

Genetic diversity and genome wide population structure analysis of a grain amaranth (*Amaranthus hypochondriacus*) diversity panel using Genotyping by Sequencing

Dinesh Pandey

dinesh.pandey@gbpuat-cbsh.ac.in

Govind Ballabh Pant University of Agriculture and Technology

Rashmi Chauhan

Govind Ballabh Pant University of Agriculture and Technology

Sharat Prabhakaran

Govind Ballabh Pant University of Agriculture and Technology

Apoorv Tiwari

Sam Higginbottom University of Agriculture, Technology & Sciences

Dinesh Joshi

Vivekananda Parvatiya Krishi Anusandhan Sansthan

Rahul Chandora

ICAR-NBPGR (Regional Station)

Gohar Taj

Govind Ballabh Pant University of Agriculture and Technology

Tarannum Jahan

Govind Ballabh Pant University of Agriculture and Technology

S. P. Singh

Govind Ballabh Pant University of Agriculture and Technology

J. P. Jaiswal

Govind Ballabh Pant University of Agriculture and Technology

Anil Kumar

Rani Lakshmi Bai Central Agricultural University

Article

Keywords: *Amaranthus hypochondriacus*, Pseudocereal, Genotyping By Sequencing, Exotic and Indigenous accessions, Molecular genetic diversity, Population structure, SNP-carrying genes, Gene

ontology.

Posted Date: August 18th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3252287/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on September 30th, 2025. See the published version at <https://doi.org/10.1038/s41598-025-01626-7>.

Abstract

Amaranth, a gluten-free pseudo-cereal, is grown, cultivated, and adapted in diverse ecological zones all over the world. In recent years, increased attention has been paid on its nutritional eminence, particularly in terms of its seed protein, which specifically contains high levels of the essential amino acids. This important food and nutritional security crop's gene pool, which is diverse, needs to be characterized and genotyped in order to broaden its genetic base and to boost production. Genotyping by sequencing data of diverse germplasm serve useful tool for identification of important candidate genes associated with important traits of plants. In the present study, genotyping-by-sequencing (GBS) approach was used to characterize a genetically diverse collection of 192 Amaranth accessions. This resulted in an estimated 41,931 single-nucleotide polymorphisms (SNPs) segregating across the entire collection and several thousand SNPs segregating within every accession. A model-based population structure analysis reveals the presence of three subpopulations among the Amaranth accessions, which are in parallel with the results of phylogenetic analysis. A total of 1796 gene ontology (GO) terms were assigned to SNP-carrying genes for three main categories: biological process, cellular component, and molecular function. High-throughput genotyping and sequencing data generated, will be very useful not only for breeders for further enhancement of Amaranth but also for molecular biologists for isolation and identification of nutritionally important genes from Amaranth, which can be used for biofortification programs.

Introduction

Out of the several thousand known edible plant species, just over 30 crops provides 95% of the world food supplies and four major crops rice, maize, wheat and potato provides 60–70% global dietary requirements (Joshi et al., 2018). The shrinking portfolio of world food basket raises serious concerns on how major crops alone can ascertain the food and nutritional security of the ever-increasing population. Further, the major crops (rice, wheat maize and potato) are the reservoir of carbohydrates providing high amount of calories but lacking the essential amino acids and minerals for a balanced nutrition (Kumar et al., 2021). Therefore, diversifying the food basket by incorporation of long forgotten ancient pseudocereals with unparallel nutritional profile is eventually a leading strategy to address the global nutritional security and combat the health infirmities and prevalent hidden hunger (Joshi et al., 2020).

Grain amaranth, *Amaranthus hypochondriacus*, is one such pseudocereal with tremendous but under-explored source of nutraceutical properties and has the potential to diversify the cropping systems around the world. Amaranth grains are known for high concentration of essential amino acids, gluten free nature of protein, low glycemic index and exceptional mineral profile compared to the major crops (Parihar et al., 2021). Additionally, amaranth is the rich source of several functional compounds and bioactive flavonoids which includes squalene, tocopherols, tocotrienols, sterols and other lipophilic constituents (Rastogi and Shukla, 2013). Several therapeutic studies have confirmed the role of these bioactive compounds in reducing the risk of cardiovascular disease, cancer, obesity and type 2 diabetes

(Silva-Sanchez et al., 2008; Khamar and Jasrai, 2014; Reyad-ul-Ferdous et al., 2015). Furthermore, the C₄ photosynthetic pathway and ability to withstand harsh environmental conditions and fragile ecosystems makes this crop an ideal candidate for climate resilient agriculture and an alternative food grain for the development of sustainable agriculture systems.

Despite the immense potential, efforts in order to bred highly nutritive and productive cultivars for the amaranth have been limited. The critical impediments hampering the progress of breeding for agronomic and nutraceutical traits in grain amaranth are the poorly characterized germplasm collections at the genome level and the undeciphered genetic mechanisms governing these traits. There are only scattered germplasm holdings of grain amaranth maintained in *in situ* conditions across the world (Joshi et al., 2018). To some extent, these collections have been assessed for morphological and agronomic traits in different agro-ecological conditions and utilized in breeding programme. Genetic diversity has been investigated in these collections with a variety of genetic markers; however, most of them are at a very small scale with very less number of PCR based markers (Chandi *et al.*, 2013; Khaing et al., 2013; Stefunova *et al.*, 2015; Gelotar et al., 2019).

High-throughput genomic technologies can now be used even in crops such as Amaranth (Clouse et al., 2016; Lightfoot et al., 2017), which are still considered subsistence crops, to analyze molecular markers, transcript sequences, gene space, genome structural variations, and genome sequences. Low sequencing costs have motivated the development of GBS, a high-density marker genotyping technology that is robust, cost-effective, and highly multiplexed. Studies of genetic variation in a wide variety of species are possible using GBS by analyzing a large number of SNPs. There are several advantages to using SNPs in plant genetic analysis, including their widespread genomic distribution, co-dominant inheritance, reproducibility, and location on particular chromosomes (Brookes, 1999; Cho et al., 1999; Gupta et al., 2001; Sachidanandam et al., 2001; Rafalski, 2002; Gupta and Rustgi, 2004). High throughput technologies of sequencing and genotyping are expected to improve breeding progress in developing countries in several ways: For instance, these technologies will facilitate selection of parents for new crosses and introduce novel alleles from exotic germplasm through high-resolution germplasm fingerprinting with low coverage sequencing or SNP genotyping. As a result of a lack of characterization of the amaranth gene pool at genome level, it is difficult to find genetically diverse accessions that have beneficial traits to be used in breeding programs. Furthermore, the crop has not been extensively investigated for the purpose of understanding its diversity, taxonomy, and evolution at genome level. To date, there are only few reports on utilization of next generation SNP markers for determining genetic structure of amaranth diversity collections (Thapa et al., 2021; Jamalluddin et al., 2022). Therefore, a detailed genome level characterization of genetic diversity and population structure analysis of the diverse amaranth germplasm collections would provide insights on genetic relationship between the different accessions.

In the present study, we used GBS technology to genotype a diverse amaranth germplasm collection. The objectives of this study were to: (1) generate SNP markers by GBS technology and evaluate their characteristics; (2) determine the population structure and genetic diversity of the amaranth diversity

panel and (3) Functional annotation of the generated SNPs and assigning the gene ontology (GO) terms to SNP-carrying genes in amaranth. Such studies would provide comprehensive knowledge on the value of locally adapted and exotic germplasm and allow for its effective utilization in breeding programs to develop nutritionally rich and stress tolerant cultivars.

Results

High-Throughput Genotyping of *Amaranthus hypochondriacus* accessions was performed by Genotyping-by-sequencing analysis using 192 genotypes, which comprised of a set of 118 (IC) accessions from India and 74 (EC) accessions from different regions of the world (Fig. 1). ApeKI restriction enzyme was used for preparing the library because it is methylation sensitive. Methylation-sensitive restriction enzymes have been shown to be effective in selecting genomic DNA for exonic gene-containing regions and reducing genomic regions with repeat elements (Fellers, 2008). Single-nucleotide polymorphisms were called using the TASSEL GBS pipeline (Glaubitz et al., 2014).

Amaranthus hypochondriacus v2.1 downloaded from (https://phytozome-next.jgi.doe.gov/info/Ahypochondriacus_v2_1) was used as reference genome. UNEAK commands are run as TASSEL plugins via the command line (Linux or Mac operating systems). A total of around 400 million raw reads were generated and filtered. A detailed final SNP statistic, showing number of heterozygous, homozygous and total single-nucleotide polymorphisms (SNPs) per accession is given in Supplementary Table S1. A 6599654 SNPs were identified among the accessions of which 5,477,353 were homozygous and 1,111,501 were heterozygous.

The SNP having maximum frequency of allele occurrence is G/C with an overall allele occurrence of 10,041, with a frequency of about 23.946 per cent. The SNP having least frequency of occurrence is T/G with an overall occurrence of 1017, with a 2.425 per cent frequency. Single-nucleotide polymorphism statistics in the entire dataset showing number and percentage frequency of allele occurrence is detailed in Table 2.

In plant genomes, due to biased mutational processes (e.g. cytosine deamination), the ratio of transitions to transversions for bona fide polymorphisms is expected to be much greater than one. In the present study, total number of transitions were 14488 as compared to 27327 number of transversions resulting in the transition/transversion ratio of 0.54 for *Amaranthus* which represents no bias towards transition or transversion type of base substitutions. In previous studies, this ratio has been reported as high as 2.5 in maize (Jiao et al., 2012) and 2.4 in *Arabidopsis sp.* (Ossowski et al., 2010) which is due to higher rate of transition than that of transversion. The data regarding the rates of transitions and transversions are graphically represented in Fig. 2.

Number of SNPs per chromosome ranged from 1315 (chromosome 16) – 4361 (chromosome 1). The total number of SNPs present in each chromosome is given in Fig. 3a. A heat map portraying SNP density per chromosome at one Mb window size is shown in Fig. 3b.

Triangle plot for pairwise LD between marker sites was worked out. To visualize LD throughout the genome, heat maps were produced based on pairwise R^2 estimates, and corresponding p-values were calculated using permutations for all marker pairs (Fig. 4.a). The R^2 values for each marker pair are presented in the bottom half of the heat map and are represented by shades of red increasing in intensity in equal increments of 0.1 from 0.0 (white) to 1.0 (red). The linkage disequilibrium decay was calculated using the linkage disequilibrium values as revealed by TASSEL 5.0. The average size of haplotype blocks was estimated at 295883 bp with the help of the LD values per marker pair obtained from TASSEL 5.0 followed by analysis using RStudio (Fig. 4.b).

Phylogenetic Analysis

In the present study, the genetic diversity analysis was done among the global collection of 192 Amaranth genotypes. The phylogenetic tree was created through neighbor-joining method of TASSEL software. In the resulting dendrogram, the genotypes segregated into 3 major clusters A, B and C (Fig. 5). Cluster C consisted of only the Indian genotypes, Clusters B also consisted of Indian genotypes except one exotic genotype whereas Cluster A was comprised of exotic and Indigenous genotypes. The cluster A, the largest cluster contained 159 genotypes, 73 out of the total 74 exotic genotypes were present in this cluster along with 86 indigenous genotypes. Cluster A can be broadly categorised into genotypes of 8 sub clusters, namely, A1, A2, A3, A4, A5, A6, A7 and A8. A1 consisted of 15 genotypes of which three are exotic form China, Germany and France, rest 12 are indigenous accessions 7 from Maharashtra, 4 from Himachal Pradesh and one from Gujarat. A2 consisted of 13 accessions 6 exotic accessions – 2 from China, 2 USA, 1 Germany and 1 from Russia. Also 7 indigenous accessions – 4 from Maharashtra, one each from Uttarakhand, Himachal Pradesh and Meghalaya. A3 consisted of 10 accessions of which 3 were exotic one each from U.S.A., Germany and Poland. The remaining 7 indigenous genotypes were 3 from Maharashtra, one each from Himachal Pradesh and Uttarakhand, remaining two unknowns. A4 consisted of 24 accessions of which majority were exotic (17) – 6 from Germany, 5 from U.S.A. and 4 Russia, one each from France and Latvia. It also contained 6 indigenous accession, 2 each from Himachal, Jammu & Kashmir and one each from Madhya Pradesh, Uttarakhand and Meghalaya. Sub cluster A5 consisted of 26 accessions, of which 10 are exotic, 3 from the U.S.A., 2 from Germany, one each from Russia, Nepal, China, Kenya and South Africa. Among the indigenous, 8 from Himachal Pradesh, 3 from Maharashtra, 2 Uttarakhand and one each from, Sikkim, Jammu & Kashmir and Madhya Pradesh. A6 contained the maximum individuals under subclusters of cluster A, 38 accessions and also the maximum number of exotic genotypes 19. 5 from U.S.A., 4 from Russia, 3 from Netherlands, 2 each from China, Latvia, Germany, one from Poland. Among the indigenous, 6 genotypes each from Maharashtra and Himachal Pradesh, 3 from Gujarat, one each from Rajasthan, Sikkim, Madhya Pradesh and Uttarakhand. Subcluster A7 consisted a total of 21 accessions, of which 5 were exotic accessions, three from the U.S.A. and one each from Latvia and Benin. Among the indigenous, 5 were from Maharashtra, 4 from Madhya Pradesh, 3 from Himachal Pradesh, one each from Uttarakhand, Sikkim, Gujarat and Karanataka. The final sub cluster of cluster A, A8, had 12 accessions of which 5 were

exotic- 4 from U.S.A., one from Germany. In case of indigenous genotypes, 2 Maharashtra and one each from Sikkim, Meghalaya, Uttarakhand, Himachal Pradesh, 1 Madhya Pradesh.

Cluster B comprised of 7 *A. hypochondriacus* genotypes, all except one were indigenous genotypes. The exception was an exotic genotype EC-120051 from China. The indigenous genotypes were from Maharashtra, Uttarakhand and Madhya Pradesh.

Cluster C contained 26 genotypes, which were all indigenous, which can be broadly categorized into two groups. 10 from Maharashtra, 4 from Sikkim, 2 each from Uttarakhand, Himachal Pradesh, Madhya Pradesh, Gujarat, one each from Karnataka, Goa and Chhattisgarh.

Population Structure Analysis

To infer the right number of subpopulations (K), k was set from 1 to 10 and 3 runs for each k value were performed. In each run, a burn-in of 5000 iterations was followed with an extra 50000 replications. To choose the best k value, an impromptu statistic ΔK based on the rate of change in the log probability of data between successive K values was used. The maximum ΔK value was observed at $K = 3$. The inferred ancestry at $K = 3$ suggested that the *Amaranthus* genotypes were grouped into three subpopulations (Fig. 6). As expected, there was a decent correlation between the geographic pattern observed in the phylogenetic analysis and the population structure identified using STRUCTURE software with some prominent exceptions.

Since the ancestries of most of the genotypes were not known earlier, an admixture model was used to infer their ancestry. A model, which allows for hybrids is more flexible and often provides a better fit for complicated population structure. Figure 6 displays the Proportional allocations of each of 192 *A. hypochondriacus* accessions to K ancestral gene pools inferred using ADMIXTURE (Alexander et al., 2015), under $K = 3$, shown as a bar plot in Fig. 7.

Principal Component Analysis (PCA):

Two-dimensional PCA of 192 accessions has further indicated the separation of different accessions in three major groups and it also indicated that different accessions of *A. hypochondriacus* overlapped with each other as per their phylogenetic relatedness (Fig. 8.)

Functional Annotation of Single-Nucleotide Polymorphisms

GO terms were assigned to *A. hypochondriacus* SNP-carrying genes, proteins annotated in GO terms. A total of 2815 GO terms were collected that were most frequently related to biological processes (1107) followed by molecular function (984) and cellular components (724). The gene ontology assignment categorization of SNP carrying genes into three main categories is portrayed in Fig. 9.

Among the biological processes, leaving out the unknown and uncharacterized genes, maximum number of genes corresponded to metabolic process (39.65%) and cellular process (30.8%), followed by localization (8.03%). In case of molecular function, binding and catalytic activity proteins were found to

have the highest share with 43.5% and 43.2%. In cellular component category, it's in the proportion of cell (19.06), cell part (19.06), membrane (17.54), organelle (13.25), protein contain complex (12.84), Gene ontology (GO) term assignment for different categories of *A.hypochondriacus* single-nucleotide polymorphism carrying genes is shown in Fig. 10.

Discussion

A breeder's ability to manage crop genetic resources efficiently depends greatly on having adequate knowledge of existing genetic diversity. Molecular, biochemical, and morphological analysis are all possible methods for analyzing genetic diversity. Since the above methods have limitations, next-generation sequencing technologies (such as Genotyping by sequencing or GBS) are being applied to a wide range of crop species to study diversity among population in precise manner. In addition to providing a powerful and rapid way to identify the genetic basis of agriculturally important traits, precise phenotyping of agronomically significant traits in combination with genotyping help in predicting the breeding value of individuals in breeding populations. GBS has now made it possible for underutilized pseudocereal crop like Amaranth to conduct a variety of genetic analyses that were previously too expensive and unaffordable. The GBS based approach generate data which are very useful in conducting a number of analyses including Genome wide association mapping (GWAS), marker based identification of QTL, Germplasm diversity analysis, germplasm fingerprinting. There were three major clusters identified based on the cluster analysis of genotypes based on geographical distribution. However, the dendrogram did not classify genotypes unequivocally based on country of origin. The main reason for this could be attributed to broad diversity of accessions of *A. hypochondriacus* used in the present study. Another reason is that different genotypes or accessions of *A. hypochondriacus* have spread to different geographical locations from it's one centre of origin i.e. Mexico but some of them retained the same genetic background. The domestication of Amaranth took place in USA around 6000 years ago (He et al 2013). The observation that majority of Asian and European Amaranth accessions were clustered with accessions from USA suggests that Amaranth it spread to Europe and Asia after it's domestication in USA. The latter is supported by a report which states that Amaranth was introduced and grown in Germany in sixteenth century and spread in Europe by 17th century (Rastogi and Shukla, 2020). It then spread to Himalayan region and interior china in 19th century. Its cultivation started in China, India Africa, Europe, North and South America in 20th Century. Cluster A comprises of Indigenous accessions largely clustered with accessions from Germany, China and USA. It is therefore obvious that most of accessions from India may have evolved from the accessions of these countries. Cluster B contains one accession from China EC-120051 clustered with accessions from India. Remaining six accessions of China clustered with other Indian accessions along with exotic accessions from Germany, Russia and USA. It suggests the spread of Amaranth from Europe to China from which it spread to India due to close geographical proximity between two countries and possible gene flow. Cluster C contains only indigenous accessions which may have evolved separately due to divergence from ancestral accession of China or Europe. Since indigenous accessions are grouped together with accessions from different countries, therefore it suggests that allelic reshuffling may have occurred due to exchange of

Indian and exotic genetic resources followed by hybridization and selection. The division of a plant population into genetically pure and admixed lines is an important feature of population structure analysis. In the present investigation, the population diversity analysis was done based upon the geographical origins of different accessions, which led to division of entire collection of accessions in three subpopulations. The accessions in cluster A has shown admixture of 54% with most of these accessions are from USA. The presence of only one Chinese accession along with Indian accessions in cluster B and six Chinese accession in cluster C indicate some level of admixture with accessions from China. In order to improve Amaranth breeding programme, there were made crosses between well adapted US accessions and European or Asian accessions. The presence of different alleles in accessions of Cluster A and Cluster B as observed in the present study might have resulted due to recombination and reshuffling of alleles due to hybridization between accessions of diverse geographical origin. The admixture between US accessions and Indian accessions suggests that many accessions of Amaranth may have resulted due to intercrossing between Indian and US accessions as a part of breeding strategy. However, it also suggests that some accessions have also been derived from some European ancestors. The presence of Chinese accession in cluster B also suggests the possible gene flow from Chinese accession to Indian accessions as a result of breeding activities. However, Cluster C contains only Indian accessions, which are grouped together but fall in a separate cluster. This has indicated that these indigenous accessions might have evolved either from Chinese or US accessions followed by their complete divergence to form a separate population. The principal component analysis (PCA) further supports the phylogenetic relatedness among different Exotic and Indigenous accessions of Amaranth. In the present study most of the snp carrying genes of Amaranth which have been functionally annotated belong to different metabolic and cellular processes. Using functionally annotated SNPs, it may be possible to establish marker–trait associations and identify genes controlling agriculturally important traits.

Materials and Methods

Plant Material

The diversity panel comprised of 192 accessions representing grain amaranth, spanning from countries of origin, landraces, improved cultivars and genetic stocks, which were procured from NBPGR, Shimla (Table 1). This collection comprised of accessions which broadly fall in to two categories viz. Exotic collection (EC) and Indigenous collection (IC). The 118 IC accessions represents the released cultivars, and landraces of grain amaranth from 13 states of India including North western Himalayan region. The IC accessions represent all the amaranth growing agro-ecological conditions (tropical, temperate and semi-arid) prevalent in the country. The 74 EC accessions assembled in the panel belongs to USA, Germany, Russian Federation, China, Lativa, Netherlands, France, Poland, Nepal, Benin, Kenya and South Africa. All the genotypes of the diversity panel are derived from self-fertilizing lineages, and as such, heterozygosity is not expected due to the extreme self-pollinating behavior of the crop.

DNA Extraction and Quantification

DNA was extracted from seeds of 192 genotypes of Amaranth using CTAB based protocol. After extraction, the quality and quantity of extracted DNA was checked on 0.8% agarose gel. 1 kb DNA ladder was loaded as control to check the quality and quantity of DNA. In order to verify that restriction enzymes can cut genomic DNA without being inhibited, restriction digestion test using *HindIII* restriction enzyme was carried out on every genomic DNA sample. This helps in ascertaining the purity and quality of genomic DNA for restriction-site associated DNA sequencing library preparation, hence ensuring better results. DNA was then quantified using the Nanodrop Spectrophotometer 1000 (Thermo Fisher Scientific).

Genotyping by Sequencing

The genomic DNA of 192 *Amaranthus hypochondriacus* accessions were subjected to GBS (Genotyping by Sequencing) analysis. DNA samples were normalized to a final concentration of 5 µg for GBS. Lyophilized DNA samples were sent to NXGenBio Life Sciences, New Delhi, for GBS analysis. GBS libraries were prepared following Elshire *et al.* (2011) with minimal modification; Libraries were prepared with ApeKI restriction enzyme for genomic DNA digestion followed by ligation with ApeKI and common adapters. The purified 192-plex final DNA library was quantified using a Bioanalyzer (Agilent Technologies) and were sequenced on a single lane of Illumina HiSeq™ X10 platform (Illumina® Inc., San Diego, CA, USA) using V4 sequencing chemistry.

FASTX Toolkit (version 0.0.13) was used to separate sample from combined data based on barcodes.fastx_barcode_splitter.pl (FASTXToolkit), a program which reads FASTA/FASTQ file and splits it into several smaller files, based on barcode matching. FastQC (version 0.11.5, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) were used for Quality Check. Universal Illumina Adapters was removed using trim_galore (version 0.6.2, https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), a wrapper script to automate quality and adapter trimming as well as quality control. *Amaranthus hypochondriacus* v2.1 downloaded from (https://phytozome-next.jgi.doe.gov/info/A_hypochondriacus_v2_1) was used as a reference genome (Lightfoot, D.J. *et al.*, 2017). Reads were mapped against the reference genome using the MEM algorithm of BWA (version 0.7.5) with default parameters.

Single-Nucleotide Polymorphism Calling and Filtering

Variant calling was done using UGBS- GATK pipeline (version 3.6, <https://gatk.broadinstitute.org/hc/enus>). The GATK is the industry standard for identifying SNP and indels in germline DNA and RNAseq data. In addition to the variant callers themselves, the GATK also includes many utilities to perform related tasks such as processing and quality control of high-throughput sequencing data. Variants were filtered and indels were removed using vcftools (version 0.1.17, <http://vcftools.sourceforge.net/>). All SNPs were filtered at minor allele frequency 5% MAF and 20% missing rate (minimum 80% sample have that particular SNPs). A workstation with 250 GB of RAM with 32-core Intel Xeon processor was used for all the bioinformatics work and consequent analyses. The GBS data was obtained in the form of a hapmap file format.

Population Structure Analysis:

In order to undertake research on population structure and evaluate multilocus genotypic data (192 accessions) we have utilized the window-based software package known as Structure, which is openly available. Initially, the hap map file was changed into a variant call format file by TASSEL software and then the PGD Spider V.2.0.8.2 software was used to convert it into the structure format so that it could be used as an input file for the Structure program. For estimation of exact population structure, K's from 1 to 10 (with 10 iterations) were run and the LnP(D) value was used to group all the genotypes. For each run, a burn-in of 5000 iterations was followed by an additional 50,000 replications.

Linkage disequilibrium and haplotype block:

Triangle plot for pairwise LD between marker sites were carried out. To visualize LD throughout the genome, heat maps were produced based on pairwise R^2 estimates, and corresponding p-values were calculated using permutations for all marker pairs (Fig. 4.b). The R^2 values for each marker pair are presented in the bottom half of the heat map and are represented by shades of red increasing in intensity in equal increments of 0.1 from 0.0 (white) to 1.0 (red). The linkage disequilibrium decay was calculated using the linkage disequilibrium values calculated using TASSEL 5.0. The average size of haplotype blocks was estimated using RStudio.

Phylogenetic Analysis:

The phylogenetic analysis for the samples was performed with the help of the graphical user interface of TASSEL V. 5.0 standalone using neighbor-joining algorithm formed with node name.

Principal Component Analysis:

Principal component analysis (PCA) is useful for this since it allows check how similar or unlike each sample is in a given dataset, or how the samples in the collection are related to one another. The PCA was carried out using PCA function in RStudio.

Functional Annotation of Single-Nucleotide Polymorphisms:

The hap map file, which includes flanking sequences for each of the filtered SNPs, was used to infer the possible roles of the SNPs. This most recent iteration of the Plant RefSeq (reference sequence) database is now accessible via a locally hosted BLAST server. Each SNP's flanking sequence was compared to the Plant RefSeq database using a local BLAST. The server housing the standalone BLAST has 32 cores and 256 GB of RAM configured for multi-threading mode. Data were collected and combined after obtaining the BLAST result, which included Gene IDs of the corresponding species. Each of these SNPs was given a gene ontology identifier (GO ID) based on a search of the GO database, which may have resulted in a single GO ID or several IDs.

Conclusions

Amaranth accessions have been classified into three subpopulations based on several thousand SNPs segregating within each accession, and these SNPs have also revealed phylogenetic relationships among them and have provided a valuable dataset. In the future, Amaranth breeding could be guided by these data. With our SNP genotyping, breeders will be able to select parents and introduce alleles from exotic germplasm into new crosses. Amaranth grains, being very rich in proteins and minerals such as Phosphorous, Calcium, Potassium as well as important phytochemicals such as squalene, are becoming very popular among the health conscious people. Having developed a useful set of SNPs, we are trying to find out SNPs associated with high grain protein or calcium content in future investigations. As the genome of *A. hypochondriacus* is already deciphered, such investigations would pave way to identify genes/QTLs linked to high protein, minerals through GWAS. It is therefore anticipated that these high-throughput genotyping and sequencing technologies would contribute to bettering breeding techniques and raising the caliber and output of these minor crops.

Declarations

Disclosure of potential conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Funding

The present research work was supported by Department of Biotechnology, Govt. of India in the form of research grants (BT/BI/01/069/2018).

Compliance with Ethical Standards:

Guidelines and legislation

All plant experiments were conducted in accordance with relevant institutional, national, and international guidelines and legislation.

Disclosure of potential conflicts of interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

Research involving Human Participants and/or Animals

Not applicable.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Author contributions

Conceptualization [Dinesh Pandey] Methodology [Rashmi Chauhan, Tarannum Jahan], Germplasm collection and analysis [Dinesh Joshi, Rahul Chandora] Research data analysis and interpretation [Sharat Prabhakaran and Apoorv Tiwari] Writing - Original draft preparation [Dinesh Pandey], Writing- Review and Editing [Rashmi Chauhan, Sharat Prabhakaran, Apoorv Tiwari, Dinesh Joshi] Funding Acquisition [Dinesh Pandey, Anil Kumar] Resources [J.P.Jaiswal, S.P.Singh, Gohar Taj]

Ethics approval and consent to participate

Not applicable

Consent for publication

All the authors have consented for publication of content.

Availability of data and materials

The NGS data of the 192 genotypes in the present study will be made available in the NCBI SRA database.

References

1. Alexander DH, Shringarpure SS, Novembre J, Lange K. 2015. ADMIXTURE 1.3 Software Manual. <http://software.genetics.ucla.edu/admixture/download.html>.
2. Brookes AJ. 1999. The essence of SNPs. *Gene* 234:177–186.
3. ChandiA, Milla-Lewis SR, Jordan DL, York AC, Burton JD,Zuleta MC *et al.*2013. Use of AFLP Markers to AssessGenetic Diversity in Palmer Amaranth (*Amaranthuspalmeri*) Populations from North Carolina and Georgia. *Weed Sci* 61: 136–145.
4. Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E *et al.* 1999. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. *Nat Genet* 23:203–207.
5. Clouse JW, Adhikary D, Page JT, Ramaraj T, Deyholos MK, Udall JA, Fairbanks DJ, Jellen EN, Maughan PJ. 2016. The amaranth genome: genome, transcriptome, and physical map assembly. *Plant Genome* 9:1–14
6. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES *et al.* 2011. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379. doi:10.1371/journal.pone.0019379.
7. Fellers JP. 2008.Genome Filtering Using Methylation-Sensitive Restriction Enzymes with Six Base Pair Recognition Sites. *The Plant Genome*. 1:146–152.
8. Gelotar MJ, DharajiyaDT, Solanki SD, Prajapati NN,Tiwari KK. 2019. Genetic diversity analysis and molecular characterization of grain amaranth genotypes using inter simple sequence repeat (ISSR) markers. *Bulletin of the National Research Centre*. 43:103
9. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES. 2014. TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLoS ONE* 9:e90346.

10. Gupta PK, Rustogi S. 2004 Molecular markers from the transcribed/expressed region of the genome in higher plants. *Funct. Integr. Genomics* 4:139–162.
11. Gupta PK, Roy JK, Prasad M. 2001. Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci.*80:524–535.
12. Jamalluddin N, Massawe FJ, Mayes S, HoWK, Symonds RC. 2022. Genetic diversity analysis and marker-trait associations in *Amaranthus* species. *PLoS ONE* 17(5): e0267752
13. Jiao Y, Zhao H, Ren L *et al.* 2012. Genome-wide genetic changes during modern breeding of maize. *Nat Genet* 44: 812–815.
14. Joshi DC, Sood S, Hosahatti R, Kant L, Pattanayak A, Kumar A, Yadav D, Stetter MG. 2018. From zero to hero: the past, present and future of grain amaranth breeding. *Theoretical and Applied Genetics* 131(9): 1,807– 1,823
15. Joshi DC, Zhang K, Wang C, Chandora R, Khurshid M, Li J, He M, Georgiev MI and Zhou M. 2020. Strategic enhancement of genetic gain for nutraceutical development in buckwheat: a genomics-driven perspective. *Biotechnology Advances* 39: 107479.
16. Khaing AA, Moe KT, Chung JW, Baek HJ *et al.* 2013. Genetic diversity and population structure of the selected core set in *Amaranthus* using SSR markers. *Plant Breeding*. 132: 165–173.
17. Khamar R, Jasrai YT. 2014. Nutraceutical analysis of amaranth oil, avocado oil, cumin oil, linseed oil and neem oil. *Int. J. Bioassays* 3:2090–2095.
18. Kumar A, Tripathi MK, Joshi D, Kumar V. 2021. *Millets and Millet Technology*. Springer, Singapore.
19. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. 2017. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biol* 15:74.
20. Ossowski S, Korbinian S, José I. *et al.* 2010. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327: 92–94
21. Parihar M, Kumar A, Bisht JK, Bhinda MS, Shyamnath, Meena RP, Mondal T, Joshi DC, Bijarniya H, Singh S and Kant L. 2021. Reviving the forgotten food network of potential crops to strengthen nutritional and livelihood security in North Western Himalayas. *Indian Journal of Agronomy* 66:S44–S59
22. Rafalski A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr Opin. Plant Biol* 5:94–100.
23. Rastogi A, Shukla S. 2013. Amaranth: a new millennium crop of nutraceutical values. *Crit Rev Food Sci Nutr* 53:109–125.
24. Reyad-ul-Ferdous M, Shamim D, Shahjahan M, Sharif T, Mukti M. 2015. Present biological status of potential medicinal plant of *Amaranthus viridis*: a comprehensive review. *Am J ClinExp Med* 3:12–17.

25. Sachidanandam R, Weissman D, Schmidt SC, *et al.* 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.
26. Silva-Sanchez C, Barba de la Rosa AP, Leon-Galvan MF, De Lumen BO, De Leon-Rodriguez A, Gonzalez de Mejia E. 2008. Bioactive peptides in amaranth (*Amaranthushypochondriacus*) seed. *J Agric Food Chem* 56:1233–1240.
27. Štefúnová V, Bežo M, Žiarovská J, Ražná K. 2015. Detection of the genetic variability of amaranthus by RAPD and ISSR markers. *Pakistan Journal of Botany*, 47: 1293–1301.
28. Thapa R, Edwards M, Blair MW. 2021. Relationship of Cultivated Grain Amaranth Species and Wild Relative Accessions. *Genes*, 12:1849.

Tables

Table 1. Details of global diversity panel of grain *Amaranthus* selected along with their geographical origin.

S.No.	Accessions	Geographical Origin
1	IC-35421	Anjangaon, Amravati, Maharashtra
2	IC-35422	Anjangaon, Amravati, Maharashtra
3	IC-35423	Anjangaon, Amravati, Maharashtra
4	IC-35424	Anjangaon, Amravati, Maharashtra
5	IC-35428	Anjangaon, Amravati, Maharashtra
6	IC-35430	Anjangaon, Amravati, Maharashtra
7	IC-35431	Anjangaon, Amravati, Maharashtra
8	IC-35492	Karasaon, Amravati, Maharashtra
9	IC-35493	Nagpur, Maharashtra
10	IC-35361	Patur, Balapur, Akola, Maharashtra
11	IC-35362	Patur, Balapur, Akola, Maharashtra
12	IC-35367	Patur, Balapur, Akola, Maharashtra
13	IC-35368	Patur, Balapur, Akola, Maharashtra
14	IC-35370	Patur, Balapur, Akola, Maharashtra
15	IC-35371	Patur, Balapur, Akola, Maharashtra
16	IC-35372	Patur, Balapur, Akola, Maharashtra
17	IC-35374	Patur, Balapur, Akola, Maharashtra
18	IC-35375	Patur, Balapur, Akola, Maharashtra
19	IC-35377	Patur, Balapur, Akola, Maharashtra
20	IC-35378	Patur, Balapur, Akola, Maharashtra
21	IC-35383	Patur, Balapur, Akola, Maharashtra
22	IC-35414	Anjangaon, Amravati, Maharashtra
23	IC-35416	Anjangaon, Amravati, Maharashtra
24	IC-35419	Anjangaon, Amravati, Maharashtra
25	IC-35420	Anjangaon, Amravati, Maharashtra
26	IC- 35376	Patur, Balapur, Akola, Maharashtra
27	IC-35417	Anjangaon, Amravati, Maharashtra
28	IC-35363	Patur, Balapur, Akola, Maharashtra

29	IC-35364	Patur, Balapur, Akola, Maharashtra
30	IC-35413	Anjangaon, Amravati, Maharashtra
31	IC-35432	Anjangaon, Amravati, Maharashtra
32	IC-35381	Patur, Balapur, Akola, Maharashtra
33	IC-35426	Anjangaon, Amravati, Maharashtra
34	IC-35380	Patur, Balapur, Akola, Maharashtra
35	IC-35379	Patur, Balapur, Akola, Maharashtra
36	IC-35373	Patur, Balapur, Akola, Maharashtra
37	IC-35369	Patur, Balapur, Akola, Maharashtra
38	IC-26264	Akola, Maharashtra
39	IC-35366	Patur, Balapur, Akola, Maharashtra
40	IC-35382	Patur, Balapur, Akola, Maharashtra
41	IC-35365	Patur,Balapur,AKOLA,Maharashtra
42	IC-35415	Anjangaon,,Amravati ,Maharashtra
43	IC-467910	Kurthla, Chamba, Himachal Pradesh
44	IC-36831	Panjain, Mandi, Himachal Pradesh
45	IC-36833	Bago, Shimla, Himachal Pradesh
46	IC-36834	Kathog, Shimla, Himachal Pradesh
47	IC-38039	Shimla, Himachal Pradesh
48	IC-38054	Shimla, Himachal Pradesh
49	IC-38160	Mandi, Himachal Pradesh
50	IC-38161	Mandi, Himachal Pradesh
51	IC-38192	Mandi, Himachal Pradesh
52	IC-38193	Mandi, Himachal Pradesh
53	IC- 17935	Jeori, Shimla, Himachal Pradesh
54	IC-17936	Narkanda, Shimla, Himachal Pradesh
55	IC-36835	Sariwan,, Shimla, Himachal Pradesh
56	IC-38162	Mandi, Himachal Pradesh
57	IC-38057	Shimla, Himachal Pradesh

58	IC-36830	Mangoli, Mandi, Himachal Pradesh
59	IC-329587	Tindi Udaipur, Lahaul&Spiti, Himachal Pradesh
60	IC-36832	Moag, Shimla, Himachal Pradesh
61	IC-38055	Shimla, Himachal Pradesh
62	IC-38163	Mandi, Himachal Pradesh
63	IC-38164	Mandi, Himachal Pradesh
64	IC-38056	Shimla, Himachal Pradesh
65	IC-38166	Mandi, Himachal Pradesh
66	IC - 38058	Shimla, Himachal Pradesh
67	IC-106354	Chackloh, Hamirpur, Himachal Pradesh
68	IC-329588	Tindi Udaipur, Lahaul & Spiti, Himachal Pradesh
69	IC-38159	Mandi, Himachal Pradesh
70	IC-38065	Shimla, Himachal Pradesh
71	IC-547510	Jwasa,, Uttarkashi, Uttarakhand
72	IC-547511	Paughat,, Uttarkashi, Uttarakhand
73	IC-582932	Atala, Utarkashi, Uttarakhand
74	IC-582935	Atala, Utarkashi, Uttarakhand
75	IC-47436	Narayan Ashram, Pithoragarh, Uttarakhand
76	IC-47437	Narayan Ashram, Pithoragarh , Uttarakhand
77	IC-24263	Sama,, Almora, Uttarakhand
78	IC-355789	Palvana, Almora, Uttarakhand
79	IC-24264	Naehti, Loharkhot, Almora, Uttarakhand
80	IC-24266	Mollikhal, Salt block, Bageshwar, Uttarakhand
81	IC-24265	Saling, Loharkhot, Bageshwar,Uttarakhand
82	IC-264790	Bhatraj, Nainital, Uttarakhand
83	IC-41997	Mulai, Betul, Madhya Pradesh
84	IC-41998	Ambada,, Betul, Madhya Pradesh
85	IC-41999	Ambada, Betul, Madhya Pradesh
86	IC-42000	Pagara near Pachmarhi,, Hoshangabad, Madhya Pradesh

87	IC-21795	Odagi, Singrauli, Madhya Pradesh
88	IC-21804	Karangia, Dindori, Madhya Pradesh
89	IC-21938	Shahpura, Dindori, Madhya Pradesh
90	IC-21808	Odagi,Singrauli, Madhya Pradesh
91	IC-21931	Kasturi, Dindori, Madhya Pradesh
92	IC-42001	Pagara near Pachmarhi, Hoshangabad, Madhya Pradesh
93	IC41996	Multai ,Betul, Madhya Pradesh
94	IC21930	Bichhiya, Vidisha, Madhya Pradesh
95	IC-41769	Ligtam, East Sikkim, Sikkim
96	IC-41767	S. Rigu,, East Sikkim, Sikkim
97	IC-41768	Khechepessi, west Sikkim, Sikkim
98	IC-21947	Gharghoda,, Raigarh, Sikkim
99	IC-21961	Marwahi, Bilaspur, Sikkim
100	IC-21967	Gharghoda, Raigarh, Sikkim
101	IC-21973	Raigarh, Sikkim
102	IC-21960	Marwahi, Bilaspur, Sikkim
103	IC-35774	Isaran, Ahmdabad, Gujarat
104	IC-35780	Isaran, Ahmdabad, Gujarat
105	IC-35777	Isaran, Ahmadabad, Gujarat
106	IC-35775	Isaran, Ahmadabad, Gujarat
107	IC-35776	Isaran,Ahmadabad, Gujarat
108	IC-35770	Isaran, Ahmadabad, Gujarat
109	IC-35520	Salai, Mehsana ,Gujarat
110	IC -20303	Kruin, Mawryngkneng,,East Khasi Hills, Meghalaya
111	IC-20304	Puringmawryngkneng, East Khasi Hills, Meghalaya
112	IC-20306	Zigzag, East Garo Hills, Meghalaya
113	IC-258370	Norbugh, Badgam, Jammu & Kashmir
114	IC-361853	Thakur, Kapran,, Anantnag, Jammu & Kashmir
115	IC-328559	Bani, Bashli, Kathua, Jammu & Kashmir

116	IC-26270	Malkapur, BIDAR, Karnataka
117	IC-26266	Bijapur, Bijapur, Karnataka
118	IC-26272	Nanora, Bicholim, North Goa, Goa
119	IC-396983	Ladpura,, Bhilwara, Rajasthan
120	IC-21939	Tapkara,Raigarh, Chattisgarh
121	NAIP 42	Unknown
122	NAIP 46	Unknown
123	NAIP 49	Unknown
124	EC-338765	United States Of America
125	EC-146535	United States Of America
126	EC-146508	United States Of America
127	EC-146537	United States Of America
128	EC-146541	United States Of America
129	EC-146536	United States Of America
130	EC-146506	United States Of America
131	EC-146510	United States Of America
132	EC-146509	United States Of America
133	EC-146539	United States Of America
134	EC- 146543	United States Of America
135	EC-146493	United States Of America
136	EC-146494	United States Of America
137	EC-146533	United States Of America
138	EC-146538	United States Of America
139	EC-146542	United States Of America
140	EC-146532	United States Of America
141	EC-146540	United States Of America
142	EC-146507	United States Of America
143	EC145973	United States Of America
144	EC146534	United States Of America

145	EC-289386	United States Of America
146	EC-289393	United States Of America
147	EC-328885	Germany
148	EC328875	Germany
149	EC-328873	Germany
150	EC-328896	Germany
151	EC-328883	Germany
152	EC-328881	Germany
153	EC-328880	Germany
154	EC-328878	Germany
155	EC-328877	Germany
156	EC-328874	Germany
157	EC-328879	Germany
158	EC-328876	Germany
159	EC-328886	Germany
160	EC-328891	Germany
161	ES- 519526	Russia
162	EC519523	Russia
163	EC-519527	Russia
164	EC-519512	Russia
165	EC-519511	Russia
166	EC-519510	Russia
167	EC-519509	Russia
168	EC-519541	Russia
169	EC-519542	Russia
170	EC-519544	Russia
171	EC-345800	China
172	EC-120051	China
173	EC-120048	China

174	EC-120049	China
175	EC-120052	China
176	EC-120050	China
177	EC-345797	China
178	EC-333744	Latvia
179	EC-333747	Latvia
180	EC333745	Latvia
181	EC333748	Latvia
182	EC-150190	Netherlands
183	EC-150191	Netherlands
184	EC-150192	Netherlands
185	EC-32881	France
186	EC-351945	France
187	EC-157314	Poland
188	EC-157313	Poland
189	EC-18862	Nepal
190	EC-151544	Benin
191	EC-223672	Kenya
192	EC-583624	South Africa

Table 2. Single-nucleotide polymorphism statistics in the entire dataset showing number and percentage frequency of allele occurrence, where (A/C) indicate that A is a reference allele and C is the alternate allele.

S.No.	Alleles	Allele Occurrence	Percentage frequency of allele occurrence
1	G:C	10041	23.946
2	C:G	9938	23.701
3	C:T	3869	9.227
4	G:A	3823	9.117
5	T:C	3423	8.163
6	A:G	3413	8.14
7	T:A	1553	3.704
8	A:T	1514	3.611
9	G:T	1151	2.745
10	C:A	1114	2.657
11	A:C	1075	2.564
12	T:G	1017	2.425

Figures



Figure 1

Global geographical distribution of 192 accession of *Amaranthus hypochondriacus* used in this GBS study.

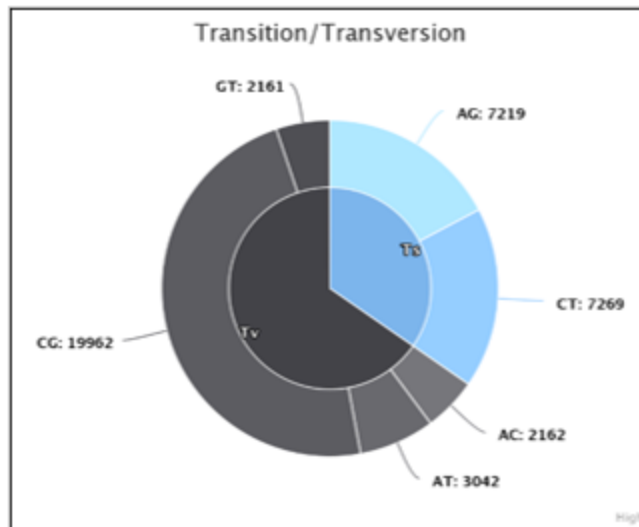


Figure 2

The rates of various substitution mutations grouped into transitions and transversions.

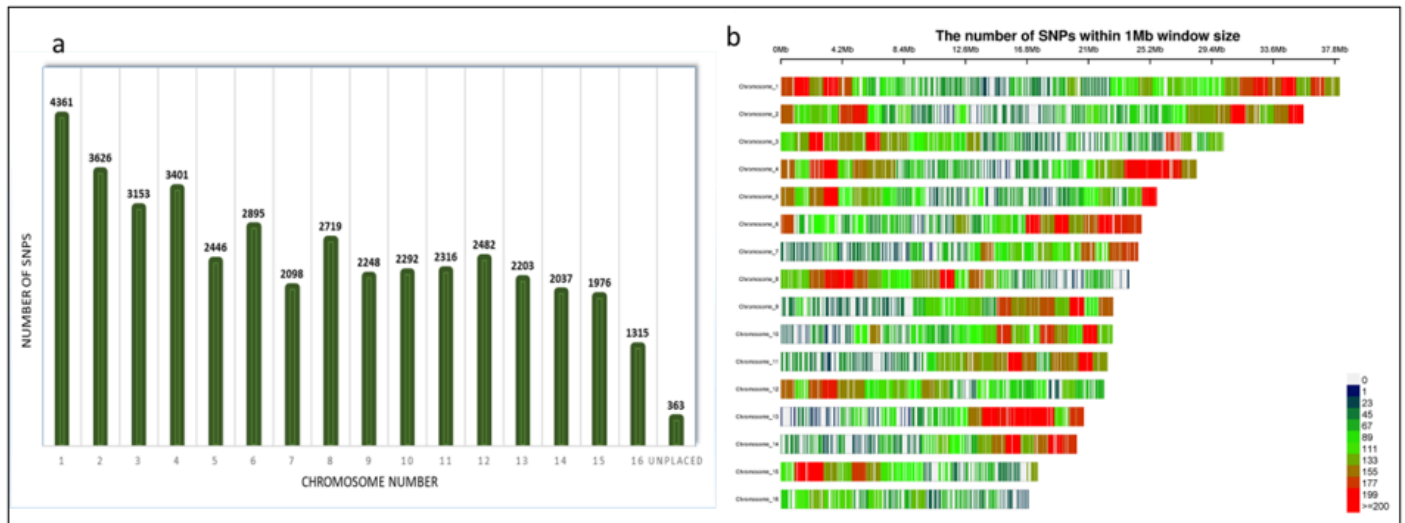


Figure 3

(a) The total number of SNPs in each chromosome of *A. hypochondriacus*.(b) Heatmap portraying the densities of SNPs on individual chromosomes within one Mb window size.

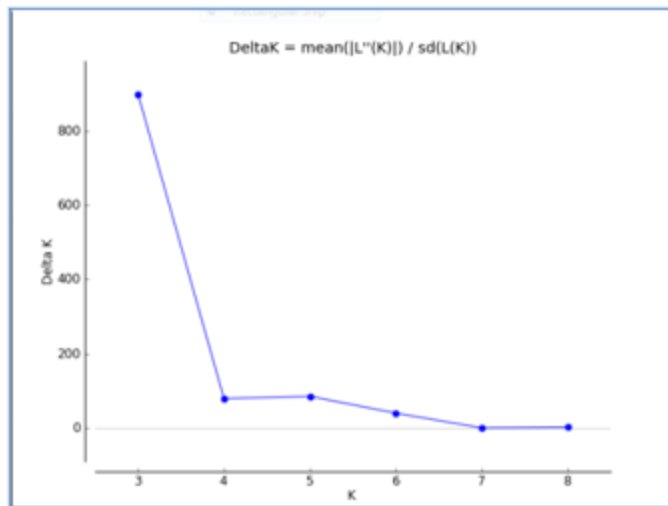


Figure 6

Identification of the appropriate subpopulation number (K): Subpopulation number (K) against delta K and the maximum K value observed at K=3

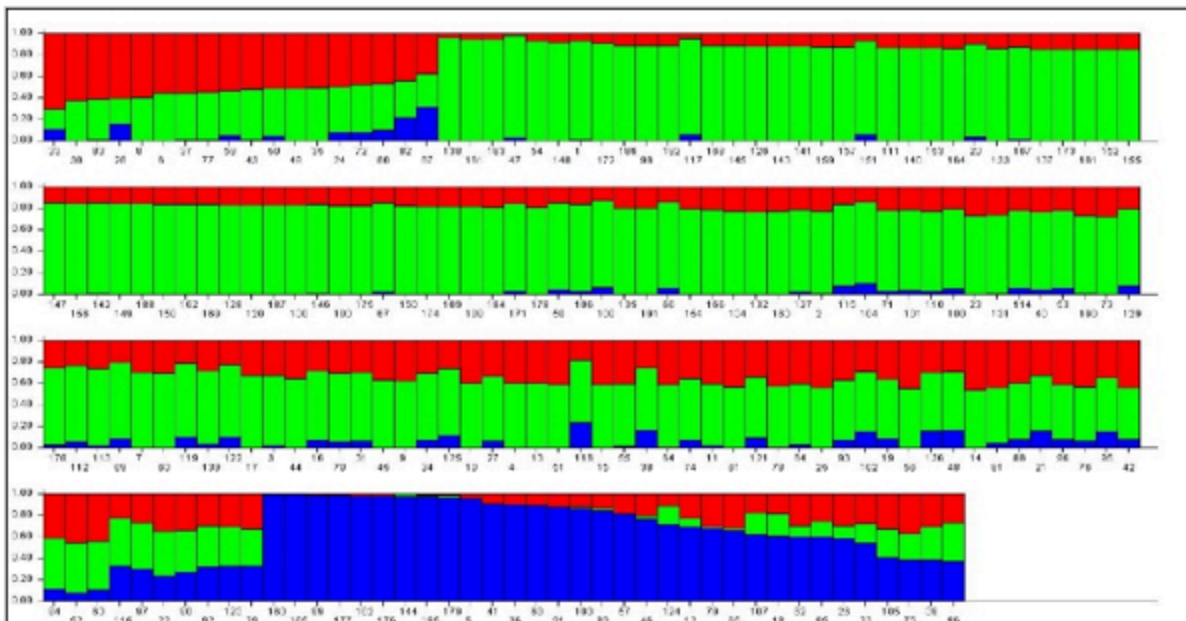


Figure 7

Proportional allocations of each of 192 *Amarathus hypochondriacus* accessions to K ancestral gene pools inferred using ADMIXTURE (Alexander et al., 2015), under K = 3, shown as a bar plot.

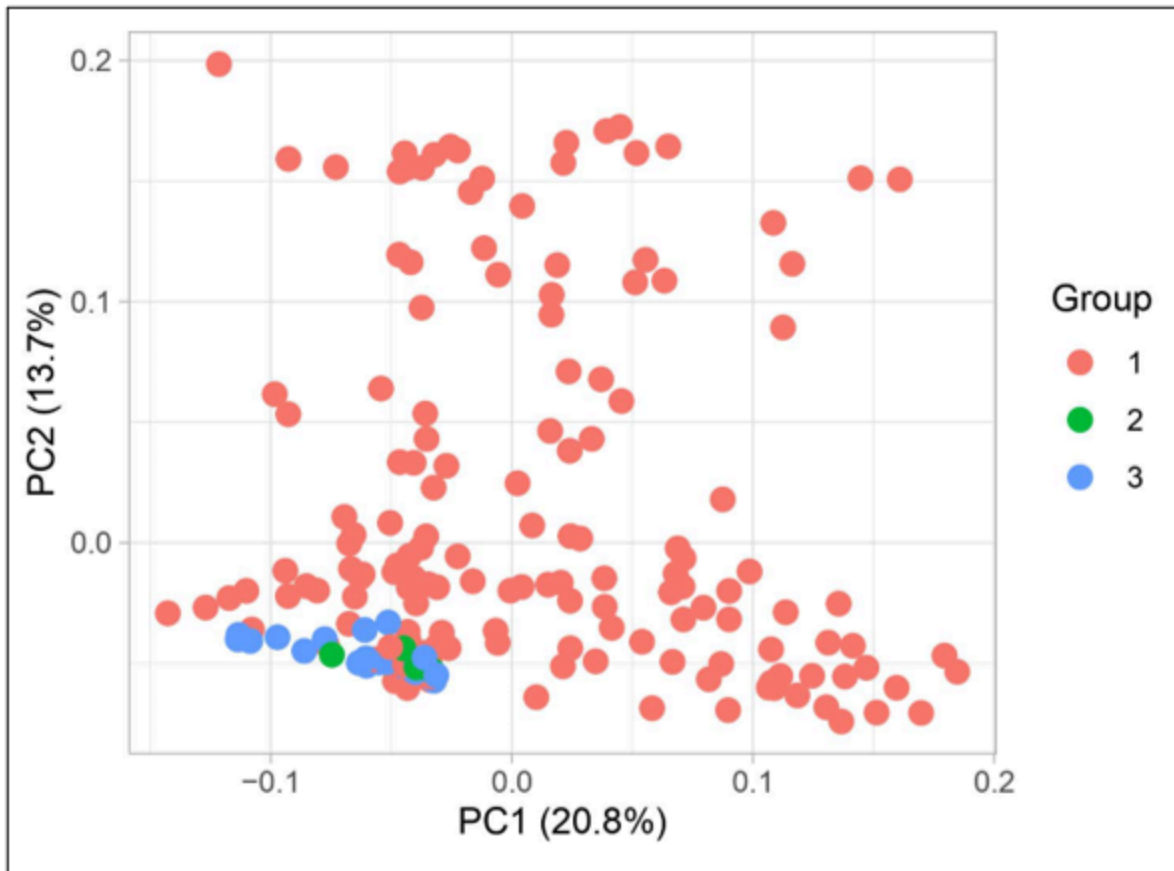


Figure 8

2-D PCA of 192 *Amarathus hypochondriacus* accession based on 41,931 filtered single nucleotide polymorphisms. The axes represent the first two principal components (PCs).

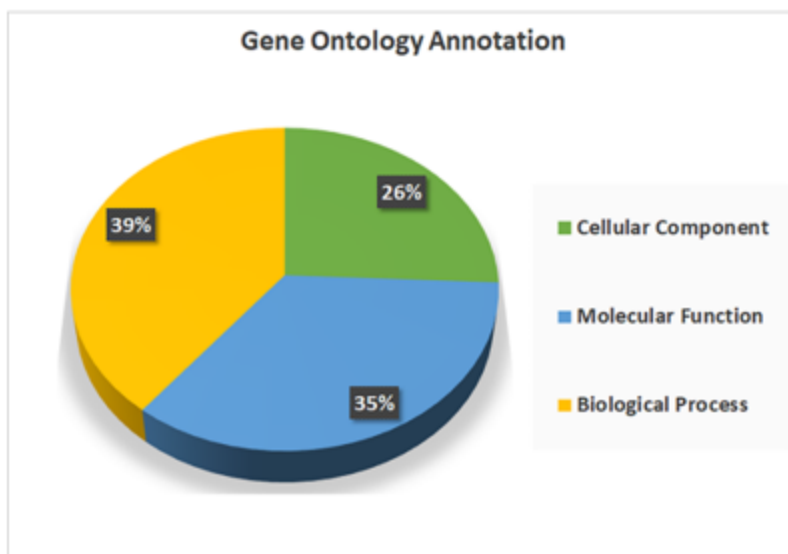


Figure 9

Gene ontology annotation categorization.

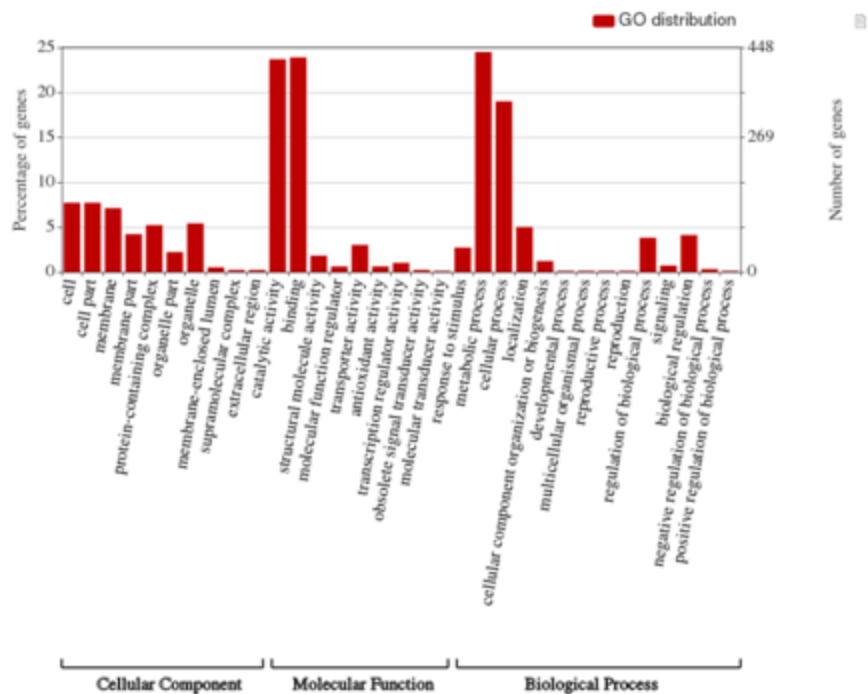


Figure 10

Gene ontology (GO) term assignment of *A. hypochondriachus* single-nucleotide polymorphism carrying genes classified into three main categories: biological process, cellular component, and molecular function.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTableS1.docx](#)