Extended Report

Transgene analysis and integration site sequencing of a transgenic CHO cell line with the vector X

Prepared for:

Customer name:

Internal project number:
Quote number:

Version: 1
Date:

Goal

In this study, 1 transgenic CHO cell line with the vector xxxxx sequence was analyzed.

The aim of this analysis was to:

1. Study the vector integrity:
    1) Determine the presence of sequence variants and their allele frequency.
    2) Determine the presence of vector-vector breakpoints that represent concatemerization of multiple copies of the vector and/or structural rearrangements in a single vector sequence.
2. Identify vector integration site(s) and breakpoint sequences between the vector and genome.
3. Assess the presence of structural variants surrounding the vector integration site(s).
4. Estimate the copy number of the vector.

An overview of the TLA technology and technical details of the performed analyses is provided in the manual "Introduction to the terminology and methods used in TLA analyses_v2".

Summary

Sample	Vector integrity	Integration site(s)	Structural variants at the integration site	Copy number estimation
MCB	6 sequence variants, 2 structural variants	Chr3: 169,680,259 - 169,680,260	no	3-5

Conclusion

In MCB 1, 3-5 copies of the vector have integrated in chr 3. 6 sequence variants and 2 structural variants are found within the integrated vector sequence.

Abbreviations

Abbreviation	Full name
CHO	Chinese Hamster Ovary
TLA	Targeted Locus Amplification
PCR	Polymerase Chain Reaction
NGS	Next-Generation Sequencing
bp	Base pair
DNA	Deoxyribonucleic acid
Html	HyperText Markup Language
P	Primer
BWA-MEM	Burrows-Wheeler Aligner-Maximal Exact Match

Methods

TLA, sequencing and data mapping

Viable frozen CHO-K1 cells were used and processed according to Cergentis' TLA protocol (de Vree et al. Nat Biotechnol. Oct 2014). An overview of the TLA technology and technical details of the performed analyses is provided in the manual "Introduction to the terminology and methods used in TLA analyses_v2".

TLA was performed with 2 independent primer sets specific for the vector sequence (Table 1).

Table 1: Primers used in TLA analysis

Primer set	Name/Viewpoint	Direction	Binding position	Sequence
1	Amp	Rv	132	X
		Fw	256	X
2	GOI	Rv	2,745	X
		Fw	3,186	X

PCR products were purified, library prepped using the Illumina Nextera flex protocol and sequenced on an Illumina sequencer.

In short, the Nextera DNA Flex library prep kit uses a bead-based transposome complex to tagment genomic DNA by fragmenting and adding adapter tag sequences. Following the tagmentation step, a limited-cycle PCR step adds Nextera DNA Flex-specific index adapter sequences to the ends of a DNA fragment. The Sample Purification Bead cleanup step then purifies libraries for use on an Illumina sequencer (source: Nextera™ DNA Flex Library Prep reference guide, Document # 1000000025416 v01). The resulting library contains samples with unique barcodes (10-base Illumina indexes) for each sample and each primer set. Library is sent for sequencing (paired-end 149 bases) on the NextSeq. The NGS reads were aligned to the vector sequence and host genome.

Alignment of sequencing reads

The sequencer produces a runfolder in each sequencing run containing the base call information, settings and information about the sequencing run and images of the flowcell taken during the 2x 149 cycles of base calling and 2x 10 cycles barcode reading. This runfolder along with the barcodes of the TLA samples are used as input for bcl2fastq tool from Illumina, to convert base call information to read information for each TLA sample in paired-end FASTQ files, this process is called demultiplexing. Bcl2fastq generates an html report which describes/summarizes metrics about the basecalling and the demultiplexing that has been performed for both the complete Illumina run and for each TLA sample. That gives an impression about how the run has performed, the amount of sequenced reads that are assigned to each TLA sample/barcode and the quality of the bases that have been sequenced. Due to overall good quality of the data generated all aligning reads are included. After the conversion to FASTQ files, reads were mapped using BWA-MEM (Li et al. Bioinformatics, 2010 [PMID: 20080505]), version 0.7.15-r1140, settings mem -B 7.

The Chinese Hamster CriGri-PICRH 1.0 genome assembly GCF_003668045.3 was used as host reference genome sequence.

Table 2: Quality matrix for sequencing run

Sample	Number of reads	Read length (bp)	% reads mapped to vector*	% reads mapped to genome*	% >= Q30 Bases**	Mean Quality Score***
MCB p1	1,519,218	149	XXX	XXX	XXX	XXX
MCB p2	1,811,122	149	XXX	XXX	XXX	XXX

*split reads can be assigned to both the vector and genome, therefore a sum of the percentage reads mapped to vector and percentage reads mapped to genome > 100% is possible.
**% >= Q30 bases: the percentage of sequenced bases that have a quality score 30 or higher.
***Mean Quality Score: the average quality score of the sequenced bases.

Sequence variants detection

The presence of sequence variants is determined using samtools mpileup (samtools version 1.3.1) (Li et al. Bioinformatics, Jun 2009 [PMID: 19505943], Li et al. Bioinformatics, Nov 2011 [PMID: 21903627]). Only read-bases with a minimal Q-score of 20 (Base call accuracy of >99%) are used for the detection.

Sequence variants are reported that meet the following criteria:

- allele frequency (relative amount of reads with the variant, compared to total coverage on the variant position) of at least 5%;
- the variant is present in the data of both primer-sets;
- for at least one of the primer-sets the coverage is >=30X;
- the variant is identified in both forward and reverse aligning sequencing reads;
- low frequency variants (between 5-20% mutant allele frequency) are not found with similar frequencies in an unrelated control.

Structural variants detection

Breakpoint sequences consisting of two parts of the vector, are identified using a proprietary Cergentis script. Breakpoints resulting from the TLA procedure itself are recognized by the restriction enzyme-specific sequence at the junction site and removed.

Vector-vector breakpoint sequences are reported that meet the following criteria:

- the breakpoint sequence is present in >1% of the reads at the position of the fusion;
- the breakpoint sequence is observed in data of all primer sets, unless the data provides a clear explanation why the fusion is not found in one of the data sets;
- the breakpoint sequence is not present in unrelated control sample(s);
- visual inspection of the breakpoint sequence in an NGS data browser is performed to remove fusions that are sequencing artefacts, e.g. breakpoints found at hairpin structures or low-complexity regions.

Integration site detection

Integration sites are detected based on a) coverage peak(s) in the genome and b) the identification of breakpoint sequences between the vector sequence and host genome.

Results MCB

Vector integrity

Figure 1 depicts the NGS coverage across the vector sequence using primer set 1. Same results were obtained with primer set 2.

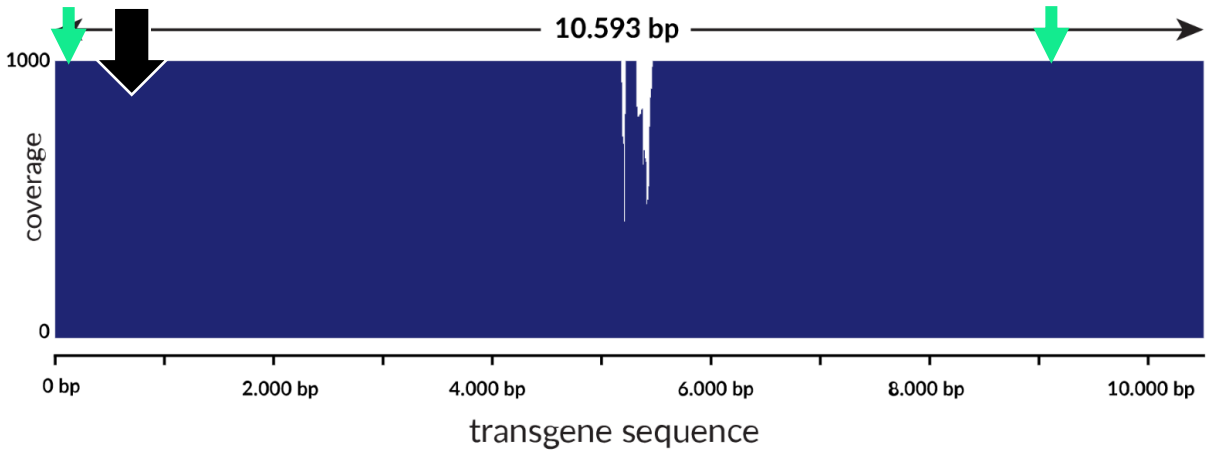

Figure 1: NGS sequencing coverage across the vector with primer set 1. The black arrow indicates the primer location. The green arrows indicate the locations of the identified vector-genome breakpoint sequences (described below). Y-axis is limited to 1000x. In an actual report the data of all primer sets will be presented.

High coverage is observed across the complete vector sequence Vector: 1-10,250. Local dips in coverage are due to GC rich regions that are less efficiently sequenced.

Sequence variants and structural variants were called in the covered regions.

Sequence variants

Detected sequence variants are presented in table 3. All sequence variants at or near 100% mutation frequency were detected in this sample as well as in the deviant control and most likely represent deviations present in the provided reference sequence of the vector before its introduction into the sample.

Table 3: Identified sequence variants

				Primer set 1		Primer set 2	
Region	Position	Ref	Mut	Cov	%	Cov	%
Amp	141	A	C	21,254	20	788	25
GOI	1,013	T	-4AGTT	1,881	100	7,501	100
GOI	2,956	T	C	1,278	27	34,122	21
GOI	5,698	A	+1G	751	18	2,221	15
Backbone	9,487	T	G	1,358	100	1,523	99
Backbone	10,037	G	A	2,145	20	854	20

'+' indicates an insertion; '-' indicates a deletion.

Column 1 (Region): the region where the variant is found within the reference sequence. Column 2 (Position): position within the reference sequence. Column 3 (Reference): nucleotide present in the reference sequence at this position. Column 4 (Mutation): observed mutation. Column 5-8: quantitative measurements are presented for each primer set in each sample. Column 5 and 7 (Cov, short for Coverage): total number of reads that map to this position in the data generated with either primer set 1 (column 5) or 2 (column 7). Column 6 and 8 (%): percentage of reads containing the mutation (=mut/cov*100%) in the data of primer set 1 (column 6) or 2 (column 8).

Vector concatemerization and structural variants

The identified vector-vector breakpoint sites are shown in table 4. In total, 2 structural variants were identified. Intact reads were also found at all positions indicating that (partial) vector sequences have concatemerized. These structural variants were only identified in this sample and not in the independent control.

Table 4: Vector-vector breakpoints

								# of reads with fusion		% of reads with fusion			
Break-point	Vector		Vector		Orientation of the breakpoint	Hom	Insert	p1	p2	p1-pos1	p1-pos2	p2-pos1	p2-pos2
1	→	6,945	7,657	←	tail to tail	1	-	100	1,452	17	16	11	22
2	←	14	5,051	→	head to head	-	2	1,025	859	12	22	21	17

Column 1 (Breakpoint): breakpoint number. Column 2 (Vector): orientation and position of the left side of the breakpoint. Column 3 (Vector): position of the right side of the breakpoints and orientation. Column 4 (Orientation of the breakpoint): orientation of the breakpoint. Column 5 (Hom, short for homology): number of bases of homology found between the sequence at the left and right side. The homologous bases are not included when determining the positions as represented in columns 2 and 3. Column 6 (Insert): number of novel bases that are inserted at the breakpoint site. Column 7-10 (#of reads with fusion/% of reads with fusion): 2 quantitative measurements are presented for each primer set: a) the absolute number of reads in which each breakpoint is found b) relative number of reads (%) that contain the breakpoint, for example p1-pos1 = 6 and p1-pos2 = 8, means that of all reads that aligned to pos1, 6% contained the fusion, and of all reads that aligned to pos2 8% contained the fusion. Please note, the number of reads counted for each breakpoint is a slight underestimate of the actual number of reads that contained the breakpoint, this is because breakpoints are only counted if both sides of the breakpoint can be mapped. If the sequence on one of the sides is too short to be mapped, it is not counted. Relative frequency with a % higher than 100 is sometimes encountered. This occurs on non-unique sequences (repetitive sequences in genome or vector).

The left side of the fusion is in red, the right side of the fusion is in blue, any homologous bases are in purple and any inserted bases are black.

1) vector:6,945 (tail) fused to vector:7,567 (tail) with 1 homologous base

ATCGGTTTAACAACGGTTAAGCGTTAGTTCCTTGAATCGAAACTTTGGTAACATGTAGCTAGGCTAATG
CATATGCAATGGATTCGAGACTAATGACCCTTAGGCCTAATTAGGGCTAGAGTCTCGAGAGCATTGGG
ATATCGCGCGGCCTTAGGGACTCTCGGAGACTGGAGCTCAGAGATTTCGGCGATACGCGATATCGGT

2) vector:14 (head) fused to vector:5,051 (head) with 2 inserted bases

GGGTCTAGGGACTGATCGGGATGCCCTGGACTAGGATAGCTAGCTTTTACAAACCCACAATGGATTA
GAAATCCGAATAATGGGGATTACCCCCTAGATCGAAATTTCGAAAGTGGGAGATCGCGTCAGAAGCT
AACGAAGGGATCGCATAGAGAGGACTCGGCTAGAGAGATCGCAGATCGAGATCGAACGTACATCGAT
CAGTCGACTGA

Integration sites

Whole genome coverage plot

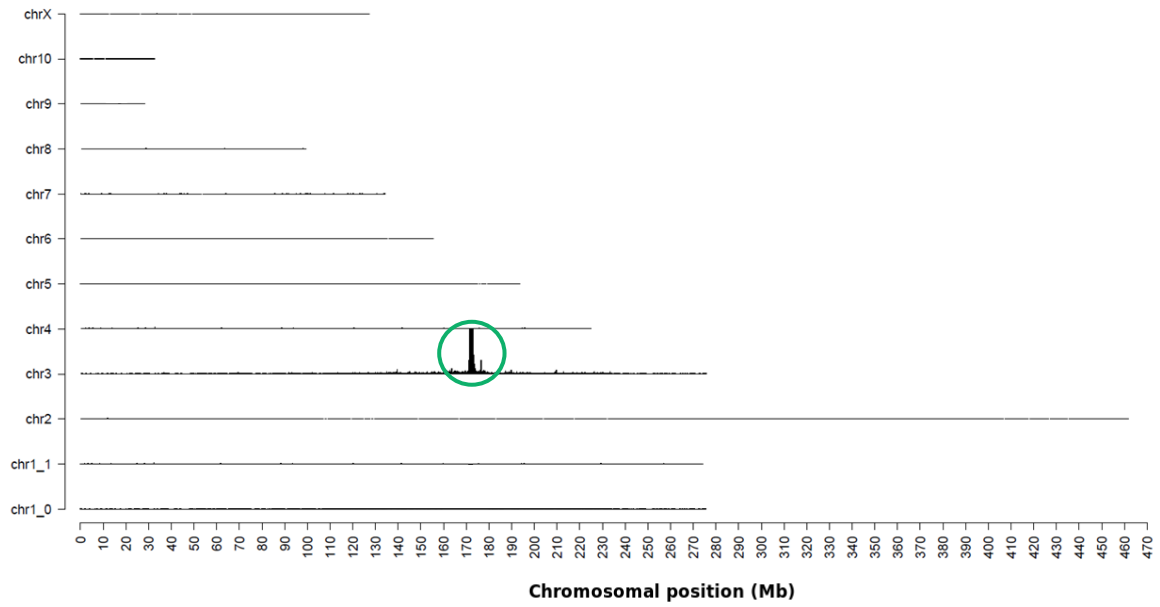

Figure 2: TLA sequence coverage across the Chinese Hamster genome using primer set 1. The chromosomes are indicated on the y-axis, the chromosomal position on the x-axis. Identified integration site is encircled in green. In an actual report the data of all primer sets will be presented.

As shown in figure 2, the vector has integrated on chromosome 3.

Locus-wide coverage

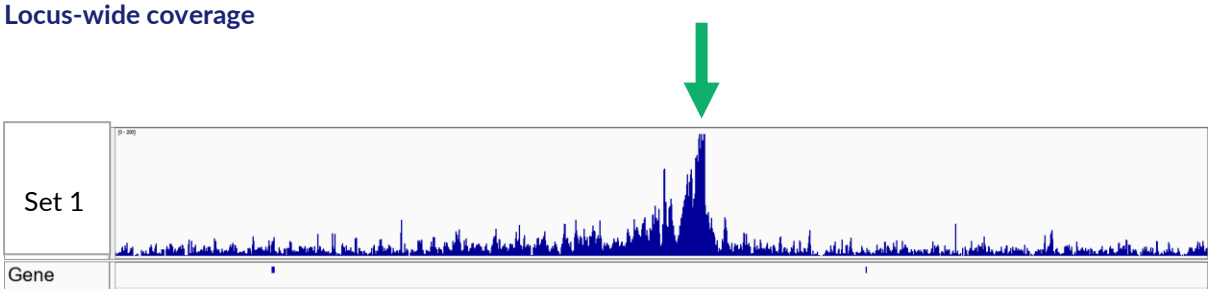

Figure 3: TLA sequence coverage (in blue) across the vector integration locus, chr3:169,530,000-169,830,000. The black arrow indicates the location of the breakpoint sequences. Y-axis is limited to 200x. In an actual report the data of all primer sets will be presented.

Coverage is observed across the vector integration site as shown in figure 3.

Breakpoint sequences

The following breakpoint sequences were identified marking the vector integration:

5' integration site:
chr3:169,680,259 (tail) fused to Vector: 4 (head) with 5 inserted bases
ATTGCACGTACGTACGTTTGGCAAACACTGTGCCTCGACTGCCGTCGGCGTAACGTCAGCTAGTTTAC
CCTGTTGTACACACTGTGATAGGATGGTCGAATCGATGCTAAGCTTCGTAAATCGATATCGATCGTAG
CTATGCTAGGGTCGCC

3' integration site:
Vector: 9,527 (tail) fused to chr3:169,680,260 (head) with 3 bases homology
CACTATGGGTACGTACGTTATATCCCTGATCGTGCTCGTAGCTGCCTGCTAAGCTAGCTGATGCTGCC
GCTTGTTGTACACTTAGGACTGTGATAGCTACGTCGTAAGCTGCTCGATGCTAGATCGCTAGCGGCGG
CTAGCTAGTGGCTGAGT

The coverage profile in figure 4 shows that no genomic rearrangements have occurred in the region of the integration site.

From this data it is concluded that the vector has integrated chr3: 169,680,259 - 169,680,260. According to Refseq, there are no genes annotated here.

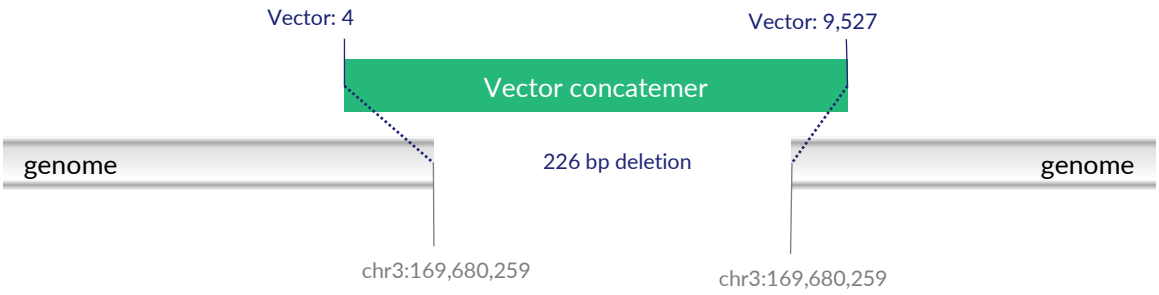

Figure 4: Schematic representation of the integration site.

Copy number estimation

In the MCB sample, the coverage on the vector-side is 4-5 times higher than on the genome-side of the integration site. 1 integration site and 2 vector-vector breakpoints are found. The copy number is estimated to be 3-5 copies.

Table 4: Observed coverage in generated data of genome and vector mapped reads at positions that were identified as breakpoint between vector and genome

		Primer set 1		Primer set 2	
Region	Position	Cov	Ratio	Cov	Ratio
Genome	chr3:169,680,259	3,303	3.5	177	4.2
Vector	4	11,558		745	
Genome	chr3: 169,680,260	356	3.9	453	3.5
Vector	9,527	1,400		1,586	

QC information

Sample and Study details

Sample receipt date
Condition of sample at receipt
Start date in the lab
Sequencing run
Date data analysis
Deviations from the protocol
TLApp version

Study Personnel

Lab technician
Data Analyst
QC Analysis and Report

Quality control

The results are independently verified and reviewed and are an accurate and complete representation of the study. TLA processing of cells, NGS sequencing, and data analysis (except for copy number estimation) are ISO/IEC 17025:2017 accredited by the Dutch Accreditation Council RvA, Registration number L671.

Scientific-approval
Date
Signature

*Report will be signed after reviewed by customer