Leveraging genetic variants for personalized and allele-specific sgRNA design

Kathleen C. Keough[1,2], Svetlana Lyalina[2,3], Michael P. Olvera[2], Sean Whalen[2], Bruce R. Conklin[2,4], Katherine S. Pollard[2,5,6]
[1]*Pharmaceutical Sciences and Pharmacogenomics Graduate Program at the University of California, San Francisco, California, USA*
[2]*Gladstone Institutes, San Francisco, California, USA*
[3]*Bioinformatics Graduate Program at the University of California, San Francisco, California, USA*
[4]*Departments of Biostatistics, Medicine, Ophthalmology and Pharmacology, University of California, San Francisco, California, USA*
[5]*Department of Epidemiology & Biostatistics, Institute for Human Genetics, Quantitative Biology Institute, and Institute for Computational Health Sciences, University of California, San Francisco, California, USA*
[6]*Chan Zuckerberg Biohub, California, USA*

Abstract

The CRISPR/Cas system is a highly specific genome editing tool capable of distinguishing alleles differing by even a single base pair. However, current tools only design guides (sgRNAs) for a reference genome, not taking into account individual variants which may generate, remove, or modify CRISPR/Cas sgRNA sites. Here we describe ExcisionFinder, a tool for designing personalized and allele-specific sgRNAs for genome editing. We leverage >2,500 human genomes to identify optimized pairs of sgRNAs that can be used for human therapeutic editing in large populations in the future.

Main Text

The CRISPR/Cas genome-editing system is highly specific, with the ability to discriminate between similar genomic sites, even alleles, based on a single nucleotide difference[1]. In order to target a genomic region with the CRISPR system, a single-guide RNA (sgRNA) must be designed that is specific to the region of interest. While current sgRNA design tools incorporate various data relating to predicted efficiency and specificity such as epigenetic marks and chromatin accessibility[2–4], sgRNAs are typically designed using reference genomes (GRCh38 for human or GRCm38 for mouse). Since sgRNAs are often used on cell lines or organisms with many nucleotide differences from the reference (e.g., on average 0.1% of a human genome[5]), ignoring allelic variation can impact sgRNA efficacy and specificity[6,7]. To address this challenge, we developed ExcisionFinder, a software tool that designs personalized and allele-specific sgRNAs for individual genomes, identifies pairs of sgRNAs to generate excisions likely to block expression of a gene, and leverages patterns of shared variation from >2,500 human genomes to design sgRNA pairs for that will have the greatest utility in a target population.

Incorporating genetic variation into sgRNA design enables allele-specific and personalized CRISPR experiments. Personalized design involves accounting for variants that disrupt or generate sgRNA sites in a given genome. When the goal is to target both alleles, homozygous non-reference variants change the set of sgRNA sites, while heterozygous variants may prevent one allele from being targeted. Allele-specific editing involves leveraging heterozygous variants in sgRNA design and harnessing the specificity of the CRISPR system to intentionally target only one allele. This establishes new therapeutic possibilities, including genome surgery to treat dominant negative disease by excising only the disease causing copy of a gene, an approach which rescues healthy phenotypes in cell and animal models of dominant negative diseases including Huntington's disease[8] and retinitis pigmentosa[9,10].

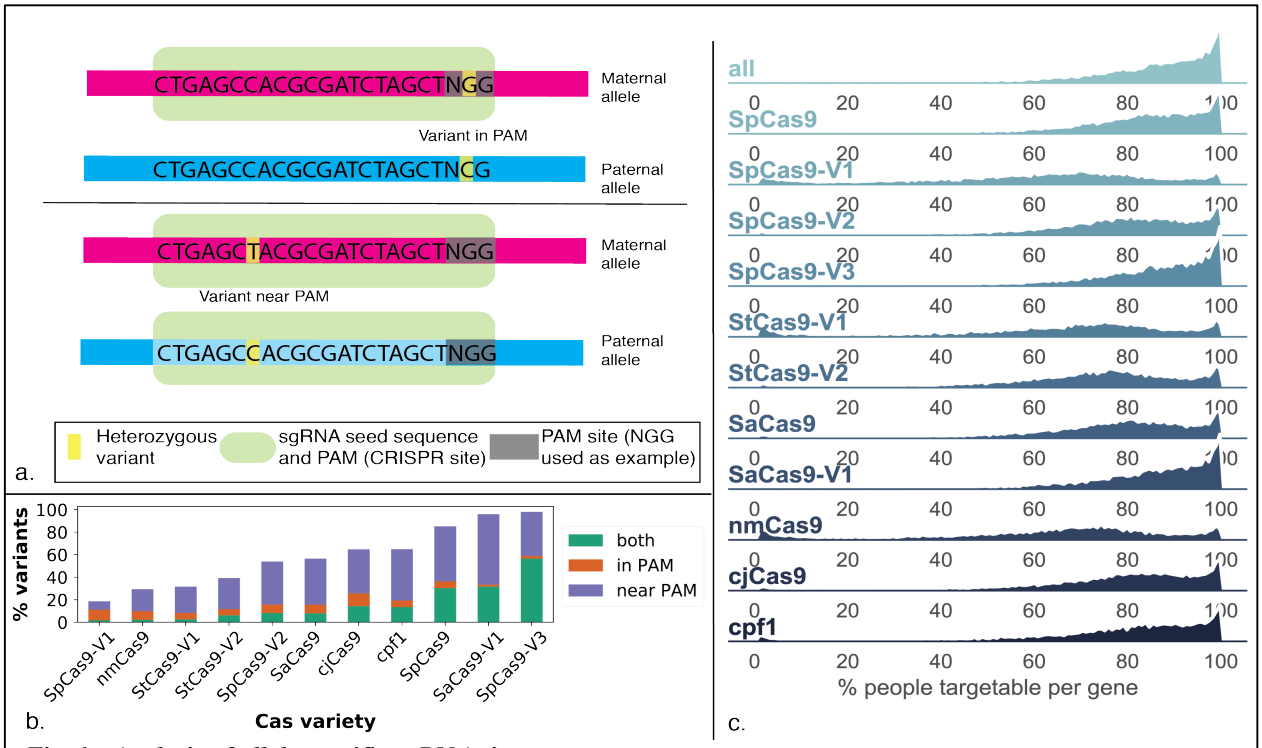


Fig. 1. Analysis of allele specific sgRNA sites

A) A heterozygous variant may generate an allele-specific CRISPR site by being located in a PAM site or in the proximal sgRNA recognition sequence. If a variant is homozygous, it similarly may generate or destroy a novel sgRNA site in an individual. B) Most variants annotated by the 1000 Genomes Project (1KGP) are in or near a PAM site, and thus may generate or destroy a sgRNA site depending on which variants are present in an individual's genome. C) This faceted density plot shows the percentage of putatively targetable 1KGP individuals (2,504 total individuals) per protein-coding gene for 11 types of Cas nuclease.

A genetic variant can impact sgRNA sites by being located in or near a protospacer adjacent motif (PAM site), potentially generating or eliminating sgRNA sites in an individual. Accordingly, a heterozygous variant can generate an allele-specific sgRNA site by enabling differentiation of sgRNAs between alleles[11,12] (Figure 1a). Because Cas nucleases have different PAM sequences, a variant may create an allele-specific sgRNA site for one Cas but not another. We analyzed 11 Cas types (Supplementary Table 1) and ~81 million genome-wide variants annotated by the 1000 Genomes Project[13] (1KGP), finding that most heterozygous variants generate allele-specific sgRNA sites for at least one Cas type, even when considering only variants in PAMs, which are putatively more specific[1] (Figure 1b). The likelihood that a variant creates an allele-specific sgRNA site differs across Cas nucleases (range: 19-98%), is positively correlated with PAM frequency in the reference genome (Pearson rho=0.9, p=0.04), and is negatively correlated with PAM size (Pearson rho=-0.9, p=0.05). These results emphasize the importance of accounting for heterozygous and homozygous non-reference variants in sgRNA design.

We next focused on pairing sgRNAs to generate genomic excisions to eliminate protein production from one allele of a gene. We used these data to approximate how effective allele-specific editing could be in eliminating disease alleles, determining the key obstacles that need to be overcome for successful therapeutic editing. Given a Cas nuclease, an estimated maximum distance between the two sgRNAs on the haplotype to be excised, and allele-specific sgRNA sites, it is possible to classify genes (or other genomic elements) as putatively targetable or not (Supplementary Figure 1). We use the term putatively

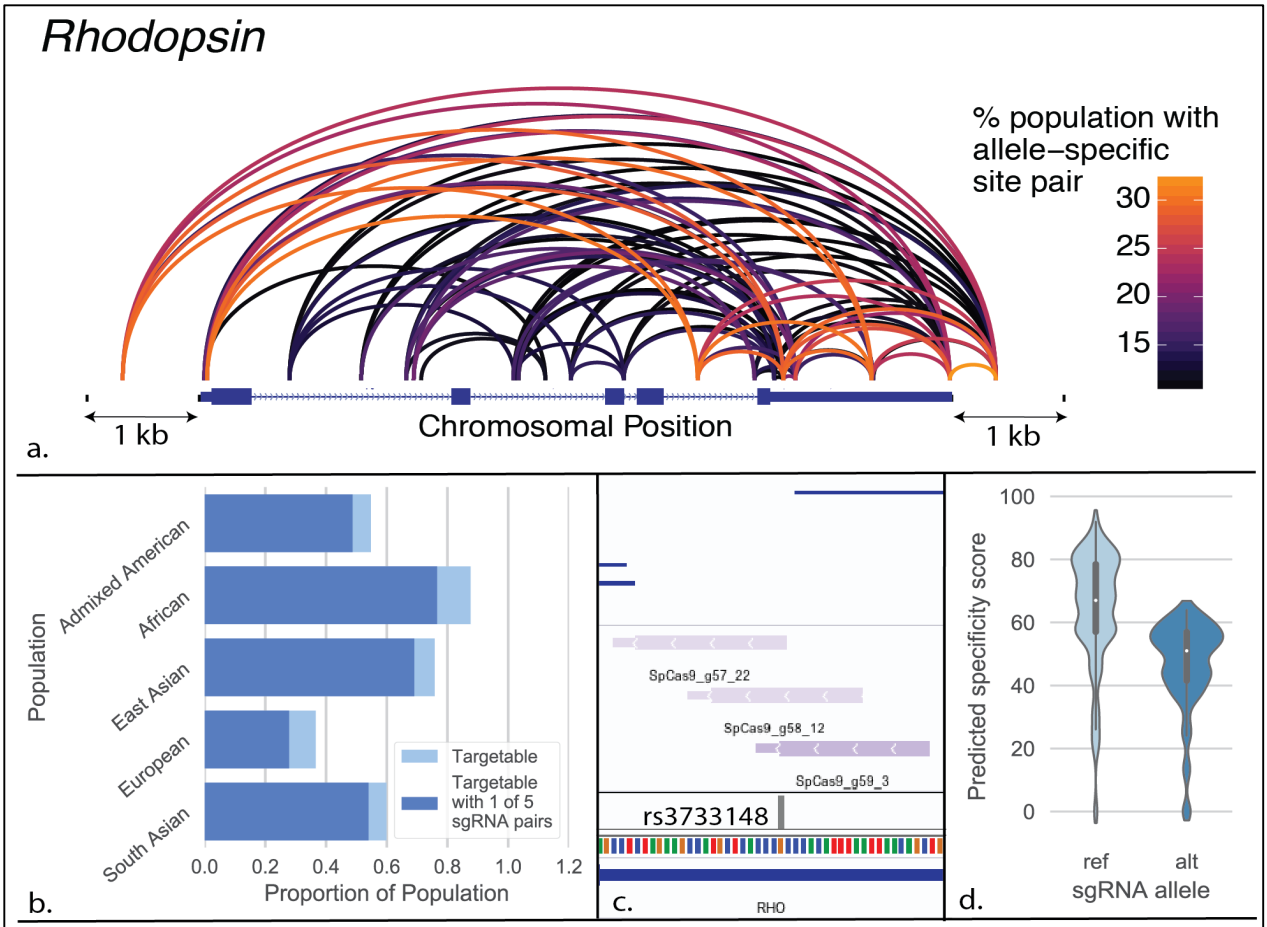


Fig. 2: Targeting pairs of allele specific polymorphisms

A) 82 pairs of allele-specific sgRNA sites for SpCas9 are shared by at least 10% of 1KGP in the gene RHO including 1 kb flanking the gene. B) Most targetable individuals have at least one of five SpCas9 sgRNA pairs optimized to be highly shared among the 1KGP cohort in the gene RHO plus the 1 kb regions flanking the gene. C) Allele-specific (upper portion) and personalized (lower portion) guides for SpCas9 designed around the variant rs3733148, which is heterozygous in WTC and occurs in the last exon of RHO. The horizontal blue bars indicate personalized, non-allele-specific sgRNAs. Lavender (middle horizontal bars) indicates allele-specific sgRNAs. The vertical grey bar indicates the variant. Colored blocks indicate specific bases of the genetic sequence. The horizontal blue bar at the bottom indicates the exon. Allele-specific sgRNAs incorporate the variant in order to use it to differentiate between the two alleles in WTC, whereas personalized, but non-allele-specific, guides avoid the heterozygous variant as it will render a mismatch between the sgRNA and one of WTC's alleles. D) All possible allele-specific sgRNAs for SpCas9 in RHO plus the flanking 1 kb sequences were designed and scored for predicted specificity using CRISPOR. In the violinplot showing the score distribution for sgRNAs designed against the reference and alternate alleles per each heterozygous variant, the inner boxplot denotes the quartiles of each dataset.

targetable, because it will not always be possible to cut specifically at a site and excision of a coding exon will not always stop expression[14]. If we choose a maximum distance of 10 kilobases (kb) between sgRNAs and consider 11 Cas varieties, the average individual from 1KGP is putatively targetable for allele-specific excision at 77% of protein-coding genes. This rate is evenly distributed across chromosomes but varies by Cas nuclease and gene (Figure 1c). For genes that are not targetable, additional allele-specific sgRNA sites may be found by leveraging non-coding variants up- and down-stream of the gene or even in distal enhancers for the gene. Genome-wide, we found that by simply including the 5 kb flanking regions of each gene, we can increase the mean targetable protein-coding genes per individual from 77% to 85%. We conclude that allele-specific excision is applicable to the vast majority of genes in a genome.

We compared a paired-guide to single-guide strategy (at least one allele-specific sgRNA site in a coding exon) for allele-specific gene knockout in the individual human genome of an induced pluripotent stem cell (iPSC) line commonly used for disease modeling, WTC[15]. A paired-guide strategy is valuable as it allows gene disruption using non-coding variants, which occur more frequently than coding variants. We found that a paired-guide strategy rendered more than double the number of genes putatively targetable compared to a single-guide strategy (Supplementary Figure 2. Furthermore, genes that were targetable with a single- and not paired-guide approach tended to have less than two heterozygous variants in the gene, indicating lack of multiple variants as the reason a paired-guide strategy fails. Thus, these genes likely could also be putatively targetable with a paired-guide strategy by simply incorporating flanking, promoter, or other regulatory regions. We therefore recommend paired-guides for allele-specific gene excision.

Due to haplotype structure, variants that impact sgRNA sites are often shared among large proportions of the individuals within and sometimes between populations. Sharing varies by population, as individuals with shared ancestry will share haplotypes with specific sets of variants. For example, the gene *RHO* (with 1 kb flanking on each side), which can cause dominant negative retinitis dystrophy[10], has 82 pairs of sgRNA sites that are shared by >10% of all 1KGP individuals, and the number and composition of these pairs varies within the major populations comprising 1KGP (Figure 2a, Supplementary Figure 3. We therefore developed an algorithm to identify allele-specific sgRNA guide pairs for a given gene that cover the maximum number of individuals in a population; these have the broadest therapeutic potential, similar to designing a drug to treat as many people as possible. Specifically, our method seeks to cover the most people with the fewest sgRNA pairs using their shared heterozygous variants, similar to the set cover problem[16]. Applying our method to the *RHO* locus (genic region plus 1 kb flanking non-coding sequence), we discovered five allele-specific sgRNA pairs that putatively excise one allele of the gene in ~88% of 1KGP individuals with at least two variants, or 57% of the overall 1KGP population (Figure 2b). Thus, optimized pairs of sgRNAs may be used to study or treat genetic diseases in large groups, potentially eliminating the need to develop new sgRNA pairs for each patient, with practical implications for the development of genome surgery as a field.

We incorporated these methods into ExcisionFinder, an open-source software tool (Supplementary Figure 4). This tool designs personalized and allele-specific sgRNAs for unique individuals and cohorts, given their genetic variants, and optimizes sgRNA pairs to cover many individuals based on shared variants. As an example, we applied ExcisionFinder to the gene *RHO* plus 1 kb flanking in the genome of WTC. We identified all sgRNAs for SpCas9 that would be missed using the GRCh38 reference genome, demonstrating how allele-specific sgRNAs leverage heterozygous variants while personalized sgRNAs avoid heterozygous variants but incorporate homozygous variants (Figure 2c). We integrated the specificity scoring capabilities of CRISPOR[4] to enable users to stratify guides by that metric as desired (Figure 2d). To our knowledge, this is the first CRISPR/Cas computational resource that enables personalized and allele-specific genome editing. The ExcisionFinder toolkit and tutorials are available along with the database of annotated 1KGP variants at https://github.com/keoughkath/ExcisionFinder. Because sgRNA specificity is critical for all applications of CRISPR, these resources are an important step towards effective deployment of CRISPR-based technologies with diverse genomes, including but not limited to research and therapeutic development for once incurable dominant negative diseases.

Acknowledgements

We are grateful for valuable scientific input from Anthony Moore, and members of the Pollard and Conklin labs. K.C.K received support from the Claire Giannini Fund and the UCSF Discovery Fellows Program. We thank Anders Riutta from the Gladstone Bioinformatics Core for his Github troubleshooting expertise. We are grateful to Maximilian Haeussler for his assistance and feedback with the integration of CRISPOR. B.R.C and K.S.P. were supported by the Gladstone Institutes. K.S.P. received support from the National

Institutes of Health (UM1HL098179, P01HL089707, R01MH109907). B.R.C received support from the National Institutes of Health (U01HL100406, P01HL089707, R01HL130533, 1R01EY028249).

Author Contributions

K.C.K, S.W., B.R.C., and K.S.P. conceived the project. K.C.K, S.W., S.L., B.R.C., and K.S.P. designed the experiments, K.C.K, S.L., M.P.O., B.R.C., and K.S.P. analyzed data. K.C.K, B.R.C., and K.S.P. wrote the paper with editing from all other authors.

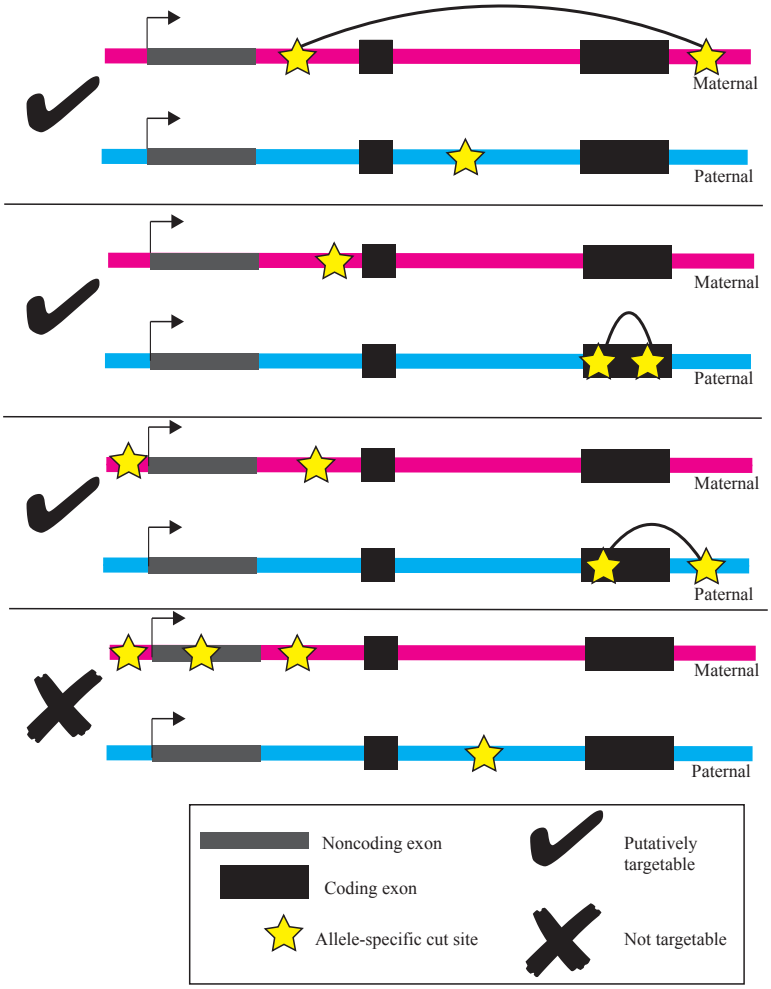
Competing Interests

B.R.C. is a founder of Tenaya Therapeutics, a company focused on finding treatments for heart failure, including the use of CRISPR interference to interrogate genetic cardiomyopathies. B.R.C. and K.S.P. hold equity in Tenaya, and Tenaya provides research support for heart failure related research to B.RC and K.S.P.

References

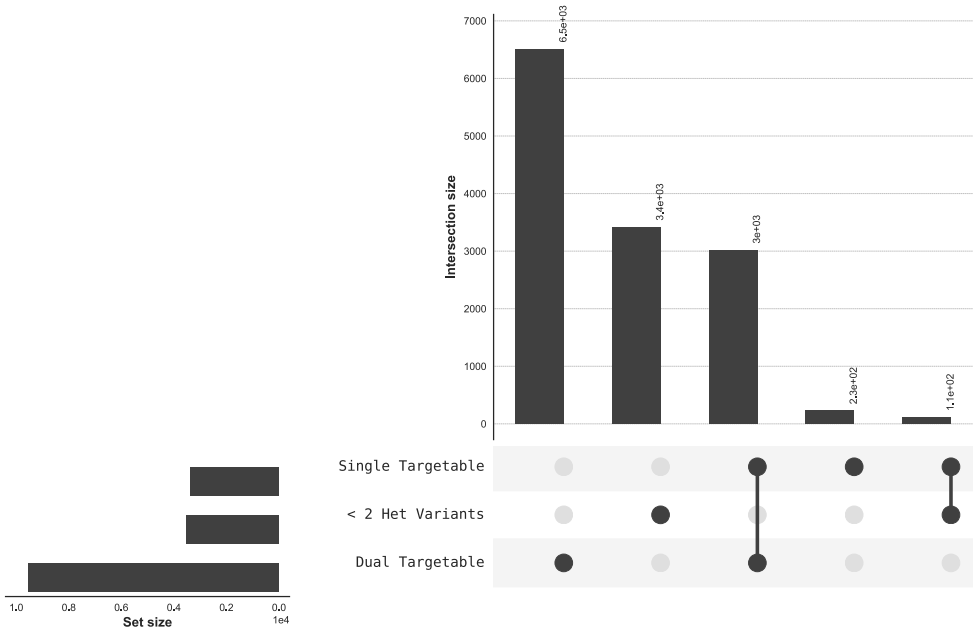
1. Kathleen A. Christie, David G. Courtney, Larry A. DeDionisio, C. C. S. & Shyamasree De Majumdar, Laura C. Mairs, M. A. N. & C. B. T. M. Towards personalised allele-specific CRISPR gene editing to treat autosomal dominant disorders. *Sci. Rep.* 1–11 (2017). doi:10.1038/s41598-017-16279-4
2. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat. Biotechnol.* **32,** 1262–7 (2014).
3. Horlbeck, M. A., Gilbert, L. A. & Villalta, J. E. Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. **9,** 1–20 (2016).
4. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17,** 148 (2016).
5. National Institutes of Health. NIH Curriculum Supplement Series. *Biological Sciences Curriculum Study* (2007). doi:10.1371/journal.pone.0075601
6. Scott, D. A. & Zhang, F. Implications of human genetic variation in CRISPR-based therapeutic genome editing. *Nat. Med.* **23,** 1095–1101 (2017).
7. Lessard, S. *et al.* Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E11257–E11266 (2017).
8. Shin, J. W. *et al.* Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. *Hum. Mol. Genet.* **25,** 4566–4576 (2016).
9. Gao, X. *et al.* Treatment of autosomal dominant hearing loss by in vivo delivery of genome editing agents. *Nature* **553,** 217–221 (2018).
10. Bakondi, B. *et al.* In Vivo CRISPR/Cas9 Gene Editing Corrects Retinal Dystrophy in the S334ter-3 Rat Model of Autosomal Dominant Retinitis Pigmentosa. *Mol. Ther.* **24,** 556–563 (2015).
11. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8,** 2281–2308 (2013).
12. Christie, K. A. *et al.* Towards personalised allele-specific CRISPR gene editing to treat autosomal dominant disorders. *Sci. Rep.* **7,** 16174 (2017).
13. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
14. Chen, X. *et al.* Dual sgRNA-directed gene knockout using CRISPR/Cas9 technology in Caenorhabditis elegans. *Sci. Rep.* **4,** 7581 (2014).
15. Drubin, D. G. & Hyman, A. A. Stem cells: the new "model organism". *Mol. Biol. Cell* **28,** 1409–

1411 (2017).
16. Clarkson, K. L. in *Lecture Notes in Computer Science* 246–252 (1993).
17. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493-6 (2004).
18. Mitchell, S., OSullivan, M. & others. PuLP: a linear programming toolkit for python. *Univ. Auckl.* (2011).
19. Van der Auwera, G. A. *et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline*. *Current Protocols in Bioinformatics* (2013). doi:10.1002/0471250953.bi1110s43
20. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
21. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the 9th Python in Science Conference* (eds. van der Walt, S. & Millman, J.) 51–56 (2010).
22. Stéfan van der Walt, S. C. C. and G. V. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13,** 22–30 (2011).
23. Francesc Alted, I. V. and others. PyTables: Hierarchical Datasets in Python.
24. Cock, P. J. A. *et al.* Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25,** 1422–1423 (2009).
25. Shirley, M. D., Ma, Z., Pedersen, B. S. & Wheelan, S. J. *Efficient 'pythonic' access to FASTA files using pyfaidx*. (PeerJ Inc., 2015).



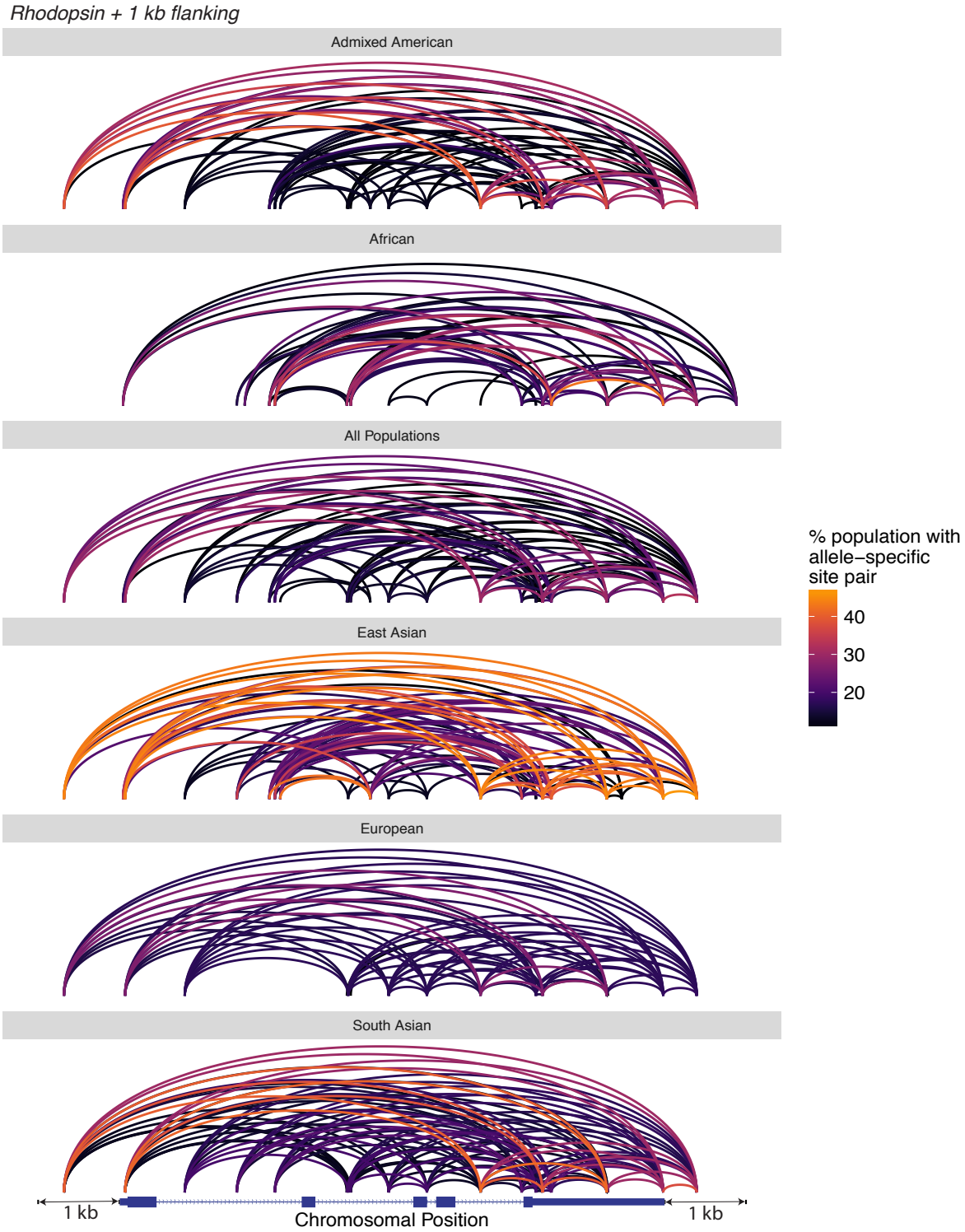
Supplementary Figure 1

A pair of allele-specific sgRNA sites is defined at putatively targetable if their predicted excision will disrupt at least one protein-coding exon.



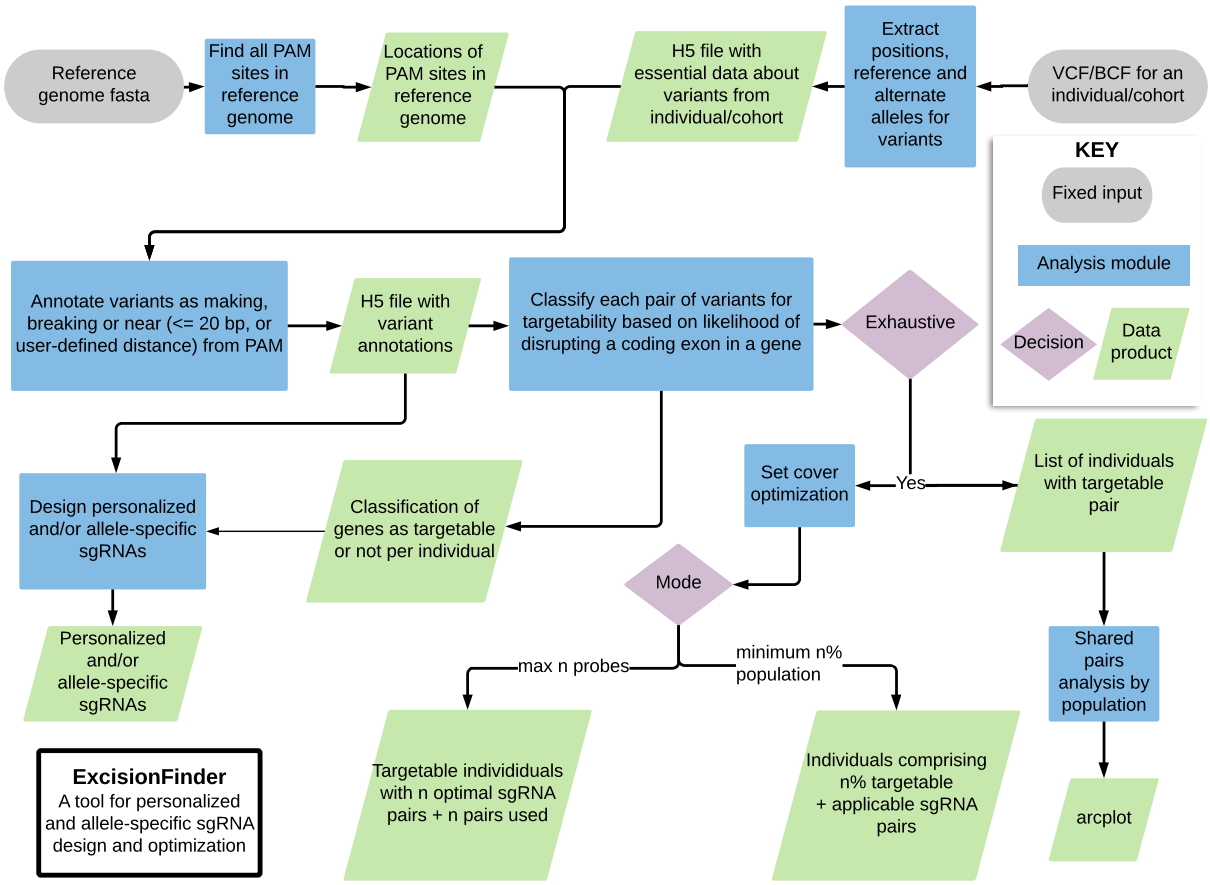
Supplementary Figure 2

Many more genes are targetable in the genome of WTC with a paired (dual)- as opposed to single-guide strategy. The number of variants in a gene is influential in determining targetability. Many genes that are not dual- or single-guide targetable have very few variants, and the genes that are only targetable with a single-guide approach compared to a dual-guide approach also tend to have fewer variants. All 11 Cas varieties are considered in this analysis.



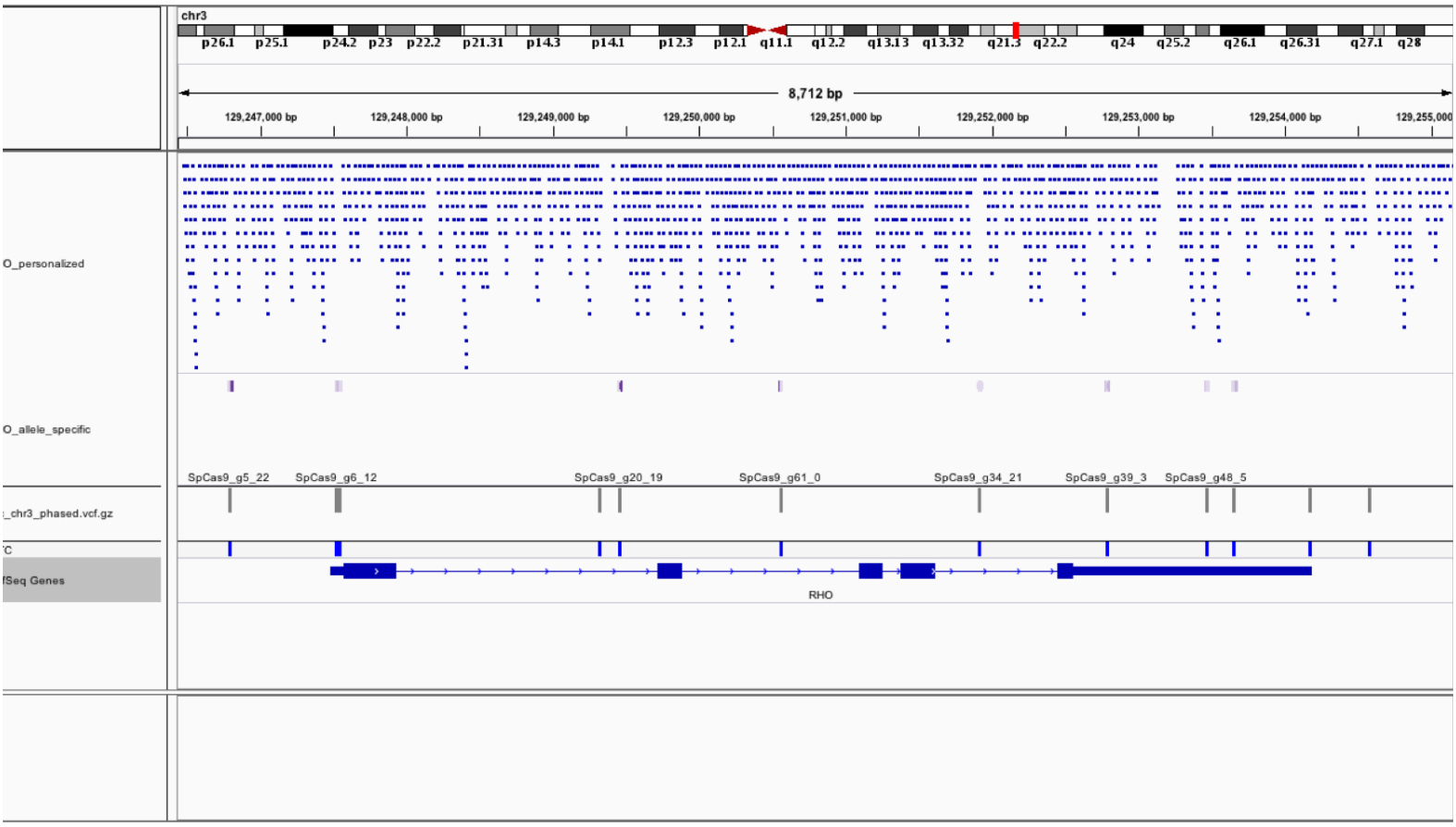
Supplementary Figure 3

Shared pairs of sgRNAs per locus vary by population. We show allele-specific sgRNA site pairs shared by at least 10% of each population for SpCas9 in the gene *RHO* plus the 1 kb flanking regions in the five super-populations in the 1KGP as well as the overall 1KGP cohort.



Supplementary Figure 4

Flowchart for the ExcisionFinder software tool.



Supplementary Figure 5

Integrative Genomics Viewer track view of personalized (in blue, upper track) and allele-specific (lavender, middle track) sgRNAs for SpCas9 in the gene *RHO* plus 1 kb flanking in WTC. Allele-specific guides are shaded according to position of the variant in the guide, with variants closer to the PAM being darker based on their putative greater specificity. The track labeled WTC denotes heterozygous variants in WTC in this locus. The bottom track shows the RefSeq annotation for this gene in GRCh37.

Common name(s)	Abbreviation	PAM	Properties
SpCas9	SpCas9	NGG	*Streptococcus pyogenes* (Sp) Cas9., most widely used version with dozens of variants using same PAM, e.g. eSpCas9, SpCas9-HF1, eSpCas9 1.1 and more (Jinek et al. 2012)
SpCas9 VRER Variant	SpCas9-V1	NGCG	Version of SpCas9 with alternative targeting range (Kleinstiver et al. 2015)
SpCas9 EQR Variant	SpCas9-V2	NGAG	Version of SpCas9 with alternative targeting range (Kleinstiver et al. 2015)
SpCas9 VQR Variant	SpCas9-V3	NGAN or NGNG	Version of SpCas9 with wider targeting range (Kleinstiver et al. 2015)
SaCas9	SaCas9	NNGRRT	*Staphylococcus aureus* (Sa) Cas9. Small relative to SpCas9, (Horvath et al. 2008, Jiang et al. 2013)
SaCas9 KKH Variant	SaCas9-V1	NNNRT	Version of SaCas9 with 2 to 4-fold increased targeting range relative of SaCas9 (Kleinstiver et al. 2015)
nmCas9	nmCas9	NNNNGATT	*Neisseria meningitidis* (Nm) Cas9, with different PAM site (Hou et al. 2013)
cpf1	cpf1	TTTN	Multiple variations, notably opposite orientation system and sticky-end cut rather than blunt. Multiple species exist, including from *Acidaminucoccus* and *Lachnospiraceae.* (Zetsche et al. 2015)
StCas9 1	StCas9-V1	NNAGAA	*Streptococcus thermophilus* (St) Cas9. Smaller relative of SpCas9. Increased specificity. (Kleinstiver et al. 2015, Muller et al. 2016)
StCas9 2	StCas9-V2	NGGNG	*Streptococcus thermophilus* (St) Cas9. Smaller relative of SpCas9. Increased specificity. (Muller et al 2016)
cjCas9	cjCas9	NNNNACA	*Campylobacter jejuni* Cas9. Smallest Cas9 ortholog to date, easy to package (Kim et al. 2017)

Supplementary Table 1

11 types of Cas enzyme were evaluated, each of which has a distinct PAM site.

Supplementary Table 2

All possible personalized (non-allele-specific) sgRNAs for SpCas9 in *RHO* plus the 1 kb flanking sequences for WTC. WTC has no homozygous variants in this region, thus allele frequency and variant-related columns are blank. However, the sgRNAs are designed to avoid the 9 heterozygous variants that WTC has in this region, as is visible in Supplementary Figure 5.

Supplementary Table 3

All possible allele-specific sgRNAs for SpCas9 in *RHO* plus the 1 kb flanking sequences for WTC. This table demonstrates the output of designing guides for a region in an individual region with ExcisionFinder. The tool has multiple options for customizing this output to more easily facilitate experimental design, such as translation of sequence to RNA, different guide lengths, and more.

Supplementary Table 4

An example subset of variant annotations for the first 100 variants on chromosome 1 from 1KGP.

Methods

PAM occurrence in the human reference genome

PAM frequency

The ExcisionFinder tool includes a script enabling scanning of a reference genome fasta file for existing PAM sites. We used this to identify PAM sites in the reference human genomes GRCh37 and GRCh38.

PAM size

PAM sizes were equated as the sum of non-N (A, C, G or T) bases in a PAM site. Thus "NGG" for SpCas9 would have size 2, and "NNGRRT" for SaCas9 would have size 4.

ExcisionFinder analysis of the 1000 Genomes cohort

Annotation of variants

Genetic variants were determined to generate or destroy an allele-specific sgRNA site if they were proximal to or in a PAM site (Figure 1a). Sufficient proximity to a PAM site was defined for this study as 20 base pairs based on the common length of sgRNA recognition sequences. For all Cas varieties this was the 20 base pairs 5' of the PAM, except for cpf1 for which it was 3' of the PAM. The sgRNA design tools that are part of ExcisionFinder allow different user-defined sgRNA lengths and addition of Cas enzymes and PAMs. There is evidence to suggest that genetic variants that generate or destroy a PAM are more likely to lead to allele-specific Cas activity compared to those in the seed sequence[1]; ExcisionFinder thus provides options to differentiate between CRISPR sites in a PAM site versus the sgRNA recognition sequence. All variants genome-wide were annotated for the 1KGP cohort and are available for querying, an example subset of these data for the first 100 variants annotated by 1KGP on chromosome 1 in reference genome GRCh37 is available in Supplementary Table 4.

Generation of gene set

The gene set analyzed was compiled using the canonical transcripts for RefSeq gene annotations for human reference genome GRCh37 and GRCh38 downloaded using the UCSC table browser[17], and filtered for genes with at least one coding exon. When non-protein-coding genes were excluded, 15,199 genes were evaluated for GRCh37, and 16,143 for GRCh38. Values reported in the text are for GRCh37 unless stated otherwise, but analyses were conducted for both reference genomes with similar results.

Allele-specific putative gene targetability genome-wide

Putative allele-specific targetability of a gene is defined here as whether a gene contains a pair of allele-specific sgRNA sites for at least one of the 11 Cas enzymes evaluated that are less than 10 kb apart on the same haplotype in an individual that will disrupt a coding exon (Supplementary Figure 1). This metric was calculated for each protein-coding gene for all 2,504 1KGP individuals.

Set cover analysis

In order to determine optimal pairs of sgRNAs to cover large groups of people in a particular gene, we applied set cover optimization which we implemented using the Python package PuLP[18]. The aim was to maximize the number of individuals from the 1KGP for whom a user-supplied maximum number of sgRNA pairs would putatively target a given gene. This script can also be used to determine a minimum percentage of people to be covered by a set of sgRNA pairs.

WTC sequencing

The genome for the iPSC line WTC[15] was sequenced by the Allen Cell Science Institute. Analysis and variant calls in the reference genome GRCh37 were done according to GATK version 3.7 best practices[19] and phased using Beagle version 4.1 with default settings[20].

WTC targetability analysis

Variant annotation procedures were the same as in the 1KGP analysis. The same genes lists used in the 1KGP analysis were analyzed in WTC, except when specified in the text, for the cases of 5 kb flanking the gene *RHO*, or when analyzing targetability for all genes + 5 kb flanking vs. genic region only.

Packages used

Python

Docopt was used for handling of command-line arguments. Pandas[21] version 0.21.0 and NumPy[22] version 1.13.3 and elements of the standard Python distribution sys, os, and regex were used for multiple aspects of data analysis. PuLP[18] version 1.6.8 was used for set cover analysis. PyTables[23] was used for data management. Biopython[24] and pyfaidx[25] were used for Fasta processing. Scripts from CRISPOR[4] were integrated into ExcisionFinder to facilitate specificity scoring of sgRNAs.

R

Packages used to generate arcplots included viridis version 0.5.1, viridisLite version 0.3.0, igraph version 1.1.2, ggraph version 1.0.0, ggplot2 version 2.2.1, reshape2 version 1.4.3, dplyr version 0.7.4, tidyr version 0.7.2, and readr version 1.1.1.

Bioinformatics

Bcftools versions 1.5 and 1.6 were used to manipulate VCF and BCF files.

Scripts

Scripts were written in Python version 3.6.1, R version 3.3.2 and Bash version 3.2.57.

Data Availability

1KGP phase 3 data were downloaded from the 1KGP website (http://www.internationalgenome.org/). The reference GRCh37 and GRCh38 genome data were downloaded from the UCSC genome browser. The 1KGP analysis dataset has been made available for public access online at (http://lighthouse.ucsf.edu/public_files_no_password/excisionFinderData_public/1kgp_dat/).

Code Availability

All data processing and analysis scripts are located at github.com/keoughkath/ExcisionFinder. This repository also houses the tools described here for allele-specific CRISPR sgRNA design as an open-source resource.