CALITAS: a CRISPR-Cas-aware ALigner for *In silico* off-TArget Search

Tim Fennell[1], Deric Zhang[2], Meltem Isik[2], Tongyao Wang[2], Gregory Gotta[2], Christopher J. Wilson[2], **Eugenio Marco**[2]

[1]Fulcrum Genomics, [2]Editas Medicine, Inc. | 11 Hurley Street, Cambridge, MA 02141

Introduction

CRISPR/Cas-based medicines are being developed to treat serious diseases and their safety evaluations must include specificity assessments

In silico off-target prediction methods are used together with experimental methods like GUIDE-seq and Digenome-seq to assemble lists of candidate off-target sites

Current *in silico* off-target prediction tools lack critical features such as addition of unlimited bulges on one or both strands within an alignment, flexible PAM or PAMless searches with tolerability for PAM mismatches, output of a single best alignment without the need for further bed file manipulation, and de novo inclusion of variants via a user supplied VCF file

To address these issues, we have developed CALITAS, a freely available, fast, software package that uses a modified and tuned version of the classic Needleman-Wunsch algorithm. CALITAS can align gRNA sequences to user-provided regions, and also perform genome wide on- and off-target searches

Features

- Similar penalties for mismatches, gRNA and DNA gaps
- Unlimited gaps (or bulges) can be aligned on both strands
- User-defined maximum number of gRNA mismatches and gaps
- Mismatches in the PAM are tolerated
- Ability to use multiple PAM sequences or no PAM
- Option to produce either the single best alignment per off-target site or all alignments meeting mismatch/gap limits
- Ability to set base pair overlap cutoff for differentiating unique adjacent alignments
- Ability to align to user-provided regions or search genome wide
- Ability to align against alternate alleles in the reference, via user-provided VCF files, for example from the 1000 Genomes Project

Results

CALITAS uses a modified Needleman-Wunsch algorithm

CALITAS uses net cost parameters to give similar penalties for mismatches, gRNA gaps and DNA gaps (something not found in standard NW aligners)

User-tunable net cost parameters

```
guide-mismatch-net-cost = -120
guide-gap-net-cost = -121
genome-gap-net-cost = -122
pam-mismatch-net-cost = -260
```

Internal NW penalties

```
guide match = 60
PAM match = 130
guide mismatch = -60
gRNA gap = -61
DNA gap = -2
PAM mismatch = -130
```

Alignment example	Internal NW calculation	Net cost	CALITAS score (Higher is better)
Perfect Match tttnAGGAAACTTCTGGCAGGACC |||||||||||||||||||| TTTCAGGAAACTTCTGGCAGGACC 4 PAM matches + 20 guide matches	4 * 130 + 20 * 60	0	**1720**
1 Mismatch tttnAGGAAACTTCTGGCAGGACC |||||.||||||||||||| TTTCAGGAAACTTCTGGCAGGACC 4 PAM matches + 19 guide matches 1 guide mismatch	4 * 130 + 19 * 60 − 60	−120	**1600**
1 gRNA gap (genome bulge) tttnAGG-AAACTTCTGGCAGGACC |||||||~|||||||||||| TTTCAGGTAAACTTCTGGCAGGACC 4 PAM matches + 19 guide matches 1 gRNA gap	4 * 130 + 19 * 60 − 61	− 121	**1599**
1 DNA gap (guide bulge) tttnAGGAAACTTCTGGCAGGACC |||||||~|||||||||||| TTTCAGGA-ACTTCTGGCAGGACC 4 PAM matches + 18 guide matches 1 DNA gap	4 * 130 + 18 * 60 − 2	− 122	**1598**
1 PAM mismatch tttnAGGAAACTTCTGGCAGGACC |.||||||||||||||||||| TATCAGGAAACTTCTGGCAGGACC 3 PAM matches + 20 guide matches 1 PAM mismatch	3 * 130 + 20 * 60 − 130	− 260	**1460**

CALITAS can predict off-targets from genomic variants from a VCF file

A CALITAS alignment to hg38 + variants

Step 1

Identify individual variants/alleles that are present above some threshold frequency (e.g. 1%)

Filtered variants 1000 Genomes: rs75468119:54989257:A>T:0.019, rs2371098:54989270:A>G:0.629

Step 2

Identify putative short-range haplotypes by linking variants that are within 1 guide length of each other

hg38

```
TCAGAAATGAGATAGATCTGGGGAAGGGACTGAG
          T          G
rs75468119 A>T:0.019   rs2371098 A>G:0.629
```

Step 3

Assemble modified sequences that include the individual variants and/or haplotype variants

hg38 + variants

```
TCAGAAATGAGATAGTTCTGGGGAAGGGACTGAG AF:0.019
TCAGAAATGAGATAGATCTGGGGAAGGGGCTGAG AF:0.629
TCAGAAATGAGATAGTTCTGGGGAAGGGGCTGAG AF:0.019
```

Step 4

Search the modified sequences for putative alignments

1000 Genomes alignment

```
ATTGAGATAG-TGTGGGGAAGnrg
|.|||||||~|.||||||||||||
TCAGAAATGAGATAGATCTGGGGAAGGGGCTGAG
2 guide mismatches + 1 gRNA gap AF:0.019

ATTGAGATAG-TGTGGGGAAGnrg
|.|||||||~|.||||||||||||
TCAGAAATGAGATAGTTCTGGGGAAGGGGCTGAG
2 guide mismatches + 1 gRNA gap AF: 0.629
```

B CALITAS alignment to hg38

hg38 alignment

```
ATTGAGATAG-TGTGGGGAAGnrg
|.|||||||~|.||||||||||.
TCAGAAATGAGATAGATCTGGGGAAGGGACT
```

2 guide mismatches + 1 gRNA gap + 1 PAM mismatch

(A) Shown are the steps CALITAS uses to build short-range haplotypes from variants from a VCF file, assign allele frequencies, and search for putative alignments. Shown are two alignments with 2 guide mismatches and 1 gRNA gap (B) CALITAS alignment in the same region as in (A) but in the reference genome hg38, where the alignment has 2 guide mismatches, 1 gRNA gap and 1 PAM mismatch

AsCas12a has fewer predicted off-targets by CALITAS, followed by SaCas9 and SpCas9

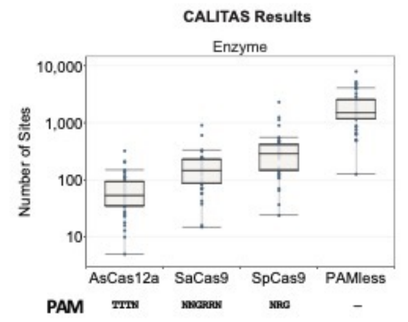
We used CALITAS to make *in silico* predictions for 41 gRNAs with PAMs for AsCas12a, SaCas9 and SpCas9

AsCas12a PAM: TTTC

SaCas9 / SpCas9 PAM: AGG

TTTCAGGAAACTTCTGGCAGGACCAGGGAT

Sites up to 3 Mismatches plus Gaps

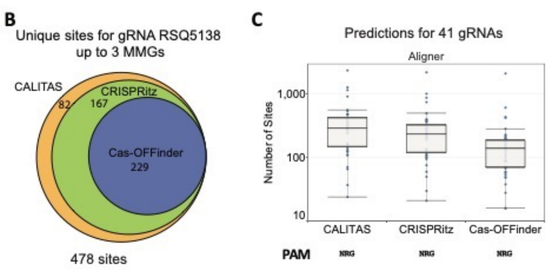


Number of CALITAS-identified off-target sites up to 3 MMGs for 41 guides, using the PAMs: TTTN for AsCas12a, NNGRRN for SaCas9, NRG for SpCas9 and no PAM for PAMless

CALITAS predicts more off-target sites than CRISPRitz or Cas-OFFinder and simplifies handling of off-target results

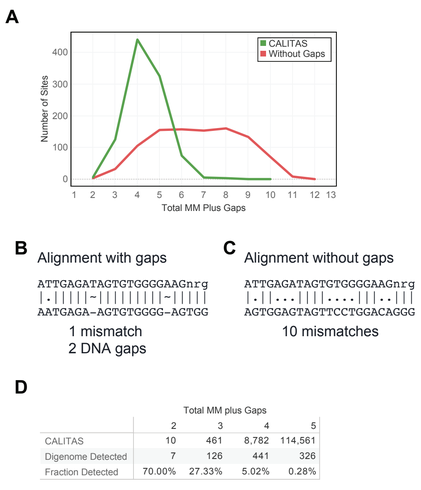
A Predictions for gRNA RSQ5138

Aligner	Parameters	Total Alignments Produced	Total Alignments Produced up to 3 MMGs	Total Unique Sites up to 3 MMGs
CALITAS	max-total-diffs=3	478	478	478
Cas-OFFinder	Up to 3 MM + 3 Gaps	56,042	395	229
CRISPRitz	Up to 3 MM + 3 Gaps	2,381,219	903	396



(A) Number of alignments for CALITAS, Cas-OFFinder and CRISPRitz when making predictions for gRNA RSQ5138 up to 3MMGs. Shown are the initial number of alignments, the number of alignments up to 3 mismatches and gaps (MMGs) and the number of unique sites up to 3 MMGs. (B) Proportional area Venn diagram showing the overlap between the predicted unique sites for CALITAS, CRISPRitz and Cas-OFFinder and for gRNA RSQ5138 up to 3MMGs. (C) Comparison of the predictions of CALITAS, CRISPRitz and Cas-OFFinder for the number of off-target sites up to 3 MMGs for 41 guides, using the PAM NRG for SpCas9

Multiple gaps are present in off-targets



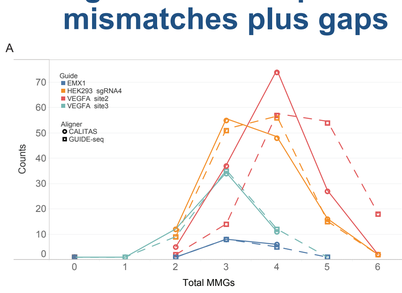
D

Total MM plus Gaps

	2	3	4	5
CALITAS	10	461	8,782	114,561
Digenome Detected	7	126	441	326
Fraction Detected	70.00%	27.33%	5.02%	0.28%

(A) Classification of 987 Digenome-seq sites for gRNA RSQ5138 as a function of the number of MMGs when aligning the gRNA sequence using CALITAS with default parameters (green line) or with parameters set to prevent gaps (red line). (B) and (C), representative alignment with 1 mismatch and 2 gaps obtained using CALITAS (B), or with 10 mismatches using CALITAS with parameters to prevent gaps (C). (D) table showing the total number of *in silico* identified off-target sites by CALITAS for different number of MMGs, the number of Digenome-seq sites, and the fraction of CALITAS-identified sites detected by Digenome-seq

CALITAS aligns GUIDE-seq sites with fewer mismatches plus gaps



B

Total Mismatches plus Gaps

Guide		0	1	2	3	4	5
EMX1	CALITAS	1	0	6	342	8,366	129,574
	Guide-Seq Detected	1	0	3	7	5	0
	Fraction Detected	100.00%		50.00%	2.05%	0.06%	0%
HEK293 sgRNA4	CALITAS	1	0	41	1,942	16,042	166,747
	Guide-Seq Detected	1	0	14	58	50	9
	Fraction Detected	100.00%		34.15%	2.99%	0.31%	0.01%
VEGFA site2	CALITAS	1	0	26	483	11,713	72,396
	Guide-Seq Detected	1	0	6	38	76	27
	Fraction Detected	100.00%		23.08%	7.87%	0.65%	0.04%
VEGFA site3	CALITAS	1	1	89	2,518	45,605	283,577
	Guide-Seq Detected	1	1	13	36	9	0
	Fraction Detected	100.00%	100.00%	14.61%	1.43%	0.02%	0%

(A) Shown is the classification of the GUIDE-seq sites found for four guides as a function of the number of MMGs when aligning each gRNA sequence using CALITAS (solid lines and circles) or as aligned in Tsai et al. without including gaps (dashed lines and squares). (B) table showing for four different gRNAs and different number of MMGs, the number of *in silico* predicted off-target sites by CALITAS, the number of GUIDE-seq detected sites, and the fraction of CALITAS-identified sites detected by GUIDE-seq

Conclusions

CALITAS is a new state-of-the-art aligner useful for *in silico* prediction of CRISPR/Cas on- and off-target sites

CALITAS is freely available and can be downloaded at https://github.com/editasmedicine/calitas

Alignments using experimentally discovered Digenome-Seq off-target sites show the importance of including multiple gaps

Comparison with CRISPRitz and Cas-OFFinder shows that CALITAS' off-target site list is more comprehensive, likely due to better gap handling within the alignments

CALITAS along with biochemical and cellular assays (like Digenome-seq and GUIDE-seq, respectively) provides a streamlined workflow for off-target discovery. Off-target sequencing panels can be quickly made after guide selection using CALITAS, selecting sites with three or fewer mismatches plus gaps. Followed by experimental discovery of off-target sites with higher number of mismatches plus gaps, which are found with much lower relative frequency

See publication for more details: Fennell T, Zhang D, Isik M, Wang T, Gotta G, Wilson CJ, Marco E. CALITAS: A CRISPR-Cas-aware ALigner for *In silico* off-TArget Search. CRISPR J. 2021 Apr;4(2):264-274. doi: 10.1089/crispr.2020.0036

Acknowledgements

We would like to thank Mike Dinsmore for assistance in reviewing the code, and the following Editas teams for supporting this project: Sequencing, Screening, Sample Management, Informatics, and Scientific Communications

Disclosures:
Employee of Fulcrum Genomics: T.F.
Employees and Shareholders of Editas Medicine: D.Z., M.I., T.W., G.G., C.J.W., E.M.