

# Genetic diversity among cultivated beets (*Beta vulgaris*) assessed via population-based whole genome sequences

Paul J. Galewski (New Corresponding Author) (✉ [paulgalewski@gmail.com](mailto:paulgalewski@gmail.com))

J. Mitchell McGrath (Former Corresponding Author)

USDA-ARS Midwest Area <https://orcid.org/0000-0002-7365-9901>

---

## Research article

**Keywords:** sugar beet, table beet, fodder beet, leaf beet, chard, genome wide analysis, crop diversity, crop differentiation

**Posted Date:** January 10th, 2020

**DOI:** <https://doi.org/10.21203/rs.2.15867/v3>

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

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on March 2nd, 2020. See the published version at <https://doi.org/10.1186/s12864-020-6451-1>.

# Abstract

Background: Diversification on the basis of utilization is a hallmark of *Beta vulgaris* (beet), as well as other crop species. Often, crop improvement and management activities are segregated by crop type, thus preserving unique genome diversity and organization. Full interfertility is typically retained in crosses between these groups and more traits may be accessible if the genetic basis of crop type lineage were known, along with available genetic markers to effect efficient transfer (e.g., via backcrossing). *Beta vulgaris* L. ( $2n = 18$ ) is a species complex composed of diverged lineages (e.g., crop types), including the familiar table, leaf (chard), fodder, and sugar beet crop types. Using population genetic and statistical methods with whole genome sequence data from pooled samples of 23 beet cultivars and breeding lines, relationships were determined between accessions based on identity-by-state metrics and shared genetic variation among lineages. Results: Distribution of genetic variation within and between crop types showed extensive shared (e.g. non-unique) genetic variation. Lineage specific variation (e.g. apomorphy) within crop types supported a shared demographic history within each crop type, while principal components analysis revealed strong crop type differentiation. Relative contributions of specific chromosomes to genome wide differentiation were ascertained, with each chromosome revealing a different pattern of differentiation with respect to crop type. Inferred population size history for each crop type helped integrate selection history for each lineage, and highlighted potential genetic bottlenecks in the development of cultivated beet lineages. Conclusions: A complex evolutionary history of cultigroups in *Beta vulgaris* was demonstrated, involving lineage divergence as a result of selection and reproductive isolation. Clear delineation of crop types was obfuscated by historical gene flow and common ancestry (e.g. admixture and introgression, and sorting of ancestral polymorphism) which served to share genome variation between crop types and, likely, important phenotypic characters. Table beet was well differentiated as a crop type, and shared more genetic variation within than among crop types. The sugar beet group was not quite as well differentiated as the table beet group. Fodder and chard groups were intermediate between table and sugar groups, perhaps the result of less intensive selection for end use.

## Background

*Beta vulgaris* L. (beet) is an economically important plant species consisting of several distinct cultivated lineages (*B. vulgaris subsp. vulgaris*) These lineages, or “crop types,” include sugar beet, table beet, fodder beet, and chard. The crop types have been adapted for specific end uses and thus exhibit pronounced phenotypic differences. Crop type lineages breed true, indicating a genetic basis for these phenotypes. Cultivated beets likely originated from wild progenitors of *B. vulgaris subsp. maritima*, also called “sea beet” (Biancardi et al. 2012). It is widely accepted that beet populations were first consumed for leaves. The earliest evidence for lineages with expanded roots occurs in Egypt around 3500 BC. The root types and the origin of the enlarged root is thought to have occurred in the Near East (Iraq, Iran, and Turkey) and spread west (Europe) (Zossimovich 1940). Interestingly, beet production for roots as an end use was first described along trade routes across Europe. Historically, Venice represented a major European market of the Silk Road and facilitated the distribution of eastern goods across Europe

(Kuzmina 2008). Table beet has been proposed to have been developed within Persian and Assyrian gardens (Goldman and Navazio 2002). Whether this specifically corresponds to the origin of the expanded root character or a restricted table beet phenotype remains unknown. In fact, early written accounts regarding the use of root vegetables often confused beet with turnip (*Brassica rapa*).

Hybridization between diverged beet lineages has long been recognized as a source of genetic variability available for the selection of new crop types and improving adaptation (Schukowsky 1950 cited in Winner et al. 1993, Cooke and Scott 1993). In 1747, Margraff was the first to recognize the potential for sucrose extraction from beet. Achard, a student of Margraff, was the first to describe specific fodder lineages that contained increased quantities of sucrose and the potential for an economically viable source of sucrose for commoditization (Winner 1993). In 1787, Abbe de Commerell suggested red mangle (fodder) resulted from a red table beet/chard hybrid and that the progenitors of sugar beet arose from hybridizations between fodder and chard lineages (Fischer 1989, Ford-Lloyd 1995). Louise de Vilmorin (1816-1860), a French plant breeder, first detailed the concept of progeny selection in sugar beet, a method of evaluating the genetic merit of lineages based on progeny performance (Gayon and Zallen 1998). Vilmorin used differences in specific gravity as a measure to select beet lineages and increase sucrose content. This approach led to increases in sucrose concentration from ~4% in fodder beet to ~18% in current US hybrids (reviewed in McGrath and Fugate 2012).

*B. vulgaris* is a diploid organism ( $2n = 18$ ) with a predicted genome size of 758 Mb (Arumuganathan and Earle 1991). Chromosomes at metaphase exhibit similar morphology (Paesold et al. 2012). The first complete reference genome for *B. vulgaris* (e.g., RefBeet) provided a new perspective regarding the content of the genome (e.g., annotated gene models, repeated sequences, and pseudomolecules) (Dohm et al. 2014). This research confirmed whole genome duplications and generated a broader view of genome evolution in the Eudicots, Caryophyllales, and *Beta*. The EL10.1 reference genome (Funk et al., 2018) represents a contiguous chromosome scale assembly resulting from a combination of PacBio long-read sequencing, BioNano optical mapping and Hi-C linking libraries. Together, EL10.1 and RefBeet provide new opportunities for studying the content and organization of the beet genome. Resequencing of important beet accessions has the potential to characterize the landscape of variation and inform recent demographic history of beet, including the development of crop types and other important lineages.

Population genetic inference leveraging whole genome sequencing (WGS) data have proven powerful tools for understanding evolution from a population perspective (Storz 2005, Lynch 2009, Casillas and Barbadilla 2017). Knowledge of the quantity and distribution of genetic variation within a species is critical for the conservation and preservation of genetic resources in order to harness the evolutionary potential required for the success of future beet cultivation. Recent research has revealed the complexity of relationships within *B. vulgaris* crop types (Andrello et al. 2017). Studies have shown sugar beet is genetically distinct and exhibits reduced diversity compared to *B. vulgaris subsp. maritima*. Geography and environment are major factors in the distribution of genetic variation within sugar beet accessions in the US (McGrath et al. 1999). Furthermore, spatial and environmental factors were evident in the complex

distribution of genetic variation in wide taxonomic groups of *Beta* (Andrello et al. 2016), which include the wild progenitors of cultivated beet.

Here we present a hierarchical approach to characterize the genetic diversity of cultivated *B. vulgaris* using pooled sequencing of accessions representing the crop type lineages. These accessions contain a wide range of phenotypic variation including leaf and root traits, distinct physiological/biochemical variation in sucrose accumulation, water content, and the accumulation and distribution of pigments (e.g., betaxanthin and betacyanin). These phenotypic traits, along with disease resistance traits, represent the major economic drivers of beet production. Developmental genetic programs involved in cell division, tissue patterning, and organogenesis likely underlie the differences in root and leaf quality traits observed between crop types. Improvement for these traits as well as local adaptation and disease resistance occurs at the level of the population. Pooled sequencing provides a means to characterize the diversity of important beet lineages and survey the nucleotide variation, which has utility in marker-based approaches across a diverse community of breeders and researchers interested in *B. vulgaris*. Pooled sequencing works in synergy with both the reproductive biology of the crop as well as the means by which phenotypic diversity is evaluated (e.g., population mean phenotype) and beets are improved through selection. The genetic control of important beet traits, currently unknown, will help prioritize existing variation and access novel sources of trait variation in order to address the most pressing problems related to crop productivity and sustainability.

## Results

Twenty-five individuals from each of the 23 *B. vulgaris* accessions were chosen to represent the cultivated *B. vulgaris* crop types (Table 1 and Figure 1). Leaf tissue was pooled, DNA extracted and sequenced using the Illumina 2500 in paired end format. On average,  $61.84 \pm 12.22$  GB of sequence data was produced per accession, with an average depth of 81.5X. After processing for quality, reads were aligned to the EL10.1 reference genome. Approximately 20% of bases were discarded owing to trimming of low-quality base calls and adapter sequences. Biallelic SNP and lineage-specific variants were used to estimate the quantity and organization of genome-wide variation within these *B. vulgaris* populations and groups (e.g., species, crop types, and accessions). On average 90.74% of the filtered reads aligned to the EL10.1 reference genome. A total of 14,598,354 variants were detected across all accessions, and 12,411,164 (85.0%) of these were classified as a SNP, and of these 10,215,761 (82.3%) were biallelic. Thus, most SNP variants appeared to be biallelic, as only 2,718,205 (18.6%) of the SNP variants were characterized as multiallelic. After filtering for read depth ( $n \geq 15$ ), 8,461,457 biallelic SNPs remained for computational analysis. Insertions and deletions (indels) were called using GATK (370,260) (Table 2), which served to reduce false variants resulting from misalignments. This represented a large reduction from the 2,187,190 indels called using the bcftools pipeline.

AMOVA was performed in order to quantify the distribution of variation within and among cultivated *B. vulgaris* crop types. The results showed no strong population subdivision with respect to crop type. The variation shared among crop types (99.37%), far exceeded the variation apportioned between crop type

lineages (0.40%). The variation detected between accessions within a crop type was also low (0.23%) (Table 3). This result suggested a small proportion of the total variation is unique to any given accession. This was confirmed by the low quantity of lineage-specific variation (LSV) detected, evaluated in a hierarchical fashion. Lineages were defined as individual accessions, crop types, and species (Table 2). In total, 600,239 variants (4.0%) were unique and fixed within a single accession. The accumulation of variation on specific chromosomes for each accession was informative (Table 4). Individual accessions of sugar beet contained a large quantity of LSV on Chromosome 6 relative to other sugar beet chromosomes and indicated that either divergent selection or drift has occurred on this sugar beet chromosome. The variety, 'Bulls Blood' (BBTB), contained the greatest amount of LSV detected, 8,893 indels and 79,236 SNP variants (Table 2). Table beet accessions contained the most LSV overall which suggested Table Beet is the most divergent of the crop type (Table 4).

Within the crop types, 10,661 variants were crop type specific and were not found within any other crop type. Of these, 8,098 were characterized as SNPs and 1,963 as indels. The number of SNP LSV detected within sugar beet, table beet, fodder beet, and chard crop types were as follows: 3,317, 1,379, 643, and 3,359, respectively (Table 2). Indel LSV detected for the crop types were 342, 558, 205, and 858, respectively (Table 2). Diversity contained within the species, crop type, and individual accessions was estimated using expected heterozygosity ( $2pq$ ) (Table 2 and Figure 2). Expected *heterozygosity* ( $2pq$ ) varied from 0.027 in our inbred reference EL10 sugar beet accession to 0.253 in the recurrent selection sugar beet breeding population GP9. Within the crop types, the mean expected *heterozygosity* for sugar beet was 0.207, table beet = 0.147, fodder beet = 0.221, and chard = 0.216 (Table 2). Interestingly, chard contained the most LSV of the crop types yet showed high diversity ( $2pq$ ), suggesting unique variation supports the divergence of this lineage.

The expected heterozygosity ( $2pq$ ) for accessions such as EL10 and W357B was low. This was expected owing to inbreeding via the presence of self-fertility alleles in these two accessions. These accession EL10 was excluded from further analysis due to the fact that the sequence data was derived from a single individual. Interestingly, the variety 'Bulls Blood' lacked variation relative to other beet accessions, and it is likely that recent selection underlies this result (Chris Becker, personal communication). The variation in diversity estimates as measured by expected heterozygosity ( $2pq$ ) suggested the level of diversity is highly dependent on the breeding system, selection for end use traits and  $N_e$  size.

The variation detected was used to cluster accessions in two ways: (1) a hierarchical clustering based on relationship coefficients estimated using the quantity of shared variation between accessions, and (2) a principal components analysis using allele frequency in each accession, estimated using an IBS (Identity by State) approach. The resulting dendrogram and heatmap showed that the table beet crop type was the only group to have strong evidence (e.g., high relationship coefficients and bootstrap values) supporting it as a unique group harboring significant variation (Table 5). Likewise, the green (LUC and FGSC) and red (RHU and Vulcan) chard accessions showed evidence for two distinct groups (Figure 3). Sugar beet lineages with known pedigree relationships and high probability for shared variation (e.g., SR98/2 and EL51) also had strong evidence, which supports the delineation of population structure on the basis of

shared variation. Additionally, the clade composed of SP7322, SR102, GP10, and GP9 resolved in a similar fashion.

PCA used genome-wide allele frequency estimates for individual accessions. The first principal component (PC1) explained 75.6% of the variance in allele frequency and separated the table beet crop type from the other crop types. The second component (PC2) explained 15.25% of the variance (Figure 4). Sugar and table beets appeared the most divergent and were able to be separated along both dimensions. Chard and fodder crop types were distinguishable but appeared less divergent. Allele frequency estimates analyzed on a chromosome-by-chromosome basis demonstrated that specific chromosomes cluster the accessions by crop type (Figure 5). Chromosomes 3, 8, and 9 appear to be important for the divergence between sugar beet and other crop types. All chromosomes were able to separate table beet with the exception of Chromosomes 7 and 9.

Finally, using our population genomic data we tested a composite likelihood method to estimate historical effective population size ( $N_e$ ) to infer demographic histories for crop type lineages. Table beet appears to have a distinct history in this respect as well as one or more demographic separations when compared with the other three lineages. Trends in historical effective population sizes ( $N_e$ ) for fodder and sugar groups were quite similar to each other, and no early divergence was detected between them. The chard group appeared to share early demographic history with the fodder/sugar group but showed a different trend later, suggesting it diverged early with respect to the other crop types (Figure 6). The demographic history of *B. vulgaris* crop type correlates well with historical evidence (e.g., records of antiquity, archeological evidence, and scientific literature) detailing the development of distinct crop type lineages (Table 6).

## Discussion

The accessions sampled here represent divergent lineages used in the cultivation of beet. All have notable breeding histories, which has served to capture and fix genetic variation resulting in predictable phenotypes characteristic of each lineage (e.g. accession or crop type). The organization and distribution of genetic variation within and between accessions reflects the historical selection and evolutionary pressures experienced as these crop types and varieties were developed. Pooled sequencing allowed us to make the cogent genomic comparisons that informs the history of beet development, from ancestral gene pools and domestication to the development of varieties and germplasm within modern breeding programs. Using population genomic data, we were able to support *B. vulgaris* as a species complex, uncover genomic variation associated with development of beet crop types, and gain fundamental insight into the natural history of beet.

Two biological groups could be identified with high confidence using these data, a table beet group and a group encompassing chard, fodder beet, and sugar beet. Previous research, which used genetic markers to cluster crop types, reported similar findings (Mangin et al. 2015, Andrello et al. 2016). The strong evidence for a unique table beet group hints at both genetic drift, resulting from reproductive isolation, as

well as positive selection for end use (Figure 3, Figure 4, Figure 6). In general, selection and drift act to change allele frequency within a population (Hedrick 2005), but the effects are relative to the effective population size ( $N_e$ ) of the populations under selection. Effective population size is an important consideration because it relates to the standing genetic diversity within populations (Crow and Denniston 1988, Waples 1990). The patterns of variation resulting from drift and selection are distinct. For example, table beet accessions had low diversity ( $2pq$ ) relative to other crop types (Table 2), and the ability to separate table beet accessions on the basis of allele frequency is suggestive of selection (Figure 4 and 5). Relationship coefficients, on the other hand, highlight the differences in the quantity of shared variation within and between crop types (Table 5 and Figure 3), suggesting table beet may have been less connected to other crop types historically. Allele frequency showed signals of differentiation distributed across all chromosomes for table beet (Figure 5), likely reflecting both selection and drift. The low quantity of shared variation between crop types did not support long term phylogeographic explanations for the differentiation observed. Long periods of geographic isolation can produce barriers to reproduction, further reinforcing isolation and divergence of populations (Palumbi 1994). This appears not to be the case in cultivated beet, as experimental hybrids between crop types show few barriers to hybridization and produce viable progeny, which does not suggest a large degree of chromosomal variation between the groups. The creation of segregating populations from crosses between sugar and table beet crop types support this observation (McGrath et al. 2007, Laurent et al. 2007).

The lesser degree of separation between chard, fodder, and sugar crop types may be the result of increased connectivity (e.g., historical gene flow) between these lineages versus table beet. High gene flow exerts a homogenizing effect on the diversity contained within populations and increases the quantity of shared variation. This may explain a lack of clear delineation of these crop types using genome-wide markers. Fodder and sugar crop types could be separated using allele frequency (Figure 4) but clusters based on shared variation were less clear (Figure 3). This was not unexpected given the known history between these lineages. The development of fodder lineages that accumulate sucrose have occurred in recent history (~200 years), giving rise to the progenitor of sugar beet, the 'White Silesian' (Fischer 1989, Winner 1993). Phenotypic divergence between species is attributed more to indel variation than to SNP variation owing to their greater consequences on gene expression and gene regulation (Chen et al. 2009). This phenomenon may be visible in population divergence as well as speciation. The high quantity of shared variation between sugar and fodder crop types (Table 5) and the low quantity of indel LSV detected within sugar and fodder crop types (Table 2) likely reflects a shared demographic history relative to comparisons between other crop types (Figure 6). Interestingly, chard contained the most LSV of the crop types yet showed high diversity ( $2pq$ ), suggesting some unique variation supports the divergence of this lineage. The larger quantity of shared variation between the sugar beet, fodder beet, and chard crop types versus table beet (Table 5) suggests differences in the extent and timing of gene flow between lineages.

Chard is hypothesized as the first crop type developed from diverse ancestral *B. vulgaris subsp. maritima* populations (Biancardi et al. 2012, Winner 1993). This is supported by the high level of diversity ( $2pq$ ) (Table 2 and Figure 2), a high quantity of LSV (Table 2), and interesting trends in the demographic history

(Figure 6). The clear delineation of two distinct chard groups (Figure 3) suggests major differences in genome composition between the two groups and a unique demographic history for each chard lineage. The chards share similar leaf morphology but the roots of the red chard group were enlarged and had fewer 'sprangles' (e.g. adventitious roots branching from the tap root) relative to the green chard accessions but not to the extent as the root types (e.g. sugar, fodder, and table). This may reflect introgressions between the red chard and a root type, and potentially an unintended consequence of chard improvement for color traits.

The enlarged tap root character appears to have been first developed in table beet lineages (Biancardi et al. 2012), but the expanded root character is shared across crop type lineages. This suggests several hypotheses: (1) the root character in fodder beet reflects the introgression of this character from a table beet to a chard background and represents a single source for this character (Zossimovich 1940), (2) an ancestral population gave rise to the root character and diverged into fodder and table lineages, (3) the enlargement evolved several times and contributes to the diversity in shape and form. Historically, it appears admixture, hybridization, and introgression were fundamental to the development of beet lineages. Schukowsky (1950) suggested that the broad adaptation of beet to novel growing environments may be due to variation accumulated in geographically diverse ancestral populations and shared via admixture and gene flow between lineages. Trait variation in wild relatives is becoming increasingly important for crop adaptation to a changing growing environment (Takuno et al. 2015). Distinguishing between sorting ancestral variation and introgression events remains a challenge in population genomic analysis but could yield important insight into beet crop type development, and other cultivated species as well.

The beet crop types have appeared to have diverged by selection. The variation in allele frequency of bi-allelic SNPs for beet accessions was able to distinguish the crop types (Figure 4). This suggests that the allele frequency data contains signal related to historical selection (Figure 5). Sugar and table beet appear to be the most diverged, which is consistent with large breeding efforts for each of these crop types. Allele frequency data analyzed on a per chromosome basis demonstrated that only specific chromosomes can differentiate on the basis crop type. Ostensibly the presence of variation located on specific chromosomes is under positive selection for end use, leading to an accumulation of lineage-specific differences including those linked to defining phenotypic characters. In fact, many quantitative trait loci studies support the fact that specific regions along chromosomes contain the variation that ultimately influences phenotype (Doerge 2002). Interestingly, even small amounts of variation can have profound effects on phenotypic variation (Doebley and Stec 1993, Meyer and Purugganan 2013). Allele frequency estimates for specific chromosomes as well as the variation in lineage-specific variation for crop type on specific chromosomes suggests a small degree of total genome variation explains beet crop type differences. Given the support for crop type relationships based on allele frequency and degree of shared variation, it appears the divergence of beet crop types occurred in the presence of high gene flow. Population divergence in the presence of gene flow produces distinct patterns of variation with respect to selection (Martin et al. 2013). Cryptic relationships within other species complexes have been explained by various models including the islands-of-differentiation model (Waples 1998, Bickford et al. 2007).

Admixture and introgression events may have served to share genetic variation across cultivated beet accessions and crop type lineages, which in turn, created challenges for the clear delineation of subpopulations. This is confounded by the fact that, as lineages evolve, a lesser quantity of variation with greater agricultural importance contributes to our notion of economic and agronomic value. Resolving the degree to which historical admixture and introgression has contributed to the development of beet crop type will require more in-depth analysis of the variation at nucleotide level within local chromosome regions.

## Conclusions

Beet crop types are important lineages which exhibit both genetic and phenotypic divergence. Sufficient support for treatment of these groups as significant biological units was present from *de novo* clustering of beet accessions. It would appear selection for end use qualities and genetic drift were major factor in the observed differentiation between lineages and explains the apportionment of genetic variation between crop types at distinct chromosome locations. Common ancestry and admixture and introgression likely maintained levels of genetic variation between crop types and reflects a complex demographic history between crop types. The majority of genetic variation detected in beet crop types was biallelic SNPs, but lineage specific variation may have had a greater role in crop diversification, with table beet showing the greatest degree of differentiation. Most variation is held within the species (as represented by the crop types here), and only a small amount of the total variation is partitioned within individual crop types. Understanding the history of beet crop type diversification, in terms of the evolution of genomes and traits within and between crop types, will help to identify and recover a genetic basis for crop type phenotypes. Directed molecular breeding approaches may be developed to incorporate novel traits from other crop types and wild populations.

## Methods

### *Beta vulgaris* accessions and sequencing

Twenty-three beet accessions were sequenced to 80X coverage relative to the predicted 758 Mb *B. vulgaris* genome using a pooled sequencing approach. The accessions are representative of the four recognized crop types and capture the range of phenotypic diversity found within cultivated beet (Table 1). Accessions were grown in the greenhouse and leaf material was harvested from 25 individuals per accession. Leaf material, one young expanding leaf of similar size from each individual within an accession, was combined, homogenized, and DNA was extracted using the Macherey-Nagel NucleoSpin Plant II Genomic DNA extraction kit (Bethlehem, PA). Libraries were prepared using the Illumina TruSeq DNA Nano Library Preparation Kit. Libraries were QC'd and quantified using a combination of Qubit dsDNA HS, Caliper LabChipGX HS DNA and Kapa Biosystems Illumina Library Quantification qPCR assays. Each set of 8 libraries were pooled in equimolar amounts. Each of these pools was loaded on four (4) lanes of an Illumina HiSeq 2500 High Output flow cell (v4). Sequencing was done using HiSeq SBS reagents (v4) in a 2x125bp paired end format. Base calling was performed by Illumina Real Time

Analysis (RTA) v1.18.64 and output of RTA was demultiplexed and converted to FastQ format with Illumina Bcl2fastq v1.8.4. The resulting reads were assessed for quality using FastQC (Andrews 2010), library bar-code adapters were removed, and reads were trimmed according to a quality threshold using TRIMMOMATIC (Bolger et al. 2014) invoking the following options (ILLUMINACLIP:adapters.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). These filtered reads were used for downstream analysis.

### *Data processing and variant detection*

Variants for each accession were called by aligning the filtered reads to the EL10.1 reference genome assembly (Funk et al. 2018) using bowtie2 v2.2.3 (options -q -phred33-quals -k 2 -x) (Langmead and Salzberg 2012). The resulting alignment files were sorted and merged using SAMtools version 0.1.19 (Li et al. 2009). SNP variants were called for each accession using BCFtools (Li 2011), filtered for mapping quality (MAPQ >20) and read depth ( $n > 15$ ), and then combined using VCFtools (Danecek et al. 2011). The combined data was again filtered to obtain biallelic sites across all accessions. Indels were evaluated using the Genome Analysis Toolkit (GATK) haplotype caller (McKenna et al. 2010) following best practices (<https://www.broadinstitute.org/partnerships/education/broade/best-practices-variant-calling-gatk-1>). The 'mpileup' subroutine in SAMtools was then used to quantify the alignment files and extract allele counts. Allele frequency was estimated within each accession for SNP loci identified as biallelic across all accessions. Population parameters were then estimated using allele frequencies within each accession such that ( $p + q = 1$ ), where p was designated as the allele state of the EL10.1 reference genome and q, the alternate, detected in each sequenced accession. Expected heterozygosity ( $2pq$ ), also termed gene diversity (Nei 1987), was used to compare diversity contained within each accession.

### *AMOVA*

Analysis of molecular variance (AMOVA) was used to assess the distribution of genetic variation within the species (Excoffier et al. 1992). AMOVA was performed using the *ade4* package in R (Thioulouse et al. 1997) following the approach for pooled sequence data outlined in Gompert et al. (2010).

### *Crop type relationships*

Biallelic SNPs were used to calculate pairwise relationship coefficients between accessions using an identity by state (IBS) approach within the Kinship Inference for Association Genetic Studies (KING) package (Manichaikul et al. 2010). Neighbor joining trees were generated in order to extract bootstrap support along branches of our phylogram. In total 100 replications were used and analysis was carried out using the ape package (Analyses of Phylogenetics and Evolution) in R (Paradis and Schliep 2004).

### *Principle components analysis (PCA)*

PCA was carried out in R using singular value decomposition function, svd() in R.

### *Population size history*

Composite likelihood methods were used to estimate historical population sizes and infer demographic history from genome sequences of each accession using the program SMC++ version 1.12.1 (Terhorst et al. 2016) invoking the commands (smc++ estimate -o analysis/ 1.25e-8) to estimate historical population size and (smc++ split -o split/ pop1/model.final.json pop2/model.final.json) to estimate the joint demography between populations. A mutation rate of 1.25e-8 was assumed used based on the Arabidopsis mutation rate predicted to be between 10e-7 and 10e-8.

### *Lineage-specific variation*

Lineage-specific variation (LSV), defined as homozygous private variation (e.g., apomorphy), was extracted from the merged VCF file containing variants for all accessions. Variants that were fixed within a particular accession or assemblage of accessions (lineage), and not detected within any other lineage, were considered LSV. Variant files representing LSV were produced for each lineage in a hierarchical fashion (e.g., species, crop type and accessions). LSV was then evaluated with respect to lineage as well as its distribution along chromosomes.

## **List Of Abbreviations**

LSV - Lineage specific variation

WGS – Whole genome sequencing

SNP – Single nucleotide polymorphism

Indel – Insertion/deletion

*2pq* – expected heterozygosity

IBS - Identity by State

PCA – Principle components analysis

Ne - effective population size

GATK – Genome analysis toolkit

AMOVA - Analysis of molecular variance

## **Declarations**

### *Ethics approval and consent to publish*

This report does not involve the use of any animal or human data or tissue, and does not contain data from any individual person, and thus these aspects are not applicable.

### *Consent for publication*

Both authors consent to publication of this manuscript.

### *Availability of data and materials*

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. Whole genome sequences for the reported accessions have been deposited in NCBI under the BioProject Accessions PRJNA563463 (population whole genome sequences) and PRJNA413079 (EL10 genome assembly). Code available at [https://github.com/BetaGenomeNinja/BMC\\_Genetic-diversity\\_Beets](https://github.com/BetaGenomeNinja/BMC_Genetic-diversity_Beets) and data sets including vcf files and the allele frequency matrix is available via Data Dryad (<https://doi.org/10.5061/dryad.sbcc2fr2t>).

### *Competing interests*

The authors declare that they have no competing interests.

### *Funding*

Funding was provided by USDA-ARS CRIS 3635-21000-011-00D and the Beet Sugar Development Foundation, Denver, CO, USA, who had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

### *Authors' contributions*

PJG generated and analyzed the data and wrote the draft manuscript, and JMM conceived of the sequencing approach and sampling of representative accessions. PJG and JMM contributed equally to the interpretation of findings, and both authors read and approved the final manuscript.

### *Acknowledgements*

The authors would like to thank Andy Funk for his unswerving support and helpful discussions.

## **References**

- Andrello, M., K. Henry, P. Devaux, B. Desprez, and S. Manel, 2016 Taxonomic, spatial and adaptive genetic variation of *Beta* section *Beta*. *Theor. Appl. Genet.* 129: 257–271.
- Andrello, M., K. Henry, P. Devaux, D. Verdelet, B. Desprez *et al.*, 2017 Insights into the genetic relationships among plants of *Beta* section *Beta* using SNP markers. *Theor. Appl. Genet.* 130: 1857–1866.
- Andrews, S., 2010 FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>

- Arumuganathan, K., and E. D. Earle, 1991 Nuclear DNA content of some important plant species. *Plant Mol. Biol. Report.* 9: 208–218.
- Biancardi, E., L. W. Panella, and R. T. Lewellen, 2012 *Beta maritima: The origin of beets*. Springer, New York, NY.
- Bickford, D., D. J. Lohman, N. S. Sodhi, P. K. L. Ng, R. Meier *et al.*, 2007 Cryptic species as a window on diversity and conservation. *Trends Ecol. Evol.* 22: 148-155
- Bolger, A. M., M. Lohse, and B. Usadel, 2014 Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120.
- Casillas, S., and A. Barbadilla, 2017 Molecular population genetics. *Genetics* 205: 1003–1035.
- Chen, J. Q., Y. Wu, H. Yang, J. Bergelson, M. Kreitman *et al.*, 2009 Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* 26: 1523–1531.
- Cooke, D. A., and R. K. Scott, 1993 *The Sugar Beet Crop*. Chapman and Hall Publishers, London.
- Crow, J. F., and C. Denniston, 2006 Inbreeding and variance in effective population numbers. *Evolution* 42: 482–495.
- Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks *et al.*, 2011 The variant call format and VCFtools. *Bioinformatics* 27: 2156-2158.
- Doebley, J., and A. Stec, 1993 Inheritance of the morphological differences between maize and teosinte: Comparison of results for two F2 populations. *Genetics* 134: 559–570.
- Doerge, R. W., 2002 Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* 3: 43–52.
- Dohm, J. C., A. E. Minoche, D. Holtgräwe, S. Capella-Gutiérrez, F. Zakrzewski *et al.*, 2014 The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). *Nature* 505: 546–549.
- Excoffier, L., P. E. Smouse, and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics* 131: 479–491.
- Fischer, H. E., 1989 Origin of the “Weisse Schlesische Rübe” (white Silesian beet) and resynthesis of sugar beet. *Euphytica* 41: 75–80.
- Ford Lloyd, B. V., 1995 Sugarbeet, and other cultivated beets. In: *Evolution of crop plants* (J. Smartt & N. W. Simmonds, Eds.). Longman Scientific & Technical, Essex, U.K.

Funk, A., P. Galewski, and J.M. McGrath, 2018 Nucleotide-binding resistance gene signatures in sugar beet, insights from a new reference genome. *Plant J.* 95: 659-671.

Gayon, J., and D. T. Zallen, 1998 The role of the Vilmorin Company in the promotion and diffusion of the experimental science of heredity in France, 1840-1920. *Journal of the History of Biology* 31: 241-262.

Goldman, I. L., and J. P. Navazio, 2002 History and breeding of table beet in the United States. *Plant Breed. Rev.* 22: 357–388.

Gompert, Z., M. L. Forister, J. A. Fordyce, C. C. Nice, R. J. Williamson *et al.*, 2010 Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. *Molecular Ecology* 19: 2455–2473.

Hedrick, P., 2005 *Genetics of Populations*. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Kuzmina, E. E., 2008 *The Prehistory of the Silk Road*. University of Pennsylvania Press, Philadelphia, Pennsylvania.

Langmead, B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9: 357-359.

Laurent, V., Devaux, P., Thiel, T., Viard, F., Mielordt, S., Touzet, P., and Quillet, M., 2007 Comparative effectiveness of sugar beet microsatellite markers isolated from genomic libraries and GenBank ESTs to map the sugar beet genome. *Theor. Appl. Genet.* 115: 793-805.

Li, H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987-2993.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078-2079.

Lynch, M., 2009 Estimation of allele frequencies from high-coverage genome-sequencing projects. *Genetics* 182: 295–301.

Mangin, B., F. Sandron, K. Henry, B. Devaux, G. Willems *et al.*, 2015 Breeding patterns and cultivated beets origins by genetic diversity and linkage disequilibrium analyses. *Theor. Appl. Genet.* 128: 2255-2271.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale *et al.*, 2010 Robust relationship inference in genome-wide association studies. *Bioinformatics* 26: 2867-2873.

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters *et al.*, 2013 Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.* 23: 1817-1828.

McGrath, J. M., C. A. Derrico, and Y. Yu, 1999 Genetic diversity in selected, historical US sugarbeet germplasm and *Beta vulgaris* ssp. *maritima*. *Theor. Appl. Genet.* 98: 968–976.

- McGrath, J. M., D. Trebbi, A. Fenwick, L. Panella, B. Schulz *et al.*, 2007 An open-source first-generation molecular genetic map from a sugarbeet × table beet cross and its extension to physical mapping. *Crop Sci.* 47: S27-S44.
- McGrath, J.M., and K.K. Fugate, 2012 *Analysis of Sucrose from Sugar Beet*. In: *Preedy VR, editor. Dietary Sugars: Chemistry, Analysis, Function and Effects. Food and Nutritional Components in Focus No. 3.* (V. R. Preedy, Ed.). Royal Society of Chemistry Publishing, Cambridge, UK.
- McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20: 1297–1303.
- Meyer, R. S., and M. D. Purugganan, 2013 Evolution of crop species: Genetics of domestication and diversification. *Nat. Rev. Genet.* 14: 840-852.
- Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.
- Paesold, S., D. Borchardt, T. Schmidt, and D. Dechyeva, 2012 A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution. *Plant J.* 72: 600–611.
- Palumbi, S. R., 2003 Genetic divergence, reproductive isolation, and marine speciation. *Annu. Rev. Ecol. Syst.* 25: 547-572.
- Paradis, E., and Schliep K., 2018 ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35: 526-528.
- Schukowsky, P.M. (1950). *The Cultivated Plants and their Relatives* (in Russian). Moscow.
- Storz, J. F., 2005 Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol. Ecol.* 14: 671–88.
- Takuno, S., P. Ralph, K. Swart, R. J. Elshire, J. C. Glaubitz *et al.*, 2015 Independent molecular basis of convergent highland adaptation in maize. *Genetics* 200: 1297-1312.
- Terhorst, J., Kamm, J.A., Song, Y.S., 2016 Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* 49: 303–309.
- Thioulouse, J., D. Chessel, S. Dolédec, and J. M. Olivier, 1997 ADE-4: A multivariate analysis and graphical display software. *Stat. Comput.* 7: 75-83.
- Waples, R. S., 1990 Conservation genetics of Pacific salmon. II. Effective population size and the rate of loss of genetic variability. *J. Hered.* 81: 267–276.

Waples, R. S., 1998 Separating the wheat from the chaff: Patterns of genetic differentiation in high gene flow species. *J. Hered.* 89: 438–450.

Winner, C., 1993 History of the crop. In: *The Sugar Beet Crop* (D.A. Cooke and R. K. Scott, eds.) pp 1-35. Chapman and Hall Publishers, London.

Zossimovich, V. P., 1940 *Wild species and origin of cultivated beets.* p. 17-44. Sveklovodstvo.

## Tables

Due to technical limitations, the tables have been placed in the Supplementary Files section.

Table 1: List of materials for sequencing.

Table 2: SNP and Indel variation in cultivated *B. vulgaris*. Gene diversity ( $2pq$ ) indicates the diversity and expected genetic variation within populations.

Table 3: Analysis of molecular variance (AMOVA).

Table 4: Number of lineage-specific SNP and indel variants along chromosomes.

Table 5: Pairwise relationship matrix. Relationship coefficients are indicated above the diagonal, the number of shared variants is indicated below the diagonal, and the number of variants is given on the diagonal.

Table 6: Historical time line highlighting evidence of beet utilization.

## Supplemental File Legends

Supplemental Figure 1: Insert size distribution for PE sequencing libraries for *B. vulgaris* accession C869. (max = 64,496,131, min = 32, median = 440, and standard deviation = 511,068)

Supplemental Figure 2: Size distribution for indels detected within cultivated *B. vulgaris* accessions.

## Figures

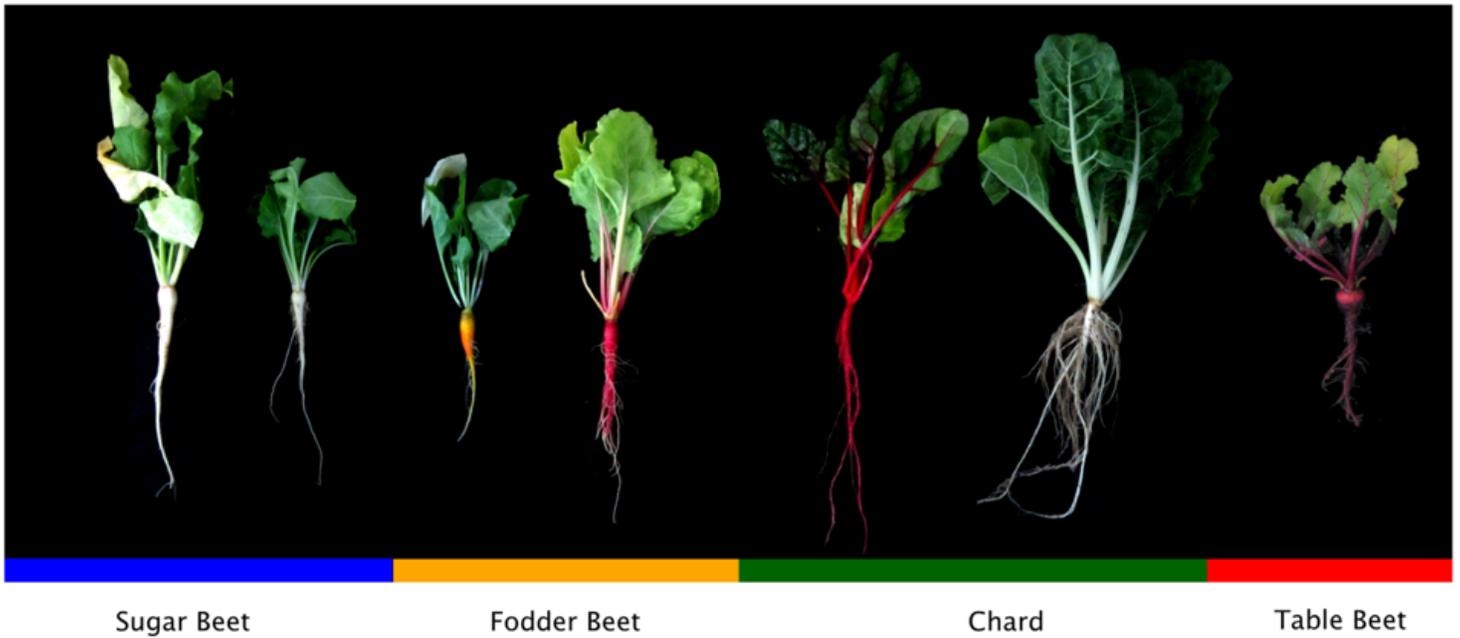


Figure 1

Phenotypes of *B. vulgaris* showing crop type characteristics are distinguishable by 9-weeks of age. Color bars refer to crop type in subsequent figures.

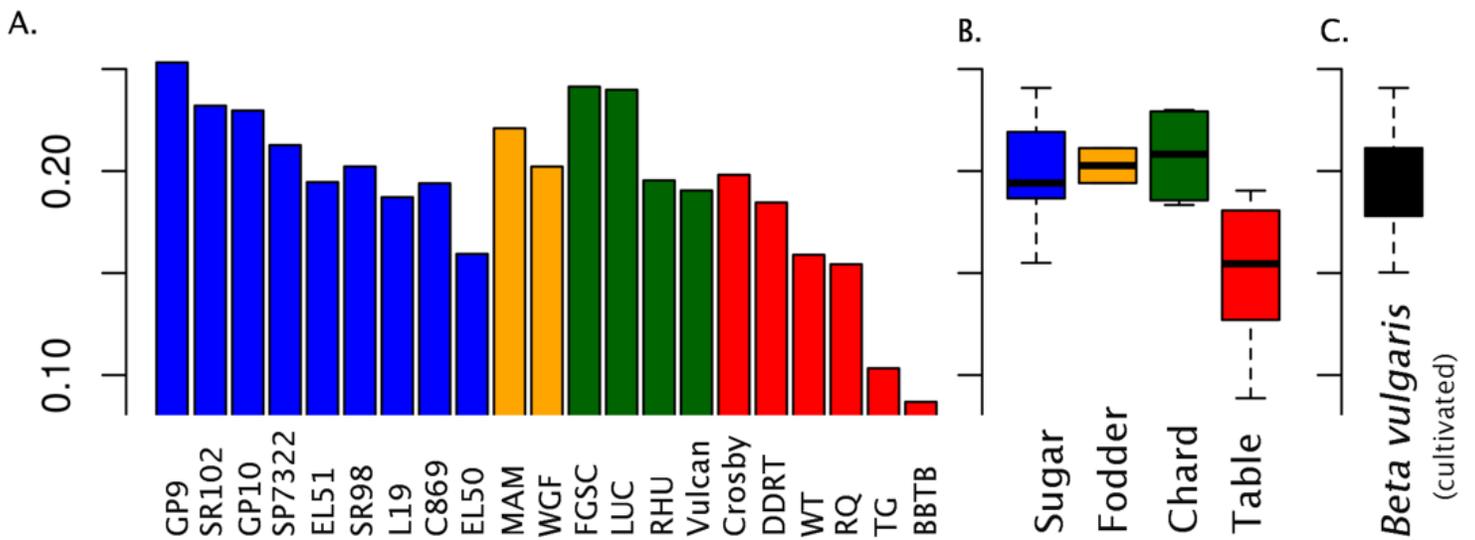
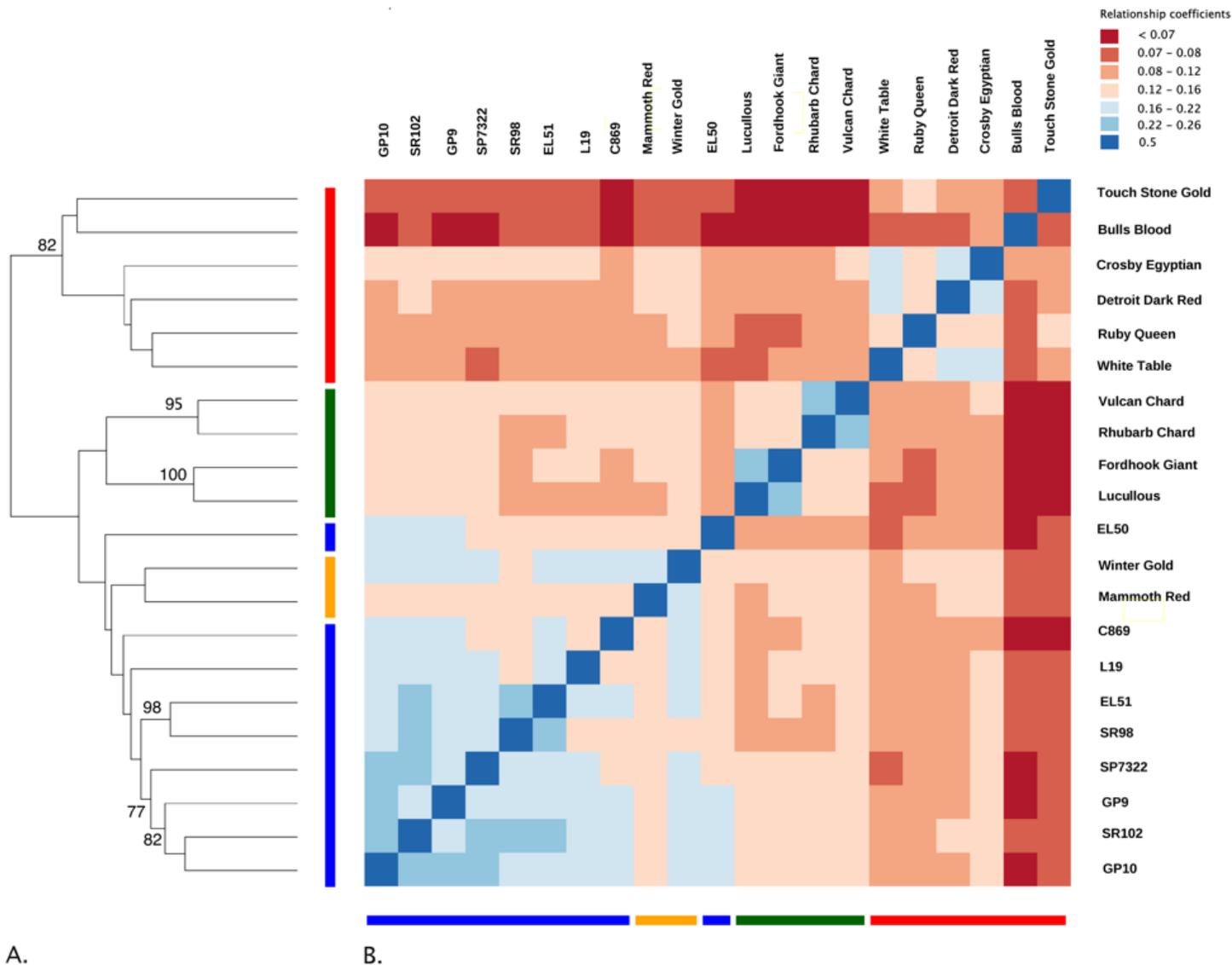


Figure 2

Gene diversity/expected heterozygosity ( $2pq$ ) of *B. vulgaris* lineages. (A) Populations, (B) Crop types, and (C) Species. Colors are coded according to Panel B and values are present in Table 2



**Figure 3**

Lineage relationships inferred by hierarchical clustering of pairwise relationship coefficients. (A) Dendrogram reflects support for clusters. Branch lengths indicate relationship coefficients between lineages, high (blue) and low (red). (B) Heatmap shows relationship coefficient values for all comparisons. Colors at the bottom and left of heat map represent crop type, sugar beet (blue), fodder beet (orange), chard (green), table beet (red).

PCA Plot (Allele Frequency)

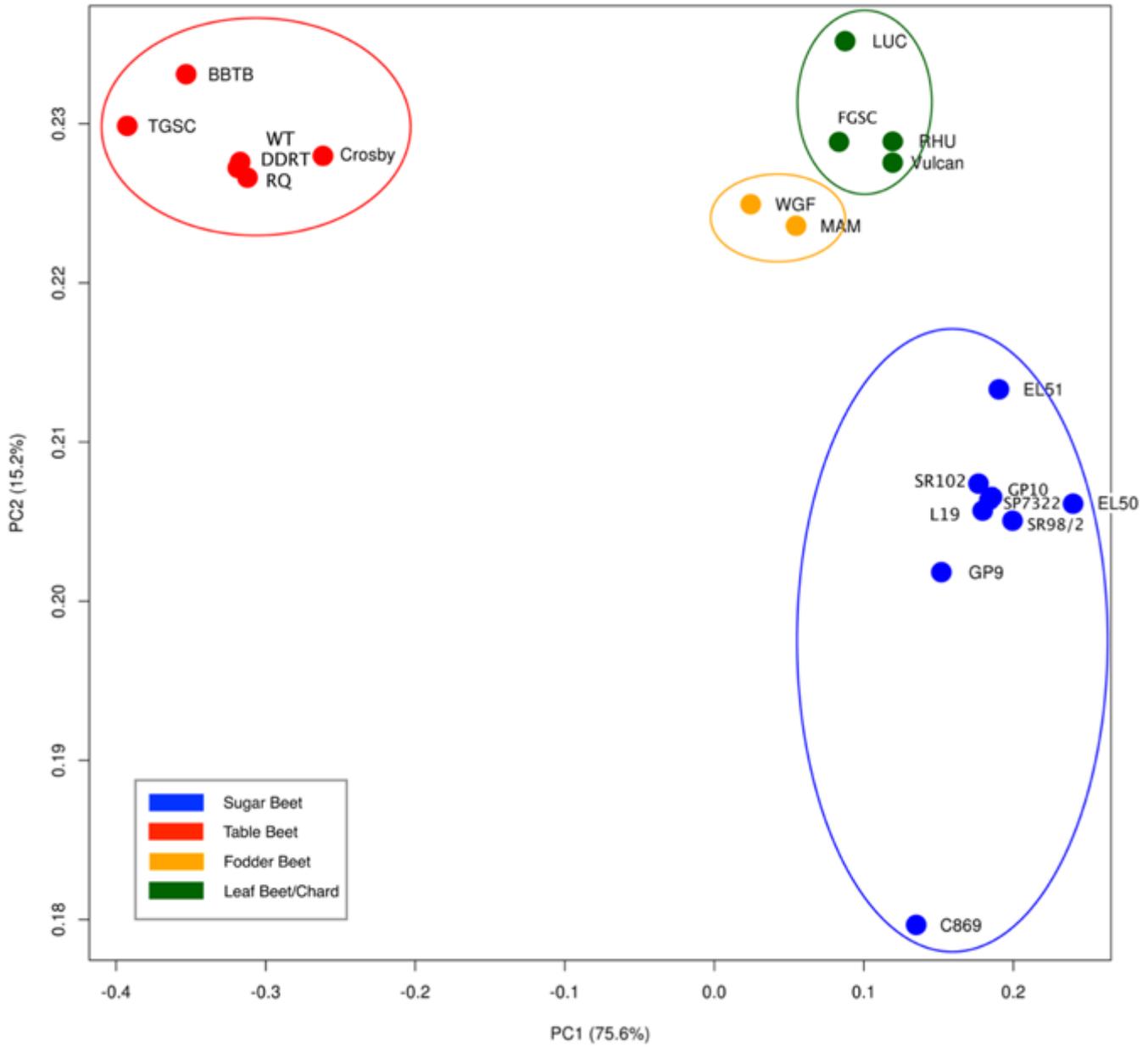
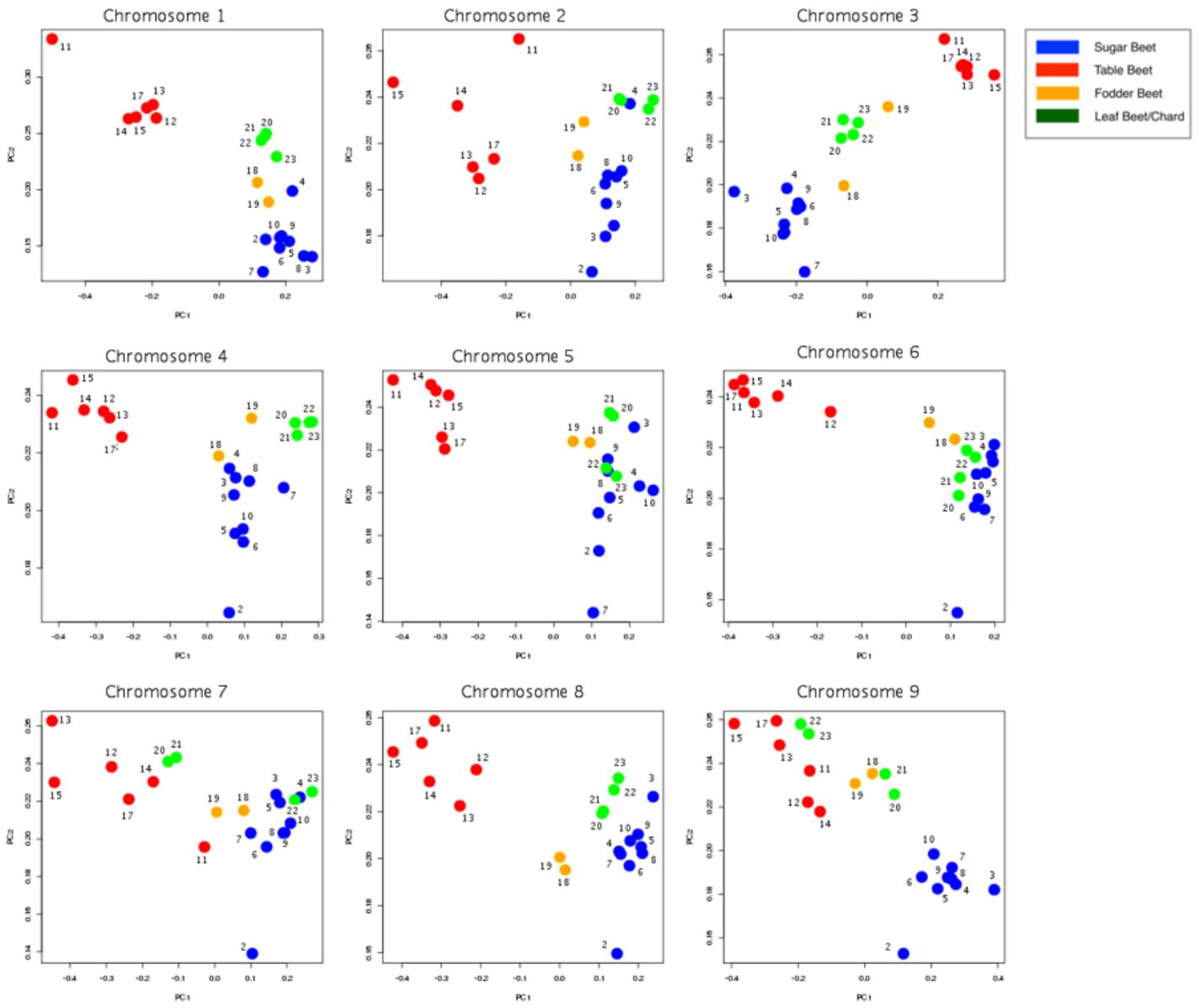


Figure 4

PCA plot showing the separation of crop types using genome-wide allele frequency data.



**Figure 5**

PCA plot showing the separation of crop types using allele frequency data on a chromosome by chromosome basis. Colors group crop types as in Figure 4.

## SMC++ Analysis

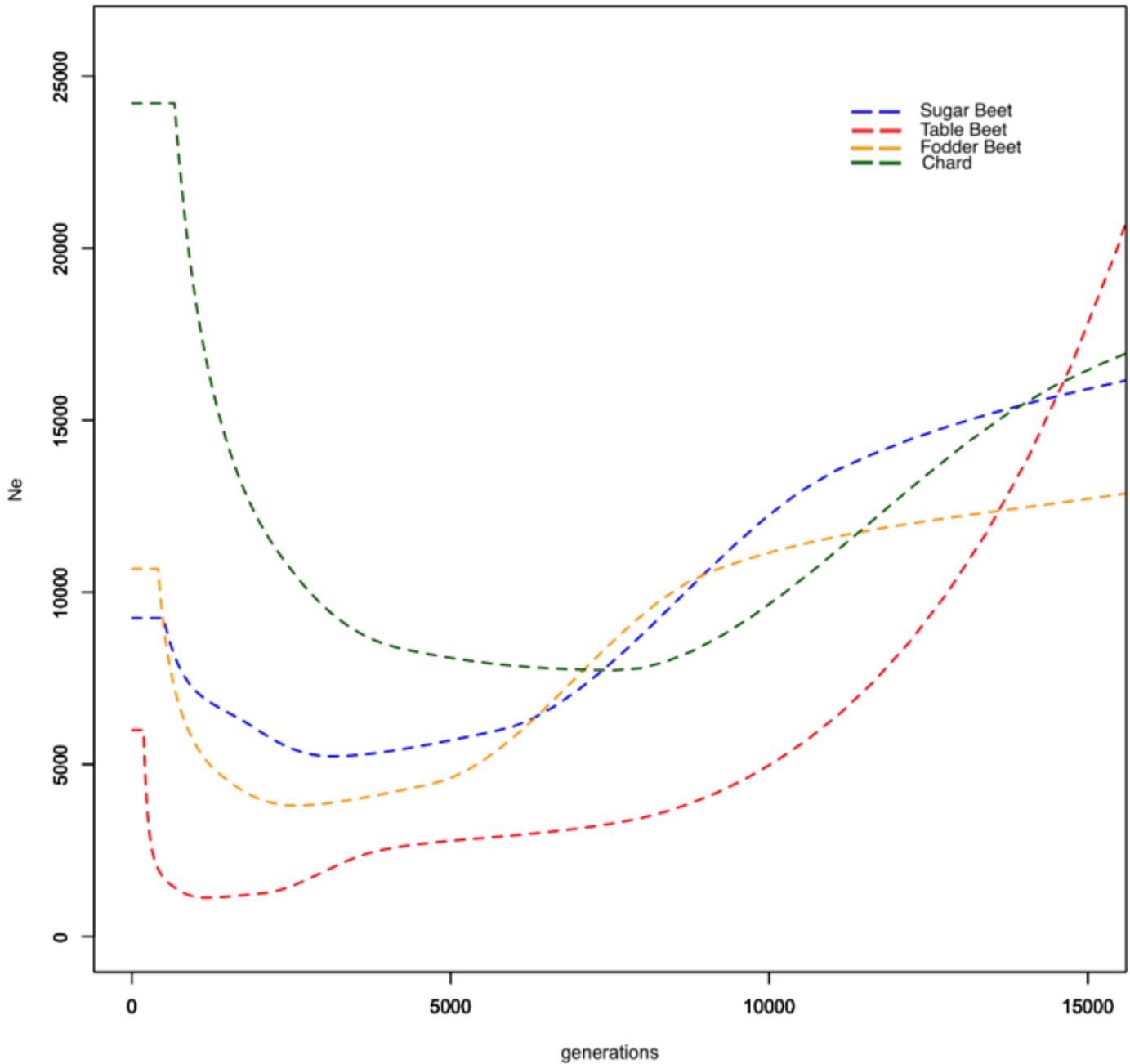


Figure 6

Inferred historical  $N_e$  of *B. vulgaris* crop types using the program SMC++. Colors group crop types. Red = table beet, blue = sugar beet, green = chard (leaf beet), yellow = fodder beet.

## Supplementary Files

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

- [SupplementalFigure2.jpg](#)

- [SupplementalFigure1.jpg](#)
- [BMCGenomicsPaulCh1.submitted927.xlsx](#)