Mini review

Next generation sequencing technologies to explore the diversity of germplasm resources: Achievements and trends in tomato

Pasquale Tripodi
CREA Research Centre for Vegetable and Ornamental Crops, Via dei Cavalleggeri 25, 84098 Pontecagnano Faiano (SA), Italy

Article info
Article history:
Received 17 October 2022
Received in revised form 10 November 2022
Accepted 10 November 2022
Available online 13 November 2022

Keywords:
Tomato
Next generation sequencing
SNP array
Genotyping by sequencing
Resequencing
Pangenome

Abstract
Tomato is one of the major vegetable crops grown worldwide and a model species for genetic and biological research. Progress in genomic technologies made possible the development of forefront methods for high-scale sequencing, providing comprehensive insight into the genetic architecture of germplasm resources. This review revisits next-generation sequencing strategies and applications to investigate the diversity of tomato, describing the common platforms used for SNP genotyping of large collections, de novo sequencing, and whole genome resequencing. Significant findings in evolutionary history are outlined, thus discussing how genomics has provided new hints about the processes behind domestication. Finally, achievement and perspectives on pan-genome construction and graphical pan-genome development toward precise mining of the natural variation to be exploited for breeding purposes are presented.

© 2022 The Author. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Contents
1. Introduction ................................................................. 6250
2. Global germplasm resources in tomato ......................................................... 6251
3. Advances in next generation sequencing in the past twenty years .................. 6251
4. Genotyping array .......................................................... 6252
5. Reduce representation libraries for SNP discovery ...................................... 6254
6. Whole genome sequencing and resequencing projects .................................. 6254
7. Long-read technologies for de novo sequencing and pan-genome investigations ... 6256
8. Future perspectives .......................................................... 6256
CRediT authorship contribution statement ...................................................... 6256
Declaration of Competing Interest ................................................................. 6256
Acknowledgements ....................................................................................... 6256
References ................................................................................................. 6256

1. Introduction

Tomato (Solanum lycopersicum L.) is one of the most economically important vegetable crops, with a production of ~190 million tons on a surface of over 5 million hectares it ranks among the first 12 crops in the World, accounting for ~17% of the entire vegetable production [1]. Thanks to the content of health beneficial compounds and its double attitude for consumption as fresh or processed products [2], this crop plays a central role in human nutrition. Since the beginning of 90s, tomato has been a model for plant breeding and genetics studies through the development of dense libraries of molecular markers [3], genetic and physical maps [4], and experimental mapping population [5], being among vegetables, pioneering for dissecting the basis of quantitative traits and for map-based cloning of genes and quantitative trait loci (QTLs) [6].

The completion of the whole genome sequence in 2012 [7], led to a rapid evolution of cutting-edge technologies for next-generation sequencing toward the discovery and massive screening of thousands or millions of single nucleotide polymorphisms
(SNPs) evenly distributed across the genome. The transition from second to third-generation sequencing methods allowed a deeper scale of analysis, maximizing the throughput and generating vast quantities of data [8] thus providing a unique opportunity to explore the gene pool. So far, thousands of morphologically diverse wild and cultivated accessions have been developed during the domestication and post-domestication processes and further refined in decades of intense breeding [9]. This germplasm represents a valuable source of variation that has been the target of several genomics diversity investigations.

After a brief revisitation of tomato germplasm resources, this review discusses the role and the impact that next-generation sequencing platforms have had in the investigation of genomic diversity. Particular emphasis is given to: a) reduced represented libraries such as restriction site-associated DNA sequencing (RAD-seq) [10] and genotyping-by-sequencing (GBS) [11]; b) SNP genotyping arrays; c) short and long-read sequencing methods. Main findings and perspectives are discussed, highlighting the future actions to take in order to fully grasp the potential held in germplasm collections.

2. Global germplasm resources in tomato

Tomato belongs to the large Solanaceae family, which includes over 90 genera and nearly 3000 species widely distributed in the tropical and temperate world regions [12]. According to the most recent taxonomic classification, the section Lycopersicon of the Solanum genus comprises the domesticated (S. lycopersicum) and semi-domesticated (S. lycopersicum var. cerasiforme) tomato, as well as 12 wild relatives, all originating from different habitats of Latin America and characterized by a wide diversity expressed at molecular and morpho-physiological levels [13,14]. Among wild species, six are distinguished for the green fruits with the presence or absence of longitudinal stripes (S. carneum, S. corneii, S. chmelinskii, S. habrochaites, S. pennellii and S. neorickii), three for the green-purple colour (S. chilense, S. peruvianum, S. huylaseyen), the rest with yellow-green (S. cheesmaniae) and orange (S. galapasgoense) colours [15]. S. pimpinellifolium, the closest wild relative of the cultivated tomato presents red fruits at maturity. Wild relatives have been proven a valuable source of genetic diversity and a reservoir of genes conferring resistance to biotic and biotic stresses, being used as founder lines in several breeding programs [15]. The evolution of the modern tomato from its wild progenitor is a sum of complex mechanisms triggered during the domestication phase and related to the change of mating systems, migrations, and selection of causative mutations [16,17]. These processes known as domestication syndrome shaped the architecture of plants, growth habitus, and the morphology of the edible part [18]. Selective breeding and cultivation in specific niches led to the development of numerous cultivars, adapted to specific environments, thus favouring the establishment of highly variable phenotypes. Market needs and consumer preferences furtherly drove the rapid evolution of varieties addressed to processing and fresh consumption [19]. Currently, a wide varietal panorama exists in tomato represented by thousands of morphologically diverse ancient and modern varieties including heirlooms, landraces, elite cultivars, and hybrids [20].

Preserving these germplasm collections is crucial for facing challenges of food and agricultural systems in the coming decades, related to demographic trends, climate changes, and the need for more sustainable use of resources [21]. It is estimated that more than 80 thousand accessions are kept ex-situ across genebanks located around the globe [22]. The largest collection is held at the World Vegetable Center (WorldVeg, Tainan, Taiwan) [23], comprising over 8,200 cultivated tomato, mostly cultivars and landraces and several hundred accessions of wild species. Another vast collection is maintained at the Plant Genetic Resources Unit of the United States Department of Agriculture (USDA-PGRU, Geneva, NY, USA) [24] and consists of more than 6,600 accessions sampled in different areas of North, Centre, and South America and includes both vintage and modern varieties as well as several breeding lines. The Tomato Genetic Resource Center at the University of Davis (TGRC, CA, USA) [25] is well-known for maintaining over 1,000 wild accessions and a similar number of monogenic mutants for plant development, fruit features, and disease resistance. Furthermore, this genebank holds numerous pre-breeding stocks consisting of the different experimental mapping populations developed through interspecific breeding such as introgression lines, backcross, and recombinant inbred lines. Other main genebanks located in Europe (IPK, Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany; CGN, Centre for Genetic Resources, The Netherlands; INRAE-GAFL, French National Institute for Agricultural Research Genetics and improvement of Fruit and Vegetables, Avignon, France) [26–28] store and distribute thousands of individuals, including scientific core collections developed for specific purposes (e.g., resequenced genotypes, intra- and interspecific populations, multiparental populations, etc.). Leading-edge technologies for sequencing and genotyping and advanced computational tools opened new frontiers for exploring crop genetic diversity. The gained knowledge is paramount for planning strategies for the management, conservation, and improvement of germplasm resources. The genomic investigations later described helped to dissect the genetic structure of tomato, providing new insight into its evolutive history and mechanisms of domestication.

3. Advances in next generation sequencing in the past twenty years

The first progress in DNA sequencing was made by the group of Frederick Sanger in the late 70’s who developed the homonymous method based on the use of specific chain-terminating fluorescently-labelled di-deoxynucleotides that were incorporated in the single-strand DNA chain and then separated through electrophoresis-based methods [29]. The subsequent automation enabled the development of first-generation sequencers (e.g., ABI prism) with different flexibility, capable of analyzing 4 to 96 samples per single run [30]. Despite the high quality of sequencing was able to produce sequences up to 1000 bp in length, the laboriousness and sluggish methodological procedure as well as the limited quantity of data in the unit of time made this method not proper for high throughput applications. Indeed, Sanger enabled the completion of small genomes [31–33] and was used for the initiation of large sequencing projects (e.g., human genome) [34], while remaining poorly exploited for the investigation of larger and more complex genomes.

Fragment analysis implemented in capillary sequencing systems was used to detect microsatellite polymorphisms. In tomato, this strategy has been widely applied, mostly in small collections with a defined number of markers [35–38] thus generating low data points and limiting the informativeness at genome-scale (Fig. 1).

Technological advancement accelerated the development of next-generation sequencing, which represented a step ahead of the traditional Sanger sequencing method, increasing efficiency, and ensuring a greater genome coverage with higher throughput at reduced costs [39]. Based on read length (i.e., number of base pairs sequenced from a DNA fragment) NGS applications are categorized into second and third generation technologies [40].

Second-generation or short-read sequencing is based on different hybridization or synthesis-based techniques [41] in which mil-
of DNA fragments, after libraries preparations steps, are sequenced in parallel, ensuring fragments of 250–800 bp length and a throughput higher than 25 gigabases per day [42]. The first outstanding NGS machine was launched in 2005 by the 454 Life Sciences Roche [43] opening the era of pyrosequencing [44] which was later discontinued in 2013 [41]. During 2006–2007, two other platforms were opened to the community: SOLiD (Thermo-Fisher Scientific) who proposed the alternative sequencing by ligation [45], and Solexa later acquired by Illumina which developed the adapter-bridge amplification [46] providing a greater read length and depth compared to SOLiD [47] although this latter has remained competitive for the cost for base. Ion Torrent was the last striking second-generation platform [48] based on the sequencing by synthesis approach without the use of fluorescence or luminescence techniques, thus reducing costs and speeding up sequencing process. Despite its competitiveness, constraints occur in the detection of homopolymers making difficult the unravel repetitive sequences. Nowadays, Illumina is the most widely used technology for genotyping studies in tomato since offers several options for genome and transcriptome sequencing and resequencing with outputs varying from a few gigabases up to 6 terabases [49]. Both reduced representation libraries sequencing methods and SNP array later discussed rely on Illumina technology.

Third-generation or long-read sequencing made a big leap with respect to the previous approaches, offering a faster and more accurate method of sequencing by reducing errors and improving the resolution of assembling in repetitive regions [50]. Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are the two technologies implementing long-read. The former adopts the single molecule real-time (SMRT) technology based on a DNA polymerase that drives both the reaction in absence of polymerase chain reaction (PCR) amplification and the real-time capture of fluorescence signals once labelled nucleotides are incorporated into the DNA [51,52]. The latter uses a membrane with bio-pores of nanoscale diameters in which a differential electric potential is applied, hence allowing the detection of nucleotides through pores during the sequencing process. The advantages of SMRT PacBio include accuracy and low occurrence of bias and epigenetic events’ identification, thus enabling the high resolution of sequencing in complex genome regions [53]. Nanopore sequencing offers the same advantages of PacBio, although the error rate is reported to be higher [41,54]. Furthermore, Nanopore technologies are fast, affordable, and flexible in terms of space requirements given the small size of the sequencing devices. Additional methods corroborate long-read for the completion of genome sequencing through a better ordering of positioning of contigs (Optical Mapping) [55] or inferring spatial chromatin conformation on chromosomes (HiC) [56]. In tomato long-read sequencing has been applied both for generating high-quality reference genomes and for pan-genome projects.

4. Genotyping array

In plants, array platforms based on hybridization on solid supports of customized oligonucleotide (allele-specific) probes, have
been developed for discovering and genotyping thousands of single-nucleotide variants [57]. Arrays rely on several steps that include fragmentation of DNA and its amplification by PCR, followed by labeling with dyes, and hybridization with specific probes [58]. Variations of the protocol may occur depending on the technology used. The first platform allowing the design of customized probes for SNP genotyping of tomato germplasm was the Illumina Golden Gate. Shirasawa and collaborators [39] from the information retrieved from the genome sequence of six cultivated lines, developed an assay of 1,536 SNPs and performed a population diversity and genome-wide association study (GWAS) for major agronomic and morphological traits in 663 tomato accessions (Table 1). Corrado and colleagues [60] developed a panel of 177 customized SNPs to infer the population structure of 214 cultivated tomato genotypes. Both investigations represented a step ahead for the population genetic studies in tomato, allowing marker analysis at a deeper genome-scale. Golden gate has been discontinued by Illumina in 2014 being replaced by other approaches such as the SolCAP that nowadays is the most used array in tomato. SolCAP has been designed by the Solanaceae Coordinated Agricultural Project Consortium through transcriptome sequences of six tomato accessions including four cultivated genotypes for fresh-market and processing (S. lycopersicum), one cherry type (S. lycopersicum var. cerasiforme) and one wild relative (S. pimpinellifolium), generating 62,576 non-redundant SNP [61,62]. The array, designed using 10,000 probes, ensured the presence of SNPs in coding regions throughout the entire genome, and with minimal overlaps for a total of 8,784 SNPs, 7,720 of which had high quality after applying filtering criteria. SoCAP has been the genotyping platform used in several genetic diversity studies. Sim and collaborators [63] investigated a diverse set of 426 accessions, mostly vintage cultivars and modern varieties for the fresh market or for processing. Analysis of principal components confirmed separation of the different groups essayed, highlighting positive selection within candidate genes responsible of fruit morphology. Blanca and collaborators [64] investigated the diversity of S. pimpinellifolium and S. l. cerasiforme to better define the role of the two species during the domestication of tomato. Based on ancestry analysis from 7,414 SNPs on 272 individuals retrieved from the COAMV (Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana) genebank, it was possible to suggest that S. pimpinellifolium was the wild progenitor of the cultivated tomato, while S. l. cerasiforme was an intermediated species which had been diversified in different areas of Andean region and Mesoamerica, where the pre-domestication and the domestication occurred, respectively. Sauvage and colleagues [65] performed association mapping in 163 individuals of S. lycopersicum, S. pimpinellifolium and S. l. cerasiforme retrieved from the germplasm collection maintained at INRAE-GAFL. Through 5,995 single-nucleotide polymorphisms spanning the whole genome, 44 loci associated with sugar and organic acids were identified. Population structure revealed two main groups, separating the domesticated from the wild species, thus confirming the previous study [64]. Furthermore, the extensive linkage disequilibrium found by authors, suggested a combined effect of domestication and selective breeding that led to a reduction of nucleotide diversity and the elimination of recombinant lineages. The SolCAP array has offered the possibility to merge data from different studies. Blanca and collaborators [66] performed a meta-analysis by combining the information from the three above-mentioned studies [63–65], thus generating a large dataset of 1,008 accessions that included both cultivated and wild tomatoes as well as interspecific hybrids (Table 1). The enlargement of the germplasm panel shed light on patterns of genetic structure based on fruit morphology, geographical provenance, climate zones, cultivar classification, allowing, furthermore, to refine inconsistencies in passport data by revealing that several cherry tomatoes were incorrectly classified as S. l. cerasiforme. Authors investigated the frequency of main genes cloned in tomato and involved in the transition from small and round to large shape (fw2.2, fw3.2, H, ovate, fas, sun) finding the absence or very low frequency in S. pimpinellifolium accessions while a higher frequency was found in cultivated and/or cerasiforme genotypes. Finally, new clues were provided for the evolutionary history of cultivated tomatoes suggesting that the major genetic bottleneck occurred prior to the introduction of tomato in Europe coinciding with the migration of S. l. cerasiforme from Northern Perú and Ecuador to Mesoamerica. Recently, Schouten and collaborators [67], investigated the genetic diversity of modern cultivars commercialized in the Netherlands highlighting how breeding programs in the past decades have increased the diversity in a two-stage step, the former to transfer resistances to pathogens, the latter to improve fruit quality and aroma. Based on 7,661 SNPs, authors captured the Nei’s genetic diversity index (H) of cultivars released since 1960, detecting furthermore the introgressions from wild relatives. Results allow determining a boost of diversity since 1970, thus establishing how the increase in diversity due to introgressions exceeded the possible decrease caused by selection. Beyond genetic diversity studies, the SolCAP array has been demonstrated to be a powerful genotyping technology for GWAS for fruit morphological traits [68], agromonic traits [69], quality

| NGS platform | Method/Application | Species | N° of individuals | N° of SNP markers | Reference |
|-------------|-------------------|---------|------------------|------------------|-----------|
| S. lycopersicum, S. lycopersicum var. cerasiforme, S. caranum, S. chmielewskii, S. cheesmaniae, S. chilense, S. galapagoense, S. habrochaites, S. neoricii, S. pennelli, S. pimpinellifolium, S. persicaria, S. huaylasense | 34,550 | [73] |
| Golden Gate | Golden Gate | SL, SP, SLc, SPm, Spv | 633 | 1,536 | [59] |
| Illumina Golden Gate | Golden Gate | SL | 214 | 177 | [60] |
| Illumina Infinium | SolCAP | SL | 426 | 7,720 | [63] |
| Illumina Infinium | SolCAP | SL, SLc, SPm | 272 | 7,414 | [64] |
| Illumina Infinium | SolCAP | SL, SLc, SPm | 163 | 5,995 | [65] |
| Illumina Infinium | SolCAP | SL, SLc, SPm, SC, SN, SC, SL × SPm, SL × SP | 121 | 7,672 | [68] |
| Illumina Infinium | SolCAP | SL, SLc, SC, SL, SLc | 600 | 7,720 | [71] |
| GeneTitan Affimetrix | 51 K Axiom® | SL | 96 | 16,782 | [72] |
| GeneTitan Affimetrix | 51 K Axiom® | SL | 162 | 34,550 | [73] |
| Illumina HiSeq 2500 | ddRAD | SL | 288 | 32,799 | [77] |
| Illumina HiSeq 2000 | GBS | SL, SH, SC, SP, SY | 1254 | 64,943 | [78] |
| Illumina HiSeq 2500 | SPET | SL, SPm, SP, SA, SC, SN, SH, SY | 400 | 12,002 | [81] |

*S. lycopersicum, S. lycopersicum var. cerasiforme, S. caranum, S. chmielewskii, S. cheesmaniae, S. chilense, S. galapagoense, S. habrochaites, S. neoricii, S. pennelli, S. pimpinellifolium, S. persicaria, S. huaylasense.*

---

Table 1
Different next generation sequencing (NGS) platforms and genotyping methods to investigate the diversity of tomato collections.
study in terms of both assayed accessions and number of markers, and hybridization breeding [82] (Table 2). A mean genome coverage of 2.9 million read pairs per sample were produced with GBS sequencing, leading to a total of 1036 million raw reads, corresponding to an average of 3.5 million read pairs for 288 tomato genotypes, whereas 3700 million reads corresponding to an average of 2.9 million read pairs per sample were produced with GBS on 1254 accessions. This result highlights how the double digestion and paired-end sequencing increase the genome coverage [79].

Another innovative and flexible approach for genotyping patented by NuGEN® Technologies (San Carlos, CA, USA) is the single primer enrichment technology (SPET). SPET combines the main features of RRLs and arrays, allowing the discovery of thousands of genome-wide polymorphic SNPs through the sequencing of specific sites using customized probes (primers) [80]. This method offers notable advantages compared to reduced representation libraries. Indeed, while GBS and ddRADseq detect the genetic variation randomly, often within repetitive and/or intergenic regions, SPET offers the advantage to sample the genetic variation within the gene space improving the possibility to identify causative polymorphisms. The only application of SPET in tomato has been reported by Barchi and collaborators [81] who designed 5,000 probes based on data from 236 resequenced genomes [82,83]. The SPET panel comprising 12,002 SNPs mostly localized in gene-rich chromosome regions was used to genotype 400 cultivated and wild accessions developed in the frame of the G2P-SOL international consortium [84]. The results converged with the acquired knowledge on phylogenetic relationships between tomato species, increasing the intra-specific discrimination, thus demonstrating the reliability for further domestication and genome-wide association studies. It is expected a wide use of SPET technology for the management of large collections and to resolve any mislabelling particularly for the accessions maintained in genebanks.

6. Whole genome sequencing and resequencing projects

The first reference genome was released by the International Tomato Genome Consortium in 2012 [7]. The full sequence of S. lycopersicum cultivar Heinz 1706 and the draft of S. pimpinellifolium were presented, combining both first and second next-generation sequencing technologies. The cultivated tomato genome was assembled in 91 scaffolds and covered 760 MB of the estimated 900 MB for a total of 34,727 predicted protein-coding genes. Comparative analysis with the wild ancestor and other members of the family Solanaceae, identified two triplication events that reshaped the functionality of genes underlying key fruit traits. Since then, efforts have been pushed forward genome resequencing projects. The first attempt was performed by Causse and collaborators in 2013 [85] on 4 S. lycopersicum and 4 S. l. cerasiforme genotypes. The Illumina GAIIx platform generated 970 million paired-end reads mostly of 101 bp for an average sequence depth of coverage of 11.2×. A total of over 4 million unique SNPs, ~128 thousand InDels and 1,686 putative copy-number variations were detected, thus providing novel genomic variation to use for QTL and gene mapping. A year later, the 100 Tomato Genome Sequencing Consortium resequenced the genome of 84 cultivated and wild tomato accessions and produced de novo assemblies for the reference Heinz 1706 and for three wild parents of crossing populations (S. habrochaites LycC4, S. pennelli LA0716 and S. arcanum LA2157) by combining Illumina HiSeq 2000 and 454 FLX technology, thus opening new perspectives for genetic diversity, phylogenetics and hybridization breeding [82] (Table 2). A mean genome coverage of 36-fold per accession was found, with a number of SNPs ranging from 0.5 to 4 million in the cultivated accessions up to giving novel insight into the role of the Mediterranean basin as a secondary centre of diversity from which many breeds were generated and then exported to the rest of the world.

Another array containing 51,912 candidate SNPs was developed by resequencing 96 big-fruited hybrids with a mean depth of 1.9x through the Illumina HiSeq2000 [72]. The 51 K Axiom array allowed a higher SNP coverage and has been mostly used to harness the diversity of commercial cultivars for genomic selection and GWAS [72,73]. No further hints have been provided for the genetic structure and phylogenetic relationships of cultivated and wild tomatoes.

5. Reduce representation libraries for SNP discovery

Reduced-representation libraries (RRLs) are methods based on restriction enzyme digestion of a target genome for generating small fragments to sequence and discovering, after several computational steps, a thorough catalogue of genetic markers [74]. Although the approach is focused on reducing the overall portion of the genome to be sequenced, it provides high coverage for each target sequence, thus enabling powerful genotyping of many individuals with limited costs [75]. Based on modifications in the different steps of digestion, ligation, pooling, amplification, and amplicons size, RRLs can be grouped into three categories: complexity reduction of polymorphic sequences (CRoPS); restriction-site associated DNA (RAD-seq) techniques (e.g., ddRAD, mbRAD, 2bRAD, ezRAD); genotyping by sequencing (GBS) [75,76]. Unlike other species, in tomato few investigations of germplasm diversity involved the application of RRLs. Double digest Restriction-site associated DNA (ddRADseq) and GBS have been the most used methods for high-density genotyping of large collections. Both rely on the same libraries-preparation procedures, although GBS is more flexible involving fewer steps than ddRADseq and more enzyme choice options [74,75]. ddRADseq has been applied for population structure analysis of a representative panel of local varieties and heirlooms. The study, performed by Esposito and collaborators [77], concerned 288 accessions mostly retrieved from the Mediterranean area and including a large portion of long shelf-life cultivars. The double enzyme digestion (Mbol and Sphi) protocol enabled the discovery of over 240 thousand SNPs, of which 32 thousand after the application of filtering criteria were used to infer phylogenetic relationships and ancestry of individuals through Bayesian and non-parametric computations. The analyses highlighted specific clusters that reflected the origin of cultivars, thus suggesting new clues on the existence of a specific fraction of diversity within the long shelf-life gene pool. Indeed, within long shelf-life germplasm, loci under selection were detected in genomic regions regulating fruit ripening and response to main abiotic stress.

Genotyping by sequencing through single enzyme restriction (ApeKI) was adopted in a large investigation performed in a collection of 1254 accessions including European traditional and modern varieties, early domesticated accessions, and wild relatives. Among the total 448,121 variants discovered, 64,943 were selected by filtering for quality criteria [78]. GBS results revealed a low level of genetic diversity in the European traditional gene pool when compared to the wild and Mesoamerican one, displaying, in agreement with ddRADseq data [77], the presence of main clusters that included long shelf-life accessions of Spanish and Italian origin. Furthermore, within the European varieties, 298 highly variable positions underlying major morphological characteristics selected by breeders were discovered, showing how the existing variability in tomato has been created through breeding from original material with narrowed genetic diversity. So far, this is the broadest study in terms of both assayed accessions and number of markers,
10–11 million in wild species, indicating the low level of nucleotide diversity enclosed in domesticated tomatoes due to genetic erosion. Furthermore, a higher number of non-synonymous SNPs was found in cultivated tomatoes when compared to wild relatives due to breeding activities and selective pressure that led to differential allele fixation. After a few months, additional resequencing data for 360 tomato accessions were released [83]. Overall, 2.6 trillion base pairs of sequence, with a median depth of 5.7× and a final set of about 12 million SNPs and 1.3 million small indels were used to reconstruct the evolutionary history of tomato breeding. The authors corroborated previous findings that support *S. pimpinellifolium* as the ancestor of cultivated tomato and *S. l. cerasiforme* as the intermediated species that led the transition from wild to cultivated tomato. In this study, a new link between the semi-wild ancestor and modern tomato was found. The origin of *S. l. cerasiforme* has been reported to be antecedent to domestication, diverging from the wild *S. pimpinellifolium* thousands of years before humans were present in Latin America. Authors also found two ancestral groups for *S. l. cerasiforme* based on their provenance in South and Central America regions. Therefore, this species may have originated in Ecuador as a wild and then spread in other surrounding regions finally toward Central America. During this transition, native medium-size fruits were selected, however, some traits were lost once the partially domesticated forms were spread northward. By reshaping the role of the semi-domesticated berry tomato, this study opened new questions on the movement of this nightshade species in the origin centre highlighting the importance to explore the gene flow during the different steps of domestication and identify lost alleles, hence opening new frontiers for plant breeding.

| NGS Platform                  | Method/Application | Species*                   | N° of individuals | Reference |
|-------------------------------|--------------------|----------------------------|-------------------|-----------|
| Illumina HiSeq 2000 & 454 FLX technology | Resequencing       | SL, SLC, SP, SH, SA, Spm, SG | 877           | [82]      |
| Illumina HiSeq 2000           | Resequencing       | SL, SH, SCh, SG, SPv, SN, SCI, SP, SLc | 360       | [83]      |
| Illumina GAIix                | Resequencing       | SL, SLc                     | 8              | [85]      |
| Illumina HiSeq 2000           | Resequencing       | SL, SLc, SPM               | 476h           | [86]      |
| Illumina NextSeq             | Resequencing       | SL, SLc, SPM               | 160            | [88]      |
| PacBio, Illumina NextSeq 500, HiCt, ONT* | De novo sequencing | SPM                       | 2              | [91]      |
| PacBio Sequel II, HiCt, Bianano | De novo sequencing | SL                         | 1              | [92]      |
| PacBio HiFi and ONT Nanopore  | De novo sequencing | SL                         | 1              | [93]      |
| PacBio Sequel and Illumina HiSeq 2000 | De novo sequencing | SPM, SLc                    | 2              | [94]      |
| Illumina MiSeq, PacBio, HiCt  | De novo sequencing | SPM, SLc                    | 1              | [95]      |
| Illumina GAIix, HiSeq2000, Nextseq | Resequencing       | SL, SLC, SPM, SCh, SG       | 725d           | [89]      |
| Illumina Truseq, ONT         | Pangenome + De novo sequencing | SL, SLC, SPM, SCh, SG       | 100         | [96]      |
| PacBio Sequel II and HiCt    | Graph-Pangenome    | SL, SLC, SLP, SCh, SG       | 838h           | [99]      |

* SL = *S. lycopersicum*, SLC = *S. lycopersicum* var. *cerasiforme*, SA = *S. arcanum*, SCh = *S. cheiranthus*, SCI = *S. chilense*, SG = *S. galapagoense*, SH = *S. habrochaites*, SLP = *S. lycopersicoides*, SN = *S. neorickii*, SP = *S. pennelli*, SPM = *S. pimpinellifolium*, SPV = *S. peruvianum*

| Method/Application                     | Species* | N° of individuals | Reference |
|---------------------------------------|----------|-------------------|-----------|
| De novo sequencing                    | SPM, SLc | 2                 | [91]      |
| De novo sequencing                    | SL       | 1                 | [92]      |
| De novo sequencing                    | SL       | 1                 | [93]      |
| De novo sequencing                    | SPM, SLc | 2                 | [94]      |
| De novo sequencing                    | SLP      | 1                 | [95]      |
| Pangenome + De novo sequencing        | SL, SLC, SPM, SCh, SG | 100 | [96] |
| Graph-Pangenome                      | SL, SLC, SLP, SCh, SG | 838h | [99] |

*Including sequences of 561 previously resequenced genomes [7,82,83,85,86,100,102].

*Including sequences of 806 previously resequenced genomes [82,83,85,86,89,96,100].

Table 2

Next generation sequencing (NGS) approaches for resequencing, *de novo* genome sequencing and pangenome in tomato.
this first pan-genome represents the next step to precisely mining the natural variation to be exploited for breeding purposes.

7. Long-read technologies for de novo sequencing and pan-genome investigations

The first results of the applications of long-read sequencing in tomato have been presented from 2020 onward. PacBio, ONT, HiCi and Bionano have been used singularly or in combination to generate high-quality reference genomes for individual cultivated lines or wild relatives, including S. pimpinellifolium acc. LA2093 and LA1589 [91], S. pimpinellifolium acc. LA1670 and S. lycopersicum var. cerasiforme acc. LA1673 [92], S. lycopersicum cv. Moneyberg [93], S. lycopersicum cv. Heinz 1706 [94], and S. lycopersicoides acc. LA2951 [95] (Table 2). Genomes generated by long-read sequencing guaranteed a sequencing depth 5 times higher than those of the short-read, beyond 100x. By linking PacBio to ONT was possible to improve the assembly contiguity at the whole chromosome level avoiding the occurrence of any gap and ensuring, furthermore, scaffold chromosome lengths ten time higher than those generated by second-generation technologies [92].

The new references allowed to detect with high precision the existing structural variants (SVs) between wild and cultivated species, determining their role in the evolution of key traits during domestication and breeding, thus facilitating their use as potential targets for the improvement of important horticultural traits [91]. The breakthrough of third-generation sequencing in tomato genomics has been achieved for pan-genome investigations, providing a broader view of structural and genetic changes underlying the diversity. Alonge and colleagues [96], applied nanopore sequencing to detect structural variations in 100 tomato accessions and released de novo assemblies for 14 reference genomes combining both short- and long-read sequencing. ONT produced a total of ~8 Teraborades of long-read data with a minimum of 40x genome coverage. In total over 238 thousand SVs were detected, mostly insertions and deletions. In agreement with previous studies [82,89] the wild species had more structural variants compared to the cultivated one, as a result of the loss of diversity during the evolution of tomato. Within the cultivated tomato, different clusters were detected reflecting a level of admixture with wild and semi-domesticated accessions, thus suggesting the occurrence of many SVs still undiscovered. Indeed, authors detected several SV associated with resistances introgressed from wild relatives. Functional analysis allowed to link the SV to three major domestication and breeding traits: the smoky flavour which negatively influences the consumer preferences, sb1 responsible of partial suppressor of branching and fw3.2 a major QTL controlling fruit mass. The two latter were found associated to hidden duplications that altered gene expression, thus reshaping the phenotypic diversity. The study opened new insights into the possibility of integrating the study of pan-genomes with long-read sequencing to reveal, with high reliability, the presence and function of structural variants in complex genomes.

The release of graphical pan-genome is the latest milestone achieved in tomato genomics. The concept of graph pan-genomes has been firstly applied in bacteria, aiming to represent the genome of a species rather than a single accession [97]. Thanks to the quality of long-read sequences and progresses in bioinformatics, this approach is increasingly applied in crop species [98]. In 2022, Zhou and collaborators [99], developed a large pan-genome consisting of 832 genotypes. Authors integrated 32 de novo assemblies obtained through PacBio and HiCi technologies and 806 previously sequenced assemblies, of which 100 using ONT [96] and 706 using short-read [82,83,85,86,89,100]. This approach allowed to generate high-quality assemblies representing as much as possible the whole diversity occurring in tomato. The graph pan-genome spanned 1007.56 Mbp for a total of 51,155 genes, 28.3% of which new respect the reference genome. In total, over 19 million variants were detected including ~18 million SNPs, ~1.5 million indels, and about 200 thousand SVs [99]. The authors then tested the potentiality of the graph pan-genome for the identification of variants associated with metabolic traits. Compared to single reference genomes, graph pan-genome offered a higher mapping resolution in GWA studies, enabling the detection of the missing heritability and providing comprehensive coverage of all genetic variants. Such advancement paves the way for new opportunities for genomic-assisted breeding and for the selection of target traits. Although some constraints for comparative analysis, genome mapping, and its visualization still occurs, it is expected that graphical pan-genomes will be obtained for several other species in further years thanks to the development of novel dedicated analytic tools [101].

8. Future perspectives

During the past decades, remarkable advancements have been done in the development of next-generation sequencing technologies, making possible a deep knowledge of the tomato genome. The decline of costs and the solutions offered by diverse NGS platforms paved the way for the exploration of the diversity enclosed in the germplasm collection. The research community is benefiting from the knowledge to understand the role of genes underlying major traits of agricultural interest and for genomic-assisted breeding. Although in tomato high-quality genomes at deep sequencing scale have been obtained, the complete genomic information on the entire existing variability is still lacking. Among Solanaceous, such efforts have been done only in pepper [103]. Unlocking the diversity enclosed in over 10,000 existing varieties will provide novel opportunities for harnessing the potential stored in germplasm resources, facilitating their use in genetic improvement programs, and accelerating gene discovery and transfer.

CRediT authorship contribution statement

Pasquale Tripodi: Conceptualization, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

There isn’t any competing interest to declare.

Acknowledgements

This research was supported by the European Union Horizon 2020 Research and Innovation program for funding this research under grant agreement No 774244 (Breeding for Resilient, Efficient and Sustainable Organic Vegetable Production; BRESOV).

References

[1] FAOSTAT (2021). FAOSTAT. Available online at: http://www.fao.org/faostat/en/ (accessed October 16, 2022).

[2] Gerszberg A, Hnatuszko-Konka K, Kowalczyk T, Kόnonowicz AKK. Tomato (Solanum lycopersicum L.) in the service of bio-technology. Plant Cell Tiss Org Cult 2015;120:881–902. https://doi.org/10.1007/s11240-014-0664-4

[3] Tanksley, S. D., Canal, M. W., Prince, J. P., De Vicente, M. C., Bonierbale, M. W., Broun, P., ... & Martin, G. (1992). High density molecular linkage maps of the tomato and potato genomes. Genetics, 132(4), 1141-1160.genetic and physical maps.

[4] Haanstra JWP, Wye C, Verbakel H, Meijer-Dekens F, Van den Berg P, Odinot P, et al. An integrated high-density RFLP-AFLP map of tomato based on two
Sauvage C, Cañizares J, Cordero L, Pascual L, Diez MJ, Nuez F. Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. Plant One 2012;7(10):e48198.

Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, et al. Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol 2014;165(3):1120–32.

Blanca J, Montero-Pau J, Sauvage C, Bauchet G, Illa E, Diez MJ, et al. Genetic variation in tomato, from wild ancestors to contemporary breeding accessions. BMC Genomics 2015;16(1):1–19.

Schouten H, Tikunov Y, Verkerke W, Finkers R, Bovy A, Bai Y, et al. Breeding has increased the diversity of cultivated tomato in The Netherlands. Front Plant Sci 2019;10:1606.

Sacco A, Ruggieri V, Parisi M, Festa G, Rigano MM, Picarella ME, et al. Use of modern tomato breeding germplasm for deciphering the genetic control of agronomical traits by Genome Wide Association study. Theor Appl Genet 2017;130(5):875–89.

Ruggieri V, Francone G, Sacco A, D’Alessandro A, Rigano MM, Parisi M, et al. An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC Plant Biol 2014;14(1):1–15.

Roozbahaianzadeh R, de Maagd RA, Lammers M, Molthoff J, Meijer-Dekens F, et al. Exploring a reservoir genome of tomato core collection for phenotypic and genotypic variation in plant growth and fruit quality traits. Genes 2020;11(11):1278.

Yamamoto E, Matsuana H, Omogi A, et al. A simulation-based breeding design that uses whole-genome prediction in tomato. Sci Rep 2016;6:19454. https://doi.org/10.1038/srep19454.

Kim M, Nguyen TTP, Ahn JH, Kim GJ, Sim SC. Genome-wide association study identifies QTL for eight fruit traits in cultivated tomato (Solanum lycopersicum L.). Hortic Res 2021;8.

Wickland DP, Battu G, Hudson KA, et al. A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow. GB-eaSy BMC Bioinformatics 2017;18:586. https://doi.org/10.1186/s12859-017-0200-6.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genomic-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 2011;12(7):499–510.

Poland JA, Rife TW. Genotyping-by-sequencing for plant breeding and population structure analysis in cultivated tomato provides new insights into the genetic diversity of Mediterranean ‘da serbo’ type long shelf-life germplasm. Hortic Res 2020;7:1–11. https://doi.org/10.1038/s41438-020-00353-6.

Blanca J, Cañizares J, Cordero L, Pascual L, Diez MJ, Nuez F. Variation revealed by SNP genotyping and morphology provides insight into the origin of the tomato. Nat Commun 2020;11:5817. https://doi.org/10.1038/s41467-020-19682-0.

Takes H, Shirasawa K, Kuzuraba K, Toyoda A, Matsuzawa Y, Iioka S, et al. De novo genome assembly of two tomato ancestors, Solanum pimpinellifolium and Solanum lycopersicum var. cerasiforme, by long-read sequencing. DNA Res 2021;28(1):dss029.

van Rens W, Schmidt MH, Effgen S, DB, Wang Y, Zaidan MWAM, et al. A chromosome scale tomato genome built from complementary PacBio and Nanopore sequences alone reveals extensive linkage drag during breeding. Plant J 2022;110:522–3.

Su X, Wang B, Feng X, et al. A high-continuity and annotated tomato reference genome. BMC Genomics 2021;22:858. https://doi.org/10.1186/s12864-021-08212-x.

Powell AF, Feder A, Li J, Schmidt MHW, Courtenay L, Aliseek S, et al. A Solanum lycopersicoides reference genome facilitates insights into tomato specialized metabolism and immunity. Plant J 2022.

Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. Cell 2020;182(1):145–61.

Tettelin H, Maximiuk V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial “pan-genome”. Proc Natl Acad Sci 2005;102(39):13950–5.

Della Coletta R, Qiu Y, Ou S, et al. How the pan-genome is changing crop genomes and empowers tomato breeding. Nature 2022;606:527–34. https://doi.org/10.1038/s41586-022-04808-9.

Bolger A et al. The genome of the stress-tolerant wild tomato species Solanum pennellii. Nat Genet 2014;46:1034–8. https://doi.org/10.1038/ng.3046.

Edwards D, Batley J. Graph panogenomes find missing heritability. Nat Genet 2022;54:919–20. https://doi.org/10.1038/d41588-022-01099-8.

Strickler SR et al. Comparative genomics and phylogenetic discordance of cultivated tomato and close wild relatives. PeerJ 2015;3:e793.

Tripodi P, Rabanus-Wallace M, Barchi L, Kale S, Esposito S, Acquadro A, et al. Global range expansion history of pepper (Capsicum spp.) revealed by over 10,000 genebank accessions. Proc Natl Acad Sci 2021;118(34).

Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z, et al. Genomic analyses provide insights into the history of tomato breeding. Nat Genet 2014;46:1220–6. https://doi.org/10.1038/ng.3117.