# *MEDICAGO TRUNCATULA* GENOME SEQUENCE: RELEASE 1.0
**(July 17, 2006 Assembly)**

## OVERVIEW

***Genome assembly.*** As of July 2006, a total of 1,996 BACs had been sequenced and 1,826 were used in the construction of pseudomolecules, comprising 186.2 Mbp of non-redundant genome sequence (Table 1). BACs excluded from the assembly included contaminants and clones completely contained within the sequences of other, longer ones. Of the BAC clones in the assembly, 1,151 clones (122 megabase pairs, Mbp) were finished to phase 3, 371 clone (37 Mbp) to the level of ordered and oriented (phase 2), and 304 (30 Mbp) to the level of shotgun (phase 1). The BACs reside on 348 sequence contigs that combine to form 275 scaffolds (Table 1). Together with 138 singleton BACs, the scaffolds span 211.1 Mbp of genome sequence. These sequences, contigs and scaffolds were used to generate a pseudo-golden path (pgp) and an initial release of pseudomolecules for all eight chromosomes. In the assembly there are 331 gaps plus 89 BAC clones (9 contigs and 68 singletons) not yet anchored to the genetic map. Known BACs extend into one or both ends of 215 of these gaps. The average size of BAC scaffolds is 699 kbp and the N50 is 505 kbp, with sizes ranging from 87 kbp (the shortest singleton) to slightly more than 4 Mbp (Table 1). Based on the estimated size of the *Medicago truncatula* (*Mt)* genome (450-560 Mbp), the assembly so far constitutes 38-47% of the entire *Mt* genome, with coverage of the protein coding portion notably higher.

**Table 1. Assembly of the *M. truncatula* genome sequence**

| Chromosome | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | unanchored | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| BACs | 193 | 209 | 257 | 238 | 244 | 117 | 238 | 241 | 89 | 1,826 |
| Singletons | 14 | 3 | 8 | 23 | 0 | 4 | 10 | 8 | 68 | 138 |
| Contigs | 39 | 39 | 55 | 46 | 39 | 32 | 36 | 53 | 9 | 348 |
| Scaffolds | 31 | 31 | 42 | 35 | 18 | 30 | 31 | 48 | 9 | 275 |
| Gaps[1] | 44 | 34 | 50 | 57 | 18 | 33 | 40 | 55 | 76 | 407 |
| Sequence (bp) | 19,213,636 | 20,756,037 | 26,058,061 | 24,268,153 | 26,457,404 | 12,706,461 | 23,279,352 | 24,552,347 | 8,930,397 | 186,221,848 |
| Pseudomolecule[2] (bp) | 22,438,636 | 23,156,037 | 28,918,061 | 28,813,153 | 29,652,404 | 15,021,461 | 25,989,352 | 28,057,347 | 8,965,397 | 211,011,848 |
| Pseudomolecule[3] (bp) | 26,838,636 | 26,456,037 | 34,068,061 | 34,513,153 | 31,352,404 | 18,321,461 | 29,989,352 | 33,557,347 | 16,565,397 | 251,661,848 |
| Scaffold avg (bp) | 642,595 | 730,854 | 667,974 | 696,600 | 1,647,356 | 475,318 | 789,065 | 553,010 | 260,269 | 699,173 |
| Scaffold N-50 (bp) | 516,101 | 577,517 | 546,101 | 481,525 | 1,353,066 | 408,186 | 579,556 | 415,353 | 224,805 | 504,962 |

1 Gaps between scaffolds
2 Sequence found on all scaffolds and singletons, but not including 100,000 bp spacers inserted into gaps
3 Sequence found on all scaffolds and singletons including 100,000 bp inserted into gaps

***Anchoring to the genetic map.*** The sequence assembly was anchored to the existing genetic map of *Mt* though the use of 851 BAC-based simple sequence repeats, 121 BAC-based cleaved amplified polymorphic sites, and 16 single nucleotide polymorphisms using two previously described mapping populations (Thoquet et al. 2002; Choi et al. 2004). It has been reported previously that the parents of the population described in Choi et al. (2004) (Jemalong X 'A20') display a balanced translocation between chromosomes 4 and 8 (Kulikova et al. 2004). As expected, the resulting map displayed numerous biases in marker segregation and suppressed recombination on the south ends of chromosomes 4 and 8 (*Mt*-4S and *Mt*-8S). The second linkage map (Jemalong x DZA315) was free of distortion, enabling the placement of BAC-based markers accurately in these regions. Altogether, a total of 1,774 BACs comprising 177.2 Mbp (97.2% of the total) could be anchored to *M. truncatula*'s eight chromosomes by one or more DNA markers or by physical overlap with a neighboring anchored BAC clone (Table 1). The remaining 68 BACs could not be anchored, but are included in most calculations described below. Genetic marker order was collinear with the inferred genome sequence 91% of the time for intervals greater than 1 cM (centimorgan) in length. The accuracy of the sequence assembly was also validated by comparison with available BES data, where 97% of paired ends agreed with the predicted BAC clone order.

***Sequence accuracy.*** The accuracy of sequenced clones was estimated by comparing overlapping phase 3 BAC data coming from different sequencing centers. There were a total of 256 cases totaling 5.5 Mbp of such overlapping BAC sequence. Because most of the initial (seed) BAC sequencing took place at the University of Oklahoma, most cases of overlapping BACs

involved at least one Oklahoma BAC. Altogether, sequences between different centers were identical 99.99% of the time, reaching the 1 in 10,000 Bermuda standard for finished sequence.

***Estimates of genome coverage.*** To examine the nature of the genome sequenced so far and to contrast it to the unsequenced portion, the sequence assembly was compared to an earlier pilot whole genome shotgun sequence (WGS) (Mun et al. 2006). Although the earlier WGS achieved only ~0.1X coverage, it provided an unbiased representation of the overall genome. Because comparisons based on phase 1 and 2 BACs were likely to miss some alignments and thus underestimate coverage, these sequences were not included in the comparison, leaving 122 Mbp of phase 3 sequence. At a stringency of >50% coverage and >95% identity, 25.9% of WGS sequences were captured by phase 3 sequence. This value fell to 20.9% at >80% coverage and >95% identity. Extrapolating, these results indicate that the entire *Mt* genome lies between 471 and 583 Mbp in size, in close agreement with previous estimates.

To estimate the proportion of genes contained within the sequence assembly the fraction of high quality tentative consensus sequences (HQTCs) in the TIGR *M. truncatula* Gene Index version 8.0 captured by the *Mt* assembly was calculated. HQTCs were defined as those at least 500 nucleotides in length with at least five EST members, altogether numbering 9,396. Again, phase 1 and 2 BAC sequences were excluded from the comparison. At a stringency of >50% coverage and >95% identity, 37.2% of all HQTCs aligned with phase 3 BAC sequence. Increasing the stringency to >80% coverage and >95% identity lowered the percent aligned to 35.3%. Extending the estimates to the entire sequence assembly, including phase 1 and 2 BACs along with those at phase 3, these results translate to 55-58% of all *Mt* genes captured by the current genome sequence assembly.

## MATERIALS AND METHODS

***Genome assembly.*** A single accession of *Mt* cultivar 'Jemalong' commonly known as 'A17' provided DNA for three BAC libraries as a basis for hierarchical, BAC-by-BAC sequencing of the gene-rich euchromatin (Nam et al. 1999). To construct pseudomolecules, BACs whose sequences were entirely contained within other clones plus any contaminants of non-*Mt* origin were first eliminated. All remaining BACs were run through the MUMMer (Delcher et al. 2002) to find overlap regions. In this analysis, regions were required to have at least 2,000 bp and 99% identity between two sequences to be considered an overlap. For phase 2 and 3 BACs, only the terminal overlap was considered and the redundant region between two BACs was removed to form a contiguous contig sequence. For overlaps involving phase 1 BACs, the terminal criterion was not required due to the draft nature of the BAC sequences and 5,000 Ns were incorporated into the contig sequence to denote the presence of a phase 1 BAC. All contig sequences were then joined or extended to form scaffolds through the use of paired BES. BLAT (Kent 2002) was used to align all BESs against the BACs. Only low copy BES pairs (both ends hitting fewer than two times) were used and parameters for this analysis were 99% identity and 90% coverage of the BES. In cases where scaffolds could be formed by paired BES, 50,000 Ns were inserted between neighboring contigs to construct scaffolds. The final step of assembly was anchoring and ordering scaffolds onto the eight chromosomes by reference to genetic maps composed of DNA-based genetic markers (Thoquet et al. 2002; Choi et al. 2004). All scaffolds or singleton BACs that could not be anchored to the genetic map were collected together onto an unanchored chromosome 0. A spacer of 100,000 Ns was inserted between scaffolds that could not be spanned by any BAC or paired BES.

***Gene annotation.*** Annotation was carried out by the International *Medicago* Genome Annotation Group (IMGAG) (Town 2006). BAC sequences were first masked using the in-house repeat database and then subjected to gene prediction by FGENESH using a *Mt* matrix (http://www.softberry.com/berry.phtml). All available *Mt* ESTs were mapped to BACs using the Program to Assemble Spliced Alignments (PASA), requiring 95% identity over 90% of the EST length (Haas et al. 2003). The FGENESH predictions that had been enhanced by PASA-based

cDNA/EST updates were included as additional evidence for the Eu'Gene pipeline, which combines both intrinsic (*ab initio* predictions) and extrinsic (cDNA/EST and protein matches) data to produce its gene structure predictions (http://medicago.toulouse.inra.fr/imgag_egn/cgi-bin/egn_getinfo.cgi).

Gene predictions were then assigned a confidence level according to the nature of the evidence supporting their structure: **F** - complete coverage by FL-cDNA(s) or EST(s); **E** - expressed sequence match(es); **H** - homology to protein(s) from *Mt* or other species; **I** - predicted by *ab initio* algorithm(s); **L** - low confidence gene calls, in other words, gene calls not in categories **F**, **E**, or **H** and also less than 100 aa in length. The predicted proteins were then searched against the INTERPRO collection of databases and against a comprehensive non-redundant in-house protein database at TIGR. Automatic assignment of gene function was based first on the most significant domain match(s) in INTERPRO or, failing that, on the best match in the TIGR database.

## REFERENCES

Choi HK, Kim D, Uhm T, Limpens E, Lim H, et al. (2004a) A sequence-based genetic map of *Medicago truncatula* and comparison of marker collinearity with *M. sativa*. *Genetics* 166: 1463-1502.

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucl Acids Res* 30: 2478-2483.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, et al. (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31: 5654-5666.

Kent WJ (2002) BLAT -- the BLAST-like alignment tool. *Genome Res* 12: 656-664.

Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR, et al. (2004) Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. *Chromosoma* 113: 276-283.

Mun JH, Kim DJ, Choi HK, Gish J, Debelle F, et al (2006) Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. *Genetics* 172: 2541-2555.

Nam Y, Penmetsa RV, Endre G, Uribe P, Kim D, Cook DR (1999) Construction of a bacterial artificial chromosome library of *Medicago truncatula* and identification of clones containing ethylene-response genes. *Theor Appl Genet* 98: 638-646.

Thoquet P, Gherardi M, Journet EP, Kereszt A, Ane JM, et al. (2002) The molecular genetic linkage map of the model legume *Medicago truncatula*: an essential tool for comparative legume genomics and the isolation of agronomically important genes. *BMC Plant Biol* 2:1.

Town CD (2006) Annotating the genome of *Medicago truncatula*. *Curr Opin Plant Biol* 9: 122-127.

## *Medicago* Genome Sequencing Consortium

**University of Oklahoma**: Bruce A. Roe[1] (Principal Investigator), Mandi M. Aycock[1], Aleksandra Davis[1], Shweta V. Deshpande[1], Amy Dickey[1], Anh P. Do[1], Trang P. Do[1], Mounir Elharam[1], Bart N. Ford[1], Ying Fu[1], Axin Hua[1], Honggui Jia[1], Huarong Jiang[1], Yi Jing[1], Steve M Kenton[1], Xiangfei Kong[1], Hongshing C. Lai[1], Jennifer Lewis[1], Shaoping Lin[1], Chelsea McIntosh[1], Rose Morales-Diaz[1], Fares Z Najar[1], Ying Ni[1], Majesta S. O'Bleness[1], Goldameir O. Osisanya[1], Angela C Prescott[1], Sulan Qi[1] , Baifang Qin[1], Chunmei Qu[1], Jiaxi Quan[1], Iryna F. Sanders[1], Steve B. Shaull[1], Ruihua Shi[1], Stephen Snow[1], Lin Song[1], Janice Te[1], Chunyan Wang[1], Keqin Wang[1], Ping Wang[1], Doug White[1], Jim D. White[1], Graham B. Wiley[1], Junjie Wu[1], Yanbo Xing[1], Xi Xu[1], Weihong Xu[1], Limei Yang[1], Ziyun Yao[1], Yang Ye[1], Peng Zhang[1], Liping Zhou[1]; **The Institute for Genomic Research (TIGR)**: Erin Monaghan[2], Tamara S. Arbogast[2], Foo Cheung[2], Luke J. Tallon[2], Erin E. Hine[2], Ryan Althoff[2], Douglas Fadrosh[2], Tamara Feldblyum[2], Kristin A. Wilhelm[2], Kimberly M. Arfsten[2], Kristine M. Jones[2], Yongli Xiao[2], Hean Koo[2], Jyoti Shetty[2], Lakshmi Viswanathan[2], Steve Ferriera[2], Christopher D. Town[2] (Principal Investigator); **Wellcome Trust Sanger Institute**: Christine Nicholson[3], Mario Caccamo[3], Karen Holt[3], Matthew Jones[3], Giselle Kerry[3], Christine Lloyd[3], Lucy Matthews[3], Sarah Pelan[3], Carol Scott[3], Sarah Sims[3], Sean J Humphray[3], Jane Rogers[3] (Principal Investigator); **Genoscope**: Francis Quétier[4] (Principal Investigator), Sylvie Samain[4], Patrick Wincker[4], Béatrice Ségurens[4], Stéphanie Fouteau[4], Arnaud Couloux[4], Agnès Viollet[4], Anne Berger[4], Claude Scarpelli[4], Jean Weissenbach[4]; **John Innes Centre (JIC)**: Giles Oldroyd[5]; **University of Minnesota**: Bing-Bing Wang[6], Xiaohong Wang[6], Roxanne Denny[6], Jay Vasdewani[6], Ben Chacko[6], Eric Boehlke[6], Ryan Bretzel[6], Min Wang[6], Ernest F. Retzel[6], Nevin D. Young[6] (Principal Investigator); **Laboratoire des Interactions Plantes-Microorganismes**: Jérôme Gouzy[7], Olivier Saurat[7], Anne-Marie Dudez[7], Philippe Bardou[7], Céline Noirot[7], Frédéric Debellé[7] (Principal Investigator); **Munich Information Center for Protein Sequences (MIPS)**: Manuel Spannagl[8], Remy Bruggmann[8], Gabi Kastenmueller[8], Octave Noubibou[8], Klaus F.X. Mayer[8] (Principal Investigator); **Gent University**: Lieven Sterck[9], Stephane Rombauts[9], Eric Bonnet[9], Yves Van de Peer[9] (Principal Investigator); **Wageningen University**: Rene Geurts[10] (Principal Investigator), Carolien Franken[10], Marijke Hartog[10], Ton Bisseling[10]; **University of California, Davis**: Douglas R. Cook[11] (Principal Investigator), Dongjin Kim[11], Jeong-Hwan, Mun[11], John C. Gish[11], JongMin Baek[11]; **USDA-ARS, Iowa State University**: Steven B. Cannon[12] (Principal Investigator), Michelle A. Graham[13], Ethalinda K.S. Cannon[12]; **Max Planck Institute for Plant Breeding Research**: Heiko Schoof[14] (Principal Investigator), Anika Jöcker[14]; **Laboratoire de Biométrie et Intelligence Artificielle**: Thomas Schiex[15]; **Agricultural Biotechnology Center, Szeged**: György B. Kiss[16] (Principal Investigator), Péter Kaló[16]; **National Center for Genome Resources**: Gregory D. May[17], Joann Mudge[17]; **Samuel Roberts Noble Foundation**: Richard Dixon[18]

**Genome Sequence Annotation and Analysis**: Eric Bonnet[8], Steven B. Cannon[12], Foo Cheung[2], Jérôme Gouzy[7], Michelle A. Graham[13], Anika Jöcker[14], Klaus F.X. Mayer[8], Erin Monaghan[2], Christine Nicholson[3], Stéphane Rombauts[9], Thomas Schiex[15], Heiko Schoof[14], Manuel Spannagl[8], Lieven Sterck[9], Christopher D. Town[2], Bing-Bing Wang[6], Xiaohong Wang[6]

**Genome Sequencing Steering Committee**: Bruce A. Roe[1], Douglas R. Cook[11], Frédéric Debellé[7], Rene Geurts[10], Klaus F.X. Mayer[8], Giles Oldroyd[5], Francis Quétier[4], Jane Rogers[3], Christopher D. Town[2], Nevin D. Young[6]

[1] Advanced Center for Genome Technology, Department of Chemistry and Biochemistry, Stephenson Research and Technology Center, 101 David L. Boren Blvd, Rm 2107, University of Oklahoma, Norman, OK 73019, USA

[2] The Institute for Genomic Research, A Division of the J. Craig Venter Institute, 9712 Medical Center Drive, Rockville, Maryland 20850, USA

[3] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK

[4] Genoscope/Centre National de Séquençage , 2, rue Gaston Crémieux, CP 5706, 91057 Evry Cedex, France

[5] Department of Disease and Stress Biology, John Innes Centre, Norwich NR4 7UH, UK

[6] Department of Plant Pathology, University of Minnesota, 495 Borlaug Hall, St. Paul, MN 55108, USA

[7] Laboratoire des Interactions Plantes-Microorganismes, INRA UMR441, CNRS UMR2594, chemin de Borde Rouge, Auzeville, F-31320 Castanet-Tolosan, France

[8] Munich Information Center for Protein Sequences/Institute for Bioinformatics, GSF Research Center for Environment and Health, Ingolstädter Landstr.1, Neuherberg, Germany

[9] Department of Plant Systems Biology, VIB, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[10] Laboratory of Molecular Biology, Department of Plant Science, Wageningen University, Dreijenlaan 3, 6703HA Wageningen, Netherlands

[11] Department of Plant Pathology, University of California, One Shields Ave, Davis, CA 95616, USA

[12] USDA-ARS Corn Insects and Crop Genetics Research Unit and Department of Agronomy, Iowa State University, Ames, IA, 50011, USA

[13] Virtual Reality Applications Center, Iowa State University, Ames, IA, 50011, USA

[14] Max Planck Institute for Plant Breeding Research, Plant Computational Biology, Carl von Linné Weg 10, 50829 Köln, Germany

[15] Laboratoire de Biométrie et Intelligence Artificielle, INRA UR875, chemin de Borde Rouge, Auzeville, F-31320 Castanet-Tolosan, France

[16] Agricultural Biotechnolgy Center, Gödöllő, Hungary, 2. Biological Research Center, Gödöllő, Hungary

[17] National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505 USA

[18] Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Pky. Ardmore, OK 73401, USA