Medicago sativa cv. Mercedes genome sequence

Sébastien Carrere, Jerome Gouzy, Frédéric Debellé, William Marande, Bernadette Julier, Philippe Barre

04 November 2020

Medicago sativa cv. Mercedes genome sequence

Sébastien Carrère[1], Jérôme Gouzy[1], Frédéric Debellé[1], William Marande[3], Bernadette Julier[2] & Philippe Barre[2]

Abstract

An alfalfa (or lucerne) genome reference sequence is an essential tool for breeding of this major legume species. A clone of Flemish origin has been sequenced and the genome assembly has been carried out with NRGene protocols. A total of almost 190 000 scaffolds have been generated and this genome assembly reaches 2.6 Gb (80% of the 3.2 Gb expected). Genome annotation has provided 233 049 protein-coding genes and 36 752 non-protein coding genes. A genome portal based on JBrowse has been developed for searching the annotated genome (https://medicago.toulouse.inra.fr/MsatMercedes-NRGENE-20181029/).

Introduction

Alfalfa or lucerne (*Medicago sativa*) is an autotetraploid species (2n = 4x = 32) whose total genome size approximates 3.2 Gb. The allogamy of this species generates a high heterozygosity that complexifies genome assembly. The genome of *Medicago truncatula*, a wild diploid related model species, has long been used as a reference sequence for alfalfa genomic studies. Indeed, the two species are highly syntenic (Choi et al 2003; Julier 2003). For example, GBS reads have been mapped on *M. truncatula* to identify polymorphic markers (Li et al 2014; Annicchiarico et al 2016; Julier 2018). Thirty percent of reads are being mapped only, probably because the non-coding portions are poorly conserved, so the marker density is not as high as expected.

A reference alfalfa genome sequence is needed to progress in alfalfa genetics and genomics. In EUCLEG project (www.eucleg.eu) funded by European Union, a genome sequence has been obtained and is made available on this website: (https://medicago.toulouse.inra.fr/MsatMercedes-NRGENE-20181029/).

[1] Université de Toulouse, INRAE, CNRS, LIPM, Castanet-Tolosan, France
[2] INRAE, P3F, Lusignan, France
[3] INRAE-CNRGV, Castanet-Tolosan, France
Correspondance: sebastien.carrere@inrae.fr
DOI : 10.25794/ksvh-td05, https://doi.org/10.25794/KSVH-TD05

Material and method

The alfalfa clone Mercedes2.11 that has been used as a parent of a mapping population (Julier et al. 2003), has been chosen. This clone originates from the French variety Mercedes, a representative of the Flemish (Flamande) germplasm from North of France.

The sequencing has been conducted by the private company NRGene. NRGene has set up a whole genome sequencing and assembly service that had given excellent results on several species with a complex genome (wheat, maize, Italian ryegrass, strawberry, ...).The service is based on a classical genome sequencing (that provided by Illumina) with special conditions that enhance the quality of the sequences. More importantly, NRGene has developed a genome assembly software (DeNovoMAGICTM-3.0) that enables the genome assembly of the most complex genomes (including heterozygous and polyploid ones as alfalfa is). High-quality DNA of Mercedes2.11 was provided to NRGene.

Description of the assembly

The second generation sequencing offered a mean depth coverage of 131, considering a monoploid genome of x = 0.8 Gb (2n = 4x). In addition, third generation sequencing was obtained with a coverage of 17. After assembling, the total number of scaffolds was 189 167, obtained from 296 844 contigs (Table 1). This genome assembly is far from the 32 expected pseudomolecules, probably because of high heterozygosity that hampers genome assembly. However, this genome assembly reaches 2.6 Gb (3.2 Gb expected), meaning that 80% of the genome has been obtained.

Table 1. Scaffold statistics summary

	# of scaffolds	Min length (nt)	Max length (nt)	N50 (nt): 50% of the assembly is contained in scaffolds larger than this value	# of scaffolds larger than N50	N90 (nt): 90% of the assembly is contained in scaffolds larger than this value	# of scaffolds larger than N90	Mean scaffold length	Median scaffold length	Assembly size (nt)	# of Ns
All scaffolds	189 167	1 000	15 212 363	1 289 735	547	3 578	28 273	13 976	1 577	2 643 922 996	45 316 006
Scaffolds ≥ 5000nt	18 154	5 000	15 212 363	1 500 367	436	157 915	1 895	128 742	8 656	2 337 185 696	44 954 302

Table 2. Busco statistics (Busco3: release 3.0.2, embryophyta_odb9 dataset)

Busco type	Number	Percent
Complete BUSCOs (C)	1380	95.8%
Complete and single-copy BUSCOs (S)	154	10.7%
Complete and duplicated BUSCOs (D)	1226	85.1%
Fragmented BUSCOs (F)	16	1.1%
Missing BUSCOs (M)	44	3.1%
Total BUSCO groups searched	1440	

To further estimate the completeness of the genome, we used two independent complementary approaches. Mapping of full-length mRNA transcripts obtained by PacBio single-molecule long-read sequencing technology (IsoSeq transcripts, Chao et al., 2019) indicated that more than 98% of these transcripts can be mapped on the assembly at ≥95% identity. In addition, BUSCO (Benchmarking Universal Single Copy Orthologs, Simão et al., 2015) analysis of the genome assembly indicated that 93 % of the genes in a BUSCO set of single copy genes were found in the assembly (Table 2). Most of these genes are present in 1 to 4 copies (Table 3). This is in accordance to the autotetraploidy of alfalfa. It also confirms that the assembling procedure was able to recover the 4 haplotypes when they were different.

Table 3. Estimation of the number of gene copies

Number of gene copies	Number of genes identified
1	212
2	284
3	393
4	442
5	5
6	2
>6	4

The structural and functional annotation process of the *M.sativa* Mercedes genome has been based on an automatic pipeline, the integrative gene finder EuGene that integrates more than 200 RNA-seq datasets (Sallet et al. 2019). As a total, we found 233 049 protein-coding genes and 36 752 non-protein coding genes.

Web portal and genome browser

This genome portal has been developed to browse and search genome annotation. It is built upon JBrowse genome browser and ElasticSearch search engine.

We annotated protein coding genes using various methods:

- Blastp *vs. A. thaliana*, *M. truncatula*, *P. vulgaris, L. japonica*, *G. max* proteomes
- InterProsScan
- Blast2Go.

All these annotation results have been merged into annotation sheets that are searchable using keywords or accession numbers through the genome portal.

The portal link:

https://medicago.toulouse.inra.fr/MsatMercedes-NRGENE-20181029/

Acknowledgements

This alfalfa genome assembly in part of EUCLEG project (www.eucleg.eu). This project has received funding from the European Union's Horizon 2020 Programme for Research & Innovation under grant agreement n°727312.

References

Annicchiarico P, Nazzicari N, Ananta A, Carelli M, Brummer EC (2016) Assessment of cultivar distinctness in alfalfa: a comparison of genotyping-by-sequencing, simple-sequence repeat marker, and morphophysiological observations. Plant Genome 9:1–12

Chao, Y., Yuan, J., Guo, T. et al. (2019) Analysis of transcripts and splice isoforms in Medicago sativa L. by single-molecule long-read sequencing. Plant Mol Biol 99, 219–235 https://doi.org/10.1007/s11103-018-0813-y

Haitao Chen, Yan Zeng, Yongzhi Yang, Lingli Huang, Bolin Tang, He Zhang, Fei Hao, Wei Liu, Youhan Li, Yanbin Liu, Xiaoshuang Zhang, Ru Zhang, Yesheng Zhang, Yongxin Li, Kun Wang, Hua He, Zhongkai Wang, Guangyi Fan, Hui Yang, Aike Bao, Zhanhuan Shang, Jianghua Chen, Wen Wang, Qiang Qiu (2020) Allele-aware chromosome-level genome assembly and efficient transgene-free genome editing for the autotetraploid cultivated alfalfa. Nature communications 11:2494

Julier B, Flajoulot S, Barre P, Cardinet G, Santoni S, Huguet T, Huyghe C (2003) Construction of two genetic linkage maps in cultivated tetraploid alfalfa (*Medicago sativa*) using microsatellite and AFLP markers. BMC Plant Biol 3:9

Julier B, Lambroni P, Delaunay S et al (2018) Use of GBS markers to distinguish among lucerne varieties, with comparison to morphological traits. Mol Breed 38:133

Li XH, Wei YL, Acharya A, Jiang QZ, Kang JM, Brummer EC (2014) A saturated genetic linkage map of autotetraploid alfalfa (*Medicago sativa* L.) developed using genotyping-bysequencing is highly syntenous with the *Medicago truncatula* genome. G3-Genes Genomes Genetique 4: 1971–1979

Sallet, Erika ; Gouzy, Jerome ; Schiex, Thomas (2019) EuGene: An Automated Integrative Gene Finder for Eukaryotes and Prokaryotes. Methods in molecular biology.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212. doi:10.1093/bioinformatics/btv351