Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci

Samuel Lessard[a,b], Laurent Francioli[c,d], Jessica Alfoldi[c,d], Jean-Claude Tardif[a,b], Patrick T. Ellinor[d,e], Daniel G. MacArthur[c,d], Guillaume Lettre[a,b], Stuart H. Orkin[f,g,h,i,j,1], and Matthew C. Canver[f,g,h,i,1]

[a]Research Center, Montreal Heart Institute, Montréal, QC H1T 1C8, Canada; [b]Department of Medicine, Faculty of Medicine, Université de Montréal, Montréal, QC H3T 1J4, Canada; [c]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114; [d]Program in Medical and Population Genetics, Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA 02142; [e]Cardiovascular Research Center, Massachusetts General Hospital, Charlestown, MA 02129; [f]Division of Hematology/Oncology, Boston Children's Hospital, Boston, MA 02115; [g]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115; [h]Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138; [i]Department of Pediatrics, Harvard Medical School, Boston, MA 02115; and [j]Howard Hughes Medical Institute, Boston, MA 02115

Contributed by Stuart H. Orkin, November 2, 2017 (sent for review August 18, 2017; reviewed by Jacob Corn, Charles A. Gersbach, and Shengdar Tsai)

The CRISPR-Cas9 nuclease system holds enormous potential for therapeutic genome editing of a wide spectrum of diseases. Large efforts have been made to further understanding of on- and off-target activity to assist the design of CRISPR-based therapies with optimized efficacy and safety. However, current efforts have largely focused on the reference genome or the genome of cell lines to evaluate guide RNA (gRNA) efficiency, safety, and toxicity. Here, we examine the effect of human genetic variation on both on- and off-target specificity. Specifically, we utilize 7,444 whole-genome sequences to examine the effect of variants on the targeting specificity of ∼3,000 gRNAs across 30 therapeutically implicated loci. We demonstrate that human genetic variation can alter the off-target landscape genome-wide including creating and destroying protospacer adjacent motifs (PAMs). Furthermore, single-nucleotide polymorphisms (SNPs) and insertions/deletions (indels) can result in altered on-target sites and novel potent off-target sites, which can predispose patients to treatment failure and adverse effects, respectively; however, these events are rare. Taken together, these data highlight the importance of considering individual genomes for therapeutic genome-editing applications for the design and evaluation of CRISPR-based therapies to minimize risk of treatment failure and/or adverse outcomes.

CRISPR-Cas9 | off-target specificity | on-target specificity | human genetic variation | therapeutic genome editing

The clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system holds enormous potential for therapeutic genome editing to treat a wide spectrum of genetic diseases (1–8). Human CRISPR-Cas9 clinical trials have already been initiated (9, 10) and are likely to increase in number in the future. The development of successful therapies not only requires treatment efficacy but also requires that patient safety remain paramount. This requires assessing for toxicity related to reagent delivery, to the genome-editing reagents themselves, and to off-target effects. Significant progress has been made to aid in off-target prediction (11, 12) and for unbiased genome-wide off-target detection (13–19). From a therapeutic genome-editing perspective, these methods for unbiased genome-wide off-target detection are often limited by their reliance on the reference genome or the genome of the cells used for study to evaluate guide RNA (gRNA) efficiency, safety, and toxicity; however, newer methods circumvent limitations imposed by use of the reference genome through direct sequencing of target site regions to screen each individual patient (17).

Numerous efforts have been made to document human genetic variation. For example, the 1000 Genomes Project (1000G) database consists of 2,504 whole-genome sequences (WGSs) from 26 populations spanning Africa (AFR), East Asia (EAS), Europe (EUR), South Asia (SAS), and the Americas (AMR) (20). On average, individual genomes within the database deviated from the reference genome at 4.1–5.0 million sites. The majority of variants in an individual genome were common with only 1–4% of variants having a frequency <0.5%. Notably, these deviations included 2,100–2,500 structural variants per genome. The median number of variants varied across populations; however, the order of magnitude re-mained unchanged (Dataset S1). In total, across all individuals/populations studied, ∼64 million autosomal variants were identified with a frequency <0.5%, ∼12 million with a frequency between 0.5% and 5%, and ∼8 million with a frequency >5% (20).

Recent work has demonstrated the utility of considering variants when designing CRISPR genome-editing experiments (21).

Significance

CRISPR-Cas9 holds enormous potential for therapeutic genome editing. Effective therapy requires treatment to be efficient and safe with minimal toxicity. The sequence-based targeting for CRISPR systems necessitates consideration of the unique genomes for each patient targeted for therapy. We show using 7,444 whole-genome sequences that SNPs and indels can reduce on-target CRISPR activity and increase off-target potential when targeting therapeutically implicated loci; however, these occurrences are relatively rare. We further identify that differential allele frequencies among populations may result in population-specific alterations in CRISPR targeting specificity. Our findings suggest that human genetic variation should be considered in the design and evaluation of CRISPR-based therapy to minimize risk of treatment failure and/or adverse outcomes.

Author contributions: M.C.C. conceived this study; S.L., S.H.O., and M.C.C. designed this study; S.L. and M.C.C. performed research; L.F., J.A., J.-C.T., P.T.E., D.G.M., and G.L. contributed whole-genome sequencing data; S.L. and M.C.C. analyzed data; and S.L., G.L., S.H.O., and M.C.C. wrote the paper.

Reviewers: J.C., Innovative Genomics Initiative; C.A.G., Duke University; and S.T., St. Jude Children's Research Hospital.

The authors declare no conflict of interest.

Data deposition: A list of individual aggregate off-target scores for samples from the 1000 Genomes Project and their local scores and a list of individual aggregate off-target scores for samples from the French Canadian dataset and their local scores are available for download at the following link: www.mhi-humangenetics.org/en/resources.

[1]To whom correspondence may be addressed. Email: stuart_orkin@dfci.harvard.edu or matthew_canver@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1714640114/-/DCSupplemental.

This included analysis evaluating the effect of variants on gRNA te or destroy protospacer adjacent motif (PAM) sequences (21). Therapeutic genome-editing reagents will encounter a unique genome for each patient seeking treatment. Therefore, we sought to evaluate whether variants should be considered for clinical translation of CRISPR-based therapies. We hypothesized that human genetic variation may alter CRISPR-Cas9 targeting on- and off-target specificity at therapeutically implicated loci. We further hypothesized that personalized off-target events could exist that would predispose patients to treatment failure and/or adverse outcomes. Finally, we investigated whether population-specific variants would also predispose patients to treatment failure and/or adverse outcomes.

Results

Therapeutic Loci and Off-Target Score Calculation. To evaluate these hypotheses, we identified a comprehensive list of 23 human-genome and 7 viral-genome therapeutic targets based on literature mining for loci previously targeted by CRISPR-Cas9 for therapy (Table 1). These loci have been demonstrated to be amenable therapeutic targets for CRISPR-based strategies to elicit nonhomologous end joining (NHEJ) repair or homology-directed repair (HDR). A list of gRNAs was generated by designing all gRNAs targeting the indicated regions or using the gRNAs from previous studies (Dataset S2). gRNAs for both NHEJ and HDR applications were designed to include all gRNAs within the relevant exon(s) for coding region targets and ±100 bp for noncoding targets. It is important to note

Table 1. Summary of therapeutically implicated loci

Gene/virus	Target	Coordinates (hg19)	Disease	Repair	Refs.
Gene					
ALCAM	All exons	chr3:105085753–105295744	HIV-1 infection	NHEJ	63
BCL11A	Enhancer	chr2:60722309–60722472	β-Hemoglobinopathies	NHEJ	64
B2M	All exons	chr15:45003686–45010357	Hypoimmunogenic cells for transplantation	NHEJ	65
CCR5	Exon 1	chr3:46414395–46415452	HIV infection	NHEJ	65
CEP290	Intron 26	chr12:88494861–88495060	Leber's congenital amaurosis type 10	NHEJ	2, 66
CXCR4	Exon 2	chr2:136872440–136873482	HIV-1 infection	NHEJ	67
HLA-A	Exon 3	chr6:29911046–29911320	Hypoimmunogenic cells for transplantation	NHEJ	68
PCSK9	Exons 1–2	chr1:55505512–55509707	Cardiovascular disease	NHEJ	69
PDCD1	Exon 1	chr2:242800916–242800990	Tumor immunotherapy	NHEJ/HDR	70, 71
PSIP1	Exons 2, 12, 14	chr9:15468629–15510186	HIV-1 infection	NHEJ	72
TPST2	All exons	chr22:26921712–26992681	HIV-1 infection	NHEJ	63
TRAC,TRBC1, TRBC2	Exon 1	chr14:23016448–23016719; chr7:142498738–142499111; chr7:142498726–142499111	T cell immunotherapy	NHEJ	73–75
SLC35B2	All exons	chr6:44221838–44225308	HIV-1 infection	NHEJ	63
ADA	Intron 6/exon 7	chr20:43251649–43251819	Adenosine deaminase severe combined immunodeficiency (ADA-SCID)	HDR	76
ALB	Intron 1	chr4:74270125–74270832	Lysosomal storage disease, hemophilia A, B	HDR	77
CFTR	Exon 10	chr7:117199519–117199709	Cystic fibrosis	HDR	78
COL7A1	Exons 2, 3, 14, 15, 54, 117	chr3:48602217–48631981	Epidermolysis bullosa	HDR	79
CYBB	Exon 7	chrX:37658207–37658337	X-linked chronic granulomatous disease	HDR	80
DMD	Exons/intron 45–55	chrX:31533884–32250573	Duchenne's muscular dystrophy	HDR	81, 82
FANCC	Intron 4	chr9:97934216–97934415	Fanconi anemia	HDR	83
F9	Intron 1	chrX:138613012–138619169	Hemophilia B	HDR	84
FAH	Exon 8/intron 8	chr15:80464492–80464690	Hereditary tyrosinemia type I	HDR	85
HBB	Exon 1	chr11:5248162–5248251	Sickle cell disease	HDR	86, 87
IL2RG	Exon 5	chrX:70329079–70329240	X-linked severe combined immunodeficiency (X-SCID)	HDR	88
SERPINA1	Intron 4/exon 5	chr14:94844848–94845047	α-1-Antitrypsin deficiency	HDR	89
Virus					
Cytomegalovirus	Viral genome	—	Congenital defects, disease in immuno-compromised individuals	NHEJ	49
Epstein bar virus	Viral genome	—	Infectious mononucleosis, malignancies	NHEJ	49
Hepatitis B virus	Viral genome	—	Hepatitis B	NHEJ	51, 90–92
Herpes simplex virus type 1	Viral genome	—	Cold sores, keratitis	NHEJ	49
HIV-1	Viral genome (LTR)	—	HIV-1 infection	NHEJ	50
Human papilloma virus	E6–E7 oncogenes	—	Cervical carcinoma	NHEJ	93
JC virus	T antigen	—	Progressive multifocal leukoencephalopathy	NHEJ	52

that HDR efficiency decreases as a function of the distance between the variant and the double-strand break site (22). We also identified gRNAs targeting viral genomes using the same approach (Table 1). This analysis resulted in a list of 2,481 gRNAs targeting human genomic regions and 484 gRNAs targeting viral genomic regions. In addition, 128 nontargeting gRNAs were included as negative controls. Using a previously published aggregate off-target score (23, 24), we calculated aggregate off-target scores using the reference genome for all gRNAs (Fig. 1*A*). For a given gRNA, a "local" off-target score is calculated for each genomic match (from 0 to four mismatches). The summation of all local off-target scores for a given gRNA results in a genome-wide off-target score, termed "aggregate off-target score" (see *Materials and Methods* for additional details). For this off-target scoring method (range, 0–100), higher scores indicate lower off-target cleavage potential and lower scores indicate higher off-target cleavage potential. Importantly, the probability of cleavage decreases with increased number of mismatches (25). Therefore, sites with higher numbers of mismatches, such as three or four mismatch sites, may not actually result in off-target cleavage despite prediction. To reflect this, sites with more mismatches (i.e., three or more) penalize the aggregate off-target score much less than sites with fewer (i.e., one, two) mismatches. Nontargeting gRNAs were designed without any perfect genomic matches and were also chosen based on having an aggregate off-target score >90%, suggesting that genomic cleavage is unlikely. While cleavage mediated by a nontargeting gRNA is still possible at sites with genomic mismatches, it is expected that these occurrences are rare.

Single-Nucleotide Polymorphisms Can Create Novel Off-Target Sites. We first investigated whether single-nucleotide polymorphisms (SNPs) altered the number of off-target sites in the genome using 7,444 WGSs from three different datasets: 1000 Genomes Project phase 3 (1000G) (n = 2,504) (20), a subset of the gnomAD database [an updated and expanded version of the ExAC dataset (26)] (n = 2,938), and a French Canadian (FC) dataset (n = 2,002) (27) (see *Materials and Methods* for additional details). Notably, the FC dataset is a founder population with increased genetic homogeneity. Fewer variants suggested a decreased probability to create or alter off-target sites a priori.

SNPs can alter off-target sites by increasing or decreasing the number of mismatches between a genomic region and the gRNA sequence. In addition, SNPs can create (alter NHG or NGH from reference genome to become an NGG motif) or destroy (alter reference genome NGG motif to NHG or NGH sequence) PAM sequences (H = A, C, or T). Creation of PAM sequences may generate new loci for off-target cleavage while destruction of PAM sequences potentially removes loci for off-target cleavage. SNPs present within the 1000G database led to the creation of 11,585,879 new NGG PAM sequences (4.1% of total PAMs in the reference genome, 11,585,879/281,005,914) and led to the destruction of 22,182,468 PAM sequences (7.9% of total PAMs in the reference genome, 22,182,468/281,005,914). To determine the number of PAMs per haploid genome within the 1000G dataset, the number of created PAMs was added and the number of destroyed PAMs was subtracted from the total number of NGG motifs in the reference autosomal genome (n = 281,005,914). Interestingly, the number of NGG motifs per haploid genome was

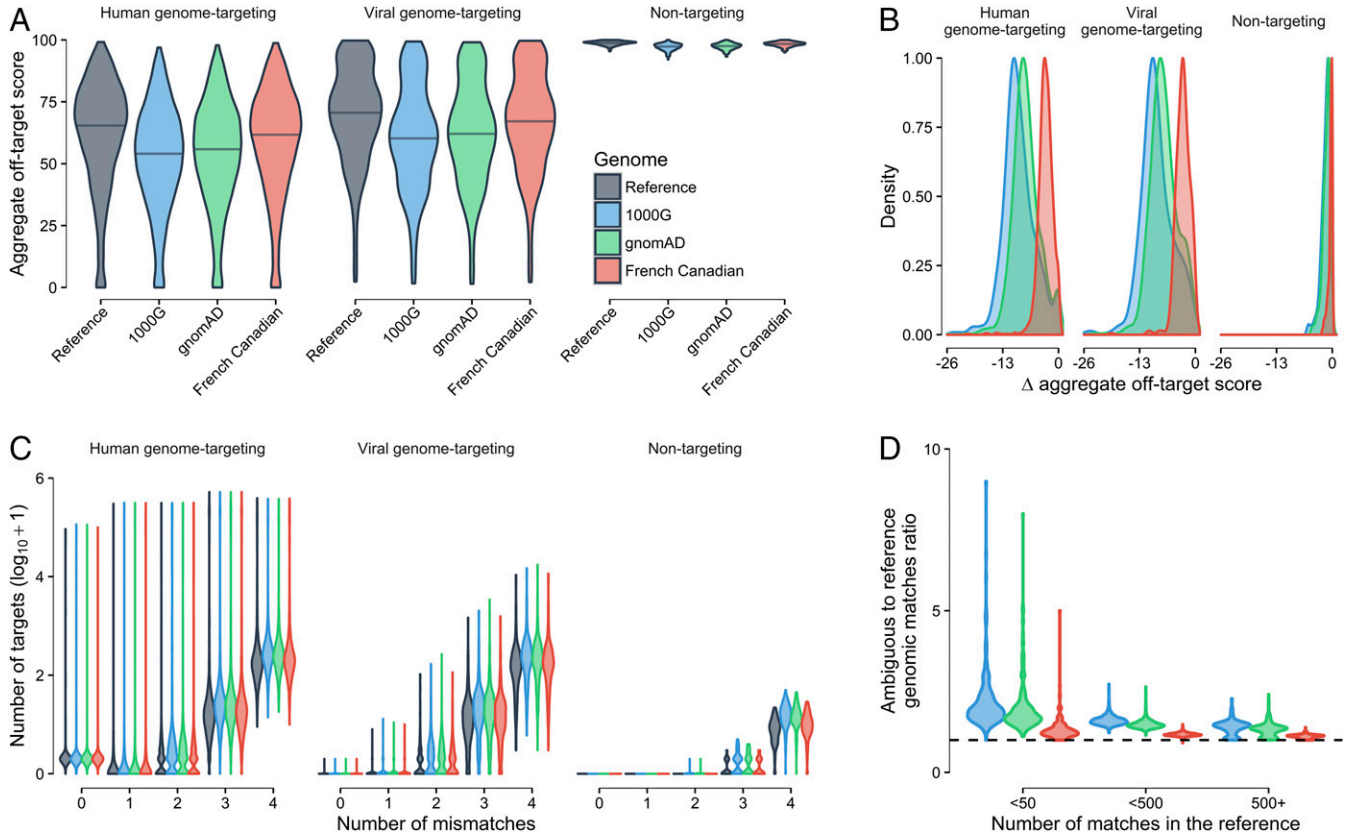


Fig. 1. Off-target scores using the ambiguous genome approach. (*A*) Distribution of aggregate off-target scores in the reference and ambiguous genomes for human-genome–targeting, viral-genome–targeting, and nontargeting gRNAs. (*B*) Change in aggregate off-target score between ambiguous and reference genomes. (*C*) Distribution of off-target sites by number of mismatches. (*D*) Ratio of the number of off-target sites in ambiguous genomes compared with the reference genome stratified by the number of off-target sites in the reference genome. The *y* axis shows the ratios for each gRNA, whereas the *x* axis shows the number of off-target sites in the reference genome.

similar to the reference genome, with a mean increase of 34 NGG motifs (median, −42 NGG motifs). However, the number of NGG motifs varied across individual haploid genomes (SD, ±1,559 NGG motifs). The number of NGG motifs also varied across populations, with individuals of European descent showing the largest reduction (−1,327 ± 922, mean ± SD) and individuals of African ancestry displaying the largest increase as compared to the reference genome (1,429 ± 1,142, mean ± SD, Fig. S1).

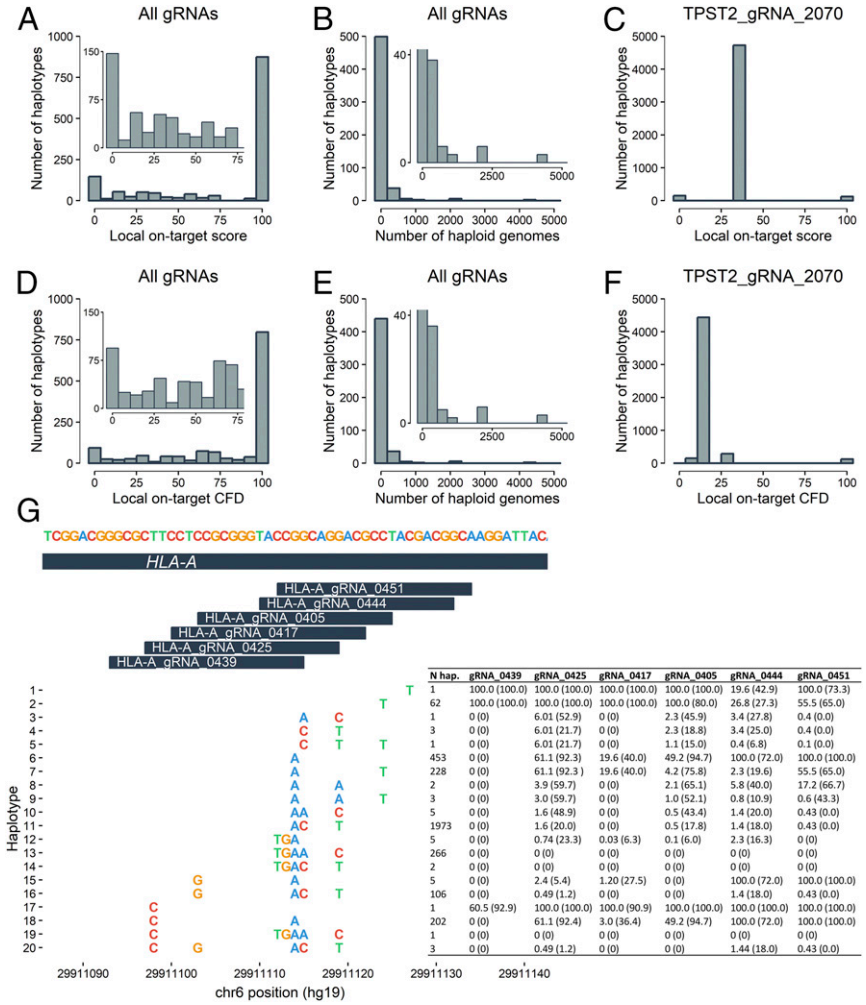
To further investigate the effect of SNPs within the 1000G, gnomAD, and FC datasets, we created an "ambiguous genome" by replacing each SNP position by an International Union of Pure and Applied Chemistry (IUPAC) ambiguity code to account for all possible SNP alleles. For example, an A > C SNP would be replaced by the ambiguity code "M" (M = A or C). With this replacement strategy, both alleles can map to the SNP locus without penalty. Therefore, all possible matches upstream of an NGG motif were identified in the reference and ambiguous genomes (up to four mismatches), which were used to calculate off-target scores. Interestingly, we observed a reduction in aggregate off-target scores for all three datasets when comparing the reference and ambiguous genomes, which suggested increased off-target cleavage potential (Fig. 1 *A* and *B*). The largest decrease in aggregate off-target scores was associated with the 1000G dataset (Fig. 1*A* and Dataset S3). For human genome-targeting gRNAs, the mean reductions in aggregate off-target scores were 9.3%, 7.8%, and 3.1% for the 1000G, gnomAD, and FC datasets, respectively (Fig. 1 *A* and *B*). For viral-genome–targeted gRNAs, the mean reductions of aggregate off-target scores were 9.1%, 7.5%, and 3.0%, respectively (Fig. 1 *A* and *B*). These data were consistent with the reduced genetic diversity within the FC founder population. The decreased mean aggregate off-target scores predominantly resulted from an increased number of low-scoring off-target sites, which are those with two to four mismatches from the reference gRNA sequence (Fig. 1*C*).

Notably, the ratio of the number of ambiguous genome to reference genome off-target sites suggested that gRNAs with fewer off-target sites in the reference genome displayed higher ratios (i.e., increased number of off-target sites in the ambiguous genome compared with the reference genome); however, this may reflect increased noise in the ratio because new off-target sites contribute more to the ratio for gRNAs with fewer off-target sites in the reference genome (Fig. 1*D*). As expected, nontargeting gRNAs showed the smallest decrease in aggregate off-target score (Fig. 1 *A* and *B*). Of the 2,327 (2,327/2,481, 93.8%) human-genome–targeted gRNAs with only one predicted perfect genomic match (single on-target site with zero mismatches) in the reference genome, 14 gRNAs (0.6%) were predicted to have new perfect off-target matches in the ambiguous genome. Of note, *HLA-A*_gRNA_0422 had an off-target score of 80.7% in the reference genome, but only 27.5% when considering gnomAD SNPs with the number of off-target sites being increased by 75 including three off-target sites with 0 or one mismatch. *HLA-A*_gRNA_0422 showed the largest score reduction when considering gnomAD and 1000G SNPs followed by consideration of the FC dataset. The total number of human-genome–targeted gRNAs with a reduction in aggregate off-target score of >30% was 23 (23/2,327, 1.0%), 17 (17/2,327, 0.7%), and 5 (5/2,237, 0.3%), respectively, for the 1000G, gnomAD, and FC datasets. For viral-genome–targeted gRNAs, these reductions were 3 (3/484, 0.6%), 5 (5/484, 1.0%), and 2 (2/484, 0.4%), respectively, for the same datasets. Overall, these results suggest that SNPs can create novel off-target sites and reduce the number of mismatches in existing off-target sites, thus increasing the potency of the associated off-target site. Notably, reduced alteration of off-target potential among the FC dataset suggested that the effect of genetic variation is likely dependent on the extent of the individual genetic diversity. Taken together, these data suggest that SNPs can increase the off-target cleavage potential for gRNAs and further suggest that an increased number of SNPs is likely to increase the off-target cleavage potential.

Variants Alter gRNA On-Target Specificity for Human-Genome–Targeting gRNAs. The ambiguous genome approach offered an initial assessment of the effect of variants on targeting specificity; however, the ambiguous genome analysis approach is limited because it does not consider haplotypes present in the population, it does not discriminate allele frequencies, and it does not include insertion/deletions (indels). To address these limitations, we selected the subset of gRNAs with an aggregate off-target score of ≥80% in the reference genome and tested these against every possible haplotype in the 1000G dataset including both SNP and indel variants. This analysis was restricted to gRNAs with ≥80% aggregate off-target scores because gRNAs with low aggregate off-target scores are unlikely to be considered for therapeutic applications. Therefore, this subset included 481 human-genome–targeting, 150 viral-genome–targeting, and 128 nontargeting gRNAs.

Using this approach, we first investigated on-target sites for each human-genome–targeting gRNA, which identified 263 gRNAs (263/481, 54.7%) with on-target sites harboring variants from the 1000G dataset (Fig. 2 *A*–*F* and Dataset S4). These gRNAs targeted 83 different regions after aggregating SNPs based on proximity into local haplotypes (Fig. 2*G* offers an example of a single region with multiple SNPs in close proximity). These regions were composed of 310 unique haplotypes from the 1000G dataset with a mean of 3.7 different haplotypes per region. In total, 58.6% (n = 793/1,353) of gRNA–target haplotype pairs were predicted to yield a perfect match (perfect local targeting score) and perfect cutting frequency determination (CFD) score, which is another score for the assessment of gRNA activity and off-target cleavage potential (25). On the other hand, 27.8% (n = 376/1,353) of gRNA–haplotype pairs yielded a local on-target score below 50%, and 20.9% (n = 283/1,353) of gRNA–haplotype pairs resulted in a CFD below 50% (Fig. 2 *A* and *D*). These affected sites (i.e., sites with SNPs at their target site resulting in local on-target score <50% or CFD <50%) belonged to 176 (176/263, 66.9%) and 139 (139/263, 52.9%) of human-genome–targeting gRNAs, respectively. In total, 16.3% (n = 43/263) of gRNAs had at least one target haplotype yielding a null local on-target score or null CFD, where null signifies a local on-target score or CFD of zero. The frequency of null on-target haplotypes in the 1000G ranged from 0.02% (n = 1/5,008) to 39.4% (n = 1,973/5,008) with a mean of 2.0% (median, 0.06%). Similarly, the frequency of imperfect haplotypes [mismatch(es) at on-target site] was highly biased toward singletons (Fig. 2 *B* and *E*). Nonetheless, 15.6% (n = 41/263) of gRNAs with SNPs at their on-target sites were predicted to have a local on-target score or local CFD <100% in 50 samples/individuals or more. For instance, *TPST2*_gRNA_2070 (chr22:26,937,299–26,937,349; hg19) had a local on-target score of 38.7% (CFD, 13.6%) in 4,439 (88.6%) haploid genomes (Fig. 2 *C* and *F*). Six gRNAs targeted the *HLA-A* region (chr6:29,910,958–29,911,176; hg19; Fig. 2*G*). This region included nine SNPs implicated in 20 unique haplotypes (excluding the reference). Of note, a haplotype (haplotype #10, Fig. 2*G*) present in 39.4% (n = 1,973/5,008) of samples abrogated the target site for all six gRNAs. In 71 of 92 (77%) null haplotypes, the null score was due to an altered PAM site.

We repeated this type of haplotype-based analysis using human-genome–targeted gRNAs in samples/individuals from the FC dataset and identified 155 human-genome–targeting gRNAs targeting 243 different haplotypes in 26 unique genomic regions (Fig. S2 and Dataset S5). Here, we found 430 (430/2,844, 15.1%) and 317 (317/2,844, 11.1%) gRNA–haplotype pairs that reduced the local on-target score or local CFD below 50%, respectively. In total, 9.7% (n = 15/243) of gRNAs had null local on-target scores in at least one haplotype. These haplotypes had a mean frequency of 1.5% (n = 59.4/4,004; median = 2.5%). In total,



N hap.	gRNA_0439	gRNA_0425	gRNA_0417	gRNA_0405	gRNA_0444	gRNA_0451
1	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	19.6 (42.9)	100.0 (73.3)
62	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)	100.0 (80.0)	26.8 (27.3)	55.5 (65.0)
1	0 (0)	6.01 (52.9)	0 (0)	2.3 (45.9)	3.4 (27.8)	0.4 (0.0)
3	0 (0)	6.01 (21.7)	0 (0)	2.3 (18.8)	3.4 (25.0)	0.4 (0.0)
1	0 (0)	6.01 (21.7)	0 (0)	1.1 (15.0)	0.4 (6.8)	0.1 (0.0)
453	0 (0)	61.1 (92.3)	19.6 (40.0)	49.2 (94.7)	100.0 (72.0)	100.0 (100.0)
228	0 (0)	61.1 (92.3)	19.6 (40.0)	4.2 (75.8)	2.3 (19.6)	55.5 (65.0)
2	0 (0)	3.9 (59.7)	0 (0)	2.1 (65.1)	5.8 (40.0)	17.2 (66.7)
3	0 (0)	3.0 (59.7)	0 (0)	1.0 (52.1)	0.8 (10.9)	0.6 (43.3)
5	0 (0)	1.6 (48.9)	0 (0)	0.5 (43.4)	1.4 (20.0)	0.43 (0.0)
1973	0 (0)	1.6 (20.0)	0 (0)	0.5 (17.8)	1.4 (18.0)	0.43 (0.0)
5	0 (0)	0.74 (23.3)	0.03 (6.3)	0.1 (6.0)	2.3 (16.3)	0 (0)
266	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
2	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
5	0 (0)	2.4 (5.4)	1.20 (27.5)	0 (0)	100.0 (72.0)	100.0 (100.0)
106	0 (0)	0.49 (1.2)	0 (0)	0 (0)	1.4 (18.0)	0.43 (0.0)
1	60.5 (92.9)	100.0 (100.0)	100.0 (90.9)	100.0 (100.0)	100.0 (100.0)	100.0 (100.0)
202	0 (0)	61.1 (92.4)	3.0 (36.4)	49.2 (94.7)	100.0 (72.0)	100.0 (100.0)
1	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
3	0 (0)	0.49 (1.2)	0 (0)	0 (0)	1.44 (18.0)	0.43 (0.0)

Fig. 2. Variants can reduce gRNA targeting efficiency. (*A*) Distribution of on-target scores for human-genome–targeting gRNAs for each possible target haplotype. (*B*) Distribution of samples/individuals carrying haplotypes predicted to be targeted with a local on-target score of <100%. (*C*) Distribution of local on-target scores for the gRNA *TPST2*_gRNA_2070. (*D*) Distribution of on-target CFDs for human-genome–targeting gRNAs for each possible target haplotype. (*E*) Distribution of samples/individuals carrying haplotypes predicted to be targeted with a CFD of <100%. (*F*) Distribution of CFDs for the gRNA *TPST2*_gRNA_2070. (*G*) Example of haplotypes at the *HLA-A* locus. *Inset* plots with a restricted *y*-axis range are shown for *A*, *B*, *D*, and *E* for easier visualization of data.

2.9% (n = 7/243) of gRNAs targeted at least one null haplotype seen in more than 40 samples (~1% frequency) clustered in five unique regions. The most common null haplotype was present in 22.6% of haploid genomes (n = 906/4,004; chr22:26936744–26936919; hg19; *TPST2*_gRNA_2144). In total, 69% (n = 87/126) of null haplotypes were due to an altered PAM sequence. Overall, these results suggest that genetic variants in the on-target site can dramatically affect gRNA targeting specificity and efficiency, particularly if the variants are located within PAM sequences.

Characteristics of Off-Target Sites. We then assessed the global characteristics of off-target sites for all human-genome–targeting, viral-genome–targeting, and nontargeting gRNAs with an aggregate off-target score ≥80% in the reference genome. The 1000G and FC dataset variants altered 21,981 and 10,348 off-target sites, respectively, targeted by gRNAs from Dataset S2 in the reference genome (Datasets S6 and S7). 1000G- and FC-derived variants created an additional 23,316 and 8,773 unique off-target sites, respectively. In both cases, ~8% (n = 1,767/23,316 and 699/8,773) were solely due to novel PAM sites created by variants. On the other hand, variants overlapping PAMs destroyed matches at 13.8% (n = 3,039/21,981) and 10.5% (n = 1,084/10,348) of reference sites. When variants altered the underlying off-target site sequence, the median change in local off-target score due to variants was ±0.03% in the 1000G and FC datasets and the mean difference in local CFD was ±2.0% in both datasets. These small changes in local off-target and CFD scores were predominantly due to the large number of off-target sites with four mismatches, which accounted for >92% of off-target sites in both the 1000G and FC datasets. In the 1000G dataset, 73.2% (n = 556/759) of the gRNAs with an aggregate off-target score ≥80% in the reference genome had new off-target sites with less than four mismatches, 10.1% (n = 77/759) with less than three mismatches, and 0.5% (n = 4/759) with less than two mismatches. The new mismatches were at 1,531, 84, and 5 unique sites, respectively (Dataset S8). In the FC dataset, 49.3% (n = 374/759) of the gRNAs with an aggregate off-target score ≥80% in the reference genome had new off-target sites with less than four mismatches, 4.9% (n = 37/759) with less than three mismatches, and 0.7% (n = 5/759) with less than two mismatches. The new mismatches were at 599, 45, and 5 unique sites, respectively (Dataset S8).

Indels can theoretically have a higher impact on reference sequence than SNPs, potentially resulting in the creation of novel/altered off-target sites. To investigate the effect of indels, we examined

regions with off-target sites consisting of indel-only haplotypes in the 1000G dataset to consider indels independently from SNPs. This analysis identified 184 sites with 69.6% (128/184) of them with novel off-target sites. This was a modest enrichment compared with haplotypes consisting of SNPs-only (n = 21,169/34,636; 61.1%; Fisher exact test, $P = 0.019$), although this was not significant in the FC dataset (indels-only: 357/522, 68.4%; SNPs-only: 7,777/11,897, 65.4%; $P = 0.16$). However, 7.6% (n = 27/357) of indel-mediated off-target site alteration in the FC dataset had less than four mismatches, whereas that ratio was 0.8% for SNPs (n = 66/7,777; Fisher exact test, $P = 1.1 \times 10^{-15}$), suggesting that indels are more likely to create more potent novel off-target sites. Similarly, sites in the reference genome with less than four mismatches were more likely to be completely destroyed by indels than by SNPs (1000G odds ratio = 10.5, $P = 9.5 \times 10^{-4}$; FC odds ratio = 20.6, $P = 4.1 \times 10^{-12}$).

Notably, of the 45,297 (21,981 + 23,316) different off-target sites, 5,633 (12.4%) were covered by structural variants annotated in the 1000G dataset, suggesting that these could also modify the probability of off-target effects. For instance, an *HLA-B* gene haplotype (chr6:31,238,852–31,238,965; hg19) was a strong off-target site for *HLA-A*_gRNA_0422 in 29.1% (n = 1,459/5,008) of haplotypes (local off-target score, 60.5–68.3%; local CFD, 100). In total, 6.9% (n = 344/5,008) haploid genomes had a deletion of this region, completely removing this target site in the 1000G dataset. On the other hand, three samples (3/5,008, 0.06%) had a duplication covering this site (chr6:31,131,451–31,272,307; hg19), thus increasing the number of off-target sites.

Assessing Off-Target Effects Using Personal Genomes. We calculated aggregate off-target scores for all gRNAs (human-genome–targeting, viral-genome–targeting, and nontargeting) in the 1000G and FC datasets (Fig. 3). In the vast majority of cases (n = 3,753,205/3,796,064 gRNA–haploid genome pairs, 98.9%), the individual gRNA aggregate off-target score was ≥80%. Nonetheless, this accounted for 42,859 haploid genome–gRNA pairs with a score <80%, implicating 62 gRNAs and virtually all samples (Fig. 3*A*). The FC dataset showed similar statistics, with 99.3% (n = 2,823,851/2,842,840) of haploid genome–gRNA having a score ≥80% (Fig. 3*B*). Again, all samples had an aggregate off-target score of <80% for any one of 23 gRNAs. Consistently, the mean reduction in aggregate off-target score was −0.03% and −0.01% for the 1000G and FC datasets, respectively, when examining 758 and 710 gRNAs with at least one overlapping variant at an off-target site, respectively (Fig. 3 *C* and *D*). Only seven gRNAs for the 1000G dataset and one gRNA for the FC dataset had a reduction in aggregate off-target score of more than 5% in at least one individual. Four gRNAs showed a very strong reduction in score (>15% reduction in aggregate off-target score) in at least one haplotype in the 1000G dataset (Table 2). *ALB*_gRNA_0837 had an aggregate off-target score between 82.4% and 84.2% in 83% of haplotypes (4,153/5,008). However, the aggregate off-target score was reduced below 46% in the remaining 17% samples (855/5,008). This was primarily due to a single off-target site on chromosome 11 (chr11:100,402,414–100,402,433; hg19) (Table 2). In the reference genome, this region on chromosome 11 contains two mismatches (local off-target score, 2.4%), one of which is rescued by rs11560892 (C > G) matching to position 18 of the gRNA sequence. The remaining mismatch is not predicted to alter targeting (local off-target score, 100%), thus creating a potent off-target site.

Similarly, *HLA-A*_gRNA_0451 had an aggregate off-target score of >88.6% in the majority of haplotypes (n = 5,004/5,008, 99.8%) (Table 2). However, four haplotypes showed an aggregate off-target score of 48%, which was mainly due to a rescued off-target site on chromosome 13 (chr13:33,591,178–33,591,197; hg19; Table 2). Notably, this region falls in the coding region (exon 1) of

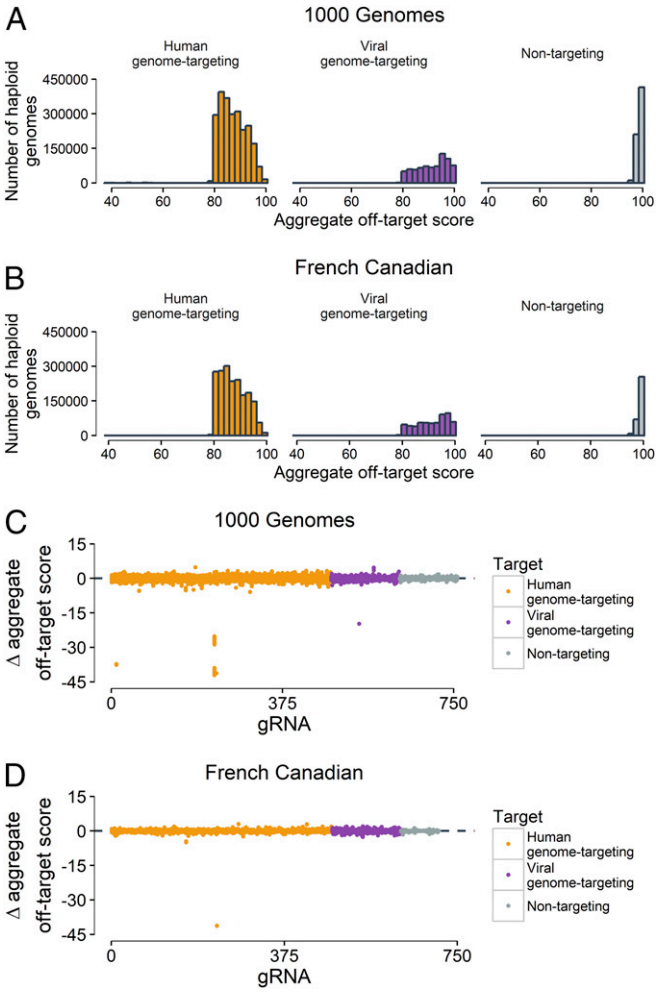


Fig. 3. Variants can increase the risk of off-target effects. (*A*) Distribution of aggregate off-target scores for each 1000G haplotype. (*B*) Distribution of aggregate off-target scores for each FC haplotype. (*C*) Difference in aggregate off-target scores for each 1000G haploid genome and the reference genome. The *x* axis corresponds to different gRNAs, and each dot represents the difference in score of each haploid genome in the 1000G dataset. The figure includes 758 gRNAs with at least one match with overlapping variants. (*D*) Difference in aggregate off-target score for each FC haploid genome and the reference genome. The *x* axis corresponds to different gRNAs and each dot represents the difference in score of each haploid genome in the FC dataset. The figure includes 710 gRNAs with at least one match with overlapping variants.

the *KL* gene, a gene associated with hyperphosphatemic familial tumoral calcinosis (28); these haplotypes belonged to four unique haplotypes. Although all of the four samples/individuals had one copy of the on-target site that had a perfect match, they also carried a copy whereby the on-target site was predicted to have very low (n = 3/4; local on-target score, 0.4%) or reduced (n = 1/4; local on-target score, 55.5%) activity, making these individuals at potentially increased risk of both treatment failure and adverse effects due to off-target cleavage (Fig. 2*G*).

One individual was a carrier of the G allele of rs552139758 (A > G), which created a novel PAM sequence on chromosome 14 (chr14:52,120,745–52,120,765; hg19) and a novel off-target site for *HIV-1*_gRNA_0196 (Table 2). This PAM-creation site on chromosome 14 falls within intron 1 of the *FRDM6* gene and results in reduction of the aggregate off-target score for *HIV-1*_gRNA_0196 from >80.8 to 61.9%. In addition, *HLA-A*_gRNA_0422 had an off-target site in the coding sequence of *HLA-C* (Table 2) whereby two haplotypes showed local off-target scores >60% in 1,459

Table 2. Representative example of off-target sites created by variants present in the 1000 Genomes database

Seq. pos.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	PAM 1	PAM 2	PAM 3	Local score	CFD	*N*	Freq.
Score penalty	0	0	0.014	0	0	0.395	0.317	0	0.389	0.079	0.445	0.508	0.613	0.851	0.732	0.828	0.615	0.804	0.685	0.583							
ALB_gRNA_0837; chr11:100402414–100402433 (-) (hg19)																											
Guide	T	A	A	T	T	T	T	C	T	T	T	T	G	C	G	C	A	C	T	A	N	G	G				
Reference	T	A	A	A	T	T	T	C	T	T	T	T	G	C	G	C	A	G	T	A	T	G	G	2.39	30.3	4149	82.85
Hap. 1																		C						100	63.3	855	17.07
Hap. 2									G									C						3.89	33.9	1	0.02
Hap. 3															A			C						2.5	41.6	2	0.04
Hap. 4														T				C						1.29	18.2	1	0.02
HLA-A_gRNA_0451; chr13:33591178–33591197 (-); KL CDS (hg19)																											
Guide	T	G	C	C	G	T	C	G	T	A	G	G	C	G	T	C	C	T	G	C	N	G	G				
Reference	A	G	C	C	G	C	C	G	T	A	G	G	C	G	T	C	C	A	G	C	A	G	G	3.83	92.9	4981	99.46
Hap. 1													A											0.78	19.5	23	0.46
Hap. 2						T																		100	100	4	0.08
HIV-1_gRNA_0196; chr14:52120745–52120765 (+); FRMD6 intron (hg19)																											
Guide	C	A	G	C	A	G	T	T	C	T	T	G	A	A	G	T	A	C	T	C	N	G	G				
Reference	C	A	G	C	A	G	T	T	C	T	T	G	C	A	G	T	A	C	T	C	A	A	G	0	0	5002	99.88
Hap. 1	T																							0	0	5	0.10
Hap. 2																						G		38.7	38.5	1	0.02
HLA-A_gRNA_0422; chr6:31238925–31238944 (+); HLA-C CDS (hg19)																											
Guide	C	T	C	T	C	A	A	C	T	G	C	T	C	C	G	C	C	T	C	A	N	G	G				
Reference	C	T	C	T	C	A	G	C	T	G	C	T	C	C	G	C	C	G	C	A	C	G	G	1.25	66.7	1200	23.96
Hap. 1																		T						68.3	100.0	1123	22.42
Hap. 2																		A						1.25	50.0	663	13.24
Hap. 3						C												A						0.27	46.4	569	11.36
Hap. 4						C												T						2.16	92.9	690	13.78
Hap. 5						C	A											T						60.5	92.9	336	6.71
Hap. 6						C	A											A						1.20	46.4	252	5.03
Hap. 7																			T					0.14	18.4	161	3.21
Hap. 8						C	A																	1.20	61.9	7	0.14
Hap. 9						C																		0.27	61.9	5	0.010
Hap. 10																		A	T					0.14	13.8	2	0.004

Variants included in the chr11:100402414–100402433 (hg19) haplotype are rs566289682, rs555981507, rs181027193, and rs11560892. Variants included in the chr13:33591178–33591197 (hg19) haplotype are rs200611452 and rs116289670. Variants present in the chr14:52120745–52120765 (hg19) haplotype are rs532153306 and rs552139758. Sites displaying mismatches with the gRNA sequence are shown in red, whereas sites where variants rescue the gRNA sequence are highlighted in blue. CFD, cutting frequency determination; Freq., frequency; Hap., haplotype; PAM, protospacer adjacent motif; Seq. pos., sequence position.

haploid genomes. This is also an example where the CFD score (66.7) was a better predictor of off-target potential than the local off-target score (1.25%), in that the former was less affected by the presence of SNPs (Table 2).

Additionally, instances were identified with an improved mean aggregate off-target score due to variants within the 1000G dataset (mean Δ aggregate off-target score of >0; Fig. 3*C*). The distribution of gRNAs with a Δ aggregate off-target score of >0 was shifted toward 0, suggesting that variants were more likely to decrease the aggregate off-target score within each haploid genome (Fig. S3). Notably, F9_gRNA_1349 had a 4.8% increase in aggregate off-target score albeit only in one haploid genome. In total, 727 (14.5%, 727/5,008) haploid genomes displayed a Δ aggregate off-target score of >0 for HSV-1_gRNA_0057, with 38 (0.8%, 38/5,008) displaying a Δ aggregate off-target score of >4%. HSV-1_gRNA_0079 had a Δ aggregate off-target score of >0 in all haploid genomes, although the gain was limited due to the gRNA already having an aggregate off-target score of 98.9 in the reference genome (0.89 mean Δ aggregate off-target score). Overall, the magnitude of off-target score increase was likely blunted in this analysis since we selected for gRNAs with high aggregate off-target scores (≥80%). Finally, we investigated whether some populations were more at risk for off-target effects than other populations in the 1000G dataset. We calculated the difference in scores between each sample and the reference for each gRNA. In total, 721 of 758 (95.1%) of gRNAs showed significant differences in scores between populations after adjusting for multiple comparisons (Kruskal–Wallis, $P < 6.6 \times 10^{-5}$). Overall, African-ancestry populations showed the largest reduction in scores compared with the reference population (Δ aggregate off-target score, −0.0346; SD, 1.0745), while Europeans populations displayed the smallest changes (Δ aggregate off-target score, −0.0216; SD, 0.9882) (Dataset S9). This is consistent with increased genetic diversity observed in populations of African ancestry. Taken together, variants may predispose a subset of individuals to adverse events for CRISPR-mediated therapeutic genome editing.

Discussion

CRISPR technology holds enormous potential for clinical translation as therapy for a wide array of genetic disorders. Historically, gene therapy clinical trials have demonstrated that a small subset of patients may experience adverse events (29). Our data suggest that variants may contribute to both treatment failure of CRISPR-based therapies as well as predispose individuals to adverse outcomes due to personalized off-target effects; however, the effect of variants on on- and off-target specificity is not unique to CRISPR genome editing, but also extends to other genome-editing platforms including zinc finger nucleases and TAL effector nucleases. Notably, we identified variant-induced off-target sites in coding sequence. This type of situation potentially offers an adverse clinical outcome if such sites are located within genes with important roles for cellular function (e.g., tumor suppressor genes). It may be advisable for safety considerations to exclude gRNAs with predicted off-target sites within or near important genes such as tumor suppressors even if they have three or four mismatches. As such, these data may suggest the utility of WGS for patients before therapeutic genome-editing treatments. WGS data would allow for in silico on- and off-target analysis, which may identify patients predisposed to treatment failure and/or adverse outcomes before therapy initiation. Notably, given the creation/alteration of off-target sites in noncoding sequence, WGS would likely be required for this analysis as opposed to whole-exome sequencing. Minimally, our results suggest that on-target sites should be investigated by conventional Sanger sequencing to assure maximal gRNA efficiency. Alternatively, in vitro unbiased

genome-wide off-target detection methods can be employed (13–19). It is also possible to overcome adverse events by using enhanced-specificity/high-fidelity versions of SpCas9 (30–32), by using other methodologies to enhance targeting specificity (25, 33–36), and/or by furthering the understanding of cleavage kinetics to help minimize nuclease exposure to reduce off-target potential (37). However, variants that create potent off-target sites (e.g., novel zero or one mismatch sites) are likely to be problematic even in the setting of improved specificity techniques. Furthermore, enhanced-specificity/high-fidelity nucleases are only available for SpCas9 at present.

It is important to note that our study only considered NGG-restricted gRNAs compatible with SpCas9 (or enhanced-specificity versions such as SpCas9-HF1/eSpCas9/HypaCas9) (30–32); however, the effect of variants altering on- and off-targeting specificity is unlikely to be restricted to SpCas9 and will likely affect all CRISPR nucleases considered for therapeutic genome-editing applications (38). In addition, given the wide array of genetic or viral diseases that could be targeted by genome-editing approaches, we have evaluated only a small subset of the possible therapeutic loci; however, we have identified gRNAs with variant-induced reduction in predicted gRNA efficacy at on-target sites and variant-induced creation of potent off-target sites. This finding appears unlikely to be specific to the chosen loci and more likely to be a generalized phenomenon. However, it is important to note that our data suggest that these findings are rare, which is consistent with previous work (38).

The FC dataset was included to evaluate for population-specific effects due to novel variants present and/or variants present at differential allele frequencies within a specific population. Deleterious population-specific effects were not overtly observed in this dataset; however, stratification of the 1000G dataset by population demonstrated population-specific effects for on- and off-target specificity. The minimal population-specific effects observed in the FC dataset are consistent with its increased genetic homogeneity as a founder population and thus fewer differences with the reference genome. Notably, founder populations are associated with fewer variants; however, the variants are often more frequent. The increased frequency of particular variants may become problematic for therapeutic genome editing if certain high-frequency variants alter on-target sites and/or create high-potency off-target sites. Taken together, differential variant frequencies within populations are likely to contribute to population-specific effects for CRISPR-based therapeutic targeting.

As the understanding of Cas9 binding continues to unfold and aid in determination of off-target loci (39) and factors affecting accessibility of these sequences [e.g., nucleosomes (40)], it may be possible to refine in silico off-target analysis beyond sequence-only to further predict if off-target sites and/or variant-induced off-target sites are likely to result in off-target cleavages. In particular, future analysis would benefit from differentiating between the requirements for Cas9 binding vs. Cas9 cleavage (32, 41); incorporation of this type of information would likely increase the reliability of identifying off-target sites with a high probability of cleavage.

To minimize the possibility of variants affecting gRNAs in development for clinical translation, it may be useful to consider variants at the gDNA design stage. For example, publicly available variant databases [e.g., dbSNP, dbVAR, ExAC (26), 1000G (20)] may be examined during gRNA design to create variant-aware gRNAs (21). In silico analysis, such as presented in this manuscript, can also be used to aid gRNA selection for clinical translation. In addition, gRNAs derived from the reference genome or variant-aware gRNAs can be tested in diverse cell lines or primary cells to evaluate for toxicity. One might also evaluate a therapy-optimized CRISPR gRNA using patient-derived induced pluripotent stem cells differentiated to the relevant lineage, which could represent a viable paradigm for empiric evaluation of variant-induced effects on CRISPR targeting; however, this approach could be compromised by somatic mosaicism, which has been detected in many individuals across numerous tissue types (42). The somatic mutation rate has been estimated to be $\sim 10^{-9}$/nucleotide/cell division (43, 44). Further estimates suggested 3,500–8,900 cell divisions for cells such as lymphocytes, lymphoblastoid cell lines, or colonic mucosae in ~65-y-old individuals (44). Therefore, it is conceivable that somatic variants may limit the ability to evaluate on- and off-target sites using any of the suggested methods, particularly for individuals with advanced age. Of note, estimates of the germline mutation rate have varied widely with estimates above and below the somatic mutation rate (43, 45, 46). Interestingly, somatic mosaicism could also be exploited for CRISPR-based therapy, such as for cancers with genomic amplifications, through induction of apoptosis due to numerous double-strand breaks (47, 48).

Taken together, our analysis suggests the necessity for preclinical studies to consider variants at the gDNA design stage and/or to validate more than one gRNA for clinical translation to increase the likelihood of providing safe, effective, and personalized therapeutic options for all patients regardless of genotype. In summary, our data suggest that human genetic variation alters on- and off-target specificity for CRISPR-based therapeutic genome editing. Therefore, it will be prudent to account for patient-specific genomes in on- and off-target analyses as CRISPR-based therapies approach the clinic.

Materials and Methods

gRNA Design. gRNAs were designed using publicly available tools (11) and/or identified in previously published studies (Table 1 and Dataset S2). gRNAs for both NHEJ and HDR applications were designed to include all gRNAs within the relevant exon(s) for coding region targets and ±100 bp for noncoding targets. Human genome (hg19) was used to obtain gene-based sequences. Viral sequences utilized were as follows: EBV (49): KC207813.1 human herpesvirus 4 strain Akata, complete genome;

CMV (49): KF297339.1 human herpesvirus 5 strain TB40-E clone Lisa, complete genome;

HSV1 (49): JN555585.1 human herpesvirus 1 strain 17, complete genome;

HPV E6E7: LC193821.1 human papillomavirus type 16 DNA, complete genome, isolate: FT001;

HIV1 (50): AF105229.1 cloning vector pHR′-CMVLacZ, complete sequence;

HBV (51): AF305422.1 synthetic construct hepatitis B virus 1.28-mer overlength sequence; EU570069.1 hepatitis B virus isolate 1-B24, complete genome; FJ899793.1 hepatitis B virus isolate C122-2, complete genome; V01460.1 hepatitis B virus (strain ayw) genome;

JCV (52): NC_001699.1 JC polyomavirus, complete genome.

CRISPOR (11) was used to obtain gRNA efficiency scores from Fusi et al. (53), Chari et al. (54), Xu et al. (55), Doench et al. (25, 56), Wang et al. (57), Moreno-Mateos et al. (58), Housden et al. (59), Prox. GC (60), -GG (61), and Out-of-Frame (62).

Calculation of Off-Target and CFD Scores. Off-target scores were calculated as previously described (11, 23, 24). Briefly, the number and position of mismatches between gRNA–DNA were calculated with scores ranging from 0 (nontargeting) to 1 (perfect match), which was termed the "local off-target score." Based on this analysis, sequences with a score >0 were considered potential off-targets. For sequences with more than four mismatches, a score of 0 was assigned. An aggregate off-target score from all possible local off-targets was calculated according to Sanjana et al. (23):

$$S_{\text{guide}} = \frac{100}{100 + \sum_{i=0}^{n} S_{\text{hit}}(h_i)}.$$

In this equation, n signifies the number of potential off-target "hits" and $S_{\text{hit}}(h_i)$ is the targeting score of the possible off-target sequence h_i. Therefore, a "local" off-target score was calculated for each genomic match (from 0 to four mismatches) for a given gRNA. The summation of all local off-target scores for a given gRNA resulted in a genome-wide off-target score, termed "aggregate off-target score." For this off-target scoring method (score range, 0–100), higher scores indicate lower off-target cleavage potential and lower scores indicate higher off-target cleavage potential.

CFD scores were calculated as previously described (21, 25). Briefly, percent activity values are provided in Doench et al. (25) for all possible gRNA–DNA

mismatches. These percent activity values can be multiplied together in the setting of multiple gRNA–DNA mismatches (25). When a local off-target score or local CFD was calculated at the on-target site, it is referred to as an "on-target score."

Nontargeting gRNA Design. In total, 128 gRNAs were used as negative controls. The 128 gRNAs were previously designed to lack perfect matches within the genome and have an aggregate off-target score of >90% based on the calculation described in *Calculation of Off-Target and CFD Scores* (21).

PAM Creation and Destruction Analysis. To determine the total number of PAMs in the genome for SpCas9, all NGG motifs were identified on the sense and antisense strands using the matchPattern function from the Biostrings package for all 22 autosomes. Destroyed PAMs were defined as GG sites that were overlapped by a SNP (this analysis was performed on both strands). In these cases, the reference allele was G so alternative alleles destroyed the GG motif (i.e., altered reference genome NGG motif to NHG or NGH sequence). Created PAMs were defined by identification of all SNPs with an alternative allele of a G and that were preceded or followed by a G nucleotide, thus creating a GG motif (i.e., altered NHG or NGH sequence to become an NGG motif; this analysis was performed on both strands). To determine the number of PAMs per haploid genome, we generated all possible haploid genomes from the 1000 Genomes dataset by inserting the alternative allele of SNPs at each site carried by the samples. We then counted the total number of NGG motifs for each haploid genome.

Genomic Coordinates. All genomic coordinates displayed are hg19. Coordinates for viral genomes are not displayed.

WGS Data. In total, 7,444 WGSs were obtained for analysis. These data were obtained from the 1000G database (n = 2,504) (20), a subset from the genome aggregation database [the gnomAD dataset, an updated and expanded version of the ExAC dataset (26); n = 2,939], and Low-Kam et al. (27) (n = 2,002). All three datasets were sequenced at low coverage (<15×).

Ambiguous Genome Analysis. Variants were downloaded from the 1000G phase 3 dataset (n = 2,504) (20). Data from two other whole-genome sequencing datasets were also accessed: an FC (n = 2,002) dataset from the Montreal Heart Institute biobank (27) and a subset of the GnomAD dataset (n = 2,938). An ambiguous genome was built using a custom R (version 3.2.0) script based on the R package Biostrings (version 2.38.4). Human genome sequences were obtained using the BSgenome (version 1.38.0) package BSgenome.Hsapiens.UCSC.hg19.masked (version 1.3.99), applying the default masks (assembly gaps and intracontig ambiguities). Each nucleotide was replaced at the SNP positions by an IUPAC ambiguity code to account for all possible SNP alleles. For example, an A→C SNP would be replaced by the ambiguity code "M," so that both alleles can map to the SNP location without penalty.

For each gRNA, all possible matches were identified in the reference and ambiguous genomes using the Biostring *matchPDict* function allowing up to four mismatches. Only matches upstream of an NGG motif and that had less than five ambiguities were considered. The restriction of less than five ambiguities was imposed so that ambiguities did not overinflate the number of matches. For each match, the targeting score was calculated as described in Sanjana et al. (23) using mismatch penalties from Hsu et al. (24), as well as the CFD score (25) (see *Calculation of Off-Target and CFD Scores* for more detail). For each gRNA, we reported the number of matches for each mismatch category, the aggregated score, and mean, median, SD, and 10th, 25th, 75th, and 90th percentiles of the CFD score.

On-Target Haplotype Analysis. To measure on-target effects in personal genomes, each SNP and indel in the 1000G and FC datasets overlapping the predicted on-target sites (including the PAM) was considered. The sequences 22 bp on either side of each variant were identified and overlapping target sites were merged to create local haplotypes. Genomic sequences were created based on existing haplotypes in the datasets and tested whether they were targeted by gRNAs using the Biostring *matchPDict* function (up to four mismatches). In total, 481 human-genome–targeting, 150 viral-genome–targeting, and 128 nontargeting gRNAs with aggregate off-target scores ≥80% in the reference genome were investigated. Only matches upstream of an NGG PAM were considered valid matches. The number of mismatches, off-target scores, and CFD scores were calculated as above for each match.

Δ Aggregate Off-Target Score. The "Δ aggregate off-target score" was calculated as the difference between the reference aggregate off-target score and each sample's aggregate off-target score.

Off-Target Haplotype Analysis. To measure off-target effects in personal genomes, each SNP and indel in the 1000G and FC datasets was considered. Sequences 22 bp on either side of each variant were identified and overlapping sequences were merged to create local haplotypes. Genomic sequences were created based on existing haplotypes in the datasets and tested whether they were targeted by gRNAs using the Biostring *matchPDict* function (up to four mismatches). Only matches upstream of an NGG PAM were considered valid matches. The number of mismatches, off-target scores, and CFD scores were calculated as above for each match. Each sample (individual) was then separated into haploid genomes and the aggregate off-target score was calculated given the individual's haplotypes:

$$s_{g,i} = \sum_{j}^{n} s_{g,i,j} + s_{g,\text{nonvariable}},$$

where $s_{g,i,j}$ is the off-target site score of the jth of n off-target sites of gRNA g in haplotypes of the haploid genome i. $s_{g,\text{nonvariable}}$ represents the sum of all local off-target scores in nonvariable regions of the genome (not overlapped by variants) for gRNA g. The aggregate off-target score of guide g ($Z_{g,i}$) in the haploid genome i is given by the following:

$$Z_{g,i} = 100 \times \frac{100}{100 + s_{g,i}}.$$

Off-Target Analysis Computational Tool. The computational tool ("CRISPR Off-Target Tool," version 2.0.1) used to perform the off-target analysis as well as its source code are available for download at www.mhi-humangenetics.org/en/resources.

ACKNOWLEDGMENTS. We thank Daniel E. Bauer for helpful discussions. We thank all participants and staff of the André and France Desmarais Montreal Heart Institute (MHI) Hospital Cohort. J.-C.T. holds the Canada Research Chair in Personalized Medicine and is funded by Genome Canada and Genome Quebec. G.L. is funded by Genome Canada and Genome Quebec, the Canada Research Chair Program, and the MHI Foundation. S.H.O. is supported by National Heart, Lung, and Blood Institute Award P01HL032262 and National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Award P30DK049216 (Center of Excellence in Molecular Hematology). M.C.C. is supported by NIDDK Award F30DK103359.

1. Cox DB, Platt RJ, Zhang F (2015) Therapeutic genome editing: Prospects and challenges. *Nat Med* 21:121–131.
2. Maeder ML, Gersbach CA (2016) Genome-editing technologies for gene and cell therapy. *Mol Ther* 24:430–446.
3. White MK, Khalili K (2016) CRISPR/Cas9 and cancer targets: Future possibilities and present challenges. *Oncotarget* 7:12305–12317.
4. Barrangou R, Doudna JA (2016) Applications of CRISPR technologies in research and beyond. *Nat Biotechnol* 34:933–941.
5. Prakash V, Moore M, Yáñez-Muñoz RJ (2016) Current progress in therapeutic gene editing for monogenic diseases. *Mol Ther* 24:465–474.
6. Carroll D (2016) Genome editing: Progress and challenges for medical applications. *Genome Med* 8:120.
7. Doudna JA (2015) Genomic engineering and the future of medicine. *JAMA* 313:791–792.
8. Cornu TI, Mussolino C, Cathomen T (2017) Refining strategies to translate genome editing to the clinic. *Nat Med* 23:415–423.
9. Cyranoski D (2016) CRISPR gene-editing tested in a person for the first time. *Nature* 539:479.
10. Sheridan C (2017) CRISPR therapeutics push into human testing. *Nat Biotechnol* 35:3–5.
11. Haeussler M, et al. (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* 17:148.
12. Zhang X-H, Tee LY, Wang X-G, Huang Q-S, Yang S-H (2015) Off-target effects in CRISPR/Cas9-mediated genome engineering. *Mol Ther Nucleic Acids* 4:e264.
13. Tsai SQ, et al. (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* 33:187–197.
14. Kim D, et al. (2015) Digenome-seq: Genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods* 12:237–243, 1.
15. Frock RL, et al. (2015) Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* 33:179–186.
16. Yan WX, et al. (2017) BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat Commun* 8:15058.
17. Tsai SQ, et al. (2017) CIRCLE-seq: A highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat Methods* 14:607–614.
18. Park J, et al. (2017) Digenome-seq web tool for profiling CRISPR specificity. *Nat Methods* 14:548–549.

19. Cameron P, et al. (2017) Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat Methods* 14:600–606.
20. Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526:68–74.
21. Canver MC, et al. (2017) Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nat Genet* 49: 625–634.
22. Paquet D, et al. (2016) Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* 533:125–129.
23. Sanjana NE, Shalem O, Zhang F (2014) Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* 11:783–784.
24. Hsu PD, et al. (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 31:827–832.
25. Doench JG, et al. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* 34:184–191.
26. Lek M, et al.; Exome Aggregation Consortium (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291.
27. Low-Kam C, et al. (2016) Whole-genome sequencing in French Canadians from Quebec. *Hum Genet* 135:1213–1221.
28. Hamosh A, et al. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 30:52–55.
29. Bosley KS, et al. (2015) CRISPR germline engineering—the community speaks. *Nat Biotechnol* 33:478–486.
30. Slaymaker IM, et al. (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science* 351:84–88.
31. Kleinstiver BP, et al. (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529:490–495.
32. Chen JS, et al. (2017) Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* 550:407–410.
33. Tycko J, Myer VE, Hsu PD (2016) Methods for optimizing CRISPR-Cas9 genome editing specificity. *Mol Cell* 63:355–370.
34. Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK (2014) Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* 32:279–284.
35. Guilinger JP, Thompson DB, Liu DR (2014) Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. *Nat Biotechnol* 32: 577–582.
36. Ran FA, et al. (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154:1380–1389.
37. Rose JC, et al. (2017) Rapidly inducible Cas9 and DSB-ddPCR to probe editing kinetics. *Nat Methods* 14:891–896.
38. Scott DA, Zhang F (2017) Implications of human genetic variation in CRISPR-based therapeutic genome editing. *Nat Med* 23:1095–1101.
39. Boyle EA, et al. (2017) High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc Natl Acad Sci USA* 114: 5461–5466.
40. Horlbeck MA, et al. (2016) Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife* 5:e12677.
41. Sternberg SH, LaFrance B, Kaplan M, Doudna JA (2015) Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* 527:110–113.
42. O'Huallachain M, Karczewski KJ, Weissman SM, Urban AE, Snyder MP (2012) Extensive genetic variation in somatic human tissues. *Proc Natl Acad Sci USA* 109: 18018–18023.
43. Milholland B, et al. (2017) Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* 8:15183.
44. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
45. Besenbacher S, et al. (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* 6:5969.
46. Conrad DF, et al.; 1000 Genomes Project (2011) Variation in genome-wide mutation rates within and between human families. *Nat Genet* 43:712–714.
47. Aguirre AJ, et al. (2016) Genomic copy number dictates a gene-independent cell response to CRISPR-Cas9 targeting. *Cancer Discov* 6:914–929.
48. Munoz DM, et al. (2016) CRISPR screens provide a comprehensive assessment of cancer vulnerabilities but generate false-positive hits for highly amplified genomic regions. *Cancer Discov* 6:900–913.
49. van Diemen FR, et al. (2016) CRISPR/Cas9-mediated genome editing of herpesviruses limits productive and latent infections. *PLoS Pathog* 12:e1005701.
50. Hu W, et al. (2014) RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. *Proc Natl Acad Sci USA* 111:11461–11466.
51. Liu X, Hao R, Chen S, Guo D, Chen Y (2015) Inhibition of hepatitis B virus by the CRISPR/Cas9 system via targeting the conserved regions of the viral genome. *J Gen Virol* 96:2252–2261.
52. Wollebo HS, et al. (2015) CRISPR/Cas9 system as an agent for eliminating polyomavirus JC infection. *PLoS One* 10:e0136046.
53. Fusi N, Smith I, Doench J, Listgarten J (2015) In silico predictive modeling of CRISPR/Cas9 guide efficiency. bioRxiv:10.1101/021568.
54. Chari R, Mali P, Moosburner M, Church GM (2015) Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* 12:823–826.
55. Xu H, et al. (2015) Sequence determinants of improved CRISPR sgRNA design. *Genome Res* 25:1147–1157.
56. Doench JG, et al. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol* 32:1262–1267.
57. Wang T, Wei JJ, Sabatini DM, Lander ES (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343:80–84.
58. Moreno-Mateos MA, et al. (2015) CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* 12:982–988.
59. Housden BE, et al. (2015) Identification of potential drug targets for tuberous sclerosis complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Sci Signal* 8:rs9.
60. Ren X, et al. (2014) Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in *Drosophila*. *Cell Rep* 9:1151–1162.
61. Farboud B, Meyer BJ (2015) Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. *Genetics* 199:959–971.
62. Bae S, Kweon J, Kim HS, Kim J-S (2014) Microhomology-based choice of Cas9 nuclease target sites. *Nat Methods* 11:705–706.
63. Park RJ, et al. (2017) A genome-wide CRISPR screen identifies a restricted set of HIV host dependency factors. *Nat Genet* 49:193–203.
64. Canver MC, et al. (2015) BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* 527:192–197.
65. Mandal PK, et al. (2014) Efficient ablation of genes in human hematopoietic stem and effector cells using CRISPR/Cas9. *Cell Stem Cell* 15:643–652.
66. Ruan GX, et al. (2017) CRISPR/Cas9-mediated genome editing as a therapeutic approach for Leber congenital amaurosis 10. *Mol Ther* 25:331–341.
67. Wilen CB, et al. (2011) Engineering HIV-resistant human $CD4^+$ T cells with CXCR4-specific zinc-finger nucleases. *PLoS Pathog* 7:e1002020.
68. Torikai H, et al. (2013) Toward eliminating HLA class I expression to generate universal cells from allogeneic donors. *Blood* 122:1341–1349.
69. Ding Q, et al. (2014) Permanent alteration of PCSK9 with in vivo CRISPR-Cas9 genome editing. *Circ Res* 115:488–492.
70. Beane JD, et al. (2015) Clinical scale zinc finger nuclease mediated gene editing of PD-1 in tumor infiltrating lymphocytes for the treatment of metastatic melanoma. *Mol Ther* 23:1380–1390.
71. Schumann K, et al. (2015) Generation of knock-in primary human T cells using Cas9 ribonucleoproteins. *Proc Natl Acad Sci USA* 112:10437–10442.
72. Fadel HJ, et al. (2014) TALEN knockout of the PSIP1 gene in human cells: Analyses of HIV-1 replication and allosteric integrase inhibitor mechanism. *J Virol* 88:9704–9717.
73. Torikai H, et al. (2012) A foundation for universal T-cell based immunotherapy: T cells engineered to express a CD19-specific chimeric-antigen-receptor and eliminate expression of endogenous TCR. *Blood* 119:5697–5705, and erratum (2015) 126:2527.
74. Provasi E, et al. (2012) Editing T cell specificity towards leukemia by zinc finger nucleases and lentiviral gene transfer. *Nat Med* 18:807–815.
75. Eyquem J, et al. (2017) Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. *Nature* 543:113–117.
76. Joglekar AV, et al. (2013) Integrase-defective lentiviral vectors as a delivery platform for targeted modification of adenosine deaminase locus. *Mol Ther* 21:1705–1717.
77. Sharma R, et al. (2015) In vivo genome editing of the albumin locus as a platform for protein replacement therapy. *Blood* 126:1777–1784.
78. Firth AL, et al. (2015) Functional gene correction for cystic fibrosis in lung epithelial cells generated from patient iPSCs. *Cell Rep* 12:1385–1390.
79. Sebastiano V, et al. (2014) Human COL7A1-corrected induced pluripotent stem cells for the treatment of recessive dystrophic epidermolysis bullosa. *Sci Transl Med* 6: 264ra163, and erratum (2014) 6:267er8.
80. De Ravin SS, et al. (2017) CRISPR-Cas9 gene repair of hematopoietic stem cells from patients with X-linked chronic granulomatous disease. *Sci Transl Med* 9:eaah3480.
81. Ousterout DG, et al. (2015) Multiplex CRISPR/Cas9-based genome editing for correction of dystrophin mutations that cause Duchenne muscular dystrophy. *Nat Commun* 6:6244.
82. Maggio I, Liu J, Janssen JM, Chen X, Gonçalves MAFV (2016) Adenoviral vectors encoding CRISPR/Cas9 multiplexes rescue dystrophin synthesis in unselected populations of DMD muscle cells. *Sci Rep* 6:37051.
83. Osborn MJ, et al. (2015) Fanconi anemia gene editing by the CRISPR/Cas9 system. *Hum Gene Ther* 26:114–126.
84. Li H, et al. (2011) In vivo genome editing restores haemostasis in a mouse model of haemophilia. *Nature* 475:217–221.
85. Yin H, et al. (2016) Therapeutic genome editing by combined viral and non-viral delivery of CRISPR system components in vivo. *Nat Biotechnol* 34:328–333.
86. DeWitt MA, et al. (2016) Selection-free genome editing of the sickle mutation in human adult hematopoietic stem/progenitor cells. *Sci Transl Med* 8:360ra134.
87. Dever DP, et al. (2016) CRISPR/Cas9 β-globin gene targeting in human haematopoietic stem cells. *Nature* 539:384–389.
88. Genovese P, et al. (2014) Targeted genome editing in human repopulating haematopoietic stem cells. *Nature* 510:235–240.
89. Yusa K, et al. (2011) Targeted gene correction of α1-antitrypsin deficiency in induced pluripotent stem cells. *Nature* 478:391–394.
90. Lin S-R, et al. (2014) The CRISPR/Cas9 system facilitates clearance of the pntrahepatic HBV templates in vivo. *Mol Ther Nucleic Acids* 3:e186.
91. Dong C, et al. (2015) Targeting hepatitis B virus cccDNA by CRISPR/Cas9 nuclease efficiently inhibits viral replication. *Antiviral Res* 118:110–117.
92. Ramanan V, et al. (2015) CRISPR/Cas9 cleavage of viral DNA efficiently suppresses hepatitis B virus. *Sci Rep* 5:10833.
93. Kennedy EM, et al. (2014) Inactivation of the human papillomavirus E6 or E7 gene in cervical carcinoma cells by using a bacterial CRISPR/Cas RNA-guided endonuclease. *J Virol* 88:11965–11972.