Insights into opium poppy (*Papaver* spp.) genetic diversity from genotyping-by-sequencing analysis

Uyen Vu Thuy Hong[1,2], Muluneh Tamiru-Oli[1,2,*], Bhavna Hurgobin[1,2], Christopher R. Okey[3], Artur R. Abreu[3], Mathew G. Lewsey[1,2,*]

[1]La Trobe Institute for Agriculture and Food, La Trobe University, AgriBio Building, Bundoora, VIC 3086, Australia

[2]Australian Research Council Research Hub for Medicinal Agriculture, La Trobe University, AgriBio Building, Bundoora, VIC 3086, Australia

[3]Palla Pharma Ltd, Docklands, VIC 3008, Australia.

***Authors for correspondence:**
Mathew G. Lewsey: m.lewsey@latrobe.edu.au
Muluneh Tamiru-Oli: m.oli@latrobe.edu.au

ABSTRACT

Opium poppy (*Papaver somniferum*) is one of the world's oldest medicinal plants and a versatile model system to study secondary metabolism. However, our knowledge of its genetic diversity is limited, restricting utilization of the available germplasm for research and crop improvement. We used genotyping-by-sequencing to investigate the level of genetic diversity and population structure in a collection of poppy germplasm consisting of 91 accessions originating in 30 countries of Europe, North Africa, America, and Asia. We identified five genetically distinct subpopulations. Although the accessions were from geographically diverse regions, no strong association was detected between geographic origin and subpopulation at regional and sub-regional levels. However, most accessions obtained from the same country were grouped together as genetically

distinct, likely a consequence of the restriction on movement of poppy germplasm. Phylogenetic analysis identified clades that were largely consistent with the subpopulations detected. Clades could be differentiated broadly based on capsule dehiscence, plant stature and seed weight, traits believed to be associated with poppy domestication. Our study determined that the genetically diverse collection was likely composed of both wild and cultivated forms. This has direct implications for germplasm management and utilization of the available diversity for genetic improvement of poppy.

Keywords: domestication, GBS, genetic diversity, opium poppy, phylogeny, population structure

INTRODUCTION

Opium poppy (*Papaver somniferum* L.) is one of the oldest cultivated plant species. Archaeological evidence shows that poppy has been cultivated and used for thousands of years, dating back to the earliest Neolithic ages (Kritikos & Papadaki, 1967; Askitopoulou *et al.*, 2002; Bernáth & Németh, 2009; Salavert *et al.*, 2018). However, its origin and domestication history has remained unclear until recently. Several lines of evidence, based mainly on archaeological data and geographical distribution of cultivated and wild species, suggest the Mediterranean as the centre of poppy origin and domestication (Salavert *et al.*, 2020; Jesus *et al.*, 2021). Changes in capsule and seed sizes and capsule indehiscence, which is the retention of seed in the capsules, are believed to be among the domestication-related traits in poppy (Zohary *et al.*, 2012). Currently, poppy is widely cultivated, both as a licit and illicit crop, in Asia, Europe, Oceania and South America as a major source of several benzylisoquinoline alkaloids (BIAs) for the pharmaceutical industry and for the clandestine production of heroin (Beaudoin & Facchini, 2014; Tamiru-Oli *et al.*, 2018; INCB/UN, 2021). Poppy seeds are also used in the food industry in baking and pressed for extraction of edible oil, while the plant is grown for ornamental purposes in some countries due to its attractive flowers (Labanca *et al.*, 2018).

Considerable poppy genetic diversity has been reported in several countries including India, Turkey, Czech Republic, and Australia (Saunders *et al.*, 2001; Celik *et al.*, 2016; Lahiri *et al.*, 2018; Srivastava *et al.*, 2020; Svoboda *et al.*, 2020). Germplasm collections of varying sizes exist in some of these countries (Bajpai *et al.*, 2001; Prajapati *et al.*, 2002; Celik *et al.*, 2016; Lahiri *et al.*, 2018). Additionally, a substantial number of poppy genetic resources are currently maintained as seeds in global genebanks. The Leibniz Institute of Plant Genetics and Crop Plant Research (IPK)

genebank in Germany has over 1,100 accessions of poppy that were collected worldwide (Börner, 2006). A collection of similar size is maintained at the Institute of Protection of Biodiversity and Biological Safety in the Slovak University of Agriculture (Brezinova *et al.*, 2009). Germplasm collections provide the genetic and phenotypic diversity used in crop breeding and development. They are also vital resources for research aimed at dissecting the genetic and molecular basis of essential plant processes including secondary metabolism. However, these resources are largely underexploited in poppy because the genetic diversity of the available germplasm has not been studied in detail. Most available reports are based on studies that either assessed germplasm from a single country or used a limited number of classical DNA markers such as amplified fragment length polymorphisms (AFLPs), random amplified polymorphic DNAs (RAPDs) and simple sequence repeat markers (SSRs) (Saunders *et al.*, 2001; Acharya & Sharma, 2009; Celik *et al.*, 2014; Celik *et al.*, 2016).

Molecular markers have been instrumental for the study of genetic diversity and population structure of germplasm collections. Such studies generate information vital for both germplasm conservation and utilization of these resources for genetic improvement of crops. As the most abundant types of sequence variations in plant genomes, single nucleotide polymorphisms (SNPs) are suitable for several applications that require high-density and genome-wide markers including genetic diversity and population structure analyses, QTL mapping and map-based cloning (Batley & Edwards, 2007; Kumar *et al.*, 2012). Recent advances in next generation sequencing (NGS) have greatly reduced the cost of genome sequencing, allowing the generation of unlimited number of molecular markers. Genotyping-by-sequencing (GBS) is a low-cost and fast method for genome-wide SNP discovery and mapping and reduces the complexity of genomes by generating smaller fragments via restriction digestion (Elshire *et al.*, 2011). This makes the method suitable for the study of species with large genomes such as poppy. The *P. somniferum* genome has been sequenced recently, and repetitive elements make up 70% of the estimated 2.8 Gb genome (Guo *et al.*, 2018; Li *et al.*, 2020). The availability of this reference genome is believed to facilitate the genetic analysis of poppy and the mapping and isolation of genes or genomic regions associated with traits of interest.

In this work, we employed GBS analysis to characterize 91 poppy accessions consisting of two *Papaver* species, *P. somniferum* and *P. setigerum*, and originating from diverse geographic regions that covered 30 countries of Europe, North Africa, America, and Asia. *P. setigerum* is commonly thought to be the direct ancestor of *P. somniferum* and has historically also been used for alkaloid production (Hammer, 1981). We provide a genome-wide assessment of the genetic diversity and

population structure of the accessions using GBS and shed light on their phylogenetic relationships. We also show the lack of strong association between genetic clusters and geographic origin at regional and sub-regional levels, which is likely a consequence of the strictly regulated international law that governs the production and movement of opium poppy. Data generated here will allow effective management of poppy genetic resources and facilitate the genetic improvement of opium poppy through the development molecular markers, selection of parents and the identification of useful traits.

MATERIALS AND METHODS

Plant materials and morphological characterization

Seeds of 95 *Papaver* accessions from diverse geographical origins were obtained from the global poppy germplasm collection maintained at Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Genbank in Germany (Table **S1**). Of these, 91 successfully germinated accessions were used for GBS analysis. The accessions were morphologically diverse materials belonging to the two main *Papaver* species: *P. somniferum* (88 accessions) and *P. setigerum* (3 accessions) (Fig. **1**). Seeds were sown in 200 mm pots using a standard potting mix. Two pots (each a replication) were used per line and plants were thinned to two per pot after germination. Plants were grown in a glasshouse under constant temperature (22 °C day and night) and photoperiod (17 hr/7 hr light/dark). At full maturity, data was recorded on plant height (from ground level to the top of capsule of the tallest branch), number of branches (main stem plus side branches), and capsule and seed characteristics. Capsules were harvested for alkaloid profiling at the dry capsule stage. For measuring seed wight, we first removed debris and counted 1000 clean seeds using the Contador seed counter fitted with feed container no. 3 for fine seeds (Pfeuffer GmbH, Germany).

DNA isolation, GBS library preparation and sequencing

For DNA extraction, ~50-100 mg fresh young leaves were harvested in liquid nitrogen and ground to a fine powder using a TissueLyser II system (Qiagen). Genomic DNA was extracted using the CTAB (cetyl trimethylammonium bromide) method (Murray & Thompson, 1980) with the following minor modification. Following grinding of the samples but prior to addition of the extraction buffer, samples were washed using a 0.1 M HEPES (N-2-hydroxyetylpiperazine-N'-ethanesulphonic acid) buffer (pH 8.0) containing 1% polyvinylpyrrolidone, 0.9% L-ascorbic acid, and 2% 2-mercaptoethanol to remove polysaccharides and phenolic compounds. DNA quality

was checked by agarose gel electrophoresis and quantified with a NanoDrop spectrophotometer ND-1000 version (Thermo Fisher Scientific, Wilmington, DE, USA). For GBS library construction, the double digest RAD-seq (ddRAD) based library preparation protocol was used (Peterson *et al.*, 2012). The protocol included the following steps: DNA digestion with two restriction enzymes, ligation of barcoded adapters compatible with restriction sites overhang, size selection of pooled digested-ligated fragments using Blue Pippin, and amplification of library via PCR using indexed primers. For protocol optimization, eight double digested libraries were prepared from a pool of 3 representative samples (PAP 630, PAP 696 and a commercial cultivar) using the restriction enzyme combinations *Pst*I/*Msp*I, *Pst*I/*Mse*I, *Pst*I/*Nla*III, *Pst*I/*Hpy*CH4IV, *Eco*RI/*Msp*I, *Eco*RI/*Mse*I, *Eco*RI/*Nla*III and *Eco*RI/*Hpy*CH4IV. The library generated using *Eco*RI and *Nla*III was sequenced on the Illumina NextSeq500 platform (Illumina, San Diego, CA, USA) with the standard protocol for single-end reads in 150-cycle mid-output mode at the Australian Genome Research Facility (Melbourne, Australia).

Mapping and SNP calling

Raw sequence data were demultiplexed and sorted using the "process_radtags" function in STACKS v2.41 with the default parameters: "--inline-index" for barcode option and "--renz_1 EcoRI --renz_2 NlaIII" for enzymes option (Catchen *et al.*, 2013). After trimming barcode sequences, all trimmed reads (150 bp) were checked for quality, and low-quality reads (with quality score of less than 10) and "no RadTag" reads were removed. The filtered reads were aligned to the draft poppy genome sequence retrieved from NCBI (GCA_003573695.1_ASM357369v1) using BWA v0.7.17 and SAMtools v1.9 (Li & Durbin, 2009; Li *et al.*, 2009; Guo *et al.*, 2018). SNP calling was performed using the refmap.pl pipeline of STACKS v2.41. All the 91 accessions were treated as a single population. SNP loci that were mapped with reads from less than 90% of the accessions sequenced (--min-samples-per -pop 0.9) were excluded from further analysis. The density and distribution of the filtered SNPs across poppy chromosomes were determined using BEDTools v2.26.0 and SnpEff v4.1l (Quinlan & Hall, 2010; Cingolani *et al.*, 2012). To visualize SNP density, the "MVP.report" function with 1Mb non-overlapping window from rMVP package was performed in R studio (Team, 2013; Team, 2020; Yin *et al.*, 2021).

Population structure analysis

The population structure of the accessions was assessed using three methods. First, the accessions were clustered using identity-by-state (IBS) to determine the relationship between accessions based on the proportion of shared alleles between pairs of individuals in PLINK v1.07 (Purcell, S. *et al.*,

2007). Second, for elucidation of the genetic structure and identification of the optimum number of subgroups, we removed the outlier PAP 400 and applied the Bayesian information criterion (BIC) and Bayesian clustering - STRUCTURE analyses to the remaining 90 accessions. BIC analysis as a nonparametric method was performed with the adegenet package v2.1.3 in R studio based on 131,039 SNP dataset (Jombart, 2008; Jombart & Ahmed, 2011; Team, 2013). The best number of subgroups/subpopulations was determined as the *K*-means corresponding to the lowest BIC score using the "find.clusters" function. To determine the relationship between the subpopulations, we carried out discriminant analysis of principal components (DAPC). A cross validation function "Xval.dapc" was used to determine the optimal number of PCs to be retained.

Third, to investigate the population structure in detail, the admixture within the accessions was determined using Bayesian clustering based on a Bayesian Markov Chain Monte Carlo model (MCMC) implemented in STRUCTURE v2.3.4 (Pritchard *et al.*, 2000). With the assumption that SNPs at the same RAD locus are linked, --write_single_snp flag (in STACKS) was applied to ensure that only one SNP per RAD locus was used for STRUCTURE analysis. To determine the most likely number of subpopulations (clusters), four independent runs with 500,000 iterations and a 150,000-step burn-in period were performed for each *K* from 1 to 10. The output was obtained by structureHarvester v0.6.93 using the maximum estimated log-likelihood [log(P(X|K)] model and the highest ΔK in Evanno method (Pritchard *et al.*, 2000; Evanno *et al.*, 2005; Earl & vonHoldt, 2012). After determining the most probable *K* values, ten runs of 500,000 iterations followed by a 150,000 step burn-in were performed using STRUCTURE for each *K*. Additionally, for each optimal *K*, CLUMPP was used to generate individual and population Q matrices from the membership coefficient matrices of the ten replicates obtained from STRUCTURE (Jakobsson & Rosenberg, 2007). Bar plots were generated using DISTRUCT software (Rosenberg, 2004).

Genetic diversity and differentiation

Common measures of genetic diversity including private allele number (AP), percentage of polymorphic loci (%Poly), observed and expected heterozygosity (H_O and H_E), nucleotide diversity (π) and inbreeding coefficient (F_{IS}) were calculated for the five subpopulations using the "populations" function in STACKS v2.41 (Catchen *et al.*, 2013). The genetic differentiation between the subpopulations was calculated based on pairwise population differentiation (F_{ST}) values from GENODIVE v3.05 (Meirmans, 2020). Significance levels ($\alpha = 0.05$) of the F_{ST} values were determined by running 999 permutations and assessing this against a Bonferroni-adjusted *P*-value to account for multiple testing. The correlation matrix was visualized using the corrplot package in R studio (Team, 2020; Wei & Simko, 2021). To determine the distribution of genetic

variation, analysis of molecular variance (AMOVA) was performed using GENODIVE v3.05 (Meirmans, 2020). Significance level was tested using 999 permutations.

Phylogenetic analysis

To understand the evolutionary relationships among the poppy accessions, maximum-likelihood (ML) phylogenetic analysis was performed using IQ-TREE software (Nguyen *et al.*, 2015). For constructing the ML phylogenetic tree, the unlinked SNPs generated using the "populations" function in STACKS were exported as PHYLIP format (--phylip-var) to IQ-TREE software. Fast model-selection method (-m MFA) and ultrafast bootstrap (-bb 1000) were applied for accurate phylogenetic estimates and branch supports in IQ-TREE (Kalyaanamoorthy *et al.*, 2017; Hoang *et al.*, 2018). Phylogenetic trees from this analysis were visualized using iTOL v4 (Letunic & Bork, 2019).

Alkaloid profiling

Alkaloid content of 90 accessions was measured according the protocol used by Dittbrenner and colleagues (Dittbrenner *et al.*, 2009). Only accessions that produced enough capsules for analysis were included. Alkaloids were also analysed in four accessions that had failed to germinate during the first trial and subsequently lacked GBS data (Table **S1**). For testing the relationship between alkaloid and genetic diversity, we conducted principal component analysis (PCA) using TASSELv5.2.73 and plotted the graph using ggplot2 package in R studio (Bradbury *et al.*, 2007; Wickham, 2016; Team, 2020).

Genome size and ploidy analysis

The genome size and ploidy level of selected accessions was estimated using flow cytometry analysis of propidium iodide (PI)-stained nuclei isolated from poppy leaves. Nuclei of tomato (*Solanum lycopersicum*), which has genome size of ~900 Mb (Consortium, 2012), were simultaneously isolated from leaves, stained and analysed with poppy nuclei as an internal reference standard. Nuclei were isolated with the Galbraith lysis buffer using a modified protocol from Gutzat and Scheid as follows: (1) Chopping fresh young leaf tissue (0.5 g) using double-sided razor blades in 2 mL ice-cold lysis buffer, (2) filtering the homogenate through a 40 μm nylon mesh, (3) adding 2.5 μL of RNase (10 mg/mL) to 500 μL filtered homogenate, then incubating on ice for 10 minutes, (4) centrifugation at 400 g for 3 minutes, removing the supernatant and resuspending the pellet gently in 1 mL lysis buffer, then incubating on ice for 15 minutes, and (5) filtering the homogenate again through a 40 μm nylon mesh (Galbraith *et al.*, 1983; Gutzat & Scheid, 2020).

For nuclei staining, 25 μL PI (1 mg/ml) was added into 500 μL nuclei solution to get a final concentration of 50 μg/mL. Samples were screened on a CytoFLEX S flow cytometer (Beckman Coulter, Brea, CA, USA). The PerCP-A fluorescence intensity of G1 and G2 phase cells of internal standard and samples was used to estimate genome size and ploidy level of the samples.

RESULTS

Optimization of a poppy GBS protocol for SNP discovery

To generate a sufficient SNP dataset for genetic diversity analysis, we first optimized a GBS protocol for opium poppy. GBS remains a method of choice for genome-wide SNP detection in non-model species and species with large genomes. It is however yet to be utilized for opium poppy. A critical step in GBS protocols is the reduction of genome complexity using restriction enzymes (REs), and the two-enzyme GBS protocol uses a combination of a rare- and common-cutting REs (Poland *et al.*, 2012). Although some RE combinations are often used in plants, the optimal combinations need to be determined for the genome of each species (Glaubitz *et al.*, 2014). To select the optimal enzyme combination for opium poppy, we prepared eight double digested libraries from a pool of 3 representative samples using the enzyme combinations *Pst*I/*Msp*I, *Pst*I/*Mse*I, *Pst*I/*Nla*III, *Pst*I/*Hpy*CH4IV, *Eco*RI/*Msp*I, *Eco*RI/*Mse*I, *Eco*RI/*Nla*III and *Eco*RI/*Hpy*CH4IV. We then compared the absence of visible repeat regions within the size selection area (280 bp-375bp) and level of amplifications to select *Eco*RI/*Nla*III as the optimal combination for opium poppy GBS library preparation (Fig. **S1**).

The multiplexed pool of *Eco*RI/*Nla*III-based GBS libraries representing 91 accessions was sequenced to generate a total of 103,802,122 raw 150-bp single end reads (15.57 Gb data, average 1.14 million reads per sample, Table **S2**). We aligned 103,601,279 reads to the poppy reference genome after filtering, with a read alignment rate of 97% to 99% (Table **S2,S3**). A total 165,363 SNPs were identified at 76,407 loci and present in ≥90% of accessions. These SNPs were evenly distributed across the 11 chromosomes and unplaced scaffolds of opium poppy (Fig. **2a,b**). The 165,363 SNPs were predicted to have 165,786 effects, of which 149,536 (90.2%) were found in intergenic regions, while the remaining 16,250 (9.8%) corresponded to genic regions (Table **S4**). This optimized protocol will allow researchers to rapidly apply GBS to unlimited number of poppy accessions at reduced cost, allowing detailed characterization of the available germplasm.

Assessing genetic relatedness of 91 *Papaver* accessions

Next, we set out to determine the relationship amongst the 91 *Papaver* accessions from a broad geographic range. The accessions, which originated from 30 countries in four continents, were primarily *P. somniferum* (88 accessions) but also included 3 *P. setigerum* accessions (Table **S1**). We first used pairwise comparisons of the 165,363 filtered SNPs for hierarchical clustering based on the identity-by-state algorithm (Purcell, Shaun *et al.*, 2007). Four clusters plus a single accession (PAP 400) far from all others were identified (Fig. **3**). Cluster 1 contained the three *P. setigerum* accessions, while the *P. somniferum* accessions, except PAP 400, were grouped into three distinct clusters. Although PAP 400 was labelled as a *P. somniferum* accession, our result shows that it belongs neither to *P. somniferum* nor *P. setigerum*. PAP 400 was also morphologically distinct from both *Papaver* species (Fig. **4a**).

To study the relationships between *P. somniferum*, *P. setigerum* and PAP 400, we determined genome sizes and ploidy levels of PAP 400 and representative *P. somniferum* and *P. setigerum* accessions by flow cytometry. We estimated the genome size of *P. somniferum* to be ~3.04 Gb, which is slightly larger a previous ~2.87 Gb estimate (Guo *et al.*, 2018) (Fig. **4b**). The genome of *P. setigerum* was estimated to be ~4.9 Gb, indicating the *P. setigerum* genome is close to twice the genome size of *P. somniferum* (Fig. **4c,f**). This was similar to previous genome size estimates and supports reports that *P. setigerum* is tetraploid (2n = 4× = 44) with chromosomes smaller in size compared with the diploid (2n = 2× = 22) *P. somniferum* (Malik *et al.*, 1979; Wakhlu & Bajwa, 1987; Srivastava & Lavania, 1991). PAP 400 had a slightly smaller genome size than the diploid *P. somniferum* (Fig. **4d,e**). Taken together, our results suggested that PAP 400 is likely a different species and a case of mislabelling, which can occur in seedbanks during plant establishment and storage (Akpertey *et al.*, 2020). This is plausible considering that IPK holds seeds of other *Papaver* species in its collection.

Population structure and genetic diversity amongst *Papaver* accessions

Understanding the genetic structure of populations is useful for germplasm conservation and plant breeding. Patterns of genetic diversity are also informative about the history of domestication of crops. To infer population structure, we analysed the 90 *Papaver* accessions, removing the outlier PAP 400 and associated data (131,039 SNPs remaining). Five clusters (subpopulations) were inferred at the lowest Bayesian information criterion (BIC) score (Fig. **5a**). To understand the genetic relationships between the five subpopulations, we carried out DAPC. Ten principal components (PCs) were retained (with 47.75% of the variance conserved) by the cross-validation

function, which gave four discriminant eigenvalues (Fig. **5b**). Cluster 1 was comprised of all three *P. setigerum* accessions, while Clusters 2, 3, 4, and 5 consisted of 12, 4, 21, and 50 *P. somniferum* accessions, respectively. The wide separation between Cluster 1 and the other clusters on the DAPC plot illustrated the extensive genetic difference between *P. somniferum* and *P. setigerum*.

We investigated population structure in greater detail and detected admixture within the 90 poppy accessions by applying the admixture model in STRUCTURE, clustering the accessions using 49,166 unlinked SNPs (Pritchard *et al.*, 2000). The most probable number of clusters detected was four to eight (Fig. **S2**). Bar plots for each optimal K value, with the accessions sorted following the DAPC clustering, illustrated that the clustering pattern was similar to the DAPC analysis result (Fig. **5c**). At K = 5, Cluster 1 was composed of all three *P. setigerum* accessions, which are genetically distinct from the other accessions (Fig. **5c**). Interestingly, Cluster 2 was a less genetically diverse group of 12 accessions of broad geographic origins extending from North Africa to East Asia, indicative of germplasm exchanges in the past. Cluster 3 contained the four accessions from North Korea with moderate level of admixtures from Clusters 2 and 5. Cluster 4 was composed of 21 diverse accessions with a high level of admixtures from Clusters 2, 3 and 5. Most of the accessions in Cluster 4 originated from western and Mediterranean regions of Europe. Considering that opium poppy was domesticated in the western Mediterranean, from where it spread to north and central Europe, this admixture might be due to ongoing gene flow between wild and domesticated forms. Cluster 5 contained 50 accessions with low level of admixtures from all the other clusters (Salavert *et al.*, 2020; Jesus *et al.*, 2021). Subpopulations were strongly differentiated, as shown by pairwise calculation of genetic differentiation (F_{ST}), suggesting low levels of allele sharing (Fig. **5d**; Table **1**).

There were noticeable differences in genetic diversity between the five subpopulations (clusters) (Table **2**). The number of private/unique alleles (AP) ranged from 790 (Cluster 3) to 37,157 (Cluster 1; *P. setigerum*), calculated from the 131,039 SNP dataset. This further confirmed the genetic distinctiveness of *P. setigerum*. The percentage of polymorphic loci varied from 3.93% (Cluster 3) to 25.05% (Cluster 5). The level of observed heterozygosity (H_O) was highest for Cluster 1 (0.166) and lowest for Cluster 2 (0.009). H_O was lower than the expected heterozygosity (H_E) for all subpopulations except Cluster I (*P. setigerum*), indicating high levels of inbreeding in *P. somniferum*. This finding was supported by the higher inbreeding coefficients (F_{IS}) for the *P. somniferum* subpopulations (0.061 to 0.377). F_{IS} was negative for Cluster 1, possibly a consequence of an excess of the observed heterozygotes. The highest nucleotide diversity (π) was observed in Cluster 1 (0.219) and the lowest in Cluster 3 (0.040). The genetic variations were both due to

differences between (41.6%) and within (44.8%) subpopulations, determined using analysis of molecular variance (AMOVA; Table **3**). Taken together, our findings suggest that long-term selection and geographical isolation have contributed to differentiation of the subpopulations.

Geography is an important variable that drives population divergence because it is assumed that genetic differences generally increase with geographic distance (Sexton *et al.*, 2014). To test the hypothesis that geographic distance has driven the genetic differences in the poppy populations of our study, we combined the population structure analyses with information on the country of origin of each accession (Fig. **6**). Overall, no strong association was detected between the geographic origin of accessions and population groups at regional and sub-regional levels. However, the North Korean accessions formed a distinct population. Additionally, some accessions obtained from the same countries were grouped together as genetically similar. Examples included accessions from Japan, Morocco, Belgium, Switzerland, Germany, Mongolia, Russia, Australia, Bulgaria, and Czechoslovakia (Fig. **6**). We interpret these results as reflecting the highly restricted movement of poppy germplasm between different countries or regions and the controlled circumstances under which poppies are cultivated.

Variability in alkaloid content across genetically distinct poppy accessions

Papaver cultivars and accessions exhibit notable variability in the quantity and composition of their alkaloid contents (Dittbrenner *et al.*, 2009; Shukla *et al.*, 2010). Understanding the genetic basis of this variability would improve knowledge of alkaloid biosynthesis and may provide tools for breeding and synthetic biology. We examined this by quantifying both the major (morphine, codeine, thebaine) and minor (papaverine and opripavine) alkaloids in dry capsules of the 90 accessions that reached full maturity (Fig. **7**). We observed considerable variation among the accessions in alkaloid content and composition (Fig. **7**). Total alkaloid content ranged from 0.125 (PAP 795) to 1.610 (PAP 784) g/100 g DW, whereas morphine content varied between 0.072 (PAP 229A, *P. setigerum*) and 1.416 (PAP 784) g/100 g DW. Codeine and thebaine contents ranged from 0.002 (PAP 795) to 0.342 (PAP 151) and 0.000 to 0.336 (PAP 719) g/100 g DW, respectively (Fig. **7a**). Eight of the accessions analysed (9%) did not have a detectable level of thebaine. Papaverine and oripavine content ranged from 0.000 to 0.077 and 0.102 g/100 g DW, respectively. We identified 18 accessions with undetectable levels of papaverine and six with undetectable oripavine. These results demonstrate that considerable diversity in total alkaloid contents exists in the accessions studied.

The relative abundance of individual alkaloids also varied across accessions. Morphine was the most abundant alkaloid in 85 of the 90 accessions (94.4%), ranging from 32.3% to 96% (of total alkaloids) (Fig. **7b**). Codeine was the most abundant alkaloid (over 50% of total alkaloids) in only three accessions (PAP 150, PAP 151 and PAP 152), though 14 accessions had codeine abundances of 20% (of total alkaloids) or more. Notably, all three high codeine accessions were collected in Morocco. Thebaine, papaverine or oripavine were the most abundant alkaloid in none of the accessions. Thebaine composition ranged from 0.0% to 24.4% (of total alkaloids), with the highest in PAP 719. The proportions of papaverine ranged from 0.0% to 9.4% and oripavine from 0.0% to 29.9% (of total alkaloids).

We tested for relationships between genetic and chemical diversity by applying principal component analysis based upon the alkaloid profiles of the accessions, then compared subpopulation clustering patterns (Fig. **7c** and Fig. **S3**). The accessions of subpopulations 4 and 5 had the most similar alkaloid composition, demonstrated by PCA (PC1 and PC2 accounting for 92.4% of variability, Fig. **7c**). The *P. setigerum* accessions were clearly separated from all others, while accessions of subpopulation 2 had the most diverse and distinct alkaloid composition among the *P. somniferum* accessions. Accessions with the highest codeine proportions, including PAP 150 (53.9%), PAP 151 (52.0%), PAP 152 (50.4%), PAP 200 (42.2%), PAP 739 (35.6%), PAP 149 (33.8%) and PAP 354 (28.7%), were clearly separate from the others (Fig. **7c**). Accessions of Clusters 4 and 5 had a similar alkaloid profile. These results suggested a congruent pattern of genetic and chemical diversity in poppy. For a crop like poppy that is highly valued for its secondary metabolites, understanding patterns of metabolic variation within the context of natural genetic diversity is important both for conservation and utilization of the available germplasm.

Phylogenetic relationships of opium poppy

To understand the phylogenetic relationships between *P. somniferum* accessions, we constructed a maximum-likelihood phylogeny from unlinked SNPs (49,160 variable sites and 25,917 parsimony informative sites) using the IQ-tree software package. The three *P. setigerum* accessions were used as outgroups (Nguyen *et al.*, 2015). Three main clades and a number of outgroups were identified, which mostly reflected the four *P. somniferum* subpopulations identified by STRUCTURE analysis (Fig. **5b** and **8**). Clade I was composed of accessions from subpopulation 4 (Fig. **8**). Clade II contained all the 12 accessions from subpopulation 2, all the four accessions from subpopulation 3, and two accessions from subpopulation 5. Accessions from subpopulation 2 and 3 formed monophyletic sub-clades within Clade II. Clade III included 45 accessions of subpopulation 5. The length of the phylogeny branches suggested that the most genetic divergence happened within

accessions of Clade I, which is consistent with the earlier STRUCTURE analysis indicating that the highest amount of genetic variation was in subpopulation 4 (Fig. **5b** and **8**).

We examined morphological and agronomical traits that could potentially distinguish the various clades. We observed differences in capsule and seed morphology (Fig. **8**; Fig. **S4,S5**). All the 16 accessions in Clade I produced poricidal capsules (capsules with apical pores), similar to capsules of *P. setigerum* accessions. Two of the 18 accessions in Clade II (PAP 766 and PAP 767) produced poricidal capsules and 16 (~88.9%) indehiscent (closed) capsules. All the 45 accessions of Clade III, except two (PAP 466 and PAP 483), produced indehiscent capsules. We also found that variations in the agronomic-related traits of plant height, number of branches per plant and seed weight were broadly consistent with the subpopulations and phylogenetic groups identified (Fig. **9a**). Plant height ranged from 53.0 cm (PAP 229A) to 137.25 cm (PAP 156), branch number ranged from 1.75 (PAP 184) to 11.25 (PAP 229A), and 1000-seed weight from 0.165 g (PAP 255) to 0.815 g (PAP 733). The *P. setigerum* accessions and accessions in Clade 1 were generally dwarf, highly branching and produced lighter seeds (Fig. **9a**). Contrastingly, accessions in Clade 3 were taller, less branching and produced heavier seeds. We tested if the association between plant height, branch number seed weight was significant. We found a significant negative correlation between number of branches and height ($r=-0.48$, $p < 0.001$) and number of branches and 1000-seed weight ($r=-0.70$, $p < 0.001$), while 1000-seed weight and height were significantly positively corelated ($r=0.49$, $p < 0.001$) (Fig. **9b-d**). The loss of a wind-based seed dispersal mechanism through a transition from poricidal to indehiscent capsules is believed to be among the changes in morphological traits that occurred during poppy domestication (Zohary *et al.*, 2012). Furthermore, seed size, height and branching habit are important characteristics to distinguish *Papaver* species (Hrishi, 1960; Jesus *et al.*, 2021). Accordingly, we conclude that the accessions we studied included both wild and domestic variants.

DISCUSSION

The poppy germplasm currently available in several genebanks is a rich potential source of useful alleles. Accurate genotyping of this germplasm is crucial not only for maintenance of the available genetic diversity, but also to facilitate the identification and deployment of promising genotypes/alleles for genetic improvement of the crop. In the present study, we optimized a GBS protocol for poppy and applied GBS-based analysis of SNP markers for the assessment of genetic

relationships in diverse poppy accessions from two species. This optimized protocol provides a rapid and cost-effective method for genotyping of unlimited number of accessions.

The phylogenetic and taxonomic relationships of *Papaver* species, particularly that of the cultivated *P. somniferum* and the wild species *P. setigerum*, is still debated (Hammer, 1981; Srivastava & Lavania, 1991; Liu *et al.*, 2020; Salavert *et al.*, 2020). Some treat the two as separate species, while others consider *P. setigerum* as a subspecies of *P. somniferum* based on morphology and alkaloid profiles (Hammer, 1981; Liu *et al.*, 2020). Our findings confirmed the genetic separation of the two species (Fig. **3-5**). This result was expected, because it has been shown previously that *P. setigerum* is distinct from but phylogenetically closer to *P. somniferum* than all other *Papaver* species (Lane *et al.*, 2018; Liu *et al.*, 2020). *P. setigerum* is also considered the putative progenitor of cultivated *P. somniferum* (Hammer, 1981; Jesus *et al.*, 2021). Our result based on genome size analysis supports previous reports that *P. somniferum* is diploid (2n = 2× = 22) and *P. setigerum* tetraploid (2n = 4× = 44) (Fig. **4**) (Malik *et al.*, 1979; Wakhlu & Bajwa, 1987). These data do not support the hypothesis of a direct origin of *P. somniferum* from *P. setigerum*. However, they raise interesting questions about the possible relationships between the two species. Considerable homology exists between the two genomes, suggesting they may have a common origin (Malik *et al.*, 1979). Interspecific crosses are possible between the two species despite some meiotic abnormalities observed at the F1 generation (Malik *et al.*, 1979; Singh *et al.*, 1998). Notably, a diploid form of *P. setigerum* has been described (Hammer & Fritsch, 1977; Hammer, 1981). However, our data suggest that all the three accessions included in our study are tetraploids. For a detailed analysis of the genetic relationship of these two species, more samples from *P. setigerum*, including both the diploid and tetraploid forms, need to be studied.

The loss of dispersal mechanisms, increases in seed or organ size, changes in plant stature and reduced seed dormancy are among traits associated with domestication in plants (Ross-Ibarra *et al.*, 2007; Chen *et al.*, 2015). The capsules of many *Papaver* species and related genera dehisce primarily by small apical pores near the top of the capsule, facilitating seed dispersal by wind (Carolan *et al.*, 2006; Zumajo-Cardona *et al.*, 2018). Consequently, it is assumed that poricidal capsules are associated with wild forms and genotypes that produce indehiscent capsules, which retain seed in the capsules to full maturity and harvest, were selected for during poppy domestication (Zohary *et al.*, 2012). Our phylogenetic analysis determined that the three *P. setigerum* accessions and all the 16 accessions in Clade I produced poricidal capsules. Contrastingly, 14 of the 18 accessions in Clade II (77.8%) and 45 of the 47 accessions in Clade III (95.7%) had indehiscent capsules (Fig. **8**). Interestingly, *P. setigerum* accessions and accessions in Clade I were

also small in stature (dwarf and branching) and produced lighter seeds, while those in Clade III were relatively taller, less branching and with heavier seeds (Fig. **9**). Additionally, thirty-five of the 90 accessions in our study have distinct names, thirty-one of which (88.5%) had indehiscent capsules (Fig. **S4**, Table **S1**). Crop landraces, also known as local or farmers' varieties, are commonly referred to by descriptive names to distinguish them from one-another and these names may reflect intentional selection (Berg, 2009). Taken together, we hypothesize that most of the dwarf and branching accessions with poricidal capsules and lighter seeds were likely sampled from natural populations, while those taller accessions with fewer branches, indehiscent capsules and heavier seeds are cultivated forms. Their phylogenetic relationships seem to suggest their transition from wild to cultivated forms.

The accessions we studied were highly diverse in their alkaloid profiles. This suggests poppy has experienced extensive selection and breeding to improve alkaloid profiles of the plant. Although morphine was the dominant alkaloid in most of the accessions, we also identified accessions with codeine levels of up to 54% (of total alkaloids). Both natural and induced mutants with altered alkaloid profiles have been instrumental to elucidate the molecular mechanisms underlying differences in alkaloid contents in poppy (Millgate *et al.*, 2004; Winzer *et al.*, 2012; Pathak *et al.*, 2013; Chaturvedi *et al.*, 2014; Guo *et al.*, 2018). The transcriptional regulation of alkaloid biosynthesis in poppy has not been studied in detail. Diverse chemotypes are potential sources of gene expression or enzyme variants with differing activities that affect the alkaloid biosynthesis pathway. These are vital resources to understand the biochemical and genomic regulation of alkaloid biosynthesis. This is also key for breeding commercial lines with specific alkaloid abundances or desired alkaloid profiles. The development of the *top1* mutant, a high-thebaine and high-oripavine variety, was a significant commercial breakthrough that allowed the production of thebaine from morphine-free plants (Millgate *et al.*, 2004). Thebaine is used for the semi-synthesis of painkillers oxycodone and hydrocodone and the anti-opioid addiction drugs buprenorphine and naltrexone (Chen *et al.*, 2018). Similarly, the development of high codeine and morphine-free poppy varieties would allow the direct plant-based production of codeine while preventing the illicit synthesis of the morphine-derived heroin (Hagel & Facchini, 2010).

Our study demonstrates the utility of GBS for genetic analyses in opium poppy. Many hundreds of poppy germplasm accessions are available from genebanks worldwide. These are an immense potential resource if fully exploited. Application of GBS to this complete population, complemented by the currently available reference genome sequence of poppy, would drive further studies of the diversity, domestication, and phylogenetic relationships of opium poppy and

related species. Data generated from such studies would also enable selection of suitable parents, identification of genes associated with traits of interest and development of markers for the genetic improvement of the crop.

Table 1. Pairwise genetic differentiation (F_{ST}) between five subpopulations of opium poppy accessions calculated from 49,166 single nucleotide polymorphism loci. F_{ST} values are given below the diagonal and those in bold represent significant differences. Corrected *P* values are given above the diagonal (*, $\alpha \leq 0.005$, with Bonferroni correction for multiple comparisons).

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1		0.005	0.033	0.002	0.001
Cluster 2	0.587		0.001	0.001	0.001
Cluster 3	0.591	0.550		0.001	0.001
Cluster 4	0.407	0.371	0.272		0.001
Cluster 5	0.627	0.571	0.447	0.235	

Table 2. Measures of diversity for 90 *Papaver* accessions from five subpopulations calculated from 49,166 single nucleotide polymorphism loci.

Population	N	AP	%Poly	H_O	H_E	π	F_{IS}
Cluster 1	3	37,157	8.15	0.231	0.166	0.219	-0.015
Cluster 2	12	3,937	11.54	0.009	0.061	0.064	0.189
Cluster 3	4	790	3.93	0.009	0.034	0.040	0.061
Cluster 4	21	14,335	24.28	0.018	0.126	0.130	0.377
Cluster 5	50	15,714	25.05	0.016	0.076	0.077	0.313

N, number of individuals; Ap, private alleles; %Poly, percentage of polymorphic loci; H_O, observed heterozygosity; H_E, expected heterozygosity; π, nucleotide diversity; F_{IS}, inbreeding coefficient

Table 3. Analysis of molecular variance for 90 opium poppy accessions based on 49,166 single nucleotide polymorphism markers.

Source of variation	% Variation	*F*-value	*P*-value
Between subpopulations	41.6	0.416	0.001*
Between accessions within subpopulations	44.8	0.767	0.001*
Within accessions	13.6	0.864	

*$\alpha \leq 0.05$

ACKNOWLEDGMENTS

UTVH received a PhD scholarship from La Trobe University Graduate Research School. Work in the Lewsey lab is funded by the Australian Research Council Industrial Transformation Hub in Medicinal Agriculture (IH180100006) and a Commonwealth Scientific and Industrial Research Organisation SIEF STEM+ Fellowship with Palla Pharma Ltd. We thank La Trobe's Bioimaging Platform for support with genome size analysis.

AUTHORS' CONTRIBUTION

MTO, MGL, UVTH designed the study. UVTH, MTO, CRO, ARA conducted the experiments. UVTH, MTO, BH analysed the data. CRO, ARA provided research materials. UVTH, MTO, MGL interpreted the results and wrote the paper. All co-authors read and approved the final manuscript.

REFERENCES

Acharya HS, Sharma V. 2009. Molecular characterization of opium poppy (Papaver somniferum) germplasm. *American Journal of Infectious Diseases* **5**(2): 148-153.

Akpertey A, Padi FK, Meinhardt L, Zhang D. 2020. Effectiveness of Single Nucleotide Polymorphism markers in genotyping germplasm collections of Coffea canephora using KASP assay. *Frontiers in plant science* **11**: 2300.

Askitopoulou H, Ramoutsaki IA, Konsolaki E 2002. Archaeological evidence on the use of opium in the Minoan world. *International Congress Series*: Elsevier. 23-29.

Bajpai S, Gupta A, Gupta M, Kumar S. 2001. Inter-relationships between morphine and codeine in the Indian genetic resources of opium poppy. *Journal of herbs, spices & medicinal plants* **8**(4): 75-81.

Batley J, Edwards D 2007. SNP applications in plants. *Association mapping in plants*: Springer, 95-102.

Beaudoin GA, Facchini PJ. 2014. Benzylisoquinoline alkaloid biosynthesis in opium poppy. *Planta* **240**(1): 19-32.

Berg T. 2009. Landraces and folk varieties: a conceptual reappraisal of terminology. *Euphytica* **166**(3): 423-430.

Bernáth J, Németh É 2009. Poppy. *Oil crops*: Springer, 449-468.

Börner A. 2006. Preservation of plant genetic resources in the biotechnology era. *Biotechnology Journal: Healthcare Nutrition Technology* **1**(12): 1393-1404.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**(19): 2633-2635.

Brezinova B, Macak M, Eftimova J. 2009. The morphological diversity of selected traits of world collection of poppy genotypes (genus Papaver). *Journal of Central European Agriculture* **10**(2): 183-192.

Carolan JC, Hook IL, Chase MW, Kadereit JW, Hodkinson TR. 2006. Phylogenetics of Papaver and related genera based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnL intron and trnL–F intergenic spacers. *Annals of Botany* **98**(1): 141-155.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: an analysis tool set for population genomics. *Molecular ecology* **22**(11): 3124-3140.

Celik I, Camci H, Kose A, Kosar FC, Doganlar S, Frary A. 2016. Molecular genetic diversity and association mapping of morphine content and agronomic traits in Turkish opium poppy (Papaver somniferum) germplasm. *Molecular Breeding* **36**(4): 46.

Celik I, Gultekin V, Allmer J, Doganlar S, Frary A. 2014. Development of genomic simple sequence repeat markers in opium poppy by next-generation sequencing. *Molecular Breeding* **34**(2): 323-334.

Chaturvedi N, Singh M, Shukla AK, Shasany AK, Shanker K, Lal RK, Khanuja SP. 2014. Comparative analysis of Papaver somniferum genotypes having contrasting latex and alkaloid profiles. *Protoplasma* **251**(4): 857-867.

Chen X, Hagel JM, Chang L, Tucker JE, Shiigi SA, Yelpaala Y, Chen H-Y, Estrada R, Colbeck J, Enquist-Newman M. 2018. A pathogenesis-related 10 protein catalyzes the final step in thebaine biosynthesis. *Nature chemical biology* **14**(7): 738-743.

Chen YH, Gols R, Benrey B. 2015. Crop domestication and its impact on naturally selected trophic interactions. *Annual Review of Entomology* **60**: 35-58.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**(2): 80-92.

Consortium TG. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**(7400): 635.

Dittbrenner A, Mock H, Börner A, Lohwasser U. 2009. Variability of alkaloid content in Papaver somniferum L. *Journal of Applied Botany and Food Quality* **82**(2): 103-107.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources* **4**(2): 359-361.

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE. 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS one* **6**(5): e19379.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular ecology* **14**(8): 2611-2620.

Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E. 1983. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**(4601): 1049-1051.

Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. *PloS one* **9**(2): e90346.

Guo L, Winzer T, Yang X, Li Y, Ning Z, He Z, Teodor R, Lu Y, Bowser TA, Graham IA. 2018. The opium poppy genome and morphinan production. *Science* **362**(6412): 343-347.

Gutzat R, Scheid OM 2020. Preparing chromatin and RNA from rare cell types with fluorescence-activated nuclear sorting (FANS). *Plant Epigenetics and Epigenomics*: Springer, 95-105.

Hagel JM, Facchini PJ. 2010. Dioxygenases catalyze the O-demethylation steps of morphine biosynthesis in opium poppy. *Nature chemical biology* **6**(4): 273-275.

Hammer K. 1981. Problems of Papaver somniferum-classification and some remarks on recently collected European poppy land-races. *Die Kulturpflanze* **29**(1): 287-296.

Hammer K, Fritsch R. 1977. Zur Frage nach der Ursprungsart des Kulturmohns Papaver somniferum L. *Die Kulturpflanze* **25**(1): 113-124.

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution* **35**(2): 518-522.

Hrishi NJ. 1960. Cytogenetical studies on Papaver somniferum L. and Papaver setigerum DC. and their hybrid. *Genetica* **31**(1): 1-130.

INCB/UN. 2021. *Report of the International Narcotics Control Board 2020*: United Nations Publications.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**(14): 1801-1806.

Jesus A, Bonhomme V, Evin A, Ivorra S, Soteras R, Salavert A, Antolin F, Bouby L. 2021. A morphometric approach to track opium poppy domestication. *Scientific reports* **11**(1): 1-11.

Jombart T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**(11): 1403-1405.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**(21): 3070-3071.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**(6): 587-589.

Kritikos PG, Papadaki S. 1967. *The history of the poppy and of opium and their expansion in antiquity in the eastern Mediterranean area*: UN.

Kumar S, Banks TW, Cloutier S. 2012. SNP discovery through next-generation sequencing and its applications. *International journal of plant genomics* **2012**.

Labanca F, Ovesna J, Milella L. 2018. Papaver somniferum L. taxonomy, uses and new insight in poppy alkaloid pathways. *Phytochemistry reviews* **17**(4): 853-871.

Lahiri R, Lal R, Srivastava N, Shanker K. 2018. Genetic variability and diversity in Indian germplasm of opium poppy (Papaver somniferum L.). *Journal of applied research on medicinal and aromatic plants* **8**: 41-46.

Lane AK, Augustin MM, Ayyampalayam S, Plant A, Gleissberg S, Di Stilio VS, Depamphilis CW, Wong GK-S, Kutchan TM, Leebens-Mack JH. 2018. Phylogenomic analysis of Ranunculales resolves branching events across the order. *Botanical Journal of the Linnean Society* **187**(2): 157-166.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**(W1): W256-W259.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li Q, Ramasamy S, Singh P, Hagel JM, Dunemann SM, Chen X, Chen R, Yu L, Tucker JE, Facchini PJ. 2020. Gene clustering and copy number variation in alkaloid metabolic pathways of opium poppy. *Nature communications* **11**(1): 1-13.

Liu L, Du Y, Shen C, Li R, Lee J, Li P. 2020. The complete chloroplast genome of Papaver setigerum and comparative analyses in Papaveraceae. *Genetics and molecular biology* **43**.

Malik C, Mary T, Grover I. 1979. Cytogenetic studies in Papaver. V. Cytogenetic studies on P. somniferum* P. setigerum hybrids and amphiploids. *Cytologia* **44**(1): 59-69.

Meirmans PG. 2020. genodive version 3.0: Easy-to-use software for the analysis of genetic data of diploids and polyploids. *Molecular Ecology Resources* **20**(4): 1126-1131.

Millgate AG, Pogson BJ, Wilson IW, Kutchan TM, Zenk MH, Gerlach WL, Fist AJ, Larkin PJ. 2004. Morphine-pathway block in top1 poppies. *Nature* **431**(7007): 413-414.

Murray M, Thompson WF. 1980. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* **8**(19): 4321-4326.

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**(1): 268-274.

Pathak S, Lakhwani D, Gupta P, Mishra BK, Shukla S, Asif MH, Trivedi PK. 2013. Comparative transcriptome analysis using high papaverine mutant of Papaver somniferum

reveals pathway and uncharacterized steps of papaverine biosynthesis. *PloS one* **8**(5): e65622.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. 2012. Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one* **7**(5): e37135.

Poland JA, Brown PJ, Sorrells ME, Jannink J-L. 2012. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PloS one* **7**(2): e32253.

Prajapati S, Bajpai S, Singh D, Luthra R, Gupta MM, Kumar S. 2002. Alkaloid profiles of the Indian land races of the opiumpoppy Papaver somniferum L. *Genetic Resources and Crop Evolution* **49**(2): 183-188.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**(2): 945-959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, De Bakker PI, Daly MJ. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics* **81**(3): 559-575.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**(3): 559-575.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**(6): 841-842.

Rosenberg NA. 2004. distruct: a program for the graphical display of population structure. *Molecular Ecology Notes* **4**(1): 137-138.

Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. *Proceedings of the National Academy of Sciences* **104**(suppl 1): 8641-8648.

Salavert A, Martin L, Antolín F, Zazzo A. 2018. The opium poppy in Europe: exploring its origin and dispersal during the Neolithic. *Antiquity* **92**(364).

Salavert A, Zazzo A, Martin L, Antolín F, Gauthier C, Thil F, Tombret O, Bouby L, Manen C, Mineo M. 2020. Direct dating reveals the early history of opium poppy in western Europe. *Scientific reports* **10**(1): 1-10.

Saunders JA, Pedroni MJ, Penrose LD, Fist AJ. 2001. AFLP analysis of opium poppy. *Crop Science* **41**(5): 1596-1601.

Sexton JP, Hangartner SB, Hoffmann AA. 2014. Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution* **68**(1): 1-15.

Shukla S, Yadav HK, Rastogi A, Mishra BK, Singh SP. 2010. Alkaloid diversity in relation to breeding for specifi c alkaloids in opium poppy (Papaver somniferum L.). *Czech Journal of Genetics and Plant Breeding* **46**(4): 164-169.

Singh S, Shukla S, Khanna K, Dixit B, Banerji R. 1998. Variation of major fatty acids in F8 generation of opium poppy (Papaver somniferum× Papaver setigerum) genotypes. *Journal of the Science of Food and Agriculture* **76**(2): 168-172.

Srivastava A, Gupta S, Shanker K, Gupta N, Gupta AK, Lal R. 2020. Genetic diversity in Indian poppy (P. somniferum L.) germplasm using multivariate and SCoT marker analyses. *Industrial Crops and Products* **144**: 112050.

Srivastava S, Lavania U. 1991. Evolutionary DNA variation in Papaver. *Genome* **34**(5): 763-768.

Svoboda P, Vašek J, Vejl P, Ovesná J. 2020. Genetic features of Czech blue poppy (Papaver somniferum L.) revealed by DNA polymorphism. *Czech Journal of Food Sciences* **38**(3): 198-202.

Tamiru-Oli M, Premaratna SD, Gendall AR, Lewsey MG. 2018. Biochemistry, Genetics, and Genomics of Opium Poppy (Papaver somniferum) for Crop Improvement. *Annual Plant Reviews online*: 1177-1219.

Team R 2020. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA, 2020.

Team RC. 2013. R: A language and environment for statistical computing.

Wakhlu A, Bajwa P. 1987. Cytological analysis in embryogenic callus cultures and regenerated plants of Papaver somniferum L.(opium poppy). *Cytologia* **52**(3): 631-638.

Wei T, Simko V 2021. R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.90).

Wickham H. 2016. ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York, USA, pp. 260.

Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, Li Y, Meade F, Teodor R, Vaistij FE. 2012. A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. *Science* **336**(6089): 1704-1708.

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, et al. 2021. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics*.

Zohary D, Hopf M, Weiss E. 2012. *Domestication of Plants in the Old World: The origin and spread of domesticated plants in Southwest Asia, Europe, and the Mediterranean Basin*: Oxford University Press on Demand.

Zumajo-Cardona C, Pabón-Mora N, Ambrose BA. 2018. Duplication and Diversification of REPLUMLESS–A Case Study in the Papaveraceae. *Frontiers in plant science* **9**: 1833.

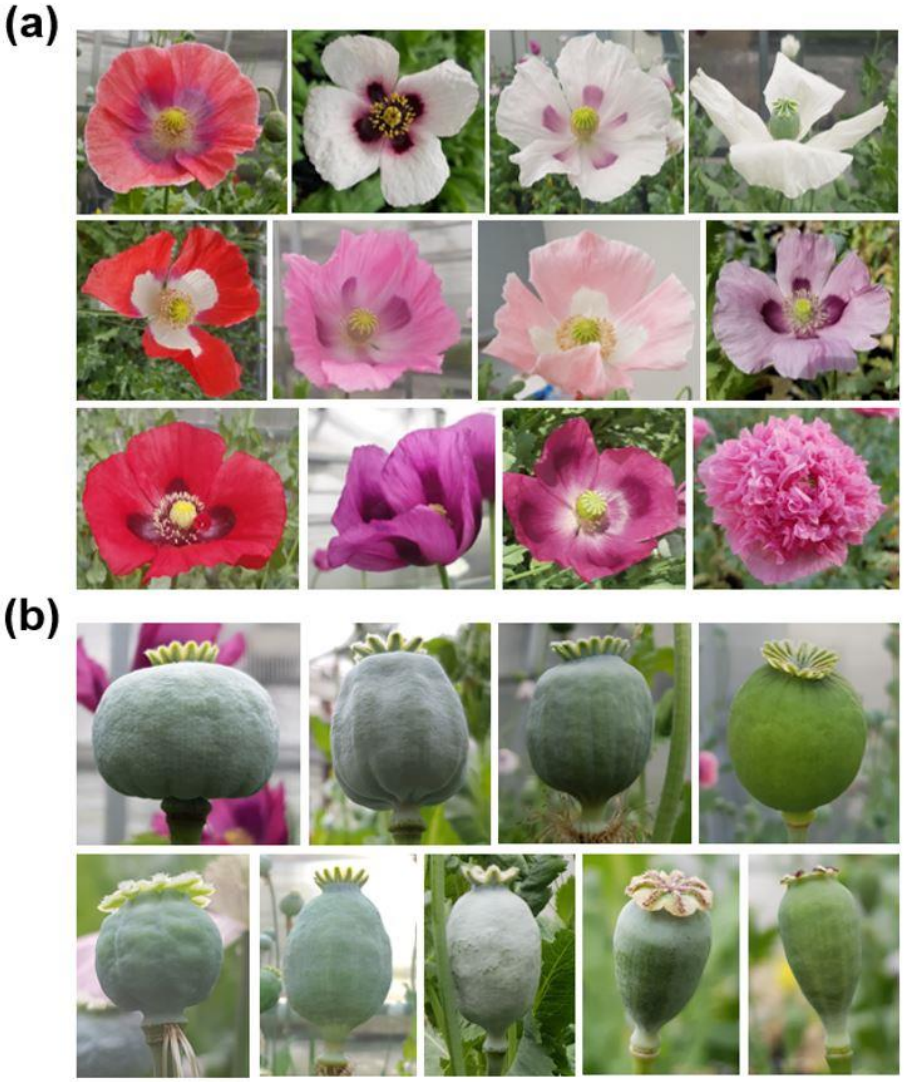

Fig. 1 Morphological diversity among the studied poppy (*Papaver*) accessions. **(a)** Distinct floral morphologies identified in the collection. **(b)** Diversity in capsule shape and size.

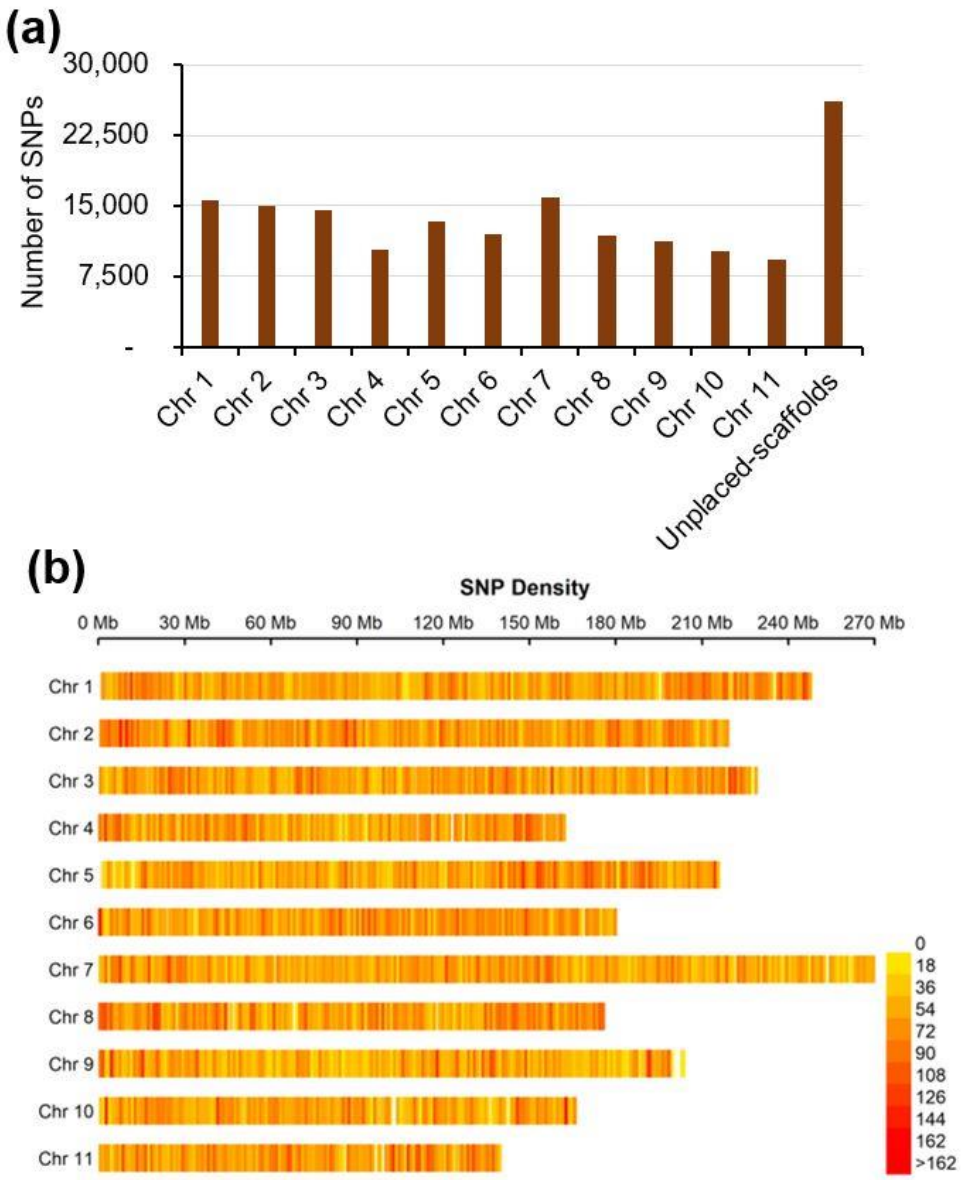

Fig. 2 Distribution and density of poppy (*Papaver*) GBS SNPs across the poppy genome. **(a)** Distribution of the of 165,363 SNP markers identified across the 11 chromosomes and unplaced scaffolds of *P. somniferum*. **(b)** SNP density across the 11 *P. somniferum* chromosomes. Number of SNPs in each 1 Mb non-overlapping window are shown.

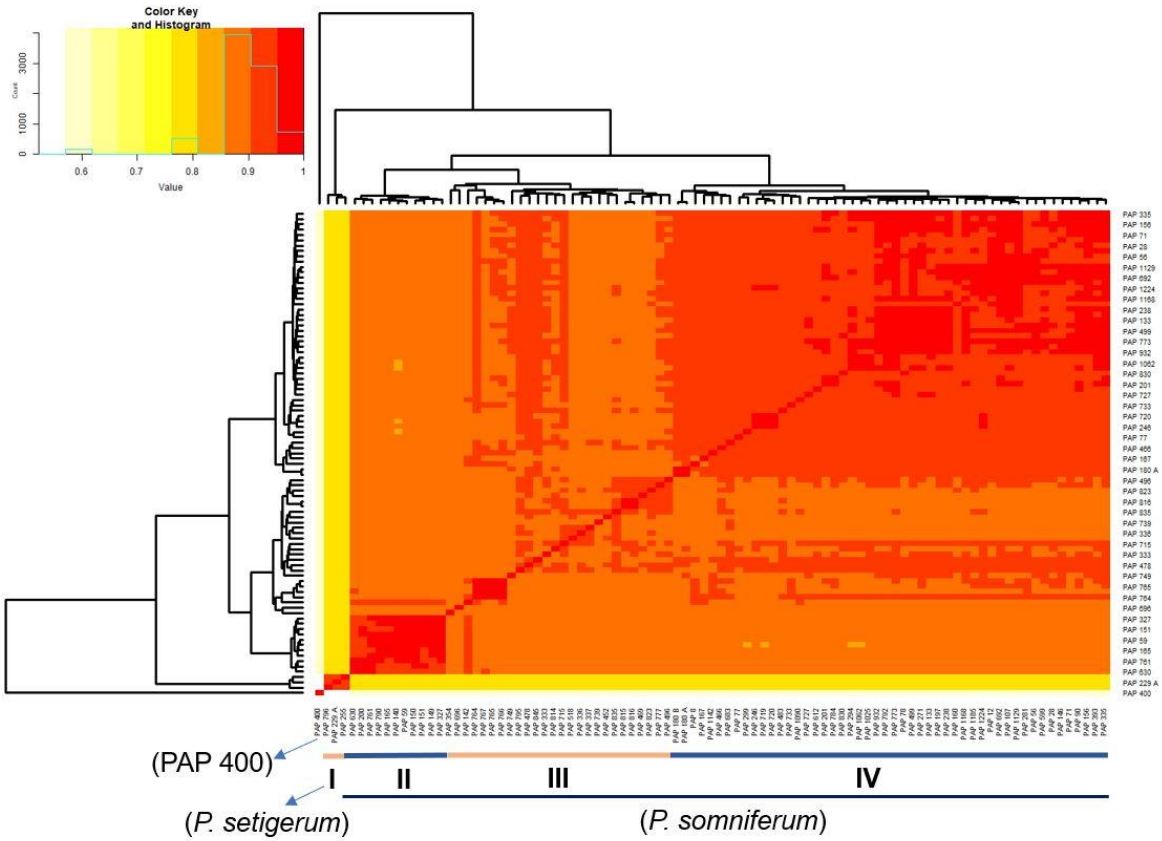

Fig. 3 Hierarchical cluster dendrogram and heatmap based on pairwise identity-by-state (IBS) values from 165,363 SNP data representing the genetic relationships among 91 *Papaver* accessions. The four clusters identified are shown with numbers from I - IV. Degree of relatedness is indicated by colours from light yellow (no relatedness) to red (strong relatedness).

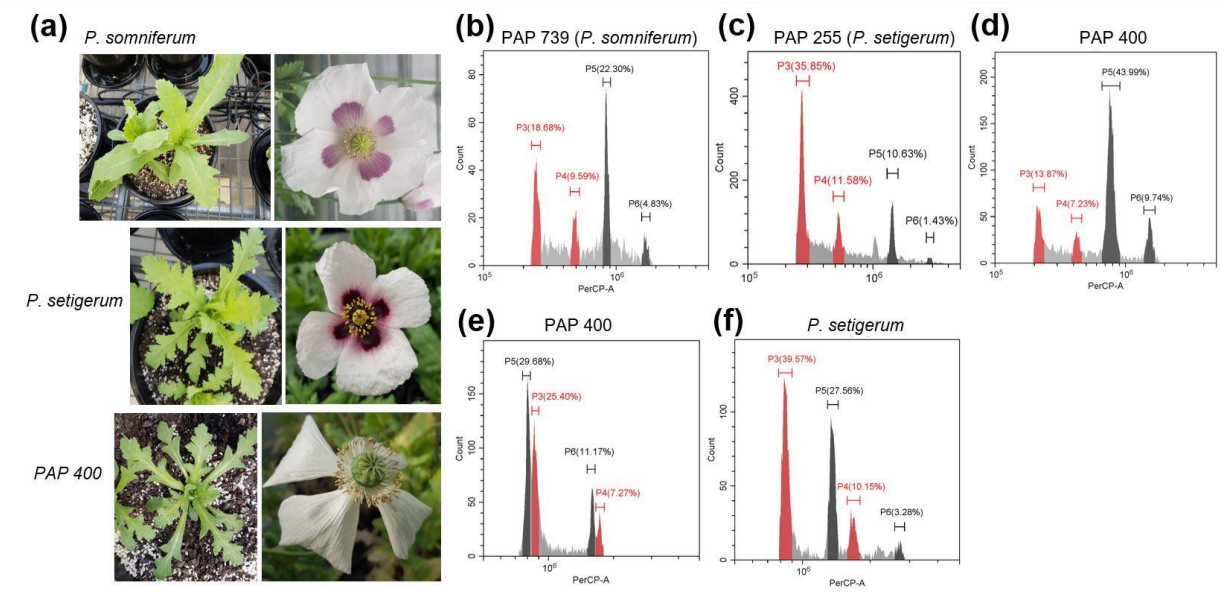

Fig. 4 PAP 400 is morphologically distinct from *P. somniferum* and has a smaller genome. **(a)** Leaf and flower morphology of PAP 400, *P. somniferum* and *P. setigerum*. **(b-d)** Fluorescence histograms of PI-stained nuclei from *Papaver* and tomato (internal reference standard, genome size of ~900 Mb) leaves isolated, stained and analysed simultaneously. G1 and G2 peaks of PAP 400, PAP 739 (*P. somniferum*) and PAP 255 (*P. setigerum*) are shown in black, while G1 and G2 peaks of tomato are shown in red. **(e,f)** Fluorescence histograms of PI-stained nuclei isolated from PAP 400 and *P. setigerum* with PAP 518 (*P. somniferum*) used as internal standard. G1 and G2 peaks of PAP 400 and *P. setigerum* were shown in black. The red histograms represent G1 and G2 peaks of *P. somniferum* (internal standard).

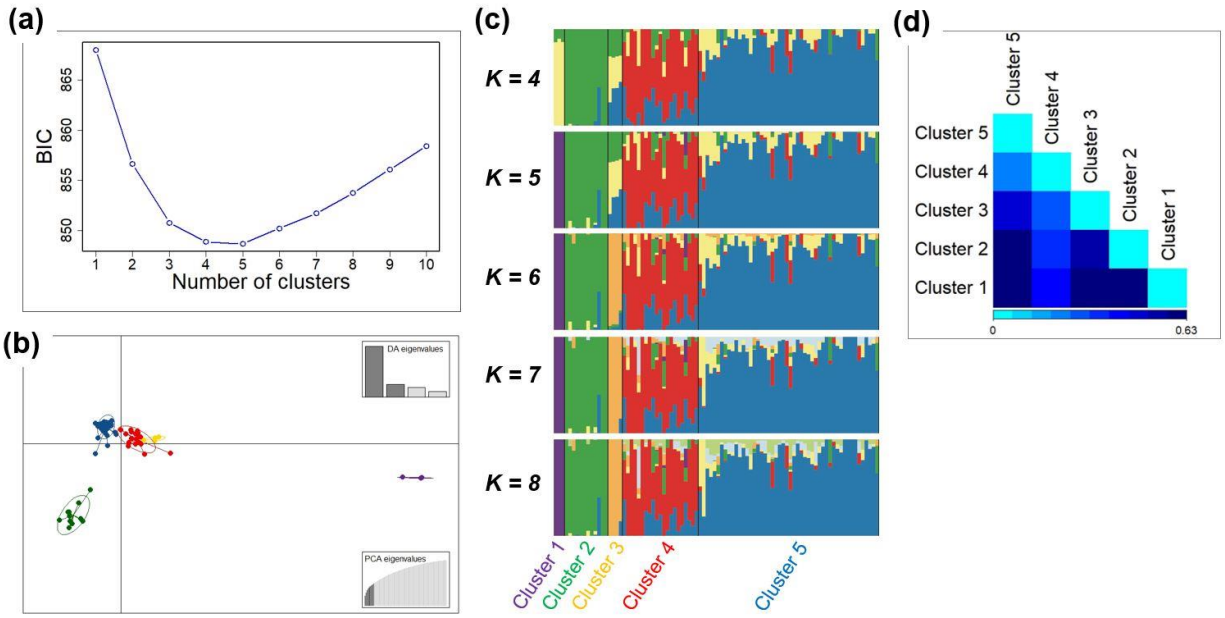

Fig. 5 Population structure and differentiation of 90 *Papaver* accessions. **(a)** Bayesian Information Criterion (BIC) values for different number of clusters (sub-populations) using a dataset of 131,039 SNPs. **(b)** Discriminant analysis of principal components (DAPC) for 90 *Papaver* accessions using a dataset of 131,039 SNPs. Ten PCs and four discriminant eigenvalues were retained during analyses to describe the relationship between the clusters. The axes represent the first two Linear Discriminants (LD). Each circle represents a cluster, and each dot represents an accession. **(c)** Genetic structure of 90 *Papaver* accessions estimated by STRUCTURE using various K values. Each accession is represented by a vertical bar divided into K-coloured segments that represent the accession's estimated common fraction in the K clusters. The accessions are sorted according to DAPC clusters. **(d)** Pairwise population differentiation estimated using F_{ST}.

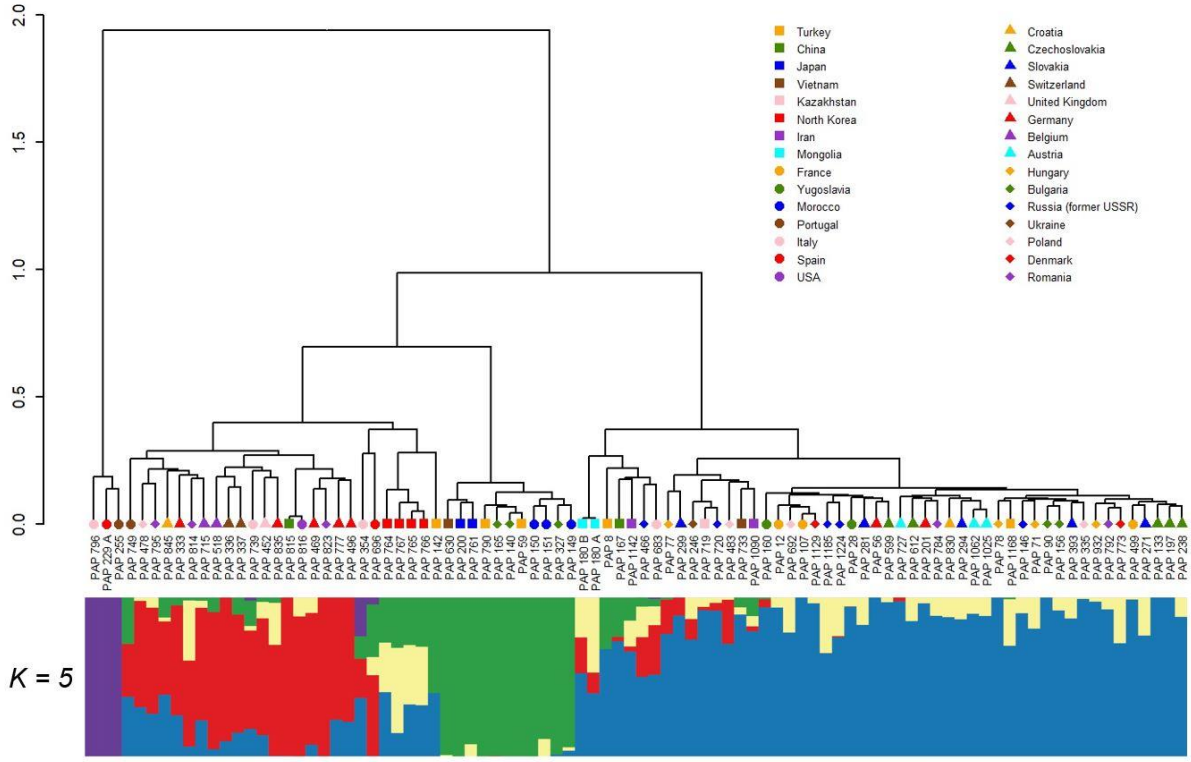

Fig. 6 Genetic diversity and structure of 90 *Papaver* accessions. IBS clustering showing the genetic relationship among the accessions (top panel) and bar plots describing the subpopulations obtained by STRUCTURE at $K = 5$ (lower panel). The accessions are sorted according to IBS clusters. Country of origin of the accession is indicated by differently coloured shapes.

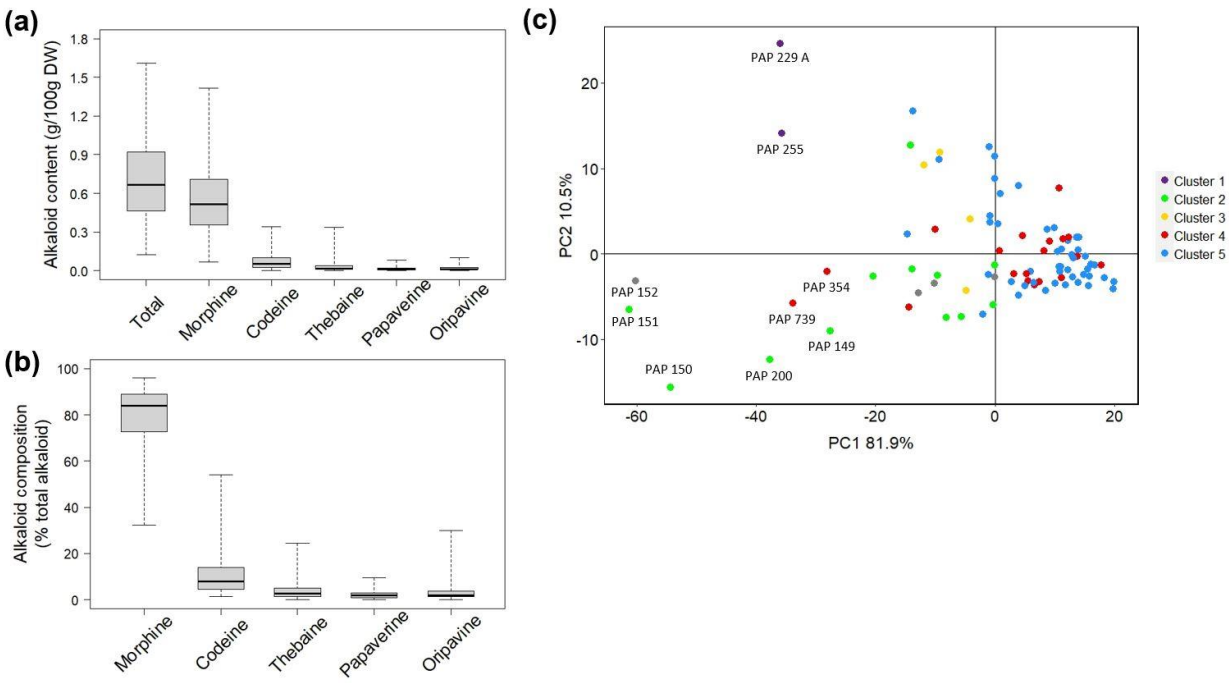

Fig. 7 Diversity in alkaloid **(a)** content and **(b)** composition of poppy capsules. **(c)** Two-dimensional PCA score plot showing the diversity in alkaloid composition. Axes are labelled with the percentage of variation explained by the two PCs. Colours represent the five subpopulations (clusters) identified by STRUCTURE analysis (**Fig. 5**). Accessions for which no GBS data was available are indicated with grey dots. The *P. setigerum* accessions (PAP 229A and PAP 255) and the six accession with the highest codeine levels (PAP 151, PAP 152, PAP 150, PAP 200, PAP 739, PAP 149, and PAP 353) are indicated.

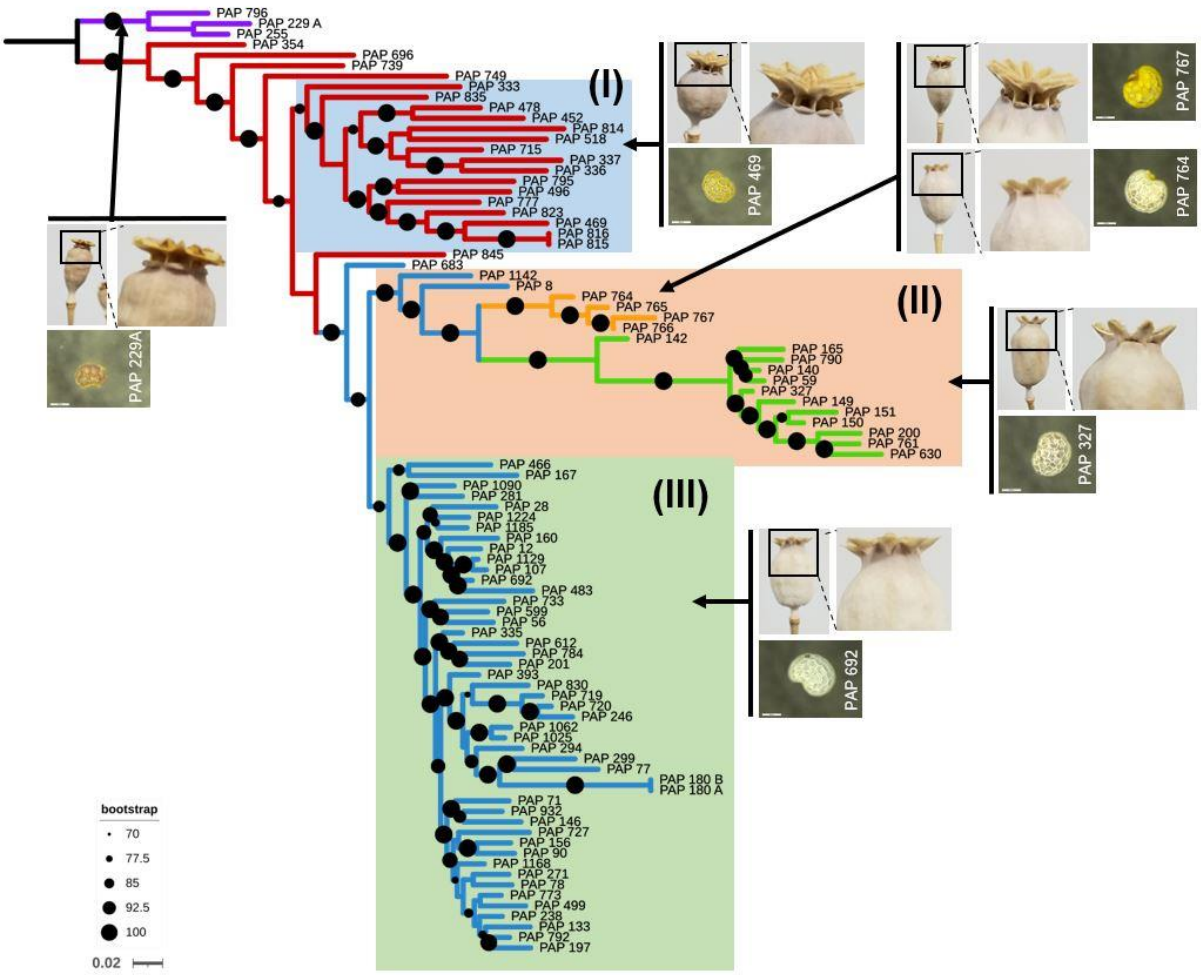

Fig. 8 A maximum-likelihood phylogenetic tree of 90 *Papaver* accessions based on 49,166 GBS SNP markers. Bootstrap supports are shown for nodes over 70% (1,000 replicates) with different sized black circles. *P. setigerum* accessions (purple node, PAP 796, 229A, 255) were used as outgroups. The three clades identified (I, II, III) are shown with boxes in different colours. The five subpopulations (clusters) identified by STRUCTURE and DAPC analyses (**Fig. 5**) are indicated by nodes with different colours (Cluster 1, purple; Cluster 2, green; Cluster 3, orange; Cluster 4, red; Cluster 5, blue). Pictures of capsule and seed morphologies of selected accessions from the different clusters are shown with names of the accessions.

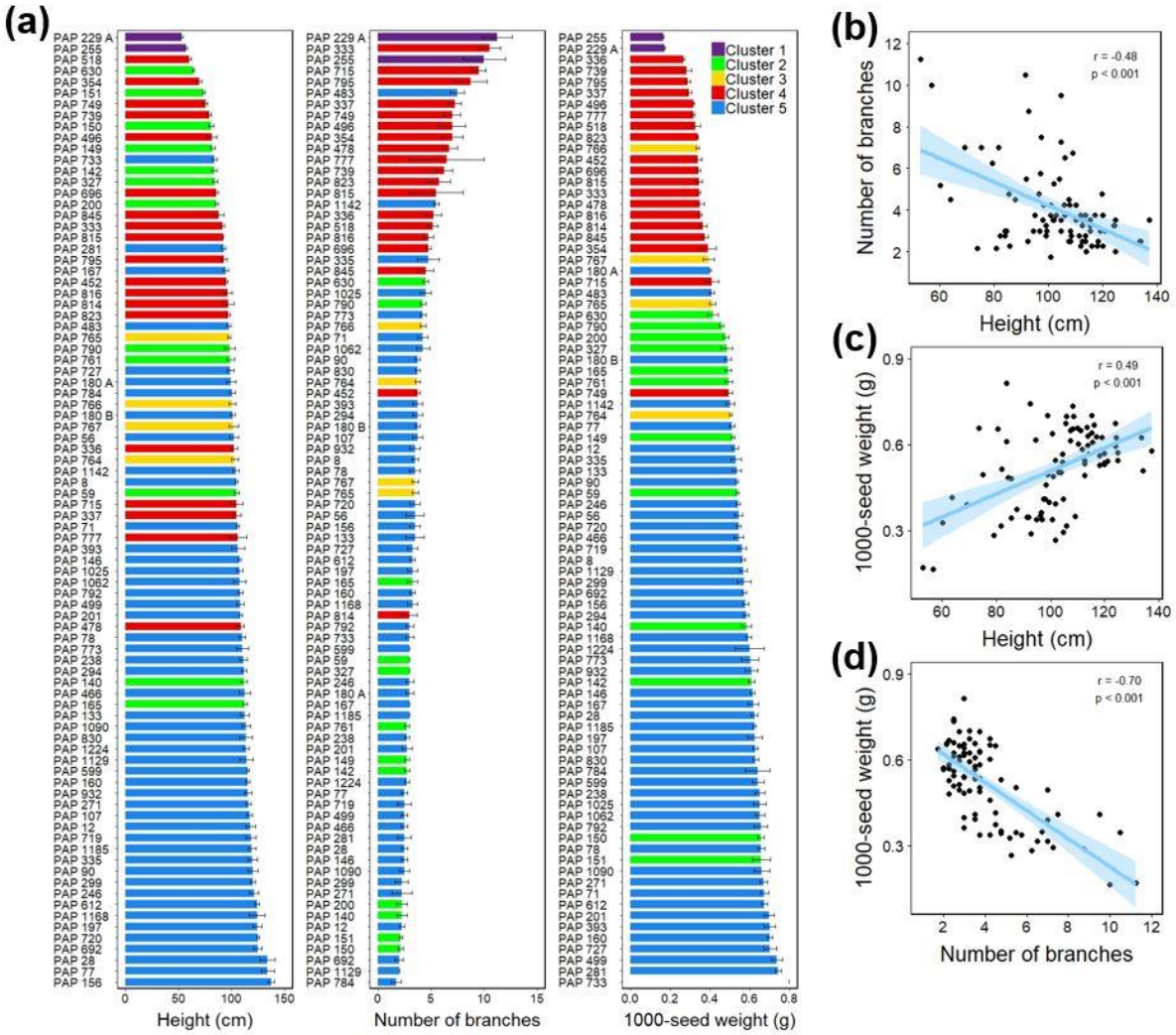

Fig. 9 Variations in selected agronomic-related traits in *Papaver* accessions. (**a**) The diversity in plant height (left), number of branches per plant (middle) and 1000-seed weight (right) among the accession studied. The five subpopulations (clusters) identified by STRUCTURE and DAPC analyses (**Fig. 5**) are indicated by bars with different colours. Data shown are mean ±SE. The accessions were arranged in ascending (height and 1000-seed wight) or descending (number of branches) order of trait values and the order varies for the different traits. (**b-d**) Scatter plots showing the relationships between plant height, number of branches and seed weight. Linear regression lines are shown in blue, and the regions shaded in light blue represent the 95% confidence intervals. Pearson correlation coefficients (*r*) and their significance level (*p*) are shown.