Report

Full Clonality Assessment Report for Transgene analysis and integration site sequencing using TLA and qPCR of 1 CHO cell line sample with the vector sequence

Prepared for:	Company name Company address
Customer name:	Name Position within a company email
Internal project number:	XXX
Quote number:	XXX
Version:	1
Date:	26-Jun-2025

Goal

In this study, 1 transgenic CHO with the vector XXX sequence were analyzed.

The aim of this analysis was to:

A) Characterize the original MCB:

1. Study the vector integrity:
   - Determine the presence of sequence variants and their allele frequency.
   - 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. (Optional)

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 transgene & integration site TLA analyses & ddPCR_v3".

Summary

Sample	Vector Integrity	Integration site(s)	Structural variants at the integration site	Copy number estimation (optional)	Clonality
MCB	6 sequence variants, 2 structural variants	Chr3: 169,680,259 - 169,680,260	-	3-5	monoclonal

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.

Abbreviations

Abbreviation	Full name
bp	Base pair
BWA-MEM	Burrows-Wheeler Aligner-Maximal Exact Match
CHO	Chinese Hamster Ovary cell
Cq	Quantification Cycle
DNA	Deoxyribonucleic acid
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
Hom	Homologous bases
Html	HyperText Markup Language
Ins	Insert/novel bases
NGS	Next-generation Sequencing
MCB	Master Cell Bank
P	Primer
PCR	Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
TLA	Targeted Locus Amplification

Methods

TLA, sequencing and data mapping

Viable frozen CHO-K1 cells were used and processed according to the published 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 transgene & integration site TLA analyses & ddPCR_v3".

TLA was performed with 4 independent primer sets specific for the vector sequence, 2 independent primer sets per enzyme (Table 1a&b).

Table 1a: Primers used in TLA analysis using NlaIII

Primer set	Name/View point	Direction	Binding position	Sequence
1	AMP	RV	132	X
		FW	256	X
2	GOI	RV	2,745	X
		FW	3,186	X

Table 1b: Primers used in TLA analysis using DpnII

Primer set	Name/View point	Direction	Binding position	Sequence
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 (SPB) 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 (dual 10-base Illumina indexes) for each sample and each primer set. Library is sent for sequencing (paired-end 2x149 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 (BCL) 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 base calling 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 number 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 bwa mem -M -t 4 -B 7 -w 33 -O 5 -E 2 -T 33 -Y.

The Chinese Hamster CriGri-PICRH1.0 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	1,363,956	149	78	66	90.94	35.80
MCB p2	754,067	149	55	95	90.03	34.64
MCB p3	1,363,956	151	78	66	90.94	35.80
MCB p4	754,067	151	55	95	90.03	34.64

*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 1.11) (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 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 Solvias 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 both 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 primers set 2-4.

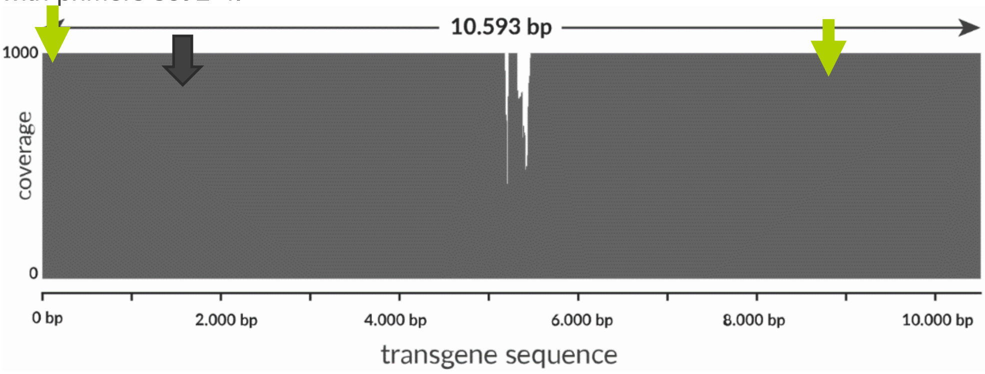

Figure 1: NGS sequencing coverage (in grey) across the vector. Black arrows indicate the primer locations. The green arrows indicate the locations of the identified vector-genome breakpoint sequences (described below). The vector map is shown on the bottom. Y-axes are limited to 1000x. In an actual report the data of all primer sets will be presented.

High coverage is observed across the complete vector sequence Vector: 1-10,593, indicated by the grey areas in Figure 1, demonstrating that this sequence is integrated in this sample 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. Please note that using the 5% filtering criteria, the reported allele frequencies for an individual sequence variant represent the fraction of all the occurrences of that variant among all vector copies integrated in all loci in the entire cell population.

Table 3: Identified sequence variants

Column 1 (Region): the region where the variant is found within the reference sequence. Column 2 (pos): position within the reference sequence. Column 3 (ref): nucleotide present in the reference sequence at this position. Column 4 (mut): observed mutation. Column 5-12: quantitative measurements are presented for each primer-set in each sample Column 5, 7, 9, 11 (Cov = 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) and 4 (column 12).

				Primer set 1		Primer set 2		Primer set 3		Primer set 4	
Region	Pos	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

Vector concatemerization and structural variants

The identified vector-vector breakpoint sites are shown in table 4. In total, 2 structural variants were identified indicating concatemerization. Breakpoint site 2 is in close proximity to the linearization site at position #. Using TLA it is not possible to determine the exact order of (partial) copies and to confirm the presence of at least one complete copy. In samples with multiple vector copies the number of vector-vector junctions may be underestimated.

Table 4: Vector-vector breakpoints

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. Column 7-18 (#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, on position 1 (pos1) and on position 2 (pos2). For example, 6 at p1-pos1 and 8 at p1-pos2 mean that of all reads that aligned to pos1, 6% contained the breakpoint, and of all reads that aligned to pos2 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).

								# of reads with fusion				% of reads with fusion							
Break-point	Vector		Vector		Orientation of the breakpoint	Hom	Ins	p1	p2	p3	p4	p1 pos1	p1 pos2	p2 pos1	p2 pos2	p3 pos1	p3 pos2	p4 pos1	p4 pos2
1	→	6,945	7,657	←	tail to tail	1	-	100	180	1,452	450	23	17	28	16	17	11	22	14
	←	14	5,051	→	head to head	-	2	1,025	1,512	859	15	18	12	22	16	21	19	17	23

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

ATCGGTTTAACAACGGTTAAGCGTTAGTTCCTTGAATCGAAACTTTGGTAACATGTAGCTAGGCTA
ATGCATATGCAATGGATTCGAGACTAATGACCCTTAGGCCTAATTAGGGCTAGAGTCTCGAGAGC
ATTGGGATATCGCGCGGCCTTAGGGACTCTCGGAGACTGGAGCTCAGAGATTTCGGCGATACG
CGATATCGGT

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

GGGTCTAGGGACTGATCGGGATGCCCTGGACTAGGATAGCTAGCTTTTACAAACCCACAATGGA
TTAGAAATCCGAATAATGGGGATTACCCCCTAGATCGAAATTTCGAAAGTGGGAGATCGCGTCAG
AAGCTAACGAAGGGATCGCATAGAGAGGACTCGGCTAGAGAGATCGCAGATCGAGATCGAACG
TACATCGATCAGTCGACTGA

Integration sites

Whole genome coverage plot

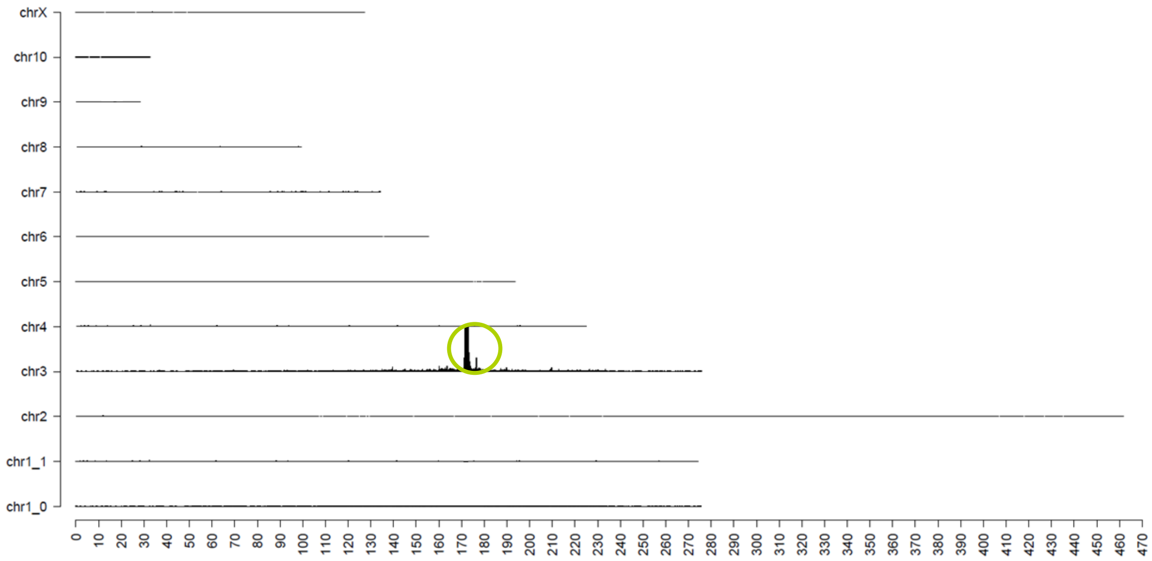

Figure 2: TLA sequence coverage across the Chinese Hamster genome using primer set 1. The chromosomes are indicated on the y-axis, the chromosomal position on the x-axis. Identified integration site is encircled in green. In an actual report the data of all primer sets will be presented.

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

Locus-wide coverage

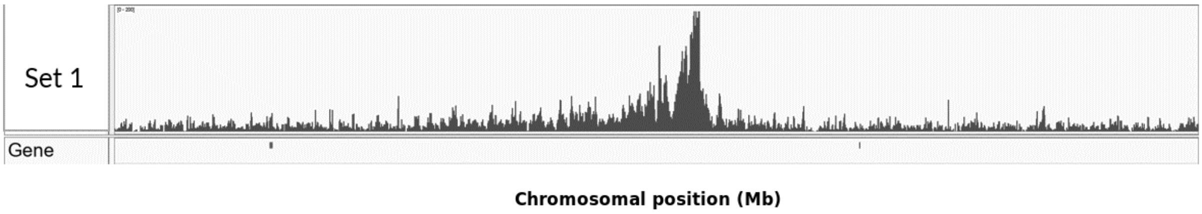

Figure 3: TLA sequence coverage (in grey) across the vector integration locus, chr3:169,530,000-169,830,000. The green arrow indicates the location of the breakpoint sequences. Y-axis is limited to 200x. In an actual report the data of all primer sets will be presented.

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
ATTGCACGTACGTACGTTTGGCAAACACTGTGCCTCGACTGCCGTCGGCGTAACGTCAGCTAGTTTACCCTGTTGTACACACTGTGATAGGATGGTCGAATCGATGCTAAGCTTCGTAAATCGATATCGATCGTAGCTATGCTAGGGTCGCC

3’ integration site:
Vector: 9,527 (tail) fused to chr3:169,680,260 (head) with 3 bases homology
CACTATGGGTACGTACGTTATATCCCTGATCGTGCTCGTAGCTGCCTGCTAAGCTAGCTGATGCTGCCGCTTGTTGTACACTTAGGACTGTGATAGCTACGTCGTAAGCTGCTCGATGCTAGATCGCTAGCGGCGGCTAGCTAGTGGCTGAGT

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 at chr3:169,680,259 -169,680,260 as shown in Figure 4. According to the RefSeq, there are no genes annotated here.

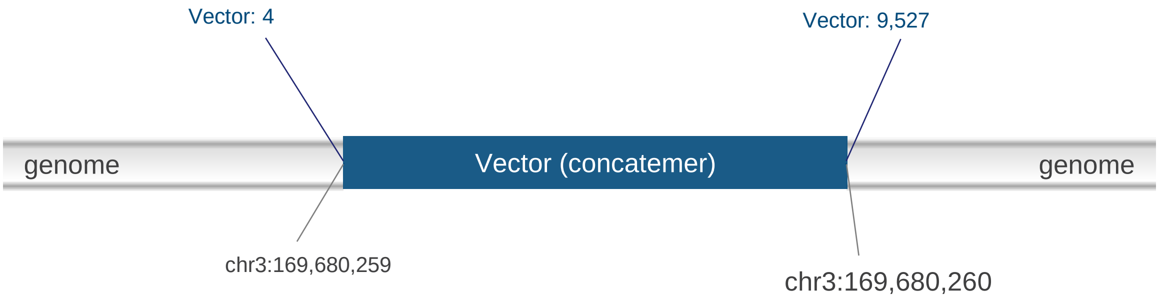

Figure 4: Schematic representation of the integration site.

Copy number estimation (optional)

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.

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 5 for the breakpoint and breakpoint sequences.

As a negative control, the parental cell line was used. This negative control sample does not contain the vector 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	ATTGCACGTACGTACGTTTGGCAAACACTGTGCCTCGACTGCCGTCGGCGTAACGTCAGCTAGTTTACCCTGTTGTACACACTGTGATAGGATGGTCGAATCGATGCTAAGCTTCGTAAATCGATATCGATCGTAGCTATGCTAGGGTCGCC
MCB-specific Breakpoint 2: Vector: 9,527 (tail) fused to chr3:169,680,260 (head) with 3 bases homology	CACTATGGGTACGTACGTTATATCCCTGATCGTGCTCGTAGCTGCCTGCTAAGCTAGCTGATGCTGCCGCTTGTTGTACACTTAGGACTGTGATAGCTACGTCGTAAGCTGCTCGATGCTAGATCGCTAGCGGCGGCTAGCTAGTGGCTGAGT

Methods

qPCR of breakpoint sequences

Three-color TaqMan qPCR was performed a on Bio-Rad CF96x machine, using protocol and primers indicated in table 6 and table 7 respectively. The qPCR plate set-up is shown in table 8.

A single well reaction containing primer sets for each breakpoint as well as 1 housekeeping gene (GADPH) was performed for each subclone and control sample. Specific run protocols and primers are described below.

Housekeeping gene GAPDH 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	95°C	2 minutes
Cycle 2:	40x	95°C	10 seconds
		60°C	20 seconds
		72°C	15 seconds
		95°C	2 minutes
	Collect data and analyze		

Table 7: Primers and Probes used for qPCR

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

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

		1	2	3	4	5	6	7	8	9	10	11	12
A	Fusion 1 Fusion 2 GAPDH	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
B	Fusion 1 Fusion 2 GAPDH	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
C	Fusion 1 Fusion 2 GAPDH	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
D	Fusion 1 Fusion 2 GAPDH	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
E	Fusion 1 Fusion 2 GAPDH	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
F	Fusion 1 Fusion 2 GAPDH	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
G	Fusion 1 Fusion 2 GAPDH	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
H	Fusion 1 Fusion 2 GAPDH	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 Postive control	Negative Control	Water Control

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 2 tested breakpoints and 1 housekeeping gene in each sample and the controls. The 93 derived subclones were positive for both tested MCB specific breakpoints as well as the DNA control (GAPDH), as can be observed by a Cq value above 0 for all 3 breakpoints. The control results were as expected, namely the positive control (MCB) was positive for all MCB specific breakpoints and DNA control (GAPDH). The negative control was negative for the MCB specific breakpoints and positive for DNA control (GAPDH). The water control was negative for all.

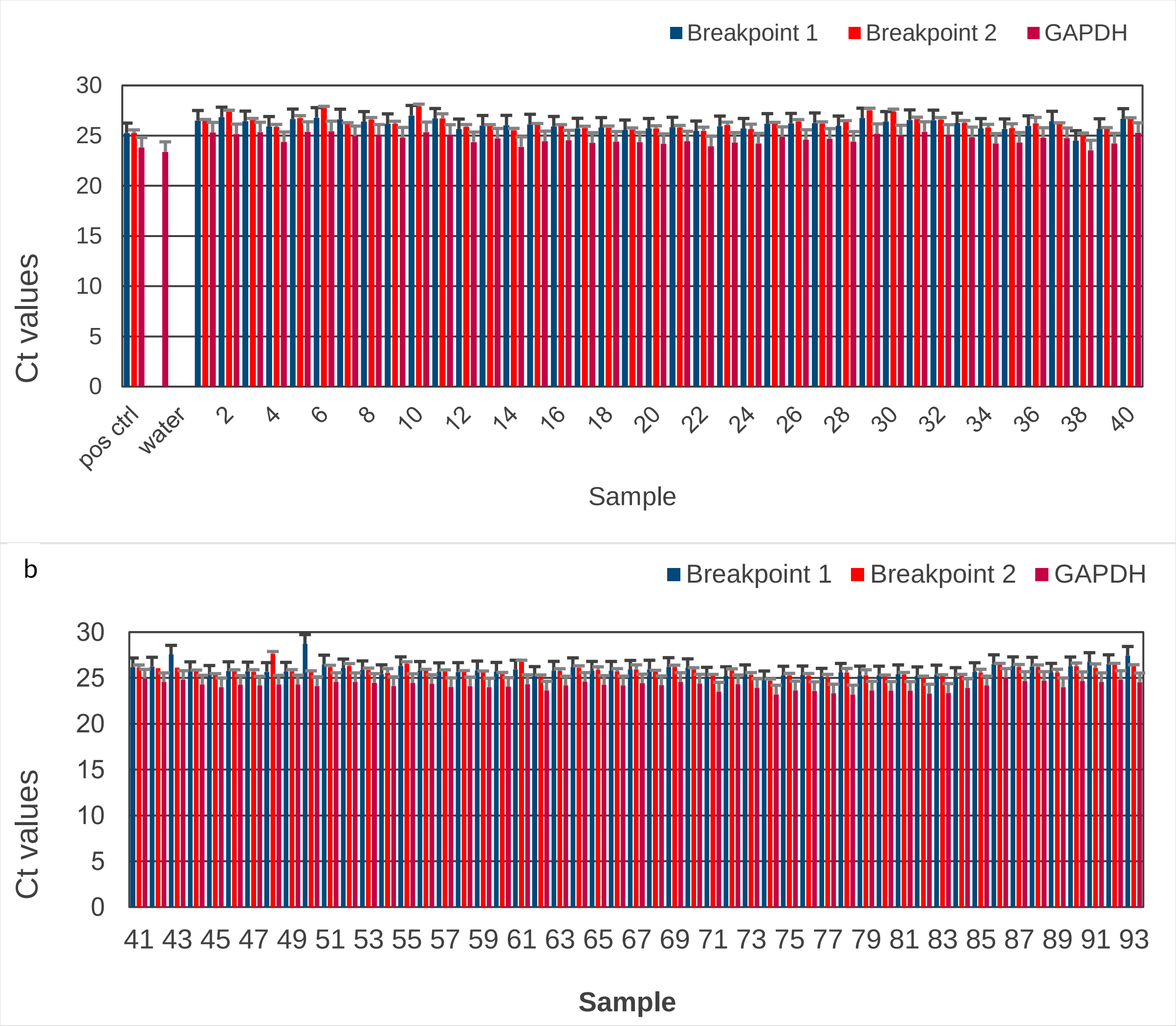

Figure 5a&b: Graphical representation of the average Cq Values and Standard Deviation for qPCR of subclones (n=3)

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 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
Date data analysis
Deviations from the protocol
TLApp version:

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

Study Personnel

Lab technician
Lab technician qPCR
Data Analyst
QC Analysis and Report

TESTING
RvA L 671

Quality control

The results are independently verified and reviewed and are an accurate and complete representation of the study. 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. The copy number estimation and clonality assessment is not included in the scope of this accreditation.

Scientific approval
Date
Signature

*Report will be signed after reviewed by customer

Appendix A

White paper

Targeted Locus Amplification and NGS combined with qPCR-based breakpoint analysis for the assurance of monoclonality in recombinant cell lines

Judith GM Bergboer[1], Martijn JE Kelder[1], Max van Min[1], Nika Tuta[2], Mitja Crček[2], Matjaz Vogelsang[2]

[1] Cergentis, Utrecht, The Netherlands

[2] Novartis, Biologics Technical Development TRD, Mengeš, Slovenia

CERGENTIS
COMPLETE GENE SEQUENCING

NOVARTIS

Abstract

Recombinant protein therapeutics are routinely produced in Chinese hamster ovary (CHO) cells. Minimizing the heterogeneity within a Master Cell Bank (MCB) allows for a well-controlled process that is capable of the consistent manufacture of a product. Regulatory authorities therefore expect that clonal CHO cell lines are used. In this paper, we describe a rapid, reliable and cost-effective assessment of the probability of clonal derivation of recombinant cell populations by combining TLA and NGS with MCB-specific breakpoint qPCR assays and statistical analyses.

Introduction

Recombinant protein therapeutics, or biologics, are an important class of pharmaceuticals for which Chinese hamster ovary (CHO) cells are the most commonly used expression system. The process of developing a CHO cell line expressing a specific recombinant therapeutic is well-established: expression vector(s) encoding the transgene(s) of the therapeutic agent as well as a selectable marker are transfected into the host cell[1]. The resulting culture is a heterogeneous pool of cells that is, in case of random integration, typically the product of multiple independent integrations of expression vector(s) in the CHO genome. The next step is to select and grow those candidates that stably produce highest titers of the protein of interest. The top clone in this process leads to generation of the Master Cell Bank (MCB), which is used in the manufacturing of recombinant biologics.

A clonally derived MCB helps to ensure a robust production process and consistent product quality; FDA guidance[2] instruct cloning the cell substrate "*from a single cell progenitor*" during cell line development, while the EMA guidance stipulates that "*the cell substrate to be used for the production of the monoclonal antibodies should be a stable and continuous monoclonal cell line*[3]. Regulatory authorities therefore request a high assurance of clonality[4,5,6].

The FDA has recommended that two-rounds of limiting dilution cloning (LDC) at sufficiently low seeding densities (≤0.5 cells/well) provide acceptable probability that a cell line is clonal[4,7]. Other approaches have been developed and used either in combination with limiting dilution or as stand-alone methods[8]. These approaches include, but are not limited to, use of the ClonePix system[9], flow cytometry-mediated single cell sorting[10,11] and automated cell imaging systems[12]. Some ongoing clinical programs however employ legacy cell lines that were created before the industry had access to such practices and methods and may not satisfy current regulatory expectations for clonality when filing for market access.

Supporting evidence can be requested at several stages (e.g. IND or BLA) in the filing process. To provide supporting evidence the following additional tests can be considered: sub-clone analysis whereby a vial of the Master Cell Bank is plated as single cells (using LDC), expanded, and characterized using phenotypic analyses (e.g. cell doubling time, specific productivity etc.), product quality testing and genotypic analyses (e.g. fluorescence in situ hybridization (FISH) or Southern blotting) to evaluate individual integration sites[13].

Targeted Locus Amplification (TLA) combined with next-generation sequencing (NGS) allows for complete characterization of integration sites and the integrated transgene/vector sequence in any species[14]. This technology has been widely adapted by the pharmaceutical industry in various phases of CLD[15,16,17,18].

In this paper we describe a general and cost-effective approach to analytically assess the probability of monoclonal derivation of recombinant cell populations (a similar approach has been presented by Aebischer-Gumy *et al.*[16]). Using TLA combined with NGS, unique genetic features of the MCB can be identified, e.g. the breakpoint sequence between genome and the plasmid which characterizes the integration site or vector-vecor junctions of an integrated concatemer. Clonally derived cell populations generated from the MCB can be analyzed by qPCR for the presence or absence of these unique genetic features. Compared to other cited technologies, such as Southern blotting and fluorescence in situ hybridization (FISH), qPCR breakpoint analyses allow for the analysis of a large number of monoclonal-derived cell populations for unique MCB-specific breakpoints. The methods and statistical analyses described in this paper therefore enable an efficient assessment of the probability of clonality.

Whilst we here describe the analysis of a CHO cell bank, the approach equally well applies to other cell types used in the production of biopharmaceuticals or viruses, such as those of human (e.g. HEK293) and murine (NS0 and Sp2/0) origins.

Material and methods

Cell line generation

A stable monoclonal antibody (mAb)-producing Chinese hamster ovary (CHO) cell bank was generated using the DHFR/MTX selection process. Briefly, a linearized expression vector encoding the heavy and light chain genes was transfected into CHO parental cells via electroporation. Transfected cells were grown in selective growth medium at 36.5ºC and 10% CO2 to recover stable integrants (i.e cells that have integrated the expression vector into their genome). Pools were single cell cloned with a limiting dilution approach combined with imaging. Cells were seeded at a final density of 0.3 vc/well in 96 well plates. Cells were expanded and assessed for productivity, growth and product quality. The top clone was used to generate primary seed lot (PSL), which was further scaled up to generate the Master Cell Bank (MCB).

TLA/NGS

TLA followed by NGS, as well as bioinformatical analyses were performed on PSL vial as described[14]. Region of interest was targeted using the transgene-specific primer set. TLA products were sequenced on an Illumina sequencer generating paired-end, 2x150 bp reads. Mapping was performed using BWA-SW (Smith-Waterman algorithm[19]) with the Chinese hamster genome sequence (GCF_003668045.1 assembly) as reference genome.

Analytical subcloning and DNA isolation

MCB vial was thawed and cells were cultivated in serum-free medium at 36.5ºC and 10% CO2 before single-cell isolation was carried out with Cytena single-cell printer and imaging. Cells were dispensed into 96-well plates prefilled with 100µl serum-free growth medium. Cell imaging was performed at days 0, 1, 10, and 18 after single-cell deposition. Clones were expanded in 24DW plates and 200uL of each culture was used for DNA extraction using KAPA Express Extract kit following manufacturer's instructions. The remaining culture for each subclone was cryopreserved.

Quantitative PCR

DNA extracts were assessed for the presence of MCB specific integration site using qPCR. TaqMan assays targeting genome-vector junction site and a CHO-genome region (*GLUC* region was used to control for successful extraction of gDNA), respectively, were custom designed by Applied Biosystems. Quantitative PCR was performed in 10 µL total reaction volume using 2X TaqMan Universe PCR master mix. The following thermal parameters were used: UNG nuclease activation at 50°C for 2 minutes and initial denaturation at 95°C for 10 minutes, followed by 40 cycles at 95°C for 15 seconds, and 60°C for 1 minute. Only DNA extracts with $Ct_{GLUC}<30$ (indication of successful extraction of gDNA) were considered in the interpretation. Every reaction was performed in triplicates.

Statistical assessment of probability of clonality

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)[20]) presents a method 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 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 the confidence interval used.

Table 2 shows the effects of increasing N on the probability of clonality using a 95% confidence interval. Supplementary **Table 1** shows the effects of increasing N on the probability of clonality using a 90%, 95% and 99% confidence interval, respectively.

Table 2: Calculations of one sided 95% confidence intervals for clonal derivation.

Number of clonally derived populations tested and found to conform	One-sided 95% confidence interval for non-clonal derivation	One-sided 95% confidence interval for clonal derivation
1	0.95	0.05
2	0.776	0.224
3	0.632	0.368
4	0.527	0.473
5	0.451	0.549
10	0.259	0.741
20	0.139	0.861
50	0.058	0.942
60	0.049	0.951
75	0.039	0.961
93	0.032	0.968
100	0.03	0.97
186	0.016	0.984

Results

TLA sites

A single integration site was observed in the analyzed cell culture, suggesting the occurrence of a single integration event of the transgene into genome of the CHO cell, from which the MCB originates. The integration site was observed on chromosome 9 in the Chinese hamster genome assembly and the genome-vector junction sequences described in **Table 3** were identified.

Junction	Sequence
NW_020822657.1 (picr_41_new): 15329882 (+) – transgene:17(+)	5’- TTTCAAGGCCTAGGGTAACACGTTTGGAATCAACTTCTTGTCTGCCAG AGACGGTGTCTAT||*NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN* *NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN* *GC*
transgene:11290(+) – NW_020822657.1 (picr_41_new): : 15329887 (+)	5’- *NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN* *NNNNNNNNNNNNNN*||AAATGTCTGTTTTTAGGTGGCAGACTTGTTTG GGGGGCAGAGTCTTGCTATGTGGGGACCTGGCTAGCTTGGAATACTAT AT

Table 3: Nucleotide sequences at genome-vector junction sites. Sequence orientation is presented with (-) or (+); transgene sequence is presented in italics, junction is marked (||). Transgene and CriGri genome positions (bp) at junctions are indicated. Sequences of TaqMan assays for detection of identified junction sites are marked in *green* (forward and reverse primer) and *blue* (probe). True transgene sequence at junction sites is consealed (ie. Ns are used instead).

Next, the presence of the identified, MCB-specific integration site was assessed in DNA extracts from 60 analytical subclones using TaqMan assays targeting CriGri_Chr9:3752476-transgene:17 junction (*ASSAY1*) and *GLUC* region in the Chinese hamster genome, respectively. The presence of the NW_020822657.1 (picr_41_new): 15329882 – transgene:17 genome-vector junction site was confirmed in all analyzed subclones (**Figure 1**). This finding supports the monoclonal origin of the analyzed mAb producing cell bank at over 95% probability and with 95% confidence.

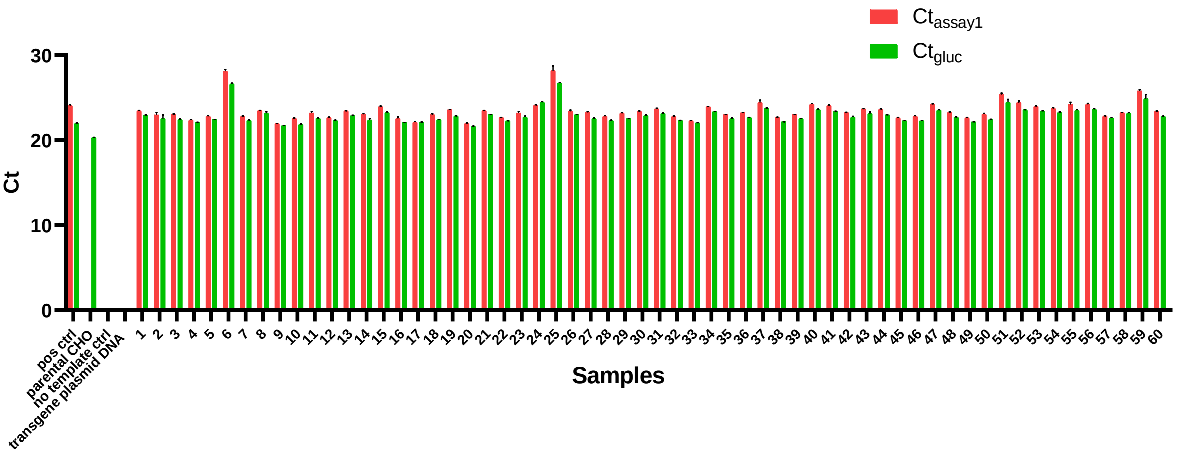

Figure 1: Confirming presence of MCB-specific junction site in all analyzed 60 MCB-derived analytical subclones The performance of both TaqMan assays was adequate as the following acceptance criteria were met: successful amplification of *GLUC* and *ASSAY1* in MCB (pos ctrl); successful amplification of *GLUC* in non-transfected parental CHO cell line (neg ctrl); no amplification in either the transgene plasmid DNA sample or no template control. Amplification of *ASSAY1* and *GLUC* was confirmed in all 60 analytical subclones.

Conclusion and Discussion

Our work demonstrates that the use of TLA followed by NGS allows a detailed analysis of integrated transgenes, transgene integration sites and the identification of unique genetic features in a specific cell line. Cell bank homogeneity was assessed by testing populations clonally derived from the cell bank for the presence or absence of identified genetic features.

The intrinsic plasticity of the CHO genome[21,22,23] can result in the loss of specific genetic sequences of the MCB in subclones. This highlights the advantage of the analysis of at least 2 MCB specific breakpoints (**Table 4**). Clones with negative qPCR results can also be further evaluated using TLA to determine if they do share the MCB integration site. In addition, the evaluation of a subset of subclones over time using TLA and NGS provides information about the genetic stability of the integration site and integrated transgene sequences, which are key for a stable recombinant therapeutic protein production process.

Table 4: Potential outcomes from TLA and qPCR breakpoint analysis experiment

Event	Cause	Solution
Not all integration sites are identified in original MCB	Integration sites with partial integrated vector present in MCB	Perform TLA with multiple primer sets
MCB-specific breakpoint is not confirmed in at least one analytical subclone	Genetic instability of the subclone or MCB is not clonal	Use at least 2 breakpoints in the qPCR breakpoint analysis or Evaluate other MCB-specific integration site (if present) or Perform TLA on 'negative' subclone

In conclusion, we have described a cost-effective approach to analytically assess the probability of clonal derivation of recombinant cell populations, by combining TLA and NGS with MCB-specific breakpoint qPCR assays and statistical analyses.

References

1. Wurm FM. (2004) Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol.* 22:1393–1398.

2. ICH FDA (1998). Topic Q5D Quality of Biotechnological Products: Derivation and Characterisation of Cell Substrates Used for Production of Biotechnological/Biological Products. *CPMP/ICH/294/95*

3. EMA (2016). Development, production, characterisation and specifications for monoclonal antibodies and related products. *EMA/CHMP/BWP/532517/2008*

4. Kennett S. (2014). Establishing Clonal Cell Lines – A Regulatory Perspective Black Cell, Blue Cell, Old Cell, New Cell? *WCBP*

5. Novak R. (2017). Regulatory perspective on the evaluation of clonality of mammalian cell banks. *CDER/OPQ/OBP/DBRRI*

6. Welch J. (2017). Tilting at clones: A regulatory perspective on the importance of "Clonality" of mammalian cell banks. *CDER/OPQ/OBP/DBRRIV*

7. Wu P *et al.* (2018) Tools and methods for providing assurance of clonality for legacy cell lines. *Cell Culture Engineering XVI*

8. Gross A *et al.* (2015) Technologies for single-cell isolation. *Int J Mol Sci.* 16:16897–16919

9. Newman ENC and Whitney D. (2007) Rapid automated selection of mammalian cell colonies by cell surface protein expression. *Nat Methods*. 4, 462

10. Misaghi S *et al.* (2016) Slashing the timelines: Opting to generate high-titer clonal lines faster via viability-based single cell sorting. *Biotechnol Prog.* 32:198–207

11. DeMaria CT *et al* (2007). Accelerated clone selection for recombinant CHO CELLS using a FACS-based high-throughput screen. *Biotechnol Prog.* 23:465–472

12. Evans K *et al.* (2015). Assurance of monoclonality in one round of cloning through cell sorting for single cell deposition coupled with high resolution cell imaging. *Biotechnol Prog.* 31:1172–1178

13. Rawatt R. (2016) Regulatory Consideration for Biotechnology Products: Clonality of the Production Cell Bank. *Informa Life Sciences Annual Cell Line Development and Engineering Conference, 11 – 13 April, 2016*

14. De Vree P *et al.* (2014). Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping *Nature Biotechnology* 32: 1019-1025

15. Kaas CS *et al.,* (2015) Deep Sequencing Reveals Different Compositions of mRNA Transcribed From the F8 Gene in a Panel of FVIII-producing CHO Cell Lines. *Biotechnol J.* 10:1081-1089

16. Aebischer-Gumy C *et al.* (2018) Analytical Assessment of Clonal Derivation of eukaryotic/CHO Cell Populations. *J Biotechnol.* 20;17-26.

17. Boyd D *et al.* (2018) Isolation and Characterization of a Monoclonal Antibody Containing an Extra Heavy-Light Chain Fab Arm. *MAbs* 10:346-353.

18. Aeschlimann SH *et al.* (2019) Enhanced CHO Clone Screening: Application of Targeted Locus Amplification and Next-Generation Sequencing Technologies for Cell Line Development. *Biotechnol J.* 14:e1800371

19. Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler Transform. *Bioinformatics* 1:589-595.

20. ASTM E2334-09 (2018), Standard Practice for Setting an Upper Confidence Bound for a Fraction or Number of Non-Conforminga items, or a Rate of Occurrence for Non-Conformities, Using Attribute Data, When There is a Zero Response in the Sample, *ASTM International, West Conshohocken, PA*

21. Barnes LM *et al.* (2003). Stability of protein production from recombinant mammalian cells. Biotechnol. Bioeng. 81:631–639

22. Vcelar S *et al.* (2018). Changes in chromosome counts and patterns in CHO cell lines upon generation of recombinant cell lines and subcloning. Biotechnol. J. 13: e1700495

23. Wurm, F (2013). CHO quasispecies—implications for manufacturing processes. Processes 1:296–311

24. Frye C et al., (2016). Industry view on the relative importance of "clonality" of biopharmaceutical-producing cell lines. Biologicals 44, 117-122

Supplementary material

Supplementary Table 1: Calculations of one sided 90%, 95% and 99% confidence intervals for clonal derivation.

Number of clonally derived populations tested and found to conform	One sided 90% confidence interval for clonal derivation	One sided 95% confidence interval for clonal derivation	One sided 99% confidence interval for clonal derivation
1	0.100	0.050	0.010
5	0.631	0.549	0.398
10	0.794	0.741	0.631
20	0.891	0.861	0.794
30	0.926	0.905	0.858
40	0.944	0.928	0.891
50	0.955	0.942	0.912
60	0.962	0.951	0.926
70	0.968	0.958	0.936
80	0.972	0.963	0.944
90	0.975	0.967	0.950
100	0.977	0.970	0.955
200	0.989	0.985	0.977

Contact

Address: Cergentis B.V.
Yalelaan 62
3584 CM Utrecht
The Netherlands

Website: www.cergentis.com

Phone: +31 30 760 1636

General: info@cergentis.com
Sales: sales@cergentis.com