# Monsanto Arabidopsis Polymorphism and Ler Sequence Collections (Historical Record)

TAIR used to provide access to a collection of predicted Arabidopsis single-nucleotide polymorphisms (SNP) and small insertions/deletions (INDELs) between the publicly available Columbia (Col) sequence and Landsberg erecta (Ler) sequence generated by Monsanto. In addition to the polymorphisms, the Ler sequence was released. The data were kindly submitted by Dr. Steve Rounsley.

The activities previously handled by Cereon Genomics, a wholly-owned subsidiary of Monsanto Company, were relocated to St. Louis, Missouri on June 30, 2002. Any Agreements entered into with Cereon Genomics are now administered by Monsanto Company. This website text has been changed to substitute Monsanto Company in place of any references to Cereon Genomics.

These data were accessible only to non-profit institutions, universities, and colleges after user registration. TAIR reproduces the release and usage notes below for historical record. The files are no longer provided for access.



# Polymorphism Releases

There are now 37,344 SNPs, 18,579 InDels, and 747 Large Indels, a total of 56,670 polymorphisms.

## Release 1: Released May 3, 2000

**SNPs and INDELs**
There are 39315 entries (from 981 BACs). 25,274 SNPs (Single Nucleotide Polymorphism) and 14,041 INDELs (small insertions/deletions).
**Large INDELs**
632 INDELs > 100bp found in 341 of the 980 BACs used for Release 1.
Size range: Min 101bp Max 38Kb
This collection of INDELs are larger than 100bp and were omitted from the original polymorphism release due to an increased level of false positives contained within. Repetitive sequence is the major cause for such false positives, and the larger the gap between two matches, the more likely one of the matches is due to a match against a repeated region.
The data is being provided with these caveats so please use this with caution. There are however true positives in this dataset, particularly the INDELs at the lower end of the size distribution, and many people had requested access to these for the purposes of specific studies such as transposon analysis.

## Release 2: Released on November 16, 2000

**SNPs and INDELs**
Release 2 contains 4476 predicted polymorphisms from 124 BAC clones that have been sequenced by the AGI. 1,633 INDELs and 2843 SNPs were found. This is an 11% increase over the previous collection of approximately 39,000.
The remainder of the Columbia BACs are currently being processed and there will be a final update of the Monsanto collection shortly after this is completed.
**Large INDELs**
72 Large INDELs found in 43 of the 124 BACs used for Release 2.
Also provided here are INDELs greater than 100 bp. These should be treated with caution, as they are more likely to be the result of artifacts of the analysis method. However, many will still be true insertions/deletions and may therefore be of interest for certain kinds of analysis.

## Release 3: Released on March 21, 2001

**SNPs and INDELs**
Release 3 contains 12175 predicted polymorphisms from 396 BAC clones that have been sequenced by the AGI. 2905 INDELs, 9227 SNPs, and 43 Large INDELs were found. This is an 27% increase over the previous collection of approximately 45,000. The total number of polymorphisms now is 56,670.

## Column Headings

1. SNP_Name: CER(Monsanto)+Monsanto's internal ID number e.g. CER454879
2. Chromosome: 1 thru 5
3. BAC_Name: in standard AGI format, ordered by position in chromosome
4. BAC_Accession: GenBank/EMBL/DDBJ accession number
5. BAC_Length: in bp
6. Left_Coor*: The coordinate of the base to the left of the polymorphic location. See below.
7. Right_Coor*: The coordinate of the base to the left of the polymorphic location. See below.
8. SNP_Type: SNP (Single Nucleotide Polymorphism) or IND (Insertion/Deletion)
9. IND_Size: size of insertion or deletion in Columbia. Col/Ler. e.g. -4/4 means a 4bp deletion in Columbia and 4/-4 means a 4bp insertion in Col. Left blank for SNPs.
10. SNP_Base: changed nucleotide. Col/Ler e.g. A/T means A in Columbia and T in Landsberg. Left blank for INDELs
11. Left_Flank: 20 bp directly to the left of the polymorphic location
12. Right_Flank: 20bp directly to the right of the polymorphic location. See Note below.
13. Restriction_sites (Col): restriction sites in Col from the SNP/IND
14. Restriction_sites (Ler): restriction sites in Ler from the SNP/IND

*Coordinates:
For a SNP - the two coordinates flank the polymorphic base.
For an insertion in Columbia, the two coordinates flank the inserted sequence.
For a deletion in Columbia, the deletion occurs between the two coordinates listed.

## Notes

20mers are provided for locating the correct coordinates just in case the BAC sequence changes. Also note that the 20mers were supposed to be directly flanking the changed nucleotide, but the right flanking sequence may be off by one base for SNPs - but not for INDs.

Please note that the coordinates used in the datafiles refer to the originally submitted BAC sequence. Many BAC sequences at GenBank have been edited by the AGI groups in order to produce finished chromosome records. This involves removing overlapping regions, and flipping some clones in order to produce a consistent direction along the chromosome. In addition, AGI groups may make alterations at any time to the submitted sequence in order to correct errors. This can also cause the original coordinates to be inaccurate.

In order to access the original BAC sequence, you need to use the link provided in the current GenBank record. The link will look something like this:

"**COMMENT**: On Dec 16, 1999 this sequence version replaced gi:5729683"

## Usage

The flanking sequence provided in the Monsanto data files attempts to provide an alternative way to locate the polymorphism. The 20mers can be used to BLAST against the Arabidopsis genome to identify the specified location. There are some caveats to keep in mind when doing this:

- 1. This sequence should help find the appropriate location in the BAC of interest. It is not necessarily unique to the genome. It may also match other BACs in the genome, but these are not important for locating the polymorphism.
- 2. If the 20mer matches more than once in the BAC of interest, try using the other 20mer as well and combining the results. You can also use TAIR's PatMatch, which allows you to put in the polymorphic sequence as well as its approximate length in between the two 20mer set.
- 3. If the 20mer does not find a match in the BAC of interest, it could be that the editing mentioned above may have moved this location to a neighboring BAC. In this case, check your search results against the neighboring BACs.
- 4. If it still does not match, beware that using the default BLAST parameters does not always work well with such a small query sequence. Several things can increase your chances of finding a match in the BAC sequence of interest.
    - A. Use a smaller database. An example would be a species specific collection at NCBI, or the TAIR BLAST server selecting only Arabidopsis genomic sequences > 10kb
    - B. Do not filter for low complexity.
    - C. Increase the mismatch penalty to -8. This should force identical matches.
- 5. If multiple hits to the same BAC occur - do not panic. Remember, many INDELs are caused by a different copy number of a direct repeat. The flanking sequence may therefore hit multiple places. The best bet here is to pick primers several hundred bases either side of this general region.

# Ler Sequence Release

There are 81,306 sequence entries from a single-pass shotgun sequencing. It contains approximately 95 Mb of sequence.

**Released on Feb 28, 2002**

The Landsberg erecta genomic sequence provided is the result of a low coverage whole genome shotgun sequencing project carried out at Monsanto. After removal of mitochondrial and chloroplast sequences, almost 500,000 traces containing 263Mb of sequence were used for assembly. This resulted in 50,262 contigs and 31,044 singletons totalling about 92Mb.

# Reference

# General Liability Statement

**Monsanto SNP and Ler Sequence Collection**

Thank you for visiting this Website. Monsanto reserves the right to change the Agreement at any time, and you agree that your use of the Monsanto Information shall be subject to the terms and conditions set forth in the Agreement at the time the Monsanto Information is downloaded.

This disclaimer of liability applies to any damages or injury caused by any failure of performance, error, omission, interruption, deletion, defect, delay in operation or transmission, computer virus, communication line failure, theft or destruction or unauthorized access to, alteration of, or use of record, whether for breach of contract, tortious behavior, negligence or under any other cause of action.

You specifically acknowledge that Monsanto is not liable for your defamatory, offensive, infringing or illegal materials or conduct, or that of third parties, and Monsanto reserves the right to remove such materials from the Website without liability.

The contents of the Website pages, including, but not limited to text, graphics and icons, are copyrighted materials owned or controlled by Monsanto and may contain Monsanto Company's name, trademarks, service marks and trade names. Under the terms of the Agreement, you may download one copy of these materials on any single computer and print a copy of the materials for the uses permitted by the Agreement. No other permission is granted to you to print, copy, reproduce, distribute, transmit, upload, download, store, display in public, alter or modify these materials. No permission is granted here to you to use Monsanto's icons, site address or other means to hyperlink other Internet sites with any page in the Website, except as provided under "Proprietary Notices" section of the Agreement. Monsanto may make improvements and/or changes in the Monsanto Information accessible from the Website, including the terms and conditions of your use of this Website, without liability.

THE MATERIALS AND INFORMATION YOU FIND ON THE WEBSITE ARE PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING WITHOUT LIMITATION ANY WARRANTY FOR INFORMATION, SERVICES OR PRODUCTS PROVIDED THROUGH OR IN CONNECTION WITH THE WEBSITE AND ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, EXPECTATION OF PRIVACY OR NON-INFRINGEMENT. Some jurisdictions do not allow the exclusion of implied warranties, so the above exclusion may not apply to you.