Manual

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

For reports delivered before December 2021, please click here
for version 1 of this manual.

CERGENTIS
Genetic Engineering

TLA and sequencing

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

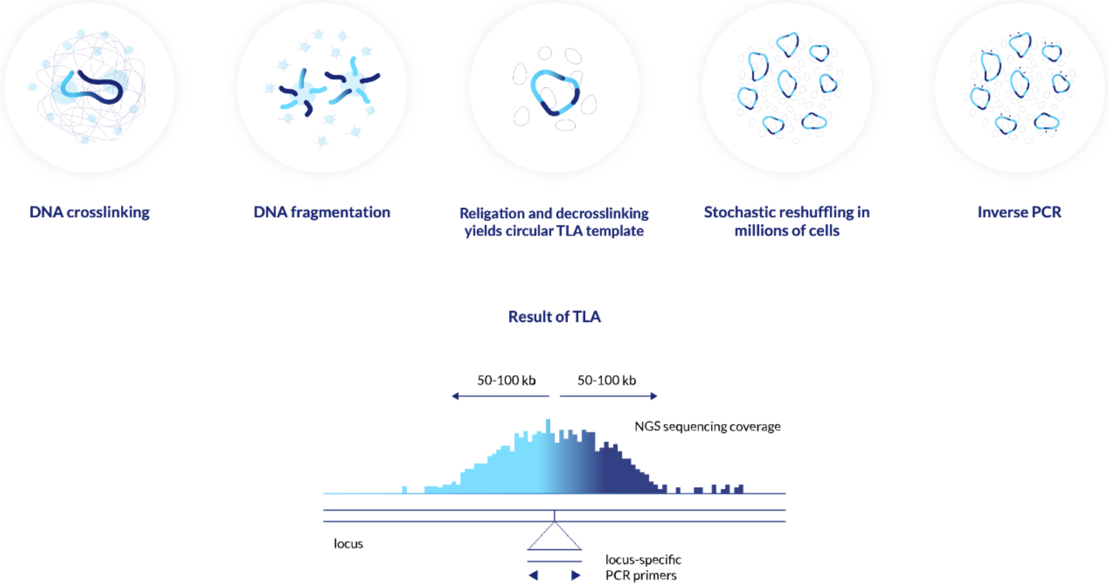

DNA crosslinking
DNA fragmentation
Religation and decrosslinking yields circular TLA template
Stochastic reshuffling in millions of cells
Inverse PCR
Result of TLA
50-100 kb
50-100 kb
NGS sequencing coverage
locus
locus-specific PCR primers

Figure 1: Schematic representation of TLA technology.

Using one TLA primer pair complementary to a sequence within a vector containing a transgene, sequence information is generated across the entire vector and its integration site(s) (**Figure 2**). Cergentis 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.

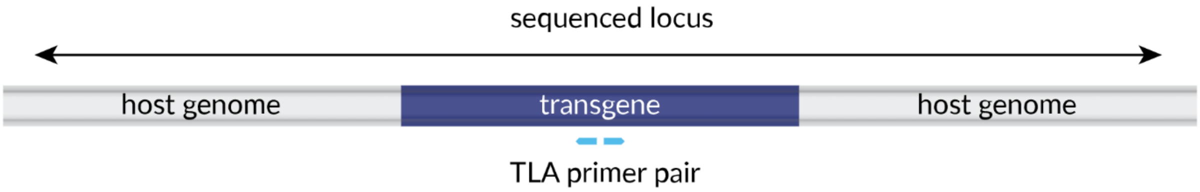


Figure 2: TLA-based transgene and integration site sequencing using transgene specific primer pairs.

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, 2x151 bases) is performed on an Illumina System. For a standard analysis, ~1 mln reads are generated per PCR product.

Alignment of sequencing reads

The Illumina® System, together with the bcl2fastq Conversion Software (Illumina) perform 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 a backbone. Cergentis recommends sending the full vector reference sequence in order to optimally assess its integration and the presence of sequence variants.

Coverage is defined as the number of NGS reads that cover a sequence (vector sequence or a 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/UTRs 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).

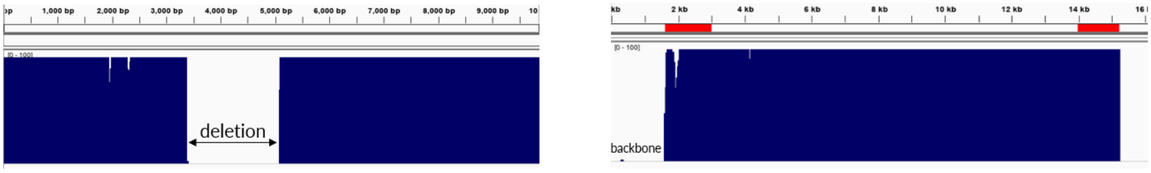


Figure 3: NGS sequencing coverage across the vector in the random (left) and targeted integration (right, homology arms are shown in red).

In samples with virus-mediated gene transfer, the obtained complete vector coverage might be a result of the 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 the 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 complementary to sequences common to all 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 (SNPs, 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/structural variants.

Sequence variants detection

The presence of sequence variants is determined using samtools mpileup (samtools version 1.3.1) (Li *et al.* Bioinformatics, Jun 2009 [PMID: 19505943], Li 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 and 20% for primary cells;
- The variant is present in the data generated on that sample 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;
- Low frequency variants (between 5-20% mutant allele frequency) are not found with similar frequencies in a independent control (Cergentis recommends the use of a 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 that map to the indicated position;
- Percentage (%) - percentage of reads containing the mutated allele (=mut/cov*100%).

Please note that using these 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.
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. In TLA we also have breakpoints around restriction sites, because of the DNA digestion and ligation that is performed in the experiment, the sequences that are in proximity but from different origins fuse together at these restriction sites, these are TLA induced fusions. The reads that are indicative of a fusion sequence are called fusion reads, split reads, breakpoint reads or chimeric reads.

There are three different breakpoint sequences derived from fusion reads that are reported in the results tables. 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.

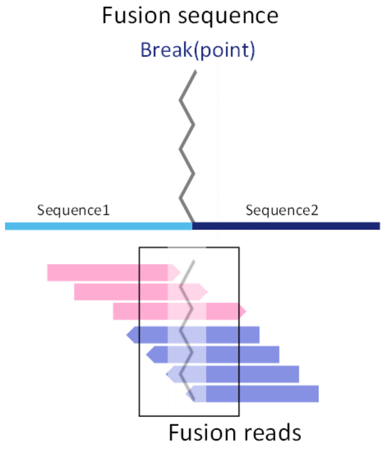


Figure 4: Schematic overview of a fusion or breakpoint sequence. Light blue and the dark blue sequence (on top) fuse together creating a fusion sequence. The point where the sequences fuse together is called the breakpoint. Based on the alignments of fusion reads (pink reads in forward orientation, pink reads in reverse orientation from the sequencer) you can reconstruct the fusion sequence and determine where sequence 1 and sequence 2 originate from.

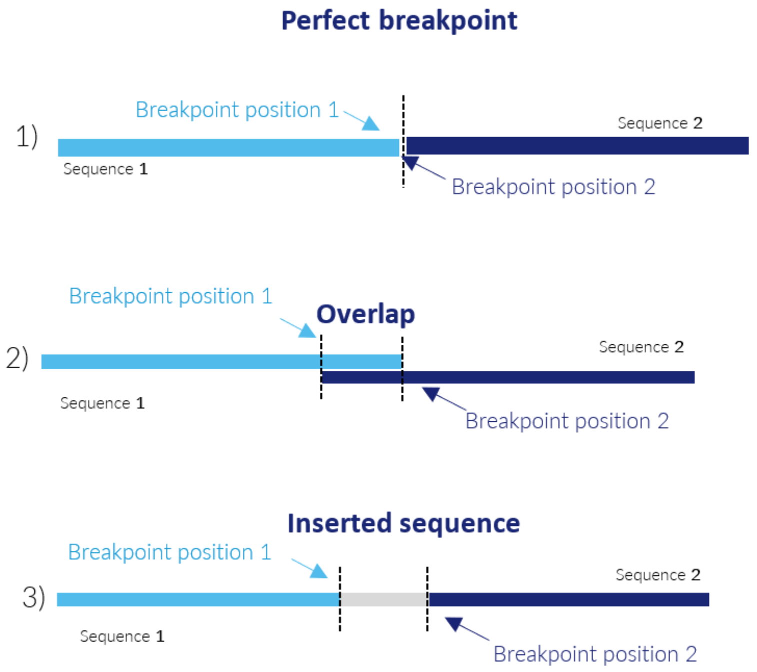
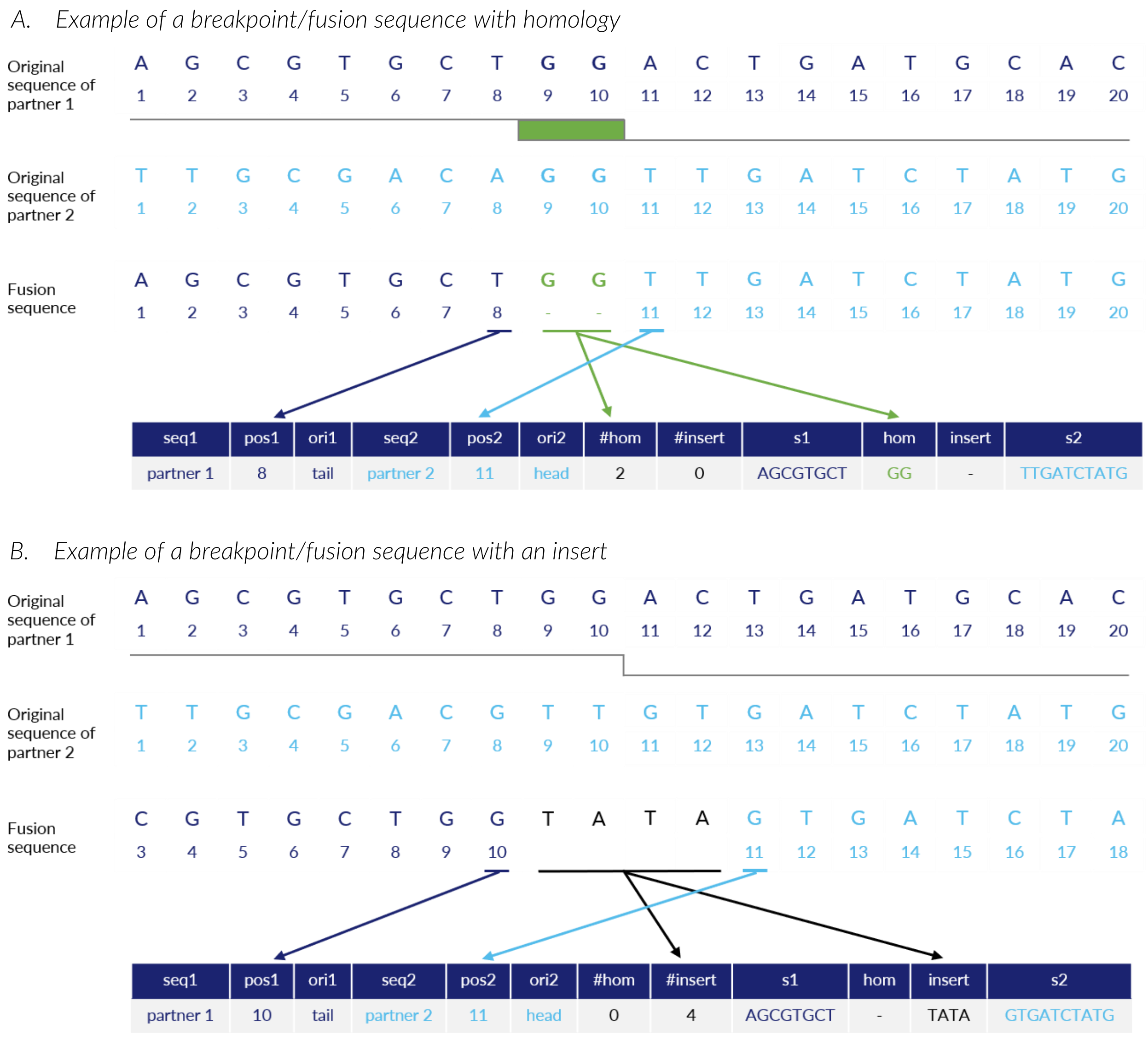


Figure 5: Schematic examples of different fusions. Example 1; perfect breakpoint, 2; fusion with homologous or overlapping bases, 3; fusion with novel or inserted bases.

The second type of breakpoint sequences that are reported in the table 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**). A part of sequence 1 also aligns to sequence 2. From the overlapping bases it is not known if they belong to sequence 1 or sequence 2. The breakpoint positions that are reported are the bases before or after the overlap because these are unique to sequence 1 and sequence 2 and the overlap is reported as homology between sequence 1 and sequence 2.

The third type of breakpoint sequences in the fusion tables 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 just a few bases or up to 50 bases and sometimes larger. The inserted sequence can be a known and aligned sequence or an unaligned sequence from unknown source / origin.



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

Original sequence of partner 1	A	G	C	G	T	G	C	T	G	G	A	C	T	G	A	T	G	C	A	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Original sequence of partner 2	T	T	G	C	G	A	C	A	G	G	T	T	G	A	T	C	T	A	T	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Fusion sequence	A	G	C	G	T	G	C	T	G	G	T	T	G	A	T	C	T	A	T	G
	1	2	3	4	5	6	7	8	-	-	11	12	13	14	15	16	17	18	19	20

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*

Original sequence of partner 1	A	G	C	G	T	G	C	T	G	G	A	C	T	G	A	T	G	C	A	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Original sequence of partner 2	T	T	G	C	G	A	C	G	T	T	G	T	G	A	T	C	T	A	T	G
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Fusion sequence	C	G	T	G	C	T	G	G	T	A	T	A	G	T	G	A	T	C	T	A
	3	4	5	6	7	8	9	10					11	12	13	14	15	16	17	18

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' and 'insert', 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 Cergentis 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 artefacts, e.g. breakpoints found at hairpin structures or low-complexity regions.

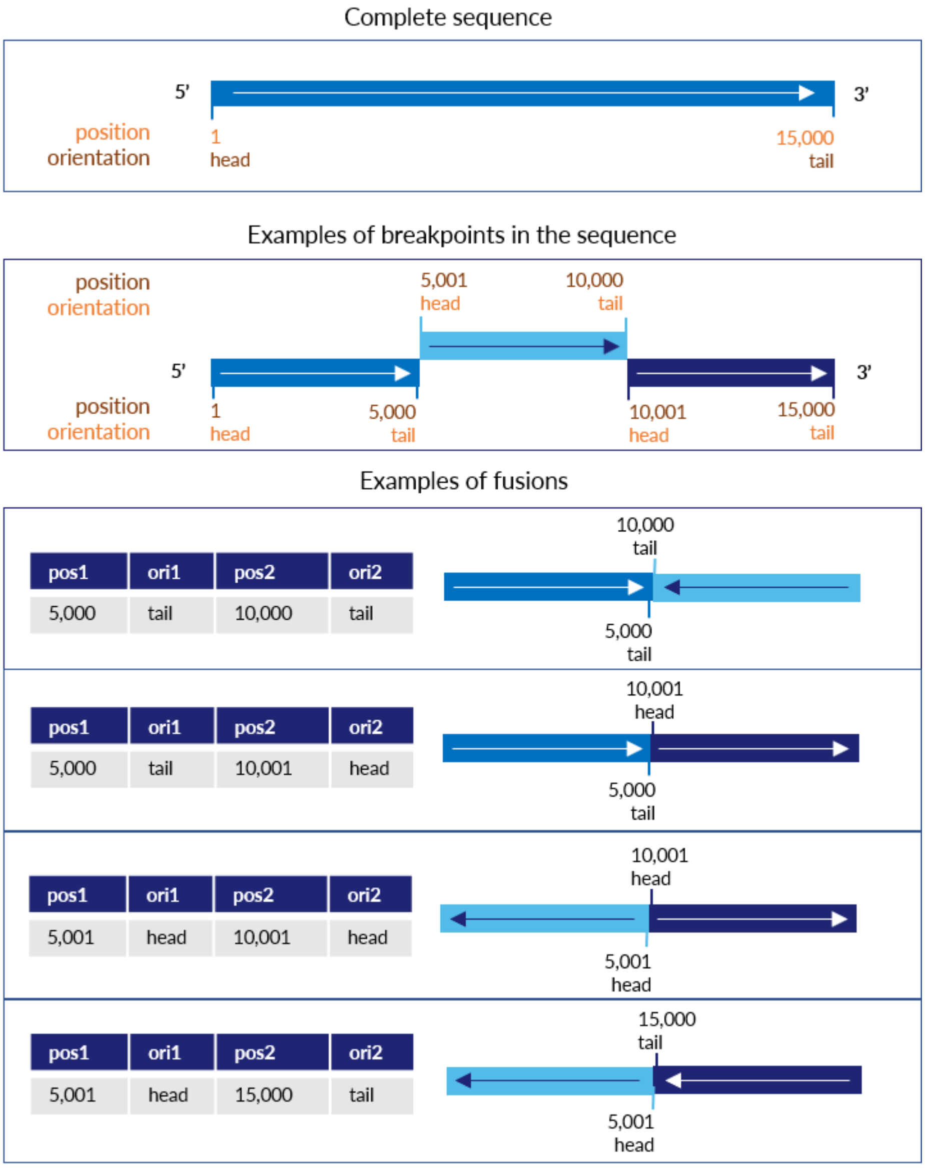
In heterogeneous samples, the detection of structural variants within individual integration sites is not possible. Only very abundant events like loss of the specific (backbone) sequences in the majority of integrations is identified.

Details regarding vector-vector breakpoint sequences

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

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

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.



pos1	ori1	pos2	ori2
5,000	tail	10,000	tail

pos1	ori1	pos2	ori2
5,000	tail	10,001	head

pos1	ori1	pos2	ori2
5,001	head	10,001	head

pos1	ori1	pos2	ori2
5,001	head	15,000	tail

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).

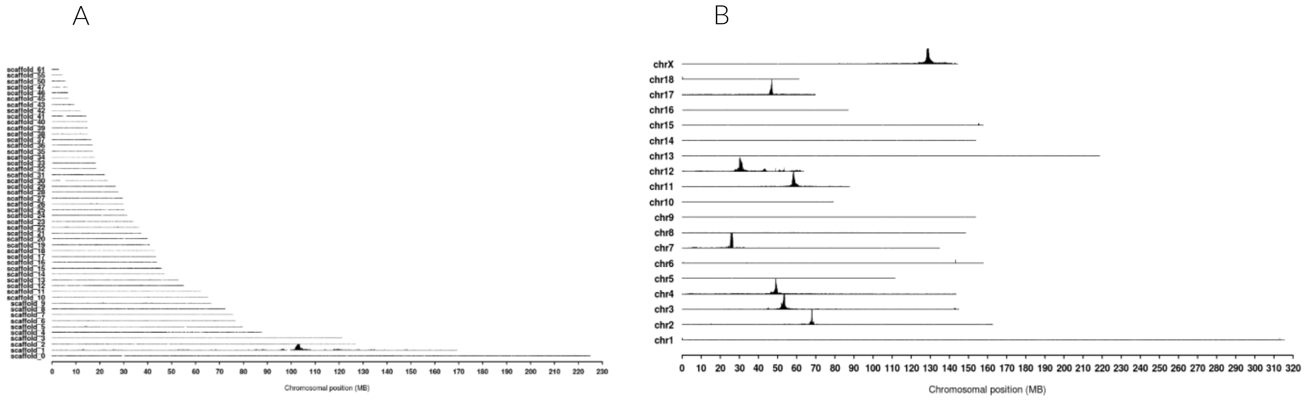


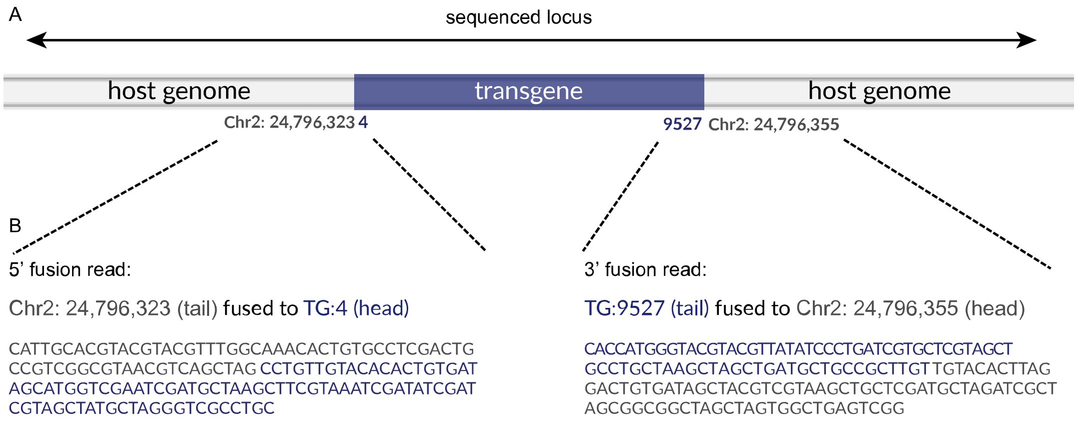
Figure 8: TLA sequence coverage across the Chinese Hamster Ovary (CHO) **(A)** and mouse **(B)** genomes. The top 50 covered scaffolds (A) and all chromosomes (B) are indicated on the y-axis, the chromosomal/scaffold position on the x-axis. High coverage peaks represent a single integration (A) and 8 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).



5' fusion read:

Chr2: 24,796,323 (tail) fused to TG:4 (head)

CATTGCACGTACGTACGTTTGGCAAACACTGTGCCTCGACTG
CCGTCGGCGTAACGTCAGCTAG CCTGTTGTACACACTGTGAT
AGCATGGTCGAATCGATGCTAAGCTTCGTAAATCGATATCGAT
CGTAGCTATGCTAGGGTCGCCTGC

3' fusion read:

TG:9527 (tail) fused to Chr2: 24,796,355 (head)

CACCATGGGTACGTACGTTATATCCCTGATCGTGCTCGTAGCT
GCCTGCTAAGCTAGCTGATGCTGCCGCTTGT TGTACACTTAG
GACTGTGATAGCTACGTCGTAAGCTGCTCGATGCTAGATCGCT
AGCGGCGGCTAGCTAGTGGCTGAGTCGG

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 quantification of integration events (i.e. sequence variants and vector-vector fusions) in that locus.

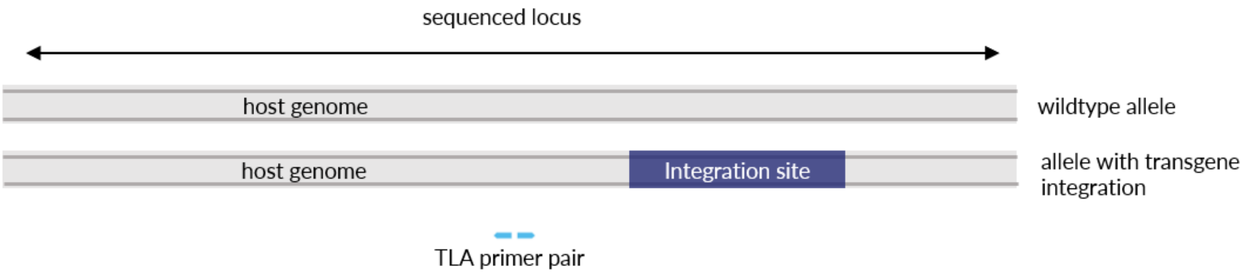


Figure 10: TLA-based analyses of individual integration sites. A TLA analysis with a primer pair in close proximity to the integration site provides sequence information across the entire locus.

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 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 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%).

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. 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 in integration sites

The presence of structural variants (genomic rearrangements) 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.

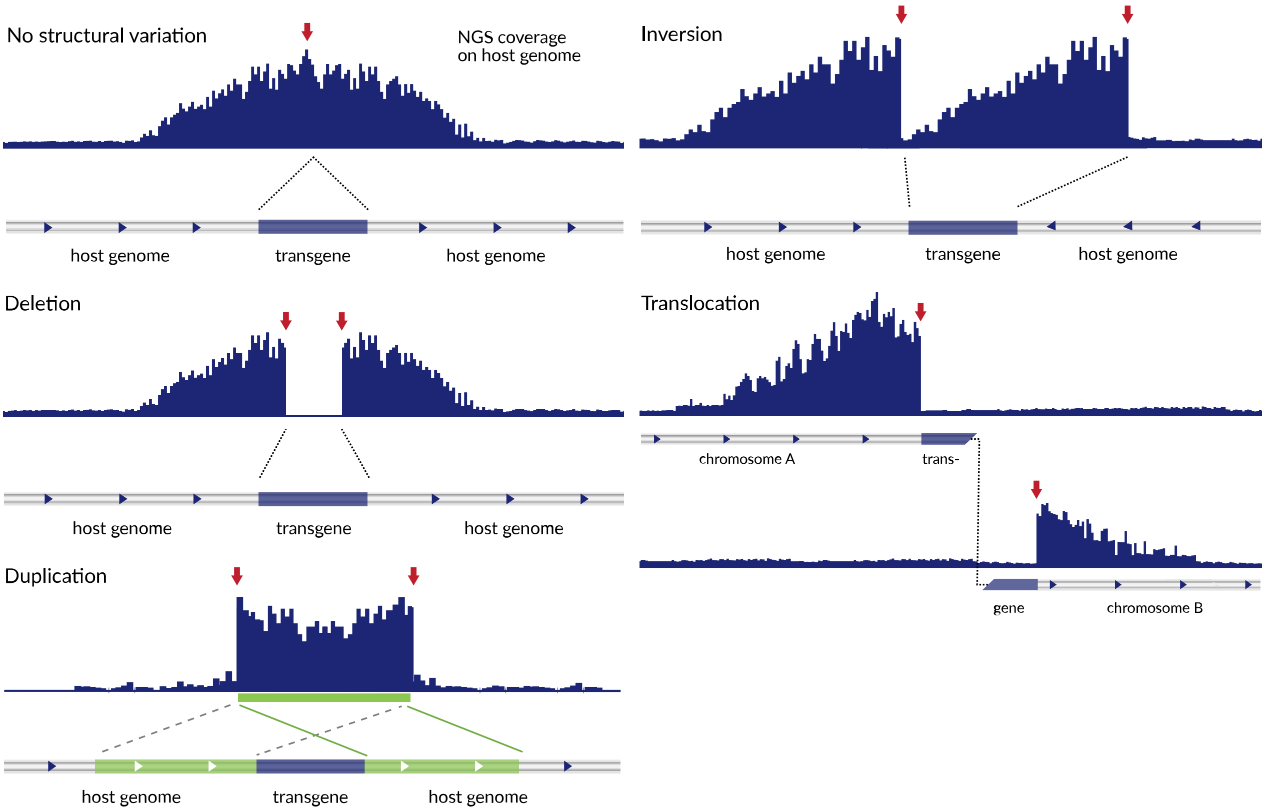


Figure 11: Schematic representation of the changes in the coverage profile and vector-genome breakpoint sequences (red arrows) that are observed in the most common types of rearrangements.

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).

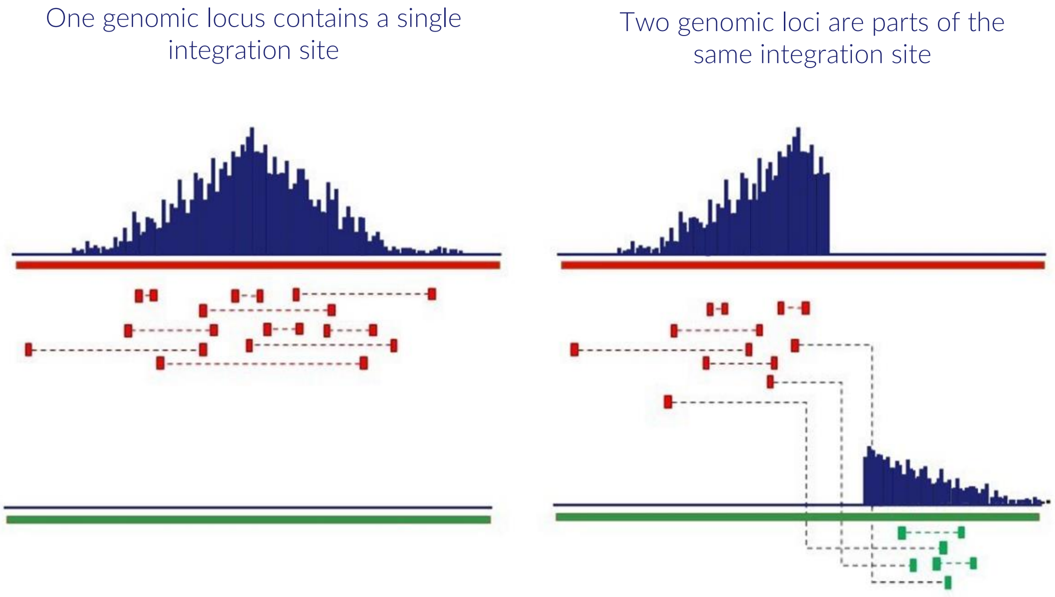

One genomic locus contains a single integration site
Two genomic loci are parts of the same integration site

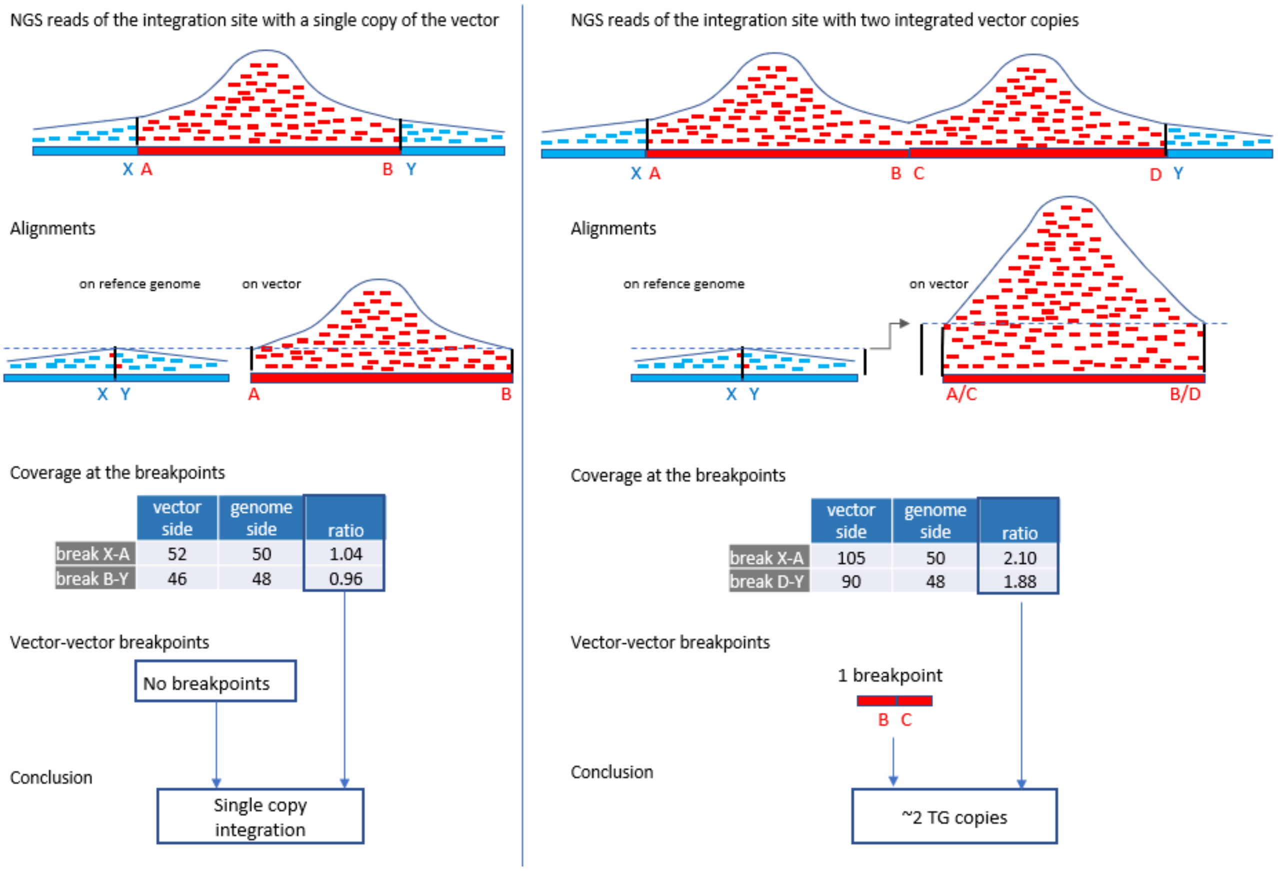
Figure 12: Schematic representation of the use of paired-end information in integration site analyses.

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.

TLA analyses do provide breakpoint sequences that can be used in (quantitative) PCR to quantify the abundance of identified integration sites in a population of cells. Based on TLA data itself, copy numbers can only be estimated.

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


NGS reads of the integration site with a single copy of the vector
X A
B Y
Alignments
on refence genome
on vector
X Y
A
B
Coverage at the breakpoints
Vector-vector breakpoints
No breakpoints
Conclusion
Single copy integration
NGS reads of the integration site with two integrated vector copies
X A
B C
D Y
Alignments
on refence genome
on vector
X Y
A/C
B/D
Coverage at the breakpoints
Vector-vector breakpoints
1 breakpoint
B C
Conclusion
~2 TG copies

	vector side	genome side	ratio
break X-A	52	50	1.04
break B-Y	46	48	0.96

	vector side	genome side	ratio
break X-A	105	50	2.10
break D-Y	90	48	1.88

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).

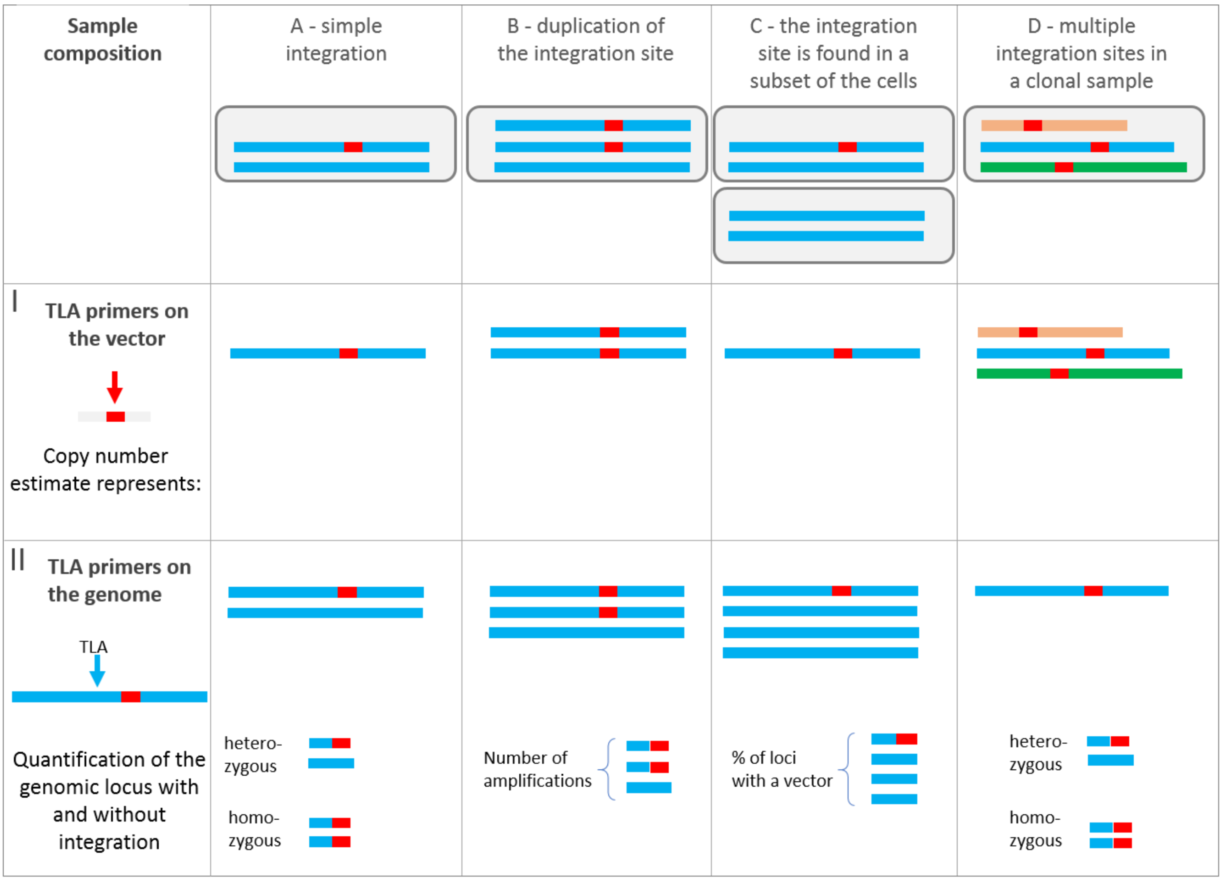
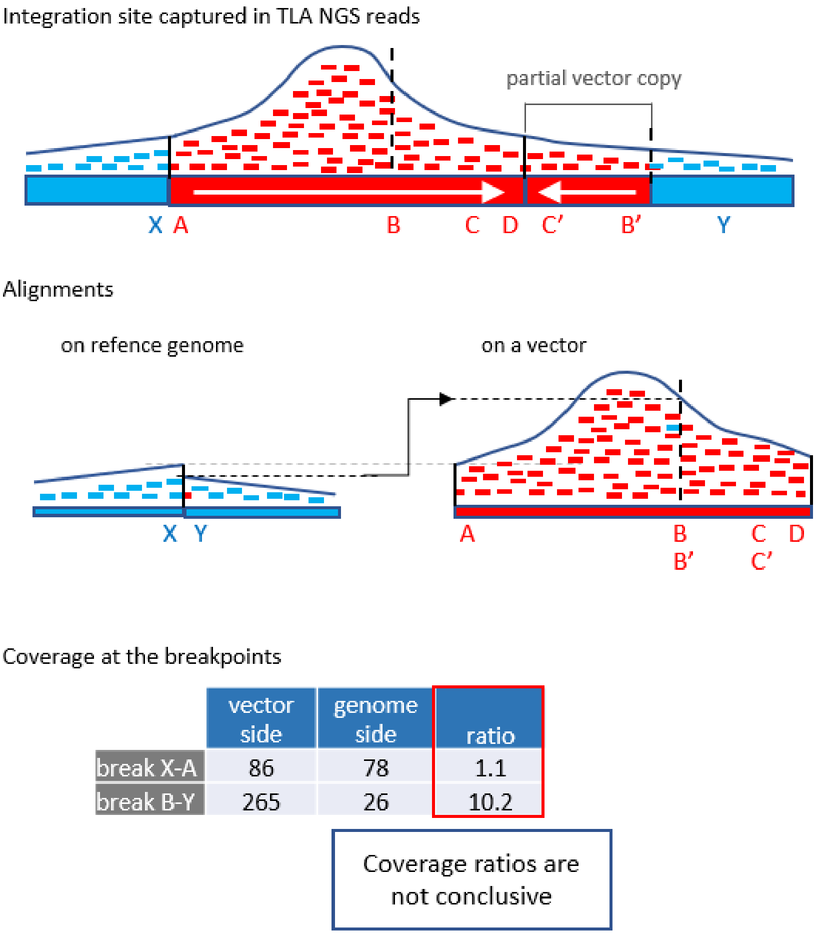

Sample composition
A - simple integration
B - duplication of the integration site
C - the integration site is found in a subset of the cells
D - multiple integration sites in a clonal sample
I TLA primers on the vector
Copy number estimate represents:
II TLA primers on the genome
TLA
Quantification of the genomic locus with and without integration
hetero-zygous
homo-zygous
Number of amplifications
% of loci with a vector
hetero-zygous
homo-zygous

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.

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

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.



	vector side	genome side	ratio
break X-A	86	78	1.1
break B-Y	265	26	10.2

Coverage ratios are not conclusive

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.