GM plants compared to the baseline; a whole genome sequencing approach

CGM 2015-07

ONDERZOEKSRAPPORT

GM plants compared to the baseline; a whole genome sequencing approach

Henk J. Schouten, Elio Schijlen, Jan Schaart, Henri van de Geest, Sofia Papadimitriou, M.J.M. Smulders, Richard Finkers, Gabino Sanchez Perez

Wageningen University and Research Centre

This report was commissioned by COGEM. The content of this publication is the sole responsibility of the author and does not necessarily reflect the views of COGEM.

Dit rapport is in opdracht van de Commissie Genetische Modificatie (COGEM) samengesteld. De mening die in het rapport wordt weergegeven is die van de auteur en weerspiegelt niet noodzakelijkerwijs de mening van de COGEM.

Met dank aan dr. ing. J. Schaart, Wageningen Universiteit en Research Centrum, voor de foto van tomatenplanten op de omslag

Content

Preface

For the environmental risk assessment of market applications of genetically modified (GM) crops several studies are submitted by the applicant. One of these studies involves the molecular characterisation of the introduced changes, using techniques such as Southern blot, PCR, and sequence analysis of the inserted sequence and flanking regions by means of genomic walking.

With the introduction of novel methodologies as next generation sequencing (NGS), (re-)sequencing of complete plant genomes is becoming available at relatively low costs. This provides a new tool for the molecular characterisation of GM crops, as genomic sequences of GM plants and their non-GM conventional counterpart can be compared in detail.

COGEM has commissioned a research project to address the question what the value of NGS could be for the risk assessment of GM plants. The first main question was whether genomic differences observed between a GM plant and its non-GM counterpart are due to the genetic modification process, which does not only involve the insertion of foreign DNA but also specific culturing steps. The second main question was how the spectrum of induced changes during the genetic modification process compares to the natural baseline variation in the species. Finally, a third, more practical question was if NGS could replace the elaborate screening of material for (multiple) inserts and their location by means of Southerns and PCR.

The present report addresses these questions in two ways. In the first section, the natural genomic variation of several plant species is reviewed in a desk study. The second part employs an experimental approach, in which genetically modified *Arabidopsis thaliana* and tomato were molecularly characterised by NGS, screening for new mutations across the entire genome using bio-informatics tools.

In Arabidopsis, the mutation frequency introduced by genetic modification using flower dip was very low and did not differ significantly from a previous report. Small deletions were predominant in the transformants (6 out of 8 mutations in total). Also a small 'splinter' fragment of the insert was detected.
The tomato plants were modified using *Agrobacterium*. Much higher mutation frequencies were found compared to the Arabidopsis floral dip experiment. Furthermore, the number of mutations varied among replicate transformant lines, which is presumably due to the method of *in vitro* regeneration rather than the genetic modification per se. No evidence for clustering of mutations in certain genomic areas was found.
The desk study further revealed that the genetic variation induced by genetic modification comprises a small fraction in comparison with the baseline genomic variation that is available for conventional plant breeding. This holds for both Arabidopsis and tomato.

Furthermore, the experimental part shows that NGS can be a suitable tool for the molecular characterisation of GM plants, provided that the sequencing is sufficiently deep, a good reference genome of the species is available, and the sequences of putative transgenic inserts and genomic borders are confirmed by conventional directed sequencing.

This report is of assistance for dealing with the risk assessment of GM crops which are molecular characterised with new sequence methodologies. The main conclusion is that the results do not indicate that new elements should be added to the present risk assessment procedure, and that NGS is an alternative method for the screening of the GM plant on presence and location of the inserted DNA.

Prof. dr. P.H. van Tienderen, Chairman Advisory Committee

ADVISORY COMMITTEE

Prof. dr. P.H. van Tienderen, member of COGEM
Dr. J.M. Kooter, Vrije Universiteit Amsterdam, member of COGEM
Dr. D.C.M. Glandorf, GMO office, National Institute of Public Health and the Environment
A. T. A. Box, B. ASc., COGEM Secretariat

Summary

The goals of the current project were:

1. Description of the baseline of natural variation in genomes of plants, used in conventional plant breeding. Different sources of variation in conventional breeding were distinguished, i.e.
   a. Variation in DNA sequences of wild germplasm, old cultivars, land races and current cultivars. This is variation due to accumulation of mutations as a result of evolution, and human-mediated selection. This variation is represented in the breeding material of the crop;
   b. Mutations that do occur spontaneously during propagation and breeding of the crop. This variation is due to natural mutations in vegetatively and generatively propagated plant material. We mean here the current dynamics of plant genomes, on top of the chromosomal crossovers that occur during meiosis;
2. Evaluation of changes in the genome due to tissue culture and regeneration (= somaclonal variation);
3. Evaluation of changes in the genome due to transformation;
4. Comparison of changes in the genome due to transformation (goal 3) with the baseline of natural variation (goal 1) and somaclonal variation (goal 2);
5. Discussion of additional biosafety questions that may arise when whole genome sequences of GM plants can be obtained;
6. Suggestions for using the whole genome sequence data of GM plants for the environmental risk assessment of these plants. Here, the more practical question was addressed whether the elaborate screening of GM plants for inserts, and sequences of DNA flanking these inserts, could be performed by means of next generation sequencing (NGS), instead of Southerns and genome walking kits.

Both existing literature and new experimental data are presented.

The genetic variation in *Arabidopsis thaliana*, rice and tomato.

<u>*A. thaliana*</u>. Although *A. thaliana* is not grown as a commercial crop, it is the most widely studied plant species for unravelling genetics. A wealth of information on genetic variation is publically available for this species. Therefore we included this species in this study. Cao et al. (2011) re-sequenced 80 strains of *A. thaliana* representing the genetic diversity present in eight populations across the native range of the species in Eurasia, spanning various climates and elevations. They identified nearly 5 million (4,902,039) single nucleotide polymorphisms (SNPs) across the 80 strains. This represents, on the average, one SNP per 23 bp, taking all 80 strains into account. Most SNPs were not restricted to one strain only, but were found in at least two strains. More than 800,000 (810,467) small insertions/deletions (indel 1-20 bp) were detected in the 80 accessions, which is on the average one small indel per 140 bp. They detected at least 174,789 structural variants, of which 49% were detected in more than one strain. In the reference genome of *A. thaliana*, 31,189 transposable element insertions have been annotated. From these insertions, 80% showed evidence of being partially or completely absent from the genome of at least one of the 80 sequenced strains. This underlines the variability of these elements. Cao et al. discovered SNPs in more than 6,000 (6,197) genes that altered start codons, introduced premature stop codons, extended the open reading frame of

the reference sequence, or affected splice donor or acceptor sites. These were named ‘drastic mutations’.

Rice. Xu et al. (2011) sequenced 40 cultivated accessions from the major groups of rice (*Oryza sativa*) and 10 accessions of their wild progenitors (*O. rufipogon* and *O. nivara*). They obtained 6.5 million SNPs, which is about 1.6 million more than Cao et al. (2011) detected in the 80 wild *A. thaliana* accessions. Huang et al. (2012) performed a similar study, but increased the number of accessions considerably from 40 to 1,529 accessions, yielding nearly 8 million (7,970,359) non-singleton SNPs.

In addition to ‘copying errors’ in the DNA during cell divisions leading to SNPs and small indels, and unequal crossing-overs during meiosis leading to large insertions or deletions, another class of mutations is caused by activity of transposable elements. Transposable elements constitute a large portion of eukaryotic genomes. Jiang et al. (2003) showed activity of a DNA transposon called *miniature Ping* (*mPing*), and of another family of transposable elements called *Pong,* in rice. Kikuchi et al. (2003) observed the efficient excision of *mPing* and reinsertion into new loci in the rice genome. Naito et al. (2006) showed bursts of the *mPing* elements from ~50 to ~1000 copies in rice genomes. In spite of these bursts, these transposable elements did not kill their hosts, and appeared to be significantly underrepresented in exons and introns.

Tomato. Wageningen UR coordinated a re-sequencing initiative in tomato. 84 tomato genotypes were selected, including 11 varieties, 43 land races and 30 wild species, representing different tomato types, such as cherry, beef, round, pink and heirloom types. All accessions can be crossed with cultivated tomato. The data presented here have been published by Aflitos et al. (2014), but also additional analyses are reported. The SNP counts for tomato cultivars (*S. lycopersicum* and *S. lycopersicum* var. *cerasiforme*) were relatively low (between 200K and 4.5M; 850K on average; 1 SNP / 800 bp on average), when compared to the reference genome of *S. lycopersicum* Heinz 1706. For members of the wild species, not belonging to the species *S. lycopersicum*, SNP numbers increased sharply, to 8 to 10 Million SNPs per accessions (approximately 1 SNP / 80 bp). We consistently observed in all accessions a significant higher SNP frequency in intergenic regions compared to genic regions. Approximately, 89.5% of the SNPs were detected in intergenic regions. In genic regions the SNP frequency was highest in the non-coding regions: 7.5% mapped in introns. The 5’ and 3’ UTRs showed more polymorphisms than the internal introns, which in turn had a higher SNP frequency compared to exons. From the polymorphisms in exons, 32.4% appeared to be synonymous while 40% was non-synonymous. The remaining of the polymorphisms resulted in different effects, such as a frame shift, loss or gain of a start or stop codon, etc.

We also detected on chromosome 9 in six *S. lycopersicum* accessions an introgression from a wild species. This finding is in agreement with the known presence of alleles for a gene conferring resistance to Tobacco Mosaic Virus (TMV). However, the large size of the introgressed chromosomal part was a surprise to many breeding scientists involved in the project. Apparently, a very large part (~75%) of chromosome 9 has been co-introgressed with the resistance gene on this chromosome. From previous research it appeared that the fragment, containing the TMV locus is actually inverted. The consequence of this is that this fragment does not recombine anymore after its initial introgression. This example clearly illustrates the increase of the genetic variation in commercial cultivars as a result of introgression breeding.

The variation discussed thus far was identified on the basis of the tomato reference genome of *S. lycopersicum* cv Heinz 1706. However, the genetic differences between the tomato wild species and the reference genome are much larger than described above, because of 1) the differences in genome size. Genome size estimates of *S. lycopersicum* and *S. pennellii* are 0.95 and 1.23 pg respectively. Sequences that do not exist in the reference genome of Heinz, could not be compared to Heinz in the described analysis. 2) Sequence reads were mapped rather stringently. If more than two mismatches occurred, reads were discarded even though this might have represented valid sequence variation. Whereas 96 % of the reads from *S. lycopersicum* accessions could be mapped to the reference genome of cv Heinz, which also belongs to *S. lycopersicum,* only 53% of the reads from the wild species could be mapped to this reference genome. This implies that nearly half of the number of reads was deviating too much for being mapped. These reads were not included in the SNP frequency estimations shown above. Therefore, the genetic variation between *S. lycopersicum* and the sequenced wild species is far larger than suggested by the SNP counts. At the same time it should be kept in mind that all (re)sequenced wild species are crossable with *S. lycopersicum*, and can be used by breeders in conventional breeding programs.

It can be concluded that the genetic variation in *A. thaliana,* rice and tomato is enormous. When focusing on SNPs only, the detected genetic variation within these three groups is about 5 million, 9 million and > 10 million SNPs respectively.

The dynamic nature of plant genomes

Mutations after seed propagation. In order to gain insight into the mutation rate of seed-propagated plants, Ossowski et al. (2010) re-sequenced the genomes of five *A. thaliana* lines that were derived from one mother plant, and had been maintained in a glasshouse by single-seed descent for 30 generations. They identified and validated 99 base substitutions and 17 insertions and deletions, so in total 116 mutations. Their results imply a spontaneous mutation rate of 7×10^{-9} base substitutions and 1.4×10^{-9} indels per site per generation. This is equivalent to 2.3 spontaneous mutations per plant per generation.

Mutations after *in vitro* propagation (somaclonal variation). Somaclonal variation is a phenomenon that results in the phenotypic variation of plants regenerated from tissue culture or cell culture. It can be a result of activity of retrotransposons or other genetic mutations. It may also be caused by epigenetic modifications, such as adding or removal of methylation at DNA sites (Müller et al., 1990; Smulders and de Klerk 2011). Here we focus on mutations in the DNA sequence itself, observed by whole genome sequencing.

Jiang et al. (2011) showed that the mutation rate in *Arabidopsis* regenerated from *in vitro* culture was 60 to 350 times higher compared to the mutation rate observed in sexually propagated *Arabidopsis.* These regenerated plants were not transformed nor treated with *Agrobacterium.* Miyao et al. (2012) analysed the whole-genome sequences of three rice plants that were independently regenerated from a cell culture that originated from a single seed stock. The frequency of base substitutions was estimated to be 1.74×10^{-6} per site per regeneration. This is nearly 250 times higher than the base substitution frequency of 7×10^{-9} in the 30 sexual generations experiment of Ossowski et al. (2010). Ignoring that Miyao et al. worked with rice and Ossowski et al. with *A. thaliana*, these experiments indicate that base pair substitution occurs far more frequently during cell culture and/or regeneration than during seed propagation.

Miyao et al. (2012) also studied the activity of transposons. Among the 43 examined transposons, only one appeared to be active, i.e. *Tos17*. In one regenerated line, 10 new insertions of this transposon were detected. In an earlier study by Hirochika et al. (1996), the transposition of *Tos17* during tissue culture was already reported. The longer the period rice had been in tissue culture, the more transpositions of this endogenous copia retrotransposon were detected. Sabot et al. (2012) found that not only the retrotransposon family *Tos17* showed transpositions during tissue culture of rice, but at least 13 TE families appeared to be active in the used genotype, causing 34 new insertions.

Summarizing, genome wide mutations occur during tissue culture and cell culture, like during seed propagation. Although the types of mutations were similar, the frequency during in vitro culture was about 250 times higher compared to seed propagation. Also transposons were more active, enhancing the somaclonal variation. As the genetic modification process contains usually a tissue or cell culture phase and a regeneration phase, these phases can contribute to increased frequencies of spontaneous mutations compared to seed propagation.

Mutations associated with transformation

Examples in literature. There are a few published examples of genomes of GM crops that have been fully sequenced, and compared to the parental genomes. Kawakatsu et al. (2013) sequenced a transgenic rice line, obtained by *Agrobacterium*-mediated transformation, and found nearly 200 mutations compared to the parent. These mutations were spread among the genome. Based on the mutation rate and on the mutation pattern among the whole genome, Kawakatsu et al. conclude that the mutations detected in the transgenic rice line compared to its parent, were caused by somaclonal variation during *in vitro* culture. Alignment of the non-GM parental line to the Nipponbare reference genome of rice, revealed > 500 times more polymorphisms between these two non-GM genomes, compared to the mentioned ~ 200 mutations in the GM cultivar. The mentioned natural variation in rice according to Huang et al. (2012) is even about 40,000 times larger than the ~ 200 mutations in the GM cultivar. This illustrates that the frequency of induced mutations during transformation and regeneration was very small compared to naturally occurring variation in rice.

New experimental data on tomato. For assessment of the mutation frequency of GM plants due to *A. tumefaciens* mediated transformation, we developed a gm tomato plant and a non-gm tomato plant, both regenerated from tissue culture from a common parental inbred parent. We re-sequenced the genomes of the parent and both derived plants, and identified the mutations in the gm and non-gm offspring compared to their common parent. This experiment was performed in three biological replicates. We detected 274±217 mutations per plant in the gm-plants, and 97±3 mutations per plant in the non-gm regenerants. The T-DNA insertions themselves were not included in these mutation frequencies. The mentioned mutation frequencies were not significantly different, due to the large variation in the number of mutations in the gm-plants. However, the number of small deletions tended to be higher in transformants (14±5 deletion per plant) compared to non-gm plants (9±2 deletions per plant). The locations of the mutations in the genomes were not associated to the locations of the T-DNA insertions.

New experimental data on *Arabidopsis*. We also re-sequenced five *A. thaliana* transformants that were obtained from one parental plant using flower dip, so without regeneration from tissue culture. Only eight mutations were found (ignoring the T-DNA insertions themselves and deletions at the insertion sites), varying from 0 to 4 mutations per plant. The average of 1.6 mutations per plant is not significantly different from the frequency of spontaneous mutations during seed propagation in *A. thaliana* as determined by Ossowski et al. (2010).

Remarkably, six out of these eight mutations in the transgenic *A. thaliana* plants were small deletions, and the remaining two mutations were SNPs. However, in the spontaneous mutations in *A. thaliana* detected by Ossowski et al. the number of small deletions was 10 times lower compared to the number of SNPs. So, both in tomato and in *A. thaliana, A. tumefaciens*-mediated transformation seemed to have raised the number of small deletions, even when disregarding deletions at the T-DNA insertions themselves.

Comparison of the mutation frequencies in tomato and Arabidopsis. The mutation frequencies in gm tomato plants varied strongly, and were on the average about 250 times higher than in gm *A. thaliana* plants. The assembled genome size of tomato is about 6 times larger than for *A. thaliana*, and only partially explains this difference. The main cause for this difference must lie in the *in vitro* culture and regeneration in tomato, which was more mutagenic than the flower dip and seed production in *A. thaliana*.

Comparison of natural variation (baseline) with mutation frequencies due to transformation

The observed number of SNPs in tomato cultivars, when compared to the reference genome of the cultivar 'Heinz', varied from 200K to 4.5M. This variation is >250 times higher compared to mutations due to transformation and regeneration in our tomato experiments. The number of SNPs per accession of a related species used in conventional breeding was even >20,000 times higher than the number of mutations per transformant.

The number of mutations in the *A. thaliana* transformants was very small compared to the accumulated variation during evolution of this species: Cao et al. (2011) identified nearly 5.9 million small polymorphisms, including 810 K small insertions/deletions across 80 *A. thaliana* strains.

We conclude that the frequency of genome wide mutations due to transformation is very low compared to the frequency of polymorphism between cultivars, or genetic variation in breeding germplasm. *The variation due to transformation is well within the baseline of natural variation between cultivars, when ignoring the T-DNA insertion sites themselves.*

Small and large deletions at the insertion sites

In the five *A. thaliana* transformants we detected 12 T-DNA insertions sites. In 8 out these insertions 12 sites, small deletions and one very large (>700 kb) deletion were found, besides the T-DNA. From this observation we conclude that deletions in the genomic DNA at the T-DNA insertion sites were common. These deletions at the T-

DNA insertion site were not included in the 8 mutations in gm *Arabidopsis* plants mentioned earlier. When including the deletions at the T-DNA sites, then the number of deletions in the five gm *A. thaliana* plants raises from 6 to 14, so on average 2.8 deletions per plant. Ossowski et al. (2010) detected on average 0.13 deletions per plant per generation.

The observed frequency of deletions in the tomato and *A. thaliana* transformants at and far from the T-DNA insertions were higher than in the non-gm plants. The experiment does not answer the question whether *A. tumefaciens* favours double strands breaks in order to create more opportunities for T-DNA integration, or whether it interferes with DNA repair, leading to genome wide development of deletions. However, the frequency of deletions due to transformation is still very low compared to the natural frequency of deletions in cultivars within a species.

Splinter discovered

By means of scrutinizing the NGS-data, we discovered a splinter of 50 basepairs in a transgenic *A. thaliana* plant. This splinter was a small part of the *gfp*-gene that originated from the used T-DNA. The presence was confirmed by PCR. This splinter was not detected in the other four transformed *A. thaliana* plants. The splinter was inserted into an intron of a gene, not leading to a change in the protein, according to gene prediction software. As far as we know, this is the first report on occurrence of a ‘splinter’ in a transgenic plant. Such a splinter would probably be overlooked when using Southern blotting techniques, due to its small size.

NGS for discovery of T-DNA insertions

Next Generation Sequencing (NGS) appeared to be a suitable technique for discovery of T-DNA insertions in gm plants, is far more sensitive than Southern blotting, and cheaper. Also small partial insertions (splinters) can be detected, that may be overlooked by Southern blotting and PCR. NGS can also be used for positioning these insertions in the genome, in case a reference genome sequence is available for the plant species. Furthermore, NGS can be used for revealing the sequences of the inserts and flanking DNA. However, several conditions have to be taken into account, which are discussed in this report. Particularly, we recommend validation of the putative inserts by using an additional approach, such as amplification of the inserts and their flanking DNA, followed by sequencing, preferably using a technique that provides long sequencing reads, such as PacBio. We think that under these conditions NGS will outperform and is suitable to replace Southern blotting and genome walking from T-DNA.

Recommendations

In view of the high sensitivity and relatively low costs of NGS, we recommend using NGS for molecular characterisation of T-DNA insertions and flanking DNA in gm plants. However, we do not recommend a compulsory use of the obtained NGS data for identification of mutations elsewhere in the genome, as the mutation rate associated with the transformation is very small compared to the natural genetic variation (baseline level), when disregarding the T-DNA insertions and their insert sites.

1. Introduction

1.1 Molecular characterisation of a GM plant from local to genome-wide sequencing?

For authorisation for import or cultivation of a genetically modified (gm) plant event in Europe, a dossier has to be delivered in order to evaluate potential environmental risks of that event. Molecular characterization is part of the dossier. At the genomic level, the molecular characterization includes an evaluation of the number of inserts, the DNA sequence of the insert(s) in the plant, the DNA sequences of the regions flanking the insert, evaluation of possible interruption of endogenous genes by the insert, homology of the DNA sequences of junctions of flanking DNA and inserts to known genes that code for toxins or allergens, possible introduction of backbone vector DNA, etc. (EFSA, 2011). This characterisation at the DNA level has been based on 'classical' molecular techniques such as Southern blotting for the number of inserts and vector backbone integrations, PCR and sequencing of inserts, and genome-walking tools for revealing of flanking DNA sequences, etc. This characterisation does not necessarily include all changes that occurred in the genome.

Next generation sequencing techniques allow nowadays sequencing of numerous DNA fragments in parallel at relatively low prizes, including (re)sequencing of whole genomes, which may replace Southern blots. Recently, a paper described the molecular characterisation of a GM plant, using whole genome sequence of the GM plant (Kawakatsu et al., 2013). This does not only provide information on the insert and its flanking DNA, but also on the whole genome. This whole genome sequence can be compared with the whole genome sequence of the non-GM parental plant. Any deviations in the genome of the GM plant compared to the parental genome can thus be discovered. These deviations could be a result of the transformation process itself, or can be a consequence of somaclonal variation, i.e. spontaneous mutations that occur during tissue culture, regeneration and propagation of the GM plant.

However, also the genome of the parental cultivar changes over time, due to accumulation of natural mutations, including nucleotide substitutions, insertions, deletions, translocations, movement of transposons. These spontaneous mutations will only occasionally lead to a visible change in the phenotype.

The company Monsanto indicated that in future applications for import or cultivation of GM plant events into Europe, it will use whole genome sequence information, and select the sequences of the insert and flanking sequences rather than the classical molecular characterisation by Southern blots (see for method Kovalic et al, 2013). The COGEM would like to know how to use this information for the molecular characterisation in view of the environmental risk assessment (ERA). Currently, the guidance document from EFSA for molecular characterisation of GM crops does not mention whole genome sequence data but refers to Southern blotting techniques, and other more classical techniques (EFSA, 2011), restricted to the inserts and flanking DNA.

Further, the report at hand discusses the added value of the whole genome sequence information for the environmental risk assessment (ERA) on top of the molecular characterisation, in view of the genetic variation within the baseline.

Epigenetic modifications may play a role too, but these are outside the scope of this project.

1.2 The baseline

In the environmental risk evaluation, potential adverse effects of the new traits are compared to effects of non-GM plants, and the natural range of these traits in cultivars and germplasm that can be used for conventional breeding of that crop. This natural variation is named the baseline. The baseline in a narrow sense refers to the natural variation that occurs in non-GM plant cultivars, whereas the baseline in a wider sense includes also the non-GM germplasm that can be used in conventional breeding of the crop, such as crossable wild relatives of the crop. The EFSA uses the narrow definition for the natural variation, and refers to the variation among commercial non-GM varieties, rather than the non-GM germplasm that can be used for conventional breeding of the crop (EFSA, 2010).

1.3 Whole genome sequencing (WGS)

At the start of the second millennium, in 2000, the first whole genome sequence of a plant was published, i.e. of *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000). A year later, the first human whole genome sequence was published, and in the subsequent year the whole genome sequences of two rice cultivars followed (Yu et al, 2002; Goff et al, 2002). All these sequences were based on the so-called Sanger sequencing technology (Hamilton and Buell, 2012). However, thanks to the emerging next generation sequencing technologies, and dramatic cost reductions of these high throughput techniques (Muers, 2011; Edwards et al, 2013), the number of crops for which a whole genome sequence has been published, has increased tremendously since 2008 (Fig 1).

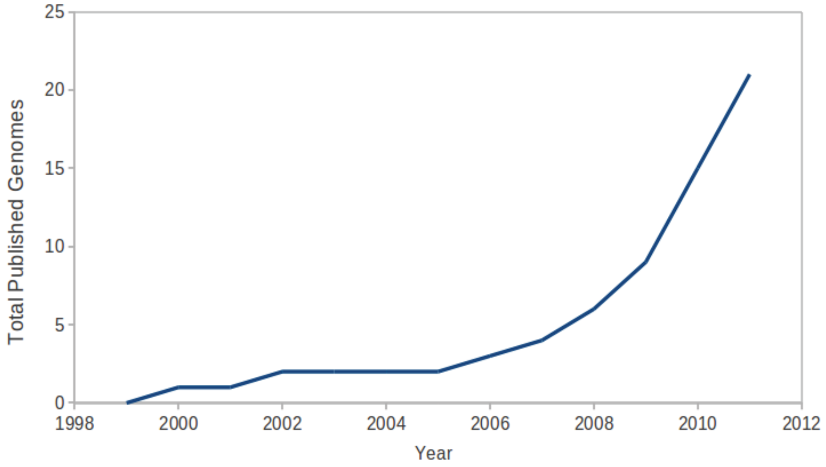


Figure 1. Number of published genomes of plant species. Since 2011 this number continued to increase exponentially.
Source: http://genomevolution.org/wiki/index.php/Sequenced_plant_genomes).

1.4 Re-sequencing

These whole genome sequences have provided reference genomes for so-called re-sequencing projects. Huang et al. (2011) re-sequenced the genomes of 150 inbred lines from a cross of two parental rice genotypes. This led to an extremely detailed genetic map, showing very precisely recombination sites that had occurred in the parents during meiosis.

Following the ambition of the 1000 human genomes project (http://www.1000genomes.org), a 1001 genomes project for *Arabidopsis thaliana* was initiated (Weigel and Mott, 2009; Cao et al, 2011).

A third example is tomato. After the sequencing of the tomato genome (Sato et al., 2012), Wageningen University and Research Centre, has re-sequenced 150 tomato genomes. Recently, Lin et al. (2014) resequenced 360 tomato genotypes.

Another example is the re-sequencing of three rice genotypes after regeneration. This revealed the level of somaclonal variation that arose during in vitro culture and regeneration (Miyao et al., 2011). The observed molecular spectrum was similar to that of the spontaneous mutations in *Arabidopsis thaliana*. However, the base change ratio was estimated to be 248-fold higher than the spontaneous mutation rate of *A. thaliana.*

In the human genome research, the number of re-sequenced genomes will proceed to grow at a high speed, as appears from the announcement of the 100K genomes project: The Genomic Medicine Centres are due to start work in January 2015, with sequencing of 100,000 human genomes, to be completed by the end of 2017, using Illumina sequencing technology (MIT Technology Review, 14 July 2014; http://www.bioworld.com/content/100k-genomes-project-moving-forward-medicine-centers-open-2015).

1.5 Goals of the project

The goals of the current project are:

7. Describe the baseline of natural variation in genomes of plants, used in conventional plant breeding. Different sources of variation in conventional breeding will be distinguished, i.e.
    a. Variation in DNA sequences of wild germplasm, old cultivars, land races and current cultivars. This is variation due to accumulation of mutations as a result of evolution, and human-mediated selection. This variation is represented in the breeding material of the crop;
    b. Mutations that still do occur spontaneously during propagation and breeding of the crop. This variation is due to natural mutations in vegetatively and generatively propagated plant material. We mean here the current dynamics of plant genomes, on top of the chromosomal crossovers that occur during meiosis;
8. Evaluate changes in the genome due to tissue culture and regeneration (= somaclonal variation);
9. Evaluate changes in the genome due to transformation;
10. Compare changes in the genome due to transformation (goal 3) with the baseline of natural variation (goal 1) and somaclonal variation (goal 2);
11. Discuss additional biosafety questions that may arise when whole genome sequences of GM plants can be obtained;
12. Provide suggestions for using the whole genome sequence data of GM plants for the environmental risk assessment of these plants.

Both existing literature and new experimental data will be presented.

1.6 Limitations of this study

1.6.1 Only WGS studies

There are numerous papers that describe variation of germplasm for plant breeding programs. This variation describes polymorphism at the DNA level, represented by DNA markers such as AFLP and SSR, variation at the RNA level e.g. represented by cDNA-AFLP, micro-arrays or RNA-Seq, variation at the protein level, metabolic variation, and self-evidently phenotypic variation. The current report is confined to natural variation of breeding germplasm revealed by whole genome sequencing (WGS). Probably we are currently at the onset of a wealth of papers, describing natural WGS variation between sexually compatible plants. However, at this moment the number of papers on this topic is still limited. These publications are discussed here.

1.6.2 Three genera

The second limitation of this study is the choice of plant species. We limit this study to plant genera for which information is available or produced in this study for the natural variation revealed by WGS, and information on mutation rates after sexual reproduction, after in vitro culture and regeneration, and after genetic transformation. Based on these criteria, this study focuses on the three genera *Arabidopsis,* rice and tomato.

1.6.3 Mutation breeding not included

Induced mutations by chemicals or irradiation treatments have been used also widely in plant breeding. The freely accessible FAO/IAEA website contains a database of > 3000 plant varieties derived from induced mutations (http://www-infocris.iaea.org/MVD/default.htm). The mutation breeding, however, is not discussed in this report.

2. The natural variation of germplasm for plant breeding programs

2.1 Arabidopsis thaliana (literature)

80 natural strains. Although *A. thaliana* is not grown as a commercial crop, it is the most widely studied plant species for unravelling genetics and plant genomes. It is the model plant for studying the relationship between genomes and phenotypic traits. Also, it is the first plant species of which a whole genome sequence was analysed (Arabidopsis Genome Initiative, 2000). Currently, the so-called 1001 genome project is undertaken for *A. thaliana* (Weigel and Mott, 2009), that aims at re-sequencing 1001 *A. thaliana* accessions, and the results of the re-sequencing the genomes of the first 80 strains of this species have been published (Cao et al, 2011). These 80 strains were chosen to represent the genetic diversity present in eight populations across the native range of the species in Eurasia, spanning various climates and elevations, from the high mountains of Central Asia to the European Atlantic Coast, and from North Africa to the Arctic Circle.

SNPs and small indels. The authors first predicted single-nucleotide polymorphisms (SNPs) and small insertions and deletions, ranging from 1 to 20 bp hereafter called small indels, in unique regions on the basis of single-read alignments against the 119 Mb reference genome of *A. thaliana*, excluding 6.8 Mb of highly repetitive regions.
In the remaining 112 million base pairs, Cao et al. identified nearly 5 million (4,902,039) SNPs across the 80 strains (Figure 2). This represents, on the average, one SNP per 23 bp, taking all 80 strains into account. Most SNPs were not restricted to one strain only, but were found in at least two strains. As few as 56 accessions were sufficient to detect 98% of all SNPs shared between geographical regions. Although a large number of rare SNPs remained to be discovered, Cao et al. captured a substantial majority of common SNPs. The number of discovered private SNPs per accession, not present in any of the other studied accessions, was between 322 and 93,199. More than 800,000 (810,467) small insertions/deletions (indel 1-20 bp) were detected in the 80 accessions, which is on the average one small indel per 140 bp.

Structural variants. Cao et al. sequenced fragments of DNA from both ends of the fragments, giving sequence reads that were physically close together on the genome, and in opposite directions. These twins of reads are named ‘paired ends’. Cao et al. aligned the paired ends of the accessions to the reference genome, for evaluation of structural variants. In case of structural variants, the paired ends map not as ‘twins’ but map at different locations or in unexpected orientations. This discordant mapping of paired ends revealed structural variants (SVs) of at least 20 bp in length. Because of uncertainty associated with the precise locations of SV ends, the number of unique SV deletions was more difficult to determine, but conservatively this number was at least 174,789, of which 49% were detected in more than one strain.

Transposable elements. In the reference genome of *A. thaliana*, 31,189 transposable element insertions have been annotated. From these insertions, 80% showed evidence of being partially or completely absent from the genome of at least one of the 80 sequenced strains. This underlines the variability of these elements.

Copy-number variants. Also inferred copy-number variants (CNVs) of minimum length of 1,000 bp were evaluated. Cao et al. detected more than 1,000 (1,059) CNVs covering 2.2 Mb of the reference genome, of which 393 overlapped with coding sequences. This indicates also copy number variants for genes.

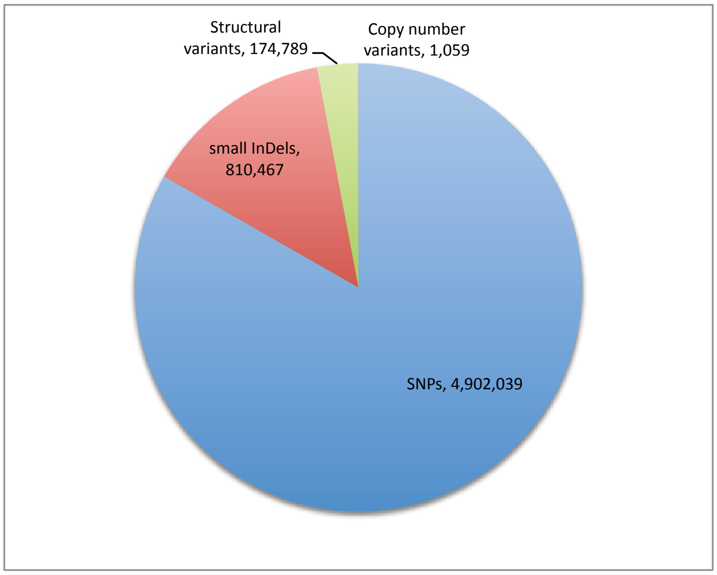


Figure 2. Number of DNA polymorphisms among 80 *Arabidopsis thaliana* accessions, for the haploid genome of 112 Mbp, excluding repetitive DNA. Most variants were detected in more than one accession. Data from Cao et al. (2011).

'Drastic' mutations in genes. Cao et al. evaluated whether changes in DNA sequences were likely to have functional consequences based on their annotation. Across the 80 accessions, they discovered SNPs in more than 6,000 (6,197) genes that altered start codons, introduced premature stop codons, extended the open reading frame of the reference sequence, or affected splice donor or acceptor sites. 4,263 genes had a premature stop in at least one accession, and 2,793 in two or more accessions. Furthermore, 4,525 larger SV deletions overlapped coding sequences of 2,247 genes by at least 50 bp, thus seriously affecting the functions of these putative genes. These SNPs and SV imply drastic mutations in more than 8000 genes. Therefore Cao at el. named these mutations 'drastic mutations'. Finally 27,167 small indels in open reading frames have the potential to cause frame shifts. However, multiple indels in *A. thaliana* may restore the correct frame, and thus the number of frame shifts that drastically change coding sequences is likely to be much smaller. The drastic mutations were most abundant in NB-LRR genes. This family represents a highly dynamic group of resistance gene analogs, involved in resistance to pests and diseases.

An additional interesting observation was that the ratio of deleterious to tolerated mutations was negatively correlated with the effective population size. In small populations, deleterious mutations could accumulate at intermediate frequencies as a result of less effective purifying selection compared to large populations. This confirms that inbreeding depression apparently was a more serious threat to these marginal populations compared to large populations.

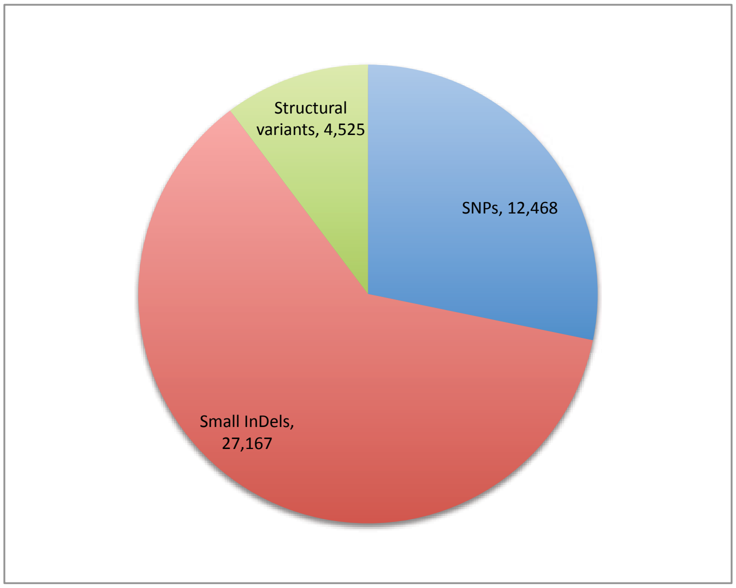


Figure 3. Number of mutations in genes in the 80 A. thaliana accessions, probably affecting the genes' functions drastically. The more subtle mutations are not included. Data from Cao et al. (2011).

More subtle mutations. Cao et al. (2011) also analyzed more subtle mutations in genes and between genes, comparing the 80 accessions. They distinguished mutations in the UTRs (untranslated regions), introns, in intergenic regions, deleterious non-synonymous mutations likely to change the genes function heavily, and tolerated non-synonymous mutations in coding regions that probably do not have a strong impact on the gene function. Whereas 20 % of the genes showed premature stop codons in one or more accessions, approximately 30 % of the genes showed a deleterious non-synonymous mutation. About 50 % of the genes had one or more tolerated non-synonymous mutations in one or more accession. The authors did not report explicitly on the frequencies of synonymous mutations.

Restoring gene disruptions. Gan et al. (2011) sequenced the genomes and transcriptomes of 18 natural *A. thaliana* accessions. On the basis of the reference annotation, one-third of protein-coding genes were predicted to be disrupted in at least one accession. However, re-annotation of each genome revealed that alternative gene models often restore coding potential. This implies that drastic mutations may not always knock out the mutated genes completely, but that the genes may be restored to some extent.

Variation in expression. The expression of nearly half of the expressed genes varied among seedlings from the 18 accessions (Gan et al., 2011). Deviation of expression was frequently associated with *cis* variants within 6 kilobases, as were intron retention alternative splicing events.

We can conclude that the genetic variation of the species *A. thaliana, growing at various sites around the globe, is enormous*. In the course of evolution millions of mutations have occurred, lost and added, and still continues to accumulate, implying a wide baseline.

2.2 Brassicaceae (literature)

Conserved noncoding sequences. *Arabidopsis* belongs to the Brassicaceae family. Haudry et al. (2013) compared genomes of plants belonging to nine different species within the Brassicaceae. In the overwhelming amount of variation, they searched for conservation. The detected conservation across orthologous bases suggested that at least 17% of the *A. thaliana* genome is under selection, with nearly one-quarter of the sequence under selection lying outside of coding regions. The authors localized approximately 90,000 conserved noncoding sequences (CNSs). Population genomics analyses of *A. thaliana* and *Capsella grandiflora* confirmed that most of the identified CNSs are evolving under medium to strong purifying selection. From binding studies it appeared that many CNSs are regions bound by transcription factors, and therefore are transcriptional regulatory elements (Haudry, 2013).

Huang et al. (2013) re-sequenced ten elite accessions of *Brassica napus* (oilseed rape; 2n = AACC). The aim was to develop SNP arrays for genetic association studies of breeding material of this crop. The progenitors of this polyploid species are *Brassica rapa* (2n = AA) and *B. oleracea* (2n = CC). A total of 1600 million paired-end reads of 75-bp or 100-bp read length were aligned against the references genomes of *B. rapa* and *B. oleracea*, at a sequencing depth for each variety from 5.3× to 37.5×. The authors discovered in the ten elite accessions of *B. napus* nearly 900,000 bi-allelic SNPs throughout the genome. They found in these elite lines more than 36,000 putative amino acid variants in 13,552 protein-coding genes.

Although the genome of this species is about 10 times larger than the genome of *A. thaliana,* the number of SNPs in *B. napus* for the SNP arrays is 'only' about 18 % compared to the number of detected SNPs in the 80 natural accessions in *A. thaliana.* The main reasons are:

1. The strong filtering applied by Huang et al. (2013) for selection of SNPs that are suitable for SNP arrays. They removed all reads that aligned at more than one locus. Also they excluded 78 million of reads that were supported by less than four reads. Further they removed more than 6 million SNPs that were heterozygous in at least one individual, and another 7 million SNPs with minor allele frequencies. This illustrates that the filtering has a very dominant impact on the number of remaining SNPs. Because of the very high frequency of SNPs, the filtering for development of SNP arrays could be very stringent, reducing the amount of SNPs to a few per cent only compared to the initially detected set of SNPs. But still sufficient SNPs remained that could be used for SNP arrays.
2. Further it has to be taken into account that Huang et al. (2013) used elite lines that passed strong selection by canola breeders in view of commercial value of the lines, whereas Cao et al. (2011) used natural accessions from very different environments in Eurasia. Wild germplasm covers more genetic variation compared to commercial cultivars of the crop.
3. Huang et al. (2013) used 10 accessions, whereas Cao used 80 accessions. Cao et al. (2011) mentioned that at least 67 accessions were needed to identify 98% of all non-private SNPs.

4. Although *B. napus* and *A. thaliana* belong to the same family, they are two different species in life history and demography: species may vary considerably in genetic variation within species.

The re-sequencing study by Huang et al. (2013) on *B. napus* aiming at selection of SNPs for SNP arrays, illustrates that the strong filtering for SNP arrays does not allow a comprehensive description of the whole genome variation of breeding germplasm including wild material. These re-sequencing studies are not meant to provide that comprehensive description, but aim at detection of informative SNPs markers that can be used for studies on associations between SNPs and phenotypic traits. For this reason, other papers on re-sequencing aiming at development of SNP-arrays are not discussed here.

We conclude that although NGS studies are suitable for elucidating the genetic variation of germplasm of breeders, most NGS studies are more focused on discovery of SNPs, which is usefull in breeding programs in view of molecular markers for phenotypic traits. Although a large majority of the genome wide variation is then neglected, still many thousands of SNPs remain that allow genome wide association studies (GWAS) or the creation of high-density genetic linkage maps.

2.3 Rice (literature)

Xu et al. (2011) sequenced 40 cultivated accessions from the major groups of rice (*Oryza sativa*) and 10 accessions of their wild progenitors (*O. rufipogon* and *O. nivara*) to >15 × raw data coverage. They obtained 6.5 million SNPs, which is about 1.6 million more than Cao et al. (2011) detected in the 80 wild *A. thaliana* accessions (Figure 2). Using these SNPs, the authors detected thousands of genes with significantly lower diversity in cultivated compared to wild rice.

Huang et al. (2012) performed a similar study, but increased the number of accessions considerably from 40 to 1,529 accessions. However, the sequence depth was far lower. They sequenced 446 geographically diverse accessions of the wild rice species *Oryza rufipogon,* the immediate ancestral progenitor of cultivated rice. This sequencing was performed with two-fold genome coverage. Aligning the reads against the rice reference genome sequence, revealed a total of 5,037,497 non-singleton SNPs (Huang et al. 2012), which is a similar number of SNPs detected in the 80 wild *A. thaliana* accessions (Figure 2). On top of these *O. rufipogon* accessions, Huang et al. (2012) analysed the genomes of more than 1000 (1,083) cultivated *O. sativa* varieties, both from the *indica* and *japonica* groups, to construct a comprehensive map of rice genome variation. These varieties were sequenced at one-fold genome coverage. The total number of sequenced *O. rufipogon* and *O. sativa* accessions was 1,529, yielding nearly 8 million (7,970,359) non-singleton SNPs. The authors used these SNPs for an in-depth study on the history of domestication of rice, such as reduction in nucleotide diversity and altered allele frequency in the domestication loci.

Bin Han (National Center for Gene Research, China) presented at the PAG-meeting in San Diego in January 2014 the 3000 Rice Genome Project of the Global Rice Science Partnership (GRiSP). The number of accessions doubled compared to Huang et al. (2012). Further the sequencing depth improved to 14 x on the average.

They detected 18.9 million SNPs, which is nearly four times higher than in the 80 *A. thaliana* accession (Figure 2). We have to realize that in this rice study not only the number of accessions was higher compared to the Arabidopsis study, but also more than one species were sequenced, and furthermore the genome of rice is nearly three times larger than of *A. thaliana*, i.e. 374 Mbp and 134 Mbp respectively. The detected SNP frequency in *A. thaliana* was one SNP per 23 bp (Cao et al., 2011), and in rice one SNP per 20 bp (Bin Han et al., PAG meeting), which is in the same order of magnitude.

2.4 Tomato (re-analysis of NGS data for this report)

In the last few years, WUR coordinated a large re-sequencing initiative in tomato, entitled "The 150 tomato genome re-sequencing initiative" (http://www.tomatogenome.net). Eighty-four out of the 150 tomato genotypes are old varieties, land races and related wild species (Fig. 4), representing different tomato types, such as cherry, beef, round, pink and heirloom types. Ten old varieties, 43 land races and 30 wild accessions were included. All accessions can be crossed with cultivated tomato. The remaining individuals were genotypes belonging to a population of recombinant inbred lines (RIL population) from a cross between *S. lycopersicum* cv. Moneymaker and *S. pimpinellifolium*.

The 84 tomato genotypes were selected to represent the genetic variation in an initial collection of > 7000 tomato accessions. They were sequenced using Illumina HiSeq 2000. The average coverage per accession was 36.7±2.3 X. Sequence reads of accessions were mapped against the reference genome of *S. lycopersicum* cv. Heinz 1706 v2.40. The cultivar 'Heinz' is a representative of old cultivars.

For three wild species the reads were assembled *de novo*. The rationale for these three new assemblies was that the re-sequenced genomes differ sometimes significantly from the published reference genome. For example, the genome size of *S. pennellii* (one of these three *de novo* sequenced genomes) is estimated to be 1.3x larger than the sequenced genome of the cultivated tomato (*S. lycopersicon*). It is expected that the expansion of the *S. pennellii* genome is mainly due to the accumulation of retro-transposons, but possibly also due to new genes compared to the reference genome. These sequencing and re-sequencing efforts provide detailed information on the natural variation at the genome level in tomato cultivars and germplasm.

Here we summarize the main results from that project regarding natural genetic variation in the tomato cultivars and wild germplasm, as it provides very detailed information about the natural variation (baseline). The recent publication from Aflitos et al. (2014) has been based on the same data, and therefore there is some overlap between the paper of Aflitos et al. and the results presented here. However, in view of this report, additional analyses have been carried out that are not presented by Aflitos et al. (2014).

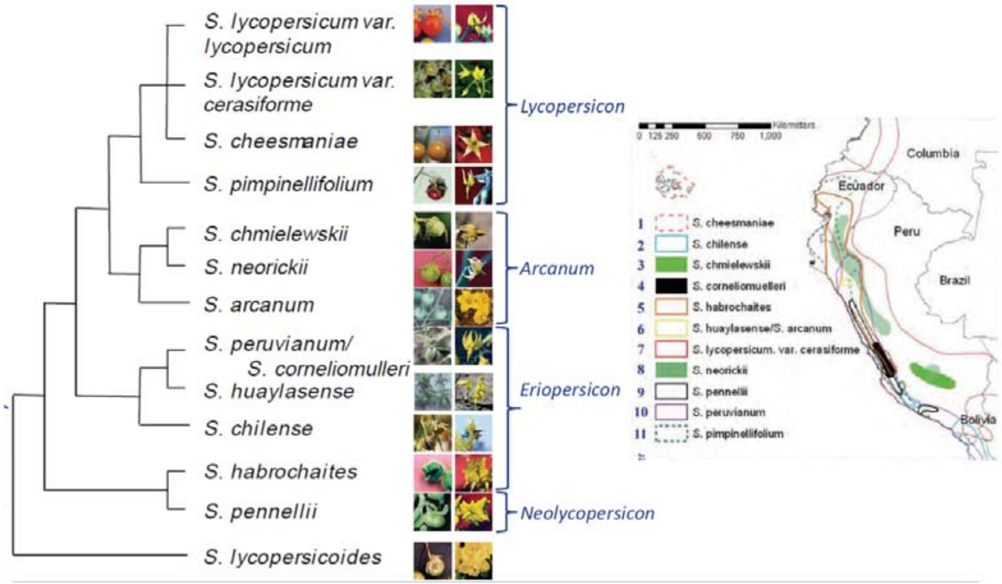


Figure 4. Schematic overview of likely species relationships and geographical origin in the plant clade *Solanum* section *Lycopersicon*. The consensus phylogenetic tree is based on four independent genetic studies from Spooner et al. (2005). It includes three recently described species by splitting the *S. peruvianum* complex (Peralta et al. 2005). Placement of *S. cheesmaniae*, *S. arcanum* and *S. habrochaites* is tentative. Tomato (wild) species are classified in four larger sections (*Lycopersicon*, *Arcanum*, *Eriopersicon*, and *Neolycopersicon* section). The outgroup species *S. lycopersicoides* is included for reference.

2.4.1 Allelic variation

SNPs

To assess the sequence diversity in *Solanum* section *Lycopersicon,* the SNPs were classified and quantified for each of the 84 accessions. The SNP counts for tomato cultivars (*S. lycopersicum* and *S. lycopersicum* var. *cerasiforme*) were relatively low (between 200K and 4.5M; 850K on average; 1 SNP / 800 bp on average) and gradually increase for accessions of *S. galapagense* and *S. cheesmaniae* (3.3M on average), and *S. pimpinellifolium* (4.6M on average). For members of the wild species, not belonging to the species *S. lycopersicum*, SNP numbers increased sharply (Fig. 5), to 8 to 10 Million SNPs per accessions (approximately 1 SNP / 80 bp). This is in agreement with their more distant position in the phylogenetic tree for the tomato clade (Peralta *et al*, 2008; Fig. 4).

The sequence diversity in the *S. lycopersicum* group appears rather large at first sight (the blue bars in Figure 5). However, the majority of the accessions belong to the category 'old cultivars and landraces'. Old cultivars are product of plant breeding by companies, whereas landraces are the result of selection by (small) farmers and growers, with local adaption to the needs of these growers. These accessions show limited variation (200K to 500K SNPs). The second category (>500K SNPs) includes more deviating landraces, tomato accessions collected from their wild habitat, and tomato varieties that resulted from introgression breeding. For example, *S. lycopersicum* var. *cerasiforme* LA1479 was collected from its natural habitat (Santa Cecilia, Ecuador) and is one of the most deviating *S. lycopersicum* accession sequenced, showing 4.3M SNPs compared to cultivar 'Heinz' (1 SNP/170 bp; almost

5 times higher than on average in cultivars). The number of SNPs in the wild species compared to 'Heinz' was on average 20 times higher than in cultivars.

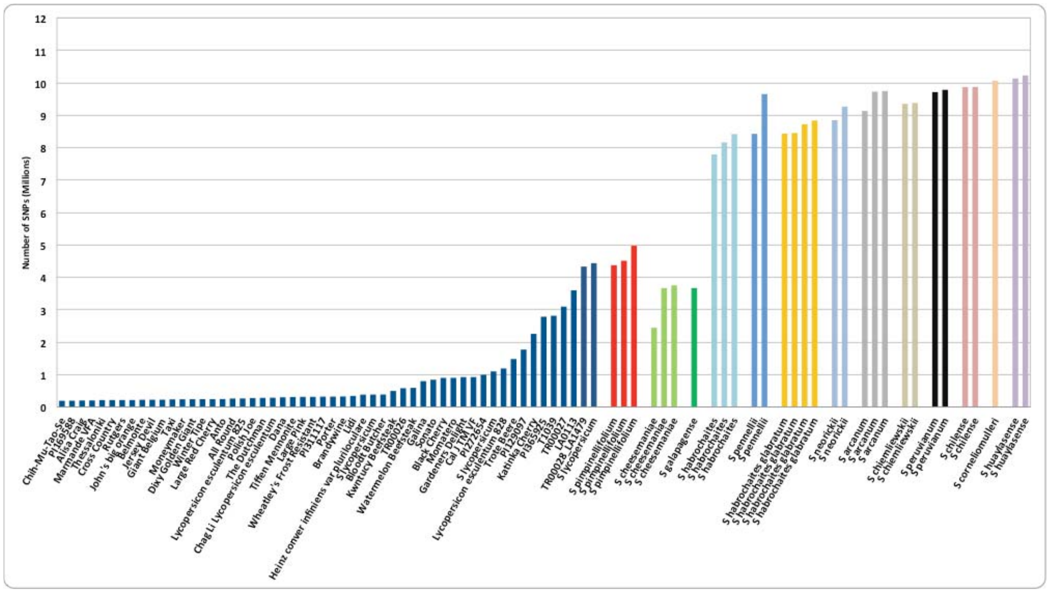


Figure 5. SNP and (small) INDEL variation in the re-sequenced accession in comparison to the genome reference sequence of cultivar Heinz 1706. Accession names are indicated below the horizontal axis. Colour-coded bars indicate species groups. The dark blue bars represent the *S. lycopersicum* germplasm. From: Aflitos et al. 2014; Finkers pers. communication.

When compared to the *S. lycopersicum* Heinz 1706 SL2.40 annotated genome, we consistently observed in all accessions a significant higher SNP frequency in intergenic regions compared to genic regions. Approximately, 89.5%±3% of the SNPs were detected in intergenic regions. In genic regions the SNP frequency was highest in the non-coding regions: 7.5%±2.2% mapped in introns. The 5' and 3' UTRs showed more polymorphisms than the internal introns, which in turn had a higher SNP frequency compared to exons. From the polymorphisms in exons, 32.4%±15.2% appeared to be synonymous while 40%±8.3% was non-synonymous (fig. 6). The remaining of the polymorphisms resulted in different effects, such as a frame shift, loss or gain of a start or stop codon, etc.

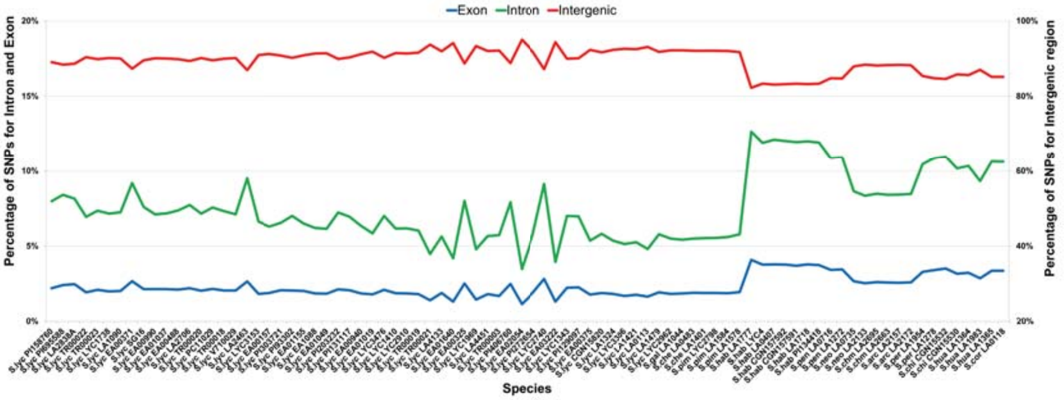


Figure 6. Genome-wide SNP ratio for 84 accessions. SNP classes are colour coded as indicated. Accession IDs and SNPs percentage in linear scale are indicated on the x-axis and y-axis, respectively. From: Aflitos et al. 2014; Finkers pers. communication.

A striking finding is the ratio between non-synonymous and synonymous SNPs (dN/dS). For crops, non-synonymous SNPs outnumber synonymous SNPs while the opposite is generally true for wild species (Fig. 7). Although we currently have no clear explanation for the relatively higher dN frequency in crop tomatoes, it might partly be the result of (introgression) breeding that maintains SNPs under positive selection (favouring amino acid replacements). This question requires a more detailed follow-up on the level of the individual genes.

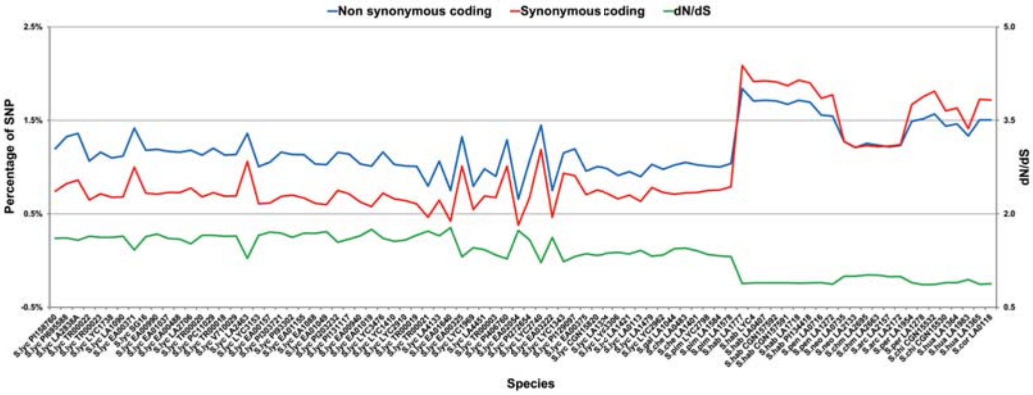


Figure 7. Non-synonymous (dN) and synonymous (dS) SNPs in tomato accessions and related wild species. The dN and dS percentage, and dN/dS ratio relative to the total number of SNPs per accession is indicated in the left and right vertical axis respectively. From: Aflitos et al. 2014; Finkers pers. communication.

Copy Number Variation

Also copy number variations were observed in the re-sequenced genomes. An in-depth analysis of this type of genomic variation has not yet been conducted.

Traces of introgression breeding

Related tomato wild species are exploited to improve (commercial) tomato cultivars for traits such as disease resistance or tolerance to adverse abiotic conditions. These introgressions can easily be detected from re-sequencing data by visualizing SNP & INDEL rate in relation to the position on the genome (fig. 8). For this, variation was counted per 10KB bin and plotted along the physical position. Regions with a high number of variants (plotted in green) can easily be discriminated from regions showing low variability (black). *S. lycopersicum* accessions (plotted starting left) can be easily distinguished from related wild species (bright green block on the right). Six *S. lycopersicum* accessions, containing an introgression from a wild species (7, 12, 15, 74, 75, and 85) can be detected too. This finding is in agreement with the known presence of alleles for a gene conferring resistance to Tobacco Mosaic Virus (TMV). However, the large size of the introgressed chromosomal part was a surprise to many breeding scientists involved in the project. Apparently, a very large part of chromosome 9 has been co-introgressed with the resistance gene on this chromosome. From previous research, we know that the fragment, containing the TMV locus is actually inverted. The consequence of this is that this fragment does not recombine anymore after its initial introgression. Also, the resistance gene is located in a heterochromatic area with a low gene frequency, which, as a rule of thumb, hardly recombine in the tomato crop. Nevertheless, we do notice that the TMV introgression differs in size in some cultivars, as can be observed at the borders of the introgression in Figure 8. These borders are in the gene-rich euchromatin. In the context of the tomato genome, this can mean several hundreds of genes either from the wild species or from tomato, which could be very relevant for the phenotype of the genotype. This example clearly illustrates the increase of the genetic variation in commercial cultivars as a result of introgression breeding.

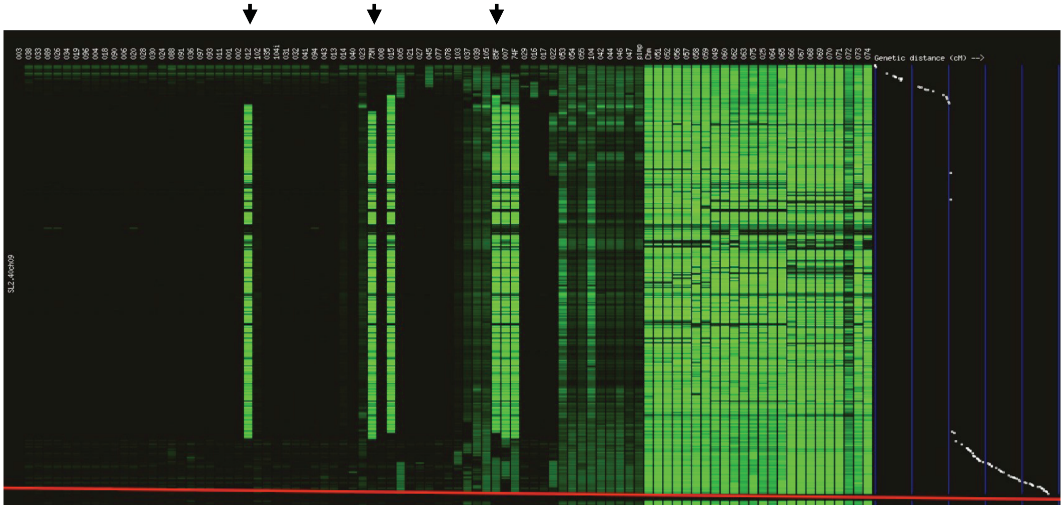


Figure 8. Visualisation of traces of introgression breeding on chromosome 9 in tomato. The vertical axis represents the physical positions on Chromosome 9. At the left side, the horizontal axis represents the accessions, shown in Figure 5. The number of allelic variants per bin of 10KB was counted. Regions with a high number of deviations from the reference genome of 'Heinz' are shown in green, and regions containing low numbers of SNPs are displayed in black. The *S. lycopersicum* accessions (plotted starting left) can be easily distinguished from the sequenced wild species (bright green block on the right). Six *S. lycopersicum* accessions, containing an introgression from a wild species can be recognized too, in the middle of the figure. The plot at the right side visualizes the physical distance (y-axis) vs. the genetic distance (x-axis). Ends of chromosomes are showing a lot of recombination while the centre (heterochromatin) does not recombine at all.) From: Aflitos et al. 2014; Finkers pers. communication.

2.4.2 Structural variation

The *de novo* (draft) assemblies of three related wild species enabled investigating the occurrence and extent of structural variation between these species and the tomato genome reference of *S. lycopersicum* cv. Heinz SL2.40. We searched for colinearity breaks in contigs obtained from the *de novo* assemblies that could be indicative for differences in chromosomal organization between tomato and related wild relatives (Fig. 9). To minimize the detection of false positives in erroneously assembled contigs, we applied an alignment threshold of 90% sequence identity. Furthermore, we only considered aligned scaffolds containing segments in opposite orientations. Such intra-contig inversions were observed in each of the 12 chromosomes at multiple positions for each of the reference species. We found 534, 393 and 393 putative inversions for *S. habrochaites* LYC4, *S. arcanum* LA2157 and S*. pennellii* LA716 scaffolds, respectively. The number of rearrangements was highest for chromosome 1 and lowest for chromosomes 2 and 11. We were able to size 53 putative small intra-scaffold inversions (23, 28 and 2 putative rearrangements for *S. habrochaites* LA2157, *S. arcanum* and S. *pennellii,* respectively). Analysis of homology of the sequence surrounding the inversion indicated that at least nine cases appeared related to *Mu* or *hAT* transposons. In four cases we found overlapping rearranged segments between *S. arcanum* and *S. habrochaites*. Szinay *et al*. (2012) and Peters *et al*. (2012) showed that in the *Solanaceae* also large-scale rearrangements of several megabases in size can occur. However, we are currently unable to study these, because: 1) The proper ordering and orientation of contigs into megabase sized scaffolds depends on the availability of genetic maps and physical maps, which are currently lacking for the three *de novo* sequenced genomes; and 2) The N50 contig sizes for *de novo* assemblies of *S. arcanum*, *S. habrochaites* and *S. pennellii* do not exceed 50kb. We cannot yet detect such rearrangements above the 200 kb range.

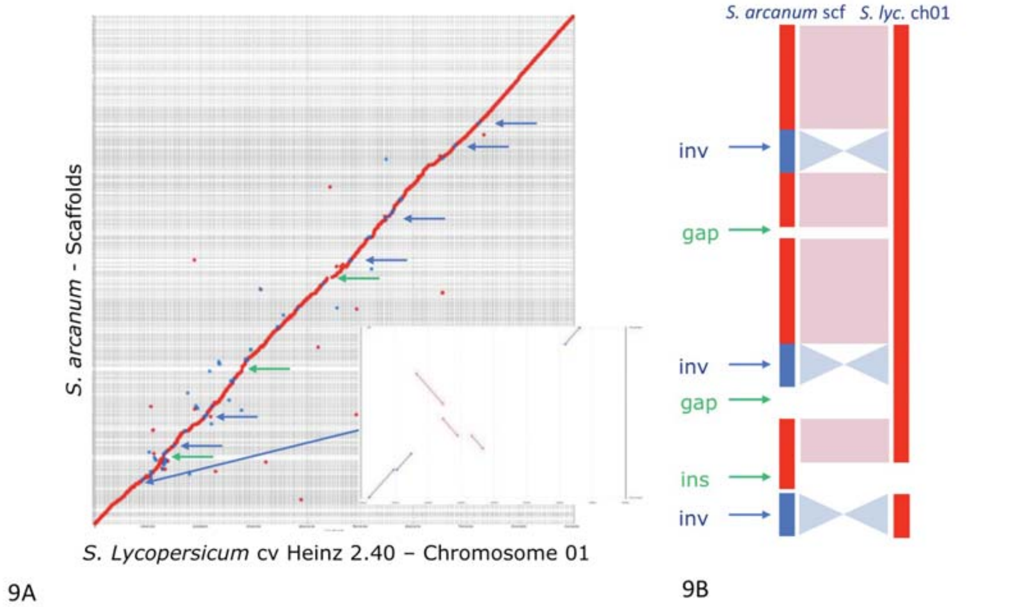


Figure 9. A) Dot plot, comparing scaffolds of the *de novo* assembly of *S. arcanum* LA2157 with its corresponding 50 MB region in the assembly of *S. lycopersicum* cv. Heinz 1706 SL2.40 and B) cartoon explaining the used colours and codes for inversions and Insertions or deletions of large pieces of DNA. The alignment (A) indicates the presence of several small inversions between both genomes (indicated with blue arrows). For clarity, the dot plot of one of such segment is enlarged showing the reversed orientation of three of the contigs. Also, larger gaps (insertions/deletions in *S. arcanum* vs. *S. lycopersicum*; indicated with green arrows) can be observed. From: Aflitos et al. 2014; Finkers pers. communication.

2.4.3 Variation in the analysis yet to be identified

The variation discussed thus far was identified on the basis of the tomato reference genome of *S. lycopersicum* cv Heinz 1706. However, the genetic differences between the tomato wild species and the reference genome are much larger than described above. Two main factors are contributing to this. The first factor is the differences in genome size. Genome size estimates of *S. lycopersicum* and *S. pennellii* are 0.95 and 1.23 pg respectively (http://data.kew.org/cvalues/). Sequences that do not exist in the reference genome of Heinz, could not be compared to Heinz in the described analysis. The second factor is that sequence reads were mapped rather stringently. If more than two mismatches occurred, reads were discarded even though this might have represented valid sequence variation. The reason for these strict criteria is the computational time required to analyse these datasets, which increases exponentially when allowing more variation.

Whereas 96 % of the reads from *S. lycopersicum* accessions could be mapped to the reference genome of cv Heinz, which also belongs to *S. lycopersicum,* only 53% of the reads from the wild species could be mapped to this reference genome. This implies that nearly half of the number of reads was deviating too much for being mapped. These reads were not included in the SNP frequency estimations shown above. Therefore, the genetic variation between *S. lycopersicum* and the sequenced wild species is far larger than suggested by the SNP counts. At the same time it should be kept in mind that all (re)sequenced wild species are crossable with *S. lycopersicum*, and can be used by breeders in conventional breeding programs, and therefore belong to the baseline.

The *de novo* assemblies of three genomes of wild species created three new reference genomes that can be used for mapping reads from other species (Fig. 10). The read mapping frequency increased when reads from wild species were mapped against a new reference genome from a closer relative. For *S. arcanum* and S. *habrochaites* 72.87%±7.87% and 78.74%±15,63% of the reads, respectively correctly mapped against the *S. arcanum* LA2157 and *S. habrochaites* LYC4 reference genome, whereas 55.37%±9.29% of *S. pennellii* LA1272 reads correctly mapped to the *S. pennellii* LA716 reference.

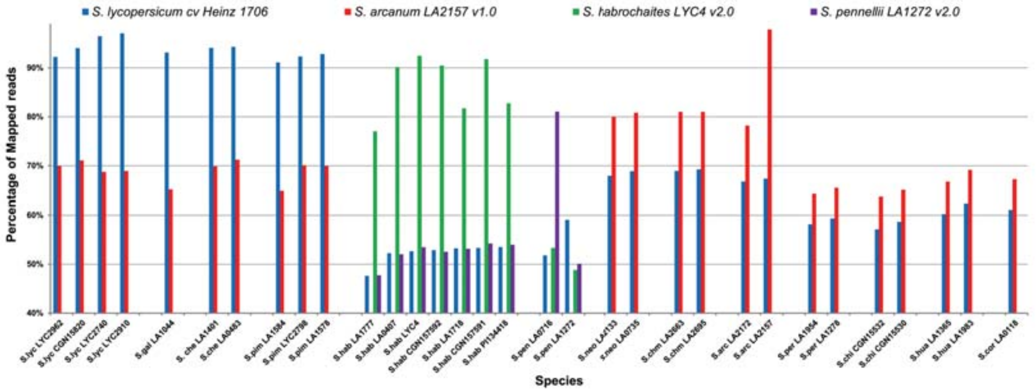


Figure 10 Percentage of reads from *S. lycopersicum*, *S. arcanum*, *S. habrochaites* and *S. pennellii* accessions mapped against four reference genomes. From: Aflitos et al. 2014; Finkers pers. communication.

We conclude that, when using the existing tomato cultivars as the baseline, the baseline variation is up to 4.5M SNPs, most likely affecting many genes and responsible for the different phenotypic traits. When the sexually compatible wild tomato species that are used in breeding are included in the baseline, the genetic variation and associated SNPs is even higher. Also structural variation occurs, such as large inversions. Such inversion can lead to introgressed chromosomal fragments from wild species in tomato cultivars of more than half a 'wild' chromosome, harbouring the desired introgressed trait, but also many thousands of SNPs and other types of polymorphisms.

2.5 Tomato (literature)

Recently, Lin et al. (2014) re-sequenced 360 tomato accessions in China. After aligning the reads to the tomato reference genome ('Heinz'), they generated a final set of 11,620,517 SNPs and 1,303,213 small indels (insertions/deletions shorter than 5 bp), so ~ 13 M polymorphisms compared to 'Heinz'. Still numerous polymorphisms were not taken into account, as only the SNPs and small indels of the reads that aligned to the reference genome were included. The reads that deviated, and thus harbored a high level of polymorfism, were ignored. Also repetitive DNA was left out. However, the order of magnitude of SNPs from this Chinese project is comparable to the SNP numbers that we detected (Ch. 2.3).

3. The dynamic nature of plant genomes (literature)

3.1 Mutation rates in seed propagated crops

3.1.1 Mutation rates of Arabidopsis thaliana in a greenhouse

The previous chapter describes the genetic variation that accumulated during evolution and was partly lost due to natural selection and/or domestication. However, plant genomes are not static, but change continuously, due to mutations.

In order to gain insight into the mutation rate of seed-propagated plants, Ossowski et al. (2010) re-sequenced the genomes of five *A. thaliana* lines that were derived from one mother plant, and had been maintained in a glasshouse by single-seed descent for 30 generations. By comparison of these five genomes, the mutations that had occurred during the 30 generations could be revealed. Ossowski et al. used the reference strain Col-0, for which a high-quality genome was published in 2000, which was improved further in 2008 (Ossowski et al., 2008). The depth of sequence coverage was between 23X and 31X in each of the five lines, using Illumina sequencing. Base substitutions were called if one of the lines differed from all others. If a sequence deviation occurred in more than one line, this could have been a result of heterozygosity of the starting plant, and these polymorphisms were not regarded as spontaneous mutations during the 30 generations. Ossowski et al. identified and validated as mutations 99 base substitutions and 17 insertions and deletions, so in total 116 mutations. Validation was performed by means of Sanger sequencing. Their results imply a spontaneous mutation rate of 7×10^{-9} base substitutions per site per generation.

For base mutations, there are six possibilities (A:T→T:A; A:T→G:C; A:T→C:G; C:G →G:C; C:G→T:A; C:G→A:T). The occurrence of these six possibilities appeared to be very biased towards C:G→T:A transitions. This type of base substitution happened 7 times more frequently than the other base substitutions. The authors explain this by deamination of methylated cytosines and ultraviolet light–induced mutagenesis (Ossowski et al., 2010).

The estimated rate of small deletions (1-3 bp in size) was 0.6×10^{-9} per site per generation, and for small insertions 0.3×10^{-9} per site per generation. So small deletions occurred twice as frequent compared to small insertions. Deletions larger than 3 bp occurred nearly as frequent as small deletions, i.e. at a frequency of 0.5×10^{-9} per site per generation. They removed on average 800 ±1900 bp per event. This implies a gradual, small decrease in genome size over the 30 generations.

The authors looked at the frequencies of mutations in intergenic regions, introns, synonymous substitutions in coding regions, non-synonymous substitution in coding regions, shift of reading frame for short indels, or gene deletion for large deletions, in untranslated regions (UTRs) and transposable elements. Assuming that only nonsynonymous mutations and indels affecting coding regions are likely to affect fitness, the diploid genomic rate of mutations affecting fitness would be 0.2 ± 0.1 per genome per generation. This estimation is higher but not significantly different from the estimated deleterious mutation rate of 0.1 by Schultz et al. (1999), who measured fitness decrease at the phenotypic level after 10 generations of accumulation of spontaneous mutations in 1,000 inbred lines of the *A. thaliana*.

Summarizing, Ossowski et al. (2010) estimated that the mutation rate in seed-propagated *A. thaliana* plants is approximately 7 × 10^{-9} base substitutions and 1.4 × 10^{-9} indels per site per generation. This is equivalent to 2.3 spontaneous mutations per plant per generation.

3.1.2 Activity of transposable elements in rice

In addition to 'copying errors' in the DNA during cell divisions leading to SNPs and small indels, and unequal crossing-overs during meiosis leading to large insertions or deletions, another class of mutations is caused by activity of transposable elements. Transposable elements constitute a large portion of eukaryotic genomes. Many studies have been performed on transposable elements. Here an example is briefly mentioned. Based on whole genome sequences of two rice genomes, Jiang et al. (2003) showed activity of a DNA transposon called *miniature Ping* (*mPing*). The size of this transposable element is 430 bp and it is present in 70 copies in one sequenced genome (*O. japonica*) and 14 copies in the other sequenced genome (*O. indica*). In cells that showed activity of *mPing,* activity of another family of transposable elements was found too, called *Pong*. New inserts of both families were especially detected in low copy regions of the rice genome. Since the domestication of rice *mPing* was amplified preferentially in cultivars adapted to environmental extremes (Jiang et al., 2003). Kikuchi et al. (2003) observed the efficient excision of *mPing* and reinsertion into new loci in the rice genome. Naito et al. (2006) showed bursts of the *mPing* elements from ≈ 50 to ≈1000 copies in rice genomes. In spite of these bursts, these transposable elements did not kill their hosts, and appeared to be significantly underrepresented in exons and introns. Remarkably, *mPing* showed a preference for insertion into 5' flanking sequences of genes, sometimes leading to up-regulation of these genes, and sometimes render them stress inducible (Naito et al., 2009).

3.1.3 Horizontal transfer of transposable elements

El Baidouri et al. (2014) analyzed 40 sequenced eukaryotic genomes, representing the major plant families, and detected 32 cases of horizontal transfer of retrotransposons between sexually incompatible species, including monocots and dicots, such as between palm and grapevine, tomato and bean, or poplar and peach. When extrapolating to all plant species, this would imply more than 2 million events of horizontal gene transfer of retrotransposons between plants. They also showed that these TEs have remained functional after their transfer, occasionally causing a transpositional burst. This suggests that plants can relatively frequently exchange genetic material through horizontal transfer. The authors did not investigate the underlying mechanisms of this horizontal gene transfer, but name different possible causes, such as pathogens that may play a role as vector of TEs, but this is still rather speculative.

3.2 Somaclonal variation

3.2.1. Somaclonal variation in *Arabidopsis*

Somaclonal variation is a phenomenon that results in the phenotypic variation of plants regenerated from tissue culture or cell culture. It can be a result of activity of retrotransposons or other genetic mutations. It may also be caused by epigenetic modifications, such as adding or removal of methylation at DNA sites (Müller et al., 1990; Smulders and de Klerk 2011). Here we focus on mutations in the DNA sequence itself, observed by whole genome sequencing.

Jiang et al. (2011) showed that the mutation rate in *Arabidopsis* regenerated from *in vitro* culture was 60 to 350 times higher compared to the mutation rate observed in sexually propagated *Arabidopsis.* These regenerated plants were not transformed nor treated with *Agrobacterium.* The researchers observed a distinctive molecular pattern of base substitutions, insertions and deletions, differing from the mutation pattern of seed propagated plants. This research is described in more detail below.

Jiang et al. generated regenerants of Arabidopsis via a two-stage culture of root explants. First, callus was formed on roots grown on auxin-rich medium. Subsequently, from the callus cell mass, new roots and shoot were induced on medium with specific auxin and cytokinin concentrations. The researchers produced 28 regenerants from a single *Arabidopsis* (Col-0) root. These regenerants were self-pollinated, to generate R1 families. Variant phenotypes were detected in 8 of the 28 lineages. Two of these were not stably heritable, suggesting an epigenetic cause. However, in six lineages (so 21% of the regenerants), phenotypes were stably heritable and segregated within the R1 families, with segregation ratios approximating to Mendelian expectations for single-gene recessive mutations. Five individual R1 plants and the progenitor P1 plant were sequenced, and analyzed for changes in the R1 plants compared to the P1 plant. They detected 152 mutations, comprising 131 single-base substitution (SBSs) and 21 small indels (≤ 2 bp). The mutations were evenly spread among the chromosomes. The authors detected only the homozygous mutations, as a consequence of their filtering of reads. They expected that only about 25% of regenerant mutations would have become homozygous in the R1 generation. Therefore they estimated the actual number of regenerant mutations by multiplying the detected mutations by four (Jiang et al., 2011). Taking this correction into account, the somaclonal mutation frequencies varied between 4 x 10^{-7} and 24 x 10^{-7} mutations per site in the five regenerant lineages. The mutation rate thus increased between 60 x and 340 x in the regenerant lineages compared to the mutation in sexually propagated *Arabidopsis*, which was 7 x 10^{-9} (Ossowski et al. 2010).

The most common category of regenerant mutations were single-base substitutions. Jiang et al. subdivided these SNPs in mutations between the purines (A, G) or between the pyrimidines (C, T) called transitions, and mutations from purines to pyrimidines or reverse, called transversions. Whereas the frequency of transitions was similar to the frequency of transversions in the regenerants (transitions : transversions = 63 : 68 = 0.92), Ossowski et al. (2010) detected a 2.4 higher frequency of transitions compared to transversions. This bias towards transitions in sexually propagated Arabidopsis was caused by a relatively high frequency the change from C:G to T:A (Ossowski et al. 2010). This type of transition was not biased in the regenerants of Jiang et al.

In addition, regenerant plants carried an elevated frequency of indel mutations compared to sexually propagated plants. Nearly all regenerant indels occurred in the number of repeats of a mononucleotide (AAA… or TTT…) or dinucleotide SSR (ATATAT.. or CTCTCT…). These are probably a result of slippage of the DNA polymerase during DNA replication (Schlötterer, 2000). However, next generation sequencing is relatively poor in determining indels in simple sequence repeat regions, and therefore not very suited for reliable estimations of changes in number of repeats in these stretches.
Jiang et al. also searched for structural variants, but in spite of exhaustive searches, no large indels or chromosomal rearrangements were detected in any of the five paired-end sequenced R1 plants. Although the Arabidopsis genome contains multiple mobile genetic elements, Jiang et al. did not detect any activity of these transposons in the regenerants.
Out of the 152 mutations in the regenerants, 29 occurred in protein-coding sequence. All these mutations were SNPs. Out of these 29 mutations, 17 were non-synonymous mutations that altered the amino acid sequence of proteins. As mentioned, eight regenerants showed stable phenotypic changes in their lineages. Jiang et al. deciphered molecularly the non-synonymous mutations in the genes, causing the mutant phenotypes. None of the indels affected protein-coding sequence, suggesting that SNPs may be a major genetic cause of somaclonal variation.

The callus phase growth and/or in vitro regeneration from tissue culture might be inherently mutagenic. This is supported by previous observations that somaclonal variant phenotype frequencies increase in proportion to the duration during which cells are maintained in tissue culture (Lee et al., 1987).

3.2.2. Somaclonal variation in rice

Miyao et al. (2012) analyzed the whole-genome sequences of three rice plants that were independently regenerated from a cell culture that originated from a single seed stock. The period of cell culture was 5 months. Many SNPs and indels were detected and validated in the genomes of the three regenerated plants. The frequency of base substitutions was estimated to be 1.74 x 10^{-6} per site per regeneration. This is nearly 250 times higher than the base substitution frequency of 7 x 10^{-9} in the 30 sexual generations experiment of Ossowski et al. (2010). Ignoring that Miyao et al. worked with rice and Ossowski et al. with *A. thaliana*, these experiments indicate that base pair substitution occurs far more frequently during cell culture and/or regeneration than during seed propagation.
The observed molecular spectrum was similar to that of the spontaneous mutations in *A. thaliana* (Miyao et al. 2012). On top of these base substitutions, Miyao et al. studied the activity of transposons. Among the 43 examined transposons, only one appeared to be active, i.e. Tos17. In one regenerated line, 10 new insertions of this transposon were detected.

In an earlier study by Hirochika et al. (1996), the transposition of *Tos17* during tissue culture was already reported. The longer the period rice had been in tissue culture, the more transpositions of this endogenous copia retrotransposon were detected. The transcript (the RNA) of *Tos17* was only detected under tissue culture conditions. This indicates that the transposition of Tos17 is mainly regulated at the transcriptional level (Hirochika et al. 1996). The insertion sites of *Tos17* appeared to be preferably in coding regions, probably leading to deleterious mutations.

Sabot et al. (2012) also studied the activity of transposable elements (TEs) during tissue culture of rice, using paired-end deep sequencing. Although most TE-related sequences correspond to inactive, degenerated elements due to silencing and deletions, at least 13 TE families appeared to be active in the used genotype, causing 34 new insertions. So, not only the retrotransposon family *Tos17* showed transpositions during this study, but also other TE families. Lin et al. (2012) showed dramatic differences between rice cultivars for tissue culture-induced mobility of *Tos17*: Two rice cultivars showed mobilization of *Tos17*, the third line showed total immobility of the element, and the fourth line did not contain the element. They made crosses between these cultivars, and observed that immobility was dominant over mobility. They conclude that the tissue culture-induced mobility of *Tos17* in rice is under complex genetic and epigenetic control (Lin et al., 2012).

Ngezahayo et al. (2009) showed that the activation of the transposable element *mPing* in rice during tissue culture is correlated with changes in cytosine methylation at the elements' flanks and random genomic loci. Cytosine methylation is a major epigenetic modification in most eukaryotes, serving as a genome defense system including taming activity of transposable elements (Ngezahayo et al. 2009). Reduction of methylation during tissue culture or other stress conditions may therefore activate transposable elements, such as *mPing.*

Sabot et al. (2011) studied the transpositional landscape of a rice mutant derived from an *in vitro* callus culture. The activity of the TEs was studied through paired-end mapping of a fourfold coverage of the genome using Illumina technology. They showed that at least 13 TE families were active, causing 34 new insertions. *mPing* was one of these 13 TE families.

An illustrative example of the impact of somaclonal variation has been unintentionally provided by a T-DNA insertion mutant library for rice (Chang et al, 2012): After the completion of the rice genome-sequencing project, the rice research community proposed to characterize the function of every predicted gene in rice by 2020. In view of this goal, large-scale T-DNA insertion mutant populations were generated in China, consisting of 372,346 mutant rice lines. Phenotyping of these lines, and determination of the loci where the T-DNA was inserted, could unravel the biological functions of the DNA sequences that were disrupted by the T-DNA inserts. However, the generation of rice T-DNA insertion mutations by *A. tumefaciens*-mediated transformation involves a tissue culturing process, and hence gives somaclonal variation, on top of the mutations due to the T-DNA insertions. The authors describe that less than 5% of the observed phenotypic alterations were actually caused by T-DNA insertions (Chang et al., 2012). This illustrates that the great majority (> 95%) of the phenotypic changes in T-DNA rice lines were caused by somaclonal variation, and only a small minority by disruption of genes by T-DNA inserts.

Summarizing, genome wide mutations occur during tissue culture and cell culture, like during seed propagation. Although the types of mutations were similar, the frequency was much higher compared to seed propagation, e.g. about 250 times higher. Also transposons were more active, enhancing the somaclonal variation. As the genetic modification process contains usually a tissue or cell culture phase and a regeneration phase, these phases can contribute to increased frequencies of spontaneous mutations compared to seed propagation.

4. Mutations associated with transformation (literature)

There are only a few studies on whole genome sequences of transformed plants. There are a few studies on transgenic rice and one study on transgenic papaya. Although we focus this review to Arabidopsis, rice and tomato, the papaya case is still included here, due to the absence of similar data for Arabidopsis and tomato in the current literature.

4.1 Transgenic rice

The whole genome of a transgenic rice line, obtained by *Agrobacterium*-mediated transformation, was sequenced by Kawakatsu et al. (2013). They aligned the reads to the sequence of the used *Agrobacterium* C58 strain, but did not detect any reads that aligned with the *Agrobacterium* sequence, indicating that no portions of the *Agrobacterium* genome had been integrated into the transgenic rice. Further, they mapped the reads against the Nipponbare reference genome of rice, and screened for structural variant (SV) candidates in the transgenic rice line by comparing the computed distances of the paired-end reads with their locations within the Nipponbare reference. A total of 149 insertions, 280 deletions, 12 inversions, and 39 instances of inter-chromosome swapping were called. However, manual validation of these potential SVs revealed that all of the detected insertions, inversions, and inter-chromosome swappings were false positives. This underlines the importance of careful checking of SV candidates.

Also they compared the genome of the transgenic rice with the sequence of the parental genome. This revealed one deletion of 810 bp in the transgenic line in a region with transposon-like elements. This deletion could not be explained by transposon translocation. In the 10-kb region around the deletion no annotated genes were detected, and the authors regard this as a result of somaclonal variation, and not caused by *Agrobacterium* infection. They detected also 28 small deletions/insertions of 1 or 2 bp and one of 5 bp, mainly in mono- or dinucleotide repeats. Usually, changes in the number of microsatellite repeats (SSRs) are attributed to slippage during DNA replication (see above).

Further they detected 167 SNPs spread among all chromosomes, when comparing the transgenic line with its parent. When comparing the transgenic line with the Nipponbare reference genome of rice, 93110 SNPs were found, which is more than 500 time higher than the 167 SNPs between the transgenic line and its parent.

The total number of mutations between the transgenic line and its parent was 196, which means a mutation rate of 5.5 x 10^{-7}. These mutations were distributed among the whole genome and did not preferentially accumulate near the T-DNA insertion site. The pattern of nucleotide base substitution in the transgenic line was consistent with somaclonal variation (Kawakatsu et al. 2013).

As discussed in Chapter 3, Miyao et al. (2012) reported a somaclonal mutation rate of 1.7 x 10^{-6} in rice after 5 months of cell culture, which means a mutation rate of 0.85 x 10^{-7} per week *in vitro* culture. In the rice transformation process by Kawakatsu et al.

the *in vitro* period lasted approximately 2 months (callus induction: 1 month, *Agrobacterium* infection: 3 days, and selection: 1 month). The mutation rate was 0.68 x 10^{-7} per week. This is similar to the rate detected by Miyao et al. *in vitro* cultures of rice, although Kawakatsu et al. also transformed the plant, whereas Miyao et al. and Jiang et al. only applied *in vitro* cultures without transformation. Based on the mutation rate and on the mutation pattern among the whole genome, Kawakatsu et al. conclude that the mutations detected in the transgenic rice line were caused by somaclonal variation during *in vitro* culture. No genomic damage, caused by *Agrobacterium* infection was detected, other than the T-DNA insertion.

The mutations in the transgenic rice line, as compared to its parent, changed the amino acid sequences of 11 genes (Kawakatsu et al., 2012). The amino acid sequences of the japonica rice cultivars Koshihikari and Omachi differed in 1017 and 794 genes when comparing to Nipponbare. This underlines that the natural variation between cultivars was far higher than the mutation frequency due to somaclonal variation in the transgenic line. In spite of the changes in amino acid sequences in 11 genes, Kawakatsu et al. did not observe phenotypic changes in the transgenic line, although they admit that some phenotypic changes under specific circumstances could not be excluded.

Kawakatsu et al. (2012) also compared the transcriptomes of the transgenic line and its parent. As the transgenic line has been meant as a seed-based edible vaccine against Japanese cedar pollinosis and contains, as GM, two major pollen allergens, Cry j 1 and Cry j 2, the authors took mRNA from developing endosperm. Since activities of promoters used for transgene expression were highest at 10 – 20 days after flowering, they chose this stage for the analysis. They identified 28 genes that were differentially expressed in the transgenic line compared to the parent. Two of these genes were the inserted transgenes. The authors explained the increased expression of 11 genes as a reaction on the accumulation of recombinant proteins. One gene that showed higher expression was 27 bp downstream the T-DNA insert, suggesting that the insertion caused this increased expression. The researchers detected no SNPs or small indels within 2 kb upstream of any differentially expressed gene, suggesting that DNA mutations had little effect on the endosperm transcriptome. However, epigenetic modifications or downstream effects of some polymorphic genes leading to differential expression might be possible (Kawakatsu et al., 2012).

4.2 Transgenic papaya

Ming et al. (2008) sequenced the genome of the virus-resistant tropical fruit tree papaya (*Carica papaya*), containing insertions of the transgene coding for coat-protein of the papaya ringspot virus. They analysed the sites of the three inserts that resulted from biolistic transformation. Remarkably, five of the six flanking sequences of the three insertions appeared to be nuclear DNA copies of papaya chloroplast DNA fragments. The integration of the transgenes into chloroplast DNA-like sequences may be related to the observation that transgenes produced either by *Agrobacterium*-mediated or biolistic transformation are often inserted in AT-rich DNA, as is the chloroplast DNA of papaya and other land plants (Ming et al., 2008). The alternative explanation that papaya chloroplast DNA was co-inserted into the nuclear DNA, together with the transgene, was not mentioned by the authors. As the authors did not compare the genome of the transgenic papaya to the genome of the non-transgenic parental papaya, this alternative explanation could not be verified.

A second remarkable observation was that four of the six insert junctions had sequences that match topoisomerase I recognition sites. During replication the double stranded DNA is unwound, but becomes overwound ahead of the replication fork, leading to tension that could eventually halt DNA replication. To overcome this problem, topoisomerase cuts one or both strands of the DNA double helix, allowing turning of the DNA and removal of rotation tension. After the rotation, the topoisomerase re-anneals the cut strands. Apparently, the transgene was preferably inserted at these topoisomerase I recognition sites, taking advantage of the breakpoints during DNA replication.

As the non-transgenic genome was not sequenced, the authors did not describe changes in the remainder of the genome, more distant from the inserted DNA. So, mutations due to transformation could not be detected, apart from the inserted transgenes. The authors looked for fragments of the transgene and the vector in the transgenic papaya, but apart from the three mentioned insertions they could not provide conclusive evidence on presence or absence of such fragments in the transgenic papaya.

5. Genetic changes due to somaclonal variation and transformation in tomato and Arabidopsis (experiments)

5.1 Goals of experiments

Experiments were carried out with tomato and Arabidopsis, with the following aims:

1. Evaluate changes in the genome due to tissue culture and regeneration (= somaclonal variation);
2. Evaluate changes in the genome due to transformation;
3. Compare frequency of changes in the genome due to transformation with the baseline of variation in breeding germplasm, commercial cultivars, and somaclonal variation.

These experiments have been performed with two plant species, i.e. tomato and *Arabidopsis thaliana.*

5.2 Materials and Methods

5.2.1 Tomato

Experimental setup

The experimental setup is illustrated in Fig. 1.

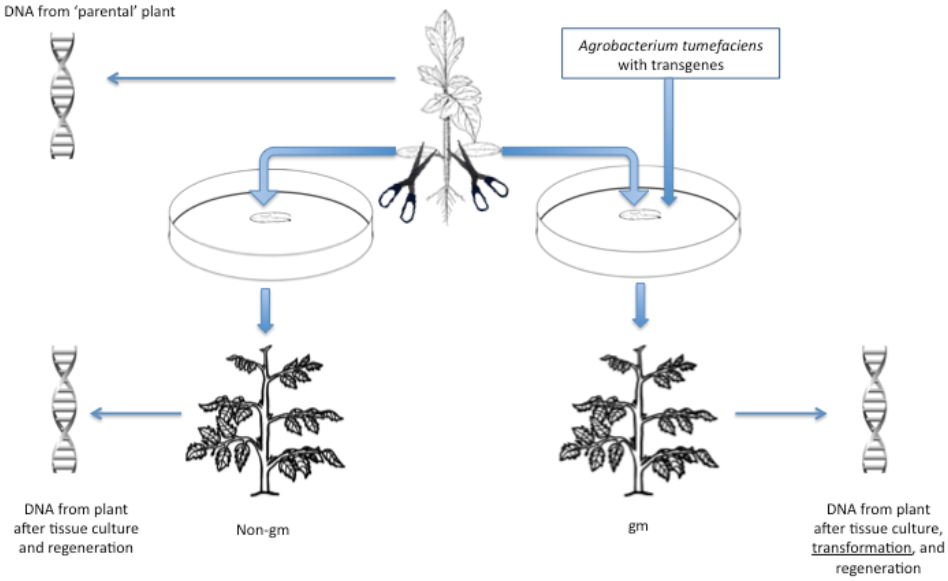

DNA from 'parental' plant
Agrobacterium tumefaciens with transgenes
DNA from plant after tissue culture and regeneration
Non-gm
gm
DNA from plant after tissue culture, transformation, and regeneration

Figure 11. Cartoon of one replication of the tomato experiment.

The experiment was performed in three replicates, leading to three 'trios', totally 9 plants.

Parental seedlings

To obtain seedlings for the transformation and regeneration experiment, seeds harvested from a single plant of the tomato cultivar Moneymaker were used. One hundred seeds were disinfected in 1% NaOCl and germinated on germination medium which is MS-medium (Murashige and Skoog, 1962) at half strength, 10 g/l sucrose and 8 g/l micro agar in plastic containers. Each container contained 10 seeds that were all individually numbered. For synchronisation of germination the containers with seeds were incubated for 4 days at 4°C in the dark after which they were transferred to 24°C (16h light, 8 h dark).

Ten days after transfer to 24°C, the seeds had germinated and from each seedling the two cotyledons were harvested and cut in 2x2 explants. The seedlings were kept for further development of leaf tissue for DNA isolation. Cotyledon explants originating from the same seedling were marked with the same number as the seedling. The explants from one cotyledon of a seedling were used for transformation and explants from the other cotyledon were used for regeneration.

Transformation

For transformation, the explants were placed with the adaxial side down on shoot induction medium (SIM; MS salts, Nitsch vitamins (Nitsch and Nitsch, 1969), 30 g/l sucrose, 6 g/l micro agar, 1.5 mg/l zeatin-riboside and 0.2 mg/l IAA) medium which was supplemented with acetosyringone (100 μM final concentration) (50 explants/plate) and precultured for 2 days at 24°C under 16 h light/8h dark condition. After preculture, the cotyledon explants were infected with an *Agrobacterium tumefaciens* strain (AGL1; Lazo et al., 1991) harbouring the binary plasmid pBINplus (van Engelen et al. 1995) by incubating the explants for 15-20 minutes in the *Agrobacterum* suspension (OD 600nm = 0.3). As T-DNA, the vector only contained the *nptII* gene for kanamycin resistance under control of the *nos* promoter and the *nos* terminator. The T-DNA was flanked by the T-DNA borders (van Engelen et al. 1995). The size of the T-DNA was 3404 bp.

After infection the explants were placed back on the SIM plates and co-cultured with *Agrobacterium*. After a co-cultivation period of 2 days at 24°C, 16 h light/8h dark, the explants were transferred to fresh SIM which was supplemented with 50 mg/l kanamycin for selection of transgenic regenerants and 125 mg/l cefotaxim for elimination of Agrobacteria and cultivated at 24°C, 16 h light/8h dark. After three weeks the explants were transferred to fresh medium and three weeks later shoots that regenerated on the cotyledon explants were harvested and propagated in containers with shoot propagation medium (MS salts and vitamins, 30 g/l sucrose, 8 g/l micro agar), which was supplemented with 50 mg/l kanamycin and 125 mg/l cefotaxim.

Regeneration

For regeneration of the explants without a transformation treatment, the cotyledon explants were placed with the adaxial side down on SIM, which was refreshed after 3 weeks, similarly as the explants that were meant for transformation. Shoots that appeared on the cotyledon explants were harvested and propagated in containers with shoot propagation medium (MS salts and vitamins, 30 g/l sucrose, 8 g/l micro agar).

DNA isolation

Genomic DNA was isolated from ‘in vitro’ grown leaf material from three different seedlings and from the transformants and regenerants that were regenerated from cotyledons from these seedlings using a CTAB-based DNA isolation method basically as described by Doyle and Doyle (1987).

Figure 12 shows some illustrative pictures of the experiment.

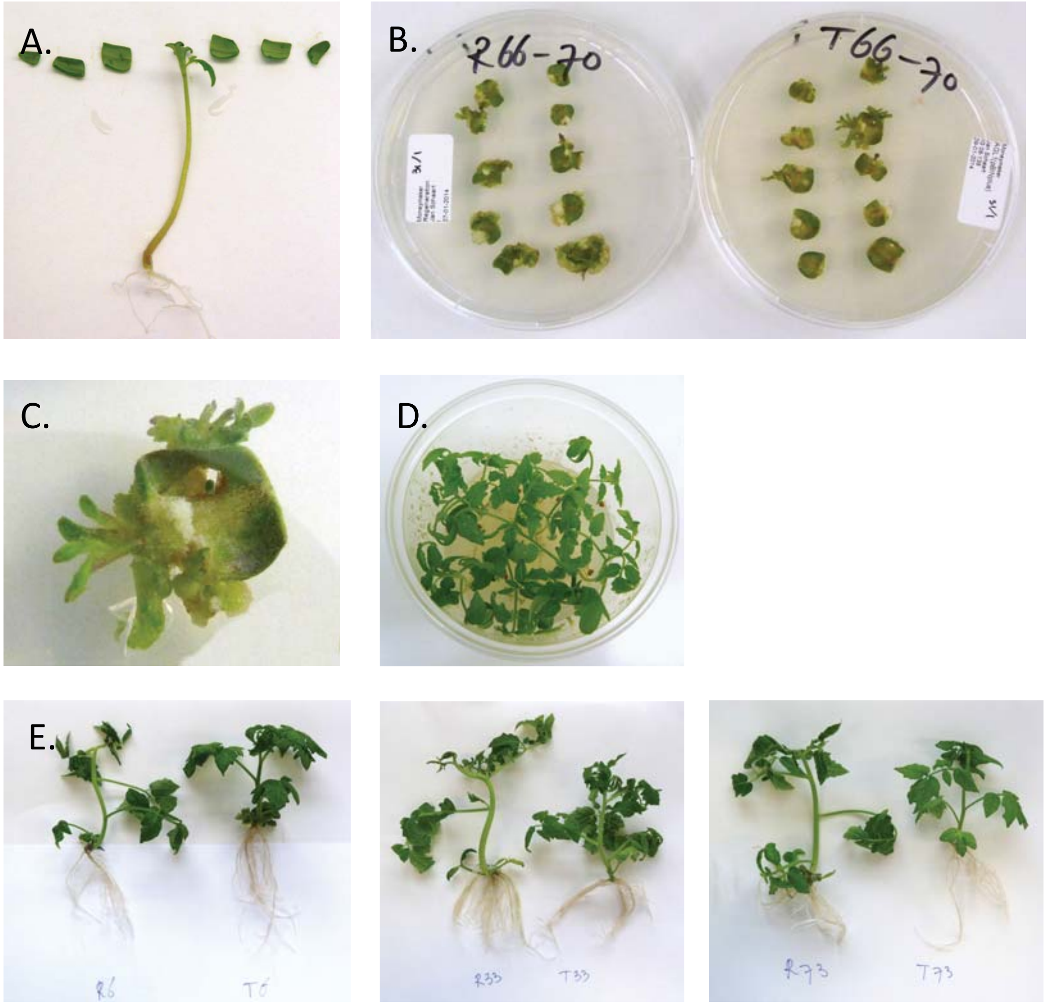


Figure 12. Transformation of tomato. A. Cotyledons are cut in pieces; B. From the four cotyledons pieces (explants) per plant, two were used for regeneration and two for transformation + regeneration; C. Detail of explant with regenerating tissue; D. Parental plants that continued growing after removal of the cotyledons; E. Pairs of regenerants and transformants per parental plant in three biological replicates.

5.2.2 Arabidopsis

The experimental setup for the experiment with *A. thaliana* is depicted in Figure 13.

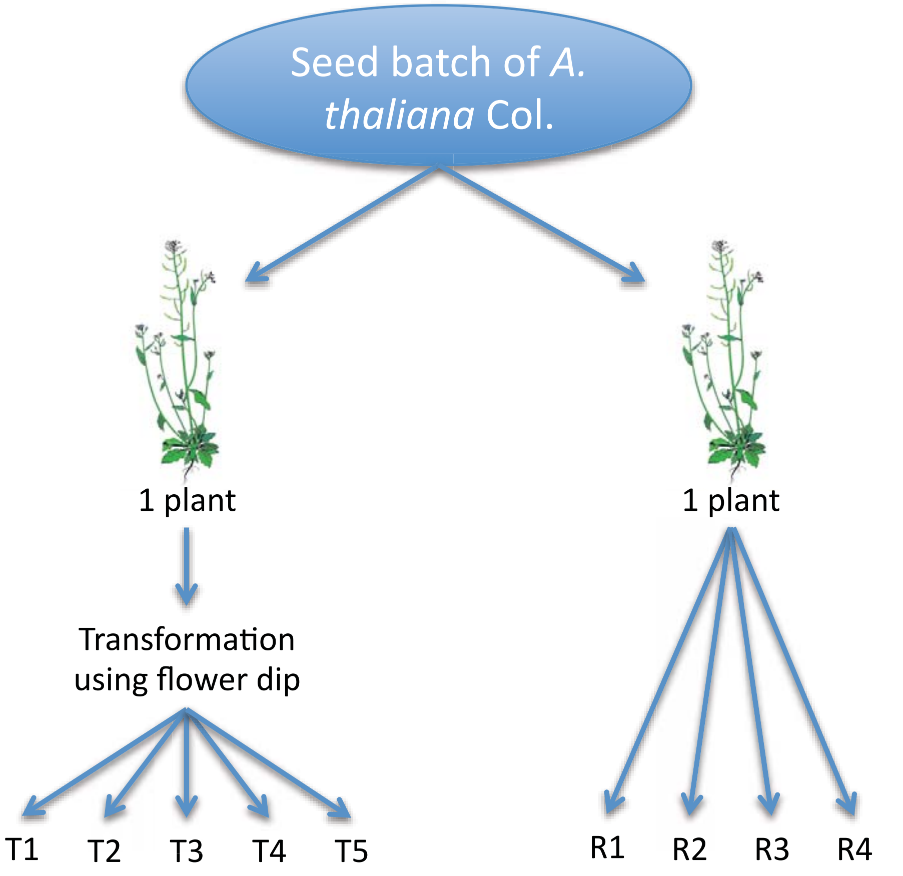

Seed batch of *A. thaliana* Col.
1 plant
1 plant
Transformation using flower dip
T1
T2
T3
T4
T5
R1
R2
R3
R4

Figure 13. Setup of the experiment for estimation of the mutation frequency after transformation using flower dip, compared to seed propagation without transformation.

Generation of the construct

A reporter construct was generated with the 3.7 kb promoter region of the *A. thaliana* gene SAUR8 (AT2G16580) using the GatewayTM system (Invitrogen). The fragment was amplified from Col-0 genomic DNA, and recombined into pDONR207. The entry vector was subsequently recombined with the destination vector pBGWFS7 (basta resistance) (Karimi et al., 2002). The resulting vector was transformed into *Agrobacterium tumefaciens* strain LBA4404 using electroporation (Weigel & Glazebrook, 2006). The size of the T-DNA was 8379 bp.

Floral dip transformation

For floral dip transformation, Col-0 seeds were sown in square pots. Transformation with the construct containing *Agrobacterium* suspension was subsequently performed using the floral dip method as described by Clough and Bent (1998). Seeds were harvested from single plants, and sown separately on 1/2MS plates (pH 5.8), containing 9 g/l agar and 15 mg/l basta (phosphinothricin). One of the parent plants gave rise to five basta resistant seedlings, which were selected for DNA extraction and sequencing.

Reference plants

From the same seed batch, a plant was grown without floral dip. From this plant, seeds were harvested and sown. From four of these progeny plants, descending from the same mother plant, DNA was extracted for sequencing.

DNA isolation

Genomic DNA was isolated from five transformant descending from one mother, and from four non-transformed seedling descending from one other mother plant, using a CTAB-based DNA isolation method as described by Doyle and Doyle (1987).

DNA from the four non-transformed seedlings was pooled at equal quantities per seedling.

5.2.3 Sequencing

Sequencing

DNA samples were random sheared into smaller pieces with average size of approximately 500 bp using a Covaris E210 sonicator. Sheared DNA fragments were used for preparation of individual indexed libraries suitable for Illumina HiSeq sequencing using the Illumina TruSeq Nano DNA LT Sample Preparation Kit. Quality control of final libraries was done on an Agilent 2100 Bioanalyzer DNA100 chip and concentration was measured using a Qubit fluorometer. Final libraries had an average insert size of approximately 600 to 650 bp. Libraries were pooled and analysed on an Illumina HiSeq 2000 sequencer using 2x100 nt Paired End sequencing. After completion of the sequencing run the pools were de-multiplexed and assigned to original samples using Casava 1.8.2 software.

Data analysis

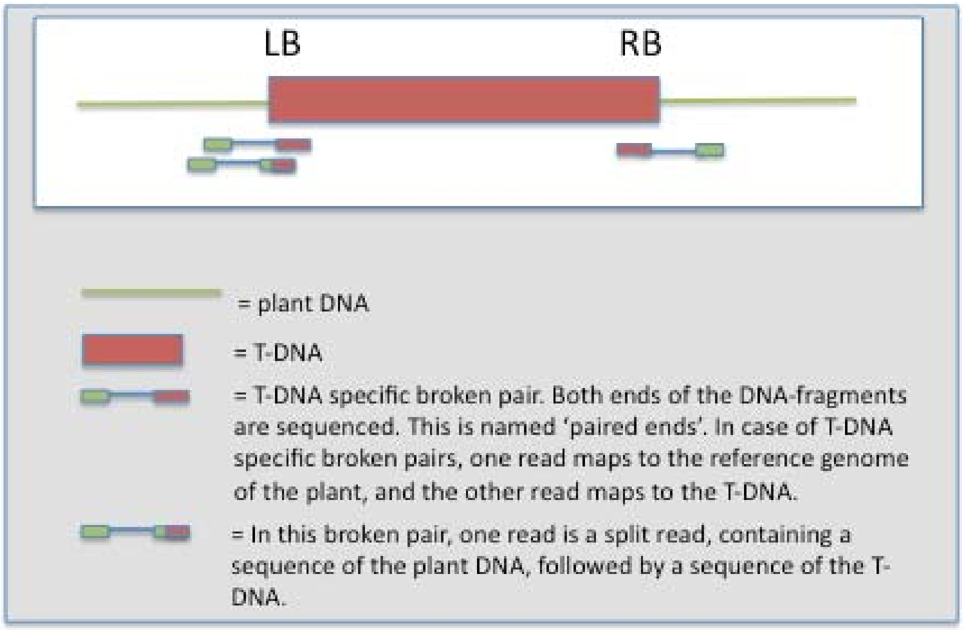
Short sequencing reads require the presence of multiple overlapping segments, referred to as coverage or depth, in order to map and cover efficiently the underlying genomic structure. Short reads NGS technologies require high coverage to increase the chance of producing overlapping sequences. In this project we aimed for at least 25x average coverage over the whole genome per (diploid) sample. For the pooled sample of the four non-transformed *A. thaliana* plants, we aimed for at least 100x genome coverage.

However, short reads have the disadvantage that they are more difficult to map uniquely to a reference genome, especially in highly repetitive areas. This problem can be reduced when additional sequence information is obtained by either longer sequence reads or Paired End sequencing. Paired End sequencing is the production of short sequence reads from both ends of a larger fragment with a well-defined (insert) size. Paired End reads can be mapped more reliably as the reads should map as intact pair to the reference genome, and with a given distance between the two reads within each pair.

Therefore, in this study we generated 100 nt long Paired End sequencing reads derived from both ends of approximately 600 to 650 bp DNA fragments. Large insertions, deletions or rearrangements of a genome, compared to the reference genome, can be detected by analysing:

1. so called split reads. These refer to partially mapped reads;
2. broken read pairs, i.e. reads from a pair that map at an unexpected distance from one another, or on different contigs or in an unexpected orientation.

This is illustrated in the cartoon below (Fig. 14). For analysis of the T-DNA insertions and their positions in the genomes, we filtered the reads that contained vector-DNA sequences. As there appeared to be multiple T-DNA insertions per plant, we searched for split reads and broken pairs that contained sequences of both plant DNA and vector DNA. That allowed to obtain insertion specific sequence information.



= plant DNA

= T-DNA

= T-DNA specific broken pair. Both ends of the DNA-fragments are sequenced. This is named 'paired ends'. In case of T-DNA specific broken pairs, one read maps to the reference genome of the plant, and the other read maps to the T-DNA.

= In this broken pair, one read is a split read, containing a sequence of the plant DNA, followed by a sequence of the T-DNA.

Figure 14. Cartoon, explaining the terms 'broken pair' and 'split read', obtained after analysis of paired-end sequences.

For the analysis of the tomato lines we used the published tomato genome *Solanum esculentum* cv. Heinz 1706 V2.40 as reference. For the analysis of the *A. thaliana* lines we used the published *A. thaliana* genome TAIR10 as reference.
All read mappings, variant callings, comparisons and filtering steps were done using CLC Genomics workbench 7.03 software and / or BWA mappings combined with command line scripts for further downstream filtering.

5.3 Results

5.3.1 Tomato

Sequence data:
After read quality trimming, sequence reads were mapped to the tomato *Solanum lycopersicum* cv Heinz 1706 reference genome, version 2.40. Although this is regarded as a high quality reference genome it should be mentioned that this assembly is incomplete. Whereas the expected tomato genome size is approximately 950 Mb, the total assembly size is only 781 Mb, encompassing 22390 smaller and larger gaps and including over 49 million ambiguous bases (N's). Success rate of read mapping depends on several features including the quality and completeness of the reference genome, similarity between mapped sequences and reference sequence and the complexity of the genome, especially with respect to repetitive sequences.

After mapping, all samples showed an average coverage of at least 25 fold and a similar covered fraction of the reference genome (0.94) indicating the production of highly comparable datasets (Table 1).

For detection of split reads, at least 50% of the read length should align to the reference genome or T-DNA. This allowed detection of putative T-DNA inserts. In the aligning parts, 90% similarity was required, to allow single nucleotide variant (SNV) detection.

Table 1. Statistics of tomato sequence data and read mappings.

sample ID	Sequence reads	reads after trimming	Trimmed Read Length	Covered fraction of reference	Reference bases >10x coverage *	Average coverage
6	222996002	213688868	98.07	0.94	730931825	26.73
6 regenerant	221116204	211687936	98.13	0.94	730910666	26.49
6 transformant	357913790	343463242	98.14	0.94	733996071	43.00
33	219100590	210170862	98.07	0.94	730885479	26.29
33 reg.	302814808	290996544	98.28	0.94	733515961	36.47
33 transf.	298315300	289103420	98.64	0.94	733487901	36.36
73	211928334	202782377	98.0	0.94	730162154	25.35
73 reg.	356629696	344129451	98.42	0.94	734015303	43.2
73 transf.	393851500	378747589	98.22	0.94	734206295	47.45

Nucleotide variant detection specific in transformants

Each replication contained three plants: a parent, a regenerant, and a transformant. We called such a group a trio, and since we had three replications, in total there were nine plants.

For each 'trio' parent-regenerant-transformant we performed variant (SNV) callings of the three plants compared to the tomato Heinz reference genome with different setting using CLC software. As cv Moneymaker was used for the experiment, rather than cv Heinz, a huge amount (on average 1.8 Million) of variants was identified compared to cv Heinz 1706. We repeated multiple mappings with different settings in order to exclude these common variants that were shared by all re-sequenced plants. Comparison of called variants of parent, regenerant and transformant within each 'trio' was subsequently done indirectly (via Heinz as a reference) as well as pair wise.

The general setup contained five subsequent steps as explained below (Figure 15);

1. Stringent settings were used for variant callings of the 'test line' (for example transformant) and less stringent criteria for calling variants within the two other (parent and regenerant) plants.
2. Obtained variants in the transformant were compared to the obtained parental variants, keeping unique transformant versus parent variants only.
3. These transformant versus parent variants were compared to variants found in regenerants. Common variants were excluded resulting in transformant-specific variants only.

4. These transformant-specific variants were additionally filtered using coverage (0.5 to 2 times the coverage compared to the overall genome coverage of the plant), forward/reverse fraction (0.4 to 0.6) and heterozygosity as criteria. Homozygous SNVs were excluded as we used primary transformants, and it is very unlikely that two identical mutations occur simultaneously at a locus on both homologous chromosomes. Homozygous SNVs more likely were inherited from the parental plant.
5. At least ten variants were manually checked to determine the true/false fraction.

It should be remembered that the size of the assembled reference genome is 795 Mb, whereas the expected genome size is 950 Mb. This implies that part of the genome, probably consisting mostly of repetitive elements, is not covered, and any changes in that part of the genome in the regenerants or transformants is not included our analysis. In addition, the approach we used here may overlook large deletions (> 10 bp).

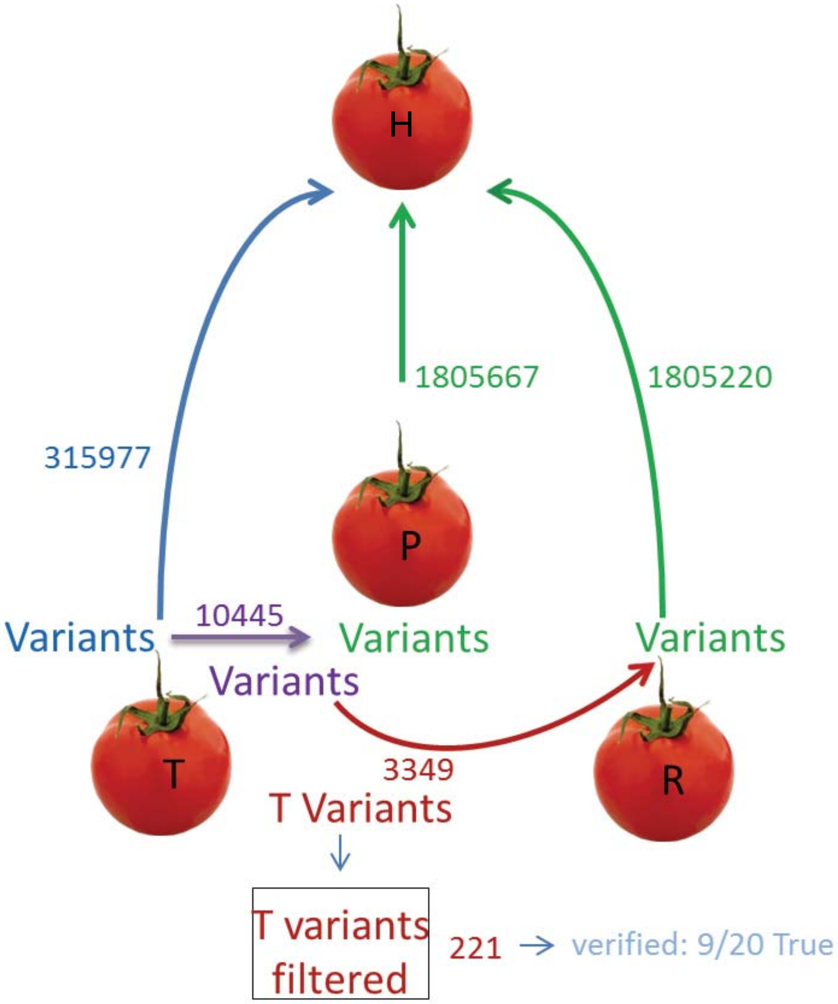


Figure 15. Schematic workflow for variant callings (here for unique variants found in transformant 33) according the following steps: 1) Variants were called using more (blue) or less (green) stringent settings against the reference genome of Heinz, depicted at the top (H). 2) Identified variants were compared between transformant and parent (purple). 3) Subsequent

remaining transformant-parent variants were compared to variants found in the regenerant (red). 4) The resulting transformant-specific ('T') variants were filtered and a subset was manually verified. The shown numbers are for trio 33 only and represent the one-way transgenic unique variants only, as example.

Results of single nucleotide variant (SNV) detection steps from Figure 15 are given below in numbers for all trio plant combinations using corresponding colour codes for each single step. T= Transformant; P = Parent; R = regenerant

Plant	SNVs	Plant	SNVs	Plant	SNVs
6 T	311664	6 P	1797779	6 R	1794220
33 T	315977	33 P	1805667	33 R	1805220
73 T	315377	73 P	1795476	73 R	1805220

Results of variant detection step 2: Identified variants were compared between transformant and parent (purple).

Transformant compared to Parent	SNVs
6 T relative to 6 P	10751
33 T relative to 33 P	10445
73 T relative to 73 P	15862

Results of nucleotide variant detection step 3: Variants called uniquely in transformants compared to parental plants and regenerants:

Trio	SNVs present in transformant (T), not present in P or R	Filtering[1])	Manual check[2])	Estimated unique SNVs in transformants[3])
6	5405	113	3/20 True	17
33	3349	221	9/20 True	99
73	5867	881	16/20 True	705

[1]) Filter settings: 0.5 < average coverage < 2, Forward/reverse fraction 0.4 to 0.6, exclude homozygous SNVs.

[2]) Alignments of sequence reads from all plants within each trio were visual inspected at positions of a subset of 20 randomly chosen variants. This resulted in true or false judgement of the corresponding transformant specific variant detection.

[3]) (Number of filtered SNVs) x (Percentage validated true changes) as best estimate for the number of true SNVs.

Nucleotide variant detection specific in regenerants

For the variant callings specific for regenerants we used exactly the same approach as described above for transformant specific variant detection, except that we now compared stringent called variants from the regenerants and compared those to variants found using less stringent settings within the corresponding parents and transformants.

Results of nucleotide variant detection steps are given below in numbers for all trio plant combinations using same colour codes for each single step.

Plant	SNVs	Plant	SNVs	Plant	SNVs
6 T	**1804547**	6 P	1797779	**6 R**	**308971**
33 T	**1814062**	33 P	1805667	**33 R**	**313871**
73 T	**1814062**	73 P	1795476	**73 R**	**313868**

Regenerant compared to Parent	SNVs
6 R relative to 6 P	7386
33 R relative to 33 P	10080
73 R relative to 73 P	12754

Trio	SNVs present in regenerant, not present in P or R	Filtering	Manual check	Estimated unique SNVs in regenerants
6	2100	201	10/20 True	101
33	3044	182	10/20 True	91
73	2647	183	11/20 True	101

The estimations of numbers of SNVs is depicted in Figure 16. The putatively low number of SNVs in transformant 6 and the putatively high number in transformant 73 is remarkable, and requires a deeper analysis.

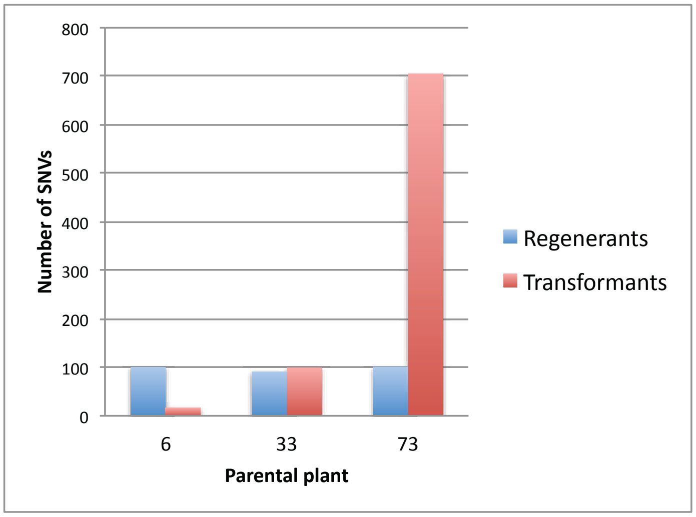


Figure 16. The estimated number of single nucleotide variants (SNVs) in transformants and regenerants, grouped per parental plant.

In Fig. 17 the relative frequencies of the different types of mutations are shown for the regenerants and transformants in pie charts. This figure indicates that the number of indels (insertions/deletions) is higher in the transformants (23%) compared to the regenerants (13%).

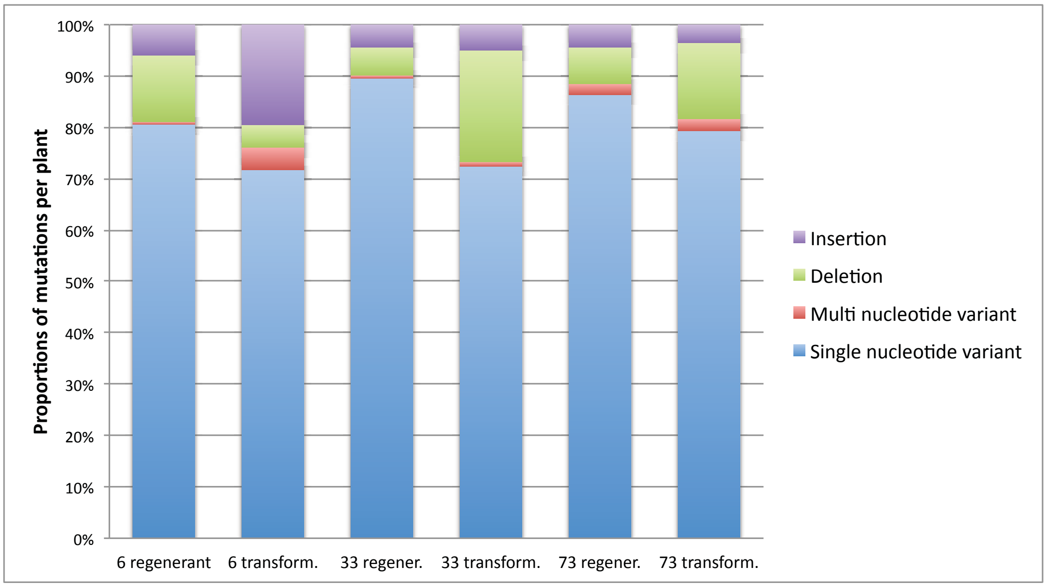


Figure 17. Different types of mutations in the tomato regenerants and transformants.

Detection of T-DNA insertions

To find the locations of the T-DNA insertions in the genomes of the transformants, we mapped the reads to both the tomato cv Heinz genome V2.4, the T-DNA and pBinplus vector backbone. The reads referred to approximately 98 bp sequences from both ends of a DNA fragment of approximately 500 bp. The two ends of one fragment are named ‘paired ends’. When mapping these ends to the reference genome, the far majority of the ends mapped as ‘twins’ at the expected distance of approx. 500 bp on the reference genome. We specifically looked for broken read pairs and partial mapped reads on the T-DNA, and used these reads to locate putative insertions in the tomato genome. In case of a broken pair, one end maps at the reference genome, and the other end at the T-DNA. In case of so-called split reads, part of the read mapped on the reference genome, and part on the T-DNA (Fig. 14).

Although this strategy worked well in case of complete single insertions of the T-DNA from Left border to Right border, still uncertainties existed in case of multiple insertion events, especially when reads were present that may be derived from putatively partial insertion events in the same transformant. Therefore, in addition, mapped read coverage information of the T-DNA was used to detect putative multiple copy numbers and/or partial insertion events.

Read mappings of identified putative insertion locations were visually inspected and judged on coverage, and on presence and frequency of broken pairs and split reads as criteria.

In some cases we detected only broken pairs and/or split reads derived from one end only of the T-DNA construct, indicating the insertion of an invertedly repeated T-DNA or insertion of the border sequence only.

Finally in some cases we detected broken read pairs at loci at a very low frequency. Possibly, these broken pairs are indicating chimeric tissue. A regenerated tomato plant may have originated from more than one cell, differing in T-DNA insertions. Each cell may have led to a plant part, giving chimerism.

Several identified insertion locations were highly reliable. However, also some uncertain positions remained in the transformants. In five cases, the T-DNAs seem to

be incomplete (Table 2), but this is not sure yet. In the current analysis it was not feasible to provide reliable estimates of the T-DNA sizes, and the DNA-sequences of the inserts, due to the fact that there were more than one insertions per plant, making it impossible to univocally assign T-DNA reads to specific inserts. The reads were about 100 bp long, the paired-ends were several hundreds of basepairs apart, but the size of the complete T-DNA was 3404 bp. Further, the transformants contained multiple insertions per plant. This combination made assembly of the individual T-DNA insertions impossible. Only read and read-pairs that contained both T-DNA and flanking plant DNA could be positioned reliably.
A summary of the identified putative T-DNA insertion sites for the three transgenic tomato plants is given below. T-DNA range corresponds to the part of the T-DNA of which there is evidence based on broken read pairs and/or split reads for insertion at given location.

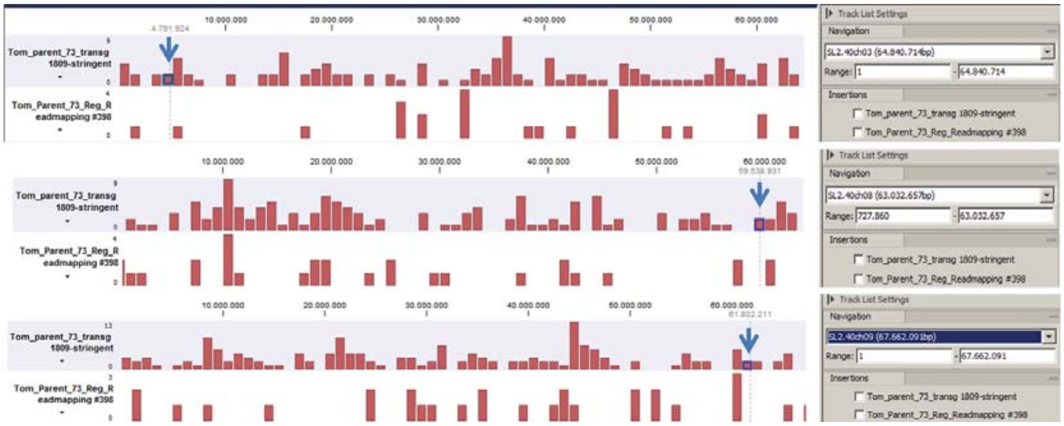
The size of the T-DNA in the vector was 3404 bp, and it contained only the *nptII* gene for kanamycin resistance. The size of the pBinPlus vector backbone was 8993 bp.

Table 2. Putative T-DNA insertions in the three tomato transformants.

Transformant	Approximate positions according analysis of 'broken pairs'	Number of broken pairs / split reads	Putative type of insertion
6	ch02: 46,164,321-46,165,114	36/13	Nearly complete T-DNA.
6	ch06: 32,108,348-32,109,191	42/21	Partial insertion?
6	ch11: 40,216,120-40,217,631	42/21	Partial insertion?
33	ch 02: 32,231,674-32,231,970 Coverage @ insertion= 25x, average=26x	31/12	Probably an inverted repeat
33	ch 03: 45,136,864-46,567,271 Coverage @ insertion= 37x, average=36x	5/0	Partial insertion? Chimeric plant?
33	ch 07: 23,325,727 Coverage @ insertion= 48x, average=32	5/0	Chimeric plant?
33	ch 07: 54,309,452 Coverage @ insertion= 67x, average=36	3/0	Chimeric plant?
73	ch03: 4,758,089-4,759,408	97/24	Complete T-DNA
73	ch08: 59,547,256-59,548,590	92/36	Partial insertion?
73	ch09: 61,831,503-61,833,024	74/20	Partial insertion?

Spatial distribution of T-DNA insertions and mutations among the genomes

Due to the relatively high number of SNVs identified in transformant 73, we investigated whether these variants were clustered near the identified insertion sites. Figure 18 shows that the distribution of variants over the genome appeared to be random. No clear clustering of variants at or near T-DNA insertion sites was observed.



Track List Settings
Navigation
SL2.40ch03 (64.840.714bp)
Range: 1 - 64.840.714
Insertions
□ Tom_parent_73_transg 1809-stringent
□ Tom_Parent_73_Reg_Readmapping #398

Track List Settings
Navigation
SL2.40ch08 (63.032.657bp)
Range: 727.860 - 63.032.657
Insertions
□ Tom_parent_73_transg 1809-stringent
□ Tom_Parent_73_Reg_Readmapping #398

Track List Settings
Navigation
SL2.40ch09 (67.662.091bp)
Range: 1 - 67.662.091
Insertions
□ Tom_parent_73_transg 1809-stringent
□ Tom_Parent_73_Reg_Readmapping #398

Figure 18. Representation of uniquely identified SNVs in transformant 73 and regenerant 73 nearby identified T-DNA insertion locations on chromosome 3, 8 and 9 (upper, middle and lower panel respectively). Note that for each of the chromosomes a slightly different region size was selected for the sliding window including the variants. Frequency of variants is represented as red bars. Identified insertion sites are indicated by a blue arrow in transformants and a dotted line towards the corresponding chromosome location in the regenerant plant.

The figure below shows the spread of the different types of mutations in this transformant 73 among the whole genome. This transformant has the highest number of mutations. Also the positions of the three insertions of T-DNA are given. Also this figure shows that there is no clustering of mutations near the T-DNA insertions. This holds for all types of mutations.

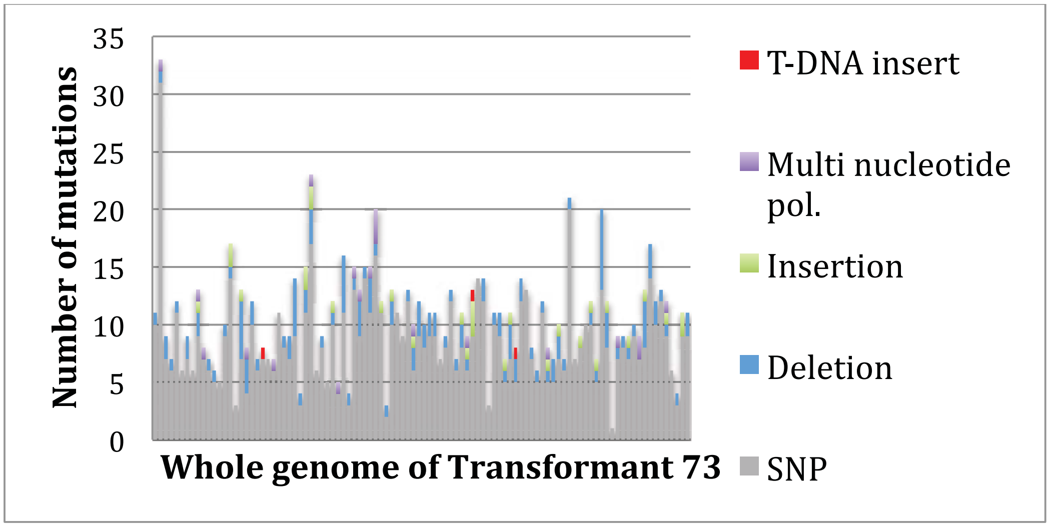


Figure 19. The frequencies of the different types of mutations spread among the whole genome of the transformant 73. The positions of the three T-DNA insertions are shown in red. The mentioned insertions and deletions are small (<10 bp).

5.3.2 Arabidopsis

Sequence data:

High quality sequence reads were selected and mapped to the *A. thaliana* Columbia (Col-0) TAIR10 reference genome. All samples had an average coverage of at least 25X and a high fraction of the reference genome covered (>0.99) indicating the production of highly comparable datasets (Table 3). The coverage for the non-transformed pool was well above the minimum intended coverage of 100 X.

Table 3. Number of reads and coverage per *A. thaliana* transformant At_T1 to At_T5. WT1-4pool refers to the pool of four wild type *A. thaliana* plants.

sample ID	Sequence reads	reads after trimming	Read Length	% mapped reads	Average Coverage
At_T1	61683376	59480648	98.1	99.6%	48.53
At_T2	33049036	32103568	98.42	99.5%	26.25
At_T3	31968584	30902759	98.2	99.8%	25.29
At_T4	56915326	54937185	98.1	99.5%	44.77
At_T5	45006154	43359470	98.1	99.7%	35.4
WT1-4pool	158153422	152300479	98.1	99.5%	124.1

Nucleotide variant detection specific in transformants

Due to the large numbers of variants observed for each line (average 5362 ± sd 123) and control pool (29706) when compared to the TAIR reference genome we concluded that the used Arabidopsis plants deviated significantly from the published genome sequence of Columbia Col-0. Also in this experiment we therefore repeated multiple mappings with different settings in order to exclude the common variants among the transformants and wild type plants.

For each transgenic plant we executed a stringent variant calling compared to the reference genome TAIR and compared the identified variants to the less stringently called variants found within the non-transformed pool. Variants called for each of the transformants were then compared to exclude common variants and the remaining 29 variants were all manually checked. Of these 29 variants, eight variants appeared to be specific for a transformant. These SNVs are shown in Table 4. SNVs that were shared among transformants were inherited from the common parent, and were not a result of transformation.

Table 4. SIngle nucleotide variants (SNVs) detected in five transformants of *A. thaliana*. We did not discover SNVs in At_T2 and At_T5.

Plant	Region	Type	Reference	Allele	Length	Within gene	Exon/ Intron	Zygosity
At-T1	Chr2_11595707..11595709	Deletion	GTG	-	3	AT2G27130	Exon	Hetero
At_T1	Chr2_18967242..18967245	Deletion	TTCC	-	4	-	-	Hetero
At_T1	Chr5_23122680..23122684	Deletion	GGGTA	-	5	AT5G57120	Exon	Hetero
At_T1	Chr5_24549125	Deletion	C	-	1	AT5G60990	Exon	Hetero
At_T3	Chr5_5539059	SNP	A	T	1	AT5G16850	Intron	Hetero
At_T4	Chr2_4023449	SNP	G	C	1	-	-	Hetero
At_T4	Chr4_652049..652050	Deletion	AC	-	2	-	-	Hetero
At_T4	Chr5_18641150..18641153	Deletion	GTAG	-	4	-	-	Hetero

The frequency of SNVs equals 8 in 5 plants, so on average 1.6 SNVs per transformant. Three SNVs occurred in an exon. Two out of these three led to a frame shift, which may be deleterious for the coded protein. As the mutations were heterozygous, the wild type alleles were still present. Fitness costs are therefore unlikely in these plants. Progeny plants that would be homozygous for a mutation, however, might show fitness effects.

Remarkably, 6 of the 8 SNVs consist of small deletions of several nucleotides, and only two SNPs were detected.

For comparison, we analysed the variants within the untransformed pool WT1-4. For this we now executed the variant calling of the pooled untransformed plants using stringent criteria and compared the identified variants to the less stringent called variants found within each transformed plant. After filtering, 51 putative variants remained. These were visually inspected, and only one SNV appeared to be true.

During visual inspection of T-DNA insertions, we identified two more SNPs close to a T-DNA insertion that were not identified using the approach described above, as these SNPs were present in broken read pairs. Broken read pairs were excluded in SNV calling. These will be discussed below.

Determination of T-DNA insertions

To determine the location of the T-DNA inserts, we mapped the reads to both the *A. thaliana* Columbia genome TAIR10 and the vector sequence (T-DNA containing pSaur8, and vector pBGWFS7). We specifically looked for broken read pairs and partial mapped reads on the gene construct ends, and used these reads to locate putative insertion locations on the genome. The used gene construct contained a promoter from *A. thaliana.* This appears also from the unequal coverage of the T-DNA. Indeed, mapping selected reads of this high coverage T-DNA region pointed clearly to chromosome 2 (Fig 20). Consequently this part of the T-DNA was excluded for all other downstream analysis.

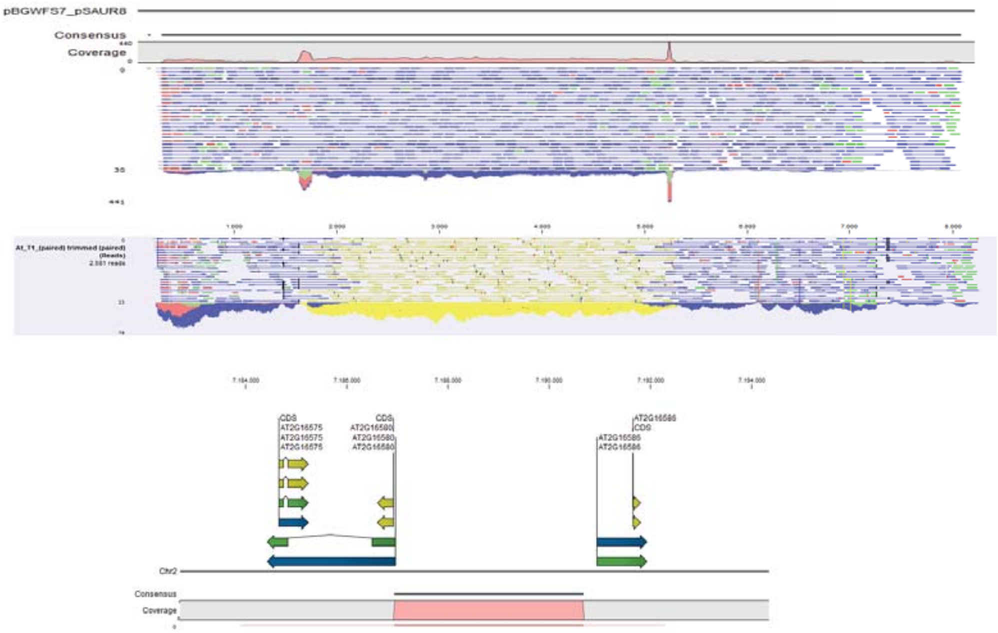
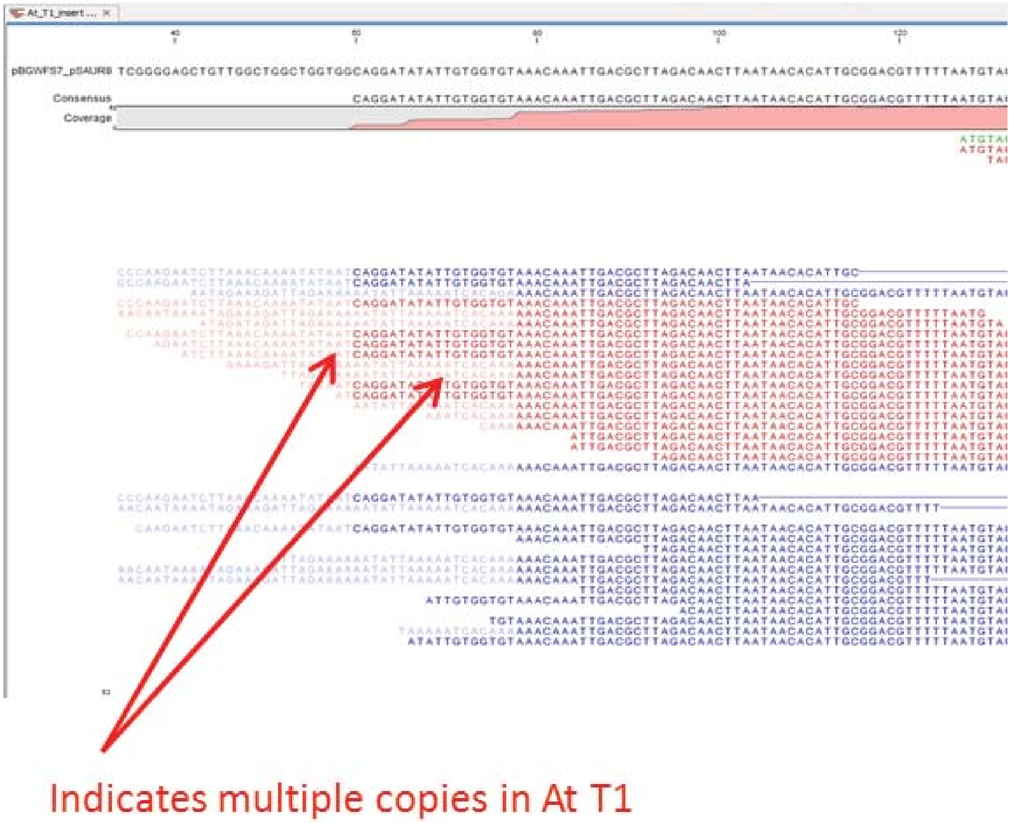


Figure 20. Arabidopsis sequence in the T-DNA. Upper part: At_T1 reads mapped to the pBGWFS7-pSAUR8 gene construct showing unequal coverage. Middle part: Selected reads from the upper part mapped to the *Arabidopsis* genome including the gene construct as reference. Reads derived from the higher coverage region in the upper part mapped (yellow) to *A. thaliana* TAIR10 chromosome 2. Lower part: Annotation of *A. thaliana* chr 2 region with the promoter region used in the gene construct.

Based on the different starting locations of split reads that mapped both partially to the T-DNA and partially to different genomic regions (Figure 21) we identified multiple insertion events within one transgenic plant.


At_T1_insert ...
40
60
80
100
120
pBGWFS7_pSAUR8 TCGGGGAGCTGTTGGCTGGCTGGTGGCAGGATATATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACACATTGCGGACGTTTTTAATGTA
Consensus CAGGATATATTGTGGTGTAAACAAATTGACGCTTAGACAACTTAATAACACATTGCGGACGTTTTTAATGTA
Coverage

Indicates multiple copies in At T1

Figure 21. Reads of transformant T1, mapped to the left end of the T-DNA. The figure illustrates that two different T-DNA insertions had occurred in this plant. The normal font refers to T-DNA, and the transparent font to genomic DNA of the host. Reads of intact mapped read pairs are in blue, reads of broken pairs in red.

For each transgenic plant we selected split reads and mapped these back to the reference genome to find positions of the insertions. Identified putative insertion positions were verified manually using heterozygous coverage of broken read pairs and split reads as criteria. As we used primary transformants, the insertions should be heterozygous, rather than homozygous. At_T1 and At_T2 contained four reliable T-DNA insertions, At_T3 contained two of them, while At_T4 and At_T5 contained one or two reliable insertion positions. In addition we used command line scripts to search for read pairs of which one read mapped to the T-DNA and the corresponding read of the pair mapped to the genome, so broken pairs. This enabled us to identify additional putative insertion sites. Also these were manually verified using visualization of read mappings by CLC genomics software. Remarkably, visual verification of identified insertion sites revealed several events where the T-DNA insertion resulted in a genomic deletion (four examples are given in Figure 22-25). The insertion event of At_T5 on chromosome 1 even associates with a large (~736 Kb) heterozygous deletion including many genes (Fig. 13). A summary of the identified T-DNA insertion sites for the five transgenic plants is given below in Table 5.

In addition, in three cases, multiple T-DNA border-spanning paired-ends were identified but only spanning one border sequence. This may suggest the presence of partial insertions or inverted repeated insertions. Further research is needed to investigate these insertion events in detail.

Table 5. Putative T-DNA insertion sites, detected by analyzing split reads using CLC* and/or broken read pairs using command line scripts‡.

Plant	Insertion location	Observation during analysis of broken pairs	Visual inspection	Insertion in a gen?	Putative type of insertion
At_T1	chr 1 25247859‡	Broken pairs (BP) two-sided*), split reads, heterozygous, deletion 25247859---864	Confirmed	-	Normal T-DNA
	chr 3 1738367* (1735553‡)	BP one-sided*), split reads, heterozygous	Confirmed	AT3G05830: Exon	Partial insertion?
	chr 5 6216771* (6220800‡)	BP two-sided, split reads, heterozygous, deletion 6216771---781	Confirmed	-	Normal T-DNA
	chr 5 8704509‡	BP two-sided, split reads, heterozygous, deletion 8704509---520	Confirmed	-	Normal T-DNA
At_T2	chr 1 3601046‡	Low coverage, BP one-sided, split reads, heterozygous, deletion 3601046-4347	Confirmed	AT1G10820: Exon	Partial? Chimeric?
	chr 2 12291247* (12290971‡)	BP two-sided though biased, split reads, heterozygous	Confirmed	-	No partial insertion, however a deviating pattern
	chr 2 12598512*	BP two-sided, split reads, heterozygous, deletion 12598512---545	Confirmed	-	Normal T-DNA
	chr 2 16310954‡	Low coverage, BP two-sided, split reads, heterozygous, evidence for small deletion around 16311370*	Confirmed	AT2G39080: Intron	Chimeric?
At_T3	chr 2 15559497	BP two-sided, split reads, heterozygous	Confirmed	AT2G37040: Intron	Normal T-DNA
	chr 3 23000811* (23000640‡)	BP two-sided, split reads, heterozygous	Confirmed	-	Normal T-DNA
At_T4	chr 2 272121‡	Only few BP	Insertion could not be confirmed. Probably an artefact.	AT2G01600	Partial insertion? Artefact?
	chr 3 45795* (45264‡)	BP two-sided, split reads, heterozygous, deletion 45795---877	Confirmed	-	Normal T-DNA
At_T5	chr 1 29443606‡	BP two-sided, split reads, heterozygous, large deletion *29443606---30180091	Confirmed	-	Normal T-DNA. Big deletion in plant DNA
	Chr 3 13632986‡	BP two-sided, split reads, but low coverage at repeat border	Questionable	-	No partial insertion, however a deviating pattern

*) Two-sided means that both ends of the T-DNA are detected in the broken pairs. This indicates a (nearly) full T-DNA insertion. One-sided means that only one end of

the T-DNA is detected in the broken pairs and split reads. This might be caused by a small insertion (splinter), or to an inverted repeat.

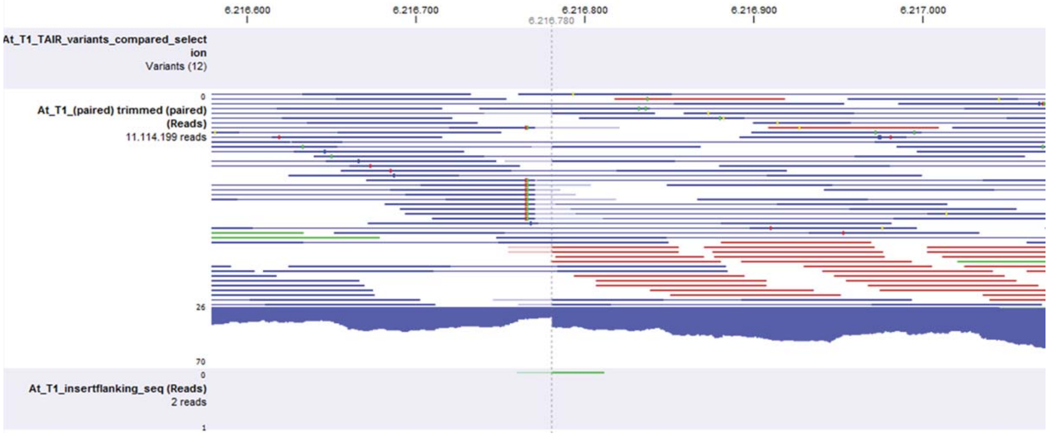

6.216.600
6.216.700
6.216.800
6.216.780
6.216.900
6.217.000
At_T1_TAIR_variants_compared_selection
Variants (12)
0
At_T1_(paired) trimmed (paired) (Reads)
11.114.199 reads
26
70
0
At_T1_insertflanking_seq (Reads)
2 reads
1

Figure 22. Example of a deletion at the T-DNA insertion site at chromosome 5 of transformant At_T1. Reference positions of Chromosome 5 are shown on top. The dotted vertical line represents the identified insertion location. Dark blue lines represent intact read pairs. The light blue lines in between indicate non-sequenced DNA fragment of the intact pair. Green and red lines represent forward and reverse oriented reads of broken pairs respectively. The blue graph represents coverage of mapped reads. The clear sudden coverage drop at the insertion location indicates a heterozygous genomic deletion at the T-DNA insertion site. This is confirmed by the sudden misalignment (light colours) of split reads left of the dotted line. One single split read at the bottom (in green) perfectly aligned to the genome sequence at the right side of the read, and perfectly to the T-DNA at the left side of the read.

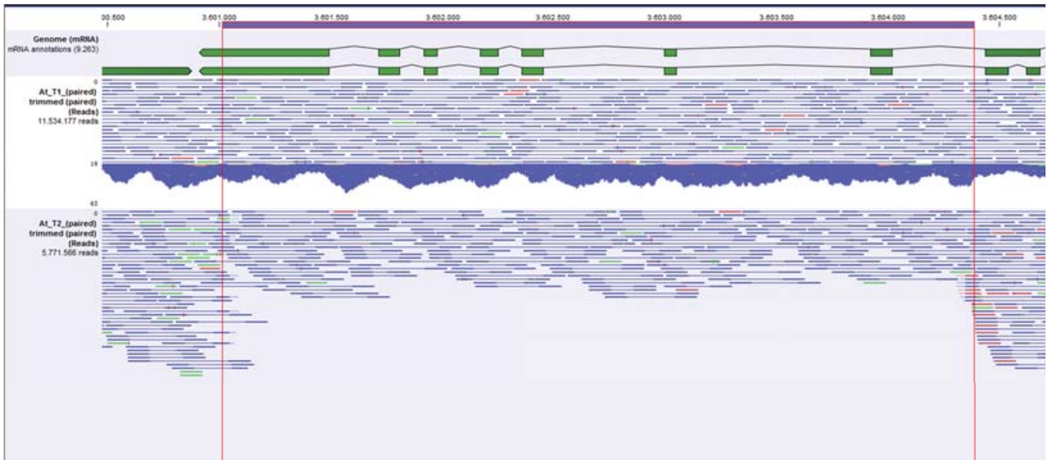


Figure 23. Example of a deletion at a T-DNA insertion site in transformant At-T2. Read mappings at chromosome 1 of transformant At_T1 (upper panel) show equal coverage distribution, while those of transformant At_T2 (lower panel) show a clear drop in coverage. This sudden coverage drop occurs between the identified T-DNA insertion borders thereby resulting in a 3301 bp deletion within gene locus AT1G10820.

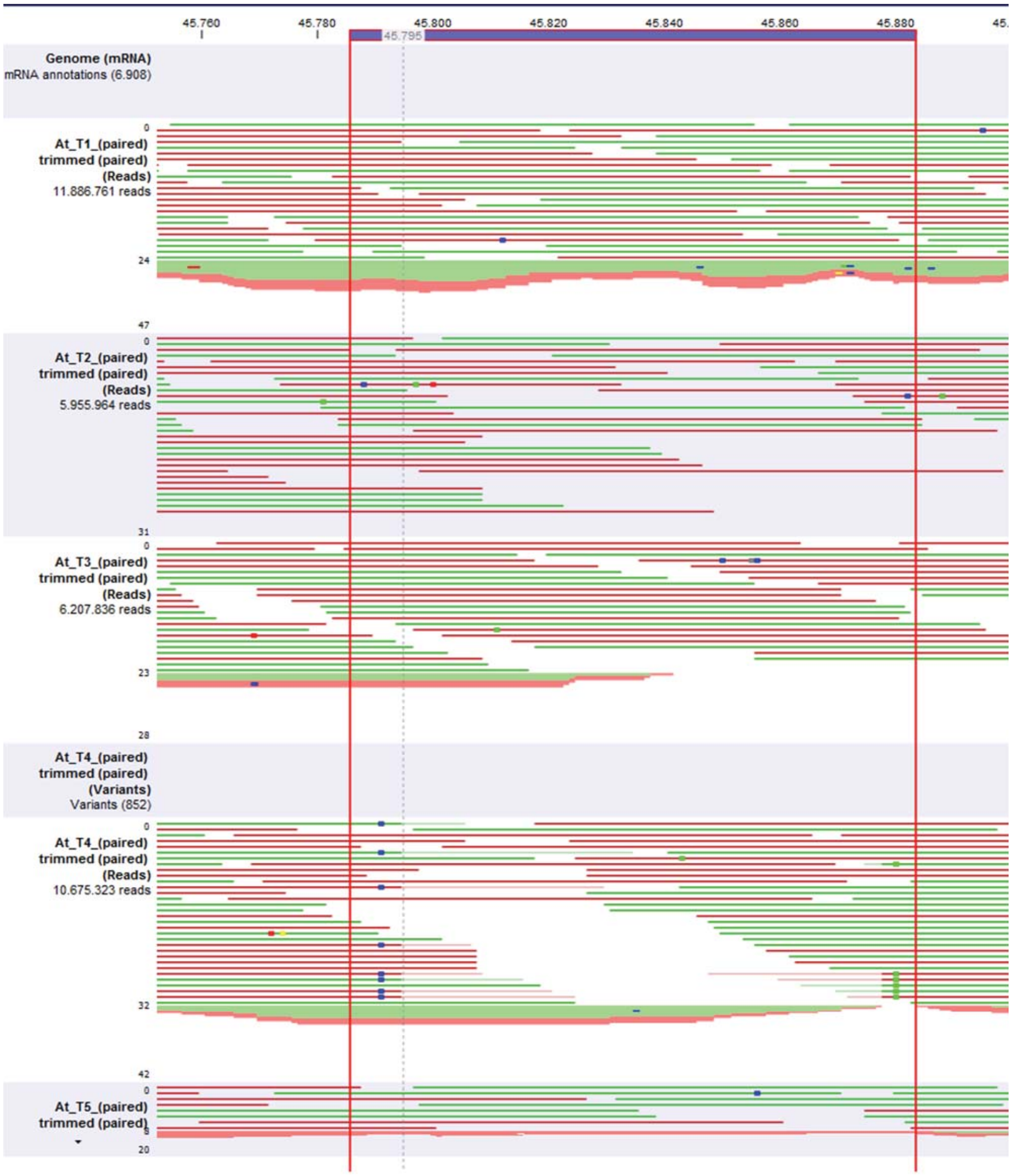


Figure 24. Example of a deletion at T-DNA insertion site in transformant At-T4. In this Figure read pairs are disconnected resulting in single reads that map in forward (green) or reverse (red) orientation, also for intact pairs. Read mappings on a selection of chromosome 3 are shown for all five transformants. Insertion of T-DNA in transformant At_T4 resulted in an 82 bp deletion starting at insertion site on chromosome 3, position 45795 (dotted vertical line). The T-DNA sequences are indicated in a light color. This deletion is not observed in At_T1 to 3 that are given as reference. In addition two SNPs correlating with the T-DNA insertion in At_T4 were identified. These are an A->T SNP at 45791 (blue dots in reads to the left of the T-DNA insertion site) and an A->C SNP at position 45881 (green dots in reads to the right of the T-DNA insertion site). Red lines indicate the area of interest.

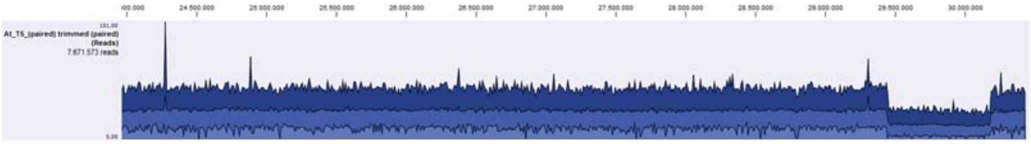


Figure 25. The large deletion at a T-DNA insertion site in transformant At-T5. Mapped read coverage on a selection of chromosome 1 is plotted as maximum (dark blue), average (blue) and minimum (light blue) coverage. The T-DNA insertion is at the large (~736 kb) heterozygous deletion, clearly visible as a ~50% drop of mapped read coverage between locations 29443606 and 30180091. Close inspection of broken pairs confirmed the presence of the T-DNA at this deletion.

5.3.3 'Splinters'

Bioinformatic analysis

We filtered for broken pairs and split reads that contained both plant DNA sequences and vector sequences. These were mapped to 1) the reference genome and 2) the vector sequence (T-DNA and backbone), using scripts on the command line level in Unix. The obtained alignments were visualized using CLC-software. This was done for the individual insertions in the five transgenic *Arabidopsis* plants and in the three transgenic tomato plants.

One clear small insertion, that we named a splinter, was discovered in At_T2 of *A. thaliana.* The splinter consisted of 50 bp of a gene, coding for green fluorescent protein (gfp), originating from the used T-DNA. The 50 bp of the splinter gave a perfect alignment in the middle of the T-DNA, so not at the right border or left border. This small insertion was detected in this transformant only, and not in the four other transgenic *A. thaliana* plants. Further, the insertion appeared to be in a heterozygous state, confirming that it was inserted in one chromosome during the transformation, and not in the homologous chromosome, as expected. The insertion was detected in 12 split reads that mapped to Chr. 2, around position 16.311.370 of the chromosome. Nine out of these 12 reads started in the plant genome, and changed in the gfp sequence, and continued in the plant genome (Figures 26, 27). The remaining three split reads contained one plant sequence and one T-DNA sequence, fully confirming the findings from the mentioned nine split reads. The insertion site shows an 11 bp deletion in the plant genome. In this deletion, the 50 bp splinter was inserted. As the splinter was inserted in a reverse orientation, the read sequences of Figure 26 should be converted into the inverted complement sequences (the sequences of the complementary strands), for revealing the resemblance with the sequences in Figure 27. Scrutinizing the sequence information revealed that 6 basepairs 'filler DNA' was inserted between the splinter and the plant DNA.
The sequence of the splinter is
ATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCG.
The insertion occurred in an intron of gene AT2G39080.

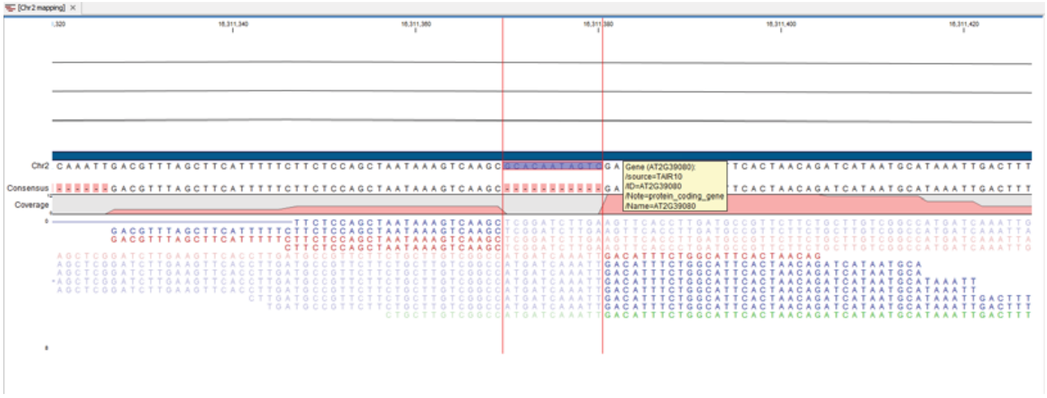


Figure 26. Split-reads from Transformant T2 of *A. thaliana*, that mapped on Chr. 2. The light letters represent bases that do not resemble the sequence of Chr 2. The normal font letters display the bases that do match with the plant sequence. The two vertical red lines show the insertion site and a deletion of 11 bp plant DNA.

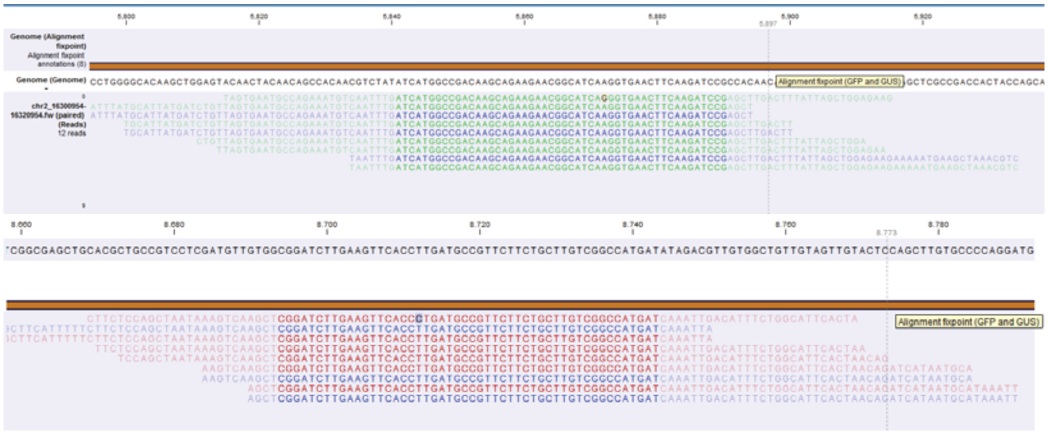

5,800
5,820
5,840
5,860
5,880
5,900
5,920
Genome (Alignment fixpoint)
Alignment fixpoint annotations (8)
Genome (Genome)
chr2_16300954-16320954.fw (paired) (Reads) 12 reads
Alignment fixpoint (GFP and GUS)
8.660
8.680
8.700
8.720
8.740
8.760
8.780
8.773
Alignment fixpoint (GFP and GUS)

Figure 27. The split-reads from T2 that mapped to Chr. 2 of *A. thaliana* are mapped here to the vector DNA of the used *A. tumefaciens* strain. The split-reads map to the the gfp-part of the T-DNA. The light parts do not align to the vector, but do match with the plant DNA's sequence. As the splinter was inserted in the reverse orientation, the read sequences were displayed in the inverted complement sequences in the lowest figure, for revealing the resemblance with the sequences in Figure 26.

Complexity of the bioinformatic analysis

We were not able to find conclusive evidence for other splinters in the transgenic *A. thaliana* and tomato plants, although there were some alignments that could not be easily explained. A detailed analysis of all individual insertions was performed, but further scrutinizing or new data appeared to be required to understand each insertion. The analysis was complex, because multiple insertions were present in each plant, making assembling of vector sequences per plant difficult. The insertions occurred in different sites, and included repeated T-DNA insertions, involving inverted repeats too. Insertions in repetitive DNA or a poor local quality of the reference genomes, also hampered the analysis. However, the 50 bp splinter described above, appeared to be very clear and consistent, because the twelve split reads contained the gfp splinter sequence. Nine out of these 12 split reads contained the full gfp splinter, flanked at both sides by plant DNA.

PCR confirmation of the splinter

In order to exclude the unlikely event that an artefact during NGS library preparation led to the 50 bp splinter, we performed PCRs for *A. thaliana* transformant At_T2, using an Agilent Bioanalyzer, starting from DNA before library preparation. Primers were designed that were specific for the flanking DNA and the splinter (Table 6). Although also unspecific bands occurred (Table 6D), the PCRs showed the expected amplicons at the expected sizes. The bioinformatic analysis indicated that the splinter is present in heterozygous state, which makes sense, as one copy of Chr 2 of the primary transformant should contain the splinter, whereas the homologous chromosome should not harbour this T-DNA fragment. The PCR results confirm this, as appears from Sample 5 in Table 6. One band represents the amplicons with native plant DNA only, whereas the ~50 bp larger band contains the same fragment + the inserted splinter. The difference in expected size between the band with splinter and the band without splinter is not exactly 50 bp but approximately 50 bp, as a small deletion did occur in the plant DNA at the insertion site, and a few bp filler DNA was added besides the splinter, as described in detail above.

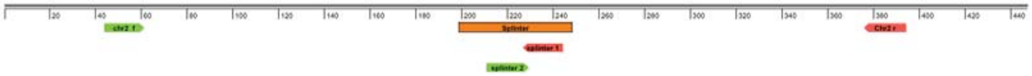
We conclude that the presence, genomic position, and size of the 50 bp splinter is confirmed by PCR.

Table 6. PCRs for verification of the presence of the 50 bp splinter in *A. thaliana* plant 2 (At_T2); A. The primer sequences; B. the positions of the primers of the chromosome containing the splinter. In view of readability, not the absolute genomic positions are shown (around 16.311.380), but local positions with smaller numbers (0 to 440); C. expected and observed amplicons sizes; D. digital representation of the observed amplicons after PCR, using Agilent Bioanalyzer. The bands that resemble the expected bands are indicated by arrows.

A.

Primer name	Primer sequence
splinter 1	TGGCCGACAAGCAGAAGA
splinter 2	TTCACCTTGATGCCGTTCT
chr2 F	TGCTGCATTCCTGATCCG
chr2 R	ATGTGATCTTTTGTGCTCC

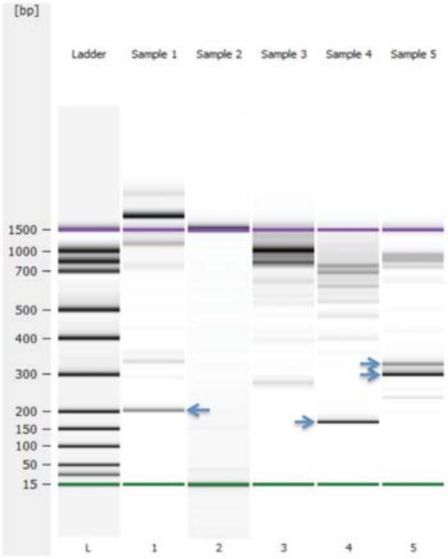
B.



C.

Primer 1	Primer 2	Sample	Amplicon size (bp)	
			Expected	**Found**
Chr2F	Splinter 1	1	201	~ 200 bp and faint bands
Chr2R	Splinter 1	2	no	no
Chr2F	Splinter 2	3	no	many bands
Chr2R	Splinter 2	4	184	~180 bp and vague faint bands
Chr2F	Chr2R	5	with splinter: 351 without splinter: 305	~350 bp, ~300 bp and faint bands

D.



5.4 Discussion and conclusions from the experimental part

To estimate the mutation frequency due to transformation (including selection for kanamycin-resistance), we re-sequenced five *A. thaliana* transformants that were obtained from one parental plant after flower dip, so without regeneration. The number of mutations (next to the T-DNA insertions themselves) varied from 0 to 4 mutations per plant, 1.6 mutations per plant on average. The frequency of the detected SNVs is not significantly higher than the frequency of spontaneous mutations in *A. thaliana* during the 30 generations experiment of Ossowski et al. (2010) seed propagation, which was 2.3 spontaneous mutations per plant per generation. When comparing Tables 4 and 5, it appears that there was no spatial association between the positions of the T-DNA insertions and the locations of the SNVs on the chromosomes. This indicates that the transformation event was not mutagenic to the remainder of the genome, although the transformants may have had stress due to kanamycin in the selection medium.

The frequencies and positions of the SNVs on the genomes of the transformants indicate that these SNVs are not a result of the transformation itself, but a result of spontaneous mutations, occurring also in non-transformants. As the number of evaluated transformants was relatively low, it cannot be excluded that the selection for kanamycin resistance might have slightly increased the mutation frequency.

However, at the T-DNA insertion sites themselves we detected one big deletion in one transformant (Fig. 25). This deletion was next to the T-DNA insertion, and its size was 736 kb. Also in the DNA flanking other T-DNA insertions several smaller or larger deletions were detected. Examples are shown in Figs. 22 to 24. Apparently, T-DNA insertions can be accompanied with small or large deletions in the plant DNA right at the site of the inserted T-DNA.

Labra et al. (2004) also looked for genomic DNA changes in transgenic *Arabidopsis* plants after flower dip, using AFLP (amplified fragment length polymorphism) and RAMP (random amplified microsatellite polymorphism). No statistically relevant genomic modifications in transgenic plants were detected, as compared with control untreated plants. This result is consistent with our findings. As the number of mutations that we found is very low, it is unlikely that they can be detected with marker platforms such as AFLP and RAMP. However, Labra et al. did detect variation in callus-derived *A. thaliana* plants. They concluded that “somaclonal variation is essentially correlated with the stress imposed by the *in vitro* cell culture, rather than with the integration of a foreign gene”.

We also re-sequenced a pool of four non-transformant *A. thaliana* plants, obtained from one mother plant. We detected only one mutation. However, the pooling of four plants, the heterozygous nature of mutations, and the unequal distributions of the reads among the different plants, made it very difficult to filter for mutations in a reliable way. Therefore we regard the estimation of the mutation frequency of the plants in the pool not reliable. The estimations from Ossowski et al. (2010) are far more reliable, and that is why we use those estimates, even though the experimental design is slightly different.

We also developed GM tomato plants and non-GM regenerants, derived from common parents, and re-sequenced these plants. We detected about 100 mutations per plant in the non-GM regenerants compared to the parents (Fig. 16). In the transformants, the number of mutations varied strongly, from 17 to 700. The low number of mutations in one transformant is remarkable, as one would expect at least a similar mutation frequency as in the regenerants. Also the high number of mutations in Transformant 73 (Fig. 16) is striking. We investigated the positions of the mutations in the transformant and this shows that the mutations were randomly distributed across the genomes, also for Transformant 73 and not necessarily close to the inserted T-DNA. There was no significant difference between the mutation frequencies of the regenerants compared to the transformants. We do not know whether the higher variation in the frequencies of mutations in the transformants compared to the lower variation in the regeneration is caused by the transformation process, including the selection for kanamycin resistance.

The analyses of the T-DNA insertion sites in *A. thaliana* showed that genomic DNA is often deleted at the T-DNA insertion. These deletions were sometimes small (2 to 11 bp), but also larger deletions were detected (33 bp, 82 bp), and in one case a large deletion of 736 kb, including several genes, was found. It should be realized that in the transformants the T-DNA insertions and the deletions are heteozygous. Disruption or loss of important genes (Table 5) may be compensated by the homologous intact chromosomes. However, if after seed propagation progeny plants segregate for these mutations, progeny homozygous for the T-DNA and for the deletions will arise, and it may display possible adverse fitness effects from the disruption or deletion of genes.

We did not analyse in detail the sizes of inserted T-DNAs, as that was outside the scope of this project. However, we obtained evidence for the presence of a small insertion of T-DNA that we named a splinter (paragraph 5.3.3). Such a small insertion of 50 bp is most likely not detectable by regular Southern blotting, but only detectable by NGS. Information on presence of small insertions is relevant in view of methods for the molecular characterisation of inserts for authorisation of GM cultivars.

In tomato, we found indications of T-DNA insertions at a coverage well below 50% (Tables 2). We hypothesize that the low coverage may result from chimerism in the transformants. For DNA isolation, all leaves of a transformant were picked, and pooled per plant. Chimerism, where only a sector of a transformant carries a specific T-DNA insertion, could explain the low number of reads for some inserts. This issue is not relevant for seed-propagated progeny of transformants.

6. Discussion

6.1 Mutation frequencies in GM genomes compared to the natural DNA variation between cultivars and breeding germplasm

Recently, 84 tomato genotypes were re-sequenced, including cultivars and wild species that are used in conventional tomato breeding. The number of SNPs (single nucleotide polymorphisms) accumulated in the tomato cultivars, when compared to the reference genome of 'Heinz', varied from 200K to 4.5M, 850K on average. The number of mutations in the tomato transformants and regenerants varied around 100 SNPs per plant, up to about 700 SNPs per plant. This means that the variation between cultivars is far higher (> 250 times higher) compared to mutations due to transformation and regeneration. The number of SNPs per wild accession for conventional breeding was even >10,000 times higher than the number of mutations per transformant. This means that the induced variation in the genome due to transformation and regeneration in tomato is a tiny fraction of what is present as natural variation between tomato cultivars and germplasm for tomato breeding, when ignoring the T-DNA insertion itself.

In *A. thaliana* we detected on the average 1.6 mutations per plant in the floral dip transformants compared to their parents. This is not significantly different from the mutation frequency in seed propagation without transformation, i.e. 2.3 spontaneous mutations per plant per generation (Ossowski et al., 2010). We did detect more deletions than SNPs, but the absolute number of differences is very small, so we cannot determine whether this has anything to do with the transformation procedure, or is an artifact of our stringent quality criteria during the analysis if the paired-end reads. The level of variation is very small, compared to the accumulated variation during evolution of this species. Cao et al. identified nearly 5 million (4,902,039) SNPs across the 80 *A. thaliana* accessions (Figure 2).

A similar conclusion can be drawn for rice, where 196 mutations were detected in a transgenic rice line compared to its parent (Kawakatsu et al. 2013), whereas nearly 8 million SNPs were detected when comparing rice cultivars (Huang et al. 2012).

We conclude that the number of genome-wide mutations detected in transgenic plants when compared to their parental plants is very low compared to variation among cultivars, and the same types of mutations are found as in conventionally bred varieties. Therefore it does not make sense to request identifications of all mutations in GM cultivars for an environmental risk assessment, as the variation is well within baseline of variation among cultivars.

However, at the T-DNA loci small deletions and also one very large deletion (> 700 kb) in a *A. thaliana* transformants were found. This is a common phenomenon (Ossowski et al., 2010), although the number of small insertions and deletions (indels) in the tomato and *A. thaliana* transformants was higher than in non-GM plants (Fig. 17, Table 4). The frequency of the deletions found in transformed plants is still small compared to the natural frequency of indels in cultivars within a species.

Latham et al. (2006) and Wilson et al. (2006) reviewed the phenomenon of mutations in plants caused during transformation. The authors focussed in particular on mutations and chromosomal rearrangements flanking the insertion of the T-DNA. In their view, these mutations pose a risk regarding biosafety. The transgenic plant should in their opinion be as identical to its parent as possible. Therefore, they recommend extended backcrossing for elimination of genome-wide mutations, and sequencing of flanking DNA of 50 kbp at each side of the insertion, and discarding of plants that show any mutation in the flanking DNA compared to the parent plant. These and other precautions should ensure that transformation-induced mutations will not impact on biosafety. However, Latham et al. (2006) and Wilson et al. (2006) did not mention mutation breeding of plants, that have led to more than 3200 plant varieties (Maluszynski et al., 2000; Ahloowalia et al., 2004; http://www-infocris.iaea.org/MVD/default.htm), and contain numerous mutations compared to their parents. If spontaneous mutants would also be included, the number of varieties would expand strongly. The induced mutant varieties have been developed in 175 plant species, including rice, wheat, barley, cotton, rapeseed, sunflower, grapefruit, apple, banana, and many other species. They are released world-wide and many millions of people eat and use products of these varieties, without molecular characterisation of the mutations, or imposed removal of off-target mutations by repeated back-crosses. One could react with a proposal to put also plants from induced mutations under strict safety regulations. This would make sense only if the mentioned >3200 varieties would have induced more frequently biosafety problems than varieties from cross breeding. However, we are not aware of any biosafety problem caused by an induced mutation of a released variety. Apparently, the common thorough evaluation of induced mutants at the phenotypic level by the breeders suffices (Schouten & Jacobsen, 2007).

Further, Latham et al. (2006) and Wilson et al. (2006) did not mention that in conventional breeding, traits from wild germplasm are introduced into cultivars by means of crosses and backcrosses with an elite cultivar. During this process, chromosomal parts from the wild germplasm are introduced. These chromosomal parts may harbour hundreds of unknown “wild” alleles and thousands of deviations in the DNA sequence compared to the original elite cultivar. These thousands of natural deviations can be regarded as thousands of mutations. We all use and eat such cultivars for many decades (Schouten & Jacobsen, 2007). The precautionary measures proposed by Latham et al. [1] and Wilson et al. [2] for GM plants regarding detection of mutations are not in balance at all with common practice in conventional plant breeding. It is unscientific to propose screening flanking DNA of 50 kbp at each side of the insertion, requiring discarding of plants that show any changes there, but simultaneously accepting plants with hundreds or thousands of unknown but probably more dramatic DNA changes after irradiation or introgression. Proposing a ban on mutations caused by gene transformation for the sake of biosafety indicates a blind spot for the safety of numerous mutations induced by conventional breeders for more than 70 years, and introgression of unknown chromosomal parts from wild germplasm since centuries (Schouten & Jacobsen, 2007).

6.2 Whole genome sequencing and environmental risk evaluation of GM plants

6.2.1 NGS data can be used for the molecular characterisation of the inserts for the ERA

An essential part of the environmental risk assessment (ERA) is the molecular characterisation of the insert(s) in the GM plant. The EFSA provide guidance for this part (EFSA Journal 2011; 9(5): 2150, p 10):

> ***Information on the sequences actually inserted/deleted or altered***
>
> *The applicant should provide the following information:*
> *a) size and copy number of all detectable inserts, both complete and partial. This is typically determined by Southern analysis* (underlining by HJS)*. Probe/restriction enzyme combinations used for this purpose should provide complete coverage of sequences that could be inserted into the host plant, such as any part of the plasmid/vector, or any remaining carrier or foreign nucleic acid introduced in the GM plant. The Southern analysis should span the entire transgenic locus/loci as well as the flanking sequences and include all appropriate controls;*
> *b) organisation and sequence of the inserted genetic material at each insertion site;*
> *c) in the case of deletion(s), size and function of the deleted region(s), whenever possible;*
> *d) sub-cellular location(s) of insert(s) (in nuclear, plastid, or mitochondrial chromosomes, or maintained in a non-integrated form) and methods for its determination;*
> *e) sequence information for both 5‟ and 3‟ flanking regions at each insertion site, with the aim of identifying interruptions of known genes. Bioinformatic analyses should be conducted using up-to-date databases with the aim of performing both intra-species and inter-species similarity searches. In the case of GM plants containing stacked events, applicants should assess the safety of potential interactions between any unintended modifications at each insertion site;*
>
> *Open Reading Frames (ORFs) present within the insert and spanning the junction sites. The ORFs should be analysed between stop codons, not limiting their lengths. Bioinformatic analyses should be conducted to investigate possible similarities with known toxins or allergens using up-to-date databases.*

As appears from this description and from the GMO applications submitted to EFSA, Southern blot analysis is a standard part of the molecular characterisation of the inserts in the GM plants for Europe. However, banding patterns in Southern blots are not always clear and cannot always be interpreted unambiguously. For finding flanking host DNA, PCR-based chromosome walking approaches are commonly used. This can be cumbersome too, especially in case of complex inserts, or in case of inserts that are similar to native sequences (Vanblaere et al., 2013). NGS may be helpful here.

We mention three papers describing the use of NGS for the characterization of T-DNA inserts in GM plants.

Kovalic et al. (2012)

Kovalic et al. (2012) described the use of next generation sequencing for molecular characterisation of two GM soya genotypes, one containing a single T-DNA insertion, and one containing a more complex insert. They compared their approach with the Southern blot analysis, and showed that next generation sequencing can efficiently replace the Southern blots in this context. They deep-sequenced (>70 X) the GM genotype, using Illumina sequencing, covering the whole genome, and selected the reads that contained sequences significantly similar to that of the transformation plasmid. The reads that contained both DNA from the plasmid and from the plant were called junctions (in our report we have called them split reads). These junctions were at the borders of the inserts. Alignments of these junctions allowed determination of the number of inserts, the presence or absence of unintended insertions, and flanking host DNA sequences. Furthermore, Kovalic et al. designed primers for amplification of the inserts, including about 1 kb flanking DNA at each side of the insert, and used them for validation of the insertion sites that were predicted by the NGS sequence analysis, by sequencing, precise analysis of the insert, and checking the integrity and organisation at that site. For comparison, they sequenced the same site in the non-GM parent, using the primers that annealed to the flanking DNA.
Kovalic et al. stated that the NGS approach is a kind of digital version of the Southern blot analysis, but has several advantages:

1. Experimentally simpler; no experimental variation based on T-DNA composition, whereas Southern blots require optimization per T-DNA;
2. Greater ease of data interpretation and greater reproducibility;
3. The costs are less than 50% compared to Southern blot analysis.

Both the GM genotype with a single insert and the GM genotype containing the complex insert were analysed successfully using NGS, according to Kovalic et al.

The authors took several precautions:

1. In order to minimize the number of reads that contained foreign DNA, care was taken to remove potential surface contaminants on the plant material, such as laboratory dust, bacteria, etc., and dedicated clean equipment was used;
2. The non-GM parent was sequenced too, in order to reduce the number of false positive 'junctions' that occur in endogenous plant DNA, as the native DNA of the plant may harbour sequences that are similar to parts of the sequences of the plasmid DNA;
3. The sequencing was deep (> 70 X).

Wahler et al. (2013)

Wahler et al. (2013) used NGS to characterise the insert in a GM rice genotype (LLRice62). They started from the assumption that an unauthorized event was analysed, for which the plasmid or T-DNA used was not known, but the parental genome had been sequenced already. Wahler et al. sheared the genomic DNA of the GM line into a library of fragments of 300 to 400 bp. The nucleotide sequences of the 5'- and 3'-ends were determined by paired-end sequencing. The reads were approximately 75 bp per end, at sequencing depth of 65 X on the average. Because the construct was not known but the sequenced parental genome was available, the NGS reads were aligned only to the parental genome sequence. When one read of a pair could be mapped to the reference genome, but the other read not, Wahler et al. named the non-mapped read an orphan read. In this way more than 20,000 putative 'inserts' were detected. For filtering these 'inserts' Whaler et al. assumed that (1) insertions from transformation are generally larger than 100 bp, i.e. larger than a single read and (2) that they exhibit homology to genetic elements typically used in plant transformation. For the latter selection criteria, they aligned the orphan reads to sequences of the pCAMBIA-1300 vector.

Sets of filtered, overlapping orphan reads were then used to assemble parts of the inserts. During an iterative process, new orphan reads were detected, leading eventually to a full description of the insert. In the insert the authors recognized sequences similar to the pCAMBIA-1300 vector, and a gene cassette of the bar gene under the control of the CaMV 35S promoter and the CaMV 35S terminator. This resembled the description of the LLRice62 in an application for EFSA.

A potential weak point in this analysis strategy, in our view, is the fact that they needed to use the sequence of the pCAMBIA-1300 vector. Although this vector was not used for the transformation of LLRice62, it contained very similar sequences to the used vector. If a very deviating vector would have been used, or biolistic methods without plasmids, the filtering of orphan reads would have been problematic, and the characterization of the GM event may have been unsuccessful.

However, the current report deals with using NGS for EFSA dossiers for which all sequence information on the used T-DNA is available. This deviates from the goal addressed by Wahler et al. (2013), i.e. detection and characterization of unauthorized events.

Yang et al. (2013)

The third paper on using NGS for molecular characterization is from Yang et al. (2013). They described three bioinformatics approaches, i.e. 1) an approach when the sequence of the used plasmid and T-DNA are known beforehand, 2) an approach when the plasmid and T-DNA are not known, but a library of commonly used plasmids and transgenes and promoters is used instead, and 3) an approach when no *a priori* knowledge is available on the sequences of the insert(s). In other words, they address the potential weak point of the strategy followed by Wahler et al. (2013). For the current report, the first case is relevant, i.e. 'the sequence of the used plasmid and T-DNA are known beforehand'.

For all three approaches, a reference genome of the species should be available. Further, for each approach a validation of the outcome was required, which they carried out using PCR covering the insert(s) and Sanger sequencing of these PCR fragments.

For the analysis, they used paired ends that could be categorized into five types (Fig. 28).

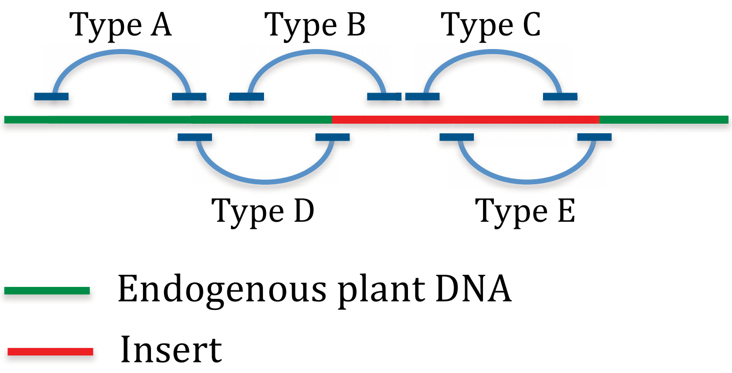


Figure 28. The five types of paired ends as distinguished by Yang et al. (2013) in their NGS analysis of GM rice.

For testing the bioinformatics approach (1) when the sequence of the plasmid is known beforehand, Yang et al. analyzed a GM rice event (T1c-19) after Illumina paired end sequencing at 24 X sequencing depth, 90 bp per end. They found eight putative insertions. Two out of these eight turned out to be real insertions. One putative insertion was placed on a wrong chromosome, due to repetitive sequences on that chromosome and on the correct chromosome. Five putative insertions could not be confirmed by PCR and Sanger sequencing. This underlines the importance of the validation by means of PCR and Sanger sequencing, when using the approach of Yang et al. They did not find backbone sequences from the plasmid used for *A. tumefaciens* mediated transformation.
After this analysis, Yang et al. confirmed that all inserts were covered, by means of counting the number of reads containing insert sequences, and comparing these to the total number of reads.

They tested also approach (2) when the plasmid and T-DNA were not known, but a library of sequences of commonly used plasmids, promoters and transgenes was used instead. They used the same GM rice event (T1c-19), but ignored the knowledge about the used plasmid and T-DNA. Rather, they used a library of sequences that were used in other GM crops. They again found the inserts, but they admit that the success depends on the completeness of the constructed GMO library.

They also tried approach (3), when no *a priori* knowledge about the insert was present at all. They tested this for the same rice event, and for mimicked transgene insertions from soya. The insertions could partially be detected.

Yang et al. emphasize that the NGS approaches are superior to Southern analyses and PCR techniques. They illustrate this by showing that state-of-art Southern analysis of the same GM rice line revealed only one insert, while NGS discovered that two inserts that were nearly identical (in size) were present.

The authors mention that for intragenic plants, which do not contain ‘foreign’ DNA but do contain ‘species-own’ DNA inserts, it will be extremely difficult to identify the inserts, especially when it is not known which intragenic sequences have been used. The same holds for cisgenic plants, that do not harbor foreign genes. Determining the baseline for intragenic events is a challenge, as it would require also the characterization of the level of (small) structural variants across a set of cultivars.

6.2.2 'Splinters'

In Arabidopsis we discovered one clear splinter, originating from the T-DNA used in the transformation. This splinter was 50 bp long, and appeared to be a fragment of the *gfp*-gene. This fragment is derived from the middle part of the T-DNA and not close to the left border or right border of the T-DNA As far as we know, this is the first report on occurrence of a 'splinter' in a transgenic plant.

In our case, only 50 bp of the *gfp*-gene was inserted. The coding region of a full *gfp*-gene is ~717 bp. It is very unlikely that the 50 bp insertion will result in a functional protein. From this perspective, such a rare splinter probably does not have a phenotypic effect that differs from a normal mutational event, so without any consequences regarding biosafety. In our case, the 50 bp fragment was inserted into an existing gene. Even so, no change in the coded protein was predicted, using FGENESH software from Softberry, as the splinter was inserted into an intron, and will be spliced out, together with the native intron, according to the gene prediction software. In this case, no phenotypic effect is expected. However, if it were inserted into an exon, a change in the predicted protein would occur.

Small splinters may be overlooked by conventional methods, such as Southern blotting and PCR. However, NGS is suitable to discover splinters.

However, during analysis of the NGS-data, one has to be aware of the possible occurrence of small T-DNA insertions, as one may disregard these insertions, and regard them as noise or artefacts.

Splinters may get lost in seed-propagated GM crops or when the transgenic trait/locus is crossed into other varieties, because:

1. Splinters, as all other T-DNA insertions, are present in heterozygous state in the primary transformant. During backcrossing of the transformant, approximately 50 % of the progeny will inherit the splinter, and the other 50 % will not harbour the splinter, meaning a loss of the splinter in half of the plants. Repeated backcrossing can reduce the probability of presence of the splinter. In case the splinter is genetically linked to the full T-DNA insertion (so present on the same chromosome pair as the full T-DNA insertion), and the progeny plants are selected for presence of the GM-trait, the likelihood of presence of the splinter will be higher than 50 % if the splinter and full T-DNA are present on the same chromosome (so in coupling phase), and lower than 50 % if the splinter is present on the homologous chromosome compared to the full T-DNA (so in repulsion phase).
2. In case of selfing of the transformant , approximately 50 % of the progeny will inherit the splinter. Repeated selfing can either lead to loss or to fixation of the splinter.

In this context it is relevant to mention that a transgenic event is evaluation by EFSA. This specific event can be used for crosses with other genotypes, leading to new GM varieties, if the construct is inherited to these varieties. As example, MON810 is a corn event that is approved for cultivation in the EU. Already more than 100 corn varieties have been derived from this event (pers. comm. B. Glandorf). If an event would contain a splinter, then derived varieties may have lost this splinter.

6.2.3 We recommend using NGS data for molecular characterisation of transformants, taking a series of conditions into account

Based on our own experience using NGS for detection of T-DNA insertions in transformants in Arabidopsis and tomato (Chapter 5), and on the examples described by others (Kovalic et al. 2012, Yang et al. 2013, and Wahler et al. 2013), we conclude that NGS outperforms Southern blotting techniques for detecting transgenic sequences in T-DNA. Overall, NGS is more sensitive for finding fragments of foreign DNA compared to Southern blotting. However, in order to achieve this higher sensitivity we have several remarks and recommendations:

1. The sequencing should be sufficiently deep. We used >25 X to detect insertions that were present. The plants were heterozygous for the insertion and the insertions were thus covered by a sequencing depth of >12.5 X on average. We regard this sufficient for positioning for the insertion, but low for an accurate description of the insertions. We recommend to use <u>> 25 X per haploid genome</u> for a homozygous diploid plant species. Polyploid and heterozygous transgenic plants require a much higher coverage and/or combining differently sizes libraries in order to be able to deal with homologous and paralogous genes, structural variants, etc. (but this was beyond the scope of our study). Kovalic et al. (2012), Wahler et al. (2013), and Yang et al. (2013) used 60 X, 65 X and 24 X coverage respectively in homozygous genomes, so far more than we used in our study;
2. This coverage recommendation refers to a read length of approximately 100 bp, and paired-ends, using Illiumina HiSeq. Longer reads, e.g. of 300 nt using MiSeq sequencing technology, or very long reads using PacBio technology would be even a better strategy for detection of split reads (so reads that cover both the T-DNA and the plant genome);
3. Preferably a good reference genome of the species should be available. Also it is recommended to re-sequence the non-GM comparator, with special attention to the insertion sites. This non-GM comparator can also be used to filter out false negative alignments of T-DNA reads to the plant genome;
4. The DNA sequence of the transgenic construct, used for transformation, should be known beforehand. This is the same requirement as for Southern blotting;
5. If the T-DNA is composed of native DNA, as is the case for intragenics and cisgenesis, and not known beforehand, then detection of the inserts using NGS or Southern blotting will be very difficult, as also natural rearrangements and gene duplication may have occurred. If the sequence of the T-DNA composed of native DNA is known beforehand, as is the case for dossiers that have to be evaluated by EFSA and COGEM, it might be feasible to detect the number and the sites of insertions using NGS, but we did not try this in the current study;
6. For a reliable, detailed description of insert sites, we strongly recommend that the NGS-derived insertions should be considered as putative inserts, which should be validated preferably by long-range PCR of the inserts in the transformant, followed by sequencing of the amplicons, using long range sequencing techniques such as PacBio. The PacBio reads should preferably cover the whole insert and flanking DNA. We recommend the use of long range sequencing techniques for the amplicons, in view of the possibly duplicated nature of the insert and/or flanking DNA.

7. Repeats in the insert may hamper a correct assembly of the reads, especially when using sequence reads shorter than the repeat. When a confirmation step is made, using PCR, an artefact may occur during the PCR, which leads to the same artefact as during assembly of repeats, and therewith enforces the misinterpretation rather than corrects the artefact. This is illustrated in the Fig. 29. Note that PCR is prone to creating artefacts like miss-priming resulting in a-specific PCR products, priming non-unique regions resulting in multiple PCR products, or truncated elongation resulting in strand switches which lead to chimeric PCR products. These possibilities should be considered and must be excluded. Similar errors can occur too when analyzing repetitive native DNA;
8. Organisations such as BGGO, COGEM and EFSA should be able to check the quality of the molecular characterisation. In view of this, a company that submits the request for market approval of the GM cultivar, should to provide the raw NGS data, and the bioinformatics pipeline that was used to analyse the sequencing data, in case the organisations would like to validate or outsource the validation to a third party. This should allow others to re-perform the analysis, based on the provided sequencing data, if there would be any need for doing such an analysis.
9. NGS is suitable for detection of splinters that might be overlooked with conventional methods, such as Southern Blotting, PCR, and genome walking tools.

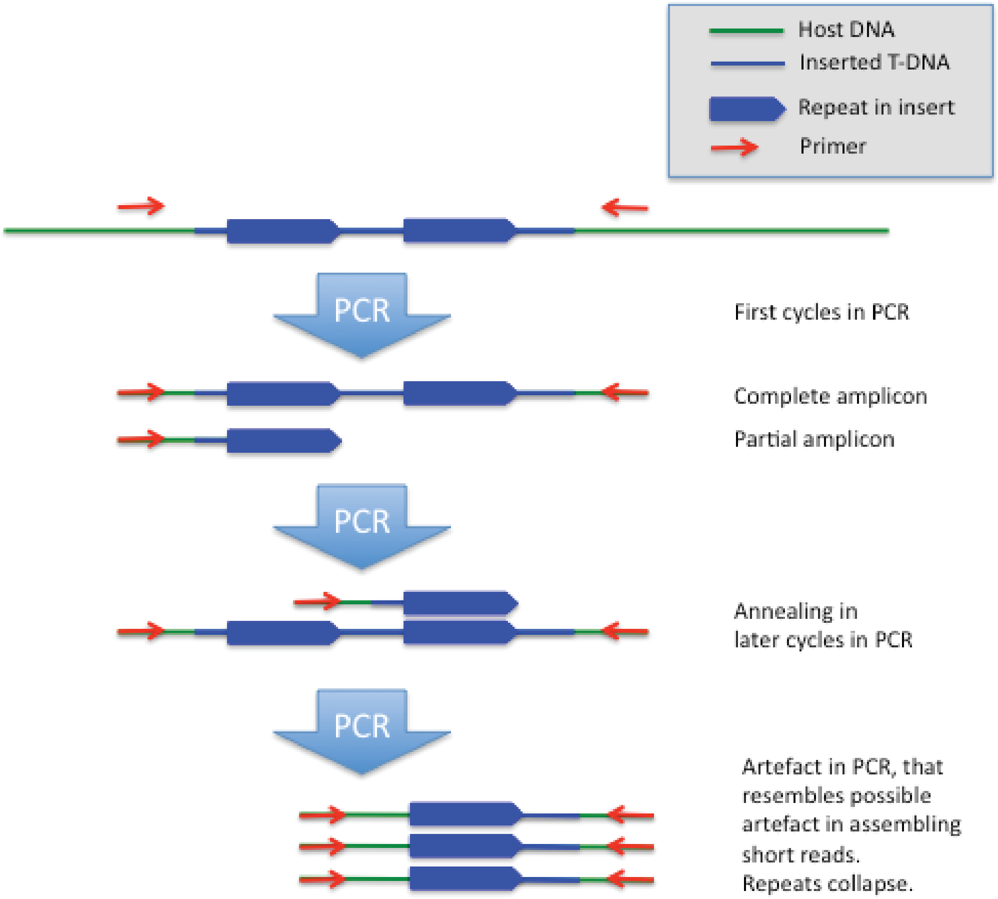


Figure 29. Illustration of a possible PCR artefact leading to collapse of repeats. Unfortunately, a similar artefact may be created when short reads are assembled into a contig. One should be alert to this kind of errors, e.g. by performing additional PCRs, using internal primers, or using an excess of primers and a low number of PCR cycles with long elongation times, followed by sequencing individual DNA stretches, e.g. using PacBio sequencing.

7. References

Ahloowalia BS, Maluszynski M, and Nichterlein K. (2004). Global impact of mutation-derived varieties, *Euphytica*, 135: 187–204.

Aldridge S, Huggett B, Jayaraman KS, et al. (2008) 1000 Genomes project. Nat Biotechnol 26:256–256. doi: 10.1038/nbt0308-256b

Atwell S, Huang YS, Vilhjálmsson BJ, et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631. doi: 10.1038/nature08800

Baidouri El M, Carpentier MC, Cooke R, et al. (2014) Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res 24:831–838. doi: 10.1101/gr.164400.113

Beilstein MA, Nagalingum NS, Clements MD, et al. (2010) Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. PNAS 18724–18728.

Cao J, Schneeberger K, Ossowski S, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956–963. doi: 10.1038/ng.911

Chang Y, Long T, Wu C (2012) Effort and Contribution of T-DNA Insertion Mutant Library for Rice Functional Genomics Research in China: Review and Perspective. Journal of Integrative Plant Biology 953–966.

Clough SJ, Bent AF (2008) Floral dip: a simplified method for Agrobacterium-mediated transformation ofArabidopsis thaliana. Plant Journal 16:735–743. doi: 10.1046/j.1365-313x.1998.00343.x

De Buck S, Podevin N, Nolf J, et al. (2009) The T-DNA integration pattern in Arabidopsis transformants is highly determined by the transformed target cell. The Plant Journal 60:134–145. doi: 10.1111/j.1365-313X.2009.03942.x

EFSA (2011) Guidance for risk assessment of food and feed from genetically modified plants. EFSA Journal 2011; 9(5):2150 1–37. doi: 10.2903/j.efsa.2011.2150

Ghedira R, Buck S, Ex F, et al. (2013) T-DNA transfer and T-DNA integration efficiencies upon Arabidopsis thaliana root explant cocultivation and floral dip transformation. Planta 238:1025–1037. doi: 10.1007/s00425-013-1948-3

Goff SA (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. japonica). Science 296:92–100. doi: 10.1126/science.1068275

Hamilton JP, Robin Buell C (2012) Advances in plant genome sequencing. Plant Journal 70:177–190. doi: 10.1111/j.1365-313X.2012.04894.x

Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M (1996) Retrotransposons of rice involved in mutations induced by tissue culture. Proc Natl Acad Sci USA 93:7783–7788.

Huang S, Deng L, Guan M, et al. (2013) Identification of genome-wide single nucleotide polymorphisms in allopolyploid crop Brassica napus. BMC Genomics 14:717. doi: 10.1093/bioinformatics/bth088

Huang X, Feng Q, Qian Q, et al. (2009) High-throughput genotyping by whole-genome resequencing. Genome Res 19:1068–1076. doi: 10.1101/gr.089516.108

Huang X, Kurata N, Wei X, et al. (2012a) A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501. doi: 10.1038/nature11532

Huang X, Zhao Y, Wei X, et al. (2012b) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32–39. doi: 10.1038/ng.1018

Initiative AG (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 796–815.

Jiang C, Mithani A, Gan X, et al. (2011) Regenerant Arabidopsis lineages display a distinct genome-wide spectrum of mutations conferring variant phenotypes. Current Biology 1385–1390.

Jiang N, Bao Z, Zhang X, et al. (2003) An active DNA transposon family in rice. Nature 421:163–167. doi: 10.1038/nature01214

Karimi M, Inzé D, Depicker A (2002) GATEWAY™ vectors for Agrobacterium-mediated plant transformation. Trends Plant Sci 7:193–195. doi: 10.1016/S1360-1385(02)02251-3

Kawakatsu T, Kawahara Y, Itoh T, Takaiwa F (2013) A whole-genome analysis of a transgenic rice seed-based edible vaccine against cedar pollen allergy. DNA Research. doi: 10.1093/dnares/dst036

Kikuchi K, Terauchi K, Wada M, Hirano H-Y (2003) The plant MITE mPing is mobilized in anther culture. Nature 421:167–170. doi: 10.1038/nature01218

Kovalic D, Garnaat C, Guo L, et al. (2012) The use of next generation sequencing and junction sequence analysis bioinformatics to achieve molecular characterization of crops improved through modern biotechnology. The Plant Genome Journal 5:149–163. doi: 10.3835/plantgenome2012.10.0026

Labra M, Vannini C, Grassi F, et al. (2004) Genomic stability in Arabidopsis thaliana transgenic plants obtained by floral dip. Theor Appl Genet 109:1512–1518. doi: 10.1007/s00122-004-1773-y

Latham, J. R., Wilson, A. K., & Steinbrecher, R. A. (2006). The Mutational Consequences of Plant Transformation. *Journal of Biomedicine and Biotechnology*, *2006*(2), 25376–7. http://doi.org/10.1155/JBB/2006/25376

Lee M, Phillips RL (1987) Genetic variants in progeny of regenerated maize plants. Genome 29:834–838. doi: 10.1139/g87-142

Lin C, Lin X, Hu L, et al. (2012) Dramatic genotypic difference in, and effect of genetic crossing on, tissue culture-induced mobility of retrotransposon Tos17 in rice. Plant Cell Rep 31:2057–2063. doi: 10.1007/s00299-012-1316-y

Lynch M (2010) Evolution of the mutation rate. Trends in Genetics 26:345–352. doi: 10.1016/j.tig.2010.05.003

Maluszynski M, K. Nichterlein, L. van Zanten, and B. S. Ahloowalia (2000). “Officially released mutant varieties—the FAO/IAEA database,” *Mutation Breeding Review* 12:1– 84.

Ming R, Hou S, Feng Y, et al. (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452:991–996. doi: 10.1038/nature06856

Miyao A, Nakagome M, Ohnuma T, et al. (2012) Molecular Spectrum of Somaclonal Variation in Regenerated Rice Revealed by Whole-Genome Sequencing. Plant and Cell Physiology 53:256–264. doi: 10.1093/pcp/pcr172

Muers M (2011) Technology: Getting Moore from DNA sequencing. Nature Reviews Genetics 12:586–587. doi: 10.1038/nrg3059

Müller E, Brown PTH, Hartke S, L rz H (1990) DNA variation in tissue-culture-derived rice plants. Theor Appl Genet. doi: 10.1007/BF00224228

Naito K, Cho E, Yang G, et al. (2006) Dramatic amplification of a rice transposable element during recent domestication. Proceedings of the National Academy of Sciences 103:17620–17625. doi: 10.1073/pnas.0605421103

Naito K, Zhang F, Tsukiyama T, et al. (2009) Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature 461:1130–1134. doi: 10.1038/nature08479

Nakano M, Nomizu T, Mizunashi K, et al. (2006) Somaclonal variation in Tricyrtis hirta plants regenerated from 1-year-old embryogenic callus cultures. Scientia Horticulturae 110:366–371. doi:

10.1016/j.scienta.2006.07.026

Ngezahayo F, Xu C, Wang H, et al. (2009) Tissue culture-induced transpositional activity of mPing is correlated with cytosine methylation in rice. BMC plant biology 9:91. doi: 10.1186/1471-2229-9-91

Ossowski S, Schneeberger K, Clark RM, et al. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res 18:2024–2033. doi: 10.1101/gr.080200.108

Ossowski S, Schneeberger K, Lucas-Lledo JI, et al. (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science 327:92–94. doi: 10.1126/science.1180677

Sabot F, Picault N, El-Baidouri M, et al. (2011) Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant Journal 66:241–246. doi: 10.1111/j.1365-313X.2011.04492.x

Sato S, Tabata S, Hirakawa H, et al. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641. doi: 10.1038/nature11119

Schlötterer, C. (2000). Evolutionary dynamics of microsatellite DNA. Chromosoma 109(6): 365-371

Schouten, H. J., & Jacobsen, E. (2007). Are mutations in genetically modified plants dangerous? *Journal of Biomedicine and Biotechnology*, *2007*(7), 82612. http://doi.org/10.1155/2007/82612

Schultz ST, Lynch M, Willis JH (1999) Spontaneous deleterious mutation in Arabidopsis thaliana. Proceedings of the National Academy of Sciences 96:11393–11398. doi: 10.1073/pnas.96.20.11393

Smulders MJM, G-J de Klerk (2011) Epigenetics in plant tissue culture. Plant Growth Regulation 63: 137-146. Doi: 10.1007/s10725-010-9531-4 (open access)

Stroud H, Ding B, Simon SA, et al. (2013) Plants regenerated from tissue culture contain stable epigenome changes in rice. eLife Sciences. doi: 10.7554/eLife.00354

The 100 Tomato Genome Sequencing Consortium, Aflitos S, Schijlen E, et al. (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant Journal 136-148. doi: 10.1111/tpj.12616

Vanblaere T, Flachowsky H, Gessler C, Broggini GAL (2014) Molecular characterization of cisgenic lines of apple “Gala” carrying the Rvi6 scab resistance gene. Plant Biotechnol J 12:2–9. doi: 10.1111/pbi.12110

Wahler D, Schauser L, Bendiek J, Grohmann L (2013) Next-Generation Sequencing as a tool for detailed molecular characterisation of genomic insertions and flanking regions in genetically modified plants: a pilot study using a rice event unauthorised in the EU. Food Analytical Methods 6:1718–1727. doi: 10.1007/s12161-013-9673-x

Weigel D, Mott R (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol 10:107. doi: 10.1186/gb-2009-10-5-107

Wilson A, J. Latham, and R. A. Steinbrecher (2006), "Transformation-induced mutations in transgenic plants: analysis and biosafety implications," *Biotechnology and Genetic Engineering Reviews*, vol. 23: 209–237.

Maluszynski M, K. Nichterlein, L. van Zanten, and B. S. Ahloowalia, "Officially released mutant varieties—the FAO/IAEA database," *Mutation Breeding Review*, vol. 12, pp. 1– 84, 2000.

Xu X, Liu X, Ge S, et al. (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111. doi: 10.1038/nbt.2050

Yang L, Wang C, Holst-Jensen A, et al. (2013) Characterization of GM events by insert knowledge adapted re-sequencing approaches. Sci Rep. doi: 10.1038/srep02839

Yu J (2002) A Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica). Science 296:79–92. doi: 10.1126/science.1068037

Acknowledgements

We are grateful indeed to the guidance committee from COGEM for the valuable input, discussions, and constructive comments on the report. We also acknowledge Prof Gerco Angenent and Dr Marian Bemer for developing and providing *A. thaliana* material. We thank Dr Jan-Peter Nap for the discussions about this project, and Bas te Lintel Hekkert for performing the Illumina Sequencing. Also we thank Dr. Clemens van de Wiel for critically reading this report.