

## Unexpected CRISPR off-target mutation pattern *in vivo* are not typically germline-like

Zhiting Wei<sup>1, #</sup>, Funan He<sup>2, #</sup>, Guohui Chuai<sup>1</sup>, Hanhui Ma<sup>3,\*</sup>, Zhixi Su<sup>2,\*</sup>, Qi Liu<sup>1,\*</sup>

1 Department of Central Laboratory, Shanghai Tenth People's Hospital; Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, China.

2 Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China.

3 Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, U.S.

\* Corresponding authors:

# These authors contribute equally to this work

Email addresses:

Hanhui Ma: Hanhui.Ma@umassmed.edu

Zhixi Su: zxsu@fudan.edu.cn

Qi Liu: qiliu@tongji.edu.cn

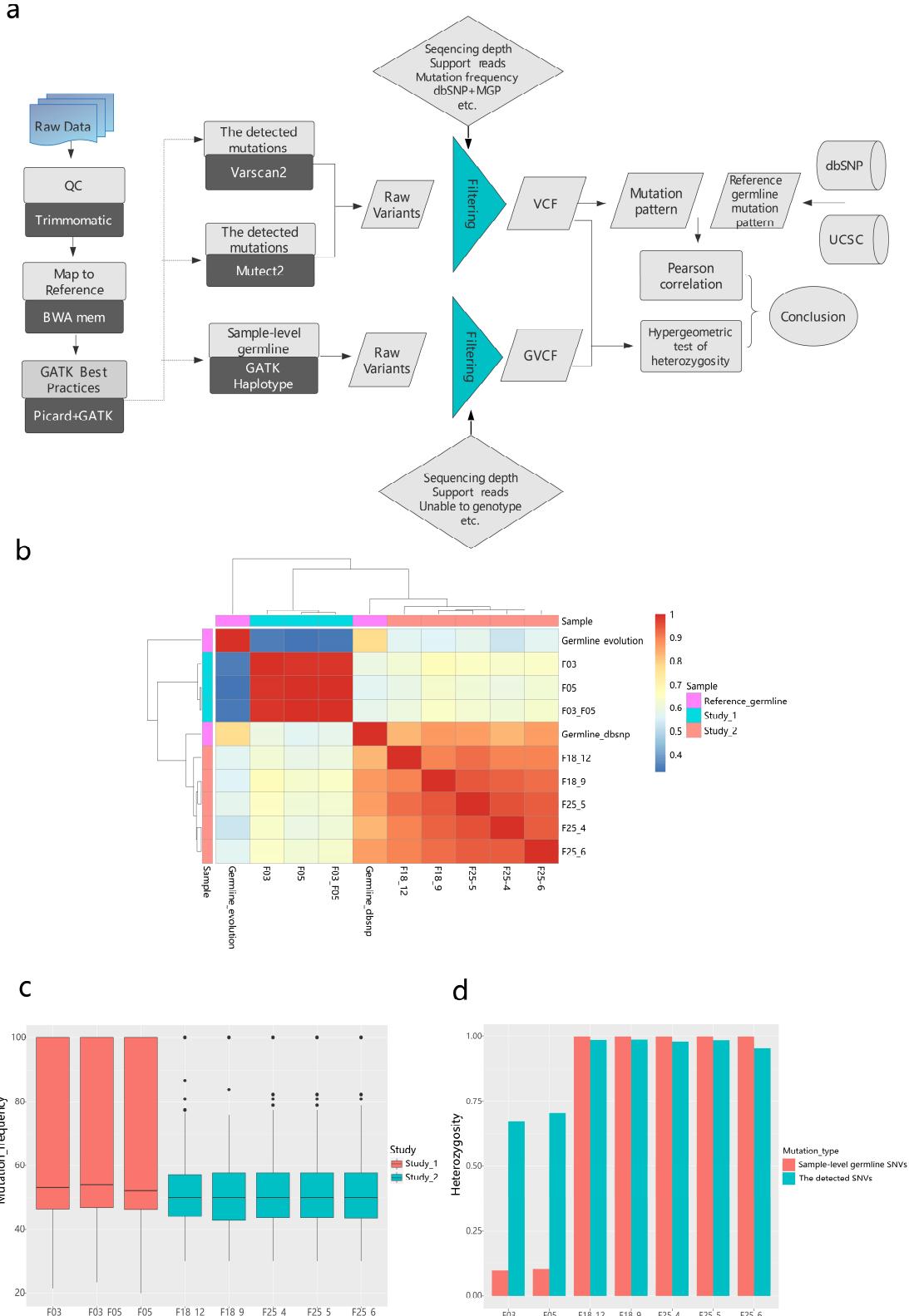
### To the Editor:

Schaefer et al.<sup>1</sup> (referred to as **Study\_1**) recently presented the provocative conclusion that CRISPR-Cas9 nuclease can induce many unexpected off-target mutations across the genome that arise from the sites with poor homology to the gRNA. As Wilson et al.<sup>2</sup> pointed out, however, the selection of a co-housed mouse as the control is insufficient to attribute the observed mutation differences between the CRISPR-treated mice and control mice. Therefore, the causes of these mutations need to be further investigated. In 2015, Iyer et al.<sup>3</sup> (referred to as **Study\_2**) used Cas9 and a pair of sgRNAs to mutate the *Ar* gene *in vivo* and off-target mutations were investigated by comparison the control mice and the offspring of the modified mice. After analyzing the whole genome sequencing (WGS) of the offspring and the control mice, they claimed that off-target mutations are rare from CRISPR-Cas9 engineering. Notably, their study only focused on indel off-target mutations. We re-analyzed the WGS data of these two studies and detected both single nucleotide variants (SNVs) and indel mutations.

Because these two studies draw relatively opposite conclusions on the off-targeting of the CRISPR system on the whole genome, the origins or causes of the mutations need to be cautiously examined. Here we performed a computationally evolutionary investigation (**Figure 1a**) to re-analyze the WGS data of these two studies with a direct comparison of the experiment design and analysis results (**Table 1** and **Supplementary Table 1**). The computational framework designed for the above analysis can accurately

infer the likelihood of the origins of these mutations, *i.e.*, whether these mutations are germline-like or not. This controversial issue arises substantially arguments but is unresolved<sup>1-4</sup>. Here our analysis concluded that the so-called unexpected SNVs pattern (**Supplementary Notes** and **Figure 1a**) in **Study\_1** are not typically germline-like, while for **Study\_2**, the detected SNVs are in fact germline mutations.

Thousands of genomic mutations were found from both studies based on our computational framework (**Supplementary Table 2**). Following our strict criteria to filter SNVs with low mutation frequency and indels overlapping with the UCSC Genome Browser<sup>5,6</sup> short tandem repeat, we confirmed a low number of so-called unexpected mutations claimed in **Study\_1**<sup>1</sup>(**Supplementary Table 2**). A great deal of filtered mutations are potentially false positives arising from spontaneous mutations, sequencing and Burrows-Wheeler Aligner(BWA)<sup>7</sup> alignment errors. This leads us believe that the calling of CRISPR-induced mutations especially indels should be carefully performed to avoid false positives. We integrated two computational methods to infer the putative origins of these detected SNVs, by (1) quantitatively analyzing the similarity of the mutation pattern<sup>8-11</sup> (**Supplementary Table 3**) between the reference germline SNVs (derived from well-curated public databases<sup>5,9</sup>; **Figure 1a** and **Supplementary Notes**) and the detected SNVs derived from **Study\_1** and **Study\_2** followed our re-analysis pipeline. The rationale is that for a particular organism (e.g., mouse), its germline mutation pattern should be evolutionarily conserved; (2) we applied a hypergeometric distribution to test the difference of mutation heterozygosity between the sample-level germline SNVs (detected using *GATK HaplotypeCaller*; **Figure 1a** and **Supplementary Notes**) and the detected SNVs.



**Figure 1 (a)** Overview of our computational framework to re-analyze the WGS data of these two studies. **(b)** Heatmap of the similarities of the mutation pattern between the reference germline SNVs and the detected SNVs. **(c)** The detected SNVs mutation frequency in these two studies. (F03 and F05 are the samples from Study\_1, F03\_F05 SNVs is the intersection SNVs of these two samples. F18\_9, F18\_12, F25\_4, F25\_5 and F25\_6 are the samples from Study\_2). **(d)** Heterozygosity of the detected SNVs and the sample-level germline

SNVs.

As seen in the heatmap (**Figure 1b**), the detected SNV mutation pattern correctly cluster the samples from different studies. In addition, the mutation pattern of samples from **Study\_2** highly correlated with the reference dbSNP germline mutation pattern ( $R^2 \sim 0.85$ ,  $P\text{-value} < 2.2\text{e-}16$ ), while the correlation between the samples from **Study\_1** and the reference dbSNP germline was weaker ( $R^2 \sim 0.55$ ,  $P\text{-value} \sim 1\text{e-}10$ ). Although the correlation of mutation pattern in either study with germline at evolution scale are relatively low, we found that the samples from **Study\_2** correlated stronger with germline mutation pattern than those of samples from **Study\_1**.

The heterozygosity of the so-called unexpected SNVs of the samples in **Study\_1** was 67.2% and 70.4%, respectively (**Figure 1c**), while that of the sample-level germline SNVs was much lower (**Figure 1c**). We reasoned that if all the detected SNVs were germline, the heterozygosity of such SNVs would be nearly identical to that of the sample-level germline SNVs. Our statistical test, however, indicates a significant difference between the heterozygosity of them ( $p\text{-value}=0$ ; **Supplementary Notes**), proving that the overall derived mutation pattern in **Study\_1** is not germline-like. Note that Lareau et al. and Kim et al.<sup>4,6</sup> recently demonstrated that the two CRISPR-Cas9 treated mice (F03, F05) in **Study\_1** are actually more closely related to each other genetically than to the control mouse, proving that these mutations are most likely pre-existing variants. Our analysis result further presented that there are unusual SNVs arose in **Study\_1** combined with pre-existing variants, and the overall mutation pattern *in vivo* are not germline-like. In contrast, in **Study\_2**, we cannot reject the null hypothesis that the detected SNVs were sampled from the germline background. We computationally validated that the majority of the detected mutations in **Study\_2** are germline, which is consistent with their original notation that off-target somatic mutations are rare in their Cas9-modified mice<sup>3</sup>.

Because the same CRISPR genome-editing technology led to different conclusions in two studies, we performed a direct comparison of these two studies to evaluate the possible causes for the differences (**Table 1** and **Supplementary Table 1**). First, the experimental protocols differ, where the protocol used in **Study\_1** is not a routine and well-accepted CRISPR knockout protocol (sgRNA plasmid + Cas9 protein). The potentiality to induce unusual mutation patterns with such protocol are waiting to be further explored. Secondly, the whole genome sequencing of the founder mouse should be performed before CRISPR knockout. The founder genetic variation should be carefully examined in both studies, since such genetic variation, which may contain in the zygotes, can confound the target sites of certain sgRNAs more than others. This information should be integrated into the study for sgRNA selection to ensure safety<sup>12</sup>.

In summary, we demonstrated that the unexpected CRISPR off-target mutation pattern in **Study\_1** are not typically germline-like. Some of unusual and unidentified mutations may arise in **Study\_1**, but the real reasons remain to be explored. Based on

the available data and a direct comparison of the two studies, we presented two possible reasons and future re-analysis directions that may contribute to such different conclusions. To characterize the authentic CRISPR-mediated mutations, we are required to have appropriate controls to rule out other sources of mutations, which will be needed for benchmarking of targeting safety of CRISPR-based gene therapy.

**Table 1. A direct comparison of Study\_1 and Study\_2**

Mechanism	sgRNA	Experiment	Sequencing	OT-sites in repair-related region	DNA	Detected mutations in DNA repair-related gene	Evaluate
			sample	in gene		in DNA repair-related gene	Cas9 activity
<b>Study_1</b>	HDR	CCAACCTAAGTAGCAGAAAG	sgRNA plasmid +Cas9 protein +ssODNA	F0 founder	None	None	NO
<b>Study_2</b>	NHEJ	GGAGGCAGCTGCTCTCAGGG GGCGGCCACCGCTCCTGGCA	sgRNA +Cas9 mRNA	+Cas9 F1 offspring	None	None	YES

OT-sites: off-target sites;

## Competing financial interests

The authors declare no competing financial interests.

## Acknowledgments

This work was supported by the National Major Research and Innovation Program of China (Grant No. 2016YFC1303205, SQ2017YFSF090222), National Natural Science Foundation of China (Grant No. 61572361), Shanghai Rising-Star Program (Grant No. 16QA1403900), and Shanghai Natural Science Foundation Program (Grant No. 17ZR1449400).

## References:

- 1 Schaefer, K. A. *et al.* Unexpected mutations after CRISPR-Cas9 editing in vivo. *Nature methods* **14**, 547-548, doi:10.1038/nmeth.4293 (2017).
- 2 Wilson, C. J. The experimental design and data interpretation in “Unexpected mutations after CRISPR–Cas9 editing in vivo” by Schaefer et al. are insufficient to support the conclusions drawn by the authors. *bioRxiv* 153338 (2017)
- 3 Iyer, V. *et al.* Off-target mutations are rare in Cas9-modified mice. *Nature methods* **12**, 479, doi:10.1038/nmeth.3408 (2015).
- 4 Kim, S.-T. *et al.* Questioning unexpected CRISPR off-target mutations in vivo. doi:10.1101/157925 (2017).
- 5 Doerks, T., Copley, R. R., Schultz, J., Ponting, C. P. & Bork, P. The Human Genome Browser

at UCSC. *Genome research* **12**, 47-56, doi:10.1101/ (2002).

6 Mijuskovic, M. A Streamlined Method for Detecting Structural Variants in Cancer Genomes by Short Read Paired-End Sequencing. doi:10.1371/journal.pone.0048314.t001.

7 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

8 Zhu, Y. O., Sherlock, G. & Petrov, D. A. Extremely Rare Polymorphisms in *Saccharomyces cerevisiae* Allow Inference of the Mutational Spectrum. *PLoS genetics* **13**, e1006455, doi:10.1371/journal.pgen.1006455 (2017).

9 Smigielski, E. M., Sirotnik, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research* **28**, 352-355 (2000).

10 Chen, F. C., Chen, C. J., Li, W. H. & Chuang, T. J. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome research* **17**, 16-22, doi:10.1101/gr.5429606 (2007).

11 Carlson, J. *et al.* Extremely rare variants reveal patterns of germline mutation rate heterogeneity in humans. *bioRxiv*, 108290 (2017).

12 David A Scott & Feng Zhang, Implications of human genetic variation in CRISPR-based therapeutic genome editing, *Nature Medicine*, 10.1038/nm.4377 (2017).