Final Report

Study Title:

Genetic characterization of the XXXXX cell line

Prepared for:	Company name Company address
Customer representative:	Name Position within a company email
Study identification:	XXXX/XXXX-XXXX
Version:	1

Index

1 Summary

Using the Cergentis' Targeted Locus Amplification (TLA) methodology followed by Next-Generation Sequencing (NGS) and data mapping in combination with droplet digital PCR (ddPCR) copy number determination, sample XXX from the XXXX cell line containing the XXXXX sequence was analyzed. Genetic characterization was based on transgene and vector integrity, identified integration site and integrated number of copies.

The studied sample showed 1 integration site on chromosome 6. Data mapping revealed 1 sequence variants and 4 structural variants in the introduced vector sequences. The copy number showed 8 partial vector copies.

2 Study information

2.1 Sponsor

Name	XXXX
Address	XXXXX
Telephone	XXXX
E-mail	XXX@XXXX
Sponsor representative:	XXXXX - Project Management

2.2 Test facility

Name	Cergentis BV
Address	Yalelaan 62, 3584 CM Utrecht, The Netherlands
Telephone	+31 30 760 16 36
E-mail	cheryl.dambrot@cergentis.com
Representative	Cheryl Dambrot – Study Director

2.3 Sample information

Sample name (label on shipped tube)	XXXXX
Description	X tubes labelled with sample code containing 10 million xxx cell line cells were received. The Sponsor was responsible for characterization and identification of the samples and control.
Storage condition	The sample was received on dry ice and stored in an ultra low temperature freezer at -80°C until sample preparation.
Provided vector sequence	XXXXX sequence provided by the Sponsor, see Appendix 1).

2.4 Study personnel

Name Lab technician	xxxx
Name Data Analysis	xxx
Name Data Quality Control	xxxx
Name Study Director	Cheryl Dambrot
Name Director QA and RA	Maaike van der Weij

2.5 Study dates

Date sample receipt	29-Nov-2022
Experimental start date	06-Dec-2022
Experimental completion date	16-Dec-2022
Mapping Completion date	17-Dec-2022
Analysis Completion date	04-Jan-2023
Quality Control date	07-Jan-2023
Reporting completion date	See signature date of study director

2.6 Administrative details

Internal Project nr	XXXX / 202X - XXXX
Raw data reference	run 23-XXX
Software version	TLApp version 1.X.X
Deviations from standard procedures	none

3 Abbreviations

Abbreviation	Full name
Bp	Base pair
BWA-MEM	Burrows-Wheeler Aligner-Maximal Exact Match
DNA	DeoxyriboNucleic Acid
FW	Forward
Html	HyperText Markup Language
NGS	Next-generation Sequencing
Set X	Primer set X
PCR	Polymerase Chain Reaction
ddPCR	Droplet Digital PCR
RV	Reverse
TLA	Targeted Locus Amplification

4 Purpose of the study

The study was performed to genetically characterize the Sponsor's cell sample in accordance with the recommendations for characterization of expression constructs in eukaryotic cells, ICH topic Q5B (Analysis of the Expression Construct in Cell Lines Used for Production of r-DNA Derived Protein Products). Using the TLA technology, transgene and integration site analysis was performed. The locations of the vector integration in the genetically modified XXX cell line cells as well as the integrity of the integrated vector sequence was determined. Furthermore, the copy number of the vector was determined using ddPCR.

5 Methods

5.1 TLA technology

The TLA technology is described by De Vree et al., Nature Biotechnology 32(10), 1019-1025 (2014). Technical details are also provided in the Cergentis Manual: TLA terminology & methods (Appendix 9). Briefly, genomic DNA is crosslinked, fragmented and circular DNA fragments are generated. The locus of interest is amplified and sequenced with NGS technology, and the sequence data are subsequently analyzed.

The generated sequence data is used to a) determine the presence of sequence variants and their allele frequency in the integrated vector sequence, b) to determine the presence of vector-vector breakpoints or backbone sequence that represent concatemerization of multiple copies of the vector and/or structural rearrangements in a single vector sequence, c) to identify vector integration site or sites and breakpoint sequences between the vector and genome and d) to assess the presence of structural variants surrounding the vector integration site(s) in the host genome.

5.2 Study outline

5.2.1 TLA and sequencing

The vial was thawed to room temperature, cells were counted and viability was measured. A total of XX million viable xxxx cell line cells were collected from the cryovial, cross-linked and fragmented by enzymatic digestion. The DNA was circularized by ligation and amplified by PCR with primer pairs specific for the genetic locus of interest. The primer sequences (see Table 1a and b) were based on the complete vector sequence containing the transgene provided by the Sponsor (see Section 2.3).

The PCR products were purified, and library prepped using the protocol in the Nextera™ DNA Flex Library Prep reference guide, Document # 1000000025416 v01. The resulting libraries contained products with unique barcodes (dual 10-base Illumina indexes) for each sample and each primer set. The libraries were sequenced (paired-end 2x149 bases) on the NextSeq (Illumnia®) system.

Table 1a: Primers used in TLA analysis for NlaIII

Primer set	Name/VP	Direction	Binding position *Vector name*	Sequence
1				
2				

Table 1b: Primers used in TLA analysis for DpnII

Primer set	Name/VP	Direction	Binding position *Vector name*	Sequence
3				
4				

5.2.2 Alignment of sequencing reads

The NextSeq system 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 2x149 cycles of base calling and 2x10 cycles barcode reading. The information is demultiplexed into readable information that is suitable for aligning.

Overall good quality of the data was generated (Appendix 2), so all aligning reads were included. After conversion to FASTQ files, reads were mapped using data mapping software BWA-MEM (Li et al. Bioinformatics, 2010 [PMID: 20080505]). The NGS reads were aligned to the vector sequence and host genome. For reference of the used vector sequence see Section 2.3. Since the samples originate from a XXX cell line, the XXXX genome was used as host reference genome sequence.

5.2.3 Sequence variants detection in vector sequence

The presence of sequence variants was evaluated by comparison with the data in the vector reference file as provided by the Sponsor, using the tool "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 pre-set criteria, as described in the Cergentis Manual: TLA terminology & methods (Appendix 9):

- 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 all 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,

5.2.4 Structural variants detection in vector sequence

A breakpoint sequence is a sequence containing the breakpoint of, and fusion between, two sequences originating from different origin compared to the reference sequence (see Cergentis Manual, TLA Terminology & methods in Appendix 9 for further details). Vector-Vector breakpoint sequences consisting of two parts of the vector, are identified using a proprietary Cergentis script as available in the TLApp (for version see Section 2.6) dedicated to detection of breakpoints in TLA sequencing data. Breakpoints resulting from the TLA procedure itself are recognized by the restriction enzyme-specific sequence at the junction site and artefacts, e.g. breakpoints found at hairpin structures or low complexity regions, are removed.

Vector-vector breakpoint sequences are reported that meet the following pre-set criteria, as described in the Cergentis Manual, TLA Terminology & methods (Appendix 9):

- the breakpoint sequence is present in >1% of the reads at the position of the fusion,
- the breakpoint sequence is observed in data of both primer-sets,

5.2.5 Integration site detection

Integration sites are detected based on a) coverage peaks in the genome and b) the identification of breakpoint sequences between the vector sequence and host genome as presented in the paper of De Vree et al., Nature Biotechnology 32(10), 1019-1025 (2014) and in the Cergentis Manual, TLA Terminology & methods in Appendix 9.

5.2.6 Copy number determination

ddPCR assays have been designed to determine the copy number of the vector elements *X* and X present in the vector (Appendix 1 and Table 2). Additionally, two assays on the endogenous *gene 1* and *gene 2* have been purchased (dHsaCP2500313 and dHsaCP2500315Bio-Rad, Hercules, CA) to serve as genomic reference. The two independent reference genes were used because modified cell lines may not always be stably diploid for their complete genome. The copy number was determined as the ratio of the total copies for the vector element *X* or *X* to the total copies for the genomic reference gene, multiplied by 2 for the copies per cell for the genomic reference gene. For an accurate copy number assessment of this vector, restriction enzymes (NlaIII and HaeIII) have been used in the ddPCR reactions to fragment the vector concatemers in these cell line samples. Testing was performed on gDNA of the transgenic cell line samples with 5-Units of these restriction enzymes in the reaction, according to the manufacturer's instructions (QX200 Droplet Digital PCR System; Bio-Rad, Hercules, CA).

Table 2: ddPCR assays

Target	Oligo type	Sequence
Element X	FW primer	
	RV primer	
	Probe	
Element X	FW primer	
	RV primer	
	Probe	

6 Results

6.1 TLA and sequencing

For each of the studied samples two independent datasets were generated by TLA followed by NGS sequencing from which the Quality Scores indicate that the minimum requirements are met. Q30 scores over 80 and Mean quality scores over 30 are considered high quality data scores. See Appendix 2 for the quality scores and sequencing run details.

6.2 Alignment of sequencing reads

The generated data sets, comprising of NGS reads, were mapped against the provided vector sequence in the Vector integrity section, and to the host reference genome, in the integration site section. Appendix 2 shows the percentages of reads mapped to the vector and to the genome.

6.3 Vector integrity

6.3.1 Results of NGS read coverage across the vector sequence

Appendix 3 depicts the obtained NGS coverage per base for the studied sample across the integrated vector sequence. The figures show the number of reads mapping to each individual position, across the provided vector sequence for each of the four primer sets.

Coverage is observed across the complete vector sequence Vector: 1-xxx, demonstrating that the entire sequence is integrated in the genome of the sample.

6.3.2 Sequence variants in vector sequence

Comparison of the mapped reads (Appendix 3) with the provided vector sequence (Appendix 1) using the criteria in section 5.2.3 revealed X sequence variants in the integrated vector sequence (Appendix 4).

The integrated sequence was identical to the provided vector reference, with the exception of one mutation. This C to A sequence variant in the Vector's annotated neomycin resistance cassette was detected in this sample.

6.3.3 Structural variants in vector sequence

Comparison of the mapped reads with the provided vector sequence using the criteria in section 5.2.3 revealed4 structural variants present in the integrated vector sequence.

The vector-vector breakpoint sites that were evaluated according to the criteria in section 5.2.4. A total of 4 structural variants were identified (see Appendix 5). For all structural variants, intact reads were also found at all positions of the vector-vector breakpoints indicating that (partial) vector sequences have concatemerized.

6.4 Integration sites

The integration sites were evaluated using the criteria in section 5.2.5. Whole genome coverage plots were generated using data obtained with primer sets 1, 2, 3 and 4. The plots showed 1 integration site, namely on chromosome 6 (see Appendix 6).

The vector integration site in chromosome 6 is at chr6:69,721,102-69,721,205 and shows that a genomic deletion has occurred in the genomic region of the integration site. The 200 bp genomic sequence in between the two identified breakpoints is deleted. The identified breakpoints are located in intron 7 of of LMBRD1 (Appendix 7).

6.5 Copy number determination

Using the criteria described in section 5.3.6 the copy number is determined. Approximately 300 total copies have been found for the reference genes *1* and *2* (Appendix 7; Figure 1). This number of copies corresponds well to the input amount of approximately 1 ng gDNA based on the average weight of 6.5 pg for the xxxx genome (Piovesan et al., *BMC research notes* 2019). These results indicate that the cells are diploid and therefore, these genomic genes were used as references to calculate the copy number of the vector elements *X* and *Y*. The total copies for the vector elements *X* and *Y* were higher than those for the reference genes (*1* and *2*), indicating that the cell line samples contain more than 2 copies per cell for these vector elements. A copy number of 8 for *X* and 7 for *X* has been found in each cell line sample. The discrepancy between X and *Y* copy numbers can be explained by the presence of vector-vector breakpoints that have been identified by TLA analysis previously. Hence, the cell line sample contains 8 partial copies of vector X (Appendix 7).

7 Conclusion

The sample X from the cell line X showed 1 integration site. 1 sequence variants and X structural variants in the integrated vector sequence were found. The copy number was determined to be 8 partial copies.

8 Approval

The scope of accreditation for ISO/IEC 17025:2017, accredited by the Dutch Accreditation Council RvA, Registration number L671, entails all analytical services including, determination of the integrity of the transgene vector sequence; determination of the vector integration site(s) and breakpoint sequences between the vector and genome, determination of the presence of structural variants surrounding the vector integration site(s), next generation sequencing (NGS) and bio-informatic data analysis, next generation sequencing (NGS) and bio-informatic data analysis. The ddPCR copy number determination, found in sections, 5.2; 5.3.6 and 6.5 and appendices 7 and 8 is not included in the scope of this accreditation.

Scientific approval	Cheryl Dambrot – Study Director
Date	
Signature	

QA approval	Maaike van der Weij – Director QA and RA
Date	
Signature	

Addendum

Appendix 1 Vector sequence and annotation
Appendix 2. Quality Matrix for sequencing run
Appendix 3. NGS read coverage across the vector sequence
Appendix 4. Sequence variants in vector sequence
Appendix 5. Structural variants in vector sequence
Appendix 6. Whole genome wide coverage plots
Appendix 7. Integration site information
Appendix 8. Copy number determination details
Appendix 9. Cergentis Manual: TLA terminology & methods

Appendix 1

Vector sequence and annotation

Vector Name:

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Figure 1: Annotation of vector (name) sequence

Appendix 2

Quality matrix for sequencing run

Table 1 shows the number of reads obtained for each of the datasets in the study. For the sample 4 datasets are generated, one for each primer set (set 1-set 4). The reads are mapped to the vector and host genome sequence, the percentage of reads mapped to each is shown per dataset. The quality scores show that the generated data is of high quality. A high quality score implies that a base call is more reliable and less likely to be incorrect. For base calls with a quality score of Q30, the error probability is 0.001, meaning that one base call in 1,000 called bases is predicted to be incorrect. Q30 scores over 80 and Mean quality scores over 30 are considered high quality data scores. All data was of high quality.

Table 1: 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***
sample set 1	1,577,910	149	66	75	85.16	32.82
sample set 2	1,325,929	149	62	80	82.24	33.15
Sample set 3	1,277,910	149	58	76	83.12	30.24
Sample set 4	1,115,829	149	63	77	84.56	31.45

*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.

Appendix 3

NGS read coverage across the vector sequence

In Figures 1 the read coverage is expressed in number of reads mapped for Primer set 1 (Set 1), Primer set 2 (Set 2), Primer set 3 (Set 3) and Primer set 4 (Set 4) for sample X. On the X-axes is the vector map displayed and the Y-axes indicate the number of NGS read coverage, upper limit is specified in figure legend. Black bold arrows indicate the primer locations.

High coverage is observed across the complete vector sequence Vector: 1-xxxx, indicated by the grey areas in figures 1, demonstrating that the entire sequence is integrated in the genome.

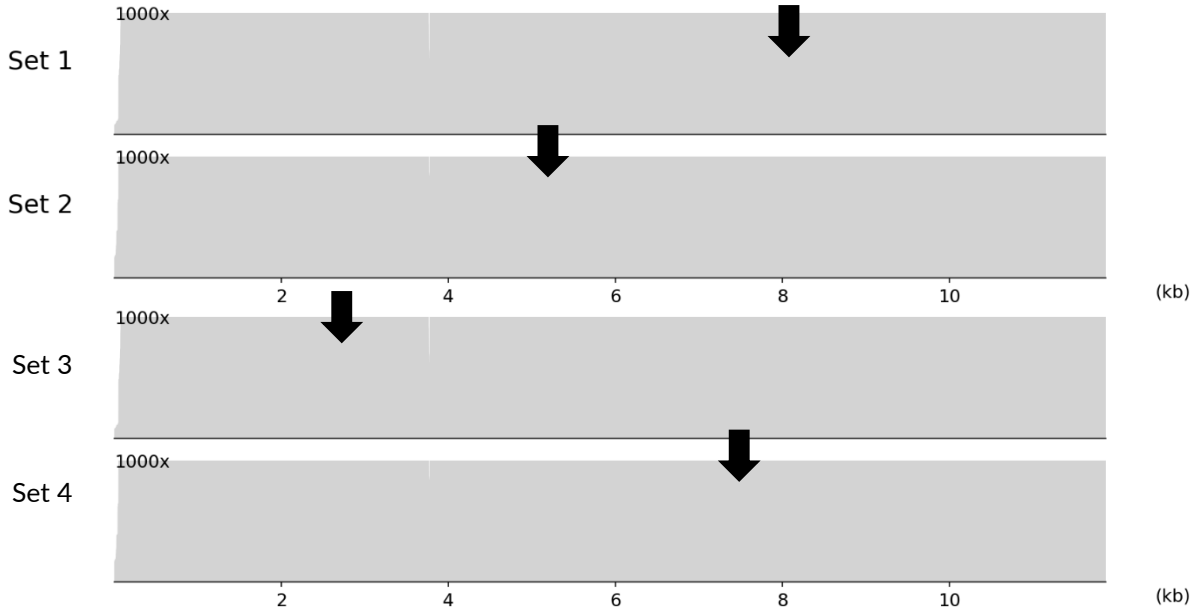

Figure 1: sample X, Y-axes are limited to 1,000x read coverage.

Appendix 4

Sequence variants in vector sequence

Comparison of the mapped reads (Appendix 3) with the provided vector sequence (Appendix 1) using 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 per enzyme,
- for at least one of the primer-sets the coverage is >=30X,
- the variant is identified in both forward and reverse aligning sequencing reads,

Table 1 defines the details regarding the identified sequence variant in the integrated vector sequence for the sample.

Table 1: Identified sequence variant

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, 7 (Cov = coverage): total number of reads that map to this position in the data generated with either primer set 1 (column 5), set 2 (column 7) set 3 (column 9), set 4 (column 11). Column 6 , 8 , 10 and 12 (%): percentage of reads containing the mutation (=mut/cov*100%) in the data of primer set 1 (column 6) set 2 (column 8), set 3 (column 10) and set 4 (column 12).

1. Region	2. Pos	3. Ref	4. Mut	5. Cov	6. %	7. Cov	8. %	9. Cov	10. %	11. Cov	12. %
NeoR	4,990	C	A	2,568	25	23,690	35	1,668	20	4,790	30

Appendix 5

Structural variants in vector sequence

Comparison of the mapped reads (Appendix 3) with the provided vector sequence (Appendix 1) using 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,

Details regarding the identified structural variants in the integrated vector sequence for sample X are shown in Table 1. For all structural variants, intact reads were also found at all positions of the vector-vector breakpoints indicating that (partial) vector sequences have concatemerized. The actual breakpoint sequences are presented in Figure 1. Please note, the number of reads counted for each breakpoint is a slight underestimate of the actual number of reads that contained the breakpoint, 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).

Table 1: Vector-vector breakpoint details

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 = 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. Columns 7-10 (#of reads with fusion primer set 1 (Set 1), primer set 2 (Set 2), primer set 3 (Set 3) and primer set 4 (Set 4) absolute number of reads in which the breakpoint is found. Columns 11-18: (% of reads with fusion): the percentage of fusions at that position compared to the total number of reads that align to that position.

							# of reads with fusion				% of reads with fusion							
1	2		3		4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Break-point	Vector		Vector		Orientation of the breakpoint	Hom	Ins	Set 1	Set2	Set3	Set3	Set1 pos1	Set1 pos2	Set2 pos1	Set2 pos2	Set3 pos1	Set3 pos2	Set4 pos1	Set4 pos2
1	←	123	4,040	←	head to tail	5	-	65	63	65	63	2	0	1	0	2	0	1	0
2	→	500	2,753	←	tail to tail	3	-	245	45	245	45	1	15	0	3	1	15	0	3
3	→	3,500	6,178	→	tail to head	2	-	83	133	83	133	2	1	3	2	2	1	3	2
4	→	4,600	9,806	→	tail to head	2	-	55	428	55	428	1	1	3	3	1	1	3	3

In Figure 1 the sequences corresponding to the breakpoints presented in Table 1 are provided.

1) vector:123 (head) fused to vector:4,040 (tail) with 5 homologous bases
TGCATCGTACGTTGGCCAATCGTCGTCTAGCTGTGTCTGCTTGTCGGCTAGTCCGATGGCACCGTGC
GTCAGGGTCCAAGGTTCA

2) vector:500 (tail) fused to vector:2,753 (tail) with 3 homologous bases
GTGTGTGTGCGACACGTCGTACAGTCAGTGCAGTGCAGTTTGTCCAAGGTTCCGGGCATATGCCC
GGTTGGTTGAAGAGTCAGTGCAGGTGTGCAGTGTGTCAGTGCGTGATGAAACCCTGGGCCC

3) vector:3,500 (tail) fused to vector:6,178 (head) with 2 homologous bases
CATATGTGTACACACGTGTGTCAGTGCCAAATTGGGCATGCAGTGCGTGTGTCATATTAGGGTGTG
TGTAAAGTGTGCCCCGTGTGCAGTGCGTACGTGCAGTGTGTCAAACCCGGTTGGT

4) vector:4,600 (tail) fused to vector:9,806 (head) with 2 homologous bases
TGTGAAACGTGTGCCCAAGTGTCCAGTGCCAGTGCATTGGTGTGTGCCGTGAAACGGAGTTTGTGTA
AAGGTGTCCGTGGTACGTGCCGTATGCAGTGCAGTGCATGTGAGTGTCAGT

Figure 1: Sequences corresponding to the breakpoints presented in Table 1. Red underlined the left side of the breakpoint sequence, in blue italic the right side of the breakpoint sequence is shown. Bases in between can be homologous (shared) in purple and encircled or inserted (novel) as stated above each sequence. Inserted (novel) bases were not found.

Appendix 6

Whole genome wide coverage plots

Figures 1 till 4 show the NGS read coverage across the genome sequence for sample X with the four primer sets. The chromosomes are indicated on the y-axis, the chromosomal positions on the x-axis. The data of primer sets 1-4 show the same integration sites, namely at chr6. The identified integration sites is encircled in green.

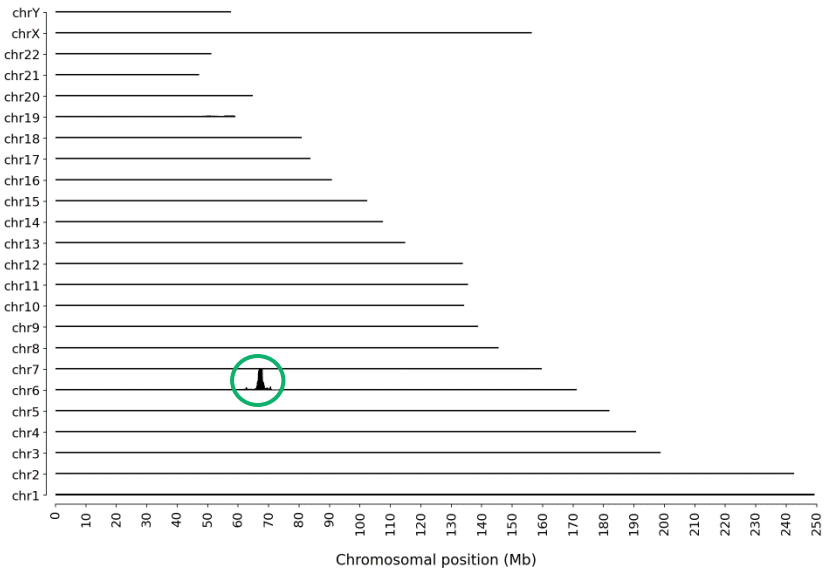

Figure 1: TLA sequence coverage across the xxxx genome using primer set 1 in sample X.

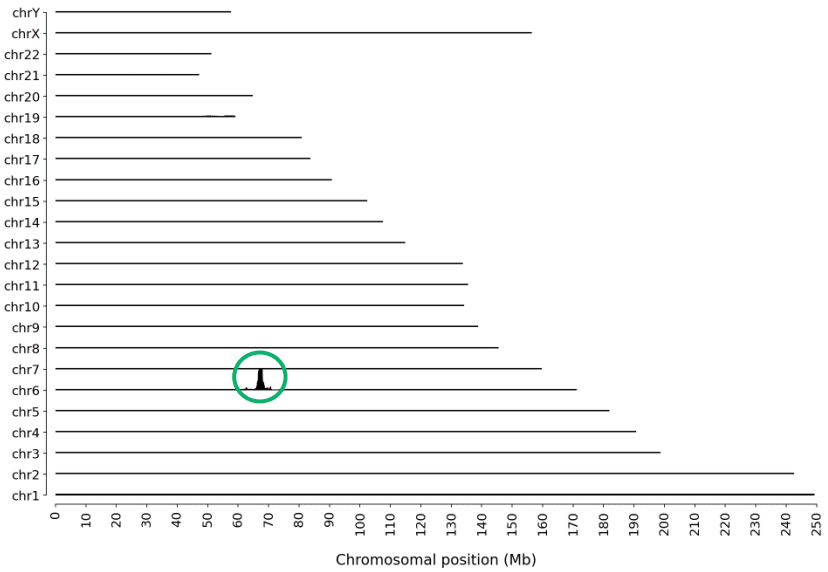

Figure 2: TLA sequence coverage across the xxxx genome using primer set 2 in sample X.

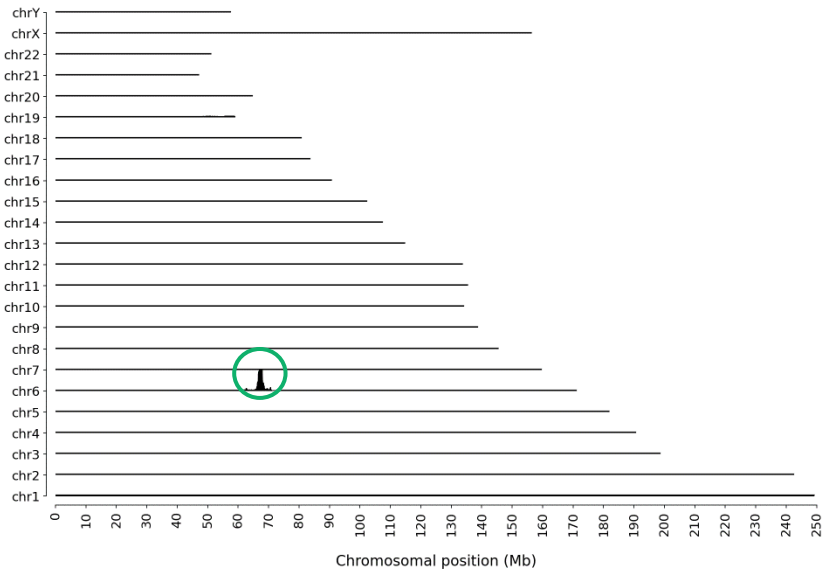

Figure 3: TLA sequence coverage across the xxx genome using primer set 3 in sample X.

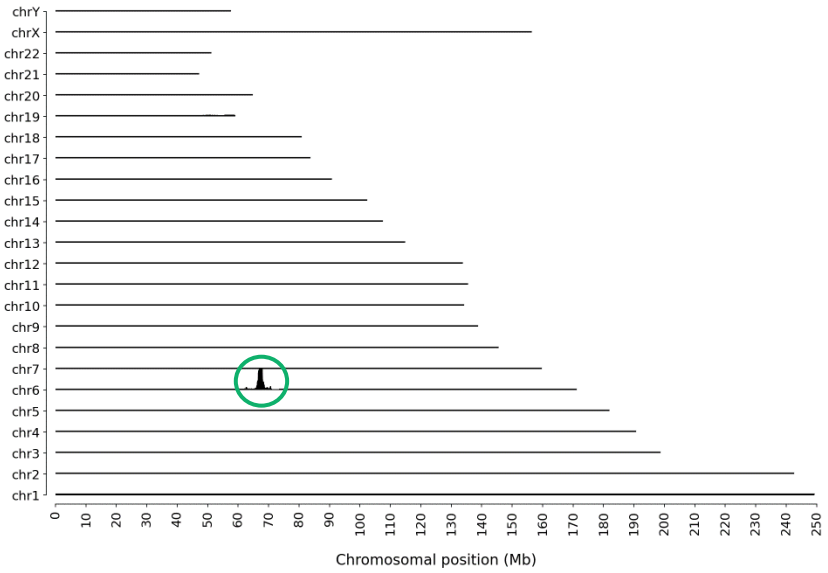

Figure 4: TLA sequence coverage across the xxxx genome using primer set 4 in sample X.

Appendix 7

Integration site information

In Figure 1 the regional view is presented. The actual breakpoint sequences are provided below the figures. The read coverage is expressed in number of reads mapped for primer set 1 (Set 1), primer set 2 (Set 2), primer set 3 (Set 3) and primer set 4 (Set 4) for the sample X respectively, on the specified region in the human genome containing the integration site in chromosome 6 as shown in the legend. On the X-axes the genome location is marked. The Y-axes indicate the number of NGS reads, which are given in a linear scale with an upper limit specified in the figure legend. The regional views provide information on the host genome rearrangements and the genes annotated at the identified breakpoint positions.

I. Read coverage (regional view)

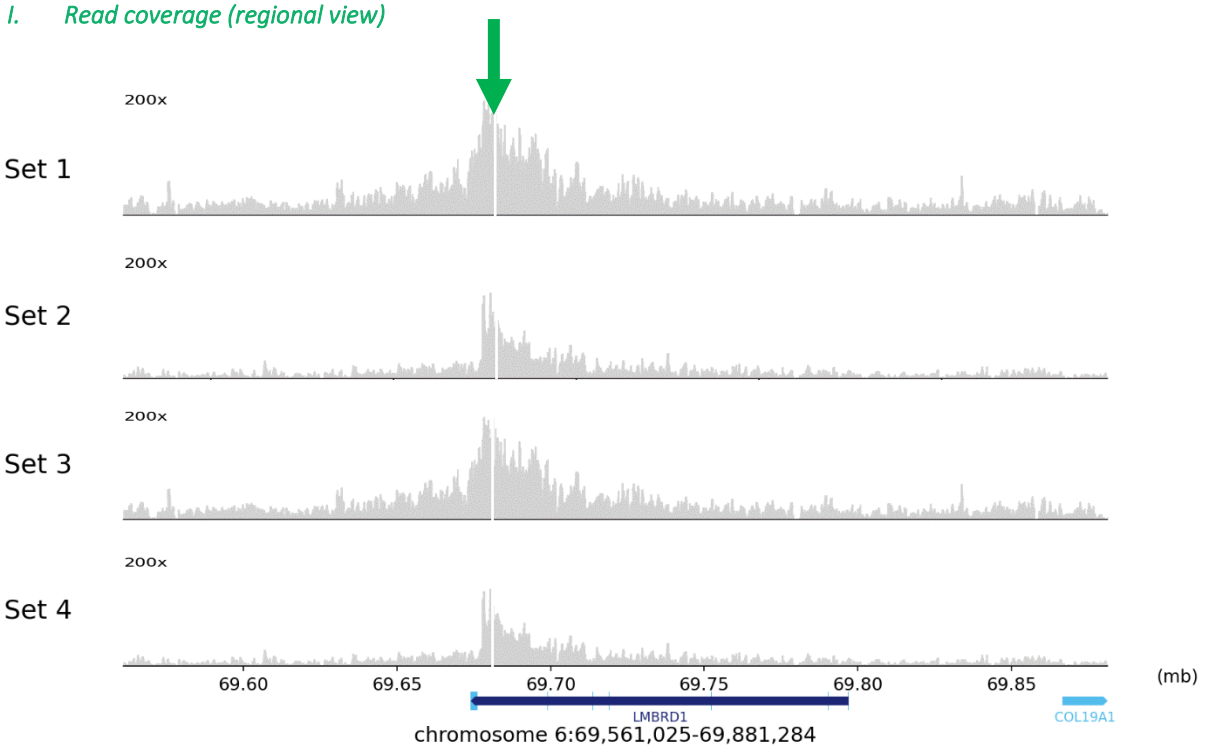

Figure 1: TLA sequence coverage (in grey) across the vector integration locus, chr6:69,561,025-69,881,284 of the sample. The green arrows indicate the location of the breakpoint sequences (1 and 2). Y-axes are limited to 200x.

II. Breakpoint sequences marking the integration site in chromosome 6

The following breakpoint sequence was identified marking the vector integration in chromosome 6, at the position of green arrow in Figures 1.

1. chr6:69,721,202 (tail) fused to Vector:545 (tail) with 4 homologous bases
AATGCTCTGGAATCCTAGGTAAACTCAAAAGGCAGTCTAGGAAACAAGGACTGCAATTCCTAGGCAAC
TCCTAGTGCTTCTGAGGTGCCCAATTTGTGACACGTGACGTCAGTGTGAAACCCAACACACGTGTGTT
GTTGTAAAGTGTGGCGTATGTGCAGCCCC

2. Vector:6,408 (had) fused to chr6:69,721,205 (head) with 3 inserted bases
GTCGTGATGTGTGTGTAAACCGGTTCCAATTGGTTCCAATTGTGTGAATTTGCAGTGAACTAGGCTGG
GCTTACAGTGGACTAGGGTGGCATGTGACCCAGGGAGACAACAGCTAAGGGAGTGCTTGCACCATTC
CTCTCCCAACCCCATGAAGTGCAGCTCACCTCAACAAAGGTGACTCTTTCCTTTGGCCTGAGGAAAGG

The breakpoint sequences and the coverage profiles in figure 1 show that a genomic deletion has occurred in the region of the integration site. The 200 bp genomic sequence in between the two identified breakpoints is deleted.

From this data it is concluded that the vector has integrated at chr6:69,721,102-69,721,205 and shows that a genomic deletion has occurred in the region of the integration site. The 200 bp genomic sequence in between the two identified breakpoints is deleted. The identified breakpoints are located in intron 7 of *LMBRD1*.

Appendix 8

Copy number details

In Figures 1 and 2 the results of the copy number are presented. Approximately 300 total copies have been found for the reference genes 1 and 2 (Figure 1). This number of copies corresponds well to the input amount of approximately 1 ng gDNA based on the average weight of 6.5 pg for the xxxx genome (Piovesan et al., BMC research notes 2019). These results indicate that the cells are diploid and therefore, these genomic genes were used as references to calculate the copy number of the vector elements X and Y. The total copies for the vector elements X and Y were higher than those for the reference genes (1 and 2), indicating that the cell line samples contain more than 2 copies per cell for these vector elements. A copy number of 8 for X and 7 for X has been found in each cell line sample. The discrepancy between X and Y copy numbers can be explained by the presence of vector-vector breakpoints that have been identified by TLA analysis previously. Hence, the cell line sample contains 8 partial copies of vector X.

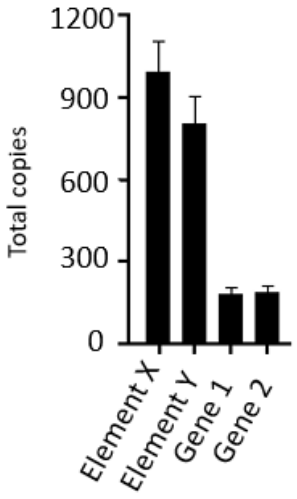

Figure 1: The total copies of the vector elements *X* and *Y* and genomic reference genes *1* and *2* for the sample. The data represents mean ± SD of three experiments performed in duplicate.

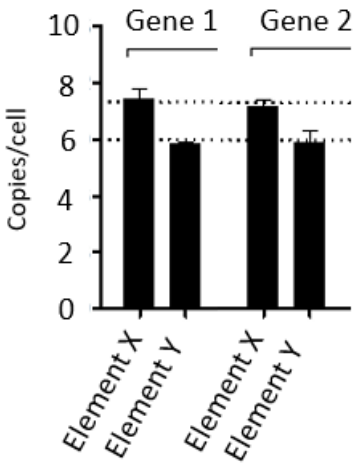

Figure 2: The copies per cell for the vector elements *X* and *Y* as compared to the genomic reference genes *1* and *2* for sample. The data represents the mean ± SD of three experiments performed in duplicate.

Appendix 9

Cergentis Manual: TLA terminology & methods