Genome-wide characterization of genetic diversity and population structure of *Valeriana officinalis* L.

Maja Boczkowska (✉ m.boczkowska@ihar.edu.pl)

Plant Breeding and Acclimatization Institute <https://orcid.org/0000-0001-8691-410X>

Katarzyna Bączek

Warsaw University of Life Sciences - SGGW

Olga Kosakowska

Warsaw University of Life Sciences

Anna Rucińska

Polish Academy of Sciences Botanical Garden - Center for Biological Diversity Conservation

Wiesław Podyma

Plant Breeding and Acclimatization Institute - NRI

Zenon Węglarz

Warsaw University of Life Sciences

Research article

Keywords: *Valeriana officinalis*, population genomics, population structure, conservation genomics, diversity, DArTseq,

Posted Date: June 6th, 2019

DOI: <https://doi.org/10.21203/rs.2.10069/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: *Valeriana officinalis* L. is one of the most important medicinal plant with a mild sedative, nervine, antispasmodic and relaxant effect. Despite a substantial number of studies on this species, population genomics has not yet been analyzed. The main aim of this study was: characterization of genetic variation of natural populations of *V. officinalis* in Poland and comparison of variation of wild populations and the cultivated form using Next Generation Sequencing based DArTseq technique. We also would like to establish foundations for genetic monitoring of the species in the future and to develop genetic fingerprint profile for samples deposited in gene bank and in natural sites in order to assess the degree of their genetic integrity and population structure preservation in the future.

Results: The major and also the most astounding result of our work is the low level of observed heterozygosity of individual plants from natural populations despite the fact that the species is widespread in the studied area. Inbreeding, in naturally outcrossing species such as valerian, decreases the reproductive success. The analysis of the population structure indicated the potential presence of metapopulation in a broad area of Poland and the formation of a distinct gene pool in Bieszczady Mountains. The results also indicate the presence of individuals of the cultivated form in natural populations in the region where the species is cultivated for the needs of the pharmaceutical industry and this could lead to structural and genetic imbalance in wild populations.

Conclusions: The DArTseq technology can be applied effectively in genetic studies of *V. officinalis*. The genetic variability of wild populations is in fact significantly lower than assumed. Individuals from the cultivated population are found in the natural environment and their impact on wild populations should be monitored.

Background

Common valerian (*Valeriana officinalis* L.) has a long tradition of use as a medicinal plant [1]. *Valerianae radix* (commonly known as valerian root) is one of the best-tested plant materials in terms of pharmacology and clinical aspects. The raw material contains compounds from several chemical groups, including essential oil (according to European Pharmacopoeia not less than 0.5%), valepotriates, and sesquiterpenic acids (according to European Pharmacopoeia not less than 0.17%) with valerenic and acetoxvalerenic acids considered as species-specific marker compounds. Valerian root and the extracts thereof are used in the states of anxiety or sleeping disorders. Their efficacy relates to an interplay between different groups of compounds rather than with individual compounds [2].

Plant description

V. officinalis belongs to the genus *Valeriana* L. that is a part of the family *Valerianaceae*, now considered also as a part of the *Caprifoliaceae* Juss. in the order *Dipsacales* Juss. ex Bercht. & J.Presl. Approximately 200 species are listed within this cosmopolitan genus, several of which are of commercial interest. *V. officinalis* occurs natively throughout Eurasia, except for arctic and desert zones [3]. It

occupies fresh habitats such as pastures, meadows, ditches, scrub, open woodland, along margins of watercourses but also dry calcareous soils. Common valerian is a highly morphologically differentiated perennial with short, straight rhizome, roots and stolons as underground organs. Leaves are sessile, pinnately separated with 6-10 pairs of lance-shaped leaflets. Flowers are numerous, hermaphrodite, arranged in terminal cymose inflorescence with white or pink five-lobed corolla [4]. *V. officinalis sensu lato* constitutes a collective species with controversial subordinate taxonomy and phylogeny. Besides a high degree of developmental plasticity, it is also characterized by a varying ploidy level ($2n=14$, $2n=28$, $2n=42$ or $2n=56$) [5-7].

Problem

Wild growing medicinal and aromatic plants (MAPs) are widely available source for discovering new drugs. Herbal drug industry has been expanding very intensively in recent years. Currently, in developing countries simple herbal products are used as a primary source of medicines, while in western countries they are used as an alternative or complementary to conventional medicine.

However, according to World Health Organization (WHO) data, still about 80% of people depend on pharmaceuticals of plant origin [8]. Out of about 422 000 known plant species, about 12.5%, have been used for the medical purposes. It is estimated that about four thousands of herbal plants species are threatened with extinction [9]. It is caused by both overexploitation of raw materials from natural sites and reduction of natural habitats due to uncontrolled deforestation as well as rapid development of agriculture and urban areas [10]. Climate change also threatens medicinal plants, which, like all other living members of the biosphere, are not immune to them. Particularly endemic species are vulnerable to changes caused by global warming in the geographical regions or ecosystems which they inhabit [11]. The conservation of medicinal plant resources should take place in several directions by *in situ* and *ex situ* conservation strategies, good agricultural practices and sustainable use solutions [12]. Awareness of genetic diversity levels significantly improves the efficiency of genetic resources conservation, sustainable use and accelerates breeding activities that lead to the production of forms suitable for cultivation. Genomic studies of MAPs remain behind model and important cultivated plant species.

Genotyping

There is a lack of complex molecular genetic studies of *V. officinalis* as well as for the majority of medicinal plants. Next Generation Sequencing (NGS) technology could provide efficient toolkit for genotyping studies of medicinal plants. NGS, that is based on massive parallel sequencing, has revolutionized plant science over the last years. This ultra-high-throughput, scalable and high-speed technology can be applied to organisms having an unknown genome sequence. Large amount of Single Nucleotide Polymorphism (SNP) markers can be obtained through the simplified genome sequencing like restriction site associated DNA sequencing (RADSeq), genotyping-by-sequencing (GBS) or other related techniques [13-15]. DArTseq (Diversity Array Technology sequencing) is one such method. DArT was first developed at the beginning of 2000's and has improved both throughput and accuracy in the detection of DNA polymorphisms without the need for prior information on DNA sequences through hybridization and

solid-state surfaces [16]. Today's DArTseq technique is the result of a combination of DArT and NGS [17]. The use of combinations of endonucleases that target low-copy DNA sites rather than repetitive DNA fragments leads to an effective reduction in genome complexity. The resultant representation of the genome is composed of both constant and polymorphic fragments across individuals. Until now, DArTseq has been applied to many plant species, especially cultivated plants, due to its high throughput, genomic coverage and transferability. Relatively few papers on its use for genotyping of wild and medicinal plant species have been published. It has been successfully used in *Eucalyptus* L'Hér [18], *Cochlearia polonica* A.Fröhl. and *Ligularia sibirica* (L.) Cass. (Rucińska unpublished data) research

Aim of the study

The aims of this study are *i*) evaluation of the usefulness of the DArTseq technique for genotyping of *V. officinalis*; *ii*) characterization of genetic variation of natural populations of *V. officinalis* in Poland; *iii*) comparison of variation of wild populations and the cultivated form; *iv*) the establishment of foundations for genetic monitoring of the species in the future; *v*) the development of genetic fingerprint profile for samples deposited in gene banks and in natural sites in order to assess the degree of their genetic integrity and population structure preservation in the future.

Methods

Plant material

The objects of the study were 19 wild growing common valerian populations (accessions) originated from natural sites in Poland (Mazowieckie, Świętokrzyskie, Kujawsko-Pomorskie and Podkarpackie Voivodeships) (Fig. 1). The formal identification of the plant material was done by Katarzyna Bączek i.e. gene bank curator of the medicinal and aromatic plants collection. The species is not under legal protection and therefore no formal consent for seed collection was required. Seeds were collected from each population separately (min. 30 individual plants representing the whole range of the population). The seeds from plants within single population were mixed to obtain a representative sample. Reference material was a cultivated form of common valerian 'Lubelski' originated from Martin Bauer company. Voucher specimens were deposited at herbarium of Department of Vegetables and Medicinal Plants, WULS-SGGW. Populations' seeds are kept in the National Centre for Plant Genetic Resources (Polish GeneBank) [45].

In Table 1, geographical coordinates (latitude and longitude) were given, followed by the determination of common valerian abundancy, according to Braun-Blanquet scale. In February 2016, seeds were sown in greenhouse, into multi-pots (with hole diameters 2-3 cm) filled with peat substrate, to obtain seedlings. Each population was represented by 25 plants. In the end of May 2016, the seedlings were planted in the experimental field to obtain a field *ex situ* collection.

DNA extraction and genetic analysis

Before planting the seedlings in the field (in May 2016) young, healthy leaves were taken from 10 plants from each population and from 20 plants from the cultivated form 'Lubelski'. The collected biological material was dried in presence of active silica gel. Until DNA isolation, the dried material was stored in -18°C in vacuum sealed packages. The tissue has been crushed into a fine powder in a bead mill (Mixer Mill MM 200, Retsch). Isolation of DNA was conducted using the Genomic Mini AX Plant kit (A&A Biotechnology, Poland) according to the manufacturer's protocol. DNA integrity was evaluated visually after electrophoresis in 2% agarose gel. Purity and concentration were evaluated spectrophotometrically (NanoDrop ND-1000). A set of 188 isolates with appropriate qualitative and quantitative parameters was sent for DArTseq analysis to Diversity Arrays Technology Laboratory, University of Canberra, Australia.

Data analysis

DArTseq markers analysis

The results of genotyping were generated as tables listing silicoDArT and SNP markers. SilicoDArTs markers (dominant) that are analogous to microarray DArTs are derived from the sequencing of fragments from the genomic representation and are associated with variability within the restriction sites. SNP markers (co-dominant) are single nucleotide polymorphisms detected in sequenced fragments of genome representations. Both markers were then scored as binary data, indicating presence (1) or absence (0) of a marker in the genomic representation of each sample, as described by Cruz et al. [17]. In the next step both types of markers were tested for reproducibility (%), call rate (%) and polymorphism information content (PIC).

Genetic diversity analyses

The following coefficients were calculated for SNP markers: observed heterozygosity (H_o), Nei unbiased genetic diversity (uH_e), and inbreeding coefficient (F_{IS}) according to the following equation:

Due to technical limitations, Equation 1 has been placed in the Supplementary Files section.

For SilicoDArT markers only Nei unbiased genetic diversity (uH_e) was calculated. Genetic distance, for both types of markers, was measured by using Jaccard coefficient [46] according to the formula:

Due to technical limitations, Equation 2 has been placed in the Supplementary Files section.

where a is the number of variables present in both individuals being compared, b is the number of variables present in the first individual and absent from the second individual and c is the number of variables present in the second individual and absent from the first one. The symmetric genetic distance

matrices were used to assess the amount of variation among the assigned regional groupings by Analysis of Molecular Variance (AMOVA) with tests including 999 permutations. In order to test for a correlation between two genetic distance matrices and between genetic and geographical distances (in km) among populations, Mantel tests were performed (computing 10000 permutations).

Population structure analysis

Genetic structure was assessed using Agglomerative Hierarchical Clustering (AHC) according to Ward's method, Principal Coordinate Analysis (PCoA) and model-based Bayesian clustering algorithms. Admixture and correlated allele frequencies model was used to determine the number of clusters (K) in the range from one to ten. Ten independent runs were set for each K value with a burn-in period of 10 000 and 50 000 Markov Chain Monte Carlo replications after burn-in with no prior information about the origin of individuals. The ΔK method was used to determine the most probable value of K.

Joint analysis

Generalized Procrustes Analysis (GPA) was used to obtain consensus configuration of SNP and SilicoDArT results and to reduce the scale effect.

Software

All above mentioned analyses were performed using the Microsoft Excel 2016/2018, R packages (vegan, dplyr, ape, scatterplot3d), XLSTAT Ecology (Addinsoft, Inc., Brooklyn, NY, USA), GenAlEx 6.501 [47], STRUCTURE v2.3.4 [48] and CLUMPAK [49]. The data analysis was performed within the framework of the Computational Grant (G72-19) of the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw (ICM UW).

Results

Marker quality analysis

SNP markers

A total of 65 510 SNP markers with an average reproducibility of 99.6% were obtained using the DArTseq method for 188 individuals of *V. officinalis*. Over 98% of SNPs had reproducibility above 95%, of which 55 101 was completely reproducible (Fig. 2a). The call rate was in the range from 20% to 100% (Fig. 2b). Approximately 63% of SNPs had a call rate above 75%. When the quality criteria were applied, i.e. reproducibility above 90% and call rate above 95%, 23 507 SNP markers were assigned to subsequent analysis. The polymorphism information content (PIC) was in the range of 0.01-0.5 with an average of 0.181 and a median of 0.129. As much as 40% of SNP markers had a PIC below 0.1, and only 1% had a maximum value of 0.5 (Fig.2c).

SilicoDArT markers

A total of 61 633 silicoDArT markers with an average reproducibility of 99.8% were obtained. All markers had the repeatability above 95% and as many as 58 6244 had the repeatability equal to 100% (Fig. 2d). The call rate ranged from 0.8 to 1 (Fig. 2e). About 80% of silicoDArT had a call rate above 90%. After applying quality criteria, 41 925 markers were approved for further analysis. The PIC value ranged from 0.01-0.5 to 0.170 and the median 0.123. The frequency analysis showed that over 40% of markers had a very low PIC, i.e. below 0.1, and only 0.02% had a maximum PIC value (0.5).

Genetic diversity

SNP markers

In individual populations, the percentage of polymorphic loci ranged between 31-61% (Fig. 3a). The occurrence of unique alleles populations ranged from 0 to 1059 (Fig. 3b). No unique alleles were found in two populations (PL 403186 and PL 403187) originating from the Świętokrzyskie Voivodeship. 22% of alleles that occurred in wild populations were not represented in the cultivated form. The contribution of SNP markers unique for the regions is shown in Fig. 3c. The observed heterozygosity (H_o) ranged from 0.196 (PL 401941) to 0.354 (PL 504920) (Fig. 3d). The mean heterozygosity of wild populations was 0.250. The expected heterozygosity (uH_e) varied between 0.274 (PL 403192) and 0.348 (PL 403183), with an average 0.315 for wild populations (Fig. 3d). The lowest value of inbreeding coefficient (F_{IS}) was found for cultivated form PL 504920 (-0.104) while the highest for population PL 401941 (0.354) (Fig. 3d). In general, in wild populations the coefficient reached values above zero, which indicates the presence of inbreeding. The population PL 401945 from the Świętokrzyskie Voivodeship was the closest to the state of Hardy-Weinberg equilibrium. Negative F-value for the cultivated form indicates excessive heterozygosity. The genetic distance according to Jaccard's coefficient between the examined individuals ranged from 0.08 (individuals from population PL 406738 and PL 403192) to 0.380 (individuals from population PL 401941 and PL 406738). The mean distance between individuals from natural populations was 0.239.

SilicoDArT markers

The percentage of polymorphic silicoDArT loci in the investigated samples was 17-49% (Fig.3a). Unique fragments were found in different numbers between 20 (PL 401938) and 2282 (PL 504920) (Fig. 2b). The cultivated form did not include 21 010 fragments, which were present in natural populations. The contribution of unique fragments in specific regions was presented in Fig. 3c. The Nei unbiased genetic diversity coefficient (uH_e) ranged between 0.328 and 0.409 for populations PL 403192 and PL 403183, respectively, and it showed a strong positive correlation (0.807; $p < 0.0001$) with uH_e calculated for SNP markers. The mean genetic distance between 188 individuals was 0.861. The maximum distance (0.941) was between individuals from population PL 401941 and PL 403192, while the minimum distance was 0.294 for individuals from population PL 406738 and PL 403192.

Mantel test

The comparison of genetic distance matrices obtained for SNPs and silicoDArTs showed a strong uphill (0.752; $p < 0.0001$) linear relationship of them (Fig. 4a). The test in which genetic distance matrices between populations were collated with a matrix of geographical distance showed a moderate positive relationship for both SNPs and silicoDArTs (0.304 and 0.329; $p < 0.0001$ respectively) (Fig. 4b, c).

AMOVA

The Analysis of Molecular Variance (AMOVA) for SNP markers of all samples showed that 23% of the variance in allele frequency could be attributed to differences between populations (Fig. 5). In case when only wild populations were analyzed, the value of differentiation among populations (Φ_{PT}) coefficient was 0.199 ($p < 0.001$). The hierarchical AMOVA of wild populations revealed that 78% of total variance was found among individuals, 16% was found among populations while the rest of variation (6%) was among regions. The genetic variation between the cultivated form and wild populations was 29%. AMOVA results of silicoDArTs indicated greater intrapopulation than interpopulation variation (~83% vs ~17%). The Φ_{PT} coefficient for the natural populations exclusively was 0.166 ($p < 0.001$). The hierarchical AMOVA of wild populations confirmed that the largest fraction of variation (82%) was within populations, while the interpopulation variance was 13% and inter-regions variation was restricted to 5%. 16% of the variation was found between the cultivated form and the natural populations.

Population structure

SNP markers

The Agglomerative Hierarchical Clustering based on SNP markers showed the presence of four main groups (Fig. 6a). The first one, most distant, was composed of 20 individuals. It consisted of the cultivated form individuals and four individuals from population PL 406738 and one of PL 403192. The second cluster consisted of 55 individuals, which came from six populations collected in the Podkarpackie Voivodeship. Nine specimens from Mazowieckie Voivodeship formed the third and the smallest group. The last group included as many as 104 individuals. Its core was constituted by populations located in the Świętokrzyskie Voivodeship. In addition, it contained the population from Kujawsko-Pomorskie and four individuals from the Mazowieckie Voivodeship.

The first three main coordinates of Principal Coordinate Analysis (PCoA) explained only 24.5% of the total variability (Fig. 6b, c). In the 1 and 2 coordinate system, the presence of four groups, i.e. the cultivated form, population originating from the Podkarpackie Voivodeship with a separate population PL 403183 and a large group consisting of populations from the other three Voivodships, was clearly visible. The third coordinate allowed to isolate as well as a group of individuals from Mazowieckie Voivodeship.

In order to further elaborate the genetic structure of the populations, a model-based Bayesian clustering was conducted using the Structure 2.3.4 software. In order to find the suitable number of clusters (K) it was tested in the range from 1 to 10 and was plotted against ΔK which showed a sharp peak for $K=5$ (Fig. 6d), hence, five distinct groups were found to contribute significant genetic information across

tested accessions. The sub-populations were denoted as POP1 – POP5, and each sub-population contributed in 10%, 5%, 56%, 21% and 7% of total *V. officinalis* genetic makeup, respectively (Fig. 6d). The proportion of membership of individuals in each population was illustrated in the bar plot of the population assignment test in structure analysis (Fig 6e, f). Genotypes that the estimated proportion of membership (Q) scored >0.80 were considered as pure while <0.80 as admixed. It suggested that all cultivated form's individuals were assigned entirely in POP1, the same as four ones from population PL 406738 and one from PL 403192 that were originated from the Mazowieckie Voivodeship. The POP2 was comprised of seven individuals from both Mazovian populations. The remaining six plants from that region showed intermediated genetic composition and were determined as heterogeneous. POP3 dominated in wild populations originating from Świętokrzyskie and Kujawsko-Pomorskie Voivodships, although in some part of the accessions which collection sites were located further to the South, the presence of about 10-20% of genetic information from POP5 was also noted. Most of populations collected in Bieszczady Mountains i.e. Podkarpackie Voivodship comprised admixtures of POP4 and POP3, however population PL 403183 represented POP4 and POP5 genetic makeup.

SilicoDArT markers

Ward's hierarchical grouping showed the presence of five main clusters (Fig. 7a). The groups were identified in a very similar way to the above-mentioned SNP analysis. The only difference corresponded to the classification of PL 403183 from Podkarpackie Voivodship as a separate group. Also, PCoA results for silicoDArT data showed high agreement with those reported for SNP markers (Fig. 7b, c). The three main coordinates explained 14.9% of the total variation.

Bayesian cluster analysis of silicoDArT markers using the ΔK method to determine optimal K values suggested that the most likely number of clusters was two (Fig. 7d), hence two distinct groups were found, and they correspond to cultivated and wild populations (Fig. 7e). The subpopulations were determined as POP1 and POP2 and constituted 11% and 89% of the total genetic information, respectively (Fig. 7f). The structure of the lower order for K=3 showed the existence of third group composed of populations originating from Podkarpackie Voivodship (Fig. 7g, h).

Joint analysis

Generalized Procrustes Analysis was performed based on the first ten principal coordinates obtained from PCoA for SNP and silicoDArT data. The most effective reduction of the total variability was scaling and rotation ($p < 0.0001$). The residuals by individuals after transformation were in the range 0.0001-0.009. On average, the highest value was observed in the cultivated form. The residuals by configuration after transformation were equal (0.169) indicating that both methods gave consensus results. A consensus test allowed to determine that the observed Rc (results of the consensus test) value was significantly higher than 95% of the results that were obtained when permuting data. The eigenvalues showed that the first three axes represented 53.56% of variability. The variability splitted between the methods showed that the first axis explained a bit more variation of SNP than silicoDArT data. The first three GPA factors explained 53.5% of total variance. Finally, scatterplots presented the consensus results.

Due to the high consistency of the results obtained for SNP and silicoDArT analysis, as before, five major groups could be identified. There was a clearly visible distinctiveness between the cultivated form and natural populations. Among the wild populations, the ones coming from the Bieszczady Mountains were distinct and moreover they showed additional internal differentiation. Also, a part of individuals from the Masovian region forms a separate aggregation. On the other hand, the population originating from Kujawsko-Pomorskie Voivodeship does not differ from the large group of populations collected in the Świętokrzyskie Voivodeship (Fig. 8).

Discussion

V. officinalis is a species commonly used in herbal medicine. It is widespread in Europe and Asia and is currently not considered to be endangered. Despite the great interest of researchers, this species has not yet received genomic characterization. Previous studies concerned mainly the content and specification of active compounds and their use for therapeutic purposes [1, 19-22]. Several studies using molecular markers, such as Amplified Fragment Length Polymorphism (AFLP) or Random Amplification of Polymorphic DNA (RAPD) were performed for this species, but none of them concerned the intraspecies diversity analysis [23-25]. In this context, *Valeriana wallichii* DC (syn. *Valeriana jatamansi* Jones) is much better recognized [26-28]. The NGS high throughput sequencing technology, i.e. RADseq, had already been used in the *Valerianaceae* phylogenetic research, but it only concerned the species native to South America and the *V. officinalis* was not included in it [29].

Utility of DArTseq markers

To the best of our knowledge, in the research presented in this paper, the NGS technology was used for the first time for the analysis of intraspecies variation within the *Valeriana* genus. Until now, the used here, DArTseq technique has been mainly applied for the analysis of crop and animal species [30-35]. For 188 individuals of *V. officinalis*, representing 19 wild populations and one cultivated form, over 65 000 SNP markers and over 61 000 silicoDArT markers with a very high reproducibility coefficient were obtained using this technique. For comparison, in *Triticum durum* Desf. DArTseq analysis allowed to identify more than 20 000 SNPs and 56 000 silicoDArTs [36]. In phylogenetic studies of the genus *Secale*, this technique allowed to identify about 14 thousand SNPs, and in studies of rye inbreeding lines almost 15 thousand silicoDArT and five thousand SNP markers were obtained [37, 38]. In the population genomics of *Eucalyptus* species 54 000 SNP markers were recorded [18]. In earlier studies of *V. officinalis* using AFLP markers only 311 fragments were generated. The PIC value obtained in the presented study for both SNP and silicoDArT markers was quite low and amounted to 0.181 and 0.123, respectively. It was a derivative of a significant fraction of monomorphic loci. In the above-mentioned cereals studies the PIC value was significantly higher and amounted to 0.302 and 0.265 in wheat and 0.37 and 0.22 in rye for SNPs and silicoDArTs, respectively [36, 38]. However, it should be noted that the genome of *V. officinalis* is more than 2.5 times smaller than the genome of *Secale cereale* L. and more than 4 times smaller than the genome of *T. durum* [23, 39, 40]. On the other hand, similar to the our (~ 0.18) PIC values were obtained in the study of *Ctenophorus caudicinctus*, *Litoria ewingii* and *Litoria paraewingii* the

Australian lizard and frogs [35]. Based on the above data, therefore, it can be concluded that the DArTseq technique is efficient and useful in the genetic analysis of *V. officinalis*.

Diversity

Due to the lack of earlier studies on genetic variability in the species *V. officinalis*, it is difficult to refer the level of parameters obtained in the presented paper to a species-specific diversity pattern. A substantial fraction of monomorphic or low polymorphic loci and a negligible proportion of unique alleles/loci indicates very homogenous genetic background of the examined natural populations. Low level of observed heterozygosity and positive values of the inbreeding coefficient indicate a prevalence of mating between closely related individuals. Considering that *V. officinalis* produces rhizome and has the ability to clonal propagation, there is a high probability that adjacent individuals at natural sites are in fact ramets clusters, i.e. a group of clones formed as a result of vegetative proliferation of a genet (mother plant). They can still be linked to the mother plant or exist separately as a result of the division of genet [41]. Moreover, according to the research of Konon and Novikova [42], in valerian the mechanism of self-incompatibility appears frequently to be absent. Thus pollination with pollen of the same genotype is possible, although fertilization with foreign pollen is promoted by the delayed development of the stigma in relation to the stamens [25]. This is also confirmed by the lack of correlation between the level of variation and the abundance of *V. officinalis* in the natural sites. Moreover, in the cultivated form, the level of genetic variability is significantly higher than in natural populations, and the inbreeding coefficient indicates the excess of heterozygotes even though the cultivated form was developed as a result of targeted selection. The higher variability of the cultivated form results probably from the farming practice and pollination biology of this species. Valerian has a high outcrossing rate of 76.5 to 97.7% under field conditions, as shown by studies of Penzkofer et al. [25]. Under the cultivation conditions for herbal raw material, the plants are unearthed in the second year of growing, and plantations are established from the seedlings. The result is that no clonal propagation takes place on plantations, i.e. a single plant is a single genet, and neighboring plants will probably have a different genetic makeup as a result of mixing of seeds before sowing and random planting of seedlings.

Gene pool

Despite, to a large extent, the homogeneous genetic background of natural populations, based on the analysis of the population structure, the differentiation of gene pools typical for particular regions of the country is clearly visible. It is most noticeable in the case of Podkarpackie voivodship, where there occurs an isolation due to the presence of the Bieszczady Mountains. However, population structure also indicates that there is no complete isolation of natural populations since in each of the regions some admixture from the adjacent gene pool(s) appears which is also indicated by the fact that the correlation between genetic and geographical distance was rather limited. The presence of the fifth group previously marked as POP5 at varying degrees in populations originating from Podkarpackie and Świętokrzyskie Voivodships may indicate that an additional gene pool may be present in Southern Poland. This may be facilitated by the topography of that part of the country, i.e. the presence of highlands and high

mountains (Tatra Mountains) with a more severe climate. The presence of *V. officinalis* on natural sites in that region has been proven [3, 41]. From the point of view of the genetic structure of the species population, the similarity of genotypes occurring in the relatively remote Świętokrzyskie and Kujawsko-Pomorskie Voivodships also seems to be interesting. This may indicate that a metapopulation with free gene flow occurs in a significant part of Poland. This hypothesis is supported by the widespread occurrence of valerian across the country, however, its verification will be possible only after investigating a larger number of natural populations, especially from regions not included in this study [3]. In the populations inhabiting the Mazowieckie Voivodeship, the presence of individuals of the genotype corresponding to the cultivated form 'Lubelski' was recognized. This is probably due to the presence of valerian plantations in the area. In other words, we are dealing with the uncontrolled spread of the cultivated form beyond the plantation area. Nonetheless, it is crucial that there is no evidence of crossbreeding wild and cultivated specimens. Valerian cultivation is very widespread and common in the Lubelskie Voivodeship, so in that area the probability of contamination of natural populations by the cultivated form is extremely high. Unfortunately, no environmental monitoring is carried out in this context.

Vulnerabilities

Although *V. officinalis* is still considered to be a commonly occurring species not only in Poland, but also throughout its entire range of distribution, due to the loss and degradation of habitats, the abundance of this species may start to decline rapidly. Climate change and increasing periods of drought are not without significance here [43]. Considering the low level of heterozygosity and definitely lower than assumed level of genetic variation, this species may face the threshold of extinction in a very short time as inbreeding decreases the reproductive capacity of outbreeding species [44]. An additional risk which is not considered is the uncontrolled release of cultivated forms into the environment. Since valerian is a native species, no environmental monitoring is carried out in this area, especially that without the use of molecular techniques it is basically not possible. However, it should be expected that cultivated forms developed by selection and not representing a complete genetic makeup of the species will have also negative impact on the size of the gene pool. The cultivation of medicinal plant species, on the one hand, enhances the protection of wild populations which are not drained to obtain raw herbal resources, and on the other hand, may result in disturbance of the structure and genetic balance of wild populations through the uncontrolled spread of individuals of cultivated forms beyond the plantation area. Based on the results presented here, it is impossible to assess to what extent this threat is real and what long-term effects it will have. However, there is no doubt about the existence of this problem and this issue should be explored further in the future.

Conclusions

- The DArTseq technology can be applied effectively in genetic studies of *officinalis*.
- The genetic variability of wild populations is in fact significantly lower than assumed.

- Individuals from the cultivated population are found in the natural environment and their impact on wild populations should be monitored.
- The received genetic fingerprints data will allow to monitor the condition of wild populations in the future and to assess whether the genetic integrity of the studied accessions has been maintained during gene bank storage.

Abbreviations

AFLP - Amplified Fragment Length Polymorphism

AMOVA - Analysis of Molecular Variance

DArT - Diversity Arrays Technology

DNA - deoxyribonucleic acid

F_{IS} - inbreeding coefficient

Φ_{PT} - differentiation among populations coefficient

GBS - genotyping-by-sequencing

GPA - Generalized Procrustes Analysis

H_o - observed heterozygosity

K - suitable number of clusters

MAPs - medicinal and aromatic plants

NGS - Next Generation Sequencing

P - calculated probability

PCoA - Principal Coordinate Analysis

PIC - polymorphism information content

POP1-POP5 – population1 – population5

Q - proportion of membership

RADseq - restriction site associated DNA sequencing

RAPD - Random Amplification of Polymorphic DNA

Rc - Results of the consensus test

SNP - Single Nucleotide Polymorphism

uHe - Nei unbiased genetic diversity

WHO - World Health Organization

WULS-SGGW - Warsaw University of Life Sciences (SGGW)

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets used and analyzed during the current study are available from the Open Science Framework database and are publicly available under link: <https://osf.io/76342/>. The materials used during the current study are available from National Centre for Plant Genetic Resources, Radzików, Poland according to the Standard Material Transfer Agreement. Voucher specimens are deposited in a publicly available herbarium of Department of Vegetables and Medicinal Plants, WULS-SGGW.

Competing interests

The authors declare that they have no competing interests.

Funding

The molecular studies were financed by the National Science Centre from grant „MINIATURA I” no. 2017/01/X/NZ9/01170. Acquisition of seeds from wild populations was supported by the Ministry of Agriculture and Rural Development of Poland in the form of the Multi-annual program: 2015-2020 “Establishment of a scientific basis for biological progress and preservation of plant genetic resources as a source of innovation in order to support sustainable agriculture and food security of the country” coordinated by Plant Breeding and Acclimatization Institute (IHAR) - National Research Institute. The calculations were performed at the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw (ICM UW) within the framework of Computational Grant No. G72-19.

Author Contributions

MB, KB and ZW designed the study. KB, OK and ZW collected the seeds and established an *ex situ* collection. MB isolated DNA analyzed the data and wrote the paper. ZW and WP were responsible for the management of the Multi-annual program: 2015-2020. MB was responsible for the management of project no. 2017/01/X/NZ9/01170 and Computational Grant No. G72-19. KB, OK, AR, ZW and WP contributed to the final manuscript review and editing. AR and MB were responsible for the graphic presentation of the results.

Acknowledgements

The authors would like to thank Professor Monika Rakoczy-Trojanowska for her support in preparing the project proposal.

References

1. Patočka J, Jakl J: **Biomedically relevant chemical constituents of *Valeriana officinalis***. *J Appl Biomed* 2010, **8**(1):11-18.
2. Wichtl M: **Herbal drugs and phytopharmaceuticals: a handbook for practice on a scientific basis**: Medpharm GmbH Scientific Publishers; 2004.
3. Zając M, Zając A: **Elementy geograficzne rodzimej flory Polski: The geographical elements of native flora of Poland**. In.: Inst. Botaniki Univ. Jagiellońskiego; 2009.
4. Bone K, Mills S: **Principles and practice of phytotherapy: modern herbal medicine**: Elsevier Health Sciences; 2012.
5. Evstatieva L, Handjieva N, Popov S, Pashankov P: **A biosystematic study of *Valeriana officinalis* (*Valerianaceae*) distributed in Bulgaria**. *Plant Syst Evol* 1993, **185**(3-4):167-179.
6. Hidalgo O, Vallès J: **First record of a natural hexaploid population for *Valeriana officinalis*: genome size is confirmed to be a suitable indicator of ploidy level in the species**. *Caryologia* 2012, **65**(3):243-245.
7. Skalińska M: **Polyploidy in *Valeriana officinalis* Linn. in relation to its ecology and distribution**. *Bot J Linn Soc* 1947, **53**(350):159-186.
8. Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, Sundaresan V: **DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market**. *Plant Biotechnol J* 2016, **14**(1):8-21.
9. Schippmann U, Leaman DJ, Cunningham A: **Impact of cultivation and gathering of medicinal plants on biodiversity: global trends and issues**. *Biodiversity and the ecosystem approach in agriculture, forestry and fisheries* 2002.
10. Mesfin K, Tekle G, Tesfay T: **Assessment of threatening factors of medicinal plant species in Samre district, south-eastern Tigray, northern Ethiopia**. *J Med Plant Res* 2013, **1**(4):38-42.
11. Das M, Jain V, Malhotra S: **Impact of climate change on medicinal and aromatic plants**. *Indian J Agric Sci* 2016, **86**:1375-1382.

12. Chen S-L, Yu H, Luo H-M, Wu Q, Li C-F, Steinmetz A: **Conservation and sustainable use of medicinal plants: problems, progress, and prospects.** *Chin Med J* 2016, **11**(1):37.
13. Poland JA, Brown PJ, Sorrells ME, Jannink J-L: **Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach.** *PloS one* 2012, **7**(2):e32253.
14. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA: **Harnessing the power of RADseq for ecological and evolutionary genomics.** *Nat Rev Genet* 2016, **17**(2):81.
15. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML: **Genome-wide genetic marker discovery and genotyping using next-generation sequencing.** *Nat Rev Genet* 2011, **12**(7):499.
16. Jaccoud D, Peng K, Feinstein D, Kilian A: **Diversity arrays: a solid state technology for sequence information independent genotyping.** *Nucleic Acids Res* 2001, **29**(4):e25-e25.
17. Cruz VMV, Kilian A, Dierig DA: **Development of DArT Marker Platforms and Genetic Diversity Assessment of the U.S. Collection of the New Oilseed Crop *Lesquerella* and Related Species.** *PloS one* 2013, **8**(5):e64062.
18. Rutherford S, Rossetto M, Bragg JG, McPherson H, Benson D, Bonser SP, Wilson PG: **Speciation in the presence of gene flow: population genomics of closely related and diverging *Eucalyptus* species.** *Heredity* 2018:1.
19. Barton DL, Atherton PJ, Bauer BA, Moore Jr DF, Mattar BI, LaVasseur BI, Rowland Jr KM, Zon RT, LeLindqwister NA, Nagargoje GG: **The use of *Valeriana officinalis* (valerian) in improving sleep in patients who are undergoing treatment for cancer: a phase III randomized, placebo-controlled, double-blind study: NCCTG Trial, N01C5.** *J Support Oncol* 2011, **9**(1):24.
20. Circosta C, De Pasquale R, Samperi S, Pino A, Occhiuto F: **Biological and analytical characterization of two extracts from *Valeriana officinalis*.** *J Ethnopharmacol* 2007, **112**(2):361-367.
21. Letchamo W, Ward W, Heard B, Heard D: **Essential oil of *Valeriana officinalis* L. cultivars and their antimicrobial activity as influenced by harvesting time under commercial organic cultivation.** *J Agric Food Chem* 2004, **52**(12):3915-3919.
22. Singh N, Gupta A, Singh B, Kaul V: **Quantification of valerenic acid in *Valeriana jatamansi* and *Valeriana officinalis* by HPTLC.** *Chromatographia* 2006, **63**(3-4):209-213.
23. Bressler S, Klatte-Asselmeyer V, Fischer A, Paule J, Dobeš C: **Variation in genome size in the *Valeriana officinalis* complex resulting from multiple chromosomal evolutionary processes.** *Preslia* 2017, **89**:41-61.
24. Pant M, Nailwal TK, Tewari LM, Kumar S, Kumari P, Kholia H, Tewari G, Campus D: **Molecular characterization of *Valeriana* species with PCR, RAPD and SDS PAGE.** *Nature Sci* 2009, **7**(7):41-49.
25. Penzkofer M, Seefelder S, Heuberger H: **Estimation of outcrossing rates using genomic marker and determination of seed quality parameters in *Valeriana officinalis* L. sl under field conditions.** *Euphytica* 2018, **214**(5):81.
26. Rajkumar S, Singh SK, Nag A, Ahuja PS: **Genetic Structure of Indian Valerian (*Valeriana jatamansi*) Populations in Western Himalaya Revealed by AFLP.** *Biochem Genet* 2011, **49**(9-10):674-681.

27. Singh SK, Katoch R, Kapila RK: **Genetic and biochemical diversity among *Valeriana jatamansi* populations from Himachal Pradesh.** *Sci World J* 2015, **2015**.
28. Sundaresan V, Sahni G, Verma R, Padalia R, Mehrotra S, Thul ST: **Impact of geographic range on genetic and chemical diversity of Indian valerian (*Valeriana jatamansi*) from northwestern Himalaya.** *Biochem Genet* 2012, **50**(9-10):797-808.
29. Gonzalez LA: **Phylogenetics and mating system evolution in the Southern South American *Valeriana* (*Valerianaceae*).** University of New Orleans; 2014.
30. Edet OU, Gorafi YS, Nasuda S, Tsujimoto H: **DARTseq-based analysis of genomic relationships among species of tribe *Triticeae*.** *Sci Rep* 2018, **8**.
31. Zaitoun SYA, Jamous RM, Shtaya MJ, Mallah OB, Eid IS, Ali-Shtayeh MS: **Characterizing Palestinian snake melon (*Cucumis melo* var. *flexuosus*) germplasm diversity and structure using SNP and DARTseq markers.** *BMC Plant Biol* 2018, **18**(1):246.
32. Alam M, Neal J, O'Connor K, Kilian A, Topp B: **Ultra-high-throughput DARTseq-based silicoDART and SNP markers for genomic studies in macadamia.** *PloS one* 2018, **13**(8):e0203465.
33. Schultz AJ, Cristescu RH, Littleford-Colquhoun BL, Jaccoud D, Frère CH: **Fresh is best: Accurate SNP genotyping from koala scats.** *Ecol Evol* 2018, **8**(6):3139-3151.
34. Nguyen NH, Premachandra H, Kilian A, Knibb W: **Genomic prediction using DART-Seq technology for yellowtail kingfish *Seriola lalandi*.** *BMC Genomics* 2018, **19**(1):107.
35. Melville J, Haines ML, Boysen K, Hodgkinson L, Kilian A, Smith Date KL, Potvin DA, Parris KM: **Identifying hybridization and admixture using SNPs: application of the DARTseq platform in phylogeographic research on vertebrates.** *R Soc Open Sci* 2017, **4**(7):161061.
36. Baloch FS, Alsaleh A, Shahid MQ, Çiftçi V, de Miera LES, Aasim M, Nadeem MA, Aktaş H, Özkan H, Hatipoğlu R: **A whole genome DARTseq and SNP analysis for genetic diversity assessment in durum wheat from central fertile crescent.** *PloS one* 2017, **12**(1):e0167821.
37. Al-Beyroutiová M, Sabo M, Slezia P, Dušínský R, Birčák E, Hauptvogel P, Kilian A, Švec M: **Evolutionary relationships in the genus *Secale* revealed by DARTseq DNA polymorphism.** *Plant Syst Evol* 2016, **302**(8):1083-1091.
38. Targońska-Karasek M, Bolibok-Brągoszewska H, Rakoczy-Trojanowska M: **DARTseq genotyping reveals high genetic diversity of polish rye inbred lines.** *Crop Sci* 2017, **57**(4):1906-1915.
39. Bauer E, Schmutzer T, Barilar I, Mascher M, Gundlach H, Martis MM, Twardziok SO, Hackauf B, Gordillo A, Wilde P: **Towards a whole-genome sequence for rye (*Secale cereale* L.).** *Plant J* 2017, **89**(5):853-869.
40. Bennett MD, Leitch IJ: **Nuclear DNA amounts in angiosperms.** *Ann Bot* 1995:113-176.
41. Kostrakiewicz-Gierałt K: **The variability of population and individual traits of medicinal plant *Valeriana officinalis* L. var. *officinalis* Mikan under different site conditions.** *Period Biol* 2018, **120**(1):41-50.
42. Konon N, Novikova N: **Reaction of *Valeriana officinalis* to inbreeding.** *Rastitel'nye resursy* 1981.

43. Ruosteenoja K, Markkanen T, Venäläinen A, Räisänen P, Peltola H: **Seasonal soil moisture and drought occurrence in Europe in CMIP5 projections for the 21st century.** *Clim Dyn* 2018, **50**(3-4):1177-1192.
44. Brook BW, Tonkyn DW, O'Grady JJ, Frankham R: **Contribution of inbreeding to extinction risk in threatened species.** *Conserv Ecol* 2002, **6**(1).
45. **EGISET a database system for documentation of NCPGR collections.**
46. Jaccard P: **Étude comparative de la distribution florale dans une portion des Alpes et des Jura.** *Bull Soc Vaudoise Sci Nat* 1901, **37**:547-579.
47. Peakall R, Smouse PE: **GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update.** *Bioinformatics* 2012, **28**(19):2537-2539.
48. Hubisz MJ, Falush D, Stephens M, Pritchard JK: **Inferring weak population structure with the assistance of sample group information.** *Mol Ecol Resour* 2009, **9**(5):1322-1332.
49. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I: **Clumpak: a program for identifying clustering modes and packaging population structure inferences across K.** *Mol Ecol Resour* 2015, **15**(5):1179-1191.

Tables

Table 1 The list of surveyed accessions. The abundance according to Braun-Blanquet scale. Accession number in accordance with the database of National Centre for Plant Genetic Resources (Polish GeneBank).

Population no.	Region - Voivodeship	Site	Accession no.	Latitude	Longitude	Abundance
1	Kujawsko-Pomorskie	Szamocin	401930	N 53 03 102	E 017 11 506	4
2	Podkarpackie	Karlików	401935	N 49 26 752	E 022 03 516	2
3		Turzańsk	403179	N 49 21 325	E 022 09 286	3
4		Płonna	403181	N 49 25 056	E 022 07 692	2
5		Wola Sękowa	403182	N 49 39 298	E 021 59 682	2
6		Droga na Dobrą	403183	N 49 36 334	E 022 12 478	1
7		Hroszówka	403184	N 49 41 278	E 022 15 037	1
8	Świętokrzyskie	Rembów	403186	N 50 43 330	E 020 31 539	2
9		Łukowa	403187	N 50 43 059	E 020 31 474	4
10		Brzeziny	401938	N 50 46 849	E 020 32 340	3
11		Włochy	401939	N 50 52 805	E 020 56 694	3
12		Kostomłoty II	401941	N 50 93 333	E 020 56 673	2
13		Przełom	401942	N 51 01 020	E 020 26 141	2
14		Smyków 1	403190	N 51 01 025	E 020 28 164	3
15		Droga na Jacentów	401945	N 51 04 365	E 020 23 061	4
16		Ruda Malenicka	401946	N 51 14 630	E 020 22 354	3
17		Radoszyce	403191	N 51 06 126	E 020 15 914	3
18	Mazowieckie	Kabaty 1	406738	N 52 07 282	E 021 05 541	2
19		Kabaty 2	403192	N 52 07 254	E 021 05 421	2
20	Cultivated form 'Lubelski'	Płońsk	504920	N 52 81 170	E 020 18 830	-

Figures

Figure 1

Geographical regions from which the wild populations of *V. officinalis* were originated (marked by green).
The map was created using MS Excel 2018 supported by Bing © HERE, MSFT, Microsoft, Wikipedia.

Figure 2

Distribution of DArTseq data for quality parameters: a) SNP data reproducibility; b) SNP data call rate; c) SNP polymorphic information content (PIC) distribution; d) silicoDArT data reproducibility; e) silicoDArT data call rate; f) silicoDArT PIC distribution.

Figure 3

The results of diversity analysis: a) the percentage of polymorphic loci within the populations; b) contribution of unique alleles/loci in the studied populations; c) percentage of unique alleles/loci in the examined regions. Accession numbering in accordance with Tab. 1

Figure 4

Mantel correlation test between a) SNP and silicoDArT markers; b) geographical distance and genetic distance based on SNP markers; c) geographical distance and genetic distance based on DArTseq markers.

Figure 5

Results of the AMOVA framework for SNP and silicoDArT markers

Figure 6

Results of population structure analysis based on SNP markers. a) Hierarchical grouping by Ward's method performed based on the Jaccard's distance matrix; b) plot of the first two axis of Principal Coordinate Analysis; c) plot of the first and the third axis of PCoA; d) the results of ΔK measure obtained from CLUMPAK software; e) the results of 10 000 iterations of STRUCTURE software with $K = 5$ where K is the number of groups assumed; each vertical bar represents single plant. The length of the colored

segment shows the estimated proportion of membership of that sample to each group. f) participation of five groups in total genetic information. Accession numbering in accordance with Tab. 1

Figure 7

Results of population structure analysis based on silicoDART markers. a) Hierarchical grouping by Ward's method performed based on the Jaccard's distance matrix; b) plot of the first two axis of Principal Coordinate Analysis; c) plot of the first and the third axis of PCoA; d) the results of ΔK measure; e) the

results of 10 000 iterations of STRUCTURE software with $K = 2$ where K is the number of groups assumed; each vertical bar represents single plant. The length of the colored segment shows the estimated proportion of membership of that sample to each group; f) participation of two groups in total genetic information; g) the results of STRUCTURE software with $K = 3$; h) participation of three groups in total genetic information. Accession numbering in accordance with Tab. 1

Figure 8

Results of joint analysis of SNP and silicoDArT data through Generalized Procrustes Analysis. a) 2d plot of the first two axis of GPA; b) 2d plot of the first and the third axis of GPA

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Equation1.png](#)
- [Equation2.png](#)