Crop-to-wild geneflow: environmental risk assessment for the release of genetically modified sorghum in Kenya

Mutegi Evans

Thesis submitted in fulfilment of the requirements for the degree Philosophiae Doctor in Plant Breeding in the Faculty of Natural and Agricultural Sciences, Department of Plant Sciences, University of the Free State, Bloemfontein

May 2009

Promoter:
Prof. Maryke Labuschagne

Co-promoters:
Prof. Liezel Herselman
Dr. Fabrice Sagnard

Table of contents

University declaration

I declare that this thesis hereby submitted by me for the Philosophiae Doctor degree at the University of the Free State is my own independent work and has not previously been submitted by me at another university/faculty. I further cede copyright of the thesis in favour of the University of the Free State.

Signed on ……. at the University of the Free State, Bloemfontein, South Africa.

Signature:

Name: Mutegi Evans

Acknowledgements

King Solomon once wrote; "……*Of making many books there is no end, and much study wearies the body"* (Ecclesiastes 12:12). I am therefore, first and foremost grateful to God for constantly renewing my physical, mental and spiritual strength throughout this study. To Him be all glory and honour.

This thesis formed part of a project titled "Environmental risk assessment of genetically engineered sorghum in Mali and Kenya", initiated by the International Crops Research Institute for Semi-Arid Tropics (ICRISAT) with funds from the United States Agency for International Development (USAID) through the Biotechnology and Biosafety Interface (BBI) and Plant Biosafety Systems (PBS) programme. I am grateful to the management of ICRISAT for granting me a PhD graduate fellowship under this project and to my employer, Kenya Agricultural Research Institute (KARI) for granting me three years paid study leave to undertake the programme. Very special mention goes to the project principle investigator and thesis co-promoter, the late Dr. Fabrice Sagnard (ICRISAT), who until his untimely death on 18th November 2008 spared no effort, time or passion to offer much needed scientific guidance and moral support.

I extend my sincere gratitude to Profs. Maryke T. Labschagne (promoter) and Liezel Herselman (co-promoter) for their enthusiastic support, scientific guidance and well appreciated inputs in reviewing the drafts of this thesis. I would like to thank Dr. Santie de Villiers (ICRISAT) for her passionate encouragement and logistical support at every stage of this work, but above all for generously agreeing to review the draft thesis chapters. I am very grateful to Dr. Monique Deu, Centre de Coopération Internationale du Recherche Agronomique pour le Développement (CIRAD, France) for her timely assistance with data cleaning and analysis, and for her enlightening comments and suggestions on the draft thesis chapters. I wish to sincerely thank Dr. Kassa Semagn, formerly of ICRISAT, whose expertise on molecular marker data acquisition and analysis has contributed immensely towards completion of this study. I also extend my sincere gratitude to Dr. Dan Kiambi (ICRISAT) for his steadfast encouragement and scientific support at every stage of this study.

My sincere gratitude goes to several other persons and institutions for contributing in various different ways towards the accomplishment of this study:

- Farmers from Turkana, coastal, western/Nyanza and eastern regions of the country, for providing the bulk of the sorghum samples and accompanying information for this study.
- Mr. Ben M. Kanyenji, project coordinator KARI, whose immense knowledge of Kenyan sorghum, as well as logistical support was instrumental in sample acquisition trips in the various parts of the country.
- The Officer In-charge, National Genebank of Kenya-KARI, Mr. Zachary K. Muthamia for providing extra sorghum samples for this study.
- Caroline N. Mwongera (ICRISAT) for her able assistance with data and sample collection within the intensive study site in the lower parts of Meru South District.
- Michael Kimani, formerly of ICRISAT for patiently and painstakingly introducing me to practical molecular techniques: DNA extraction, PCR amplification and genotyping. His assistance in the laboratory towards acquisition of the molecular data used in this study is deeply acknowledged.
- Mr. Joseph Kamau (KARI, Muguga), Mr. Charles Marangu (KARI, Embu), Mr. Benard Rono (KARI, Embu) and Moses Muraya (PhD student, University of Hohenheim) for their useful inputs during the sample collection trips in the various parts of the country.
- The Biosciences Eastern and Central Africa (BeCA) hub located within the International Livestock Research Institute (ILRI), Nairobi, for granting me bench space and access to the state of the art genotyping facilities.
- Mrs. Sadie Geldenhuys (University of the Free State, South Africa) for her logistical and administrative support during my studies and especially during my periods of stay at the university.

My sincerely heartfelt gratitude goes to my dear wife, Rosemary Njeri and my lovely daughter, Faith Kendi for bearing with my many days and hours absent from home as I pursued my dream. Their constant encouragement and prayers are highly appreciated. Lastly, I am grateful to my mother Ciambuba M'Njaruba, my brothers and sister for their prayers and supportive encouragement.

Dedication

In memory of Dr. Fabrice Sagnard who passed away on 18th November 2008 after a brave battle with cancer. Dr. Sagnard was one of my co-promoters and the principle investigator of the USAID-BBI/PBS funded project on “Environmental risk assessment of genetically engineered sorghum in Kenya and Mali” upon which this thesis was based. Dr. Sagnard was an excellent population geneticist and paid particular attention to farmer practices and *in situ* conservation of plant genetic resources. His passion, warmth and deep scientific insights will be missed by many.

Quote

All science is concerned with the relationship of cause and effect. Each scientific discovery increases man's ability to predict the consequences of his actions and thus his ability to control future events.

Laurence J. Peter

List of abbreviations and acronyms

â	Rousset's genetic distance between individuals
A^p	Number of private alleles
A^r	Number of rare alleles
A^t	Total number of alleles
ABI	Applied Biosystems
ACZs	Agro-climatic zones
AFLP	Amplified fragment length polymorphism
AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
BC	Before Christ
BBI	Biotechnology and Biosafety Interface
BeCA	The Biosciences Eastern and Central Africa
blog	Regression slope
bp	Base pair(s)
Bt	*Bacillus thuringiensis*
CBSU	Computational Biology Service Unit
cDNA	Coding DNA
CIRAD	Centre de Coopération Internationale du Recherche Agronomique pour le Développement (French Agricultural Research Centre for International Development)
cm	Centimetre(s)
cpDNA	Chloroplast DNA
CTAB	Cetyl Trimethyl Ammonium Bromide
d_{ij}	Dissimilarity among genotypes i and j
DNA	Deoxyribonucleic acid
dNTP	2'-deoxynucleoside 5'-triphosphate
EDTA	Ethylenediaminetetraacetic acid
F	Fallow field
F_1	First filial generation
F_{ij}	Kinship coefficients
F_{IS}	Fixation index of individuals relative to the sub-population

F_{IT}	Fixation index of individuals relative to the total population
F_{ST}	Fixation index of sub-population relative to the total population/total fixation index
FAM	5-Carboxyfluorescine
FAO	Food and Agriculture Organisation of the United Nations.
FAOSTAT	FAO statistical database
G_{ST}	Nei's total fixation index
GM	Genetically modified
GPS	Global positioning system
H	Test statistic for the Kruskal-Wallis test
H_e	Expected heterozygosity/gene diversity
H_o	Observed heterozygosity
ha	Hectare(s)
HCl	Hydrochloric acid
HWE	Hardy-Weinberg equilibrium
ICRISAT	International Crops Research Institute for Semi-Arid Tropics
ILRI	International Livestock Research Institute
ISS	Intensive study site
K	Number of unknown populations/genetic clusters
KARI	Kenya Agricultural Research Institute
KCl	Potassium chloride
km	Kilometre(s)
LE	Linkage equilibrium
LIZ	ABI internal size standard for sequences up to 500 bp
LOD	Log odds
m	Metre(s)
m	Migration (geneflow) rate
M	Molar
masl	Metres above sea level
MCMC	Monte Carlo Markov Chain
mg	Milligram(s)
$MgSO_4$	Magnesium sulphate
min	Minute(s)
ml	Millilitre(s)

mm	Millimetre(s)
mM	Millimolar(s)
Mt.	Mount
mtDNA	Mitochondrial DNA
NaCl	Sodium chloride
Ne	Effective population size
NED	2,7,8-benzo-5-fluoro-2,4,7-trichloro-5-carboxyfluorescein
ng	Nanogram(s)
NJ	Neighbour-joining
Nm	Number of effective immigrants per generation
PBS	Plant Biosafety Systems
PCoA	Principle coordinate analysis
PCR	Polymerase chain reaction
PET	An ABI fluorescent dye
pH	Measure of acidity/basicity
PIC	Polymorphic information content
PNAS	Proceeding of the National Academy of Sciences of the USA
P(X|K)	Probability of X given K
Q_i	Mean proportion of estimated genome originating from a particular cluster
q_i	Proportion of an individual's genome in a particular genetic cluster
r^2	Coefficient of determination
r_{ij}	Relative kinship coefficient
R_s	Allelic richness
R_{ST}	Slatkin's total fixation index
RAPD	Random amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
RNAse	Ribonuclease
rpm	Revolutions per minute
s	Second(s)
SF	Sorghum field
SN-G	Semi natural, grassland habitat
SN-R	Semi natural, riverine habitat
SPAGeDi	Spatial pattern analysis of genetic diversity

ssp	Subspecies
SSR	Simple sequence repeat
Taq	*Thermus aquaticus*
TE	Tris/EDTA buffer
Tris-HCl	Tris (hydroxymethyl) aminomethane hydrochloride
USA	United States of America
USAID	United States Agency for International Development
UV	Ultraviolet
V	Volt(s)
v/v	Volume/volume
VIC	2'-chloro-7'-phenyl-1,4-dichloro-6-carboxyfluorescein
w/v	Weight/volume
β	Beta
θ	Weir and Cocherham's total fixation index
µl	Microlitre(s)
µM	Micromolar(s)
$\prod_{taxon}^{S}$	Private allelic richness
%	Percentage(s)
ΔK	Delta K
°C	Degrees centigrade

List of tables

List of figures

List of articles published and presentations delivered from thesis project

E. Mutegi, F. Sagnard, M. Labuschagne and L. Herselman 2009. Assessing the genetic structure and crop-wild geneflow in the *Sorghum bicolor* gene pools of Kenya using microsatellites. A presentation made at the ILRI Graduate Fellow Forum 2009. ILRI, Nairobi, 5-6 March 2009.

E. Mutegi, F. Sagnard, M. Labuschagne and L. Herselman 2009. Crop-to-wild geneflow: Environmental risk assessment and implications for the release of GM sorghum in Kenya. Poster presented at the BecA - ILRI Hub and Syngenta Foundation for Sustainable Agriculture Partnership Conference: ***From technology to product development for the African farmer.*** BecA - ILRI Hub, Nairobi, April 29, 2009.

Chapter 1

General introduction

Recent advances in biotechnology have culminated in the genetic engineering of many crops of economic importance. Undoubtedly this new technology has profound potential to improve the ever increasing demand for food globally, more so in the developing countries. Nonetheless, critical concerns have also been raised about the potential risks posed by genetically modified (GM) crops on the environment. Foremost among these concerns is the potential escape of transgenes from cultivated crops to their wild and weedy relatives through geneflow. The possible harmful consequences of such escape are the evolution of more aggressive weeds in agricultural systems, the generation of more invasive species in natural habitats, the gradual replacement of wild gene pools by cultivated ones and in some extreme cases, the extinction of crop wild relative populations (Conner *et al*. 2003; Ellstrand 2003; Haygood *et al*. 2003; Chen *et al*. 2004; Johnston *et al.* 2004). Scientific assessment of these potential environmental risks is an integral part of biosafety regulations and therefore precedes any decision to release a GM crop.

Sorghum (*Sorghum bicolor* (L.) Moench) is Africa's second most important cereal in terms of both area harvested and annual production. According to the latest global statistics (FAO 2008), Africa contributed over 60% to the total land area dedicated to cultivation of sorghum. There is no doubt therefore that sorghum occupies an important position as a dietary staple for millions of people, especially in arid and semi-arid lands of Africa and Asia. The important socio-economic position enjoyed by sorghum makes it a necessity for production enhancement programmes in developing countries of Africa. Advances in genetic engineering offer potentially promising tools for augmenting traditional approaches to sorghum crop improvement. Plausible progress has been achieved towards developing and optimising protocols for transferring genes into sorghum using electroporation, agrobacterium and microprojectile bombardment techniques (Casas *et al*. 1997; Zhao *et al*. 2000; Gao *et al*. 2005; Howe *et al*. 2006). Sorghum has therefore been successfully engineered against stem fungal rot, stalk borer

and for high lysine content expression (Casas *et al.* 1993; 1997; Zhu *et al.* 1998; Krishnaveni *et al.* 2000; Zhao *et al.* 2000; Girijashankar *et al.* 2005; Ayoo 2008). In light of the now increasing deployment of GM crops in developing countries (James 2007), it seems that it is only a matter of time before transgenic sorghum is deployed into Africa's predominantly traditional agro-ecosystems. A probable case in point is the on-going initiative aimed at deploying nutritionally enhanced transgenic sorghum to subsistence farmers on the continent, under the auspices of the African Biofortified Sorghum project (Zhao 2008). It is essential that such deployment be preceded by studies that characterise potential environmental risks especially with ecological and agronomical characteristics of Africa's traditional agro-ecosystems in mind.

Sorghum was domesticated in Africa and is a critical component of food security for more than 100 million people on the continent today. Several wild relatives of cultivated sorghum are found in Africa, both in natural habitats and as weeds in farmers' fields. Spontaneous, morphologically-intermediate plants between cultivated sorghum and its wild relatives have been reported in and near sorghum fields in Africa (Dogget and Majisu 1968; Baker 1972; De Wet 1978; Dogget and Prasada Rao 1995; Tesso *et al.* 2008; Mutegi et al. 2009). Moreover, crop-to-wild hybridisation in sorghum has been implicated in the origin of at least one noxious weed, *Sorghum x almum* Parodi and in enhanced weediness and invasiveness of another, johnsongrass (*Sorghum halepense* (L.) Pers) (Ellstrand *et al.* 1999). Prediction of the extent and direction of introgression between sorghum and its wild and weedy relatives is thus an important part of environmental risk assessment of transgenic sorghum. Such studies have not yet been reported in Africa, the centre of origin and diversity for sorghum.

Kenya borders Sudan and Ethiopia in the north and is therefore located on the southern outskirt of the north-eastern quadrant of Africa, where sorghum is believed to have been first domesticated (De Wet 1978; Dogget 1988). Kenya's sorghum gene pool is therefore represented by both cultivated and crop wild relatives (Clayton and Renvoize 1982). Sorghum is grown in all but one of the country's eight provinces. The crop is grown for subsistence mainly by small scale farmers under traditional farming systems. In Kenya sorghum is a particularly important part of the dietary needs of many people in the mid-altitude western region, semi-arid lowland lands of the eastern region and the expansive arid zones of northern Rift Valley. Although there have been attempts to breed and

introduce improved varieties in most of these growing areas, large sets of local landraces still dominate cultivated sorghum diversity. Although the work of Dogget and Majisu (1968) documented some morphological evidence of hybridisation between cultivated and wild sorghum in Kenya and in the neighbouring countries of Uganda and Tanzania, detailed evidence based on the more reliable molecular analysis is still outstanding. Moreover, neither the variability of genetic introgression in different agro-ecological zones of the country, nor the direction of geneflow between particular cultivated varieties and ecotypes of wild sorghums has been investigated. Such information is urgently needed in order to assist policy makers' deliberations on the potential impact of growing transgenic sorghum in Kenya. Furthermore, information on the extent and genetic structure of both the cultivated and wild sorghum gene pools in Kenya is lacking, but is important both for effective conservation and crop improvement programmes.

Approaches based on population genetics theories and molecular markers have proved to be effective in investigating genetic structure and geneflow in different wild-weedy-domesticate complexes. Examples include rice (*Oryza sativa* L.) (Kuroda *et al*. 2006), maize (*Zea mays* L.) (Fukunaga *et al*. 2005), buckwheat (*Triticum aestivum* L.) (Konishi and Ohnishi 2007), lima bean (*Phaseolus lunatus* L.) (Martinez-Castillo *et al*. 2007), common beans (*Phaseolus vulgaris* L.) (Papa and Gepts 2003; Zizumbo-Villarreal *et al*. 2005), pejipaye palm (*Bactris gasipaes* Kunth) (Couvreur *et al*. 2006; Ugalde *et al*. 2008), sugar beet (*Beta vulgaris* L.) (Desplanque *et al.* 1999; Arnaud *et al*. 2003) and squash (*Cucurbita argyrosperma* Huber ssp. *argyrosperma* and *C. moschata* Duchesne*)* (Montes-Hernandez and Eguiarte 2002).

The present study employed microsatellite or simple sequence repeat (SSR) markers to generate allelic data for use in genetic distance-based and model-based population genetics approaches with a view to understand the genetic structure and relationships between cultivated and wild sorghum in Kenya. Furthermore, the extent and direction of geneflow between the two taxa at country, regional and landscape scales was quantified. Outputs from the work will contribute to development of biosafety regulations and guidelines for the introduction of transgenic sorghum in the country by answering the following questions:

(i) What is the extent of genetic diversity in cultivated and wild sorghum gene pools in Kenya?

(ii) What is the structure of the genetic diversity at national, regional and landscape scales?

(iii) What evolutionary factors shape the observed structure?

(iv) What are the genetic and evolutionary relationships between cultivated and wild sorghum gene pools?

(v) Is there geneflow between cultivated and wild sorghum? If so, what is the prevalent direction?

The goal of the study was therefore to contribute to the understanding of the geneflow related environmental risks of releasing genetically modified sorghum into Kenya's agro-ecosystem and contribute to biosafety, conservation and utilisation decisions regarding sorghum in the country. Specific goals are to conduct a comparative phylogeography survey of wild and cultivated sorghums in Kenya and to characterise the amount of introgression at local scale.

References

Arnaud F, Viard F, Delescluse M, Cuguen J. 2003. Evidence for geneflow via seed dispersal from crop to wild relatives in *Beta vulgaris* (Chenopodiaceae): consequences for the release of genetically modified crop species with weedy lineages. Proceedings of the Royal Society B: Biological Sciences 270:1565-1571.

Ayoo LMK. 2008. Genetic transformation of Kenyan sorghum (*Sorghum bicolor* L. Moench) with anti-fungal genes and response to *Collectotrichum sublineolum* infection. PhD Dissertation, University of Hamburg, Germany. pp.116.

Baker HG. 1972. Human influences on plant evolution. Economic Botany 26:32-43.

Casas AM, Kononowicz AK, Haan TG, Hang L, Tomes DL, Bressan RA, Hasegawa PM. 1997. Transgenic sorghum plants obtained after microprojectile bombardment of immature inflorescences. Invitro Cellular and Developmenta Biology-Plant 33:92-100.

Casas AM, Kononowicz AK, Zehr BU, Tomes TD, Axtell DJ, Butler GL, Bressan AR, Hasegawa PM. 1993. Transgenic sorghum plants via microprojectile bomabardment. Proceeding of the National Academy of Sciences of the USA (PNAS) 90:11212-11216.

Chen LJ, Lee DS, Song PZ, Suh HS, Lu B. 2004. Geneflow from cultivated rice (*Oryza sativa*) to its weedy and wild relatives. Annals of Botany 93:67-73.

Clayton WD, Renvoize RD. 1982. Poaceae. Flora of Tropical East Africa, Part 3. AA Balkema Rotterdam, Netherlands, pp. 731.

Conner AJ, Glare TR, Nap JP. 2003. The release of genetically modified crops into the environment. Part II. Overview of ecological risk assessment. The Plant Journal 33:19-46.

Couvreur TLP, Billotte N, Risterucci AM, Lara C, Vigouroux Y, Ludena B, Pham JL, Pintaud JC. 2006. Close genetic proximity between cultivated and wild *Bactris gasipaes* Kunth. revealed by microsatellite markers in Western Ecuador. Genetic Resources and Crop Evolution 53:1361-1373.

De Wet JMJ. 1978. Systematics and evolution of sorghum sect. Sorghum (Gramineae). American Journal of Botany 65:477-484.

Desplanque B, Boudry P, Broomberg K, Saumitou-Laprade P, Cuguen J, Van Dijk H. 1999. Genetic diversity and geneflow between wild, cultivated and weedy forms of *Beta vulgaris* L. (Chenopodiaceae), assessed by RFLP and microsatellite markers. Theoretical and Applied Genetics 98:1194-1201.

Dogget H. 1988. Sorghum. Longman Scientific and Technical Essex, England, pp. 512.

Dogget H, Majisu BN. 1968. Disruptive selection in crop development. Heredity 23:1-23.

Dogget H, Prasada Rao KE. 1995. Sorghum. In: Smartt J, Simmonds NW (Eds.), Evolution of Crop Plants. Longman Group, Burnt Mill. pp. 180.

Ellstrand NC. 2003. Dangerous liaisons - when cultivated plants mate with their wild relatives. Johns Hopkins University Press Baltimore MD, pp. 244.

Ellstrand NC, Prentice HC, Hancock JF. 1999. Geneflow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics 30:539-563.

FAO. 2008. FAOSTAT. http://faostat.fao.org.

Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, Sanchez G, Liu K, Buckler ES, Doebley J. 2005. Genetic diversity and population structure of teosinte. Genetics 169:2241-2254.

Gao Z, Xie X, Ling Y, Muthukrishnan S, Liang GH. 2005. *Agrobacterium tumefaciens*-mediated sorghum transformation using a mannose selection system. Plant Biotechnology Journal 3:591-599.

Girijashankar V, Sharma KK, Swathisree V, Prasad LS, Bhat BV, Royer M, Narasu ML, Altosaar I, Seetharama N. 2005. Development of transgenic sorghum for insect resistance against the spotted stem borer (*Chilo partellus*). Plant Cell Reports 24:513-522.

Haygood R, Ives AR, Andow DA. 2003. Consequences of recurrent geneflow from crops to wild relatives. Proceedings of the Royal Society of London 270:1879-1886.

Howe A, Shirley S, Dweikat I, Fromm M, Clemente T. 2006. Rapid and reproducible *Agrobacterium*-mediated transformation of sorghum. Plant Cell Reports 25:751-758.

James C. 2007. Global status of commercialised Biotech/GM crops: 2007. ISAAA Brief No. 37 (http://www.isaaa.org).

Johnston J, Blancas L, Borem A. 2004. Geneflow and its consequences: a case study of Bt Maize in Kenya. In: Hilbeck A, Andow DA (Eds.), Environmental risk assessment of genetically modified organisms: Vol. 1 A case study of Bt Maize in Kenya. CAB International, Wallinford, UK. pp. 207.

Konishi T, Ohnishi O. 2007. Close genetic relationship between cultivated and natural populations of common buckwheat in the Sanjiang area is not due to recent geneflow between them: An analysis using microsatellite markers. Genes and Genetic Systems 82:53-64.

Krishnaveni S, Jeoung J, Muthukrishnan S, Liang G. 2000. Transgenic sorghum plants constitutively expressing a rice chitinase gene show improved resistance to stalk rot. Journal of Genetic Breeding 55:151-158.

Kuroda Y, Kaga A, Tomooka N, Vaughan DA. 2006. Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Molecular Ecology 15:959-974.

Martinez-Castillo J, Zizumbo-Villarreal D, Gepts P, Colunga-GarciaMarin P. 2007. Geneflow and genetic structure in the wild-weedy-domesticated complex of *Phaseolus lunatus* L. in its mesoamerican center of domestication and diversity. Crop Science 47:58-66.

Montes-Hernandez S, Eguiarte LE. 2002. Genetic structure and indirect estimates of geneflow in three taxa of *Cucurbita* (Cucurbitaceae) in western Mexico. American Journal of Botany 89:1156-1163.

Mutegi E, Sagnard F, Muraya M, Kanyenji B, Rono B, Mwongera C, Marangu C, Kamau J, Parzies H, de Villiers S, Semagn K, Traore PS, Labuschagne M. 2009. Ecogeographical distribution of wild, weedy and cultivated *Sorghum bicolor* (L.) Moench in Kenya: implications for conservation and crop-to-wild gene flow. Genetic Resources and Crop Evolution (DOI 10.1007/s10722-009-9466-7)

Papa R, Gepts P. 2003. Asymmetry of geneflow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. Theoretical and Applied Genetics 106:239-250.

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J, Marx D, Bothma G, Ejeta G. 2008. The potential for crop-to-wild geneflow in sorghum in Ethiopia and Niger: A geographic survey. Crop Science 48:1425-1431.

Ugalde JAH, Urpi JM, Nunez OR. 2008. Genetic diversity and kin relationships among wild and cultivated populations of the pejibaye palm (*Bactris gasipaes*, Palmae) using microsatellite markers. Revista de Biologia Tropical 56:217-245.

Zhao Z. 2008. The Africa biofortified sorghum project - Applying biotechnology to develop nutritionally improved sorghum for Africa. Proceedings of the 11th IAPTC&B Congress, August 31-18, 2006 Beijing, China. p. 273-277.

Zhao Z, Cai T, Tagliani L, Wang N, Pang H, Rudert M, Schroeder S, Hondred D, Pierce D. 2000. *Agrobacterium*-mediated sorghum transformation. Plant Molecular Biology 44:789-798.

Zhu H, Muthukrishnan S, Krishnaveni S, Wilde G, Jeoung J, Liang G. 1998. Biolistic transformation of sorghum using a rice chitinase gene. Journal of Genetic Breeding 52:243-252.

Zizumbo-Villarreal D, Colunga-GarciaMarin P, de la Cruz EP, Gado-Valerio P, Gepts P. 2005. Population structure and evolutionary dynamics of wild-weedy-domesticated complexes of common bean in a Mesoamerican region. Crop Science 45:1073-1083.

Chapter 2

Literature review

2.1 Introduction

Sorghum [*S. bicolor* (L.) Moench] is the world's fifth most produced cereal crop after maize (*Z. mays* L.), rice (*O. sativa* L., *O. glaberrima* Steud.), wheat (*Triticum* spp.) and barley (*Hordeum vulgare* L.). In 2007, the world planted 43.8 million ha of sorghum, with over 80% of the area devoted to the crop being found in Africa and Asia (Figure 2.1) (FAO 2008). Sorghum forms an important dietary component of many people globally, with the most significant contribution being in the arid and semi-arid lands in many African and Asian countries. In Kenya, sorghum is ranked third in importance among other cereals. In 2007 alone over 100000 ha of land were devoted to sorghum production in Kenya (Figure 2.2), with majority of the producers being small scale farmers.

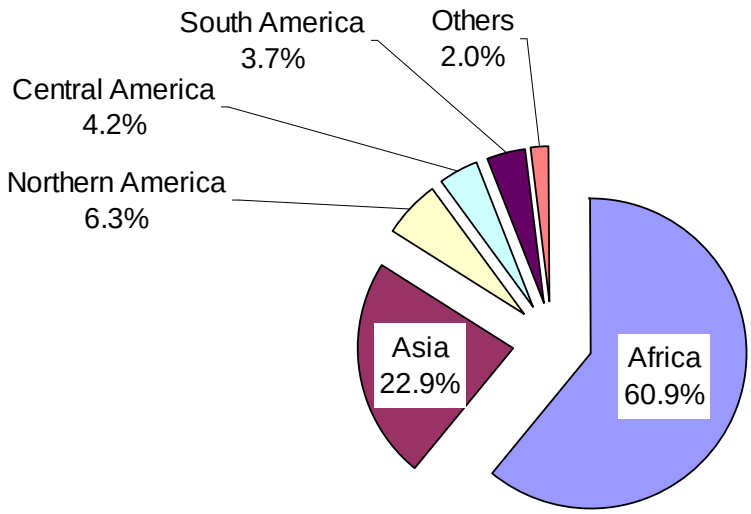

Figure 2.1 Proportion of area devoted globally to sorghum production in the year 2007 (Source: FAO 2008).

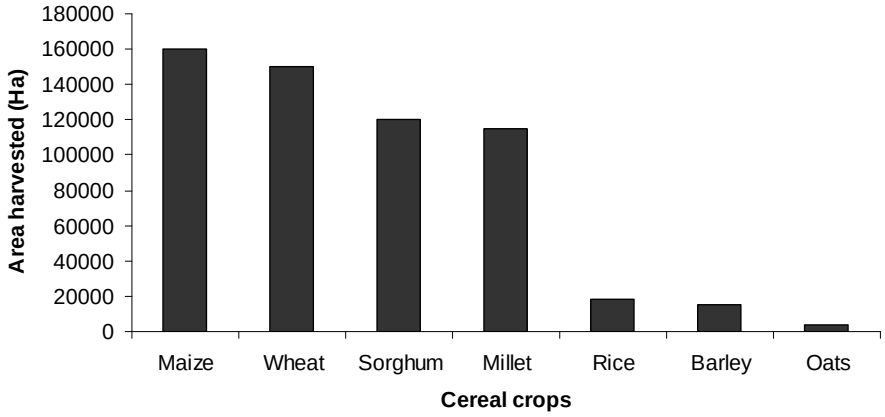

Figure 2.2 Area devoted to sorghum and other cereal crops in Kenya in 2007 (Source: FAO 2008).

2.2 Taxonomy of cultivated and wild sorghum

Sorghum Moench is a large and heterogeneous genus belonging to the Andropogoneae tribe in the botanical family Poaceae. The genus is divided into five sub-generic sections: Eusorghum, Parasorghum, Heterosorghum, Chaetosorghum and Spitosorghum (Figure 2.3). The primary and secondary gene pools of sorghum, which include the cultivars and their wild and weedy relatives (Harlan and De Wet 1971), are classified within the section Eusorghum. Three species are recognised in this section: (i) *S. halepense*, a member of the secondary gene pool, is a rhizomatous perennial weedy taxa and a native of Eurasia, but now introduced in warm temperate regions of the world, (ii) *S. propinquum* (Kunth) Hitchc, a member of the primary gene pool, is a rhizomatous perennial weedy species with distribution mainly in south-east Asia and (iii) *S. bicolor*, the most important member of the primary gene pool, is indigenous to Africa and comprises all cultivars of sorghum, their wild progenitors, as well as morphologically stabilised weedy forms that are thought to be derivatives of crop-wild introgression (De Wet 1978).

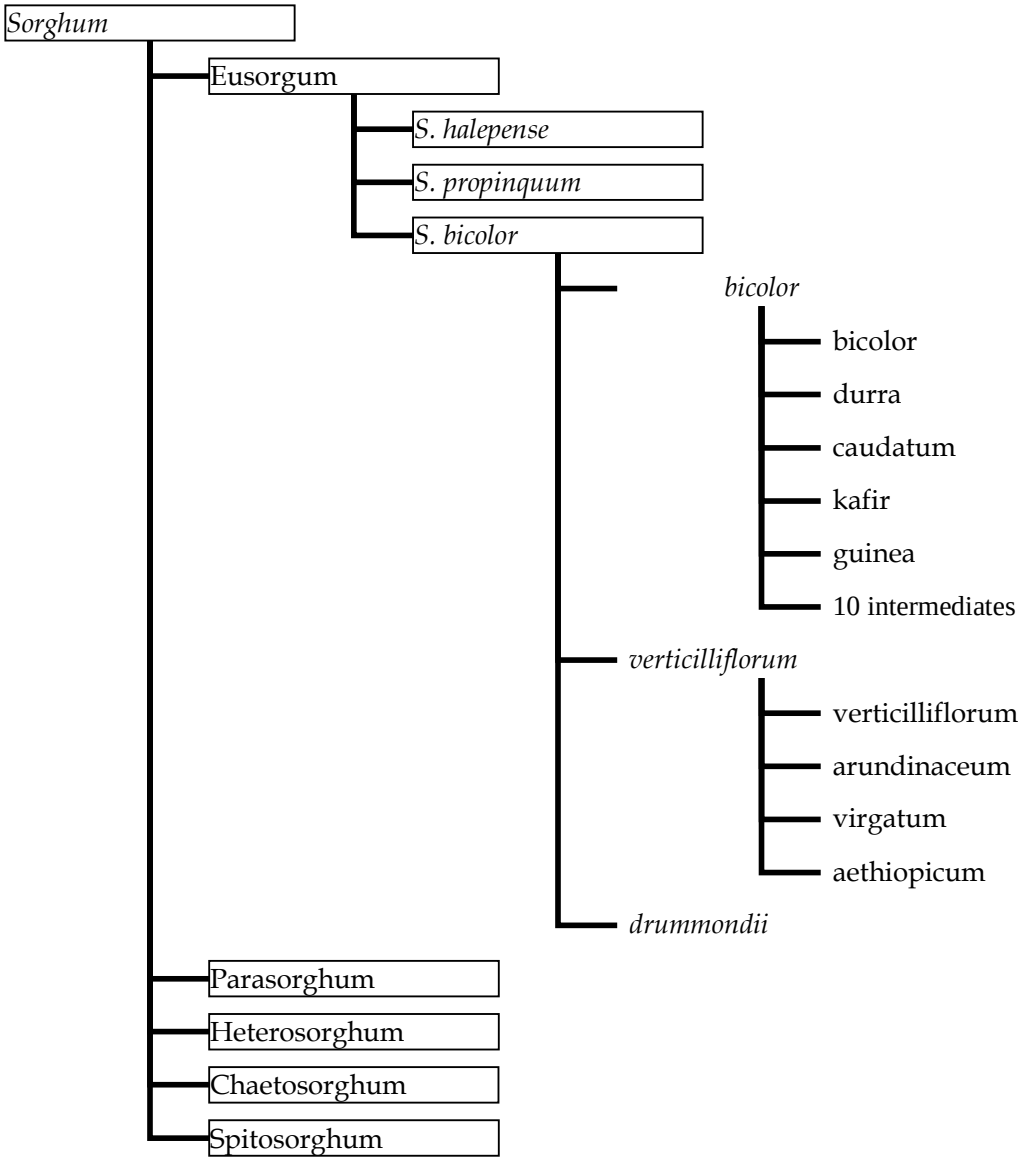

Figure 2.3 Schematic presentation of the taxonomy of the genus *Sorghum*.

Cultivated sorghum and its proposed wild progenitors have been classified under a single species, *S. bicolor*, within which three sub-specific categories are recognised: ssp. *bicolor*, ssp. *verticilliflorum* (Steud.) and ssp. *drummondii* (Steud.) (Harlan and De Wet 1972; Dogget 1988) (Figure 2.3). All cultivars of sorghum are encompassed within ssp. *bicolor*, in which five basic and ten intermediate races are further recognised on the basis of spikelet and panicle morphology: (i) race bicolor is characterised by open panicles and long clasping glumes that usually enclose the elliptic grain at maturity, (ii) race kafir is characterised by more or less compact panicles with elliptic sessile spikelets and glumes

that tightly clasp the usually longer grain at maturity, (iii) race caudatum is characterised by open to compact panicles, with grains that are flat on one side and distinctively curved on the opposite and shorter glumes that leave the grains exposed, (iv) race durra has characteristic compact panicles, flattened and ovate sessile spikelets and lower glume that is wrinkled near the middle and (v) race guinea is characterised by large, open panicles with pendulous branches and glumes that are long, widely open and with a conspicuous awn. Pairwise hybridisation of the five basic races has further given rise to ten intermediate races of cultivated sorghum (Harlan and De Wet 1972; Dogget 1988).

The closest wild relatives of cultivated sorghum are found in Africa. These are all encompassed within the ssp. *verticilliflorum*, formally ssp. *arundinaceum*. Based mainly on variations in plant habit, leaves, inflorescence and ecogeographical distribution, members of the ssp. *verticilliflorum* have been classified into four botanical races: arundinaceum, verticilliflorum, aethiopicum and virgatum. Race arundinaceum comprises of robust and tall forest grasses that occur mainly in humid and sub-humid West Africa. It has large leaves and a broad loose panicle with pendulous branches. Race verticilliflorum is the most widespread of all the wild sorghum in sub-Saharan Africa. It is abundant in dry savannas and differs from arundinaceum mainly by having panicles with non-pendulous branches. Race aethiopicum is found in the southern margin of the Sahara desert. Members of this race are shorter than those of arundinaceum and verticilliflorum and are further characterised by smaller panicles that have erect to sub-erect branches. Race virgatum is a slender desert grass occurring from central Sudan to Egypt, mostly along stream banks and irrigation ditches. The four races are so closely related morphologically that they do not deserve formal taxonomic status and are considered to be essentially well defined ecotypes (De Wet *et al.* 1970; De Wet and Huckabay 1971; De Wet 1978; Dogget 1988).

Moreover, all races of cultivated *S. bicolor* ssp. *bicolor* and those of its closest wild relative *S. bicolor* ssp. *verticilliflorum* are inter-fertile and hybridise to produce highly variable weedy types within and around sorghum fields (Harlan and De Wet 1972). Morphologically stabilised derivatives of this introgression between the two congeners have therefore been classified under *S. bicolor* ssp. *drummondii* (Harlan and De Wet 1972; Dogget 1988).

2.3 Sorghum domestication

While it is unequivocally agreed that sorghum plants are African in origin, there have been divergent views on where and when its domestication occurred. For example, Murdock (1959) postulated that sorghum was domesticated in West Africa around river Niger by the Mande peoples some 4500 BC and that the crop was introduced from there to the Sudan by 4000 BC. Based on comparative morphological studies and numerical taxonomy, De Wet and Huckabay (1967) suggested that sorghum was domesticated independently from local wild relatives of the crop in three regions: Ethiopian region, tropical West Africa and South East Africa. Harlan (1975) used the distribution of races of sorghum in Africa to conclude that the initial domestication of sorghum occurred in a long belt across central Africa, perhaps running through contemporary Ethiopia, Sudan and Chad. Mann *et al.* (1983) hypothesised that the origin and domestication of sorghum took place in north-eastern Africa, perhaps in the expanse of the land today recognised as Ethiopia and Sudan, approximately some 5000 years ago. Based on his hypothesis on the development and spread of agriculture in Africa, as well as on the distribution of sorghum races, Dogget (1988) concluded that sorghum was first cultivated by the Cushites in the Ethiopian highlands of East Africa.

2.4 Genetic relationships within and among cultivated and wild sorghum

Genetic compatibility between domesticated and wild populations in regions of sympatric occurrence often leads to wild-weedy-domesticate hybrid complexes (Ellstrand *et al.* 1999; Ellstrand 2003b) as a result of introgressive hybridisation from domesticated populations to wild ones and *vice versa*. Crop wild relatives have long been recognised as valuable sources of new variation and potentially novel genes and therefore constitute important genetic resources for plant breeding and conservation programmes. In the advent of GM crops, however, wild-weedy-domesticate hybrid complexes have become a source of growing biosafety concern due to their potential to facilitate transgene escape. Knowledge on the population structure can provide insight into many important evolutionary and ecological properties of a species. In particular it can be used to estimate geneflow, a critical component for assessing the potential environmental risks posed by GM crops (Snow and Moran-Palma 1997; Conner *et al*. 2003; Haygood *et al*. 2003; Cleveland and Soleri 2005; Chapman and Burke 2006; Thies and Devare 2007; Chandler and Dunwell 2008). Furthermore, understanding the

extent and partitioning of the genetic diversity of a crop and its wild relatives is critical to effective conservation and use of its genetic resources.

Several authors have documented the potential that exists in the wild sorghum gene pool with regard to providing new sources of resistance and adaptation in breeding (Sharma and Franzmann 2001; Gurney *et al.* 2002; Kamala *et al.* 2002; Komolong *et al.* 2002; Rao Kameswara *et al.* 2003; Jordan *et al.* 2004; Dillon *et al.* 2007; Hajjar and Hodgkin 2007). That wild and cultivated sorghums are inter-fertile and grow in sympatry in many agro-ecosystems of sub-Saharan Africa has long been documented (Dogget and Majisu 1968; Dogget 1988; Dogget and Prasada Rao 1995). For example, Dogget and Majisu (1968) studied the relationship between wild and cultivated sorghum in East African countries including Kenya and provided morphological evidence of wild-weedy-domesticate hybrid complexes. However, neither the variability of genetic introgression in different agro-ecological zones, nor the direction of geneflow between the diverse cultivated varieties and ecotypes of wild sorghums in Africa has been investigated. In many parts of Africa a wide assortment of intermixed cultivars of sorghum are grown in scattered cultivated plots often in close proximity to its wild and weedy forms in both cultivated fields and intermediate habitats such as field margins and fallow plots (Dogget 1988; Barnaud *et al.* 2007; Tesso *et al.* 2008).

Molecular evidence of genetic introgression between wild and cultivated sorghum has been documented for the *S. bicolor-S. halepense* (johnsongrass) complex in the United States (Arriola and Ellstrand 1996; Morrell *et al.* 2005) and within the *S. bicolor* species (Aldrich and Doebley 1992; Aldrich *et al.* 1992; Casa *et al.* 2005). However, these results can only be considered as indicative and far from conclusive with regard to development of comprehensive biosafety regulations for introduction of GM sorghum in Africa's agro-ecosystems. The work by Aldrich and Doebley (1992), for example, focused only on 56 accessions of cultivated and wild sorghum originating from nine African and two Asian countries. Kenya's sorghum gene pool in their study was represented by only six accessions of unspecified geographical origin, with only two being of the wild form (Aldrich and Doebley 1992; Aldrich *et al.* 1992). More extensive sampling is definitely necessary for more in-depth studies on the genetic relationships within and between cultivated and wild sorghum populations in Kenya, if practical biosafety, conservation and germplasm utilisation questions are to be answered.

There has been a substantial effort over the last two decades to characterise at diverse spatial scales, the levels and patterns of genetic diversity within cultivated sorghum by means of morphological, biochemical and molecular markers (Morden *et al.* 1989; Deu *et al.* 1994; Menkir *et al.* 1997; Dje *et al.* 1998; 1999; 2000; Ayana *et al.* 2000b; Kong *et al.* 2000; Ghebru *et al.* 2002; Menz *et al.* 2004; Abu-Assar *et al.* 2005; Folkertsma *et al.* 2005; Kamala *et al.* 2006; Barnaud *et al.* 2007; Perumal *et al.* 2007; Barnaud *et al.* 2008; Deu *et al.* 2008; Sagnard *et al.* 2008). In comparison, only a few studies have been directed towards expanding knowledge on genetic structure of wild sorghum and on the genetic and evolutionary relationships between wild and cultivated sorghum (Aldrich and Doebley 1992; Cui *et al.* 1995; Deu *et al.* 1995; Ayana *et al.* 2000a; Casa *et al.* 2005).

2.4.1 Extent and organisation of diversity in cultivated sorghum

Early genetic diversity studies on cultivated sorghum using isozymes, showed genetic diversity to be substantially spatially structured at global scale (Morden *et al.* 1989; Ollitrault *et al.* 1997) but not at national (Ayana *et al.* 2001) or local (Dje *et al.* 1998; 1999) scales. Interestingly, local scale studies using sorghum samples collected *in situ,* contradicted findings of the global scale studies using *ex situ* collections, by revealing more variation within than between accessions. In both cases however, the system of racial classification for sorghum as proposed by Harlan and De wet (1972) was not substantiated, perhaps due to low marker resolution. As noted by Aldrich and Doebley (1992), isozyme markers are restricted to a limited number of coding regions of the genome and can detect only those mutations that cause changes in protein mobility. With the advent of DNA marker technologies, studies employing nuclear restriction fragment length polymorphism (RFLP) analysis revealed concordance between genetic differentiation and racial classification in cultivated sorghum (Deu *et al.* 1994; 1995; 2006). A study by Casa *et al.* (2005) using 98 SSRs loci on 73 genebank accessions of cultivated sorghum, however, seemed to contradict these previous findings by revealing little evidence of genetic differentiation among racial groups. Reasons for this disagreement are not clear although diverse composition of the germplasm evaluated (i.e. selection of accessions to include racial and geographic diversity in sorghum), the occurrence of pollen or seed flow and/or recent divergence were suggested (Casa *et al.* 2005).

Recent diversity studies on cultivated sorghum have relied heavily on SSR markers and *in situ* collections in an attempt to elucidate the evolutionary process that shape patterns of genetic diversity at regional, national and local spatial scale (Barnaud *et al.* 2007; Deu *et al.* 2008; Sagnard *et al.* 2008). Such information is critical for both *in situ* and *ex situ* conservation, as well as for utilisation of plant genetic resources. Recently, Deu *et al.* (2008) used 28 SSR markers to conduct a genetic diversity survey on 484 sorghum samples collected in 79 villages across Niger. They detected high levels of genetic diversity that was differentiated along sorghum botanical races, geographical distribution and ethnic groupings of farmers, but poorly along climatic zones. The high levels of diversity was explained by convergence of three sorghum races (guinea, caudatum and durra), while spatial distribution patterns of different linguistic groups that were associated with cultivation of specific races explained the observed geographic and ethnic genetic structure (Deu *et al.* 2008).

In a local scale genetic diversity survey, Barnaud *et al.* (2007) used 14 SSR markers to characterise 21 landraces of sorghum collected at village level among the Duupa farmers in northern Cameroon. Despite observing that farmers grew a mixture of an average of 12 landraces per field, significant genetic differentiation between landraces was obtained, perhaps due to some form of barrier to inter-landrace geneflow and seed selection by farmers (Barnaud *et al.* 2007). Furthermore, inbreeding was observed to vary among landraces, a further suggestion that different mating systems among landraces might exist. Subsequent studies (Barnaud *et al.* 2008) have indeed revealed variable, though extensive outcrossing rates among landraces. The authors further postulated that the selection exerted by farmers was a key parameter for determining the fate of new genetic combinations from the outcrossing events and thus in the patterns of genetic differentiation among landraces (Barnaud *et al.* 2008).

In a recent study, Sagnard *et al.* (2008) used 12 SSR markers and 1518 samples collected *in situ* from Burkina Faso, Mali, Niger and in the same village in Cameroon sampled by Barnaud *et al.* (2007; 2008) to characterise the evolutionary forces that shaped genetic diversity of cultivated sorghum at multiple spatial scales. Their work suggested that the genetic variability in a variety is mainly the result of two factors: (i) its mating system and (ii) genetic drift arising from a limited number of reproductive individuals either at the time the variety is introduced into a household or each year

when farmers select their seeds from only a fraction of their harvest. Furthermore, they found no evidence of spatial genetic structure among villages separated by more than 30 km. This suggested that traditional seed exchange systems in West Africa operate at local scale. When they compared genetic diversity between countries, Niger was found to be genetically richer than Mali despite the fact that the latter grows sorghum in a larger agro-climatic range than the former. These results demonstrated that the diversity of human groups acted together with the agro-ecological factors to shape the structure of sorghum genetic diversity (Sagnard *et al.* 2008). Both should thus be taken into account in designing plant genetic resources conservation and crop improvement programmes.

2.4.2 Extent and organisation of diversity in wild sorghum

In the case of wild sorghum, Morden *et al.* (1990) used 90 genebank accessions originating from Africa, India and Thailand to conduct one of the first surveys on allozyme variation among wild congeners of cultivated sorghum in the sub-generic section Eusorghum. Their work failed to provide any clear taxonomic differentiation among species of the section as proposed by De Wet (1978), possibly due to a combination of low levels of marker polymorphism and insufficient sampling of *S. halepense* and *S. x almum.* (Morden *et al.* 1990). Their work further compared the allozymic variation of S. *bicolor* ssp. *verticilliflorum* with that of cultivated *S. bicolor* spp. *bicolor* from previous work (Morden *et al.* 1989) to reveal higher levels of diversity in the wild gene pool compared to cultivated sorghum. In a similar study, Aldrich and Doebley (1992) undertook nuclear and chloroplast DNA (cpDNA) RFLP analysis focusing on the geographical and racial diversity represented in cultivated sorghum (ssp. *bicolor*) and its proposed wild progenitor (ssp. *verticilliflorum*). Along with observing a clear genetic differentiation between cultivated and wild sorghum, they found higher levels of nuclear diversity within the latter compared to the former. Moreover, the nuclear diversity of cultivated sorghum was found to be well encompassed within the wild sorghum gene pool. They further observed that nuclear diversity of the wild sorghum gene pool from north-eastern Africa was comparatively closer to that of cultivated sorghum. Cui *et al.* (1995) made similar observations in their RFLP analysis study on cultivated and wild genebank accessions originating from Africa, Asia and the USA. Considered together, these results strongly favour the taxonomic classification of Harlan and De Wet (1972) and the hypothesis that sorghum

was domesticated from *S. bicolor* ssp. *verticilliflorum* in the north-eastern quadrant of Africa (Harlan *et al.* 1976).

Within wild sorghum, the work by Aldrich and Doebley (1992) further observed nuclear genetic differentiation along geographical distribution but not along racial classification. On the contrary, cpDNA analysis revealed neither spatial genetic differentiation nor genetic separation between cultivated and wild gene pools. Introgressive hybridisation between the cultivated and wild gene pools was thought to be behind the apparent genetic homogeneity between cultivated and wild sorghum (Aldrich and Doebley 1992). With the exception of one cultivated sub-race (*guinea margartiferum*), similar observations were made by Deu *et al.* (1995) in their study on comparative genetic diversity of cultivated and wild sorghum using mitochondrial DNA (mtDNA) markers. According to Deu *et al.* (1995), however, interspecific and interracial genetic homogeneity was as a result of a common mitochondrial background due to recent common ancestry. Their conclusions are in line with the hypothesis by Harlan *et al.* (1976) that domestication occurred from *S. bicolor* ssp. *verticilliflorum*, followed by diversification in cultivated sorghum in different geographic areas under different environmental and human selection pressures. In a comparative genetic study, Casa *et al.* (2005) used SSR markers to quantify and characterise diversity in a panel of gene bank accessions of cultivated and wild sorghum and established that landraces retained up to 86% of the diversity observed in wild sorghums. Genetic differentiation between cultivated and wild populations was found to be moderate while little evidence was available for racial differentiation in wild forms (Casa *et al.* 2005). All these studies were, however, characterised by poor racial and geographical representation, a fact that could have an effect on the observed level of racial and geographical differentiation.

Ayana *et al.* (2000a) used random amplified polymorphic DNA (RAPD) markers to investigate the extent and partitioning of genetic diversity in *S. bicolor* ssp. *verticilliflorum* collected *in situ* from five regions of Ethiopia, one of Africa's presumed homes for sorghum domestication (Harlan *et al.* 1976; Dogget and Prasada Rao 1995). Overall, they observed genetic diversity in wild sorghum to be low and non-spatially differentiated. Contrary to observations made by previous studies using both isozyme and RFLP makers and SSR markers (Morden *et al.* 1990; Aldrich and Doebley 1992) at global scale, Ayana *et al.* (2000a) found the overall genetic diversity of wild sorghum in

Ethiopia to be lower than what had been observed in cultivated forms using similar marker systems in another study (Ayana *et al.* 2000b). Reduction of wild sorghum populations through habitat destruction and fragmentation by human activities was thought to be the leading cause of these observations (Ayana *et al.* 2000a).

From the preceding review, it is evident that only few genetic diversity surveys have been conducted on wild sorghum using samples collected *in situ* from Africa, the centre of origin and diversity for cultivated sorghum. By observing a lack of spatial genetic differentiation in wild sorghum and less diversity in wild compared to cultivated sorghum, the study of Ayana *et al.* (2000a) is in contradiction with observations made using gene bank collections at global scale (Aldrich and Doebley 1992; Cui *et al.* 1995; Deu *et al.* 1995; Casa *et al.* 2005). This may indicate that global scale diversity study outcomes may be misleading if directly extrapolated to national and/or local scale for the management of conservation, breeding and/or biosafety programmes. Further studies at national and local scale using exhaustively sampled *in situ* collections of cultivated and wild sorghum, accompanied by detailed information on locations, growing environments and farmer practices, will be necessary if more meaningful evolutionary inferences of biosafety, conservation and utilisation purposes are to be made. There are still many gray areas with regard to genetic and evolutionary relationships between cultivated and wild sorghum that need to be explored. For example, while a number of authors (Morden *et al.* 1989; Aldrich and Doebley 1992; Cui *et al.* 1995; Ghebru *et al.* 2002; Casa *et al.* 2005) have speculated that introgressive hybridisation between cultivated and wild sorghum has been responsible for introducing new alleles into cultivated forms, little empirical evidence has been presented on the level and direction of geneflow between the two congeners. In Kenya this is compounded by the fact that little has been reported on the extent and partitioning of genetic diversity in cultivated as well as wild sorghum gene pools.

2.5 Advances in developing genetically modified sorghum

Optimised protocols for genetically transforming sorghum based on either agrobacterium or particle bombardment techniques are now in place (Casas *et al.* 1993; 1997; Zhao *et al.* 2000; Gao *et al.* 2005; Howe *et al.* 2006). Successful genetic engineering of sorghum has been reported for chitosanase and/or chitinase gene against fungal diseases (Zhu *et al.* 1998; Krishnaveni *et al.* 2000; Ayoo 2008), *Bacillus thuringiensis* (Bt) genes against stalk

borer (Girijashankar *et al*. 2005) and alpha-hordothionin protein gene originating from barley (*H. vulgare*) for high lysine content (Zhao *et al*. 2000). It is of interest to note that Kenyan sorghum landraces were used in the work of Ayoo (2008). Similar efforts to transform sorghum are underway, prominent among them, the initiative by the African Biofortified Sorghum project whose aim is to deploy a nutritionally enhanced and more digestible transgenic sorghum to subsistence farmers in Africa (Zhao 2008). There is current an urgent need therefore to generate science-based geneflow data in the wild-weedy-domesticate complex of *S. bicolor* for use by biosafety regulators in Africa with regard to testing and commercially releasing transgenic sorghum. In Kenya for example, a biosafety law was enacted at the beginning of this year paving way for on-farm deployment of transgenic crops, one of which could soon be sorghum.

2.6 Crop-to-wild geneflow and its potential consequences

Geneflow involves the movement and incorporation of genes between gene pools of populations (Futuyuma 1998). Along with genetic drift, selection and mutation, geneflow represents one of the main evolutionary forces shaping gene frequencies in diverse populations (Futuyuma 1998; Neal 2004). In plants, geneflow can occur via movement of pollen and hybridisation or by direct movement of seed or vegetative propagules such as stolons, rhizomes, stem cuttings, roots, crowns and bulbs. Pollen dispersal is the main mode by which flowering plants exchange genes and thus the chief mechanism of geneflow between populations of the same species or sexually compatible relatives (Levin and Kerster 1974).

Geneflow between crops and their wild relatives has been taking place since the dawn of agriculture (Ellstrand *et al*. 1999; Haygood *et al*. 2003). Out of the world's 13 most important food crops, 12 were identified to hybridise with their wild relatives somewhere within their agricultural range (Ellstrand *et al.* 1999). In a similar but more expanded review, Warwick and Stewart (2005) identified that only four out of the 25 globally important crops did not have sexually compatible weedy relatives. Armstrong *et al.* (2005) reviewed the potential of 123 temperate crops widely grown in New Zealand to hybridise with indigenous and introduced relatives. They found that 54% of the crops were reproductively compatible with at least one other indigenous or naturalised species, while a further 10% had at least some limited reproductive

compatibility with wild relatives. Such hybrids need to be only partially fertile to be able to mediate geneflow (Haygood *et al.* 2003) through introgressive hybridisation.

At least five factors must be satisfied before hybridisation and subsequent introgression can take place. Firstly, the two taxa in question must be situated near enough for pollen exchange to occur. Secondly, the populations in consideration must overlap at least partially in flowering time, to allow pollen from one population to find a mate in the other. Thirdly, for two taxa to hybridise, it is necessary that they share pollinators, a condition that is most easily satisfied for wind-pollinated species. The fourth condition necessary for hybridisation between two taxa is reproductive compatibility. Finally, the resultant F_1 hybrid must be viable and at least partially fertile, to allow for the introgression of alleles from one taxa into the other through backcrossing (Conner *et al*. 2003; Haygood *et al*. 2003; Chandler and Dunwell 2008).

Since the mid-1990's when the first transgenic crops were released, the land area under these crops has continued on a path of growth to over 120 million hectares worldwide, with approximately 13.3 million farmers in 25 countries growing transgenic crops today (James 2008). Almost equally expanding is scientific concern on the potential harmful consequences of transgene escape into wild and weedy relative populations (Snow and Moran-Palma 1997; Wei *et al*. 1999; Haygood *et al*. 2003; Armstrong *et al*. 2005; Chapman and Burke 2006; Auer 2008; Chandler and Dunwell 2008; Jhala *et al*. 2008). One of the possible consequences of transgene escape is replacement of wild genes by crop genes through genetic assimilation, the overall result of which is genetic erosion in wild populations (Haygood *et al*. 2003; Chapman and Burke 2006; Chandler and Dunwell 2008). If the resulting crop-wild hybrids are of lower fitness in comparison to their parents, wild populations may shrink and potentially become locally extinct. This is because smaller populations are more vulnerable to habitat disruption, inbreeding depression and other risks (Ellstrand and Elam 1993; Levin *et al*. 1996; Mooney and Cleland 2001; Haygood *et al*. 2003). Crop-wild hybrids can also potentially become invasive if they carry more fitness than their parents. Invasiveness in natural habitats is a conservation problem due to the threat posed to other members in the ecosystems. Similarly, crop-weed hybrids with enhanced fitness in farmlands can potentially evolve into more aggressive weeds leading to agricultural losses. Geneflow from cultivated crops to their wild relatives has been implicated in the evolution of more aggressive

weeds in seven out of the 13 most important crops (Ellstrand *et al*. 1999; Warwick and Stewart 2005). As a result of these concerns, different approaches have been used to investigate the potential of transgene escape via crop-to-wild geneflow in many crop species (Arias and Rieseberg 1994; Arriola and Ellstrand 1996; Desplanque *et al.* 1999; Montes-Hernandez and Eguiarte 2002; Papa and Gepts 2003; Chen *et al.* 2004; Andersen *et al.* 2005; Martinez-Castillo *et al.* 2007; Johnson and Galloway 2008; Tesso *et al.* 2008; Mutegi *et al.* 2009).

2.7 Methods of estimating geneflow

Geneflow can be estimated using direct or indirect methods. Direct approaches can be categorised into three: (i) measurement of pollen dispersal from point to source, (ii) measurement of gene dispersal from point and block sources and (iii) paternity analysis of progeny in sink populations. Indirect approaches on the other hand infer geneflow from the population genetic theory (Levin and Kerster 1974; Slatkin 1985a; Slatkin and Barton 1989; Niegel 1997; Whitlock and Mccauley 1999; Ellstrand 2003a; Pearse and Crandall 2004).

2.7.1 Direct approaches to geneflow estimation

2.7.1.1 Pollen dispersal from point to source

In this approach, pollen movement is estimated either through inferences from pollinator foraging distances (Marr *et al*. 2000; Osborne and Williams 2001; Chapman *et al*. 2003; Cresswell and Osborne 2004; Ferrari *et al*. 2006) or the dispersal of pollen analogues (Smouse *et al*. 2001; Gaudeul and Till-Bottraud 2004; Adler and Irwin 2006; Johnson and Galloway 2008), or both (Cresswell *et al*. 1995; Cresswell 2000; Marr *et al*. 2000). Such approaches have been used to extrapolate geneflow from GM crops to wild and weedy relatives. For example, Cresswell and Osborne (2004) inferred bumblebee pollen-mediated geneflow rates ranging from 2 - 8% between GM oilseed rape (*Brassica napus* L.) and its volunteer and feral populations in a fragmented agricultural landscape at distances ranging from 9 - 18 m.

2.7.1.2 Gene dispersal from point to source

The second method of geneflow estimation uses synthetic or sometimes natural stands of plants, as experimental pollen source populations. The source population, whose

plants will usually bear a unique genetic marker, is then surrounded at variable distances by pollen trap plants which are usually fixed for alternate alleles. After harvest, seedlings from seed borne by the sink population are progeny tested either by morphological or molecular markers depending on the system used. The fraction of progeny in the recipient population bearing the specific pollen-donor marker gives the hybridisation rate and hence estimates of potential geneflow between the two. Such studies have been conducted to estimate the potential crop-to-wild geneflow in a number of crops including lettuce (*Lactuca sativa* L.) (Andrea and Guadagnuolo 2008), maize (*Z. mays*) (Ellstrand *et al*. 2007), rice (*O. sativa*) (Arriola and Ellstrand 1996), sorghum [*S. bicolor*) (Arriola and Ellstrand 1996; Chen *et al*. 2004) and sunflower (*Helianthus annuus* L.) (Arias and Rieseberg 1994).

In the study on crop sorghum (*S. bicolor* ssp. *bicolor*) and its weedy tetraploid congener, johnsongrass (*S. halepense*), Arriola and Ellstrand (1996) observed spontaneous hybridisation rates varying from 0 - 100% at distances ranging from 0.5 - 100 m from the crop. Such studies have, however, not been conducted in Africa where sorghum was domesticated and continues to share sympatry with many of its closest wild and weedy relatives. The only similar study involving sorghum in Africa was by Schmidt and Bothma (2006) who looked at crop-to-crop geneflow risks in South Africa using a male sterile recipient population. They obtained average hybridisation rates of 0.06 - 2.5% at distances ranging from 13 - 158 m. Using mathematical models, Schmidt and Bothma (2006) estimated that geneflow from cultivated sorghum could take place up to a maximum distance of 700 m away. They reasoned that if the pollen source was a transgenic sorghum, presence of fertile crop and/or wild-weedy sorghum at these distances could lead to introgression of transgenes.

Estimation of geneflow from such experiments is, however, constrained by the artificial nature of the systems. Relatively uniform plants in monospecific experimental plots may be poor analogues of *in situ* populations where variations in plant sizes, spacing, phenology and the presence of other species that share pollen are expected to play a role in pollen dispersal (Ellstrand 1992). This may especially be true for small holder traditional mixed cropping systems that might characterise sorghum farming in many parts of Africa (Dogget and Prasada Rao 1995; Barnaud *et al.* 2008; Tesso *et al.* 2008).

Furthermore, such experimental systems suffer from difficulties of measuring long-distance dispersal events occurring beyond the study plot (Ellstrand 1992).

The common lesson from studies carried out to estimate geneflow from pollen dispersal and via spontaneous hybridisation is that pollen tends to be dispersed close to the source, with the frequency of pollinations declining with distance. Moreover, studies have shown that spontaneous hybridisation rates and therefore pollen mediated crop-to-wild geneflow will vary with species, populations, genotypes, environments and even seasons (Levin and Kerster 1974; Hamrick 1987; Ellstrand 2003a).

2.7.1.3 Estimates of geneflow through paternity analysis

Paternity analysis attempts to assign paternal parents to seeds collected from known maternal parent plants using genetic markers (Allison and Lewis 1993; Sork *et al*. 1999). Paternity analysis approaches have found wide use in investigations on both historical and contemporary geneflow, especially at landscape scale (Sork *et al.* 1999; Kameyama *et al.* 2000; Smouse *et al.* 2001; Austerlitz *et al.* 2004; Hardy *et al.* 2004; Slavov *et al.* 2005; Burczyk *et al.* 2006; Johannessen *et al.* 2006; van Treuren *et al.* 2006; Bacles and Ennos 2008). Generally, the first step in these investigations is to determine the multi-locus genotype of all reproductively mature plants in a population plus a sample of their progeny using highly polymorphic genetic markers. In a second step, paternity is inferred using either (i) paternity exclusion or (ii) likelihood-based paternity assignment approaches (Adams *et al.* 1992; Sork *et al*. 1999; Bernasconi 2003).

Paternity exclusion approaches generally compare genotypes of a sample of progeny to those of potential male parents within a well defined local population. Progeny whose multi-locus genotypes do not match those of all parents within the locality are assumed to result from pollen originating from outside the study locality. Using SSR and amplified fragment length polymorphism (AFLP) molecular markers in paternity exclusion approaches, for example, van Treuren *et al.* (2006) compared the multi-locus genotype of a rejuvenated ryegrass (*Lolium perenne* L.) gene bank accession to that of its parental sample to infer at least 1.6% cases of cross-accession contamination through pollen-mediated geneflow. In another study, Kameyama *et al.* (2000) used six SSR markers to investigate fine scale patterns of geneflow in a 150 x 70 m quadrant

containing 18 flowering plants of an endemic tree shrub, *Rhododendron matternichii* var. *hondoense*. The authors compared SSR genotypes of the seed offspring to the 18 potential parents to conclude that 20 - 30% of the seeds were sired by pollen originating from outside the quadrant.

In a field study aimed at investigating the potential geneflow from transplastomic oilseed rape (*B. napus*) via hybridisation with its sexually compatible weedy relative (*B. rapa* L.), Johannessen *et al*. (2006) reported hybridisation rates ranging from 0.4 - 5.3% with *B. napus* as the paternal parent. They observed hybridisation to decline with an increase in the number of crop plants relative to the weeds. While paternity exclusion approaches have the advantage of simplicity due to largely straight forward algorithms and potentially unambiguous results, they can underestimate geneflow due to successful immigrant gametes whose genotype is indistinguishable from locally produced gametes (Devlin and Ellstrand 1990). Furthermore, the method requires exhaustive sampling of potential parents and subsequent use of markers with high exclusion power. Even with considerable multi-locus polymorphism, unambiguous exclusion of paternity in large populations may still pose a major challenge for this approach (Allison and Lewis 1993; Parker *et al*. 1998; Slavov *et al*. 2005).

Prior to development of alternative genetic markers such as the SSRs, allozymes were the multi-locus genetic markers of choice in most paternity exclusion studies (Allison and Lewis 1993). As the need for sampling larger population sizes grew, however, allozyme markers proved to be inadequate in their exclusion power, necessitating the development of a number of maximum likelihood-based methods of assigning paternity (Meagher 1986; Thompson and Meagher 1987). Likelihood-based methods basically consider the statistical probability that a given male is the paternal parent, given the genotypes of the known maternal parent and offspring. The simplest of these is the 'most likely' method in which genotype probability scores (log odds, LOD) are calculated for potential parent-offspring pairs in a given study population and the male with the highest LOD score is identified as the maximum likelihood paternal parent (Meagher 1986). Another likelihood-based approach is the fractional paternity method (Devlin *et al*. 1988; Nielsen *et al*. 2001) which, unlike the 'most likely' method, assigns a proportion of parentage to even progeny with multiple, equally likely parents. Other modifications of the 'most likely' method use *a priori* information on factors affecting

the mating success in the study population such as spatial distances, fecundity and phenological overlap to shape paternity likelihood estimates (Meagher 1986; Adams *et al*. 1992). Likelihood methods have particularly found popular use in estimation of the level and patterns of contemporary geneflow at landscape scale using multi-locus genotype models (Sork *et al.* 1999). Multi-locus likelihood approaches have thus been used for among other studies, estimation of pollen movement distances within and among populations and approximation of the rate of pollen immigration into populations at landscape scale (Hardy *et al*. 2004; Burczyk *et al*. 2006; Hanaoka *et al*. 2007).

2.7.2 Indirect geneflow estimation methods

Indirect methods infer geneflow from population genetics theory and associated parameters such as the fixation index (F_{ST}) and its analogues (Wright 1951; Nei 1973; Weir and Cockerham 1984; Slatkin 1985a; Weir and Hill 2002), Bayesian model-based genetic admixture (Pritchard *et al*. 2000), coalescent theory-based maximum likelihood (Kingman 1982; Beerli and Felsenstein 1999; 2001) and spatial autocorrelation analysis (Sokal and Oden 1978a; 1978b; Barbujani 1987; Escudero *et al*. 2003). This group of indirect approaches generally rely on statistical models that have been developed to use either allele frequency distributions or the genealogy of DNA sequences, to indirectly infer geneflow. The first step in applying the approaches is to sample organisms from the target population and depending on the genetic marker system used, data on the spatial distribution of alleles, chromosomal segments or phenotypic traits is generated to calculate statistical parameters. Following estimation, these statistical parameters are related to a model of geneflow and of other mechanisms of genetic evolution in order to draw inferences on, not only the level and pattern of geneflow, but on other evolutionary factors and dynamics of the population as well (Slatkin 1985a; Ellstrand 1992; Niegel 1997).

2.7.2.1 Fixation index based geneflow estimation

Geneflow has commonly been inferred from the existing population structure by use of various statistical methods that indirectly estimate the parameter, *Nm*, the average number of effective immigrants per generation, in a population. Most species exhibit spatial variation in both morphology and gene frequency, which is dictated by a balance between forces tending to create local genetic differentiation and those tending to produce homogeneity. Geneflow is the evolutionary force that opposes genetic

differentiation, while mutation, genetic drift due to finite population sizes and natural selection are the main forces that favour genetic differentiation of local populations. Observed patterns of population genetic differentiation can consequently be applied in models based on population genetics theory to predict historical geneflow (Slatkin 1987; Futuyuma 1998; Hartl 2000).

Wright (1951) developed a method of estimating geneflow using the fixation index of population subdivision relative to the total population (F_{ST}), with an assumption of an Island model of population structure. This model assumes equilibrium between the homogenising effects of geneflow and the disruptive effects of genetic drift, uniform population size and constant rates of geneflow over space and time. The average number of effective migrants per generation, *Nm* is determined from the following expression (Wright 1951):

$$F_{ST} \approx \frac{1}{(4Nm + 1)}$$

where F_{ST} is defined as genetic variance among (sub)populations in the total sample and therefore a measure of the genetic differentiation among populations. This model has been further extended through the development of other methods for deriving a measure of population differentiation under different assumptions (Nei 1973; Weir and Cockerham 1984; Slatkin 1985b). Slatkin (1985a) proposed the parameter R_{ST}, which uses frequency of "rare alleles" to derive population genetic differentiation, while the parameter G_{ST} was proposed by Nei (1973) to measure genetic differentiation in populations using multi-locus and multi-allelic genetic data. The other commonly used population genetic differentiation parameter is θ which takes care of unequal sample sizes in the multi-locus and multi-allelic genetic data (Weir and Cockerham 1984).

The F_{ST}-based approaches have been widely used to study population structure and historical rates of geneflow in diverse populations of both plant and animal species. For example, Cui and co-workers (1995) used this approach to study phylogenetic relationships and geneflow between cultivated and wild sorghum. They used genotypic data based on nuclear and cytoplasmic RFLP analysis of 42 cultivated accessions and 12 wild sorghum genotypes to suggest low levels of genetic differentiation and high rates of introgression between cultivated and wild sorghum. Like in most studies using gene

bank accessions, however, their samples were not exhaustive and rates of geneflow calculated from F_{ST} values may have been overestimated.

The F_{ST}-based approaches have come under increased criticism (Sork *et al*. 1999; Whitlock and Mccauley 1999; Pearse and Crandall 2004) mainly due to the underlying population structure assumptions of constant population size, symmetrical and constant inter-population geneflow rates and persistent populations for periods sufficient to achieve equilibrium between geneflow and genetic drift (Wright 1951; 1978). Whitlock and Mccauley (1999) for example, argued that these underlying assumptions are biologically unrealistic and likely to be violated by most populations, such that there is no quantitative information to be gained about dispersal using gene frequency data. The authors recommend that F_{ST} and its variants be limited to the study of the extent of population genetic structure, arguing that rarely they translate to accurate estimates of geneflow. Their view was supported by Pearse and Crandall (2004) who argued that due to the simple assumptions, F_{ST}-based approaches are likely to be biased estimates of geneflow. Examples include conservation genetic assessment studies where most populations and species of conservation concern are small and/or have recently declined in size, experienced fragmentation or be otherwise disturbed. Similarly, Sork *et al.* (1999) pointed out that since F_{ST} (and its analogues) represent a summary statistic for a set of sub-populations, it provides little insight into landscape-level processes that differentially affect spatial patterns of genetic structure and rates of geneflow among sub-populations.

Criticism on the F_{ST}-based approaches, coupled with advances in molecular data collection and statistical analysis, have led to the development of complementary methods for studying population genetic structure and to indirectly estimate geneflow with far less demographic assumptions. These methods can be grouped into two broad categories: (i) Coalescent or genealogical–based approaches that rely on the ancestral genealogy of DNA sequences to infer geneflow (Beerli and Felsenstein 1999; 2001; Beerli 2006) and (ii) Bayesian model-based multi-locus approaches that use transient gametic disequilibrium information to infer geneflow (Pritchard *et al*. 2000; Falush *et al*. 2003; Baudouin *et al*. 2004; Beaumont and Rannala 2004; Beerli 2006). There is currently an increased interest to use these alternate methods, especially in answering practical conservation and evolutionary genetics questions.

2.7.2.2 Coalescent-based approaches

In principle, coalescent-based methods analyse DNA sequence lineages in populations, in order to make inferences on their demographic and evolutionary history. Several population-level characteristics, including migration (geneflow) rate (*m*) (Beerli and Felsenstein 1999; 2001), effective population size (*Ne*) (Strobeck 1983) and population growth rate (Kuhner *et al*. 1998) are estimated, either simultaneously or separately. The maximum likelihood coalescent method implemented in the software MIGRATE (Beerli and Felsenstein 1999) for example, is a two-population model that can be used to simultaneously estimate *Ne* and *m*, by analysing the genealogical relationships among alleles both within and among populations. The same authors later extended this method (Beerli and Felsenstein 2001) to allow for the estimation of multi-directional geneflow and effective population size for more than two sub-populations. Such approaches have lately attracted interest in the study of geneflow and other parameters of population dynamics in a number of plant species, including *Araucaria angustifolia* (Bert.) O. Ktze. *(*Araucariaceae*)* (Stefenon *et al*. 2008), *B. cretica* (Guss.) (Brassicaceae) (Edh *et al*. 2007), *Lupinus microcarpus* (Benth) Jeps. (Fabaceae) (Drummond and Hamilton 2007) and serpentine sunflower (*H. exilis*) (Sambatti and Rice 2006).

For example, Stefenon *et al*. (2008) recently used the coalescent-based maximum likelihood method of Beerli and Felsenstein (1999) to estimate geneflow between two neighbouring populations of the sub-tropical confer species, *A. angustifolia* (Berth.). Their study revealed estimates of effective migration rate about one migrant per generation, a scenario congruent with low inter-population geneflow in the species. The study of Edh and co-workers (2007) provides a good example of determining population structure and geneflow using the coalescent-based methods. The authors used nuclear and chloroplast SSRs to reveal high levels of differentiation (F_{ST} = 0.628 and 1.00, respectively) accompanied by low rates of geneflow (*Nm* = 0.286) among six endemic populations of *B. cretica* in the Aegean islands of Greece. Like the traditional F_{ST}-based methods, coalescent-based approaches are based on assumptions such as constant population size, constant mutation rate, constant migration rate and a drift-migration balance. Furthermore, due to the complex nature of performing simultaneously computation parameters such as *Ne* and *m*, caution is required in interpreting the results from the computer software (Beebee and Rowe 2008).

2.7.2.3 Bayesian model-based approaches

One of the approaches that has received great interest from researchers is the method developed by Pritchard *et al.* (2000) and expanded by Falush *et al.* (2003). This approach uses multi-locus genotypic data in Bayesian statistics to infer population structure, assign individuals to populations, as well as identify migrants and admixture in individuals. The approach assumes a *K* number of unknown populations (genetic clusters) in the study sample, each characterised by a set of allele frequencies at each locus. Each individual in the sample is probabilistically assigned to a cluster or jointly to two or more clusters if their genotypes indicate they are admixed. The loci are assumed to be at Hardy-Weinberg (HWE) and linkage equilibrium (LE) (Pritchard *et al.* 2000). A growing number of authors have recently used this method to study population genetic structure and geneflow in many plant species (Andersen *et al.* 2005; Garris *et al.* 2005; Breton *et al.* 2006; Kuroda *et al.* 2006; Mariac *et al.* 2006; Curtu *et al.* 2007; Konishi and Ohnishi 2007; Martinez-Castillo *et al.* 2007). Bayesian model-based methods have been used only to a limited extent in investigating genetic structure in cultivated sorghum (Barnaud *et al*. 2007) but not in the wild sorghum gene pools or in investigating introgression between the two reproductively compatible congeners.

Examples in other plant species include the study of Garris *et al.* (2005), which used allelic data from 169 nuclear SSRs and sequence information from two cpDNA loci to infer and assign 234 accessions of rice (*O. sativa*) into five distinct genetic groups that corresponded to five recognised varietal groups in the crop. In another study, Curtu *et al.* (2007) used isozyme and SSR data to decipher population structure and the extent of hybridisation among six sympatrically distributed species of oak (*Quercus* ssp.) in west-central Romania. In the study, the Bayesian model-based analysis (Pritchard *et al.* 2000) revealed four distinct genetic clusters each corresponding to one of the four co-occurring species and further demonstrated pair-wise interspecific introgression levels between the six oak species that ranged from 1.7 - 16.2%.

Similarly, these approaches have been used to study population structure and geneflow between domesticated crops and their wild and weedy relatives. In a comparative genetic structure study involving a taxa closely related to sorghum, Mariac *et al.* (2006) used 24 SSR loci to genotype 46 individuals of wild pearl millet [*Pennisetum glaucum* (L.) R. Br. ssp. *monodii*] and 426 individuals of cultivated pearl millet (*P. glaucum* ssp.

glaucum) in Niger. They used the Bayesian model-based analysis to demonstrate strong differentiation between cultivated and wild pearl millet taxa and at the same time reveal reciprocal introgression between the two taxa. In another similar study Andersen *et al.* (2005) investigated geneflow to the wild sea beet [(*Beta vulgaris* L. ssp. *maritima* (L.) Arcangeli)] from cultivated sugar beet [(*B. vulgaris* L. ssp. *vulgaris*) (L.) Arcangeli)] with a view of contributing to biosafety guidelines for seed production in GM sugar beet in Denmark. They used eight SSR loci to genotype individuals from 12 sea beet populations originating from Denmark, six more from neighbouring countries and four sugar beet lines originating from Denmark. At least seven individuals showed triallelic genotypes at one or two SSR loci, an indication of hybridisation between the two taxa. Further admixture analysis by Bayesian model-based method (Pritchard *et al.* 2000), however, revealed no evidence of introgression among cultivated and wild beet subspecies; perhaps underlining that introgression beyond the first hybridisation is not extensive (Andersen *et al.* 2005).

A more recent example is the work of Martinez-Castillo *et al*. (2007) which used nine SSR loci to study the extent and direction of geneflow in the wild-weedy-domesticate complex of lima bean (*Phaseolus lunatus*) under traditional agricultural systems in four regions of Mexico, one of the centres of origin and diversity of the crop. Bayesian admixture analysis revealed a three-fold higher level of geneflow via introgressive hybridisation from the domesticated to the wild populations than in the opposite direction (Martinez-Castillo *et al*. 2007). This observation led them to conclude that genetic assimilation of the wild lima bean by its domesticated counterpart was a possibility and a potential route for transgene escape in the centre of origin and diversity of the crop.

Combined, these results serve to demonstrate that approaches based on Bayesian statistical inference of population structure and analysis of admixture events (Pritchard *et al.* 2000; Falush *et al.* 2003) are effective in assessing the extent of genetic differentiation and introgression within and between populations of diverse taxa. The methods therefore offer practical solutions to questions on conservation, evolution and biosafety with regard to understanding the partitioning of genetic diversity within and between crops and their wild relatives and in elucidating the level and direction of inter-taxa and inter-population geneflow. Such studies focusing on comparative genetic

structure and introgression between cultivated and wild sorghum gene pools have not yet been reported but are important for contributing to their conservation, breeding and biosafety strategies in Africa.

2.7.2.4 Spatial autocorrelation analysis

Spatial autocorrelation analysis is a set of statistical procedures designed to detect and quantify correlation between samples for a given variable (e.g. allelic frequency or morphological measurements) as a function of spatial distance (Sokal and Oden 1978a; 1978b; Heywood 1991). In principle, every possible pair of sample locations *i* and *j* ($i \neq j$) is assigned a distance class and used to estimate spatial dependency for the variable in consideration. The approach is commonly used to investigate non-random spatial distribution of genotypes due to isolation-by-distance, the restricted geneflow among individuals. Isolation-by-distance leads to higher genetic similarity among neighbouring individuals than among more distant individuals. Examples of spatial autocorrelation coefficients include the traditional Moran's I (Moran 1950) and Gearly's c (Geary 1954) indices, but also the relatively recent Kinship coefficients (Loiselle *et al*. 1995; Ritland 1996; Rousset 2000; Hardy 2003). The coefficient of kinship between two individuals A and B is the probability that a gene taken at random from A, at a given locus, may be identical in descent to a gene taken at random from B at the same locus. Kinship coefficients are then defined as $F_{ij} \equiv (Q_{i,j} - Q)/(1-Q)$, where Q represents the probabilities of identity in state between two genes: $Q_{i,j}$ for random genes between individuals *i* and *j* and Q for random genes within the reference population sampled. The pairwise kinship coefficients thus measure the correlation in frequencies of homologous alleles between individuals (Barbujani 1987; Heywood 1991). Unlike the former spatial genetic structure statistics, the kinship coefficients have a well-developed foundation in population genetics theory and thus provide a natural means of combining data both over multiple alleles at a locus and across loci to obtain a more powerful test of genetic structure (Heywood 1991; Smouse and Peakall 1999).

Results of a spatial autocorrelation analysis are presented in the form of correlograms, graphic displays in which the values of autocorrelation coefficients are plotted against distance classes. Tests of significance are obtained for each distance class by a randomisation process under the null hypotheses of random relationship among samples of a population. (Sokal and Oden 1978a; 1978b; Heywood 1991; Escudero *et al.* 2003).

Many studies have used spatial autocorrelation approaches to investigate spatial genetic structure and geneflow within and among populations of diverse plant species (Fenster *et al.* 2003; Papa and Gepts 2003; Kuroda *et al.* 2006; Shimono *et al.* 2006; Myers *et al.* 2007; Hu *et al.* 2008). For example Papa and Gepts (2003) performed spatial autocorrelation analysis based on AFLP data on 382 individuals consisting of 18 wild and 13 domesticated common beans (*P. vulgaris*) from different regions of Mexico. The authors observed lower levels of isolation-by-distance among domesticated populations compared to those of its wild/weedy counterpart. This was attributed to higher levels of farmer driven seed-mediated geneflow among the populations of domesticated common beans. Kuroda and co-authors (2006) analysed 616 individuals of wild soyabean consisting of 77 populations collected from its entire distribution range in Japan, using 20 SSR markers. Spatial autocorrelation analysis of the data revealed an isolation-by-distance pattern with possibility of geneflow among populations as far apart as 100 km.

2.8 Conclusions

Sorghum is a crop of global importance and contributes directly to the dietary and nutritional needs of over 100 million resource challenged people in Africa. Africa is the center of origin, domestication and diversification for sorghum. Diverse cultivar types of the crop still share sympatry with its closest wild and weedy relatives in many agro-ecosystems of Africa. The two are inter-fertile giving rise to a heterogeneous wild-weedy-domesticate complex, which taxonomically classified together in a single species, *S. bicolor*. The socio-economic importance enjoyed by the crop makes it an obvious target for genetic engineering with a view to overcoming some of its productivity and nutritional bottlenecks. Nonetheless, fear abounds that wild and weedy relatives of sorghum will most likely act as conduits for escape of transgenes for GM sorghum into the environment with possible negative consequences in agricultural and natural environments such as increased weediness, invasiveness, genetic erosion and population extinctions. There is currently an urgent need to characterise the environmental risks for the deployment of GM sorghum in Africa, in order to develop scientifically supported biosafety regulations and guidelines. Understanding the extent and direction of movement of genes between the domesticated and wild and/or weedy relative populations of sorghum is the first step in characterising the potential environmental risks of escaped transgenes. Such information is currently lacking, but is critically needed for science based decision making by biosafety regulators in Africa. The wild-weedy-domesticate

complex of sorghum in Africa constitutes important genetic resources for crop breeding programmes as it is a rich source of important traits: pest/disease resistances, tolerance to abiotic stresses and yield enhancement. Nevertheless, information on the extent and organisation of genetic diversity in the wild-weedy-complex of sorghum is limited but important for formulating effective genetic resources conservation and utilisation strategies. In the current study, SSR molecular markers were combined with approaches based on population genetics theory to analyse samples of wild-weedy-domesticate complex of *S. bicolor* at a country and local scale. The main aim of the study was to elucidate the extent and direction of geneflow between cultivated and wild sorghum within the traditional agro-ecosytems in Kenya. In addition, the study aimed at expanding knowledge on the extent and organisation of genetic diversity within and among cultivated and wild sorghum genepools in the country. Results from the study are expected to attract practical applications, first in formulating biosafety regulations and guidelines for testing and commercially releasing GM sorghum and secondly in formulating appropriate strategies for conservation and utilisation genetic resources of cultivated sorghum and its wild-weedy relatives.

2.9 References

Abu-Assar AH, Uptmoor R, Abdelmula AA, Salih M, Ordon F, Friedt W. 2005. Genetic variation in sorghum germplasm from Sudan, ICRISAT, and USA assessed by simple sequence repeats (SSRs). Crop Science 45:1636-1644.

Adams WT, Griffin AR, Moran GR. 1992. Using paternity analysis to measure effective pollen dispersal in plant populations. The American Naturalist 140:762-780.

Adler LS, Irwin RE. 2006. Comparison of pollen transfer dynamics by multiple floral visitors: Experiments with pollen and fluorescent dye. Annals of Botany 97:141-150.

Aldrich PR, Doebley J. 1992. Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics 85:293-302.

Aldrich PR, Doebley J, Schertz KF, Stec A. 1992. Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics 85:451-460.

Allison AS, Lewis PO. 1993. Reproductive traits and male fertility in plants: Empirical approaches. Annual Review of Ecology and Systematics 24:331-351.

Andersen NS, Siegismund HR, Meyer V, Jorgensen RB. 2005. Low level of geneflow from cultivated beet (*Beta vulgaris* L. ssp *vulgaris*) into Danish populations of sea beet (*Beta vulgaris* L. ssp *maritima* (L.) Arcangeli). Molecular Ecology 14:1391-1405.

Andrea LD, Guadagnuolo R. 2008. Hybridization rates between lettuce (*Lactuca sativa*) and its wild relative (*L. serriola*) under field conditions. Environmental Biosafety Research 7:61-71.

Arias DM, Rieseberg LH. 1994. Geneflow between cultivated and wild sunflowers. Theoretical and Applied Genetics 89:655-660.

Armstrong TT, Fitzjohn RG, Newstrom LE, Wilton AD, Lee WG. 2005. Transgene escape: what potential for crop-wild hybridization? Molecular Ecology 14:2111-2132.

Arriola PE, Ellstrand NC. 1996. Crop-to-weed geneflow in the genus *Sorghum* (Poaceae): Spontaneous inter specific hybridization between Johnsongrass, *Sorghum halepense*, and crop sorghum, *S. bicolor*. American Journal of Botany 83:1153-1160.

Auer C. 2008. Ecological risk assessment and regulation for genetically-modified ornamental plants. Critical Reviews in Plant Sciences 27:255-271.

Austerlitz F, Dick CW, Dutech C, Klein EK, Oddou-Muratorio S, Smouse PE, Sork VL. 2004. Using genetic markers to estimate the pollen dispersal curve. Molecular Ecology 13:937-954.

Ayana A, Bekele E, Bryngelsson T. 2000a. Genetic variation in wild sorghum (*Sorghum bicolor* ssp *verticilliflorum* (L.) Moench) germplasm from Ethiopia assessed by random amplified polymorphic DNA (RAPD). Hereditas 132:249-254.

Ayana A, Bryngelsson T, Bekele E. 2000b. Genetic variation of Ethiopian and Eritrean sorghum (*Sorghum bicolor* (L.) Moench) germplasm assessed by random amplified polymorphic DNA (RAPD). Genetic Resources and Crop Evolution 47:471-482.

Ayana A, Byngelsson T, Bekele E. 2001. Geographic and altitudinal allozyme variation in sorghum (*Sorghum bicolor* (L.) Moench) landraces from Ethiopia and Eritrea. Hereditas 135:1-12.

Ayoo LMK. 2008. Genetic transformation of Kenyan sorghum (*Sorghum bicolor* L. Moench) with anti-fungal genes and response to *Collectotrichum sublineolum* infection. PhD Dissertation, University of Hamburg, Germany. pp. 116.

Bacles CFE, Ennos RA. 2008. Paternity analysis of pollen-mediated geneflow for *Fraxinus excelsior* L. in a chronically fragmented landscape. Heredity 101:368-380.

Barbujani G. 1987. Autocorrelation of gene frequencies under isolation by distance. Genetics 117:777-782.

Barnaud A, Deu M, Garine E, Mckey D, Joly HI. 2007. Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. Theoretical and Applied Genetics 114:237-248.

Barnaud A, Trigueros G, Mckey D, Joly HI. 2008. High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? Heredity 101:445-452.

Baudouin L, Piry S, Cornuet JM. 2004. Analytical Bayesian approach for assigning individuals to populations. Journal of Heredity 95:217-224.

Beaumont MA, Rannala B. 2004. The Bayesian revolution in genetics. Nature 5:251-261.

Beebee JC, Rowe G. 2008. An introduction to molecular ecology. Oxford University Press. pp. 400.

Beerli P. 2006. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics 22:341-345.

Beerli P, Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763-773.

Beerli P, Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in *n* subpopulations by using a coalescent approach. Proceeding of the National Academy of Sciences of the USA (PNAS) 98:4563-4568.

Bernasconi G. 2003. Seed paternity in flowering plants: an evolutionary perspective. Perspectives in Plant Ecology, Evolution and Systematics 6:149-158.

Breton C, Tersac M, Berville A. 2006. Genetic diversity and geneflow between the wild olive (oleaster, *Olea europaea* L.) and the olive: several Plio-Pleistocene refuge zones in the Mediterranean basin suggested by simple sequence repeats analysis. Journal of Biogeography 33:1916-1928.

Burczyk J, Adams WT, Birkes DS, Chybicki IJ. 2006. Using genetic markers to directly estimate geneflow and reproductive success parameters in plants on the basis of naturally regenerated seedlings. Genetics 173:363-372.

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S. 2005. Diversity and selection in Sorghum: simultaneous analyses using simple sequence repeats. Theoretical and Applied Genetics 111:23-30.

Casas AM, Kononowicz AK, Haan TG, Hang L, Tomes DL, Bressan RA, Hasegawa PM. 1997. Transgenic sorghum plants obtained after microprojectile bombardment of immature inflorescences. In vitro Cellular and Developmental Biology-Plant 33:92-100.

Casas AM, Kononowicz AK, Zehr BU, Tomes TD, Axtell DJ, Butler GL, Bressan AR, Hasegawa PM. 1993. Transgenic sorghum plants via microprojectile bomabardment. Proceeding of the National Academy of Sciences of the USA (PNAS) 90:11212-11216.

Chandler S, Dunwell JM. 2008. Geneflow, risk assessment and the environmental release of transgenic plants. Critical Reviews in Plant Sciences 27:25-49.

Chapman MA, Burke JM. 2006. Letting the gene out of the bottle: the population genetics of genetically modified crops. New Phytologist 170:429-443.

Chapman RE, Wang J, Bourke AFG. 2003. Genetic analysis of spatial foraging patterns and resource sharing in bumble bee pollinators. Molecular Ecology 12:2801-2808.

Chen LJ, Lee DS, Song PZ, Suh HS, Lu B. 2004. Geneflow from cultivated rice (*Oryza sativa*) to its weedy and wild relatives. Annals of Botany 93:67-73.

Cleveland DA, Soleri D. 2005. Rethinking the risk management process for genetically engineered crop varieties in small-scale, traditionally based agriculture. Ecology and Society 10:1-33.

Conner AJ, Glare TR, Nap JP. 2003. The release of genetically modified crops into the environment. Part II. Overview of ecological risk assessment. The Plant Journal 33:19-46.

Cresswell JE. 2000. A comparison of bumblebees' movements in uniform and aggregated distributions of their forage plant. Ecological Entomology 25:19-25.

Cresswell JE, Bassom AP, Bell SA, Collins SJ, Kelly TB. 1995. Predicted pollen dispersal by honey-bees and three species of bumble-bees foraging on oil-seed rape: A comparison of three models. Functional Ecology 9:829-841.

Cresswell JE, Osborne JL. 2004. The effect of patch size and separation on bumblebee foraging in oilseed rape: implications for geneflow. Journal of Applied Ecology 41:539-546.

Cui YX, Xu GW, Magill CW, Schertz KF, Hart GE. 1995. RFLP-based assay of *Sorghum bicolor* (L) Moench. genetic diversity. Theoretical and Applied Genetics 90:787-796.

Curtu A, Gailing O, Finkeldey R. 2007. Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. BMC Evolutionary Biology 7:218.

De Wet JMJ. 1978. Systematics and evolution of sorghum sect. Sorghum (Gramineae). American Journal of Botany 65:477-484.

De Wet JMJ, Harlan JR, Price EG. 1970. Origin of variability in the Spontanea complex of *Sorghum bicolor*. American Journal of Botany 57:704-707.

De Wet JMJ, Huckabay JP. 1967. The origin of *Sorghum bicolor*. II. Distribution and domestication. Evolution 21:787-802.

De Wet JMJ, Huckabay JP. 1971. Origin and domestication of *Sorghum bicolor*. Economic Botany 95:128-134.

Desplanque B, Boudry P, Broomberg K, Saumitou-Laprade P, Cuguen J, Van Dijk H. 1999. Genetic diversity and geneflow between wild, cultivated and weedy forms of *Beta vulgaris* L. (Chenopodiaceae), assessed by RFLP and microsatellite markers. Theoretical and Applied Genetics 98:1194-1201.

Deu M, Gonzalezdeleon D, Glaszmann JC, Degremont I, Chantereau J, Lanaud C, Hamon P. 1994. RFLP diversity in cultivated sorghum in relation to racial differentiation. Theoretical and Applied Genetics 88:838-844.

Deu M, Hamon P, Chantereau J, Dufour P, Dhont A, Lanaud C. 1995. Mitochondrial-DNA diversity in wild and cultivated sorghum. Genome 38:635-645.

Deu M, Rattunde HFW, Chantereau J. 2006. A global view of genetic diversity in cultivated sorghum using a core collection. Genome 49:168-180.

Deu M, Sagnard F, Chantereau J, Calatayud C, Herault D, Mariac C, Pham JL, Vigouroux Y, Kapran I, Traore PS, Mamadou A, Gerard B, Ndjeunga J, Bezancon G. 2008. Niger-wide assessment of *in situ* sorghum genetic diversity with microsatellite markers. Theoretical and Applied Genetics 116:903-913.

Devlin B, Ellstrand NC. 1990. The development and application of a refined method for estimating geneflow from angiosperm paternity analysis. Evolution 44:248-259.

Devlin B, Roeder K, Ellstrand NC. 1988. Fractional paternity assignment: theoretical development and comparison to other methods. Theoretical and Applied Genetics 76:369-380.

Dillon SL, Shapter FM, Henry RJ, Cordeiro G, Izquierdo L, Lee LS. 2007. Domestication to crop improvement: Genetic resources for Sorghum and Saccharum (Andropogoneae). Annals of Botany 100:975-989.

Dje Y, Ater M, Lefebvre C, Vekemans X. 1998. Patterns of morphological and allozyme variation in sorghum landraces of northwestern Morocco. Genetic Resources and Crop Evolution 45:541-548.

Dje Y, Forcioli D, Ater M, Lefebvre C, Vekemans X. 1999. Assessing population genetic structure of sorghum landraces from North-western Morocco using allozyme and microsatellite markers. Theoretical and Applied Genetics 99:157-163.

Dje Y, Heuertz M, Lefebvre C, Vekemans X. 2000. Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. Theoretical and Applied Genetics 100:918-925.

Dogget H. 1988. Sorghum. Longman Scientific and Technical Essex, England, pp. 512.

Dogget H, Majisu BN. 1968. Disruptive selection in crop development. Heredity 23:1-23.

Dogget H, Prasada Rao KE. 1995. Sorghum. In: Smartt J, Simmonds NW (Eds.), Evolution of Crop Plants. Longman Group, Burnt Mill. pp. 180.

Drummond CS, Hamilton MB. 2007. Hierarchical components of genetic variation at a species boundary: population structure in two sympatric varieties of *Lupinus microcarpus* (Leguminosae). Molecular Ecology 16:753-769.

Edh K, Widen B, Ceplitis A. 2007. Nuclear and chloroplast microsatellites reveal extreme population differentiation and limited geneflow in the Aegeas endemic *Brassica cretica* (Brassicaceae). Molecular Ecology 16:4972-4983.

Ellstrand NC. 1992. Geneflow among seed plant populations. New Forests 6:241-256.

Ellstrand NC. 2003a. Current knowledge of geneflow in plants: implications for geneflow. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 358:1163-1170.

Ellstrand NC. 2003b. Dangerous liaisons - when cultivated plants mate with their wild relatives. Johns Hopkins University Press Baltimore MD, pp. 244.

Ellstrand NC, Elam DR. 1993. Population genetic consequences of small population size: Implications for plant conservation. Annual Review of Ecology and Systematics 24:217-242.

Ellstrand NC, Garner LC, Hegde S, Guadagnuolo R, Blancas L. 2007. Spontaneous hybridization between maize and teosinte. Journal of Heredity 98:183-187.

Ellstrand NC, Prentice HC, Hancock JF. 1999. Geneflow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics 30:539-563.

Escudero A, Iriondo JM, Torres ME. 2003. Spatial analysis of genetic diversity as a tool for plant conservation. Biological Conservation 113:351-365.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587.

FAO. 2008. FAOSTAT. http://faostat.fao.org.

Fenster CB, Vekemans X, Hardy OJ. 2003. Quantifying geneflow from spatial genetic structure data in a metapopulation of *Chamaecrista fasciculataI* (Leguminosae). Evolution 57:995-1007.

Ferrari MJ, Bjornstad ON, Partain JL, Antonovics J. 2006. A gravity model for the spread of a pollinator-borne plant pathogen. American Naturalist 168:294-303.

Folkertsma RT, Frederick H, Rattunde HFW, Chandra S, Raju GS, Hash CT. 2005. The pattern of genetic diversity of Guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. Theoretical and Applied Genetics 111:399-409.

Futuyuma JD. 1998. Evolutionary Biology. Sinauer Associates Inc. Sunderland, Mass, pp. 763

Gao Z, Xie X, Ling Y, Muthukrishnan S, Liang GH. 2005. *Agrobacterium tumefaciens*-mediated sorghum transformation using a mannose selection system. Plant Biotechnology Journal 3:591-599.

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S. 2005. Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631-1638.

Gaudeul M, Till-Bottraud I. 2004. Reproductive ecology of the endangered alpine species *Eryngium alpinum* L. (Apiaceae): Phenology, gene dispersal and reproductive success. Annals of Botany 93:711-721.

Geary RC. 1954. The contiguity ratio and statistical mapping. The Incorporated Statistician 5:115-145.

Ghebru B, Schmidt RJ, Bennetzen JL. 2002. Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. Theoretical and Applied Genetics 105:229-236.

Girijashankar V, Sharma KK, Swathisree V, Prasad LS, Bhat BV, Royer M, Narasu ML, Altosaar I, Seetharama N. 2005. Development of transgenic sorghum for insect resistance against the spotted stem borer (*Chilo partellus*). Plant Cell Reports 24:513-522.

Gurney AL, Press MC, Scholes JD. 2002. Can wild relatives of sorghum provide new sources of resistance or tolerance against *Striga* species? Weed Science 42:317-324.

Hajjar R, Hodgkin T. 2007. The use of wild relatives in crop improvement: a survey of developments over the last 20-years. Euphytica 156:1-13.

Hamrick JL. 1987. Geneflow and distribution of genetic variation in plant populations. In: Urbanska KM (Ed.), Differentiation Patterns in Higher Plants. Academic press, London. pp. 67.

Hanaoka S, Yuzurihara J, Asuka Y, Tomaru N, Tsumura Y, Kakubari Y, Mukai Y. 2007. Pollen-mediated geneflow in a small, fragmented natural population of *Fagus crenata*. Canadian Journal of Botany 85:404-413.

Hardy OJ. 2003. Estimation of pairwise relatedness between individuals and characterization of isolation by distance processes using dominant genetic markers. Molecular Ecology 12:1577-1588.

Hardy OJ, Gonzalez-Martinez SC, Freville H, Boquien G, Mignot A, Colas B, Olivieri I. 2004. Fine-scale genetic structure and gene dispersal in *Centaurea corymbosa* (Asteraceae) I. Pattern of pollen dispersal. Journal of Evolutionary Biology 17:795-806.

Harlan JR. 1975. Crops and man. American Scociety of Agronomy Madison, pp. 284.

Harlan JR, De Wet JMJ. 1971. Toward a rational classification of cultivated plants. Taxon 20:509-517.

Harlan JR, De Wet JMJ. 1972. A simplified classification of cultivated sorghum. Crop Science 12:172-177.

Harlan JR, De Wet JMJ, Stemler ABL. 1976. Origins of African plant domestication. Mouton The Hague, Netherlands, pp. 498.

Hartl DL. 2000. A primer of population genetics. Sinauer Associates, Inc. Sunderland, Massachusetts, pp. 221.

Haygood R, Ives AR, Andow DA. 2003. Consequences of recurrent geneflow from crops to wild relatives. Proceedings of the Royal Society of London 270:1879-1886.

Heywood JS. 1991. Spatial analysis of genetic variation in plant populations. Annual Review of Ecology and Systematics 22:335-355.

Howe A, Shirley S, Dweikat I, Fromm M, Clemente T. 2006. Rapid and reproducible *Agrobacterium*-mediated transformation of sorghum. Plant Cell Reports 25:751-758.

Hu Y, Zhu Y, Zhang QY, Xin HL, Qin LP, Lu BR, Rahman K, Zheng HC. 2008. Population genetic structure of the medicinal plant *Vitex rotundifolia* in China: Implications for its use and conservation. Journal of Integrative Plant Biology 50:1118-1129.

James C. 2008. Global status of commercialised Biotech/GM crops: 2008 The First Thirteen Years, 1996 to 2008. ISAAA Brief No. 39 (http://www.isaaa.org).

Jhala AJ, Hall LM, Hall JC. 2008. Potential hybridization of flax with weedy and wild relatives: An avenue for movement of engineered genes? Crop Science 48:825-840.

Johannessen MM, Andersen BA, Jorgensen RB. 2006. Competition affects geneflow from oilseed rape ([female]) to *Brassica rapa* ([male]). Heredity 96:360-367.

Johnson LMK, Galloway LF. 2008. From horticultural plantings into wild populations: movement of pollen and genes in *Lobelia cardinalis*. Plant Ecology 197:55-67.

Jordan J, Butler D, Henzell B, Drenth J, McIntyre L. 2004. Diversification of Australian sorghum using wild relatives, new directions for a diverse planet. Proceedings of the 4th International Crop Science Congress (http://www.cropscience.org.au/icsc2004/poster/3/3/1/986_jordand.htm).

Kamala V, Bramel PJ, Sivaramakrishnan S, Chandra S, Kannan S, Harikrishna S, Rao DM. 2006. Genetic and phenotypic diversity in downy-mildew-resistant sorghum (*Sorghum bicolor* (L.) Moench) germplasm. Genetic Resources and Crop Evolution 53:1243-1253.

Kamala V, Singh SD, Bramel PJ, Rao DM. 2002. Sources of resistance to downy mildew in wild and weedy sorghums. Crop Science 42:1357-1360.

Kameyama Y, Isagi Y, Naito K, Nakagoshi N. 2000. Microsatellite analysis of pollen flow in *Rhododendron metternichii* var. *hondoense*. Ecological Research 15:263-269.

Kingman JFC. 1982. On the genealogy of large populations. Journal of Applied Probability 19:27-43.

Komolong B, Chakraborty S, Ryley M, Yates D. 2002. Identity and genetic diversity of the sorghum ergot pathogen in Australia. Australian Journal of Agriculture Research 53:621-628.

Kong L, Dong J, Hart GE. 2000. Characteristics, linkage-map positions, and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple-sequence repeats (SSRs). Theoretical and Applied Genetics 101:438-448.

Konishi T, Ohnishi O. 2007. Close genetic relationship between cultivated and natural populations of common buckwheat in the Sanjiang area is not due to recent geneflow between them: An analysis using microsatellite markers. Genes and Genetic Systems 82:53-64.

Krishnaveni S, Jeoung J, Muthukrishnan S, Liang G. 2000. Transgenic sorghum plants constitutively expressing a rice chitinase gene show improved resistance to stalk rot. Journal of Genetic Breeding 55:151-158.

Kuhner MK, Yamato J, Felsenstein J. 1998. Maximum likelihood estimation of population growth rates based on the coalescent. Genetics 149:429-434.

Kuroda Y, Kaga A, Tomooka N, Vaughan DA. 2006. Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Molecular Ecology 15:959-974.

Levin DA, Francisco-Ortega J, Jansen RK. 1996. Hybridization and the extinction of rare plant species. Conservation Biology 10:10-16.

Levin DA, Kerster HW. 1974. Geneflow in seeds plants. Evolutionary Biology 7:139-220.

Loiselle BA, Sork VL, Nason J, Graham C. 1995. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). American Journal of Botany 82:1420-1425.

Mann JA, Kimber C, Miller FR. 1983. The origin and early cultivation of sorghums in Africa. Bulletin 1454, Texas Agriculture Experiment Station, College Station, TX, USA.

Mariac C, Luong V, Kapran I, Mamadou A, Sagnard F, Deu M, Chantereau J, Gerard B, Ndjeunga J, Bezanon G, Pham JL, Vigouroux Y. 2006. Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. Theoretical and Applied Genetics 114:49-58.

Marr DL, Leebens-Mack J, Elms L, Pellmyr O. 2000. Pollen dispersal in *Yucca filamentosa* (Agavaceae): The paradox of self-pollination behavior by *Tegeticula yuccasella* (Prodoxidae). American Journal of Botany 87:670-677.

Martinez-Castillo J, Zizumbo-Villarreal D, Gepts P, Colunga-GarciaMarin P. 2007. Geneflow and genetic structure in the wild-weedy-domesticated complex of *Phaseolus lunatus* L. in its mesoamerican center of domestication and diversity. Crop Science 47:58-66.

Meagher TR. 1986. Analysis of paternity within a natural population of *Chamaelirium luteum*. 1. Identification of most-likely male parents. The American Naturalist 128:199-215.

Menkir A, Goldsbrough P, Ejeta G. 1997. RAPD based assessment of genetic diversity in cultivated races of sorghum. Crop Science 37:564-569.

Menz MA, Klein RR, Unruh NC, Rooney WL, Klein PE, Mullet JE. 2004. Genetic diversity of public inbreds of sorghum determined by mapped AFLP and SSR markers. Crop Science 44:1236-1244.

Montes-Hernandez S, Eguiarte LE. 2002. Genetic structure and indirect estimates of geneflow in three taxa of *Cucurbita* (Cucurbitaceae) in western Mexico. American Journal of Botany 89:1156-1163.

Mooney HA, Cleland EE. 2001. The evolutionary impact of invasive species. Proceeding of the National Academy of Sciences of the USA (PNAS) 98:5446-5451.

Moran PAP. 1950. Notes on continuous stochastic phenomena. Biometrika 37:17-23.

Morden CW, Doebley JF, Schertz KF. 1989. Allozyme variation in old world races of *Sorghum bicolor* (Poaceae). American Journal of Botany 76:247-255.

Morden CW, Doebley JF, Schertz KF. 1990. Allozyme variation among the spontaneous species of *Sorghum* section Sorghum (Poaceae). Theoretical and Applied Genetics 80:296-304.

Morrell PL, Williams-Coplin TD, Lattu AL, Bowers JE, Chandler JM, Patterson AH. 2005. Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. Molecular Ecology 14:2143-2154.

Murdock GP. 1959. Staple subsistence crops of Africa. Geographical Review 50:521-540.

Mutegi E, Sagnard F, Muraya M, Kanyenji B, Rono B, Mwongera C, Marangu C, Kamau J, Parzies H, de Villiers S, Semagn K, Traore PS, Labuschagne M. 2009. Ecogeographical distribution of wild, weedy and cultivated *Sorghum bicolor* (L.) Moench in Kenya: implications for conservation and crop-to-wild gene flow. Genetic Resources and Crop Evolution (DOI 10.1007/s10722-009-9466-7)

Myers ER, Chung MY, Chung MG. 2007. Genetic diversity and spatial genetic structure of *Pinus strobus* (Pinaceae) across an island landscape inferred from allozyme and cpDNA markers. Plant Systematics and Evolution 264:15-30.

Neal D. 2004. Introduction to population biology. Cambridge University Press Cambridge, UK, pp. 393.

Nei M. 1973. Analysis of gene diversity in subdivided populations. Proceeding of the National Academy of Sciences of the USA (PNAS) 70:3321-3323.

Niegel JE. 1997. A Comparison of alternative strategies for estimating geneflow from genetic markers. Annual Review of Ecology and Systematics 28:105-128.

Nielsen R, Mattila DK, Clapham PJ, Palsboll PJ. 2001. Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic Humpback whale. Genetics 157:1673-1682.

Ollitrault P, Noyer J, Chantereau J, Glaszmann JC. 1997. Structure génétique et dynamique des variétés traditionnelles de sorgho au Burkina Faso.Gestion des ressources génétiques des plantes en Afrique des Savanes. Chirat, St-Just-la-Pendue. pp. 240.

Osborne JL, Williams IH. 2001. Site constancy of bumble bees in an experimentally patchy habitat. Agriculture Ecosystems and Environment 83:129-141.

Papa R, Gepts P. 2003. Asymmetry of geneflow and differential geographical structure of molecular diversity in wild and domesticated common bean (*Phaseolus vulgaris* L.) from Mesoamerica. Theoretical and Applied Genetics 106:239-250.

Parker PG, Allison AS, Schug MD, Booton GC, Fuerst PA. 1998. What molecules can tell us about populations: choosing and using a molecular marker. Ecology 79:361-382.

Pearse DE, Crandall KA. 2004. Beyond F-ST: Analysis of population genetic data for conservation. Conservation Genetics 5:585-602.

Perumal R, Krishnaramanujam R, Menz MA, Katile S, Dahlberg J, Magill CW, Rooney WL. 2007. Genetic diversity among sorghum races and working groups based on AFLPs and SSRs. Crop Science 47:1375-1383.

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

Rao Kameswara N, Reddy LJ, Bramel PJ. 2003. Potential of wild species for genetic enhancement of some semi-arid food crops. Genetic Resources and Crop Evolution 50:707-721.

Ritland K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research 67:175-185.

Rousset F. 2000. Genetic differentiation between individuals. Journal of Evolutionary Biology 13:58-62.

Sagnard F, Barnaud A, Deu M, Barro C, Luce C, Billot C, Rami JF, Bouchet S, Dembele D, Pomies V, Calatayud C, Rivallan R, Joly H, Brocke KV, Toure A, Chantereau J, Bezancon G, Vaksmann M. 2008. Multi-scale analysis of sorghum genetic diversity: Understanding the evolutionary processes for *in situ* conservation. Cahiers Agricultures 17:114-121.

Sambatti JBM, Rice KJ. 2006. Local adaptation, patterns of selection, and geneflow in the Californian serpentine sunflower (*Helianthus exilis*). Evolution 60:696-710.

Schmidt M, Bothma G. 2006. Risk assessment for transgenic sorghum in Africa: crop-to-crop geneflow in *Sorghum bicolor* (L.) Moench. Crop Science 46:790-798.

Sharma HC, Franzmann BA. 2001. Host-plant preference and oviposition response of the sorghum midge, towards wild relatives of sorghum. Journal of Applied Entomology 125:109-114.

Shimono A, Ueno S, Tsumura Y, Washitani I. 2006. Spatial genetic structure links between soil seed banks and above-ground populations of *Primula modesta* in subalpine grassland. Journal of Ecology 94:77-86.

Slatkin M. 1985a. Geneflow in natural populations. Annual Review of Ecology and Systematics 16:393-430.

Slatkin M. 1985b. Rare alleles as indicators of geneflow. Evolution 39:53-65.

Slatkin M. 1987. Geneflow and the geographic structure of natural populations. Science 236:787-792.

Slatkin M, Barton NH. 1989. A comparison of three indirect methods for estimating average levels of geneflow. Evolution 43:1349-1368.

Slavov GT, Howe GT, Gyaourova AV, Birkes DS, Adams WT. 2005. Estimating pollen flow using SSR markers and paternity exclusion: accounting for mistyping. Molecular Ecology 14:3109-3121.

Smouse PE, Dyer RJ, Westfall RD, Sork VL. 2001. Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. Evolution 55:260-271.

Smouse PE, Peakall R. 1999. Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. Heredity 82:561-573.

Snow AA, Moran-Palma P. 1997. Commercialization of transgenic plants: potential ecological risks. BioScience 47:86-96.

Sokal RR, Oden NL. 1978a. Spatial autocorrelation in biology 1. Methodology. Biological Journal of the Linnean Society 10:199-228.

Sokal RR, Oden NL. 1978b. Spatial autocorrelation in biology 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10:249.

Sork VL, Nason J, Campbell DR, Fernandez JF. 1999. Landscape approaches to historical and contemporary geneflow in plants. Trends in Ecology and Evolution 14:219-224.

Stefenon VM, Gailing O, Finkeldey R. 2008. The role of geneflow in shaping genetic structure of the subtropical cinifer species *Araucaria angustifolia.* Plant Biology 10:356-364.

Strobeck C. 1983. Estimation of the neutral mutation rate in a finite population from DNA sequence data. Theoretical Population Biology 24:160-172.

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J, Marx D, Bothma G, Ejeta G. 2008. The potential for crop-to-wild geneflow in sorghum in Ethiopia and Niger: A geographic survey. Crop Science 48:1425-1431.

Thies JE, Devare MH. 2007. An ecological assessment of transgenic crops. Journal of Development Studies 43:97-129.

Thompson EA, Meagher TR. 1987. Parental and sib likelihoods in genealogy reconstruction. Biometrics 43:585-600.

van Treuren R, Goossens PJ, Sevcikova M. 2006. Variation in effective pollination rates in relation to the spatial and temporal distribution of pollen release in rejuvenated perennial ryegrass. Euphytica 147:367-382.

Warwick SI, Stewart CN. 2005. Crops come from wild plants - How domestication, transgenes, and linkage together shape ferality. In: Gressel J (Ed.), Crop Ferality and Volunteerism. CRC Press, Boca Raton, Florida. pp. 30.

Wei W, Qian YQ, Ma KP. 1999. Geneflow between transgenic crops and their wild related species. Acta Botanica Sinica 41:343-348.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Weir BS, Hill WG. 2002. Estimating F-Statistics. Annual Review of Genetics 36:721-750.

Whitlock MC, Mccauley DE. 1999. Indirect measures of geneflow and migration: FST [ne] 1/ (4Nm + 1). Heredity 82:117-125.

Wright S. 1951. The genetical structure of populations. Annals of Eugenics 15:323-354.

Wright S. 1978. Evolution and the genetics of populations: Variability within and among natural populations. University of Chicago Press Chicago, pp. 590.

Zhao Z. 2008. The Africa biofortified sorghum project - Applying biotechnology to develop nutritionally improved sorghum for Africa. Proceedings of the 11th IAPTC&B Congress, August 31-18, 2006 Beijing, China. p. 273-277.

Zhao Z, Cai T, Tagliani L, Wang N, Pang H, Rudert M, Schroeder S, Hondred D, Pierce D. 2000. *Agrobacterium*-mediated sorghum transformation. Plant Molecular Biology 44:789-798.

Zhu H, Muthukrishnan S, Krishnaveni S, Wilde G, Jeoung J, Liang G. 1998. Biolistic transformation of sorghum using a rice chitinase gene. Journal of Genetic Breeding 52:243-252.

Chapter 3

Compared phylogeography of wild and cultivated sorghum in Kenya using microsatellites

3.1 Abstract

This study examined the diversity, structure, geneflow and evolutionary relationships within and among cultivated and wild genotypes of *S. bicolor* in Kenya. A total of 439 individuals comprising 329 cultivated and 110 wild sorghum genotypes were analysed using 24 microsatellite markers. Two hundred and ninety five alleles were obtained across all loci and individuals, with 257 alleles being detected in the cultivated sorghum gene pool and 238 alleles in the wild sorghum gene pool. The cultivated sorghum gene pool harboured significantly less genetic diversity than the wild sorghum gene pool, a suggestion that sorghum domestication was accompanied by a genetic bottleneck. The overall genetic differentiation between the cultivated and wild sorghum gene pools as estimated by principle coordinate analysis, neighbour-joining, fixation index, analysis of molecular variance and Bayesian model-based analyses was low, although levels of divergence varied in different sorghum cultivation regions of the country. The close genetic proximity has probably arisen primarily due to historical and contemporary geneflow between the two congeners. Variability in the level of genetic proximity between cultivated and wild sorghum among growing regions may reflect differences in the extent of the geneflow probably due to differences in farmer practices. Spatial autocorrelation analysis revealed a strong spatial genetic structure in both cultivated and wild sorghum at country scale, with individuals within a radius of about 180 km showing a close genetic relationship with one another. In cultivated sorghum significant spatial genetic structure was exhibited in eastern/central, western/Nyanza and north-eastern, but not in the coastal and Turkana regions. Finally, the genetic distance between pairs of cultivated and wild sorghum individuals increased significantly with logarithmic distance, a pattern typical of the isolation-by-distance model which suggested that crop-to-wild geneflow in sorghum can be predicted by the geographic distance between individuals.

3.2 Introduction

Phylogeography (Avise *et al.* 1987) is a field of study concerned with the principles and processes governing the geographic distribution of genetic variation, especially within and among conspecific populations. In its purest form, empirical phylogeographic analyses deal with the geographic distributions within and among alleles whose phylogenetic relationships are known or can be estimated. In the broad sense, however, phylogeographic analyses can encompass empirical or theoretical treatments that consider the evolutionary relatedness aspects of the spatial distribution of any genetic trait(s), whether morphological, behavioural, molecular or otherwise (Avise 1998; 2000). A phylogeographic analysis affords insight into the historical, contemporary and biotic factors that influence the genetic structuring of populations.

Plants demonstrate a wide range of diversity in their morphology, adaptation and ecology, the product of many years of evolutionary divergence and diversification. Characterising the extent and partitioning of this diversity across populations and geographic regions of a target taxon, coupled with an understanding of the mechanisms through which it arises, has been of interest to plant genetic resources conservation and crop improvement programmes alike (Frankel and Hawkes 1975; Frankel *et al.* 1995). Information about genetic relationships within and among populations of target taxa can help conservation and breeding programme managers to focus effort and resources on truly distinctive groups. Furthermore, geneflow, including that among diverged taxa, is recognised as an evolutionary factor that strongly influences intraspecific population genetic structure (Ellstrand 1992). With the advent of GM crops there are mounting concerns that transgenes will escape to sexually compatible wild and weedy relatives via geneflow, potentially leading to increased invasiveness, weediness, genetic erosion and in extreme cases extinction of populations (Ellstrand 1992; Snow and Moran-Palma 1997; Bhatia and Mitra 2003; Conner *et al.* 2003; Haygood *et al.* 2003; Cleveland and Soleri 2005; Thies and Devare 2007; Auer 2008; Chandler and Dunwell 2008). Estimation of historical crop-to-wild geneflow through approaches such as phylogeographic analyses is therefore of great significance for the development and implementation of biosafety guidelines.

Sorghum is one of the world's most important cereals. It is especially important for food security in arid and semi-arid lands of Africa and Asia. Out of the total land area grown

with sorghum globally in the year 2007 for example, Africa's contribution was in excess of 60% (FAO 2008). Besides its use as a cereal crop, in Africa, sorghum is extensively used for fodder, construction material, brooms, syrup and beer. In Kenya, sorghum is grown in all but one administrative province. It is an important food security crop and dietary staple in the country's arid and semi-arid lands. A wide diversity of sorghum landraces is cultivated under equally diverse agro-climatic conditions and practices by subsistence farmers in different communities of Kenya. Moreover, morphologically and geographically diverse wild relatives of domesticated sorghum in the primary and tertiary gene pools are known to occur in the country (Clayton and Renvoize 1982). Wild relatives of sorghum are recognised as broad genetic base reservoirs and potential sources for resistance and adaptation traits in breeding programmes (Gurney *et al.* 2002; Kamala *et al.* 2002; Reed *et al.* 2002; Rao Kameswara *et al.* 2003; Rich *et al.* 2004).

Domestication of sorghum is thought to have commenced around 4000 - 3000 BC in the region that corresponds today to southern Sudan and Ethiopia (De Wet 1978; Dogget 1988). Cultivated sorghum (*S. bicolor* ssp. *bicolor*) is taxonomically conspecific with its wild progenitor (*S. bicolor* ssp. *verticilliflorum*) and the stabilised weedy derivative of their hybridisation (*S. bicolor* ssp. *drummondii*). Harlan and De Wet (1972) divided cultivated sorghum into five basic races (bicolor, caudatum, durra, kafir and guinea) and ten intermediate races, on the basis of panicle and spikelet morphology. Based on inflorescence structure, plant habit and geographic distribution, wild sorghum have on the other hand been classified into four ecogeographic races or ecotypes: aethiopicum, arundinaceum, verticilliflorum and virgatum (De Wet *et al.* 1970; De Wet 1978; Dogget 1988). All subspecies of *S. bicolor* are inter-fertile under sympatric conditions, leading to a continuum of wild-weedy-domesticate complex forms that co-occur within many sorghum growing parts of Africa (Dogget and Majisu 1968; Dogget 1988; Tesso *et al.* 2008; Mutegi *et al.* 2009). Moreover, cultivated and wild sorghum occupy diverse ecological landscapes and have over the years been subjected to diverse biotic and abiotic selection pressures across their geographic range. Wide genetic diversity is therefore anticipated in the landraces of cultivated sorghum and their wild-weedy relatives in Africa.

Early studies relied on morphological descriptors and numeric taxonomic approaches to lay a foundation for genetic relationships between cultivated sorghum and its wild-weedy

congeners (Liang and Casady 1966; De Wet and Huckabay 1967; Murty *et al.* 1967). Perhaps the most comprehensive of these is the work done by De Wet and Huckabay (1967), who used type specimens and original collections of cultivated and wild sorghum respectively, to show a clear genetic separation between the two gene pools, with the weedy forms occupying a somewhat intermediate position. Levels and patterns of diversity within and among cultivated and wild sorghum gene pools have since then been characterised using either allozyme or DNA markers leading to increased insight into their evolutionary and genetic relationships (Morden *et al.* 1990; Aldrich and Doebley 1992; Aldrich *et al.* 1992; Cui *et al.* 1995; Deu *et al.* 1995; Casa *et al.* 2005). Four major observations are worth noting from these previous studies: (i) there is low to moderate genetic differentiation among cultivated and wild sorghum gene pools, (ii) portions of the wild gene pools most genetically similar to cultivars originate in central-north-eastern Africa, (iii) genetic diversity is greater in the wild gene pool than in the cultivated one, with the diversity of the domesticate encompassed in the wild and (iv) introgression among cultivated and wild gene pool takes place, but the frequency is thought to be low enough to allow distinct genetic constitutions. The first three of these observations are in line with the views that cultivated sorghum arose from *S. bicolor* ssp. *verticilliflorum* and that domestication of sorghum most likely took place in the eastern-central African region (Harlan and Stemler 1976; Dogget 1988).

With regard to introgression between cultivated and wild sorghum, little is known on the extent, patterns and direction. At the same time, previous results were all obtained mostly using genebank collections and may need validation using exhaustive samples obtained *in situ* at different spatial scales, especially in Africa. Attempts have been made to use *in situ* collected samples but such studies have been limited to separate investigations of genetic diversity and structure in either cultivated sorghum (Dje *et al.* 1998; 1999; Ayana *et al.* 2000b; Ayana *et al.* 2001; Ghebru *et al.* 2002; Barnaud *et al.* 2007; Deu *et al.* 2008; Sagnard *et al.* 2008;) or its closest wild relatives (Ayana *et al.* 2000a). As well remarked by Dogget (1988) ‘carefully designed sampling of geographical area by geographical area, using appropriate descriptors and analytical methods, will shed a lot more light on the history of the sorghum crop and also on that of wild forms’.

This study applied microsatellite markers to analyse cultivated sorghum and its closest wild relatives sampled at a national scale in Kenya, with a view of elucidating patterns of

diversity within and among the two congeners and shedding more light on their genetic and evolutionary relationships. Specifically this work sought to answer the following questions:

(i) What is the comparative extent of diversity within cultivated and wild gene pools of sorghum at country and regional scale level?

(ii) Are cultivated and wild sorghum gene pools genetically differentiated?

(iii) How is genetic diversity in cultivated and wild sorghum gene pools partitioned among growing regions and agro-climatic zones of Kenya?

(iv) Is diversity in cultivated and wild sorghum gene pools spatially structured?

(v) Has geneflow between cultivated and wild sorghum taken place in the past?

3.3 Materials and methods

3.3.1 Material collection

Cultivated and wild sorghum seed samples were collected in farmer fields in the crop's four main growing areas of Kenya: (i) Turkana, which is situated in the northern parts of the Rift Valley bordering Sudan and Ethiopia; (ii) western/Nyanza region covering Kisii Highlands and the lowland parts around Lake Victoria; (iii) eastern/central region covering the Highlands east of Mt. Kenya and the much drier lowlands of the larger Meru, Kitui and Machakos administrative Districts; and (iv) coastal areas of the country including Taita Hills and the adjacent areas as well as the farming systems in the Indian Ocean hinterlands (Figure 3.1). Three collection trips were undertaken between June 2006 and July 2007 in order to capture differences in cropping seasons amongst the four growing regions. A sample sheet was used to record the passport data associated with each collected sample. Farmer knowledge on cultivated varieties as well as on wild and weedy sorghum distribution, ecology and dynamics was recorded. The geographic coordinates and elevation data associated with each collection point were recorded using a handheld global positioning system (GPS) (eTrex Summit HC, Garmin). Additional georeferenced samples were acquired from the National Genebank of Kenya to cover the north-eastern region of the country bordering Ethiopia and parts of central Rift Valley where sorghum is grown but which were not covered in the collection trips. In total, 439 genotypes comprising 110 wild and 329 cultivated sorghum varieties were sampled. Overall, the samples were representative of Kenya's sorghum growing agro-climatic and ethno-linguistic diversity. The highest number of samples for cultivated sorghum was

collected in eastern/central (90), followed in a decreasing order by western/Nyanza (72), Rift Valley (55), Turkana (42), Coast (36) and north-eastern (34) regions. In the wild sorghum, the highest number of samples was sampled in eastern/central (41), followed in a decreasing order by Coast (39), Turkana (17) and western/Nyanza (13) regions.

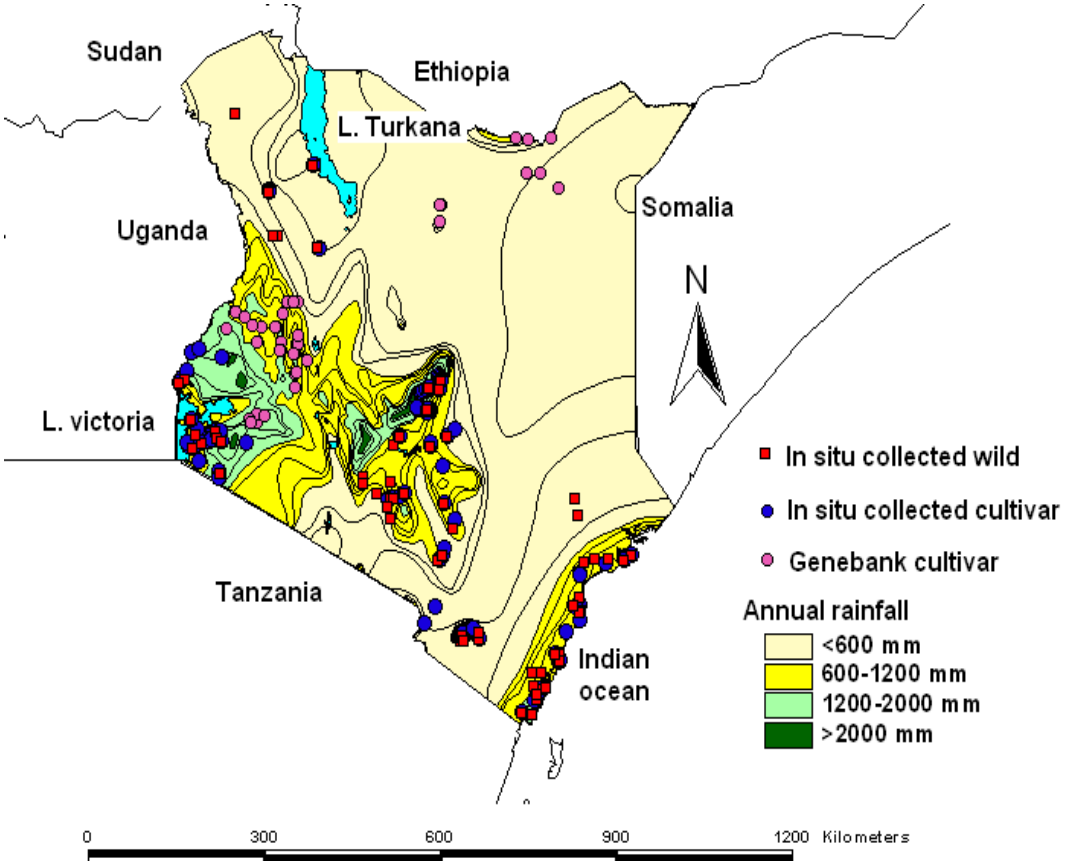

Figure 3.1 Map of Kenya showing the sources of collection for cultivated and wild sorghum and the associated annual rainfall classes.

3.3.2 DNA isolation

At least five seeds from each sample of cultivated and wild sorghum were grown for two weeks in the laboratory in potted plastic trays before DNA isolation. To break seed dormancy in wild sorghum, glumes were removed using a scalpel blade and seeds soaked overnight in water at 35°C. To ensure sufficient light throughout the germination and growth period, trays were placed on a bench next to a glass widow. Plants were watered regularly to ensure normal growth. Only one seedling per sample was used to extract total genomic DNA from leaves using a modified version of the high throughput Cetyl Trimethyl Ammonium Bromide (CTAB) method described by Mace *et al.* (2003) and a

GenoGrinder (Geno/Grinder 2000, Spex Certiprep USA). The approach of sampling one individual per sample was necessitated by the need to maximize the number of different accessions of cultivated and wild sorghum to be genotyped countrywide. The approach has proved sufficient to detect large-scale inter-sample evolutionary trends in crops and/or their wild relatives (e.g. Matsuoka *et al.* 2002, Fukunaga *et al.* 2005, Mariac *et al.* 2006; Deu *et al.* 2008) provided that the number of loci is sufficient. Sufficient CTAB buffer [3% (w/v) CTAB, 1.4 M NaCl, 20 mM (hydroxymethyl) aminomethane hydrochloride (Tris-HCl), 20 mM ethylenediaminetetraacetic acid (EDTA), pH 8.0, 0.17% (v/v) β-mercaptoethanol], corresponding to 450 µl per sample, was dispensed into a glass bottle and incubated at 65^{o}C in a water bath. Freshly harvested leaves (3 - 5 cm) were placed in a 96-well deep-plate, containing two 4 mm steel grinding balls. To each sample, 450 µl warm CTAB buffer was added before the plate was sealed with a fitted silicone mat. The plate was loaded into the GenoGrinder for grinding at a setting of 500 strokes/min for 10 - 15 min. After grinding, samples were incubated for 10 min at 65^{o}C in a water bath with occasional mixing.

The plate was removed from the water bath and 450 µl of chloroform-isoamylalcohol (24:1) added to each sample. The plate was gently inverted twice to mix the samples before centrifuging at 10000 rpm for 10 min. A fixed volume of 400 µl of the upper aqueous layer was for each sample transferred to freshly labelled 1.5 ml Eppendorf tubes. In order to precipitate the DNA, 280 µl of isopropanol (stored at -20°C) was added to each sample and the tubes inverted twice to mix. The tubes were subsequently centrifuged at 10000 rpm for 10 min, before discarding the supernatant and air-drying the DNA pellet for 30 min. To wash the DNA pellets, 500 µl 70% (v/v) ethanol was added into the tubes and centrifuged at 10000 rpm for 3 min. The supernatant was discarded and the DNA pellet air-dried for 30 min, following which 100 µl of low-salt TE [10 mM Tris-HCl, 0.1 mM EDTA (pH 8.0)] was added to re-suspend the DNA. A volume of 1 µl RNAse (10 mg/ml) was added to each sample before incubating in an oven at 37°C for 1 hour.

Following extraction, the quality of the DNA was checked by loading 3 µl of each sample on ethidium bromide stained 0.8% (w/v) agarose gel and electrophoresis at 100 V for 1 hour. The quality of the DNA was determined by visualising the gel under ultraviolet (UV) light. The concentration of the DNA was determined using the NanoDrop 3300 spectrophotometer (Applied Biosystems), by loading 1 µl of each sample. Samples with

low DNA concentration (less than 10 ng/µl) were re-extracted. Upon quantification, all samples were normalised to a final concentration of 10 ng/µl by adding variable volumes of double distilled water to a final volume of 100 µl.

3.3.3 PCR amplification and genotyping

Thirty SSR markers (Appendix 1) were analysed using the M13-tailed primer method (Schuelke 2000) to label amplicons for visualisation on an ABI 3730 (Applied Biosystems) capillary sequencer. Forward primers were 5'-tailed with a 19-base pair (bp) M13 universal sequence, 5'-CACGAGCTTGTAAAACGACXXXXXXXXXXXXX-3', where the X's denote microsatellite-specific primer sequences (See Appendix 1 for details).

Polymerase chain reaction (PCR) was performed in 10 µl reaction volumes, containing 5 µl of template DNA, 0.2 units of Amplitaq Gold *Taq* DNA polymerase (Applied Biosystems), 1X PCR buffer (10 mM Tris-HCl pH 8.3, 50 mM KCl, 1.5 mM $MgSO_4$), 0.16 mM dNTPs, 2 µM sequence-specific reverse primer, 0.04 µM 5'-M13 tailed sequence-specific forward primer and 0.16 µM 5'-fluorescently labelled M13 universal sequence primer in a GeneAmp PCR system 9700 thermocycler (Applied Biosystems). The M13 universal sequence primer was 5'-tagged with VIC, NED, FAM or PET fluorescent dyes in order to facilitate post-PCR multiplexing. This allowed for the co-loading of the 30 primer products in ten sets of three primer products each.

The PCR programme was as described by Folkertsma *et al.* (2005): initial denaturation at 94°C for 15 min followed by 10 cycles of 94°C for 15 s, annealing for 20 s at touchdown temperatures declining from 61 to 51°C and extension at 72°C for 30 s; this was followed by 35 cycles of 94°C for 10 s, 54°C for 20 s and 72°C for 30 s; and a final extension step of 20 min at 72°C. Reliability of amplification for each primer set was confirmed by resolving 3 µl of each of the PCR products on a 2% (w/v) agarose gel at 100 V for 45 min. Subsequently, 3 µl of each of the PCR products was denatured at 94^{o}C in 8 µl of Hi-Di formamide with 0.13 µl of GeneScan 500 LIZ internal size standard (Applied Biosystems). PCR products were subjected to capillary electrophoresis in an ABI Prism 3730 DNA Analyser (Applied Biosystems) for allele detection. Allelic data was scored using the software Genemapper 3.7 (Applied Biosystems). A sorghum standard positive control, BTX623 was amplified and genotyped alongside each sample plate to ascertain the reproducibility of allele scoring. Amplitaq Gold DNA polymerase, like many other

DNA polymerases is known to catalyse the addition of a single nucleotide (usually Adenosine) to the 3'-end of a fraction of the amplicons potentially leading to inconsistent allele calls. The software AlleloBin (Prasanth *et al.* 2006) was therefore used to correct single bp allele call inconsistence in the data for each run.

3.3.4 Data analysis

3.3.4.1 Extent of genetic diversity

Standard parameters of genetic diversity, including total number of alleles (A^t), number of rare alleles (A^r, alleles with a frequency < 5% per group), private alleles (A^p, alleles unique to a group), observed heterozygosity (H_o) and unbiased expected heterozygosity or gene diversity (H_e) were computed for cultivated and wild sorghum using the software GENETIX 4.05 (Belkhir *et al.* 2004). The genetic diversity parameters were calculated and compared within and among cultivated and wild sorghum at four levels: country scale, region of origin, agro-climatic zone and altitudinal class. Agro-climatic zones were defined according to the agro-climatic zone map of Kenya (Sombroek *et al.* 1982) which recognises seven zones using a moisture index based on annual rainfall expressed as a percentage of potential evaporation (Appendix 2). The seven agro-climatic zones are: I (humid), II (sub-humid), III (semi-humid), IV (semi-humid to semi-arid), V (semi-arid), VI (arid) and VII (very arid). The GPS based altitude value associated with each sample was used to define four altitudinal classes: "0 - 500 m", "from 500 - 1000 m", "from 1000 - 1500 m" and "greater than 1500 m". Since the observed number of alleles in a sample is highly dependent on sample size, the programmes FSTAT (Goudet 2002) was employed to compute the mean allelic richness across all loci (R_s) for each defined level of genetic structure. In addition the software HP-RARE 1.2 (Kalinowski 2005) was used to compute and compare the private allelic richness ($\prod^{S}_{taxon}$) in the cultivated and wild sorghum gene pools. The two programmes implement the rarefaction statistical method first used by Hulbert (1971) to estimate species diversity. The method allows for unbiased comparisons among populations of unequal sample sizes by calculating a standardised estimate of allelic richness for a fixed sample size. The principle is to estimate the expected number of alleles in a sub-sample of 2n genes given that 2N (N>n) genes have been sampled. In FSTAT, n is fixed for the smallest number of individuals genotyped per locus in a sample to calculate overall allelic richness as:

$$Rs = \sum_{i=1}^{n} \left[\frac{\left(\frac{2N - Ni}{2n} \right)}{\left(\frac{2N}{2n} \right)} \right]$$

where N_i is the number of alleles of type i among the 2N genes.

The programme HP-RARE performs rarefaction on private alleles using a hierarchical sampling design to estimate the allelic richness among populations of balanced sample size. The private allelic richness of a population, Π^{S}_{taxon}, is the expected number of private alleles in a sample of size S taken from the population. It is estimated by:

$$\prod^{S}_{taxon} \cong \sum_{i=1}^{m} u^{S}_{i,taxon}$$

Where *taxon* is either a population or a set of populations that have been placed in the same category, $u^{S}_{i,taxon}$ is the probability that the i^{th} allele will only be found in the subsample from the taxon indicated. A balanced sample size of 82 individuals (or 164 genes) for both cultivated and wild sorghum was used in the computation. This was equivalent to the smallest number of individuals genotyped across all loci.

Differences in $R_{s,}$ $\prod^{S}_{taxon}$ and H_e among the various levels were assessed for significance using Wilcoxon's signed-rank test as implemented in the software GenStat (VSN International Ltd. 2007).

3.3.4.2 Genetic relationships within and among cultivated and wild sorghum

To investigate the genetic relationships among cultivated and wild sorghum populations at national and region of origin scale, the neighbour-joining (NJ) cluster analysis and principle coordinate analysis (PCoA) algorithms implemented in the software DARwin 5.0 (Perrier *et al.* 2003) were used.

In each case, a genetic dissimilarity matrix was first computed from multilocus allelic data for pairs of individual plants according to the simple matching procedure:

$$d_{ij} = 1 - \frac{1}{L} \sum_{I=1}^{L} \frac{M_I}{\pi}$$

where d_{ij} is the dissimilarity among genotypes i and j, L is number of loci, π is the ploidy of the taxa being investigated and M_I is the number of matching alleles for locus I. The pairwise deletion option was chosen to ensure that dissimilarity calculations were done only for pairs of genotypes where allelic scores were obtained for at least 70% of all loci. Eighteen samples which exhibited too many missing data points were eliminated from the final dissimilarity calculation. The dissimilarity matrix for the remaining 421 samples was subsequently used as input in the NJ and PCoA procedures.

3.3.4.3 Cultivated and wild sorghum genetic structure

In order to explore the genetic differentiation (acquisition of different allele frequencies) within and among cultivated and wild sorghum genepools at various spatial scales, four complimentary approaches were used: F-statistics, Bayesian model-based clustering, analysis of molecular variance (AMOVA) and spatial autocorrelation.

3.3.4.3.1 F-statistics

F-statistics (Wright 1951; 1978) and their unbiased estimators (Weir and Cockerham 1984) are commonly used to estimate genetic differentiation under a hierarchical population structure model. A population is said to have a hierarchical structure if it can be subdivided into sub-populations that can be grouped into progressively inclusive levels in which, at each group, the next lower levels are included or nested within the next higher ones (Hartl and Clark 1997). Classically, three F-statistics parameters, namely F_{IS}, F_{ST} and F_{IT} are defined for three hierarchical levels of a population under the assumptions of random union of gametes according to the HWE. Due to the departure from random mating in many natural populations, the parameters F_{IS} and F_{IT} are measures of the level of inbreeding within sub-populations and the overall population, respectively, where positive values indicate a deficiency of heterozygotes and negative values indicate an excess of heterozygotes. In turn the parameter F_{ST} uses allelic frequency information to measure identity of individuals within sub-populations as compared to individuals from other sub-populations within the total population and thus measures the degree of genetic

differentiation among sub-populations. The value for F_{ST} ranges from 0 (indicating no genetic differentiation among sub-populations) to a theoretical maximum of 1 (indicating that sub-populations are fixed for different alleles). In practice though, the observed F_{ST} value is usually much lower than 1, even in highly differentiated populations (Hartl and Clark 1997). The three F-statistics parameters are related in the following expression according to Wright (1951):

$$(1 - F_{IT}) = (1 - F_{IS}) (1 - F_{ST})$$

Under the assumption of similar (and low) mutation pressure in different populations, the F_{ST} of neutral marker loci is primarily determined by the balance between random genetic drift and geneflow. Consequently, the degree of differentiation among populations at neutral marker loci as estimated by the F_{ST} parameter, indicates the expected degree of population differentiation as a result of the combined effect of genetic drift and geneflow (Hartl and Clark 1997).

The software GENETIX 4.04 was used to compute estimates of the F-statistics within and between cultivated and wild sorghum following the unbiased method of Weir and Cockerham (1984). This method takes into account errors due to sampling alleles within loci and those due to sampling individuals within sub-populations. The significance of the F-statistics parameters was tested using the permutations (10000) procedure.

3.3.4.3.2 Analysis of molecular variance (AMOVA)

Analysis of molecular variance (AMOVA) is a framework designed by Excoffier *et al.* (1992) and later extended by Michalakis and Excoffier (1996) to study hierarchical partitioning of molecular variation within a species by incorporating information of DNA haplotype divergence into analysis of variance format. The input data is derived from a matrix of squared-distances among all pairs of haplotypes. The approach estimates variance components and F-statistics analogs, designated as Φ-statistics, reflecting the correlation of haplotypic diversity at different levels of both genetic and non-genetic hierarchical sub-divisions as determined by each investigation (Excoffier *et al.* 1992). Permutation approaches are used to test the significance of the variance components and Φ-statistics, thus eliminating the normality assumption necessary under conventional analysis of variance (ANOVA) analysis but inappropriate for molecular data.

Investigation of the genetic structure of cultivated and wild sorghum was therefore further undertaken by an AMOVA analysis using the software ARLEQUIN 3.11 (Excoffier *et al.* 2005). The total variance among genotypes was partitioned into variance among populations and within populations, the populations being defined on the basis of sorghum type (cultivated vs. wild), geographic (region of origin), agro-climatic zone and altitudinal range criteria. The significance of partitioning of the genetic variance components was tested using 10000 permutations.

3.3.4.3.3 Model-based cluster analysis

In addition to the analyses discussed above, the Bayesian model-based clustering method implemented in the software STRUCTURE 2.23 (Pritchard *et al.* 2000) was used to explore genetic structure in the samples of cultivated and wild sorghum. The method takes a sample of genotypes and uses the assumption of HWE and LE in loci to simultaneously find the number of populations, K, that best fits the data and the number of individual assignments that minimise HWE and LE in those sub-populations. Thus without any prior information on population sampling design, STRUCTURE provides an estimate of the number of sub-populations. The method is useful in identifying geneflow events since individuals whose genotypes indicate admixture are assigned jointly to two or more populations. The basic admixture model with unlinked loci and uncorrelated allele frequencies was used. The assumed number of populations (K) varied from two to 10, with 10 replicate runs per K, a burn-in period length of 500000 and a post-burning simulation length of $1x10^6$. No *a priori* population information was used. Due to the high computational requirement of STRUCTURE, all analyses were performed using the web resources of the Computational Biology Service Unit (CBSU) from Cornell University (http://cbsuapps.tc.cornell.edu/structure.aspx). The 'true K' for successive K values (K = 2 to K = 10) was estimated using the *ad hoc* statistic ΔK (Evanno *et al.* 2005) which is based on the second order rate of change of P(X|K), the probability of the data with respect to a given K. According to Evanno and co-workers (2005) the modal value of the distribution of ΔK is located at the real K. The modal value was illustrated graphically by plotting the ΔK values against successive K values. Following the identification of the 'true K', the run showing the highest P(X|K) value was considered in drawing a barplot of the proportion of an individual's genome assigned to each of the clusters using the software R (R Development Core Team 2007).

3.3.4.3.4 Spatial autocorrelation analysis

To investigate the spatial structure of genetic diversity in cultivated and wild sorghum at the country scale, spatial autocorrelation analysis was performed as implemented in the software SPAGeDi (Hardy and Vekemans 2002). A geographic distance matrix was generated from the latitude/longitude coordinates associated with each sample using the software Geographic Distance Matrix Generator version 1.2.3 (http://biodiversityinformatics.amnh.org/open_source/gdmg). In each case, 20 distance classes were defined such that there were approximately equal numbers of pairwise comparisons in each class. Within each class, the relative kinship coefficient (r_{ij}) was estimated using the method of Ritland (1996). This index represents the correlation in allelic states between homologous genes and weighs allele distribution by the inverse of allele frequency, thus giving more weight to rare alleles. This way the approach results in lower sampling variance, hence, is powerful for detecting genetic structure (Hardy and Vekemans 2002). The cultivated sorghum samples obtained from the genebank for north-eastern and Rift Valley regions did not have corresponding wild sorghum samples and were therefore eliminated from the analysis to allow for comparison. The significance of the estimated values of kinship coefficient and regression slope was tested by permuting individuals among locations 1000 times. The relationship between genetic relatedness and geographic distance was visualised in correlograms using the software R, with the 95% confidence interval envelope under the null hypothesis of no spatial structure indicated. The same approach was further used to test for spatial genetic structure in cultivated sorghum within the various growing regions. The number of distance classes was set to eight for all regions except in north-eastern and Turkana where, due to a limitation in the number of sampling sites, six and four distance classes were used, respectively. Wild sorghum was not considered for this type of analysis due to limitations in the number of sampling sites within most of the regions.

Further, the method of Rousset (2000) was used to indirectly infer the extent of gene dispersal between cultivated and wild sorghum individuals. This approach is based on the analytical model of isolation-by-distance, which predicts that the genetic distance between individuals (â) (Rousset 2000) increases approximately linearly with the logarithm of spatial distance. Rousset's measure of genetic distance, â , was computed for each pair of individuals using the programme SPAGeDi (Hardy and Vekemans 2002).

Ten distance classes consisting of approximately equal numbers of individual pairwise genetic distance comparisons were defined. Subsequently, the pairwise genetic distance estimates were regressed on the logarithm of spatial distance, providing a regression slope (*blog*) and an estimate of the coefficient of determination (r^2). The significance of the regression slope was tested by a randomisation procedure whereby individuals were permuted among locations 1000 times to asses the distribution of the slope values under the null hypothesis of no correlation between geographic and genetic distance. P-values were estimated as the proportion of this distribution lying higher than the observed slope value. The programme R was used to plot estimates of the pairwise genetic distance between individuals of cultivated and wild sorghum against logarithmic spatial distance, with the regression line shown.

3.4 Results

3.4.1 Seedling growth, DNA yield, PCR amplification and genotyping

Most seeds germinated within three to four days reaching a height of 3 - 5 cm within 14 days (Figure 3.2). The DNA extraction protocol used in this study yielded good quality DNA (Figure 3.3) with the concentration ranging from 2.76 - 1069.3 ng/µl. All except one (xtxp-295) SSR markers successfully amplified DNA in most of the samples. Of all the SSR markers that amplified, data on five markers (msbCIR-223, msbCIR-283, xgap265, xtxp278 and xtxp-289) was eliminated from the subsequent analysis due to unspecific amplification. Both the failure to amplify and the unspecific amplification may reflect non optimal annealing temperatures due to the touch down temperatures used during PCR amplification.

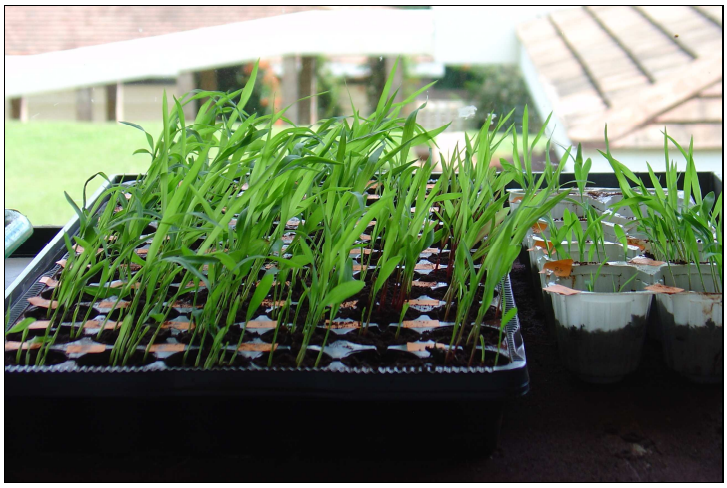

Figure 3.2 Image of two week old cultivated and wild sorghum seedlings growing in potted trays in the laboratory.

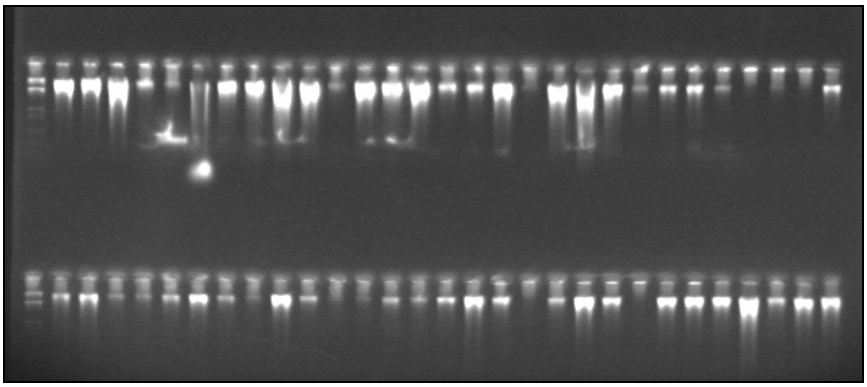

Figure 3.3 Agarose gel electrophoresis image showing the quality of the sorghum DNA extraction. The first well of each row represents a molecular size marker, whereas the rest represent sorghum DNA samples.

3.4.2 Extent of genetic diversity in sorghum

Genetic diversity estimates for the entire sorghum collection and separately for cultivated and wild sorghum are presented in Table 3.1. In total, 295 alleles were detected using 24 SSR markers on a total of 439 cultivated and wild sorghum genotypes, with close to 70% of these being rare alleles (present in less than 5% of the genotypes). The number of alleles detected for the 329 cultivated sorghum genotypes was 257, out of which 173 (67%) were rare and 15 (5%) private alleles. In comparison, 238 alleles were observed for

the 110 wild sorghum genotypes, with 122 (51%) being rare and 13 (5%) private alleles. The cultivated gene pool was observed to harbour lower genetic diversity than the wild gene pool, based on significantly lower mean allelic richness ($P \leq 0.05$), private allelic richness ($P \leq 0.05$) and gene diversity ($P < 0.001$) values.

Table 3.1 Comparative genetic diversity estimates for Kenya's sorghum gene pool

Gene pool	N	A^t	A^r (< 5%)	A^p	R_s	$\prod^{S}_{taxon}$	H_e	H_o
Wild	110	238	122	13	9.68	2.60	0.6872	0.1762
Cultivated	329	257	173	15	8.63	1.22	0.5876	0.1100
Overall	439	295	204	-	12.12	-	0.6304	0.1267

N = Number of samples, A^t = Total number of alleles, A^r = Number of rare alleles, A^p= Number of private alleles, R_s = allelic richness, $\prod^{S}_{taxon}$ = Private allelic richness, H_e = Expected (unbiased) heterozygosity, H_o = Observed heterozygosity.

Table 3.2 presents comparative estimates of the genetic diversity parameters for the cultivated and wild sorghum gene pools at three levels of possible genetic structure: regions of origin, agro-climatic zones and altitudinal class. In cultivated sorghum, the proportion of rare alleles within regions ranged from 34 (31.8%) in the north-eastern to 86 (52.4%) in western/Nyanza regions, whereas in wild sorghum the range was 82 (83.7%) in western/Nyanza to nine (9.5%) in Turkana regions. The number of rare alleles for cultivated sorghum was generally comparable among the various agro-climatic and altitudinal classes. In wild sorghum, no rare alleles were observed in either the more humid agro-climatic zones (sub-humid to humid) or the more elevated altitudinal classes (>1500 m), largely due to the low number of samples analysed. In the most arid agro-climatic zones (arid to very arid) the relative proportion of rare alleles was notably low (5.9%) for the wild sorghum gene pool. In cultivated sorghum, the highest number of private alleles was detected in eastern/central region among geographic regions, in the arid to very arid zone among agro-climatic zones and in the 1000 - 1500 m above sea level among altitudinal zones. In comparison, the highest number of private alleles in wild sorghum was detected at the coast among geographic regions, in the semi-humid to semi-arid zone among agro-climatic zones and in the 0 - 500 m above sea level among altitudinal zones.

Table 3.2 Genetic diversity estimates for the sorghum gene pool at various structuring factors

	Structuring Factor	N	A^t	A^r (< 5%)	A^p	^1R_s	H_e	H_o
Cultivated	**Regions (22)**							
	Turkana	42	95	41	10	3.45	0.3396	0.0965
	western/Nyanza	72	164	86	18	5.34	0.4875	0.0905
	eastern/central	90	169	85	21	5.42	0.5595	0.1742
	Coast	36	141	48	5	5.59	0.5510	0.1200
	Rift Valley	55	134	61	12	4.80	0.4549	0.0710
	north-eastern	34	107	34	4	4.26	0.4190	0.0505
	Agro-climatic zones (26)							
	I: humid	33	125	48	6	5.10	0.4765	0.0977
	II: sub-humid	40	141	57	9	5.49	0.5023	0.0905
	III: semi-humid	51	157	74	12	5.83	0.5299	0.0912
	IV: semi-humid to semi-arid	75	180	97	12	6.26	0.5673	0.1444
	V: semi-arid	50	143	59	9	5.50	0.5538	0.1501
	VI-VII: arid to very arid	80	152	78	14	5.03	0.5318	0.0785
	Altitudinal range (31)							
	0 - 500 m	46	145	58	6	5.81	0.5287	0.0982
	500 - 1000 m	119	182	90	18	5.99	0.5978	0.1198
	1000 - 1500 m	114	206	131	28	6.61	0.5415	0.1147
	>1500 m	50	147	69	14	5.72	0.4742	0.0862
Wild	**Regions (10)**							
	Coast	39	195	77	46	6.00	0.6806	0.1287
	eastern/central	41	164	55	18	5.12	0.6215	0.1813
	Turkana	17	95	9	7	3.67	0.5165	0.2381
	western/Nyanza	13	98	82	6	3.92	0.4836	0.2179
	Agro-climatic zones (9)							
	I: humid	3	53	0	1	-	0.2993	0.3194
	II: sub-humid	10	107	0	3	4.93	0.5266	0.2056
	III: semi-humid	17	137	13	4	5.78	0.6508	0.0956
	IV: semi-humid to semi-arid	39	178	70	26	5.65	0.6326	0.1834
	V: semi-arid	21	154	32	13	5.76	0.6272	0.1572
	VI-VII: arid to very arid	20	118	7	9	4.70	0.5668	0.2166
	Altitudinal range (18)							
	0 - 500 m	33	190	68	36	7.19	0.6779	0.1621
	500 - 1000 m	36	157	57	13	6.09	0.6151	0.1976
	1000 - 1500 m	32	166	70	11	5.97	0.6133	0.1759
	>1500 m	9	97	0	4	-	-	-

N = Number of samples, A^t = Total number of alleles, A^r = Number of rare alleles, A^p = Number of private alleles, R_s = Allelic richness, H_e = Expected (unbiased) heterozygosity, H_o = Observed heterozygosity, [1]calculated based on a minimum sample of gene copies (shown in brackets for each group).

3.4.3 Geographical variation in cultivated and wild diversity

The allelic richness for cultivated sorghum was significantly lower in Turkana (mean R_s = 3.45) compared to all other regions ($4.26 \leq R_s \leq 5.59$; $P \leq 0.05$), where no significant inter-regional differences in allelic richness were evident. In contrast, gene diversity was significantly higher in the eastern/central region (mean H_e = 0.5595) compared to the western/Nyanza (mean H_e = 0.4875; $P = 0.005$), Rift Valley (mean H_e = 0.4549; $P < 0.001$), north-eastern (mean H_e = 0.4190; $P = 0.003$) and Turkana (mean H_e = 0.3396; $P < 0.001$) regions. The coastal region appeared to harbour the most genetically rich pool of wild sorghum, as exhibited in a mean allelic richness of 6.00, that was significantly higher than those of Turkana (R_s = 3.67; $P < 0.001$), the eastern/central (R_s = 5.12; $P = 0.004$) and western/Nyanza (R_s = 3.92; $P < 0.001$) regions. This trend was also revealed in the gene diversity estimates.

3.4.4 Environmental variation in cultivated and wild sorghum diversity

Cultivated sorghum from the semi-humid to semi-arid agro-climatic zone exhibited significantly higher allelic richness and gene diversity estimates (R_s = 6.26; H_e = 0.5673) compared to the other six zones ($5.03 \leq R_s \leq 5.83$; $0.5538 \geq H_e \geq 0.4765$; $0.002 \geq P \leq 0.042$). In the wild sorghum gene pool, the humid agro-climatic zone was not considered for the comparative genetic diversity due to insufficient sample size (only three samples). For the remaining zones, wild sorghum genetic diversity (both H_e and R_s) generally appeared to be less diverse in the drier and wetter zones compared to zones of moderate humidity.

Within altitudinal ranges, cultivated sorghum allelic richness ranged from 5.72 (>1500 masl) to 6.61 (1000 - 1500 masl) and gene diversity estimates from 0.4742 (>1500 masl) to 0.5978 (500 - 1000 masl). Generally, cultivated sorghum genotypes originating from mid altitude (500 - 1000 masl) to high altitude (1000 - 1500 masl) appeared to harbour the most diverse pool of cultivated sorghum as estimated by allelic richness and gene diversity. In contrast, cultivated sorghum exhibited less genetic variability beyond 1500 masl both in terms of allelic richness and gene diversity. In wild sorghum, allelic richness values ranged from 5.97 (1000 - 1500 masl) to 7.19 (0 - 500 masl) and gene diversity values from 0.6133 (1000 - 1500 masl) to 0.6779 (0 - 500 masl). Wild sorghum generally exhibited a trend of greater diversity at lower altitudes and less diversity at more elevated altitudes.

3.4.5 Genetic relationships within and among cultivated and wild sorghum

Figures 3.4 and 3.5 show the genetic relationships among cultivated and wild sorghum gene pools based on first (axis 1 and 2) and second (axis 1 and 3) planes of the PCoA, respectively. The three main axes of the PCoA explained 7.6%, 6.1% and 5.3% of the total variability respectively. Generally, the separation between cultivated and wild sorghum gene pools is low. The first two planes of the PCoA clearly separated the north-eastern cultivated sorghum from both the cultivated and wild counterparts from the other regions (Figure 3.4). Similar trends were observed for the second plane of the PCoA where cultivated and wild sorghum from Turkana as well as wild sorghum from the coastal regions were separated from the other regions (Figure 3.5).

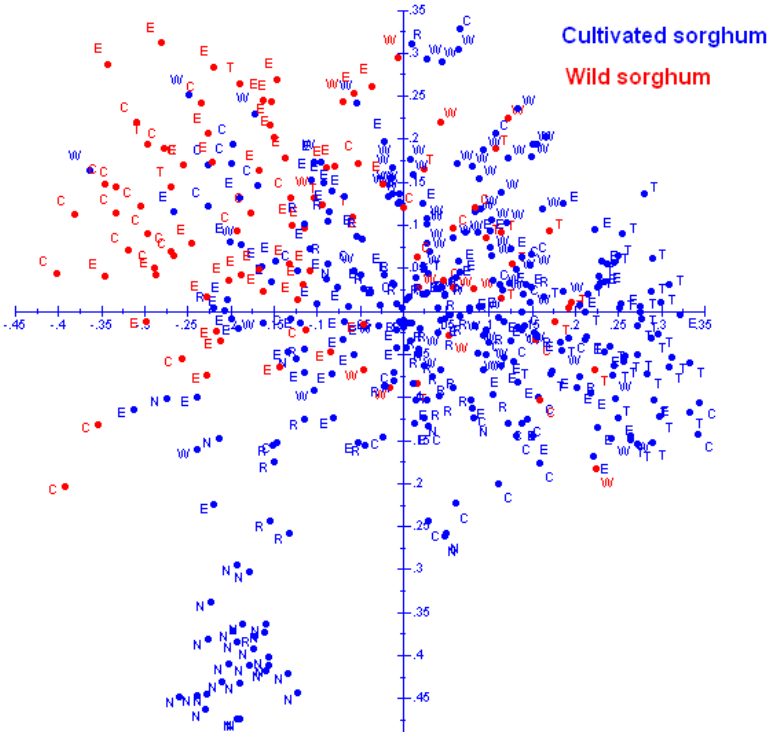

Figure 3.4 Biplot of the axis 1 and 2 of the principle coordinate analysis based on the dissimilarity of 24 SSR markers for cultivated and wild sorghum. Letters E, C, N, R and W represent eastern/central, coast, north-eastern, Rift Valley and western/Nyanza regions, respectively.

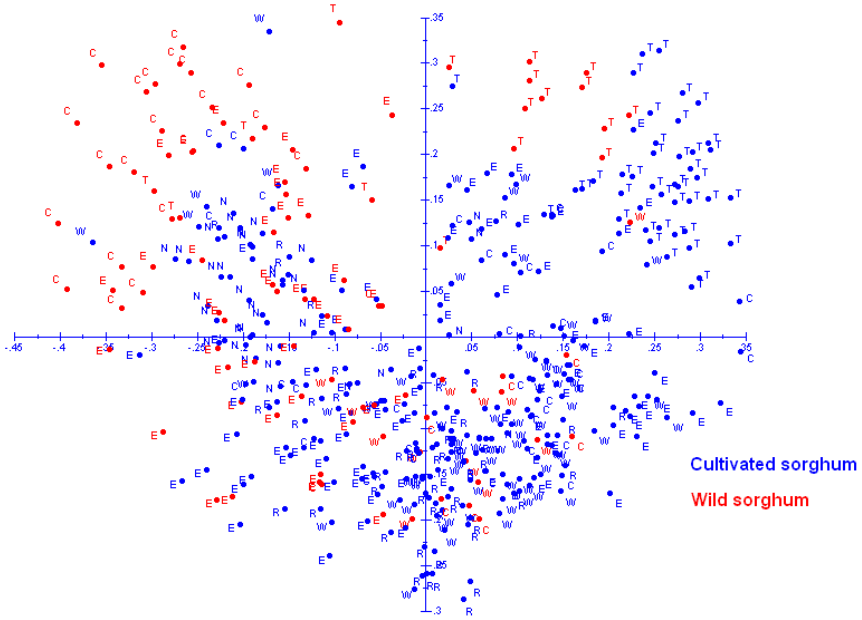

Figure 3.5 Biplot of the axis 1 and 3 of the principle coordinate analysis based on the dissimilarity of 24 SSR markers for cultivated and wild sorghum. Letters E, C, N, R and W represent eastern/central, coast, north-eastern, Rift Valley and western/Nyanza regions, respectively.

The genetic relationships within cultivated and wild sorghum gene pools as determined by NJ analysis are presented in Figures 3.6 and 3.7, respectively. For cultivated sorghum, genotypes were assigned into four broad groups (Figure 3.6; cluster A-D) of largely mixed geographic origin. Cluster A consisted of most of the Rift Valley sorghum and some individuals from the western/Nyanza and eastern/central regions. Cluster B consisted of four major sub-groups with membership drawn from all sampled regions. Notably, most of the north-eastern collections were clustered together in one of the sub-groups in cluster B and so was a group of eastern/central dual-season sorghum genotypes cultivated only in the highlands of Mt. Kenya. Cluster C included a sub-group consisting of all but one of the Turkana genotypes, as well as collections from western/Nyanza, eastern/central, Rift Valley and coastal regions. Finally, cluster D was predominated by genotypes from western/Nyanza alongside a few individuals from all other regions except north-eastern and Turkana (Figure 3.6).

Wild sorghum genotypes were grouped into two but rather genetically proximal clusters (Figure 3.7). Cluster A consisted of most of the coastal genotypes, a large proportion of the eastern/central collections, a few collections from Turkana and a single collection from the western/Nyanza region. Cluster B consisted of most of the Turkana genotypes in a distinct sub-group, plus genotypes from the western/Nyanza, eastern/central and coastal regions (Figure 3.7).

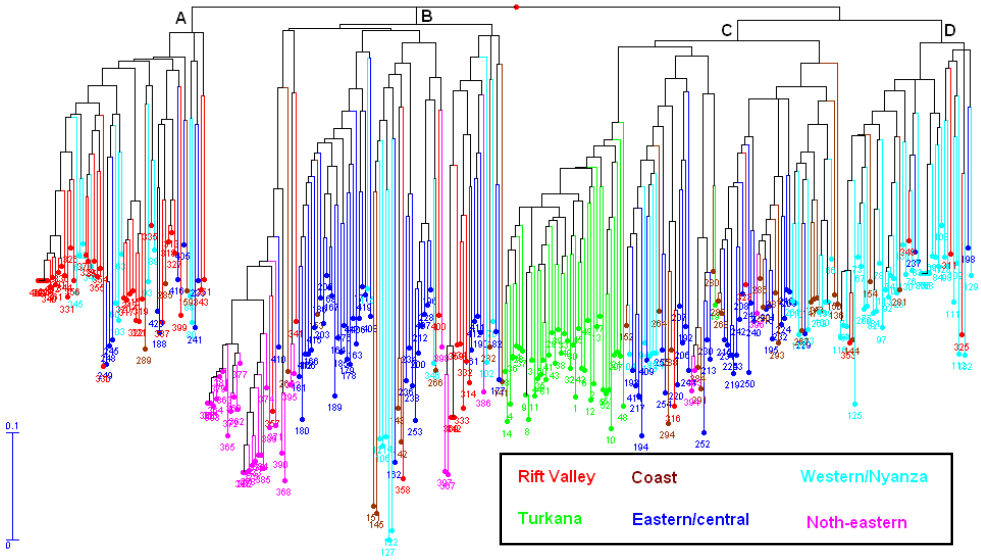

Figure 3.6 **Neighbour-joining cluster analysis dendrogram showing the genetic relationship among cultivated genotypes in Kenya. The dendrogram was generated based on data from 24 SSR markers.**

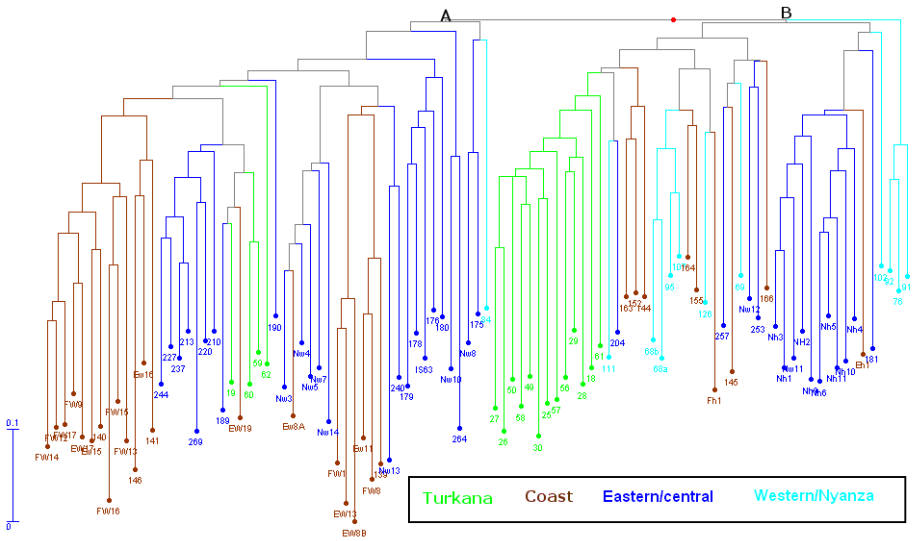

Figure 3.7 **Neighbour-joining cluster analysis dendrogram showing the genetic relationship among wild/weedy genotypes in Kenya. The dendrogram was generated based on data from 24 SSR markers.**

3.4.6 Cultivated and wild sorghum genetic structure

3.4.6.1 FST-based genetic differentiation

Consistent with previous results of PCoA, genetic differentiation between cultivated and wild sorghum based on F_{ST} was low but highly significant ($F_{ST} = 0.062$; $P < 0.001$) (Table 3.3).

Table 3.3 **F_{ST}-based genetic differentiation of the sorghum gene pool at various levels**

Sorghum gene pool	Differentiation level	Mean F_{ST}
Overall	Sub-specific (cultivated vs wild sorghum)	0.062***
Cultivated	Geographic regions	0.187***
	Agro-climatic zones	0.077***
	Altitudinal classes	0.066***
Wild	Geographic regions	0.097***
	Agro-climatic zones	0.054***
	Altitudinal classes	0.062***

*** Highly significant P-value ($P < 0.001$)

Within the cultivated sorghum gene pool, genetic differentiation was high among geographic regions (F_{ST} = 0.187; P < 0.001) and moderate among agro-climatic zones (F_{ST} = 0.077, P < 0.001) and altitudinal ranges (F_{ST} = 0.066; P < 0.001) (Table 3.3). In contrast, similar levels of moderate genetic differentiation were observed in wild sorghum among geographic regions (F_{ST} = 0.097; P < 0.001), agro-climatic zones (F_{ST} = 0.054) and altitudinal ranges (F_{ST} = 0.062; P < 0.001) (Table 3.3). It is to be noted that cultivated sorghum showed two-fold higher genetic differentiation among geographic regions than wild sorghum.

Table 3.4 shows estimates of pairwise genetic differentiation, F_{ST}, within cultivated and wild sorghum at geographical level. Overall, all F_{ST} values were significantly greater than zero (0.001 > P ≤ 0.05). Genetic differentiation among cultivated and wild sorghum gene pools was generally lower within (F_{ST}: 0.03 - 0.18) than among (F_{ST}: 0.05 - 0.33) regions. Within regions, the highest genetic differentiation between cultivated and wild sorghum gene pools was recorded in Turkana, while the least was recorded in western/Nyanza (F_{ST} = 0.03). The genetic differentiation was comparable and moderate between cultivated and wild sorghum gene pools in the coastal (F_{ST} = 0.11) and eastern (F_{ST} = 0.10) regions. Both cultivated and wild sorghum showed variable levels of genetic differentiation across regions, with the former generally exhibiting greater F_{ST} values than the latter. The lowest level of genetic differentiation (F_{ST} = 0.03) in cultivated sorghum was observed between the eastern/central and coastal region populations and the highest (F_{ST} = 0.44) among the Turkana and north-eastern populations. Notably, a high level of inter-region genetic similarity was exhibited between eastern/central and coastal regions (F_{ST} = 0.03), western/Nyanza and coastal regions (F_{ST} = 0.05) and eastern/central and western/Nyanza (F_{ST} = 0.07). In contrast, Turkana and north-eastern cultivated populations appeared to be clearly distinct both between each other and among the rest of the cultivated populations. Similar trends were revealed in wild populations, with substantial inter-regional similarities among coastal, eastern/central and western regions (0.06 ≤ F_{ST} ≤ 0.10), that was coupled with substantial distinctiveness of Turkana populations in relation to those from other regions (0.13 ≤ F_{ST} ≤ 0.17, P < 0.001). Comparisons among sorghum types showed closer genetic proximity among cultivated and wild sorghum genotypes within than between other regions for Turkana and western/Nyanza genotypes, but not for the coastal and eastern/central genotypes.

Table 3.4 Estimates of pairwise F_{ST} among collections of cultivated and wild sorghum within and among different geographical regions. Letters E, C, N, R and W represent eastern/central, coast, north-eastern, Rift Valley and western/Nyanza regions, respectively

		Cultivated						Wild			
		C	**E**	**N**	**R**	**T**	**W**	**C**	**E**	**T**	**W**
Cultivated	**C**	*	0.03	0.24	0.11	0.25	0.05				
	E		*	0.24	0.09	0.24	0.07				
	N			*	0.30	0.44	0.31				
	R				*	0.31	0.13				
	T					*	0.27				
	W						*				
Wild	**C**	0.11	0.13	0.23	0.17	0.31	0.14	*	0.06	0.13	0.10
	E	0.10	0.10	0.26	0.14	0.31	0.13		*	0.13	0.09
	T	0.15	0.17	0.33	0.24	0.18	0.20			*	0.17
	W	0.05	0.07	0.31	0.13	0.32	0.03				*

Surprisingly, coastal and eastern/central cultivated genotypes were genetically closer to wild populations from western/Nyanza than to those found in the same regions. Notably, cultivated sorghum populations from both north-eastern and Turkana generally showed a high degree of differentiation from wild populations of sorghum. Interestingly, there appeared to be more divergence among cultivated and wild sorghum gene pools in Turkana (F_{ST} = 0.18) than among Turkana and eastern/central wild sorghum genotypes (F_{ST} = 0.13).

Pairwise F_{ST} values within and among agro-climatic zones for both cultivated and wild sorghum failed to reveal any consistent patterns of differentiation, except in populations from the more the arid zones (VI - VII) which seemed more genetically diverged from the rest (Table 3.5).

Table 3.5 Estimates of pairwise $^{1}F_{ST}$ among collections of cultivated and wild sorghum within and among different agro-climatic zones

		Cultivated						Wild				
		I	**II**	**III**	**IV**	**V**	**VI-VII**	**I**	**III**	**IV**	**V**	**VI-VII**
Cultivated			**0.0**	**0.0**	0.0	0.0						
	I	*	**1**	**1**	4	9	0.16					
				0.0	0.0	0.0						
	II		*	**0**	3	8	0.13					
					0.0	0.0						
	III			*	2	6	0.12					
						0.0						
	IV				*	2	0.11					
	V					*	0.11					
	VI-VII						*					
Wild			0.0	0.0	0.0	0.1			**0.0**	0.0	0.0	
	II	0.08	6	6	4	0	0.17	*	**3**	4	4	0.10
			0.1	0.1	0.0	0.1				0.0	0.0	
	III	0.13	2	1	9	2	0.17		*	2	1	0.07
			0.0	0.0	0.0	0.1					0.0	
	IV	0.07	6	5	6	0	0.15			*	4	0.11
			0.1	0.1	0.1	0.1						
	V	0.14	3	3	1	6	0.19				*	0.09
			0.1	0.1	0.1	0.1						
	VI-VII	0.17	6	5	3	5	0.12					*

[1]Non significant values (P > 0.05) in bold. I = humid, II = sub-humid, III = semi-humid, IV = semi-humid to semi-arid, V = semi-arid and VI-VII = arid to very arid.

Table 3.6 presents estimates of F_{ST} among cultivated and wild sorghum collections within and among altitudinal ranges. Low to moderate levels of genetic differentiation (F_{ST}: 0.03 - 0.12; $P < 0.001$) were recorded for cultivated sorghum among altitudinal ranges, whereas in wild sorghum the level of genetic differentiation at altitudinal level was generally moderate (F_{ST}: 0.05 - 0.11; $P \leq 0.01$). Genetic differentiation among cultivated and wild sorghum populations was variable, ranging from low to moderate within altitudinal classes (F_{ST}: 0.05 - 0.16; $P < 0.001$) and moderate to great among altitudinal

classes (F_{ST}: 0.09 - 0.20; P < 0.001). Overall, the pairwise F_{ST} values did not reveal a consistent pattern in levels of differentiation within and among cultivated and wild sorghum gene pools across altitudinal ranges.

Table 3.6 Estimates of pair wise F_{ST} among collections of cultivated and wild sorghum within and among different altitudinal (masl) ranges

		Cultivated				Wild			
		< 500	**500 - 1000**	**1000 - 1500**	**>1500**	**< 500**	**500 - 1000**	**1000 - 1500**	**> 1500**
Cultivated	**< 500**	*	0.03	0.06	0.12				
	500 - 1000		*	0.06	0.10				
	1000 - 1500			*	0.07				
	>1500				*				
Wild	**< 500**	0.13	0.10	0.10	0.15	*	0.06	0.06	0.05
	500 - 1000	0.11	0.11	0.13	0.18		*	0.07	0.11
	1000 - 1500	0.10	0.09	0.05	0.09			*	0.06
	>1500	0.20	0.15	0.14	0.16				*

3.4.6.2 Analysis of molecular variance (AMOVA)

The outcome of partitioning of genetic diversity within and between cultivated and wild gene pools at national, geographical and environmental (agro-climatic and altitudinal) levels using the AMOVA procedure is presented in Tables 3.7 - 3.10. Among cultivated and wild sorghum gene pools, the analysis showed all variance components to be highly significant and the bulk of variation to be partitioned within (93.6%) rather than among (6.4%) the two congeners (Table 3.7). This apparent low level of differentiation between cultivated and wild sorghum gene pools was also observed in the preceding PCoA and F_{ST} genetic analyses procedures. The variance components were highly significant for the partitioning of cultivated and wild sorghum diversity at geographic level (Table 3.8).

There was much less differentiation among than within geographic regions for both cultivated and wild sorghum as indicated by much higher partitioning of the genetic variation within than among regions. Notably, differentiation among geographic regions was 1.7-fold higher in cultivated than in wild sorghum. Significant genetic structure was revealed in cultivated and wild sorghum gene pools both at agro-climatic (Table 3.9) and altitudinal (Table 3.10) level. There was by far more genetic variation within than among the two environmental factors in both cultivated and wild sorghum.

Table 3.7 Analysis of molecular variance among and within cultivated and wild sorghum

Source of variation	Sum of squares	Variance components	Percentage variation
Among sorghum types	166.548	0.50701	6.4 (P < 0.001)
Within sorghum types	6076.477	7.40010	93.6 (P < 0.001)
Total	**6243.024**	**7.90712**	

Table 3.8 Analysis of molecular variance within and among geographic regions for cultivated and wild sorghum

	Source of variation	Sum of squares	Variance components	Percentage variation
a. Cultivated				
	Among regions	735.651	1.40794	19.2 (P < 0.001)
	Within regions	3635.380	5.94047	80.8 (P < 0.001)
	Total	**4371.031**	**7.34841**	
b. Wild				
	Among regions	160.120	0.96419	11.2 (P < 0.001)
	Within regions	1545.325	7.68130	88.8 (P < 0.001)
	Total	**1705.445**	**8.64550**	

Table 3.9 Analysis of molecular variance within and among agro-climatic zones regions (ACZs) for cultivated and wild sorghum

	Source of variation	Sum of squares	Variance components	Percentage variation
a. Cultivated				
	Among ACZs	339.8	0.608	8.5 (P < 0.001)
	Within ACZs	4031.2	6.586	91.5 (P < 0.001)
	Total	**4371.0**		
b. Wild				
	Among ACZs	122.2	0.601	6.9 (P < 0.001)
	Within ACZs	1574.3	8.091	93.1 (P < 0.001)
	Total	**1696.5**		

Table 3.10 Analysis of molecular variance within and among altitudinal classes regions for cultivated and wild sorghum

	Source of variation	Sum of squares	Variance components	Percentage variation
a. Cultivated				
	Among altitudinal classes	247.1	0.518	7.2 (P < 0.001)
	Within altitudinal classes	4123.9	6.717	92.8 (P < 0.001)
	Total	**4371.0**		
b. Wild				
	Among altitudinal classes	94.1	0.630	7.4 (P < 0.001)
	Within altitudinal classes	1467.9	7.934	92.6 (P < 0.001)
	Total	**1562.0**		

3.4.6.3 Bayesian model-based cluster analysis

Consistent with the PCoA, F_{ST} and AMOVA results, the Bayesian model-based cluster analysis at K = 2 failed to identify distinct differentiation among cultivated and wild sorghum gene pools (Figure 3.8). Figure 3.9a-c presents the best K clusters as identified by the Evanno's *ad hoc* ΔK method: K = 5, K = 7 and K = 2 for the combined sorghum gene pool, cultivated gene pool and wild sorghum gene pool, respectively.

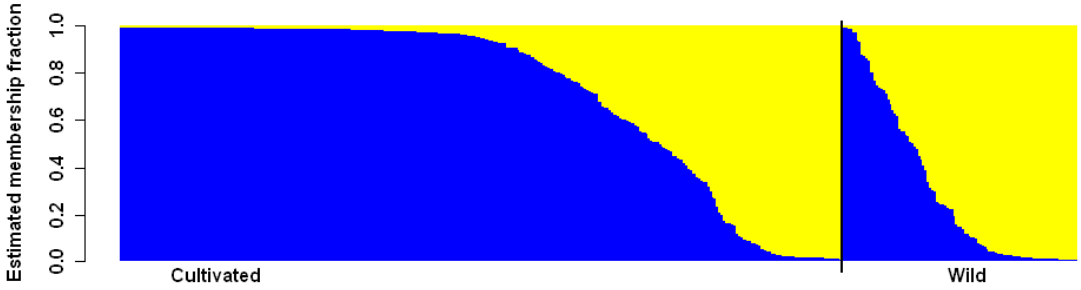

Figure 3.8 **Estimated population structure at K = 2 for the entire sorghum gene pool ordered by type and membership fraction. Each individual is represented by a vertical line, which is partitioned into coloured segments that represent the individual's membership fraction in K clusters.**

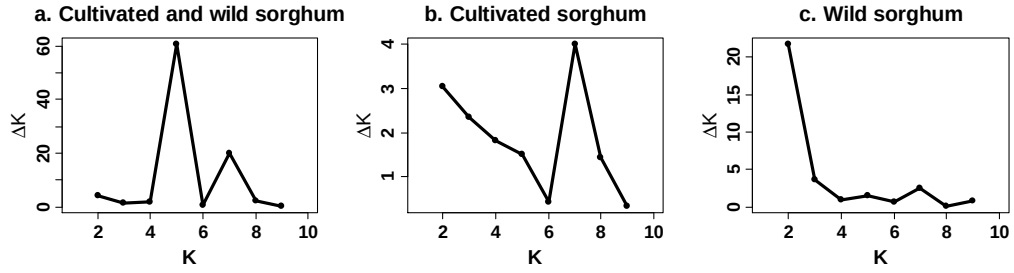

Figure 3.9 **A plot of Evanno's *ad hoc* ΔK statistic against different possible values for K. The modal value indicates the most probable value of K: (a) K = 5 for entire sorghum gene pool, (b) K = 7 for cultivated sorghum gene pool and (c) K = 2 for wild sorghum gene pool.**

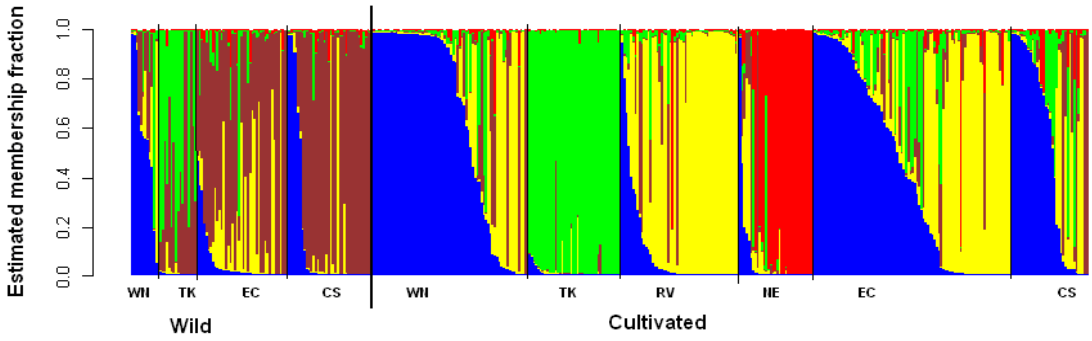

Figure 3.10 Estimated population structure at K = 5 for cultivated and wild sorghum ordered by type and geographic region. Each individual is represented by a vertical coloured line, which is partitioned into coloured segments that represent the individual's membership fraction in K clusters. Letters EC, CS, NE, RV and WN represent eastern/central, coast, north-eastern, Rift Valley and western/Nyanza regions, respectively.

At K = 5 for the entire sorghum gene pool, wild sorghum populations were separated largely into four genetic groups, the largest (brown) of which was least shared with the cultivated sorghum individuals (Figure 3.10). Cultivated sorghum on the other hand was generally represented in all five clusters, with north-eastern and Turkana collations being restricted largely into single distinct groups. Notably, close genetic proximity among cultivated and wild sorghum collections from western/Nyanza zones was suggested by assignment of individuals from both in similar clusters. For the rest of the regions, except Turkana, cultivated and wild sorghum collections seemed to largely share at least two genetic clusters. In Turkana, wild sorghum collections were split into two clusters, one of which (brown) was not shared with its cultivated counterpart (Figure 3.10).

In wild sorghum, Evanno's *ad hoc* ΔK method identified K = 2 as the optimal level of genetic structure (Figure 3.9c). All 13 wild collections from western/Nyanza were largely assigned to one of these two groups (blue), whereas most of the Turkana collections were assigned to the other (Figure 3.11). Wild collections from eastern/central and those from the coastal region were assigned into both clusters, with the largest number of individuals in both cases sharing the same genetic group (blue) as the western/Nyanza collections.

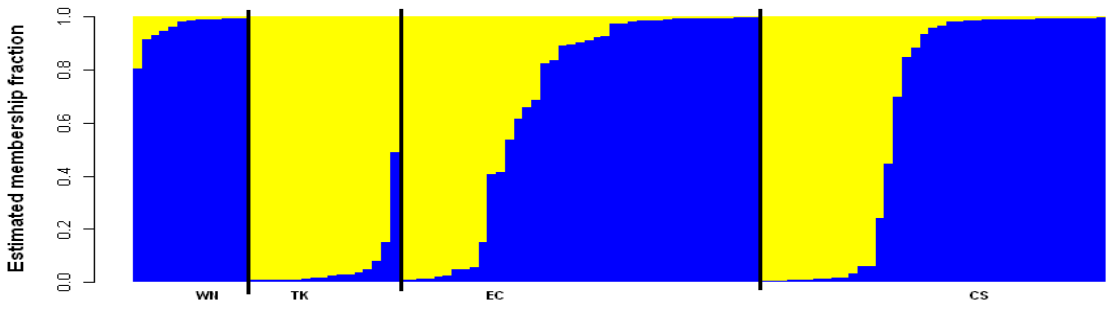

Figure 3.11 Estimated population structure for wild sorghum gene pool at K = 2, ordered by geographic region and membership fraction. Letters EC, CS, NE, RV and WN represent eastern/central, coast, north-eastern, Rift Valley and western/Nyanza regions, respectively.

Evanno's ΔK method determined the optimum number of genetic clusters in cultivated sorghum to be seven (Figure 3.9b). A high level of genetic differentiation within regions was supported by the observed assignment of individuals from a single region into several clusters, apart from Turkana and to a large extent the north-eastern region, where cultivated sorghum individuals were largely grouped into a single genetic cluster (Figure 3.12). Furthermore, in terms of the proportion of cluster membership in each individual, Turkana and to a large extent the north-eastern cultivated sorghum appeared to share little or no ancestry with their counterparts from other regions. Cultivated sorghum from these other regions, in contrast, showed a high degree of shared cluster membership, even though Rift Valley collections appeared predominantly to belong to one of the clusters (pink) (Figure 3.12).

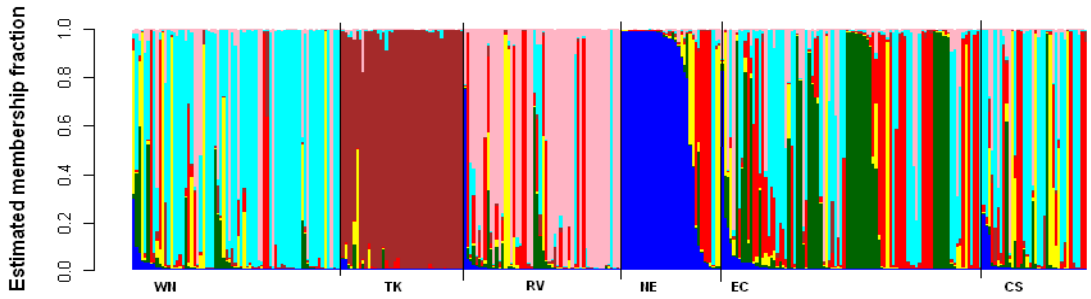

Figure 3.12 Estimated population structure at K = 7 for cultivated gene pool ordered by geographic regions. Each individual is represented by a vertical coloured line, which is partitioned into coloured segments that represent the individual's membership fraction in K clusters. Letters EC, CS, NE, RV and WN represent eastern/central, coast, north-eastern, Rift Valley and western/Nyanza regions, respectively.

3.4.6.4 Spatial genetic structure

Outcome of spatial autocorrelation analyses in cultivated and wild sorghum is presented as correlograms in Figure 3.13. In both cultivated and wild sorghum, there was a clear decrease in pairwise relatedness among individuals with increasing geographical distance, a reflection of strong spatial genetic structure. Cultivated sorghum had a mean regression slope (*blog*) value of -0.015 ($P < 0.001$) and a coefficient of determination value (r^2) of 0.045, while for wild sorghum, the mean regression slope (*blog*) value was -0.017 ($P < 0.001$), while the coefficient of determination (r^2) value was 0.055. The kinship coefficient values were positive and significant within a range of about 180 km in the two conspecifics. Furthermore, negative and significant kinship coefficient values were clearly evident in wild sorghum beyond 600 km, while in cultivated sorghum significant negative values did not show a consistent pattern. When spatial autocorrelation analysis was performed within regions for cultivated sorghum, only Turkana and coastal regions failed to show evidence of spatial genetic structure (Figure 3.14a-f). Eastern/central and north-eastern regions showed significant positive kinship coefficients at both short-distances and long-distances, whereas Rift Valley and western/Nyanza were characterised only by short-distance positive kinship coefficients.

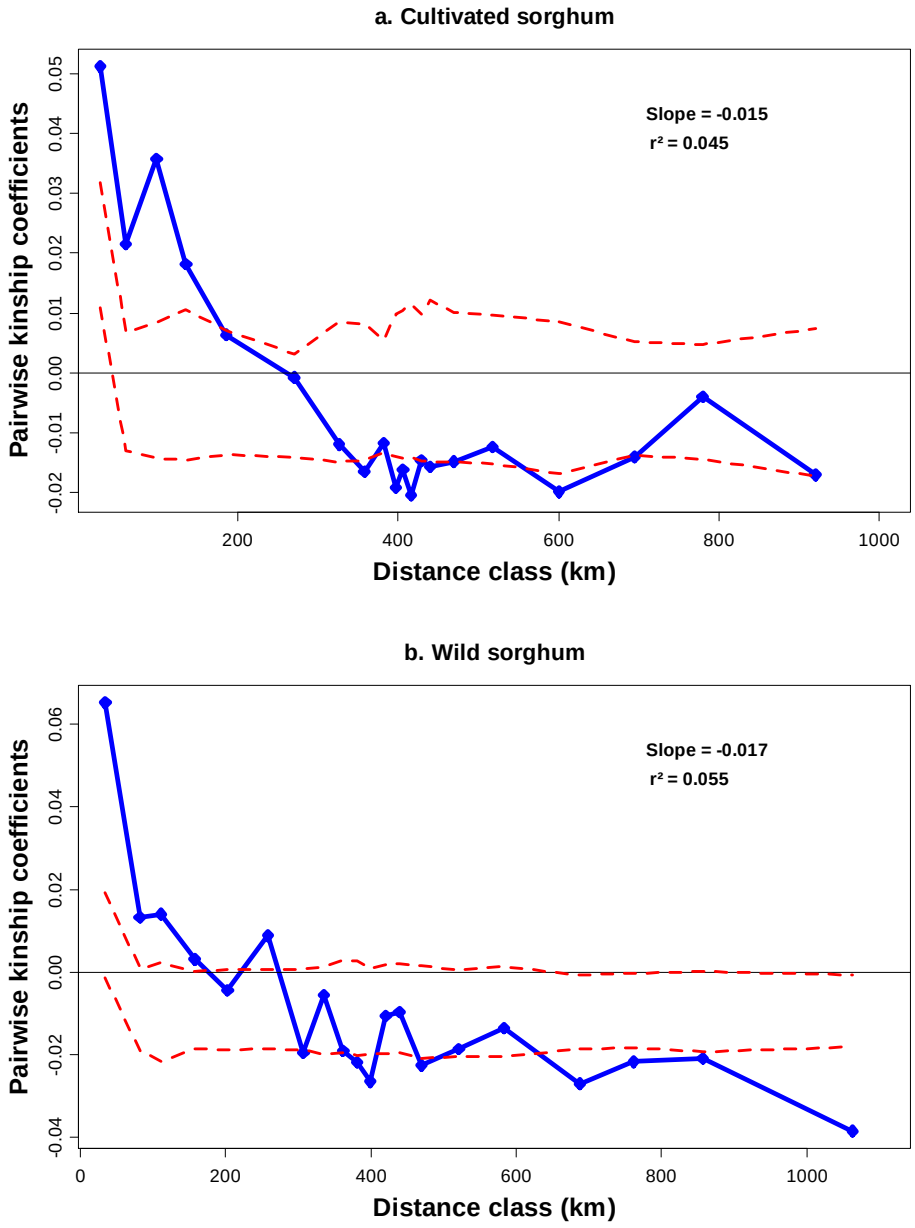

Figure 3.13 **Correlograms for spatial patterns of genetic differentiation in cultivated (a) and wild (b) sorghum genotypes based on Ritland's pairwise kinship coefficient of individuals. The dashed lines represent upper and lower 95% confidence limit envelopes around the null hypothesis of no spatial structure.**

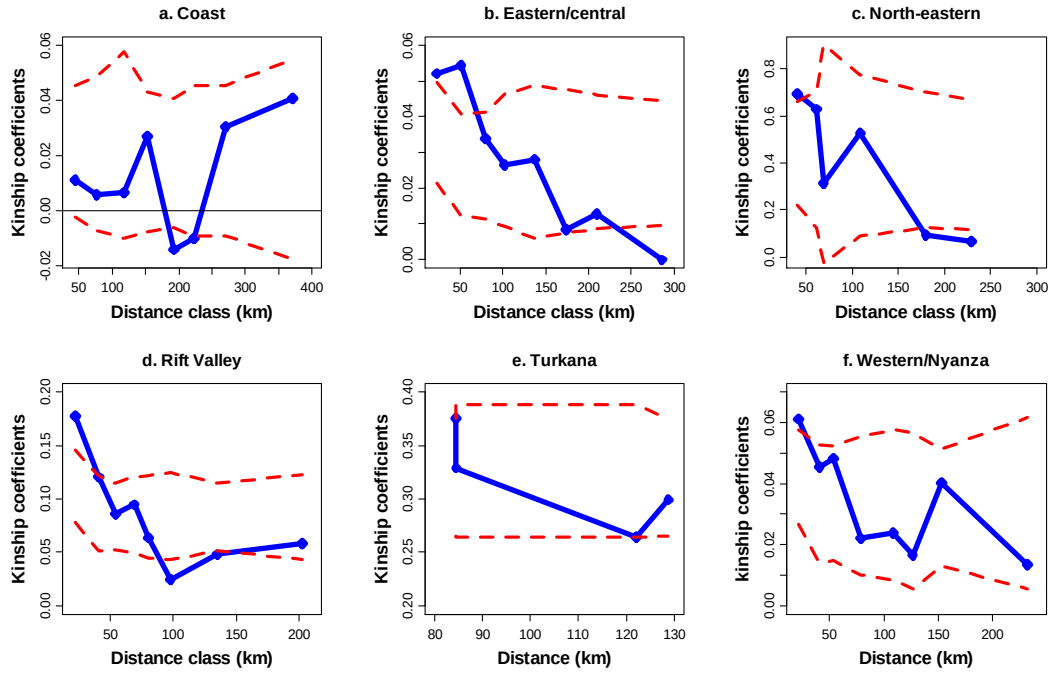

Figure 3.14 Correlograms for spatial patterns of genetic differentiation in cultivated sorghum genotypes for coast (a), eastern/central (b), north-eastern (c), Rift Valley (d), Turkana (e) and western/Nyanza (f) regions based on Ritland's pairwise kinship coefficient of individuals. The dashed lines represent upper and lower 95% confidence limit envelopes around the null hypothesis of no spatial structure.

3.4.7 Spatial analysis of cultivated-wild sorghum genetic distance

The genetic distance between pairs of cultivated and wild sorghum individuals increased linearly with logarithmic spatial distance (Figure 3.15), demonstrating a typical pattern of isolation-by-distance. The regression of pairwise crop-to-wild sorghum genetic distance gave a significant positive slope (*blog* = 0.149, Permutation test: $P \leq 0.001$, $r^2 = 0.028$).

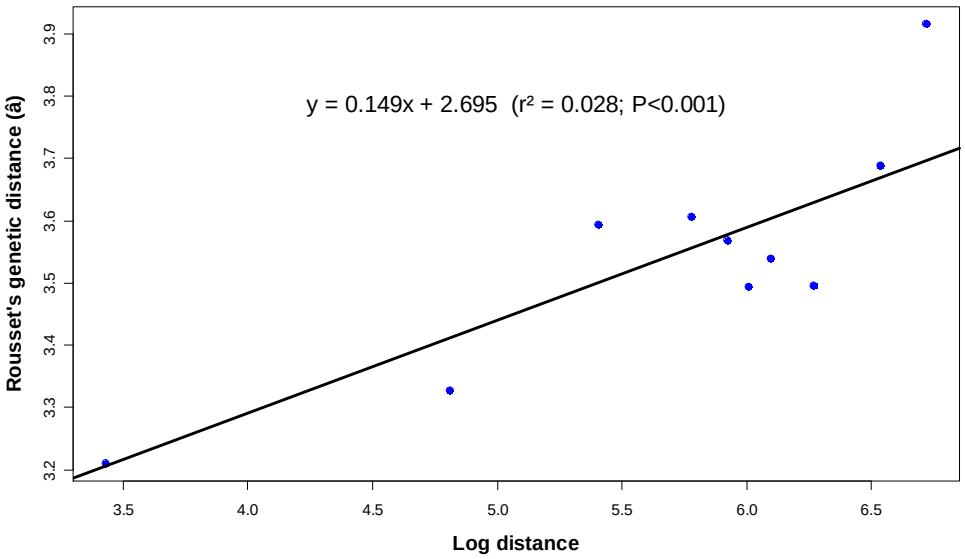

Figure 3.15 A plot of Rousset's genetic distance among cultivated and wild sorghum in relation to isolation distance (in km).

3.5 Discussion

3.5.1 Extent of genetic diversity in cultivated and wild sorghum

Mean gene diversity across the 24 SSR markers for cultivated sorghum in Kenya (H_e = 0.59) is similar to values reported for microsatellites in Niger (H_e = 0.61) by Deu *et al.* (2008) and in South Africa (H_e = 0.60) by Uptmoor *et al.* (2003), but slightly lower than values estimated for Eritrea by Ghebru *et al.* (2002) and for Morocco by Dje *et al.* (1999). In the wild sorghum gene pool, the mean gene diversity estimated for Kenya across the 24 SSR markers (H_e = 0.69) was higher than the gene diversity estimated for a set of accessions selected to represent a wide geographic sampling in Africa (H_e = 0.59) by Casa *et al.* (2005). As noted by Deu and co-workers (2008), however, comparisons of the magnitude of genetic diversity between different studies is difficult as it may be complicated by differences in among other: (i) underlying sampling schemes (single plant vs DNA bulk), (ii) the number of SSR surveyed, (iii) size of SSR repeats (e.g. di-, tri-, or tetra-nucleotide) and (iv) location of the SSR on the genome (coding vs non-coding regions).

During the process of domestication, evolutionary processes of founder effect, population bottleneck and artificial selection are all expected to reduce genetic diversity of the crop in relation to its wild progenitor (Ladizinsky 1999; Gepts 2004). This view was supported in the present study by findings that cultivated sorghum harboured lower genetic diversity (in terms of overall allelic richness, private allelic richness and gene diversity) than its proposed wild progenitor. Findings were consistent with previous comparisons among cultivated and wild sorghum using various genetic markers (Morden *et al.* 1990; Aldrich and Doebley 1992; Cui *et al.* 1995; Casa *et al.* 2005). The large proportion of rare alleles revealed in the entire sorghum collection supported data previously reported for sorghum germplasm with diverse phenotypes and geographic origin (Casa *et al.* 2005). The significantly higher private allelic richness in the wild sorghum relative to its cultivated counterpart is of great importance to the conservation and utilisation of sorghum genetic resources. These findings support the widely held view that crop wild relatives are potential sources of important and unique genes for crop improvement programmes and therefore deserve special attention in plant genetic resources conservation and utilisation strategies.

The considerable variability in proportions of rare alleles and in the extent of allelic richness and gene diversity observed in both cultivated and wild sorghum across geographic regions and across environmental factors (agro-climatic and altitudinal variation) may point to differences in random genetic drift. Adverse climatic conditions and the associated possible reduction in effective population sizes, may also explain the differences. For example, the proportion of rare alleles, the level of allelic richness and gene diversity estimates in cultivated sorghum were lowest in Turkana and north-eastern regions. Historical and/or recent loss of some farmer saved seeds among planting seasons through crop failure may have led to reduced effective population sizes and with it increased loss of rare alleles. In particular, the role of drought in the loss of farmer saved seeds has long been recognised. For instance, Morgan (1974) documented drought related recurrent loss of farmer saved seeds among Turkana farmers as far back as in the early 1970's. Both Turkana and north-eastern region fall within the drought-prone arid to very arid agro-climatic zones (VI and VII) of Kenya, which experience less than 200 mm of rainfall annually. In addition, differences in genetic diversity among regions may be a reflection of regional level differences in farmer selection pressures and practices, the

consequence of which might be fixation of different genes or combination of genes in different landraces.

3.5.2 Genetic structure and relationships in cultivated sorghum

Genetic differentiation appeared high among geographic regions in cultivated sorghum (F_{ST} = 0.19), but moderate among agro-climatic zones (F_{ST} = 0.08) or altitudinal classes (F_{ST} = 0.07). Similar trends were reported for cultivated sorghum in Niger (Deu *et al.* 2008), with moderate differentiation among regions (F_{ST} = 0.07) and lower differentiation among annual rainfall classes (F_{ST} = 0.03). These findings seem to suggest that geographic isolation had played a bigger evolutionary impact on the genetic structure of cultivated sorghum than environmental factors. The main evolutionary forces responsible for producing genetic structure in plant populations are geneflow, selection associated with environmental heterogeneity and random genetic drift (Hartl and Clark 1997; Neal 2004). Because most SSR loci are presumably selectively neutral, environmental factors are expected to offer minimum contribution to the observed genetic structure. In contrast, geographic isolation limits the level of geneflow among populations and should contribute majorly to contemporary genetic structure. Furthermore, AMOVA partitioning revealed 80.8% of the genetic diversity to reside within regions, whereas 19.2% was attributable to differentiation among regions. Past studies on geographic patterns of genetic diversity distribution in cultivated sorghum at country scale (Ayana *et al.* 2001; Ghebru *et al.* 2002; Nkongolo and Nsapato 2003) have reported similar trends and may be explained by non-random geneflow within regions. This may be the case especially for localities within a region that are substantially isolated from each other to warrant minimised levels of seed exchange among farmers.

Clustering patterns in cultivated sorghum based on NJ and Bayesian model-based cluster analyses were similar but failed to show clear-cut grouping according to the geographic origin of genotypes. These results are generally consistent with previous findings showing a weak differentiation among regions in other African countries: Ethiopia and Eriteria (Ayana *et al.* 2000b; Ghebru *et al.* 2002), Malawi (Nkongolo and Nsapato 2003) and Niger (Deu *et al.* 2008). With the exception of Turkana and the north-eastern regions where cultivated sorghum constituted separate groups, samples from the other regions were assigned into three other non-exclusive genetic groups (as shown by the Bayesian-model based analysis). Some of the genotypes presented high levels of admixture.

Intermingling of individuals of different geographic origin in the NJ tree and in the clusters revealed by Bayesian-model based analysis may reflect contemporary and/or historical geneflow among the geographic regions through both formal and informal seed exchange systems (Mutegi *et al.* 2009). For example, the clustering pattern exhibited by western/Nyanza individuals suggests that the region might have played an important role in the dispersal of cultivated sorghum in other parts of the country except in the Turkana and north-eastern regions. This hypothesis was further supported by results of pairwise F_{ST} in cultivated sorghum which revealed western/Nyanza to be generally less genetically differentiated from all regions except Turkana and north-eastern regions.

The strong spatial structure of genetic variation found in Kenyan cultivated sorghum may ascribe to long-distance (up to 180 km) seed exchange. In the present study, a number of farmers originally from the western/Nyanza and eastern/central regions were for example noted to have migrated along with their sorghum landraces into the coastal region. Medium to long distance seed exchanges among regions, mostly through inter-ethnic marriage relationships, but also appreciably through formal distribution of improved varieties via government and non-governmental extension systems (Mutegi *et al.* 2009) may have contributed to the observed pattern. Another plausible explanation for the strong spatial genetic structure observed in cultivated sorghum in the present study is non-random distribution of the racial groups. For example, a recent study on cultivated sorghum in Niger by Deu and co-workers (2008) found a strong relationship between genetic relatedness and geographical distance within a range less than 200 km. These authors postulated that non-random distribution of genetically distinct races, coupled with restricted seed exchange over short distances (less than 100 km) within races were the factors underlying the observed spatial genetic structure. The present study revealed variable patterns of spatial structure in the genetic diversity of sorghum within regions. The correlograms' isolation-by-distance pattern revealed in the eastern/central and north-eastern regions could be explained by a combination of short-distance seed exchanges among farmers and non-random spatial distribution of sorghum landrace types within regions. Mutegi *et al.* (2009) reported sorghum seed systems in Kenya to be mainly traditional where farmers' relatives and local markets constitute important sources of seed. The lack of spatial genetic structure in the Turkana and coastal sorghum was consistent with random spatial distribution of alleles, probably due to extensive seed exchange within the two regions. In Turkana for example, it was established through

farmer interviews that since the early 1980's local sorghum landraces mostly from the southern part of the Turkana district have been multiplied and distributed by different non-governmental organisations across different farming areas in the region.

3.5.3 Genetic structure and relationships in wild sorghum

Like in cultivated sorghum, genetic differentiation in wild sorghum was on the basis of F_{ST} estimates higher among regions (F_{ST} = 0.097) than among agro-climatic zones (F_{ST} = 0.054) or among altitudinal classes (F_{ST} = 0.062). This suggested that wild sorghum diversity was structured more along geographical than environmental factors. Differentiation among geographic regions based on AMOVA was statistically significant, but as was the case in cultivated sorghum, the bulk of diversity (88.8%) was revealed to reside within regions compared to 11.2% found among regions. The lower level of differentiation in wild sorghum among regions was further corroborated by pair-wise F_{ST} estimates which were generally moderate, ranging from 0.06 (coast vs eastern/central) to 0.17 (Turkana vs western/Nyanza). Similarly, NJ and Bayesian-model based analyses confirmed the poor differentiation among regions in wild sorghum, with genotypes failing to group according to their geographic origin. These results generally conform to the only other published country-scale diversity analysis study in wild sorghum (Ayana *et al.* 2000a). The researchers found 88% of the total genetic variation in Ethiopian wild sorghum to reside within regions of origin, whereas 12% was partitioned among regions of origin. This could be explained either by inter-regional geneflow, recent common descent or a combination of the two factors. Although in the present study most of the regions were well isolated from each other geographically, long-distance seed dispersal of wild sorghum could be effected via the exchange of cultivated sorghum seed across regions. For example, a putative hybrid sorghum individual obtained from one of the sorghum fields at the coast whose owner had migrated from the western/Nyanza region was observed to cluster with some wild sorghum individuals from western/Nyanza.

In plant natural populations, spatial distribution of genetic variation is primarily determined by seed and pollen dispersal, habitat distribution, micro-environmental selection and genetic drift (Levin and Kerster 1974; Epperson 1993). The significant positive spatial autocorrelation within a range of about 180 km, coupled with a significant negative spatial autocorrelation from about 600 km found in wild sorghum was typical of clinal spatial genetic structure. Among the factors that could explain this kind of

correlogram profile, one could perhaps consider seed-mediated and/or pollen-mediated geneflow. Migration (seed and pollen dispersal with or without interbreeding) causes similarity between neighbouring populations, whereas distant populations differ for the studied autocorrelation coefficient (Sokal and Oden 1978; Epperson 1993; 2004). The significant long-distance negative kinships observed beyond 600 km may reflect founder effects (through long-distance dispersal and establishment of new populations), coupled with geneflow among initially differentiated neighbouring populations that resulted in local autocorrelation. Surprisingly, significant positive spatial autocorrelation was observed within approximately 180 km in both cultivated and wild sorghum, suggesting that largely similar evolutionary factors could underlie the spatial genetic structure of the two congeners (Sokal and Oden 1978). There are two plausible explanations for this identical pattern of spatial genetic structure: (i) inadvertent dispersal and establishment of wild sorghum seed via cultivated sorghum seed systems and (ii) pollen-mediated crop-wild geneflow in sites of sympatric occurrence.

3.5.4 Genetic relationships among cultivated and wild sorghum individuals

There was no distinct separation among cultivated and wild sorghum genotypes as revealed by PCoA and Bayesian analyses, a suggestion of close genetic proximity between the two gene pools. These results appeared to contradict past studies using allozyme (Aldrich *et al.* 1992), RFLP (Cui *et al.* 1995) and microsatellite (Casa *et al.* 2005) markers, where a distinct separation among cultivated and wild sorghum was reported. All these previous studies acquired samples from *ex situ* collections and thus the extent of range overlap between cultivated and wild sorghum could not be ascertained. In contrast, most of the samples of cultivated and wild sorghum analysed in the present study were sympatrically growing *in situ* within and around sorghum fields. The close genetic proximity among cultivated and wild sorghum in the present study may therefore reflect important historical geneflow between the two congeners. This interpretation was further supported by the low level of genetic structure revealed between cultivated and wild sorghum using the F_{ST}, AMOVA and Bayesian-model cluster analyses. At the same time, results of pair-wise F_{ST} analysis between cultivated and wild sorghum across regions showed generally greater crop-wild semblance within than among regions. Finally, genetic distance between cultivated and wild sorghum individuals at country-scale was observed to increase with increasing distance with a positive and significant slope. Cultivated and wild sorghum are inter-fertile, with natural hybrids among ssp.

bicolor and ssp. *verticiliflorum* within and around cultivation in Africa being well documented (Dogget and Majisu 1968; Dogget and Prasada Rao 1995; Tesso *et al.* 2008). Even in the present study, several putative hybrid plants, whose morphological attributes were observed to be intermediate between cultivated and wild sorghum were encountered in several of the visited localities across regions.

The extent of differentiation between cultivated and wild sorghum gene pools across regions as revealed by pairwise F_{ST} was shown to vary substantially across regions. These results seem to suggest that the extent of introgression between the two congeners varies across regions, probably due to regional differences in farmer practices such as weed management. In western/Nyanza for example, most farmers were observed to tolerate weedy putative crop-wild hybrids in their sorghum fields even after harvest. In addition to possibly enhancing hybridisation between cultivated and wild sorghum on-farm (due to overlapping flowering stages) such a practice could lead to increased populations of wild sorghum in the subsequent growing season. Not surprisingly, genetic differentiation among cultivated and wild sorghum was least in western/Nyanza region compared to the other regions.

Most north-eastern cultivated sorghum genotypes showed a distinct genetic separation both from the other cultivated sorghum and from the wild sorghum of other geographic regions. This possibly suggested that the evolutionary history of north-eastern cultivated sorghum was separate from that of cultivated and wild sorghum gene pools from the other geographic regions. This pool of cultivated sorghum may be part of the Ethiopian sorghum gene pool, having originated largely from the Boran agro-pastoralist ethnic group who dominate the upper eastern and north-eastern provinces of Kenya and whose distribution spans across the Kenyan-Ethiopian border. The Turkana cultivated and wild sorghum genotypes also showed a tendency to form a distinct group from their counterparts in the other regions, possibly a reflection of their independent evolutionary history. The Turkana region is located toward northern Uganda and southern Sudan, raising the possibility that the Turkana sorghum gene pool might have originated from either of these two countries. Furthermore, both the Turkana and north-eastern regions are geographically relatively remote in relation to other sorghum growing regions and therefore expected to have experienced minimal if any cross-regional pollen and/or seed-mediated genetic exchange. Interestingly this appeared to hold true even among the two

regions, even though they are themselves geographically relatively close. The two regions are separated by the Lake Turkana and Mt. Kulal, both of which could have acted as physical barriers to seed-mediated geneflow.

3.6 Conclusions and recommendations

This study revealed higher levels of genetic diversity in the wild compared to its cultivated congener, which suggests that the process of domestication reduced levels of genetic variation in sorghum. Wild sorghum is thus a potential source of novel genes for broadening the genetic base of cultivated sorghum. Furthermore, levels of genetic diversity within the two *S. bicolor* conspecifics differed significantly among geographic regions. Adequate measures need to be put in place for systematic conservation of these important genetic resources using complementary *ex situ* and *in situ* approaches. Such approaches could benefit from further studies on the extent and partitioning of diversity in the country, with wild sorghum samples from natural habitats away from cultivated lands and from the regions not sampled in the present study (Rift Valley and north-eastern) being included.

This study revealed genetic diversity in cultivated and wild sorghum pools to be structured more at geographic level than at agro-climatic or altitudinal level. This probably suggests that geneflow and genetic drift have played a more central evolutionary role than environmental/adaptive selection in shaping the contemporary genetic structure of both the crop and its wild progenitor in Kenya. Furthermore, there was greater partitioning of diversity within than among geographic regions for both cultivated and wild sorghum, with the proportion of diversity distributed among regions being greater in the former than in the latter. This possibly reflects greater uniformity within regions in the cultivated sorghum as compared to the wild sorghum, probably due to higher levels of within-region geneflow in the former (through seed exchange by farmers) than the latter.

Cultivated and wild sorghum showed strong spatial genetic structure at country scale, probably due to human-mediated long-distance seed dispersal in the two congeners. The strong spatial structure revealed in cultivated and wild sorghum may in addition reflect non random spatial organisation of distinct genetic types since some landraces and/or ecotypes may be found in some regions and not others. Further work is however needed to confirm this hypothesis since the racial classification of the various cultivated and wild

sorghum samples used in this study was not considered. At regional scale, different patterns of spatial genetic structure were demonstrated in the cultivated gene pool, probably a reflection of variations in the underlying evolutionary forces such as seed-mediated geneflow.

Important historical geneflow between cultivated and wild sorghum is strongly suggested by two findings in this study: (i) low level of divergence between cultivated and wild sorghum gene pools and (ii) significant isolation-by-distance between pairs of cultivated and wild sorghum individuals. The level of divergence between cultivated and wild sorghum varied among geographic regions, probably a reflection of intra-region differences in the level of crop-to-wild geneflow. Differences in farmer practices such as weedy sorghum management and/or seed selection are some of the factors that could explain this inter-region variation in the extent of crop-to-wild geneflow. Furthermore, the pattern of increased genetic similarity between geographically close pairs of cultivated and wild sorghum individuals relative to isolated ones (isolation-by-distance) as revealed in this study is of further biosafety significance. It suggests that crop-to-wild geneflow in sorghum is spatially predictable, with the risk of transgene escape into wild-weedy relatives of the crop being higher within and around cultivated fields compared to natural habitats away from cultivation. Overall, this study suggests that deployment of GM sorghum in Kenya will lead to escape and persistence of transgenes into wild-weedy sorghum relatives, with the rate of crop-to-wild geneflow being variable among growing regions. The consequences of such transgenes escape and persistence is not known and needs to be characterised so as to formulate adequate biosafety guidelines and regulations for testing and commercially releasing GM sorghum in the country.

Finally, this study brings to the fore the urgent need to conduct systematic collection and *ex situ* conservation of existing wild-weedy sorghum genetic resources in the country as a safeguard against unpredictable consequences of transgeneflow from GM sorghum. Such effort should include landraces of the crop, which still form the bulk of cultivated sorghum in all the crop's growing regions in Kenya.

3.7 References

Aldrich PR, Doebley J. 1992. Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics 85:293-302.

Aldrich PR, Doebley J, Schertz KF, Stec A. 1992. Patterns of allozyme variation in cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics 85:451-460.

Auer C. 2008. Ecological risk assessment and regulation for genetically-modified ornamental plants. Critical Reviews in Plant Sciences 27:255-271.

Avise JC. 1998. The history and purview of phylogeography: a personal reflection. Molecular Ecology 7:371-379.

Avise JC. 2000. Phylogeography: The history and formation of species. Harvard University Press Massachusetts, USA, pp. 447.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Niegel JE, Reeb CA, Saunders NC. 1987. Intraspecific phylogeography - the mitochondrial-DNA bridge between population genetics and systematics. Annual Review of Ecology and Systematics 18:489-522.

Ayana A, Bekele E, Bryngelsson T. 2000a. Genetic variation in wild sorghum (*Sorghum bicolor* ssp *verticilliflorum* (L.) Moench) germplasm from Ethiopia assessed by random amplified polymorphic DNA (RAPD). Hereditas 132:249-254.

Ayana A, Bryngelsson T, Bekele E. 2000b. Genetic variation of Ethiopian and Eritrean sorghum (*Sorghum bicolor* (L.) Moench) germplasm assessed by random amplified polymorphic DNA (RAPD). Genetic Resources and Crop Evolution 47:471-482.

Ayana A, Byngelsson T, Bekele E. 2001. Geographic and altitudinal allozyme variation in sorghum (*Sorghum Bicolor* (L.) Moench) landraces from Ethiopia and Eritrea. Hereditas 135:1-12.

Barnaud A, Deu M, Garine E, Mckey D, Joly HI. 2007. Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. Theoretical and Applied Genetics 114:237-248.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F. 2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier (France),

Bhatia CR, Mitra R. 2003. Consequences of geneflow from genetically engineered crops. Current Science 84:138-141.

Casa AM, Mitchell SE, Hamblin MT, Sun H, Bowers JE, Paterson AH, Aquadro CF, Kresovich S. 2005. Diversity and selection in Sorghum: simultaneous analyses using simple sequence repeats. Theoretical and Applied Genetics 111:23-30.

Chandler S, Dunwell JM. 2008. Geneflow, risk assessment and the environmental release of transgenic plants. Critical Reviews in Plant Sciences 27:25-49.

Clayton WD, Renvoize RD. 1982. Poaceae. Flora of Tropical East Africa, Part 3. AA Balkema Rotterdam, Netherlands, pp. 731.

Cleveland DA, Soleri D. 2005. Rethinking the risk management process for genetically engineered crop varieties in small-scale, traditionally based agriculture. Ecology and Society 10:1-33.

Conner AJ, Glare TR, Nap JP. 2003. The release of genetically modified crops into the environment. Part II. Overview of ecological risk assessment. The Plant Journal 33:19-46.

Cui YX, Xu GW, Magill CW, Schertz KF, Hart GE. 1995. RFLP-based assay of *Sorghum bicolor* (L) Moench. genetic diversity. Theoretical and Applied Genetics 90:787-796.

De Wet JMJ. 1978. Systematics and evolution of sorghum sect. Sorghum (Gramineae). American Journal of Botany 65:477-484.

De Wet JMJ, Harlan JR, Price EG. 1970. Origin of variability in the Spontanea complex of *Sorghum bicolor*. American Journal of Botany 57:704-707.

De Wet JMJ, Huckabay JP. 1967. The origin of *Sorghum bicolor*. II. Distribution and domestication. Evolution 21:787-802.

Deu M, Hamon P, Chantereau J, Dufour P, Dhont A, Lanaud C. 1995. Mitochondrial-DNA diversity in wild and cultivated sorghum. Genome 38:635-645.

Deu M, Sagnard F, Chantereau J, Calatayud C, Herault D, Mariac C, Pham JL, Vigouroux Y, Kapran I, Traore PS, Mamadou A, Gerard B, Ndjeunga J, Bezancon G. 2008. Niger-wide assessment of *in situ* sorghum genetic diversity with microsatellite markers. Theoretical and Applied Genetics 116:903-913.

Dje Y, Ater M, Lefebvre C, Vekemans X. 1998. Patterns of morphological and allozyme variation in sorghum landraces of northwestern Morocco. Genetic Resources and Crop Evolution 45:541-548.

Dje Y, Forcioli D, Ater M, Lefebvre C, Vekemans X. 1999. Assessing population genetic structure of sorghum landraces from North-western Morocco using allozyme and microsatellite markers. Theoretical and Applied Genetics 99:157-163.

Dogget H. 1988. Sorghum. Longman Scientific and Technical Essex, England, pp. 512.

Dogget H, Majisu BN. 1968. Disruptive selection in crop development. Heredity 23:1-23.

Dogget H, Prasada Rao KE. 1995. Sorghum. In: Smartt J, Simmonds NW (Eds.), Evolution of Crop Plants. Longman Group, Burnt Mill. pp. 180.

Ellstrand NC. 1992. Geneflow by pollen: Implications for plant conservation genetics. Oikos 63:77-86.

Epperson BK. 1993. Recent advances in correlation analysis of spatial patterns of genetic variation. Evolutionary Biology 27:95-155.

Epperson BK. 2004. Multilocus estimation of genetic structure within populations. Theoretical Population Biology 65:227-237.

Evanno S, Regnaut S, Goudet J. 2005. Detecting the number of cluster of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611-2620.

Excoffier L, Laval LG, Schneider S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131:479-491.

FAO. 2008. FAOSTAT. http://faostat.fao.org.

Folkertsma RT, Frederick H, Rattunde HFW, Chandra S, Raju GS, Hash CT. 2005. The pattern of genetic diversity of Guinea-race *Sorghum bicolor* (L.) Moench landraces as revealed with SSR markers. Theoretical and Applied Genetics 111:399-409.

Frankel OH, Brown AHD, Burdon JJ. 1995. The conservation of plant diversity. Cambridge University Press. Cambridge, UK, pp. 299.

Frankel OH, Hawkes JG. 1975. Crop genetic resources for today and tommorrow. Cambridge University Press Cambridge, UK, pp. 492.

Fukunaga K, Hill J, Vigouroux Y, Matsuoka Y, Sanchez G, Liu K, Buckler ES, Doebley J. 2005. Genetic diversity and population structure of teosinte. Genetics 169:2241-2254.

Gepts P. 2004. Crop domestication as a long-term selection experiment. In: Jannick J (Ed.), Plant Breeding Reviews, Volume 24, Part 2: Long-term selection: Crops, Animals, Bacteria. John Wiley & Sons, Inc. pp. 23.

Ghebru B, Schmidt RJ, Bennetzen JL. 2002. Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. Theoretical and Applied Genetics 105:229-236.

Goudet J. 2002. FSTAT, a program to estimate and test gene diversity and fixation indices. (version 2 9 3 2).

Gurney AL, Press MC, Scholes JD. 2002. Can wild relatives of sorghum provide new sources of resistance or tolerance against *Striga* species? Weed Science 42:317-324.

Hardy OJ, Vekemans X. 2002. SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology 2:618-620.

Harlan JR, De Wet JMJ. 1972. A simplified classification of cultivated sorghum. Crop Science 12:172-177.

Harlan JR, Stemler ABL. 1976. The races of sorghum in Africa. In: Harlan JR, De Wet JMJ, Stemler ABL (Eds.), Origins of African Plant Domestication. Mouton, The Hague, Paris. pp. 478.

Hartl DL, Clark G. 1997. Principles of population genetics. Sinauer Associates, Inc. Sunderland, pp. 452.

Haygood R, Ives AR, Andow DA. 2003. Consequences of recurrent geneflow from crops to wild relatives. Proceedings of the Royal Society of London 270:1879-1886.

Hulbert SH. 1971. The nonconcept of species doversity: a critique and alternative parameters. Ecology 52:577-586.

Kalinowski S. 2005. HP-RARE 1.0: a computer program for performing rarefaction on measures of allelic richness. Molecular Ecology 5:187-189.

Kamala V, Singh SD, Bramel PJ, Rao DM. 2002. Sources of resistance to downy mildew in wild and weedy sorghums. Crop Science 42:1357-1360.

Ladizinsky G. 1999. Plant evolution under domestication. Kluwer Academic Publishers London, pp. 254.

Levin DA, Kerster HW. 1974. Geneflow in seeds plants. Evolutionary Biology 7:139-220.

Liang GHL, Casady AJ. 1966. Quantitative presentation of the systematic relationships among twenty-one sorghum species. Crop Science 6:76-79.

Mace EM, Buhariwalla HK, Crouch JH. 2003. A high-throughput DNA extraction protocol for tropical molecular breeding programs. Plant Molecular Biology Reporter 21:459a-459h.

Mariac C, Luong V, Kapran I, Mamadou A, Sagnard F, Deu M, Chantereau J, Gerard B, Ndjeunga J, Bezanon G, Pham JL, Vigouroux Y. 2006. Diversity of wild and cultivated pearl millet accessions (*Pennisetum glaucum* [L.] R. Br.) in Niger assessed by microsatellite markers. Theoretical and Applied Genetics 114:49-58.

Matsuoka Y, Vigouroux Y, Goodman M-M, Sanchez J, Buckler E, Doebley J (2002) A single domestication for maize shown by multilocus microsatellite genotyping. Proceeding of the National Academy of Sciences of the USA (PNAS) 99:6080–6084

Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061-1064.

Morden CW, Doebley JF, Schertz KF. 1990. Allozyme variation among the spontaneous species of *Sorghum* section Sorghum (Poaceae). Theoretical and Applied Genetics 80:296-304.

Morgan WTW. 1974. The south Turkana expedition: Scientific papers X. Sorghum gardens in south Turkana: Cultivation among a nomadic pastoral people. The Geographical Journal 140:80-93.

Murty BR, Arandchalam V, Saxena MBL. 1967. Classification and catalogue of world collection of *Sorghum*. Indian Journal of Genetics and Plant Breeding 27:1-74.

Mutegi E, Sagnard F, Muraya M, Kanyenji B, Rono B, Mwongera C, Marangu C, Kamau J, Parzies H, de Villiers S, Semagn K, Traore PS, Labuschagne M. 2009. Ecogeographical distribution of wild, weedy and cultivated *Sorghum bicolor* (L.) Moench in Kenya: implications for conservation and crop-to-wild gene flow. Genetic Resources and Crop Evolution (DOI 10.1007/s10722-009-9466-7)

Neal D. 2004. Introduction to population biology. Cambridge University Press Cambridge, UK, pp. 393.

Nkongolo KK, Nsapato L. 2003. Genetic diversity in *Sorghum bicolor* (L.) Moench accessions from different ecogeographical regions in Malawi accessed with RAPDs. Genetic Resources and Crop Evolution 50:149-156.

Perrier X, Flori A, Bonnot F. 2003. Data analysis methods. In: Hamon P, Seguin M, Perrier X, Glaszmann JC (Eds.), Genetic diversity of cultivated tropical plants. Enfield, Science Publishers, Montpellier. pp. 76.

Prasanth V, Chandra S, Jayashree B, Hoisington D. 2006. AlleloBin - A program for allele binning of microsatellite markers based on the the algorithm of Idury and Cardon (1997). ICRISAT International Crops Research Institute for the Semi-Arid Tropics.

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

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org).

Rao Kameswara N, Reddy LJ, Bramel PJ. 2003. Potential of wild species for genetic enhancement of some semi-arid food crops. Genetic Resources and Crop Evolution 50:707-721.

Reed JD, Ramundo BA, Claflin LF, Tuinstra MR. 2002. Analysis of resistance to ergot in sorghum and potential alternate hosts. Crop Science 42:1135-1138.

Rich PJ, Grenier U, Ejeta G. 2004. Striga resistance in the wild relatives of sorghum. Crop Science 44:2221-2229.

Ritland K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research 67:175-185.

Rousset F. 2000. Genetic differentiation between individuals. Journal of Evolutionary Biology 13:58-62.

Sagnard F, Barnaud A, Deu M, Barro C, Luce C, Billot C, Rami JF, Bouchet S, Dembele D, Pomies V, Calatayud C, Rivallan R, Joly H, Brocke KV, Toure A, Chantereau J, Bezancon G, Vaksmann M. 2008. Multi-scale analysis of sorghum genetic diversity: Understanding the evolutionary processes for *in situ* conservation. Cahiers Agricultures 17:114-121.

Schuelke M. 2000. An economic method for the fluorescent labelling of PCR fragments. A poor man's approach to genotyping for research and high throughput diagnostics. Nature Biotechnology 18:233-234.

Snow AA, Moran-Palma P. 1997. Commercialization of transgenic plants: potential ecological risks. BioScience 47:86-96.

Sokal RR, Oden NL. 1978. Spatial autocorrelation in biology 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10:249.

Sombroek WC, Braun HMH, van der Pour BJA. 1982. Explanatory soil map and agro-climatic zone map of Kenya. Report E1:1-56.

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J, Marx D, Bothma G, Ejeta G. 2008. The potential for crop-to-wild geneflow in sorghum in Ethiopia and Niger: A geographic survey. Crop Science 48:1425-1431.

Thies JE, Devare MH. 2007. An ecological assessment of transgenic crops. Journal of Development Studies 43:97-129.

Uptmoor R, Wenzel W, Friedt W, Donaldson G, Ayisi K, Ordon F. 2003. Comparative analysis on the genetic relatedness of *Sorghum bicolor* accessions from Southern Africa by RAPDs, AFLPs and SSRs. Theoretical and Applied Genetics 106:13161-1325.

VSN International Ltd. 2007. GenStat Discovery Edition 3. VSN International Ltd. Hernel Hempstead, UK.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Wright S. 1951. The genetical structure of populations. Annals of Eugenics 15:323-354.

Wright S. 1978. Evolution and the genetics of populations: Variability within and among natural populations. University of Chicago Press Chicago, pp. 590.

Chapter 4

Estimation of the extent of crop-to-wild geneflow in sorghum at local scale in Kenya

4.1 Abstract

The extent and direction of geneflow in the wild-weedy-domesticate complex of *S. bicolor* was investigated at a local scale within traditional farming systems in the eastern slopes of Mt. Kenya, Meru South district. A total of 483 individuals consisting of 15 wild-weedy sorghum and 22 cultivated sorghum populations were genotyped using 10 polymorphic SSR markers. Two complimentary approaches were used to analyse geneflow. The Bayesian model-based admixture method showed recent geneflow between cultivated and wild sorghum to be asymmetric, with the magnitude being almost nine-times higher from the crop to its wild-weedy relatives than *vice versa*. This could be explained by three non-mutually exclusive factors: on-farm population size differences between the crop and its wild-weedy relatives that favour pollen flow from the former to the latter, negative selection by farmers against wild-weedy sorghum alleles in maternal crop parents, and/or natural outcrossing differences between cultivated and wild-weedy sorghum. Furthermore, significant differences were revealed among farms in the extent of crop-to-wild geneflow. The F-statistics based analysis further revealed intermediate levels of historical/long-term geneflow (Nm = 0.66) between cultivated and wild sorghum, with highly significant and substantial differentiation between the two congeners (F_{ST} = 0.27; $P \leq 0.001$). Overall, the findings in this study suggest that wild-weedy relatives growing sympatrically with GM sorghum will most likely act as bridges for the escape and diffusion of transgenes into the surrounding cultivated and natural habitats. The findings are important for informing decisions on biosafety regulations and guidelines both for testing and releasing transgenic sorghum in Kenya's traditional farming systems.

4.2 Introduction

The global area planted with GM crops has continued to grow steadily over the last decade. In 2008 alone, the global area of GM crops grew by 9.4% to 125 million hectares. In the same year significant progress was achieved in Africa, with Egypt and Burkina-Faso joining South Africa in commercialising GM crops (James 2008). This continued increase has been driven by a variety of potential benefits, including increased yields, easier and more effective weed/pest control, lower consumer prices, a wider variety of produce available throughout the year and production of nutrient enriched staple crops. These benefits notwithstanding, growth in wide-scale commercialisation of GM crops has been accompanied by equally increasing concerns, among other, potential harmful effects to the environment. One of the major environmental concerns is the possibility that transgenes will escape via geneflow into sympatric wild and/or weedy populations and depending on the nature of the engineered traits, lead to harmful agricultural and/or ecological consequences. In agricultural lands, it is feared that escape of transgenes may confer yield enhancement or resistance to herbicides, diseases or insect pests, potentially altering the fitness of non-cultivated species. This might render existing weeds more aggressive and difficult to control or enhance weediness in species that are currently not considered noxious weeds. In natural habitats, the concern is that crop-wild hybrids harbouring enhanced fitness may become invasive, potentially leading to genetic erosion and in the worst case scenario, the extinction of entire recipient populations through demographic and/or genetic swamping. It is argued that the resulting economic and environmental damage could potentially equal or by far outweigh any economic benefits achieved through genetic transformation of the crops (Rogers and Parkes 1995; Snow and Moran-Palma 1997; Conner *et al.* 2003; Hancock 2003; Haygood *et al.* 2003; Celis *et al.* 2004; Poppy 2004; Hails and Morley 2005; Chapman and Burke 2006; Cantamutto and Poverene 2007; Auer 2008; Chandler and Dunwell 2008; Hooftman *et al.* 2008; Kumar *et al.* 2008; Schoen *et al.* 2008).

For geneflow to take place between a crop and its wild-weedy relative, four factors must be satisfied, at least in part: (i) overlap in distribution range, (ii) overlap in flowering times, (iii) shared pollinators and (iv) cross-compatibility. Although many crops are known to co-occur with their wild and weedy relatives and hybrids between them have been reported, hybridisation alone does not correspond to geneflow. Knowledge on introgression, the permanent incorporation of genes from one set of differentiated

populations into another through recurrent crossing, is further required in order to establish the extent to which transgenes are likely to persist in wild-weedy populations. This is because introgression of crop genes in wild populations beyond F_1 hybrids might be prevented or slowed down by hybrid sterility or breakdown in early hybrid generations (Ellstrand *et al.* 1999; Jarvis and Hodgkin 1999; Ellstrand 2003; Armstrong *et al.* 2005; Warwick and Stewart 2005). In general, introgression is governed both by internal genetic interactions determining reproductive compatibility and external (ecological) factors affecting selection on hybrids in their environment (Arnold 1997; Jenczewski *et al.* 2003).

The enhanced possibility of repeated hybridisation cycles leading to allele introgression from crops to wild and/or weedy relatives in centres of origin and diversity for specific crops has long been recognised (Anderson 1949; De Wet and Harlan 1975; Arnold 1997). These events occur first, due to overlapping distribution and secondly, as a result of permeable reproductive barriers between crop species and their progenitors. Domestication is thought to be a recent process in evolutionary scales, for complete reproductive isolation to have taken place between crops and their progenitors (Simmonds 1992). Cultivated sorghum (*S. bicolor* ssp. *bicolor*) originated in Africa (Harlan and Stemler 1976) and still occurs sympatrically with its proposed wild progenitor (*S. bicolor* ssp. *verticilliflorum*). The two are thought to be reproductively inter-fertile, the product of which is a highly heterogeneous hybrid derivative. Stabilised forms of the putative hybrid derivative have been classified under a separate subspecies as *S. bicolor* ssp. *drummondii* (Harlan and De Wet 1972; De Wet 1978) and are known to co-occur as persistent weeds with the wild types in sorghum and other cereals fields and in intermediate habitats such as fallows, field margins and roadsides in many parts of Africa, including Kenya (Dogget and Majisu 1968; Dogget 1988; Tesso *et al.* 2008; Mutegi *et al.* 2009). The three subspecies of *S. bicolor* are thought to form an inter-fertile crop-wild-weed complex in traditional agro-ecosystems of sorghum in Africa (De Wet 1978; Dogget 1988; Tesso *et al.* 2008) and are likely conduits for escape and spread of transgenes from GM sorghum (Arriola and Ellstrand 1996; 1997; Morrell *et al.* 2005).

Notable progress has been made towards developing and optimising sorghum genetic transformation protocols (Casas *et al.* 1993; 1997; Zhao *et al.* 2000; Gao *et al.* 2005; Howe *et al.* 2006). Successful genetic engineering of sorghum has been reported for chitosanase and/or chitinase gene against fungal diseases (Zhu *et al.* 1998; Krishnaveni *et*

al. 2000; Ayoo 2008), Bt genes against stalk borer (Girijashankar *et al*. 2005) and alpha-hordothionin protein gene originating from barley (*H. vulgare*) for high lysine content (Zhao *et al.* 2000). It is of interest to note that Kenyan sorghum landraces were used in the work of Ayoo (2008). Additional efforts to transform sorghum are underway, prominent among them, the initiative by the African Biofortified Sorghum project whose aim is to deploy a nutritionally enhanced and more digestible transgenic sorghum to subsistence farmers in Africa (Zhao 2008). There exists an urgent need therefore to characterise the extent and patterns of geneflow between cultivated sorghum and its closest wild-weedy relatives within the context of an African traditional sorghum agro-ecosystem.

Substantial work has been undertaken on the potential for spontaneous hybridisation and introgression between cultivated sorghum, *S. bicolor* and its noxious weedy congener, *S. halepense* or johnsongrass (Arriola and Ellstrand 1996; 1997; Morrell *et al.* 2005). Arriola and Ellstrand (1996) documented evidence of spontaneous crop-to-weed hybridisation between *S. bicolor* and *S. halepense* under experimental field conditions. They determined that incidences and rates of hybridisation vary with distance of the weed from the crop, location of study site and year of study. The authors concluded that transgenes introduced into crop sorghum can escape cultivation through interspecific hybridisation with johnsongrass. In order to provide an insight into the potential fate of transgenes in weedy populations, Arriola and Ellstrand (1997) compared fitness of johnsongrass x crop sorghum hybrids with that of non-hybrid johnsongrass under agricultural conditions. No significant increase or decrease in fitness was observed for hybrid weeds, a result that led them to reason that transgenes that are neutral or beneficial to johnsongrass would most likely persist in populations growing in agricultural conditions under continued geneflow from the crop. These conclusions were supported by the study of Morrell *et al*. (2005), which went a step further to investigate the extent of introgression or long-term persistence of sorghum cultivar alleles in weedy populations of johnsongrass. Putative cultivar-specific alleles were found to be present in up to 32.3% of individuals in johnsongrass populations with a history of exposure to cultivated sorghum, an indication that introgression between cultivated and its weedy relative does occur and is frequent (Morrell *et al*. 2005). Besides, lower frequencies of cultivar-specific alleles were observed at a smaller number of loci in populations of johnsongrass without recent exposure to cultivated sorghum, a further suggestion that introgressed cultivar alleles may

disperse across long distances (Morrel *et al.* 2005). The authors concluded that johnsongrass populations potentially offer conduits for escape and widespread dissemination of transgenes from cultivated sorghum. From these series of experiments it is plausible to conclude that, depending on the nature of the engineered trait, introduction of transgenic sorghum into agricultural lands already invaded by johnsongrass, is a serious biosafety risk.

Such detailed crop-to-wild geneflow risk assessment studies are yet to be reported for the domesticated sorghum and its African closest wild and weedy relatives. With the now growing possibility of transgenic sorghum deployment on the continent (e.g. Zhao 2008), however, significant scientific attention is being directed towards the subject as evidenced by studies in a number of countries (*www.ifpri.org/pbs/pdf/bbiprojects.pdf*). For example, Tesso *et al.* (2008) examined the current distribution of cultivated sorghum and its wild and weedy relatives in Ethiopia and Niger. Cultivated sorghum was found to overlap its distribution range and flowering dates with its wild-weedy relatives in many growing regions in the two countries. Many putative hybrids were observed, especially in Ethiopia. The authors anticipated current gene transfer from cultivated sorghum to wild and weedy sorghum populations in Ethiopia and Niger to be widespread (Tesso *et al.* 2008). Similarly, in a study closely related to the present one, Mutegi *et al.* (2009) more recently reported widespread co-occurance of cultivated and wild-weedy sorghum in all major growing regions of Kenya. They documented putative crop-wild hybrids to be frequently present in the sorghum growing regions. The authors conluded that crop-wild geneflow in sorghum is likely to occur in many agro-ecosystems of Kenya.

In the present study, microsatellite markers were used to analyse cultivated and wild-weedy sorghum samples collected at a local scale within a heterogeneous traditional sorghum agro-ecosystem in the eastern parts of Kenya, with a view of estimating the extent and direction of introgression between the two congeners. Outputs from this study should offer insight to biosafety regulators on the potential geneflow risks for testing and/or releasing transgenic sorghum within the crop's traditional cropping systems in Kenya. This will be achieved by answering three main questions:

(i) Is there geneflow between cultivated and wild sorghum?

(ii) If so, what is the extent and direction of introgression between the two taxa?

(iii) Does the rate of introgression vary among farms?

4.3 Materials and methods

4.3.1 Study site selection

An 8 km x 8 km intensive study site (ISS) was established on the eastern slopes of Mt. Kenya, on the easternmost limit of Meru South administrative district. Selection of the site was motivated by among other factors, the importance of sorghum to the farming communities, occurrence of wild and weedy sorghums within and around cropping fields, great environmental heterogeneity and variability of farmers' cropping systems and practices. The ISS is home to four ethno-linguistic groups, namely the *Tharaka*, *Mwimbi*, *Chuka* and *Muthambi*. Farming in the region is exclusively rain fed, small scale and largely for subsistence. The site experiences a bi-modal rainfall regime, with the short, more reliable rain from October to December and long, less reliable rain from March to May. Subsequently, the short rains provide the major cropping season with planting around November and harvesting in late January/February, while the long rains correspond to more or less the minor season whose planting is in late February/March and harvesting in July/August. Moreover, the site is characterised by an altitudinal gradient which ranges between 1050 to 750 masl and correspondingly from the maize-tobacco cropping zone to the sorghum-pearl millet cropping zone. Thus although sorghum is cultivated in the entire ISS, its importance diminishes with increased altitude. Two main types of sorghum varieties are grown within the ISS: (i) single season varieties that complete their cycle within one season and can thus be planted twice a year and (ii) the ratooning varieties that are planted in the short rain season, cut approximately 20 cm above ground at the end of the same season, before completing the cycle in the subsequent long rain season.

4.3.2 Household selection and sample collection

In order to establish a protocol for collecting ISS representative samples of cultivated and wild-weedy sorghum populations, farmer interviews were conducted in 372 households across the ISS. Information on among other geographical coordinates, elevation, varieties of all crops grown and on presence and abundance of wild-weedy populations was collected. Farm-level crop and variety diversity information was subsequently used to select 44 households, representative of the cropping system variability within the ISS. The selected households were visited for quantitative categorisation data and sample collection. First, a hand-held GPS system (Geoxm 2005, Trimble) was used to map

farmers' fields in each of the selected households. Subsequently, two, 1 m wide transects were performed across each mapped field in different directions to record data on abundance of crop sorghum varieties and wild-weedy sorghum. Twenty two landraces of cultivated sorghum (each consisting of five panicles) corresponding to the ISS varietal diversity were collected alongside 15 sympatric wild-weedy sorghum populations (each consisting of 25 panicles) in 13 out of the 44 farms. Two more wild sorghum populations, each consisting of 25 panicles were collected from near natural habitats next to two of the farms. For each of the wild-weedy sorghum populations, single, mature panicles were collected from 25 mother plants selected randomly across each of the habitats: sorghum field, fallow field and semi-natural. For each of the 22 sympatric cultivated sorghum samples in each site, five mature panicles of each farmer named variety were randomly sampled from the fields. In total, 110 wild-weedy sorghum individuals and 373 individuals of cultivated sorghum were analysed.

4.3.3 DNA extraction, PCR amplification and genotyping

From each of the collected panicles of cultivated and wild sorghum populations, five randomly selected seeds were planted in the laboratory and subsequently one seedling per panicle used for DNA extraction. Growth conditions, DNA extraction, PCR amplification and genotyping were performed as described in Chapter 3. The 10 SSR markers used to analyse the samples are shown in Table 4.1. The markers were a subset of the 24 SSRs used in Chapter 3, selected based on two criteria: (i) ability to genetically differentiate country-scale cultivated and wild sorghum pools ($F_{ST} \geq 0.02635$) or (ii) high polymorphic information content (PIC) values (≥ 0.70540) (Table 4.1).

Table 4.1 List of microsatellite loci used in the assay

Locus name	Repeat Motif	[a]F_{ST}	[b]PIC
Xcup53	$(TTTA)_5$	0.27098	0.27775
Xtxp278	$(TTG)_{12}$	0.16386	0.33641
mSbCIR246	$(CA)_{7.5}$	0.10013	0.38927
mSbCIR248	$(GT)_{7.5}$	0.08459	0.67394
Xtxp57	$(GT)_{21}$	0.02635	0.86278
Xtxp12	$(CT)_{22}$	0.00940	0.91309
SbAGB02	$(AG)_{35}$	0.00734	0.70540
Xtxp320	$(AAG)_{20}$	0.00359	0.84515
mSbCIR238	$(AC)_{26}$	0.00345	0.84421
Xgap206	$(AC)_{13}/(AG)_{20}$	0.01730	0.86687

[a] = Fixation index; the degree of differentiation among sub-populations, [b] = Polymorphic information content.

4.3.4 Data analysis

4.3.4.1 Bayesian admixture approach

The extent and direction of introgression between cultivated and wild sorghum was estimated using the Bayesian model-based method implemented in the programme STRUCTURE (Pritchard *et al.* 2000; Falush *et al.* 2003). The programme uses Bayesian methods with Monte Carlo Markov Chain (MCMC) to infer the population structure of individual genotypes in a sample by assuming a model in which K populations (where K may be unknown) are each characterised by a set of allele frequencies at each locus. Individuals are assigned to discrete populations, or jointly to two or more populations if their genotypes indicate they are admixed. The model assumes that within populations, the loci are at HWE and LE. Principally, individuals are assigned to K populations in such a way as to achieve this. Besides its application in demonstrating the presence of population structure, STRUCTURE can be used to assign individuals to populations and to identify migrants and admixed individuals, thus the extent and direction of introgression among populations. STRUCTURE estimates the ancestry of individuals using four main models: (i) no admixture model assumes that individuals are distinctly from one population or another, (ii) the admixture model allows individuals to be of mixed ancestry and estimates the proportion of each individuals genome that can be traced to each of the inferred K populations, (iii) the linkage model is a generalised admixture model to deal with "admixture linkage disequilibrium" or the correlations that arise between linked markers in recently admixed populations and (iv) the prior

information model which allows some or all individuals to be assigned to pre-defined populations (e.g. biological categories) and is unlike the default mode which uses only genetic information to learn about population structure (Pritchard *et al.* 2000; Falush *et al.* 2003).

To investigate the extent of divergence between cultivated and wild-weedy gene pools, the population structure in the entire SSR data set was first analysed using the default parameter settings of STRUCTURE 2.2. The models used assumed 'admixture ancestry', uncorrelated allele frequencies and assigned no pre-defined populations (cultivated and wild-weedy) to the individuals. The analysis assumed populations (K) from K = 1 to K = 10. The method of Evanno *et al.* (2005) was used to determine the most likely number of clusters. This approach uses an *ad hoc* statistic, ΔK, based on the second-order rate of change of the posterior probability of K (PX|K) between successive K values. Using this approach the best population divergence was obtained at K = 2, corresponding to the cultivated and wild-weedy sorghum gene pools. The level and direction of admixture between the cultivated and wild-weedy sorghum gene pools was consequently inferred through an additional STRUCTURE run at K = 2, with the population information (cultivated and wild-weedy) for each individual being considered (USEPOPINFO option). The prior probability that each individual had pure ancestry from its assigned population (cultivated or wild-weedy) was set at 0.95 using the option MIGRPRIOR = 0.05. To estimate the hypothesis that an individual had mixed ancestry within the past three generations, the option GENSBACK = 3 was set. Specifically, this investigated the probability that an individual had a parent, a grandparent or a great grandparent in the alternate population. The extent of the admixture between cultivated and wild-weedy sorghum was further estimated and compared among three selected farms. This was in an attempt to answer the question whether geneflow between the two sorghum congeners varied among farms. Each of the selected farms had a minimum of three cultivar types (or landrace names) and at least one co-occurring wild-weedy sorghum population. The landraces in each farm were pooled into a single population and analysed in STRUCTURE against the sympatric population(s) of wild-weedy sorghum. All STRUCTURE analyses were performed using the high performance computing resources of the CBSU from Cornell University (http://cbsuapps.tc.cornell.edu/structure.aspx). The initial analysis included 10 runs per K value, whereas the subsequent runs at K = 2

included five runs per K value. The programme was each time run with a burn-in of 500000 MCMC iterations followed by 10^6 iterations of data collection.

The extent and direction of introgression (ISS and farm-level) was investigated by assessing the average proportion of membership (Q_i) of the sampled populations (i.e. ISS-level: cultivated vs. wild-weedy; farm-level: landrace pool vs. wild-weedy) in each of the inferred clusters at K = 2. The proportion of an individual's genome belonging to each of the clusters, q_i, was used to assess the assignment of cultivated and wild-weedy sorghum genotypes. An individual was assumed to be assigned to a cluster if it had at least 95% of its genome ($q_i \geq 0.95$) in the particular cluster, with individuals with less ($q_i < 0.95$) assumed to be jointly assigned to the clusters (due to admixture or hybridisation). The proportion of wild-weedy sorghum individuals' genome in the cluster associated with cultivated sorghum (cluster one) was used as a measure of crop-to-wild geneflow. The non-parametric Kruskal-Wallis test was used in the programme GenStat (VSN International Ltd. 2007) to test the null hypothesis of no difference in the extent of crop-to-wild geneflow among the four selected farms. Pairwise comparisons of crop-to-wild geneflow among the four farms was summarised graphically using notched box plots in the programme R (R Development Core Team 2007). In addition to providing a visual appraisal of data, notched box plots allow for approximate hypothesis testing. Specifically, if the notches about the medians do not overlap in a side by side comparison of two or more box plots, the medians are, roughly significantly different at about 95% confidence interval (McGill *et al.* 1978). Values of $p_{1,}$ the mean proportion of wild-weedy sorghum individual's genome in cluster one, were first normalised through log transformation before generating the box plots.

4.3.4.2 Fixation index (F_{ST}) - based analysis

The fixation index of population sub-division relative to the total population (F_{ST}) was used to estimate historical geneflow between cultivated and wild sorghum using the programme GENETIX (Belkhir *et al.* 2004), according to the equation $Nm \approx 0.25(1-F_{ST})/F_{ST}$ (Wright 1951). The parameter *Nm* infers the average number of effective migrants per generation.

4.3.4.3 Genetic structure

The genetic differentiation among populations in cultivated and wild-weedy sorghum genotypes at household/parcel level was analysed using three approaches: (i) F-statistics [F_{ST}, F_{IS}, F_{IT}; (Weir and Cockerham 1984)] was estimated using GENETIX (Belkhir *et al.* 2004), (ii) an analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992; Michalakis and Excoffier 1996) was performed using ARLEQUIN 3.11 (Excoffier *et al.* 2005) and spatial autocorrelation analysis (Sokal and Oden 1978a; 1978b) using the relative kinship coefficient (r_{ij}) method as implemented in the programme SPAGeDi (Hardy and Vekemans 2002).

4.4 Results

4.4.1 Sample collection

In total, 15 wild-weedy populations consisting of 25 panicles each and 22 populations of at least 12 farmer-named varieties of cultivated sorghum, each consisting of five panicles, were collected (Figure 4.1 and Table 4.2). Figure 4.1 shows the geographic distribution of the sampled sites, together with the household-level perception on the density of wild-weedy sorghum. Wild-weedy sorghum was present in the majority (89.5%) of the 372 farms, with farmers perceiving the density as either absent (10.5% of the respondents), abundant (13.7% of the respondents) or scarce (75.8% of the respondents). However, no clear spatial trend on the wild-weedy sorghum presence was observed (Figure 4.1).

Table 4.2 presents a list of the sampled cultivated and wild sorghum populations, with the associated provenance information. Most of the wild-weedy populations were collected from either sorghum fields (66%) or the crops fallow fields (20%). Semi-natural habitats in the neighbourhood of cultivated fields were represented by only two wild-weedy populations: (i) a riverine habitat located some 3 m from the nearest sorghum field and (ii) a grassland habitat located some 20 m away from the nearest sorghum. The farmers from sites where samples were collected from grew between one and six sorghum landraces with an average of three. Two traditional landraces were the most commonly grown: (i) “Kathirigwa”, a single season variety and (ii) “Mugana” a dual season/ratoon sorghum type. Populations were sampled from an altitudinal range of 817 - 1032 masl (Table 4.2).

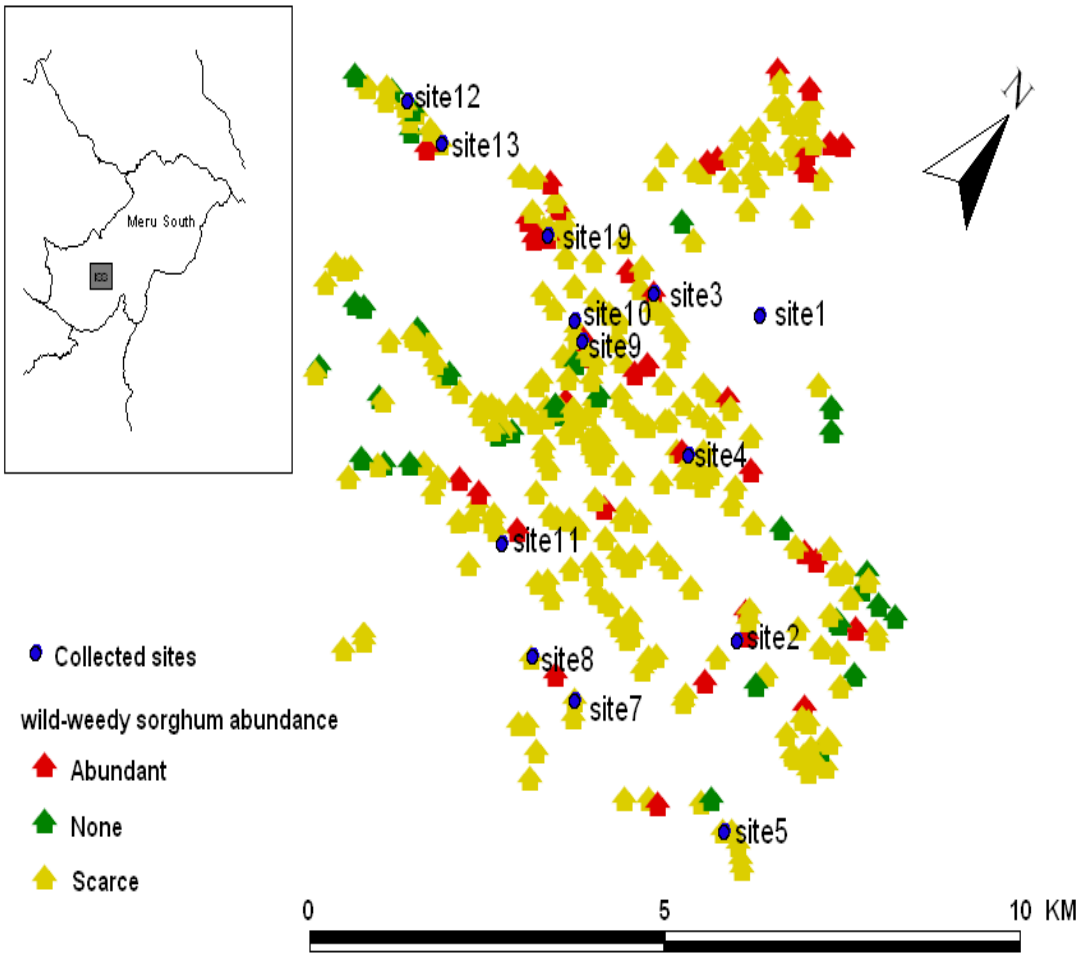

Figure 4.1 **Map of the intensive study site showing distribution of the 372 visited households, farmer perception on wild sorghum abundance and collection sites for studied populations.**

Table 4.2 List of collected cultivated and wild sorghum populations

Site ID	Collection No.	Population	[1]Types	[2]Variety names	[3]Habitat	Altitude (masl)
1	IS2	1	W		F	841
	IS3	2	C	Ciamaguru		
	IS4	3	C	Muruge		
	IS5	4	C	Kaguru		
	IS6	5	C	Mucuri		
	IS7	6	C	Kathirigwa		
	IS8	7	C	Mugana		
2	IS20	8	W		SF	820
3	IS25z	9	W		SF	884
	IS26	10	C	Musalama		
	IS27	11	C	Kaguru		
	IS28	12	C	Muruge		
	IS29	13	C	Mugana		
4	IS30	14	W		F	846
	IS32	15	C	Mugeta		
	IS33	16	C	Mugana		
5	IS36	17	W		SF	817
6	IS39	18	W		SF	840
7	IS43	19	W		SF	869
	IS44	20	C	Kathirigwa		869
8	IS45	21	C	Kathirigwa		890
	IS46	22	C	Serena		
	IS47	23	W		SF	
9	IS51	25	W		F	874
10	IS53	25	C	Kathirigwa		869
	IS54	26	W		SF	
11	IS55	27	W		SF	884
	IS56	28	C	Mugana		
	IS57	29	C	Munyerege		
	IS58	30	C	Kathirigwa		
12	IS60	31	W		SN-R	1032
13	IS61	32	W		SF	1012
19	IS63	33	W		SF	936
	IS64	34	C	Mugana		
	IS66	35	W		SN-G	
	IS65A	36	C	Kathirigwa		
	IS65B	37	C	Kathanta		

[1]Sorghum type: C-cultivated, W-wild-weedy; [2]Farmer-named cultivar names; [3]Habitat: F-Fallow, SF-Sorghum field, SN-R – Semi natural riverine, SN-G – Semi natural grassland.

4.4.2 Recent geneflow

The initial STRUCTURE analysis identified K = 2 to be the most appropriate number of populations, according to the method described by Evanno and co-workers (2005) (Figure 4.2).

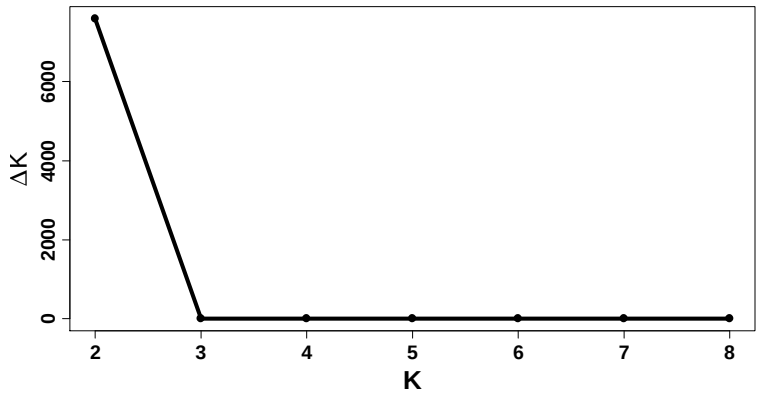

Figure 4.2 Evanno's ΔK statistic for K = 2 to K = 8. The modal value is at K = 2.

The proportion of each individual's genome in each of the two clusters (q_i) is presented in the form of a bar plot in Figure 4.3. Cluster 1 (red) corresponded to the cultivated sorghum gene pool, with 92% of the individuals having a $q_i \geq 0.95$, whereas cluster 2 corresponded to the wild-weedy sorghum gene pool with 53% of the individuals having a $q_i \geq 0.95$. These results indicate that the wild-weedy populations consisted of more admixed individuals (close to six-fold higher) than their cultivated counterparts.

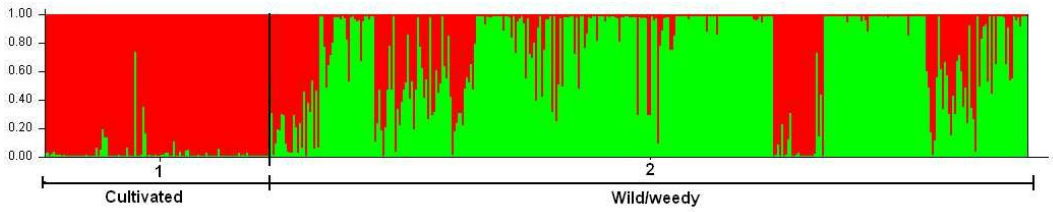

Figure 4.3 Bar plot of the estimated genetic structure at K = 2 using the default STRUCTURE parameters with the individuals ordered by sorghum type. Each individual is represented by a vertical line which is partitioned into coloured segments that represent its proportion of genome in K (coloured) clusters.

The mean proportion of estimated ancestry in each of the two genetic groups (cluster one and two) for the cultivated and wild sorghum gene pools is presented in Table 4.3. For the wild-weedy sorghum gene pool, the mean proportion of estimated ancestry in cluster one (corresponding to the cultivated sorghum pool) was substantial (Q_1 = 0.278). Contrastingly, the mean proportion of ancestry for cultivated sorghum in cluster two (corresponding to the wild-weedy sorghum pool) is 0.032. These results indicate that the inferred local scale geneflow was almost nine times higher from cultivated sorghum populations to the sympatric wild-weedy populations than *vice versa*.

Table 4.3 Mean proportion of estimated ancestry in each of the K = 2 clusters for cultivated and wild sorghum gene pools

Population	Q_1 cultivated pool (cluster one)	Q_2 wild pool (cluster two)
Cultivated	0.968	0.032
Wild-weedy	0.278	0.722

Farm-level estimates of the mean proportion of ancestry in cluster one and two for the pool of sorghum landraces and the co-occurring wild-weedy relative population(s) are presented in Table 4.4. The mean proportion of wild-weedy sorghum genome originating in cultivated sorghum (cluster 1) ranged from Q_1 = 0.091 to Q_1 = 0.583. The Kruskal-Wallis test rejected the null hypothesis of no differences in the level of crop-to-wild geneflow among the four selected farms (H = 42.2; P < 0.001). The mean proportion of cultivated sorghum genome estimated to originate from the wild-weedy sorghum gene pool (cluster 2) (a measure of wild-to-crop geneflow) was low ($0.001 \leq Q_2 \leq 0.066$) (Table 4.4). These results confirmed the ISS scale observation that geneflow was higher from cultivated sorghum to its wild-weedy relatives than *vice versa* (Table 4.3).

Table 4.4 Farm-level mean proportion of estimated ancestry (Q_i) for the pool of cultivated sorghum and the co-occurring wild-weedy sorghum population(s)

Farm/site	[1]Sorghum type	Number of individuals	Distance from cultivated sorghum	Mean proportion of individual genome (Qi)	
				Cluster 1	Cluster 2
	Cultivated	30		0.934	0.066
1	Wild-weedy (F)	25	5 m	0.091	0.909
	Cultivated	20		0.999	0.001
3	Wild-weedy (SF)	25	Within	0.382	0.618
	Cultivated	15		0.999	0.001
11	Wild-weedy (SF)	25	Within	0.576	0.424
	Cultivated	15		0.999	0.001
	Wild-weedy (SN-G)	25	20 m	0.214	0.786
19	Wild-weedy (SF)	25	Within	0.583	0.417

[1] Wild-weedy sorghum habitat in parentheses: F - Fallow field, SF - Sorghum field and SN-G - semi-natural grassland.

Figure 4.4 summarises the pairwise differences among the selected farms in the mean proportion of genome originating from cultivated sorghum gene pool (cluster one). The notches around the median on each rectangle box display the significance ($P \leq 0.05$) of differences in the extent of crop-to-wild among farms. According to McGill *et al.* (1978) box plots whose notches do not overlap in a pairwise side by side comparison provide evidence of differences ($P \leq 0.05$) in the sample median. Results in the present study for example revealed significantly higher levels of crop-to-wild geneflow in farm 11 compared to all the other farms except farm 19. In contrast, the extent of crop-to-wild geneflow was significantly lower in farm 1 compared to all the other farms (Figure 4.4). Not surprisingly, the extent of crop-to-wild geneflow was generally higher among wild-weedy populations acquired from sorghum fields compared to those from either fallow fields or semi-natural habitats. In farm 19 for example, the magnitude of crop-to-wild geneflow was clearly higher (almost three-fold) in the wild-weedy sorghum population acquired from a sorghum field compared to that acquired from a semi-natural habitat

(Table 4.4; Figure 4.4). The length of a box plot portrays the variability of a sample. The extent of crop-to-wild geneflow was thus variable among individuals within farms, with the highest level being revealed in farm 11 and the least in farm 19.

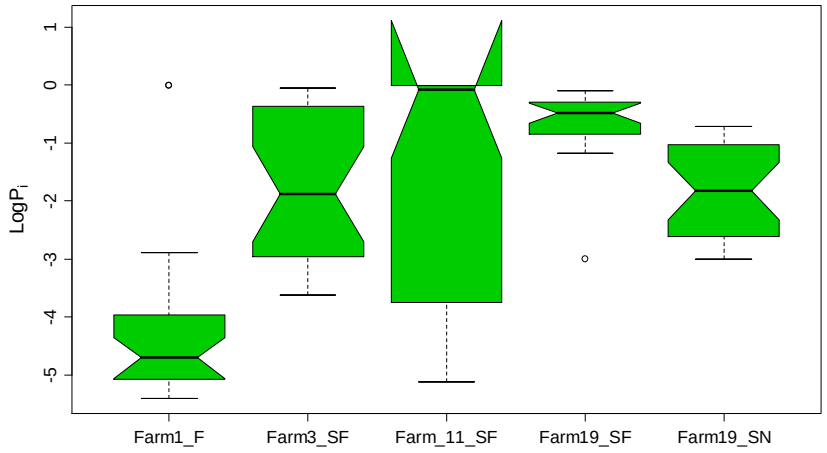

Figure 4.4 **Notched box plots showing farm-level differences in the proportion of wild-weedy sorghum genome originating from cultivated sorghum (crop-to-wild geneflow). Wild-weedy sorghum habitats are represented by the letters: F-Fallow field, SF-sorghum field and SN-Semi-natural habitat.**

4.4.3 Historical/long-term geneflow

When the F_{ST}-based approach was used to investigate the level of local scale geneflow between cultivated and wild-weedy sorghum populations, an overall *Nm* value of 0.66 was obtained. The overall divergence between cultivated and wild-weedy sorghum populations was substantial and highly significant (F_{ST} = 0.274; $P \leq 0.001$), an outcome that was congruent with the STRUCTURE based separation of the two congeners into distinct genetic groups (Figure 4.3).

4.4.4 Genetic structure

Cultivated and wild-weedy sorghum populations were highly differentiated, as indicated by substantial and significant F_{ST} values (0.392 and 0.354, respectively; $P \leq 0.001$; 5000 permutations) (Table 4.5). These results were supported by AMOVA, with 43.5% ($P \leq 0.001$) and 36.2% ($P \leq 0.001$) of the total diversity in cultivated and wild sorghum, respectively, being attributed to variation between populations (Table 4.6). The bulk of the diversity in the two congeners was thus partitioned within populations. High levels of

overall inbreeding in cultivated sorghum and its wild-weedy relatives were suggested by mean overall F_{IT} values of 0.718 and 0.710, respectively (Table 4.5). The average F_{IS} was 0.536 for cultivated sorghum and 0.550 for wild-weedy sorghum suggesting that substantial inbreeding occurs within populations of landraces and their wild-weedy relatives, on-farm.

Table 4.5 Estimates of F-statistics[1] for populations of cultivated and wild sorghum

Population	[2]F_{ST}	[3]F_{IT}	[4]F_{IS}
Cultivated sorghum	0.392 (0.052)	0.718 (0.049)	0.536 (0.065)
Wild-weedy sorghum	0.354 (0.034)	0.710 (0.027)	0.550 (0.034)

[1]Jacknife over loci – Standard deviation in parenthesis; [2] a measure of genetic differentiation; [3] a measure of overall inbreeding in the studied gene pools of cultivated and wild sorghum; [4] a measure of inbreeding within sub-populations (accessions).

Table 4.6 AMOVA partitioning of diversity within and among cultivated and wild-weedy sorghum populations

	Source of variation	Sum of squares	Variance components	% of variation	P-Value
Cultivated sorghum	Among populations	244.14	1.11	43.5	≤0.001
	Within populations	262.77	1.44	56.5	≤0.001
	Total	**506.92**	**2.55**		
Wild sorghum	Among populations	682.28	1.06	36.2	≤0.001
	Within populations	1200.96	1.86	63.7	≤0.001
	Total	**1883.25**	**2.93**		

When the local scale spatial gene structure in cultivated and wild sorghum populations was explored using an autocorrelation analysis according to Ritland (1996), there was no evidence of the an isolation-by-distance pattern (Figure 4.5).

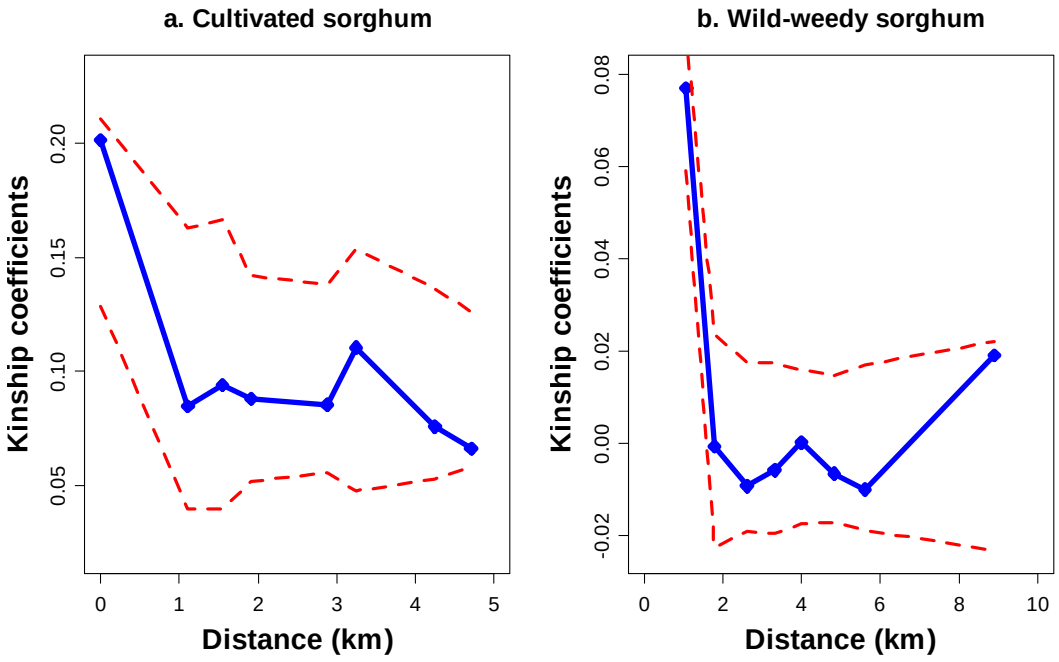

Figure 4.5 Correlograms for spatial patterns of genetic differentiation in (a) cultivated sorghum and (b) wild-weedy sorghum based on Ritland's pairwise kinship coefficients. The dashed lines represent the upper and lower 95% confidence limit envelopes around the null hypothesis of no spatial structure (random geneflow).

4.5 Discussion

4.5.1 Extent and direction of geneflow

The present study has demonstrated geneflow in the *S. bicolor* wild-weedy-domesticate complex at a local scale in a traditional cropping system. The estimated overall magnitude of recent introgression was nine times higher from the crop to the wild-weedy gene pool of sorghum than in the reverse direction. These results were generally consistent with past studies on hybridisation and introgression within the wild-weedy-domesticate complex of sorghum (Dogget and Majisu 1968; Aldrich and Doebley 1992; Arriola and Ellstrand 1996; 1997; Morrell *et al.* 2005), although directional extent of the introgression, especially under traditional sorghum farming systems in Africa has not been reported before. Dogget and Majisu (1968) performed a hybrid index analysis on samples of a *S. bicolor* wild-weedy-domesticate complex in East Africa using morphological markers. They speculated that samples genetically intermediate between cultivated and wild

sorghum were derivatives of hybridisation between the two congeners. The study of Arriola and Ellstrand (1996) used isozyme markers and progeny testing to conclude that spontaneous interspecific hybridisation between *S. bicolor* and the noxious weedy relative *S. halepense* (johnsongrass) can and does occur at substantial and measurable rates. Morrel and his co-workers (2005) used sorghum genomic and coding DNA (cDNA)-based RFLP probes to study crop-to-weed introgression in naturalised populations of johnsongrass in the USA. The authors found putative cultivar-specific alleles to be present at higher frequencies in populations of johnsongrass adjacent to sorghum cultivation fields, than in those away from or outside sorghum cultivation regions.

The asymmetrical crop-to-wild/weed geneflow estimated in the present study might have resulted from three none mutually exclusive explanations. First, on-farm differences in population size between cultivated and wild-weedy sorghum may favour pollen flow from the former. Routine weed control and management practices by farmers are expected to reduce the number of wild-weedy sorghum individuals relative to their domesticated counterparts. Under such circumstances, cultivated sorghum is expected to produce larger pollen densities relative to their wild-weedy relatives and therefore a higher pollen flow toward the limited wild-weedy plants within and around fields. Secondly, farmer seed selection may reduce the probability of wild gene introgression into the domesticated gene pool. Farmers may easily recognise and select against putative crop x wild hybrids with morphological features different from the cultivated sorghum maternal parents. In this study farmers classified putative crop x hybrids together with the wild-weedy sorghum under the name *Munya wa maguna*, which translates to the monkey sorghum. Most farmers were able to easily distinguish the various wild-weedy forms from their cultivated relatives using morphological features such as loose panicles, smaller and highly shattering seeds, high number of tillers and narrower leaves and stems. Finally, asymmetrical crop-to-wild-weedy geneflow in the *S. bicolor* complex may reflect higher natural outcrossing rates in the wild-weedy gene pools relative to its cultivated counterpart. For example, higher rates of natural outcrossing in the weedy Sudan grass (*S. bicolor* ssp. *sudanensis*) (up to 61%) compared to the grain type sorghum (up to 26%) have previously been reported in the USA (Pedersen *et al*. 1998).

The extent of crop-to-wild introgression was observed to vary significantly among farms, probably a reflection of differences among farmers in their crop and weed management

practices. For example, differences among farmers in management of wild-weedy sorghum may include: (i) frequency and timing of the weeding and (ii) extent to which farmers tolerate wild-weedy sorghum populations perceived to have no effect on the development of sorghum and other crops. Such information was not expressly collected in the present study but could be important in explaining the dynamics of crop-to-wild geneflow at farm level. For example, during the field surveys in the in the present study, the frequency of wild-weedy sorghum occurrence among farms was observed to vary among sites, both within sorghum fields and in the surrounding habitats such as fallows, field margins and terraces. In principle, practices affecting the level of spatial and/or flowering overlap between cultivated sorghum and its wild-weedy relatives can be expected to have an impact on the extent of crop-to-wild geneflow. For example, greater magnitude of crop-to-wild geneflow was revealed in the wild-weedy sorghum population found growing together with its cultivated congener, compared to the wild-weedy sorghum population found in a semi-natural grassland habitat on farm 19. The lower levels of crop-to-wild geneflow in the semi-natural habitat may be explained by its relative spatial isolation (20 m) from the nearest field of cultivated sorghum. Previous studies on crop-to-weed (Arriola and Ellstrand 1996) and crop-to-crop (Schmidt and Bothma 2006) geneflow in sorghum have demonstrated the rate of spontaneous hybridisation in the crop to decrease with distance.

4.5.2 Historical geneflow

Wright (1951) proposed that less than one effective migrant per generation ($Nm < 1$) is sufficient to allow population differentiation through fixation and genetic drift, whereas $Nm > 1$ is sufficient to overcome the effects of genetic drift and thus reduce the level of population differentiation. In the present study, the estimated level of historical geneflow between cultivated and wild sorghum was moderate ($Nm = 0.66$) and insufficient to overcome the effect of selection and genetic drift and thus divergence between the two congeners ($F_{ST} = 0.274$). This may be explained at least in part by continuous disruptive selective pressure (exerted by the farmers in cultivated environments and by natural selection in semi-natural habitats) against crop x wild hybrids and products of their introgressive hybridisation. Disruptive selection involves simultaneously selecting more than one level of a character in a population and is thought to have played a major role early in the domestication of sorghum. For example, Dogget (1988) hypothesised that a disruptive selection situation existed and continues to exist in the plots of many African small-holders, where wild sorghum and wild x cultivated hybrids are often found.

4.5.3 Genetic structure in cultivated and wild sorghum

The substantial and highly significant differentiation revealed among populations of both cultivated sorghum (F_{ST} = 0.392; AMOVA: 43.5% variation among populations) and its wild-weedy progenitor (F_{ST} = 0.354; AMOVA: 36.2% variation among populations), suggested the existence of some form of barriers to geneflow. Among such barriers, one should perhaps consider limited overlap in flowering times and differences in panicle morphology (Dje *et al.* 2004; Barnaud *et al.* 2007; 2008) among the various landraces and wild-weedy types. In cultivated sorghum, the degree of differentiation among populations (accessions) in the present study was within the range of what was previously reported among 12 landraces (F_{ST} = 0.36) in a local scale study in Cameroon (Barnaud *et al.* 2007), but lower than what was reported both by Dje *et al.* (2000) among 25 accessions (F_{ST} = 0.68) in a world collection and by Ghebru *et al.* (2002) among 28 Eritrean accessions (F_{ST} = 0.50). The lower F_{ST} values in the local scale studies could be explained by higher geneflow rates that potentially exists among landraces on-farm, where a number of landraces are usually planted closely mixed in the fields (Barnaud *et al.* 2007). Furthermore, sorghum seed systems in the present study were noted to be informal in nature, with farmers either saving their own seed, obtaining it from neighbours and relatives and in some instances buying the seed from local informal markets. Such practices may enhance the magnitude of geneflow among landraces and thus contribute to reduced genetic differentiation among cultivated sorghum accessions.

Wright's (1951) F-statistics parameters F_{IS} and F_{IT} can be used as indices of the level of inbreeding in sub-populations (within accessions) and in the overall population (entire cultivated and wild sorghum gene pool), respectively. The high level of inbreeding inferred overall (based on F_{IT} values) and within accessions (based on F_{IS} values) for both cultivated and wild sorghum in the present study was consistent with the predominantly self-pollinated mating systems associated with the sorghum gene pool. Similar results have been reported previously in cultivated sorghum (Dje *et al.* 1999; Barnaud *et al.* 2007). Dje and co-workers (1999) compared populations of cultivated sorghum accessions among different fields at a regional scale in Morocco using SSRs to report mean F_{IT} and F_{IS} values of 0.675 and 0.635, respectively. Barnaud and co-workers (2007) compared different landraces at a local scale in a village in Cameroon using SSRs to report mean F_{IS} value of 0.68. Under a mixed mating system model, the expression $s = 2F_{IS}/(1+F_{IS})$ (Weir 1996), a selfing rate of s = 0.70 and s = 0.71 would be expected in

cultivated and wild sorghum, respectively. The selfing rate value for cultivated sorghum fell within the range of what has previously been reported from experimental field data by Ellstrand and Foster (2009) based on progeny assays (mean s = 0.71) and by Ollitrault *et al.* (1997) based on a study on Guinea-race landraces in Burkina-Faso (mean s = 0.81).

Isolation-by-distance will generate positive and significant correlation between geographic distance (or its natural logarithm) and multilocus estimates of pairwise genetic relatedness (kinship coefficient). This relationship was not significant in the present study suggesting that genetic differentiation among populations/accessions in cultivated and wild sorghum at the local scale in the eastern slopes of Mt. Kenya is not directly related to physical distance. In cultivated sorghum, this may be explained by spatially random seed exchange among relatives and/or friends and/or purchase of seed from local markets. In the wild-weedy sorghum, lack of isolation-by-distance among populations may be explained either by human mediated secondary seed dispersal in seed lots and manure, or by grasing animals.

4.6 Conclusions and recommendations

This study has demonstrated asymmetric geneflow from cultivated sorghum to its wild-weedy relatives at a local scale in a traditional farming system in Kenya. The first implication for these findings is that geneflow from cultivated sorghum (whether GM or otherwise) will most likely greatly reduce genetic diversity (genetic erosion) in the sympatric wild and/or weedy populations. It would be interesting to investigate the extent to which crop-to-wild geneflow has impacted genetic diversity in wild-weedy relatives of sorghum. Such an investigation would demand sampling of wild sorghum from both cultivated and natural habitats away from cultivation for comparison. Secondly, asymmetric crop-to-wild geneflow suggests that wild-weedy relatives of sorghum growing sympatrically with deployed GM sorghum will most likely act as bridges for the escape of transgenes into surrounding cultivated and natural habitats. Whether or not such escape will lead to extinctions, increased weediness and/or invasiveness will depend on the nature of the transgene trait: whether neutral, detrimental, or beneficial in the ecological and genomic environment of the recipient population (Ellstrand *et al.* 1999). For example, Arriola and Ellstrand (1997) demonstrated similar levels of fitness between the noxious weed johnsongrass and progenies of its hybridisation with cultivated sorghum. The authors predicted the likely persistence of neutral or beneficial transgenes

in populations of johnsongrass growing in agricultural lands under continued geneflow from the crop. Such hybrid fitness studies however need to be extended to GM sorghum in order to generate empirical data for the possible effect of transgenes in ecological and/ or genomic environment of wild-weedy sorghum relatives.

Furthermore, this study has demonstrated differences in the extent of crop-to-wild geneflow among farms, probably a reflection of farmer differences in their crop management practices. These findings emphasise the need to take farmer practices into consideration when formulating biosafety regulation and guidelines for testing and/or releasing GM sorghum. Such guidelines could benefit from further studies that incorporate farmer information on among other the local seed systems, weed management and history of the sampled fields. Finally, sorghum landraces and their sympatric wild-weedy relatives constitute important genetic resources for sorghum breeding programmes and deserve conservation attention. Conservation strategies should take into consideration the substantial genetic differentiation demonstrated in this study between cultivated and wild-weedy sorghum gene pools as well as among their respective populations on-farm.

4.7 References

Aldrich PR, Doebley J. 1992. Restriction fragment variation in the nuclear and chloroplast genomes of cultivated and wild *Sorghum bicolor*. Theoretical and Applied Genetics 85:293-302.

Anderson E. 1949. Introgressive hybridization. John Wiley & Sons, Inc. New York, pp. 109.

Armstrong TT, Fitzjohn RG, Newstrom LE, Wilton AD, Lee WG. 2005. Transgene escape: what potential for crop-wild hybridization? Molecular Ecology 14:2111-2132.

Arnold ML. 1997. Natural hybridization and evolution. Oxford University Press Oxford, U.K., pp. 232.

Arriola PE, Ellstrand NC. 1996. Crop-to-weed geneflow in the genus *Sorghum* (Poaceae): Spontaneous inter specific hybridization between johnsongrass, *Sorghum*

halepense, and crop sorghum, *S. bicolor*. American Journal of Botany 83:1153-1160.

Arriola PE, Ellstrand NC. 1997. Fitness of interspecific hybrids in the genus Sorghum: Persistence of crop genes in wild populations. Ecological Applications 7:512-518.

Auer C. 2008. Ecological risk assessment and regulation for genetically-modified ornamental plants. Critical Reviews in Plant Sciences 27:255-271.

Ayoo LMK. 2008. Genetic transformation of Kenyan sorghum (*Sorghum bicolor* L. Moench) with anti-fungal genes and response to *Collectotrichum sublineolum* infection. PhD Dissertation, University of Hamburg, Germany. pp. 116.

Barnaud A, Deu M, Garine E, Mckey D, Joly HI. 2007. Local genetic diversity of sorghum in a village in northern Cameroon: structure and dynamics of landraces. Theoretical and Applied Genetics 114:237-248.

Barnaud A, Trigueros G, Mckey D, Joly HI. 2008. High outcrossing rates in fields with mixed sorghum landraces: how are landraces maintained? Heredity 101:445-452.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F. 2004. GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5000, Université de Montpellier II, Montpellier.

Cantamutto M, Poverene M. 2007. Genetically modified sunflower release: Opportunities and risks. Field Crops Research 101:133-144.

Casas AM, Kononowicz AK, Haan TG, Hang L, Tomes DL, Bressan RA, Hasegawa PM. 1997. Transgenic sorghum plants obtained after microprojectile bombardment of immature inflorescences. In vitro Cellular and Developmental Biology-Plant 33:92-100.

Casas AM, Kononowicz AK, Zehr BU, Tomes TD, Axtell DJ, Butler GL, Bressan AR, Hasegawa PM. 1993. Transgenic sorghum plants via microprojectile bomabardment. Proceeding of the National Academy of Sciences of the USA (PNAS) 90:11212-11216.

Celis C, Scurrah M, Cowgill S, Chumbiauca S, Green J, Franco J, Main G, Kiezebrink D, Visser RGF, Atkinson HJ. 2004. Environmental biosafety and transgenic potato in a centre of diversity for this crop. Nature 432:222-225.

Chandler S, Dunwell JM. 2008. Geneflow, risk assessment and the environmental release of transgenic plants. Critical Reviews in Plant Sciences 27:25-49.

Chapman MA, Burke JM. 2006. Letting the gene out of the bottle: the population genetics of genetically modified crops. New Phytologist 170:429-443.

Conner AJ, Glare TR, Nap JP. 2003. The release of genetically modified crops into the environment. Part II. Overview of ecological risk assessment. The Plant Journal 33:19-46.

De Wet JMJ. 1978. Systematics and evolution of sorghum sect. Sorghum (Gramineae). American Journal of Botany 65:477-484.

De Wet JMJ, Harlan JR. 1975. Weeds and Domesticates: Evolution in the man-made habitat. Economic Botany 29:99-108.

Dje Y, Forcioli D, Ater M, Lefebvre C, Vekemans X. 1999. Assessing population genetic structure of sorghum landraces from North-western Morocco using allozyme and microsatellite markers. Theoretical and Applied Genetics 99:157-163.

Dje Y, Heuertz M, Ater M, Lefebvre C, Vekemans X. 2004. *In situ* estimation of outcrossing rate in sorghum landraces using microsatellite markers. Euphytica 138:205-212.

Dje Y, Heuertz M, Lefebvre C, Vekemans X. 2000. Assessment of genetic diversity within and among germplasm accessions in cultivated sorghum using microsatellite markers. Theoretical and Applied Genetics 100:918-925.

Dogget H. 1988. Sorghum. Longman Scientific and Technical Essex, England, pp. 512.

Dogget H, Majisu BN. 1968. Disruptive selection in crop development. Heredity 23:1-23.

Ellstrand NC. 2003. Dangerous liaisons - when cultivated plants mate with their wild relatives. Johns Hopkins University Press Baltimore MD, pp. 244.

Ellstrand NC, Foster KW. 2009. Impact of population structure on the apparent outcrossing rate of grain sorghum (*Sorghum bicolor).* Theoretical and Applied Genetics 66:323-327.

Ellstrand NC, Prentice HC, Hancock JF. 1999. Geneflow and introgression from domesticated plants into their wild relatives. Annual Review of Ecology and Systematics 30:539-563.

Evanno S, Regnaut S, Goudet J. 2005. Detecting the number of cluster of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14:2611-2620.

Excoffier L, Laval LG, Schneider S. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1:47-50.

Excoffier L, Smouse PE, Quattro JM. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567-1587.

Gao Z, Xie X, Ling Y, Muthukrishnan S, Liang GH. 2005. *Agrobacterium tumefaciens*-mediated sorghum transformation using a mannose selection system. Plant Biotechnology Journal 3:591-599.

Ghebru B, Schmidt RJ, Bennetzen JL. 2002. Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. Theoretical and Applied Genetics 105:229-236.

Girijashankar V, Sharma KK, Swathisree V, Prasad LS, Bhat BV, Royer M, Narasu ML, Altosaar I, Seetharama N. 2005. Development of transgenic sorghum for insect resistance against the spotted stem borer (*Chilo partellus*). Plant Cell Reports 24:513-522.

Hails RS, Morley K. 2005. Genes invading new populations: a risk assessment perspective. Trends in Ecology and Evolution 20:245-252.

Hancock JF. 2003. A framework for assessing the risk of transgenic crops. Bioscience 53:512-519.

Hardy OJ, Vekemans X. 2002. SPaGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology 2:618-620.

Harlan JR, De Wet JMJ. 1972. A simplified classification of cultivated sorghum. Crop Science 12:172-177.

Harlan JR, Stemler ABL. 1976. The races of sorghum in Africa. In: Harlan JR, De Wet JMJ, Stemler ABL (Eds.), Origins of African Plant Domestication. Mouton, The Hague, Paris. pp. 478.

Haygood R, Ives AR, Andow DA. 2003. Consequences of recurrent geneflow from crops to wild relatives. Proceedings of the Royal Society of London Series B-Biological Sciences 270:1879-1886.

Hooftman DAP, Gerard J, Oostermeijer B, Marquard E, den Nijs HCM. 2008. Modelling the consequences of crop-wild relative geneflow: a sensitivity analysis of the effects of outcrossing rates and hybrid vigour breakdown in Lactuca. Journal of Applied Ecology 45:1094-1103.

Howe A, Shirley S, Dweikat I, Fromm M, Clemente T. 2006. Rapid and reproducible *Agrobacterium*-mediated transformation of sorghum. Plant Cell Reports 25:751-758.

James C. 2008. Global status of commercialised Biotech/GM crops: 2007. ISAAA Brief No. 39 (http://www.isaaa.org).

Jarvis DI, Hodgkin T. 1999. Wild relatives and crop cultivars: detecting natural introgression and farmer selection of new genetic combinations in agroecosystems. Molecular Ecology 8:S159-S173.

Jenczewski E, Ronfort J, Chevre A. 2003. Crop-to-wild geneflow, introgression and possible fitness effects of transgenes. Environmental Biosafety Research 2:9-24.

Krishnaveni S, Jeoung J, Muthukrishnan S, Liang G. 2000. Transgenic sorghum plants constitutively expressing a rice chitinase gene show improved resistance to stalk rot. Journal of Genetic Breeding 55:151-158.

Kumar V, Bellinder RR, Brainard DC, Malik RK, Gupta RK. 2008. Risks of herbicide-resistant rice in India: A review. Crop Protection 27:320-329.

McGill R, Tukey J, Larsen W. 1978. Variations of box plots. The American Statistician 32:12-16.

Michalakis Y, Excoffier L. 1996. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061-1064.

Morrell PL, Williams-Coplin TD, Lattu AL, Bowers JE, Chandler JM, Patterson AH. 2005. Crop-to-weed introgression has impacted allelic composition of johnsongrass populations with and without recent exposure to cultivated sorghum. Molecular Ecology 14:2143-2154.

Mutegi E, Sagnard F, Muraya M, Kanyenji B, Rono B, Mwongera C, Marangu C, Kamau J, Parzies H, de Villiers S, Semagn K, Traore PS, Labuschagne M. 2009. Ecogeographical distribution of wild, weedy and cultivated *Sorghum bicolor* (L.) Moench in Kenya: implications for conservation and crop-to-wild gene flow. Genetic Resources and Crop Evolution (DOI 10.1007/s10722-009-9466-7)

Ollitrault P, Noyer J, Chantereau J, Glaszmann J. 1997. Structure g´en´etique et dynamique des vari´et´es traditionnelles de sorgho au Burkina-Faso. Gestion des ressources g´en´etiques des plantes en Afrique des savanes, IER-BRG Solagral, Bamako, Mali. p. 231-240.

Pedersen J, Toy JJ, Johnson B. 1998. Natural outcrossing of Sorghum and Sudangrass in the central great plains. Crop Science 38:937-939.

Poppy GM. 2004. Geneflow from GM plants - towards a more quantitative risk assessment. Trends in Biotechnology 22:436-438.

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

R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (http://www.r-project.org).

Ritland K. 1996. Estimators for pairwise relatedness and individual inbreeding coefficients. Genetical Research 67:175-185.

Rogers HJ, Parkes HC. 1995. Transgenic plants and the environment. Journal of Experimental Botany 46:467-488.

Schmidt M, Bothma G. 2006. Risk assessment for transgenic sorghum in Africa: crop-to-crop geneflow in *Sorghum bicolor* (L.) Moench. Crop Science 46:790-798.

Schoen DJ, Reichman JR, Ellstrand NC. 2008. Transgene escape monitoring, population genetics, and the law. Bioscience 58:71-77.

Simmonds NW. 1992. Evolution of crop plants. Longman Scientific and Technical New York, pp. 339.

Snow AA, Moran-Palma P. 1997. Commercialization of transgenic plants: potential ecological risks. BioScience 47:86-96.

Sokal RR, Oden NL. 1978a. Spatial autocorrelation in biology 1. Methodology. Biological Journal of the Linnean Society 10:199-228.

Sokal RR, Oden NL. 1978b. Spatial autocorrelation in biology 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society 10:249.

Tesso T, Kapran I, Grenier C, Snow A, Sweeney P, Pedersen J, Marx D, Bothma G, Ejeta G. 2008. The potential for crop-to-wild geneflow in sorghum in Ethiopia and Niger: A geographic survey. Crop Science 48:1425-1431.

VSN International Ltd. 2007. GenStat Discovery Edition 3. VSN International Ltd. Hernel Hempstead, UK.

Warwick SI, Stewart CN. 2005. Crops come from wild plants - How domestication, transgenes, and linkage together shape ferality. In: Gressel J (Ed.), Crop Ferality and Volunteerism. CRC Press, Boca Raton, Florida. pp. 30.

Weir B. 1996. Genetic data analysis II: Methods for discrete population genetic data. Sinauer Associates Sunderland, Massachusetts, pp. 445.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Wright S. 1951. The genetical structure of populations. Annals of Eugenics 15:323-354.

Zhao Z. 2008. The Africa biofortified sorghum project - Applying biotechnology to develop nutritionally improved sorghum for Africa. Proceedings of the 11th IAPTC&B Congress, August 31-18, 2006 Beijing, China. p. 273-277.

Zhao Z, Cai T, Tagliani L, Wang N, Pang H, Rudert M, Schroeder S, Hondred D, Pierce D. 2000. *Agrobacterium*-mediated sorghum transformation. Plant Molecular Biology 44:789-798.

Zhu H, Muthukrishnan S, Krishnaveni S, Wilde G, Jeoung J, Liang G. 1998. Biolistic transformation of sorghum using a rice chitinase gene. Journal of Genetic Breeding 52:243-252.

Chapter 5

General conclusions and recommendations

Sorghum was domesticated in Africa and is the world's fifth most important cereal, both in terms of production and area dedicated to cultivation. In Africa, sorghum is the second most important cereal; supporting the dietary and nutrition requirements of over 100 million, mostly resource challenged rural people in the continent's arid and semi-arid lands. There is now a growing need to increase the supply of food in sub-Saharan Africa, in an effort to cope with an ever increasing human population. The prevailing urgent need to increase productivity and nutritional quality of Africa's staple crops such as sorghum cannot therefore be gainsaid. Advances in modern genetic engineering techniques offer promising tools (in addition to traditional breeding approaches) for overcoming some of the productivity and nutritional bottlenecks in sorghum. Although GM sorghum is not yet available commercially, optimised transformation protocols have been developed and successfully applied on the crop for various transgenic traits including insect resistance, herbicide tolerance, anti-fungal attack and increased lysine content.

The increasing prospect for deployment of transgenic sorghum into Africa's traditional farming systems is exemplified by the African Biofortified Sorghum project. This project is a consortium of nine institutions working to develop and deploy into Africa, transgenic sorghum varieties that are easier to digest and contain higher levels of vitamins A and E and the essential amino acid, lysine. Nonetheless, as is the case with other GM crops, there are growing concerns that through crop-to-wild geneflow, transgenes may escape from GM sorghum to its wild and/or weedy relative populations in cultivated and/or natural habitats. For example, there are environmental concerns that escaped transgenes may confer fitness traits into crop-wild/weedy hybrids leading to increased weediness, invasiveness, genetic erosion and in some severe cases extinction of entire populations. These concerns have stimulated a growing body of investigations on geneflow related environmental risk assessment studies in a number of crops and their wild and/or weedy relatives. The ultimate goal is to aid biosafety decision-making for testing and/or commercially releasing GM crops.

In February 2009, the Government of Kenya enacted a biosafety law paving the way for the commercial production of GM crops. There is currently an urgent need therefore to formulate scientifically sound biosafety regulations and guidelines. Understanding the extent and direction of movement of genes between the domesticated and wild and/or weedy relative populations is the first step in characterising the potential environmental risks of escaped transgenes. Such studies are yet to be reported for sorghum in Africa, even though diverse forms of the crop and its wild and weedy relatives are known to still co-exist widely within and near agricultural lands. The present study was therefore primarily motivated by the current need to characterise the extent and direction of geneflow between cultivated sorghum and its wild or weedy relatives under traditional farming systems in Kenya. Estimates of the extent and direction of geneflow in the wild-weedy-domesticate complex of *S. bicolor* in the current study have provided a useful foundation for formulating biosafety regulations and guidelines for testing and/or commercially releasing GM sorghum in Kenya.

Using a combination of distance and model-based population genetics approaches, the current study has demonstrated that geneflow takes place between cultivated and wild-weedy sorghum populations under traditional farming systems in Kenya. At the same time, this study has established that geneflow between the two congeners is asymmetric, with up to nine-fold greater movement of genes from cultivated sorghum to its wild-weedy relatives than in the reverse direction. Furthermore, the present study has estimated the extent of geneflow between cultivated and wild sorghum to vary among growing regions at national level and among farms and habitats at local scale. The biosafety implication of these findings is that wild-weedy sorghum populations growing within and/or near GM sorghum fields will most likely act as conduits for the escape of transgenes into agricultural and nearby natural habitats. Biosafety regulators would benefit from further studies on the likely genetic and/or ecological consequences of such escape in the recipient wild-weedy sorghum populations. Nonetheless, the asymmetric crop-to-wild geneflow revealed in the present study is a pre-requisite for genetic swamping of wild-weedy alleles with those from cultivated sorghum populations. The consequence of such swamping is genetic erosion in recipient wild-weedy sorghum populations. There exists a need to verify the extent to which such genetic erosion has already taken place within and near cultivated sorghum fields.

Cultivated sorghum landraces and their wild and/or weedy relatives constitute important genetic resources and should be conserved now for present and future crop improvement programmes. Such efforts should ideally be preceded by an understanding of the extent and partitioning of this diversity within and among the wild-weedy-domesticate complex of sorghum populations across ecological and geographic ranges. Such information is currently lacking for Kenya but undoubtedly important in aiding conservation and breeding programmes to focus effort and resources on truly distinctive groups. It was the aim of this study therefore, to characterise the genetic structure and relationships within and between cultivated and wild sorghum in Kenya, including understanding some of the underlying evolutionary mechanisms.

The current study has revealed higher levels of overall genetic diversity in Kenya's wild-weedy sorghum gene pool compared to the cultivated counterparts. These findings suggest that the process of domestication was accompanied by genetic bottlenecks as cultivated sorghum was selected, diversified and established from a limited number of founder individuals. Furthermore, these findings emphasise the importance of crop wild-weedy relatives as potential sources of unique and novel genes, especially towards broadening the genetic base in the cultivated counterparts. Unfortunately, these important genetic resources are under immense threat of loss due to a number of biotic and abiotic factors: (i) habitat fragmentation and destruction, (ii) climate change and its associated negative environmental impacts and (iii) as has clearly been demonstrated in the present study, crop-to-wild geneflow and its associated genetic erosion. The need to systematically collect and conserve these important genetic resources for breeding and other potential future uses cannot be overemphasised.

At the country scale, the present study revealed substantial differences in the amount of genetic diversity among regions both in cultivated sorghum and its wild-weedy relatives. Based on the AMOVA, the current study further revealed greater partitioning of genetic diversity within rather than among the growing regions at country scale, in the two congeners. This would suggest that future germplasm exploratory and collection undertakings will have to sample more intensively within the different sorghum growing regions in the country, in an effort to obtain samples that are representative of the available genetic variation. However, most wild-weedy sorghum samples analysed in the study were sampled from within and/or near agricultural lands. Furthermore, no wild-

weedy sorghum samples from the Rift Valley and north-eastern regions of the country were analysed. Strategies to systematically conserve the wild-weedy-domesticate complex of sorghum in Kenya could benefit from additional genetic diversity studies using a more comprehensive sampling scheme. Particularly, wild sorghum samples from natural habitats (e.g. protected areas) away from sorghum cultivation should form part of such studies. Phenotypic traits which may be economically important to sorghum improvement need to be exploited and evaluated in association with molecular markers, to determine whether the genetic diversity patterns observed in the neutral markers reflect similar patterns in the agro-morphologically useful alleles.

Finally, based on F_{ST}, AMOVA and Bayesian model-based analyses at country scale, this study has revealed poor geographic differentiation of genetic diversity in cultivated as well as wild sorghum gene pools. This outcome suggests that geographical isolation distances are not sufficient to prevent geneflow among the growing regions, possibly due to long-distance seed exchanges among farmers. For the wild sorghum gene pool, the findings were rather surprising, but could reflect inadvertent human-mediated long-distance movement of wild seeds via cultivated sorghum seed systems. This hypothesis was given further credence by findings of significant positive spatial autocorrelation of genetic relatedness (within approximately 180 km) at country scale in the two congeners. These findings are of significance in formulation of biosafety strategies for GM sorghum in the country. The implication is that existing sorghum seed systems are capable of long-distance dispersal of transgenic sorghum and/or derivatives for their hybridisation within the wild-weedy-domesticate complex. Such factors should be considered in formulating biosafety regulations and guidelines for testing and/or commercially releasing GM sorghum in Kenya.

Summary

The important socio-economic position enjoyed by sorghum has made it a target for genetic modification to enhance productivity and/or nutritional quality. However, there are growing environmental concerns that through geneflow, transgenes might escape from genetically modified (GM) sorghum to its sexually compatible wild and/or weedy relatives, with possible negative effects. Characterising the levels and dynamics of geneflow in the wild-weedy-domesticate complex of sorghum in traditional agro-ecosystems of Africa, including Kenya, is of interest to biosafety regulators. This study used approaches based on population genetics theory to (i) characterise the genetic structure of the wild-weedy-domesticate complex of *S. bicolor* at a country scale in Kenya and (ii) to estimate the extent and direction of geneflow between cultivated and wild-weedy populations at a local scale in the country. The structure and dynamics of diversity was first estimated and compared at country scale by genotyping 329 cultivated and 110 wild sorghum individuals using 24 microsatellite markers. Subsequently, the magnitude and direction of geneflow between the two congeners was estimated at a local scale by analysing 483 individuals comprising of 15 wild-weedy and 12 cultivated sorghum populations using 10 microsatellite markers. Overall, cultivated sorghum harboured lower diversity than its wild counterpart. Levels of genetic diversity in cultivated and wild sorghum differed significantly among regions, with most of the diversity being partitioned more within than among the crop's growing regions. There were generally low levels of differentiation within and between cultivated and wild sorghum at country scale, but the extent of crop-to-wild genetic proximity varied among sorghum growing regions. These findings may reflect important historical geneflow between cultivated sorghum and its progenitor, with the level of crop-to-wild genetic exchange varying among regions. At local scale, the extent of geneflow was approximately nine times higher from cultivated sorghum to its wild-weedy relatives than *vice versa*. Additionally, the extent of crop-to-wild geneflow varied significantly among farms. Overall, this study suggests that deployment of GM sorghum in Kenya's agro-ecosystems will most likely lead to movement of transgenes into sympatric populations of its wild-weedy relatives.

Keywords: autocorrelation, diversity, genetic structure, introgression, phylogeography, *Sorghum bicolor*, spatial analysis, wild-weedy-domesticate complex

Opsomming

Die belangrike sosio-ekonomiese posisie van sorghum het dit die ideale gewas vir genetiese modifikasie vir produktiwiteit en/of voedingskwaliteitsverbetering gemaak. Daar is egter groeiende kommer oor die omgewing, aangesien geenvloei transgeen ontsnapping vanaf geneties gemodifiseerde (GM) sorghum na wilde en/of onkruid kruisingsversoenbare verwante kan veroorsaak, met moontlike negatiewe effekte. Die karakterising van die vlakke en dinamika van geenvloei in die wilde-onkruidagtige-verboude kompleks van sorghum in tradisionele agro-ekosisteme van Afrika, Kenia ingesluit, is van belang vir bioveilgheidsreguleerders. Hierdie studie het die benadering van populasiegenetikateorie gebruik om (i) die genetiese struktuur van die wilde-onkruidagtige-verboude kompleks van *S. bicolor* binne die hele Kenia te karakteriseer en (ii) om die hoeveelheid en rigting van geenvloei tussen verboude en wilde-onkruidagtige populasies op ‘n plaaslike vlak binne die land te bepaal. Die struktuur en dinamika van diversiteit is eers bepaal en toe nasionaal vergelyk deur 329 verboude en 110 wilde sorghum individue met 24 mikrosatelliet merkers. Daarna is die hoeveelheid en rigting van geenvloei tussen die twee groepe op plaaslike skaal bepaal deur toetsing van 483 individue bestaande uit 15 wilde-onkruidagtige en 12 verboude sorghum populasies met 10 mikrosatelliet merkers te evalueer. In die geheel het verboude sorghum minder diversiteit as die wilde groep getoon. Vlakke van genetiese diversiteit in verboude en wilde sorghum het betekenisvol tussen streke verskil, met meer diversiteit binne as tussen streke. Daar was oor die algmeen lae vlakke van differensiasie binne en tussen verboude en wilde sorghum op nasionale vlak, maar die gewas-na-wilde genetiese afstand het tussen sorghum produksie areas gewissel. Hierdie bevindinge kan die belangrike historiese geenvloei binne verboude sorghum en die bron van afkoms reflekteer, met die vlak van gewas-na-wilde uitruiling wat verskil tussen streke. Op ‘n plaaslike vlak was die hoeveelheid geenvloei ongeveer nege keer hoër van verboude sorghum na die wilde-onkruidagtige verwantes as andersom. Verder het die vlak van gewas-na-wilde geenvloei betekenisvol tussen plase verskil. In die geheel dui hierdie studie aan dat die vrystelling van GM sorghum in Kenia se agro-ekosisteem waarskynlik na die beweging van transgene na simpatriese populasies van die wilde-onkruidagtige verwantes sal lei.

Sleutelwoorde: diversiteit, filogeografie, genetiese struktuur, introgressie, outokorrelasie, ruimtelike analise, *Sorghum bicolor*, wilde-onkruidagtige-verboude kompleks

Appendices

Appendix 1 List of the 24 microsatellite loci and their overall variation in the entire Kenyan sorghum genepool

Locus name	Core motif	Forward primer sequence (5'-3')	Reverse primer sequence(5'-3')	Size range	No. of alleles	H_e	H_o
1. gpsb123	(AC)7 (GA)5	M13-ATAGATGTTGACGAAGCA	GTGGTATGGGACTGGA	284-316	8	0.7895	0.1235
2. mSbCIR238	(AC)26	M13-AGAAGAAAAGGGGTAAGAGC	CGAGAAACAATTACATGAACC	64-112	23	0.8826	0.2044
3. mSbCIR240	(TG)9	M13-GTTCTTGGCCCTACTGAAT	TCACCTGTAACCCTGTCTTC	101-113	7	0.5864	0.1117
4. mSbCIR246	(CA)7.5	M13-TTTTGTTGCACTTTTGAGC	GATGATAGCGACCACAAATC	87-109	9	0.3993	0.0857
5. mSbCIR248	(GT)7.5	M13-GTTGGTCAGTGGTGGATAAA	ACTCCCATGTGCTGAATCT	84-122	14	0.7544	0.3890
6. mSbCIR262	(CATG)3.25	M13-GCACCAAAATCAGCGTCT	CCATTTACCCGTGGATTAGT	210-242	7	0.5433	0.0730
7. mSbCIR276	(AC)9	M13-CCCCAATCTAACTATTTGGT	GAGGCTGAGATGCTCTGT	220-238	9	0.4521	0.0958
8. mSbCIR300	(GT)9	M13-TTGAGAGCGGCGAGGTAA	AAAAGCCCAAGTCTCAGTGCTA	98-122	10	0.6278	0.1412
9. SbAGB02	(AG)35	M13-CTCTGATATGTCGTTGTGCT	ATAGAGAGGATAGCTTATAGCTCA	91-143	25	0.8359	0.1443
10. Xcup02	(GCA)6	M13-GACGCAGCTTTGCTCCTATC	GTCCAACCAACCCACGTATC	185-203	7	0.5147	0.1329
11. Xcup14	(AG)10	M13-TACATCACAGCAGGGACAGG	CTGGAAAGCCGAGCAGTATG	201-241	18	0.6536	0.0843
12. Xcup53	(TTTA)5	M13-GCAGGAGTATAGGCAGAGGC	CGACATGACAAGCTCAAACG	178-202	7	0.4760	0.2329
13. Xcup61	(GAG)7	M13-TTAGCATGTCCACCACAACC	AAAGCAACTCGTCTGATCCC	186-210	8	0.5241	0.1000

Locus name	Core motif	Forward primer sequence (5'-3')	Reverse primer sequence(5'-3')	Size range	No. of alleles	H_e	H_o
14. Xcup63	(GGATGC)4	M13-GTAAAGGGCAAGGCAACAAG	GCCCTACAAAATCTGCAAGC	135-153	4	0.3551	0.0616
15. Sb4-72	(AG)16	M13-TGCCACCACTCTGGAAAAGGCTA	CTGAGGACTGCCCCAAATGTAGG	176-214	13	0.8040	0.0890
16. Xisep0310	CCAAT(4)	M13-TGCCTTGTGCCTTGTTTATCT	GGATCGATGCCTATCTCGTC	148-218	11	0.2864	0.0182
17. Xtxp010	(CT)14	M13-ATACTATCAAGAGGGGAGC	AGTACTAGCCACACGTCAC	129-151	12	0.8233	0.0882
18. Xtxp012	(CT)22	M13-CGTCTTCTACCGCGTCCT	CATAATCCCACTCAACAATCC	165-209	23	0.9286	0.1525
19. Xtxp015	(TC)16	M13-CACAAACACTAGTGCCTTATC	CATAGACACCTAGGCCATC	201-235	16	0.8262	0.1800
20. Xtxp057	(GT)21	M13-GGAACTTTTGACGGGTAGTGC	CGATCGTGATGTCCCAATC	217-265	23	0.8998	0.1499
21. Xtxp114	(AGG)8	M13-CGTCTTCTACCGCGTCCT	CATAATCCCACTCAACAATCC	191-239	8	0.4786	0.0824
22. Xtxp136	(GCA)5	M13-GCGAATAGCATCTTACAACA	ACTGATCATTGGCAGGAC	233-239	3	0.4350	0.0886
23. Xtxp320	(AAG)20	M13-TAAACTAGACCATATACTGCCATGATAA	GTGCAAATAAGGGCTAGAGTGTT	248-305	19	0.8684	0.1234
24. Xtxp40	(GGA)7	M13-CAGCAACTTGCACTTGTC	GGGAGCAATTTGGCACTAG	103-142	11	0.4368	0.0881
25. mSbCIR223	(AC)6	CGTTCCAATGACTTTTCTTC	GCCAATGTGGTGTGATAAAT		Eliminated		
26. mSbCIR283	(CT)8 (GT)8.5	TCCCTTCTGAGCTTGTAAAT	CAAGTCACTACCAAATGCAC		Eliminated		
27. Xgap206	(AC)13/(AG)20	HEX-ATTCATCATCCTCATCCTCGTAGAA	AAAAACCAACCCGACCCACTC		Eliminated		
28. Xtxp265	(GAA)19	GTCTACAGGCGTGCAAATAAAA	TTACCATGCTACCCCTAAAAGTGG		Eliminated		
29. Xtxp278	(TTG)12	GGGTTTCAACTCTAGCCTACCGAACTTCCT	ATGCCTCATCATGGTTCGTTTTGCTT		Eliminated		
30. Xtxp278	(TTG)12	GGGTTTCAACTCTAGCCTACCGAACTTCCT	ATGCCTCATCATGGTTCGTTTTGCTT		Eliminated		

H_e = Gene diversity (Expected heterozygosity), H_o = Observed heterozygosity

Appendix 2 Agro-climatic zones in Kenya with moisture index and annual rainfall

Agro-climatic zones	Classification	Moisture index (%)	Annual rainfall (mm)
I	Humid	>80	1100 - 2700
II	Sub-humid	65 – 80	1000 - 1600
III	Semi-humid	50 – 65	800 - 1400
IV	Semi-humid to semi-arid	40 – 50	600 - 1100
V	Semi-arid	25 – 40	450 - 900
VI	Arid	15 – 25	300 - 550
VII	Very arid	< 15	150 - 350