Example Report

Transgene analysis and integration site sequencing of 1 human T-cell sample containing provirus X

Prepared for:	Company name
	Company address
Customer:	Name
	Position
	Email
Internal project number:	X
Quote number:	X
Version:	1
Date:	X

Goal

In this study, 1 sample of human transgenic T-cells with the proviral X sequence was analyzed.

The aim of this analysis was to:

1. Study the provirus integrity:
    1) Determine the presence of sequence variants and their allele frequency.
    2) Determine the presence of provirus-provirus breakpoints that represent concatemerization of multiple copies of the provirus and/or structural rearrangements in a single proviral sequence.
2. Identify provirus integration site(s) and breakpoint sequences between the provirus and genome.
3. Characterize the distribution of the integration sites in the genome (annotated genes vs intergenic regions).

An overview of the TLA technology and technical details of the performed analyses is provided in the manual "Introduction to the terminology and methods used in TLA analyses_v2".

Summary

Sample	Provirus integrity	Number of integration sites
Sample A	6 sequence variants, no structural variants	1,078

Conclusion

In Sample A, 1,078 different integration sites are detected and the data indicate that the sample may contain many more. Out of them, 5% were found in gene exons, 48% in gene introns and 8% within 1 kb upstream the genes. 6 sequence variants and no structural variants are found within the integrated provirus.

TLA, sequencing and data mapping

Viable frozen human T-cells were used and processed according to Cergentis' TLA protocol (de Vree et al. Nat Biotechnol. Oct 2014). An overview of the TLA technology and technical details of the performed analyses is provided in the manual "Introduction to the terminology and methods used in TLA analyses_v2".

TLA was performed with 2 independent primer sets specific for the proviral sequence (Table 1).

Table 1: Primers used in TLA analysis

Primer set	Name	Direction	Binding position in provirus X	Sequence
1	psi+pack	RV	892	X
	psi+pack	FW	1,057	X
2	GOI	RV	2,894	X
	GOI	FW	3,253	X

The NGS reads were aligned to the proviral sequence and host genome. The human hg38 genome was used as host reference genome sequence.

Integration site detection in heterogeneous samples

In the studied samples, the transduction is performed using a retroviral provirus, and the breakpoints are expected to occur at the boundaries of the LTR sequences. In provirus X, the genome-provirus fusion events are analysed at the following locations:

Position A: 1 (head), the beginning of the 5'LTR and the provirus. Sequence NNNNN represents the identified genomic sequence:
NNNNN_AATGAAAGACCCCACCTGTAGGTTTGGCAAGCTAGCTTAAGTAACGCCATTTTGCAAGGCAT

Position B: 3,842 (tail), the end of the 3'LTR and the provirus. Sequence NNNNN represents the identified genomic sequence:
TCGCTGTTCCTTGGGAGGGTCTCCTCTGAGTGATTGACTACCCGTCAGCGGGGGTCTTTCA_NNNNN

These breakpoint sequences are filtered further:

- The breakpoints at the selected positions +/- 5 nucleotides were allowed due to microhomology between the proviral sequence and the genome.
- 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 provirus and the genome are filtered out.
- The breakpoint sequences aligning to both 5'-LTR and 3'-LTR are checked, and the duplicates are removed.
- 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 fusions are reported, so that the total number of identified integrations is slightly over-estimated. Here, 4 such integration sites out of 1,078 total sites (0.4%) are found.
- Optional: the breakpoint sequences present in an independent sample (if available) can be filtered out.

Results Sample 1

Provirus integrity

Figure 1 depicts the NGS coverage across the proviral sequence using primer sets 1 and 2.

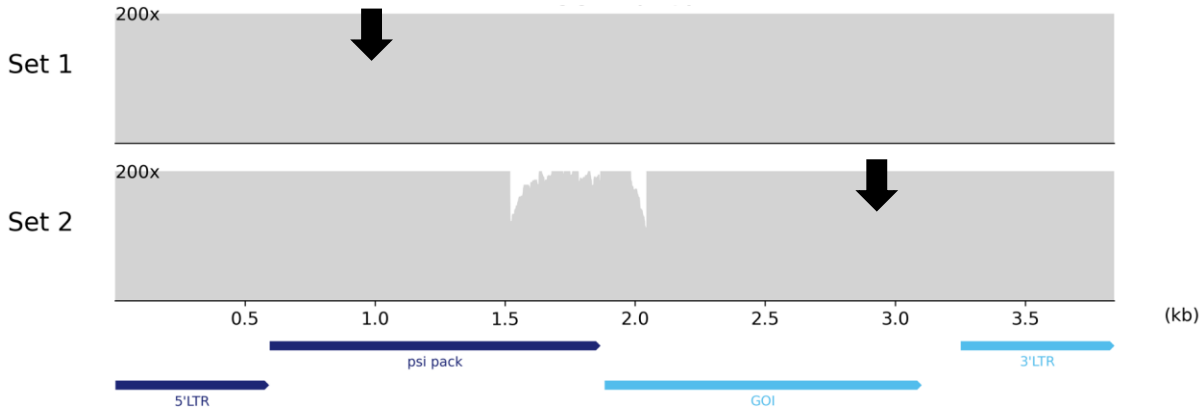

Figure 1a: NGS sequencing coverage (in grey) across the provirus. Black arrows indicate the primer location. The provirus map is shown on the bottom. Y-axes are limited to 200x.

High coverage is observed across the complete proviral sequence: 1-3,842 within the LTRs.

Because the sample is very heterogeneous, the coverage profile is not informative about individual integration sites, but the data show that all parts of the provirus are present in the sample and there is no evidence that specific regions are very frequently lost.

Sequence variants and structural variants were called in the covered regions.

Sequence variants

Detected sequence variants are presented in table 2. In very heterogeneous samples 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 virus 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 a negative 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 provirus.

Table 2: Identified sequence variants

					Primer set 1		Primer set 2	
Category	Region	Position	Reference	Mutation	Coverage	%	Coverage	%
A	Psi pack	720	A	C	967	100	1354	100
A	GOI	1967	T	-4AGTT	1181	100	1542	100
B	GOI	2956	T	C	1578	50	1262	56
B	GOI	3054	A	+1G	1631	52	1147	50
A	Backbone	3132	T	G	1845	100	1098	100

Provirus concatemerization and structural variants

No structural variants were identified in this sample. The heterogeneity of the sample prohibits the detection of variants within individual integration sites.

Integration sites

Whole genome coverage plot

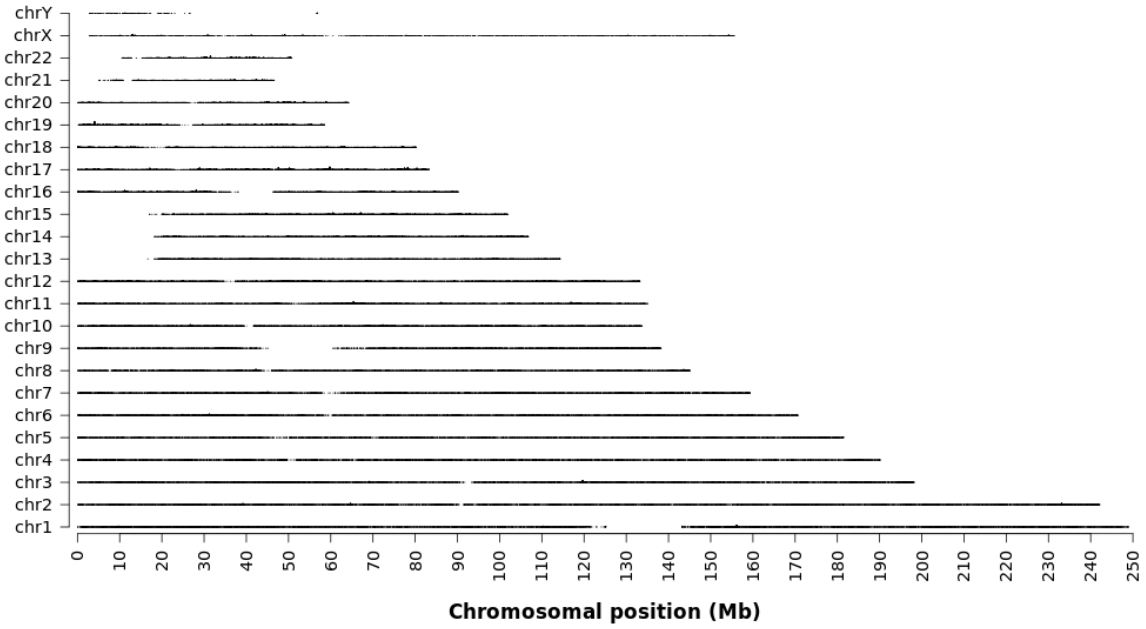

Figure 2: TLA sequence coverage across the human genome using primer set 1. The chromosomes are indicated on the y-axis, the chromosomal position on the x-axis.

As shown in figure 2, no large coverage peaks were seen on the genome due to heterogeneity of the sample, as reported below. When many integrations are present in the cell population and the coverage is viewed on a genome-wide scale, the coverage per integration site is too limited and no peak is seen. The identification of the integration sites is based on the analysis of the provirus-genome breakpoint sequences at the specified positions in the provirus (as described above).

Similar results were obtained with primer set 2.

Breakpoint sequences

1,078 integration sites were detected at the expected locations in the provirus (table 3). The accompanying results tables list the genome-provirus fusions that were identified. In this example report, a selection of the integration sites is shown in the Addendum.

Table 3: Identified integration sites

Primer set	Position A Proviral: 1 (+)	Position B Proviral: 3842/592 (-)	Total
Set 1	287	160	
Set 2	143	488	
Total per position	430	648	1,078

Most of the fusion sites are present in a limited number of reads (1-20) with no overlap between the data of 2 primer sets. Therefore, the analyzed sample is very heterogeneous.

The data do not show evidence for an integration site that occurs more frequently than other sites.

Gene annotation at the integration sites (optional deliverable, extra costs apply)

The integration sites were identified in all chromosomes in intergenic regions, introns and exons as shown in Figure 3.

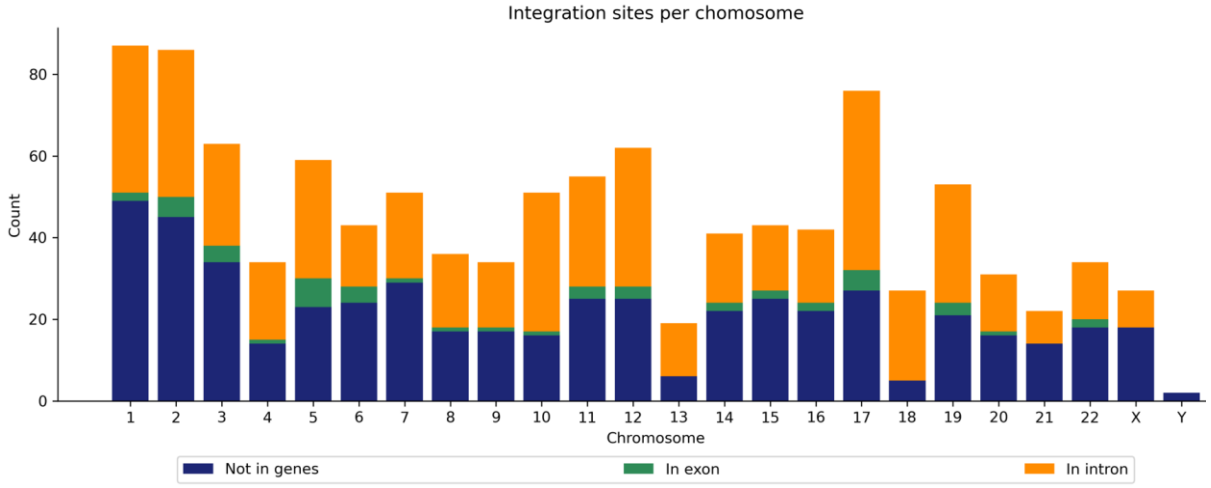

Figure 3: Distribution of the integration sites across all chromosomes.

Out of total 1,078 integration sites, 50 (5%) were found in gene exons, 514 (48%) in gene introns and 91 (8%) within 1 kb upstream the genes (Figure 4).

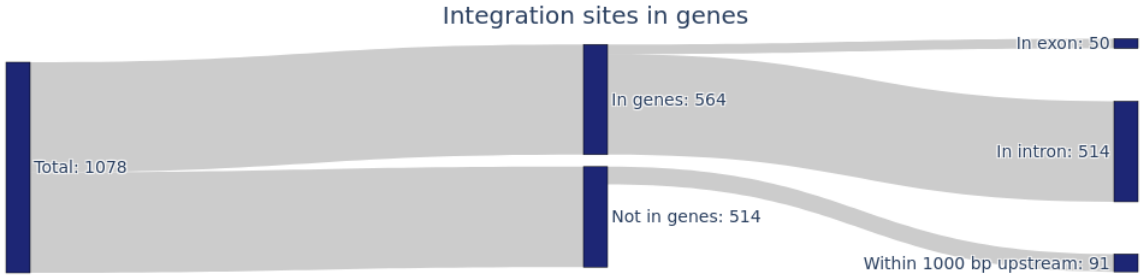

Figure 4: Distributions of the integration sites between the gene exons, introns, promoters (1 kb upstream region*) and intergenic regions.

* The size of the gene promoters/regulatory regions can be defined by the customer.

Addendum

The detailed description of 25 integration sites in shown in Table 4 which only illustrates the Excel tables accompanying the report. Typically, the full list of the identified integration sites will only be shown in the Excel tables.

Table 4: Positions and sequences of the integration sites

Column 1 (#seq1): Provirus name. Column 2 (pos1): position of the left side of the breakpoint (within the provirus). Column 3 (ori1): orientation of the left side of the breakpoint. Columns 4-5 (seq2-pos2): position of the right side of the breakpoints (on the genome). Column 6 (ori2): orientation of the right side of the breakpoint. Columns 7-8 (# of reads, set 1 and set 2): the absolute number of reads in which each breakpoint is found in the data of primer set 1 and set 2. Column 9 (Multiple hits): number assigned to each set of similar s2 sequences. Column 10 (count): total number of entries in a set of breakpoints with similar s2 sequences. Column 11 (s1): sequence of the left side of the breakpoint, only 10 bp are shown here. Column 12 (Hom): bases of homology found between the sequences at the left and right side. The homologous bases are not included when determining the positions as represented in columns 2 (pos1) and 5 (pos2). Column 13 (Ins): novel bases that are inserted at the breakpoint site. Column 14 (s2): sequence of the right side of the breakpoint, only 10 bp are shown here. Column 15 (in gene): gene annotated at pos2 (column 5). Column 16 (Intron/exon): location of the breakpoint in the introns or exons of the annotated genes.

#seq1	pos1	ori1	seq2	pos2	ori2	# of reads set 1	# of reads set 2	Multiple hits	count	s1	Hom	Ins	s2	in gene	Intron exon
Provirus	3	+	chr5	131383660	-	5	0			...GGGTCTTTCA	-	-	ACCCCAGCTC...	CDC42SE2	Intron
Provirus	3	+	chr10	73495069	+	3	0			...GGGTCTTTCA	-	-	GTGTGGGTGT...	PPP3CB	Intron
Provirus	3842	-	chr8	132759768	-	6	0			...GGGTCTTTCA	-	-	GATTGATTGA...	TMEM71	Intron
Provirus	3	+	chr3	52287259	+	0	2			...GGGTCTTTCA	-	-	AGCCCAGTCA...		
Provirus	3840	-	chr11	12232501	-	0	2	1	2	...GGGGGTCTTT	CA	-	ATCTGCCCAC...	MICAL2	Intron
Provirus	3842	-	chr11	113656168	-	0	2	1	2	...GGGTCTTTCA	-	-	ATCTGCCCAC...		
Provirus	3	+	chr11	14769930	+	1	0			...GGGTCTTTCA	-	-	ATTTAATCAT...	PDE3B	Intron
Provirus	3842	-	chr2	174592045	+	2	0			...GGGTCTTTCA	-	-	CTTCAATGTG...	WIPF1	Intron
Provirus	3	+	chr22	49943698	-	19	0			...GGGTCTTTCA	-	-	GAGACAATAG...		
Provirus	3842	-	chr16	25016647	-	2	0			...GGGTCTTTCA	-	-	CTTTTTTGTT...		
Provirus	3	+	chr9	75150898	-	0	5			...GGGTCTTTCA	-	-	AGAAAAGGTA...		
Provirus	3	+	chr4	25859374	+	0	6			...GGGTCTTTCA	-	-	CCGTGACTTA...	SEL1L3	Intron
Provirus	3	+	chr17	18314853	+	0	2			...GGGTCTTTCA	-	-	GCGCCCACCG...	TOP3A	Exon
Provirus	3	+	chr15	22265692	+	5	0			...GGGTCTTTCA	-	-	CACTCATCTC...		
Provirus	3842	-	chr1	219657790	-	2	0			...GGGTCTTTCA	-	-	CATGGTCTTC...		
Provirus	3	+	chr11	18014006	+	0	16			...GGGTCTTTCA	-	-	AGCTCTTACA...		
Provirus	3841	-	chr22	18657648	-	1	0	2	2	...GGGGTCTTTC	A	-	CATGAAGAAA...		
Provirus	3841	-	chr22	18711782	+	1	0	2	2	...GGGGTCTTTC	A	-	CATGAAGAAA...		
Provirus	3	+	chr1	169688951	+	0	3			...GGGTCTTTCA	-	-	ATTTGAACAG...		
Provirus	4	+	chr7	21353711	+	0	1			...GGGGTCTTTC	A	-	CAATCTGTGC...		
Provirus	3841	-	chr17	35028544	-	0	7			...GGGGTCTTTC	A	-	CTGAGAAAAA...	RFFL	Intron
Provirus	3	+	chr22	36333708	-	15	0			...GGGTCTTTCA	-	-	CTGGAATGAA...	MYH9	Intron
Provirus	3842	-	chr19	10427866	+	20	0			...GGGTCTTTCA	-	-	GCTCATGCCC...	PDE4A	Intron
Provirus	3842	-	chr9	19161267	+	0	1			...GGGTCTTTCA	-	-	ATGCCTGGAA...		
Provirus	3	+	chr2	25273304	-	0	1			...GGGTCTTTCA	-	-	CCCTGTCTCT...	DNMT3A	Intron

Sample and Study details

Sample receipt date	X
Condition of sample at receipt	X
Start date in the lab	X
Sequencing run	X
Date data analysis	X
Deviations from the protocol	X
TLApp version:	X

Study Personnel

Lab technician	X
Data Analyst	X
QC Analysis and Report	X

Quality control

The results are independently verified and reviewed and are an accurate and complete representation of the study. TLA processing of cells, NGS sequencing, and data analysis are ISO/IEC 17025:2017 accredited by the Dutch Accreditation Council RvA, Registration number L671.

Scientific approval	X
Date	X
Signature	X