Article

Propagation Fidelity and Kinship of Tomato Varieties 'UC 82' and 'M82' Revealed by Analysis of Sequence Variation

Vladimir Cambiaso [1] , **Gustavo Rubén Rodríguez** [1] **and David Merrill Francis** [2,*]

1   Instituto de Investigaciones en Ciencias Agrarias de Rosario (IICAR-CONICET-UNR), Cátedra de Genética, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Campo Experimental Villarino, Santa Fe, S2125ZAA Zavalla, Argentina; cambiaso@iicar-conicet.gob.ar (V.C.); grodrig@unr.edu.ar (G.R.R.)

2   Department of Horticulture and Crop Sciences, The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, OH 44691, USA

*   Correspondence: francis.77@osu.edu

Abstract: Plant varieties are named and released based on distinct, unique and stable characteristics but may be maintained separately by genebanks or stock centers under separate accession identification numbers. Genetic heterogeneity of the original variety, genetic drift, failure to exclude cross pollination, and propagation error may erode the integrity of genetic resources. The availability of resequencing and genotyping data for duplicate samples enables an analysis to clarify the relationship between specific varieties or independently curated accessions of the same variety while also assessing the fidelity of germplasm maintenance. We accessed both Single Nucleotide Polymorphism (SNP) array genotypes and resequencing data for two important tomato varieties 'UC 82' and 'M82' that have been maintained as separate accessions in collections as important resources for the research and breeding communities. Our analysis of these data suggests that polymorphism rates from resequencing of cultivated tomato are overestimated in the literature due to heterozygous calls caused by either sequence error or coalignment of repetitive sequences. We defined a set of 32,352 robust SNPs from a file containing data for all samples and we compared the distribution of data with SNPs called from a genotyping array. For both analyses, intravariety variation was found in haplotype blocks, with the same haplotypes identified using SNPs detected from array and sequence data. The distribution pattern of variation across the entire genome sequence was similar for both 'UC 82' and 'M82'. Overall, the differences between distinct accessions of a variety were nearly as great as the differences between 'UC 82' and 'M82'. The similarities between 'UC82' and 'M82' range from 99.33% to 99.74% and are highly consistent with a common pedigree and shared selection from partially inbred progeny. The data also suggest that these tomato genetic resources have been propagated with high fidelity.

Keywords: *Solanum lycopersicum*; high-density genotyping; whole genome resequencing; genetic resources; haplotypes

1. Introduction

High-density genotyping and genome sequence data are increasingly available for large numbers of crops and their wild relatives. The availability of genotyping and resequencing data for duplicate accessions of the same variety in public data bases permits questions to be addressed about the fidelity of propagation and also the genetic relationship, or kinship, between varieties. Genetic heterogeneity within individual accessions of a named variety has been investigated for major crops like soybean [1], maize [2], and the model species *Arabidopsis thaliana* [3]. At the time the reference sequence for the tomato variety 'Heinz 1706' was released [4] a large Single Nucleotide Polymorphism (SNP) genotyping

array was available as a commercial resource [5–7]. Subsequently resequencing data from several large [8–11] and small scale projects [12,13] became available as did a one thousand variety data set based on the SNP array [14]. These genetic and genomic data were used to provide insight into the origin, domestication and breeding history of the tomato [6,8,9,14,15]. There is a growing set of information for duplicate samples of the same variety, which has been maintained by independent groups as distinct accessions. These data provide both experimental controls and provide a resource to ask specific questions about genetic relationships within tomato germplasm collections. We sought to use sequence and SNP data to examine the relationship or kinship between tomato varieties and assess the fidelity of maintenance or propagation within accessions of the same varieties.

In tomato, the *Solanum lycopersicum* L. variety 'M82' (also referred to as 'M82-1-8' and 'M82-1-8 (VF)') has become an important resource for genetic studies. 'M82' is used as a parent for stable introgression line populations [16,17], which serve as a resource for the discovery of quantitative trait loci (QTL) [18–21], map-based cloning [22,23], and genomic studies seeking to link form and function [24,25]. The relationship between 'M82' and the variety 'UC 82' has been the subject of debate. 'UC 82' was released based on yield and fruit setting ability as well as properties favorable to bulk handling at the beginning of the mechanical harvest era but prior to the wide spread use of hybrid tomatoes in the canning industry [26]. Its pedigree is well described in the variety release notification and elsewhere [26,27]. Published studies describe 'M82' as a selection from 'UC 82' [28,29], despite the latter being described as a stable variety [27]. In an analysis of population structure based on SNP data, 'M82' and 'UC 82' clustered together with varieties having pedigrees tracing to California breeding programs [6]. The lack of a published pedigree for 'M82' contributes to a vague understanding of its origin.

We accessed independent data sets based on Illumina genotyping and next generation sequencing for 'UC 82' and 'M82' accessions in order to examine the relationship between the two varieties and variability within accessions of the same variety. Genome-wide analysis demonstrated high propagation fidelity among individual accessions of the same variety and a close relationship between these important and widely used tomato varieties.

2. Materials and Methods

2.1. Plant Material

SNP genotyping and whole genome sequence data from different sources of 'UC 82' and 'M82' were obtained (Table S1). 'M82' accession LA3475 was provided from the Charles Rick Tomato Genetics Resource Center (TGRC), UC Davis, and maintained at The Ohio State University (OSU). Other sources of 'M82' were maintained in Israel (e.g., ERS484046) or were unlisted (e.g., TS-228, see below). Stocks from 'UC 82' accession LA1706 were also obtained from the TGRC and maintained separately at the Genebank of the Institute for the Conservation and Improvement of Agricultural Biodiversity (COMAV) and OSU.

2.2. SNP Genotyping Data from the Illumina "Solanaceae Coordinated Agricultural Project" (SolCAP) Array

Genotyping was performed using the commercially available Illumina SNP array (Illumina, Inc., San Diego, CA USA) for tomato developed as part of the "Solanaceae Coordinated Agricultural Project" or SolCAP, as described in previous studies [14]. Data for 7720 SNPs are available from Table S2 from the original manuscript [14]. Samples with genotype data from SNP array were UC 82-COMAV, UC 82-OSU, and M82-OSU (Table S1). SNP data from 'Heinz 1706', the reference genome variety for tomato, were also used to provide a parallel reference for the resequencing data set. The close wild relative *Solanum pimpinellifolium*. L. LA1589 provided an outgroup for analysis.

SNPs for 'M82' and 'UC 82' accessions were arranged according to physical position in the genome and clustered using the R pvclust 2.0-0 package [30]. The SNP data were used to estimate the percentage of residual heterozygosity for each accession and to calculate the proportion of shared and

distinct SNPs in pairwise comparisons between the accessions. The physical size of genomic regions between contiguous SNPs relative to reference assembly version SL3.0 [31] was used to determine the divergent proportion relative to the total genome sequence among accessions.

2.3. Whole Genome Resequencing Data

Sequence Read Archive (SRA) files for the *Solanum lycopersicum* varieties 'M82' and 'UC 82' were downloaded from the National Center for Biotechnology Information (NCBI) SRA database and from the European Nucleotide Archive. Samples TS-3 (BioSample: SAMN03004944; SRA: SRS698748), TS-228 (BioSample: SAMN03005071; SRA: SRS698858) and ERS484046 (Project number: PRJEB6302; Sample accession: SAMEA2610542) correspond to 'M82'. Based on database passport information, DNA for TS-3 was derived from an accession of 'M82' from Israel while the source for accession TS-228 is unlisted. DNA for ERS484046 was derived from LA3475 sourced from the TGRC. Sample TS-171 (BioSample: SAMN03005032; SRA: SRS698829) corresponds to 'UC 82' sourced in the United States. We refer to these samples as M82-TS3, M82-TS228, M82-ERS484046, and UC82-TS171 (Table S1).

2.4. Sequence Alignment and SNP Calling

SRA were converted into FastQ format using the fastq-dump tool from the SRA toolkit [32]. Quality control of the FastQ files was performed using FastQC version 0.11.4 [33]. The four samples analyzed were derived from different sequencing chemistry and resulted in different read lengths and quality. M82-TS3 reads averaged 50 base pairs (bp) with data Illumina 1.5 encoded and Phred quality scores on a scale of 0–62, while M82-ERS484046, M82-TS228, and UC82-TS171 reads data were Sanger/Illumina 1.9 encoded and averaged 100 bp with Phred quality scores on a scale of 0–93. The wrapper tool Trim Galore! version 0.3.7 [34] was used to remove adapters and trim reads. Trimming was performed according to samples, with paired reads shorter than 60 bp removed from sample M82-ERS484046, M82-TS228, and UC82-TS171 and shorter than 20 bp removed from sample M82-TS3. Quality of reads was filtered using Phred+64 for the M82-TS3 sample and Phred+33 for M82-ERS484046, M82-TS228, and UC82-TS171 to impose a uniform quality filter at a Phred score of 20 corresponding to 99% accuracy.

The trimmed FastQ files were aligned against the *S. lycopersicum* L. cv. 'Heinz 1706' reference assembly version SL3.0 [31] using Bowtie 2 version 2.3.2 [35]. During the alignment, the "–very-sensitive-local" option was used. This option sets a combination of parameters aimed to cover the speed, sensitivity, and accuracy of the process. The ending "-sensitive" implies a slower progression but more sensitive and accurate. In Bowtie 2 the "-local" mode of alignment was used to trim some characters from one or both ends when doing so maximized the alignment score. The option "–phred64-quals" was used to align M82-TS3, while the option "–phred33-quals" was used for the other three samples, consistent with appropriate options based on the sequencing chemistries and Illumina encoding.

The aligned files in the Sequence Alignment Map (SAM) format were sorted, labeled, and converted to a compressed binary version (BAM files) using Picard software version 2.18.14 [36]. Duplicate records were located using Picard. The resulting BAM files were analyzed with Qualimap version 2.2.1 [37]. Stats and graphical representations of the obtained coverage across the entire genome sequence were extracted.

Three different approaches were followed to perform whole genome sequence comparison among the four sequence samples. For both approaches, the HaplotypeCaller tool from GATK version 4.0.9.0 [38,39] was used to obtain Genomic Variant Calling Format (gVCF) files for each sample. These files had records for all sites whether there is a variant call present or not. The CombineGVCFs tool from GATK version 4.0.9.0 [38,39] was used to merge gVCFs into a single gVCF. The function GenotypeGVCFs in GATK version 4.0.9.0 [38,39] was used to perform joint genotyping for all samples included in each gVCF, resulting in Variant Call Format (VCF) files. As a first approach we compared the four samples in pairwise combinations, so six different VCF files were obtained. The second

approach resulted in a single VCF containing all four samples. Finally, we also compared each sample separately against the reference and therefore a single VCF for each sample was generated. SNPs were extracted from VCF files with the SelectVariants tool from GATK version 4.0.9.0 [38,39] using the "–select-type-to-include" option. Default parameters for filtering SNPs were set to use the VariantFiltration tool from GATK version 4.0.9.0 [38,39]. The "–remove-filtered-all" option from the VCFtools version 0.1.15 [40] was used to remove all sites previously filtered with the VariantFiltration tool. The Linux "grep" command with the "-v" option was used to filter out lines in the VCF files containing missing data (marked as "/") for at least one sample.

Since regions with extremely low depth of coverage are unreliable and regions with extremely high depth of coverage are correlated with highly repetitive sequences and the alignment of those reads may not be accurate, descriptive statistics from the read depth of coverage at each SNP were calculated to determine the best cut off criteria. We excluded regions with less than 4× coverage as lower coverage does not allow correction for sequence error. We also filtered based on a maximum read depth of coverage for each sample to exclude repetitive sequences, which may be mapped on top of the same reference leading to an overestimation of polymorphism usually as heterozygosity. Maximum depth of coverage was set as $d+4\sqrt{d}$ where d was the mean depth of coverage [41]. The SnpSift filter command from SnpSift version 4.0 [42] was used to filter the VCF files by depth of coverage.

2.5. SNP Density Comparison from Resequencing Data

Variation and genotype calling was performed for each sample to calculate the SNP density relative to the reference genome sequence. Single VCFs were used to estimate the percentage of residual heterozygosity for each sample. VCF files obtained from both approaches were used to calculate the proportion of shared and distinct SNPs by chromosome in each pairwise comparison between the four samples. The VCF containing data for all four samples permitted a calculation of the SNPs shared by the four samples relative to the reference genome sequence. Genomic regions between contiguous SNPs were used to determine the divergent proportion from the total genome sequence among samples. Whole genome patterns of SNP distributions were visualized using "CircosVCF" [43].

3. Results

3.1. SNP Genotyping Data Analysis

Data derived from the SolCAP Illumina array were analyzed for 7720 SNPs for two 'UC 82' accessions, an 'M82' accession, and 'Heinz 1706'. The percentage of missing data ranged from 0.76% to 1.27% and percentage of heterozygous SNPs ranged from 0.08% to 0.19% respectively. The low residual heterozygosity detected for all samples was expected for highly inbred lines. Heterozygous SNPs from the original selection are likely to be fixed and have an equal probability to be fixed as polymorphic or reference during each generation of self-pollination. After removal of missing data and heterozygous SNPs, a total of 7578 SNPs were retained for further analysis. SNPs were clustered in seven genomic regions distributed on chromosomes 1, 6, 7, 9, 10, and 11 (Table 1). Differences within and between variety accessions were described based on the physical length of the haplotypes. As a proportion of the genome sequence flanked by polymorphic SNPs, the UC 82-COMAV and UC 82-OSU differ by 0.18% (Table 1), with 92% of polymorphic SNPs clustered on the top and the bottom of Chromosome 11. The comparison between 'UC 82' accessions and 'M82' reveals differences that range between 0.14% and 0.19%, and 0.26% when considering both samples together (Table 1).

Table 1. Genomic regions carrying polymorphic Single Nucleotide Polymorphisms (SNPs) between 'UC 82' and 'M82' accessions based on the SolCAP Illumina array data.

	Polymorphic Regions[z]			Polymorphic SNPs between			
Chromosome	Start Position	End Position	Region Size (bp)	UC 82 Accessions	UC 82-COMAV and M82-OSU	UC 82-OSU and M82-OSU	UC 82 and M82[y]
1	80,898,673	81,065,071	166,398	3	-	3	3
6	37,721	37,722	1[x]	1	1	-	1
7	60,221,252	60,222,929	1677	2	2	-	2
9	71,304,279	71,953,325	649,046	-	8	8	8
10	59,730,745	59,731,037	292	-	2	2	2
11	7,180,897	7,728,808	547,911	25	25	-	25
	52,038,184	52,836,721	798,537	48	-	48	48
Total			2,163,862	79	38	61	89
Percentage different relative to the total genome sequence (%)[w]				0.18	0.14	0.19	0.26

[z] Genomic region flanked by contiguous polymorphic SNPs. [y] Considering all polymorphic regions present in at least one pairwise comparison. [x] A single polymorphism was detected in Chromosome 6. [w] Calculated as the proportion from the reference genome sequence version SL3.0 (total size 828,349,174 bp) covered by polymorphic regions.

When pairwise comparisons were performed, fewer polymorphic SNPs were identified between 'UC 82' and 'M82' accessions (38 and 61 SNPs out of 7578) relative to those detected within accessions of 'UC 82' (79 SNPs out of 7578; Table 1). In all cases over 66% of the polymorphic SNPs were clustered on Chromosome 11. Cluster analysis across all polymorphic SNPs, placed UC 82-COMAV closer to the M82-OSU sample at alpha = 0.98. This clustering was not consistent across chromosomes, however. For SNPs on chromosomes 6 and 7 the UC 82-OSU and M82-OSU samples clustered together. For chromosomes 9 and 10, 'UC 82' samples clustered together. On Chromosome 11, markers between 7.18 and 7.72 Mb cluster UC 82-OSU and M82-OSU samples, and between 52.03 and 52.83 Mb, 48 SNPs cluster the UC 82-COMAV with M82-OSU. In a separate analysis, SNPs from the SolCAP array were used to repeat the comparisons using 'Heinz 1706' and LA1589 as outgroups. The results again showed that the UC 82-COMAV and M82-OSU accessions were more closely clustered than the two 'UC 82' accessions (Figure S1).

3.2. Sequence Alignment and SNP Calling

Sequence data from the four SRA files covered 55.30−89.08 percent of the 828,349,174 bp SL3.0 reference genome sequence to a depth of coverage of at least 4× (Table 2). Since different percentages of the genome sequence were covered, the total number of detected variants differed among samples (Table 2).

Table 2. Whole genome sequence alignment and variant calling relative to tomato reference genome sequence assembly version SL3.0 for samples M82-ERS484046, M82-TS3, M82-TS228, and UC82-TS171.

Sample	Number of Mapped Reads	Genome Coverage (%)[z]	Mean Depth of Coverage (×)	Raw		Filtered[y]		Hetetozygous SNPs[x]	
				Indels	SNPs	Indels	SNPs	Number	%
M82-ERS484046	511,813,113	89.1	50.1	169,490	1,257,174	162,203	385,236	14,082	3.8
M82-TS3	62,544,931	55.3	4.6	33,423	627,865	33,396	261,324	4593	5.1
M82-TS228	142,963,559	70.4	16.5	97,189	1,165,307	91,969	299,680	24,266	13.9
UC82-TS171	68,783,839	61.4	8.0	93,344	1,074,041	90,628	362,639	30,228	11.0

[z] Percentage of the tomato reference genome sequence (assembly version SL3.0) covered by at least 4× depth of short reads. [y] Number of detected variants after filtration. [x] Number and percentage of heterozygous SNPs identified from filtered SNPs.

Figure 1 presents a schematic representation of genomic regions with SNP calls. We identified 32,352 robust SNPs from a total of 1,886,601 SNPs detected in a single four-way VCF containing data for all samples. Only 414,820 SNPs passed the default filtering parameters from GATK. When SNPs with missing data in one sample were removed we retained 212,068 informative SNPs for all samples.

Filtering by depth of coverage reduced the total number of SNPs to 46,981 SNPs. Of the remaining SNPs 496 were unmapped (located on Chromosome 0) and 14,133 were heterozygous. The percentage of heterozygous SNPs detected when comparing each sample separately to the reference, ranged from 3.8% to 13.9% (Table 2) and could reflect true residual heterozygosity, sequence error, or duplicate genes that were clustered together in the alignment (Figure 1A,C). Distinguishing residual heterozygosity from sequencing, alignment or SNP calling errors is difficult, even when following a conservative assembly pipeline. Since the mating system of tomato favored elimination of heterozygous SNPs, we therefore defined the set of robust SNPs based on homozygous regions identified after hard-filtering was applied (Figure 1B).
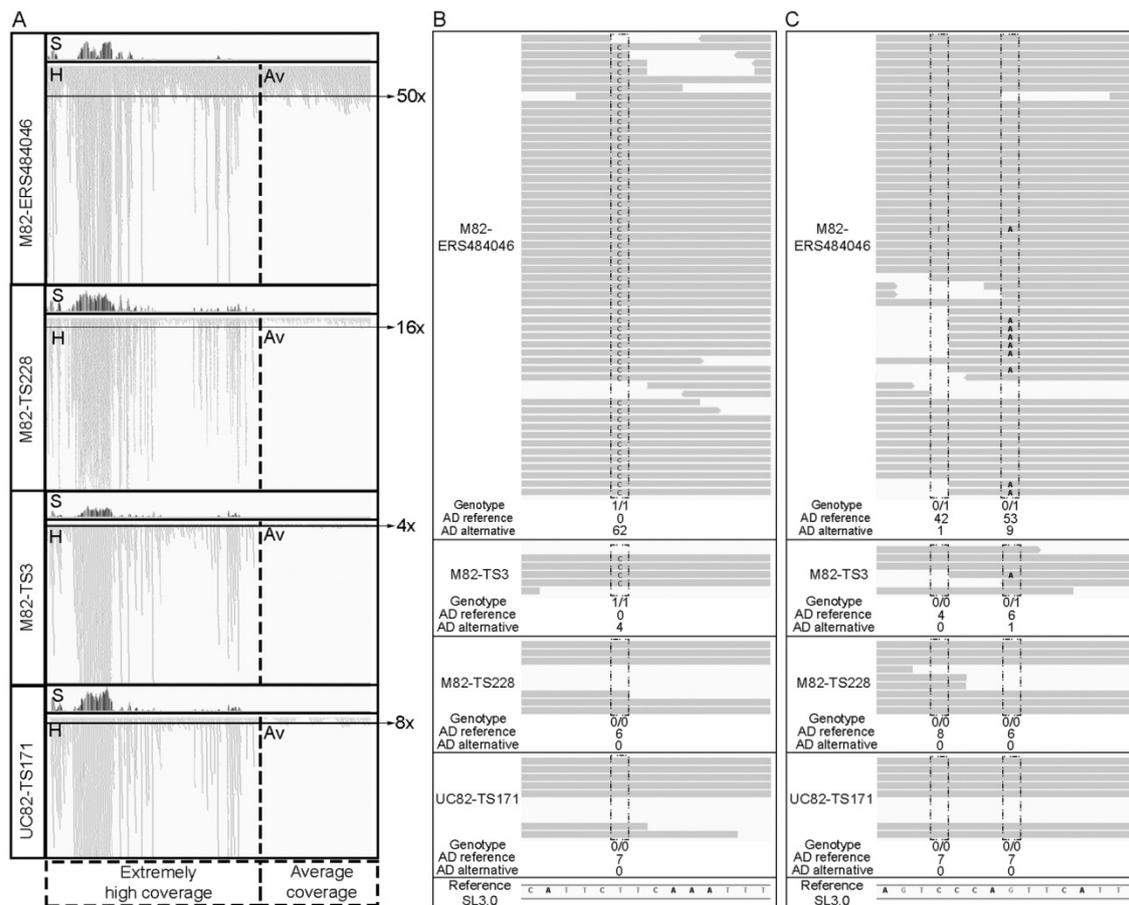
Figure 1. Schematic representation of genomic regions with SNP calls. (**A**) Graphical representation of number of detected SNPs (S) in regions with extremely high (H) and average (Av) coverage. (**B**) Genomic region with a robust homozygous SNP call. (**C**) Genomic region with heterozygous SNP calls after hard-filtering. Solid gray bars represent reads supporting bases homozygous to the reference. Detected SNPs are represented with letters A, T, and C. Genotype assigned: 0/0 as homozygous to the reference and 0/1 heterozygous. The allelic depths (AD) represent the numbers of reads supporting each called base.

3.3. SNP Density Comparison from Sequence Data

The genome sequences of M82-ERS484046, M82-TS3, M82-TS228, and UC82-TS171 were highly similar. The same conclusions were drawn whether we considered pairwise VCF comparisons or the four-way VCF comparison. The advantage of joining samples in a single four-way VCF was the ease of performing analysis of the samples based on exactly the same positions, and we focused on results from this comparison. Using the set of robust SNPs, differences within 'M82' accessions ranged from 50 to 165 SNPs or only 0.15%–0.51% of the genome sequence (Table S2). The range of differences across all

12 chromosomes were between 0% and 1.27% for the M82-ERS484046 vs. M82-TS228 comparison, 0.02% and 12.30% for M82-ERS484046 vs. M82-TS3 and 0.08% and 13.27% for M82-TS3 vs. M82-TS228 (Table S2). The greatest variation within 'M82' pairwise comparisons was localized on chromosomes 7 and 9 (Table S2). Comparisons of UC82-TS171 with the three 'M82' accessions displayed a range from 229 to 277 SNPs or 0.71%–0.86% of the genome sequence (Table S2). The range of polymorphic SNPs across all 12 chromosomes was of 0.02%–21.41%, 0.12%–22.32%, and 0.03%–21.10% when comparing UC82-TS171 with M82-ERS484046, M82-TS228 and M82-TS3, respectively (Table S3). For all pairwise comparisons the greatest differences in percentage of the homozygous SNPs were detected on Chromosome 1 and the largest number of polymorphic SNPs was identified on Chromosome 11 (Table S3).

3.4. SNP Allele Comparison across the Entire Genome

The distribution pattern of SNPs across the entire genome sequence for all samples was highly similar (Figure 2). A clear enrichment of SNPs was found on chromosomes 4, 5, 11, and 12, and to a lesser degree on chromosomes 2, 6, 7, and 10. Only a few specific regions on chromosomes 1, 3, 8, and 9 contained variation relative to the reference (Figure 2).
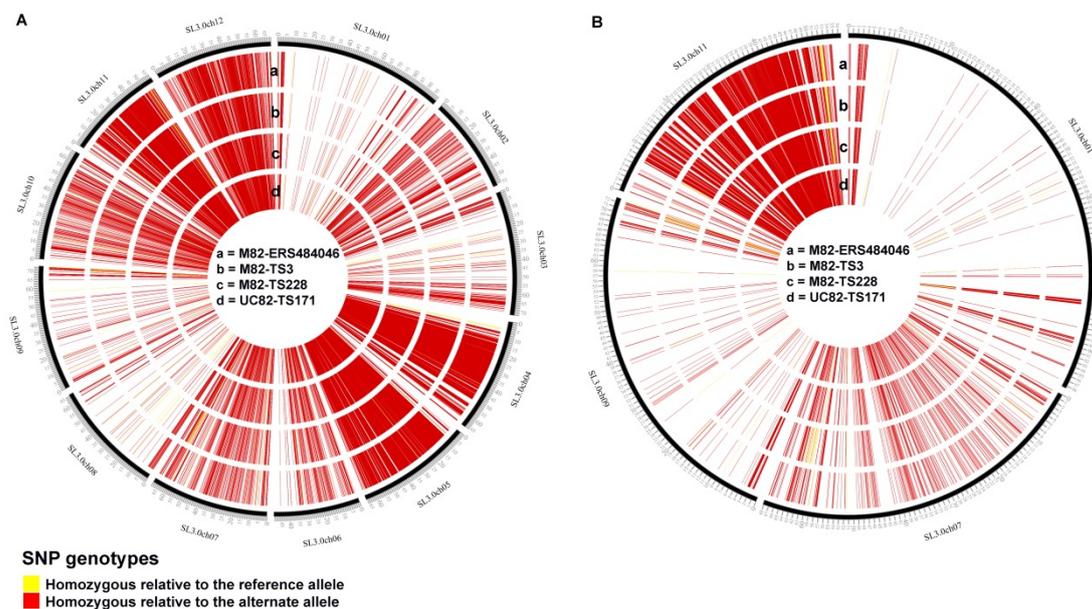

SNP genotypes
- Homozygous relative to the reference allele
- Homozygous relative to the alternate allele

Figure 2. (**A**) SNP allele comparison across the entire genome sequence among samples M82-ERS484046, M82-TS3, M82-TS228, and UC82-TS171. (**B**) SNP allele comparison magnification of most diverse chromosomes.

3.5. Comparison of Polymorphic Genomic Regions

Figure 2B shows a close-up of SNP allele comparisons on chromosomes with maximum differences within 'M82' and between 'M82' and 'UC 82' samples (Tables S2 and S3). Regions on the bottom of chromosome 7 and 9 clearly differentiated M82-TS3 from the other 'M82' samples, while UC82-TS171 shared SNPs with M82-ERS484046 and M82-TS3 on Chromosome 7 and with M82-TS228 on Chromosome 9. Two regions located on the bottom of Chromosome 1 and 11 differentiate UC82-TS171 from all 'M82' samples (Figure 2B).

To determine the proportion of the total genome sequence for which within accession variation was detected, genomic regions with at least three consecutive SNPs were summarized in Table 3. Intravariety variation for 'M82' accessions covered from 0.0001% to 0.43% of the total genome sequence when pairwise comparisons were made (Table 3). These results suggest that samples M82-ERS484046 and M82-TS228 were highly similar. The overall proportion of the tomato genome sequence that varied within 'M82' accessions was 0.43% with the most polymorphic region located on bottom of

Chromosome 7 (SL3.0ch07:60,125,464-62435965 bp) and 9 (SL3.0ch09:71180818-72418523 bp; Table 3). Divergence between 'UC 82' and 'M82' accessions detected by pairwise comparisons cover from 0.33% to 0.59% of the total genome sequence (Table 3). Differences between all 'M82' samples and UC82-TS171 were mostly located on the bottom of Chromosome 1 (SLch01:80,660,651-81,107,229 bp) and 11 (SLch11:52,035,195-52,843,385 bp). When we considered all the differential regions detected between 'UC 82' and 'M82', they covered 0.67% of the total genome sequence. The total number of SNPs detected between 'M82' samples and UC82-TS171 was higher than those detected within 'M82'. For two pairwise comparisons between 'UC 82' and 'M82' (M82-ERS484046 vs. UC82-TS171 and M82-TS228 vs. UC82-TS171), the proportion of the total genome sequence covered was smaller than 'M82' comparisons with M82-TS3 (M82-ERS484046 vs. M82-TS3 and M82-TS228 vs. M82-TS3; Table 3).

3.6. Haplotype Identification Based on SNP Genotyping and Whole Genome Resequence Data

Analyses based on independent genotyping or resequencing data identify the same genomic regions that differentiate 'UC 82' and 'M82' accessions. To compare and visualize these regions, the genotypic data from variety 'Heinz1706' was extracted from the Illumina SolCAP array and used to determine reference and alternate alleles for samples UC 82-COMAV, UC 82-OSU, and M82-OSU. The allelic information relative to the reference was then used to guide a schematic representation of the chromosome regions for the accessions analyzed based on both genotypic and resequencing data (Figure 3). From this visualization, three different haplotypes could be distinguished among all 'UC 82' and 'M82' accessions. The first haplotype was represented by samples UC 82-OSU and UC 82-TS171; the second by UC 82-COMAV and M82-TS3; and the third by M82-ERS484046, M82-TS228 and M82-OSU.
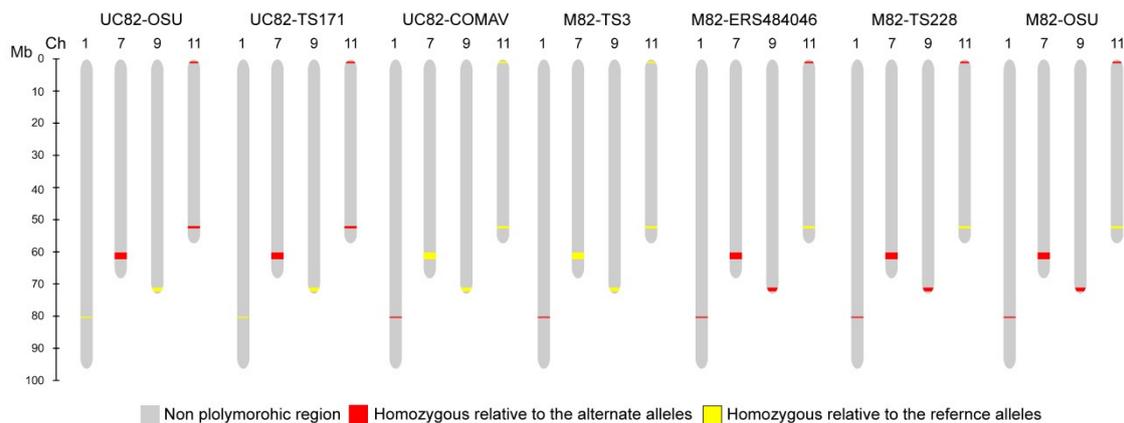

Figure 3. Schematic representation of the differential chromosome regions between 'UC 82' and 'M82' accessions.

Table 3. Genomic regions carrying polymorphic SNPs within 'M82' and between 'UC 82' and 'M82' samples based on the whole genome sequence alignment and SNP calling analysis.

	Differential Regions[z]			Polymorphic SNPs within M82 Samples			Polymorphic SNPs between M82 and UC 82 Samples			
Ch[y]	Start Position	End Position	Region Size (bp)	M82-ERS484046 vs. M82-TS228	M82-ERS484046 vs. M82-TS3	M82-TS228 vs. M82-TS3	M82-ERS484046 vs. UC82-TS171	M82-TS228 vs. UC82-TS171	M82-TS3 vs. UC82-TS171	UC 82 vs. M82[x]
1	80,660,651	81,107,229	446,578	-	-	-	71	66	66	71
3	23,871,750	23,871,769	19	5	-	5	-	5	-	5
4	4,575,994	5,372,152	796,158	-	-	-	5	5	5	5
5	27,883,416	27,884,031	615	3	-	3	-	3	-	3
	45,459,503	45,459,508	5	3	-	3	-	3	-	3
7	60,125,464	62,435,965	2,310,501	-	29	29	-	-	28	29
9	71,180,818	71,538,753	357,935	-	14	14	14	14	-	14
	72,124,815	72,418,523	293,708	-	19	19	19	19		19
10	42,637,890	42,637,909	19	4	-	4	-	4	-	4
	59,465,991	59,518,850	52,859	-	5	5	5	5	-	5
11	7,181,778	7,700,078	518,300	-	16	16	-	-	16	16
	52,035,195	52,843,385	808,190	-	-	-	78	78	78	78
Total			5,584,887	15	83	98	192	202	193	202
Percentage different relative to the total genome sequence (%)[w]				0.0001	0.43	0.43	0.33	0.33	0.59	0.67

[z] Genomic region covered by at least three contiguous polymorphic SNPs. [y] Chromosome. [x] Considering all differential regions present in at least one pairwise comparison. [w] Calculated as the proportion from the reference genome sequence version SL3.0 (total size 828,349,174 bp) covered by differential regions.

4. Discussion

In order to compare variation between varieties and within accessions of the same variety we defined 32,352 robust SNPs in tomato. An important step in defining these SNPs was the elimination of heterozygous calls. The problem of accurate identification of heterozygosity is complicated by both a sequencing error and inaccurate read mapping in next-generation sequencing technology and has been widely discussed [44,45]. Typically, these problems are minimized by invoking strict coverage parameters for depth of coverage, uniformity of calling, and quality scores. Our analysis followed GATK [39,46] best practices. The HaplotypeCaller uses local de-novo assembly to reassemble the reads in variable regions, a strategy that provides more accuracy. A graphical inspection of the depth of coverage from all genotypes revealed regions in common with extremely high concentration of reads (Figure 1A), probably caused by repetitive DNA or gene-families, which has been artificially condensed in the reference. The regions with higher depth were on Chromosome 1 (between positions SL3.0ch01:23,806,000 and 23,812,500), Chromosome 2 (between positions SL3.0ch02:14,175,000 and 14,190,000), and Chromosome 11 (between positions SL3.0ch11:20,552,400 and 20,553,600). These were also regions with a high frequency of heterozygous SNP calls. Based on results from the analysis of the SolCAP Illumina array, we expected fewer heterozygous SNPs than detected from resequencing data, suggesting many of these are false positives. To assess the importance of this issue we altered filtering conditions, only for samples with similar mean depth of coverage (M82-TS3, M82-TS228, and UC82-TS171; Table 2), comparing both a maximum depth of 15× and 30×, with the rationale that a higher coverage would permit more error due to duplicate regions of the genome sequence. The two filters differed in total SNP calls with 51,535 at 30× and 38,826 at 15×, suggesting that error was introduced due to alignment to repetitive sequences. We therefore applied a cut off criteria for maximum read depth coverage to filter false positive SNP calling due to duplicate sequences clustered together in the alignment. The filter criteria were determined for each sample as d+4√d where d is the mean depth of coverage [41]. After this filtering was applied, the remaining heterozygous SNPs should be more reflective of true residual heterozygosity, but a visual inspection of these SNPs revealed heterozygous calls (0/1) when only one SNP was detected out of 30 reads (Figure 1C). Given the homozygosity of tomato, such calls were likely errors. Such errors appeared to occur independently of depth of coverage, and we assumed these errors might be caused by the Bayes' likelihood method used to estimate genotypes.

The genome sequences of 'UC 82' and 'M82' accessions were highly similar. Both, the SNP density and the SNP allele comparison across the entire genome sequence for all samples presented highly similar distribution patterns. A distribution of chromosomal variant density between 'M82' and the 'Heinz 1706' reference genome sequence was previously published [10] and is similar to the SNPs distribution presented in Figure 2, for both 'M82' and 'UC 82' samples. Furthermore, the high density of homozygous SNPs relative to the alternate allele detected on chromosomes 4, 5, and 11 for the four samples (Figure 2) were consistent with putative *S. pimpinellifolium* introgressions into 'M82' previously detected [10]. Variation within 'M82' accessions ranged from 0.0001% to 0.43% of the total genome sequence and between 'UC 82' and 'M82' accessions the range was 0.33%–0.59% when pairwise comparisons are made (Table 3). Considering all of the differential regions detected between 'UC 82' and 'M82' accessions, the divergence covered 0.67% of the total genome sequence (Table 3).

The same genomic regions differentiate 'UC 82' and 'M82' accessions whether considering genotyping array or resequencing data. From the visualization of these regions, three different haplotypes could be distinguished among all 'UC 82' and 'M82' accessions (Figure 3). A first haplotype consisting of 'UC 82' accessions, a second of 'M82' accessions and a third haplotype was shared by one sample from each variety.

The genome-wide intra-variety variation of 0.18% and 0.43% we detected within 'UC 82' and 'M82' accessions, respectively, suggests a high fidelity of propagation for these varieties within the tomato genetics community. These percentages compare favorably to the genome-wide differences of 2.3% for the maize variety 'B73' and 3.1% for the soybean variety 'Williams 82', two references for

agricultural genetics, genomics and breeding [1,2]. Differences for the model plant standard *Arabidopsis* 'Col-0' were considerably higher at 20% [3]. Differences that we detected within the same accessions were consistent with drift as regions of residual heterozygosity were resolved and no evidence for outcrossing or introgression were detected.

For specific comparisons and regions of the genome sequences, the within accession variation was greater than between 'UC 82' and 'M82'. A shared history or kinship for these varieties provides an explanation for these patterns of variation. The historical record is unclear as to whether 'M82' and 'UC 82' are independent selections made out of the same four-way cross or independent selections from the same partially inbred line from that same cross. The 'M82' release notice states, "initial cultivation of the variety 'M82-1-8 (VF)' was at the University of California, Davis, by Dr. M.A. Stevens" and, further, "selections and improvement in uniformity were made by Y. Rudich, Faculty of Agriculture, Rehovot" (translated from the Hebrew) [47]. The analysis presented here suggests that the two varieties were selected from the same partially inbred line, rather than from independent F_2 plants. Based on the published pedigree for 'UC 82' [26,27] there is a disproportionate contribution from VF 65-2, suggesting that independent F_2 plants will share more than 50% of their genomes by descent. Translating pedigree into SNP percentages is difficult as all parents were not represented in sequence or genotyping databases and we could not assume they were 100% different. However, observations across genotyping data show a narrow range of 0.14%–0.19% differences in pairwise comparisons. Considering all differential regions present in at least one comparison, the accessions differed at 0.26% of the genome sequence (they share 99.74%). Sequence analysis showed similar proportions, a range of 0.33%–0.59% differences in pairwise comparisons and a 0.67% (shared 99.33%) when considering all differential regions present in at least one pairwise comparison. These ranges were highly consistent with a common pedigree and shared selection between 'UC 82' and 'M82'.

To place the analysis presented here into context based on findings in other plant species, it is useful to consider results from maize, soybean, and *Arabidopsis*. In the introduction of their paper, describing sequence variation between accessions of the *Zea mays* inbred 'B73', Liang and Schnable [2] review the implications of genetic variation that may have occurred in genetic stocks as a result of contamination, accidental out-crossing, residual heterozygosity, or mutation. They note that resequencing of single plants from a source of the *Arabidopsis thaliana* accession 'Columbia-0' identified variation that accounted for nearly 20% of the total reference genome sequence [3]. They also note that two sources of the *Glycine max* variety 'Williams 82', the reference genome accession for soybean, contains 3.1% polymorphism [1]. The differences in 'Williams 82' accessions originated from the differential segregation of polymorphic chromosomal regions following introgression, backcrossing, and single-seed descent during the breeding process. In maize, 13,360 high confidence SNPs in 27 RNA-seq samples for 'B73' were described [2]. This number was 4.8-fold lower than the number expected from a non-B73 accession. Their analysis defined robust SNPs and found that 97.7% of the maize gene space was consistent with a single haplotype across all 27 'B73' samples. Over half of the SNPs locate within blocks on specific regions in different chromosomes, and were likely introgressions. The remaining SNPs were distributed over all ten chromosomes and were either new mutations, or reflect segregation and fixation of heterozygous loci from the original 'B73' accession [2]. The three genotypes, 'Williams 82', 'B73', and 'Columbia-0' were among the most widely studied by their respective plant research communities, thus understanding the distribution and causes of variation between independent accessions had important implications to the respective research communities.

5. Conclusions

A high fidelity of propagation for two tomato varieties was demonstrated based on two independent data sets. The close relationship between 'UC 82' and 'M82' accessions was confirmed, and the similarities were highly consistent with a common pedigree and shared selection. We also demonstrated how sequencing technologies and bioinformatics analysis could be successfully used by plant breeders to differentiate haplotypes between and within varieties.

References

1. Haun, W.J.; Hyten, D.L.; Xu, W.W.; Gerhardt, D.J.; Albert, T.J.; Richmond, T.; Jeddeloh, J.A.; Jia, G.; Springer, N.M.; Vance, C.P.; et al. The Composition and Origins of Genomic Variation among Individuals of the Soybean Reference Cultivar Williams 82. *Plant Physiol.* **2011**, *155*, 645–655. [CrossRef] [PubMed]

2. Liang, Z.; Schnable, J.C. RNA-seq based analysis of population structure within the maize inbred B73. *PLoS ONE* **2016**, *11*, e0157942. [CrossRef] [PubMed]

3. Shao, M.-R.; Shedge, V.; Kundariya, H.; Lehle, F.R.; Mackenzie, S.A. Ws-2 introgression in a proportion of *Arabidopsis thaliana* Col-0 stock seed produces specific phenotypes and highlights the importance of routine genetic verification. *Plant Cell* **2016**, *28*, 603–605. [CrossRef] [PubMed]

4. The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **2012**, *485*, 635–641. [CrossRef] [PubMed]

5. Sim, S.-C.; Durstewitz, G.; Plieske, J.; Wieseke, R.; Ganal, M.W.; Van Deynze, A.; 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]

6. Sim, S.-C.; Van Deynze, A.; Stoffel, K.; Douches, D.S.; Zarka, D.; Ganal, M.W.; Chetelat, R.T.; Hutton, S.F.; Scott, J.W.; Gardner, R.G.; et al. High-density SNP genotyping of tomato (*Solanum lycopersicum* L.) reveals patterns of genetic variation due to breeding. *PLoS ONE* **2012**, *7*, e45520. [CrossRef]

7. 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 J.* **2012**, *5*, 17–29. [CrossRef]

8. Lin, T.; Zhu, G.; Zhang, J.; Xu, X.; Yu, Q.; Zheng, Z.; Zhang, Z.; Lun, Y.; Li, S.; Wang, X.; et al. Genomic analyses provide insights into the history of tomato breeding. *Nat. Genet.* **2014**, *46*, 1220–1226. [CrossRef]

9. Aflitos, S.; Schijlen, E.; de Jong, H.; de Ridder, D.; Smit, S.; Finkers, R.; Wang, J.; Zhang, G.; Li, N.; Mao, L.; et al. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.* **2014**, *80*, 136–148.

10. Bolger, A.; Scossa, F.; Bolger, M.E.; Lanz, C.; Maumus, F.; Tohge, T.; Quesneville, H.; Alseekh, S.; Sørensen, I.; Lichtenstein, G.; et al. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **2014**, *46*, 1034–1038. [CrossRef]

11. Gao, L.; Gonda, I.; Sun, H.; Bao, K.; Tieman, D.M.; Fish, T.L.; Stromberg, K.A.; Sacks, G.L.; Theodore, W.; Foolad, M.R.; et al. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **2019**, *51*, 1044–1051. [CrossRef] [PubMed]

12. Tranchida-Lombardo, V.; Aiese Cigliano, R.; Anzar, I.; Landi, S.; Palombieri, S.; Colantuono, C.; Bostan, H.; Termolino, P.; Aversano, R.; Batelli, G.; et al. Whole-genome re-sequencing of two Italian tomato landraces reveals sequence variations in genes associated with stress tolerance, fruit quality and long shelf-life traits. *DNA Res.* **2018**, *25*, 149–160. [CrossRef] [PubMed]

13. Cambiaso, V.; Pratta, G.R.; Pereira da Costa, J.H.; Zorzoli, R.; Francis, D.M.; Rodríguez, G.R. Whole genome re-sequencing analysis of two tomato genotypes for polymorphism insight in cloned genes and a genetic map construction. *Sci. Hortic. (Amsterdam)* **2019**, *247*, 58–66. [CrossRef]

14. Blanca, J.; Montero-Pau, J.; Sauvage, C.; Bauchet, G.; Illa, E.; Díez, M.J.; Francis, D.; Causse, M.; van der Knaap, E.; Cañizares, J. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genom.* **2015**, *16*, 1–19. [CrossRef] [PubMed]

15. Razifard, H.; Ramos, A.; Della Valle, A.L.; Bodary, C.; Goetz, E.; Manser, E.J.; Li, X.; Zhang, L.; Visa, S.; Tieman, D.; et al. Genomic Evidence for Complex Domestication History of the Cultivated Tomato in Latin America. *Mol. Biol. Evol.* **2020**. [CrossRef] [PubMed]

16. Eshed, Y.; Abu-Abied, M.; Saranga, Y.; Zamir, D. Lycopersicon esculentum lines containing small overlapping introgressions from L. pennellii. *Theor. Appl. Genet.* **1992**, *83*, 1027–1034. [CrossRef] [PubMed]

17. Eshed, Y.; Zamir, D. A genomic library of Lycopersicon pennellii in L. esculentum: A tool for fine mapping of genes. *Euphytica* **1994**, *79*, 175–179. [CrossRef]

18. Eshed, Y.; Zamir, D. Introgressions from Lycopersicon pennellii can improve the soluble-solids yield of tomato hybrids. *Theor. Appl. Genet.* **1994**, *88*, 891–897. [CrossRef]

19. Eshed, Y.; Zamir, D. An Introgression Line Population of Lycopersicum pennellii in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated QTL. *Genetics* **1995**, *141*, 1147–1162.

20. Eshed, Y.; Zamir, D. Less-than-additive epistatic interactions of quantitative trait loci in tomato. *Genetics* **1996**, *143*, 1807–1817.

21. Lippman, Z.B.; Semel, Y.; Zamir, D. An integrated view of quantitative trait variation using tomato interspecific introgression lines. *Curr. Opin. Genet. Dev.* **2007**, *17*, 545–552. [CrossRef] [PubMed]

22. Fridman, E.; Pleban, T.; Zamir, D. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 4718–4723. [CrossRef] [PubMed]

23. Fridman, E.; Carrari, F.; Liu, Y.; Fernie, A.R.; Zamir, D. Zooming in on a quantitative interspecific introgressions. *Science* **2004**, *305*, 1786–1790. [CrossRef] [PubMed]

24. Sun, L.; Rodriguez, G.R.; Clevenger, J.P.; Illa-Berenguer, E.; Lin, J.; Blakeslee, J.J.; Liu, W.; Fei, Z.; Wijeratne, A.; Meulia, T.; et al. Candidate gene selection and detailed morphological evaluations of fs8.1, a quantitative trait locus controlling tomato fruit shape. *J. Exp. Bot.* **2015**, *66*, 6471–6482. [CrossRef] [PubMed]

25. Soyk, S.; Lemmon, Z.H.; Oved, M.; Fisher, J.; Liberatore, K.L.; Park, S.J.; Goren, A.; Jiang, K.; Ramos, A.; van der Knaap, E.; et al. Bypassing negative epistasis on yield in tomato imposed by a domestication gene. *Cell* **2017**, *169*, 1–14. [CrossRef]

26. Stevens, M.A.; Dickinson, G.L.; Aguirre, M.S. *UC82—A High Yielding Processing Tomato*; UC-Davis Bull. Veg. Crop. Ser. 183; Vegetable Crops Department, University of California: Davis, CA, USA, 1976.

27. Stevens, M.A.; Dickinson, G.L.; Aguirre, M.S. UC-82 Pedigree. *Rep. Tomato Genet. Coop.* **1977**, *27*, 65.

28. Lawson, D.M.; Lunde, C.F.; Mutschler, M.A. Marker-assisted transfer of acylsugar-mediated pest resistance from the wild tomato, Lycopersicon pennellii, to the cultivated tomato, Lycopersicon esculentum. *Mol. Breed.* **1997**, *3*, 307–317. [CrossRef]

29. Carter, J.D.; Pereira, A.; Dickerman, A.W.; Veilleux, R.E. An Active Ac/Ds Transposon System for Activation Tagging in Tomato Cultivar M82 Using Clonal Propagation. *Plant Physiol.* **2013**, *162*, 145–156. [CrossRef]

30. Suzuki, R.; Shimodaira, H. Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **2006**, *22*, 1540–1542. [CrossRef]

31. Sol Genomics Network International Tomato Genome Sequencing Project Release Version 3.0. Available online: https://solgenomics.net/organism/Solanum_lycopersicum/genome (accessed on 8 April 2020).

32. Sequence Read Archive Submissions Staff Using the SRA Toolkit to Convert .Sra Files into other Formats. Available online: https://www.ncbi.nlm.nih.gov/books/NBK158900/ (accessed on 8 April 2020).

33. Andrews, S. FastQC a Quality Control Tool for High Throughput Sequence Ddata. Available online: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 8 April 2020).

34. Babraham Bionformatics Group Trim Galore! Babraham Institute (UK). Available online: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed on 8 April 2020).

35. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef] [PubMed]

36. Broad Institute Picard Tools. Available online: http://broadinstitute.github.io/picard/ (accessed on 8 April 2020).

37. García-Alcalde, F.; Okonechnikov, K.; Carbonell, J.; Cruz, L.M.; Götz, S.; Tarazona, S.; Dopazo, J.; Meyer, T.F.; Conesa, A. Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics* **2012**, *28*, 2678–2679. [CrossRef] [PubMed]

38. McKenna, A.; Hanna, M.; Banks, E.; Sivachenko, A.; Cibulskis, K.; Kernytsky, A.; Garimella, K.; Altshuler, D.; Gabriel, S.; Daly, M.; et al. The genome analysis toolkit: A mapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **2010**, *20*, 1297–1303. [CrossRef] [PubMed]

39. DePristo, M.A.; Banks, E.; Poplin, R.; Garimella, K.V.; Maguire, J.R.; Hartl, C.; Philippakis, A.A.; del Angel, G.; Rivas, M.A.; Hanna, M.; et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **2011**, *43*, 491–498. [CrossRef] [PubMed]

40. Danecek, P.; Auton, A.; Abecasis, G.; Albers, C.A.; Banks, E.; DePristo, M.A.; Handsaker, R.E.; Lunter, G.; Marth, G.T.; Sherry, S.T.; et al. The variant call format and VCFtools. *Bioinformatics* **2011**, *27*, 2156–2158. [CrossRef]

41. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **2014**, *30*, 2843–2851. [CrossRef]

42. Cingolani, P.; Patel, V.M.; Coon, M.; Nguyen, T.; Land, S.J.; Ruden, D.M.; Lu, X. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Front. Genet.* **2012**, *3*, 1–9. [CrossRef]

43. Drori, E.; Levy, D.; Rahimi, O. Genome analysis CircosVCF: Circos visualization of whole-genome sequence variations stored in VCF files. *Bioinformatics* **2017**, *33*, 1392–1393. [CrossRef]

44. Raffaele, S.; Farrer, R.A.; Cano, L.M.; Studholme, D.J.; MacLean, D.; Thines, M.; Jiang, R.H.Y.; Zody, M.C.; Kunjeti, S.G.; Donofrio, N.M.; et al. Genome evolution following host jumps in the irish potato famine pathogen lineage. *Science* **2010**, *330*, 1540–1543. [CrossRef]

45. Kosugi, S.; Natsume, S.; Yoshida, K.; MacLean, D.; Cano, L.; Kamoun, S.; Terauchi, R. Coval: Improving Alignment Quality and Variant Calling Accuracy for Next-Generation Sequencing Data. *PLoS ONE* **2013**, *8*, e75402. [CrossRef]

46. Van der Auwera, G.A.; Carneiro, M.O.; Hartl, C.; Poplin, R.; del Angel, G.; Levy-moonshine, A.; Jordan, T.; Shakir, K.; Roazen, D.; Thibault, J.; et al. From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **2013**, *11*, 1–33.

47. Rudich, Y.; Harel, S.; Geisenberg, H. M-82-1-8 (VF) A new variety of tomatoes for canning. *Hassadeh* **1977**, *57*, 846–850.