Völundr User Guide

# Introduction

The Völundr bioinformatics pipeline is intended to analyze sgRNA distributions in cell populations for a modified synthetic lethal type assay. Völundr was initially described in a publication looking at genes synthetic lethal with loss of polymerase theta (Feng et al., 2019). This pipeline has been used on libraries containing approximately 3900 sgRNAs up to libraries containing approximately 11,500 sgRNAs. To use this pipeline, it is recommended that the sgRNA library contain at least 500 control sgRNAs. These can either be non-targeting or designed to target intragenic regions. A final important note. This pipeline was designed to result in a low number of false positive calls. Because of the constraints to accomplish that, Völundr was not designed or intended to analyze whole genome screens.

To install and use Völundr read through the requirements listed below and get a working copy of Python installed first. Make sure you are doing all this on a Linux box that is 64-bit architecture. There are currently two methods to install this. The first is to clone or download the package to a location that you have access to. The second is to clone the package and then install using Python setuptools and the setup.py script. Unless otherwise noted ALL sequence position numbering uses the 0 based Python convention.

# CRISPR Library Considerations

Any lentiviral CRISPR vector should work. This pipeline has been used successfully with libraries in the lentiCRISPR v2 (Addgene plasmid # 52961 ; http://n2t.net/addgene:52961 ; RRID:Addgene\_52961, (Sanjana et al., 2014) vector and destabilized Cas9 vectors (Addgene plasmid # 90086 ; http://n2t.net/addgene:90086 ; RRID:Addgene\_90086, (Senturk et al., 2017)). The library should contain at least 500 control sgRNA’s. It is recommended to include 800 to 1000 control sgRNA’s in the library. These can be non-targeting as described in (Feng et al., 2019), non-gene targeting as described in (Morgens et al., 2017) or a combination of both.

**IMPORTANT: The sgRNA’s must have unique names. The recommended convention is Gene\_000 where Gene is the target gene and 000 is a sequential number. Also, the sgRNA controls must be named Organism\_000 where Organism is the target organism such as Mouse or Human. Yes, capitalize the first letter.**

The next consideration for library design is the number of sgRNA’s per gene. Five appears to be the absolute minimum. Ten is highly recommended. This is needed because of variations in the efficacy of the sgRNA’s. Some will not result in efficient cutting. Some will result in very efficient repair. Some will produce mutations in the target that are more or less severe than anticipated. When using 10 sgRNA’s it is not uncommon to find that 2 are always under or overrepresented and two are more neutral than the group as a whole. By having 10 sgRNA’s this becomes much less of an issue during analysis.

The final consideration is how many sgRNA’s to include in the library. This analysis pipeline requires the sgRNA diversity to be balanced and for each of the sgRNA’s to be present at >150 copies at the beginning of the experiment. To allow for variations in transduction and passage steps, the recommended target is >200 copies. For a library of 4,000 sgRNA’s that is 800,000 cells or about 5.2 µg of genomic DNA for the preparation of the sequencing library. While this pipeline has been successfully used on a library containing approximately 11,500 sgRNA’s (Feng et al., 2019), it is recommended to keep the count under 6,000.

# Sequencing Library Considerations

This is a brief explanation of what the sequencing libraries look like and what that means for the analysis. Figure 1 shows a schematic of what the libraries will look like no matter which vector is used. Figure 1A shows the relative location of the PCR primers to the U6 promoter and sgRNA. If possible, place the 3’ end of the forward primer with in 15 to 20 nucleotides of sgRNA sequence. Place the 3’ end of the reverse primer in the first good position ≥ 150 nucleotides 3’ of the forward primer binding site. Figure 1B shows the expected product from the first round PCR. A second round PCR is used to add the sequencing adapters and, in this example, Illumina indices (Figure 1C). The final sequencing library is shown in Figure 1D.

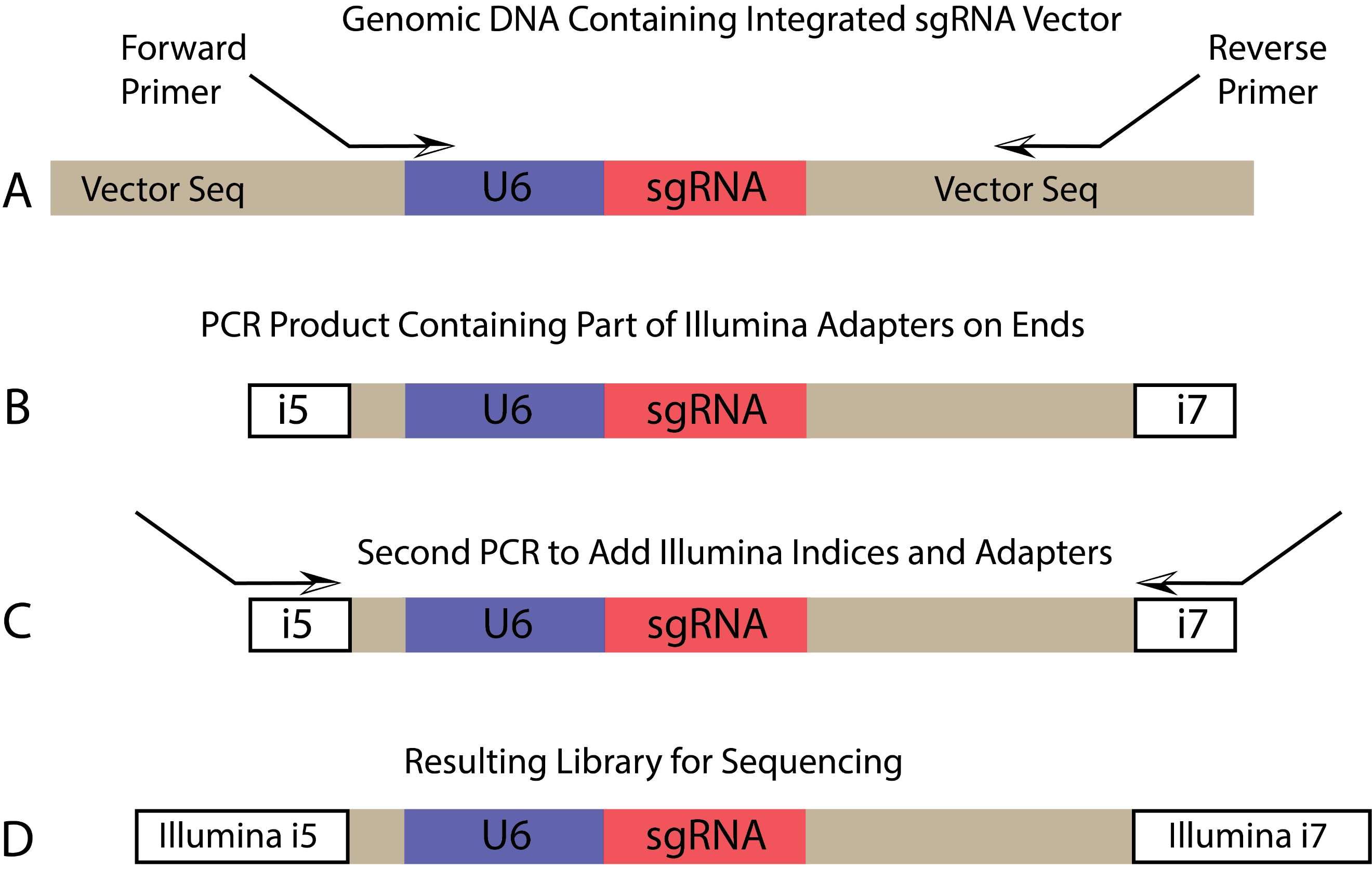


Figure 1

## Sequence Diversity.

These sequencing libraries will have essentially zero diversity. That means for every cycle until the sgRNA in read 1 or read 2 the nucleotide will be the same. For Illumina platforms this is will result in a sequencing failure. To resolve this, we use a pool of four forward and four reverse primers phased 0 to 3 in the first PCR.

# Prerequisites for Installing and Running Völundr

1. Linux OS, tested on RHEL 7.x, Scientific Linux 7.x, and CentOS 7.x. Will possibly run on a Mac OS, although it has not been tested. Will not run on Windows because ScarMapper uses Pysam to parse the reference FASTA file.
2. Ideally ≥45 Gb RAM. The amount of RAM depends on the number of reads being processed and the number of parallel jobs.
3. >4 CPUs/threads. Processes one library per CPU/thread in parallel. The number of parallel jobs possible is CPUs/threads minus 1.
4. The minimum available drive space required is approximately five times the size of the gzipped FASTQ file.
5. Python ≥3.6
6. Required Python packages in no particular order, latest version of each
   1. scipy
   2. natsort
   3. pysam
   4. python-magic
   5. pathos
   6. numpy
   7. python-Levenshtein
   8. setuptools

## Installation

1. Download or clone Völundr from GitHub [https://github.com/pkMyt1/Volundr.git] to a location you have read/write access. Do not attempt to install using the setup.py file.
2. Test installation by moving up to the Völundr directory and executing <python3 volundr.py> (without the <>). You should get the error message below.
   1. usage: volundr.py [-h] --options\_file OPTIONS\_FILE  
       volundr.py: error: the following arguments are required: --options\_file

# Required Files

Völundr requires several support files to run, these are listed and discussed below. Examples of all but the genome reference can be found in the docs folder. The location of these doesn’t matter if you have read/write access.

**FASTQ file.** This can be gzipped or uncompressed. When using the Illumina dual indexing system, it is often less expensive to do paired end sequencing however Völundr only uses read 1.

**Sample Manifest file.** This is a tab delimited text file. Any blank lines or lines beginning with a # are ignored. In the docs folder there is a Microsoft Excel template for this file (*Volundr Sample Manifest Template.xlsx*). To create the sample manifest either copy columns A through D of the template sheet and past them into a text editor or export the template sheet as a tab delimited text file. The example sheet in this file goes over how to fill out the template sheet. The first column, containing the index names, is derived from the first column of the sheet that corresponds to the type of index used. As shown in the example sheet, it is very important that all indices are unique. Völundr will confirm that all indices are unique and will exit if they are not. Second critical item; DO NOT use white space in the sample names. White space is stripped from everything when the file is read. The final point; keep in mind that dog ≠ Dog.

**Master index file.** This is a tab delimited text file. Any blank lines or lines beginning with a # are ignored. To create this file, copy the first three columns from the sheet in the Microsoft Excel template (*Volundr Sample Manifest Template.xlsx*) that corresponds to your library indexing method. When using the Illumina dual indexing approach, the master index file should look like table 1. Custom indexing methods are possible. This requires tweaking the block of code in Völundr that processes the FASTQ file. Contact the author for help on doing this.

|  |  |  |
| --- | --- | --- |
| **# Index\_ID** | **Forward (D7s)** | **Reverse (D5s)** |
| D501+D701 | ATTACTCG | TATAGCCT |

**Table 1.** First two lines from the master index file for Illumina dual indexing.

**Target file.** This is a tab delimited text file. Any blank lines or lines beginning with a # are ignored. This file contains the name and sequence of the sgRNA’s in the library. The first two columns in the file are used by Völundr. Any additional columns are ignored. There are three different sgRNA sequence formats allowed. These cannot be mixed. Table 2A is from the target file used in Feng et al. (Feng et al., 2019). Table 2B is the simplest example where the sequence shown is the sgRNA sequence and they are all the same length. Table 2C is a more complex case where the length of the sgRNA sequences is variable. This example was reported in need reference for this and used in Feng et al. (Feng et al., 2019).

**A.**

|  |  |  |
| --- | --- | --- |
| # Gene | Oligo\_Seq | mouse\_DDR\_ID |
| Aicda\_1 | GTGGAAAGGACGAAACACCGACCATTTCAAAAATGTCCGCGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGG | mDDR\_1 |

**B.**

|  |  |  |
| --- | --- | --- |
| # Gene | Oligo\_Seq | sgRNA ID |
| AMFR\_1 | GCTGCGCGGAAACTGCCCTG | Vaziri\_E3\_1 |

**C.**

|  |  |  |  |
| --- | --- | --- | --- |
| # Gene | Oligo\_Seq | Code | Ensembl ID |
| Ackr2\_261.1 | GGAACCGACGGTGGTGAG | MPA | ENSMUSG00000044534 |

**Table 2.** Three different allowed sequence formats for the target file. A. Sequence of the oligonucleotide used to clone the sgRNA. In this example the sgRNA sequence begins at position 20 and is 20 nucleotides long. B. Example where only the sgRNA sequence is in the target file. C. This is an example where only the sgRNA sequence is listed but the length of the sgRNA sequences is variable.

NOTE: The name (Gene column) of all control sgRNA’s in this file must begin with the target organism listed in the –Species option in the Options\_File. For example, --Species Mouse will have control sgRNA’s Mouse\_123 where 123 can be a number or anything else as long as there is no white space.

**Options file.** This is a tab delimited text file. Any blank lines or lines beginning with a # are ignored. This file provides all the user configurable parameters Völundr requires. There are two blank options files in the doc folder. One is for doing the sgRNA target search and the other is for doing statistical analysis on the data from the target search. There is also a GUI available [https://github.com/pkMyt1/Volundr\_GUI.git] for writing the options file. As explained here, it is also used as a bash shell file to start Völundr. It is not necessary to use it as a shell file. The parts of the options file are described in detail below.

## Options File

This must be a formatted as a tab delimited text file. In the docs folder you will find options file templates *run\_Volundr\_Target\_Search.sh* and *run\_Volundr\_Statistics.sh* for doing a sgRNA target search and analysis respectively. The options file is a shell script file that starts the program and passes parameters to the program. It can also be used to pass commands to computing cluster load balancers such as SLURM or LSF. The options for each file are described separately below.

### Target Search Options File

--TargetSearch

This must be True. Options are True or False.

--Statistics

This must be False. Options are True or False.

--FASTQ1

Full path to FASTQ1

--Master\_Index\_File

Full path to the master index file described above.

-- SampleManifest

Full path to the sample manifest file.

--Target\_File

Full path to the target file.

--Working\_Folder

Full path to a working folder. This is where Völundr will write the output files.

--Verbose

Default value is INFO. Set verbosity level. Options are INFO, DEBUG, ERROR. For general runs leave this at INFO. DEBUG will limit the reads analyzed to 2 million.

--Job\_Name

Provide a name for a run. This name will be incorporated into the output files simplifying identification. For this reason avoid special characters such as (`~!@#$%^&(\*}{). The job name must NOT contain white space.

--Spawn

Defines how many parallel jobs to create. Each library defined in the sample manifest file will be processed in its own job. Max setting should be number of CPUs/threads – 1. Minimum value is 1.

--Demultiplex

Default value is False. Allowed values are True or False. Defines if demultiplexed FASTQ files are written. If the files are written, they are automatically compressed with gzip with a compression setting of 9.

--Species

Current allowed values are Mouse or Human. Make sure the control sgRNA names in the target file matches this.

--Platform

Currently Illumina, Ion, or Custom. This tells Völundr how to identify and process the indices. Note: Custom does not work out of the box. Requires minor code changes to match the indexing method used.

-- Analyze\_Unknowns

Default value is False. Allowed values are True or False. If set to True, Völundr will do a target search on the reads that do not contain identifiable indices.

-- RevComp

Default value is False. Allowed values are True or False. If the sgRNA’s are on the opposite strand as FASTQ1 then set to True.

-- Target\_Mismatch

Default is 1. Any integer is allowed. Defines the maximum number of mismatches allowed during the sgRNA target search. Generally, one should not exceed a value of 3.

--Min\_Length

Default is 120. Minimum sequence length of read to analyze. This value is important for Ion platforms and possibly Custom designs. Ion platforms have the index as part of the read and will yield sequences shorter than expected. Custom designs could also have the index as part of the read.

--Target\_Length

Default is 20. Allowable values are any integer or “Variable”, no quotes. This depends on which type of target file you have. Those described in Table 2A and B have fixed length sgRNA sequences. If using them set this value to that length. The type in Table 2C has a variable length sgRNA and requires “Variable”.

--Target\_Start

Default value is 20. Allowed values are any integer. For targets from Table 2B enter 0.

--Index\_Mismatch

The default value is 1. Allowed values are any integer. Defines how many allowable mismatches when demultiplexing.

--Target\_Padding

The default value is 2. Allowed values are any integer. If Völundr cannot locate the anchor sequence (see below) this defines how many nucleotides to allow for some “wiggle” when attempting to salvage the search for the sgRNA.

--Expected\_Position

The default value is 61. Allowed values are any integer. This is the nucleotide in the read where the sgRNA sequence is expected to start. Note: Völundr uses Python 0 based numbering here so the first nucleotide is position 0 not position 1.

--AnchorSeq

The default value is AAACACCG. Allowed values are any nucleotide string, all caps. This is a sequence chosen by the user that is used to “anchor” the sgRNA search. Here is an example for defining the anchor sequence and its relationship to the sgRNA. The sequence in yellow is the forward primer binding site shown in Figure 1A. The first nucleotide is the expected start site of the sequence. The NNN in red is the sgRNA sequence. The **AAACACCG** is the sequence used for the anchor.

CGATTTCTTGGCTTTATATATCTTGTGGAAAGGACG**AAACACCG**NNNNNNsgRNANNNNNNNNNGTTTTAGAGCTAGAAATAGCAA

--AnchorMismatch

The default value is 1. Allowed values are any integer. Defines how many allowable mismatches for the anchor sequence search.

--AnchorStart

The default value is 35. Allowed values are any integer. Nucleotide position in the read where Völundr will start searching for the anchor sequence.

-- AnchorSop

The default value is 65. Allowed values are any integer. Nucleotide position in the read where Völundr will stop searching for the anchor sequence and try to find the sgRNA sequence based on the expected position.

--Index\_Mismatch

The default value is 1. Allowed values are any integer. Defines how many allowable mismatches when demultiplexing.

--OutputPositionFrequencyData

Default is False. Allowed values are True or False. Determines if additional files are written. These may or may not be useful.

### Statistics Options

--TargetSearch

This must be False. Options are True or False.

--Statistics

This must be True. Options are True or False.

--DataFiles

Full path to the folder containing the counts files. Can be the same as the working folder

--Master\_Index\_File

Full path to the master index file described above.

-- SampleManifest

Full path to the sample manifest file.

--Target\_File

Full path to the target file.

--Working\_Folder

Full path to a working folder. This is where Völundr will write the output files.

--Verbose

Default value is INFO. Set verbosity level. Options are INFO, DEBUG, ERROR. For general runs leave this at INFO. DEBUG will limit the reads analyzed to 2 million.

--Job\_Name

This must be the same job name used during the target search.

--Species

Current allowed values are Mouse or Human. Make sure the control sgRNA names in the target file matches this.

--Control\_Sample

Sample to be used as the experimental control. Generally, this would be wild-type cells or untreated cells. This is a sample name from the sample manifest file.

--Library\_Control

Sample to be used as the library control. This sample is used to find and mask sgRNA’s that are underrepresented in the library. This could be the plasmid used to package the library, a no Shield1 control when using the destabilized Cas9 libraries, or a day 1 transduced cell population. Make sure the name matches a sample name in the sample manifest file.

--Bad\_sgRNA\_Percentile

Default is 2.5, value is a float. This defines the lower percentile limit for a guide to be included in the analysis. The value is calculated from the normalized (TCnorm) values of the library control.

--UpperPercentile

Default is 97.5, value is a float. Part of the statistical tests. This is the upper limit for the empirically derived null distribution. It comes from random sampling the permutations of the control sgRNA values.

--LowerPercentile

Default is 2.5, value is a float. Part of the statistical tests. This is the lower limit for the empirically derived null distribution. It comes from random sampling the permutations of the control sgRNA values.

-- PermutationCount

Default is 100000. Value is an integer. This defines the number of permutations Völundr will run to determine the empirically derived null distribution.

--Alpha

Default is 0.1, value is a float. This is the error rate used in the Benjamini, Krieger and Yekuteli two stage linear step-up procedure with estimation of number of true hypotheses. It is left up to the user to determine if a multiple sample correction is necessary or not. The output always presents the original p-value as well as the corrected p-value.

--Target\_Mismatch

Default is 1. The value is an integer. From the counts files, how many mismatches are going to be allowed for the analysis. Must be a value that is in the raw counts files.

--Write\_TDnorm\_Log2\_sgRNA\_Control\_File

Default is False. Allowed values are True or False. When True Völundr will output a file containing the Log2 transformed TDnorm values for the sample control.

--Write\_TDnorm\_Log2\_sgRNA\_Sample\_File

Default is False. A llowed values are True or False. When True Völundr will output a file containing the Log2 transformed TDnorm values for each of the samples from the sample manifest.

--Write\_Log2\_sgRNA\_File

Default is False. Allowed values are True or False. When True Völundr will output a single file containing the Log2 transformed TDnorm values for the samples from the sample manifest.

--Write\_Permuted\_Log2\_Data\_File

Default is False. Allowed values are True or False. When True Völundr will the data from each permutation for deriving the empirical null distribution.

## Output Files

Völundr writes different output files depending on which module is being run and which options the user has chosen. The files will always be written to the working folder. All output files are tab delimited text files.

### Target Search

In the docs folder there are several Microsoft Excel files beginning with the “*Synthetic Lethal Template*”. These templates can be used to have the summary and target counts data in a single file.

#### <Job\_Name>\_summary.txt

This file is always written. The file contains summary information about the target search run. In the docs folder, file “*Example\_summary.txt”* shows what the output looks like.

#### <Job\_Name>\_<Sample\_Index\_Name>\_target\_counts.txt

These files are always written. One file for each sample in the sample manifest. In the docs folder, file “*Example\_target\_counts.txt*” shows the output from a run allowing 1 mismatch in the sgRNA sequence. If two mismatches were stipulated, then there would be an additional column showing the counts for two mismatches. These files are required for the statistics module. DO NOT delete these text files.

#### <Job\_Name>\_<Sample\_Index\_Name>\_Target\_Position\_Freq.txt

This is an optional output file. One of these is written for each sample in the sample manifest. In the docs folder the “*Example\_Target\_Position\_Freq.txt*” file shows a typical set of data. It contains frequency distributions of total anchor sequence found and where, total sgRNA sequence found and where, as well as total reads with no sgRNA and where the searches began. Intended for debugging purposes.

### Statistics

All files written here are tab delimited text files. The statistics module has the option to output several files of intermediate data. As stated for the options above, generally these are only helpful when debugging.

#### <Job Name>\_<Library Control>\_Masked\_Targets.txt

This file is always written. It contains information about which sgRNA’s did not pass the cutoff supplied in the options file. In the docs folder, “*Example\_Masked\_Targets.txt*” shows what it looks like. Keep in mind, it is possible that this file will be empty if all sgRNA’s pass.

#### <Job Name>\_<Sample Name>\_KS3\_Log2\_Delta\_Genes.txt

This file is always written. It contains the results of the statistical tests. These results are always the delta of the experimental sample with the sample control. In the docs folder, “*Example\_KS3\_Log2\_Delta\_Genes.txt*” shows what the output file looks like. Hopefully most of the data should be self-explanatory. The “*Upper Null Set Limit*” is the value corresponding to the upper limit in the options file for the empirical null distribution. The “*Lower Null Set Limit*” is the value corresponding to the lower limit in the options file for the empirical null distribution. Table 3 shows the first two rows from the example file.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene | Log2 | Original p-value | Corrected p-value | Reject Null Hypothesis |
| Aicda | 0.247071 | 0.452936 | 0.826048 | False |

**Table 3.**

# References

Feng, W., Simpson, D.A., Carvajal-Garcia, J., Price, B.A., Kumar, R.J., Mose, L.E., Wood, R.D., Rashid, N., Purvis, J.E., Parker, J.S.*, et al.* (2019). Genetic determinants of cellular addiction to DNA polymerase theta. Nature communications *10*, 4286.

Morgens, D.W., Wainberg, M., Boyle, E.A., Ursu, O., Araya, C.L., Tsui, C.K., Haney, M.S., Hess, G.T., Han, K., Jeng, E.E.*, et al.* (2017). Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. Nature communications *8*, 15178.

Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods *11*, 783-784.

Senturk, S., Shirole, N.H., Nowak, D.G., Corbo, V., Pal, D., Vaughan, A., Tuveson, D.A., Trotman, L.C., Kinney, J.B., and Sordella, R. (2017). Rapid and tunable method to temporally control gene editing based on conditional Cas9 stabilization. Nature communications *8*, 14370.