Off-targets analysis report

Corre Guillaume

2025-11-04

Contents

Working directory : **/media/DATA/projects/GENETHOFF/test_dataset**

Pipeline version: **V1.0**

The analysis was performed **with** bulge tolerance.

The quantification is based on **UMI counts** .

All parameters can be found in Configuration settings at the end of this document.

1 Description of libraries

Table 1: Table 1

library	Cells	Genome	gRNA	gRNA sequence	PAM	Cas	type	orientation
VEGFA_s1_K562_neg	K562	GRCh38p14	VEGFA_s1	GGGTGGGGGGAGTTTGCTCC	NGG	Cas	guideseq	negative
VEGFA_s1_K562_pos								positive

2 Statistics of reads processing

Table 2: Table 2

library	Demultiplexed	ODN_match	Filtered
VEGFA_s1_K562_neg	2,297,044 (100%)	2,116,070 (92.12%)	1,894,227 (82.46%)
VEGFA_s1_K562_pos	1,175,378 (100%)	1,092,756 (92.97%)	1,077,959 (91.71%)

Read-pairs with length greater than **25** bp (both of the pair) were considered for analysis.

All percentages represents % of demultiplexed reads

3 Reads alignment, cut sites calling and clustering

Summary of the alignment step, calling step and clustering of cutting sites including :

- The number of reads aligned on the genome
- The number of UMIs detected (estimation of total number of events)
- The number of unique ODN insertion sites
- The number of clusters.

Table 3: Table 3

				Clusters count			
library	Reads	UMIs	Insertions	Clusters	With gRNA match ..	And ..	count
VEGFA_s1_K562_neg	588,312	26,074	7,400	7,322	132	.. 2 PCR orientations	0
						.. 2 ODN orientations	8
						.. In Oncogene	0
VEGFA_s1_K562_pos	432,763	18,879	5,039	5,018	119	.. 2 PCR orientations	0
						.. 2 ODN orientations	3
						.. In Oncogene	0

Cut sites were identified from the alignment start position of R2 reads.

- Reads were aggregated if they share the exact same start position and the same UMI sequence.
- UMI were corrected using the **Adjacency** method with a Hamming distance tolerance of **1**.
- Positions/UMI with more than **3** reads were considered for next step.

Clusters are defined as a group of ODN insertion sites within a distance smaller than **100** bp and characterized by :

- Match of the crRNA sequence with less than **6** edits withing the cluster boundaries +/- **50** bp ("With gRNA match . . ." in table below)
- Presence of reads aligning in both directions, indicating both ODN orientations ("2 ODN orientations" in table below)
- Presence of reads from 2 PCR orientations if available ("2 PCR orientations" in table below)
- Number of clusters overlapping an oncogene.

Clusters with more than **2** total UMIs were considered.

Table 4: Table 4

library	Position	Cut offset	Edits_gRNA	Edits_PAM	N_UMI_cluster	Relative_abundance	Rank
VEGFA_s1_K562_neg	chr6:43769560	-1	0	0	369	31.98	1
VEGFA_s1_K562_pos	chr6:43769560	-2	0	0	115	19.39	1

4 Best match(es)

For each library, the best match(es) are clusters with the minimal number of edits in the gRNA and PAM sequences (which may not necessarily be 0).

- “Position” : Theoretical cutting site based on gRNA alignment to gDNA and nuclease offset.
- “Cut offset” : Difference between theoretical cutting site and most frequent cutting site in the cluster.
- Edits* : Number of INDEL+mismatches
- “Relative Abundance” : Contribution of cluster abundance in % of total UMIs count. Only clusters with gRNA match with less than **6** edits and more **2** UMIs are considered.
- Rank: Rank of the best candidate(s) based on UMI count.

4.1 Rank-abundance curve

The rank-abundance curve provides insights into relative cluster abundance - a steep curve indicates dominance by a few clusters, while a shallow curve suggests more even distribution among different clusters.

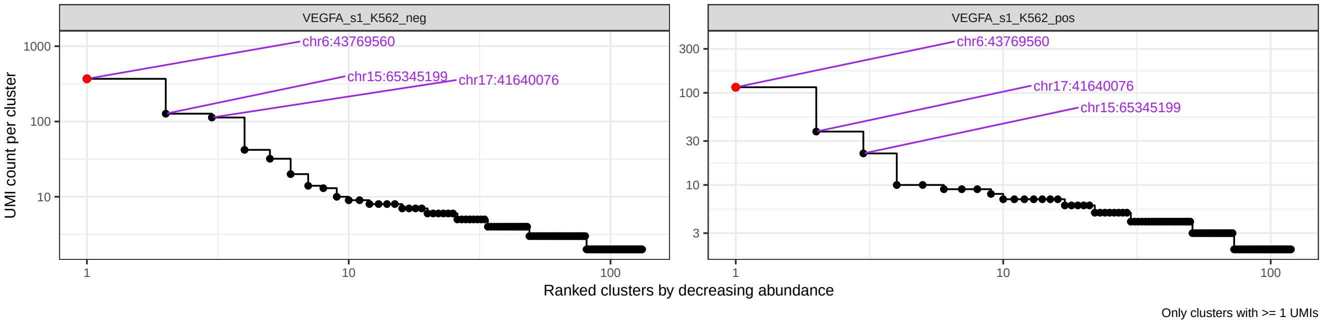

Figure 1: Figure 1

- Only clusters with gRNA match and more than **1** total UMIs are plotted.
- `Red` dots correspond to clusters with `minimal edits` (see table 4).
- Top3 most abundant clusters (UMI counts) are labeled.

4.2 Distribution of cut sites around best candidates position

For each best candidate cluster, plot the UMI count detected around the gRNA theoretical cut site (dashed line) for each cut site, in the forward or reverse ODN orientation (blue and red bars respectively).

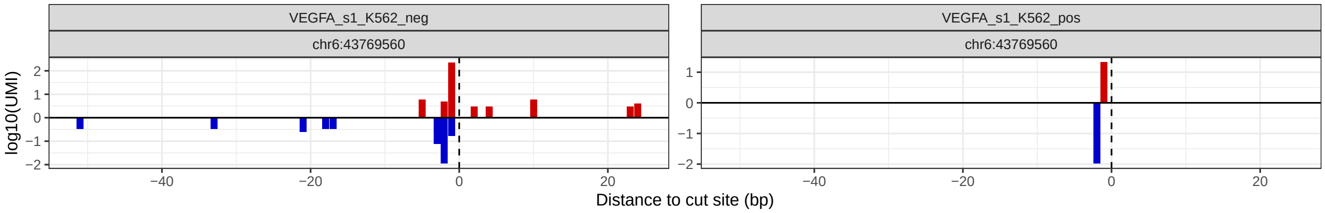

Figure 2: Figure 2

5 Genome distribution of clusters

This figure represents the distribution of unique clusters `with gRNA` match per chromosome, colored by `prediction` status. Only clusters with number of edits (INDELs and substitutions) smaller or equal to **6** are considered.

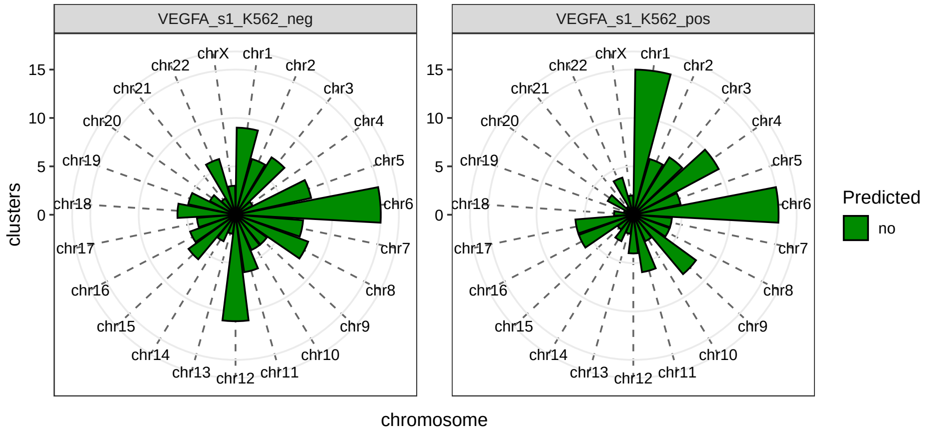

Figure 3: figure 3

This figure represents the total number of UMIs (cells) per chromosome from clusters with gRNA match, colored by `prediction` status. Only clusters with number of edits (INDELs and substitutions) smaller or equal to **6** are considered.

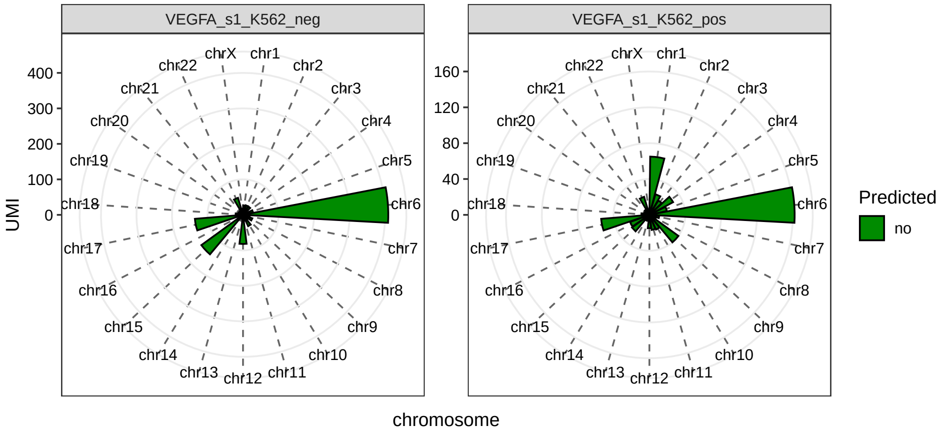

Figure 4: figure 4

6 Configuration settings

```
author: "Guillaume CORRE, PhD"
affiliation: "Therapeutic Gene Editing - GENETHON, INSERM U951, Paris-Saclay University, Evry, France"
contact: "gcorre@genethon.fr"
version: 'V1.0'

clean_intermediates_files: 'FALSE'
skip_demultiplexing: 'FALSE'
quantification: "umi" # use "umi" or "fragment" for the quantification (only if rescue_R2 is 'no")
rescue_R2: "FALSE"     # Rescue R2 reads if R1 is too short or the pair doesn't align properly.

## Library informations
sampleInfo_path: "test_datasheet.csv"
```

```
read_path:  ""
R1: "Undetermined_S0_L001_R1_001.fastq.gz"
R2: "Undetermined_S0_L001_R2_001.fastq.gz"
I1: "Undetermined_S0_L001_I1_001.fastq.gz"
I2: "Undetermined_S0_L001_I2_001.fastq.gz"


################################################
## path to references
genome:
  GRCh38p14:
    fasta: "/media/DATA/references/human/GRCh38p14/gencode/fasta/GRCh38.primary_assembly.genome.chr.fa"
    index: "/media/DATA/references/human/GRCh38p14/gencode/index/GRCh38.primary_assembly.genome.chr" # path to r
    annotation: "/media/DATA/references/human/GRCh38p14/gencode/annotations/gencode.v49.primary_assembly.basic.a
    oncogene_list: "/media/DATA/projects/GENETHOFF/02-ressources/OncoList_OncoKB_GRCh38_2025-07-04.tsv"
  mouse:
    fasta: ""
    index: ""
    annotation: ""
################################################

################################################
minLength: 25 ## Minimal read length after trimming, before alignment
################################################

## Alignement
################################################
aligner: "bowtie2"   ## Aligner to use (bowtie2 or bwa)
minFragLength: 100         # Minimal fragment length after alignment
maxFragLength: 1500        # Maximal fragment length after alignment

################################################

################################################
# post alignment
minMAPQ: 20                       # Min MAPQ score to keep alignments
UMI_hamming_distance: 1         # min distance to cluster UMI using network-based deduplication, use [0] to keep
UMI_deduplication: "Adjacency"  # method to correct UMI (cluster or Adjacency)
UMI_pattern: "NNWNNWNN"
UMI_filter: "TRUE"              # If TRUE, remove UMIs that do no match the expected pattern [FALSE or TRUE]
UMI_side: "3"                   # position of the UMI in the I2 read sequence: 5 = UMI+BC, 3 = BC+UMI
################################################


################################################
## Off targets calling
tolerate_bulges: "TRUE"           # whether to include gaps in the gRNA alignment (this will change the gap pena
max_edits_crRNA: 6                # filter clusters with less or equal than n edits in the crRNA sequence (edits
ISbinWindow: 100                  # insertion sites closer than 'ISbinWindow' will be clustered together
minReadsPerUMI: 3                 # Min number of reads per UMIs (>=)
minUMIPerIS: 2                    # Min number of UMI per IS (>=)
slopSize: 50                      # window size (bp) around IS (both directions) to identify gRNA sequence (ie 5
min_predicted_distance: 100       # distance between cut site and predicted cut site to consider as predicted
################################################

################################################
# reporting
max_clusters: 100                 # max number of cluster alignments to report
minUMI_alignments_figure: 1       # filter clusters with more than n UMI in the report alignment figure (set to
################################################
```

```
# Prediction
################################################
SWoffFinder:
  path: "/opt/SWOffinder" ## Path to SWoffinder on your server (downloaded from https://github.com/OrensteinLab/
  maxE: 6                 # Max edits allowed (integer).
  maxM: 6                 # Max mismatches allowed without bulges (integer).
  maxMB: 6                # Max mismatches allowed with bulges (integer).
  maxB: 3                 # Max bulges allowed (integer).
  window_size: 100
################################################



################################################
# Sequences for the trimming steps

guideseq:
  positive:
    R1_trailing: "GTTTAATTGAGTTGTCATATGT"
    R2_leading: "ACATATGACAACTCAATTAAAC"
    R2_trailing: "AGATCGGAAGAGCGTCGTGT"
  negative:
    R1_trailing: "ATACCGTTATTAACATATGACAACTCAA"
    R2_leading: "TTGAGTTGTCATATGTTAATAACGGTAT"
    R2_trailing: "AGATCGGAAGAGCGTCGTGT"



iguideseq:
  positive:
    R2_leading: "ACATATGACAACTCAATTAAACGCGAGC"
    R2_trailing: "AGATCGGAAGAGCGTCGTGT"
    R1_trailing: "GCTCGCGTTTAATTGAGTTGTCATATGT"
  negative:
    R1_trailing: "TCGCGTATACCGTTATTAACATATGACAACTCAA"
    R2_leading: "TTGAGTTGTCATATGTTAATAACGGTATACGCGA"
    R2_trailing: "AGATCGGAAGAGCGTCGTGT"



olitagseq:
  positive:
    R1_trailing: "GGGGTTTAATTGAGTTGTCATATGTT"
    R2_leading: "AACATATGACAACTCAATTAAACCCC"
    R2_trailing: "TCCGCTCCCTCG"
  negative:
    R1_trailing: "CCCATACCGTTATTAACATATGAC"
    R2_leading: "GTCATATGTTAATAACGGTATGGG"
    R2_trailing: "TCCGCTCCCTCG"

tagseq:
  positive:
    R1_trailing: "TGCGATAACACGCATTTCGCATAA"
    R2_leading: "CTTATGCGAAATGCGTGTTATCGCA"
    R2_trailing: "AGATCGGAAGAGCGTCGTGT"
  negative:
    R1_trailing: "ATCTCTGAGCCTTATGCGAAATGC"
    R2_leading: "CGCATTTCGCATAAGGCTCAGAGAT"
    R2_trailing: "AGATCGGAAGAGCGTCGTGT"
################################################
```

7 R session informations

R version 4.4.3 (2025-02-28) Platform: x86_64-conda-linux-gnu Running under: Debian GNU/Linux 11 (bullseye)

Matrix products: default BLAS/LAPACK: /home/gcorre/miniconda3/envs/GENETHOFF/lib/libopenblasp-r0.3.30.so; LAPACK version 3.12.0

locale: [1] LC_CTYPE=fr_FR.UTF-8 LC_NUMERIC=C
[3] LC_TIME=fr_FR.UTF-8 LC_COLLATE=fr_FR.UTF-8
[5] LC_MONETARY=fr_FR.UTF-8 LC_MESSAGES=fr_FR.UTF-8
[7] LC_PAPER=fr_FR.UTF-8 LC_NAME=C
[9] LC_ADDRESS=C LC_TELEPHONE=C
[11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C

time zone: Europe/Paris tzcode source: system (glibc)

attached base packages: [1] stats graphics grDevices utils datasets methods base

other attached packages: [1] UpSetR_1.4.0 yaml_2.3.10 kableExtra_1.4.0 lubridate_1.9.4 [5] forcats_1.0.1 stringr_1.5.2 dplyr_1.1.4 purrr_1.1.0
[9] readr_2.1.5 tidyr_1.3.1 tibble_3.3.0 ggplot2_4.0.0
[13] tidyverse_2.0.0 rmdformats_1.0.4

loaded via a namespace (and not attached): [1] generics_0.1.4 xml2_1.4.0 stringi_1.8.7 hms_1.1.4
[5] digest_0.6.37 magrittr_2.0.4 evaluate_1.0.5 grid_4.4.3
[9] timechange_0.3.0 RColorBrewer_1.1-3 bookdown_0.45 fastmap_1.2.0
[13] plyr_1.8.9 ggrepel_0.9.6 gridExtra_2.3 viridisLite_0.4.2 [17] scales_1.4.0 textshaping_1.0.3 cli_3.6.5 rlang_1.1.6
[21] withr_3.0.2 tools_4.4.3 tzdb_0.5.0 vctrs_0.6.5
[25] R6_2.6.1 lifecycle_1.0.4 pkgconfig_2.0.3 pillar_1.11.1
[29] gtable_0.3.6 glue_1.8.0 Rcpp_1.1.0 systemfonts_1.3.1 [33] xfun_0.53 tidyselect_1.2.1 rstudioapi_0.17.1 knitr_1.50
[37] farver_2.1.2 htmltools_0.5.8.1 labeling_0.4.3 rmarkdown_2.30
[41] svglite_2.2.1 compiler_4.4.3 S7_0.2.0