9 Resequencing

Pietro Gramazio, Santiago Vilanova and Jaime Prohens

Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV), Universitat Politècnica de València, Camino de Vera 14, 46022 Valencia, Spain

E-mail: jprohens@btc.upv.es

Abstract

The next-generation sequencing revolution is allowing the whole-genome resequencing (WGRS) of hundreds or even thousands of accessions for staple crops and model species. With the release of their reference genome, progressively also other plants species are undertaking WGRS projects for a broad variety of studies. In common eggplant (*Solanum melongena*), although a first draft of the reference genome sequence has been published, no resequencing studies have been performed so far. In this chapter, we present first results of the resequencing of eight accessions, seven of common eggplant and one of the wild relative *S. incanum*, that correspond to the parents of a MAGIC population that is currently under development using the newly developed eggplant genome sequence presented in chapter 7 of this book. Over 10 million polymorphisms were identified among the accessions, 90% of them in the wild related *S. incanum*, confirming the genetic erosion of the cultivated common eggplant. Among the MAGIC population parents, the common polymorphisms distribution pattern along the chromosomes has revealed possible footprints of ancestral introgressions from interspecific crosses. The set of polymorphisms has been extensively annotated and currently is being used for further analyses in order to efficiently genotype the ongoing MAGIC population and to dissect important agronomic and morphological traits. The information provided in this first resequencing study in eggplant will be extremely helpful to assist plant breeding to develop new improved and resilient varieties to face future threats and challenges.

Keywords

Resequencing, *Solanum melongena*, *Solanum incanum*, MAGIC population, polymorphisms

Introduction

The release of the first drafts and complete genomes of the most important cultivated crops has represented a scientific milestone in plant biology. For the first time ever, genome sequences have provided a vast amount of information for a comprehensive analysis of the genome structures, genes and repetitive elements, among others (Huang et al., 2013). However, the information retrieved from the reference genome sequence is not sufficient to provide a comprehensive picture of the structural and allelic variation of a species or of its related genepool materials (Schatz et al., 2014). The continuing improvements in sequencing technologies coupled with the significant decrease of the sequencing costs have opened the way to the whole-genome resequencing (WGRS) of hundreds of cultivated accessions and wild relatives for model crops and the most important economically cultivated crops such as *Arabidopsis thaliana*, tomato, rice, soybean, or cotton, among others (Weigel and Mott, 2009; Xu et al., 2012; Aflitos et al., 2014; Zhou et al., 2015; Du et al., 2018). In fact, WGRS can extremely speed up the challenging task of reconstructing the "pan-genome puzzle" of a species through the identification of global polymorphisms and gene variations, genomic structural variations, gene copy number, or copy number variation (Jha et al., 2016). With WGRS, the natural variation of a crop can be can be easily captured through the identification of millions of robust polymorphisms among accessions, allowing to perform forward genetics techniques and genome-wide association studies, and thus unravelling the genetic base of complex traits of agronomic

importance (Ogura and Busch, 2015). WGRS can also shed light on the history of a crop and identify the genetic diversity bottlenecks occurred during the domestication and the genes that are associated with this process. For example, Zhou et al. (2015) were able to detect 230 selective sweeps and 162 selected copy number variants associated with ten genomic regions and nine domestication traits in addition to the identification of 13 previously uncharacterized loci for agronomic traits in soybean including oil content, plant height and pubescence form. Moreover, using this approach it is possible to associate genes and traits with geographical areas revealing how populations and subpopulations have adapted to a specific geographic area (Qi et al., 2013). Ultimately, the reconstruction and identification of the different stages of breeding history and artificial selection by WGRS provide new and most efficient tools and strategies for future crop breeding and biotechnology (Jiao et al., 2012).

Resequencing in eggplant

To our knowledge, no resequencing studies have been published in eggplant (*Solanum melongena* L.) so far. In fact, despite the economic importance of this crop, which ranks fifth among vegetables in total worldwide production (Faostat, 2016), and its role to guarantee food security in tropical and subtropical regions, few genomic studies have been performed in eggplant and its wild relatives (Gramazio et al., 2018). The disparity between eggplant and other important cultivated crops for genomic data and information is still large, although some efforts are being done to narrow the gap. In comparison, in tomato several resequencing studies have been published, including the resequencing of 360 cultivated and wild relatives accessions representing several geographical origins, consumption types and improvement statuses (Lin et al., 2014), 84 tomato accessions and related wild species to explore genetic variation (Aflitos et al., 2014), experimental populations (Causse et al., 2013; Kevei et al., 2015, Zhang et al., 2018), elite cultivars (Kobayashi et al., 2014; Jung et al., 2016), mutants (Shirasawa et al., 2016), abiotic stress tolerance (Tranchida-Lombardo et al., 2018), among others.

The lack of resequencing studies in eggplant might be, in part, due to the unavailability of a high-quality reference genome sequence and the corresponding annotation. Up to now, the eggplant research community can rely on just a draft eggplant genome published in 2014, which is fragmented in 33,873 scaffolds covering 833.1 Mb (ca. 74% of the eggplant genome) and where 85,446 genes were predicted (Hirakawa et al., 2014). Thus, mapping a WGRS dataset onto this eggplant genome sequence could lead to a loss of valuable information about the target of the study. A new high-quality eggplant genome sequence, obtained by the "Italian Eggplant Genome Sequencing Consortium" (http://www.eggplantgenome.org/), has been presented and will soon be released (see chapter 7 in this book, Barchi et al., 2016). Based on the statistics presented, this new reference genome is much less fragmented and the number of genes annotated is about 35,000, very similar to those described in tomato (Tomato Genome Consortium, 2012). The imminent availability of this high-quality reference genome will foster genomic studies as it occurred in other cultivated species. In fact, our group, which had access to this improved genome sequence thanks to fruitful collaborations with the members of the "Italian Eggplant Genome Sequencing Consortium", took the opportunity to use this valuable information to assist several research lines, including a re-sequencing study. Among them, we have performed the WGRS of eight accessions that correspond to the parents of a MAGIC (Multi-parent advanced generation inter-cross) population that we are currently developing. Seven out of the eight parents correspond to common eggplants (*S. melongena*) from different geographic areas. These accessions are phenotypically very diverse, showing substantial differences in fruit size, fruit shape, fruit color, calyx prickliness, and many other agronomic and morphological traits. The eighth parent is a *S. incanum* accession, a wild species from the secondary genepool of common eggplant (Syfert et al., 2016). *Solanum incanum* is very interesting for eggplant breeding since has been reported as a powerful source of phenolic compounds, showing contents several times higher than common eggplant (Stommel and Whitaker, 2003; Prohens et al., 2013), and is tolerant to some abiotic and abiotic stresses, mainly drought

(Knapp et al., 2013). The specific *S. incanum* accession (MM577) used for this WGRS has been extensively characterized in other studies for several traits (Stommel and Whitaker, 2003; Gisbert et al., 2011; Salas et al. 2011; Meyer et al. 2015). In addition, MM577 and one of the seven MAGIC parent, AN-S-26, have been used to build an interspecific genetic linkage map to locate the candidate genes involved in the chlorogenic acid biosynthesis pathway and other candidate genes of agronomic interests, as well as, the candidate genes involved in the fruit flesh browning (Gramazio et al., 2014). Subsequently, this mapping population was used to develop the first introgression line population in eggplant genepool (Gramazio et al., 2017).

The main goal of this WGRS project was to provide a large set of molecular markers among the eight founder lines to efficiently assist the genotyping of the first MAGIC population in eggplant, as well as, to dissect the genetic base of complex traits of agronomic importance in eggplant, detect potential introgressions associated to domestication and geographical areas and ultimately provide tools to clarify the eggplant evolutionary history and enhance eggplant breeding.

High-throughput Sequencing and mapping

To generate a large set of high-resolution SNPs distributed throughout the genome, a whole-genome sequencing approach was adopted. After preparing the Illumina paired-end libraries of 300 bp, the MAGIC parents were sequenced on two lines of an Illumina HiSeq 4000 sequencer. The sequencing produced over 100 Gb of data with a range of 150 to 220 million raw reads per sample (Figure 1). Less than 3% of the raw reads were discarded after the trimming and cleaning process and the remaining clean reads were mapped onto the high-quality reference genome (Barchi et al., 2016). Over 80% of the reads were successfully mapped with an overall coverage of around 20x. Thanks to the newly improved Illumina platforms, it becomes affordable to have a good coverage also for genomes of medium size as *S. melongena* (around 1.2 Gb). In fact, just a few years ago, the most common mapping coverage in WGRS projects was around 10x or less, including for small/medium genomes as rice, tomato or barley (Causse et al., 2013; Zhou et al., 2015, Xu et al., 2012). Although single molecule real time sequencing (SMRT) platforms are becoming more popular for de novo whole-genome sequencing, the Illumina platform is by far the most frequently selected sequencing technology for WGRS studies, especially for high-quality and completed genomes, since in this scenario the short read length is not a limitation and the higher throughput compared to other technologies is preferred. New Illumina platforms, like NovaSeq 6000 System, can give an impressive sequencing output up to 6Tb of data that correspond to around 20 billion paired-end reads and thus may further decrease sequencing costs, which may foster resequencing studies in eggplant.

Variant calling, distribution and annotation

Over 10 million polymorphisms were identified among the eight MAGIC parents resequenced, most of which were SNPs. While among the *S. melongena* accessions the variants identified were around one million per accession, for the *S. incanum* accession the number of variants was over 9 million (Figure 2). This large difference in polymorphisms between cultivated and wild relative species is quite common for the most economically important and staple crops, where artificial selection for important breeding traits and the seeking to uniformity for commercial varieties have dramatically increased their genetic erosion (Aflitos et al., 2014; Zhou et al., 2015). Before the advent of the next-generation sequencing era, the development of reliable molecular markers was not an easy and inexpensive task. In consequence, many crops, in particular no-model crops, had been neglected from research studies and molecular-marker assisted (MAS) selection. Common eggplant was one of them and just a few years ago the gap with other economic important crops has been narrowed thanks to the first genomic studies performed (Gramazio et al., 2018).

The variants detected were divided into 10 Mbp-sized bins in order to identify similar patterns of polymorphisms distribution and associate them with potential common ancestral introgressions. Figure 3 shows an example of the distribution of homozygous SNPs for the chromosome 6. It is very clear that the common eggplant accessions C, D, E and G presented a similar SNP distribution pattern from the beginning of the chromosome 6 to about 25 Mbp and then until 60 Mbp the accessions C and D shared other common peaks while the accessions E and G did not present the same ones. This similar SNPs distribution represented by these peaks may be a footprint of an old interspecific introgression from a common eggplant relative. Then the accessions C and D, which are from the same geographical area and probably shared a recent ancestor, could have incorporated an additional introgression resulted from another hybridization event. An alternative hypothesis could be that the accessions E and G might have lost part of the introgression during the domestication events.

In addition, the variants were annotated and classified by impact (high, low, moderate or modifier), by functional class (missense, nonsense or silent mutation), by the type (start lost, stop gained, stop lost, and others) and region affected (intergenic, intron, exon, and others), as well as, DNA substitution mutations (transitions and transversion) and amino acids changes. At the time this chapter is being written many analyses are being performed using the information generated in this WGRS study like repetitive elements, copy number variations (CNVs), relationship analyses among the accessions, or search for candidate genes of important agronomic traits.

Conclusions

The combination of the decreasing cost of sequencing and the availability of high-quality sequencing genomes are boosting resequencing studies even for no-model plant species, including eggplant. Although for model plants like *Arabidopsis thaliana* (Weigel and Mott, 2009) or important staple crops like rice (Guo et al., 2014) thousands of accessions have been resequenced during the last decade, the first resequencing studies in other species of scientific or economic interest are, little by little, being published. The potential of resequencing to interrogate the whole genome of eggplant and identify structural and functional variation among accessions makes it a great powerful analysis and inquiry strength. Furthermore, its high versatility of approaches and strategies allows answering many scientific and technical questions, including allele and variants discovery, germplasm genomic characterization, domestication history, or dissecting agronomic-associated loci for plant breeding, among others. The first resequencing efforts performed in eggplant can boost the gathering of genomic data from the germplasm of this species and wild relatives, which may be pivotal to develop a new generation of improved eggplant varieties adapted to present and future challenges in eggplant production and fruit quality.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No 677379 (G2P-SOL project: Linking genetic resources, genomes and phenotypes of Solanaceous crops) and from Spanish Ministerio de Economía, Industria y Competitividad and Fondo Europeo de Desarrollo Regional (grant AGL2015-64755-R from MINECO/FEDER).

References

Aflitos S, Schijlen E, De Jong H et al (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J 80:136–148. doi:10.1111/tpj.12616

Barchi L, Delledonne M, Lanteri S et al (2016). A high quality eggplant genome sequence: a new tool for the analysis of the Solanaceae family evolution and for the molecular deciphering of complex traits. Paper presented at 20th EUCARPIA General Congress Plant Breeding: The Art of Bringing Science to Life, Zurich, 29 August - 1 September 2016

Causse M, Desplat N, Pascual L et al (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. BMC Genomics 14:791. doi:10.1186/1471-2164-14-791

Du X, Huang G, He S et al (2018) Resequencing of 243 diploid cotton accessions based on an updated A genome identifies the genetic basis of key agronomic traits. Nat Genet 50:796–802. doi:10.1038/s41588-018-0116-x

Faostat (2016) http://www.fao.org/faostat. Accessed 23 July 2018

Gisbert C, Prohens J, Raigón MD et al (2011) Eggplant relatives as sources of variation for developing new rootstocks: Effects of grafting on eggplant yield and fruit apparent quality and composition. Sci Hort 128:14-22

Gramazio P, Prohens J, Plazas M et al (2014) Location of chlorogenic acid biosynthesis pathway and polyphenol oxidase genes in a new interspecific anchored linkage map of eggplant. BMC Plant Biol 14:350. doi:10.1186/s12870-014-0350-z

Gramazio P, Prohens J, Plazas M et al (2017) Development and genetic characterization of advanced backcross materials and an introgression line population of *Solanum incanum* in a *S. Melongena* background. Front Plant Sci. doi:10.3389/fpls.2017.01477

Gramazio P, Prohens J, Plazas M et al (2018) Genomic tools for the enhancement of vegetable crops: A case in eggplant. Not Bot Horti Agrobot Cluj-Napoca. doi:10.15835/nbha46110936

Guo L, Gao Z, Qian Q (2014) Application of resequencing to rice genomics, functional genomics and evolutionary analysis. Rice 7:4. doi:10.1186/s12284-014-0004-7

Hirakawa H, Shirasawa K, Miyatake K et al (2014) Draft genome sequence of eggplant (*Solanum melongena* L.): The representative *Solanum* species indigenous to the old world. DNA Res 21:649–660

Huang X, Lu T, Han B (2013) Resequencing rice genomes: An emerging new era of rice genomics. Trends Genet 29:225–232

Jha U, Barh D, Parida S et al (2016) Whole-genome resequencing: Current status and future prospects in genomics-assisted crop improvement. In: Khan MS, Khan IA, Barh D (ed) Applied Molecular Biotechnology: The Next Generation of Genetic Engineering, CRC Press, Boca Raton, p 187–211

Jiao Y, Zhao H, Ren L et al (2012) Genome-wide genetic changes during modern breeding of maize. Nat Genet 44:812–815. doi:10.1038/ng.2312

Jung YJ, Nou IS, Cho YG, et al (2016) Identification of an SNP variation of elite tomato (*Solanum lycopersicum* L.) lines using genome resequencing analysis. Hortic Environ Biotechnol 57:173–181

Kevei Z, King RC, Mohareb F et al (2015) Resequencing at ≥40-fold depth of the parental genomes of a *Solanum lycopersicum* × S. *pimpinellifolium* recombinant inbred line population and characterization of frame-shift Indels that are highly likely to perturb protein function. G3: Genes, Genomes, Genetics. doi:10.1534/g3.114.016121

Knapp S, Vorontsova MS, Prohens J (2013) Wild relatives of the eggplant (*Solanum melongena* L.: Solanaceae): new understanding of species names in a complex group. PLoS One 8:e57039

Kobayashi M, Nagasaki H, Garcia V et al (2014) Genome-wide analysis of intraspecific DNA polymorphism in 'micro-tom', a model cultivar of tomato (*Solanum lycopersicum*). Plant Cell Physiol 55:445–454.

Lin T, Zhu G, Zhang J et al (2014) Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46:1220–1226. doi:10.1038/ng.3117

Meyer RS, Whitaker BD, Little DP et al (2015) Parallel reductions in phenolic constituents resulting from the domestication of eggplant. Phytochemistry 115:194–206. doi:10.1016/j.phytochem.2015.02.006

Ogura T, Busch W (2015) From phenotypes to causal sequences: using genome wide association studies to dissect the sequence basis for variation of plant development. Curr Opin Plant Biol. 23:98–108

Prohens J, Whitaker BD, Plazas M et al (2013) Genetic diversity in morphological characters and phenolic acids content resulting from an interspecific cross between eggplant, *Solanum melongena*, and its wild ancestor (*S. incanum*). Ann Appl Biol 162:242–257. doi: 10.1111/aab.12017

Qi J, Liu X, Shen D et al (2013) A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. Nat Genet 45:1510–1515. doi:10.1038/ng.2801

Salas P, Prohens J, Seguí-Simarro JM (2011) Evaluation of androgenic competence through anther culture in common eggplant and related species. Euphytica 182:261–274. doi:10.1007/s10681-011-0490-2

Schatz MC, Maron LG, Stein JC et al (2014) Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. Genome Biol 15:506. doi: 10.1186/s13059-014-0506-z

Shirasawa K, Hirakawa H, Nunome T et al (2016) Genome-wide survey of artificial mutations induced by ethyl methanesulfonate and gamma rays in tomato. Plant Biotechnol J 14:51–60. doi: 10.1111/pbi.12348

Stommel JR, Whitaker BD (2003) Phenolic Acid Content and composition of eggplant fruit in a germplasm core subset. J Amer Soc Hort Sci 128:704–710. doi:10.1002/mnfr.200600067

Syfert MM, Castañeda-Álvarez NP, Khoury CK et al (2016) Crop wild relatives of the brinjal eggplant (*Solanum melongena*): Poorly represented in genebanks and many species at risk of extinction. Am J Bot 103: 635–651

Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–41. doi: 10.1038/nature11119Tranchida-Lombardo V, Aiese Cigliano R, Anzar I et al (2018) 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 Research 25:149–160

Tranchida-Lombardo V, Aiese Cigliano R, Anzar I et al (2018) 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 Research 25:149–160

Weigel D, Mott R (2009) The 1001 Genomes Project for Arabidopsis thaliana. Genome Biol 10:107. doi: 10.1186/gb-2009-10-5-107

Xu X, Liu X, Ge S et al (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105–111. doi: 10.1038/nbt.2050

Zhang S, Yu H, Wang K et al (2018) Detection of major loci associated with the variation of 18 important agronomic traits between *Solanum pimpinellifolium* and cultivated tomatoes. Plant J 95:312–323. doi: 10.1111/tpj.13952

Zhou Z, Jiang Y, Wang Z et al (2015) Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol 33:408–414. doi: 10.1038/nbt.3096

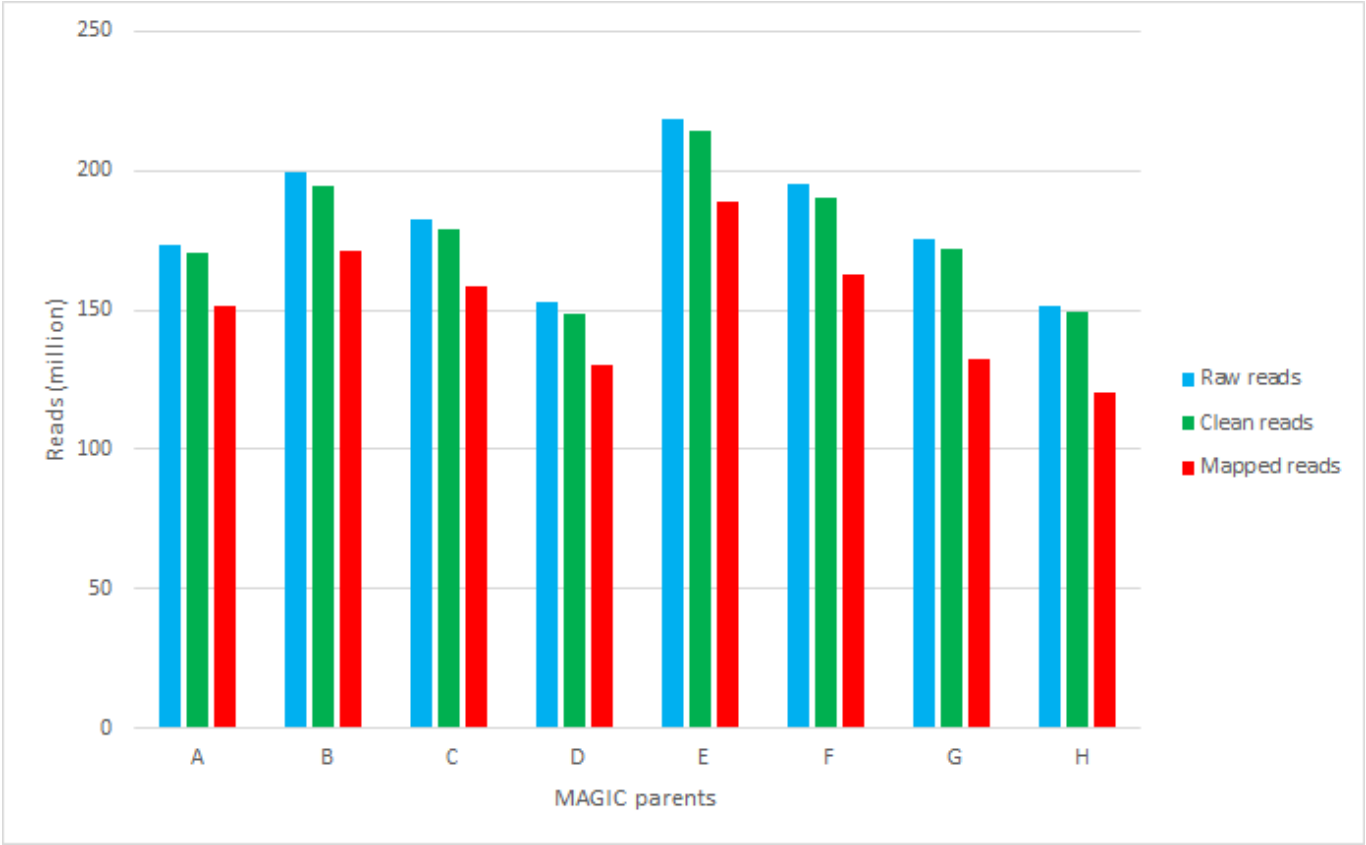

Figure 1. Statistics of the sequencing of the eight MAGIC parents and the read mapping onto the eggplant reference genome. The codes A to G correspond to *S. melongena*, while the code H to *S. incanum*.

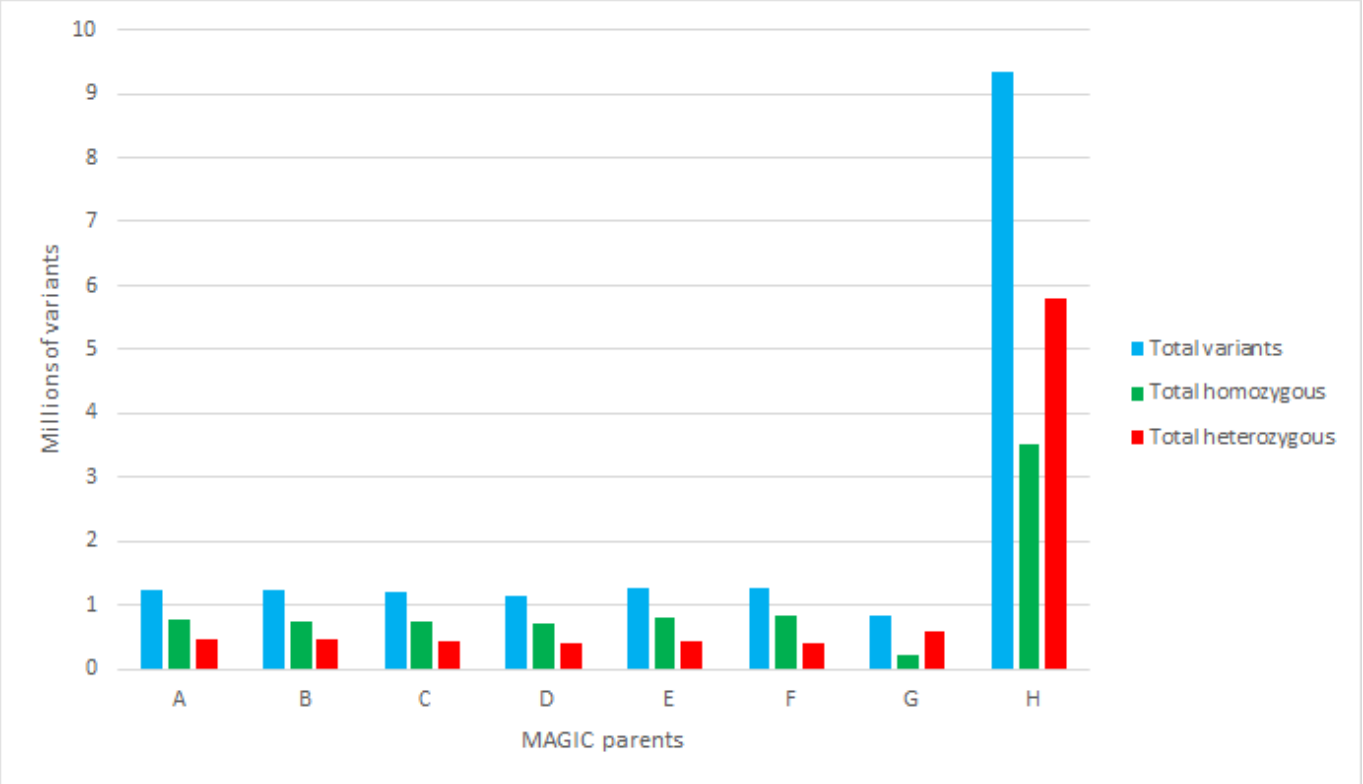

Figure 2. Statistics of the variants identified in the eight MAGIC parents a. The codes A to G correspond to *S. melongena*, while the code H to *S. incanum*.

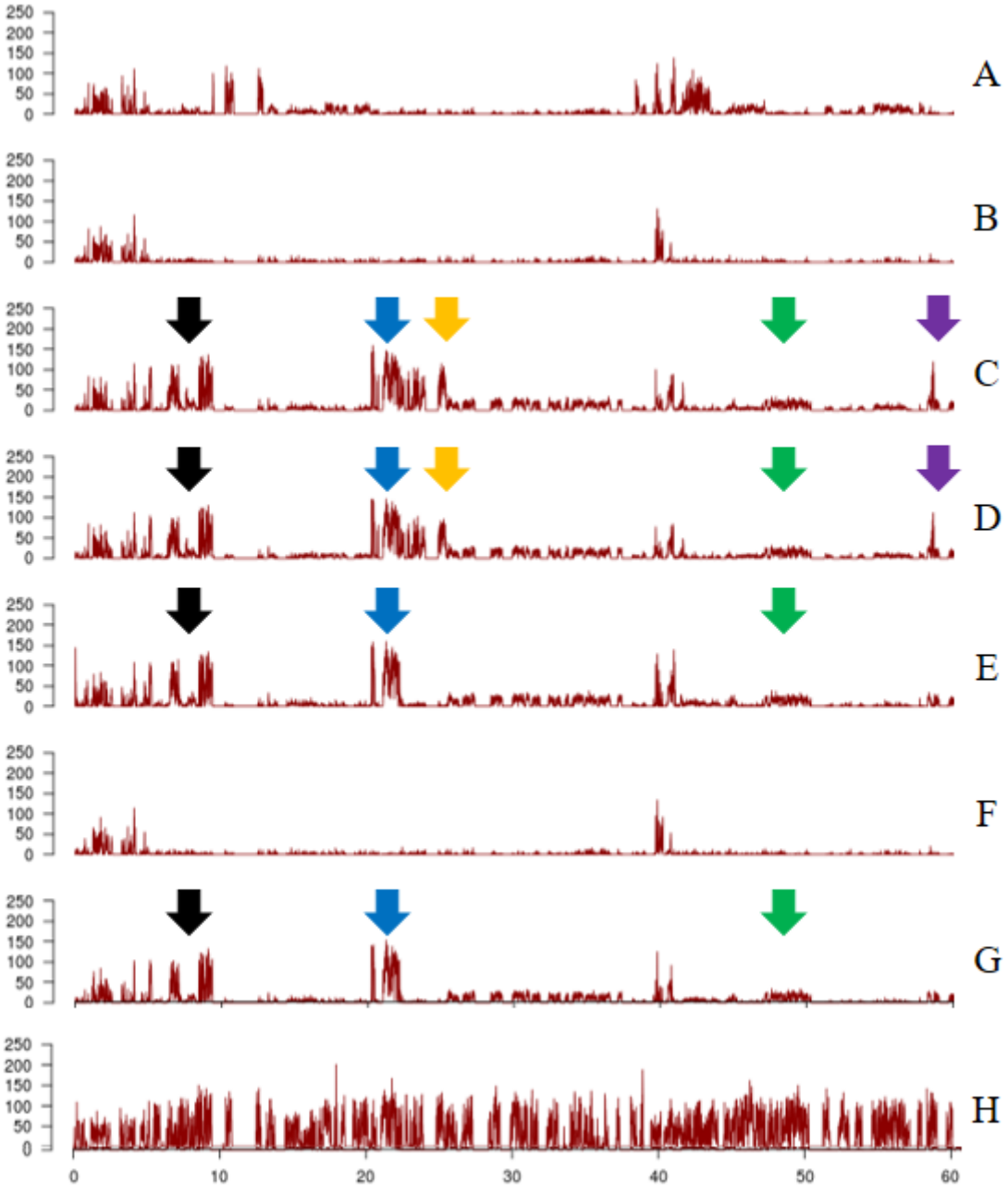

Figure 3. Distribution of homozygous variants along the first part of chromosome 6 divided into 10 Mbp-sized bins (in red). The x-axis represents the Mbp of chromosome 6 and y-axis the number of homozygous SNPs identified. The arrows of the same color indicate the similar SNP distribution pattern. The codes A to G correspond to *S. melongena*, while the code H to *S. incanum*.