Final Report

Study Title:

Genetic stability 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 Targeted Locus Amplification (TLA) methodology followed by Next-Generation Sequencing (NGS) and data mapping in combination with droplet digital PCR (ddPCR) copy number determination, two samples of XXXX cell line containing the XXXXX sequence were analyzed to determine genetic stability. Evaluation was based on comparison of the results on transgene and vector integrity, identification of integration site and copy number of samples for two different culture stages of the cell lines.

The studied samples showed one (1) integration site for vector X, namely on chromosome 6. Data mapping revealed one (1) sequence variant and four (4) structural variants in the introduced vector sequence. The copy number showed eight (8) partial vector copies. Based on the obtained data it is concluded that the studied cell bank is genetically stable with respect to the integrated sequences as well as the genomic regions of the integration sites and copy number.

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

2.3 Sample information

Sample Code on vial received	Cell line	Sample Name	Passage information	Number of vials received	Content of each vial	Abbreviation used in report
Company 1	X	XXX MCB		5	10 million x cells	MCB
Company 2	X	XXX EPC		5	10 million x cells	EPC

The Sponsor was responsible for characterization and identification of the samples and control.

Storage condition	The samples were received on dry ice and stored in an ultra low temperature freezer (temp -80°C) until sample preparation.
Provided vector sequence	XXXXX sequence provided by the Sponsor (see Appendix 1).

2.4 Study Personnel

Name Lab technician TLA	X
Name Lab technician TLA	X
Name Data Analysis TLA	X
Name Data Quality Control TLA	X
Name Data Analysis ddPCR	X
Name Data Quality Control ddPCR	X
Name Study Director	Cheryl Dambrot
Name Head of QA	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 TLA	04-Jan-2023
Analysis Completion date ddPCR	25-Jan-2023
Quality Control date	27-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 22-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
ddPCR	Droplet Digital PCR
DNA	DeoxyriboNucleic Acid
EPC	End of Production
FW	Forward
Hom	Homologous bases
Html	HyperText Markup Language
Ins	Inserted bases
MCB	Master Cell Bank
NGS	Next-generation Sequencing
PCR	Polymerase Chain Reaction
RV	Reverse
Set X	Primer set X
TLA	Targeted Locus Amplification

4 Purpose of the study

The study was performed to evaluate the genetic stability of 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 and ddPCR, transgene, integration site and copy number 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. The results of the samples of the MCB and EPC were compared to draw conclusions regarding the genetic stability of the XXX cell bank.

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 Solvias 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 ddPCR Technology

Droplet digital PCR (ddPCR) is a relatively new technology that enables the precise quantification of targeted nucleic acids in samples. It measures absolute quantities (total copies per reaction) by counting the nucleic acid molecules encapsulated in discrete, volumetrically defined, water-in-oil droplet partitions. By dividing the number of total copies (per reaction) of the target by the number of total copies of the genomic references and multiplying this by the assumed copies per cell of the genomic references, copies per cell can be calculated for each target. Therefore, ddPCR can be used for copy number variation analysis. The results with the samples are compared to each other using the standard deviation in the experiment as well as precision of the method. Technical details are provided in the Appendix 8.

5.3 Study outline

5.3.1 TLA and sequencing

The cell suspension was thawed to room temperature, cells were counted and viability was measured. Six (6) million viable cells were collected from the cryovial, cross-linked and then 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 1) 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				

5.3.2 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.

5.3.3 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 1.11) (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 Solvias 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

5.3.4 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 Solvias 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 Solvias 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 artifacts, 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 Solvias 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,

5.3.5 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 Solvias Manual, TLA Terminology & methods in Appendix 9.

5.3.6 Copy number determination

The ddPCR assay has been designed on Element Y present in the vector X (see Table 1 for primer and probe sequences and the Appendix 8 for further details). Additionally, two assays (set of primers and probe) on the endogenous targets A and B have been purchased to serve as genomic references. For an accurate copy number assessment, a restriction enzyme was used in the ddPCR reactions to fragment (potential) vector concatemers. Testing was performed on freshly extracted gDNA of the transgenic cell line samples with 5-Units of restriction enzyme HaeIII according to the manufacturer's instructions (QX200 Droplet Digital PCR System; Bio-Rad, Hercules, CA).

To determine the number of copies the following calculation was used:

'Copy number = (total copies of Element Y / total copies of genomic reference gene) *assumed number of genomic reference gene copies'

To determine the assumed number of genomic reference genes copies, first the following calculation was performed:

'The assumed number of genomic reference gene copies = total detected copies of one genomic reference gen / total detected copies of the other genomic reference gene'

If this ratio is within the range of 0.8-1.2, the assumption is made that the references are present in an equal number of copies. The actual assumed copy number of the references is thus based on the calculated ratio and publicly available information about the ploidy of the relevant cell type.

Table 2: Primer sets and probe sequences of the ddPCR assays

Target	Oligo type	Sequence
Element Y	FW primer	
	RV primer	
	Probe	
Endogenous target A	FW primer	
	RV primer	
	Probe	
Endogenous target B	FW primer	
	RV primer	
	Probe	

6 Results

6.1 TLA and sequencing

For each of the studied samples four 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 indicating that the vector and its integration locus have been amplified and sequenced.

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. In both samples coverage is observed across the complete vector sequence Vector: 1-xxx, demonstrating that the entire sequence is integrated in the genome of the samples.

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 5.3.3 revealed one (1) sequence variant in the integrated vector sequence for both samples (Appendix 4).

6.3.3 Structural variants in vector sequence

Comparison of the mapped reads with the provided vector sequence using the criteria in section 5.3.3 revealed four (4) structural variants present in the integrated vector sequence for both samples (see Appendix 5).

6.4 Integration sites

The presence of integration sites was evaluated using the criteria in section 5.3.5. Whole genome coverage plots were generated using data obtained with primer sets 1, 2, 3 and 4. In both samples the plots showed one (1) integration site, namely on chromosome 6 (see Appendix 6). Further analysis into the integration site locus, show the same pattern in both sample with a vector integration site in at chr6:69,721,102-69,721,205 as well as a genomic deletion in the genomic region of the integration site. The 102 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 vector copy number was determined using ddPCR. The ratio of the total number of copies for the genomic reference genes A and B was determined to be within the range specified described in section 5.3.6 for both samples. This implies that both genes are present in equal copy numbers in this cell line (Appendix 8, Figure 1). With the understanding that X cells are diploid for most of their genome, it is assumed that two copies are present of each genomic reference gene. Element Y has a higher number of total copies than the number of total copies for the genomic reference genes for both samples (Appendix 8, Figure 1). The average copies/cell of the Element Y was 8 copies/cell for the MCB sample and 8 copies/cell for the EPC sample (Appendix 8, Figure 2). The copy number values of the element Y for samples MCB and EPC are within previously determined intermediate precision of the assay (3-5%) and thus considered the same with both reference genes.

Table 3: Copy number of Element Y using Target A and *B* as reference

	Copies/cell (Mean±SD)	
	Element Y	
	Targe A	**Target B**
Sample MCB	8.03 ± 0.09	7.95 ± 0.08
Sample EPC	7.97 ± 0.03	8.03 ± 0.06

Data represent the mean ± SD of one experiment performed in triplicate.

7 Conclusion

Both samples from the cell line X showed one (1) integration site. One (1) sequence variant and four (4) structural variants in the integrated vector sequence were found. For all samples the copy number was determined to be eight (8) partial copies. Based on the obtained data it is concluded that the studied cell bank is genetically stable with respect to the integrated sequences as well as the genomic regions of the integration sites and copy number.

8 Approval

The following analytical services performed in this study are in scope of accreditation for ISO/IEC 17025:2017, accredited by the Dutch Accreditation Council RvA, Registration number L671. The 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 is not included in the scope of this accreditation.

Scientific approval	Cheryl Dambrot – Study Director
Date	
Signature	

QA approval	Maaike van der Weij – Head of QA
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. Solvias Manual, TLA Terminology & methods

Appendix 1

Vector sequence and annotation

Vector Name:

XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXX

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 each sample 4 datasets are generated, one for each primer set (set 1, set 2, set 3 and set4, see Table 2: Primers used in TLA analysis, in section 5.3 of the main report). 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 MCB set 1	1,577,910	149	66	75	85.16	32.82
sample MCB set 2	1,325,929	149	62	80	82.24	33.15
Sample MCB set 3	1,277,910	149	58	76	83.12	30.24
Sample MCB set 4	1,115,829	149	63	77	84.56	31.45
Sample EPC set 1	1,565,910	149	67	77	85.16	32.82
sample EPC set 2	1,454,929	149	63	79	82.24	33.15
Sample EPC set 3	1,324,910	149	60	74	83.12	30.24
Sample EPC set 4	1,200,829	149	62	78	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.

NGS read coverage across the vector sequence

In Figures 1 and 2 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 MCB and EPC. 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 and 2, demonstrating that the entire sequence is integrated in the genome.

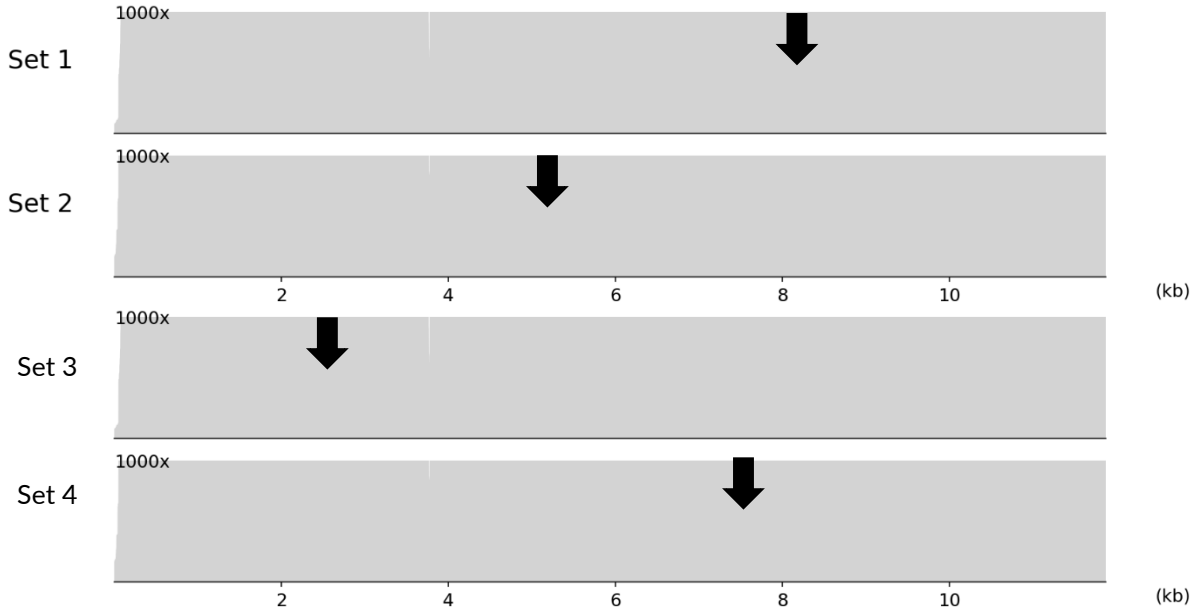


Figure 1: sample MCB, Y-axes are limited to 1,000x read coverage.

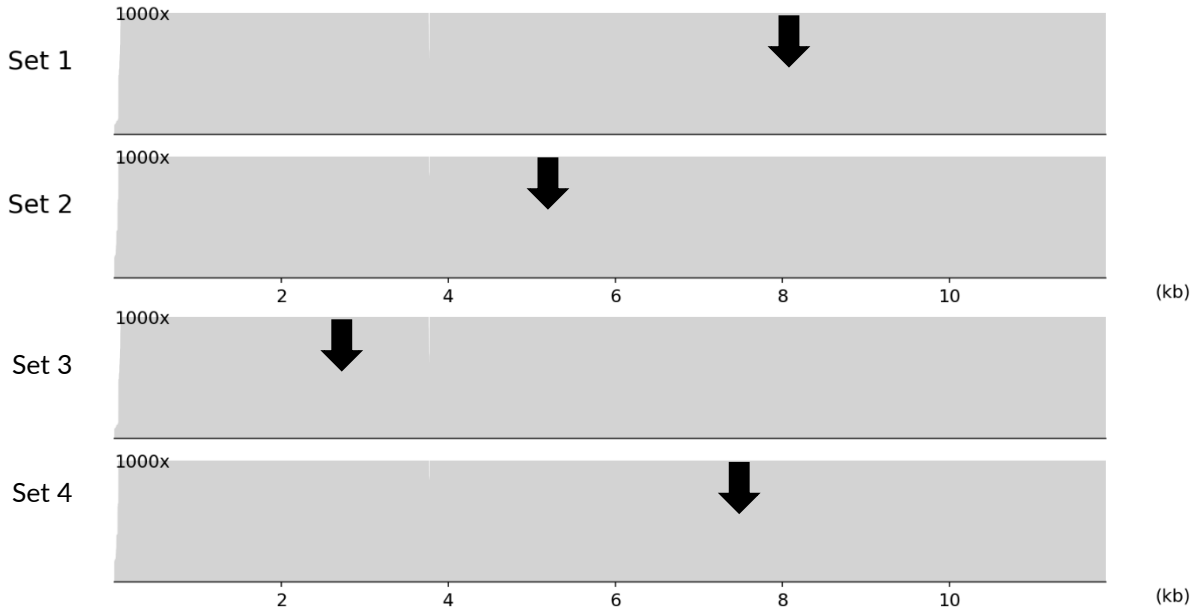


Figure 2: sample EPC, Y-axes are limited to 1,000x read coverage.

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).

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

Sample EPC											
1. Region	2. Pos	3. Ref	4. Mut	5. Cov set 1	6. % set 1	7. Cov set 2	8. % set 2	9. Cov set 3	10. % set 3	11. Cov set 4	12. % set 4
NeoR	4,990	C	A	2,458	23	22,590	30	1,568	22	4,580	25

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 MCB and EPC 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 (next page). 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.

Sample MCB

							# 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	-	63	61	64	62	2	0	1	0	2	0	1	0
2	→	500	2,753	←	tail to tail	3	-	230	43	232	41	1	13	0	3	1	16	0	3
3	→	3,500	6,178	→	tail to head	2	-	40	123	78	120	2	1	3	2	2	1	3	2
4	→	4,600	9,806	→	tail to head	2	-	65	415	43	425	1	1	3	3	1	1	3	3

Sample ECP

							# 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	-	63	61	64	62	2	0	1	0	2	0	1	0
2	→	500	2,753	←	tail to tail	3	-	230	43	232	41	1	13	0	3	1	16	0	3
3	→	3,500	6,178	→	tail to head	2	-	40	123	78	120	2	1	3	2	2	1	3	2
4	→	4,600	9,806	→	tail to head	2	-	65	415	43	425	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
TGCATCGTACGTTGGCCAATCGTCGTCTAGCTGTGTCT**GCTTG***TCGGCTAGTCCGATGGCAC*
CGTGCGTCAGGGTCCAAGGTTCA

2) vector:500 (tail) fused to vector:2,753 (tail) with 3 homologous bases
GTGTGTGTGCGACACGTCGTACAGTCAGTGCAGTGCAGTTTGTCCAAGGTTCCGGGCA**TAT***G*
CCCGGTTGGTTGAAGAGTCAGTGCAGGTGTGCAGTGTGTCAGTGCGTGATGAAACCCTGGGC
CC

3) vector:3,500 (tail) fused to vector:6,178 (head) with 2 homologous bases
CATATGTGTACACACGTGTGTCAGTGCCAAATTGGGCATGCAGTGCGTGTGTCATA**TT***AGGGT*
GTGTGTAAAGTGTGCCCCGTGTGCAGTGCGTACGTGCAGTGTGTCAAACCCGGTTGGT

4) vector:4,600 (tail) fused to vector:9,806 (head) with 2 homologous bases
TGTGAAACGTGTGCCCAAGTGTCCAGTGCCAGTGCATTGGTGTGTGCCGTGAAAG**GA***GTTTG*
TGTAAAGGTGTCCGTGGTACGTGCCGTATGCAGTGCAGTGCATGTGAGTGTCAGT

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 boxed or inserted (novel) as stated above each sequence. Inserted (novel) bases were not found.

Appendix 6

Whole genome wide coverage plots

Figures 1 till 8 show the NGS read coverage across the genome sequence for sample MCB (figure 1-4) and sample EPC (figure 5-8) 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 site is encircled in green.

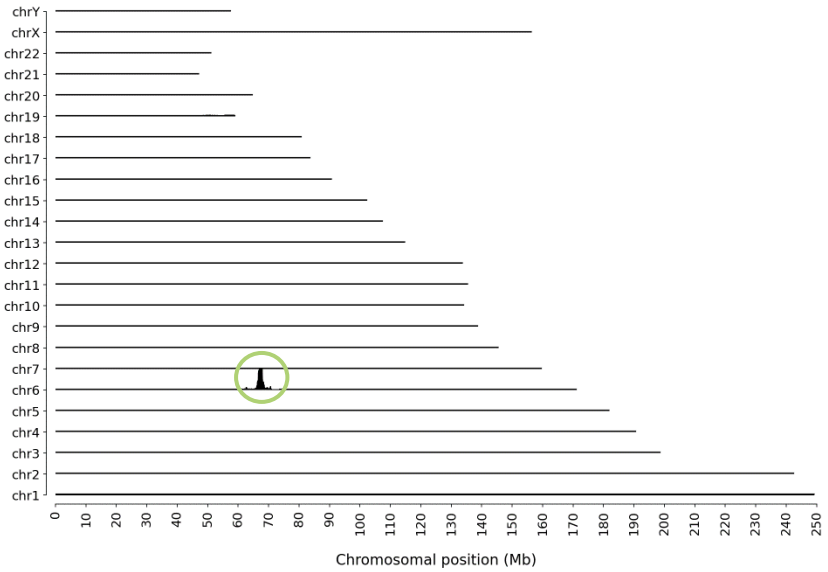


Figure 1: TLA sequence coverage across the xxxx genome using primer set 1 in sample MCB.

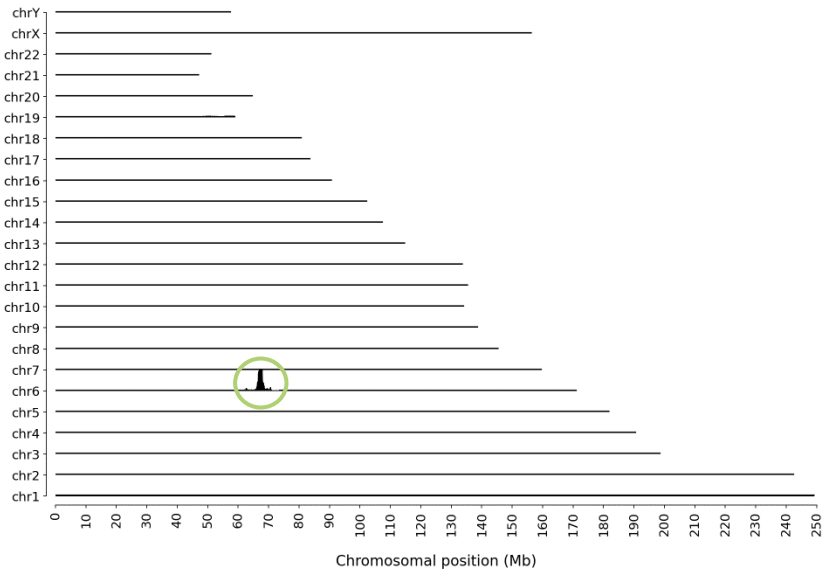


Figure 2: TLA sequence coverage across the xxxx genome using primer set 2 in sample MCB.

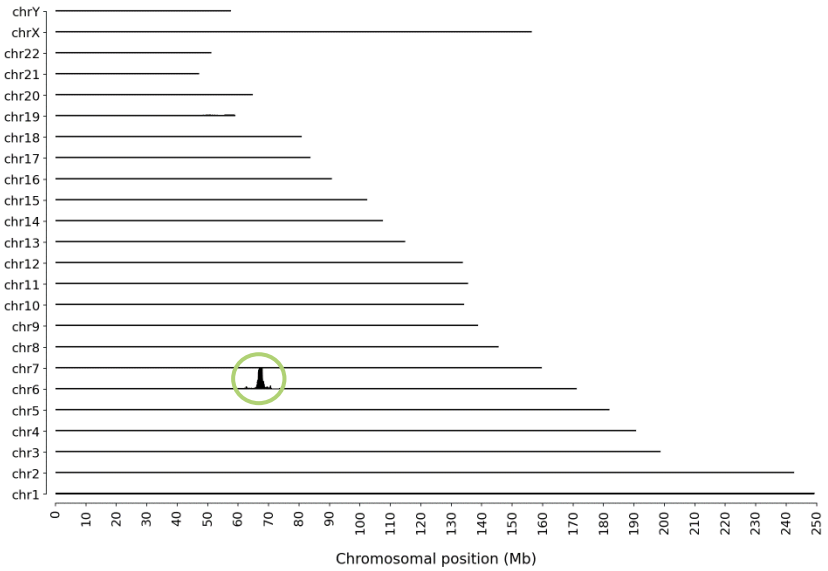


Figure 3: TLA sequence coverage across the xxx genome using primer set 3 in sample MCB.

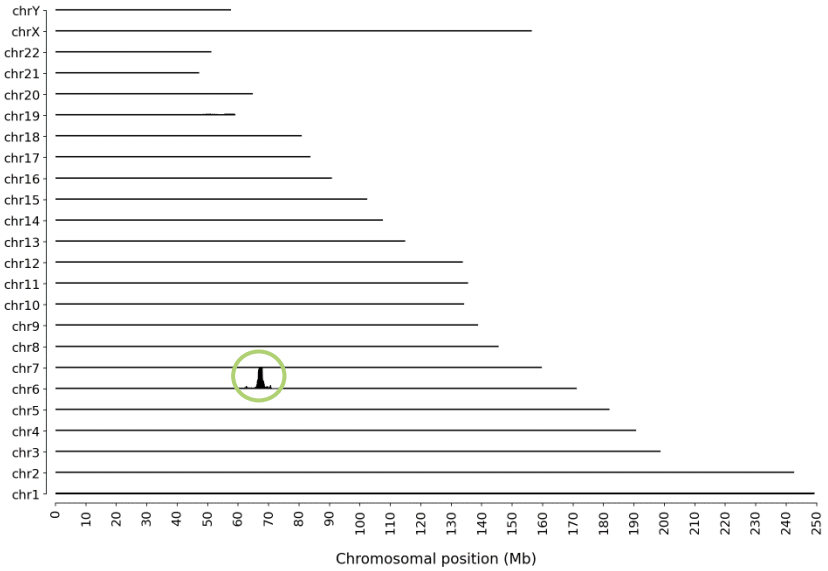


Figure 4: TLA sequence coverage across the xxxx genome using primer set 4 in sample MCB.

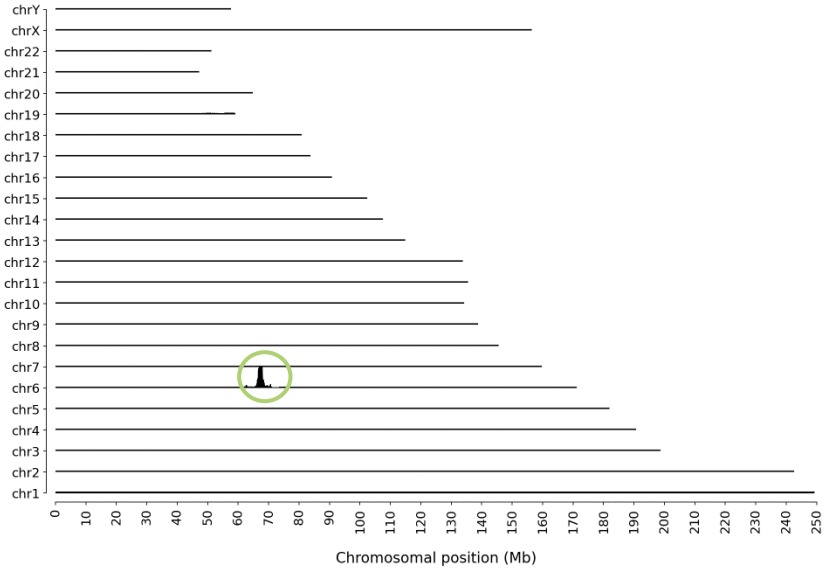


Figure 5: TLA sequence coverage across the xxxx genome using primer set 1 in sample EPC.

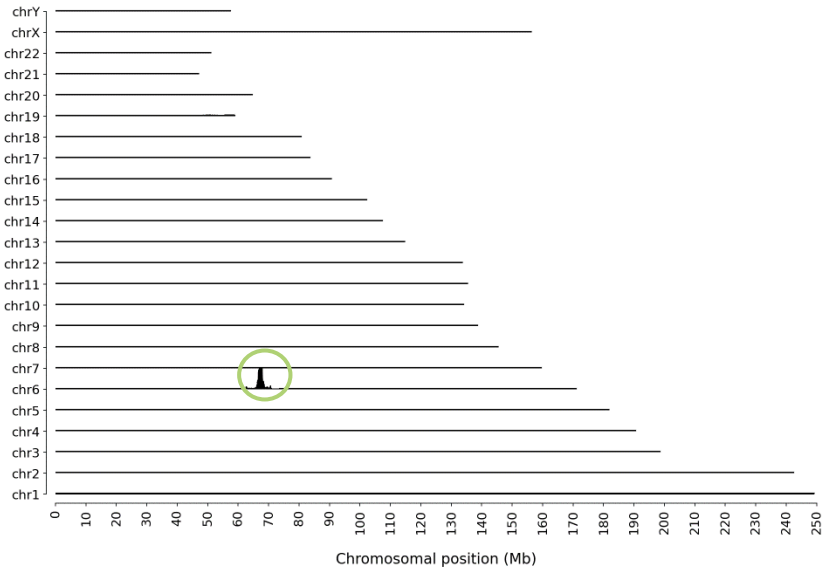


Figure 6: TLA sequence coverage across the xxxx genome using primer set 2 in sample EPC.

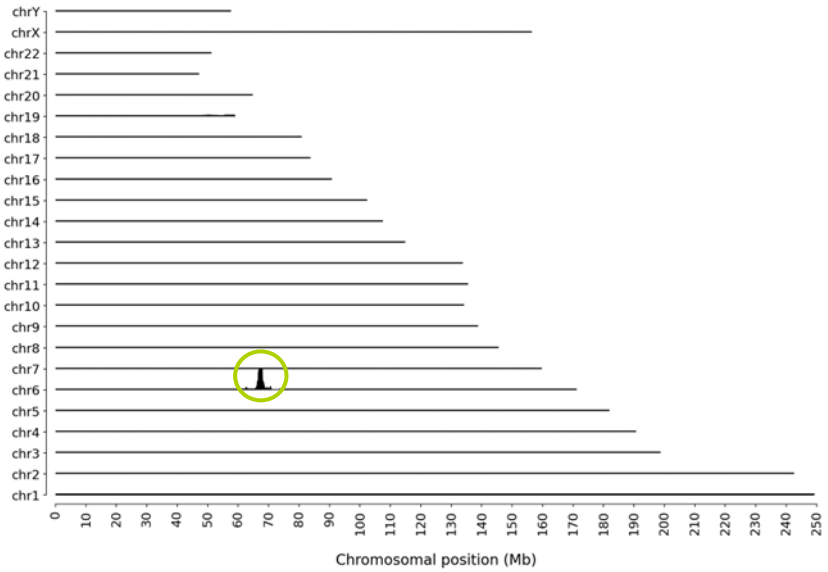


Figure 7: TLA sequence coverage across the xxx genome using primer set 3 in sample EPC.

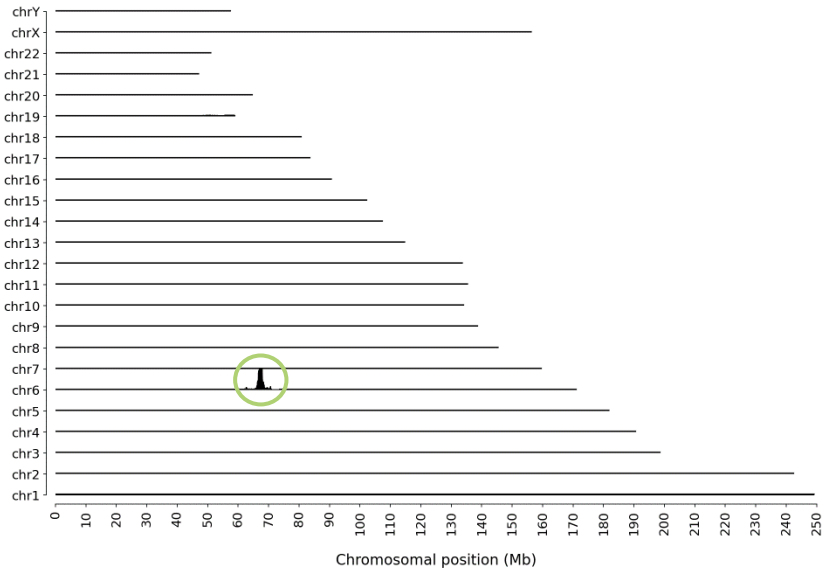


Figure 8: TLA sequence coverage across the xxxx genome using primer set 4 in sample EPC.

Appendix 7

Integration site information

In Figure 1 and 2 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 MCB and EPC respectively, on the specified region in the XXX 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)

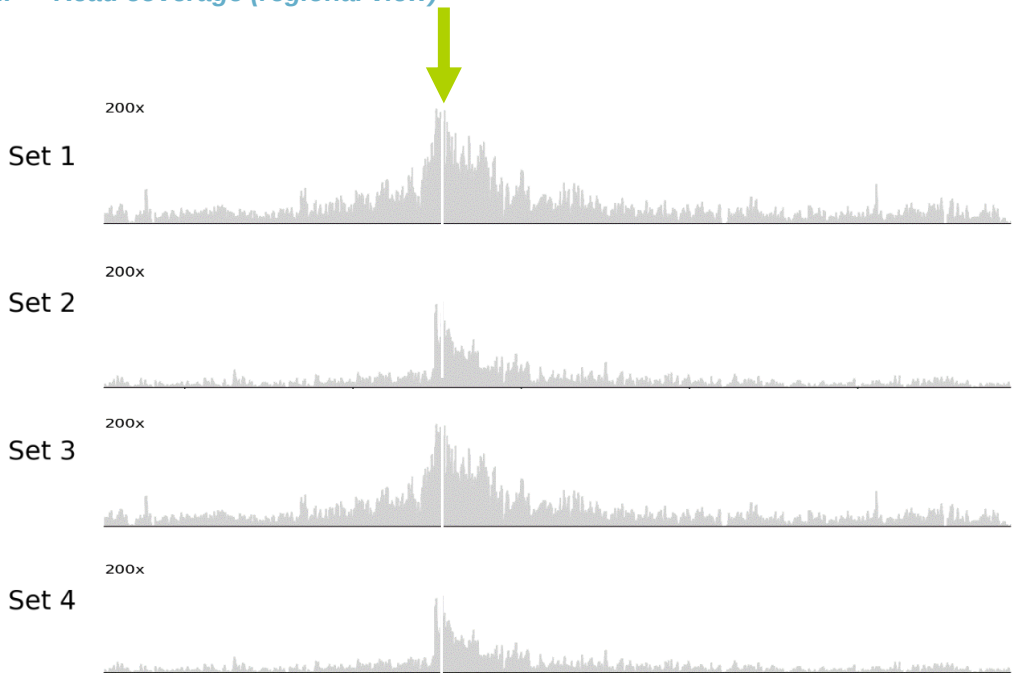


Figure 1: TLA sequence coverage (in grey) across the vector integration locus, chr6:69,561,025-69,881,284of the sample MCB. The green arrow indicates the location of the breakpoint sequences (1 and 2). Y-axes are limited to 200x.

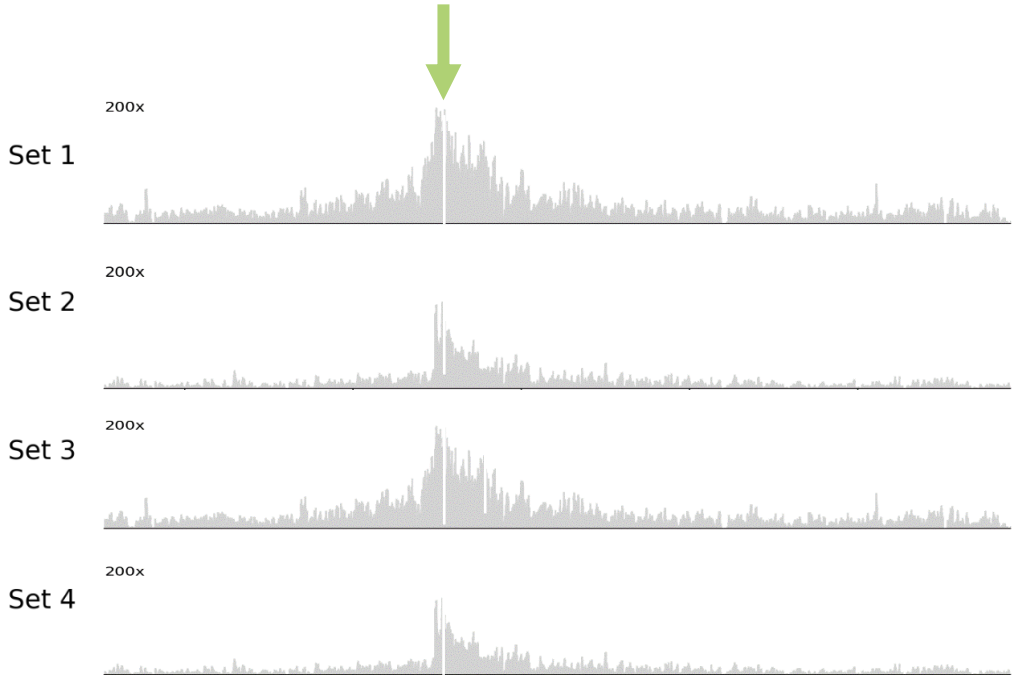


Figure 2: TLA sequence coverage (in grey) across the vector integration locus, chr6:69,561,025-69,881,284 of the sample EPC. The green arrows indicate the location of the breakpoint sequences (1 and 2). Y-axes are limited to 200x.

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 and 2.

1. chr6:69,721,102 (tail) fused to Vector:545 (tail) with 4 homologous bases
AATGCTCTGGAATCCTAGGTAAACTCAAAAGGCAGTCTAGGAAACAAGGACTGCAATTCCTAGGC
*AACTCCTA*GTGCTTCTGAGGTGCCCAATTTGTGACACGTGACGTCAGTGTGAAACCCAACACAC
GTGTGTTGTTGTAAAGTGTGGCGTATGTGCAGCCCC

2. Vector:6,408 (head) fused to chr6:69,721,205 (head) with 3 inserted bases
GTCGTGATGTGTGTGTAAACCGGTTCCAATTGGTTCCAATTGTGTGAATTTGCAGTGAA*CTAGGC*
TGGGCTTACAGTGGACTAGGGTGGCATGTGACCCAGGGAGACAACAGCTAAGGGAGTGCTTGC
ACCATTCCTCTCCCAACCCCATGAAGTGCAGCTCACCTCAACAAAGGTGACTCTTTCCTTTGGCC
TGAGGAAAGG

The breakpoint sequences and the coverage profiles in figures 1 and 2 show that a genomic deletion has occurred in the region of the integration site. The 102 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 identified breakpoints are located in intron 7 of *LMBRD1*.

Appendix 8

Copy number details using ddPCR

In Figures 1 and 2, the results of the copy number determination for MCB and EPC are presented. Based on the total amount of copies of the reference genes, it is assumed that each cell contains two copies of target A and B reference genes, with the assumption that the cell line is near diploid for most of the genome. The copy number of Element Y was determined using target A and B as the genomic reference genes. For both samples, the total copies for Element Y were higher than the total copies for the genomic references target A and B (Figure 1). An average of 8 copies per cell of the MCB, EPC sample were found for Element Y (Figure 2). The copy number values of element Y for both samples are within previously determined intermediate precision of the assay (3-5%) and thus considered the same with both reference genes.

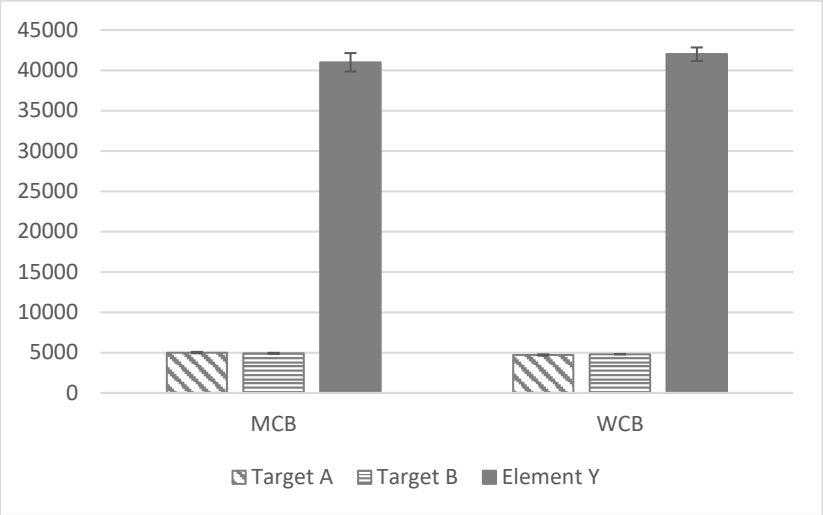


Figure 1: Total amount of copies of the genomic references A and B, and Element Y for samples MCB and EPC. The data represents the mean ± SD of the experiment performed in triplicate.

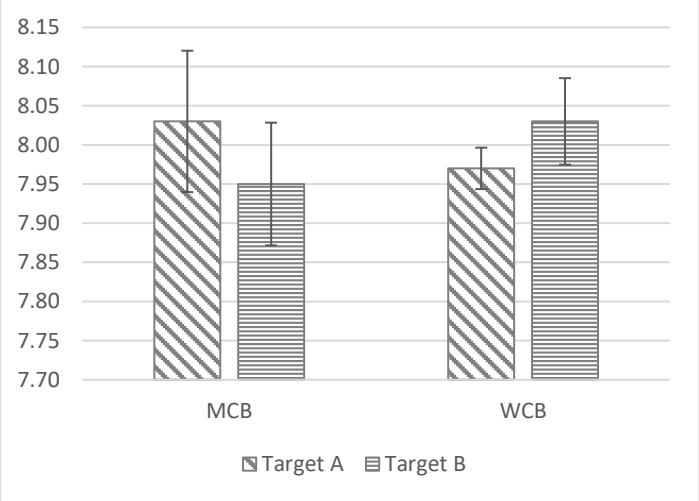


Figure 2: The copies per cell of the GOI as compared to the genomic genes Target A and B, for samples MCB, EPC. The data represents the mean ± SD of the experiment performed in triplicate.

Appendix 9

Solvias Manual, TLA Terminology & methods

Manual

Introduction to the terminology and methods used in transgene & integration site TLA analyses & ddPCR

solvias

Table of Contents

TLA and sequencing

The samples submitted for TLA analysis are used and processed according to Solvias' Targeted Locus Amplification (TLA) protocol, developed by Cergentis (de Vree *et al*. Nature Biotechnology 2014). A summary of the TLA technology is shown in **Figure 1**.

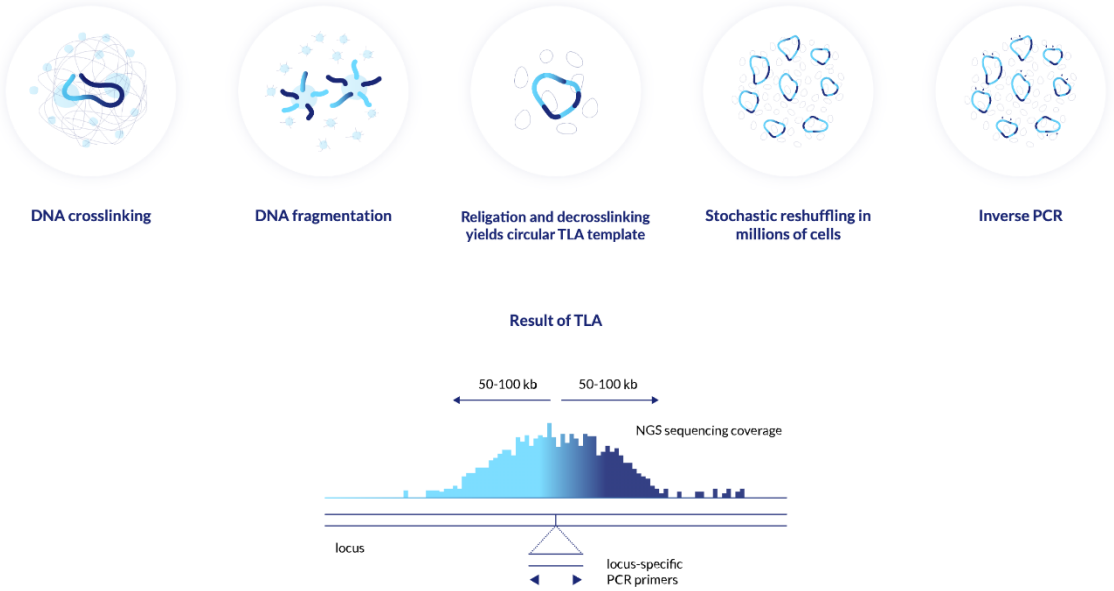


Figure 1: Schematic representation of TLA technology.

In brief, the TLA technology uses the physical proximity of sequences within a locus as the bases of selection. First, DNA is crosslinked, fragmented and re-ligated to generate the circular fragments of DNA. Then, using one TLA primer pair complementary to a sequence within a locus / vector containing a transgene, sequence information is generated across the entire vector and its integration site(s) (**Figure 2**). Solvias recommends the use of two vector-specific primer pairs complementary to two different sequences within a vector. The primer sets are used in individual TLA amplifications and subsequent analysis. This results in two independent data sets.

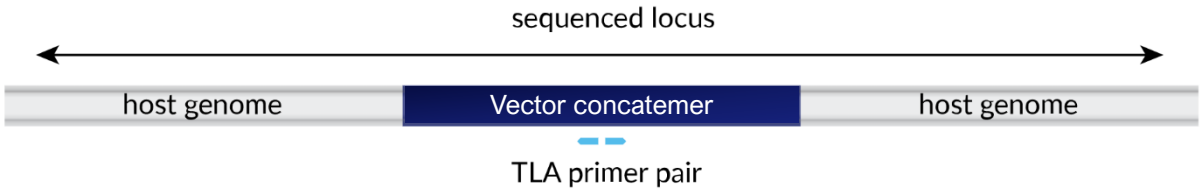


Figure 2: TLA-based transgene and integration site sequencing using transgene specific primer pairs.

PCR products are purified, library prepped using the Nextera® DNA Flex Library Prep protocol (Illumina), and pooled. The resulting pool contains unique barcodes (Nextera® DNA CD Indexes, Illumina) for each PCR product. NGS sequencing (paired-end, 2X149 bases) is performed on an Illumina System. For a standard analysis, ~1 million reads are generated per PCR product.

Alignment and processing of sequencing reads

The Illumina® System, together with the bcl2fastq Conversion Software (Illumina), performs base calling and demultiplexing (converting the base call information into the read information). Using the barcode information, paired-end FASTQ files are generated for each individual amplification of a TLA sample. Reads are mapped to the vector sequence and host genome using BWA-MEM, version 0.7.15-r1140, settings bwa mem -M -t 4 -B 7 -w 33 -O 5 -E 2 -T 33 -Y (Li H., 2013, arXiv:1303.3997).

Vector sequence coverage

Definition of vector and transgene coverage

Vector refers to the entire sequence that was used to modify the cells, so including a transgene/region of interest and backbone. Solvias recommends sending the full vector reference sequence in order to optimally assess its integration and the presence of sequence and structural variants.

Coverage is defined as the number of NGS reads that cover a base (in a vector sequence or genomic locus). The coverage determines the sensitivity with which sequence and/or structural variants can be detected.

Partial/complete vector integrations

A gap in vector coverage indicates partial integrations of the vector caused by (un)intended deletion(s) in the original vector or introduced during integration (**Figure 3**, left panel).

In case homology arms/ITRs/LTRs were used for (targeted) integrations, TLA data will, depending on how clean such integrations have been, show the presence or absence of coverage on the backbone sequences (**Figure 3**, right panel).

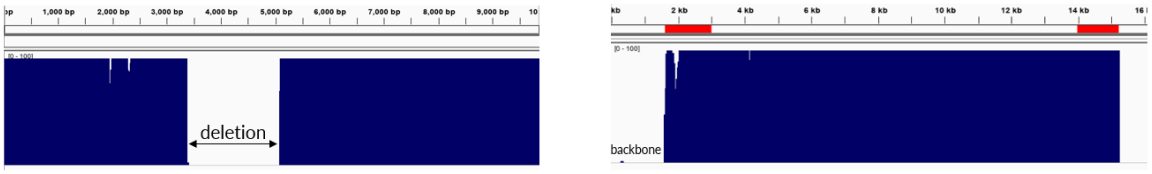


Figure 3: NGS sequencing coverage across the vector in the random (left) and targeted integration (right, homology arms are shown in red).

In samples with virus-mediated gene transfer, the obtained complete vector coverage might be a result of amplification of the episomal vector DNA remaining in the cells. In this case, the identification of the numerous integration sites in heterogeneous samples might be not possible due to loss of sequencing capacity on the non-integrated sequences. The experimental set-up can therefore be adjusted to eliminate episomal DNA before integration site enrichment.

Multiple vectors and co-integrations

Depending on the nature of multiple integrated vector sequences, TLA primers can be used on sequences common to all integrated vectors, or primers can be designed that are specific for each individual vector. The specific TLA primer sets allow identification of vector-specific integration sites or demonstrate co-integration of individual vectors.

TLA analyses can be used to sequence and identify (un)expected co-integrations of (partial) vector sequences, unknown vector sequences and/or of other DNA sequences (e.g. *E. coli*). Detailed analyses of unexpected sequences can be performed by mapping generated NGS data on appropriate reference (genome) sequences.

Large vectors

In the analysis of large vectors (>50 kb), additional TLA primer pairs can be added to ensure sufficient sequencing coverage is generated on the vector and its integration site(s).

Variations in vector sequence

TLA allows identification of sequence variants (single nucleotide variants, insertions and deletions of one or several nucleotides), structural variants within the integrated vector, and its concatemerization. Heterogeneity of the sample, as well as copy number of the integrated vector and sequencing depth will affect the sensitivity of calling sequence and structural variants.

Sequence variants

The presence of sequence variants is determined using samtools mpileup (samtools version 1.11) (Li *et al.* Bioinformatics, Jun 2009 [PMID: 19505943], Li H. Bioinformatics, Nov 2011 [PMID: 21903627]). Only read-bases with a minimal Q-score of 20 (Base call accuracy of >99%) are used for the detection.

Sequence variants are reported that meet the following criteria:

- Allele frequency (relative amount of reads with the variant, compared to total coverage on the variant position) of at least 5% for cell lines/cells in culture and 20% for primary cells;
- The variant is present in the data generated with at least 2 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 (variants present in >95% reads of the same orientation in the data of at least 1 primer set with ≥30x mutated bases are filtered out);
- The variant is not located in a repetitive sequence: variants surrounded by ≥4 bases of the mutated base are removed unless they show >50% frequency in at least 1 primer sets; InDels followed by ≥6 of the same bases/set of bases are removed unless they show >50% frequency in at least 1 primer sets;
- Low frequency variants (between 5-20% mutant allele frequency) are not found with similar frequencies in an independent control (Solvias recommends the use of an independent control for more reliable filtering).

Sequence variants are reported in a table which includes the following columns:

- Region - annotated region in the reference sequence that was used for mapping;
- Position - position within the reference sequence;
- Reference - nucleotide present in the reference sequence at this position;
- Mutation - observed mutation, “+” indicates an insertion, “-” indicates a deletion right downstream of the reference nucleotide;
- Coverage - total number of reads passing Q20 that map to the indicated position;
- Percentage (%) - percentage of reads containing the mutation (=mut/cov*100%).

Please note that using these filtering criteria, the reported 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.

In very heterogeneous samples (non-clonal cell population with viral/transposon-mediated integrations), detection of variants in individual integration sites is not possible. The identified variants are then categorized as follows:

- A. variants that occur in all samples with >80% mutant allele frequency represent general deviations that were present in the supplied reference sequence of the vector before its introduction in the cells.
- B. variants that are found in all samples with 5 - 80% mutant allele frequency can indicate heterogeneity in the sequence of the virus that was used. Variants in this category that have low allele frequencies (<20%) can also represent systematic sequencing errors. These errors can be filtered out by including an independent control for the analysis or by performing independent validation experiments.
- C. sample specific variants found in XX, but not in YY or ZZ with 5 - 100% mutant allele frequency represent specific mutations that occurred in this sample in 5 -100% of the integrated vector.

Vector concatemerization and structural variants

Structural variants within the integrated vector sequence are identified by detecting vector-vector breakpoint sequences.

Terminology

A breakpoint sequence or fusion sequence is a sequence containing the breakpoint of and fusion between, two sequences originating from different origins compared to the reference sequence (**Figure 4**). This can be from two different genes/genomic regions, two vector regions where structural variation takes place or a vector sequence that fuses to a genomic sequence. In case of integration sites, it can also happen when deletions, insertions and inversions take place around the integration site. The reads that are indicative of a fusion sequence are called fusion reads, split reads, breakpoint reads or chimeric reads. In TLA procedure, DNA digestion and ligation results in the procedure-induced fusions of the sequences that are in proximity but are of different origin. These TLA induced fusions are not analyzed.

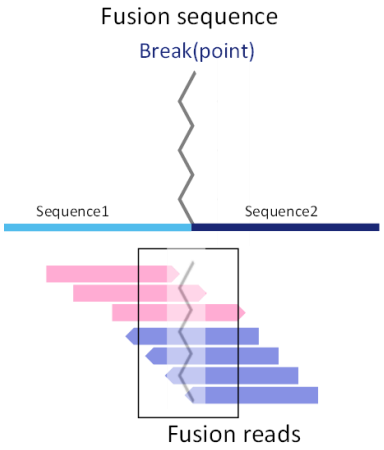


Figure 4: Schematic overview of a fusion or breakpoint sequence. Light blue and the dark blue sequence (on top) fuse together creating a fusion sequence. The point where the sequences fuse together is called the breakpoint. Based on the alignments of fusion reads (pink reads in forward orientation, pink reads in reverse orientation from the sequencer) you can reconstruct the fusion sequence and determine where sequence 1

There are three different breakpoint sequences derived from fusion reads that are reported. The first one (**Figure 5**, number 1) is when there is a perfect junction between sequence 1 (light blue) and sequence 2 (dark blue) and the exact breakpoint positions where the two sequences fuse together is known.

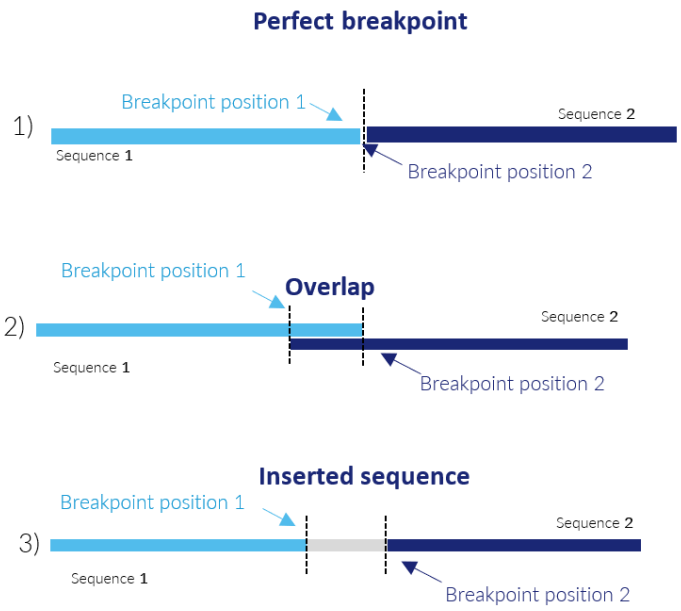
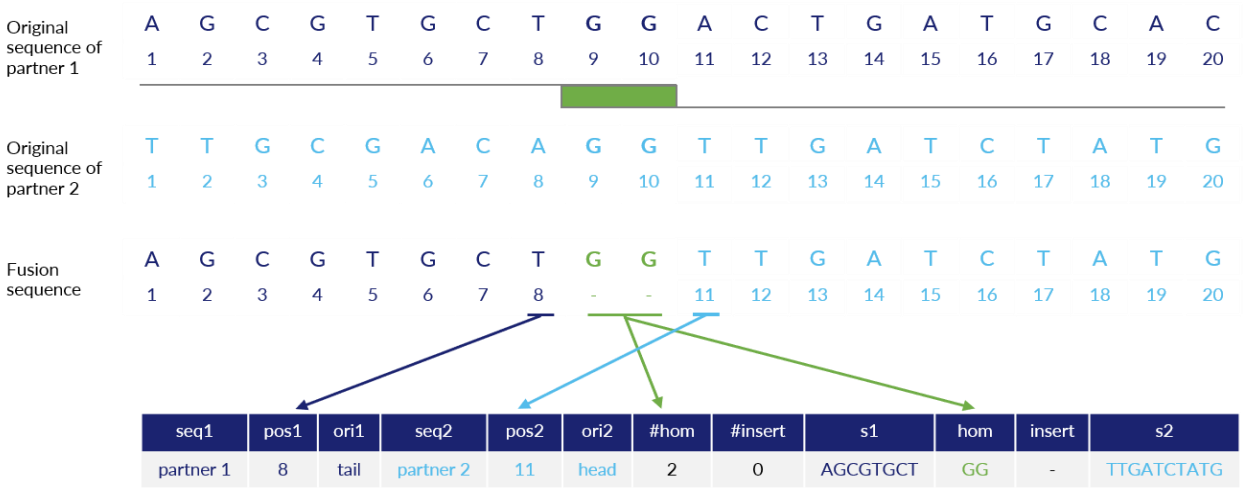


Figure 5: Schematic examples of different fusions. 1 - perfect breakpoint, 2 - fusion with homologous or overlapping bases, 3 - fusion with novel or inserted bases.

The second type of breakpoint sequences are the fusion sequences where there is homology and thus an overlap between sequence 1 and sequence 2 in the alignments of the fusion reads (**Figure 5**, number 2, **Figure 6A**). Bases from sequence 1 also align to sequence 2. With the current data set is not known if the overlapping bases belong to sequence 1 or sequence 2. Thus the overlapping based are removed from the reported breakpoint positions and reported as homology between sequence 1 and sequence 2.

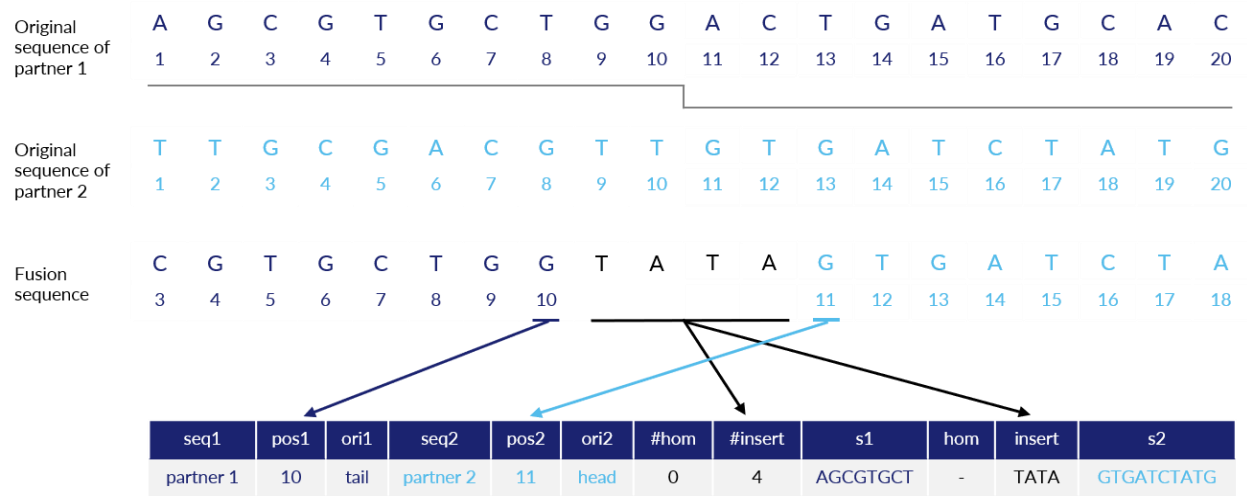
The third type of breakpoint sequences are the fusion sequences that have an inserted sequence between sequence 1 and sequence 2 (**Figure 5**, number 3, **Figure 6B**). The length of the inserted sequence can be varying in size (just a few bases or even50 bases or more). The inserted sequence can be an unaligned sequence from unknown source / origin or align (partially) to the vector or other known sequence (genomic).

A. *Example of a breakpoint/fusion sequence with homology*



seq1	pos1	ori1	seq2	pos2	ori2	#hom	#insert	s1	hom	insert	s2
partner 1	8	tail	partner 2	11	head	2	0	AGCGTGCT	GG	-	TTGATCTATG

B. *Example of a breakpoint/fusion sequence with an insert*



seq1	pos1	ori1	seq2	pos2	ori2	#hom	#insert	s1	hom	insert	s2
partner 1	10	tail	partner 2	11	head	0	4	AGCGTGCT	-	TATA	GTGATCTATG

Figure 6: Explanation of the terms 'homology' (A) and 'insert' (B), used to describe details of the sequences found at vector-vector and vector-genome breakpoints.

Detection of a vector-vector breakpoint sequence

Breakpoint sequences consisting of two parts of the vector, are identified using a proprietary Solvias script. Breakpoint sequences 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 one of the positions of the fusion;
- The breakpoint sequence is observed in data of at least two primer sets, unless the data provides a clear explanation why it is not found in one of the data sets;
- The breakpoint sequence is not present in independent control sample(s) (if included);
- Visual inspection of the breakpoint sequence in a NGS data browser is performed to remove those that are sequencing artifacts, e.g. breakpoints found at hairpin structures or low-complexity regions.

In heterogeneous samples, the detection of structural variants within individual integration sites is not possible. Only very abundant events like loss of the specific (backbone) sequences in the majority of integrations is identified.

Details regarding vector-vector breakpoint sequences

Vector-vector breakpoints and integration site breakpoints (see below) are described using the following terms:

- The number of homologous, common, or inserted, novel, bases at the junction site (for details see **Figure 6**).
- “head” and “tail” orientations of the vector/genome fragments at the junction site (for details see **Figure 7**).

Please note that based on TLA data is it not possible to reconstruct the order and size of individual copies in the integrated concatemer sequence.

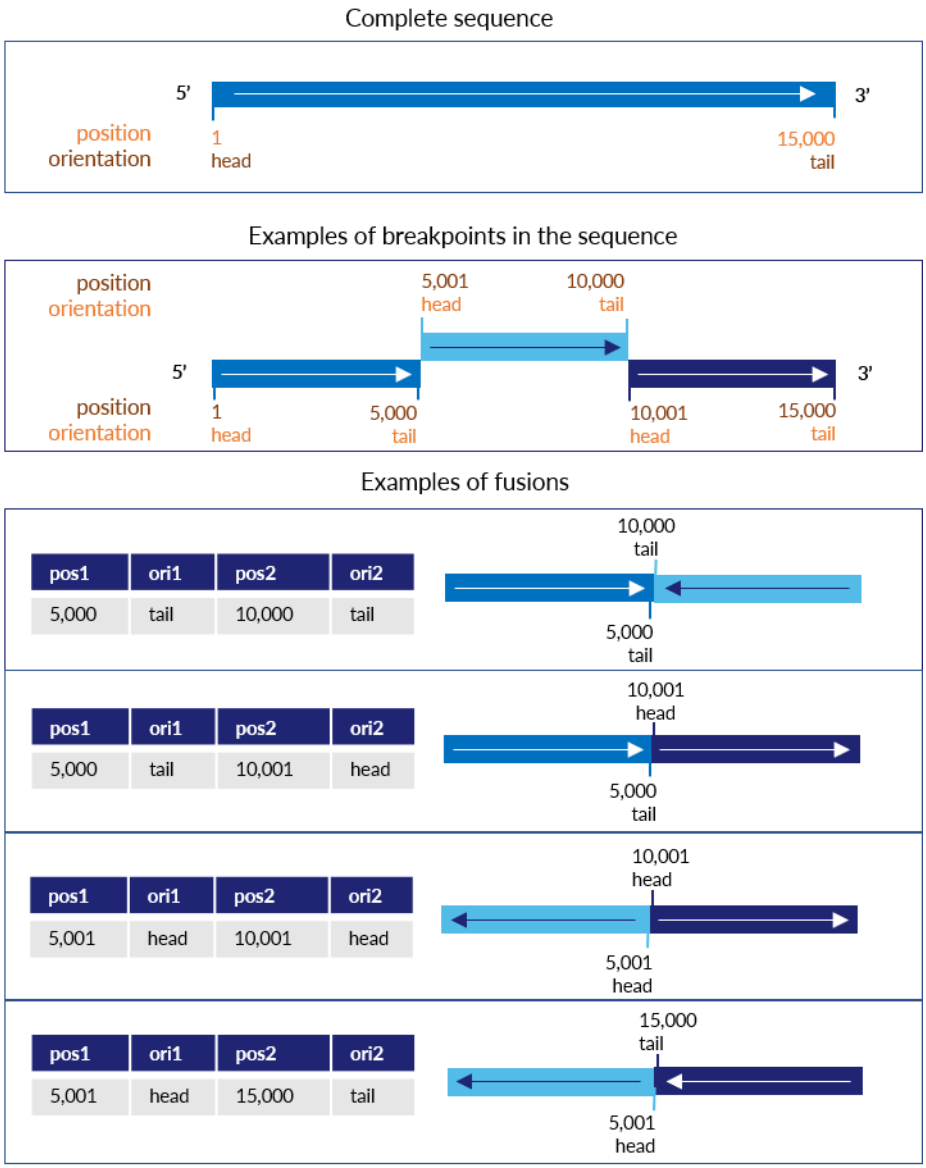


Figure 7: Schematic representation of the terms 'head' and 'tail', used to describe the orientation of sequences found at vector-vector and vector-genome breakpoints.

Integration site detection

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

Coverage peak(s) in the genome

TLA results in high coverage across the genomic positions of vector integration sites. Integration sites are therefore clearly visible in whole genome coverage plots (**Figure 8**).

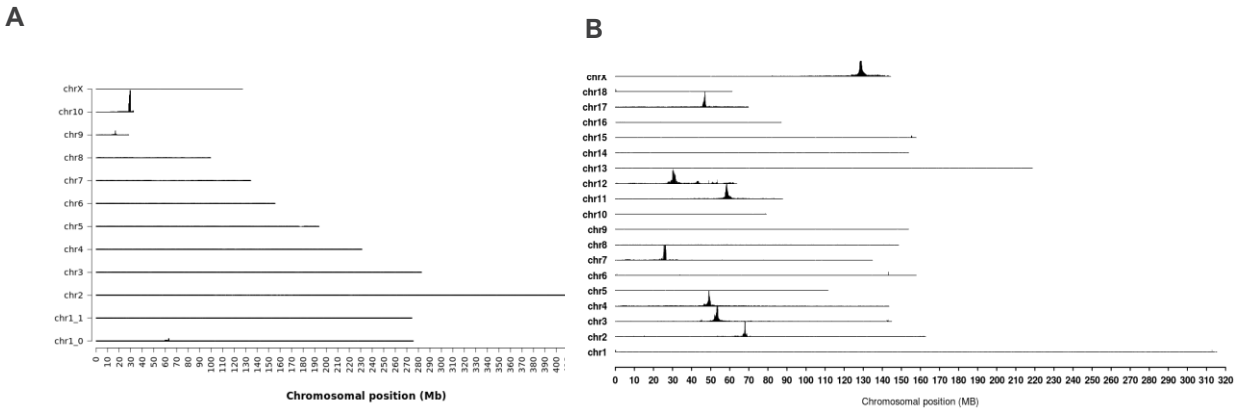


Figure 8: TLA sequence coverage across the Chinese Hamster Ovary (CHO) (A) and mouse (B) genomes. The chromosomes are indicated on the y-axis, the chromosomal position on the x-axis. High coverage peaks represent a single integration (A) and 8 integration sites (B) of a vector.

Partial vector integrations

Partial integrations can be identified if the integrated sequence contains a primer binding sequence.

Breakpoint sequences

A genome-vector breakpoint sequence will consist of a combination of a vector sequence and genomic sequence. Breakpoint sequences are reported in the same manner as vector-vector breakpoint sequences (**Figure 9**).

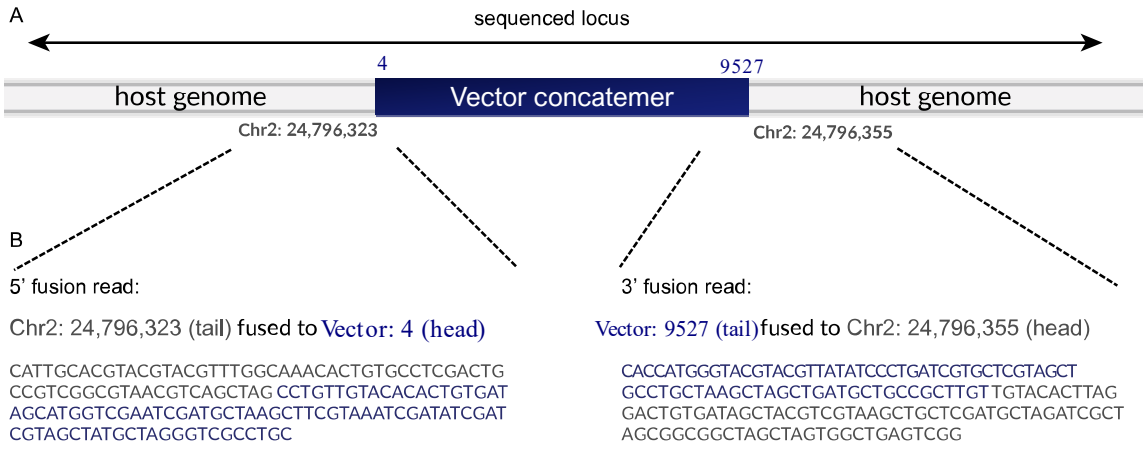


Figure 9: Vector-genome breakpoint sequences. A. Schematic depiction of a vector carrying a transgene and integrated into the host genome at the indicated positions on chromosome 2. B. Breakpoint sequences as depicted in a TLA report. For each breakpoint, the exact sequence as well as the relative orientation of the partner is provided.

Targeted sequencing of individual integration sites

TLA analysis with primers complementary to a wild-type sequence next to an individual integration site will provide sequence information across both transgenic and wild-type alleles (**Figure 10**). This enables determination of zygosity and characterization of the vector integrity (i.e. sequence variants and vector-vector fusions) in that locus.

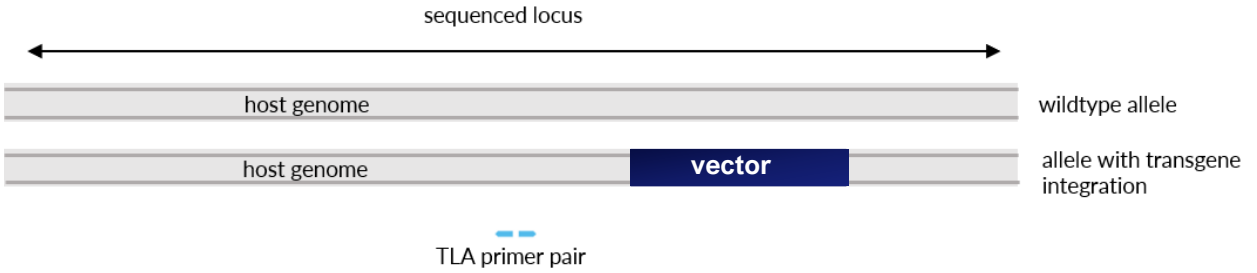


Figure 10: TLA-based analyses of individual integration sites. A TLA analysis with a primer pair in close proximity to the integration site provides sequence information across the entire locus on all alleles.

Integration sites in heterogeneous samples

Detection of the integration sites

In heterogeneous samples, the coverage per integration site that occurred in an individual cell or a small subset of the analysed cells is limited and no integration peak is seen on a genome-wide scale. Therefore, only the breakpoint reads containing vector-genome breakpoint sequences are used for the analyses.

When transduction is performed using a viral/transposon vector, the breakpoints are expected to occur at the boundaries of the LTR/ITR sequences. The breakpoints at these locations are selected and further filtered according to the following criteria:

- The reads in which the part that aligns to the genome is <20 bp are removed to ensure specificity.
- The reads with the wrong orientation at the expected location are filtered out
- The breakpoint sequences containing any inserted, novel, bases between the vector and the genome are filtered out (if the integration mechanism is not designed to generate/allow insertions at the integration sites).
- The duplicates due to homology of LTRs are removed.
- The breakpoint sequences present in the independent control (if included) are filtered out.
- For some breakpoint sequences, two or more hits are found on the genome and it cannot be determined which is the real integration site. In these cases, all breakpoints are reported, so that the total number of identified integrations is slightly over-estimated (typically 1-5%).

Interpretation of the data

After 5-6 million cells are crosslinked, fragmented and circularized, TLA amplification is performed on a subset of the total input, ~100,000 cells. Therefore, if a sample contains rare integration sites, only a subset of the integration sites is present in the amplification reaction. In addition, only a subset of the amplified TLA circles includes vector-genome fusion sequences and only a subset of amplified DNA sequences is sequenced. Therefore, TLA analyses of heterogeneous samples with unique integration sites that occur infrequently, results in identification of numerous integration sites, but each of them is found in the data of only one primer set and with the low number of sequencing reads. Additional amplifications and/or additional sequencing can be considered for a more complete list of the identified integrations. However, increasing the amount of input material per amplification is not required.

Identification of structural variants at integration sites

The presence of structural variants (genomic rearrangements) at the integration site is assessed by analysing the coverage profile and positions of vector – genome breakpoint sequences (**Figure 11**). If the genome sequence is intact, TLA will generate a bell-shaped coverage profile around the integration site. Deviations in this profile indicate genomic rearrangements with respect to the reference sequence.

If complex structural rearrangements are identified, paired-end sequencing analyses can be used to determine which breakpoints occur in physical proximity to each other and constitute the same integration site.

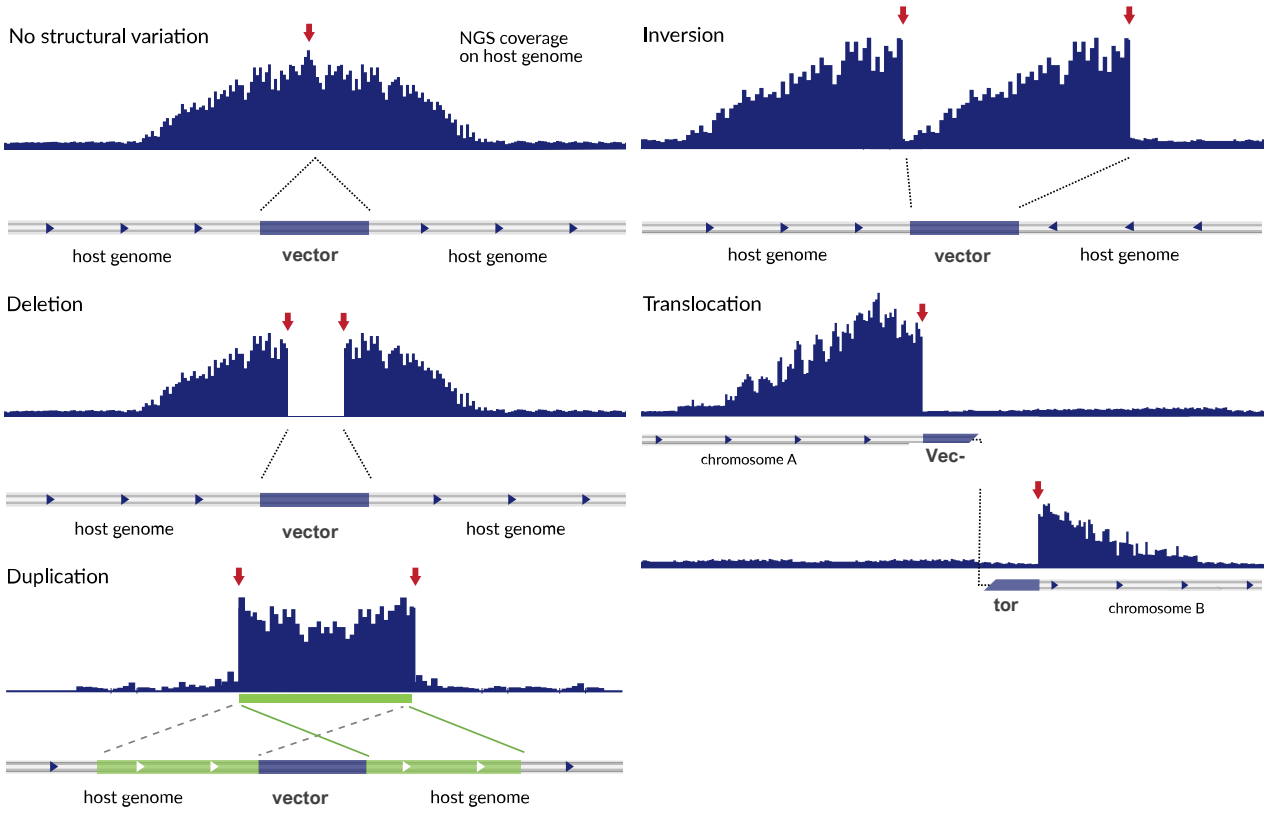


Figure 11: Schematic representation of the changes in the coverage profile and vector-genome breakpoint sequences (red arrows) that are observed in the most common types of rearrangements.

Paired-end information and paired-end analyses

TLA samples are sequenced in paired-end: generated TLA amplicons are sheared to ~600 bp fragments and then both the first 149 bp and the last 149 bp of these fragments are sequenced. The two sequences resulting from the same DNA fragment are called “mates”. Both sequences are used in the analyses, and the mating information is kept. The two mates within a pair will originate from the same physical DNA locus. TLA paired-end sequence information can thus be used to assess which vector and/or integration site sequences originate from the same allele (**Figure 12**).

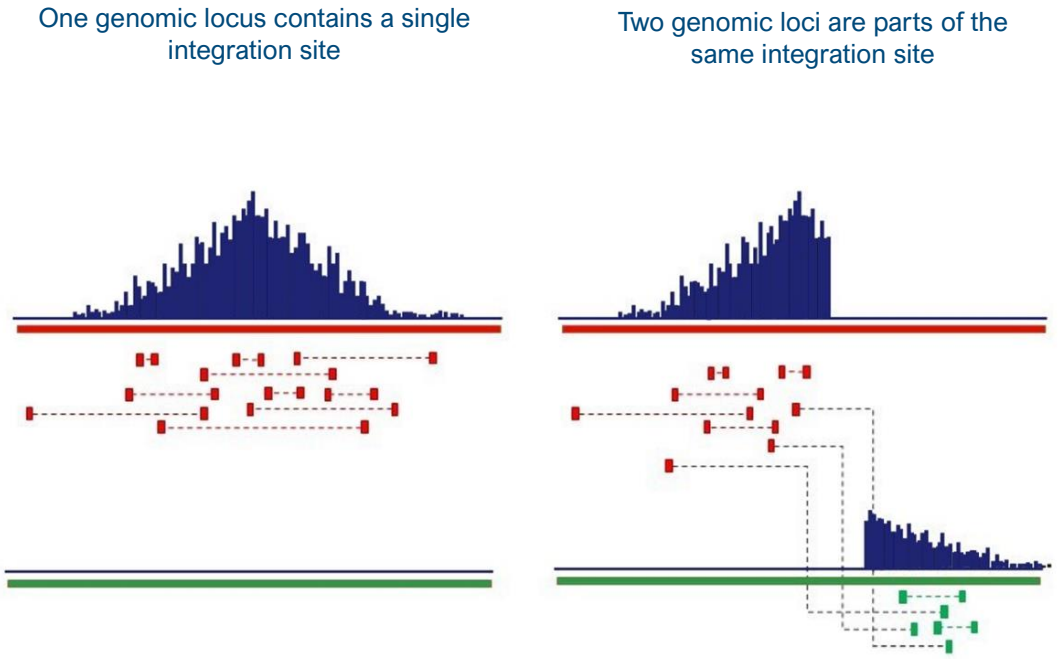


Figure 12: Schematic representation of the use of paired-end information in integration site analyses.

Copy number estimation

TLA is a targeted sequencing method. Routine vector and integration site sequencing analyses using a vector specific primer pair therefore only provide information about the present vector copies and do not provide information about the abundance of cells without the vector sequence.

Based on TLA data itself, copy numbers can only be estimated. For more accurate copy number determination ddPCR is recommended (see below).

Copy number estimations are based on three variables: the number of integration sites, the number of vector-vector junctions in a concatemer, and the ratio of the coverage on the vector-side and genome-side of the integration site (**Figure 13**).

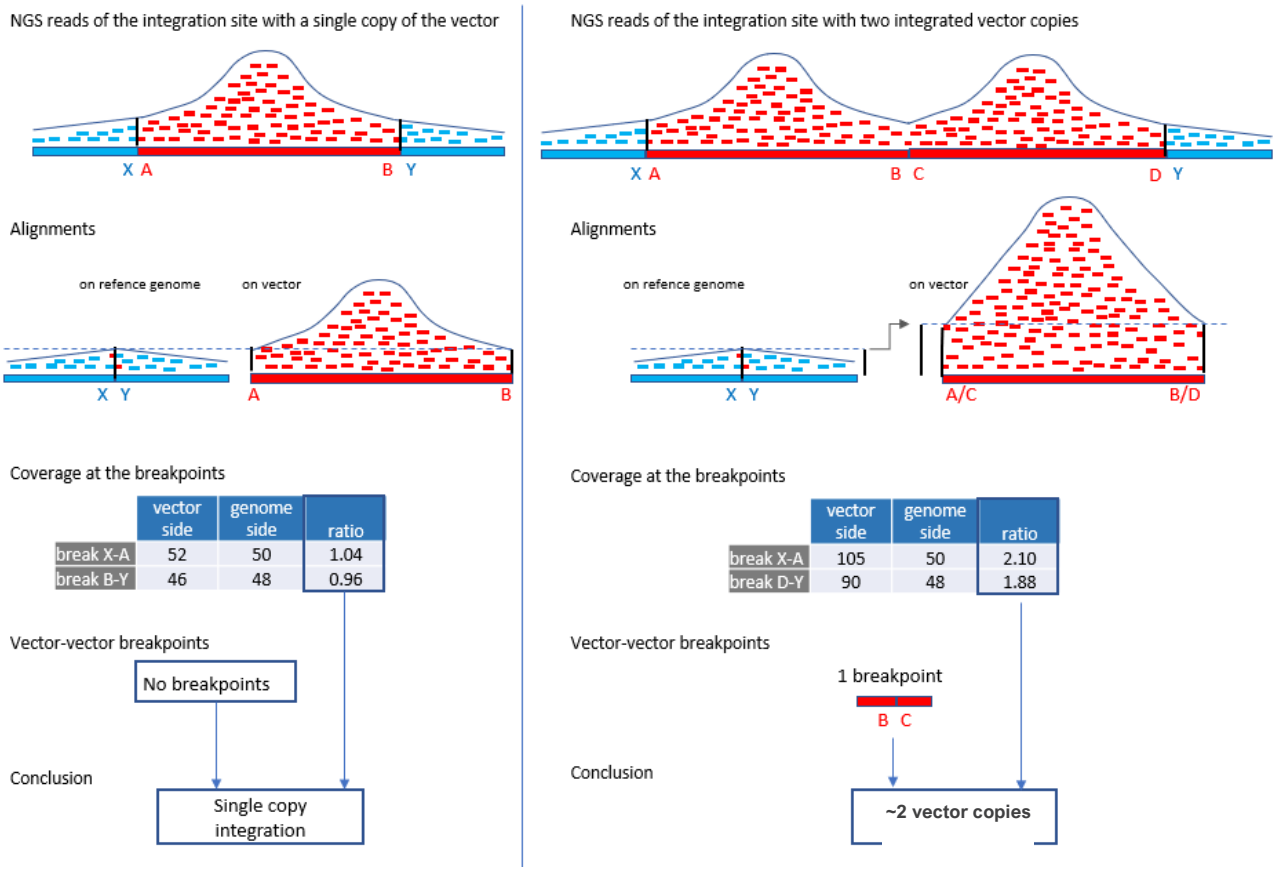


Figure 13: Schematic representation of vector copy number estimation based on TLA data.

If multiple integration sites are identified, standard TLA analyses cannot determine whether these occur in the same or different cells in the analysed sample. Standard TLA analyses can also not identify duplications of complete genomic loci/chromosomes in which vectors have integrated.

Therefore, the provided estimation gives information about the copy number present at an individual locus, which is not necessarily the same as the copy number per cell (**Figure 14-I**). In some cases, a TLA amplification from the genomic side of the integration site(s) can provide more detailed information on the vector copy number per cell (**Figure 14**).

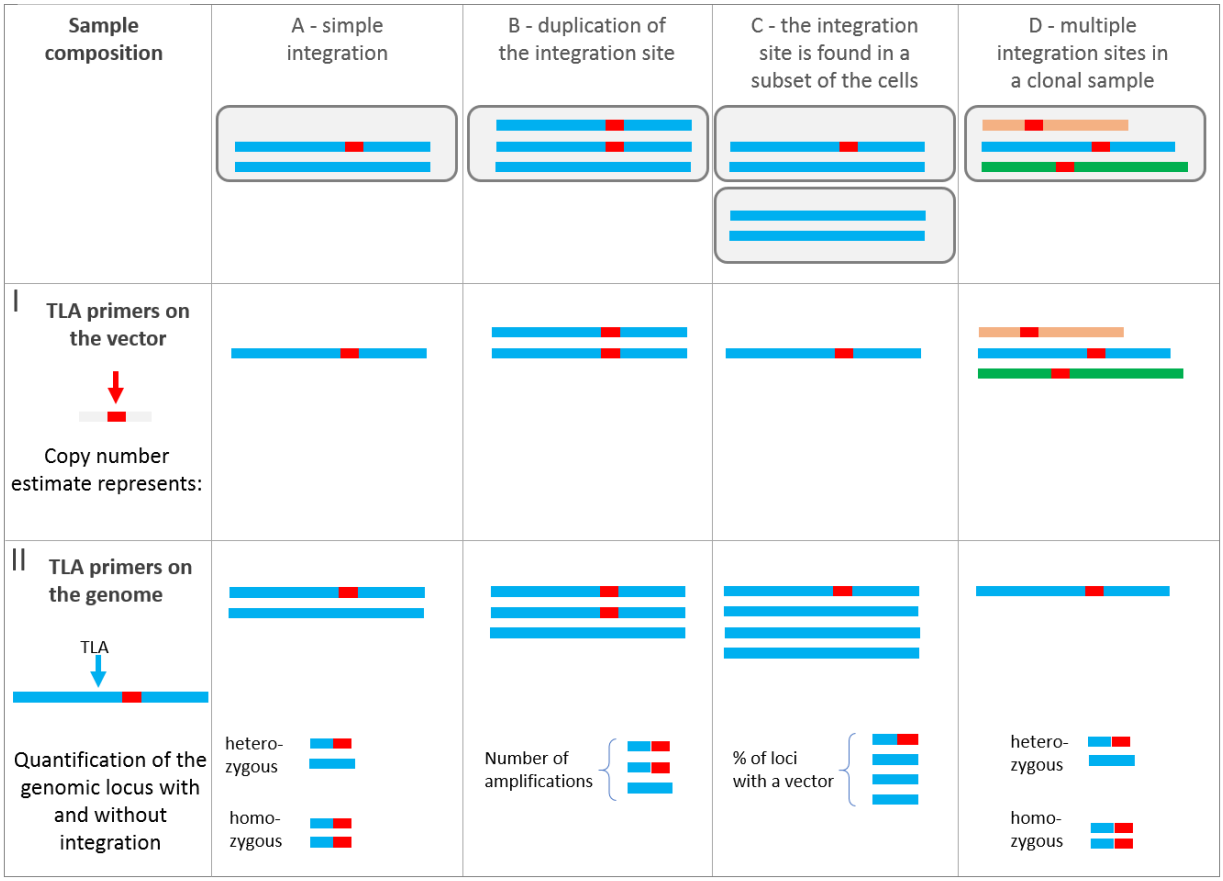


Figure 14: Copy number estimations in different sample types. I) Using vector specific primer sets, only the allele carrying the vector is interrogated. Based on this data alone, one cannot distinguish situations A, B and C. II) TLA amplifications performed from the genomic side of the integration site provides information on the relative frequency of alleles carrying a vector versus wild-type alleles. The combination of the information obtained with TLA amplifications targeting the vector and the genome can resolve the different sample types. Note that in a more complex situations (D), copy number estimations are not always possible.

Rearranged copies of the integrated vector can influence the measurements of the coverage ratio, as illustrated in **Figure 15**. Also, if many vector fragments are concatemerized, it can be difficult to estimate how many vector copies they jointly represent. Therefore, copy number estimations of high copy number vector integrations with complex vector-vector junctions are less accurate.

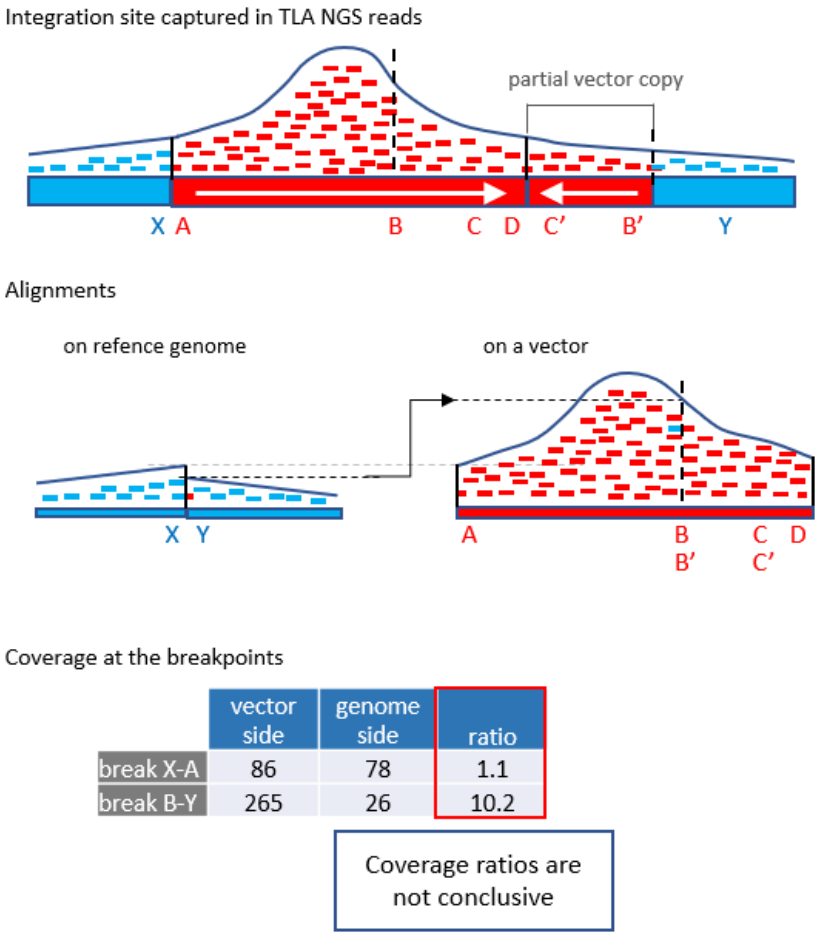


Figure 15: The effect of vector rearrangements on copy number estimations. If rearranged copies of the vector are present, the number of NGS reads produced on the breakpoint position in the vector can differ between the copies of the vector. In those cases, the total NGS coverage obtained on the breakpoint position is not a good representation of the number of vector copies. In a simple case represented here, the copy number can still be deducted, but in more complex cases this becomes increasingly difficult.

In heterogeneous samples with numerous integration sites, a copy number per integration site cannot be estimated based on the TLA data.

Accurate copy number assessment

Digital droplet PCR (ddPCR) allows accurate copy number quantifications and is performed to determine e.g.:

- vector copy number in clonal cell banks
- vector copy number stability over time
- average vector copy number in heterogeneous cell populations (like CAR engineered T-cells)
- presence of residual plasmid
- endogenous gene copy number

ddPCR workflow

ddPCR can be performed either in conjunction with TLA-based service or as a stand-alone experiment. ddPCR assays are designed on one or multiple selected regions of interest (i.e. vector element or endogenous gene) as well as two reference genes. Due to uncertainty about ploidy in modified cell lines two independent reference genes are used. For further confirmation of the ploidy of the cell line and improved vector copy number quantifications, an additional assay can be designed around a vector-genome breakpoint sequence in the modified cell line (previously determined by TLA analysis) as vector breakpoint sequences in random integration events are prone to occur on one allele.

gDNA is either extracted in house from viable frozen or crosslinked cells or gDNA received already isolated. The ddPCR reactions are set up as a duplex or multiplex PCR where a primer/probe pair targeting the region of interest is combined with another primer/probe pair targeting the reference gene tagged with different fluorophores, FAM and HEX are the most used fluorophores. DNA digestion is performed directly in the ddPCR reaction mixture (buffer, dNTPs, primers, and probes) using a restriction enzyme, according to the protocol supplied by Bio-Rad. The mixture is randomly distributed in 20,000 droplets in the QX200 Droplet Generator®. Thereafter, PCR amplification is performed in all droplets simultaneously. The number of droplets being negative and positive for the fluorophore signal is measured using a QX200 Droplet Reader and analyzed with the Bio-Rad QuantaSoft software (QX manager regulatory edition). The Bio-Rad QuantaSoft software fits the fraction of positive droplets to a Poisson algorithm providing the total copies of the region of interest and the reference genes.

Copy number quantifications

The copy number is determined as the ratio of the total copies for the region of interest to the total copies for the reference gene, multiplied by the number of copies for the reference (**Figure 16**). The latter is known, assumed or determined as follows: similar total copies are expected for the reference genes if the cell line sample has a stable diploid or polyploid for their complete genome. The ploidy of the cell line sample may not be previously determined and therefore unknown. In such cases, the cell line sample is considered diploid presenting with 2 copies of the reference genes unless any literature stating otherwise is available. Alternatively, the copy number is inferred based on the ddPCR assay for the breakpoint sequence of the vector integration site (**Figure 16**). For the latter, the copy number of the reference gene is determined as the ratio of the total copies for the reference gene to the total copies for the breakpoint sequence, assuming there is 1 copy of the breakpoint sequence per cell.

$$\text{copy number region of interest (i.e. vector element)} = \left(\frac{\text{total copies region of interest}}{\text{total copies reference gene}}\right) \times 2 \text{ (assuming a diploid genome)}$$

$$= \left(\frac{\text{total copies region of interest}}{\text{total copies breakpoint sequence}}\right) \times 1 \text{ (assuming integration site on 1 allele)}$$

Figure 16: The formulas used to calculate the copy number of the region of interest using genomic ploidy or integration site breakpoint as reference.

Example of ddPCR results

In the following example, the vector copy number was determined for two related samples (MCB and EOPC). The copy number of the vector elements X and Y was determined using ddPCR. Endogenous gene 1 and gene 2 were used as reference genes. In addition, a ddPCR assay for the breakpoint found during TLA analysis was included. The total copies of reference genes 1 and 2 were 3 times higher than the total copies for the breakpoint sequence of each sample (**Figure 17A**), resulting in approximately 3 copies per cell for these reference genes as compared to the breakpoint sequence. Hence, the results suggest that these cell line samples are triploid.

The total copies for the vector elements X and Y were 3 and 4 times higher than the total copies of the reference genes and 9 and 12 times higher than the breakpoint sequence (**Figure 17A**). This indicates the presence of 9 copies per cell for vector element X and 12 copies per cell for vector element Y in each cell line sample (**Figure 17B**). The difference in copy number between the elements indicates partial copies are also present in this sample.

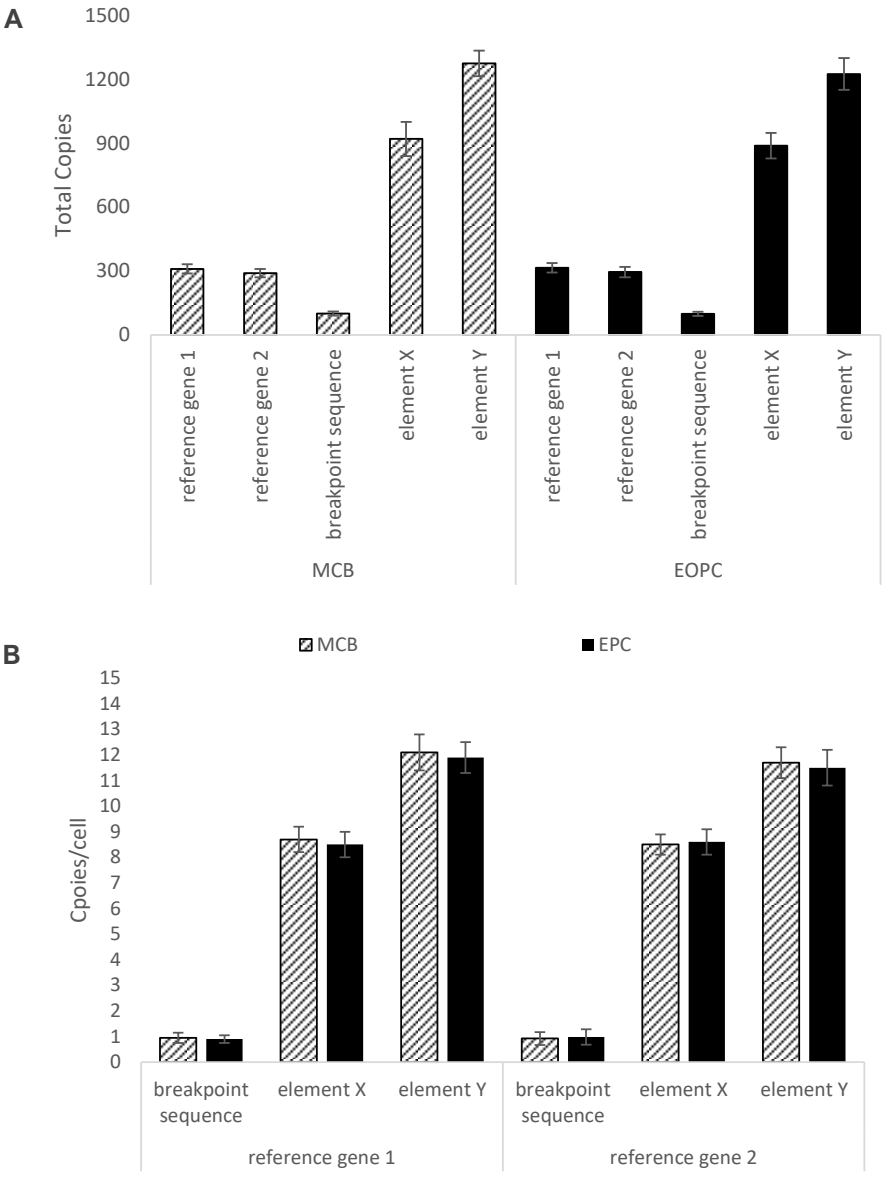


Figure 17: A. The total copies of the breakpoint sequence, genomic reference genes and element X and Y for MCB, and EOPC. B. The copies per cell of the elements X and Y as compared to the genomic reference genes 1 and 2 for MCB, and EOPC. The data represents mean ± SD of three experiments performed in duplicate.