RESEARCH ARTICLE

Pseudogene-Mediated Gene Conversion After CRISPR-Cas9 Editing Demonstrated by Partial *CD33* Conversion with *SIGLEC22P*

Benjamin C. Shaw and Steven Estus*

Abstract

Although gene editing workflows typically consider the possibility of off-target editing, pseudogene-directed homology repair has not, to our knowledge, been reported previously. Here, we employed a CRISPR-Cas9 strategy for targeted excision of exon 2 in *CD33* in U937 human monocyte cell line. Candidate clonal cell lines were screened by using a clinically relevant antibody known to label the IgV domain encoded by exon 2 (P67.6, gem-tuzumab). In addition to the anticipated deletion of exon 2, we also found unexpected P67.6-negative cell lines, which had apparently retained *CD33* exon 2. Sequencing revealed that these lines underwent gene conversion from the nearby *SIGLEC22P* pseudogene during homology repair that resulted in three missense mutations relative to *CD33*. Ectopic expression studies confirmed that the P67.6 epitope is dependent upon these amino acids. In summation, we report that pseudogene-directed homology repair can lead to aberrant CRISPR gene editing.

Introduction

The CRISPR-Cas9 system has revolutionized gene editing.¹ In this process, a single guide RNA (sgRNA) directs Cas9 endonuclease to cleave DNA at a sequence-specific site. The DNA cleavage results in a double-strand DNA break (DSB) that is repaired by either homology-directed repair (HDR) or non-homologous end joining (NHEJ). The former results in targeted integration of DNA sequence, while the latter typically results in gene disruption through the introduction of insertions or deletions (indels). To generate a targeted knock-in or knock-out, an HDR template of exogenous DNA is often supplied as part of the process.² Alternatively, endogenous HDR templates have been described, including *HBD* sequence being incorporated into *HBB* and sequence from one allele of *HPRT* being incorporated into the other allele.^{3,4} However, to our knowledge, HDR directed by a pseudogene has not been previously reported.

CD33 genetic variants, including rs12459419, have been associated with a reduced risk of Alzheimer's disease (AD) in genome-wide studies.^{5–7} We and others subsequently identified rs12459419 as a functional single

nucleotide polymorphism (SNP) that increases the proportion of *CD33* lacking exon 2 (*D2-CD33*).^{8–12} This exon encodes the ligand-binding IgV domain of this member of the sialic acid-binding immunoglobulin-type lectin (SIGLEC) family.¹³ Hence, while the extracellular portion of *CD33* normally includes an IgV and IgC₂ domain, *D2-CD33* encodes a protein with only the IgC₂ domain.¹⁴ *CD33* inhibits microglial activity through its immunomodulatory tyrosine inhibitory motif (ITIM) and ITIM-like domains, which recruit protein tyrosine phosphatases, SHP1 and SHP2, to impact intracellular calcium flux, phagocytosis, and microglial migration.^{9,11,12,15–20}

Given that the AD-protective rs12459419 increases *D2-CD33* at the expense of *CD33*, the prevailing theoretical mechanism has been that rs12459419 reduces AD risk through decreased *CD33* function. However, recent findings that a *bona fide* loss of function indel, rs201074739, is not associated with AD risk has led to this hypothesis being revised to suggest that rs12459419 and its related *D2-CD33* isoform represent a gain of function.^{13,21,22} The gain-of-function mechanism and localization of *D2-CD33* protein remain heavily debated.^{8,9,11,13,17,20–25}

Department of Physiology and Sanders-Brown Center on Aging, University of Kentucky, Lexington, USA.

*Address correspondence to: Steve Estus, PhD, Sanders-Brown Center on Aging, 789 S. Limestone St., Lexington, KY 40536, USA, Email: steve.estus@uky.edu

Here, we sought to generate a model of physiologic D2-CD33 expression by using CRISPR-Cas9 to excise *CD33* exon 2 in the U937 human monocyte cell line. During these experiments, we identified a subset of cells that apparently underwent HDR directed by the *SIGLEC22P* pseudogene, located 13.5 kb away from *CD33*. Although the *SIGLEC22P* pseudogene shares approximately 87% identity over 1,800 bp with *CD33*, this gene conversion was detected because three nucleotides in *SIGLEC22P* differ from those within the targeted *CD33* exon 2 and result in three missense amino acids in CD33: p.N20K, p.F21I, and p.W22R. Hence, we report pseudogene-directed gene conversion as a mechanism for unanticipated CRISPR mutations.

Methods

Cell lines and antibodies

U937 and HEK293 cell lines were obtained from American Type Culture Collection (ATCC). U937 cells were cultured in RPMI 1640 with HEPES (Gibco 22400-089) supplemented with 10% fetal bovine serum (FBS), defined (HyClone, GE Healthcare SH30070.03); 50 IU/mL penicillin, 50 μ g/mL streptomycin (Gibco 15070-063); and 2 μ M L-glutamine (Gibco A2916801). HEK293 cells were cultured in Eagle's Minimum Essential Medium, ATCC formulation (ATCC 30-2003) supplemented with 10% FBS, defined (HyClone, GE Healthcare SH30070.03); 50 IU/mL penicillin, 50 μ g/mL streptomycin (Gibco 22400-089). Cells were maintained at 37°C in a 5% CO₂ in air atmosphere. The U937 cell line has been reported as either diploid or triploid at chromosome 19, which contains *CD33*.^{26,27} Antibodies, concentrations, and CD33 domains targeted are shown in Table 1.

CRISPR-Cas9 gene editing

CRISPR reagents were purchased from Integrated DNA Technologies (IDT). sgRNAs and *Streptococcus pyogenes* Cas9 protein (IDT 1081059) were incubated at a 1:1 molar ratio (0.5 nmol each) at room temperature for 10 min to form ribonucleotide-protein complexes (RNPs). The

sgRNA sequences targeting *CD33* exon 2 were 5'-TCCA TAGCCAGGGCCCCTGT and 5'-GCATGTGACAGGT GAGGCAC.²⁵ U937 cells were washed three times in phosphate-buffered saline (PBS; Gibco 10010-023) and re-suspended in complete Nucleofector Kit C (Lonza Biosciences VCA-1004) media (10⁶ cells per transfection) with 5 μ L electroporation enhancer (IDT 1075916) and RNPs. Cells were electroporated using a Nucleofector IIb device (Lonza Biosciences) under protocol V-001 and immediately added to a 12-well plate with 1.5 mL complete media and cultured for 2 weeks.

Cell sorting and flow cytometry

Edited U937 cells were washed in PBS with 5% heat-inactivated FBS (Gibco 10082-147), re-suspended at 10⁶ cells/mL, and then treated with Human TruStain FcX blocker (BioLegend 422302). Cell sorting was carried out in azide-free buffers; for flow cytometry, 0.02% sodium azide was included in all buffers. Cells were stained with HIM3-4-FITC and P67.6-BV711 for 1 h on ice, washed twice with Hank's Balanced Salt Solution (HBSS), and then stained with Fixable Viability Dye eFluor780 (Invitrogen 65-0865-18). Cells were resuspended in HBSS (Gibco 24020-117) with 5% heat-inactivated FBS (Gibco 10082-147) for sorting. Viable cells were gated using scatter and viability exclusion stain, sorted as either HIM3-4⁺ P67.6⁺, HIM3-4⁺ P67.6⁻, or HIM3-4⁻ P67.6⁻, and collected in complete media. At 48 h post sort, cells were split using limiting dilution on a 96-well plate at an average density of 0.5 cells/well and expanded until sufficient cell numbers for analysis were achieved, which was after approximately 8 weeks. Clones were screened by flow cytometry again prior to polymerase chain reaction (PCR) and sequence analysis.

PCR screening and cloning

Genomic DNA from CRISPR-edited U937 clones was isolated with a DNeasy kit (Qiagen 69506) as per the manufacturer's instructions. A portion of *CD33* was amplified with Q5 High-Fidelity DNA Polymerase (New

Table 1. Antibodies Used in This Study

Clone	Conjugate	Vendor	Catalog #	Lot	Dilution/Use	CD33 domain targeted
P67.6	BV711	BioLegend	366624	B302694	2.5 μ g/mL FC	IgV
P67.6	Unconjugated	BioLegend	825601	B258818	5 μ g/mL IF	IgV
WM53	BV711	BioLegend	303424	B253729	5 μ g/mL FC	IgV
WM53	Unconjugated	BioLegend	96281	B274701	5 μ g/mL IF	IgV
HIM3-4	FITC	BioLegend	303304	B284834	20 μ g/mL FC	IgC ₂
PWS44	Unconjugated	Leica	NCL-L-CD33	6024275	2 μ g/mL IF	IgC ₂
H110	Unconjugated	Santa Cruz	sc-28811	K2211	1 μ g/mL IF	Intracellular
Goat anti-Mouse	AlexaFluorPlus 488	Invitrogen	A32723	TF266577	10 μ g/mL IF	
Goat anti-Rabbit F(ab) ₂	AlexaFluor 568	Invitrogen	A21069	2087701	10 μ g/mL IF	

IF, immunofluorescence; FC, flow cytometry.

England BioLabs M0439L) using forward primer 5'-CACAGGAAGCCCTGGAAGCT and reverse primer 5'-GAGCAGGTCAGGTTTTTGGGA (Invitrogen). *SIGLE C22P* was amplified similarly with forward primer 5'-GCACCTCAGAGTGAAGGAC and reverse primer 5'-GAAGGGGTGACTGAGGTACA. Thermocycling parameters were as follows: 98°C for 1 min; 98°C for 15 s, 66°C for 15 s, 72°C for 45 s, 32 cycles; 72°C for 2 min, 25°C hold. PCR products were separated on a 0.8% agarose-TBE gel, purified using a Monarch gel extraction kit (New England BioLabs T1020L), and sequenced by a commercial company (ACGT, Inc.). The three missense mutations identified were introduced into a previously described pcDNA3.1-*CD33*-V5/HIS vector using a QuikChange XL kit (Agilent 200517) with forward primer 5'-GCACTTGCAGCCGGATTTTTGGATC CATAGCCAGGGCC and reverse primer 5'-GGCCCTG GCTATGGATCCAAAAATCCGGCTGCAAGTGC (Invitrogen) to generate the pcDNA3.1-*KIRCD33*-V5/HIS vector, transformed into TOP10 *Escherichia coli* (Invitrogen C404003), isolated using a Plasmid Plus Midiprep Kit (Qiagen 12945), and verified by sequencing (ACGT, Inc.).⁸

Gene expression by quantitative PCR

Quantitative PCR was used to quantify expression of total *CD33* and *D2-CD33*, as previously described.¹⁴ Briefly, primers corresponding to sequences within exons 4 and 5 were used to quantify total *CD33* expression (forward, 5'-TGTTCCACAGAACCCAACAA-3'; reverse, 5'-GGCTG TAACACCAGCTCCTC-3'), as well as primers corresponding to sequences at the exon 1–3 junction and exon 3 to quantify the *D2-CD33* isoform (forward, 5'-CCCTGC TGTGGGCAGACTTG-3'; reverse, 5'-GCACCGAGGAG TGAGTAGTCC-3'). PCR was conducted using an initial 2 min incubation at 95°C, followed by cycles of 10 s at 95°C, 20 s at 60°C, and 20 s at 72°C. The 20 μ L reactions contained 1 μ M each primer, 1 $\times$ PerfeCTa SYBR Green Super Mix (Quanta Biosciences), and 20 ng cDNA. Experimental samples were amplified in parallel with serially diluted standards that were generated by PCR of cDNA using the indicated primers, followed by purification and quantitation by UV absorbance. Results from samples were compared relative to the standard curve to calculate copy number in each sample. Assays were performed in triplicate and normalized to expression of ribosomal protein L32 (*RPL32*) as the housekeeping gene.^{14,28}

HEK293 transfection

HEK293 cells were seeded at approximately 70% confluency 24 h before transfection. Cells were then transfected with Lipofectamine 3000 with Plus Reagent (Invitrogen

L3000001) as per the manufacturer's instructions, 250 ng plasmid per well in eight-well glass chamber slides (MatTek CCS-8) for immunofluorescence or 1,000 ng per well in 12-well plates (Corning 3513) for flow cytometry. Cells were transfected with either the previously described wild-type *CD33* vector (pcDNA3.1-*CD33*-V5/HIS), pcDNA3.1-*KIRCD33*-V5/HIS, or no vector control. Cells were incubated for 24 h before analysis by flow cytometry or immunofluorescence and confocal microscopy.

Confocal immunofluorescence microscopy

Transfected HEK293 cells were fixed with 10% neutral buffered formalin (Thermo Fisher Scientific SF100-4) for 30 min and then blocked and permeabilized for 30 min with 10% goat serum (Sigma–Aldrich S26-LITER), 0.1% Triton X-100 (Thermo Fisher Scientific BP151-500) in PBS (Fisher BioReagents BP665-1). Primary and secondary antibodies were diluted in the same blocking and permeabilization buffer and incubated at room temperature for 90 min. Cells were washed three times in blocking and permeabilization buffer between primary and secondary antibodies, and three times in PBS prior to coverslip mounting with Prolong Glass with NucBlue mounting media (Invitrogen P36981) and high-tolerance No. 1.5 coverglass (ThorLabs CG15KH1). Images were acquired using a Nikon A1R HD inverted confocal microscope with a 60 $\times$ oil objective and NIS Elements AR software.

Statistical analyses

Analyses were performed using GraphPad Prism v8.4.2. Gene expression data were analyzed by one-way analysis of variance followed by Dunnett's multiple comparisons to unedited U937 cells.

Results

CRISPR-Cas9-mediated *CD33* exon 2 deletion leads to loss of P67.6 epitope

To generate an *in vitro* model of *D2-CD33*, we targeted exon 2 for deletion by using guide RNAs corresponding to sequences in the flanking introns, as previously described.²⁵ Cells were transfected, maintained for 2 weeks, and sorted according to Figure 1. Live cells were gated by light scatter (Fig. 1A). Singlet events were identified (Fig. 1B), and cells were sorted into separate tubes based on *CD33* phenotypes (Fig. 1C), with unstained cells shown for reference (Fig. 1D). *CD33* immunophenotype was determined with antibodies P67.6 and HIM3-4, which target epitopes in IgV and IgC₂ that are encoded by exon 2 and exon 3, respectively. *CD33* domains targeted by each antibody used in this study are shown in

FIG. 1. Sorting CRISPR-Cas9 *CD33* exon 2-edited cells reveals to two cell surface phenotypes. Gates are labeled in each panel and correspond to the following panel, with percent of parent gate shown. **(A)** Forward scatter (FSC) by side scatter gating was used to gate out dead cells ($FSC^{low} SSC^{high}$). **(B)** Singlet events were selected along the FSC-Area by FSC-Height diagonal. **(C)** Populations were identified as either unedited ($HIM3-4^{+} P67.6^{+}$), potential D2-CD33 or KO-CD33 ($HIM3-4^{-} P67.6^{-}$), or unexpected ($HIM3-4^{+} P67.6^{-}$). **(D)** Unstained control (blue) shown on top of the sorted cells (red) for reference. This unexpected population occurred at approximately a 1-to-9 frequency with respect to the D2- or KO-CD33 cells, or 10% of all edited cells. Color images are available online.

Table 1. We found that of the 396,789 sorted cells, 91.7% fell outside of the unedited cell gate, and we presume these cells contain a CRISPR-mediated change in *CD33* sequence. Clonal cell lines were established from bulk collection of the gates drawn in Figure 1C. These cell lines were subsequently re-examined by flow cytometry with the same P67.6 and HIM3-4 antibodies. While unedited cells showed robust labeling by both HIM3-4 and P67.6 (Fig. 2A), edited cell lines showed strong labeling by HIM3-4 but not P67.6 (Fig. 2B), or no labeling by either HIM3-4 or P67.6 (Fig. 2C). Data are representative of three independently established cell lines for each phenotype. Since D2-CD33 protein is not readily apparent on the cell surface,^{8,21,23–25,29} we expected that the latter cells (Fig. 2C) were candidates for exon 2 excision, which was confirmed

by a PCR product of the appropriate size (Fig. 2D, right) and by sequencing. However, cell lines with robust cell surface HIM3-4 but not P67.6 labeling were unexpected. Screening by the size of the PCR amplicon with primers corresponding to exon 1 and exon 3 suggested that exon 2 was still present (Fig. 2D, middle). Sequencing of this PCR fragment revealed that the $HIM3-4^{+} P67.6^{-}$ clones contained three apparent SNPs in exon 2 compared to the unedited wild-type (WT-CD33) U937 cell line. These SNPs have an identical minor allele frequency (MAF) = 9.86×10^{-5} and are indexed as rs3987761, rs3987760, and rs35814802.³⁰ Introduction of these SNPs results in changes in three consecutive amino acids (p.N20K, p.F21I, p.W22R), which we refer to as KIR-CD33. The nonsynonymous amino acids are the 4th–6th amino-terminal residues of the mature protein.

FIG. 2. CRISPR-Cas9 editing of *CD33* exon 2 leads to loss of P67.6 epitope. **(A)** Unedited U937 cells display robust P67.6 and HIM3-4 labeling. **(B)** Edited U937 clone that is robustly labeled by HIM3-4 but not P67.6. **(C)** Edited U937 clone that is not labeled by either HIM3-4 or P67.6. The depicted results in **(B)** and **(C)** are representative of at least three clonal cell populations established for each phenotype. **(D)** Genomic DNA polymerase chain reaction (PCR) of *CD33* exon 1 to exon 3 of the above cell lines. PCR products at 789 and 428 bp correspond to the expected sizes for the presence and absence of exon 2, respectively. **(E)** *KIR-CD33* mutations do not affect total *CD33* gene expression as determined by quantitative PCR (qPCR), but removal of exon 2 increases total *CD33* gene expression by 39.8%. **(F)** *KIR-CD33* mutations do not affect splicing efficiency of exon 2 as determined by qPCR, but removal of exon 2 at the genomic level increases exon 1–exon 3 junction to 100% of total *CD33* expression. Data from **(E)** and **(F)** analyzed by one-way analysis of variance followed by Dunnett's multiple comparisons test to unedited control. ** $p < 0.01$; **** $p < 0.0001$. Color images are available online.

EMBOSS and PSORT II predict cell surface localization for KIR-CD33 and no change in the signal peptide cleavage site.^{31,32} Consistent with typical cell-surface localization, HIM3-4 labeled both CD33 and KIR-CD33 in a similar fashion (Fig. 2A and B). Notably, total *CD33* gene expression and exon 2 splicing in KIR-CD33 cells does not differ from that of unedited cells (Fig. 2E and F). *CD33* expression was increased in D2-CD33 cells (Fig. 2E), which exclusively express the *D2-CD33* isoform (Fig. 2F).

Validation of the CRISPR-induced in-frame disruption of the P67.6 epitope

To demonstrate rigorously that the lack of P67.6 labeling was solely due to the KIR mutations, we introduced the

KIR mutations into a previously described *CD33* expression vector.⁸ *CD33* and *KIR-CD33* vectors were transfected into HEK293 cells, which do not naturally express *CD33*, and the cells processed for flow cytometry (Fig. 3) and immunofluorescent confocal microscopy (Fig. 4). For flow cytometry, cells were labeled with P67.6 or WM53, each of which was conjugated to the same fluorochrome to facilitate direct comparisons (Fig. 3). Importantly, these antibodies both target exon 2, are both mouse IgG1 κ , and have comparable degrees of labeling. In the *CD33* HEK293 cells, labeling with both WM53 and P67.6 correlated well with HIM3-4 (Fig. 3A and B). However, in the *KIR-CD33* HEK293 cells, labeling with WM53 but not P67.6 correlated with HIM3-4 labeling, as P67.6 labeling was

FIG. 3. Loss of P67.6 epitope is in-frame and preserves cell surface expression. (A) and (B) WT-CD33 or (C) and (D) KIR-CD33 expressing HEK293 cells were labeled with HIM3-4 and either (A) and (C) WM53 or (B) and (D) P67.6. HIM3-4⁺ cells identified in the "Transfected" gate (A–D) were gated to show (E) WM53 or (F) P67.6 binding in transfected cells. Color images are available online.

FIG. 4. Loss of P67.6 epitope does not alter other common CD33 epitopes. **(A–C)** WT-CD33 or **(D–F)** KIR-CD33 HEK293 cells were labeled with H110 (red) and either **(A)** and **(D)** WM53 (green), **(B)** and **(E)** P67.6 (green), or **(C)** and **(F)** PWS44 (green) and DAPI (blue). Color images are available online.

not apparent (Fig. 3C and D). Further gating on the HIM3-4⁺ cells shows that WM53 labeling is not affected by the *KIR-CD33* mutation (Fig. 3E), while P67.6 labeling in *KIR-CD33* cells is comparable to non-transfected control cells (Fig. 3F). These results were confirmed with immunofluorescent confocal microscopy using an array of anti-CD33 antibodies (Fig. 4). The *CD33*-transfected HEK293 cells showed consistent double labeling between an antibody against a cytoplasmic epitope (H110) and either WM53, P67.6, or PWS44 (Fig. 4A–C). For the *KIR-CD33* cells, robust co-labeling was observed between H110 and WM53 or PWS44 (Fig. 4D and F). However, P67.6 labeling of *KIR-CD33* cells was not detected (Fig. 4E). We thus conclude that the residues identified—p.N20, p.F21, and p.W22—are necessary for P67.6 binding and that these changes are the reason that this antibody failed to label the CRISPR-edited cells. Although P67.6 has been humanized and used clinically, prior studies have not mapped its epitope at this resolution.^{17,33,34}

Identification of *SIGLEC22P* as a HDR template for *CD33*

The three nucleotide changes (Fig. 5, red) are 23–27 bp away from the putative Cas9 cleavage site (Fig. 5, scissors), were not consecutive, and were present and identical in each of the clones with the KIR-CD33 phenotype.

Since this was unlikely due to chance, we hypothesized that this was due to HDR from elsewhere in the genome. A search of a 50 bp sequence centered on the *KIR* mutations revealed that this sequence occurs in the *SIGLEC22P* pseudogene, which is located 13.5 kb away from *CD33*. Further investigation found an extended region of homology between *CD33* and *SIGLEC22P* that flanked the Cas9 cleavage site. Indeed, the first 500 bp of the genes share 97% identity, including a 143 bp region of otherwise complete identity centered on the *KIR* mutations (Fig. 5, underline). The *SIGLEC22P* pseudogene exon 2 contains 11 additional mutations as well as two intronic mutations relative to *CD33*. None of these mutations were detected in the CRISPR-edited cell lines, indicating the region used for repair was limited to, at most, the 143 bp region surrounding the *KIR* mutations near the Cas9-induced DSB. We interpret these results as indicating that upon DSB at the beginning of exon 2 in *CD33* (Fig. 5, blue), the *SIGLEC22P* pseudogene was used as a repair template because of the strong sequence homology to *CD33* (Fig. 5, underlined). In this process, three *SIGLEC22P*-specific mutations were introduced into *CD33*, resulting in missense mutations in three adjacent codons and thus *KIR-CD33*. This indicates an in-frame ectopic gene conversion using pseudogene sequence. Two lines of evidence indicate that the *KIR-CD33* cell lines were homozygous for the *KIR*

FIG. 5. Alignment of cell line sequencing data. Unedited U937 cells at *top*, followed by KIR-CD33 sequence, D2-CD33 sequence, and reference *SIGLEC22P* sequence. Cas9 cleavage site marked by scissor icons. Mismatches from the unedited *CD33* sequence are denoted by either the differing base (KIR-CD33) or dash in the case of a gap (D2-CD33). Intron 1 in *black*, exon 2 in *blue*. The *SIGLEC22P* region used as a repair template is *underlined*. Mutations introduced into *CD33* in *red*. A representative post-PCR sequencing chromatogram from one clonal KIR-CD33 is shown at the mutation site with clear single peaks demonstrating homozygosity. Color images are available online.

mutation. First, the DNA sequence chromatogram showed only clear single peaks through the KIR sequence (Fig. 5, chromatogram). Second, cell populations with an intermediate level of P67.6⁺ labeling were not detected (Fig. 1C). Further, sequencing of the *SIGLEC22P* region in the KIR-CD33 clones revealed that the *CD33* sequence was not present, which we interpret to mean that *SIGLEC22P* was used as a repair template rather than a crossover event.

Discussion

We show here that the DSB repair pathways initiated after *S. pyogenes* Cas9 cleavage can lead to ectopic gene conversion from a pseudogene in a mitotic human cell line. This gene conversion resulted in an in-frame chimeric protein, wherein <150 bp of sequence from a nearby pseudogene replaced the targeted gene sequence (Fig. 5). Indeed, this distance is consistent with meiotic gene conversion events observed by Jeffreys *et al.*, who showed that gene conversion occurs through relatively short tracts, with a mean length between 55 and 290 bp.³⁵ Given that the *SIGLEC22P* locus in these KIR-CD33 cells lacks any detectable *CD33* sequence, we interpret this as further evidence of a gene conversion rather than a mitotic crossover event. We speculate that this gene conversion occurs *in trans*, that is, from the intact chromosome, as a *cis*-mediated repair would more likely result in a gene fusion with intergenic deletion as has been previously reported.^{36,37} This type of deletion event also occurs naturally, in the absence of Cas9 DSBs, as in

the case of *SIGLEC14* deletions.³⁸ Gene conversion *in trans* after CRISPR-induced DSBs has been demonstrated previously.⁴ We were surprised by the unexpectedly high frequency of conversion observed here; this pseudogene conversion occurred at approximately 1 in 10 edited cells. Exogenous double-stranded regions of homology as short as 58 bp have been used *in vitro* to introduce mutations through CRISPR-Cas9/HDR mechanisms.² Coincidentally, this ectopic gene conversion disrupted the epitope of a well-validated antibody, P67.6 (Fig. 6), known clinically as gemtuzumab. Using transiently transfected HEK293 cells expressing the chimeric protein, we demonstrated that these mutations are sufficient to abrogate P67.6 binding, providing the most precise epitope mapping to date of this clinically relevant antibody.

The *KIR-CD33* mutations—rs3987761, rs3987760, and rs35814802—are indexed in dbSNP and gnomAD with minor allele frequencies <10⁻⁴ and are roughly equivalent across populations.³⁰ We considered the possibility that these mutations, while rare, occur naturally and are clinically relevant. The *SIGLEC* family of genes are undergoing rapid evolution in many species, including humans.³⁹ This rapid evolution has resulted in the pseudogenization of many *SIGLECs*, including the *CD33* pseudogene *SIGLEC22P*. These same bases are indexed as SNPs in *SIGLEC22P* (rs997169007, rs1049597792, and rs1005338799), also have identical frequencies (MAF = 2.1 × 10⁻⁴) in the gnomAD database, and the

FIG. 6. Model of pseudogene repair mechanism and anti-CD33 antibody binding sites. CD33 is normally a transmembrane, cell surface receptor with one IgV domain and one IgC₂ domain. The KIR-CD33 mutation introduced by pseudogene directed-repair abrogates P67.6, but not WM53, binding. IgC₂ domain antibodies HIM3-4 and PWS44, and the intracellular domain H110 antibody bind both CD33 and KIR-CD33. D2-CD33 is not readily apparent on the cell surface in CRISPR-Cas9 edited U937 cells, implying that under physiologic expression, the D2-CD33 protein is retained in an intracellular vesicle. Overall, pseudogene repair of a CRISPR-Cas9-targeted gene can disrupt the binding of well-validated antibodies without introducing a frameshift, and protein-level expression alone is not sufficient for knockout confirmation. Created with Biorender. Color images are available online.

major and minor alleles are the inverse of the *KIR-CD33* SNPs.³⁰ The region of homology identified here between *CD33* and *SIGLEC22P* is 143 bp—substantially shorter than the 250 bp reads upon which the gnomAD and dbSNP databases are built.⁴⁰ While it is possible these are *bona fide* variants in *CD33*, the most likely explanation is that the *SIGLEC22* pseudogene sequences have been mapped incorrectly to *CD33* by algorithms. These mis-

sense SNPs are also not recorded in the BeatAML variant database, which records known AML-associated functional variants such as missense SNPs, further underscoring the low probability that these are true *CD33* variants.⁴¹

Off-target editing is a frequent concern in gene editing workflows, including CRISPR-Cas9. However, gene conversion is often overlooked as a potential confound. For instance, within *CD33*, off-target editing of *SIGLEC22P*,

resulting in a 14 kb deletion, and subsequent *SIGLEC22P/CD33* gene fusion has been previously reported, which may have been the result of a mitotic crossover.^{36,37} Neither report noted a gene conversion between *SIGLEC22P* and *CD33*. Mitotic crossover initiated by CRISPR-Cas9 cleavage has also been reported in the *HPRT* locus, resulting in a 36 kb crossover.⁴ Gene conversion via CRISPR-Cas9 between a 101 bp homologous region in *HBD* and *HBB* has also been reported. Notably, these genes were also the first reported gene conversion event in humans.^{3,42} To our knowledge, this is the first report of pseudogene-mediated gene conversion during the CRISPR-Cas9 editing process. Our results demonstrate the need for rigorous screening in studies which rely on gene editing, and analysis of the region flanking the expected cut site for homology and possible gene conversion. Since the human genome contains 8,000–12,000 pseudogenes and approximately 3,400 genes in the human genome have known pseudogenes, pseudogene-directed homology repair is a potentially considerable confound.^{43,44} Approximately 84% of putative pseudogenes are estimated to be located on a different chromosome than their parent gene, while the remaining 16% have a mean intergenic distance of 1.8 Mb (median 1.3 Mb).⁴⁵ Whether the gene conversion described here occurs in *cis* or *trans*, or if this event is impacted by intergenic distance, is unclear. As pseudogenes are by definition nonfunctional, they may be overlooked as irrelevant during the design process. Given the pseudogene conversion described here, there is a clear need to analyze the sequence of the edit, rather than relying on gene or protein expression. This is especially relevant, given that CRISPR-mediated *CD33* knockout allografts and autografts have been proposed as a potential AML treatment strategy in conjunction with chimeric antigen receptor (CAR) T cells.^{36,37} In this strategy, autologous transplantation of *CD33*^{null} hematopoietic stem and progenitor cells (HSPCs) would reconstitute the myeloid system, while engraftment of CAR-T cells would provide long-term surveillance against any surviving *CD33*⁺ AML cells. Mitotic gene conversion, and presumably the herein described pseudogene-mediated gene conversion, has been demonstrated after CRISPR-Cas9 editing in primary human cells between *HBD* and *HBB*.³ Screening these HSPCs with the standard gemtuzumab antibody, however, could lead to engraftment of some KIR-*CD33* cells as well.

While off-target editing of a nonfunctional pseudogene may have limited impact on downstream results, incorporating elements from a nonfunctional pseudogene into a target gene may lead to deleterious mutations that abrogate function of the target gene entirely, for instance the introduction of a premature stop codon. This is especially important to consider for designs that incorporate

unedited cells that have undergone sorting and single-cell cloning as a control. We also speculate that this pseudogene-mediated repair will reduce the efficiency of creating a gene disruption if the disruption site has homology in a pseudogene. The breakage site may be repaired by a pseudogene with complete identity at the DSB, requiring more clones to be screened to find a gene disruption. By coincidence, our initial screen for editing included an antibody that overlapped the KIR sequence. This gene conversion, at the protein level, is masked when using an alternative antibody (WM53) targeting the same domain and is not apparent by PCR alone. Combining the data from Figures 2D and F and 3C, one could reason that these cells were unedited, as there are no apparent differences in PCR fragment size, gene expression, splicing, or cell surface protein expression, and thus incorrectly assume that sequencing is unnecessary. We conclude that in addition to off-target Cas9-editing confounds, researchers should be aware of the potential for pseudogene-directed homology repair.

Acknowledgments

We would like to thank Yuriko Katsumata for bioinformatics support, Ann M. Stowe for flow cytometry support, the University of Kentucky Light Microscopy Core Facility for help with image acquisition and processing, and the University of Kentucky Flow Cytometry and Immune Monitoring Core Facility, supported in part by the Office of the Vice President for Research, the Markey Cancer Center and an NCI Center Core Support Grant (P30CA177558).

Author Disclosure Statement

The authors declare no conflicts of interest.

Funding Information

This work was supported by F99NS120365 to B.C.S. from the National Institute of Neurological Disorders and Stroke and by RF1AG059717-01S1 and R21AG068370 to S.E. from the National Institute on Aging.

References

1. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012;337:816–821. DOI: 10.1126/science.1225829.
2. Renaud J-B, Boix C, Charpentier M, et al. Improved genome editing efficiency and flexibility using modified oligonucleotides with TALEN and CRISPR-Cas9 nucleases. *Cell Rep* 2016;14:2263–2272. DOI: 10.1016/j.celrep.2016.02.018.
3. Javidi-Parsijani P, Lyu P, Makani V, et al. CRISPR/Cas9 increases mitotic gene conversion in human cells. *Gene Ther* 2020;27:281–296. DOI: 10.1038/s41434-020-0126-z.
4. Susani L, Castelli A, Lizier M, et al. Correction of a recessive genetic defect by CRISPR-Cas9-mediated endogenous repair. *CRISPR J* 2018;1:230–238. DOI: 10.1089/crispr.2018.0004.

5. Hollingworth P, Harold D, Sims R, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet* 2011;43:429–435. DOI: 10.1038/ng.803.
6. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet* 2019;51:404–413. DOI: 10.1038/s41588-018-0311-9.
7. Naj AC, Jun G, Beecham GW, et al. Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat Genet* 2011;43:436–441. DOI: 10.1038/ng.801.
8. Malik M, Chiles IIIJ, Xi HS, et al. Genetics of CD33 in Alzheimer's disease and acute myeloid leukemia. *Hum Mol Genet* 2015;24:3557–3570. DOI: 10.1093/hmg/ddv092.
9. Raj T, Ryan KJ, Replogle JM, et al. CD33: increased inclusion of exon 2 implicates the Ig V-set domain in Alzheimer's disease susceptibility. *Hum Mol Genet* 2014;23:2729–2736. DOI: 10.1093/hmg/ddt666.
10. Schwarz F, Springer SA, Altheide TK, et al. Human-specific derived alleles of CD33 and other genes protect against postreproductive cognitive decline. *Proc Natl Acad Sci U S A* 2016;113:74–79. DOI: 10.1073/pnas.1517951112.
11. Gričic A, Serrano-Pozo A, Parrado Antonio R, et al. Alzheimer's disease risk gene CD33 inhibits microglial uptake of amyloid beta. *Neuron* 2013;78:631–643. DOI: 10.1016/j.neuron.2013.04.014.
12. Bradshaw EM, Chibnik LB, Keenan BT, et al. CD33 Alzheimer's disease locus: altered monocyte function and amyloid biology. *Nat Neurosci* 2013;16:848–850. DOI: 10.1038/nn.3435.
13. Estus S, Shaw BC, Devanney N, et al. Evaluation of CD33 as a genetic risk factor for Alzheimer's disease. *Acta Neuropathol* 2019;138:187–199. DOI: 10.1007/s00401-019-02000-4.
14. Malik M, Simpson JF, Parikh I, et al. CD33 Alzheimer's risk-altering polymorphism, CD33 expression, and exon 2 splicing. *J Neurosci* 2013;33:13320–13325. DOI: 10.1523/JNEUROSCI.1224-13.2013.
15. Balaian L, Zhong RK, Ball ED. The inhibitory effect of anti-CD33 monoclonal antibodies on AML cell growth correlates with Syk and/or ZAP-70 expression. *Exp Hematol* 2003;31:363–371. DOI: 10.1016/s0301-472x(03)00044-4.
16. Hernández-Caselles T, Martínez-Esparza M, Pérez-Oliva AB, et al. A study of CD33 (SIGLEC-3) antigen expression and function on activated human T and NK cells: two isoforms of CD33 are generated by alternative splicing. *J Leukoc Biol* 2006;79:46–58. DOI: 10.1189/jlb.0205096.
17. Pérez-Oliva AB, Martínez-Esparza M, Vicente-Fernández JJ, et al. Epitope mapping, expression and post-translational modifications of two isoforms of CD33 (CD33M and CD33m) on lymphoid and myeloid human cells. *Glycobiology* 2011;21:757–770. DOI: 10.1093/glycob/cwq220.
18. Walter RB, Raden BW, Zeng R, et al. ITIM-dependent endocytosis of CD33-related Siglecs: role of intracellular domain, tyrosine phosphorylation, and the tyrosine phosphatases, Shp1 and Shp2. *J Leukoc Biol* 2008;83:200–211. DOI: 10.1189/jlb.0607388.
19. Paul SP, Taylor LS, Stansbury EK, et al. Myeloid specific human CD33 is an inhibitory receptor with differential ITIM function in recruiting the phosphatases SHP-1 and SHP-2. *Blood* 2000;96:483–490.
20. Bhattacherjee A, Rodrigues E, Jung J, et al. Repression of phagocytosis by human CD33 is not conserved with mouse CD33. *Commun Biol* 2019;2:450. DOI: 10.1038/s42003-019-0698-6.
21. Bhattacherjee A, Jung SJ, Ho M, et al. The CD33 short isoform is a gain-of-function variant that enhances Aβ1-42 phagocytosis in microglia. *Mol Neurodegener* 2021;16:19. DOI: 10.1186/s13024-021-00443-6.
22. Wissfeld J, Nozaki I, Mathews M, et al. Deletion of Alzheimer's disease-associated CD33 results in an inflammatory human microglia phenotype. *Glia* 2021;99:1393–1412. DOI: 10.1002/glia.23968.
23. Siddiqui SS, Springer SA, Verhagen A, et al. The Alzheimer's disease-protective CD33 splice variant mediates adaptive loss of function via diversion to an intracellular pool. *J Biol Chem* 2017;292:15312–15320. DOI: 10.1074/jbc.M117.799346.
24. Godwin CD, Laszlo GS, Wood BL, et al. The CD33 splice isoform lacking exon 2 as therapeutic target in human acute myeloid leukemia. *Leukemia* 2020;34:2479–2483. DOI: 10.1038/s41375-020-0755-7.
25. Humbert O, Laszlo GS, Sichel S, et al. Engineering resistance to CD33-targeted immunotherapy in normal hematopoiesis by CRISPR/Cas9-deletion of CD33 exon 2. *Leukemia* 2019;33:762–808. DOI: 10.1038/s41375-018-0277-8.
26. Lee J-Y, Lee C-H, Shim S-H, et al. Molecular cytogenetic analysis of the monoblastic cell line U937. *Cancer Genet Cytogenetics*. 2002;137:124–132. DOI: 10.1016/s0165-4608(02)00565-4.
27. Shipley JM, Sheppard DM, Sheer D. Karyotypic analysis of the human monoblastic cell line U937. *Cancer Genet Cytogenet* 1988;30:277–284. DOI: 10.1016/0165-4608(88)90195-1.
28. Grear KE, Ling IF, Simpson JF, et al. Expression of SORL1 and a novel SORL1 splice variant in normal and Alzheimer's disease brain. *Mol Neurodegener* 2009;4:46. DOI: 10.1186/1750-1326-4-46.
29. Gbadamosi MO, Shastri VM, Hylkema T, et al. Novel CD33 antibodies unravel localization, biology and therapeutic implications of CD33 isoforms. *Future Oncol* 2021;17:263–277. DOI: 10.2217/fo-2020-0746.
30. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–443. DOI: 10.1038/s41586-020-2308-7.
31. Nakai K, Horton P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* 1999;24:34–36. DOI: 10.1016/s0968-0004(98)01336-x.
32. Rice P, Longden I, Bleasby A. EMBOS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276–277. DOI: 10.1016/s0168-9525(00)02024-2.
33. Mortland L, Alonzo TA, Walter RB, et al. Clinical significance of CD33 nonsynonymous single-nucleotide polymorphisms in pediatric patients with acute myeloid leukemia treated with gemtuzumab-ozogamicin-containing chemotherapy. *Clin Cancer Res* 2013;19:1620–1627. DOI: 10.1158/1078-0432.CCR-12-3115.
34. Chauhan L, Shin M, Wang YC, et al. CD33_PGX6_Score predicts gemtuzumab ozogamicin response in childhood acute myeloid leukemia: a report from the children's oncology group. *JCO Precis Oncol* 2019;3:PO.18.00387. DOI: 10.1200/po.18.00387.
35. Jeffreys AJ, May CA. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* 2004;36:151–156. DOI: 10.1038/ng1287.
36. Borot F, Wang H, Ma Y, et al. Gene-edited stem cells enable CD33-directed immune therapy for myeloid malignancies. *Proc Natl Acad Sci U S A* 2019;116:11978–11987. DOI: 10.1073/pnas.1819992116.
37. Kim MY, Yu K-R, Kenderian SS, et al. Genetic inactivation of CD33 in hematopoietic stem cells to enable CAR T cell immunotherapy for acute myeloid leukemia. *Cell* 2018;173:1439–1453.e1419. DOI: 10.1016/j.cell.2018.05.013.
38. Yamanaka M, Kato Y, Angata T, et al. Deletion polymorphism of SIGLEC14 and its functional implications. *Glycobiology* 2009;19:841–846. DOI: 10.1093/glycob/cwp052.
39. Padler-Karavani V, Hurtado-Ziola N, Chang YC, et al. Rapid evolution of binding specificities and expression patterns of inhibitory CD33-related Siglecs in primates. *FASEB J* 2014;28:1280–1293. DOI: 10.1096/fj.13-241497.
40. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 2013;45:580–585. DOI: 10.1038/ng.2653.
41. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* 2018;562:526–531. DOI: 10.1038/s41586-018-0623-z.
42. Slightom JL, Blechl AE, Smithies O. Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. *Cell* 1980;21:627–638. DOI: 10.1016/0092-8674(80)90426-2.
43. Zhang Z, Carriero N, Zheng D, et al. PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* 2006;22:1437–1439. DOI: 10.1093/bioinformatics/btl116.
44. Pei B, Sisu C, Frankish A, et al. The GENCODE pseudogene resource. *Genome Biol* 2012;13:R51. DOI: 10.1186/gb-2012-13-9-r51.
45. Sisu C, Pei B, Leng J, et al. Comparative analysis of pseudogenes across three phyla. *Proc Natl Acad Sci U S A* 2014;111:13361–13366. DOI: 10.1073/pnas.1407293111.