ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2018) **16**(4); 296–305
doi:10.1017/S1479262117000314

Genetic diversity within African tomato using next generation sequencing

Grace W. Mungai[1], Willis Owino[1]*, Jane Ambuko[2], J. J. Giovannoni[3], A. B. Nyende[1] and G. Michuki[4]

[1]*Jomo Kenyatta University of Agriculture and Technology, P. O. Box 62000–00200, Nairobi, Kenya,* [2]*University of Nairobi, P.O. Box 29053-00625 Nairobi, Kenya,* [3]*Boyce Thompson Institute of Plant Research, Cornell University, 533 Tower Road Ithaca, New York14853, USA and* [4]*The Africa Genomics Centre and Consultancy Ltd P.O.BOX 381-00517 Nairobi, Kenya*

Received 7 March 2017; Accepted 27 October 2017 – First published online 15 March 2018

Abstract

Full potential of African tomato has not been tapped due to lack of information regarding its characterization. The aim of this work was to study the diversity of 17 African tomato landraces collected from *Solanaceae* gene bank – Tanzania. Evaluation was done using Complete Random Block Design. Morphological data collected were subjected to GenStat's and Darwin6 software. RNA was extracted from leaf samples, fruits at three ripening stages using modified Trizol method and sequencing done using Illumina sequencing platform. The raw reads were filtered and analysed using the Bioinformatics tools. Phenotypically, the landraces clustered into three clusters dendrogram representation. Clustering was attributed by phenotypic variation. Analysis of variance showed significant phenotypic variations among the landraces ($P < 0.05$). A total of 115,965 validated single nucleotide polymorphisms (SNPs) were mined from the 303,754,051 high-quality filtered reads. Molecular characterization showed significant variation within the landraces at fruit development stages. Unlike the phenotypic variation, phylogenetic tree representation grouped the 17 landraces according to their geographical location with some landraces from different countries grouping together. The findings of this study reveal significant morphological variation among African tomato contributed by plant height, leaf blade length, leaf blade width and fruit width. Positive correlation between fruit width and yield ($r = 0.93$, $P < 0.01$) was observed. Results of this study reveal that there is admixture of landraces from various geographical locations. Morphological characterization of African tomato can only lay a foundation but it does not reveal genetic diversity. The transcriptome SNP analysis revealed significant variation among the African tomato according to their geographical location.

Keywords: biodiversity, Illumina, morphological, phylogenetic

Introduction

The African tomato is an important fruit and vegetable crop. It is widely used in salads as well as for culinary purposes. The fruit contains significant amounts of vitamin A and C, lycopene, β-carotene, magnesium, iron, phosphorus, potassium, riboflavin, niacin, sodium and thiamine with antioxidant properties and potential beneficial health effects (Zhang *et al.*, 2010).

The African tomato exhibits broad diversity and can be produced in peripheral areas. The identification of variability among landraces is essential to the maintenance and utilization of germplasm resources (Shirasawa *et al.*, 2013). The study and evaluation of germplasm is importance for current, future agronomic and genetic improvement of the tomato crop (Reddy *et al.*, 2013). Morphological and molecular markers can be used to identify and estimate the genetic diversity of plants.

*Corresponding author. E-mail: willis@agr.jkuat.ac.ke

Table 1. African tomato landraces used in this study

Code No.	Acc. No.	Country of origin	Suggested naming of the samples inclusive of two phenotypes
1At	V1005987	Morocco	1A_V1005987_Mor_kid_red
2at	V1006833	Ethiopia	2at_V1006833_Eth_obl_red
4at	V1005872	Morocco	4at_V1005872_Kid_red
5at	VI005878	Morocco	5at_V1005878_Mor_ro_red
6at	RV102114	Tanzania	6at_Rv102114_Tanz_ro_red
7at	V1007108	S. Africa	7at_V1007108_S.Afr_obl_red
8at	Tindi 050580	Kenya	8at_Tindi_050580_Ken_ro_yel
9at	RV102112	Madagascar	9at_Rv102112_Mad_ov_pin
10at	Tindi 050589	Kenya	10at_Tindi_050589_kenya_round_yellow
11at	V1006838	Ethiopia	11at_V1006838_Ethiopia_round_red
12at	V1006842	Ethiopia	12at_V1006842_Eth_ro_yel
13at	V1006826	Morocco	13at_v1006826_Mor_kid_red
15at	V1005874	Ethiopia	15at_v1005874_Eth_ro_red
16at	V1030380	Mauritius	16at_v1030380_Mau_ov_red
17at	V1006892	S. Africa	17at_v1006892_S.Afr_ov_pink
18at	V1035028	S. Africa	18at_v1035028_S.Africa_ro_red
19at	V1005875	Morocco	19at_v1005875_Mor_ro_red

In Africa, tomato production has been greatly affected by biotic and abiotic stresses. However, African tomato species which are adapted to harsh growing conditions may possess genes for adaptation to these biotic and abiotic stresses. In Africa, there are diverse *Solanaceae* species whose phenotypic and genotypic traits are neither characterized nor documented. Lack of molecular markers that detect differences between best breeding lines of tomato has prevented a detailed study of most qualities of economic importance within genetic backgrounds that are relevant to plant breeders, growers and processors. Morphological and genetic diversity of the African tomato is therefore required to progress the genetic resource base for future crop improvement programmes. Single nucleotide polymorphisms (SNPs) can be mined from sequence data to characterize allelic variation, genome-wide mapping and for marker-assisted selection (Yang *et al.*, 2004). This study was done to characterize the African tomato landraces using morphological and molecular biology tools.

Materials and methods

Morphological characterization

Sampling sites and sampling

Seventeen African tomato landraces were collected from the *Solanaceae* gene (Table 1) conservation stations at the Asian Vegetable Center, Regional Center for Africa (AVRDC- RCA), and Arusha, Tanzania.

Viability check and pre-germination

Ten seeds of each landrace were planted on petri-dishes with wet paper for 10 d ensuring the paper did not dry out during this time. All African landraces with 70% and above germination rates were used for pre-germination on trays containing peat moss media in the greenhouses at the Jomo Kenyatta University (JKUAT) Institute of Biotechnology Research laboratory (IBR).

After 4 weeks, germinated seeds were transplanted in potting bags containing well-mixed forest soil, and manure in the ratio of 3:1 and placed in the IBR greenhouse.

Experimental design and layout

Complete Random Block Design (CRBD) was used to set up an evaluation plot in an open field at the JKUAT farm. The 17 landraces were sown in three blocks each containing three plots. Six replicates of each landrace were grown in each plot but data were collected of six samples from each block such that each landrace had a total of 18 plants, with a total of 306 plants from all 17 African Tomato landraces.

Phenotypic characterization

Measurements and observations were taken from six tagged individual plants selected from 18 plants of each landrace. Phenotyping was carried out using nine quantitative and nine qualitative traits to estimate the levels of variation among the African tomato landraces. Vegetative data were collected when 50% of the plants had flowered

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

(Table 2), while the fruit data were collected at mature green, mature breaker and mature red stages.

Data analyses

Analysis of variance (ANOVA) was carried out to determine genetic diversity of the measured nine quantitative traits. Means for each trait were separated by the least significant difference at ($P < 0.05$). Phenotypic correlation coefficients were computed to examine the degree of association among the quantitative traits. Multivariate ANOVA was conducted to reveal the patterns of phenotypic diversity of quantitative traits studied. Means of each quantitative character were standardized before subjecting to principal component analysis (PCA) as was suggested by Reddy *et al.* (2009). The standardized data of nine quantitative traits were then used as an input for the PCA biplot loading and cluster analysis. An agglomerative, hierarchical cluster classification technique with average linkage strategy was performed. Statistical analysis was done using GenStat Discovery, Edition 4.

Molecular characterization

Sample collection

Seeds from the 17 African tomato landraces were planted at the Boyce Thompson Institute for plant research at Cornell University, USA. Leaf samples after the 3 weeks and fruit samples at the three fruiting stages (mature green, mature breaker and mature red/yellow) were used.

RNA extraction

Leaf and fruit sample were collected using sterile forceps and immediately kept in well-labelled falcon tubes containing liquid nitrogen. This was followed by RNA extraction using the modified Trizol method by Kumar *et al.* (2011). The leaf and fruit RNA was quantified using a nanodrop and its quality checked by viewing the gel through the UV light.

Library construction

RNA extraction was followed by library construction using the modified protocol by Zhong *et al.*, (2011). The library quality and quantity was checked by viewing through the UV light and cubit equipment, respectively. The libraries were later multiplexed using different barcodes to make a lane.

Sequencing

The multiplexed libraries were then sent to the Biotechnology Resource Centre (BRC) at the Cornell University in the USA for sequencing using the Illumina platform.

Phenotypic data collection

Data were collected at both vegetative and reproduction stages. Vegetative data included stem colour, petiole colour, leaf base shape, leaf colour, plant growth habit, height of the plant at 50% fruiting, and at reproductive stage included fruit shape, colour and texture.

Computational data analyses

Differential gene expression. Filtering of the primer, adaptor, ribosomal RNA (rRNA) was done using the next generation sequencing (NGSQ) tool kit to have filtered high-quality reads. Ribosomal contamination was filtered from the high-quality RNA –seq reads using Ribopicker v 0.4.3 (Wang *et al.*, 2010; Lee *et al.*, 2014). The non-rRNA Fastq sequences were used for differential gene expression. TopHat software was used to align the non-rRNA to *Solanum lycopersicum* SL2.5.,0 genome from Ensembl Genomes, for initial assembly, this is because Tophat identifies splice junctions and handles assembly of reads to reference genome even where big gaps (introns) are present. This is important for gene expression in coding regions only. Splice junctions occur between an intron and exon, it was also used to convert the non-rRNA sequence to a BAM file, the Cufflink was used to assemble the transcripts followed by the Cuff Compare, which compared two or more transcripts, the compared transcripts were merged using the Cuff Merge and Cuffdiff was used for differential gene expressions. The Cummerbund was used to plot the abundance of the differential genes expressed while the R Studio was used to visualize the graphs charts and tables (Fig. 1).

SNPs mining

Separate and adapter/barcode trimmed sequences from Illumina were checked for quality using NGS tool kit (Lee *et al.*, 2014) and high-quality reads filtered. Ribosomal contamination was filtered from the high-quality RNA – seq reads using Ribopicker v 0.4.3 (Nielsen *et al.*, 2011; Lee *et al.*, 2014). The reads were aligned to the *S. lycopersicum* SL2.50 genome from Ensembl Genomes using STARv 2.3.0 (Nielsen *et al.*, 2011; Lee *et al.*, 2014) using default settings. STAR v 2.3.0 was used because nucleotide polymorphism mining was done from both exons and introns yielding to a SAM alignment file, to the SAM file, read groups were added, duplicate reads removed, reads sorted by coordinates and the file converted to BAM file and indexed using Picard – tools v2.1.1 (Nielsen *et al.*, 2011; Lee *et al.*, 2014). The genome analysis tool kit unified genotyper v2.8-1 (GATK) was used to call SNPs in all the samples, resulting in a multi-sample variant call format (VCF) file (Nielsen *et al.*, 2011; Lee *et al.*, 2014). Default parameters were used for SNP calling in GATK with HaplotypeCaller set at phred-scaled confidence threshold of 20. Annotation and prediction of effects and variants on genes in the VCF file was done using

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

Table 2. Morphological characterization at vegetative and reproductive stages at the JKUAT greenhouse

	Landrace	PH	Growth habit	ST	LBL	LBW	LVC	LBC	PC	LBS	FS	FC	FL	FW	FT
1	V10050580	40	determ	L.P	6.3	3.6	Purple	Green	Purple	Asy	Round	Yellow	2.4	6.3	Smooth
2	TINDI/050589	52	determ	P	7	2.9	Purple	Green	G + P	Asy	Round	Yellow	3.2	6.3	Smooth
3	V1006826	66.2	Determ	Green	16.7	6.4	L.Green	Green	G + P	Asy	Kidney	Red	9.5	22.7	Ridged
4	RV102112	68.4	Determ	P + G	13.4	7.3	L.Purple	Green	G + P	Asy	Oval	Pink	6.2	11.1	Smooth
5	V1006833	74.2	indeterm	Green	12.6	4.9	Green	Green	G + P	Sym	Oblong	Red	8.2	11.8	Ridged
6	V10035028	85.4	indeterm	Green	9.7	5.4	Green	Green	G + P	Asy	Round	Red	7.6	15.2	Ridged
7	V1005878	93	indeterm	Green	9.1	5.7	Green	Green	G + P	Asy	Round	Red	8.8	18.1	Ridged
8	V1005872	94.8	indeterm	Green	13.8	7.8	G + P	D.Green	G + P	Sym	Kidney	Red	10	19.9	Ridged
9	V1006892	95.6	indeterm	Green	10.6	6.4	L.Green	L.Green	G + P	Asy	Oval	Pink	4.6	8.3	Smooth
10	LO5942	96.2	indeterm	Green	14.2	6.9	L.Green	Green	G + P	Asy	Oval	Red	6.7	12.7	Smooth
11	V1006838	96.4	indeterm	Green	11.9	6.8	Green	L.Green	G + P	Sym	Round	Red	5.3	12	Ridged
12	V1005987	96.8	indeterm	Green	9.8	6	L.Green	Green	G + P	Asy	Kidney	Red	10.5	21.1	Ridged
13	RV02114	100.6	indeterm	P + G	10.6	5.7	L.Green	Green	G + P	Asy	Round	Red	5	10.2	Smooth
14	V1005875	106.4	indeterm	P + G	12.3	6.5	L.Green	Green	G + P	Asy	Round	Red	8.1	18.3	Ridged
15	V1006842	106.6	indeterm	Green	10.5	6.6	L.Green	Green	Green	Asy	Round	Yellow	9.1	19.7	Smooth
16	V1006828	109	indeterm	Green	10.6	6.5	Green	Green	G + P	Asy	Round	Red	5.7	10.5	Smooth
17	V1007108	120.2	indeterm	Green	9	4.5	L.Green	Green	G + P	Asy	Oblong	Red	8.8	12.3	Ridged

PH, plant height; determ, determinate; indeterm, indeterminate; ST, stem colour; LBL, leaf blade length; LW, leaf width; LVC, leaf vein colour; LBC, leaf blade colour; PC, petiole colour; LBS, leaf base shape; FS, fruit shape; FC, fruit colour; FL, fruit length; FW, fruit width; FT, fruit texture; FY, fruit yield; G + P, green and purple stripes; ASY, asymmetrical; Sym, symmetrical. Significant differences were observed in the various morphological characteristics evaluated at the vegetative and reproductive stages.

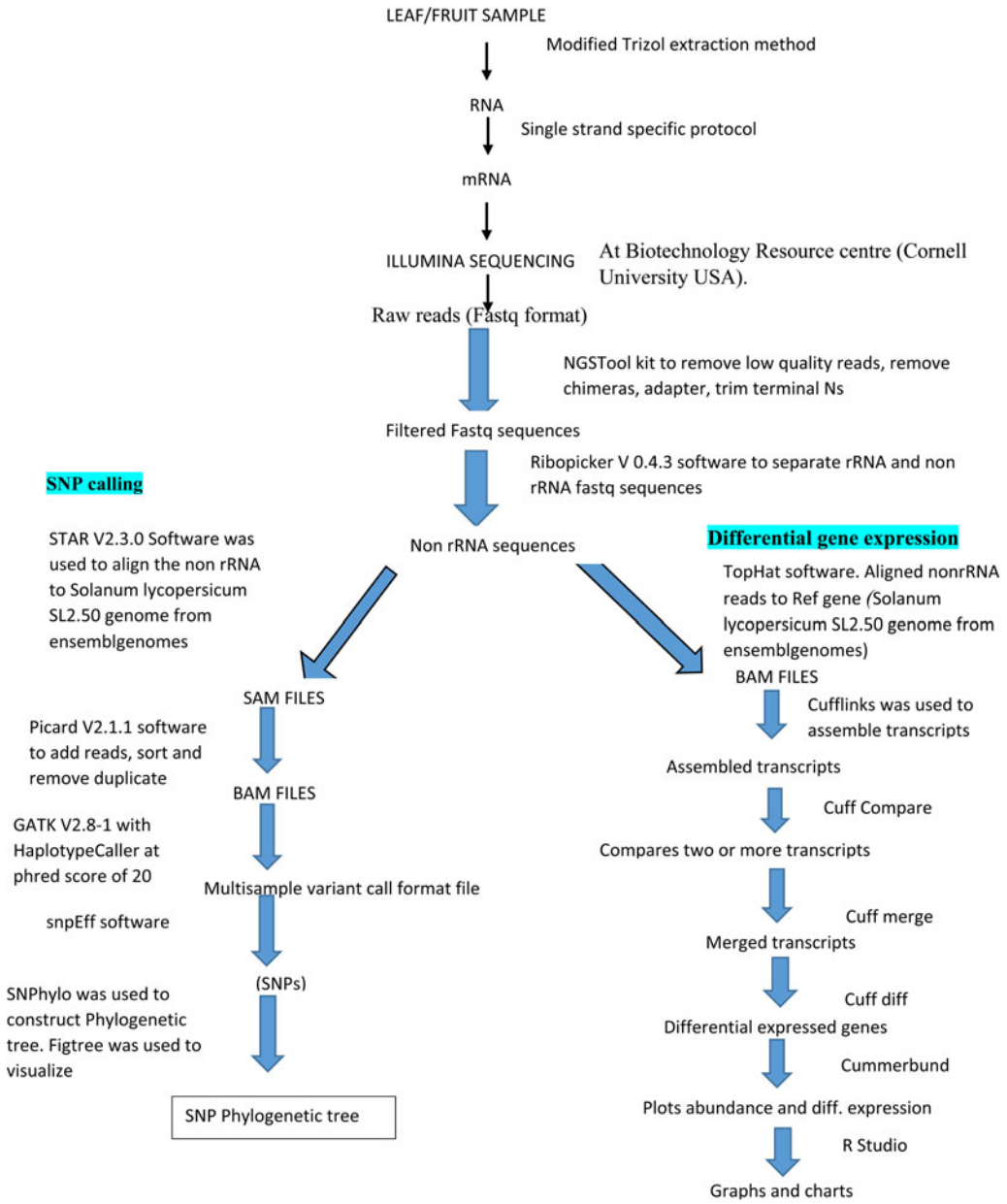

Fig. 1. Workflow showing SNP mining steps and differential gene expression.

snpEff and SNP phylogenetic tree constructed with SNPhyo (Wang *et al.*, 2010; Dewey, 2011; Lee *et al.*, 2014). The generated tree was visualized using Figtree (Fig. 1).

Results

Phenotypic variation

Stem colour: 12 landraces had green stems, while three, one and one landraces had purple and green colour, purple and light purple colours, respectively (Table 2).

Plant growth habit and leaf colour

Variation was observed on the plant growth habit with 13 accessions having the indeterminate growth habit and only four accessions having determinate growth habit. Leaf colour varied from green to light green and dark green (Table 2).

Petiole colour

Variation was observed in the African tomato with most landraces (16) showing the presence of anthocyanin (Table 2).

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

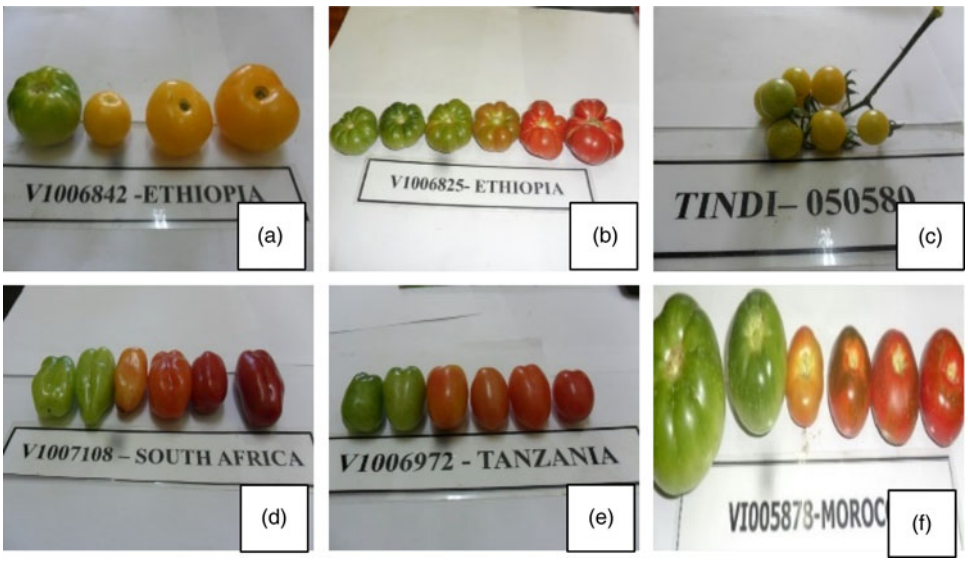

Fig. 2. African tomato was either round (nine landraces – a, b, c, f), oval (three – e), oblong (two – d) or kidney (three – b) in shape. The surface was either smooth (eight – a, c, d, e, f) or ridged (nine – b, d). African tomato was either red (12 – b, d, f,) (accessions), pink (two – e) or yellow (three – a, c) in colour when ripe.

Leaf margins, leaf vein leaf lobbing margins and leaf base shape

The African tomato showed either asymmetrical leaf base shape (14 landraces) or heart leaf base shape (three landraces) at the point where the stalk meets the leaf (Table 2).

Fruit shape: The African tomato exhibited diversity in shapes including – three landraces showed kidney shape, two landraces showed oblong shape, three had oval shape while most landraces (nine) had a round shape (Fig. 2 and Table 2).

Fruit colour: African tomato had red colour (12). Two landraces had pink, while two landraces were yellow in colour (Fig. 2 and Table 2).

Fruit surface texture

African tomato had either a ridged (nine landraces) or smooth (eight landraces) fruit surface (Fig. 2 and Table 2).

Fractional analysis presentation of the African tomato

The fraction analysis shows that the African tomato is wide diverse. However, overlapping of the landraces was observed showing that some landraces are closely associated, for instance, landraces V1005878 and L05942 and V1006833 and RV102112. Other landraces clustered far from others showing high divergence, while Tindi 050589 and Tindi 050590 are closely located but far from other landraces (online Supplementary Fig. S1).

Dendrogram presentation

The 17 African landraces grouped into three major clusters with many sub clusters using morphological data (online Supplementary Fig. S2). Cluster 1 had four accessions, cluster 2 had two accessions while cluster 3 had 11 accessions. Main contributing factor in cluster 1 was green and purple stripes petiole colour and same green leaf blade colour; with all the four accessions having green colour.

Cluster 2 had landraces that had similar phenotypes, i.e. Tindi 050589 and Tindi 050590 which were of cherry type grouped closely using morphological data. These two landraces had same fruit size (cherry), same fruit colour (yellow), same fruit shape (round) same fruit surface texture (smooth) and were both indeterminate.

Cluster 3 comprised of landraces which were closely related in leaf blade colour, stem colour and fruit texture. Examples of these landraces included V1005872, V1005987, V1006842 and V1005874.

PCA of the quantitative traits

The first seven principal components (PC1, PC 2, PC3, PC4, PC5, PC6 and PC7) analysed covered 68.51% variation within the 14 dimensions generated (online Supplementary Table S1). The quantitative traits that contribute more to PC1 were fruit length, width internode height and leaf length accounting for 20.89% of the total variation. The PC2 accounted for 11.63% of the total variation due to fruit length and fruit mass; PC3, PC4, PC5 and PC6 accounted for 9.54, 7.92, 6.75, 6.16% of total variation due

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

petal number, plant height, petal number and plant width, respectively; PC7 accounted for 5.52% due to leaf length, leaf width, plant height and plant width (online Supplementary Table S1).

Simple matrix correlation of the phenotypic traits

There was significant positive correlation between leaf blade length ($r = 0.72$, $P < 0.01$), leaf width and fruit yield and plant height and plant width ($r = 0.446$, $P < 0.01$); significant positive correlation was also observed in leaf blade colour and petiole colour ($r = 0.48$, $P < 0.01$). Other parameters which showed significant positive correlation included fruit colour and fruit shoulder colour ($r = 0.761$, $P < 0.01$), fruit length and fruit width ($r = 0.64$, $P < 0.01$), fruit length and fruit yield ($r = 0.65$, $P < 0.01$) and between fruit width and fruit yield ($r = 0.93$, $P < 0.01$) (online Supplementary Table S2).

There was a negative correlation observed between stem colour and internode colour ($r = -0.34$, $P < 0.01$), stem colour and fruit yield ($r = -0.28$, $P < 0.01$); fruit texture had a negative correlation with fruit yield ($r = -0.37$, $P < 0.01$); others with negative correlation included leaf blade length and fruit shoulder colour, leaf length and petiole numbers, internode height and fruit texture, fruit width and fruit texture and fruit yield at $r = -0.26$, -0.27, -0.29, -0.39, $P < 0.01$, respectively (see online Supplementary Table S2).

Morphological characterization using qualitative traits

Phenotypic diversity for individual qualitative traits revealed a high degree of variation among the studied landraces (see online Supplementary Table S3) using the Shannon–Weaver diversity index to estimate phenotypic diversity of eight qualitative traits studied.

The highest phenotypic diversity index (H') for traits studied recorded was 0.99 in petiole colour, stem colour and vein colour with a total mean phenotypic diversity index of 7.89. Substantial variation was observed in stem colour and vein colour (see online Supplementary Table S3).

Qualitative morphological parameters showed a close relationship between the 17 landraces. Morphological features used in the delimitation of the accessions were the presence or absence of ridge on the fruit, fruit shape and colour, leaf orientation and general fruit morphology. Fruit morphology is the major qualitative character used in the identification of the selected African tomatoes

Molecular characterization

Raw reads were received from the BRC in FASTQ format (see online Supplementary Fig. S3). A procedure to identify SNPs diversity included pre-processing the sequence data and filtering low-quality bases, mapping reads to the Tomato reference genome and post-processing of the alignment results in order to find the effect of variation.

Pre-, post-processing and alignment

Initially the sequencing quality was scrutinized using FastQC tool. NGSQC toolkit was used to filter the low-quality reads and discard the primer/adapter contaminated reads with default parameters (see online Supplementary Fig. S3) according to Patel and Jain (2012). After filtering based on the quality score, 90.8 million reads before fruiting, 91.6 million reads at mature green stage, 84.2 million reads at mature breaker and 82.4 million reads mature red were retained and used for further analysis. Short sequencing reads were mapped to the annotated tomato reference genome (*S. lycopersicum* GCF-000188115.3_SL2.5.0) using TopHat with the default parameters. Properly mapped reads were separated from the unmapped reads using Filter SAM by setting the flag values in SAMtool. Among the 90.8 M reads in before fruiting stage, about 80.89–94.87% reads were properly mapped to the tomato genome. In the 91.6 M reads in mature green stage, 74.9–94.3% mapped to the tomato genome. In mature breaker, 73.64–94.64% of the 84.2 M reads mapped to the tomato genome, while 73.59–94.64% of the 82.4 M mapped to the tomato genome (Pabinger *et al.*, 2013).

Variant calling

SNP calling and annotation. A total of 115,965 SNPs and 689 multiallelic SNPs were mined from all the 17 African tomato landraces used in this study (Table 3). The annotation was performed based on genomic location and the SNPs and were distributed in exonic and splicing region.

Analysis of differentially expressed genes. A total of 140,909 differentially expressed genes were mined from the 17 landraces used in this study, 4000 genes were differentially expressed in V1005987, 4640 from V1006833, 2787 from V1005872, 7065 from V1005878, 3586 from RV102114, 10,161 from V1007108, 9269 from Tindi 050580, 7125 from RV102112, 7374 from Tindi 050589, 13,028 from V1006838, 11,854 from V1006842, 11,515 from V1006826, 11,033 from V1005874, 8275 from V1030380, 9538 from V1006892, 8513 from V10035028 and 11,146 from V1005875 (see online Supplementary Table S4).

A total of 115,965 SNPs were discovered in the 17 landraces. These SNPS were discovered across all the 12 chromosomes as a result of insertions and deletions (Table 3).

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

Table 3. Total number of SNPs at mined from all the 17 African tomato landraces

					First ALT					
Chromosome no.	Indels	TS	TV	TS/TV	TS	TV	TS/TV	SNPs	No. of sites	Mult-allelic SNPs
1	2949	9590	6393	1.50	9563	6341	1.51	15,983		
2	1529	6352	4380	1.45	6331	4348	1.46	10,679	265	51
3	1798	6605	4485	1.47	6589	4440	1.48	11,029	405	59
4	1588	6808	4555	1.49	6783	4458	1.52	11,241	422	120
5	1022	3926	2685	1.46	3906	2637	1.48	6543	277	68
6	1613	6704	4440	1.51	6688	4411	1.52	11,099	340	42
7	1200	3552	2533	1.40	3535	2498	1.42	6033	310	51
8	1114	3944	2718	1.45	3929	2681	1.47	6610	251	51
9	1427	6889	4595	1.46	6665	4541	1.47	11,206	307	76
10	1106	4192	2914	1.44	4176	2864	1.46	7040	288	65
11	1544	7356	4691	1.57	7342	4632	1.59	11,974	338	72
12	1036	3855	2707	1.42	3847	2681	1.43	6528	243	34
Totals								115,965		689

Fig. 3. Phylogenetic tree showing diversity in the 17 landraces as a result of variation in gene expressions at four different fruiting stages.

Phylogenetic tree representation of the SNPs showed the 17 clustered according to their geographical locations (Fig. 3 and see online Supplementary Fig. S4). However, some landraces from different geographical regions clustered closely. For instance, V1030380, an oval red landrace from Mauritius, grouped closely with RV102112, an oval pink landrace from Madagascar, and Tindi 050580, a round yellow Kenyan landrace. V1005872, a kidney-shaped red landrace from Morocco, grouped closely to an oblong red Ethiopian V1006833. V1007108, an oblong-shaped red landrace from South Africa, grouped closely to Tindi 050580, a round yellow Kenyan landrace, and V1006838, a round red Ethiopian landrace.

Discussion

The results of the clustering analysis using the Darwin's 6 software showed that branching occurred at a very low phenon line, which suggests a broad to overall similarities among all landraces; this can be attributed to hybridization and ability of tomato to self-pollinate (Lawal *et al.*, 2015).

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

In this study, the first seven principal components (PC1, PC 2, PC3, PC4, PC5, PC6 and PC7) analysed covered 68.51% variation within the 14 dimensions generated (see online Supplementary Table S1). According to Chatfield and Collins (1980), components with an eigenvalue of <1 should be eliminated so that fewer components be dealt with; moreover, eigenvalues >1 are considered significant.

Strong positive correlations were observed in this study using the simple correlation matrix (see online Supplementary Table S2); similar results were obtained by previous findings of Kisua *et al.* (2015). Strong positive correlation on yield, leaf width and plant diameter would contribute to the quantity of food synthesized by the plant during photosynthesis, the plant width could serve well in H_2O and translocated food from aerial part of the plant. This finding was in agreement with the findings of Shafiei (2015) that parameters with strong positive correlation could be used in a breeding programme.

The significant positive correlation between fruit width and fruit yield ($r = 0.93$, $P < 0.01$) can be used in selection of more promising genotypes; similar results were obtained by Santos *et al.* (2017) who observed that this correlation can be used to recognize the heaviest fruit in the field using simpler instruments which can be of great importance since it benefits works aimed at genotype selection.

As expected, there was a negative correlation between plant height and plant yield ($r = -0.003$, $P < 0.01$), which according to Santos *et al.* (2017) may indicate an effect of competition among fruits for photo assimilates and hence a dilution effect brought about by the increased or reduced yield (see online Supplementary Table S2).

The existence of high variability as shown by diversity values recorded in online Supplementary Table S3 indicates that the diversity among the landraces is due to variation in qualitative traits. Overall, a high value of (H') represents a diverse and equally distributed classes for an individual trait. On the contrary, a lower value that indicates less diversity since Shannon–Weaver diversity index accounts for abundance and evenness of a population present in a community according to Hirakawa *et al.* (2013).

The number of SNPs tabulated in Table 3 shows a Ti/Tv ratio ranging from 1.4 to 1.57. Previously, Ni *et al.* (2012), Wencai *et al.* (2004) and Sathya *et al.* (2015) proclaimed that Ti/Tv ratio for a random variation resulting from systematic errors in the sequencing technology, alignment artefacts and data processing failures should be close to 0.5. In this study, transition to transversion ratio ranged from 1.40 to 1.57, a difference of 0.17 (Table 3). The SNPs were mined across all the 12 chromosomes at varying numbers. Giovannoni, (2007) also observed similar results from his work on tomato fruit ripening with variation occurring in all the 12 chromosomes and that these variations were caused by either deletions or insertions.

SNPs diversity was mainly contributed by geographical locations unlike the morphological characterization which grouped the landraces according to fruit shape, colour and size. SNPs diversity also revealed population admixture among specific landraces from Kenya, South Africa, Ethiopia, Morocco and Madagascar (Fig. 3 and see online Supplementary Fig. S4). This is in agreement with what Wu *et al.* (2015) and Hamilton *et al.* (2012) found out that environmental variables can have an impact on the movement of gametes and individuals among natural populations, hence affecting gene flow patterns. This may also lead to spatial and progressive dispersal of genetic variation and evolutionary advancement of regular populations. In conclusion, morphological description is important in initial characterization of African tomato landraces with similar phenotypes, i.e. Tindi 050589 and Tindi 050590 grouped together using morphological traits.

There was a significant variation among African tomato contributed by vegetative growth stages of the African tomato landrace like plant height, leaf blade length, leaf blade width and fruit width. Substantial variation among the 17 African tomato landraces was observed in the reproductive stages of the African tomato landraces, i.e. fruit colour, fruit shape, fruit texture, leaf base and leaf lade colour. However, transcriptome SNP analysis revealed significant variation among the African tomato according to their geographical location indicating that morphological characterization of African tomato can only lay a foundation but it does not reveal genetic diversity. While transcriptome analysis goes beyond the phenotypic traits and showed which of the landraces from different geographical locations had been mixed.

It was found out that environmental variables can have an impact on gene flow patterns, which may influence spatial and progressive dispersal of genetic variation and evolutionary advancement of regular populations. This study represents an important step forward in genomics, genetics and for the breeding of cultivated tomato.

Recommendations

Next generation sequence should be used to fully characterize and unveil the responsible genes for the unique traits in the African tomato landraces for breeding purposes.

Acknowledgement

The support for this research work was provided by U.S. Agency for International Development through the Partnerships for Enhanced Engagement in Research (PEER) program Sub-Grant Number: PGA-2000003426 to Prof. Willis Owino and Prof. James Giovannoni. We also acknowledge the support from National Commission for

https://doi.org/10.1017/S1479262117000314 Published online by Cambridge University Press

Science Technology and Innovation (NACOSTI). The authors would like to thank Dr. Tsvetelina Stoilova of AVRDC- World Vegetable Center in Arusha, Tanzania for the provision of a number of the tomato accessions. We also thank the Boyce Thomson Institute for Plant Research (BTI), Cornell University, USA for hosting GWM during her research study visit and technical assistance in the current study.

Supplementary material

The supplementary material for this article can be found at https://doi.org/10.1017/S1479262117000314

References

Chatfield C and Collins AJ (1980) *Introduction to Multivariate Analysis*. London: Chapman and Hall.

Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with next-generation sequencing data. *BMC Bioinformatics* 12: 323. doi: 10. 1186/1471-2105-12-323.

Giovannoni JJ (2007) Fruit ripening mutants yield insights into ripening control. *Current Opinion in Plant Biology* 10: 283–289.

Hamilton JP, Sim SC, Stoffel K, Van DA, Buell CR and Francis DM (2012) Single nucleotide polymorphism discovery in cultivated tomato via sequencing by synthesis. *Plant Genome* 5: 17–29.

Hirakawa H, Shirasawa K, Ohyama A, Fukuoka H, Aoki K, Rothan C, Sato S, Isobe S and Tabata S (2013) Genome-wide SNP genotyping to infer the effects on gene functions in tomato. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*. 20: 221–233. doi: 10.1093/ dnares /dst005.

Kisua J, Mwikamba K, Makobe M and Muigai A (2015) Genetic diversity of sweet and grain sorghum populations using phenotypic markers. *International Journal of Biosciences* 6: 34–46.

Kumar R, Tyagi AK and Sharma AK (2011) Genome-wide analysis of auxin response factor (ARF) gene family from tomato and analysis of their role in flower and fruit development. *Molecular Genetics and Genomics* 285: 245–260.

Lawal IO, Grierson DS and Afolayan AJ (2015) Phytochemical and antioxidant investigations of a Clausena anisate hook, a South African medicinal plant. *African Journal of Traditional, Complementary and Alternative medicines* 12: 28–37. doi: 10.4314/ajtcam.v12i1.5.

Lee TH, Guo H, Wang X, Kim C and Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* 15: 162.

Ni Y, Hall AW, Battenhouse A and Iyer VR (2012) Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. *BMC Genetics* 13: 46.

Nielsen PR, Albrechtsen A and Song YS (2011) Genotype and SNP calling from or without a reference genome. *BMC Bioinformatics* 12: 323. doi: 10.1371/journal.pone.0037558

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J and Trajanoski Z (2013) A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* 15: 256–278. 256–78. doi: 10.1093/bib/bbs086.

Patel RK and Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation *PLoS ONE*, https://doi.org/10.1371/journal.pone.0030619.

Reddy BR, Reddy MP, Reddy DS and Begum H (2013) Correlation and path analysis studies for yield and quality traits in tomato (*Solanum lycopersicum* L.) IOSR. *Journal of Agriculture and Veterinary Science (IOSRJVS)* 4: 56–59.

Reddy BV, Ramesh SR, Reddy PS and Kumar AA (2009) Genetic enhancement for drought tolerance in sorghum. *Plant Breeding Reviews* 31: 189–222.

Santos PC, Alexandre P, Marta SM, Almy JC and Daniele LR (2017) Relationship between yield and fruit quality of passion fruit C0 progenies under different nutritional levels. ISSN 0100-2945. doi: 10.1590/0100-29452017691.

Sathya B, Akila PD and Gopal RK (2015) NGS meta data analysis for identification of SNP and INDEL patterns in human airway transcriptome: 4–9 sequencing data. *PLoS ONE* 7: e30619.

Shafiei SM 2015 Apolarity for determinants and permanents of generic matrices. *Journal of Commutative Algebra* 7: 89–123. doi: 10.1216/JCA-2015-7-1-89. https://projecteuclid.org/euclid.jca/1425307760.

Shirasawa K, Ishii K, Kim C, Ban T, Suzuki M, Ito T, Muranaka T, Kobayashi M, Nagata N, Isobe S and Tabata S (2013) Development of capsicum EST-SSR markers for species identification and in silico mapping onto the tomato genome sequence. *Molecular Breeding* 31: 101–110.

Wang K, Li M and Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 38, e164. doi: 10.1093/nar/gkq603.

Wencai Y, Xiaodong B, Eileen K, Christina E, Sophien K, Esther van der K and David F (2004) Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. *Molecular Breeding* 14: 21–34.

Wu Z, Yu D, Wang Z, Li X and Xu X (2015) Great influence of geographic isolation on the genetic differentiation of *Myriophyllum spicatum* under a steep environmental gradient. *Scientific Reports* 5: 15618. doi: 10.1038/srep15618.

Yang WC, Bai XD, Kabelka E, Eaton C, Kamoun S, vander Knaap E and Francis D. (2004) Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum*. By computer aided analysis of expressed sequence tags. *Molecular Breeding* 14: 21–34.

Zhang X, Sebastiani P, Liu G, Schembri F, Dumas Y, Langer E, Alekseyev E, O'Connor Y, Brooks LD and Spira M (2010) Similarities and differences between smoking-related gene expression in nasal and bronchial epithelium. *Physiological Genomics* 41: 1–8.

Zhong S, Joung JG, Zheng Y, Chen Y, Liu B, Shao Y, Xiang JZ, Fei Z and Giovannoni JJ (2011) High-throughput Illumina strand-specific RNAsequencing library preparation. *Cold Spring Harbor Protocols* 8: 940–949.