Case Study

Full Clonality Assessment

1. Background

Chinese hamster ovary (CHO) cells are widely used to produce biotherapeutics. The product gene is introduced via a vector in a process called transfection, resulting in heterogeneous cells. Single-cell cloning is then performed and high-producing cell lines are selected for use. To ensure consistent quality, regulatory authorities require proof of clonality, meaning the cells must originate from a single transfected ancestor cell.

2. Objectives

In this study, one transgenic CHO cell line with the vector X sequence was analyzed. The aims of this analysis was to (1) characterize the master cell bank (MCB) and (2) assess its clonality.

3. Methods

MCB Characterization

Targeted Locus Amplification (TLA), sequencing, and data mapping were performed on viable frozen CHO-K1 cells. Four vector specific primer sets were selected, allowing for the amplification of the vector and the locus of interest (table 1). PCR products were purified, library prepped using the Illumina's Nextera DNA Flex library prep protocol and sequenced on an Illumina sequencer. The sequencing run generated base call information, which was converted into FASTQ files via Illumina's bcl2fastq tool in a process called demultiplexing. After the conversion to FASTQ files, reads were mapped to either the vector sequence or host genome using BWA-MEM version 0.7.15-r1140, settings mem -B7. The Chinese Hamster CriGri-PICRH 1.0 genome assembly GCF_003668045.3 was used as host reference genome sequence. Quality control metrics of the sequencing run were recorded (table 2). The presence of sequence variants is determined using samtools mpileup (samtools version 1.3.1) and reported if they met the criteria on table 3. Structural variants are verified by 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 if meet criteria on table 3. Integration sites were identified based on coverage peaks and breakpoint sequences between vector and host genomes.

MCB Clonality Assessment

Specific breakpoints were identified in the MCB using TLA. These breakpoints are exclusive to the MCB and serve as genetic markers. To detect these markers, qPCR probes and primers that specifically target the MCB breakpoints were used. A single well reaction containing primer sets for each breakpoint as well as 2 housekeeping genes (GADPH and ACTB) was performed on the control sample and on 93 subclones deriving from the MCB. GAPDH and ACTB are 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 MCB specific breakpoints should be positive in the positive control, while 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.

Table 1

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

Table 2

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

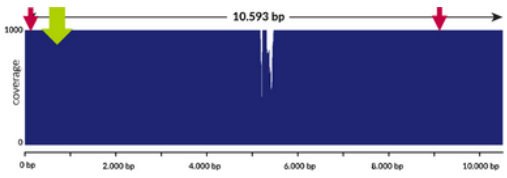
Table 3

Reporting criteria
Sequence variants
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
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 sequencing artefacts

4. Results

Vector integrity

NGS sequencing coverage across the vector with primer set 1. The green arrow indicates the primer location. The pink arrows indicate the locations of the identified vector-genome breakpoint sequences. Same results were obtained with primer set 2-4.



Sequence variants

Detected sequence variants are presented in table 4. All sequence variants at or near 100% mutation frequency were detected in this sample and most likely represent deviations present in the provided reference sequence of the vector before its introduction into the sample.

Table 4

Region	Position	Ref	Mut	Primer set 1		Primer set 2		Primer set 3		Primer set 4	
				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

In total, 2 structural variants were identified. The left side of the fusion is in red, the right side of the fusion is in blue, any homologous bases are dark blue, and inserted bases are green.

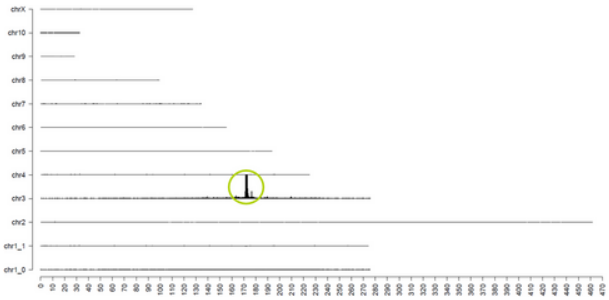
							# of reads with fusion				% of reads with fusion			
Break-point	Vector		Vector		Orientation	Hom	Insert	p1	p2	p3	p4	p1	p2	p3	p4
1	→	6,945	7,657	←	tail to tail	1	-	100	180	1,452	450	23_17	28_16	17_11	22_14
2	←	14	5,051	→	head to head	-	2	1,025	1,512	859	15	18_12	22_16	21_19	17_23

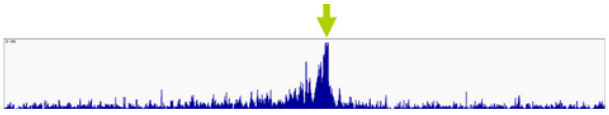
vector:6,945 (tail) fused to **vector:7,567 (tail)** with **1 homologous base**: ATCGGTTTAACAACGGTTAAGCGTTAGTTCCTTGAATCGAAACTTTGGTAACAT GTAGCTAGGCTAATGCATATGCAATGGATTCGAGACTAATGACCCTTAGGCCTAATTAGGGCTAGAGTCTCGAGAGCATTGGGATATCGCGCGGCCTTAG GGACTCTCGGAGACTGGAGCTCAGAGATTTCGGCGATAC GCGATATCGGT

vector:14 (head) fused to **vector:5,051 (head)** with 2 **inserted bases**: GGGTCTAGGGACTGATCGGGATGCCCTGGACTAGGATAGCTAGCTTTTACAAAC CCACAATGGATTAGAAATCCGAATAATGGGGATTACCCCCTAGATCGAAATTTCGAAAGTGGGAGATCGCGTCAGAAGCTAACGAAGGGATCGCATAGAG AGGACTCGGCTAGAGAGATCGCAGATCGAGATCGAACGTACATCGATCAGTCGACTGA

Integration sites

TLA sequence coverage across the genome using primer set 1, and across the vector integration locus, chr3:169,530,000-169,830,000 is shown below.





The following breakpoint sequences were identified:

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

ATTGCACGTACGTACGTTTGGCAAACACTGTGCCTCGACTGCCGTCGGCGTAACGTCAGCTAG
TTTACCCTGTTGTACACACTGTGATAGGATGGTCGAATCGATGCTAAGCTTCGTAAATCGATAT
CGATCGTAGCTATGCTAGGGTCGCC

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

CACTATGGGTACGTACGTTATATCCCTGATCGTGCTCGTAGCTGCCTGCTAAGCTAGCTGATG
CTGCCGCT**TGT**TGTACACTTAGGACTGTGATAGCTACGTCGTAAGCTGCTCGATGCTAGATCG
CTAGCGGCGGCTAGCTAGTGGCTGAGT

From this data it is concluded that the vector has integrated at **chr3: 169,680,259 - 169,680,260**. No genomic rearrangements have occurred in the region of the integration site.

Copy number estimation

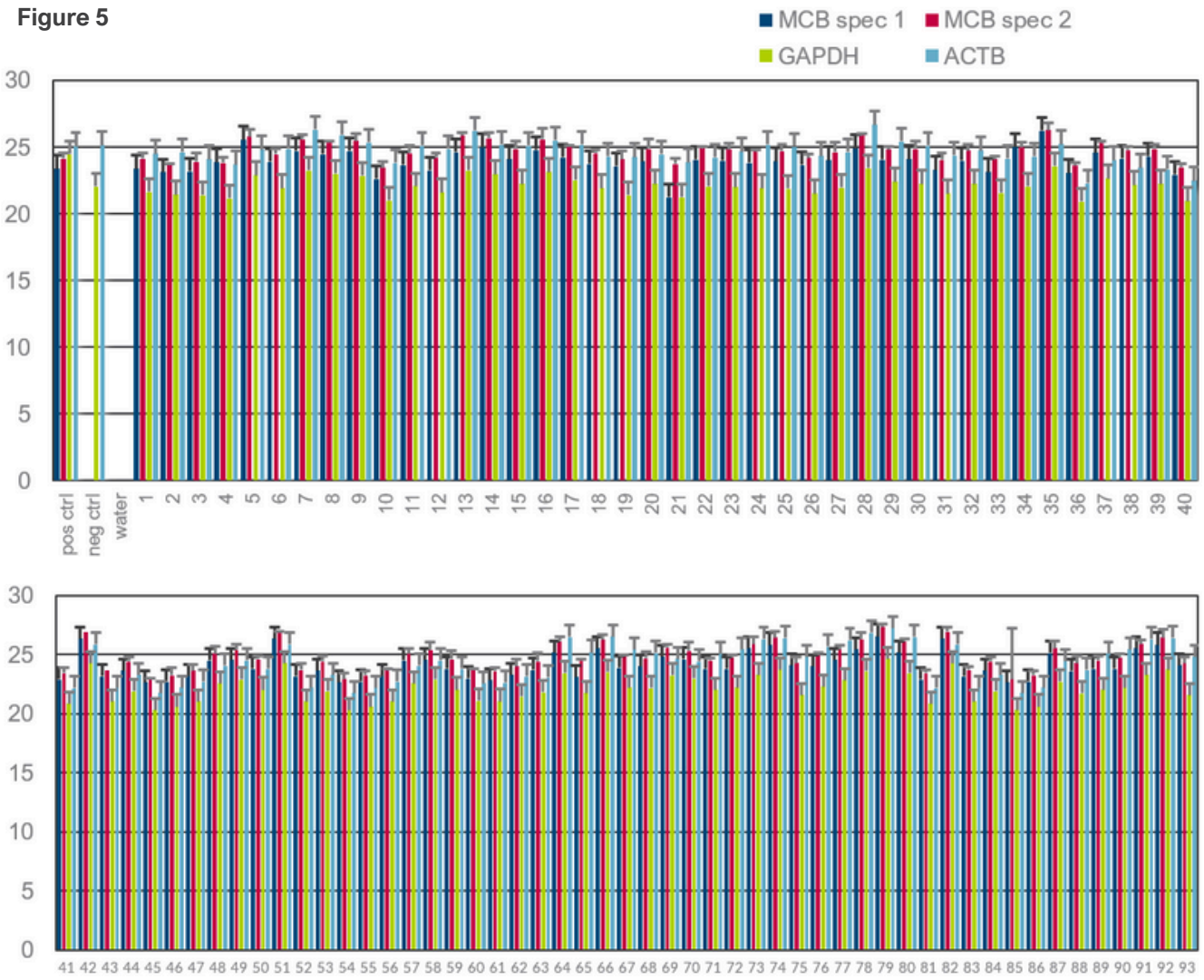
In the MCB sample, the coverage on the vector-side is 4-5 times higher than on the genome-side of the integration site. 1 integration site and 2 vector-vector breakpoints are found. The copy number is estimated to be 3-5 copies.

Region	Position	Primer set 1		Primer set 2		Primer set 3		Primer set 4	
		Cov	%	Cov	%	Cov	%	Cov	%
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	

qPCR of breakpoint sequences

Figure 5 shows the average Cq values and standard deviation of triplicates for the 2 MCB specific breakpoints and 2 housekeeping genes tested. 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 tested 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. The results (N=93 tested population and 93 positives, C=0.95 confidence interval) give one sided confidence interval for clonal derivation = 0.976.

Figure 5



5. Conclusion

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 X-MCB specific breakpoint sequences. These findings support the monoclonal origin of the analyzed MCB at over 95% probability and 95% confidence.

solvias

Why partner with us?

- CDMO/CRO
- Founded in 1999
- 800+ team members
- 175+ PhD-level scientists
- GMP, GLP, ISO9001 certified
- 22.5K sqm of lab capacity
- 700+ customers worldwide
- 6 centers of excellence