Title

Bioinformatics analysis
of *Glycine max* transformation event FG72

Critical Confidential Information removed (CCI Version)
Only for use in Australia/New Zealand

Author

Steven Verhaeghe

Completed on

June 28th, 2010

Testing Facility

Bayer BioScience N.V.
BioAnalytics
Molecular Characterization
Technologiepark 38
B-9052 Gent
Belgium

STUDY IDENTIFICATION PAGE

Study initiation date:	June 5th , 2008
Study completion date:	October 27th, 2009
Test Facility Address:	Bayer BioScience N.V. BioAnalytics Technologiepark 38 9052 Gent – Belgium Tel: +32 9-243 04 11 Fax: +32 9-224 06 94
Molecular Characterization Manager:	Dirk Nennstiel Address see Test Facility Tel: +32 9-243 04 39 Fax: +32 9-224 06 94 e-mail: dirk.nennstiel@bayercropscience.com
Study Manager:	Steven Verhaeghe Address see Test Facility Tel: +32 9-243 05 55 Fax: +32 9-224 06 94 e-mail: steven.verhaeghe@bayercropscience.com
Sponsor:	Regulatory Affairs
Sponsor representative:	Russell Essner Global Regulatory Affairs Manager P.O. Box 12014 2 T.W. Alexander Drive Research Triangle Park, NC 27709 – United States Tel: +1 919 549 2171 Fax: +1 919 549 2892 e-mail: russ.essner@bayercropscience.com

ABBREVIATIONS

3’	3-prime
5’	5-prime
aa	amino acids
BLASTx	Basic Local Alignment Search Tool comparing nucleotide sequences translated in the 6 reading frames with protein sequences
bp	base pairs
CDS	CoDing Sequence
DNA	DeoxyNucleic Acid
FGENESH	Find GENES using Hidden markov model
ORF	Open Reading Frame
polyA	poly-Adenylation signal sequence
RBS	Ribosome Binding Site
RNA	RiboNucleic Acid
T-DNA	Transfer or Transgenic DeoxyriboNucleid Acid
TSSP	Transcription Start Site Prediction

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

SUMMARY

Bayer CropScience has introduced a *2mepsps* gene construct, encoding tolerance to glyphosate, and a *hppdPf W336* gene construct, encoding tolerance to isoxaflutole, in *Glycine max* by means of direct gene transfer of soybean cells. *Glycine max* transformation event FG72 contains two partial 3'histonAt sequences in a head to head orientation, followed by 2 complete T-DNA copies arranged in a head to tail orientation. Upon integration of the FG72 insert into the *Glycine max* genome, a genomic region translocated to a new position, which is joined by 158 bases of Ph4a748 promoter sequences (Verhaeghe, 2010a; Verhaeghe, 2010b).

In this study, a bioinformatics analysis was performed on the transgenic locus sequences of soybean event FG72 to assess the presence of potential newly created open reading frames (ORFs) leading to the production of proteins. *Glycine max* wild type sequences were analysed to determine whether regulatory sequences, endogenous soybean genes and/or ORFs were interrupted by the insertion of the transgenic DNA sequences in the transformation event FG72 or by the translocation of genomic sequences.

These bioinformatics analyses, using the current databases and bioinformatics tools, indicate that it is highly unlikely that the predicted newly created ORFs will lead to the expression of proteins. Three putative promoter regions interrupted upon transformation are predicted. No newly created genes or genes interrupted upon transformation are predicted. A prolonged transcript of the second copy of the *2mepsps* gene is predicted.

1. Introduction

Bayer CropScience has introduced a *2mepsps* gene construct, encoding tolerance to glyphosate, and a *hppdPf W336* gene construct, encoding tolerance to isoxaflutole, in *Glycine max* by means of direct gene transfer of soybean cells.

Glycine max transformation event FG72 contains two partial 3'histonAt sequences in a head to head orientation, followed by 2 complete T-DNA copies arranged in a head to tail orientation. Upon integration of the FG72 insert into the *Glycine max* genome, a genomic region translocated to a new position, which is joined by 158 bases of Ph4a748 promoter sequences (Figure 2; Verhaeghe, 2010a; Verhaeghe 2010b).

In this study, a bioinformatics analysis is performed on the transgenic locus sequences of soybean event FG72 to assess the presence of potential newly created open reading frames (ORFs) leading to the production of proteins. *Glycine max* wild type sequences are analysed to determine whether regulatory sequences, endogenous soybean genes and/or ORFs are interrupted by the insertion of the transgenic DNA sequences or by the translocation of genomic sequences.

2. Bioinformatics analysis methods

2.1. Homology search for known functional genes or proteins

The Basic Local Alignment Search Tool (BLAST; Altschul *et al*., 1997) finds regions of local similarity between sequences. BLASTx compares the six-frame theoretical translation products of the nucleotide query sequence (both strands) against a protein sequence database. This analysis is performed on non transgenic *Glycine max* sequences in order to search for homology with known genes.

Appendix 14 describes the databases used in the similarity searches performed in this study. The used databases are all copies or compilations of public databases.

2.2. Gene prediction and open reading frame search

2.2.1. Search for open reading frames

The ORF search was performed using the GetORF search program from the EMBOSS (European Molecular Biology Open Software Suite) tools. In a first analysis, an ORF was defined as a region between two translation stop codons (TAA, TAG, TGA) with a minimum size coding for three amino acids. In a second GetORF analysis, an ORF was defined as a region between a start codon (ATG) and a stop codon (TAA, TAG, TGA) with a minimum size coding for three amino acids.

Only ORFs crossing a junction or insertion point, i.e. newly created ORFs and ORFs interrupted upon transformation, are reported.

2.2.2. Search for potentially expressed genes

FGENESH, used for gene structure prediction (Softberry Inc., version 2.4), predicts exons and introns by statistical sequence analysis and transcription start sites and poly-adenylation signals by homology search with known consensus sequences from plants.

Only genes crossing a junction or insertion point, i.e. newly created genes and genes interrupted upon transformation, are reported. The reported FGENESH prediction results were obtained using the parameters described in Figure 1.

```
FGENESH parameters:
   - Dicot plants (Arabidopsis)
   - Print mRNA
   - Print exons
   - Consider also exons with donor splicing site of GC type
```

Figure 1. Parameters used for reported FGENESH prediction results

2.3. Prediction of regulatory elements

2.3.1. Prediction of core promoter sequences

TSSP is a pattern-finding tool used to search for core promoter (TATA-box) and enhancer sequences listed in the RegSite Database (version 4, Softberry Inc.). Promoters relevant for predicted newly created ORFs containing a start codon, newly created promoters and interrupted promoters are reported.

2.3.2. Prediction of putative ribosome binding sites

A consensus sequence has been determined for the ribosome binding site (RBS) based on a bioinformatics analysis of nucleotide frequencies at positions flanking the translation start codon of dicotyledon and monocotyledon plant genes (Joshi *et al.,* 1997). This sequence (aaaaaaaA(A/C)aATGGCtacta(c/t)ta) has been shown to be important for initiation and efficiency of translation (Gallie *et al.,* 1987). The -3 and +4 positions (where the A of ATG is +1) are considered as the most important in determining a favorable context of initiator ATG.

The ATG context sequences of the newly created ORFs containing a start codon are compared with this RBS consensus sequence.

3. Query sequences

3.1. Soybean event FG72 sequences

To perform the bioinformatics analysis on soybean event FG72 sequences, three fragments were used as query sequences (Figure 2):

- fragment FG72-TR (17806 bp, Appendix 1) containing the inserted DNA with the 5' (fragment d) and 3' flanking sequences (fragment b). This fragment contains 6 junctions:
    - junction 1 between the 5' flanking sequence and the first partial 3'histonAt sequence.
    - junction 2 between 2 partial 3'histonAt sequences.
    - junction 3* between the second partial 3'histonAt sequence and the first complete T-DNA copy.
    - junction 4* between the 2 complete T-DNA copies.
    - junction 5 between the second complete T-DNA copy and the filler DNA.
    - junction 6 between the filler DNA and the 3' flanking sequence.

- fragment FG72-TL1 (2217 bp, Appendix 2) containing the 5' end of the translocated region (fragment c) and the sequence flanking this region (fragment b). This fragment contains 1 junction:

* The bases in junctions 3 and 4 consist both of 6 basepairs that can be originating both from pSF10 plasmid sequences downstream of 3'histonAt and from pSF10 plasmid sequences upstream of 3' nos.

- junction 7 between the 5’ end of the translocated region and its flanking sequence.

- fragment FG72-TL2 (2439 bp, Appendix 3) containing the 3’ end of the translocated region (fragment c), 158 bp of Ph4a748 promoter sequence and the sequence flanking this promoter sequence (fragment a). This fragment contains 2 junctions:
  - junction 8 between the 3’ end of the translocated region and Ph4a748 promoter sequence.
  - junction 9 between Ph4a748 promoter sequences and its flanking sequence.

3.2. Non transgenic Glycine max sequences

The sequences of three *Glycine max* wild type regions, interrupted upon transformation, were used as query sequences (Figure 2):

- fragment JACK-WT1 (2303 bp, Appendix 4) comprising the 5’ integration site of the FG72 insert. This fragment contains 1 insertion point:
  - insertion point 1 between the 5’ pre insertion locus sequence and the 5’ end of the translocating region.

- fragment JACK-WT2 (2991 bp, Appendix 5) comprising the 3’ integration site of the FG72 insert. This fragment contains 2 insertion points:
  - insertion point 2 between the 3’ end of the translocating region and bases deleted upon transformation.
  - insertion point 3 between bases deleted upon transformation and the 3’ pre insertion locus sequence.

- fragment JACK-WT3 (2212 bp, Appendix 6) comprising the reintegration site of the translocating sequences. This fragment contains 2 insertion points:
  - insertion point 4 between the 5’ flanking sequence of the reintegration site and bases deleted upon transformation.
  - insertion point 5 between bases deleted upon transformation and the 3’ flanking sequence of the reintegration site.

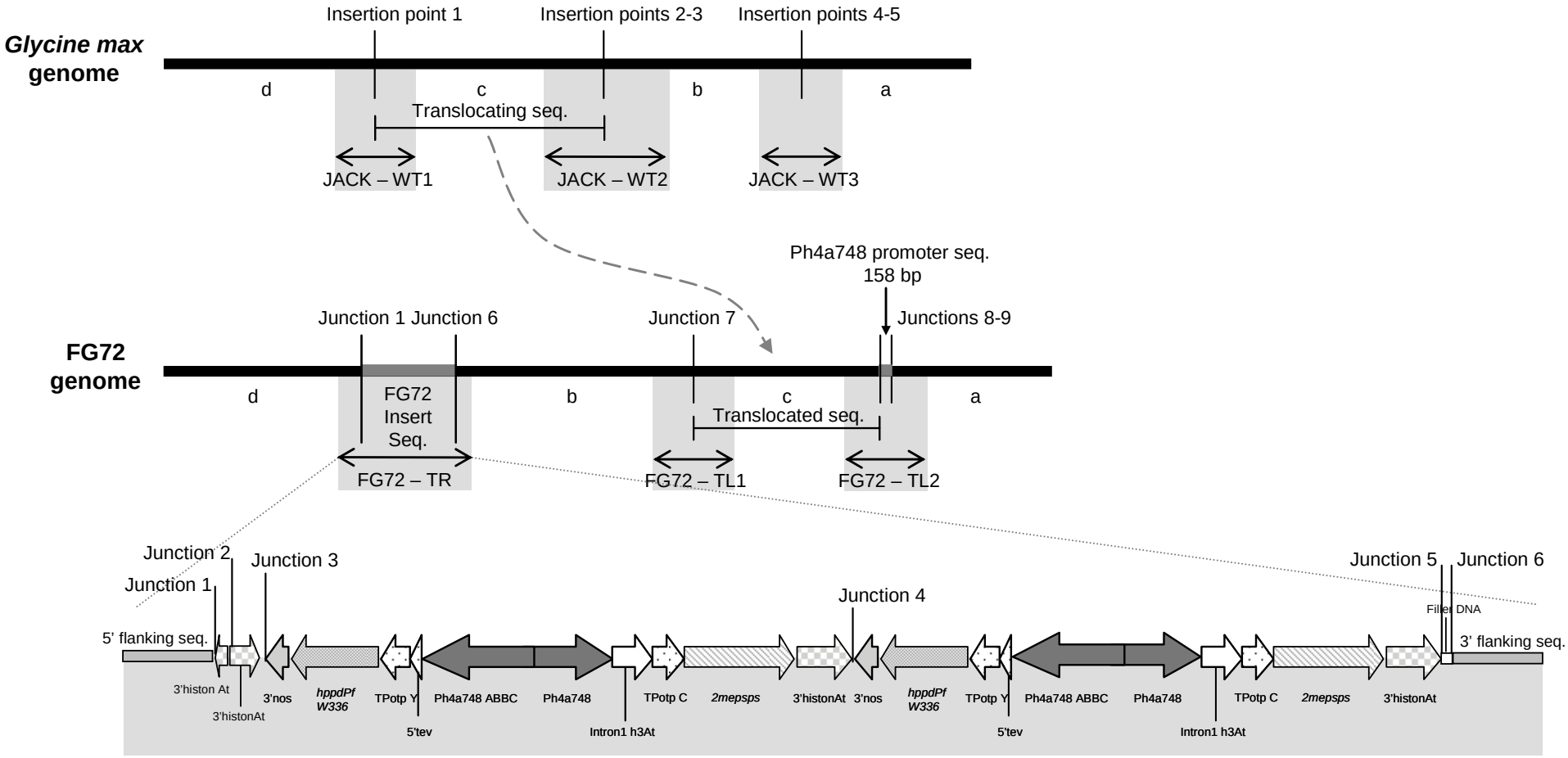

Figure 2. Schematic overview of the analysed FG72 transgenic and non-transgenic sequences

4. Results and discussion

4.1. Soybean event FG72 sequences

All results of the performed bioinformatics analysis on FG72 sequences are depicted schematically in Figure 3 to Figure 7.

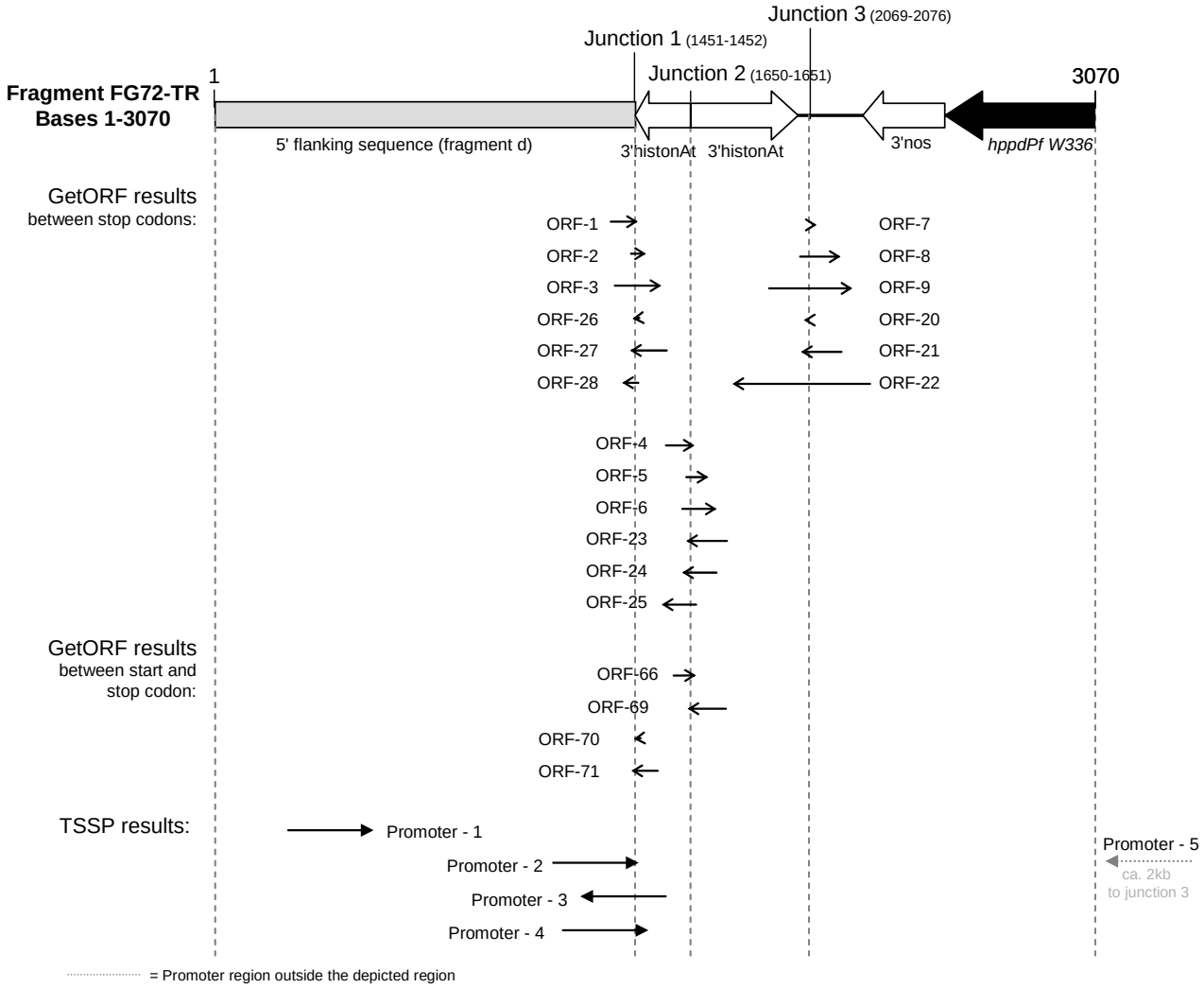

Figure 3. Schematic overview of junctions 1, 2 and 3 (fragment FG72-TR) with indication of bioinformatics analysis results

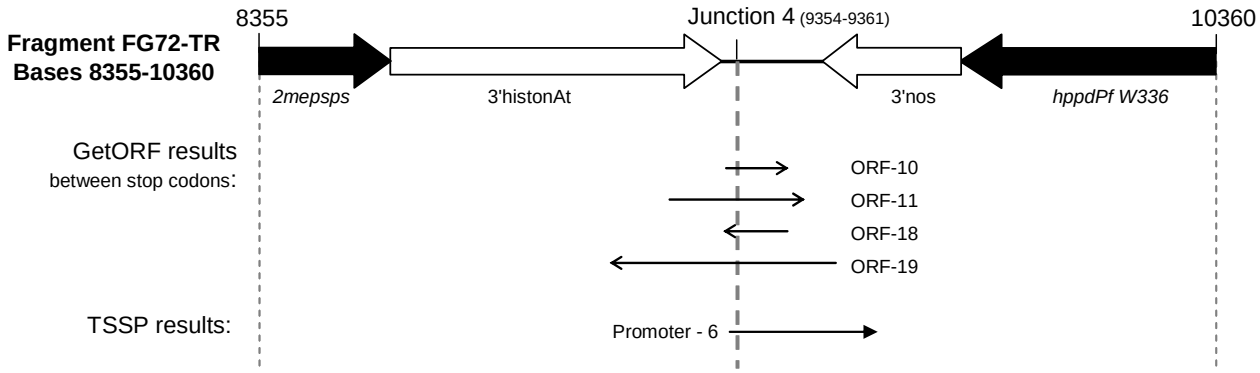

Figure 4. Schematic overview of junction 4 (fragment FG72-TR) with indication of bioinformatics analysis results

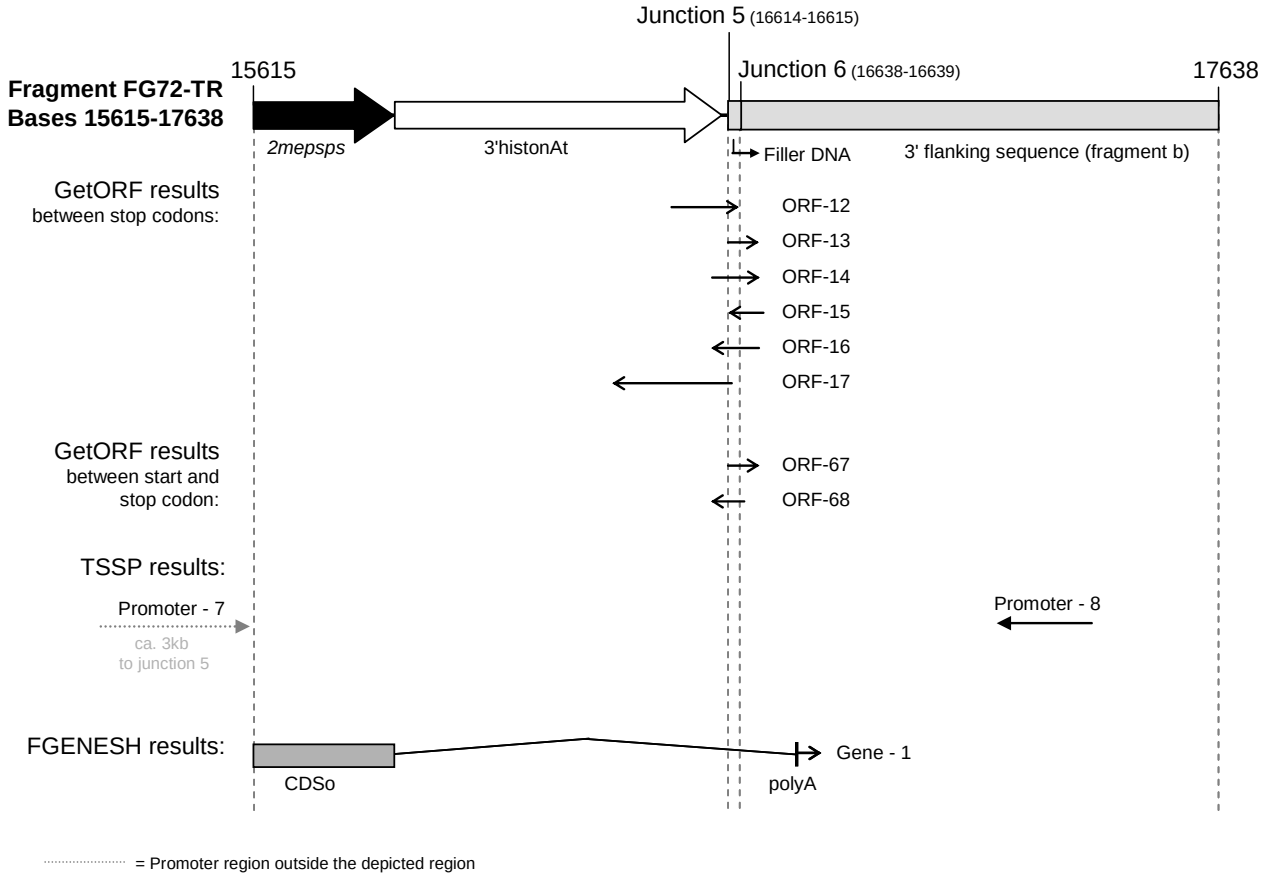

Figure 5. Schematic overview of junctions 5 and 6 (fragment FG72-TR) with indication of bioinformatics analysis results

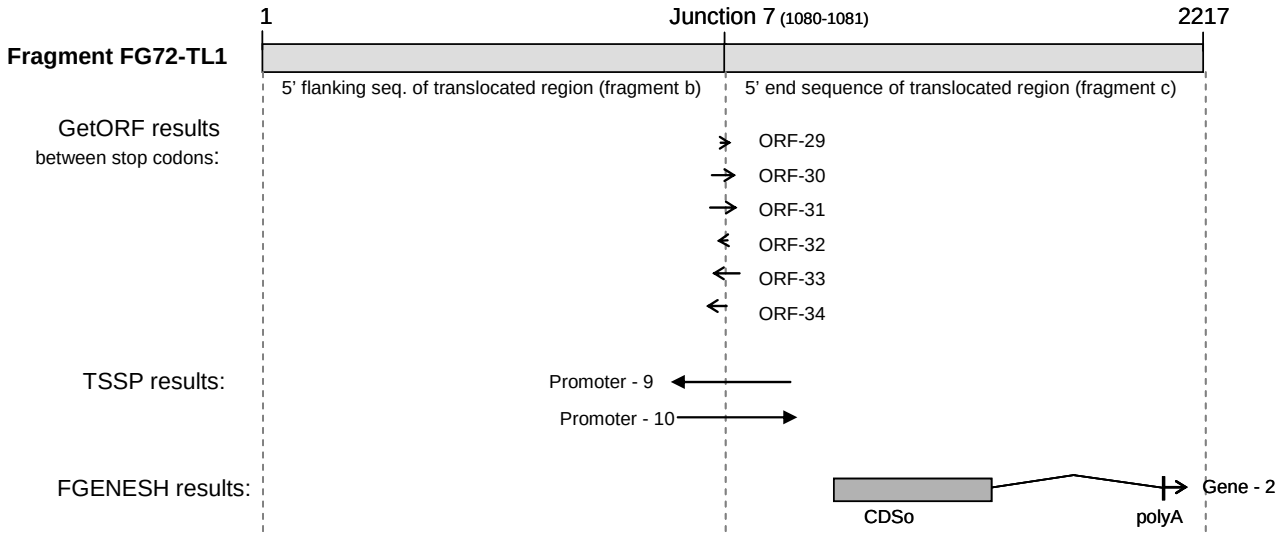

Figure 6. Schematic overview of junction 7 (fragment FG72-TL1) with indication of bioinformatics analysis results

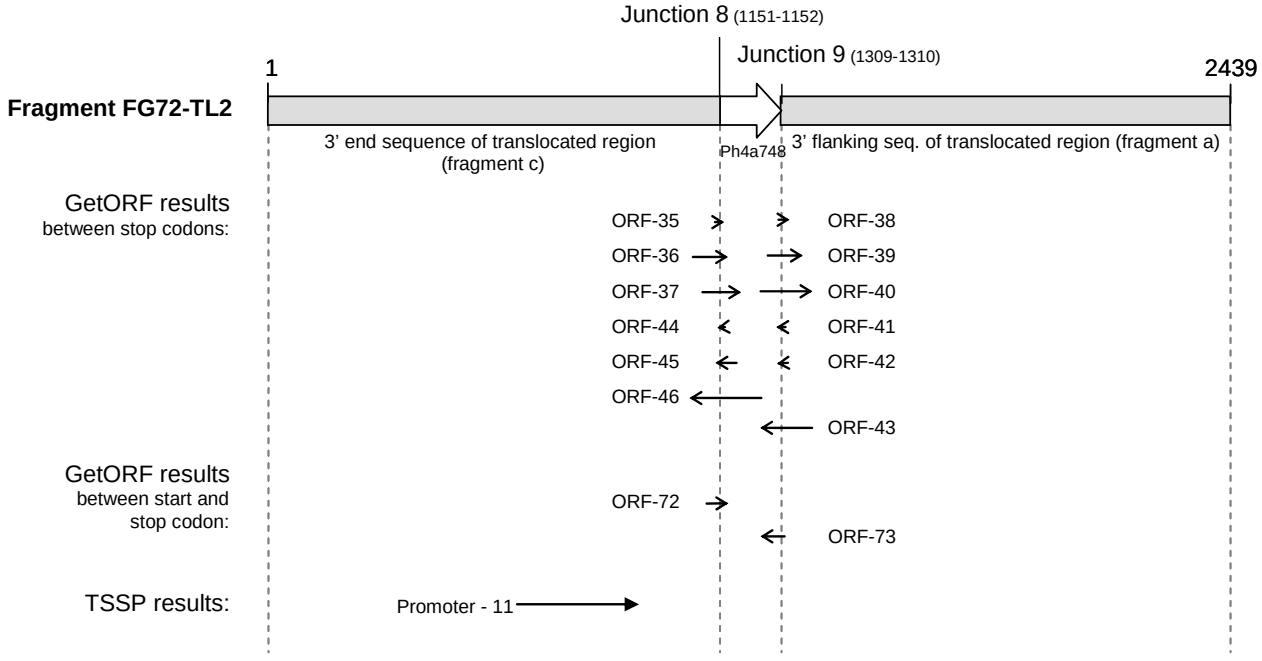

Figure 7. Schematic overview of junctions 8 and 9 (fragment FG72-TL2) with indication of bioinformatics analysis results

GetORF identified 46 newly created ORFs defined between two stop codons (ORF-1 till ORF-46, Appendix 7) and 8 newly created ORFs defined between a start and a stop codon (ORF-66 till ORF-73, Appendix 8).

The bases comprised in junction regions 3 and 4 (*i.e.* 6bp) can be originating both from sequences downstream of 3'histonAt and from sequences upstream of 3'nos of the transforming plasmid pSF10. As a result, ORFs starting or ending in these junction regions are not newly created and only ORFs spanning the complete junction are newly created ORFs. Therefore, 4 predicted ORFs, defined between two stop codons and starting or ending in junction region 4, are not newly created and are not reported.

FGENESH predicted 5 genes:

- Two genes corresponding to the *hppdPf W336* genes, both preceded by TPotp Y. FGENESH results are not shown.
- Two genes corresponding to the *2mepsps* genes, both preceded by TPotp C. FGENESH results of the first copy of the *2mepsps* gene are not shown. The second copy of the *2mepsps* gene was predicted with a poly-adenylation signal in the 3' flanking sequences, which would lead to a prolonged transcript spanning junctions 5 and 6 (Gene–1, Figure 5). Appendix 11 shows the predicted mRNA, exon and protein sequence of this gene.
- One gene (Gene-2, Figure 6) corresponding to a putative cysteine protease (cfr. BLASTx result on the non transgenic Glycine max sequences, section 4.2, Appendix 15). This gene is positioned in the 5' end sequences of the translocated region and is not newly created. Appendix 12 shows the predicted mRNA, exon and protein sequence of this gene.

The ATG context sequence of the newly created ORFs containing a start codon (i.e. ORFs 66 till 73) were compared with the RBS consensus sequence. In Table 1, the most important nucleotides of the RBS consensus sequence are shown in capital letters (Joshi *et al.*, 1997); nucleotides similar to the consensus sequence are marked in bold.

Most of the essential nucleotides are absent for ORFs 66 till 73, indicating that translation of these ORFs is very unlikely.

Table 1: Overview of the ATG context analysis results

Consensus sequence	aaaaaaaAMa	ATG	GCtactayta
ORF-66	Critical Confidential Information removed		
ORF-67			
ORF-68			
ORF-69			
ORF-70			
ORF-71			
ORF-72			
ORF-73			

(M = A or C and Y = C or T)

Throughout the FG72 sequences, 11 relevant promoter regions were predicted by TSSP (Table 2):

- Promoter regions 2, 3, 4, 6, 9 and 10 are spanning a junction and are newly created.
- Promoter regions 1, 2, 4, 5, 7, 8 and 11 are in front of one or more predicted newly created ORFs containing a start codon (Table 2). Since the ATG context analysis indicated that translation of these ORFs is very unlikely, it is unlikely that these promoters are biologically active.

Table 2. Overview of predicted promoter regions in FG72 sequences

Fragment	Promoter region	Position in fragment (bp)	Spanning junction	Promoter in front of newly created ORF(s)
FG72-TR	Promoter-1	301 → 600	/	ORF-66
	Promoter-2	1166 → 1464	Junction 1	ORF-66
	Promoter-3	1260 ← 1558	Junction 1	/
	Promoter-4	1197 → 1488	Junction 1	ORF-66
	Promoter-5	4158 ← 4434	/	ORF-69, ORF-70, ORF-71
	Promoter-6	9348 → 9647	Junction 4	/
	Promoter-7	13315 → 13615	/	ORF-67
	Promoter-8	17192 ← 17459	/	ORF-68
FG72-TL1	Promoter-9	965 ← 1248	Junction 7	/
	Promoter-10	970 → 1264	Junction 7	/
FG72-TL2	Promoter-11	679 → 975	/	ORF-72

4.2. Non transgenic Glycine max sequences

All results of the performed bioinformatics analysis on non transgenic *Glycine max* sequences are depicted schematically in Figure 8 to Figure 10.

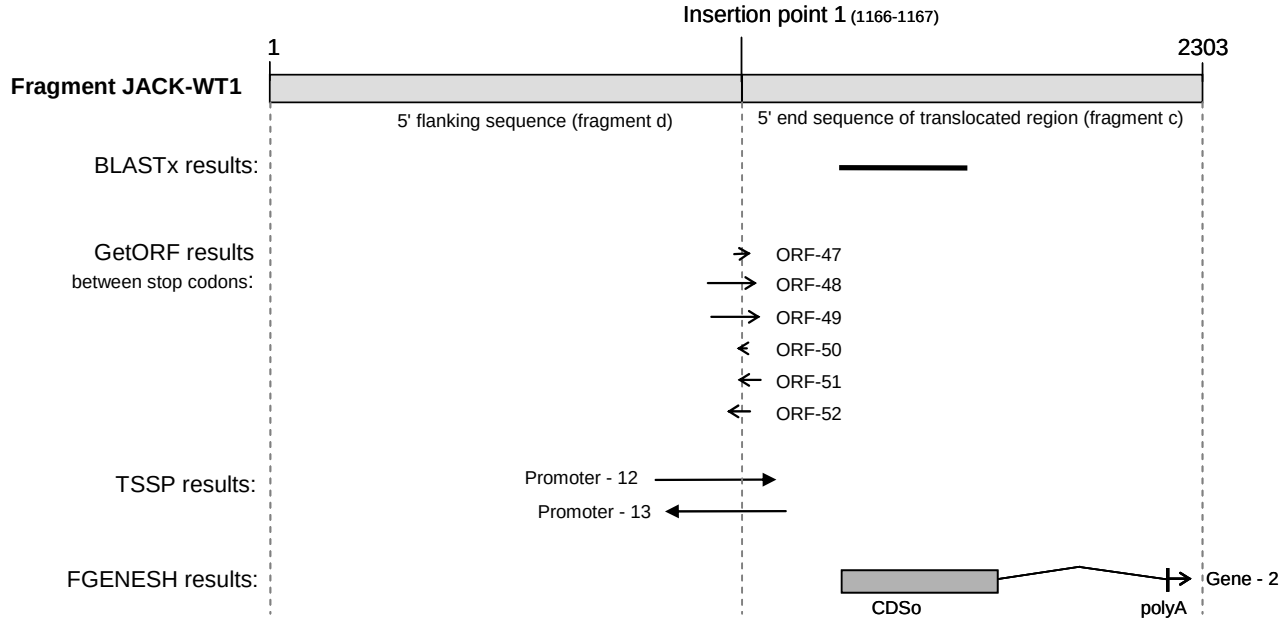

Figure 8. Schematic overview of junction 10 (fragment JACK-WT1) with indication of bioinformatics analysis results

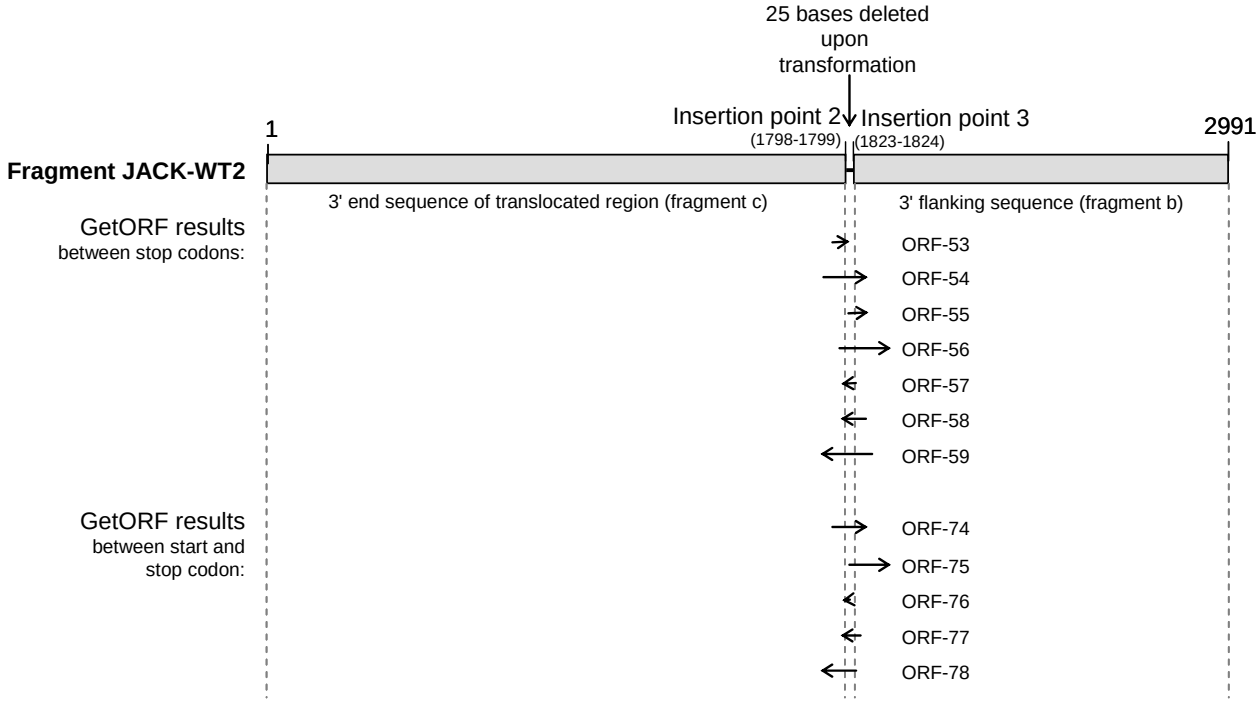

Figure 9. Schematic overview of junctions 11 and 12 (fragment JACK-WT2) with indication of bioinformatics analysis results

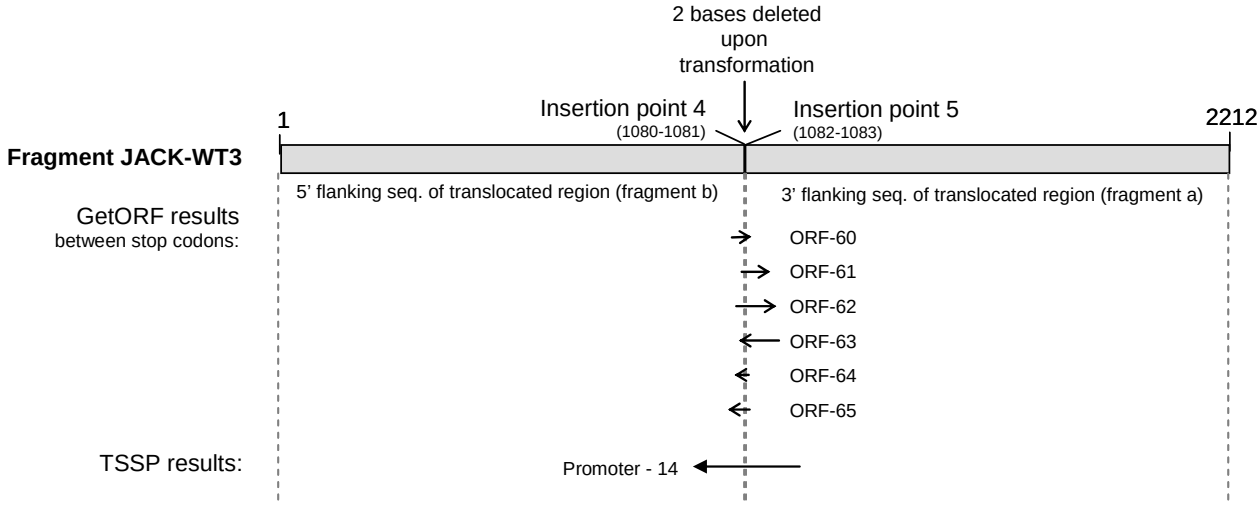

Figure 10. Schematic overview of junctions 13 and 14 (fragment JACK-WT3) with indication of bioinformatics analysis results

BLASTx homology search identified homology of the 5' end sequence of the translocated region with part of a putative cysteine protease (Figure 8). Details on this BLASTx hit using the Uniprot database are given in Appendix 15. Similar results are obtained using the Dad and Genpept databases (results not shown). This protein is not interrupted upon transformation. No other relevant known functional genes, interrupted upon transformation, could be identified using BLASTx.

GetORF identified 19 interrupted ORFs defined between two stop codons (ORF-47 till ORF-65, Appendix 9) and 5 interrupted ORFs defined between a start and a stop codon (ORF-74 till ORF-78, Appendix 10).

No genes crossing the insertion points were predicted by FGENESH. One gene is predicted (Gene-2, Figure 8) in the 5' end sequence of the translocated region. This gene corresponds to the BLASTx result on the non transgenic *Glycine max* sequences is 100% identical to the Gene-2 predicted in the FG72 sequences. This predicted Gene-2 is not newly created. Appendix 13 shows the predicted mRNA, exon and protein sequence of this gene.

Three promoter regions spanning an insertion point are predicted throughout the non transgenic sequences (Table 3). These promoter regions are interrupted upon transformation.

Table 3. Overview of predicted promoter regions in non transgenic *Glycine max* sequences

Fragment	Promoter region	Spanning insertion point	Position in fragment (bp)
JACK-WT1	Promoter-12	Insertion point 1	955 → 1251
JACK-WT1	Promoter-13	Insertion point 1	975 ← 1275
JACK-WT3	Promoter-14	Insertion points 4-5	965 ← 1209

5. Conclusion

A bioinformatics analysis was performed on the transgenic locus of FG72 to detect open reading frames (ORFs) and regulatory elements, created by the insertion of the transgenic DNA sequences in the transformation event FG72 or by the translocation of genomic sequences. Several *in silico* tools were used taking into account the current scientific knowledge on gene expression.

These bioinformatics analyses, using the current databases and bioinformatics tools, indicate that it is highly unlikely that the predicted newly created ORFs will lead to the expression of newly created proteins. Three putative promoter regions interrupted upon transformation are predicted. No newly created genes or genes interrupted upon transformation are predicted. A prolonged transcript of the second copy of the *2mepsps* gene is predicted.

6. References

N°	DART N°	Report N°	Author(s), year, title, source, edition, pages
1.	M-222480-01-1	-----	Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. - 1997 - Gapped BLAST and PSI-BLAST: a new generation of protein database search programs - Nucleic Acids Res. vol. 25, pages 3389-3402.
2.	M-222482-01-1	-----	Gallie D.R., Sleat D.E., Watts J.W., Turner P.C. and Wilson T.M.A. (1987) A comparison of eukaryotic viral 5'-leader sequences as enhancers of mRNA expression *in vitro.* Nucleic Acids Res. vol. 15, pages 8693-8711.
3.	M-222486-01-1	-----	Joshi C.P., Zhou H., Huang X. and Chiang V.L. (1997) Context sequences of translation initiation codon in plants. Plant Mol. Biol. vol. 35, pages 993 – 1001.
4.			Verhaeghe S. (2010a) Detailed insert characterization of *Glycine max* transformation event FG72 by Southern blot analysis.
5.			Verhaeghe S. (2010b) Full DNA sequence of event insert and integration site of *Glycine max* transformation event FG72.

APPENDICES

Appendix 1. Nucleotide sequence of fragment FG72-TR

Critical Confidential Information removed

Feature	Position	Corresponding position in pSF10
5' flanking sequence (fragment d)	bp 1 → bp 1451	/
T-DNA sequences	bp 1650 → bp 1452 bp 1651 → bp 2070 bp 2071 → bp 9360 bp 9355 → bp 16614	bp 9834 → bp 10032 bp 9948 → bp 10372 bp 3075 → bp 10365 bp 3080 → bp 10339
3'histonAt	bp 1650 → bp 1452	bp 9834 → bp 10032
3'histonAt	bp 1651 → bp 2029	bp 9948 → bp 10326
3'nos	bp 2548 → bp 2257	bp 3553 → bp 3262
hppdPf W336	bp 3625 → bp 2549	bp 4630 → bp 3554
TPotp Y	bp 3997 → bp 3626	bp 5002 → bp 4631
5'tev	bp 4138 → bp 3998	bp 5143 → bp 5003
Ph4a748ABBC	bp 5428 → bp 4139	bp 6433 → bp 5144
Ph4a748	bp 5429 → bp 6443	bp 6434 → bp 7448
Intron1 h3At	bp 6444 → bp 6924	bp 7449 → bp 7929
TPotp C	bp 6925 → bp 7296	bp 7930 → bp 8301
2mepsps	bp 7297 → bp 8634	bp 8302 → bp 9639
3'histonAt	bp 8635 → bp 9321	bp 9640 → bp 10326
3'nos	bp 9828 → bp 9537	bp 3553 → bp 3262
hppdPf W336	bp 10905 → bp 9829	bp 4630 → bp 3554
TPotp Y	bp 11277 → bp 10906	bp 5002 → bp 4631
5'tev	bp 11418 → bp 11278	bp 5143 → bp 5003
Ph4a748ABBC	bp 12708 → bp 11419	bp 6433 → bp 5144
Ph4a748	bp 12709 → bp 13723	bp 6434 → bp 7448
Intron1 h3At	bp 13724 → bp 14204	bp 7449 → bp 7929
TPotp C	bp 14205 → bp 14576	bp 7930 → bp 8301
2mepsps	bp 14577 → bp 15914	bp 8302 → bp 9639
3'histonAt	bp 15915 → bp 16601	bp 9640 → bp 10326
Filler DNA	bp 16615 → bp 16638	/
3' flanking sequence (fragment b)	bp 16639 → bp 17806	/
Junction 1	bp 1451 → bp 1452	/
Junction 2	bp 1650 → bp 1651	/
Junction 3	bp 2069 → bp 2076	/
Junction 4	bp 9354 → bp 9361	/
Junction 5	bp 16614 → bp 16615	/
Junction 6	bp 16638 → bp 16639	/

Appendix 2. Nucleotide sequence of fragment FG72-TL1

Critical Confidential Information removed

Feature	Position
5' flanking sequence of the translocated region (fragment b)	bp 1 → bp 1080
5' end sequence of the translocated region (fragment c)	bp 1081 → bp 2217
Junction 7	bp 1080 → 1081

Appendix 3. Nucleotide sequence of fragment FG72-TL2

Critical Confidential Information removed

Feature	Position
3' end sequence of the translocated region (fragment c)	bp 1 → bp 1151
Ph4a748 sequence	bp 1152 → bp 1309
3' flanking sequence of the translocated region (fragment a)	bp 1310 → bp 2439
Junction 8	bp 1151 → bp 1152
Junction 9	bp 1309 → bp 1310

Appendix 4. Nucleotide sequence of fragment JACK-WT1

Critical Confidential Information removed

Feature	Position
5' flanking sequence of FG72 (fragment d)	bp 1 → bp 1166
5' end of the translocated sequence (fragment c)	bp 1167 → bp 2303
Insertion point 1	bp 1166 → 1167

Appendix 5. Nucleotide sequence of fragment JACK-WT2

Critical Confidential Information removed

Feature	Position
3' end of the translocated sequence (fragment c)	bp 1 → bp 1798
Bases deleted upon transformation	bp 1799 → bp 1823
3' flanking sequence of FG72 (fragment b)	bp 1824 → bp 2991
Insertion point 2	bp 1798 → bp 1799
Insertion point 3	bp 1823 → bp 1824

Appendix 6. Nucleotide sequence of fragment JACK-WT3

Critical Confidential Information removed

Feature	Position
5' flanking sequence of the translocated region (fragment b)	bp 1 → bp 1080
Bases deleted upon transformation	bp 1081 → bp 1082
3' flanking sequence of the translocated region (fragment a)	bp 1083 → bp 2212
Insertion point 4	bp 1080 → bp 1081
Insertion point 5	bp 1082 → bp 1083

Appendix 7. FG72 sequences - GetORF prediction results (ORF defined between two stop codons, 3aa)

Critical Confidential Information removed

Appendix 8. FG72 sequences - GetORF prediction results (ORF defined between a start and a stop codon, 3aa)

Critical Confidential Information removed

Appendix 9. Non transgenic sequences - GetORF prediction results (ORF defined between two stop codons, 3aa)

Critical Confidential Information removed

Appendix 10. Non transgenic sequences - GetORF prediction results (ORF defined between a start and a stop codon, 3aa)

Critical Confidential Information removed

Appendix 11. FGENESH results of Gene-1

Critical Confidential Information removed

Appendix 12. FGENESH results of Gene-2 on FG72 sequences

Critical Confidential Information removed

Appendix 13. FGENESH results of Gene-2 on non transgenic sequences

Critical Confidential Information removed

Appendix 14. Database definitions when using BLASTx

Database	Posted date of database	Date of analysis	Number of letters in database	Number of sequences in database
Dad	Apr 9, 2009	May 14, 2009	999,999,968	3,149,053
Genpept	Apr 9, 2009	May 14, 2009	999,999,578	3,282,066
Uniprot	Apr 8, 2009	May 14, 2009	999,999,356	2,681,453

Parameters used for analysis:
- Matrix: BLOSUM62
- Gapped Alignment: Yes
- Expect: 1
- Filter: none
- Descriptions: 100
- Alignments: 100

Dad

Dad (DDBJ Aminoacid sequence Database) is a protein database translated from the DDBJ (DNA Data Bank of Japan) which is officially certified to collect DNA sequences from researchers and to issue the internationally recognized accession number to data submitters. This database exchanges the collected data with EMBL/EBI and GenBank/NCBI on a daily basis.

Uniprot

UniProt (Universal Protein Resource) is the world's most comprehensive catalog of information on proteins. It is a central repository of amino acid sequence and function created by joining the information contained in Swiss-Prot, TrEMBL and PIR. UniProt has three components, each optimized for different uses. The UniProt Knowledgebase (UniProtKB) is the central access point for extensive curated protein information, including function, classification and cross-reference.

Genpept

GenPept is a genetic sequence databank which contains translated protein-coding sequences. Previous releases of GenPept were produced by translating the GenBank flat file release. Beginning with Release 70, GenBank entries include a translation of each valid CDS. This information is associated with each CDS through the addition of a translation qualifier in the GenBank Feature Table. The GenPept amino acid sequence is simply copied from the value of the translation qualifier.

Appendix 15. Results of BLASTx analysis

Critical Confidential Information removed