Quantitative evaluation of chromosomal rearrangements in primary gene-edited human stem cells by preclinical CAST-Seq

Giandomenico Turchiano[1,2,#,*], Geoffroy Andrieux[3,4], Georges Blattner[1,2,#], Valentina Pennucci[1,2], Julia Klermund[1,2], Gianni Monaco[1,2], Sushmita Poddar[1,2,$], Claudio Mussolino[1,2], Tatjana I. Cornu[1,2], Melanie Boerries[3,4,5,6], Toni Cathomen[1,2,6,*]

[1] Institute for Transfusion Medicine and Gene Therapy, Medical Center – University of Freiburg, Freiburg, Germany.

[2] Center for Chronic Immunodeficiency (CCI), Medical Center – University of Freiburg, Freiburg, Germany.

[3] Institute of Medical Bioinformatics and Systems Medicine, University of Freiburg, Freiburg, Germany

[4] German Cancer Consortium (DKTK) and German Cancer Research Center (DKFZ), Partner site Freiburg, Freiburg, Germany

[5] Comprehensive Cancer Center Freiburg (CCCF), Medical Center – University of Freiburg, Freiburg, Germany

[6] Faculty of Medicine, University of Freiburg, Freiburg, Germany

[#]current address: Infection, Immunity and Inflammation Program, Great Ormond Street Institute of Child Health, University College London, London, UK

[$]current address: Department of Biosystems Science and Engineering (D-BSSE), ETH Zürich, Basel, Switzerland

* Correspondence:
Toni Cathomen, toni.cathomen@uniklinik-freiburg.de (lead contact)
Giandomenico Turchiano, g.turchiano@ucl.ac.uk

Abstract

Genome editing with programmable nucleases has shown great promise for clinical translation but also revealed the risk of genotoxicity caused by chromosomal translocations or the insertion of mutations at off-target sites. Here, we describe CAST-Seq, an innovative assay to identify and quantify chromosomal aberrations derived from on- and off-target activities of CRISPR-Cas nucleases or TALENs. CAST-Seq also detected novel types of chromosomal rearrangements, including homology-mediated translocations that are mediated by homologous recombination. Depending on the employed designer nuclease, translocations occurred in 0–0.5% of gene-edited human stem cells and some 20% of target loci harbored gross aberrations. In conclusion, CAST-Seq analyses are particularly relevant for therapeutic editing of stem cells to enable a thorough risk assessment before clinical application of gene editing products.

Keywords

chromosomal aberrations; chromosomal rearrangements; clinical risk assessment; CRISPR-Cas; designer nucleases; gene editing; off-target activity; off-target effects; programmable nucleases; translocations

Introduction

Targeted genome editing has been successfully employed to genetically modify various human cell types or organs for therapeutic applications (Carroll, 2014). Zinc-finger nucleases (ZFNs) (Alwin et al., 2005; Porteus and Baltimore, 2003; Smith et al., 2000; Urnov et al., 2005), transcriptional activator-like effector nucleases (TALENs) (Cermak et al., 2011; Miller et al., 2011; Mussolino et al., 2011), and CRISPR-Cas9 nucleases (Gasiunas et al., 2012; Jinek et al., 2012; Cho et al., 2013; Cong et al., 2013; Mali et al., 2013) have been used in clinical trials to target inborn disorders, infectious diseases or certain types of cancer (Bailey and Maus, 2019; Cornu et al., 2017). Before applying genome editing in transplantable cells or *in vivo* however, engineered nucleases need to be carefully evaluated with respect to activity and specificity. Both parameters are fundamental for safe translation in order to maintain genome integrity of edited cells and to reduce the risk of oncogenic transformation, often referred to as genotoxicity (Kim et al., 2019; Tsai and Joung, 2016).

Designer nuclease induced off-target (OT) activity can lead to short insertion/deletion (indel) mutations, large chromosomal deletions and inversions, as well as chromosomal translocations. Much effort has been invested in increasing the safety of genome editing tools in the past years, leading to better designer nuclease platforms with improved specificities (Kim et al., 2019; Tsai and Joung, 2016). Nonetheless, a thorough preclinical assessment of their specificity is a clearly stated requirement by the regulatory authorities (Cathomen et al., 2019). Ideally, the applied tests combine high sensitivity with high specificity and allow scientists not only to measure designer nuclease-induced mutagenesis but also chromosomal aberrations (Kim et al., 2019; Tsai and Joung, 2016). Computer-based prediction algorithms (*in silico* methods), *in vitro* tests and cell-based assays rely on next-generation sequencing (NGS) and are typically employed in a two-step process: a predictive ‘screening test’ is used first to identify potential OT sites, followed by a ‘confirmatory test’ to validate the predictions in the gene-edited, clinically relevant cell type.

In silico prediction algorithms are based on well-defined parameters, including similarity to the target sequence (Kim et al., 2019; Tsai and Joung, 2016). They represent a fast and cheap screening method, but often miss critical OT sites. In contrast, experimental methods, such as BLISS (Yan et al., 2017), CIRCLE-Seq (Tsai et al., 2017), Digenome-Seq (Kim et al., 2015), DISCOVER-Seq (Wienert et al., 2019), GUIDE-Seq (Tsai et al., 2015), and SITE-Seq (Cameron et al., 2017), enable a more or less unbiased identification of OTs but are more laborious and sometimes lack specificity and/or sensitivity (Kim et al., 2019). *In vitro* methods tend to be highly sensitive in detecting genomic sites that are cleaved *in vitro*, but a large proportion of those *in vitro* OTs cannot be confirmed in gene-edited patient cells (Kim et al., 2019), suggesting that *in vitro* assays overestimate the number of relevant OTs in a cell. Cell-based assays, such as BLISS, DISCOVER-Seq and GUIDE-Seq, detect OTs by tagging the DNA double-strand breaks (DSB) created by the engineered nuclease by different means. They seem specific and sensitive but do not identify gross chromosomal aberrations induced by programmable nucleases (Kosicki et al., 2018). To this end HTGTS (Frock et al., 2015) and UDiTaS (Giannoukos et al., 2018) were developed. These methods are based on linker-mediated (LM) or linear amplification-mediated (LAM)-PCR (Mueller and Wold, 1989; Schmidt et al., 2007) and identify translocation events to the known on-target site. However, these methods cannot be applied in a clinical context as they are not quantitative, their sensitivity is unknown, and they can only identify particular types of genomic aberrations.

To overcome these limitations, we established CAST (chromosomal aberrations analysis by single targeted LM-PCR)-Seq. CAST-Seq exploits locus-specific decoy primers that decreased the background noise and allowed us to identify with high sensitivity both OT-mediated translocations (OMTs) as well as chromosomal aberrations mediated by on-target activity of designer nucleases. The latter group includes large deletions and inversions at the target site as well as chromosomal rearrangements that we termed homology-mediated translocations (HMTs). HMTs represent a novel class of chromosomal aberrations mediated by a homologous recombination (HR)-based mechanism. Moreover, CAST-Seq is a quantitative assay that is performed directly in the clinically relevant cell type, so rendering NGS-based 'confirmatory tests' redundant. Depending on the designer nuclease and the target site, chromosomal rearrangements occurred in up to 1.6% of edited stem cells and up to 20% of on-target loci harbored large deletions or other chromosomal aberrations.

Results

Experimental setup and bioinformatics

Various chromosomal rearrangements can be induced by nuclease activity at the on-target and/or OT sites, respectively (**Fig. 1a**). Notably, cleavage at on-target and OT sites induce large (>250 bp) deletions/inversions at the cleavage sites as well as OT-mediated translocation (OMT) with balanced or unbalanced outcome. Moreover, we postulated that on-target cleavage will induce HR-based HMT events between the on-target site and a locus that shares a certain degree of homology with the on-target region. CAST-Seq was developed to identify and quantify these chromosomal aberrations with high sensitivity by mapping the chromosomal sequences fused to one arm of the target site (**Fig. 1b**). First, the genomic DNA was fragmented to an average length of 350 bp, followed by linker ligation, and three PCR steps: the first PCR reaction is performed with a 'bait primer' binding to the on-target sequence, a 'prey primer' that recognizes the linker sequence, and 'decoy primers' that bind the target sequence to prevent on-target amplification (**Fig. 1b**, **Fig. S1**, **Table S1**). The second and third PCR steps introduce adaptors and barcode sequences for NGS (**Fig. 1b**).

All data reported in this study were produced in primary CD34-positive human hematopoietic stem and progenitor cells (HSPCs). CRISPR-Cas nucleases, comprising both wild-type and a high-fidelity (HiFi) variant of Cas9 (Vakulskas et al., 2018), were targeted to either the *CCR5* locus (sites #1 and #2) or two previously described sites in *FANCF* and *VEGFA* (Tsai et al., 2017; Tsai et al., 2015). Moreover, a TALEN targeting the *HBB* locus was included (Patsali et al., 2019). The cells were pre-activated for two days before transferring CRISPR-Cas nucleases as ribonucleoprotein (RNP) complexes or TALENs in the form of mRNA (**Fig. 1c**). Genomic DNA was collected at different time points to evaluate the kinetics and dynamics of DNA repair and chromosomal alterations. On-target activity, as measured by indel formation, was in the range of 48–89% with cell viabilities ranging from 70–99% (**Fig. 1d-e, Fig. S2**).

A novel bioinformatic pipeline was developed to map the chromosomal regions that were fused to the target site as a result of chromosomal rearrangements (**Fig. S3a**, Methods). Because the DNA ends of a nuclease-induced DSB can be processed differently before translocation (Chiarle et al., 2011; Roukos et al., 2013), the distinctive translocation fusion points together with the linker ligation points were used as 'unique molecular signature' to compute the number of individual translocation events, called 'hits'. An *in silico* random library of 10,000 random reads was created to define the statistical likelihood of two or more hits to fall within the same 'cluster'. A cutoff of 2,500 bp was identified as a conservative threshold to define a cluster that contains events mediated by the same trigger (p<0.05, **Fig. S3b**). To classify these events as OMT or HMT, the region surrounding the fusion point was screened either for the presence of sequences that share homology to the 23-bp target site (OMT, **Fig. S3c**) or that contain a stretch of at least 25 bp of homology to the on-target region within a 5 kb window (HMT, **Fig. S3d**). A 25 bp homology stretch was reported to be sufficient to mediate HR in mammalian cells (Ayares et al., 1986). For some events both criteria were met (ambiguous OMT/HMT annotation). If neither an OT nor a homology stretch could be identified, the chromosomal aberration was considered to be prompted by a natural break site (NBS). This classification was verified by target deep amplicon sequencing on some of the sites (**Table S2**).

Chromosomal aberrations induced by a CCR5 targeting CRISPR-Cas9 nuclease

Genomic DNA extracted from CD34+ cells nucleofected with an RNP targeting *CCR5* target site #1 ($CCR5^{\#1}$), was subjected to CAST-Seq that included two decoy primers (**Fig. 2a**). A qualitative and quantitative examination revealed that most retrieved CAST-Seq hits included chromosomal aberrations within $CCR5^{\#1}$, i.e. large deletions or inversions, as well as acentric and dicentric translocations with the homologous chromosome within a ~15-kb region surrounding the on-target site (**Fig. 2b-c, Fig. S4**). The second most frequent class of events involved large ~15-kb deletions between *CCR5* and an ambiguous OMT/HMT site in the *CCR2* locus. Targeted deep sequencing revealed that the number of indels at the *CCR2* site did not differ significantly from untreated controls (**Fig. 2d**), implicating that the observed chromosomal aberrations were induced by HMT and not OMT. Further sequencing confirmed indels at $CCR5^{\#1}$ (>80%) as well as at the identified five OMT sites and one ambiguous OMT/HMT site (0.2–11%), confirming that those events represent OMTs (**Fig. 2c-d**). For three of the six sites, indel frequencies were significantly ($p<0.01$) lower when HiFi Cas9 was used (**Fig. 2d**). CAST-Seq also exposed rare, rather complex repair outcomes, in which small portions of target region sequences were duplicated and/or inverted, or in which partial sequence insertions from homologous regions (i.e. *CCR2*) were observed (**Fig. 2e**). A quantitative Circos plot summarizes the CAST-Seq results and indicates classification and locations of the chromosomal rearrangements (**Fig. 2f**). The relation between the number of CAST-Seq hits (**Fig. 2c, Table S3**) and the number of absolute translocation events was determined by droplet digital PCR (ddPCR). A *STAT3* specific amplicon was used to normalize the values. The squared correlation coefficient of 0.96 confirmed a linear correlation (**Fig. 2g**) and allowed us to infer that one CAST-Seq hit corresponds to about 10 factual chromosomal aberrations events (as determined by ddPCR). Moreover, knowing the number of input genomes (~150,000 haploid genomes in 500 ng of DNA) and the relation between CAST-Seq hits and actual chromosomal rearrangements, we calculated the limit of detection of CAST-Seq to be in the range of 1 chromosomal aberration per 7,500 cells. Moreover, we were able to calculate the absolute number of rearrangements per cell, which amounted to 1.6% of affected alleles for the *CCR5/CCR2* HMT and 0.01–0.5% of alleles for the four OMTs shown in **Fig. 2g**.

Chromosomal aberrations induced by other engineered nucleases

In order to substantiate the HMT phenomena, we included a TALEN pair targeting *HBB* in proximity to the related *HBD* locus (Patsali et al., 2019), and designed a second gRNA targeting *CCR5* ($CCR5^{\#2}$) in a region that differs more substantially from the *CCR2* paralogue sequence. CAST-Seq analysis was further performed on CD34+ cell samples edited with CRISPR-Cas9 nucleases targeting sites in *FANCF* and *VEGFA* (Tsai et al., 2017; Tsai et al., 2015). The TALEN pair targeting *HBB* induced aberrations at the on-target site, including ~6-kb deletions between *HBB* and the closely related *HBD* locus (**Fig. 3a**, **Table S2**). For the $CCR5^{\#2}$ nuclease, we observed as predicted a notably reduced amount of HMT hits compared to the $CCR5^{\#1}$ targeting CRISPR nuclease (**Fig. 3b, Fig. 2f**, **Table S2**). To corroborate that HMT is mediated by the HR machinery, K562 cells were nucleofected with $CCR5^{\#1}$-targeting RNPs or with *HBB* TALEN-encoding mRNA in the presence or absence of B02, a well-characterized inhibitor of RAD51 (Huang et al., 2011; Ward et al., 2015). While HMT events between *CCR5-CCR2* and *HBB-HBD* were significantly reduced, the OMT between *CCR5* and Chr.13 was not affected by the drug (**Fig. S5**), clearly connecting the HR machinery to HMT.

CD34+ cells edited with CRISPR nucleases targeting *VEGFA* and *FANCF* confirmed that the *FANCF* nuclease was highly specific, only revealing chromosomal anomalies at the on-target site but no further translocations (**Fig. 3c**). On the other hand, the nuclease targeting *VEGFA* returned multiple OMT and ambiguous HMT/OMT events (**Fig. 3d**). These results were in good agreement with OT analyses returned by GUIDE-Seq (Tsai et al., 2015) and CIRCLE-Seq (Tsai et al., 2017) (**Fig. S6**) as well as indel detection by target amplicon sequencing (**Table S2**). Further NGS analysis confirmed that CAST-Seq classified correctly six out of seven aberrations in *VEGFA*-edited cells while no indels were found at the two OMT events in *FANCF*-edited HSPCs (**Table S3**). These two rare aberrations are in proximity to the on-target site and thus likely represent large deletions. On the other hand, of the top ranked OTs identified by GUIDE-Seq and CIRCLE-Seq, only 1 (0.29% indels; *FANCF* at 74%) out of 8 OTs was cleaved in *FANCF*-edited HSPCs and 5 (0.27–6.6% indels; *VEGFA* at 67%) of 13 OTs in *VEGFA*-edited HSPCs (**Table S3**), suggesting that a majority of GUIDE-Seq and CIRCLE-Seq predicted OTs were not cleaved in CD34+ cells. Furthermore, two verified OTs did not induce translocations, as confirmed by direct PCR (data not shown), and were therefore not detected by CAST-Seq. In conclusion, CAST-Seq is able to identify qualitatively as well as quantitatively CRISPR-Cas and TALEN induced chromosomal rearrangements that could be reliably classified in OMT, HMT, and NBS-induced aberrations.

Qualitative and quantitative changes in chromosomal rearrangements over time

To assess the stability of the chromosomal aberrations, CAST-Seq was performed at different time points. Zooming in on a ~33 kb range around the target site visualized the full extent of the large deletions around the on-target sites and enabled us to evaluate the dynamic changes in the edited cell population (**Fig. 4a-c**). Since CAST-Seq has a dictated sequencing orientation, it is possible to nominate the orientation of the chromosomal rearrangements and hence to unveil large inversions, deletions and insertions, as well as chromosomal translocations between non-homologous and homologous chromosomes, including the formation of acentric or dicentric chromosomes (see also **Fig. 1a, Fig. S4**). For all of our samples a gradual qualitative and quantitative loss of CAST-Seq reads over the time course of two weeks was observed, which was most pronounced for the *CCR5*$^{\#1}$ sample (**Fig. 4a-c, Fig. S7**). This decrease, which was also observed at OMTs, likely reflecting the loss of cells with unbalanced translocations and suggesting that some rearrangements are subjected to negative selective pressure. On the other hand, the quantitative loss of CAST-Seq hits was rather modest **(Fig. 4d-f)**. Of note, genome editing with HiFi Cas9 abolished some OMTs but had – as expected – no impact on HMTs (**Fig. 4d-e**, **Table S2**). Expression of the *HBB* targeting TALENs induced chromosomal rearrangements at the on-target site, including the nominated deletion events spanning up to 15 kb and the 6 kb deletions between *HBB* and *HBD* (**Fig. 4c, f**). Targeted amplicon sequencing confirmed minimal OT activity at the *HBD* locus (**Table S2**), suggesting that these aberrations were synergistically triggered by OMT and HMT (see also **Fig. S5**). Collectively, these experiments confirm qualitatively and quantitatively various chromosomal rearrangements around the on-target site. Based on the number of CAST-Seq hits and the ddPCR analysis (**Fig. 2g**), this number can mount to up to 1 chromosomal rearrangement per 100 cells if a related gene is in proximity to the target site.

Kinetics of DNA repair

A ddPCR-based assay was established to follow the repair kinetics at the on-target site (**Fig. 5a**). Primers flanking the cleavage sites (*CCR5*$^{\#1}$, *CCR5*$^{\#2}$, *HBB*) were designed to detect copy number variations (CNVs) based on loss of primer binding sites in the case of non-homologous translocations or large deletions. EvaGreen – rather than specific probes – was chosen to quantify the number of alleles even when the loci were altered by nuclease-induced indel mutations. Two ddPCR amplicons placed on either side of the central 'edge amplicon' were designed to distinguish between non-homologous translocations and large deletions, and further amplicons that amplified either distal regions on the target chromosome or on two other chromosomes (*RAD1*, *STAT3*) were used to normalize the values (**Fig. 5a**).

One day after transfer of the engineered nucleases, a large fraction (30-45%) of the CRISPR-Cas or TALEN induced chromosomal breaks were either not rejoined due to continuous designer nuclease activity or subject to large deletions or translocations (**Fig. 5b-d**, **Fig. S8**). The CNVs plateaued after day 4, suggesting completed DSB repair and/or a selective loss of cells with chromosomal aberrations that affect viability or proliferation. At these later time points, the difference in CNV between 'edge amplicons' (red) and 'flanking amplicons' (black, grey) specifies the nature of the chromosomal aberration: large deletion (decrease in 'flanking amplicons') or chromosomal translocation (decrease in 'edge amplicons'). This distinction is evident when comparing the *CCR5*$^{\#1}$ targeting nuclease to the more specific designer nucleases targeting *CCR5*$^{\#2}$ at day 14 (**Fig. 5b-d**). The distal ddPCR amplicons did not show considerable CNVs, confirming no gross chromosomal loss of information. These ddPCR data were then used to normalize the T7E1 assays (**Fig. 1d**) or targeted amplicon sequencing results (**Table S2**), which cannot detect gross chromosomal rearrangements (**Fig. 5e-g**). In conclusion, about 12-22% of the target alleles in these gene-edited CD34+ cells either harbored large chromosomal deletions or were subject to translocations. Notably, the quantitative ddPCR data were in good agreement with data shown in **Fig. 2** and **Fig. 4**, and confirmed that well-designed nucleases provoke considerably less chromosomal translocations.

As part of future preclinical risk assessments, careful investigations of the translocation events must be carried out. An initial analysis considering a wide region that can be potentially affected by the chromosomal rearrangements highlights the presence of some proto-oncogenes in proximity to some of the identified chromosomal translocations (**Table S2**), further corroborating the value of CAST-Seq.

Discussion

Genome-wide methods to detect off-target activities of ZFNs, TALENs, and CRISPR-Cas nucleases have been pivotal for characterizing and improving the specificities of these engineered nucleases (Kim et al., 2019; Tsai and Joung, 2016). Here, we present CAST-Seq as a novel and sensitive methodology that enables scientists to detect, categorize and quantify chromosomal rearrangements prompted by on-target as well as off-target activities of designer nucleases. Unlike previously described assays, CAST-Seq is (i) quantitative, (ii) able to discover on-target genomic deletions, (iii) detect previously undescribed types of chromosomal aberrations, such as homology-mediated translocations, and (iv) is performed directly in the clinically relevant cell type.

The linear correlation between the numbers of CAST-Seq hits in a cluster and the number of chromosomal rearrangements in a specified region confirmed the quantitative nature of the method and revealed its high sensitivity. Quantification is based on the fact that chromosomal breaking points in combination with the adapter ligation site create a unique molecular signature, which enabled us to compute the number of individual translocations, to group them into clusters that are likely prompted by the same trigger, and to quantify the frequencies of such events. Our data revealed that up to 0.5% of cells harbor bona-fide chromosomal translocations, up to 1.6% of cells reveal HMT events if a closely related gene is present, and some 20% of cells contain gross chromosomal aberrations at on-target sites.

Bradley and colleagues previously reported significant designer nuclease-associated on-target mutagenesis in primary murine cells (Kosicki et al., 2018). Our study confirmed such large 10-kb deletions/inversions at the on-target sites and extend these observations both qualitatively and quantitatively by combining CAST-Seq analysis with a ddPCR strategy. Translocations with the second on-target site on the homologous chromosome that led to acentric and dicentric chromosomes are somewhat less frequent but still prominent, suggesting that chromosomal rearrangements at the on-target site seem difficult to avoid. On the other hand, bona-fide chromosomal translocations can be largely averted by smart choice of the target site and by the use of highly specific CRISPR-Cas9 nucleases, such as shown for the *CCR5*$^{\#2}$ and *FANCF* targeting nucleases.

While a few studies reported translocations between two nuclease-induced cleavage events (Brunet et al., 2009; Chiarle et al., 2011; Frock et al., 2015; Giannoukos et al., 2018; Kosicki et al., 2018), we found that DSBs is just one of the factors that drives chromosomal aberrations. Our data demonstrate for the first time that regions that share substantial homology to the on-target region are subject to RAD51-dependent chromosomal rearrangements, even if they do not contain an off-target site. In particular, when targeting a locus that is flanked by a closely related gene (or pseudogene), the likelihood of inducing chromosomal rearrangements is high. Up to 1.6% of cells contained large 15-kb deletions between *CCR5* and *CCR2* although we did not detect off-target activity in *CCR2*. In order to prevent such unwanted chromosomal aberrations, it is prudent to avoid target loci that share stretches of sequence homology with chromosomal regions somewhere else in the genome whenever possible.

As genome editing for clinical applications is further developed, it is paramount to co-develop (pre)clinical risk assessment tools to carefully monitor the introduced genetic changes. Our study confirms previous reports that indicate a high false-positive rate of *in vitro* assays (Kim et al., 2019). We did not find indels in gene-edited HSPCs at a majority of the top ranked OTs predicted by GUIDE-Seq and/or CIRCLE-Seq, indicating that CAST-Seq did not miss OT-triggered translocations but rather that these sites were not cleaved in HSPCs. It is important to keep in mind that GUIDE-Seq and CIRCLE-Seq are forecast tools that predict where OTs may occur in the clinically relevant cell type. CAST-Seq, on the other hand, identifies and quantifies the chromosomal rearrangements that occurred after editing of the clinically relevant cell type.

The sensitivity of CAST-Seq (1 chromosomal aberration in ~7,500 cells) can be further improved by employing higher amounts of genomic input DNA or by performing CAST-Seq in both directions. As shown in the IGVs, the deletion profile around the on-target site is not evenly distributed, suggesting that additional on-target site aberrations can be detected if CAST-Seq was performed from either side. Of note, the sensitivity of CAST-Seq is already higher than the 0.1% detection limit of NGS-based amplicon sequencing, which is typically used to detect indel mutations at predicted off-target sites. This means that, at least theoretically, CAST-Seq may be able to detect OMTs at sites that cannot be identified by NGS-based sequencing approaches. Additionally, with few adaptations in decoy primer design, CAST-Seq can be adjusted to assess chromosomal rearrangements after HR-based genome editing with a donor template.

In the absence of biological tests to read out the consequences of the genetic insults elicited by designer nucleases, one has to rely on surrogate methods such as CAST-Seq. Because translocations are a hallmark of cancerous cells, chromosomal rearrangements may constitute a first oncogenic 'hit' in stem cells, which have a long replicative lifespan and may become neoplastic with time. Such a scenario is reminiscent of the activation of proto-oncogenes by retroviral insertion in CD34+ cells, which caused leukemia in some of the early gene therapy trials (Baum et al., 2004). On the other hand, it is important to keep in mind that as for indel mutations and as for the majority of retroviral insertion sites, most translocations may be biologically inert. CAST-Seq will be helpful to identify the rare, potentially tumorigenic chromosomal rearrangements as part of a preclinical risk management.

In conclusion, CAST-Seq is a novel, highly sensitive NGS-based assay for the identification and quantification of unintended chromosomal rearrangements that occur in addition to the more typical indel mutations at OT sites. CAST-Seq will not only allow researchers to qualitatively identify chromosomal rearrangements but also to quantitatively track and enumerate the clonal expansion of translocation events over time. CAST-Seq is hence especially important in therapeutic genome editing settings, where chromosomal aberrations need to be carefully monitored to assess and mitigate the clinical risk associated with the use of a specific engineered nuclease.

Acknowledgments

We thank Adrian Thrasher for support, Marianna Romito for the production of TALEN mRNA, and all members of our laboratories for constructive discussions and suggestions.

This work was supported by the German Federal Ministry of Education and Research (IFB–01EO0803 to TCa, TIC, and CM; coNfirm–01ZX1708F to MB; MIRACUM–01ZZ1801B to MB) and the German Research Foundation (SFB 850 to MB).

Author Contributions

Conceptualization (GT, GA, MB, TCa), development or design of methods (GT, GA, VP, TCa), programming and software development (GA, GT, GM), performing experiments (GT, GA, GB, VP, JK, SP), data analysis (all authors), writing manuscript (TCa, GT, GA), supervision (TCa, GT, MB, TIC, CM), and funding acquisition (TCa, MB, TIC, CM).

Declaration of Interests

TCa and TIC have sponsored research collaborations with Cellectis and Miltenyi Biotec. All other authors report no conflicts of interest.

Methods

Cell culture and transfection. Cryopreserved human CD34$^+$ HSPCs derived from cord blood (AllCells, Cat.# CB008F; StemCell-Technologies, Cat.# 70008.3) were thawed and cultivated in a density range of 0.5–1 x 10^6/ml at 37°C, 5% CO_2 in GMP-grade CellGro media (CellGenix, Cat.# 20802-0500) complemented with TPO (100 ng/ml; Peprotech Cat.# 300-18), Flt-3 (300 ng/ml; Immunotools, Cat.# 60100864), SCF (300 ng/ml; Immunotools Cat.# 11343327), IL-3 (60 ng/ml; Immunotools Cat.# 11340035), 20 mg/ml streptomycin and 20 U/ml penicillin (Sigma Cat.# P0781). Cell viability was determined by flow cytometric scatter blot analysis or on a NucleoCounter NC-250 (Chemotech, Denmark) using Solution 18 AO•DAPI staining (Chemometec, Cat.# 910-3018). For nucleofection, 5–7 x 10^5 CD34+ cells were resuspended in 20 µL of P3 solution (Lonza, Cat.# V4XP-3032) and mixed with previously assembled RNPs or TALEN mRNAs. For RNP assembly, 6 µg of Cas9 protein (PNA Bio, Cat.# CP02; IDT Alt-R® S.p. HiFi Cas9 Nuclease V3, 1081060) were complexed for five minutes with 225 pmol of gRNA (three 2'-

O-methyl phosphorothioate linked nucleotides at either 5’ and 3’-end; Synthego, USA). TALEN encoding mRNA was generated as previously described (Patsali et al., 2019). Nucleofection was performed with a 4D-Nucleofector (Lonza, Germany), program CA137 in a 16-well cuvette format.

K562 cells were cultivated at a density of 0.5–1 x 10^6/ml at 37°C, 5% CO_2 in RPMI1640 medium supplemented with GlutaMAX™ (ThermoFisher, Cat.# 61870010), 10% FBS (PAN-Biotech, Cat.# P40-47500), 0.1 mg/ml streptomycin and 100 U/ml penicillin (Sigma Cat.# P0781). For transfection, 1.5 x 10^6 cells were resuspended in 20 µL of SF solution (Lonza, Cat.# V4XC-2032), mixed with TALEN mRNA or RNPs (see above), and nucleofected with program FF120 in a 16-well cuvette format. Cells were transferred to medium containing either the RAD51 inhibitor B02 (Merck, Cat.# SML0364) at a final concentration of 10 µM or DMSO as a control. Fresh B02 was added 24 hours later, and cells harvested on day 4 to extract genomic DNA using QIAamp® DNA Blood Mini Kit (Qiagen, Cat.# 51306).

CAST-Seq sample preparation*.* Genomic DNA of at least 5 x 10^5 cells was isolated at the indicated time points using QIAamp® DNA Blood Mini Kit (Qiagen, Cat.# 51306) and fragmented by sonication (M220 Focused Ultrasonicator) or enzymatic digestion (NEBNext® Ultra™ II FS DNA Library Prep Cat.# E7805S) to obtain average fragment lengths of 350 bp. The DNA was end repaired and a protruding 3’-A nucleotide was added according to the manufacturer’s instructions (NEBNext® Ultra™ II FS DNA Library Prep Cat.# E7805S) in order to enable the ligation of a linker sequence with a protruding 3’-T. After DNA purification (Qiagen, PCR purification kit, Cat.# 28104), two rounds of PCR utilizing Q5 polymerase (NEB, Cat.# M0493S) was performed using the following conditions: 20 cycles at 95°C for 15 s, 63°C (first reaction) or 68°C (second reaction) for 20 s, 72°C for 20 s. The first reaction was performed with primers complementary to the linker sequence and to a sequence in close proximity to the on-target site. One or two decoy primers were introduced to reduce full-length amplification of the fragments which contain the on-target sequence without a chromosomal aberration event. The second PCR utilized nested primers to reduce the amount of mis-amplified fragments, while the third PCR introduced the barcoded Illumina adapter for sequencing (NEB, NEBNext Multiplex Oligos for Illumina, Cat.# E7335). Denatured amplicons at 6-10 pM were loaded into the Illumina MiSeq Reagent Kit V2-500 cycle (Illumina, Cat.# MS-102-2003) according to the manufacturer instruction. Oligonucleotide sequences are reported in **Suppl. Table 4**.

Bioinformatic analysis. ***Alignment:*** Mate paired reads from Illumina miSeq sequencing were merged using FLASH software (Magoc and Salzberg, 2011). BBmap was used for filtering and trimming as follow: merged reads containing the designer nuclease target site were filtered-in, whereas PCR mispriming products reads were filtered-out. Linker sequences, Illumina adapter sequences, targeted elongation sequence and bad quality reads were trimmed. Selected reads were aligned to the human genome GRCh38 (hg38) using Bowtie2 (Langmead and Salzberg, 2012) and the very-sensitive preset parameters to maximize the alignment accuracy. To reduce the probability of finding false positives, aligned reads with good mapping quality (MAPQ >15) were selected. The aligned BAM file was converted into bed file using BEDTools (Quinlan and Hall, 2010) (see **Suppl. Table 5**).

Deduplication/cluster definition: Reads located on the same coordinates were considered as PCR-derived duplicates and therefore deduplicated. To cope with translocation point or linker ligation sequencing/alignment biases, we added a tolerance of +/-3 bp. Hence, all reads within this +/-3 bp window were deduplicated and the total amount of reads was stored to quantify the translocation event identified as a “hit”. High hit density regions were determined using a random set of regions of the human genome to estimate distance distribution between two consecutive elements. A threshold distance of 2,500 bp achieved a significant p-value ($p<0.05$) in all tested samples. Subsequently, consecutive hits separated by less than 2,500 bp were merged into clusters, representing all putative translocation sites. When comparing more than one replicate for a sample, two proximal clusters could be merged during the bioinformatic process (*CCR5/CCR2* and *HBB/HBD*), and the individual clusters were manually recovered by re-setting the borders. Finally, the significance of the identified clusters was evaluated compared to a non-treated control sample using a Fisher’s exact test. Significance threshold was set for adjusted p-value (Benjamini-Hochberg) below 0.05.

Translocation event classification: Translocation sites were classified into three groups: Off-target-mediated translocation (OMT), homology-mediated translocation (HMT), and naturally occurring break site (NBS)-derived translocations. To assess statistical significance of the groups, a set of 10,000 random region sequences of 500 bp length was chosen over the entire human genome. These sequences were later used to calculate p-value from the empirical cumulative distribution. For OMT, translocation sites were aligned to the on-target sequence. A nucleotide substitution matrix using +1 and -1 as weights for match and mismatch, respectively, was built (**Suppl. Table 6**). Gaps were allowed with the same penalty weight as mismatch. A pairwise alignment from Biostrings R Package with “local-global” type of alignment was used. OT alignment scores were calculated for identified translocation sites and random sequences. For HMT, the longest common substring (LCS) between left and right 2500 bp flanking regions around the translocation site, and the known 5 kb window around the expected OT, was calculated for identified translocation sites and random sequences. A 25 bp LCS threshold obtained a significant p-value, i.e. longer than the top 5% of LCS in random sequences, and was reported to be sufficient to initiate HR. Finally, every single translocation site was categorized as follow: OMT if OT alignment score was higher than the top 5% scores on random sequences; HMT if LCS longer than 25 bp; NBS otherwise (see **Suppl. Table 7**).

Annotation. Selected translocation sites were annotated with the nearest gene or gene region (e.g. promoter, exon, intron, etc.), based on distance to transcriptional start site (TSS) reported in the Bioconductor Annotation Package TxDb.Hsapiens.UCSC.hg38.knownGene (**Suppl. Table 7**). The set of genes that is located within a window of 100 kb around the translocation site is reported, specifically highlighting cancer-related genes based on the OncoKB database (Chakravarty et al., 2017).

Molecular analyses. T7E1 assay was performed and analyzed as previously described (Dreyer et al., 2015). For ddPCR, 150-550 ng of genomic DNA were digested with 5 U of HindIII HF or AvrII (NEB) at 37°C for 30 min to reduce sample viscosity. After digestion, either 100 ng (translocation) or 20 ng (large deletion) of digested genomic DNA were added to the ddPCR reaction mix containing QX200TM EvaGreen ddPCR Supermix TM (Bio-Rad, Cat.# 1864034). Each reaction was complexed with 100 nM of primers and loaded into the QX200 Droplet Generator (Bio-Rad). The generated droplets were transferred to a 96-well PCR plate (Bio-Rad, Cat.# 12001925) and the plate sealed with a PX1 PCR plate sealer (Bio-Rad). For all assays, endpoint PCR was performed: lid preheat at 95°C for 5 min, 50 cycles of 95°C for 30 s, 62°C for 60 s, 72°C for 2 min, followed by 5 min at 4°C and 5 min at 90°C (ramping rate set to 2°C/s). Data was acquired in a QX200 Droplet Reader and results analyzed with QuantaSoftTM Analysis Pro (Bio-Rad). Results were considered significant if at least 10,000 droplets/20 µl reaction were generated. To calculate the frequencies of translocations, the ddPCR values were first corrected for noise (subtraction of value of untreated matched control) and then normalized for the amount of genomic input DNA using an internal control (*STAT3*). To calculate the frequencies of ‘large deletions’ and ‘other aberrations’, the ddPCR values were first corrected for noise (subtraction of value of untreated matched control) and then normalized for the amount of genomic input DNA by dividing the number by the average of the two values obtained for the control genes (*RAD1*, *STAT3*). The average value from 5’ and 3’ assays was used to determine the fraction of large deletions. The fraction of ‘other aberrations’ was calculated by subtracting the fraction of large deletions from the ‘Edge’ value. The indel percentage from T7E1 assay was recalculated based on the formula: (100-(large deletion x 100)-(translocation x 100)) x indel%.

For validation of CAST-Seq, some HMT and/or OMT sites were analyzed by NGS. PCR primers were designed to amplify a 300–430 bp genomic segment comprising the putative HMT and/or OMT sites. The amplicons were subjected to end-repair, adaptor ligation and an indexing PCR using NEBNext® Ultra™ II DNA library prep kit for Illumina, as described above. The denatured amplicons were loaded at 6-10 pM into the Illumina MiSeq Reagent Kit V2 - 500 cycle (Illumina, Cat.# MS-102-2003) according to the manufacturer’s instructions. The FASTQ files were analyzed for indels using the command line version of CRISPResso (Pinello et al., 2016), considering 40 bp around the supposed cleavage or translocation site but disregarding substitutions. The derived indel proportion of the treated sample was compared to the corresponding values of the untreated sample in a one-tailed Z-test, and corrected with the standard deviation of untreated sample values in order to account for variability of measurements. All oligonucleotide sequences are reported in **Suppl. Table 4**.

Data availability. All data generated or analyzed during this study are included in this published article and its supplementary data files.

Supplemental Information

- Figures S1 to S8
- Tables S1 to S7

References

Alwin, S., Gere, M.B., Guhl, E., Effertz, K., Barbas, C.F., 3rd, Segal, D.J., Weitzman, M.D., and Cathomen, T. (2005). Custom zinc-finger nucleases for use in human cells. Mol Ther *12*, 610-617.

Ayares, D., Chekuri, L., Song, K.Y., and Kucherlapati, R. (1986). Sequence homology requirements for intermolecular recombination in mammalian cells. Proc Natl Acad Sci U S A *83*, 5199-5203.

Bailey, S.R., and Maus, M.V. (2019). Gene editing for immune cell therapies. Nat Biotechnol *37*, 1425-1434.

Baum, C., von Kalle, C., Staal, F.J., Li, Z., Fehse, B., Schmidt, M., Weerkamp, F., Karlsson, S., Wagemaker, G., and Williams, D.A. (2004). Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. Mol Ther *9*, 5-13.

Brunet, E., Simsek, D., Tomishima, M., DeKelver, R., Choi, V.M., Gregory, P., Urnov, F., Weinstock, D.M., and Jasin, M. (2009). Chromosomal translocations induced at specified loci in human stem cells. Proc Natl Acad Sci U S A *106*, 10620-10625.

Cameron, P., Fuller, C.K., Donohoue, P.D., Jones, B.N., Thompson, M.S., Carter, M.M., Gradia, S., Vidal, B., Garner, E., Slorach, E.M., *et al.* (2017). Mapping the genomic landscape of CRISPR-Cas9 cleavage. Nat Methods *14*, 600-606.

Carroll, D. (2014). Genome engineering with targetable nucleases. Annu Rev Biochem *83*, 409-439.

Cathomen, T., Schule, S., Schussler-Lenz, M., and Abou-El-Enein, M. (2019). The Human Genome Editing Race: Loosening Regulatory Standards for Commercial Advantage? Trends Biotechnol *37*, 120-123.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res *39*, e82.

Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H.*, et al.* (2017). OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol *2017*, 1-16.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J.*, et al.* (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107-119.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol *31*, 230-232.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A.*, et al.* (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819-823.

Cornu, T.I., Mussolino, C., and Cathomen, T. (2017). Refining strategies to translate genome editing to the clinic. Nat Med *23*, 415-423.

Dreyer, A.K., Hoffmann, D., Lachmann, N., Ackermann, M., Steinemann, D., Timm, B., Siler, U., Reichenbach, J., Grez, M., Moritz, T.*, et al.* (2015). TALEN-mediated functional correction of X-linked chronic granulomatous disease in patient-derived induced pluripotent stem cells. Biomaterials *69*, 191-200.

Frock, R.L., Hu, J., Meyers, R.M., Ho, Y.J., Kii, E., and Alt, F.W. (2015). Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nat Biotechnol *33*, 179-186.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A *109*, E2579-2586.

Giannoukos, G., Ciulla, D.M., Marco, E., Abdulkerim, H.S., Barrera, L.A., Bothmer, A., Dhanapal, V., Gloskowski, S.W., Jayaram, H., Maeder, M.L.*, et al.* (2018). UDiTaS, a genome editing detection method for indels and genome rearrangements. BMC Genomics *19*, 212.

Huang, F., Motlekar, N.A., Burgwin, C.M., Napper, A.D., Diamond, S.L., and Mazin, A.V. (2011). Identification of specific inhibitors of human RAD51 recombinase using high-throughput screening. ACS Chem Biol *6*, 628-635.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816-821.

Kim, D., Bae, S., Park, J., Kim, E., Kim, S., Yu, H.R., Hwang, J., Kim, J.I., and Kim, J.S. (2015). Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. Nat Methods *12*, 237-243, 231 p following 243.

Kim, D., Luk, K., Wolfe, S.A., and Kim, J.S. (2019). Evaluating and Enhancing Target Specificity of Gene-Editing Nucleases and Deaminases. Annu Rev Biochem *88*, 191-220.

Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol *36*, 765-771.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Magoc, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957-2963.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823-826.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J.*, et al.* (2011). A TALE nuclease architecture for efficient genome editing. Nat Biotechnol *29*, 143-148.

Mueller, P.R., and Wold, B. (1989). In vivo footprinting of a muscle specific enhancer by ligation mediated PCR. Science *246*, 780-786.

Mussolino, C., Morbitzer, R., Lutge, F., Dannemann, N., Lahaye, T., and Cathomen, T. (2011). A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. Nucleic Acids Res *39*, 9283-9293.

Patsali, P., Turchiano, G., Papasavva, P., Romito, M., Loucari, C.C., Stephanou, C., Christou, S., Sitarou, M., Mussolino, C., Cornu, T.I.*, et al.* (2019). Correction of IVS I-110(G>A) beta-thalassemia by CRISPR/Cas-and TALEN-mediated disruption of aberrant regulatory elements in human hematopoietic stem and progenitor cells. Haematologica *104*, e497-e501.

Pinello, L., Canver, M.C., Hoban, M.D., Orkin, S.H., Kohn, D.B., Bauer, D.E., and Yuan, G.C. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. Nat Biotechnol *34*, 695-697.

Porteus, M.H., and Baltimore, D. (2003). Chimeric nucleases stimulate gene targeting in human cells. Science *300*, 763.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Roukos, V., Voss, T.C., Schmidt, C.K., Lee, S., Wangsa, D., and Misteli, T. (2013). Spatial dynamics of chromosome translocations in living cells. Science *341*, 660-664.

Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., Ball, C., Pilz, I., Braun, S., Glimm, H., and von Kalle, C. (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat Methods *4*, 1051-1057.

Smith, J., Bibikova, M., Whitby, F.G., Reddy, A.R., Chandrasegaran, S., and Carroll, D. (2000). Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. Nucleic Acids Res *28*, 3361-3369.

Tsai, S.Q., and Joung, J.K. (2016). Defining and improving the genome-wide specificities of CRISPR-Cas9 nucleases. Nat Rev Genet *17*, 300-312.

Tsai, S.Q., Nguyen, N.T., Malagon-Lopez, J., Topkar, V.V., Aryee, M.J., and Joung, J.K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat Methods *14*, 607-614.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P.*, et al.* (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol *33*, 187-197.

Urnov, F.D., Miller, J.C., Lee, Y.L., Beausejour, C.M., Rock, J.M., Augustus, S., Jamieson, A.C., Porteus, M.H., Gregory, P.D., and Holmes, M.C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. Nature *435*, 646-651.

Vakulskas, C.A., Dever, D.P., Rettig, G.R., Turk, R., Jacobi, A.M., Collingwood, M.A., Bode, N.M., McNeill, M.S., Yan, S., Camarena, J.*, et al.* (2018). A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. Nat Med *24*, 1216-1224.

Ward, A., Khanna, K.K., and Wiegmans, A.P. (2015). Targeting homologous recombination, new pre-clinical and clinical therapeutic combinations inhibiting RAD51. Cancer Treat Rev *41*, 35-45.

Wienert, B., Wyman, S.K., Richardson, C.D., Yeh, C.D., Akcakaya, P., Porritt, M.J., Morlock, M., Vu, J.T., Kazane, K.R., Watry, H.L.*, et al.* (2019). Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. Science *364*, 286-289.

Yan, W.X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M.W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L.*, et al.* (2017). BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. Nat Commun *8*, 15058.

Figure Legends

Figure 1. Experimental overview. (**a**) Schematic representation of chromosomal aberrations induced by on-target and off-target activities of designer nucleases. (**b**) CAST-Seq library preparation. Simultaneous activity of designer nucleases at an on-target (blue) and an off-target (yellow) site can induce e.g. a reciprocal translocation (black arrow). In most cases, no translocation will happen (right). Genomic DNA of untreated and gene-edited cells is randomly fragmented and end-repaired to add a 3'-A overhang, which is used for ligation of a short linker (red). 1st PCR is performed with bait and prey primers (open arrows) binding to the target site and the linker, along with 'decoy' primers (filled arrows). 2nd PCR with nested primers adds adaptors that are used in 3rd PCR to add barcodes. (**c**) Experimental overview. (**d**) On-target activity. Indel frequencies were determined by T7E1 assay 4 days post-transfection. (**e**) Cell viability. Viability was examined 24 h post-electroporation.

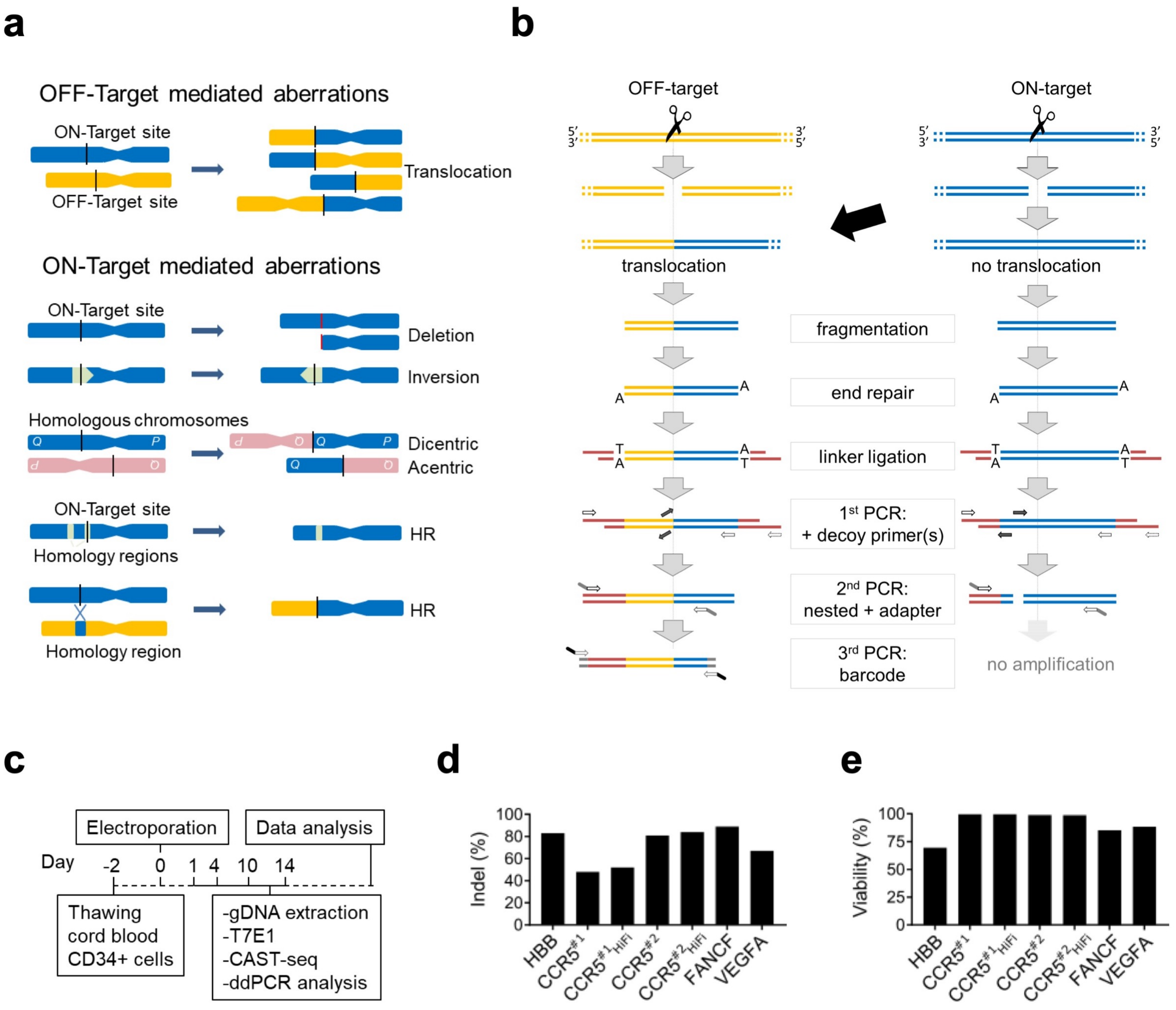
Figure 2. CAST-Seq analysis of *CCR5*$^{\#1}$ targeting CRISPR-Cas9 nuclease. (**a**) Schematic of decoy strategy. Prey and bait primers bind to linker (red) and on-target site (blue) to amplify chromosomal aberrations. Decoy primers bind in close proximity to on-target site but opposite to bait primer in order to prevent the formation of full-length amplicons at non-modified target sites (left). (**b**) Qualitative CAST-Seq analysis. Integrative Genomics Viewer (IGV) plots illustrate CAST-Seq reads surrounding the target site within a window of 33 kb. Mapped CAST-Seq reads are represented by bars (only top 7 lines shown). Blue and red bars indicate sequences aligning to negative or positive strand, respectively. Coverage, i.e. the number of mapped reads, is indicated on the middle, gene locations on the bottom. Positions of on-target site and *CCR2* HMT cluster are emphasized by dotted lines. (**c**) Target site alignment. Reference *CCR5*$^{\#1}$ target site is shown on top (N, any nucleotide; R, purine). Mismatched nucleotides and deletions/insertions (-1/1) are highlighted. Number of hits are listed on the left, categories on right. (**d**) Indel analysis. Targeted deep amplicon sequencing was performed on identified HMT and/or OMT sites of genomic DNA harvested 4 days after gene editing with Cas9 or HiFi-Cas9. Statistically significant differences are indicated by **** (p<0.0001; Z-test corrected by standard deviation calculated on untreated (UT) cells). (**e**) Graphical representation of some rare complex rearrangements found at on-target site. *CCR2* (pink) and *CCR5* (grey) derived sequences (top) or a long stretch of an inverted/duplicated *CCR5* sequence (grey, bottom). (**f**) Visualization of chromosomal rearrangements. Circos plot shows on-target site cluster (ON, green), OMT (red), HMT (blue), NBS (grey), or ambiguous OMT/HMT (yellow). From outer to inner layer: black rectangles show DNA location of the translocation sites. Grey rectangles represent coding TSS to TES coordinates of *CCR5* and *CCR2* genes. Red ring indicates alignment score against gRNA sequence, with significant score

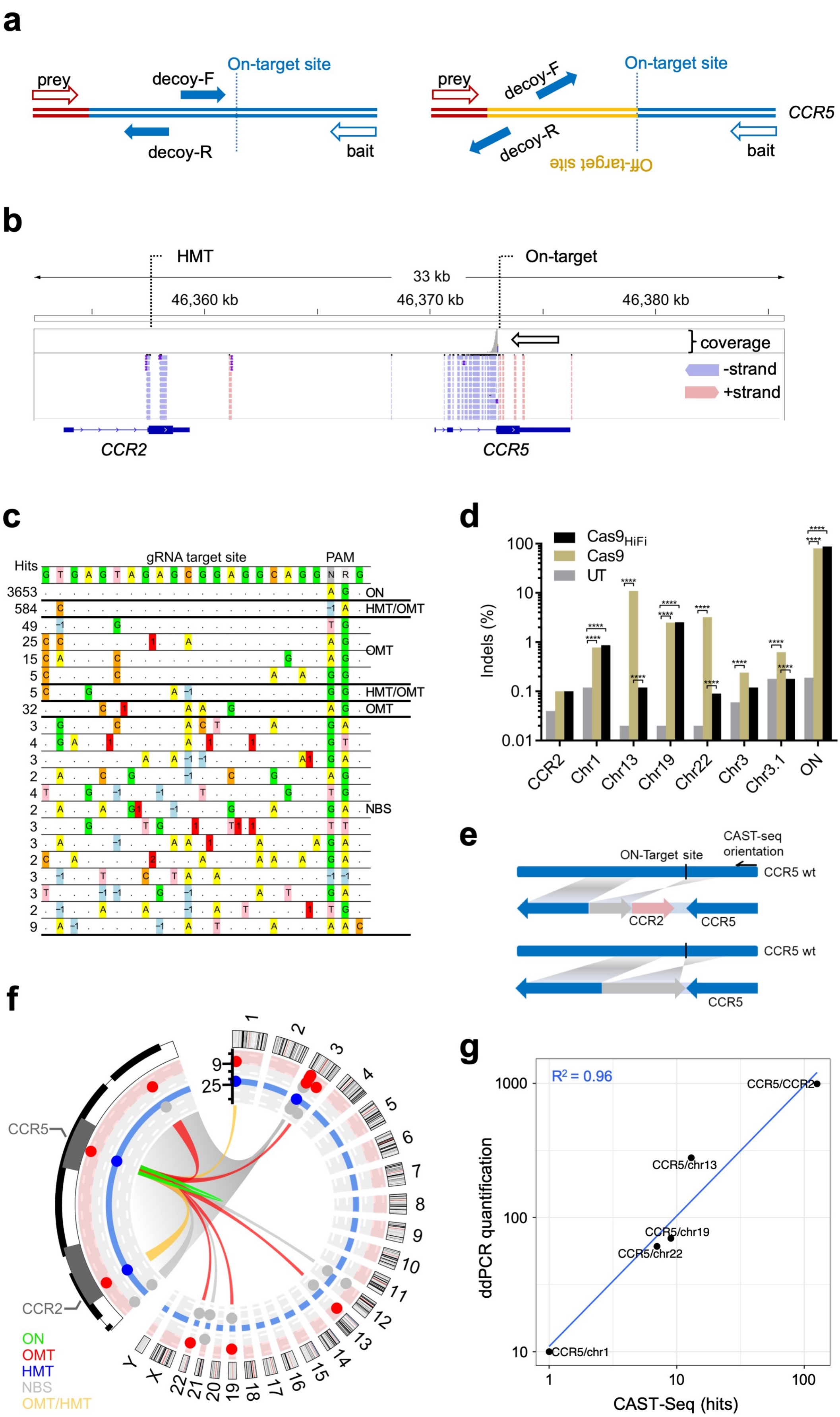
accentuated by red dots. Blue ring indicates length of sequence homology, with significant lengths emphasized by blue dots. *CCR5* target region is enlarged on the left. Arcs highlight the identified translocations between OT and other sites. (**g**) Quantification. The number of chromosomal rearrangements quantified by CAST-Seq or ddPCR are represented in scatter plot. Linear regression line (blue) and squared correlation coefficient (R^2) are indicated.

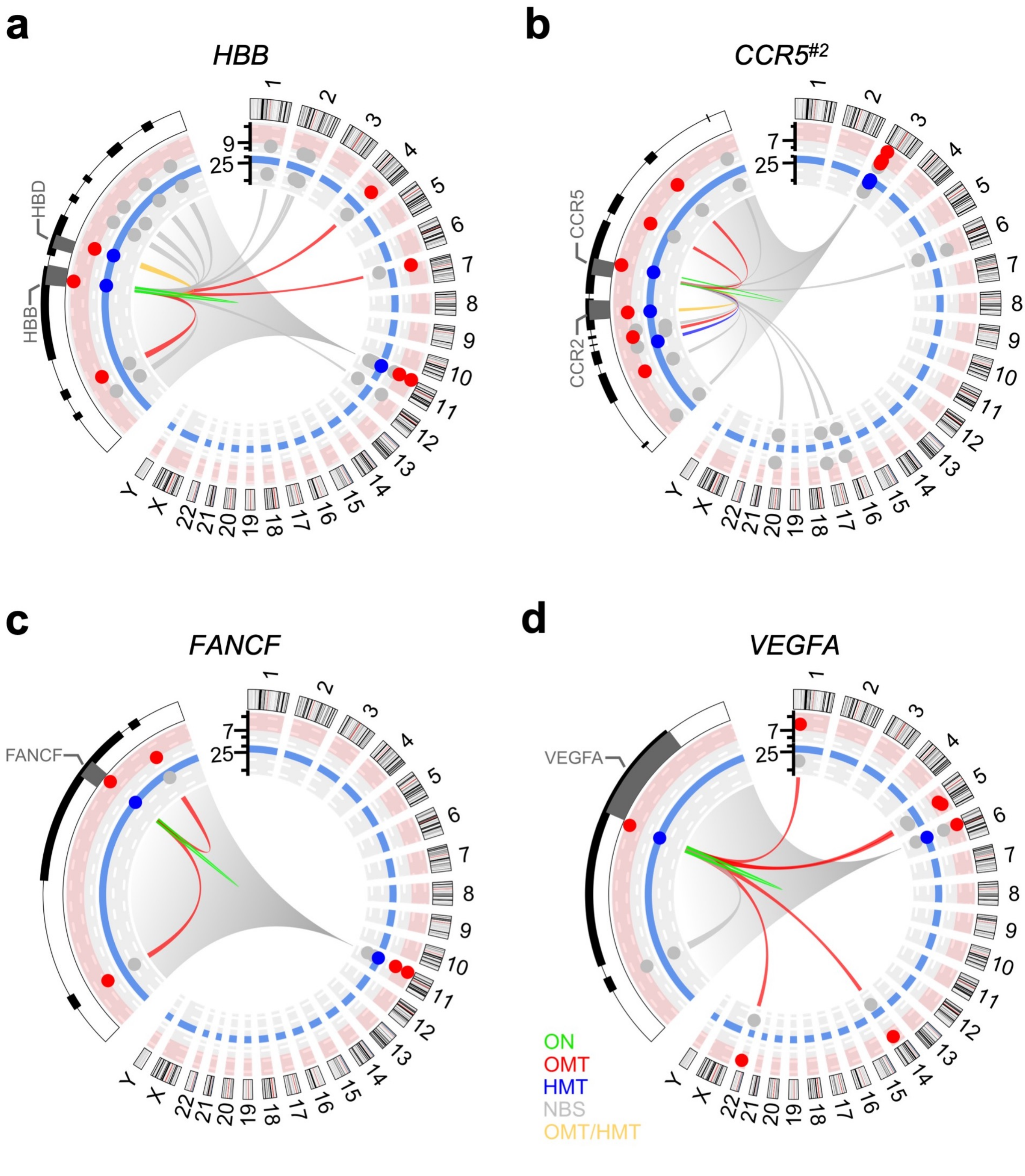
Figure 3. CAST-Seq analysis of CRISPR-Cas9 or TALEN targeted genomic sites. (**a-d**) Visualization of chromosomal aberrations. Circos plots summarize CAST-Seq analysis of *HBB* targeting TALEN pair (a) as well as CRISPR-Cas9 targeting *CCR5*$^{\#2}$ (b), *FANCF* (c) and *VEGFA* (d).

Figure 4. Dynamics. (**a-c**) Qualitative visualization. Integrative Genomics Viewer (IGV) plots show target region, *CCR5*$^{\#1}$ (a), *CCR5*$^{\#2}$ (b) and *HBB* (c), within a window of 33 kb. Only top rows are shown. White arrows indicate bait orientation and dotted vertical lines the on-target site. Harvesting time in days post-electroporation (D1, D4, D14) is indicated on the left. (**d-f**) Quantitative analysis. Plots show number of clustered CAST-Seq hits for D1 to D14 samples of CRISPR-Cas targeting *CCR5*$^{\#1}$ (d) and *CCR5*$^{\#2}$ (e) or TALEN targeting *HBB* (f). Cluster category (HMT and/or OMT) is indicated.

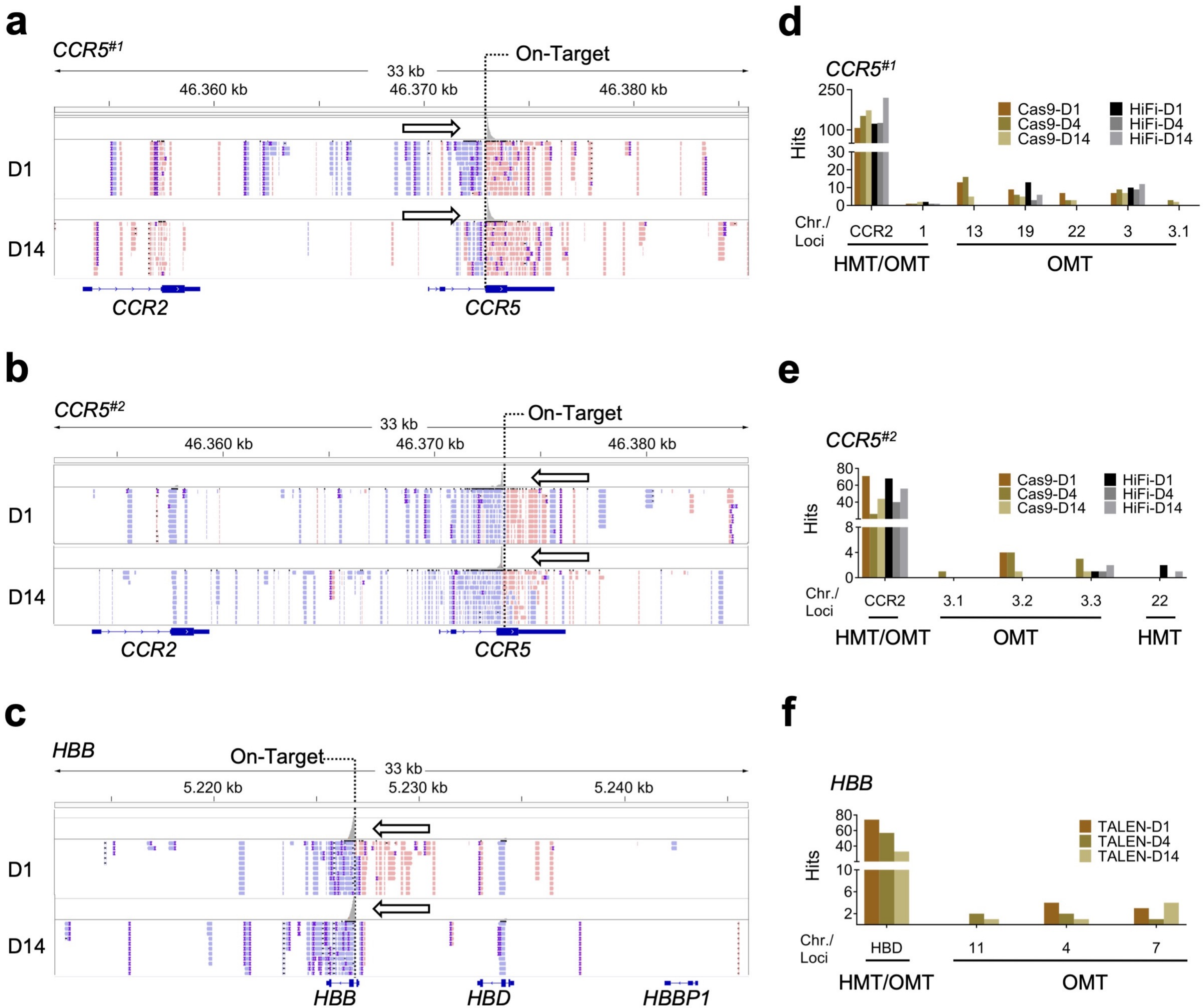
Figure 5. DNA repair kinetics and quantification of chromosomal aberrations. (**a**) ddPCR strategy. The 'edge amplicon' (~200 bp) encompass the cleavage site and is flanked by 5' or 3' amplicons to either site of the target site. Translocation are expected to reduce the amount of edge amplicon products, while large deletions will also reduce the quantity of the flanking amplicons. Amplicons positioned at the telomeric side (telo.) and the opposite chromosome arm (q arm) relative to the target site, as well as two control amplicons (cto.) on other chromosome, were used to establish the relative change of amplifiable on-target copies. (**b-d**) Variation of target site copy numbers. Plots show relative copy number variation (CNV) of amplifiable target sites in CD34+ cells edited with CRISPR-Cas targeting *CCR5*$^{\#1}$ (b) or *CCR5*$^{\#2}$ (c), or with a TALEN targeting *HBB* (d), at different time points (day 1 to day 14) after transfection. (**e-g**) Data summary. ddPCR results were used to normalize (Norm.) the indel frequencies determined by T7E1 assay for D4 time points. 'Large deletion' denotes the relative decrease of the average number of flanking amplicons while 'other aberrations' is specified as the relative difference between the number of edge amplicons and the average number of flanking amplicons.


a
OFF-Target mediated aberrations
ON-Target site
OFF-Target site
Translocation
ON-Target mediated aberrations
ON-Target site
Deletion
Inversion
Homologous chromosomes
Q
P
d
O
Dicentric
Acentric
ON-Target site
Homology regions
HR
HR
Homology region
b
OFF-target
ON-target
5′
3′
translocation
no translocation
fragmentation
end repair
A
T
linker ligation
1st PCR:
+ decoy primer(s)
2nd PCR:
nested + adapter
3rd PCR:
barcode
no amplification
c
Electroporation
Data analysis
Day
-2
0
1
4
10
14
Thawing
cord blood
CD34+ cells
-gDNA extraction
-T7E1
-CAST-seq
-ddPCR analysis
d
Indel (%)
100
80
60
40
20
0
HBB
CCR5#1
CCR5#1HiFi
CCR5#2
CCR5#2HiFi
FANCF
VEGFA
e
Viability (%)
100
75
50
25
0
HBB
CCR5#1
CCR5#1HiFi
CCR5#2
CCR5#2HiFi
FANCF
VEGFA

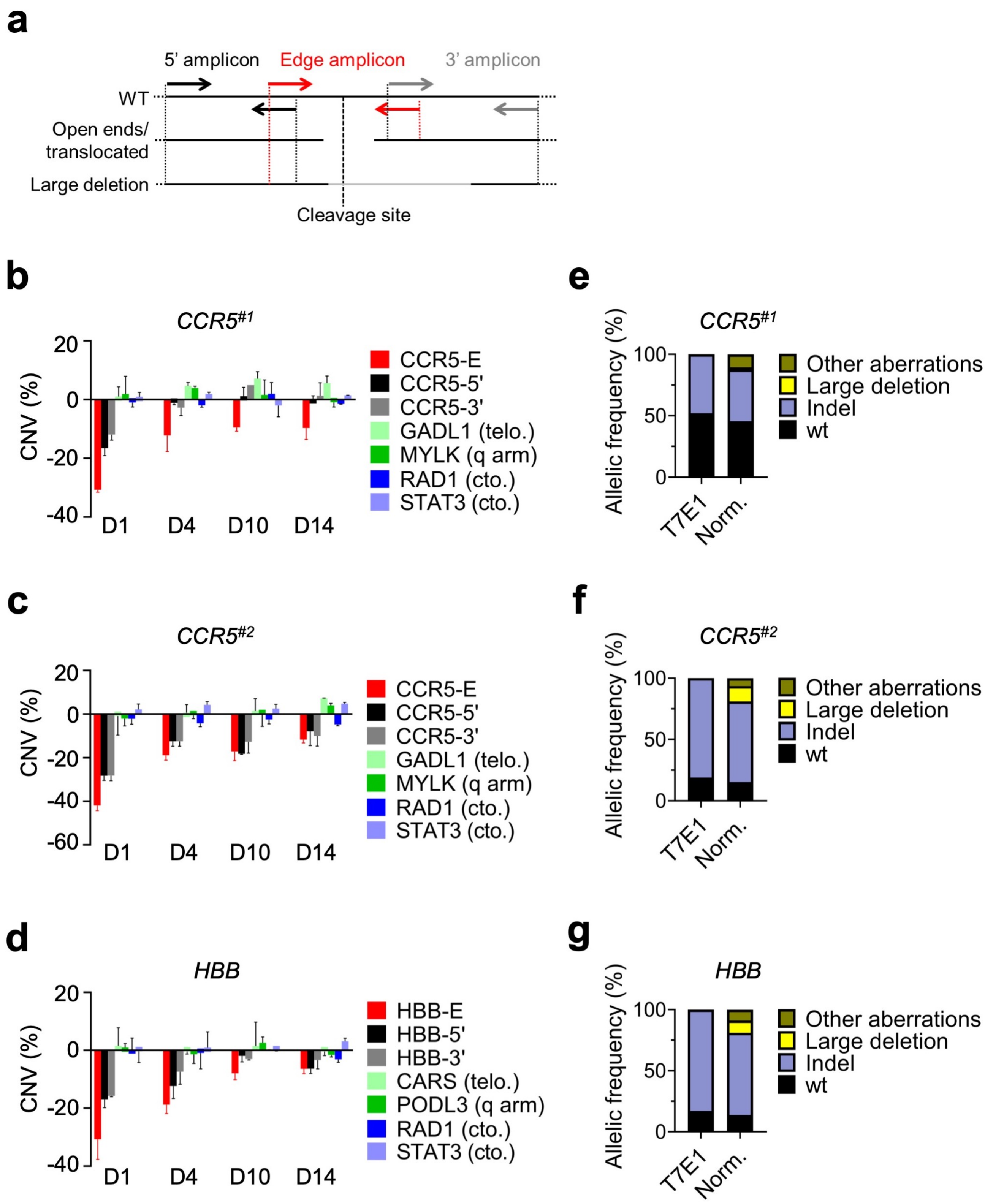

a
prey
decoy-F
On-target site
decoy-R
bait
Off-target site
CCR5
b
HMT
On-target
33 kb
46,360 kb
46,370 kb
46,380 kb
coverage
-strand
+strand
CCR2
CCR5
c
Hits
gRNA target site
PAM
ON
HMT/OMT
OMT
NBS
d
Cas9HiFi
Cas9
UT
Indels (%)
CCR2
Chr1
Chr13
Chr19
Chr22
Chr3
Chr3.1
ON
e
ON-Target site
CAST-seq orientation
CCR5 wt
CCR2
CCR5
f
ON
OMT
HMT
NBS
OMT/HMT
g
R² = 0.96
CCR5/CCR2
CCR5/chr13
CCR5/chr19
CCR5/chr22
CCR5/chr1
ddPCR quantification
CAST-Seq (hits)



a
HBB
HBD
HBB
b
CCR5#2
CCR5
CCR2
c
FANCF
FANCF
d
VEGFA
VEGFA
ON
OMT
HMT
NBS
OMT/HMT



a
CCR5#1
On-Target
33 kb
46.360 kb
46.370 kb
46.380 kb
D1
D14
CCR2
CCR5
b
CCR5#2
On-Target
33 kb
46.360 kb
46.370 kb
46.380 kb
D1
D14
CCR2
CCR5
c
HBB
On-Target
33 kb
5.220 kb
5.230 kb
5.240 kb
D1
D14
HBB
HBD
HBBP1
d
CCR5#1
Hits
Cas9-D1
Cas9-D4
Cas9-D14
HiFi-D1
HiFi-D4
HiFi-D14
Chr./ Loci
CCR2
1
13
19
22
3
3.1
HMT/OMT
OMT
e
CCR5#2
Hits
Cas9-D1
Cas9-D4
Cas9-D14
HiFi-D1
HiFi-D4
HiFi-D14
Chr./ Loci
CCR2
3.1
3.2
3.3
22
HMT/OMT
OMT
HMT
f
HBB
Hits
TALEN-D1
TALEN-D4
TALEN-D14
Chr./ Loci
HBD
11
4
7
HMT/OMT
OMT


a
5’ amplicon
Edge amplicon
3’ amplicon
WT
Open ends/ translocated
Large deletion
Cleavage site
b
CCR5#1
CNV (%)
CCR5-E
CCR5-5'
CCR5-3'
GADL1 (telo.)
MYLK (q arm)
RAD1 (cto.)
STAT3 (cto.)
D1 D4 D10 D14
c
CCR5#2
d
HBB
HBB-E
HBB-5'
HBB-3'
CARS (telo.)
PODL3 (q arm)
e
f
g
Allelic frequency (%)
Other aberrations
Large deletion
Indel
wt
T7E1
Norm.

Supplemental Information

Quantitative evaluation of chromosomal rearrangements in primary gene-edited human stem cells by preclinical CAST-Seq

Giandomenico Turchiano[1,2,#,*], Geoffroy Andrieux[3,4], Georges Blattner[1,2,#], Valentina Pennucci[1,2], Julia Klermund[1,2], Gianni Monaco[1,2,§], Sushmita Poddar[1,2,$], Claudio Mussolino[1,2], Tatjana I. Cornu[1,2], Melanie Boerries[3,4,5,6], Toni Cathomen[1,2,6,*]

*To whom correspondence should be addressed:
Toni Cathomen, toni.cathomen@uniklinik-freiburg.de
Giandomenico Turchiano, g.turchiano@ucl.ac.uk

This PDF file includes:

- Materials and Methods
- Figures S1 to S8
- Tables S1 to S7

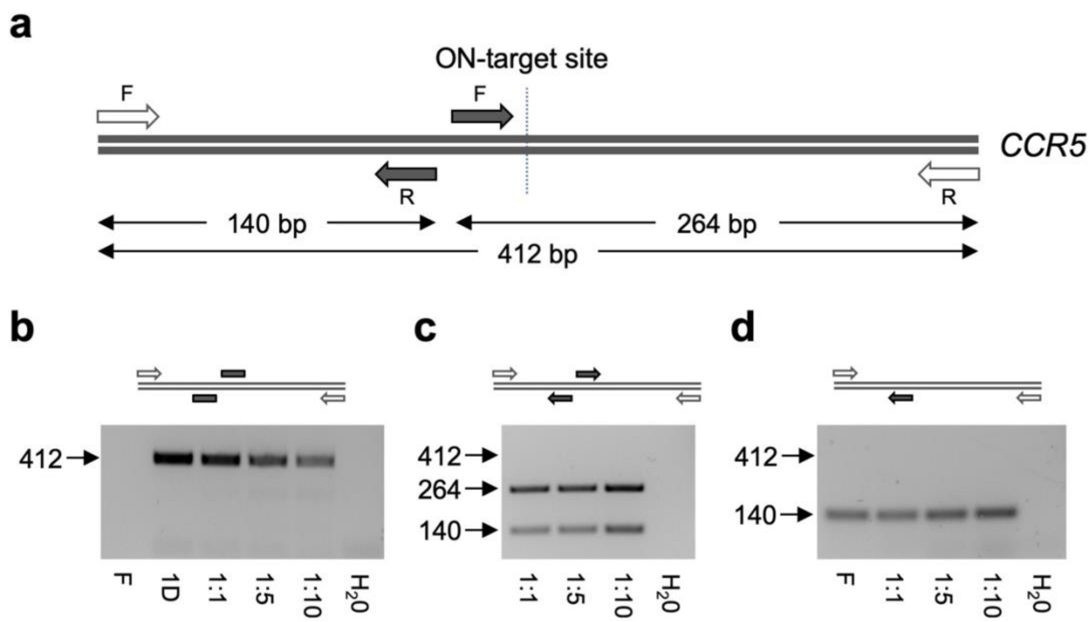


Figure S1. Effect of decoy primers. (**a**) Schematic of decoy primer test system. Effect of decoy primers (filled arrows) was tested on the *CCR5*[#1] locus using two locus-specific primers (open arrows) that amplify a 412 bp fragment. The expected amplicon lengths are indicated. F, forward primer; R, reverse primer. (**b**) Effect of blocked decoy primers. PCR was performed with *CCR5* primers in combination with decoy primers that were blocked by 3' phosphorylation (filled bars). F, reaction with only CCR5 forward primer with blocked reverse decoy primer; 1D, only one of the two decoy primers was used; H20, no template in reaction. 1:1; 1:5 and 1:10 reflect the ratio of CCR5 primers to decoy primers. (**c**) Effect of non-blocked decoy primers. PCR was performed as above with non-blocked decoy primers. H20, no template in reaction. 1:1; 1:5 and 1:10 reflect the ratio of CCR5 primers to decoy primers. (**d**) Effect of single decoy primer. PCR was performed as above with a single non-blocked decoy primer. F, CCR5 forward primer in combination with reverse decoy primer. Amplicons lengths are indicated on the left, all primer sequences are indicated in Suppl. Table 3.

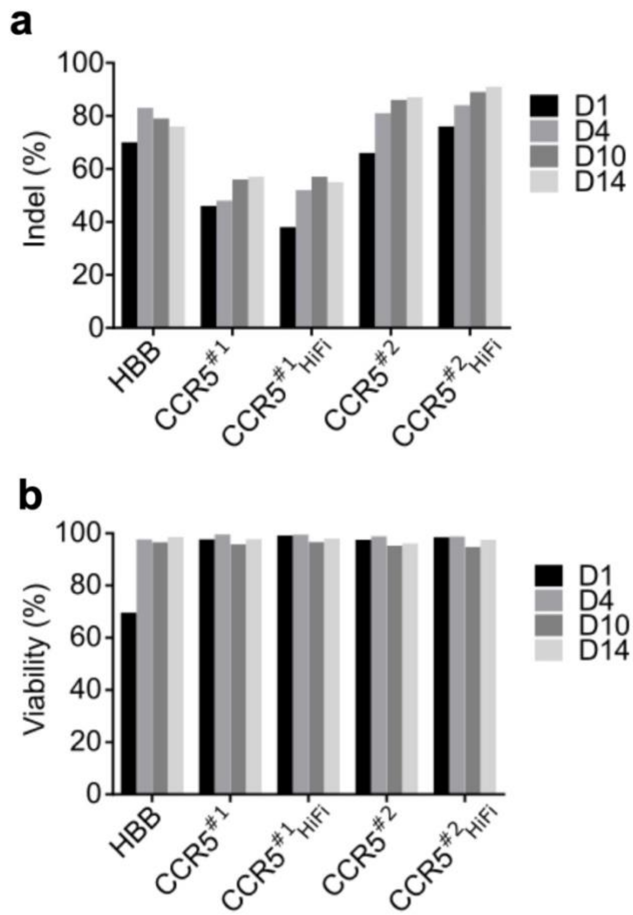


Figure S2. On-target activity and cell viability. (**a**) On-target activity. Indel frequencies were determined by T7E1 assay 1 to 14 days post-transfection. (**b**) Cell viability. Viability was examined 24 h post-electroporation by flow cytometry after staining cells with DAPI and acridine orange.

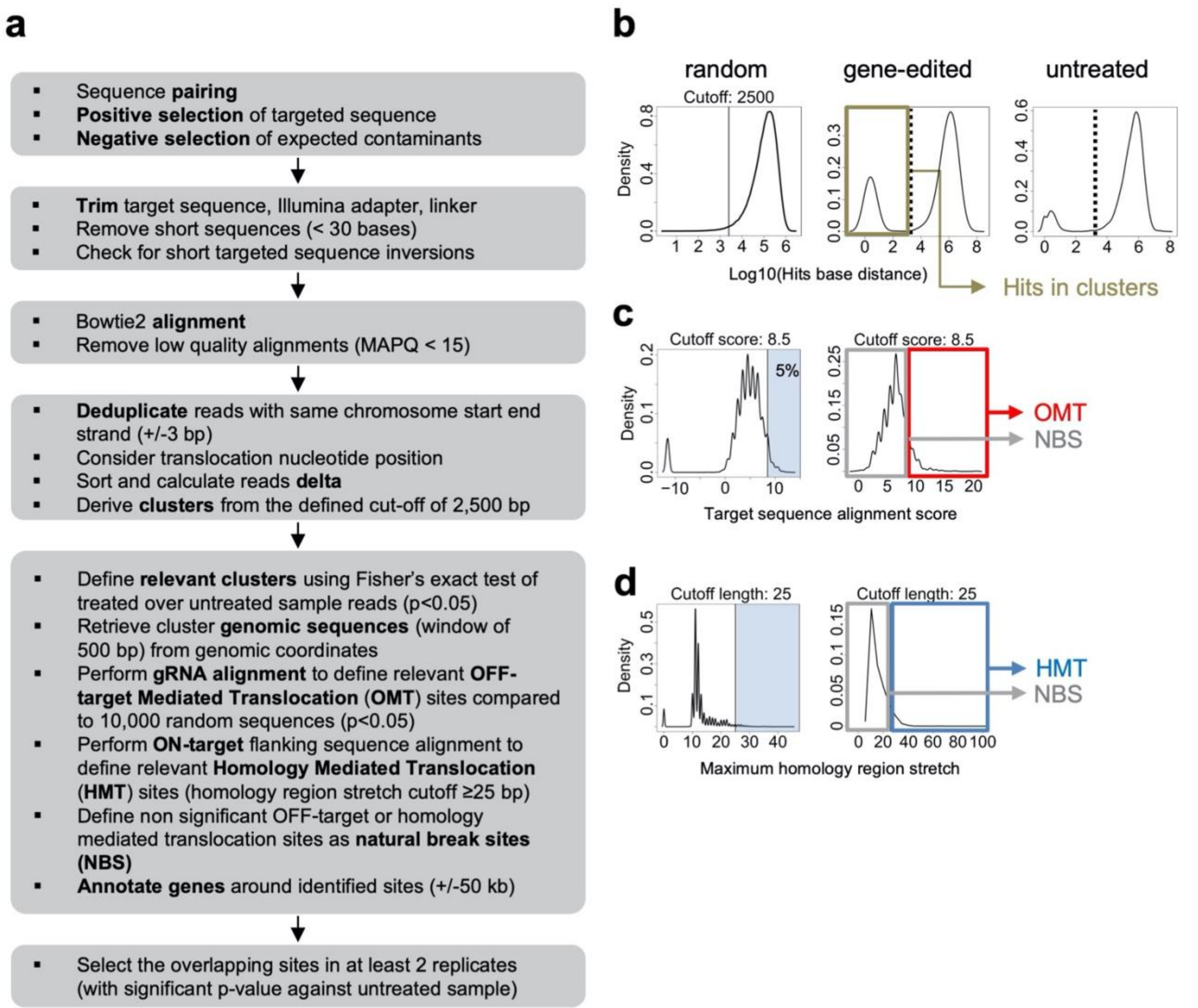
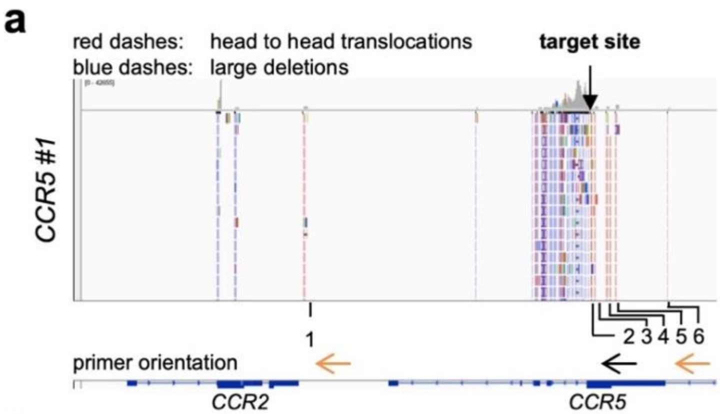
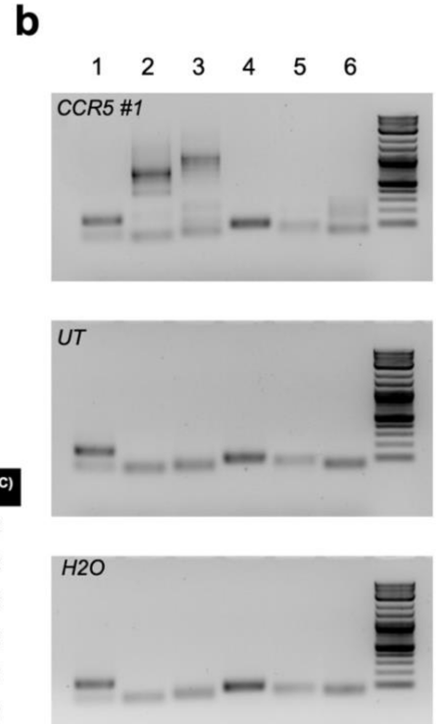


Figure S3. Bioinformatics pipeline. (**a**) Overview. FASTQ files derived from NGS were processed according to the overview. The boxes group the main steps in the bioinformatics flow: pairing and filtering, trimming, alignment, cluster definition, cluster analysis, filtering. (**b**) Read base distance. In order to calculate the likelihood of a read to fall into a cluster by chance rather than a designer nuclease provoked event, the CAST-Seq sample from gene edited cells was compared to an *in silico* created random read library that contains the same number of reads. The distribution of the distance of consecutive reads is shown on a logarithmic scale. In this example, the 2,500-bp threshold line describes an area of <5% in the random library, meaning that the likelihood of a read to fall into one cluster by chance is smaller than 5% (p<0.05). CAST-Seq analysis from untreated cells is shown as a control. (**c**) Target sequence alignment score. A 500-bp genomic region surrounding these translocation sites was compared against 10,000 random sequences of 500-bp. Every site was aligned to the designer nuclease target sequence. If the target sequence alignment score of the site was higher than the 5% best score in the random sequences, the event was classified as off-target mediated translocation (OMT). (**d**) Maximum homology region stretches. The longest common homologous substring between the target region and the translocation region was searched within a 5 kb window surrounding the translocation site. If the homologous substring length was longer than 24 bp, the event was classified as homology-mediated translocation (HMT). All other events were categorized as natural occurring breaking site (NBS)-derived translocation.





c

Site	Primer sequence (5'→3')	CCR5 primer sequence (5'→3')	T_A (°C)
1	CTCCCTGTGTCTTCTCATTC	AGGTAGATGTCAGTCATGCT	62
2	TTGGCAATGTGCTTTTGGAA	AGGTAGATGTCAGTCATGCT	62
3	CAATAACTTGATGCATGTGAAGG	AGGTAGATGTCAGTCATGCT	62
4	GCAATTTGTCTCAGGAAAAGG	AGGTAGATGTCAGTCATGCT	62
5	CCCAGAAGGGGACAGTAAGAAGG	GCAGGTAGATGTCAGTCATGCT	67
6	AGCAAACACAGCATGGACGAC	GCAGGTAGATGTCAGTCATGCT	67

Figure S4. Validation by PCR. (**a**) IGV visualization. CAST-Seq analysis reporting the locations of large deletions (blue ticks) and dicentric/large inversion mutations (red ticks). The validation primers were designed to bind to the negative strand (orange arrows), and PCR amplification product is expected only when an inversion occurred with the on-target site (black arrow). Position of the PCR primers is indicated by 1 to 6. (**b**) PCR analysis. 45-cycle PCR was performed with 50 ng of genomic DNA and reactions resolved on a 1% agarose gel. *CCR5*$^{\#1}$ gene edited sample was compared to untreated (UT) sample and a no-template (H_2O) control. (**c**) Primers. Primer sequences and their used annealing temperature (T_A°C) in the 6 reactions is indicated.

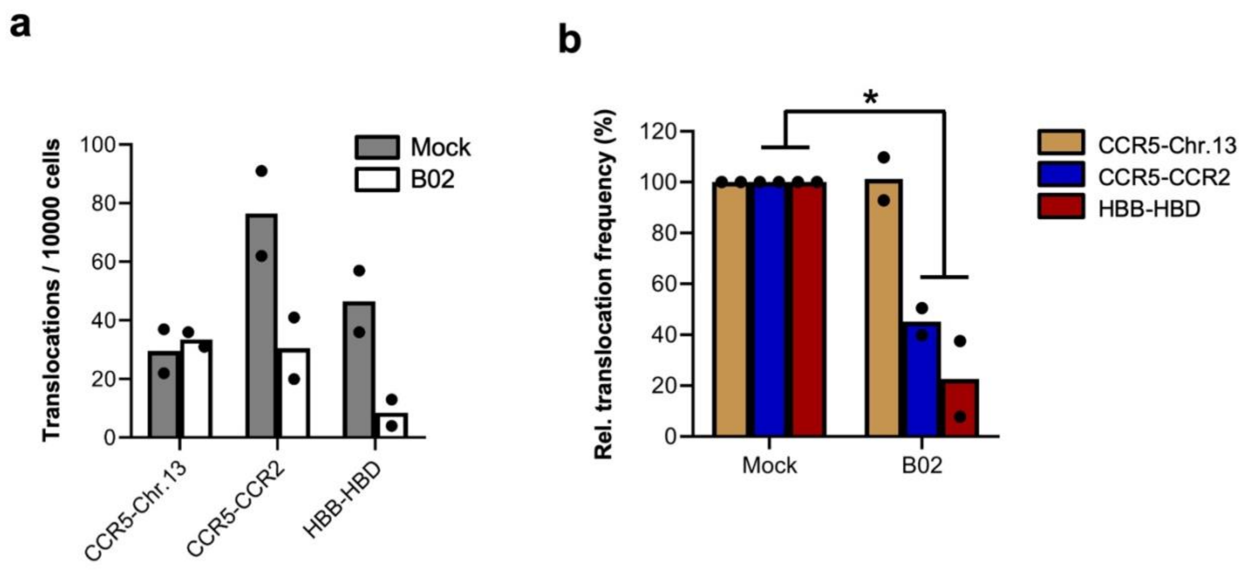


Figure S5. Effect of RAD51 inhibition. K562 cells were nucleofected with *CCR5*$^{\#1}$-targeting RNPs or with *HBB* TALEN-encoding mRNA in the presence or absence of B02, a well-characterized inhibitor of RAD51 [33, 34]. The number of translocation events was determined by ddPCR (n=2). The (**a**) absolute and (**b**) the relative number of translocations as compared to mock (DMSO)-treated samples is indicated. Paired t-test was performed on HMT events (*CCR5-CCR2* and *HBB-HBD*) comparing B02-treated vs. mock-treated samples. Statistically significant difference is indicated by * (p<0.05).

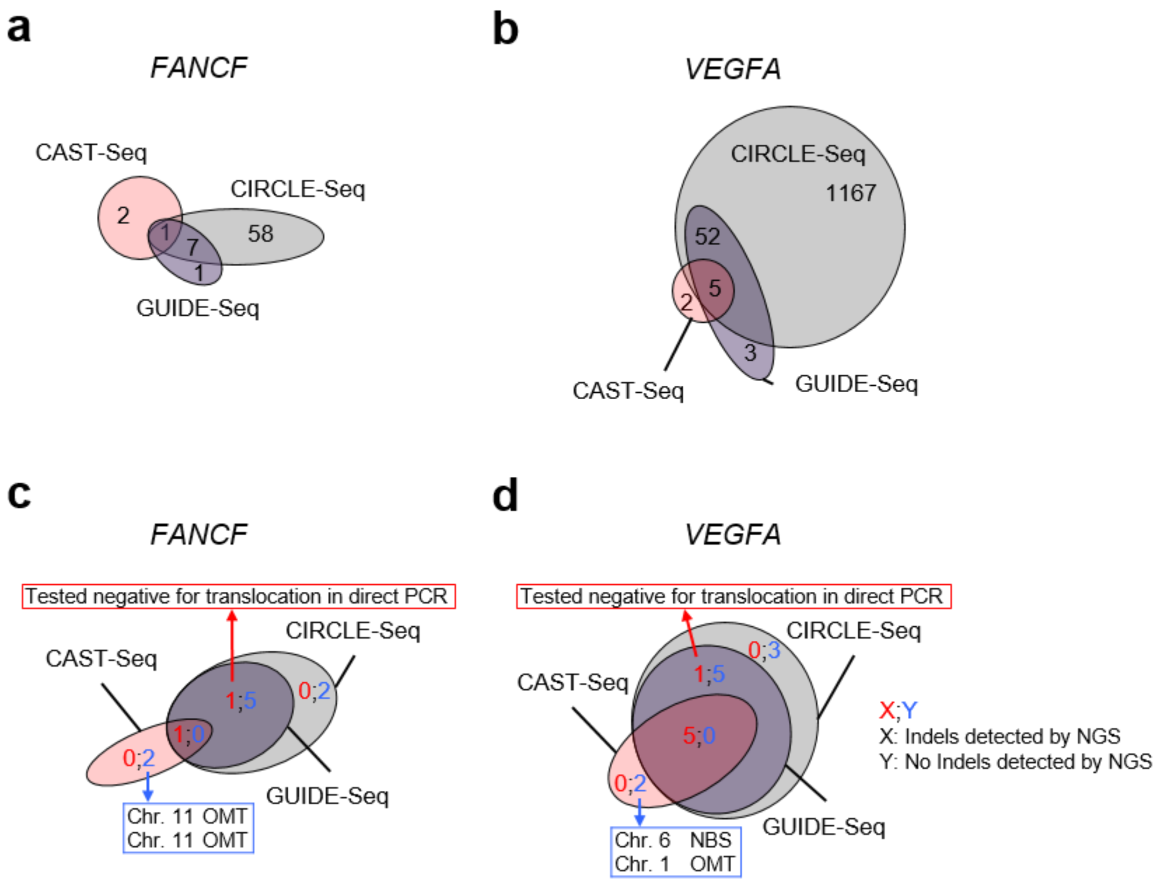


Figure S6. Comparison of CAST-Seq with GUIDE-Seq and CIRCLE-Seq in Venn diagrams. Data obtained from CAST-Seq analysis of CRISPR-Cas9 nucleases targeting *FANCF* (**a**) or *VEGFA* (**b**) were compared with published GUIDE-Seq [18] and CIRCLE-Seq [21] data. A subgroup of GUIDE-Seq and CIRCLE-Seq sites, namely the top 6 *FANCF* **(c)** and top 11 *VEGFA* **(d)** OTs that were tested for indels by NGS (see Suppl. Table 3), were compared with the according CAST-Seq data. The two sites that were positive for indels but not detected by CAST-Seq, were checked for translocations in a direct PCR on edited genomic DNA with a pair of primers binding the on-target and the supposed translocation site. The result of this PCR is indicated in the red box. The four sites that were negative for indels but identified by CAST-Seq are specified in the blue box.

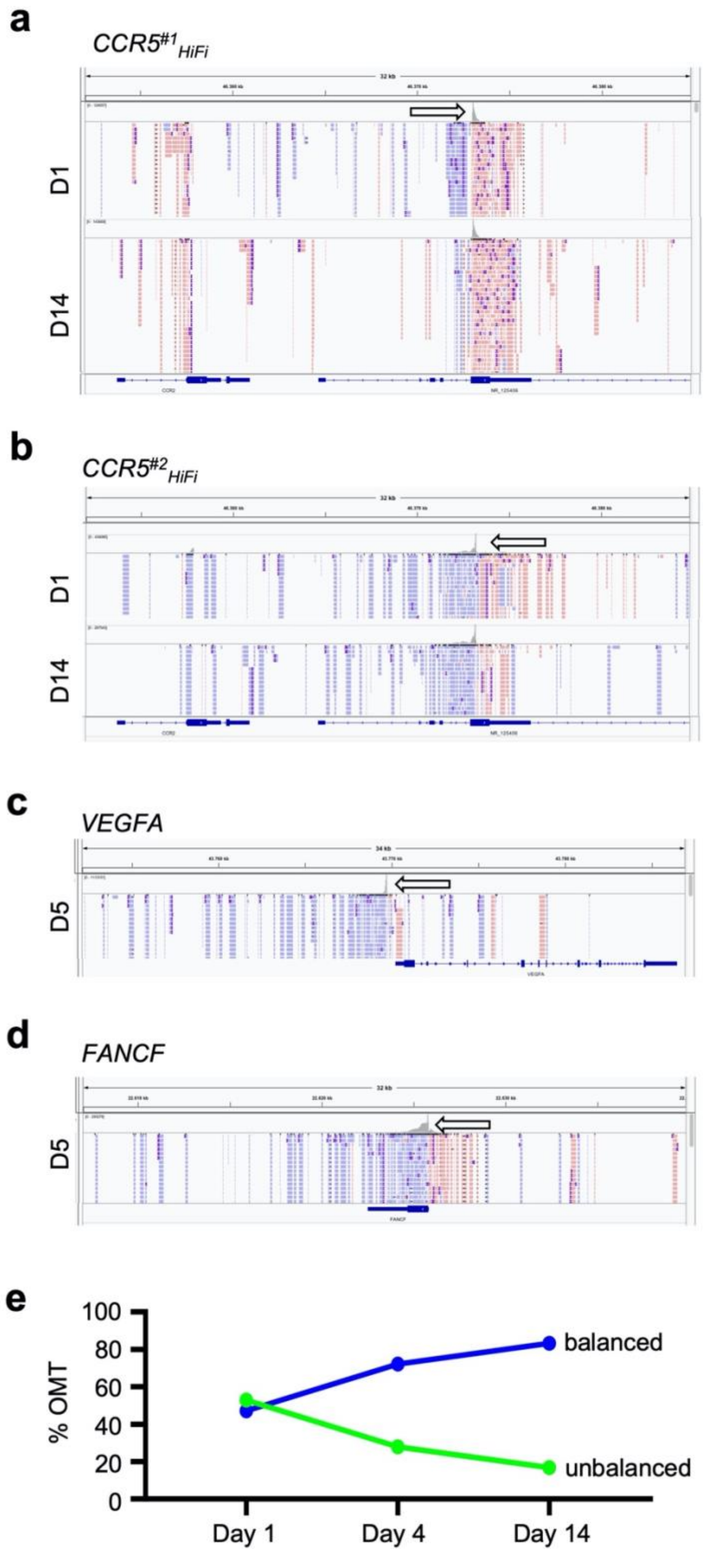


Figure S7. Qualitative visualization of CAST-Seq analysis. **(a-d)** IGV plots showing target regions, *CCR5*$^{\#1}$ (a), *CCR5*$^{\#2}$ (b), VEGFA (c), and *FANCF* (d), within a window of 32–34 kb. Only top rows are shown. Bait orientation (arrow) and harvesting times (D1, D5, D14) are indicated. Mapped CAST-Seq reads are represented by bars. Blue and red bars indicate reads aligned with negative or positive strand, respectively. Gene locus is revealed on the bottom. **(e)** Loss of cells with unbalanced translocations. Shown is an evaluation of *CCR5#1* nuclease derived OMTs with regard to balanced vs. unbalanced translocations.

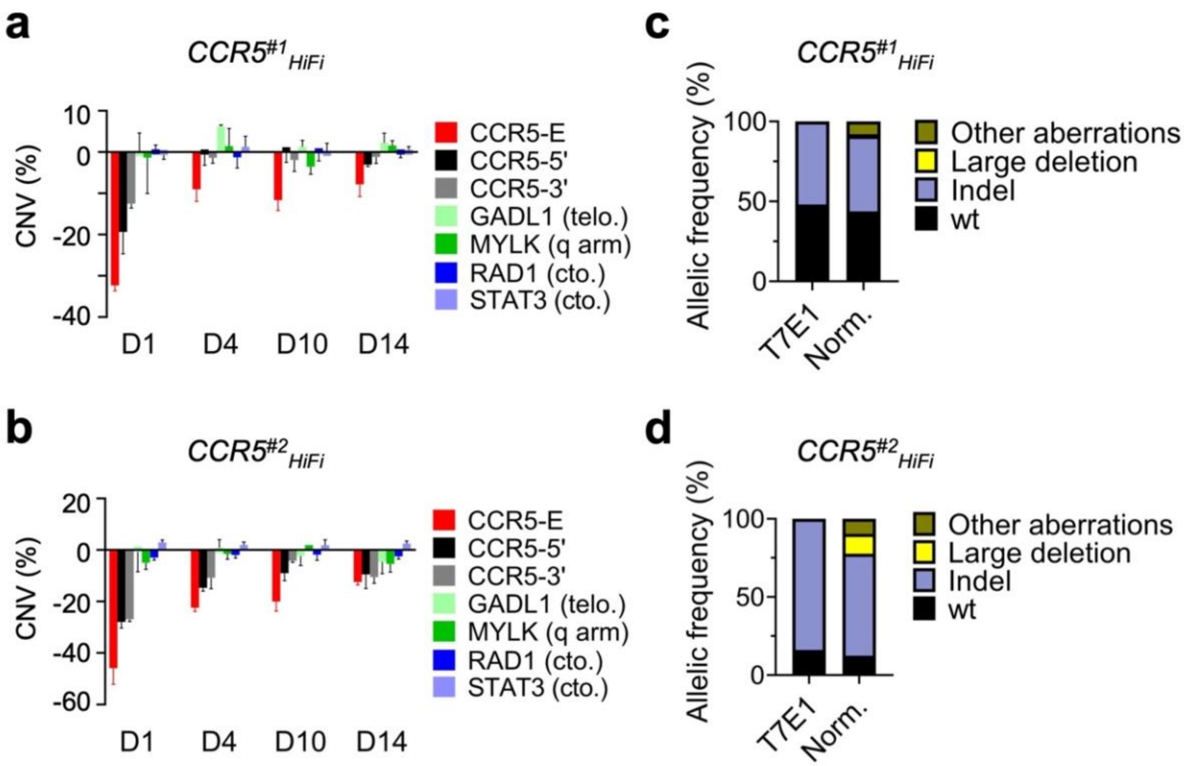


Figure S8. Quantification of chromosomal aberrations at target site. (**a, b**) Variation of target site copy numbers. Plots show relative copy number variation (CNV) of amplifiable target sites in CD34+ cells edited with CRISPR-Cas$_{\mathrm{HiFi}}$ targeting *CCR5#1* (b) or *CCR5#2* (c), at different time points (day 1 to day 14) after transfection. (**c, d**) Data summary. ddPCR results were used to normalize (Norm.) the indel frequencies determined by T7E1 assay for D4 time points. 'Large deletion' denotes the relative decrease of the average number of flanking amplicons while 'Other aberrations' is specified as the relative difference between the number of edge amplicons and the average number of flanking amplicons.

Supplementary Tables

Target	OFF-target reads (average fold change)*	stdev
VEGFA	5,1	±0.4
FANCF	5,0	±0.2

**comparing number of OFF-target reads between samples with or without decoy primers*

Table S1. **Impact of decoy primers.** To assess the impact of the decoy primers on the signal-to-noise ratio of CAST-Seq, side-by-side analyses were performed in the presence or absence of decoy primers in a biological duplicate. Data is based on all reads in clusters identified by CAST-Seq performed on genomic DNA isolated from CD34+ HSPCs that were edited with CRISPR-Cas9 nucleases targeting either *VEGFA* or *FANCF*. The fold change was calculated using the formula:

$$\frac{(total\ reads\ in\ cluster\ \text{‘}decoy\text{’} - reads\ in\ ON\ target\ cluster\ \text{‘}decoy\text{’}) \div total\ reads\ decoy}{(total\ reads\ in\ cluster - reads\ in\ ON\ target\ cluster) \div total\ reads}$$

SEE EXCEL Table

Table S2. CAST-Seq sites. Listed are all sites identified by CAST-Seq in CD34+ HSPCs edited with the mentioned TALEN or CRISPR-Cas9 nuclease (wildtype or HiFi): chromosomal locations (Chr., Start, End); total number of reads and hits; D1-D14 Hits; CAST-Seq category; cluster location with respect to gene annotations; distance to the closest transcriptional start site; closest transcribed element; transcripts within a window of 100 kb (oncogenes highlighted in red); putative target sequence; NGS data indicating percent of indels and number of reads in treated or untreated samples; statistical significance calculated with Z-test.

				NGS data					comparison of CAST-Seq vs. GUIDE-Seq and CIRCLE-Seq							
Sample	Chr.	Start	End	sample indel%	sample reads	control indel%	control reads	signifi-cance	platform	CAST-Seq rank	CAST-Seq category	CAST-Seq hits	GUIDE-Seq rank	Guide-Seq reads	CIRCLE-Seq rank	CIRCLE-Seq reads
FANCF	3	35071167	35072189	0,01	9940	0,05	6003		CIRCLE			/			6	140
FANCF	6	143060942	143060965	N/A	N/A	N/A	N/A	/	GUIDE			/	6	101		
FANCF	10	37664255	37664277	0,28	20061	0,17	7605		GUIDE/CIRCLE			/	8	77	3	218
FANCF	10	42914565	42914588	0,01	59248	0,02	50745		GUIDE/CIRCLE			/	3	524	2	298
FANCF	10	71703362	71703384	0,02	48613	0,01	40467		GUIDE/CIRCLE			/	7	78	4	198
FANCF	11	22566592	22569448	0,03	17224	0,03	22992		CAST	3	OMT	16				
FANCF	11	22597178	22635739	74,20	142577	0,40	80274	****	GUIDE/CIRCLE/CAST	1	ON	15498	1	4816	1	382
FANCF	11	22638629	22640545	0,14	16666	0,16	45669		CAST	2	OMT	20				
FANCF	12	117055848	117055871	0,13	36807	0,21	12019		CIRCLE			/			5	168
FANCF	17	80950160	80950183	0,08	49962	0,10	39915		GUIDE/CIRCLE			/	4	150	29	20
FANCF	18	8707523	8707546	0,29	52593	0,06	33969	****	GUIDE/CIRCLE			/	2	2099	12	70
FANCF	X	87100159	87100182	0,09	23665	0,14	28529		GUIDE/CIRCLE			/	5	125	15	66
VEGFA	1	48227172	48227470	0,12	84888	0,13	76732		CAST	7	OMT	2				
VEGFA	2	10233330	10233352	0,06	31713	0,06	41585		CIRCLE			/			9–11	218
VEGFA	3	194276088	194276111	1,88	47173	1,72	35774		GUIDE/CIRCLE			/	8	1315	100–104	104
VEGFA	5	11312195	11312217	N/A	N/A	N/A	N/A	/	CIRCLE			/			3	326
VEGFA	5	90145132	90145163	2,95	12307	0,03	14256	****	GUIDE/CIRCLE/CAST	4	OMT	27	2	2559	9–11	218
VEGFA	5	116098968	116098980	6,65	1022	0,76	1179	****	GUIDE/CIRCLE/CAST	5	OMT	14	4	2200	82–87	112
VEGFA	6	43741428	43743269	0,12	40686	0,11	31599		CAST	6	NBS	12				
VEGFA	6	43745782	43785917	67,03	29075	0,25	34963	****	GUIDE/CIRCLE/CAST	1	ON	2444	3	2440	24–25	176
VEGFA	7	2880012	2880034	0,19	51898	0,17	54580		CIRCLE			/			5	256
VEGFA	8	142809394	142809416	0,67	24062	0,57	32875		CIRCLE			/			8	226
VEGFA	8	143124420	143124442	N/A	N/A	N/A	N/A	/	CIRCLE			/			7	236
VEGFA	10	97000824	97000847	0,01	51586	0,02	34652		GUIDE/CIRCLE			/	7	1437	36	150
VEGFA	11	69083657	69083680	0,02	58363	0,02	38398		GUIDE/CIRCLE			/	6	1535	180–184	78
VEGFA	11	115887381	115887403	N/A	N/A	N/A	N/A	/	CIRCLE			/			2	346
VEGFA	14	61612048	61612071	1,03	55663	1,01	27523		GUIDE/CIRCLE			/	9	1170	6	250
VEGFA	14	64765817	64765839	N/A	N/A	N/A	N/A	/	CIRCLE			/			9–11	218
VEGFA	14	65102424	65102466	7,70	111899	0,87	131081	****	GUIDE/CIRCLE/CAST	2	OMT	63	1	3125	4	314
VEGFA	14	73886777	73886799	0,27	50519	0,02	22905	****	GUIDE/CIRCLE			/	11	790	1	352
VEGFA	19	40055953	40055976	0,18	20641	0,12	23961		GUIDE/CIRCLE			/	10	796	15	206
VEGFA	22	37266767	37266791	1,70	64635	0,24	62021	****	GUIDE/CIRCLE/CAST	3	OMT	35	5	1997	68–70	120

Table S3. Comparison of CAST-Seq, GUIDE-Seq and CIRCLE-Seq. Considered were all sites identified by CAST-Seq in CD34+ HSPCs edited with CRISPR-Cas9 targeting *FANCF* or *VEGFA*, as well as top 6 *FANCF* and top 11 *VEGFA* OTs predicted by GUIDE-Seq in edited K562 cells and CIRCLE-Seq. Listed are: chromosomal locations (Chr., Start, End); NGS data (percent of indels and number of reads from treated and untreated samples as well as relative statistical significance calculated with the Z-test); comparison (CAST-Seq rank, CAST-seq Category, total number of CAST-Seq hits in cluster; GUIDE-Seq rank, number of GUIDE-Seq reads, CIRCLE-Seq rank, number of CIRCLE-Seq reads).

Target	Function		ID	Sequence 5'-3'
CCR5#1	Target sequence		CCR5 #1	GTGAGTAGAGCGGAGGCAGGAGG
CCR5#2			CCR5 #2	CAATGTGTCAACTCTTGACAGGG
HBB			HBB TALEN L	TGATAGGCACTGACTCTCT
			HBB TALEN R	TAAGGGTGGGAAAATAGAC
VEGFA (site 3)			VEGFA site 3	GGTGAGTGAGTGTGTGCGTGTGG
FANCF			FANCF	GGAATCCCTTCTGCAGCACCTGG
CCR5#1	T7E1	For	2067	GTGGACAGGGAAGCTAGC
		Rev	3911	GCAGGTAGATGTCAGTCATGCT
CCR5#2	T7E1	For	3779	CTGGTCATCCTCATCCTG
		Rev	3780	AGACCTTCTTTTTGAGATCTGG
HBB	T7E1	For	3517	TGAGGAGAAGTCTGCCGTTAC
		Rev	3518	CAGCTCACTCAGTGTGGC
VEGFA	T7E1	For	3765	TCCAGATGGCACATTGTCAG
		Rev	3756	CGAGGAGGGAGCAGGAAAGT
FANCF	T7E1	For	4076	GGGCCGGGAAAGAGTTGCTG
		Rev	4077	GCCCTACATCTGCTCTCCCTCC
CCR5#1	CAST-seq PCR I	bait	4034	AGGTAGATGTCAGTCATGCT
		decoy for	4036	ATCAATGTGAAGCAAATCGCA
		decoy rev	4037	AGGGCTCCGATGTATAATAATTG
	CAST-seq PCR II	bait nested	4035	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCTCTTCAGCCTTTTGCAGTTTATCAG
CCR5#1 For (telomeric side)	CAST-seq PCR I	bait	4272	GGATTATCAAGTGTCAAGTCC
		decoy for	3779	CTGGTCATCCTCATCCTG
		decoy rev	4261	AAAACCAAAGATGAACACCAGT
	CAST-seq PCR II	bait nested	4262	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTATACATCGGAGCCCTGCCA
CCR5#2	CAST-seq PCR I	bait	4284	AAACACAGCATGGACGAC
		decoy for	4285	CCAGTGGGACTTTGGAAATAC
		decoy rev	4286	GCATAGTGAGCCCAGAAG
	CAST-seq PCR II	bait nested	4288	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGGAGGATGATGAAGAAGATTCCAGAG
HBB	CAST-seq PCR I	bait	4396	GTTGGTATCAAGGTTACAAGAC
		decoy for	4395	CTGCTGGTGGTCTACC
	CAST-seq PCR II	bait nested	4397	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGACCAATAGAAACTGGGCATGTGG
VEGFA	CAST-seq PCR I	bait	4382	GAGAGGGACACACAGATC
		decoy for	4380	CGTCTTCGAGAGTGAGGAC
		decoy rev	4381	CTGCTCGCTCCATTCAC
	CAST-seq PCR II	bait nested	4383	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTACACAGATCTATTGGAATCCTGGAGTG
FANCF	CAST-seq PCR I	bait	4362	GTTCCAATCAGTACGCAG
		decoy for	4360	CTTGAGACCGCCAGAAG
		decoy rev	4361	CACTACCTACGTCAGCAC
	CAST-seq PCR II	bait nested	4363	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCCGTCTCCAAGGTGAAAGC
Linker prey primers	CAST-seq PCR I	initial PCR	4032	GTAATACGACTCACTATAGGGC
	CAST-seq PCR II	nested PCR	4033	ACACTCTACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGGCTCCGCTTAAGGGAC
Linker oligo		positive strand	4038	GTAATACGACTCACTATAGGGCTCCGCTTAAGGGACT
		negative strand	4039	P-GTCCCTTAAGCGGAGC-NH3
CCR5#1	ddPCR CCR5/chr13	For	1 CCR5g3C13	CTGATGTGTGGCAGTTTGGGAC
		Rev	3911	GCAGGTAGATGTCAGTCATGCT
		For	4318	ACAGATTTTCCACTGCGTGG

	ddPCR CCR5/chr19	Rev	3911	GCAGGTAGATGTCAGTCATGCT
	ddPCR CCR5/chr22	For	4328	CATCACCTGAGTCATAGGGAAG
		Rev	3911	GCAGGTAGATGTCAGTCATGCT
	ddPCR CCR5/chr1	For	#91_G3OT_chr1_31944151_F	GAGGTTTCAAGCCCCATGTC
		Rev	3911	GCAGGTAGATGTCAGTCATGCT
	ddPCR CCR5/CCR2	For	4325	ATCCACAACATGCTGTCCAC
		Rev	3911	GCAGGTAGATGTCAGTCATGCT
	ddPCR STAT3	For	2851	ACTCTCACGGACGAGGAGC
		Rev	2852	CAGTTTTCTAGCCGATCTAGGCAG
CCR5#1	NGS Chr 13_24886251	For	2 CCR5g3C13	CCCACCAACAACAAAGTGAGGTGA
		Rev	1 CCR5g3C13	CTGATGTGTGGCAGTTTGGGAC
	NGS Chr 22_29074056	For	4327	CCGCTACAAGAGGCTATACG
		Rev	4328	CATCACCTGAGTCATAGGGAAG
	NGS Chr 1_31943321	For	#91_G3OT_chr1_31944151_F	GAGGTTTCAAGCCCCATGTC
		Rev	#92_G3OT_chr1_31944151_R	CCCGGAATTCACAGCTTCAC
	NGS Chr 19_35351318	For	4317	TGTACTTACGGGAAGGAGGAG
		Rev	4318	ACAGATTTTCCACTGCGTGG
	NGS Chr 3_46300625	For	#53_G3OT_chr3_46300375_F	CCTGTGTCAGGGTGGATTAG
		Rev	#54_G3OT_chr3_46300375_R	GAACAAGTATCAAAAGCAAGCCAG
	NGS Chr 3_46326033	For	#5_G3OT_chr3_46325980_F	GGGGCTCTATTAGGTTGTCATATAC
		Rev	#6_G3OT_chr3_46325980_R	CTGCTCTCACTAGATCCCTG
	NGS Chr 3_46331320	For	#93_G3WT_chr3_46331320_F	GGTGAGGAGACTGAAGGAAC
		Rev	#94_G3WT_chr3_46331320_R	GGCTGATGAGTACCACCAC
	NGS Chr 3_46339180	For	#7_G3OT_chr3_46339882_F	AGAACAGCAAGGGAGAGGTC
		Rev	#8_G3OT_chr3_46339882_R	CAATTGCAAATTGTGCATTTTTTGCAG
	NGS Chr 3_46347637	For	#23_G3OT_chr3_46348363_F	GTGAAGCCGTCTGGTTCTTAAC
		Rev	#24_G3OT_chr3_46348363_R	GTGTGGAGGACAACTCCTTTG
	NGS Chr 3_46352118	For	4235	ATCCACAACATGCTGTCCAC
		Rev	4236	GCACATTGCATTCCCAAAGAC
	NGS Chr 3_46360211	For	2813	GCAGCAAACCTTCCCTTCACTAC
		Rev	4280	TGCTCTTCAGCCTTTTGCAGTTTATCAG
	NGS Chr 3_46382097	For	#3_G3OT_chr3_46382675_F	CGACCACACTCCCATTTCTTG
		Rev	#4_G3OT_chr3_46382675_R	CCCCACCTTTTCCTGTAGAAC
	NGS Chr 3_138188178	For	#31_G3OT_chr3_138187958_F	CAAGTCTGTGCGGCTTCTATC
		Rev	#32_G3OT_chr3_138187958_R	CAGTAACTTTCATTCCTGGTCCTG
CCR5#1 HiFi	NGS Chr 1_31944401	For	#91_G3OT_chr1_31944151_F	GAGGTTTCAAGCCCCATGTC
		Rev	#92_G3OT_chr1_31944151_R	CCCGGAATTCACAGCTTCAC
	NGS Chr 19_35352351	For	4317	TGTACTTACGGGAAGGAGGAG
		Rev	4318	ACAGATTTTCCACTGCGTGG
	NGS Chr 3_46332407	For	#95_G3HIFI_chr3_46332407_F	CCTTCCCTCAGTGCCAATATC
		Rev	#96_G3HIFI_chr3_46332407_R	GTCACTGAAAGCTCCAAGCTC
	NGS Chr 3_46352334	For	4325	ATCCACAACATGCTGTCCAC
		Rev	4326	GCACATTGCATTCCCAAAGAC
	NGS Chr 3_46359979	For	2813	GCAGCAAACCTTCCCTTCACTAC
		Rev	4280	TGCTCTTCAGCCTTTTGCAGTTTATCAG
	NGS Chr 3_46379509	For	#3_G3OT_chr3_46382675_F	CGACCACACTCCCATTTCTTG
		Rev	#4_G3OT_chr3_46382675_R	CCCCACCTTTTCCTGTAGAAC
	NGS Chr 3_46395111	For	#97_G3HIFI_chr3_46395111_F	AGCCCTAAAGAACAGTGAGGAG
		Rev	#98_G3HIFI_chr3_46395111_R	CTATCTGGTAAACCAGGACCTTC
CCR5#2	NGS Chr 3_46344665	For	#23_399WT_chr3_46344665_F	CCCCACTGCTTATAGGCTG
		Rev	#24_399WT_chr3_46344665_R	CTTTTGTGTTCCAAGGTGTTAGTC
	NGS Chr 3_46350403	For	#13_399OT_chr3_CCR2_46357840_F	GCTGGTCGTCCTCATCTTAATAAAC
		Rev	#14_399OT_chr3_CCR2_46357840_R	GGGCCACAGACATAAACAGAATC
	NGS Chr 3_46361830	For	399_FwdOn_chr3_46337599	CTGGTCATCCTCATCCTGATAAAC
		Rev	399_RevOn_chr3_46337599	TGACTGTATGGAAAATGAGAGCTG
	NGS Chr 3_46393754	For	#3_399OT_chr3_46393504_F	GGGAGAGATTAGCCTTTGGTG
		Rev	#4_399OT_chr3_46393504_R	CCGCTTAGCTATGTGGACAAG
	NGS Chr 3_46416360	For	#5_399OT_chr3_46416110_F	TACCCATCCCACAGTGCTATTAC
		Rev	#6_399OT_chr3_46416110_R	TTCCAACTGTAGGTTAACATGCTC
CCR5#2 HiFi	NGS Chr 22_21876484	For	#19_399HIFI_chr22_21876484_F	CCAGCATTGACTCCTCCTTC
		Rev	#20_399HIFI_chr22_21876484_R	TGGGTCACATGGTTCTCTTG
	NGS Chr 3_46315040	For	#15_399HIFI_chr3_46311487_F	CCTGGTGGCTTGCTACTATTC
		Rev	#16_399HIFI_chr3_46311487_R	AAGCAGTACAGAGACAGCTATG
	NGS Chr 3_46337559	For	#13_399OT_chr3_CCR2_46357840_F	GCTGGTCGTCCTCATCTTAATAAAC
		Rev	#14_399OT_chr3_CCR2_46357840_R	GGGCCACAGACATAAACAGAATC
	NGS Chr 3_46361075	For	399_FwdOn_chr3_46337599	CTGGTCATCCTCATCCTGATAAAC
		Rev	399_RevOn_chr3_46337599	TGACTGTATGGAAAATGAGAGCTG
	NGS Chr 3_46416360	For	#17_399HIFI_chr3_46416360_F	GTGGTGACATGTATTATGCTTACAC
		Rev	#18_399HIFI_chr3_46416360_R	GCCTACTCTGTCATTCCAACTG

FANCF	NGS Chr 18_8707523	For	chr18_8707523_FANCF_11	CCAGTCCTTTGTAAGCATCCAG
		Rev	chr18_8707523_FANCF_12	ACAGCTCAAATCACATAACCCAC
	NGS Chr 11_22566592	For	#5_FANCF_chr11_22566592_F	GGCTGCTACTGGGAATGTTAAG
		Rev	#6_FANCF_chr11_22566592_R	CCTGCAGAATACTGTAGCTGAC
	NGS Chr 11_22597178	For	SP_FANCF_Fwd	GGGCCGGGAAAGAGTTGCTG
		Rev	SP_FANCF_Rev	GCCCTACATCTGCTCTCCCTCC
	NGS Chr 11_22638629	For	#3_FANCF_chr11_22638629_F	GGACACTATGCAACTGATGGAC
		Rev	#4_FANCF_chr11_22638629_R	AGACCTACCCTTATCCCTGAC
	NGS Chr 3_35071167	For	chr3_35071167_FANCF_23	CTTCAAACCCTGAAGCTGCAATC
		Rev	chr3_35071167_FANCF_24	AGTGCTGGGTAGTGAATGTAAATG
	NGS Chr 10_37664255	For	chr10_37664255_FANCF_25	GAAAGCTCCAGCTAGAACAAGATG
		Rev	chr10_37664255_FANCF_26	CCAGTGAGACCAGTTTGAGAC
	NGS Chr 10_42914565	For	chr10_42914565_FANCF_13	CCAAAGGAGAACTCTCATAGGTG
		Rev	chr10_42914565_FANCF_14	CCAGTGAGACCAGTTTGAGAC
	NGS Chr 10_71703362	For	chr10_71703362_FANCF_21	GGCTTCTTTGCCTCCTGTTC
		Rev	chr10_71703362_FANCF_22	TCAGGTATAAGCCCTCGTGAC
	NGS Chr 17_80950160	For	chr17_80950160_FANCF_15	GGGTACAGTTCTGCGTGTTG
		Rev	chr17_80950160_FANCF_16	GACAGGTGCTCAGACAGAAG
	NGS Chr X_87100159	For	chrX_87100159_FANCF_17	CCCTAGCCATGGAGCAATC
		Rev	chrX_87100159_FANCF_18	GGAACTAGAGCCTCGAGTAGTG
	NGS Chr 12_117055848	For	chr12_117055848_FANCF_27	TACTCTGCTATCAAACACTAGCAC
		Rev	chr12_117055848_FANCF_28	CTCTCCTTGCTACATGCTGTG
VEGFA	NGS Chr 7_2880012	For	chr7_2880012_VEGFA_53	GTGTGCATGTATCTGTGCATGAC
		Rev	chr7_2880012_VEGFA_54	CACTTGTGCAAATGCACTTGTC
	NGS Chr 2_10233330	For	chr2_10233330_VEGFA_49	GCAGTTGGTGCTGTGAAAG
		Rev	chr2_10233330_VEGFA_50	GGCTCAACAACCTGCTCAC
	NGS Chr 22_37266767	For	#3_VEGFA_chr22_37266767_F	CCTGGCCCATTTCTCCTTTG
		Rev	#4_VEGFA_chr22_37266767_R	CCAATACCCAGGTATCCGTG
	NGS Chr 19_40055953	For	chr19_40055953_VEGFA_47	CTCCCTACTGGGGACATTTTC
		Rev	chr19_40055953_VEGFA_48	GACGACCTAGCTGGGTAAG
	NGS Chr 6_43741428	For	#9_VEGFA_chr6_43741420_F	CCAGCTACCAGTTGTAAAAGGAC
		Rev	#10_VEGFA_chr6_43741420_R	GGGTCTGCATTTGAACCATAAAC
	NGS Chr 6_43745782	For	VEGFA_FwdOn_chr6_43745782_F	GAAGCAACTCCAGTCCCAAATATG
		Rev	VEGFA_RevOn_chr6_43745782_F	GGAGCAGGAAAGTGAGGTTAC
	NGS Chr 1_48227172	For	chr1_48226922_VEGFA_37	CCCTGCTGATCTTGTTGATGTC
		Rev	chr1_48226922_VEGFA_38	CGTGCACATACATTCGCAAAG
	NGS Chr 14_61612048	For	chr14_61612048_VEGFA_45	CCTCACTTAGTCTTCAGTAAGCAC
		Rev	chr14_61612048_VEGFA_46	TGCAGAAGCAGGAGATGTTTG
	NGS Chr 14_65102424	For	SP_Chr14_65102424_Fwd	GAGGGGGAAGTCACCGACAA
		Rev	SP_Chr14_65102424_Rev	TACCCGGGCCGTCTGTTAGA
	NGS Chr 11_69083657	For	chr11_69083657_VEGFA_43	CACCTCTAGCTCTGCATTTCTTTG
		Rev	chr11_69083657_VEGFA_44	GACCCTGACAGAAAGGCAAG
	NGS Chr 14_73886777	For	chr14_73886777_VEGFA_61	CGTCAACGAATTAGCTGACCTG
		Rev	chr14_73886777_VEGFA_62	GGGTACTACCTAACCGAGGAG
	NGS Chr 5_90145132	For	#5_VEGFA_chr5_90145132_F	ACCTAATTGATGCAGTTTGGCTC
		Rev	#6_VEGFA_chr5_90145132_R	CCTCATTTAGGCCCACAAAATTTC
	NGS Chr 10_97000824	For	chr10_97000824_VEGFA_41	GGCTGACAGTACTTCATGGTTG
		Rev	chr10_97000824_VEGFA_42	AGCAAATTGCGCCATAGCTG
	NGS Chr 5_116098968	For	#7_VEGFA_chr5_116098968_F	GCTAGATACTGAGGAAAGACTGTG
		Rev	#8_VEGFA_chr5_116098968_R	CTGGTCAGAGGGTACAACTTTTAG
	NGS Chr 8_142809394	For	chr8_142809394_VEGFA_57	GAGGATGCGAGTGTGGTG
		Rev	chr8_142809394_VEGFA_58	CCATCCCACTGGTGTCATC
	NGS Chr 3_194276088	For	chr3_194276088_VEGFA_39	CTGCCAGGAAAACAGAGGTC
		Rev	chr3_194276088_VEGFA_40	CCTTTCTAAGGCACGAGTCAG
TALEN HBB	NGS Chr 11_5203025	For	#19_TALENHBB_chr11_5203025_F	GTTGCCACCATAGAGACTATCAG
		Rev	#20_TALENHBB_chr11_5203025_R	CAACATTCCAGACAGTGCTCAG
	NGS Chr 11_5211158	For	3518	CAGCTCACTCAGTGTGGC
		Rev	3517	TGAGGAGAAGTCTGCCGTTAC
	NGS Chr 11_5231460	For	3520	AGTGCAGCTCACTCAGCT
		Rev	3519	TGAGGAGAAGACTGCTGTCAA
	NGS Chr 11_5242606	For	#31_TALENHBB_chr11_5242606_F	GTTCCCTCATCCAAAAACACTCAG
		Rev	#32_TALENHBB_chr11_5242606_R	GCTCACGGATGACCTCAAAG
	NGS Chr 11_5254108	For	#29_TALENHBB_chr11_5254108_F	GCGGCTAAAAGACCAGAAAGATAC
		Rev	#30_TALENHBB_chr11_5254108_R	GGGCTTAGACACCAGTCTC
	NGS Chr 11_5262322	For	#27_TALENHBB_chr11_5262322_F	CAAATGGCCATCAGCGATATAATG
		Rev	#28_TALENHBB_chr11_5262322_R	TCGGTCAGTTCAAGTAATTTTGTTG
	NGS Chr 7_8835280	For	#9_HBBOT_chr7_8835030_F	AGAAATTGAGCATAATGGTGGGAG
		Rev	#10_HBBOT_chr7_8835030_R	GCGATCCTGACTCACTGTAAC
		For	#5_HBBOT_chr4_124782108_F	TCAGCTATTCCTGGGTGATTAGAG

	NGS Chr 4_124782358	Rev	#6_HBBOT_chr4_124782108_R	CAAGACACCACTGATACATCCTG
	NGS Chr 1_158843625	For	#7_HBBOT_chr1_158843375_F	ACCAGGAAGAAGTGGGTCTTG
		Rev	#8_HBBOT_chr1_158843375_R	CACTGTGGTGTGATGAGAAGAG
CCR5#1	ddPCR-Edge	For	4281	TTATTATACATCGGAGCCCTGCCAA
		Rev	4280	TGCTCTTCAGCCTTTTGCAGTTTATCAG
	ddPCR-5'	For	4282	AGTTTGCATTCATGGAGGGCAAC
		Rev	4283	GGCAGGGCTCCGATGTATAATAATTG
	ddPCR-3'	For	4115	CATGCTGGTCATCCTCATCCTG
		Rev	4141	CCCAGAAGGGGACAGTAAGAAGG
CCR5#2	ddPCR-Edge	For	4072	TCCTTCTTACTGTCCCCTTCTGG
		Rev	4073	AGCAAACACAGCATGGACGAC
	ddPCR-5'	For	4115	CATGCTGGTCATCCTCATCCTG
		Rev	4141	CCCAGAAGGGGACAGTAAGAAGG
	ddPCR-3'	For	4142	ATCGATAGGTACCTGGCTGTCG
		Rev	4114	GTATGGAAAATGAGAGCTGCAGGTG
CCR5#1 & CCR5#2	ddPCR-GADL1 (Telomere)	For	4064	TGCCAAGGCATCTTACCTCTTCC
		Rev	4065	GCATCTGGTCTTCTGCTACACTGG
	ddPCR-MYLK (q arm)	For	4068	CAGCCTTGTGATTCATGCTGTCC
		Rev	4069	GGACTCACCTTCTACTGTCAACTCC
HBB	ddPCR-Edge	For	4478	AGACCAATAGAAACTGGGCATGTGG
		Rev	4479	ATCACTAAAGGCACCGAGCACT
	ddPCR-5'	For	4472	GGCTCATGGCAAGAAAGTGCTC
		Rev	4473	CAGTGCAGCTCACTCAGTGTG
	ddPCR-3'	For	4470	CTGAGGAGAAGTCTGCCGTTAC
		Rev	4471	CCACATGCCCAGTTTCTATTGGT
	ddPCR-CARS (Telomere)	For	4474	GGGCCAGGGAAGTGTATGATG
		Rev	4475	ACAGACATCAGTGCCATTGCG
	ddPCR-PODL1 (q arm)	For	4476	GCAGGTTCAGTCCCTCTTGG
		Rev	4477	TGCTTGGCCTATGGACAGTTG
Common Target	ddPCR-RAD1 (ctl.)	For	4143	CCTTCAGCTCTGTGGTGACG
		Rev	4144	CCCTTCTCAGCAAAGTCCCTG
	ddPCR-STAT3 (ctl.)	For	2851	ACTCTCACGGACGAGGAGC
		Rev	2852	CAGTTTTCTAGCCGATCTAGGCAG

Table S4. **Primer and linker sequences.** Listed are all deoxyoligonucleotides used to perform CAST-Seq, T7E1 assay, ddPCR or direct PCRs. CRISPR-Cas9 target sites are reported with PAM in bold, the split *HBB* TALEN binding sequence is indicated for both subunits in 5’-3’ orientation.

Software	Version	Usage
FLASH (https://ccb.jhu.edu/software/FLASH/)	v1.2.11	pairing reads
Bbmap (https://jgi.doe.gov/data-and-tools/bbtools/)	38.22	selection of designer nuclease target sites, linker and adapter trimming
Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)	2.3.4.2	Alignment to hg38 genome
samtools (http://samtools.sourceforge.net)	1.9	SAM to BAM conversion
bedtools (https://bedtools.readthedocs.io/en/latest/)	v2.27.1	BAM to Bed conversion, random sequences generation…

Table S5. **Software.** Listed are all software used for CAST-Seq.

	A	C	G	T	M	R	W	S	Y	K	V	H	D	B	N	indel
A	1	-1	-1	-1	0	0	0	-1	-1	-1	-0,3333	-0,3333	-0,3333	-1	-0,5	-1
C	-1	1	-1	-1	0	-1	-1	0	0	-1	-0,3333	-0,3333	-1	-0,3333	-0,5	-1
G	-1	-1	1	-1	-1	0	-1	0	-1	0	-0,3333	-1	-0,3333	-0,3333	-0,5	-1
T	-1	-1	-1	1	-1	-1	0	-1	0	0	-1	-0,3333	-0,3333	-0,3333	-0,5	-1
M	0	0	-1	-1	0	-0,5	-0,5	-0,5	-0,5	-1	-0,3333	-0,3333	-0,6667	-0,6667	-0,5	-1
R	0	-1	0	-1	-0,5	0	-0,5	-0,5	-1	-0,5	-0,3333	-0,6667	-0,3333	-0,6667	-0,5	-1
W	0	-1	-1	0	-0,5	-0,5	0	-1	-0,5	-0,5	-0,6667	-0,3333	-0,3333	-0,6667	-0,5	-1
S	-1	0	0	-1	-0,5	-0,5	-1	0	-0,5	-0,5	-0,3333	-0,6667	-0,6667	-0,3333	-0,5	-1
Y	-1	0	-1	0	-0,5	-1	-0,5	-0,5	0	-0,5	-0,6667	-0,3333	-0,6667	-0,3333	-0,5	-1
K	-1	-1	0	0	-1	-0,5	-0,5	-0,5	-0,5	0	-0,6667	-0,6667	-0,3333	-0,3333	-0,5	-1
V	-0,3333	-0,3333	-0,3333	-1	-0,3333	-0,3333	-0,6667	-0,3333	-0,6667	-0,6667	-0,3333	-0,5556	-0,5556	-0,5556	-0,5	-1
H	-0,3333	-0,3333	-1	-0,3333	-0,3333	-0,6667	-0,3333	-0,6667	-0,3333	-0,6667	-0,5556	-0,3333	-0,5556	-0,5556	-0,5	-1
D	-0,3333	-1	-0,3333	-0,3333	-0,6667	-0,3333	-0,3333	-0,6667	-0,6667	-0,3333	-0,5556	-0,5556	-0,3333	-0,5556	-0,5	-1
B	-1	-0,3333	-0,3333	-0,3333	-0,6667	-0,6667	-0,6667	-0,3333	-0,3333	-0,3333	-0,5556	-0,5556	-0,5556	-0,3333	-0,5	-1
N	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-0,5	-1
indel	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	na

IUPAC code

A	Adenine	M	A or C	V	A or C or G
C	Cytosine	R	A or G	H	A or C or T
G	Guanine	W	A or T	D	A or G or T
T (or U)	Thymine (or Uracil)	Y	C or T	B	C or G or T
		S	G or C		
		K	G or T	N	any base

Table S6. **Scoring matrix.** Scoring matrix of nucleotide substitution used for the alignment of translocation sites against the target site sequence, including weights for mismatch and bulges (insertions/deletions). IUPAC code is used. A, adenine; C, cytosine; G, guanine; T (or U), thymine (or uracil); R, A or G; Y, C or T; S, G or C; W, A or T; K, G or T; M, A or C; B, C or G or T; D, A or G or T; H, A or C or T; V, A or C or G; N, any base.

Software	Version	Location	Usage
BSgenome.Hsapiens.UCSC.hg38	1.4.1	http://bioconductor.org/packages/release/data/annotation/html/BSgenome.Hsapiens.UCSC.hg38.html	get sequence from genomic coordinates
Biostrings	2.46.0	https://bioconductor.org/packages/release/bioc/html/Biostrings.html	align sequence to guide-RNA
ChIPseeker	1.14.2	https://bioconductor.org/packages/release/bioc/html/ChIPseeker.html	gene annotation of translocation sites
TxDb.Hsapiens.UCSC.hg38.knownGene	3.2.2	https://bioconductor.org/packages/release/data/annotation/html/TxDbHsapiens.UCSC.hg38.knownGene.html	known gene coordinates and gene regions
org.Hs.eg.db	3.5.0	https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html	match gene symbol and entrez ID
biomaRt	2.34.2	https://bioconductor.org/packages/release/bioc/html/biomaRt.html	retrieve oncogene TSS

Table S7. **R packages.** Listed are the R packages used for CAST-Seq.