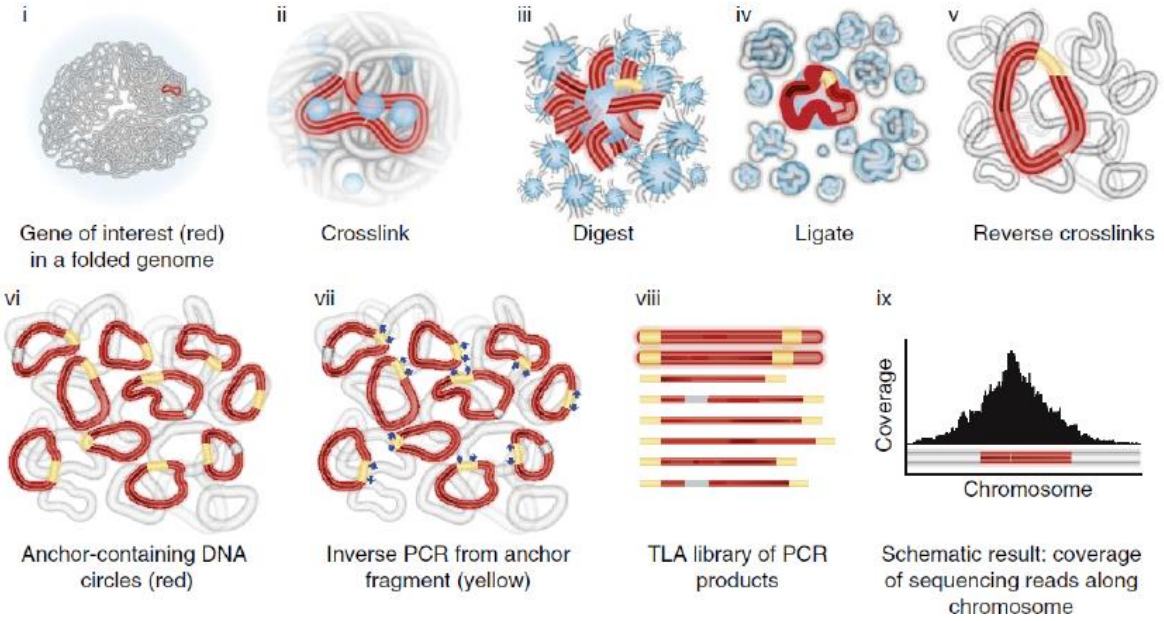
Transgene Mapping Analysis

Tg(Hu PRNP BAC)2632

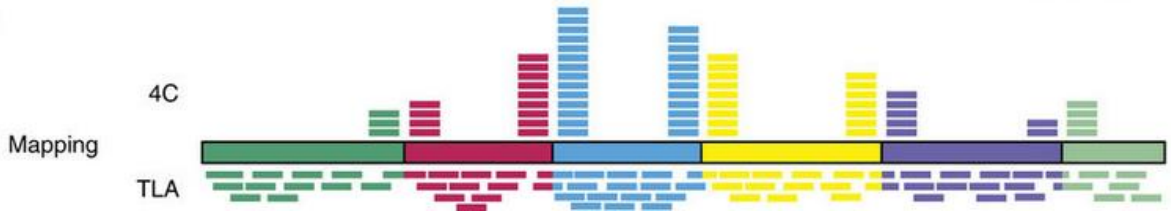
November 20, 2018

TLA technology



Targeted sequencing using TLA.
Neighboring sequences that form a gene or genetic locus (red) are in close spatial proximity (i) and therefore are preferentially crosslinked (ii). Digestion with a frequently cutting enzyme (iii) and ligation (iv) results in large DNA circles composed of multiple crosslinked restriction fragments (v). Different copies of a locus (from different cells) result in DNA circles composed of co-captured restriction fragments. Limited trimming (with a compatible but less frequently cutting enzyme) and ligation creates PCR-amplifiable DNA circles (vi). Fragments captured with a fragment of interest (the anchor sequence, yellow) are selectively PCR-amplified with anchor-specific inverse PCR primers (blue arrows) (vii). The resulting sample (viii) is highly enriched for locus-specific sequences and can be processed with standard library preparation procedures for next-generation sequencing. Mapped reads originate from the locus of interest and collectively span tens of kilobases (ix).

TLA technology



In TLA, the entire restriction fragments are amplified and sequenced, whereas 4C analyzes the ends of the fragments. Sequencing reads are shown as colored blocks (top and bottom). TLA data can thus be used to build contigs representing the sequence of a genetic locus of interest.

De Vree et al.,
Targeted sequencing by proximity ligation for comprehensive variant detection and local haplotyping.
Nat Biotechnol. 2014 Oct;32(10):1019-25. doi: 10.1038/nbt.2959. Epub 2014 Aug 17.

Experimental setup

A splenocyte sample from one male (M33833.2; line 17003) of the Tg(Hu PRNP BAC)2632 transgenic line has been analyzed by TLA technology.

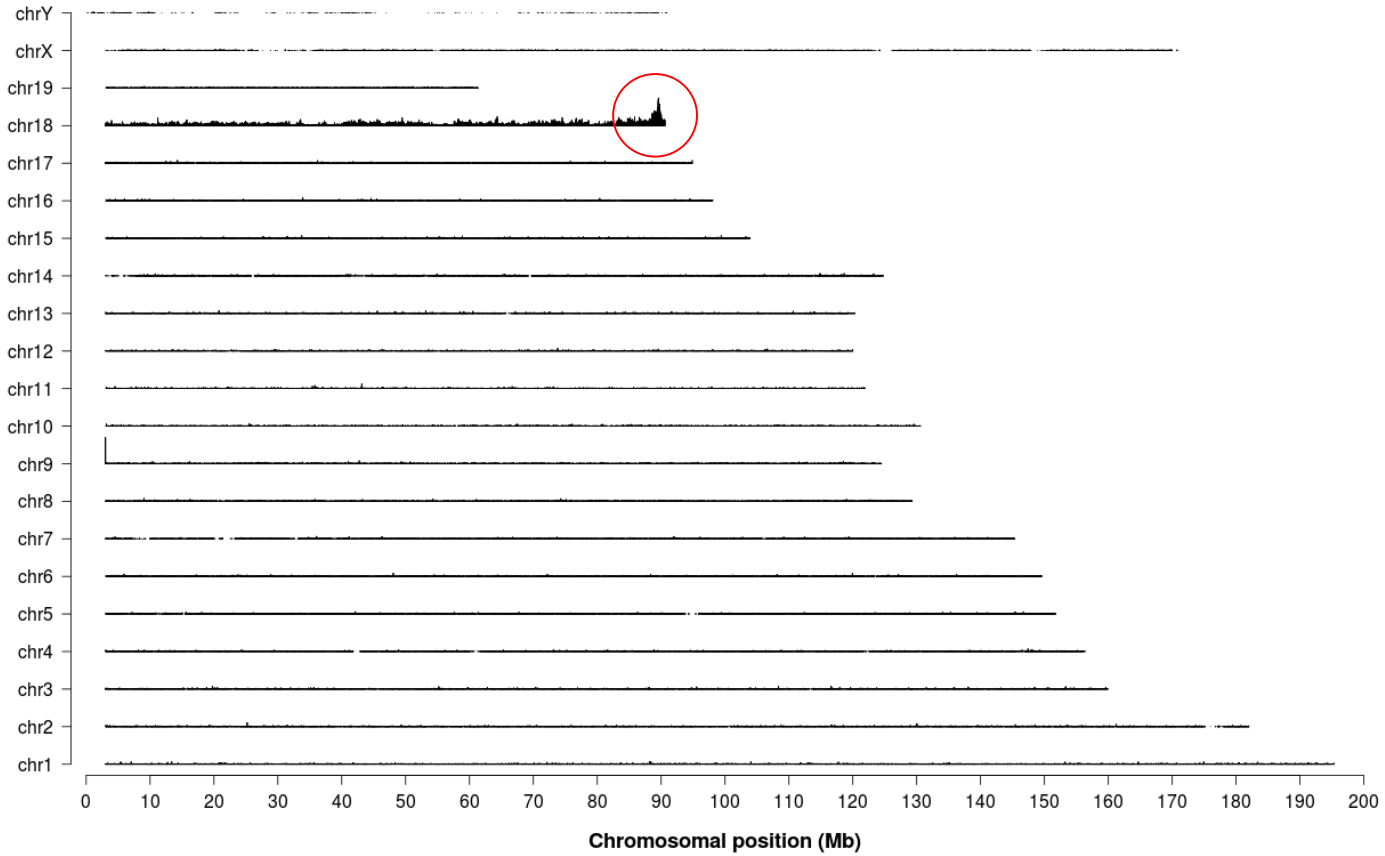
2 primer sets were designed on the transgene :

set 1	fw	CCTTTCTTGCTTTATTAGATCCA
	rv	TGCGGATCATTCACTGGTA
set 2	fw	CGTCCAACATCAATACAACC
	rv	CAGTTTCATTTGATGCTCGA

The primer sets were used in individual TLA amplifications. PCR products were purified and library prepped using the Illumina Nextera flex protocol and sequenced on an Illumina sequencer. Reads were mapped using BWA-SW (Li et al. Bioinformatics, 2010 [PMID: 20080505]), version 0.7.15-r1140, settings bwasw -b 7. The NGS reads were aligned to the TG sequence and host genome. The mouse mm10 genome was used as host reference genome sequence.

Integration site analysis

- TLA sequence coverage



1: TLA sequence coverage across the mouse genome using primer set 2. The chromosomes are indicated on the y-axis, the chromosomal position on the x-axis. Similar results were obtained with primer set 1.

As shown in figure 1, the TG has integrated in chromosome 18.

Integration site analysis

- Locus wide coverage

Dok6

2: TLA sequence coverage across the TG integration locus; mouse chr18:89,344,107-90,025,894. Sequence coverage (in grey) generated with primer set 1 (top) and primer set 2 (bottom). The red arrows point toward the breakpoint sequences identified with both primer sets. Y-axis is limited to 10x

Integration site analysis

- Integration site spanning read

The following reads were identified marking the TG integration:

5' integration site: Chr18:89,628,561 (tail) fused to TG: 44,070 (tail) with 3 homologous bases

CTGCTATACTCAGAGATCAGTTAAATGATTAAGCATCATCAGATTATCTTCCTTTTATAGCAGATGGGA
ATAAATACAGAAAACCACTTTGTGCCAGACTTGGGGGTGTGGTGGCTGGATTTTAGGAGCTATATCC
AGAATGGAAACACGG

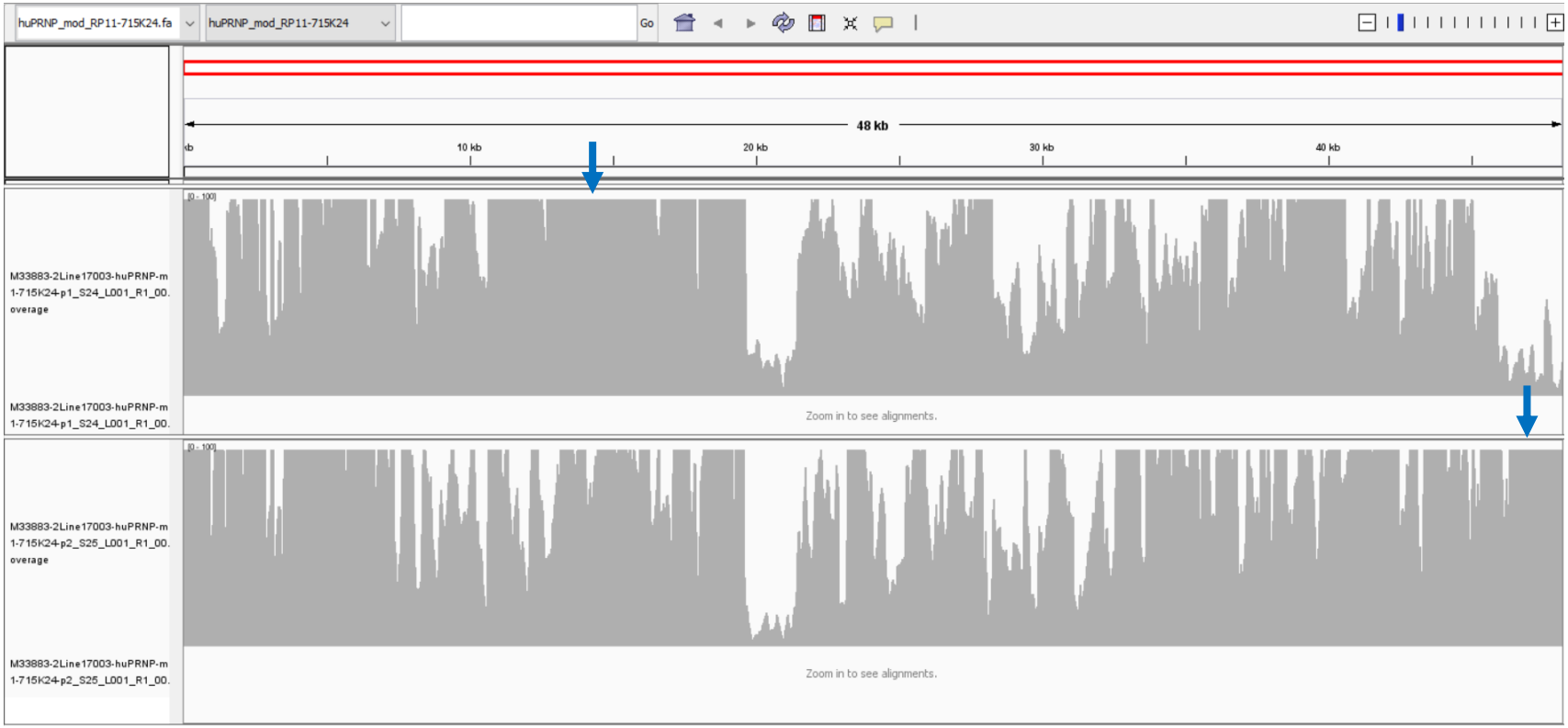
5' integration site: TG:45,936 (tail) fused to Chr18: 89,676,017 (head) with 3 homologous bases

CCCTGAGTGTACACCTTGAAATGGCGAATCTCATGTTACGGGATGTCAATATTGAGTAAGAAATTTTT
ACAGGATAAAAGAAAGTTAGTAAGCAC

Because of the presence of a CATG site close to the breakpoint, validation is recommended.

The coverage profile in figure 2 shows a 47kb deletion has occurred in the region of the integration site. From this data it is concluded that the TG has integrated in mouse chromosome chr.18:89,628,561-89,676,017. According to the refseq this is in the intronic region between exons 1 and 2 of *Dok6*.

TG sequence



3: NGS sequencing coverage across the TG with primer set 1 (top) and primer set 2 (bottom). Blue arrows indicate primer location. Y-axis is limited to 100x. Good coverage is observed across almost the complete TG sequence TG 1- 48,422. No coverage is seen between TG: 1,458-1,502 and TG: 8,164-8,183 due to low complexity regions. Low coverage is observed between TG: 19,661-21,441 due to a GC rich region. Local dips in the coverage profile are due to GC rich regions.

TACONIC

Table 1:Small variants in TG.

SNVs 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 20%,
- the variant is present in the data of all primer sets with coverage in the region,
- for at least one of the primer-sets the coverage is >=30X,
- the variant is identified in both forward and reverse aligning sequencing reads,
- low frequency variants (between 5-20% mutant allele frequency) are not found with similar frequencies in a negative control (if included).

			Primer set 1		**Primer set 2**	
Position	Reference	Mutation	Coverage	%	Coverage	%
895	C	T	73	100	122	100
1921	A	G	88	99	120	98
2659	A	G	22	95	82	99
3574	T	C	108	100	153	100
4775	G	A	94	99	181	97
4866	A	G	83	99	130	98
6955	C	G	77	100	95	100
7968	G	A	106	100	81	100
9143	C	T	75	100	36	97
11821	A	-3TTG	103	88	51	78
18645	A	G	165	100	103	100
21573	T	-1A	59	97	39	100
35890	C	T	112	98	73	96
41399	T	-5CAAAA	119	56	173	83
42801	C	G	111	99	85	100
44766	C	-1A	89	25	133	28

The presence of SNVs is determined using samtools mpileup (samtools version 1.3.1) (Li et al. Bioinformatics, Jun 2009 [PMID: 19505943], Li et al. Bioinformatics, Nov 2011 [PMID: 21903627]).

Structural variation

Fusion sequences consisting of two parts of the T, are identified using a proprietary Cergentis script. Fusions resulting from the TLA procedure itself are recognized by the restriction enzyme-specific sequence at the junction site and removed.

TG-TG fusions are reported that meet the following criteria:
- the TG-TG fusion is present in >1% of the reads at the position of the fusion,
- the TG-TG fusion is observed in data of both primer-sets, unless the data provides a clear explanation why the fusion is not found in one of the data sets,
- the TG-TG fusion is not present in negative control sample(s) (if included),
- visual inspection of the TG-TG fusions in an NGS data browser is performed to remove fusions that are sequencing artefacts, e.g. fusions found at hairpin structures or low-complexity regions.

A total of 2 TG-TG fusions were found within the TG, indicating concatemerization of the TG sequence:

1. TG: 1 (head) fused to TG: 48,422 (tail)

GTCTCCAGCTGCATGACTCTTAAAGCGGCCGCGGGCCCGAGCTTAAGTAACTAACTAACAGGAAG
AGTTTGTAGAAACGCAAAAAGGCCATCCGTCAGGATGGCCTTCTGCTTAGTTTGATGCCTGGCAG
TTTATGGCGGGCGTCCTGCCCGCCACC

2. TG: 10 (head) fused to TG: 45,959 (tail) with 8 homologous bases

GAATCTGTCTCCAGCTGCATGACTCTTAAAGCGGCCGCAATAGCACCATTGGGGTAATTCCCGTAA
CATGAGATTCGCCATTT

Copy number

An exact copy number cannot be determined using TLA. However, an estimation can be made based on the number of integration sites, number of fusion reads and the ratio of coverage on the TG and genome integration site.

In this sample, the coverage on the TG is much higher than on the genomic side of the integration (roughly 10-20 times). 2 TG-TG fusions are found. The copy number is estimated to be 10-20 copies.

Summary

- The TG has integrated into the intronic region between exon 1 and 2 of the *Dok6* gene (chr 18: 89,628,561-89,676,017)
- The integration event led to a 47kb deletion of mouse genome sequence.
- A number of SNVs has been identified in the TG sequence of this sample.
- The copy number is estimated to be 10-20 copies.

Proposed genotyping assays

5’ integration site:

5'int_wt_rv	AGATCTTTCCTACCTTTTCG
5'int_tg_rv	TATAGCTCCTAAAATCCAGC
5'int_fw	TCAAATACTACTGTTCTGCA
wt amplicon:	251 bp
tg amplicon:	189 bp

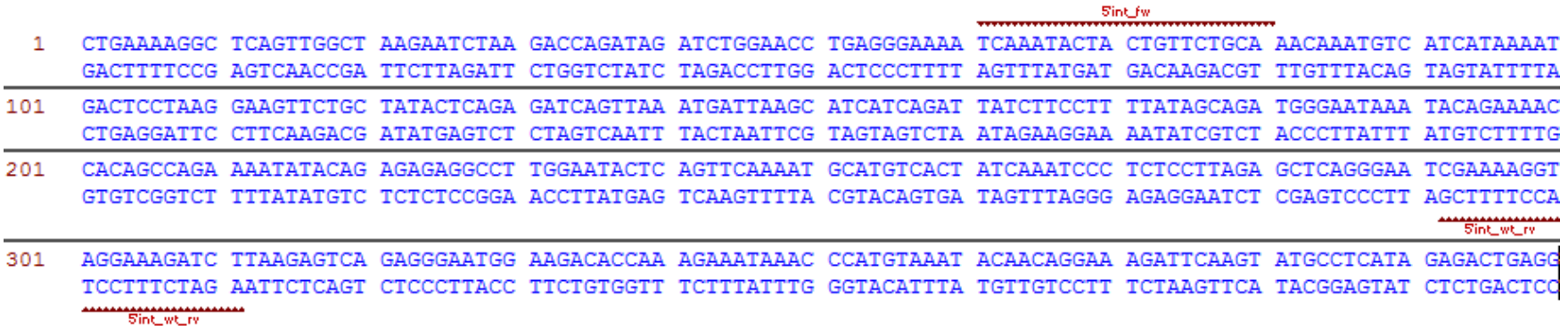
3’ integration site:

3'int_wt_fw	ACAAGGTAATAGTTTCATGC
3'int_tg_fw	CCCTGAGTGTACACCTTGAA
3'int_rv	GGTCTTAAGTACCGGTCTTT
wt amplicon:	344 bp
tg amplicon:	188 bp

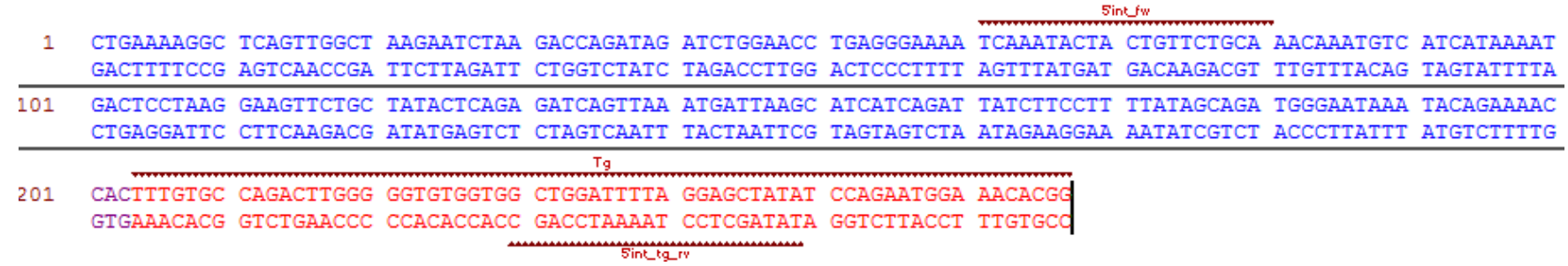
Proposed genotyping assays

5’ integration site:

wt



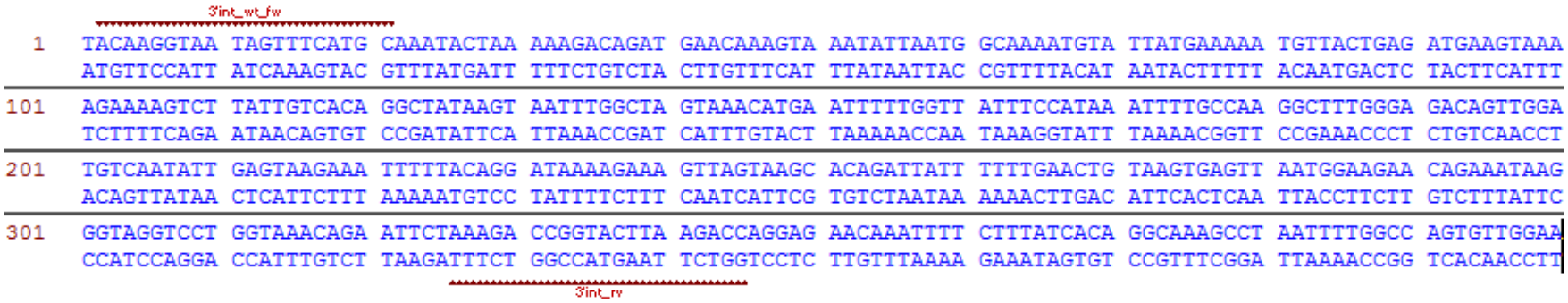
tg



Proposed genotyping assays

3' integration site:

wt



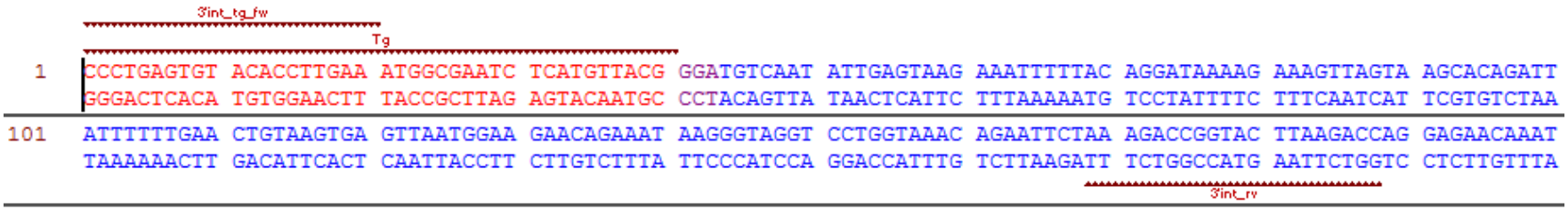

```
      3'int_wt_fw
  1   TACAAGGTAA TAGTTTCATG CAAATACTAA AAAGACAGAT GAACAAAGTA AATATTAATG GCAAAATGTA TTATGAAAAA TGTTACTGAG ATGAAGTAAA
      ATGTTCCATT ATCAAAGTAC GTTTATGATT TTTCTGTCTA CTTGTTTCAT TTATAATTAC CGTTTTACAT AATACTTTTT ACAATGACTC TACTTCATTT

101   AGAAAAGTCT TATTGTCACA GGCTATAAGT AATTTGGCTA GTAAACATGA ATTTTTGGTT ATTTCCATAA ATTTTGCCAA GGCTTTGGGA GACAGTTGGA
      TCTTTTCAGA ATAACAGTGT CCGATATTCA TTAAACCGAT CATTTGTACT TAAAAACCAA TAAAGGTATT TAAAACGGTT CCGAAACCCT CTGTCAACCT

201   TGTCAATATT GAGTAAGAAA TTTTTACAGG ATAAAAGAAA GTTAGTAAGC ACAGATTATT TTTTGAACTG TAAGTGAGTT AATGGAAGAA CAGAAATAAG
      ACAGTTATAA CTCATTCTTT AAAAATGTCC TATTTTCTTT CAATCATTCG TGTCTAATAA AAAACTTGAC ATTCACTCAA TTACCTTCTT GTCTTTATTC

301   GGTAGGTCCT GGTAAACAGA ATTCTAAAGA CCGGTACTTA AGACCAGGAG AACAAATTTT CTTTATCACA GGCAAAGCCT AATTTTGGCC AGTGTTGGAA
      CCATCCAGGA CCATTTGTCT TAAGATTTCT GGCCATGAAT TCTGGTCCTC TTGTTTAAAA GAAATAGTGT CCGTTTCGGA TTAAAACCGG TCACAACCTT
                                                  3'int_rv
```

tg




```
      3'int_tg_fw
                  Tg
  1   CCCTGAGTGT ACACCTTGAA ATGGCGAATC TCATGTTACG GGATGTCAAT ATTGAGTAAG AAATTTTTAC AGGATAAAAG AAAGTTAGTA AGCACAGATT
      GGGACTCACA TGTGGAACTT TACCGCTTAG AGTACAATGC CCTACAGTTA TAACTCATTC TTTAAAAATG TCCTATTTTC TTTCAATCAT TCGTGTCTAA

101   ATTTTTTGAA CTGTAAGTGA GTTAATGGAA GAACAGAAAT AAGGGTAGGT CCTGGTAAAC AGAATTCTAA AGACCGGTAC TTAAGACCAG GAGAACAAAT
      TAAAAAACTT GACATTCACT CAATTACCTT CTTGTCTTTA TTCCCATCCA GGACCATTTG TCTTAAGATT TCTGGCCATG AATTCTGGTC CTCTTGTTTA
                                                                                  3'int_rv

201   TTTCTTTATC ACAGGCAAAG CCTAATTTTG GCCAGTGTTG GAA
      AAAGAAATAG TGTCCGTTTC GGATTAAAAC CGGTCACAAC CTT
```