

Supplementary information for

“Unexpected mutations after CRISPR-Cas9 editing in vivo” are most likely pre-existing SNPs and not nuclease-induced mutations

Caleb A. Lareau,^{*} Kendell Clement,^{*} Jonathan Y. Hsu,^{*} Vikram Pattanayak, J. Keith Joung,⁺ Martin J. Aryee,⁺ Luca Pinello⁺

Overview

This document contains six key Supplementary notes to describe the data processing and analysis for our reanalysis of the data published by Schaefer *et al.* 2017¹ describing unexpected mutations attributed to CRISPR-Cas9. We’ve also included several figures and tables to further elucidate our analysis workflow and results. We note that our code, analysis, and workflow is accessible online at http://aryeelab.org/crispr_mutation_reanalysis

Notes

- 1) Supplementary Note 1: **Variant calling and reproducibility of original results**
- 2) Supplementary Note 2: **Genetic relatedness trees.**
- 3) Supplementary Note 3: **Closest off-target analysis.**
- 4) Supplementary Note 4: **Motif enrichment analysis.**
- 5) Supplementary Note 5: **Analysis of shared indels between F03 and F05 mice.**
- 6) Supplementary Note 6: **Closest common variant analysis.**

Tables

- 1) Supplementary Table 1: **Heterozygosity proportions in the mice.**

Figures

- 1) Supplementary Figure 1: **Reproduction of reported variant calls.**
- 2) Supplementary Figure 2: **Distance distributions of Schaefer *et al.* reported SNVs to closest Cas-OFFinder predictions.**
- 3) Supplementary Figure 3: **Motif enrichment at variants identified through the cancer pipeline.**
- 4) Supplementary Figure 4: **Empirical evaluation of indel concordance in F03 and F05 mice.**
- 5) Supplementary Figure 5: **Base substitution comparison between F03 and F05.**
- 6) Supplementary Figure 6: **Reported variant proximity to dbSNP loci.**

Supplementary Note 1

Though differing slightly from the authors' alignment and variant calling methodology,(CITE) we reprocessed the raw sequencing data from the three mice using the Genome Analysis Toolkit (GATK) best practices pipeline.² Specifically, each mouse was genotyped using the Haplotype Caller functionality in GATK version 3.7. For the set of mice, we filtered the full set GATK calls to 70,059 variants that were covered with at least 23x in each sample (as suggested by Schaefer et al.), had a minimum quality of 300, had only two alleles, and where two samples shared the same genotype and the third was different. 31,078 of these SNVs overlapped with known dbSNP calls, and 38,981 were absent from a GATK annotation, which we term "novel." Notably, the original manuscript failed to provide a complete genotyping analysis of these three mice. Supplementary Table 1 provides an overview of the number of variants called including the proportion of heterozygous variants, which clearly deviates from 0% as suggested by the authors in their initial rebuttal.

To assess the concordance of our analysis with the previously published results, we replicated what we term the "cancer pipeline" analysis where somatic variants were called using the F03 and F05 mice as "tumor" samples and the FVB mouse as the "normal" sample. Variants corresponding to the reported dbSNP annotation were filtered, and the final set of variants called per mouse under this framework was the intersection of variants inferred by Lofreq, Mutect, and Strelka as previously suggested. We found that 91.1% and 93.5% (Supplementary Figure 1) of the originally reported variants could be replicated in our analysis with differences owing to variation in the bioinformatics analyses where the authors deviated from GATK best practices. These deviations include the exact choice of aligner. Moreover, as the authors reported differential coverage for these mice (50x for F03 and F05 but 30X for FVB), we downsampled the two treated mice to 30x coverage to infer that this analysis pipeline was robust to the variation in the sequencing depth as 97% of the reproduced variants in the full data were also called in the downsampled (Supplementary Figure 1).

While the use of these somatic variant calling pipelines elucidate some degree of the shared mutations within F03 and F05 distinct from FVB, we determined that this analysis framework was underpowered as only 7.8% of the true differences shared between the treated mice relative to the control inferred by GATK were inferred from the cancer analysis pipeline (Figure 1D). Moreover, when permuting the normal and "tumor" (Cas9-treated) tissue types associated with the cancer pipeline, we observed a similar proportion of heterozygous variants (Supplementary Table 1).

Supplementary Note 2

To succinctly visualize the genetic relatedness between these three mice given the mutational landscape associated with them, we computed genetic relatedness trees (shown in Figure 1E) using an approach as outlined at:

http://userweb.eng.gla.ac.uk/cosmika.goswami/snp_calling/SNPCalling.html. Specifically, we selected SNVs called by GATK with at least 23 reads in each sample (the cutoff used by Schaefer et al.¹), at least 300 quality, and one sample having a different genotype than the other two.

We first created a representative variant fasta pseudo-sequence for all three samples using these filtered GATK variant calls by concatenating the alternate allele from each SNV (if present) to the end of the sequence representing the sample. These sequences were necessarily aligned because one nucleotide is added to the end of the pseudo-sequence for each sample regardless of whether the genotype is reference or an alternate allele.

Next, we used FastTree to infer the approximately-maximum-likelihood phylogenetic tree for the three samples. In newick tree format, this is represented as: (F03:0.13789, F05:0.13635, FVB:0.22576);. The images rendered in Figure 1E reflect the exact numeric distances computed by the tool. Further details of these scripts are available online.

Supplementary Note 3

To validate the original finding reported by Schaefer et al. that no DNA sequences resembling the purported CRISPR-Cas9 on-target site, we used Cas-OFFinder to perform in-silico off-target predictions with mismatches up to 6bp and DNA/RNA bulges up to 2bp.³ We plotted the log10 minimum distance from the Schaefer et al. SNVs to the closest predicted off-target sites, and also sampled sets of permuted common dbSNPs to establish a background model. Our results confirmed that there was no clear deviation of the in-silico predicted off-targets near Schaefer et al. SNVs compared to randomly sampled common dbSNPs, suggesting a lack of sequence-specificity associated with the reported variants in this finding.

Supplementary Note 4

To test if the specific co-localization of the thousand of SNPs shared between F03 and F05 treated mice could be explained by DNA sequence preference, we performed de-novo motif analysis using HOMER.⁴ Motifs were inferred using sequences of 100bp centered at the SNPs obtained with the cancer pipeline proposed by Schaefer *et al.* (mm10 reference genome) for the groups F03 vs FVB and F05 vs FVB as target sets. As background set, we used sequences of 100bp centered at any dbSNPs variant recovered by GATK in any of the three mice. From this analysis, no single enriched motif was shared between F03 and F05.

Supplementary Note 5

To further support our inference of increased treated genetic relatedness, we analyzed the extent of insertion and deletion (indel) mutations shared between the F03 and F05 mice. We were able to identify all F03 and F05 indels outlined in the paper, however, an additional 2 and 4 non-exonic indels for F03 and F05, respectively, were also detected with our processing steps. When filtering for common indel locations between the genome-edited mice, we found 118 shared indels (117 shared indels in Schaefer *et al.*). Importantly, we did not require the exact indel mutation to match, but only required the indels to exist within 20bp of each other. Strictly looking at indel size, there was complete agreement across all the shared indels between F03 and F05 (Supplementary Figure 4, blue). Notably, the mutated sequence also completely matched across these 118 shared indels.

CRISPR-Cas9-induced double stranded breaks result in a heterogeneous indel population. Recent studies have suggested that population-level indel distributions are nonrandom, however, single indel events are still stochastic in nature.⁵ (With this in mind, we sought to develop a statistical test to assess the probability of observing occurrences where two independent genome cleavage events via CRISPR-Cas9 result in identical indel lengths. Assuming that these highly specific off-target indels, as suggested by the authors, follow similar mechanisms of DNA repair after CRISPR-Cas9-induced DSBs and that indel events in the F03 and F05 mice are independent, we approximated the probability of realizing the observed data from an empirically defined background model. We construct empirical indel size distributions based on deep sequencing data of 96 unique gRNAs from van Overbeek *et al.*,⁵ sample a gRNA from this group, and finally sample two indel sizes from the distribution to simulate paired occurrences of indel formation. We simulate this process 118 times to visualize the simulation on the scale of the observed data (Supplementary Figure 4, red).

For a conservative approximation of this indel concordance observation, we chose the least heterogeneous indel distribution (maximum of the binomial probabilities) across all the indel size distributions and performed a binomial test to assess the significance of observing 118 paired indel size matches ($p < 1.5 \times 10^{-42}$). From these simulations, we suggest that the 118 observed identical indel mutations between the F03 and F05 mice were independent events due to nuclease-induced double-strand DNA breaks. The more likely explanation for this striking perfect alignment of indel mutations is pre-existing and shared private indel mutations between the genome-edited mice.

Supplementary Note 6

Finally, we tested to see if the SNVs discovered by Schaefer et al. were found to be close to annotated common variant SNVs that showed the same pattern as the Schaefer et al. SNVs. We selected 31,078 common variants in dbSNP that were in our filtered GATK SNVs. We noted that 41.7% of these dbSNP variants were the same between F03 and F05, and differed in FVB. We identified which of these dbSNPs was closest to the 1,373 SNPs called by the cancer pipeline, and found that 54.4% of SNVs from the cancer pipeline were closest to dbSNP variants that were shared between F03 and F05. This is significantly higher than would be expected given the distribution of these SNVs in all dbSNPs (two-sided Fisher's Exact Test, $p < 2 \times 10^{-19}$).

Supplementary Table 1

SNPs	All GATK SNPs*			Overlap w/Cancer Pipeline F03/F05**		Overlap w/Permuted Cancer Pipeline***
	F03	F05	FVB	F03	F05	FVB vs (F03/F05)
# Heterozygous calls	732263	664436	580338	781	781	136
# Total Calls	7087153	6972046	6809736	1256	1256	222
% Heterozygous	10.33%	9.53%	8.52%	62.18%	62.18%	61.26%

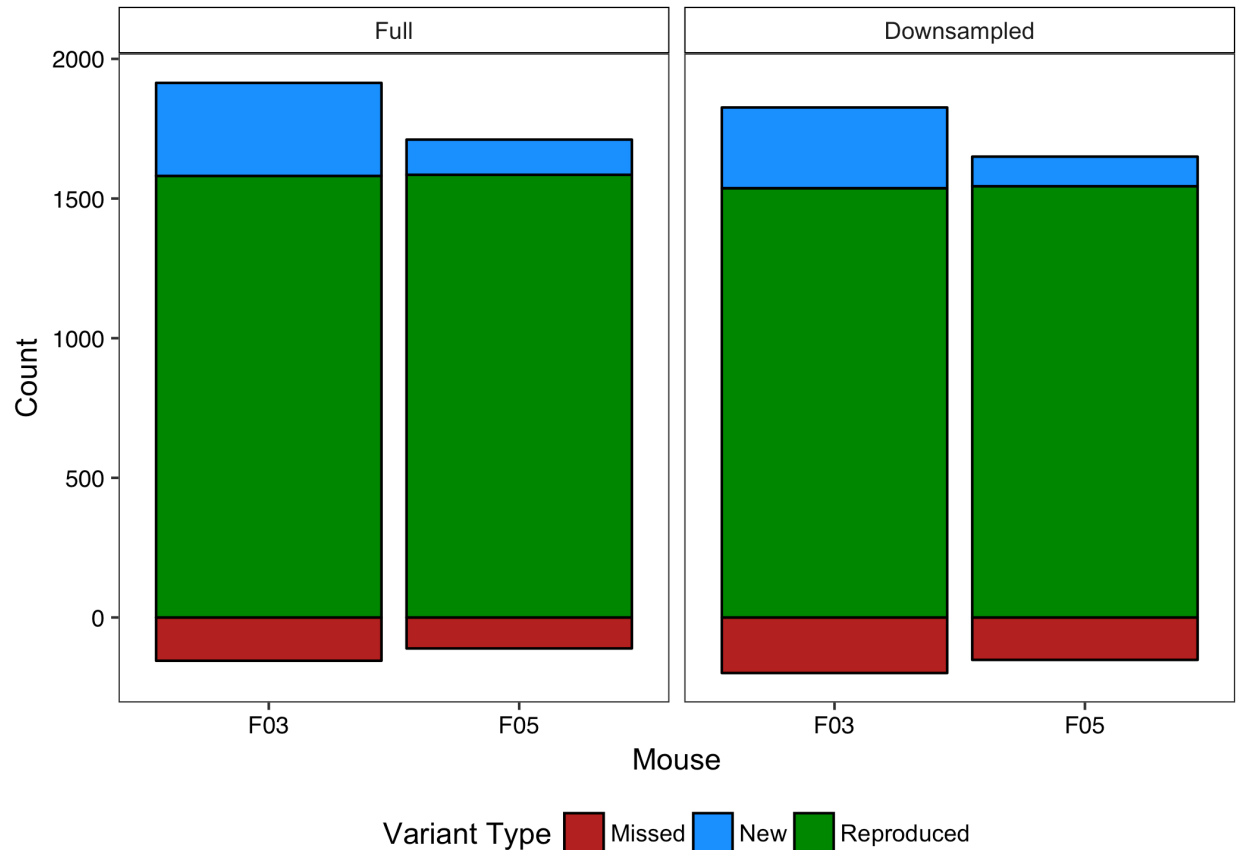
Heterozygosity proportions in the mice. Schaefer *et al.* suggest that allelic heterozygosity provides evidence for Cas9 off-target activity. After calling SNPs using the GATK Best Practices pipeline (see methods), we found that 8.5% of SNPs in the FVB were heterozygous, similar to heterozygosity rates of the F03 and F05 samples (10.33% and 9.53% respectively, and these rates were lower when downsampling the uneven coverage). At GATK SNPs that overlapped with the intersection of F03 and F05 SNPs as called by the cancer variant calling pipeline, we confirm that approximately 60% of SNPs are heterozygous, which is similar to the rates reported in (the Schaefer *et al.* rebuttal table). Although this rate is noticeably high, we permuted the samples in the cancer variant calling pipeline to call SNVs in FVB as compared to either F03 or F05 as a “normal” tissue. At the intersection of these FVB-F03 and FVB-F05 SNPs, we find an equally high rate of heterozygosity (61.96%), suggesting that the observed heterozygous genotype frequency is likely an artifact of the application of Schaefer *et al.*’s methodology and is independent of Cas9 activity.

*All GATK SNPs were called using the standard GATK Best Practices pipeline, then filtered for those with quality scores of at least 20, and at least 10 reads

**Overlap w/Cancer Pipeline F03/F05 are GATK Filtered SNPs that overlap with SNPs called using the intersection of F03-FVB and F05-FVB from Strelka, Lofreq, and MuTect

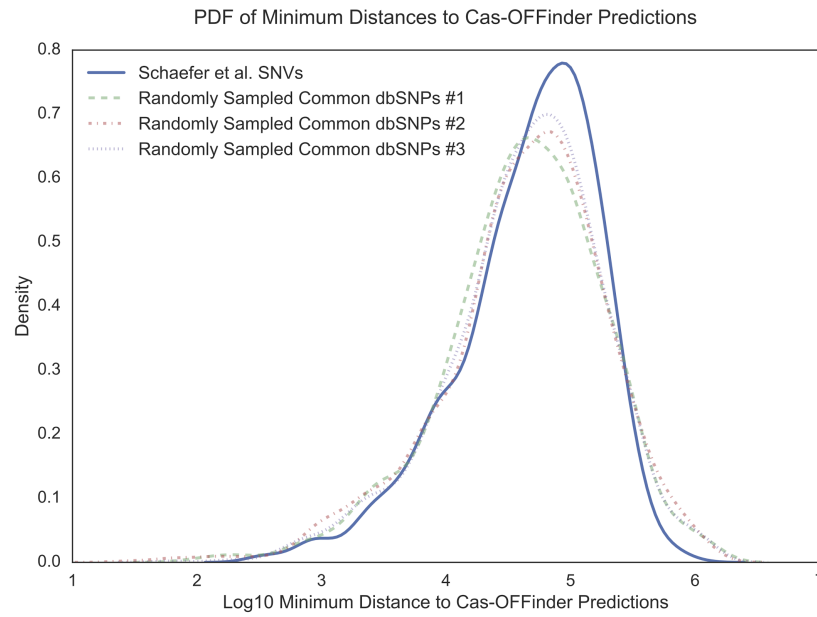
***Overlap w/Permuted Cancer Pipeline are GATK Filtered SNPs that overlap with SNPs called using the intersection of FVB-F03 and FVB-F05 from Strelka, LoFreq, and Mutect. In other words, these variants are called through these intersecting pipelines described in Schaefer *et al.* and represent

Supplementary Figure 1



Reproduction of reported variant calls. Variants were called using the intersection of three somatic cancer mutation calling software frameworks for each mouse using FVB as a “normal” tissue. We were able to reproduce 91.1% and 93.5% of the previously reported variants (green; left) though our analysis framework differed slightly as we employed GATK best practices. The number of mutations called was and concordant with the previously reported variants was similar for an identical iteration of the pipeline ran on downsampled data for the F03 and F05 mice (right), indicating that the differential coverage reported by the authors did not have a substantial impact on the analysis.

Supplementary Figure 2



Distance distributions of Schaefer *et al.* reported SNVs to closest Cas-OFFinder predictions.

Log10 minimum distances were plotted for all Schaefer *et al.* reported SNVs to the closest Cas-OFFinder in-silico off-target prediction (mismatch up to 6bp, DNA/RNA bulges up to 2bp). Three independent randomly sampled common dbSNPs were also plotted to assess the background of the analysis.

Supplementary Figure 3

Homer *de novo* Motif Results (HOMER_DENOVO_treatedF05_normalFVB.bed.filt_vs_GATK.out.dbSNP.bed/)

[Known Motif Enrichment Results](#)

[Gene Ontology Enrichment Results](#)

If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into [STAMP](#)

More information on motif finding results: [HOMER](#) | [Description of Results](#) | [Tips](#)

Total target sequences = 1667

Total background sequences = 85087

* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-12	-2.942e+01	3.49%	1.10%	26.6bp (29.3bp)	PB0051.1_Osr2_1/Jaspar(0.637) More Information Similar Motifs Found	motif file (matrix)
2		1e-12	-2.938e+01	1.56%	0.23%	29.2bp (29.1bp)	MAFK/MA0496.1/Jaspar(0.734) More Information Similar Motifs Found	motif file (matrix)
3		1e-12	-2.795e+01	0.72%	0.03%	17.5bp (29.1bp)	CRZ1(MacIsaac)/Yeast(0.588) More Information Similar Motifs Found	motif file (matrix)
4 *		1e-11	-2.726e+01	4.87%	1.99%	26.5bp (30.0bp)	Nr2e1/MA0676.1/Jaspar(0.805) More Information Similar Motifs Found	motif file (matrix)
5 *		1e-11	-2.719e+01	26.23%	19.13%	27.8bp (29.5bp)	PABPN1(RRM)/Homo_sapiens-RNCMP00157-PBM/HughesRNA(0.738) More Information Similar Motifs Found	motif file (matrix)
6 *		1e-11	-2.645e+01	7.04%	3.48%	28.0bp (28.7bp)	AZF1/MA0277.1/Jaspar(0.715) More Information Similar Motifs Found	motif file (matrix)
7 *		1e-11	-2.572e+01	1.81%	0.36%	28.8bp (29.8bp)	MeI2(dmmpmm)/Papatsenko/fly(0.867) More Information Similar Motifs Found	motif file (matrix)
8 *		1e-10	-2.426e+01	0.42%	0.00%	26.3bp (15.6bp)	HNRNP2(RRM)/Homo_sapiens-RNCMP00160-PBM/HughesRNA(0.664) More Information Similar Motifs Found	motif file (matrix)
9 *		1e-10	-2.426e+01	0.42%	0.00%	23.7bp (23.1bp)	dHNF4(NR)/Fly-HNF4-ChIP-Seq(GSE73675)/Homer(0.632) More Information Similar Motifs Found	motif file (matrix)
10 *		1e-10	-2.426e+01	0.42%	0.00%	25.7bp (10.3bp)	Pp_0206(RRM)/Physcomitrella_patens-RNCMP00206-PBM/HughesRNA(0.655) More Information Similar Motifs Found	motif file (matrix)

Homer *de novo* Motif Results (HOMER_DENOVO_treatedF03_normalFVB.bed.filt_vs_GATK.out.dbSNP.bed/)

[Known Motif Enrichment Results](#)

[Gene Ontology Enrichment Results](#)

If Homer is having trouble matching a motif to a known motif, try copy/pasting the matrix file into [STAMP](#)

More information on motif finding results: [HOMER](#) | [Description of Results](#) | [Tips](#)

Total target sequences = 1861

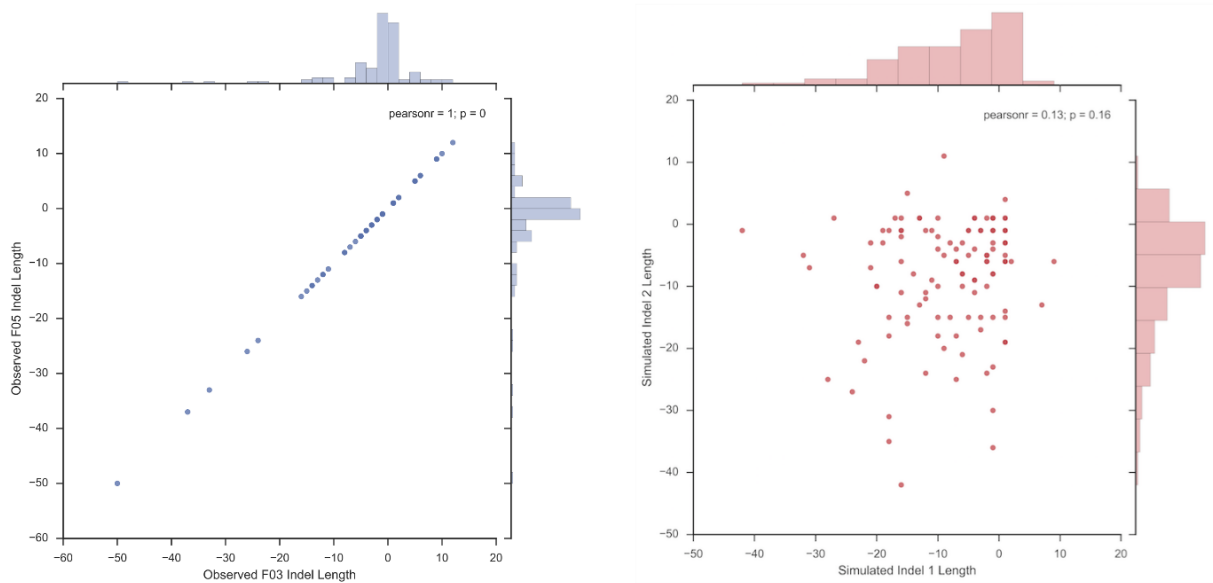
Total background sequences = 86198

* - possible false positive

Rank	Motif	P-value	log P-value	% of Targets	% of Background	STD(Bg STD)	Best Match/Details	Motif File
1		1e-16	-3.825e+01	1.02%	0.05%	22.4bp (27.6bp)	ONECUT3/MA0757.1/Jaspar(0.855) More Information Similar Motifs Found	motif file (matrix)
2		1e-16	-3.768e+01	1.24%	0.10%	18.8bp (27.0bp)	SFP1/SacCer-Promoters/Homer(0.747) More Information Similar Motifs Found	motif file (matrix)
3		1e-15	-3.656e+01	12.79%	7.29%	27.0bp (29.5bp)	S334(RRM)/Danio_rerio-RNCMP00224-PBM/HughesRNA(0.792) More Information Similar Motifs Found	motif file (matrix)
4		1e-15	-3.478e+01	2.74%	0.68%	25.6bp (27.4bp)	br-Z1(dmmpmm)/Down/fly(0.732) More Information Similar Motifs Found	motif file (matrix)
5		1e-14	-3.384e+01	1.45%	0.18%	23.7bp (28.0bp)	PB0133.1_Hic1_2/Jaspar(0.726) More Information Similar Motifs Found	motif file (matrix)
6		1e-14	-3.287e+01	0.97%	0.06%	25.5bp (28.1bp)	Rbpj1(?)Panc1-Rbpj1-ChIP-Seq(GSE47459)/Homer(0.690) More Information Similar Motifs Found	motif file (matrix)
7		1e-13	-3.206e+01	1.24%	0.13%	29.1bp (24.2bp)	PB0114.1_Egr1_2/Jaspar(0.668) More Information Similar Motifs Found	motif file (matrix)
8		1e-13	-3.069e+01	0.54%	0.01%	23.6bp (22.1bp)	SRSF10(RRM)/Homo_sapiens-RNCMP00088-PBM/HughesRNA(0.739) More Information Similar Motifs Found	motif file (matrix)
9		1e-12	-2.955e+01	12.25%	7.37%	26.1bp (30.8bp)	Tb_0230(RRM)/Trypanosoma_brucei-RNCMP00230-PBM/HughesRNA(0.751) More Information Similar Motifs Found	motif file (matrix)
10		1e-12	-2.782e+01	1.34%	0.20%	26.3bp (29.4bp)	PB0203.1_Zfp691_2/Jaspar(0.649) More Information Similar Motifs Found	motif file (matrix)

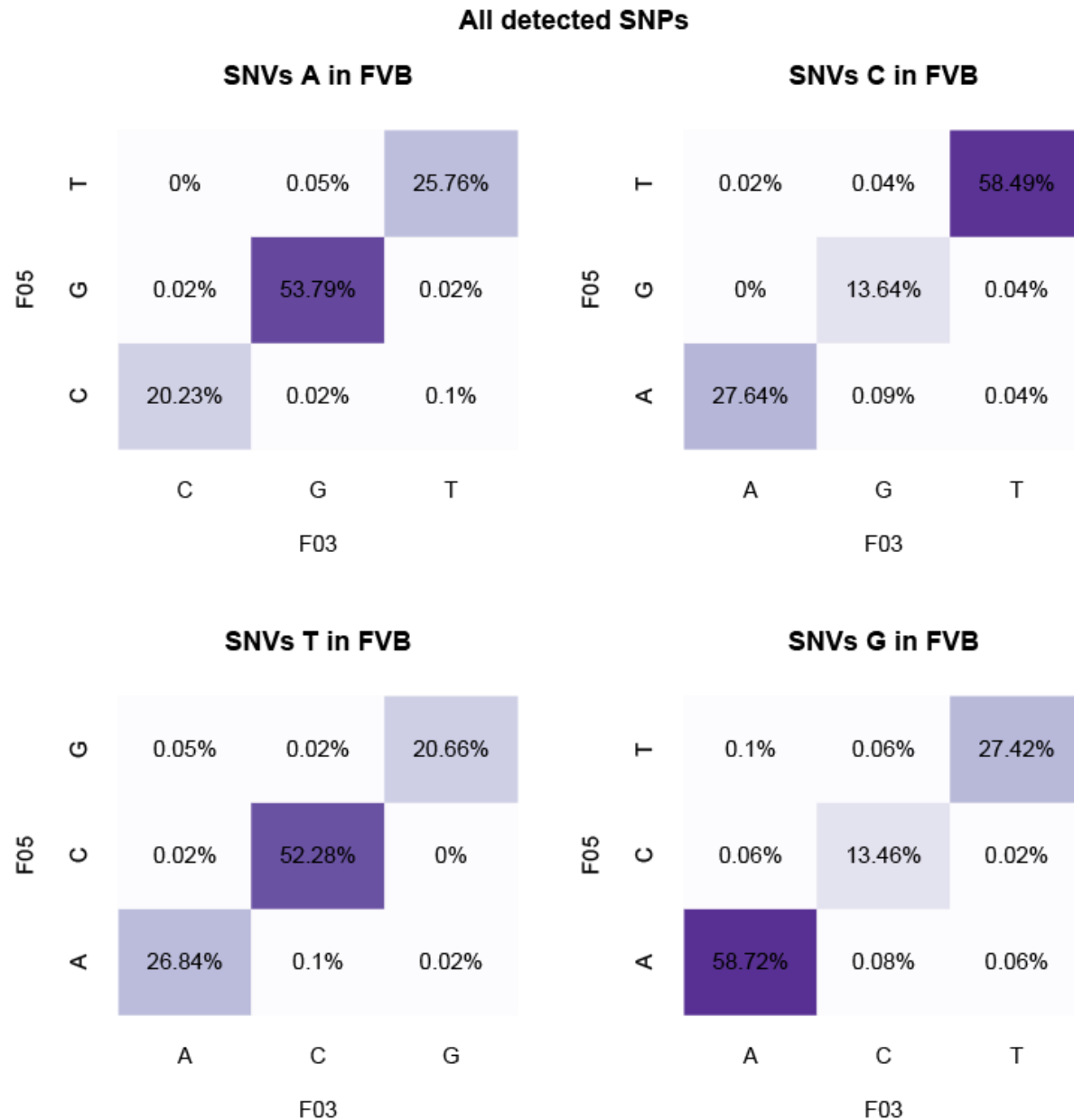
Motif enrichment at variants identified through the cancer pipeline. Motif sequences were inferred using HOMER. Motifs were inferred using sequences of 100bp centered at the SNPs obtained with the cancer pipeline proposed by Schaefer *et al.* (mm10 reference genome) for the groups F03 vs FVB and F05 vs FVB as target sets. No enriched motifs were present for both SNP sets, living no clear mechanism for potential shared DNA sequence recognition by Cas9.

Supplementary Figure 4



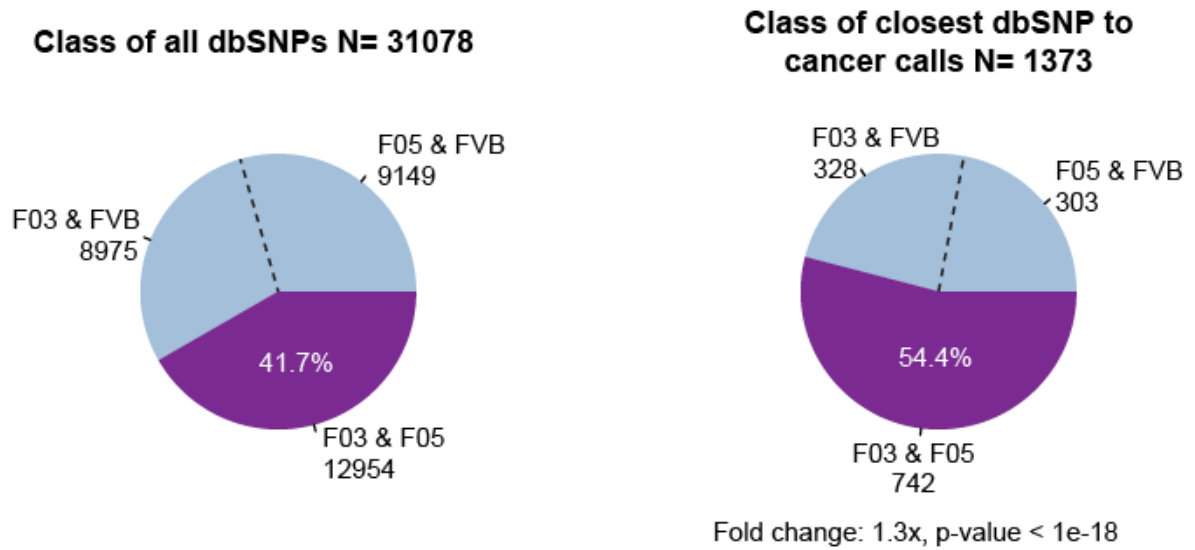
Empirical evaluation of indel concordance in F03 and F05 mice. We sought to quantify the empirical probability that 118 indels shared between F03 and F05 mice would all be the same length (left). From empirical indel distributions described in van Overbeek *et al.*, we simulated paired indel occurrences (right) to highlight the heterogeneity of the indel outcome process. With the least heterogeneous indel size distribution, we use the binomial test to assess the significance of observing 118 paired indel size matches ($p = 1.5 \times 10^{-42}$).

Supplementary Figure 5



SNP substitution comparison between F03 and F05. Heatmaps showing the overlap of the nucleotide observed at a SNVs in F03, and the nucleotide observed in F05 at the same location. In each heatmap, only SNVs that are homozygous reference, and then show a variant in both F03 and F05 are shown. Notably, the density along the diagonal shows that a vast majority (> 99%) of variant loci shared in F03 and F05 also share the same nucleotide base.

Supplementary Figure 6



Characterization of characterized mutations called from variant calling pipelines. Left: Pie chart showing the fraction of dbSNP variants called by the GATK pipeline that have the same genotype in F03 and F05 (but not in FVB). Right: Variants called by the cancer pipeline were mapped to the closest dbSNP variant. This pie chart shows the fraction of these dbSNP variants that had the same genotype in F03 and F05. The variants that are closest to calls from the cancer pipeline are significantly enriched for variants whose genotype is shared between F03 and F05.

References

1. Schaefer, K.A. et al. Unexpected mutations after CRISPR-Cas9 editing in vivo. *Nat Methods* **14**, 547-548 (2017).
2. Van der Auwera, G.A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11-33 (2013).
3. Bae, S., Park, J. & Kim, J.S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473-1475 (2014).
4. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
5. van Overbeek, M. et al. DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. *Mol Cell* **63**, 633-646 (2016).