## Target plot

Here you can observe for each sample the most frequent deleted bases. You can make either separate plots or combine them into one.

## Sample Info

Here you can assess how many reads per sample are included in the final analysis. You can toggle between the fraction of reads or the absolute number of reads. The top graph provides you with information regarding the correct, not merged and failed filter reads.

The middle graph displays the fraction of correct reads from the total or from the merged reads. The fraction of correct reads from total varies in our hands between 0.4 – 0.9. This is typically caused by sequence errors and/or by primer dimers. A good rule of thumb is to look at various samples from the same amplicon and look for deviations either in ‘read number’ or in ‘fraction correct’.

The lower graph gives you another view on the number of reads that are regarded as correct or not.

## Templated insertions plot

Experimental plot that displays the fraction of DELINS, TINS and their subtype. In addition this plot shows where the templated insertions (TINS) originate from. FW and RC correspond to forward and reverse complement orientation. That is whether the origin of the TINS is found in the reference DNA in forward or reverse complement orientation. These plots have the following options:

* *Select sort:* TINS can be sorted based on start or end position
* *Set y-*value: the maximum y-value is based on unset, the maximum y value of all plots or on the max of a junction type. The maximum y value allows for comparison between genotypes.
* *Set calculated position:* The origin of the insert can be plotted relative to the event junction (set at 0) or relative to the reference. In the latter option you can assess for example if TINS originate preferentially from one side or the other

## Outcome plot

This is the most experimental plot of SIQPlotteR. It is not described in the paper, but the goal is to directly compare the outcome of multiple samples. This can be done by setting one or more samples as a control. Currently there are a few possibilities to make comparison plots:

* Line: the top X outcomes from the controls are plotted with their respective frequencies. Any sample that deviates (by number of SDs) is shown in color.
* Heat map: the top X outcomes from the controls are plotted as a heat map with the remaining samples.
* Pca: principal component analysis. Here the top X outcomes from the controls are used to asses differences between the samples. Similar samples are clustered closer to each other. Underneath a vector plot is shown which displays the contribution of each event class.
* XY: User can select which outcome type to compare. Samples outside the set number of SDs are shown with a label (only for the type on the X axis)
* Umap: similar to the pca analysis, only a umap reduction is performed on the topX outcomes. Like the pca plot, this plot tries to capture differences between samples in a single XY plot. The idea is that the further samples lie apart, the more different they are.

There are some settings that can be tried out, but these settings might change in new versions. The most important settings are:

* *Select Control Sample(s)*: this greatly determines the plots, so ensure that you select multiple control samples here, preferably technical replicates of wild-type samples.
* *Set number of outcomes:* This setting greatly determines the plots. It also takes into account the selected mutation types. We find that especially wild-type fractions vary per experiment, so they are often deselected.

## About

Here you can find details on SIQPlotteR such as the current version and contact information.

# Troubleshooting

Please see if your issue is already listed here and you can hopefully solve it. If not, please contact us via Twitter: @RobinVSchendel or via github [issues](https://github.com/RobinVanSchendel/SIQ/issues):

## SIQ does not start on MacOS

This is likely because MacOS is preventing SIQ from starting. Please hold ‘Control’ and click on SIQ\_1.1.jar (or the current version). Please look carefully at the steps described [here](#_MacOS).

## SIQ does not work on files that need merging of paired-end reads

The program Flash that is required for you operating system seems to be not working. When starting SIQ, FLASH is copied to the directory that SIQ is running in. It might be that the program does not work for your operating system. Please check the output of the following commands to see if Flash is working:

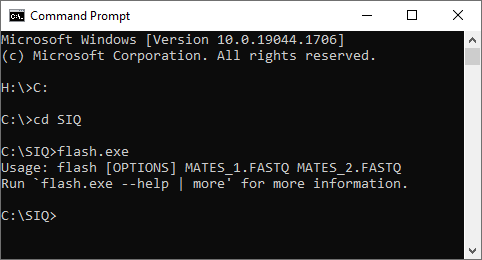


Figure 26 FLASH is working correctly (Windows)

Or in Linux:

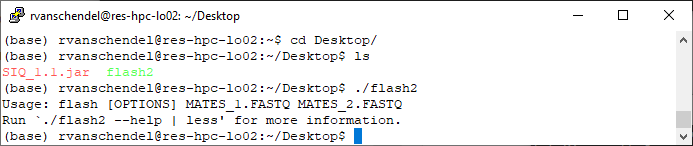


Figure 27 FLASH is working correctly (Linux)