Direct profiling of genome-wide dCas9 and Cas9 specificity using ssDNA mapping (CasKAS)

Georgi K. Marinov[1,#], Samuel H. Kim[2], S. Tansu Bagdatli[1], Alexandro E. Trevino[3,4], Josh Tycko[1], Tong Wu[5], Lacramioara Bintu[4], Michael C. Bassik[1,6], Chuan He[5,7,8], Anshul Kundaje[1,9], and William J. Greenleaf[1,10,11,12,#]

[1] *Department of Genetics, School of Medicine, Stanford University, Stanford, CA 94305, USA*
[2] *Cancer Biology Programs, School of Medicine, Stanford University, Stanford, CA 94305, USA*
[3] *Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA*
[4] *Department of Bioengineering, Stanford University, Stanford, CA 94305, USA*
[5] *Department of Chemistry and Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL, 60637, USA*
[6] *Chemistry, Engineering, and Medicine for Human Health (ChEM-H), Stanford University, Stanford, CA, 94305, USA*
[7] *Department of Biochemistry and Molecular Biology and Institute for Biophysical Dynamics, The University of Chicago, Chicago, IL, 60637, USA*
[8] *Howard Hughes Medical Institute, The University of Chicago, Chicago, IL, 60637, USA*
[9] *Department of Computer Science, Stanford University, Stanford, CA 94305, USA*
[10] *Department of Applied Physics, Stanford University, Stanford, CA 94305, USA*
[11] *Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA, 94305, USA*
[12] *Chan Zuckerberg Biohub, San Francisco, California, USA*
[#] *Corresponding author*

CRISPR-mediated genome and epigenome editing is a transformative technology but its practical application has often been plagued by off-target activities of guide RNAs. Accordingly, numerous methods have been developed to map sgRNA specificity genome-wide, all of which, however, are cumbersome and/or expensive, and most are not applicable to catalytically dead CRISPR enzymes. We have developed a novel quick, inexpensive and straightforward such assay based on directly mapping the single-stranded DNA structures formed upon CRISPR binding ("CasKAS"), the application of which we demonstrate in both *in vitro* and *in vivo* contexts.

CRISPR-based methods for editing the genome and epigenome have emerged as a highly versatile way for manipulating the genetic makeup and regulatory states of cells. Much hope is invested in CRISPR's potential to transform medical practice by enabling direct editing out of pathogenic and engineering in of therapeutic sequence variants. CRISPR has also become a standard tool for discovery in fundamental biomedical research, e.g. in the form of high-throughput massively parallel CRISPR screens[1].

The presence of significant off-target effects for many guide RNAs (sgRNAs), extending beyond the simple cases of direct matches to the sgRNA, presents a major hurdle to fully realizing this potential. Off-targets are particularly problematic for medical applications, where risks of negative consequences for a patient's health need to be minimized as fully as possible.

To address this problem, numerous approaches have been developed to experimentally map off-targets genome-wide. Methods such as Digenome-seq[2] look for particular types of cut sites around target sequences in whole-genome sequencing data; however, deep whole-genome sequencing is still quite expensive to carry out. Assays such as BLESS[3], GUIDE-seq[4], HTGTS[5], DSBCapture[6], BLISS[7], SITE-seq[8], CIRCLE–seq[9], TTISS[10], INDUCE-seq[11], and CHANGE-seq[12] aim to instead directly map Cas9 cleavage events; however, they all involve some combination of complex and laborious molecular biology protocols and non-standard reagents, and have not been widely adopted as a result. Other methods have also been published, such as DISCOVER-seq[13], which maps DNA repair activity by applying ChIP-seq against the MRE11 protein; however, they, as well as earlier applications of ChIP-seq to map catalytically dead dCas9 occupancy sites genome-wide[14,15] suffers from background and specificity issues associated with the ChIP procedure. Most recently, long-read sequencing has been adapted to the problem of Cas9 specificity profiling, in the form of SMRT-OTS and Nano-OTS[16], but the cost of these methods is relatively high while their throughput is comparatively low.

Various computational models have been trained to predict off-targets genome-wide[17,18]. However, these exhibit far from perfect accuracy, while in many situations, especially within clinical contexts, direct experimental evidence is needed to meet as high an epistemic standard as possible.

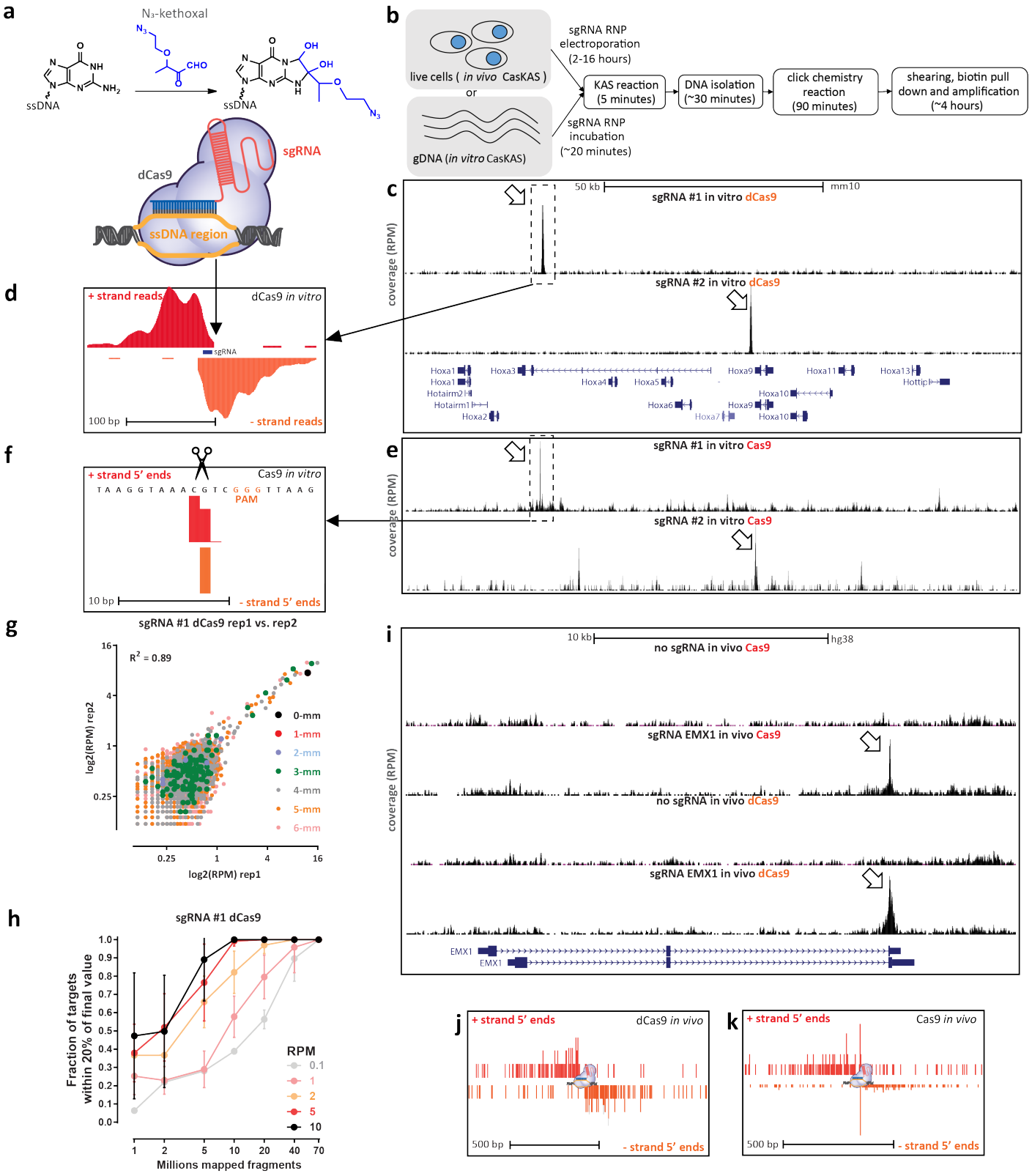


Figure 1: CasKAS maps dCas9- and Cas9-mediated strand invasion and cleavage events genome-wide *in vitro* and *in vivo*. (a) CasKAS is based on the KAS-seq assay for mapping ssDNA structures. N_3-kethoxal covalently modifies unpaired guanine bases (while having no activity for G bases paired within dsDNA). Strand invasion by Cas9/dCas9 carrying an sgRNA results in the formation of a ssDNA structure, which can be directly identified

(legend continued on next page)

Thus at present there is still a need for faster, more accessible and versatile methods for mapping CRISPR off targets. Here we address this problem by adapting the recently developed KAS-seq[19] assay for mapping single-stranded DNA (ssDNA) structures (**k**ethoxal-**a**ssisted **s**sDNA sequencing[19]) to the CRISPR context (Figure 1a-b). KAS-seq is based on the specific labeling of unpaired G bases by N_3-kethoxal, to which biotin can then be added using click chemistry. After shearing, biotinylated DNA can be specifically enriched for and sequenced.

When a Cas9-sgRNA ribonucleoprotein (RNP) is engaged with its target site, the sgRNA invades the DNA double helix, forming a ssDNA structure on the other strand (Figure 1a). We reasoned that KAS-seq should be able to map such sites in a highly specific manner. We carried out an initial *in vitro* experiment using mouse genomic DNA (gDNA), purified dCas9 and two sgRNAs targeting the *Hoxa* locus.

Strikingly, we observed strong and very specific peaks at the expected target sites for each sgRNA (Figure 1c). Detailed examination of dCas9 CasKAS profiles around the predicted sgRNA target sites revealed strand coverage asymmetry patterns similar to those observed for ChIP-seq around transcription factor binding sites[20] (Figure 1d), indicating that enrichment derives from the sgRNA target site itself and confirming the utility of N_3-kethoxal for mapping dCas9 occupancy sites. We term the assay "CasKAS".

We then reasoned that CasKAS should be able to also capture active Cas9 complexed with DNA, as the enzyme is thought to remain associated with DNA for some time after cleaving it REF. We carried out Cas9 CasKAS experiments with the same sgRNAs and again observed enrichment at the expected on-target sites (Figure 1e). Remarkably, examination of Cas9 CasKAS read profiles around the on-target site showed that the 5' ends of reads are tightly phased around the expected cut site, with one cut position on one strand and two to three such positions on the other (Figure 1f), consistent with the previously known patterns of Cas9 cleavage REF. CasKAS therefore provides target specificity profiles for both active and catalytically dead Cas9 versions (we note that, curiously, the genome-wide profiles show some differences between the two versions of the enzyme; Supplementary Figure 1).

In vitro CasKAS data is highly reproducible between replicates (Figure 1g), and between 10 and 20 million mapped reads, i.e. a modest sequencing depth, are generally sufficient to capture off-target specificity profiles (Figure 1h).

We observed similar results with two mouse sgRNAs targeting the *Nanog* locus (Supplementary Figure 2) and with two human sgRNA ("EMX1" and "VEGFA"; Supplementary Figures 3 and 4). We find no enrichment using components of the RNP in isolation – sgRNAs, dCas9 or Cas9 (Supplementary Figure 3).

Next we tested the application of CasKAS *in vivo*. Live cells contain a substantial amount of ssDNA due to ongoing active transcription and other processes[19], so the *in vivo* CasKAS signal is expected to be a mixture of the ssDNA associated with the RNP and endogenous ssDNA. We carried out KAS-seq experiments in HEK293 cells transfected with EMX1 or VEGFA RNPs, as well as negative, no-guide controls, which provide information about endogenous ssDNA profiles. At the EMX1 gene, which happens not to be active in HEK293 cells, we observe strong peaks at the expected site (Figure 1i), an asymmetric read profile around it for dCas9 (Figure 1j), and a substantial degree of read phasing at the cut site, similar to what is observed *in vitro* for active Cas9 (Figure 1g). The VEGFA gene is active in HEK293 cells, but the dCas9/Cas9 CasKAS signal is still readily identifiable as an addition to the endogenous ssDNA enrichment pattern (Supplementary Figure 5). These results demonstrate the utility of CasKAS for profiling CRISPR specificity both *in vitro* and *in vivo*

We then examined the genome-wide specificity of sgRNAs as measured by CasKAS. We focus on the mouse sgRNA #1 for demonstration purposes as it displayed a substantial number of off-targets yet that number was also sufficiently small for all of them to be examined directly. We called peaks *de novo* (see Methods for details) without relying on off-target prediction algorithms, then manually

using N_3-kethoxal. (b) Outline of *in vivo* and *in vitro* CasKAS. For in *in vitro* CasKAS, gDNA is incubated with a dCas9/Cas9 RNP, then N_3-kethoxal is added to the reaction; for in *in vivo* CasKAS, cells are transfected with an RNP, then treated with kethoxal. DNA is then purified, click chemistry is carried out, DNA is sheared, labeled fragments are pulled down with streptavidin beads, and sequenced. (c and d) Mapping of dCas9 targets *in vitro*. (c) Mouse gDNA was incubated with dCas9 RNPs carrying one of two sgRNAs targeting the mouse *HOXA* locus. Highly specific labeling is observed at the expected target location of each sgRNA. (d) Asymmetric strand distribution of *in vitro* dCas9 CasKAS reads around the sgRNA target site. (e and f) Mapping of Cas9 targets *in vitro*. (e) Mouse gDNA was incubated with Cas9 RNPs carrying one of same two sgRNAs targeting the mouse *HOXA* locus. (f) The distribution of 5' read ends around targets sites in *in vitro* CasKAS datasets shows direct capture of the intermediate cleavage state. (g) Reproducibility of in vivo dCas9 CasKAS datasets. Shown are RPM values for 500bp windows centered on the top ~7,000 predicted target sites for the "sgRNA #1" in two *in vitro* CasKAS experiments. Off-target sites are color-coded by the number of mismatches relative to the sgRNA. (h) CasKAS requires a moderate sequencing depth of 10-20 $\times$ 10^6 reads to accurately rank potential off-targets. (i-k) *In vitro* CasKAS maps Cas9 and dCas9 target sites. (i) Shown are CasKAS experiments with Cas9 and dCas9 and with the EMX1 sgRNA or with no sgRNA (negative control) (j) Assymmetric 5' end distribution around target sites in dCas9 *in vivo* CasKAS. (k) In *in vivo* Cas9 CasKAS, a mixture distribution is observed between phased cleavage sites and broader ssDNA labeling.

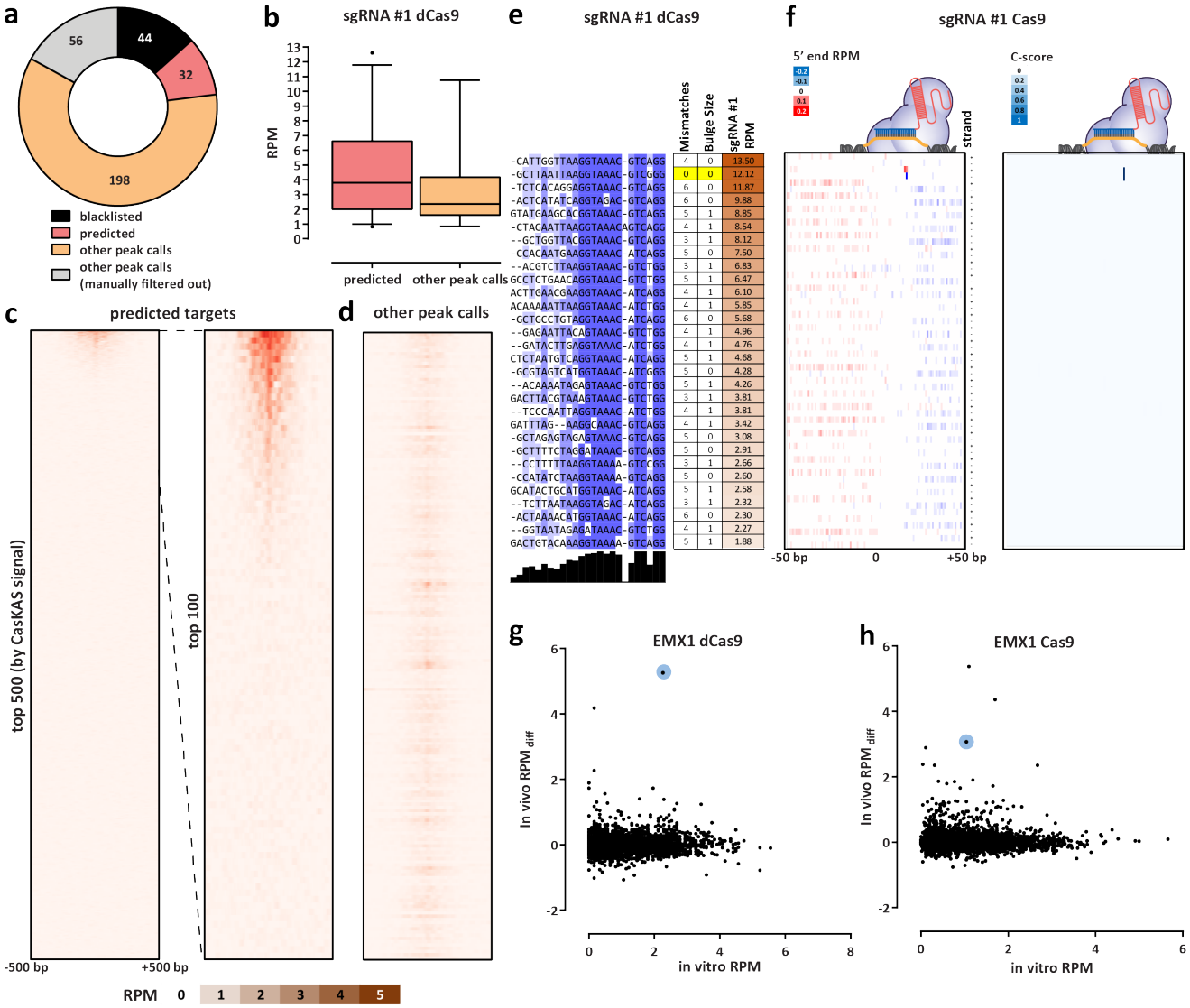


Figure 2: CasKAS profiles sgRNA specificity genome-wide. (a) Summary of de novo peak calls for sgRNA #1 (using MACS2) (b) CasKAS signal is stronger over predicted off-target sites, but legitimate interactions are also found elsewhere in the genome. (c) CasKAS profile over predicted (by Cas-OFFinder) off-target sites for sgRNA #1 with dCas9 (all such sites and focusing only on the top 100 ranked by dCas9 CasKAS signal). (d) CasKAS profile over peak calls outside predicted (by Cas-OFFinder) off-target sites for sgRNA #1 with dCas9. (e) Determinants of sequence specificity as measured by dCas9 CasKAS (for sgRNA #1). PAM-distal regions of the sgRNA are less constrained than its PAM-proximal parts. The on-target sgRNA is highlighted in yellow. (f) Active Cas9 signal read profiles can be used to distinguish off-targets associated with cutting from those where only binding occurs. Shown are the same off-target sites as in (e) and the plus- and minus-strand active Cas9 5' end profiles around the sgRNA. In this case (sgRNA #1), only the on-target site shows a Cas9 CasKAS pattern indicating cleavage; at the other sites even active Cas9 likely only binds but does not cut. A simple cutting score metric ("*C*-score") based on multiplying the 5' end forward- and reverse-strand profiles can be used to quantify cutting vs. binding. (g and h) Comparison between *in vitro* and *in vivo* CasKAS signal over predicted off-target sites for the EMX1 sgRNA. *In vivo* CasKAS is quantified as the difference in read per million (±500 bp of the sgRNA site) between the sgRNA KAS-seq and the no-guide control KAS-seq ("RPM_{diff}). The on-target site is shown in blue.

curated the set of peaks (Figure 2a). Remarkably, while we find 32 peaks at predicted off-target sites, we also find 192 (i.e. ∼6× as many) additional manually curated peaks; while these peaks exhibit generally lower CasKAS signal (Figure 2b), they all appear to be genuine sites of occupancy as they display proper peak shape characteristics (see Sup-

plementary Figure 14 for details). Most of the predicted (in total ∼7,500) off-target sites for this sgRNA do not show substantial occupancy by dCas9 CasKAS (Figure 2c-d).

Sequence comparison of the occupied predicted off-target sites allowed us to evaluate determinants of Cas9 specificity (Figure 2e). Consistent with previous reports REF, the PAM-distal region is under much less constraint than the PAM-proximal one. We observed a similar picture with the other sgRNAs we profiled, in both mouse and human (Supplementary Figures 6-9 and Supplementary Figures 10-13).

Interesting patterns were observed when carrying out a similar analysis of the peaks not associated with predicted off-target sites (Supplementary Figure 15) – at numerous sites, robustly occupied by dCas9, the number of mismatches and the size of "bulge" regions is much larger than what is considered permissible by off-target prediction algorithms; this likely explains the much larger number of such sites relative to the set of occupied predicted off-targets.

A remarkable observation emerged from the comparison of dCas9 and Cas9 results. We devised a simple metric for evaluating the degree of read phasing at cut sites (a "*C*-score"; see Methods for details), and used it to estimate the degree of cutting by Cas9. Strikingly, while the on-target site exhibits the second highest dCas9 CasKAS signal, and even though all off-target sites show binding by CasKAS, only the on-target site displays cutting activity (Figure 2f). The behavior of other sgRNAs varies (Supplementary Figures 6-9 and 16), with some showing multiple cut sites. Thus the combination of dCas9 and Cas9 CasKAS (or even Cas9 CasKAS alone) is a powerful tool for detecting not only the binding but also the cutting specificity of sgRNAs

Finally, we compared *in vitro* and *in vivo* CasKAS profiles (Figure 2g-h). We find many fewer strongly enriched sites in *in vivo* datasets than *in vitro*, with the on-target site being either the top (for dCas9) or among the top (for Cas9) sites in vivo.

In conclusion, we have presented CasKAS, a new, simple and robust method for mapping the specificity of active and catalytically dead versions of CRISPR enzymes. CasKAS has numerous advantages over existing tools while also opening up new possibilities for studying CRISPR biology. CasKAS is very simple as it requires no specialized molecular biology protocols and takes just a few hours *in vitro* (and a similar amount of time after harvesting cells *in vivo*), and it also cheap as it actively and strongly enriches for off-targets. It measures strand invasion by CRISPR rather than association with DNA, a biochemically more specific event. We compared *de novo* called CasKAS peaks to those generated by other means, and while we found large sets of peaks unique to each method, those found only by CasKAS contained much higher fractions of predicted off-target sites than those unique to other methods (Supplementary Figure 17). CasKAS can be used to profile the specificity of all types of DNA-targeting CRISPR proteins as it does not rely on measuring DNA cleavage or modification. It can be applied in primary cells as what is measured is physical association with DNA and not the outcome of CRISPR activity that may only be detectable after cell division. A limitation of CasKAS is the requirement that a G nucleotide is present within the sgRNA sequence, as without it there would be no kethoxal labeling; however, only a small fraction (≤5%) of sgRNAs in the human genome lack any Gs (Supplementary Figure 18), with the median number being four to five, thus most guides are not significantly affected. Another minor limitation of the current *in vitro* protocol is that labeling is carried out on high molecular weight (HMW) DNA and samples are sheared one by one, which limits throughput. We have explored using presheared and end-repaired DNA (to minimize kethoxal labeling of Gs on sticky ends generated by sonication), with comparable results to using HMW DNA; further optimization should allow the parallel high-throughput plate-based profiling of the specificity of very large numbers of sgRNAs.

In addition to being highly valuable for off-target profiling *in vitro* and in previously difficult to assay settings such as primary cells, we expect CasKAS to provide fruitful insights into the mechanisms and dynamics of *in vivo* CRISPR action (taking advantage of finely controllable CRISPR systems such as vfCRISPR [21]), and the influence of transcriptional, regulatory, and epigenetic and other functional genomic contexts on CRISPR activity.

References

1. Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**(6166):80–84.
2. Kim D, Bae S, Park J, Kim E, Kim S, Yu HR, Hwang J, Kim JI, Kim JS. 2015. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods* **12**(3):237–243
3. Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, Karaca E, Chiarle R, Skrzypczak M, Ginalski K, Pasero P, Rowicka M, Dikic I. 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods* **10**(4):361–365.
4. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, Wyvekens N, Khayter C, Iafrate AJ, Le LP, Aryee MJ, Joung JK. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**(2):187–197.
5. Frock RL, Hu J, Meyers RM, Ho YJ, Kii E, Alt FW. 2015. Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* **33**(2):179–186.
6. Lensing SV, Marsico G, Hänsel-Hertsch R, Lam EY, Tannahill D, Balasubramanian S. 2016. DSBCapture:

in situ capture and sequencing of DNA breaks. *Nat Methods* **13**(10):855–857.

7. Yan WX, Mirzazadeh R, Garnerone S, Scott D, Schneider MW, Kallas T, Custodio J, Wernersson E, Li Y, Gao L, Federova Y, Zetsche B, Zhang F, Bienko M, Crosetto N. 2017. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat Commun* **8**:15058.
8. Cameron P, Fuller CK, Donohoue PD, Jones BN, Thompson MS, Carter MM, Gradia S, Vidal B, Garner E, Slorach EM, Lau E, Banh LM, Lied AM, Edwards LS, Settle AH, Capurso D, Llaca V, Deschamps S, Cigan M, Young JK, May AP. 2017. Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat Methods* **14**(6):600–606.
9. Tsai SQ, Nguyen NT, Malagon-Lopez J, Topkar VV, Aryee MJ, Joung JK. 2017. CIRCLE–seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat Methods* **14**(6):607–614.
10. Schmid-Burgk JL, Gao L, Li D, Gardner Z, Strecker J, Lash B, Zhang F. 2020. Highly Parallel Profiling of Cas9 Variant Specificity. *Mol Cell* **78**(4):794–800.e8.
11. Dobbs FM, van Eijk P, Fellows MD, Loiacono L, Nitsch R, Reed SH. 2020. Precision digital mapping of endogenous and induced genomic DNA breaks by INDUCE-seq. *bioRxiv* 2020.08.25.266239
12. Lazzarotto CR, Malinin NL, Li Y, Zhang R, Yang Y, Lee G, Cowley E, He Y, Lan X, Jividen K, Katta V, Kolmakova NG, Petersen CT, Qi Q, Strelcov E, Maragh S, Krenciute G, Ma J, Cheng Y, Tsai SQ. 2020. CHANGE-seq reveals genetic and epigenetic effects on CRISPR-Cas9 genome-wide activity. *Nat Biotechnol* **38**(11):1317–1327.
13. Wienert B, Wyman SK, Richardson CD, Yeh CD, Akcakaya P, Porritt MJ, Morlock M, Vu JT, Kazane KR, Watry HL, Judge LM, Conklin BR, Maresca M, Corn JE. 2019. Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. *Science* **364**(6437):286–289.
14. Wu X, Scott DA, Kriz AJ, Chiu AC, Hsu PD, Dadon DB, Cheng AW, Trevino AE, Konermann S, Chen S, Jaenisch R, Zhang F, Sharp PA. 2014. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol* **32**(7):670–676.
15. Kuscu C, Arslan S, Singh R, Thorpe J, Adli M. 2014. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol* **32**(7):677–683.
16. Höijer I, Johansson J, Gudmundsson S, Chin CS, Bunikis I, Häggqvist S, Emmanouilidou A, Wilbe M, den Hoed M, Bondeson ML, Feuk L, Gyllensten U, Ameur A. 2020. Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. *Genome Biol* **21**(1):290.
17. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z, Wilen C, Orchard R, Virgin HW, Listgarten J, Root DE. 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**(2):184–191.
18. Perez AR, Pritykin Y, Vidigal JA, Chhangawala S, Zamparo L, Leslie CS, Ventura A. 2017. GuideScan software for improved single and paired CRISPR guide RNA design. *Nat Biotechnol* **35**(4):347–349.
19. Wu T, Lyu R, You Q, He C. 2020. Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity *in situ*. *Nat Methods* **17**(5):515–523.
20. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9):1813–1831.
21. Liu Y, Zou RS, He S, Nihongaki Y, Li X, Razavi S, Wu B, Ha T. 2020. Very fast CRISPR on demand. *Science* **368**(6496):1265–1269.
22. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**(3):R25.
23. Feng J, Liu T, Qin B, Zhang Y, Liu XS. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc* **7**(9):1728–1740.
24. Amemiya HM, Kundaje A, Boyle AP. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**(1):9354.
25. Bae S, Park J, Kim JS. 2014. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**(10):1473–1475.
26. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5):1792–1797.
27. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. 2009. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**(9):1189–1191.

Methods

Guide RNA sequences

Guide RNAs were obtained from IDT ("sgRNA #1" and "sgRNA #2") or from Synthego (all others).

The following sgRNA sequences were used in this study:

1. "sgRNA #1": XXX
2. "sgRNA #2": XXX
3. "EMX1_Tsai": `GAGTCCGAGCAGAAGAAGAA`
4. "VEGFA-site1": `GGGTGGGGGGAGTTTGCTCC`
5. "Nanog-sg2": `GATCTCTAGTGGGAAGTTTC`
6. "Nanog-sg3": `GTCTGTAGAAAGAATGGAAG`

Guide RNAs were dissolved to a concentration of 100 μM using nuclease-free 1× TE buffer and stored at –20 °C.

In vitro CasKAS

In vitro CasKAS experiments were executed as follows.

First, 1 μL of each synthetic sgRNA were incubated at room temperature with 1 μL of recombinant purified dCas9 (MCLab dCAS9B-200) for 20 minutes. The RNP was then incubated with 1 μg of gDNA at 37 °C for 10 minutes.

The KAS reaction was then carried out by adding 1 μL of 500 mM N_3-kethoxal (ApeXBio A8793). DNA was immediately purified using the MinElute PCR Purification Kit (Qiagen 28006), and eluted in 87.5 or 175 μL 25mM K_3BO_3.

In vivo CasKAS

For *in vivo* CasKAS experiments, HEK293T cells were seeded at 400,000 cells/well into a 6-well plate the day before RNP transfection. Media was exchanged 2 hours before transfection. For each well, 6,250 ng of Cas9 (MCLAB CAS9-200) or dCas9 (MCLAB dCAS9B-200) and 1,200 ng sgRNA was complexed with CRISPRMAX (XX FROM?? XX) reagent in Opti-MEM (XX FROM?? XX) following manufacturer's protocol. After incubation at room temperature for 15 minutes, the RNP solution was directly added to each well and gently mixed. The cells were incubated with the RNP complex for 14 hours at 37 °C. To harvest and perform kethoxal labeling, media was removed and room temperature 1× PBS was used to wash the cells. Cells were then dissociated with trypsin XXX TRYPSIN SOLUTION??XXX, trypsin was quenched with media, cells were pelleted at room temperature, and then resuspended in 100 μL of media supplemented with 5 M N_3-kethoxal. Cells were incubated for 10 minutes at 37 °C with shaking at 500 rpm in a Thermomixer. Cells were then pelleted by centrifuging at 500 *g* for 5 minutes at 4 °C. Genomic DNA was then extracted using the Monarch gDNA Purification Kit (NEB T3010S) following the standard protocol but with elution using 85 μL 25 mM K_3BO_3 at pH 7.0.

Click reaction, biotin pull down and library generation

The click reaction was carried out by combining 175 μL purified and sheared DNA, 5 μL 20 mM DBCO-PEG4-biotin (DMSO solution, Sigma 760749), and 20 μL 10× PBS in a final volume of 200 μL or 87.5 μL purified and sheared DNA, 2.5 μL 20 mM DBCO-PEG4-biotin (DMSO solution, Sigma 760749), and 10 μL 10× PBS in a final volume of 100 μL. The reaction was incubated at 37 °C for 90 minutes.

DNA was purified using AMPure XP beads (50 μL for a 100 μL reaction or 100 μL for a 200 μL reaction), beads were washed on a magnetic stand twice with 80% EtOH, and eluted in 130 μL 25mM K_3BO_3.

Purified DNA was then sheared on a Covaris E220 instrument down to ∼150-400 bp size.

For streptavidin pulldown of biotin-labeled DNA, 10 μL of 10 mg/mL Dynabeads MyOne Streptavidin T1 beads (Life Technologies, 65602) were separated on a magnetic stand, then washed with 300 μL of 1× TWB (Tween Washing Buffer; 5 mM Tris-HCl pH 7.5; 0.5 mM EDTA; 1 M NaCl; 0.05% Tween 20). The beads were resuspended in 300 μL of 2× Binding Buffer (10 mM Tris-HCl pH 7.5, 1 mM EDTA; 2 M NaCl), the sonicated DNA was added (diluted to a final volume of 300 μL if necessary), and the beads were incubated for ≥15 minutes at room temperature on a rotator. After separation on a magnetic stand, the beads were washed with 300 μL of 1× TWB, and heated at 55 °C in a Thermomixer with shaking for 2 minutes. After removal of the supernatant on a magnetic stand, the TWB wash and 55 °C incubation were repeated.

Final libraries were prepared on beads using the NEBNext Ultra II DNA Library Prep Kit (NEB, #E7645) as follows. End repair was carried out by resuspending beads in 50 μL 1× EB buffer, and adding 3 μL NEB Ultra End Repair Enzyme and 7 μL NEB Ultra End Repair Enzyme, followed by incubation at 20 °C for 30 minutes (in a Thermomixer, with shaking at 1,000 rpm) and then at 65 °C for 30 minutes.

Adapters were ligated to DNA fragments by adding 30 μL Blunt Ligation mix, 1 μL Ligation Enhancer and 2.5 μL NEB Adapter, incubating at 20 °C for 20 minutes, adding 3 μL USER enzyme, and incubating at 37 °C for 15 minutes (in a Thermomixer, with shaking at 1,000 rpm) .

Beads were then separated on a magnetic stand, and washed with 300 μL TWB for 2 minutes at 55 °C, 1000 rpm in a Thermomixer. After separation on a magnetic stand, beads were washed in 100 μL 0.1 × TE buffer, then resuspended in 15 μL 0.1 × TE buffer, and heated at 98 °C for 10 minutes.

For PCR, 5 μL of each of the i5 and i7 NEB Next sequencing adapters were added together with 25 μL 2× NEB Ultra PCR Mater Mix. PCR was carried out with a 98 °C incubation for 30 seconds and 12 cycles of 98 °C for 10 seconds, 65 °C for 30 seconds, and 72 °C for 1 minute, followed by incubation at 72 °C for 5 minutes.

Beads were separated on a magnetic stand, and the supernatant was cleaned up using 1.8× AMPure XP beads.

Libraries were sequenced in a paired-end format on a Illumina NextSeq instrument using NextSeq 500/550 high output kits (2×36 cycles).

Data processing

Demultipexed fastq files were mapped to the `hg38` assembly of the human genome or the `mm10` version of the mouse genome as 2×36mers using Bowtie[22] with the following settings: `-v 2 -k 2 -m 1 --best --strata -X 1000`. Duplicate reads were removed using `picard-tools` (version 1.99).

Browser tracks generation, fragment length estimation, TSS enrichment calculations, and other analyses were carried out using custom-written Python scripts (`https://github.com/georgimarinov/GeorgiScripts`). The `refSeq` set of annotations were used for evaluation of enrichment around TSSs.

Peak calling

Peak calling on *in vitro* binding datasets was carried out using version 2.1.0 of MACS2[23] with default settings.

Peaks were then compared against the ENCODE set of "blacklisted" regions[24] to filter out likely artifacts.

Sequence analysis

Guide RNA off-target predictions were obtained from Cas-OFFinder[25]

Multiple sequence alignments of sgRNA sequences and their off-targets were generated using MUSCLE[26] and visualized using JalView[27].

Quantification

Cutting score calculation

The Cas9 cutting C-score was calculated as follows.

First, basepair-level Read-Per-Million (RPM) profiles for mapped read 5' ends were generated separately for the forward and reverse strands. Then the C-score was calculated by multiply the forward and reverse strand profiles (summed over a running window of 3 bp):

$$C-\text{score}_{c,i} = \sum_{j=i-1}^{j=i+1} RPM^{+}_{c,j} \times \sum_{j=i-1}^{j=i+1} RPM^{-}_{c,j} \qquad (1)$$

Where c, i indicate the coordinates by chromosome and position.

Data availability

Sequencing reads for the datasets described in this study are available from GEO accessiong XXX SUBMIT XXX.

Author contributions

G.K.M. conceptualized the study, performed initial *in vitro* CasKAS experiments, analyzed data, and wrote the manuscript with input from all authors. S.H.K. developed the *in vivo* CasKAS protocol and performed *in vivo* CasKAS experiments. S.T.B. carried out *in vitro* CasKAS optimization. A.E.T. and J.T. supplied sgRNAs and designed off-target profiling experiments. A.E.T. carried out off-target analysis for mouse sgRNAs. T.W. provided key reagents. W.J.G., A.K., C.H. M.C.B. and L.B. supervised the study.

Acknowledgments

This work was supported by NIH grants (P50HG007735, RO1 HG008140, U19AI057266 and UM1HG009442 to W.J.G., 1UM1HG009436 to W.J.G. and A.K., 1DP2OD022870-01 and 1U01HG009431 to A.K.), the Rita Allen Foundation (to W.J.G.), the Baxter Foundation Faculty Scholar Grant, and the Human Frontiers Science Program grant RGY006S (to W.J.G). W.J.G is a Chan Zuckerberg Biohub investigator and acknowledges grants 2017-174468 and 2018-182817 from the Chan Zuckerberg Initiative. Fellowship support provided by the Stanford School of Medicine Dean's Fellowship (G.K.M.), the Siebel Scholars, the Enhancing Diversity in Graduate Education Program and the Weiland Family Fellowship (A.E.T.).

XXX MORE XXX

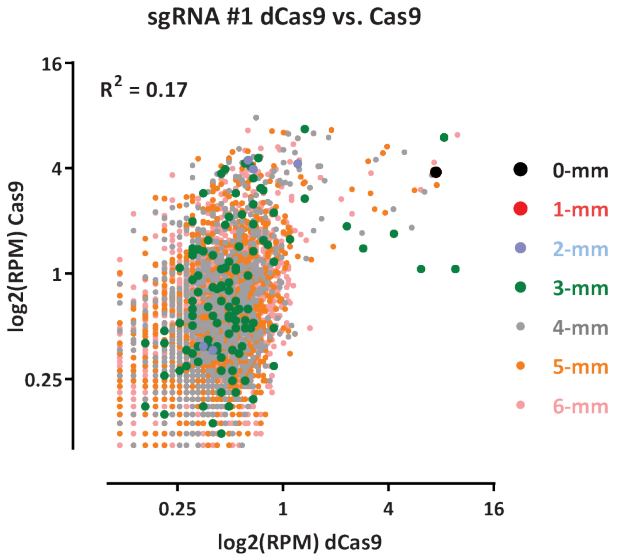
The authors would like to thank Zohar Shipony and members of the Greenleaf, Kundaje, and Bassik labs for helpful discussion and suggestions regarding this work.

Competing interests

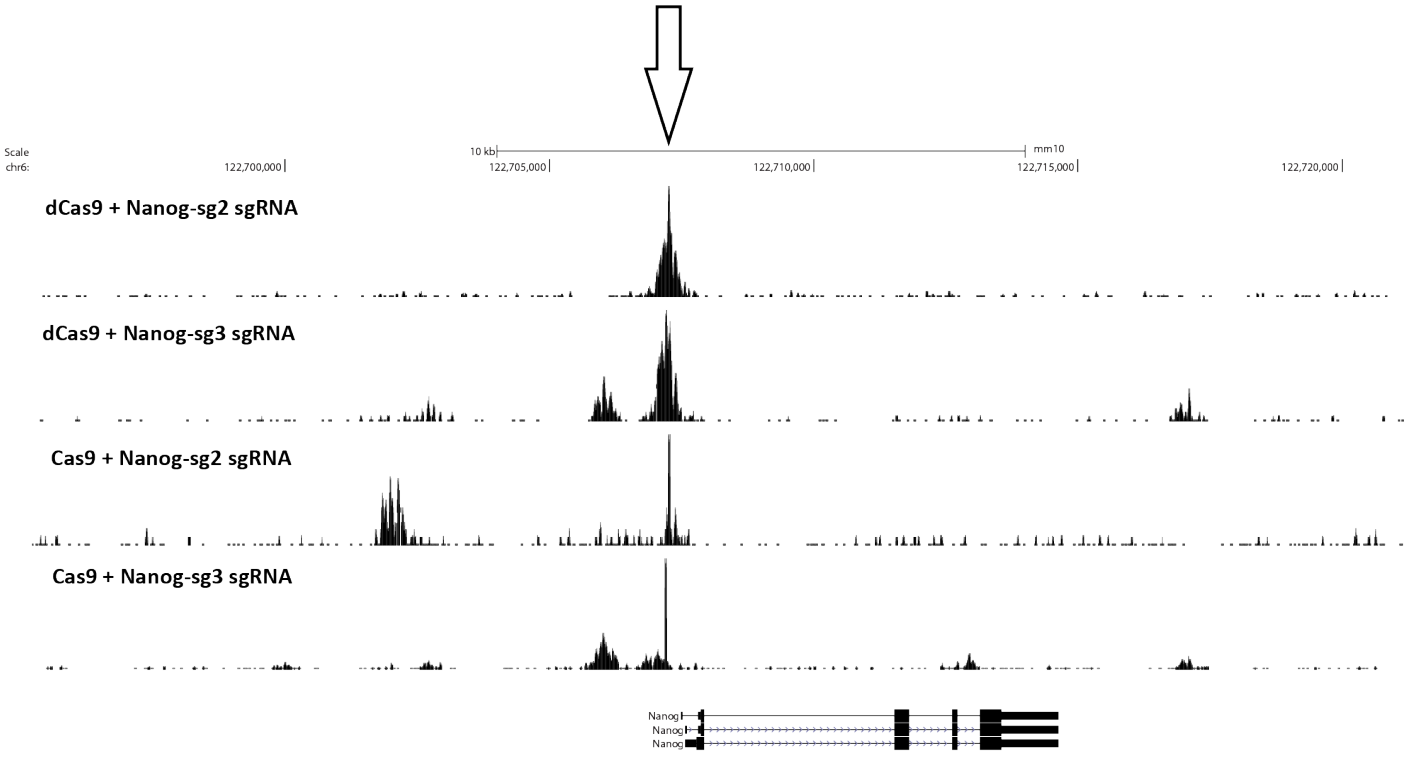
The authors declare no competing interests.

Supplementary Materials

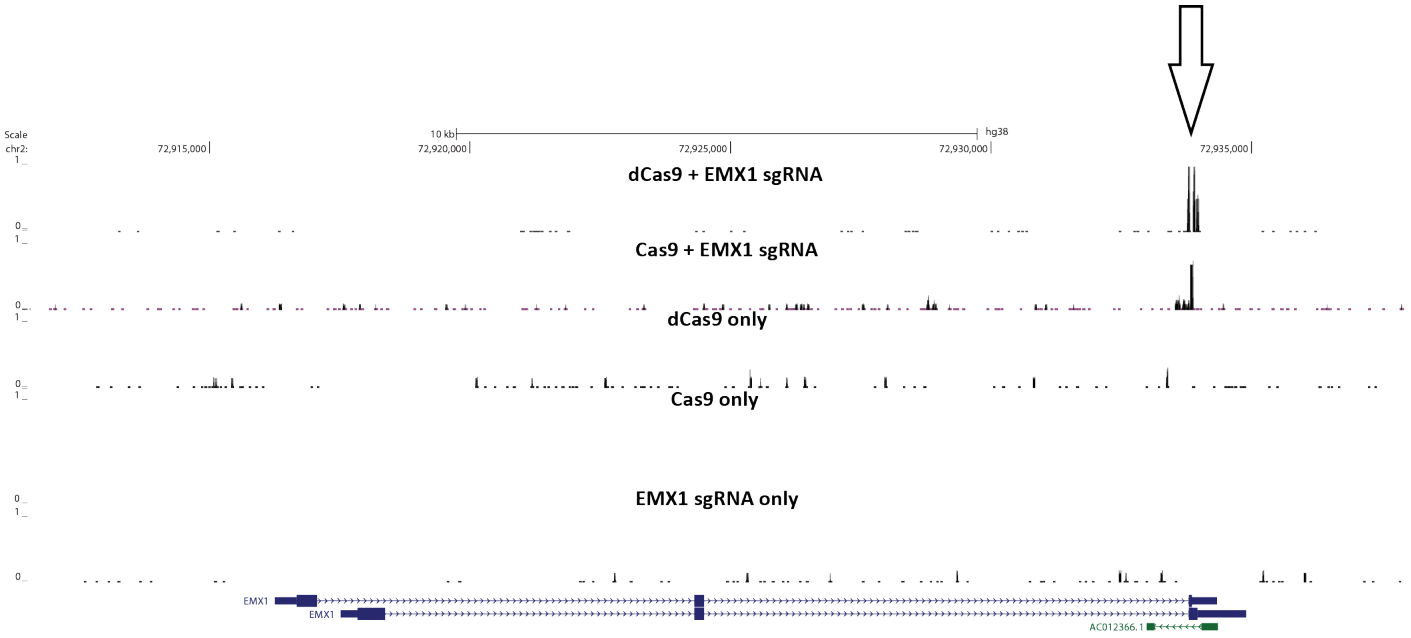
Supplementary Figures



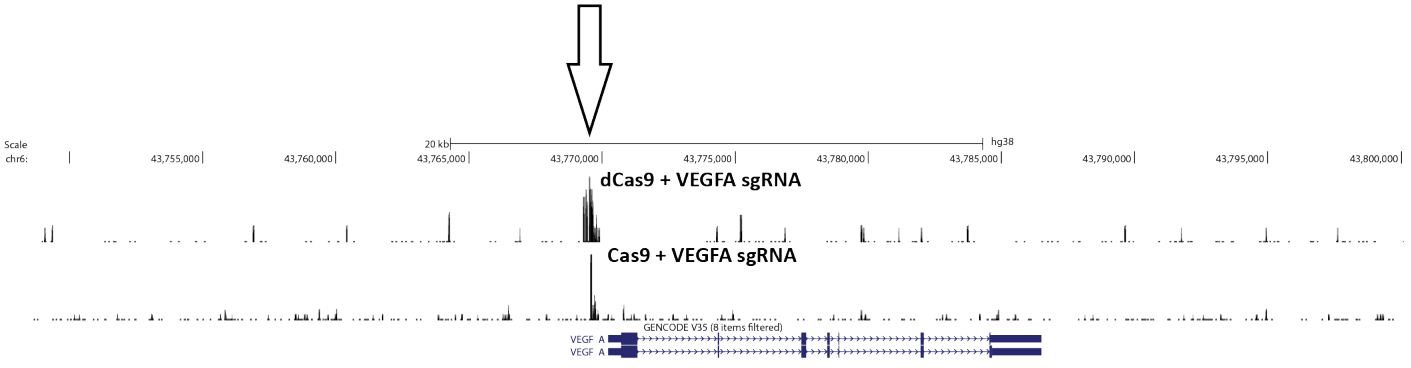
Supplementary Figure 1: Correspondence between *in vitro* dCas9 and active Cas9 CasKAS profiles for the mouse sgRNA #1 guide.



Supplementary Figure 2: *In vitro* dCas9 and Cas9 CasKAS profiles around the mouse *Nanog* locus using the "Nanog-sg2" and "Nanog-sg3" sgRNAs.

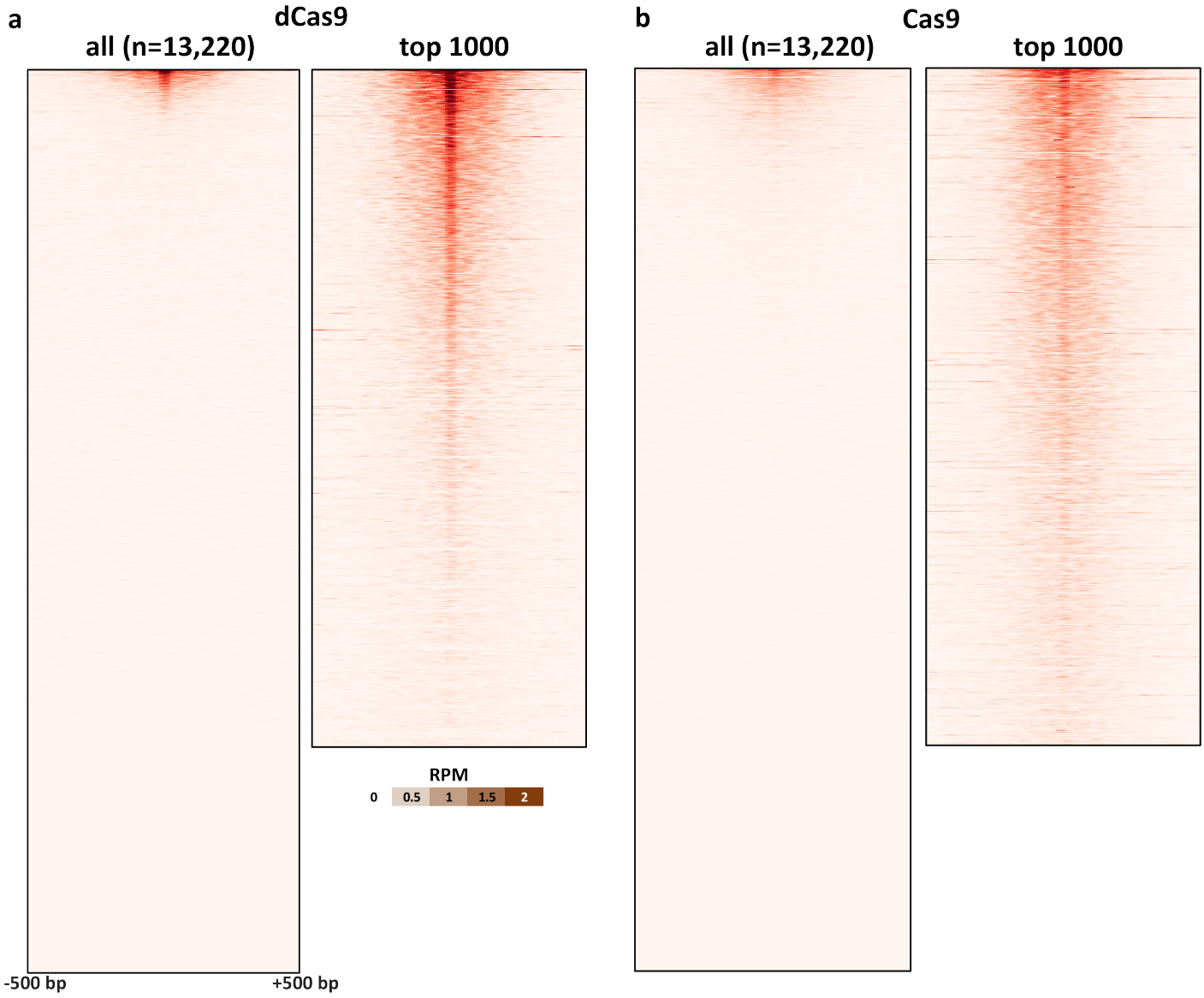


Supplementary Figure 3: CasKAS signal *in vitro* is specific to the activity of the dCas9/Cas9 protein combined with its sgRNA. CasKAS was carried out with the EMX1 sgRNA and with the following combinations of protein and sgRNA: dCas9 + sgRNA, Cas9 + sgRNA, dCas9 alone, Cas9 alone, or sgRNA alone.

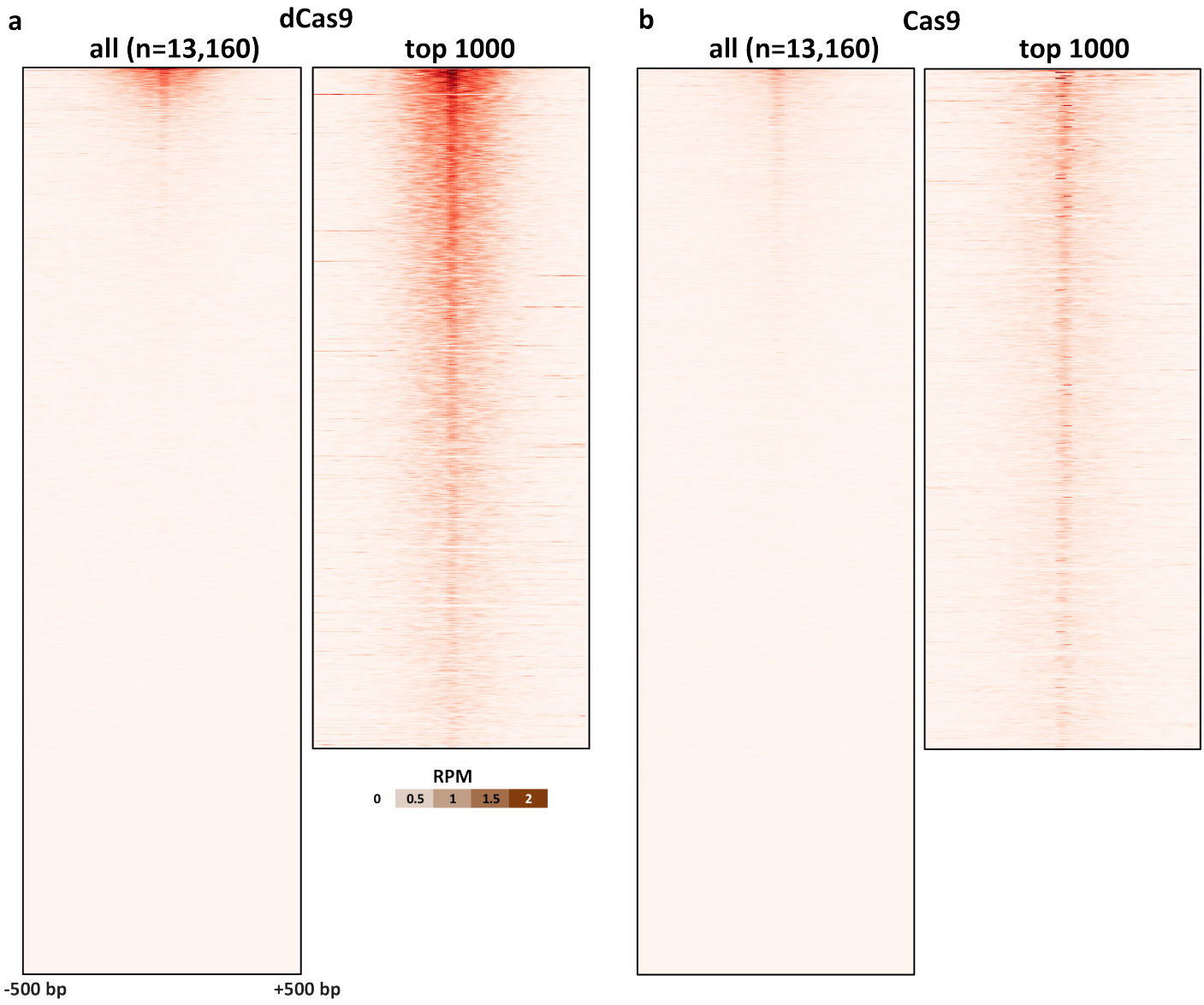


Supplementary Figure 4: CasKAS signal *in vitro* around the *VEGFA* gene with the VEGFA sgRNA.

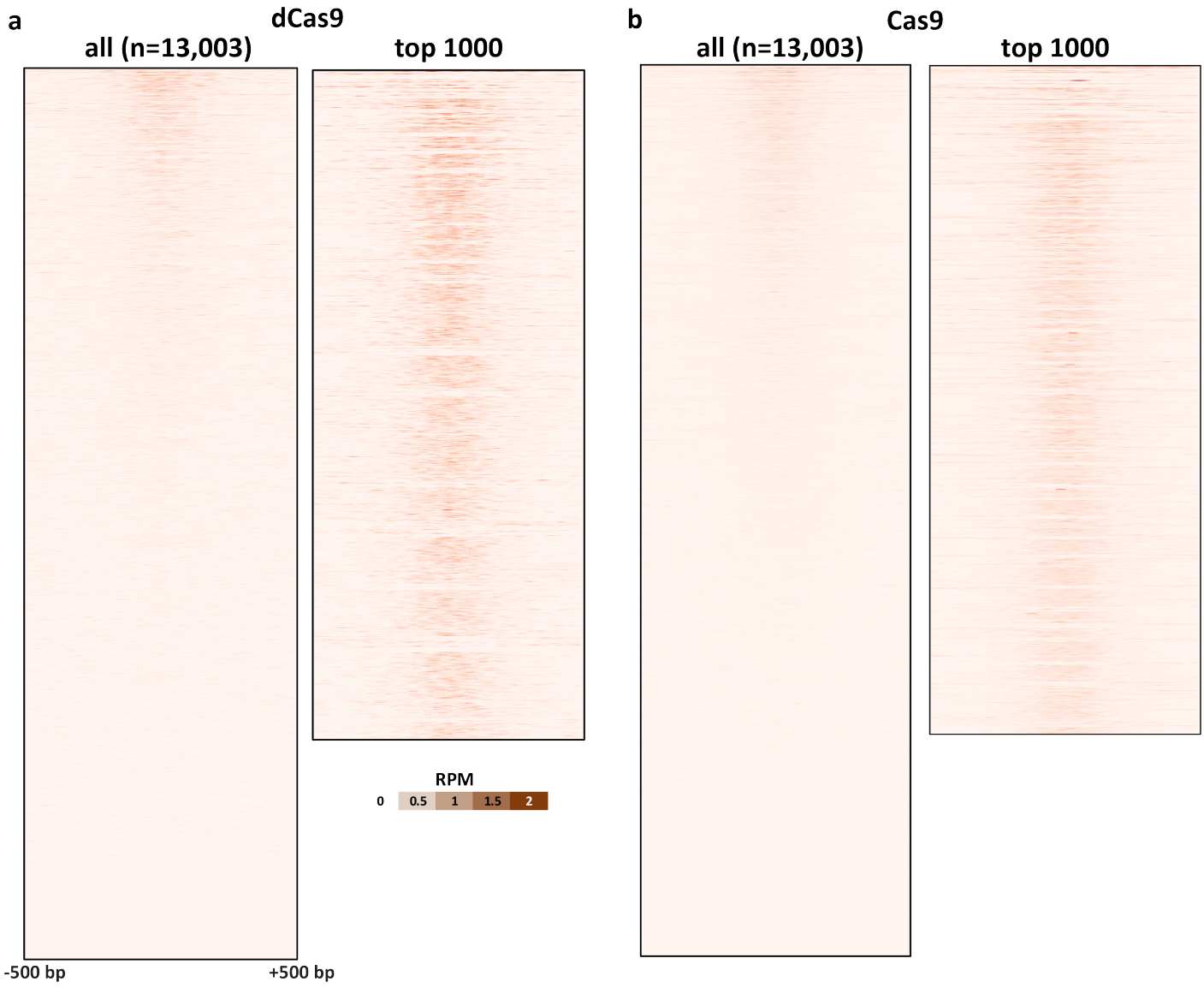
Supplementary Figure 5: CasKAS signal *in vivo* around the *VEGFA* gene with the VEGFA sgRNA.



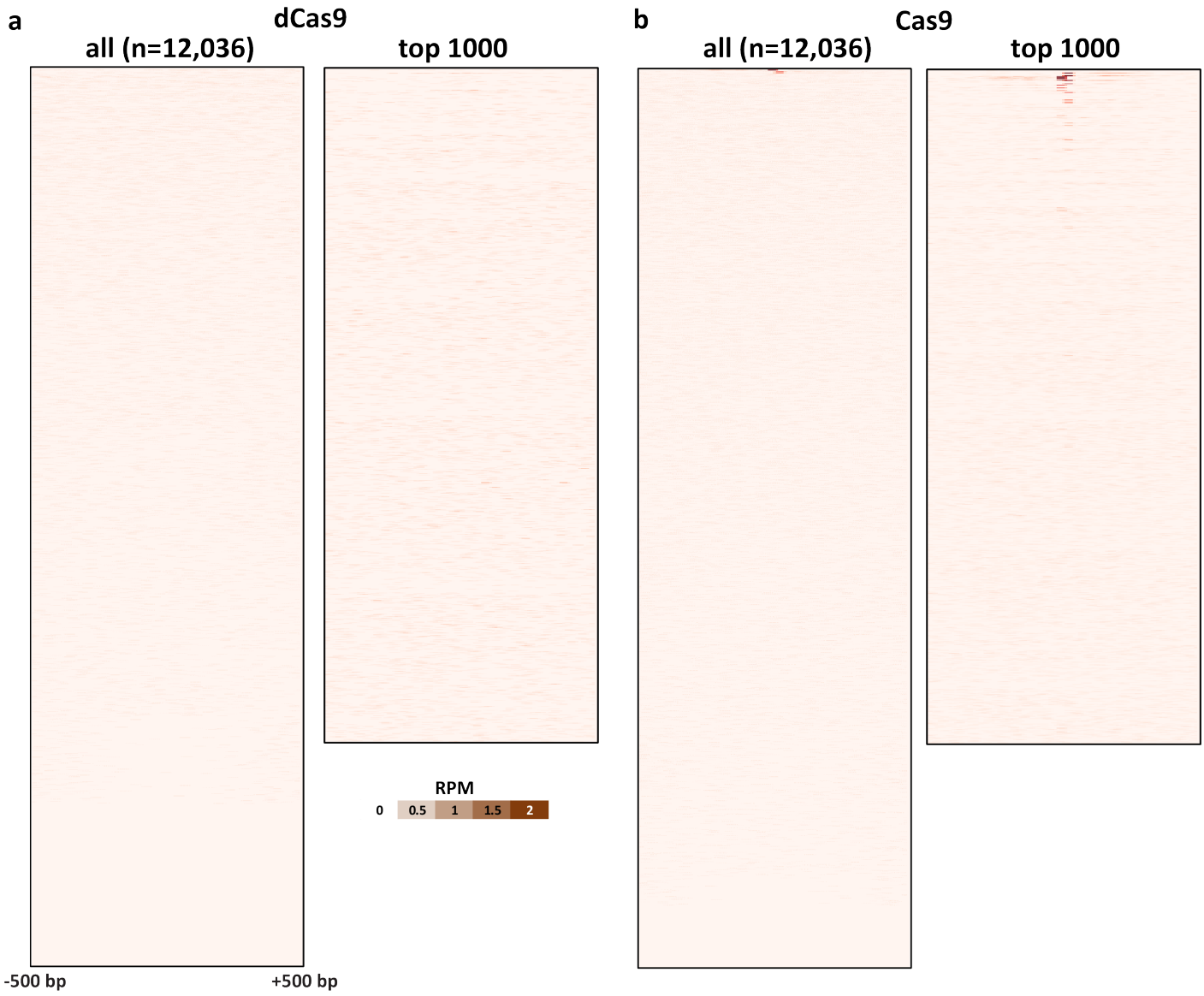
Supplementary Figure 6: *In vitro* dCas9 and Cas9 CasKAS profiles for the "Nanog-sg2" sgRNA. CasKAS profiles are shown for all off-target sites predicted by Cas-OFFinder as well as for the top 1000 sites (ranked by CasKAS RPM values over the ±500bp region around the sgRNA target site).



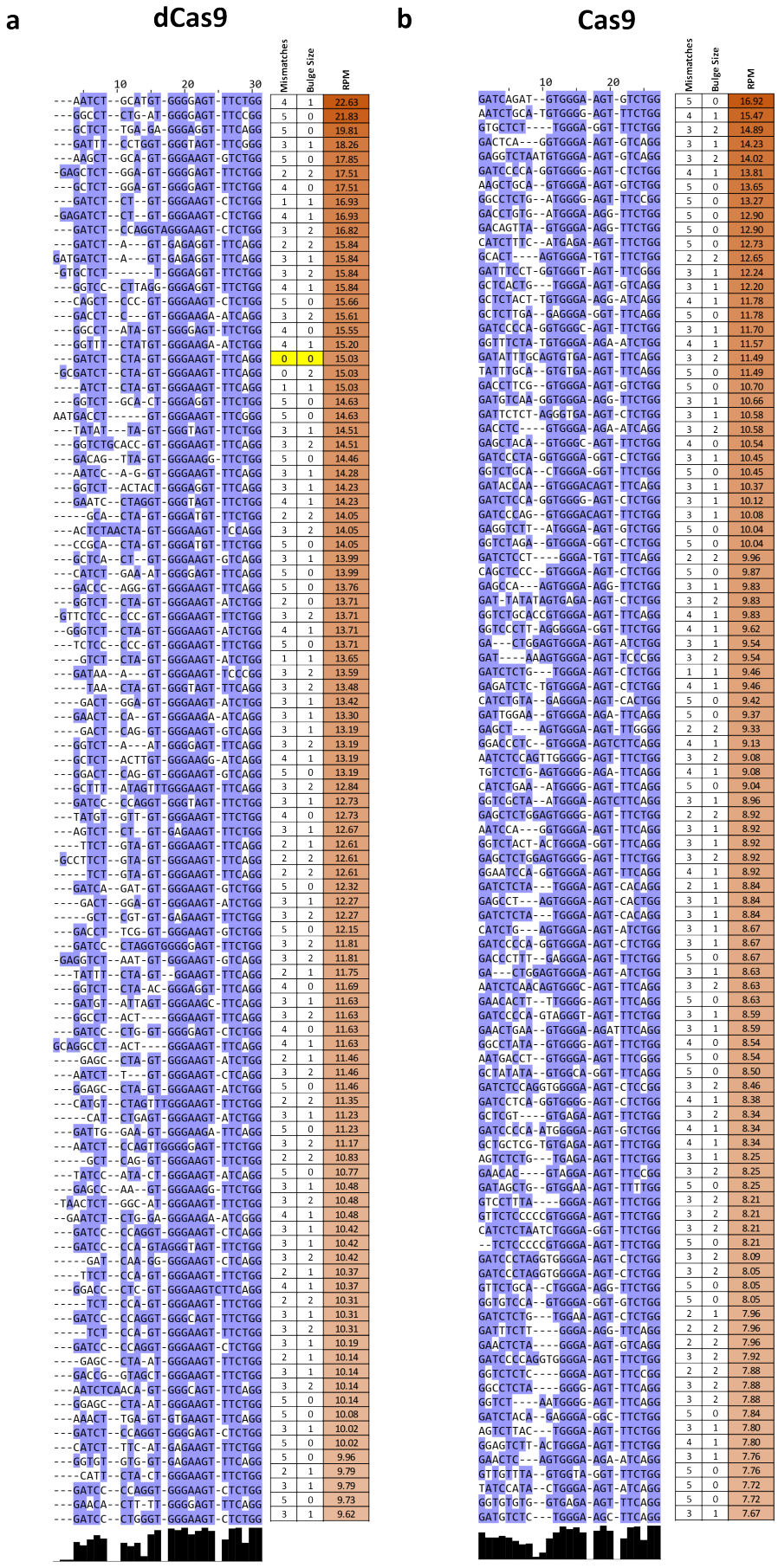
Supplementary Figure 7: ***In vitro*** **dCas9 and Cas9 CasKAS profiles for the "Nanog-sg3" sgRNA**. CasKAS profiles are shown for all off-target sites predicted by Cas-OFFinder as well as for the top 1000 sites (ranked by CasKAS RPM values over the ±500bp region around the sgRNA target site).



Supplementary Figure 8: ***In vitro*** **dCas9 and Cas9 CasKAS profiles for the "EMX1_Tsai" sgRNA**. CasKAS profiles are shown for all off-target sites predicted by Cas-OFFinder as well as for the top 1000 sites (ranked by CasKAS RPM values over the ±500bp region around the sgRNA target site).

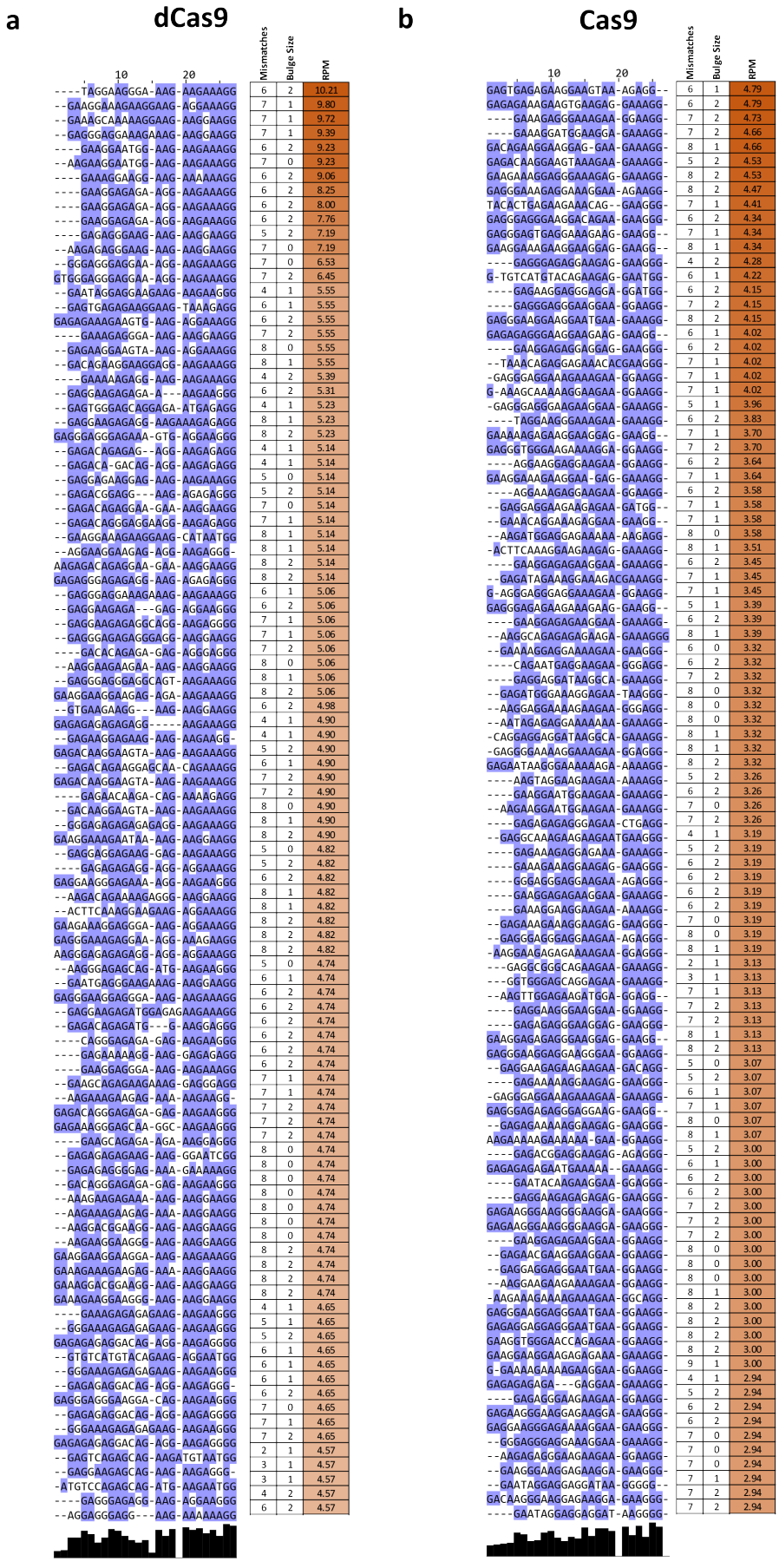


Supplementary Figure 9: ***In vitro*** **dCas9 and Cas9 CasKAS profiles for the "VEGFA-site1" sgRNA**. CasKAS profiles are shown for all off-target sites predicted by Cas-OFFinder as well as for the top 1000 sites (ranked by CasKAS RPM values over the ±500bp region around the sgRNA target site).

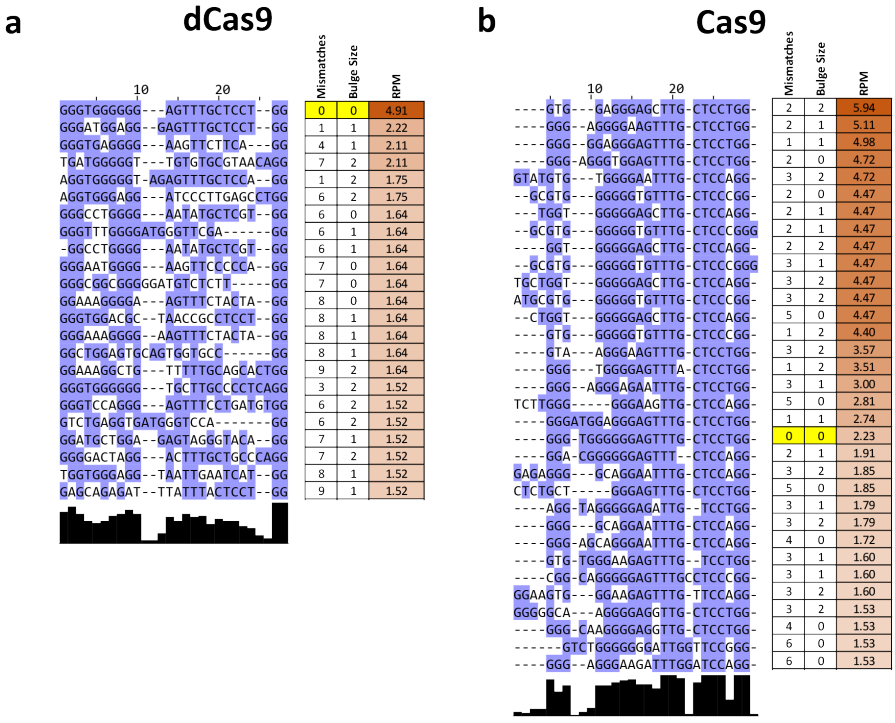


Supplementary Figure 10: Multiple sequence alignment of off-target sites identified by *in vitro* dCas9 and Cas9 CasKAS for the "Nanog-sg2" sgRNA. Shown are the top 100 off-target sites as predicted by Cas-OFFinder and ranked by CasKAS signal. The on-target site (if within the top 100) is highlighted in yellow.

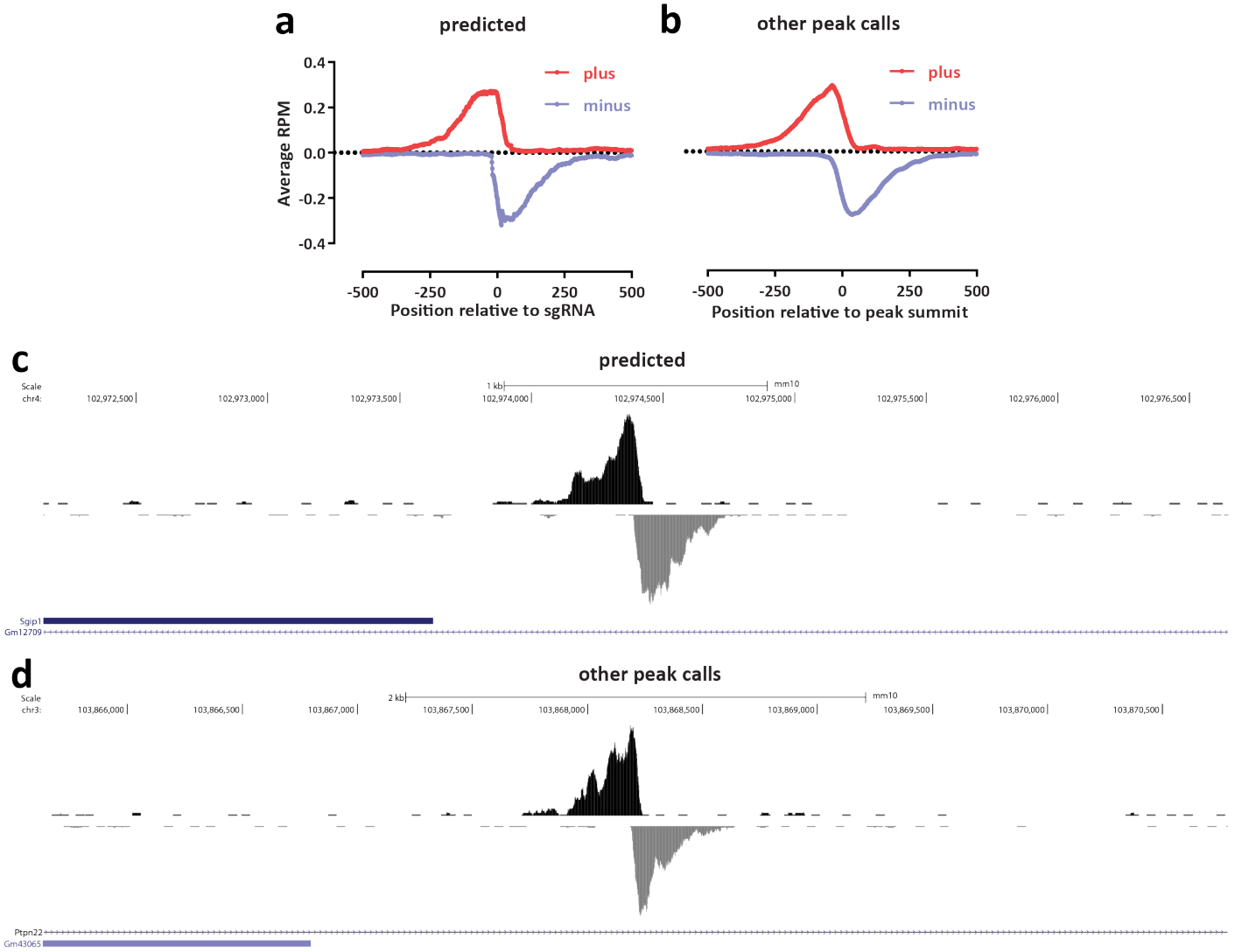
Supplementary Figure 11: Multiple sequence alignment of off-target sites identified by *in vitro* dCas9 and Cas9 CasKAS for the "Nanog-sg3" sgRNA. Shown are the top 100 off-target sites as predicted by Cas-OFFinder and ranked by CasKAS signal. The on-target site (if within the top 100) is highlighted in yellow.



Supplementary Figure 12: Multiple sequence alignment of off-target sites identified by *in vitro* dCas9 and Cas9 CasKAS for the "EMX1_Tsai" sgRNA. Shown are the top 100 off-target sites as predicted by Cas-OFFinder and ranked by CasKAS signal. The on-target site (if within the top 100) is highlighted in yellow.

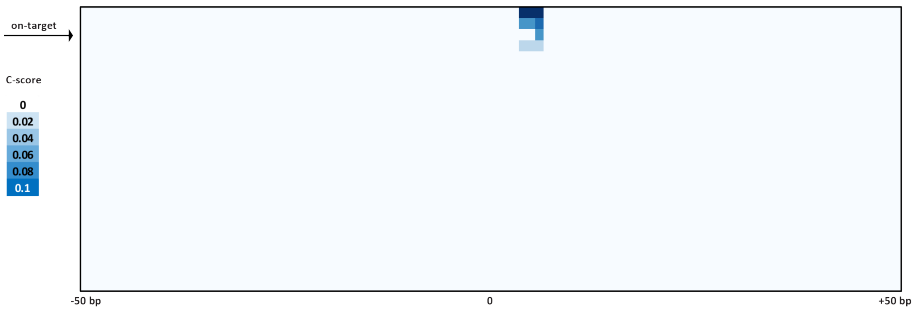


Supplementary Figure 13: Multiple sequence alignment of off-target sites identified by *in vitro* dCas9 and Cas9 CasKAS for the "VEGFA-site1" sgRNA. Shown are the all target sites with RPM ≥ 1.5 as predicted by Cas-OFFinder and ranked by CasKAS signal. The on-target site (if within the top 100) is highlighted in yellow.

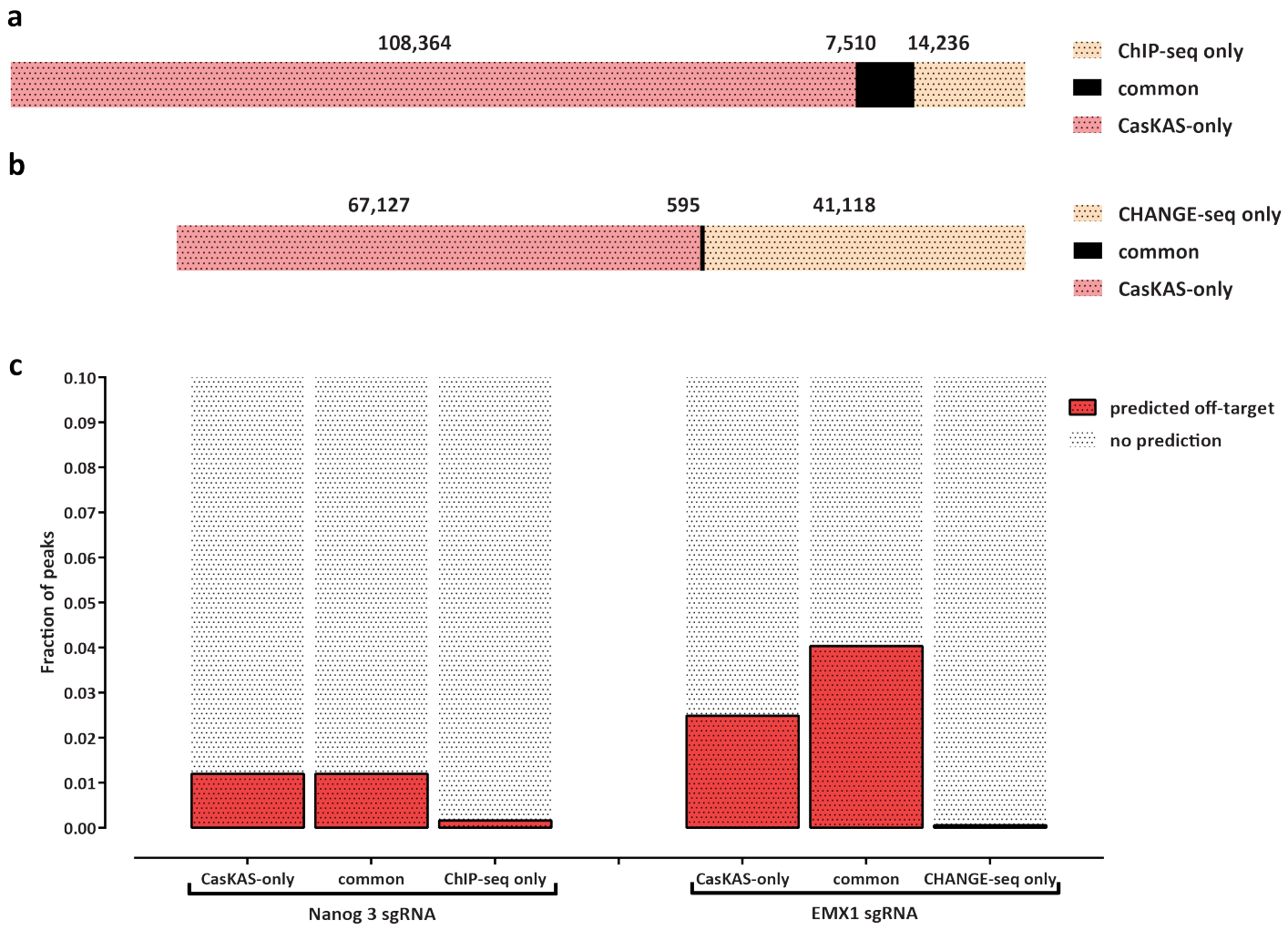


Supplementary Figure 14: CasKAS identifies proper off-target sites that are missed by sgRNA prediction algorithms. Shown is *in vitro* dCas9 CasKAS for the "sgRNA #1" sgRNA. Peaks were called *de novo* using MACS2, then intersected with Cas-OFFinder off-target prediction, and the outersect was manually filtered to exclude obvious artifacts based on peak shape (e.g. arising from repetitive elements in the genome). (a) Aggregate forward- and reverse-strand profiles around off-target sites predicted by Cas-OFFinder (centered on the sgRNA); (b) Aggregate forward- and reverse-strand profiles around sites not predicted by Cas-OFFinder (centered on the MACS2 peak summit); (c) Example UCSC Genome Browser snapshot of a CasKAS read profile around an off-target site predicted by Cas-OFFinder; (c) Example UCSC Genome Browser snapshot of a CasKAS read profile around an off-target site not predicted by Cas-OFFinder. Both predicted and identified through peak calling sites exhibit the expected asymmetric read distribution around a fixed occupancy point (the sgRNA-dCas9 RNP complexed with DNA).

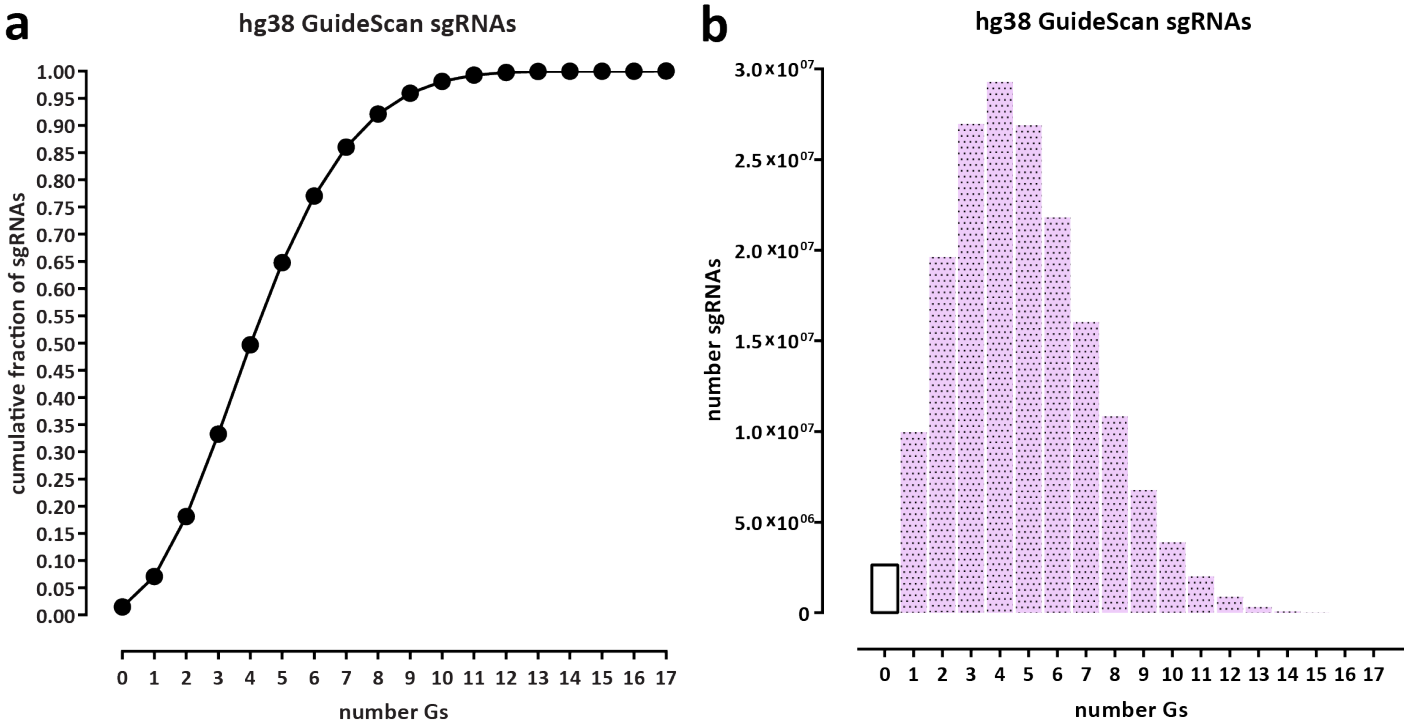
Supplementary Figure 15: Multiple sequence alignment of off-target sites identified by *in vitro* dCas9 and Cas9 CasKAS for the "sgRNA #1" sgRNA outside the list of predicted off-targets by Cas-OFFinder. MACS2 peak calls were manually filtered to exclude artifactual peaks, then the sequence of the ±50-bp region around the peak summit was used as input to the multiple sequence alignment, together with the sgRNA itself.



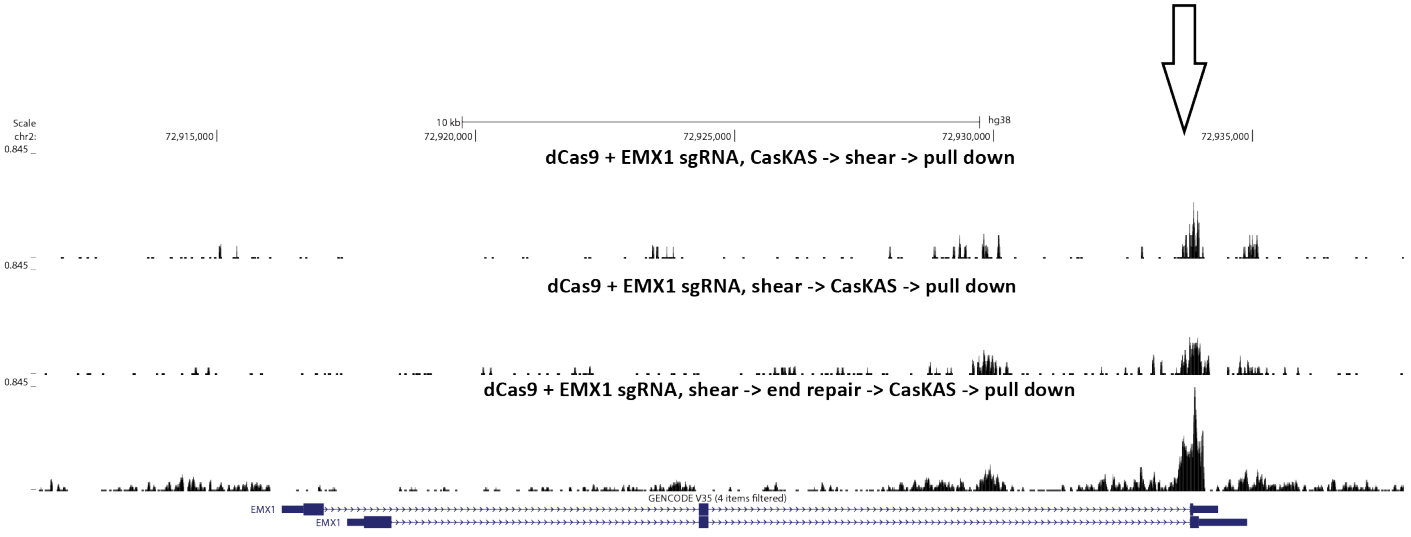
Supplementary Figure 16: Cutting profiles around on- and off-target sites for the VEGFA sgRNA. Four sites where cleavage is observed are identified within the list of predicted off-targets.



Supplementary Figure 17: Comparing *in vitro* dCas9 results to using ChIP-seq and CHANGE-seq for off-target profiling. Shown is the overlap between MACS2 peak calls for the Nanog-sg3 sgRNA with Nanog ChIP-seq dataset (SRR1168384 from GEO accession ID GSE54745) in (a) and the EMX1 sgRNA with EMX1 CHANGE-seq (SRA accession SRX8227890) in (b). The fraction of peaks common or unique to each assay that are predicted to be off-targets for each sgRNA by Cas-OFFinder is shown in (c).



Supplementary Figure 18: Most sgRNAs in the human genome contain multiple G nucleotides and are thus subject to labeling by N_3-kethoxal. Statistics were calculated for all valid sgRNAs as defined by GuideScan[18] (a) Cumulative fraction of sgRNAs. (b) Absolute number of sgRNAs.



Supplementary Figure 19: CasKAS can be performed on pre-sheared DNA. CasKAS was performed *in vitro* using the EMX1 sgRNA, first, conventionally, by carrying out the CasKAS reaction, then isolating and shearing genomic DNA, and also by pre-shearing the DNA and carrying out the CasKAS reaction on the fragmented DNA. The concern in that case is that the presence of sticky ends containing Gs and unprotected from the action of the N_3-kethoxal would lower the background. This problem can be addressed by carrying out end repair on the sheared DNA prior to the CasKAS reaction.