

Genetic diversity of the cultivated vanilla in Madagascar

Rasoamanalina Rivo Onisoa Léa

Mars 2023

Thèse présentée en vue de l'obtention du grade de docteur en sciences
agronomiques et ingénierie biologique

EARTH AND LIFE INSTITUTE

UNIVERSITE CATHOLIQUE DE LOUVAIN

THESIS EXAMINING BOARD

SUPERVISOR: Prof. Pierre Bertin

PRESIDENT OF THE JURY: Prof. Xavier Draye

MEMBERS OF THE JURY: Prof. Mondher El Jaziri

Dr. Khaled Mirzaei

Prof. Renate Wesselingh

“But by the grace of God, I am what I am”

(1Kor. 15:10a)

To my dear parents, my lovely daughter Jemima and husband Daniel

Acknowledgements

I would like to sincerely thank my supervisor, Prof. Pierre BERTIN for welcoming me into his research team, for the confidence, his crucial advice and precious support throughout the whole process of the thesis.

I am deeply grateful to Prof. Mondher EL JAZIRI, who is the pioneer of the present research project PRD Vanille 2016- ARES and made this thesis possible through his intervention and the financial support provided by ARES.

My deep appreciation to Dr. Khaled MIRZAEI for his assistance in the lab work and data processing. Thank you for the moments of fun, the teas, and sorry for your broken leg because of my samples!!!

I am very thankful to Angel Raphael Ramirez RAMIREZ for his kind and interesting discussions and exchanges during the data processing.

I would like to thank Prof. Xavier DRAYE and Prof. Renate WISSELINGH, as well as Prof. Mondher EL JAZIRI and Dr. Khaled MIRZAEI, for agreeing to serve on the jury and reading and improving the manuscript.

I acknowledge the contribution of the companies Takasago and Rama Export for their valuable support in making possible the data collection, by providing expert people on the site of studies, allowing us to get access easily to farmers and vanilla plantations.

My thanks also go to Prof. Tsiry RASAMIRAVAKA and Dr. Christian RABEMANTSOA for hosting the DNA extraction in Antananarivo in their labs, and Dr. Aro Vonjy RAMAROSANDRATANA for facilitating some administrative processes and sharing ideas.

I would also like to thank Grazella ANDRIANAMBININA for contributing to the fieldwork within the research ethics.

I would like to express my sincere gratitude to ARES-CCD for providing financial support for my PhD program, within the project “PRD – Vanille Madagascar, 2016”. This support has been instrumental in enabling me to

Acknowledgements

pursue my research goals and to contribute to the advancement of knowledge in my field.

Many thanks to the PEPA team for being around during those several years. You were always there during the sad and happy moments and always helped me to feel like at my home country!!!

I am deeply grateful to my mother and father (deceased). What I have become now is from your infinite love. Thank you so much, mom, for helping me with my data checking for several days and nights, for accompanying me in the field, and staying awake with me for several nights of DNA extraction. Special thanks to my sisters, brother, friend Mino, family in Low, and my second family in Antalaha for their love, prayers, and support. Finally, my deep appreciation goes to my lovely daughter and husband for their support, understanding, and huge patience during these years.

Summary

Natural vanillin extracted from vanilla pods is the most common flavour in the world, mainly used for foods, drinks and also in perfumery and cosmetics. The producing countries are far from able to meet the growing demand. Vanilla production is limited to some tropical areas such as Madagascar. More than 80,000 Malagasy families rely directly on vanilla production, which constitutes the major part of exportations of the country. However, the usual position of Madagascar as the “head” of providers in terms of volume and quality is currently threatened due to several issues, the most frequent being instability in quantity and quality of the final product. Fungal attacks have been present in the plantations since before 1932 causing a loss of about one fifth of the production in the 1950s. A breeding program (from 1944 to 2000) has offered promise for a sustainable vanilla crop by demonstrating the possibility of creating resistant vanilla varieties. Unfortunately, the program has been abandoned and since then, neither genetic study nor agronomical improvement has been conducted on the Malagasy germplasm. The genetic diversity of the most cultivated vanilla species – *Vanilla planifolia* – has been previously reported to be very narrow due to its restricted initial diversity from a handful Mexican cuttings used as unique genetic basis in all areas of introduction. The classical markers such as isozymes, RAPD, RFLP, microsatellites; previously used did not successfully explain the variation within *V. planifolia*.

The aim of the present work was first to evaluate the current genetic resources of the unexplored cultivated vanilla in Madagascar, in the attempt of supporting future conservation and breeding programs. Secondly, in order to address the knowledge gap about vanilla, the study sought to understand the contribution of space and environment parameters in the genetic variation of the crop. For these purposes, we used the single nucleotide polymorphism (SNP) markers, a high throughput and the most abundant marker across genomes, successfully used in several crop plants with low diversity and recently used in vanilla.

A large-scale prospect was conducted to cover as much diversity present in Madagascar as possible.

Our sampling included 246 accessions collected from the four major vanilla production regions distributed from north to south and northwest of the country (Manakara, SAVA, Mananara Nord and Ambanja). This sampling covered various environmental conditions for vanilla cultivation in Madagascar. Biological materials were collected from 120 plantations, with 30 vanilla fields from each region, distributed between low, medium, and high altitudes. A total of 46 accessions from a local germplasm collection built from the ancient breeding program was also included. This latter was first used to test the effectiveness of the SNP markers in vanilla genetic study and to choose the genotyping tool that fitted better to vanilla. ddRAD-seq protocol was used to generate DNA sequences data and the three computing tools (Stacks, Bcftools and GATK) most widely used for genotyping were compared through descriptive genetic statistics, population structure and job running time in SNP calling. Then, the genetic diversity and the structure of the Malagasy germplasm was evaluated using the developed pipeline. Lastly, using redundancy analysis, the contribution of space and environment on the SNP variation was computed based on the geographical coordinates and the information about soil, temperature and precipitation of the accession location from QGIS-based data.

In general, the present study showed that SNP is an effective marker that can be used for inter- and intra-species levels of vanilla studies. BCFtools was demonstrated to be appropriate to vanilla SNP genotyping due to its ability to genotype highly heterozygous genomes under a reasonable running time. The developed SNP markers were genotyped with enough depth of coverage and were proportionally distributed among the 14 vanilla chromosomes. The set of targeted accessions were genetically segregated according to their respective taxonomic/phenotypic groups with a number of private alleles in each group. The success of the chosen marker was also underlined by the revealed structuration within species.

In particular, our results revealed a clear genetic structure of the Malagasy germplasm and highlighted original varieties, mainly arising from the past breeding program. The 17,948 developed high-quality SNPs distinguished the 246 accessions from the field into five genetic groups – *V. planifolia*, *V. pompona*, Big Vanilla and two intermediate groups with a similar phenotype intermediate between the two first groups. Two additional genetic groups were identified in the local collection (Vanilla Banane and Tsivaky). Intra-species genetic structuration was also revealed from the current study. *V. planifolia* accessions were structured into 3 major genetic groups with high variability (9,281 SNPs). However, the genetic distance between groups was low as previously reported. *V. pompona* accessions were also grouped according to their provenance in the SAVA region, the genetic differentiation according to altitude was evidenced.

Redundancy analysis detected a low contribution of spatial and environmental parameters to the variation of SNPs. Therefore, many other unincluded parameters may contribute to the diversity of SNPs. Our results highlighted the clear genetic structuration of the cultivated vanilla in Madagascar with the presence of original varieties. The data provided through this study offer valuable information for future breeding and management programs.

Contents

Acknowledgements.....	i
Summary.....	iii
List of figures.....	xi
List of tables.....	xiii
List of acronyms	xv
Introduction.....	1
Justification	1
Research goals.....	3
Outline of the thesis	4
Chapter 1.....	7
Literature review.....	7
1.1 Taxonomy and morphology	7
1.1.1 Taxonomy.....	7
1.1.2 Botanical description.....	9
1.2 Ecology and geographical distribution	11
1.2.1 Ecology of vanilla	11
1.2.2 Origin and distribution	11
1.3 Economical importance of vanilla in the world.....	12
1.4 Vanilla in Madagascar.....	12
1.4.1 Economical importance.....	12
1.4.2 Evolution history.....	14
1.4.3 Vanilla cultivation practices	17
1.4.4 Production constraints	20
1.5 The aromatic vanilla	21
1.6 Genetic diversity	22
1.6.1 The crucial role of genetic diversity.....	22
1.6.2 Genetic diversity analysis approaches.....	24

1.6.3 State of knowledge on the genetics of <i>Vanilla</i>	31
Chapter 2.....	39
Diversity and structure assessment of the genetic resources in a germplasm collection from a vanilla breeding program in Madagascar	39
Abstract	39
2.1 Introduction	40
2.2 Materials and methods	42
2.2.1 Plant material and datasets	42
2.2.2 DNA isolation, library preparation and sequencing.....	42
2.2.3 Quality assessment, cleaning, and demultiplexing.....	44
2.2.4 Alignment to the reference genome	44
2.2.5 SNP calling and filtering.....	44
2.2.6 Basic statistics of genetic variation and population structure	46
2.3 Results	46
2.3.1 ddRAD sequencing, cleaning, and sequencing.....	46
2.3.2 Comparison between Stacks, BCFtools and GATK	47
2.3.3 Common SNP discovery	47
2.3.4 Genome-wide genetic variation	48
2.3.5 Population structure and genetic relationships.....	48
2.4 Discussion.....	52
2.5 Conclusions	55
Chapter 3.....	59
Genome-wide genetic variation and population structure in the cultivated vanilla of Madagascar revealed by ddRAD-seq	59
Abstract	59
3.1 Introduction	60
3.2 Materials and methods	63
3.2.1 Plant material.....	63

3.2.2 Genomic data.....	66
3.2.3 RAD analysis	67
3.2.4 Population structure analysis.....	67
3.2.5 Genetic diversity and population differentiation.....	68
3.2.6 Functional annotation of SNP associated genes.....	68
3.3 Results	69
3.3.1 ddRAD sequencing.....	69
3.3.2 Genetic structure of the vanilla population in Madagascar	70
3.3.3 Geographical distribution of species and cultivars	75
3.3.4 Genetic diversity and differentiation among vanilla genetic groups revealed by PCA and UPGMA	75
3.3.5 Functional annotation	77
3.4 Discussion.....	79
3.4.1 ddRAD-seq, a suitable tool for vanilla genetic resources characterization	79
3.4.2 Vanilla population in Madagascar is segregated into six genetic groups.....	80
3.4.3 High diversity within the genetic groups of vanilla cultivated in Madagascar	84
3.5 Conclusions	85
Chapter 4.....	89
Low influence of environment and spatial distribution on genetic variability of cultivated vanilla in Madagascar	89
Abstract	89
4.1 Introduction	90
4.2 Materials and methods	93
4.2.1 Genetic data.....	93
4.2.2 Spatial data	96

Contents

4.2.3 Environmental data	97
4.2.4 Disentangling the contribution of environment and geography to genetic variation	97
4.2.5 Contribution of environmental variables to SNP variation	98
4.3 Results	98
4.3.1 Genetic data	98
4.3.2 Spatial and environmental data according to sampling sites	101
4.3.3 Contribution of individual explanatory variables in structuring genetic divergence	102
4.3.4 Range of variation of the significant environmental variables among the genetic groups	106
4.3.5 Variance partitioning by RDA	107
4.4 Discussion	108
Chapter 5	113
General discussion, conclusions and Perspectives	113
5.1 General discussion and conclusions	113
5.1.1 SNP generated from ddRAD-sequencing technology – useful markers for the genetic structuration of cultivated vanilla	114
5.1.2 Genetic resources in Madagascar have shown high genetic structuration with the occurrence of some original genotypes	116
5.1.3 SNP markers reveal intra species genetic structuration within <i>Vanilla</i>	118
5.2. Research contribution	119
5.3 Perspectives	120
References	121
Appendix A	153
Appendix B	159

List of figures

1.1. <i>V. planifolia</i> plant, <i>V. planifolia</i> flower and longitudinal section of the flower.....	10
1.2. Evolution of the exported vanilla volumes from 1961 to 2020 from Madagascar, Indonesia, and globally	13
2.1. Diagram of final library.....	44
2.2. Density distribution generated from ddRAD sequencing of the 23,701 common loci detected by Stacks, BCFtools, and GATK.....	48
2.3. First and second principal components of the PCA of the 45 vanilla samples with the filtered SNPs from Stacks, BCFtools , GATK and the common SNPs detected by all three protocols together.....	49
2.4. Genetic structure of the 45 cultivated vanilla accessions revealed by ADMIXTURE	50
2.5. UPGMA tree of the 45 individuals from the common loci detected by Stacks, BCFtools, and GATK together.....	52
3.1. Sampling distribution of the 272 studied vanilla accessions.	66
3.2. Distribution of density of SNPs by SNP coverage depth.	69
3.3. Principal Component Analysis (PCA) of the 246 accessions.	71
3.4. Dendrogram generated by UPGMA cluster analysis of the 246 accessions.....	72
3.5. Population structure of the 246 vanilla samples based on the 17,948 SNP markers (ADMIXTURE).	73,74
3.6. Genomic location of the 17.948 SNPs.	78
3.7. Top-BLAST hits species distribution.	79
3.8. Distribution of Gene Ontology (GO) classification of missense SNPs flanking regions.....	79
4.1. Sampling distribution of the 208 studied vanilla accessions from farmer fields	95
4.2. Principal Component Analysis (PCA) of the 208 samples taken from farmer fields according to the two first eigenvectors	99
4.3. Genetic structure of the 154 accessions of the <i>V. planifolia</i> using the UPGMA	100
4.4. Principal Component Analysis (PCA) of 208 accessions.....	101

List of figures

- 4.5.** Biplot of the two first RDA axes (dotted lines) with the 208 pooled accessions (a) and 154 *V. planifolia* accessions (b). 104
- 4.6.** Variation of the seven environmental variables by regional genetic cluster among the 208 pooled accessions. 107

List of tables

1.1. Description of the documented progenitors from the ancient breeding program.....	16
1.2. Composition of the typical mature green pod in Madagascar.....	21
1.3. Composition of processed vanilla pod.	22
2.1. Descriptive statistics on ddRAD sequencing and the genetic diversity of the 45 studied vanilla accessions from the Rama Export collection.....	47
2.2. Comparison of the three variant callers: Stacks, BCFtools, and GATK	47
3.1. Distribution and morphological group of the 272 vanilla samples.	64
3.2. Morphological description of phenotypic groups.....	65
3.3. Analysis of molecular variance among five clusters detected by PCA and UPGMA analyses.....	75
3.4. Pairwise comparison (Fst) between clusters using the 17,948 SNPs..	76
3.5. Descriptive statistics of genetic diversity within each cluster defined PCA and UPGMA	77
4.1. Distribution and morphological group of the 208 vanilla samples. ..	96
4.2. Pairwise comparison (Fst) between the four zones using the 9,281 genotyped SNPs from <i>V. planifolia</i> accessions.....	100
4.3. Significance of individual environmental variables in SNP variation among the 208 accessions and within the 154 accessions of <i>V. planifolia</i>	103
4.4. Contribution of environmental variables alone, spatial variables alone, spatially constrained environmental variables, and the intersection between both spatial and environmental predictors into SNP variation	108

List of acronyms

AMOVA	Analysis of molecular variance
bp	Base pairs
BAM	Binary alignment map
BWA	Burrows-Wheeler Aligner
CTAB	Cetyltrimethylammonium bromide
CV	Cross-validation
dbMEM	Distance-based Moran's eigenvector maps
dd-RAD	Double digest restriction-site associated DNA
DNA	Deoxyribonucleic acid
DP	Depth
EDTA	Ethylenediaminetetraacetic acid
FAOSTAT	Organization statistical database
GATK	Genome analysis toolkit
GBS	Genotyping by sequencing
GC	Guanine-cytosine
GO	Gene ontology
He	Expected heterozygosity
Ho	Observed heterozygosity
IBS	Identity by state
ISSR	Inter simple sequence repeats
HPC	High performance computing
MAEP	Ministry of Agriculture, livestock and fisheries
Min-maf	Minimum minor allele frequency
NGS	Next-generation sequencing
Pa	Private allele
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
Q	Read quality
RAPD	Randomly amplified polymorphic DNA
RDA	Redundancy analysis
REs	Restriction enzymes

List of acronyms

RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RRS	Reduced representation sequencing
SAM	Sequence Alignment Map
SAVA	Region including four districts – Sambava, Antalaha, Vohemara, and Andapa
SD	Standard deviation
SNP	Single nucleotide polymorphisms
SRSL	Short-read library sequencing
TE	Tris-EDTA buffer
UPGMA	Unweighted pair group method with arithmetic mean
V.	<i>Vanilla</i>
VCF	Variant call format
VEP	Ensembl variant effect predictor
WGS	Whole genome sequencing

Introduction

Natural vanillin is the most common flavour extracted from vanilla pods, mainly used to flavour foods and drinks, and is widely used in perfumery and cosmetics. The producing countries are far from able to meet the growing demand. Alternatives such as synthetic vanilla, produced by biotechnological and chemical processing, have been developed (Banerjee & Chattopadhyay, 2019). Natural vanilla production is limited to some tropical areas, among which Madagascar still dominates the global production. However, the usual label of Madagascar as the “head” of providers in terms of volume and quality is currently threatened due to several issues, the most frequent being instability in quantity and quality. In addition, fungal attacks have been present in the plantations since before 1932 (Dequaire, 1976). One fifth of the production was recorded to have been lost in the 1950s due to *Fusarium oxysporum* f. sp. *radicis-vanillae* (Tonier, 1960), and plantations collapsing are currently detected in some regions. Natural vanilla is labelled as “bio” and no chemical product is allowed for its production. Apparently, effective biological treatments against *F. oxysporum* stem and root rot do not yet exist. The ancient vanilla breeding effort in Madagascar (1944-2000; Grisoni & Nany, 2021) offers promise for a sustainable vanilla crop by demonstrating that introgression of resistance genes is feasible. Unfortunately, the program has been abandoned in late of 2000, since then no genetic study or improvement has been conducted. Particularly, the understanding of the genetic resources present in Madagascar is limited. For this reason, we focus the present study on evaluating vanilla genetic resources in Madagascar.

Justification

The international vanilla market is usually largely dominated by Madagascar. Vanilla is the most important crop in this country, bringing 25% of its total income from the whole exportation value and 5-7% of its gross domestic product (Ministry of Industry, Trade, and Handwork, 2019-2021). The production is entirely carried out by small growers. More than 80,000 families rely directly on vanilla production, and the vanilla industry

INTRODUCTION

generates a variety of indirect income streams, such as transportation, tourism, and trade.

Vanilla flavour is extracted from the commercial vanilla pods – *Vanilla planifolia* Andrews – that belongs to the genus *Vanilla* and the *Orchidaceae* family. About 130 vanilla species – mostly wild – have been recorded, among which 30 have been reported to wear aromatic fruits (Soto-Arenas, 1999). In addition to *V. planifolia*, *V. X tahitensis* is also cultivated for commercial interest, though much less significantly. Other species, such as *V. pompona* and *V. odorata*, exist within vanilla plantations or are harvested in the wild but are not economically important. *V. planifolia* is the only species grown on a large scale in Madagascar. However, other phenotypes uncharacterized genetically exist in the field.

Genetic diversity within the cultivated vanilla is narrow. Cultivated vanilla around the world arose from the domestication of a unique and small population of *V. planifolia* from Mesoamerica (Bory et al., 2008a). Because of this domestication and diffusion process, which is exclusively done by vegetative propagation, commercial vanilla has a very limited genetic diversity despite that assisted sexual propagation is the rule for pod production and that crossings are feasible (Bory et al., 2008b; Minoo et al., 2008). For Madagascar, the cultivated vanilla arose from a very small Mexican subpopulation of *V. planifolia* that was transited to Reunion Island around 1819, from where a number of cuttings were first introduced in Madagascar in 1870. Vanilla has always been ordinarily replanted by cuttings in different areas where the dispersal pathway is farmer-to-farmer without any particular organization for genetic diversity management. The limited origin suggests that the genetic diversity of the cultivated vanilla in Madagascar could also be very limited. However, phenotypic variations are detected in the field. These observations need to be scientifically clarified.

From 1944 to 2000, a vanilla breeding program mainly based in Ambohitsara, Antalaha in the north-eastern (SAVA) region has been active (Dequaire 1976; Tonnier, 1960). In view of producing disease-resistant and high vanillin content varieties, several crossing strategies were conducted,

using different progenitors. Thousands of hybrids were created from the crossing between local and imported, wild and cultivated genitors. Presumably, some of the imported and/or created genotypes escaped into the fields. Unfortunately, the program was abandoned after about 50 years due to political issues. Some genotypes have been now rescued in a collection built by a local company (Ramanandraibe Exportation S.A., known as Rama Export). Since the dropout of the breeding program, no vanilla genetic improvement has been applied and the genetic resources from the research centre in Ambohitsara appear to be dispersed or lost. It is therefore urgent to evaluate the genetic diversity present in Madagascar in order to establish conservation strategies and genetic improvement for sustainable vanilla production.

Research goals

The overall objective of this thesis is to provide an overview of the available genetic resources and gather information about the gene pool to establish, for the first time, a genetic database of cultivated vanilla in Madagascar, in order to contribute to the future conservation and agronomical improvement of vanilla in Madagascar. The following investigations were therefore conducted:

- a- Develop an effective molecular pipeline able to detect point mutations (such as SNPs) and adapted to the genetic context of vanilla (low genetic diversity, highly heterozygous);
- b- Provide the genetic characterization of cultivated vanilla genetic resources in Madagascar:
 - build a set of markers that could help to identify the genetic structure of the vanilla germplasm of Madagascar
 - evaluate the genetic diversity rescued *ex situ* from the breeding program efforts in Ambohitsara, Antalaha
 - evaluate the *in situ* genetic diversity by covering the vanilla cultivation areas as large as possible
 - understand the genetic structure of the local germplasm

- c- Quantify the relationship between genetic variation and environmental conditions (ecology and geography) to better understand its adequation with the requirements of vanilla development.

Practically, large prospects across vanilla areas in Madagascar were conducted under different geographic and ecological distributions. The collected samples, including those from the Rama Export collection, were sequenced using the ddRAD approach. The three widely used pipelines for SNP genotyping were compared to choose the one that fits the best for the vanilla genome study and is more practical for later analysis. The high-quality SNPs were then used to study the genetic structure of vanilla in Madagascar. Finally, QGIS-based data were also used to investigate the relationships between SNP variation and eco-geographical parameters.

Outline of the thesis

The **first chapter** of this thesis provides a general background on cultivated vanilla. Its biology, ecology, importance and vanilla cultivation practices is described. In addition, a general context about molecular markers advancement in genetic diversity investigation with their application in vanilla studies is presented.

The **second chapter** reports on an investigation conducted within a local germplasm collection. Beside the study of the genetic diversity of the rescued resources, the samples are also used to compare different pipelines for large-scale data analysis.

The **third chapter** presents the genetic structure of the samples collected from the major areas of vanilla cultivation in Madagascar. The possible origin of the current state of vanilla diversity is discussed in this part. An initiation of functional annotation of the detected SNP related-genes is also presented.

Vanilla cultivation in Madagascar benefits a large distribution under a diversity of environmental conditions. The **fourth chapter** investigates the relationship between the environment, geography, and vanilla genetic diversity.

The **last part** presents an overview of the results and discusses the findings of the current research and will design some perspectives.

Chapter 1

Literature review

1.1 Taxonomy and morphology

1.1.1 Taxonomy

The word “vanilla” is routinely used to designate either the plant itself, pods, or the aroma from the cultivated vanilla. It originated from the Spanish word vainilla or, “little pod” that means a long and slender fruit (Azeez, 2008; Arditti, 2009). The widely cultivated *Vanilla planifolia* was baptized by Andrews in 1808 (*Vanilla planifolia* Jacks. ex Andrews, Fouché & Jouve, 1999). The genus *Vanilla* contains about 130 recognized species and belongs to the largest subfamily Vanilloideae, of the monocot Orchidaceae family. The latest update for the vanilla classification was conducted by Soto-Arenas and Cribb (2010), in which the genus was divided into *Vanilla* (with membranaceous leaves and inflorescence not clearly distinguished from the stem) and *Xanata* (leafy or leafless plants with inflorescence neatly differentiated from the stem) subgenera.

The *Vanilla* subgenus includes Neotropics species which are reorganised into two morphological groups segregated by the petal traits – the *V. mexicana* group (*V. costaricensis*, *V. guianensis*, *V. inodora*, *V. martinezii*, *V. methonica*, *V. mexicana*, *V. oroana* and *V. ovata*) and the *V. parviflora* group (*V. angustipetala*, *V. parviflora*, *V. verrucosa*, *V. bertonensis*, *V. bradei*, *V. edwallii* and *V. organensis*).

The second subgenus *Xanata* contains 2 sections – *Xanata* and *Tethya*. Cultivated and aromatic species are included in the *Xanata* section containing American species (West Indies, Tropical and South Americas). This section includes six morphological groups – *V. palmarum* group (*V. bicolor*, *V. palmarum* and *V. savannarum*), *V. trigonocarpa* group (*V. espondae*, *V. hartii*, *V. sprucei* and *V. trigonocarpa*), *V. planifolia* group (*V. appendiculata*, *V. bahiana*, *V. cristagalli*, *V. denticulata*, *V. dubia*, *V. dungsii*, *V. fimbriata*, *V. helleri*, *V. insignis*, *V. odorata*, *V. phaeantha*, *V. planifolia*, *V. ribeiroi*, *V. schwackeana*, *V. tahitiensis* and *V. uncinata*), *V.*

penicillata group (*V. penicillata*), *V. hostmanii* group (*V. cribbiana*, *V. dressleri*, *V. gardneri*, *V. hostmannii* and *V. ruiziana*) and the *V. pompona* group (*V. calyculata*, *V. chamissonis*, *V. columbiana*, *V. grandiflora*, *V. pompona*, *V. pseudopompona* and *V. vellozii*). The second section *Tethya*, includes species distributed in Asia and Africa. The 50 species within this section are classified into 12 groups – *V. kinabaluensis* group (*V. abundiflora*, *V. giullianetii*, *V. kaniensis*, *V. kempteriana*, *V. kinabaluensis*, *V. klabatiensis*, *V. ovalis*, *V. platyphylla*, *V. ramificans*, *V. siamensis*, *V. seranica*, *V. sumatrana*, *V. utteridgei* and *V. wariensis*), *V. albida* group (*V. albida*, *V. andamanica*, *V. havilandii*, *V. montana*, *V. moonii*, *V. sanjappae* and *V. yersiniana*), *V. annamica* group (*V. annamica* and *V. somai*), *V. griffithii* group (*V. griffithii* and *V. palembanica*), *V. borneensis* group (*V. borneensis* and *V. diabolica*), *V. imperialis* group (*V. grandifolia*, *V. imperialis*, *V. ochyrae* and *V. polylepis*), *V. françoisii* group (*V. françoisii* and *V. coursii*), *V. chalotii* group (*V. chalotii*, *V. nigerica* and *V. seretii*), *V. africana* group (*V. acuminata*, *V. africana*, *V. crenulata*, *V. cucullata*, *V. hallei*, *V. heterolopha*, *V. ramosa* and *V. zanzibarica*), *V. barbellata* group (*V. bakeri*, *V. barbellata*, *V. claviculata*, *V. dilloniana* and *V. poitaei*), *V. aphylla* group (*V. aphylla*, *V. calopogon* and *V. wightii*) and *V. phalaenopsis* group (*V. decaryana*, *V. humblotii*, *V. madagascariensis*, *V. phalaenopsis*, *V. perrieri*, *V. roscheri* and *V. walkeriae*).

Among the hundreds of species of the genus *Vanilla*, only two are regularly cultivated: *V. planifolia* (principal source of commercial vanilla, a Mexican vanilla) and the hybrid species *V. x tahitensis* (from Tahiti). Other species such as *V. pompona* and *V. odorata* can be harvested from the wild or cultivated, but are of no economic importance. Interspecific hybridization can be easily obtained between different wild and cultivated vanilla species. The occurrence of spontaneous hybridization within the genus *Vanilla* has also been documented (Lubinsky et al. 2008). The most investigated case is the cultivated *V. tahitensis*, from which several hypothetical origins have arisen. The presumed recently recognized parents are *V. planifolia* and *V. odorata*. Other successful hybridizations were also reported, for instance, those between *V. claviculata* and *V. barbellata* (Nielsen, 2000); *V. planifolia* and *V. aphylla* (Divakaran et al.,

2006); *V. planifolia* and *V. palmarum* (Li et al., 2020); *V. planifolia* and *V. pompona* (Belanger & Havkin-Frenkel, 2010; Grisoni & Nany, 2021); *V. x tahitensis* x *V. pompona* (Lepers-Andrzejewski et al., 2012). Other successful hybridizations have been documented from the breeding program in Madagascar (Grisoni & Nany, 2021). However, some vanilla hybrids appear to be self-sterile, presumably from parents with different ploidy levels (Grisoni & Nany, 2021). Self or cross-pollination of *V. planifolia* provides beans. However, selfed seeds were reported to provide sterile grains in some instances (Correll, 1948). The characterization of the resulting hybrids is still poor, and even unexplored.

1.1.2 Botanical description

The succulent vine of the commercial *V. planifolia* can reach up to 20 m in length, with a general light green color. The stem is cylindrical in cross section, usually simple, rarely branched, and presents nodes every 5–20 cm from where roots, leaves, and flowers arise. Vanilla has two types of roots – aerial, that anchor to the host plants or other supports for vertical growth; and terrestrial, that go into the soil for water and nutrient uptake. The subsessile leaves are arranged into alternate spirals. The oblong, thick and smooth leaves vary from 10–20 cm in length and 3–7 cm in width. The vanilla plant flowers once a year, and the first bloom appears about three years after planting. Flowers are greenish-yellow and stand on a green pedicel of 4–5 cm. Vanilla shares the typical flower morphology of orchids, with three petals positioned between three sepals. It has two normal petals beside the central one called “labellum” (Figure 1.1). Vanilla is a hermaphroditic plant. However, the male and female parts of the flower are separated by an organ called “rostellum”, which prevents natural pollination.

Flowers on the same bunch appear and bloom subsequently, and the blooming time is short, about 24 hours within which pollination should be conducted. The developmental stages of vanilla beans require 8 to 9 months from fertilization, depending on region. The mature bean of about 10–23 cm with a diameter of 1 cm is green and slightly yellowish, cylindrical with 3 irregular sides. At maturity, the dehiscent fruit releases lipidic grains. The harvesting time varies between regions. For instance, in

Madagascar, it happens between June to August depending on the zone. The mature bean will then follow several steps of the curing process to activate enzymatic reactions for aromatic compound synthesis in order to release the aromatic vanilla bean.

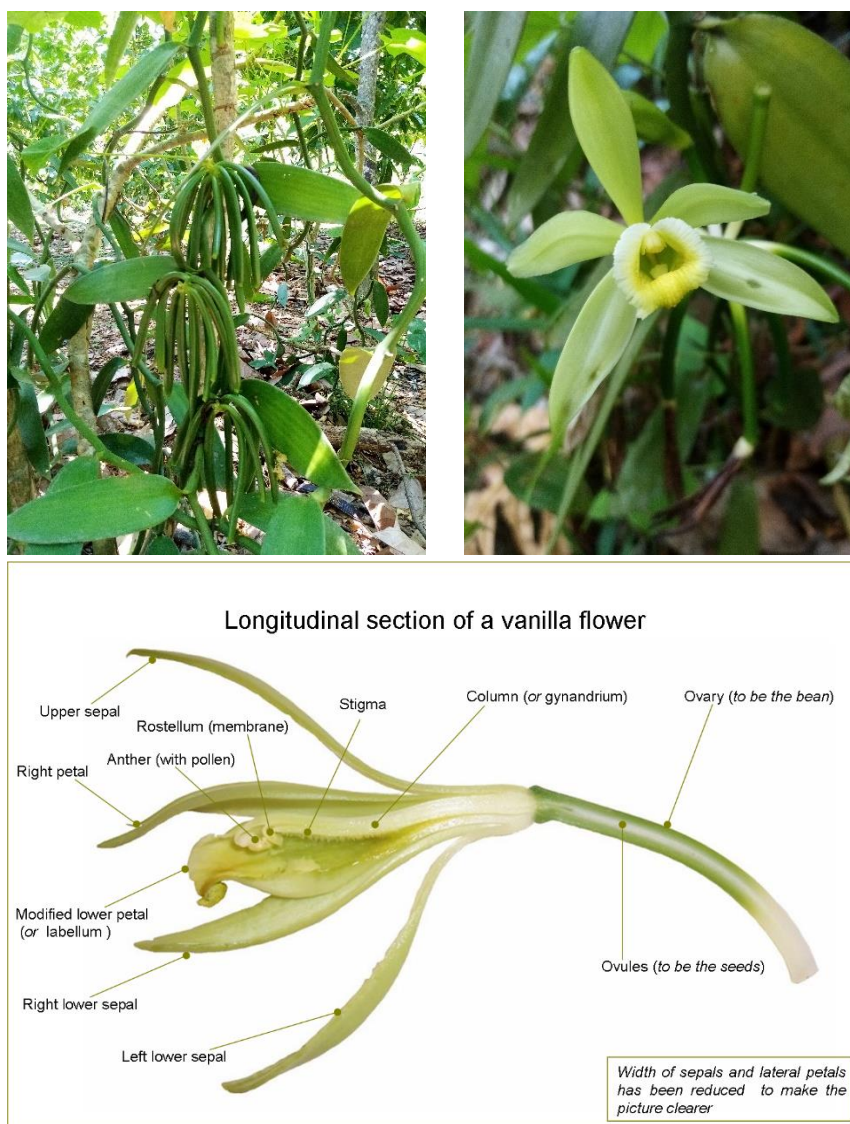


Figure 1.1. *V. planifolia* plant with 2 months of beans (top left), *V. planifolia* flower (top right) and longitudinal section of the flower (bottom, Van & Kerke, 2014).

1.2 Ecology and geographical distribution

1.2.1 Ecology of vanilla

Vanilla has a relatively long-life cycle. Plants from seed germination produce their first flowers by the sixth year (Grisoni & Nany, 2021), against the third in case of vegetative propagation. This does not favour the sexual multiplication of vanilla, especially by farmers.

The optimal conditions for the orchid plant are humid (80%), warm (24–31°C), and partially shady (about 50%) from sea level to 1500 m of altitude. A rainfall of about 150–300 cm per year is ideal for vanilla development (Azeez, 2008). The distribution of rainfall throughout the year is important. An excess of water could favour the development of infections causing stem and root rot, and a lack of water could result in a delay in flowering time and a reduction in the number of flowers. Vanilla can grow on the surface of a variety of well-drained soils, though light, porous, and humus-rich substrates, such as sandy, ferritic, mulch, hay, or plant debris, are preferable.

1.2.2 Origin and distribution

Vanilla cultivation originated in Mexico, where it was grown before the 1500s for local consumption, particularly to flavor drinks made from cocoa and other species (Rao & Ravishankar, 2000). The Aztec conquest by the Spanish marked the beginning of the history of vanilla toward an international use that started in the European countries. In the early 1500s, vanilla beans were introduced for the first time in Europe via Spanish conquistadors (Lubinsky et al., 2010). The growing interest gained by vanilla beans in Europe and the US led to several attempts to grow it in Europe and other regions from the 1700s (Havkin-Frenkel & Belanger, 2018). The lack of pollinators constrained the production (Rao & Ravishankar, 2000). The discovery of hand pollination by Prf. Charles Morren of Liège in 1838 shed light on the possibility of producing vanilla outside its area of origin (Fouché & Jouve, 1999). However, European production was still unsuccessful due to inadequate climatic conditions. The international market only started about 176 years ago, since the effective method of hand pollination was set up in the Reunion Island by

Edmond Albius in 1841 (Bory et al., 2008a). Cuttings were gained in a large geographical distribution in tropical areas of Africa, Asia and South America – Java (1773), Reunion Island (1827), Madagascar (1870), Seychelles (1866), Haiti (1893), Uganda (1912), and India in 1835 (Anandan, 2004).

1.3 Economical importance of vanilla in the world

V. planifolia and the few other vanilla cultivated species are the only edible crop species within the Orchidaceae family. Vanilla is mainly cultivated for the beans. The very popular vanilla flavour is the second most expensive spice after saffron (Banerjee & Chattopadhyay, 2019), with a price that can reach up to 600 USD. The global trade of natural vanilla has been estimated to account for 510 million USD in 2018 (Zion market research, 2022). Vanilla is now largely used in foods, drinks, medicine, pharmacy, perfumes, and cosmetics. Vanilla has also been linked to antimicrobial (Bulpitt, 2005) and antioxidant properties (Rojas-López & Cañizares-Macías, 2013). About 98,000 hectares are currently recorded to be occupied by vanilla plantations (FAOSTAT, 2022), with a total of 32,118 tons of green pods, equal to 8,047 tons of cured vanilla. Madagascar and Indonesia account for the majority of global supply, with 2,975 and 2,306 tons, respectively, in 2020. The remaining parts are produced by other countries such as Mexico (589 t), Papua New Guinea (495 t), China (433 t), Turkey (360 t), Uganda (185 t), and Tonga (180 t). Global natural vanilla production is not sufficient to fill the growing demand in America, Europe, and certain Asian countries such as Japan. The natural vanillin is today facing competition from inexpensive synthetic vanillin such as chemical synthesis from ligneous materials (Gallage & Møller, 2015).

1.4 Vanilla in Madagascar

1.4.1 Economical importance

The global vanilla market has usually been dominated by Madagascar since long time ago. The history of Madagascar on the global market started

before 1929 with a production of more than 1,000 tons, to reach over 3000 tons in 2013. However, this production fluctuates mainly due to climatic and political issues (Figure 1.2, FAOSTAT). The majority of the cured vanilla goes to France (34.2%), USA (28.8%), Germany (13.4%) and Canada (9.4%) (Ministry of Agriculture and Livestock and Fishing, MAEP 2017). Vanilla is the first export product in value for Madagascar. It accounts for one-fourth of the country's total exports, and for 5 to 7% of GDP (Ministry of Industry, Trade, and Handwork, 2019-2021). During the French colonialism period, vanilla production was focused on the north-east, in the SAVA region. Currently, vanilla cultivation is practiced on the eastern coast from the south to the north-east, and in the north-west of Madagascar on more than 74,800 hectares (FAOSTAT, 2022) including 11 regions from the 22 political subdivisions (Anosy, Atsimo Atsinana, Ihorombe, Haute Matsiatra, Vatovavy, Fito Vinany, Atsinanana, Alaotra Mangoro, Analanjirofo, SAVA, Diana, and Sofia). In Madagascar, the thousands of tons of pods are exclusively produced by small growers. Vanilla involves over 80,000 families that depend directly on the production (growers). In addition to the direct income from the production (from plantation to beans conditioning), post-harvest operations (collection, curing, conditioning), and vanilla play important roles in promoting tourism, transport, and trade, especially during the open period of the local market for vanilla.

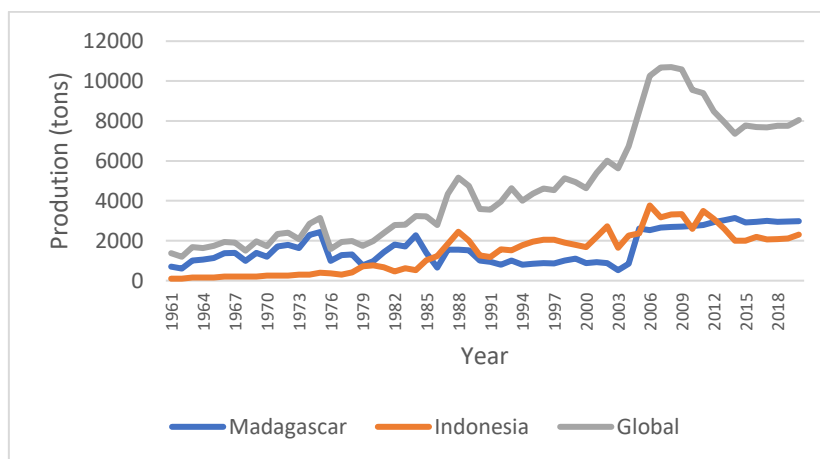


Figure 1.2. Evolution of the exported vanilla volumes from 1961 to 2020 from Madagascar, Indonesia, and globally (data from FAOSTAT, accessed in October 2022).

1.4.2 Evolution history

V. planifolia, which is the widely cultivated vanilla in Madagascar, arose from a few Mexican cuttings that passed through Reunion Island. However, the introduction of vanilla in Madagascar is poorly documented. Historically, French people introduced vanilla cuttings to Madagascar and promoted vanilla production during the colonial period. According to Grisoni et al. (2020), *V. planifolia* was introduced from France to the Indian Ocean in 1822 through the French Reunion Island. The vanilla cuttings spread throughout the surrounding islands several decades after the discovery of hand pollination on Reunion Island in 1841. Vanilla was first introduced in Madagascar in 1870 on Nosy Be, a Malagasy island in the north-west; then in 1880 on Sainte Marie Island, a Malagasy island on the east coast; and finally, in Vatomandry in 1891. The arrival time of cuttings in the SAVA (a region including four districts – Sambava, Antalaha, Vohemara, and Andapa) – the main zone of cultivation –, is unknown. However, the history of vanilla in the SAVA region might be closely related to the colonial period during which regional exploitation politics were established by France (1897–1946). For the better exploitation of each zone, potential and locally adapted crop were specified for each Malagasy region: for example, cacao in the north west, coffee and litchi in the south east, cloves in the middle east, rice in the middle east and middle west, and vanilla in the north east. The SAVA was the first main area of vanilla production in Madagascar, and was for a long time considered as the "capital" of production due to the success of Madagascar on the global market. The vanilla crop is now extensively practiced along the eastern coast of the country and in the north-west where environmental conditions correspond to its requirements. Presumably, introduced vanilla existed in other areas, such as in the south-east, around the same time as in SAVA, but was not exploited due to politics (personal communication from farmers). In addition, the economic seduction of the SAVA due to vanilla production attracted migration from several regions, especially for vanilla-related fieldwork (hand-pollination, gardening...), which contributed to the cutting dispersal across the country. In addition, after the colonial period, vanilla gained many areas, and cuttings moved from farmer to farmer without any particular management. Since then, despite

the success of Madagascar in vanilla production, the vegetatively propagated crop has been constrained by the arrival of the pathogen attacks, mainly *Fusarium oxysporum* f. sp. *radicis-vanillae* (Koyyappurath et al., 2016). In the 1950s, one-fourth of the production was recorded to have been lost from the damage caused by *Fusarium* (Grisoni & Nany, 2021). A pioneering vanilla breeding program was established in 1950 to control the disease (Dequaire, 1976). The program aimed mainly to improve the resistance properties against pathogens of *V. planifolia*, as well as other potential traits such as high vanillin content, indehiscence, early maturation, through interspecific crossings with progenitors bringing the desirable traits. Several imported and local species were used for cross-breeding. Table 1.1 presents the panel of genitors and the crossing configurations reviewed by Grisoni & Nany (2021). Several thousands of genotypes were produced from crossings based on *V. planifolia*. Self-fertilization, simple, double, three-way crossings were conducted. Self-fertile and self-sterile genotypes were produced. Field tests were performed in 10 communes within the SAVA region. Tsitaitra and Magnitry Ampotony were the most important retained varieties.

CHAPTER 1 Literature review

Table 1.1. Description of the documented progenitors from the ancient breeding program in Madagascar (Grisoni & Nany., 2021).

Species	Accession name	Origin	Year of introduction	Main traits
<i>V. abundiflora</i>	VSP093	Colombia (Bogota)		
<i>V. coursii</i>	coursii	Madagascar		
<i>V. crenulata</i> cf	VSP074	Benin (Nirouli)	1952	
<i>V. francoisii</i>	Francoisii	Madagascar		
<i>V. madagasacariensis</i>	madagasacariensis	Madagascar		Drought resistance, hardiness
<i>V. phaeantha</i>	Ph069	West Indies	1951	Resistance to fusariose
<i>V. phaeantha</i> cf	VSP033	France (Auteuil)	1950	Resistant
<i>V. planifolia</i>	F.0	Madagascar	1950	Aromatic fruit
<i>V. planifolia</i>	F1 (Pasteur1)	Madagascar	1943	
<i>V. planifolia</i>	F Antalaha56	Madagascar	1956	
<i>V. pompona</i>	P065	Madagascar	Comoros	
<i>V. pompona</i>	Pompona pure	West Indies	1950	Very resistant
<i>V. pompona</i>	P005	Mauritius		Resistance to fusariosis, vegetative growth
<i>V. pompona</i>	P050	France (Nogent)		
<i>V. pompona</i>	P104	Moye Bampaperu		
<i>V. tahitensis</i>	Tahiti 24	Tahiti		
<i>V. tahitensis</i>	Tahiti Haapape	Tahiti	1950	Indehiscent fruit, earliness
<i>V. walkeriae</i>	Walkeriae	Sri Lanka	1952	
<i>V. zanzibarica</i>	Zanzibarica	Zanzibar	1950	
<i>Vanilla</i> sp.	VSP039	Africa (Mission Jean Fe'lix)		
<i>Vanilla</i> sp.	VSP008	Cameroon (Nyon et Sanaga)		
<i>Vanilla</i> sp.	VSP071	Trinidad		
<i>Vanilla</i> sp.	VSP103	Brazil		
<i>Vanilla</i> sp.	VSP106	Brazil (Pernambuco)		
<i>Vanilla</i> sp.	VSP108	Bolivia via JB Rio de Janeiro	1963	
<i>Vanilla</i> sp.	VSP109	Bolivia via JB Rio de Janeiro	1963	

1.4.3 Vanilla cultivation practices

Vanilla is a perennial vine plant that requires support, or “tutor”, for proper development. Support is usually a living host tree. Dead trees and concretes are sporadically used for intensive production in some countries. Cultivation of vanilla is widely practiced in open fields, although production within shade houses has recently been applied in some countries (Leclercq et al., 2001; Havkin-frenkel & Belanger, 2018; Quiro, 2018; Exley, 2019). The following information relies on our own observation and farmer explanations gathered during our field survey from 2017 to 2020. The overall practice is almost similar across the regions. However, operations are slightly adapted according to the regional environmental conditions. Vanilla production involves five typical steps – land preparation; vine support acquisition; cutting plantation; maintenance; pollination; and harvesting.

Land and tutors

In Madagascar, vanilla is cultivated from the sea level to over 1500 meters, on flat, light, and steep slopes, from the sea to the inland, on the east coast and in the northwest. Vanilla field surfaces are generally less than 1 hectare, usually arising either from a modified primary forest or from a post-agricultural land with or without regenerated secondary forest. Basically, farmers in Madagascar do not practice land preparation activities (such as ploughing and tillage). The most common pre-plantation tasks consist in the adjustment of shady trees and light ground clearing. The used vine tutors are exclusively living trees that play four major roles—supporting the vanilla vine, filtering sunlight, maintaining humidity, and providing humus for the vanilla plant through dead leaves. Depending on the species availability, shady plants and tutor composition vary slightly from one region to another. Natural forest tree species are used in some areas when the plantation is installed from a modified primary forest. However, the majority of vanilla supports are cultivated. Tutors are usually planted before the rainy season, from August to October. After 6 to 12 months, 1.5 m of straight tree branches with a diameter of about 5 cm can support cuttings. *Pachira aquatica* (water chestnut, Guiana chestnut and Malabar chestnut), *Jatropha curcas* (jatropha) and the leguminous *Glirici-*

-dia sepium (gliricidia) are the common cultivated species. They are fast-growing, well adapted for pruning, and improve soil quality (especially the leguminous trees). Usually, rare fruits or forest trees co-exist with cultivated ones, which helps in filtering sunlight. For instance, in the northwest, such as in Ambanja, cacao plants, which is the main crop of the region, are the common shady species with a few other fruit trees (mango, jack fruit, breadfruit) and vanilla climbs usually on the cultivated *Jatropha curcas*, *Pachira aquatica* or rarely *Albizia lebbbeck* (bonara). In Mananara Nord in the north-east, the main vanilla plant supporter is the cultivated *Pachira aquatica*, with some *Jatropha curcas*, *Gliricidia sepium*, and some rare indigenous forest species. The clove trees (*Zyzygium aromaticum*), which are among the main cash crops of the region, are almost always present at a higher level with other species like litchis, jackfruits, and mangos. In the southeast, such in Mananara, vanilla plantations of low altitude (for example, Nosy Varika) are remarkably planted in slightly cleared natural forests of 5–6 m high. A second stratum of high *Combretaceae* trees serves as light filtering. The plantations in the center of Manakara are dominated by cultivated trees such as jatropha, water chestnut and gliricidia, with some natural species and/or fruit trees like litchi and coffee trees (*Coffea canephora*). In SAVA, the most commonly used tutors are *Gliricidia sepium*, *Jatropha curcas*, and *Pachira aquatica*. Likewise in other regions, fruit trees (mango, jackfruit, litchi) are almost always present as an additional sunlight filter. Some plantations in intermediate zones are installed within cleared natural forests. Cultivations within mountains of bamboo forest exist from the middle to high altitudes.

Cuttings

Vanilla has a long history of vegetative propagation despite the occurrence of sexual fertility. There is no particular propagule management for vanilla in Madagascar; cuttings move from farmer to farmer. Typically, cuttings are distributed through familial and social relationships. Inheritance is the classical way in which parents build the first vanilla field for their sons since they are ready after leaving school. However, sharing between housemates is also common in all regions, where that between communes

and districts is rare. For a few years, cutting selling has started to be detected during the market days in several areas. A vanilla “seed” consists of a cutting of about 1.5 m in length. After being removed from a healthy and vigorous mother plant, the vine is left for several days in a humid and shady place. This will give more flexibility to the stem before its training to the tutor and also helps dry the injury from cutting that could provide infection since it is in contact with the soil. The period of plantation is usually after the harvesting period. Three to four nodes without leaves are covered by a light soil or mulch, and the remaining part is attached to the support.

Maintenance

Vanilla is a highly sensitive plant that requires continuous attention for a better yield. Maintenance involves the surroundings of the plant and the plant itself. Herbaceous plants are generally cleaned twice a year (from January to September), and the resulting debris is deposited around the plant. To strengthen the vanilla plant, suspended stems are directed to partially get the soil for a marcotting-like system called "bouclage". Vanilla production is entirely organic. Occasional use of fertilizers is detected in some areas; for example, in Ambanja: dry feces of zebu, pods of cocoa, decayed wood, coconut mulch. Clove leaf waste after essential oil extraction are rarely used in Mananara Nord. Frequent prospect is performed to detect infected vines or fruits and insects. Until now, physical elimination has been the common practice to fight against these enemies.

Hand pollination and harvesting

Vegetative development of vanilla takes about 3–4 years before releasing the first flower. An axillary bud develops into an inflorescence where blooming is subsequent. The longevity of flowers is less than 24 hours. Usually, blooming starts early in the morning and the flower starts to wither in the afternoon. Natural pollinators, *Euglossa* bees, are endemic to Mexico, Central and South America (Bory & al., 2008a). For commercial purposes, vanilla is predominantly hand-pollinated even within its area of origin. Successful pollination should be conducted in the morning. Frequent prospecting is therefore crucial to ensure production. Flowering

time varies slightly between zones, from the middle of August to January. Pollination consists of raising the rostellum (usually by using a wooden needle) that separates the male from the female parts of the flower to put the pollen cluster in contact with the female organ and putting both in contact using a stick. Most farmers choose to limit the per-bunch number of pollinated flowers to 8–12 for better bean quality (size) and also to allow a longer life for the plant. Beans are mature after 8 to 9 months after pollination, depending on the region. The bean is ready for harvesting when its color turns yellowish and the sides are not significant, i.e., they become irregularly cylindrical. The harvested beans will follow several processes of pre-treatment and curing to release the aromatic vanilla bean.

1.4.4 Production constraints

Beside the political, security, and environmental (flooding, cyclones, etc.) issues, the agronomical constraints of vanilla cultivation in Madagascar are mainly the attacks of fungi, viruses, and insects. Among the prospects we conducted, several symptoms were detected in the field, such as stem and root rot, and partial drought on the aerial roots and stem that caused the death of the entire plant and the whole plantation. Such heavy damages were encountered in the middle altitudes of Mananara Nord, where infection waves move field-by-field. The occurrence of insects locally called "Kakamenaloha", "Voapanetroka", "Sababaka", and "Besoroka" was also detected in several areas. As reported by growers, these insects release a substance that causes the drop of concerned flowers and buds. Some insect larvae develop inside the vanilla stem and cause a partial death of the plant. Premature dropping of vanilla beans has been also a common issue in different regions. Chemical pesticides are never used. A few sanitary strategies have been adopted by growers, such as removal of infected plants, physical elimination of insects and larvae, agroecological system (plantation of chillies, rarely practiced), use of an organic mix of pesticides (chilis, sisal, ash, etc.), and the control of cutting movement from one zone to another (case of Manakara). The extent of the issues is higher in Mananara Nord and the SAVA regions compared to that of Ambanja and Manakara. Phytopathology is beyond the scope of our study,

although further studies of these threats are highly suggested. As shown by the former breeding program in Madagascar, genetic improvement is one of the possible approaches to overcome these issues.

1.5 The aromatic vanilla

Most of the green vanilla pods, especially the commercial *V. planifolia*, are flavorless before the curing processes. Through this latter which natural enzymatic reactions are activated to release a broad range of aroma and flavor components. Table 1.2 presents the typical contents of 100g of mature bean of the widely commercial *V. planifolia* produced in Madagascar (Odoux & Brillouet, 2009).

Table 1.2. Composition of the typical mature green pod in Madagascar (Odoux & Brillouet, 2009).

Fraction	g/100g	Components (%)
Fiber	40	Lignin (62), cellulose (27), hemicellulose (11)
Lipids	12	Linoleic acid (54), oleic acid (10), palmitic acid (10)
Sugars	10	Sucrose (80), glucose (15), fructose (5)
Guclovanillin	10	
Minerals	10	Potassium (28), calcium (10), magnesium (2)
Organic acids	5	Citric (50), malic (30)
Proteins	3	

The post-harvest treatment of vanilla beans varies from one country to another; artisanal practices may even slightly differ among different areas of the same region. The skills and operations in the curing process often influence the composition of the cured vanilla pod beside the genetic nature of the plant, the environmental influences and the maturity stage of the harvested bean. The curing process contains four major common steps – killing, sweating, drying, and conditioning. The killing phase consists of stopping the vegetative growth of the living pod and disrupting the cell structure to allow the initiation of enzymatic and non-enzymatic reactions. The purpose of pod killing is to put the flavor precursors in contact with the enzyme catalyzers of the conversion reactions. Pod killing is usually performed under sun exposure, in an oven, or by quickly dipping into hot water (Frenkel et al., 2010). Sweating is among the most crucial steps of a

post-harvesting operation that aims to find the optimal conditions for the enzymatic reactions through which precursor molecules are translated into vanillin and related compounds. At the end of sweating, the pod color becomes dark brown due to the oxidation of polyphenolic compounds. During drying, the moisture of the bean is reduced from 600-700g/kg to approximately 250-300g/kg. The last step – conditioning – consists of storing the cured beans inside rolled paper or closed boxes. During this phase, numerous chemical and biological processes, including oxidative degradation, esterification, and etherification, produce different volatile fragrance components (Frenkel et al., 2010).

The cured vanilla pod contains several components, including volatile and non-volatile constituents. Natural vanillin (4-hydroxy-3-methoxybenzaldehyde) is the main chemical flavor component extracted from vanilla beans and widely used as a flavor (Walton et al., 2003). However, the vanillin pathway is still at the heart of debate among vanilla researchers. Several hypotheses have been raised and no conclusive pathway has been determined (Dixon, 2010; Gallage et al., 2014; Yang et al., 2017). Genes involved in the aromatic compounds are still miss explored (Rao et al., 2014). Table 1.3 shows the main components of cured vanilla pods (Rao et al., 2014; Solano-De La Cruz et al., 2019; Li et al., 2022).

Table 1.3. *Composition of processed vanilla pod (Rao & Ravishankar, 2000).*

Component	g/kg
Vanillin	20
Vanillic acid p	1
<i>p</i> -Hydroxybenzaldehyde	2
<i>p</i> -Hydroxybenzyl methyl ether	0.2
Sugars	250
Fat	150
Cellulose	150-300
Minerals	60
Water	350

1.6 Genetic diversity

1.6.1 The crucial role of genetic diversity

Genetic diversity refers to all genetic variation across genotypes within or between species (Bhanu, 2017). Variations in morphology and physiology

are mostly caused by genetic polymorphism. Genetic diversity allows populations or species to cope with environmental changes across an evolutionary time span (Freeman & Herron, 2007). For agricultural species, investigating genetic diversity is crucial for applied purposes such as conservation and breeding programs. Data related to diversity will help to monitor and assess the population and then set efficient strategies for genetic resource conservation. Ensuring sustainable agriculture is one of the direct downstream benefits of conservation. Genetic diversity is the basis for any crop improvement that allows breeders to create novel and enhanced cultivars with desirable properties through selection (Swarup et al., 2021). Beside the direct commercial purposes (nutritional qualities, flavour, etc.), the development of resilient varieties to biotic (diseases and pests) and abiotic (drought, salinity, photosensitivity, etc.) stresses has become more important in the context of climate change. However, several anthropogenic and natural drivers may lead to genetic diversity loss (Khoury et al., 2022). Breeding practices that target a number of traits and use unique or few potential genotypes will favour certain alleles and increase homogeneity between individuals. Populations with narrow genetic diversity are more vulnerable to environmental stresses and may lead to the loss of the crop ability to cope with the changing environment, a feature that has become crucial nowadays. Domestication and artificial selection also reduces genetic diversity by locally favouring certain alleles (Yamasaki et al., 2007; Allaby et al., 2019). Anthropogenic pressure on biodiversity habitats has long been the leading cause of genetic diversity loss (Huston, 1979). Regarding the natural drivers of genetic diversity loss, the mating system and spatial distribution are among the main cases. Unlike sexual reproduction, the asexual mode is a common factor of narrowing the genetic background, for instance in the case of the cultivated *V. planifolia*. Species having a large geographical distribution are likely to be genetically diverse due to a reduction in gene flow. Isolation by distance has been reported to increase genetic diversity in crop species (Pusadee et al., 2009; Magige et al., 2022).

1.6.2 Genetic diversity analysis approaches

Several marker systems have been developed and used to disentangle the genetic diversity. Markers are considered fundamental in crop improvement where they play a wide range of roles, including material evaluation such as assessment of diversity degree, cultivar identification, and purity assessment (Djè et al., 2000; Becerra et al., 2010; Henry, 2012); tracking genomic regions under selection (Hasan et al., 2021); and studying heterosis (Stuber et al., 1992; Zhang et al., 1996; Rajendrakumar, 2015). Overall, the subsequent widely used marker systems are classified into four major types – morphological, cytological, biochemical, and molecular markers. Until now, the taxonomy of the *Vanilla* genus has been mostly based on morphological markers, especially flower and leaf (presence/absence) descriptions (Soto Arenas & Cribb, 2013; Soto Arenas & Dressler, 2013) and on its geographical distribution (Nielsen, 2000; Szlachetko & Veyret, 1994; Householder et al., 2010; Ferreira et al., 2020).

Morphological descriptors constitute a classical marker system mostly used in taxonomy (Will & Rubinoff, 2004; Soto Arenas & Cribb, 2013) and early cultivar selection (Karaköy et al., 2014; El-denary et al., 2016). Despite the advancement of other markers, morphological descriptors are still unavoidable and have diverse applications. They are a visual-based marker system; their low cost and ease of use are their main advantages. However, for large-scale observations, they become time-consuming and expensive. Some morphological traits are not stable (size, color of some organs, etc.) under different environmental conditions and could make discrimination difficult, thus not suitable as descriptors. Morphological descriptors may also vary from one species to another. For instance, leaf shape is a good descriptor for oak species (Borazan & Babaç, 2003), but not for cultivated vanilla.

Cytological markers include mainly the variation in structure and organization of chromosomes: size, volume, and length. These markers have been used to explore intra and inter-specific diversity among several crop species such as ginger (Nayak et al., 2005), maize (Albert et al., 2010), potato (Das et al., 2010), etc. However, the low resolution of this marker

system has been reported to be the major limitation in revealing genetic diversity (Bhanu, 2017).

Biochemical markers such as isozymes detect diversity at a functional gene level. The variation in protein sequence products of genes is identified by electrophoresis and specific staining (Liu & Cordes, 2004). Genetic variation is quantified by the presence/absence and allele frequencies (Liu & Cordes, 2004). These markers are easy to use and cheap. They were successfully used in intraspecific diversity analysis of several crop species (Hamrick & Godt, 1997). However, these markers are constrained by their small number, environmental sensitivity, low resolution, lack of heterozygosity, and the high quantity and quality of material needed. In addition, some variations at the DNA level do not affect protein sequences (Liu & Cordes, 2004; Govindaraj et al., 2015).

Molecular markers involve organelle and genomic DNA/RNA variations. These markers have become attractive in genetic diversity studies due to the large number of choices offered with plenty of configurations and allow researchers to find a compromise between the goal, the targeted sample context, cost, and time effectiveness. Molecular markers are therefore more flexible and help to solve the ambiguity revealed by other markers such as morphological descriptors. However, several studies combine both morphology and molecular markers to evaluate the genetic diversity of crop species (Pham et al., 2011; Naz et al., 2013; Ammar et al., 2015). Based on the approach by which markers are generated, molecular markers are classified into hybridization-based and polymerase chain reaction-based (PCR-based). Restriction fragment length polymorphisms (RFLPs) are among the early developed molecular markers that belong to the hybridization-based markers (Botstein et al., 1980). RFLP relies on restriction enzymes that cut the DNA at every recognition site and the use of a probe specific to a target DNA sequence. The length and number of DNA fragments are compared among individuals, populations, or species. These variations result from the differences in DNA sequences due to mutations or DNA rearrangement that provide differences in restriction site locations. Despite its codominant property, the power of RFLP in detecting genetic diversity is low due to the low polymorphism of the ma-

-rkers (Liu & Cordes, 2004). In addition, the markers require prior knowledge of the genome for the probe design. These markers were primarily designed for the human genome (De Martinville et al., 1982) before their application to plant species (Weber & Helentjaris, 1989; Devos et al., 1993; Dubcovsky et al., 2001).

PCR-based methods include rapid amplified polymorphic DNA (RAPDs), amplified fragment length polymorphisms (AFLPs), microsatellites with their relatives, and single nucleotide polymorphism (SNP) markers, among others. For RAPDs, a couple of short-length (8–10 bp) primers are used to amplify any DNA regions that may present the corresponding primer binding sites. Genetic variation is computed by scoring the presence or absence of amplified bands, usually on agarose gel electrophoresis. RAPD was reported to be powerful in detecting polymorphism, easy to use, and do not require prior knowledge of the genome (Lynch & Milligan, 1994). However, as a dominant marker, it is unable to distinguish homozygous from heterozygous loci. In addition, amplified fragments from different loci of the same size are assigned as a single locus. RAPD also suffers from low reproducibility (Liu & Cordes, 2004).

Amplified fragment length polymorphisms (AFLPs) is a combination of RFLP and RAPD as it combines both DNA digestion using restriction enzymes and fragment amplification by PCR. The RFLP step relies on the digestion of a target DNA molecule by one or more restriction enzymes tailored to the DNA's carrying restriction sites. AFLP is applicable without previous information about the target species, as adaptors can be used as primer sites (Vos et al., 1995). AFLP differs from the previous markers in its high polymorphism (Govindaraj et al., 2015). However, it is more expensive and still a dominant marker.

Microsatellites or simple sequence repeats (SSRs; Selkoe & Toonen, 2006) have been very popular for a few decades and are still in use regardless of the advancement in genotyping systems. SSRs are short tandem repeats of 2–6 bp. Compared to the previous cited markers, microsatellites were reported to be co-dominant, abundant with a higher rate of polymorphism and evenly distributed across the genome, making them very attractive in

genetic related field studies (Kalia et al., 2011). The requirement of the genome sequence for the primer design is the major weakness of this marker (Govindaraj et al., 2015) .

SNP marker identification has been developed to capture genomic polymorphism at a finer scale (Syvänen, 2001). SNPs can have 4 alleles (4 types of nucleotides), but usually these are biallelic. SNPs are therefore less polymorphic compared to the microsatellite, which is a multi-allelic marker. However, SNPs are much more abundant along the genome, enabling high throughput with lower error rates and ease of scoring (Flanagan & Jones, 2019). In the last decades, SNP markers have gained popularity in genetic studies. Helyar et al. (2011) suggest that SNPs are an appropriate marker for population genetic studies since they are the most common type of variation in the genome. Both coding (exons) and non-coding (e.g. introns, 5'UTR, and 3'UTR) sections of the genomes contain SNPs. In addition, with the advancement of technology, SNPs have become much more powerful compared to other markers – high degree of resolution, high reproducibility, low cost of genotyping, and time-effectiveness (Thomson, 2014).

Despite a long history of discovery, SNP markers were first genotyped utilizing chip technology in the late 1990s (Liu & Cordes, 2004). With the first generation of sequencing typically represented by the Sanger sequencing (Sanger et al., 1977) and the Maxam-Gilbert sequencing (Maxam & Gilbert, 1977) that sequences a small number of longer sequences until 1000 bp, SNP development was constrained by its cost. To address this issue, researchers developed the second or next generation sequencing (NGS; Behjati & Tarpey, 2013). NGS is much more flexible due to its variety of configurations that allow it to sequence both long sequences for whole genome sequencing (WGS) and short sequences to reveal SNPs for genetic comparison. This latter gives the opportunity to choose genome coverage depending on the goal of study and the cost. Three main platforms were therefore developed – Roche 454 (pyrosequencing; Rothberg & Leamon, 2008), SOLiD (ligation sequencing; Hedges et al., 2011) and the Illumina (sequencing by synthesis; Meyer & Kircher, 2010).

Comparing the whole genomes of every sample should be the best way to capture all the information within the target population. The costs and time involved, however, make this still impractical nowadays. Reduced representation sequencing (RRS) has therefore been proposed by NGS to sequence up to thousands of samples at once for a reasonable time and cost (Davey et al., 2011). The high-throughput sequencing developed by NGS allows studies of model and non-model organisms, with or without prior genomic knowledge (Garvin et al., 2010). This approach consists of subsampling homologous genomic regions across the studied individuals using restriction enzymes (REs) for digestion. RRS includes the restriction site-associated DNA sequencing approaches – mainly the genotyping-by-sequencing (GBS; Elshire et al., 2011) and relatives such as RAD-seq (Miller et al., 2007), 2b-RAD (Wang et al., 2012), ezRAD (Toonen et al., 2013) and ddRAD (Peterson et al., 2012). Overall, the principle of these methods relies on the restriction enzyme used to cut the genome at the specific recognition sites that will generate several thousands of markers (Davey et al., 2011). These methods differ mainly in the enzyme used for digestion, the used adaptors (oligo-sequences containing sequencing primer binding sites and sites designed to attach the DNA fragment to the flow cell of the sequencer machine) and the application or not of size selection. For example, GBS uses 2 REs and barcoded adaptors without size selection. The original RAD-seq relies on a unique RE to cut DNA, and shaking is performed to further shorten fragments. It uses an Y type adaptor for sequencing and size selection is not applied. The 2b-RAD uses enzymes of the IIB type that can digest DNA for 33 to 36 bp length. There is no required size selection. Like previous RRSs, dd-RADseq uses 2 REs and barcoded adaptors. However, it uses size selection to reduce the set of loci for sequencing. ddRAD-seq is a well-documented protocol that has gained popularity in studies of trait-association, population genetics and plant breeding because of high throughput and low-cost genotyping (Shirasawa et al., 2016; Jaiswal et al., 2019; Laila et al., 2019; Sudan et al., 2019; Esposito et al., 2020; Natarajan et al., 2020).

The protocol does not require prior knowledge of genome and as a reduced representation sequencing, ddRAD-seq aims to reduce the comp-

-lexity of the genome by subsampling specific regions across the genome making it a popular approach especially for species with complex genome and repetitive regions like strawberry (Natarajan et al., 2020), wheat (Saintenac et al., 2013), or sunflower (Pegadaraju et al., 2013). This tool was successfully used in some Orchid species (Roy et al., 2017; Gargiulo et al., 2021) and not yet applied on vanilla. For NGS analysis, SNP accuracy relies on enough depth of coverage that is the most important parameter to distinguish true polymorphism from errors, *i.e.* the greater the reads supporting a called variant, the more reliable the SNP (Wall et al., 2014). Bias can be due to the miss sampling among the two alleles in heterozygous locus and leading to incorrect homozygous calls and vice versa by sequencing error, homozygous genotype can be declared erroneously as heterozygous (Nielsen et al., 2012). Setting enough threshold value for SNP calling is hence capital for stringency of data for any analysis. The minimum value set varies among studies and the choice of the value depends greatly on the amount and quality of data from sequencing. Nielson et al., 2012 suggested a coverage greater to 15x to confidently avoid bias. However, lower coverage (between 4 to 10) were widely used in several studies (Pegadaraju et al., 2013; Lène et al., 2018; Wang et al., 2019). Sequencing configuration and SNP genotyping pipelines are greatly involved in the SNP output. Each approach does not provide an identical SNP panel. Choosing a suitable platform is often challenging and depends on several parameters, such as the goal itself, the genetic background of the target species (cross-pollinated or clonally propagated; high or low expected level of diversity; etc.). Depending on the cost, researchers can choose the amount of data, *i.e.*, the genome proportion to be sequenced and the depth of coverage, which will impact the data quality (the higher the number of times a locus is sequenced, the better its quality). Despite

the high throughput technology used in the marker generation, no genetic marker, including SNP markers, is completely error-proof (Andrews et al., 2016; Crotti et al., 2019). Reasonability in the whole post-sequencing analysis is therefore required to ensure that the discovered SNP markers reflect the real diversity among the studied samples. Basic steps for seque-

-nce data processing are common for every approach. The first step includes data assessment, separation of pooled reads by the specific barcodes (demultiplexing), and cleaning (Andrews et al., 2016). The latter is based on the quality of the reads. Low quality reads, ambiguous barcodes, and reads with uncalled bases are discarded. In addition, sequence sections with an overall low quality could be trimmed. Several tools have been developed for such tasks, including FastQC (Andrews, 2010), NGS QC Toolkit (Patel & Jain, 2012), HTQC (Yang et al., 2013), fastp (Chen et al., 2018) for quality assessment, Stacks (Catchen et al., 2013a), Trimmomatic (Bolger et al., 2014) for sequence trimming, Stacks (Catchen et al., 2013a), DeML (Renaud et al., 2015) for read demultiplexing. SNP markers are genotyped after sequence pre-filtering. A reference genome-based analysis can be performed if the whole genome of the target or closely related species is available; otherwise, a *de novo* analysis could also be conducted. The advantage of the use of a reference genome is mainly to get the map of sequences on the target species' genome. Several standard pipelines are now freely available for SNP genotyping (Nielsen et al., 2011). Practically, they do differ mainly on processing time, required bioinformatic skills, amount of output data, SNP coverage, error rate, and sometimes in specificity, like heterozygous level detection. Researchers can choose a suitable pipeline based on the goal of study, facilities, and the data itself. A pipeline that works for one species is not necessarily efficient for another. For example, the Stacks pipeline is good enough to study the genetic diversity of *Sarcophilus harrisii* (Wright et al., 2019) but not for vanilla due to the downside of detecting highly heterozygous loci as discussed hereunder. Intuitively, testing several pipelines on dozens of samples will be safer to pick the right pipeline that could respond to the biology of the species. The last step involves SNP filtering to separate true alleles from false positive alleles. Both extremes (too hard and too relaxed) of filtering could also lead to erroneous interpretation of the data (Catchenet al., 2013b; Mastretta-Yanes et al., 2015). Reasonable filtering should be therefore applied.

1.6.3 State of knowledge on the genetics of *Vanilla*

1.6.3.1 *The genome of vanilla*

Genomic data for vanilla is not yet fully elucidated. The complexity of the vanilla genome made its study challenging. The vanilla genome has been reported to undergo the common endoreplication mechanism in orchids and other plant species (Piet et al., 2022). This is characterized by the nuclear DNA content doubling without division in somatic cells, through different steps (Lee et al., 2010). This phenomenon has been reported to happen only on a part of the DNA content, i.e., a fraction is not endoreplicated (fraction F), and another fraction (fraction P) is subject to endoreplication (Brown et al., 2017). The chromosome number in *V. planifolia* is still a subject of debate. Intra-individual variation in chromosome number has been detected in vanilla (Nair & Ravindran, 1994; Bory et al., 2008c). The first whole genome sequencing of *V. planifolia* identified a diploid genome with 14 chromosomes (Hasing et al., 2020), similar to that detected by Nair (2n = 28, 2013) using chromosome counts, and a very recent study distinguished 16 chromosomes (Piet et al., 2022). Other previous studies have identified chromosome counts that indicate that the vanilla has a diploid genome with 2n = 32 for most of the studied accessions (Zur et al., 1930; Bory et al., 2008c). Occurrence of autopolyploidy (autotriploidy and autotetraploidy) has been noticed within *V. planifolia* accessions to explain morphological variation on the Reunion island (Bory, et al., 2008c). Variation in ploidy level (di, tri, and tetraploid) has also been documented in *V. tahitensis* (Lepers-Andrzejewski et al., 2011).

1.6.3.2 *Genetic diversity analysis in vanilla*

Most of the markers described above (1.6.2) have been applied in taxonomy and diversity studies of vanilla. The Vanilla genus has been classified and taxonomically revised primarily using morphological markers, with a few uses of chloroplast and ribosomal single genes (Cameron et al., 1999; Cameron, 2004; Cameron & Carmen Molina, 2006; Bouetard et al., 2010; Soto Arenas & Dressler, 2013). These taxonomies set the position of the genus within the Orchidaceae family and classified vanilla species into different sub-genera and groups as described earlier

(1.1.1). Seven enzymes used by Nielsen & Siegismund (1999) have not been able to separate three species of vanilla from each other, *V. claviculata*, *V. barbellata*, and *V. dilloniana*. These isozymes have not been able to relocate vanilla hybrids, for example, those putatively descended from *V. claviculata* and *V. barbellata* (Nielsen, 2000). RAPD (Besse et al., 2004; Sreedhar et al., 2007; Minoo et al., 2008; Verma et al., 2009), AFLP (Duval et al., 2006; Lubinsky et al., 2008; Minoo et al., 2008; Lepers-Andrzejewski et al., 2012) and SSR markers (Sreedhar et al., 2007; Bory et al., 2008d; Gigant et al., 2012; Fajardo et al., 2014; Ramos-Castellá et al., 2017) have also been used in several studies of relationships within and between vanilla species. All studies highlighted the effectiveness of the marker at the inter-species comparison level. Most of the markers were not suitable for intra-specific studies. However, ISSR and SNP markers have been demonstrated to separate some *V. planifolia* phenotypes from each other. For example, some cultivated and wild Mexican *V. planifolia* morphotypes have been separated using ISSR markers (Ramos-Castellá et al., 2017; Villanueva-Viramontes et al., 2017). The use of SNP markers on cultivated *V. planifolia* revealed slight differentiation among accessions (Hu et al., 2019; Chambers et al., 2021). These investigations confirm the low genetic base in the cultivated *V. planifolia* species, as discussed hereunder, and the effectiveness of SNP markers in detecting the fine scale genomic low variability.

It is recognized that the cultivated vanilla in its secondary areas of distribution arose from a narrow genetic base (Lubinsky et al., 2008; Minoo et al., 2008; Verma et al., 2009). Vanilla's current global spread is due to the vegetative multiplication of a few cuttings from Mexico via colonizers (Lubinsky et al., 2008). As a result, only a portion of the original gene pool has been presumably introduced in other areas of cultivation. The propagation of handfuls of cuttings coupled with the vegetative reproduction mode might reduce the genetic diversity in time within the cultivated vanilla (Besse et al., 2004; Bory et al., 2008b). However, morphological variation within the *V. planifolia* species has been reported. For instance in Mexico, four *V. planifolia* morphotypes have been distinguished in the field based on stem and leaf color – Mansa (= Dura),

Rayada (= Variegata= acamaya), Oreja de Burro and Albo-marginata (Soto Arenas & Dressler, 2013b). The worldwide cultivated *V. planifolia* is believed to derive from Mansa, the commonly cultivated type in Mexico (Soto Arenas & Dressler, 2013b). Rayada is distinguished from others by the presence of yellow stripes on stems and leaves. However, Soto Arenas & Dressler (2013) did not find any genetic difference between this type and Mansa, using the ITS region of nuclear ribosomal DNA. Albo-marginata is characterized by leaves with white margins; it was suggested that this type originated from a mutation of the Mansa type in Madagascar (Soto Arenas & Dressler, 2013b). However, as mentioned above, its presence in Mexico has been now documented (Chambers, 2019). Oreja de Burro is a self-incompatible morphotype (Castillo Martinez & Engleman, 1993). It is important to note that these morphological variants are not separate species, but rather variations within the same species, *V. planifolia*. The different morphotypes may be the result of genetic mutations or environmental factors that affect the expression of genes in different parts of the plant that may result in the appearance of distinct morphological variants within a single plant species. Variation within *V. planifolia* has also been discovered on Reunion Island, including the phenotypes Classique (the common vanilla with flat, light green leaves and pods that gradually narrow) and Mexique (= Bleu, dark blue), which have curved sides and dark bluish leaves with a central gutter and produce cylindrical pods. Other minor morphotypes, such as Aiguille, Grosse Vanille, Sterile, and Variegata, have been discovered in addition to these two major cultivated types (Bory et al., 2008c). Despite the morphological diversity of *V. planifolia* in Mexico and the Indian Ocean, authors studying these morphotypes have commonly concluded that they present little genetic variation (Bory et al., 2008c; Lubinsky et al., 2008). Using the RLFP markers, the genetic differentiation of these variants, including the above cited Mexican morphotypes, has revealed very low diversity. Only the type Aiguille, a possible result of sexual reproduction, has been distinguished from the others (Bory et al., 2008c). Several hypotheses have been raised to explain the origin of the variations within *V. planifolia*, including spontaneous mutation, sexual recombination (Aiguille), natural seedling, epigenetic variation, and polyploidization (Bory et al., 2008c). The autopolyploidy of

V. planifolia in the Indian Ocean has partially explained the phenotype variation, with the phenotypes Sterile and Grosse vanille being tri- and tetraploid (Bory et al., 2008c). Enzymatic profile variation has also been reported among *V. planifolia* accessions in India (Sasikumar, 2010).

1.6.3.3 Germplasm conservation

In Mesoamerica, efforts are being made to conserve the *Vanilla* species *in situ*, particularly for its primary gene pool. However, this diversity is threatened mainly due to forest loss and the over-extraction of resources for replanting. In addition, areas under conservation do not fully cover small and dispersed species (Bory et al., 2008b; Watteyn et al., 2020). Some initiatives on farmer-based conservation (*circa situm*) through agroforestry systems are ongoing in Columbia. The application of this approach has been reported to be challenging as the conservation of each type is highly dependent on its economic value to the farmer (Bramel & Frey, 2021).

In response to threats to vanilla biodiversity loss, several strategies for short- and long-term ex-situ conservation have been developed and implemented. Vanilla germplasm is maintained in different forms, including living plants, *in vitro* culture, seeds, and cryoconservation. Bramel & Frey (2021) have reviewed the global efforts on vanilla conservation and found that many *Vanilla* species are not yet under conservation, most species are endangered, and several species have few accessions in collection. Among the 118 recorded species of *Vanilla*, 54 have been identified as being under conservation – 18 for *Xanallata*, 31 for *Tephya* and 5 for *Membranaceae* sections. These species contain a total of 1,946 accessions recorded from 221 institutions (national genebanks, universities, botanical gardens and private collections). Most of the conservation systems are constrained by vulnerability to biotic and abiotic stresses, cost, time, the socio-economic context, and the effectiveness of the approach. Improvements and research have been suggested to set effective and long-term conservation strategies (Grisoni et al., 2007). To improve vanilla germplasm conservation, protocols have been shared, including those for seed *in vitro* germination (Dequaire, 1976; Porrás-Alfaro & Bayman, 2007), cryoconservation (Gonzalez-Arnan et al., 2009;

Hernández-Ramírez et al., 2020) and tissue culture (Abebe et al., 2009; Zuraida et al., 2013).

The following chapter is focused on exploring the genetic resources of cultivated vanilla from the ancient vanilla breeding program in Madagascar. The studied samples were therefore collected from a local collection (the Rama Export collection). In addition to the genetic diversity investigation, the current pilot study was mainly conducted to develop the ddRAD protocol in vanilla and to establish the whole pipeline for data processing prior to the national scale study. The DNA extraction was conducted at the Fundamental and Applied Biochemistry Department, Université d'Antananarivo, Madagascar, and the remaining parts were realized at the Université catholique de Louvain, except the DNA sequencing (GENEWIZ, Germany). The paper was realized with the collaboration of the following researchers:

Rasoamanalina Rivo Onisoa Léa¹, Mirzaei Khaled¹, Mondher El Jaziri², Ramirez Ramirez Angel Rafael^{1,3}, Bertin Pierre¹

¹Earth and Life Institute - Agronomy - Université catholique de Louvain – Belgium ; pierre.bertin@uclouvain.be

²Laboratory of Plant Biotechnology, Université libre de Bruxelles, Belgium; Mondher.El.Jaziri@ulb.be

³Faculty of Agroforestry, University of Guantánamo, Cuba; aramirez@cug.co.cu

Chapter 2

Diversity and structure assessment of the genetic resources in a germplasm collection from a vanilla breeding program in Madagascar

Abstract

A breeding program of aromatic vanilla, dating back to 1944, was conducted in Ambohitsara, Antalaha, SAVA - Madagascar. Imported, local, wild and cultivated vanillas were used as progenitors and thousands of hybrids were generated. However, no genetic evaluation of this germplasm has been conducted and, after the end of the project (2000), these genetic resources seem to have been dispersed or lost. Here, we aim to investigate the genetic diversity and structure of rescued genotypes held in a local collection in Antalaha. ddRAD-seq protocol was applied that provided 865 million read sequences from 56 accessions. The filtered 23,701 concordant common SNPs detected by the three widely used tools (Stacks, BCFtools, GATK) for SRSL were used for germplasm evaluation. Clustering analysis segregated samples into five genetic groups: *Vanilla planifolia*, *Vanilla pompona*, hybrid Tsitaitra, Vanille Banane, and the phenotype Tsivaky. The overall genetic diversity was high ($H_o = 0.39$). The presence of private alleles in all groups (from 487 to 2,866) indicated that populations were diverging and represented a large gene pool that could be useful for future breeding efforts. Conservation of the remaining resources and further prospect is therefore urgent. The comparison of the three pipelines showed that Stacks was less suitable than both other ones for genetic diversity study for vanilla due to its low capacity in detecting heterozygous loci. BCFtools and GATK provided closer results and performed better. Among those two, BCFtools is much more reasonable in time and material requirement with good enough output and can be suggested for genome-wide genetic investigation in orchids.

2.1 Introduction

The genus *Vanilla* is included into the Orchidaceae family within monocots and contains about 130 species mostly wild, among which 30 have been reported to wear aromatic fruits (Soto-Arenas, 1999). However, only the widely cultivated and commercial *Vanilla planifolia* and the far less important *V. X tahitensis* present commercial interest (Lepers-Andrzejewski et al. 2012). Other species such as *V. pompona* and *V. odorata* exist within vanilla plantations or are harvested in the wild but are not economically important.

Cultivated vanilla around the world arose from the domestication of a unique population of *V. planifolia* from Meso America (Bory et al., 2008b). Because of this domestication and diffusion process, commercial vanilla has very limited genetic diversity (Bory et al., 2008c; Minoo et al., 2008). In addition, asexual propagation makes this crop very uniform genetically and, therefore, vulnerable to disease infection. Besides, obtaining vanilla with high vanillin content and disease resistance is challenging. Breeding programs are limited by (i) the few available germplasm resource databases; (ii) the asexual propagation commonly used by farmers; (iii) the complexity of the genome (Brown et al., 2017); and (iv) the relatively long-life cycle of vanilla. In fact, the first flowers only appear around the third year of cuttings and by the seventh year after seed germination.

Regardless of the high economic value of vanilla (the second most expensive spice after saffron), few breeding efforts have been made. One of them was conducted in Madagascar, dating back to the middle of the twentieth century (1944). In view of producing disease-resistant and high vanillin content varieties, several crossing strategies were conducted between introduced, local, cultivated and wild species. The improvement aimed to keep the potential traits of *V. planifolia*, especially the aroma profile, and thus used this species as a basis in different crossing configurations and mainly attempted to introduce the resistance and hardness traits from other species such as *V. phaeantha*, *V. pompona*, *V. madagascariensis*, *V. abundiflora*, *V. coursii*, *V. francoisii*, *V. walkeriae*, and

V. zanzibarica (Grisoni & Nany, 2021). Additionally, *V. tahitensis* was also used to introduce the bean indehiscence trait into *V. planifolia*. The intra- and inter-specific mating and trait selection led to a thousand hybrids from which “Tsitaitra” and “Magnitry Ampotony”, reported as hybrids of *V. planifolia* x *V. tahitensis* cv *Haapape* and *V. planifolia* x *V. pompona*, were retained for disease resistance and vanillin content, respectively. Unfortunately, the program was abandoned in 2000, about 50 years after its beginning and the genetic resources seem to have been dispersed or lost and the program is poorly documented. Since then and as far as we know, no research has been conducted on the ancient breeding program. It is, therefore, urgent to assess the remaining resources for the conservation of the breeding materials and for future genetic improvement. To fill this gap, the present study aims to investigate the genetic diversity and structure of the genotypes kept within a local collection in Antalaha, in the SAVA region, in the north-eastern part of Madagascar. The vanilla collection was established in Antalaha by Ramanandraibe Exportation SA, known as Rama Export (a local firm exporting vanilla), and containing a number of genotypes mainly collected from the abandoned vanilla research station with rare plants from other places within the SAVA region.

Among molecular techniques, next-generation sequencing (NGS) has become a popular practice in conservation programs for diversity investigation (Ekblom et al., 2011; Hunter et al., 2018). NGS is a high-throughput genotyping that allows sequencing of several individuals at once for a reasonable time (Kumar et al., 2012; Shirasawa et al., 2016). Due to their cost-effectiveness and flexibility, genotyping-by-sequencing (GBS) and restriction associated-site (RAD) protocols are now very attractive NGS methods. Double digest RAD-seq (ddRAD-seq) is one of the several short-read library sequencing (SRSL) within the RAD family that targets homologous regions across the genome and relies on two restriction enzymes. This protocol was used in the present study since it has been successfully used in several crop species having complex genomes like vanilla (Saintenac et al., 2013; Feng et al., 2020; Natarajan et al., 2020).

2.2 Materials and methods

2.2.1 Plant material and datasets

Fifty-six vanilla plants maintained in the Rama Export collection and a local farmer in Antalaha, SAVA region, Madagascar were used in this study. The germplasm panel includes remaining progenitors and hybrids from the ancient breeding program (in Ambohitsara, Antalaha) pertaining to five groups easily identified phenotypically: *V. planifolia*, *V. pompona*, hybrids labelled as Tsitaitra and some other varieties, so-called Tsivaky and Vanille Banane. The variety Tsitaitra has been documented as a result of *V. planifolia* X *V. pompona* crossing. However, information regarding the putative parents of Tsivaky and Vanille Banane is missing.

2.2.2 DNA isolation, library preparation and sequencing

Total genomic DNA from young and fresh leaves was extracted using the CTAB-based protocol established by Doyle et Doyle (1987) with some modifications. Briefly, to release the cell content, 400 mg of leaf were crushed in 1 ml of CTAB buffer solution preheated to 65 ° C (0.1 M Tris pH8, 20 mM EDTA pH8.5, 1.4 M NaCl, 2% CTAB, 0.5 mM Na₂S₂O₅). β-mercaptoethanol and PVP 1% each were added right before extraction. The solution was then incubated at 65 ° C for 45 min with shaking every 5 min. DNA was separated from cell debris using chloroform / isoamyl alcohol (24:1, v/v) followed by centrifugation. This step was repeated twice to avoid contamination. The DNA in solution was subsequently precipitated with 3M sodium acetate, pH 5.2 (1/10 volume) and 2/3 volume isopropanol stored at room temperature. Then, the DNA pellet was washed twice with cold 70% ethanol. After evaporation of residual alcohol, DNA was resuspended in 30 µl of TE1x solution mixed with RNase A DNase free to remove molecular RNAs. Finally, the DNAs were stored at -20 °C.

For library preparation, DNA integrity was first evaluated on 1% agarose gel electrophoresis and quantity was assessed with Qubit (2.0 Fluorometer, Invitrogen, Carlsbad, CA, USA). These tasks were performed to ensure the success of the DNA sequencing as DNA degradation can affect the accuracy of the data. In addition, libraries require a minimum

amount of DNA to ensure sufficient coverage and accuracy during sequencing. In order to avoid sequencing bias through over- or underrepresentation of samples, DNA concentration was then normalized to 600ng in 30µl for each sample. For each individual sample, genomic DNA was digested with two restriction enzymes ECoRI HF (NEB) and NlaIII (NEB) by incubating at 37 ° C for 18 hours. This combination of a rare cutter that recognizes longer and less common restriction site, ECoRI HF and a frequent cutter that recognizes shorter and more common restriction site, NlaIII) was chosen to improve the efficiency, accuracy, and uniformity of the generated short sequences. Digestion success was assessed on 1% of agarose gel. Then 200ng of double digested DNA were ligated with two adapters at each restriction cut site using T4 DNA ligase (NEB). Both adapters are designed to be compatible with each cut site overhang and serve for later amplification step as they are complementary to the Illumina PCR oligo-sequences. The first adapter is barcoded, the identifier is specific for each individual to allow to identify reads from each sample during data processing. The second adapter is common for all individual in the same library (Figure 2.1). The barcoded samples in each set were then pooled together. To sequence short representative DNA fragments, we selected adapter-ligated fragments ranging from 300 to 600 bp. DNA fragment size selection was performed at the Genomics Core in Leuven, Belgium. A specific index was added by PCR at the common adapter side of size selected fragments. The index allows to pool together two or more sets of individuals. The prepared libraries were sent to GENEWIZ (Germany) company for sequencing using Illumina NovaSeq 250 bp_length paired end sequencing platform.

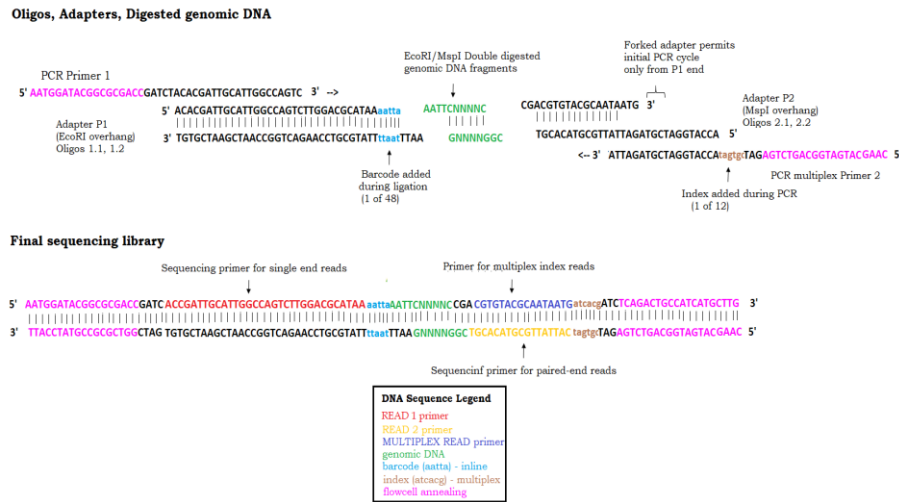


Figure 2.1. Diagram of final library (Peterson & al., 2012).

2.2.3 Quality assessment, cleaning, and demultiplexing

Raw sequence reads were demultiplexed with Stacks v2.53 (Catchen et al., 2011) using an individual's dual index-barcode information. Read quality was then checked using Fastqc v0.11.9 (Andrews, 2010). To retain only fragments of good quality, reads were trimmed from 220 bases and those with uncalled base or base quality lower than 30 were discarded.

2.2.4 Alignment to the reference genome

Paired end reads were aligned to the published reference genome (Hasing et al., 2020) using the Burrows-Wheeler Aligner (BWA) v0.7.15 (Li, 2013). The BWA mem function was run with its default parameters as described by (Rochette et al., 2017). SAMtools v.1.6 (Danecek et al., 2021) was used to convert the aligned reads from SAM to BAM format and sort the reads by position for variant calling. Mapping statistics were computed by VCFtools v4.2 (Danecek et al., 2011).

2.2.5 SNP calling and filtering

Many tools have been designed to call variants from wide-genome sequencing. Some of them provide outputs closer between themselves than other ones. some of them may be prone to bias and generate unreliable results (Stift et al., 2019). Tools differ mainly in the algorithms used to estimate genotype and in accuracy. Choosing the best pipeline is

often challenging. To get out high-quality variants from the sequenced reads, the three pipelines widely used for SRS genotyping were selected, i.e. Stacks (Catchen et al., 2013a), BCFtools (Danecek et al., 2021, <https://samtools.github.io/bcftools/bcftools.html>), and GATK (Poplin et al., 2018) and compared to each other. Concordant common sites detected by these approaches were used for downstream analysis.

Stacks relies on building catalogues of loci from aligned reads across all individuals, from which variants are called, based on a Bayesian maximum-likelihood statistical model (Hohenlohe et al., 2010; Catchen et al., 2011). The catalogue of loci was built using the gstacks program implemented in Stacks v2.53 with default parameters except the mapping quality set to greater than 30. Preliminary filtering was applied on raw SNPs using the included population module (r 80, min_maf 0.01). BCFtools calls variants based on calculated genotype likelihood. It starts by generating genotype likelihoods at each genomic position (BCFtools mpileup function) and uses this to call genotypes (call function). GATK uses a Bayesian probabilistic model to calculate genotype likelihoods. This pipeline was run with its default parameters. It involves preprocessing of the aligned reads (picard), per-individual variant calling (HaplotypeCaller), genotype data merging (genotypeGVCF) into a single VCF file, and filtering (Danecek et al., 2011). In each of the three pipelines, only reads with a mapping quality greater than Q30 were used for genotyping. The SNPs were further filtered according to the following criteria: present in at least 80% of individuals; minimum allele frequency of 0.01, minimum and maximum depth for a site: 10 and 250, respectively; minimum and maximum depth for a genotype per individual: 10 and 250, respectively. This task was performed by VCFtools v4.2 (Danecek et al., 2011), which was again used to identify common SNPs between pipelines. Consecutive SNPs (within 10 bp window) were removed and the set of SNPs was used for genetic diversity and structure analysis. SNP location and function were identified with VEP (McLaren et al., 2016). to further validate the SNPs.

2.2.6 Basic statistics of genetic variation and population structure

Genetic diversity and clustering were analysed using the concordant and common filtered SNPs detected by the three compared pipelines. The genetic diversity among accessions was assessed by computing the expected (H_e) and observed heterozygosity (H_o), fixation indices, inbreeding coefficient (F_{is}), and private alleles (P_a) through GAD v.1.1 (Lewis & Zaykin, 2002). ADMIXTURE v1.02 (Alexander & Lange, 2009). model-based cluster estimation with K ranging from 1 to 10 and 1000 bootstrap replicates was used to identify genetically homogeneous groups within the population. Cross-validation (CV) values were used to identify the appropriate K -value. The tool evaluates the fit of different clustering scenarios and the results were summarized by simple graphs of K , where the lowest peak indicates the theoretically best number of groups. Principal Component Analysis (PCA) and UPGMA tree (based on IBS distance) were also used to visualize genetic relationships between accessions using TASSEL v5 (Bradbury et al., 2007) and ggplot2 of R v4.2.0 (Kronthaler & Zöllner, 2021) for plotting. PCA was also applied on data from each pipeline separately.

2.3 Results

2.3.1 ddRAD sequencing, cleaning, and sequencing

A total of 865 million reads were generated from Illumina NovaSeq sequencing of the 56 vanilla samples. After cleaning and filtering the raw sequence data, 72.5% of the reads were retained, with a mean of 11 million cleaned and high-quality ($Q > 30$) reads per sample, and eleven individuals with a high rate of missing data were removed for downstream analysis, thus leading to 45 individuals (Table 2.1). The genome-wide GC content was 34%. An average of 99.3% (SD 1.0) of total reads were physically mapped to the reference genome that covered about 13 % of the genome.

Table 2.1. Descriptive statistics on ddRAD sequencing and the genetic diversity of the 45 studied vanilla accessions from the 23 701 common SNPs. N: number of individuals; He: expected heterozygosity; Ho: observed heterozygosity; Fis: inbreeding coefficient; Pa: private alleles.

Accession type	N	He	Ho	Fis	Pa
<i>V. planifolia</i>	10	0.17	0.26	-0.51	1 118
<i>V. pompona</i>	8	0.31	0.50	-0.69	1 101
Tsitaitra	22	0.27	0.47	-0.71	1 195
Tsivaky	3	0.24	0.38	-0.90	2 866
Vanille Banane	2	0.30	0.34	-0.55	487
Total*/mean	45*	0.26	0.39	-0.68	-

2.3.2 Comparison between Stacks, BCFtools and GATK

Stacks detected the highest number of SNPs, i.e. 316 398. BCFtools and GATK provided a number of markers quite closer to each other, i.e. 80,888 and 126,391 respectively. The missing data rates were of the same order of magnitude; however, Stacks had a greater value than both others. Stacks showed the lowest loci coverage value (33x), while loci called by BCFtools and GATK showed higher sequencing coverage (64.6x, 51.4x). Stacks appeared to call for less heterozygous genotypes compared to the other pipelines (Table 2.2).

Table 2.2. Comparison of the three variant callers: Stacks, BCFtools, and GATK. Time CPU: running time for variant calling (hours); Nb: number of filtered SNPs; He: expected heterozygosity; Ho: observed heterozygosity.

Pipeline	Time CPU	Nb of filtered SNPs	% missing data (SD)	Loci coverage: Mean (SD)	He	Ho
Stacks	2	316 398	17 (2.5)	33.37 (17.8)	0.12	0.14
BCFtools	39	80 888	14 (5.3)	64.6 (23.4)	0.24	0.34
GATK	316	126 391	13 (5.7)	51.4 (15.6)	0.23	0.30

2.3.3 Common SNP discovery

After comparing variants called by the three pipelines, 34,219 filtered concordant common SNPs were obtained (Annex, Figure B.2). The 23,701

resulting high-quality SNPs were physically mapped to the genome, which were proportionally distributed across the 14 vanilla chromosomes (Hasing et al., 2020), i.e., the per-chromosome number of SNPs appeared to be proportional to its size (Figure 2.2). A total of 2,998 of these variants were located in coding regions (Annex, Table A.1).

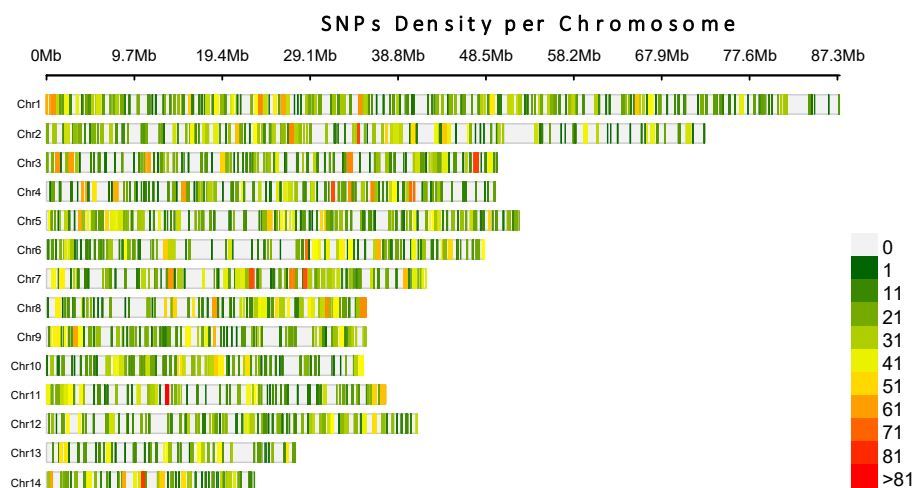


Figure 2.2. SNPs density distribution along the chromosomes generated from ddRAD sequencing of the 23,701 common loci detected by Stacks, BCFtools, and GATK.

2.3.4 Genome-wide genetic variation

Using the 23,701 common SNPs, the revealed genetic diversity of the studied vanilla accessions was high, with an observed heterozygosity (H_o) of 0.39. *V. pompona* and the hybrid Tsitaitra group presented the highest values (0.50 and 0.47), and *V. planifolia* exhibited the lowest one (0.26). The occurrence of private alleles (alleles that occur only within a given group) was noticed in each phenotypic group, with the largest being observed within the Tsivaky phenotype (2,866). Negative F_{is} was obtained for all groups suggesting an excess of heterozygosity (Table 2.1).

2.3.5 Population structure and genetic relationships

Based on the three first axes of PCA, the five phenotypic groups were neatly separated from each other by BCFtools, GATK and the common SNPs (Figure 2.3). Three groups can be separated by PC 1: the group of *V. pompona* accessions, the group containing both Tsitaitra and Vanille Ba-

-nane and the last one containing both *V. planifolia* and Tsivaky. Using PC 2, Tsitaitra and Vanille Banane were then separated from each other, and similarly *V. planifolia* from Tsivaky by PC 3. With Stacks, however, the phenotypic groups were mixed except Vanille Banane (first and second components: Figure 2.3, first and third components: Annex, Figure B.1).

The 23,701 common SNPs were used to analyze the genetic structure of the 45 vanilla accessions with ADMIXTURE and UPGMA. The best K-value detected by the cross-validation error plot was K=7 (Figure 2.4 a). This result was consistent with the PCA and UPGMA (Figure 2.5). Indeed, with K=7, the five phenotypic groups clearly segregated from each other (Figure 2.4 b) and two sub-groups appeared within *V. pompona* and Tsitaitra phenotypic groups, respectively.

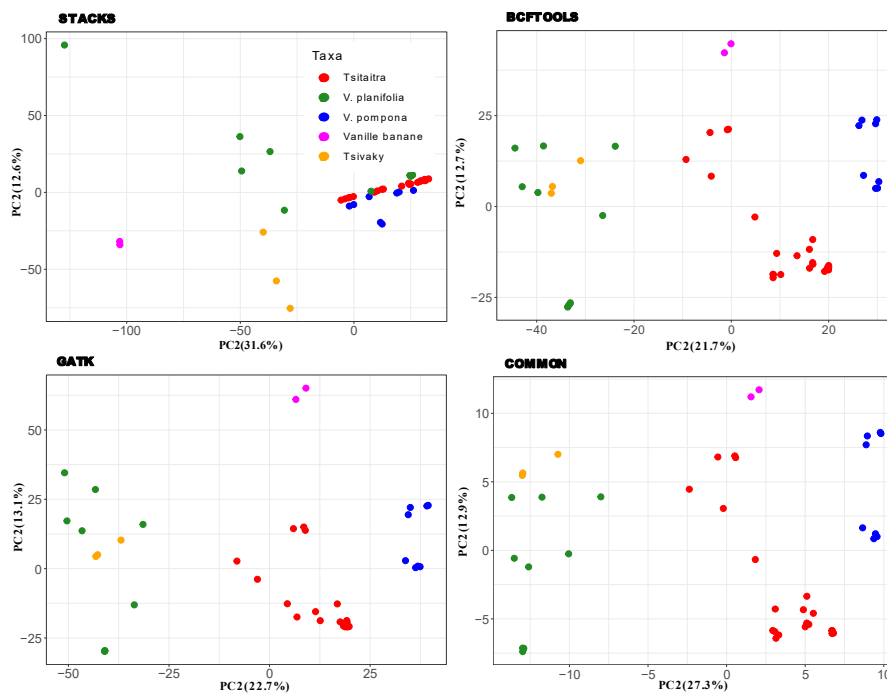


Figure 2.3. First and second principal components of the PCA of the 45 vanilla samples with the filtered SNPs from Stacks (316,398), BCFtools (80,888), GATK (126,391) and the 23,701 common SNPs detected by all three protocols together.

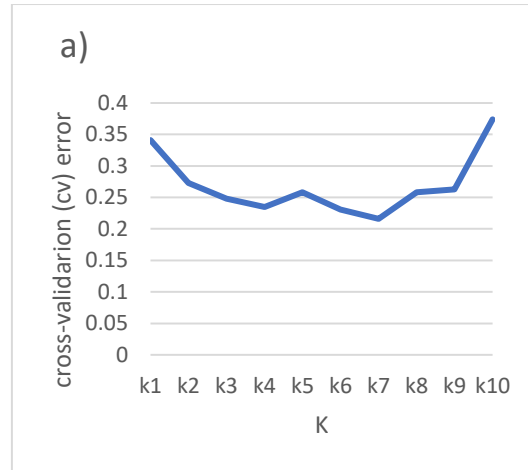
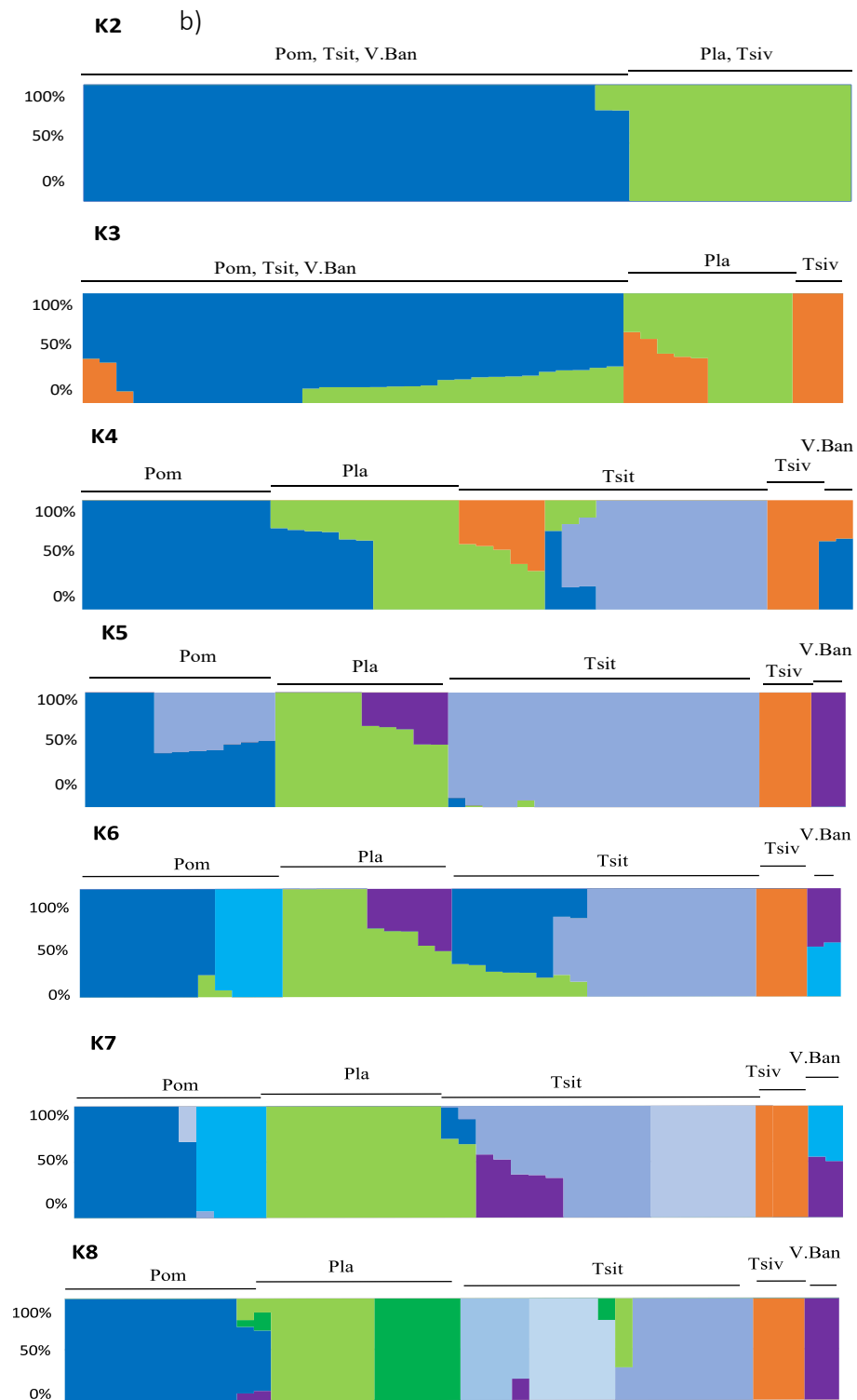


Figure 2.4. Genetic structure of the 45 cultivated vanilla accessions revealed by ADMIXTURE with the 23,701 common SNPs detected by Stacks, BCFtools, and GATK. a) Cross validation (CV) error plot, the lowest value indicates the appropriate number of populations ($K=7$). b) Estimated population structure of the 45 individuals from ADMIXTURE. Pom (*V. pompona*), Pla (*V. planifolia*), Tsit (Tsitaitra), VBan (Vanille Banane), Tsiv (Tsivaky) (NEXT PAGE).



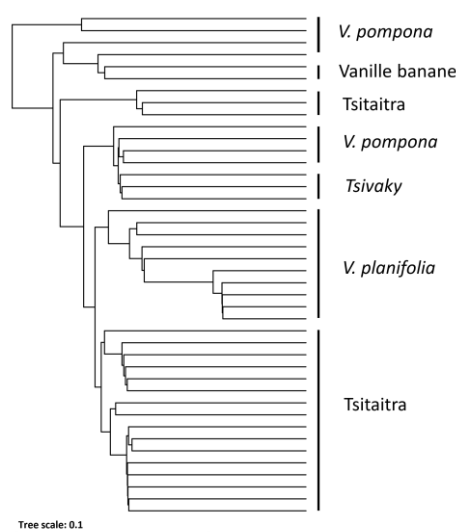


Figure 2.5. UPGMA tree of the 45 individuals based on IBS distance from the 23701 common loci detected by Stacks, BCFtools, and GATK together

2.4 Discussion

Understanding the genetic diversity and the structure of *ex-situ* populations is crucial for better management of resources and for genetic improvement programs. Furthermore, for conservation purposes, an important role in germplasm screening, organization, collection establishment, understanding the survival of germplasm and growth pattern becomes possible (Pillon et al., 2007). The genome-wide identified SNPs were used to investigate the genetic diversity and structure of the 45 samples from the vanilla collection in Antalaha, Madagascar.

The GC content of the 865 million reads generated from Illumina sequencing data was 34%, which is within the range estimated for monocots (33.6–48.9%, according to Šmarda et al., 2014) and slightly higher than the value previously reported for *V. planifolia* (30.8%, Hasing et al., 2020). High GC content has been shown to play an important role in the biology and ecology of plants, such as adaptation to challenging environmental conditions (Šmarda et al., 2014).

Pipelines widely used in population genetic diversity and structure studies

include Stacks, BCFtools, and GATK were used. To get high-quality SNPs from raw data, we compared the performance of these tools by computing the genetic diversity and structure by PCA and by evaluating a few other parameters such as data quality and computing time. Despite the high number of loci detected, Stacks did not clearly reveal any structure within the germplasm. When compared to Stacks, the success of BCFtools and GATK may be attributed to their ability to identify more likely heterozygous loci (Li, 2011; Li & Wren., 2014). Presumably, this difference is due to differences in the algorithms used for variant calling. Stacks uses a population-based approach, whereas GATK uses a variant calling model based on Bayesian statistics. In the first approach, variants are identified by comparing the frequency of alleles across the population, and variants that are present in multiple individuals are more likely to be real. This approach is generally more computationally efficient and faster than model-based approaches. A model-based approach, such as the one used by GATK, uses a statistical model to estimate the likelihood of observing different genotypes at each genomic position. In this approach, the model takes into account factors such as sequencing depth, base quality, and sequencing error rates to calculate the likelihood of each possible genotype. The genotype with the highest likelihood is then selected as the most probable genotype at that position. This approach is computationally intensive and requires more processing power. Regarding the effectiveness of BCFtools, data is scarce despite its widespread use. Up to date, Stacks has been successfully used in several studies (Saenz-Agudelo et al., 2015; Feng et al., 2020; Esposito et al., 2020), but not in vanilla, as stated in the present study. Vanilla is a highly heterozygous species (Hu et al., 2019), and only tools good enough to detect this level of heterozygosity are suitable. Our research supports the notion that preliminary evaluation of several pipelines for a given dataset is crucial for investigating genetic diversity. GATK and BCFtools produced outputs close to each other, but GATK required more time and memory than BCFtools (Table 2.2). Considering the respective initial number of SNPs, the number of SNPs common to the three pipelines was quite low, indicating the possible presence of bias and particularity of each approach. SNPs generated by BCFtools and GATK have shown comparable output (Ho, PCA). Therefore,

BCFtools combined with stringent filtering can be suggested as a good enough option for genome-wide genetic investigation in vanilla, being reasonable in time and material requirement with good enough output.

The common 23,701 identified SNPs by the three tools were evenly distributed across the 14 vanilla chromosomes (Figure 2.1). The number of markers within each chromosome was proportional to its size, indicating the effectiveness of the ddRAD-seq protocol for vanilla genetic investigation and suggesting its adequacy on a large scale, for example at the national scale study. Thus, these markers were used to assess the diversity of the 45 accessions from the ancient breeding program. Among the accessions, the highest observed heterozygosity was found for *V. pompona* (0.50) while *V. planifolia* had a much lower value (0.26). This is fully consistent with the previously reported low genetic diversity of *V. planifolia* (Bory et al., 2008d). Intense selection on a specific trait, particularly the aroma profile, may result in the loss of polymorphism on certain loci. High heterozygosity was obtained from the hybrid individuals (Tsitaitra) compared to *V. planifolia*. The hybrid Tsitaitra is the result of crossing between the two main species *V. planifolia* and *V. pompona*. The high number of private alleles in all groups suggests that populations are diverging. The high number of specific alleles represented a large gene pool that could be useful for a future breeding program.

Vegetatively propagated plants are reported to show an excess of heterozygosity (Elias et al., 2001; Ge et al., 2005), including vanilla (Hu et al., 2019). From the core selected SNPs, all accessions exhibited negative inbreeding coefficient (F_{is}) values. This result indicated pronounced heterozygosity. Likewise, heterosis may be the result of the higher observed heterozygosity than expected (Table 2.2).

Clustering analysis could reveal the relationship between individuals. Genotypic data was used to cluster vanilla samples. The first two axes of the principal component analysis (PCA) explained altogether 40.2 % of the total SNP variation. This variation allowed to distinguish the structure within the studied germplasm. Indeed, *V. planifolia*, *V. pompona*, hybrid individuals (Tsitaitra), Vanille Banane and Tsivaky were all separated from

each other. This result was consistent with the results obtained with ADMIXTURE ($k = 7$). In addition, the three approaches (PCA, UPGMA and ADMIXTURE) altogether confirmed the occurrence of two sub-groups in both *V. pompona* and Tsitaitra groups. The segregation of accessions within the Tsitaitra group may result from different crosses applied during the breeding program. Regarding the *V. pompona* group, genetic separation among accessions from different geographical origins has been previously reported (Bory et al., 2008b), a result confirmed by our study as discussed in the next chapter (Chapter 3). The origin of introduced *V. pompona* progenitors from different geographical areas (West Indies, France, Comoros, Mauritius and Moye Bampamperu, Grisoni & Nany, 2021) supports this hypothesis. Tsitaitra is a very robust vanilla vine, resistant to *Fusarium sp.* and to challenging environmental conditions. In general, this variety shows thick and tough leaves with a robust stem, mostly with bigger flowers compared to that of *V. planifolia*. Its pods are bigger and longer, but with lower vanillin content. This variety has been reported to result from different crossings between *V. planifolia* and *V. pompona* (Grisoni & Nany, 2021), and the overall result suggests that most of the Tsitaitra hybrids could result from a single back cross of *V. planifolia* x *V. pompona* with *V. pompona*. The variety Tsivaky is a robust vanilla plant, with a general dark green color with a cylindrical, indehiscent, and very big bean with a higher vanillin content than that of *V. planifolia*. The flower of Tsivaky is very specific with the presence of teeth-like hairiness on the labellum. We do not have any information about its progenitors, but our results showed that the group is close to *V. planifolia*. Based on its flower shape, a wild species such as *V. pilifera* could have contributed to its genomic composition, and also *V. tahitensis* for the pod indehiscence and earliness behaviour. UPGMA shows that this variety is close to *V. pompona*. Its robustness may thus come from this later.

2.5 Conclusions

A successful ddRAD protocol generated 23 701 reliable SNPs that were equally distributed along the 14 chromosomes of vanilla. The genetic ap-

-proach used clearly distinguishes *V. planifolia* from *V. pompona* and allows us to relocate the hybrids (Tsitaitra) and the other phenotypes (Tsivaky and Vanille Banane) among the previous species. The investigated collection has a diversity of interesting traits such as vanillin content, bean indehiscence, yield, and fungi and viral resistance. The detected high-quality SNPs could be used in conservation strategies and trait association studies and breeding implementation in the future. Despite the long list of introduced and generated genetic material for the 50 years of research in Antalaha, the evaluation indicates a very low number of remaining resources. Conservation of these accessions is therefore urgent for moving forward in the future breeding program. In addition, research on a larger scale, such as a national scale, to gather as much information as possible about Madagascar's vanilla gene pool is also necessary.

Once the whole protocol has been successfully developed, we collected data on a national scale. The goal was to obtain a general overview of the genetic resources of the cultivated vanilla in Madagascar. To do so, we targeted potential regions for prospecting by taking into account the diversity, geographical and ecological variation. The four major areas of vanilla cultivation in Madagascar were therefore considered as described hereunder. The data collection was possible with the intervention of the Rama Export company framework in different regions. They facilitated our contact with famers and kindly provided technicians who helped us with the site sampling based on their experience. As in the previous chapter, the DNA extraction was conducted at the Fundamental and Applied Biochemistry Department, Université d'Antananarivo, Madagascar and the remaining parts were realized at the Université catholique de Louvain, except the DNA sequencing (GENEWIZ, Germany). The paper was realized with the collaboration of the following researchers:

Rasoamanalina Rivo Onisoa Léa¹, Mirzaei Khaled¹, Mondher El Jaziri², Tsiry Rasamiravaka³, Aro Vonjy Ramarosandratana⁴, Bertin Pierre¹

¹Earth and Life Institute - Agronomy - Université catholique de Louvain - Belgium

²Laboratory of Plant Biotechnology, Université libre de Bruxelles, Belgium

³ Fundamental and Applied Biochemistry Department, Université d'Antananarivo, Madagascar

⁴Department of Plant Biology and Ecology, Université d'Antananarivo, Antananarivo, Madagascar

Chapter 3

Genome-wide genetic variation and population structure in the cultivated vanilla of Madagascar revealed by ddRAD-seq

Abstract

The international vanilla market is largely dominated by Madagascar where the commercial vanilla – *Vanilla planifolia* – arose from a very small Mexican subpopulation that transited through the Reunion Island about 150 years ago. Some genotypes were also spread from an ancient breeding program that started in 1944 and may be still found within plantations, where some morphological variability may be encountered. To provide an overview on the extent of genetic variation of cultivated vanilla in Madagascar, 272 accessions from the four major areas of production, under different regional (SAVA, Ambanja, Mananara Nord and Manakara regions) and ecological distributions were collected. The accessions were characterized morphologically and ddRAD-seq protocol was used to generate genetic data. Five groups were identified on the basis of morphology: Vanille Banane, Big Vanilla, *V. planifolia*, *V. pompona* and a group intermediate to both previous ones. The 17,948 identified high-quality SNPs were proportionally distributed across the 14 chromosomes of *V. planifolia* with an average coverage of 43.8X and an overall high genetic diversity ($H_o = 0.44$). PCA and UPGMA tree analysis based on SNP data revealed that Vanille Banane and Big Vanilla were the most distant groups. Then the morphologically intermediate group was genetically separated into two sub-groups. Results highlighted differentiation between accessions related to their taxonomy, phenotype, regional and ecological distribution. Genetic structuration of accessions of *V. pompona* was in accordance to their region of provenance, and, in the SAVA, to altitude. Big Vanilla occurred only at low altitude of Mananara Nord, whereas the two intermediate sub-groups occur mostly at low and high altitudes of Manakara. We emphasized that the ancient breeding program

and anthropogenic mediation play an important role in the current state of distribution of genetic resources in Madagascar. This is the first genetic investigation of the cultivated vanilla in Madagascar providing high-quality SNPs and relevant data on genetic structure and diversity that will offer valuable information for future breeding and management programs.

3.1 Introduction

The genus *Vanilla* contains about 130 species which are essentially wild, with few cultivated species including *Vanilla pompona*, *V. tahitensis* and *V. planifolia*. The latter which is native to Central America is the most economically relevant. The public interest for vanilla has emerged in the early 1500s after the Spanish conquest of the Aztec and thereafter gained the pantropical distribution, subsequently to several attempts of cultivation in Europe (Lubinsky et al., 2010). This recent crop holds on the market the second rank of most expensive spices in the world after saffron (Parthasarathy et al., 2008). Since 1929, the international vanilla market has been largely dominated by Madagascar. The big Island benefits from ideal environmental conditions for vanilla cultivation in a large area extending from the south to the northeast, as well as in the north-western part. Vanilla is the most valuable crop for Madagascar, bringing 25% of income from the whole exportation value and 7% of gross domestic product (Ministry of Agriculture and Livestock and Fishing, MAEP 2018). Hence, much attention has been paid on the socio-economic aspects related to vanilla industry. However, genetic characterization of vanilla appears to be behind priorities.

The cultivated vanilla in Madagascar arose from a very small Mexican subpopulation of *V. planifolia* that was transited through Reunion Island (formerly Bourbon island) around 1819, from where a number of cuttings were first introduced into the surrounded Malagasy Islands, Nosy Bé (1870), then Sainte Marie (1880) before reaching in 1891 the SAVA region (made up of four districts: Sambava, Antalaha, Vohemar and Andapa) (Grisoni & Nany, 2021). These successive transfers of small number of cuttings of vanilla to Madagascar is in favour of possible bottlenecks within

the germplasm through time. Vanilla has always been ordinarily propagated vegetatively by cuttings and the dispersal pathway is farmer-to farmer without any particular organization for genetic diversity management. Between 1944 and 2000, a vanilla breeding center based in Ambohitsara (Antalaha district in SAVA region) in the North East has been active in enhancing the vanilla germplasm (Dequaire 1976; Tonnier, 1960). The main objectives of this ancient breeding program were to obtain disease-resistant, high vanillin content and pod indehiscent vanilla plants. More than 20,000 hybrids were created from inter- and intra-specific crossings of local and imported progenitors including *V. planifolia* for aroma, *V. phaeantha* and *V. pompona* for disease resistance, as well as *V. madagascariensis* for drought resistance and hardiness. Few other wild species are listed as progenitors such as *V. abundiflora*, *V. coursii*, *V. francoisii*, *V. walkeriae*, *V. zanzibarica* and *V. sp.* (Grisoni & Nany, 2021). Nevertheless, this list is still far from being complete as most documents containing information about crossings were hand-made and were lost during successive cyclones hitting and damaging the breeding center infrastructures (information received from the current local manager). About ten selected varieties were surveyed by multilocal tests within the SAVA region and the breeding center has retained the two most promising varieties for vanillin content and disease resistance, named as “Magnitry Ampotony” and “Tsitaitra”, respectively reported as hybrids of *V. planifolia* x *V. tahitensis* cv *Haapape* and *V. planifolia* x *V. pompona*. Presumably, varieties were released spontaneously among local farmers in the surroundings of the research center. Unfortunately, from the thousands of new generated lines, only a small number of poorly documented hybrids has remained in this breeding center. Thanks to the vanilla collection established by the Ramanandraibe Exportation SA (Known as Rama Export, a local company exporting vanilla) in Antalaha district, a limited number of remaining genotypes mainly collected from the abandoned vanilla breeding center as well as rare plants from other places within the SAVA region are now under conservation.

Morphological variability of the clonally propagated cultivated vanilla in Madagascar is found in the field for pod traits (length, size and vanillin con-

-tent) with existing variability from site to site and within a same vanilla plantation. However, as far as we know, no study has ever been conducted on the characterization of genetic resources of cultivated vanilla in Madagascar. Since vanilla is vegetatively propagated, variations due to genetic recombination are very scarce. The most likely hypothesis explaining the biological origin of this clonal diversity is the accumulation of point mutations. Previous studies using classical molecular marker systems such as isozymes (Nielsen et al., 1999; Nielsen 2000), AFLP (Duval et al., 2006; Bory et al., 2008b; Leperz et al., 2012), RAPD (Besse et al., 2004; Sreedhar et al., 2007; Minoo et al., 2008; Verma et al., 2009), ISSR (Sreedhar et al., 2007; Bory et al., 2008d; Verma et al., 2009; Fajardo et al., 2014; Ramos et al., 2017) and plastid genes (Cameron et al., 1999; Cameron 2004; Cameron et al., 2006; Bouetard et al., 2010) were not successful to estimate the genetic variability within *V. planifolia*, due to their technical limitation such as the low transferability and low representativeness of the genome. The endoreplication phenomenon (Brown et al., 2017) makes the diploid ($2n = 32$) or the polyploid genome in some vanilla cultivars) quite complex. This genetic complexity combined with the relatively long-life cycle of vanilla makes its genetic improvement difficult.

High throughput sequencing is nowadays available and offers the possibility of investigating a much higher proportion of the genome (Buermans et al., 2014). Next generation sequencing (NGS) approaches offer a large choice of platform for genomic studies including reduced representation sequencing libraries (RRSL) such as restriction site-associated DNA sequencing (RAD-seq, Andrews et al., 2016) and genotyping-by-sequencing (GBS, Elshire et al., 2011). In addition to the reduction of genome complexity by subsampling specific regions, its high throughput and cost-effective properties, make RRSL an attractive tool for genetic studies. RAD-seq and its derivatives have been successfully used on crop species with complex genome including sunflower, wheat, whitespot giant arum, sweet potatoes, and strawberries (Pegadaraju et al., 2013; Saintenac et al., 2013; Gao et al., 2017; Feng et al., 2020; Natarajan et al., 2020) and even in some wild orchids species such as *Dendrobium*

densiflorum, *Cymbidium aloifolium*, *Rhynchostylis retusa* and *Cypripedium calceolus* (Roy et al., 2017; Gargiulo et al., 2021). Still, this approach has never been tested to assess the genetic diversity of cultivated vanilla.

The objective of the present study was therefore to investigate the genetic structure and diversity of Malagasy cultivated vanilla at a national scale. Accordingly, a panel of 272 cultivated vanilla accessions covering the major vanilla production zones of Madagascar, under different geographical and ecological conditions, was sampled, mainly from farmer fields, and from the Rama Export collection and used for genome-wide analysis using double digest restriction associated DNA sequencing or ddRAD-seq (Peterson et al., 2012).

3.2 Materials and methods

3.2.1 Plant material

A diversity panel of 272 cultivated vanilla accessions was established and used for the present study. Among these, 235 were collected from farmer fields in the four major areas of vanilla production in Madagascar, and labelled in order to establish an *in situ* collection. The remaining 37 accessions were taken from the Rama Export *ex situ* collection. To cover the highest diversity as possible, an oriented sampling approach was applied in the field survey. Areas of study were distributed from the south to the north-eastern coast and the north western part of the country to cover the different environmental conditions encountered for vanilla cultivation in Madagascar (Figure 3.1). The four major vanilla production areas – SAVA, Manakara, Mananara Nord and Ambanja – were thus targeted for sampling based on our prior knowledge of the areas and the information provided by our key informant from Rama Export. Biological material was collected from a total of 120 plantations, where the per-zone 30 vanilla fields were distributed between low (0-100m), medium (100-150m) and high elevation (>150m), with 10 plantations each.

Identification of the different collected accessions was based on phenotypical traits including shape, color and size of the leaf and stem, ge-

-ometrical shape of transversal section of bean as well as flower color and size (Table 3.1, Table 3.2). According to their provenance, a number of synonymous local names exist for the different phenotypes. In the present study, a single appellation was kept for each type as used hereunder. Hence, accessions were preliminary grouped according to their morphology – those taxonomically distinct from each other such as *V. planifolia* and *V. pompona*, and accessions phenotypically distinguishable from each other as Vanille Banane, Big vanilla, Tsivaky and an intermediate group. This latter was assigned for accessions showing intermediate morphological traits between *V. planifolia* and *V. pompona*.

Table 3.1. Distribution and morphological group of the 272 collected vanilla accessions. Low (L), medium (M) and high altitude (H).

Geographical provenance		Morphological group					
		V. planifolia	V. pompona	Interme-diate	Big Vanilla	Vanille Banane	Tsivaky
Manakara	L	17	4	8	0	0	0
	M	12	2	1	0	0	0
	H	12	0	14	0	0	0
Mananara Nord	L	20	1	3	4	0	0
	M	18	2	0	0	0	0
	H	14	0	0	0	0	0
SAVA	L	14	4	0	0	0	0
	M	10	12	2	0	0	0
	H	12	4	0	0	0	0
Ambanja	L	15	2	3	0	0	0
	M	11	0	2	0	0	0
	H	12	0	0	0	0	0
Rama Export collection		24	4	7	0	1	1
Total		191	35	40	4	1	1

3.2 Materials and methods

Table 3.2. Morphological description of phenotypic groups.

Phenotype group	Leaves	Stem	Flowers	Pods
V. planifolia	Oblong, light green, 10 to 20 cm long 4 to 6 cm wide, flat surface	Light green, internode 10 to 15 cm	Yellowish green about 5 cm of length, green ovary pedicel	Cylindric, maturity 8 to 9 months
V. pompona	Lanceolate, dark green, long and large: up to 25 cm long and 10 cm wide, median groove emphasized	Dark green, internode up to 20 cm, vigorous	Pale yellowish, with the inside face trends to dark yellow, big size compared to that of V. planifolia	Triangular section, big size, low vanillin content compared to that of V. planifolia, maturity around 9 months
Intermediate group*	Light to dark green, short to long, from 8 to 20 cm long and 4 to 8 cm wide, thick and rigid, with medium grooves	Light to dark green, short to long from 6 to 15 cm, vigorous, resistant to <i>Fusarium</i> and poor environmental conditions	Color almost the same as that of V. planifolia but thicker and larger	Irregularly cylindric, short to long, vanillin content lower compared to that of V. planifolia, maturity 9 to 10 months
Big Vanilla	Light to dark green, long up to 30 cm and up to 11 cm wide, very thick and rigid, flat	Light to dark green, internode varying from 7 to 15 cm, section two to three-fold bigger that of V. planifolia	Dark yellow, thick, large and long (up to 9 cm)	-

*Also called as “Tsitaitra” (= “resistant plant”) from an ancient breeding (at Ambohitsara, Antalaha, SAVA) program and within its surrounding areas.

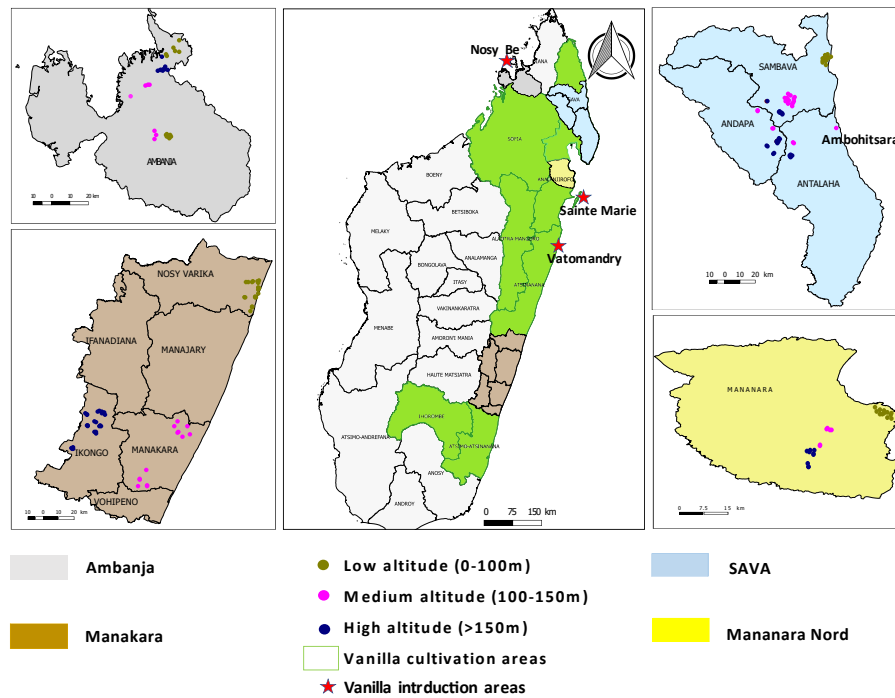


Figure 3.1. Sampling distribution of the 208 studied vanilla accessions from farmer fields. The colored delimited areas correspond to vanilla cultivation areas and the four major vanilla production zones corresponding to four administrative regions, are those colored in grey, brown, blue and yellow. Vanilla introduction areas to Madagascar are marked by red stars – Nosy Be (1870), Sainte Marie (1880) and Vatomandry (1891).

3.2.2 Genomic data

Genomic data preparation was similar as in the previous chapter. However, we chose the NovaSeq platform for sequencing after setting the best compromise between the number of studied samples, the amount of data and the cost. Briefly, total genomic DNA was isolated using Doyle's CTAB-based protocol (Doyle et Doyle, 1987). Double di-gest RAD libraries were prepared using *ECOR*I and *NLA*III restriction enzymes, according to the protocol described by Peterson et al., (2012). Prior to purification and size selection (300-600bp), barcoded adaptors were subsequently attached to the digested fragments using DNA ligase. Amplified fragments were then pooled into libraries and were sequenced on Illumina HiSeq 150 bp of length (GENEWIZ, Germany).

3.2.3 RAD analysis

Prior to genetic comparison, pooled reads were first separated by individuals through their respective dual index-barcode information using the *process_radtags* program from Stacks v2.4 (Catchen et al., 2011). Only reads with quality score above 30 with unambiguous barcode were retained for downstream analysis for a stringent data. Paired end reads were aligned to the published reference genome (Hasing et al., 2020) using Burrows-Wheeler Aligner (BWA) v0.7.15 (Li & Durbin, 2009a). BWA mem function was run with its default parameters as described by Rochette et al., (2017). SAMtools v.1.6 (Li et al., 2009b) was used to convert resulting aligned reads in SAM format into BAM format and sort reads by position for variant calling using BCFtools mpileup and BCFtools call v1.13 with default settings (Li et al, 2009b). Then Vcftools pipeline was used to filter raw genotype by setting the minimum allele frequency (MAF) at 0.04, minimum and maximum depth for a site 10 and 250 (min-mean-DP, max-mean-DP), minimum and maximum depth for a genotype per individual (minDP, maxDP) to 10 and 250 and only sites present in a least 80% of individuals were considered (r). Further SNP filtering was done to discard consecutive SNPs. All jobs for data processing were run on the CÉCI HPC (<https://www.cec-hpc.be>).

3.2.4 Population structure analysis

Principal component analysis (PCA), unweighted pair group method with arithmetic mean (UPGMA), and ADMIXTURE were used. For PCA, component contribution based on covariance matrix was computed by TASSEL v5 (Bradbury et al., 2007) and accessions distribution was then plotted using ggplot2 of RStudio v4.2.0 (Kronthaler et al., 2021). TASSEL was used for UPGMA to calculate the genetic similarity matrix between each sample based on the allele sharing distance IBS – a modified Euclidian distance – and the dendrogram was constructed to visualize genetic relationship between accessions. The tree was drawn with Itol v5 (<https://itol.embl.de>; Letunic et al., 2021) from the computed UPGMA dendrogram. To further study the population structure, ADMIXTURE was applied to estimate the membership coefficient of each accession to each group using the maximum likelihood estimates from multiple loci. For ac-

-cession grouping, hypothetical K ranging from 1 to 10 were tested with 1000 bootstrap replicates for Cross-Validation (CV) error estimation. The most appropriate K-value was obtained by assessing CV scores. The lowest CV value indicates the best K.

3.2.5 Genetic diversity and population differentiation

The per-group (based on PCA and UPGMA analysis) genetic diversity was assessed by the following metrics with GDA v1.1 (Lewis et al., 2002): expected and observed heterozygosities (H_e , H_o), fixation indices and inbreeding coefficient (F_{is}). A pairwise comparison of groups was performed to measure genetic distance through an exact F_{st} test (Weir et al., 2011) using Arlequin v3.5.2.2 (Excoffier et al., 2010), with 10 000 permutations and 0.05 allowed missing data level. The F_{st} result was visualized on a F_{st} heatmap constructed with RStudio v4.2.0 using corrplot package (Wei et al., 2017). To compare the hierarchical genetic variance at different levels, an analysis of molecular variance (AMOVA) was performed with Arlequin v3.5.2.2 (1000 permutation steps for significance level for variant components computation). The conversion of the filtered SNP dataset in VCF format containing SNPs (Danecek et al., 2011) to different formats for different purposes was done with PGDSpider v2.1.150 (Lischer 2011) whenever necessary.

3.2.6 Functional annotation of SNP associated genes

To investigate in which function the detected SNPs are involved, the genomic location and consequences of the variants were first detected using VEP (McLaren et al., 2016). We focused only on SNPs present in the coding regions as they are involved in protein structure, function and expression (Hunt et al., 2009). Missense SNPs flanking sequences were extracted from these coding regions. To explore the gene function of the region surrounding missense SNPs, gene including these variants were first extracted by running BLASTX (Kiesewetter et al., 1993) search against the NCBI non-redundant database (E-value $1e-6$) using BLAST+/2.11.0-gompi-2019b (Camacho et al., 2009) on CÉCI cluster. Gene function annotation was then conducted by BLAST2GO v6.0 (Conesa et al., 2008). Gene Ontology (GO) terms classifies gene function into biological processes, mo-

-lecular components and cellular components.

3.3 Results

3.3.1 ddRAD sequencing

The sequencing of the 272 accessions generated a total of 240.5 Gb containing 2.27 billion raw reads, corresponding to an average of 8.3 ± 4.6 million reads per individual (mean $\pm$ SD). After filtering for high base quality ($Q > 30$) and reads with unambiguous barcodes, 79.5% (1.8 billion) of high quality-reads corresponding to 6.6 ± 3.9 million per individual (mean $\pm$ SD) were retained. Twenty-six accessions with high rate of missing data were removed from the analysis. The 17,948 resulting high-quality SNPs from the remaining 246 accessions were proportionally distributed across the 14 chromosomes according to size (Annex, Figure B.3). The depth of coverage of each variant ranged from 15.5X to 247.8 X (Figure 3.2) with an average of 43.8X. The maximum SNP density corresponded to a depth coverage of 26X.

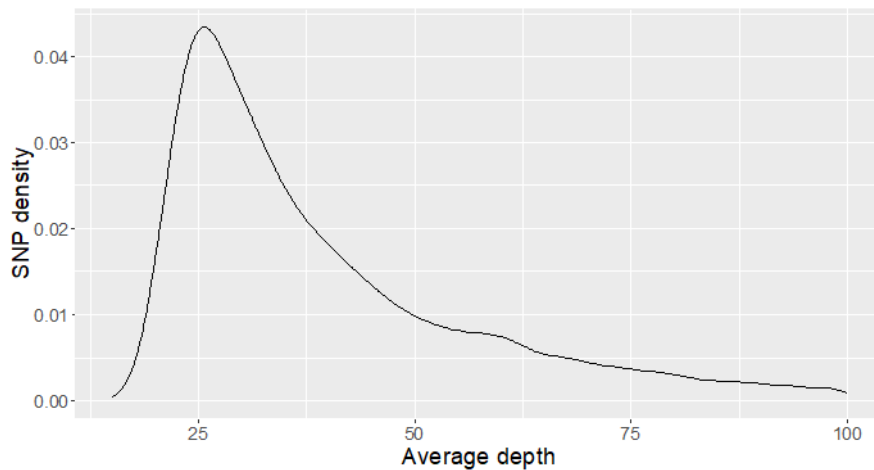


Figure 3.2. Distribution of density of SNPs by SNP coverage depth. The minimum depth was 15.5X and the mean 43.8X.

3.3.2 Genetic structure of the vanilla population in Madagascar

The 17,948 high-quality SNPs were used to infer genetic structure and relationship within the vanilla germplasm composed of 246 vanilla accessions. The two first components of the PCA explained 47.41 and 7.01% of the total variation, respectively (Figure 3.3). The 246 accessions were segregated into six distinct clusters. The phenotype Vanille Banane, a self-sterile plant within the Rama Export collection was clearly the most phenotypically different compared to the others and clustered far away from the other groups. Likewise, Big vanilla, collected from farmers, was also morphologically distinct from the others and clustered separately. All accessions phenotypically classified as *V. pompona* were clustered together, and so did the accessions of *V. planifolia*. However, the accessions phenotypically classified as intermediate – presumably known as Tsitaitra in the surroundings of the ancient research station, including the Rama Export collection – were divided into two groups (intermediate 1 and intermediate 2).

The results from UPGMA (Figure 3.4) were completely consistent with the PCA where, first, Vanille Banane and Big Vanilla were denoted as the two most distant clusters from the others. Then UPGMA separated the *V. pompona* accessions, followed by the two intermediate phenotypic groups and finally all the *V. planifolia* accessions, including Tsivaky. In *V. pompona*, the accessions were clearly clustered according to their region of provenance and, inside the SAVA region, according to the altitude. Such a correspondence between geographical and genetical data was not found in the other phenotypic groups.

According to the cross validation (CV) error plot, the optimal number of ancestral groups in ADMIXTURE was K=3 (Figure 3.5-a). The first group (1, green) comprised all the *V. planifolia* accessions, the second one (2, blue), the *V. pompona* accessions, and the third one (3, orange), the big Vanilla accessions. The intermediate clusters 1 and 2 of the PCA and UPGMA methods were most often a mixture of ADMIXTURE groups 1 and 3. With k=4, the group 1 *i.e.* *V. planifolia* obtained with K=3 was separated into two

sub-groups, suggesting the possible occurrence of at least two genetic origins in the collected *V. planifolia* accessions.

The intermediates clusters 1 and 2 revealed by PCA and UPGMA appeared to be admixed between *V. planifolia* and Big Vanilla, and between the two *V. planifolia* groups and *V. pompona*, respectively. Similarly, Vanille Banane appeared to share its genetic background between *V. planifolia* and Big Vanilla (Figure 3.5-b). The phenotype Tsivaky always grouped together with the *V. planifolia* group in ADMIXTURE.

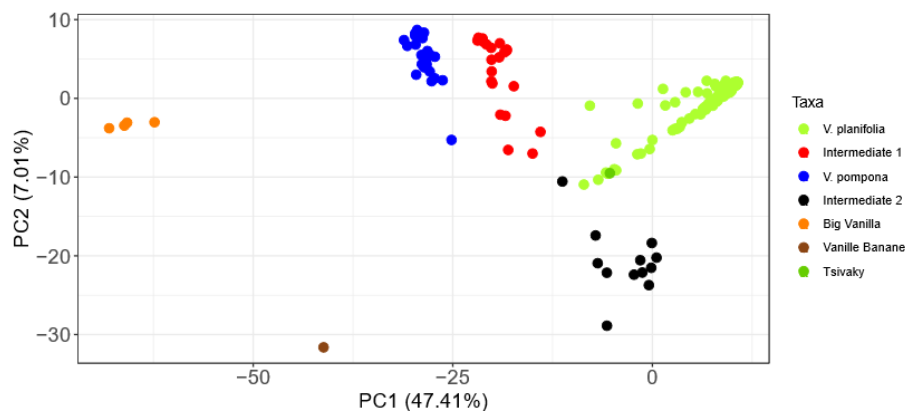


Figure 3.3. Principal Component Analysis (PCA) of the 246 accessions according to the two first eigenvectors with the 17,946 markers. Color attribution was based on phenotype identification, with the intermediate phenotypic group further split in two colors on the basis of the PCA.

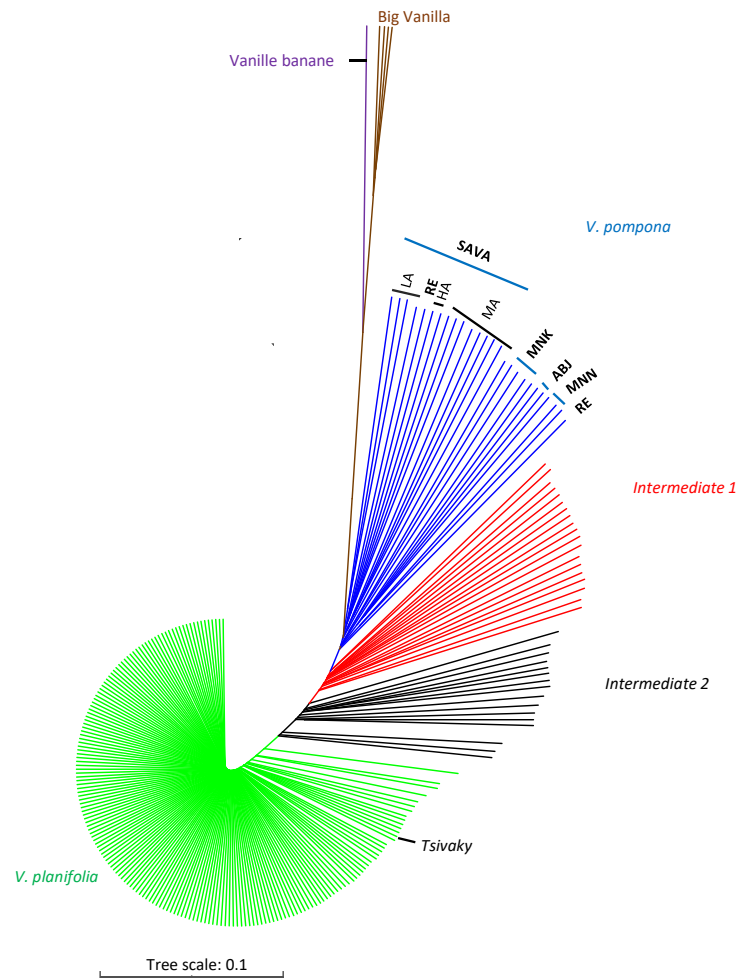


Figure 3.4. Dendrogram generated by UPGMA cluster analysis illustrating genetic relationship among the analysed vanilla germplasm composed of 246 accessions. Manakara (MNK), Ambanja (ABJ), Mananara nord (MNN), Rama Export collection (RE), low altitude (LA), medium altitude (MA), high altitude (HA). Color attribution was based on phenotype identification, with the intermediate phenotypic group further split in two colors on the basis of the PCA and UPGMA.

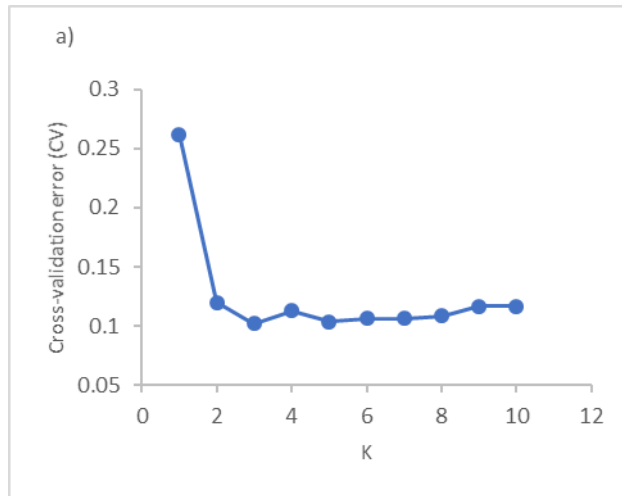


Figure 3.5a. Population structure of the 246 vanilla accessions based on the 17,948 SNP markers. Cross validation (CV) error plot to identify the optimal number of ancestral groups. The lowest CV value indicates the optimal value of K, $K = 3$. Here $K=4$ fitted better to the data.

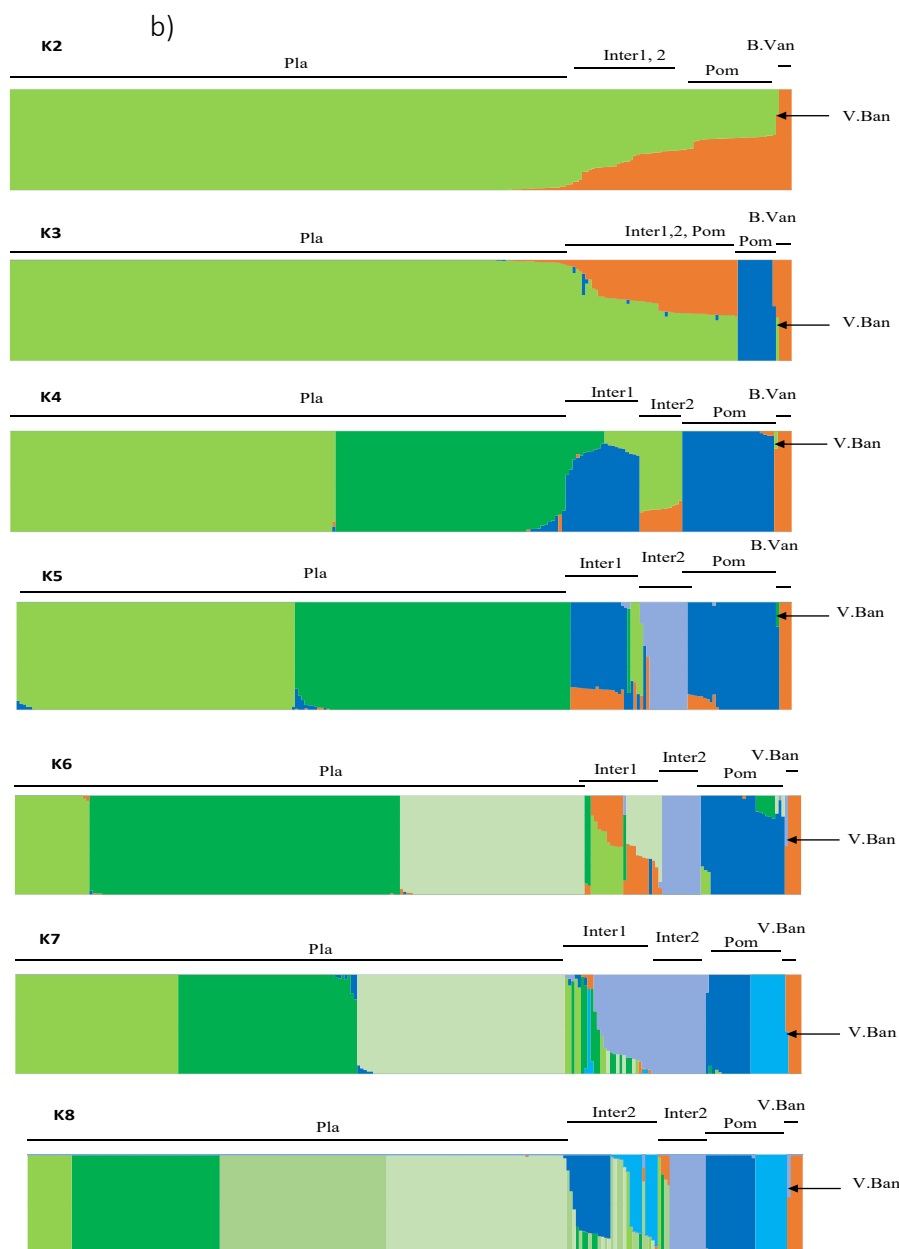


Figure 3.5b. Population structure of the 246 vanilla accessions based on the 17,948 SNP markers. Plot of membership coefficients. Each accession is represented by a vertical line where the color proportion corresponds to its membership coefficient to a group (ADMIXTURE). The classification into Pla (*V. planifolia*), Pom (*V. pompona*), V.Ban (Vanille Banane), BigV (Big Vanilla), Inter1 (Intermediate 1), Inter2 (Intermediate2) was based on phenotypic data, PCA and UPGMA results.

3.3.3 Geographical distribution of species and cultivars

Provenance-based distribution was observed in identified genetic groups through genetic structure analysis methods such as PCA, UPGMA and ADMIXTURE, except for the *V. planifolia* group. Following these analyses, *V. planifolia* was the largest group and distributed in all regions without geographical structuration based on the identified SNPs. Big Vanilla was mainly observed in the low altitude of Mananara Nord.

Genetic variation within the *V. pompona* genetic group was structured according to their geographical and ecological distribution. Samples from the four sites of study were separated from each other and, for the SAVA region, genetic variation of accessions was in agreement to their distribution among the three altitudinal levels (Figure 3.4).

The two intermediate groups (Figure 3.5-b) were mostly found in the Manakara Region. The group intermediate 1 occurred mostly at the low altitude and intermediate 2 only at high altitude.

3.3.4 Genetic diversity and differentiation among vanilla genetic groups revealed by PCA and UPGMA

The variance among and within genetic groups revealed by PCA and UPGMA was evaluated using AMOVA based on the entire set of genotyped SNPs (Table 3.3). Vanille Banane and Tsivaky were not included in the following analysis as each contained a single individual only. This analysis highlighted a much higher genetic variance within (72.0%) than among groups (28.0%).

Table 3.3. Analysis of molecular variance among the five groups detected by PCA and UPGMA. Data from 245 vanilla accessions and 17,948 SNPs.

Source of variation	degree of freedom	Sum of squares	Variance components	Percentage of variation (%)	Fixation indices	p-values
Among groups	4	36171.9	168.9	28.0	0.28	0.000+-0.000
Within groups	485	196120.8	434.8	72.0	-	-

To compute the genetic distances between the genetic groups identified by PCA and UPGMA, the pairwise F_{st} was applied (Table 3.4). Again, Vanille Banane and Tsivaky were not included. The weighted F_{st} values were all highly significant (p -values = 0). The genetic distance between *V. pompona* and *V. planifolia* was intermediate (0.32). Similarly to the pattern from PCA and UPGMA tree, Big Vanilla was the most distant group with the highest F_{st} values when compared to the remaining groups, where the maximum was with *V. planifolia* (F_{st} = 0.68). On the opposite, *V. pompona* and the intermediate 1 showed the lowest genetic distance (F_{st} = 0.014)

Table 3.4. Pairwise comparison (F_{st}) between groups using the 17,948 SNPs. Groups were defined according to PCA (Figure 3.3) and UPGMA (Figure 3.4) analyses.

Genetic group	<i>V. planifolia</i>	<i>V. pompona</i>	Intermediate 1	Intermediate 2	Big Vanilla
<i>V. planifolia</i>	0				
<i>V. pompona</i>	0.32*	0			
Intermediate 1	0.24*	0.01*	0		
Intermediate 2	0.16*	0.15*	0.12*	0	
Big Vanilla	0.69*	0.20*	0.30*	0.52*	0

* p -value=0

Finally and with respect to genetic diversity, a focus was made on the heterozygosity. The highest genetic diversity was observed in *V. pompona* ($H_o=0.62$) while Big Vanilla had the lowest ($H_o=0.29$) (Table 3.5). The *V. planifolia* group showed a moderate value ($H_o=0.31$). A pronounced excess of observed heterozygosity compared to the expected one was detected within all groups, *i.e.* more heterozygous individuals than expected under the assumptions of the Hardy-Weinberg equilibrium. Likewise, the coefficient of inbreeding was negative (Table 3.5).

Table 3.5. Descriptive statistics of genetic diversity within each group defined by PCA and UPGMA. (N) number of accessions, (He) expected heterozygosity, (Ho) observed heterozygosity, (FIS) inbreeding coefficient.

Genetic groups	N	He	Ho	FIS
<i>V.planifolia</i>	62	0. 17	0. 31	-0.84
<i>V.pompona</i>	26	0. 37	0. 62	-0.68
Intermediate 1	19	0.35	0. 56	-0.66
Intermediate 2	16	0.29	0. 44	-0.56
Big Vanilla	4	0.20	0. 29	-0.27
Mean	-	0. 27	0. 44	-0.62

3.3.5 Functional annotation

From the 17,948 identified SNPs, 3,367 (18,8 %) were located into genic regions, including missense (8.4%), splice region variants (0.9%) and synonymous variants (9.7%) (Figure 3.6). The 1,506 missenses SNPs were used for functional annotation as they present mutations within coding regions and therefore might affect amino acid sequence and hence induce structural and/or functional changes in the resulting protein. Flanking regions around these variants were extracted from 150 base length from both sides. In total, 1,192 (79.34%) sequences were successfully blasted against the non-redundant NCBI database with the e-value threshold of 10^{-6} from which 750 matched to the Blast2GO database and had one or more gene ontology (GO) annotations from the same database. The remaining amount not blasted are presumably specific to the *V. planifolia* genome.

The top hits species distribution from BLASTX searches of homologous protein sequences in UniProt were dominated by the Orchidaceae family with *Dendrobium catenatum* (556 hits) and *Phalaenopsis equestris* (287 hits). Note that no *V. planifolia* samples were present in this database. Other monocot species had few hits such as *Elaeis guineensis* (62) and *Phoenix dactylifera* (47), *Ananas comosus* (24), *Musa acuminata* (22), *Asparagus officinalis* (17) and *Oryza sativa japonica* (8) (Figure 3.7).

Annotated sequences were involved in three functional classes based on the associated GO terms: 565 sequences involved in biological processes, 617 related to molecular functions and 556 linked to cellular components. For biological processes, the analysis highlighted metabolic processes involving macromolecules (232), compounds (149) and cellular nitrogen compounds (139). Interestingly, as far as aroma of vanilla beans is concerned, 128 genes were connected to metabolic processes involved in aromatic compounds. The molecular functions comprised catalytic (transferase activity, hydrolase activity...) and binding (organic cyclic compound binding, heterocyclic compound binding ...) activities. Finally, the cellular components were distributed in organelles, membranes, intracellular anatomical structures and cytoplasm parts (Figure 3.8).

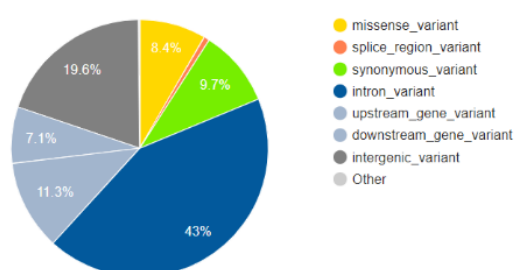


Figure 3.6. Genomic location of the 17,948 SNPs.

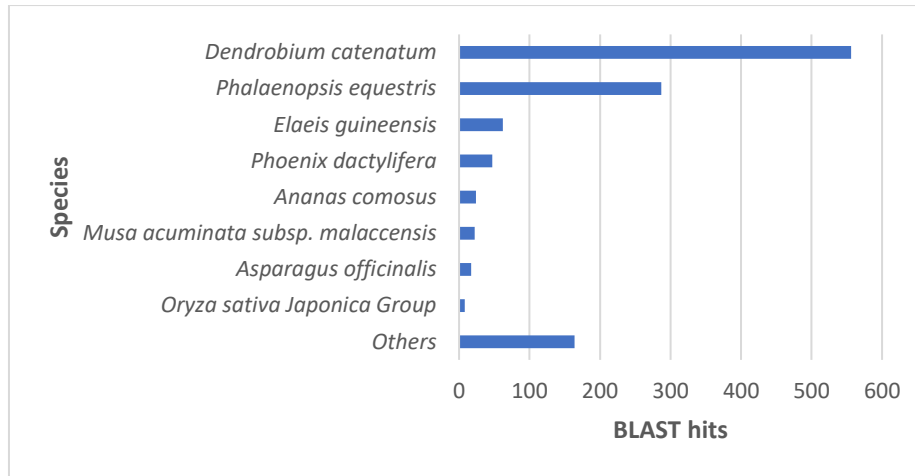


Figure 3.7. Top-BLAST hits species distribution.

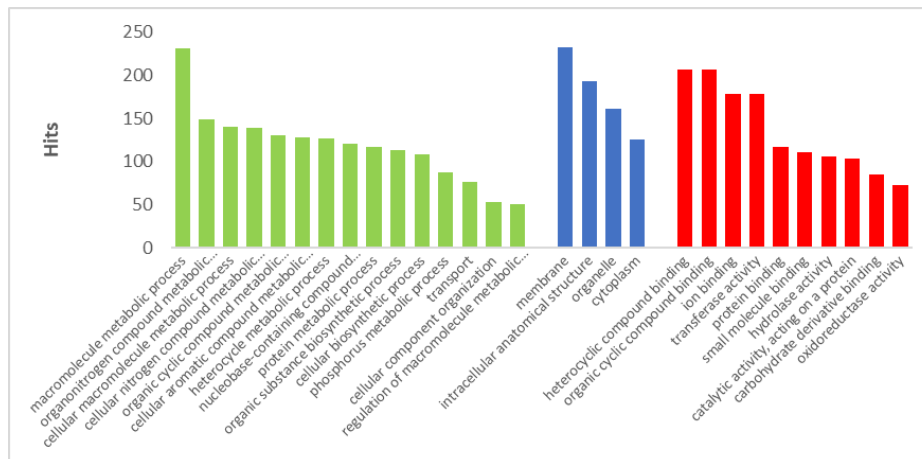


Figure 3.8. Distribution of Gene Ontology (GO) classification of missense SNPs flanking regions. GO terms were classified into three functional classes: biological process in green (565 sequences), molecular functions in blue (617) and cellular components in red (556).

3.4 Discussion

3.4.1 ddRAD-seq, a suitable tool for vanilla genetic resources characterization

ddRAD-seq belongs to the RAD-seq family that enables the sequencing of targeted genome fraction and the identification of nucleotides polymorphism in a large amount of samples at once (Peterson et al., 2012;

Puritz et al., 2014, Kess et al., 2016). Recently, RAD-seq have been widely used in more than 23 species, including a number of important crop species, such as wheat, sunflower, rice, tomato etc... (Saintenac et al. 2013; Pegadaraju et al. 2013; Fujino et al., 2022; Esposito et al., 2020). This tool was demonstrated to be powerful in population genetic studies (Gompert et al., 2010; Catchen et al., 2013b). In the present study, SNPs were filtered with minimum depth of 10X. The minimum coverage of the 17,948 resulting SNPs was 15x (higher than the threshold set), the average 43.8X and the maximum SNP density was found at about 26X coverage (Figure 3.2). The high-quality score of reads (Q>30) used for analysis should be one of the main key parameters of this result. In addition, these SNPs were proportionally distributed among chromosomes (Annex, Figure B.3). These results highlight the effectiveness of ddRAD-seq approach to investigate large-scale vanilla study.

3.4.2 Vanilla population in Madagascar is segregated into six genetic groups

Genetic variability assessment is the basis for any breeding program implementation, successful conservation and efficient use of plant genetic resources. With the advance of technologies, molecular markers offer a range of choices to distinguish individuals from each other. Compared to other crop species, *in situ* cultivated vanilla suffers from a lack of genetic knowledge. Most efforts have been mainly devoted to accessions from collections (Minoo et al., 2008; Verma et al., 2009; Alomia et al., 2021) and data on genetic resources of field cultivated vanilla are missing, especially in Madagascar where no genetic study has been conducted since its introduction from Reunion island. To investigate the population genetics resources of the cultivated vanilla in Madagascar, a panel of 246 accessions collected from the major vanilla production areas was sequenced using ddRAD protocol. Sequences were compared between all individuals that provided 17,948 high quality polymorphic variants used for the genetic differentiation exploration. Genetic analysis clustered the analysed cultivated vanilla germplasm into six groups, *V. planifolia*, *V. pompona*, Vanille Banane, Big Vanilla and two intermediate groups.

V. pompona and *V. planifolia* clusters (Figures 3.3, 3.4) were represented by accessions from everywhere at different altitude levels. The occurrence of *V. planifolia* in all regions is not surprising as it is cultivated for commercial purpose and cuttings moved from one place to another without any particular management. Breeding strategies might be one of the causes of segregating pattern of *V. planifolia* accessions revealed by ADMIXTURE (Figure 3.5-b, K=4). As breeders seek more diverse individuals in mating materials, it is likely that several hybridization events occurred, involving crosses between several parents of *V. planifolia*, *V. pompona* and possibly close relatives, resulting in segregating individuals and separate groups. However, the two intermediate groups are very close to each other with an F_{st} close to zero. Individuals within *V. planifolia* group appeared to be very similar at the genetic level (Figure 3.3, 3.4), as reported by previous studies in reunion Island (Bory et al., 2008c). Vegetative propagation way may have contributed to maintain a high level of homogeneity of the commercial vanilla at the population level.

Beside *V. planifolia*, *V. pompona* was also detected in almost all regions, with lower representation, most frequently however in the SAVA region. This species has been documented as one of the introduced progenitors during the breeding program at the Ambohitsara breeding station (Dequaire 1976, Grisoni & Nany, 2021) and a possible release of cuttings from the research centre probably happened from close to distant areas. This pattern shows a model of vanilla cutting dispersal due to human mediation in Madagascar without any formal configuration but done farmer by farmer which can reach even the southern part of the Island. The distribution of *V. pompona* accessions indicated geographical and environmental distribution related to genetic variation as shown by the UPGMA tree (Figure 3.4). Investigating the link between genetic variation and ecological parameters may hence be interesting to understand vanilla growth requirements.

Phenotypically, Big Vanilla is morphologically distinct from *V. planifolia* and *V. pompona* with leaf, stem and flower sizes about two-folds bigger from which it took its local name. The large flower is dark yellowish releasing long and larger beans than that of the two other species with, respectively,

yellowish green and pale yellowish flowers. Big Vanilla phenotype is similar to that of *Vanilla palmarum* (Soto-Arnas et al., 2013), a vanilla species presents in the South of America as epiphyte plant of palms and was reported to early diverge from other American species such as *V. planifolia* and *V. pompona* (Bory et al., 2008d, Arenas et al., 2013). The relationship between Big Vanilla, *V. planifolia* and *V. pompona* revealed by the 17,948 SNPs is consistent with the previous reports with *V. palmarum*, *V. planifolia* and *V. pompona* using AFLP and microsatellite markers (Duval et al., 2006; Bory et al., 2008b; Bory et al., 2008d). Curiously, farmers pointed the occurrence of beans without hand-pollination for Big Vanilla as the released flowers are inaccessible (epiphyte plant of high tree) and the same feature was reported for *V. palmarum* as the rare Vanilla species mainly self-pollinated and producing banana-like beans (Arenas et al., 2013). However, *V. palmarum* is not listed in the introduced progenitors in Madagascar (Grisnoni et Nany., 2021). Interestingly, the Vanille Banane phenotype from the Rama Export collection was grouped close to Big Vanilla group, and is shown to share about 75% of its genetic background from this group and the remaining from *V. planifolia* (Figure, 3.5-b). Morphologically, Big Vanilla and Vanille Banane are not very different from each other, but Vanille Banane is self-sterile, possibly resulting from a crossing between *V. planifolia* and the possible polyploid Big Vanilla. Hence, Big Vanilla might be one of the introduced genitors used during the breeding program in Madagascar but possibly missing in the available documents (Grisoni et al., 2021). Genetic comparison between Big Vanilla and *V. palmarum* is required to confirm the present hypothesis. The narrow distribution of Big Vanilla could result from choices of farmers regarding the economic value of this species. However, the self-pollination trait could be economically potential and useful in breeding program to reduce bean production cost.

Likewise Big Vanilla, the intermediate groups were seen over limited and specific geographic ranges mainly in the low (Intermediate 1) and particularly in high (Intermediate 2) altitudes of Manakara. Different hypotheses could raise from the genetic pattern of these admixed groups. First, based on their respective genetic backgrounds revealed by ADMIX-

-TURE (Figure 3.5), recent gene flow between *V. planifolia* and the two other groups (*V. pompona* and Big Vanilla) could be an explanation. Rare natural pollination of *V. planifolia* by social bees (*Melipona* sp.) and hummingbirds (*Cynnis* sp.) has been reported (Donovan 1948) in Mexico where the pollination rate is very low. The lack of these specific pollinators in vanilla area of introduction does not favour this hypothesis in the present result. However, Big Vanilla was informed to produce very few mature beans without human contribution. A possible natural mating hence, partially occurs. Still, this needs to be taken with caution and be confirmed by assessment. Nevertheless, from all accessions collected in Mananara Nord, no one belonging to the Intermediate 2 group was tracked around this area (at least from our sampling), this admixed group occurs specifically in the south-eastern part of Madagascar. Accordingly, gene flow provides low support for this hypothesis. Second, a computation limitation may also have been involved in this pattern. Since few accessions are included in the intermediate groups, a bias could have occurred and could have impacted sample assignment to one or the other group. However, the geographical distribution of accessions of each group is very specific, and the same outcome was shown by PCA (Figure 3.3) and UPGMA (Figure 3.4). Consequently, sampling bias does not appear to justify the present pattern.

The accession labelled as “Tsitaïtra” (= “resistant plant”) from the breeding program fell into both intermediate groups. Basically, Tsitaïtra is one of the outcomes of several years of work from the vanilla breeding station in Ambohitsara to introduce the resistance potentiality of *V. pompona* into the widely commercial *V. planifolia* (Bouriquet 1946; Grisoni & Nany, 2021). But, after the loss and disorder caused by the almost yearly cyclone hitting Madagascar combined to the poor available document, initial label is lost and all hybrids are now designed as Tsitaïtra. This later is characterized by robustness and overall thicker and larger size with lower aromatic profile compared to *V. planifolia* and was documented as a back cross of the hybrid *V. planifolia* x *V. pompona* with *V. planifolia* (Grisoni & Nany, 2021). The genetic profile of the Intermediate 1 genetic group suggests that they are likely back-crossed with *V. pompona*. Different cross-

-sing strategies were tested (personal communication from ancient technician involved in the program) between the two parents and possibly most of remaining cultivated Tsitaitra are the one close to *V. pompona*. And, the occurrence of the so-called Tsitaitra in the Intermediate 2 group confirms that Tsitaitra is not necessarily a hybrid from *V. planifolia* and *V. pompona*. We therefore favour the hypothesis of the breeding program history coupled to anthropogenic mediation as the main driver of the distribution pattern. Considering the long list of introduced genotypes and the number of resulting hybrids from the breeding program, a huge part of genotype is presumably lost. Conservation of the remaining genotypes is therefore urgent.

3.4.3 High diversity within the genetic groups of vanilla cultivated in Madagascar

PCA and UPGMA tree (Figures 3.3 and 3.4) clustered the vanilla cultivated in Madagascar into six major groups: *V. planifolia*, *V. pompona*, Vanille Banane, Big Vanilla and two intermediate groups. Weighted F_{st} was used to further validate the genetic similarity between the defined groups. Frankham et al., 2002 considered a F_{st} value greater than 0.15 as a significant threshold to separate two groups from each other. F_{st} values (Table 3.4) confirmed most of the groups of cultivated vanilla in Madagascar identified through PCA and UPGMA approaches and confirmed the distant relationship of Big Vanilla with the remaining groups with high values of F_{st} . AMOVA indicated that most of the variation occurred within (72.0%) than among groups (18.0%) with high genetic diversity ($H_o = 0.44$) (Tables 3.3, 3.5).

Diversity of *V. planifolia* in Madagascar is much higher ($H_o=0.44$) than that reported in America ($H_o=0.12$, Hu et al., 2019) and slightly higher than that found in the in Reunion Island ($H_o=0.28$, Favre et al., 2022) Similarly, *V. pompona* and the intermediate group 1 showed so far higher values ($H_o=0.62$, $H_o=0.56$, respectively) of genetic diversity than those reported outside, where the H_o were equal to 0.02 (Favre et al. 2022) and 0.15 (Hu et al. 2019) (Table 3.5). For all groups, a noticeable excess of heterozygosity was observed. Expected heterozygosity was lower than the

observed one, likewise the inbreeding coefficient was always negative. Our result is consistent with the excess of heterozygosity statement in most cultivated plants with vegetative way of multiplication (Elias et al., 2001; Ge et al., 2005), and for vanilla (Philip et al., 1986; Hasing et al., 2020). Several studies pointed the asexual reproduction of vanilla as a possible explanation of the excess of heterozygotes through which accumulation of mutations across generations maintain or increase heterozygosity level that will be fixed until sexual reproduction occurs (Balloux et al., 2003; Alberto et al., 2005; Ruggiero et al., 2005).

3.5 Conclusions

We demonstrated that ddRAD-seq is an efficient approach for vanilla study. The 17,948 detected genome-wide SNPs provide useful information for genetic research on cultivated vanilla and helpful for a better knowledge of the genetic resources of the Malagasy vanilla. The present result could be used as starting point and tool for the next step in conservation strategies implementation and future breeding program.

To further understand the extent of local adaptation in the genetic diversity of the cultivated vanilla in Madagascar, a study investigating the contribution of environment and spatial distribution on SNP variation was carried out on the set of accessions analysed in Chapter 3, except those from the collection, as reported in the following chapter. We used remote sensed-based data, using geographical coordinates of each accession for the environmental variables data acquisition. This study will provide additional information for future conservation planning.

Chapter 4

Low influence of environment and spatial distribution on genetic variability of cultivated vanilla in Madagascar

Abstract

Cultivated vanilla is originated from Mexico and has been introduced into many other tropical areas. Few cuttings of cultivated vanilla were introduced to Madagascar via Reunion Island, around 1870. Despite its narrow genetic basis, its recent introduction in Madagascar and its mode of propagation by cuttings, some phenotype and vanillin content variations have been reported in the country. Vanilla has recently benefited studies related to genetics and aromatic compounds. However, the relationship between genetic variation and environment has not been investigated. This approach could help in conservation planning and for sustainable production. To quantify the contribution of environment and spatial distribution in genetic diversity of the cultivated vanilla in Madagascar, 208 accessions were collected from farmers' fields, within the four major areas of vanilla cultivation in the country. The studied samples belong to five previously recognized genetic groups: *Vanilla planifolia*, *V. pompona*, Big Vanilla and two intermediate groups, presumably of hybrid origin. No genetic difference was detected between the four zones among the *V. planifolia* accessions, suggesting that a sampling drift hypothesis from the location of the original introduction could not explain the SNP variation. Only about 6.2 % of the genetic variation was detected to be related to environment and space. A redundancy analysis (RDA) was then applied on the 208 accessions as well as on the subset of 154 *V. planifolia* accessions, with both spatial and environmental variables taken as possible explanatory variables of SNP variation. When spatial distribution was taken alone, again no relationship with SNP variation was detected in the 208 pooled accessions. However, the SNP variation was found to be weakly correlated with spatial variables

when applied to *V. planifolia* alone. When environment variables were taken alone, a small but significant pattern of local adaptation was detected in both accession samples. This pattern was further enhanced when environmental variables were constrained by spatial distribution. Precipitation variables were identified as the key drivers of *V. planifolia* SNP variation. SNP variation within the cultivated vanilla in Madagascar could therefore be partially attributed to environmental variation, and the introduced vanilla is likely a pooled population within which spatial structuration is challenging. The present results increase our knowledge of the genetic diversity and evolution of the cultivated vanilla in Madagascar.

4.1 Introduction

Understanding the evolutionary response of species to environmental changes is fundamental to conservation strategies. According to the theory of evolution, spatially heterogeneous environments can promote local adaptation of genotypes (Leimu & Fischer, 2008). Selection by environmental factors may influence the genetic divergence of species (Wang & Bradburd, 2014). Evidence of local adaptation has been demonstrated in both animals (Fraser et al., 2011; Wei et al., 2020) and plants (Leimu & Fischer, 2008). For crop plants, most efforts have been devoted to the wild-related species (Pournosrat et al., 2018; Gibson & Moyle, 2020; Chang et al., 2022). Despite the few investigations conducted on crops, these later were however taken as a model to understand adaptation and diversification (Olsen & Wendel, 2013; Shaaf et al., 2014). Investigating the adaptation of crop species growing under different conditions will help selecting the most suitable genotypes, therefore mitigating the impact of climate change on agriculture and ensuring sustainable production. The cultivated vanilla can be used for such studies as it is an introduced plant from a single environment into many other tropical areas. Recently, cultivated vanilla has benefited studies related to genetics, such as diversity studies and the reference genome sequencing (Sreedhar et al., 2007; Ramos-Castellá et al., 2017; Hasing et al., 2020; Chambers et al., 2021; Favre et al., 2022) and to the chemistry of its aromatic compounds (Boyce et al., 2003; Rao et al., 2014; Yang et al.,

2017). However, the investigation on the relationship between genetic variation and eco-geography appears to be left behind. The cultivated vanillas, mostly *Vanilla planifolia* with a few other species – *V. pompona*, *V. tahitensis* – belong to the genus *Vanilla* within the *Orchidaceae* family. As with other orchids, a partially shady area (about 50%, Quiro, 2018) is the most suitable habitat for proper development. Vanilla is a vine plant that prefers warm, (24°C to 30°C, Rao & Ravishankar, 2000) and moist (80%> humidity, Baqueiro-Peña & Guerrero-Beltrán, 2017) environments. It needs between 1,500–3,000 mm of well distributed annual rainfall (Azeez, 2008). It can grow easily from 0 to 1500 masl on various substrates, generally humid soils (basically sandy and lateritic) containing humus.

Madagascar is one of the major vanilla producers across the world. The large optimal areas for vanilla cultivation, coupled with the high number of actively involved farmers in the bean production, play an important role in the position of Madagascar on the global vanilla market. Madagascar has been considered a model area for wild species diversification studies due to the richness of environmental gradients between regions (Vences et al., 2009). The particularity of Madagascar relies mainly on its biogeographical isolation and the topography that congruently shape contrasting climate and vegetation. Madagascar benefits humidity from the Indian ocean which invades the eastern slopes of the mountain chains that separates the island in two major latitudinal bioclimatic zones (Vences et al., 2009) : a tropical rainforest stretching on the eastern part and a tropical dry vegetation (forest, bush, spines) – with mangroves along the coastal area – on the western part, with an exception on the north west that benefits the humidity from the wind from the western sea with a humid forest. Many studies were hence conducted on diversification pattern of species in Madagascar (Andrianoelina et al., 2006; Vences et al., 2009; Brown et al., 2015; Andriamihaja et al., 2021). As a humidity-dependant plant, vanilla is now cultivated eastwards from the south to the northeast, and in the particularly humid area of the north-west.

As previously mentioned (Chapter 1, 1.4.1), the cultivated *V. planifolia* in Madagascar arose from a Mexican subpopulation that transited through Reunion Island. The three main known introductions are (Figure 4.1): first,

via the north-western Malagasy island Nosy Be in 1870; second, through the east coast island of Sainte Marie in 1880; third, the middle east coast in Vatomandry in 1891 (Grisoni & Nany, 2021). The arrival time of cuttings in the SAVA – the main zone of cultivation – is not clear, but presumably happened during the colonial period (1897–1946) during which the SAVA region was targeted for vanilla production. A partial indication of the *V. planifolia* origin is therefore known. However, information regarding the genetic bases of the introduced genotypes is still missing. Introduced cuttings presumably arose from different genotypes being part of the original Mexican gene pool. In addition, during the pioneering vanilla breeding program that was established in Ambohitsara, Antalaha – SAVA region, from 1950 to 2000 to control vanilla diseases (Bouriquet, 1946; Tonnier, 1960; Dequaire, 1976), several genotypes belonging to several species were imported. These genotypes and their hybridization with the local genotypes have enriched the vanilla genetic pool in Madagascar. The vanilla cropping is now extensively practiced along the eastern coast of the country and in the north-west under different climatic and topographic conditions. The dispersal of the cuttings has been done spontaneously by the farmers. However, there is a gap in the data regarding the extent of this phenomenon, the number of farmers involved, the amount and the uniformity of the starting genetic material that has been transported to the different regions.

The path of vanilla introduction in Madagascar by subsequent subsampling should have favoured a bottleneck in the genetic diversity of *V. planifolia*. This may lead to reduced genetic variation within the imported *V. planifolia* in Madagascar (as shown in Chapter 3) and to decrease the species' capacity for local adaptation under selection pressure. Recently introduced populations with a small number of individuals are particularly susceptible to random genetic drift at the beginning of introduction due to their small effective size (Allendorf, 1986). Introduced species have been shown in the past to frequently adapt under novel environmental settings, which is also true for vegetatively propagated species (Oduor et al., 2016; Simón-Porcar et al., 2021; Liu et al., 2020), making them a good model for investigations on local adaptation that takes place over brief time scales

(Colautti & Lau, 2015). In addition, the genetic variation of vanilla between sites in Madagascar may be influenced by human intervention in vanilla cutting dispersal, which promotes genotypes differently in different regions. Understanding the evolution pattern of diversity *V. planifolia* since its introduction in Madagascar will increase the necessary knowledge for the conservation planning of the crop species.

The scope of the present study was to assess the contribution of environment and spatial distribution in SNP variation within the cultivated vanilla in Madagascar. To do so, 208 accessions collected in farmer fields from the four major areas of vanilla cultivation – SAVA (a combination of four districts – Sambava, Antalaha, Vohemar, and Andapa), Mananara Nord, Manakara, and Ambanja – were used. The studied samples belong to five previously recognized genetic groups: *V. planifolia*, *V. pompona*, Big Vanilla and two intermediate groups presumably of hybrid origin (Chapter 3). Genetic distances between accessions from the studied zones were first computed. Then, redundancy analysis (RDA), a multivariate pipeline that allows to study the relationship between a multivariate response variable matrix and one or several multivariate explanatory matrices of variables, was used. This approach was successfully used in several investigations of isolation by distance (IBD) and isolation by environment (IBE) on genetic variation (Lasky et al., 2012; Riordan et al., 2016; Wanderley et al., 2018; Chang et al., 2022; Gao et al., 2021). First, we looked at the independent contributions of spatial distribution and environmental variables in SNP variation. Then, we identified their colinear effect.

4.2 Materials and methods

4.2.1 Genetic data

In order to cover an as much heterogenous environment as possible and collect much diversity within the cultivated vanilla of the country, sampling sites *i.e.* farmer fields were distributed within the four major vanilla production areas (zones), corresponding to administrative regions: SAVA, Mananara Nord, Ambanja and Manakara. Within each zone, prospects

inside three altitudinal zones – low (0-100m), medium (100-150m), and high (>150m) – were done when available (Figure 4.1). The same set of accessions as those analysed in Chapter 3, excluding those from the Rama Export Collection, was used in the present study. A total of 208 accessions collected from farmers were therefore included in the present study. These accessions belong to five previously recognized genetic groups: *Vanilla planifolia*, *V. pompona*, Big Vanilla and two intermediate groups, presumably of hybrid origin (Chapter 3). SNP data were generated according to the protocol detailed in our previous study (Chapter 2, 2.2.2). For further analyses as described hereunder, SNP calling was performed independently for either the entire set of 208 accessions or *V. planifolia* accessions alone. As RDA is very sensitive to missing data, row SNPs were filtered for 20, 10 and 5% of allowed missing data, and the rate of 5% was retained. SNPs within linkage disequilibrium were removed ($r^2 > 9$). The resulting SNP data were converted into allele counts using plink (Purcell et al., 2007). To further analyse the SNP variation among the studied areas, we computed a genetic pairwise comparison (F_{st}) among accessions from each zone in *V. planifolia*, the most representative group of the studied area.

4.2 Materials and methods

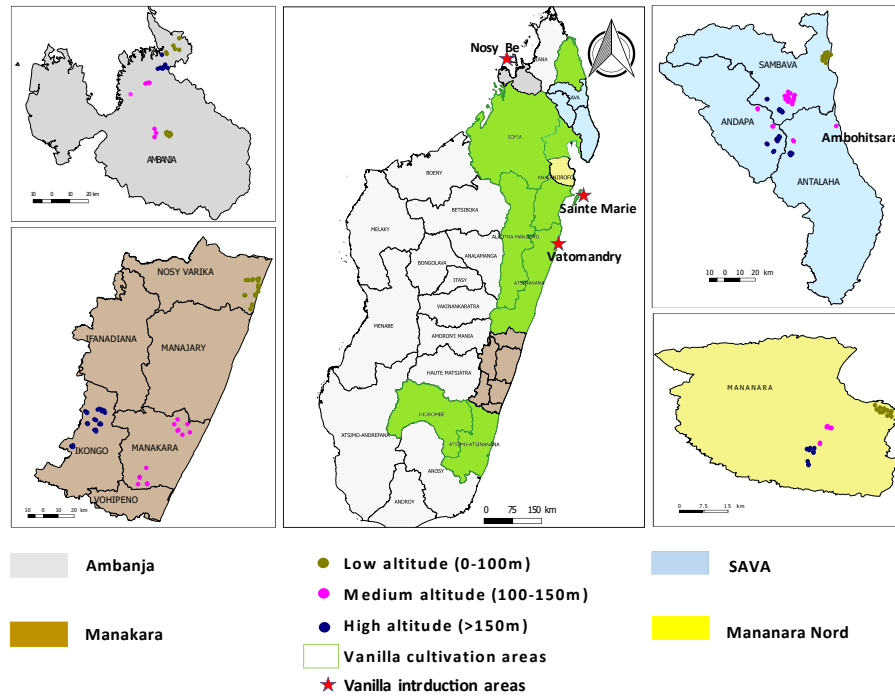


Figure 4.1. Sampling distribution of the 208 studied vanilla accessions from farmer fields. The colored delimited areas correspond to vanilla cultivation areas and the four major vanilla production zones corresponding to four administrative regions, are those colored in grey, brown, blue and yellow. Vanilla introduction areas to Madagascar are marked by red stars – Nosy Be (1870), Sainte Marie (1880) and Vatomandry (1891).

Table 4.1. Distribution and phenotypical groups of the 208 vanilla samples. Low (L), medium (M) and high altitude (H). From the 272 accessions sequenced originally, 246 were analysed in Chapter 3. From these accessions, the 38 from the Rama Export collection were excluded for the present study.

Geographical provenance		<i>V. planifolia</i>	<i>V. pompona</i>	Intermediate	Big Vanilla
Manakara	L	17	4	8	0
	M	0	0	0	0
	H	12	0	11	0
Mananara Nord	L	20	1	3	4
	M	18	2	0	0
	H	14	0	0	0
SAVA	L	14	4	0	0
	M	10	9	2	0
	H	12	2	0	0
Ambanja	L	15	2	2	0
	M	11	0	0	0
	H	11	0	0	0
Total = 208		-	154	24	26

4.2.2 Spatial data

The geographic coordinates (latitude and longitude) of each accession were recorded by GPS. These data were used to generate spatial variables in RDA using distance-based Moran's eigenvector maps – dbMEM – (Borcard et al., 2011; Dray et al., 2022). dbMEM is a PCA-based approach designed for spatial data where the geographical structure of the dataset is decomposed into a number of eigenvectors. These later are used as proxy variables in uni- or multivariate analyses that allow to compute the effect of spatial distribution in the response data matrix (Borcard et al., 2011). The dbMEM axes were calculated from the recorded latitude and longitude of each accession. The Euclidian distance matrix was first built using the *dist* function with the Soda package in R (Chambers, 2020). Uncorrelated vectors were then obtained from this matrix using the *dbmem* function of the *adespatial* package in R (Dray et al., 2022). Forward selection was applied to retain only relevant and significant vectors that allowed to explain spatial structure of the genotypes with 10,000 permutations (Crabot et al., 2019). This analysis was performed with the *ordistep* function within *vegan* in R (Oksanen, 2022).

4.2.3 Environmental data

Climate, soil, vegetation index, topography, and the geographic distance of each accession from the ancient breeding program location in Ambohitsara (Antalaha, SAVA), were here assigned as environmental data. A total of twenty-seven environmental variables were compiled for each accession (Annex, Table A.2). The 19 standard climatic variables were obtained from the WorldClim 2.0 database (Hijmans et al., 2005) with the highest resolution, 30 arcseconds [~ 1 km]. The 5 soil variables were extracted for 5 cm depth from the SoilGrids 2.0 database (Poggio et al., 2021) through the Web Coverage Services (WCS) of QGIS 3.22 (QGIS, 2022). The normalized difference vegetation index (NDVI) was also included using the Copernicus Global Land Service database (CGLS, Buchhorn et al., 2020), and data extraction was conducted with QGIS. Finally, site elevation recorded by GPS and the distance of each accession from an ancient breeding program of cultivated vanilla were also included in the environmental variables. The data were centered and scaled prior to analysis. Similarly to the treatment of spatial variables, forward selection was performed to retain only significant environmental variables for RDA analysis.

4.2.4 Disentangling the contribution of environment and geography to genetic variation

To compute the contribution of both environment and spatial data to SNP variation, variance partitioning was performed using RDA. This latter is a form of canonical analysis that allows to quantify the correlation between a multivariate response matrix with up to thousands of variables – such as SNP data – and one or more multivariate matrices of explanatory variables (Borcard et al., 2011). Here, the response variable matrix was represented by the SNP data, displayed as allele counts. The matrices of explanatory variables contained the forward selected environmental and spatial data. Canonical axes were computed by the ordination of explanatory variables that best explain response variables using PCA by multivariate linear regression. The pipeline enables partitioning of the total genetic variation explained by the model into the parts explained independently by environ-

-ment and space as well as their collinear contributions through the partial RDA option (Borcard et al., 2011). In addition to the variance partitioning among the 208 accessions altogether, variance partitioning was also computed for the main cultivated vanilla species, *i.e.* *V. planifolia* group, as its sampling is representative of the studied zone and hence allows to assess the environmental and distribution factors effect on SNP variation. The function *rda* was used to run RDA, and significance of each model was tested at 10,000 permutations using the *anova.cca* function of the R package *vegan* (Oksanen, 2022).

4.2.5 Contribution of environmental variables to SNP variation

The relative contribution of the significant environmental variables retained after forward selection was estimated by computing their respective R^2 adjusted. This was conducted at both global and within group levels. R^2 adjusted was chosen to avoid inflation of R^2 due to the number of predictor variables.

4.3 Results

4.3.1 Genetic data

Similarly to the results obtained on 246 accessions in Chapter 3, the 17,516 filtered SNP markers clustered the 208 analysed accessions into their respective genetic groups using PCA (Figure 4.2), as in the previous study using the same approach (Chapter 3.3.2, Figure 3.5) – *V. planifolia*, *V. pompona*, Big Vanilla and the two intermediate groups (Intermediate 1 and 2). The two first axes captured 47.57% of SNP variation among the 208 accessions, slightly lower than that of the 246 accessions including accessions from the Rama Export collection (54.24%), presumably due to the excluded samples from the collection and the further filtered SNPs.

For the *V. planifolia* group, a total of 9,281 SNPs was genotyped. According to UPGMA, a certain level of genetic structuration was detected within this *V. planifolia* (Figure 4.3). Although the clustering pattern did not show a clear cut-off between zones, some clusters contained a majority of acce-

-ssions from a given zone, such as Ambanja, Manakara and SAVA. Furthermore, each zone presented a number of specific alleles, with a total of 75 (Annex, Table A.3): Mananara Nord (43), Manakara (26), Ambanja (5) and SAVA (1). However, we did not find any evidence of genetic differentiation between these zones (Table 4.2). Genetic distances (Annex, Table A.4) between the 5 genetic groups from the 208 accession was still of the same magnitude as previously reported in Chapter 3, Table 3.4.

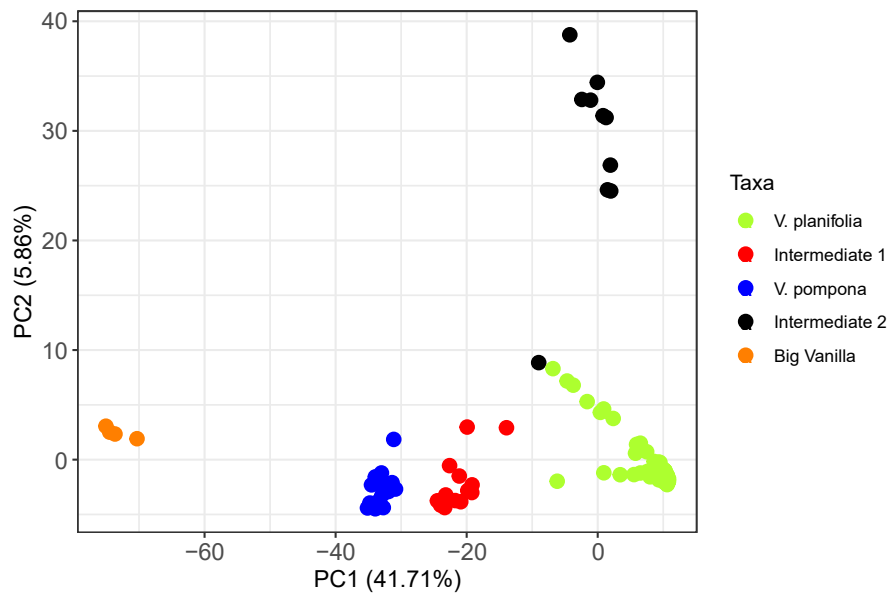


Figure 4.2. Principal Component Analysis (PCA) of the 208 accessions taken from farmer fields according to the two first eigenvectors with the 17,516 markers. Color attribution was based on phenotype identification.

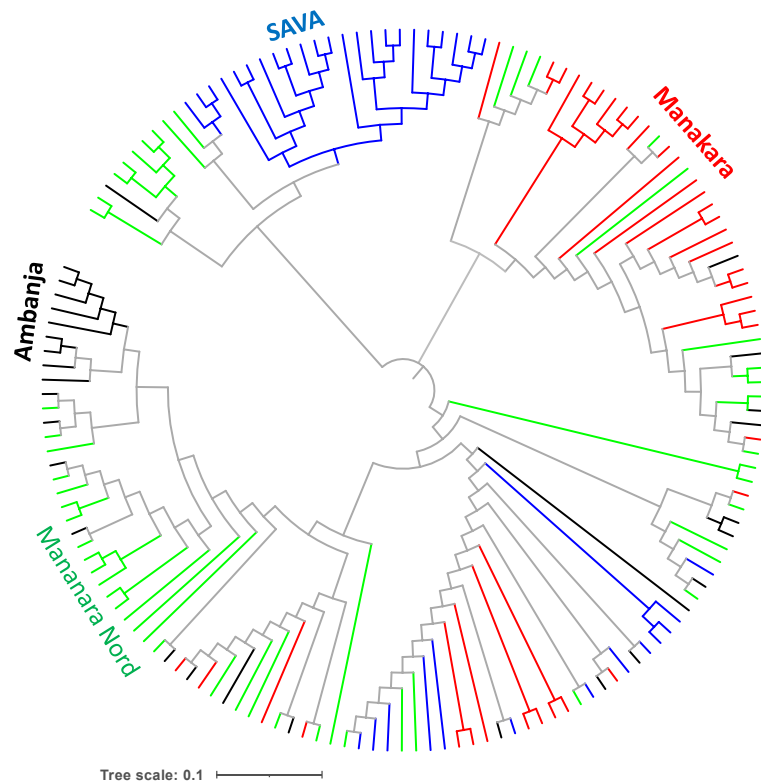


Figure 4.3. Genetic structure of the 154 accessions of *V. planifolia* using the UPGMA with the 9,281 SNPs. The color of accessions was attributed according to their provenance – black (Ambanja), green (Mananara Nord), red (Manakara) and blue (SAVA).

Table 4.2 Pairwise comparison (F_{st}) between the four zones using the 9,281 genotyped SNPs from *V. planifolia* accessions.

	Ambanja	Mananara Nord	Manakara	SAVA
Ambanja	1			
Mananara Nord	-0.015	1		
Manakara	-0.013	-0.009	1	
SAVA	-0.010	-0.011	-0.008	1

4.3.2 Spatial and environmental data according to sampling sites

Only significant and non-redundant variables that RDA found in the SNP variation contribution were kept via forward selection and used for downstream analysis. Four axes from the dbMEM variables (1, 4, 5, and 6) were found to be sufficient in explaining the spatial organization of the 208 accessions from farms. Environmental data were clearly clustered by site (Figure 4.4). Seven environmental variables were identified to significantly predict genetic diversity among all accessions: mean temperature of warmest quarter, mean temperature of wettest quarter, precipitation of driest month, temperature annual range, soil pH, precipitation of wettest quarter and sand content.

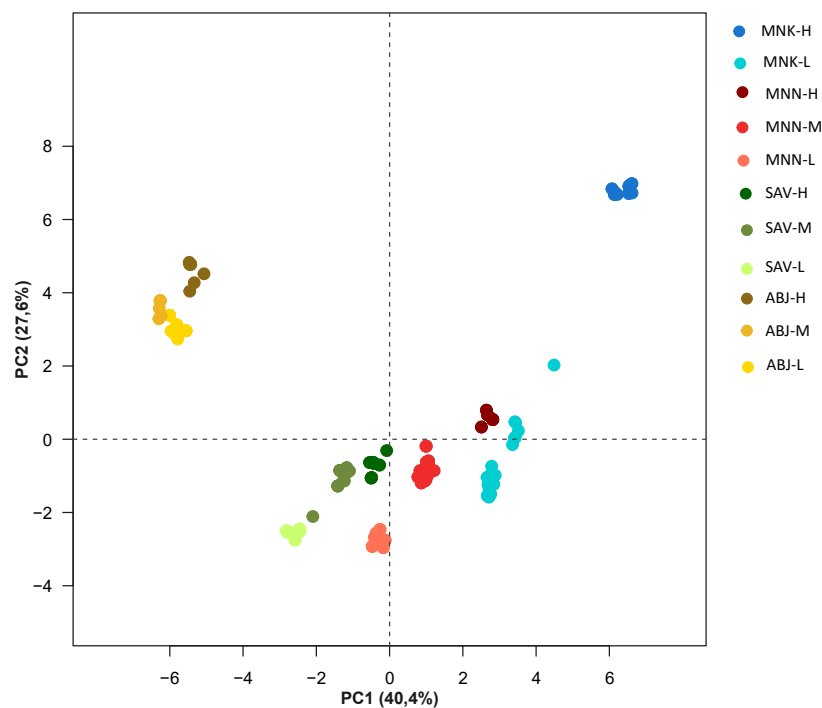


Figure 4.4. Principal Component Analysis (PCA) of 208 accessions according to the two first eigenvectors with the 27 environmental variables. Color attribution was based on sites (altitudinal level within each zone). Manakara (MNK), Mananara Nord (MNN), SAVA (SAV), Ambanja (ABJ), low altitude (L), medium altitude (M), high altitude (H).

4.3.3 Contribution of individual explanatory variables in structuring genetic divergence

The mean temperatures of the warmest and wettest quarters had the highest R^2 adjusted contribution values of 6.2 and 5.7%, respectively. Precipitation of the driest month, temperature annual range, pH, precipitation of the wettest quarter, and sand content were also involved but with less contribution, ranging from 1.4 to 3.6% (Table 4.3, Figure 4.5a).

Figure 4.5a suggests that intermediate groups 1 and 2 are genetically quite close to each other, likely indicating similar ecological conditions between them. This distinguishes them from Big Vanilla, which is found in the northern part of Madagascar and experiences different ecological conditions. While *V. pompona* is present in multiple regions, the genetic distance with the two intermediate groups is not as evident as it is with other Big Vanilla (Figure 4.5a).

Regarding the *V. planifolia* group, eight climatic variables entirely related to precipitation were identified to contribute to the SNP variation (Table 4.3). The strongest predictors were precipitation seasonality and precipitation of the wettest quarter (Figure 4.5b). Genetic structuration of *V. planifolia* was detected within Ambanja and Mananara Nord regions.

Table 4.3. Significance of individual environmental variables in SNP variation among the 208 accessions (global comparison, combining all genetic groups, using the 17,516 SNPs) and within the 154 accessions of *V. planifolia* using RDA (with the 9,281 SNPs).

variable	Code	R ² adjusted	F	p-value
Global comparison				
Mean temperature of warmest quarter	MTWaQ	0.062	2.05	0.020
Mean temperature of wettest quarter	MTWeQ	0.057	5.33	0.002
Precipitation of driest month	PDrM	0.036	2.03	0.040
Temperature annual range	TAR	0.031	1.93	0.04
Soil pH	pH	0.027	2.08	0.038
Precipitation of wettest quarter	PWeQ	0.022	2.63	0.010
Sand content	Sand	0.014	3.95	0.004
<i>V. planifolia</i>				
Precipitation seasonality	PS	0.043	1.37	0.002
Precipitation of warmest quarter	PWaQ	0.041	1.25	0.020
Precipitation of wettest month	PWeM	0.039	1.57	0.002
Precipitation of driest month	PDrQ	0.032	1.92	0.002
Annual precipitation	AP	0.035	1.488	0.002
Precipitation of driest quarter	PDrM	0.026	1.58	0.002
Max temperature of warmest month	MaxTWM	0.022	1.95	0.002
Precipitation of wettest quarter	PWeQ	0.016	3.57	0.002

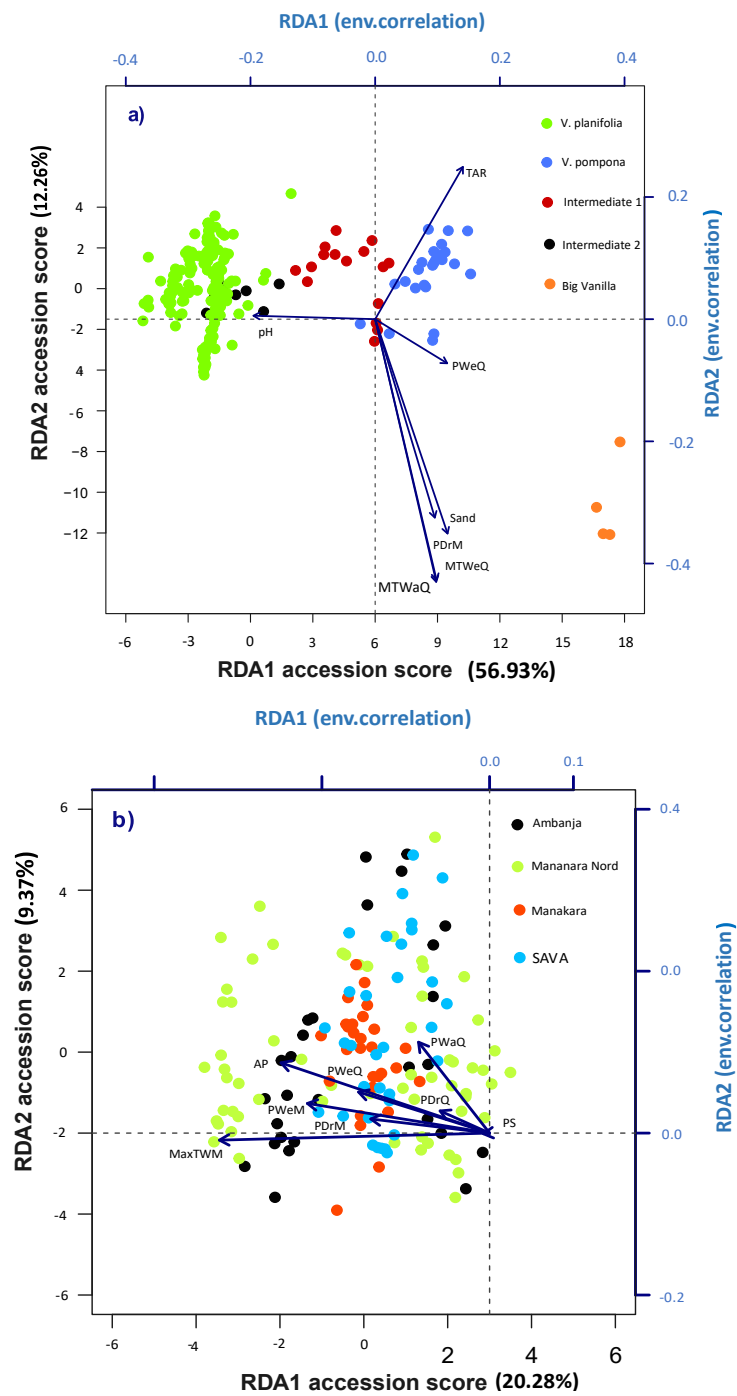


Figure 4.5. *Biplot of the two first RDA axes (dotted lines) with the 208 pooled accessions (a) and 154 V. planifolia accessions (b) loading scores (black axes, left and bottom) and the selected environmental variables (blue axes, right and top), using respectively the 17,516 and 9,281 called SNPs. RDA axes represent the ordination of the forward selected environmental variables (dark blue arrows) constrained by the selected spatial vectors. Length and direction of arrows indicate the strength in driving SNP variation and the relationship between each environmental variable, respectively. The color of accessions (points) was attributed from the genetic group for a) and from the provenance i.e. the four zones for b). TAR (temperature annual range), pH (soil pH), sand (soil sand content), PWeM (precipitation of wettest month), PDrQ (precipitation of driest quarter), PS (precipitation seasonality), PWaQ (precipitation of warmest quarter), PWeQ (precipitation of wettest quarter), PDrM (precipitation of driest months), MTWeQ (mean temperature of wettest quarter), MTWaQ (mean temperature of warmer quarter), AP (annual precipitation), MaxTWem (Max Temperature of Warmest Month). Additional information on the loading values of each vector on the two first RDA axis is provided in Annex, Table A.5.*

4.3.4 Range of variation of the significant environmental variables among the genetic groups

The genetic group Intermediate 2 showed the most distant environmental preferences compared to the remaining groups (Figure 4. 6). This group is grown in a zone with the highest temperature annual range, precipitation of the wettest quarter, and sand content, and the lowest mean temperature of the wettest and driest quarters, precipitation in the driest month, and soil pH. The four other groups had more or less equal preferences between themselves. *V. pompona* was most likely associated with sites with lower precipitation of the wettest quarter, and Big Vanilla with a zone with significantly lower mean temperature of the wettest and driest quarters, precipitation in the driest month, and soil pH. *V. planifolia* did not appear to contain any extreme values in any parameter.

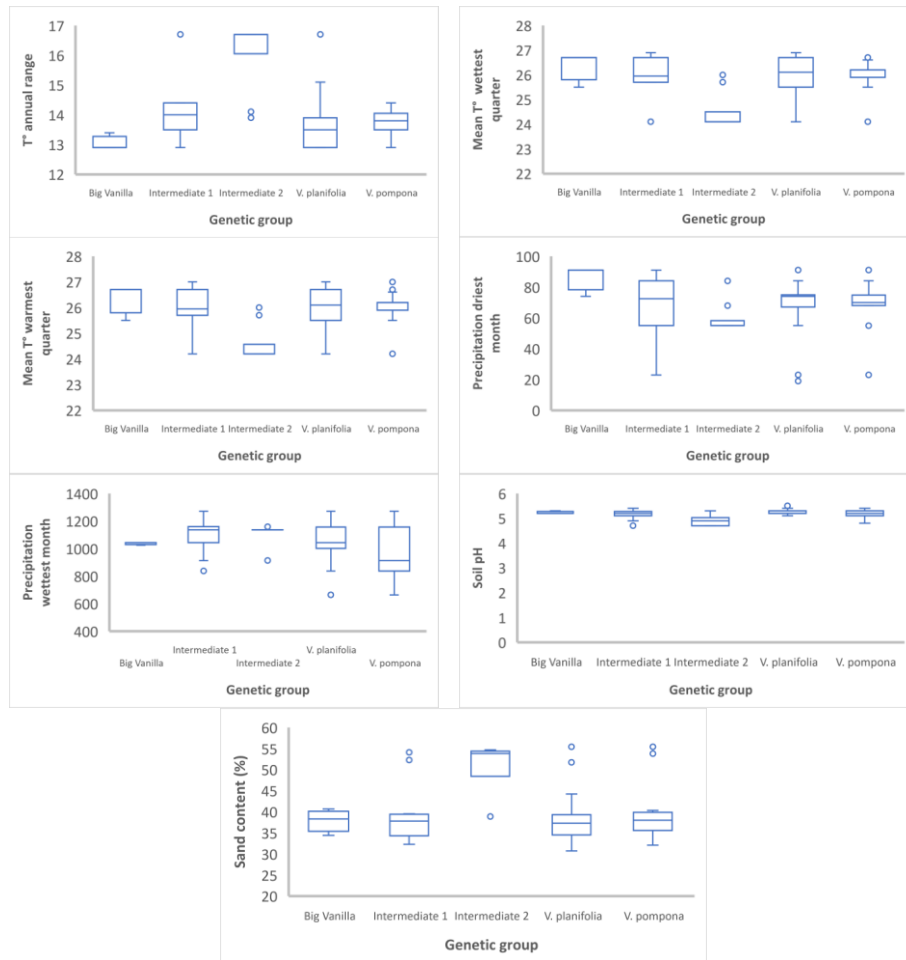


Figure 4.6. Variation of the seven environmental variables by regional genetic cluster among the 208 pooled accessions.

4.3.5 Variance partitioning by RDA

Variance partitioning by RDA was performed to quantify how much environment and spatial distribution and their collinearity influence genetic variation among the 208 accessions and within the *V. planifolia* group. Overall, the correlations between genetic variation and predictors were low but significant at both levels of comparison. The SNP variation explained by the pure environmental variables was slightly higher in the 208 accessions taken together (global comparison, 3.1%) compared to that within the *V. planifolia* group (2.24%). Spatial variables alone did not show any effect in the global comparison. Among the *V. planifolia* accessions,

the contributions of both spatial distribution (2.21%) and the environmental variables (2.24%) in the genetic variation were similar to each other. When the contributions of the environment were analyzed with spatial distribution in mind (*i.e.* spatially constrained environmental variables), they improved to 6.2 and 4.37% in the global comparison and within *V. planifolia*, respectively. The unexplained SNP variation was high in both comparisons, at 93,7% and 94,8%, respectively (Table 4.4).

Table 4.4. Contribution of environmental variables alone [a], spatial variables alone [b], spatially constrained environmental variables [c], intersection between both spatial and environmental predictors [d] into SNP variation *i.e.* the combined effect of the two variables on the proportion of variance not accounted for by either pure environment or pure spatial distribution effects. Analyses were conducted among the 208 accessions altogether (global comparison) using the 17,516 SNPs and within the *V. planifolia* group using the 9,281 SNPs called from the 154 studied accessions.

Variable	R ²	R ² adjusted	p-value
Global comparison (all vanilla accessions)			
Environnemental variables alone [a]	0.056	0.031	0.001
Spatial variables alone [b]	0.073	0.001	0.272
Spatially constrained environmental variables [c]	0.093	0.062	0.001
Intersection between b and c [d]	NA	0.031	NA
Total [a] + [b] + [d]	NA	0.063	NA
Residual	NA	0.937	NA
<i>V. planifolia</i> accessions			
Environnemental variables alone [a]	0.0940	0.0224	0.001
Spatial variables alone [b]	0.0682	0.0212	0.001
Spatially constrained environmental variables [c]	0.0714	0.0437	0.001
Intersection between b and c [d]	NA	0.0086	0.001
Total [a] + [b] + [d]	NA	0.0522	NA
Residual	NA	0.9476	NA

4.4 Discussion

Introduced species have been shown in the past to frequently change and adapt under novel environmental settings (Oduor et al., 2016; Liu et al., 2020), making them a good model for investigations on local adaptation that take place over brief time scales (Colautti & Lau, 2015). Evidence of short-time local adaptation has been previously reported in vegetatively

propagated plant species (Simón-Porcar et al., 2021). However, asexual reproduction may contribute to the slowdown of adaptive evolution in plants due to its conservative property of genetic structure of the starting material (Myles et al., 2011). In addition, the number of generations and the plant life-cycle length may influence local adaptation, i.e., long-lived perennial species may be less able to adapt to a changing environment than annuals and short-lived perennials (Anderson et al., 2011). Understanding the relationship between genetic variation and local adaptation in crop plants will help in detecting well adapted genotypes for a given environment. Cultivated vanilla is a perennial, vegetatively propagated plant, recently introduced to the Indian Ocean, where some phenotypic diversity within plantations is noticed. In Madagascar, natural vanilla production is widely practiced over a large spatial range, under a diversity of environmental conditions such as climate, soil, and altitude. Here, we aim to investigate the possible occurrence of local adaptation for the recently imported crop plant in Madagascar. To disentangle the relative effects of environment and spatial distribution on genetic variation, we performed RDA using a panel of accessions collected from vanilla fields within the major vanilla production areas under different environmental conditions. A relatively small but significant local adaptation pattern was identified in both pooled 208 accessions containing the five genetic groups and within the *V. planifolia* genetic group. The pattern of low contribution of independent environment and spatial distribution was previously reported in wild species with sexual reproduction such as wild tomato (respectively 3% and 2%, Gibson & Moyle, 2020) and Malagasy wild vanilla (3.4% and 0%, Andriamihaja et al., 2021), using RDA with SNP markers. However, these species have shown high correlations with environmental variables when constrained by space, with respectively 17 and 42.8%. The current study found that the majority of SNP variation was not explained by the investigated factors, implying that other drivers may be importantly involved. This suggests that other parameters not included in the current analysis could contribute to the genetic diversification, for instance, predictors varying at smaller scale (such as micro-climate), the farmer's influence in cutting dispersal, as well as stochastic or other sources of variation related neither to climatic envi-

-ronment nor agricultural practices. Vanilla seed dispersal, like that of other cultivated plants, is done through human mediation; particularly in Madagascar, cuttings movement is done farmer-by-farmer without any particular organization. The extent of this dispersion is unknown. As a result, sampling drift may occur. However, as shown in Table 4.2, no genetic differentiation was detected between zones, providing low support for sampling drift. The present result suggests that the cultivated *V. planifolia* in Madagascar constitutes a population of pooled accessions in which there is a very weak spatial structuration, and the existing SNP variation could be attributed to the small contribution of environmental variables.

Seven environmental variables were identified as significant, non-redundant factors that could be involved in the structuration of vanilla genetic clusters. Overall, these variables are related to temperature, precipitation and soil. By disentangling the individual effects of each environmental variable, those related to temperature (mean temperature of the wettest quarter and mean temperature of the warmest quarter) explained more SNP variation among the entire samples than soil and precipitation (Table 4.3). The existence of a correlation between spatial and altitude variables with gradients of temperature supports this finding (Annex, Table A.6). Temperature has been shown to influence genetic diversity in both model and non-model plant species (Poncet et al., 2010; Bothwell et al., 2013; Fang et al., 2013; Hsieh et al., 2013; Huang et al., 2015). It has previously been pointed out to play an important role in orchid distribution (Kolanowska, 2013; Kolanowska, 2013; Tsiftsis & Djordjević, 2020; Charitonidou et al., 2021). Regarding the seven predictors identified as involved in local adaptation of vanilla, the intermediate group 2 has shown the most extreme preference, *i.e.*, either the highest (TAR, PWeQ) or the lowest (MTWeQ, MTWaQ, PDrM, and the soil pH) values. This group is found in the high-altitude region of Manakara, which presents unique environmental conditions compared to the other sampling sites. Due to its proximity to the highlands, its climatic conditions tend to be similar those of the highlands. The commercial *V. planifolia* appeared to easily develop in heterogeneous conditions (Figure 4.6) and

V. pompona is related to sites with lower PWeQ, suggesting that species or genotypes respond differently to environmental conditions.

The computed genetic variation between the 154 accessions of *V. planifolia* was mostly predicted by variables related to precipitation. This latter has been previously identified as a key driver of SNP variation in other plant species (Jiang et al., 2019; Gao et al., 2021). The orchid's nature, i.e., the dependence of *V. planifolia*'s to humidity, could be linked to this pattern. The relatedness of orchids to precipitation variables has been demonstrated in other orchid species (Ongaro et al., 2018), as well as in *V. planifolia* (Armenta-Montero et al., 2022).

We can conclude that despite the recent introduction of vanilla in Madagascar, some local adaptation events were detected. Our previous report (Chapter 2, 2.4) stated that the GC content estimated for vanilla in Madagascar (34%) was higher than that previously reported for *V. planifolia* (30.8%, Hassing et al., 2020). GC content has been demonstrated to be involved in environmental adaptation (Šmarda et al., 2014).

Chapter 5

General discussion, conclusions and Perspectives

5.1 General discussion and conclusions

Productivity is a matter on the global supply, particularly for crops that can only be grown in specific regions. The latter is the case for vanilla, which is grown only in tropical and humid regions. The current vanilla supply is still far to fulfil the global demand. Commercial synthetic vanilla is therefore available and easily accessible on the market. Madagascar and Indonesia still control the majority of the global natural vanilla market. Product quality is also becoming increasingly important in efforts to improve crops through plant breeding (Hefferon, 2015). Consumers worldwide are starting to worry about food quality. Yet, breeding for better plant quality is a complex undertaking (Munck, 2009; Wen et al., 2016). So, the investigation on tools more efficient and effective that will help in selection for better quality products is crucial nowadays.

Madagascar began vanilla cultivation from a few cuttings introduced since 1870, similarly to other areas of introduction in the world. Several additional genotypes have been produced or introduced to the country through a breeding program from 1944 to 2000 with the aim of improving the genetic resistance of *V. planifolia* in view of fighting against the fusariosis that caused a loss of about one fifth of the entire Malagasy production since the 1950s (Tonier, 1960). The under exploration of these materials and the lack of understanding regarding their molecular characterisation motivated the work developed under this doctoral thesis. The knowledge generated by the present research project is an attempt to support the conservation and the promotion of the use of Madagascar's genetic resources for future breeding effort.

In this thesis, molecular markers were used to evaluate and characterize the genetic diversity of Malagasy vanilla population under cultivation (farmers' populations), as well as that under conservation within a local

collection build from the ancient breeding program. This latter was also used as a pilot study to set up the whole pipeline from the lab work to the data processing prior the large-scale investigation.

The main achievements of this PhD project were:

- (1) Providing further evidence of effective molecular pipeline in structuring population with low genetic diversity, such as the case of the cultivated vanilla. ddRAD sequencing was picked for the present study to generate molecular data as the protocol has been previously successfully used in other species with similar genetic contexts. Our observations also bring awareness for the need to test data processing tools especially for the SNP markers for any species in order to select the most appropriate one depending on the genomic context of the species. BCFtools was suggested to be good enough to genotype SNP markers for vanilla from ddRAD data. The approaches of SNP development were further explored in Chapter 2.
- (2) Generating useful data through the integration of genotypic and ecological characterization gathered throughout Chapter 2, 3 and 4. The output from the genetic characterization of *ex* and *in situ* germplasm generates valuable data to support an efficient and effective management of the available genetic resources in future breeding activities.

5.1.1 SNP generated from ddRAD-sequencing technology – useful markers for the genetic structuration of cultivated vanilla

Prior to the large-scale population genetic exploration of the cultivated vanilla in Madagascar, a pilot study was conducted to set up the whole protocol, from the field through the lab work to the data processing. The aim of the latter was to find a suitable pipeline for the vanilla genome. The three widely used tools for genetic diversity analysis were hence compared – Stacks, BCFtools, and GATK. Using the 56 *ex situ* vanilla samples, we found that BCFtools, GATK and the SNP set detected commonly by the three pipelines altogether provided comparable results. However, GATK required much more run time. Hence, BCFtools was picked for the analysis

and could be suggested for large-scale studies due to its reliability and reasonable requirements. On the contrary, the widely used Stacks tool was not suitable for our dataset. It is worth noting that Stacks is a powerful and versatile tool that has been successfully used in many population genetics studies across a wide range of organisms (Manching et al., 2017; Esposito et al., 2020; Feng et al., 2020). However, like any software, it has its limitations, and researchers need to carefully consider whether it is the best tool for their particular study system and research questions. All the three tools are popular in population genetics analyses; however, they are designed for different types of data and analyses. GATK and BCFtools tend to perform well when analysing large datasets. This is because they can handle complex datasets that may contain large numbers of variants. In addition, vanilla accessions are highly heterozygous, and the low capacity of Stacks in genotyping heterozygous loci was not convenient in disentangling vanilla genetic diversity. The result suggests that one tool working for a given species will not necessarily work for another species. Taking into account the genomic context of the target species is therefore useful, and comparing different pipelines is intuitively safer.

A set of high-quality SNPs has been developed from the present study. The reliability of these markers has been validated through their ability to distinguish the phenotype groups. Accessions belonging to each phenotypic group were segregated into corresponding genetic groups and intra species structuration was revealed as discussed hereunder. In addition to the deep coverage (mean = 43X), the markers were proportionally distributed along the 14 chromosomes of *V. planifolia*. The pronounced heterozygosity previously reported for vanilla was also confirmed by our results, with a negative F_{is} value in all genetic groups. This is the first application of the ddRAD protocol to the population diversity of vanilla. These results highlight the effectiveness of the ddRAD-seq approach in the genetic investigation of vanilla and can be useful in a large-scale vanilla study. The power of SNP markers compared to other popular markers such as microsatellites has been previously reported in other species with asexual reproduction such as sweet potato (Feng et al., 2020) and *Citrus sp.* (Trifa et al., 2013). The strength of the SNP markers

rely on their high density, high variability, accuracy and ability to detect rare alleles (Syvänen, 2001; Helyar et al., 2011; Flanagan & Jones, 2019). SNP markers are densely distributed across the genome, with an average density of one SNP every 1-2 kilobases. In populations with low genetic diversity, where genetic variation is expected to be concentrated in particular genomic areas, this high density enables the detection of genetic variation. SNPs are highly variable and can occur at a very high frequency in the genome. Due to the great diversity, even differences across closely related populations can be detected. SNP genotyping is a highly accurate and reproducible technique, which can be performed on high-throughput genotyping platforms. This makes it possible to analyse large numbers of individuals and loci, which can help to increase statistical power and detect subtle differences in the genetic composition of the population. SNP markers can also detect rare alleles that are not present in other individuals or populations. This can be particularly important in populations with low genetic diversity, where rare alleles may be more prevalent and provide important information about the genetic history and diversity of the population.

5.1.2 Genetic resources in Madagascar have shown high genetic structuration with the occurrence of some original genotypes

The current study evaluated on a large scale the genetic diversity of cultivated vanilla in the field in Madagascar for the first time. A panel of 246 accessions collected from farms including four local major vanilla production areas and few accessions from a local collection was studied. This germplasm included two known species (*V. planifolia* and *V. pompona*) and four other morphotypes that had not been genetically characterized previously. The latter included the Big vanilla phenotype, a presumed *V. palmarum* species; a morphotype presenting morphological traits between *V. planifolia* and *V. pompona* and was therefore called "intermediate" in the present study; the locally so called "Tsivaky" and "Vanille Banane" types. These species/phenotypes are present in both farms and the collection except Tsivaky and Vanille Banane which are only present in the collection. The 17,948 high quality genotyped SNP markers

clearly structured the taxonomic/phenotypic groups into separated genotypic groups using different approaches of genetic diversity study such as PCA, UPGMA, ADMIXTURE and Fst. The intermediate groups were identified as hybrids, positioned between *V. planifolia* and *V. pompona* and were neatly distinguished from the other morphotypes. The almost exclusive vegetative propagation that results from horticultural practices (cuttings) is a key element in the structuration and evolution of vanilla diversity. Indeed, it fixes highly heterozygous genetic structures, thus limiting gene flows. With the use of SNP data, the majority of vanilla accessions in Madagascar are grouped in accordance to their external references (taxonomic/phenotypic grouping).

Apart from the initial polymorphism resulting from multiple introductions, the main mechanisms of diversification within Malagasy germplasm were supposedly crossings of genotypes through the ancient breeding program and spontaneous mutations. Furthermore, in addition to the *V. planifolia* initially introduced in Madagascar by colonizers and the few additional local wild genotypes, most of genotypes present in Madagascar have been introduced through the breeding program from 1944 to 2000. It has been reported that few thousands of hybrids had been developed during this ancient breeding program. Unfortunately, most of them have apparently disappeared with some exhaust to the field. The variety called “Tsitaitra” is among the few promising varieties created during this program. It resulted from a crossing between *V. planifolia* and *V. pompona* (Grisoni et al, 2021). Several crossing ways have been reported between these two genitors and all the products are currently denoted as “Tsitaitra”. Our result highlights that the remaining Tsitaitra present in Madagascar should result from a back cross between *V. planifolia* and *V. pompona* with *V. pompona*, but also from another a back cross between *V. planifolia* and Big vanilla with Big Vanilla which is presumably a *V. palmarum* species. These phenotypes are what we call “intermediate” group that were genetically segregated into two groups. These genotypes are original and so constitute an added value for the Malagasy germplasm. They have been reported to present potentiality in phytopathology resistance and robustness. These genotypes could present more diverse agronomical

traits that could be evaluated tightly for vanilla trait improvement. This study highlights the importance of setting up, in the future, fine agronomic and physiological evaluations of the different original genotypes of Madagascar in comparison with other varieties. Priority should be given to the response to biotic and abiotic stresses and agronomic behaviour under suitable environmental conditions.

5.1.3 SNP markers reveal intra species genetic structuration within *Vanilla*

In addition to the structure of the global genetic resources present in Madagascar, the genetic structure within the main cultivated species *V. planifolia* was investigated. The relationship between the 154 sequenced *V. planifolia* accessions was explored. The study found a high level of variation within the species, with a large number of polymorphic SNP variants identified (9,281). Our results suggest that SNP markers are more informative than classical markers previously used in *V. planifolia*. For instance, the ITS region of nuclear ribosomal DNA and RFLP did not successfully separate the four morphotypes in Mexico (Soto Arenas & Dressler, 2013, Bory et al. (2008b). A similar result has been recently reported using the same marker type on *V. planifolia* collected from a geographically broader scale (mostly from America and Asia) (Chambers et al., 2021). Three closer groups called as “types” have been distinguished from which, clear geographical delimitation was not revealed (Chambers et al., 2021). Genetic structuration among *V. planifolia* accessions was revealed despite the vegetative propagation of vanilla and the narrow genetic basis of the introduced material. However, there was no clear distinction among geographic zones, possibly due to the movement of cuttings between areas through exchanges between farmers. Though, we observed regional domination in the clustering of *V. planifolia* accessions, suggesting that there may be some evolutionary process occurring, that are resulting in genetic diversification among the accessions. This hypothesis is supported by the detected amount of environmental parameters contribution in the SNP variation in this study (discussed in chapter 4). A study using SNP markers on vanilla accessions from large collections in Reunion and French Polynesia did not revealed any genetic

structuration among vegetatively propagated *V. planifolia* accessions (Favre et al., 2022). This result supports the occurrence of adaptation phenomenon within vanilla.

Morphological variation among *V. planifolia* accessions was observed within plantations. However, we were mostly limited to the color and size of the vegetative part of the plant. Due to constraints mainly related to time, we did not have the opportunity to assess other stable traits such as beans and flowers. For this reason, we did not include the morphological description in the present study. A comparison with the documented morphotypes on Reunion Island (Blue, Grosse Vanille, Aiguille, Classique, and Sterile; Bory et al., 2008b) could enhance the knowledge of the genetic resources present in the Indian Ocean.

Intra-specific structuration was also detected within the *V. pompona* species, where accessions were structured according to their provenance. Interestingly, in the SAVA region, altitudinal distribution was identified. Using AFLP markers, a similar pattern was also reported previously on a large geographical range of accessions of *V. pompona* (Brazil, French Polynesia, Guadeloupe, French Guyana, Reunion Island, etc.) (Bory et al., 2008b). The present findings confirm the effectiveness of SNPs markers in the infra species genetic diversity studies.

5.2. Research contribution

The effectiveness of ddRAD-seq in population genetic studies of cultivated vanilla was demonstrated, and it could be used for investigation at a larger scale. Since the introduction of vanilla in Madagascar and its long history in global vanilla production, this study provides the first data based on molecular markers. Our study has drawn a picture of the available genetic resources in Madagascar. A total of 17,948 wide-genome high-quality SNPs was developed from the present study. The overall high genetic diversity combined with the high number of specific alleles represents a valuable gene pool of vanilla. Coupled with the diversity of interesting agronomical traits of the genotypes, such as vanillin content, bean indehiscence, yield, and fungi and viral resistance, the detected high-qua-

-lity SNPs will be useful in trait association research. Recorded and developed data from the present research will be useful material in the establishment of conservation strategies and future vanilla production improvement in Madagascar.

5.3 Perspectives

Conservation of genetic resources is urgent. *Ex situ* and/or *in situ* conservation of the genetic resources is highly suggested in order to rescue the remaining genotypes. The information from the current study that is already available along with the locations of the accessions will be essential to launch and focus this effort.

Once conservation is established, several studies for the ultimate goal of sustainable vanilla production could be planned. The present study provided a molecular characterization of the cultivated vanilla in Madagascar. Agronomical description of the identified genotypes will be among the major next steps before identification of candidate genes for potential use in the vanilla improvement program.

The developed markers in the present study will be available for future analyses of other accessions and could help in assisting selection that could accelerate breeding programs.

References

- Abebe, Z., Mengesha, A., Teressa, A., & Tefera, W. (2009). Efficient in vitro multiplication protocol for *Vanilla planifolia* using nodal explants in Ethiopia. *African Journal of Biotechnology*, 8(24), 6817–6821
- Albert, P. S., Gao, Z., Danilova, T. V., & Birchler, J. A. (2010). Diversity of chromosomal karyotypes in maize and its relatives. *Cytogenetic and Genome Research*, 129(1–3), 6–16. <https://doi.org/10.1159/000314342>
- Alexander, D. H., & Lange, K. (2009). Admixture 1 . *Software Manual*
- Alberto, F., Gouveia, L., Arnaud-Haond, S., Pérez-Lloréns, J. L., Duarte, C. M., & Serrão, E. A. (2005). Within-population spatial genetic structure, neighbourhood size and clonal subrange in the seagrass *Cymodocea nodosa*. *Molecular Ecology*, 14(9), 2669–2681. <https://doi.org/10.1111/j.1365-294X.2005.02640.x>
- Allaby, R. G., Ware, R. L., & Kistler, L. (2019). A re-evaluation of the domestication bottleneck from archaeogenomic evidence. *Evolutionary Applications*, 12(1), 29–37. <https://doi.org/10.1111/eva.12680>
- Alexander, D. H., & Lange, K. (2009). Admixture 1 . 02 Software Manual.
- Allendorf, F. W. (1986). Genetic drift and the loss of alleles versus heterozygosity. *Zoo Biology*, 5(2), 181–190. <https://doi.org/10.1002/zoo.1430050212>
- Ammar, M. H., Alghamdi, S. S., Migdadi, H. M., Khan, M. A., El-Harty, E. H., & Al-Faifi, S. A. (2015). Assessment of genetic diversity among faba bean genotypes using agro-morphological and molecular markers. *Saudi Journal of Biological Sciences*, 22(3), 340–350. <https://doi.org/10.1016/j.sjbs.2015.02.005>
- Anandan, A. (2004). *Vanilla: The Green Gold*. Sura Books
- Anderson, J. T., Willis, J. H., & Mitchell-Olds, T. (2011). Evolutionary genetics of plant adaptation. *Trends in Genetics*, 27(7), 258–266. <https://doi.org/10.1016/j.tig.2011.04.001>
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolu-

References

- tionary genomics. *Nature Reviews Genetics*, 17(2), 81–92. <https://doi.org/10.1038/nrg.2015.28>
- Andrews, S (2010). FastQC: a quality control tool for high throughput sequencing data. Available online: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (accessed on July 2022)
- Andriamihaja, C. F., Ramarosandratana, A. V., Grisoni, M., Jeannoda, V. H., & Besse, P. (2021). Drivers of population divergence and species differentiation in a recent group of indigenous orchids (*Vanilla* spp.) in Madagascar. *Ecology and Evolution*, 11(6), 2681–2700. <https://doi.org/10.1002/ece3.7224>
- Andrianoelina, O., Rakotondraoelina, H., Ramamonjisoa, L., Maley, J., Danthu, P., & Bouvet, J. M. (2006). Genetic diversity of *Dalbergia monticola* (Fabaceae) an endangered tree species in the fragmented oriental forest of Madagascar. *Biodiversity and Conservation*, 15(4), 1109–1128. <https://doi.org/10.1007/s10531-004-2178-6>
- Arditti J. *Vanilla*. In: *Micropropagation of orchids*. Vol. 2. 2nd ed. Hoboken, NJ: Blackwell Publishing; 2008. p. 1239-90
- Arditti J. *Vanilla*. In: *Micropropagation of orchids*. Vol. 2. 2nd ed. Hoboken, NJ: Blackwell Publishing; 2008. p. 1239-90
- Armenta-Montero, S., Menchaca-García, R., Pérez-Silva, A., & Velázquez-Rosas, N. (2022). Changes in the Potential Distribution of *Vanilla planifolia* Andrews under Different Climate Change Projections in Mexico. *Sustainability (Switzerland)*, 14(5). <https://doi.org/10.3390/su14052881>
- Azeez S. *Vanilla*. In: Parthasarathy VA, Chempakam B, Zachariah TJ, editors. *Chemistry of spices*. UK: CAB International; 2008. p. 287-311
- Balloux, F., Lehmann, L., & De Meeûs, T. (2003). The population genetics of clonal and partially clonal diploids. *Genetics*, 164(4), 1635–1644.
- Becerra, V., Mario Paredes, C., Carmen Rojo, M., Díaz, L. M., & Blair, M. W. (2010). Microsatellite marker characterization of Chilean common bean (*Phaseolus vulgaris* L.) germplasm. *Crop Science*, 50(5), 1932–1941. <https://doi.org/10.2135/cropsci2009.08.0442>
- Banerjee, G., & Chattopadhyay, P. (2019). Vanillin biotechnology: the perspectives and future. *Journal of the Science of Food and Agriculture*, 99(2), 499–506. <https://doi.org/10.1002/jsfa.9303>

- Baqueiro-Peña, I., & Guerrero-Beltrán, J. Á. (2017). Vanilla (*Vanilla planifolia* Andr.), its residues and other industrial by-products for recovering high value flavor molecules: A review. *Journal of Applied Research on Medicinal and Aromatic Plants*, 6, 1–9. <https://doi.org/10.1016/j.jarmap.2016.10.003>
- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of Disease in Childhood: Education and Practice Edition*, 98(6), 236–238. <https://doi.org/10.1136/archdischild-2013-304340>
- Belanger, F. C., & Havkin-Frenkel, D. (2010). Molecular Analysis of a Vanilla Hybrid Cultivated in Costa Rica. *Handbook of Vanilla Science and Technology*, 256–265. <https://doi.org/10.1002/9781444329353.ch15>
- Besse, P., Da Silva, D., Bory, S., Grisoni, M., Le Bellec, F., & Duval, M. F. (2004). RAPD genetic diversity in cultivated vanilla: *Vanilla planifolia*, and relationships with *V. tahitensis* and *V. pompona*. *Plant Science*, 167(2), 379–385. <https://doi.org/10.1016/j.plantsci.2004.04.007>
- Bhanu, A. N. (2017). Assessment of Genetic Diversity in Crop Plants - An Overview. *Advances in Plants & Agriculture Research*, 7(3), 279–286. <https://doi.org/10.15406/apar.2017.07.00255>
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Borazan, A., & Babaç, M. T. (2003). Morphometric leaf variation in oaks (*Quercus*) of Bolu, Turkey. *Annales Botanici Fennici*, 40(4), 233–242
- Borcard, D., Gillet, F., & Legendre, P. (2011). Numerical Ecology with R. In *Numerical Ecology with R*. <https://doi.org/10.1007/978-1-4419-7976-6>
- Bory, S., Da Silva, D., Risterucci, A. M., Grisoni, M., Besse, P., & Duval, M. F. (2008d). Development of microsatellite markers in cultivated vanilla: Polymorphism and transferability to other vanilla species. *Scientia Horticulturae*, 115(4), 420–425. <https://doi.org/10.1016/j.scienta.2007.10.020>
- Bory, Severine, Lubinsky, P., Risterucci, A. M., Noyer, J. L., Grisoni, M., Duval, M. F., & Besse, P. (2008b). Patterns of introduction and diversification of *Vanilla planifolia* (Orchidaceae) in Reunion Island

References

- (Indian Ocean). *American Journal of Botany*, 95(7), 805–815.
<https://doi.org/10.3732/ajb.2007332>
- Bory, Séverine, Catrice, O., Brown, S., Leitch, I. J., Gigant, R., Chiroleu, F., Grisoni, M., Duval, M. F., & Besse, P. (2008c). Natural polyploidy in *Vanilla planifolia* (Orchidaceae). *Genome*, 51(10), 816–826.
<https://doi.org/10.1139/G08-068>
- Bory, Séverine, Grisoni, M., Duval, M. F., & Besse, P. (2008a). Biodiversity and preservation of vanilla: Present state of knowledge. *Genetic Resources and Crop Evolution*, 55(4), 551–571.
<https://doi.org/10.1007/s10722-007-9260-3>
- Bothwell, H., Bisbing, S., Therkildsen, N. O., Crawford, L., Alvarez, N., Holderegger, R., & Manel, S. (2013). Identifying genetic signatures of selection in a non-model species, alpine gentian (*Gentiana nivalis* L.), using a landscape genetic approach. *Conservation Genetics*, 14(2), 467–481. <https://doi.org/10.1007/s10592-012-0411-5>
- Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Botstein. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms. *Am J Hum Gen*, 32, 314–331.
papers2://publication/uuid/0B80518E-A22B-41F3-BE43-171F51007E42
- Bouetard, A., Lefeuvre, P., Gigant, R., Bory, S., Pignal, M., Besse, P., & Grisoni, M. (2010). Evidence of transoceanic dispersion of the genus *Vanilla* based on plastid DNA phylogenetic analysis. *Molecular Phylogenetics and Evolution*, 55(2), 621–630.
<https://doi.org/10.1016/j.ympev.2010.01.021>
- Bouriquet, G. (1946). Le Vanillier et la Vanille à Madagascar. *Revue Internationale de Botanique Appliquée et d'agriculture Tropicale*, 26(286), 398–404. <https://doi.org/10.3406/jatba.1946.1981>
- Boyce, M. C., Haddad, P. R., & Sostaric, T. (2003). Determination of flavour components in natural vanilla extracts and synthetic flavourings by mixed micellar electrokinetic capillary chromatography. *Analytica Chimica Acta*, 485(2), 179–186. [https://doi.org/10.1016/S0003-2670\(03\)00413-6](https://doi.org/10.1016/S0003-2670(03)00413-6)
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19), 2633–2635.

- <https://doi.org/10.1093/bioinformatics/btm308>
- Bramel, & Frey. (2021). *Global strategy for the conservation and use of vanilla genetic resources. The International Treaty on plant genetic resources for food and agriculture*
- Brown, K. A., Parks, K. E., Bethell, C. A., Johnson, S. E., & Mulligan, M. (2015). Predicting plant diversity patterns in madagascar: Understanding the effects of climate and land cover change in a biodiversity hotspot. *PLoS ONE*, 10(4), 1–12. <https://doi.org/10.1371/journal.pone.0122721>
- Brown, S. C., Bourge, M., Maunoury, N., Wong, M., Bianchi, M. W., Lepers-Andrzejewski, S., Besse, P., Siljak-Yakovlev, S., Dron, M., & Satiat-Jeunemaître, B. (2017). DNA Remodeling by strict partial endoreplication in orchids, an original process in the plant Kingdom. *Genome Biology and Evolution*, 9(4), 1051–1071. <https://doi.org/10.1093/gbe/evx063>
- Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N., Masiliunas, D., Linlin, L., Herold, M., & Fritz, S. (2020). Copernicus Global Land Service: Land Cover 100m: Collection 3: epoch 2015: Globe (Version V3.0.1). *Remote Sensing*, 12(6), 1044
- Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1842(10), 1932–1941. <https://doi.org/10.1016/j.bbadis.2014.06.015>
- Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1842(10), 1932–1941. <https://doi.org/10.1016/j.bbadis.2014.06.015>
- Bulpitt, C. J. (2005). The uses and misuses of orchids in medicine. *QJM - Monthly Journal of the Association of Physicians*, 98(9), 625–631. <https://doi.org/10.1093/qjmed/hci094>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, 10, 1–9. <https://doi.org/10.1186/1471-2105-10-421>
- Cameron, K. M. (2004). Utility of plastid psaB gene sequences for inves-

References

- tigating intrafamilial relationships within Orchidaceae. *Molecular Phylogenetics and Evolution*, 31(3), 1157–1180. <https://doi.org/10.1016/j.ympev.2003.10.010>
- Cameron, K. M., & Carmen Molina, M. (2006). Photosystem II gene sequences of psbB and psbC clarify the phylogenetic position of Vanilla (Vanilloideae, Orchidaceae). *Cladistics*, 22(3), 239–248. <https://doi.org/10.1111/j.1096-0031.2006.00102.x>
- Cameron, K. M., Chase, M. W., Whitten, W. M., Kores, P. J., Jarrell, D. C., Albert, V. A., Yukawa, T., Hills, H. G., & Goldman, D. H. (1999). A phylogenetic analysis of the Orchidaceae: Evidence from rbcL nucleotide sequences. *American Journal of Botany*, 86(2), 208–224. <https://doi.org/10.2307/2656938>
- Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., & Postlethwait, J. H. (2011). Stacks: Building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, 1(3), 171–182. <https://doi.org/10.1534/g3.111.000240>
- Catchen, J. M., Bassham, S., Wilson, T., Currey, M., O'Brien, C., Yeates, Q., & Cresko, W. A. (2013b). The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, 22(11), 2864–2883. <https://doi.org/10.1111/mec.12330>
- Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). StackCatchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013a). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–40. <http://doi.org/10.1111/mec.12354s>: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140. <https://doi.org/10.1111/mec.12354.Stacks>
- Chambers, A. H. (2019). Vanilla (Vanilla spp.) Breeding. *Advances in Plant Breeding Strategies: Industrial and Food Crops: Volume 6*, 707-734
- Chambers, A. J. M. (2020). *Package ‘SoDA’ NeedsCompilation yes*
- Chambers, A., Cibrián-Jaramillo, A., Karremans, A. P., Moreno Martinez, D., Hernandez-Hernandez, J., Brym, M., Resende Jr, M. F., Moloney, R., Sierra, S. N., Hasing, T., Alomia, Y. A., Hu, Y., & Genotyping Consortium, V. (2021). *Genotyping-By-Sequencing diversity analysis of international Vanilla collections uncovers hidden diversity and*

- enables plant improvement*.
<https://doi.org/10.1016/j.plantsci.2021.111019>
- Chang, C. W., Fridman, E., Mascher, M., Himmelbach, A., & Schmid, K. (2022). Physical geography, isolation by distance and environmental variables shape genomic variation of wild barley (*Hordeum vulgare* L. ssp. *spontaneum*) in the Southern Levant. *Heredity*, 128(2), 107–119. <https://doi.org/10.1038/s41437-021-00494-x>
- Charitonidou, M., Kougioumoutzis, K., & Halley, J. M. (2021). An orchid in retrograde: Climate-driven range shift patterns of *ophrys helenae* in Greece. *Plants*, 10(3), 1–18. <https://doi.org/10.3390/plants10030470>
- Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 34(17), i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
- Chen, X., Hoon Lim, S., Man Wong, S., Hwa Lee, Y., Kuo, J., Wing Yam, T., & Lin, J. J. (1999). Amplified fragment length polymorphism analysis of vandaceous orchids. *Plant Science*, 141(2), 183–189. [https://doi.org/10.1016/S0168-9452\(98\)00245-3](https://doi.org/10.1016/S0168-9452(98)00245-3)
- Colautti, R. I., & Lau, J. A. (2015). Contemporary evolution during invasion: Evidence for differentiation, natural selection, and local adaptation. *Molecular Ecology*, 24(9), 1999–2017. <https://doi.org/10.1111/mec.13162>
- Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, 2008. <https://doi.org/10.1155/2008/619832>
- Conesa, A., & Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics*, 2008. <https://doi.org/10.1155/2008/619832>
- Correll, D. S. (1953). Vanilla-its botany, history, cultivation and economic import. *Economic Botany*, 7(4), 291–358
- Crabot, J., Clappe, S., Dray, S., & Datry, T. (2019). Testing the Mantel statistic with a spatially-constrained permutation procedure. *Methods in Ecology and Evolution*, 10(4), 532–540. <https://doi.org/10.1111/2041210X.13141>

References

- Crotti, M., Barratt, C. D., Loader, S. P., Gower, D. J., & Streicher, J. W. (2019). Causes and analytical impacts of missing data in RADseq phylogenetics: Insights from an African frog (*Afrivalus*). *Zoologica Scripta*, 48(2), 157–167. <https://doi.org/10.1111/zsc.12335>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), 1–4. <https://doi.org/10.1093/gigascience/giab008>
- Das, A. B., Mohanty, I. C., Mahapatra, D., Mohanty, S., & Ray, A. (2010). Genetic variation of Indian potato (*Solanum tuberosum* L.) genotypes using chromosomal and RAPD markers. *Crop Breeding and Applied Biotechnology*, 10(3), 238–246. <https://doi.org/10.1590/s1984-70332010000300009>
- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7), 499–510. <https://doi.org/10.1038/nrg3012>
- De Martinville, B., Wyman, A. R., White, R., & Francke, U. (1982). Assignment of the first random restriction fragment length polymorphism (RFLP) locus (D14S1) to a region of human chromosome 14. *American Journal of Human Genetics*, 34(2), 216–226
- Dequaire, J. (1976). L'amélioration du vanillier à Madagascar. *Journal d'agriculture Tropicale et de Botanique Appliquée*, 23(7), 139–158. <https://doi.org/10.3406/jatba.1976.3246>
- Devos, K. M., Atkinson, M. D., Chinoy, C. N., Francis, H. A., Harcourt, R. L., Koebner, R. M. D., Liu, C. J., Masojć, P., Xie, D. X., & Gale, M. D. (1993). Chromosomal rearrangements in the rye genome relative to that of wheat. *Theoretical and Applied Genetics*, 85(6–7), 673–680. <https://doi.org/10.1007/BF00225004>
- Divakaran, M., Babu, K. N., Ravindran, P. N., & Peter, K. V. (2006). Inter-

- specific hybridization in vanilla and molecular characterization of hybrids and selfed progenies using RAPD and AFLP markers. *Scientia Horticulturae*, 108(4), 414–422. <https://doi.org/10.1016/j.scienta.2006.02.018>
- Dixon, R. A. (2010). Vanillin Biosynthesis - Not as Simple as it Seems? *Handbook of Vanilla Science and Technology*, 292–298. <https://doi.org/10.1002/9781444329353.ch18>
- Djè, Y., Heuertz, M., Lefebvre, C., & Vekemans, X. (2000). Assessment of genetic diversity within and among germplasm accessions in {...}. *TAG Theoretical and Applied {...}*, 918–925. <http://www.springerlink.com/index/B4Y4J9ND3T9QWYB4.pdf>
- Doyle, J. J., & Doyle, J. L. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* (Vol. 19, Issue 1, pp. 11–15)
- Dray, A. S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., & Wagner, H. H. (2022). *Package ‘adespatial’*. <https://doi.org/10.1890/11-1183.1>.Maintainer>
- Dubcovsky, J., Ramakrishna, W., Sanmiguel, P. J., Busso, C. S., Yan, L., Shiloff, B. A., & Bennetzen, J. L. (2001). *Plantphysiology_125_1342.Pdf*. 125(March), 1342–1353
- Duval M.F., Bory S., Andrzejewski S., Grisoni M., Besse P., Causse S., Charon C., Dron M., Odoux E., & W M. (2006). Diversité génétique des vanilliers dans leurs zones de dispersion secondaire\n. *Les Actes Du BRG*, 6, 181–196.
- Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1–15. <https://doi.org/10.1038/hdy.2010.152>
- El-denary, M. E., Ali, R. A. M., & Gomaa, S. E. (2016). Morphological Markers for Early Selection in Strawberry Breeding Programs. *Alexandria Science Exchange Journal: An International Quarterly Journal of Science Agricultural Environments*, 37(April-June), 102–115. <https://doi.org/10.21608/asejaiqsae.2016.2098>
- Elias, M., Penet, L., Vindry, P., McKey, D., Panaud, O., & Robert, T. (2001). Unmanaged sexual reproduction and the dy-namics of genetic dive-

References

- diversity of a vegetatively propagated crop plant, cassava (*Manihot esculenta* Crantz), in a traditional farming system. *Molecular Ecology*, 10(8), 1895–1907. <https://doi.org/10.1046/j.0962-1083.2001.01331>
- Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, 6(5), 1–10. <https://doi.org/10.1371/journal.pone.0019379>
- Esposito, S., Cardi, T., Campanelli, G., Sestili, S., Díez, M. J., Soler, S., Prohens, J., & Tripodi, P. (2020). ddRAD sequencing-based genotyping for population structure analysis in cultivated tomato provides new insights into the genomic diversity of Mediterranean ‘da serbo’ type long shelf-life germplasm. *Horticulture Research*, 7(1). <https://doi.org/10.1038/s41438-020-00353-6>
- Excoffier, L. and Lischer, H. E. L. (2010). An Integrated Software Package for Population Genetics Data Analysis. *Molecular Ecology Resources*, 10, 564–567
- Exley, R. (2019). *Vanilla Production in Australia Climatic Regions of Australia Suitable for Vanilla*. *Ridley 1912*, 147–156
- Fajardo, C. G., de Almeida Vieira, F., & Molina, W. F. (2014). Interspecific genetic analysis of orchids in Brazil using molecular markers. *Plant Systematics and Evolution*, 300(8), 1825–1832. <https://doi.org/10.1007/s00606-014-1009-9>
- Fang, J. Y., Chung, J. Der, Chiang, Y. C., Chang, C. Te, Chen, C. Y., & Hwang, S. Y. (2013). Divergent Selection and Local Adaptation in Disjunct Populations of an Endangered Conifer, *Keteleeria davidiana* var. *formosana* (Pinaceae). *PLoS ONE*, 8(7), 1–16. <https://doi.org/10.1371/journal.pone.0070162>
- Favre, F., Jourda, C., Grisoni, M., Piet, Q., Rivallan, R., Dijoux, J. B., Hascoat, J., Lepers-Andrzejewski, S., Besse, P., & Charron, C. (2022). A genome-wide assessment of the genetic diversity, evolution and relationships with allied species of the clonally propagated crop *Vanilla planifolia* Jacks. ex Andrews. *Genetic Resources and Crop Evolution*. <https://doi.org/10.1007/s10722-022-01362-1>
- Feng, J., Zhao, S., Li, M., Zhang, C., Qu, H., Li, Q., Li, J., Lin, Y., & Pu, Z. (2020). Genome-wide genetic diversity detection and population structure

- analysis in sweetpotato (*Ipomoea batatas*) using RAD-seq. *Genomics*, 112(2), 1978–1987. <https://doi.org/10.1016/j.ygeno.2019.11.010>
- Ferreira, A. W. C., Franken, E. P., & Pansarin, E. R. (2020). Confirmation of the presence of *vanilla hartii* rolfe (Orchidaceae, vanilloideae) in Brazil. *Check List*, 16(4), 951–956. <https://doi.org/10.15560/16.4.951>
- Flanagan, S. P., & Jones, A. G. (2019). The future of parentage analysis: From microsatellites to SNPs and beyond. *Molecular Ecology*, 28(3), 544–567. <https://doi.org/10.1111/mec.14988>
- Fouché, J. G., & Jouve, L. (1999). *Vanilla planifolia* : history , botany and culture in Reunion island To cite this version : *Vanilla planifolia* : history , botany and culture in Reunion island. *EDP Sciences*, 19(8), 689–703
- Fraser, D. J., Weir, L. K., Bernatchez, L., Hansen, M. M., & Taylor, E. B. (2011). Extent and scale of local adaptation in salmonid fishes: Review and meta-analysis. *Heredity*, 106(3), 404–420. <https://doi.org/10.1038/hdy.2010.167>
- Freeman, S., & Herron, J. (2007). *Evolutionary analysis Chapter 3 - Darwinian Natural Selection Chapter 11 - Sexual Selection Chapter 12 Kin Selection and Social Behaviour* (p. 576). <http://uncw.edu/people/bruce/hon210/pdfs/mccartneychaps.pdf>
- Frenkel, C., Ranadive, A. S., Vázquez, J. T., & Havkin-Frenkel, D. (2010). Curing of Vanilla. *Handbook of Vanilla Science and Technology*, 79–106. <https://doi.org/10.1002/9781444329353.ch6>
- Gallage, N. J., Hansen, E. H., Kannangara, R., Olsen, C. E., Motawia, M. S., Jørgensen, K., Holme, I., Hebelstrup, K., Grisoni, M., & Møller, B. L. (2014). Vanillin formation from ferulic acid in *Vanilla planifolia* is catalysed by a single enzyme. *Nature Communications*, 5(May). <https://doi.org/10.1038/ncomms5037>
- Gallage, N. J., & Møller, B. L. (2015). Vanillin-bioconversion and bioengineering of the most popular plant flavor and its de novo biosynthesis in the vanilla orchid. *Molecular Plant*, 8(1), 40–57. <https://doi.org/10.1016/j.molp.2014.11.008>
- Gao, Y., Yin, S., Wu, L., Dai, D., Wang, H., Liu, C., & Tang, L. (2017). Genetic diversity and structure of wild and cultivated *Amorphophallus paeonifol-*

References

- lius populations in southwestern China as revealed by RAD-seq. *Scientific Reports*, 7(1), 2–11. <https://doi.org/10.1038/s41598-017-14738-6>
- Gao, J., Liu, Z. L., Zhao, W., Tomlinson, K. W., Xia, S. W., Zeng, Q. Y., Wang, X. R., & Chen, J. (2021). Combined genotype and phenotype analyses reveal patterns of genomic adaptation to local environments in the subtropical oak *Quercus acutissima*. *Journal of Systematics and Evolution*, 59(3), 541–556. <https://doi.org/10.1111/jse.12568>
- Gargiulo, R., Kull, T., & Fay, M. F. (2021). Effective double-digest RAD sequencing and genotyping despite large genome size. *Molecular Ecology Resources*, 21(4), 1037–1055. <https://doi.org/10.1111/1755-0998.13314>
- Garvin, M. R., Saitoh, K., & Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: A technical review. *Molecular Ecology Resources*, 10(6), 915–934. <https://doi.org/10.1111/j.1755-0998.2010.02891.x>
- Ge, X. J., Liu, M. H., Wang, W. K., Schaal, B. A., & Chiang, T. Y. (2005). Population structure of wild bananas, *Musa balbisiana*, in China determined by SSR fingerprinting and cpDNA PCR-RFLP. *Molecular Ecology*, 14(4), 933–944. <https://doi.org/10.1111/j.1365-294X.2005.02467.x>
- Gibson, M. J. S., & Moyle, L. C. (2020). Regional differences in the abiotic environment contribute to genomic divergence within a wild tomato species. *Molecular Ecology*, 29(12), 2204–2217. <https://doi.org/10.1111/mec.15477>
- Gigant, R., Brugel, A., de Bruyn, A., Risterucci, A. M., Guiot, V., Viscardi, G., Humeau, L., Grisoni, M., & Besse, P. (2012). Nineteen polymorphic microsatellite markers from two African *Vanilla* species: Across-species transferability and diversity in a wild population of *V. humblotii* from Mayotte. *Conservation Genetics Resources*, 4(1), 121–125. <https://doi.org/10.1007/s12686-011-9489-1>
- Gish, W., & States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature genetics*, 3(3), 266–272
- Gompert, Z., Forister, M. L., Fordyce, J. A., Nice, C. C., Williamson, R. J., & Alex Buerkle, C. (2010). Bayesian analysis of molecular variance in

- pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology*, 19(12), 2455–2473. <https://doi.org/10.1111/j.1365-294X.2010.04666.x>
- Gonzalez-Arnan, M. T., Lazaro-Vallejo, C. E., Engelmann, F., Gamez-Pastrana, R., Martinez-Ocampo, Y. M., Pastelin-Solano, M. C., & Diaz-Ramos, C. (2009). Multiplication and cryopreservation of vanilla (*Vanilla planifolia* 'Andrews') Maria. *In Vitro Cellular and Developmental Biology - Plant*, 45(5), 574–582. <https://doi.org/10.1007/s11627-009-9242-6>
- Govindaraj, M., Vetriventhan, M., & Srinivasan, M. (2015). Importance of genetic diversity assessment in crop plants and its recent advances: An overview of its analytical perspectives. *Genetics Research International*, 2015(Figure 1). <https://doi.org/10.1155/2015/431487>
- Grisoni, M., Moles, M., Besse, P., Bory, S., Duval, M. F., & Kahane, R. (2007). Towards an international plant collection to maintain and characterize the endangered genetic resources of Vanilla. *Acta Horticulturae*, 760(September 2017), 83–90. <https://doi.org/10.17660/actahortic.2007.760.9>
- Grisoni, M., & Nany, F. (2021). The beautiful hills: half a century of vanilla (*Vanilla planifolia* Jacks. ex Andrews) breeding in Madagascar. *Genetic Resources and Crop Evolution*, 68(5), 1691–1708. <https://doi.org/10.1007/s10722-021-01119-2>
- Hamrick, J. L., & Godt, M. J. W. (1997). Allozyme diversity in cultivated crops. *Crop Science*, 37(1), 26–30. <https://doi.org/10.2135/cropsci1997.0011183X003700010004x>
- Hasan, N., Choudhary, S., Naaz, N., Sharma, N., & Laskar, R. A. (2021). Recent advancements in molecular marker-assisted selection and applications in plant breeding programmes. *Journal of Genetic Engineering and Biotechnology*, 19(1), 1–26. <https://doi.org/10.1186/s43141-021-00231-1>
- Hasing, T., Tang, H., Brym, M., Khazi, F., Huang, T., & Chambers, A. H. (2020). A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nature Food*, 1(12), 811–819. <https://doi.org/10.1038/s43016-020-00197-2>
- Havkin-frenkel, D., & Belanger, F. C. (2018). Mexican Vanilla Production. *Handbook of Vanilla Science and Technology*, 1–26.

References

- <https://doi.org/10.1002/9781119377320.ch1>
- Hedges, D. J., Guettouche, T., Yang, S., Bademci, G., Diaz, A., Andersen, A., Hulme, W. F., Linker, S., Mehta, A., Edwards, Y. J. K., Beecham, G. W., Martin, E. R., Pericak-Vance, M. A., Zuchner, S., Vance, J. M., & Gilbert, J. R. (2011). Comparison of three targeted enrichment strategies on the Solid sequencing platform. *PLoS ONE*, 6(4), 1–8. <https://doi.org/10.1371/journal.pone.0018595>
- Helyar, S. J., Hemmer-Hansen, J., Bekkevold, D., Taylor, M. I., Ogden, R., Limborg, M. T., Cariani, A., Maes, G. E., Diopere, E., Carvalho, G. R., & Nielsen, E. E. (2011). Application of SNPs for population genetics of nonmodel organisms: New opportunities and challenges. *Molecular Ecology Resources*, 11(SUPPL. 1), 123–136. <https://doi.org/10.1111/j.1755-0998.2010.02943.x>
- Henry, R. J. (2012). Molecular Markers in Plants. *Molecular Markers in Plants*. <https://doi.org/10.1002/9781118473023>
- Hernández-Ramírez, F., Dolce, N., Flores-Castaños, O., Rascón, M. P., Ángeles-Álvarez, G., Folgado, R., & González-Arno, M. T. (2020). Advances in cryopreservation of vanilla (*Vanilla planifolia* Jacks.) shoot-tips: assessment of new biotechnological and cryogenic factors. *In Vitro Cellular and Developmental Biology - Plant*, 56(2), 236–246. <https://doi.org/10.1007/s11627-020-10069-w>
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15), 1965–1978. <https://doi.org/10.1002/joc.1276>
- Hohenlohe, P. A., Bassham, S., Etter, P. D., Stiffler, N., Johnson, E. A., & Cresko, W. A. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, 6(2). <https://doi.org/10.1371/journal.pgen.1000862>
- Householder, E., Janovec, J., Mozambique, A. B., Maceda, J. H., Wells, J., Valega, R., ... & Christenson, E. (2010). Diversity, natural history, and conservation of Vanilla (Orchidaceae) in Amazonian wetlands of Madre de Dios, Peru. *Journal of the Botanical Research Institute of Texas*, 227–243.
- Hsieh, Y. C., Chung, J. D., Wang, C. N., Chang, C. T., Chen, C. Y., & Hwang,

- S. Y. (2013). Historical connectivity, contemporary isolation and local adaptation in a widespread but discontinuously distributed species endemic to Taiwan, *Rhododendron oldhamii* (Ericaceae). *Heredity*, 111(2), 147–156. <https://doi.org/10.1038/hdy.2013.31>
- Hu, Y., Resende, M. F. R., Bombarely, A., Brym, M., Bassil, E., & Chambers, A. H. (2019). Genomics-based diversity analysis of *Vanilla* species using a *Vanilla planifolia* draft genome and Genotyping-By-Sequencing. *Scientific Reports*, 9(1), 1–16. <https://doi.org/10.1038/s41598-019-40144-1>
- Huang, C. L., Chang, C. Te, Huang, B. H., Chung, J. Der, Chen, J. H., Chiang, Y. C., & Hwang, S. Y. (2015). Genetic relationships and ecological divergence in *Salix* species and populations in Taiwan. *Tree Genetics and Genomes*, 11(3). <https://doi.org/10.1007/s11295-015-0862-1>
- Hunt, R., Sauna, Z. E., Ambudkar, S. V., Gottesman, M. M., & Kimchi-Sarfaty, C. (2009). Silent (synonymous) SNPs: should we care about them? *Methods in Molecular Biology* (Clifton, N.J.), 578(1), 23–39. https://doi.org/10.1007/978-1-60327-411-1_2
- Hunter, M. E., Hoban, S. M., Bruford, M. W., Segelbacher, G., & Bernatchez, L. (2018). Next-generation conservation genetics and biodiversity monitoring. *Evolutionary Applications*, 11(7), 1029–1034. <https://doi.org/10.1111/eva.12661>
- Huston, M. (1979). A General Hypothesis of Species Diversity Author (s): Michael Huston Source : The American Naturalist , Vol . 113 , No . 1 (Jan . , 1979), pp . 81-101 Published by : The University of Chicago Press for The American Society of Naturalists Stable URL : h. *The American Naturalist*, 113(1), 81–101
- Jaiswal, V., Gupta, S., Gahlaut, V., Muthamilarasan, M., Bandyopadhyay, T., Ramchiary, N., & Prasad, M. (2019). Genome-Wide Association Study of Major Agronomic Traits in Foxtail Millet (*Setaria italica* L.) Using ddRAD Sequencing. *Scientific Reports*, 9(1), 1–11. <https://doi.org/10.1038/s41598-019-41602-6>
- Jiang, X. L., Gardner, E. M., Meng, H. H., Deng, M., & Xu, G. B. (2019). Land bridges in the Pleistocene contributed to flora assembly on the continental islands of South China: Insights from the evolutionary history of *Quercus championii*. *Molecular Phylogenetics and Evolution*,

References

- 132(August 2018), 36–45.
<https://doi.org/10.1016/j.ympev.2018.11.021>
- Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., & Dhawan, A. K. (2011). Microsatellite markers: An overview of the recent progress in plants. *Euphytica*, 177(3), 309–334. <https://doi.org/10.1007/s10681-010-0286-9>
- Karaköy, T., Baloch, F. S., Toklu, F., & Özkan, H. (2014). Variation for selected morphological and quality-related traits among 178 faba bean landraces collected from Turkey. *Plant Genetic Resources: Characterisation and Utilisation*, 12(1), 5–13. <https://doi.org/10.1017/S1479262113000208>
- Kess, T., Gross, J., Harper, F., & Boulding, E. G. (2016). Low-cost ddRAD method of SNP discovery and genotyping applied to the periwinkle *Littorina saxatilis*. *Journal of Molluscan Studies*, 82(1), 104–109. <https://doi.org/10.1093/mollus/eyv042>
- Khoury, C. K., Brush, S., Costich, D. E., Curry, H. A., de Haan, S., Engels, J. M. M., Guarino, L., Hoban, S., Mercer, K. L., Miller, A. J., Nabhan, G. P., Perales, H. R., Richards, C., Riggins, C., & Thormann, I. (2022). Crop genetic erosion: understanding and responding to loss of crop diversity. *New Phytologist*, 233(1), 84–118. <https://doi.org/10.1111/nph.17733>
- Kiesewetter, S., Macek Jr, M., Davis, C., Curristin, S. M., Chu, C.-S., Graham, C., Shrimpron, A. E., Witt, D., Crystal, R., & Cutting, G. (1993). © 1993 Nature Publishing Group <http://www.nature.com/naturegenetics> Nature Genetics, 3, 73–96.
- Kolanowska, M. (2013). Niche Conservatism and the Future Potential Range of *Epipactis helleborine* (Orchidaceae). *PLoS ONE*, 8(10), 1–8. <https://doi.org/10.1371/journal.pone.0077352>
- Koyyappurath, S., Atuahiva, T., Le Guen, R., Batina, H., Le Squin, S., Gautheron, N., Edel Hermann, V., Peribe, J., Jahiel, M., Steinberg, C., Liew, E. C. Y., Alabouvette, C., Besse, P., Dron, M., Sache, I., Laval, V., & Grisoni, M. (2016). *Fusarium oxysporum* f. sp. *radicis-vanillae* is the causal agent of root and stem rot of vanilla. *Plant Pathology*, 65(4), 612–625. <https://doi.org/10.1111/ppa.12445>
- Kronthaler, F., & Zöllner, S. (2021). Data Analysis with RStudio. In *Data*

- Analysis with RStudio*. <https://doi.org/10.1007/978-3-662-62518-7>
- Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. *International Journal of Plant Genomics*, 2012. <https://doi.org/10.1155/2012/831460>
- Laila, R., Park, J. I., Robin, A. H. K., Natarajan, S., Vijayakumar, H., Shirasawa, K., Isobe, S., Kim, H. T., & Nou, I. S. (2019). Mapping of a novel clubroot resistance QTL using ddRAD-seq in Chinese cabbage (*Brassica rapa* L.). *BMC Plant Biology*, 19(1), 1–9. <https://doi.org/10.1186/s12870-018-1615-8>
- Lasky, J. R., Des Marais, D. L., McKay, J. K., Richards, J. H., Juenger, T. E., & Keitt, T. H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: The roles of geography and climate. *Molecular Ecology*, 21(22), 5512–5529. <https://doi.org/10.1111/j.1365-294X.2012.05709.x>
- Lee, H. O., Davidson, J. M., & Duronio, R. J. (2010). Institute of Human Virology 2009 Journal Club Series Suzanne Gartner , PhD “ An Introduction to Endoreplication ” 725 West Lombard Street 3rd Floor Light Well Refreshments Endoreplication : polyploidy with purpose. *Genes & Development*, 23, 2461–2477. <https://doi.org/10.1101/gad.1829209.results>
- Leclercq Le Quillec, F., Riviere, C., and Lagorce, A. 2001. Spread of *Cymbidium* mosaic potexvirus and potyviruses in vanilla plants grown in shade houses in Reunion Is-land. *Fruits* 56:249-260
- Leimu, R., & Fischer, M. (2008). A meta-analysis of local adaptation in plants. *PLoS ONE*, 3(12), 1–8. <https://doi.org/10.1371/journal.pone.0004010>
- Lène, Hé, R., Benot, B., Darrier, Navarro, J., Kitt, J., Déric Choulet, F., Leveugle, M., Duarte, J., Davassi, A., Balfourier, F., Le Paslier, M.-C., Lie Berard, A., Brunel, D., Feuillet, C., Poncet, C., Sourdille, P., & Paux, E. (2018). High throughput SNP discovery and genotyping in hexaploid wheat. *PLoS ONE*, 3, 1–19. <https://doi.org/10.1371/journal.pone.0186329>
- Lepers-Andrzejewski, S., Causse, S., Caromel, B., Wong, M., & Dron, M.

References

- (2012). Genetic linkage map and diversity analysis of tahitian vanilla (*Vanilla xtahitensis*, orchidaceae). *Crop Science*, 52(2), 795–806. <https://doi.org/10.2135/cropsci2010.11.0634>
- Lepers-Andrzejewski, S., Siljak-Yakovlev, S., Brown, S. C., Wong, M., & Dron, M. (2011). Diversity and dynamics of plant genome size: An example of polysomaty from a cytogenetic study of Tahitian vanilla (*vanilla xtahitensis*, Orchidaceae). *American Journal of Botany*, 98(6), 986–997. <https://doi.org/10.3732/ajb.1000415>
- Letunic, I., & Bork, P. (2021). Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1), W293–W296. <https://doi.org/10.1093/nar/gkab301>
- Lewis, P. O., & Zaykin, D. (2002). GDA User ' s Manual Authors ' Contact Information.
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21), 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Oxford University Press* <http://github.com/lh3/bwa>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- 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. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20), 2843–2851. <https://doi.org/10.1093/bioinformatics/btu356>
- Li, J., Demesyeux, L., Brym, M., & Chambers, A. H. (2020). Development of species-specific molecular markers in *Vanilla* for seedling selection of hybrids. *Molecular Biology Reports*, 47(3), 1905–1920. <https://doi.org/10.1007/s11033-020-05287-9>

- Li, H., & Durbin, R. (2009a). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Li, H., & Wren, J. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, 30(20), 2843–2851. <https://doi.org/10.1093/bioinformatics/btu356>
- Li, X., Wang, X., Weng, G., Zhu, M., & Yan, K. (2022). Genome-wide identification and characterization of the fatty acid desaturase gene family in *Vanilla planifolia*. *South African Journal of Botany*, 150, 813–820. <https://doi.org/10.1016/j.sajb.2022.08.041>
- Lischer, H. (2011). PGDSpider version 2.1.1.5 (May 2018) An automated data conversion tool for connecting population genetics and genomics programs Manual PGDSpider ver 2.1.1.5. *Bioinformatics*, 28(2), 298–299. http://www.cmpg.unibe.ch/software/PGDSpider/PGDSpider_manual_vers_2-1-1-5.pdf
- Liu, W., Zhang, Y., Chen, X., Maung-Douglass, K., Strong, D. R., & Pennings, S. C. (2020). Contrasting plant adaptation strategies to latitude in the native and invasive range of *Spartina alterniflora*. *New Phytologist*, 226(2), 623–634. <https://doi.org/10.1111/nph.16371>
- Liu, Z. J., & Cordes, J. F. (2004). DNA marker technologies and their applications in aquaculture genetics. *Aquaculture*, 238(1–4), 1–37. <https://doi.org/10.1016/j.aquaculture.2004.05.027>
- Lubinsky, P., Bory, S., Hernández Hernández, J., Kim, S. C., & Gómez-Pompa, A. (2008). Origins and dispersal of cultivated vanilla (*Vanilla planifolia* Jacks. [Orchidaceae]). *Economic Botany*, 62(2), 127–138. <https://doi.org/10.1007/s12231-008-9014-y>
- Lubinsky, P., Romero-González, G. A., Heredia, S. M., & Zabel, S. (2010). Origins and Patterns of Vanilla Cultivation in Tropical America (1500–1900): No Support for an Independent Domestication of Vanilla in South America. *Handbook of Vanilla Science and Technology*, 117–138. <https://doi.org/10.1002/9781444329353.ch8>

References

- Lischer, H. (2011). PGDSpider version 2.1.1.5 (May 2018) An automated data conversion tool for connecting population genetics and genomics programs Manual PGDSpider ver 2.1.1.5. *Bioinformatics*, 28(2), 298–299. http://www.cmpg.unibe.ch/software/PGDSpider/PGDSpider_manual_vers_2-1-1-5.pdf
- Lynch, M., & Milligan, B. G. (1994). Analysis of population genetic structure with RAPD markers. *Molecular Ecology*, 3(2), 91–99. <https://doi.org/10.1111/j.1365-294X.1994.tb00109.x>
- Magige, E. A., Fan, P. Z., Wambulwa, M. C., Milne, R., Wu, Z. Y., Luo, Y. H., Khan, R., Wu, H. Y., Qi, H. L., Zhu, G. F., Maity, D., Khan, I., Gao, L. M., & Liu, J. (2022). Genetic Diversity and Structure of Persian Walnut (*Juglans regia* L.) in Pakistan: Implications for Conservation. *Plants*, 11(13), 1–14. <https://doi.org/10.3390/plants11131652>
- Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, 15(1), 28–41. <https://doi.org/10.1111/1755-0998.12291>
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–564. <https://doi.org/10.1073/pnas.74.2.560>
- McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1). <https://doi.org/10.1186/s13059-016-0974-4>
- Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, 5(6). <https://doi.org/10.1101/pdb.prot5448>
- Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, 17(2), 240–248. <https://doi.org/10.1101/gr.5681207>
- Minoo, D., Jayakumar, V. N., Veena, S. S., Vimala, J., Basha, A., Saji, K. V.,

- Nirmal Babu, K., & Peter, K. V. (2008). Genetic variations and interrelationships in *Vanilla planifolia* and few related species as expressed by RAPD polymorphism. *Genetic Resources and Crop Evolution*, 55(3), 459–470. <https://doi.org/10.1007/s10722-007-9252-3>
- Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., Prins, B., Reynolds, A., Chia, J. M., Ware, D., Bustamante, C. D., & Buckler, E. S. (2011). Genetic structure and domestication history of the grape. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9), 3530–3535. <https://doi.org/10.1073/pnas.1009363108>
- Nair, R., & Ravindran, P. (1994). *Somatic association of chromosomes and other mitotic abnormalities in Vanilla planifolia (Andrews)*, *Caryologia*, 47(1), 65–73. <https://doi.org/10.1080/00087114.1994.10797284>
- Natarajan, S., Hossain, M. R., Kim, H. T., Denison, M. I. J., Ferdous, M. J., Jung, H. J., Park, J. I., & Nou, I. S. (2020). ddRAD-seq derived genome-wide SNPs, high density linkage map and QTLs for fruit quality traits in strawberry (*Fragaria x ananassa*). *3 Biotech*, 10(8), 1–18. <https://doi.org/10.1007/s13205-020-02291-5>
- Nayak, S., Naik, P. K., Acharya, L., Mukherjee, A. K., Panda, P. C., & Das, P. (2005). Assessment of genetic diversity among 16 promising cultivars of ginger using cytological and molecular markers. *Zeitschrift Fur Naturforschung - Section C Journal of Biosciences*, 60(5–6), 485–492. <https://doi.org/10.1515/znc-2005-5-618>
- Naz, S., Zafrullah, A., Shahzadhi, K., & Munir, N. (2013). Assessment of genetic diversity within germplasm accessions in tomato using morphological and molecular markers. *Journal of Animal and Plant Sciences*, 23(4), 1099–1106.
- Nielsen, L. R. (2000). *Sw . and K barbellata Rchb . f . pollination experimental data*. 285–302
- Nielsen, L. R., & Siegmund, H. R. (1999). Interspecific differentiation and hybridization in *Vanilla* species (Orchidaceae). *Heredity*, 83(5), 560–567. <https://doi.org/10.1038/sj.hdy.6885880>
- Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and

References

- SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6), 443–451. <https://doi.org/10.1038/nrg2986>
- Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*, 7(7). <https://doi.org/10.1371/journal.pone.0037558>
- O'Donnell, M. S., & Ignizio, D. A. (2012). Bioclimatic Predictors for Supporting Ecological Applications in the Conterminous United States. *U.S Geological Survey Data Series 691*, 10
- Odoux, E., & Brillouet, J. M. (2009). Anatomy, histochemistry and biochemistry of glucovanillin, oleoresin and mucilage accumulation sites in green mature vanilla pod (*Vanilla planifolia*; Orchidaceae): A comprehensive and critical reexamination. *Fruits*, 64(4), 221–241. <https://doi.org/10.1051/fruits/2009017>
- Oduor, A. M. O., Leimu, R., & van Kleunen, M. (2016). Invasive plant species are locally adapted just as frequently and at least as strongly as native plant species. *Journal of Ecology*, 104(4), 957–968. <https://doi.org/10.1111/1365-2745.12578>
- Oksanen, J., Blanchet, F. G., & Maintainer, H. W. (2019). Package “vegan” Title Community Ecology Package. *Community ecology package*, 2, 1–297
- Olsen, K. M., & Wendel, J. F. (2013). Crop plants as models for understanding plant adaptation and diversification. *Frontiers in Plant Science*, 4(AUG). <https://doi.org/10.3389/fpls.2013.00290>
- Ongaro, S., Martellos, S., Bacaro, G., De Agostini, A., Cogoni, A., & Cortis, P. (2018). Distributional pattern of sardinian orchids under a climate change scenario. *Community Ecology*, 19(3), 223–232. <https://doi.org/10.1556/168.2018.19.3.3>
- Parthasarathy, V. A., Chempakam, B., & Zachariah, T. J. (2008). Chemistry of spices. In *Chemistry of Spices*. <https://doi.org/10.4327/jsnfs1949.32.267>
- Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2). <https://doi.org/10.1371/journal.pone.0030619>

- Pegadaraju, V., Nipper, R., Hulke, B., Qi, L., & Schultz, Q. (2013). De novo sequencing of sunflower genome for SNP discovery using RAD (Restriction site Associated DNA) approach. *BMC Genomics*, 14(1). <https://doi.org/10.1186/1471-2164-14-556>
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE*, 7(5). <https://doi.org/10.1371/journal.pone.0037135>
- Pham, T. D., Geleta, M., Bui, T. M., Bui, T. C., Merker, A., & Carlsson, A. S. (2011). Comparative analysis of genetic diversity of sesame (*Sesamum indicum* L.) from Vietnam and Cambodia using agro-morphological and molecular markers. *Hereditas*, 148(1), 28–35. <https://doi.org/10.1111/j.1601-5223.2010.02196.x>
- Philip, V. J., & Nainar, S. A. Z. (1986). Clonal Propagation of *Vanilla planifolia* (Salisb.) Ames using Tissue Culture. *Journal of Plant Physiology*, 122(3), 211–215. [https://doi.org/10.1016/S0176-1617\(86\)80119-5](https://doi.org/10.1016/S0176-1617(86)80119-5)
- Pillon, Y., Qamaruz-Zaman, F., Fay, M. F., Hendoux, F., & Piquot, Y. (2007). Genetic diversity and ecological differentiation in the endangered fen orchid (*Liparis loeselii*). *Conservation Genetics*, 8(1), 177–184. <https://doi.org/10.1007/s10592-006-9160-7>
- Piet, Q., Droc, G., Marande, W., Sarah, G., Bourgois, L., Zucca, J., Dron, M., Besse, P., & Grisoni, M. (2022). A chromosome-level , haplotype-phased *Vanilla planifolia* genome highlights the challenge of partial endoreplication for accurate whole- genome assembly. 1–15. <https://doi.org/10.1016/j.xplc.2022.100330>
- Poggio, L., De Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *Soil*, 7(1), 217–240. <https://doi.org/10.5194/soil-7-217-2021>
- Poncet, B. N., Herrmann, D., Gugerli, F., Taberlet, P., Holderegger, R., Gielly, L., Rioux, D., Thuiller, W., Aubert, S., & Manel, S. (2010). Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. *Molecular Ecology*, 19(14), 2896–2907. <https://doi.org/10.1111/j.1365-294X.2010.04696.x>
- Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O.,

References

- Van der Auwera, G. A., ... & Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178
- Porras-Alfaro, A., & Bayman, P. (2007). Mycorrhizal fungi of Vanilla : diversity, specificity and effects on seed germination and plant growth . *Mycologia*, 99(4), 510–525. <https://doi.org/10.1080/15572536.2007.11832545>
- Pournosrat, R., Kaya, S., Shaaf, S., Kilian, B., & Ozkan, H. (2018). Geographical and environmental determinants of the genetic structure of wild barley in southeastern Anatolia. *PLoS ONE*, 13(2), 1–19. <https://doi.org/10.1371/journal.pone.0192386>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, 23(24), 5937–5942. <https://doi.org/10.1111/mec.12965>
- Pusadee, T., Jamjod, S., Chiang, Y. C., Rerkasem, B., & Schaal, B. A. (2009). Genetic structure and isolation by distance in a landrace of Thai rice. In *Proceedings of the National Academy of Sciences of the United States of America* (Vol. 106, Issue 33, pp. 13880–13885). <https://doi.org/10.1073/pnas.0906720106>
- QGIS. (2022). *QGIS Desktop 3.16 User Guide QGIS Project*. 142
- Quiro, E. V. (2018). Vanilla Production in Costa Rica. *Handbook of Vanilla Science and Technology*, Ocampo 1987, 41–51. <https://doi.org/10.1002/9781119377320.ch3>
- Rajendrakumar, P. (2015). Sorghum Molecular Breeding. *Sorghum Molecular Breeding*. <https://doi.org/10.1007/978-81-322-2422-8>
- Ramos-Castellá, A. L., Iglesias-Andreu, L. G., Martínez-Castillo, J., Ortíz-García, M., Andueza-Noh, R. H., Octavio-Aguilar, P., & Luna-Rodríguez, M. (2017). Evaluation of molecular variability in germplasm of vanilla (*Vanilla planifolia* G. Jackson in Andrews) in Southeast Mexico: Implications for genetic improvement and conservation. *Plant Genetic Resources*: 15(4), 310-320

- Rao, S. R., & Ravishankar, G. A. (2000). Vanilla flavour: Production by conventional and biotechnological routes. *Journal of the Science of Food and Agriculture*, 80(3), 289–304. [https://doi.org/10.1002/1097-0010\(200002\)80:3<289::aid-jsfa543>3.0.co;2-2](https://doi.org/10.1002/1097-0010(200002)80:3<289::aid-jsfa543>3.0.co;2-2)
- Rao, X., Krom, N., Tang, Y., Widiez, T., Havkin-Frenkel, D., Belanger, F. C., Dixon, R. A., & Chen, F. (2014). A deep transcriptomic analysis of pod development in the vanilla orchid (*Vanilla planifolia*). *BMC Genomics*, 15(1), 1–12. <https://doi.org/10.1186/1471-2164-15-964>
- Renaud, G., Stenzel, U., Maricic, T., Wiebe, V., & Kelso, J. (2015). DeML: Robust demultiplexing of Illumina sequences using a likelihood-based approach. *Bioinformatics*, 31(5), 770–772. <https://doi.org/10.1093/bioinformatics/btu719>
- Riordan, E. C., Gugger, P. F., Ortego, J., Smith, C., Gaddis, K., Thompson, P., & Sork, V. L. (2016). Association of genetic and phenotypic variability with geography and climate in three southern California oaks. *American Journal of Botany*, 103(1), 73–85. <https://doi.org/10.3732/ajb.1500135>
- Rochette, N. C., & Catchen, J. M. (2017). Deriving genotypes from RAD-seq short-read data using Stacks. *Nature Protocols*, 12(12), 2640–2659. <https://doi.org/10.1038/nprot.2017.123>
- Ruggiero, M. V., Capone, S., Pirozzi, P., Reusch, T. B. H., & Procaccini, G. (2005). Mating system and clonal architecture: A comparative study in two marine angiosperms. *Evolutionary Ecology*, 19(5), 487–499. <https://doi.org/10.1007/s10682-005-0292-x>
- Rojas-López, A., & Cañizares-Macías, M. P. (2013). Antioxidant Capacity in Vanilla Extracts Obtained by Applying Focused Microwaves. *Food and Nutrition Sciences*, 04(08), 244–253. <https://doi.org/10.4236/fns.2013.48a030>
- Rothberg, J. M., & Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nature Biotechnology*, 26(10), 1117–1124. <https://doi.org/10.1038/nbt1485>
- Roy, S. C., Moitra, K., & De Sarker, D. (2017). Assessment of genetic diversity among four orchids based on ddRAD sequencing data for conservation purposes. In *Physiology and Molecular Biology of Plants* (Vol. 23, Issue 1, pp. 169–183). <https://doi.org/10.1007/s12298-016-0401-z>

References

- Ryan, P., Valentin, R.-R., Mark, A. D., Tim, J. F., Mauricio, O. C., Geraldine, A. V. der A., Davird, E. K., Laura, D. G., Ami, L.-M., David, R., Khalid, S., Joel, T., Sheila, C., Chris, W., Monkol, L., Stacey, G., Mark, J. D., Ben, N., Daniel, G. M., & Eric, B. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 1–22
- Saenz-Agudelo, P., DiBattista, J. D., Piatek, M. J., Gaither, M. R., Harrison, H. B., Nanninga, G. B., & Berumen, M. L. (2015). Seascape genetics along environmental gradients in the Arabian Peninsula: Insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, 24(24), 6241–6255. <https://doi.org/10.1111/mec.13471>
- Saintenac, C., Jiang, D., Wang, S., & Akhunov, E. (2013). Sequence-based mapping of the polyploid wheat genome. *G3: Genes, Genomes, Genetics*, 3(7), 1105–1114. <https://doi.org/10.1534/g3.113.005819>
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12), 5463–5467
- Sasikumar, B. (2010). Vanilla Breeding - a Review. *Agricultural Reviews*, 31(2), 139–144
- Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecology Letters*, 9(5), 615–629. <https://doi.org/10.1111/j.1461-0248.2006.00889.x>
- Shaaf, S., Sharma, R., Kilian, B., Walther, A., Özkan, H., Karami, E., & Mohammadi, B. (2014). Genetic structure and eco-geographical adaptation of garlic landraces (*Allium sativum* L.) in Iran. *Genetic Resources and Crop Evolution*, 61(8), 1565–1580. <https://doi.org/10.1007/s10722-014-0131-4>
- Shirasawa, K., Hirakawa, H., & Isobe, S. (2016). Analytical workflow of double-digest restriction site-associated DNA sequencing based on empirical and in silico optimization in tomato. *DNA Research*, 23(2), 145–153. <https://doi.org/10.1093/dnares/dsw004>
- Simón-Porcar, V. I., Silva, J. L., & Vallejo-Marín, M. (2021). Rapid local adaptation in both sexual and asexual invasive populations of monkeyflowers (*Mimulus* spp.). *Annals of Botany*, 127(5), 655–668. <https://doi.org/10.1093/aob/mcab004>
- Snoussi Trifa, H. (2013). Diversité génétique intra et interspécifique des

- porte-greffes d'agrumes utilisés en Tunisie (Doctoral dissertation, INAT)
- Šmarda, P., Bureš, P., Horová, L., Leitch, I. J., Mucina, L., Pacini, E., Tichý, L., Grulich, V., & Rotreklová, O. (2014). Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proceedings of the National Academy of Sciences of the United States of America*, 111(39), E4096–E4102. <https://doi.org/10.1073/pnas.1321152111>
- Solano-De La Cruz, M. T., Adame-García, J., Gregorio-Jorge, J., Jiménez-Jacinto, V., Vega-Alvarado, L., Iglesias-Andreu, L. G., Escobar-Hernández, E. E., & Luna-Rodríguez, M. (2019). Functional categorization of de novo transcriptome assembly of *Vanilla planifolia* Jacks. potentially points to a translational regulation during early stages of infection by *Fusarium oxysporum* f. sp. *vanillae*. *BMC Genomics*, 20(1), 1–15. <https://doi.org/10.1186/s12864-019-6229-5>
- Soto Arenas, M. A. (1999). Filogeografía y recursos genéticos de las vainillas de México. *México, DF*
- Soto Arenas, M. A., & Cribb, P. (2013). A new infrageneric classification and synopsis of the genus *Vanilla* Plum. ex mill. (Orchidaceae: Vanillinae). *Lankesteriana*, 9(3), 355–398. <https://doi.org/10.15517/lank.v0i0.12071>
- Soto Arenas, M. A., & Dressler, R. L. (2013). A revision of the Mexican and Central American species of *Vanilla* plumier ex miller with a characterization of their region of the nuclear ribosomal DNA. *Lankesteriana*, 9(3). <https://doi.org/10.15517/lank.v0i0.12065>
- Sreedhar, R., Venkatachalam, L., Roohie, K., & Bhagyalakshmi, N. (2007). Molecular Analyses of *Vanilla planifolia* Cultivated in India using RAPD and ISSR Markers. *Orchid Science and Biotechnology*, 1(1), 29–33
- Stift, M., Kolář, F., & Meirmans, P. G. (2019). Structure is more robust than other clustering methods in simulated mixed-ploidy populations. In *Heredity* (Vol. 123, Issue 4, pp. 429–441). <https://doi.org/10.1038/s41437-019-0247-6>
- Stuber, C. W., Lincoln, S. E., Wolff, D. W., Helentjaris, T., & Lander, E. S. (1992). Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers.

References

- Genetics*, 132(3), 823–839.
<https://doi.org/10.1093/genetics/132.3.823>
- Sudan, J., Singh, R., Sharma, S., Salgotra, R. K., Sharma, V., Singh, G., Sharma, I., Sharma, S., Gupta, S. K., & Zargar, S. M. (2019). DdRAD sequencing-based identification of inter-genepool SNPs and association analysis in *Brassica juncea*. *BMC Plant Biology*, 19(1), 1–15. <https://doi.org/10.1186/s12870-019-2188-x>
- Swarts, N. D., Sinclair, E. A., Krauss, S. L., & Dixon, K. W. (2009). Genetic diversity in fragmented populations of the critically endangered spider orchid *Caladenia huegelii*: Implications for conservation. *Conservation Genetics*, 10(5), 1199–1208. <https://doi.org/10.1007/s10592-008-9651-9>
- Swarup, S., Cargill, E. J., Crosby, K., Flagel, L., Kniskern, J., & Glenn, K. C. (2021). Genetic diversity is indispensable for plant breeding to improve crops. *Crop Science*, 61(2), 839–852. <https://doi.org/10.1002/csc2.20377>
- Syvänen, A. (2001). Principles of SNP-genotyping methods. *Nature Reviews Genetics*, 2(December), 930–942
- Szlachetko, D., & Veyret, Y. (1994). Deux espèces nouvelles de *Vanilla* (Orchidaceae) de Guyane française. *Bulletin Du Muséum National d'histoire Naturelle. Section B, Adansonia*, 16(2), 219–223
- Thomson, M. J. (2014). High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breeding and Biotechnology*, 2(3), 195–212. <https://doi.org/10.9787/pbb.2014.2.3.195>
- Tonnier, J. P. (1960). *La fusariose du vanillier à Madagascar*. 138. *Rapport du laboratoire du vanillier de l'Ivoloïna*
- Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: A simplified method for genomic genotyping in non-model organisms. *PeerJ*, 2013(1), 1–15. <https://doi.org/10.7717/peerj.203>
- Tsiftsis, S., & Djordjević, V. (2020). Modelling sexually deceptive orchid species distributions under future climates: the importance of plant–pollinator interactions. *Scientific Reports*, 10(1), 1–12. <https://doi.org/10.1038/s41598-020-67491-8>

- Vences, M., Wollenberg, K. C., Vieites, D. R., & Lees, D. C. (2009). Madagascar as a model region of species diversification. *Trends in Ecology and Evolution*, 24(8), 456–465. <https://doi.org/10.1016/j.tree.2009.03.011>
- Verma, P. C., Chakrabarty, D., Jena, S. N., Mishra, D. K., Singh, P. K., Sawant, S. V., & Tuli, R. (2009). The extent of genetic diversity among *Vanilla* species: Comparative results for RAPD and ISSR. *Industrial Crops and Products*, 29(2–3), 581–589. <https://doi.org/10.1016/j.indcrop.2008.11.006>
- Villanueva-Viramontes, S., Hernández-Apolinar, M., Carnevali Fernández-Concha, G., Dorantes-Euán, A., Dzib, G. R., & Martínez-Castillo, J. (2017). Wild *Vanilla planifolia* and its relatives in the Mexican Yucatan Peninsula: Systematic analyses with ISSR and ITS. *Botanical Sciences*, 95(2), 169–187. <https://doi.org/10.17129/botsoci.668>
- Van, S., & Kerke, D. E. (2014). Systematics of African Vanilla Orchids Linking Morphological Classification and Dna-Based Phylogeny
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T. Van De, Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M., & Zabeau, M. (1995). AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research*, 23(21), 4407–4414. <https://doi.org/10.1093/nar/23.21.4407>
- Wall, J. D., Tang, L. F., Zerbe, B., Kvale, M. N., Kwok, P. Y., Schaefer, C., & Risch, N. (2014). Estimating genotype error rates from high-coverage next-generation sequence data. *Genome Research*, 24(11), 1734–1739. <https://doi.org/10.1101/gr.168393.113>
- Walton, N. J., Mayer, M. J., & Narbad, A. (2003). Vanillin. *Phytochemistry*, 63(5), 505–515. [https://doi.org/10.1016/S0031-9422\(03\)00149-3](https://doi.org/10.1016/S0031-9422(03)00149-3)
- Wanderley, A. M., Machado, I. C. S., de Almeida, E. M., Felix, L. P., Galetto, L., Benko-Iseppon, A. M., & Sork, V. L. (2018). The roles of geography and environment in divergence within and between two closely related plant species inhabiting an island-like habitat. *Journal of Biogeography*, 45(2), 381–393. <https://doi.org/10.1111/jbi.13137>
- Wang, I. J., & Bradburd, G. S. (2014). Isolation by environment. *Molecular Ecology*, 23(23), 5649–5662. <https://doi.org/10.1111/mec.12938>
- Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nature Methods*,

References

- 9(8), 808–810. <https://doi.org/10.1038/nmeth.2023>
- Wang, M., Tu, L., Yuan, D., Zhu, D., Shen, C., Li, J., Liu, F., Pei, L., Wang, P., Zhao, G., Ye, Z., Huang, H., Yan, F., Ma, Y., Zhang, L., Liu, M., You, J., Yang, Y., Liu, Z., ... Zhang, X. (2019). Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nature Genetics*, 51(2), 224–229. <https://doi.org/10.1038/s41588-018-0282-x>
- Watteyn, C., Fremout, T., Karremans, A. P., Huarcaya, R. P., Azofeifa Bolaños, J. B., Reubens, B., & Muys, B. (2020). Vanilla distribution modeling for conservation and sustainable cultivation in a joint land sparing/sharing concept. *Ecosphere*, 11(3). <https://doi.org/10.1002/ecs2.3056>
- Weber, D., & Helentjaris, T. (1989). Mapping RFLP loci in maize using B-A translocations. *Genetics*, 121(3), 583–590. <https://doi.org/10.1093/genetics/121.3.583>
- Wei, S., Li, Z., Momigliano, P., Fu, C., Wu, H., & Merilä, J. (2020). The roles of climate, geography and natural selection as drivers of genetic and phenotypic differentiation in a widespread amphibian *Hyla annectans* (Anura: Hylidae). *Molecular Ecology*, 29(19), 3667–3683. <https://doi.org/10.1111/mec.15584>
- Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., & Zemla, J. (2017). R package “corrplot”: Visualization of a Correlation Matrix. *Statistician*, 56, 316–324. <https://github.com/taiyun/corrplot%0Ahttps://github.com/taiyun/corrplot/issues>
- Weir, B., & Cockerham, C. (2011). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6), 1358–1370. <http://www.jstor.org/stable/2408641>
- Whitehead I., M. I., Hillard, T. C., & Crook, D. (1990). The role and use of progestogens. *Obstetrics and Gynecology*, 75(4), 59S-76S. <https://doi.org/10.1097/00006250-199004001-00013>
- Will, K. W., & Rubinoff, D. (2004). Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. *Cladistics*, 20(1), 47–55. <https://doi.org/10.1111/j.1096-0031.2003.00008>

- Wright, B., Farquharson, K. A., McLennan, E. A., Belov, K., Hogg, C. J., & Grueber, C. E. (2019). From reference genomes to population genomics: Comparing three reference-aligned reduced-representation sequencing pipelines in two wildlife species. *BMC Genomics*, 20(1), 1–11. <https://doi.org/10.1186/s12864-019-5806-y>
- Yamasaki, M., Wright, S. I., & McMullen, M. D. (2007). Genomic screening for artificial selection during domestication and improvement in maize. *Annals of Botany*, 100(5), 967–973. <https://doi.org/10.1093/aob/mcm173>
- Yang, H., Barros-Rios, J., Kourteva, G., Rao, X., Chen, F., Shen, H., Liu, C., Podstolski, A., Belanger, F., Havkin-Frenkel, D., & Dixon, R. A. (2017). A re-evaluation of the final step of vanillin biosynthesis in the orchid *Vanilla planifolia*. *Phytochemistry*, 139, 33–46. <https://doi.org/10.1016/j.phytochem.2017.04.003>
- Yang, X., Liu, D., Liu, F., Wu, J., Zou, J., Xiao, X., Zhao, F., & Zhu, B. (2013). HTQC: A fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics*, 14(1), 2–5. <https://doi.org/10.1186/1471-2105-14-33>
- Zhang, Q., Zhou, Z. Q., Yang, G. P., Xu, C. G., Liu, K. D., & Saghai Maroof, M. A. (1996). Molecular marker heterozygosity and hybrid performance in indica and japonica rice. *Theoretical and Applied Genetics*, 93(8), 1218–1224. <https://doi.org/10.1007/BF00223453>
- Zion Market Research. Vanilla market. <https://www.zionmarketresearch.com/report/vanilla-market>. Consulted on October, 2022
- Zur, G. E., Orchidaceen, C. D. E. R., Hoffmann, K. M. A. X., & Einleitung, I. (1930). (*Eingegangen am 21. Januar 1930.*) I. *Einleitung*. 1915
- Zuraida, A., Hassan, K., Liyana, F., Nazreena, O., Sembok, W., Zaliha, W., Mohd, C., Che, Z., Zamri, Z., & Sreeramanan, S. (2013). A Simple and Efficient Protocol for the Mass Propagation of *Vanilla planifolia*. *In Vitro*, 2013(September), 1685–1692

Appendix A

Table A.1. Genomic location of the 21 701 common SNPs detected by Stacks, Bcftools, and GATK.

Consequence type	Count
Intron SNP	11 039
Intergenic SNPs	5 077
Upstream gene SNPs	2 627
Downstream gene SNPs	1 934
Missense SNPs	1 444
Synonymous SNPs	1 385
Splice region SNPs	169
Stop gained SNPs	16
Stop lost SNPs	3
Splice donor SNPs	3
Splice acceptor SNPs	1
Start lost	1

Appendix A

Table A.2. List of the 27 environmental variables. Data from WorldClim and SoilGrids were acquired at ~ 30 arcseconds [~1 km] resolution. Each variable from WorldClim is described in (O'Donnell & Ignizio, 2012).

Variable	Code	Database	Unit
Annual Mean Temperature	AMT	WorldClim	°C
Mean Diurnal Range	MDR	WorldClim	°C
Isothermality	ISO	WorldClim	
Temperature Seasonality	TS	WorldClim	°C
Max Temperature of Warmest Month	MaxTWM	WorldClim	°C
Min Temperature of Coldest Month	MinTCM	WorldClim	°C
Temperature Annual Range	TAR	WorldClim	°C
Mean Temperature of Wettest Quarter	MTWeQ	WorldClim	°C
Mean Temperature of Driest Quarter	MTDrQ	WorldClim	°C
Mean Temperature of Warmest Quarter	MTWaQ	WorldClim	°C
Mean Temperature of Coldest Quarter	MTCoQ	WorldClim	°C
Annual Precipitation	AP	WorldClim	mm
Precipitation of Wettest Month	PWeM	WorldClim	mm
Precipitation of Driest Month	PDrM	WorldClim	mm
Precipitation Seasonality	PS	WorldClim	mm
Precipitation of Wettest Quarter	PWeQ	WorldClim	mm
Precipitation of Driest Quarter	PDrQ	WorldClim	mm
Precipitation of Warmest Quarter	PWaQ	WorldClim	mm
Precipitation of Coldest Quarter	PCoQ	WorldClim	mm
Clay content	Clay	SoilGrids	%
Nitrogen content	N	SoilGrids	%
pH	pH	SoilGrids	
Sand content	Sand	SoilGrids	%
Silt content	Silt	SoilGrids	%
NDVI (normalized difference vegetation index)	NDVI	CGLS	
Distance from the ancient breeding program	D.Bree	GPS	
Altitude	ALT	GPS	

Table A.3. Identified private alleles in each of the four studied zones using the 9,281 SNPs from the 154 accessions of *V. planifolia* group.

Locus	Allele	Frequency	Zone	Locus	Allele	Frequency	Zone
SNP 1	G	0.105	ABJ	SNP 39	T	0.025	MNN
SNP 2	C	0.111	ABJ	SNP 40	G	0.025	MNN
SNP 3	C	0.053	ABJ	SNP 41	T	0.025	MNN
SNP 4	T	0.065	ABJ	SNP 42	A	0.026	MNN
SNP 5	A	0.107	ABJ	SNP 43	A	0.028	MNN
SNP 6	A	0.107	MNK	SNP 44	A	0.026	MNN
SNP 7	C	0.0600	MNK	SNP 45	A	0.026	MNN
SNP 8	G	0.068	MNK	SNP 46	C	0.083	MNN
SNP 9	T	0.057	MNK	SNP 47	A	0.024	MNN
SNP 10	A	0.055	MNK	SNP 48	T	0.024	MNN
SNP 11	T	0.068	MNK	SNP 49	T	0.027	MNN
SNP 12	A	0.060	MNK	SNP 50	G	0.023	MNN
SNP 13	A	0.095	MNK	SNP 51	T	0.051	MNN
SNP 14	A	0.113	MNK	SNP 52	C	0.025	MNN
SNP 15	G	0.071	MNK	SNP 53	T	0.024	MNN
SNP 16	C	0.068	MNK	SNP 54	A	0.026	MNN
SNP 17	T	0.057	MNK	SNP 55	C	0.024	MNN
SNP 18	G	0.057	MNK	SNP 56	A	0.025	MNN
SNP 19	A	0.057	MNK	SNP 57	C	0.025	MNN
SNP 20	T	0.071	MNK	SNP 58	G	0.026	MNN
SNP 21	C	0.055	MNK	SNP 59	T	0.026	MNN
SNP 22	G	0.068	MNK	SNP 60	G	0.025	MNN
SNP 23	T	0.060	MNK	SNP 61	A	0.032	MNN
SNP 24	T	0.068	MNK	SNP 62	T	0.025	MNN
SNP 25	C	0.055	MNK	SNP 63	C	0.027	MNN
SNP 26	T	0.088	MNK	SNP 64	T	0.028	MNN
SNP 27	G	0.071	MNK	SNP 65	A	0.032	MNN
SNP 28	G	0.053	MNK	SNP 66	T	0.032	MNN
SNP 29	T	0.076	MNK	SNP 67	A	0.02	MNN
SNP 30	C	0.057	MNK	SNP 68	A	0.033	MNN
SNP 31	A	0.071	MNK	SNP 69	A	0.024	MNN
SNP 32	C	0.060	MNK	SNP 70	C	0.034	MNN
SNP 33	C	0.025	MNN	SNP 71	A	0.023	MNN
SNP 34	A	0.037	MNN	SNP 72	G	0.024	MNN
SNP 35	T	0.024	MNN	SNP 73	A	0.026	MNN
SNP 36	T	0.026	MNN	SNP 74	A	0.025	MNN
SNP 37	C	0.025	MNN	SNP 75	A	0.096	SAV
SNP 38	A	0.025	MNN				

Appendix A

Table A.4. Pairwise comparison (*Fst*) between the 5 genetic groups using the 17,516 genotyped SNPs from the 208 polled accessions.

	V. <i>planifolia</i>	V. <i>pompona</i>	Intermediate1	Intermediate2	Big Vanilla
<i>V. planifolia</i>	0				
<i>V. pompona</i>	0.264	0			
Intermediate1	0.174	0.170	0		
Intermediate2	0.180	0.013	0.137	0	
Big Vanilla	0.615	0.176	0.286	0.49	0

Table A.5. Pairwise comparison (*Fst*) between the 5 genetic groups using the 17,516 genotyped SNPs from the 208 polled accessions.

	V. <i>planifolia</i>	V. <i>pompona</i>	Intermediate1	Intermediate2	Big Vanilla
<i>V. planifolia</i>	0				
<i>V. pompona</i>	0.264	0			
Intermediate1	0.174	0.170	0		
Intermediate2	0.180	0.013	0.137	0	
Big Vanilla	0.615	0.176	0.286	0.49	0

Table A.6. Correlation matrix of climatic and spatial variables. Here, only the retained variables from forward selection from RDA were employed.

dbME M6	dbME M5	dbME M4	dbME M1	Sand1	pH1	pWeQ	PDrM	MTWa Q	MTWe Q	TAR	elevati on	
0	0.484	0	0.110	0.415	0	0	0	0	0	0.506	1	elevati on
0	0	0	0.352	0.716	0	0.437	0	0	0	1		TAR
0.359	0	0.080	0	0	0.5741	0	0	0.998				MTWe Q
0.337	0	0.056	0	0	0.580	0	0	1				MTWa Q
0.195	0	0.370	0.137	0	0	0	1					PDrM
0.019	0	0	0.410	0.164	0.018	1						pWeQ
0.320	0	0	0	0	1							pH1
-0.548	0.432	0.107	0.433	1								Sand1
0	0	0	1									dbME M1
0	0	1										dbME M4
0	1											dbME M5
1												dbME M6

Appendix B

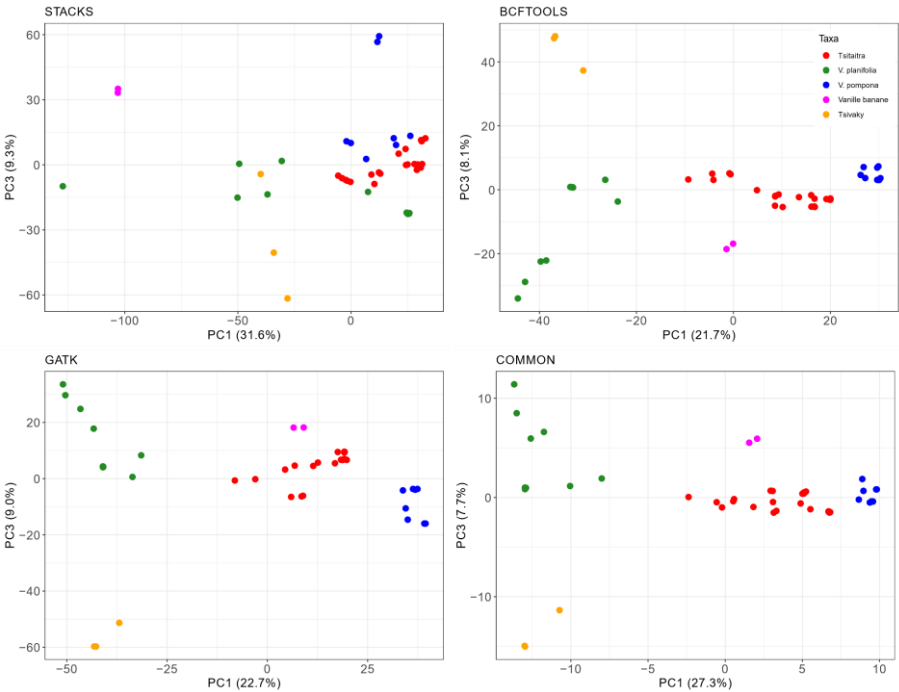


Figure B.1. First and third principal components of the PCA analysis of the 45 vanilla accessions from Stacks, Bcftools, GATK and the 23701 common SNPs detected by the three protocols.

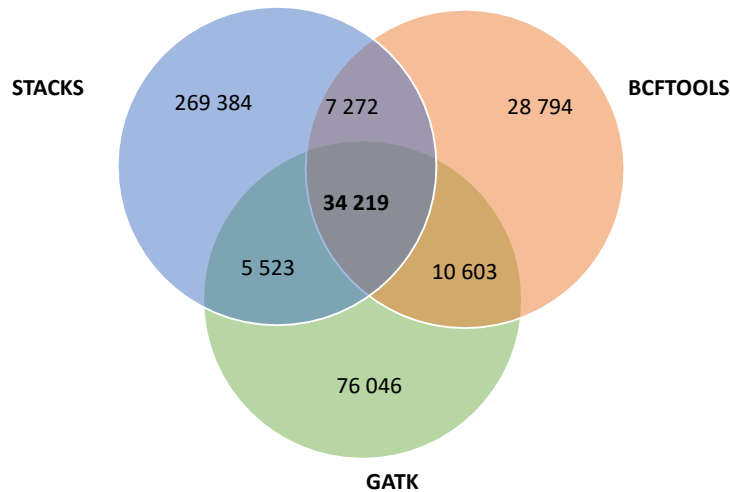


Figure B.2. Venn diagram of called and shared genotype loci from Stacks, Bcftools and GATK pipelines, before filtering of the successive common SNPs to all three pipelines together.

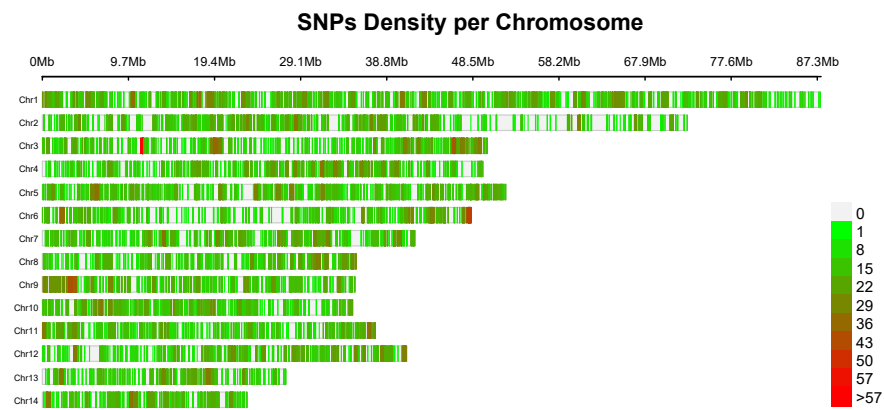


Figure B.3. distribution of the 17,948 SNPs across the 14 chromosomes of *Daphna* reference genome (Hasing & al., 2020).