Towards a high-throughput method to measure genetic diversity in grassland

Author(s):
Loera-Sánchez, Miguel A.; Studer, Bruno; Kölliker, Roland

Publication date:
2019-06

Permanent link:
https://doi.org/10.3929/ethz-b-000353858

Originally published in:
Grassland Science in Europe 24

Towards a high-throughput method to measure genetic diversity in grassland

Loera-Sánchez M.A., Studer B. and Kölliker R.
Molecular Plant Breeding, Institute of Agricultural Sciences, ETH Zurich, Universitaetstrasse 2, 8092 Zurich, Switzerland

Abstract

Despite its economic and ecological relevance, the genetic diversity of grasses and legumes is rarely assessed in large-scale biodiversity surveys. This is due to the fact that morphology-based methods and low-throughput molecular markers are not well-suited for large-scale assessments. In addition, most grass and legume species lack the genomic information needed to develop high-throughput DNA-based methods. We hypothesize that a set of semi-conserved nuclear loci will provide enough sequence information to efficiently assess species richness and genetic diversity in mixed-species samples of grasses and legumes. We followed a targeted sequencing approach to enrich 611 nuclear loci from multiple genotypes of 16 economically relevant forage species. The target loci showed increasing within-species diversity as more genotypes were analysed, suggesting that they can be used in an amplicon-based method to measure genetic diversity on a large scale. Furthermore, some loci were also able to discriminate between species, which is a key feature for applications in mixed-species samples.

Keywords: probe capture, amplicon sequencing, grassland species, genetic diversity, high-throughput genotyping, species mixtures

Introduction

The genetic diversity within each plant species of a community is a valuable asset for grasslands. High levels of plant genetic diversity may stabilize and increase plant population yields (Abbott *et al.,* 2017). Furthermore, the genetic diversity of grasses and legumes, the most economically relevant plant families in grasslands, provides the basis for breeding superior forage crop cultivars.

Utilizing and protecting genetic diversity in grasslands requires detailed measurements on a large scale. However, biodiversity assessments of plants have so far focused largely on species richness, due mainly to technical limitations for measuring genetic diversity (Taberlet *et al.,* 2012). Developments in the area of high-throughput DNA sequencing and genotyping offer new possibilities to assess genetic diversity and species richness on a larger scale.

We hypothesize that a set of nuclear loci will allow us to measure genetic diversity in grass and legume populations and, at the same time, discriminate species in mixed samples. In this work, we assessed the within-species genetic diversity and the between-species discrimination potential of 611 nuclear loci. Those 611 loci are shared by a phylogenetically diverse set of plants, which comprises two grasses (*Lolium perenne* and *Brachypodium distachyon*), two legumes (*Trifolium pratense* and *Glycine max*), as well as *Arabidopsis thaliana*, *Theobroma cacao*, *Solanum lycopersicum*, and *Vitis vinifera* but occur only once in each genome. Our aim is to find the 10 to 20 most sequence-diverse loci, which we will use to produce an amplicon-based method to detect genetic diversity and species richness in grasslands on a large scale.

Materials and methods

Genomic DNA was extracted from plant individuals of 16 forage species (*Arrhenatherum elatius, Alopecurus pratensis, Cynosurus cristatus, Dactylis glomerata, Festuca pratensis, F. rubra, Lolium multiflorum, L. perenne, Phleum pratense, Poa pratensis, Trisetum flavescens, Lotus corniculatus, Medicago*

sativa, Onobrychis viciifolia, Trifolium pratense, and *T. repens*). Each species was represented by 5 plants from three cultivars. 80 dual-indexed Illumina libraries were prepared for each plant, and 16 additional libraries were prepared with the pooled DNA of all plants from each species. The fragment size of the libraries was ~550 bp. Sequence capture was performed on the libraries with a custom MYbaits® v4 kit (Arbor Biosciences, MI, USA). The 100-nt long baits were designed to target 611 nuclear loci. The enriched libraries were sequenced (Illumina MiSeq) and separate *de novo* assemblies were made with one library of each species. Quality-controlled reads were mapped to their respective *de novo* assemblies. We used *k*-mer richness to measure sequence diversity at each locus in an increasing number of genotypes of the same species. *K*-mer calculations were performed with *k=25* and a coverage cut-off of 10. Clustering based on k-*mer* composition was done with *kWIP* (Murray *et al.*, 2017).

Results and discussion

A total of 13,827,916 high-quality paired reads were obtained, of which 2,199,853 reads (15.91%) were on-target (Table 1), a rather low proportion. However, the mean depth at every on-target nucleotide position for all libraries was 33.39X, which is enough to capture the sequence diversity of the different alleles of the target loci in diploid plants. In contrast, the mean off-target sequencing depth was 2.38X, which indicates that off-target reads come from random pieces of the genome, while the on-target reads are concentrated on the loci for which the sequence capture baits were designed. In addition, legume samples had more on-target reads than grass samples (1,583,144 reads vs 616,709), even though grasses represented the majority of all the sequenced libraries. Thus, the overall low read mapping rate could have been due to hybridization issues when pooling samples from different species on the same sequence capture reaction.

A direct relationship between *k*-mer richness at each locus and number of plants (i.e. distinct genotypes) was observed (Figure 1A). As more plants are considered for analysis, the *k*-mer richness at each locus increases, indicating that these loci are suitable to detect within species diversity. Clustering of the reads mapping to one of the 611 loci (uce-11004050; Figure 1B) allowed to separate all individual legume plants according to the respective species. For grasses, the separation was less clear, although some species were almost completely resolved. Inconsistencies in species clustering can be due to differences in *k*-mer sequencing depth, since grasses had lower on-target mean sequencing depth than legumes did.

Table 1. Sequence capture summary results.

Group of libraries	Raw reads	QC reads	On-target reads	On-target mean depth	Off-target mean depth
Grasses	10,373,596	9,821,623	616,709 (6.28%)	15.78X	2.15X
Legumes	4,170,860	4,006,293	1,583,144 (39.52%)	59.92X	3.48X
Total	14,544,456	13,827,916	2,199,853 (15.91%)	33.39X	2.38X

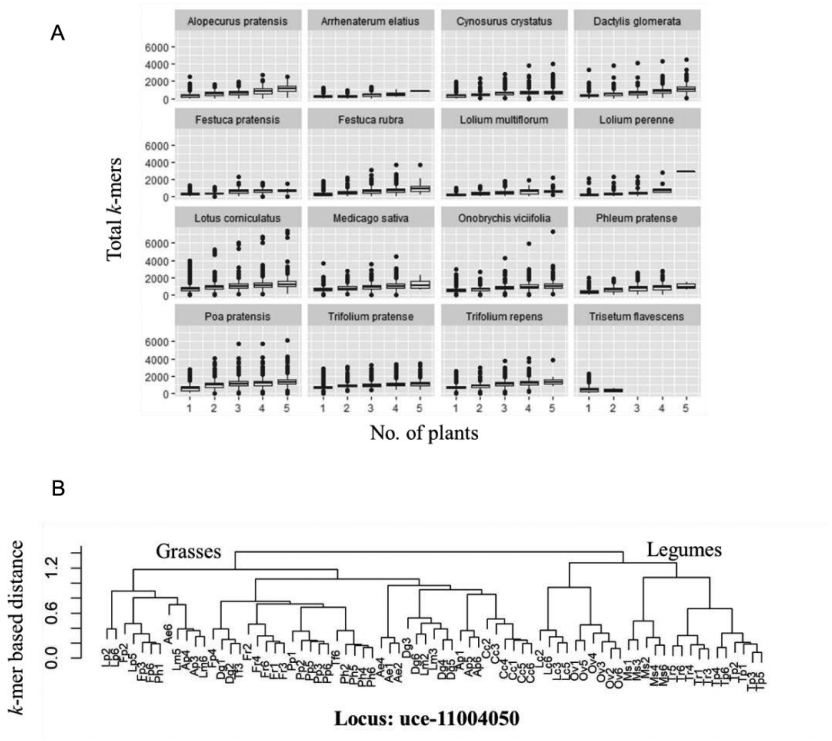

Figure 1. (A) The sequencing reads of one to five individual plants were categorized by locus and *k*-mer richness was calculated. Each data point corresponds to one locus. Each subplot corresponds to one species. Only *k*-mers with sequencing depth higher than 10× were considered. (B) Clustering of individual plants based on *k*-mer composition for locus uce-11004050.

Conclusion

Overall, the targeted 611 loci reflect the genetic diversity present in the different genotypes that were analysed for each species. Some loci, like the locus uce-11004050, are also able to distinguish species from each other. Our results constitute the first steps towards an amplicon-based method to measure genetic diversity in grassland plants. The loci displaying highest levels of between-species discrimination success and highest within-species diversity will be selected for further analysis.

References

Abbott J. M., Grosberg R.K., Williams S.L., and Stachowicz J.J. (2017) Multiple dimensions of intraspecific diversity affect biomass of eelgrass and its associated community. *Ecology* 98, 3152-64.

Murray K.D., Webers C., Ong C.S., Borevitz J., and Warthmann N. (2017) KWIP: The k-mer weighted Inner product, a *de novo* estimator of genetic similarity. *PLOS Computational Biology* 13, e1005727.

Taberlet P., Zimmermann N.E., Englisch T., Tribsch A., Holderegger R., Alvarez N., Niklfeld H., *et al.* (2012) Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. *Ecology Letters* 15, 1439-48.