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 follow 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 $\geq$ 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 $\geq$ 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

Abyzov, A., Urban, A.E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. *21*, 974–984.

Chiang, C., Jacobsen, J.C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R.E., Kirby, A., Lindgren, A.M., Rudiger, S.R., et al. (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat. Genet. *44*, 390–397, S1.

Feuk, L. (2010). Inversion variants in the human genome: role in disease and genome architecture. Genome Med. *12*, 11.

Layer, R.M., Hall, I.M., and Quinlan, A.R. (2012). LUMPY: A probabilistic framework for structural variant discovery. arXiv:1210.2342v1 [q-bio.GN].

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN].

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Peters, D.T., Cowan, C.A., and Musunuru, K. (2013). Genome editing in human pluripotent stem cells. StemBook [Internet]. Cambridge (MA): Harvard Stem Cell Institute; 2008–.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Stransky, N., Egloff, A.M., Tward, A.D., Kostic, A.D., Cibulskis, K., Sivachenko, A., Kryukov, G.V., Lawrence, M.S., Sougnez, C., McKenna, A., et al. (2011). The mutational landscape of head and neck squamous cell carcinoma. Science *333*, 1157–1160.

Talkowski, M.E., Ernst, C., Heilbut, A., Chiang, C., Hanscom, C., Lindgren, A., Kirby, A., Liu, S., Muddukrishna, B., Ohsumi, T.K., et al. (2011). Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am. J. Hum. Genet. *88*, 469–481.

Talkowski, M.E., Rosenfeld, J.A., Blumenthal, I., Pillalamarri, V., Chiang, C., Heilbut, A., Ernst, C., Hanscom, C., Rossin, E., Lindgren, A.M., et al. (2012). Sequencing chromosomal abnormalities reveals neurodevelopmental loci that confer risk across diagnostic boundaries. Cell *149*, 525–537.