Ethiop. J. Agric. Sci. 30(1) 49-68 (2020)

Restriction Site Associated DNA Sequencing based Single Nucleotide Polymorphism Discovery in Selected Tef (*Eragrostis tef*) and Wild *Eragrostis* Species

Dejene Girma[1,2], Gina Canarozzi[1], Annett Weichert[1] and Zerihun Tadele[1,3,4]

[1] *Institute of Plant Sciences, University of Bern, Altenbergrain 21, 3013 Bern, Switzerland*

[2] *Ethiopian Institute of Agricultural Research, National Agricultural Biotechnology Research Center P.O. Box 249 Holetta, Ethiopia; [3]Addis Ababa University, Institute of Biotechnology, P.O. Box 32853, Ethiopia; [4]Center for Development and Environment (CDE), University of Bern, Bern 3012, Switzerland*

አህፅሮት

ጤፍ በኢትዮጵያ የተገኘ እና በኢትዮጵያዊያን አርሶአደሮች ለብዙ አመታት ሲመረት የኖረ ዘርፈ ብዙ ጥቅሞች ያሉት የሰብል አይነት ነዉ፡፡ ጤፍ የሚመረተውና ጥቅም ላይ የሚውለው በአብዛኛው በኢትዮጵያ በመሆኑ በአለም አቀፍ ደረጃ ጤፍ ላይ የሚደረጉ ምርምሮች ውሱን ናቸው፡፡ በተለይ የጤፍን ምርምር ለማሳለጥ የተዘጋጁ የሞለኪውላር ሳይንስ ግብአቶች ውሱን በመሆናቸው በሌሎች ሰብሎች ላይ ጥቅም ላይ የዋሉ የዲ ኤን ኤ ሲኩዌንሲንግ ቴክኖሎጂን መሰረት ያደረጉ የምርምር አቅጣጫዎችን መከተል ወሳኝ ነው፡፡ እዚህ በቀረበው የምርምር ስራ የዲ ኤን ኤ ሲኩዌንሲንግ ቴክኖሎጂን መሰረት ያደረገን የምርምር ስልት በመከተል ከአገሪቱ የተለያዩ አካባቢዎች የተሰበሰቡ አርባ ሁለት የጤፍ ዝርያዎች፣ አንድ ሚውታንት ላይን እና ሁለት ለጤፍ ቅርብ የሆኑ ዋይልድ ፊላቲቭስ ላይ ጥናት ተደርጎ የጤፍን ምርምር በዘመናዊ መልኩ የሚያግዙ ግብአቶች ተገኝተዋል፡፡

Abstract

Genome-wide knowledge about the nature and extent of genetic diversity present in tef (Eragrostis tef), the most consumed food grain in Ethiopia is limited. Adopting next generation sequencing (NGS) protocols to enhance its genomics and breeding is essential. Here, we applied the Restriction Site Associated DNA (RAD) sequencing protocol and surveyed the genomes of 43 tef landraces, one mutant line and two wild Eragrostis species. After mapping sequencing reads to the de novo assembled unitag and the tef reference genome, a total of 9,024 and 58,735 high quality single nucleotide polymorphisms (SNPs) were identified, respectively. We identified greater number of SNPs and greater nucleotide diversity in the two wild Eragrostis species than in the tef landraces. The tef landrace populations in this study were poorly differentiated with F_{ST} values of 0.015. In the phylogenetic analysis, grouping of the landraces was not consistent with the area of collection, but few localized grouping of the landraces was evident, probably showing the communality of tef seed use across geographical boundaries. The improved tef varieties show reduced genetic diversity compared to the landraces and were all grouped into one cluster reflecting the nature of tef breeding which largely targets common genomic regions. We suggest that future work needs to aim beyond common genomic regions. The work presented here is a valuable addition to the growing molecular resources developed for tef genetic improvement.

Introduction

Owing to their central importance for global food security, much of the world food crops such as wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.), barley (*Hordium* *vulgare*), sorghum (*Sorghum bicolar* L.) and maize (*Zea maize*) have been studied in greater detail with the genomes of each species sequenced. Tef (*Eragrostis tef*) is one such key food security crop to millions of people in East Africa. The crop is known for being the major part of the daily meal for millions of people in Ethiopia. Its resilience to poor growth condition and highest market price compared to the major cereals are some of the qualities that make tef the top food security crop.

As the demand for high yielding and lodging tolerant improved tef varieties has increased, the need to assist the conventional tef breeding and the tef genomics research with modern genomic tools have become apparent. Tools are available, such as those used on similar crops, model, and non-model plants.

Genome-wide identification of polymorphisms among individuals within a species is crucial to studying the genetic basis of phenotypic differences and for elucidating the evolutionary history of the species (Srivastava, Wolinski, and Pereira 2014). For this purpose, single nucleotide polymorphisms (SNPs) are becoming increasingly used. A number of methods have been developed for the discovery and genotyping of SNPs including TaqMan and SNPlex SNP genotyping (De la Vega *et al.* 2005), microarray (Gunderson *et al.* 2005) TILLING (Targeting Induced Local Lesions IN Genomes) (McCallum *et al.* 2000) temperature gradient capillary electrophoresis (Hsia *et al.* 2005) and primer-guided nucleotide incorporation assay (Syvanen *et al.* 1990). Alternative SNP discovery methods that employ next-generation sequencing technologies such as the Restriction site-associated sequencing (RAD-Seq) have been developed in recent years and have flourished because of their practicality and low cost.

Modern plant breeding have evolved from conventional breeding to molecular breeding (Gepts and Hancock 2006). Selection within breeding populations differs at various breeding stages, so genetic diversity present in released cultivars of a crop may vary (Rauf *et al.* 2010).

Genetic polymorphism varies among species and within genomes, and has important implications for the evolution and conservation of species (Ellegren and Galtier 2016). Allelic polymorphism and heterozygosity are among the common measures of genetic diversity within a population. On the other hand, genetic variation among populations is frequently measured using fixation index (F*st*) and genetic distance such as Nei's D (Fu 2015). In order to understand the changes in these genetic diversity parameters, genome-wide diversity scans can be conducted. The focus of this study was, therefore, to use the RAD-seq protocol with the Illumina sequencing platform to discover SNPs and genetically characterize the germplasm panel composed of 46 germplasm coming from three species (*E. tef*, *E. minor* and *E. curvula*). The result reported here could stimulate further genomics research on *Eragrostis* species to facilitate their use in tef breeding and genomics research.

Materials and Methods

Germplasm panel

A panel of forty-five tef germplasm was used. The panel included thirty-nine accessions spanning four different areas of collection and obtained from the Gene Bank at the Ethiopian Biodiversity Institute (EBI) (www.ebi.gov.et/), Ethiopia (Table 1), three improved varieties: *Tsedey* (DZ-Cr-37), an improved variety developed through inter-specific hybridization and that which was used to generate the reference tef genome sequence, *Simada* (DZ-Cr-385) and *Magna* (DZ-01-196), all received from Debre Zeit Agricultural Research Center, Ethiopia. *Kegne* (3774-13), a mutant line derived from the *Tsedey* variety was obtained from the Institute of Plant Sciences, University of Bern, Switzerland and two wild relatives (viz *E. curvula* and *E. minor*) were obtained from the United States Department of Agriculture, Agricultural Research Service (USDA-ARS) (https://www.ars.usda.gov/). The tef accessions were collections from diverse agro-ecological regions ranging in altitude from 1000 m to 2860 ml. These accessions were collected from farmers’ fields and/or market places, and represent locally adapted varieties. The three commercially released varieties are the products of extensive breeding through selection and hybridization. The list of the regions and approximate area of collection of the germplasm with the corresponding altitudes (m) is given in Table 1. The accessions as well as the accompanying data were obtained from The Ethiopian Biodiversity Institute (EIB). Based on this data, we grouped the accessions into four major areas of collection. Accessions collected from the North East and those collected from the South East/Central part, each contain nine accessions. Similarly, accessions collected from the North West (thirteen accessions) and accessions collected from South West (five accessions).

Genomic DNA extraction

Seeds of individual genotypes were grown in pots in the growth room at the Institute of Plant Sciences, University of Bern, Switzerland under 12hr light and 12hr dark conditions. Genomic DNA was extracted from 100 mg of leaf tissue obtained from four-week-old individual plants using the CTAB (Chua *et al.* 1990; Doyle and Dickson 1987) protocol with some modifications. Samples were normalized to 20 ng/µl concentration for library preparation and DNA quality and quantity was checked using 1.5% Agarose gel electrophoresis.

Table 1. Information on the germplasm panel used in this study

Accession code	Code number	Region/type	Area of collection	Altitude (m)
Northeast				
234375	11	*Tigray/Adwa*	*Mehakelegnaw*	1000
242568	13	*Tigray/Adwa*	*Mehakelegnaw*	1380
212602	14	*Amahara S.Wollo*	*Tenta*	2690
243492	15	*Amahara S.Wollo*	*Tenta*	2935
212603	16	*Amahara S.Wollo*	*Meqdela*	2750
212592	21	*Amahara S.Wollo*	*Kola-Temben*	2010
235326	25	*Tigray/Wukro*	*Woqro*	2860
243488	39	*Amahara/S.Wollo*	*Qalu*	2180
243515	40	*Tigray/Temben*	*Degu-Temben*	2580
Southeast/Central				
215356	4	*Oromia/Bale*	*Gololcha*	2500
229984	7	*Oromia/Bale*	*Goro*	2120
55100	12	*Oromia/Harerghe*	*Chiro*	2030
237742	17	*Oromia/Bale*	*Adaba*	2380
237687	26	*Oromia/S.Shewa*	*Dendi*	2150
230771	30	*Oromia/Borena*	*Moyale*	1200
237125	31	*Oromia/N.Shewa*	*Kewot*	1360
230586	33	*Oromia/Bale*	*Ginir*	1450
237695	37	*Oromia/W.Shewa*	*Ambo*	2390
Northwest				
229759	3	*Amahara E.Gojam*	*Enbese Sar Mider*	2610
229770	6	*Amahara E.Gojam*	*Awebel*	2700
55062	9	*Amahara E.Gojam*	*Enemay*	2560
236529	10	*Amahara W.Gojam*	*Denbecha*	2060
55184	18	*Amahara W.Gojam*	*Bure Wenberema*	2590
212708	19	*Amahara N.Gondar*	*Wegera*	2800
212715	20	*Amahara S.Gondar*	*Fogera*	2100
212706	22	*Amahara E.Gojam*	*Enarj Enawega*	2600
228969	23	*Amahara E.Gojam*	*Gozamn*	2480
229758	24	*Amahara E.Gojam*	*G.Siso Enese*	2500
229763	35	*Amahara/E.Gojam*	*Enbise Sar Mider*	2610
55185	36	*Amahara /Agew Awi*	*Banja*	2580
212700	38	*Amahara/E.Gojam*	*Debay Telategen*	2540
Southwest				
212930	1	*SNNP*/N.Omo*	*Bonke*	2250
236091	2	*SNNP/Hadiya*	*Limo*	2240
212923	5	*SNNP/Hadiya*	*Konteb*	2300
202949	8	*SNNP/Hadiya*	*Goro*	1120
225751	28	*SNNP/Omo*	*Arbaminch*	1100
236088	29	*SNNP/Omo*	*Humbo*	1450
241674	32	*SNNP/Bench Maji*	*Konso*	1460
2225761a	34	*SNNP/N.Omo*	*Kucha*	1290
Kegne	41	Mutant	NA	NA
DZ-Cr-37 (*Tsedey*)	42	Improved	NA	NA
DZ-Cr-196 (*Magna*)	43	Improved	NA	NA
DZ-Cr-385	44	Improved	NA	NA
E. curvula	46	Wild	NA	NA
E. minor	48	Wild	NA	NA

**SNNP (Southern Nations Nationalities and Peoples) Region*

RAD library preparation and sequencing

The DNA library preparation for RAD sequencing was performed by Floragenex, Inc. (Eugene, OR, USA) following the protocol described by (Baird *et al.* 2008). Genomic DNA from the 45 samples was digested with the restriction endonuclease SbfI-HF and processed into RAD libraries. Briefly, 200 ng of genomic DNA was digested for 60 min at 37°C in a 50 µL reaction with 20 units (U) of SbfI-HF (New England Biolabs [NEB]). After digestion, samples were heat-inactivated for 20 min at 65°C followed by addition of 2.0 µL of 100 nM P1 Adapter(s), a modified Solexa© adapter (Illumina, *Inc*.). PstI P1 adapters each contained a unique multiplex sequence index (barcode) which is read as the first four nucleotides of the Illumina sequence read. One-hundred nM P1 adaptors were added to each sample along with 1 µL of 10 mM rATP (Promega), 1 µL 10× NEB Buffer 4, 1.0 µL (1000 U) T4 DNA Ligase (high concentration, Enzymatics, *Inc*), and 5 µL H_2O which was then incubated at room temperature (RT) for 20 min. Samples were again heat-inactivated for 20 min at 65°C, pooled and randomly sheared with a Bioruptor (Diagenode) to an average size of 400 bp. Samples were then run on a 1.5% agarose (Sigma), 0.5 X TBE gel, and DNA fragments in the range of 250 bp to 500 bp were isolated using a MinElute Gel Extraction Kit (Qiagen). End blunting enzymes (Enzymatics, *Inc*) were then used to polish the ends of the DNA.

Samples were then purified using a MinElute column (Qiagen, *Inc*) and 15 U of Klenow exo− (Enzymatics, *Inc*) was used to add adenine (Fermentas) overhangs on the 3′ end of the DNA at 37°C. After subsequent purification, 1 µL of 10 µM P2 adapter, a divergent modified Solexa© adapter (Illumina, *Inc*.), was ligated to the obtained DNA fragments at 4°C. Samples were again purified and eluted in 15 µL. The eluate was quantified using a Qubit fluorimeter and 10 ng of this product was used in PCR amplification with 50 µL Phusion Master Mix (NEB), 5 µL of 10 µM modified Solexa© Amplification primer mix (Illumina, *Inc*.) and up to 100 µL H_2O. Phusion PCR settings followed product guidelines for a total of 18 cycles. Samples were gel purified by excising DNA fragments ranging from the 300 to 550 bp size range, and diluted to 10 nM. Sequencing was performed on one lane of an Illumina GAIIx/HiSeq2000 (Illumina, *Inc* San Diego, CA).

Short read processing and mapping

Single-end raw reads of all the genotypes were stripped off their barcodes and quality filtering was performed using the FastQC (Patel and Jain 2012) software and based on FastQC report the reads were trimmed leaving 81 bases with Phred quality score of at least 20 for mapping and downstream analysis. First, a working assembly (called unitag assembly) composed of 14,035 unitags was generated from the reads of one of the tef landraces with the highest number of reads using custom perl scripts (Floragenex, *Inc*). We then mapped the trimmed reads to the indexed genome with the Bowtie (reference) algorithm with a maximum of three nucleotide mismatches and one gap between the reads and the reference. Alignment files in SAM/BAM (Sequence Alignment Map) format were generated. Subsequently, the reads that mapped to more than one position in the reference genome and reads that did not map to the reference were filtered out from the BAM files and only reads mapped to a single physical position in the genome were used for SNP calling.

SNP calling and analysis

SAM/BAM files were further processed using SAMtools (Li *et al.* 2009), VCFtools (Danecek *et al.* 2011) and Stacks (Catchen *et al.* 2011) software for SNP calling and data summarization. Variant positions produced in a Variant Call Format (VCF); a text file format that contains meta-information lines, a header line, and then data lines containing information about a position of the variant in the genome were filtered by setting a low quality cutoff of Q20 and sites that are only biallelic with MAF > 0.05 and with 80% coverage were generated for further analyses using the VCFtools (Danecek *et al.* 2011). The SNPs were categorized based on type (as transitions and transversions) using SAMtools while population parameters were estimated from each genotype using the Populations function of the Stacks (Catchen *et al.* 2011) software.

Phylogenetic analysis

To assess the genetic and geographic relationships among the germplasm, a maximum likelihood phylogenetic tree was constructed. First, the SNP dataset in VCF format was converted into RAxML format using the PDGSpider software (Lischer and Excoffier 2012). The RAXML formatted files were used as input for the RAxML program under the general time reversible model of nucleotide evolution and the gamma model of rate variation (GTRGAMMA) to generate the maximum likelihood phylogenetic tree using 100 bootstrap replicates. The best tree was then visualized using the SplitsTree4 (Huson and Bryant 2006).

Estimating nucleotide diversity and population differentiation

Genome-wide nucleotide diversity (π) was computed using the Stacks (Catchen *et al.* 2011) software. We estimated population differentiation using mean values of Wright's F_{ST}. Populations were split into six sub-populations and coded as numbers whereby pop1 = landraces from North West, pop2 = landraces from North and North East, pop3 = landraces from central and South East, pop4 = landraces from South West, pop5 = improved varieties, and pop6 = wild *Eragrostis* species. Data analyses involving read mapping and SNP calling were performed at the Vital-IT (http://www.vital-it.ch) Center for high-performance computing of the Swiss Institute of Bioinformatics (SIB) (http://www.sib.swiss), University of Bern (www.ips.unibe.ch) and on a personal computer.

VerifyingRAD tags containing the SNPs

For verifying the RAD tags containing the SNPs, we searched the VCF file containing quality SNPs and picked a SNP and its corresponding position. We then extracted the RAD tag containing the SNP. In total, we extracted three RAD tags containing three SNPs that are present either only in the wild *Eragrostis* species or in tef. Using BLASTN 2.2.18+, we searched each of the TAGs in the tef genome. The matching scaffolds were extracted and primers were designed to amplify the TAGs containing the SNPs (S1**Table x**). We amplified each TAG using PCR and the resulting products were sequenced. The sequences in fasta format were used to make a multiple sequence alignment using the online alignment tool (Clustal Omega) from the European Molecular Biology Laboratory EMBL-EBI at (https://www.ebi.ac.uk). A maximum likelihood tree with 100 bootstrap iterations was inferred using the MEGA 7.0.16 program under the General Time Reversible Model (Nei and Kimar 2000). The phylogenetic tree was then compared with the phylogenetic tree generated by using sequences from the RAD sequencing.

Results

RAD tag sequencing enables genome-wide SNP discovery from the tef landraces

The sequencing of the SbfI library generated over 113 million single-end reads corresponding to 11 Gbp of sequences (Table 2). The number of raw reads ranged from 975,666 to 5,207,049 with over 3 million reads generated per individual germplasm. The sequencing quality of the majority of the reads was generally high with *Phred* scores above 20 for most of the reads. After trimming bad quality sequences, reads with 75 bp length were retained for the subsequent analysis. To identify SNPs genome-wide, the trimmed sequence reads were aligned to the unitag (*de novo* assembly) and the tef reference genome, which have genome lengths of 339 Mb and 642 Mb, respectively. Map files were generated, sorted and indexed as most downstream analysis tools only work with sorted and indexed map files.

Table 1. Summary of RAD tag sequencing and SNP discovery.

Category	**Number**
Summary of the RAD-seq	
Samples analyzed	45
Total number of raw Illumina sequence reads obtained	113,313,748
Sequence reads per sample (range)	975,666-5,207,049
Reads per sample (mean)	2.5 x 10^6
SNPs from the Unitag	
-Raw SNPs	11,598
-SNPs retained after quality filtering	9,024
-Biallelic sites with MAF > 0.05 and with no missing data	956
-Ts/Tv[b]	1.4
SNPs from the tef reference genome	
-Raw SNPs	81,599
-SNPs retained after quality filtering	58,735
-Biallelic sites with MAF > 0.05 and with no missing data	12,553
-Ts/Tv[b]	1.3

[b]The ratio of Ts (transitions) / Tv (transversions).

The figures were generated by mapping reads to the de novo assembled genome (Unitag assembly) and the tef reference genome. The germplasm included 1 mutant line, 39 tef landraces, 3 improved tef varieties and 2 wild Eragrostis species

A total of 11, 598 raw SNPs were identified by using the Unitag (Table 2) and 81,599 SNPs by mapping the reads to the tef reference genome. We were interested to know the difference in SNP numbers between the tef and the wild *Eragrostis* species. The number of SNPs was higher in the two wild species *E. curvula* and *E. minor* than in the tef landraces (Table S1) showing that the wild species are more diverse than cultivated tef. The smallest number of SNPs was identified from the improved variety DZ-Cr-196 (800 SNPs) and the mutant line *Kegne* (900 SNPs).

Transitions are more prevalent in the tef genome

We found that of the identified SNPs, 58.8% were transitions and 41.2% were transversions. The transitions were split with 48% (A↔G) and 52% (C↔T) while the transversions were 20% (A↔T), 24% (A↔C), 27% (G↔T), and 29% (C↔G) with transitions to transversions (Ts:Tv) ratio of 1.4 showing that transitions are more prevalent in the tef genome compared to transversions.

Genetic diversity within the tef landrace populations

In order to know the extent of genetic diversity in the germplasm panel, we computed genome-wide estimates of nucleotide diversity for each sub-population using the populations function of the Stacks (Catchen *et al.* 2011) software. We found that mean nucleotide diversity values were smallest $\pi = 0.004$ for the tef landraces followed by $\pi = 0.007$ for the improved tef varieties and $\pi = 0.021$ for the wild *Eragrostis* species.

Poor genetic differentiation in the tef landrace populations provides rationale for utilizing variation at the inter-specific level

To examine the genetic divergence between populations, we computed Wright's fixation index (F_{ST}) (Wright 1951) using the Populations function of the Stacks (Catchen *et al.* 2011) software. Since the tef landrace populations in this study are collections from diverse agro-ecological zones, we wondered if they are genetically differentiated.

Mean F_{ST} value between the landrace sub-populations was 0.002, suggesting lack of differentiation while as expected the landraces and the wild *Eragrostis* species were differentiated with mean F_{ST} values of 0.515 (Table 3). It is interesting to note that the landrace sub-populations were poorly differentiated $F_{ST} = 0.015$ from the improved varieties, and that together with the previously published result (Zhu *et al.* 2012) supports the hypothesis that the current tef improvement process (mainly based on selection from the landraces) has small effect on the global genetic make-up of the landraces. This poor genetic divergence between tef sub-populations provides rationale for utilizing variation at the inter-specific level.

Principal component analysis shows a clear separation of the wild *Eragrostis* species from the tef landraces

To generate a visual summary of the SNP dataset, we performed principal component analysis (PCA) in R. The first principal component sufficiently explained most of the total variation in the dataset (44.61%) while 7.18% of the variation was explained by the second principal component (Fig. 1). The two wild *Eragrostis* species *E. curvula* and *E. minor* cluster far away from the tef landraces and form a discrete cluster. On the other hand, the tef landraces fell into one major cluster with the mutant line and the three improved tef varieties forming a genetic continuum (the bottom four points in Fig. 1). This result shows the power of PCA analysis to detect population substructure from genome-wide SNP datasets.

Table 3. Mean pair-wise Wright's fixation index (F_{ST}) estimates.

	1	2	3	4	5	6
1	0	0.004[b]	0.006	0.003	0.014	0.515
2		0	0.003	0.003	0.016	0.365
3			0	0.002	0.014	0.366
4				0	0.015	0.358
5					0	0.285
6						0

[b] *Mean F_{ST} estimates among population pairs*

Values are given for each of the six populations. 1 = landraces from North West, 2 = landraces from North East, 3 = landraces from South East, 4 = landraces from South West, 5 = improved varieties, and 6 = wild Eragrostis species. Detailed description about each population including geographical location of the collection is presented in Table 1.

All estimates are significant at (P < 0.001) level of significance

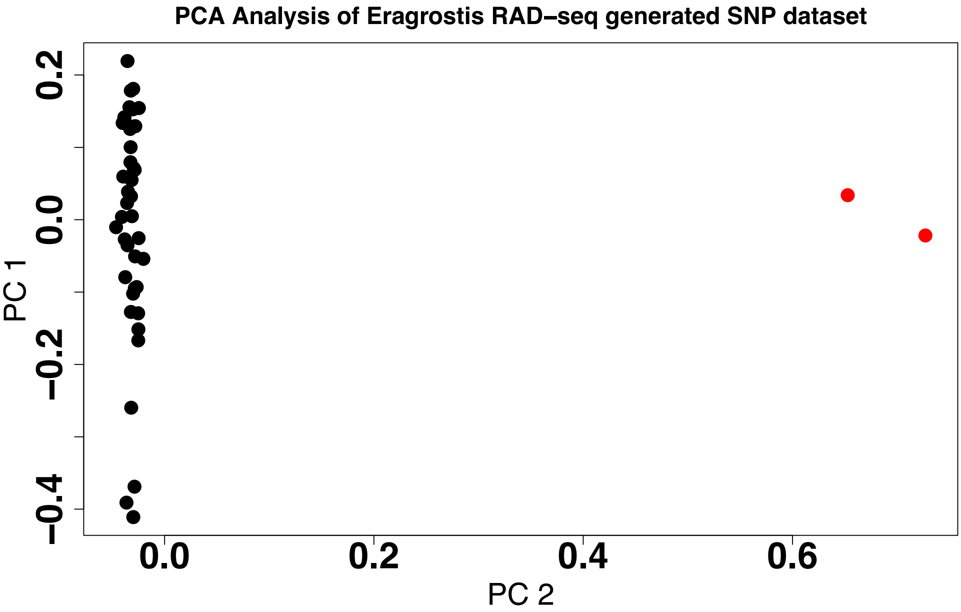

Figure 1. PCA was performed on the SNP dataset obtained from 45 individual germplasm containing no missing data

The first two principal components were plotted and clearly show the separation of the tef landraces from the two wild Eragrostis species (red dots). The bottom four points correspond to the mutant line Kegne, the three improved tef varieties (DZ-Cr-196, DZ-Cr-37, DZ-Cr-385 and) and their clustering pattern reflects their genetic similarity as a direct effect of the genetic improvement process.

Phylogenetic relationships

To assess the relationships among the individuals in the panel and visualize the inferred relationships in the form of a phylogenetic tree, a maximum likelihood tree with 100 bootstrap iterations was inferred using the RaxML program under GTRGAMMA model (Stamatakis 2014). The tree shows that the tef landraces, the improved tef varieties and the mutant line were all clustered into one big clade (clade F) (Fig. 2) suggesting genetic similarity, while the two wild species *E. curvula* and *E. minor* clustered together into a separate clade, clade G. This grouping is consistent with the results of the PCA analysis (Fig. 1). Within the clade consisting the entire tef landraces, we find pockets of clusters such as clade E, involving all the improved tef varieties and the mutant line (which was developed from DZ-Cr-37), reflecting the nature of the tef breeding process, which targets common agronomic traits. Clade A and clade D represent collections from the North West. Clade A is composed of Accession 212592 and Accession 212603 both from *Wollo* and Accession 229770 collected from *Gojam*. On the other hand, Clade D consists of Accessions 229758 (*Gojam*), Accessions 212708 (*Gondar*) and Accessions 243488) (*Wollo*). These areas are very close to each other that farmers in these areas might be using the same germplasm.

The majority of the accessions in Clade B are collection from *Gojam*, *Gondar* and *Adwa* all located in the North West. Clade C is comprised of two accessions, Accession 229984 and Accession 215356 both collections from *Bale*, South East. Clade E is composed of a mosaic of accessions collected from almost all collection sites that were targeted by this study and includes Accession 237695 (*Shewa*), Accession 234375 (*Adwa*), Accession 230771 (*Borena*), Accession 236091 (*Hadiya*), Accession 2225761a (*Omo*), Accession 237742 (*Bale*), Accession 55100 (*Harerghe*) and Accession 230771 (*Borena*) which are collected from the southern part of the country. Given the dynamic informal cereal seed system in the country, which is marked by deliberate movement and sharing of seeds by farmers between neighboring regions and beyond oftentimes, it appears difficult to completely assign a germplasm to one location. However, there are landraces typical to a region that are popular and identified by local given names that reflect their inherent properties.

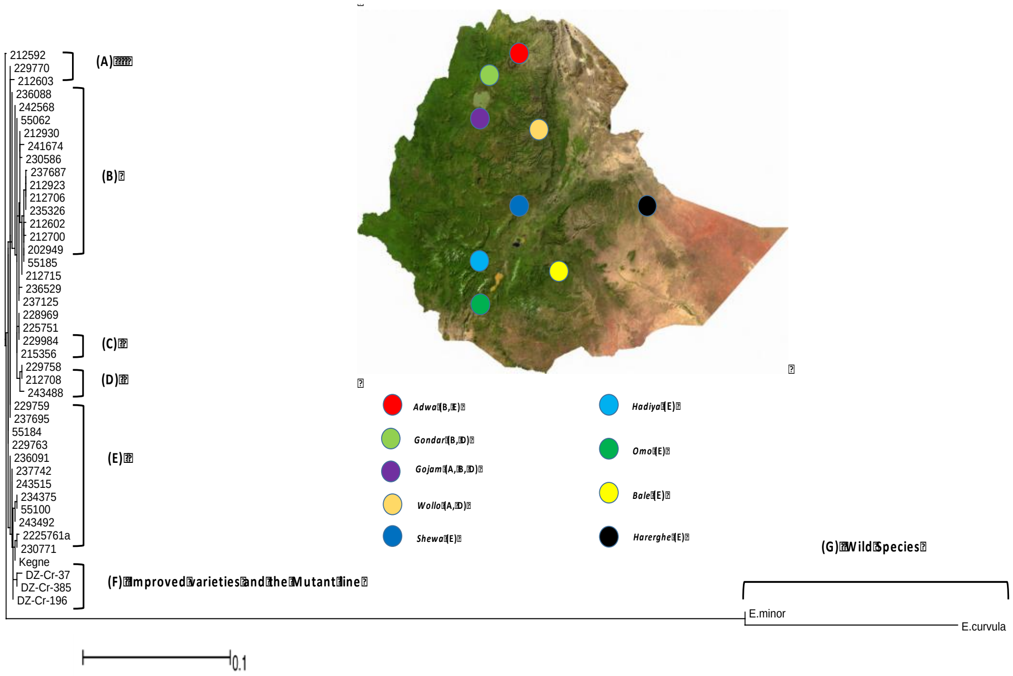

Figure 1. Phylogenetic tree inferred by using the RAxML program from the analyses of the SNP. We used the dataset that contained biallelic sites with no missing data.

The scale bar (bottom) reflects evolutionary distance, measured in units of substitutions per nucleotide site. The map shows the approximate areas of collection and the clade where representative accessions are found. The map is divided into North West (Gondar, Gojam), North East (Adwa, Wollo), South East (Harerghe and Bale) and South West (Hadiya and Omo). Source: (https://commons.wikimedia.org/wiki/Atlas_of_Ethiopia).

Sequences from the RAD tags and PCR gave similar phylogenetic trees

For verifying the RAD tags containing the SNPs, we amplified the RAD tags containing three selected SNPs using primers specifically designed for this purpose (see methods and Table S3) and the resulting products were sequenced. The phylogenetic analysis based on the maximum likelihood method with MEGA 7.0.16 software program under the General Time Reversible Model (Nei and Kimar 2000) shows all the tef genotypes as a clade (37A representing DZ-Cr-37) and 196A representing DZ-01-196) (Fig. 3) and supported the clade we see in the previous tree (Fig. 2). Moreover, the wild *Eragrostis* species grouped outside the tef clade as previously reported in similar studies (Ingram *et al.*, 2003 and Girma *et al.*, 2018). The primers developed here (S3 Table) could be used to differentially amplify regions of the genome in tef and wild *Eragrostis* species.

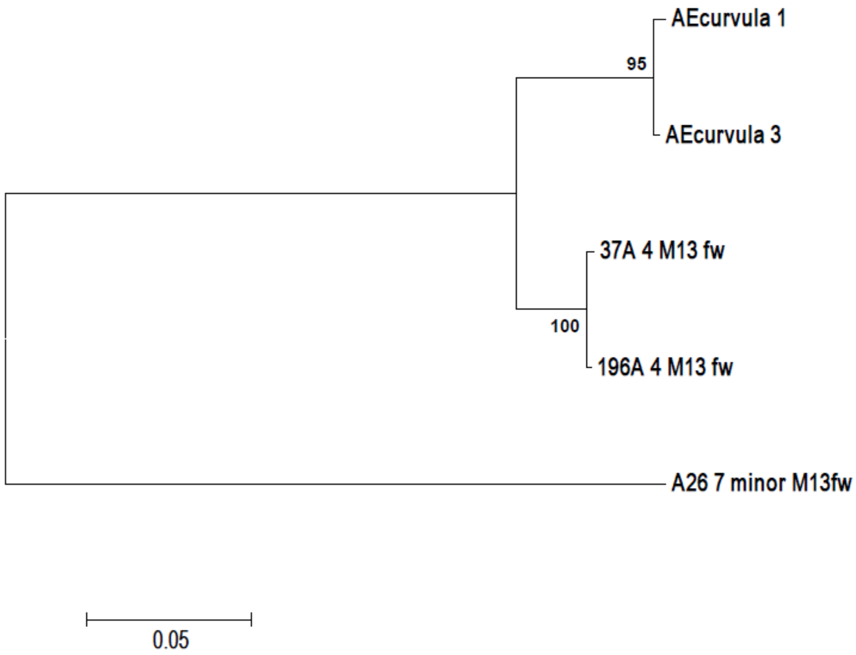

Figure 3. Maximum likelihood phylogenetic tree obtained using PCR amplified RAD tags containing SNPs.

The tree was inferred with MEGA 7.0.16 software program under the General Time Reversible Model (Nei and Kimar 2000) with 100 bootstrap iterations. The numbers at the edge of each branch are bootstrap values.

Discussion

Genomic resources for tef have started to accumulate in the last two decades. However, next generations sequencing based studies have not yet been reported. Here, we applied an NGS-based protocol called, the RAD-seq for the first time to a germplasm panel comprised of tef landraces, improved tef varieties, a mutant line and two wild *Eragrostis* species. We followed two approaches to map the sequencing reads and to discover single nucleotide polymorphisms (SNPs).

Following the *de novo* assembly approach, we have identified 9,024 SNPs. In contrast, following the read-to-reference mapping approach, we have identified 58,735 SNPs. The availability of the tef reference genome has boosted our ability to capture more variability from our germplasm panel. The phenotypic diversity of the tef landraces has been exhaustively studied in the last three decades (Assefa et al 2010). At the genomic level, however, our understanding of the genetic diversity of the tef landraces is still at its juvenile stage. We compared the number of SNPs identified in the tef landraces to that discovered in the wild *Eragrostis* species. The tef landraces had almost half the number of SNPs identified in the wild species. This suggests that the tef germplasm, which have undergone through years of selection have become more homogenous than the wild species that show higher variability. Domestication is often associated with a reduction in the genetic variation of domesticated plants compared to their wild progenitors (Doebley, Gaut, and Smith 2006). Our study is in agreement with this and with similar findings in other crops, such as those reported in sorghum (Mace *et al.* 2013), soybean (Chung *et al.* 2014; Lam *et al.* 2010), rice (Krishnan S, Waters, and Henry 2014), barley (Morrell *et al.* 2014; Zeng *et al.* 2015), sunflower (Liu and Burke 2006) and peach (Cao *et al.* 2014). To widen the narrow genetic base of the tef improvement, we propose that the wild *Eragrostis* species that could harbor novel variability deserve the attention of tef breeders.

The extent of genetic diversity in a population is often measured by nucleotide diversity (π), which is expressed as the average number of nucleotide differences per site between any two randomly chosen DNA sequences (alleles) sampled (Nei and Li 1979). The low nucleotide diversity values $\pi = 0.004$ for the tef landraces followed by $\pi = 0.007$ for the improved tef varieties compared to the wild *Eragrostis* species $\pi = 0.021$ agrees with the evidence for reduced nucleotide diversity among populations of selfing taxa such as Arabidopsis $\pi = 0.007$ (Innan *et al.* 1996) and Solanum $\pi = 0.001$ (Baudry *et al.* 2001). However, we speculate that the selfing alone may not be responsible for the low nucleotide diversity we observe in tef, and that additional factors such as the breeding process, which is based on narrow genetic base may play a role. The higher nucleotide diversity values in the wild *Eragrostis* species suggest that the wild species are more diverse and may harbor unique variability useful for use in tef breeding.

Nucleotide substitutions in the form of transitions (A↔G and C↔T) or transversions (A↔C), (A↔T), (G↔C), and (G↔T) occur during evolution (Jukes, 1987) and the rate ratios of transitions to transversions (Ts/Tv) are estimated by pairwise sequence comparison and joint likelihood analysis (Yang and Yoder 1999). We found that transitions were more prevalent (58.8%) than transversions (41.2%) in the tef genome. Such selective bias in transitions over transversions is consistent with findings in other crops such as hexaploidy wheat (Hussein et al 2018).

Owing to the nature of restriction enzymes, RAD sequencing preferentially targets orthologous sequence fragments across genomes and hence generates comparative genomic data suitable for phylogenetic analysis (Rubin, Ree, and Moreau 2012). Although relatively new to molecular systematics, the use of RAD-seq data for constructing interspecific phylogenetic trees has been demonstrated (Rubin, Ree, and Moreau 2012). Molecular phylogenetic analysis on tef landraces is scarce and the ones that we know were based on data from single genes (the nuclear waxy and the *rps16* plastid gene) assayed on five tef landraces and thirty wild *Eragrostis* species. In contrast, our analysis is based on genome-scale SNP data generated from 39 tef landraces, 3 improved varieties, a mutant line and two wild *Eragrostis* species. The phylogenetic analysis grouped the wild species; *E. curvula* and *E. minor* into one cluster and the improved varieties DZ-Cr-37, DZ-Cr-385, DZ-Cr-196 and the mutant line *Kegne* into a separate cluster but within the tef landraces cluster (Fig. 3.2). The grouping of improved varieties into one clade suggests that the tef breeding process may have targeted common genomic regions and in a narrow genetic base of selection.

The cultivation of tef is typically characterized by the diffusion and use of seeds across geographic boundaries. We captured this feature in the current phylogenetic analyses; with the tef landraces, partly showing a grouping based either on common area of collection or communality of germplasm use. For instance, Accession 212529 and Accession 212603 are collections from *Wollo* and were grouped in clade C with Accession 229770 from *Gojam*. Geographically, these areas are very close and the farming communities in that area are known to have a lot in common with a pronounced exchange of cereals seeds (source?). To evaluate the phylogenetic accuracy of the current grouping, we compared our tree with previously published tef phylogenies. We observe that our phylogenetic tree generally agrees with most of the trees that showed close intra-specific phylogenetic relationships (Assefa, Merker, and Tefera 2003a, 2003b; Assefa *et al.* 2001).

Conclusion

The present study provides a genome-wide SNP data from four germplasm groups in the genus *Eragrostis*, namely the tef landraces, improved tef varieties, mutants and the wild *Eragrostis* species. We have identified thousands of SNPs representing the first SNP data set obtained from the tef germplasm to date. We presented genome-scale evidence for the low nucleotide diversity in the tef germplasm as well as poor population differentiation between tef landraces and the improved varieties. Overall, the tef landraces show some sub-population division due to geographic distribution, but they also exhibit common distribution due to the movement and communal use of seeds. We provide, for the first time, an analysis of intra- and inter-specific phylogenetic relationships in tef and the wild *Eragrostis* species using genome-scale sequence data. However, given the scale of this study, a better understanding of the phylogenetic relationships in the genus *Eragrostis* may require the analysis of the entire wild *Eragrostis* species or the systematic investigation of the species suggested as close relatives. Considering its key role as a food security crop in Ethiopia and as a lifestyle food alternative in the West, more molecular resources need to be developed and the use of the presented dataset to inform future genomics assisted population genomics and breeding in tef is worthwhile.

Acknowledgements

This work was done as part of a PhD project and was supported by the Syngenta Foundation for Sustainable Development and the University of Bern.

Supplementary tables

Table 4.. Summary of SNP statistics. Using SAMTools, the raw SNP dataset was filtered to contain only biallelic SNPs and SNPs with Phred quality score more than 20

Accession number	Sequencing code	Raw SNPs	SNPs	SNPs Q >20
234375	KDG-11	1486	1350	1183
242568	KDG-13	1608	1451	1197
212602	KDG-14	1457	1280	1048
243492	KDG-15	1669	1483	1240
212603	KDG-16	1200	1059	874
212592	KDG-21	1354	1210	1014
235326	KDG-25	2973	2779	2001
243488	KDG-39	1473	1310	1114
243515	KDG-40	1356	1225	1043
215356	KDG-4	1776	1605	1350
229984	KDG-7	1670	1484	1229
55100	KDG-12	1603	1451	1257
237742	KDG-17	1833	1681	1389
237687	KDG-26	1535	1362	1095
230771	KDG-30	1972	1826	1423
237125	KDG-31	1365	1214	1037
230586	KDG-33	1340	1189	1037
237695	KDG-37	1483	1312	1144
229759	KDG-3	1508	1344	1144
229770	KDG-6	1598	1444	1203
55062	KDG-9	1799	1637	1340
236529	KDG-10	1519	1380	1114
55184	KDG-18	1527	1355	1114
212708	KDG-19	1456	1296	1109
212715	KDG-20	1248	1118	929
212706	KDG-22	1566	1456	1187
228969	KDG-23	1486	1330	1070
229758	KDG-24	1548	1379	1165
229763	KDG-35	1355	1202	1014
55185	KDG-36	1595	1450	1185
212700	KDG-38	1860	1728	1354
212930	KDG-1	1336	1183	976
236091	KDG-2	1580	1428	1120
212923	KDG-5	1367	1204	997
202949	KDG-8	2043	1897	1454
225751	KDG-28	1601	1440	1194
236088	KDG-29	1889	1722	1364
241674	KDG-32	1461	1327	1118
2225761a	KDG-34	1532	1389	1137
Improved tef varieties				
3774-13/*Kegne*	KDG-41	900	833	623
DZ-Cr-37 (*Tsedey*)	KDG-42	1783	1628	1303
DZ-Cr-196 (*Magna*)	KDG-43	800	753	600
DZ-Cr-385	KDG-44	3175	2646	1954
Wild *Eragrostis* species				
E. curvula	KDG-46	7610	7358	5047
E. minor	KDG-48	5304	5179	3245
Total		81599	74377	58735

Table 5. Summary of the mapping statistics. After mapping the raw reads to the tef reference genome, we filtered out both unmapped reads and reads that mapped at multiple positions and retained the reads that only mapped 1X using this specification from the SAMTools software (samtools view -b -F 4)

Accession code	Sequencing code	Raw reads	Mapped 1X	Unmapped	Mapped > 1X	Mapping rate
234375	KDG-11	1720898	1087292	36527	597079	96.76%
242568	KDG-13	2871916	1261422	75713	1534781	97.36%
212602	KDG-14	1824236	806796	46775	970665	97.44%
243492	KDG-15	2322092	1034426	66915	1220751	97.12%
212603	KDG-16	1901504	849782	45627	1006095	97.60%
212592	KDG-21	1553546	689843	39424	824279	97.46%
235326	KDG-25	3766541	1696464	107218	1962859	97.15%
243488	KDG-39	1858208	828049	45146	985013	97.57%
243515	KDG-40	975666	430956	25687	519023	97.37%
215356	KDG-4	2400909	1058535	58173	1284201	97.58%
229984	KDG-7	2735865	1211358	72793	1451714	97.34%
55100	KDG-12	2210305	976608	56742	1176955	97.43%
237742	KDG-17	3529959	1591510	102375	1836074	97.10%
237687	KDG-26	2633221	1164815	65699	1402707	97.50%
230771	KDG-30	5062996	2262378	129720	2670898	97.44%
237125	KDG-31	1689630	758996	38683	891951	97.71%
230586	KDG-33	1076786	479872	28137	568777	97.39%
237695	KDG-37	1945437	871450	42518	1031469	97.81%
229759	KDG-3	1833711	817321	44008	972382	97.60%
229770	KDG-6	2617552	1163934	70101	1383517	97.32%
55062	KDG-9	3851483	1697686	101602	2052195	97.36%
236529	KDG-10	2443998	495407	61471	1295235	97.48%
55184	KDG-18	2078340	928436	50798	1099106	97.56%
212708	KDG-19	1281127	576889	33177	671061	97.41%
212715	KDG-20	1532258	699051	36929	796278	97.59%
212706	KDG-22	5104362	2297462	127318	2679582	97.51%
228969	KDG-23	3464399	1555724	90439	1818236	97.39%
229758	KDG-24	1407508	637016	35421	735071	97.48%
229763	KDG-35	1151924	507028	27248	617648	97.63%
55185	KDG-36	3980561	1712094	102723	2165744	97.42%
212700	KDG-38	4175991	1851918	116211	2207862	97.22%
212930	KDG-1	1434017	649511	33524	750982	97.66%
236091	KDG-2	1678387	745207	41176	892004	97.55%
212923	KDG-5	1431565	617275	37944	776346	97.35%
202949	KDG-8	4992141	2250678	124016	2617447	97.52%
225751	KDG-28	1670098	748622	41652	879824	97.51%
236088	KDG-29	3582641	1632917	86592	1863132	97.58%
241674	KDG-32	1553129	700257	39360	813512	97.47%
2225761a	KDG-34	1946955	507028	27248	617648	97.63%
Improved tef varieties						
Kegne	KDG-41	1920783	869224	49469	1002090	97.42%
DZ-Cr-37	KDG-42	3606132	1622407	89940	1893785	97.51%
DZ-Cr-196	KDG-43	2637999	1189732	58974	1389293	97.76%
DZ-Cr-385	KDG-44	2482658	1105569	102577	1274512	95.87%
Wild species						
E. curvula	KDG-46	5207049	2369051	2719094	118904	47.26%
E. minor	KDG-48	2759150	357234	1782746	619170	35.39%

Table 6. Sequences of the forward and reverse primers for polymorphic SNP markers that confirmed the identity of SNPs generated using the rad sequencing and the PCR among the tef genotypes and the wild species

RAD Tag Name	Scaffold and location	Forward primer	Reverse primer	Expected size	SNP position in VCF file
RADid_0000008_depth_26	Et_scaffold11432.4062-4940.r.fasta	GAAGCCCAGGATCACGGACG	CTACTCCTCATCTTCTTCCCCATCG	819	SNP (C/A) at position 24
RADid_0000777_depth_289	Et_scaffold2807.65674-66553 Et_scaffold4798.43429-44308	CTCGACTGATTGACTGGCTCCTC	CCTCACCTCCATCAAAGTAGCTCAGG	659	SNP (A/G) at position 56
RADid_0000736_depth_349	Et_scaffold9483.78943-79822.r	CCTCAGCACCAAGACCGACG	CAACACCGCATCCTTTTCAATAAGC	879	SNP (T/C) at position 36
RADid_0003498_depth_77	Et_scaffold3099.22932-23811	AATCTCTCTTTCTGTTTCTTCGGTCG	GTTTGATGTGTGCGGTGCC	562/299	SNP (C/A) at position 36
RADid_0002273_depth_38	Et_scaffold12691.1249-2170.r Et_C7554131.r.fasta	CATCAGTGTTTCCGTCGATTCAACC	TGTAAATGAACAGGCAGGGATCAGG	163	SNP (T/G) at position 31

References

Assefa K, A Merker, and H Tefera. 2003a. 'Inter simple sequence repeat (ISSR) analysis of genetic diversity in tef [Eragrostis tef (Zucc.) Trotter]', *Hereditas*, 139: 174-83.

Assefa K, A Merker, and H Tefera. 2003b. 'Multivariate analysis of diversity of tef (Eragrostis tef (Zucc.) Trotter) germplasm from western and southern Ethiopia', *Hereditas*, 138: 228-36.

Assefa K, H Tefera, A Merker, T Kefyalew, and F Hundera. 2001. 'Variability, heritability and genetic advance in pheno-morphic and agronomic traits of tef [Eragrostis tef (Zucc.) Trotter] germplasm from eight regions of Ethiopia', *Hereditas*, 134: 103-13.

Baird NA, PD Etter, TS Atwood, MC Currey, AL Shiver, ZA Lewis, EU Selker, WA Cresko, and EA Johnson. 2008. 'Rapid SNP discovery and genetic mapping using sequenced RAD markers', *PLoS One*, 3: e3376.

Baudry E, C Kerdelhue, H Innan, and W Stephan. 2001. 'Species and recombination effects on DNA variability in the tomato genus', *Genetics*, 158: 1725-35.

Cao K, Z Zheng, L Wang, X Liu, G Zhu, W Fang, S Cheng, P Zeng, C Chen, X Wang, M Xie, X Zhong, X Wang, P Zhao, C Bian, Y Zhu, J Zhang, G Ma, C Chen, Y Li, F Hao, Y Li, G Huang, Y Li, H Li, J Guo, X Xu, and J Wang. 2014. Comparative population genomics reveals the domestication history of the peach, Prunus persica, and human influences on perennial fruit crops. *Genome Biol*, 15: 415.

Catchen JM, A Amores, P Hohenlohe, W Cresko, and JH Postlethwait. 2011. 'Stacks: building and genotyping Loci de novo from short-read sequences', *G3 (Bethesda)*, 1: 171-82.

Chua KY, CR Doyle, RJ Simpson, KJ Turner, GA Stewart, and WR Thomas. 1990. 'Isolation of cDNA coding for the major mite allergen Der p II by IgE plaque immunoassay', *Int Arch Allergy Appl Immunol*, 91: 118-23.

Chung WH, N Jeong, J Kim, WK Lee, YG Lee, SH Lee, W Yoon, JH Kim, IY Choi, HK Choi, JK Moon, N Kim, and SC Jeong. 2014. 'Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes', *DNA Res*, 21: 153-67.

Danecek P, A Auton, G Abecasis, CA Albers, E Banks, MA DePristo, RE Handsaker, G Lunter, GT Marth, ST Sherry, G McVean, R Durbin, and Group Genomes Project Analysis. 2011. 'The variant call format and VCFtools', *Bioinformatics*, 27: 2156-8.

De la Vega FM, KD Lazaruk, MD Rhodes, and MH Wenz. 2005. 'Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP Genotyping Assays and the SNPlex Genotyping System', *Mutat Res*, 573: 111-35.

Doebley JF, BS Gaut, and BD Smith. 2006. 'The molecular genetics of crop domestication', *Cell*, 127: 1309-21.

Doyle Jeff J and Elizabeth Dickson. 1987. 'Preservation of Plant Samples for DNA Restriction Endonuclease Analysis', *Taxon*, 36: 715-22.

Ellegren Hans and Nicolas Galtier. 2016. Determinants of genetic diversity. Nature Reviews Genetics volume 17, pages 422–433 (2016).

Fu Yong-Bi. 2015. Understanding crop genetic diversity under modern plant breeding Theor Appl Genet. 2015; 128(11): 2131–2142. doi: 10.1007/s00122-015-2585-y

Gunderson KL, FJ Steemers, G Lee, LG Mendoza, and MS Chee. 2005. 'A genome-wide scalable SNP genotyping assay using microarray technology', *Nat Genet*, 37: 549-54.

Hsia AP, TJ Wen, HD Chen, Z Liu, MD. Yandeau-Nelson, Y. Wei, L. Guo, and P. S. Schnable. 2005. 'Temperature gradient capillary electrophoresis (TGCE)--a tool for the high-throughput discovery and mapping of SNPs and IDPs', *Theor Appl Genet*, 111: 218-25.

Hussain M, MA Iqbal, BJ Till, M-u-Rahman .2018. Identification of induced mutations in hexaploid wheat genome using exome capture assay. PLoS ONE 13(8): e0201918. https://doi.org/10.1371/journal.pone.0201918

Huson DH and D Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23: 254-67.

Innan H, F Tajima, R Terauchi, and NT Miyashita. 1996. 'Intragenic recombination in the Adh locus of the wild plant Arabidopsis thaliana', *Genetics*, 143: 1761-70.

Krishnan S, Gopala, Daniel LE Waters, and Robert Henry. 2014. Australian Wild Rice Reveals Pre-Domestication Origin of Polymorphism Deserts in Rice Genome, *PLoS One*, 9: e98843.

Lam HM, X Xu, X Liu, W Chen, G Yang, F L Wong, MW Li, W He, N Qin, B Wang, J Li, M Jian, J Wang, G Shao, J Wang, SS Sun, and G Zhang. 2010. 'Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection', *Nat Genet*, 42: 1053-9.

Li H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, and Subgroup Genome Project Data Processing. 2009. 'The Sequence Alignment/Map format and SAMtools', *Bioinformatics*, 25: 2078-9.

Lischer HE and L Excoffier. 2012. 'PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs', *Bioinformatics*, 28: 298-9.

Liu A and JM Burke. 2006. 'Patterns of nucleotide diversity in wild and cultivated sunflower', *Genetics*, 173: 321-30.

Mace ES, S Tai, EK Gilding, YLi, PJ Prentis, L Bian, BC Campbell, W Hu, DJ Innes, X Han, A Cruickshank, C Dai, C Frere, H Zhang, CH Hunt, X Wang, T Shatte, M Wang, ZSu, JLi, X Lin, ID Godwin, DR Jordan, and J Wang. 2013. 'Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum', *Nat Commun*, 4: 2320.

McCallum CM, L Comai, EA Greene, and S Henikoff. 2000. 'Targeting induced local lesions IN genomes (TILLING) for plant functional genomics', *Plant Physiol*, 123: 439-42.

Morrell PL, AM Gonzales, KK Meyer, and MT Clegg. 2014. 'Resequencing data indicate a modest effect of domestication on diversity in barley: a cultigen with multiple origins', *J Hered*, 105: 253-64.

Nei and Kumar 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nei M and WH Li. 1979. 'Mathematical model for studying genetic variation in terms of restriction endonucleases', *Proc Natl Acad Sci U S A*, 76: 5269-73.

Patel RK and M Jain. 2012. 'NGS QC Toolkit: a toolkit for quality control of next generation sequencing data', *PLoS One*, 7: e30619.

Rauf S, JT da Silva, AA Khan, A Naveed. 2010. Consequences of plant breeding on genetic diversity. Int J Plant Breed. 41:1–21.

Rubin BE, RH Ree, and CS Moreau. 2012. 'Inferring phylogenies from RAD sequence data', *PLoS One*, 7: e33394.

Srivastava, K Subodh., Pawel Wolinski, and Andy Pereira. 2014. 'A Strategy for Genome-Wide Identification of Gene Based Polymorphisms in Rice Reveals Non-Synonymous Variation and Functional Genotypic Markers', *PLoS One*, 9: e105335.

Stamatakis, Alexandros. 2014. 'RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies', *Bioinformatics*.

Syvanen AC, K Aalto-Setala, L Harju, K Kontula, and H Soderlund. 1990. 'A primer-guided nucleotide incorporation assay in the genotyping of apolipoprotein E', *Genomics*, 8: 684-92.

Wright S. 1951. 'The Genetical Structure of Populations.', *Ann Eugenic*, 15: 323-54.

Yang Z and AD. Yoder. 1999. 'Estimation of the transition/transversion rate bias and species sampling', *J Mol Evol*, 48: 274-83.

Zeng X, H Long, Z Wang, S Zhao, Y Tang, Z Huang, Y Wang, Q Xu, L Mao, G Deng, X Yao, X Li, L Bai, H Yuan, Z Pan, R Liu, X Chen, Q WangMu, M Chen, L Yu, J Liang, D DunZhu, Y Zheng, S Yu, Z LuoBu, X Guang, J Li, C Deng, W Hu, C Chen, X TaBa, L Gao, X Lv, YB. Abu, X Fang, E Nevo, M Yu, J Wang, and N Tashi. 2015. 'The draft genome of Tibetan hulless barley reveals adaptive patterns to the high stressful Tibetan Plateau', *Proc Natl Acad Sci U S A*, 112: 1095-100.

Zhu Q, SM Smith, M Ayele, L Yang, A Jogi, SR Chaluvadi, and J L Bennetzen. 2012. 'High-throughput discovery of mutations in tef semi-dwarfing genes by next-generation sequencing analysis', *Genetics*, 192: 819-29.