nature medicine

Implications of human genetic variation in CRISPR-based therapeutic genome editing

David A Scott[1–3] & Feng Zhang[1–4]

CRISPR–Cas genome-editing methods hold immense potential as therapeutic tools to fix disease-causing mutations at the level of DNA. In contrast to typical drug development strategies aimed at targets that are highly conserved among individual patients, treatment at the genomic level must contend with substantial inter-individual natural genetic variation. Here we analyze the recently released ExAC and 1000 Genomes data sets to determine how human genetic variation impacts target choice for Cas endonucleases in the context of therapeutic genome editing. We find that this genetic variation confounds the target sites of certain Cas endonucleases more than others, and we provide a compendium of guide RNAs predicted to have high efficacy in diverse patient populations. For further analysis, we focus on 12 therapeutically relevant genes and consider how genetic variation affects off-target candidates for these loci. Our analysis suggests that, in large populations of individuals, most candidate off-target sites will be rare, underscoring the need for prescreening of patients through whole-genome sequencing to ensure safety. This information can be integrated with empirical methods for guide RNA selection into a framework for designing CRISPR-based therapeutics that maximizes efficacy and safety across patient populations.

The development of CRISPR-based RNA-guided endonucleases, such as Cas9 and Cpf1, for eukaryotic genome editing has sparked intense interest in the use of this technology for therapeutic applications[1–3]. In contrast to small-molecule therapies, which target highly conserved active sites in proteins, therapies designed to target particular DNA sequences must take into account genetic variation among patient populations. If this variation disrupts the therapy target site, it can affect the efficacy of a CRISPR-based therapeutic; if it generates off-target candidate sites, it can affect the safety of a CRISPR-based therapeutic. Previously, it has been reported that genetic variation in cell lines can alter Cas9 targeting[4], but there has been limited effort to comprehensively and systematically evaluate this phenomenon in large human populations.

As CRISPR-based therapies advance toward human clinical trials, it is important to consider how natural genetic variation in the human population may affect the results from these trials and even patient safety. Recently, large-scale sequencing data sets from the Exome Aggregation Consortium (ExAC) and 1000 Genomes Project have provided an unprecedented view of the landscape of human genetic variation[5–8]. These data sets have captured nearly all common variants in the human population and contain deep coverage of rare variants[5,8], enabling evaluation of the effects of human variation on therapeutic genome editing in diverse human populations. Here we use these data sets to determine the impact of population genetic variation on therapeutic genome editing with *Streptococcus pyogenes* Cas9 (SpCas9), the SpCas9 variants VQR and VRER, *Staphylococcus aureus* Cas9 (SaCas9), and *Acidaminococcus* sp. Cpf1 (AsCpf1)[1–3,9,10]. We found extensive variation likely to impact the efficacy of these enzymes and propose that unique, patient-specific off-target candidates will be one of the main challenges in ensuring the safety of these therapeutics. These results provide a framework for designing CRISPR-based therapeutics, highlight the need to develop multiple guide RNA–enzyme pairs for each target locus, and suggest that pretherapeutic whole-genome sequencing will be required to ensure uniform efficacy and safety for treatment across patient populations.

RESULTS

Human genetic variation impacts choice of Cas enzyme

To date, two families of class 2 (single-effector) CRISPR nucleases, Cas9 and Cpf1, have been harnessed for eukaryotic genome editing[1–3,11]. Both Cas9 and Cpf1 are programmed by guide RNAs, which direct cleavage of DNA targets that are complementary to the guide RNA target and flanked by a short protospacer-adjacent motif (PAM) specific to each endonuclease[3,12] (**Fig. 1a**). Mismatches between the guide RNA and its DNA target have been shown to decrease RNA-guided endonuclease activity, and deviation from the canonical PAM sequence often completely abolishes nuclease activity[13–16]. The recently released ExAC data set, with variants from 60,706 individuals, contains on average one variant for every 8 nucleotides (nt) in the human exome[5]. To assess the impact of this variation on guide RNA efficacy, we used the ExAC data set to catalog variants present among all possible targets in the human reference exome that either (i) disrupt the target PAM sequence or (ii) introduce mismatches between the guide RNA and the genomic DNA, which we collectively term 'target variation' (**Fig. 1a**).

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to F.Z. (zhang@broadinstitute.org).

Received 21 November 2016; accepted 20 June 2017; published online 31 July 2017; doi:10.1038/nm.4377

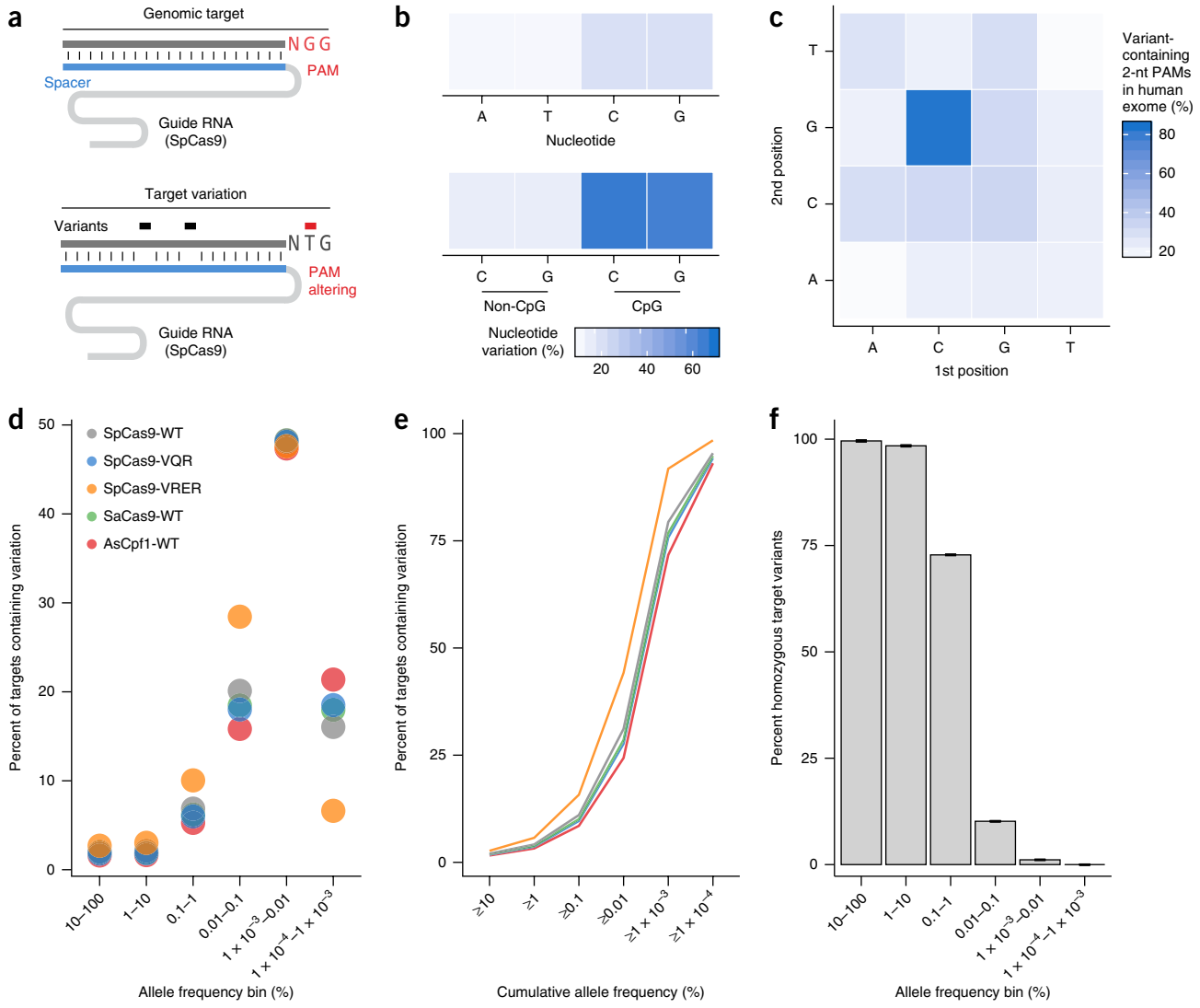

Figure 1 Human genetic variation substantially impacts the efficacy of RNA-guided endonucleases. (**a**) Schematic of a genomic target (consisting of a spacer and an adjacent PAM element) and guide RNA, with and without target variation. (**b**) Fraction of bases for individual nucleotides containing variation in the ExAC data set. (**c**) Fraction of 2-nt PAM motifs altered by variants in the ExAC data set. (**d**) Percentage of target sequences with variation at different allele frequencies for each CRISPR endonuclease. (**e**) Cumulative percentage of target sequences containing variants greater than or equal to each allele frequency for each CRISPR endonuclease. (**f**) Fraction of target sequences containing homozygous variants at different allele frequencies. The mean and s.e.m. is shown for five CRISPR endonucleases.

In addition to two orthologs of Cas9 (SpCas9 and SaCas9) and Cpf1, a number of SpCas9 variants have been engineered as tools for genome editing, with each using a different PAM[1–3,9,10] (**Table 1**). Consideration of multiple enzymes with different PAM requirements will increase the number of available genomic targets for therapeutic loci. We therefore assessed variation at each PAM in the human exome for SpCas9-WT (wild type; 5′-NGG-3′), SpCas9-VQR (5′-NGA-3′), SpCas9-VRER (5′-NGCG-3′), SaCas9-WT (5′-NNGRRT-3′), and AsCpf1-WT (5′-TTTN-3′) (where N is any nucleotide and R is A or G)—all of which are currently being considered as candidate enzymes for development as CRISPR-based therapeutics. The recently reported enhanced-specificity Cas9 (eSpCas9) and high-fidelity SpCas9 (SpCas9-HF) have the same 5′-NGG-3′ PAM as SpCas9-WT and are thus not considered separately here[17,18]. For each nuclease, we determined the fraction of exonic PAMs containing variants that alter PAM recognition either through abolishing the existing PAM or creating a new one relative to the reference genome. In the ExAC population, the total fraction of targets containing PAM-altering variants was similar for all enzymes (21–35%) except SpCas9-VRER, for which 80% of targets were affected by PAM-altering variants (**Table 1** and **Supplementary Fig. 1**). The PAM for SpCas9-VRER contains a CpG motif, which has been shown to be highly mutable[5]. In accordance with these results, we found that CG was the most highly mutable 2-nt PAM motif in the human exome, and 66% of cytosine and guanine nucleotides present in CpG motifs showed variation in the 60,706 ExAC individuals[5] (**Fig. 1b,c** and **Supplementary Table 1**). These results suggest that enzymes using PAMs containing CG motifs are considerably more affected by target variation in the human genome.

Low-variation regions of the human exome are more reliably targeted

We extended our analysis to determine the fraction of all possible targets of SpCas9-WT, SpCas9-VQR, SaCas9-WT, and AsCpf1-WT in

Table 1 Fraction of targets containing PAM-altering variants for five Cas enzymes

			Whole-exome PAM variation by allele frequency (%)						
Protein	PAM	Orientation	≥10%	≥1	≥0.1	≥0.01	≥0.001	Total	*n*
AsCpf1-WT	TTTN	Left	0.15	0.26	0.61	1.81	8.91	21.04	2,702,056
SpCas9-VQR	NGA	Right	0.11	0.25	0.69	2.28	11.39	23.19	9,838,603
SpCas9-WT	NGG	Right	0.16	0.37	1.13	3.82	17.46	32.61	10,286,445
SaCas9-WT	NNGRRT	Right	0.23	0.44	1.16	3.68	17.29	34.68	1,938,911
SpCas9-VRER	NGCG	Right	0.77	1.86	5.79	20.72	66.67	80.16	981,524

n specifies the number of protein-coding targets in the human exome for each enzyme. Orientation refers to the location of the PAM relative to the guide-RNA-complementary region of the target, and allele frequencies refer to the percentage of the ExAC population containing a particular PAM-altering variant for a given target in the human exome.

the human exome that contain variants. We found that 93–95% of targets contained variants in the ExAC data set that are likely to alter the efficiency of target cleavage (**Fig. 1d**,**e** and **Supplementary Table 2**), and most target variation occurring at frequencies of <0.1% was heterozygous (**Fig. 1d**,**f** and **Supplementary Table 2**).

The ExAC data set is large enough that it provides near-comprehensive coverage of variants at allele frequencies of ≥0.01% in the population (i.e., variation that will exist in at least 1 out of 10,000 alleles in the population)[5]. Hence, we used this data set to compile a list of exome-wide target sites for SpCas9-WT, Cas9-VQR, SaCas9-WT, and AsCpf1-WT lacking variants occurring at an allele frequency of ≥0.01% (referred to as 'platinum' targets; whole-exome platinum targets for each enzyme are given in **Supplementary Data**). These platinum targets should be efficacious in ≥99.99% of the population (**Fig. 2**; target variation <0.01%).

For further analysis, we selected 12 therapeutically relevant genes, including those that are currently the focus of therapeutic development: *CEP290*, *CFTR*, *DMD*, *G6PC*, *HBB*, *IDUA*, *IL2RG*, *PCSK9*, *PDCD1*, *SERPINA1*, *TTR*, and *VEGFA* (**Supplementary Fig. 2**; platinum targets for each enzyme for these 12 genes are given in **Supplementary Data**). For these 12 genes, approximately two-thirds of the possible exonic targets met our platinum criteria, with *PCSK9* containing the smallest fraction of platinum targets (50%) (**Supplementary Table 3**). This finding suggests that, for most genomic regions, ample platinum targets will exist that can be considered when beginning the process of therapeutic target selection.

We observed that both high-variation targets and platinum targets clustered along exons for each of the 12 genes examined. For example, all targets in the 5′ half of *PCSK9* exon 4 were platinum, whereas very few platinum targets existed for exon 5 (**Fig. 2c**). However, even for regions of high-frequency variation, such as *PCSK9* exons 1–4, it was still possible to find small numbers of platinum targets for some enzymes (**Fig. 2c**). This observation for *PCSK9* is representative of the findings for the other genes investigated in this study and suggests that considering multiple enzymes with distinct PAM requirements increases the likelihood of finding a platinum target. In the event that a genomic region of interest contains variation that cannot be avoided, it will be necessary to design multiple guide RNAs, each tailored to accommodate the presence of high-frequency (≥0.01% allele frequency) variants.

Low-frequency off-target candidates predominate in large populations

A second major consideration in CRISPR-based therapeutics is safety, which can be improved by designing guide RNAs with minimal potential off-target activity. Unbiased investigation of genome-wide CRISPR nuclease activity suggests that most off-target activity occurs at loci with ≤3 mismatches with respect to the guide RNA sequence[9,13,19–24]. Current approaches for Cas9 target selection rank off-target candidates found in the reference human genome by both the number and position of guide RNA mismatches, under the assumption that loci containing ≤3 mismatches or containing PAM-distal mismatches are more likely to be cleaved[13–15]. However, in a population of individuals, this strategy is complicated by the existence of multiple haplotypes (sets of variants that co-occur), which will have different positions or numbers of mismatches at candidate off-target sites (**Fig. 3a**). To assess the predicted safety of a guide RNA within a population, we used the 1000 Genomes data set, which includes phased single-nucleotide variant calls for 2,504 individuals[8]. From these data, we reconstructed allele-specific whole-genome sequences for each individual. In contrast with the much larger ExAC data set, which collapses all variants, the 1000 Genomes data set contains information about haplotypes, which enabled us to identify off-target sites in the population arising from single or multiple variants in an individual haplotype. For platinum targets in the 12 genes considered here, we quantified off-target candidates (defined as genomic loci with ≤3 mismatches with respect to a given guide RNA) arising from all 1000 Genomes haplotypes (**Supplementary Data**). In this relatively small population of 2,504 individuals, more than half of the haplotypes containing off-target candidates were present in ≥10% of individuals (**Fig. 3b**). However, for haplotypes present in <10% of individuals, the number of off-target candidates for each guide RNA increased with decreasing haplotype frequency (**Fig. 3b**). This trend indicates that, for large populations, most unique off-target candidates for a given guide RNA will differ between individuals, as shown by the rise in the cumulative number of off-target candidates for an individual guide RNA accompanying decreasing allele frequency (**Fig. 3c**).

Avoiding high-frequency off-target candidates should maximize population safety

For individual guide RNAs in the 12 genes we analyzed, we found that the number of off-target candidates for SpCas9-WT, SpCas9-VQR, SaCas9-WT, and AsCpf1-WT varied from 0 to greater than 10,000 in the 1000 Genomes population (**Fig. 3d**). Much of this disparity reflects how unique or repetitive an individual target sequence is within the human genome. For instance, SaCas9-WT, which has a longer PAM and hence fewer genomic targets, had on average fewer off-target candidates per guide RNA. Of the 12 genes we considered, some contain more repetitive regions relative to the rest of the human genome, as reflected by increased numbers of targets with high off-target candidate counts (**Fig. 4a**). For example, within *PCSK9* exons 2–5, we observed that platinum targets with high and low numbers of off-target candidates tended to cluster in regions of sequence that were repetitive and unique within the genome, respectively (**Fig. 4b**). This pattern held true for all 12 genes studied. Interestingly, within repetitive regions of exons, we did identify small numbers of platinum targets with substantially reduced quantities of

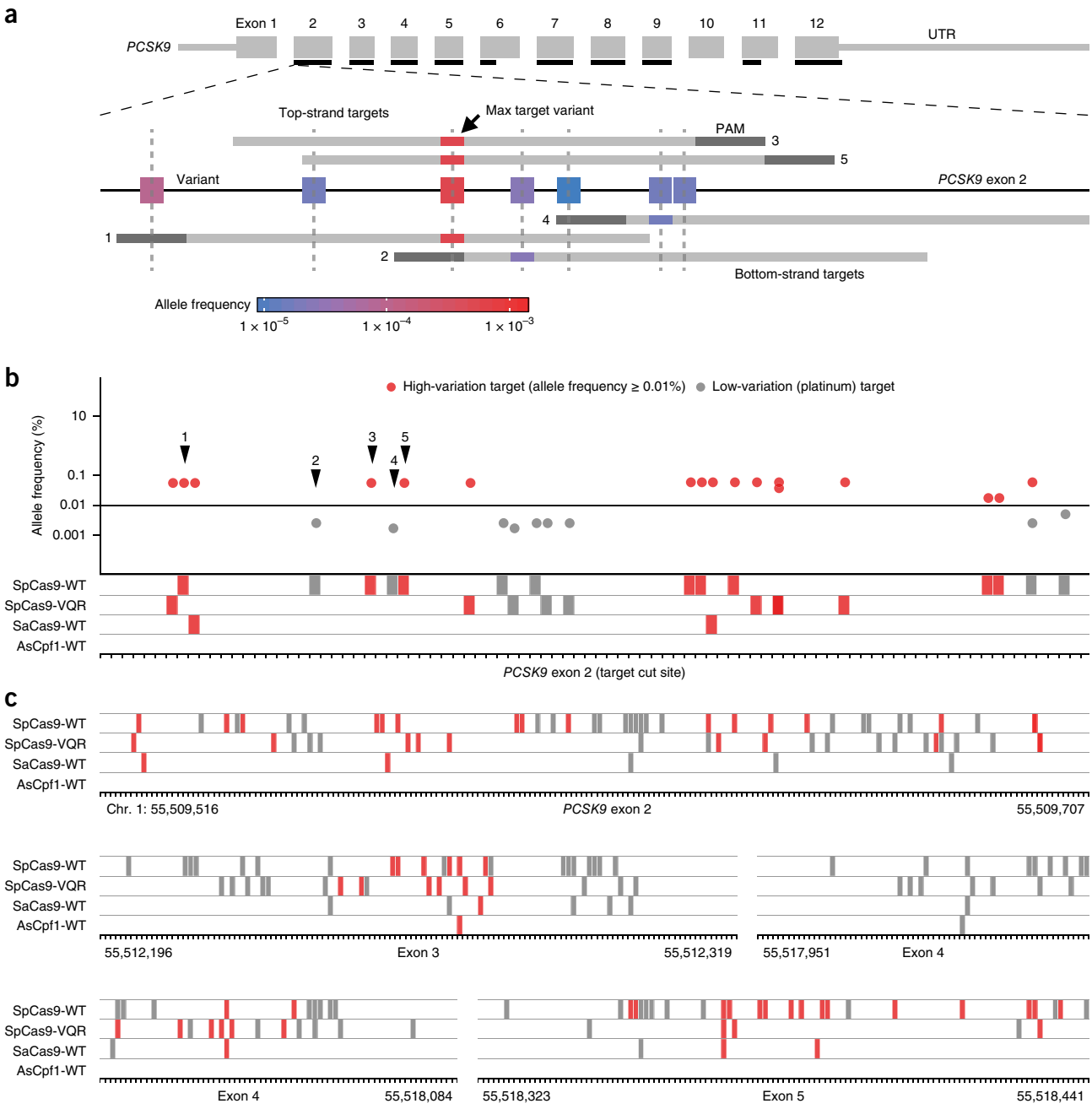

Figure 2 Selection of platinum targets maximizes population efficacy. (**a**) Schematic showing variation in target sequences within exon 2 of *PCSK9*, with regions of high coverage in the ExAC data set indicated by black lines below the exons. Variants for a short region of *PCSK9* exon 2 are highlighted along with five targets for SpCas9-WT. Top-strand targets (PAM on the right) are shown above the region of *PCSK9* exon 2, and bottom-strand targets (PAM on the left) are shown below. The frequency of the highest frequency variant in the ExAC data set intersecting each target sequence is indicated by color and is used as the target variation frequency. Variants that do not affect target recognition by the endonuclease (such as a high-frequency variant intersecting the N in the 5′-NGG-3′ PAM of target 2) do not affect targeting efficiency and have been excluded. Targets 2 and 4 have the lowest variation of the five targets shown. (**b**) Frequency of target variation plotted by cut-site position for target sequences spanning the start of *PCSK9-001* exon 2, with the five targets shown in **a** indicated by arrows. The horizontal line at 0.01% separates platinum targets (gray) from targets with high variation (red). The classification for each target is depicted below for each enzyme (gray or red boxes). (**c**) Classification (same colors as in **b**) for each enzyme of targets spanning exons 2–5 of the *PCSK9-001* isoform.

off-target candidates. These findings further support the notion that using multiple enzymes with distinct PAM requirements should enhance both safety and efficacy by increasing the number of available targets for therapeutically relevant genomic loci.

Additionally, in a population, the number of off-target candidates at a given locus is compounded by the existence of multiple haplotypes, the number of which will increase with the size of the population. Hence, for each off-target candidate present in a high-frequency haplotype,

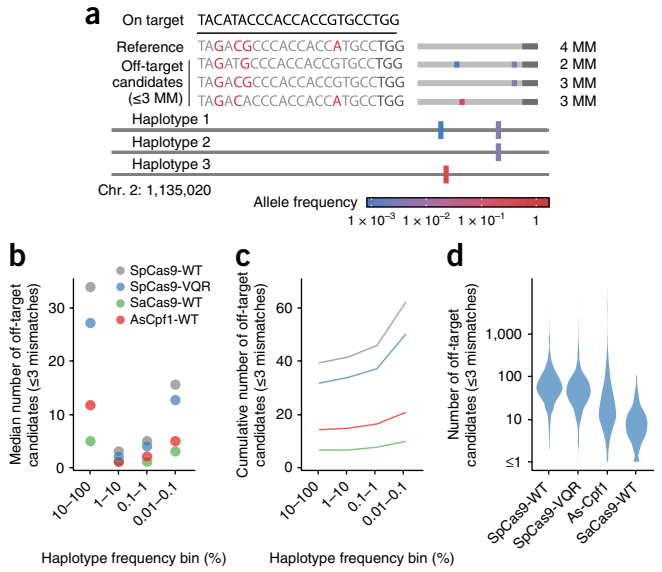

Figure 3 Human genetic variation substantially impacts the safety of CRISPR endonuclease therapeutics. (**a**) Schematic of off-target candidates arising owing to multiple different haplotypes. MM, mismatches. (**b**) Number of off-target candidates present in the 1000 Genomes data set for each CRISPR endonuclease for the 12 therapeutically relevant genes at different allele frequencies. (**c**) Distribution of the number of off-target candidates per platinum target for each CRISPR endonuclease. (**d**) Distribution of the number of off-target candidates per enzyme for the four CRISPR endonucleases studied here.

in a large population multiple lower-frequency haplotypes are likely to exist that may lead to different gene-editing outcomes. Thus, minimizing the number of off-target candidates occurring in high-frequency haplotypes is of critical importance for the selection of therapeutic guide RNAs. The current 1000 Genomes data set provides comprehensive coverage of alleles occurring at a frequency of up to 0.1% in the population (considered to be the lower bound for high-frequency variants), allowing the identification of platinum targets with minimal off-target candidates in high-frequency haplotypes in the human population[5,8]. Use of the enhanced-specificity enzymes eSpCas9 and Cas9-HF1 will further reduce the likelihood of cleavage at off-target candidate sites, but it will still remain important to avoid target regions that are repetitive or have off-target candidates in high-frequency haplotypes[17,18].

Consideration of patient populations

Genome-editing therapies are currently being designed for a range of applications, including treatment of rare genetic diseases (for example, Leber's congenital amaurosis) and common conditions (for example, high cholesterol), and therapeutic augmentation whereby a genetic change increases the efficacy of a treatment (for example, *PDCD1* knockout for enhanced immunotherapy). Each of these applications will have a unique patient population with its own landscape of genetic variation, and this can be considered when choosing therapeutic targets. For example, Tay–Sachs disease occurs in Ashkenazi populations at more than ten times the rate that it occurs in the general population[25]; because of the shared genetic heritage among people with Tay–Sachs disease, there will be fewer variants in the population of affected individuals and those that are present will occur at a higher frequency. On the other hand, populations of patients who have diverse genetic backgrounds will contain large numbers of variants that occur at high frequencies in a subset of the population but at low frequencies overall. The 1000 Genomes Project provides demographic information, including sex and ancestry, for each individual, so we used these data to explore how much off-target candidate variation for a given individual was explained by population demographics. For all off-target candidates for the guide RNAs targeting the 12 genes considered here, we performed principal-component analysis (PCA) and found that the first five principal components separated individuals most effectively by continent but also by subcontinent and sex (**Fig. 4c** and **Supplementary Figs. 3–5**). We found that these first five principal components accounted for 12% of the variation in off-target candidates occurring at a frequency of <100% among members of the population, indicating that the safety and efficacy of therapeutics can be enhanced by designing therapeutic targets for subpopulations of patients with specific variants.

DISCUSSION

Ideally, personalized genomic medicine would utilize tailored RNA-guided endonuclease therapeutics for each patient. However, in most cases, the cost and time required to obtain regulatory approval for each individualized therapeutic would be prohibitive given the current regulatory framework. Instead, a small number of carefully chosen enzyme–guide combinations may be developed and tested to provide a suite of potential therapeutics for a particular patient population. Current methods for selecting targets and guides typically rely solely on sequence information from the human reference genome and criteria obtained from empirical tests of efficacy. However, although guide RNA efficacy does vary and can be difficult to predict, if therapeutic targets are selected on the basis of efficacy alone, those therapies run the risk of being mired in a clinical trial with confounding results and/or undesirable outcomes due to human genetic variation.

Our findings regarding the impact of genetic variation on Cas endonuclease activity can be integrated with empirical methods to streamline the design and testing of genome-editing therapeutics in a consolidated framework (**Fig. 4d**). First, when possible, regions of low variation should be targeted, which will ensure maximal efficacy across a patient population with diverse genetic backgrounds. Second, guide RNAs need to be selected to minimize the number of off-target candidates occurring on high-frequency haplotypes in the patient population to reduce the likelihood of off-target mutations resulting in oncogenic events or undesirable side effects. Third, assessing the amount of low-frequency variation present in the patient population can be helpful for estimating the number of guide RNA–enzyme combinations required to effectively and safely treat the anticipated patient population. This will be particularly important when designing targets for use within specific populations. For example, for treatment of common diseases, more guide RNA–enzyme combinations will need to be developed given the breadth of the natural genetic variation in the patient population. Fourth, *in silico* screening and empirical assays[9,19–21,26–28] to assess target efficacy and genome-wide specificity should be used to identify the optimal guide RNA–enzyme combinations from the pool of selected guide RNAs. In the event that no high-efficacy guides are found, the number of guide RNA–enzyme combinations should be increased and tailored to the presence of multiple independent high-frequency haplotypes. The safety of selected combinations of guide RNA and enzyme should be evaluated through unbiased whole-genome off-target detection in relevant cell lines (ideally patient specific). Combinations that pass all of these filters should then be moved forward for regulatory approval. Finally,

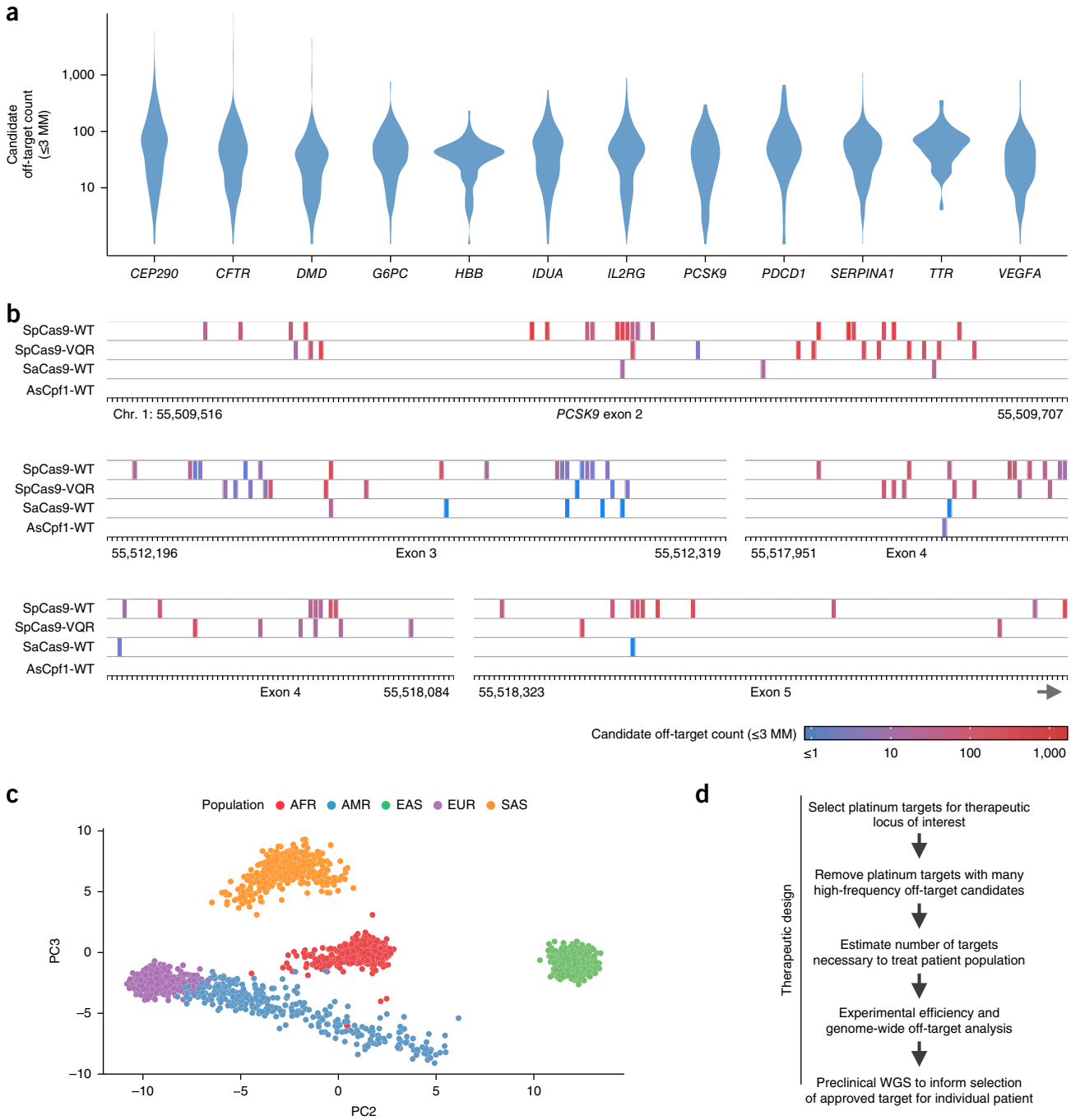

Figure 4 Gene- and population-specific variation inform therapeutic design. (**a**) Distribution of the number of off-target candidates per platinum target for the 12 therapeutically relevant genes studied here. (**b**) Total off-target candidates for platinum targets spanning exons 2–5 of *PCSK9-001* are shown for each CRISPR endonuclease. (**c**) PCA separating 1000 Genomes individuals (each represented by a dot) into superpopulations on the basis of individual-specific off-target profiles for platinum targets spanning the 12 therapeutically relevant genes. PC2 and PC3 are shown. AFR, African; AMR, admixed American; EAS, East Asian; EUR, European; SAS, South Asian. (**d**) Proposed framework for identifying therapeutic guides that maximize efficacy and safety. WGS, whole-genome sequencing.

pretherapeutic whole-genome sequencing of individual patients will be needed to select a single approved guide RNA–enzyme combination for treatment that is a perfect match to the patient's genome and free of patient-specific off-target candidates.

The selection of specific targets to pursue for therapeutic development will also depend on the type of gene edit desired. For example, gene-knockout strategies (which are being pursued using *PDCD1* for immunotherapy and *PCSK9* for treatment of cardiovascular disease) have many guide choices, and researchers can choose from a range of low-frequency target regions within or near the gene of interest and then select the guide within that subset that provides the most efficient gene knockdown. Other diseases can be addressed by removal of single or multiple causal variants, and therapies are being developed that aim to remove mutated segments of genes to restore function

(for example, *CEP290* for Leber's congenital amaurosis and *DMD* for Duchenne muscular dystrophy). While this strategy for therapeutic intervention also affords some flexibility in target selection, researchers will be limited to working with a single enzyme because two guide RNAs are needed for each gene, reducing the number of target options; this strategy carries the risk of doubling the number of potential off-target candidates as well. Finally, homology-directed repair (HDR) is being used for correction of disease-causing mutations affecting mitotically active cells in the body (such as *SERPINA1* for alpha-1 antitrypsin and *CFTR* for cystic fibrosis). In HDR strategies, Cas nucleases are used to cleave the target gene typically within 10–20 nt of the desired integration site, greatly restricting the targetable range. However, considering SpCas9, SpCas9-VQR, SaCas9, and AsCpf1, a target is present at approximately every 4 nt in the human exome, which should allow selection of a low-variation region even in situations with a narrow target range, which would be required by HDR.

Continued technological development will deliver more powerful and precise systems for therapeutic genome editing. Beyond nuclease-based strategies, new approaches that leverage the programmable DNA-binding activity of CRISPR-based enzymes to direct DNA base–modifying enzymes, such as base editing[29], also promises to further expand therapeutic options. Finally, well-designed clinical trials that are carried out efficiently and smoothly will be central to addressing the regulatory and ethical challenges facing therapeutic genome editing. Failure to anticipate the genetic diversity in patient populations will confound clinical trials and may lead to adverse outcomes. Our analysis of the impact of human genetic variation on CRISPR–Cas-based therapeutics provides a toolset and resources that will increase the efficacy and safety of these therapies, ultimately moving them more quickly toward the clinic.

URLs. ExAC, http://exac.broadinstitute.org/; 1000 Genomes Project, http://www.internationalgenome.org/.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank R. Macrae, L. Francioli, S. Jones, J. Strecker, D. Cox, I. Slaymaker, and W. Yan for helpful discussions and insights. F.Z. is a New York Stem Cell Foundation–Robertson Investigator. F.Z. is supported by the US National Institutes of Health through the National Institute of Mental Health (5DP1-MH100706 and 1R01-MH110049); the National Science Foundation; the New York Stem Cell Foundation; the Howard Hughes Medical Institute; the Simons Foundation; the Paul G. Allen Family Foundation; the Vallee Foundation; the Skoltech–MIT Next-Generation Program; James and Patricia Poitras; Robert Metcalfe; and David Cheng. The computer code and resources related to this work are available through the Zhang laboratory website (http://www.genome-engineering.org/) and GitHub (http://github.com/fengzhanglab).

AUTHOR CONTRIBUTIONS

D.A.S. and F.Z. conceived the study; D.A.S. performed all experiments and analyses; D.A.S. and F.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
2. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
3. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR–Cas system. *Cell* **163**, 759–771 (2015).
4. Yang, L. *et al.* Targeted and genome-wide sequencing reveal single nucleotide variations impacting specificity of Cas9 in human stem cells. *Nat. Commun.* **5**, 5507 (2014).
5. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
6. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
7. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
8. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
9. Ran, F.A. *et al. In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186–191 (2015).
10. Kleinstiver, B.P. *et al.* Engineered CRISPR–Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
11. Makarova, K.S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
12. Garneau, J.E. *et al.* The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67–71 (2010).
13. Hsu, P.D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).
14. Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat. Biotechnol.* **31**, 839–843 (2013).
15. Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR–Cas nucleases in human cells. *Nat. Biotechnol.* **31**, 822–826 (2013).
16. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L.A. RNA-guided editing of bacterial genomes using CRISPR–Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).
17. Slaymaker, I.M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
18. Kleinstiver, B.P. *et al.* High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
19. Tsai, S.Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR–Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
20. Frock, R.L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.* **33**, 179–186 (2015).
21. Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR–Cas9 off-target effects in human cells. *Nat. Methods* **12**, 237–243, 1, 243 (2015).
22. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
23. Kleinstiver, B.P. *et al.* Genome-wide specificities of CRISPR–Cas Cpf1 nucleases in human cells. *Nat. Biotechnol.* **34**, 869–874 (2016).
24. Kim, D. *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat. Biotechnol.* **34**, 863–868 (2016).
25. Lu, Y.-F., Goldstein, D.B., Angrist, M. & Cavalleri, G. Personalized medicine and human genetic diversity. *Cold Spring Harb. Perspect. Med.* **4**, a008581 (2014).
26. Cameron, P. *et al.* SITE-Seq: a genome-wide method to measure Cas9 cleavage. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.043 (2017).
27. Tsai, S.Q. *et al.* CIRCLE-seq: a highly sensitive *in vitro* screen for genome-wide CRISPR–Cas9 nuclease off-targets. *Nat. Methods* **14**, 607–614 (2017).
28. Yan, W.X. *et al.* BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.* **8**, 15058 (2017).
29. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. & Liu, D.R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).

ONLINE METHODS

Data sets. Our target variation analysis was performed using the Exome Aggregation Consortium (ExAC) data set from 60,706 globally diverse individuals[5].

Our investigation of off-target candidates was performed using the 1000 Genomes Project phase 3 data set containing phased whole-genome sequences from 2,504 globally diverse individuals[8].

Whole-exome target variation analysis. We included all targets containing the canonical protospacer adjacent motif (PAM) sequences for the CRISPR enzymes SpCas9-WT (5′-NGG-3′), SpCas9-VQR (5′-NGA-3′), SpCas9-VRER (5′-NGCG-3′), SaCas9 (5′-NNGRRT-3′), and AsCpf1 (5′-TTTN-3′) for all exons in GENCODE release 19 (GRCh37.p13) annotated as protein-coding and having an average coverage of at least 20 reads per ExAC sample (**Supplementary Table 2**). For analysis of variation in these targets, we included all missense or synonymous variants passing quality filtering, as described previously[5], in the ExAC data set. Because the publicly available ExAC data set includes only summary information for each variant, it was not possible to determine whether multiple variants occurring in a single genomic target occur in different haplotypes. Hence, we calculated target variation frequency as the maximum frequency of variants in an individual target. While this approach accurately approximates the variation of most targets in the population, it does underestimate the variation frequency for rare targets containing multiple high-frequency variants existing on separate haplotypes. Platinum targets were defined as those with a maximum variant frequency of <0.01% in the ExAC population.

Off-target candidate analysis. Phased haplotypes included in the 1000 Genomes phase 3 data set were used to create whole-genome allele-specific references for 2,504 individuals. We included in our analysis all single-nucleotide polymorphisms passing quality filtering in the 1000 Genomes phase 3 data set, as described previously[8]. Up to 100 protein-coding platinum targets for each therapeutically relevant gene (as available; **Supplementary Table 3**), including *CEP290*, *CFTR*, *DMD*, *G6PC*, *HBB*, *IDUA*, *IL2RG*, *PCSK9*, *PDCD1*, *SERPINA1*, *TTR*, and *VEGFA*, were selected for the proteins SpCas9-WT, SpCas9-VQR, SaCas9, and AsCpf1. Targets for each gene were searched against the references for each of the 2,504 individuals included in the 1000 Genomes Project to profile candidate off-target sites specific to each individual. All PAM sequences associated with nuclease activity were included in the off-target analysis for each enzyme as follows: SpCas9-WT (5′-NGG-3′, 5′-NAG-3′), SpCas9-VQR (5′-NGAN-3′, 5′-NGNG-3′), SaCas9 (5′-NNGRRT-3′), and AsCpf1 (5′-TTTN-3′). For the purpose of this study, off-target candidates are defined as unintended genome-wide targets for a specific guide RNA–enzyme combination with ≤3 mismatches with the guide RNA protospacer.

Demographics analysis. We performed a principal-component analysis (PCA) using the 1000 Genomes data set, taking into account for each individual the presence or absence of off-target candidates for each target included in our analysis of the 12 therapeutically relevant genes present in less than 100% of the individuals comprising the 1000 Genomes data set (n = 46,362 off-target candidates; PCA computed using the R prcomp function). Superpopulation groups included the following: AFR, African; AMR, admixed American; EAS, East Asian; EUR, European; SAS, South Asian. Population groups included the following: CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; CHS, Southern Han Chinese; CDX, Chinese Dai in Xishuangbanna, China; KHV, Kinh in Ho Chi Minh City, Vietnam; CEU, Utah residents (CEPH) with northern and western European ancestry; TSI, Toscani in Italia; FIN, Finnish in Finland; GBR, British in England and Scotland; IBS, Iberian population in Spain; YRI, Yoruba in Ibadan, Nigeria; LWK, Luhya in Webuye, Kenya; GWD, Gambian in Western Divisions in The Gambia; MSL, Mende in Sierra Leone; ESN, Esan in Nigeria; ASW, Americans of African ancestry in the southwestern United States; ACB, African Caribbeans in Barbados; MXL, Mexican ancestry from Los Angeles, United States; PUR, Puerto Ricans from Puerto Rico; CLM, Colombians from Medellín, Colombia; PEL, Peruvians from Lima, Peru; GIH, Gujarati Indian from Houston, Texas; PJL, Punjabi from Lahore, Pakistan; BEB, Bengali from Bangladesh; STU, Sri Lankan Tamil from the UK; ITU, Indian Telugu from the UK.

A **Life Sciences Reporting Summary** for this paper is available.

Data availability. Tables including all platinum targets in the human exome are freely available as a supplement to this manuscript (**Supplementary Data**).

All computer code used in this work is freely available from https://github.com/fengzhanglab/CRISPR-Human_Variation_Nature_Medicine_manuscript. The ExAC data were downloaded from the following ftp site: ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3.1. The 1000 Genomes data were acquired from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502.

doi:10.1038/nm.4377

nature research

Corresponding Author: Zhang, Feng

Date: Jun 7, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Whole exome sequencing data from all 60,706 individuals in the Exosome Aggregation Consortium (ExAC; http://exac.broadinstitute.org) and whole genome sequencing data from all 2,504 individuals in the 1000 Genomes dataset (http://www.internationalgenome.org) were used in the preparation of this work.

2. Data exclusions

   Describe any data exclusions.

   No data exclusions were made in the preparation of this work.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   This analysis using the ExAC and 1000 Genomes datasets is reproducible using code available at: https://github.com/fengzhanglab/CRISPR-Human_Variation_Nature_Medicine_manuscript.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   The complete ExAC and 1000 Genomes populations were used for these analyses without subdivision into experimental groups.

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   No blinding was necessary for the preparation of this work.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

   For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or the Methods section if additional space is needed).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒	A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly.
☒	☐	A statement indicating how many times each experiment was replicated
☐	☒	The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒	A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒	The test results (e.g. *p* values) given as exact values whenever possible and with confidence intervals noted
☐	☒	A summary of the descriptive statistics, including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒	Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Nature Medicine: doi:10.1038/nm.4377

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Code for this study is publicly available at: https://github.com/fengzhanglab/CRISPR-Human_Variation_Nature_Medicine_manuscript.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	Code for this study is publicly available at: https://github.com/fengzhanglab/CRISPR-Human_Variation_Nature_Medicine_manuscript, and all supplementary tables are made available online as supplements included with the manuscript.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	N/A

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	N/A

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	Details on the ExAC population are available at: http://exac.broadinstitute.org, and for the 1000 Genomes dataset at: http://www.internationalgenome.org.

Nature Medicine: doi:10.1038/nm.4377