

Biogeographical and Genomic Analysis of *Eleusine* Species

by

Adekola Oluwatosin Owoyemi

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science
Auburn, Alabama
December 11, 2021

Keywords: Species Distribution Model (SDM), Structural Variant (SV), *Eleusine*

Copyright 2021 by Adekola O. Owoyemi

Approved by

Leslie Goertzen, Chair, Associate Professor of Biological Sciences
Courtney Leisner, Assistant Professor of Biological Sciences
Kevin Burgess, Assistant Professor of Biological Sciences

Abstract

Eleusine Gaertn. (Poaceae, subfamily Chloridoideae), is a small taxon of closely related and distinct diploids and tetraploids endemic to Africa that have been scrutinized from vegetative, floral, cytological, and molecular evidence with a sustained interest in their phylogeny and adaptations, partly due to the economic and ecological impacts of a super crop (*E. coracana*) and a weed species (*E. indica*) in the genus. Studies to elucidate the genotypic and phenotypic relationships in *E. coracana* have always involved Single Nucleotide Polymorphisms (SNPs), although recent studies show that SNPs do not capture large genomic variations that equally contribute to phenotypic differences. In this thesis, I used environmental data to characterize the eco-geographical distribution of the different *Eleusine* species in Africa and investigated structural variations in *E. coracana*. Using Maximum Extent modeling software (Maxent), I characterized possible environmental predictors for the presence of *Eleusine* species in Africa based on collection records on Global Biodiversity Information Facility (GBIF) and 33 bioclimatic and soil data. Furthermore, I analyzed publicly available, paired-end, whole-genome *E. coracana* sequences from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) repository for structural variants and their genomic distribution with custom bash and R scripts created with freely available bioinformatics tools. Maxent modeling revealed a high degree of variation in the probability of *Eleusine* species on the African continent and indicated possible suitable environments in new locations. There is a need to corroborate these environmental distribution findings with known locality records (e.g., herbarium records) and field verifications. Whole genome sequence (WGS) analysis revealed a high occurrence of Structural Variants (SV) in *Eleusine coracana* with 455 inversions, 18,990 duplications, and 103,338 deletions variants detected. This high incidence of deletion and duplication events are consistent

with SV analyses in other plants, especially polyploids. In addition, substantiating identified genomic variations in *E. coracana* in combined multiple approaches involving other high-performance SV callers would be helpful for more robust prediction and reduce error calls. Hopefully, identified variants lay the groundwork for future analyses identifying structural genomic variations. These approaches in this research together present the first data uncovering environmental preferences and genomic variation influences in *Eleusine* and can help our understanding of the genus.

Acknowledgements

My sincere appreciation goes to the staff, faculty, and students of the Biological Science Department for contributing one way or the other to the successful completion of my program in Auburn university. Special thanks to:

Dr Leslie Goertzen—my committee chair, conversation with you before and through my MS polished my thoughts and made this research possible. Thank you for providing direction to my research aims. I appreciate your interest in my academic pursuit and your wonderful support, from the very first email

Dr Leisner Courtney— for agreeing to be on my committee, for providing insights and feedbacks on my analysis, for your encouragement, and timely support, and for making my defense enjoyable

Dr Kevin Burgess— for agreeing to be on my committee, for insightful feedbacks and suggestions and for making my defense enjoyable

Dr Katelyn Lawson— for a fantastic GIS for Biologist class which inspired the SDM analysis and for your expertise with SDM analysis when I needed it the most

Dr Tonia Schwartz—for teaching Functional genomics course where I familiarized with genomics analysis and provide the much-needed insight and referral when I was stuck

Dr Laurie Stevison—for insights on structural variant analysis and showing me what a one-line of bash code could do; I wished I attended one of your classes

Dr David Young—for computational support on the Alabama Super Computer and being patient with my long list of demands

Dr Nathan Hall—although we never shared time in the Goertzen's lab, you selflessly gave your time and insights and broadened my knowledge of genome analysis. Thanks for leaving a blazing trail in the lab that I could follow

Kirby Norell—thanks for being available all time and doing much more than providing administrative support

Francesco Moen—thanks for being the one that I could always call on while I was still learning

Giovani Rossi—the stranger who was like my family from my first day at Auburn University, I felt at home having you as my friend

Sope Adeniji—my friend, partner, gist buddy, and code-graph-writing-peer reviewer of life. Thanks for nurturing my dreams

Joel Adebola—your coming made everything fun

My family members back home in Nigeria—thanks for your constant morale support

Table of Contents

Abstract	ii
Acknowledgments.....	iv
Table of Contents	v
List of Tables	vi
List of Figures	vii-viii
List of Abbreviations	ix
Chapter 1: General introduction.....	1
Research Objective and Experimental Context	4
Chapter 2: Geospatial Characterization and Distribution Mapping of Eleusine Species in Africa	5
Introduction	5
Materials and Methods	7
Results	18
Discussion	36
Chapter 3: Structural Variation Analysis of Eleusine coracana whole-genome- sequences	43
Introduction	43
Methods.....	46
Results	52
Discussion	65
Chapter 4: General Conclusions	70
References	71
Appendix 1. Code used for selecting mapping extent for each species in ArcGIS	79

List of Tables

Table 2.1: Summary of downloaded environmental layer maps.....	11
Table 2.2: Broad soil categorization used in Soil Atlas of Africa downloaded from the Joint Research Centre-European Soil Data Centre (ESDAC)	13
Table 2.3: Summary of <i>Eleusine</i> herbarium records downloaded from GBIF	15
Table 2.4: logistic threshold cutoff values from maximum training sensitivity plus specificity (maximum value = 1) for the full Africa extent Maxent models.....	19
Table 2.5 logistic threshold cutoff values from maximum training sensitivity plus specificity (maximum value = 1) for the narrow Africa extent Maxent models	25
Table 2.6: Number of substantial environmental variables for the full and the narrow extent Maxent models	33
Table 3.1 BioProject accession numbers and numbers of SRAs per each downloaded for analysis from NCBI database	48
Table 3.2 quality of reads of <i>E. coracana</i> sequences downloaded from NCBI before and after trimming	53
Table 3.3 Summary of significant ($p \leq 0.05$) GO:Process functional annotation of genes overlapping identified structural variations	63

List of Figures

Figure 2.1 Distribution of nine <i>Eleusine</i> species across the African continent	8
Figure 2.2 Maxent settings used for modeling distribution	17
Figure 2.3 Binary maps of Eleusine species distribution from full extent model	20-21
Figure 2.4 AUC Plots of testing and training omission and predicted area varies with the cumulative threshold for full extent models.....	22-23
Figure 2.5 Receiver operating curves (ROC) for training and test data plots for full extent models.....	24
Figure 2.6 Binary maps of Eleusine species distribution from narrow extent model	26-27
Figure 2.7 AUC Plots of testing and training omission and predicted area varies with the cumulative threshold for narrow extent models	28
Figure 2.8 Receiver operating curves (ROC) for training and test data plots for narrow extent models.....	29
Figure 2.9 Binary maps of projected Eleusine species distribution from narrow extent model	30-32
Figure 2.10 Tree plots of relative contributions of major environmental variables ($\geq 1\%$) in full (i) and the narrow (ii) Maxent models	34-36
Figure 3.1 Summary chart of bioinformatics pipeline for structural variants identification analysis in <i>Eleusine coracana</i>	47
Figure 3.2 Number of structural variant events found for each whole-genome sequence ...	55
Figure 3.3 Number of high confidence structural variant events found for each whole-genome sequence after merging complete overlaps	56
Figure 3.4 Boxplot showing the size distribution of the length of structural variation Events	57-59
Figure 3.5 Boxplot showing the number of evidence (number of reads, number of split	

reads) supporting SV calls in each whole-genome sequence.....	60-62
Figure 3.6 Genomic Distribution of Structural Variants in the <i>Eleusine coracana</i> Genome as viewed in IGV	61
Figure 3.7 Number of genes overlapping identified structural variants events	62
Figure 3.8 GO functional annotation of genes overlapping identified deletion structural variation events	63
Figure 3.9 GO functional annotation of genes overlapping identified duplication structural variation events	64
Figure 3.10 GO functional annotation of genes overlapping identified inversion structural variation events	64

List of Abbreviations

SNP	Single Nucleotide Polymorphism
NCBI	National Center for Biotechnology Information
SRA	Sequence Read Archive
WGS	Whole Genome Sequencing
SV	Structural Variant
SDM	Species Distribution Models
GBIF	Global Biodiversity Information Facility
DEM	Digital Elevation Model
AAEAC	Africa Albers Equal Area Conic
ESDAC	Joint Research Centre-European Soil Data Centre
AUC	Area Under the Curve
ROC	Receiver Operator Characteristic
NGS	Next-Generation sequencing
CNV	Copy-Number Variation

Chapter 1: General Introduction

Grasses (Poaceae), with over 11,500 known species (Duvall *et al.*, 2007; Shchapova, 2012; Christenhuys and Byng, 2016), are the fifth most species-rich group of flowering plants. It includes crop, pasture, and weed species adapted to all key landmasses from warm and cold (Kellogg, 2001; Strömberg, 2011). Within the grass lineage, the PACMAD clade (subfamilies Panicoideae, Aristidoideae, Chloridoideae, Micrairoideae, Arundinoideae, Danthonioideae) consists of closely related species with paramount and ecological and economic importance that have evolved the efficient carbon-fixing, C₄ photosynthesis several times and are well adapted to open vegetation (Cotton *et al.*, 2015; Soreng *et al.*, 2015).

Among the PACMAD grasses, the subtribe Eleusininae Dumort. (Poaceae: Chloridoideae: Cynodonteae) is a morphologically diverse group of about 231 species and 27 genera Eleusininae (Peterson *et al.*, 2015; Soreng *et al.*, 2017; Muchut *et al.*, 2017) occurring primarily at low latitudes in Africa, Asia, Australia, and the Americas (Peterson *et al.*, 2010; Peterson *et al.*, 2015). Generally, they are morphologically characterized as having diverse paniculate inflorescences (Muchut *et al.*, 2017) and mostly exhibit C₄ leaf anatomy (Ellis, 1984).

Eleusine species are herbaceous plants with flattened culms (or stems) erect, prostrate, or angled and flattened at the internode. They have a digitate or sub-digitate inflorescence with spikes arranged into a terminal whorl. Each spike has many laterally compressed spikelets, usually disarticulated at maturity, with the fruit (grain) being unusual, among grasses, ornamented, and enclosed by a thin pericarp (Phillips 1972). The genus, *Eleusine* Gaertn. (Poaceae, subfamily Chloridoideae), is a small taxon of closely related and distinct tufted annuals or perennials that sometimes have rhizomes or stolons (Phillips, 1972; Peterson *et al.*, 2021).

Species in the genus include diploids and tetraploids based on a haploid chromosome number of $n = 8, 9,$ and 10 . Cytological studies suggested showed $n = 9$ as the basic chromosome number in *Eleusine* with $n = 8$ arising from aneuploidy and $n = 10$ arising from a gain in chromosome number (Hiremath and Chennaveeraiah, 1982). *E. coracana* is an allotetraploid ($2n = 4x = 36$, genome formula AABB) that is morphologically similar to both *E. indica* ($2n = 2x = 18$, AA) and *E. africana* ($2n = 4x = 36$, AABB). The other tetraploid in the genus is *E. kigeziensis* ($2n = 4x = 38$, AADD). *E. floccifolia* ($2n = 2x = 18$, BB), *E. intermedia* ($2n = 2x = 18$, AB), *E. jaegeri* ($2n = 2x = 20$, DD), *E. multiflora* ($2n = 2x = 16$, CC) and *E. tristachya* are diploids. *E. semisterilis*, known only from type specimen, is cytologically unknown and probably extinct (Phillips, 1972; Phillips, 1995).

Over decades, the number of *Eleusine* species and their relationships have been scrutinized from vegetative, floral cytological, and molecular evidence. However, the genus is incontestably monophyletic (Kennedy 1957, Philip 1972, De Wet *et al.* 1984, Ganeshaiyah & Umarshaanker 1980, Gasser and Vegetti 1997, Bisht and Mukai 2000, Bisht and Mukai 2002, Neves *et al.* 2005, Liu *et al.* 2011). Current recognition of the species in the genus is essentially shown by Philip (1972). He grouped *E. africana* and *E. indica* as subspecies while extensively describing the vegetative and floral morphology and the life cycle of 9 predominant members found in Africa. This arrangement has been followed by rearranging relationships in the taxon, especially in identifying *E. africana*, *E. coracana*, and *E. indica* as distinct species. Recently, *E. poiflora*, formerly closely related to the *Coelachyrum* genus, was added to the group (Peterson *et al.*, 2021).

Pieces of evidence from cytological, biochemical, and molecular sources reveal that *E. indica* is maternally related to the AA genomes in *E. coracana* and *E. africana* (Hilu, 1995; Werth *et al.* 1994; Liu *et al.*, 2011; Peterson *et al.*, 2015; Soreng *et al.*, 2017; Peterson *et al.*, 2021).

Morphological and genetic proximities between *E. coracana* and *E. africana* also suggest gene flow occurs between them in nature, and probably *E. coracana* originated from *E. africana* through selection (Chennaveeraiah and Hiremath 1974; Hilu and deWet 1976). From ribosomal DNA similarities, Bisht and Mukai (2000, 2001) suggested that *E. floccifolia* is the paternal progenitor for *E. africana* and *E. coracana*. However, this claim has been refuted from nuclear Internal Transcribed Spacers (ITS) and plasmid *trnT-trnF* (region between Threonine and Phenylalanine of chloroplast tRNA gene) sequences (Neves *et al.*, 2005). The sister relationship between *E. indica* and *E. tristachya* and between *E. floccifolia* and *E. jaegeri* are widely accepted from biochemical and genetic evidence (Liu *et al.* 2011; Hiremath and Chennaveeraiah 1982; Hiremath and Salimath 1991; Hilu and Johnson 1992; Peterson *et al.*, 2015; Peterson *et al.*, 2021). Recent plastid phylogeny groups the three tetraploid species and with the *E. indica*–*E. tristachya* clade under a common ancestor. (Liu *et al.* 2014). The close relationship of *Eleusine* species and *Coelachyrum poiflorum* in molecular studies (Liu *et al.*, 2011; Liu *et al.* 2014; Peterson *et al.*, 2015; Soreng *et al.*, 2017) influenced its transition as a member of the group. The evolutionary relationship in the genus is still largely unresolved as paternal progenitor(s) remains unknown (Liu *et al.*, 2011; Liu *et al.*, 2014; Zhang *et al.*, 2019).

East Africa is the center of *Eleusine* diversity, and 9 of the 11 species are found in Africa. Eight species are endemic to Africa (Phillips, 1972; Liu *et al.*, 2011; Peterson *et al.*, 2021). One species, *E. tristachya*, is native to the New World. Generally, *Eleusine* species records are confined to East Africa, occupying a narrow range at high altitudes (Phillips, 1972; Liu *et al.*, 2011). *E. indica* (L.) Gaertn. is documented as a pantropical and introduced weed from all continents except Australia and Antarctica (Phillips, 1995; Liu and Peterson, 2010). *E. coracana* is widely known

for cultivation in sub-Saharan Africa and Asia. The newly added *E. poiflora* extends from southwest Asia into Somalia and Djibouti.

Research Objective and Experimental Context

There is sustained interest in understanding the study of *Eleusine* species which has reached economic and ecological impacts. Variations in temperature and availability of water (majorly from anthropogenic led climate change), with decreasing soil fertility and rising pest and disease occurrence, have led to a stress-induced loss in plant yield (Dhankher and Foyer 2018; Chaudhry and Sidhu, 2021). Identification and adoption of climate-resilient crops (crops with enhanced tolerance to stress) are recognized as a coping mechanism for threats to future food security. *E. coracana* (called finger millet), a historic orphan cereal with modern interest cultivated for grain and fodder, is highly nutritious, adaptable to diverse environments, and drought and disease tolerant. Furthermore, *E. indica* is a widespread weed, notorious for being hard to control due to its high reproductive capacity, herbicide resistance, and wide tolerance to various environments (Holm *et al.*, 1977; Chen *et al.*, 2015). Understanding the complex genetics and traits of *Eleusine* species has enormous benefits for agriculture and other industries.

Therefore, my research objectives are to:

1. characterize the geographical distribution of the different *Eleusine* species in Africa and;
2. investigate genomic structural variations in *E. coracana*.

Chapter 2: Geospatial Characterization and Distribution Mapping of Eleusine Species in Africa

Introduction

Savanna, characterized by the abundance of grasses with widely spaced trees that do not form a canopy, makes up about 50% of the African continent's land surface (Belsky, 1994; Scholes and Archer, 1997). Africa savanna has a rich floristic and physiognomic diversity, and the C₄ grasses are a significant component of their structure (Pasturel *et al.*, 2016; Still *et al.*, 2003).

One exciting group among the C₄ grasses exhibiting significant morphological and ecological diversity in Africa is the *Eleusine* Gaertn. (*Poaceae*, subfamily *Chloridoideae*) genus. It is a taxon of eleven annuals and perennial that includes an essential historical crop (*E. coracana*), a ubiquitous weed (*E. indica*), and other wild-growing individuals (*E. africana*, *E. floccifolia*, *E. intermedia*, *E. jaegeri*, *E. kigeziensis*, *E. multiflora*, and *E. tristachya*).

Eleusine is mainly African (at least eight species; Phillips, 1972), and one species, *E. tristachya*, is endemic to the New World. Generally, all species reportedly occupy a range of habitats from low to high altitudes in Africa (Phillips, 1972; Liu *et al.*, 2011). *E. indica* (L.) Gaertn. is documented as a pantropical and introduced weed in all continents except Australia and Antarctica (Holm *et al.*, 1977; Phillips, 1995; Liu and Peterson, 2010). *E. coracana* (finger millet) is widely known for cultivation in sub-Saharan Africa and Asia. The newly added *E. poiflora* extends from southwest Asia into Somalia and Djibouti.

There is a sustained interest in understanding the biology of *Eleusine* species with reaching economic and ecological impacts. The present and increasing variations in temperature and available water (majorly from anthropogenic led climate change), with decreasing soil fertility and rising pest and disease occurrence, have caused stress-induced losses in plant yield (Dhankher and Foyer 2018; Chaudhry and Sidhu, 2021). Finger millet has been identified as a climate-resilient

crop (crops with enhanced tolerance to stress), thus, a coping mechanism for threats to future food security.

Plant species distribution is mainly associated with water availability, especially at latitudes closer to the equator, where the sun's radiant energy is abundant (Hawkins *et al.*, 2003). Africa spans the equator stretching from the northern temperate to southern temperate zones, and most of the continent is in the tropics. Thus, Africa lands are among the most vulnerable ecosystems to climate change and increasing human pressure (Sala *et al.* 2000, Parr *et al.* 2014). Studies on broad environmental correlates of grassland in Africa (Pasturel *et al.*, 2016; Bocksberger *et al.*, 2016). However, knowledge about specific habitat requirements and the distribution of plant species is lacking. At present, environmental distribution analysis is presently unavailable for *Eleusine* species.

Species Distribution Models (SDMs) correlate environmental conditions (predictor variables) with locations where an organism has been observed (Guisan & Thuiller 2005). SDM uses identified suitable environment layers to predict potential habitats where the species can occur. Maps of potential habitat suitability aid in the species environmental management by identifying potential restoration and protection sites and can lead to the discovery of new populations (Hernandez *et al.* 2006). There are various methods for modeling species distribution. One standard method is the use of presence-only data, which relies only on location records for where the species has never been recorded (Pearce & Boyce 2006).

Digital herbarium records are available for *Eleusine* species collections in Africa on the Global Biodiversity Information Facility (GBIF). These are broadly presence-only data that are useful in modeling distribution. Understanding the factors that determine the present geographic

distribution of *Eleusine* species helps identify and predict potential suitable environmental conditions.

In this study, I characterized the eco-geographical patterning of the different *Eleusine* species in Africa. I utilized available climatic, soil, vegetation, and digital elevation model (DEM) map to gain insights and make substantial predictions about the probability of presence, potential species habitat, and environmental correlates for each species in the genus within Africa. Understanding the distribution, ecology, and population dynamics of *Eleusine* species in Africa could provide insights into the history and relationships in the genus.

Materials and Methods

This study is a broad characterization of environmental predictors for *Eleusine* species in Africa based on collection records on GBIF (the Global Biodiversity Information Facility). Bioclimatic and soil data were used to find out possible indicators for the presence of *Eleusine* species in Africa. Eight *Eleusine* species are recognized as native to Africa, and one species, *E. tristachya*, is endemic to the new world (Phillips 1972). All species have been documented in Africa, and their habitats range from the dry highlands of East Africa to low-lying coastal areas. Figure 2.1 shows a preview of the location of *Eleusine* species collections, created with ArcGIS® pro software by Esri.

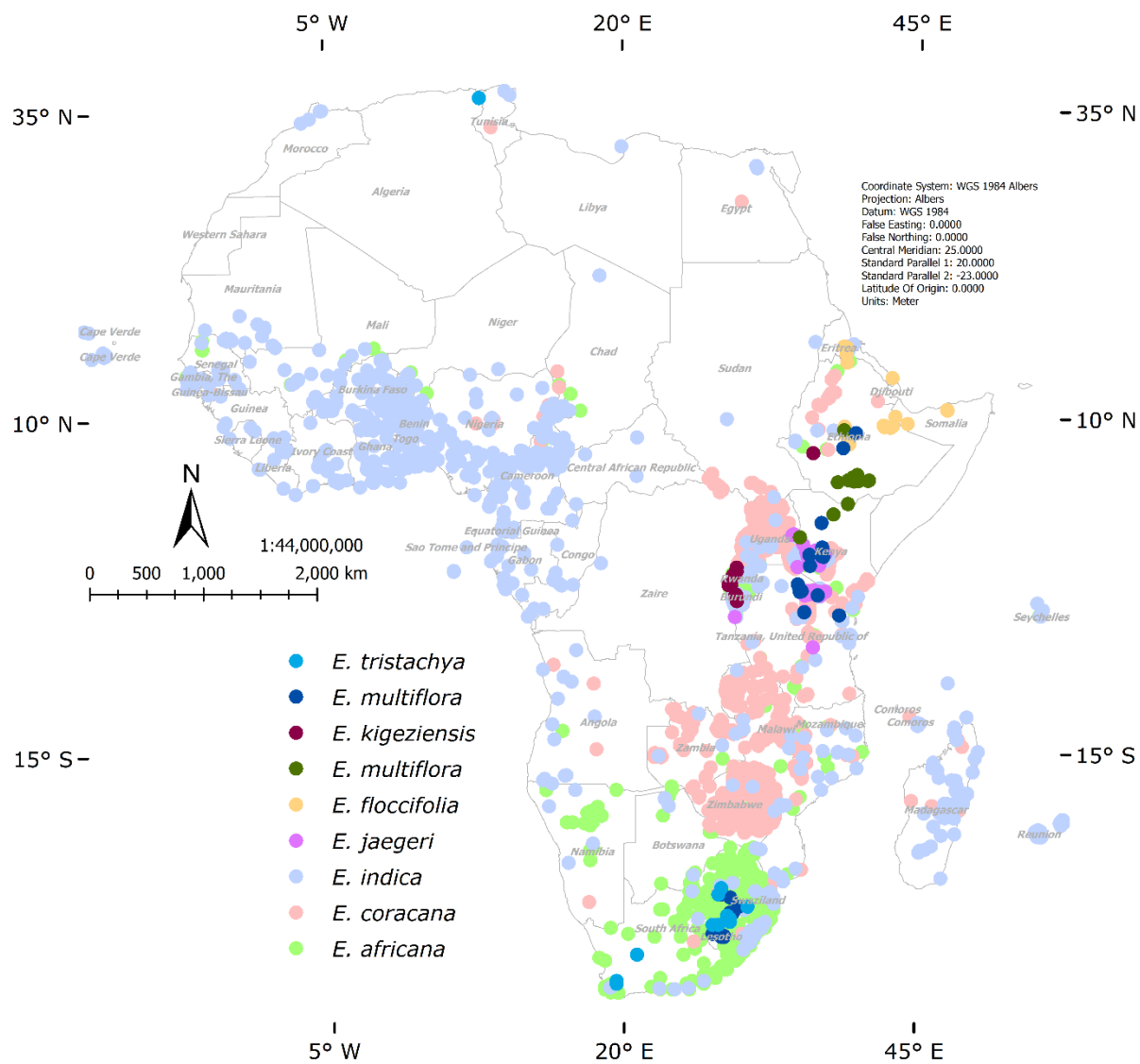


Figure 2.1: Distribution of nine *Eleusine* species across the African continent. A total of 2,961 locations were extracted for Africa collections from the Global Biodiversity Information Facility records (GBIF; <https://doi.org/10.15468/dl.e83gx8> accessed November 24, 2020).

Species extents and Predictor Layers

Africa covers about 12 million sq mi from latitude 37.354722 to latitude -31.854167 and longitude -17.520277 to 51.464444. However, the recorded range on *Eleusine* species in Africa differs among species, with 3 (*E. coracana*, *E. africana*, and *E. indica*) occupying broad ranges

across sub-Saharan Africa and other species confined to east and southern Africa. *E. tristachya* is reported in South Africa and Algeria.

In modeling distribution for *Eleusine* species in Africa, shapefile for Africa, containing South Sudan and Abyei, was downloaded from openAFRICA (https://africaopendata.org/pt_BR/dataset/africa-shapefiles/resource/04ed7565-614d-473e-88b9-2e9208c5cece). This was used to define the map extent for each species in ArcGIS pro using two approaches. First, a full Africa map was used for all species to represent all possible environments in Africa accurately. Using the full extent implies that each species could have dispersed anywhere across the continent. It also means that the whole continent has been considered for sampling.

In the second approach, the modeling extent for each species was narrowed to include only countries where they were reported in GBIF records, reducing artifacts of prediction statistics when modeling with Maxent as advised by Phillips, S. J. (2017). Narrowing the modeling extents also enabled distribution projection to areas where species have not been reported. In this approach, selected countries for a species broadly represent environments where the species have been found. It also implied that the species had dispersed anywhere across the extent. Mapping extent for each species was extracted with the Dissolve tool after selecting the countries from the Attribute Table of the Africa map with SQL commands (Appendix 1)

Each map extent was projected to the Africa Albers Equal Area Conic (AAEAC) projection using the Project Raster tool. AAEAC is a regional scale projection for Africa, with each cell having an equal area. The projection's x-y measurement is in meters, the same units as the z-axis for the elevation layer and its derivatives.

Available general habitat suitability indicators—reported as eco-physiologically meaningful environmental variables affecting the distribution of plants, such as climatic (rainfall and temperature), soil (type and characteristics) (Mod et al., 2016)—were used to model *Eleusine* species distribution in Africa.

To characterize each collection point, I used a set of 33 environmental variables that include climatic and edaphic factors. Nineteen bioclimatic variable maps and accompanying digital elevation model (DEM) were downloaded from WorldClim Version2 (Fick and Hijmans, 2017). These are 30 seconds spatial resolution biologically meaningful variables maps derived from historical monthly temperature and rainfall values for 1970 to 2000. WorldClim biological variables include annual trends, such as mean annual temperature, annual precipitation, and seasonal variables, such as annual range in temperature and precipitation. They also include extreme or limiting environmental factors, such as the temperature of the coldest and warmest months and precipitation of the wet and dry quarters and are often used for modeling species distribution (Fick and Hijmans, 2017). In addition to bioclimatic variables, ten soil property maps for Africa were downloaded from the iSDAsoil dataset (Hengl et al. 2021) soil property for collection points. The iSDAsoil datasets are publicly available Soil Information System raster maps for Africa. At 30-m (1 arc second) spatial resolution, iSDAsoil data offers a higher resolution scale resolution than the smaller scale bioclimatic data from WorldClim v2. However, environment mapping extent definitions in ArcGIS purposely excludes a temporal mismatch and conforms the higher resolution data to the spatial characteristics of the higher resolution by summarizing adjoining areas. Accordingly, these analyses involve a static distribution of species occurrence and thus, can be relaxed (Pacifci *et al.*, 2019).

Furthermore, the Soil Atlas of Africa was downloaded with permission from the Joint Research Centre-European Soil Data Centre (ESDAC) to characterize soil type for the study area. According to the soil metadata, the map presents the soil map of Africa and contains the dominant WRB Reference Soil Group and associated qualifiers. The shapefile map comes with a comprehensive pdf document that details the different soil types (abbreviated as SU_WRB1 in the attribute table). Table 2.1 shows a summary of downloaded environmental variable maps.

Table 2.1: Summary of downloaded environmental layer maps. iSDAsoil data 30-m (1 second) resolution scale was conformed to bioclimatic variables resolution in ArcGIS to prevent temporal mismatch by summarizing adjoining areas.

Symbol	Environmental Layer	Resolution	Source
BIO1	Annual Mean Temperature	30 seconds	WorldClim version 2.1
BIO2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	30 seconds	WorldClim version 2.1
BIO3	Isothermality (BIO2/BIO7) ($\times 100$)	30 seconds	WorldClim version 2.1
BIO4	Temperature Seasonality (standard deviation $\times 100$)	30 seconds	WorldClim version 2.1
BIO5	Max Temperature of Warmest Month	30 seconds	WorldClim version 2.1
BIO6	Min Temperature of Coldest Month	30 seconds	WorldClim version 2.1
BIO7	Temperature Annual Range (BIO5-BIO6)	30 seconds	WorldClim version 2.1
BIO8	Mean Temperature of Wettest Quarter	30 seconds	WorldClim version 2.1
BIO9	Mean Temperature of Driest Quarter	30 seconds	WorldClim version 2.1
BIO10	Mean Temperature of Warmest Quarter	30 seconds	WorldClim version 2.1
BIO11	Mean Temperature of Coldest Quarter	30 seconds	WorldClim version 2.1
BIO12	Annual Precipitation	30 seconds	WorldClim version 2.1
BIO13	Precipitation of Wettest Month	30 seconds	WorldClim version 2.1
BIO14	Precipitation of Driest Month	30 seconds	WorldClim version 2.1
BIO15	Precipitation Seasonality (Coefficient of Variation)	30 seconds	WorldClim version 2.1
BIO16	Precipitation of Wettest Quarter	30 seconds	WorldClim version 2.1
BIO17	Precipitation of Driest Quarter	30 seconds	WorldClim version 2.1
BIO18	Precipitation of Warmest Quarter	30 seconds	WorldClim version 2.1
BIO19	Precipitation of Coldest Quarter	30 seconds	WorldClim version 2.1
NA	Digital elevation model (DEM)	30 seconds	WorldClim version 2.1
NA	Soil pH for Africa at 0–20cm depth intervals	30 m	iSDAsoil dataset
NA	Soil pH for Africa at 20-50cm depth intervals	30 m	iSDAsoil dataset
NA	Soil organic carbon for Africa at 0–20cm depth intervals	30 m	iSDAsoil dataset
NA	Soil organic carbon for Africa at 20-50cm depth intervals	30 m	iSDAsoil dataset
NA	Soil total carbon for Africa at 0–20cm depth intervals	30 m	iSDAsoil dataset
NA	Soil total carbon for Africa at 20-50cm depth intervals	30 m	iSDAsoil dataset
NA	Soil total organic Nitrogen for Africa at 0–20cm depth intervals	30 m	iSDAsoil dataset
NA	Soil total organic Nitrogen for Africa at 20-50cm depth intervals	30 m	iSDAsoil dataset
NA	Soil effective Cation Exchange Capacity (eCEC) for Africa at 0–20cm depth intervals	30 m	iSDAsoil dataset
NA	Soil effective Cation Exchange Capacity (eCEC) for Africa at 20-50cm depth intervals	30 m	iSDAsoil dataset

Spatial data layers were created from downloaded environmental variables in ArcGIS pro. First, the DEM layer was clipped with the Extract-By-Mask tool using Africa polygon shapefile as the modeling extent with the following settings—output projection, Africa Albers Equal Area Conic (AAEAC), extent, Africa polygon shapefile. All other options were left at default. The clipped elevation layer, composed of 27,666 columns and 25,702 rows, and a cell size of 311, was used as the modeling extent for other bioclimatic and soil characteristics maps. In doing this, the output projection, extent, snap raster, and cell size options were all set to the DEM output. This step guaranteed that all final layers had the same cell size, spatial reference, and extent (number of rows and columns) and minimized runtime errors in the Maxent program.

Slope and aspect layers were extracted from the DEM layer to determine if they affect the distribution of *Eleusine* species. The attribute table of the soil type map was improved with information about each soil type from the accompanying document using the Join tool. The shapefile was then converted to raster with the Polygon to Raster tool using soil types as the classification criteria (Table 2.2). All environment layers were saved and exported to ASCII (.asc) files and used as covariates for distribution modeling.

Table 2.2: Broad soil categorization used in Soil Atlas of Africa downloaded from the Joint Research Centre-European Soil Data Centre (ESDAC)

Assigned Value	Soil type
-1	Cells with no data
0	Water/
1	Calcisols
2	Durisols
3	Kastanozems
4	Fluvisols
5	Cambisols
6	Regosols
7	Vertisols
8	Leptosols
9	Solontez
10	Luvisols
11	Nitisols
12	Solonchaks
13	Gypsisols
14	Planosols
15	Arenosols
16	Phaeozems
17	Andosols
18	Plinthosols
19	Acrisols
20	Gleysols
21	Lixisols
22	Histosols
23	Ferralsols
24	Alisols
25	Stagnosols
26	Chernozems
27	Umbrisols
28	Podzols
29	Technosols

Species Presence Data

An up-to-date herbarium data on *Eleusine* collections worldwide was downloaded from GBIF (November 24, 2020, <https://doi.org/10.15468/dl.e83gx8>). The 57,241 global *Eleusine* records were parsed with a custom R script. A summary is presented in Table 2.3.

Africa collection records were filtered from the total records. After removing samples present in living collections and not representative of native climates, 5,892 records, for which no GPS coordinates were provided, were extracted, and their coordinates were determined with GEOlocate software (Rios and Bart 2010) (online) using the accompanying description of the collection location. The first coordinates were chosen for records with multiple suggestions in GEOlocate. Three hundred and one different coordinates were added through GEOlocate. The large dataset ensures analyzable representation for each species. All coordinates were checked for general accuracy, with duplicates and default placements removed. Each final record represented a unique herbarium collection and, if best-collecting practices were followed, should constitute a population of 20 individuals or more. For this analysis, each collection was treated as presence-only data and representative of flowering plant populations at a specific point in time. Individual *Eleusine* species record was extracted to an independent table and exported as a CSV file for processing in ArcGIS pro.

In ArcGIS, decimal degrees geocoordinates were converted to spatial data in geographic coordinate system (GCS) projection by converting CSV files to XY data points. They were then projected to the same projection (AAEAC) as the predictor layers in ArcGIS. The Add-XY-Coordinate tool was used to generate the equivalent XY coordinates for the new projection. The final table was exported as CSV files and used as point data for *Eleusine* species occurrence after deleting unneeded columns in MS-Excel 365.

Table 2.3: Summary of *Eleusine* herbarium records downloaded from GBIF (accessed November 24, 2020) and parsed in R to remove collection records which are impossible to locate as well as duplicate geocoordinates

Total Number of Specimen Downloaded from GBIF	57,241
Total record from Africa	15,933
Total number of recorded African species	9
Number of Presence Records Per Species	
<i>E. africana</i>	370
<i>E. coracana</i>	1321
<i>E. floccifolia</i>	24
<i>E. indica</i>	1086
<i>E. multiflora</i>	19
<i>E. jaegeri</i>	88
<i>E. kigeziensis</i>	13
<i>E. multiflora</i>	28
<i>E. tristachya</i>	12

Distribution Modeling and Statistical Analysis

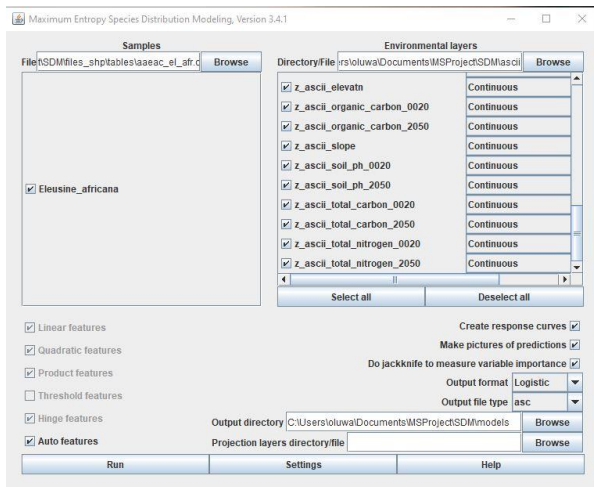
The distribution models for 9 *Eleusine* species were individually analyzed, with full Africa extent and narrowed extent, in Maximum Entropy Species Distribution Modelling (Maxent) version 3.4.4 (Nick et al., 2011) using custom settings (Fig. 2.2). Maxent, written in java, utilizes maximum entropy in modeling species distributions from the presence and environmental data (Phillips et al. 2006). Maxent software was run in Windows 10 ® environment on an 8th gen Intel Core i7 CPU with 16Gb RAM. The default java headspace was raised to 6144Mb in system settings to ensure adequate allocation of computational resources for running Maxent. The memory option was also increased in the BAT file provided with Maxent download and the modified BAT file used to start the program.

CSV files of presences were added in the samples window and the ASCII (.asc) environmental layers in the environmental layers window. All ASC files were in the same folder and thus automatically added to Maxent by choosing one. The following boxes were checked—

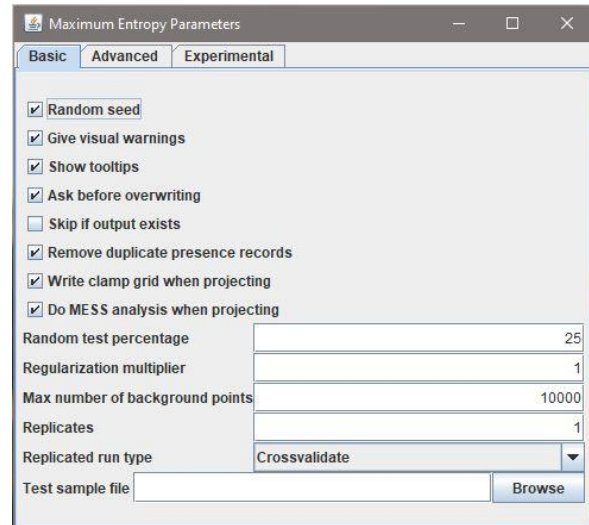
Create Response Curves, Make Pictures of Predictions, and Do jackknife. Output format was set to as Logistic, and output file type, ASC. The desired Output Directory folder for the Maxent output was specified. The full Africa extent layers folder was selected for models with narrow extent to make projections from the modeling. Other feature-types option in the bottom left of the homepage was left at default.

In the Settings window, the Random Seed box was checked, and a value of 25 was entered in the random test percentage box. The Max number of background points was set to 10,000, and cross-validation was checked to simulate enough random sampling for species with a small presence dataset. The default options were accepted in the Advanced tab but saved plot data was ticked to explore the output. The experimental tab was left as is, but the “write background predictions” box was checked to output variables used to calculate sensitivity and specificity, and therefore the TSS (True Skill Statistic) in the model.

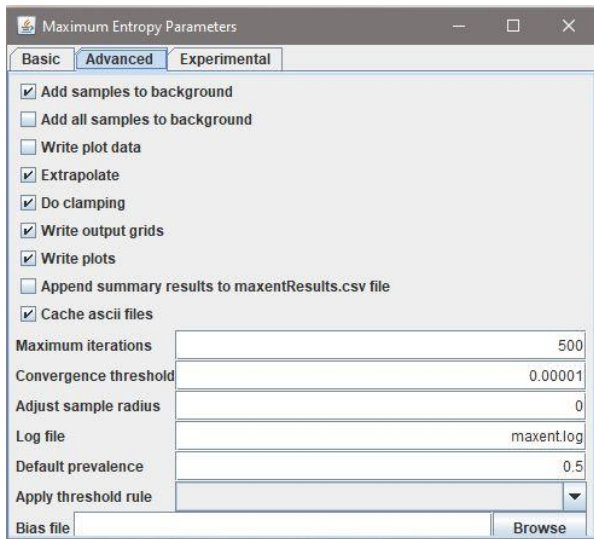
Maxent creates transformations of the covariates, called features, extract a sample of background locations, and contrast them against the presence locations. The logistic format of the output, introduced in version 3 to make it easier to interpret Maxent output, is better calibrated and works the same way as the raw output previously used.



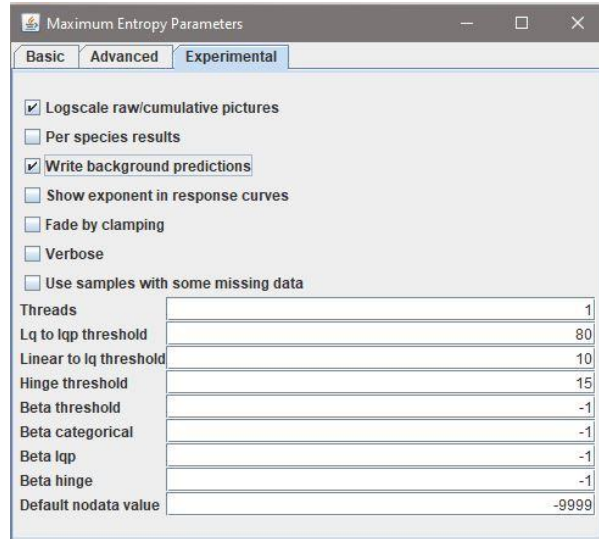
A



B



C



D

Figure 2.2: Maxent settings used for modeling distribution of *Eleusine* species in Africa showing custom options selected in the: main interface (A), basic setting interface (B), advanced setting interface (C), and experimental setting interface (D).

Potential Species Distribution in Africa Predictions

The potential distributions of *Eleusine* species were estimated by using the narrowed Maxent model to make projections over bioclimatic, edaphic, and topographic layers of full Africa extent. The projection layers were prepared in ArcGIS as described for full distribution extent layers. For Maxent to recognize these layers appropriately, they were saved as ASC files using the

same name as corresponding predictor layers. The folder containing all projection layers was selected in the main interface of Maxent under the Projection Layer directory/file option.

Viewing Distribution Models

Binary maps showing areas suitable and not suitable for *Eleusine* species distribution were prepared from ASC output files of the Maxent models using the logistic threshold maximum training sensitivity plus specificity cutoff values in ArcMap 10.7.1. This required building pyramids with defaults settings to allow for proper display and reasonable resolution of the ASC files, defining projection for proper spatial reference, and computing histograms of probabilities to render the logistic output.

Results

Maxent models indicated conditions typical of where species were found, and usually, output distribution extends to areas where they have not been reported but are suitable for the species. Maxent ran distribution modeling on all *Eleusine* species with the 33 environmental layers provided successfully. However, due to spatial incorrectness, the program removed few presence data from *E. africana*, *E. coracana*, and *E. indica*. The results generally show a high degree of variation among the species in their probability of occurrence on the continent.

Model Predictions

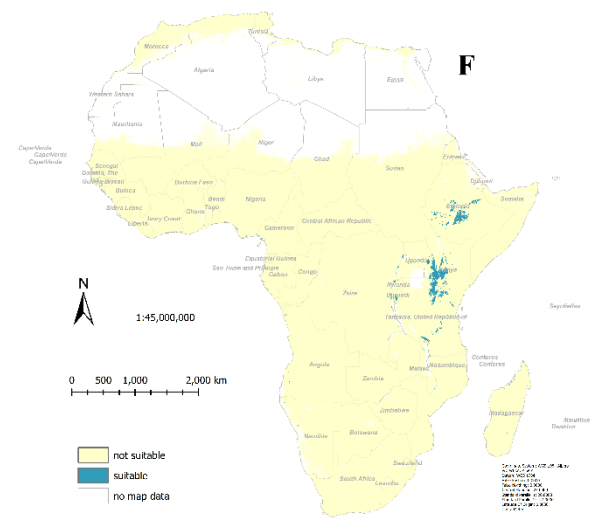
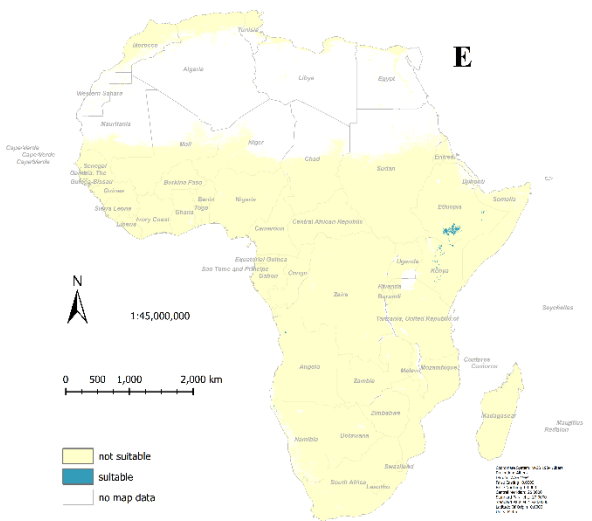
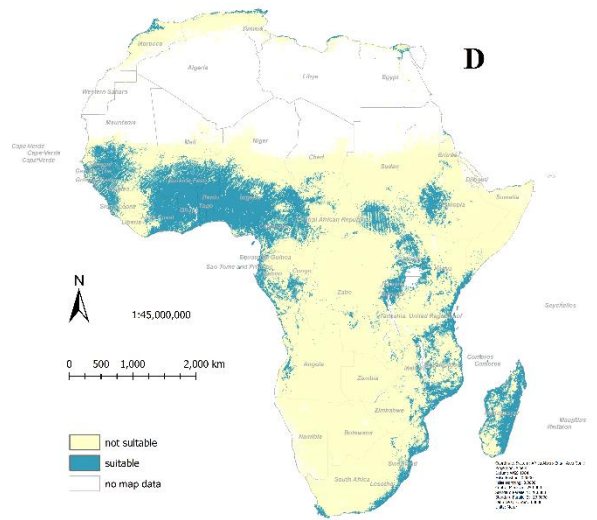
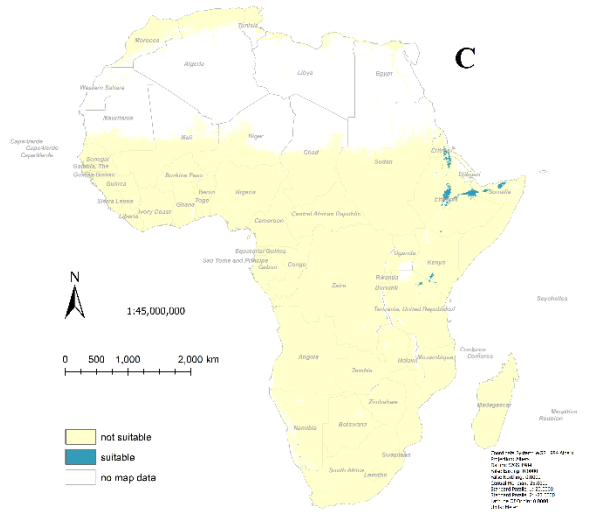
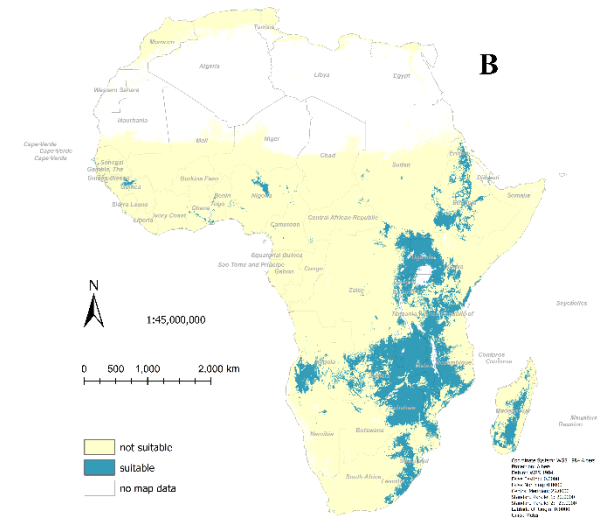
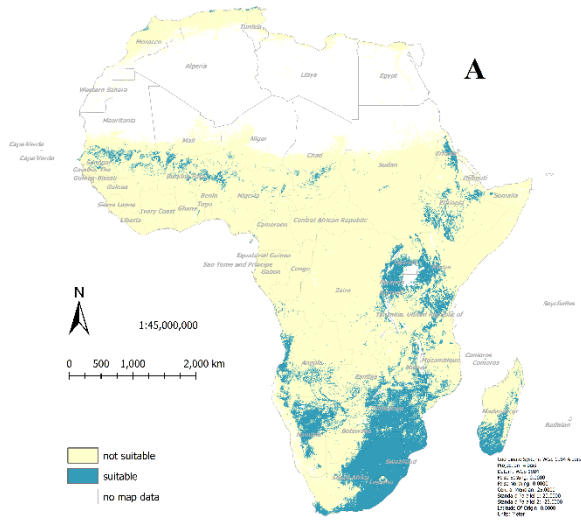
Full Africa Extent Maxent Models

The generated habitat distributions in the model using full Africa extent are constrained to only areas where the species have not been reported. Binary maps, presented in Figure 2.3, shows areas suitable and not suitable for species distribution using the logistic threshold cutoff values from maximum training sensitivity plus specificity (Table 2.4). *E. africana* and *E. coracana*

occupy an extensive range in eastern and southern Africa, with some scattered locations in the West. On the contrary, *E. indica* is more distributed in west Africa than in the east and south. *E. intermedia*, *E. floccifolia*, *E. jaegeri*, and *E. kigeziensis* are uniquely constrained to east Africa. *E. multiflora* has a unique patch in the east and south of the continent. *E. multiflora* appears typically distributed in south and north Africa distributions, especially closer to the coasts. Importantly, all countries where a species have not been reported (as shown in Fig 2.3) typically show no probabilities of occurrence.

Table 2.4: Logistic threshold cutoff values from maximum training sensitivity plus specificity (maximum value = 1) for the full Africa extent Maxent models containing thirty-three environmental variables.

Species	Maximum training sensitivity plus specificity	P-value
<i>E. africana</i>	0.200	9.15e-41
<i>E. coracana</i>	0.219	0.00e+00
<i>E. floccifolia</i>	0.382	1.83e-12
<i>E. indica</i>	0.312	0.00e+00
<i>E. intermedia</i>	0.692	2.63e-05
<i>E. jaegeri</i>	0.111	4.05e-37
<i>E. kigeziensis</i>	0.221	3.87e-05
<i>E. multiflora</i>	0.037	1.01e-08
<i>E. tristachya</i>	0.182	1.25e-01



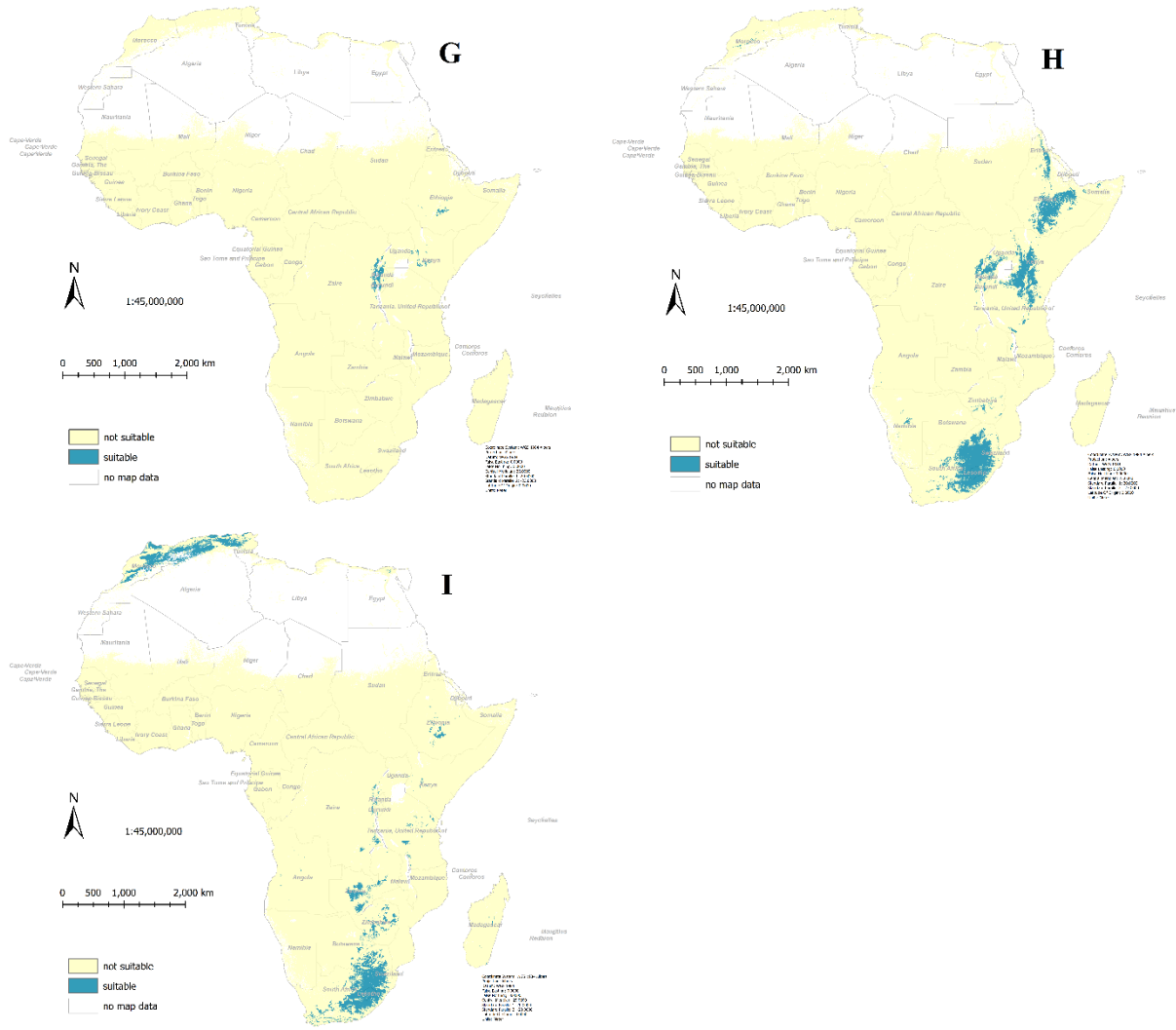
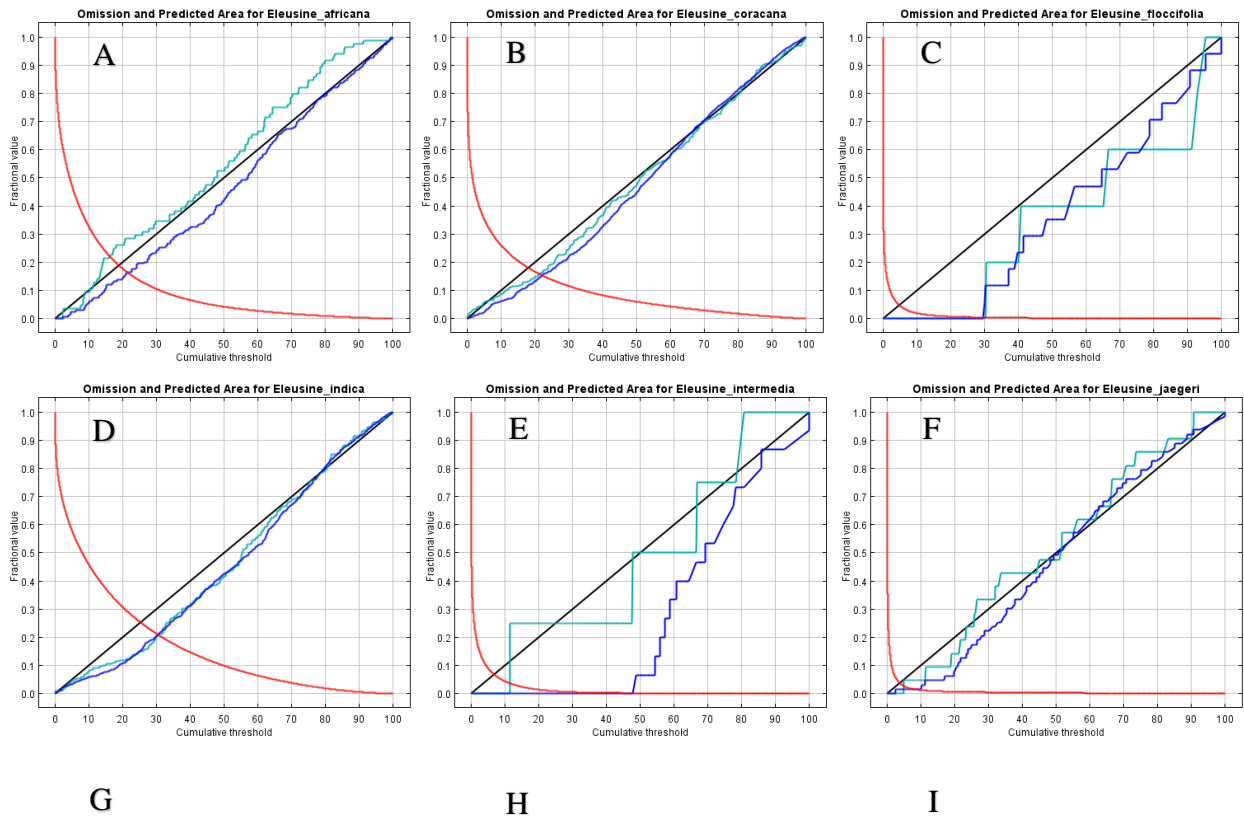


Figure 2.3: Binary maps showing *Eleusine* species distribution (using the logistic threshold cutoff values from maximum training sensitivity plus specificity) as predicted with full Africa extent in Maxent models using thirty-three environmental variables. The dark color indicates areas with a high probability of predictions (suitable). The light color indicates areas with a low probability of predictions (not suitable) (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Performance measures of full Africa extent models

The performances of the potential distributions of *Eleusine* species prediction models were provided in Maxent by calculating the area under the curve (AUC) of the receiver operator characteristic (ROC). Two types of plots were provided. The first plots show how testing

(turquoise) and training (blue) omission and predicted area varies with the cumulative threshold. The omission on test samples is mainly close to the predicted omission rate, showing the model ran as expected in Maxent. However, the line graphs for species with fewer presence data (*E. floccifolia*, *E. intermedia*, *E. kigeziensis*, and *E. tristachya*) were not as smooth as those with more extensive data sizes (figure 2). A strong predictive performance of Maxent model can be seen for species such as *E. africana*, *E. coracana*, *E. indica*, and *E. jaegeri*, with high number of presence data points, but the predictive performance of the model was weak for species fewer data points such as *E. floccifolia*, *E. intermedia*, *E. kigeziensis*, and *E. tristachya* (Fig. 2.4).



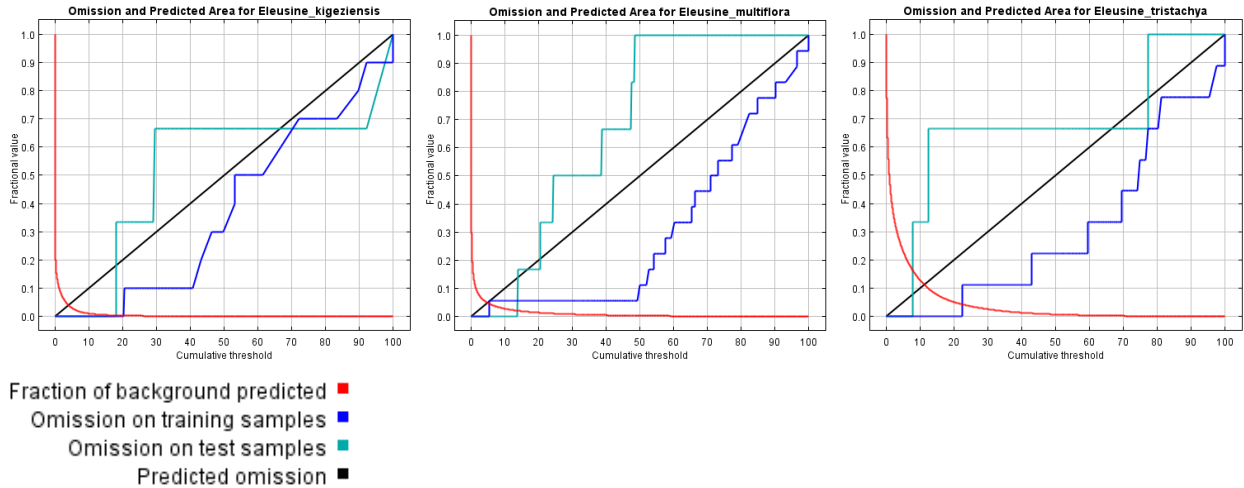


Figure 2.4: Plots show how omission (testing and training) and predicted area varies with the cumulative threshold for the full Africa extent Maxent models containing thirty-three environmental variables. (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

In the second plot, Receiver operating curves (ROC) for training and test data (Fig. 2.5) are plots of sensitivity (the proportion of true positives) versus 1-specificity (proportion of false negatives) over the whole range of threshold values between 0 and 1. The training plot (red line) indicates the fit of the model to the training data, while the test plot (blue line) indicates the fit of the model to the test data (predictive power) (Philips, 2017). The ROC values for the full Africa extent Maxent models are greater than 0.8 for all species.

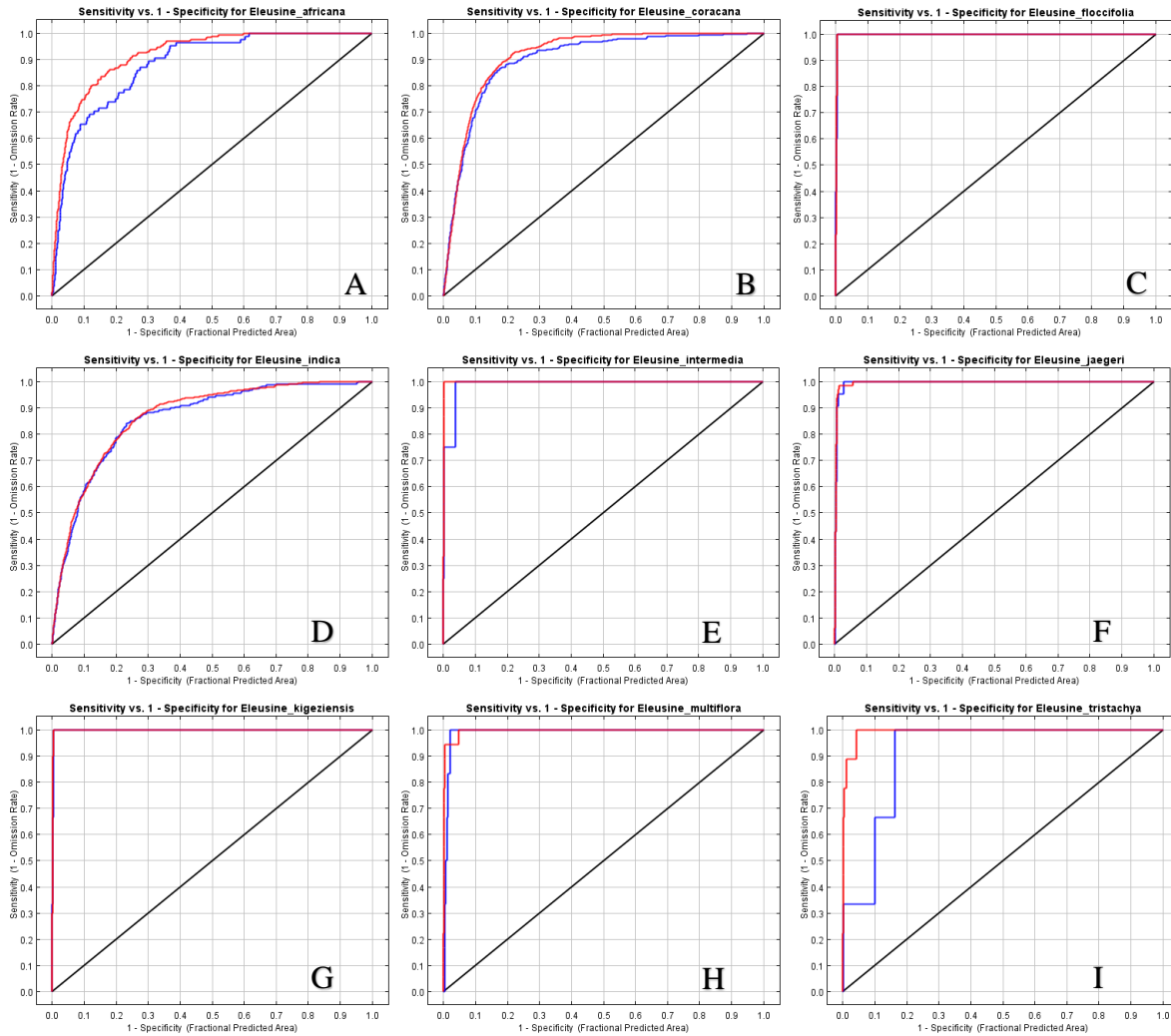


Figure 2.5: Receiver operating curves (ROC) for training and test data plots (sensitivity—the proportion of true positives versus 1-specificity—the proportion of false negatives over the whole range of threshold values between 0 and 1) for the full Africa extent Maxent models containing thirty-three environmental variables. (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Narrowed Extent Maxent Distribution Models

Generated habitat distributions in the model using the narrowed extent are more constrained than those using full Africa extent. Binary maps showing areas suitable and not suitable for species distribution using the logistic threshold cutoff values from maximum training sensitivity plus specificity (Table 2.5) are presented in Figure 2.6. The logistic threshold cutoff

values from maximum training sensitivity plus specificity (Table 2.5) ranged from 0.177 in *E. jaegeri* to 0.629 in *E. tristachya*.

Table 2.5: Logistic threshold cutoff values from maximum training sensitivity plus specificity (maximum value = 1) for the narrow extent Maxent models containing thirty-three environmental variables.

Species	Maximum training sensitivity plus specificity	P-value
<i>E. africana</i>	0.281	1.21e-51
<i>E. coracana</i>	0.302	0.00e+00
<i>E. floccifolia</i>	0.434	1.00e+00
<i>E. indica</i>	0.391	0.00e+00
<i>E. intermedia</i>	0.567	9.94e-09
<i>E. jaegeri</i>	0.177	4.51e-28
<i>E. kigeziensis</i>	0.278	4.00e-04
<i>E. multiflora</i>	0.251	1.57e-03
<i>E. tristachya</i>	0.629	1.76e-01

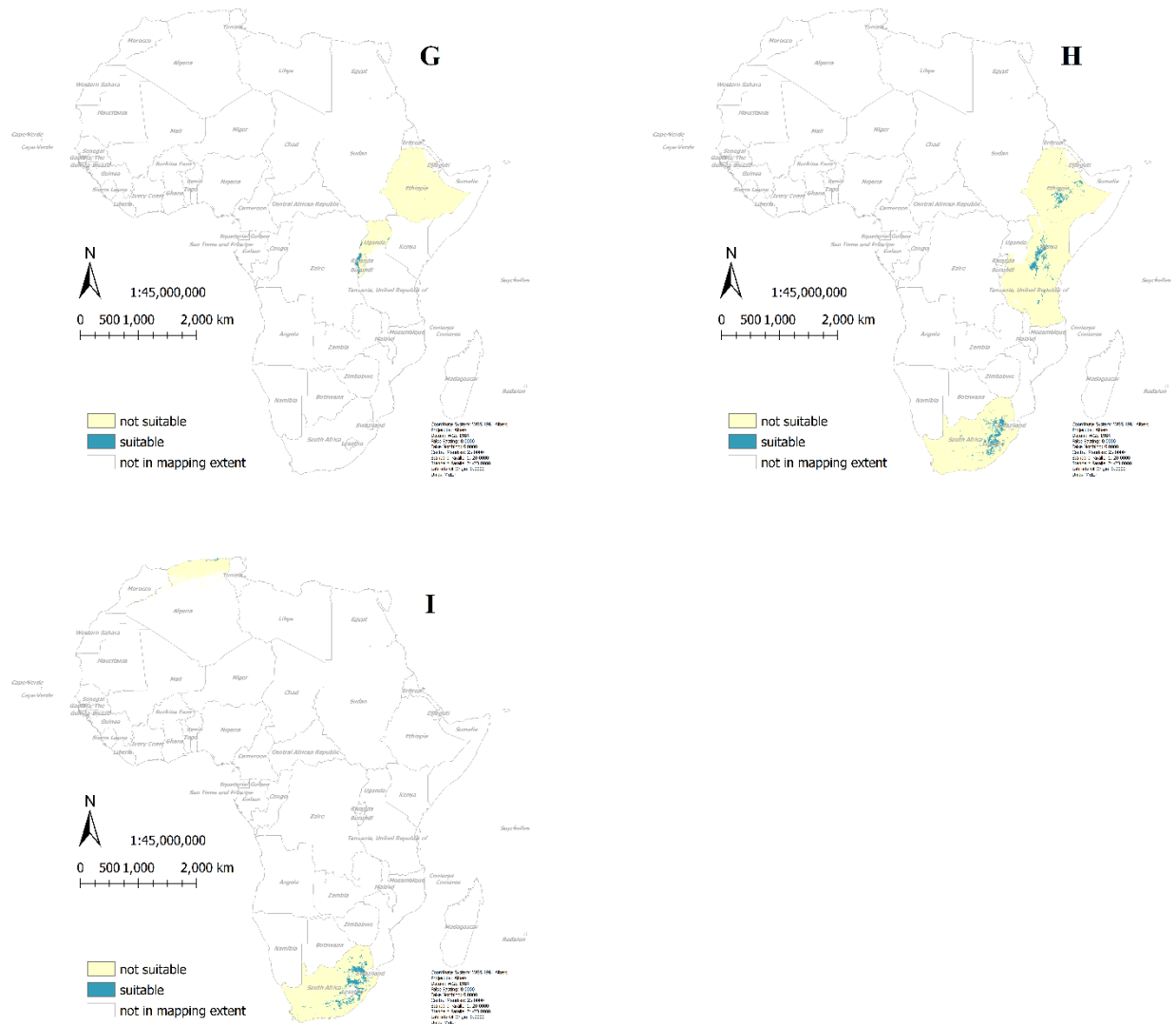


Figure 2.6: Binary maps showing *Eleusine* species distribution (using the logistic threshold cutoff values from maximum training sensitivity plus specificity) as predicted by the narrow Africa extent Maxent models containing thirty-three environmental variables. The dark color indicates areas with a high probability of predictions (suitable). The light color indicates areas with a low probability of predictions (not suitable) (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Performance measures of narrow Africa extent models

Testing and training omission plots of the Maxent narrow extent models are presented in

Figure 2.7. The omission on test samples is mainly close to the predicted omission rate, showing the model ran as expected in Maxent. The predictive performance plots are similar to

corresponding Maxent full Africa extent models. The strength of the predictive performance of Maxent model is stronger for species such as *E. africana*, *E. coracana*, *E. indica*, and *E. jaegeri*, with high number of presence data points, and weaker for species with fewer data points such as *E. floccifolia*, *E. intermedia*, *E. kigeziensis*, and *E. tristachya*.

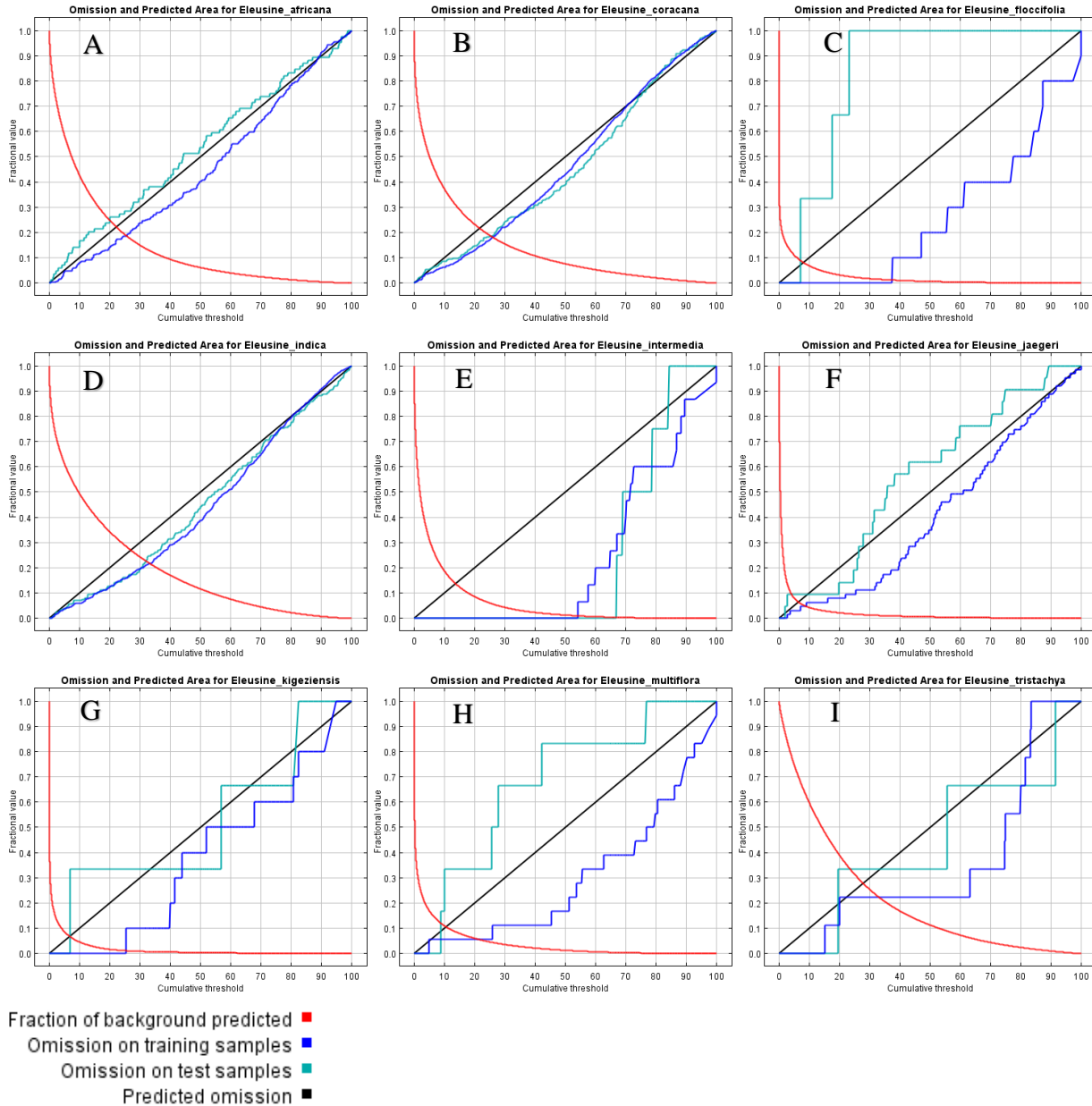


Figure 2.7: Plots show how omission (testing and training) and predicted area varies with the cumulative threshold for the narrow Africa extent Maxent models containing thirty-three environmental variables. (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Receiver operating curves (ROC) for training and test data of Maxent narrow extent models are presented in Figure 2.8. The predictive power of the models is high as the ROC values are greater than 0.8 for all species.

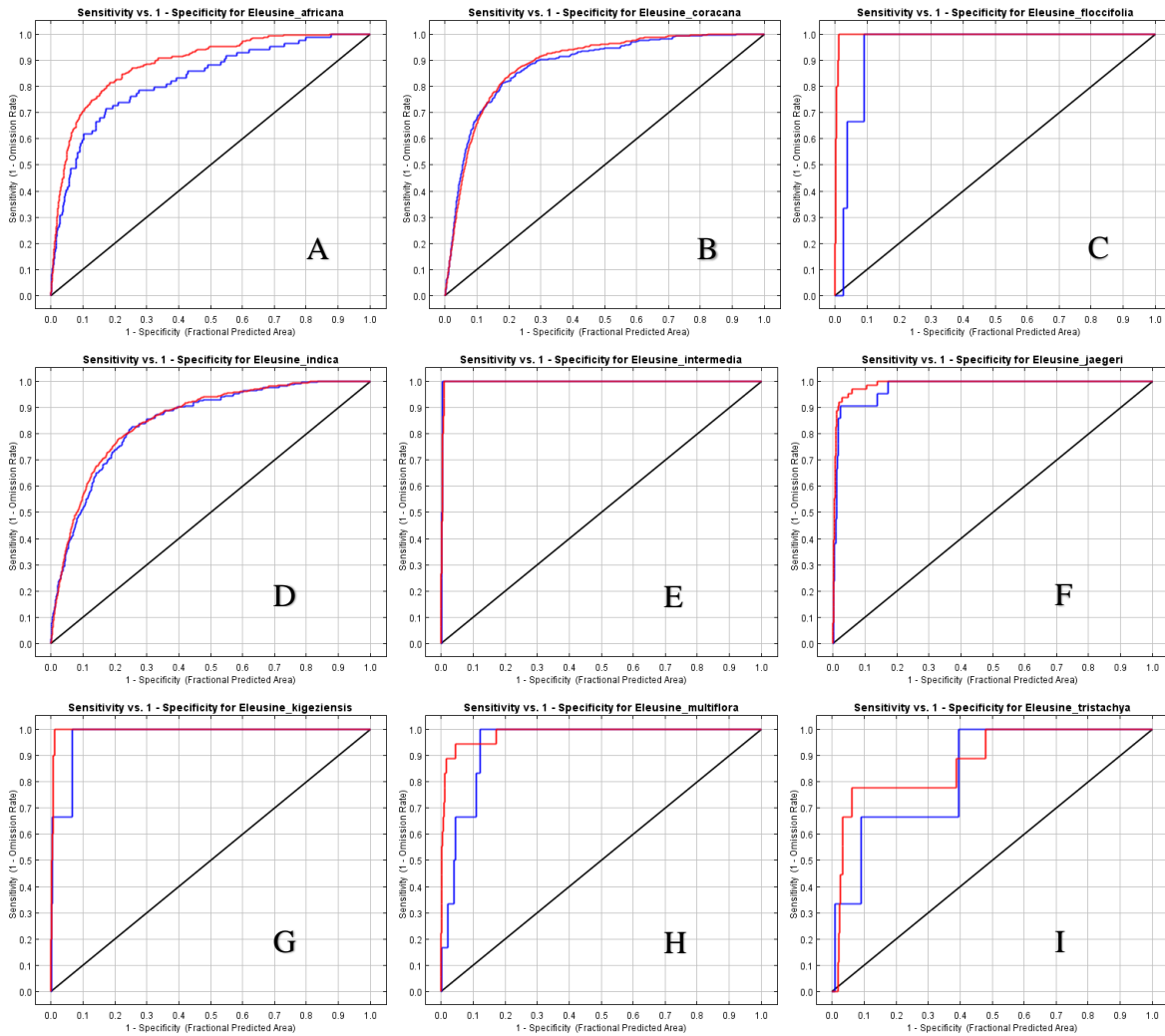
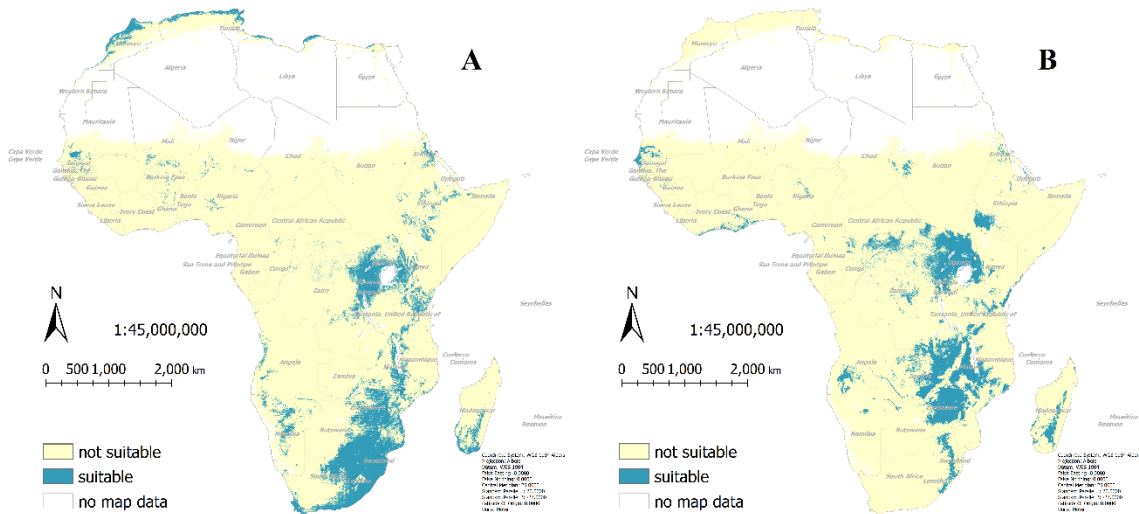
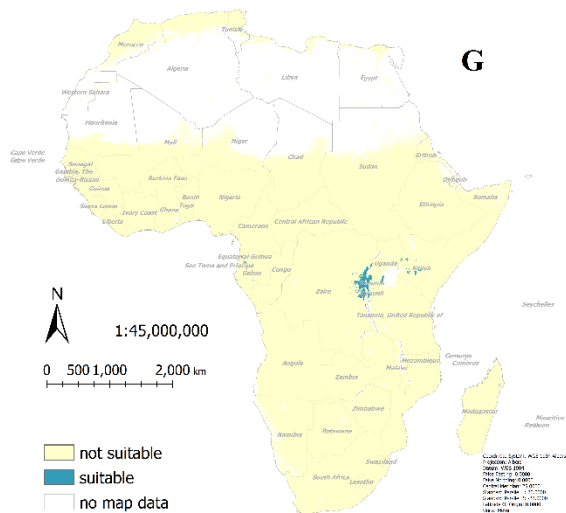
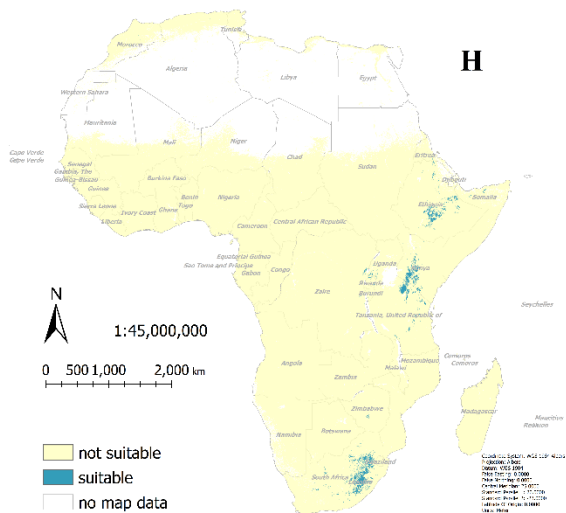
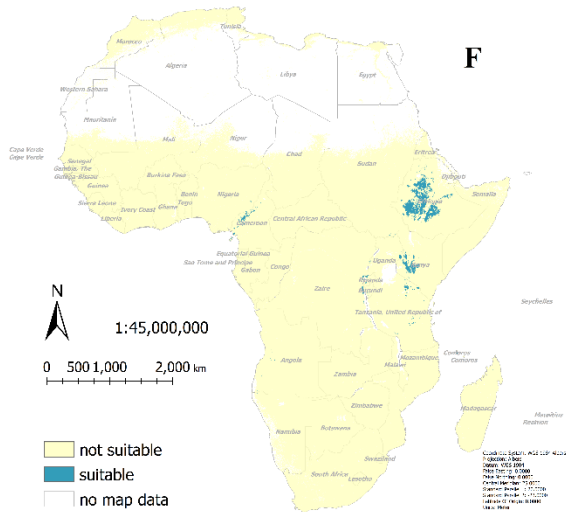
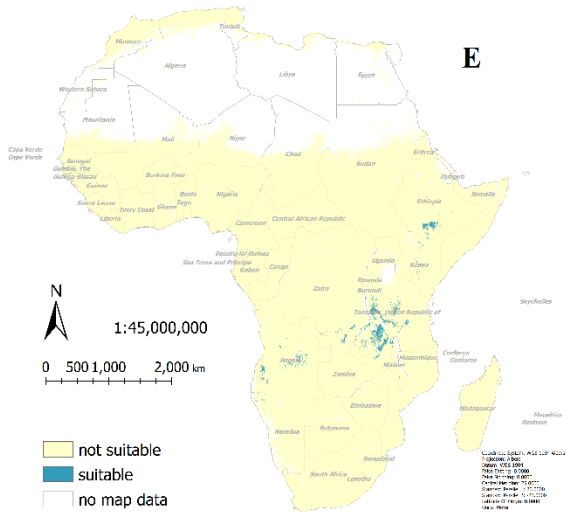
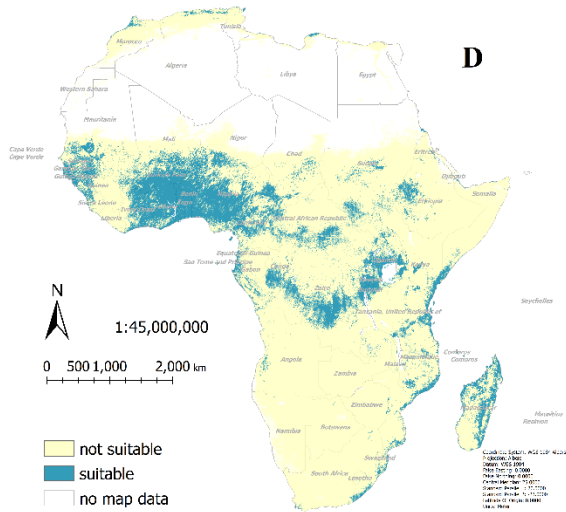
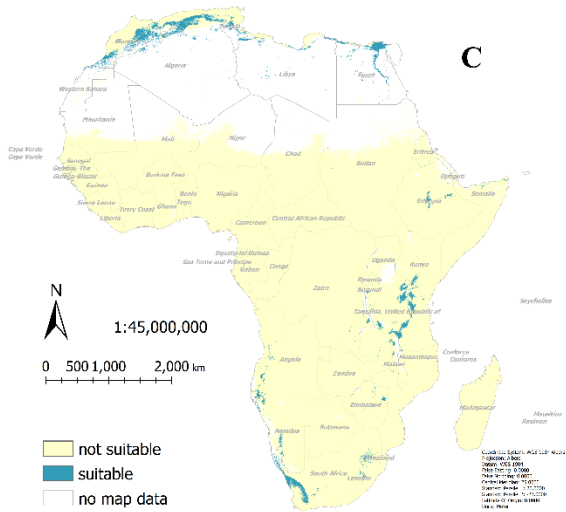


Figure 2.8: Receiver operating curves (ROC) for training and test data plots (sensitivity—the proportion of true positives versus 1-specificity—the proportion of false negatives over the whole range of threshold values between 0 and 1) for the narrow extent Maxent models containing the 33 environmental variables (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Narrow Extent Predictions

I used the narrow extent Maxent model to investigate the potential distribution range of *Eleusine* species in Africa. The model predictions shown in Figure 2.9 indicate that potential *Eleusine* species occurrence extends to other parts of Africa where they have not been reported mainly for species with small known ranges. These are the binary maps of the narrow extent Maxent model projections for *Eleusine* species onto the thirty-three environmental variables for Africa, using the logistic threshold cutoff values from maximum training sensitivity plus specificity. Warmer colors show areas with better-predicted conditions. *Eleusine africana* and *E. floccifolia* occurrence are predicted to extend to the northern parts of the continent. The probability of occurrence is the least for *E. multiflora*.





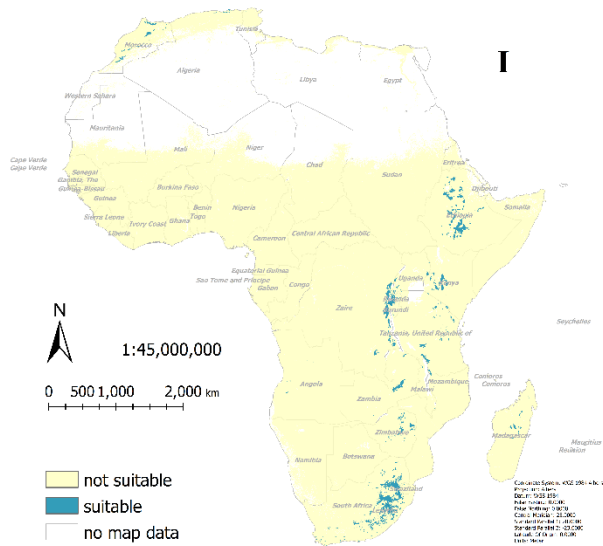


Figure 2.9: Binary maps showing projected *Eleusine* species distribution (using the logistic threshold cutoff values from maximum training sensitivity plus specificity) as predicted by the narrow Africa extent Maxent models containing the 33 environmental variables. The dark color indicates areas with a high probability of predictions (suitable). The light color indicates areas with a low probability of predictions (not suitable) (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Analysis of variable contributions

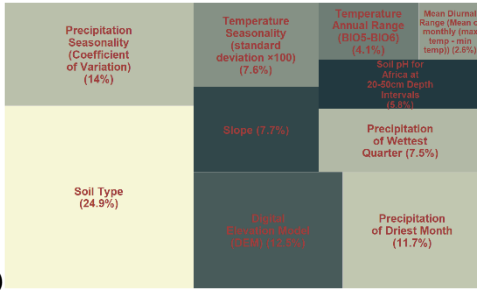
Table 2.6 shows the number of environmental variables with a relative contribution greater than or equal to one percent. *E. coracana* and *E. indica*, have the highest number (20) of substantial environmental factor in the full extent model. *E. indica* has the highest number of contributing environmental factors in the narrow extent model. Distribution models for other species had between ten to fourteen predictor variables, except *E. multiflora* with eight predictor variables in the narrow extent model and *E. tristachya* with only for relevant variables in both models. The relative contributions of the substantial (greater than 1%) environmental variables used in the full and narrow extent Maxent models are presented side by side for each species in Figure 2.10. There are many major overlaps in the contribution of the environmental variables used for predictions in the two models. Different biologically relevant aspects of temperature and precipitation are the

most consistent predictor variable in the two models for *Eleusine* species. However, elevation featured as a high (> 10% relative contribution) contributing factor for *E. coracana*, *E. floccifolia*, *E. indica*, *E. intermedia*, *E. jaegeri*, and *E. multiflora* in the full extent models. The observed high contribution of elevation is only true for *E. coracana*, *E. indica*, *E. jaegeri*, and *E. multiflora* in the narrow extent models. Soil type has a high contribution to the distribution modeling of *E. intermedia* and *E. tristachya* in the full extent and the distribution modeling of *E. floccifolia*, *E. intermedia*, *E. kigeziensis*, *E. multiflora*, and *E. tristachya* in the narrow extent model. The relative contribution of slope was high (>10 %) for *E. africana* and *E. intermedia* only in the narrow extent models.

Table 2.6: Number of substantial (with relative contribution greater than or equal to $\geq 1\%$) environmental variables for the full and the narrow extent Maxent models.

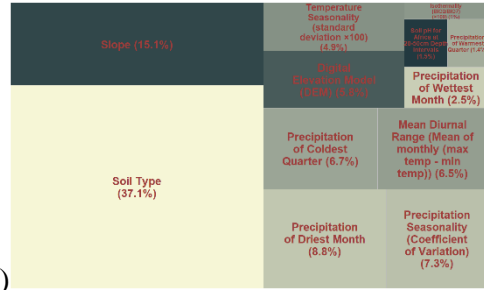
Species	Number of substantial ($\geq 1\%$) environmental variables	
	Full Extent Model	Narrow Extent Model
<i>E. africana</i>	13	14
<i>E. coracana</i>	20	13
<i>E. floccifolia</i>	10	13
<i>E. indica</i>	20	17
<i>E. intermedia</i>	12	10
<i>E. jaegeri</i>	12	13
<i>E. kigeziensis</i>	17	13
<i>E. multiflora</i>	12	8
<i>E. tristachya</i>	4	4

Full

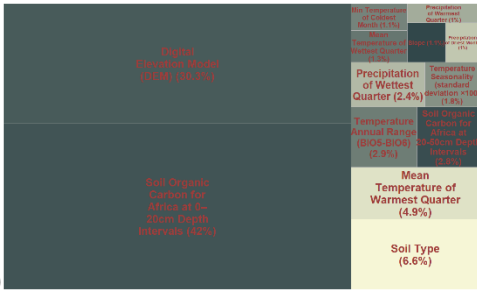


E(i)

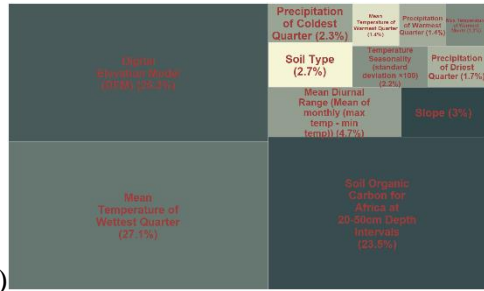
Narrow



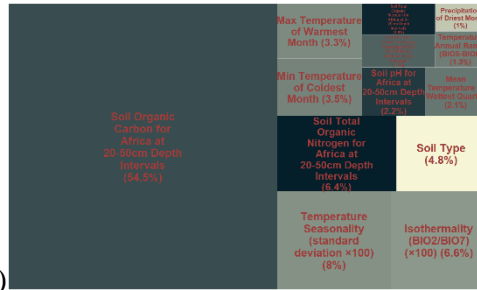
E(ii)



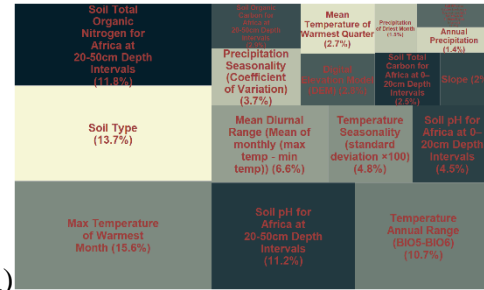
F(i)



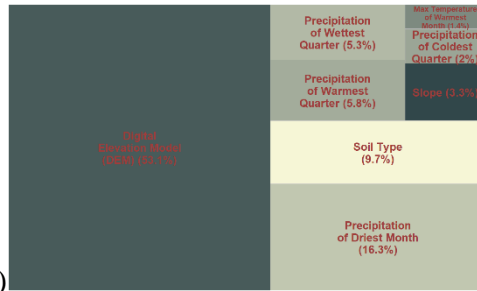
F(ii)



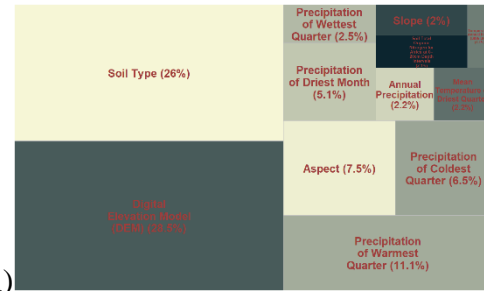
G(i)



G(ii)



H(i)



H(ii)

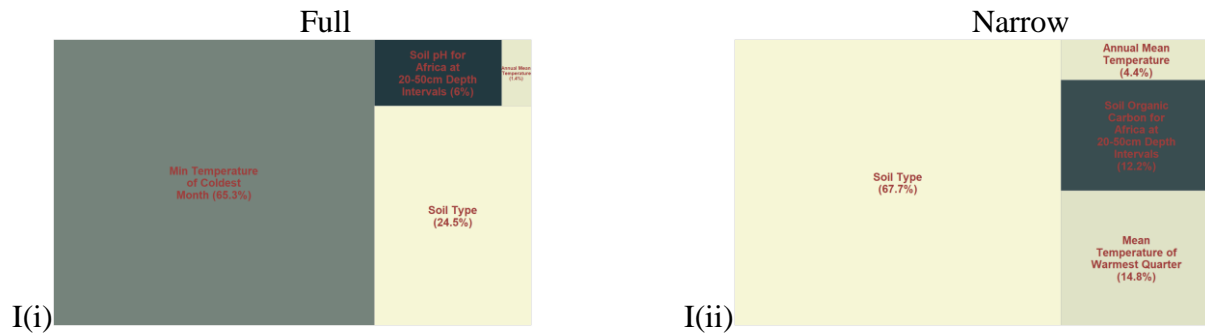


Figure 2.10: Tree plots of the relative contributions of major environmental variables (relative contribution greater than or equal to $\geq 1\%$) to the full (i) and the narrow (ii) Maxent models. (A) *E. africana*, (B) *E. coracana*, (C) *E. floccifolia*, (D) *E. indica*, (E) *E. intermedia*, (F) *E. jaegeri*, (G) *E. kigeziensis*, (H) *E. multiflora*, and (I) *E. tristachya*.

Discussion

Model predictions

This assessment is the first environmental distribution modeling specific to *Eleusine* species. Thus, it hopefully lays the foundation for more profound environmental and geographic studies and potentially inform agricultural and conservation planning. The broad choice of environmental predictor variables was to explicitly consider all variables relevant to species distribution. This approach is a recommended first step to identifying and eliminating ineffective variables and comprehensively selecting applicable environmental predictors based on high contribution level and expert knowledge for SDM of target species (Lin and Chiu, 2020). Environmental distributions were modeled with Maximum Entropy software. Maxent is a species distribution modeling software suitable for presence-only data used in this analysis (Elith *et al.*, 2010). One advantage of using a maximum modeling algorithm over more straightforward statistical tools, such as logistic regression, is that it reduces the impact of interactions that could occur among these variables (Phillips *et al.*, 2006). This approach has allowed new insights into how climatic and soil variables may have influenced the distribution of *Eleusine* in Africa.

Full Africa extent maxent models

Binary maps were generated for the full extent models in this study with the logistic threshold cutoff values from maximum training sensitivity plus specificity. These values, which vary for each species and model, are based on Maxent's probability of prediction of occurrences. The maximum sum of sensitivity and specificity has been considered consistent in producing results and one of the best threshold selection methods with presence-only data when random points are used instead of genuine absences (Liu et al., 2005; Liu et al., 2016).

The binary maps for the full extent models reflect the limited distributions of *Eleusine* species described by Phillips (1972). Species with abundant records, such as the wild occurring *E. africana*, *E. indica*, and the cultivated *E. coracana*, have somewhat unique distributions that do not overlap. The probability of occurrence for *E. africana* covers a large extent in eastern and southern Africa, with a continuous patch in the northern parts of West Africa. An unusual prediction for *E. africana* is that the full Africa model also suggests north Africa as a highly suitable environment. This northerly Africa occurrence of *E. africana* has never been reported. *E. indica*, in contrast to *E. africana*, is highly probable in west Africa, with patches of occurrence along the edges of east and southeast Africa and in Madagascar. Even though *E. indica* is widely known as a tropical and subtropical weed (Phillips, 1972, 1995; Liu and Peterson, 2010); Maxent model shows a low probability of occurrence in many places in tropical Africa. It is interesting that although *E. coracana* is reported as widely cultivated from the west to the east in Africa (Phillips, 1972; Liu et al., 2011), the observed distribution predictions in this study reflects a mainly eastern Africa cultivation with an isolated occurrence in West Africa (Nigeria and Senegal). This prediction is somewhat consistent with the description of *E. coracana* as evolved and adapted to east Africa (Liu and Peterson, 2010; Phillips, 1972; Liu et al., 2011). The observed pattern of occurrence could be due to cultivation and consumption preferences. Nevertheless, this

result may also suggest that west Africa has a limited suitable environment for cultivating finger millet and that the environment preferences of *E. coracana* may be different from other adaptable millets like *Digitaria*—known to west Africa (National Research Council, 1996). The probabilities of occurrence of the constrained *E. intermedia*, *E. floccifolia*, *E. jaegeri*, and *E. kigeziensis* are high around the locations where they have been reported in the east and southeast Africa. The distribution patterns show a unique patch for each species and largely reflect the described distributions by Phillips (1972). However, *E. multiflora* showed an extended environmental preference beyond Eritrea, Ethiopia, Kenya, and Tanzania to neighboring Uganda in the east and South Africa, Namibia, and Zimbabwe in the south. The eastern occurrence suggests that it could be present or, at least, could thrive outside its presently known range. Similarly, the probability of occurrence of *E. tristachya* (a south American endemic and only widely reported in South Africa) is high in North Africa, and it indicates that the species could, possibly, thrive in north Africa, particularly closer to the coasts.

Performance measures of full Africa extent models

The predictive performances of the potential distributions of *Eleusine* species modeled by Maxent with the full Africa map show increasing improvements with an increasing number of location records. The plots of omission (training and testing) against the cumulative threshold were increasingly closer to the prediction versus cumulative threshold graphs (AUC graphs). Generally, the receiver operating curves (ROC) show values greater than 0.8 for all species indicating a high fit of the model to the data (predictive power) (Phillips, 2017). Like the AUC plots, the ROC curves improved with increasing sample size. Maxent has been cited as robust in modeling distribution for species with small occurrence datasets (generally less than 100 locations) (Hernandez et al. 2006; Papes and Gaubert 2007; Phillips et al. 2006). However, as shown in the model performance

plots, Bean et al. (2012) reported that prediction accuracy is affected by small sample sizes. Therefore, inferences should be made with caution when dealing with small sample sizes.

Narrow extent Maxent distribution models

The predicted species distribution pattern generated by Maxent with the narrow extents are generally similar to corresponding distribution maps in the full Africa extent models. The species distribution maps generated are only for the countries where presence records exist and exclude countries where species have not been reported. Each species prediction shows a slimmer geographical distribution pattern than the corresponding full extent model. The advantage of limiting background is that it increases the likelihood of a more ecologically realistic distribution because it reduces artifacts of prediction statistics when modeling with Maxent (Elith *et al.*, 2010; Phillips, S. J., 2017).

Performance measures of narrow Africa extent models

Although the narrow extent models are expected to be more realistic due to their more limited background than the full extent models (Elith *et al.*, 2010), their predictive performances modeled by Maxent are mainly like the full extent models. They show increased improvements with an increasing number of location records. The plots of omission (training and testing) against the cumulative threshold were close to the plot of prediction versus cumulative threshold (AUC graphs) in *E. africana*, *E. coracana*, and *E. indica* with many occurrence records. The receiver operating curves (ROC) values were also greater than 0.8 for all species and indicated a high fit of the model to the data (predictive power) (Phillips, 2017).

Narrow extent predictions

One other advantage of limiting background used in the narrow extent model is that it was possible to contrast reported areas with and unoccupied environments and make predictions of the likely

distribution of *Eleusine* in areas where they have never been reported. In this analysis, Maxent projections were broadly consistent with the distribution models of the full extent models. For example, the northerly occurrence predicted for *E. africana* in the full extent model was affirmed by the narrow extent projections providing more support for the environmental suitability of the region for *E. africana*. Additionally, *E. coracana* shows a low probability of occurrence or suitability in west Africa, albeit with new areas that may support cultivation identified along the coast. Projecting the narrow model extent of *E. tristachya* to a complete map of Africa shows a much different contrast to the full Africa model. The high northerly probability predicted in the full Africa model is mainly absent in the narrow model projections. It suggests that the high probability shown in the full extent model could be due to artifacts from the more prominent environmental background (Elith *et al.*, 2010; Phillips, S. J., 2017). It is, however, interesting to note the new suitable environments—in the east and southeast Africa—identified from the narrow extent projections, especially for the highly constrained *E. intermedia*, *E. floccifolia*, *E. jaegeri*, and *E. kigeziensis*. Identifying novel suitable environments particularly emphasizes the usefulness of species modeling to the management and conservation concerns of endemic *Eleusine* species (Elith *et al.*, 2010; Phillips, S. J., 2017).

Analysis of variable contributions

The relative contributions of environmental variables are considerably similar in the full and narrow extent modeling approaches. The observed variations in environmental predictor contributions may result from artifacts from the more extensive background in the full extent model. These similarities in the composition of contributing environmental factors could be due to the robustness of Maxent modeling software in modeling species distribution prediction with

presence-only data (Hernandez *et al.* 2006; Papes and Gaubert 2007; Elith *et al.*, 2010; Phillips, 2017).

Biologically relevant aspects of temperature and precipitation are the most consistent predictor variable in the two models for *Eleusine* species. This reflects the theory that plant species distribution is mainly associated with water availability, especially at latitudes closer to the equator, where the sun's radiant energy is abundant (Hawkins *et al.*, 2003). The observed high contribution is also consistent with reported correlation between climate and grass distribution (Hartley, 1950). The occurrence of elevation as a high (> 10%) contributing factor for many of the *Eleusine* species in Africa is in tandem with their known occurrence and cultivation at high latitude. West Africa generally has a low altitude, and the high contribution of elevation to the distribution modeling of *E. coracana* may help explain its limited cultivation in the region.

Conclusion and future recommendations

A larger sample size will be required to better comprehend *Eleusine* species' distribution. My analyses included all possible geocoordinates on GBIF for *Eleusine* in Africa, but the model statistics were poor for species with low data points. It is fair to highlight that the precision of geocoordinates used in this study was inconsistent. Some values were given precise to the three-hundredths degree, and others seemed collected to one-ninth degree precision and then approximated. Thus, field validation is an essentially critical next step in validating the results of these models (Rebelo & Jones 2010).

Furthermore, field validation should include factors that were not available to this study and which could be ecologically meaningful to the adaptation and distribution of *Eleusine* species. These include biotic interactions, disturbance, and topography/land use data (Mod *et al.*, 2016). These concerns show the need to collaborate with known locality records (e.g., herbarium records)

to carry field verifications. It is also essential to do carefully repeated sampling to determine that the target species is genuinely present or absent from a locality in building strong distribution models. Sufficient repeated field observations would help adopt a distribution model that accounts for imperfect detections of large-scale analysis. This is invaluable for identifying new populations, defining environmental characteristics, and is helpful for habitat restoration and conservation efforts of wild *Eleusine* species.

Chapter 3: Structural Variation Analysis of *Eleusine coracana* whole-genome sequences

Introduction

The uniqueness and similarities of plant species' habitat, growth, and reproduction can be traced to their genomes. Unlocking the information in the structure, organization, evolution, and function of plant genomes will advance our understanding of plant biology and help crop improvement. Genomics can help us find correlations between genomic variations and observed traits (Edwards and Batley, 2004).

Next-generation DNA sequencing (NGS) technology with reduced cost has increased the quality and diversity of publicly available plant genomic resources and since completing the primary genomic sequence of *Arabidopsis thaliana*. The availability of high-quality data has facilitated the development of tools for analyzing genomic data and the integration of information from the field of omics. Genome analyses include identifying genes and gene products and elucidating functional relationships between genotype and phenotype using whole-genome sequencing and re-sequencing data (Edwards and Batley, 2004; Li *et al.*, 2009). Single Nucleotide Polymorphisms (SNPs) (or variants) studies have dominated NGS plant genetic variants identification in genetic mapping and genome-wide association. However, recent studies show that SNPs do not capture large genomic variations that equally contribute to phenotypic differences (Saxena *et al.*, 2014; Francia *et al.*, 2015).

Genomic structural variants (SV) are large sequence differences in a genome relative to a reference genome. SVs could be a loss (deletion) or gain (duplication) in copy number, a change in orientation (inversion), or chromosomal location (translocation) of a sequence (Medvedev *et al.* 2009; Escaramís *et al.* 2015). These changes can lead to loss or variation in gene dosage. SV analyses in humans show that structural variants account for more variations in base pairs than

SNPs (Alkan *et al.*, 2011; Baker 2012; Sudmant *et al.* 2015). SVs are large and possibly altering gene structure, dosage, or location (Layer *et al.*, 2014). Variation in a gene copy number has been called copy-number variation (CNV) and missing regions in some individuals relative to others, called copy-number variation (CNV) (Schiessl *et al.*, 2018).

Several studies have also interrogated the association between structural variants and plant phenotypes. They reveal that SVs overlap and enrich abiotic stress response genes, protein-coding genes, and disease resistance genes in soybean, rice, potato, and Arabidopsis (Cook *et al.*, 2012; Fuentes *et al.*, 2019; Kyriakidou *et al.*, 2019; Zmienko *et al.*, 2020). SVs have also been linked to boron tolerance in barley (Sutton *et al.*, 2007). Furthermore, Li *et al.* (2016) found variations of PAVs informative for assessing patterns of genetic diversity in Glycine spp.

Plant genomes contain many repetitive regions, and many plants have multiple ploidy (multiple copies of entire chromosomes) levels (diploid tetraploids, hexaploids, and others). Ploidies are from spontaneous genome duplication (autopolyploidy) or hybridization of chromosomes from different species (allopolyploidy). SVs can arise through these duplication events, with the eventual differential loss of duplicated genes (Iovene *et al.*, 2013).

Bioinformatics tools for identifying structural variants from high throughput sequencing short read data utilize one of the following approaches. The first method involves inferring from discordantly mapped paired-reads whose distances are significantly different from the predetermined average insert size in the paired-end mapping approach (or RP) (Sindi *et al.* 2009). Second, using the position and distance between fragments of a read independently aligned to the reference genome to determine structural variants in split-read mapping approach (or SR) (Schröder *et al.* 2014). Read depth approach (or RD) uses the correlation between sequencing depth coverage and the frequency of a genomic region (Abyzov *et al.*, 2011; Duitama *et al.* 2014;

Smith *et al.* 2015). Finally, the *de novo* assembly approach (or AS) reconstructs DNA fragments (contigs) from short reads and compares them to a reference genome to infer SVs (Rizk *et al.* 2014; Yang *et al.* 2015). No single method can detect the total genomic structural variations. However, the highest resolution studies of SVs can be achieved using a *de novo* assembly-based approach; this is computationally intensive for large individuals.

Due to the complexity of structural variants and their occurrence in repetitive regions, discovering structural variation (SV) from whole-genome sequencing data is better with a combination of approaches and prior knowledge. (Rausch *et al.* 2012; Layer *et al.* 2014; Mohiyuddin *et al.* 2015). *LUMPY* is an SV discovery framework that utilizes signals from read-pair, split-read, read-depth jointly. *LUMPY* yields improved sensitivity and performed well in calling SVs of diverse sizes, especially when a low coverage data signal is reduced owing (Layer *et al.*, 2014; Kosugi *et al.*, 2019).

Finger millet (*Eleusine coracana* L. Gaertn.) is a historical, nutritional crop, particularly in Asia and Africa. It is a self-fertilized allotetraploid ($2n = 4x = 36$) annual considered a hardy crop due to its wide adaptability. It is a drought and disease-tolerant crop and has been reported to have an extended shelf life (Parashuram *et al.*, 2011). However, unlike wheat and other popular grains, *E. coracana* has remained unpopular due to its coarse texture and intense seed coat color (Sood *et al.*, 2018). Recently, there has been an increased interest in adapting finger millet as an economically viable, super future crop, with studies to elucidate the genetic architecture and decipher the relationship between genotype and phenotype in finger millet.

Recently, there is an explosion in the number of high-quality whole-genome sequencing (WGS) data and transcriptomics data for finger millet accessions on the National Center for Biotechnology Information (NCBI). Furthermore, the recent availability of a draft genome

sequence (*E. coracana* genome v1.1 on Phytozome13, https://phytozome-next.jgi.doe.gov/info/Ecoracana_v1_1) makes it possible to analyze genomic variations in finger millet. Presently, over a hundred whole-genome sequence data of *E. coracana* are publicly available on NCBI. These are global collections from various finger millet accessions released by various Bioprojects. Some of these WGS data were created for SNPs analysis and genome building.

Here, I investigated genetic variations in the accessions by identifying structural variants in 116 WGS from NCBI with *LUMPY*. I determined the distribution and functional genomic impact of SV regions by analyzing genes overlapping with SVs. Identifying and understanding the distribution of SVs in *E. coracana* could assist researchers in the identification of novel resistance genes and improve current breeding efforts.

Materials and Methods

Structural variants and their genomic distribution in *E. coracana* were analyzed by downloading publicly available, paired-end, whole-genome sequences from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) repository. The accessions were analyzed with custom bash and R scripts created based on freely available bioinformatics tools (Fig. 3.1).

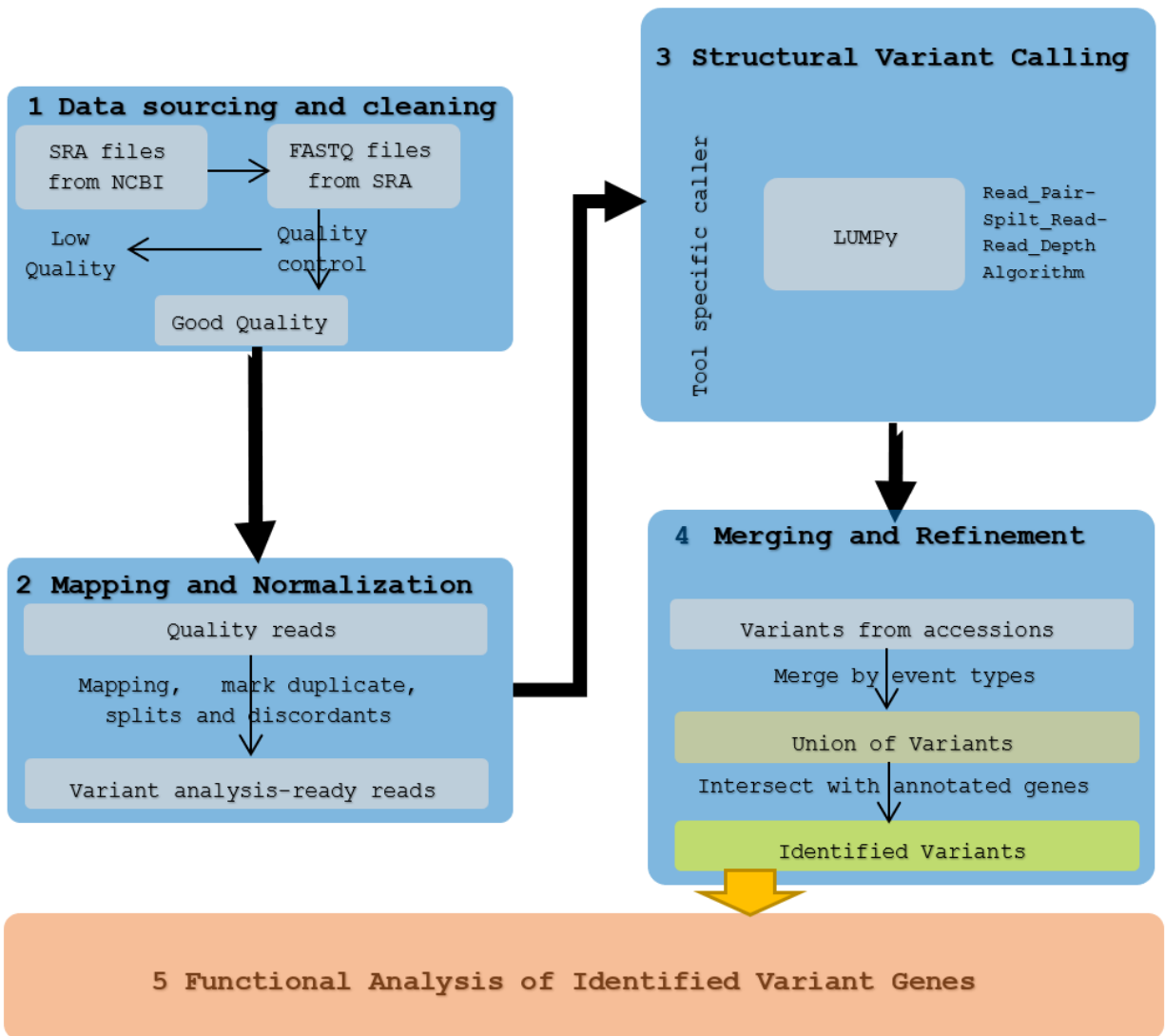


Figure 3.1: Summary chart of bioinformatics pipeline for structural variants identification analysis in *Eleusine coracana*

Step 1: Data Sourcing and Cleaning

Raw data archive files from NCBI

To identify *E. coracana* genomes for structural variant analysis, I searched for publicly available paired-end whole-genome sequence on NCBI (accessed June 24, 2021) using the scientific name of the species — ‘*Eleusine coracana*’ as the search term under the Taxonomy section. The SRA sequence link from the results table showed 347 sequences filtered with the

following parameters: Source-DNA; Type-genome; Library-paired and Strategy-genome. After manually removing plastid genomes from the search result, the SRA table was downloaded from the SRARunSelector. This table contained 116 samples of Illumina reads from 5 NCBI BioProject databases (<http://www.ncbi.nlm.nih.gov/bioproject>) under the accession numbers PRJNA383952, PRJDB5606, PRJNA338521, PRJNA377606, and PRJNA610152 (Table 3.1). The SRA Rutable was used to download sequences to the Alabama Supercomputer for analysis.

Table 3.1: BioProject accession numbers and numbers of SRAs per each downloaded for analysis from NCBI database

SN	BioProject Accession Number	Number of SRA in BioProject	Data Source
1	PRJDB5606	9	Beijing Genomics Institute, China
2	PRJNA338521	6	University of Agricultural Sciences, India
3	PRJNA377606	11	University of Zurich, Zurich
4	PRJNA610152	88	The University of Trans-Disciplinary Health Sciences and Technology, India
5	PRJNA383952	2	University of Agricultural Sciences, India

FASTQ files from Raw Reads

To obtain *FASTQ* files in identified SRA, I created a custom bash script. SRA files are compressed files suitable for archiving sequences. In the script, the list of SRA files to download was prepared by pulling them from the first column of the SraRunTable. After that, the SRA were downloaded files from NCBI database using the prefetch command of *sra2.10.9* toolkit (<https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software>) with the --option-file option set to the list of SRA list, and the --max-size option to 50G (due to the large size of some of the SRA files). The SRA files were discarded after extracting the *fastq* files with the fasterq-dump command in *sra2.10.9*. The total size of the SRA files was 4.4terabytes.

Data Filtering and Quality Analysis

To obtain high quality data for my analysis, I performed quality assessment, trimming, and QC result aggregation of downloaded sequences in a sequence of custom scripts. In the scripts, downloaded sequences were first assessed for quality with *FastQC*v.0.11.9 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc>), and thereafter trimmed accordingly with Trimmomatic v.0.39 (Bolger et al., 2014), after which the reads were reevaluated for quality. In trimming, Illumina adapter sequences (in a custom list of adapters) and the leading and trailing sequences with a base quality of less than 20 were taken out. Reads with less than 40 bases and reads with local drops in average base quality less than 22 measured within a six-base sliding window were also removed. Quality assessment results were aggregated by BioProjects using *MultiQC*v1.7 (Ewels *et al.*, 2016). All 116 accessions had more than 10 million reads following the quality-based filtering; therefore, none was excluded from further analysis.

Step 2: Read Mapping and Normalization Procedures

Read mapping and duplicate tagging

To realign reads in FASTQ files to the respective regions they likely originated from, I downloaded the recently published *E. coracana* reference genome assembly v1.1 (Devos *et al.*, 2021) from Phytozome and mapped the trimmed reads to it using the *Burrows-Wheeler Aligner* (BWA) in 3 steps. First, the reference genome was indexed using `bwa index` command and default parameters. The trimmed forward and reverse reads were then mapped to the indexed genome with the faster and more accurate `bwa-mem` algorithm v.07.12 (Li and Durbin, 2010) using default parameters in the second step. In post-alignment, aligned outputs were sorted with *SAMTools*v.1.11 (Li et al., 2009). *samblasterv*.0.1.24 (Faust and Hall, 2014) was used to remove duplicates (`--excludeDups` option) and to tag discordant and split reads (`--addMateTags` and -

maxSplitCount options) to reduce variant analysis complexity and runtime. After this, the SAM files were compressed to BAM files using samtools with default parameters.

Excluding high coverage regions

To improve the quality of SV calls and reduce false positives from high coverage regions, I identified and excluded regions with very high coverage were using two custom python scripts by Ryan Layer (<https://github.com/arq5x/lumpy-sv>) for *LUMPy* structural variant analysis. First, the `get_coverages.py` (modified) script was used to find the min, max, mean, and standard deviation coverages of the split reads and paired-end bam files and create coverage profiles for the bam files. I chose to exclude regions that have more than five times the standard deviation coverage from the output. The `get_exclude_regions.py` (modified) script was used to create the `exclude.bed` files. `get_coverages.py` and `get_exclude_regions.py` were called bam files from a custom bash script.

Step 3: Structural Variant Discovery Pipeline with *LUMPy* Express

To detect structural variants in the `bwa-mem` aligned 116 WGS samples, I called the `lumpyexpress` module of *LUMPy*v.0.3.1 (Layer et al., 2014) on the samples independently in a custom bash script. In calling `lumpyexpress`, I used the defaults parameters for a single sample with pre-extracted splitters and discordant. *LUMPy* is a probabilistic framework for structural variant discovery based on read-depth, read-pair, and split-read density. High coverage regions identified in the previous step were provided to the software with the `-x` option to reduce artifacts. *LUMPy* produced a *VCF* 4.2 specification file with a raw catalog of 4 structural variant events (deletion—DEL, duplication—DUP, inversion—INV, and breakpoints—BND) filtered for precision and type of event in downstream analysis.

Step 4: Structural Variant Merging, Refinement and Filtering

To obtain high quality calls, I parsed the structural variant files with a custom script as follows. First, precise variant calls (high confidence calls) were separated from imprecise calls for each *VCF* file generated per sample. Next, bed files with chromosome, start position, stop position, and the number of support (read depth and split reads) and length were created for each precise variant calls separated by events into individual files. A union of all structural variants in all samples was created per event by combining the individual event bed files of the samples. LUMPY identified many overlapping structural variants. The variants were merged using *BEDtools.2.26.0* (Quinlan and Hall, 2010) merge command and sorted by chromosome number and start position to remove redundancy.

Distribution of SVs relative to gene models

The overlap of identified structural variants (by events and samples) with the intergenic regions in *E.coracana* was carried out using the *BEDtools* intercept command. Structural variant events were intersected with the genic regions in the Eleusine genome v1.1 annotation file (gff3) (Devos *et al.*, 2021). The gene names of intersected regions were pulled from the description text file provided with the genome. Graphical representations of the distribution and genes of the structural variant in the genome were prepared with *IGVv.2.9.4* (Robinson *et al.*, 2011).

Graphical charts of different analyses statistics

To visualize and to understand variation or show relationships between variables, graphical representation of the structural variant results, the number of events discovered per sample, the boxplot of the number of supports for the calls, and the number of genes overlapping each event type were created in *R v4.11* (R Core Team, 2021).

Step 5: Annotation and Analysis of Structural Variation Genes

To understand the potential impact of identified structural variants on genes, the *TOPGO* package (Alexa and Rahnenfuhrer, 2021) of *R* software was used for the functional analysis of trait-related genes. The analysis pipeline, written in a custom *R* script, required the *GO.db*, *biomaRt*, and *Rgraphviz* libraries. The GO enrichment analyses for Biological Process, Molecular Function, and Cellular Component of deleted, duplicated, and inverted genes were carried out with the GO:IDs using the `annFUN.gene2GO` function. The gene annotation list of the GO:IDs were created by retrieving them from Phytozome 13 (Goodstein et al., 2012) BioMarts (database). The pipeline tested for significance between genes that overlap structural variants and the total genes in the *E. coracana* genome v1.1 (Devos et al., 2021) using the `weight01` algorithm with `fisher` (default). The GO annotation analysis results were saved to file, and the hierarchical plots of enriched GO terms were plotted. The `geom_bar` function of `ggplot2` was used to generate the histograms of the top enriched terms.

Results

Detection of structural variations using whole-genome re-sequencing data

One hundred and sixteen whole genome sequences (WGSs), downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) repository under five BioProjects, were mapped to the newly published *E. coracana* reference genome v1.1 on Phytozome 13 to detect structural variations. Available metadata indicated that sequences were generated from total DNA extract of young leaves of 93 different *Eleusine* accessions (grown in greenhouse conditions) with Illumina sequencer.

Data Filtering and Quality Analysis

Table 3.2 provides a summary of sequence quality (by BioProjects) before and after trimming. Briefly, about 5 to 45 % of reads were removed from the sequences. BioProject PRJNA610152 had the highest quality sequences, and only 5% of the reads were removed. None of the downloaded WGS was excluded from the analysis after trimming.

Table 3.2: Bioprojects, number, number of reads and quality of reads of *E. coracana* sequences downloaded from NCBI before and after trimming. Quality is grouped by Bioprojects with MultiQC

SN	Bioproject	Number of SRA	Data Source	Number of reads before trim (in millions)	Read length before trimming	Number of reads after trim (in millions)	Read length after trimming	Percentage after trimming
1	PRJDB5606	9	Beijing Genomics Institute, China	50 – 210	100/150	50-190	92-142	~90
2	PRJNA338521	6	University of Agricultural Sciences, India	12-97	60-150	7-78	60-150	~55
3	PRJNA377606	11	University of Zurich, Zurich	10-190	81-300	3-163	84-225	~63
4	PRJNA610152	88	The University of Trans-Disciplinary Health Sciences and Technology, India	25-70	100/125	18-63	97-123	~95
5	PRJNA383952	2	University of Agricultural Sciences, India	2	250	2	225	90

Number of structural variants detected

This SV analysis pipeline used *LUMPY* (Layer et al., 2014), a software that calls structural variants based on three read signatures (read pair, read depth, and split read). *LUMPY* detected between 0 and 32,176 in each accession (Fig. 3.2). Most of the events detected were breakpoint events which were not further analyzed in this study due to their complexities. Inversion events were the least reported structural variant type in *E. coracana*. Imprecise calls were also removed from further analyses to improve the accuracy of the SVs. Many of the detected calls were found

to be overlapping and thus merged into one. The number of events recognized in each accession was reduced based on the quality control, merging overlapping events, and exclusion of complex BND events (Fig. 3.3). The size distributions of these SVs and distribution of the number of supports are shown in Figures 3.4 and 3.5, respectively.

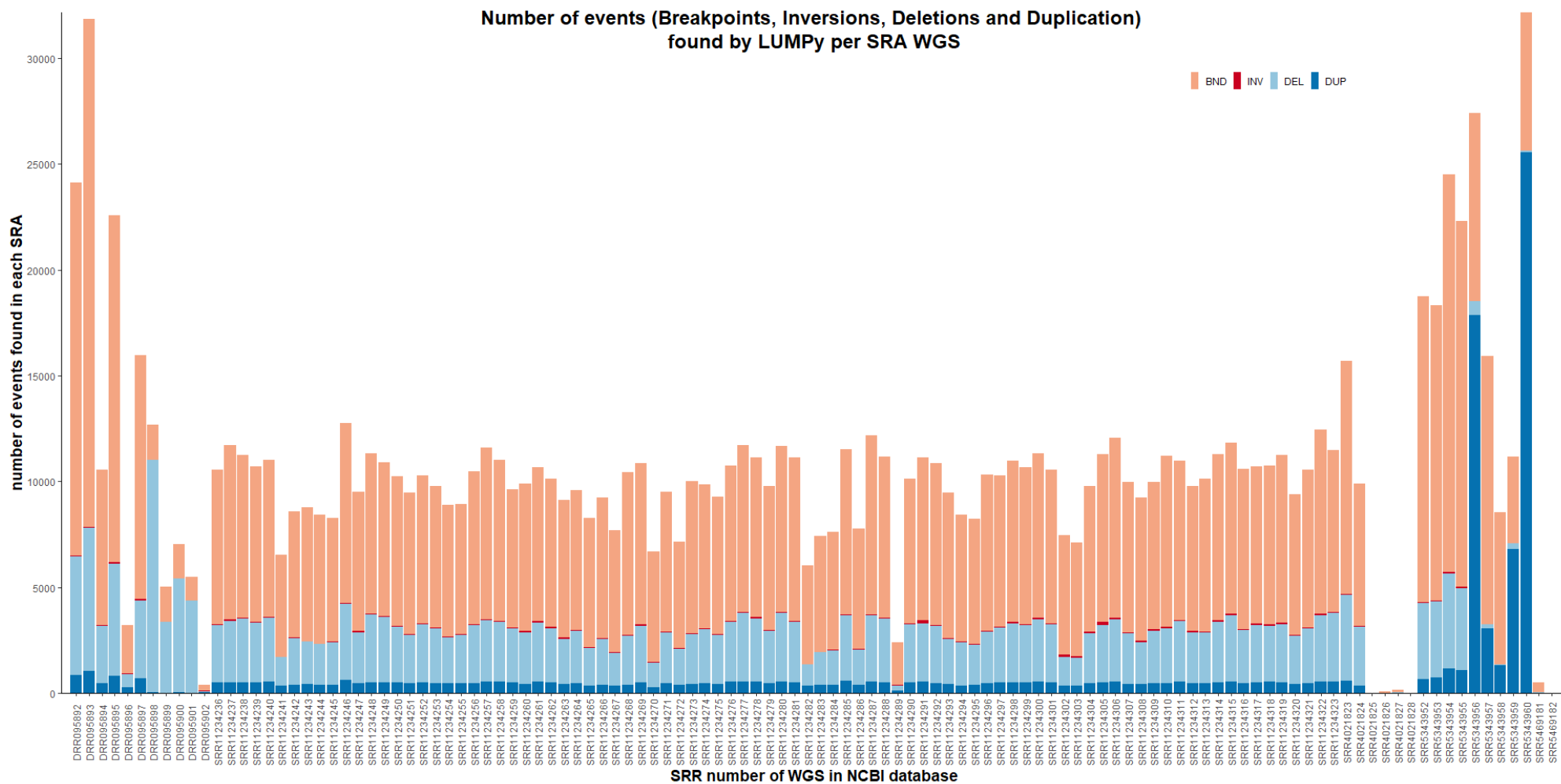


Figure 3.2: Number of structural variant events found for each whole-genome sequence downloaded from NCBI database when compared to the *E.coracana* v1 genome.

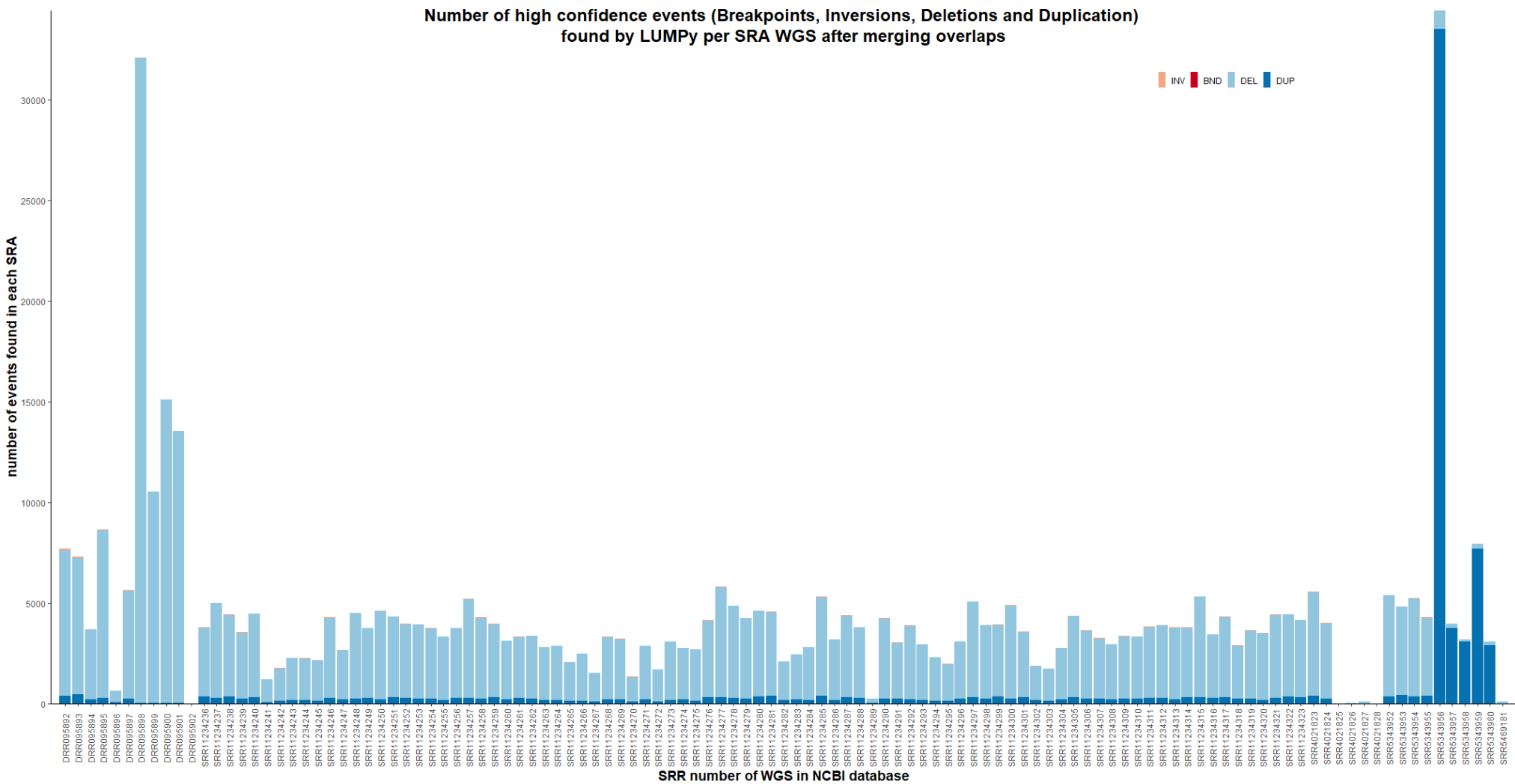
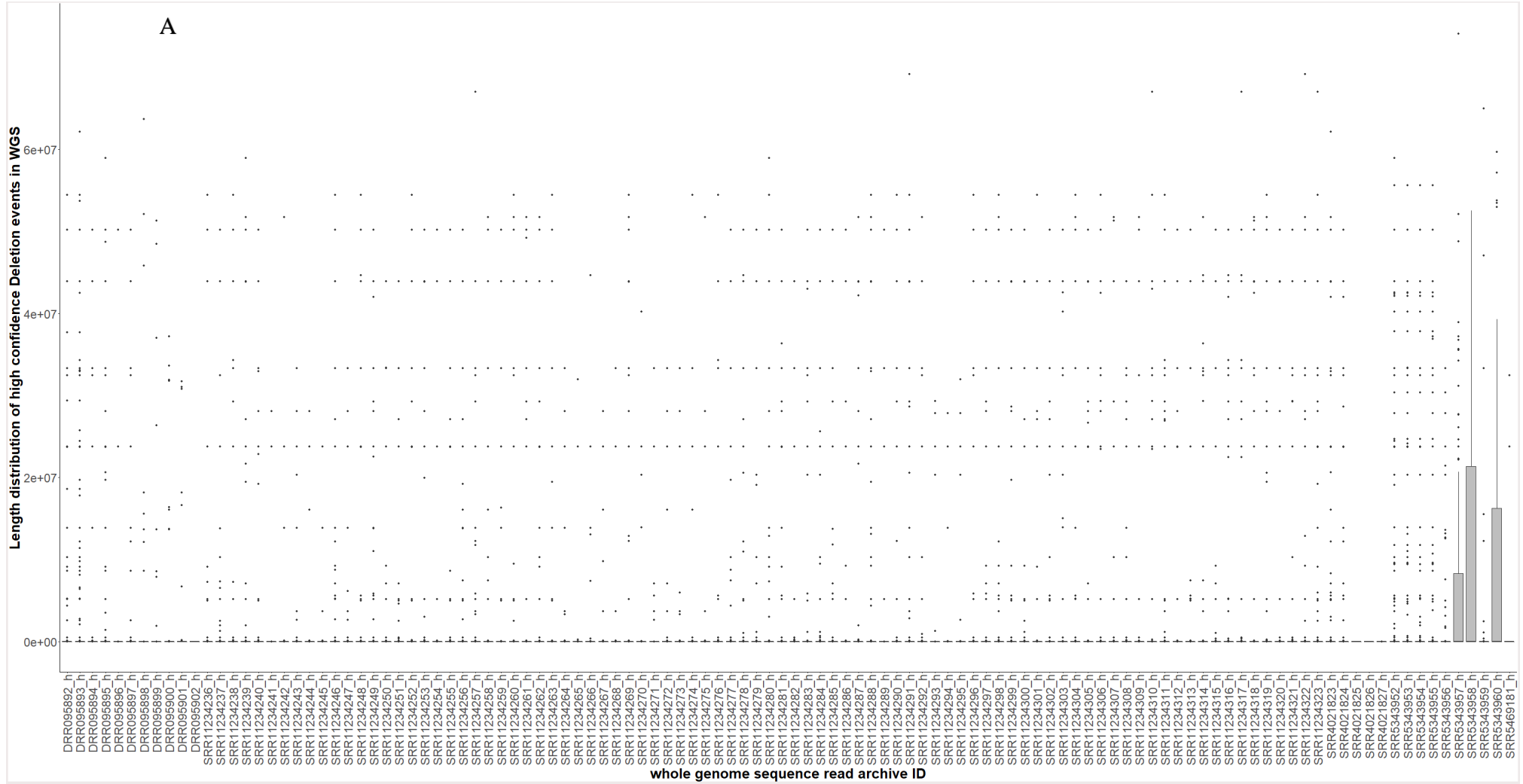
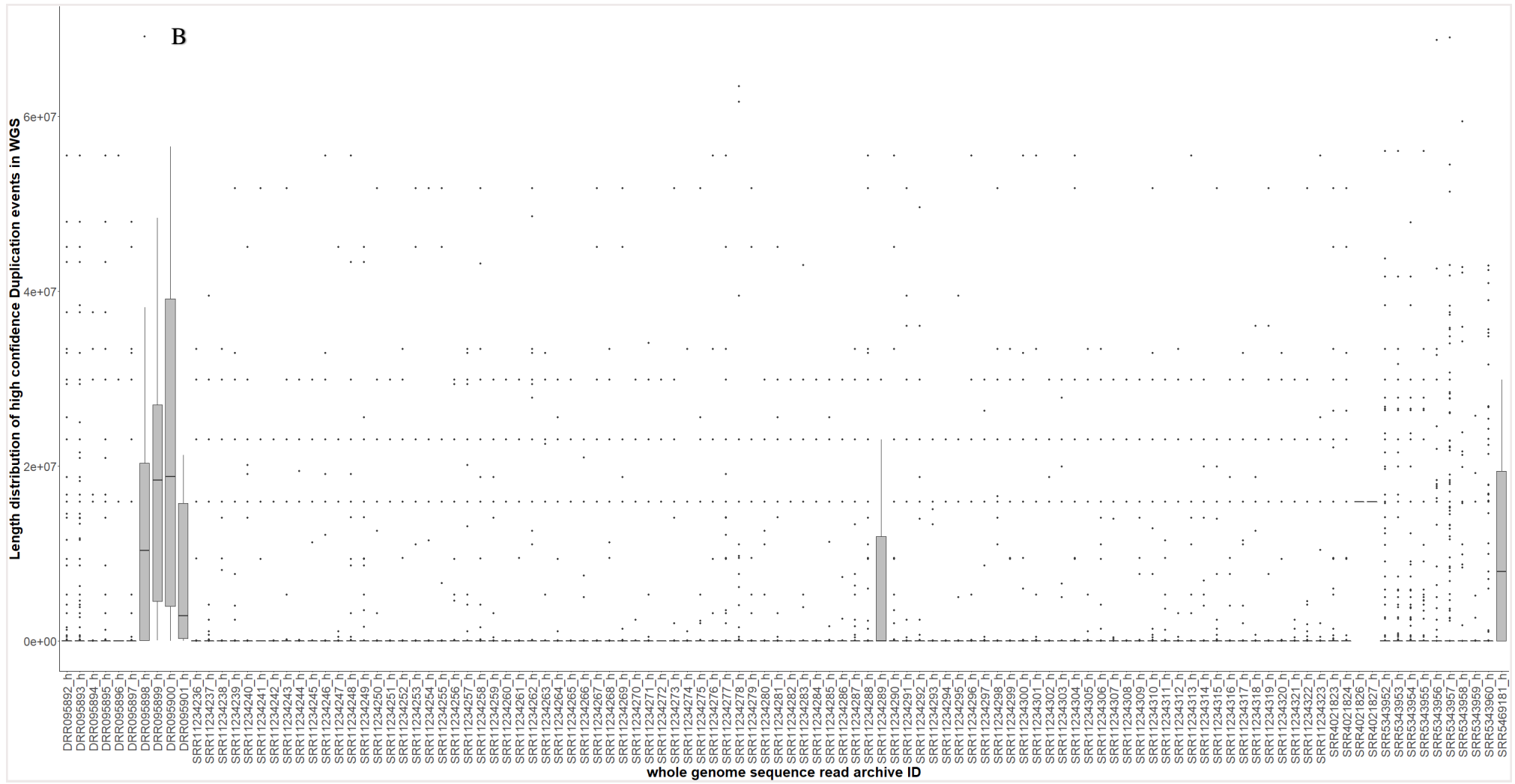


Figure 3.3: Number of high confidence structural variant events found for each whole-genome sequence after merging complete overlaps.





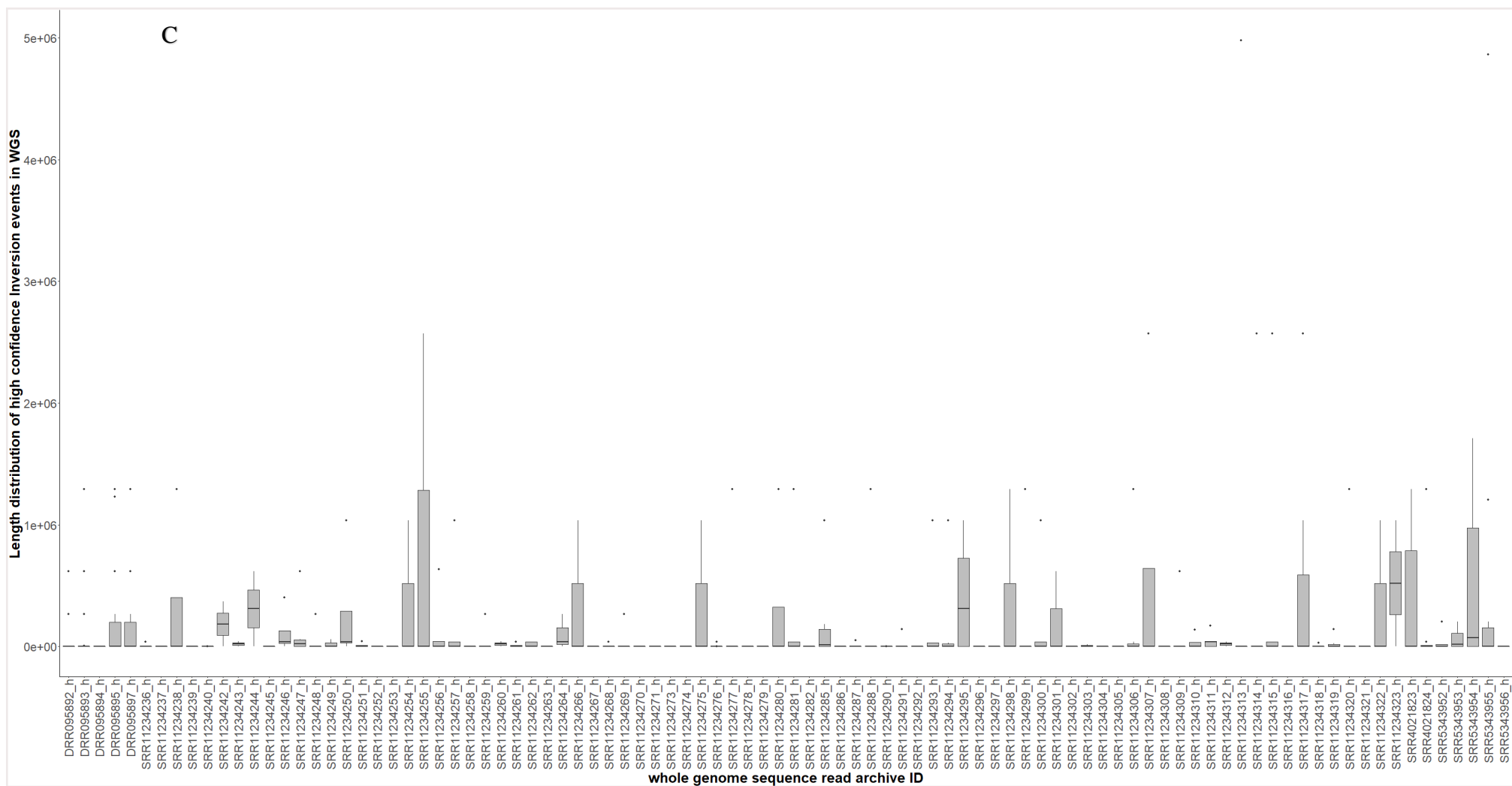
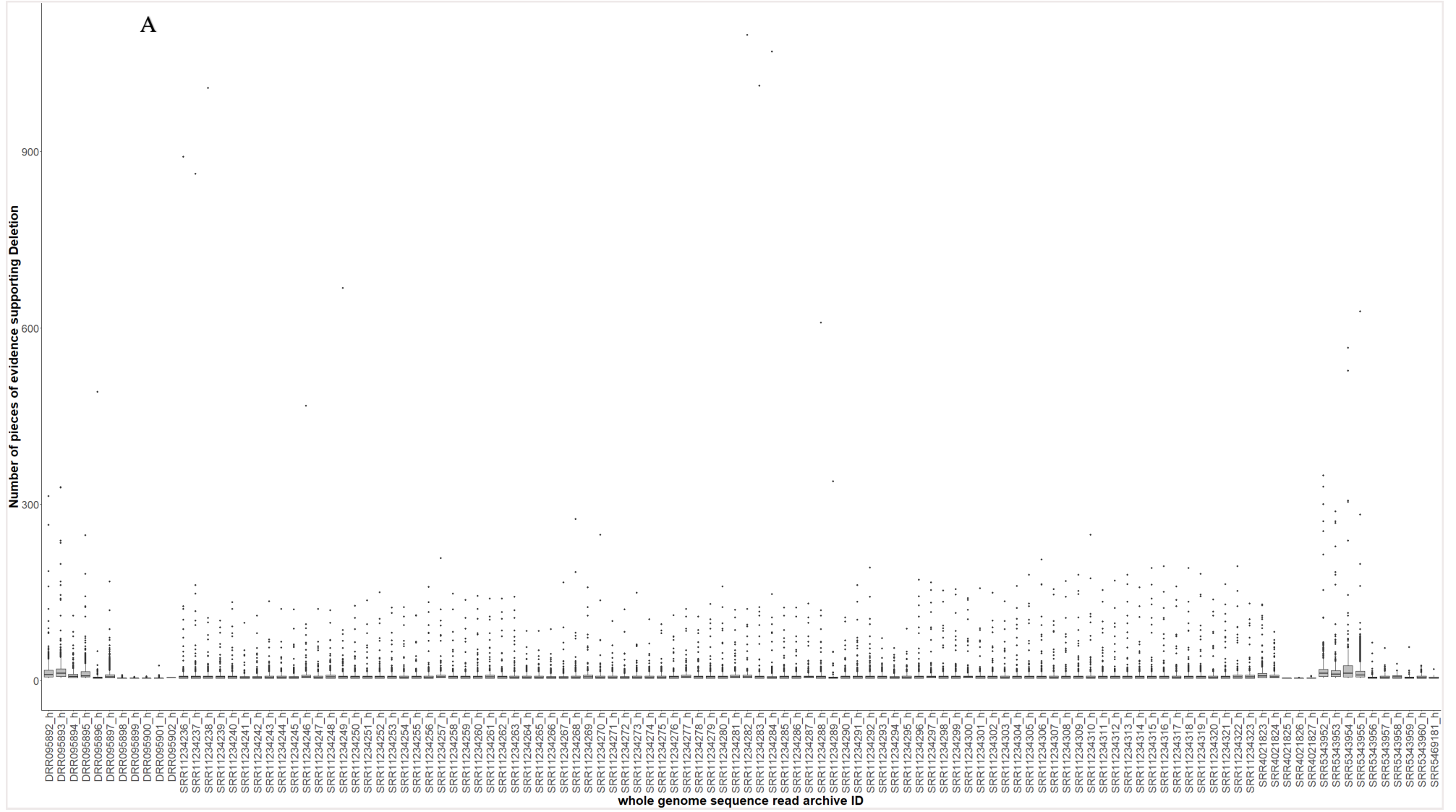
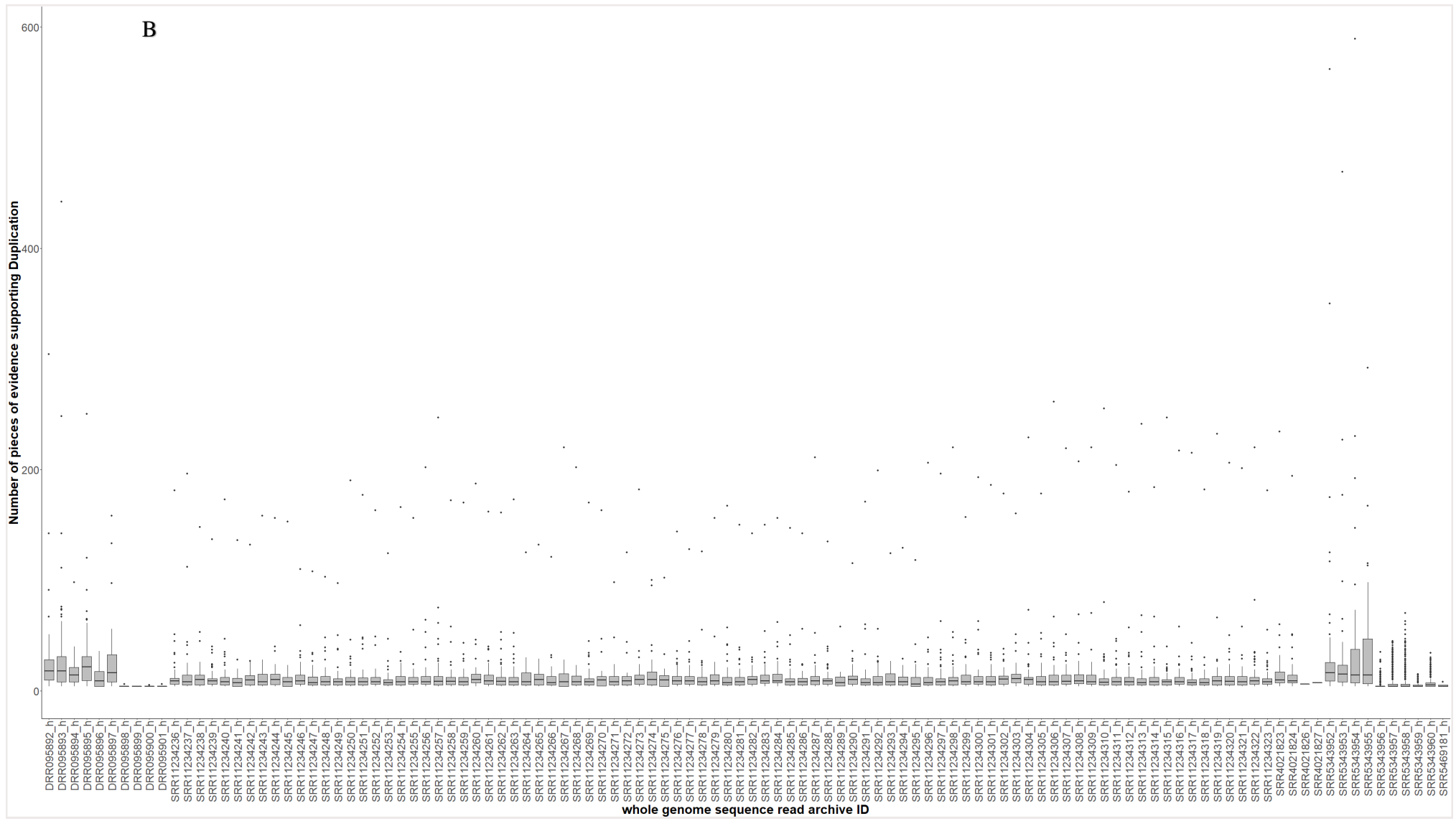


Figure 3.4: Boxplot showing the size distribution of the length of structural variation events (Deletion (A), Duplication (B), and Inversion(C)) found for each whole-genome sequence downloaded from the NCBI database.





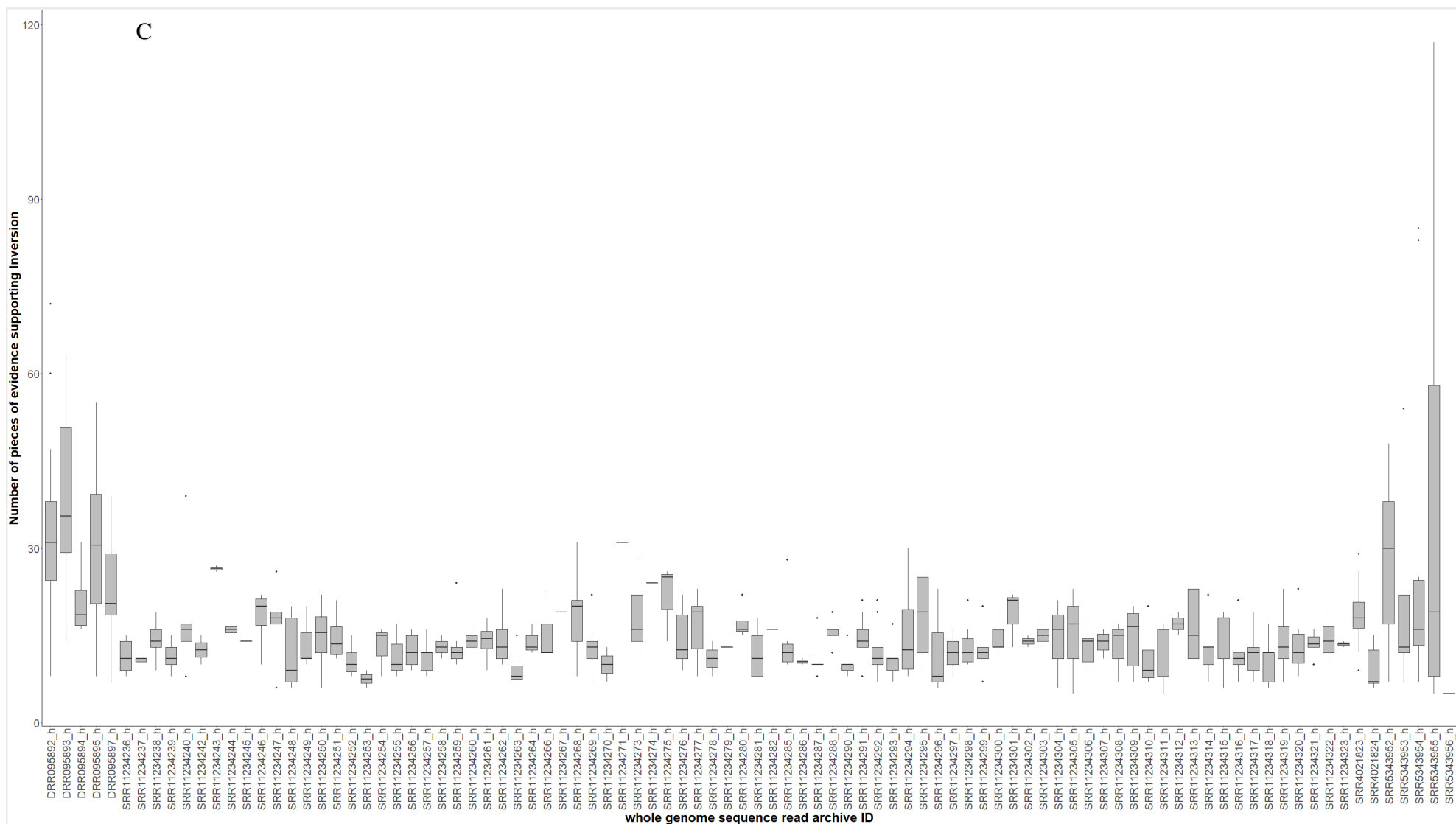


Figure 3.5: Boxplot showing the number of evidence (number of reads, number of split reads) supporting Deletion (A), Duplication (B), and Inversion(C) events found for each whole-genome sequence downloaded from the NCBI database.

Identification of Deletion, Duplication, and Inversion Events

Structural variation events (deletion, duplication, and inversion) present in all *E. coracana* were identified by combining all high confidence variants discovered by *LUMP*y in all accessions, followed by merging regions that overlapped to reduce redundancy. This procedure resulted in 93 inversions, 1,922 duplications, and 3,344 deletions. The distribution of these structural variants in the newly drafted genome as viewed in *Integrative Genomics Viewer* (IGV) (Robinson et al. 2017) is presented in Figure 3.6.

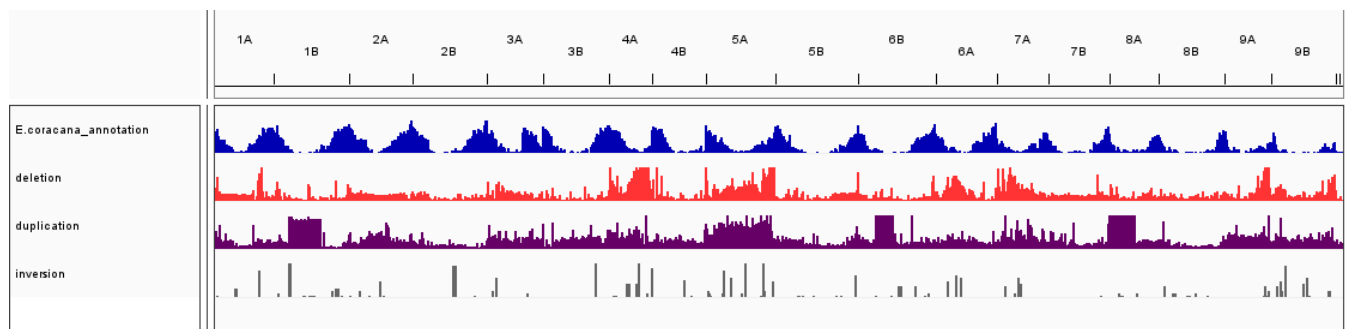


Figure 3.6: Genomic Distribution of Structural Variants (deletion, duplication, and inversion) in the *Eleusine coracana* Genome viewed in IGV.

The interplay between the Structural Variation and Genes

There are 48,883 identified genic regions in the recently published *E. coracana* genome v1.1. Intersecting identified structural variants with genic regions showed that 41,238 and 40,747 genes overlapped deletion and duplication events, respectively. Inversion events have low genic overlap in finger millet—324 genic regions (Fig. 3.7).

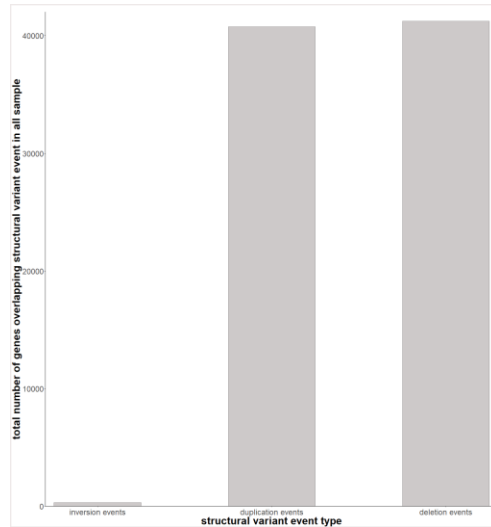


Figure 3.7: Number of genes overlapping identified structural variants in *Eleusine coracana* genome (inversion-324, deletion-41,238 and duplication-40,747, total number of genes in genome-48,883).

Functional impact of SV-overlapped genes

The assessment of the functional impact of discovered structural variations in Gene Ontology (GO) annotation indicated possible effects of variants on overlapped genes. A summary of the functional impact is shown in Table 3.3, categorized into all three primary genes GO categories—molecular function (MF), biological process (BP), and cellular component (CC). A complete list of GO:IDs affected is available in Supplementary Table. The deletion and duplication affected biological processes included several metabolic, biosynthetic, and transport processes (Fig. 3.8 and 3.9). Duplication events also affect biotic stress response and phosphorylation processes. Molecular functions affected by deletion and duplication events include ATP binding, protein tyrosine kinase, and transporter activity. The impact of inversion events was low for biological processes and molecular functions (Fig 3.10). All three events showed little influence on cellular components (Fig. 3.8, 3.9 and 3.10).

Table 3.3: Summary of significant ($p \leq 0.05$) GO:Process functional annotation of genes overlapping identified structural variations in *Eleusine coracana* under Biological Process (BP), Molecular Function (MF), and Cellular Component (CC)

GO:Process	SV Type	Number of Significant Processes	Number of Significant Genes
BP	Deletion	18	19013
BP	Duplication	15	8495
BP	Inversion	11	22
MF	Deletion	17	10102
MF	Duplication	20	8573
MF	Inversion	17	32
CC	Deletion	3	3442
CC	Duplication	1	36
CC	Inversion	2	2

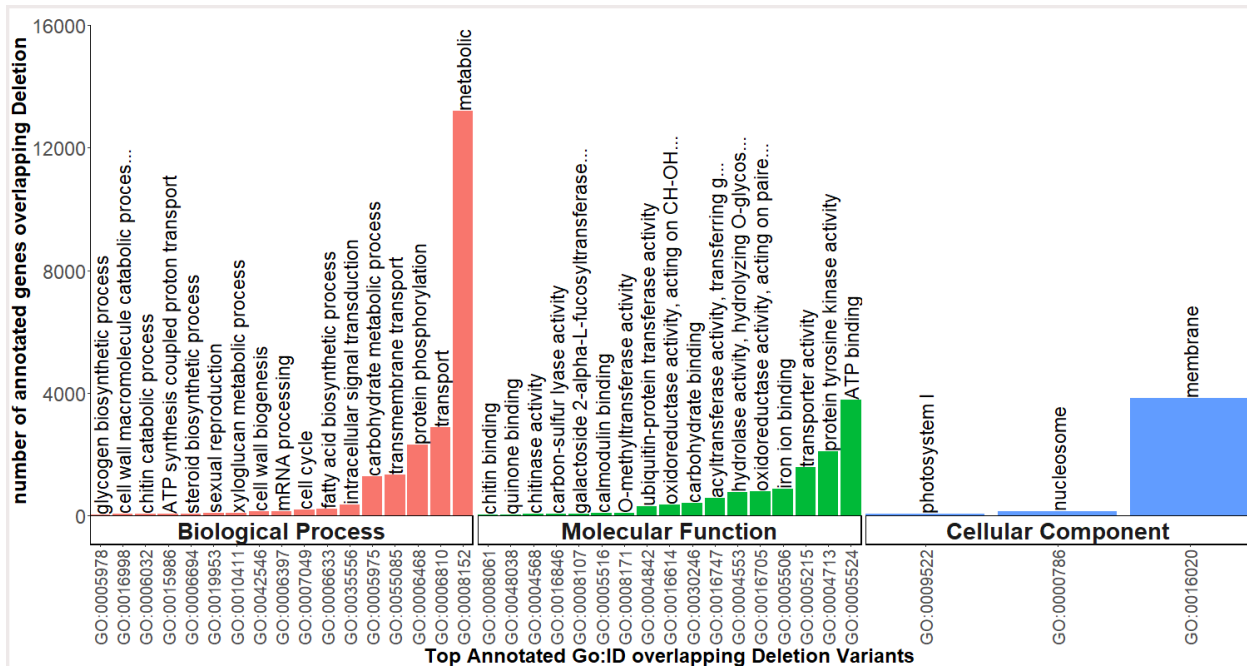


Figure 3.8: GO functional annotation of significant ($p < 0.05$) genes overlapping identified deletion variation events in *Eleusine coracana* under Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

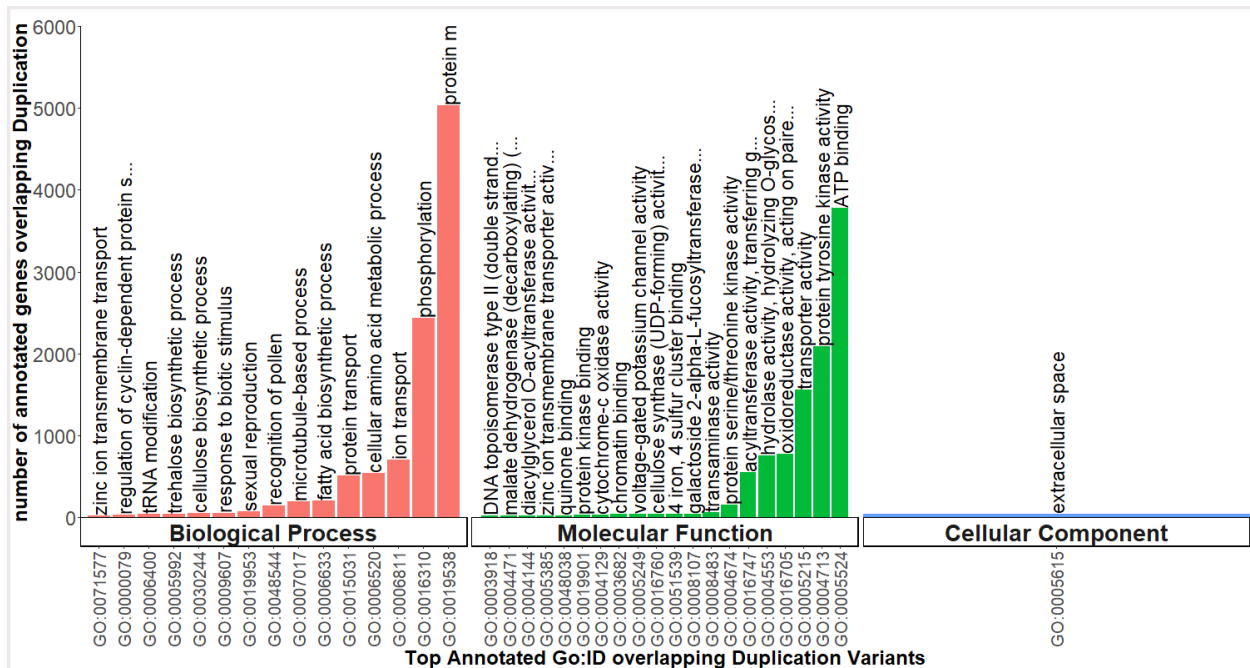


Figure 3.9: GO functional annotation of significant ($p < 0.05$) genes overlapping identified duplication variation events in *Eleusine coracana* under Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

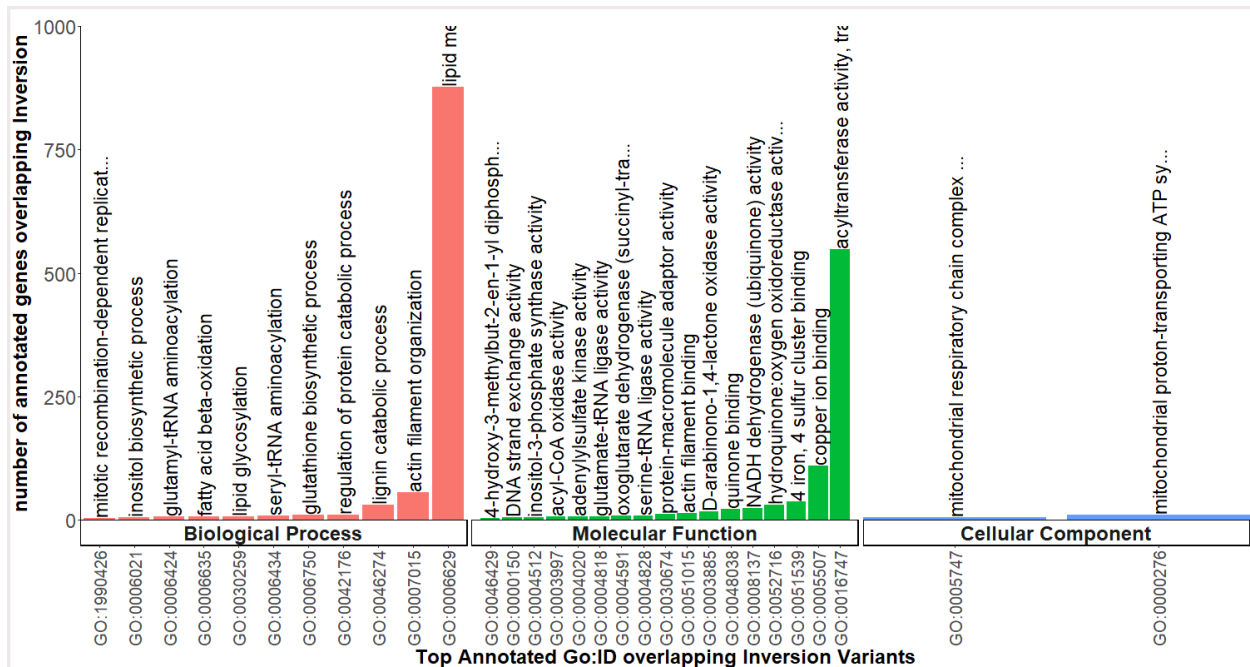


Figure 3.10: GO functional annotation of significant ($p < 0.05$) genes overlapping identified duplication variation events in *Eleusine coracana* under Biological Process (BP), Molecular Function (MF), and Cellular Component (CC).

Discussion

Detection of structural variations in *E. coracana* using whole-genome re-sequencing data

This analysis is the first analysis of structural variants in *E. coracana*, and it utilized 116 "short reads" whole genome sequences (WGS) generated by high-throughput sequencing from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA). To date, our knowledge of genomic variations in *E. coracana* has primarily been about Single Nucleotide Polymorphisms (SNPs). However, it has become evident that SNPs do not capture long genetic variations, which may affect the dosage and presence of genes leading to phenotypic diversity within a species (Alkan et al., 2011; Baker 2012; Sudmant et al. 2015; Layer et al., 2014; Cook et al., 2012; Fuentes et al., 2019; Kyriakidou et al., 2019; Zmienko et al., 2020). Genomic structural variants are crucial to consider when uncovering the genetic basis of observable plant traits (Cook et al., 2012; Fuentes et al., 2019; Kyriakidou et al., 2019; Zmienko et al., 2020). In a series of bash and R scripts, I performed SV analysis of *E. coracana* by removing low-quality reads from downloaded WGS and mapping them to the newly published draft *E. coracana* genome on Phytozome 13. My analysis utilized *LUMPy*, a probabilistic framework for calling SVs based on read-pair, read-depth, and split read signals. *LUMPy* has been identified as a high-performing SV-caller (Layer et al., 2014; Kosugi et al., 2019). WGS analyzed in this study were generated from young leaves of ninety-three finger millet accessions. There are over 10,000 documented accessions of *E. coracana*, so the number of accessions represented in this study is small (Sood *et al.*, 2019). Results show SVs are important in *E. coracana*, and it would be interesting to incorporate more accessions in future analysis.

Data Filtering and Quality Analysis

Downloaded WGS retained at least 55 % of their reads or more than 5 million reads after trimming lower quality reads and adapter sequences; therefore, all WGS were represented in mapping and SV detection. However, I could not extract any information about SVs in 5 WGS which had about 55% reads remaining after trimming. Although the reason for no detection was not verified, it is likely due to low quality and reduced number of reads, as reported in their copy number variation (CNV) analysis of *Arabidopsis thaliana* (Zmienko *et al.*, 2020).

Number of structural variants detected.

On average, *LUMP*y made about ten thousand raw structural variants calls in this analysis. The raw SVs detected do not include coverage regions greater than five standard deviations in each WGS, removed to reduce false positives SV calls as recommended (Li, 2014; Layer *et al.*, 2014). However, it is possible that excluding high coverage areas reduced the sensitivity of SV calls. Filtering raw calls and merging overlapping events in each WGS cut the number of identified variants to five thousand on average. The filtering process also excluded translocation events (identified as BND by *LUMP*y) due to their complexity and the inability to differentiate the events contained within them in this analysis. Difficulty in differentiating BNDs was also reported as a challenge in the SV analysis of rice ((Fuentes *et al.*, 2021). The number of high confidence SVs retained is close to the reported eight thousand SV in the papaya genome (Liao *et al.*, 2021) but incomparable to the approximately 1.5 million SV events found in the rice genome (Fuentes *et al.*, 2019).

Identification of Deletion, Duplication, and Insertion Events

Structural variants events were merged across WGS, followed by the combination of overlapping regions. The merger resulted in ninety-three inversions, 1,922 duplications, and 3,344 deletions variants. The reported number of structural variants per event is in finger millet is

consistent with the number of SV reported for papaya (Liao *et al.*, 2021). Furthermore, the high incidence of deletion and duplication events is consistent with theoretical predictions that polyploidy increases the likelihood of occurrence of genomic structural variants, and that the path to diploidization involves the loss, retention, or maintenance of duplicate genes due to increased sequence similarity (Adams and Wendel, 2005; Schiessl *et al.*, 2018). The higher number of deletions indicates that deletions were very common in the finger millet genome. Theory suggests that hybridization leading to activation of genes and promoting unequal crossing over are causally responsible for high deletion variants in allopolyploid genomes like *E. coracana* (Otto, 2007). The detected low inversion events are consistent with other plant studies, and the low records have been explained as a likely result of purging these events from essential genes due to their deleterious effects (Zmienko *et al.*, 2019; Hämälä *et al.*, 2021; Liao *et al.*, 2021; Fuentes *et al.*, 2021). The distribution of the events in IGV viewer shows that each chromosome has a reasonably equal amount of SV for each event, suggesting that each chromosome may have been subjected to a similar selection process (Liao *et al.*, 2021). However, the distribution of SVs along the genome has an uneven coverage, suggesting that functional constraints may have interacted with the abundance of SVs and impacted their distribution (Otto, 2007; Zmienko *et al.*, 2020).

The interplay between the Structural Variation and Genes

The intersection of identified structural variants with the genic region in the *E. coracana* draft genome shows three-hundred and twenty-four genes overlap inversion events, 40,747 gene coding regions overlap duplication events and 41,238 gene coding regions deletion events across the finger millet chromosomes. The overlap with gene coding regions further strengthens the hypothesis that SVs pose structural and functional constraints on genes and affects their dosages. Structural variant simulation studies predict that SVs tend to accumulate deleterious variants and

thus may constrain adaptation (Berdan *et al.*, 2021). Most of these SVs may be signatures of selection and adaptation in *E. coracana* accessions.

Functional impact of SV-overlapped genes

The GO annotations of gene overlapping SVs reveal that metabolic biosynthetic and transport processes are top biological processes affected by deletion and duplication. A high overlap is observed in the significant gene categories under biological function and molecular component categories of deletion and duplication events. This functional gene overlap between deletion and duplication events may suggest that duplication events could have lessened the effects of deletion variants. In addition, key processes critical to biotic stress responses, which might play important roles in environmental adaptability, were also highlighted in duplication events in *E. coracana*. Further investigation and analysis of genes present in significant categories would provide an opportunity to understand domestication, diversification, and adaptation in *E. coracana* and provide resources for developing molecular markers (Schiessl *et al.*, 2018).

Conclusion and future recommendations

There is increasing attention to the role of structural variants in plant species diversification and adaptation. The number of plant species for which SV regions have been identified at the genome-wide scale has proliferated within the last decade (Cook *et al.*, 2012; Fuentes *et al.*, 2019; Kyriakidou *et al.*, 2019; Zmienko *et al.*, 2020). This study, hopefully, lays the groundwork for identifying structural genomic variations that can help our understanding and improvement of *E. coracana*. It is crucial to analyze the WGS used in this study in combined multiple approaches with other high-performance SV callers like *PINDEL* and *DELLY* and compare the results. Combined multiple approaches are fundamental to producing a more robust prediction and reducing error calls from *LUMPy*.

Furthermore, an integrative study that would involve detailed characterization and validation of identified structural variants and their impact on gene dosages would help identify and develop desired agronomic traits. A recent extensive genome-wide genotyping of 423 finger millet landraces, using the same genome assembly used in this study, identified 8,778 SNPs. Identified SNPs were used to analyze patterns of divergence and population structure (Bančič *et al.*, preprint 2021). There are at least 10,000 recorded accessions of *E. coracana*, and only three accessions in this study were used in the SNP study. Generally, SVs show a similar population structure with SNPs, albeit with weaker signals. Although not covered in this study, it will be interesting to investigate the similarities between the distribution genomic variation and population divergence in SNPs and SVs analyses of *E. coracana*. Structural Variants and transposable elements (TEs) reportedly have similar high sequence genomic distribution, and it has been suggested that SVs are products of TE activity. Therefore, it would also be noteworthy to incorporate how genomic distribution of SVs correlates with TEs in the *E. coracana* genome.

Chapter 4: General Conclusion

In this thesis, I predicted bioclimatic and edaphic factors that affect the distribution of *Eleusine* species in Africa using the full Africa map extent and extent narrowed to countries in the collection record. Maxent worked quite similarly to a large degree in the two extents; however, the narrow extent had the advantage of identifying likely suitable environments. Further understanding of the distribution pattern and factor is hinged on collaborating with known locality records for field verifications. It is essential to carefully repeat sampling in determining realistic environmental factors and in building strong distribution models. Good, repeated field observations would help adopt a distribution model that accounts for imperfect detections of large-scale analysis.

Secondly, I investigated structural variations in the allotetraploid, *E. coracana*. The results show a high incidence of structural variants in the *E. coracana* genome, overlapping essential genes in critical biological processes such as metabolic and biotic stress adaptations. The result suggests they play an essential role in evolution, growth, and development. It is necessary to corroborate the findings with other high-quality SV callers. Furthermore, I recommend investigating identified variants in future genomic variations analyses targeting crop improvement in *E. coracana*.

Overall, the research approaches used in this thesis underscore the usefulness of public data for plant research, and they demonstrate the possibility of analyzing extensive data using computational biology tools. The approaches together present the first data uncovering environmental preferences and genomic variation influences in *Eleusine* and can help our understanding of the genus.

References

- Abyzov, A., Urban, A.E., Snyder, M. and Gerstein, M., 2011. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome research*, 21(6), pp.974-984.
- Adams, K.L. and Wendel, J.F., 2005. Polyploidy and genome evolution in plants. *Current opinion in plant biology*, 8(2), pp.135-141.
- Alkan, C., Coe, B.P. and Eichler, E.E., 2011. Genome structural variation discovery and genotyping. *Nature Reviews Genetics*, 12(5), pp.363-376.
- Baker, M., 2012. Structural variation: the genome's hidden architecture. *Nature methods*, 9(2), pp.133-137.
- Bancic, J., Odeny, D.A., Ojulong, H.F., Josiah, S.M., Buntjer, J., Gaynor, R.C., Hoad, S.P., Gorjanc, G. and Dawson, I.K., 2021. Genomic and phenotypic characterization of finger millet indicates a complex diversification history. *bioRxiv*.
- Bean, William T., Robert Stafford, and Justin S. Brashares. "The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models." *Ecography* 35, no. 3 (2012): 250-258.
- Belsky, A.J., 1994. Influences of trees on savanna productivity: tests of shade, nutrients, and tree-grass competition. *Ecology*, 75(4), pp.922-932.
- Bisht, M.S. and Mukai, Y., 2000. Mapping of rDNA on the chromosomes of Eleusine species by fluorescence in situ hybridization. *Genes & genetic systems*, 75(6), pp.343-348.
- Bisht, M.S. and Mukai, Y., 2001. Genomic in situ hybridization identifies genome donor of finger millet (*Eleusine coracana*). *Theoretical and Applied Genetics*, 102(6-7), pp.825-832.
- Bisht, M.S. and Mukai, Y., 2002. Genome organization and polyploid evolution in the genus *Eleusine* (Poaceae). *Plant Systematics and Evolution*, 233(3), pp.243-258.
- Bocksberger, G., Schnitzler, J., Chatelain, C., Daget, P., Janssen, T., Schmidt, M., Thiombiano, A. and Zizka, G., 2016. Climate and the distribution of grasses in West Africa. *Journal of Vegetation Science*, 27(2), pp.306-317.
- Chaudhry, S. and Sidhu, G.P.S., 2021. Climate change regulated abiotic stress mechanisms in plants: a comprehensive review. *Plant Cell Reports*, pp.1-31.
- Chen, J.C., Huang, H.J., Wei, S.H., Zhang, C.X. and Huang, Z.F., 2015. Characterization of glyphosate-resistant goosegrass (*Eleusine indica*) populations in China. *Journal of Integrative Agriculture*, 14(5), pp.919-925.

- Chennaveeraiah, M.S. and Hiremath, S.C., 1974. Genome analysis of *Eleusine coracana* (L.) Gaertn. *Euphytica*, 23(3), pp.489-495.
- Christenhusz, M. J., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, 261(3), 201-217.
- Cook, D.E., Lee, T.G., Guo, X., Melito, S., Wang, K., Bayless, A.M., Wang, J., Hughes, T.J., Willis, D.K., Clemente, T.E. and Diers, B.W., 2012. Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *science*, 338(6111), pp.1206-1209.
- Cotton, J.L., Wysocki, W.P., Clark, L.G., Kelchner, S.A., Pires, J.C., Edger, P.P., Mayfield-Jones, D. and Duvall, M.R., 2015. Resolving deep relationships of PACMAD grasses: a phylogenomic approach. *BMC plant biology*, 15(1), pp.1-11.
- De Wet, J.M.J., Rao, K.P., Brink, D.E. and Mengesha, M.H., 1984. Systematics and evolution of *Eleusine coracana* (Gramineae). *American journal of Botany*, 71(4), pp.550-557.
- Dhankher, O.P. and Foyer, C.H., 2018. Climate resilient crops for improving global food security and safety.
- Duitama, J., Quintero, J.C., Cruz, D.F., Quintero, C., Hubmann, G., Foulquie-Moreno, M.R., Verstrepen, K.J., Thevelein, J.M. and Tohme, J., 2014. An integrated framework for discovery and genotyping of genomic variants from high-throughput sequencing experiments. *Nucleic acids research*, 42(6), pp.e44-e44.
- Duvall, M. R., Davis, J. I., Clark, L. G., Noll, J. D., Goldman, D. H., & Sánchez-Ken, J. G. (2007). Phylogeny of the grasses (Poaceae) revisited. *Aliso: A Journal of Systematic and Evolutionary Botany*, 23(1), 237-247.
- Edwards, D. and Batley, J., 2004. Plant bioinformatics: from genome to phenome. *TRENDS in Biotechnology*, 22(5), pp.232-237.
- Ellis, R.P., 1984. *Eragrostis walteri*—a first record of non-Kranz leaf anatomy in the sub-family Chloridoideae (Poaceae). *South African Journal of Botany*, 3(6), pp.380-386.
- Elith, Jane, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, and Colin J. Yates. "A statistical explanation of MaxEnt for ecologists." *Diversity and distributions* 17, no. 1 (2011): 43-57.
- Escaramís, G., Docampo, E. and Rabionet, R., 2015. A decade of structural variants: description, history and methods to detect structural variation. *Briefings in functional genomics*, 14(5), pp.305-314.

- Francia, E., Pecchioni, N., Policriti, A. and Scalabrin, S., 2015. CNV and structural variation in plants: prospects of NGS approaches. In *Advances in the understanding of biological sciences using next generation sequencing (NGS) approaches* (pp. 211-232). *Springer*, Cham.
- Fuentes, R.R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J.F., Mohiyuddin, M., Wing, R.A., McNally, K.L., Tatarinova, T., Grigoriev, A. and Mauleon, R., 2019. Structural variants in 3000 rice genomes. *Genome research*, 29(5), pp.870-880.
- Ganeshiah, K.N. and Shaanker, R.U., 1982. Evolution of reproductive behaviour in the genus *Eleusine*. *Euphytica*, 31(2), pp.397-404.
- Gasser, M. and Vegetti, A.C., 1997. Inflorescence typology in *Eleusine indica* and *Eleusine tristachya* (Poaceae). *Flora*, 192(1), pp.17-20.
- Guisan, A. and Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. *Ecology letters*, 8(9), pp.993-1009.
- Hämälä, T., Wafula, E.K., Guiltinan, M.J., Ralph, P.E., dePamphilis, C.W. and Tiffin, P., 2021. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proceedings of the National Academy of Sciences*, 118(35).
- Hartley, W. "The global distribution of tribes of the Gramineae in relation to historical and environmental factors." *Australian Journal of Agricultural Research* 1, no. 4 (1950): 355-372.
- Hawkins, B.A., Field, R., Cornell, H.V., Currie, D.J., Guégan, J.F., Kaufman, D.M., Kerr, J.T., Mittelbach, G.G., Oberdorff, T., O'Brien, E.M. and Porter, E.E., 2003. Energy, water, and broad-scale geographic patterns of species richness. *Ecology*, 84(12), pp.3105-3117.
- Li, Heng. "Toward better understanding of artifacts in variant calling from high-coverage samples." *Bioinformatics* 30, no. 20 (2014): 2843-2851.
- Hengl, T., Miller, M.A., Križan, J., Shepherd, K.D., Sila, A., Kilibarda, M., Antonijević, O., Glušica, L., Dobermann, A., Haefele, S.M. and McGrath, S.P., 2021. African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, 11(1), pp.1-18.
- Hernandez, P.A., Graham, C.H., Master, L.L. and Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5), pp.773-785.
- Hilu, K.W. and De Wet, J.M.J., 1976. Domestication of *Eleusine coracana*. *Economic Botany*, 30(3), pp.199-208.

- Hilu, K.W. and Johnson, J.L., 1992. Ribosomal DNA variation in finger millet and wild species of Eleusine (Poaceae). *Theoretical and Applied Genetics*, 83(6-7), pp.895-902.
- Hilu, K.W., 1995. Evolution of finger millet: evidence from random amplified polymorphic DNA. *Genome*, 38(2), pp.232-238.
- Hiremath, S.C. and Chennaveeraiah, M.S., 1982. Cytogenetical studies in wild and cultivated species of Eleusine (Gramineae). *Caryologia*, 35(1), pp.57-69.
- Hiremath, S.C. and Salimath, S.S., 1991. Quantitative nuclear DNA changes in Eleusine (Gramineae). *Plant Systematics and Evolution*, 178(3), pp.225-233.
- Holm, L.G., Plucknett, D.L., Pancho, J.V. and Herberger, J.P., 1977. The World's Worst Weeds. *The world's worst weeds*.
- Iovene, M., Zhang, T., Lou, Q., Buell, C.R. and Jiang, J., 2013. Copy number variation in potato—an asexually propagated autotetraploid species. *The Plant Journal*, 75(1), pp.80-89.
- Kellogg, E. A. (2001). Evolutionary history of the grasses. *Plant physiology*, 125(3), 1198-1205.
- Kennedy-O'Byrne, J., 1957. Notes on African grasses: XXIX. A new species of Eleusine from tropical and South Africa. *Kew Bulletin*, pp.65-72.
- Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. and Kamatani, Y., 2019. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome biology*, 20(1), pp.1-18.
- Kyriakidou, M., Achakkagari, S.R., López, J.H.G., Zhu, X., Tang, C.Y., Tai, H.H., Anglin, N.L., Ellis, D. and Strömvik, M.V., 2020. Structural genome analysis in cultivated potato taxa. *Theoretical and Applied Genetics*, 133(3), pp.951-966.
- Layer, R.M., Chiang, C., Quinlan, A.R. and Hall, I.M., 2014. LUMPY: a probabilistic framework for structural variant discovery. *Genome biology*, 15(6), pp.1-19.
- Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K. and Wang, J., 2009. SNP detection for massively parallel whole-genome resequencing. *Genome research*, 19(6), pp.1124-1132.
- Li, Y.F., Hong, H.L., Li, Y.H., Chang, R.Z. and Qiu, L.J., 2016. The identification of presence/absence variants associated with the apparent differences of growth period structures between cultivated and wild soybeans. *Journal of integrative agriculture*, 15(2), pp.262-270.
- Liao, Z., Zhang, X., Zhang, S., Lin, Z., Zhang, X. and Ming, R., 2021. Structural variations in papaya genomes. *BMC genomics*, 22(1), pp.1-13.

- Lin, C.T. and Chiu, C.A., 2020. Comparison of predictor selection procedures in species distribution modeling: a case study of *Fagus hayatae*. *Cerne*, 26, pp.172-182.
- Liu, C., Newell, G. and White, M., 2016. On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and evolution*, 6(1), pp.337-348.
- Liu, C., White, M. and Newell, G., 2009, July. Measuring the accuracy of species distribution models: a review. In *Proceedings 18th World IMACs/MODSIM Congress. Cairns, Australia* (Vol. 4241, p. 4247).
- Liu, Q. and Peterson, PM, 2010. Advances in systematics of adaptively radiated Eleusine Gaertn.(Poaceae). *Journal of Tropical and Subtropical Botany*.
- Liu, Q., Jiang, B., Wen, J. and Peterson, P.M., 2014. Low-copy nuclear gene and McGISH resolves polyploid history of *Eleusine coracana* and morphological character evolution in *Eleusine*. *Turkish Journal of Botany*, 38(1), pp.1-12.
- Liu, Q., Triplett, J.K., Wen, J. and Peterson, P.M., 2011. Allotetraploid origin and divergence in *Eleusine* (Chloridoideae, Poaceae): evidence from low-copy nuclear gene phylogenies and a plastid gene chronogram. *Annals of Botany*, 108(7), pp.1287-1298.
- Medvedev, P., Stanciu, M. and Brudno, M., 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature methods*, 6(11), pp.S13-S20.
- Mod, H.K., Scherrer, D., Luoto, M. and Guisan, A., 2016. What we use is not what we know: environmental predictors in plant distribution models. *Journal of Vegetation Science*, 27(6), pp.1308-1322.
- Mohiyuddin, M., Mu, J.C., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H. and Lam, H.Y., 2015. MetaSV: an accurate and integrative structural-variant caller for next generation sequencing. *Bioinformatics*, 31(16), pp.2741-2744.
- Muchut, S.E., Pilatti, V., Uberti-Manassero, N., Vegetti, A.C. and Reinheimer, R., 2017. Inflorescence diversity in subtribe Eleusininae (Poaceae: chloridoideae: Cynodonteae). *Flora*, 228, pp.50-59.
- National Research Council. *Lost crops of Africa: volume I: grains*. National Academies Press, 1996.
- Neves, S.S., Swire-Clark, G., Hilu, K.W. and Baird, W.V., 2005. Phylogeny of *Eleusine* (Poaceae: Chloridoideae) based on nuclear ITS and plastid trnT-trnF sequences. *Molecular phylogenetics and evolution*, 35(2), pp.395-419.
- Otto, S.P., 2007. The evolutionary consequences of polyploidy. *Cell*, 131(3), pp.452-462.

- Pacifici, K., Reich, B.J., Miller, D.A. and Pease, B.S., 2019. Resolving misaligned spatial data with integrated species distribution models. *Ecology*, 100(6), p.e02709.
- Papeş, M., and Philippe Gaubert. "Modelling ecological niches from low numbers of occurrences: assessment of the conservation status of poorly known viverrids (Mammalia, Carnivora) across two continents." *Diversity and distributions* 13, no. 6 (2007): 890-902.
- Parashuram, D.P., Jayaram Gowda, S.R. and Mallikarjun, N.M., 2011. Heterosis and combining ability studies for yield and yield attributing characters in finger millet (*Eleusine coracana* (L.) Gaertn.). *Electronic Journal of Plant Breeding*, 2(4), pp.494-500.
- Parr, C.L., Lehmann, C.E., Bond, W.J., Hoffmann, W.A. and Andersen, A.N., 2014. Tropical grassy biomes: misunderstood, neglected, and under threat. *Trends in ecology & evolution*, 29(4), pp.205-213.
- Pasturel, M., Alexandre, A., Novello, A., Dièye, A.M., Wélé, A., Paradis, L., Cordova, C. and Hély, C., 2016. Grass physiognomic trait variation in African herbaceous biomes. *Biotropica*, 48(3), pp.311-320.
- Pearce, J.L. and Boyce, M.S., 2006. Modelling distribution and abundance with presence-only data. *Journal of applied ecology*, 43(3), pp.405-412.
- Peterson, P.M., Romaschenko, K. and Johnson, G., 2010. A classification of the Chloridoideae (Poaceae) based on multi-gene phylogenetic trees. *Molecular Phylogenetics and Evolution*, 55(2), pp.580-598.
- Peterson, P.M., Romaschenko, K., Herrera Arrieta, Y. and Vorontsova, M.S., 2021. Phylogeny, classification, and biogeography of *Afrotrichloris*, *Apochiton*, *Coelachyrum*, *Dinebra*, *Eleusine*, *Leptochloa*, *Schoenefeldia*, and a new genus, *Schoenefeldiella* (Poaceae: Chloridoideae: Cynodonteae: Eleusininae). *Journal of Systematics and Evolution*.
- Phillips, S., 1995. Vol. 7: Poaceae (Gramineae). Addis Ababa: Addis Ababa University.
- Phillips, S.J., Anderson, R.P. and Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4), pp.231-259.
- Phillips, S.M., 1972. A survey of the genus *Eleusine* Gaertn. (Gramineae) in Africa. *Kew Bulletin*, pp.251-270.
- Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V. and Korbel, J.O., 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18), pp.i333-i339.

- Rebello, H. and Jones, G., 2010. Ground validation of presence-only modelling with rare species: A case study on barbastelles *Barbastella barbastellus* (Chiroptera: Vespertilionidae). *Journal of Applied Ecology*, 47(2), pp.410-420.
- Rizk, G., Gouin, A., Chikhi, R. and Lemaitre, C., 2014. MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics*, 30(24), pp.3451-3457.
- Sala, O.E., Chapin, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A. and Leemans, R., 2000. Global biodiversity scenarios for the year 2100. *science*, 287(5459), pp.1770-1774.
- Saxena, R.K., Edwards, D. and Varshney, R.K., 2014. Structural variations in plant genomes. *Briefings in functional genomics*, 13(4), pp.296-307.
- Schiessl, S.V., Katche, E., Ihien, E., Chawla, H.S. and Mason, A.S., 2019. The role of genomic structural variation in the genetic improvement of polyploid crops. *The Crop Journal*, 7(2), pp.127-140.
- Scholes, R.J. and Archer, S.R., 1997. Tree-grass interactions in savannas. *Annual review of Ecology and Systematics*, 28(1), pp.517-544.
- Schröder, J., Hsu, A., Boyle, S.E., Macintyre, G., Cmero, M., Tothill, R.W., Johnstone, R.W., Shackleton, M. and Papenfuss, A.T., 2014. Socrates: identification of genomic rearrangements in tumour genomes by re-aligning soft clipped reads. *Bioinformatics*, 30(8), pp.1064-1072.
- Shchapova, A. I. (2012). Evolution of the basic chromosome number in Poaceae Barnh. *Russian Journal of Genetics: Applied Research*, 2(3), 252-259.
- Sindi, S., Helman, E., Bashir, A. and Raphael, B.J., 2009. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12), pp.i222-i230.
- Smith, S.D., Kawash, J.K. and Grigoriev, A., 2015. GROM-RD: resolving genomic biases to improve read depth detection of copy number variants. *PeerJ*, 3, p.e836.
- Sood, S., Joshi, D.C., Chandra, A.K. and Kumar, A., 2019. Phenomics and genomics of finger millet: current status and future prospects. *Planta*, 250(3), pp.731-751.
- Soreng, R.J., Peterson, P.M., Romaschenko, K., Davidse, G., Teisher, J.K., Clark, L.G., Barberá, P., Gillespie, L.J. and Zuloaga, F.O., 2017. A worldwide phylogenetic classification of the Poaceae (Gramineae) II: An update and a comparison of two 2015 classifications. *Journal of Systematics and Evolution*, 55(4), pp.259-290.
- Soreng, R.J., Peterson, P.M., Romaschenko, K., Davidse, G., Zuloaga, F.O., Judziewicz, E.J., Filgueiras, T.S., Davis, J.I. and Morrone, O., 2015. A worldwide phylogenetic

- classification of the Poaceae (Gramineae). *Journal of Systematics and Evolution*, 53(2), pp.117-137.
- Still, C.J., Berry, J.A., Collatz, G.J. and DeFries, R.S., 2003. Global distribution of C3 and C4 vegetation: carbon cycle implications. *Global biogeochemical cycles*, 17(1), pp.6-1.
- Strömberg, C. A. (2011). Evolution of grasses and grassland ecosystems. *Annual review of Earth and planetary sciences*, 39, 517-544.
- Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H.Y. and Konkel, M.K., 2015. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571), pp.75-81.
- Sutton, T., Baumann, U., Hayes, J., Collins, N.C., Shi, B.J., Schnurbusch, T., Hay, A., Mayo, G., Pallotta, M., Tester, M. and Langridge, P., 2007. Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science*, 318(5855), pp.1446-1449.
- Werth, C.R., Hilu, K.W. and Langner, C.A., 1994. Isozymes of Eleusine (Gramineae) and the origin of finger millet. *American Journal of Botany*, 81(9), pp.1186-1197.
- Yang, R., Nelson, A.C., Henzler, C., Thyagarajan, B. and Silverstein, K.A., 2015. ScanIndel: a hybrid framework for indel detection via gapped alignment, split reads and de novo assembly. *Genome medicine*, 7(1), pp.1-12.
- Zhang, H., Hall, N., Goertzen, L.R., Chen, C.Y., Peatman, E., Patel, J. and McElroy, J.S., 2019. Transcriptome Analysis Reveals Unique Relationships Among Eleusine Species and Heritage of Eleusine coracana. *G3: Genes, Genomes, Genetics*, 9(6), pp.2029-2036.
- Zmienko, A., Marszalek-Zenczak, M., Wojciechowski, P., Samelak-Czajka, A., Luczak, M., Kozłowski, P., Karłowski, W.M. and Figlerowicz, M., 2020. AthCNV: A map of DNA copy number variations in the Arabidopsis genome. *The Plant Cell*, 32(6), pp.1797-1819.

Appendix 1. code used for selecting mapping extent for each species in ArcGIS

```
## code for Eleusine africana
ADM0_NAME = 'Angola' Or ADM0_NAME = 'Botswana' Or ADM0_NAME = 'Burkina Faso' Or ADM0_NAME =
'Burundi' Or ADM0_NAME = 'Cameroon' Or ADM0_NAME = 'Chad' Or ADM0_NAME = 'Eswatini' Or ADM0_NAME
= 'Ethiopia' Or ADM0_NAME = 'Gambia' Or ADM0_NAME = 'Kenya' Or ADM0_NAME = 'Lesotho' Or ADM0_NAME
= 'Madagascar' Or ADM0_NAME = 'Malawi' Or ADM0_NAME = 'Mali' Or ADM0_NAME = 'Mozambique' Or
ADM0_NAME = 'Namibia' Or ADM0_NAME = 'Nigeria' Or ADM0_NAME = 'Rwanda' Or ADM0_NAME = 'Senegal'
Or ADM0_NAME = 'Seychelles' Or ADM0_NAME = 'South Africa' Or ADM0_NAME = 'Tanzania, United Republic
of' Or ADM0_NAME = 'Uganda' Or ADM0_NAME = 'Zambia' Or ADM0_NAME = 'Zimbabwe'

## code for Eleusine coracana
ADM0_NAME = 'Angola' Or ADM0_NAME = 'Burkina Faso' Or ADM0_NAME = 'Burundi' Or ADM0_NAME =
'Cameroon' Or ADM0_NAME = 'Comoros' Or ADM0_NAME = 'Ethiopia' Or ADM0_NAME = 'Guinea-Bissau' Or
ADM0_NAME = 'Kenya' Or ADM0_NAME = 'Madagascar' Or ADM0_NAME = 'Malawi' Or ADM0_NAME = 'Mozambique'
Or ADM0_NAME = 'Nigeria' Or ADM0_NAME = 'Rwanda' Or ADM0_NAME = 'South Africa' Or ADM0_NAME =
'South Sudan' Or ADM0_NAME = 'Tanzania, United Republic of' Or ADM0_NAME = 'Uganda' Or ADM0_NAME
= 'Zambia' Or ADM0_NAME = 'Zimbabwe'

## code for Eleusine floccifolia
ADM0_NAME = 'Eritrea' Or ADM0_NAME = 'Ethiopia' Or ADM0_NAME = 'Somalia'

## code for Eleusine indica
ADM0_NAME = 'Angola' Or ADM0_NAME = 'Benin' Or ADM0_NAME = 'Botswana' Or ADM0_NAME = 'Burkina
Faso' Or ADM0_NAME = 'Burundi' Or ADM0_NAME = 'Cape Verde' Or ADM0_NAME = 'Cameroon' Or ADM0_NAME
= 'Central African Republic' Or ADM0_NAME = 'Congo' Or ADM0_NAME = 'Côte d'Ivoire' Or ADM0_NAME
= 'Equatorial Guinea' Or ADM0_NAME = 'Eritrea' Or ADM0_NAME = 'Eswatini' Or ADM0_NAME = 'Ethiopia'
Or ADM0_NAME = 'Gabon' Or ADM0_NAME = 'Ghana' Or ADM0_NAME = 'Guinea' Or ADM0_NAME = 'Guinea-
Bissau' Or ADM0_NAME = 'Kenya' Or ADM0_NAME = 'Liberia' Or ADM0_NAME = 'Madagascar' Or ADM0_NAME
= 'Malawi' Or ADM0_NAME = 'Mali' Or ADM0_NAME = 'Mauritania' Or ADM0_NAME = 'Mauritius' Or
ADM0_NAME = 'Mayotte' Or ADM0_NAME = 'Morocco' Or ADM0_NAME = 'Mozambique' Or ADM0_NAME =
'Namibia' Or ADM0_NAME = 'Nigeria' Or ADM0_NAME = 'Rwanda' Or ADM0_NAME = 'Senegal' Or ADM0_NAME
= 'Seychelles' Or ADM0_NAME = 'Sierra Leone' Or ADM0_NAME = 'South Africa' Or ADM0_NAME = 'South
Sudan' Or ADM0_NAME = 'Tanzania, United Republic of' Or ADM0_NAME = 'Togo' Or ADM0_NAME =
'Tunisia' Or ADM0_NAME = 'Uganda' Or ADM0_NAME = 'Zambia' Or ADM0_NAME = 'Zimbabwe'

## code for Eleusine intermedia
ADM0_NAME = 'Ethiopia' Or ADM0_NAME = 'Kenya' Or ADM0_NAME = 'Somalia'

## code for Eleusine jaegeri
ADM0_NAME = 'Kenya' Or ADM0_NAME = 'Tanzania, United Republic of' Or ADM0_NAME = 'Uganda'

## code for Eleusine kigeziensis
ADM0_NAME = 'Burundi' Or ADM0_NAME = 'Ethiopia' Or ADM0_NAME = 'Rwanda' Or ADM0_NAME = 'Uganda'

## code for Eleusine multiflora
ADM0_NAME = 'Ethiopia' Or ADM0_NAME = 'Kenya' Or ADM0_NAME = 'Lesotho' Or ADM0_NAME = 'South
Africa' Or ADM0_NAME = 'Tanzania, United Republic of'

## code for Eleusine tristachya
ADM0_NAME = 'Algeria' Or ADM0_NAME = 'South Africa'
```