Dna Research, 2017, 24(1), 81–91
doi: 10.1093/dnares/dsw050
Advance Access Publication Date: 8 December 2016
Full Paper

OXFORD

Full Paper

Exploiting the great potential of Sequence Capture data by a new tool, SUPER-CAP

Valentino Ruggieri[1,2], Irantzu Anzar[2], Andreu Paytuvi[2], Roberta Calafiore[1], Riccardo Aiese Cigliano[2,*], Walter Sanseverino[2,*], and Amalia Barone[1]

[1]Department of Agricultural Sciences, University of Naples Federico II, Via Università 100, 80055 Portici (NA), Italy, and [2]Sequentia Biotech SL, Calle Compte d'Urgell, 240, 08035 Barcelona, Spain

*To whom correspondence should be addressed. Tel. 0034 930107368. Email: raiesecigliano@sequentiabiotech.com (R.A.C.); wsanseverino@sequentiabiotech.com (W.S.)

Edited by Dr. Mikita Suyama

Received 25 May 2016; Editorial decision 21 October 2016; Accepted 26 October 2016

Abstract

The recent development of Sequence Capture methodology represents a powerful strategy for enhancing data generation to assess genetic variation of targeted genomic regions. Here, we present SUPER-CAP, a bioinformatics web tool aimed at handling Sequence Capture data, fine calculating the allele frequency of variations and building genotype-specific sequence of captured genes. The dataset used to develop this in silico strategy consists of 378 loci and related regulative regions in a collection of 44 tomato landraces. About 14,000 high-quality variants were identified. The high depth (>40×) of coverage and adopting the correct filtering criteria allowed identification of about 4,000 rare variants and 10 genes with a different copy number variation. We also show that the tool is capable to reconstruct genotype-specific sequences for each genotype by using the detected variants. This allows evaluating the combined effect of multiple variants in the same protein. The architecture and functionality of SUPER-CAP makes the software appropriate for a broad set of analyses including SNP discovery and mining. Its functionality, together with the capability to process large data sets and efficient detection of sequence variation, makes SUPER-CAP a valuable bioinformatics tool for genomics and breeding purposes.

Key words: target enrichment, sequence reconstruction, heterozygous variants, web tool analysis

1. Introduction

Identifying sequence polymorphisms responsible for phenotypic variation and understanding the genetic basis of complex traits have been two major challenges of plant molecular genetics since the development of the first molecular markers up to the impressive high-throughput genomics technologies available today.[1,2] The process of genotype-phenotype association is one of the central goals in the path towards plant improvement, and this requires the use of all the genetic and genomic information available for a given individual and/or population. The association of the genomic variation with traits of interest requires the reliable detection and the systematic investigation of the entire spectrum of DNA variability, including single nucleotide polymorphisms (SNPs), insertions/deletions (INDELs) as well as copy number variations (CNVs) and presence/absence of variations (PAV). Substantial progress towards this goal was made in the last few years. Large germoplasm collections have been characterized using SSR,[3,4] SNP markers and SNP arrays.[5–9] However, most of these approaches had known disadvantages that limited the

power of the detection, in particular because the SNPs used in these studies were selected from a limited number of divergent sources and were commonly chosen to exceed a minimum frequency of the rare allele. Whole genome sequencing has also been used to explore individual variation at the genomic level in plants[10–13] but, due to its high cost and complexity in data analysis, it is not expected to be widely applied for investigation of variants underlying specific traits of interest at a sufficient coverage. Recently, the development of microarray based or liquid-based genomic selection methods, commonly referred to as 'Sequence Capture', provided an affordable way to produce high-quality variants, thus solving most of the ascertained bias reported.[14–16] Combining the recently developed targeted sequence enrichment with Next-Generation Sequencing (NGS) technologies, Sequence Capture methodology represents a powerful strategy for enhancing data generation to assess genetic variation of target regions, and it is likely to replace PCR as the main target enrichment method in both plants and animals.[17] In addition, setting the experiments to obtain a sufficient depth of coverage would support the detection of rare functional variants and allow the discovery of complete haplotypes of genes, as well as CNV and PAV. With the ever-decreasing costs of sequencing and the advances in Sequence Capture technologies, these approaches are nowadays largely applied in different fields. Thanks to its high reliability and moderate cost per experiment, Sequence Capture is becoming an affordable diagnostic tool for medical and personalized medicine purposes.[18] Despite its large employment in the medical field, the use of Sequence Capture in plant science is still emerging. Sequence Capture experiments were carried out for few species, including maize,[19] strawberry,[20] rapeseed canola,[21,22] black cottonwood,[23] potato,[24] wheat,[25,26] medicago[17] and cassava.[27]

The work herein proposed, which investigates the sequence variation of a group of 378 genes and the related regulative regions in a collection of 44 landraces, represents the first study of Sequence Capture in tomato species. Although software tools are available for variant calling and variant mining, their application is not straightforward, requiring users to install various packages and to convert data into different formats. This lack of easily accessible software pushed us to propose a web-based tool, named SUPER-CAP (http://supercap.sequentiabiotech.com/), to boost quick, proficient and affordable analysis of sequence capture data. This tool, benefitting from SUPER-W pipeline[28] and combining different software and customizable procedures, assists the user in sequence capture experiments from the identification of single-nucleotide variants (SNVs) and small insertions and deletions (INDELs) to the reconstruction of genotype-specific (hereafter referred to as 'private') sequences/gene features for each target region of each sample.

In addition, in this study, an explorative investigation on different aspects affecting sequence capture data was also carried out. For instance, since one of the major challenges in the enrichment/capture technologies are to avoid spurious variants calling for heterozygous (He) loci, this work evaluated how the depth of the reads and the procedure set-ups impact the variant calling of He variants. In fact, being the tomato a highly homozygous (Ho) species, it represents a good model for testing He variant calling into the gene set considered. In addition, spurious callings are frequently referred to multi-copy gene families (duplicated region/high homologous regions).[29] The gene set considered in this study represents a large range of variability in terms of gene families. This allowed for suitable assessment and estimation of the effective presence of He variants among the multi-copy genes.

As a final point, the variant set was also used to investigate: (i) the level and the type of sequence polymorphism in captured genes across the 44 tomato samples, (ii) the pattern of variants distribution among the lines, with respect to the identification of rare variants, (iii) the number of genes CNV, PAV and (iv) the functional annotation of the variants in order to prioritize the study of highly relevant variants. The work herein proposed represents a framework for easy Sequence Capture-related experiments that will promote similar studies, as well as in differing species.

2. Material and methods

2.1. Plant material and selection of target genes

A panel of 44 genetically diverse *Solanum lycopersicum* genotypes was selected from a wide collection of tomato landraces[30] available at the University of Naples, department of Agricultural Sciences. Plants were grown in a greenhouse under controlled conditions at the aforementioned Department; detailed information, including source and geographic origin for each genotype, is presented in Supplementary Table S1.

Genomic DNA from the 44 genotypes was isolated from young tomato leaves using the DNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA). DNA quantity and purity/degradation was firstly checked on 1% agarose gel. In order to fit the standard parameters for Sequence Capture analyses, DNA concentration and quality were determined by Nanodrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and Qubit fluorometer (Life Technologies, Darmstadt, Germany) according to the manufacturer's requirements.

An inventory of 378 candidate genes representing possible targets for antioxidant metabolism in tomatoes was compiled from published literature, previous research[31,32] and exploration of metabolic pathway databases (LycoCyc at http://solgenomics.net/and KEGG at http://www.genome.jp/kegg/). Genes were also selected to represent different gene family sizes. In total 235 out of the 378 genes completely represent 25 gene families. In particular, 171 genes belong to eight large gene families (we declared a big family if composed by more than 10 copies in the genome), 40 belong to seven medium gene families (between nine and four gene copies), and 24 belong to 10 small gene families (two or three copies). Six out of the 378 genes represent single copy genes. Moreover, in order to take into account variations in the regulatory region of each selected gene, a 3 Kbp promoter region was also included in the analysis. For genes with an intergenic distance shorter than 3 Kbp, only the related shorter portion of the promoter was considered. Supplementary Table S2 shows detailed information of the genomic regions considered.

2.2. Sequence capture design and sequencing

Probe design and gene enrichment were performed following the protocol provided with the solution-based Roche NimbleGen SeqCap EZ Library (Roche-NimbleGen, Madison, WI, USA). *Loci* coordinates of the genic and promotor regions were identified and submitted to Roche Diagnostics for probe design using NimbleDesign and SignalMap software (http://www.nimblegen.com/products/software/index.html). This probe set contains probes with up to 20 close matches in the genome as determined by the Sequence Search and Alignment by Hashing Algorithm (http://www.sanger.ac.uk/resources/software/ssaha/). Acco rding to NimbleGen specification, we considered a probe to match the genome if the variation ratio was <0.05. Following the SeqCapEZ protocol, 44 paired-end libraries were prepared using the Illumina Kapa Library Prep Kit and were multiplexed and sequenced with HiSeq 1,500 (read size:100 bp) according to Illumina specifications.

2.3. SUPER-CAP usage and procedures

SUPER-CAP is a bioinformatics tool specifically designed for accurate mapping, variant calling and reconstruction of private sequences using the variations detected from Sequence Capture data. SUPER-CAP includes an updated version of SUPER-W (release 4).[28] SUPER-CAP is a user-friendly web tool which only needs two input files to work: the captured region file in BED format and the filtered NGS reads in FASTQ format. The main steps and procedure underlying the SUPER-CAP tool, including the calling variants, the variants filtering and the targeted sequence reconstruction are presented in Figure 1.

SUPER-W has been modified to specifically handle sequence capture data. In this new version, SUPER-W uses as input the BED file of capture probe design (it can also handle a whole exome) and the filtered NGS reads. The first step is to map all the samples against a reference genome (specified by the user) with BWA (version 0.7.5; options used: mem:BWA-MEM algorithm).[33] The mapped files are processed for PCR duplicates (Picard, MarkDuplicates tool, version 1.118), filtered for quality (minimum quality required: 30), sorted and indexed.[34] The resulting BAM file is then used for the variant calling step. Small variations (SNPs and short deletion and insertion polymorphisms), large variations (deletions, inversions and duplications), or both, can be set for detection. The small variations are called with SAMtools[34] through an accurate and sensitive double calling step, while the structural variants are detected using LUMPY (version 0.0.11; options used: -mv 4 –tt 0 –pe –sr).[35]

Once the variant calling step has been successfully completed, statistics on sequence capture experiment are calculated. The statistics report can be considered as a checkpoint, allowing the user to check the IN/OFF target read count, which highlights the specificity or sensitivity of the sequence capture experiment.

Raw genomic variants annotated by SUPER-W are then classified and filtered out. As default, only variants with a coverage >6× and a Phred quality >30 were used for subsequent analysis. Furthermore, variants overlapping each other are removed from downstream analysis as they cause a low confidence region. Moreover, a special option has been released to allow calling variants according to customized allele frequency (AF). As default value, variants with an AF between 0 and 0.2 are considered Ho for the reference allele, variants with a AF between 0.4 and 0.6 are considered He while variants with an AF between 0.8 and 1 are considered Ho for the alternative allele. Variants with an AF ranging out of boundaries set (i.e. between 0.2 and 0.4 and between 0.6 and 0.8) are called with an alert comment in the output VCF file. Coverage and quality filters as well as the AF parameters can be manually modified by the user. Finally, the filtered variants were classified according to type (SNP or

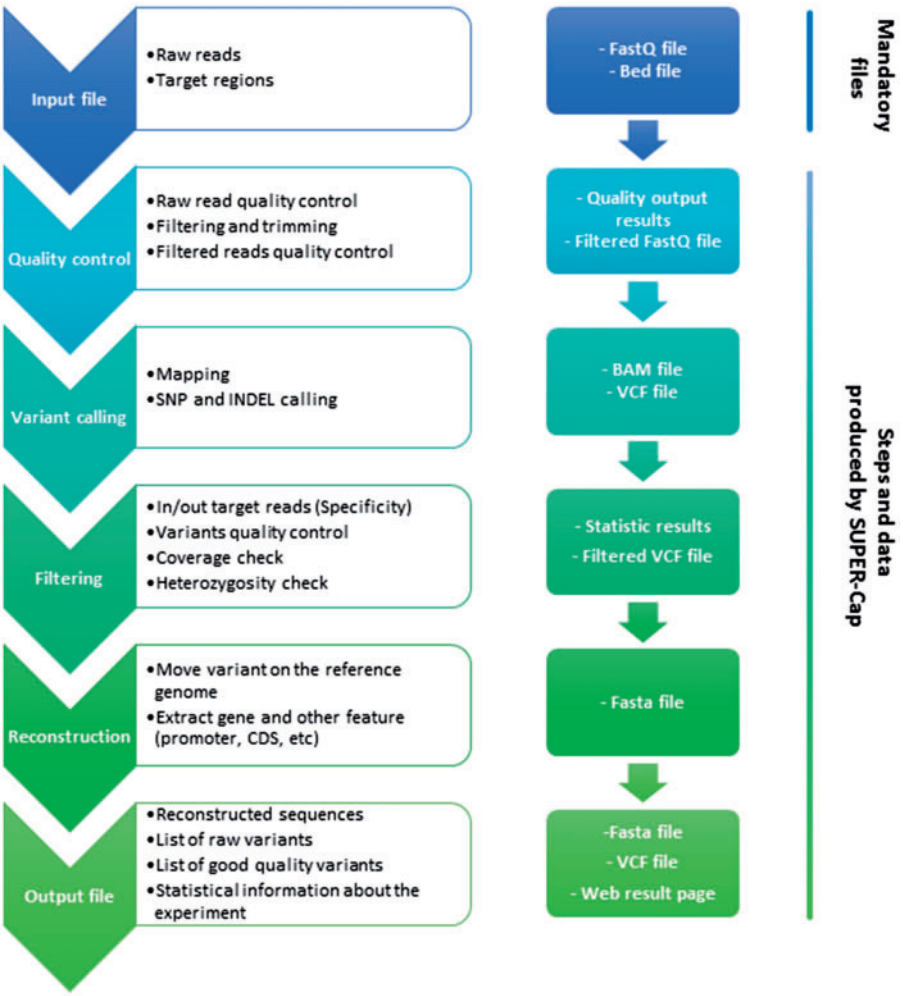


Figure 1. The SUPER-CAP tool. The main steps and procedures underlying the tool are graphically reported in the workflow.

INDEL). For SNP variants the breakdown into transitions and transversions was also determined. To improve the variant features, SnpSift and SnpEff[36] were implemented in the workflow, which allows imputing the gene region in which the variants fall (promoter, intron, exon, Untranslated regions (UTRs)) and the putative impact of the variants on protein functionality.

The high-quality variants, the reference genome and the BED file of the captured regions were used to perform a targeted sequence reconstruction. Using bcftools[37] and BEDtools[38] utilities, a new consensus sequence was created by applying the filtered variants to the reference sequence. He SNPs were introduced using IUPAC ambiguity codes and He INDELs were discarded. Through a liftover process, which recalculates the gene coordinates after the insertion or deletion of INDELs, a FASTA file with reconstructed sequences was created. Variants reconstructed were also displayed in the SUPER-CAP's browser results page, allowing the user to explore the type and the number of variants easily.

2.4. Web tool development and design

With the aim of supplying SUPER-CAP with a graphical user interface, we developed a NodeJS-based web tool (version 4.4.1. Available at https://nodejs.org/), available at 'http://supercap.sequentiabiotech.com/'. Using this tool, the user can upload the two required files (the BED file of the target regions and the FASTQ files of the sequenced reads) and choose the analysis parameters (i.e. the quality thresholds for the filtering procedure and the AF range for the He /Ho assignation). An ID is given after completion of the procedure, which is required for accessing the results page. Both the reconstructed sequences in FASTA format and the whole variants in VCF format can be downloaded from the 'Download section'. In addition, the visualization of the reconstructed sequence is also shown in an interactive graphical interface in the results page. Here the user can explore, for each gene in each genotype, the type and the number of variants that have been incorporated and the effect they produced on the reference protein. General mapping statistics, including IN/OFF targets, as well as the sensitivity of the experiment, are displayed in the same page.

2.5. Use of SUPER-CAP on tomato experimental data

The reads obtained by sequencing the tomato DNA samples were quality checked using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Then, reads were filtered for quality and trimmed with the Trimmomatic tool (version 0.30; options used: –windowsSize 3 –requiredQuality 22 –MINLEN 30).[39] By using the filtered reads, the following procedures were undertaken: reads falling in the designed regions with a PHRED quality >30 and a minimum length of 30 bp were mapped onto the tomato genome SL2.50.[40] Then, PCR duplicates and multiple mapped reads were removed by filtering on a meaningful MAPQ score of 30. Once the variants were called, in order to reduce the number of false positives calls, an additional filter was applied and only variants with coverage higher than 6× and a PHRED quality >30 were selected. The total numbers of He and Ho variants, as well as their ratio, were calculated with default parameters (0.4–0.6 as optimal AF range for the He calling). A Neighbour-joining dendrogram was built in TASSEL v.5[41] by using the detected filtered variants, and the related tree was graphically refined using the Fig-tree software (http://tree.bio.ed.ac.uk/software/figtree/). SNPs and INDELs were also structurally and functionally annotated with SnpEff v4.2[36] using the tomato iTAG2.40 annotation.

2.6. Validation of variants

To test the reliability of the adopted procedures, a validation of the variants detected was achieved by exploring publicly available data for tomato. In particular, data from a previous genotyping experiment which detected variants for 7,720 SNPs through the SolCAP array,[30,32] as well as data concerning the resequencing of 360 tomato genomes,[42] were considered. Of the 44 genotypes used in this study, 37 were common to those used in the SolCAP experiment and 9 in the 360 genome project. The validation was performed by comparing the number and the concordance of shared variations among the common genotypes. Only Ho variations were used for the validation. Intersection of the sets was carried out by intersectBed tool[38] and the comparison analysis by using VCFtools.[43]

2.7. CNV and PAV identification

The coverage and total number of reads were evaluated for each sample. Alignments were visualized using the IGV browser version 2.3.3 (http://www.broadinstitute.org/igv/). CNV was computed following a modified procedure of Schiessl *et al.* (2014)[22] using a normalized read coverage for each captured region. CNV in a given target region was assumed if the ratio of normalized coverage (genotype)/normalized coverage (all genotypes) was <0.5 or >1.5, respectively. PAV was assumed if the ratio was <0.05.

3. Results and discussion

3.1. Sequence capture experiment

We designed a custom capture probe experiment by using a customizable and cost-effective approach based on Roche-NimbleGen SeqCap EZ technology. This framework helped to set up an *in silico* approach (SUPER-CAP) with the aim of facilitating further sequence capture studies.

This study investigated the sequence variation of a group of 378 genes and related regulative regions. The total size of the sequenced region was 2,338,578 bp, of which 1,353,817 bp corresponded to genic regions and 984,761 bp to their related promoter regions. Capturing probes were developed by Roche NimbleGen to target the specific regions. The final design was non-redundant and covered about 92.6% of the targeted regions. The 7.4% of non-covered regions represents those that did not match the minimum parameters for the capture and in the majority of the cases (>95%) these were promoter regions. It has already been reported that the cover design slightly varies depending on different factors, including the commercial technology adopted.[44]

After the enrichment of the targeted regions, each capture library was sequenced in one Illumina HiSeq lane. The number of raw reads generated *per* sample varied between 1,103,352 and 2,683,868 (Table 1). Sequenced reads were analyzed in order to remove low-quality regions. After quality trimming, between 71 and 93% of the reads were retained and then used as input for SUPER-CAP. Data were then statistically evaluated for: (i) alignment rates, the number of mapped reads on the total reads, (ii) specificity, the number of reads that map to the targeted sequence and (iii) sensitivity, the percentage of targeted bases covered by sequence reads. The reads from samples 21A, 1A and 64A showed the highest alignment rates (>98%). In contrast, the lowest alignment rate was observed in genotypes 85A, 79A and 38A (<78%). The alignment success was however independent of the total number of reads. Specificity of the capture was imputed considering the number of reads in the target interval on the number of mapped reads. It showed an average value

Table 1. Alignment statistics for the 44 individual captures (Sample ID)

Sample ID	Raw reads	Mapped reads	Mapped reads Q > 30	Mapped reads Q > 30 WD	Alignment rate (%)	On target reads (%)	On target reads + 200 bp (%)	Mean depth of coverage
1A	1,344,456	1,327,567	1,269,157	1,251,088	98.74	75.45	76.06	43.2
3A	1,314,574	1,293,498	1,234,679	1,213,125	98.4	75.07	75.65	42.42
5A	1,413,706	1,391,398	1,331,163	1,310,092	98.42	75.94	76.55	45.97
8A	1,376,510	1,342,280	1,279,940	1,262,285	97.51	74.93	75.51	43.94
14A	1,580,074	1,231,066	1,142,665	1,128,063	77.91	67.22	67.75	39.67
15A	1,686,492	1,340,263	1,274,719	1,262,193	79.47	75.66	76.27	44.92
20A	1,480,340	1,162,656	1,101,676	1,056,514	78.54	71.73	72.28	40.27
21A	1,304,914	1,289,290	1,228,450	1,211,447	98.8	74.62	75.21	41.77
26A	1,103,352	871,462	830,910	821,995	78.98	75.72	76.35	30.03
27A	1,469,212	1,436,034	1,365,188	1,344,283	97.74	74.73	75.3	47.02
28A	1,354,892	1,332,311	1,264,586	1,190,895	98.33	70.39	70.87	40.76
30A	1,397,308	1,379,189	1,319,320	1,300,299	98.7	75.76	76.43	45.43
32A	1,893,224	1,480,474	1,410,253	1,394,575	78.2	74.55	75.18	48.69
34A	1,891,936	1,478,312	1,411,043	1,396,990	78.14	75.18	75.84	48.54
35A	1,788,530	1,398,636	1,333,126	1,316,721	78.2	74.51	75.16	46.02
38A	1,886,966	1,468,035	1,399,866	1,386,113	77.8	74.57	75.21	48.23
40A	2,651,822	2,575,073	2,463,126	2,432,466	97.11	76.67	77.25	88.92
41A	2,683,868	2,594,090	2,482,444	2,451,378	96.65	76.75	77.31	90.36
42A	1,450,568	1,424,424	1,348,207	1,329,068	98.2	74.28	74.84	46.26
43A	2,082,552	2,006,847	1,921,991	1,899,495	96.36	77.02	77.59	70.57
45A	2,409,462	2,316,282	2,221,361	2,193,201	96.13	76.59	77.18	80.52
57A	1,297,772	1,275,552	1,216,245	1,198,785	98.29	75	75.56	42.68
64A	1,446,368	1,429,441	1,366,044	1,346,218	98.83	75.98	76.56	47.44
66A	2,267,108	2,176,890	2,081,441	2,057,807	96.02	76.59	77.17	74.86
70A	1,372,572	1,297,126	1,223,744	1,206,911	94.5	72.75	73.28	41.49
75A	1,799,384	1,396,835	1,323,958	1,312,523	77.63	73.81	74.38	45.59
78A	2,060,104	1,992,077	1,903,906	1,876,376	96.7	76.81	77.4	72.14
79A	1,666,724	1,295,896	1,235,321	1,224,643	77.75	75.02	75.63	44.71
85A	1,879,204	1,439,306	1,364,666	1,344,083	76.59	73.83	74.42	47.2
87A	2,342,600	2,247,877	2,150,105	2,124,074	95.96	76.24	76.83	77.91
92A	2,247,742	2,165,543	2,069,466	2,043,464	96.34	75.25	75.84	73.65
93A	1,877,134	1,465,142	1,394,154	1,379,154	78.05	74.36	74.95	48.29
94A	2,105,320	2,019,814	1,927,953	1,904,794	95.94	75.59	76.25	68.51
97A	2,489,394	2,402,528	2,299,398	2,270,879	96.51	76.11	76.72	82.59
99A	2,479,650	2,380,804	2,280,336	2,251,710	96.01	76.49	77.11	81.97
102A	1,760,260	1,374,039	1,305,068	1,293,203	78.06	73.31	73.91	44.39
103A	1,969,444	1,886,176	1,780,373	1,747,969	95.77	71.5	72	61.9
105A	1,942,196	1,516,300	1,450,772	1,438,620	78.07	76.3	76.9	51.32
109A	1,835,360	1,428,227	1,356,340	1,343,895	77.82	74.8	75.39	49.28
111A	2,541,684	2,006,126	1,913,540	1,893,325	78.93	76.22	76.82	68.21
115A	1,547,164	1,229,818	1,166,492	1,099,375	79.49	70.75	71.3	41.5
117A	1,682,246	1,333,089	1,273,573	1,262,226	79.24	76.42	77.03	44.83
118A	1,665,304	1,320,669	1,260,016	1,248,760	79.3	75.09	75.72	43.63
120A	1,651,964	1,308,823	1,249,979	1,237,454	79.23	76.03	76.64	43.62
Mean	**1,806,624**	**1,602,892**	**1,527,880**	**1,505,875**	**88.72**	**74.80**	**75.40**	**53.89**

The number of raw reads (Raw reads), mapped reads (Mapped reads), mapped reads with quality Q > 30 (Mapped reads Q>30) and without duplications (Mapped reads Q > 30 WD) were reported. For each sample, the alignment rate (Alignment rate), the specificity (On target reads), the specificity including the flanking regions of 200 bp (On target reads + 200 bp) and the average depth of coverage (Mean depth of coverage) are also reported.

of 74.80%. The minimum value 67.2% was detected for the genotype 14A and the maximum (77.02%) for the genotype 43A. In addition, considering an extension of flanking regions of 200 bp, the specificity was found to vary only slightly, showing an average increase of <1%. The range, considering flanking regions, was indeed from 67.75 to 77.59%.

Even though investigators are seeking experimental designs that generate robust scientific findings at the lowest sequencing cost, an adequate depth of coverage is very important in order to reduce the variant-error rate and the assembly gaps, and to obtain the correct call of variants, in particular for the He ones.[45] It has indeed been demonstrated that for short Illumina reads a coverage comprised between 30× and 55× appears necessary to correctly identify SNVs and small INDELs with a proper degree of reliability.[45,46]

In our experiment, the normalized mean coverage of the total targeted regions showed an average value of 53.89×, allowing very accurate detection of variants. Analysing the sensitivity of the experiment (Supplementary Fig. S1), which shows how well the

targeted region is covered for reads at depths from 1× to the max depth (>150×), we observed that at a minimum coverage of 1× about 95% of the designed target regions resulted completely covered. Between 91 and 94% of the target was covered when a minimum coverage of ten reads was applied. At 20×, 76–93% of the target regions was still covered.

Although Sequence Capture experiments represent a promising technology, only few studies have used Sequence Capture in plants (<20 in PubMed in April 2016). Probably the lacking of straightforward procedures for data analysis makes this technology still emerging. For this reason, SUPER-CAP, representing a user-friendly tool, could boost the use of this technology also for plant organisms. Moreover, in this study, the high rates reached for the capture parameters, in particular for the specificity (>74%), highlighted the high efficiency reached by combing an accurate probe design with a high performant aligner included in the pipeline (BWA), for which high confidence of sequence alignments has been already reported.[19] This makes the performances obtained in this study in terms of alignment rate and sensitivity in line with performances reported for other organisms, for example humans.[44]

3.2. Variant calling and reconstruction of private sequences

SUPER-CAP allowed to detect 25,654 unfiltered variants in the targeted regions considered. Then the filtering procedure allowed to reduce the number of false positives/low-quality variants. Only variants exhibiting coverage higher than 6× and PHRED quality higher than 30 were selected. Even if a different threshold could be applied to the filtering procedure, we observed that in our experiment a depth coverage of six represented the minimum coverage to maximize the PHRED quality (Supplementary Fig. S2). Subsequent criteria were adopted to appropriately detect the number of He variants. A preliminary survey was carried out in order to estimate the number of He variants in accordance with the range of the AF (AF of mapped reads for each locus) (Supplementary Fig. S3). As expected, an inverse correlation between the range of AF and number of He variants was observed. This indicated that the set-up of the AF-range deeply affects the estimation of the He variants in the calling procedure. In order to make this sensitive phase of the variant calling easily customizable, we allow the SUPER-CAP user to set their own AF range. In this study, we set an AF range of 0.4-0.6 for the He variants, in line with a previous study in tomato.[10] At this threshold, 86,120 variants in 14,116 sites were scored. Only 5% (4,970) of the total variants in 2,543 sites resulted to be He. Looking at the distribution of the He variants, we observed that they were not uniformly distributed across the chromosomes and genotypes. Chromosome 11 harboured the highest number of He variants per Mbp of captured region (Supplementary Fig. S4) while chromosomes 10 and 0 harboured the lowest number. Considering the 44 samples analysed, the heterozygosity ranged from 1.1 to 47% (Fig. 2), with a minimum value for sample 99A and a maximum value for sample 85A. Sample

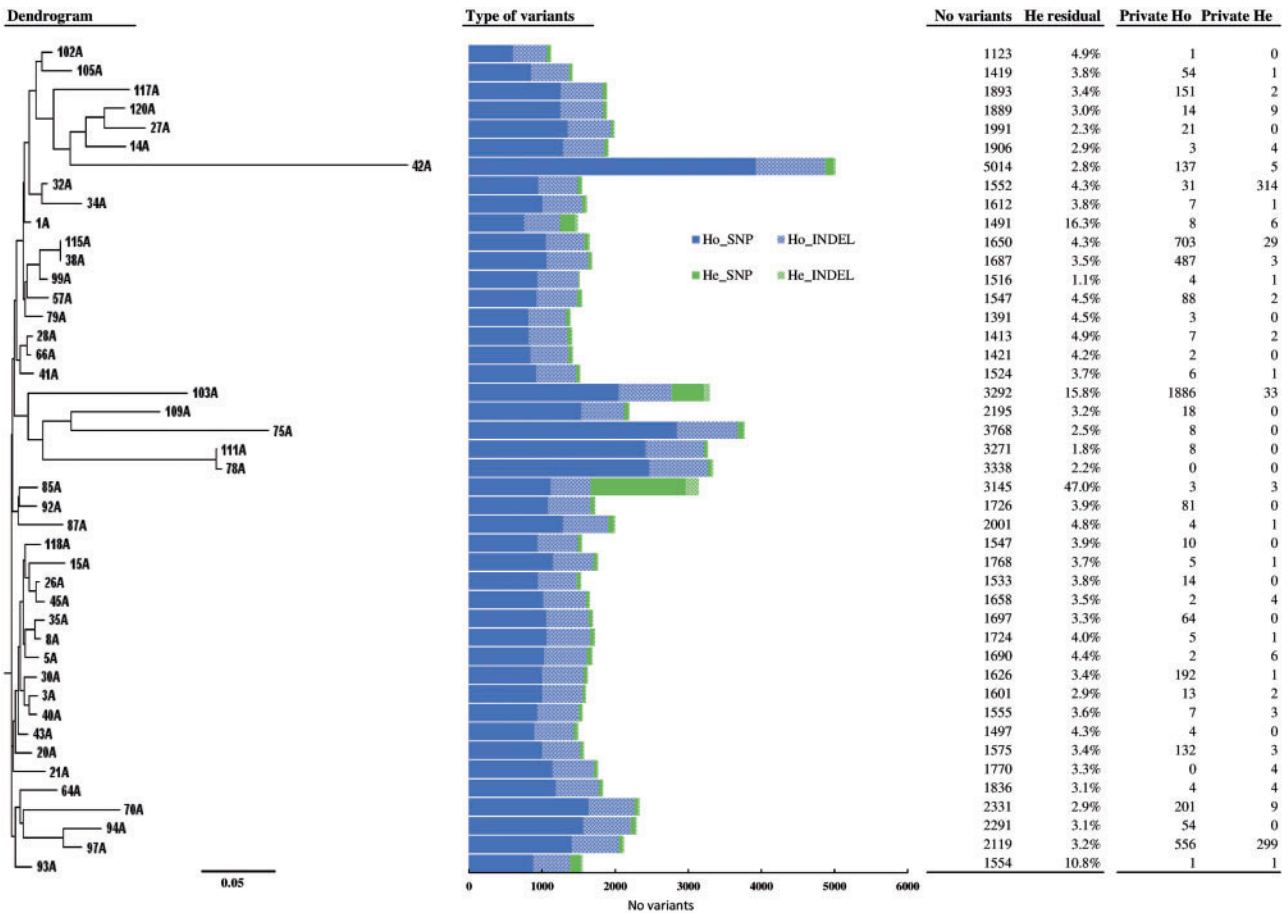


Figure 2. Distribution of variants detected for each genotype according to the type of variants. Proportion of Ho and He SNP and INDEL are reported as well as the number of private variants. A dendrogram built by using the complete set of variants shows the relationship among the genotypes.

85A showed an unexpected high percentage of residual heterozygosity, suggesting a possible sampling error of the plant material for this genotype (which could be mixed with other genotypes or to a residual segregation during the germoplasm conservation). For this reason, the sample was excluded from subsequent analyses.

Figure 2 shows the number of variants detected for each sample, with respect to the proportion of SNPs and INDELs at both Ho and He level. Although INDELs are present at lower rates than SNPs, small INDELs represent functionally important types of genomic variation.[47,48] It is reported that, compared with other sequencing technologies, Illumina does not appear to have high error rates in homopolymer regions.[49] However, in this study the INDELs surveyed were obtained after discarding positions overlapping with homopolymers >7 bp, both in the tomato reference and in resequencing sequences. This allowed to avoid possible errors due to technical bias and enabled to collect a final high-quality set of variants.

As a whole, the sample with the highest number of variants was 42A, which harbours >4,000 SNPs and about 1,000 INDELs. The sample with the lowest number of variants was 102A which exhibits 1,123 variants between SNPs and INDELs. On average, 1,357 SNPs *per* sample were scored, of which 6.8% turned out to be He. The average of INDELs *per* sample turned out to be about 600, and only 3% were He.

The high depth of coverage reached in this experiment supported the detection of rare variants. It has been suggested that rare or low-frequency variants, which are not fully captured using other conventional genotyping technology, could largely contribute to explain the effects of some traits[50] or to characterize their genetic architecture.[51] In this study 5,441 Ho and 442 He genotype-specific variants were identified. Among these, 1,965 Ho and 216 He variants resulted still private variants when compared with external public repository for SNPs in tomato (360 tomato genomes[42]). This could be of interest when designing specific SNP markers or for studying the potential functionality of specific haplotypes in the landraces collection used in this survey. However, the rare variants identified alone do not seem to explain the phylogenetic differences observed (Fig. 2). The genetic divergence was associated with the high number of private alleles only in a few samples (i.e. 103A and 97A).

The possibility of reconstructing private sequences from high-throughput resequencing data is still a challenge today.[28] For this reason, the last step in the SUPER-CAP workflow allows to specifically insert the detected variants into the sequence of each sample. This step, which integrates the single information of the variants in a genotype-specific gene-based view, enables to obtain a necessary starting point for further functional validation analyses. In this study, a total of 16,254 regions (378*43), each including the gene and the related promoter region, were reconstructed. In order to easily exploit the combined effect of multiple variants in the same protein, reconstruction of the CDS was also performed. Future works, benefitting from this effort, will evaluate the combined effect of the variants for each gene in order to find relevant association with antioxidant content in tomato.

3.3. Validation of variants

In order to validate the variants detected, we performed a survey by exploring the available resources already present for tomato. About 8,600 variant sites detected in the present study overlapped with those reported in the genome resequencing data of 360 tomatoes.[42] Most of the non-overlapping variants were excluded since they classified as rare variants in the experiment, being present only in specific samples. Moreover, by using nine common accessions, an analysis on the concordance of the variant calling was performed. Each of the nine samples shared a different number of variants, ranging from 191 to 985. Taking into account the Ho alternative variants only, an average of 95% concordance was found. Another validation was carried out by comparing SNPs detected on a tomato population genotyped using a SolCAP arrays,[30,32] which included 7,720 SNPs. Only 170 of the 7,720 SNPs fall in the targeted region considered in the present study, since the targeted regions represent ~1% (378/34,725) of the tomato genes. Of these 170 SNPs, a subset of 97 belongs to common accessions and was used for validation. On average, a concordance >95% was observed. In most cases the discrepancies were genotyped as He either on the SolCAP array (2.5%) or in the Sequence capture (0.6%). When excluding the He cases, the similarity increased to 98%.

3.4. CNV and PAV

CNV and PAV have been recently reported as sources of important phenotypic variation in plants.[52–54] The adequate depth coverage reached in this study allowed to hypothesize on the possible structural variations occurring among the samples considered. To do this, CNV and PAV analyses were performed to detect additional or deleted homologous loci of the 378 genes among the 43 samples, by analysing the normalized depth coverage of each sample. The comparison of read depths along the loci revealed at least 10 loci with a significant variation in depth in at least one line. In particular, nine genes (four Pectinesterases, two Phenylalanine ammonia-lyase, one Reductase, one Inositol-3-phosphate synthase, one Caffeoyl-CoA O-methyltransferase) showed a CNV, and one gene (Dehydroascorbate reductase) showed a PAV (Supplementary Table S3). Three genes showed a putative reduction of the CNV (ratio <0.5) while six showed an increase (ratio >1.5). The very low number of reads captured for two samples (103A and 15A) for the gene Solyc09g056180 was associated with a PAV and we assumed that this gene was deleted in the two samples. In addition, an experimental validation by genomic PCR corroborated this result, highlighting that no amplification of the Solyc09g056180 was present for these two genotypes as showed in Supplementary Figure S5. The analysis of the coverage revealed a uniform high level of coverage across all the samples for seven genes (Solyc00g027770, Solyc00g282510, Solyc00g030510, Solyc02g075620, Solyc03g042560, Solyc03g036470, Solyc03g071860). Indeed, these seven genes showed levels of coverage ranging between 104.58× and 192.48×, representing an increase of two to four fold compared with the average coverage encountered among all the genes (53.89×). This could be due to the lack of regions in the current release of tomato. A high number of reads collapsing on specific regions could be caused by unassembled/misassembled regions in the genome.[29] To validate this unexpected high coverage, we evaluated the depth of coverage of these genes in two further independent resequencing experiments (Heinz and Moneymaker) by using publicly available data (variation data from SGN databases). We observed that significant high coverage was obtained for three genes in at least one of the two resequencing experiments and one gene in both the resequencing experiments (data not shown). This corroborates that the reference genome could carry misleading sequences. An additional indication of this hypothesis, as also discussed below, come from the high ratio of He observed for these genes, showing that similar but not identical reads were mapped on the reference regions.

3.5. Heterozygous variants into gene families

He variant assignments represent one crucial difficulty in sequence experiments.[10,29,55] Since one of the causes of spurious He call was hypothesized to derive from duplicated genes, a specific investigation was carried out, aided by the high quality of variants detected by SUPER-CAP. Previous studies reported that false-positive signals could arise in regions with low complexity,[56] or result from misalignment of multiple copies of genes, paralogues, or pseudogenes[57] but, to our knowledge, no study has analysed the behaviour of the He variants in the gene family context.

In our collection, 235 out 378 genes completely represent 25 gene families. In particular, 171 genes belong to eight large gene families, 40 belong to seven medium gene families, 24 belong to 10 small gene families and six are single copy genes. In order to estimate the number of He variants in each group we evaluated (Fig. 3) the average He density (number of He variants for Kbp) for each gene family. On average, a lower number of He variants was identified in the single genes compared with the gene family groups. Significant differences were observed at *t*-test between single genes and Large, Medium and Small gene families ($P = 0.007$) but not among gene families themselves ($P > 0.05$). Although these results evidence that gene families, despite the number of copies, are more prone to produce He variants compared with single genes, the high standard deviation detected showed that differences seem to be associated with specific cases in each gene family and not due to a general behaviour. It appears evident that in each gene family only specific genes showed a very high level of heterozygosity in almost all the genotypes (Supplementary Fig. S6). In particular it was evident that Solyc00g282510, Solyc03g042560, Solyc00g027770 and Solyc01g091060 (belonging to large gene families) and Solyc12g098090 (belonging to a medium gene family) showed a skewed rate of He variants. Some possible explanations of this result were hypothesized. In particular, exploring if neighbouring regions of these genes showed a similar rate of heterozygosity (by using information from the 360 genome project), we evidenced that a similar high rate of He variants was found only for Solyc03g042560 in a region spanning ~300 Kbp.

Another source of false positive He variants might be due to duplicated regions found in the samples that are not found in the reference genome. In such cases, the variants will be due to reads from the two copies (or more) that are piled in the only copy found in the reference genome. Such biases could occur because of the limited number of genotypes on which the original reference sequence was based, or sequencing and alignment errors (Lander *et al.*, 2001). Fortunately, these cases can be highlighted looking at excesses of depth of coverage. Three out of the five genes (Solyc00g282510, Solyc00g027770, Solyc01g091060) exhibiting a high He density showed a high depth of coverage in parallel (as also evidenced in the previous paragraph), generally doubled or tripled in comparison to the average value of all the samples. This proves that the tomato reference genome might carry misleading sequences for those regions and that the He calls observed are likely to be false-positives.

3.6. Variants classification and annotation

As an additional step, the functional annotation of the variants has also been investigated by using SnpEff and SNPsift programs. The SNPs discovered were classified according to their type (transition vs transversion and insertions vs deletion) and the genomic region they were found in. A significant number of the SNPs discovered (6,098, 44.2%) were located in the genic region and, among these, about 21% were found within exons, 73% within introns, 1.73% within the 5′-UTR and 3.68% within the 3′-UTR (Fig. 4). The remaining 7,674 (55.8%) were found to be within the promoter region. Considering the number of variants per Kbp, eight variants per Kbp were found in the promoter region on average, while only 2.2 variants per Kbp were found in the exonic regions. Introns and UTRs showed more than double the number of variants per Kbp compared with exonic regions. Genic and intergenic variants were classified as transition/transversion and insertion/deletion typologies. Transitions represent approximately 64.54% of all post-filtered SNPs (11,495), with 44.2% (5,160) of them being located in genic regions. Deletions in average were more represented than insertions, particularly in promoter and intron regions. Then, the 6,098 variants of the genic regions were annotated according to their putative effect on the protein by using SnpEff v4 (Table 2). The goal of annotating variants is to provide a prediction of which ones are functionally relevant. Our detected variants were assigned to a diverse range of functional classes, with the majority (78%) classified as 'modifier', therefore without predictable effects being located in the UTR regions or in the introns. These may have little or no effect on the phenotype. Among the remaining variants, about 12% showed a 'low effect', since they

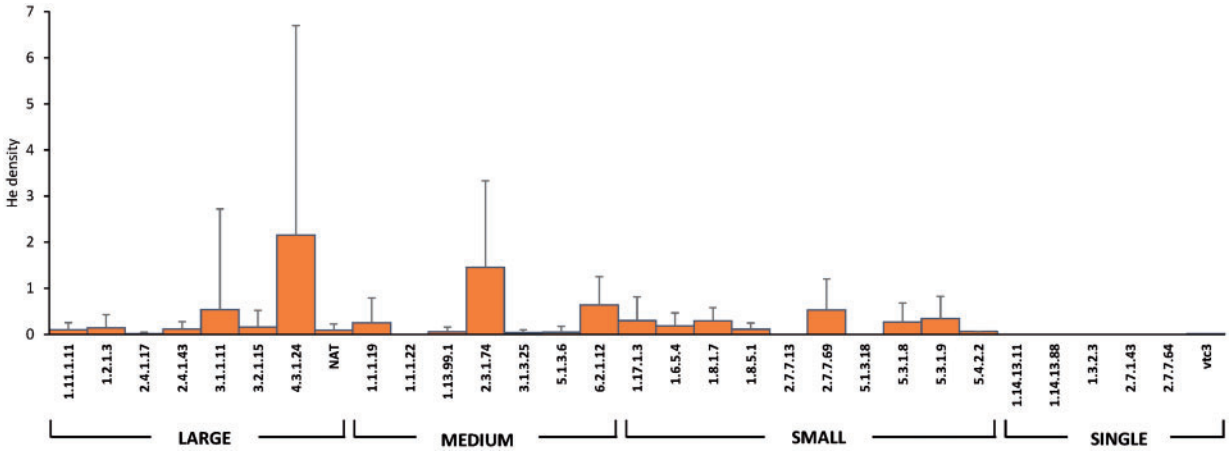
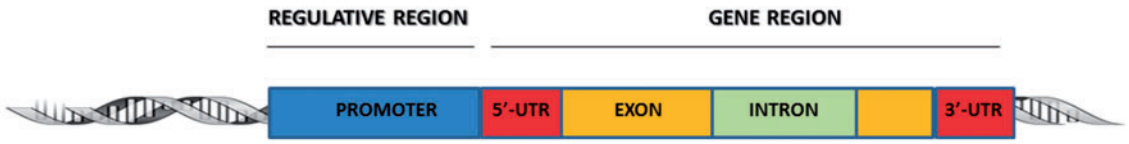


Figure 3. Average He variants density (He density) for each gene family (reported as EC number) in each category (LARGE, MEDIUM, SMALL, SINGLE). He density is expressed as the average number of He variants for 10 Kbp. Bar errors represent SD of the mean.



• **Total captured region (bp)**	972,923	23,388	575,336	791,726	48,357
• **Average length per locus (bp)**	2,567	61	1,518	2,088	127
• **No Variant sites**	7,674	104	1,309	4,460	225
• **No variants per Kbp**	7.88	5.60	2.27	5.63	4.65
• **Variant Type:**					
- No SNP	6,335	67	1,255	3,658	180
- Ts	4,148	37	808	2,315	112
- Tv	2,151	30	447	1,343	68
- No INDEL	1,339	37	54	802	45
- INS	589	13	27	324	18
- DEL	750	24	27	478	27

Figure 4. Distribution and classification of the discovered variants according to type (transition vs transversion and insertions vs deletion) and according to the genetic feature they were found in (promoter, exon, intron, UTRs).

Table 2. Distribution of 6,098 genic variants *per* type of predicted effect (PREDICTED EFFECT) on the relative protein as predicted by SNPeff.

PREDICTED EFFECT	VARIANTS No
Modifier	**4,765**
3′ UTR variant	256
5′ UTR variant	155
Intron variant	4,354
Low	**729**
5′ UTR premature start codon gain variant	13
Splice region variant & intron variant	115
Synonymous_variant	601
Moderate	**576**
Missense variant	559
Inframe deletion	7
Inframe insertion	10
High	**28**
Frameshift_variant	14
Splice acceptor variant	2
Stop_gained variant	9
Stop_lost variant	3
Total	**6,098**

were variants in coding regions that did not change the amino acid sequence (synonymous variant). About 9% showed a 'moderate' effect. These variants are predominantly non-synonymous amino acid changes (missense variant) and in few cases in-frame deletion/insertion. They are the most likely candidates for causal mutations, since they could alter the structure and function of relevant proteins. Last, about 1% of the variants were predicted to have a 'high effect'. In particular, 14 variants were predicted to cause a frameshift: two of them a splice site acceptor modification, and nine and three a start or stop codon gain or loss, respectively. On average, we observed that the percentage of INDELs with 'high effect' was higher than for SNPs, since an INDEL may rapidly cause a frameshift in the sequence. The list of the variant with a moderate or a high effect with respect to the gene they affected is provided in Supplementary Table S4.

Thus, a larger effort would be warranted to study potential links between the identified variants and trait variation in tomato, and to determine how they affect the regulation of biological pathways and processes. The variants selected or prioritized in this way would be highly preferred marker sets to be subjected to association studies using suitably larger populations, for which panels with considerable genotype and phenotype information has already been collected.[30,32]

Acknowledgements

The authors wish to thank the Genomix4Life S.r.l (http://www.genomix4life.com) for the genotyping analyses performed with ILLUMINA Infinium Technology.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by the Italian Ministry of University and Research (MIUR) (grant GenoPOMpro PON02_00395_3082360). The funders had no role in the study design, data collection or analysis, the decision to publish, or in the preparation of the article.

References

1. Edwards, D. and Gupta, P. K. 2013, Sequence based DNA markers and genotyping for cereal genomics and breeding, In: K.P., Gupta and K.R.,

Varshney (eds), *Cereal Genomics II*, Netherlands, Dordrecht: Springer, pp. 57–76.
2. Ray, S. and Satya, P. 2014, Next generation sequencing technologies for next generation plant breeding, *Front. Plant. Sci.*, **5**, 367.
3. Nicolaï, M., Cantet, M., Lefebvre, V., Sage-Palloix, A.-M. and Palloix, A. 2013, Genotyping a large collection of pepper (Capsicum spp.) with SSR loci brings new evidence for the wild origin of cultivated C. annuum and the structuring of genetic diversity by human selection of cultivar types, *Genet. Resour. Crop Evol.*, **60**, 2375–90.
4. Ranc, N., Munos, S., Santoni, S. and Causse, M. 2008, A clarified position for Solanum lycopersicum var. cerasiforme in the evolutionary history of tomatoes (solanaceae), *BMC Plant Biol.*, **8**, 130.
5. Blanca, J., Montero-Pau, J., Sauvage, C., et al. 2015, Genomic variation in tomato, from wild ancestors to contemporary breeding accessions, *BMC Genomics*, **16**, 257.
6. Chen, H., He, H., Zhou, F., Yu, H. and Deng, X. W. 2013, Development of genomics-based genotyping platforms and their applications in rice breeding, *Curr. Opin. Plant Biol.*, **16**, 247–54.
7. Filippi, C. V., Aguirre, N., Rivas, J. G., et al. 2015, Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers, *BMC Plant Biol.*, **15**, 1–12.
8. Shirasawa, K., Fukuoka, H., Matsunaga, H., et al. 2013, Genome-wide association studies using single nucleotide polymorphism markers developed by re-sequencing of the genomes of cultivated tomato, *DNA Res.*, **20**, 593–603.
9. Sim, S. C., Durstewitz, G., Plieske, J., et al. 2012, Development of a large SNP genotyping array and generation of high-density genetic maps in tomato, *PLoS One*, **7**, e40563.
10. Causse, M., Desplat, N., Pascual, L., et al. 2013, Whole genome resequencing in tomato reveals variation associated with introgression and breeding events, *BMC Genomics*, **14**, 1–14.
11. Ercolano, M. R., Sacco, A., Ferriello, F., et al. 2014, Patchwork sequencing of tomato San Marzano and Vesuviano varieties highlights genome-wide variations, *BMC Genomics*, **15**, 138.
12. Lam, H. Y. K., Clark, M. J., Chen, R., et al. 2012, Performance comparison of whole-genome sequencing platforms, *Nat. Biotech.*, **30**, 78–82.
13. Schneeberger, K., Ossowski, S., Ott, F., et al. 2011, Reference-guided assembly of four diverse Arabidopsis thaliana genomes, *Proc. Natl. Acad. Sci. U S A*, **108**, 10249–54.
14. Albert, T. J., Molla, M. N., Muzny, D. M., et al. 2007, Direct selection of human genomic loci by microarray hybridization, *Nat. Methods*, **4**, 903–5.
15. Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J. and Zwick, M. E. 2007, Microarray-based genomic selection for high-throughput resequencing, *Nat. Methods*, **4**, 907–9.
16. Olson, M. 2007, Enrichment of super-sized resequencing targets from the human genome, *Nat. Methods*, **4**, 891–2.
17. de Sousa, F., Bertrand, Y. J., Nylinder, S., Oxelman, B., Eriksson, J. S. and Pfeil, B. E. 2014, Phylogenetic properties of 50 nuclear loci in Medicago (Leguminosae) generated using multiplexed sequence capture and next-generation sequencing, *PLoS One*, **9**, e109704.
18. Rabbani, B., Tekin, M. and Mahdieh, N. 2014, The promise of whole-exome sequencing in medical genetics, *J. Hum. Genet.*, **59**, 5–15.
19. Muraya, M. M., Schmutzer, T., Ulpinnis, C., Scholz, U. and Altmann, T. 2015, Targeted sequencing reveals large-scale sequence polymorphism in maize candidate genes for biomass production and composition, *PLoS One*, **10**, e0132120.
20. Ashman, T. L., Tennessen, J. A., Dalton, R. M., Govindarajulu, R., Koski, M. H. and Liston, A. 2015, Multilocus sex determination revealed in two populations of gynodioecious wild strawberry, Fragaria vesca subsp. bracteata, *G3 (Bethesda)*, **5**, 2759–73.
21. Clarke, W. E., Parkin, I. A., Gajardo, H. A., et al. 2013, Genomic DNA enrichment using sequence capture microarrays: a novel approach to discover sequence nucleotide polymorphisms (SNP) in Brassica napus L, *PLoS One*, **8**, e81992.
22. Schiessl, S., Samans, B., Huttel, B., Reinhard, R. and Snowdon, R. J. 2014, Capturing sequence variation among flowering-time regulatory gene homologs in the allopolyploid crop species Brassica napus, *Front. Plant Sci.*, **5**, 404.
23. Zhou, L. and Holliday, J. A. 2012, Targeted enrichment of the black cottonwood (Populus trichocarpa) gene space using sequence capture, *BMC Genomics*, **13**, 1–12.
24. Uitdewilligen, J. G., Wolters, A. M., D'Hoop B. B., Borm, T. J., Visser, R. G. and van Eck, H. J. 2013, A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato, *PLoS One*, **8**, e62355.
25. Gardiner, L. J., Gawronski, P., Olohan, L., Schnurbusch, T., Hall, N. and Hall, A. 2014, Using genic sequence capture in combination with a syntenic pseudo genome to map a deletion mutant in a wheat species, *Plant J.*, **80**, 895–904.
26. Winfield, M. O., Wilkinson, P. A., Allen, A. M., et al. 2012, Targeted re-sequencing of the allohexaploid wheat exome, *Plant Biotechnol. J.*, **10**, 733–42.
27. Pootakham, W., Shearman, J. R., Ruang-Areerate, P., et al. 2014, Large-scale SNP discovery through RNA sequencing and SNP genotyping by targeted enrichment sequencing in cassava (Manihot esculenta Crantz), *PLoS One*, **9**, e116028.
28. Sanseverino, W., Henaff, E., Vives, C., et al. 2015, Transposon Insertions, Structural Variations, and SNPs Contribute to the Evolution of the Melon Genome, *Mol. Biol. Evol.*, **32**, 2760–74.
29. Fuentes Fajardo, K. V., Adams, D., Program, N. C. S., et al. 2012, Detecting false-positive signals in exome sequencing, *Hum. Mutat.*, **33**, 609–13.
30. Sacco, A., Ruggieri, V., Parisi, M., et al. 2015, Exploring a tomato landraces collection for fruit-related traits by the aid of a high-throughput genomic platform, *PLoS One*, **10**, e0137139.
31. Ruggieri, V., Bostan, H., Barone, A., Frusciante, L. and Chiusano, M. L. 2016, Integrated bioinformatics to decipher the ascorbic acid metabolic network in tomato, *Plant. Mol. Biol.*, **91** (4–5), 397–412.
32. Ruggieri, V., Francese, G., Sacco, A., et al. 2014, An association mapping approach to identify favourable alleles for tomato fruit quality breeding, *BMC Plant Biol*, **14**, 337.
33. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754–60.
34. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
35. Layer, R. M., Chiang, C., Quinlan, A. R. and Hall, I. M. 2014, LUMPY: a probabilistic framework for structural variant discovery, *Genome Biol.*, **15**, R84.
36. Cingolani, P., Patel, V. M., Coon, M., et al. 2012, Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift, *Front. Genet.*, **3**, 35.
37. Li, H. 2011, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics*, **27**, 2987–93.
38. Quinlan, A. R. and Hall, I. M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**.
39. Bolger, A. M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
40. Shearer, L. A., Anderson, L. K., de Jong, H., et al. 2014, Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome, *G3 (Bethesda)*, **4**, 1395–1405.
41. Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y. and Buckler, E. S. 2007, TASSEL: software for association mapping of complex traits in diverse samples, *Bioinformatics*, **23**, 2633–35.
42. Lin, T., Zhu, G., Zhang, J., et al. 2014, Genomic analyses provide insights into the history of tomato breeding, *Nat. Genet.*, **46**, 1220–6.
43. Danecek, P., Auton, A., Abecasis, G., et al. 2011, The variant call format and VCFtools, *Bioinformatics*, **27**, 2156–8.
44. Bodi, K., Perera, A. G., Adams, P. S., et al. 2013, Comparison of commercially available target enrichment methods for next-generation sequencing, *J. Biomol. Tech.*, **24**, 73–86.
45. Sims, D., Sudbery, I., Ilott, N. E., Heger, A. and Ponting, C. P. 2014, Sequencing depth and coverage: key considerations in genomic analyses, *Nat. Rev. Genet.*, **15**, 121–32.

46. Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., et al. 2008, Accurate whole human genome sequencing using reversible terminator chemistry, *Nature*, **456**, 53–9.
47. Cartwright, R. A. 2009, Problems and solutions for estimating indel rates and length distributions, *Mol. Biol. Evol*., **26**, 473–80.
48. Lunter, G. 2007, Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes, *Bioinformatics*, **23**, i289–96.
49. Luo, C., Tsementzi, D., Kyrpides, N., Read, T. and Konstantinidis, K. T. 2012, Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample, *PLoS One*, **7**, e30087.
50. Wang, S., Yang, X., Xu, M., et al. 2015, A rare SNP identified a TCP transcription factor essential for tendril development in cucumber, *Mol. Plant*, **8**, 1795–808.
51. Bang, S. Y., Na, Y. J., Kim, K., et al. 2014, Targeted exon sequencing fails to identify rare coding variants with large effect in rheumatoid arthritis, *Arthritis Res. Ther*., **16**, 447.
52. Belo, A., Beatty, M. K., Hondred, D., Fengler, K. A., Li, B. and Rafalski, A. 2010, Allelic genome structural variations in maize detected by array comparative genome hybridization, *Theor. Appl. Genet*., **120**, 355–67.
53. Springer, N. M., Ying, K., Fu, Y., et al. 2009, Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content, *PLoS Genet*, **5**, e1000734.
54. Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., et al. 2010, Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor, *Genome Res*., **20**, 1689–99.
55. Quail, M. A., Kozarewa, I., Smith, F., et al. 2008, A large genome center's improvements to the Illumina sequencing system, *Nat. Methods*, **5**, 1005–10.
56. Landan, G. and Graur, D. 2007, Heads or tails: a simple reliability check for multiple sequence alignments, *Mol. Biol. Evol*., **24**, 1380–3.
57. Blankenberg, D., Von Kuster, G., Coraor, N., et al. 2010, Galaxy: a web-based genome analysis tool for experimentalists, *Curr. Protoc. Mol. Biol*., Chapter 19, Unit 19.10.11-21.