Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Plant Biology

Structural variation and genome complexity: is dispensable really dispensable?

Fabio Marroni[1,3], Sara Pinosio[2,3] and Michele Morgante[1,3]

Structural variants (SVs) such as copy number variants (CNVs) and presence/absence variants (PAVs) substantially contribute to genetic variation and have an important effect on phenotypic diversity. Since unbalanced SVs are by definition sequences present only in some individuals, they have therefore been referred to as dispensable genome and are not necessary for survival, even though they may provide an important contribution to phenotypic diversity within the species. However, some multi-copy sequences of the dispensable genomes (e.g., multigene families) may be needed in a given proportion by each individual, thus belonging to a conditionally dispensable portion of the pan-genome. Another interesting aspect reported by recent studies is that the rate at which SVs are formed might be influenced by the mating system and by common environmental stresses. In conclusion the dispensable genome plays an important role in genome evolution and in the complex interplay between the genome and the environment.

Addresses
[1] Dipartimento di Scienze Agrarie e Ambientali, Università di Udine, 33100 Udine, Italy
[2] Institute of Biosciences and Bioresources, National Research Council, Via Madonna del Piano 10, 50019 Sesto Fiorentino (Firenze), Italy
[3] Istituto di Genomica Applicata (IGA), 33100 Udine, Italy

Corresponding author: Morgante, Michele (michele.morgante@uniud.it)

Current Opinion in Plant Biology 2014, **18**:31–36

This review comes from a themed issue on **Genome studies and molecular genetics**

Edited by **Kirsten Bomblies** and **Olivier Loudet**

For a complete overview see the Issue and the Editorial

Available online 16th February 2014

1369-5266/$ – see front matter,

http://dx.doi.org/10.1016/j.pbi.2014.01.003

Introduction

Genome data obtained from different plant species showed how plastic, variable and complex plant genomes are. The presence of high levels of structural variation prompted us to extend to plant species the concept of pan-genome [1], originally proposed for bacteria [2]. The pan-genome is composed of a core portion, present in all the individuals, and a dispensable portion, not present in all the individuals. The advent of next-generation sequencing (NGS) opened the possibility of re-sequencing the whole genome of several subjects [3]. This led researchers to realize the need of characterizing the genomes of plant species at a population level, since a single genome sequence is not adequate to represent all cultivars and/or populations [1,4]. Population-scale genome resequencing efforts are being performed to produce catalogues of structural variants (SVs), ultimately defining a species dispensable genome [5,6]. In the present review we will discuss methods for the detection of SVs, with special emphasis on the latest methodological developments. The dispensable portion of the genome is by definition not essential for survival, since it could be missing in at least one individual within the species. This does not equate by any mean to not having a functional relevance. Here we present the latest examples of the evolutionary and functional consequences of the SVs, representing the major components of the dispensable genome. We will then emphasize the importance of SVs (i.e., dispensable genome) for plant evolution, their possible functional role and their importance for understanding the genetic basis of phenotypic variation.

Origin and identification of SVs

It was long assumed that most of genetic variation across individuals arises from single nucleotide polymorphisms (SNPs) or small insertion/deletion polymorphisms. In recent years, the central role of structural variation has emerged [7]. SVs were originally defined as genomic alterations such as insertions, deletions, duplications, inversions and translocations covering at least 1 kb [7]. With the advent of next generation sequencing (NGS) technologies that enabled the detection of shorter alterations, the definition was adjusted to include also smaller variants [8].

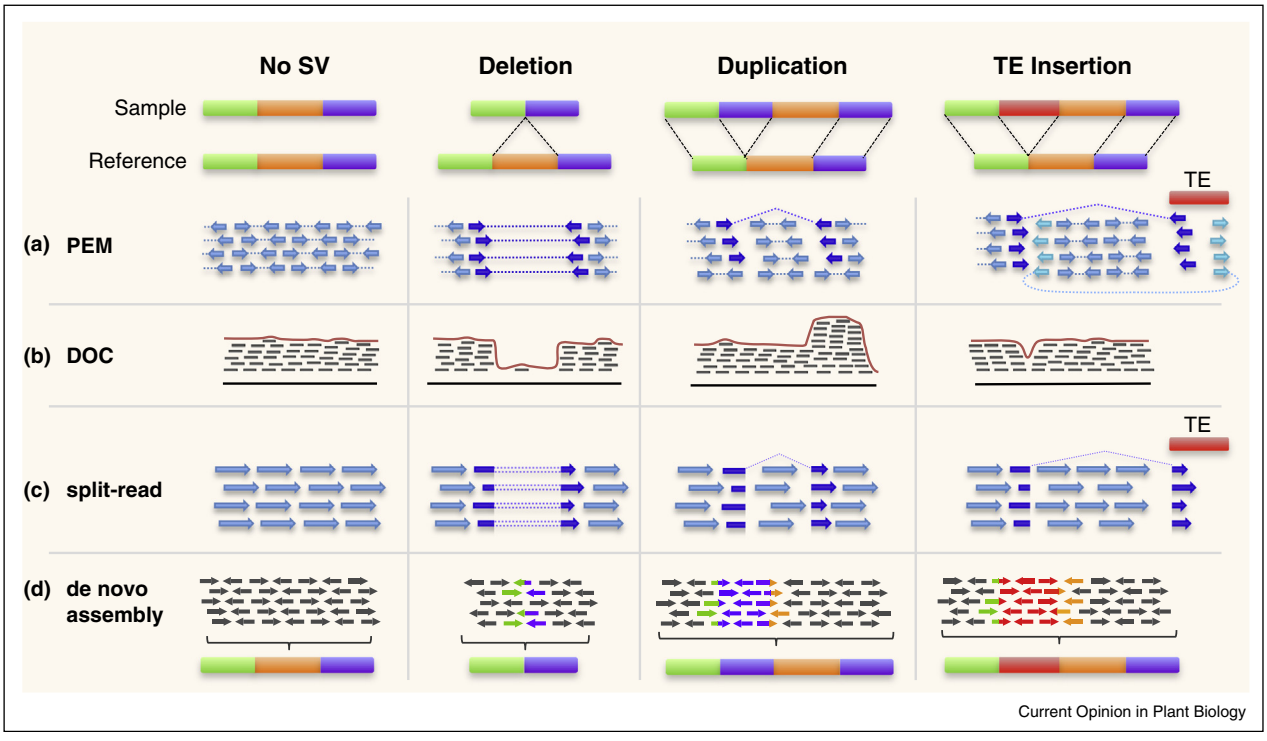
Here we will focus on two different classes of SVs that are thought to be the major contributors to phenotypic variation in plants: copy number variants (CNVs), defined as sequences that are present in different copy number among individuals and presence/absence variants (PAVs) such as sequences that are present in some individuals but completely absent in others. The contribution of CNVs and PAVs to genome diversity is significant. Intraspecific comparisons in plant species such as Arabidopsis [9••,10••], maize [11,12], soybean [13], barley [14•] and rice [15], evidenced hundreds of SVs involving several megabases of sequence.

Several mechanisms potentially lead to the formation of SVs [16]. Transposable elements (TEs) can originate insertion/deletion polymorphisms of some kbp in size and, due to their repetitive nature, can mediate ectopic recombination events leading to even larger variants [17].

The recent activity of TEs is mainly responsible for the high level of structural variation that is observed in many angiosperms such as maize [18,19], so that a large fraction of TE insertions are not shared among related species nor among conspecific individuals [20,21]. While all angiosperms sequenced so far are characterized by an evolutionarily recent activity of TEs that would presumably result in TE-related SVs being present at some level in these species, the recent genomic sequence of the first gymnosperm species, Norway spruce [22••] has revealed a very ancient TE component with most elements having inserted more than 5 Myrs ago and being shared with the related species white spruce. This provides a very different scenario for gymnosperms, where we do not expect to find TEs as the main source of SVs. Other mechanisms for the formation of SVs include non-allelic homologous recombination (NAHR) between low copy repeats (LCRs) generated during ancient duplication events [23], and Double Strand Break (DSB) and Single Strand Annealing (SSA) [14•].

Methods for the identification of SVs at a genome-wide level include microarrays, such as array comparative genomic hybridization (array CGH) and SNP arrays, and NGS. Microarray-based methods infer copy number variations by hybridizing the DNA of a sample or a population to a set of probes representing the reference genome. The main limitations of these technologies consist in the impossibility to detect copy-number differences of sequences that are not represented in the reference, the inability to provide information on the location of duplicated sequences and the low resolution in defining the breakpoints of the variations. These limitations are being overcome by NGS technologies that promise to revolutionize structural variation studies [8]. SVs can be detected from NGS data with four different approaches (Figure 1): three of them (read-pair, read-depth and split-read methods) rely on the alignment of the short reads to a reference genome, enabling the detection of different types of variations (insertions, deletions, duplications, inversions) that involve sequences represented in the

Figure 1



NGS signatures for PAVs and CNVs detection. **(a)** Paired-end mapping (PEM) signature utilizes the discordance from the expected span size and/or orientation of mapped paired-end reads to predict SVs at a resolution that is proportional to the insert-size, the read length and the physical coverage; this signal can be successfully employed for the detection of deletions and duplications and of insertions of known sequences such as annotated transposable elements (TEs), **(b)** depth of coverage (DOC) signature analyzes the local increase and decrease in sequence coverage to identify SVs that alter the number of copies of a sequence such as deletions and duplications. A slight decrease in coverage is usually observed in correspondence to a new insertion, but the signal is too weak to be captured by DOC signature, **(c)** split-read methods utilizes the 'splitted' alignment of reads spanning the breakpoint of the variant to predict different type of SVs at single base level. This strategy requires longer reads than the other methods and is less efficient in the detection of SVs in high repetitive genomic regions and **(d)** strategies on the basis of *de novo* assembly can, in principle, identify at the breakpoint resolution all type of SVs; however the assembly on the basis of short reads is still challenging especially in correspondence of repeats thus limiting the detection of variants.