Howard Hughes Medical Institute

Published as: *Methods Mol Biol*. 2015 ; 1284: 323–342.

# A User's Guide to the Arabidopsis T-DNA Insertional Mutant Collections

**Ronan C. O'Malley**[1,2], **Cesar C. Barragan**[1], and **Joseph R. Ecker**[1,2,3]

[1]Genomic Analysis Laboratory, Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd, La Jolla CA, 92037

[2]Plant Biology Laboratory, Salk Institute for Biological Studies, 10010 N. Torrey Pines Rd, La Jolla CA, 92037

[3]Howard Hughes Medical Institute

## Summary

The T-DNA sequence-indexed mutant collections contain insertional mutants for most *Arabidopsis thaliana* genes and have played an important role in plant biology research for almost two decades. By providing a large source of mutant alleles for *in vivo* characterization of gene function, this resource has been leveraged thousands of times to study a wide-range of problems in plant biology. Our primary goal in this chapter is to provide a general guide to strategies for the effective use of the data and materials in these collections. To do this, we provide a general introduction to the T-DNA insertional sequence-indexed mutant collections with a focus on how best to use the available data sources for good line selection. As isolation of a homozygous line is a common next step once a potential disruption line has been identified, the second half of the chapter will provide a step-by-step guide for the design and implementation of a T-DNA genotyping pipeline. Finally, we describe interpretation of genotyping results and include a troubleshooting section for common types of segregation distortions that we have observed. In this chapter we introduce both basic concepts and specific applications to new and more experienced users of the collections for the design and implementation of small- to large-scale genotyping pipelines.

## 1. Introduction

### The T-DNA insertion mutant libraries provide general access to alleles for gene functional studies

Mutagenesis has been a central tool for studying the genetics underlying biological traits, as phenotypic analysis of mutants provides a direct method to measure a gene's contributions to biochemical, cellular, tissue and organ characteristics. In a mutant genotype where a polymorphism alters a single gene's functional output, the isolated activity of that gene *in vivo* can be assessed by phenotypic comparison to the wild-type parental genotype. Furthermore, eukaryotic organisms harboring multiple mutations, often generated by sexual hybridization between single mutants, are valuable for characterization of more complex

Contact information for authors: Ronan C. O'Malley (omalley@salk.edu), Cesar Barragan (barragan@salk.edu), Joseph R. Ecker (ecker@salk.edu).

interactions such as epistasis, functional overlap, and sub-functionalization. Though biological assignment of gene function has always depended heavily upon phenotypic analysis of mutants, currently, only ~12% of Arabidopsis gene function assignments are based on *in vivo* characterization (1). Furthermore, while as many as 60% of Arabidopsis genes do have some inferred function, these characterizations are often based on relationships such as sequence homology to better-characterized genes and, thus, these inferred functions may often be incomplete or even inaccurate (1). With recent advances in genomic-scale tools and methods, we are beginning to see a rapid increase in the scope and quality of inferred gene functions (1, 2), but as computer models of genetic networks develop more complicated predictions about specific interactions, further characterization of mutant alleles by phenotyping will likely be required to support and extend the models (1, 3). Moreover, as a primary goal of plant research is crop improvement, mutant analysis will likely always be important as a tool for examining the *in planta* effects of the alteration of a gene function.

While new methods for targeted mutagenesis such as CRISPRs and TALENS are being developed, concerns related to specificity and off-target effects still need to be worked out in order to make these methods standard laboratory techniques (4–6). Even when robust eukaryotic genome editing tools allow for the average laboratory to inexpensively generate custom alleles, the availability of mutants for a new target gene may still be limited by the organism-specific features of transformability and lifespan. Thus, even with facile editing tools, if large numbers of genes will needed to be tested, which is likely to be the case as gene functional predictions improve, access to mutant alleles could become a bottleneck for confirmation and further characterization of predictions. One solution to the problem of immediate mutant allele access for any gene is the creation of very large collections of sequence-indexed insertion lines for an organism. A sequence-indexed mutant collection typically consists of several hundred thousand individual lines in which the precise genomic location of a mutation(s) in each line is determined by DNA sequencing. As some portion of the synthetic polymorphisms will be in or proximal to genes, these mutations commonly result in the loss or disruption of gene function. By creating a very large population of individually sequenced mutants, gene disruption alleles can be identified for almost all genes in an organism (7). Due to the value of such a resource, this approach has been applied to create sequence-indexed mutant collections in several organisms including mouse(8), zebrafish(9), Drosophila(10), and Arabidopsis (reviewed in (11)).

*Agrobacterium tumefaciens* transfer-DNA (T-DNA)-induced insertion mutant collections in *Arabidopsis thaliana*, created in the late 1990's and early 2000's as an international effort to saturate the gene-space with mutations, have been a particularly important resource for plant biology (11). The high gene-space coverage in these collections of lines is in part due to the relative ease with which T-DNA insertional mutagenesis can be used for creating large sequence-indexed collection in plants. The insertion of a T-DNA fragment into a plant host genome is a consequence of a natural transformation process where an *Agrobacterium* infection results in the transfer of a DNA fragment flanked by 25 bp border sequences (the T-DNA) from a heavily modified tumor inducing Ti plasmid into the infected plant's genome (12).Highly-efficient T-DNA transformation protocols are available for Arabidopsis (13) and because the T-DNA inserts randomly (7) and is an effective gene-disrupting mutagen, the generation of the large mutant populations required for gene-space coverage is possible. Furthermore, because the T-DNA insert contains a known DNA sequence, primers designed from the left border (LB) of the T-DNA can be used to isolate the genomic/T-DNA sequence junction in a high-throughput fashion. The genomic portion of this sequence, commonly known as the flanking sequence tag (FST), can be mapped to the genome to precisely identify the chromosomal insert location for many individual lines.

Insertional indexing of large populations of T-DNA transformant lines has been used to achieve mutant allele coverage for the majority of Arabidopsis genes. In this chapter we will describe these *Arabidopsis thaliana* T-DNA insertional mutant collections with a particular focus on the mutant collections in the Colombia (Col-0) accession: the SALK, GABI-KAT, SAIL, and WISC lines(14–17). These lines, generated by several laboratories including ours, contain in total over 260,000 individual mutant lines and represent potential disruption mutants for most *Arabidopsis thaliana* genes. In addition to these four Col-0 collections, T-DNA insertional mutants are also available in other backgrounds, such as the FLAG collection lines in Wassilewskija (WS) (18), as well as Arabidopsis transposon-insertion collection, such as the CSHL and RIKEN lines in the Landsberg erecta and Nössen accessions, respectively (19, 20). However, as the Col-0 T-DNA collections have been the most heavily utilized, we will primarily focus on the SALK, SAIL, GABI-Kat and WISC lines, which we will collectively refer to here as the T-DNA collection.

The T-DNA collection has been used as a resource for thousands of published studies to address highly-varied questions in plant biology (11). This extensive use of the Col-0 T-DNA collection is in part due to the fact that *Arabidopsis thaliana* Col-0 accession has been a primary model in plant research for several decades, and is currently the only Arabidopsis accession with a high-quality reference genome(21). Additionally, the heavy use of this collection may also be attributable to the ease with which data and seed material from the collection can be accessed by researchers. Our website, T-DNA Express (http://signal.salk.edu/cgi-bin/tdnaexpress), is the primary portal to the T-DNA line information and includes search and analysis tools (iSect) to assist in the design of experiments to effectively leverage this collection. Additionally, seed lines are also easily obtained, and can be directly ordered from seed repositories for a small charge: the Arabidopsis Biological Resource Center (ABRC: https://abrc.osu.edu/) and The European Arabidopsis Stock Centre (NASC: http://arabidopsis.info/) for U.S./Canada and Europe, respectively. These requested mutant lines generally ship to researchers within days of placing an order. In just the past fives years over 800,000 lines have been shipped from the T-DNA insertion line collections from ABRC alone (personal comm. Debbie Crist).

## 2. Materials

### 2.1 Plant Growth

1. Plastic greenhouse pots 2.5” deep (Growers Solutions #P64). Plastic Grids These plastic grids can be constructed by cutting a florescent light covering (# LP2448EGG-5) into pot-size using a bech-saw (Ridgid #14653). Soil-vermiculite 3:1 mix (Sunshine® Mix #1 / LC1) medium vermiculite (69950-V-3-6) Fertilizer (plantex #70438).

### 2.2 DNA Extraction Components

1. Tweezers, scalpel for leaf tissue extraction. 96 deep-well plate (USA Scientific # 1896-1000), Pipette tips (USA Sci. Tip One1111-1700), 2.38 mm stainless steel beads (# UX-04728-59). Paint shaker (Harbil 5G-HD Mixer # 4700510), aluminum foil tape (3M # 4380). These components will be use for tissue grinding.
2. V bottom plate (Fisher # E951040308), centrifuge (Beckman Allegra 25R # BK369434) with plate rotor (Beckman # 09T 368),
3. isopropyl alcohol (EMD # PX1835-5), 70 % ethanol, in a graduate cylinder add 70 mL of ethyl alcohol 200 proof (Pharmco # 64-17-5) then add 25 ml of sterile nano-pure water. These components will be use for DNA precipitation.

### 2.3 PCR Genotyping Components

1. dNTP mix (10 mM): Measure 6 mL of sterile nano-pure water into a conical tube. Add 1 mL each of 100 mM dATP, dTTP, dGTP, dCTP. Vortex well and store at −20C in 1 mL aliquots.
2. 10× Buffer: In a 100 mL bottle, add 34.5 mL of sterile nano-pure water. Add 10 mL of 2.5M KCl, 5 mL of 1M Tris-HCl, and 500 μL of 2.5M $MgCl_2$. Adjust the pH to 8.3, mix well and store at 4 °C
3. Primer mix: For the WT primer mix, we use 20 μM gene specific primer stocks in the forward and reverse directions. To make the mix, add 47 μL of sterile nano-pure water to a well. Mix in 1.5 μL of the forward primer and 1.5 μL of the reverse primer, for a final concentration of 0.06 μM primer mix. Vortex well and store at 4C (−20C for long-term).
4. For the T-DNA primer mix, we use a 20 μM T-DNA left border (LB) primer stock and the 20 μM gene specific reverse primer stock. For the mix add 47 μL of sterile nanopure water to a well. Add 1.5 μL of the LB primer and 1.5 μL of the reverse primer. Vortex well and store at 4 °C (−20C for long-term).
5. PCR MIX Keep it on ice: For 100 reactions, add 560 μL of sterile nanopure water to a 1.5 mL tube. Add 150 μL of 10× Buffer, 30 μL of the dNTP mix (10 mM), and 10 μL of Taq. Mix gently.
6. 96 well PCR Plate (USA Scientific 1402-9599). Gene Amp PCR System 9700(P/N N805-0200)

### 2.4 Gel Electrophoresis Components

1. 50× TAE Buffer. Measure 1 L of DI water with a graduate cylinder and pour 200 mL of it into a 2 L container and a stir bar, add 18.61 g of $Na_2EDTA$ (disodium), 242 g of Tris (Trizma) and 57.1 mL of glacial acetic acid add the remaining DI water to the container and wait for the

components to mix and dissolved adjust the pH to 8.5. Dilute to 1× TAE stock buffer.

2. 3% Agarose gel for gel electrophoresis. Place a stir bar in a conical flask, add 100 mL of 1× TAE buffer and 3 g of biological grade agarose (BioPioneer #C0009) heat the mix in a microwave oven to melt the agarose, wait for the agarose to cold down (you should be able to hold the in your hand wearing a nitrile glove) and add 2 μL of Ethidium bromide (EtBr). Pour the liquid gel into a gel multi-caster (Run One EP1019) and place the multi-caster combs in a 16+2 configuration and let it dry.
3. 3 PCR dye. in a beaker add 350 mL of DI water and stir bar, slowly add 150 mL of glycerol then add 0.08 g of Bromophenol blue (Sigma-Aldrich 114405-5G).
4. Run One unit w Blue Multi-Caster, 110V. (EP-2015 w/EP-1019) Blue Comb Multi-Caster Dual Configuration (8+1)/(16+2).
5. 5 Gel logic 200 imaging system, Kodak ID Image Analysis.

## 3. Methods

This section provides a guide to how to access the T-DNA collections using our website T-DNA express, and how to genotype the T-DNA insert to identify homozygous lines. Our genotyping protocol describes a high-throughput pipeline we have implemented in our laboratory. Individual sections will include how to select a line, design genotyping primers, grow, genotype, and interpret results. The notes include descriptions of common types of problems with the T-DNA lines and files of lines we have found to be problematic in regards to isolating a homozygous line.

### 3.1. Navigating the T-DNA Express site for identification of appropriate insertion alleles

1. In the following section we will discuss how to access T-DNA collection information using the T-DNA Express web site with a particular focus on understanding the data underlying the browser views and how to best utilize this information for line selection
2. Entering the T-DNA Express (http://signal.salk.edu/cgi-bin/tdnaexpress) in an internet browser address bar opens up a genome browser with graphical representation of genes and insertion lines associated with a genomic region (*see* Figure 1A, T-DNA Express).
3. Genes models are represented as dashed green lines where each dash marks an exon, and the 3’ most exon is shown as an arrow to indicate the directionality of coding sequence (CDS): right-pointing arrows are for genes on the Watson strand (e.g., At1g01010, *see* Figure 1A, “a”), while left-pointing arrows are for Crick-strand genes (e.g., At1g01020, *see* Figure 1A, “b”).

**4.** The T-DNA insertion FSTs are represented as a thinner arrow (below the blue chromosome) color-coded by collection. This arrow is the graphical representation of the FST BLAST match mapped to the Arabidopsis genome, currently TAIR10 (*see* Figure 1A, "c").

**5.** As additional data sets not discussed in this manuscript are also presented in the TDNA-Express browser (eg. transposon insertion collections), the "View Editor" tool indicated in Figure 1 with an red arrow can be used to limit the set to the following data types: "SALK T-DNA", "SAIL FST", "GABI-kat FST", and "Wisc FST".

**6.** Using the search box (*see* Figure 1A, "d") one can query a specific Arabidopsis Genome Initiative (AGI) gene identifier. We will primarily use the example gene At1g01010 for the remainder of this paper.

**7.** A T-DNA line can also be queried using a collection-specific identifier. As a representative T-DNA line we will use the line GABI_414G04 (*see* Figure 1A, "c") which has an insert in the first exon of At1g01010,. A search with GABI_414G04 would bring up a similar window but will be centered on the insert.

**8.** For a Watson-oriented LB insertion (when an arrow points right) the predicted T-DNA insertion position is the left most, or lower chromosomal coordinate, and for a Crick-strand insertion (when an arrow points left) the higher coordinate marks the insertion site.. A detailed description of the FST as it relates to the arrow representation on T-DNA Express is in Section 4.1.1.

**9.** To get the the genomic coordinates of the predicted insertion site in a text format click on the gene or T-DNA arrow to open the "Data View" web page. (see Figure 2). The "Data view" web page contains all the genomic coordinate data associated with the graphical representation of the gene models and FSTs seen in the "Gene view' (see Figure 2).

**10.** In Figure 2, we have annotated the shared "Gene view" and "Data view" information for At1g01010 and it's associated insertion lines by using the same lower case letter indicator in the two panels.

**11.** For the line GABI_414G04, where the left border is in the Watson orientation, the lower chromosomal location is the insertion site (*see* Figure 2, "a"). For the Crick-oriented line, SALK_128569, it is the higher chromosomal position that corresponds to the T-DNA insertion site (*see* Figure 2, "b"). However, not all inserts in the collection capture the insert and may begin on average as much as 300bp upstream of the insertion site (300 more bases back from the blunt end of the arrow). In Section 4.1.2 we provide more technical issues regarding insertion site FST location mapping.

**12.** The "Data View" page also provides the BLAST-score of the FST mapping (*see* Figure 2, "c"). The BLAST-score, tagged as "EVAL" in the

"Data View", is important for determining the confidence level of the location of mapped FST. In the examples shown in Figure 2, the e-values are very low (6e-44 to 1e-101) and thus give good confidence in the FST placement in the genome (see Figure 2, "c"). Generally, most lines do have very good FST BLAST scores, but values below 1e-15 do exists and should be recognized to avoid selecting lines with a higher false positive rate. Section 4.1.3 provides more details on judging FST score and how to easily directly inspect any FST from our database (http://natural.salk.edu/database/tdnaexpress/)).

**13.** Researchers should be aware that while T-DNA transformation is known to frequently result in multiple independent insertions in a single line. The T-DNA lines of the collections contain on average 1.13 annotated inserts per line. For the 13% of lines with more than one insert, T-DNA Express will display the first occurrence in the genome and links to the additional insert loci immediately below the genome browser window displayed as numbers ("1, 2 …") providing links to browser view for each locus. The true number of inserts per line in these collections has not yet been unequivocally established (Section 4.1.4).

**14.** In addition to searching for insertion lines by gene identifier, the T-DNA Express can also be navigated by chromosomal position (*see* Figure 1A, "a"). The transcriptional start stie of At1g01010 is chromosome 1 at position 3760. To go to this location enter "3760" in the search box and select "1" from the drop-down list on the "Chr" button. This selection will open a new window where a grey line intersecting the chromosome will show the exact base pair position you selected. This is a useful method for examining the location of genomic features, relative to T-DNA insertion sites.

**15.** The collection can also be searched using a DNA sequence BLAST on the T-DNA Express site (*see* Figure 1A, "e"). By selecting "Seq" link for At1g01010, the fasta sequence of the gene is recovered, and one can use the first 50 base pairs to test the BLAST tool. (ATGGAGGATCAAGTTGGGTTTGGGTTCCGTCCGAACGACGAGGAGCTCGTTGGTCACTAT). The BLAST results for the user-entered sequence will be presented in a separate table and will contain chromosome coordinates and BLAST match score (e-value). To display this information in a browser, click on individual BLAST results links to show genes, insertion sites, along with a graphical representation of the user-entered sequence that is aligned to the Arabidopsis genome. This tool is particularly useful for positioning and orienting primers for genotyping assays.

**16.** The site may also be searched by "function" such as a common gene name (e.g., *EIN2*), biochemical assignment (e.g., ethylene),) (*see* Figure 1A, "a"). Queries a list of linked genes with relevant genome annotations that may be associated with T-DNA insertion sites.

### 3.2 Identifying the best insertional lines for further study

1. The location of the T-DNA insertion relative to gene annotation is the best indicator of whether a gene function will be disrupted or absent in the mutant.

2. When selecting a T-DNA line, we prioritize inserts by the following order: coding sequence (CDS) exon, CDS intron, 5'UTR and promoter (500bp before the transcriptional start site, TSS). Additionally, for the T-DNA mutants disrupted in the CDS region, inserts more 5' proximal to the start codon may also have a higher likelihood of impairing gene function at a transcriptional or translational level through truncation of the mRNA or protein, and thus are preferred according to our selection criteria.

3. For the example gene At1g01010, the GABI_414G04 line is the best candidate due to its location in a CDS exon, with its location most proximal to the TSS of the gene (*see* Figure 2A, "a"). A meta-analysis of published T-DNA insertion lines supports this order in regards to its effectiveness of disruption of gene function, as 90% of CDS inserts result in a knockout, while insertions in front of the start codon (5' UTR and promoter) resulted in a both knockouts (25% of cases) and a knockdowns (67% of the cases) (22).

### 3.3 Designing genotyping primers with the iSect tools

1. Once a T-DNA lines have been selected and ordered from the ABRC (https://abrc.osu.edu/) or NASC (http://arabidopsis.info/) stock center, the next step is typically the isolation of homozygous segregants from the parental stock which may be hemizygous for the T-DNA insert,

2. A two-step PCR genotyping assay provides a powerful and scalable method to identify plants homozygous for any T-DNA insert from individual segregants (Figure 3).

3. The first PCR reaction uses a gene/genome specific primer (GSP) pair that spans the predicted T-DNA insertion site. This PCR reaction detects the presence of a wild-type copy of the gene in the plant (Figure 3). While a WT copy of the gene (the one lacking the T-DNA insertion), you will amplify a band in a wild type or heterozygous individuals. No band will amplify for a homozygous plant howerver as both copies of the gene contain the T-DNA insert whose large size prevents amplification in a PCR reaction. Thus, the absence of a wild-type PCR product is a strong indicator that the line is homozygous for the insert.

4. The second PCR reaction is performed to confirm that the candidate homozygous line contains a T-DNA insert at the predicted chromosomal location (Figure 3). This PCR reaction selectively amplifies the T-DNA/

genomic DNA junction sequence, the FST, using a combination of a left border primer and the correctly oriented GSP.(i.e. the GSP associated with the opposite strand as the left-border primer).

5. In the second PCR only the wild type copy does not contain the T-DNA/ genomic region and will not amplify, while the heterozygous and homozygous segregant samples will produce a T-DNA band. Thus, a homozygous line will show the unique pattern of −WT and +TDNA. The wild type will be +WT/−TDNA and heterozygotes are positive for both.

6. **Successful genotyping relies on the selection of good primers that flank the insert location and target the left border of the T-DNA**.

7. Our iSect tool website http://signal.salk.edu/tdnaprimers.2.html provides a database of GSP primer pairs for testing any line in the collection. For design of the GSP, a user can control primer properties using the iSect web tools, though we typically rely on the default settings that produce reliable genotyping primers.

8. If a request for a primer pair returns more than one primer set, then the requested line has more than one annotated insertion. Be sure to select the primer set that matches the chromosome location of the insert that you are interested.

9. In Table 1 we provide a list of collection-specific T-DNA left border primers that we have regularly used in-house.

10. If you prefer to design your primers by hand, you can use this link http://signal.salk.edu/isect.2.html to retrieve the gene sequence with or without introns, and use this sequence or coordinates to create primers here http://signal.salk.edu/isectprimers.html or with another third-party primer design program.

### 3.5 High-throughput DNA extraction from Arabidopsis leaf tissues

1. DNA extraction is executed using a simple, but robust and highly-scalable protocol. To increase the probabilities of isolating a homozygous specimen from a segregated population we recommend to genotype at least 16 individual plants.

2. Fill the plastic pots with soil mix and add water, place the plastic insert on top and wait a few minutes for the water to be absorbed.

3. Plant single seed from the line to genotype into every other square grid. A wet toothpick is often a good tool for this, and because we are directly seeding the soil, seed sterilization is not necessary. If germination for a line is not 100% multiple seeds can be place in a grid position and culled down to one plant after germination.

4. The identification of the individual T-DNA lines is determined by its location in a plastic grid (*see* Figure 3B). The grid provides a simple but

highly effective means to track individual segregants without the need for individual labeling of mutant lines. We typically plant only 8 plants-per-pot alternating every other grid position to avoid crowding (*see* Figure 3B). A position starting in the upper left hand corner is the first plant in a set while the label is placed on the front of the pot. This provides a visual cue for pot orientation during sample collections.

**5.** After the plants have grown to the point that they have more than four leafs, we cut off a single leaf and place it into a well of a strong 96 deep-well plate for DNA extraction. Younger leafs are preferable to older ones as they generally produce higher quality DNA in our hands. A leaf ~0.5cm in length, which weighs about ~15–20 mg, can be used, though good quality DNA can be recovered using more or less input material than this.

**6.** Once leaf tissues representing all the individuals to be genotyped have been collected in the 96 deep-well plate, add a stainless steal metal bead and 300 μL extraction buffer, seal the plate with aluminum foil tape.

**7.** To grind the tissue, place the deep-well plate(s) with your samples inside the paint shaker for 2.5 min. The tissue will be grinded by the metal bead up and down motion caused by the paint shaker. Centrifuged the deep-well plate to a separate the tissue from the DNA and proteins. The plant debris will pellet to the bottom of the plate.

**8.** Transfer 50 μL of the supernatant to a V bottom plate and add 50 μL of isopropanol, store at −20°C for 30 min.. Be careful not to dislodge any plant material at the bottom of the tube. Section 4.1.5 provides a brief description of how this can be done at quickly without having to inspect the location of the tissue pellet or supernatant.

**9.** To precipitate the DNA, centrifuge the plate and discard the supernatant by inverting the plate onto a stack of about 10 paper towels .Use a separate stack of 10 paper towels to tap-dry of any remaining solution while the plate is still inverted.

**10.** Return the plate to its normal position and add 100 μL of 75% ethanol and repeat step 9.

**11.** The plate is then inverted onto an additional stack 3–5 of paper towels, placed in a centrifuge bucket and pulse centrifuged for 5s while the plate is inverted to remove an retained ethanol. A DNA pellet cannot be typically seen at this point however there will be sufficient DNA for PCR so a researcher should not be concerned if nothing is visible in the well. We have observed that centrifugation for longer times at higher gravities (6000g) will not dislodge the precipitated DNA, so pellet loss is not a concern.

**12.** The plate is air dried for 15 minutes at room temperature, and re-suspended in 20uL of Tris/EDTA buffer (10 m Tris/1 mM EDTA). The DNA concentration will be vary among the wells but using the subsequent

PCR set up we see robust and highly consistent results so we generally do not check DNA concentrations unless significant failure rate is seen in the PCR genoytyping steps.

13. This simple protocol produces a very high quality DNA template, and is designed to allow for rapid processing of 96 plants per plate set up by multi-pipetter, so it is accessible to most laboratories.

### 3.6 The PCR genotyping assay

1. To set up the PCR reaction to span the genomic region (wild type reaction) in each required well add 7.5 μL of PCR mix, 5 μL WT primer mix, 2.5 μL DNA

2. Vortex and quick spin the plate in a centrifuge, place the PCR plate, pre heat the the thermo-cycler head before placing the PCR plate in it.

3. To set up the PCR reaction for the T-DNA-genomic region (T-DNA reaction) in a 96 well PCR plate add 7.5 μL of PCR mix, 5 μL T-DNA primer mix, 2.5 μL DNA and repeat step 2.

4. Both reactions can be run using the same thermo-cycler program.

| | |
|---|---|
| 1. 94 °C | 3 min |
| 2. 94 °C | 30 sec |
| 3. 60 °C | 30 sec |
| 4. 72 °C | 2 min |
| 5. 72°C | 10 min |
| 6  4 °C | hold |

5. Place the 3% gel in the in the RunOne unit and top up with 1× TAE buffer.

6. To analyze the PCR products add 2 μL of PCR dye to every PCR well, gentle mix pipetting up and down and load 10 μL of the product mix into the wells of the 3% gel and let it run for a minimum of 12 min. at 100 volts. As the casted gel will have 16 wells per line, using the 8 tips multichannel pipette will allocate the PCR sample every other well, this set up will allow to compare the t-dna and wt reaction side by side (see figure 3C)

7. The expected result from a genotyping assay will be approximately 1:2:1, wild type:heterozygous:homozygous, individuals in a T2 population (*see* Figure 3C).

8. As the lines available from ABRC will be T3 or later generation, one may expect to see some distortion of this ratio with a higher percentage of wild types and homozygotes, and a depletion of the hemizygotes. Section 4.2 provides troubleshooting suggestions if your genotype results differ from the expected pattern.

## 4. NOTES

### 4.1 Information regarding the FST mapping, inserts-per-line, and genotyping techniques

1. It is important to examine the direction of the T-DNA when determining the chromosomal location of the predicted insertion site. All of the FST shown on the T-DNA Express site are sequenced out of the left border of the T-DNA and extend into the genomic sequence as the left border has as a general rule been observed to be more easily recovered by PCR than the right border. Thus, for an FST that captures the entire T-DNA/genome junction, the beginning or blunt end of the T-DNA arrow represents the point between the first genomic base and the last T-DNA left border base.

2. The Sanger-sequenced FSTs predicted start sites are not always precise due to poor sequence trace information. Additionally, overlapping FST traces from multiple inserts will all be sequenced simultaneously as individual PCR bands from independent FST were not isolated. As a result only after enough sequence cycles such that only the longest FST is still incomplete will the trace be readable… We generally assume that the actual insertion site may be as much as 300 base pairs away from the annotated insertion site, particularly for position-sensitive applications like designing primers for T-DNA line genotyping. If the user wants to know the precise location of the insert one can re-sequence the line using the same approach used to originally capture the FST in the SALK collection and can isolate individual bands for sequencing if they wish to determine the sequence of more than one T-DNA in the line (23).

3. A higher e-value (lower confidence FST mapping) could be due to factors related to the length (i.e., less than 100bp) or trace quality of the FST, or mapping to genomic repeat regions, all of which can interfere with correct FST placement. To directly inspect the alignment of the FST against the genome, the original T-DNA FST can be recovered by clicking on the "Seq" link associated with the insert (*see* Figure 2, "Data View"). Alternatively, the entire list of all FST FASTA sequences can be downloaded from our site for further analysis (http://natural.salk.edu/database/tdnaexpress/).

4. The total number of inserts-per-line has been estimated to be at 1.5 based on antibiotic/herbicide selection, but this is likely to be an underestimate as these selection genes are known to be silenced in these lines (14–16).

5. The pipette tips we use are matched to our harvest/extraction plates (information for both provided in Section 2.2.1, such that a bevel on the tip will identify the depth at which supernatant can be recovered without aspirating plant debris. The primary advantage of this approach is when a 8 or 12-channel pipette is used for very rapid processing of a entire plates for high-throughput DNA extractions. The mechanics for using pipette tip bevel for this purpose are as follows: the tip is inserted at an angle such

that the end of the plastic tip is pressed to the opposite wall of the plate well. The pipette tip is then slide down the wall until the bevel is lying on top of the wall on the closer side. Other beveled pipette tip and 96-well combinations may work though those presented in the Materials section are used regularly in our laboratory for this step. Additionally it is possible to add more extraction buffer as it is unlikely to effect the quality of the PCR amplification and can be used to raise the level of the supernatant for greater separation of pellet and pipette tip.

### 4.2 Troubleshooting genotyping results

1. In this section we will describe how to troubleshoot genotyping results that fail to produce evidence of a T-DNA insert at the site, or that do contain a T-DNA but no homozygous progeny or identified.

2. The two most common types of genotyping results that do not yield homozygous lines are: i) no T-DNA insertion identified by the T-DNA/ genomic DNA PCR (though wild type does amplify) (Sections 4.2.3-9) and ii) lines which contain produce only wild-type and hemizygous plants, but no homozygous lines (Sections 4.2.10-12).

3. **The first class are lines which do not appear to contain the annotated T-DNA insertion**. Some lines in the collections are false positives. From an analysis of data from genotyping several thousand SALK and SAIL collections we have found that a small but significant proportion of sequenced insertion lines do not contain the T-DNA at the identified locus (12.6% of Salk and 14.5% of SAIL lines). This assignment is based on a lack of amplification of any T-DNA junction PCR products and presence of strong wild-type PCR products in all individuals tested (16 or greater per line) as PCR from the non-confirming lines did produce products of the predicted size from the wild-type primers, but failed to produce a T-DNA product, indicating that the primers are functional but that the locus may not contain the predicted insert.

4. Retests of these “no T-DNA” lines with a second primer pair had a very poor recovery rate, confirming that the majority of these lines do not contain the indexed T-DNA. This could be the result of problems at nearly any stage of the original mutant indexing process, though whatever the cause, the similar false positive rates seen in the SAIL and SALK collections suggest that the creation of very large indexed mutant populations is susceptible to generating some false-positives. We have created and released a full list of lines with a confirmed T-DNA insert by genotype (http://natural.salk.edu/database/tdnaexpress/ro/ ; file: _detected) and a list of lines in which we failed to identify a T-DNA genomic junction (http://natural.salk.edu/database/tdnaexpress/ro; file: tdna_not_detected). This list is not comprehensive for the collection however and many annotated inserts are not on either list.

**5.** There is one large class of insertion lines that have a high false positive rate which can be identified either computational or visually. This class of lines is recognized as multiple insertion events that have the same or nearly the same genomic insertion location and also have very similar identification numbers (e.g., Salk_128569 and Salk_128571) (see Figure 2, "b"). For these two SALK insertion lines we can observe that they share the same insertion site (chromosome 1, base pair "4298"). When multiple lines have similar identification numbers, there is a high probability that they were processed on the same 96- or 384-well plate, the formats for all molecular biology steps used in the creating of the collections. Considering the low probability that multiple lines from the same plate would have a T-DNA insert at the same genomic location, the fact that more than one-quarter of SALK and SAIL collections fit these criteria suggests that in many of these cases the lines suffered from well-to-well **cross-contamination** during the flanking sequence tag (FST) capture.

**6.** These suspected "cross-contaminated" lines should be considered as a set and not individually when one is trying to recover a line for this insertion site. We have tested over three thousand lines from sets of suspected cross-contaminated lines, and for 38% of these sets were able to identify the correct T-DNA insertion. As expected only one line in the set typically contains the predicted insert. For the example SALK lines (Salk_128569 and Salk_128571), only SALK_128569 was found to contain the insert when both were tested (see Figure 2, "b").

**7.** Of the 134,601 lines in the SALK collection, 37,536 are possible candidates for this cross-contaminated set (http://natural.salk.edu/database/tdnaexpress/ro ; file: tdna_cross-contaminates). Since we wanted to ensure high coverage of the cross-contaminate sets, this list uses liberal cutoffs (SALK number within 1000, start position within 1000) so it is likely to include some co-incident insert events from unique T-DNA as well. Like the "no T-DNA" set described above, both the SALK and SAIL collections show a similar percentage of cross-contaminated lines indicating a common problem when large numbers of lines are genotyped in parallel in a 384-well format. As this set does contain many real and valuable alleles, it is still an important resource, but users should ordered these as sets and a pre-screen for the presence of the for a line that contains the insert.

**8.** Primer failure, is always a possibility too for no T-DNA lines. If the annotated position is off and the GSP does not span the T-DNA the line will genotype as all wild-type. To control for this possibility, it is best to build a second set of primers using the iSect tools with an additional 300bp window added onto each side. If after moving the primer pair 300bp farther apart you are still unable to identify the T-DNA, it is likely that this predicted insert is a false positive and it is better not to further pursue this line.

**9.** The T-DNA transformation is a complex event and transformation induced DNA rearrangements can distort segregation results. Additionally, as the loss of some genes can result in reduced transmission or even lethality, the inability to isolate mutants for inserts for some percentage of genes. All of these factors can contribute to a non-Mendelian segregation pattern.

**10.** **In 14.8% of cases a wild type and hemizygous, but no homozygous plants are recovered** (Figure 3E and F). This set, which we will refer to as the “no homozygous” set, may be in part due to the very large scale of our genotyping assay allows less time for pursuing problematic sets so we do not generally pursue a set after 24 plants have been tested. However, we believe that it in many of these cases, the distortion in segregation may reflect a problem with the line. We would speculate that some transformation-induced genome defect could be affecting the transmission of alleles or survival of a homozygous line (see Figure 3F).

**11.** As major genomic DNA rearrangements are known to occur and have been shown to affect both plant fertility and survival in homozygous lines, some of this “no homozogous” set may be composed of these rearrangements lines (24, 25). Generally, if a rearrangement is observed, the line may still be a valuable resource for gene functional characterization as long as the phenotype can be associated to a specific gene disruption and not another effect related to the gross genomic rearrangement (24). However, rearrangements often result in depressed transmission and increased seed abortion, so suspected rearrangement lines may not provide the best material for generating higher-order mutant combinations(24).

**12.** One additional type of mutants that should also be present in this “no homozygous” set are the class of **homozygous lethal genes**; genes required for survival of the plant (26). While these homozygous lethal mutations will be part of the “no homozygous” set, the size of the current “no homozygous” set is too large to be composed solely of these. However, if you suspect that your target gene may be required for the plant’s survival, confirmation of homozygous lethality can be accomplished by demonstrating homozygous lethality from multiple independent insertion lines of that same gene and by complementation of that defect by a wild-type transgene.

## Acknowledgments

We would like to thank the entire National Science Foundation for funding of projects related to development of the Arabidopsis T-DNA insertion line resources (NSF MCB-1122250).

## References

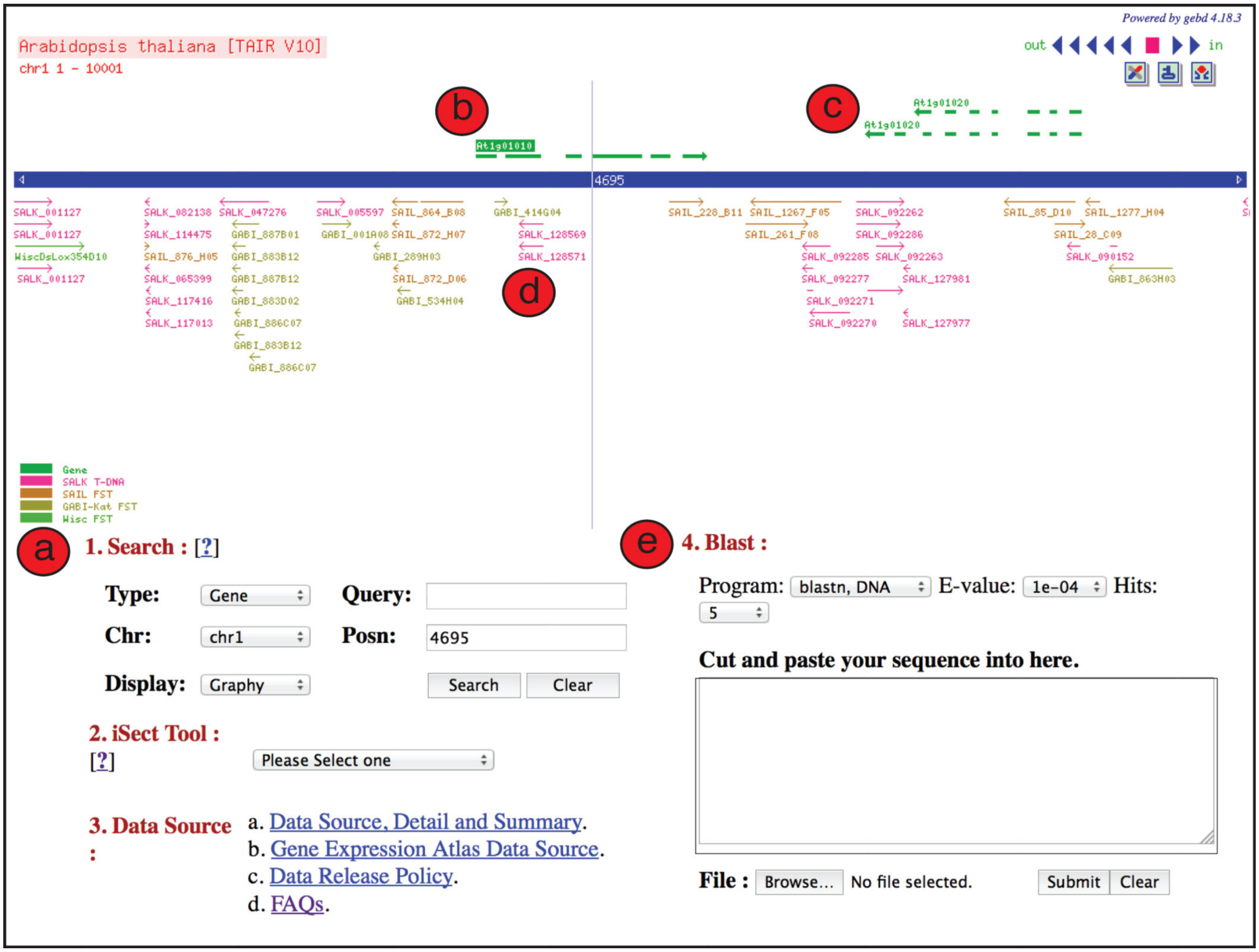
1. Rhee SY, Mutwil M. Towards revealing the functions of all genes in plants. Trends in Plant Science. 2014; 19:212–221. [PubMed: 24231067]

2. Koboldt DC, Steinberg KM, Larson DE, Wilson RK. The Next-Generation Sequencing Revolution and Its Impact on Genomics. Cell. 2013; 155:27–38. [PubMed: 24074859]
3. Carvunis A-R, Ideker T. Siri of the Cell: What Biology Could Learn from the iPhone. Cell. 2014; 157:534–538. [PubMed: 24766803]
4. Mali P, Esvelt KM, Church GM. Cas9 as a versatile tool for engineering biology. Nature Methods. 2013; 10:957–963. [PubMed: 24076990]
5. Bogdanove AJ, Voytas DF. TAL Effectors: Customizable Proteins for DNA Targeting. Science. 2011; 333:1843–1846. [PubMed: 21960622]
6. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. Nat Rev Genet. 2010; 11:636–646. [PubMed: 20717154]
7. Krysan PJ, Young JC, Sussman MR. T-DNA as an insertional mutagen in Arabidopsis. Plant Cell. 1999; 11:2283–2290. [PubMed: 10590158]
8. Elling U, Taubenschmid J, Wirnsberger G, O'Malley R, Demers S-P, Vanhaelen Q, Shukalyuk AI, Schmauss G, Schramek D, Schnuetgen F, et al. Forward and Reverse Genetics through Derivation of Haploid Mouse Embryonic Stem Cells. Cell Stem Cell. 2011; 9:563–574. [PubMed: 22136931]
9. Kettleborough RNW, Busch-Nentwich EM, Harvey SA, Dooley CM, de Bruijn E, van Eeden F, Sealy I, White RJ, Herd C, Nijman IJ, et al. A systematic genome-wide analysis of zebrafish protein-coding gene function. Nature. 2013; 496:494–497. [PubMed: 23594742]
10. Dietzl G, Chen D, Schnorrer F, Su K-C, Barinova Y, Fellner M, Gasser B, Kinsey K, Oppel S, Scheiblauer S, et al. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature. 2007; 448:151–156. [PubMed: 17625558]
11. O'Malley RC, Ecker JR. Linking genotype to phenotype using the Arabidopsis unimutant collection. Plant J. 2010; 61:928–940. [PubMed: 20409268]
12. Tzfira T, Li J, Lacroix B, Citovsky V. Agrobacterium T-DNA integration: molecules and models. Trends Genet. 2004; 20:375–383. [PubMed: 15262410]
13. Clough SJ, Bent AF. Floral dip: a simplified method forAgrobacterium-mediated transformation ofArabidopsis thaliana. The Plant Journal. 1998; 16:735–743. [PubMed: 10069079]
14. Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C, et al. A High-Throughput Arabidopsis Reverse Genetics System. The Plant Cell …. 2002; 14:2985–2994. [PubMed: 12468722]
15. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H. Genome-Wide Insertional Mutagenesis of Arabidopsis thaliana. Science. 2003; 301:653–657. [PubMed: 12893945]
16. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K. An Arabidopsis thaliana T-DNA mutagenized population (GABI-Kat) for flanking sequence tag-based reverse genetics. Plant molecular …. 2003; 53:247–259.
17. Woody ST, Austin-Phillips S, Amasino RM. The WiscDsLox T-DNA collection: an arabidopsis community resource generated by using an improved high-throughput T-DNA sequencing pipeline. Journal of plant …. 2007; 120:157–165.
18. Samson F, Brunaud V, Duchêne S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S. FLAGdb++: a database for the functional analysis of the Arabidopsis genome. Nucleic acids …. 2004; 32
19. Sundaresan V, Springer P, Volpe T, Haward S, Jones JD, Dean C, Ma H, Martienssen R. Patterns of gene action in plant development revealed by enhancer trap and gene trap transposable elements. Gene Dev. 1995; 9:1797–1810. [PubMed: 7622040]
20. Ito T, Motohashi R, Kuromori T, Mizukado S, Sakurai T, Kanahara H, Seki M, Shinozaki K. A New Resource of Locally Transposed DissociationElements for Screening Gene-Knockout Lines in Silico on the Arabidopsis Genome. Plant Physiology. 2002; 129:1695–1699. [PubMed: 12177482]
21. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000; 408:796–815. [PubMed: 11130711]
22. Wang YH. How effective is T-DNA insertional mutagenesis in Arabidopsis? Journal of Biochemical Technology. 2009; 1:11–20.
23. O'Malley RC, Alonso JM, Kim CJ, Leisse TJ, Ecker JR. An adapter ligation-mediated PCR method for high-throughput mapping of T-DNA inserts in the Arabidopsis genome. Nature Protocols. 2007; 2:2910–2917. [PubMed: 18007627]

24. Clark KA, Krysan PJ. Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines. Plant J. 2010; 64:990–1001. [PubMed: 21143679]

25. Nacry P, Camilleri C, Courtial B, Caboche M, Bouchez D. Major chromosomal rearrangements induced by T-DNA transformation in Arabidopsis. Genetics. 1998; 149:641–650. [PubMed: 9611180]

26. Lloyd J, Meinke D. A Comprehensive Dataset of Genes with a Loss-of-Function Mutant Phenotype in Arabidopsis. Plant Physiology. 2012; 158:1115–1129. [PubMed: 22247268]



**1. Search : [?]**

**Type:** Gene **Query:**

**Chr:** chr1 **Posn:** 4695

**Display:** Graphy

**2. iSect Tool : [?]** Please Select one

**3. Data Source :**

a. Data Source, Detail and Summary.
b. Gene Expression Atlas Data Source.
c. Data Release Policy.
d. FAQs.

**4. Blast :**

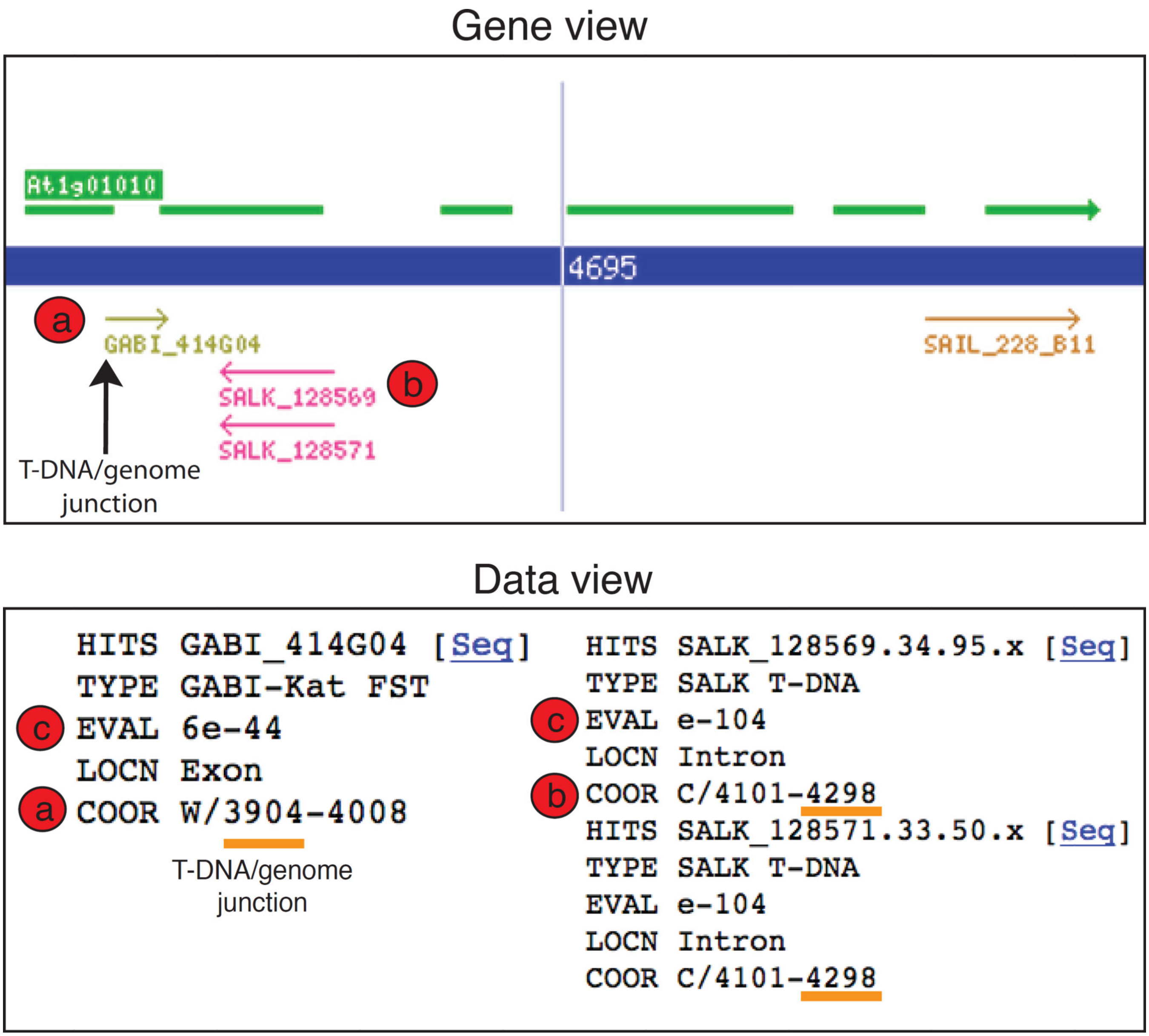
Program: blastn, DNA E-value: 1e-04 Hits: 5

**Cut and paste your sequence into here.**

**File :** No file selected.

**Figure 1. T-DNA Express features and search tools**

A view of the opening page of the T-DNA Express web site (http://signal.salk.edu/cgibin/tdnaexpress) is shown with important features annotated with a red circle. a) Indicates the search boxes used for entering gene, line, or chromosomal location for searching a line. b) The gene model for a Watson-oriented gene, and c) a Crick-oriented gene. d) The insertion line FST mapped to the Arabidopsis genome (TAIR10). e) A blast search window for using DNA sequence for searching the exact position and orientation of a feature in the genome (e.g., genotyping primer).

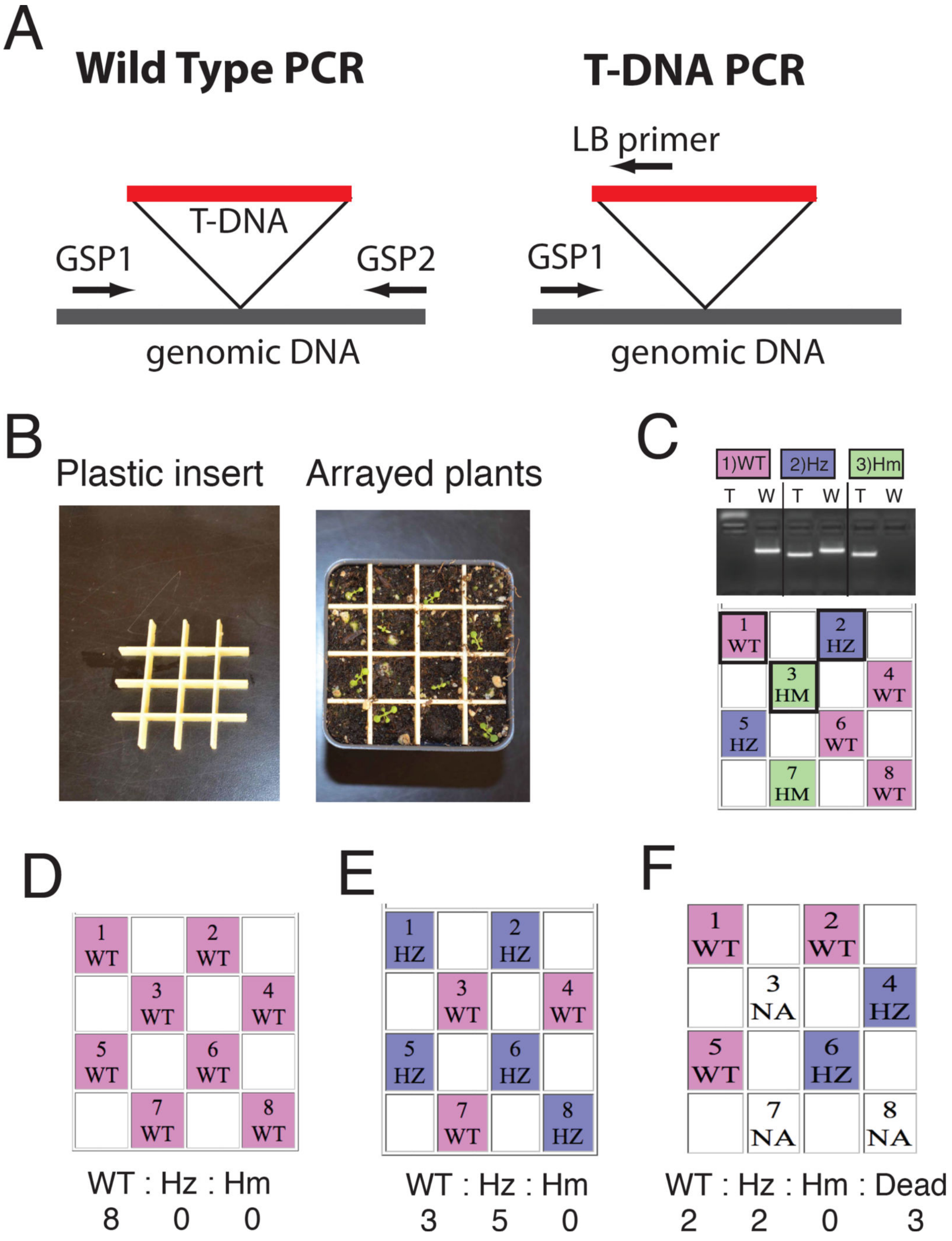


```
HITS GABI_414G04 [Seq]
TYPE GABI-Kat FST
EVAL 6e-44
LOCN Exon
COOR W/3904-4008
```

```
HITS SALK_128569.34.95.x [Seq]
TYPE SALK T-DNA
EVAL e-104
LOCN Intron
COOR C/4101-4298
HITS SALK_128571.33.50.x [Seq]
TYPE SALK T-DNA
EVAL e-104
LOCN Intron
COOR C/4101-4298
```

**Figure 2. Close-up view of gene and insert models and access to underlying annotation data**

A close-up view of a single gene (At1g01010) with associated insertion line annotation in close-up from the T-DNA Express genome browser “Gene view”. By double-clicking on a T-DNA or gene arrow representation, the underlying coordinate data used to generate the images will be shown in the “Data view” pages. a) The GABI line that is the best choice for targeting this gene for disruption. The T-DNA/genome junction site is marked on the “Gene view” for GABI_414G04, and on the “Data View” the coordinates are indicated by an orange line. b) The SALK_128569 and SALK_128571 are a pair of inserts with very similar identifier numbers and share an identical insertional coordinates underlined in the “Data view” in orange. These are a contaminated pair of lines and only one line, SALK_128569, was found to contain the actual insert. c) The BLAST score associated with the GABI_414G04.



| | | | |
|---|---|---|---|
| 1 WT | | 2 HZ | |
| | 3 HM | | 4 WT |
| 5 HZ | | 6 WT | |
| | 7 HM | | 8 WT |

| | | | |
|---|---|---|---|
| 1 WT | | 2 WT | |
| | 3 WT | | 4 WT |
| 5 WT | | 6 WT | |
| | 7 WT | | 8 WT |

| WT | Hz | Hm |
|---|---|---|
| 8 | 0 | 0 |

| | | | |
|---|---|---|---|
| 1 HZ | | 2 HZ | |
| | 3 WT | | 4 WT |
| 5 HZ | | 6 HZ | |
| | 7 WT | | 8 HZ |

| WT | Hz | Hm |
|---|---|---|
| 3 | 5 | 0 |

| | | | |
|---|---|---|---|
| 1 WT | | 2 WT | |
| | 3 NA | | 4 HZ |
| 5 WT | | 6 HZ | |
| | 7 NA | | 8 NA |

| WT | Hz | Hm | Dead |
|---|---|---|---|
| 2 | 2 | 0 | 3 |

**Figure 3. T-DNA Genotyping**

A) The two PCR reactions for genotyping an insertion line. The “Wild-Type PCR” reaction tests for the ability to amplify a genome region that will be present in wild type and heterozygous lines, but will not amplify in homozygous lines. The “T-DNA PCR” checks for the presence of a T-DNA insertion. GSP = gene/genome specific primer; LB = left-border of the T-DNA. B) A plastic grid from a florescent light cover is used as a separator for individual segregants planted form a single line. C) Typical results from a genotyping gel, with a grid matching the individual screened plants. D–E) Atypical segregation patterns observed in large-scale genotyping assays.

## Table 1

T-DNA specific primers for each of the insertion mutant collections

| Collection | PrimerID | Primer sequence |
|---|---|---|
| SALK | LB-1.3 | ATTTTGCCGATTTCGGAAC |
| SAIL | LB-1 | TAGCATCTGAATTTCATAACCAATCTCGATACAC |
| WiscDs | LB | TCCTCGAGTTTCTCCATAATAATGT |
| WiscDsLoxHs | L4 | TGATCCATGTAGATTTCCCGGACATGAAG |
| GABI-Kat | o8409 | ATATTGACCATCATACTCATTGC |