International Journal of
Molecular Sciences

MDPI

Article

Next-Generation Sequencing of Local Romanian Tomato Varieties and Bioinformatics Analysis of the *Ve* Locus

Anca-Amalia Udriște [1], Mihaela Iordachescu [1,*], Roxana Ciceoi [1] and Liliana Bădulescu [2,*]

[1] Research Center for Studies of Food Quality and Agricultural Products, University of Agronomic Sciences and Veterinary Medicine of Bucharest, 59, Mărăști Bd., 011464 Bucharest, Romania
[2] Faculty of Horticulture, University of Agronomic Sciences and Veterinary Medicine of Bucharest, 59, Mărăști Bd., 011464 Bucharest, Romania
* Correspondence: mihaela.iordachescu@qlab.usamv.ro (M.I.); liliana.badulescu@usamv.ro (L.B.)

Abstract: Genetic variability is extremely important, not only for the species' adaptation to environmental challenges, but also for the creation of novel varieties through plant breeding. Tomato is an important vegetable crop, as well as a model species in numerous genomic studies. Its genome was fully sequenced in 2012 for the 'Heinz 1706' variety, and since then, resequencing efforts have revealed genetic variability data that can be used for multiple purposes, including triggering mechanisms of biotic and abiotic stress resistance. The present study focused on the analysis of the genome variation for eight Romanian local tomato varieties using next-generation sequencing technique, and as a case study, the sequence analysis of the *Ve1* and *Ve2* loci, to determine which genotypes might be good candidates for future breeding of tomato varieties resistant to *Verticillium* species. The analysis of the *Ve* locus identified several genotypes that could be donors of the *Ve1* gene conferring resistance to *Verticillium* race 1. Sequencing for the first time Romanian genotypes enriched the existing data on various world tomato genetic resources, but also opened the way for the molecular breeding in Romania. Plant breeders can use these data to create novel tomato varieties adapted to the ever-changing environment.

Keywords: *Solanum lycopersicum* L.; NGS; genetic variability; biotic stress; Romanian tomato; *Verticillium* wilt

Citation: Udriște, A.-A.; Iordachescu, M.; Ciceoi, R.; Bădulescu, L. Next-Generation Sequencing of Local Romanian Tomato Varieties and Bioinformatics Analysis of the *Ve* Locus. *Int. J. Mol. Sci.* **2022**, *23*, 9750. https://doi.org/10.3390/ijms23179750

Academic Editor: Alina Maria Holban

Received: 19 July 2022
Accepted: 25 August 2022
Published: 28 August 2022

1. Introduction

Tomato (*Solanum lycopersicum* L.) fruits have substantial nutraceutical qualities, being an important source of fibers, lycopene and other carotenoids, vitamin C, and potassium, their consumption reducing the risk of certain cancers, cardiovascular disease, ultraviolet-light-induced skin damage, and osteoporosis [1,2]. Additionally, tomato crop production is important from an economical point of view, as it is the second most important fruit or vegetable crop next to potato (*Solanum tuberosum* L.) being cultivated worldwide [3].

Tomato belongs to the Solanaceae family, and it has been used as a model in molecular genetic and genomic studies for fruiting plants because of its diploid, compact genome (about 950 Mb) [4]. Thus, the tomato whole genome of the 'Heinz 1706' variety has been the first to be sequenced, after a 10-country and 8-year collaborative effort, using a combination of classical Sanger sequencing and the emerging new NGS sequencing, the work being completed in 2012 [5,6].

Whole-genome resequencing via Illumina platforms, based on the mechanism of SBS (sequencing by synthesis), has become the most rapid and effective method to identify the genetic variations in individuals of the same species or between related species. The various data, such as single nucleotide polymorphism (SNP), insertion and deletion (InDel), copy number variation (CNV), and structural variation (SV), obtained through resequencing is used in population genetics research and genome-wide association studies (GWAS) to investigate the mechanisms of biotic and abiotic stress resistance, to select plants and

animals for agricultural breeding programs, to identify common genetic variations among populations, and more [7–10].

Whole-genome resequencing (WGRS) analysis represents a powerful strategy for rapid identification of candidate genes responsible for traits of interests [7,11–13]. One of the fungal diseases that affects many crops, causing yield and quality losses, is *Verticillium* wilt, due to infection with *Verticillium* sp. In tomato, two linked genes providing resistance to *Verticillium dahliae* race 1, *Ve1* and *Ve2*, located in the *Ve* locus in chromosome 9, have been cloned, genes putatively encoding cell surface-like receptors [14]. Fradin et al., in 2009, showed that, out of the two genes, *Ve1* was the one actually providing the resistance to *V. dahliae* and *V. albo-atrum*. Sequence analysis of the two genes revealed that one deletion in *Ve1* resulted in the production of a truncated protein, which was present in all susceptible genotypes analyzed, thus providing a putative marker that could be used by the plant breeders to discriminate between the resistant and susceptible genotypes [15].

In the present study, we analyzed genome variation from the perspective of number and distribution of SNPs, InDels, SVs, and CNVs for eight Romanian local tomato varieties. In addition, as a case study, we analyzed the sequence of *Ve1* and *Ve2* loci in these Romanian tomato genotypes in order to determine which of them might be good candidates for future breeding of tomato varieties resistant to *Verticillium* species.

2. Results

2.1. NGS Data Analysis

2.1.1. Sequencing Data Quality Control

The genomes of eight Romanian tomato varieties were sequenced using NGS technology. Sequencing quality distribution was examined over the full length of all sequences, to detect any sites with an unusually low sequencing quality, where incorrect bases may have been incorporated at abnormally high levels. Q30 is considered a benchmark for quality in next-generation sequencing [16]. For the present sequencing, data results showed that Q30 was over 90% for all studied genotypes, and the ratio of clean data to raw data (effective rate) was around 99% (Supplementary Table S1).

2.1.2. SNP Detection, Distribution, and Mutation Frequency

SNP variations were detected in all eight studied genotypes; however, their number and distribution within the genomes varied among the genotypes.

The SNPs were distributed in all regions of the genomes: upstream, exonic (stop gain, stop loss, synonymous, nonsynonymous), intronic, splicing, downstream, upstream/downstream, intergenic, others (Supplementary Table S2). A total of 2,964,636 SNPs were identified within the genotypes studied, ranging from 223,072 in Buzău 1600 to 697,473 in Florina 44. The highest numbers of SNPs were detected in the intergenic regions for all genotypes (between 190,064/85.14% for Buzău 1600 and 623,805/89.42% for Florina 44) (Supplementary Figure S1).

For all studied genotypes, the number of transitions (ts), point mutations that change a purine nucleotide to another purine or a pyrimidine nucleotide to another pyrimidine, was higher than the number of transversions (tv), point mutations that substitute a purine for a pyrimidine or vice versa. However, the ratio ts/tv was similar for all genotypes (between 1.294 for Florina 44 and 1.394 for Ștefănești 24). For the exonic SNPs, for all genotypes studied, the number of nonsynonymous SNPs was higher than the synonymous ones, the lowest number of nonsynonymous SNPs being observed for Buzău 1600 (3010/1943, respectively, 58.42%/37.71%) and the highest for Florina 44 (5904/4062 respectively 57.51%/39.57%). Stop gain point mutations were also more numerous than stop loss ones in all genotypes under study, again the least numerous being observed in Buzău 1600 (78/23, respectively, 1,51%/0.45%), and the highest number being observed in Florina 44 (153/46, respectively, 1.49%/0.45%). The highest heterozygosity rate (‰) was observed for the genotype Kristinica (0.234), whereas the lowest rate was observed for the genotype Buzău 47 (0.115).

The distribution of the six types of SNP mutations is displayed in Figure 1. The highest number of SNPs was observed in the genotype Florina 44 for all six SNP types, and the lowest number was observed for the genotype Buzău 1600. Among the six types of SNP mutations, for all genotypes, the highest number of SNPs was the T:A > C:G type, followed by C:G > T:A, whereas the lowest number of SNPs was the C:G > G:C type.

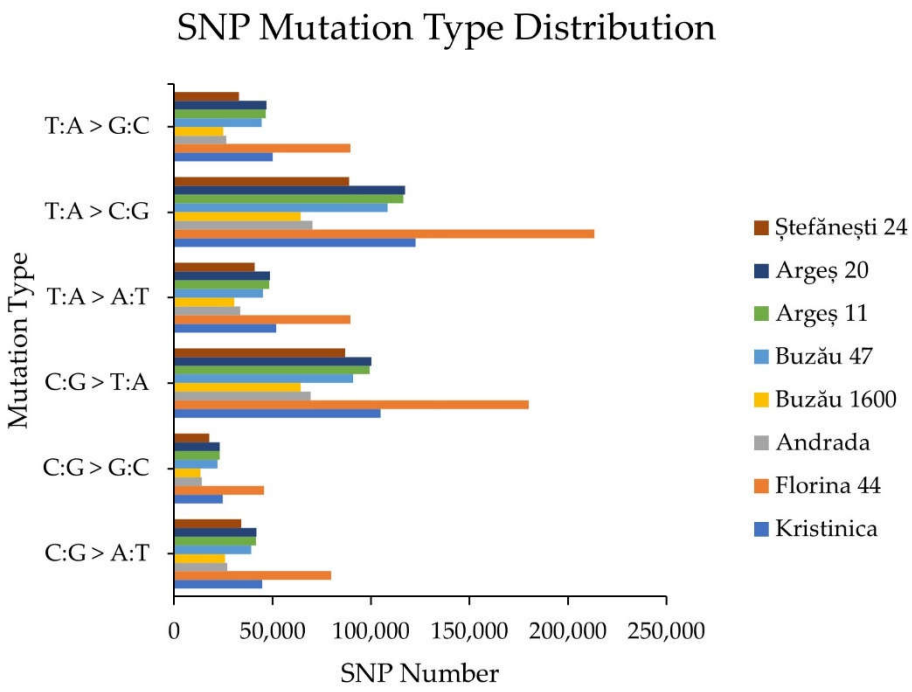

Figure 1. SNP mutation type distribution.

2.1.3. Insertion/Deletion Detection and Distribution

InDel variations were detected in all eight studied genotypes, and like in the case of SNPs, their number and distribution within the genomes varied from genotype to genotype.

The InDels were distributed in all regions of the genomes: upstream, exonic (stop gain, stop loss, synonymous, nonsynonymous), intronic, splicing, downstream, up-stream/downstream, intergenic, others (Supplementary Table S3). A visual representation of these data is visible in Supplementary Figure S2.

A total of 622,988 InDels were identified within the studied genotypes, ranging from 66,768 in Buzău 1600 to 110,138 in Florina 44. For all genotypes, the total number of insertions was roughly double the total number of deletions. The highest number of InDels was detected in the intergenic regions for all genotypes (between 47,309 in Buzău 1600 and 82,716 in Florina 44), followed by the intronic (between 8360 in Buzău 1600 and 12,236 in Florina 44), upstream (between 4584 in Argeș 20 and 6341 in Florina 44), downstream (between 2915 in Buzău 47 and 4157 in Florina 44), upstream/downstream (between 357 in Buzău 47 and 578 in Florina 44), exonic (between 338 in Buzău 1600 and Buzău 47 and 493 in Florina 44), and splicing regions (between 32 in Buzău 47 and 52 in Florina 44). For the exonic InDels, for all genotypes, the number of frameshift InDels were higher than the non-frameshift ones. The highest number of frameshift deletions was detected in the Florina 44 genotype (163), and the lowest number in the genotype Buzău 1600 (115), and for the frameshift insertions the highest number was observed in genotype Ștefănești 24 (129) and the lowest number in the genotype Buzău 47 (102). As for the non-frameshift InDels, the highest number of deletions was detected in the genotype Ștefănești 24 (101) and the lowest number in the genotype Buzău 47 (47).

InDels distribution within the coding sequence is portrayed in Figure 2. The highest number of InDels was observed for the 1 bp insertion/deletion and decreased with the increase in sequence length. Thereafter, the percentages for sequences with lengths multiples

of three bp were higher than those of other lengths. The reason for these higher percentages might be that these sequences do not cause frameshifts and, subsequently, premature STOP codons. InDels with lengths beyond 19 bp were below 1%.

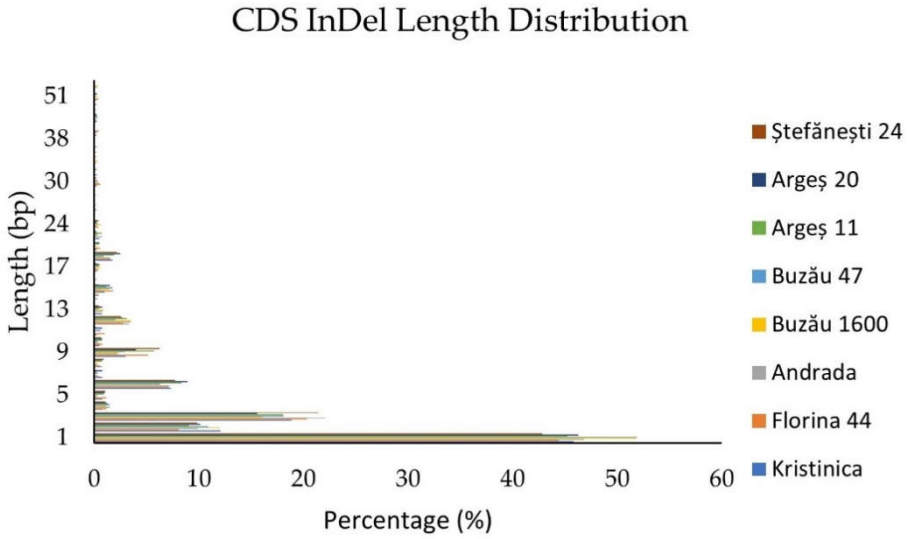

Figure 2. InDels distribution within the coding sequence.

InDels distribution within the whole genome is presented in Figure 3. The highest percentage of InDels was again observed for the 1 bp insertion/deletion and decreased gradually with the increase in sequence length.

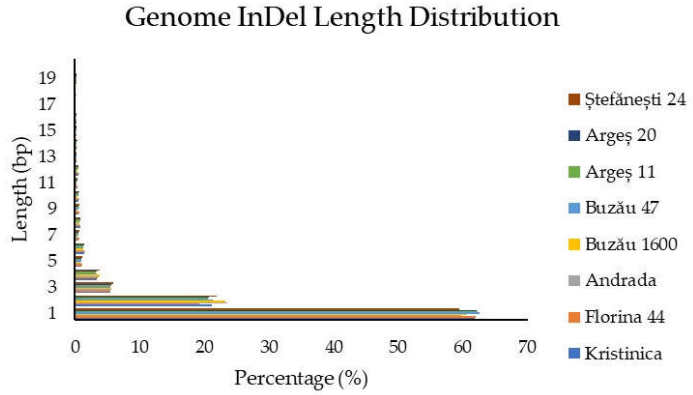

Figure 3. InDels distribution within the genome.

For the eight studied genotypes, the SNP and InDel densities on each chromosome were similar (Figure 4) and varied as follows. For the genotype Kristinica, the highest density was noted in chromosome 11, followed by chromosome 4. InDel density was also noted in chromosomes 2, 5, and 9.

For the Florina 44 variety, the highest density was observed in chromosome 4, followed by chromosome 11. For the Andrada variety, the highest density was observed in chromosome 4, but high densities were also noted for chromosomes 2, 5, 9, and 11. For the Buzău 1600 variety, the highest density was observed in chromosome 4, with high density being present as well in chromosomes 5, 9, and 11. In the case of the Buzău 47 variety, chromosome 11 presented the highest density, and high density existing as well in chromosomes 4, 5, 8, and 9. For the variety Argeș 11, the highest density was observed in chromosome 11, but high densities were also observed in chromosomes 2, 3, 4, 5, 9, and 10. For the Argeș 20 variety, again the highest density was noted in chromosome 11, with high densities in chromosomes 2, 4, 5, 7, 9, and 10. Lastly, Ștefănești 24 presented high densities in chromosomes 4, 11, and 12, and, to a lesser extent, in chromosomes 1, 5, 6, and 7.

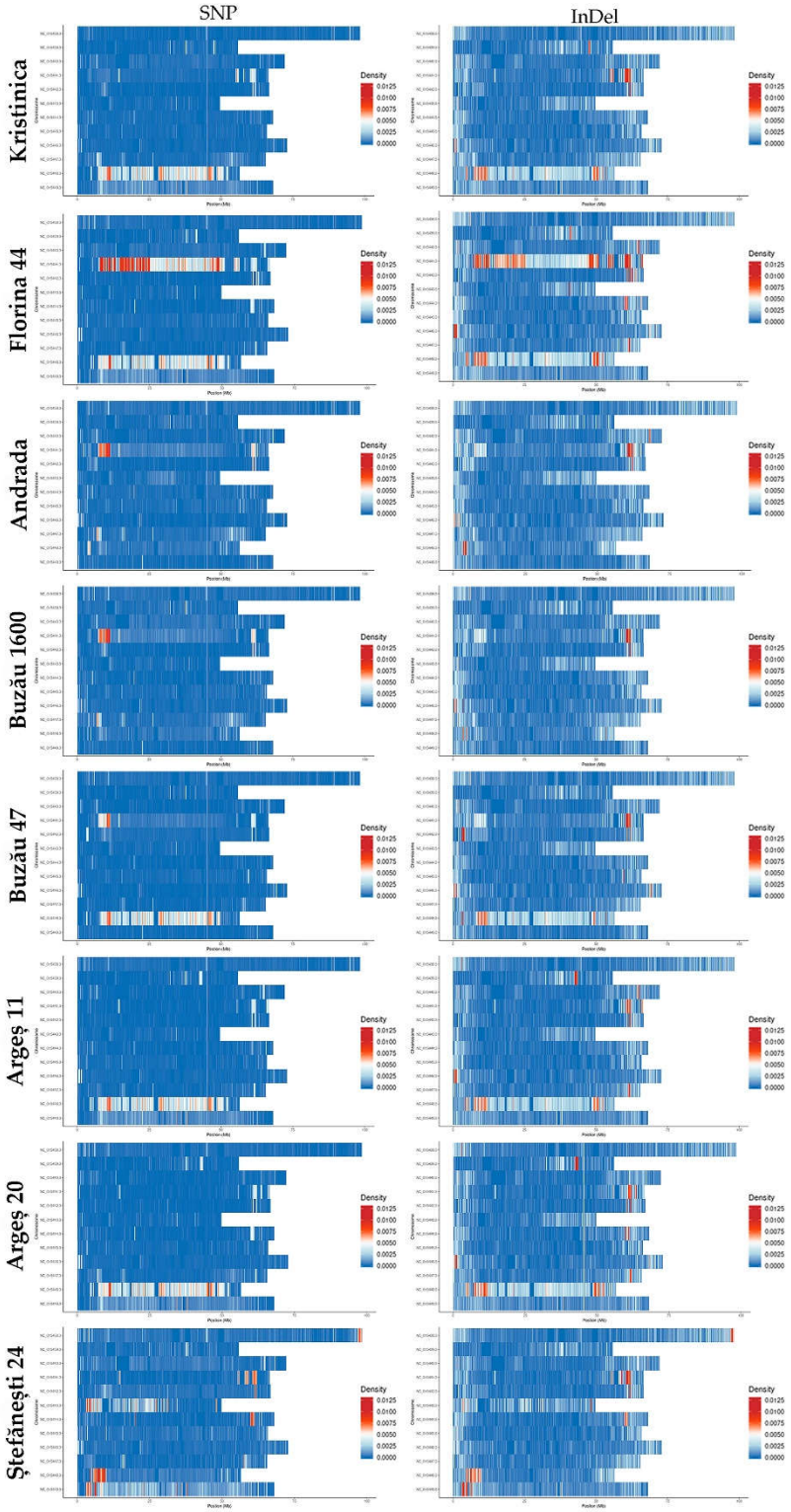

Figure 4. SNP and InDel densities per chromosome per genotype.

2.1.4. Structural Variant Detection and Annotation

Structural variations were detected in all eight studied genotypes, and their number and distribution within the genomes varied from genotype to genotype.

The SVs were distributed in all regions of the genomes: upstream, exonic, downstream, intronic, upstream/downstream, splicing, intergenic, and others (Supplementary Table S4). A visual representation of these data is visible in Supplementary Figure S3. The highest

number of SVs was observed in the intergenic regions (between 1623 in Argeș 20 and 2663 in Andrada), followed by the exonic (between 446 in Argeș 20 and 817 in Florina 44), intronic (between 96 in Buzău 47 and 169 in Andrada), upstream (between 55 in Buzău 47 and 109 in Andrada), downstream (between 32 in Argeș 20 and 74 in Andrada), upstream/downstream (between 6 in Buzău 1600 and 11 in Florina 44), and splicing (between none in Buzău 1600 and Argeș 11 and 3 in Argeș 20) regions. The distribution of the five types of SVs is visible in Figure 5.

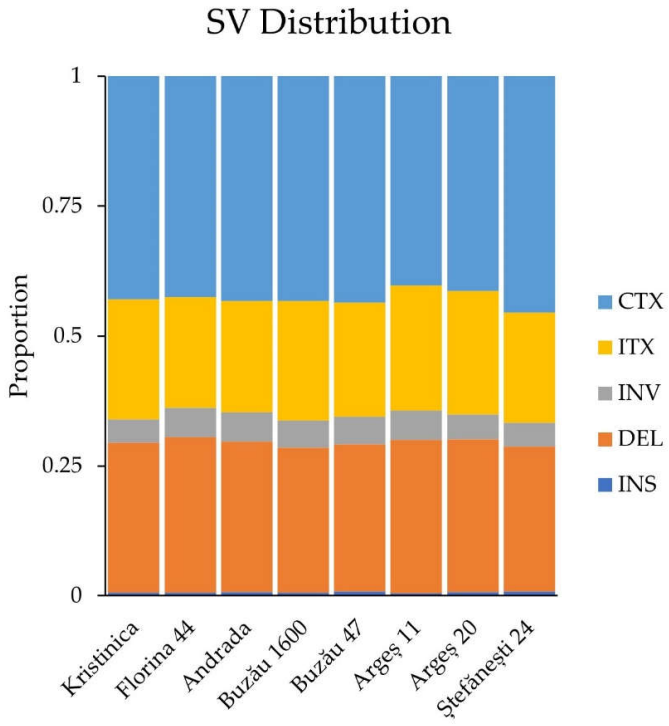

Figure 5. SV distribution within genotypes studied. CTX—interchromosomal translocation, DEL—deletion, INS—insertion, INV—inversion, ITX—intrachromosomal translocation.

The highest percentage of SVs was interchromosomal translocations, followed by the deletions, intrachromosomal translocations, inversions, and insertions. More than 40% of SVs were longer than 1200 bp. Approximately 20% of the SVs had a length of 200–300 bp. For the rest of the SVs' length-size categories, the percentages were lower than 7%. The rarest SVs were those with a length of less than 100 bp (~0.4–0.7%) (Figure 6).

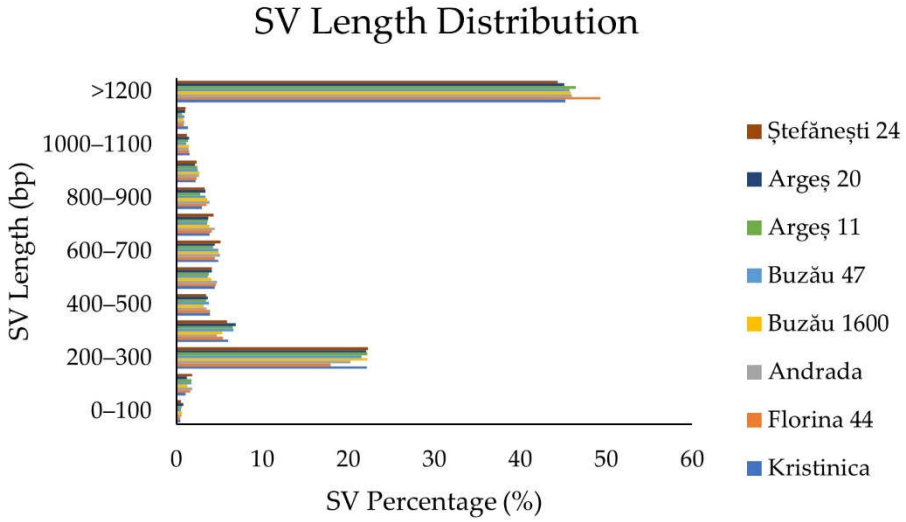

Figure 6. Structural variations' (SVs) length distribution.

2.1.5. Copy Number Variations Detection and Annotation

For all genotypes studied, the number of deletions (between 7247-Argeș 11 and 13,350-Andrada) was higher than the number of duplications (between 1265-Buzău 1600 and 1824-Argeș 20). The highest numbers of CNVs were detected within the intergenic regions, followed by the exonic regions. The lowest numbers of CNVs were noted within the upstream/downstream regions (Supplementary Table S5). A visual representation of these data is visible in Supplementary Figure S4.

2.2. Sequence Analyses of Ve1 and Ve2 Loci in Romanian Tomato Genotypes

To identify the sequences of *Ve1* and *Ve2* homologous genes in the studied genotypes, the NGS bam files containing the eight genomes were aligned with the reference genome using the Workbench software. For the *Ve1* locus, the genotypes Kristinica, Florina 44, Buzău 1600, Argeș 11, and Argeș 20 were identical with the gene from the Heinz 1706 reference genome. For the *Ve2* locus, the only genotypes that showed differences compared with the reference genome were Andrada and Buzău 47. Thereafter, the nucleotide sequences for the eight genotypes were aligned with previously published sequences of the genes.

For *Ve1*, the eight Romanian genotypes were aligned and compared with the following sequences: NC_015446.3, reference genome sequence of Heinz 1706 [17]; AF272366.2, of the Ailsa Craig genotype [14]; FJ464557.1, of the VFN-8 genotype; FJ464556.1, of the Motelle genotype; FJ464555.1, of the Moneymaker genotype; FJ464554.1, of the Craigella GCR26 genotype; and FJ464553.1, of the Craigella GCR218 genotype [15]. In the case of the *Ve1* locus, 9 SNPs were identified (Supplementary Figures S5 and S6). In the case of the first 2 SNPs, only the genotype Ailsa Craig was different, with a cytosine inserted at position 29 and also a cytosine deleted at position 35, resulting in a PMV translation instead of LWL. At position 246, the SNP presents a silent mutation, G/C. At position 380, the SNP C/A resulted in an A/D amino acid translation. The genotypes Andrada, Buzău 47, Ailsa Craig, Moneymaker, and Craigella GCR25 had a cytosine at this position, whereas the rest contained an A. At position 610, the SNP A/T translated into a T/S amino acid. Ailsa Craig, Moneymaker, and Craigella GCR26 had an adenine at this position, while the rest contained a thymine. At position 706, there is another SNP A/T, again translated into T/S; however, this time, A is present in Andrada, Buzău 47, Ștefănești 24, Ailsa Craig, Moneymaker, and Craigella GCR26. A single nucleotide deletion exists at position 1220, resulting in a premature stop codon in Andrada, Buzău 47, Ștefănești 24, Moneymaker, and Craigella GCR26. At position 1548, the SNP C/G translates into N for the varieties that do not have a deletion at position 1220. All the varieties that contain guanine at this position are producing the truncated Ve1 protein. Lastly, at position 1888, the SNP G/A translates only into D, since all the varieties that contain adenine at this position are producing the truncated Ve1 protein due to the deletion at position 1220 (Table 1).

Table 1. Sequence analysis of the *Ve1* gene.

SNP Position	DNA Sequence	Amino Acid Sequence	Genotypes									
			Heinz 1706	Craigella GCR218	Motelle	VFN-8	Ailsa Craig	Craigella GCR26	Moneymaker	Kristinica, Florina 44, Buzău 1600, Argeș 11, Argeș 20	Andrada, Buzău 47	Ștefănești 24
29/35	CCTATGGTT	PMV		-	-	-	+	-	-	-	-	-
	CTATGGCTT	LWL	+	+	+	+	-	+	+	+	+	+

Table 1. *Cont.*

SNP Position	DNA Sequence	Amino Acid Sequence	Genotypes									
			Heinz 1706	Craigella GCR218	Motelle	VFN-8	Ailsa Craig	Craigella GCR26	Moneymaker	Kristinica, Florina 44, Buzău 1600, Argeș 11, Argeș 20	Andrada, Buzău 47	Ștefănești 24
246	GTG	Silent	+	+	+	+	-	-	-	+	-	+
	GTC		-	-	-	-	+	+	+	-	+	-
380	GAC	D	-	-	+	-	+	+	-	-	+	-
	GCC	A	+	+	-	+	-	-	+	+	-	+
610	ACT	T	+	+	+	+	-	-	-	+	+	+
	TCT	S	-	-	-	-	+	+	+	-	-	-
706	ACT	T	-	-	-	-	+	+	+	-	+	+
	TCT	S	+	+	+	+	-	-	-	+	-	-
1220	TCAGAG	SE	+	+	+	+	-	-	-	+	-	-
	TAGAG	STOP	-	-	-	-	+	+	+	-	+	+
1548	AAC	N	+	+	+	+	-	-	-	+	-	+
	AAG	K *	-	-	-	-	+	+	+	-	+	-
1888	GAC	D	+	+	+	+	-	-	-	+	-	-
	AAC	N *	-	-	-	-	+	+	+	-	+	+

* The amino acids are not translated due to the STOP codon positioned upstream of these sequences. "-"/"+" denotes the absence/presence of the SNP. The underlined sequence encodes the STOP codon.

In the case of *Ve2*, the eight Romanian genotypes were aligned and compared with the following sequences: NC_015446.3, of the reference genome sequence Heinz 1706 [17]; AF365930.1, of the Ailsa Craig genotype [14]; FJ464562.1, of the genotype VFN-8; FJ464561.1, of the Motelle genotype; FJ464560.1, of the Moneymaker genotype; FJ464559.1, of the Craigella GCR218 genotype; and FJ464558.1 of the Craigella GCR26 genotype [15]. Again, 9 SNPs were identified (Supplementary Figures S7 and S8). The first SNP, G/C, at position 1385, is translated into an R/T amino acid, the Romanian genotypes Andrada and Buzău 47 being the only ones that contain threonine. The second, at position 1811, C/T, translates into A/V, with Andrada, Buzău 47, Moneymaker, Craigella GCR26, and Ailsa Craig containing thymine. At position 2761, the SNP G/A translates into D/N, with Moneymaker being the only one that has adenine. At position 2771, the SNP C/G translates into T/R, with Andrada, Buzău 47, Moneymaker, Craigella GCR26, and Ailsa Craig having guanine. At position 2893, the SNP C/T translates into P/S, with Andrada, Buzău 47, VFN-8, Motelle, and Craigella GCR218 containing cytosine. The next two SNPs, at positions 2934 and 3243, are silent. Finally, at positions 3380 and 3383, T/C translates into F/S, with only Ailsa Craig containing thymines, TTTTTT vs. TCTTCT in the other genotypes (Table 2).

Table 2. Sequence analysis of the *Ve2* gene.

SNP Position	DNA Sequence	Amino Acid Sequence	Genotypes									
			Heinz 1706	Craigella GCR218	Motelle	VFN-8	Ailsa Craig	Craigella GCR26	Moneymaker	Kristinica, Florina 44, Buzău 1600, Argeș 11, Argeș 20	Andrada, Buzău 47	Ștefănești 24
1385	ACA	T	-	-	-	-	-	-	-	-	+	-
	AGA	R	+	+	+	+	+	+	+	+	-	+
1811	GTA	V	-	-	-	-	+	+	+	-	+	
	GCA	A	+	+	+	+	-	-	-	+	-	+
2761	GAC	D	+	+	+	+	+	+	-	+	+	+
	AAC	N	-	-	-	-	-	-	+	-	-	-
2771	AGA	R	-	-	-	-	-	+	+	-	+	-
	ACA	T	+	+	+	+	+	-	-	+	-	+
2893	CCA	P	-	+	+	+	-	-	-	-	+	-
	TCA	S	+	-	-	-	+	+	+	+	-	+
2934	CTC	Silent	-	-	-	-	+	+	+	-	-	-
	CTT		+	+	+	+	-	-	-	+	+	+
3243	GGT	Silent	-	-	-	-	+	+	+	-	+	-
	GGG		+	+	+	+	-	-	-	+	-	+
3380	TTT	T	-	-	-	-	+	-	-	-	-	-
	TCT	S	+	+	+	+	-	+	+	+	+	+
3383	TTT	T	-	-	-	-	+	-	-	-	-	-
	TCT	S	+	+	+	+	-	+	+	+	+	+

"-"/"+" denotes the absence/presence of the SNP.

3. Discussion

The first tomato genome to be sequenced, Heinz 1706, provided the 'golden standard' for future resequencing efforts [6]. With the advent of NGS, more and more tomato genotypes have been wholly sequenced, enriching the knowledge at the DNA level and offering new data that can be used in subsequent studies, as well as in breeding for improved tomato varieties. The present study contributes to this growing pool of sequenced genomes with whole-genome resequencing data from eight Romanian tomato genotypes.

SNP data mined from sequenced transcriptomes and from resequenced whole genomes through next-generation sequencing have been used to study the diversity within cultivated tomato genotypes, as well as between cultivated tomatoes and wild-type relatives [12,18–22]. The present study reports almost 3 million SNPs, adding to/confirming those reported by the 100 Tomato Genome Sequencing Consortium [12] and Causse et al., 2013.

SNPs and InDels were not evenly distributed within the genome, for each genotype existing certain 'hot spots', where there was a higher density of SNPs/InDels, mostly toward the ends of chromosomes, but there were also observed wide regions with a high density of SNPs/InDels spanning almost the whole chromosome: chromosome 11 in all genotypes except Andrada, Buzău 1600, and Ștefănești 24; chromosome 4 in Florina 44; and chromosome 6 in Ștefănești 24. Interestingly, for each genotype, the 'hot spots' for SNPs

and InDels overlapped (Figure 4). The higher polymorphism towards the chromosomes' ends can be explained by the higher recombination frequency of these regions [19,23]. The broad regions with high polymorphism density were also observed in other genotypes in previously published studies, but on different chromosomes [20], most probably due to the introgressions from wild-type relatives, depending on each genotype breeding history. For instance, chromosome 11 of the Heinz 1706 genotype contains large introgressions from *S. pimpinellifolium*, having received them through disease resistance [6].

The highest numbers of SNP types were associated with T:A > C:G and C:G > T:A transitions. The prevalence of transitions as opposed to transversions has been observed in numerous other species, and is explained by the high frequency of the cytosine-to-thymine mutation following the deamination of methylated cytosine residues [24].

Structural variations, such as deletions, insertions, copy number variations, inversions, and translocations, play a major role in heritable phenotypic diversity within and between species, as they could lead to gene loss, gene duplication, and the creation of new genes [25]. If high-throughput short read sequencing is extremely efficient in detecting SNPs and small InDels, the short read length makes it difficult to characterize repetitive regions, and hence to detect efficiently structural variations [25]. Nevertheless, there are sequencing techniques that overcome these difficulties. For instance, long read nanopore sequencing significantly improves the success in identifying structural variations. If, in the present study, between ~7500 and 10,400 SV per genotype were identified, Alonge et al., 2020, identified almost 240,000 SVs in 100 tomato accessions. In addition, if in the present study, the highest numbers of SVs were translocations, in the above-mentioned study, the most common SVs were insertions and deletions.

Copy number variations are part of the structural variations. Most CNVs studied so far were those that affect protein-coding sequences, and thus result in either gains or losses of gene copies, and ultimately in the regulation of plant development and plant adaptation to environmental factors [26]. In the current study, the exonic detected CNVs were between 507 for Andrada and 673 for Argeș 20, with an average of 577, a value similar to that observed in the Causse et al. study, 2013 [20]. However, as mentioned before, with the complexity of the plant genomes added to the short read sequencing, their complete detection is difficult [26].

Sequence Analyses of the Ve1 and Ve2 Loci in the Romanian Tomato Genotypes

The *Ve* locus in tomato comprises two genes that encode proteins involved in both stress/defense and plant growth [27]. *Ve1* expression is induced by various stress conditions, both biotic and abiotic, whereas *Ve2* is constitutively expressed [27].

In the case of the *Ve1* gene, the genotypes Andrada, Buzău 47, and Ștefănești 24 present the single nucleotide deletion at position 1220 that results in the premature stop codon and putative production of truncated protein. The presence of the *Ve1* allele in these genotypes implies that they are susceptible to *Verticillium* race 1. The other genotypes are identical at the amino acid level with Motelle, VFN-8, and Craigella GCR218 genotypes, which were proved to be resistant to *Verticillium* race 1, and thus good donors of the *Ve1* allele in future breeding programs.

In the case of the *Ve2* gene, the putatively resistant genotypes (Kristinica, Florina 44, Buzău 1600, Argeș 11, and Argeș 20) are identical at the amino acid level with the resistant genotypes Motelle, VFN-8, and Craigella GCR218, except for position 965, which contains a serine instead of a proline. If, initially, it was thought that only *Ve1* had a role in plant resistance to *Verticillium*, later, it was proved that the mechanism of resistance was more complex than originally thought, and both *Ve1* and *Ve2* are involved in the process. A study where *Ve2* gene expression was suppressed via RNAi demonstrated pronounced effects on defense/stress gene expression [28]. Even though the silencing of *Ve2* does not increase the susceptibility of either resistant or susceptible genotypes to *Verticillium*, in the resistant genotypes infected with *Verticillium* race 1, the silencing induces repression of multiple genes with a role on defense/stress, whereas in the susceptible genotypes that

are missing a functional Ve1 protein, continuous Ve2 signaling is sufficient to produce a normal defense/stress response [28]. It remains to be seen in future studies if the change to serine at position 965 has a significant effect on the way the plants are coping with the *Verticillium* attack.

4. Materials and Methods

4.1. Plant Material

Eight Romanian tomato varieties, Kristinica, Florina 44, Andrada, Buzău 1600, Buzău 47, Argeș 11, Argeș 20, and Ștefănești 24, were analyzed in the present study. Tomato seeds received from the Vegetable Research and Development Station Buzău and the National Research and Development Institute for Biotechnology in Horticulture Ștefănești-Argeș were cultivated under greenhouse conditions (18–25 °C) in the Research Center for Studies of Food Quality and Agricultural Products of the University of Agronomic Sciences and Veterinary Medicine of Bucharest, Romania.

4.2. DNA Extraction

Genomic DNA was extracted from fresh leaves of tomato seedlings using an automated extraction system (InnuPure C16, Analytik Jena GmbH, Jena, Germany) based on the principle of magnetic particle separation for fully automated DNA isolation and purification. An InnuPREP Plant DNA I Kit-IPC16 (Analytik Jena GmbH, Jena, Germany) was used for genomic DNA extraction following the manufacturer's instructions. A preliminary processing step was the external lysis of the starting material. The plant sample was ground to powder in the presence of liquid nitrogen and homogenized with SLS lysis solution (containing CTAB as detergent component), proteinase K, and RNase A solution. After external lysis, the extraction proceeded with automated DNA extraction following the manufacturer's instructions. The DNA was quantified using a NanoDrop^{TM} 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE).

4.3. Sequencing, Computational Data Processing, and Sequencing Analysis

Whole-genome sequencing (WGS) was performed via an Illumina platform (NGS) by Novogene Co., Ltd., Cambridge, UK. An original image data file from the high-throughput sequencing platform Illumina was transformed to sequenced reads (raw data) by CASAVA base recognition (Base Calling) (Novogene Co., Ltd., Cambridge, UK). Raw data were stored in FASTQ (.fq) format files [29], which contain sequencing reads and corresponding base quality. The effective sequencing data were aligned with the reference sequence through the BWA (Li H. et al. 2009) software [30] (parameters: mem -t 4 -k 32 -M), and the mapping rate and coverage were counted according to the alignment results. In order to obtain clean reads, low-quality reads or reads with adaptors that would affect the quality of downstream analysis were removed (Novogene Co., Ltd., Cambridge, UK). The Phred score (Q_{phred}), the quality score of a base, was calculated using the equation $Q_{phred} = -10\log_{10}€$, where 'e' represents the sequencing error rate.

The filtered reads were mapped onto the tomato genome SL3.0 [17], used as a reference sequence. The resultant sequence alignment format files were converted to binary sequence alignment format (*.bam) files and subjected to yield a variant file including SNP information. The mapping rates of samples reflect the similarity between each sample and the reference genome. The depth and coverage are indicators of the evenness and homology with the reference genome (Novogene Co., Ltd., Cambridge, UK).

4.3.1. SNP Detection and Annotation

Individual SNP variations were detected using SAMtools with the 'mpileup -m 2 -F 0.002 -d 1000' parameter [31] (Novogene Co., Ltd., Cambridge, UK). To reduce the error rate in SNP detection, the results were filtered using two criteria: the number of support reads for each SNP was higher than 4, and the mapping quality of each SNP, calculated by the root mean square of the support reads' mapping quality, was higher than 20. There-

after, the SNPs were annotated using the ANNOVAR software [32] (Novogene Co., Ltd., Cambridge, UK) in the following categories: upstream (located within 1 kb upstream away from transcription start site of the gene), exonic (located in the exonic region), intronic (located in the intronic region), splicing (located in the splicing site, within a 2 bp range of the intron/exon boundary), downstream (located within 1 kb downstream away from transcription termination site of the gene region), upstream/downstream (located within the less than 2 kb intergenic region, which is in 1 kb downstream or upstream of the genes), intergenic (located within the more than 2 kb intergenic region), and others (located in other region). The exonic category was further split into nonsynonymous (single-nucleotide mutation with changing the amino acid sequence), synonymous (single-nucleotide mutation without changing the amino acid sequence), stop gain (a nonsynonymous SNP that leads to the introduction of a stop codon at the variant site), and stop loss (a nonsynonymous SNP that leads to the removal of the stop codon at the variant site). The genome-wide heterozygous rate for SNPs (het rate (‰)) was calculated as the ratio of heterozygous SNPs to the total number of genome bases.

Based on the type of mutations, the SNPs were classified into six categories: T:A > C:G, T:A > G:C, C:G > T:A, C:G > A:T, T:A > A:T, and C:G > G:C. For instance, the T:A > C:G mutations include mutations from T to C and A to G. When a T-to-C (T > C) mutation appears on either of the double strand, the A-to-G (A > G) mutation will be found in the same position of the other chain. Therefore, the T > C and A > G mutations were classified into a single category.

4.3.2. Insertion/Deletion (InDel) Detection and Annotation

An InDel was defined as the insertion or deletion of a DNA sequence with a length of 50 bp or less. InDels were detected using SAMTOOLS [31] with the 'mpileup -m 2 -F 0.002 -d 1000' parameter and annotated using the ANNOVAR software [32] (Novogene Co., Ltd., Cambridge, UK). The filter conditions to reduce the error rate in InDel detections were the same as with the SNP detection.

The annotation of InDels was performed using the same categories for genomic regions as the SNPs, except for the exonic region, which was subdivided into stop gain and stop loss (same as SNPs), frameshift deletion and frameshift insertion (InDel mutation changing the open reading frame with deletion or insertion), and non-frameshift deletion and non-frameshift insertion (InDel mutation without changing the open reading frame with deletion or insertion sequences of 3 or multiple of 3 bases).

Length distribution of InDels was analyzed as a percentage within the coding sequence (CDS) and within the whole genome.

4.3.3. Structural Variant Detection and Annotation

Structural variants (SVs) were defined as genomic variations with mutations of a relatively larger size, more than 50 bp, such as deletions (DEL), insertions (INS), inversions (INV), intrachromosomal translocations (ITX), and interchromosomal translocations (CTX) and were detected by the BreakDancer software [33]. SVs that were not supported by at least two pair-end read alignments were removed from further analysis. The insertions, deletions, and inversions were annotated by the ANNOVAR software [32].

4.3.4. Copy Number Variation Detection and Annotation

Copy number variations (CNV) were defined as structural variations showing deletions or duplications in the genome. Based on the reads' depth of the reference genome, the CNVnator software [34] was used to detect CNVs of potential deletions and duplications with the parameter '-call 100'. The detected CNVs were further annotated by the ANNOVAR software [32].

4.4. Sequence Analysis of the Ve Locus

Next-generation sequencing BAM files containing the nucleotide sequence data for the eight tomato varieties studied were loaded onto the Workbench software and aligned to the reference genome. For each variety, the differences in nucleotide sequence were noted. Amino acid sequences for each variety were generated using the Sequence Manipulation Suite: ORF Finder software [35].

Nucleotide sequences of *Ve* genes and amino acid sequences of corresponding putative proteins from the Romanian genotypes included in this study and sequences of *Ve* genes reported previously were aligned using the MultAlin software [36].

5. Conclusions

In the present times, we face a race between plant breeders on the one hand, who are creating new crop plant varieties that are resistant or at least tolerant to pathogens and viruses, and biotic factors on the other hand, which are constantly mutating and developing new races/strains that overcome plant resistance [37]. The resequencing of new *L. esculentum* varieties will enable researchers to link phenotypical variations to their DNA sequence variation, uncovering new information for comparative genomics studies [38]. Therefore, rather than being an end point, by bringing up novel essential data, NGS brings to light a plethora of new questions and opens up new research directions. Some of the varieties studied, such as Buzău 1600 and Buzău 47, were created between 1970 and 1980, and are still appreciated by growers and consumers alike, possessing multiple traits that would recommend them as genitors in tomato breeding [39]. One of the goals of the research founded by the Romanian Ministry of Agriculture and Rural Development, of which this study is a part of, is to create a database of Romanian cultivars/varieties' genetic variations that could be used in the future by plant breeders for selecting genitors that could donate genes encoding desirable traits. As an *in silico* case study, the survey of the *Ve* locus permitted the selection of a number of genotypes (Kristinica, Florina 44, Buzău 1600, Argeș 11, and Argeș 20) that could be donors of the *Ve1* gene conferring resistance to *Verticillium* race 1 attack, since they have amino acid sequences identical to those of proven resistant genotypes. The selected genotypes will be assessed for the confirmation of fungal resistance by artificial inoculation with different races of *Verticillium* prior to their use in plant breeding.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23179750/s1.

Author Contributions: Conceptualization, A.-A.U., M.I., R.C. and L.B.; methodology, A.-A.U. and M.I.; software, A.-A.U. and M.I.; writing—original draft, A.-A.U. and M.I.; writing—review and editing: M.I., R.C. and L.B.; funding acquisition, A.-A.U. and L.B.; project administration, A.-A.U. and L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Romanian Ministry of Agriculture and Rural Development (MADR-Bucharest) under the agricultural research and development program 2019–2022, ADER 7.2.6 project.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We thank the teams from the Vegetable Research and Development Station Buzău and the National Research and Development Institute for Biotechnology in Horticulture Ștefănești-Argeș for providing the plant material used in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Burton-Freeman, B.; Reimers, K. Tomato Consumption and Health: Emerging Benefits. *Am. J. Lifestyle Med.* **2011**, *5*, 182–191. [CrossRef]
2. Felföldi, Z.; Ranga, F.; Socaci, S.A.; Farcas, A.; Plazas, M.; Sestras, A.F.; Vodnar, D.C.; Prohens, J.; Sestras, R.E. Physico-Chemical, Nutritional, and Sensory Evaluation of Two New Commercial Tomato Hybrids and Their Parental Lines. *Plants* **2021**, *10*, 2480. [CrossRef] [PubMed]
3. Quinet, M.; Angosto, T.; Yuste-Lisbona, F.J.; Blanchard-Gros, R.; Bigot, S.; Martinez, J.-P.; Lutts, S. Tomato Fruit Development and Metabolism. *Front. Plant Sci.* **2019**, *10*, 1554. [CrossRef] [PubMed]
4. Bernatzky, R.; Tanksley, S.D. Toward a saturated linkage map in tomato based on isozymes and random cDNA sequences. *Genetics* **1986**, *112*, 887–898. [CrossRef]
5. Aoki, K.; Ogata, Y.; Igarashi, K.; Yano, K.; Nagasaki, H.; Kaminuma, E.; Toyoda, A. Functional genomics of tomato in a post-genome-sequencing phase. *Breed. Sci.* **2013**, *63*, 14–20. [CrossRef]
6. Sato, S.; Tabata, S. Tomato Genome Sequence. In *Functional Genomics and Biotechnology in Solanaceae and Cucurbitaceae Crops*; Ezura, H., Ariizumi, T., Garcia-Mas, J., Rose, J., Eds.; Biotechnology in Agriculture and Forestry; Springer: Berlin, Heidelberg, 2016; pp. 1–13, ISBN 978-3-662-48535-4.
7. Xu, X.; Bai, G. Whole-genome resequencing: Changing the paradigms of SNP detection, molecular mapping and gene discovery. *Mol. Breed.* **2015**, *35*, 33. [CrossRef]
8. Chaudhary, J.; Khatri, P.; Singla, P.; Kumawat, S.; Kumari, A.; R, V.; Vikram, A.; Jindal, S.K.; Kardile, H.; Kumar, R.; et al. Advances in Omics Approaches for Abiotic Stress Tolerance in Tomato. *Biology* **2019**, *8*, 90. [CrossRef]
9. Chaudhary, J.; Alisha, A.; Bhatt, V.; Chandanshive, S.; Kumar, N.; Mir, Z.; Kumar, A.; Yadav, S.K.; Shivaraj, S.M.; Sonah, H.; et al. Mutation Breeding in Tomato: Advances, Applicability and Challenges. *Plants* **2019**, *8*, 128. [CrossRef]
10. Adhikari, P.; Adhikari, T.B.; Louws, F.J.; Panthee, D.R. Advances and Challenges in Bacterial Spot Resistance Breeding in Tomato (*Solanum lycopersicum* L.). *Int. J. Mol. Sci.* **2020**, *21*, 1734. [CrossRef]
11. Shirasawa, K.; Kuwata, C.; Watanabe, M.; Fukami, M.; Hirakawa, H.; Isobe, S. Target Amplicon Sequencing for Genotyping Genome-Wide Single Nucleotide Polymorphisms Identified by Whole-Genome Resequencing in Peanut. *Plant Genome* **2016**, *9*, plantgenome2016.06.0052. [CrossRef]
12. The 100 Tomato Genome Sequencing Consortium; Aflitos, S.; Schijlen, E.; de Jong, H.; de Ridder, D.; Smit, S.; Finkers, R.; Wang, J.; Zhang, G.; Li, N.; et al. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.* **2014**, *80*, 136–148. [CrossRef]
13. Arafa, R.A.; Rakha, M.T.; Soliman, N.E.K.; Moussa, O.M.; Kamel, S.M.; Shirasawa, K. Rapid identification of candidate genes for resistance to tomato late blight disease using next-generation sequencing technologies. *PLoS ONE* **2017**, *12*, e0189951. [CrossRef] [PubMed]
14. Kawchuk, L.M.; Hachey, J.; Lynch, D.R.; Kulcsar, F.; van Rooijen, G.; Waterer, D.R.; Robertson, A.; Kokko, E.; Byers, R.; Howard, R.J.; et al. Tomato Ve disease resistance genes encode cell surface-like receptors. *Proc. Natl. Acad. Sci.* **2001**, *98*, 6511–6515. [CrossRef]
15. Fradin, E.F.; Zhang, Z.; Juarez Ayala, J.C.; Castroverde, C.D.M.; Nazar, R.N.; Robb, J.; Liu, C.-M.; Thomma, B.P.H.J. Genetic Dissection of *Verticillium* Wilt Resistance Mediated by Tomato Ve1. *Plant Physiol.* **2009**, *150*, 320–332. [CrossRef] [PubMed]
16. Illumina, I. Quality scores for next-generation sequencing. *Tech. Note: Inform.* **2011**, *31*.
17. Sato, S.; Tabata, S.; Hirakawa, H.; Asamizu, E.; Shirasawa, K.; Isobe, S.; Kaneko, T.; Nakamura, Y.; Shibata, D.; Aoki, K.; et al. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **2012**, *485*, 635–641. [CrossRef]
18. Hamilton, J.P.; Sim, S.-C.; Stoffel, K.; Van Deynze, A.; Buell, C.R.; Francis, D.M. Single Nucleotide Polymorphism Discovery in Cultivated Tomato via Sequencing by Synthesis. *Plant Genome* **2012**, *5*, 17–29. [CrossRef]
19. Sim, S.-C.; Durstewitz, G.; Plieske, J.; Wieseke, R.; Ganal, M.W.; Deynze, A.V.; Hamilton, J.P.; Buell, C.R.; Causse, M.; Wijeratne, S.; et al. Development of a Large SNP Genotyping Array and Generation of High-Density Genetic Maps in Tomato. *PLoS ONE* **2012**, *7*, e40563. [CrossRef]
20. Causse, M.; Desplat, N.; Pascual, L.; Le Paslier, M.-C.; Sauvage, C.; Bauchet, G.; Bérard, A.; Bounon, R.; Tchoumakov, M.; Brunel, D.; et al. Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. *BMC Genom.* **2013**, *14*, 791. [CrossRef]
21. Kim, J.-E.; Oh, S.-K.; Lee, J.-H.; Lee, B.-M.; Jo, S.-H. Genome-Wide SNP Calling Using Next Generation Sequencing Data in Tomato. *Mol. Cells* **2014**, *37*, 36–42. [CrossRef]
22. Gupta, P.; Reddaiah, B.; Salava, H.; Upadhyaya, P.; Tyagi, K.; Sarma, S.; Datta, S.; Malhotra, B.; Thomas, S.; Sunkum, A.; et al. Next-generation sequencing (NGS)-based identification of induced mutations in a doubly mutagenized tomato (*Solanum lycopersicum*) population. *Plant J.* **2017**, *92*, 495–508. [CrossRef] [PubMed]
23. Aguilar, M.; Prieto, P. Telomeres and Subtelomeres Dynamics in the Context of Early Chromosome Interactions During Meiosis and Their Implications in Plant Breeding. *Front. Plant Sci.* **2021**, *12*, 672489. [CrossRef]
24. Edwards, D.; Forster, J.W.; Chagné, D.; Batley, J. What Are SNPs? In *Association Mapping in Plants*; Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., De Silva, H.N., Eds.; Springer: New York, NY, USA, 2007; pp. 41–52, ISBN 978-0-387-36011-9.
25. Yuan, Y.; Bayer, P.E.; Batley, J.; Edwards, D. Current status of structural variation studies in plants. *Plant Biotechnol. J.* **2021**, *19*, 2153–2163. [CrossRef] [PubMed]

26. Francia, E.; Pecchioni, N.; Policriti, A.; Scalabrin, S. CNV and Structural Variation in Plants: Prospects of NGS Approaches. In *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*; Sablok, G., Kumar, S., Ueno, S., Kuo, J., Varotto, C., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 211–232, ISBN 978-3-319-17157-9.
27. Robb, E.J.; Nazar, R.N. Tomato Ve-resistance locus: Resilience in the face of adversity? *Planta* **2021**, *254*, 126. [CrossRef] [PubMed]
28. Nazar, R.N.; Xu, X.; Kurosky, A.; Robb, J. Antagonistic function of the Ve R-genes in tomato. *Plant Mol. Biol.* **2018**, *98*, 67–79. [CrossRef]
29. Cock, P.J.A.; Fields, C.J.; Goto, N.; Heuer, M.L.; Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **2010**, *38*, 1767–1771. [CrossRef]
30. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef]
31. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. 1000 Genome Project Data Processing Subgroup The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef]
32. Wang, K.; Li, M.; Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **2010**, *38*, e164. [CrossRef]
33. Chen, K.; Wallis, J.W.; McLellan, M.D.; Larson, D.E.; Kalicki, J.M.; Pohl, C.S.; McGrath, S.D.; Wendl, M.C.; Zhang, Q.; Locke, D.P.; et al. BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* **2009**, *6*, 677–681. [CrossRef]
34. Abyzov, A.; Urban, A.E.; Snyder, M.; Gerstein, M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* **2011**, *21*, 974–984. [CrossRef] [PubMed]
35. Stothard, P. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and Formatting Protein and DNA Sequences. *BioTechniques* **2000**, *28*, 1102–1104. [CrossRef]
36. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **1988**, *16*, 10881–10890. [CrossRef] [PubMed]
37. Luria, N.; Smith, E.; Reingold, V.; Bekelman, I.; Lapidot, M.; Levin, I.; Elad, N.; Tam, Y.; Sela, N.; Abu-Ras, A.; et al. A New Israeli Tobamovirus Isolate Infects Tomato Plants Harboring Tm-22 Resistance Genes. *PLoS ONE* **2017**, *12*, e0170429. [CrossRef]
38. Kumar, R.; Khurana, A. Functional genomics of tomato: Opportunities and challenges in post-genome NGS era. *J. Biosci.* **2014**, *39*, 917–929. [CrossRef] [PubMed]
39. Zamfir, B.; Hoza, D.; Vînătoru, C.; Lagunovschi, V.; Bratu, C. Research on conservation, evaluation and genetic heritage exploitation of tomato. *Sci. Papers Ser. B Horticulture* **2017**, *LXI*, 307–312.