Low Incidence of Off-Target Mutations in Individual CRISPR-Cas9 and TALEN Targeted Human Stem Cell Clones Detected by Whole-Genome Sequencing

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters.

Citation

Published Version
doi:10.1016/j.stem.2014.04.020

Accessed
April 15, 2017 9:21:51 AM EDT

Citable Link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:20481140

Terms of Use
This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP

(Article begins on next page)
Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing

Adrian Veres¹², Bridget S. Gosis¹, Qiurong Ding¹, Ryan Collins³, Ashok Ragavendran³, Harrison Brand³, Serkan Erdin³, Chad A. Cowan¹²⁵, Michael E. Talkowski³⁴⁵ & Kiran Musunuru¹²⁵⁶

1. Department of Stem Cell and Regenerative Biology, Harvard University, and Harvard Stem Cell Institute, Cambridge, Massachusetts 02138, USA

2. Department of Medicine, Harvard Medical School, Boston, MA 02115, USA

3. Molecular Neurogenetics Unit, Psychiatric and Neurodevelopmental Genetics Unit, Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA 02114, USA

4. Department of Neurology, Harvard Medical School, Boston, MA 02115, USA

5. Broad Institute, Cambridge, Massachusetts 02142, USA

6. Division of Cardiovascular Medicine, Brigham and Women’s Hospital, Boston, MA 02115, USA

Corresponding Author

Kiran Musunuru, M.D., Ph.D., M.P.H.

Harvard University

Sherman Fairchild Biochemistry Bldg 160

7 Divinity Ave

Cambridge, MA 02138, USA

E-mail: kiranmusunuru@gmail.com / Phone: (617) 496-5361 / Fax (617) 496-8351

Running Title: Whole-genome sequencing of genome-edited clones
SUMMARY

Genome editing has attracted wide interest for the generation of cellular models of disease using human pluripotent stem cells and other cell types. CRISPR-Cas systems and TALENs can target desired genomic sites with high efficiency in human cells, but recent publications have led to concern about the extent to which these tools may cause off-target mutagenic effects that could potentially confound disease-modeling studies. Using CRISPR-Cas9 and TALEN targeted human pluripotent stem cell clones, we performed whole-genome sequencing at high coverage to assess the degree of mutagenesis across the entire genome. In both types of clones, we found that off-target mutations attributable to the nucleases were very rare. From this analysis, we suggest that while some cell types may be at risk for off-target mutations, the incidence of such effects in human pluripotent stem cells may be sufficiently low to not be a significant concern for disease modeling and other applications.

Clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated (Cas) systems and transcription activator-like effector nucleases (TALENs) are recently developed genome-editing tools that target desired genomic sites in mammalian cells (Miller et al., 2011; Hockemeyer et al., 2011; Cong et al., 2013; Mali et al., 2013b; Cho et al., 2013; Jinek et al., 2013). The most commonly employed CRISPR-Cas system, derived from Streptococcus pyogenes, uses Cas9 nuclease that complexes with a guide RNA that hybridizes a 20-nucleotide DNA sequence (protospacer) immediately preceding an NGG motif (protospacer-associated motif, or PAM), resulting in a double-strand break (DSB) three basepairs (bp) upstream of the NGG (Jinek et al., 2012). TALENs bind as a pair on sequences surrounding a genomic site, positioning a dimer of FokI nuclease domains to generate a DSB at the site. The introduction of a DSB at a specified genomic site allows for modification of the site via either non-homologous end joining (NHEJ), which typically introduces an insertion or deletion (indel), or homology-
directed repair (HDR), which can be exploited to knock in a point mutation or insert a desired sequence at the site.

One important application of genome-editing technology is disease modeling (Musunuru, 2013). The ability to generate isogenic wild-type and mutant clones for phenotypic comparison would enable rigorous functional genetic studies. However, both CRISPR-Cas9 and TALENs have been demonstrated to produce off-target effects, i.e., mutagenesis at sites in the genome other than the desired on-target site (Hockemeyer et al., 2011; Mussolino et al., 2011; Fu et al., 2013; Hsu et al., 2013; Mali et al., 2013a; Pattanayak et al., 2013; Cradick et al., 2013; Cho et al., 2014). These studies have largely focused on sites with high sequence similarity to the on-target site and have documented mutagenesis rates as high as 77% for CRISPR-Cas9 and 1% for TALENs at individual off-target sites. Relatively unexplored is whether CRISPR-Cas9 or TALENs produce off-target effects at sites with low sequence similarity to the on-target site. Although the nucleases might have poor affinity and have a low probability of generating a mutation at any given single site in the genome, they might nonetheless generate a sizeable number of nonspecific mutations across the billions of basepairs of the genome in any single cell. This would significantly confound the validity of disease-modeling studies that rely upon genome-edited clones.

To date, most studies of nuclease off-target effects have been performed in aggregated pools of transformed or immortalized cultured human cells, such as HEK 293T and K562 cells, that are not well suited for disease modeling. We therefore decided to study nuclease off-target effects generated in a “real-world” application of genome editing, centered on human pluripotent stem cell (hPSC) clones being actively used for biological studies (e.g., Ding et al., 2013a).

We assessed the degree of genome-wide off-target mutagenesis in hPSC clones targeted with either CRISPR-Cas9 or TALENs. We performed whole-genome sequencing at high coverage
(60× target coverage) of ten cell lines, including nine clones we had previously generated with genome editing (Ding et al., 2013a; Ding et al., 2013b) (Figure 1): the human embryonic stem cell line HUES 9; three HUES 9 clones exposed to TALENs targeting the SORT1 gene, with one clone remaining wild-type in both alleles (clone A) and two clones bearing indels in both SORT1 alleles (clones B and C); three HUES 9 clones exposed to CRISPR-Cas9 targeting the same site in the SORT1 gene, with one wild-type clone (clone D) and two clones bearing indels in both SORT1 alleles (clones E and F); and three HUES 9 clones exposed to CRISPR-Cas9 targeting the LINC00116 gene, with one wild-type clone (clone G) and two clones bearing indels in both LINC00116 alleles (clones H and I). All of the HUES 9 clones were derived from the same stock of parental HUES 9 cells. Of note, we had found the targeting efficiency of the SORT1 TALENs to be 11%, in contrast to CRISPR-Cas9 for SORT1, which was 76%; the targeting efficiency of CRISPR-Cas9 for LINC00116 was 57% (Ding et al., 2013b).

Upon obtaining the whole-genome sequencing data, we assessed the clones for small indels, single nucleotide variants (SNVs), and structural variants (SVs), which include chromosomal inversions, rearrangements, duplications, and deletions (Supplemental Experimental Procedures). We largely focused on the identification of small indels and SVs because they comprise virtually all of the mutations introduced by NHEJ. After filtering for the small indels most likely to be true positives and to be potential off-target mutations (rather than mutations that arose in the parental cell pool) and confirmation with Sanger sequencing, we identified a total of 28 such indels across the nine experimental clones, compared against the parental HUES 9 cells as the reference. Of note, all of the previously known on-target indels (seven in total) were correctly identified by the whole-genome sequencing and filtering (Table 1 and Table S1). One of the 28 off-target indels was a frameshift in the coding sequence of ZDHHC11 (in clone I). None of the other indels lay in either the coding sequence of a gene or the expressed sequence of an annotated non-coding RNA.
None of the indels in CRISPR-Cas9 clones were within 100 nucleotides of a potential off-target site as predicted by sequence similarity—up to six mismatches—with the on-target site, and none lay near sequences that matched the on-target sites better than would be expected by chance (Figure S1). Moreover, none of the indels lay within 100 nucleotides of a sequence perfectly matching the last ten nucleotides of the protospacer with an adjacent PAM site [NGG as well as NAG, which has also been shown to be tolerated (Hsu et al., 2013; Pattanayak et al., 2013)]. Furthermore, we paid special attention to the indels that lay within five bases upstream of a potential PAM site (Table S1), where CRISPR-Cas9-mediated DSBs would be expected to occur. Although the majority of clones had a potential PAM site, none of the adjacent sequences matched the on-target site better than would be expected by chance (Figure S1).

One of the indels in a TALEN clone was located between two potential off-target binding sites as predicted by sequence similarity with the on-target sites—one with three mismatches, and the other with four mismatches—with the binding sites being 17 bp apart, within the optimal range for generating a DSB with TALENs of this type (Ding et al., 2013a) (Figure 1A). None of the other TALEN clone indels were optimally positioned near a pair of degenerate TALEN binding sites (up to five mismatches with the on-target site), and none lay near sequences that matched the on-target sites better than would be expected by chance (Figure S1).

None of the SVs and SNVs that passed our filtering criteria in CRISPR-Cas9 clones was within 100 nucleotides of a predicted off-target site. None of the variants in TALEN clones were optimally positioned near a pair of degenerate TALEN binding sites. We detected 894 unique SNVs across the nine clones (average of 100 per clone) compared to the parental HUES 9 cell line (Table 1). The SV analysis revealed two structural variants unique to an individual clone: a 5.5-kb deletion on chromosome 6 in clone F and a 261-bp segment of chromosome 4 inserted within the LINC00116 CRISPR-Cas9 on-target site on chromosome 2 in clone H (Table 1 and Table S2). Sanger sequencing confirmed that both alleles of the chromosome 4 region were
intact in clone H, signifying a duplicated insertion into the chromosome 2 on-target site rather than a balanced translocation. We speculate that due to microhomology, the chromosome 4 region was used as a repair template for a DSB at the on-target site.

Just one of the detected variants—a TALEN clone indel—seems certain to be a nuclease-mediated off-target effect. It is probable that some if not all of the other indels/SVs reflect clonal heterogeneity within the original stock of HUES 9 cells. Previous studies have documented mutagenesis occurring during the derivation and expansion of hPSCs (Hussein et al., 2011; Gore et al., 2011; Howden et al., 2011; Yusa et al., 2011). Furthermore, each clone harbored a sizable number of unique SNVs, which would not be predicted to result from NHEJ. Nonetheless, with a maximum of just two to five confirmed events in each individual clone, our results suggest that nuclease-mediated off-target effects of CRISPR-Cas9 and TALENs do not intrinsically cause a large degree of indiscriminate, nonspecific mutagenesis across the genome.

We note the limitations of this study. Even with whole-genome sequencing at high coverage, it is likely that some variants in the clones were not detected given the limitations of short-read sequencing. The small number of sequenced clones targeted at just two loci prevents generalization to all hPSC clones targeted with any CRISPR-Cas9 or TALENs of any configuration by any methodology. Furthermore, our results are not relevant to therapeutic applications targeting up to millions of cells at a time, where rare events may have deleterious consequences.

We do note that clonal heterogeneity may represent a more serious obstacle to the generation of truly isogenic cell lines than nuclease-mediated off-target effects, since each of our clones harbored a very small number of unique indels and SVs (two to five) compared to a relatively larger number of unique SNVs (average of 100) that likely arose spontaneously in culture. This suggests that even if one had in hand a genome-editing tool with perfect specificity, targeted
clones would still be likely to harbor some differences elsewhere in the genome. Rigorous studies will require whole-genome sequencing of the clones used for experiments to fully characterize their mutational profiles, or they will need to include multiple clones for each experimental condition to ensure that potential confounding by any single off-target mutation in a clone is minimized.

ACKNOWLEDGMENTS

This work was supported in part by the Harvard Presidential Scholars Fund of the Harvard Medical School MD/PhD Program (A.V.); grants R00-HL098364 (K.M.), R01-HL118744 (K.M.), R01-DK097768 (K.M.), and R00-MH095867 (M.E.T.) from the United States National Institutes of Health (NIH); the Harvard Stem Cell Institute (K.M.); and Harvard University (Q.D., K.M.). We thank Vamsee Pillalamarri, Carrie Hanscom, and the staffs of the MGH Genomics and Technology Core and NextGen Sequencing Core for technical assistance.

REFERENCES


FIGURE LEGENDS

Figure 1. On-target and Off-target Mutations

(A) HUES 9 clones targeted in the SORT1 gene with TALENs or CRISPR-Cas9.

(B) HUES 9 clones targeted in the LINC00116 gene with CRISPR-Cas9. For TALEN targeted clones, the boxes indicate the TALEN on-target and off-target binding sequences. For CRISPR-Cas9 targeted clones, the boxes indicate the 20-bp sequence matching the protospacer and the 3-bp PAM. For the on-target sites, deletions and insertions in the two alleles of each clone are indicated. For the off-target site, the mismatches with the TALEN on-target binding sequences are indicated in bold, and the deletion in one allele of the clone is indicated.
Figure 1

a

5’

\[ \text{targeting exon 3} \]

SORT1

3’

**TALEN on-target site** - chr1:109910022-109910068 (reverse)

- **HuES 9**
  - GGTAAATTATGACCTTTGGAGACATCGCTTAATCGAGATAGGTAGCTG
  - CCAAATCTGAAAACCTGTCTAGGAGTAGCTCCCCTTAG
  - wild-type

- **Clone B**
  - GGTAAATTATGACCTTTGGAGAC---C-AGCTATATCGAGATAGGTAGCTG
  - 5bp del
  - GGTAAATTATGACCTTTGGAGAC---CAAGCTATATCGAGATAGGTAGCTG
  - 4bp del

- **Clone C**
  - GGTAAATTATGACCTTTGGAGACATCGCTTAATCGAGATAGGTAGCTG
  - 1bp del
  - GGTAAATTATGACCTTTGGAGAC---TCCAAGCTATATCGAGATAGGTAGCTG
  - 5bp del

**TALEN off-target site** - chr4:126910465-126910511

- **HuES 9**
  - CATACTTGCACTTTGGAGATACAGACCAACAGATATGAGTCTGTTATCT
  - CTTAGCAACCTGAAAACTGTCTTAGTCTTCTCTTTACTAACACAAATAG
  - wild-type

- **Clone B**
  - GATACTTGCACTTTGGAGATACAGACCAACAGATATGAGTCTGTTATCT
  - 16bp del

**CRISPR on-target site** - chr1:109910022-109910068 (reverse)

- **HuES 9**
  - GGTAAATTATGACCTTTGGAGACATCGCTTAATCGAGATAGGTAGCTG
  - CCAAATCTGAAAACCTGTCTAGGAGTAGCTCCCCTTAG
  - wild-type

- **Clones E, F**
  - GGTAAATTATGACCTTTGGAGACATCGCTTAATCGAGATAGGTAGCTG
  - 1bp del
  - GGTAAATTATGACCTTTGGAGACATCGCTTAATCGAGATAGGTAGCTG
  - 1bp del

b

5’

\[ \text{targeting exon 7} \]

LINCO00116

3’

**CRISPR on-target site** - chr2:110970033-110970079 (reverse)

- **HuES 9**
  - ACTGCAAGTTGTCCGCTCTGATCTAGGCTTTGCTCTCTGAGTACTCTTGG
  - TGCAGCTCAACAGCCACAGGCTCATCGGAAAGGAAACCCTCATGAGGACC
  - wild-type

- **Clone H**
  - ACTGCAAGTTGTCCGCTCTGATCTAGGCTTTGCTCTCTGAGTACTCTTGG
  - 1bp ins
  - ACTGCAAGTTGTCCGCTCTGATCTAGGCTTTGCTCTCTGAGTACTCTTGG
  - 262bp ins

- **Clone I**
  - ACTGCAAGTTGTCCGCTCTGATCTAGGCTTTGCTCTCTGAGTACTCTTGG
  - 5bp del
  - ACTGCAAGTTGTCCGCTCTGATCTAGGCTTTGCTCTCTGAGTACTCTTGG
  - 5bp del
Table 1. Numbers of Unique On-target and Candidate Off-target Indels and Structural Variants (SVs), As Well As Unique Single Nucleotide Variants (SNVs), in TALEN and CRISPR-Cas9 Targeted Clones

<table>
<thead>
<tr>
<th>clones</th>
<th>SORT1 TALENs</th>
<th>SORT1 CRISPR-Cas9</th>
<th>LINC00116 CRISPR-Cas9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>on-target indels</td>
<td>—</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>on-target SVs</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>likely off-target indel</td>
<td>—</td>
<td>1</td>
<td>—</td>
</tr>
<tr>
<td>other candidate off-target indels</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>candidate off-target SVs</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>SNVs</td>
<td>64</td>
<td>115</td>
<td>142</td>
</tr>
</tbody>
</table>

See also Figure S1 and Tables S1–S3.

\(^a\) Homozygous for indels.

\(^b\) 261 bp duplicated insertion.
Figure S1, Related to Table 1. Minimal Edit Distances Near Indels

In each panel, the blue shaded area represents the distribution of minimal edit distances for each nuclease’s subset of 381 filtered indels, and the black line represents the expected distribution inferred from 50,000 randomly sampled parental HUES 9 line indels. Top row, minimal edit distances across a 100-bp window around a given indel. Bottom row, minimal edit distances across a 100-bp window, retaining only sequences that end in an NGG or NAG PAM sequence.
Table S1, Related to Table 1. Unique Indels Detected by Whole-genome Sequencing

<table>
<thead>
<tr>
<th>Chr</th>
<th>Pos</th>
<th>5' flanking sequence</th>
<th>Ref</th>
<th>Alternate</th>
<th>3' flanking sequence</th>
<th>SORT1 TALENs</th>
<th>SORT1 CRISPR-Cas9</th>
<th>LINC00116 CRISPR-Cas9</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>1</td>
<td>81390259</td>
<td>ATAGCTGATTAGAGATATTGCC</td>
<td>CATATGGTG</td>
<td>C</td>
<td>TCTGAGATCTTATTAGTG</td>
<td></td>
<td></td>
<td>25,11</td>
</tr>
<tr>
<td>1</td>
<td>109910034</td>
<td>GTGTTATGATTCCTCAGCTTC</td>
<td>G</td>
<td>GA</td>
<td>ATATAGCTTGGAGCTTGCAA</td>
<td>0,13</td>
<td>1,54</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>109910041</td>
<td>ATGCTCATCTCCAGATATCG</td>
<td>TTGGACT</td>
<td>TG</td>
<td>GTCGAAATGCTCATATTAC</td>
<td>9,4*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>109910043</td>
<td>ATCTCAACTCTGGAATATCG</td>
<td>TG</td>
<td>T</td>
<td>GACTGTCGAAAAGCTGATAT</td>
<td>7,17*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>109910044</td>
<td>TCTGACCTTCTGACTGATATCG</td>
<td>GGACT</td>
<td>G</td>
<td>GTCGAAATGCTCATATTAC</td>
<td>5,8*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>109910046</td>
<td>TCACTGTGATGACTTGGGTTG</td>
<td>ACTGTG</td>
<td>A</td>
<td>CAAAGTCTAATTACCAGT</td>
<td>20,8*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>170931870</td>
<td>AATTCTGCAATTTTTGAGC</td>
<td>TG</td>
<td>T</td>
<td>TTTTTTTATTATATATTAAA</td>
<td>8,17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9075788</td>
<td>CGAGAGCTTCAAGAGAGAAAGGC</td>
<td>AGCAAGCCTGCGGTT</td>
<td>A</td>
<td>AATTAAGCCCTAAACGGA</td>
<td>8,5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>76409962</td>
<td>AACAAGTCAGGTAGATACGC</td>
<td>TG</td>
<td>T</td>
<td>TGGCAATTTCTCAAAATTGT</td>
<td>14,7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>110970046</td>
<td>TGCAGCGCAGAGATACGATCG</td>
<td>AGGAAG</td>
<td>A</td>
<td>GAGGCTACTAGACGGAGAA</td>
<td>4,58</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>110970047</td>
<td>CAGCCAGCAGGATACACGAGA</td>
<td>A</td>
<td>AG</td>
<td>GCGAGGCTACTAGACGGGA</td>
<td>1,34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>136347377</td>
<td>GACCTTGAGATAATATCTGT</td>
<td>CACATGATATATATATTGCG</td>
<td>C</td>
<td>ACATATATATAAAATATG</td>
<td>12,10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>13837485</td>
<td>TGTACTTACCCACAGATTTT</td>
<td>TC</td>
<td>T</td>
<td>CCTCTGCGTCAAGACCAT</td>
<td>16,11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>19570857</td>
<td>CCACACAGTTCAAGGACACA</td>
<td>C</td>
<td>CAT</td>
<td>GAAGCCAGCCCTTCCGGGC</td>
<td>5,3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>65647862</td>
<td>TATTAGTGAACAAACATGC</td>
<td>TG</td>
<td>T</td>
<td>TTGAGTATTGATTATTG</td>
<td>11,12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>67241317</td>
<td>TGACAATCTAACTTTTTA</td>
<td>G</td>
<td>GCTT</td>
<td>CCTCTATGCACACAGAAAAG</td>
<td>11,10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>117364346</td>
<td>TTAAAAATTTTATTACTGAG</td>
<td>GAATTATTTT</td>
<td>G</td>
<td>TATATTTTATCTTGAATAT</td>
<td>9,5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>126910474</td>
<td>TTAACAGAGAAAGATACCTTG</td>
<td>GACTTTGGGTGATACAGA</td>
<td>G</td>
<td>ACAAGGATTAGATTGTGTA</td>
<td>8,4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>850612</td>
<td>CTTGAGAATGTGACGGGAGTA</td>
<td>AC</td>
<td>A</td>
<td>GACCAAGCTTACCTGCAGA</td>
<td>43,15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>32712655</td>
<td>CCAAGCCAGGTGAGAGGTTG</td>
<td>A</td>
<td>AG</td>
<td>GCAAGGCGCGCTGCCGCGGC</td>
<td>30,35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>156507271</td>
<td>TTTAAAGACCTGATGTTTTAG</td>
<td>TA</td>
<td>T</td>
<td>ACTAATGTGGCTGCTTTAG</td>
<td>13,12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>23212217</td>
<td>GTGTTGGAAGAAAGAAGAAAAAAGA</td>
<td>ATTAT</td>
<td>A</td>
<td>TTATCTTGGAACATCTCC</td>
<td>19,16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>127961691</td>
<td>ATTTGTGCTTGATCTGCGAA</td>
<td>A</td>
<td>AAG</td>
<td>AAAAAAAAATTAAAGAGAA</td>
<td>4,5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>132161273</td>
<td>AAATCCCTCTAAAATTATTGT</td>
<td>TCTAC</td>
<td>TGAAGGGATTTA</td>
<td>ATTTTCTAACACTAAATTGT</td>
<td>6,11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>33682338</td>
<td>TGACTGTGAAAGATCTTGT</td>
<td>GAA</td>
<td>G</td>
<td>ATTCCTATTTCTGAGTTGCG</td>
<td>18,14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30327724</td>
<td>ATACAATGAGTGAAGAGG</td>
<td>G</td>
<td>GTA</td>
<td>TAAATGTGGCTGCTTTAG</td>
<td>34,7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>82813787</td>
<td>TAGAGAACAGAAGGAGCTG</td>
<td>GTCTC</td>
<td>G</td>
<td>TAGAGAGATGATTTGTCG</td>
<td>18,25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>131460181</td>
<td>TGCTACTCCTCTTCTTCTAGG</td>
<td>CT</td>
<td>C</td>
<td>TGCTGTCATCTACCTTCT</td>
<td>33,8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>852647688</td>
<td>CCTCTACAAAGTTATTCTT</td>
<td>CTCT</td>
<td>C</td>
<td>TCTAGTTTACCAAAATATAGA</td>
<td>13,15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>30161597</td>
<td>CTGTTAGAGCCTCCTCTATGAG</td>
<td>CCTGTAGATGCGAGATACTGACAT</td>
<td>C</td>
<td>CAATCTCTGCTTTGAGACT</td>
<td>34,12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>71581221</td>
<td>AAAAATCATGAGATTCACAGA</td>
<td>CATATT</td>
<td>C</td>
<td>ATATCGAAAGAAACATCCAT</td>
<td>11,3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>66143591</td>
<td>AAAATATCAAGAAAGAAT</td>
<td>TTATA</td>
<td>T</td>
<td>TTGTGATTGAGCTTAGTAC</td>
<td>12,6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>50154383</td>
<td>GCCGCTGACTTCTGAGTAGT</td>
<td>TG</td>
<td>T</td>
<td>GGGGGCTCTTGGAGGATTGG</td>
<td>29,11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>39827854</td>
<td>CTGGTGACAGGACAAACACA</td>
<td>C</td>
<td>CAAAT</td>
<td>AAATACCTGAGATAATTAT</td>
<td>21,6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>28415750</td>
<td>TGTGCCCTGGAGCTAGCTGCTC</td>
<td>TTTGCG</td>
<td>T</td>
<td>TCAGGCTCCTGGGCTGCTGTA</td>
<td>50,35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Indels at on-target sites are indicated in red bold. A likely nuclease-mediated off-target indel is indicated in blue bold. An indel that lies in the coding sequence of ZDHHC11 is indicated in magenta bold. Underlines indicate potential PAMs (NGG or NAG) within five bases upstream of the indel. Columns A–I indicate reference allele counts (x) and alternate allele counts (y) in the format (x,y) for any of the clones A–I in which the called genotype included at least one copy of the alternate allele. Note: the reference allele counts for the on-target indel in clones B and C marked in red bold are not indicated wild-type alleles—rather, due to the nature of the calling algorithm, they indicate counts of alleles that do not match the alternate alleles; because clones B and C are compound heterozygotes, the reference allele counts actually represent the indels on the other alleles. All on-target and off-target indels in this Table were confirmed with Sanger sequencing.
## Table S2, Related to Table 1. Unique Structural Variants (SVs) Detected by Whole-genome Sequencing

<table>
<thead>
<tr>
<th>SV class</th>
<th>Clone</th>
<th>Treatment group</th>
<th>Chr</th>
<th>Start</th>
<th>End</th>
<th>Size</th>
<th>Split-read consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Translocation*</td>
<td>H LINC00116 CRISPR-Cas9</td>
<td>4 183998625   183998885</td>
<td>261 bp</td>
<td>chr2</td>
<td>chr4: CTGCCAGCCAGGAGTACTCCAGAA</td>
<td>CTTACATTGCTGTTGATTCTGG</td>
<td></td>
</tr>
<tr>
<td>Deletion</td>
<td>F SORT1 CRISPR-Cas9</td>
<td>6 18754456  18760023</td>
<td>5568 bp</td>
<td>chr4</td>
<td>chr2: CTGGAGGATCCCTTGAGCACAGGAGT</td>
<td>GCGAAGGCTACTAGCAGGAACAACT</td>
<td></td>
</tr>
</tbody>
</table>

A structural variant at an on-target site is indicated in red bold. Split-read consensus sequences represent the consensus of all split reads that span the breakpoint. Vertical lines denote precise breakpoint positions.

* Translocation represents a duplicated insertion from chromosome 4 that inserted into the on-target chromosome 2 site. The underlined base in the split-read consensus indicates an additional inserted basepair at the breakpoint.

* As no split-reads spanned this event, there is not precise refinement of the coordinates.
<table>
<thead>
<tr>
<th>Table S3, Related to Table 1. Numbers of Indels and Single Nucleotide Variants (SNVs) Detected by Whole-genome Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>HUES 9 A</td>
</tr>
<tr>
<td>Raw indel calls</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Post low-complexity filter</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Post homopolymeric filter</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>De novo, nuclease-specific indels</td>
</tr>
<tr>
<td>Sample-specific indels</td>
</tr>
<tr>
<td>Called reads in only one sample, &gt; 2 reads</td>
</tr>
<tr>
<td>Confirmed by Sanger sequencing</td>
</tr>
<tr>
<td>HUES 9 A</td>
</tr>
<tr>
<td>Raw SNV calls</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Post low-complexity filter</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>De novo, nuclease-specific SNVs</td>
</tr>
<tr>
<td>Sample-specific SNVs</td>
</tr>
<tr>
<td>Called reads in only one sample, &gt; 2 reads</td>
</tr>
</tbody>
</table>

a Calls in the parental HUES 9 line that were not called in the individual clone.

b Not included in this tally is an on-target indel shared by clones E and F.
SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Generation of TALEN targeted and CRISPR-Cas9 targeted clones

The targeted clones were generated as previously described (Ding et al., 2013a; Ding et al., 2013b; Peters et al., 2013). We summarize the methods below. The CRISPR-Cas9 and TALENs used were chosen for the proximity of their predicted binding sites to the desired target sites in the genes, and the TALENs were designed as an obligate heterodimer. For each CRISPR-Cas9, there were no sequences elsewhere in the genome with up to two mismatches with the 20-nucleotide target site.

TALEN genomic binding sites in SORT1 were chosen to be 15 bp in length such that the target sequence between the two binding sites was between 14 and 18 bp in length; each binding site was anchored by a preceding T base in position “0” as has been shown to be optimal for naturally occurring TAL proteins. We generated full-length TALENs harboring, in order: a N-terminal FLAG tag, a nuclear localization signal, the N-terminal portion of the TALE PthXo1 from the rice pathogen X. oryzae pv. oryzae lacking the first 176 amino acids, the engineered TAL repeat array, the following 63 amino acids from the corresponding C-terminal portion of PthXo1 and one of two enhanced FokI domains. The FokI domains used were obligate heterodimers with both the Sharkey and ELD:KKR mutations to enhance cleavage activity, engineered by PCR. Each TALEN was in a plasmid with the CAG promoter for optimal expression in human pluripotent stem cells, with the TALEN being coexpressed with a fluorescent marker [enhanced green fluorescent protein (EGFP), mCherry (Clontech), or turbo red fluorescent protein (tRFP; Evrogen)] via an intervening viral 2A sequence. For CRISPR-Cas9, we subcloned a human codon-optimized Cas9 gene with a C-terminal nuclear localization signal into the same CAG expression plasmid with EGFP, and we separately expressed the guide
RNA (gRNA) from a plasmid with the human U6 polymerase III promoter. The 20-nucleotide protospacer sequence for each gRNA was introduced using polymerase chain reaction (PCR)-based methods. The reagents used to generate these various TALEN and CRISPR-Cas9 plasmids are available through Addgene (https://www.addgene.org/talen/musunuru/ and https://www.addgene.org/crispr/musunuru/).

HUES 9 cells were grown in feeder-free adherent culture in chemically defined mTeSR1 (STEMCELL Technologies) supplemented with penicillin/streptomycin on plates pre-coated with Geltrex matrix (Invitrogen). The cells were disassociated into single cells with Accutase (Invitrogen), and 10 million cells were electroporated with 50 μg of the TALEN pair (25 μg of each plasmid) or CRISPR-Cas9 (25 μg of each plasmid) in a single cuvette and replated. The cells were collected from the culture plates 48 to 72 hours post-transfection or post-electroporation (at which point fluorescent marker expression was in decline) by Accutase treatment and resuspended in PBS. Cells expressing green and/or red fluorescent markers were collected by FACS (FACSAria II; BD Biosciences) and replated on 10-cm tissue culture plates at 15,000 cells/plate to allow for recovery in growth media.

Post-FACS, the cells were allowed to recover for 7-10 days, after which single colonies were manually picked and dispersed and replated individually to wells of 96-well plates. Colonies were allowed to grow to near confluence over the next 7 days, at which point they were split using Accutase and replica-plated to create a working stock and a frozen stock. The working stock was grown to confluence, and genomic DNA was extracted in 96-well format, followed by PCR amplification around the target site and Sanger sequencing to identify both untargeted and targeted clones. Chosen clones were expanded further for extraction of genomic DNA for whole-genome sequencing, with ~7 passages occurring between the single-cell cloning and the DNA extraction.
Identification of novel indels, single nucleotide variants, and structural variants

Genomic DNA from all ten cell lines (parental HUES 9 line, clones A–I) was extracted using the DNeasy Tissue Kit (QIAGEN) and subjected to quality assessment. The extracted DNA was sequenced as paired-end 101-nucleotide reads to a target of 60× haploid coverage on an Illumina HiSeq2000 sequencer as previously described (Stransky et al., 2011). These mate-pair libraries featured an average median fragment insert size of 329 bp and a standard deviation of 47 bp. The pair-ends reads were aligned onto the hg19 (GRCh37v. 71) human reference genome using Bowtie 2 and manipulated (deduplication, sorting, indexing) using Picard Tools, version 1.84 (http://picard.sourceforge.net). The reads have been uploaded to the NCBI Short Read Archive (SRA) and are available via the accession number SRP039576.

The Genome Analysis ToolKit, version 2.6 (McKenna et al., 2010), was used for local realignment around indels (RealignerTargetCreator, IndelRealigner), base score recalibration (BaseRecalibrator), variant calling across the ten samples (HaplotypeCaller) and variant score recalibration (VariantRecalibrator, ApplyRecalibration). Candidate indels (totalling 948,344 calls) were filtered on several criteria using Python and the PyVCF, version 0.6.0, and PyFasta, version 0.5.0, packages. First, we removed indels near low-complexity regions as defined by RepeatMasker and annotated by softmasking in hg19). Indels were considered “near” low complexity regions if any position within 10 bp or at least one third of positions within 50 bp were masked by RepeatMasker. Second, we removed indels that caused expansions or compressions of long (>6 bp) homopolymers. The effects of these filters are detailed in Table S3. By comparing indels calls in the parental HUES 9 cell line to calls for each of the clones, we can estimate false-negative rates of 4%-6% (in raw indel calls) and ~1% (after these two filters). Considering only indels that (1) were absent in the parental HUES 9 cell line and (2) were not called in samples that were treated with different nucleases (TALENs for SORT1, CRISPR-Cas9 for SORT1, CRISPR-Cas9 for LINC00116), we produced a set of 381 indels used in further
analyses. Among these 381 indels were seven on-target indels already known to be in the targeted clones via Sanger sequencing (Table S1).

We further filtered the 381 indels to identify those most likely to represent nuclease-mediated off-target effects by: (1) retaining indels for which there were called alternate alleles in only one sample, since indels generated by engineered nucleases at a given locus are extremely heterogeneous with respect to length and sequence, and it is unlikely that two independent clones would have suffered exactly the same indel at the same off-target site; and (2) retaining indels with the alternate allele present in more than two reads. This yielded a total of 53 indels. We then performed polymerase chain reaction (PCR) amplification and Sanger sequencing to confirm or refute these indels. This yielded a final list of 35 indels, including the seven on-target alleles (Table S1). Thus, at this final stage the false positive rate was 34%.

We searched the human genome for sites likely to exhibit off-target activity based on similarity to nuclease target sites. For CRISPR-Cas9, we considered two types of similar sequences: (1) any sequence within 6 (or fewer) substitutions of the 20-nt target site followed by an NRG PAM sequence and (2) any sequence matching the last 10 nt of the target site followed by an NRG PAM. Using Bowtie 1, we mapped these sequences to 14,200 and 10,935 loci of high similarity relative to the on-target SORT1 and LINC00116 sequences. Of note, by intentional design of the CRISPR-Cas9 on-target sites, there were no loci within 2 substitutions of the 20-nt target site. Except for the on-target indels, none of these genomic loci were within 100 bp of indels called in the respective samples.

For TALENs, we constructed a list of all sequences within 5 (or fewer) substitutions of either monomer’s on-target site and identified 12,301,606 genomic loci matching these sequences. We manually reviewed 142 indels occurring within 100 bp of these loci. We also identified 55,503 pairs of off-target binding sites facing each other (i.e., oriented towards each other on opposite
strands) and separated by a distance of 10-22 bp. Besides the on-target indels, only one indel occurred between the pair’s binding sites, likely representing a bona fide off-target effect.

We expanded our search to nearby off-target sites with any possible number of mismatches relative to the target sequences. We searched 100-bp windows around each indel for the sequence most closely matching the on-target site and recorded the number of mismatches of that sequence. We refer to this number as the minimal edit distance of the region near an indel. To prevent double counting, we merged the windows of indels within 100 bp of each other. For CRISPR-Cas9, we allowed for both NGG and NAG PAM sequences when counting mismatches. For TALENs, we considered every pair of sequences in the window regardless of the distance separating them. We computed minimal edit distances for the 381 indels (Figure S1, blue areas) and compared each nuclease’s distribution to background distributions determined by the minimal edit distances of 50,000 randomly chosen parental HUES 9 line indels that passed low-complexity and homopolymer filters (Figure S1, black lines). The only outliers we observed were the on-target events (minimal edit distance of 0) and the single TALEN off-target event (minimal edit distance of 7).

Candidate SNVs (totalling 3,776,763 calls) were filtered using criteria similar to the indels (Table S3). We removed SNVs near low-complexity regions and considered only SNVs (1) absent in the parental HUES 9 cell line and (2) not called in samples that were treated with different nucleases. Together, these filters produced a set of 1,742 SNVs. We applied the same final filters described above for indels; this resulted in a final list of 894 SNVs (Table 1). Using the same CRISPR-Cas9 and TALEN off-target analyses described above for indels, we determined that none of the SNVs lay in proximity to predicted off-target sites.

We sought to establish the structural variation (SV) architecture of each individual line and then compared the SV burden across technical approaches and in comparison to the parental HUES 9
line, including inversions, rearrangements, duplications, and deletions. All paired-end data were aligned with BWA-MEM, version 0.7.5a-r418 (Li, 2013), to GRCh37.71 using defaults with duplicate reads removed using Picard Tools. We used an integrated SV detection pipeline synthesized from four previously published algorithms: LUMPY, version 0.1.5 (Layer et al., 2012), DELLY, version 0.0.11 (Rausch et al., 2012), BAMSTAT, version 0.2 (Talkowski et al., 2011; Talkowski et al., 2012; Chiang et al., 2012), and CNVnator, version 0.2.7 (Abyzov et al., 2011). The principal branch of the pipeline generated a preliminary SV set by intersecting paired-end evidence from DELLY-PE and BAMSTAT with consensus split read call-sets derived from LUMPY-SR and DELLY-SR. These calls were further screened for high-confidence using mapping quality (MapQ ≥ 20) and a minimum event size equal to the mean insert size plus six times the insert size standard deviation for each that particular library (ranging from 754 bp to 866 bp; library-dependant).

Following initial filtering, we performed in silico PCR validation of split-reads supporting the event and filtered all SVs against established reference artifacts and unplaced contigs from ongoing studies in our laboratory and others (M. Talkowski, unpublished data). An analogous branch of the SV detection pipeline further supplemented these SV calls with a genome-wide focal read-depth analysis (CNVnator) and ancillary anomalous mate-pair clustering (DELLY-PE) to capture de novo CNVs. We generated a list of candidate CNVs across all libraries that passed CNVnator’s hardcoded e-value filter. We further filtered these candidate CNVs for high confidence based on CNV size, normalized read depth, and proportion of reads within the putative CNV with mapping quality ≥ 0, as consistent with CNVnator’s recommended filtering criteria. Finally, we refined these CNV calls with concordant evidence of anomalous paired-end support from DELLY-PE. As with indels, we focused on SVs and CNVs that were unique to individual clones.
Of note, we identified a pericentric inversion of chromosome 9 [inv(9)(p11.2q13)] in our consensus call set that was consistent with a previously annotated pericentric inv(9) in the HUES 9 cell line, thought to be of no clinical consequence (Feuk, 2010).

SUPPLEMENTAL REFERENCES


