High frequency *CCR5* editing in human hematopoietic stem progenitor cells protects xenograft mice from HIV infection

Supplementary Table 1

Dual-guide Edited (CCR5)					
Donor	**Total live cells Prior to EP**	**Viability prior to EP**	**Total live cells two days post EP**	**Viability two days post EP**	**% Recovery**
1	4.50E+07	95%	4.40E+07	[a]N/A (>90% when thawed for engraftment)	97.78%
2	4.00E+07	90%	3.82E+07	86%	95.50%
3	9.00E+07	94%	9.01E+07	85%	100.11%
Mock Editing (GFP)					
Donor	**Total live cells Prior to EP**	**Viability prior to EP**	**Total live cells two days post EP**	**Viability two days post EP**	**% Recovery**
1	4.50E+07	95%	6.20E+07	[a]N/A	137.78%
2	4.00E+07	90%	4.81E+07	91%	120.25%
3	4.50E+07	94%	6.02E+07	92%	133.78%

[a]Cells were cryopreserved before viability measurement

Supplementary Table 2

gRNA	**Target Sequence**	**Forward Primer Sequence**	**Reverse Primer Sequence**
TB48	CAATGTGTCAACTCTTGACA	TTATACATCGGAGCCCTGCC	TTTCGACACCGAAGCAGAGT
TB50	CGACAAAGGCATAGATGATG	GCTGTGAGGCTTATCTTCACCATC	TCGGAGTGAAGGGAGAGTTTGTC
TB8	TAAACTGAGCTTGCTCGCTC		
TB7	AGTTTACACCCGATCCACTG		

Supplementary Table 3

Gene Target	**Forward Primer Sequence**	**Reverse Primer Sequence**	**Probe**	**Fluorophore**
CCR5	GTCCATGCTGTGTTTGCTTT	CACCACCCAAGTGATCACAC	GATGATTCCTGGGAGA GACG	FAM
RPP30	Reference Gene Bio-Rad ddPCR™ Copy Number Assay ID: dHsaCP1000485			HEX

High frequency *CCR5* editing in human hematopoietic stem progenitor cells protects xenograft mice from HIV infection

Supplementary Table 4

TB50 Off-target Sites				
Chromosome	Gene Name	Genomic start	Genomic end	Off-target Sequence
chr1	LMNA	156098146	156098208	CTACAAGGGAATAGATGATG
chr2	RP11-17G11.1	152533920	152533982	CTACAAAGGCATAAATGAAG
chr3	P3H2	189956449	189956511	CTACAAAGGCATAGTTCATG
chr15	RP11-62C7.2	39035219	39035281	AGACAAAGGCCTCGATGATG
chr11	GRM5-AS1	88487883	88487945	AGACAGAGGCTTAGATGATG
chr17	KPNB1	47673627	47673689	CGAAAATGGTATAGATGATG
chrX	FAM104B	55155109	55155171	AGACAAAGCCACAGATGATG
chr4	RP11-427M20.1	133926916	133926978	TGACACAGGAATAGATGATG
chr2	CXCR1	218164267	218164329	CGATGAAGGCGTAGATGATG
chr11	APLNR	57236048	57236110	CGAAAAAGGCATAGAGGAAG
chr3	ADAMTS9-AS2	64794052	64794114	CAACAAGGGCATTGATGATG
chr8	RP11-29A2.1	5296567	5296629	TGACAAAGTAATAGATGATG
chr13	DACH1	71425420	71425482	AGACAAAGGCAGAGATGAAG
chr4	TRMT112P1	116376959	116377021	CTACAAAGGCATATGTGATG
chr19	CTD-3222D19.4	16616586	16616648	TGACAAAGGCAGAGATGAAG
chr12	SLC11A2	50995656	50995718	CAACAAAGACATTGATGATG
chr19	CEP89	32959414	32959476	CAACAAAGGAAAAGATGATG
chrX	ZDHHC15	75427692	75427754	GGACAAAGGAAGAGATGATG
chr10	RP11-476F14.1	33614046	33614108	CGATAAAGCAATAGATGATG
TB7 Off-target Sites				
Chromosome	Gene Name	Genomic start	Genomic end	Off-target Sequence
chr1	NA	191016245	191016307	AGTTTTCACCAGTTCCACTG
chr3	AC117507.1	188235173	188235235	AGTTTGCACCCGTTCCACTG
TB48 Off-target Sites				
Chromosome	Gene Name	Genomic start	Genomic end	Off-target Sequence
chr10	ADAM12	126309067	126309129	CCATGTGTCCACTCTTGGCA
chr11	TENM4	79267265	79267327	CCATATGTCTACTCTTGACA
chr17	PRKAR1A	68529541	68529603	CAAAGTCTCAACTCTTGTCA
chr7	AC007652.1	35529882	35529944	CACTGGGTCAACTCTTGAGA
chr10	LINC01435	108245291	108245353	CTATGTGTTAACTCCTGACA
chr13	CTAGE11P	75201141	75201203	CAATGTGACAACTCATGCCA
chr5	LINC01194	12801104	12801166	CTATGTTTCAACTCCTGACA
chr3	ZNF589	48263035	48263097	CAATATGTAATCTCTTGACA
chr5	CTD-2004A9.1	32668382	32668444	CTATGTGTTAACTCTTGACA
chrX	RP1-52D1.1	87340076	87340138	CAAGGTGTCAACTTCTGACA
chr12	DDX47	12823359	12823421	AAATGTGTCAACTCTTGATT
chr8	C8orf34	68660849	68660911	CTATGTGTTAACTCCTGACA

chrX	NAP1L4P2	106451667	106451729	CAATGTGGTAACTCTTGTCA
chr3	CMC1	28142472	28142534	AAATGTGTGAACTCTTGACC
chr12	RP11-153K16.2	23276517	23276579	GAATGTGTAAACTCTTGACC
chr7	RP11-667F14.1	116187375	116187437	CAGTGTGTCGACTCTTGAAA
chr14	PRKD1	29607000	29607062	TAATGTGTCAACTCATAACA
chr1	FAF1	50529073	50529135	TAGGGTGTCAACTCTTGACA
chr3	PLSCR2	146458041	146458103	CAATGTAACAACTCTTTACA
chr4	PCDH7	30917027	30917089	CAGAGTGTCAACTCCTGACA
chr7	EEF1A1P27	20580549	20580611	CAATATGTGAACTCTTGTCA
chr3	RP11-598F17.2	96248457	96248519	CAATGGGTCAGCTTTTGACA
TB8 Off-targets Sites				
Chromosome	Gene Name	Genomic start	Genomic end	Off-target Sequence
chr16	GPR139	20074350	20074412	GAAGCTGAGCTTGCTGGCTC
chr1	RGSL1	182554800	182554862	GAAACTGAGCTTGCTGCCTC
chr18	NA	66227816	66227878	TATACTGATCTTGCTAGCTC
chr6	C6orf163	87354895	87354957	TAAACAGAGCTGGCTCTCTC
chr20	SUMO1P1	53902185	53902247	TAAACTGTGCTTACTCGCTC
chr5	NA	99737598	99737660	TAAAATGAGCCTGCTCTCTC
chr5	CTD-2151A2.1	99574606	99574668	TAAAATGAGCCTGCTCTCTC

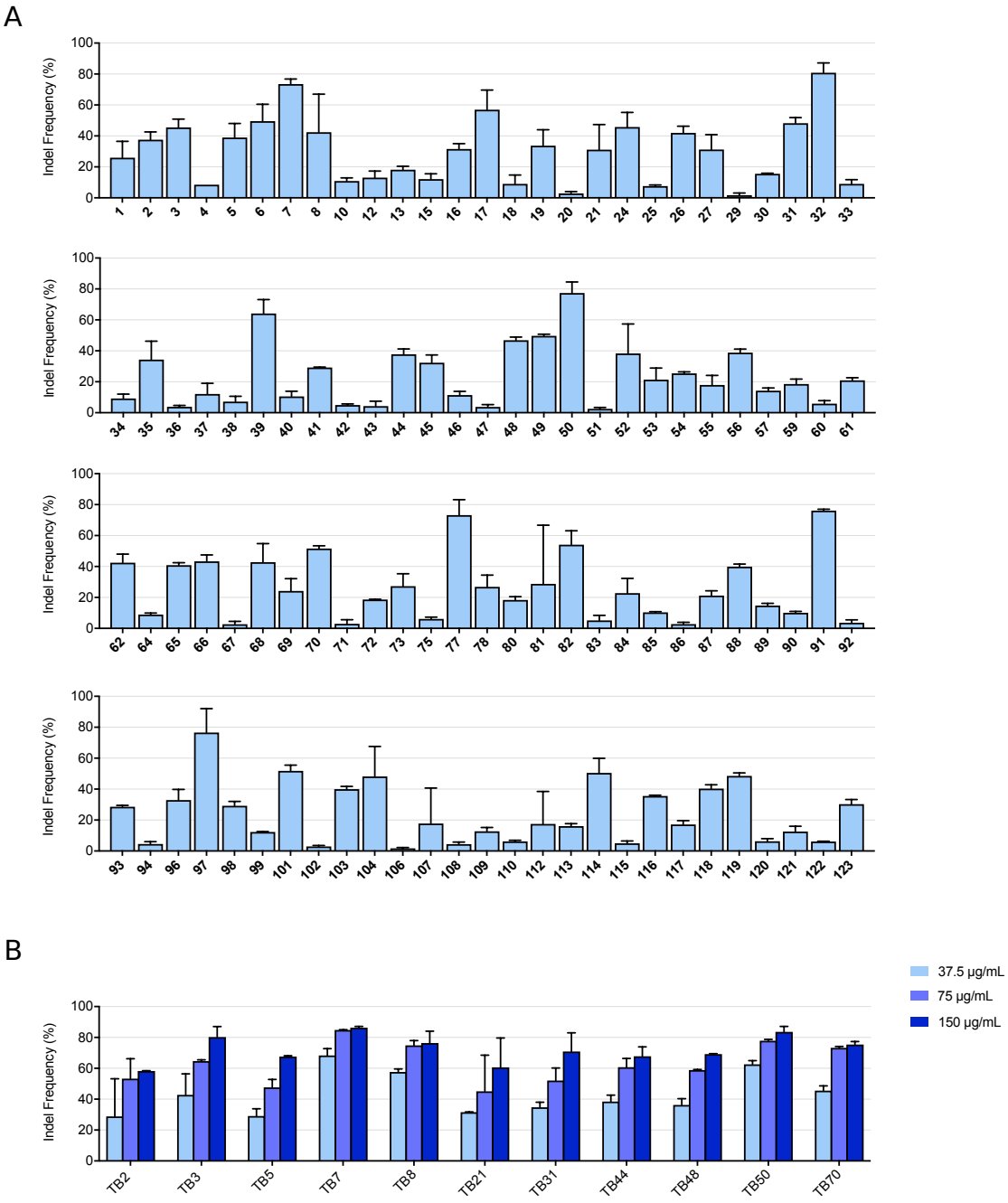
The genomic start and end coordinates are based on the GRCh38 assembly.

Supplementary Table 5

Antibodies Panels				
Antibody target	**Fluorochrome**	**Clone**	**Manufacturer**	**Catalog Number**
CD8	APC/Fire 750	SK1	BioLegend	344745
CCR5	PE/Cyanine7	J418F1	BioLegend	359107
CD4	Brilliant Violet 421	OKT4	BioLegend	317433
CD3 BV650	Brilliant Violet 650	UCHT1	BioLegend	300467
HIV Gag p24	RD1	KC57	Beckman Coulter	6604667
CD45	FITC	HI30	BioLegend	304006
CCR5	PE	HEK/1/85a	BioLegend	313707
CD235a	PE/Dazzle 594	HI264	BioLegend	349119
CD123	PE/Cyanine7	7G3	BD Biosciences	560826
CD33	Brilliant Violet 421	WM53	BioLegend	303415
CD16	Brilliant Violet 510	3G8	BioLegend	302047
CD19	Brilliant Violet 605	SJ25C1	BioLegend	363023
CD11c	Brilliant Violet 650	3.9	BioLegend	301637
CD45	Brilliant Violet 711	30-F11	BioLegend	103147
CD4	Brilliant Violet 785	OKT4	BioLegend	317441
CD14	Brilliant Ultraviolet 395	MϕP-9	BD Biosciences	563562

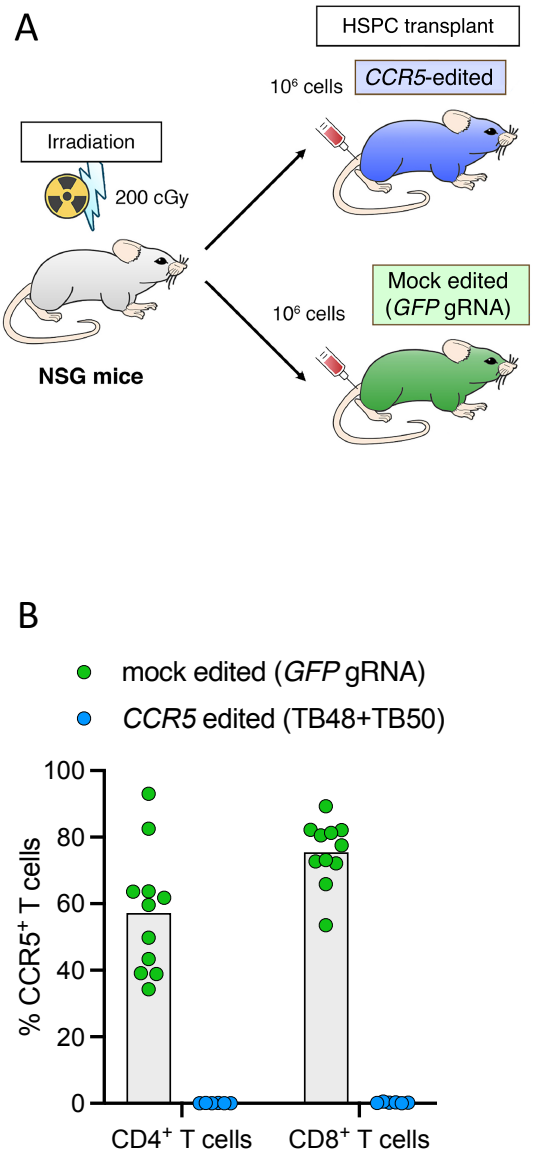
CD3	Brilliant Ultraviolet 496	UCHT1	BD Biosciences	612940
CD56	Brilliant Ultraviolet 805	NCAM16.2	BD Biosciences	749086

High frequency *CCR5* editing in human hematopoietic stem progenitor cells protects xenograft mice from HIV infection



Supplementary Fig. 1 Screening guides for editing efficiency in primary human HSPCs. Of the 123 gRNA sequences identified by the in silico predictor software, 15 were excluded for having multiple target sites within the human genome. **a** The remaining 108 gRNAs were produced through in vitro transcription and screened for editing efficiency in $CD34^+$ HSPCs from two different donors. The graphs depict the average indel frequency in HSPCs from the two donors for each of the 108 guides. **b** The 11 gRNAs demonstrating the highest editing efficiency and without homology to the *CCR2* gene were chemically synthesized, complexed with Cas9, and electroporated at 37.5, 75, or 150µg/mL concentrations using two $CD34^+$ HSPC donor sources. All bars represent the mean, and all errors bars represent ± SD.

High frequency *CCR5* editing in human hematopoietic stem progenitor cells protects xenograft mice from HIV infection



Supplementary Fig. 2 Robust ablation of CCR5 in differentiated human T cells from mice transplanted with *CCR5* edited HSPCs. **a** Schematic of humanized mouse generation. 6 – 8 week old female NSG mice were sub-lethally irradiated with 200cGy then subsequently transplanted with 1 x 10^6 mobilized $CD34^+$ HSPCs that were either mock edited (*GFP* gRNA) (n=27, green) or dual guide (TB48, TB50) edited for *CCR5* (n=27, blue) complexed to Cas9. **b** At 24-weeks post-transplant, mice were bled and human cell count, lineage frequencies, and CCR5 expression was quantified by multiparameter flow cytometry. The graph depicts the frequency of CCR5 expression in human $CD4^+$ and $CD8^+$ T cells for mice with sufficient T cell reconstitution as defined by >10 $CD4^+$ T cells per mm^3 of blood.

High frequency *CCR5* editing in human hematopoietic stem progenitor cells protects xenograft mice from HIV infection

Supplementary Text 1 - Mathematical Modeling

This Supplement provides further details of the mathematical modeling outlined in the Modeling section of the Methods in the main text. The model aims to mechanistically describe the effect of *CCR5*-editing on the probability of infection being established following an inoculation, and we use the insight gained from this model to predict the effects of *CCR5*-editing on HIV rebound in people living with HIV (PLWH).

S1.1 Parameter and Variable Definitions

Table S1.1: Definitions of symbols and notation used in this Supplementary Text along with the corresponding units and variable type.

Symbol/ Notation	Description	Units*	Variable Type
$P(A)$	Probability of event A happening		Probability
$E(B)$	Expectation of random variable B	Same as B	Numerical
$\hat{z}$	Point estimate of parameter (or parameter set) z	Same as z	Numerical
a	Animal index		Ordinal
$I_a; i_a$	The inoculation at which animal a had an outcome occur. This outcome could be either a detected infection or censoring. Note that $i_a > 0$ for all animals.		I_a is a random variable; i_a is an observation; both are ordinal
$D_a; d_a$	Indicator for if the detected outcome for animal a was infection (1) or censoring (0)		D_a is a random variable; d_a is an observation; both are logical
g	Study arm index		Ordinal
$n; n_g$	Total number of animals in the study or study arm g.	Animals	Numerical
x	Frequency of intact *CCR5* alleles in CD19$^+$		Numerical – calculated from analysis of bone marrow samples taken at necropsy
p	Probability of infection being established following an inoculation of an animal – in different sections of this Supplement, p is either a constant or a function of intact *CCR5* allele frequency in CD19$^+$ cells (x)		Probability
ξ	Event defined as all viral lineages going extinct after an inoculation		Probabilistic event
C	Number of cells infected by an inoculum	Cells	Random variable – integer
$E(C) = \rho(x)$	Expected number of cells infected by an inoculum – function of intact *CCR5* allele frequency in CD19$^+$ cells (x)	Cells	Numerical
i_{50}	Number of inoculations per animal that will result in at least 50% of animals becoming infected		Integer

$F_I(i; x_a, \boldsymbol{\theta})$	Cumulative distribution for the probability that animal a was infected on or before the i^{th} inoculation given the model parameters ($\boldsymbol{\theta}$) and intact *CCR5* allele frequency in CD19$^+$ cells (x_a)		Probability
R_0; $R_{0,H}$	Basic reproductive number; basic reproductive number of HIV in PLWH (8)		Numerical
$r(x)$	Rate of viral rebound following ART interruption (function of intact *CCR5* frequency)	week^{-1}	Numerical
r_0	Reactivation rate of latently infected cells in PLWH	week^{-1}	Numerical
ΔV_{eq}	Projected ratio of rebound set-point viral load in PLWH if they have undergone *CCR5*-editing compared to if they have not.		Numerical

**Empty cells indicate the parameter/variable is unitless.*

S1.2. Data assumptions, censoring, and inclusion

We assumed that animals were infected at the inoculation that occurred immediately prior to the detected viral load, and we censored animals who remained uninfected one week following their eighth and final inoculation. In the analysis of this experiment, we only include animals for which intact *CCR5* allele frequency was measured in CD19$^+$ cells taken from bone marrow at necropsy, as this measure was the predictor variable in our best fit mechanistic model for the probability of infection following an inoculation (outlined in Section S1.5).

S1.3. Kaplan-Meier comparison of infection probability

We assessed the significance of the differences in the time-to-infection across study arms in the *CCR5*-editing titration experiment using the *survminer* package in R.

S1.4. Estimation of infection probability for study arms of *CCR5*-editing titration experiment

In order to estimate the average probability of infection following an inoculation for each study arm, we assume that infection probability is identical for all animals within a study arm and that it is identical at each inoculation. The calculations outlined below are therefore identical for each study arm, so without loss of generality, we discuss the process of estimating the infection probability for just one study group (index g). We use a maximum likelihood estimate for this calculation, and therefore the first step is to derive the likelihood for the observed data.

To derive the likelihood function for the likelihood of observing the available data given a certain probability $p \in [0,1]$ of infection being established following an inoculation, we use the following definitions, also outlined above in Table S1.1; i_a is the inoculation at which an outcome occurred for animal a, d_a is an indicator for whether the outcome was detected infection ($d_a = 1$) or animal a was censored, without an infection having been recorded ($d_a = 0$), n is the total number of animals in this study arm, and, as mentioned above, the parameter $p \in [0,1]$ is the probability of an animal getting

infected from a given inoculation. (Note, n and p are specific to this study arm, and therefore are technically n_g and p_g. But, for simplicity of notation, we drop the g subscript in this section.) The probability of an infection being observed for animal a at the i_ath inoculation $(d_a = 1)$ is the product of the probabilities of $i_a - 1$ non-infecting inoculations and 1 infecting inoculation, which is given by $(1-p)^{i_a-1}p$. Alternatively, the corresponding probability for an animal being censored without infection after the i_ath inoculation $(d_a = 0)$, is $(1-p)^{i_a}$. These two probabilities can be combined to show that the probability of outcome d_a at inoculation i_a is given by $(1-p)^{i_a-d_a}p^{d_a}$. Therefore, the likelihood for observing the available data, given p, is:

$$\mathcal{L}(p) = \prod_{a=1}^{n}(1-p)^{i_a-d_a}p^{d_a}$$

and the log-likelihood is

$$\begin{aligned}\ln \mathcal{L}(p) &= \sum_{a=1}^{n}\big((i_a - d_a)\ln(1-p) + d_a \ln(p)\big) \\ &= \ln(1-p)\sum_{a=1}^{n}(i_a - d_a) + \ln(p)\sum_{a=1}^{n} d_a. \end{aligned} \quad \text{(S1.1)}$$

The point-estimate of p, $\hat{p}$, is the value of p that maximizes the log-likelihood, $\ln \mathcal{L}\,(p)$, in Equation (, i.e., where $\frac{d \ln \mathcal{L}}{dp} = 0$.

$$\frac{d \ln \mathcal{L}\,(p)}{dp} = -\frac{1}{1-p}\sum_{a=1}^{n}(i_a - d_a) + \frac{1}{p}\sum_{a=1}^{n} d_a \quad \text{(S1.2)}$$

When we set $\frac{d \ln \mathcal{L}}{dp}|_{p=\hat{p}} = 0$ and solve Equation (S1.2) for $\hat{p}$, there are two cases to be considered:

1) $\sum_{a=1}^{n} d_a > 0$ (i.e., at least one animal has a detected infection) and
2) $\sum_{a=1}^{n} d_a = 0$ (i.e., all animals are censored).

In the first case, when $\sum_{a=1}^{n} d_a > 0$ (i.e., there was at least one detected infection), then solving Equation (S1.2) for $\hat{p}$ gives

$$0 = -\frac{1}{1-\hat{p}}\sum_{a=1}^{n}(i_a - d_a) + \frac{1}{\hat{p}}\sum_{a=1}^{n} d_a$$

$$\frac{1}{1-\hat{p}}\sum_{a=1}^{n}(i_a - d_a) = \frac{1}{\hat{p}}\sum_{a=1}^{n} d_a$$

$$\frac{\sum_{a=1}^{n}(i_a - d_a)}{\sum_{a=1}^{n} d_a} = \frac{1}{\hat{p}} - 1$$

$$p = \frac{1}{\frac{\sum_{a=1}^{n}(i_a - d_a)}{\sum_{a=1}^{n} d_a} + 1} = \frac{\sum_{a=1}^{n} d_a}{\sum_{a=1}^{n}(i_a - d_a) + \sum_{a=1}^{n} d_a}. \quad \text{(S1.3)}$$

Note, the second derivative of the log-likelihood is

$$\frac{d^2 \ln \mathcal{L}(p)}{dp^2} = -\frac{1}{(1-p)^2}\sum_{a=1}^{n}(i_a - d_a) - \frac{1}{p^2}\sum_{a=1}^{n} d_a < 0,$$

so, the critical value of $\ln \mathcal{L}(p)$ truly is a maximum.

For the second case, where $\sum_{a=1}^{n} d_a = 0$ (i.e., no infection was ever detected), the log-likelihood simplifies to

$$\ln \mathcal{L}(p) = \ln(1-p)\sum_{a=1}^{n}(i_a - d_a) = \ln(1-p)\sum_{a=1}^{n} i_a,$$

and

$$\frac{d \ln \mathcal{L}(p)}{dp} = -\frac{1}{1-p}\sum_{a=1}^{n} i_a.$$

Therefore, $\frac{d \ln \mathcal{L}(p)}{dp} \neq 0$ for $p \in (0,1)$, but

$$\ln \mathcal{L}(p = 0) = 0$$

and taking the limit as p approaches 1 from below

$$\lim_{p \to 1^-} \ln \mathcal{L}(p) = -\infty.$$

Therefore, the maximum value of $\ln \mathcal{L}(p)$ on the interval $p \in [0, 1]$ occurs at $p = 0$ (i.e., $\hat{p} = 0$), meaning that the probability of infection being established following inoculation is 0. This makes sense, as in this case there was no evidence of a detected infection.

S1.5. Modeling the relationship between infection probability and *CCR5*-editing

In this section, we explore the mechanistic model for probability of infection following an inoculation that was presented in the main text. The first three subsections (0 to 0) focus on deriving and fitting the model and calculating the expected survival curve. The following two subsections (0 and 0) focus on using bootstrapping to estimate parameter uncertainty and prediction intervals for the survival curves.

S1.5.1. Model derivation

The mechanistic model assumes that the probability of infection following an inoculation is dependent on the availability of target cells. We note that the inoculation size would also play a role in determining probability of infection but given all inoculations in this study were of the same size, we do not explicitly include inoculation size in this model.

We assume that the expected number of cells that get infected (C) is linearly dependent on the availability of target cells, so that $E(C) = \rho(x) = bx + c$. We define $\boldsymbol{\theta} = \langle b, c \rangle$ as a vector of parameters and allow $\rho(x)$ to be represented by $\rho(x; \boldsymbol{\theta})$. We then estimate the parameters b and c, using maximum likelihood fitting.

In this analysis, we use intact *CCR5* allele frequency in CD19$^+$ cells sampled from bone marrow at necropsy as a proxy measure for target cell availability (x). Other measures of target cell availability were considered (such as CD4$^+$CCR5$^+$ cell count prior to inoculation or allele frequency in CD4$^+$ cells at necropsy) but were no more predictive than intact *CCR5* allele frequency in CD19$^+$ cells (and in general provided a worse model fit by the Akaike Information Criterion[1]).

To derive the relationship between target cell frequency and probability of infection following inoculation, we first derive the relationship between target cell frequency and the probability of a certain number of cells, m, becoming infected by the inoculation. To this end, we assume the initial number of cells that get infected (C) is Poisson distributed and, as described above, that the expected number of cells infected ($E(C) = \rho(x; \boldsymbol{\theta})$) is dependent on the frequency of target cells (x). Therefore, by definition, if x_a is the intact *CCR5* allele frequency in CD19$^+$ cells in animal a the probability of m cells getting infected from a single inoculation in animal a, $P(C = m; x_a, \boldsymbol{\theta})$ is

$$P(C = m; x_a, \boldsymbol{\theta}) = e^{-\rho(x_a;\boldsymbol{\theta})} \frac{\rho(x_a; \boldsymbol{\theta})^m}{m!}. \tag{S1.4}$$

For an infection to occur following an inoculation, at least one cell must have been infected (i.e., $C \geq 1$). Hence, the probability of an infection occurring after a single inoculation (p) is

$$\begin{aligned} p(x_a, \boldsymbol{\theta}) &= P(C \geq 1; x_a, \boldsymbol{\theta}) \\ &= 1 - P(C = 0; x_a, \boldsymbol{\theta}) \\ &= 1 - e^{-\rho(x_a;\boldsymbol{\theta})}. \end{aligned} \tag{S1.5}$$

To derive the likelihood function, we once again define i_a as the inoculation at which an event occurred for animal a (now including animals across all study arms) and d_a as an indicator for if the event was a detected infection ($d_a = 1$) or corresponded to censoring ($d_a = 0$). Then, the probability of animal a being infected (or censored) at inoculation i_a, given the parameters defined in $\boldsymbol{\theta}$ is denoted by $P(i_a, d_a; x_a, \boldsymbol{\theta})$. Thus,

$$P(i_a, d_a; x_a, \boldsymbol{\theta}) = \big(1 - p(x_a, \boldsymbol{\theta})\big)^{i_a - d_a} p(x_a, \boldsymbol{\theta})^{d_a} \tag{S1.6}$$

The log-likelihood of the full data set is therefore

$$\ln \mathcal{L}\,(\boldsymbol{\theta}) = \sum_{a=1}^{n} (i_a - d_a) \ln\big(1 - p(x_a, \boldsymbol{\theta})\big) + d_a \ln\big(p(x_a, \boldsymbol{\theta})\big).$$

We estimated the optimal parameters ($\widehat{\boldsymbol{\theta}}$) corresponding to the maximum value of $\ln \mathcal{L}(\boldsymbol{\theta})$ using the *nlm* function in R.

S1.5.2. Model fitting

We compared models where either $c = 0$, so $\rho(x) = bx$, or where c was allowed to be different from 0 via the corrected Akaike Information Criterion ($\mathrm{AIC_c}$) [1], and found that the simpler model ($\rho(x) = bx$) had a slightly lower $\mathrm{AIC_c}$ value ($\mathrm{AIC_c}$ difference = 1.99). Therefore, both models have similar statistical support, but we chose to proceed with the model

$$\rho(x) = bx \tag{S1.7}$$

due to simplicity. The best fit parameter value was $\hat{b} = 0.34$ cells infected per inoculation per intact *CCR5* allele frequency in CD19$^+$ cells.

We also note that inclusion of intact *CCR5* allele frequency in the mechanistic inoculation risk model in Equation (S1.7) results in an $\mathrm{AIC_c}$ 36 lower than that of the null hypothesis model (infection probability is independent of intact *CCR5* allele frequency and identical for all animals), and therefore has significantly more statistical support[1]. The infection probability per inoculation in this null hypothesis model was estimated using Equation (S1.3) and included all animals from the study, irrespective of study group.

S1.5.3. Expected survival curves

Once we have estimates for the probability of infection following an inoculation for each animal, we can calculate the expected survival curve. The expected survival curve for study arm g, $S_g(i)$, is defined as $S_g(i) = \frac{E\left(n_{S,g}(i)\right)}{n_g}$ where $n_{S,g}(i)$ is the number of animals in study arm g remaining uninfected after inoculation i and $n_g$is the number of animals in group g. We define $L_a(i)$ to be the indicator random variable for if animal a has been infected by ($L_a(i) = 1$) or remains uninfected at ($L_a(i) = 0$) week i. Therefore, $n_{S,g}(i)$ for study arm g is given by the number of animals in group g, n_g, minus the sum of the observations of $L_a(i)$ for the animals in group g. So, the expected value of $n_{S,g}(i)$ is

$$E\left(n_{S,g}(i)\right) = n_g - \sum_{a \in g} E\big(L_a(i)\big),$$

and it only remains to calculate $E\big(L_a(i)\big)$.

To calculate $E\big(L_a(i)\big)$, we use the model predicted probability of infection of animal a following an inoculation ($p(x_a, \boldsymbol{\theta})$ from Equations (S1.5) and (S1.7)) and define I_a as the random variable for the inoculation at which animal a is infected,

$$\begin{aligned} E\big(L_a(i)\big) &= 0 \cdot P\big(I_a > i; p(x_a, \widehat{\boldsymbol{\theta}})\big) + 1 \cdot P\big(I_a \le i; p(x_a, \widehat{\boldsymbol{\theta}})\big) \\ &= P\big(I_a \le i; p(x_a, \widehat{\boldsymbol{\theta}})\big) \\ &= 1 - \big(1 - p(x_a, \widehat{\boldsymbol{\theta}})\big)^i \end{aligned}$$

Therefore,

$$\begin{aligned} E\left(n_{S,g}(i)\right) &= n_g - \sum_{a \in g} 1 - \big(1 - p(x_a, \widehat{\boldsymbol{\theta}})\big)^i \\ &= \sum_{a \in g} \big(1 - p(x_a, \widehat{\boldsymbol{\theta}})\big)^i, \end{aligned}$$

and the fraction of animals in study arm g remaining uninfected after i inoculations (i.e., the survival curve) is

$$S_g(i) = \frac{1}{n_g} \sum_{a=1}^{n_g} \big(1 - p(x_a, \widehat{\boldsymbol{\theta}})\big)^i.$$

S1.5.4. Parameter confidence interval calculation

After estimating b based on the data ($\hat{b}$), we used bootstrapping to compute the 95% confidence interval on our estimate. We generated 1000 bootstrapped data sets in a manner that preserves the data structure. For each study arm g, ($g \in 0\%, 26\%, 54\%, 69\%$, and 96% *CCR5*-edited), n_g animals from that study arm were drawn with replacement (where n_g is the number of animals in study arm g). We then estimated b for each bootstrapped data set ($\hat{b}_k^*$ for bootstrapped data set k). The 95% confidence interval was then determined from the 2.5% and 97.5% quantiles of $\{\hat{b}_k^*\}_{k=1}^{1000}$, giving a 95% confidence interval of $(0.28, 0.43)$.

S1.5.5. Survival curve prediction bands calculation

To visualize the quality with which the model recapitulates the survival curves of each study arm, we generated 95% prediction bands for the survival curves based on the mechanistic model in Section 0 (shown in shaded regions in Figure S1-1 below). The prediction bands were generated from 1000 simulated survival curves, which incorporated the uncertainty of our parameter estimates. In providing the details of how these prediction bands were generated, we will first describe how each simulated survival curve was generated, and then describe how we incorporated the uncertainty in our parameter estimates.

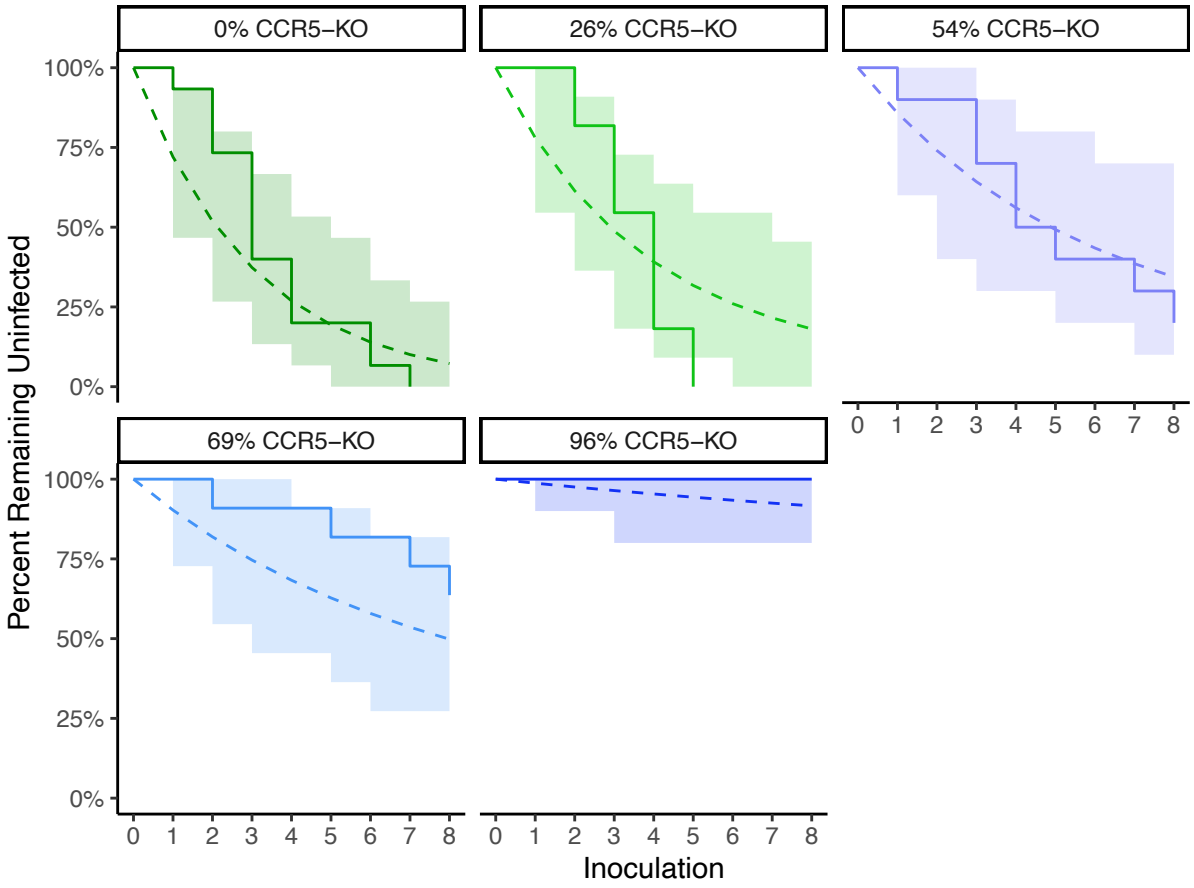


Figure S1-1: Observed vs. predicted survival curves of titration of CCR5-edited HSPCs experiment. As in Fig. 7E of the main text, Kaplan-Meier curves (dark solid lines) and corresponding expected survival curves (dark dashed lines) based on the mechanistic model are plotted. Additionally, 95% prediction intervals based on bootstrapping are indicated by the shaded regions. Each panel corresponds to a different experimental study arm.

Combining Equations (S1.5) – (S1.7), we can calculate the probability of an infection occurring after inoculation i for each animal, given their intact *CCR5* allele frequency in CD19$^+$ cells (x_a). We denote this probability by $P(I = i; b, x_a)$, where I is the random variable for infection inoculation and b is the linear coefficient defined in Section 0. We have:

$$\begin{aligned} P(I = i; x_a, b) &= \left(1 - p(x_a, b)\right)^{i-1} p(x_a, b)^1 \\ &= \left(e^{-\rho(x_a;b)}\right)^{i-1} \left(1 - e^{-\rho(x_a;b)}\right)^1 \\ &= \left(e^{-bx_a}\right)^{i-1} \left(1 - e^{-bx_a}\right)^1. \end{aligned}$$

Therefore, the probability that animal a was infected on or before the i^{th} inoculation is given by the cumulative distribution function, $F_I(i; x_a, b)$, where:

$$F_I(i; x_a, b) = P(I \leq i; x_a, b) = \sum_{s \leq i} P(I = s; x_a, b).$$

To generate a simulated data set (set of inoculations $\{i_a^*\}$ and corresponding infection indicators $\{d_a^*\}$), we stochastically simulated the inoculations at which each animal from the study became infected based on a given b (i.e., we parametrically bootstrapped the data). For this, we generated a random number, $q_a \in [0,1]$, for each animal within the study, and determined i_{q_a} such that

$$q_a \in \left[F_I\left(i_{q_a} - 1; x_a, b\right), F_I\left(i_{q_a}; x_a, b\right)\right).$$

If $i_{q_a} \leq 8$, we recorded the infection of animal a ($d_a^* = 1$) as occurring at the i_{q_a}th inoculation in the simulated data set ($i_a^* = i_{q_a}$), but if $i_{q_a} > 8$, the animal was recorded as censored after eight inoculations ($i_a^* = 8, d_a^* = 0$). Based on this simulated data set, Kaplan-Meier curves were then generated for each study arm.

In total, 1000 simulated data sets were generated, and to account for the uncertainty in the estimation of b , we selected a different estimate for b (sampled with replacement from $\{\hat{b}_k^*\}$ generated by the data bootstrapping in the previous subsection) for each simulated data set. The 95% prediction bands were then calculated as the 2.5% and 97.5% quantiles of the resultant Kaplan-Meier curves at each inoculation.

S1.6. Including dissemination effects in modeling relation between infection probability and *CCR5*-editing

In this section, we expand our model of the impact of *CCR5*-editing on infection probability (presented in Section S1.5) to include the effect that *CCR5*-editing may have on viral dissemination or ongoing replication within an animal. To differentiate the model presented in this section with that presented earlier, from now on we will refer to the model presented in Section S1.5 as the inoculation risk model (based on the "Barrier at Inoculation" mechanism described in the main text) and the model presented in this section as the dissemination risk model (based on the "Barrier at Dissemination" mechanism described in the main text).

The dissemination risk model assumes, once again, that the frequency of target cells affects the number of cells infected by the inoculation, but in addition, there is a chance that a lineage derived from an initially infected cell will go extinct. We assume that the probability of such an extinction increases as frequency of target cells decreases. Therefore, in this section, we split the process of establishment of infection following an inoculation in two phases; 1) infection of cells by the inoculation and 2) survival of at least one viral lineage initiated by cells infected by the inoculation. This section concludes by demonstrating that although the dissemination model agrees with our data, there is not sufficient evidence to support a model with dissemination effects over the model without (inoculation risk model).

The probability of m cells getting infected from an inoculation has been calculated in Equation (S1.4). Note that because the dissemination risk model consists of a different modelling framework than the inoculation risk model, we again start by assuming c may not be zero, even though we determined c is not significantly different from zero in the inoculation risk model.

In this new model we must account for the probability of all viral lineages established by initially infected cells going extinct ($P(\xi)$, where ξ is the event where all lineages go extinct). Towards this end, we adopt the standard model of viral replication within a host (i.e., three-dimensional system consisting of variables for target cell, infected cell, and virus concentrations, in which, it is assumed infected cells produce virus at a constant rate[2]), and therefore, the probability of a viral lineage established by a single infected cell going extinct is given by

$$\min\left(\frac{1}{R_0}, 1\right), \tag{S1.8}$$

where R_0 is the basic reproduction number [3-5]. By assuming loss of free virus via cell infection has a negligible effect on viral concentration, we have that R_0 is linearly dependent on the concentration of target cells[2,6]. In our model this translates into the assumption that R_0 is linearly proportional to the intact *CCR5* allele frequency. Hence, we write

$$\begin{aligned} R_0 &= R_0(x) \\ &= Rx, \end{aligned} \tag{S1.9}$$

where R is a constant. It follows from Equations (S1.8) and (S1.9) that the probability of all lineages from m infected cells going extinct is

$$P(\xi|C = m; x) = \min\left(\frac{1}{\left(R_0(x)\right)^m}, 1\right) = \min\left(\frac{1}{(Rx)^m}, 1\right). \tag{S1.10}$$

Combining Equations (S1.4) and (S1.10), we find that the probability of m cells getting infected and all lineages going extinct is

$$P(\xi\&(C = m); x, \boldsymbol{\theta}) = e^{-\rho(x;\boldsymbol{\theta})}\frac{\rho(x;\boldsymbol{\theta})^m}{m!}\min\left(\frac{1}{(Rx)^m}, 1\right). \tag{S1.11}$$

To obtain the probability of no long-term infection being established (i.e., either no cells infected or all lineages initiated by the inoculation go extinct), we sum Equation (S1.11) over all possible number of cells infected by the inoculation (m). To simplify the following equations, we will write $P\big(\xi\&(C = m)\big)$, ρ, and R_0 for $P(\xi\&(C = m); x, \boldsymbol{\theta})$, $\rho(x; \boldsymbol{\theta})$, and Rx, respectively. Assuming $R_0 = Rx > 1$,

$$\begin{aligned} P(\xi) &= \sum_{m=0}^{\infty} P\big(\xi \& (C = m)\big) \\ &= \sum_{m=0}^{\infty} e^{-\rho} \frac{\rho^m}{m!} \frac{1}{R_0^m} \\ &= e^{-\rho} \sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{\rho}{R_0}\right)^m \\ &= e^{-\rho} e^{\frac{\rho}{R_0}} = e^{-\rho \frac{R_0 - 1}{R_0}}, \end{aligned} \tag{S1.12}$$

since $\sum_{m=0}^{\infty} \frac{1}{m!} \left(\frac{\rho}{R_0}\right)^m$ is the Maclaurin series for e^{ρ/R_0}. If $R_0 = Rx \leq 1$,

$$\begin{aligned} P(\xi) &= \sum_{m=0}^{\infty} P\big(\xi \& (C = m)\big) \\ &= \sum_{m=0}^{\infty} e^{-\rho} \frac{\rho^m}{m!} 1 \\ &= e^{-\rho} \sum_{m=0}^{\infty} \frac{\rho^m}{m!} \\ &= e^{-\rho} e^{\rho} = 1 \end{aligned} \tag{S1.13}$$

since $\sum_{m=0}^{\infty} \frac{\rho^m}{m!}$ is the Maclaurin series for e^{ρ}.

Therefore, from Equations ((S1.12)and (S1.13)) we can write the probability of infection being established after an inoculation (p_d to differentiate from that of the inoculation risk model) as

$$p_d(x, \boldsymbol{\theta}) = 1 - P(\xi; x, \boldsymbol{\theta}) = \begin{cases} 1 - e^{-\rho(x;\boldsymbol{\theta}) \frac{Rx-1}{Rx}}, & Rx > 1 \\ 0, & Rx \leq 1 \end{cases} \tag{S1.14}$$

Note that, Equation (S1.14) demonstrates that the dissemination risk model predicts a threshold intact *CCR5* allele frequency ($x = \frac{1}{R}$), below which established infection is impossible.

As in Equation (S1.6), the probability of animal a being infected ($d_a = 1$) or censored ($d_a = 0$) at inoculation i_a is dependent on the probability of infection being established after inoculation. Therefore, we have

$$P(i_a, d_a; x_a, \boldsymbol{\theta}) = \big(1 - p_d(x_a, \boldsymbol{\theta})\big)^{i_a - d_a} p_d(x_a, \boldsymbol{\theta})^{d_a} \tag{S1.15}$$

but here the probability of infection is represented by p_d (as this is the dissemination risk model) and the vector of parameters is $\boldsymbol{\theta} = \langle b, c, R \rangle$. Therefore, the log-likelihood of the dissemination risk model with all data included is

$$\ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{a=1}^{n} (i_a - d_a) \ln\big(1 - p_d(x_a, \boldsymbol{\theta})\big) + d_a \ln\big(p_d(x_a, \boldsymbol{\theta})\big). \qquad \text{(S1.16)}$$

We estimated the optimal parameters ($\widehat{\boldsymbol{\theta}}$) corresponding to the maximum value of $\ln \mathcal{L}(\boldsymbol{\theta})$ using the *nlm* function from the *stats* package in R. Expected survival curves for the best fit dissemination model ($b = 0.31, c = 0.029$, and $R = 36$; 95% confidence intervals of $(0.18, 0.42)$, $(-0.025, 0.14)$, and $(9.0, 94)$, respectively) are plotted in Figure S1-2A (dashed lines), and panel B displays the corresponding probability of infection as a function of *CCR5*-editing (red line).

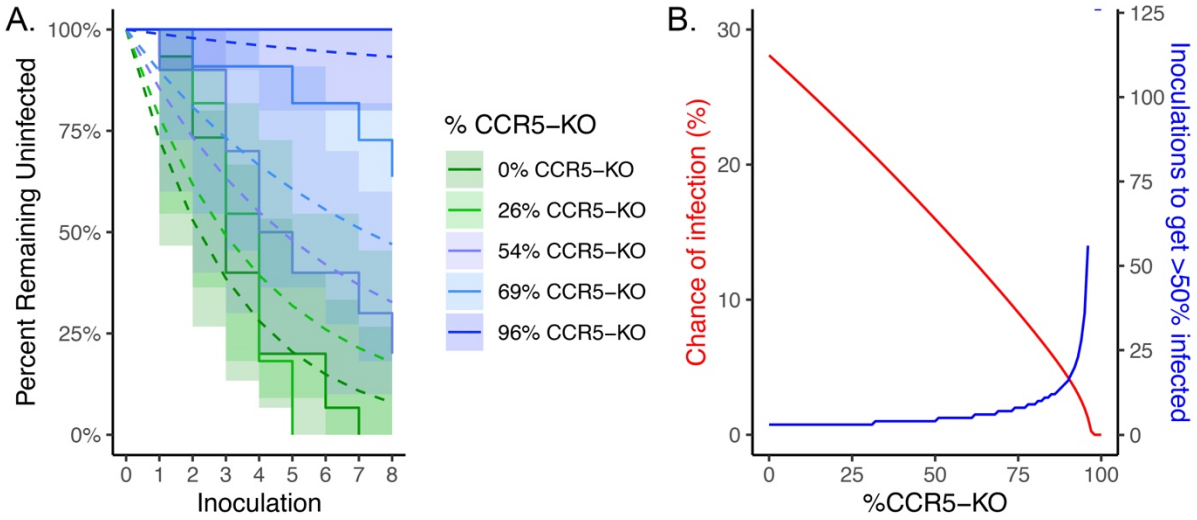


Figure S1-2: Dissemination risk model output. (A) Best fit expected survival curves (dashed lines) compared to observed survival curves (solid lines) for each study arm. Shaded regions indicate bootstrapped 95% prediction intervals for each study arm using similar methodology to that outlined in Sections S1.5.4 and S1.5.5 for the inoculation risk model. (B) Best fit probability of infection (red) and number of inoculations required to infect 50% of animals (blue; see Section S1.7) as functions of %CCR5-knock out (%CCR5-KO).

Note, that the difference in $\mathrm{AIC_c}$ between the dissemination risk model and the inoculation risk model is 3.79 (i.e., $\mathrm{AIC_c}$ for the inoculation risk model is 3.79 lower than the dissemination risk model). Therefore, although the model is consistent with the data and it is intriguing to consider the existence of a *CCR5*-editing threshold above which chronic infection is impossible, the current data does not support such a model over the simpler inoculation risk model.

S1.7. Calculation of probability of infection by eighth inoculation and number of inoculations required to infect greater than 50% of animals

Figure 7F and G in the main text displays the relationship between the number of inoculations required to infect greater than 50% of animals (blue line in Figure 7F) and the probability of infection by the eighth inoculation (black line in Figure 7G), respectively, against the percentage of *CCR5*-editing. In this section, we derive the mathematical formula for these relationships.

We note that the probability of infection per inoculation (designated here as p) is sufficient to calculate the probability of infection by the eighth inoculation and the number of inoculations required to infect greater than 50% of animals. Therefore, in this section, our derivations are only in terms of p, but to obtain the dependence on the percent of *CCR5*-editing, one merely needs to substitute in the formula from the inoculation risk model or the dissemination risk model (i.e., Equations (S1.5) and (S1.7) or p_d in Equation (S1.14), respectively, where x is the intact *CCR5* allele frequency in $CD19^+$ cells and $\boldsymbol{\theta}$ is the estimated parameter vector). For the plots in Figs 7F and G in the main text, we use the formula for p from Equations (S1.5) and (S1.7).

As indicated in Section S1.5, the probability of an animal not being infected after i inoculations is $(1-p)^i$. Thus, the probability of being infected by the end of i inoculations is

$$1-(1-p)^i,$$

and the probability of infection after eight inoculations is

$$1-(1-p)^8.$$

Additionally, the number of inoculations required to get at least 50% of animals infected, i_{50}, is the minimum i, such that

$$0.5 \leq 1-(1-p)^i.$$

Therefore,

$$\begin{aligned} (1-p)^{i_{50}} &\leq 0.5 \\ i_{50}\ln(1-p) &\leq \ln 0.5 \\ i_{50} &\geq \frac{\ln 0.5}{\ln(1-p)} \quad (\text{since } \ln(1-p) \leq 0) \end{aligned}$$

and hence

$$i_{50} = \left\lfloor \frac{\ln 0.5}{\ln(1-p)} \right\rfloor,$$

where $\lfloor \alpha \rfloor$ is the floor of α.

S1.8. Modelling the implications of *CCR5*-editing on viral rebound

High frequency *CCR5* editing in human hematopoietic stem progenitor cells protects xenograft mice from HIV infection

Figure 8B of the main text displays the impact of *CCR5*-editing on expected time-to-rebound and set-point viral load in people living with HIV (PLWH), as inferred by our modelling of the effect of *CCR5*-editing on the probability of infection following an inoculation in humanized mice. Here, we derive the equations used for the lines plotted in Fig 8B.

S1.8.1. Expected time-to-rebound

Following ART interruption, viral rebound typically occurs at a rate of once per week (expected time-to-rebound of 1 week)[7,8]. Given that *CCR5*-editing decreases the number of target cells, we assume that *CCR5*-editing may influence the time-to-rebound. To derive a formula to estimate the effect of *CCR5*-editing on expected time-to-rebound, we assume that the timing of successful rebound is exponentially distributed with a rebound rate defined as $r(x)$, which is a function of the intact *CCR5* allele frequency in CD19$^+$ cells. This means that the probability distribution function for the time-to-rebound (t), $h(t;x)$ is given by

$$h(t;x) = r(x)e^{-r(x)t},$$

and the expected time-to-rebound is

$$\begin{aligned} E(t) &= \int_{t=0}^{\infty} h(t;x)t\,dt \\ &= \int_{t=0}^{\infty} r(x)te^{-r(x)t}\,dt \\ &= \frac{1}{r(x)}\,. \end{aligned} \tag{S1.17}$$

We must now derive the expression for $r(x)$. To this end, we consider each reactivation of a latently infected cell as a new "inoculation." Therefore, the rate of successful reactivations that lead to detectible rebound is the product of the reactivation rate of latently infected cells (r_0) and the probability of a "reactivation inoculation" infecting at least one cell. To compute the probability of a "reactivation inoculation" infecting at least one cell, we use the structure of the inoculation risk model for humanized mice (i.e., Equations (S1.5) and (S1.7)). However, as we are now considering the context of rebound in PLWH, the coefficient for the expected number of cells infected (b in Equation (S1.7)) is different from the context of infection of humanized mice. Therefore, we use b_h to designate the coefficient for the expected number of cells infected following reactivation of a latent cell in PLWH, so the probability of a "reactivation inoculation" infecting at least one cell is $1 - e^{-b_h x}$. Hence, the rate of successful reactivations that lead to detectible rebound is

$$r(x) = r_0(1 - e^{-b_h x}). \tag{S1.18}$$

To estimate r_0 and b_h, we use the published rate of detectible rebound ($1\ \text{week}^{-1}$ [7,8]) and basic reproduction number ($R_0 = 8$ [6]) for PLWH. First, by substituting b_h for b in Equation (S1.7) and noting that 100% of CD19$^+$ cells contain the intact *CCR5* allele in PLWH (i.e., $x = 1$), the expected number of cells infected by the reactivation of a latently infected cell is $E(C) = \rho(1) = b_h 1 = 8$, hence $b_h = 8$. Then, substituting this into Equation (S1.18) and setting $r(1)$ (i.e., the rate of detectible rebound for people with 100% intact *CCR5* allele frequency) equal to $1\ \text{week}^{-1}$, we have

$$r(1) = 1 = r_0(1 - e^{-8*1})$$

or by rearranging

$$r_0 = \frac{1}{1 - e^{-8}}, \tag{S1.19}$$

and substituting the formula for r_0 defined in Equation (S1.19) above into Equation (S1.18) we get

$$r(x) = \frac{(1 - e^{-8x})}{(1 - e^{-8})}.$$

Thus, by Equation (S1.17), the expected time to viral rebound is

$$E(t; x) = \frac{1}{r(x)} = \frac{(1 - e^{-8})}{(1 - e^{-8x})},$$

which is the formula for the expected time-to-rebound plotted in Fig 8B (red line).

S1.8.2. Fold reduction in set-point

Along with time-to-rebound, we expect *CCR5*-editing during ART to alter set-point viral load following rebound. The blue line in Fig 8B shows the estimated fold reduction in set-point viral load compared to the percentage knock out of the *CCR5* allele and is based on the modelling used in[9]. In brief, in the standard model of within host HIV infection[2], the set-point viral load, V_{set}, is proportional to the basic reproductive number minus one. We write,

$$V_{\text{set}} = \gamma(R_0 - 1),$$

where γ is a constant. Additionally, R_0 is proportional to the frequency of target cells[2], so we can write

$$R_0 = R_{0,H}x,$$

where $R_{0,H} = 8$ is the estimated basic reproductive number for HIV[6] and x is as defined above (intact *CCR5* allele frequency in CD19$^+$ cells). Therefore, we have

$$V_{\text{set}}(x) = \gamma(8x - 1),$$

and the relative reduction in set-point viral load due to having a proportion of cells with *CCR5*-editing compared to in the absence of *CCR5*-editing is

$$\begin{aligned} \Delta V_{\text{set}} &= \frac{V_{\text{set}}(x)}{V_{\text{set}}(1)} \\ &= \frac{\gamma(8x - 1)}{\gamma(8 - 1)}. \\ &= \frac{8x - 1}{7} \end{aligned} \tag{S1.20}$$

Equation (S1.20) is plotted in blue in Fig 8B of the main text.

S1.9. Statistical packages

Computation and statistical tests were performed using R version 4.0.3.

References

1 Burnham, K. P. & Anderson, D. R. *Model selection and multimodel inference: a practical information-theoretic approach*. 2 edn, (Springer, 2002).

2 Nowak, M. A. & May, R. M. *Virus dynamics : mathematical principles of immunology and virology*. (Oxford University Press, 2000).

3 Pearson, J. E., Krapivsky, P. & Perelson, A. S. Stochastic theory of early viral infection: continuous versus burst production of virions. *PLoS Comput Biol* **7**, e1001058, doi:10.1371/journal.pcbi.1001058 (2011).

4 Whittle, P. The Outcome of a Stochastic Epidemic. *Biometrika* **42**, 116-122, doi:Doi 10.2307/2333427 (1955).

5 Allen, L. J. & Lahodny, G. E., Jr. Extinction thresholds in deterministic and stochastic epidemic models. *J Biol Dyn* **6**, 590-611, doi:10.1080/17513758.2012.665502 (2012).

6 Ribeiro, R. M. *et al.* Estimation of the initial viral growth rate and basic reproductive number during acute HIV-1 infection. *J Virol* **84**, 6096-6102, doi:10.1128/JVI.00127-10 (2010).

7 Pinkevych, M. *et al.* HIV Reactivation from Latency after Treatment Interruption Occurs on Average Every 5-8 Days—Implications for HIV Remission. *PLoS Pathog* **11**, e1005000, doi:10.1371/journal.ppat.1005000 (2015).

8 Pinkevych, M. *et al.* Modeling of Experimental Data Supports HIV Reactivation from Latency after Treatment Interruption on Average Once Every 5-8 Days. *PLoS Pathog* **12**, e1005740, doi:10.1371/journal.ppat.1005740 (2016).

9 Davenport, M. P. *et al.* Functional cure of HIV: the scale of the challenge. *Nat Rev Immunol* **19**, 45-54, doi:10.1038/s41577-018-0085-4 (2019).