Full Clonality Assessment Report

Transgene analysis and integration site sequencing using TLA and qPCR

Prepared for:

Customer name:

Internal project number:
Quote number:

Version: 1 (draft)
Date: 17-Feb-22

CERGENTIS Genetic Engineering

Goal

In this study, 1 transgenic CHO cell line with the vector xxxxx sequence was analyzed.
The aim of this analysis was to:

A) Characterize the original MCB:

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.

B) Assess the clonality of MCB

1. Determine the presence of breakpoint sites in subclones.
2. Statistical analysis of results.

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	Clonality
MCB	6 sequence variants, 2 structural variants	Chr3: 169,680,259 - 169,680,260	no	3-5	93 out of 93 subclones positive

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.

All 93 provided subclones were shown to be positive for the unique XXXX-MCB specific breakpoint sequences. These findings support the monoclonal origin of the analyzed MCB at over 95% probability and 95% confidence.

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_v1".

TLA was performed with 4 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
3	NEO	Rv	6,225	X
		Fw	6,542	X
4	GS	Rv	8,154	X
		Fw	8,499	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
MCB p3	1,499,862	149	XXX	XXX	XXX	XXX
MCB p4	1,637,441	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]).

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 all primer sets with coverage in the region,
- 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 2-4.

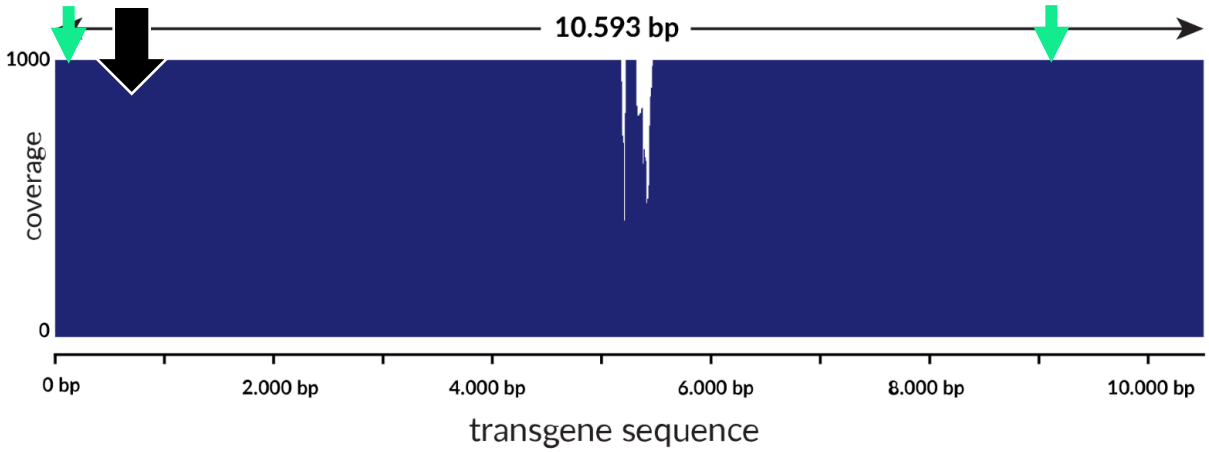

Figure 1: NGS sequencing coverage across the vector with primer set 1. The black arrow indicates the primer location. The green arrows indicate the locations of the identified vector-genome breakpoint sequences (described below). Y-axis is limited to 1000x. In an actual report the data of all primer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

High coverage is observed across the complete vector sequence Vector: 1-10,593. 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		Primer set 3		Primer set 4	
Region	**Position**	**Ref**	**Mut**	**Cov**	**%**	**Cov**	**%**	**Cov**	**%**	**Cov**	**%**
Amp	141	A	C	21,254	20	788	25	542	28	1,247	21
GOI	1,013	T	-4AGTT	1,881	100	7,501	100	1,177	99	1,001	100
GOI	2,956	T	C	1,278	27	34,122	21	2,544	19	856	17
GOI	5,698	A	+1G	751	18	2,221	15	11,521	21	2,745	23
Backbone	9,487	T	G	1,358	100	1,523	99	1,987	100	8,452	100
Backbone	10,037	G	A	2,145	20	854	20	894	19	2,421	21

'+' 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 5-12: quantitative measurements are presented for each primer-set in each sample. Column 5, 7, 9 and 11 (Cov, short for Coverage): total number of reads that map to this position in the data generated with either primer set 1 (column 5), 2 (column 7), 3 (column 9) or 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), 2 (column 8), 3 (column 10) or 4 (column 12).

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	p3	p4	p1	p2	p3	p4
1	ç	6,945	7,657	ç	tail to tail	1	-	100	180	1,452	450	23_17	28_16	17_11	22_14
2	ç	14	5,051	è	head to head	-	2	1,025	1,512	859	15	18_12	22_16	21_19	17_23

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 6_8, means that of all reads that aligned to column 2, 6% contained the breakpoint, and of all reads that aligned to column 3, 8% contained the breakpoint. 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
GAAATCCGAATAATGGGGATTACCCCCTAGATCGAAATTTCGAAGTGGGAGATCGCGTCAGAAGCT
AACGAAGGGATCGCATAGAGAGGACTCGGCTAGAGAGATCGCAGATCGAGATCGAACGTACATCGAT
CAGTCGACTGA

Integration sites

Whole genome coverage plot

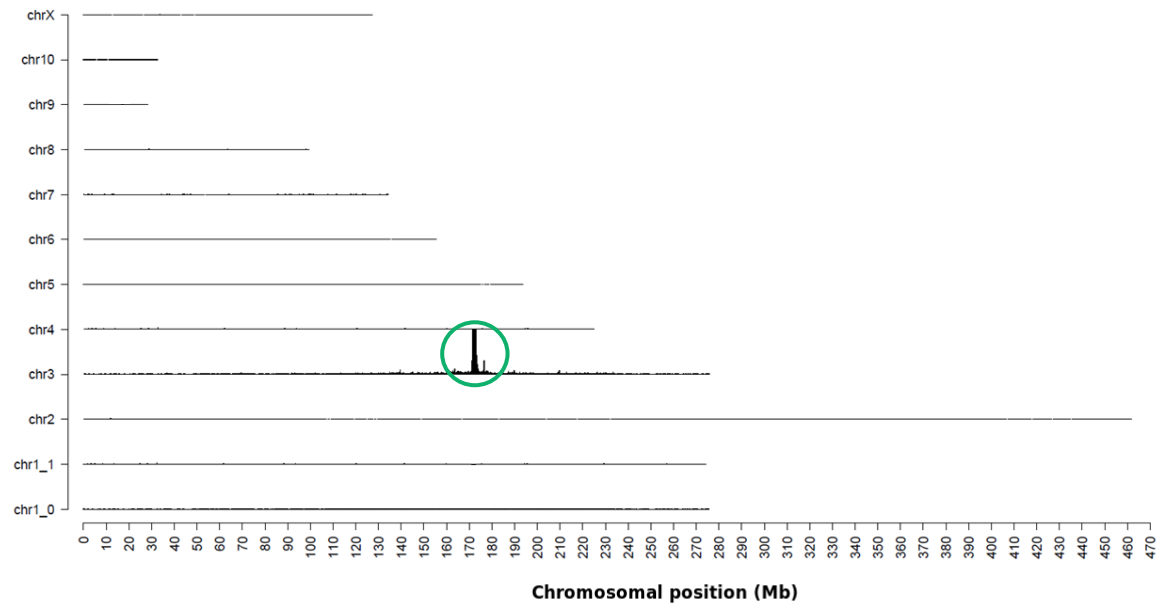

Figure 2: TLA sequence coverage across the Chinese Hamster genome using primer set 1. The chromosomes are indicated on the y-axis, the chromosomal position on the x-axis. Identified integration site is encircled in green. In an actual report the data of all primer sets will be presented.

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

Locus-wide coverage

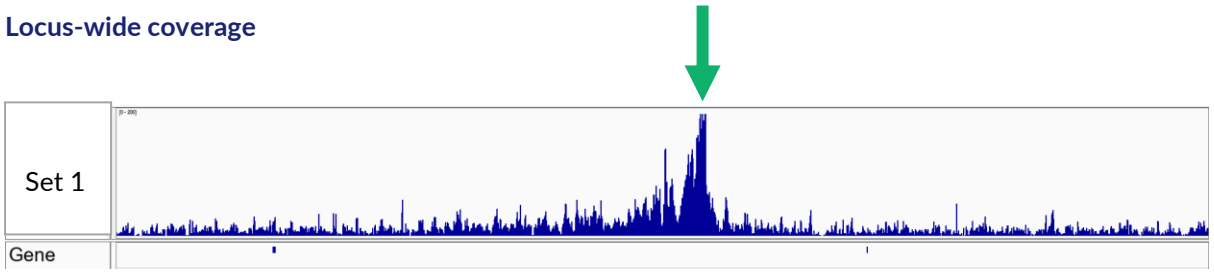

Figure 3: TLA sequence coverage (in blue) across the vector integration locus, chr3:169,530,000-169,830,000. The green arrow indicates the location of the breakpoint sequences. Y-axis is limited to 200x. In an actual report the data of all primer sets will be presented.

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 3 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.

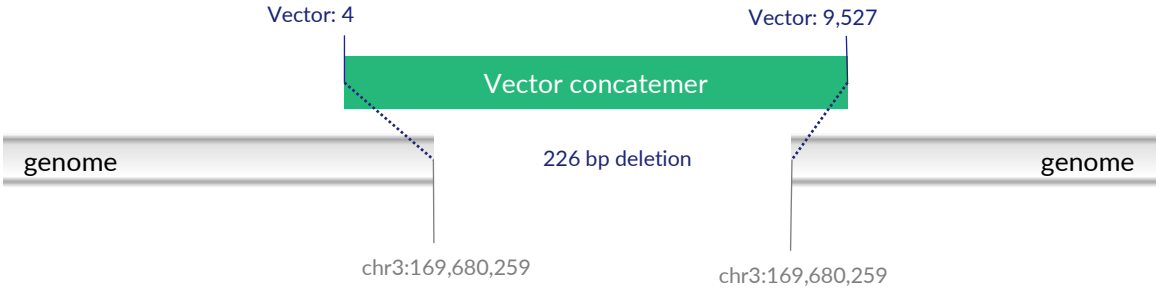

Figure 3: Schematic representation of the integration site.

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		Primer set 3		Primer set 4	
Region	**Position**	**Cov**	**Ratio**	**Cov**	**Ratio**	**Cov**	**Ratio**	**Cov**	**Ratio**
Genome	chr3:169,680,259	3,303	3.5	177	4.2	120	4.1	197	5.2
vector	4	11,558		745		489		1,025	
Genome	chr3: 169,680,260	356	3.9	453	3.5	442	4.6	3,124	3.2
vector	9,527	1,400		1,586		2,025		9,984	

Assessment of clonality of MCB

MCB specific breakpoints

MCB specific breakpoint sequences were previously identified using TLA analysis of MCB. These sequences span the borders of the integration site of MCB and are therefore unique MCB-specific breakpoints. Breakpoint specific qPCR probes with qPCR specific primer sets with fluorescent probes were designed at the breakpoint locations. See table 6 for the breakpoint and breakpoint sequences.

As a negative control, a divergent clone was used. This negative control sample does contain the vector sequence but with another integration site and therefore lacks the MCB-specific breakpoints sequences.

Table 5: Breakpoint and breakpoint sequences

Breakpoint	Sequence
MCB-specific Breakpoint 1: chr3:169,680,259 (tail) fused to Vector: 4 (head) with 5 inserted bases	ATTGCACGTACGTACGTTTGGCAAACACTGTGCC TCGACTGCCGTCGGCGTAACGTCAGCTAGTTTAC CCTGTTGTACACACTGTGATAGGATGGTCGAATC GATGCTAAGCTTCGTAAATCGATATCGATCGTAG CTATGCTAGGGTCGCC
MCB-specific Breakpoint 2: Vector: 9,527 (tail) fused to chr3:169,680,260 (head) with 3 bases homology	CACTATGGGTACGTACGTTATATCCCTGATCGTG CTCGTAGCTGCCTGCTAAGCTAGCTGATGCTGCC GCTTGTTGTACACTTAGGACTGTGATAGCTACGT CGTAAGCTGCTCGATGCTAGATCGCTAGCGGCGG CTAGCTAGTGGCTGAGT

Methodology

qPCR of breakpoint sequences

Four-color TaqMan qPCR was performed a on Bio-Rad CF96x machine, using protocol and primers indicated in table 7 and table 8 respectively. The qPCR plate set-up is shown in table 9.

A single well reaction containing primer sets for each breakpoint as well as 2 housekeeping genes (GADPH and ACTB) was performed for each subclone and control sample. Specific run protocols and primers are described below.

The Housekeeping gene GAPDH and ACTB is expected to give a signal in the positive control as well as the subclones and is an indication of DNA quality and reaction conditions. The negative control is expected to give a signal in the housekeeping gene but not with the MCB-specific breakpoint primers and probes. The water control should give no signals with all primer/probes combinations.

Every reaction was performed in triplicate. The Cq values were determined, and the average and standard deviation were calculated based on the triplicate results.

A sample is considered positive (=containing the unique sequence) if a Cq value of > 0 is obtained.

Table 6: qPCR Run Protocol

		Temperature	Duration
Cycle 1:	1x	XX°C	XX minutes
Cycle 2:	40x	XX°C	XX seconds
		XX°C	XX seconds
		XX°C	XX seconds
		XX°C	XX minutes
	Collect data and analyze		

Table 7: Primers and Probes used for qPCR

Name/View point	Direction	Sequence
MCB-specific Breakpoint 1	FW	XXX
	RV	XXX
	Probe: FAM	XXX
MCB-specific Breakpoint 2	FW	XXX
	RV	XXX
	Probe:Texas 615	XXX
GAPDH	FW	XXX
	RV	XXX
	Probe: Cy5	XXX
ACTB	FW	XXX
	RV	XXX
	Probe: HEX	XXX

Table 8: qPCR plate set-up, controls are marked in blue

		1	2	3	4	5	6	7	8	9	10	11	12
A	Fusion 1	Test Subclone A1	Test Subclone A2	Test Subclone A3	Test Subclone A4	Test Subclone A5	Test Subclone A6	Test Subclone A7	Test Subclone A8	Test Subclone A9	Test Subclone A10	Test Subclone A11	Test Subclone A12
	Fusion 2												
	GAPDH												
	ACTB												
B	Fusion 1	Test Subclone B1	Test Subclone B2	Test Subclone B3	Test Subclone B4	Test Subclone B5	Test Subclone B6	Test Subclone B7	Test Subclone B8	Test Subclone B9	Test Subclone B10	Test Subclone B11	Test Subclone B12
	Fusion 2												
	GAPDH												
	ACTB												
C	Fusion 1	Test Subclone C1	Test Subclone C2	Test Subclone C3	Test Subclone C4	Test Subclone C5	Test Subclone C6	Test Subclone C7	Test Subclone C8	Test Subclone C9	Test Subclone C10	Test Subclone C11	Test Subclone C12
	Fusion 2												
	GAPDH												
	ACTB												
D	Fusion 1	Test Subclone D1	Test Subclone D2	Test Subclone D3	Test Subclone D4	Test Subclone D5	Test Subclone D6	Test Subclone D7	Test Subclone D8	Test Subclone D9	Test Subclone D10	Test Subclone D11	Test Subclone D12
	Fusion 2												
	GAPDH												
	ACTB												
E	Fusion 1	Test Subclone E1	Test Subclone E2	Test Subclone E3	Test Subclone E4	Test Subclone E5	Test Subclone E6	Test Subclone E7	Test Subclone E8	Test Subclone E9	Test Subclone E10	Test Subclone E11	Test Subclone E12
	Fusion 2												
	GAPDH												
	ACTB												
F	Fusion 1	Test Subclone F1	Test Subclone F2	Test Subclone F3	Test Subclone F4	Test Subclone F5	Test Subclone F6	Test Subclone F7	Test Subclone F8	Test Subclone F9	Test Subclone F10	Test Subclone F11	Test Subclone F12
	Fusion 2												
	GAPDH												
	ACTB												
G	Fusion 1	Test Subclone G1	Test Subclone G2	Test Subclone G3	Test Subclone G4	Test Subclone G5	Test Subclone G6	Test Subclone G7	Test Subclone G8	Test Subclone G9	Test Subclone G10	Test Subclone G11	Test Subclone G12
	Fusion 2												
	GAPDH												
	ACTB												
H	Fusion 1	Test Subclone H1	Test Subclone H2	Test Subclone H3	Test Subclone H4	Test Subclone H5	Test Subclone H6	Test Subclone H7	Test Subclone H8	Test Subclone H9	MCB Positive control	Parental Negative Control	Water Control
	Fusion 2												
	GAPDH												
	ACTB												

Statistical assessment

The Standard Practice for Setting an Upper Confidence Bound for a Fraction or Number of Non-Conforming items of the American Society for Testing and Materials (ASTM E2334-09 (Eq. 1)) was used for the setting of a confidence interval of an unknown rate of occurrence of cells with the unique genetic event on the basis of a number of samples tested and all found to have the unique genetic event. The formula is based on: One sided 95 % (not monoclonal) = $1-\sqrt[N]{1-C}$, as published by ASTM E2334-09 (Eq. 1) approach,

The formula is therefore suited to determine the probability of clonality:

One sided confidence interval for clonal derivation = $\sqrt[N]{1-C}$, in which N is tested populations and C confidence interval used.

Results

qPCR of breakpoint sequences

Figure 5 shows the average Cq values and standard deviation of triplicates for the 3 tested breakpoints in each sample and the controls. The 93 derived subclones were positive for both tested MCB specific breakpoints as well as the DNA controls (GAPDH and ACTB), as can be observed by a Cq value above 0 for all 4 breakpoints. The control results were as expected, namely the positive control (XXX) was positive for all MCB specific breakpoints and DNA controls (GAPDH and ACTB). The negative control (parental cell line) was negative for the MCB specific breakpoints and positive for DNA controls (GAPDH and ACTB). The water control was negative for all.

Figure 5a&b: Graphical representation of the average Cq Values and Standard Deviation for qPCR of subclones and controls (n=3).

Statistics

The results (N=93 tested population and 93 positives, C=0.95 confidence interval) give

One sided confidence interval for clonal derivation = $\sqrt[93]{1-0.95}$ = 0.976.

Conclusion

All 93 provided subclones were shown to be positive for the unique xxxx-MCB specific breakpoint sequences. These findings support the monoclonal origin of the analyzed MCB at over 95% probability and 95% confidence.

QC information

Sample and Study details

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

DNA was received frozen in a Thermo Scientific Matrix 96 tube rack, Labeled gDNA xxxxx

Study Personnel

Lab technician
Lab technician qPCR
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 and clonality assessment) 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