Nanopore Cas9-targeted sequencing enables accurate and simultaneous identification of transgene integration sites, their structure and epigenetic status in recombinant Chinese hamster ovary cells

Klaus Leitner[2, *]

Krishna Motheramgari[2]

Nicole Borth[1, 2]

Nicolas Marx[1, *]

[1]BOKU University of Natural Resources and Life Sciences, Vienna, Austria

[2]Austrian Center for Industrial Biotechnology GmbH, Vienna, Austria

*These authors contributed equally

Correspondence: Nicolas Marx

University of Natural Resources and Life Sciences, Vienna

Department of Biotechnology

Muthgasse 18, 1190 Vienna, Austria

E-mail: nicolas.marx@boku.ac.at

1 Abstract

The integration of a transgene expression construct into the host genome is the initial step for the generation of recombinant cell lines used for biopharmaceutical production. The stability and level of recombinant gene expression in Chinese hamster ovary (CHO) can be correlated to the copy number, its integration site as well as the epigenetic context of the transgene vector. Also, undesired integration events, such as concatemers, truncated and inverted vector repeats, are impacting the stability of recombinant cell lines. Thus, to characterize cell clones and to isolate the most promising candidates it is crucial to obtain information on the site of integration, the structure of integrated sequence and the epigenetic status. Current sequencing techniques allow to gather this information separately but do not offer a comprehensive and simultaneous resolution.

In this study, we present a fast and robust nanopore Cas9-targeted sequencing (nCats) pipeline to identify integration sites, the composition of the integrated sequence as well as its DNA methylation status in CHO cells that can be obtained simultaneously from the same sequencing run. A Cas9-enrichment step during library preparation enables targeted and directional nanopore sequencing with up to 724x median on-target coverage and up to 153 Kb long reads. The data generated by nCats provides sensitive, detailed and correct information on the transgene integration sites and the expression vector structure, which could only be partly produced by traditional Targeted Locus Amplification-Seq data. Moreover, with nCats the DNA methylation status can be analyzed from the same raw data without prior DNA amplification.

Keywords: Nanopore Sequencing, Chinese hamster ovary, CRISPR-Cas9, Integration Site Analysis, Concatemerization, Epigenetics.

2 Introduction

The inherent genetic plasticity of CHO cells allows to integrate transgenes relatively efficiently into the genome and to genetically engineer cell lines with enhanced production phenotypes, which provides the basis for their success as protein production cell lines. In order to generate high producing clones, various cell line development (CLD) platforms have been implemented, which can be divided into two approaches. Conventional random integration (RI) protocols rely on the cell's machinery to integrate the transgene at random breakage points in the genome during DNA double strand repair. Following a selection step and optionally subsequent gene amplification, extensive screening processes are applied to identify and characterize stable producer cell clones. Since the epigenetic environment around the inserted expression vectors (EV) largely regulates the functionality of the transcriptional machinery and thus transgene expression (Marx et al., 2022), (semi-)targeted integration (TI) has emerged as an alternative cell engineering approach to overcome issues related to RI. Instead of uncontrolled, random integration of EVs into the CHO genome, insertion mediated either by transposases or RNA-guided genome editing tools, such as CRISPR-Cas9, allows site-specific TI in highly transcribed regions or single genomic loci (Lee et al., 2015, 2016; Schmieder et al., 2022; Zhao et al., 2018). These technologies are designed to mitigate the impact of epigenetic repression and control the number of transgene copies, which should result in more stable expression of the transgene. Combination of TI and methods to exchange expression cassettes, such as Recombinase-Mediated-Cassette-Exchange (RMCE), promise the standardization of cell line generation by only exchanging the inserted transgene sequence, thus retaining the cell's expression behaviour (Baumann et al., 2017; Grav et al., 2018; Nguyen et al., 2019; Pristovšek et al., 2019; Sergeeva et al., 2020). For any CLD approach - random or targeted – the identification (RI) or verification (TI) of the transgene integration sites (IS) as well as vector integrity is a fundamental step to

genetically characterize the isolated cell clones. Additionally, it is of interest whether the vectors were integrated into so-called safe-harbor regions, i.e. stable genomic regions embedded in a favorable genetic and epigenetic environment (Bandyopadhyay et al., 2019; Dhiman et al., 2020), and to analyze the intactness of the epigenetic status before and after integration.

The identification of transgene integration sites, however, is challenging as usually only the plasmid sequence prior to transfection is known. Targeted-Locus-Amplification (TLA) coupled with Next Generation Sequencing (NGS) is currently the method of choice to identify ISs of transgenes in mammalian cells and is readily used in industry (Schmieder et al., 2022; Stadermann et al., 2022). The technology is based on crosslinking live cells and proximity ligation followed by Illumina short-read NGS (de Vree et al., 2014). However, no information about the epigenetic context of integration sites can be inferred from TLA-seq data. Additionally, the turnaround time is relatively long (currently several weeks until results are available for academic labs according to the service provider) and, due to the short reads generated by Illumina sequencing, it is difficult to assemble and confirm concatemers which have, however, a significant impact on stability (Dhiman et al., 2020).

Here we present an alternative fast and accurate method to identify and characterize transgene integration sites, precise vector structure and its integrity with simultaneous DNA methylation profiling in recombinant CHO cell lines by applying nanopore Cas9-targeted sequencing (nCats). Nanopore sequencing is a long-read sequencing technology that offers the possibility to sequence long, contiguous and native DNA molecules achieving read lengths above 2Mb (Payne et al., 2019), which allows to span repetitive regions and structural variants facilitating a comprehensive assembly of genomic regions (Chao et al., 2022; Nurk et al., 2022; Zimin et al., 2022). Additionally, native DNA sequencing by nanopore allows the identification of DNA modifications (e.g. CpG methylation), which overcomes existing issues of bisulfite sequencing

(BS-seq) (**Figure 1**A). Still, due to the relatively high base-calling error rates inherent to the nanopore sequencing technology, intensive sequencing is needed to accurately map the generated reads to a reference genome. While this remains a hurdle for whole-genome approaches, the nCats enrichment strategy – by leveraging the properties of the Cas9 endonuclease - (Gilpatrick et al., 2020) enables the generation of high coverage for specific genomic regions in a single sequencing run, thus allowing to correct for base-calling errors.

To overcome low sequencing depth, nCats is making use of targeted CRISPR-Cas9 *in-vitro* cleavage of isolated genomic DNA that results in site-specific ligation of nanopore adapters to regions of interest, which greatly increases the on-target read coverage of nanopore sequencing runs. Due to the PCR-free sequencing of native DNA, nCats offers the great advantage of generating long, continuous reads which simultaneously provides information on the DNA methylation status and on the precise arrangement of multiple plasmid copies, if present, at the integration site. Importantly, by choosing CRISPR RNAs (crRNAs) that either align to the sense or antisense DNA strand, the direction of sequencing can be defined. For the identification of transgene integration sites, it would be therefore sufficient to apply nCats with crRNAs that direct the sequencing towards the 5'-junction and the 3'-junction sites (JSs) of transgenes with the neighboring chromatin. In a subsequent nCats run, crRNAs can be designed that allow to sequence from the genome into the integrated transgene sequence, thus resolving the integrity and arrangement of the expression vector. Apart from identifying or verifying transgene integration, the technique can also be used to specifically assess the epigenetic status of endogenous genes next to the integration site.

3 Materials and methods

All crRNAs and primers for PCR, qPCR or sequencing reactions are listed in **Supplementary Table 1**.

3.1 Materials and methods

CHO-K1 cells (ECACC 85051005) were cultivated in CD-CHO medium (Thermo Fisher Scientific) supplemented with 8mM L-glutamine (Sigma-Aldrich) and 0.2% Anti-Clumping Agent (Thermo Fisher Scientific). CHO-K1 $GS^{-/-}$ CTG5 cells expressing human cluster of differentiation 4 (CD4) (CHO-CD4) were generated by (Baumann et al., 2017) (cell pool CD4-CTG 5 after the third sorting round). CHO-CD4 cells were cultivated in CD-CHO medium, supplemented with 8mM L-glutamine, 700 $\mu g \cdot \mu L^{-1}$ µM G418 (Sigma-Aldrich) and 0.2% Anti-Clumping Agent. CHO-K1 $GS^{-/-}$ cells expressing the Trastuzumab (Herceptin) IgG antibody is a clonal cell line from Baumann et al. (Baumann et al., 2017) for which the CD4 expression cassette of CHO-CD4 was exchanged via RMCE with a Herceptin expression cassette. CHO-Herc cells were routinely grown in CD-CHO medium, supplemented with 8mM L-glutamine, 10 $\mu g \cdot \mu L^{-1}$ Blasticidin (Invivogen) and 0.2% Anti-Clumping Agent. Cells were grown in TPP TubeSpin® Bioreactor 50 tubes (Techno Plastic Products AG) in 10 mL medium and passaged every 3-4 days. Spin tube cultures were incubated at 37°C, 7% CO_2 and 80% humidified air at a shaking speed of 250 rpm (25 mm shaking diameter). The EV sequences of CHO-CD4 and CHO-Herc are provided in **Supplementary Figure 4 and Supplementary Figure 5**.

3.2 Design of crRNAs

crRNAs were designed with the CRISPR-Cas9 target online predictor (CRISPOR) (Concordet & Haeussler, 2018). The PAM motif 20bp-NGG Sp Cas9 and SpCas9-HF1, eSpCas9 1.1 was selected to search for suitable crRNAs. Off-target prediction was selected against the provided

CriGri-PICR reference genome (GCF_003668045.1, PICR) (Rupp et al., 2018), since the PICRH genome was not available. crRNAs were selected based on the provided “MIT specificity score” and needed to contain at least three mismatches to bin to off-targets, preferentially in the 12 bp seed region. crRNA selection was then then validated by blasting (Stothard, 2000) the sequences against the CriGri-PICRH-1.0 reference genome (GCF_003668045.3, PICRH) (Hilliard et al., 2020). Selected crRNAs were ordered from Integrated DNA Technologies (IDT) (Alt-R® CRISPR-Cas9 crRNA, 2 nmol), resuspended in TE Buffer (pH 7.5) to a final concentration of 100 µM and stored at -20°C until further use.

3.3 High molecular weight (HMW) DNA isolation

HMW DNA isolation for library preparation was performed with the Monarch® HMW DNA Extraction Kit for Cells & Blood (New England Biolabs) according to manufacturer’s instructions with the following adjustments. $5 \cdot 10^6$ cells were used as starting material for DNA isolation and the agitation speed for cell lysis was set to 2000 rpm. To achieve high DNA concentration, 70 µL of elution buffer were used to extract HMW DNA. In order to reduce the viscosity of the DNA, the eluate was further subjected to 45 min incubation at 37 °C followed by 20 times shearing with a blunt 26G needle (Circulomics). The DNA concentration was determined using a Qubit 3 fluorometer (Thermo Fisher Scientific) and was stored at 4 °C until further use.

3.4 nCats library preparation

The Cas9 Sequencing Kit (SQK-CS9109, Oxford Nanopore Technologies (ONT)) was used to perform library preparation according to manufacturer’s instructions, only for run ID 1 and run ID 7 the Ligation Sequencing Kit (SQK-LSK109, ONT) was used. When using the Ligation Sequencing Kit, the protocol and reagents described by Gilpatrick et al. were used (Gilpatrick et al., 2020). For multiple Cas9 targeting, 0.75 µL of resuspended crRNAs were pooled. To form a

guide RNA duplex, 8 µL Nuclease Free Water (NFW, Thermo Fisher Scientific), 1 µL of 100 mM tracrRNA (IDT) and 1 µL of pooled crRNAs were mixed and heated up for 5 min at 95 °C in a thermocycler. After the duplex was cooled down to RT, the RNP complex was formed by mixing 23.7 µL NFW, 3 µL Reaction Buffer (ONT), 3 µL guide RNA duplex and 0.3 µL Alt-R® S.p. Cas9 Nuclease V3 (IDT) in a PCR tube and incubated at RT for 20 min. In the meantime, the DNA ends of HMW DNA were dephosphorylated by adding 10 µg of HMW DNA with 3 µL of Reaction Buffer to a volume of 27 µL (NFW was added if needed). Then, 3 µL of Phosphatase (ONT) was added to the mixture and incubated in a thermocycler at 37 °C for 10 min, 2 min at 80 °C and cooled down to RT. 10 µL RNP complex, 1 µL dATP (ONT) and 1 µL of TAQ Polymerase (ONT) were added to the dephosphorylated DNA. Subsequently, the RNP-DNA mix was incubated at 37 °C for 15 min and 72 °C for 5 min. The cleaved DNA was cooled down at RT and transferred into a 1,5 mL DNA LoBind tube (Eppendorf). For the adapter ligation mix, 20 µL of Ligation Buffer (ONT), 3 µL of NFW, 10 µL of T4 DNA Ligase (ONT) and 5 µL of Adapter Mix (ONT) were mixed in an additional 1.5 mL DNA LoBind tube. 20 µL of the adapter ligation mixture was added to the cleaved DNA, the tube was flicked, and the remaining 18 µL were mixed with the DNA and incubated for 10 min at RT. 80 µL of SPRI Buffer (ONT) and 48 µL of AMPure XP beads (Beckman Coulter) were added to the tube. The sample was mixed by inversion and incubated at RT for 15 min. AMPure XP beads were assembled on a magnetic rack for 10 min and the supernatant was pipetted off. Then, the AMPure XP beads were resuspended in 250 µL Long Fragment Buffer (ONT) and collected on a magnetic rack for 10 min. The supernatant was pipetted off and the wash step repeated. 13 µL of Elution Buffer (ONT) was added to the AMPure XP beads and mixed by stirring, followed by 30 min incubation at RT. The eluate was transferred to a fresh DNA LoBind tube. The DNA concentration was determined with a Qubit device. Then, the nanopore flow cell was prepared according to the manufacturer’s instruction.

3.5 Nanopore Sequencing

The samples were run on R9.4.1 flow cells (ONT). Sequencing was performed with a MinION MK1B sequencing device and operated with the MinKNOW software (v4.3.25). Fast5 files were converted to fastq format using the GPU basecaller Guppy (v5.0.16) for windows with default settings. Samples were sequenced for ~ 20h. Flow cells used for a consecutive sequencing run were washed with the flow cell wash kit (EXP-WSH004) according to manufacturer's instructions. All runs and their statistics are provided in the Supplementary Data File.

3.6 Nanopore data analysis of endogenous regions

Basecalled reads in fastq format were mapped with minimap2 (v2.17) (Li, 2018) with the ('-a') option against the PICRH assembly. Samtools view (v1.15.1) (Danecek et al., 2021) was used to process the alignments to exclude unmapped reads, non-primary and supplementary alignments ('-F 2308') as well as reads with a MAPping Quality (MAPQ) < 30 ('-q 31'). To identify reads that align to the genomic regions enclosed by the crRNAs, bedtools intersect (v2.30.0) with default setting was used. On-target reads were assembled with flye (v2.9-b1779) (Kolmogorov et al., 2019) for nanopore reads, with the option ('--asm-coverage') set to 100 to confine on-target coverage. The assembly was conducted by four iterations of polishing with racon (v1.4.20) (Vaser et al., 2017). The score for matching bases ('-m') was increased to 8 whereas the score for mismatches ('-x') was decreased to minus 6. Stats_from_bam, which is included in the pomoxis program package (v0.3.4), was used with default settings to determine indels and substitutions of the polished assembly. Homopolymeric regions were determined as such when >3 of the same nucleotide were repeated.

Samtools was used to determine the total number of aligned reads, the on-target enrichment as well as the median and maximum per-base coverage of the ROI. Stats_from_bam was used to calculate the strand bias, mean read accuracy and the mean read length of on-target reads.

Bedtools intersect with the write overlap option ('-wo') was used to calculate the number of bases and the mean coverage of the target region.

The data for the coverage plots were processed using the following R packages: GenomicAlignments (v1.30.0), GenomicRanges (v1.46.1) and tidyverse (v1.3.1). On-target reads are defined as those that align within the minimal region flanked by crRNAs.

3.7 Identification of junction sites

To determine transgene-genome junction sites, sequences of the CD4 expression vector were linearized at two different sites and added as individual contigs to the PICRH assembly for analysis of the CHO-CD4 cell line (**Supplementary Figure 4**). For the CHO-Herc cell line, two additional contigs of the Herceptin expression vector were added to the PICRH assembly (**Supplementary Figure 5**). Nanopore reads were mapped against the modified reference genomes using minimap2 with default parameters to identify chimeric reads. Those reads composed of a transgenic and endogenous sequence and were retained. To determine individual junction sites, chimeric reads were grouped according to their mapping position. Reads aligning to the sense or anti-sense strand as well as the end or start coordinate of the alignment were used to allocate individual junction sites, respectively. For visual inspection, reads were plotted in the Integrative Genomics Viewer (Robinson et al., 2011) . Reads of each junction site were individually assembled and polished as described in chapter 3.6. The polished assembly of each junction site was identified by BLASTN search against the PICRH reference genome and against the expression vector to obtain the exact position of the junction site (Altschul et al., 1990). The annotation as 5'-junction and 3'-junction sites was done relative to the PICRH assembly.

3.8 Determination of integrated transgene sequences

The polished assemblies of junction sites (JS2, JS3, JS4) were further used to design crRNAs targeting the genomic portion of the junction sites. Library preparation with these crRNAs and nanopore sequencing was performed as described above. Data analysis was performed as described in 3.7 with the following adjustments. For integration site B, shasta (v0.7.0) (Shafin et al., 2020) with default settings was used to create assemblies for both recombinant cell lines, since flye (v2.9-b1779) did result in incorrect assembly of this site. All other assemblies were generated with flye.

3.9 CpG methylation frequency of endogenous and transgenic regions

CpG methylation frequency was determined with nanopolish (v0.13.3) (Simpson et al., 2017) according to the quick start manual. For analysis, supplementary and secondary alignments were excluded with samtools view ('-F 2304') from the alignment bam file. For DNA methylation analysis of endogenous regions (run IDs 1-3) or transgene integration sites (run IDs 7-9), the CriGri-PICR assembly or the assemblies of the transgene integration sites were used as reference, respectively. To obtain methylation information of all CpG patterns, the helper script calculate_methylation_frequency.py was executed with the ('-s') option. Whole genome bisulfite data of a CHO-K1 cell line (CHO-K1-8mM-Gln-MCB-exp; published by (Feichtinger et al., 2016)) mapped to the CriGri-PICR assembly was accessible from the publication of Weinguny et al. (Weinguny et al., 2021) and was processed using the bsseq R package (v4.1.2) (Hansen et al., 2012). The smoothing loess function was used to create line plots to compare nanopore with whole genome BS-Seq (WGBS) data. The coefficient of determination was calculated with the lm base R function.

3.10 Determination of crRNA off-target binding sites

The identification of off-target reads introduced by the applied crRNAs during the nCats library preparation process was performed according to (Höijer et al., 2020). In brief, insider v1.9 software with default settings was used to identify genomic positions of reads with overlapping start or end sites. Resulting bed file coordinates in close proximity were combined with bedtools merge (Quinlan & Hall, 2010) (parameters: -header -d 10) and the sequences were extracted with a ± 40 bp window around the possible crRNA site with samtools faidx (Li et al., 2009). Sequences containing unassigned bases (Ns) were filtered out. Pairwise alignment of the individual crRNAs and the extracted sequences was performed with EMBOSS needle v6.5.7 (Rice et al., 2000) with default parameters except for -gap extend, which was set to 5. According to (Höijer et al., 2020), sequences with an alignment score of >55 were considered positive binding sites. Additionally, sites without the consensus PAM motif for Cas9 (NGG) at the 3'-end were filtered out. To account for random adapter ligation, the third quartile read count of all off-target sites called by insider, excluding the called positive binding sites as well as true on-target reads, was used as a threshold to filter potential crRNA off-target sites.

3.11 Sanger sequencing of PCR products

Genomic DNA of CHO-CD4 and CHO-Herc cell lines was isolated with the DNeasy® Blood & Tissue Kit (Qiagen) according to the manufacturer's protocol. Transgene junction sites were PCR-amplified with Phusion High-Fidelity DNA Polymerase (2 U/µL) (Thermo Fisher Scientific) and respective primers with 95 °C initial denaturation for 3 min followed by 30 cycles of 98°C for 20 sec, T_a°C for 20 sec, 72°C for t_{ext} and a final extension step of 72°C for 1 min, whereas the elongation time (t_{ext}) and the annealing temperature (T_a) were individually adjusted to the amplicon size and primers used. Amplicons were purified using the DNA Clean & Concentrator®-5 kit (Zymo Research) according to the manufacturer's instructions. DNA concentration and purity

were determined with the NanoDrop™ One/OneC Microvolume UV-Vis Spectrophotometer (Thermo Fisher Scientific). PCR products were sequenced by Sanger sequencing (Sanger et al., 1977)(Eurofins Genomics) and results were analyzed with QIAGEN CLC Main Workbench.

4 Results and Discussion

4.1 Accurate genomic and epigenomic assessment of endogenous genomic regions with nanopore Cas9-targeted sequencing

For nCats library preparation, dephosphorylated HMW DNA is first subjected to CRISPR-Cas9 cleavage, generating new phosphorylated 5'-end at the cleavage site. The Cas9 endonuclease tends to stay attached to the cleaved DNA fragment not containing the PAM-sequence motif, thus shielding the 5'-end of this particular DNA fragment. The unbound DNA fragment and its phosphate group, however, is freely accessible enabling preferential ligation of nanopore adapters to these fragments and enriched sequencing of the CRISPR-Cas9-targeted region (**Figure 1**B). Additionally, the strandedness of generated nanopore reads is controlled by the complementarity of the crRNAs to the DNA strand, which allows site specific, directional sequencing of targeted DNA loci with high on-target coverage. The directionality of this approach permits to either sequence regions with single DNA strands with undefined length (unidirectional crRNAs or bidirectional crRNAs not enclosing the target strand, single-end reads) or with defined lengths (bidirectional crRNAs enclosing the target DNA strand, paired-end reads) (**Figure 1**C). While the latter results in a higher on-target coverage (Gilpatrick et al., 2020), both target sequences have to be known a priori. The former on the other hand, although resulting in lower on-target coverage, allows directed long-range sequencing without prior sequence information of neighboring genomic regions. Thus, integration sites of transgenes can be identified if appropriate

crRNAs are applied that direct the sequenced reads from the known expression cassette into the unknown genomic region.

In order to test the general functionality of nCats in CHO cells, we designed bidirectional crRNAs targeting two different endogenous genomic regions of CHO-K1 cells. We chose the β-1,4-galactosyltransferase (B4galt1) gene and the promoter of the α-1,6-fucosyltransferase (Fut8) gene, which are well described targets for CHO cell engineering (Bydlinski et al., 2018; Marx et al., 2021; Schmieder et al., 2018; Yamane-Ohnuki et al., 2004). The B4galt1 gene was targeted with crRNAs enclosing exon 1 and exon 2 (chr2: 101759735..101804787) spanning 26 Kb. For the Fut8 region, crRNAs were targeted to enclose the previously identified promoter region (chr5: 126277153..126273156) (Marx et al., 2021) spanning 18 kb. For the sequencing of the Fut8 promoter region, CHO-K1 cell pools with an artificially methylated promoter (Fut8neg) and with the endogenously demethylated, active promoter(Fut8pos) CHO-K1 cell pools to targeted DNA methylation with dCas9-DNMT3A3L (Marx et al., 2021) were used as a model to benchmark DNA methylation analysis with nCats **Figure 2**.

Resulting nanopore reads were mapped against the PICRH reference genome and on-target median coverage was calculated for the targeted regions. Median per-base coverage was 95x for the B4galt1 gene library (run ID 1), whereas coverage was up to 724x for the libraries targeting the Fut8 promoter (run ID 2 and run ID 3) (**Figure 2**J), which is comparable to TLA-Seq coverage. The higher coverage for the Fut8 promoter libraries can be attributed to the use of the more recent SQK-CS9109 Kit (instead of using SQK-LSK109 and other third-party reagents. Although the Fut8pos library (ID 2) was run on a washed, re-used flow cell with less available pores (645 pores instead of 1534 for the Fut8neg run (ID 3) and on-target coverage was significantly lower (192x) than for the Fut8neg library (724x), the percentage of mapped reads was still twice that of the B4galt1 library (fresh flow cell). As expected, the on-target reads of the respective genomic

regions started around the cut site of the used crRNAs (indicated by arrows in **Figure 1**A-B and D-E). The high on-target coverage also allowed *de novo* assembly of the targeted genomic regions resulting in an 18274 bp assembly of Fut8pos, 18242 bp assembly of Fut8neg and 26025 bp assembly of B4galt1.

An intrinsic bottleneck of nanopore sequencing is the accurate calling of homopolymeric regions (Simpson et al., 2017). In principle, this systematic error can be filtered out and a high sequencing coverage even allows to identify SNVs (Gilpatrick et al., 2020). When aligning the assembled sequences of the B4galt1 and Fut8 sequencing runs against the PICRH reference genome without further processing, 128 indels and 6 substitutions for B4galt1; 90 indels and 27 substitutions for Fut8pos and 91 indels and 29 substitutions for Fut8neg were called initially. However, after further analysis, out of the 91 indels of the Fut8neg assembly most indels (82) could be attributed to the addition of single bases at homopolymeric regions. This was similar for the B4galt1 (112 out of 128) and Fut8pos (76 out of 90) assemblies. Thus, additional bioinformatic processing does allow to refine the accuracy of nanopore assemblies. When taking the remaining indels and substitutions into account, 0.197% (translating to a Phred Quality Score Q of 27) of the bases were inaccurately called of the Fut8neg assembly (0.225% or Q of 26 for Fut8pos and 0.046% or Q of 33 for B4galt1) in comparison to the PICRH reference genome. Given that the genomic sequence of the sequenced cell line is likely different from the PICRH reference genome, the true rates are probably lower. Therefore, nanopore reads generated by nCats can provide accurate identification of genomic loci at base pair resolution.

4.2 Assessment of crRNA off-target effects with nCats

Off-target activity of crRNAs still poses a considerable problem, when applying CRISPR-Cas9 complexes for genome editing (Hsu et al., 2013) and can also affect which reads are generated by nCats. In order to validate the applied crRNAs used, we assessed potential off-targets as

reported previously (Höijer et al., 2020). From the aligned sam files, nanopore reads were checked for potential crRNA off-target binding sites. For the Fut8 promoter region, genome-wide analysis revealed 10 off-target binding sites (642 off-reads) for the Fut8neg cell line compared to 7 off-target binding sites (168 off-reads) for the Fut8pos cell line with an off-target:on-target ratio of 0.72 and 0.76, respectively (Supplementary Data). While two of the additional off-target sites of the Fut8neg sequencing run were filtered out in the Fut8pos dataset during quality control, the third additional off-target site (start at NC_048595.1: 311511597) was unique to the Fut8neg sequencing run. However, the 13 generated reads at this site were surpassing the set threshold only by one read and it might be an artefact of too permissive settings. On the other hand, the much lower read count of the Fut8pos library, might have hindered the detection of this site. For the B4galt1 sequencing run, 55 potential off-target sites and 1588 off-target reads for the crRNAs used were reported, resulting in an off-target:on-target ratio of ~13. Thus, the comparatively low on-target enrichment for this run might be, next to the use of another library prep kit, due to the high number of potential crRNA off-target reads. While the potential crRNA off-target reads did not impede the retrieval of the correct genomic locations with nCats, the high numbers highlight the importance of a careful crRNA selection process. Simultaneously, the high on-to-off-target ratio promises great and relatively simple optimization potential for the current nCats process if more accurate crRNAs are used.

Next to the impact on nCats efficiency, such off-target analysis might be a valuable tool for studies using CRISPR complexes in CHO cell research, which has oftentimes not been assessed, partly due to the limited availability of tools for detection. The identification of off-target sites by nCats could, next to the identification of on-target sites, be used to validate the off-target behavior of crRNAs selected for genome editing applications and thus improve accuracy and efficacy of these experiments. Importantly, it seems that common selection criteria for crRNAs and gRNAs used

for genome modifications in CHO cells should be reconsidered, especially in light of genome - wide CRISPR-screens.

4.3 nCats DNA methylation assessment reproduces WGBS data

While DNA sequence information of a genomic region could be also resolved by other (more laborious) methods, nCats offers the added advantage to salvage the generated raw data for more than sequence information by simultaneously analyzing the epigenetic status, specifically the DNA-methylation pattern, of the sequenced DNA. Nanopore signals of sequenced, PCR-free DNA can be used to differentiate between methylated and non-methylated CpG sites due to subtle changes of the measured electrical current (Simpson et al., 2017). Since DNA methylation is the dominant epigenetic mechanism for stable gene regulation in CHO cells (Feichtinger et al., 2016; Hernandez et al., 2019; Marx et al., 2018, 2021; Weinguny et al., 2020), the simultaneous read-out of sequencing information and epigenetic status from the same data could greatly facilitate the interpretation of suitability of stable integration sites but also that of random and targeted epigenetic editing studies. In CHO cells, whole genome nanopore methylation analysis has been used to investigate the epigenetic status between cell clones exhibiting different productivity (Chang et al., 2022). In order to validate nanopore methylation calls from nCats, we subjected the generated raw reads of the B4galt1 gene, the Fut8neg (silenced Fut8 promoter via CRISPR-based targeted methylation) and Fut8pos (native, functional Fut8 promoter) cell pools to methylation analysis with nanopolish. For the sequenced region, DNA methylation data of the native cell pools correlated well with whole genome bisulfite sequencing data (Weinguny et al., 2021) (**Figure 2**G-H), which has also been reported for other data sets (Simpson et al., 2017). For the downregulated cell pool, a strong increase of methylation was detected around the target sites for CRISPR-based DNA methylation (**Figure 2**F and **Figure 2**I), which replicated the promoter

methylation status that was analyzed before by targeted bisulfite sequencing (**Figure 2**K) (Marx et al., 2021).

The combined results provide proof-of-concept of the nCats technology's capability to accurately retrieve high coverage DNA sequence and methylation information in CHO cells.

4.4 Identification of genome-transgene junction sites of recombinant cell lines with nCats

Next, we applied nCats to identify genome-transgene junction sites of recombinant CHO cells. CHO-K1 cell lines expressing either the human cluster of differentiation 4 (CHO-CD4) receptor protein or the Trastuzumab (Herceptin) IgG antibody (CHO-Herc) were used for analysis. While CHO-CD4 was previously characterized by TLA-sequencing (Baumann et al., 2017), the CHO-Herc cell line was generated by RMCE from the CHO-CD4 cell line, thus both cell lines should have matching transgene integration sites (**Figure 3**A).

crRNAs targeting either the CMV promoter or the corresponding transgenes were designed and applied to sequence from the expression cassette into the host genome (**Figure 3**B) resulting in chimeric reads that contain information about the junction sites between the genome and the inserted transgene. nCats data of generated libraries of the CHO-CD4 cell line confirmed the genomic flanking regions of the integrated EV of integration site A (ISA; 5'-junction: NW_023276807.1:235393256, (JS1; run ID 5); 3'-junction: NW_023276807.1:235393277, (JS2; run ID 4)) as reported by TLA-sequencing (**Figure 3**C). The junction sites of the other previously reported integration site (ISB) were also retrieved by nCats, although partly in different orientations. While junction site 3 (JS3; NW_023276807.1:32641832; run ID 4) was identified by both methods as a 3'-junction site (**Figure 3**D), junction site 4 (JS4; NW_023276806.1:7791670; run ID 4) was identified by nCats as a 3'-junction site in contrast to TLA-Seq (5'-junction).

Interestingly, two additional junction sites not identified with TLA-Seq data were reported by nCats, which revealed an inverted duplication of the genomic portion of JS4 as an additional 5'-junction site (invNW_023276806.1:7784008 (JS5; run ID 4)) (**Figure 3**E). Next to the duplication of the genomic sequence, the transgenic portions of JS4 and JS5 start almost at the same vector coordinates and in the same orientation, which might have impeded the alignment of short-reads by TLA-Seq and successively the correct identification of junction sites. nCats on the other hand, due to the generation of long contiguous reads, enabled the identification of both sites highlighting the potential for accurate identification of integration sites in cell lines that frequently show chromosomal aberrations (Baik & Lee, 2017; Vcelar et al., 2017, 2018). Another additional 5'-junction was identified from chimeric reads whose genomic portions mapped to various low complexity regions. In total 37 chimeric reads were generated and after assembly of the reads, the 5'-junctions site was identified as NC_048602.1:15472407 in inverted orientation (JS6; run ID 5) based on query coverage after blasting against the PICRH reference genome. The overall lower number of on-target reads stems from the utilization of a washed and re-used flow cell. Initially only 5 of 37 reads aligned to this site, suggesting that low complexity regions can be difficult to allocate with the current mapping tools. Nevertheless, Sanger sequencing confirmed all identified junction sites, even those located in low complexity regions and assembled from a relatively low number of on-target reads (JS6) (**Figure 3**F). In concordance, the identified junction sites of the CHO-Herc cell line (run ID 6) were identical to the ones found in the CHO-CD4 cell line, which was expected due to RMCE-mediated exchange of transgenes (**Figure 3**C-E, **Supplementary Figure 6**). However, no directional reads for JS5 and JS6 were generated by nCats during this sequencing run. JS5 was partly recovered with reads that were sequenced in the "wrong" orientation, i.e. in 3'-direction from the PAM-site. These reads were generated during library preparation as Cas9 is only partially shielding the PAM-distal phosphorylated 5'-ends.

These accidental reads and the simultaneous absence of directional reads indicated that different transgene elements possibly in different orientations were present or absent at these sites in the CHO-Herc cell line after RMCE. On the other hand, a high number of short non-chimeric reads with similar length (~2.5 Kb) that aligned to the CD4 EV only were observed for the CHO-CD4 cell line (run ID 4, **Supplementary Figure 1**), which hinted at integration of multiple vector repeats. Two target sites for the used crRNAs, either in the same or in inverted orientation, would lead to sequencing of a defined length within the expression construct rather than the anticipated chimeric transgene-genome reads. Nonetheless, the data generated here reliably identified junction sites of integrated transgenes and, especially in the case of low complexity or repeat sequences, is advantageous over current short-read sequencing solutions.

4.5 nCats reliably identifies transgene integration sites and reveals prior unidentified long, concatemerized and inverted transgene repeats and its epigenetic context

While the identification of genome-transgene junction sites was successfully analyzed by nCats, the matching junction sites and the complete composition of the expression cassettes cannot be fully resolved with the current approach where reads start within the transgene sequence. While the identification of junction sites might be sufficient as a selection criterion for clones with correct targeted transgene integration(s), information on the correct orientation and functionality of the integrated vectors is crucial for the final selection of stable and high producer cell lines (Dhiman et al., 2020).

To fully identify the integrated expression vector structure, crRNAs were designed to target the 3’-genomic side of identified junction sites (JS2 (run IDs 7+9), JS3 (run IDs 8+9), JS4 (run IDs 7+9)) of the CHO-CD4 and CHO-Herc cell lines allowing to sequence from the genome into and

over the transgenic sequence. The generated reads allowed the reconstruction of three individual integration sites. The 20 bp genomic deletion between the breakpoints within integration site A identified by TLA was reproduced by nCats (**Figure 4**D&E), however, integration sites B and C differed. Here, the TLA-sequencing report states that integration site B is composed of JS4 (NW_023276806.1: 7791670) as a 5'-junction site (nCats: 3'-junction) and JS3 (NW_023276807.1:32641832). In contrast, nCats data sequenced from JS4 over the expression construct sequence identified JS5 (invNW_023276806.1:7784008) as the matching 5'-junction site to JS4, thus revealing a beforehand unidentified duplication and inversion event of chromosome 1 at ISB (invNW_023276806.1:7784008-NW_023276806.1: 7791670) (**Figure 4F&H**). A high frequency of intra-chromosomal translocations of this chromosome has been also reported for other CHO cell lines (Dhiman et al., 2020). Since the crRNAs used to sequence from JS4 over the transgene are also binding in the genomic site of JS5 but in inverted orientation, sequencing was conducted in paired-end mode (see **Figure 1**C), which is highlighted in the coverage plateau between the junction sites (α). The two additional plateaus close to JS4 are in part artefacts of short reads generated from using multiple crRNAs (γ) and the multiplication of sequencing events due the two inverted chromosomal duplication of JS4 and JS5 (β) (**Figure 4F**). nCats from JS3 into the transgenic region enabled the identification of the third integration site C (ISC; NC_048602.1:15472407-NW_023276807.1:32641832) with a JS6 as the matching 5'-junction site revealing a genomic rearrangement of chromosome 1 (JS3) with chromosome 9 (JS6) (**Figure 4G&I**). Chromosome 9 is generally associated with high genomic instability and inter-chromosomal translocations with chromosome 1 have been also reported for other cell lines (Dhiman et al., 2020). Accordingly, around ISC a 9 kbp deletion between NW_023276807.1:32613025-32622394 relative to the PICRH genome could be detected as highlighted in **Figure 4G** and **Supplementary Figure 2**. The same integration sites were

identified in the CHO-Herc cell line (runs starting at the 3'-end of JS2, JS3, JS4 (all run ID 9)) further consolidating the nCats read-out (**Figure 4D-G**).

Due to the chromosomal translocations at ISB and ISC relative to the reference genome, the continuity of the integration sites between JS5-JS4 and JS6-JS3, respectively, cannot be visualized. To illustrate the continuity of these integration sites, the previously generated chimeric reads for the identification of the junction sites (**Figure 3**D&F) were aligned against the assemblies of ISB and ISC (**Figure 4H&I**) of the CHO-CD4 cell line. Since ISA is genomically preserved, the site of integration is visible due to the drop in coverage (**Figure 4E**). It is noteworthy that - since reads are starting in the genome - both chimeric and endogenous reads will be generated if the cassette is only integrated into one of the alleles. Thus, the coverage drop at ISA of ~50% corresponds to a single copy integration event within the genome. In contrast, the reads generated by targeting the genomic portion of JS3 resulted in predominantly genomic reads (97 genomic reads vs 4 chimeric reads and 322 genomic reads vs 17 chimeric reads for CHO-CD4 and CHO-Herc, respectively), and in a marginal coverage drop at JS3 (**Figure 4G**, indicated by a dashed arrow), hinting at multiple copies of this genomic region within the analyzed cell lines. qPCR amplification of the transgenic and the genomic part of JS3 consolidates this assumption in comparison to a non-recombinant CHO-K1 cell line (**Supplementary Figure 3**). Whether transgene copy number can be accurately determined with this approach remains to be investigated, but it can be roughly estimated with nCats.

The long, contiguous nanopore reads enabled the analysis of the composition of the transgene sequences at the different integration sites (**Figure 4**A). In ISA of the CHO-CD4 cell line, the expression vector was integrated from bp 5752 to bp 4408 (inverse relative to the sequence in **Supplementary Figure 4**), which confirmed the TLA-Seq data. After RMCE treatment, the CD4 expression vector was completely removed, and the Herceptin expression vector was integrated

in the CHO-Herc cell line leaving the junction sites (JS1 and JS2) unchanged. Analysis of ISB and ISC of CHO-CD4, on the other hand, revealed several concatemerized plasmid fragments that were in part inverted, ranging from ~21 kbp (ISB) to ~ 11 kbp (ISC). Interestingly, analysis of the transgenic sequence of the CHO-Herc cell line at these integration sites revealed remaining truncated versions of the CD4 expression vector but not the Herceptin expression construct, thus showing imprecise RMCE processing at these sites. Here, only parts of the concatemerized CD4 EV sequence was removed but not exchanged with the Herceptin EV. These findings explain the inability to completely identify the junction sites (JS5 and JS6) for CHO-Herc as noted in 4.4, since the used crRNA were designed to sequence the reverse strand of the CMV promoter or the forward strand of the heavy chain within the Herceptin EV. Thus, for very complex integrations and to identify residual EVs (e.g. after RMCE), several crRNA combinations might be necessary to capture all junction sites. Still, the presented nCats approach allowed to capture all integration sites of this cell line. Importantly, for site B and C no complete sequence information of the CD4 expression construct was available from TLA-Seq data, showcasing the applicability of nCats to produce results also for complex integrations (**Figure 4C**).

The duplication and inversion of the vector fragments in the CHO-CD4 cell line explains the high number of short reads mapping only to the CD4 expression vector, as illustrated in **Supplementary Figure 1**. Due to the inversion and multiplicity of vector sequences in integration site B, Cas9-mediated adapter ligation occurred at opposing sites, thus enclosing a ca 2.5 kbp DNA stretch, which relates to the observed length of the transgenic reads. Although identification of junction sites was successful, these findings indicate a potential issue of nCats if multiple concatemerized expression vectors are integrated into the genome and only a single sequencing run to target the EVs is performed.

The difference observed between TLA-Seq data and the confirmed nCats data could be influenced by several factors. First, the cell lines were not sequenced simultaneously by TLA and nCats, but years apart and, given the genetic plasticity of CHO cells, chromothripsis comprising the integration sites might have occurred over time. Additionally, the CHO-CD4 cell line is a cell pool generated by random integration of the RMCE cassette and subsequent selection by cell sorting. Thus, subpopulations within the cell pool that were outcompeting others over time could have impacted the different read-outs. However, since CHO-Herc is a clonal cell line, extreme variability would not be expected for this data set. Despite the clonality, most IS identified in the CD4 “pool” were also present in the Herceptin producing clone, indicating that the initial sorted pool was already very homogenous. Additionally, non-chimeric reads from sequencing run ID 7 for the identification of ISB showed no chromosomal translocation at the sequenced genomic locus, which indicates the intactness of the native allele. These reads hint at genomic rearrangements at the time of transgene integration, which left ISB probably unchanged between the sequencing analyses. The overall similarity of the called junction sites between TLA-seq and nCats data (4 correct junction sites (JS1-4) and an inverted duplicate (JS5)) and the reproducibility between CHO-CD4 and CHO-Herc rather implies that the complexity of the genomic DNA sequence influenced the correct mapping of the integration sites. It is noteworthy, that JS6 is within a region of low complexity, parts of which match a plethora of other genomic regions within the PICRH genome. In combination with the presence of inverted genomic sequences (JS5 and JS6), the correctly called integration sites by nCats might simply not be identifiable with short-read sequencing technologies. Given the documented advantage of nanopore sequencing to span long repetitive regions (Nurk et al., 2022), our data highlights the robustness of nCats to identify transgene integrations even in repetitive (ISB) or low complexity (ISC) regions.

By subjecting the raw reads of the CHO-CD4 and CHO-Herc cell lines to nanopore methylation calling, we analyzed the epigenetic status of the two integration sites with the highest coverage (A and B). While ISA of both cell lines was lowly methylated, integration site B showed varying but steadily higher DNA methylation levels. When comparing the average methylation of the CMV promoter, most of the CMV promoter copies were above 50% methylated at ISB but only around 4% methylated at ISA, which suggests that the transgenes at ISB are less or not expressed. This assumption is also supported by the fact that in the Herceptin subclone, CD4 expression cannot be detected. Interestingly, the methylation status of the promoter and the gene body are rather similar, which is in contrast to observations of endogenous genes in CHO cells (Feichtinger et al., 2016). Overall, our findings suggest that cell line instability with respect to transgene expression due to concatemers could also be impacted by epigenetic mechanisms of the cell's regulatory machinery. In combination, nCats does not only provide sequence information but also insight about a regulatory layer for gene transcription at the investigated sites enabling more detailed (epi)genotypic analyses.

5 Conclusion

The generation of recombinant cell lines relies on the integration of transgenic sequences into the host genome. The number of integrated transgene copies, their integrity as well as the epigenetic environment are crucial parameters that greatly impact the level of recombinant protein expression of the generated cell lines. While several options exist to separately characterize these features, technologies that simultaneously provide information on all three are still lacking. Here, we show that Cas9-targeted nanopore sequencing can be used to accurately analyze integration sites, the structure of integrated sequences and, importantly, the chromatin status of the targeted region.

The strength of nCats and its advantage over other technologies lies in its ability to reliably and simultaneously identify repetitive and concatemerized transgene sequences, their locus of integration and simultaneously its surrounding epigenetic status. The long reads generated with nCats allow to span over transgene-genome junction sites including repetitive sequences, which might otherwise hinder assignment of short reads produced by traditional sequencing technologies. nCats reported integration sites and the structure of the expression cassette more comprehensively than TLA-seq data and allowed the identification of another, beforehand undetected integration site. The information from sequencing long native reads can also overcome existing issues associated with traditional techniques, i.e. bisulfite sequencing, that are limited in coverage as well as sequence length and exhibit PCR-induced biases. Thus, nCats presents a viable alternative to study the DNA methylation landscape of potential integration sites. It is noteworthy that previous targeted DNA (de)methylation experiments (Marx et al., 2018, 2021) would have benefited from nCats analysis.

However, nCats - and nanopore sequencing in general - still suffers from several shortcomings. Base-calling, and thus the determination of the exact transgenic sequence, is still less accurate than short-read sequencing approaches due the inherent, systematic error of nanopore technology. Thus, important information about missense or nonsense mutations in the transgenic sequence might not be called with the highest accuracy with the presented nCats approach. Additionally, with the presented approach, two sequencing runs were necessary to accurately detect both junction and integration sites, since sample multiplexing is currently not feasible with the provided library preparation kits. Consequently, high-throughput analyses are so far not realizable.

Continuous improvement of the nanopore technology as well as changes in the experimental design of nCats in particular can aid in overcoming some of these issues. According to the

provider of nanopore technologies the newest Kit 14 chemistry can produce data with quality scores of 30 and >99% read accuracy. Additionally, nCats in paired-end mode can overcome issues related to calling single-nucleotide variants (Gilpatrick et al., 2020). To further increase the coverage of target DNA, nanopore reads can be enriched by software control (Martin et al., 2022) The ability for live-base-calling of nanopore reads enables the enrichment of target DNA by controlling individual pores of the nanopore flow cell to keep or eject DNA based on the current sequence information. Although flow cells can be in principal used several times for sequencing, the data output dramatically decreases after each run. To enable the simultaneous identification of transgene integration sites, the integrity of the DNA and the methylation pattern in one sequencing run, nCats can be further improved by adjusting steps in the experimental set up. In this study, we prepared libraries with crRNAs targeting the transgene at one site for one sequencing run and prepared another for the same site but in opposing direction. Principally, mixing two separate Cas9 digests treated with either sense or anti-sense crRNAs close to the 5'- or 3'-end of the integrated expression vector, respectively, and proceeding with the sequencing work flow should generate chimeric reads for both junction sites. Those reads would provide all necessary information with a single sequencing run. Thus, the already rapid data generation (around 24h from isolating HMW DNA to first sequence results) could be further expedited while reducing sequencing costs.

6 Acknowledgements

KL, KM and NB gratefully acknowledge support by the COMET-Funding Program managed by the Austrian Research Promotion Agency FFG. Part of this study was financed by the Austrian Center of Industrial Biotechnology, a COMET K2 competence center supported by the Austrian Research Promotion Agency FFG.

The authors would like to express their gratitude to Nancy Stralis-Pavese who helped to establish the protocols for HMW DNA isolation and setting up the MinION sequencer. We would like to thank Heinz Himmelbauer for initially providing the MinION sequencer. We would also like to thank Martina Baumann for her support on the TLA-Sequencing report.

7 Author Contributions

Conceptualization: NM and NB, Acquisition of Data: NM, KL, Methodology: NM and KL, Analysis: NM, KL, KM, Writing: NM, KL, NB.

8 Raw data availability

Raw data data can be accessed at EuropeanNucleotide Archive (ENA) under project code PRJEB57448.

9 Conflict of interest

The authors declare no financial or commercial conflict of interest.

References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Baik, J. Y., & Lee, K. H. (2017). Growth Rate Changes in CHO Host Cells Are Associated with Karyotypic Heterogeneity. *Biotechnology Journal*, *13*(3), 1700230. https://doi.org/10.1002/biot.201700230

Bandyopadhyay, A. A., O'Brien, S. A., Zhao, L., Fu, H.-Y., Vishwanathan, N., & Hu, W.-S. (2019). Recurring genomic structural variation leads to clonal instability and loss of productivity. *Biotechnology and Bioengineering*, *116*(1), 41–53. https://doi.org/10.1002/bit.26823

Baumann, M., Gludovacz, E., Sealover, N., Bahr, S., George, H., Lin, N., Kayser, K., & Borth, N. (2017). Preselection of recombinant gene integration sites enabling high transcription rates in CHO cells using alternate start codons and recombinase mediated cassette exchange. *Biotechnology and Bioengineering*, *114*(11), 2616–2627. https://doi.org/10.1002/bit.26388

Bydlinski, N., Maresch, D., Schmieder, V., Klanert, G., Strasser, R., & Borth, N. (2018). The contributions of individual galactosyltransferases to protein specific N-glycan processing in Chinese Hamster Ovary cells. *Journal of Biotechnology*, *282*, 101–110. https://doi.org/10.1016/j.jbiotec.2018.07.015

Chang, M., Kumar, A., Kumar, S., Huhn, S., Timp, W., Betenbaugh, M., & Du, Z. (2022). Epigenetic Comparison of CHO Hosts and Clones Reveals Divergent Methylation and Transcription Patterns Across Lineages. *Biotechnology and Bioengineering*. https://doi.org/10.1002/bit.28036

Chao, K.-H., Zimin, A. V., Pertea, M., & Salzberg, S. L. (2022). *The first gapless, reference-quality, fully annotated genome from a Southern Han Chinese individual* (p. 2022.08.08.503226). bioRxiv. https://doi.org/10.1101/2022.08.08.503226

Concordet, J.-P., & Haeussler, M. (2018). CRISPOR: Intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research*, *46*(W1), W242–W245. https://doi.org/10.1093/nar/gky354

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

de Vree, P. J. P., de Wit, E., Yilmaz, M., van de Heijning, M., Klous, P., Verstegen, M. J. A. M., Wan, Y., Teunissen, H., Krijger, P. H. L., Geeven, G., Eijk, P. P., Sie, D., Ylstra, B., Hulsman, L. O. M., van Dooren, M. F., van Zutven, L. J. C. M., van den Ouweland, A., Verbeek, S., van Dijk, K. W., … de Laat, W. (2014). Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping. *Nature Biotechnology*, *32*(10), Article 10. https://doi.org/10.1038/nbt.2959

Dhiman, H., Campbell, M., Melcher, M., Smith, K. D., & Borth, N. (2020). Predicting favorable landing pads for targeted integrations in Chinese hamster ovary cell lines by learning stability characteristics from random transgene integrations. *Computational and Structural Biotechnology Journal*, *18*, 3632–3648. https://doi.org/10.1016/j.csbj.2020.11.008

Feichtinger, J., Hernández, I., Fischer, C., Hanscho, M., Auer, N., Hackl, M., Jadhav, V., Baumann, M., Krempl, P. M., Schmidl, C., Farlik, M., Schuster, M., Merkel, A., Sommer, A., Heath, S., Rico, D., Bock, C., Thallinger, G. G., & Borth, N. (2016). Comprehensive genome and epigenome characterization of CHO cells in response to evolutionary pressures and over time.

Biotechnology and Bioengineering, *113*(10), 2241–2253. Scopus. https://doi.org/10.1002/bit.25990

Gilpatrick, T., Lee, I., Graham, J. E., Raimondeau, E., Bowen, R., Heron, A., Downs, B., Sukumar, S., Sedlazeck, F. J., & Timp, W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology*, *38*(4), Article 4. https://doi.org/10.1038/s41587-020-0407-5

Grav, L. M., Sergeeva, D., Lee, J. S., Marin de Mas, I., Lewis, N. E., Andersen, M. R., Nielsen, L. K., Lee, G. M., & Kildegaard, H. F. (2018). Minimizing Clonal Variation during Mammalian Cell Line Engineering for Improved Systems Biology Data Generation. *ACS Synthetic Biology*. https://doi.org/10.1021/acssynbio.8b00140

Hansen, K. D., Langmead, B., & Irizarry, R. A. (2012). BSmooth: From whole genome bisulfite sequencing reads to differentially methylated regions. *Genome Biology*, *13*(10), R83. https://doi.org/10.1186/gb-2012-13-10-r83

Hernandez, I., Dhiman, H., Klanert, G., Jadhav, V., Auer, N., Hanscho, M., Baumann, M., Esteve-Codina, A., Dabad, M., Gómez, J., Alioto, T., Merkel, A., Raineri, E., Heath, S., Rico, D., & Borth, N. (2019). Epigenetic regulation of gene expression in Chinese Hamster Ovary cells in response to the changing environment of a batch culture. *Biotechnology and Bioengineering*, *116*(3), 677–692. https://doi.org/10.1002/bit.26891

Hilliard, W., MacDonald, M. L., & Lee, K. H. (2020). Chromosome-scale scaffolds for the Chinese hamster reference genome assembly to facilitate the study of the CHO epigenome. *Biotechnology and Bioengineering*, *117*(8), 2331–2339. https://doi.org/10.1002/bit.27432

Höijer, I., Johansson, J., Gudmundsson, S., Chin, C.-S., Bunikis, I., Häggqvist, S., Emmanouilidou, A., Wilbe, M., den Hoed, M., Bondeson, M.-L., Feuk, L., Gyllensten, U., & Ameur, A. (2020).

Amplification-free long-read sequencing reveals unforeseen CRISPR-Cas9 off-target activity. *Genome Biology*, *21*(1), 290. https://doi.org/10.1186/s13059-020-02206-w

Hsu, P. D., Scott, D. A., Weinstein, J. A., Ran, F. A., Konermann, S., Agarwala, V., Li, Y., Fine, E. J., Wu, X., Shalem, O., Cradick, T. J., Marraffini, L. A., Bao, G., & Zhang, F. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. *Nature Biotechnology*, *31*(9), 827–832. https://doi.org/10.1038/nbt.2647

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*(5), Article 5. https://doi.org/10.1038/s41587-019-0072-8

Lee, J. S., Grav, L. M., Pedersen, L. E., Lee, G. M., & Kildegaard, H. F. (2016). Accelerated homology-directed targeted integration of transgenes in Chinese hamster ovary cells via CRISPR/Cas9 and fluorescent enrichment. *Biotechnology and Bioengineering*, *113*(11), 2518–2523. https://doi.org/10.1002/bit.26002

Lee, J. S., Kallehauge, T. B., Pedersen, L. E., & Kildegaard, H. F. (2015). Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Scientific Reports*, *5*, 8572. https://doi.org/10.1038/srep08572

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Martin, S., Heavens, D., Lan, Y., Horsfield, S., Clark, M. D., & Leggett, R. M. (2022). Nanopore adaptive sampling: A tool for enrichment of low abundance species in metagenomic samples. *Genome Biology*, *23*(1), 11. https://doi.org/10.1186/s13059-021-02582-x

Marx, N., Dhiman, H., Schmieder, V., Freire, C. M., Nguyen, L. N., Klanert, G., & Borth, N. (2021). Enhanced targeted DNA methylation of the CMV and endogenous promoters with dCas9-DNMT3A3L entails distinct subsequent histone modification changes in CHO cells. *Metabolic Engineering*, *66*, 268–282. https://doi.org/10.1016/j.ymben.2021.04.014

Marx, N., Eisenhut, P., Weinguny, M., Klanert, G., & Borth, N. (2022). How to train your cell—Towards controlling phenotypes by harnessing the epigenome of Chinese hamster ovary production cell lines. *Biotechnology Advances*, *56*, 107924. https://doi.org/10.1016/j.biotechadv.2022.107924

Marx, N., Grünwald-Gruber, C., Bydlinski, N., Dhiman, H., Ngoc Nguyen, L., Klanert, G., & Borth, N. (2018). CRISPR-Based Targeted Epigenetic Editing Enables Gene Expression Modulation of the Silenced Beta-Galactoside Alpha-2,6-Sialyltransferase 1 in CHO Cells. *Biotechnology Journal*, *13*(10), e1700217. https://doi.org/10.1002/biot.201700217

Nguyen, L. N., Baumann, M., Dhiman, H., Marx, N., Schmieder, V., Hussein, M., Eisenhut, P., Hernández, I., Koehn, J., & Borth, N. (2019). *Novel promoters derived from Chinese Hamster Ovary cells via in silico and in vitro analysis (accepted article).*

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Payne, A., Holmes, N., Rakyan, V., & Loose, M. (2019). BulkVis: A graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics (Oxford, England)*, *35*(13), 2193–2198. https://doi.org/10.1093/bioinformatics/bty841

Pristovšek, N., Nallapareddy, S., Grav, L. M., Hefzi, H., Lewis, N. E., Rugbjerg, P., Hansen, H. G., Lee, G. M., Andersen, M. R., & Kildegaard, H. F. (2019). Systematic Evaluation of Site-Specific Recombinant Gene Expression for Programmable Mammalian Cell Engineering. *ACS Synthetic Biology*, *8*(4), 758–774. https://doi.org/10.1021/acssynbio.8b00453

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative Genomics Viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Rupp, O., MacDonald, M. L., Li, S., Dhiman, H., Polson, S., Griep, S., Heffner, K., Hernandez, I., Brinkrolf, K., Jadhav, V., Samoudi, M., Hao, H., Kingham, B., Goesmann, A., Betenbaugh, M. J., Lewis, N. E., Borth, N., & Lee, K. H. (2018). A reference genome of the Chinese hamster based on a hybrid assembly strategy. *Biotechnology and Bioengineering*, *115*(8), 2087–2100. https://doi.org/10.1002/bit.26722

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Schmieder, V., Bydlinski, N., Strasser, R., Baumann, M., Kildegaard, H. F., Jadhav, V., & Borth, N. (2018). Enhanced Genome Editing Tools For Multi-Gene Deletion Knock-Out Approaches Using Paired CRISPR sgRNAs in CHO Cells. *Biotechnology Journal*, *13*(3), 1700211. https://doi.org/10.1002/biot.201700211

Schmieder, V., Fieder, J., Drerup, R., Gutierrez, E. A., Guelch, C., Stolzenberger, J., Stumbaum, M., Mueller, V. S., Higel, F., Bergbauer, M., Bornhoefft, K., Wittner, M., Gronemeyer, P., Braig, C., Huber, M., Reisenauer-Schaupp, A., Mueller, M. M., Schuette, M., Puengel, S., … Fischer, S. (2022). Towards maximum acceleration of monoclonal antibody development: Leveraging transposase-mediated cell line generation to enable GMP manufacturing within 3 months using a stable pool. *Journal of Biotechnology*, *349*, 53–64. https://doi.org/10.1016/j.jbiotec.2022.03.010

Sergeeva, D., Lee, G. M., Nielsen, L. K., & Grav, L. M. (2020). Multicopy Targeted Integration for Accelerated Development of High-Producing Chinese Hamster Ovary Cells. *ACS Synthetic Biology*, *9*(9), 2546–2561. https://doi.org/10.1021/acssynbio.0c00322

Shafin, K., Pesout, T., Lorig-Roach, R., Haukness, M., Olsen, H. E., Bosworth, C., Armstrong, J., Tigyi, K., Maurer, N., Koren, S., Sedlazeck, F. J., Marschall, T., Mayes, S., Costa, V., Zook, J. M., Liu, K. J., Kilburn, D., Sorensen, M., Munson, K. M., … Paten, B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology*, *38*(9), Article 9. https://doi.org/10.1038/s41587-020-0503-6

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, *14*(4), Article 4. https://doi.org/10.1038/nmeth.4184

Stadermann, A., Gamer, M., Fieder, J., Lindner, B., Fehrmann, S., Schmidt, M., Schulz, P., & Gorr, I. H. (2022). Structural analysis of random transgene integration in CHO manufacturing cell lines

by targeted sequencing. *Biotechnology and Bioengineering*, *119*(3), 868–880. https://doi.org/10.1002/bit.28012

Stothard, P. (2000). The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques*, *28*(6), 1102, 1104.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Vcelar, S., Jadhav, V., Melcher, M., Auer, N., Hrdina, A., Sagmeister, R., Heffner, K., Puklowski, A., Betenbaugh, M., Wenger, T., Leisch, F., Baumann, M., & Borth, N. (2017). Karyotype variation of CHO host cell lines over time in culture characterized by chromosome counting and chromosome painting. *Biotechnology and Bioengineering*, *115*(1), 165–173. https://doi.org/10.1002/bit.26453

Vcelar, S., Melcher, M., Auer, N., Hrdina, A., Puklowski, A., Leisch, F., Jadhav, V., Wenger, T., Baumann, M., & Borth, N. (2018). Changes in Chromosome Counts and Patterns in CHO Cell Lines upon Generation of Recombinant Cell Lines and Subcloning. *Biotechnology Journal*, *13*(3), 1700495. https://doi.org/10.1002/biot.201700495

Weinguny, M., Eisenhut, P., Klanert, G., Virgolini, N., Marx, N., Jonsson, A., Ivansson, D., Lövgren, A., & Borth, N. (2020). Random epigenetic modulation of CHO cells by repeated knockdown of DNA methyltransferases increases population diversity and enables sorting of cells with higher production capacities. *Biotechnology and Bioengineering*, *117*(11), 3435–3447. https://doi.org/10.1002/bit.27493

Weinguny, M., Klanert, G., Eisenhut, P., Lee, I., Timp, W., & Borth, N. (2021). Subcloning induces changes in the DNA-methylation pattern of outgrowing Chinese hamster ovary cell colonies. *Biotechnology Journal*, e2000350. https://doi.org/10.1002/biot.202000350

Yamane-Ohnuki, N., Kinoshita, S., Inoue-Urakubo, M., Kusunoki, M., Iida, S., Nakano, R., Wakitani, M., Niwa, R., Sakurada, M., Uchida, K., Shitara, K., & Satoh, M. (2004). Establishment of FUT8 knockout Chinese hamster ovary cells: An ideal host cell line for producing completely defucosylated antibodies with enhanced antibody-dependent cellular cytotoxicity. *Biotechnology and Bioengineering*, *87*(5), 614–622. https://doi.org/10.1002/bit.20151

Zhao, M., Wang, J., Luo, M., Luo, H., Zhao, M., Han, L., Zhang, M., Yang, H., Xie, Y., Jiang, H., Feng, L., Lu, H., & Zhu, J. (2018). Rapid development of stable transgene CHO cell lines by CRISPR/Cas9-mediated site-specific integration into C12orf35. *Applied Microbiology and Biotechnology*, *102*(14), 6105–6117. https://doi.org/10.1007/s00253-018-9021-6

Zimin, A. V., Shumate, A., Shinder, I., Heinz, J., Puiu, D., Pertea, M., & Salzberg, S. L. (2022). A reference-quality, fully annotated genome from a Puerto Rican individual. *Genetics*, *220*(2), iyab227. https://doi.org/10.1093/genetics/iyab227

Figure 1: Overview of the nanopore sequencing technology and nCats applications. A) Illustration of a nanopore sequencing a DNA strand. B) nCats library preparation workflow. DNA is dephosphorylated after isolation. Cas9 cleavage at crRNA target sites (green) results in newly phosphorylated 5'-ends. Sequencing adapters are then ligated to these fragments and allow directional nanopore sequencing of a selected strand. C) Different nCats approaches for single-end or paired-end sequencing. Depending on the positioning of the Cas9-complex, DNA strands can be directionally sequenced with undefined (single-end) or defined (paired-end) length.

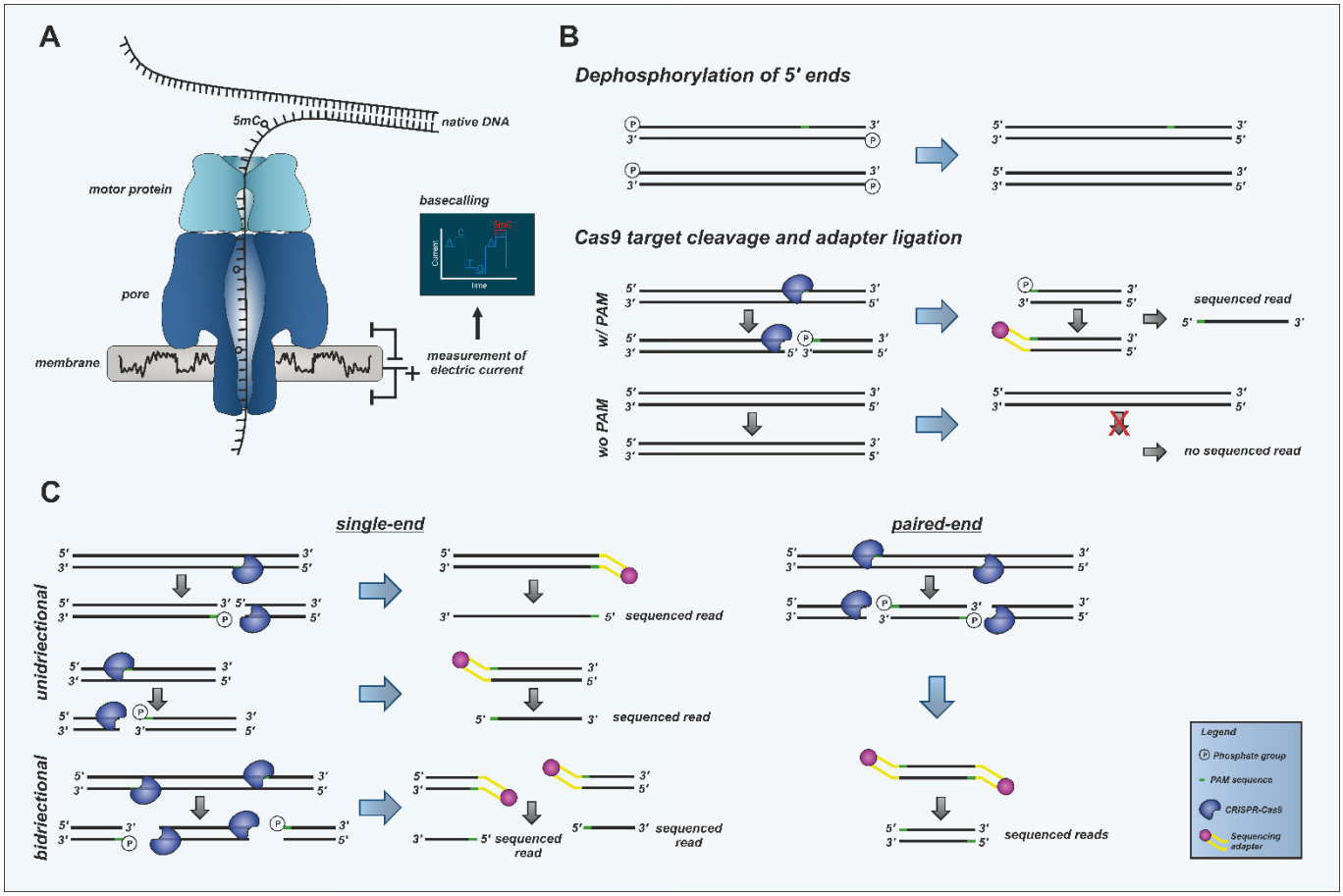
Figure 2: nCats of B4galt1 and the Fut8 promoter A)-B) Sequencing strategy of the B4galt1 gene between exon 1 and 2 and of the Fut8 promoter region C) Promoter methylated (Fut8neg) or promoter demethylated (Fut8pos) cell pools were sorted (Marx et al., 2021) and used for nCats experiments. D)-E) Per base coverage of the targeted B4galt1 gene or the Fut8 promoter region. F) DNA methylation frequency called by nCats of Fut8neg and Fut8pos cell pools and WGBS data (Weinguny et al., 2021) of the same genomic region. G)- H) Correlation of DNA methylation frequency between WGBS and nCats data for the B4galt1 and Fut8 promoter region. I) DNA methylation frequency of the Fut8 promoter region as measured with nCats and targeted BS-Seq (Marx et al., 2021). J) Run statistics summary of nCats run ID 1-3. K) Mean methylation frequency of the Fut8 promoter region, CpG Island and the region (amplicon) used for targeted BS-Seq in (Marx et al., 2021).

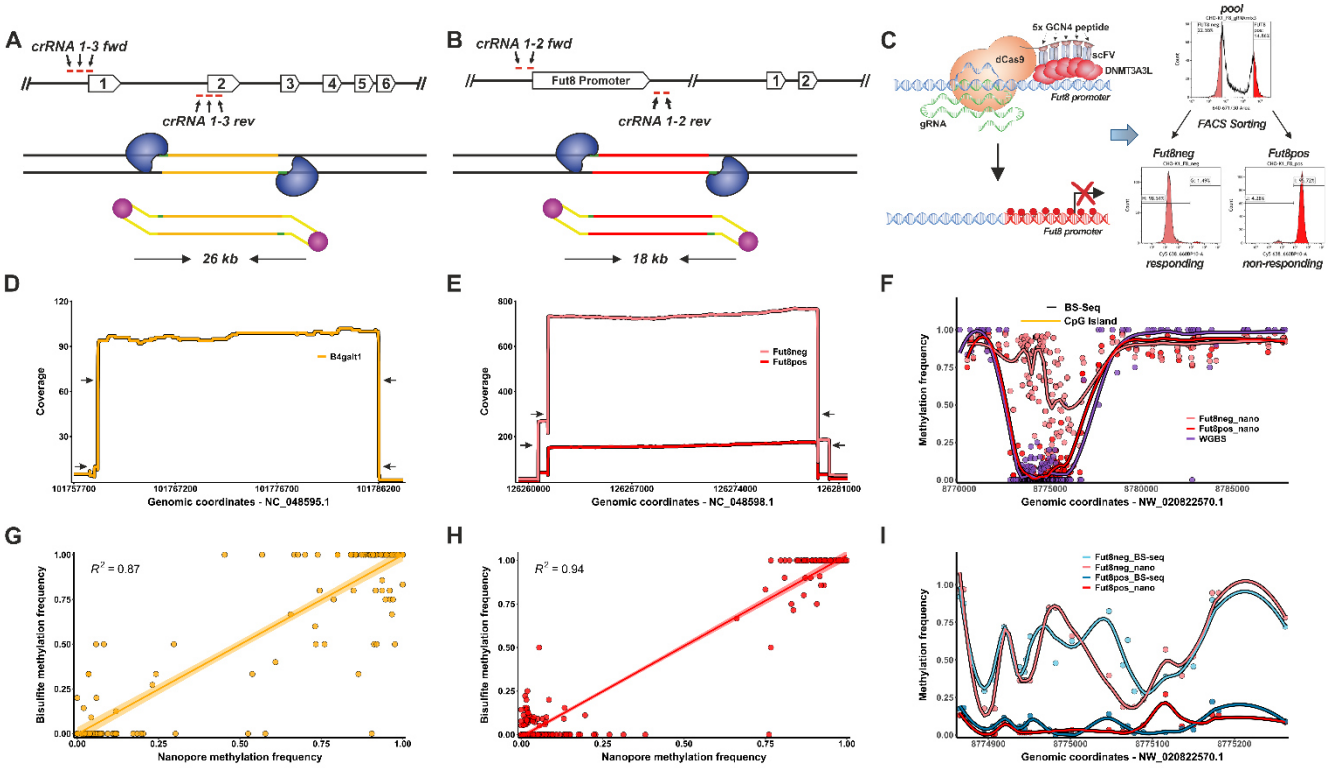
Figure 3: nCats reliably identifies junction sites of integrated transgenes. A) Cell lines CHO-CD4 and CHO-Herc were used for integration site analysis. CHO-CD4 was subjected to TLA-Seq, which identified integration site A and B in this cell line. Only the transgene sequence of ISA could be resolved. CHO-Herc was generated by RMCE of the CD4-EV by the Herceptin-EV. Integration sites of this cell line were unknown beforehand. B) Strategy for nCats sequencing to identify the transgene-genome junction sites of CHO-CD4 and CHO-Herc cell lines. crRNAs were designed to target the EV so that sequencing

commence from the transgene vector into the genome. C) Coverage plot at JS1 and JS2 of the CHO-CD4 and CHO-Herc cell line as well as Sanger sequencing data at the breakage points between the transgene vector and the genome of the CHO-CD4 cell line. D) Coverage plot at JS3 of the CHO-CD4 and CHO-Herc cell line as well as Sanger sequencing data at the breakage points between the transgene vector and the genome of the CHO-CD4 cell line. E) Coverage plot at JS4 and JS5 of the CHO-CD4 and CHO-Herc cell line as well as Sanger sequencing data at the breakage points between the transgene vector and the genome of the CHO-CD4 cell line. F) Coverage plot at JS6 of the CHO-CD4 cell line as well as Sanger sequencing data at the breakage points between the transgene vector and the genome of the CHO-CD4 cell line.

Figure 4: Identification of integration sites, the structure of the integrated transgene vectors as well their DNA methylation status by nCats. A) The identified transgene vectors and their respective elements for integration site A, B, and C of CHO-CD4 and CHO-Herc as identified with nCats. B) crRNAs were designed to bind at the genomic portion of 3'-junction sites, resulting in sequencing from these regions over the transgene into the genomic regions of the corresponding 5'-junction sites. C) Comparison between TLA-seq and nCats results for the CHO-CD4 cell line. D) Coverage plots after nCats with reads starting at JS2 for CHO-CD4 and CHO-Herc. JS1 and JS2 are marked with a black dot and the position and direction of the crRNAs used are indicated by an arrow. E) Close-up of the deleted sequence between JS1 and JS2 into which the transgene vectors are integrated. F) Coverage plots after nCats with reads starting at JS4 for CHO-CD4 and CHO-Herc. JS4 and JS5 are marked with a black dot and the position and direction of the crRNAs used are indicated by arrows. The crRNAs were binding twice (additional crRNA binding site indicated by an asterisk) since the parts of the genomic sequence of JS4 was duplicated and inverted. The coverage plateaus marked with α, β and γ. G) Coverage plots after nCats with reads starting at JS3 for CHO-CD4 and CHO-Herc. JS3 is marked with a black dot and the position and direction of the crRNAs used are indicated by an arrow. Dashed arrows show a large deletion in chromosome 1 as well as drop in coverage at the transgene integration site (tilted arrow). H) Coverage plots of chimeric reads generated from nCats sequencing from the transgene into the genome at integration site B of CHO-CD4. I) Coverage plots of chimeric reads generated from nCats sequencing from the transgene into the genome at integration site C of CHO-CD4. J) DNA methylation frequency of ISA of the CHO-CD4 cell line. The positions of the transgene and the CMV promoter are indicated below. K) DNA methylation frequency of ISA of the CHO-Herc cell line. The positions of the transgene and the CMV

promoter are indicated below. L) DNA methylation frequency of ISB of the CHO-CD4 cell line. The positions of the transgenes and the CMV promoter are indicated below. M) DNA methylation frequency of ISB of the CHO-Herc cell line. The positions of the transgene and the CMV promoter are indicated below.





J

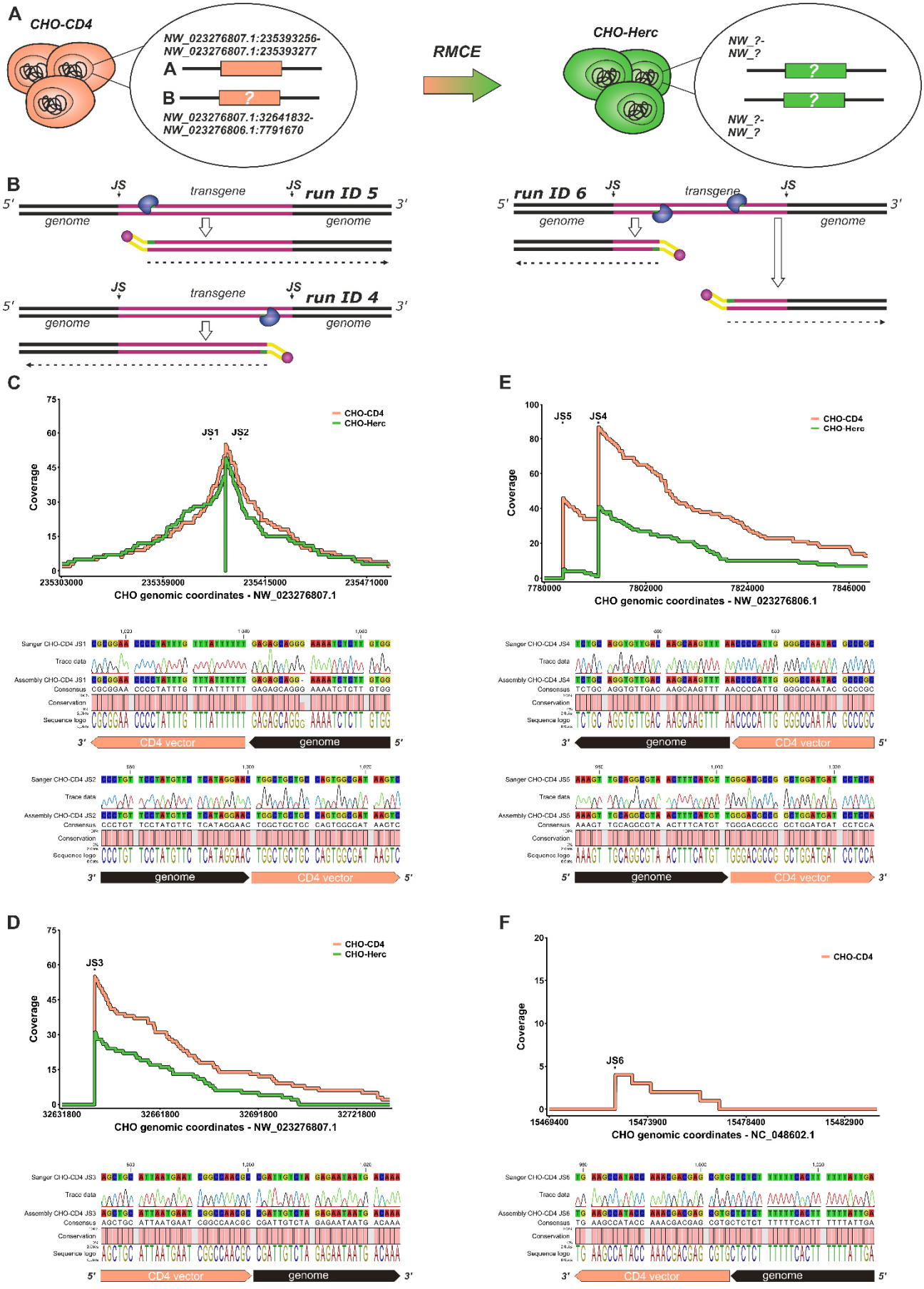
nCats stats	*B4galt1*	*Fut8pos**	*Fut8neg*
run ID	1	2	3
# aligned reads	85102	45979	243748
# on-target reads	120	232	847
on-target enrichment [%]	0.14	0.50	0.34
median coverage	95	192	724

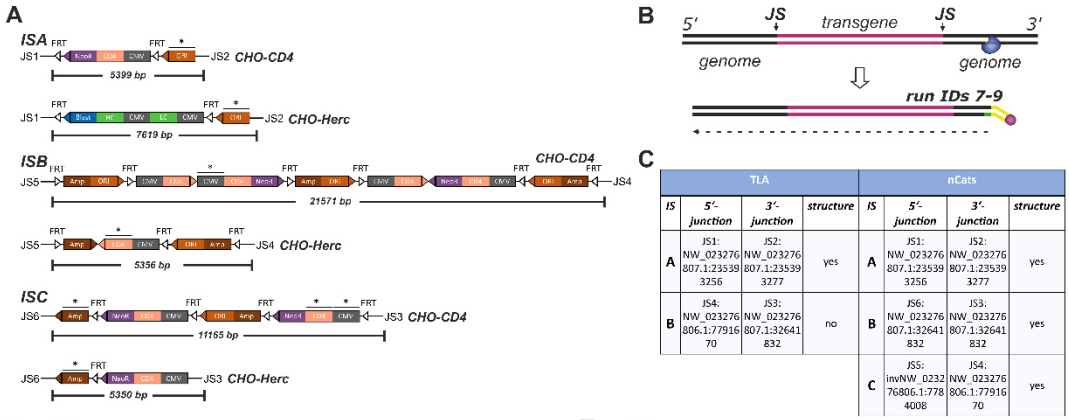
* re-used flow cell

K

Region[A]	*Coordinates*	*CHO-K1 wt*	*Fut8pos*	*Fut8neg*
Promoter	8772352-8776745	0.027[B]	0.042	0.736
CpG Island	8773435-8775516	0.023[B]	0.028	0.759
Amplicon BS-Seq	8774839-8775295	0.006[A]	0.073[A]	0.569[A]
Amplicon Nanopore		-	0.057	0.508

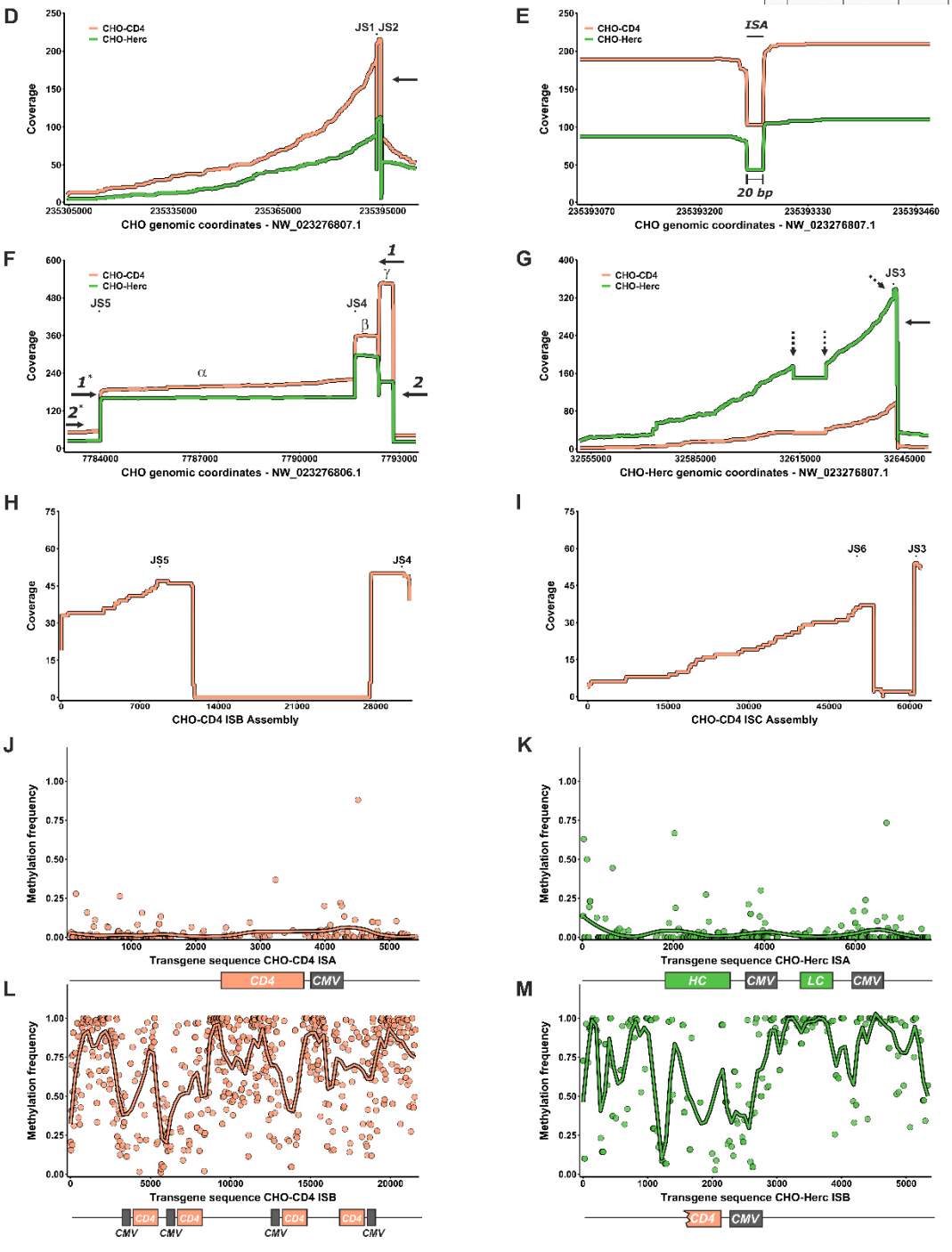
[A] Data from Marx et al., 2021, [B] data from Weinguny et al., 2020


A
CHO-CD4
NW_023276807.1:235393256-
NW_023276807.1:235393277
A
B
?
NW_023276807.1:32641832-
NW_023276806.1:7791670
RMCE
CHO-Herc
NW_?-
NW_?
?
?
NW_?-
NW_?
B
JS
transgene
JS
run ID 5
5'
3'
genome
genome
run ID 4
run ID 6
C
CHO-CD4
CHO-Herc
JS1
JS2
Coverage
CHO genomic coordinates - NW_023276807.1
Sanger CHO-CD4 JS1
Trace data
Assembly CHO-CD4 JS1
Consensus
Conservation
Sequence logo
CD4 vector
genome
Sanger CHO-CD4 JS2
Assembly CHO-CD4 JS2
E
JS5
JS4
CHO genomic coordinates - NW_023276806.1
Sanger CHO-CD4 JS4
Assembly CHO-CD4 JS4
Sanger CHO-CD4 JS5
Assembly CHO-CD4 JS5
D
JS3
CHO genomic coordinates - NW_023276807.1
Sanger CHO-CD4 JS3
Assembly CHO-CD4 JS3
F
JS6
CHO genomic coordinates - NC_048602.1
Sanger CHO-CD4 JS6
Assembly CHO-CD4 JS6



C

TLA				nCats			
IS	5′-junction	3′-junction	structure	IS	5′-junction	3′-junction	structure
A	JS1: NW_023276807.1:23539 3256	JS2: NW_023276807.1:23539 3277	yes	A	JS1: NW_023276807.1:23539 3256	JS2: NW_023276807.1:23539 3277	yes
B	JS4: NW_023276806.1:77916 70	JS3: NW_023276807.1:32641 832	no	B	JS6: NW_023276807.1:32641 832	JS3: NW_023276807.1:32641 832	yes
				C	JS5: invNW_0232 76806.1:778 4008	JS4: NW_023276806.1:77916 70	yes



10 Supporting Information

Supplementary Table 1: Sequences used for nCats library preparation, PCR, qPCR and Sanger sequencing applications.

crRNA target region	Name	Sequence (5'-3')
B4galt1		
	crRNA1_B4_fwd	GTTTCCTCTACACTTCGGAG
	crRNA2_B4_fwd	GTGTCGGCTTTCCGCCGCTG
	crRNA3_B4_fwd	GAGCAGCTCGTTGAACGAGG
	crRNA1_B4_rev	GACCCACCATGCCACAGACC
	crRNA2_B4_rev	CATGGGGCCAACTAGAGAGG
	crRNA3_B4_rev	GCTCCTGCCGGTTGCGAAAT
FUT8 promoter		
	crRNA1_F8_fwd	ACAGTTGTTACACGGCAGAA
	crRNA2_F8_fwd	AAATCCTCTGGAACCACGTG
	crRNA1_F8_rev	GCCCAACTGCAAACATGCGG
	crRNA2_F8_rev	AGGCTGATTTATTCATGCCG
CMV promoter		
	crRNA1_CMV_fwd	GCCCCATTGACGCAAATGGG
	crRNA2_CMV_fwd	GTTCCCATAGTAACGCCAAT
	crRNA1_CMV_rev	GTCCCTATTGGCGTTACTAT
	crRNA2_CMV_rev	GTACTGGGCATAATGCCAGG
	crRNA3_CMV_rev	GCCCATTTGCGTCAATGGGG
CD4 coding region		
	crRNA1_CD4_fwd	GTCGACCCCGGTGCAGCCAA
Herceptin coding region		
	crRNA1_Herc_fwd	CAATTGGTACGTGGACGGCG
	crRNA2_Herc_fwd	GCTTCTACGCCATGGACTAT
Downstream of ISA		

	crRNA1_IS_A_rev	CAGGGATTACATTCTATTAA
	crRNA2_IS_A_rev	ACCTATCACTCGgaacaaaa
Downstream of ISB		
	crRNA1_IS_B_rev	TGCCGAAAGCATATTCATTA
	crRNA2_IS_B_rev	TGTGTCGGTTACAGAACATC
Downstream of ISC		
	crRNA1_IS_C_rev	GGAGAGAGATATCCTCCGCA
	crRNA2_IS_C_rev	CTATGAAGCCCTCACCAGAT
Application	**Name**	**Sequence (5'-3')**
Junction Site PCR	**CHO-CD4 cell line**	
	PCR_JS1_fwd	TCCACTCTTTCTCACCATTCTT
	PCR_JS1_rev	CTTCTATCGCCTTCTTGACGAG
	PCR_JS2_fwd	AGCGGTATCAGCTCACTCAAAG
	PCR_JS2_rev	AACACTATCCAGCAGAAACACT
	PCR_JS3_fwd	ACAGTAGGGGTCACAATACTAG
	PCR_JS3_rev	TGCATGGCGGTAATACGGTTTG
	PCR_JS4_fwd	GCTCCTTCTCAGGCTTATATAC
	PCR_JS4_rev	TGCTTATATAGACCTCCCACCG
	PCR_JS5_fwd	TGACGAGATGCAAGAGTGGAAC
	PCR_JS5_rev	TGCTTATATAGACCTCCCACCG
	PCR_JS6_fwd	TTACGGATGGCATGACAGTAAG
	PCR_JS6_rev	TTGCTGAGGAGTGACAGATAGG
	CHO-Herc cell line	
	PCR_JS1_fwd	TCCACTCTTTCTCACCATTCTT
	PCR_JS1_rev	GTGCTTCTCGATCTGCATCCT
	PCR_JS2_fwd	TGCTGCTGGCAGTAGTAGGTG
	PCR_JS2_rev	AACACTATCCAGCAGAAACACT

	PCR_JS3_fwd	TGCATGGCGGTAATACGGTTTG
	PCR_JS3_rev	ACAGTAGGGGTCACAATACTAG
	PCR_JS4_fwd	GCTCCTTCTCAGGCTTATATAC
	PCR_JS4_rev	TGCTTATATAGACCTCCCACCG
	PCR_JS5_fwd	TTGTCAGAACAAAAGCCGTGAC
	PCR_JS5_rev	GAAGCTAGAGTAAGTAGTTCGC
	PCR_JS6_fwd	TTGCTGAGGAGTGACAGATAGG
	PCR_JS6_rev	TTACGGATGGCATGACAGTAAG
Sanger Sequencing		
	CHO-CD4 cell line	
	PCR_JS1_fwd	TCCACTCTTTCTCACCATTCTT
	PCR_JS2_rev	AACACTATCCAGCAGAAACACT
	PCR_JS3_rev	TGCATGGCGGTAATACGGTTTG
	PCR_JS4_fwd	GCTCCTTCTCAGGCTTATATAC
	PCR_JS5_fwd	TGACGAGATGCAAGAGTGGAAC
	PCR_JS6_fwd	TTACGGATGGCATGACAGTAAG
	CHO-Herc cell line	
	PCR_JS1_rev	GTGCTTCTCGATCTGCATCCT
	PCR_JS2_rev	AACACTATCCAGCAGAAACACT
	PCR_JS3_rev	ACAGTAGGGGTCACAATACTAG
	PCR_JS4_fwd	GCTCCTTCTCAGGCTTATATAC
	PCR_JS5_fwd	TTGTCAGAACAAAAGCCGTGAC
	PCR_JS6_rev	TTACGGATGGCATGACAGTAAG
qPCR of JS3	**CHO-CD4 and CHO-Herc**	
	EV - Genome	
	qPCR_JS3_EV_fwd	CTCACATTAATTGCGTTGCG

	qPCR_JS3_genome_rev	TTTGAACATACATGGACTGC
	Genome - Genome	
	qPCR_JS3_gnome_fwd	GTGCTATCAAAAGCTGAAGG
	qPCR_JS3_genome_rev	TTTGAACATACATGGACTGC

Supplementary Figure 1: Coverage of short reads mapping to the CD4 expression vector after nCats of the CHO-CD4 cell line. Arrows indicate the position and direction of crRNAs used as designed for the transfected CD4 expression vector.

Supplementary Figure 2: Coverage plots after nCats with reads starting at JS3 for CHO-CD4 and CHO-Herc. By changing the settings of the coverage function of the GenomicAlignments R package from ('drop.D.ranges=FALSE') to ('drop.D.ranges=TRUE'), the deletion of ~9kb becomes apparent at ISC.

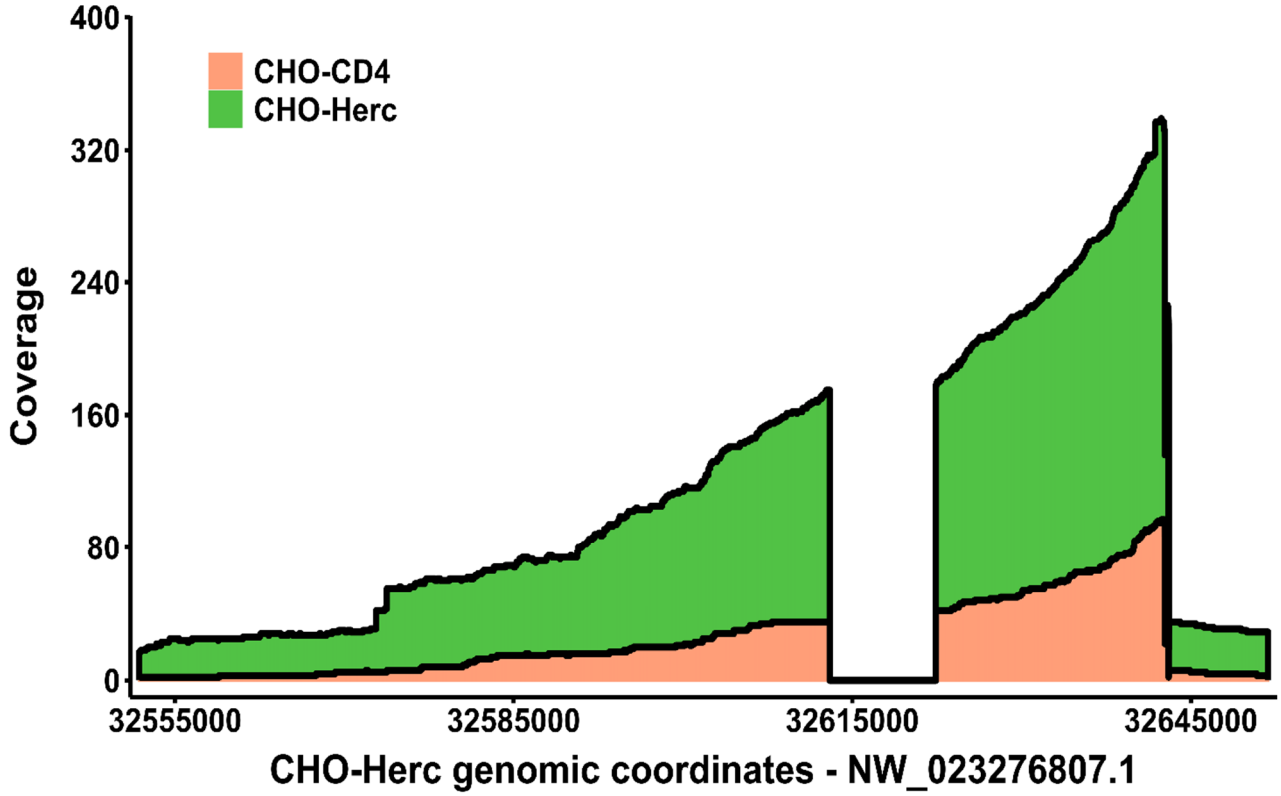
Supplementary Figure 3: Genomic qPCR of JS3 of the CHO-CD4 and CHO-Herc and the parental CHO-K1 cell line. Primers targeted either the breakpoint between the expression vector and the genome (EV-genome) or only the genomic portion (genome) of JS3. Glyceraldehyde 3-phosphate dehydrogenase was measured as a reference gene. Samples were measured in technical triplicates, error bars represent the standard deviation.

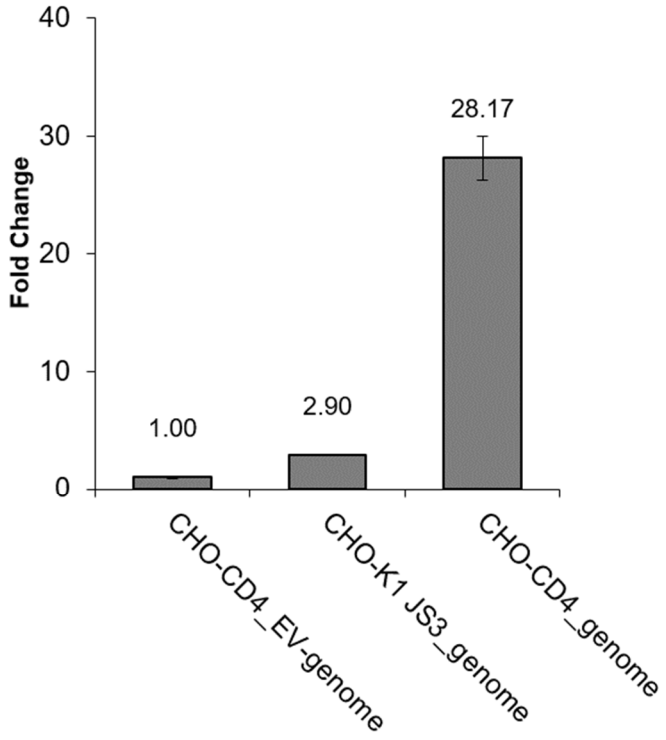
Supplementary Figure 4: Sequence of CHO-CD4 expression vector. Features of EV are highlighted by colors (CMV promoter in blue, CD4 coding region in salmon, the Kanamycin resistance gene in turquoise, the Ampicillin resistance gene in yellow, the origin of replication in pink and the FRT sites in red) and plasmid break points in large and bold bases.

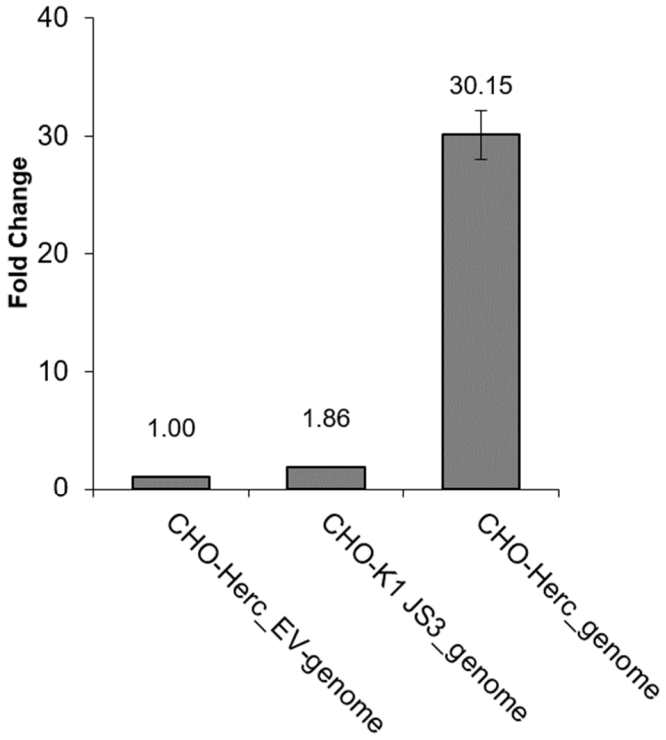
Supplementary Figure 5: Sequence of CHO-Herc expression vector. Features of EV are highlighted by colors and plasmid break points by large and bold letters. (The colors of CMV promoter, Ampicillin resistance gene, Origin of replication and FRT sites are identical to **Supplementary Figure 2**. The heavy and light antibody chain are highlighted in green and the blasticidin resistance gene in orange.)

Supplementary Figure 6: Sanger sequencing data at the breakage points between the expression vector and the genome of the CHO-Herc cell line.


250
200
150
100
50
0
Coverage
0
1500
3000
4500
6000
CD4 expression vector coordinates


400
320
240
160
80
0
Coverage
CHO-CD4
CHO-Herc
32555000
32585000
32615000
32645000
CHO-Herc genomic coordinates - NW_023276807.1





5'-

TAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAAC
GACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCAC
TTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGAC
TTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGA
TTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATG
GGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTGGTTTAGTGAACCGTCAGATCCGCTAGCGCTACCGGACTCAGATCTAATTCAAGCCCAGAG
CCCTGCCATTTCTGTGGGCTCAGGTCCCTACTGCTCAGCCCCTTCCTCCCTCGGCAAGGCCACACTGAACCGGGGAGTCCCTTTTAGGCACTTGCTTCTGGTGC
TGCAACTGGCGCTCCTCCCAGCAGCCACTCAGGGAAAGAAAGTGGTGCTGGGCAAAAAAGGGGATACAGTGGAACTGACCTGTACAGCTTCCCAGAAGAAGAG

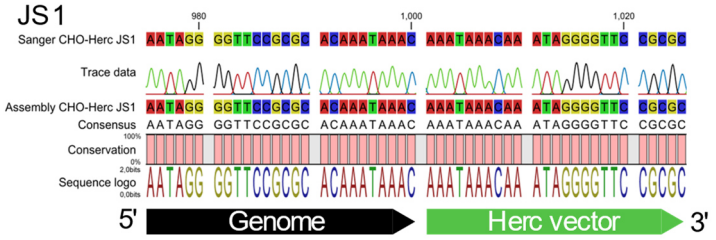
CATACAATTCCACTGGAAAAACTCCAACCAGATAAAGATTCTGGGAAATCAGGGCTCCTTCTTAACTAAAGGTCCATCCAAGCTGAATGATCGCGCTGACTCAA
GAAGAAGCCTTTGGGACCAAGGAAACTTCCCCCTGATCATCAAGAATCTTAAGATAGAAGACTCAGATACTTACATCTGTGAAGTGGAGGACCAGAAGGAGGAG
GTGCAATTGCTAGTGTTCGGATTGACTGCCAACTCTGACACCCACCTGCTTCAGGGGCAGAGCCTGACCCTGACCTTGGAGAGCCCCCCCTGGTAGTAGCCCCT
CAGTGCAATGTAGGAGTCCAAGGGGTAAAAACATACAGGGGGGGAAGACCCTCTCCGTGTCTCAGCTGGAGCTCCAGGATAGTGGCACCTGGACATGCACTGT
CTTGCAGAACCAGAAGAAGGTGGAGTTCAAAATAGACATCGTGGTGCTAGCTTTCCAGAAGGCCTCCAGCATAGTCTATAAGAAAGAGGGGGAACAGGTGGAGT
TCTCCTTCCCACTCGCCTTTACAGTTGAAAAGCTGACGGGCAGTGGCGAGCTGTGGTGGCAGGCGGAGAGGGCTTCCTCCTCCAAGTCTTGGATCACCTTTGA
CCTGAAGAACAAGGAAGTGTCTGTAAAATGGGTTACCCAGGACCCTAAGCTCCAGATGGGCAAGAAGCTCCCGCTCCACCTCACCCTGCCCCAGGCCTTGCCT
CAGTATGCTGGCTCTGGAAACCTCACCCTGGCCCTTGAAGCGAAAACAGGAAAGTTGCATCAGGAAGTGAACCTGGTGGTGATGAGAGCCACTCAGCTCCAGA
AAAATTTGACCTGTGAGGTGTGGGGACCCACCTCCCCTAAGCTGATGCTGAGCTTGAAACTGGAGAACAAGGAGGCAAAGGTCTCGAAGCGGGAGAAGGCGGT
GTGGGTGCTGAACCCTGAGGCGGGGATGTGGCAGTGTCTGCTGAGTGACTCGGGACAGGTCCTGCTGGAATCCAACATCAAGGTTCTGCCCACATGGTCGAC
CCCGGTGCAGCCAATGGCCCTGATTGTGCTGGGGGGCGTCGCCGGCCTCCTGCTTTTCATTGGGCTAGGCATCTTCTTCTGTGTCAGGTGCCGGCACTGAAGG
CGACCTGCAGCCCAAGCTTCGATCCAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTG
TGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGGTGTGGGAGGTTT
TTTAAAGCAAGTAAAACCTCTACAAATGTGGTATGGCTGATTATGATCCTGCCTCGCGCGTTTCGGTGCACTACGTGAACCATCACCCTAATCAAGTTTTTTGGGG
TCGAGGTGCCGTAAAGCACTAAATCGGAACCCTAAAGGGAGCCCCCGATTTAGAGCTTGACGGGGAAAGCCGGCGAACGTGGCGAGAAAGGAAGGGAAGAAA
GCGAAAGGAGCGGGCGCTAGGGCGCTGGCAAGTGTAGCGGTCACGCTGCGCGTAACCACCACACCCGCCGCGCTTAATGCGCCGCTACAGGGCGCGTCAGG
TGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAAT
AATATTGAAAAAGGAAGAGTCCTGAGGCGGAAAGAACCAGCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATG
CAAAGCATGCATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCAT
AGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGG
CCGCCTCGGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGGCTTTTGCAAAGATCGATCAAGAGACAGGATGAGGATCGTTTCGCAT
GATTGAACAAGATGGATTGCACGCAGGTTCTCCGGCCGCTTGGGTGGAGAGGCTATTCGGCTATGACTGGGCACAACAGACAATCGGCTGCTCTGATGCCGCC
GTGTTCCGGCTGTCAGCGCAGGGGCGCCCGGTTCTTTTTGTCAAGACCGACCTGTCCGGTGCCCTGAATGAACTGCAAGACGAGGCAGCGCGGCTATCGTGG
CTGGCCACGACGGGCGTTCCTTGCGCAGCTGTGCTCGACGTTGTCACTGAAGCGGGAAGGGACTGGCTGCTATTGGGCGAAGTGCCGGGGCAGGATCTCCTG
TCATCTCACCTTGCTCCTGCCGAGAAAGTATCCATCATGGCTGATGCAATGCGGCGGCTGCATACGCTTGATCCGGCTACCTGCCCATTCGACCACCAAGCGAA
ACATCGCATCGAGCGAGCACGTACTCGGATGGAAGCCGGTCTTGTCGATCAGGATGATCTGGACGAAGAGCATCAGGGGCTCGCGCCAGCCGAACTGTTCGC
CAGGCTCAAGGCGAGCATGCCCGACGGCGAGGATCTCGTCGTGACCCATGGCGATGCCTGCTTGCCGAATATCATGGTGGAAAATGGCCGCTTTTCTGGATTC
ATCGACTGTGGCCGGCTGGGTGTGGCGGACCGCTATCAGGACATAGCGTTGGCTACCCGTGATATTGCTGAAGAGCTTGGCGGCGAATGGGCTGACCGCTTC
CTCGTGCTTTACGGTATCGCCGCTCCCGATTCGCAGCGCATCGCCTTCTATCGCCTTCTTGACGAGTTCTTCTGAGCGGGACTCTGGGGTTCGAAATGACCGAC
CAAGCGACGCCCAACCTGCCATCACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCT
CCAGCGCGGGGATCTCATGCTGGAGTTCTTCGCCCACCCTAGGGGGAGGCTAACTGAAACACGGAAGGAGACAATACCGGAAGGAACCCGCGCTATGACGGC
AATAAAAAGACAGAATAAAACGCACGGTGTTGGGTCGTTTGTTCATAAACGCGGGGTTCGGTCCCAGGGCTGGCACTCTGTCGATACCCCACCGAGACCCCATT
GGGGCCAATACGCCCGCGTTTCTTCCTTTTCCCCACCCCACCCCCCAAGTTCGGGTGAAGGCCCAGGGCTCGCAGCCAACGTCGGGGCGGCAGGCCCTGCCA
TAGCCTCAGGTTACTCACGCGTCGATGATACCATGGAGAAGTTCCTATTCCGAAGTTCCTATTCTTCAAAAGGTATAGGAACTTCACTAGTCCCGATCCGTCGAC

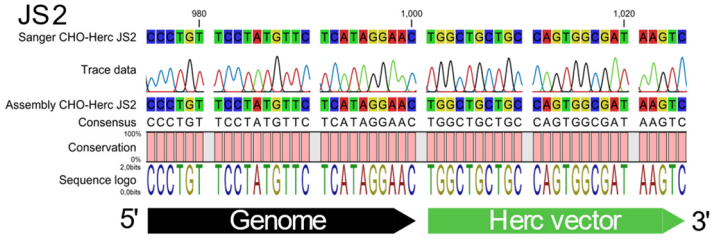
GTCAGGTGGCACTTTTCGGGGAAATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATG
CTTCAATAATATTGAAAAAGGAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAAC
GCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAG
AACGTTTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTC
AGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTG
CGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAG
CTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCC
CGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGG
TGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAA
CGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTA
ATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAA
GGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACT
CTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAG**A**TACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCC
TACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGC
AGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCA
CGCTTCCCGAAGGGAGAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTT
TATAGTCCTGTCGGGTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTT
ACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCT
CGCCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAA
TGCAGCTGGCACGACAGGTTTCCCGACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTAT
GCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTCTAGCTAGAGGTCGAGATCC
CCCGGGCTGCAGGAATTCATTGGCCCTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCCT
TCGCCAGGGGCTACCATGGAGAAGTTCCTATTCCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCAAGCTCTGGCACGACTGATCACGTATTACCGCCATGC
AT-3’

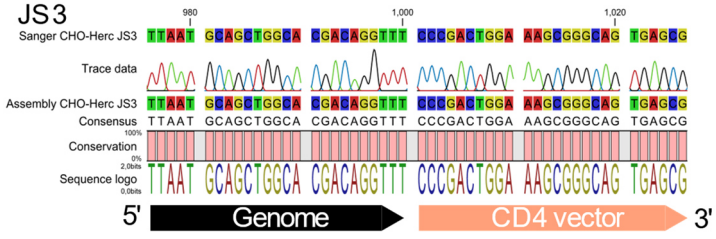
5’-

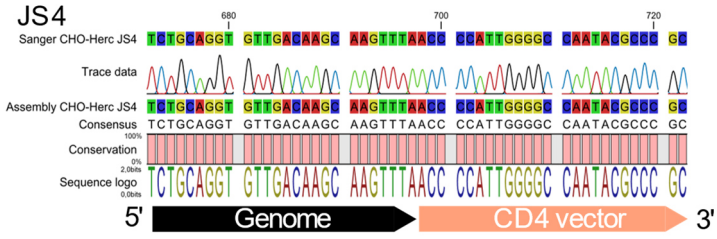
AGCTTGAGTCCTGAGGCGGAAAGAACCAGCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGC
ATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAACCATAGTCCCGCCC
CTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCGGC
CTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAAGGCCTCATCGGCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACTAAACCATGGCC
AAGCCTTTGTCTCAAGAAGAATCCACCCTCATTGAAAGAGCAACGGCTACAATCAACAGCATCCCCATCTCTGAAGACTACAGCGTCGCCAGCGCAGCTCTCTCT
AGCGACGGCCGCATCTTCACTGGTGTCAATGTATATCATTTTACTGGGGGACCTTGTGCAGAACTCGTGGTGCTGGGCACTGCTGCTGCTGCGGCAGCTGGCA
ACCTGACTTGTATCGTCGCGATCGGAAATGAGAACAGGGGCATCTTGAGCCCCTGCGGACGGTGCCGACAGGTGCTTCTCGATCTGCATCCTGGGATCAAAGC
CATAGTGAAGGACAGTGATGGACAGCCGACGGCAGTTGGGATTCGTGAATTGCTGCCCTCTGGTTATGTGTGGGAGGGCTAAGCACTTCGTGGCCGAGGAGCA
GGACTGACACGTGCTACGAGATTTCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGC
GGGGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAAATAAAGCATTTTT
TTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTATACCGTCGCTCTTCCGCTGCTTCCTCGCTCACTGACTCGCACGCGTC
GATGATACCATGGAGAAGTTCCTATTCCGAAGTTCCTATTCTTCAAAAGGTATAGGAACTTCACTAGTCCCGATCCGTCGACGTCAGGTGGCACTTTTCGGGGAA
ATGTGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAGGAAG
AGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGTAAAAGATGCTG
AAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTTTTCCAATGATGAGCACTT
TTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACTCAC
CAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGA
TCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGAC
GAGCGTGACACCACGATGCCTGTAGCAATGGCAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATG
GAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCA
TTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATA
GGTGCCTCACTGATTAAGCATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATC
CTTTTTGATAATCTCATGACCAAAATCCCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCT
GCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAG
CAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTT
ACCAGTGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTC
GTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAAGGCGGA
CAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCT
CTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTT
TTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCGCCGCAGCCGAACGACCGAGCGC
AGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCCAATACGCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGCTGGCACGACAGGTTTCCC
GACTGGAAAGCGGGCAGTGAGCGCAACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG
AATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATGATTACGCCAAGCTCTAGCTAGAGGTCGAGATCCCCCGGGCTGCAGGAATTCATTGGC
CCTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTACCCAACTTAATCGCCTTGCAGCACATCCCCCCTTCGCCAGGGGCTACCATGGAGAAG
TTCCTATTCCGAAGTTCCTATTCTCTAGAAAGTATAGGAACTTCAAGCTCTGGCACGACTGATCACGTATTACCGCCATGCATTAGTTATTAATAGTAATCAATTAC
GGGGTCGCTAGCAAATGCCCGGGCCTCGAGGGGATCTTCAATATTGGCCATTAGCCATATTATTCATTGGTTATATAGCATAAATCAATATTGGCTATTGGCCATT
GCATACGTTGTATCTATATCATAATATGTACATTTATATTGGCTCATGTCCAATATGACCGCCATGTTGGCATTGATTATTGACTAGTTATTAATAGTAATCAATTAC
GGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAA
TAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATC
ATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGCAGTACATCT
ACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTG
ACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGG
GAGGTCTATATAAGCAGAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNATGGGAGTCAAAGTTCTGTTTGCCCTGATCTGCATCGCTGTGGCCGAGGCCGACATCCAGATG
ACCCAGTCCCCCTCCAGCCTGTCTGCCTCTGTGGGCGACAGAGTGACCATCACCTGTCGGGCCTCCCAGGACGTGAACACCGCCGTGGCTTGGTATCAGCAGA
AGCCCGGCAAGGCCCCCAAGCTGCTGATCTACTCCGCCTCCTTCCTGTACTCCGGCGTGCCCTCCAGATTCTCCGGCTCTAGATCCGGCACCGACTTTACCCT
GACCATCAGCTCCCTGCAGCCCGAGGACTTCGCCACCTACTACTGCCAGCAGCACTACACCACCCCCCCCACCTTTGGCCAGGGCACCAAGGTGGAAATCAAG
CGGACCGTGGCCGCTCCCTCCGTGTTCATCTTCCCACCTTCCGACGAGCAGCTGAAGTCTGGCACCGCTTCCGTCGTGTGCCTGCTGAACAACTTCTACCCCC
GCGAGGCCAAGGTGCAGTGGAAGGTGGACAACGCCCTGCAGTCCGGCAACTCCCAGGAATCCGTGACCGAGCAGGACTCCAAGGACAGCACCTACTCCCTGT
CCTCCACCCTGACCCTGTCCAAGGCCGACTACGAGAAGCACAAGGTGTACGCCTGCGAAGTGACCCACCAGGGCCTGTCTAGCCCCGTGACCAAGTCTTTCAA
CCGGGGCGAGTGCTGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN

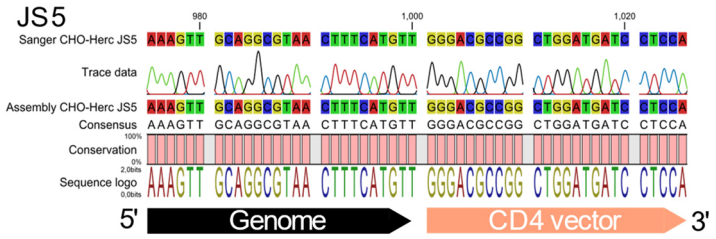
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAATAGTGTGTGGGAATTTTTTGTGTCTCTCACTC
GGAAGGACATATGGGAGGGCAAATCATTTAAAACATCAGAATGAGTATTTGGTTTAGAGTTTGGCAACATATGCCCATATGCTGGCTGCCATGAACAAAGGTTGG
CTATAAAGAGGTCATCAGTATATGAAACAGCCCCCTGCTGTCCATTCCTT**A**TTCCATAGAAAAGCCTTGACTTGAGGTTAGATTTTTTTTATATTTTGTTTTGTGTT
ATTTTTTTCTTTAACATCCCTAAAATTTTCCTTACATGTTTTACTAGCCAGATTTTTCCTCCTCTCCTGACTACTCCCAGTCATAGCTGTCCCTCTTCTCTTATGGAG
ATCCTCTAGAGGATCATCCCCGGGTACCGAGCTCGAATTCATCGATGATATCAGATCTTCAATATTGGCCATTAGCCATATTATTCATTGGTTATATAGCATAAATC
AATATTGGCTATTGGCCATTGCATACGTTGTATCTATATCATAATATGTACATTTATATTGGCTCATGTCCAATATGACCGCCATGTTGGCATTGATTATTGACTAGT
TATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGAC
CCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGAGTATTTACGGTAAACTGCCCACTTG
GCAGTACATCAAGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT
CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTT
CCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGC
GGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNGAAGTGCAGCTGGTGGAATCTGGCGGCGGACTGGTGCAGCCTGGCGGATCTCTGAGACTGTCTTGTGCCGCCTCCGGCTTCAACATCAAGGACA
CCTACATCCACTGGGTGCGACAGGCCCCTGGCAAGGGACTGGAATGGGTGGCCAGAATCTACCCCACCAACGGCTACACCAGATACGCCGACTCTGTGAAGG
GCCGGTTCACCATCTCCGCCGACACCTCCAAGAACACCGCCTACCTGCAGATGAACTCCCTGCGGGCCGAGGACACCGCCGTGTACTACTGCTCTAGATGGGG
AGGCGACGGCTTCTACGCCATGGACTATTGGGGCCAGGGCACCCTCGTGACCGTGTCCTCTGCTTCTACCAAGGGCCCCTCCGTGTTCCCTCTGGCCCCTTCC
AGCAAGTCTACCTCTGGCGGCACAGCCGCTCTGGGCTGCCTCGTGAAGGACTACTTCCCCGAGCCCGTGACAGTGTCTTGGAACTCTGGCGCCCTGACCTCCG
GCGTGCACACCTTTCCAGCTGTGCTGCAGTCCTCCGGCCTGTACTCCCTGTCCTCCGTCGTGACTGTGCCCTCCAGCTCTCTGGGCACCCAGACCTACATCTGC
AACGTGAACCACAAGCCCTCCAACACCAAGGTGGACAAGAAGGTGGAACCCAAGTCCTGCGACAAGACCCACACCTGTCCCCCTTGTCCTGCCCCTGAACTGC
TGGGCGGACCCAGCGTGTTCCTGTTCCCCCCAAAGCCTAAGGACACCCTGATGATCTCCCGGACCCCCGAAGTGACCTGCGTGGTGGTGGATGTGTCCCACGA
GGACCCTGAAGTGAAGTTCAATTGGTACGTGGACGGCGTGGAAGTGCACAACGCCAAGACCAAGCCTAGAGAGGAACAGTACAACTCCACCTACCGGGTGGTG
TCCGTGCTGACCGTGCTGCATCAGGACTGGCTGAACGGCAAAGAGTACAAGTGCAAGGTGTCCAACAAGGCCCTGCCTGCCCCCATCGAAAAGACCATCTCCA
AGGCCAAGGGCCAGCCCCGGGAACCCCAGGTGTACACACTGCCCCCTAGCCGGGAAGAGATGACCAAGAACCAGGTGTCCCTGACCTGTCTCGTGAAGGGCT
TCTACCCCTCCGATATTGCTGTGGAATGGGAGTCCAACGGCCAGCCTGAGAACAACTACAAGACCACCCCCCCTGTGCTGGACTCCGACGGCTCATTCTTCCTG
TACAGCAAGCTGACAGTGGACAAGTCCCGGTGGCAGCAGGGCAACGTGTTCTCCTGCTCCGTGATGCACGAGGCCCTGCACAACCACTACACCCAGAAGTCCC
TGTCCCTGAGCCCCGGCAAGTGANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAATAGTGTGTGGGAATTTTTTGTGTCT
CTCACTCGGAAGGACATATGGGAGGGCAAATCATTTAAAACATCAGAATGAGTATTTGGTTTAGAGTTTGGCAACATATGCCCATATGCTGGCTGCCATGAACAA
AGGTTGGCTATAAAGAGGTCATCAGTATATGAAACAGCCCCCTGCTGTCCATTCCTTATTCCATAGAAAAGCCTTGACTTGAGGTTAGATTTTTTTTATATTTTGTT
TTGTGTTATTTTTTTCTTTAACATCCCTAAAATTTTCCTTACATGTTTTACTAGCCAGATTTTTCCTCCTCTCCTGACTACTCCCAGTCATAGCTGTCCCTCTTCTCTT
ATGGAGATCTAAATCCGATAAGGATCGATCCGGA-3’

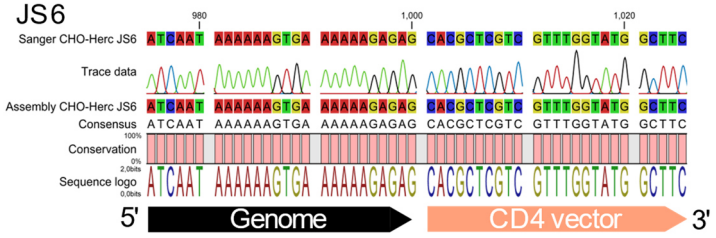

JS1
Sanger CHO-Herc JS1
Trace data
Assembly CHO-Herc JS1
Consensus
Conservation
Sequence logo
5'
Genome
Herc vector
3'


JS2
Sanger CHO-Herc JS2
Trace data
Assembly CHO-Herc JS2
Consensus
Conservation
Sequence logo
5'
Genome
Herc vector
3'


JS3
Sanger CHO-Herc JS3
Trace data
Assembly CHO-Herc JS3
Consensus
Conservation
Sequence logo
5'
Genome
CD4 vector
3'


JS4
Sanger CHO-Herc JS4
Trace data
Assembly CHO-Herc JS4
Consensus
Conservation
Sequence logo
5'
Genome
CD4 vector
3'


JS5
Sanger CHO-Herc JS5
Trace data
Assembly CHO-Herc JS5
Consensus
Conservation
Sequence logo
5'
Genome
CD4 vector
3'


JS6
Sanger CHO-Herc JS6
Trace data
Assembly CHO-Herc JS6
Consensus
Conservation
Sequence logo
5'
Genome
CD4 vector
3'