



# GDA-NT 2021 – a computer program for population genetic data analysis and assignment

Bernd Degen<sup>1,2</sup>

Received: 27 July 2021 / Accepted: 30 June 2022 / Published online: 14 July 2022  
© The Author(s) 2022

## Abstract

Data on genetic diversity and differentiation, as well as kinship between individuals, are important for the conservation of animal and plant genetic resources. Often genetic assignment is part of law enforcement of protected endangered species. The software GDA-NT 2021 is a new, freely available user-friendly Windows program that can be used to compute various measures of genetic diversity and population genetic differentiation. It further allows genetic assignment of individuals to populations and enables the calculation of kinship-coefficients and genetic distances among pairs of individuals within populations. GDA-NT 2021 specifically computes the alternative measures for population differentiation  $D_j$  and the standardized  $F_{ST}$  of Hedrick. It has more options to compute exclusion-probabilities in assignment tests, enables self-assignment tests for variable groups of individuals, and allows for information on geographic positions to be accounted for while using permutation tests to assess statistical significance.

**Keywords** *Computer program · Genetic assignment · Genetic differentiation · Genetic diversity · SNP · Statistical tests*

## Application

ORCID: 0000-0001-9082-3163.

Running title: GDA-NT 2021 computer program.

Word count: 1291.

### Declarations.

*Availability of software, data and material.*

The program, a zip-file with different input data files for demonstration, different videos that explain the use of the program and the User's manual are available on our website:

<https://www.thuenen.de/en/institutes/forest-genetics/software/gda-nt-2021>.

The conservation of plant and animal genetic resources relies on data on genetic diversity and genetic differentiation (Rodriguez-Quilon et al. 2016; Eusebi et al. 2020; Boccacci

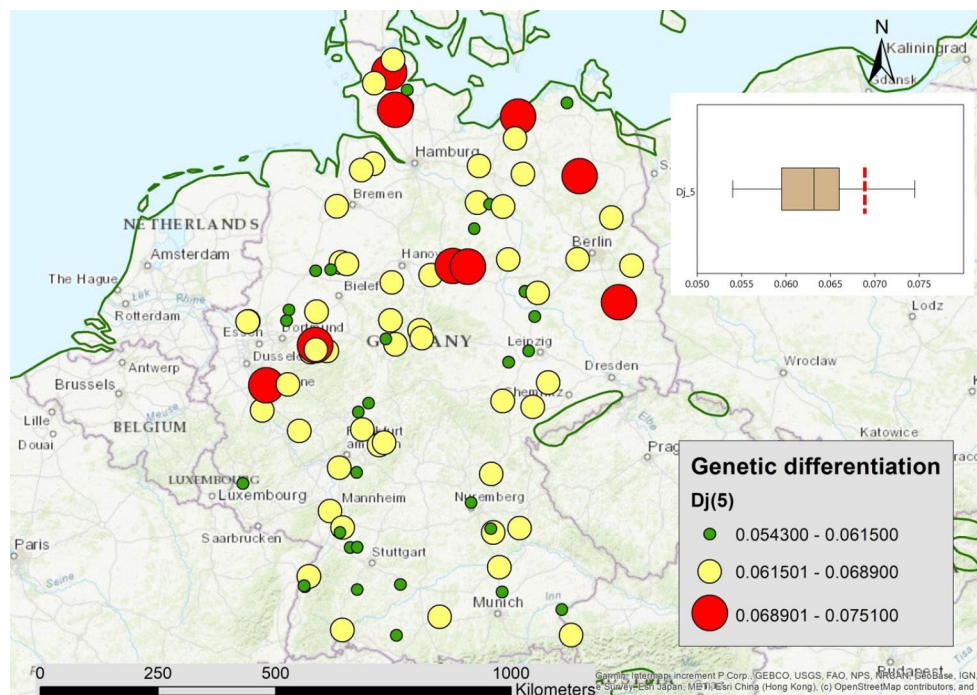
et al. 2021). Decisions on the selection of conservation units are often taken based on genetic inventories with genetic markers (Dudgeon et al. 2012; Boccacci et al. 2021). Further, the level of genetic relatedness of individuals is an important criterion for the evaluation of genetic resources (Welirmann and Bennewitz 2019). Relatedness can be estimated with the kinship-coefficient based on data of genetic markers (Han et al. 2020). In addition, genetic assignment is often used for conservation purposes, especially for law enforcement to protect species by checking the geographic origin or taxonomy of traded biological material such as seeds, timber, ivory or bush meat (Wasser et al. 2007; Degen et al. 2013).

GDA-NT stands for “Genetic data analysis and numerical tests”. The software computes various metrics of population fixation and differentiation using genetic data that are similar to other programmes, such as the Wright's  $F_{IS}$  and  $F_{ST}$  indices computed by Alequin 3.5 (Excoffier and Lischer 2010). Or it provides different genetic assignment criteria to assign or exclude reference populations as implemented in GeneClass (Piry et al. 2004). It also computes

✉ Bernd Degen  
bernd.degen@thuenen.de

<sup>1</sup> Thuenen Institute of Forest Genetics, Sieker Landstrasse 2, D-22927 Grosshansdorf, Germany

<sup>2</sup> Thuenen Institute of Forest Genetics, Sieker Landstrasse 2, D-22927 Grosshansdorf, Germany



**Fig. 1** Map visualising the genetic differentiation  $D_j(5)$  of 94 pedunculate oak locations in Germany screened at 356 nuclear SNPs. The map has been created with ArcMap 10.8 (ESRI). Each circle represents a location (stand) at which ten oak trees have been randomly sampled from the local population. This data is part of a range-wide study of pedunculate oaks in Europe (Degen et al. 2021) and has been analysed with GDA-NT 2021.  $D_j(5)$  is one of the spatially explicit measures. It computes the genetic distance of a given population to the five closest neighbour populations. The number of five neighbours has been selected according to the spatial distribution of sampled locations and represents in most cases a group of oak stands within a radius of less than 300 km. The distribution of the  $D_j(5)$ -values of the 94 locations is shown as a boxplot. The red marked stands are the 10% most differentiated ones. These extreme  $D_j(5)$ -values served as one indicator among others to identify suspicious, potentially non-local seed sources. Usually, oak populations that are not autochthones get excluded from conservation programs

allele and genotype frequencies on different aggregation levels as integrated in the R-package *pegas* (Paradis 2010). However, GDA-NT 2021 has more options to compute exclusion-probabilities in assignment tests, and does enable self-assignment tests for groups of individuals with variable group sizes. In addition, it allows the calculation of alternative measures of population differentiation, such as the standardized  $F_{ST}$  (Hedrick 2005; Meirmans and Hedrick 2011) or  $D_j$  (Gregorius and Roberds 1986; Gregorius et al. 2007), which can also integrate geographic location information to select sub-populations. The application of these alternative measures of genetic differentiation is particularly useful to address questions of conservation genetics (Prunier et al. 2017; Attu et al. 2022; Nguyen et al. 2022). Figure 1 shows the application of  $D_j$  as an indicator to identify pedunculate oak populations in Germany that are likely of foreign origin and thus should be excluded from a conservation program.

ASCII text files with diploid genetic markers (e.g., nSNPs, nSSRs) or haploid genetic markers (e.g., cpSNPs, cpSSRs) are used as input files for GDA-NT. Alternative CSV files generated by other programs such as EXCEL and R can be imported, transformed and saved as input files. The results are automatically stored as text-files and optionally

as csv-files for further data visualisation and downstream data analyses such as a detailed analysis of spatial genetic structures, e.g., using the software SGS (Degen et al. 2001) or principal component analysis (PCA) and cluster analysis based on allele frequencies with the program PAST (Hammer et al. 2001). An overview of the program features is given in Table 1. The program can handle data of up to a few hundred populations and a few hundred genetic markers (see as an example Degen et al. (2021)). GDA-NT 2021 is well suited for conservation genetics studies, where typical datasets involve the screening of many populations with a specifically selected subset of informative genetic markers. It is not developed for applications with large SNP arrays comprising thousands or millions of SNPs, but it can be used for a genetic quality check of pruned SNP sets drawn from such large SNP arrays.

GDA-NT 2021 has been programmed in visual basic and compiled for the operating system Microsoft Windows (Windows 10 and earlier versions). The program, a zip-file with different input data files for demonstration, different videos that explain the use of the program and the user's manual are available on our website:

**Table 1** Features, methods and measures implemented in the program GDA-NT 2021

Feature	Method / Measure	Reference	
<b>Genetic assignment</b>	Bayesian approach	Rannala and Mountain (1997)	
	Allele / haplotype frequencies approach	Paetkau et al. (1995)	
	Genetic distance	Gregorius (1974)	
	Exclusion probabilities	Cornuet et al. (1999)	
<b>Analysis of population data</b>	<i>Genetic composition</i>	Frequencies of alleles	
		Frequencies of single locus genotypes	
		Frequencies of multilocus genotypes / haplotypes	
	<i>Genetic variation</i>	Number of alleles ( <i>A</i> )	
		Number of single locus genotypes ( <i>NG</i> )	
		Number of multilocus genotypes ( <i>NMG</i> )	
		Allelic richness ( <i>Ar</i> )	El Mousadik and Petit (1996)
		Diversity ( <i>V</i> ) = effective number of alleles	Gregorius (1987)
		Evenness of allele frequencies ( <i>E</i> )	Gregorius (1990)
		Observed heterozygosity ( <i>Ho</i> )	Wright (1978)
	Expected heterozygosity ( <i>He</i> )		
	Fixation index ( <i>F<sub>IS</sub></i> )		
	Degree of heterozygosity		
<i>Genetic differentiation</i>	Genetic distance ( <i>DN</i> )	Nei (1972)	
	Genetic distance ( <i>GD</i> )	Gregorius (1974)	
	Differentiation of populations ( <i>Dj</i> )	Gregorius and Roberds (1986)	
	Population fixation ( <i>F<sub>ST</sub></i> )	Wright (1965)	
	Standardized population fixation ( <i>F<sub>ST-H</sub></i> )	Hedrick (2005)	
	<i>Analysis of pairs of individuals</i>	Kinship ( <i>Kin</i> )	Loiselle et al. (1995)
Multilocus genetic distance ( <i>DGM</i> )		Gregorius (1984)	

<https://www.thuenen.de/en/institutes/forest-genetics/software/gda-nt-2021>.

**Acknowledgements** I would like to thank Céline Blanc-Jolivet and Malte Mader for critical testing of the program and for helpful suggestions for its improvement. I am grateful to Hans-Rolf Gregorius and Elizabeth Gillet for detailed discussions on genetic measures and to two anonymous reviewers as well as to Dina Führmann, Katharina Liepe and Niels Müller for helpful comments on a former version of the manuscript.

**Funding** The author did not receive support from any organization for the submitted work.

**Conflicts of interest/Competing interests** The author has no relevant financial or non-financial interests to disclose. Open Access funding enabled and organized by Projekt DEAL.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Attu GAT, Frimpong EA, Hallerman EM (2022) Defining Management Units for Wild Nile Tilapia *Oreochromis niloticus* from Nine River Basins in Ghana. *Diversity-Basel*, p 14
- Boccacci P, Aramini M, Ordidge M, van Hintum TJJ, Marinoni DT, Valentini N, Sarraquigne JP, Solar A, Rovira M, Bacchetta L, Botta R (2021) Comparison of selection methods for the establishment of a core collection using SSR markers for hazelnut (*Corylus avellana* L.) accessions from European germplasm repositories. *Tree Genet Genomes* 17:14
- Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* 153:1989–2000
- Degen B, Petit RJ, Kremer A (2001) SGS - Spatial genetic software: A computer program for analysis of spatial genetic and phenotypic structures of individuals and populations. *J Hered* 92:447–449
- Degen B, Ward SE, Lemes MR, Navarro C, Cavers S, Sebbenn AM (2013) Verifying the geographic origin of mahogany (*Swietenia macrophylla* King) with DNA-fingerprints. *Forensic Sci International-Genetics* 7:55–62
- Degen B, Yanbaev Y, Mader M, Ianbaev R, Bakhtina S, Schroeder H, Blanc-Jolivet C (2021) Impact of gene flow and introgression on the range wide genetic structure of *Quercus robur* (L.) in Europe. *Forests* 12:1425
- Dudgeon CL, Blower DC, Broderick D, Giles JL, Holmes BJ, Kashiwagi T, Krueck NC, Morgan JAT, Tillett BJ, Ovenden JR (2012) A review of the application of molecular genetics for fisheries management and conservation of sharks and rays. *J Fish Biol* 80:1789–1843
- El Mousadik A, Petit R (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. *Theor Appl Genet* 92:832–839
- Eusebi PG, Martinez A, Cortes O (2020) Genomic Tools for Effective Conservation of Livestock Breed Diversity. *Diversity-Basel* 12:16
- Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10:564–567

- Gregorius H (1974) Genetic distance between populations. The concept of measurement of genetic distance. *Silvae Genetica* 23:22–27
- Gregorius HR (1984) A unique genetic distance. *Biom J* 26:13–18
- Gregorius H-R, Roberds J (1986) Measurement of genetic differentiation among subpopulations. *Theor Appl Genet* 71:826–834
- Gregorius H-R (1987) The relationship between the concepts of genetic diversity and differentiation. *Theor Appl Genet* 74:397–401
- Gregorius H-R (1990) A diversity-independent measure of evenness. *Am Nat* 136:701–711
- Gregorius HR, Degen B, König A (2007) Problems in the analysis of genetic differentiation among populations - A case study in *Quercus robur*. *Silvae Genetica* 56:190–199
- Hammer Ø, Harper DA, Ryan P (2001) Past: Paleontological statistics software package for education and data analysis. *Paleontologia Electronica* 4:1–9
- Han LD, Love K, Peace B, Broadhurst L, England N, Li L, Bush D (2020) Origin of planted *Eucalyptus benthamii* trees in Camden NSW: checking the effectiveness of circa situm conservation measures using molecular markers. *Biodivers Conserv* 29:1301–1322
- Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution* 59:1633–1638
- Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic-structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Bot* 82:1420–1425
- Meirmans PG, Hedrick PW (2011) Assessing population structure: F-ST and related measures. *Mol Ecol Resour* 11:5–18
- Nei M (1972) Genetic distance between populations. *Am Nat* 106:283–292
- Nguyen DM, Nguyen HLP, Nguyen TM (2022) Genetic structure of the endemic *Dipterocarpus condorensis* revealed by microsatellite markers. *Aob Plants*, p 14
- Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Mol Ecol* 4:347–354
- Paradis E (2010) pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics* 26:419–420
- Piry S, Alapetite A, Cornuet JM, Paetkau D, Baudouin L, Estoup A (2004) GENECLASS2: A software for genetic assignment and first-generation migrant detection. *J Hered* 95:536–539
- Prunier R, Akman M, Kremer CT, Aitken N, Chuah A, Borevitz J, Holsinger KE (2017) Isolation by distance and isolation by environment contribute to population differentiation in *Protea repens* (Proteaceae L.), a widespread South African species. *Am J Bot* 104:674–684
- Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences*, 94, 9197–9201
- Rodriguez-Quilon I, Santos-del-Blanco L, Serra-Varela MJ, Koskela J, Gonzalez-Martinez SC, Alia R (2016) Capturing neutral and adaptive genetic diversity for conservation in a highly structured tree species. *Ecol Appl* 26:2254–2266
- Wasser SK, Mailand C, Booth R, Mutayoba B, Kisamo E, Clark B, Stephens M (2007) Using DNA to track the origin of the largest ivory seizure since the 1989 trade ban. *Proc Natl Acad Sci U S A* 104:4228–4233
- Welirinnann R, Bennewitz J (2019) Key genetic parameters for population management. *Front Genet* 10:20
- Wright S (1965) The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, pp 395–420
- Wright S (1978) Variability within and among natural populations. University of Chicago Press, Chicago

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.